

T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANA BİLİM DALI



MAKİNE ÖĞRENMESİNDE SINIF DENGESİZLİĞİ İÇİN
BELİRSİZLİK YÖNLENDİRMELİ YENİDEN ÖRNEKLEME

Doktora Tezi

Fatih SAĞLAM

Danışman
Prof. Dr. Mehmet Ali CENGİZ

SAMSUN
2024

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım Doktora tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar'da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

Etik Kurul Gerekli mi?

Evet ☐ (Gerekli ise ekler kısmına ekleyiniz)

Hayır ☒

16/05/2024
Fatih SAĞLAM

TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı: MAKİNE ÖĞRENMESİNDE SINIF DENGESİZLİĞİ İÇİN
BELİRSİZLİK YÖNLENDİRMELİ YENİDEN ÖRNEKLEME

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 16/05/2024 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 5

Tek kaynak oranı : % 1 çıkmıştır.

16/05/2024
Prof. Dr. Mehmet Ali CENGİZ

ÖZET

MAKİNE ÖĞRENMESİNDE SINIF DENGESİZLİĞİ İÇİN BELİRSİZLİK YÖNLENDİRMELİ YENİDEN ÖRNEKLEME

Fatih SAĞLAM

Ondokuz Mayıs Üniversitesi

Lisansüstü Eğitim Enstitüsü

İstatistik Ana Bilim Dalı

Doktora, Şubat/2024

Danışman: Prof. Dr. Mehmet Ali CENGİZ

Makine öğrenmede sınıflama yöntemlerinin çoğu, sınıflara ait örnek sayılarının aynı olduğunu varsaymaktadır. Farklı örnek sayıları ile model kurulduğunda, sınıflama modelleri veri sayısı fazla olan sınıftan yana yanlılık göstermektedir. Böyle durumlarda kullanılan yöntemlerden birisi yeniden örneklemedir. Yeniden örnekleme ile veri üretilerek veya veri silinerek sınıflar arası denge sağlanır. Bu esnada seçilen örneklerin seçimi genellikle rastgeledir. Rastgele örnek seçimi yerine, bilgilendirici örneklerin seçilmesi daha az veri üretimi ile daha bilgilendirici veri setleri oluşturmayı sağlar. Rastgele örnek seçimi, gürültü örnekleri bölgesinde veri üretimine de yol açar. Ayrıca gürültü bulunmayan veri setinde yeni veri üretirken sınıflar arası kesişim bölgesinde veri üreterek gürültü problemine yol açabilmektedir. Bu çalışmada, aktif öğrenme kullanılarak, bilgilendirici örneklerin seçildiği belirsizlik yönlendirmeli aşırı örnekleme (UGO) yöntemi önerilmiştir. Aktif öğrenme, bilgilendirici örnekleri seçerek insan yardımıyla veri seti oluşturmak için kullanılan bir yaklaşımdır. UGO, modifiye edilmiş belirsizlik öğrenmesi ile seçilen örnekler üzerinden insana gerek olmadan sentetik veri üretimi yoluyla sınıf dengesizliğini ortadan kaldırır. İlk olarak veri üretimi öncesinde gürültü örneklerini tespit ederek belirsizlik tipinin yüksek derecede aleatorik olduğu bölgelerde veri üretilmediğinden emin olur. Epistemik belirsizliğin yüksek olduğu verilere odaklanarak bilgilendirici örnekler üretilmesini sağlar. Sınıflar arasındaki örnek sayısı farklarına bakarak dengesizliği kontrol etmek yerine, epistemik belirsizlik ortalamasını karşılaştırır. Daha sonra, veri üretim aşamasında, yeni üretilen verileri de gürültü kontrolünden geçirir ve gürültü verilerinin yerini güvenli bölgelere taşır. Sınıf epistemik belirsizlik ortalamaları arasında anlamlı fark kalmayana kadar iteratif biçimde veri üretir. UGO, gürültü üretimine neden olmaz, otomatik biçimde üretilecek veri sayısını kendi belirler ve sınıf dengesizliğini ortadan kaldırır. Yöntemin diğer yeniden örnekleme yöntemleri ile karşılaştırmalı performansları 2 simülasyon, 74 iki sınıflı ve 8 ikiden fazla sınıflı veri setinde karşılaştırılmıştır. Dengesiz doğruluk ölçütüne (IAM) göre, simülasyon ve çok sınıflı durumlarda istatistiksel olarak anlamlı başarı göstermiştir. İkiden fazla sınıflı veri setlerinde ise, sıra ortalaması daha iyi olsa da istatistiksel olarak anlamlı bir başarı göstermemiştir.

Anahtar Sözcükler: Sınıf dengesizliği, yeniden örnekleme, aktif öğrenme, belirsizlik tahmini, gürültü

ABSTRACT

UNCERTAINTY GUIDED RESAMPLING FOR CLASS IMBALANCE IN MACHINE LEARNING

Fatih SAĞLAM

Ondokuz Mayıs University

Institute of Graduate Studies

Department of Statistics

Ph.D., February/2024

Supervisor: Prof. Dr. Mehmet Ali CENGİZ

Most classification methods in machine learning assume that the number of examples of classes is the same. When a model is built with different numbers of samples, classification models show bias in favor of the class with a larger number of data. One of the methods used in such cases is resampling. Balance between classes is achieved by synthetic data generation or deleting data. The selection of samples chosen while doing this is usually random. Choosing informative samples instead of random sample selection allows creating more informative data sets with less data production. Random sample selection also leads to data generation in the region of noise samples. In addition, while producing new data in a data set that does not contain noise, it may cause a noise problem by producing data in the intersection area between classes. In this study, uncertainty guided oversampling (UGO) method is proposed, where informative samples are selected using active learning. Active learning is an approach used to create a dataset with human assistance by selecting informative examples. UGO eliminates class imbalance by generating synthetic data from selected examples with modified active learning, without the need for humans. First, it detects noise samples before data generation and makes certain that no data is generated in regions where the uncertainty type is highly aleatoric. Instead, it allows generating informative examples by focusing on data with high epistemic uncertainty. Instead of checking for imbalance by looking at differences in the number of examples between classes, it compares the average epistemic uncertainty. Then, during the data generating phase, it also checks the newly produced data for noise and relocates the noise data to safe regions. It iteratively derives data until there is no significant difference between class epistemic uncertainty means. As a result, UGO does not cause noise generation, automatically determines the number of data to be generated and eliminates class imbalance. The comparative performances of the method with other resampling methods were compared on 2 simulation, 74 two-class and 8 more than two-class data sets. According to the imbalanced accuracy measure (IAM) criterion, UGO showed statistically significant success in simulation and multi-class situations. In data sets with more than two classes, although the rank average was better, it did not show a statistically significant success.

Keywords: Machine learning, class imbalance, resampling, active learning, uncertainty estimation

ÖN SÖZ VE TEŞEKKÜR

Akademik eğitim sürecimin bir üst noktası olan doktora sürecinde yardımları için danışman hocam Prof. Dr. Mehmet Ali CENGİZ’e teşekkür ederim.

Bugünlere gelmemde en büyük pay sahibi olan aileme, hayatımın her aşamasında bana göstermiş olduğu sevgi, ilgi ve destek için minnettarım. Zor zamanlarımda bana güç verdiniz, başarılarımda ise en büyük sevincimi sizlerle paylaştım. Varlığınız ve desteğiniz olmadan bu doktora tezini tamamlayabilmem mümkün değildi.

İyi ve kötü günlerimde her zaman yanımda olan, bana inanarak ve motive ederek her daim destek olan can yoldaşım Filiz KOÇOĞLU’na sonsuz teşekkürlerimi sunmak istiyorum. Sevgin, ilgin ve sabrın olmadan bu zorlu süreci atlatmam imkansızdı. Sen benim en büyük gücüm ve ilham kaynağımsın.

Bu vesileyle, doktora tezimi tamamlamamda emeği geçen tüm arkadaşlarıma ve dostlarıma da şükranlarımı sunmak isterim.

Ve en önemlisi, bu zorlu ve uzun yolda pes etmeden ilerlememe ilham veren, her zaman kendime inanan ve potansiyelimi fark eden kendime teşekkür etmek istiyorum. Bu süreçte gösterdiğim özveri, çalışma azmi ve kararlılık beni bu noktaya taşıdı.

Fatih SAĞLAM

İÇİNDEKİLER

TEZ KABUL VE ONAYI	i
BİLİMSEL ETİĞE UYGUNLUK BEYANI	ii
TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI	ii
ÖZET	iii
ABSTRACT	iv
ÖNSÖZ VE TEŞEKKÜR	v
İÇİNDEKİLER	vii
SİMGELER VE KISALTMALAR	vii
ŞEKİLLER DİZİNİ	ix
TABLolar DİZİNİ	xi
1. GİRİŞ	1
1.1. Araştırma Konusu.....	1
1.2. Problem Tanımı	9
1.3. Önerilen Yöntem	12
1.4. Önerilen Yöntemin Katkıları ve Dezavantajları.....	14
2. YENİDEN ÖRNEKLEME	16
2.1. Aşırı örnekleme	16
2.1.1. Rastgele Aşırı Örnekleme	16
2.1.2. Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE).....	17
2.1.3. Rastgele Yürüyüş Aşırı Örnekleme (RWO)	18
2.2. Az Örnekleme.....	21
2.2.1. Rastgele Az Örnekleme	21
2.3. Hibrit Yaklaşımlar	21
2.4. İki Sınıflı Durumlarda Yeniden Örnekleme	21
3. AKTİF ÖRNEKLEME	22
3.1. Aktif Öğrenme	22
3.2. Belirsizlik Tahmini.....	24
4. ÖNERİLEN YÖNTEM	28
4.1. Modifiye Edilmiş Belirsizlik Örnekleme.....	29
4.2. Modifiye Edilmiş Marjın Belirsizlik	30
4.3. Durdurma Kriteri	33
4.4. Belirsizlik Yönlendirmeli Aşırı Örnekleme (UGO)	33
4.5. Örnekler.....	39
4.6. Gürültü Problemi Durumunda UGO	41
5. DENEYSEL TASARIM	45
5.1. Kullanılan Veri Setleri.....	45
5.1.1. Simülasyon Veri Setleri.....	45
5.1.2. İki Sınıflı Veri Setleri	47
5.1.3. İki Sınıflı Fazla Sınıflı Veri Setleri	49
5.2. Karşılaştırılan Yöntemler	50
5.3. Değerlendirme Ölçütleri.....	50
5.4. Deney Tasarım Uygulama Protokolü.....	51
6. DENEYSEL BULGULAR	54
6.1. Deney Sonuçları	54
6.1.1. Simülasyon Sonuçları	54
6.1.2. İki Sınıflı Veri Setleri için Sonuçlar	65
6.1.3. İki Sınıflı Fazla Sınıflı Veri Setleri için Sonuçlar.....	67
6.1.4. Veri Setlerinde UGO'nun Belirsizlik Üzerine Etkisi	70

7. TARTIŞMA.....	77
7.1. Deney Sonuçlarının Değerlendirilmesi	77
7.2. Yöntemin Sınırlamaları	80
7.3. Gelecekteki Çalışmalar için Öneriler	80
KAYNAKLAR	82
EKLER.....	91
Ek 1. İki sınıflı veri setlerinde genel sınıflayıcılar için IAM ortalamaları	91
Ek 2. İki sınıflı veri setlerinde genel sınıflayıcılar için IAM sıra ortalamaları	92
Ek 3. İki sınıflı veri setlerinde lineer sınıflayıcılar için IAM ortalamaları.....	94
Ek 4. İki sınıflı veri setlerinde lineer sınıflayıcılar için IAM sıra ortalamaları.....	95
Ek 5. İki sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM ortalamaları	97
Ek 6. İki sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM sıra ortalamaları.....	98
Ek 7. İki sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM ortalamaları.....	100
Ek 8. İki sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM sıra ortalamaları	101
Ek 9. İki sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM ortalamaları ..	103
Ek 10. İki sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM sıra ortalamaları.....	104
Ek 11. İki sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM ortalamaları.....	106
Ek 12. İki sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM sıra ortalamaları.....	107
Ek 13. İki sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM ortalamaları.....	109
Ek 14. İki sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM sıra ortalamaları.....	110
Ek 15. İki sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM ortalamaları	112
Ek 16. İki sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM sıra ortalamaları.....	113
Ek 17. İki sınıflı veri setlerinde genel sınıflayıcılar için IAM ortalamaları.....	115
Ek 18. İki sınıflı veri setlerinde genel sınıflayıcılar için IAM sıra ortalamaları.....	115
Ek 19. İki sınıflı veri setlerinde lineer sınıflayıcılar için IAM ortalamaları.....	115
Ek 20. İki sınıflı veri setlerinde lineer sınıflayıcılar için IAM sıra ortalamaları.....	116
Ek 21. İki sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM ortalamaları.....	116
Ek 22. İki sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM sıra ortalamaları.....	116
Ek 23. İki sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM ortalamaları.....	117
Ek 24. İki sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM sıra ortalamaları.....	117
Ek 25. İki sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM ortalamaları.....	117
Ek 26. İki sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM sıra ortalamaları.....	118

Ek 27. İkiiden fazla sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM ortalamaları	118
Ek 28. İkiiden fazla sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM sıra ortalamaları	118
Ek 29. İkiiden fazla sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM ortalamaları	119
Ek 30. İkiiden fazla sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM sıra ortalamaları	119
Ek 31. İkiiden fazla sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM ortalamaları.....	119
Ek 32. İkiiden fazla sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM sıra ortalamaları.....	120
ÖZ GEÇMİŞ.....	121



SİMGELER VE KISALTMALAR

UGO	: Aktif aşırı örnekleme
ACC	: Doğruluk oranı
AUC	: ROC eğrisi altında kalan alan
BART	: Bayesci toplamsal regresyon ağaçları
BNNET	: Bayesci sinir ağı
CART	: Sınıflama ve regresyon ağaçları
GKNB	: Gauss kernel naif Bayes
GMEAN	: Geometrik ortalama
GNB	: Gauss naif Bayes
IAM	: Dengesiz doğruluk ölçüsü
KNN	: k en yakın komşuluk
LDA	: Lineer diskriminant analizi
LogReg	: Lojistik regresyon
NNET	: Sinir ağları
QDA	: Kuadratik diskriminant analizi
RF	: Rastgele orman
ROS	: Rastgele aşırı örnekleme
RUS	: Rastgele az örnekleme
RWO	: Rastgele yürüyüş aşırı örnekleme
SMOTE	: Sentetik azınlık aşırı örnekleme tekniği
SVMLineer	: Lineer kernel destek vektör makineleri
SVMRadyal	: Radyal kernel destek vektör makineleri

ŞEKİLLER DİZİNİ

Şekil 1.1 Yansızlık miktarına göre sınıflayıcı tahmini değişimi destek vektör makinaları örneği	4
Şekil 1.2 Aşırı örneklemede gürültü problemi SMOTE örneği. a) dengesiz ve gürültüsüz veri seti, b) (a)'daki veri setine SMOTE uyguladıktan sonra, c) dengesiz ve gürültülü veri seti, d) (c)'deki veri setine SMOTE uyguladıktan sonra, e) dengesiz ve gürültüsüz veri seti, f) (e)'deki veri setine SMOTE uyguladıktan sonra.....	10
Şekil 2.1 Yeniden örnekleme türlerinin çalışma şekilleri	16
Şekil 2.2 $k = 3$ en yakın komşuya göre SMOTE ile veri üretimi süreci örneği.....	18
Şekil 2.3 RWO ile veri üretim süreci örneği.....	20
Şekil 3.1 Göl temelli örnekleme senaryosu	23
Şekil 4.1 Yeniden örnekleme için modifiye edilmiş göl temelli belirsizlik öğrenme stratejisi	30
Şekil 4.2 MA and MMA örnek üzerinde karşılaştırma.....	32
Şekil 4.3 UGO algoritması akış şeması	38
Şekil 4.4 UGOSMOTE'a ait ilk iterasyonun NNET örneği. a) Dengesiz sınıflı veri setinde gürültü örneklerinin tespiti, b) eğitilen NNET modelinin belirsizliklerine göre boyutlandırılan örnekler, belirsizlik ortalamaları, ortalama fark hipotez testi sonucu ve belirsizliklere göre seçilen örnekler, c) veri üretimi sonrası gürültü tespiti ve tespit edilen gürültülerin konum değiştirmesi	40
Şekil 4.5 Şekil 4.3'teki verinin farklı yeniden örnekleme yöntemleri uygulandıktan sonraki son halleri. UGO yöntemleri için sınıflayıcı yapay sinir ağlarıdır.....	41
Şekil 4.6 Gürültü ile performans ilişkisi için simülasyon verisi	42
Şekil 4.7 Lojistik regresyon gürültü ile performans arasındaki ilişki	43
Şekil 4.8 CART gürültü ile performans arasındaki ilişki.....	43
Şekil 4.9 GNB gürültü ile performans arasındaki ilişki.....	44
Şekil 4.10 NNET gürültü ile performans arasındaki ilişki.....	44
Şekil 5.1 Lineer olarak ayrıştırılabilen ve gürültü içeren veri seti.....	46
Şekil 5.2 Non-lineer olarak ayrıştırılabilen ve gürültü içeren veri seti	47
Şekil 6.1 Lineer ayrışan veri seti SVMLineer kontör grafikleri	55
Şekil 6.2 Lineer ayrışan veri seti LogReg kontör grafikleri.....	56
Şekil 6.3 Lineer ayrışan veri seti LDA kontör grafikleri	57
Şekil 6.4 Non-lineer ayrışan veri seti SVMRadyal kontör grafikleri.....	59
Şekil 6.5 Non-lineer ayrışan veri seti BART kontör grafiği	60
Şekil 6.6 Non-lineer ayrışan veri seti NNET kontör grafikleri.....	61
Şekil 6.7 Non-lineer ayrışan veri seti BNNET kontör grafiği.	62
Şekil 6.8 Non-lineer ayrışan veri seti KNN kontör grafikleri.....	63
Şekil 6.9 Non-lineer ayrışan veri seti GKNB kontör grafikleri.	64
Şekil 6.10 LogReg için iki sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar.....	71

Şekil 6.11 KNN için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar	72
Şekil 6.12 GKNB için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar	73
Şekil 6.13 NNET için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar	74
Şekil 6.14 RF için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar	75
Şekil 6.15 LogReg için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar	75
Şekil 6.16 KNN için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar.....	76
Şekil 6.17 GKNB için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar.....	76
Şekil 6.18 NNET için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar.....	76
Şekil 6.19 RF için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar.....	76

TABLÖLAR DİZİNİ

Tablo 2.1 SMOTE sözde kodları (tam denge)	17
Tablo 2.2 RWO sözde kodları (tam denge).	20
Tablo 4.1 UGO'un aşırı örnekleme ve aktif öğrenmedeki eksikliklere getirdiği çözümler...	28
Tablo 4.2 UGO sözde kodları	39
Tablo 5.1 Kullanılan iki sınıflı veri setleri.....	48
Tablo 5.2 Kullanılan ikiden fazla sınıflı veri setleri	49
Tablo 5.3 Karmaşıklık matrisi	50
Tablo 5.4 Çalışmada kullanılan sınıflama yöntemlerinin gruplandırılması.....	52
Tablo 5.5 Sınıflama yöntemlerinin eğitiminde kullanılan hiperparametreler ve kullanıldıkları dil paketleri	53
Tablo 6.1 İki sınıflı veri setlerinde tüm alt gruplar için IAM sıra ortalamaları	67
Tablo 6.2 İki sınıflı veri setlerinde IAM değerleri için Friedman ve ikili karşılaştırma test sonuçları	67
Tablo 6.3 İki sınıflı veri setlerinde tüm alt gruplar için IAM sıra ortalamaları.....	69
Tablo 6.4 İki sınıflı veri setlerinde IAM değerleri için Friedman ve ikili karşılaştırma test sonuçları	70

1. GİRİŞ

1.1. Araştırma Konusu

Makine öğrenmesi, bilgisayarların verilerden öğrenmesini sağlayan yapay zekanın bir alt dalıdır. Makine öğrenmesinde, bilgisayarlar örüntü keşfeder ve bu örüntüleri kullanarak tahminler yapar ve/veya kararlar verirler. Makine öğrenmesi algoritmaları, farklı tür verilerden beslenebilir. Bunlara örnek olarak metin, görüntü, ses veya sensör verileri verilebilir. Algoritmalar, tahmin veya karar hesaplamak için bu verilerden örüntü ve ilişki ortaya çıkarırlar. Makine öğrenmesi, istatistiksel öğrenme olarak da adlandırılır (Hastie vd., 2009). Genel olarak üç başlık altında toplanabilir: danışmanlı öğrenme, danışmansız öğrenme ve yarı danışmanlı öğrenme. Danışmanlı öğrenmede, gerçek etikete ait bilgiler mevcuttur ve eğitim bu etiketlerin yapısını keşfederek yapılır. Danışmansız öğrenmede, gerçek etiketlere ait bilgi yoktur ve amaç veri seti içerisindeki yapıları ortaya çıkarmaktır. Yarı danışmanlı öğrenmede ise her iki öğrenme bir arada kullanılmaktadır.

Danışmanlı öğrenme, veri setinde mevcut olan bir veya daha fazla girdi değişkenlerin, çıktı değişkenler üzerindeki etkisinin incelenmesidir. Bağımsız özdeş dağılımlı $T = \{(x_{11}, \dots, x_{p1}, y_1), \dots, (x_{1n}, \dots, x_{pn}, y_n)\}$ örneklem seti olsun. $x_i = (x_{i1}, \dots, x_{ip})$ ve y_i , i -nci örneğe ait değerler olacak şekilde $g: X \rightarrow Y$ fonksiyonu ele alınır. $f: T \rightarrow \mathbb{R}$ fonksiyonu için $g(x)$ fonksiyonu

$$g(x) = \underset{y}{\operatorname{argmax}} f(x, y) \quad (1.1)$$

şeklinde yazılabilir. Burada $X_j = \{x_{j1}, \dots, x_{jn}\}$, j -nci değişken, nitelik, açıklayıcı değişken, bağımsız değişken vb. isimlerle adlandırılır (Hastie vd., 2009). Guyon ve Elisseeff (2003), “değişken” ve “nitelik” arasındaki farkı önışlem sürecine göre belirlemiştir. Saf veri setindeki girdiler “değişken” iken, çeşitli ön işlemler sonrası eğitimde kullanılmak üzere kullanılan girdiler “nitelik” olarak adlandırılmıştır. $Y = \{y_1, \dots, y_n\}$ ise bağımlı değişken, etiket değişkeni vb. gibi isimlerle adlandırılır. Bu çalışmada X_j için Guyon ve Elisseeff’in tabirini, Y için bağımlı değişken tabirini kullanacağız.

Olasılıksal modellerde Eşitlik 1.1,

$$g(x) = P(Y|X = x, \theta) \quad (1.2)$$

şeklinde yazılabilir. Burada θ , model parametre(leri)dir. Bağımlı değişkenin (Y) kategorik bir yapıda olması durumunda, nitelikleri (x_i) bilinen bir birimin hangi

kategoriye ait olduğunu belirleme sürecine sınıflama denir. Y bağımlı değişkeni içerisindeki kategorilere de sınıf adı verilir. Sınıflar $S = \{S_1, \dots, S_k\}$ sıralı olmayan şekilde sınıf vektörü ve k , sınıf sayısı olsun. Çoğu sınıflama yönteminde $\{P(S_1|X = x_i, \theta), \dots, P(S_k|X = x_i, \theta)\}$ olasılıkları tahmin edilir. Bu olasılıklar kullanılarak

$$\hat{y}_i = g(x_i) = \underset{j}{\operatorname{argmax}} \hat{P}(S_j|X = x_i, \theta), j = 1, \dots, k \quad (1.3)$$

şeklinde sınıf tahmini yapılır. Burada $\hat{P}(S_j|X = x_i, \theta)$, $P(S_j|X = x_i, \theta)$ olasılığının tahminidir. İki sınıflı bir durum için bu tahmin

$$\hat{y}_i = \begin{cases} S_1, & \hat{P}(S_1|X = x_i, \theta) > 0.5 \\ S_2, & \hat{P}(S_2|X = x_i, \theta) \geq 0.5 \text{ veya } \hat{P}(S_1|X = x_i, \theta) \leq 0.5 \end{cases} \quad (1.4)$$

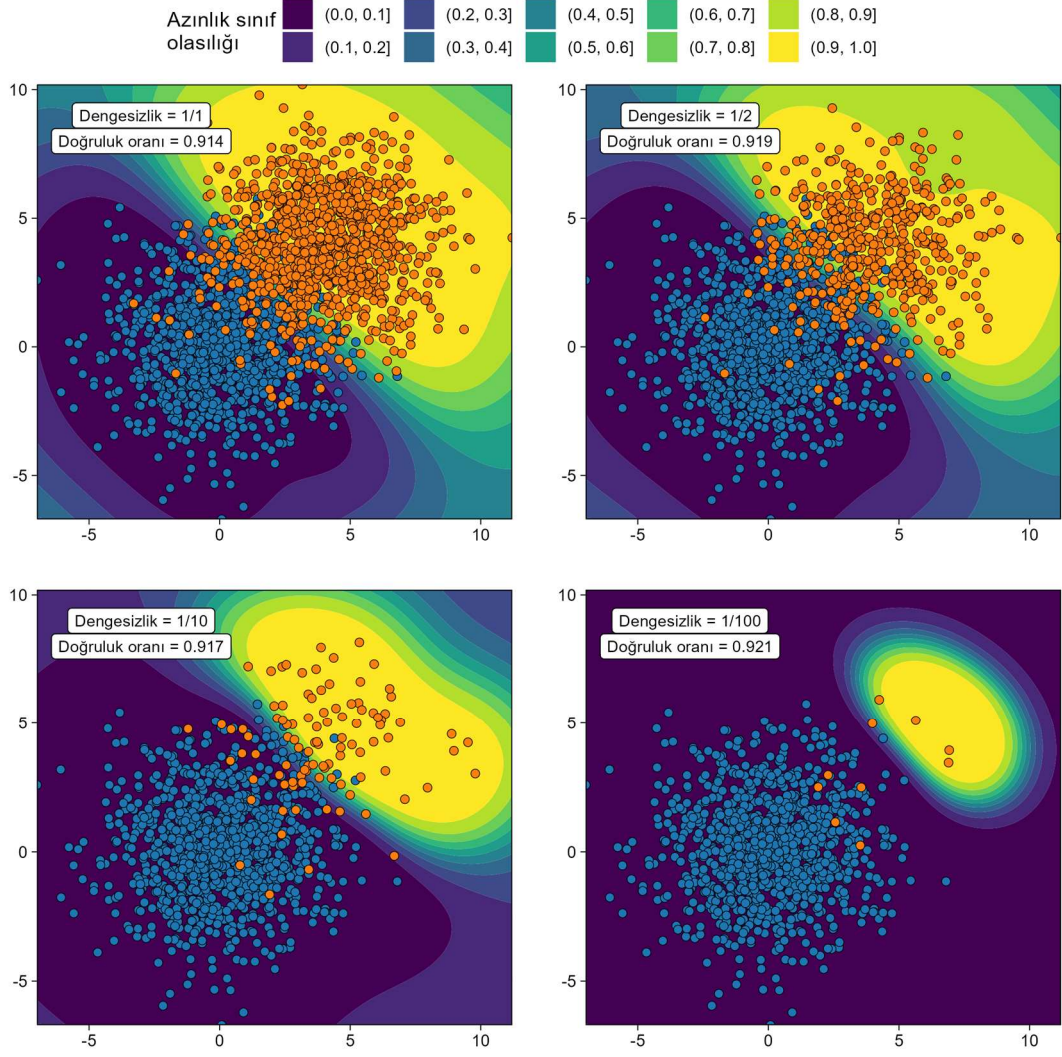
şeklinde dir. Doğru tahmin yapabilmek için veriye uyum sağlayan modele ait θ parametrelerinin tespit edilmesi gerekir. Bu, regresyonda olduğu gibi sınıflamada da kaybın minimizasyonu ile sağlanır. En bilindik kayıp fonksiyonu, çapraz entropi, diğer adıyla negatif log-olabilirlik değeridir. Çapraz entropi,

$$\text{Çapraz Entropi} = - \sum_{j=1}^k \sum_{i=1}^n y_i^{(j)} \log[\hat{P}(S_j|x_i, \theta)], \quad (1.5)$$

$$y_i^{(j)} = \begin{cases} 1, & y_i = S_j \\ 0, & y_i \neq S_j \end{cases} \quad (1.6)$$

şeklinde hesaplanır. Bunun dışında birçok kayıp fonksiyonu mevcuttur. Çapraz entropide görüldüğü üzere, tüm örneklerden gelen entropi bilgisinin toplamına göre θ model parametreleri belirlenir. Peki böyle bir durumda, sınıflardan birisi, diğerine göre daha fazla veriye sahip olursa ne olur? Sınıf sayısı fazla olan sınıftan yana tahminde bulunmak, az olan sınıftan yana tahminde bulunmaya göre daha başarılı gözükür modeller oluşturmaya neden olur. Bunun anlamı, eğitim süreci fazla örneğe sahip olan sınıftan yana yanlılık gösterecektir. Bu duruma sınıflar arası dengesizlik problemi adı verilir (Japkowicz, 2001; He and Garcia, 2009). Makine öğrenmesinde kullanılan sınıflama yöntemlerinin neredeyse hepsi, her ne kadar farklı kayıp fonksiyonlarını minimize etseler de sınıf dengesizliği probleminden etkilenmektedir (Rezvani and Wang, 2023). Sınıf dengesizliği problemine 950 sağlıklı ve 50 adet hasta kişiye ait veri seti örnek olarak verilebilir. Böyle bir veri setinde, tüm örnekler “sağlıklı” şeklinde tahminde bulunan bir sınıflayıcı %95 ACC’ye sahipken, “hasta” tahmininde bulunan sınıflayıcı %5 doğru tahmine sahip olacaktır. Bu da, sınıflayıcının eğitim sürecinde “sağlıklı” etiketinden yana yanlılık göstermesine neden olacaktır. Sınıf dengesizliğinin

gerçek hayat verilerinde sıklıkla karşılaşıyor olması da bu problemin önemini ortaya koymaktadır (Haixiang vd., 2017; Chan and Stolfo, 1998; Grzymala-Busse vd., 2000; Radivoja vd., 2004; Hilas and Mastorocostas, 2008). Sınıf dengesizliğinin nedenleri genelde zaman ve maddiyat kaybına neden olur. Uygulamalı bir örnek Şekil 1.1’de görülmektedir. Bu görselde lineer ayrışabilen iki sınıflı ve iki tesadüfi değişkenli bir veri seti simülasyonla oluşturulmuştur. Çoğunluk sınıf (mavi) örnekleri için her iki değişken örnekleri de $N(0,4)$, azınlık sınıf (turuncu) örnekleri için her iki değişken örnekleri $N(4,4)$ dağılıma sahiptir. Dengeli veri setinde her iki sınıftan da 1000 adet örnek vardır. Azınlık sınıf örneklerinden aşamalı olarak örnekler rastgele çıkartılarak dengesizlik oranı artırılmıştır. Şekil 1.1 incelendiğinde, dengesizlik oranı 1/1, 1/2, 1/10 ve 1/100 için farklı radyal kernel destek vektör makinaları sınıflayıcıları eğitilmiştir. Kontör grafiklerinde, renk barı sarıya doğru gittikçe, tahminlerdeki azınlık sınıf olasılıkları artmakta, koyu maviye gittikçe azalmaktadır. ACC ise, mevcut modelin doğru tahmin ettiği gözlem oranını vermektedir. Görüldüğü gibi dengesizlik arttıkça sınıflayıcı, azınlık sınıfa git gide daha az bir bölge tanımlamaktadır. Fakat, ACC’de ise hiçbir azalma olmamakta, hatta bir miktar artış bile görülmektedir. Bunun sebebi, çoğunluk olan sınıfın git gide daha doğru biliniyor olması ve önemsiz gözükken azınlık sınıfta yapılan hatanın, genel sonucu az etkilemesidir. Çünkü hatalı örnek azalmıştır. Buradan çıkarılacak sonuçlardan birincisi, sınıf dengesizliğinin performansı etkilediğidir. İkincisi ise ACC veya aynı anlama gelen hata oranı gibi sıklıkla kullanılan performans ölçütlerinin sınıf dengesizliği durumunda başarısız performans ölçmesidir. Sınıf dengesizliği durumunda kesinlikle alternatif ölçütler tercih edilmelidir. Alternatif ölçütlere Dördüncü bölümde daha ayrıntılı olarak değinilecektir.



Şekil 1.1 Yansızlık miktarına göre sınıflayıcı tahmini değişimi destek vektör makinaları örneği

Sınıf dengesizliği problemi çözmede kullanılan yöntemler yeniden örnekleme, maliyet duyarlı öğrenme ve topluluk algoritmaları şeklinde üç başlıkta toplanabilir (Haixiang vd., 2017).

Maliyet duyarlı öğrenmede, sınıflara farklı ağırlıkların verildiği bir maliyet matrisi oluşturulur. Bu matris ile ağırlıklı öğrenme yapılarak dengesizliğin üstesinden gelmek amaçlanır. Literatürde maliyet duyarlı öğrenme ile sınıf dengesizliği problemini çözen çalışmalar bulunmaktadır (Sun vd., 2020; Ren vd., 2022; Verbeke vd., 2023). Topluluk algoritmaları, genel olarak sınıf dengesizliğine karşı daha dayanıklıdır (Thölke vd., 2023). Bu nedenle araştırmacılar, sınıf dengesizliği durumunda topluluk algoritmalarını kullanmaktadır (Liu vd., 2024). Topluluk algoritmalarının sınıf dengesizliği durumunda kullanıldığı farklı çalışmaları görmek için Galar vd. (2011) derleme çalışması incelenebilir. Bunların dışında, yöntemlerin

bir arada kullanıldığı hibrit çalışmalar da mevcuttur. Chen vd. (2024), topluluk algoritmaları ile maliyet duyarlı öğrenmeyi bir arada kullanmışlardır.

Yeniden örnekleme, veri setindeki sınıflara ait veri miktarının artırılması veya azaltılması yoluyla denge sağlama yaklaşımıdır. Üç başlık altında toplanabilir: az örnekleme, aşırı örnekleme ve hibrit yöntemler. Az örneklemede, çoğunluk sınıfına ait örnekler çeşitli şekillerde seçilerek veri setinden atılır ve böylece denge sağlanmış olur. Rastgele az örneklemede, atılacak örnekler rastgele seçildiği gibi çıkarılacak olan örneklerin, zararlı veya az bilgilendirici olmasına dikkat edilebilir. Bu örnekler gürültü olarak adlandırılmaktadır. Az örnekleme genellikle mevcut gürültüden kurtulmak için kullanılır. Modelin kafasını karıştırabilecek örnekleri ortadan kaldıran gürültü temizleme yöntemleri de dahil olmak üzere, sınıf dengesizliğini gidermek için çeşitli az örnekleme yöntemleri mevcuttur. Tomek'in (1976) gürültü temizleme yöntemi, yoğunlaştırılmış en yakın komşu (CNN) kuralının (Hart, 1968) bir modifikasyonu olan TomekLink, en sık kullanılan yöntemlerden biridir (Krawczyk, 2016). Bir diğer popüler gürültü temizleme az örnekleme yöntemi, düzenlenmiş en yakın komşudur (ENN) (Wilson, 1972). ENN, en yakın komşulara göre riski hesaplayan ve risklerine göre gürültülü örnekleri belirleyen benzer bir yöntemdir. Bu yöntem sadece negatif gürültüleri değil aynı zamanda pozitif örnekleri de bulur. Kubat vd. (1997), dengesiz veri kümelerinde yalnızca negatif gürültülü örnekleri ortadan kaldıran tek taraflı seçim yöntemi önermiştir. Bu şekilde, tüm gürültü giderme yöntemleri bir azörnekleme yöntemi olabilir. Laurikkala (2001) komşuluk temizleme kuralı yönteminde ENN ile birlikte tek taraflı seçim yöntemini kullanmıştır. Zhang ve Mani (2003) sınıflandırıcı performansının az örnekleme yüzdesinden etkilendiğini göstermiş ve NearMiss-1, NearMiss-2 ve NearMiss-3 olarak adlandırılan ve yine en yakın komşuları temel alan üç yöntem önermişlerdir. Yöntemleri, F1 puanına göre diğer yöntemlere kıyasla başarılı sonuç vermiştir. Çoğunluk altında örnekleme tekniği (MUTE) (Bunkhumpornpat vd., 2011), en yakın komşuları kullanarak risk hesaplamasına dayanan başka bir yöntemdir. MUTE'yi önceki yöntemlerden ayıran şey, gürültüye karar verirken bir eşik parametresi kullanmalarıdır. Bunkhumpornpat ve arkadaşlarının (2017) yoğunluğa dayalı çoğunluk yetersiz örnekleme (DBMUTE), karar sınırındaki negatif örnekleri tespit eder ve pozitif veri kümesi merkezlerine çoğunluk örnek mesafesini kullanarak çıkarır. Tsai vd. (2019), negatif sınıf örneklerinin kümelenmesini ve veri kümesini filtrelemek için her sınıfa örnek seçim yöntemlerinin uygulanmasını içeren kümeleme tabanlı bir topluluk çerçevesi önerdi.

Yaklaşımları diğer yöntemlere kıyasla genel olarak daha iyi performans göstermiştir. Benzer şekilde Guzmán-Ponce vd. (2021), küme sayısını belirlemeye gerek kalmadan negatif sınıftaki aykırı değerleri tespit etmek için DBSCAN kümelemesini kullanan DBIG-US adı verilen kümeleme tabanlı bir az örnekleme yöntemi geliştirmiştir. Daha sonra bu aykırı değerleri kaldırarak az örnekleme yapmışlardır. Koziarski (2021), negatif sınıf örnekleri arasındaki bağlantılar üzerinden sentetik örnekler oluşturmuş ve denge sağlanana kadar uç noktaları verilerden çıkarmıştır. Yöntemlerine sentetik alt örnekleme tekniği (SMUTE) adını vermişlerdir. Bu çalışmaya benzer şekilde Dai vd. (2023) sınıfların örtüşen bölgelerine odaklanmıştır. Küresel benzerlikteki örtüşen örnekleri tespit etmek için matris ayrıştırmasını kullanmışlardır.

Az örneklemenin mevcut önemli bilgilerin kaybına yol açabilmesi nedeniyle literatürde aşırı örnekleme yöntemlerine, az örnekleme yöntemlerine göre daha fazla önem verilmiştir. En ünlü aşırı örnekleme yöntemi, Chawla ve arkadaşlarının SMOTE'sidir (2002). Han vd. (2005), yalnızca sınıf sınırlarında bağlantılar üreten Borderline SMOTE 1 (BLSMOTE1) ve Borderline SMOTE 2 (BLSMOTE2) adı verilen bir SMOTE varyasyonunu tanıtmıştır. Yöntemleri, F1 skoru ve gerçek pozitiflik oranına göre SMOTE'tan daha iyi performans göstermiştir. Uyarlanabilir sentetik örnekleme (ADASYN) (He vd., 2009), sınıf sınırlarına odaklanan başka bir SMOTE tabanlı yöntemdir. SVM-SMOTE (Nguyen vd., 2011), sınırları belirlemek için destek vektör makinelerinin (SVM) destek vektörlerini kullanır. Bunlarla sınıf sınırlarında yeni sentetik örnekler oluşturmada kullanılan bağlantıları belirlemiştir. Douzas vd., (2018), yeniden örnekleycinin gürültü üretmesini önlemek için k-ortalama kümeleme ve SMOTE kullanan bir aşırı örnekleme yöntemi (KmeansSMOTE) önermiştir. SMOTE, ADASYN, BLSMOTE, SVM-SMOTE ve KmeansSMOTE, Python programlama dilinin scikit-learn kütüphanesinde (Lemaître vd., 2017) mevcut oldukları için özellikle popülerdir. Pozitif örneklerin risk düzeylerini komşularına göre belirleyen ve yalnızca güvenli olanlarla bağlantı oluşturan Güvenli Düzey SMOTE (SLS) (Bunkhumpornpat vd., 2009) gibi gürültüsüz veri üretmeyi amaçlayan birkaç SMOTE tabanlı yöntem vardır. DBSMOTE (Bunkhumpornpat vd., 2012), bağlantılar aracılığıyla kümelerin merkezine giden bir yol oluşturmak için DBSCAN ve grafikleri kullanarak oluşturulan bağlantılar hakkında veri üreten, yoğunluğa dayalı bir yöntemdir. Barua vd. (2014), sınıflandırıcıların yeniden örneklemeden sonra daha hızlı öğrenmesini sağlamak için en yakın komşulara dayalı bir kümeleme yöntemi kullanarak pozitif örneklerin

ağırlıklarını belirleyen MWMOTE’u önermiştir. Yöntemleri GMEAN ve AUC açısından diğer yöntemlere göre daha iyi performans göstermiştir. Siriseriwan ve Sinapiromsaran (2016), güvenli seviye SMOTE’u (RSLS) yeniden konumlandırmayı kullanarak gürültüleri daha da güvenli hale getirmek için SLS’nin sentetik verilerinin yerini değiştirmiştir. Siriseriwan ve Sinapiromsaran (2017), belirlediği komşu sayısına göre örnekler üreten, uyarlanabilir komşu sentetik azınlık aşırı örnekleme (ANS) adı verilen başka bir yöntem önermiştir. Douzas ve Bacao (2019) geometrik SMOTE’u (GSMOTE) önermiştir. Bu yöntem, yeniden örnekleyicinin hiperparametrelerine dayalı olarak geometrik bir alan belirler ve buna dayalı olarak veri üretmektedir. Oldukça güçlü bir yöntemdir ancak hiperparametrelerine duyarlıdır. GSMOTE Python’da da mevcuttur. Radius-SMOTE (Pradipta vd., 2021), en yakın komşulara olan mesafeyi kullanır ve bu mesafenin bulunduğu bölgede yarıçap olarak örnekler oluşturur. Boosting ile SMOTE (SMOTEWB) (Sağlam and Cengiz, 2022), boosting yaklaşımını kullanarak gürültüleri belirler ve pozitif örnekleri iyi, tek başına ve kötü olarak sınıflandırır. Buna göre SMOTE ve ROS kombinasyonunu kullanarak örnekler üretir. SMOTEWB ayrıca her pozitif örnek için uygun sayıda komşunun seçilmesine yönelik bir yaklaşım da önermektedir. Bu, SMOTE tabanlı yöntemlerde k seçimini ortadan kaldırır. Dou vd.’nin NSS’si (2023), SMOTE ile oluştururken eksik verileri ele almıştır. Sentetik çoklu azınlık aşırı örnekleme (SMMO) (Damtew and Chen, 2023), veri üretirken önemli özellikleri dikkate almış ve yaklaşımları diğer yöntemlere göre daha iyi performans göstermiştir.

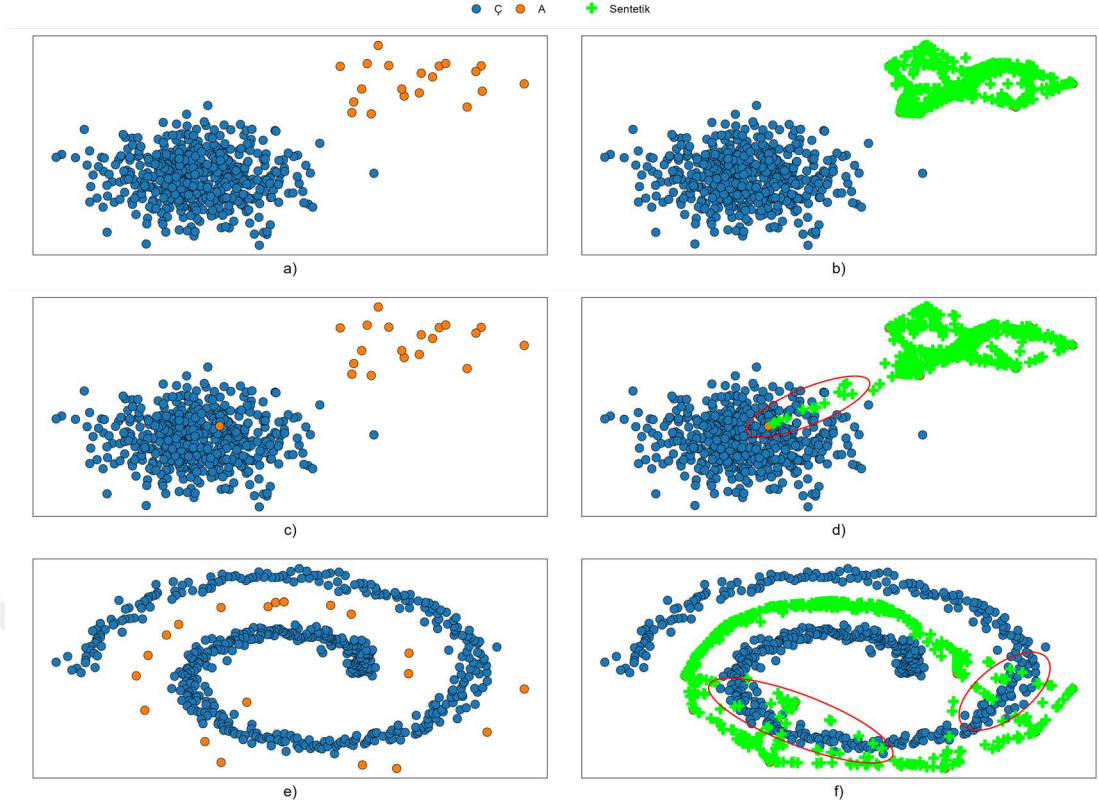
SMOTE tabanlı aşırı örneklemenin dışında örnekleme ve dağılıma dayalı çeşitli aşırı örnekleme çalışmaları bulunmaktadır. Lee (1999) yeni veriler üretmek için pozitif sınıf örneklerine Gauss gürültüsünü eklemiştir. Rastgele yürüyüş aşırı örnekleme (RWO) (Zhang and Li, 2014), orijinal verilerle aynı ortalamaya ve standart sapmaya sahip yeni veriler oluşturmak için değişkenlerin standart sapmasına dayalı olarak gürültü eklemiştir. Olasılık dağılım fonksiyonu tahminine dayalı aşırı örnekleme (PDFOS) (Gao vd., 2014), dengeyi sağlamak amacıyla yeni örnek oluşturmak için orijinal verilerin çekirdek yoğunluğunu kullanmıştır. Das vd. (2015), olasılık dağılımını tahmin etmek için Chow-Liu algoritmasını kullanan ve yeni veriler oluşturmak için Gibbs örneklemesini kullanan, RAPIDY CONVERGING Gibbs örneği (RACOG) adı verilen yeni bir aşırı örnekleme yöntemi önermiştir. Menardi ve Torelli (2014) ayrıca rastgele aşırı örnekleme örnekleri (ROSE) adı verilen yöntemlerinde olasılık dağılım fonksiyonunu belirlemek için çekirdek yoğunluğunu kullanmıştır.

ROSE'un ROS'tan daha seyrek yeniden örnekleyiciye geçmesine izin veren iki hiperparametre kullanmışlardır. ROSE ayrıca SMOTE, ROS ve RUS ile birlikte R programlama dilinin ünlü “caret” paketinde (Kuhn vd., 2020) mevcut olduğundan, iyi bilinen bir yöntemdir. SMOGN, Branco vd. (2017), SMOTE ve dağıtımına dayalı aşırı örnekleme bir karışımıdır. SMOGN, numunelere göre hangi yaklaşımın kullanılacağını belirler ve buna göre üretim yapar.

Hibrit yöntemler, aşırı örnekleme ve düşük örnekleme tekniklerinin bir kombinasyonudur ve aşırı örnekleme + az örnekleme, az örnekleme + aşırı örnekleme veya aşırı örnekleme sırasında az örnekleme gibi çeşitli biçimler alabilmektedir. Yaygın olarak kullanılan hibrit yöntemlerin örnekleri arasında SMOTE + TomekLink (Zeng vd., 2016), SMOTE + ENN (Lamari vd., 2021) ve SMOTE + NearMiss (Alamsyah vd., 2022) yer almaktadır. Aşırı örnekleme ve az örnekleme herhangi bir kombinasyonu hibrit bir yöntem olarak kullanılabilirken, spesifik aşırı örnekleme ve az örnekleme birlikte kullanılması da mümkündür. CORE (Bunkhumpornpat and Sinapiromsaran, 2015) önce gürültülü verileri kaldırır ve ardından aşırı örnekleme başlar; işbirlikçi olmayan oyun teorisini kullanarak aşırı örnekleme verileri filtrelerken (NEATER) (Almogahed and Kakadiaris, 2015) önerdikleri gürültü giderme yöntemini kullanarak önce aşırı örnekleme ve ardından az örnekleme yapar. Ağırlıklandırılmış İlişkiye Dayalı Kombinasyon Stratejisi (WERCS) (Branco vd., 2019), aşırı örneklemeden sonra az örnekleme yapan ve hem sınıflandırma hem de regresyon için kullanılabilen başka bir hibrit yöntemdir. Sowah vd. (2021), az örnekleme için kümelemeyi ve aşırı örnekleme için dışbükey kombinasyona (SNOCC) (Zheng vd., 2015) dayalı en yakın aşırı örnekleme kullanmıştır. Algoritmaları AUC, GMEAN ve Matthew korelasyon katsayısı (MCC) açısından diğer yöntemlerden daha iyi performans göstermiştir. Hussein vd. (2023), özellik alanının minimum ve maksimum değerleri arasında rastgele sentetik veriler üreten ve daha sonra daha iyi sınıflandırıcılar elde etmek için gereksiz ve kopyalanan örnekleri kaldıran, SVM'ye dayalı bir yeniden örnekleme yaklaşımı önermiştir. Kombine sentetik aşırı örnekleme ve düşük örnekleme tekniği (CSMOUTE) (Koziarski, 2021), aynı makalede önerdikleri SMUTE ve SMOTE'yu birlikte kullanmıştır. CSMOUTE alternatif yöntemlere karşı umut vaat etmiştir. Çeşitli kombinasyonlarda hibrit yeniden örnekleme popüler olmaya devam ediyor ve literatürde kullanılmaya devam etmektedir (Kou vd., 2022; Kumari and Subbarao, 2022; Puri and Kumar Gupta, 2022; Jiang vd., 2023; Park vd., 2023).

1.2. Problem Tanımı

Rezyani ve Wang'ın 2023'te yaptıkları dengesiz sınıf durumunda öğrenme için yaptıkları derleme çalışmasında, önerilen yöntemlerde hala birçok açıdan eksiklik olduğunu tespit etmişlerdir. Literatürdeki yeniden örnekleme yöntemleri incelendiğinde bazı problemler öne çıkmaktadır. Az örnekleme yöntemleri, bilgi kaybına neden olmaktadır ve çok fazla silinecek veri varsa, ters yönlü dengesizliğe yol açabilir. Aşırı örneklemede gürültü problemi vardır. Şekil 1.2, gürültü probleminin SMOTE yöntemi üzerindeki örneklerini göstermektedir. İki tür gürültü problemi görülmektedir. Şekil 1.2(a), gürültü problemi olmayan dengesiz sınıf oranına sahip bir veri setini, Şekil 1.2(b) ise SMOTE ile denge sağlandıktan sonraki veri setini göstermektedir. Burada mavi renkli noktalar çoğunluk (Ç), turuncu renkli noktalar azınlık (A) sınıfı ve yeşil renkli kalın "+" sentetik veriyi göstermektedir. Gürültü probleminin birincisi, gürültü verisi bulunan veri setlerinde daha fazla gürültü verisi üretilebiliyor olmasıdır. Şekil 1.2(c)'de biraz önceki veri setine 1 adet gürültü örneğinin eklendiği veri seti görülmektedir. Şekil 1.2(d), SMOTE ile veri üretildikten sonra, kırmızı daire ile gösterilen gürültüleri göstermektedir. Tek bir gürültü bile çok daha fazla yeni gürültü üretilmesine neden olmuştur. Gürültü olmayan veri setlerinde de problem devam etmektedir çünkü hiç var olmayan gürültülerin üretilmesi mümkündür. Şekil 1.2(e) ve 1.2(f)'de bu probleme örnek verilmiştir. Şekil 1.2(e)'deki dengesiz veri setinde hiçbir gürültü örneği bulunmamasına rağmen, Şekil 1.2(f)'de görülen SMOTE ile denge sağlanmış veri seti, çok miktarda yeni gürültü örneği içermektedir. Veri üretimi, mevcut verilerin kopyalıyor ve/veya bazı noktalarda fazla konsantre oluyorsa aşırı uyuma neden olabilir. Hibrit yöntemlerde çok fazla gürültü olduğunda, fazla gürültü silme gerçekleşmekte ve bu durumda ters yönlü dengesizliğe neden olabilmektedir. Veri üretim ve silme oranları doğru belirlenmelidir. Veri üretirken ne oranda veri üretilmesi gerektiği de bir başka problemidir. Ayrıca performanstaki kaybın dengesizlikten kaynaklanıp kaynaklanmadığı fazla araştırılmamıştır.



Şekil 1.2 Aşırı örneklemede gürültü problemi SMOTE örneği. a) dengersiz ve gürültüsüz veri seti, b) (a)'daki veri setine SMOTE uyguladıktan sonra, c) dengersiz ve gürültülü veri seti, d) (c)'deki veri setine SMOTE uyguladıktan sonra, e) dengersiz ve gürültüsüz veri seti, f) (e)'deki veri setine SMOTE uyguladıktan sonra.

Yeniden örneklemede yapılan örnekleme işlemi için, aktif öğrenme kullanılabilir. Aktif öğrenme, etiketsiz verilerdeki etiketleme maliyetini azaltmada kullanılan bir öğrenme yöntemi olduğu için, sınıf dengesizliğine çözüm bulmada ender kullanılmıştır. Aktif öğrenme kullanan yeniden örnekleme çalışmalarının öncülerinden olan Ertekin vd. (2007), aşırı örnekleme veya az örnekleme yapmak yerine aktif öğrenme kullanarak yeniden örnekleme yapıp veri seti oluşturmaktadır. Bunu yaparken, tüm modelde SVM eğiterek ulaşılması gereken destek vektör noktalarını belirlemektedirler. Daha sonra aktif öğrenme ile tüm sınıflardan belirsizlik ölçütüne göre örnekleme yapar. Tüm modeldeki destek vektörleri örnekleyene dek tekrar eder. Hız problemine çözüm bulmak için örneklem çekeceği “etiketsiz” veri sayısını kontrollü azaltır. Ertekin vd. (2007), hiperdüzleme yakın verilerin bilgilendirici olduğunu düşündüğü için test edilen etiketsiz verilerin, verinin tümünü içermesi gerekmediğini öne sürmüştür. Bu nedenle, hiperdüzleme en yakın %5 örnekten en az bir tanesini %95 olasılıkla içeren sayıda etiketsiz veri kullanmanın yeterli olduğunu düşünmüş ve bunu hesapladıklarında yaklaşık “59 numarası” nı (59

trick) hesaplamışlardır. Bu hesabın avantajı veri seti genişliğinden bağımsız olmasıdır. 59'ar etiketsiz veriden seçilen en bilgilendirici örnek, zaten seçilmesi gereken hiperdüzlem yakınlarındaki veriler olacaktır. Durdurma kriteri olarak da tüm veri setinde kurdukları modeldeki destek vektörleri kullanılmıştır. Tüm destek vektörler örneklem olarak çekildiklerinde aktif öğrenme durdurulur. Bu şekilde yeniden örnekleme yapıldığında az örnekleme problemi olan bilgi kaybı problemi yaşanmamış ve aşırı örnekleme ile gelen yapay veri artışının getirdiği yük ortadan kaldırılmış olur. Fakat Ertekin vd., veri setinin son halinin dengeli olacağına veya aktif öğrenme sürecinde çekilen örneklerin dengesine dair inceleme yapmamıştır. Bu konuda inceleme yapan Tomanek ve Hahn (2009), belirsizlik hesabında azınlık sınıfa fazla ağırlık veren bir ağırlıklı entropi ölçütü kullanmıştır. Ayrıca, aktif öğrenme esnasında uzmandan (oracle) gelen verileri, azınlık sınıf olması durumunda kopyalayarak aşırı örnekleme yapmıştır. Yöntemleri, dengesizliği dikkate almayan yöntemlere göre daha başarılı sonuçlar göstermiştir. Doucette ve Heywood (2008), çok daha basit bir aktif öğrenme yaklaşımını önerdikleri bir amaç fonksiyonuna göre kullanarak yeniden örnekleme yapmaktadır. Örnekleri seçerken, belirsizlik ölçmek yerine düzgün dağılıma göre seçmektedirler. Rastgele örneklemeden farkı, her iterasyonda, dengesizliği kontrol ederek dengeli bir veri seti oluşturduklarına emin olmaktadır. Ayrıca, ROC eğrisi altındaki alan (AUC) istatistiğine çok yakın olan Wilcoxon-Mann Whitney istatistiğini maksimize ederek veri setini belirlemektedirler.

Aktif öğrenmeyi yeniden örneklemede kullanan öncü çalışmaların yanında, daha güncel çalışmalar da mevcuttur. Yine Ertekin (2013), daha önceki (Ertekin vd., 2007) önerdiği yöntemi daha da geliştirerek, VIRTUAL isimli yeni bir yeniden örnekleme yöntemini önermiştir. Bu yöntem, diğer yöntemin aksine veri destek vektör makineleri (SVM) ile aynı çözüme yakınsayan LASVM algoritmasını kullanmaktadır. Bu algoritmanın farkı, normal sınıflayıcı oluştururken zaten aktif öğrenme mekanizmasını kullanması ve klasik SVM optimizasyon yöntemlerine göre daha hızlı olmasıdır. Aktif öğrenme mekanizmasını zaten kullandığı için, tek bir eğitim süresinde uygun bir durdurma kriteri ile yeniden örnekleme tamamlanabilir. Fu ve Lee (2013), destek vektör makinelerinde ayırıcı hiperdüzlemlerin aktif öğrenmede hala sınıf dengesizliğinden etkilendiğini göstermiştir. Ayrıca, gürültü probleminin de dikkate alınması gerektiğini belirtmiş ve seçilen örneklere olasılık ataması yapmıştır. Seçilen örneğin sorgulanması, bu olasılığa göre yanlı bir yazı tura atışı ile gerçekleşmektedir. Benzer bir başka çalışmada, Aggarwal vd. (2020), dengesiz sınıflı veri setlerinde aktif

öğrenme yaparken her örneklem çekilişinde sınıf sayısını kontrol ederek az olan sınıftan örnek çekmiştir ve sınıf dengesizliği durumunda başarılı performans elde etmiştir. Ertekin vd.'daki gibi aktif öğrenme ile örneklem çekilerek sınıf dengesizliğinden etkilenmeyen bir veri oluştururken, maruz kalınan veri seti dengesiz olduğu için bu dengesizliğe aktif öğrenme sürecinde dikkat etmek gerekir. Lee ve Seo'dur (2022), cezalandırılmış lojistik regresyon kullanarak aktif öğrenme ile azörnekleme yapmışlardır. Örnek seçmede entropi gibi belirsizlik ölçütleri yerine ortalama tahmin hata kare'si (mean squared estimation error, MSEE) kullanmışlardır.

Literatür incelendiğinde aşırı örneklemede aktif öğrenme kullanan yöntemlerde görülen eksiklikler aşağıdaki gibi özetlenebilir:

- Aşırı örnekleme yaparken mevcut veriden seçim yapılmaktadır. Bu da aşırı uyum sorununa yol açabilir. Örnek seçmek yerine, seçilen örnek üzerinden aşırı uyuma yol açmayacak biçimde sentetik veri üretimi ile bu problemi ortadan kaldırılabilir.
- Gürültülü örneklerde belirsizlik (aleatorik) yüksek olmasına rağmen, bilgilendirici değildir. Bu veriler üzerinden aşırı örnekleme, faydadan çok zarar getirecektir. Karar sınırını etkileyecek veriler veri üretimine dahil edilmemelidir.
- Sınıflar arası veri eksikliğinin gerçekten probleme yol açıp açmadığına dikkat edilmeden veri üretilmektedir. Bu, ters yönlü bir dengesizliğe yol açabilir. Dengesizliğin etkisi ölçülmelidir.
- En önemlisi, epistemik belirsizlik, sınıflayıcının veriyi nasıl öğrendiğine göre değişir ve sınıflayıcıdan sınıflayıcıya farklılık göstermektedir. Diğer çalışmalar bunu dikkate almamıştır. Sınıflayıcıya özgü belirsizlik dikkate alınmalıdır.

1.3. Önerilen Yöntem

Bu çalışmada, literatür çalışmaları ve tespit edilen eksiklikler dikkate alınarak aktif aşırı örnekleme (UGO) isimli yeni bir yeniden örnekleme yöntemi önerilmektedir. Bu yöntemde, modifiye edilmiş bir göl temelli aktif öğrenme yöntemi ile gürültü olmayan örnekler içerisinde örneklem çekilmekte ve bu çekilen örneklerden kullanıcının belirleyebildiği bir veri üretim yöntemi ile veri üretilmektedir. Modifiye edilmiş göl temelli aktif öğrenme yönteminde, klasik aktif öğrenmeden farklı olarak gerçek etiketlerin bilgisinden yararlanılmakta ve aleatorik belirsizliğin yüksek olduğu gürültülü örneklerin seçilmediğinden emin olunmaktadır.

Böylece sadece epistemik belirsizliğe odaklanılır. Epistemik belirsizliğin yüksek olduğu gürültülü örnekler, sınıf kesişim bölgelerindeki örnekler olduğu için gürültü tespitinde örneklerin komşulukları incelenmektedir. Belirli sayıda komşularının çoğunluğu, farklı sınıfa ait örneklerin kesişim bölgesindedir ve gürültüdür. En bilgilendirici örnek seçilirken, gürültü olmadığından emin olunduğu gibi örneklerin seçiminde kullanılan belirsizlik ölçütü olarak marjin yaklaşımının (Scheffer vd., 2001) mevcut senaryoya uyarlanmış hali kullanıldı. Bu yeni marjin yaklaşımında, eskisindeki gibi en yüksek olasılıklı iki sınıf arasındaki fark yerine, doğru etikete ait sınıf ile diğer en yüksek olasılıklı sınıf arasındaki fark kullanıldı. Böylece, gerçek etiket bilgisi kullanılarak sadece belirsizlik değil, doğru tahmine ne kadar uyum sağladığı da ölçülmektedir. Asıl önemlisi, belirsizliklerin hesabında hangi sınıflayıcı ile eğitim yapılacaksa o sınıflayıcı kullanılmaktadır. Böylece sınıflayıcıdan sınıflayıcıya değişecek olan epistemik belirsizliğin yaratacağı sorunun önüne geçilmiş olur. Epistemik belirsizliğin yüksek olması, belirsizliğin yüksek olduğu yerde veri eksikliği anlamına gelmektedir. Sınıflayıcı hangi bölgede veriye ihtiyaç duyuyorsa orada veri üretilmektedir. Bahsi geçen şekilde bir iterasyonda sınıflara ait gürültü olmayan verilerde belirsizlik ortalamaları hesaplanmaktadır. Sınıflara ait belirsizlik ortalamaları arasında anlamlı farklılık olup olmadığı her iterasyon sonucunda tespit edilir. Ölçülen belirsizlik epistemik belirsizlik olduğu için, ortalamalar arasında fark olmaması veri ihtiyacı arasında sınıflar arası fark olmadığı anlamına gelir. Bu nedenle veri üretimine gerek yoktur ve algoritma sona erer. Eğer belirsizlik ortalamaları arasında fark varsa, en yüksek belirsizlik ortalaması olan sınıftan veri üretilmesi gerektiğine karar verilir ve o sınıfın belirlenen oranda en belirsiz örnekleri seçilir. Veri üretimi yine bu örneklerle bağlı olarak belirlenen miktarda üretilir. Bu şekilde, gözlem sayılarına bakmadan sınıflar arasında dengesizlikten kaynaklanan bir problem olup olmadığı kontrol edilmiş olur. Sıradaki iterasyona geçmeden önce, son aşama olarak yeni gürültülü örnek üretilmediğinden emin olmak için üretilen verilerin veri setindeki komşulukları kontrol edilir. Eğer başlangıç gürültü tespitinde olduğu gibi, komşularının çoğunluğu farklı sınıf veri seti olan yeni üretilmiş veriler, gürültü verilerdir. Bu verileri silerek tekrar dengesizlik problemini ortaya çıkarma tehlikesi yaratmak yerine, bu veriler en yakın gürültü olmayan aynı sınıftaki veri noktasına taşınır, yani yer değiştirilir. Böylece gürültü üretilmediği kesinleşmiş olur ve sıradaki iterasyona geçilir. Veri üretimi, belirsizlik ortalamaları arasında anlamlı farklılık bulunmayana dek devam etmektedir. UGO, belirli bir veri üretim şekline sahip

olmadığı için bildiğimiz tüm yeniden örnekleme yöntemleri kullanılarak çalıştırılabilir.

Önerilen yöntemin katkılarını gösterebilmek için iki simülasyon ve literatürde kullanılan 82 adet gerçek ve yapay veri seti üzerinde deneyler planlandı. Simülasyon veri setlerinden birisi lineer ayrışabilen, diğeri ise lineer ayrışamayan fakat non-lineer biçimde ayrışabilen veri setleri olacak şekilde tasarlandı. Her iki simülasyon verisine, gürültü örnekleri eklenerek gürültü durumu altında yöntemlerin performansını inceleyebileceğimiz bir ortam oluşturuldu. Diğer 82 veri seti ise 74'ü iki sınıflı 8'i ikiden fazla sınıfa sahiptir. Böylece iki sınıflı ve ikiden fazla sınıflı ortamdaki performanslar da incelenebilecektir. Deneysel tasarımda önerilen yöntem (UGO) ile yeniden örneklemenin yapılmadığı durumlar, rastgele aşırı örnekleme (ROS), rastgele az örnekleme (RUS), SMOTE, RWO karşılaştırılacaktır. Önerilen yöntem, farklı veri üretme yöntemlerine göre farklılık göstereceği için tam karşılaştırma yapabilmek adına ROS, SMOTE ve RWO ile veri üretilerek üç farklı biçimde, UGOROS, UGOSMOTE ve UGORWO, kullanılacaktır. Performanslar 13 farklı sınıflama yöntemi üzerinden yapılacaktır. Bu sınıflayıcılar, Bayesci toplamsal regresyon ağaçları (Bayesian Additive Regression Trees, BART), Bayesci Sinir Ağları (Bayesian Neural Network, BNNET), C5.0, Sınıflama ve Regresyon Ağaçları (Classification and Regression Trees, CART), Gauss Kernel Naive Bayes, (GKNB), Gauss Naive Bayes (GNB), k En Yakın Komşuluk (k nearest neighbor, KNN), Lineer Diskriminant Analizi (LDA), Lojistik Regresyon (Logistic Regression, LogReg), Sinir Ağları (Neural Network, NNET), Kuadratik Diskriminant Analizi (Quadratic Discriminant Analysis, QDA), Rastgele Orman (Random Forest, randomForest), Lineer Kernel Destek Vektör Makineleri (Linear Kernel Support Vector Machines, SVMLineer), Radyal Kernel Destek Vektör Makineleri (Radial Kernel Support Vector Machines, SVMRadyal) şeklindedir. Bahsi geçen sınıflayıcılar, lineer, lineer-olmayan, sinir ağları, ağaç temelli, komşuluk temelli, yoğunluk temelli ve Bayesci yöntemler olmak üzere gruplandırılarak farklı hem genel hem de farklı sınıflama türlerindeki performansların incelenmesi amaçlanmıştır. Deneysel tasarım hakkında daha ayrıntılı bilgi Bölüm 5'te sunulmuştur.

1.4. Önerilen Yöntemin Katkıları ve Dezavantajları

UGO yöntemi ile literatürde tespit edilen eksiklikler ele alınmıştır. Önerilen yöntemin katkıları aşağıdaki gibi sıralanabilir:

- Aktif öğrenme sayesinde rastgele bölgelerde veri üretmek yerine, sadece

bilgilendirici veriler üretilir.

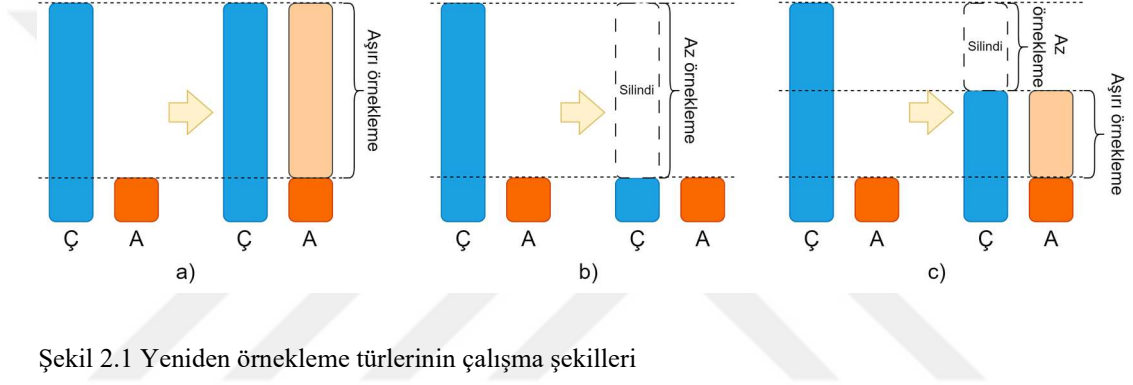
- Etkin sınıf dengesizliğinin varlığını ve miktarını veri eksikliğinin yarattığı epistemik belirsizliği dikkate alarak ölçer.
- Üretilen verilerin sınıflar arası kesişim bölgesinde üretilmediğinden emin olur.
- Ne zaman yeterince veri üretildiğine otomatik karar verir.
- Gerçek etiket bilgisinden faydalanan bir belirsizlik tahmini kullandığı için klasik aktif öğrenme yaklaşımlarına göre daha bilgilendirici örnekler çeker.

UGO ile bahsi geçen avantajları yakalamak için bazı fedakarlıklar vermek gerekmektedir. Önerilen yöntemin dezavantajları aşağıdaki gibi sıralanabilir:

- Sınıflayıcıya özgü şekilde daha uyumlu belirsizlik tahmini yaptığı için, tek bir eğitimde veri üretimini bitirmesi oldukça zordur. Her adımda model kurmaya ihtiyaç duyduğu için diğer yöntemlere göre daha uzun sürer. Bunun yaratacağı etkiyi azaltmak için, her iterasyonda daha fazla veri üretebilecek mekanizmaya sahiptir.
- Sınıflayıcıdan sınıflayıcıya farklı belirsizlik tahmini yaptığı için, yapılan belirsizlik tahminin kalitesi de sınıflayıcıya göre değişir. Kötü belirsizlik tahmini yapan (örneğin aşırı uyumlu olmaya çok meyilli) sınıflayıcılarda başarısız sonuçlar verebilir. Böyle durumlar, farklı benzer sınıflayıcılar kullanılarak problemin etkisi azaltılabilir. Örneğin: C4.5 karar ağacı için belirsizlik tahmininde rastgele orman kullanılması.
- Belirsizlik tahmininde olduğu gibi, gürültü tespiti de sınıflayıcıdan sınıflayıcıya farklılık göstermesi beklenir. Fakat önerilen yöntem sınıflayıcıya özgü gürültü tespiti yapmamaktadır. Bunun için gürültü tespiti yerine, topluluk algoritmaları ile birden fazla seçilen sınıflayıcı kullanılarak oylamalara bakılabilir. Böylece seçilen sınıflayıcıya özgü gürültülü veriler belirlenebilir. Fakat bu işlem maliyetini oldukça artıracaktır.

2. YENİDEN ÖRNEKLEME

Sınıf dengesizliği problemine çözüm olarak kullanılan yöntemlerden birisi yeniden örneklemedir. Yeniden örnekleme aşırı örnekleme, az örnekleme ve hibrit yaklaşımlar olmak üzere üç başlıkta toplanabilir. Şekil 2.1’de yeniden örnekleme türlerinin özetlenmiş süreçleri görülmektedir. Burada “A” azınlık sınıf, “Ç” çoğunluk sınıfı ifade etmektedir. Şekil 2.1(a)’da, aşırı örnekleme, Şekil 2.1(b)’de az örnekleme, Şekil 2.1(c)’de ise hibrit örnekleme gösterilmektedir. Az örneklemede, çoğunluk sınıftan örnekler çıkarılarak denge sağlanmak amaçlanmaktadır. Aşırı örneklemede ise denge, azınlık sınıfa veri eklenerek sağlanır. Hibrit yöntem ise az örnekleme ve aşırı örneklemenin birlikte kullanıldığı yaklaşımlardır.



Şekil 2.1 Yeniden örnekleme türlerinin çalışma şekilleri

2.1. Aşırı örnekleme

Aşırı örnekleme, örnek sayısı az olan sınıfa ait verilerin üretme veya kopyalama yoluyla veri setine eklenmesidir. Bu şekilde sınıf dengesizliğinin ortadan kaldırılması amaçlanır. Veri kopyalayarak aşırı örnekleme yöntemlerine örnek olarak rastgele aşırı örnekleme verilebilir. Komşular arası veri üretme yöntemi olarak sentetik azınlık aşırı örnekleme tekniği (SMOTE) ve veri setinin dağılımını dikkate alan aşırı örnekleme yöntemlerine de rastgele yürüyüş aşırı örnekleme (RWO) örnek gösterilebilir.

2.1.1. Rastgele Aşırı Örnekleme

Rastgele aşırı örnekleme, adından da anlaşılacağı gibi rastgele bir biçimde azınlık sınıftan örnekler seçerek onları kopyalama yoluna gider. Dengesizlik ortadan kalkana kadar kopyalama işlemi devam ettirilir. Basit bir yöntem olmasının yanı sıra, aynı noktalarda veri kopyaladığı için bu noktalara konsantre olacak şekilde veri üretmesi, aşırı uyum problemine neden olabilir.

2.1.2. Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE)

SMOTE, Chawla vd. (2002) tarafından önerilmiş bir yöntemdir. Belki de aşırı örnekleme yöntemlerinin en bilindik olanıdır. Azınlık sınıf örneklerinin en yakın azınlık sınıf komşularını belirler. Bu komşularla arasında lineer bir bağ oluşturur. Seçilen azınlık sınıf örneklerini kopyalamak yerine, en yakın komşuları ile arasında oluşturulan lineer bağ üzerinde rastgele veri üretir.

Şekil 2.2’de $k = 3$ en yakın komşuluk için SMOTE algoritması kullanarak veri üretiminin aşamalı bir örneği gösterilmektedir. Mavi renk çoğunluk, turuncu renk azınlık sınıfı, yeşil artı şekli de üretilen sentetik veriyi göstermektedir. Tırtıklı çizgi ise azınlık sınıf arasındaki komşuluklardır. $X^A \in X$ azınlık sınıf örneklerini, $X^C \in X$, çoğunluk sınıf örneklerini gösterebilir. $x_i^A \in X^A$ olacak şekilde $\tilde{x}_i^A$, x_i^A ’nın X^A içerisindeki en yakın k komşusundan rastgele seçilmiş olan bir komşusu olsun. λ , $Düzgün(0,1)$ dağılımlı tesadüfi bir sayı olacak şekilde $\hat{x}^A$ yeni sentetik verisi

$$\hat{x}^A = x_i^A + \lambda \times (\tilde{x}_i^A - x_i^A) \quad (2.1)$$

şeklinde üretilir. Bu işlem, veri setinde denge sağlanana dek tekrar edilir.

Tablo 2.1 SMOTE sözde kodları (tam denge)

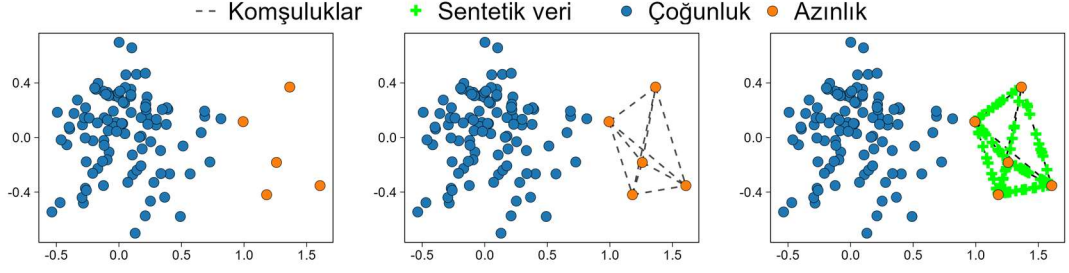
Girdi: $X_{n \times p}$, $Y_{n \times 1}$, k

1. $NN_{n^A \times k}^{A2A} \leftarrow enYakınKomşu(X^A, X^A, k)$
2. $n^{sen} \leftarrow n^C - n^A$
3. $C_{1 \times n^A} \leftarrow tekrarlar\left(\left\lceil \frac{n^{sen}}{n^A} \right\rceil, adet = n^A\right)$
4. $\hat{X}_{0 \times p}^{sen} \leftarrow$ boş matris
5. For $i = 1, \dots, n^A$
 - $k_{1 \times C_i}^{rast} \leftarrow örneklemÇek(1:k, adet = C_i, yerineKoymalı = TRUE)$
 - $NN_{1 \times C_i} \leftarrow NN_{i_{k^{rast}}^{A2A}}$
 - For $j = 1, \dots, C_i$
 - $\lambda_{1 \times 1} \leftarrow rastgeleUniform(a = 0, b = 1, adet = 1)$
 - $\tilde{x}_i^A \leftarrow X_{NN_j}^A$
 - $\hat{x}^A \leftarrow x_i^A + \lambda \times (\tilde{x}_i^A - x_i^A)$
 - $\hat{X}^A \leftarrow satırBirleştir(\hat{X}^A, \hat{x}^A)$
6. $X_{n^{yeni} \times p}^{yeni} \leftarrow satırBirleştir(\hat{X}_{n^S \times p}^A, X)$
7. $Y_{n^{yeni} \times 1}^{yeni} \leftarrow birleştir(tekrarla("A", adet = n^S), Y)$

Çıktı: $X_{n^{yeni} \times p}^{yeni}$, $Y_{n^{yeni} \times 1}^{yeni}$

Tablo 2.1, SMOTE algoritmasının sözde kodunu vermektedir. Algoritma takip edilerek, SMOTE ile denge sağlanacak veri seti üretilebilir. Algoritmanın girdisi,

$n \times p$ boyutlu X değişken matrisi, $n \times 1$ boyutlu Y bağımlı değişken vektörü ve kullanıcının belirleyeceği, genelde 5 değeri alan k en yakın komşuluk sayısıdır. NN^{A2A} tüm x_i^A için en yakın k komşu indisleri, n^{sen} üretilecek sentetik örnek sayısı ve her bir azınlık sınıf örneği üzerinden kaç adet örnek üretilmesi gerektiği belirlenir. En yakın komşuları arasından rastgele seçilen komşular kullanılarak her bir örnekten C_i adet sentetik örnek Eşitlik 2.1 kullanılarak üretilir ve yeni veri setine eklenir.



Şekil 2.2 $k = 3$ en yakın komşuya göre SMOTE ile veri üretimi süreci örneği.

$\hat{x}^A$, Şekil 2.2’de görülen komşular arasındaki doğru üzerinde üretilir ve dolayısıyla her zaman içe dönüktür. Eğitim setindeki azınlık sınıfa ait veri uzayının dışında asla veri üretmez. Bu nedenle, gerçekçi olmayan örnekler üretmez. Örnek vermek gerekirse, boy uzunluğu -2cm veya ağırlığı -2kg olan veriler üretilmesine neden olmaz. Fakat bu oluşturulan bağlar üzerinde veya bu bağlara çok yakın noktalarda, diğer sınıflara ait veriler bulunması mümkündür. Bu da normalde veri setinde olmayan gürültülü veriler üretmesine neden olabilir. Ayrıca, zaten mevcut olan gürültülü azınlık sınıf örneğinin çevresinde de veri üreteceği için çoğunluk sınıflara ait bölgelerde de azınlık sınıf verisi üretmesi olasıdır.

2.1.3. Rastgele Yürüyüş Aşırı Örneklemesi (RWO)

RWO (Zhang and Li, 2014), tek değişkenli biçimde veri üretir. Azınlık sınıflara ait verilerde her bir değişkenin standart hatasını kullanarak, seçilen örneklerin etrafında normal dağılımlı ve o değişkene ait standart hatanın karesi kadar varyanslı veri üretir. Bu yöntemin amacı, yeni üretilen verilerin ortalama ve varyansını, orijinal veri ile aynı değerlerde tutmaktır.

Aşırı örneklemede, bilinmeyen bir dağılıma sahip olan eğitim verisinde azınlık sınıf için veri üretilmesi gerekmektedir. RWO, yeni üretilen örneklerin eğitim verisi ile aynı dağılıma sahip olması için merkezi limit teoreminden faydalanır. Her bir x_i^A örneği, birden fazla değişken tarafından açıklanmaktadır ve RWO, bu değişkenlerin bağımsız olduğunu varsayar. Her bir değişkenin ortalama ve standart sapmaları μ_j^A ve

$\sigma_j^A, j = 1, \dots, p$ hesaplanır. j -nci değişkene ait gerçek ortalama ve standart sapma μ_j^A ve σ_j^A olacak şekilde n^A azınlık sınıf örnek sayısı sonsuza giderken

$$\frac{\mu_j^A - \mu_j'^A}{\sigma_j^A / \sqrt{n^A}} \sim N(0,1) \quad (2.2)$$

olur. Değişkenlerin bağımsız olduğu varsayımıyla Eşitlik 2.2 tüm değişkenler için geçerlidir. Eşitlik 2.2'de, $\mu_j'^A$ tek bırakılırsa

$$\mu_j'^A = \mu_j^A - r \times \frac{\sigma_j^A}{\sqrt{n^A}}, j = 1, \dots, p \quad (2.3)$$

şeklinde bilinmeyen dağılımın parametrelerine yaklaşılabılır. Burada $r, N(0,1)$ dağılımlı tesadüfi bir sayıdır. Buradan yola çıkarak

$$\hat{x}_{(j)}^A = x_{i(j)}^A + r \times \frac{\sigma_j^A}{\sqrt{n^A}}, i = 1, \dots, n^A, j = 1, \dots, p \quad (2.4)$$

şeklinde $\hat{x}^A$ yeni sentetik verisi üretilir. Burada $\hat{x}_{(j)}^A \sim N\left(x_{i(j)}^A, \frac{\sigma_j^A}{\sqrt{n^A}}\right)$, $\hat{x}^A$ 'nin j -nci değişken değeri, $x_{i(j)}^A$ de x_i^A 'nın j -nci değişken değeridir. Eşitlik 2.4 tüm değişkenler için ayrı ayrı uygulanır. Böylece ortaya $\hat{x}^A$ sentetik örneği çıkar.

Tablo 2.2, RWO algoritmasının sözde kodlarını vermektedir. Hiçbir hiperparametreye gerek duymayan RWO, X açıklayıcı değişken matrisini ve Y sınıf vektörünü girdi olarak alır. $n^{sen} = n^A - n^C$ adet veriyi Eşitlik 2.4'teki gibi ürettikten sonra X^{yeni} ve Y^{yeni} dengeli veri setini çıktı olarak verir.

RWO ile üretilen $\hat{x}^A$ için merkezi limit teoreminden dolayı n^A sonsuza giderken

$$E(\hat{x}_{(j)}^A) = \mu_j^A \quad (2.5)$$

ve

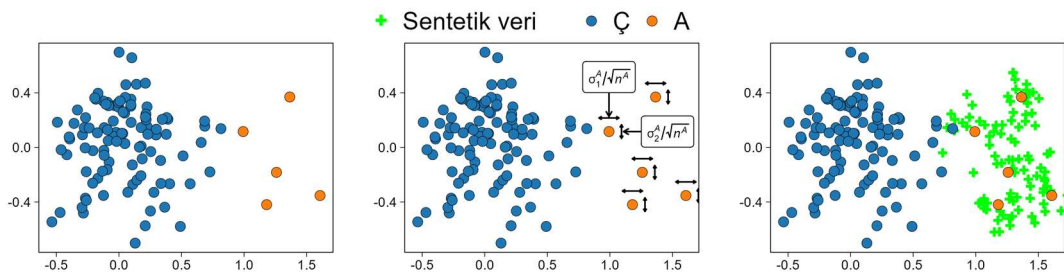
$$E(\hat{x}_{(j)}^A - \mu_j^A)^2 = (\sigma_j^A)^2 \quad (2.6)$$

olur ve ortalama ve varyanslar değişmemiş olur. Teorik ispatlarla ilgili ayrıntılı bilgi için Zhang ve Li (2014) incelenebilir. Gerçek veri setlerinde, n^A sonsuza gitmediği gibi azınlık sınıfıdır. Bu nedenle gerçek uygulamalarda ortalama ve standart sapmaların arasında anlamlı fark oluşup oluşmadığını test etmek için Zhang ve Li (2014) farklı örnek sayılarındaki uygulamalar yapmıştır. Uygulamalar sonucunda eğitim verisi ile sentetik veri setinin ortalama ve varyansları arasında istatistiksel olarak anlam olmadığını tespit etmişlerdir.

Tablo 2.2 RWO sözde kodları (tam denge).

Girdi: $X_{n \times p}, Y_{n \times 1}$ 1. $\mu_{1 \times p}^A \leftarrow p$ uzunluğunda boş vektör 2. $(\sigma^A)_{1 \times p}^2 \leftarrow p$ uzunluğunda boş vektör 3. For $j = 1, \dots, p$ $\mu_j^A \leftarrow \frac{1}{n^A} \sum_{i=1}^{n^A} x_{i(j)}^A$ $(\sigma_j^A)^2 \leftarrow \frac{1}{n^A} \sum_{i=1}^{n^A} (x_{i(j)}^A - \mu_j^A)^2$ 4. $\hat{X}_{0 \times p}^A \leftarrow$ boş matris 5. While $n^{sen} < n^C - n^A$ $i \leftarrow rastgeleseç(1, \dots, n^A)$ $\hat{x}_{1 \times p}^A \leftarrow 0_{1 \times p}, \text{ yeni veri için boş vektör}$ For $j = 1, \dots, p$ $r \leftarrow rastgeleNormal(ortalama = 0, standart sapma = 1, adet = 1)$ $\hat{x}_{(j)}^A = x_{i(j)}^A + r \times \frac{\sigma_j^A}{\sqrt{n^A}}$ $\hat{X}^A \leftarrow satırBirleştir(\hat{X}^A, \hat{x}^A)$ 6. $X_{n^{yeni} \times p}^{yeni} \leftarrow satırBirleştir(\hat{X}_{n^{sen} \times p}^A, X)$ 7. $Y_{n^{yeni} \times 1}^{yeni} \leftarrow birleştir(tekrarla("A", adet = n^{sen}), Y)$ Çıktı: $X_{n^{yeni} \times p}^{yeni}, Y_{n^{yeni} \times 1}^{yeni}$

RWO ile veri üretim süreci aşamaları Şekil 2.3'te verilmiştir. Mavi renk çoğunluk sınıfı, turuncu renk azınlık sınıfı, yeşil artı işaretleri de sentetik verileri göstermektedir. Önce, değişkenlere ait standart hatalar, yani $\sigma_j^A / \sqrt{n^A}$ değerleri hesaplanmıştır. Bu hesaplara göre tüm örnekler etrafında $N\left(x_{i(j)}^A, \frac{\sigma_j^A}{\sqrt{n^A}}\right)$ dağılımından üretilen $\hat{x}_{(j)}^A$ sentetik verileri üretilmiştir.



Şekil 2.3 RWO ile veri üretim süreci örneği.

RWO dağılım ortalama ve varyansını deęiřtirmese de, SMOTE gibi ie dnk veri retmez. Bu nedenle gereki olmayan deęerler retmesi mmkndr. Bu durum ayrıca grltl veri retmesi olasılıęını ortaya ıkarır.

2.2. Az rnekleme

Az rnekleme, oęunluk sınıf veya sınıflardan veri ıkartılarak denge saęlama yaklařımıdır. Veri sayısı azalacaęı iin eęitim sresini kısaltır. Ařırı uyum sorununa yol amaz. Fakat bilgi kaybına neden olur.

2.2.1. Rastgele Az rnekleme

Rastgele az rnekleme (ROS), adından da belli olduęu gibi az rneklemenin rastgele bir biimde yapılması iřlemidir. oęunluk sınıfından seilerek veri setinden ıkarılacak olan rnekler, rastgele biimde seilir.

2.3. Hibrit Yaklařımlar

Hibrit yntemler, az rnekleme ve ařırı rnekleme yntemlerinin bir arada kullanıldıęı yntemlerdir. Az rnekleme + ařırı rnekleme, ařırı rnekleme + az rnekleme veya ařırı rnekleme esnasında az rnekleme gibi farklı biimlerde olabilir.

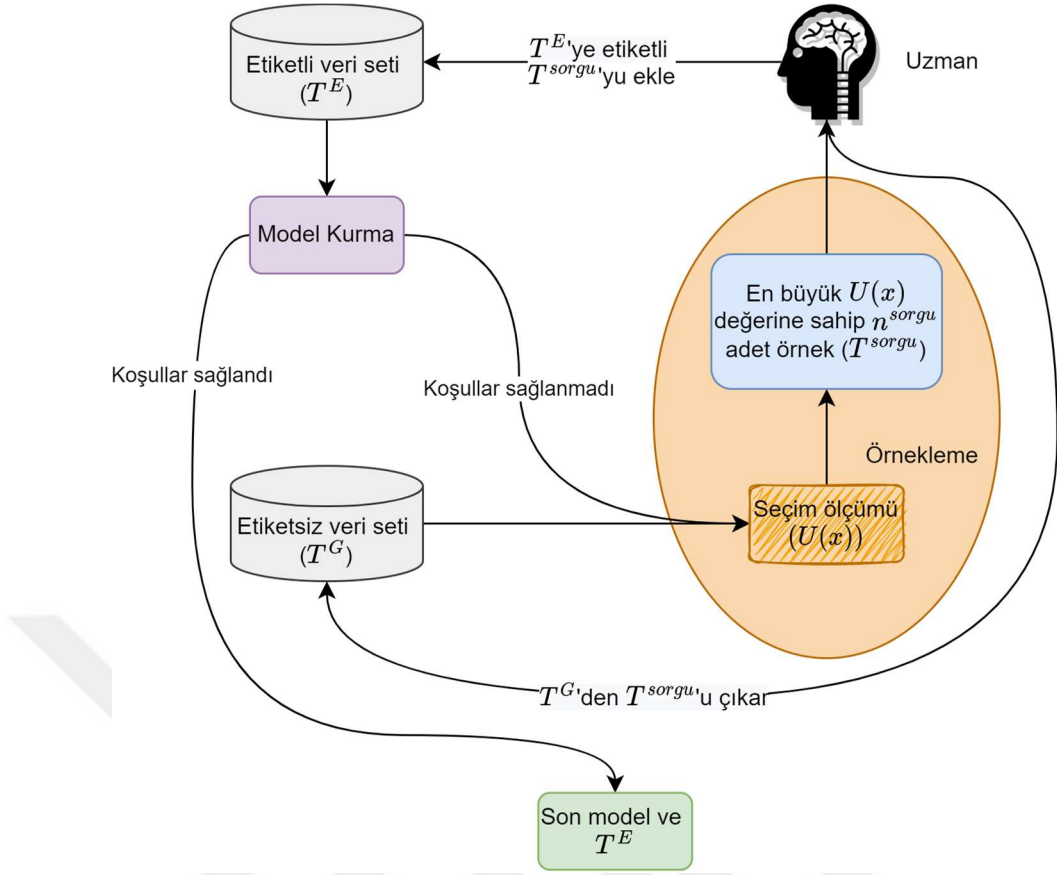
2.4. İki den Fazla Sınıflı Durumlarda Yeniden rnekleme

Blm 2’de bahsi geen tm az rnekleme ve ařırı rnekleme yntemleri, iki sınıflı veri seti erevesindedir. Bu yntemleri ikiden fazla sınıf durumuna uyarlamak olduka basittir. Az rnekleme yaparken, en dřk rnek sayısına sahip olan sınıf, azınlık sınıf olarak kabul edilir ve dięer sınıflar kullanılacak algoritmaya gre azaltılır. Ařırı rnekleme yaparken ise en yksek rnek sayısına sahip sınıf oęunluk sınıf, dięer tm sınıflar azınlık sınıfıdır. Azınlık sınıflardan seilen ynteme gre veri retilerek denge saęlanır.

3. AKTİF ÖRNEKLEME

3.1. Aktif Öğrenme

Çoğu klasik makine öğrenme yöntemi ve formal öğrenme teori modelleri, eğitim için hazırda olan verinin tamamının insanlar tarafından etiketlenebileceği varsayımına sahiptir. Fakat, birçok modern makine öğrenme uygulaması (resim ve video sınıflama, protein serisi sınıflama ve konuşma işleme), yüklü miktarda etiketlenmemiş veriye sahiptir. Aktif öğrenme, danışmanlı öğrenme yaparken uzman (oracle) ile etkileşim halinde olan bir sistemdir. Gerçek hayat problemleri farklı senaryolar gerektireceği için, çok farklı aktif öğrenme türü vardır. Bunlardan en sık kullanılanı göl temelli aktif öğrenmedir. Yapılan araştırmada, başlangıçta yüksek miktarda etiketsiz ve az miktarda etiketli veri olduğu varsayılır. Bu etiketsiz veri örnekleri “göl” olarak adlandırılır. Şekil 3.1, göl temelli aktif öğrenmenin akış şemasını göstermektedir. Mevcut etiketli veri ile model eğitilir. Daha sonra, eğitilmiş modelin etiketsiz verilerdeki belirsizlik bilgisine göre örnek veya örnekler seçilip uzmana etiketlemesi için gönderilir. Böylece, uzmanın rastgele etiketlemesine kıyasla daha az veriyle daha fazla bilgiye erişilir. Aktif öğrenme bazen “seçici örnekleme” (Cohn vd., 1994), “sorgu öğrenme” (Abe, 1998; Baum and Lang, 1992) veya “optimal deneysel tasarım” (Fedorov, 2013), olarak da adlandırılmaktadır. Sahtekarlık tespiti (Peng, 2023), resim sınıflama (Wang vd., 2016; Cao vd., 2020), elektrokatalizör keşfi (Zhong vd., 2020), kelime anlamı netleştirme (Zhu and Hovy, 2007) ve kullanıcı geri bildirimlerinde düşmanca saldırı tespiti (Kebande vd., 2021) gibi alanlarda aktif öğrenme kullanılmaktadır.



Şekil 3.1 Göl temelli örnekleme senaryosu

Sınıflama çerçevesi için göl temelli aktif öğrenmeyi ele alalım. Y , bağımlı değişken, $S = \{S_1, \dots, S_k\}$ olası sınıflar ve $y_i \in S$ ise Y 'nin i -nci değeri olsun. X , örnek uzayı için $T = (X, Y)$ veri seti T^G ve T^E şeklinde iki kısımdan oluşsun. T^G , göl veri seti ve T^E etiketli veri seti olsun ve sırasıyla n^G ve n^E örneğe sahip olsun. i -nci göl verisi

$$T_i^G = (x_i, y_i), x_i = (x_{i1}, \dots, x_{ip}), y_i \in \{?\} \quad (3.1)$$

ve i -nci etiketli veri

$$T_i^E = (x_i, y_i), x_i = (x_{i1}, \dots, x_{ip}), y_i \in S = \{S_1, \dots, S_k\} \quad (3.2)$$

olacak şekilde $g(X, Y): X \rightarrow Y$ fonksiyonu oluşturulur. H hipotez uzayı için $g \in H$ hipotez uzayı kurulmuş olur. Başlangıçta muhtemelen n^G büyük ve n^E küçük bir şekilde eğitim başlar. Aksi takdirde aktif öğrenmeye ihtiyaç olmayacaktır. T^G 'den örnek çekerken kullanılacak bir U seçim ölçümü gereklidir. $U(x_i)$ değeri, modelin, bu durumda sınıflayıcının x_i verisi üzerindeki belirsizliğini, bilgisizliğini ölçer. T^G 'deki en yüksek $U(x)$ değerine sahip n^{sorgu} adet veriden oluşan veri seti $T^{sorgu} \in T^G$ seçilir, T^G 'den çıkarılır ve uzmana gönderilir. Uzman bu örneklerin etiketini belirler. Böylece $y_i^{sorgu} \in \{?\}$ etiketlenerek

$$y_i^{sorgu} \in S = \{S_1, \dots, S_k\} \quad (3.3)$$

olur. Etiketlenen y_i^{sorgu} , T^E 'ye dahil edilir. $n^E = n^E + n^{sorgu}$ ve $n^G = n^G - n^{sorgu}$ şeklinde güncellenir. Belirlenen bir durdurma kriteri sağlanana dek bu işlem tekrar edilir.

Aktif öğrenmede dikkat edilmesi gereken noktaların başında n^{sorgu} 'nun ne kadar olacağı, durdurma kriteri ve U fonksiyonun hesaplanması gelir. Her sorgulama sürecinde, model kurulumu ve $U(x)$ ölçümü yapılacağı için düşük n^{sorgu} , yüksek maliyete neden olacaktır. Büyük n^{sorgu} ise gereğinden fazla örneğin etiketlenmesi ile sonuçlanabilir. Bu da maliyet demektir. Arada bir denge kurmak çok önemlidir. Durdurma kriterinin yanlış belirlenmesi de benzer şekilde az öğrenme ile gereksiz öğrenme arasında kalmaya neden olabilir. Az öğrenildiğinde performans düşük olacaktır. Gereksiz öğrenme ise gereksiz maliyet demektir. Gereksiz öğrenme sonucunda aşırı öğrenme de yaşanabilir.

3.2. Belirsizlik Tahmini

U fonksiyonu için literatürde farklı hesaplamalar mevcuttur. Bunlar belirsizlik örnekleme, komite sorgusu, beklenen model değişimi, beklenen hata azalımı, varyans azalımı ve yoğunluk-ağırlıklı yöntemler şeklindedir. Belirsizlik örneklemesinde öğrencinin kendinden en az emin olduğu örneği seçmek ister. Bu amaçla sınıflama modelleri için kullanılan yaklaşımlardan birisi, en az kendine güven (EAKG) (least confident) yaklaşımıdır. Seçim ölçütü

$$U_{EAKG}(x) = 1 - P(y^{(1)}|x, \theta) \quad (3.4)$$

şeklinde belirlenir. Burada $y^{(1)}$, en yüksek olasılıklı sınıf etiketidir. Ölçülen değer, en yüksek olasılıkla tahmin edilen sınıf etiketinin yüzde yüz kesinlikten ne kadar uzakta olduğudur. Bu yaklaşım, sadece en olası sınıfı dikkate alır. Diğer sınıf olasılıklarına ait bilgileri dikkate almaz. Marjin (MA) (margin) yaklaşımı (Scheffer vd., 2001), en yüksek olasılıklı iki sınıfı ele alarak $MA(x) = P(y^{(1)}|x, \theta) - P(y^{(2)}|x, \theta)$ şeklinde bir örneğin ne kadar kolay tahmin edildiğini belirler. Belirsizlik tahmini için bunun tersi alınarak

$$U_{MA}(x) = P(y^{(2)}|x, \theta) - P(y^{(1)}|x, \theta) \quad (3.5)$$

şeklinde belirsizlik hesaplanır. $y^{(2)}$, ikinci en yüksek olasılıklı sınıf etiketidir. MA yaklaşımı, sadece en olası sınıfı değil, ikinci olası sınıfı da belirsizlik hesabına dahil etmiş olur. En olası sınıf ile, ikinci en olası sınıf arasındaki fark ne kadar çoksa, model kendinden o kadar emindir. Ne kadar azsa, o kadar az emindir. Fakat, görüldüğü üzere

ikiden fazla sınıf içeren sınıflama modelleri için, EAKG’de mevcut olan düşük olasılıklı sınıfların dikkate alınmaması durumu, burada da geçerlidir.

Bunların dışında kullanılan en popüler belirsizlik ölçütü Shannon entropidir (SE) (Shannon, 1948). Shannon entropi belirsizliği $U_{SE}(x)$ ’in hesaplanması,

$$U_{SE}(x) = - \sum_{l=1}^k P(y = S_l|x, \theta) \log(P(y = S_l|x, \theta)) \quad (3.6)$$

şeklindeir. SE, diğer yöntemlerden farklı olarak tüm sınıfları hesaba katar.

Komite sorgusu (KS) stratejisi, Seung vd. (1992) tarafından önerilmiştir. KS yaklaşımında $C = \{C_1, \dots, C_{n_C}\}$ modelleri oluşturulur. Burada n_C model sayısı, C model kümesi, $C_i, i = 1, \dots, n_C$ ise komitedeki modellerden (sınıflayıcılardan) herhangi birisidir. Bu modeller birbirinden farklı modelleme yöntemleriyle kurulabilir. Her bir komite üyesi, örneğe ait tahmin konusunda oy kullanır. Ne kadar çok model ortak oy kullanmışsa, o örnek o kadar az belirsizlik içerir. Ne kadar az ortak oy mevcutsa, belirsizlik o kadar fazladır. Topluluk algoritmaları, komite sorgusu için oldukça uygun yöntemlerdir. Birbirinden farklı modeller ve uyumsuzlıklardan faydalanan topluluk algoritmalarında, model toplulukları C komitesi olarak kabul edilebilir. Abe ve Mamitsuka (1998), boosting ile sorgu ve bagging ile sorgu olacak şekilde topluluk algoritmalarını kullanmayı önermişlerdir. KS stratejisinde oy entropisi (OE) (Dagan and Engelson, 1995) ve Kullback-Leibler (KL) ayrışımı (McCallum and Nigam, 1998) en sık kullanılan ölçümlerdir. OE,

$$U_{OE}(x) = - \sum_{l=1}^k \frac{V(y = S_l)}{n_C} \log \frac{V(y = S_l)}{n_C}, \quad (3.7)$$

KL ise,

$$U_{KL}(x) = \frac{1}{n_C} \sum_{j=1}^{n_C} \sum_{l=1}^k P_{C_j}(y = S_l|x) \log \frac{P_{C_j}(y = S_l|x)}{P_C(y = S_l|x)} \quad (3.8)$$

şeklinde hesaplanır.

Beklenen model değişimi stratejisi, modele dahil edildiğinde en çok değişime neden olacak örneği seçme stratejisidir. Beklenen gradyan uzunluğu (BGU) (Settles vd., 2007), en bilindik yaklaşımdır. Olasılıksal olan ve olmayan yöntemler için uygulanabilir. x_i^{sorgu} örneği, T^E ’e eklendikten sonra yarattığı gradyan değişimini inceler. $\nabla l(T^E; \theta)$, θ parametrelerine bağlı amaç fonksiyonunun türevi ve yeni türev $\nabla l(T^E \cup \{x, y\}; \theta)$ olsun. Bu durumda BGU

$$U_{BGU}(x) = \sum_{l=1}^k P(y = S_l|x; \theta) \|\nabla l(T^E \cup \{x, y\}; \theta)\| \quad (3.9)$$

olur. Burada $\|\cdot\|$, gradyan vektörünün normudur. Yakınsama durumunda $\nabla l(T^E; \theta) \cong 0$ olacağı için, örneklerin bağımsız olduğu varsayımı altında $\|\nabla l(T^E \cup \{x, y\}; \theta)\| \cong \|\nabla l(\{x, y\}; \theta)\|$ denebilir. Bu nedenle Eşitlik 3.9'da $\|\nabla l(T^E \cup \{x, y\}; \theta)\|$ yerine $\|\nabla l(\{x, y\}; \theta)\|$ hesaplamak, daha az maliyetli olacaktır.

Beklenen hata azalımı stratejisi, genelleştirilebilir hatanın ne kadar azaldığı ile ilgilenir. $T^E \cup \{x, y\}$ 'nin T^G 'deki beklenen hatası tahmin edilmeye çalışılır ve bu şekilde en düşük beklenen gelecek hatası olan örnekler belirlenir. Yaklaşımlardan birisi, beklenen 0/1 kaybıdır ve

$$U_{0/1}(x) = \sum_{l=1}^k P(y = S_l|x; \theta) \left(\sum_{j=1}^{n_B} 1 - P(\hat{y}|x_j^B; \theta^{+\{x, S_l\}}) \right) \quad (3.10)$$

şeklinindedir. Burada $\theta^{+\{x, y\}}$, $\{x, y\}$ yeni örneği eklendikten sonra T^E veri setine ait modelin parametreleridir. Başka bir yaklaşım da beklenen log-kayıp değeridir ve

$$U_{log-kayıp}(x) = \sum_{l=1}^k P(y = S_l|x; \theta) \left(- \sum_{j=1}^{n_B} \sum_{m=1}^k P(y = S_m|x_j^B; \theta^{+\{x, S_l\}}) \log P(y = S_m|x_j^B; \theta^{+\{x, S_l\}}) \right) \quad (3.11)$$

şeklinde hesaplanır. Tüm T^B üzerinde beklenen entropinin azaltılmasına karşılık gelir.

Varyansı azaltarak dolaylı bir şekilde genelleştirilmiş hata azaltılabilir. Varyans analizi stratejisi, bu avantajı ele alır ve beklenen hata azaltımına kıyasla daha basit (bazen kapalı forma sahip şekilde) $U(x)$ hesaplayabilir. Bir modelin beklenen gelecek hatası,

$$\begin{aligned} E_T((\hat{y} - y)^2|x) &= E((y - E(y|x))^2) \\ &+ (E_{T^E}(\hat{y}) - E(y|x))^2 \\ &+ E_{T^E}((\hat{y} - E_{T^E}(\hat{y}))^2) \end{aligned} \quad (3.12)$$

şeklinde ayrışabilir (Geman vd., 1992). Burada $E_{T^E}(\cdot)$, T^E veri setindeki beklenen değeri, $E(\cdot)$, $P(y|x)$ yoğunluğu üzerindeki beklenen değeri, $E_T(\cdot)$ ise her ikisi için beklenen değeri ifade etmektedir. Denklem 23'te sağ kısımdaki ilk terim gürültüdür ve gerçek y etiketi için x biliniyorken varyans anlamına gelir. İkinci terim, yan miktardır. Üçüncü terim ise modelin varyansıdır. Üçüncü terimin minimize edilmesi, gelecek genelleştirilmiş model hatasını da minimize edecektir. Üçüncü terimi σ_y^2 olarak ifade edecek olursak varyans azalımı stratejisi için $U(x)$,

$$U_{var-azalım}(x) = \{\tilde{\sigma}_y^2\}^{+\{x^{sorgu}\}} \quad (3.13)$$

şeklinde ifade edilebilir. Burada $\tilde{\sigma}_y^2, \sigma_y^2$ varyansının tahmini ve $\{\tilde{\sigma}_y^2\}^{+\{x^{sorgu}\}}, x^{sorgu}$ eklendikten sonraki σ_y^2 tahminidir. Farklı yöntemler için $\tilde{\sigma}_y^2$ tahmini farklılık göstermektedir. Bu yaklaşım, istatistikte optimal deney tasarımı olarak geçer ve farklı yaklaşımlar mevcuttur (Fedorov, 2013).

Yoğunluk-ağırlıklı yöntemler, örnek birimleri yerine tüm girdi uzayına odaklanır. Settles ve Craven (2007) tarafından önerilen bilgi yoğunluğu (BY) yaklaşımı sadece belirsizlik olan örnekleri değil, ayrıca dağılımı temsil edebilen örnekleri seçmeyi hedefler. BY yöntemi için $U(x)$ yaklaşımı

$$U_{BY}(x) = \phi_A(x) \left(\frac{1}{n_B} \sum_{i=1}^{n_B} benzerlik(x, x_i^B) \right) \quad (3.14)$$

şeklindedir. Burada $\phi_A(x)$, herhangi bir A sorgu stratejisi için x 'in bilgilendiricilik seviyesidir. İkinci kısım ise x 'in girdi uzayındaki diğer örneklerle olan ortalama benzerlik seviyesidir. $\phi_A(x)$ için ve $benzerlik(\cdot)$ için kullanılabilecek burada değinmeyeceğimiz bir çok yöntem mevcuttur.

4. ÖNERİLEN YÖNTEM

Çalışmada önerilen aşırı örnekleme yönteminin adı “Aktif Aşırı Öğrenme”’dir (UGO). Bu yöntem ile Tablo 4.1’de bahsi geçen bazı problemlere çözümler getirilmiştir ve yeni ortama uyum sağlayabilmesi için aktif öğrenmeye düzenlemeler yapılmıştır.

Tablo 4.1 UGO’un aşırı örnekleme ve aktif öğrenmedeki eksikliklere getirdiği çözümler

Problem	Çözüm
Rastgele veri üretme	Aktif öğrenme ile bilgilendirici veri üretme
Aktif öğrenme, etiketsiz örneklerden örneklem çeker, bu nedenle gerçek sınıf bilgisinden faydalanamaz	Gerçek sınıf bilgisinden faydalanan düzenlenmiş Marjin belirsizliği
Eğitim yapılacak olan sınıflayıcıya göre belirsizlik tahminleri değişebilir	Eğitim yapılacak sınıflayıcı kullanılarak belirsizlik tahmini
Gürültü üretme	Gürültülü örnekler üzerinden veri üretilmemesi ve üretilen gürültülerin yer değiştirmesi
Sınıf dengesizliğinin yarattığı problemin tespitinin gözlem sayısına bağlı olarak değerlendirilmesi	Sınıflara ait epistemik belirsizliğin ölçülmesi ile veri eksikliğinden kaynaklı sınıflar arası dengesizliğin tespiti
Ne kadar veri üretilmesi gerektiği	Belirsizlik ortalamaları arasında fark olmayacak miktarda iteratif veri üretimi

Tablo 4.1’de bahsedilen problemlerden birisi, aşırı örneklemede, veri üretirken yapılan genellikle X^A veri setinden örneklerin seçilmesi ve seçilen örneklere ve yönteme göre veri üretilmesidir. ROS, SMOTE ve RWO’da olduğu gibi genellikle örnekler rastgele seçilir. Tüm X^A örneklerinin aynı biçimde bilgilendirici olması mümkün değildir. Bu nedenle örnek seçerken, bilgilendirici olan örneklere öncelik vermek mantıklı olacaktır. Bunu yapabilmek için göl temelli aktif öğrenme kullanılabilir. Aktif öğrenme insan ve makinenin bir arada çalıştığı bir öğrenme türüdür. İnsanın görevi, uzman (oracle) olarak etiketleri bilinmeyen verilerin etiketlerine karar vermektir. Aşırı örnekleme çerçevesinde gerçek etiketler bilindiği için biraz düzenleme ile göl temelli aktif öğrenme ile örneklem seçmek mümkündür.

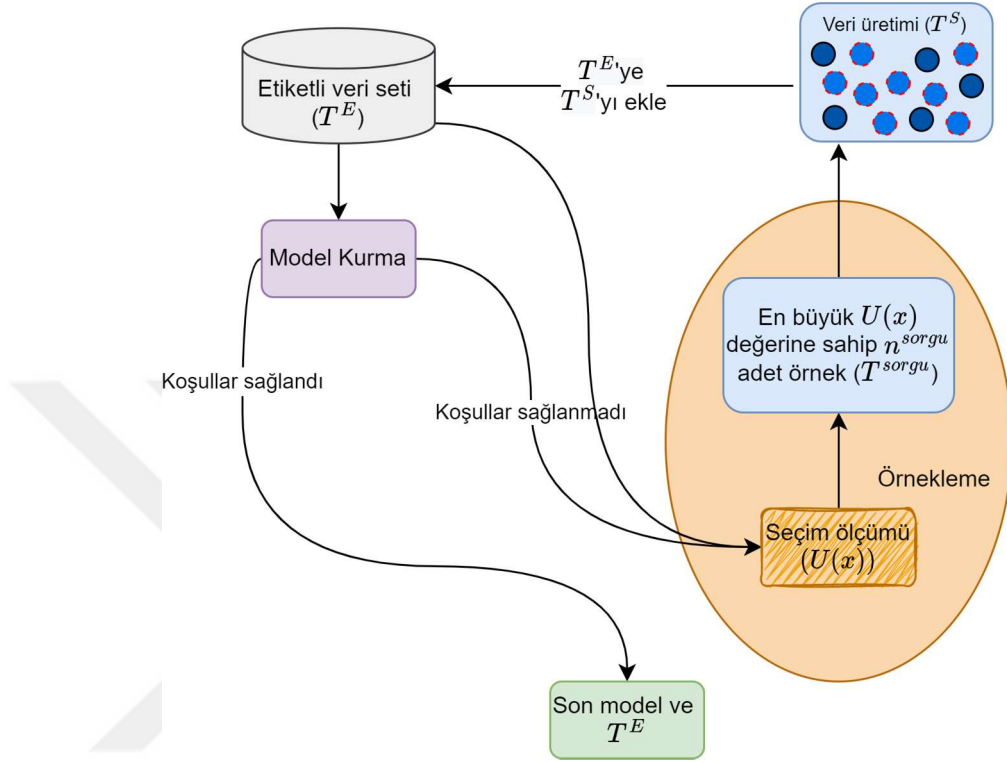
Gürültü problemi ise iki şekilde gerçekleşir. Veri setinde kesişim bölgesinde veya X^C veri setinin yoğun olduğu bölgelerde azınlık sınıf örnekleri olabilir. Böyle bir durumda aşırı örnekleme, bu gürültü örneğin olduğu yerde daha fazla gürültü üretir. Bir diğer gürültü üretimi de, yönteme bağlı olarak gerçekleşir. SMOTE gibi yöntemlerde azınlık sınıf örneklerine ait en yakın komşular arasındaki doğru veri üretilir. Fakat en yakın komşu ile kendisi arasında çoğunluk sınıf örnekleri mevcut

olabilir. Bu da belki de normalde gürültü olmayan o bölgede yeni gürültü örnekleri üretilmesine neden olur. Bu iki durumun birincisi için, öncelikle gürültülü azınlık sınıf örnekleri tespit edilerek ve o örneklerle bağlı veri üretmeyerek çözüm getirilebilir. İkinci durum için ise üretilen örneklerin gürültü olup olmadıkları kontrol edilerek gürültü olmayacakları bir konuma yerleri değiştirilebilir. Her iki öneri de UGO içerisinde uygulanmaktadır. Aktif öğrenme yaparken karşılaşılan en büyük zorluklardan birisi, durdurma kriteridir. UGO'un amacı sınıf dengesizliğinin önüne geçmek için veri üretmektir. Epistemik belirsizlik, veri eksikliğini ölçtüğü için burada sınıflara ait epistemik belirsizlik miktarları incelenebilir. UGO, sınıflara ait belirsizlik ortalamalarını hesaplar ve aralarında anlamlı fark olup olmadığını istatistiksel olarak her iterasyonda test eder. Anlamlı fark varsa, yüksek belirsizlik ortalamasına sahip sınıfın daha fazla veri üretimine ihtiyacı olduğu anlamına gelir. Bu durumda veri üretirken o sınıftan veri üretilir. Belirsizlik ortalamaları arasında anlamlı fark olmadığında üretim durur. Böylece sınıf dengesizliği, belirsizlik yönünden ölçülmüş olur. Bu hesaplamaların asıl katkısı, sınıf dengesizliğinden kaynaklanan bir performans kaybı olup olmadığının ölçebilmesidir. Sadece gözlem sayısında farklılık olması, otomatik olarak performans kaybı olacağı anlamına gelmez. Ayrıca, bu durdurma kriteri sağlanana kadar veri üretileceği için, ne kadar sayıda veri üretileceğinin kullanıcı tarafından belirlenmesine gerek yoktur. UGO, belirsizlik durumuna göre gerekli miktarda veri üretir. Ayrıca epistemik belirsizlik, modelin yetersizliğini, yani hangi örneklerde tahmin yapmakta yetersiz olduğunu, veriye ihtiyaç duyduğunu belirlediği için sınıflama yönteminden sınıflama yöntemine farklılık gösterir. Tüm sınıflama yöntemleri için geçerli bir belirsizlik tahmin yöntemi yoktur. Bu nedenle UGO, hangi sınıflayıcı kullanılacak o sınıflayıcı ile belirsizlik tahmini yapar ve örnekleri bu belirsizliklere göre seçer.

4.1. Modifiye Edilmiş Belirsizlik Örnekleme

Şekil 4.1'de önerilen aktif öğrenme yaklaşımı görülmektedir. Klasik göl temelli aktif öğrenmeden farkı, etiketsiz veri ve uzmanın olmamasıdır. $U(x)$ ile belirsizlikler doğrudan etiketli veriler üzerinden yapılır. Ardından en yüksek belirsizliğe sahip n^{sorgu} adet veri T^{sorgu} seçilir. Seçilen veriler kullanılarak veri üretilir. Veri üretimi, kullanıcının istediği şekilde yapılabilir. Örneğin seçilen veriler kopyalanabilir, SMOTE gibi bir yöntem ile örnekler arasındaki komşuluklarda veri rastgele üretilebilir veya RWO gibi bir yöntem ile değişken standart hatalarına bağlı olarak normal dağılım

ile veri üretilebilir. Üretilen sentetik veriler, T^S , T^E 'ye ve dolayısıyla T 'ye eklenir. Bu yöntemde amaç sınıf dengesizliği durumunda, aşırı örnekleme yaparak belirsizlik dengesini sağlamaktır. Durdurma kriterinin koşulları sağlanana dek bu işlemler tekrar edilir.



Şekil 4.1 Yeniden örnekleme için modifiye edilmiş göl temelli belirsizlik öğrenme stratejisi

4.2. Modifiye Edilmiş Marjın Belirsizlik

Belirsizlik, bir olayın dağılımının ve/veya dağılımdan gelecek olan olayın ne olacağından emin olunmaması durumudur. İstatistikte genellikle belirsizlik ölçümü olarak güven aralıkları, standart hata ve varyans gibi ölçümler kullanılır. Matematiksel formda gösterildiği durumlar olduğu gibi, deneyler sonucunda elde edilen ölçümlere göre de tahmin edilebilir. Tahmin edilecek olan belirsizlik, farklı sebeplerden kaynaklanıyor olabilir. Model parametre değerlerinin tam olarak bilinmiyor ve istatistiksel yöntemlerle açıklanamıyor olması, bağımsız (girdi) değişken değerlerinin sabit olmaması, kurulacak olan modelin yapısı hakkında yeterli bilgiye sahip olunmaması, kurulması amaçlanan modelin çok karmaşık olmasından kaynaklanan sayısal hatalar (örneğin modelin yakınsamaması), gözlem hatalarından kaynaklanan yanlış veriler, yetersiz veri gibi sebepler belirsizlik kaynaklarına örnek olarak gösterilebilir. Bu sebepler ile belirsizlik olarak iki kısma ayrılır: aleatorik belirsizlik ve epistemik belirsizlik. Sınıflama çerçevesinde aleatorik belirsizlik, sınıflar arası

dağılım ile ilgilidir. Epistemik belirsizlik ise tahmin edilen sınıf olasılıklarının her birinin kendi dağılımı ile ilgilidir (Dürr vd., 2020).

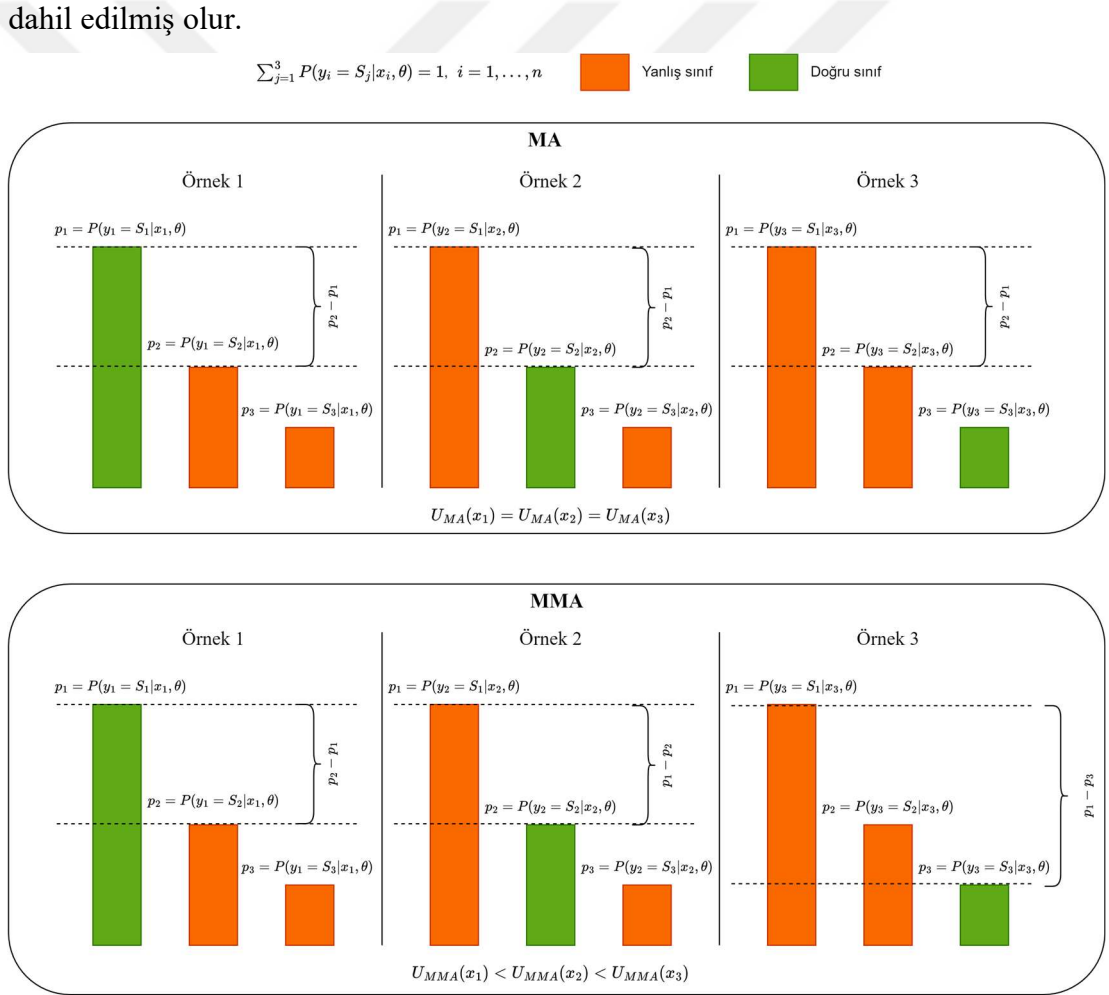
Aleatorik belirsizlik, veri belirsizliğidir veya rastgeleliğidir bu nedenle azaltılamaz (Hüllermeier and Wegeman, 2021; Wang vd., 2019; Dürr vd., 2020). Modelden bağımsız olduğu için, ölçüm yöntemlerine göre farklılık göstermemesi ve epistemik belirsizlik ile ilişkisiz olması beklenir. Fakat beklenenin aksine belirsizlik tahmin yaklaşımının ölçülen epistemik belirsizlikle birlikte aleatorik belirsizliği de etkilediği görülmüştür (Valdenegro-Toro and Mori, 2022). Aleatorik belirsizlik, verinin açıklanamayan rastgeleliğini ifade eder. Aynı deneyin tekrar tekrar yapılması sonucunda farklı çıktıların ortaya çıkmasıdır. Rastgeleliğin nedeninin açıklanamayacak olması, var olmadığı veya ölçülemeyeceği anlamına gelmez. Aleatorik belirsizlik, verideki gürültüyü ölçmek ile ilgilidir. Bu nedenle daha fazla veri elde edilse de azalmaz. Epistemik belirsizlik, yetersizlikten kaynaklanan belirsizliktir. Teoride bilinebilecek bir şeyin uygulamada bilinemediği durumları kapsar. Bu nedenle veri miktarı arttıkça epistemik belirsizlik azalır. H hipotezi altında X 'e bağlı Y 'ye giden gerçek fonksiyon $f(x)$, kullanılan yönteme bağlı olarak teorik açıdan mümkün en iyi fonksiyon $f^*(x)$ ve eldeki verilerle oluşturulmuş mevcut tahmin fonksiyonu $\hat{f}(x)$ olacak şekilde, aleatorik belirsizlik $f(x) - f^*(x)$ 'i açıklamada, epistemik belirsizlik ise $f^*(x) - \hat{f}(x)$ 'i açıklamada kullanılabilir. Bu nedenle, aleatorik ve epistemik belirsizliğin toplamı $f(x) - \hat{f}(x)$ şeklinde açıklanır. Bu da aslında hata miktarıdır. Klasik belirsizlik tahmininde, tahminin etiketsiz verilerde doğru yapılması önemlidir. Bu çalışmada, etiketsiz verilerde belirsizlik tahminine gerek olmadığı için doğrudan hata miktarı hesaba katılabilir. Eşitlik 3.5'te verilen MA belirsizlik tahmini modifiye edilerek modifiye marjın belirsizlik tahmini (MMA)

$$U_{MMA}(x_i) = P(y^{(1)'}|x_i, \theta) - P(y = y_i|x_i, \theta) \quad (4.1)$$

şeklinde oluşturulmuştur. Burada, MA'daki en yüksek olasılıklı iki sınıf arasındaki fark yerine, gerçek sınıfa ait olasılık $P(y = y_i|x, \theta)$ ile gerçek sınıfın dışındaki en yüksek olası sınıf $P(y^{(1)'}|x, \theta)$ olasılığı arasındaki fark hesaplanmıştır. MMA'da $P(y|x, \theta)$ düştükçe belirsizlik artar. Bu da hata miktarının, yani toplam belirsizlik $f(x) - \hat{f}(x)$ 'in hesaba katılmasını sağlar. $P(y^{(1)'}|x, \theta)$ olasılığının yüksek olması ise farklı sınıfları seçme ihtimalini yani belirsizliği artırır. $P(y^{(1)'}|x, \theta)$ 'nın düşük olması, diğer sınıfların olasılığının ister birbirlerine yakın olsun ister olmasın, gerçek sınıftan uzak olduğunu gösterir. Bu da, $\hat{f}(x)$ 'in $f(x)$ 'e yaklaştığının bir başka belirtisidir.

$U_{MMA}(\mathbf{x})$, -1 ile 1 arasında değer alır. Belirsizliğin -1 'e yaklaştıkça azaldığı, 1 'e yaklaştıkça arttığı şeklinde yorumlanır.

Şekil 4.2, MA ve MMA'nın üç örnek üzerindeki karşılaştırmasını göstermektedir. Turuncu renkli sütunlar, yanlış sınıfa ait olasılıkları, yeşil renkliler ise doğru sınıfa ait olasılıkları göstermektedir. Üç sınıflı bu duruma ait bu örnekler için MA belirsizlik tahminine bakıldığında, üç örnekte de eşit belirsizlik ölçümü yaptığı görülmektedir. Gerçek etiket bilgisinden faydalanan MMA ile yapılan belirsizlik tahmininde ise en yüksek belirsizlik olduğunu, en düşük belirsizliğin ise örnek 1'de olduğunu belirlemektedir. Örnek 3'te en yüksek belirsizliğin olduğunu söylemesinin nedeni, doğru sınıfa en düşük olasılığın verilmesi, Örnek 1'de ise doğru sınıfa en yüksek olasılığın verilmesidir. Böylece MMA ile hata miktarı da belirsizlik ölçümüne dahil edilmiş olur.



Şekil 4.2 MA and MMA örnek üzerinde karşılaştırma

4.3. Durdurma Kriteri

UGO, sınıf dengesizliğini epistemik belirsizlik dengesizliği olarak ele alır. Bu nedenle sınıflara ait epistemik belirsizlikleri kontrol eder ve hangi sınıf daha yüksek belirsizliğe sahipse o sınıfta veri üretir. Eğer sınıflara ait belirsizlikler arasında anlamlı fark yoksa, veri üretimine gerek yoktur ve durdurulabilir. Bu nedenle, sınıflara ait ortalama belirsizlikler

$$H_0: \bar{U}^{(1)} = \dots = \bar{U}^{(k)} \quad (4.2)$$

sıfır hipotezi ile veri üretiminin her bir iterasyonunda Welch ANOVA testi ile test edilir. H_0 reddedilemediğinde dengesizlik ortadan kalkmış olur. Böylece veri üretimi durdurulur. Welch ANOVA testinin tercih edilmesinin sebebi, genellikle 30'dan fazla örnek bulunması ve belirsizliklerin homojen dağılmamasıdır.

Bu durdurma kriterinin dışında ikinci bir durdurma kriteri vardır. Bu ikinci kriter, çok fazla veri üretilmesinin önüne geçmeyi amaçlamaktadır. İlk kritere göre UGO, normalden daha az veri üretebileceği gibi istenmeyen miktarda fazla veri de üretebilir. Bu amaçla ikinci durdurma kriteri

$$\phi \times \max(n_l) \times k - n \leq n^{sen}, l = 1, \dots, k \quad (4.3)$$

koşuludur. Bu koşul doğru ise, çok fazla veri üretilmiştir ve üretim durdurulur. Koşul yanlış ise veri üretimine devam edilir. $\phi = 1$ değeri tam denge sağlamak için gereken veri sayısına denk gelir. Çalışmada daha çok ilk kriteri kullanmak istediğimiz için $\phi = 2.5$ olarak alınmıştır. Böylece fazlasıyla veri üretilebilir.

4.4. Belirsizlik Yönlendirmeli Aşırı Örnekleme (UGO)

Çalışmada, aşırı örneklemede gürültü varlığını hesaba katan ve bilgilendirici örnekler üreten Belirsizlik Yönlendirmeli Aşırı Örnekleme (UGO) önerilmiştir. Önerilen yeniden örnekleme yönteminin algoritması ile ilgili ayrıntılı bilgiler Şekil 4.3'te görülmektedir. UGO, hem mevcut veri setindeki gürültüleri hem de yeni üretilen sentetik verilerdeki gürültüleri dikkate alır. p değişkenli n örnekli açıklayıcı değişkenler matrisi X , n uzunluğunda $S_1, \dots, S_k$ şeklinde k adet sınıflı sınıf değişkeni Y olsun. Seçilen bir makine öğrenme yöntemini temsil eden f sınıflayıcısı eğitilecek olsun. Kendi örnekleri üzerinde belirsizlik tahmini yaparken, sınıflamada kullanılan klasik belirsizlik tahminleri, epistemik ve aleatorik belirsizlik ayrımı yapmamaktadır. Bu nedenle UGO'da aleatorik belirsizliği yüksek örnekleri belirlemek amacıyla gürültü adı verilen kesişim bölgesi örnekler tespit edilmektedir. Kesişim bölgesi örneklerini tespit etmek amacıyla Tablo 4.1'de verilen *gürültüÖnTespit* algoritması

kullanılmaktadır. Bu algoritma, çok basit ve anlaşılır bir yaklaşım uygular. Her örneğin en yakın $k^{gö}$ adet komşusunu inceler. En yakın komşulara ait indislerin matrisi NN olacak şekilde i -nci örnek için $\sum_{j=1}^{k^{gö}} I(Y_{NN_{ij}} = y_i) > k^{gö}/2$ ise, örnek güvenli bölgededir ve “gürültü değil”dir. $\sum_{j=1}^{k^{gö}} I(Y_{NN_{ij}} = y_i) \leq k^{gö}/2$ ise kesişim bölgesindedir ve “gürültü” olarak nitelendirilebilir. Yani, en yakın komşularının çoğu kendi sınıfından örnekler ise gürültü değil, aksi halde gürültüdür. Burada NN_{ij} , i -nci örneğin en yakın j -nci komşusunun indisidir. Gürültü olarak nitelendirilen örnekler, kesişim bölgesinde oldukları için doğal olarak yüksek aleatorik belirsizliğe sahiptir. Bu nedenle UGO örnek üretiminde bu örnekleri kullanmaz. Tablo 4.2’de görülen *gürültüÖnTespit* algoritması, X, Y ve k^g şeklinde üç girdi argümanına sahiptir. X , gürültü tespiti incelenecek olan veri setine ait açıklayıcı değişken matrisi, Y sınıf bilgisini içeren sınıf vektörü, k^g ise bakılacak olan en yakın komşu sayısıdır. $k^g = k^{gö}$ olacak şekilde gürültü ön tespiti yapılmaktadır.

Tablo 4.2 *gürültüÖnTespit* sahte kodları

<i>gürültüÖnTespit</i> algoritması Girdi: $X_{n \times p}, Y_{n \times 1}, k^g$ 1. $NN_{n \times k^g}^{X2X} \leftarrow enYakınKomşu(X, X, k^g)$ 2. $K_{n \times 1} \leftarrow$ boş vektör 3. For $i = 1, \dots, n$ If $\sum_{j \in NN_i^{X2X}} I(Y_{NN_i^{X2X}} = y_i) > \sum_{j \in NN_i^{X2X}} I(Y_{NN_i^{X2X}} \neq y_i)$ $K_i \leftarrow$ “gürültü değil” Else $K_i \leftarrow$ “gürültü” Çıktı: $K_{n \times 1}^g$

UGO, iteratif olarak veri üretir. İterasyonun ilk adımı, f sınıflayıcısının eğitimi ile başlar. Veri üretilecek örnekleri rastgele seçmek yerine bilgilendirici örneklerin seçilmesi için f sınıflayıcısı X ve Y üzerinde $\hat{\theta}^f$ parametreleri ile eğitilir. $\hat{\theta}^f$ parametreleri sınıflayıcıya ait parametre/hiperparametreleri olan θ^f ’nin tahmincilerini temsil eder ve sınıflayıcıdan sınıflayıcıya değişiklik gösterir. Eğitim sonrası elde edilen sınıflayıcı $\hat{f}$, gürültü olmayan örnekler üzerinde belirsizlik tahmini yapmak için kullanılır. Sınıflara ait olasılık tahminleri kullanılarak U_{MMA} değerleri hesaplanır. Aleatorik belirsizliği yüksek olan örnekler kullanılmadığı için, geriye kalan örneklerde

hesaplanan U_{MMA} değerleri epistemik belirsizliği ölçer. Bu sayede veri eksikliğinden kaynaklanan ve veri artırılarak çözülebilecek belirsizlik ölçülmüş olur.

İlk durdurma kriteri burada devreye girer. Eşitlik 4.2'deki hipoteze göre sınıfların belirsizlik ortalamaları arasında fark olup olmadığı incelenir. Eğer belirsizlikler sınıflara göre anlamlı farklılık gösteriyorsa, sınıf dengesizliği olduğuna kanaat getirilir ve iterasyon devam eder. En yüksek belirsizlik ortalamasına sahip sınıf

$$c = \operatorname{argmax}_{l \in (1, \dots, k)} \bar{U}^{(l)} \quad (4.4)$$

şeklinde seçilir. $S^{(c)}$ sınıfı, en yüksek epistemik belirsizliğin bulunduğu sınıftır ve veri üretimi bu sınıftan başlar. Bu sınıfa ait örneklerden en yüksek belirsizliğe sahip $n^{sec} = n_c \times \lambda^{sec}$ adet örnek

$$i^{sec} = \operatorname{argmax}_{v \in (1, \dots, n_c), |v|=n^{sec}} \sum_{i \in v} U_i^{(c)} \quad (4.5)$$

şeklinde seçilir. Burada λ^{sec} , $[0,1]$ aralığındadır ve seçilecek örneklerin sayısını belirler. λ^{sec} ne kadar büyük olursa o kadar fazla örnek, ne kadar az olursa o kadar az örnek seçilmiş olur. $\lambda^{sec} = 1/n_c$ için her seçimde bir adet örnek seçilir ve $\lambda^{sec} = 1$ için, tüm örnekler seçilir ve belirsizlik örnekleme yapmadan veri üretilmiş olur. $|\cdot|$, “.” vektörünün uzunluğudur ve i^{sec} , en yüksek belirsizliğe sahip n^{sec} örneğin indisidir.

Veri üretilen örnekler, $X_{i^{sec}}^{(c)}$, belirlendiğine göre artık veri üretimine geçilebilir. Veri üretimi için kullanılacak yeniden örnekleme fonksiyonu olan R , girdi için en fazla $X_{i^{sec}}^{(c)}$, X , Y , $S^{(c)}$, $n^{adım}$ ve $\hat{\theta}^R$ argümanlarını gerektiren ve çıktı için en az $\hat{X}^{adım}$ çıktısını veren herhangi bir fonksiyon olabilir. Burada $X_{i^{sec}}^{(c)}$, i^{sec} ile belirlenmiş ve üzerinde/çevresinde veri üretilen örnekler, $n^{adım} = n \times \lambda^{adım}$ ve $\hat{\theta}^R$, R fonksiyonuna ait tahmini hiperparametreleridir. $\lambda^{adım}$, $[0, \infty)$ aralığındadır her adımda n^{sec} adet seçilen örnekten üretilen veri sayısı $n^{adım}$ ’ı belirlemeye yarar. $\lambda^{adım} = 0$ için veri üretilmez ve ne kadar büyük olursa, her adımda o kadar fazla veri üretilir. Küçük $\lambda^{adım}$ değerleri, daha dikkatli ve kontrollü veri üretimi sağladığı gibi çok fazla iterasyon gerekmesine ve işlem maliyetine neden olur. Büyük değerler aldıkça, fazla veri üreteceği için, iterasyon sayısını azaltır ve algoritmayı hızlandırır. Fakat her iterasyonda tek bir sınıfta çok fazla veri üretecektir. Bu durum, dengenin bir türlü sağlanamamasına neden olacağı gibi, örneklenen noktalarda çok fazla gereksiz veri üretiminden dolayı aşırı uyum problemine yol açabilir. λ^{sec} ve $\lambda^{adım}$ hiperparametreleri, UGO’ya esneklik kazandırdığı gibi sonuçları etkileyebilmektedir.

Yaptığımız denemeler, $\lambda^{sec} = 0.5$ ve $\lambda^{adim} = 0.1$ değerlerinin çoğu veri seti için başarılı ve yeterli olduğunu göstermiştir.

Tablo 4.3 *gürültüSonTespit* sözde kodları

<i>gürültüSonTespit</i> algoritması Girdi: $X_{n_A \times p}^A, X_{n \times p}, Y_{n \times 1}, k^g, S$ 1. $NN_{n_A \times k^g}^{X^A 2X} \leftarrow enYakınKomşu(X^A, X, k^g)$ 2. $K_{n_A \times 1} \leftarrow$ boş vektör 3. For $i = 1, \dots, n_A$ If $\sum_{j \in NN_i^{X 2X}} I(Y_{NN_i^{X 2X}} = S) > \sum_{j \in NN_i^{X 2X}} I(Y_{NN_i^{X 2X}} \neq S)$ $K_i \leftarrow$ “gürültü değil” Else $K_i \leftarrow$ “gürültü” Çıktı: $K_{n_A \times 1}^g, NN_{n_A \times k^g}^{X^A 2X}$
--

R ile veri üretimi sonrası $S^{(c)}$ sınıfına ait elde edilen sentetik veriler $\hat{X}^{sen}$ olsun. İkinci aşama gürültü probleminden, yani üretilen verilerde gürültü oluşması probleminden etkilenmemek için Tablo 4.3'te verilen *gürültüSonTespit* algoritması ile gürültü tespiti yapılır. Bu algoritma, X^A, X, Y, k^g ve S argümanlarını girdi olarak gerektirir. $X^A = \hat{X}^{sen}, X = X, Y = Y, k^g = k^{gs}$ ve $S = S^{(c)}$ olacak şekilde algoritma çalıştırılır ve iki adet çıktı verir. Bunlardan birincisi K^g gürültü vektörüdür ve $K^{gs} = K^g$ olarak kullanılır. İkincisi ise $NN^{X^A 2X}$ en yakın komşuluk indislerini içeren matristir ve $NN^{\hat{X}^{sen} 2X} = NN^{X^A 2X}$ olarak algoritma devam eder. Gürültü tespiti *gürültüÖnTespit* algoritmasına benzemektedir. $\hat{X}_i^{sen}$ için $\sum_{j=1}^{k^{gs}} I(Y_{NN_{ij}} = S^{(c)}) > \frac{k^{gs}}{2}$ ise “gürültü değil”dir. Aksi durumda “gürültü” olarak kabul edilir. Ardından, $\hat{X}_{K^{adim}="gürültü"}^{sen}$ örneklerinin $X_{K_{Y=S^{(c)}}="gürültü değil"}^{(c)}$ veri setindeki en yakın 1 adet komşuları tespit edilir. En yakın komşuların indislerini ifade eden vektöre $NN^{\hat{X}^{sen} 2X^{(c)}}$ dersek

$$\hat{X}_{K^{adim}="gürültü"}^{adim} = X_{NN_{K^{adim}="gürültü"}^{\hat{X}^{sen} 2X^{(c)}}}^{(c)} \quad (4.6)$$

şeklinde $\hat{X}_{K^{adim}="gürültü"}^{adim}$ sentetik örnekleri güncellenir. Yani, *gürültüSonTespit* ile tespit edilen gürültüler en yakın gürültü olmayan $S^{(c)}$ sınıfı örneğinin bulunduğu konuma yer değiştirir. Böylece gürültülü veri üretilmediği her iki gürültü problemi

türü için dikkate alınmış olur.

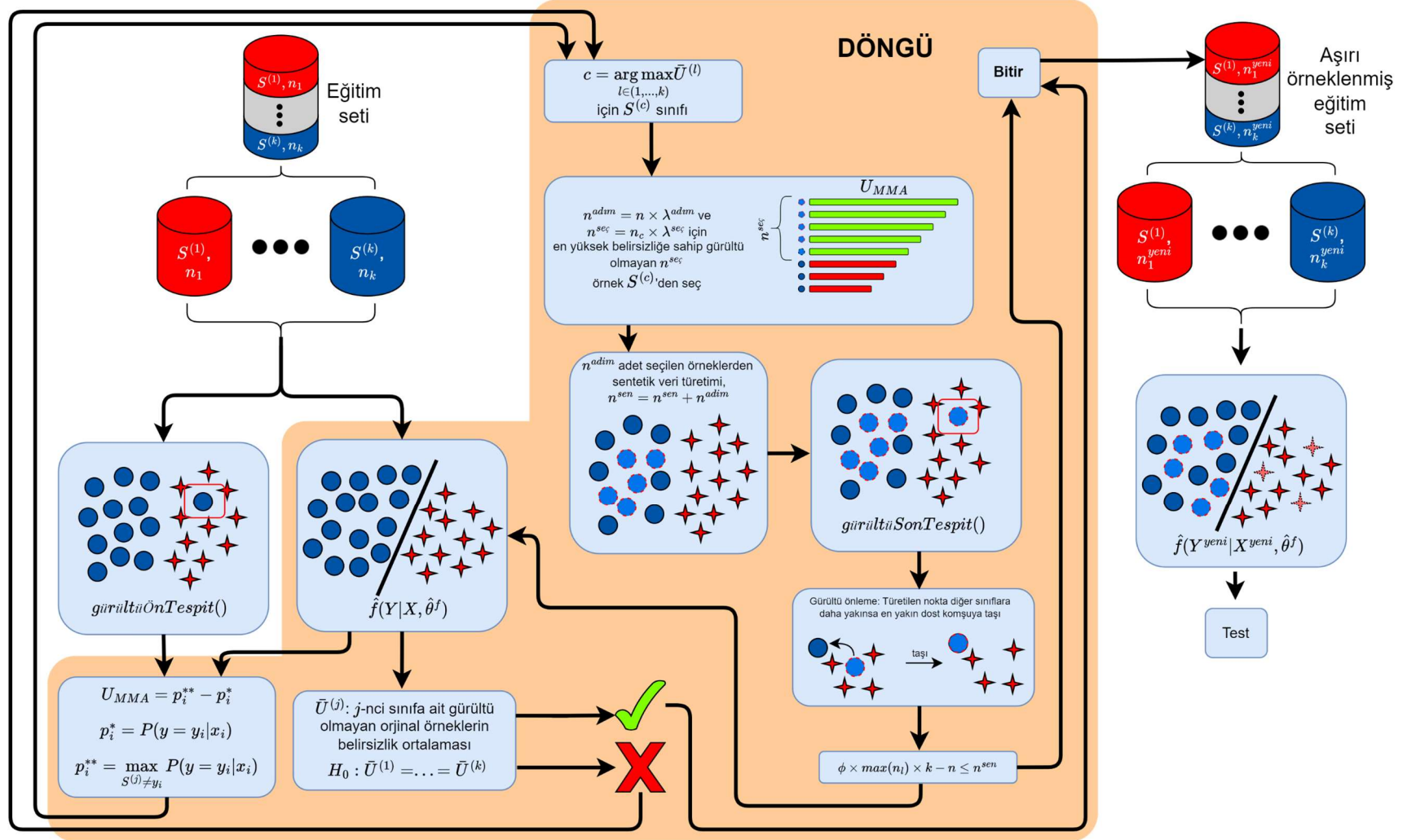
İterasyonun sonunda, Eşitlik 29’da verilen ikinci durdurma kriteri kontrol edilir. Maksimum üretim miktarına ulaşılmışsa, veri üretimi sona erer. Eğer durmamışsa yeni üretilen verilerle birlikte üretilmiş yeni veri setinde eğitime dönülür.

Tablo 4.4 UGO sözde kodları

UGO algoritması

Girdi: $X_{n \times p}, Y_{n \times 1}, f, R, \lambda^{se\check{c}}, \lambda^{adım}, \phi, k^{g\ddot{o}} = 5, k^{gs} = 1, \hat{\theta}^f, \hat{\theta}^R$

1. $K_{n \times 1}^g \leftarrow gürültüÖnTespit(X_{n \times p}, Y_{n \times 1}, k^{g\ddot{o}})$
 2. $n^{adım} \leftarrow n \times \lambda^{adım}$
 3. $\hat{X}_{0 \times p}^{sen} \leftarrow$ boş matris
 4. $\hat{Y}_{0 \times 1}^{sen} \leftarrow$ boş vektör
 5. While $\phi \times \max n_l > n^{sen}$
 - $\hat{f} \leftarrow f(X, Y, \hat{\theta}^f)$ # eğitim
 - $i_{n^{gd} \times 1}^{gd} \leftarrow hangi(K^g = "gürültü değil")$
 - $U_{n^{gd} \times 1} \leftarrow P_{\hat{f}}(y_{i_{gd}}^{(1)' | x_{i_{gd}}, \hat{\theta}^f}) - P_{\hat{f}}(y_{i_{gd}} | x_{i_{gd}}, \hat{\theta}^f)$
 - If $H_0: \bar{U}^{(1)} = \dots = \bar{U}^{(k)}$ için $p > 0.05$
 - break
 - $c \leftarrow \underset{l \in \{1, \dots, k\}}{\operatorname{argmax}} \bar{U}^{(l)}$
 - $n^{se\check{c}} \leftarrow n_c \times \lambda^{se\check{c}}$
 - $i_{n^{se\check{c}} \times 1}^{se\check{c}} \leftarrow \underset{v \in \{1, \dots, n_c\}, |v|=n^{se\check{c}}}{\operatorname{argmax}} \sum_{l \in v} U_l^{(c)}$
 - $\hat{X}_{n^{adım} \times p}^{adım} \leftarrow R(X_{i_{se\check{c}}}^{(c)}, X, Y, S^{(c)}, adet = n^{adım}, \hat{\theta}^R)$
 - $K_{n^{adım} \times 1}^{adım} \leftarrow gürültüSonTespit(\hat{X}^{adım}, X, Y, k^{gs}, S^{(c)})$
 - $NN_{n_c \times 1}^{\hat{X}^{sen} 2X^{(c)}} \leftarrow enYakınKomşu(\hat{X}_{K^{adım}="gürültü"}^{sen}, X_{K_{Y=S^{(c)}}="gürültü değil"}^{(c)}, 1)$
 - $\hat{X}_{K^{adım}="gürültü"}^{adım} \leftarrow X_{NN_{K^{adım}="gürültü"}^{\hat{X}^{sen} 2X^{(c)}}}^{(c)}$
 - $\hat{X}^{sen} \leftarrow satırBirleştir(\hat{X}^{sen}, \hat{X}^{adım})$
 - $\hat{Y}^{sen} \leftarrow birleştir(\hat{Y}^{sen}, S^{(c)})$
 6. $X_{n^{yeni} \times p}^{yeni} \leftarrow satırBirleştir(\hat{X}_{n^{sen} \times p}^{sen}, X)$
 7. $Y_{n^{yeni} \times 1}^{yeni} \leftarrow birleştir(\hat{Y}_{n^{sen} \times 1}^{sen}, Y)$
- Çıktı:** $X_{n^{yeni} \times p}^{yeni}, Y_{n^{yeni} \times 1}^{yeni}, n^{yeni} = n + n^{sen}$



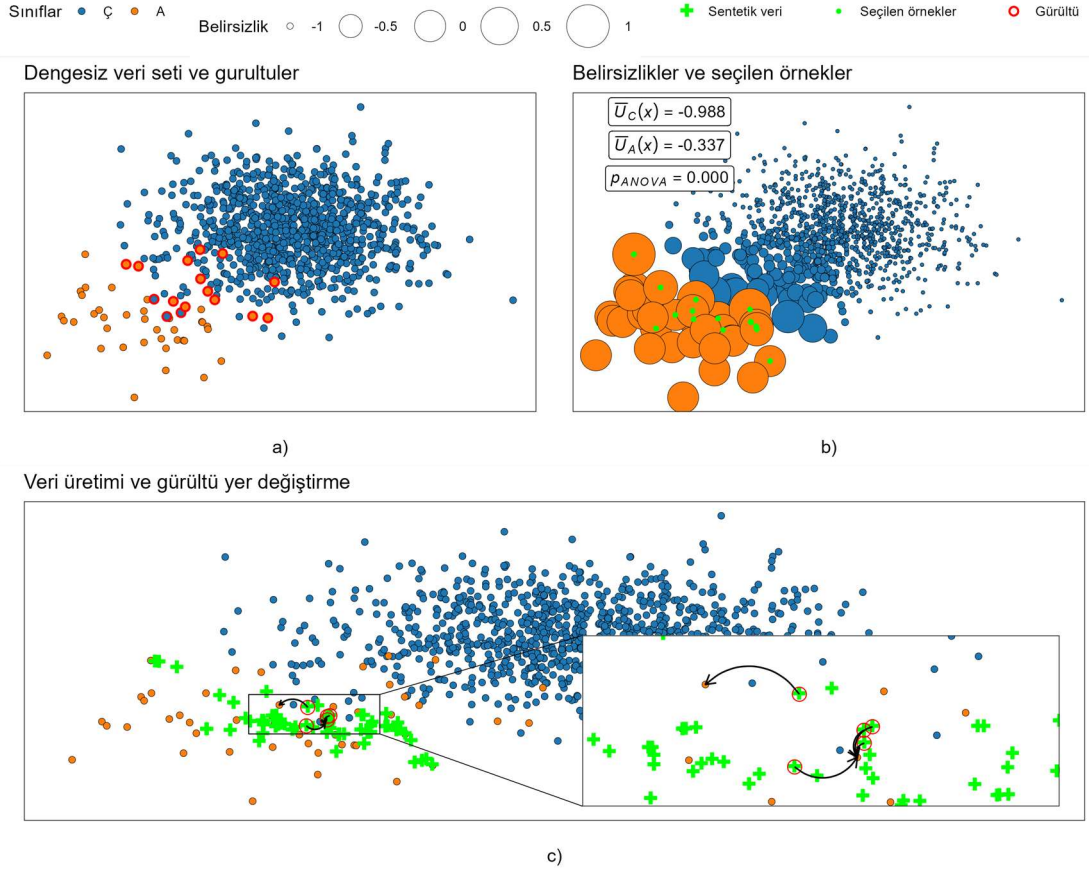
Şekil 4.3 UGO algoritması akış şeması

Tablo 4.4 UGO algoritmasının sözde kodlarını, Şekil 4.3 ise algoritmanın akış şemasını göstermektedir. Algoritma, $X, Y, f, R, \lambda^{sec}, \lambda^{adım}, \phi, k^{gö}, k^{gs}, \hat{\theta}^f$ ve $\hat{\theta}^R$ girdi argümanlarına ihtiyaç duymaktadır. Bunlardan $\hat{\theta}^f$ ve $\hat{\theta}^R$ sınıflayıcıya bağlıdır. $\lambda^{sec} = 0.5, \lambda^{adım} = 0.1, k^{gö} = 5$ ve $k^{gs} = 1$ olarak kabul edilebilir. Böylece yeterince bilgilendirici örnek seçildiğinden ve her adımda fazla hesap maliyeti katmayacağı gibi, denge kontrolü yapılabilecek miktarda küçük bir miktarda gürültü üretmeyecek şekilde veri üretilebilmiş olur. UGO algoritmasının sonucunda epistemik belirsizlik dengesi sağlanmış olan X^{yeni} ve Y^{yeni} veri seti çıktı olarak elde edilir.

4.5. Örnekler

Şekil 4.4'te dengesiz veri seti örneği verilmiştir. Bu veri setinde 2 adet bağımsız değişken ve 2 adet sınıf, $\mathcal{C}$ ve A bulunmaktadır. Çoğunluk sınıf olan $\mathcal{C}$ 'nin örnek sayısı $n^{\mathcal{C}} = 1000$ ve azınlık sınıf A 'nın örnek sayısı $n^A = 50$ 'dir. SMOTE yöntemi ile veri üreten UGOSMOTE'a ait ilk iterasyonun aşamalarını Şekil 4.4'te görmek mümkündür. Şekil 4.4(a)'da dengesiz veri setindeki gürültülerin tespiti görülmektedir. Kırmızı halka ile sarılmış olan örnekler gürültü örnekleridir. Ardından yapay sınır ağları ile eğitilen bu tüm veri seti kullanılarak U_{MMA} belirsizliği gürültü olmayan tüm örnekler için hesaplanmıştır. Şekil 4.4(b)'de noktaların büyüklükleri, belirsizlik miktarını göstermektedir. Sınıflara ait ortalama belirsizlikler $\bar{U}_{\mathcal{C}}(x) = 0.988$ iken $\bar{U}_A(x) = -0.337$ olarak hesaplanmıştır. ANOVA testi ile gruplardaki farklılık incelendiğinde, $p < 0.001$ olduğundan anlamlı farklılık olduğu görülmüştür. Bu farklılık azınlık sınıfın daha yüksek belirsizliğe sahip olduğuna işaret eder. Öyle ise seçilen c sınıfı bu durumda A olmuştur. Azınlık sınıfı örnekleri üzerinden $\lambda^{sec} = 0.5$ olacak şekilde en belirsiz örnekler seçilmiştir. Seçilen bu örnekler yeşil küçük noktalarla belirtilmiştir. Veri üretimi bu noktalar üzerinden yapılacaktır. Şekil 4.4(c)'de üretilmiş olan veriler yeşil kalın "+" sembolü şeklinde görülmektedir. Bu sentetik veriler SMOTE yöntemi ile örneklenen azınlık sınıfı verilerinin en yakın 5 komşusu arasındaki doğru üzerinde üretilmiştir. Bu nedenle gürültülü veri üretilmiş olunabilir. Bu amaçla, yeni üretilmiş verilerin de gürültü olup olmadıkları kontrol edilmiş ve yine kırmızı halkalar ile gösterilmiştir. Grafikte görülmesi zor bir kısımdan bahsettiğimiz için görselin o kısmı ayrıca büyütülmüştür. Büyütülen kısma bakarsak, gürültü sentetik örneklerinin en yakın gürültü olmayan azınlık komşuları ile aralarında bağ oluşturulduğu görülecektir. Bu gürültü örnekleri, algoritmada olduğu gibi o

noktaya taşınır ve ilk iterasyon sona erer.

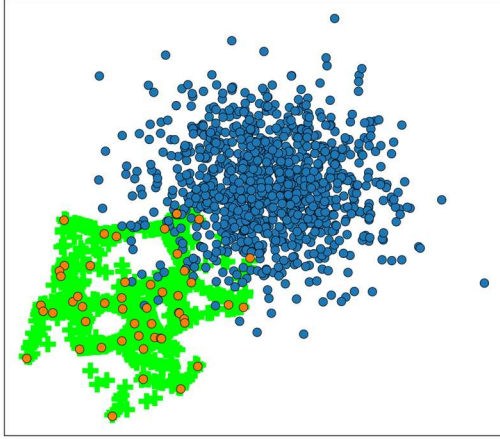


Şekil 4.4 UGOSMOTE'a ait ilk iterasyonun NNET örneği. a) Dengesiz sınıflı veri setinde gürültü örneklerinin tespiti, b) eğitilen NNET modelinin belirsizliklerine göre boyutlandırılan örnekler, belirsizlik ortalamaları, ortalama fark hipotez testi sonucu ve belirsizliklere göre seçilen örnekler, c) veri üretimi sonrası gürültü tespiti ve tespit edilen gürültülerin konum değiştirmesi

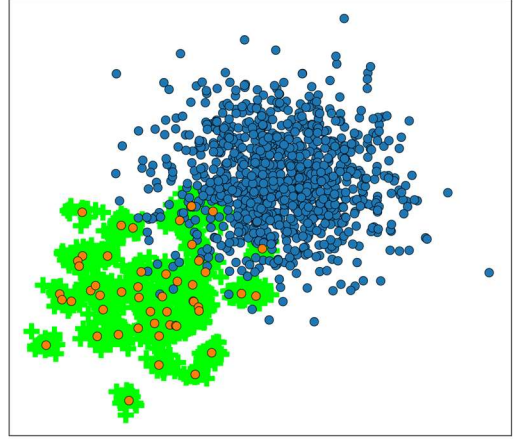
Veri setinde iterasyonlar devam ettiğinde veri setinin son hali, SMOTE, RWO ve UGORWO ile birlikte Şekil 4.5'te verilmiştir. Sol üstte SMOTE, sağ üstte RWO, sol altta UGOSMOTE ve sağ altta UGORWO ile dengelenen veri setleri görülmektedir. SMOTE ve RWO ile veri üretirken, tam denge sağlanmıştır. UGO yöntemlerinde ise durdurma kriterine göre veri üretilmiştir. Görüldüğü üzere, UGO yöntemleri çok az veri üretiminin denge sağlamak için yeterli olduğuna karar vermişlerdir. SMOTE ve RWO, $n^A = 1000$ olana dek veri üretirken UGOSMOTE ve UGORWO $n^A = 113$ verinin sınıf dengesizliği etkisinin ortadan kalkmasında yeterli olduğuna karar vermiştir. Az veri ile bunun sağlanmasının bir nedeni, tüm veri setinde değil, bilgilendirici veriler üzerinden veri üretimi yapılmasıdır. Tüm üretilen örnekler, modelin ayrışmayı anlamasına yardımcı olmaktadır. Ayrıca, UGO yöntemlerinde veri setinin son hali daha az veri içerdiği için son model eğitimi, hem daha hızlı olacaktır, hem de yakınsama sorunu yaşaması olasılığı azalacaktır. Bunların yanı sıra, SMOTE

ve RWO kesişim bölgelerinde veri üretmiştir. Bu da çok fazla gürültü içeren veri setlerine yol açmıştır. Bu durum eğitim sürecinde modelin kafasını karıştıracaktır. Fakat UGOSMOTE ve UGORWO, açıkça görülecek şekilde gürültü verisi üretmemiştir.

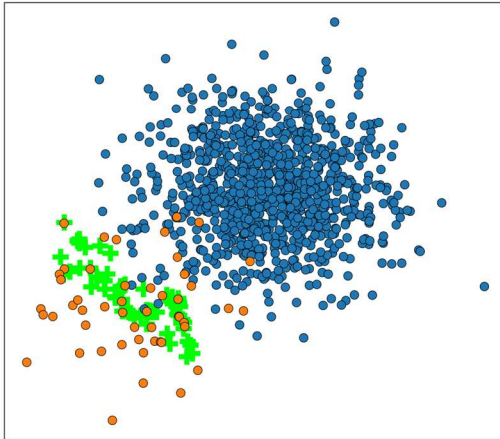
SMOTE ($n_A = 1000, n_B = 1000$)



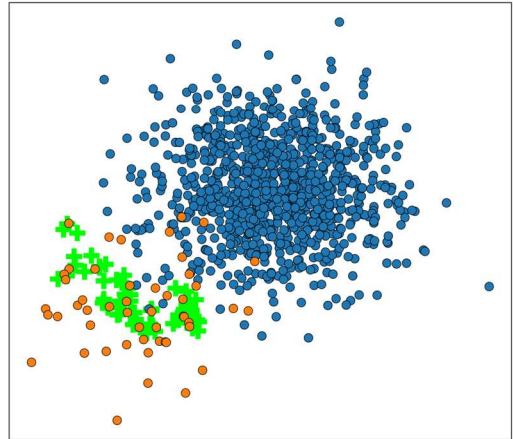
RWO ($n_A = 1000, n_B = 1000$)



UGOSMOTE ($n_A = 113, n_B = 1000$)



UGORWO ($n_A = 113, n_B = 1000$)

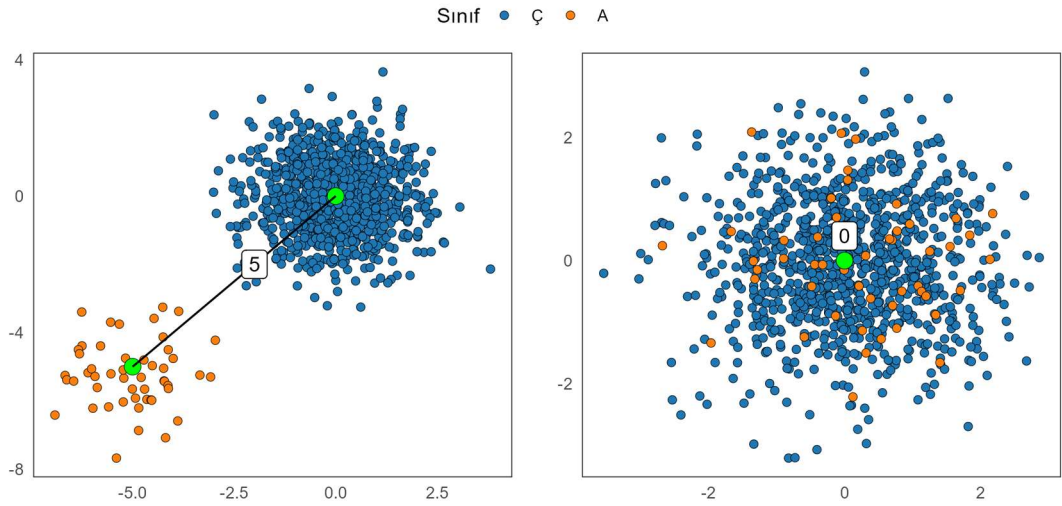


Şekil 4.5 Şekil 4.3'teki verinin farklı yeniden örnekleme yöntemleri uygulandıktan sonraki son halleri. UGO yöntemleri için sınıflayıcı yapay sinir ağlarıdır.

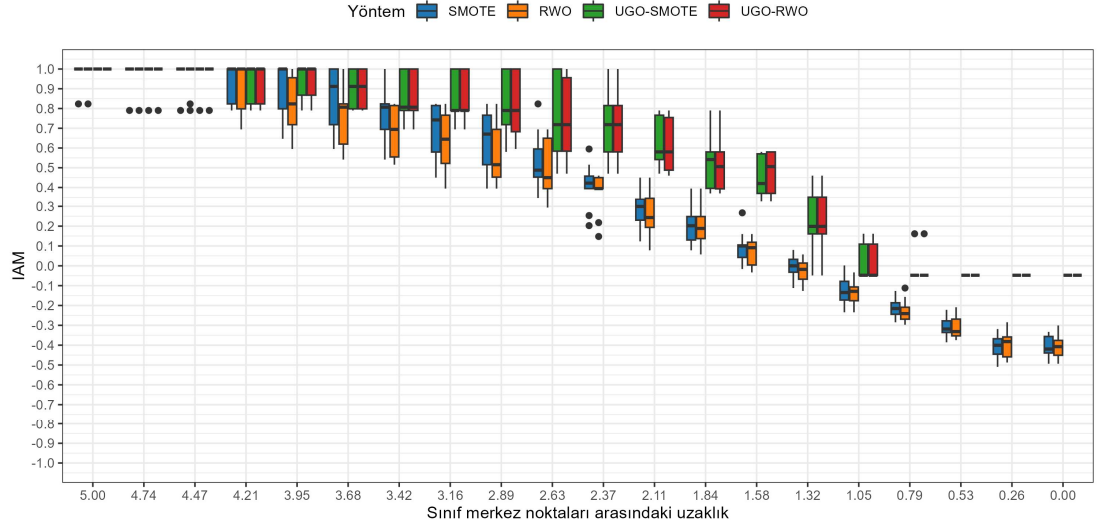
4.6. Gürültü Problemi Durumunda UGO

Gürültü seviyesi bakımından diğer yöntemlere kıyasla simülasyon veri üzerinde karşılaştırmalı performanslar incelenmiştir. Simülasyon verileri hakkında bilgi Şekil 4.6'da görülmektedir. İki değişkenli iki sınıflı veri setinde, sınıflara ait değişkenler farklı ortalamalı normal dağılımlardan çekilmiştir. Azınlık sınıf değişkenleri, $N(-5,1)$ 'den başlayarak $N(0,1)$ 'e doğru ilerleyecek şekilde, çoğunluk sınıf değişkenleri ise her zaman $N(0,1)$ olacak şekilde 20 farklı veri seti üretilmiştir.

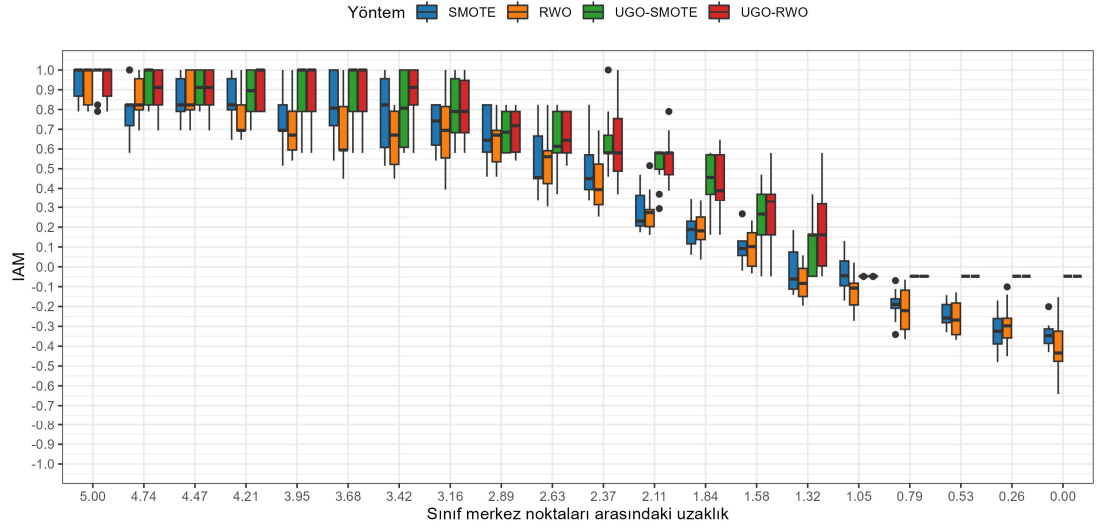
Azınlık sınıf ve çoğunluk sınıf merkez ortalamaları en az gürültülü veride (0,0) ve (-5,-5) şeklindedir. Bu durum, merkezler arası mesafesi 5 olarak adlandırılmıştır. En gürültülü durumda ise çoğunluk ve azınlık sınıf merkezleri (0,0) ve (0,0) şeklindedir ve maksimum sınıf kesişimi durumu yaşanmaktadır. En çok gürültüyü bu durum içermektedir. Azınlık sınıf merkez noktası (-5,-5)'ten (0,0)'a doğru gittikçe, dengesiz doğruluk ölçütü (imbalanced accuracy measure, IAM) performansları Şekil 4.7, 4.8, 4.9 ve 4.10'da lojistik regresyon (LogReg), CART karar ağacı ve Gauss Naif Bayes (GNB) ve Yapay Sinir Ağları (NNET) temel sınıflayıcıları için verilmiştir. Sınıflar arasındaki mesafe azaldıkça, yani sınıf kesişim bölgeleri büyüdükçe doğal gürültü artmaktadır. Gürültünün artması ile IAM değerleri arasındaki ilişki görsellerde incelendiğinde, UGO yöntemlerindeki performans kaybının diğer yöntemlere kıyasla geç yaşandığı açıkça görülmektedir. Bunun anlamı, UGO'un gürültü varlığına karşı daha dayanıklı olduğudur.



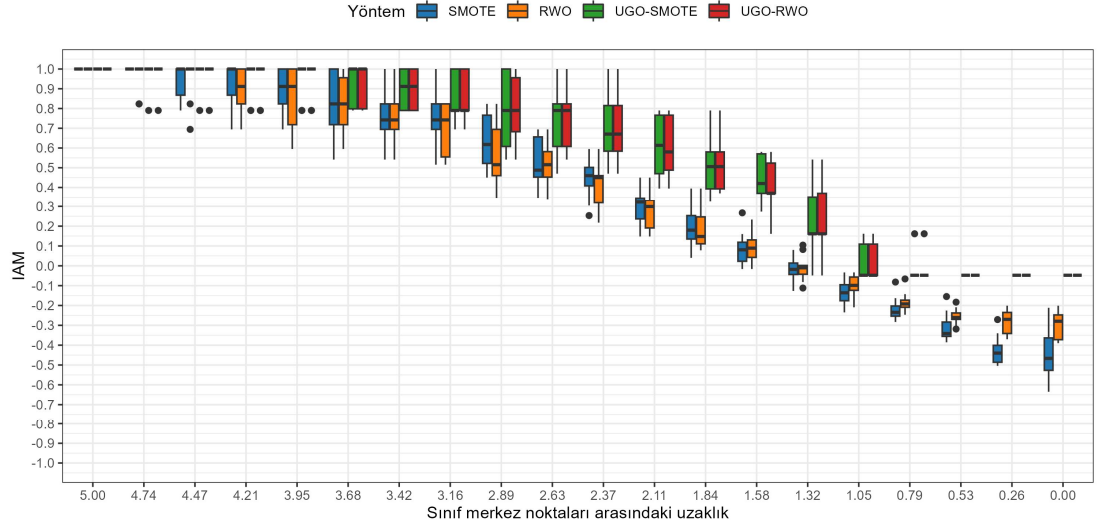
Şekil 4.6 Gürültü ile performans ilişkisi için simülasyon verisi



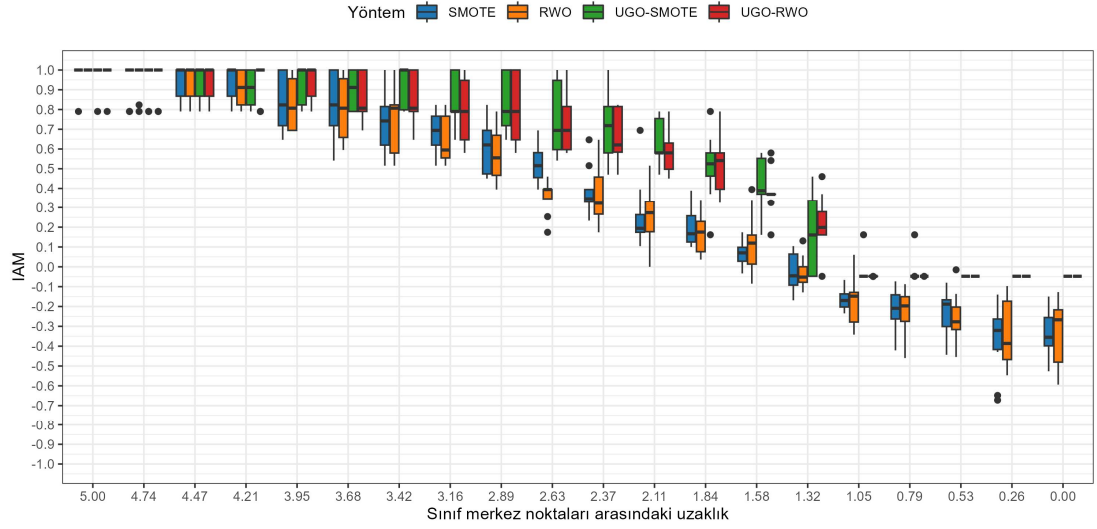
Şekil 4.7 Lojistik regresyon gürültü ile performans arasındaki ilişki



Şekil 4.8 CART gürültü ile performans arasındaki ilişki



Şekil 4.9 GNB gürültü ile performans arasındaki ilişki



Şekil 4.10 NNET gürültü ile performans arasındaki ilişki

5. DENEYSEL TASARIM

5.1. Kullanılan Veri Setleri

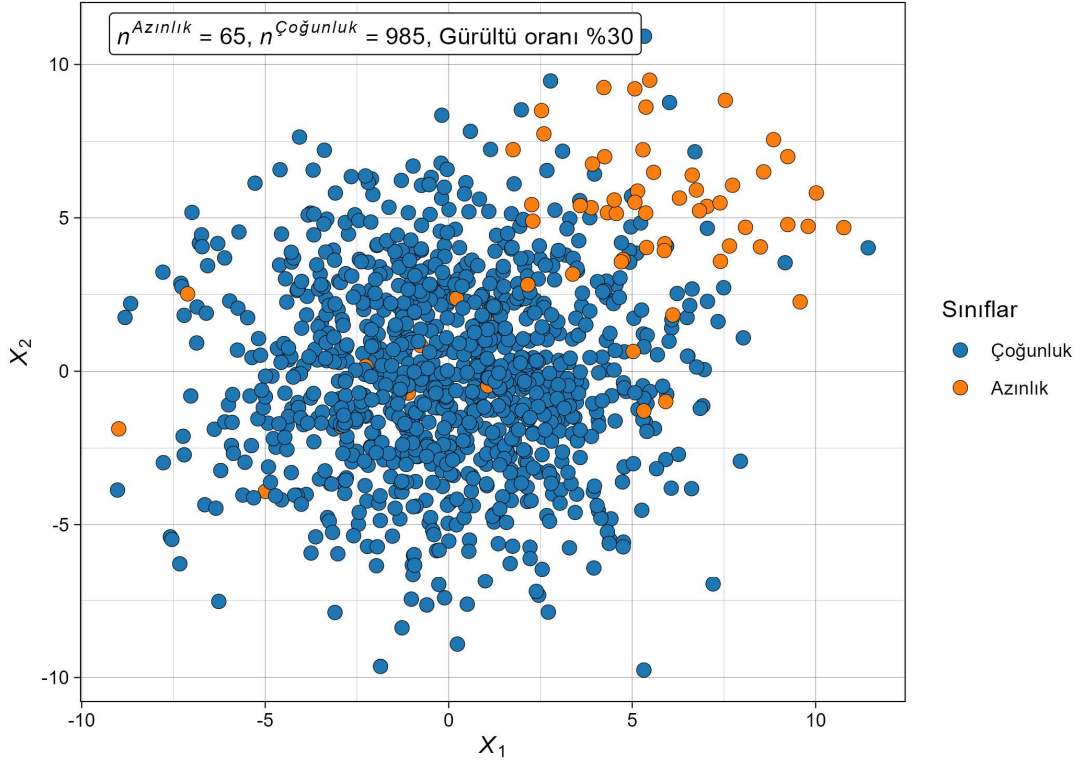
Çalışmada iki adet simülasyon verisi ve farklı çalışmalarda kullanılmış ve kıyaslama için kullanılabilecek açık erişimli 82 veri seti kullanılmıştır.

5.1.1. Simülasyon Veri Setleri

Çalışmada iki adet simülasyon veri seti kullanılmıştır. Bunlardan birincisi, lineer olarak ayrıştırılabilen iki sınıflı ve iki değişkenli bir veri seti. Diğeri ise lineer olarak ayrıştırılamayan fakat non-lineer biçimde ayrıştırılabilen bir veri setidir. Her iki seti de yüzde 30 oranında gürültü içerecek şekilde ayarlanmıştır. Bu iki veri setinin seçilmesindeki amaç, lineer sınıflayıcıların performansını ve non-lineer sınıflayıcıların performansını ayrı ayrı amaçlarına uygun ortamlarda incelemektir.

5.1.1.1. Lineer Ayrıştırılabilir Gürültü İçeren Veri Seti

Şekil 5.1’de lineer olarak ayrıştırılabilen simülasyon verisi görülmektedir. Bu verinin üretiminde, çoğunluk sınıf veri setinin değişkenlerine ait dağılımlar $N(0,3)$, azınlık sınıf değişkenlerine ait dağılımlara ise $N(6,2)$ dağılımları kullanılmıştır. Bu dağılımlara göre simüle edilen veri setinde, çoğunluk sınıf örneklerinden rastgele biçimde, $n^{Azınlık} \times 0.3$ adet örnek seçilerek azınlık sınıf olarak etiketleri değiştirilmiştir. Böylece toplam azınlık sınıf örneklerinin %30’u diğer sınıf bölgesindedir ve gürültüdür.



Şekil 5.1 Lineer olarak ayrıştırılabilen ve gürültü içeren veri seti

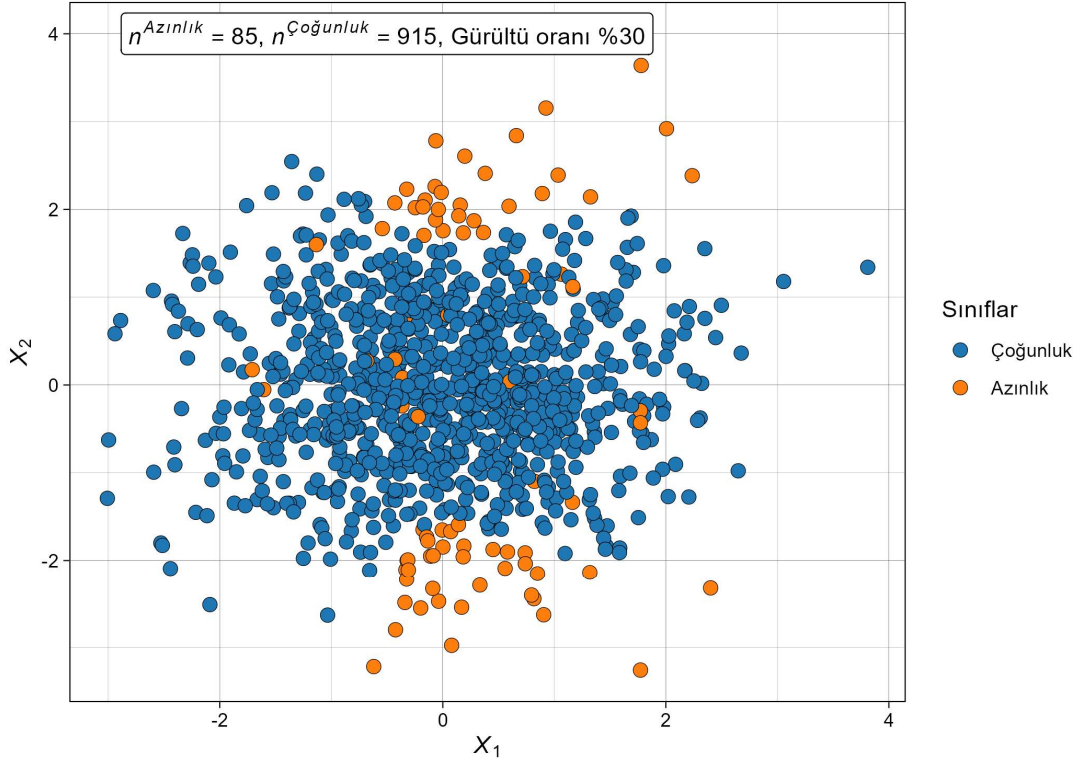
5.1.1.2. Lineer Ayrıştırılamayan Fakat Non-linear Ayrıştırılabilen Veri Seti

Doğrusal olmayan (non-linear) şekilde ayrıştırılabilen veri seti

$$Z = \sin^2 X_1 + \cos X_2 + \sin X_1 \cos X_2$$

$$Y = \left[\frac{1}{1 + \exp -Z} \right] \quad (5.1)$$

şeklinde hesaplanmıştır. Burada $X_1, X_2 \sim N(0,1)$ dağılımından rastgele çekilmiştir. $[\cdot]$ ise en yakın tam sayıya yuvarlamayı ifade etmektedir. $Y = 1$ olan sınıf, çoğunluk sınıfı, $Y = 0$ olan ise azınlık sınıfıdır. Veri üretimi bittikten sonra önceki simülasyonda olduğu gibi, çoğunluk sınıf örnekleri içerisinde rastgele biçimde $n^A \times 0.3$ adet örnek seçilip etiketleri azınlık sınıf etiketine dönüştürülerek %30 oranında gürültü olması sağlanmıştır. Elde edilen veri seti, Şekil 5.2’de görselleştirilmiştir. Burada çoğunluk sınıf örnekleri orta kısımlarda, azınlık sınıf örnekleri ise alt ve üst kısımlarda yoğunlaşmaktadır. Orta kısımlarda olan azınlık örnekleri gürültülerdir. Gürültü dışındaki örneklere bakıldığında, sınıfların lineer ayrışmasa da non-linear biçimde ayrıştığı açıktır.



Şekil 5.2 Non-lineer olarak ayrıştırılabilen ve gürültü içeren veri seti

5.1.2. İki Sınıflı Veri Setleri

Veri setleri UCI (Wang, 2023) ve KEEL (Derrac vd., 2015) veri seti havuzlarından alınmıştır. Tablo 5.1, kullanılan 74 adet veri setine dair örnek sayısı, değişken sayısı ve dengesizlik oranı hakkında bilgi vermektedir. Ayrıca, kısaltma sütununda veri setlerini işaret edecek kısaltmalar verilmiştir. n örnek sayısını, p değişken sayısını, k ise sınıf sayısını ifade etmektedir. En yüksek örnek sayısına sahip veri seti 17898 veri ile pulsar_stars iken en düşük veri sayısına sahip veri seti 192 veri ile glass_0_1_6_vs_2 verisidir. En çok değişken bulunan veri seti 54 değişken ile kddcup_guess_passwd_vs_satan iken en düşük değişken 2 değişken ile banana veri setindedir. Veri setleri toplanırken dengesizlik oranı $n^C/n^A > 5$ olmasına dikkat edilmiştir. En düşük dengesizlik oranı 5.14 ile newthyroid1 veri setinde iken en yüksek dengesizlik oranı 129.44 ile abalone19 veri setindedir.

Tablo 5.1 Kullanılan iki sınıflı veri setleri

Veri seti	Kısaltma	n	p	k	Dengesizlik (n^C/n^A)	Sınıf dağılımı
abalone_17_vs_7_8_9_10	V1	2338	9	2	39.31	2280/58
abalone_19_vs_10_11_12_13	V2	1622	9	2	49.69	1590/32
abalone_20_vs_8_9_10	V3	1916	9	2	72.69	1890/26
abalone19	V4	4174	9	2	129.44	4142/32
abalone2	V5	4177	7	2	29.71	4041/136
abalone9_18	V6	731	9	2	16.4	689/42
banana	V7	2640	2	2	9	2376/264
car_good	V8	1728	15	2	24.04	1659/69
churn	V9	5000	19	2	6.07	4293/707
ecoli_0_1_4_6_vs_5	V10	280	6	2	13	260/20
ecoli_0_1_4_7_vs_2_3_5_6	V11	336	7	2	10.59	307/29
ecoli_0_1_4_7_vs_5_6	V12	332	6	2	12.28	307/25
ecoli_0_1_vs_2_3_5	V13	244	7	2	9.17	220/24
ecoli_0_1_vs_5	V14	240	6	2	11	220/20
ecoli_0_2_3_4_vs_5	V15	202	7	2	9.1	182/20
ecoli_0_2_6_7_vs_3_5	V16	224	7	2	9.18	202/22
ecoli_0_3_4_6_vs_5	V17	205	7	2	9.25	185/20
ecoli_0_3_4_7_vs_5_6	V18	257	7	2	9.28	232/25
ecoli_0_3_4_vs_5	V19	200	7	2	9	180/20
ecoli_0_4_6_vs_5	V20	203	6	2	9.15	183/20
ecoli_0_6_7_vs_3_5	V21	222	7	2	9.09	200/22
ecoli_0_6_7_vs_5	V22	220	6	2	10	200/20
ecoli2	V23	336	7	2	5.46	284/52
ecoli3	V24	336	7	2	8.6	301/35
ecoli4	V25	336	7	2	15.8	316/20
Fishing1	V26	1182	5	2	7.82	1048/134
Fishing2	V27	1182	5	2	5.64	1004/178
glass_0_1_4_6_vs_2	V28	205	9	2	11.06	188/17
glass_0_1_6_vs_2	V29	192	9	2	10.29	175/17
glass2	V30	214	9	2	11.59	197/17
glass6	V31	214	9	2	6.38	185/29
kddcup_buffer_overflow_vs_back	V32	2233	32	2	73.43	2203/30
kddcup_guess_passwd_vs_satan	V33	1642	54	2	29.98	1589/53
kr_vs_k_one_vs_fifteen	V34	2244	20	2	27.77	2166/78
kr_vs_k_three_vs_eleven	V35	2935	20	2	35.23	2854/81
kr_vs_k_zero_one_vs_draw	V36	2901	20	2	26.63	2796/105
kr_vs_k_zero_vs_eight	V37	1460	20	2	53.07	1433/27
kr_vs_k_zero_vs_fifteen	V38	2193	20	2	80.22	2166/27
led7digit_0_2_4_5_6_7_8_9_vs_1	V39	443	7	2	10.97	406/37
newthyroid1	V40	215	5	2	5.14	180/35
page_blocks_1_3_vs_4	V41	472	10	2	15.86	444/28
poker_8_9_vs_5	V42	2075	10	2	82	2050/25
poker_8_9_vs_6	V43	1485	10	2	58.4	1460/25
pulsar_stars	V44	17898	8	2	9.92	16259/1639
Satellite2	V45	6435	36	2	9.28	5809/626
Satellite3	V46	6435	36	2	8.1	5728/707
shuttle_2_vs_5	V47	3316	9	2	66.67	3267/49
thyroid_disease	V48	3771	25	2	15.32	3540/231
vowel0	V49	988	13	2	9.98	898/90
Vowel1	V50	990	9	2	10	900/90
Vowel10	V51	990	9	2	10	900/90

Vowel2	V52	990	9	2	10	900/90
Vowel3	V53	990	9	2	10	900/90
Vowel4	V54	990	9	2	10	900/90
Vowel5	V55	990	9	2	10	900/90
Vowel6	V56	990	9	2	10	900/90
Vowel7	V57	990	9	2	10	900/90
Vowel8	V58	990	9	2	10	900/90
Vowel9	V59	990	9	2	10	900/90
winequality_red_3_9_vs_5	V60	1482	11	2	58.28	1457/25
winequality_red_4	V61	1599	11	2	29.17	1546/53
winequality_red_8_vs_6	V62	656	11	2	35.44	638/18
winequality_red_8_vs_6_7	V63	855	11	2	46.5	837/18
yeast_0_2_5_6_vs_3_7_8_9	V64	1004	8	2	9.14	905/99
yeast_0_2_5_7_9_vs_3_6_8	V65	1004	8	2	9.14	905/99
yeast_0_3_5_9_vs_7_8	V66	506	8	2	9.12	456/50
yeast_1_2_8_9_vs_7	V67	947	8	2	30.57	917/30
yeast_1_4_5_8_vs_7	V68	693	8	2	22.1	663/30
yeast_1_vs_7	V69	459	7	2	14.3	429/30
yeast_2_vs_8	V70	514	8	2	9.08	463/51
yeast3	V71	1484	8	2	8.1	1321/163
yeast4	V72	1484	8	2	28.1	1433/51
yeast5	V73	1484	8	2	32.73	1440/44
yeast6	V74	1484	8	2	41.4	1449/35

5.1.3. İkiden Fazla Sınıflı Veri Setleri

İkiden fazla sınıfa sahip 8 adet veri seti Tablo 5.2’de verilmiştir. Sınıf sayıları 3 ile 10 arasında değişiklik göstermektedir. En düşük örnek sayısı 132 ile hayes_roth verisinde, en yüksek örnek sayısı ise 4331 ile schedulingData veri setindedir. Değişken sayısı en az olan veri seti 4 ile hayes_roth iken en yüksek olan 24 ile schedulingData’dır. Dengesizlik oranı en düşük olan 1.89 ile contraceptive iken en yüksek olan 39.18 ile thyroid’dır. Sınıf dağılım sütununda her bir sütuna ait örnek sayıları büyükten küçüğe sıralı olarak verilmiştir. Çoğunluk sınıf, en yüksek örneğe sahip sınıf, azınlık sınıf ise en düşük örneğe sahip sınıftır.

Tablo 5.2 Kullanılan ikiden fazla sınıflı veri setleri

Veri seti	Kısaltma	n	p	k	Dengesizlik ($n^{coğ}/n^{az}$)	Sınıf dağılımı
contraceptive	V75	1473	9	3	1.89	629/511/333
csgo	V76	1133	20	3	6.59	560/488/85
Fishing	V77	1182	5	4	3.37	452/418/178/134
hayes_roth	V78	132	4	3	1.7	51/51/30
penbased	V79	1100	16	10	1.1	115/115/114/114/114/106/106/106/105/105
schedulingData	V80	4331	24	4	8.54	2211/1347/514/259
thyroid	V81	720	21	3	39.18	666/37/17
wine	V82	178	13	3	1.48	71/59/48

5.1. Karşılaştırılan Yöntemler

Çalışmada yeniden örnekleme yöntemleri farklı koşullar altında karşılaştırılmaktadır. Karşılaştırılacak yöntemler, yeniden örnekleme yapılmamış durum (NORES), RUS, ROS, SMOTE, RWO, UGOROS, UGOSMOTE ve UGORWO'dur. SMOTE için en yakın komşu sayısı 5 olarak kabul edilmiştir. UGO yöntemlerinde ise belirsizlik örneklemesinden faydalanabilmek için $\lambda^{seç} = 0.5$ alınmıştır. Veri üretim adımlarını takip edebilmek için düşük olacak şekilde $\lambda^{adım} = 0.1$ olarak alınmıştır.

5.2. Değerlendirme Ölçütleri

Sınıflama modellerinin performansını ölçerken sıklıkla kullanılan ölçütlerin başında ACC gelmektedir. ACC yerine hata oranı da kullanılır. Bu iki ölçüt aynı şeyi tersten ölçer. Sınıf dengesizliği durumunda ACC çoğunluk sınıftan yana yanlılık gösterir (Bekkar, vd., 2013). Örnek vermek gerekirse, 90 adet sağlıklı ve 10 adet hasta bulunan bir veri seti olsun. Bir sınıflayıcı da, hiçbir bilgiyi dikkate almadan tüm örneklere “sağlıklı” şeklinde tahminde bulunuyor olsun. Öyleyse %90 oranında yüksek bir ACC değerine sahip olacaktır. Halbuki bu model çok başarısız bir modeldir. Bu nedenle, performans hesabı yapılırken tüm sınıfları dikkate alan ölçütler kullanılmalıdır. Sınıf dengesizliği durumunda ACC ölçütüne bakarken, doğru negatif oranı ve doğru pozitif oranı ölçümleri de gözlemlenmelidir. Tek bir ölçüt ile ölçmek isteyen araştırmacılar genelde F1 skoru, Geometrik Ortalama (GMEAN), AUC (ROC eğrisi altında kalan alan), Matthew korelasyon katsayısı (MCC) gibi ölçütler kullanılmaktadır. AUC dışında kalan örnekler, Tablo 5.3'te verilen karmaşıklık matrisi yardımıyla hesaplanmaktadır. MCC ölçütü, seçilen bir sınıfı temel alarak hesaplanmaması nedeniyle iki sınıflı sınıflamada diğer ölçütlere göre daha yansızdır (Chicco, 2017; Chicco and Jurman, 2020). Fakat kullanılan bu ölçütlerin, ACC kadar olmasa da hala dengesizlik probleminden etkilendiği bilinmektedir (Zhu, 2020; Gaudreault vd., 2021).

Tablo 5.3 Karmaşıklık matrisi

		Tahmin	
		Pozitif	Negatif
		Doğru Pozitif (DP)	Yanlış Pozitif (YP)
Gerçek	Pozitif		
	Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Çok sınıflı durumlarda bu ölçütleri kullanmak için one-vs-all veya one-vs-one yaklaşımları kullanılmaktadır. One-vs-all yaklaşımında, her bir sınıfın, diğer tüm sınıflara karşı iki sınıflı varsayılan durumuna göre seçilen ölçüt hesaplanır. k adet hesaplanan ölçütün ortalaması, performansı verir. One-vs-one yaklaşımında ise, $\binom{k}{2}$ adet ikili sınıf karşılaştırmaları ile performanslar hesaplanıp ortalaması alınır. Dengesiz doğruluk ölçüsü (Imbalanced Accuracy Measure, IAM) (Mortaz, 2020), hem az sınıflı hem de çok sınıflı sınıf dengesizliği durumunda kullanılmak üzere geliştirilmiş ölçütlerden birisidir. Bu ölçüt, mevcut ölçütleri temel alarak geliştirilmiştir ve karmaşık hesaplamalara ihtiyaç duymaz. Hem ikili hem de çoklu sınıf durumunda, sınıflardan yana yanlılık göstermediği için, sınıf dengesizliğinde kullanılabilir. k sınıflı bir sınıflama için $k \times k$ büyüklüğündeki karmaşıklık matrisi ele alınarak IAM,

$$IAM = \frac{1}{k} \sum_{i=1}^k \frac{c_{ii} - \max(\sum_{j \neq i}^k c_{ij}, \sum_{j \neq i}^k c_{ji})}{\max(c_{.i}, c_{i.})} \quad (5.2)$$

şeklinde hesaplanır. Burada c_{ij} , çok sınıflı karmaşıklık matrisinin i -nci satır j -nci sütun elemanı, $c_{.i}$, i -nci sütun toplamı, $c_{i.}$ ise i -nci satır toplamıdır. IAM, $[-1,1]$ aralığında değer alır ve ne kadar 1'e yakın ise, modelin o kadar örnekleri yanlış etiketlememede iyi olduğunu gösterir. Sonuç olarak, IAM, one-vs-all ve one-vs-one gibi stratejilere ihtiyaç duymadan basit işlemle yorumlanabilir dengesizliğe dayanıklı bir performans ölçütüdür.

5.3. Deneysel Tasarım Uygulama Protokolü

Performans sonuçları IAM ölçütlerine göre raporlanmıştır. 14 farklı makine öğrenme sınıflayıcısı kullanılmıştır: BART, BNNET, CART, GKNB, GNB, KNN, LDA, LogReg, NNET, QDA, RF, SVMLineer, SVMRadyal. Raporlamalar, genel, lineer, non-lineer, sinir ağları, ağaç yapılı, yoğunluk temelli, komşuluk temelli ve Bayesci olmak üzere 8 gruba ayrılmıştır. Bu gruplar Tablo 5.4'te verilmiştir.

Tablo 5.4 Çalışmada kullanılan sınıflama yöntemlerinin gruplandırılması

Gruplar	İçerisindeki sınıflama yöntemleri
Genel	BART, BNNET, C5.0, CART, GKNB, GNB, KNN, LDA, LogReg, NNET, QDA, RF, SVMLineer, SVMRadyal
Lineer	SVMLineer, LogReg, LDA
Non-lineer	BART, BNNET, C5.0, CART, GKNB, GNB, KNN, NNET, RF, SVMRadyal
Sinir ağları	NNET, BNNET
Ağaç temelli	BART, C5.0, CART, RF
Yoğunluk temelli	GNB, GKNB, LDA, QDA
Komşuluk temelli	KNN
Bayesci	BART, BNNET

Bayesci yöntemlerde belirsizlikler hesaplanırken Valdenegro-Toro ve Mori'nin (2022) çalışması referans alınmıştır. Bu referansa göre sonsal dağılımdan 100 örnek çekilerek model ortalaması alınmıştır. Belirsizlikler bu ortalamalara göre hesaplanmıştır. Sınıflara göre varyansı sıfır olabilen değişkenler veri setlerinden otomatik olarak çıkarılmıştır. Çalışmada kullanılan sınıflama yöntemlerine ait hiperparametreler Tablo 5.5'te verilmiştir.

Performanslar 10 katlı çapraz geçerlilik ile hesaplanmıştır. Performans ölçütlerinden sınıf dengesize dayanıklı olduğu için IAM ölçütü dikkate alınmıştır. Yöntemler arasında anlamlı farklılık olup olmadığı Friedman testi ile test edilmiştir. İkili karşılaştırmalar, Bonferroni düzeltmeli bağımlı grup Wilcoxon testleri ile yapılmıştır. Tüm hipotez testleri için önem seviyesi 0.05 olarak alınmıştır.

Tablo 5.5 Sınıflama yöntemlerinin eğitiminde kullanılan hiperparametreler ve kullanıldıkları dil paketleri

Sınıflama yöntemi	Hiperparametreler	R veya Python paket(ler)i
BART	ntree = 20	BART (R)
BNNET	Gizli katman sayısı = 2 Gizli katman başına düğüm sayısı = 20 Dropout oranı = 0.25 Warm-up eğitim adımı = 100 Optimizasyon fonksiyonu = Adam Patience = 20 Aktivasyon fonksiyonu = relu	torch, torchbnn (Python)
C5.0	CF = 0.25 earlyStopping = TRUE	C50 (R)
CART	cp = 0.01 minsplit = 20	Rpart (R)
GKNB	bw tahmini = yansız çapraz geçerlilik (ucv)	Kendi kodum (R)
GNB	Yok	Kendi kodum (R)
KNN	k = 5	Caret (R)
LDA	Yok	Kendi kodum (R)
LogReg	maxit = 1000	2 sınıf için: stats (R) 2'den fazla sınıf için: nnet (R)
NNET	Gizli katman sayısı = 2 Gizli katman başına düğüm sayısı = 20 Warm-up eğitim adımı = 100 Optimizasyon fonksiyonu = Adam Patience = 20 Aktivasyon fonksiyonu = relu	torch (Python)
QDA	Yok	Kendi kodum (R)
RF	num.trees = 500 $mtry = \sqrt{p}$ max.depth = ∞ min.node.size = 1 min.bucket = 1	ranger (R)
SVMLineer	C = 0.95	kernlab (R)
SVMRadyal	C = 0.95 sigma = 0.5 (kernel parametresi)	kernlab (R)

6. DENEYSEL BULGULAR

6.1. Deney Sonuçları

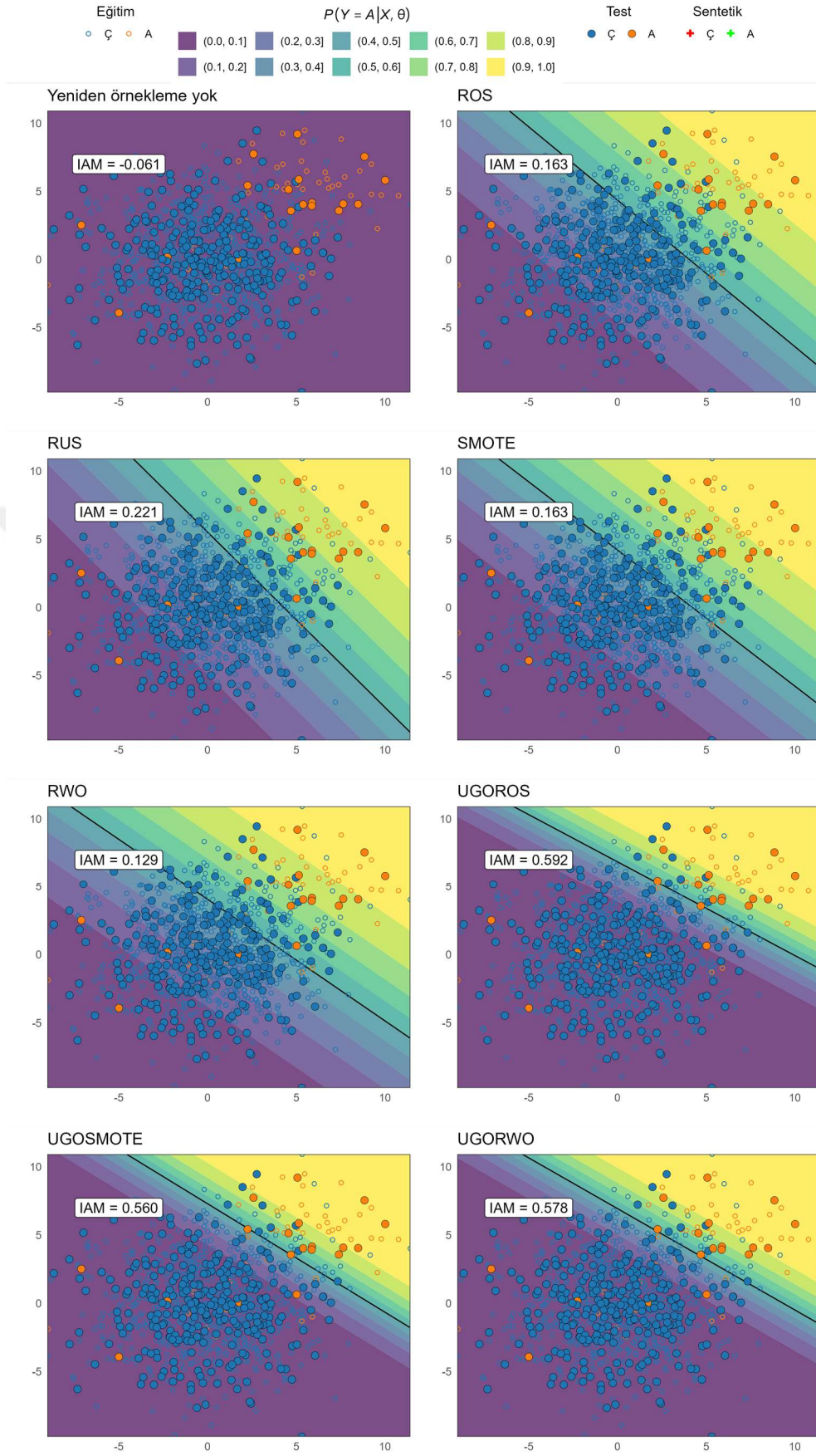
Deneyler simülasyon verileri ve literatürde sınıf dengesizliği çalışmalarında daha önce kullanılmış veri setleri kullanılarak gerçekleştirilmiştir. Simülasyon çalışması, bahsi geçen problemlerin var olduğu veri setlerine örnek olarak oluşturulmuşlardır. Literatürden alınan veri setleri ise, UGO yönteminin diğer yeniden örnekleme yöntemleri ile karşılaştırılmasının yanı sıra farklı koşullar altındaki sonuçları incelemeye ve diğer çalışmalarla karşılaştırma yapılabilmesine olanak sağlamayı amaçlamaktadır.

6.1.1. Simülasyon Sonuçları

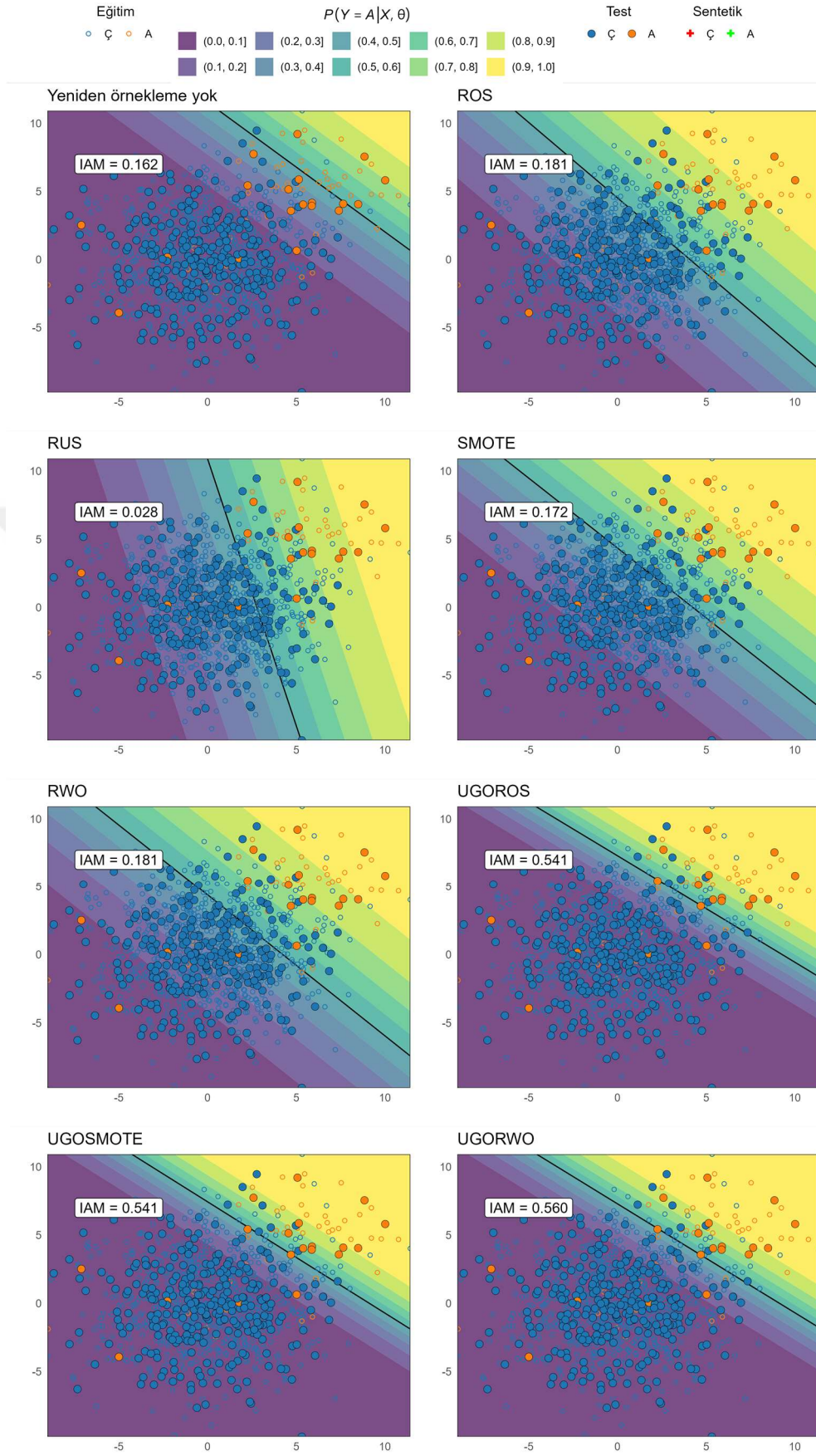
Lineer ayrıştırılabilen ve non-lineer ayrıştırılabilen olmak üzere iki farklı simülasyon verisi oluşturulmuştur. Her bir veri seti, kendi yapısına uygun sınıflayıcılar altında eğitilmiş ve UGO yöntemleri diğer yeniden örnekleme yöntemleri ile bu koşullar altında karşılaştırılmıştır.

6.1.1.1. Lineer Ayrıştırılabilen Simülasyon Veri Seti Sonuçları

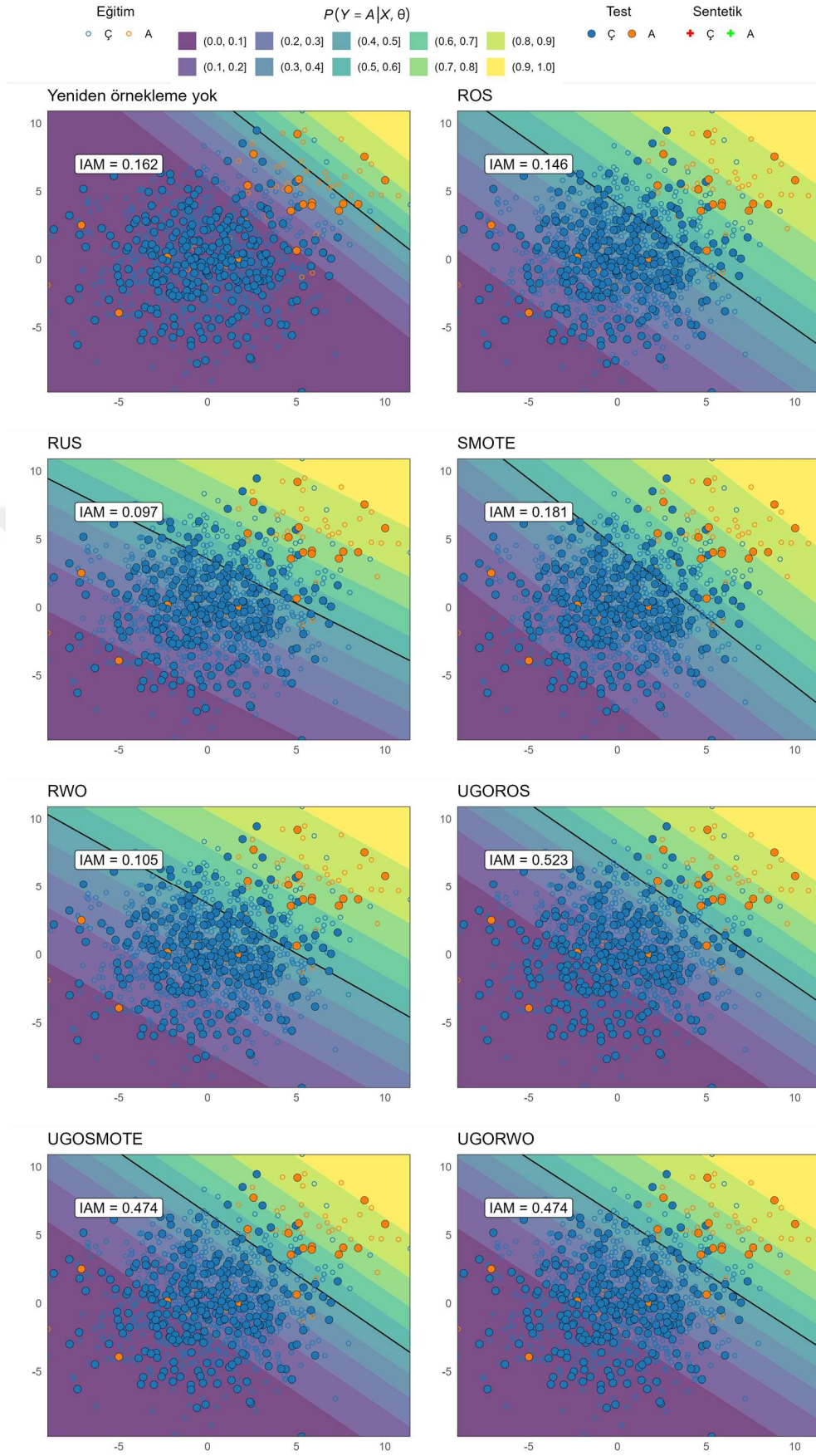
Şekil 6.1, 6.2 ve 6.3'te lineer ayrıştırılabilen simülasyon verisinde yeniden örnekleme olmayan durum, ROS, RUS, SMOTE, RWO, UGOROS, UGOSMOTE, UGORWO yöntemleri ile sırasıyla SVMLineer, LogReg ve LDA sınıflayıcılarına ait kontör grafikleri IAM performansları ile birlikte verilmiştir. Veri seti, %30 test ve %70 eğitim seti olarak iki kısma ayrılmıştır. Grafiklerde halka şeklinde gösterilen noktalar eğitim seti, için dolu noktalar ise test verisidir. “+” şeklindeki noktalar ise sentetik verileri göstermektedir. Tüm şekillerde, yeniden örnekleme yapılmadığı durumda IAM değeri 0'a yakındır. Bu da rastgele sınıflayıcıdır. ROS, RUS, SMOTE ve RWO yöntemleri denge sağlamayı amaçlarken, gürültü örneklerinden etkilendikleri için azınlık sınıftan yana ters yanlılık oluşturmuşlardır. Performansları 0.1-0.2 aralığındadır. UGO yöntemleri gürültülü örneklerden etkilenmeden veri ürettiği için IAM değerleri 0.4-0.6 aralığındadır. Tüm sınıflayıcılarda UGO yöntemleri en iyi sonucu vermiştir.



Şekil 6.1 Lineer ayrışan veri seti SVMLineer kontör grafikleri



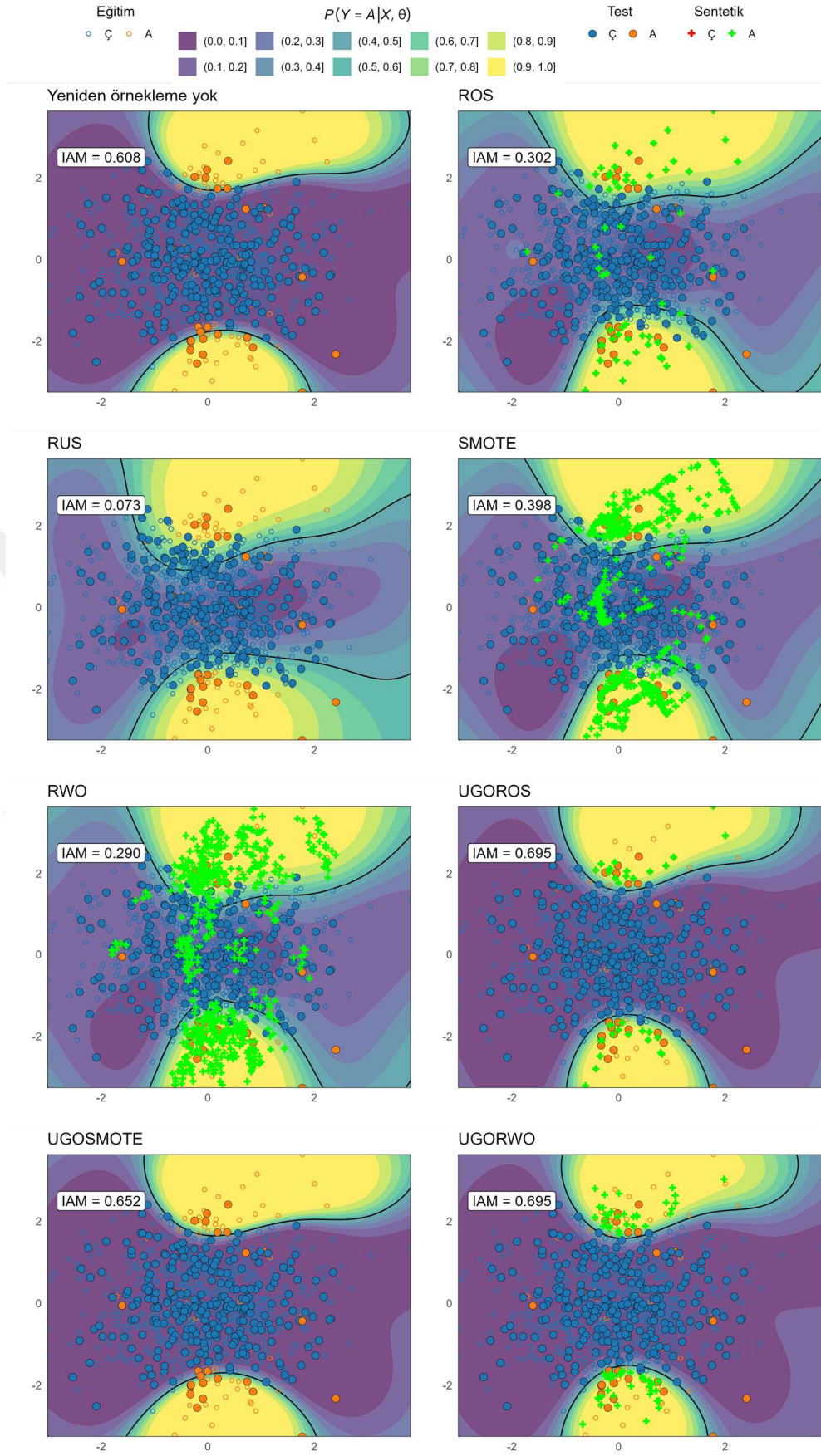
Şekil 6.2 Lineer ayrışan veri seti LogReg kontör grafikleri



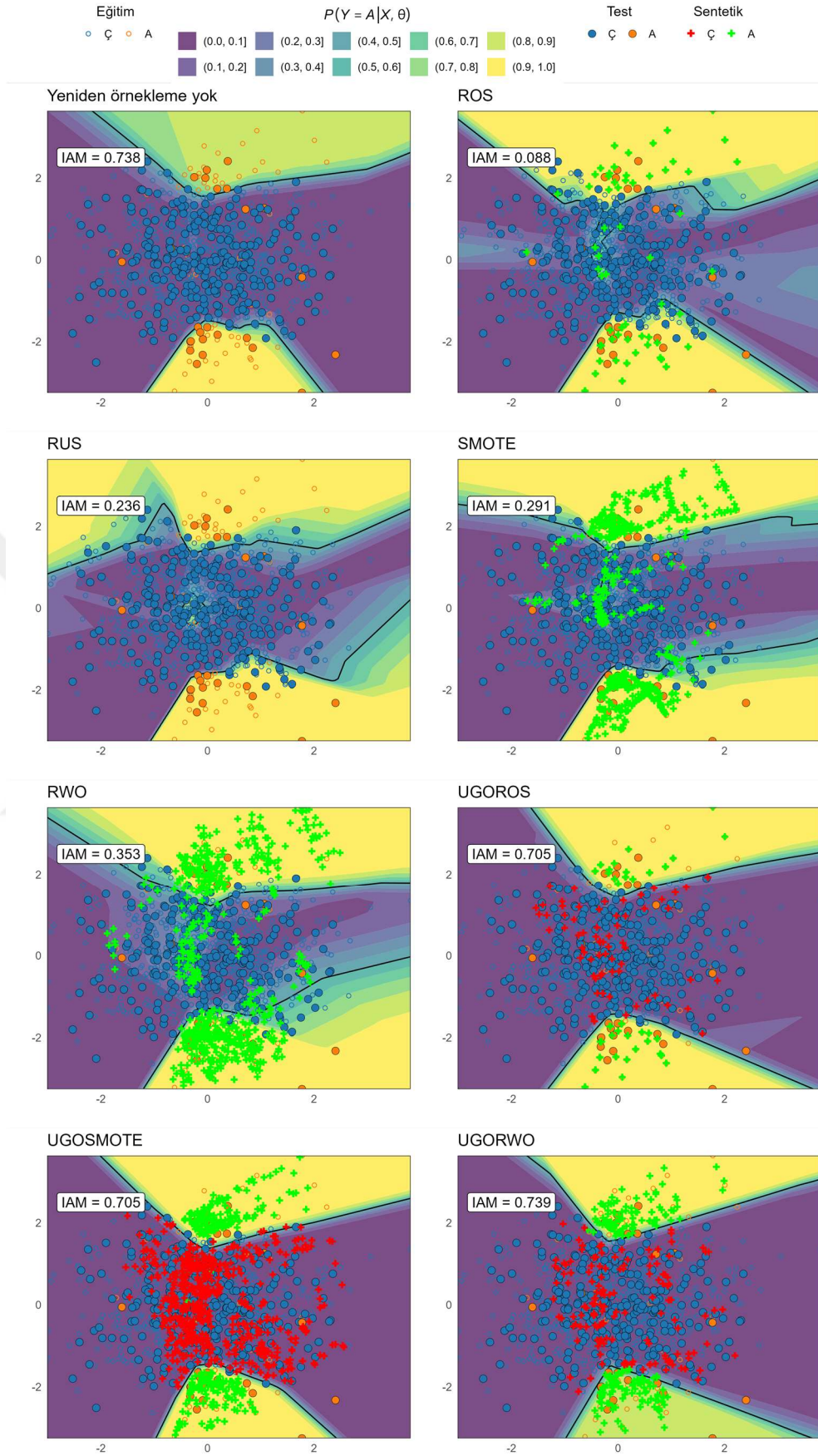
Şekil 6.3 Lineer ayrılan veri seti LDA kontör grafikleri

6.1.1.2. Non-lineer biçimde ayrıştırılabilen simülasyon sonuçları

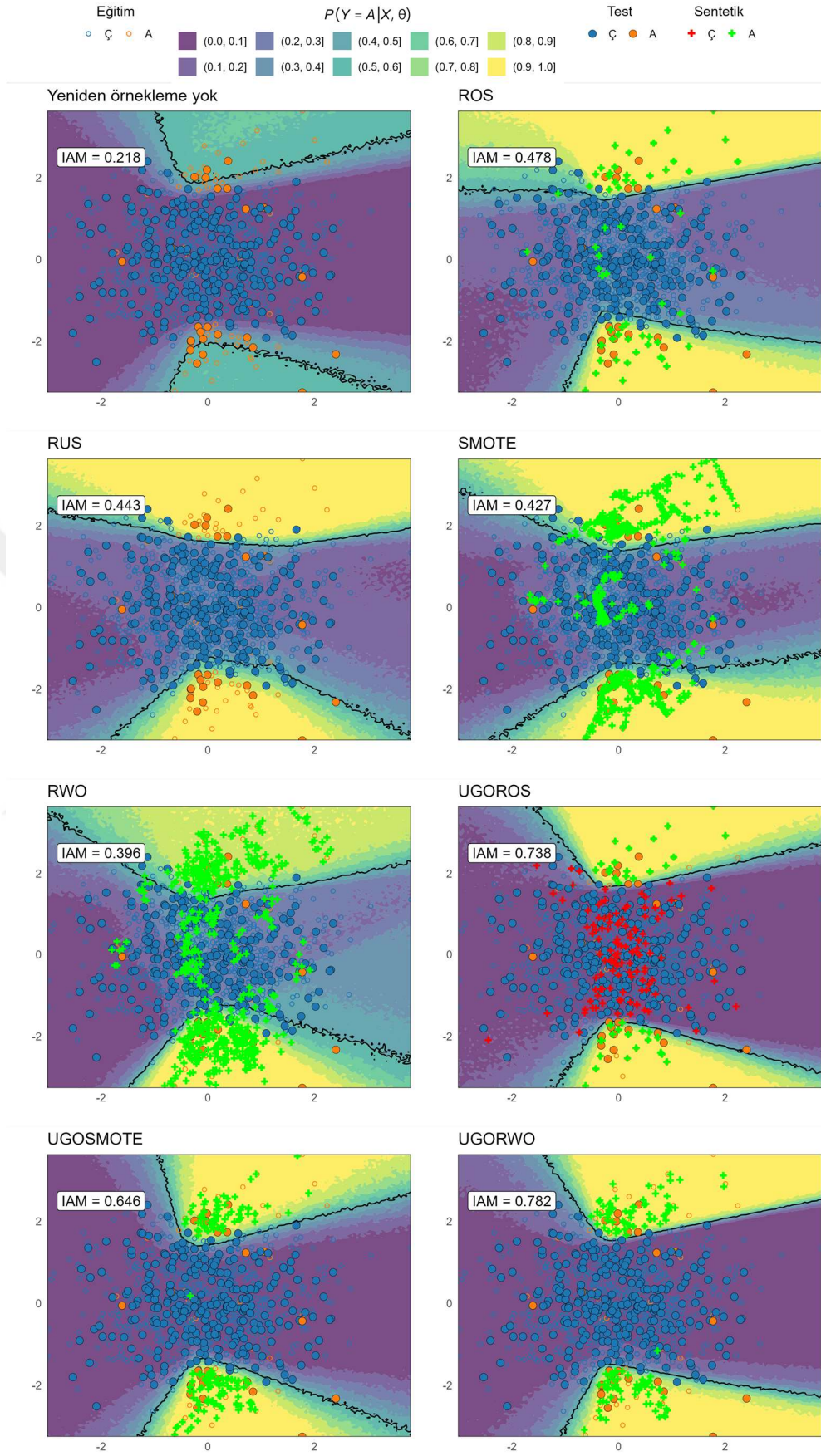
Şekil 6.4, 6.5, 6.6, 6.7, 6.8 ve 6.9’da sırasıyla SVMRadial, BART, NNET, BNNET, KNN ve GKNB sınıflama yöntemlerinin yeniden örnekleme olmayan, ROS, RUS, SMOTE, RWO, UGOROS, UGOSMOTE ve UGORWO yöntemlerinin performansları karşılaştırılmıştır. Grafiklerdeki kontör bilgisinin sararması, sınıflayıcının tahmin ettiği azınlık sınıf olma olasılığının arttığı anlamına gelir. Siyah çizgi ise 0.5 olasılık karar sınırındır. Tüm sınıflayıcılarda SMOTE, ROS ve RUS yöntemleri, yeniden örnekleme yapılmayan duruma kıyasla daha kötü IAM sonucu elde edebilmiştir. Buradan anlaşılan, gürültü durumunda bu tarz yöntemlerin ne kadar tehlikeli olmasıdır. UGO yöntemleri ise tüm sınıflayıcılarda diğer yeniden örnekleme yöntemlerinden daha başarılı olmuş ve NNET hariç tüm sınıflayıcılarda yeniden örnekleme yapılmadan yapılan eğitimden daha başarılı olmuşlardır. Gürültü örneklerinin olduğu bölgelerde klasik yeniden örnekleme yöntemlerinin ürettiği verilere ve sınıflayıcı yanıltarak azınlık sınıfı tahminde yapılmasına dikkat edelim. UGO yöntemlerinde gürültü örnekleri üretilmemiştir ve bu durumun yaşandığı bir örnek yoktur. Bahsi geçen durumlar incelendiğinde UGO yönteminin non-lineer veri setindeki başarısı görülmüştür.

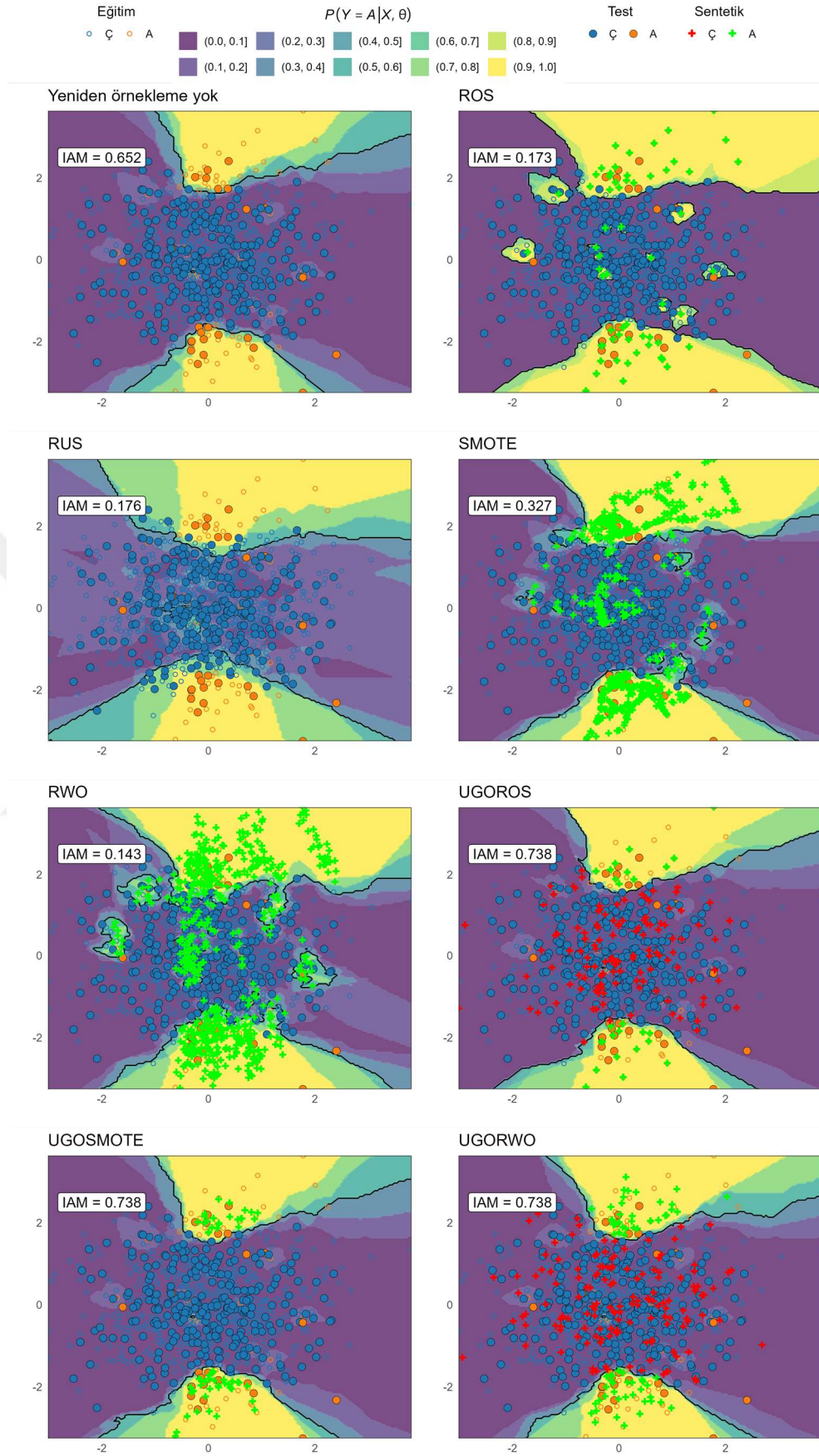


Şekil 6.4 Non-lineer ayrışan veri seti SVMRadial kontör grafikleri

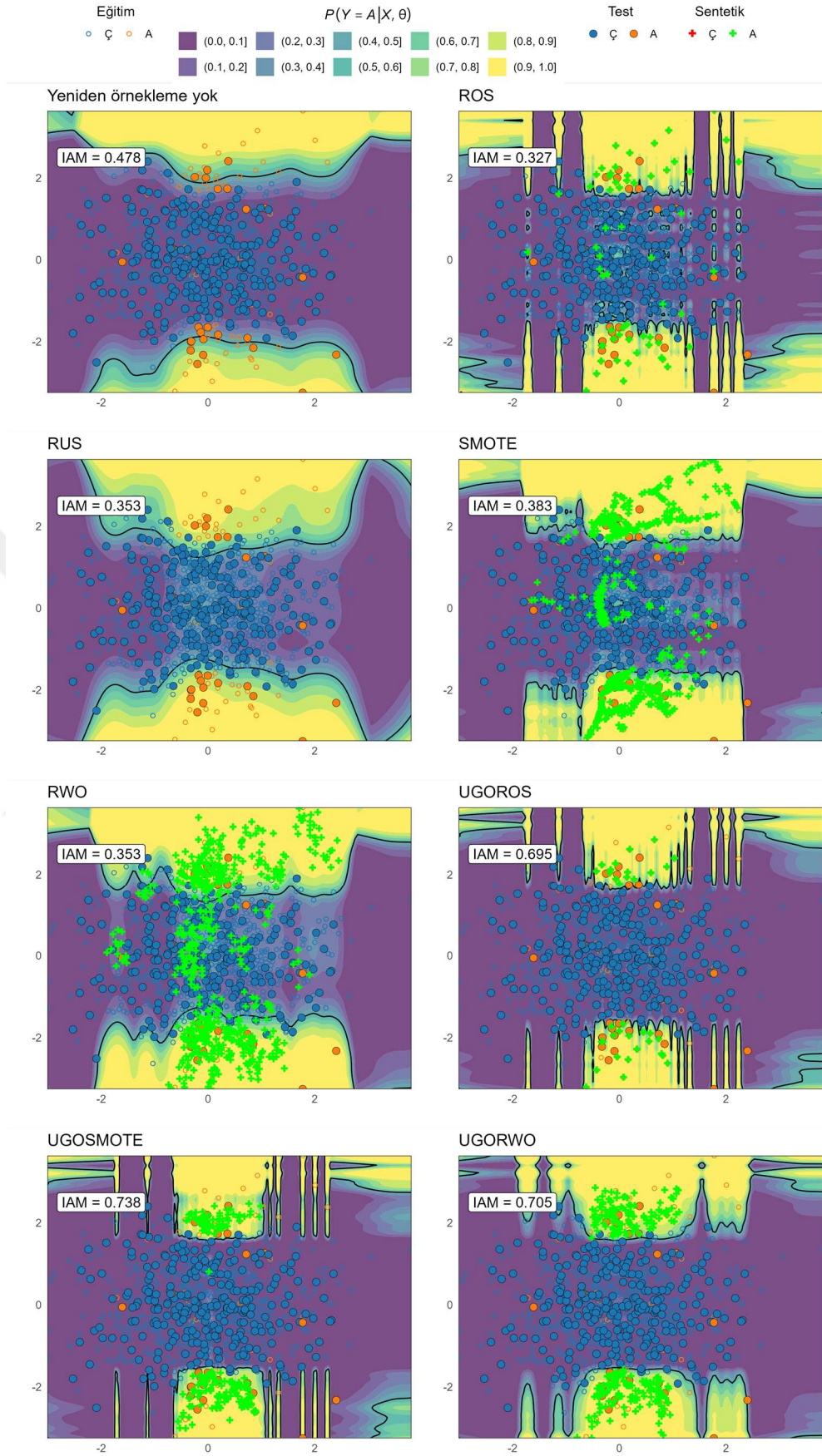


Şekil 6.6 Non-lineer ayrışan veri seti NNET kontör grafikleri





Şekil 6.8 Non-lineer ayrışan veri seti KNN kontör grafikleri.



Şekil 6.9 Non-lineer ayrıışan veri seti GKNB kontör grafikleri.

6.1.2. İki Sınıflı Veri Setleri için Sonuçlar

Çalışmada 74 adet iki sınıflı veri seti için IAM performans değerleri ve sıra ortalamaları hesaplanmıştır. IAM ortalamaları, veri setlerinde performansları görmek için, sıra ortalamaları ise yöntemler arasında karşılaştırma yapabilmek için kullanılmıştır. Tablo 5.4'te ifade edilen gruplamalara göre sonuçlar ayrılmıştır. Ek 1 ve 2 genel sınıflayıcılar için sırasıyla IAM ortalamaları ve sıra ortalamalarını göstermektedir. Koyu renkle işaretlenmiş değerler, bulunduğu veri setindeki en başarılı sonucu göstermektedir. Ek 1 incelendiğinde, UGOROS, UGOSMOTE ve UGORWO'nun ortalama IAM değerlerinin diğer yöntemlerden yüksek olduğu görülmektedir. Ek 2'deki genel ortalamalara bakıldığında, UGOROS, UGOSMOTE ve UGORWO'nun sıra ortalamalarının, diğer yöntemlere göre daha başarılı olduğu görülmüştür. En başarılı yöntem 2.997 ortalama ile UGOSMOTE iken en başarısız yöntem 7.604 ortalama ile RUS'tur. ROS yöntemi, IAM ortalamalarında 0 kez, sıra ortalamalarında ise 2 kez başarılı olmuştur. RUS yöntemi, her iki tabloda da başarılı olamamıştır. SMOTE yöntemi IAM ortalamalarında 1 kez, sıra ortalamalarında 2 kez başarılı olmuştur. RWO, IAM ortalamalarında 1, sıra ortalamalarında 0 kez başarılı olmuştur. UGOROS, IAM ortalamalarında 7 kez, sıra ortalamalarında 15 kez başarılı olmuştur. UGOSMOTE, IAM ortalamalarında 27 kez başarılı olmuş, sıra ortalamalarında ise 25 kez başarılı olmuştur. UGORWO, IAM ortalamalarında 24 kez başarılı olmuş, sıra ortalamalarında ise 19 kez başarılı olmuştur.

Lineer, non-lineer, sinir ağları, ağaç temelli, yoğunluk temelli, komşuluk temelli ve Bayesci sınıflama yöntemlerine ait IAM ortalamaları sırasıyla Ek 3, 5, 6.7, 9, 11, 13 ve 15'te, sıra ortalamaları ise sırasıyla Ek 4, 6, 8, 10, 12, 14 ve 16'da verilmiştir. Tüm gruplamalarda, UGO yöntemlerinin sıra ortalaması diğer yöntemlere göre daha düşüktür. IAM değerlerine göre en başarılı sonuçlar lineer grupta UGO-SMOE, non-lineer grupta UGORWO, sinir ağları grubunda UGOROS, ağaç yapılı grupta UGORWO, komşuluk temelli grupta UGOSMOTE ve Bayesci grupta UGOROS'a aittir. Yoğunluk temelli grupta, en başarılı IAM değerleri yeniden örnekleme için yapılmadığı duruma aittir. IAM sıra ortalamalarına göre ise en başarılı sonuçlar lineer grupta UGOSMOTE, non-lineer grupta UGORWO, sinir ağlarında UGOROS, ağaç yapılı grupta UGORWO, yoğunluk temelli grupta UGOSMOTE, komşuluk temelli grupta UGOSMOTE ve Bayesci grupta UGOROS'tur. En başarısız yöntem tüm durumlar için RUS yöntemidir.

Genel sınıflayıcılar ve diğer tüm alt gruplar için sıra ortalamaları Tablo 6.1’de, UGO’un başarısının istatistiksel olarak anlamlı olup olmadığını gösteren test sonuçları ise Tablo 6.2’de verilmiştir. Burada Friedman satırında, genel olarak yeniden örnekleme yöntemlerinin sütunda ismi geçen gruptaki sınıflayıcılar altında anlamlı fark gösterip göstermediğini test etmektedir. Tüm sınıflayıcı grupları için %5 anlamlılık seviyesinde yeniden örnekleme yöntemleri arasında anlamlı farklılık görülmüştür. İkili karşılaştırmalar ise diğer satırlardaki gibidir. Tablo 6.1 ve 6.2 incelendiğinde tüm bu sonuçların içerisinde önemli olan kısımlar şu şekilde özetlenebilir:

- Genel sınıflayıcılarda UGO ile NORES arasında anlamlı farklılık görülmüştür. UGO yöntemlerinin sıra ortalamaları anlamlı derecede NORES’ten düşüktür. UGOROS için lineer, yoğunluk temelli ve komşuluk temelli durumlarda NORES ile anlamlı farklılık görülmemiştir. Diğer durumlarda sıra ortalaması anlamlı derecede daha düşüktür. UGOSMOTE için sinir ağları ve yoğunluk temelli grupta anlamlı farklılık görülmemiştir. Diğer gruplarda sıra ortalaması anlamlı derecede daha düşüktür. UGORWO için ise lineer, sinir ağları ve yoğunluk temelli durumda NORES ile arasında anlamlı farklılık yoktur. Diğer durumlarda anlamlı derecede daha düşük sıra ortalamasına sahiptir.
- Genel sınıflayıcılarda UGOROS ile ROS arasında anlamlı farklılık vardır. Sıra ortalaması anlamlı derecede ROS’dan düşüktür. Aynı durum tüm alt gruplar için de geçerlidir.
- RUS yöntemi ile UGO yöntemleri arasında tüm sınıflayıcı gruplarında anlamlı farklılık görülmüştür. UGO sıra ortalamaları, tüm durumlarda RUS’tan anlamlı derecede düşüktür.
- SMOTE ile UGOSMOTE yöntemi arasında tüm gruplarda anlamlı farklılık görülmüştür. UGOSMOTE sıra ortalamaları anlamlı derecede SMOTE’dan düşüktür.
- RWO ile UGORWO yöntemleri arasında tüm sınıflayıcı gruplarında anlamlı farklılık görülmüştür. Tüm durumlarda UGORWO sıra ortalamaları RWO’dan anlamlı derecede daha düşüktür.

Tablo 6.1 İki sınıflı veri setlerinde tüm alt gruplar için IAM sıra ortalamaları

Grup	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
Genel	3.793	5.228	7.604	4.682	5.361	3.239	2.997	3.095
Lineer	3.806	5.619	7.441	4.721	5.910	2.838	2.574	3.090
Non-lineer	3.789	5.122	7.649	4.672	5.212	3.348	3.112	3.096
Sinir ağları	4.024	4.568	7.750	4.355	5.331	3.145	3.412	3.416
Ağaç temelli	4.476	4.561	7.826	4.524	4.965	3.356	3.231	3.061
Yoğunluk temelli	2.921	5.870	7.405	5.010	5.725	3.159	2.904	3.007
Komşuluk temelli	3.581	6.176	7.426	4.838	5.297	3.581	2.507	2.595
Bayesci	4.534	4.713	7.743	4.591	5.355	2.845	3.172	3.047

Tablo 6.2 İki sınıflı veri setlerinde IAM değerleri için Friedman ve ikili karşılaştırma test sonuçları

	Genel	Lineer	Non-linear	Sinir ağları	Ağaç	Yoğunluk temelli	Komşuluk temelli	Bayesci
Friedman	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
NORES-ROS	<0.001*	<0.001*	<0.001*	0.079	1.000	<0.001*	<0.001*	1.000
NORES-RUS	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
NORES-SMOTE	<0.001*	<0.001*	<0.001*	0.098	1.000	<0.001*	0.007*	1.000
NORES-RWO	<0.001*	<0.001*	<0.001*	<0.001*	1.000	<0.001*	<0.001*	1.000
NORES-UGOROS	<0.001*	0.125	<0.001*	0.027*	<0.001*	0.488	1.000	<0.001*
NORES-UGOSMOTE	<0.001*	0.034*	<0.001*	0.759	<0.001*	1.000	0.033*	<0.001*
NORES-UGORWO	<0.001*	0.335	<0.001*	0.893	<0.001*	1.000	0.033*	<0.001*
ROS-RUS	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
ROS-SMOTE	<0.001*	<0.001*	<0.001*	1.000	1.000	<0.001*	<0.001*	1.000
ROS-RWO	0.032*	<0.001*	1.000	0.001*	<0.001*	1.000	<0.001*	0.001*
ROS-UGOROS	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
ROS-UGOSMOTE	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
ROS-UGORWO	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RUS-SMOTE	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RUS-RWO	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RUS-UGOROS	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RUS-UGOSMOTE	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RUS-UGORWO	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
SMOTE-RWO	<0.001*	<0.001*	<0.001*	<0.001*	0.001*	<0.001*	0.240	0.002*
SMOTE-UGOROS	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
SMOTE-UGOSMOTE	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
SMOTE-UGORWO	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RWO-UGOROS	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RWO-UGOSMOTE	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
RWO-UGORWO	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*
UGOROS-UGOSMOTE	<0.001*	0.052	0.014*	1.000	1.000	0.237	<0.001*	0.640
UGOROS-UGORWO	0.165	0.538	0.008*	1.000	1.000	1.000	<0.001*	1.000
UGOSMOTE-UGORWO	1.000	<0.001*	1.000	1.000	1.000	1.000	1.000	1.000

*p<0.05

6.1.3. İki den Fazla Sınıflı Veri Setleri için Sonuçlar

Çalışmada 8 adet ikiden fazla sınıflı veri seti için IAM performans değerleri ve sıra ortalamaları hesaplanmıştır. Tablo 5.4'te görülebilecek gruplamalara göre sonuçlar ayrılmıştır. Ek 17 ve 18 genel sınıflayıcılar için sırasıyla IAM ortalamaları ve sıra ortalamalarını göstermektedir. Ek 17'deki genel ortalamalara bakıldığında, UGOROS, UGOSMOTE ve UGORWO'nun sıra ortalamalarının, diğer yöntemlere göre daha başarılı olduğu görülmüştür. En başarılı yöntem 3.313 ortalama ile UGOSMOTE iken en başarısız yöntem 6.152 ortalama ile RUS'tur. ROS, RUS, SMOTE ve RWO yöntemleri hiçbir veri setinde en başarılı yöntem olamamıştır. NORES, IAM ortalamalarında 2 kez, sıra ortalamalarında ise 1 kez başarılı olmuştur.

IAM için hem ortalama hem de sıra ortalamalarında tüm UGO yöntemleri, ROS, SMOTE ve RWO yöntemlerinden daha başarılıdır. UGOROS, IAM ortalama ve sıra ortalamalarında birer kez en başarılı yöntem olmuştur. UGOSMOTE, IAM ortalamalarında ve sıra ortalamalarında beşer kez en başarılı yöntem olmuştur. UGORWO ise IAM ortalamalarında hiç en başarılı yöntem olmazken, sıra ortalamalarında bir kez en başarılı yöntem olmuştur.

Lineer, non-lineer, sinir ağları, ağaç temelli, yoğunluk temelli, komşuluk temelli ve Bayesci sınıflama yöntemlerine ait IAM ortalamaları sırasıyla Ek 17, 19, 21, 23, 25, 27 ve 29'da, sıra ortalamaları ise sırasıyla Ek 18, 20, 22, 24, 26, 28 ve 30'da verilmiştir. IAM değerlerine göre en başarılı sonuçlar lineer grupta UGOSMOTE ve UGORWO, non-lineer grupta UGOSMOTE, sinir ağları grubunda UGORWO, ağaç yapılı grupta UGOSMOTE, yoğunluk temelli grupta UGOSMOTE, komşuluk temelli grupta UGOROS ve Bayesci grupta UGOSMOTE'a aittir. IAM sıra ortalamalarına göre ise en başarılı sonuçlar lineer grupta UGOSMOTE, non-lineer grupta UGOSMOTE, sinir ağlarında UGORWO, ağaç yapılı grupta UGOSMOTE, yoğunluk temelli grupta UGOROS, komşuluk temelli grupta UGOROS ve Bayesci grupta UGOSMOTE'tur. En başarısız yöntem tüm durumlar için RUS yöntemidir.

Genel sınıflayıcılar ve diğer tüm alt gruplar için ikiden fazla sınıflı veri setleri altında sıra ortalamaları Tablo 6.3'te, UGO'un başarısının istatistiksel olarak anlamlı olup olmadığını gösteren test sonuçları Tablo 6.4'te verilmiştir. Sinir ağları ve Bayesci gruplar dışındaki tüm sınıflayıcı grupları için %5 anlamlılık seviyesinde yeniden örnekleme yöntemleri arasında anlamlı farklılık görülmüştür. Tüm bu sonuçların içerisinde önemli olan kısımlar şu şekilde özetlenebilir:

- Genel sınıflayıcılarda sadece UGOSMOTE ile NORES arasında anlamlı farklılık görülmüştür. UGOSMOTE sıra ortalaması, NORES sıra ortalamasından genel grupta anlamlı derecede düşüktür. Yoğunluk temelli grupta NORES ile tüm UGO yöntemleri arasında anlamlı fark görülmüştür. Yoğunluk temelli gruplarda UGO yöntemlerinin sıra ortalamaları, NORES'ten anlamlı derecede düşük bulunmuştur. Diğer durumlarda NORES ile anlamlı farklılık yoktur.
- ROS ile UGOROS arasında genel ve yoğunluk temelli grupta anlamlı farklılık görülmüştür. Bu gruplarda UGOROS sıra ortalaması, ROS'tan anlamlı derecede daha düşüktür.

Tablo 6.3 İkiiden fazla sınıflı veri setlerinde tüm alt gruplar için IAM sıra ortalamaları

Grup	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
Genel	4.714	4.942	6.152	4.647	4.665	3.781	3.313	3.786
Lineer	5.500	5.042	5.917	4.292	4.625	3.708	3.250	3.667
Non-lineer	4.500	4.915	6.216	4.744	4.676	3.801	3.330	3.818
Sinir ağları	4.750	3.531	5.406	4.594	4.938	4.688	4.563	3.531
Yoğunluk temelli	4.500	4.109	6.688	4.531	4.359	4.547	3.125	4.141
Ağaç temelli	4.844	6.063	5.953	5.359	4.766	2.766	2.891	3.359
Komşuluk temelli	4.375	5.500	7.875	4.125	4.250	2.875	3.000	4.000
Bayesci	5.250	3.813	5.781	4.688	4.313	4.563	3.719	3.875

Tablo 6.4 İkiiden fazla sınıflı veri setlerinde IAM değerleri için Friedman ve ikili karşılaştırma test sonuçları

Testler	Genel	Lineer	Non-lineer	Sinir ağları	Ağaç	Yoğunluk temelli	Komşuluk temelli	Bayesci
Friedman	<0.001*	0.001*	<0.001*	0.331	<0.001*	<0.001*	0.001*	0.186
NORES-ROS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NORES-RUS	<0.001*	1.000	<0.001*	1.000	0.001*	0.510	0.219	1.000
NORES-SMOTE	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NORES-RWO	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
NORES-UGOROS	0.053	0.498	1.000	1.000	1.000	0.022*	1.000	1.000
NORES-UGOSMOTE	<0.001*	0.051	0.089	1.000	1.000	0.008*	1.000	1.000
NORES-UGORWO	0.060	0.169	1.000	1.000	1.000	0.096	1.000	1.000
ROS-RUS	<0.001*	1.000	<0.001*	0.699	0.001*	1.000	1.000	0.308
ROS-SMOTE	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ROS-RWO	1.000	1.000	1.000	1.000	1.000	0.104	1.000	1.000
ROS-UGOROS	0.012*	1.000	0.524	1.000	1.000	<0.001*	1.000	1.000
ROS-UGOSMOTE	<0.001*	0.353	0.035*	1.000	1.000	<0.001*	1.000	1.000
ROS-UGORWO	0.058	1.000	1.000	1.000	1.000	0.008*	1.000	1.000
RUS-SMOTE	<0.001*	1.000	<0.001*	1.000	<0.001*	1.000	0.219	0.812
RUS-RWO	<0.001*	0.769	0.003*	1.000	0.101	0.459	0.219	0.939
RUS-UGOROS	<0.001*	0.222	<0.001*	1.000	0.002*	<0.001*	0.219	1.000
RUS-UGOSMOTE	<0.001*	0.016*	<0.001*	1.000	<0.001*	<0.001*	0.219	1.000
RUS-UGORWO	<0.001*	0.088	<0.001*	1.000	0.002*	0.006*	0.219	1.000
SMOTE-RWO	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SMOTE-UGOROS	0.046*	1.000	0.925	1.000	1.000	0.004*	1.000	1.000
SMOTE-UGOSMOTE	0.001*	0.754	0.085	1.000	1.000	0.001*	1.000	1.000
SMOTE-UGORWO	0.205	1.000	1.000	1.000	1.000	0.046*	1.000	1.000
RWO-UGOROS	0.010*	1.000	0.307	1.000	1.000	0.023*	1.000	1.000
RWO-UGOSMOTE	<0.001*	0.165	0.004*	1.000	0.596	0.006*	1.000	1.000
RWO-UGORWO	0.002*	1.000	0.221	1.000	1.000	0.023*	1.000	1.000
UGOROS-UGOSMOTE	1.000	1.000	1.000	1.000	0.533	1.000	1.000	0.600
UGOROS-UGORWO	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
UGOSMOTE-UGORWO	0.426	1.000	0.627	1.000	0.956	1.000	1.000	1.000

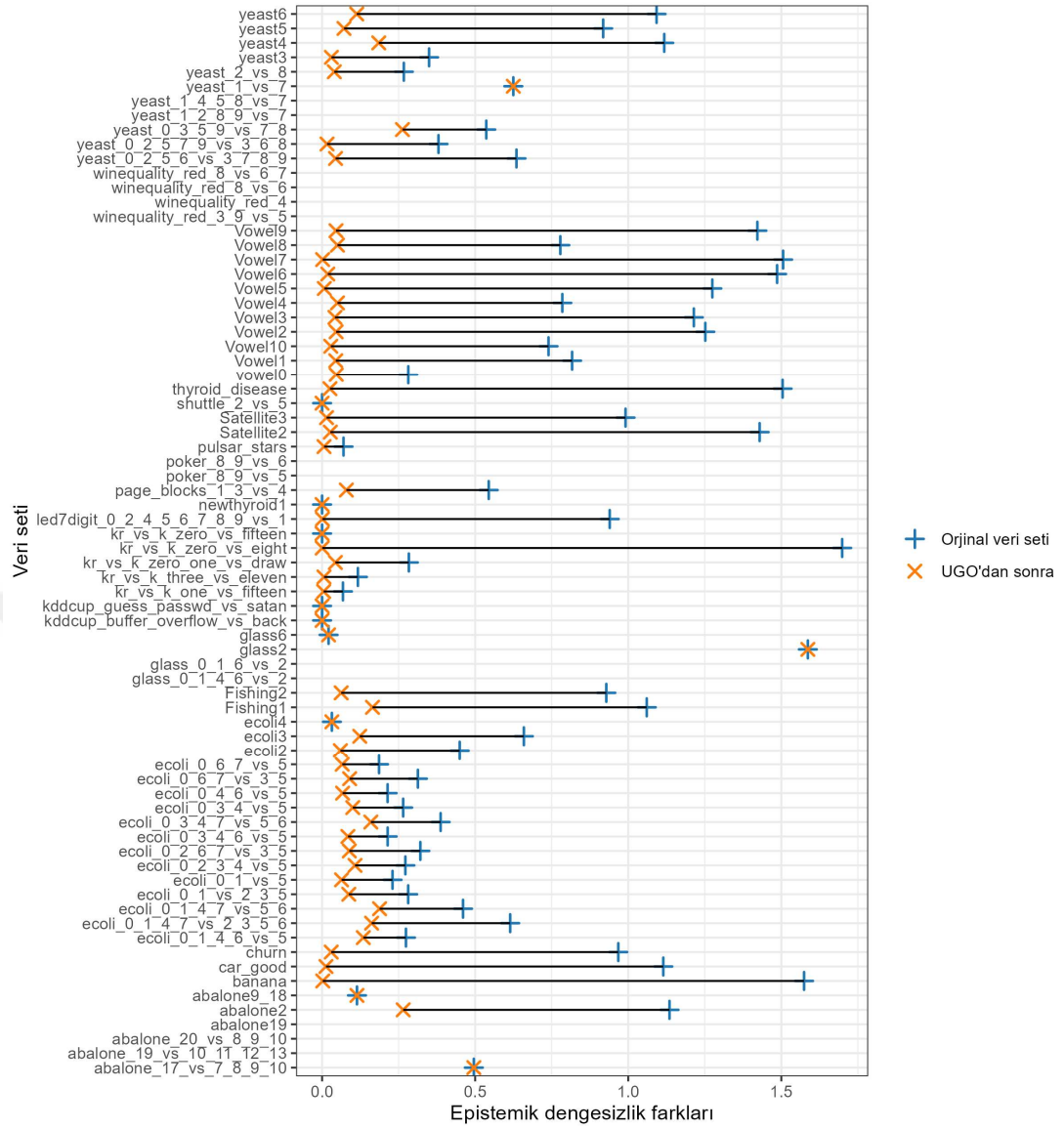
- RUS ve UGOROS yöntemleri arasında genel, non-lineer, ağaç yapılı ve yoğunluk temelli sınıflayıcı gruplarında anlamlı farklılık görülmüştür. Bu gruplarda, UGOROS yönteminin sıra ortalaması, RUS'dan anlamlı derece düşüktür. RUS ve UGOSMOTE arasında genel, lineer, non-lineer, ağaç temelli ve yoğunluk temelli sınıflayıcı gruplarında anlamlı farklılık vardır. UGOSMOTE, RUS'a kıyasla anlamlı derecede daha düşük sıra ortalamasına sahiptir. RUS ve UGORWO arasında ise genel, non-lineer, ağaç yapılı ve yoğunluk temelli gruplarda anlamlı farklılık görülmüştür ve UGORWO anlamlı derecede daha düşük sıra ortalamasına sahiptir.
- SMOTE ve UGOSMOTE arasında genel ve yoğunluk temelli gruplarda

anlamli farklılık görülmüştür. UGOSMOTE, SMOTE'a kıyasla anlamli derecede daha düşük sıra ortalamasına sahiptir.

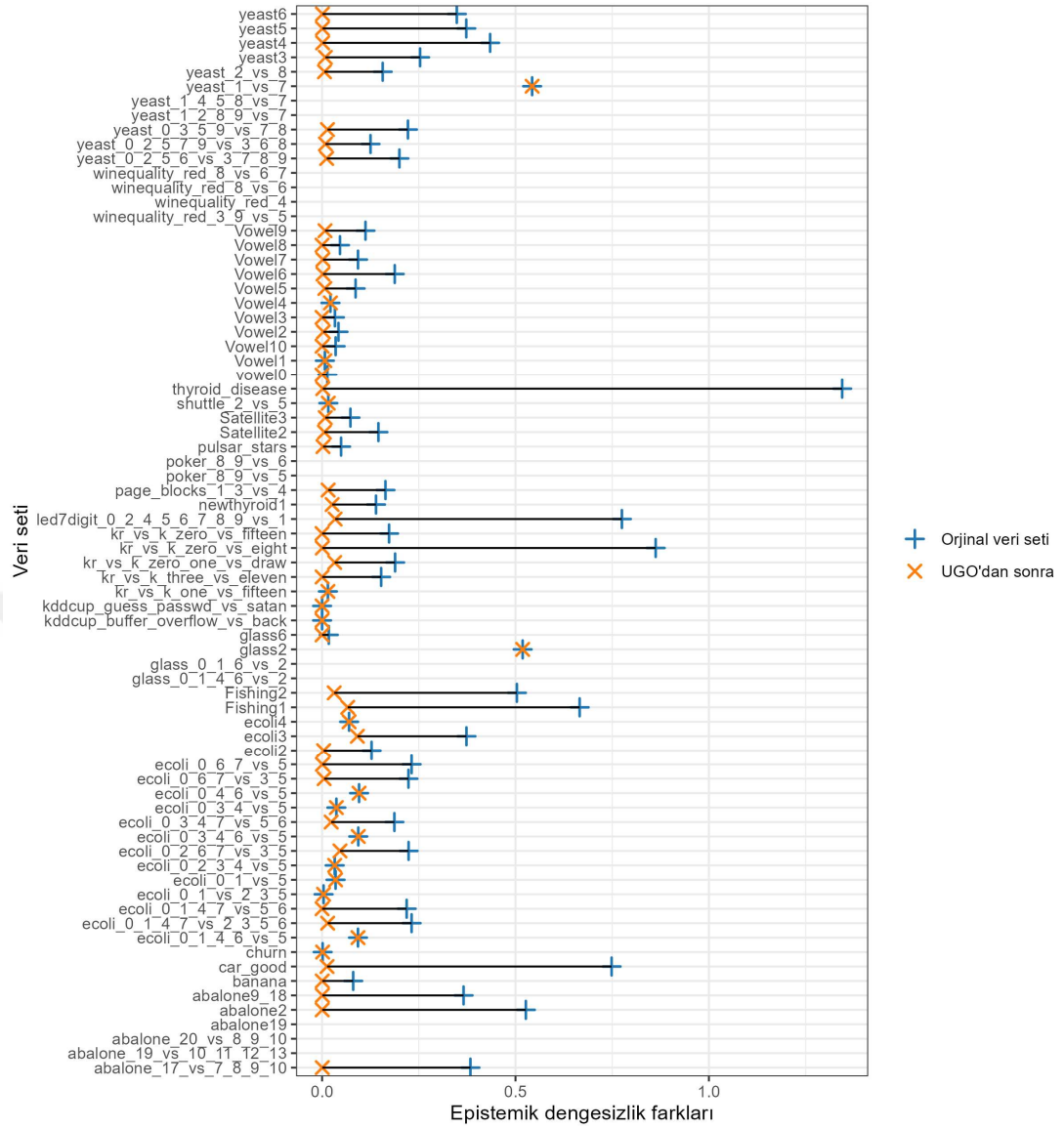
- RWO ve UGORWO arasında genel ve yoğunluk temelli gruplarda anlamli farklılık görülmüştür. UGORWO, RWO'ya kıyasla anlamli derecede daha düşük sıra ortalamasına sahiptir.

6.1.4. Veri Setlerinde UGO'nun Belirsizlik Üzerine Etkisi

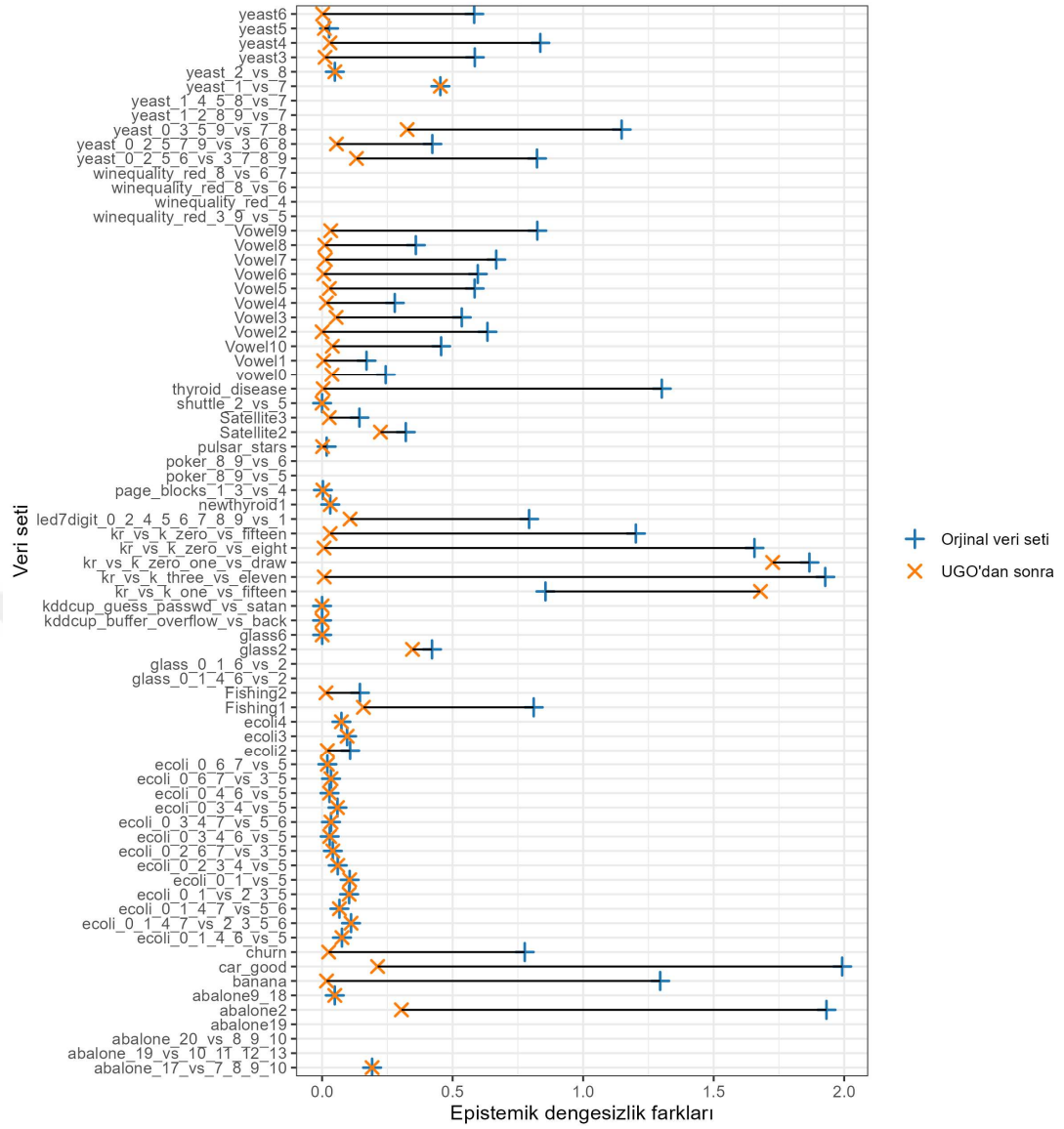
Veri setlerinde, MMA belirsizliklerindeki değişimi incelemek için, tüm veri setlerinde seçilen LogReg, KNN, GKNB, NNET ve RF sınıflayıcıları ile belirsizlik değişimleri hesaplanmıştır. İki sınıflı durumda epistemik dengesizlik, belirsizlik ortalamaları arasındaki fark olan $|\bar{U}_{MMA}^{(1)} - \bar{U}_{MMA}^{(2)}|$ ile hesaplanmıştır. Çok sınıflı durumda ise en düşük belirsizlik ortalamalı sınıfın belirsizlik ortalaması ile en yüksek belirsizlik ortalamalı sınıfın belirsizlik ortalaması arasındaki farka, yani $\max_l \bar{U}_{MMA}^{(l)} - \min_l \bar{U}_{MMA}^{(l)}$, bakılarak epistemik dengesizlik ölçülmüştür. Ölçümler sonrası, iki sınıflı veri setlerinde UGO uygulanmadan önce ve sonra epistemik dengesizlikler arasındaki değişimler Şekil 6.10, 6.11, 6.12 ve 6.13'te verilmiştir. Çok sınıflı veri setlerinde ise Şekil 6.14, 6.15, 6.16 ve 6.17'de epistemik dengesizliklerdeki değişim görülmektedir. Burada, turuncu "x" işareti UGOSMOTE uyguladıktan sonra, mavi renkli "+" işareti ise UGOSMOTE algoritmasının başlangıcında ölçülen belirsizliği göstermektedir. Bazı veri setlerinde (örneğin winequality_red_4) hiçbir epistemik dengesizlik bilgisi yoktur. Bunun nedeni, bu veri setlerindeki tüm örneklerin gürültü olarak belirlenmesidir. Belirsizlik ortalamaları, gürültü olmayan örnekler üzerinden yapıldığı için bu veri setlerinde UGO algoritması belirsizlik örnekleme aşamasına geçmemiştir. Diğer veri setlerine bakıldığında ise, hem iki sınıflı hem de çok sınıflı durumlarda epistemik dengesizliğin ya değişmediği, ya da azaldığı görülmektedir. Epistemik dengesizliğin azalması demek, sınıflara ait dengesizlik ortalamalarının birbirine yaklaşması anlamına gelmektedir.



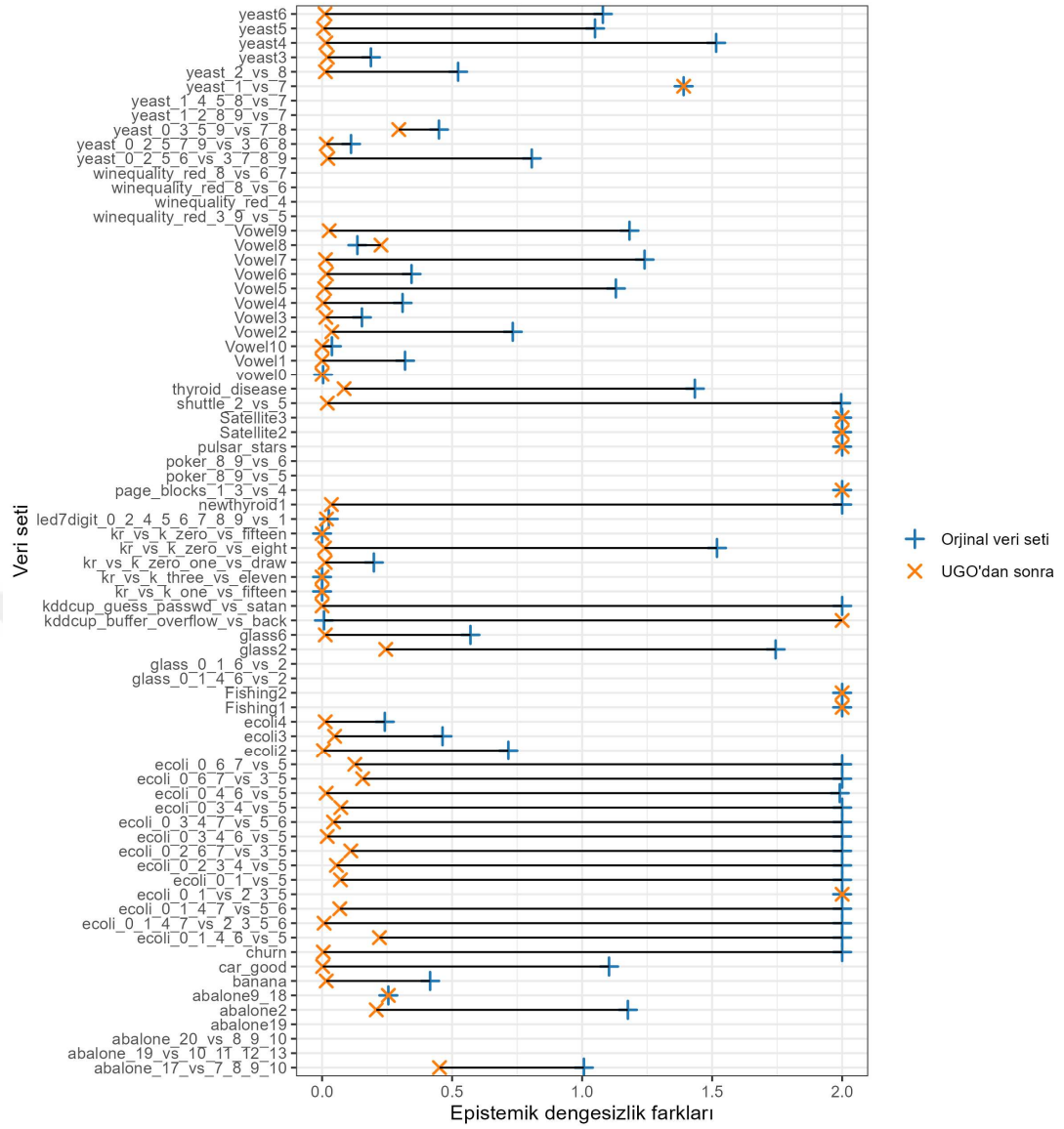
Şekil 6.10 LogReg için iki sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



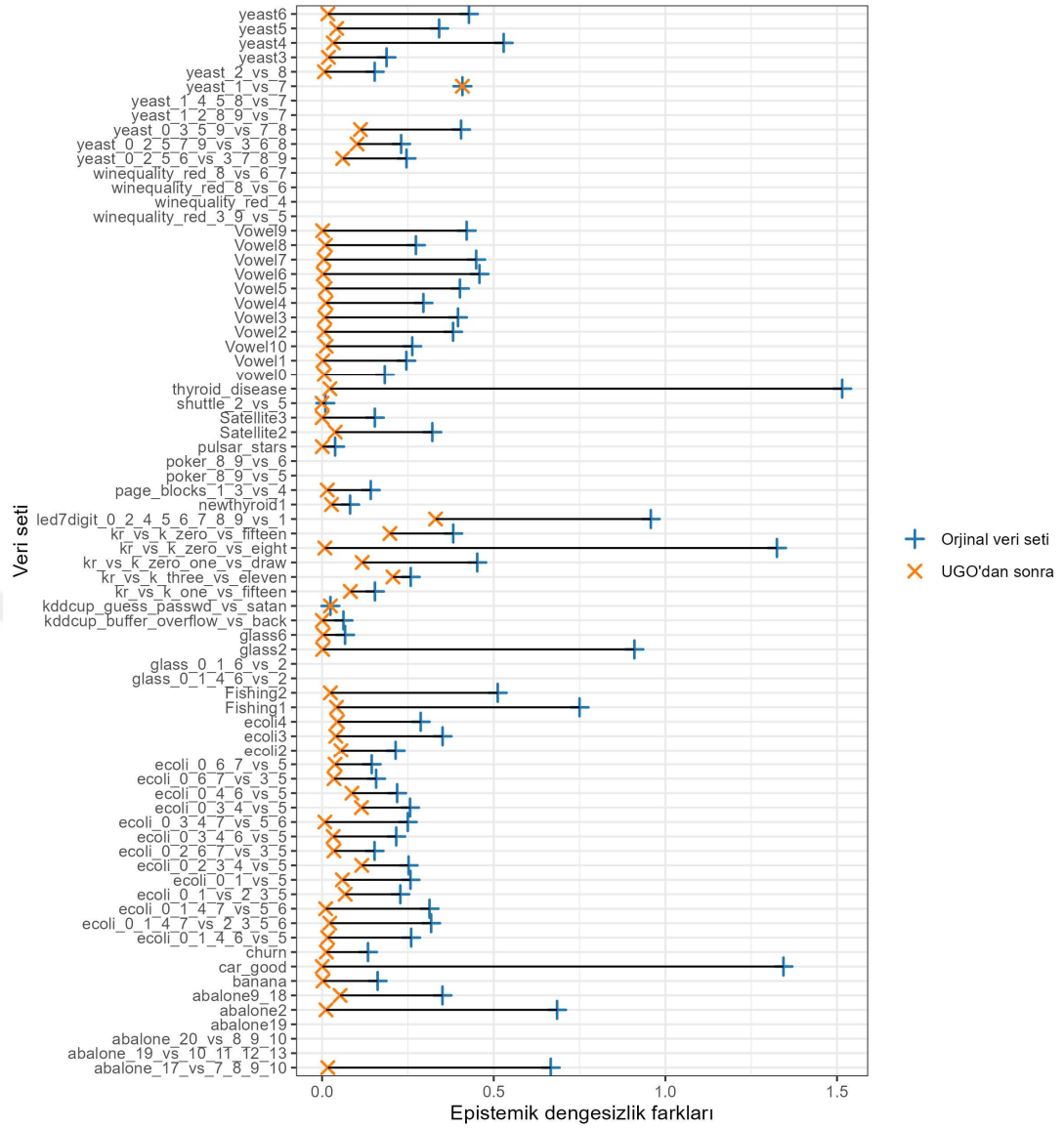
Şekil 6.11 KNN için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



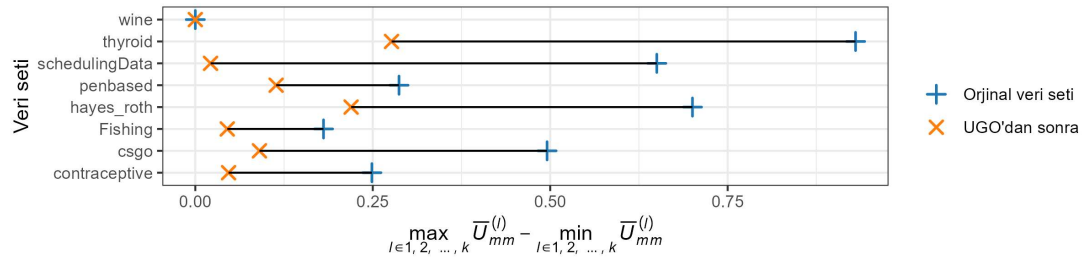
Şekil 6.12 GKNB için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



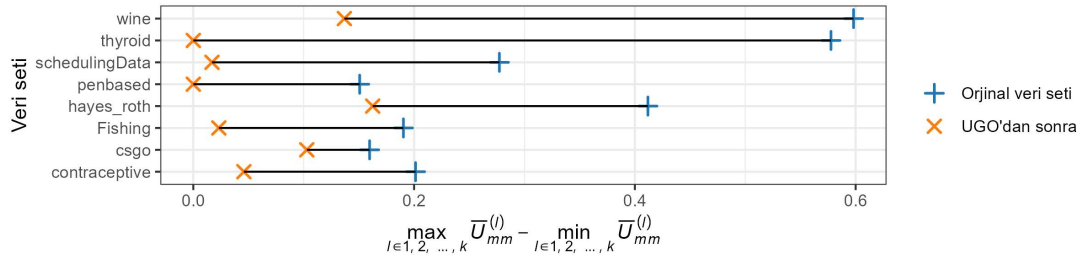
Şekil 6.13 NNET için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



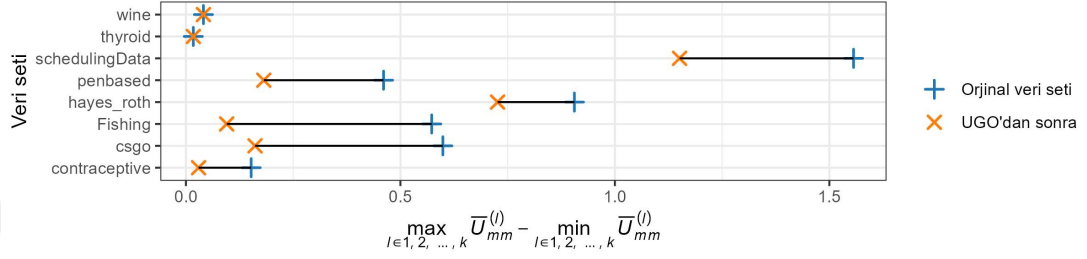
Şekil 6.14 RF için veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



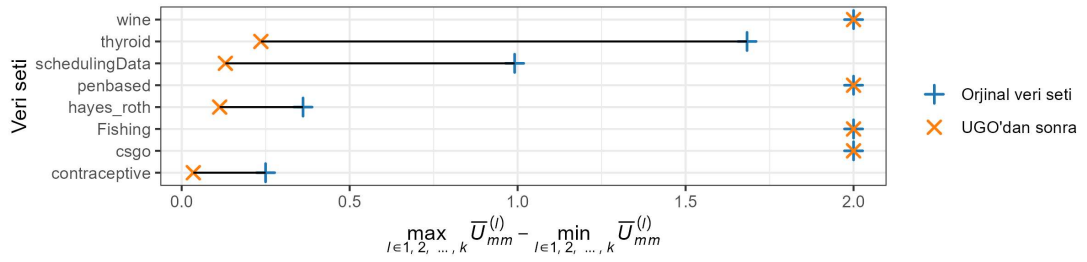
Şekil 6.15 LogReg için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



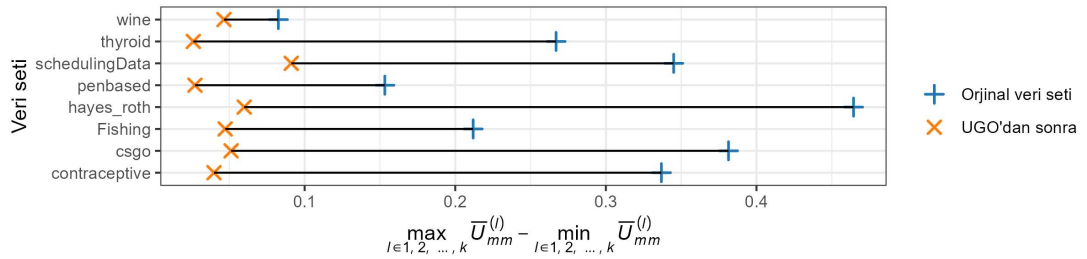
Şekil 6.16 KNN için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



Şekil 6.17 GKNB için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



Şekil 6.18 NNET için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar



Şekil 6.19 RF için ikiden fazla sınıflı veri setlerinde UGO uygulanmadan önce ve sonra sınıflar arası belirsizlikler arasındaki farklar

7. TARTIŞMA

Çalışmada önerilen UGO yöntemi, sınıflamada dengesizlik probleminde kullanılan aşırı örnekleme yöntemindeki bazı sorunlara çözümler getirmeyi amaçlamaktadır. UGO, modifiye edilmiş belirsizlik öğrenmesi kullanarak aşırı örnekleme ile veri üretilecek olan örnekleri gürültü olmadığından emin olarak yüksek belirsizlik olanlar arasından seçer ve iteratif biçimde veri üretir. Üretilen verilerin gürültü olup olmadığını tekrar kontrol eder ve gürültü ise en yakın gürültü olmayan kendi sınıfından örneğin olduğu yere taşır. Bu iterasyon, sınıflara ait ortalama U_{MMA} değerleri anlamlı farklılık göstermeyene kadar devam eder. Böylece gürültüsüz biçimde aşırı örnekleme yapılır ve dengesizliğin yarattığı problemi ortadan kaldıracak kadar miktarda veri üretilir. UGO'un bu amaçları yerine getirilip getirilmediğini görmek için simülasyon ve gerçek veri setlerinde uygulamalar yapılmıştır.

7.1. Deney Sonuçlarının Değerlendirilmesi

Performanslar çapraz geçerlilik yoluyla hesaplanmıştır. Performans ölçütü olarak, IAM tercih edilmiştir. Çoğu sınıf dengesizliği çalışmasında GMEAN, F1, AUC ve MCC gibi ölçütler kullanılmaktadır. Fakat çalışmalar, bu ölçütlerin de yanlı olduğunu göstermiştir. Bu nedenle daha güncel ve sınıf dengesizliği durumu için önerilmiş basit yorumlanabilir bir ölçüt olan IAM ölçütü kullanılmıştır.

Gürültüye olan dayanıklılığını incelemek için gürültü miktarını artırarak simülasyonlar yapılmış ve UGORWO, UGOSMOTE yöntemleri RWO ve SMOTE yöntemleri ile karşılaştırılmıştır. Karşılaştırmalar, LogReg, CART, GNB ve NNET sınıflayıcıları kullanılarak yapılmıştır. Tüm yöntemlerde gürültü miktarı arttıkça doğal olarak performans kaybı yaşanmaktadır. Bunun nedeni, kesişim bölgesinin artmasıdır. Kesişim bölgesi, aleatorik belirsizlik demektir ve veri üretmek bu problemin çözülmesi mümkün değildir. Aleatorik belirsizliği çözmenin yollarından en bilineni, veri setine yeni bilgilendirici değişkenler dahil etmektir. Sınıflayıcıların performanslarındaki düşüş incelendiğinde, UGO yöntemlerinin çok daha yavaş performans kaybına uğradığı ve bir noktadan sonra performans kaybının hiç olmadığı görülmektedir. Bunun nedeni, gürültü oranı çok fazla arttıktan sonra, UGO tüm örneklerin gürültü olduğunu düşünecek ve hiç veri üretmeyecektir. Orijinal veri setindeki performansla aynı sonucu verecektir. SMOTE ve RWO yöntemlerine bakıldığında IAM değeri 0'ın altına inerek git gide düşmektedir. Bu, rastgele sınıflayıcıdan bile daha kötü sonuçlara yol açmaktadır. Bir noktadan sonra, veri

üretilmemesi daha uygundur ve UGO bunu tespit edebilmektedir.

Lineer ve non-lineer sınıflayıcıların performanslarını incelemek için gürültü örnekleri içeren iki farklı simülasyon veri seti oluşturulmuştur. Bunlardan birincisi lineer ayrıştırılabilen bir veri seti, diğeri ise lineer ayrışmasa da non-lineer yollarla ayrıştırılabilen bir veri setidir. Böylece, lineer ve non-lineer sınıflayıcıları kendi başarılı oldukları ortamda test edilebilmiştir. Lineer ayrışan veri setinde lineer sınıflayıcılardan LogReg, SVMLineer ve LDA kullanılmıştır. Sonuçlar incelendiğinde, UGO yöntemleri, ROS, RUS, SMOTE ve RWO'ya kıyasla oldukça yüksek başarı göstermişlerdir. Non-lineer ayrışan veri setinde ise SVMRadyal, BART, NNET, BNNET, KNN ve GKNB sınıflayıcıları non-lineer sınıflayıcı olarak kullanılmıştır. Performanslar incelendiğinde, UGO yöntemleri her zaman RUS, ROS, SMOTE ve RWO'dan daha başarılı sonuçlar göstermiştir. Üretilen verilerin görselleri incelenmiş ve gürültü verileri üretilmediği de görülmüştür. Non-lineer ayrıştırılabilen veri setinde, RUS, ROS, SMOTE ve RWO yöntemlerinin, yeniden örnekleme yapılmadığı durumlardan daha düşük performans sergilemeleri dikkat çekicidir.

Simülasyon olmayan, literatürde sıklıkla kullanılan farklı özelliklere sahip 74'ü iki sınıflı, 8'i ikiden fazla sınıflı olmak üzere 82 veri seti kullanılarak sonuçlar toplanmıştır. Farklı davranış sergileyen sınıflayıcılar altındaki sonuçları daha iyi görebilmek için sınıflayıcılar genel, lineer, non-lineer, sinir ağları, ağaç temelli, yoğunluk temelli, komşuluk temelli ve Bayesci olacak şekilde gruplara ayrılmıştır. İki sınıflı veri setlerinde tüm alt gruplarda başarılı sonuca ulaşılmıştır. Burada yöntemler arasında karşılaştırma yaparken, dikkat edilmesi gereken şey UGO'un kendine özgü veri üretme yönteminin olmamasıdır. Bu nedenle, UGORWO RWO ile, UGOROS ROS ile ve UGOSMOTE SMOTE ile karşılaştırılmalıdır. İki sınıflı veri setlerinde bu sonuca her durumda ulaşılsa da ikiden fazla sınıflı veri setlerine başarı seviyesi daha düşüktür. UGO yöntemleri, iki sınıflı veri setlerinde tüm gruplarda istatistiksel olarak anlamlı olacak şekilde daha düşük sıra ortalamasına sahiptir. Bu da amacına ulaştığı ve diğer yöntemlerden daha iyi performans gösterdiği anlamına gelir. İkiden fazla sınıflı veri setlerinde, tüm sınıflayıcı gruplarında daha düşük sıra ortalamasına sahip olmasına rağmen bu başarı yeterli değildir. İstatistiksel testler sonucunda, bu başarısının anlamlı denilebilecek bir seviyede olmadığı görülmüştür. Genel sınıflayıcılar grubunda, UGO yöntemleri istatistiksel olarak anlamlı başarı göstermiştir. Fakat özellikle sinir ağları, komşuluk temelli ve Bayesci sınıflar grubunda olmak üzere birçok alt grup durumunda anlamlı başarı gösterememiştir.

Bunun olası nedenlerinden en önemlisi, UGO'un kullandığı gürültü tespit yöntemidir. Gürültü tespiti, ikiden fazla sınıflı durumlarda daha karmaşık bir yapıya dönüşmektedir. Kullanılan gürültü tespiti yöntemi, çok sınıflı veri setlerinde çok fazla gürültü örneği tespit etmektedir. Ayrıca, sadece fazla gürültü tespit etmesinin yanı sıra yanlış örnekleri de gürültü olarak tespit ediyor olabilir. Bu konuya Bölüm 7.2'de tekrar değinilecektir.

Çalışmada ortaya çıkan dikkat çekici sonuçlardan birisi, UGO dışındaki yeniden örnekleme yöntemlerinin beklenenden fazla başarısız olmasıdır. Tarawneh ve arkadaşlarının (2022) çalışmasında, 70'ten fazla aşırı örnekleme yöntemi farklı bir değerlendirme mekanizması altında değerlendirilmiş ve başarısız oldukları görülmüştür. Bu demek oluyor ki değerlendirme mekanizmasına göre yöntemlerin başarısı oldukça değişmektedir. Benzer biçimde literatürdeki en başta AUC olmak üzere F1 ve MCC gibi popüler değerlendirme kriterlerinden farklı olarak bu çalışmada IAM kullanılmıştır. Bizim IAM kullanmamızın nedeni bahsi geçen ölçütlere ait literatürdeki çalışmalar idi. Çalışmalar incelendiğinde sınıf dengesizliğinde ACC'nin kullanılmaması gerektiği konusunda literatür her ne kadar hem fikir olsa da hangi ölçütün kullanılması gerektiği konusunda hem fikir değil. AUC'un, sınıflardan bağımsız olduğundan dolayı sınıf dengesizliğinden etkilenmediğini gösteren çalışmalar (Jeni vd., 2013; Wardhani vd., 2019; Halimu vd., 2019) olduğu gibi AUC hesaplamasında integral hesabına yakınsamada sıra istatistikleri kullanılırken olasılık bilgisi kaybedildiği için dengesizliğe dayanıklı olsa da genel olarak kötü bir ölçüt olduğunu/olabileceğini söyleyen çalışmalar (Lobo vd., 2008; Jeni vd., 2013) da mevcuttur. Benzer şekilde MCC'nin yansız bir ölçüt olduğunu ve sınıf dengesizliği durumunda F1 gibi ölçütlere kıyasla başarılı olduğunu söyleyen çalışmalar olduğu gibi (Chicco vd., 2021; Halimu vd., 2019) AUC'a kıyasla daha kötü olduğunu (Halimu vd., 2019) ve hatta genel olarak çoğunluk sınıftan yana yanlı olduğunu (Zhu, 2020) söyleyen çalışmalar mevcuttur. Giriş kısmında bahsi geçen yeniden örnekleme yöntemlerinin neredeyse hepsi, en sıklıkla AUC olmak üzere birçok problemli olduğu düşünülen ölçütlere göre gösterdikleri başarı nedeniyle tanınmış yöntemlerdir. Eğer makalelerde önerilen yöntemlerde kullanılan popüler ölçütler bir şekilde sorunlu ise ve kötü bir senaryo düşünülecek olunursa geçmişte önerilmiş yeniden örnekleme yöntemlerinin aslında başarısız olduğu söylenebilir mi? Çünkü Tarawneh ve arkadaşlarının (2022) çalışmasındaki gibi sınıf dengesizliği durumunda kullanılmak üzere önerilmiş farklı bir ölçüt olan IAM kullanıldığında, klasik yeniden örnekleme

yöntemlerinin (özellikle RUS), birçok durumda hiç kullanılmamasalar daha iyi olduğu görülmüştür.

7.2. Yöntemin Sınırlamaları

UGO, örnekleme yaparken her bir iterasyonda tekrar model eğitir. Spesifik bir sınıflayıcı ile belirsizlik yapılsaydı, bunu önlemek için farklı hesaplama kolaylıkları sağlanabilirdi. Fakat, belirsizlik miktarlarının sınıflayıcıdan sınıflayıcıya olan farklılıklarını ele alabilmek için bu şekilde tercih edilmiştir. Bu da yanında yüklü bir işlem maliyetini getirmektedir.

Her sınıflayıcı için o sınıflayıcı üzerinden yapılacak olan belirsizlik tahmininin kalitesi, sınıflayıcının kalitesine bağlıdır. Örneğin aşırı uyum problemine sahip bir sınıflayıcı veya olasılık hesabı yapamayan bir sınıflayıcı, U_{MMA} hesabının kaliteini düşürecektir. Bu da hatalı örneklemeye neden olur. Bu nedenle belirsizlik tahmininde sorun çıkarabilecek sınıflayıcılarda, örnekleme için benzer başka bir sınıflayıcı kullanılması önerilmektedir. Örneğin, C4.5 karar ağacı ile eğitim yapılacaksa C5.0 ile UGO uygulanarak aşırı örnekleme yapıp, ardından C4.5 tekrar eğitim için kullanılabilir. Benzer biçimde gürültü örnekleri, sınıflayıcıdan sınıflayıcıya farklılık gösterebilir. Gürültü tespiti, aleatorik belirsizliği yüksek olan örnekleri tespit etmek için yapılmaktadır. Aleatorik belirsizlik ise yöntemin kapasitesi ile gerçeklik arasındaki farkı açıklar. Yöntemin kapasitesi de sınıflayıcıdan sınıflayıcıya değişecektir. Her sınıflayıcı için aynı gürültü tespiti yaklaşımının kullanılması, aleatorik belirsizliği düşük örneklerin seçilmesine ve yüksek olanların seçilmemesine neden olabilir.

7.3. Gelecekteki Çalışmalar için Öneriler

UGO'un çok sınıflı durumda yeterince başarı göstermemesi, gürültü tespitin çok sınıflı durumlarda başarısız olmasından kaynaklanabilir. Çalışmada kullanılan gürültü tespit yöntemi, klasik bir yaklaşımdır. Bunun yerine daha güncel yöntemler ile sonuçların nasıl değiştiği incelenebilir. Ayrıca, aleatorik belirsizliği daha iyi yakalayabilmek için sınıflayıcıya özgü bir gürültü tespiti üzerinde çalışılabilir. Bunu yapabilmek için topluluk algoritmalarından faydalanılabilir. Örneğin, torbalama ile birçok modelin kullandıkları oy uyumsuzluğuna bakılarak gürültü tespiti yapılabilir. Daha farklı olarak boosting ile örnek ağırlıkları incelenerek, yüksek ağırlıklı örnekleri gürültü olarak kabul eden bir gürültü tespiti yapılabilir. Bu topluluklar eğitim yapılacak olan sınıflayıcılardan oluşuyorsa, sınıflayıcı temelli gürültü tespiti yapılmış olacaktır. Fakat bu durum, zaten işlem maliyeti yüksek olan UGO'a daha da fazla

işlem yükü getirecektir. Buna dikkat edilmelidir.

UGO yönteminin sınıf dengesizliğinin varlığını incelerken epistemik belirsizlik üzerine yoğunlaşması, farklı konulara dikkat çekmektedir. Sınıf dengesizliği, örnek sayısına tamamen bağlı değilse; görünürde örnek sayıları eşit ve dengeli bir veri setinde bile epistemik belirsizlik açısından aslında sınıf dengesizliği olabilir mi? Sınıf dengesizliğini çözmede kullanılan yöntemlerin, uygulanmasına gerek olmayan dengesiz veri setleri olabileceği gibi, uygulanması gereken dengeli veri setleri de olabilir. Bunun için epistemik belirsizliğin sınıf dengesiyle olan ilişkisini daha ayrıntılı inceleyen daha teorik bir araştırma yapmak gerekir.

Literatürdeki çalışmalarda, AUC, F1 ve MCC gibi popüler ölçütlerde başarılı sonuç veren yöntemlerin, IAM ölçütü kullanıldığında başarısız olması, tartışmaya açıktır. Ölçütler konusundaki literatürdeki fikir ayrılığını ortadan kaldırmak için sadece eski ölçütleri değil, daha yeni ölçütleri de hesaba katan çalışmalar yapılmalıdır. Bu ölçütlerin avantaj ve dezavantajları uygulamalı ve teorik biçimde açıkça ortaya konmalı ve sınıf dengesizliği durumunda uygulama yapan araştırmacıların hangi ölçüte göre model seçeceklerine yardımcı olacak bir rehber oluşturulmalıdır. Bu çalışma, böyle bir rehber önemli derecede ihtiyaç olduğunu göstermiştir.

KAYNAKLAR

- Abe, N. & Mamitsuka, H. (1998). Query learning strategies using boosting and bagging. In *Proceedings fo the International Conference on Machine Learning (ICML)*, Morgan Kaufmann, (pp. 1-9).
- Aggarwal, U., Popescu, A., & Hudelot, C. (2020). Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1428-1437).
- Alamsyah, A. R. B., Anisa, S. R., Belinda, N. S., & Setiawan, A. (2022). SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data. *Proceedings of the International Conference on Data Science and Official Statistics, 2021*(1), 305–314. <https://doi.org/10.34123/icdsos.v2021i1.240>
- Almogahed, B. A., & Kakadiaris, I. A. (2015). NEATER: filtering of over-sampled data using non-cooperative game theory. *Soft Computing*, 19(11), 3301–3322. <https://doi.org/10.1007/s00500-014-1484-5>
- Barua, S., Islam, M. M., Yao, X., & Murase, K. (2014). MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(2), 405–425. <https://doi.org/10.1109/tkde.2012.232>
- Baum, E. B., & Lang, K. (1992, November). Query learning can work poorly when a human oracle is used. In *International joint conference on neural networks* (Vol. 8, p. 8). China: Beijing.
- Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10).
- Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: a Pre-processing Approach for Imbalanced Regression. In *Learning with Imbalanced Domains: Theory and Applications* (pp. 36–50). <http://proceedings.mlr.press/v74/branco17a/branco17a.pdf>
- Branco, P., Torgo, L., & Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343, 76–99. <https://doi.org/10.1016/j.neucom.2018.11.100>
- Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, (pp. 475–482). https://doi.org/10.1007/978-3-642-01307-2_43
- Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2011, December). MUTE: Majority under-sampling technique. In *2011 8th International Conference on Information, Communications & Signal Processing* (pp. 1-4). IEEE. <https://doi.org/10.1109/ICICS.2011.6173603>
- Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36, 664-684. <https://doi.org/10.1007/s10489-011-0287-y>
- Bunkhumpornpat, C., & Sinapiromsaran, K. (2015). CORE: core-based synthetic minority over-sampling and borderline majority under-sampling technique.

- Bunkhumpornpat, C., & Sinapiromsaran, K. (2016, May 27). DBMUTE: density-based majority under-sampling technique. *Knowledge and Information Systems*, 50(3), 827–850. <https://doi.org/10.1007/s10115-016-0957-5>
- Cao, X., Yao, J., Xu, Z., & Meng, D. (2020, July). Hyperspectral Image Classification With Convolutional Neural Network and Active Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7), 4604–4616. <https://doi.org/10.1109/tgrs.2020.2964627>
- Chan P. K., Stolfo S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 164-168). AAAI Press Manchester, UK.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- Chen, Y., Yang, X., & Dai, H. L. (2024). Cost-sensitive continuous ensemble kernel learning for imbalanced data streams with concept drift. *Knowledge-Based Systems*, 284, 111272. <https://doi.org/10.1016/j.knosys.2023.111272>
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. In *BioData Mining*, 10(1). <https://doi.org/10.1186/s13040-017-0155-3>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1). <https://doi.org/10.1186/s12864-019-6413-7>
- Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1). <https://doi.org/10.1186/s13040-021-00244-z>
- Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning*, 15(2), 201–221. <https://doi.org/10.1007/bf00993277>
- Dagan, I., & Engelson, S. P. (1995). Committee-based sampling for training probabilistic classifiers. In *Twelfth International Conference on Machine Learning* The Morgan Kaufmann series in machine learning.
- Dai, Q., Liu, J., & Shi, Y. (2023). Class-overlap undersampling based on Schur decomposition for Class-imbalance problems. *Expert Systems With Applications*, 221, 119735. <https://doi.org/10.1016/j.eswa.2023.119735>
- Das, B., Krishnan, N. C., & Cook, D. J. (2015). RACOG and wRACOG: Two probabilistic oversampling techniques. *IEEE transactions on knowledge and data engineering*, 27(1), 222-234. <https://doi.org/10.1109/tkde.2014.2324567>
- Damtew, Y. G., & Chen, H. (2023). SMMO-CoFS: Synthetic Multi-minority Oversampling with Collaborative Feature Selection for Network Intrusion Detection System. *International Journal of Computational Intelligence Systems*, 16(1). <https://doi.org/10.1007/s44196-022-00171-9>

- Derrac, J., Garcia, S., Sanchez, L., & Herrera, F. (2015). Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Valued Logic Soft Comput*, 17, 255-287.
- Doucette, J., & Heywood, M. I. (2008). GP Classification under Imbalanced Data sets: Active Sub-sampling and AUC Approximation. *Lecture Notes in Computer Science*, 266–277. https://doi.org/10.1007/978-3-540-78671-9_23
- Dou, J., Wei, G., Song, Y., Zhou, D., & Li, M. (2023). Switching Triple-Weight-SMOTE in Empirical Feature Space for Imbalanced and Incomplete Data. *IEEE Transactions on Automation Science and Engineering*, 1–17. <https://doi.org/10.1109/tase.2023.3240759>
- Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, 465, 1-20. <https://doi.org/10.1016/j.ins.2018.06.056>
- Douzas, G., & Bacao, F. (2019). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. *Information Sciences*, 501, 118-135. <https://doi.org/10.1016/j.ins.2019.06.007>
- Dürr, O., Sick, B., & Murina, E. (2020). *Probabilistic deep learning: With python, keras and tensorflow probability*. Manning Publications.
- Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007, November). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 127-136). <http://dx.doi.org/10.1145/1321440.1321461>
- Ertekin, Ş. (2013, September). Adaptive oversampling for imbalanced data classification. In *Information Sciences and Systems 2013: Proceedings of the 28th International Symposium on Computer and Information Sciences* (pp. 261-269). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-01604-7_26
- Fedorov, V. V. (2013). *Theory of optimal experiments*. Elsevier.
- Fu, J., & Lee, S. (2013). Certainty-based active learning for sampling imbalanced datasets. *Neurocomputing*, 119, 350-358. <https://doi.org/10.1016/j.neucom.2013.03.023>
- Halimu, C., Kasem, A., & Newaz, S. S. (2019, January). Empirical comparison of area under ROC curve (AUC) and Mathew correlation coefficient (MCC) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd international conference on machine learning and soft computing* (pp. 1-6). <https://doi.org/10.1145/3310986.3311023>
- Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1* (pp. 878-887). Springer Berlin Heidelberg. https://doi.org/10.1007/11538059_91
- Hart, P. (1968). The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14(3), 515-516. <https://doi.org/10.1109/TIT.1968.1054155>
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758).

New York: springer.

- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73, 220-239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284. <https://doi.org/10.1109/tkde.2008.239>
- Hilas, C. S., & Mastorocostas, P. A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems*, 21(7), 721-726. <https://doi.org/10.1016/j.knosys.2008.03.026>
- Hussein, H. I., Anwar, S. A., & Ahmad, M. I. (2023). Imbalanced Data Classification Using SVM Based on Improved Simulated Annealing Featuring Synthetic Data Generation and Reduction. *CMC-COMPUTERS MATERIALS & CONTINUA*, 75(1), 547-564. <https://doi.org/10.32604/cmc.2023.036025>
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110, 457-506. <https://doi.org/10.1007/s10994-021-05946-3>
- Japkowicz, N. (2001). Concept-learning in the presence of between-class and within-class imbalances. In *Advances in Artificial Intelligence: 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001 Ottawa, Canada, June 7-9, 2001 Proceedings 14* (pp. 67-77). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45153-6_7
- Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing imbalanced data-recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction* (pp. 245-251). IEEE. <https://doi.org/10.1109%2FACII.2013.47>
- Jiang, Z., Zhao, L., Lu, Y., Zhan, Y., & Mao, Q. (2023). A semi-supervised resampling method for class-imbalanced learning. *Expert Systems with Applications*, 119733. <https://doi.org/10.1016/j.eswa.2023.119733>
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 463-484. <https://doi.org/10.1109/TSMCC.2011.2161285>
- Gao, M., Hong, X., Chen, S., Harris, C. J., & Khalaf, E. (2014). PDFOS: PDF estimation based over-sampling for imbalanced two-class problems. *Neurocomputing*, 138, 248-259. <https://doi.org/10.1016/j.neucom.2014.02.006>
- Gaudreault, J. G., Branco, P., & Gama, J. (2021, October). An analysis of performance metrics for imbalanced classification. In *International Conference on Discovery Science* (pp. 67-77). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-88942-5_6
- Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse, W. J., & Zheng, X. (2004). An approach to imbalanced data sets based on changing rule strength. *Rough-*

- Neural Computing: Techniques for Computing with Words*, 543-553.
https://doi.org/10.1007/978-3-030-88942-5_6
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(2003), 1157-1182.
- Guzmán-Ponce, A., Sánchez, J. S., Valdovinos, R. M., & Marcial-Romero, J. R. (2021). DBIG-US: A two-stage under-sampling algorithm to face the class imbalance problem. *Expert Systems with Applications*, 168, 114301.
<https://doi.org/10.1016/j.eswa.2020.114301>
- Kebande, V. R., Alawadi, S., Awaysheh, F. M., & Persson, J. A. (2021). Active machine learning adversarial attack detection in the user feedback process. *IEEE Access*, 9, 36908-36923. <https://doi.org/10.1109/ACCESS.2021.3063002>
- Kou, G., Chen, H., & Hefni, M. A. (2022). Improved hybrid resampling and ensemble model for imbalance learning and credit evaluation. *Journal of Management Science and Engineering*, 7(4), 511-529,
<https://doi.org/10.1016/j.jmse.2022.06.002>.
- Koziarski, M. (2021). CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification. In *International Joint Conference on Neural Network*.
<https://doi.org/10.1109/ijcnn52387.2021.9533415>
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221-232.
<https://doi.org/10.1007/s13748-016-0094-0>
- Kubát, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, pp. 179-186.
- Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., & Engelhardt, A. (2020). Caret: classification and regression training. R package version 6.0-86. 2020. URL <http://CRAN.R-project.org/package=caret>.
- Kumari, M., & Subbarao, N. (2022). A hybrid resampling algorithms SMOTE and ENN based deep learning models for identification of Marburg virus inhibitors. *Future Medicinal Chemistry*, 14(10), 701–715. <https://doi.org/10.4155/fmc-2021-0290>
- Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2), 145-151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Lamari, M., Azizi, N., Hammami, N. E., Boukhamla, A., Cheriguene, S., Dendani, N., & Benzebouchi, N. E. (2021). SMOTE–ENN-Based Data Sampling and Improved Dynamic Ensemble Selection for Imbalanced Medical Data Classification. In *Advances on Smart and Soft Computing: Proceedings of ICACIn 2020* (pp. 37-49). Springer Singapore. https://doi.org/10.1007/978-981-15-6048-4_4
- Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal*,

- July 1–4, 2001, *Proceedings* 8 (pp. 63–66). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48229-6_9
- Lee, S. S. (1999). Regularization in skewed binary classification. *Computational Statistics*, 14(2), 277–292. <https://doi.org/10.1007/s001800050018>
- Lee, W., & Seo, K. (2022). Downsampling for binary classification with a highly imbalanced dataset using active learning. *Big Data Research*, 28, 100314. <https://doi.org/10.1016/j.bdr.2022.100314>
- Liu, Y., Wang, S., Sui, H., & Zhu, L. (2024). An ensemble learning method with GAN-based sampling and consistency check for anomaly detection of imbalanced data streams with concept drift. *Plos one*, 19(1), e0292140. <https://doi.org/10.1371/journal.pone.0292140>
- McCallum, A., & Nigam, K. (1998, July). Employing EM and Pool-Based Active Learning for Text Classification. In *Proceedings of the 14th International Conference on Machine Learning (ICML '98)*, pp. 350–358.
- Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1), 92–122. <https://doi.org/10.1007/s10618-012-0295-5>
- Mortaz, E. (2020). Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems*, 210, 106490. <https://doi.org/10.1016/j.knosys.2020.106490>
- Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1), 4–21. <https://doi.org/10.1504/ijkesdp.2011.039875>
- Park, Y. J., Pan, R., & Montgomery, D. C. (2023). A novel hybrid resampling for semiconductor wafer defect bin classification. *Quality and Reliability Engineering International*, 39(1), 67–80. <https://doi.org/10.1002/qre.3217>
- Peng, X., Jin, X., Duan, S., & Sankavaram, C. (2023). Active learning-assisted semi-supervised learning for fault detection and diagnostics with imbalanced dataset. *IIEE Transactions*, 55(7), 672–686. <https://doi.org/10.1080/24725854.2022.2074579>
- Pradipta, G. A., Wardoyo, R., Musdholifah, A., & Sanjaya, I. N. H. (2021). Radius-SMOTE: a new oversampling technique of minority samples based on radius distance for learning from imbalanced data. *IEEE Access*, 9, 74763–74777. <https://doi.org/10.1109/access.2021.3080316>
- Puri, A., & Kumar Gupta, M. (2022). Improved hybrid bag-boost ensemble with K-means-SMOTE–ENN technique for handling noisy class imbalanced data. *The Computer Journal*, 65(1), 124–138. <https://doi.org/10.1093/comjnl/bxab039>
- Radivoja P, Chawla NV, Dunker AK, Obradovic Z (2004) Classification and knowledge discovery in protein databases. *J Biomed Inf* 37(4):224–239. <https://doi.org/10.1016/j.jbi.2004.07.008>
- Ren, Z., Zhu, Y., Kang, W., Fu, H., Niu, Q., Gao, D., ... & Hong, J. (2022). Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data. *Knowledge-based systems*, 241, 108296. <https://doi.org/10.1016/j.knosys.2022.108296>

- Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, 110415. <https://doi.org/10.1016/j.asoc.2023.110415>
- Sağlam, F., & Cengiz, M. A. (2022). A novel SMOTE-based resampling technique through noise detection and the boosting procedure. *Expert Systems with Applications*, 200, 117023. <https://doi.org/10.1016/j.eswa.2022.117023>
- Scheffer, T., Decomain, C., & Wrobel, S. (2001, September). Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis* (pp. 309-318). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44816-0_31
- Settles, B., Craven, M., & Ray, S. (2007). Multiple-instance active learning. *Advances in neural information processing systems (NIPS 2007)*, 20.
- Seung, H. S., Oppen, M., & Sompolinsky, H. (1992, July). Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 287-294). <https://doi.org/10.1145/130385.130417>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Siriseriwan, W., & Sinapiromsaran, K. (2016). The effective redistribution for imbalance dataset: Relocating safe-level SMOTE with minority outcast handling. *Chiang Mai J. Sci*, 43(1), 234-246.
- Siriseriwan, W., & Sinapiromsaran, K. (2017). Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling. *Songklanakarin Journal of Science & Technology*, 39(5). <http://dx.doi.org/10.14456/sjst-psu.2017.70>
- Sowah, R. A., Kuditchar, B., Mills, G. A., Acakpovi, A., Twum, R. A., Buah, G., & Agboyi, R. (2021). HCBST: an efficient hybrid sampling technique for class imbalance problems. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3), 1-37. <https://doi.org/10.1145/3488280>
- Sun, Y., Sun, Y., & Dai, H. (2020). Two-stage cost-sensitive learning for data streams with concept drift and class imbalance. *IEEE Access*, 8, 191942-191955. <https://doi.org/10.1109/ACCESS.2020.3031603>
- Tarawneh, A. S., Hassanat, A. B., Altarawneh, G. A., & Almuhaimeed, A. (2022). Stop oversampling for class imbalance learning: A review. *IEEE Access*, 10, 47643-47660. <https://doi.org/10.1109/ACCESS.2022.3169512>
- Thölke, P., Mantilla-Ramos, Y. J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., ... & Jerbi, K. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, 277, 120253. <https://doi.org/10.1016/j.neuroimage.2023.120253>
- Tomanek, K., & Hahn, U. (2009, September). Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture* (pp. 105-112). <https://doi.org/10.1145/1597735.1597754>
- Tomek, I. (1976). A generalization of the k-NN rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (2), 121-126.

<https://doi.org/10.1109/TSMC.1976.5409182>

- Tsai, C. F., Lin, W. C., Hu, Y. H., & Yao, G. T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, 477, 47-54. <https://doi.org/10.1016/j.ins.2018.10.029>
- Valdenegro-Toro, M., & Mori, D. S. (2022, June). A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1508-1516). IEEE. <https://doi.org/10.1109/CVPRW56347.2022.00157>
- Verbeke, W., Olaya, D., Guerry, M. A., & Van Belle, J. (2023). To do or not to do? Cost-sensitive causal classification with individual treatment effect estimates. *European Journal of Operational Research*, 305(2), 838-852. <https://doi.org/10.1016/j.ejor.2022.03.049>
- Wang J. (2023). UCI datasets. IEEE Dataport. <https://dx.doi.org/10.21227/g4y0-sw34>
- Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., & Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338, 34-45. <https://doi.org/10.1016/j.neucom.2019.01.103>
- Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12), 2591-2600. <https://doi.org/10.1109/TCSVT.2016.2589879>
- Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019, October). Cross-validation metrics for evaluating classification performance on imbalanced data. In *2019 international conference on computer, control, informatics and its applications (IC3INA)* (pp. 14-18). IEEE. <https://doi.org/10.1109/IC3INA48034.2019.8949568>
- Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3), 408-421. <https://doi.org/10.1109/TSMC.1972.4309137>
- Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016, May). Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)* (pp. 225-228). IEEE. <https://doi.org/10.1109/icoacs.2016.7563084>
- Zhang, H., & Li, M. (2014). RWO-Sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, 20, 99-116. <https://doi.org/10.1016/j.inffus.2013.12.003>
- Zhang, J., & Zhang, I. (2003, August). kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets* (Vol. 126, No. 1, pp. 1-7). ICML.
- Zheng, Z., Cai, Y., & Li, Y. (2015). Oversampling Method for Imbalanced Classification. *Computing and Informatics / Computers and Artificial Intelligence*, 34(5), 1017-1037. <https://dblp.uni-trier.de/db/journals/cai/cai34.html#ZhengCL15>

- Zhong, M., Tran, K., Min, Y., Wang, C., Wang, Z., Dinh, C. T., ... & Sargent, E. H. (2020). Accelerated discovery of CO₂ electrocatalysts using active machine learning. *Nature*, 581(7807), 178-183. <https://doi.org/10.1038/s41586-020-2242-8>
- Zhu, Q. (2020). On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters*, 136, 71-80. <https://doi.org/10.1016/j.patrec.2020.03.030>
- Zhu, J., & Hovy, E. (2007, June). Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 783-790).



EKLER

Ek 1. İki sınıflı veri setlerinde genel sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.094	0.060	-0.102	0.062	0.049	0.138	0.152	0.120
V2	-0.020	-0.103	-0.320	-0.115	-0.137	0.004	-0.001	0.020
V3	0.141	0.078	-0.123	0.068	0.042	0.236	0.299	0.261
V4	-0.010	-0.125	-0.321	-0.139	-0.181	-0.011	-0.011	-0.001
V5	0.008	-0.051	-0.143	-0.045	-0.058	0.101	0.112	0.103
V6	0.233	0.162	-0.029	0.156	0.157	0.263	0.266	0.250
V7	0.301	0.183	0.118	0.183	0.155	0.261	0.262	0.255
V8	0.053	0.055	0.028	0.140	0.073	0.204	0.208	0.215
V9	0.307	0.226	0.177	0.270	0.249	0.339	0.329	0.314
V10	0.646	0.541	0.363	0.621	0.562	0.655	0.655	0.662
V11	0.468	0.347	0.130	0.381	0.319	0.451	0.468	0.475
V12	0.540	0.447	0.212	0.480	0.405	0.554	0.567	0.556
V13	0.607	0.474	0.291	0.516	0.471	0.628	0.629	0.624
V14	0.681	0.602	0.439	0.661	0.629	0.679	0.689	0.694
V15	0.661	0.593	0.448	0.651	0.630	0.682	0.681	0.685
V16	0.591	0.419	0.247	0.432	0.379	0.581	0.592	0.592
V17	0.621	0.562	0.409	0.620	0.583	0.642	0.638	0.651
V18	0.540	0.472	0.272	0.493	0.453	0.550	0.547	0.550
V19	0.658	0.608	0.483	0.655	0.635	0.676	0.677	0.676
V20	0.641	0.595	0.440	0.629	0.597	0.670	0.666	0.672
V21	0.615	0.437	0.291	0.448	0.398	0.604	0.601	0.604
V22	0.657	0.520	0.336	0.558	0.470	0.675	0.677	0.669
V23	0.639	0.593	0.523	0.627	0.612	0.651	0.663	0.664
V24	0.437	0.365	0.269	0.400	0.349	0.441	0.461	0.445
V25	0.719	0.608	0.412	0.659	0.608	0.725	0.734	0.727
V26	0.000	-0.030	-0.082	-0.016	-0.020	0.066	0.065	0.069
V27	0.048	0.025	-0.009	0.041	0.043	0.126	0.136	0.123
V28	0.003	0.093	-0.161	0.139	0.054	0.143	0.124	0.129
V29	-0.022	0.063	-0.184	0.078	0.005	0.123	0.099	0.111
V30	-0.005	0.107	-0.195	0.127	0.036	0.100	0.109	0.119
V31	0.728	0.696	0.613	0.697	0.700	0.729	0.736	0.729
V32	0.754	0.729	0.440	0.741	0.813	0.809	0.783	0.859
V33	0.859	0.831	0.732	0.821	0.845	0.922	0.920	0.930
V34	0.836	0.789	0.552	0.809	0.827	0.848	0.847	0.810
V35	0.686	0.642	0.286	0.659	0.669	0.748	0.729	0.700
V36	0.547	0.396	0.189	0.438	0.441	0.472	0.478	0.449
V37	-0.020	-0.077	-0.236	-0.022	-0.034	-0.010	-0.017	-0.008
V38	0.740	0.696	0.209	0.718	0.743	0.744	0.793	0.784
V39	0.642	0.388	0.300	0.502	0.493	0.522	0.562	0.546
V40	0.811	0.820	0.772	0.843	0.842	0.853	0.856	0.861
V41	0.702	0.743	0.369	0.742	0.691	0.741	0.745	0.736
V42	0.004	-0.057	-0.384	-0.043	-0.078	0.000	0.004	0.001
V43	0.156	0.119	-0.329	0.169	0.347	0.149	0.214	0.322
V44	0.781	0.698	0.699	0.689	0.705	0.765	0.745	0.770
V45	0.280	0.148	0.076	0.189	0.171	0.212	0.218	0.227
V46	0.545	0.401	0.341	0.437	0.426	0.456	0.443	0.453
V47	0.950	0.899	0.723	0.912	0.910	0.975	0.975	0.972
V48	-0.083	-0.118	-0.151	-0.109	-0.076	-0.116	-0.103	-0.088

V49	0.847	0.782	0.566	0.797	0.777	0.822	0.835	0.835
V50	0.719	0.610	0.393	0.630	0.603	0.675	0.665	0.680
V51	0.667	0.644	0.408	0.671	0.642	0.687	0.686	0.701
V52	0.493	0.467	0.212	0.495	0.457	0.531	0.532	0.530
V53	0.491	0.503	0.198	0.521	0.498	0.547	0.544	0.552
V54	0.680	0.580	0.358	0.604	0.585	0.615	0.622	0.629
V55	0.479	0.414	0.157	0.449	0.406	0.457	0.464	0.465
V56	0.391	0.340	0.051	0.376	0.341	0.387	0.397	0.399
V57	0.425	0.419	0.150	0.451	0.434	0.450	0.464	0.472
V58	0.707	0.592	0.397	0.633	0.607	0.640	0.657	0.663
V59	0.419	0.448	0.159	0.483	0.458	0.479	0.501	0.497
V60	0.057	-0.011	-0.265	-0.051	-0.026	0.122	0.136	0.147
V61	-0.005	-0.065	-0.252	-0.091	-0.053	0.013	0.025	0.038
V62	0.000	0.041	-0.167	0.027	-0.014	0.017	0.039	0.038
V63	-0.012	-0.030	-0.269	-0.052	-0.102	-0.005	0.002	-0.006
V64	0.359	0.321	0.194	0.323	0.379	0.426	0.417	0.458
V65	0.685	0.560	0.379	0.561	0.584	0.646	0.673	0.674
V66	0.132	0.084	-0.043	0.096	0.134	0.166	0.199	0.191
V67	0.117	-0.104	-0.273	-0.156	-0.103	0.171	0.175	0.182
V68	-0.043	-0.181	-0.311	-0.201	-0.193	-0.082	-0.096	-0.089
V69	0.164	0.061	-0.144	0.041	0.034	0.174	0.211	0.210
V70	0.625	0.523	0.429	0.552	0.529	0.599	0.608	0.605
V71	0.645	0.522	0.478	0.555	0.540	0.633	0.639	0.630
V72	0.159	0.065	-0.028	0.058	0.030	0.233	0.267	0.247
V73	0.564	0.451	0.235	0.466	0.440	0.510	0.508	0.498
V74	0.263	0.170	0.014	0.171	0.139	0.350	0.383	0.375
Ortalama	0.416	0.350	0.168	0.370	0.353	0.433	0.440	0.442

Ek 2. İki sınıflı veri setlerinde genel sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	4.500	4.857	7.786	4.929	5.571	2.714	2.571	3.071
V2	3.929	5.214	7.857	5.357	6.071	2.929	2.357	2.286
V3	4.500	5.286	7.714	5.000	6.143	3.000	1.929	2.429
V4	2.893	5.000	7.929	5.000	6.857	2.929	3.071	2.321
V5	4.429	5.357	7.714	5.429	6.071	2.500	2.071	2.429
V6	3.929	5.286	7.857	5.857	5.571	2.321	2.536	2.643
V7	3.071	5.500	6.357	5.000	5.714	3.643	3.214	3.500
V8	6.357	6.000	6.321	4.071	5.607	2.214	2.643	2.786
V9	5.143	5.643	6.786	4.500	4.929	2.786	3.000	3.214
V10	2.679	6.143	7.929	4.000	5.286	3.571	3.464	2.929
V11	2.857	5.643	7.929	5.071	6.643	3.143	2.500	2.214
V12	3.357	5.000	8.000	5.214	6.643	2.714	1.929	3.143
V13	3.071	5.857	7.786	5.500	6.571	2.286	2.286	2.643
V14	3.786	5.714	7.857	4.000	5.429	3.571	2.929	2.714
V15	3.964	5.786	7.857	4.143	4.929	3.464	3.214	2.643
V16	2.107	5.786	7.500	5.786	6.714	3.000	2.786	2.321
V17	3.821	6.000	7.929	3.857	5.000	3.429	3.607	2.357
V18	3.179	4.857	8.000	5.214	6.214	2.679	3.107	2.750
V19	4.143	5.429	7.929	4.143	4.929	3.250	3.107	3.071
V20	4.286	4.929	7.929	4.929	5.571	2.786	3.143	2.429
V21	2.107	5.714	7.214	6.000	6.929	2.357	2.786	2.893
V22	3.393	5.500	7.643	5.000	7.000	2.107	2.250	3.107
V23	4.071	5.429	7.500	4.429	5.714	3.857	2.643	2.357

V24	2.893	5.714	7.786	4.357	6.857	3.393	2.250	2.750
V25	3.393	5.571	7.929	5.000	6.357	3.107	2.179	2.464
V26	5.929	5.786	7.143	4.857	5.000	2.286	2.214	2.786
V27	4.429	6.357	6.929	5.286	5.214	2.357	2.357	3.071
V28	5.500	4.429	7.643	3.071	5.071	3.000	3.857	3.429
V29	5.643	4.571	7.071	3.929	5.214	2.964	3.321	3.286
V30	5.357	3.857	7.929	2.857	5.286	4.107	3.429	3.179
V31	3.643	5.071	6.571	5.107	4.821	3.750	3.393	3.643
V32	4.214	4.714	7.679	3.643	3.750	3.643	4.321	4.036
V33	4.143	4.500	6.464	5.071	3.786	3.750	4.179	4.107
V34	3.750	4.714	7.786	4.286	4.107	3.750	3.357	4.250
V35	4.321	4.286	7.643	4.893	4.750	2.857	3.607	3.643
V36	2.429	6.000	7.571	4.500	4.714	3.000	3.357	4.429
V37	4.214	5.714	8.000	3.929	4.786	2.714	3.571	3.071
V38	4.571	4.643	7.786	4.321	4.107	3.607	3.250	3.714
V39	1.893	6.179	7.500	4.286	5.143	4.357	3.071	3.571
V40	4.679	5.286	6.857	3.929	3.429	4.500	4.143	3.179
V41	4.214	3.857	7.857	3.964	5.500	3.607	3.250	3.750
V42	4.357	4.429	8.000	4.286	5.107	3.429	2.821	3.571
V43	4.393	4.821	7.750	4.250	2.964	4.321	3.893	3.607
V44	2.929	6.000	6.714	5.714	6.143	2.857	2.286	3.357
V45	2.643	5.500	7.000	3.500	4.643	4.357	4.357	4.000
V46	2.286	5.429	7.214	3.857	4.500	4.214	4.214	4.286
V47	4.500	4.821	7.286	4.607	4.679	3.286	3.464	3.357
V48	4.000	5.643	6.857	4.929	4.214	3.643	2.786	3.929
V49	3.000	5.214	7.929	4.286	5.643	2.929	3.071	3.929
V50	2.643	5.357	8.000	4.571	5.643	2.929	3.393	3.464
V51	4.643	4.929	7.929	3.857	5.000	3.071	3.786	2.786
V52	4.071	5.071	7.786	4.857	5.643	2.857	2.571	3.143
V53	5.071	4.429	7.857	4.214	5.357	3.214	3.000	2.857
V54	2.929	5.714	8.000	4.714	4.929	3.214	3.071	3.429
V55	3.929	4.643	7.857	3.500	5.643	3.714	3.500	3.214
V56	3.500	5.143	8.000	4.000	5.786	3.714	2.786	3.071
V57	3.929	4.286	7.929	4.000	4.000	4.429	3.714	3.714
V58	3.357	5.143	7.643	4.214	5.786	3.643	3.286	2.929
V59	5.071	5.000	7.786	3.929	4.786	3.643	2.786	3.000
V60	4.607	4.786	7.929	5.857	5.143	3.036	2.393	2.250
V61	4.143	5.000	7.929	5.500	4.714	3.286	3.000	2.429
V62	4.214	3.429	7.786	3.786	5.357	4.250	3.679	3.500
V63	3.179	3.071	8.000	4.857	6.286	3.321	3.286	4.000
V64	4.786	5.571	7.571	5.643	3.500	3.143	3.429	2.357
V65	2.429	6.000	7.571	5.643	5.000	3.750	3.071	2.536
V66	4.429	5.214	7.071	5.214	4.571	3.857	2.429	3.214
V67	3.500	5.786	7.429	6.500	6.000	2.500	2.071	2.214
V68	2.071	5.429	7.714	6.000	5.643	2.429	3.286	3.429
V69	3.786	5.571	7.929	5.357	5.929	3.321	2.000	2.107
V70	2.571	5.786	7.786	4.357	5.857	3.143	3.107	3.393
V71	2.500	6.357	7.286	5.357	6.000	2.857	2.571	3.071
V72	4.214	5.286	7.857	5.500	6.714	2.857	1.357	2.214
V73	2.286	5.286	7.929	4.643	5.357	3.393	3.393	3.714
V74	3.929	5.571	7.929	5.286	6.214	3.071	1.643	2.357
Ortalama	3.793	5.228	7.604	4.682	5.361	3.239	2.997	3.095

Ek 3. İki sınıflı veri setlerinde lineer sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.201	0.075	0.035	0.105	0.033	0.159	0.166	0.154
V2	-0.020	-0.136	-0.201	-0.134	-0.155	0.044	0.041	0.026
V3	0.282	0.099	0.016	0.130	0.075	0.421	0.444	0.440
V4	-0.008	-0.173	-0.215	-0.164	-0.185	-0.010	-0.009	-0.012
V5	0.076	0.024	-0.001	0.037	0.011	0.166	0.166	0.157
V6	0.398	0.271	0.188	0.312	0.220	0.446	0.461	0.390
V7	-0.100	-0.342	-0.350	-0.344	-0.345	-0.344	-0.346	-0.345
V8	-0.040	0.025	0.007	0.032	0.005	0.123	0.123	0.115
V9	-0.075	0.009	-0.001	0.011	0.014	0.084	0.084	0.087
V10	0.582	0.380	0.230	0.501	0.380	0.657	0.669	0.668
V11	0.329	0.205	0.139	0.288	0.188	0.392	0.412	0.387
V12	0.449	0.297	0.218	0.358	0.288	0.497	0.510	0.501
V13	0.516	0.352	0.250	0.413	0.344	0.582	0.585	0.581
V14	0.621	0.482	0.362	0.559	0.487	0.642	0.653	0.649
V15	0.627	0.500	0.405	0.613	0.513	0.691	0.697	0.691
V16	0.541	0.203	0.193	0.332	0.207	0.542	0.574	0.546
V17	0.600	0.496	0.350	0.597	0.487	0.694	0.693	0.696
V18	0.459	0.399	0.278	0.453	0.410	0.527	0.523	0.522
V19	0.601	0.544	0.459	0.612	0.558	0.678	0.676	0.666
V20	0.624	0.521	0.378	0.594	0.504	0.724	0.718	0.723
V21	0.559	0.269	0.221	0.380	0.229	0.552	0.565	0.568
V22	0.630	0.412	0.311	0.516	0.390	0.723	0.722	0.718
V23	0.508	0.481	0.444	0.488	0.482	0.558	0.551	0.557
V24	0.372	0.338	0.295	0.367	0.319	0.474	0.470	0.473
V25	0.751	0.617	0.421	0.678	0.605	0.741	0.759	0.746
V26	-0.047	-0.068	-0.099	-0.053	0.008	0.074	0.060	0.079
V27	-0.080	0.009	-0.031	-0.018	0.067	0.050	0.054	0.107
V28	-0.089	0.073	-0.098	0.105	0.034	0.132	0.135	0.114
V29	-0.071	0.092	-0.097	0.124	0.060	0.172	0.185	0.121
V30	-0.063	0.097	-0.097	0.105	0.063	0.129	0.109	0.121
V31	0.683	0.639	0.665	0.663	0.643	0.691	0.694	0.681
V32	0.640	0.574	0.409	0.552	0.872	0.819	0.778	0.841
V33	0.925	0.913	0.816	0.934	0.936	0.914	0.901	0.917
V34	0.919	0.866	0.638	0.864	0.859	0.867	0.862	0.865
V35	0.742	0.631	0.309	0.608	0.564	0.664	0.660	0.634
V36	0.648	0.308	0.245	0.354	0.289	0.389	0.391	0.363
V37	-0.018	-0.164	-0.188	-0.150	-0.158	-0.137	-0.144	-0.141
V38	0.921	0.766	0.318	0.767	0.767	0.847	0.841	0.847
V39	0.671	0.264	0.256	0.329	0.260	0.497	0.512	0.554
V40	0.733	0.865	0.831	0.858	0.888	0.863	0.863	0.876
V41	0.355	0.471	0.235	0.400	0.438	0.526	0.498	0.489
V42	-0.012	-0.386	-0.506	-0.387	-0.385	-0.256	-0.246	-0.263
V43	-0.017	-0.409	-0.521	-0.423	-0.406	-0.480	-0.486	-0.486
V44	0.741	0.786	0.786	0.786	0.786	0.845	0.848	0.844
V45	-0.087	-0.242	-0.245	-0.236	-0.242	-0.253	-0.255	-0.254
V46	0.255	0.010	-0.006	0.019	0.011	-0.009	-0.004	-0.008
V47	0.891	0.923	0.679	0.911	0.919	0.988	0.994	0.974
V48	-0.061	-0.057	-0.149	-0.056	-0.057	-0.057	-0.057	-0.057
V49	0.702	0.594	0.464	0.612	0.565	0.657	0.654	0.630
V50	0.463	0.328	0.292	0.334	0.325	0.354	0.362	0.342
V51	0.369	0.365	0.305	0.392	0.362	0.452	0.457	0.459

V52	0.001	0.103	0.079	0.104	0.109	0.142	0.149	0.139
V53	-0.022	0.080	0.037	0.088	0.077	0.117	0.116	0.109
V54	0.389	0.170	0.123	0.184	0.174	0.196	0.203	0.193
V55	-0.012	-0.026	-0.060	-0.019	-0.028	-0.038	-0.015	-0.036
V56	-0.089	-0.119	-0.150	-0.113	-0.117	-0.109	-0.095	-0.108
V57	-0.087	-0.171	-0.210	-0.177	-0.172	-0.202	-0.184	-0.203
V58	0.430	0.220	0.192	0.228	0.217	0.252	0.255	0.240
V59	-0.092	0.032	-0.002	0.042	0.032	0.055	0.059	0.057
V60	0.045	-0.197	-0.296	-0.195	-0.199	0.116	0.125	0.118
V61	-0.021	-0.185	-0.251	-0.186	-0.196	0.052	0.053	0.060
V62	-0.028	-0.085	-0.123	-0.087	-0.097	0.029	0.055	0.045
V63	-0.021	-0.159	-0.243	-0.158	-0.164	-0.039	-0.040	-0.054
V64	0.179	0.241	0.223	0.236	0.240	0.381	0.380	0.375
V65	0.654	0.444	0.439	0.460	0.455	0.626	0.643	0.619
V66	0.119	-0.013	-0.067	-0.013	-0.016	0.156	0.166	0.147
V67	0.014	-0.192	-0.198	-0.192	-0.187	0.127	0.130	0.119
V68	-0.043	-0.244	-0.292	-0.243	-0.248	-0.143	-0.151	-0.145
V69	0.041	-0.014	-0.046	0.014	-0.009	0.165	0.163	0.166
V70	0.547	0.469	0.405	0.462	0.443	0.551	0.584	0.570
V71	0.540	0.438	0.443	0.455	0.442	0.544	0.544	0.540
V72	0.071	0.021	-0.009	0.024	0.012	0.194	0.210	0.195
V73	0.376	0.275	0.230	0.274	0.274	0.326	0.335	0.326
V74	0.161	0.049	0.012	0.076	0.046	0.251	0.259	0.259
Ortalama	0.307	0.213	0.133	0.235	0.211	0.328	0.331	0.325

Ek 4. İki sınıflı veri setlerinde lineer sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	3.000	5.667	7.333	4.667	7.333	2.333	2.333	3.333
V2	4.000	5.667	8.000	5.333	7.000	1.333	1.667	3.000
V3	4.167	5.667	8.000	4.667	7.000	2.833	1.833	1.833
V4	2.000	6.000	8.000	5.000	7.000	3.000	2.667	2.333
V5	4.333	5.667	7.667	4.667	6.667	2.000	2.000	3.000
V6	3.333	5.667	7.667	5.333	7.333	1.667	1.667	3.333
V7	1.000	4.333	5.000	5.333	6.333	4.333	4.667	5.000
V8	8.000	6.000	6.333	4.000	5.667	1.667	1.333	3.000
V9	8.000	5.667	7.000	4.667	4.667	2.333	2.667	1.000
V10	3.000	6.667	8.000	5.000	6.333	2.333	2.333	2.333
V11	4.000	6.000	8.000	4.667	7.000	2.667	1.333	2.333
V12	3.000	6.333	8.000	5.000	6.667	3.000	1.333	2.667
V13	4.000	6.333	8.000	5.000	6.667	1.667	2.000	2.333
V14	3.667	6.333	8.000	5.000	6.667	2.000	2.000	2.333
V15	3.333	7.000	8.000	4.667	6.000	2.667	2.000	2.333
V16	2.667	7.667	6.667	5.000	6.667	2.667	1.667	3.000
V17	4.667	6.333	8.000	4.333	6.667	2.167	2.000	1.833
V18	4.000	6.667	8.000	4.667	5.667	2.333	2.333	2.333
V19	5.333	6.667	8.000	4.333	5.667	2.000	1.667	2.333
V20	4.667	5.667	8.000	4.667	7.000	2.000	2.333	1.667
V21	3.000	6.333	7.000	5.000	7.667	2.500	2.167	2.333
V22	4.000	6.000	8.000	5.000	7.000	1.500	2.000	2.500
V23	5.333	5.667	7.333	5.333	6.333	1.667	2.667	1.667
V24	4.667	5.667	7.667	4.667	7.333	1.667	2.333	2.000
V25	2.333	6.333	8.000	5.000	6.667	3.000	2.000	2.667
V26	6.667	5.667	6.667	4.667	6.333	1.333	2.667	2.000

V27	6.333	5.667	6.667	5.333	4.667	3.000	2.000	2.333
V28	7.333	4.333	7.667	3.333	6.000	2.333	2.333	2.667
V29	7.333	5.000	7.667	3.667	5.667	1.667	1.333	3.667
V30	7.000	4.333	8.000	2.667	5.667	3.000	3.333	2.000
V31	4.333	6.167	4.667	4.500	5.667	3.333	3.000	4.333
V32	3.000	6.167	8.000	5.333	3.833	2.333	3.667	3.667
V33	3.000	5.333	7.000	4.000	3.333	4.000	4.333	5.000
V34	3.333	3.500	8.000	4.333	4.333	3.833	4.333	4.333
V35	1.000	3.500	8.000	4.833	6.333	3.833	3.833	4.667
V36	1.000	6.000	8.000	4.333	7.000	2.667	2.667	4.333
V37	1.000	7.000	8.000	5.000	6.000	2.000	3.667	3.333
V38	3.000	5.000	8.000	4.667	4.333	3.500	4.000	3.500
V39	1.000	6.333	7.333	5.000	7.333	3.833	3.167	2.000
V40	6.000	4.333	5.333	3.333	2.667	5.333	5.667	3.333
V41	5.000	3.833	7.667	4.833	6.000	2.000	3.000	3.667
V42	1.000	6.333	8.000	6.000	5.667	3.000	2.333	3.667
V43	1.000	2.667	7.667	4.000	2.333	6.333	6.000	6.000
V44	5.333	6.667	5.667	6.333	6.000	2.000	1.000	3.000
V45	1.000	3.667	4.667	2.000	4.333	6.333	7.333	6.667
V46	1.000	4.000	6.667	2.000	4.333	7.333	4.667	6.000
V47	5.500	4.333	8.000	5.167	5.500	2.500	2.167	2.833
V48	7.333	4.333	7.667	2.000	3.667	3.333	4.000	3.667
V49	1.000	6.000	8.000	4.333	6.667	2.667	2.667	4.667
V50	1.000	5.667	8.000	4.667	6.000	3.333	2.667	4.667
V51	5.667	6.000	7.667	4.333	6.333	2.667	2.000	1.333
V52	7.667	5.667	7.333	5.000	4.333	1.667	1.333	3.000
V53	7.667	5.000	7.333	4.333	5.667	1.667	1.667	2.667
V54	1.000	7.000	8.000	4.667	6.000	3.333	2.000	4.000
V55	4.000	3.667	7.333	3.000	4.667	5.333	3.000	5.000
V56	1.333	6.667	8.000	4.333	6.000	4.000	2.333	3.333
V57	1.000	2.667	7.667	4.333	2.333	6.333	4.667	7.000
V58	1.333	6.000	7.000	5.333	7.000	2.667	2.667	4.000
V59	8.000	6.000	7.000	4.000	5.000	2.667	1.000	2.333
V60	3.000	6.000	8.000	5.333	6.667	2.667	1.667	2.667
V61	4.000	5.667	8.000	5.667	6.667	2.333	2.333	1.333
V62	4.000	5.667	7.000	6.333	7.000	2.333	2.000	1.667
V63	1.333	5.667	8.000	5.667	6.667	2.667	2.667	3.333
V64	6.667	5.333	7.000	5.667	5.333	2.000	1.667	2.333
V65	2.000	7.333	7.000	5.667	6.000	3.333	1.667	3.000
V66	3.000	5.667	8.000	6.000	6.333	2.000	2.333	2.667
V67	4.000	7.000	7.000	6.000	6.000	2.000	1.333	2.667
V68	1.000	5.667	8.000	5.667	6.667	2.667	3.333	3.000
V69	5.333	6.333	7.667	4.667	6.000	2.167	1.667	2.167
V70	4.667	5.000	7.667	4.667	6.667	3.000	1.333	3.000
V71	3.333	6.667	6.333	5.333	6.333	2.333	2.667	3.000
V72	5.333	5.000	7.667	5.333	6.667	2.000	1.667	2.333
V73	2.000	6.333	7.667	6.000	6.000	3.000	2.333	2.667
V74	4.333	6.000	7.667	4.667	6.333	3.000	2.333	1.667
Ortalama	3.806	5.619	7.441	4.721	5.910	2.838	2.574	3.090

Ek 5. İki sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.065	0.056	-0.140	0.050	0.053	0.133	0.148	0.110
V2	-0.020	-0.093	-0.353	-0.110	-0.132	-0.006	-0.013	0.019
V3	0.102	0.072	-0.161	0.051	0.033	0.185	0.260	0.212
V4	-0.011	-0.112	-0.350	-0.133	-0.180	-0.012	-0.012	0.003
V5	-0.011	-0.071	-0.181	-0.068	-0.077	0.083	0.097	0.089
V6	0.188	0.132	-0.088	0.114	0.140	0.213	0.213	0.212
V7	0.411	0.327	0.246	0.327	0.292	0.426	0.428	0.418
V8	0.078	0.063	0.033	0.169	0.091	0.227	0.231	0.242
V9	0.411	0.285	0.226	0.340	0.313	0.408	0.396	0.376
V10	0.663	0.584	0.400	0.654	0.612	0.654	0.651	0.661
V11	0.506	0.385	0.128	0.407	0.354	0.467	0.483	0.499
V12	0.565	0.488	0.210	0.513	0.437	0.570	0.583	0.571
V13	0.631	0.508	0.302	0.544	0.506	0.640	0.641	0.636
V14	0.697	0.634	0.460	0.689	0.668	0.689	0.699	0.706
V15	0.670	0.618	0.459	0.662	0.662	0.679	0.676	0.683
V16	0.605	0.478	0.261	0.459	0.426	0.591	0.596	0.605
V17	0.627	0.580	0.425	0.627	0.609	0.628	0.623	0.638
V18	0.562	0.491	0.270	0.503	0.465	0.556	0.554	0.557
V19	0.674	0.625	0.489	0.667	0.656	0.676	0.677	0.679
V20	0.646	0.615	0.458	0.638	0.622	0.655	0.652	0.658
V21	0.631	0.483	0.310	0.466	0.444	0.618	0.611	0.614
V22	0.664	0.549	0.344	0.570	0.492	0.662	0.665	0.655
V23	0.675	0.624	0.544	0.665	0.648	0.676	0.693	0.694
V24	0.455	0.373	0.262	0.409	0.357	0.432	0.459	0.437
V25	0.710	0.606	0.410	0.654	0.609	0.721	0.727	0.722
V26	0.012	-0.019	-0.077	-0.007	-0.028	0.064	0.067	0.066
V27	0.082	0.029	-0.003	0.057	0.037	0.147	0.158	0.127
V28	0.028	0.098	-0.178	0.148	0.060	0.146	0.120	0.133
V29	-0.008	0.055	-0.208	0.066	-0.011	0.110	0.075	0.108
V30	0.011	0.110	-0.221	0.133	0.028	0.092	0.109	0.118
V31	0.741	0.711	0.599	0.707	0.715	0.739	0.747	0.742
V32	0.785	0.772	0.449	0.792	0.797	0.806	0.785	0.864
V33	0.841	0.808	0.709	0.790	0.820	0.924	0.925	0.933
V34	0.813	0.769	0.529	0.793	0.818	0.843	0.843	0.795
V35	0.670	0.645	0.280	0.672	0.698	0.771	0.748	0.718
V36	0.520	0.419	0.173	0.460	0.482	0.495	0.501	0.473
V37	-0.020	-0.053	-0.250	0.013	0.000	0.025	0.018	0.029
V38	0.691	0.677	0.179	0.704	0.737	0.716	0.780	0.767
V39	0.635	0.423	0.312	0.550	0.557	0.529	0.576	0.544
V40	0.832	0.808	0.756	0.839	0.830	0.851	0.855	0.857
V41	0.797	0.817	0.405	0.835	0.760	0.800	0.812	0.803
V42	0.008	0.033	-0.350	0.050	0.005	0.071	0.073	0.074
V43	0.203	0.263	-0.276	0.331	0.552	0.321	0.404	0.542
V44	0.792	0.675	0.675	0.663	0.683	0.743	0.717	0.749
V45	0.381	0.254	0.164	0.305	0.284	0.339	0.347	0.358
V46	0.625	0.508	0.435	0.550	0.539	0.583	0.565	0.578
V47	0.966	0.892	0.735	0.913	0.908	0.972	0.970	0.972
V48	-0.088	-0.135	-0.152	-0.124	-0.081	-0.132	-0.116	-0.096
V49	0.887	0.833	0.593	0.848	0.835	0.867	0.884	0.892

V50	0.789	0.688	0.420	0.710	0.678	0.762	0.747	0.772
V51	0.748	0.720	0.436	0.747	0.719	0.751	0.748	0.767
V52	0.628	0.566	0.249	0.601	0.553	0.637	0.636	0.637
V53	0.632	0.619	0.243	0.639	0.612	0.664	0.660	0.673
V54	0.759	0.692	0.422	0.719	0.697	0.729	0.737	0.748
V55	0.613	0.534	0.216	0.577	0.525	0.592	0.595	0.602
V56	0.522	0.466	0.106	0.509	0.466	0.522	0.531	0.538
V57	0.564	0.580	0.249	0.622	0.600	0.628	0.641	0.656
V58	0.782	0.693	0.453	0.743	0.714	0.746	0.766	0.778
V59	0.558	0.562	0.203	0.603	0.574	0.595	0.622	0.617
V60	0.060	0.040	-0.256	-0.011	0.021	0.124	0.139	0.155
V61	0.000	-0.033	-0.252	-0.066	-0.014	0.003	0.017	0.032
V62	0.008	0.075	-0.179	0.058	0.009	0.013	0.035	0.037
V63	-0.009	0.005	-0.276	-0.023	-0.085	0.004	0.013	0.007
V64	0.408	0.343	0.186	0.347	0.416	0.439	0.427	0.481
V65	0.694	0.591	0.362	0.589	0.619	0.652	0.681	0.689
V66	0.135	0.111	-0.036	0.125	0.175	0.169	0.208	0.203
V67	0.145	-0.080	-0.293	-0.146	-0.080	0.183	0.187	0.199
V68	-0.043	-0.164	-0.316	-0.190	-0.178	-0.065	-0.081	-0.074
V69	0.198	0.081	-0.171	0.048	0.045	0.176	0.224	0.222
V70	0.647	0.538	0.436	0.576	0.552	0.613	0.614	0.614
V71	0.674	0.545	0.487	0.582	0.566	0.657	0.665	0.654
V72	0.183	0.077	-0.033	0.067	0.034	0.243	0.282	0.260
V73	0.615	0.499	0.236	0.518	0.485	0.560	0.555	0.545
V74	0.291	0.203	0.015	0.197	0.164	0.377	0.417	0.407
Ortalama	0.446	0.387	0.178	0.407	0.392	0.462	0.470	0.474

Ek 6. İki sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	4.909	4.636	7.909	5.000	5.091	2.818	2.636	3.000
V2	3.909	5.091	7.818	5.364	5.818	3.364	2.545	2.091
V3	4.591	5.182	7.636	5.091	5.909	3.045	1.955	2.591
V4	3.136	4.727	7.909	5.000	6.818	2.909	3.182	2.318
V5	4.455	5.273	7.727	5.636	5.909	2.636	2.091	2.273
V6	4.091	5.182	7.909	6.000	5.091	2.500	2.773	2.455
V7	3.636	5.818	6.727	4.909	5.545	3.455	2.818	3.091
V8	5.909	6.000	6.318	4.091	5.591	2.364	3.000	2.727
V9	4.364	5.636	6.727	4.455	5.000	2.909	3.091	3.818
V10	2.591	6.000	7.909	3.727	5.000	3.909	3.773	3.091
V11	2.545	5.545	7.909	5.182	6.545	3.273	2.818	2.182
V12	3.455	4.636	8.000	5.273	6.636	2.636	2.091	3.273
V13	2.818	5.727	7.727	5.636	6.545	2.455	2.364	2.727
V14	3.818	5.545	7.818	3.727	5.091	4.000	3.182	2.818
V15	4.136	5.455	7.818	4.000	4.636	3.682	3.545	2.727
V16	1.955	5.273	7.727	6.000	6.727	3.091	3.091	2.136
V17	3.591	5.909	7.909	3.727	4.545	3.773	4.045	2.500
V18	2.955	4.364	8.000	5.364	6.364	2.773	3.318	2.864
V19	3.818	5.091	7.909	4.091	4.727	3.591	3.500	3.273
V20	4.182	4.727	7.909	5.000	5.182	3.000	3.364	2.636
V21	1.864	5.545	7.273	6.273	6.727	2.318	2.955	3.045
V22	3.227	5.364	7.545	5.000	7.000	2.273	2.318	3.273

V23	3.727	5.364	7.545	4.182	5.545	4.455	2.636	2.545
V24	2.409	5.727	7.818	4.273	6.727	3.864	2.227	2.955
V25	3.682	5.364	7.909	5.000	6.273	3.136	2.227	2.409
V26	5.727	5.818	7.273	4.909	4.636	2.545	2.091	3.000
V27	3.909	6.545	7.000	5.273	5.364	2.182	2.455	3.273
V28	5.000	4.455	7.636	3.000	4.818	3.182	4.273	3.636
V29	5.182	4.455	6.909	4.000	5.091	3.318	3.864	3.182
V30	4.909	3.727	7.909	2.909	5.182	4.409	3.455	3.500
V31	3.455	4.773	7.091	5.273	4.591	3.864	3.500	3.455
V32	4.545	4.318	7.591	3.182	3.727	4.000	4.500	4.136
V33	4.455	4.273	6.318	5.364	3.909	3.682	4.136	3.864
V34	3.864	5.045	7.727	4.273	4.045	3.727	3.091	4.227
V35	5.227	4.500	7.545	4.909	4.318	2.591	3.545	3.364
V36	2.818	6.000	7.455	4.545	4.091	3.091	3.545	4.455
V37	5.091	5.364	8.000	3.636	4.455	2.909	3.545	3.000
V38	5.000	4.545	7.727	4.227	4.045	3.636	3.045	3.773
V39	2.136	6.136	7.545	4.091	4.545	4.500	3.045	4.000
V40	4.318	5.545	7.273	4.091	3.636	4.273	3.727	3.136
V41	4.000	3.864	7.909	3.727	5.364	4.045	3.318	3.773
V42	5.273	3.909	8.000	3.818	4.955	3.545	2.955	3.545
V43	5.318	5.409	7.773	4.318	3.136	3.773	3.318	2.955
V44	2.273	5.818	7.000	5.545	6.182	3.091	2.636	3.455
V45	3.091	6.000	7.636	3.909	4.727	3.818	3.545	3.273
V46	2.636	5.818	7.364	4.364	4.545	3.364	4.091	3.818
V47	4.227	4.955	7.091	4.455	4.455	3.500	3.818	3.500
V48	3.091	6.000	6.636	5.727	4.364	3.727	2.455	4.000
V49	3.545	5.000	7.909	4.273	5.364	3.000	3.182	3.727
V50	3.091	5.273	8.000	4.545	5.545	2.818	3.591	3.136
V51	4.364	4.636	8.000	3.727	4.636	3.182	4.273	3.182
V52	3.091	4.909	7.909	4.818	6.000	3.182	2.909	3.182
V53	4.364	4.273	8.000	4.182	5.273	3.636	3.364	2.909
V54	3.455	5.364	8.000	4.727	4.636	3.182	3.364	3.273
V55	3.909	4.909	8.000	3.636	5.909	3.273	3.636	2.727
V56	4.091	4.727	8.000	3.909	5.727	3.636	2.909	3.000
V57	4.727	4.727	8.000	3.909	4.455	3.909	3.455	2.818
V58	3.909	4.909	7.818	3.909	5.455	3.909	3.455	2.636
V59	4.273	4.727	8.000	3.909	4.727	3.909	3.273	3.182
V60	5.045	4.455	7.909	6.000	4.727	3.136	2.591	2.136
V61	4.182	4.818	7.909	5.455	4.182	3.545	3.182	2.727
V62	4.273	2.818	8.000	3.091	4.909	4.773	4.136	4.000
V63	3.682	2.364	8.000	4.636	6.182	3.500	3.455	4.182
V64	4.273	5.636	7.727	5.636	3.000	3.455	3.909	2.364
V65	2.545	5.636	7.727	5.636	4.727	3.864	3.455	2.409
V66	4.818	5.091	6.818	5.000	4.091	4.364	2.455	3.364
V67	3.364	5.455	7.545	6.636	6.000	2.636	2.273	2.091
V68	2.364	5.364	7.636	6.091	5.364	2.364	3.273	3.545
V69	3.364	5.364	8.000	5.545	5.909	3.636	2.091	2.091
V70	2.000	6.000	7.818	4.273	5.636	3.182	3.591	3.500
V71	2.273	6.273	7.545	5.364	5.909	3.000	2.545	3.091
V72	3.909	5.364	7.909	5.545	6.727	3.091	1.273	2.182
V73	2.364	5.000	8.000	4.273	5.182	3.500	3.682	4.000
V74	3.818	5.455	8.000	5.455	6.182	3.091	1.455	2.545
Ortalama	3.789	5.122	7.649	4.672	5.212	3.348	3.112	3.096

Ek 7. İki sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.070	0.035	-0.046	0.139	0.073	0.169	0.178	0.229
V2	-0.020	-0.123	-0.292	-0.087	-0.140	0.029	-0.042	0.056
V3	0.088	0.116	0.005	0.159	0.045	0.407	0.411	0.421
V4	-0.008	-0.218	-0.272	-0.162	-0.202	-0.014	-0.026	-0.006
V5	-0.030	-0.111	-0.140	-0.069	-0.087	0.107	0.152	0.134
V6	0.295	0.121	0.036	0.159	0.162	0.271	0.249	0.289
V7	0.404	0.473	0.447	0.524	0.496	0.582	0.562	0.511
V8	0.138	0.122	0.136	0.265	0.119	0.307	0.308	0.298
V9	0.469	0.472	0.407	0.471	0.419	0.554	0.556	0.556
V10	0.727	0.712	0.379	0.712	0.711	0.689	0.707	0.705
V11	0.522	0.471	0.095	0.463	0.424	0.519	0.514	0.537
V12	0.621	0.624	0.253	0.620	0.549	0.634	0.638	0.617
V13	0.620	0.562	0.130	0.513	0.549	0.617	0.629	0.624
V14	0.735	0.731	0.424	0.745	0.710	0.723	0.756	0.741
V15	0.724	0.698	0.469	0.689	0.719	0.732	0.718	0.734
V16	0.643	0.624	0.251	0.580	0.549	0.616	0.636	0.637
V17	0.678	0.697	0.448	0.676	0.667	0.695	0.690	0.713
V18	0.597	0.587	0.331	0.585	0.552	0.616	0.581	0.606
V19	0.729	0.734	0.568	0.672	0.709	0.709	0.706	0.727
V20	0.704	0.720	0.472	0.699	0.686	0.733	0.727	0.728
V21	0.638	0.599	0.340	0.584	0.577	0.650	0.619	0.640
V22	0.721	0.679	0.302	0.681	0.649	0.730	0.713	0.690
V23	0.669	0.658	0.521	0.648	0.650	0.666	0.674	0.680
V24	0.513	0.396	0.284	0.440	0.380	0.462	0.501	0.450
V25	0.734	0.720	0.474	0.707	0.640	0.735	0.745	0.724
V26	-0.053	-0.010	-0.046	0.008	0.006	0.018	0.059	0.020
V27	0.060	0.075	0.080	0.101	0.083	0.125	0.145	0.116
V28	0.225	0.147	-0.031	0.229	0.155	0.228	0.144	0.173
V29	0.209	0.123	0.197	0.220	0.175	0.221	0.140	0.156
V30	0.292	0.294	0.025	0.266	0.246	0.200	0.265	0.238
V31	0.800	0.793	0.662	0.766	0.787	0.774	0.785	0.754
V32	0.916	0.949	0.515	0.966	0.937	0.949	0.949	0.949
V33	0.956	0.964	0.892	0.901	0.964	0.964	0.956	0.964
V34	0.961	0.950	0.656	0.944	0.957	0.946	0.957	0.934
V35	0.893	0.936	0.353	0.932	0.914	0.923	0.914	0.931
V36	0.677	0.523	0.262	0.551	0.522	0.582	0.612	0.560
V37	-0.018	0.011	-0.141	0.028	0.011	0.100	0.093	0.102
V38	0.831	0.969	0.118	1.000	0.940	0.980	0.962	0.936
V39	0.756	0.487	0.265	0.665	0.444	0.534	0.634	0.545
V40	0.873	0.865	0.835	0.875	0.885	0.876	0.865	0.880
V41	0.937	0.944	0.622	0.955	0.896	0.973	0.959	0.959
V42	0.064	0.172	-0.278	0.188	0.039	0.122	0.089	0.099
V43	0.686	0.868	-0.130	0.915	0.932	0.905	0.838	0.941
V44	0.834	0.575	0.732	0.540	0.646	0.631	0.478	0.681
V45	0.443	0.232	0.229	0.317	0.263	0.260	0.309	0.341
V46	0.740	0.587	0.506	0.605	0.580	0.581	0.601	0.555
V47	1.000	1.000	0.833	1.000	1.000	1.000	1.000	1.000
V48	-0.061	-0.111	-0.123	-0.115	-0.106	-0.097	-0.066	-0.088
V49	0.981	0.987	0.634	0.990	0.983	0.991	0.985	0.977

V50	0.921	0.893	0.588	0.902	0.886	0.905	0.897	0.896
V51	0.884	0.923	0.615	0.921	0.925	0.927	0.913	0.920
V52	0.780	0.727	0.417	0.723	0.653	0.747	0.795	0.731
V53	0.850	0.885	0.433	0.891	0.860	0.888	0.874	0.880
V54	0.886	0.912	0.645	0.901	0.908	0.906	0.906	0.884
V55	0.770	0.816	0.346	0.825	0.787	0.825	0.794	0.802
V56	0.709	0.774	0.280	0.775	0.744	0.765	0.758	0.758
V57	0.764	0.865	0.432	0.863	0.853	0.844	0.829	0.846
V58	0.886	0.908	0.529	0.919	0.894	0.908	0.906	0.912
V59	0.791	0.821	0.332	0.801	0.831	0.821	0.790	0.808
V60	0.008	0.095	-0.301	0.030	0.022	0.222	0.153	0.143
V61	-0.004	-0.042	-0.268	-0.040	-0.007	0.030	0.007	-0.011
V62	-0.028	0.094	-0.130	0.092	0.042	0.045	0.011	0.069
V63	0.004	0.039	-0.239	0.005	-0.054	0.026	0.025	0.018
V64	0.452	0.328	0.184	0.308	0.442	0.494	0.448	0.477
V65	0.690	0.648	0.397	0.594	0.565	0.693	0.695	0.683
V66	0.194	0.083	-0.104	0.079	0.260	0.189	0.223	0.244
V67	0.054	-0.085	-0.314	-0.141	-0.160	0.199	0.204	0.248
V68	-0.043	-0.217	-0.341	-0.228	-0.225	-0.038	-0.076	-0.053
V69	0.195	0.164	-0.209	0.084	0.034	0.216	0.226	0.204
V70	0.639	0.613	0.368	0.594	0.592	0.642	0.626	0.628
V71	0.694	0.524	0.442	0.595	0.527	0.684	0.692	0.642
V72	0.129	0.050	-0.070	0.055	0.002	0.274	0.307	0.292
V73	0.671	0.589	0.285	0.582	0.546	0.582	0.624	0.575
V74	0.221	0.252	-0.022	0.185	0.168	0.426	0.489	0.454
Ortalama	0.515	0.488	0.243	0.494	0.472	0.540	0.536	0.538

Ek 8. İki sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	6.000	6.500	8.000	3.500	5.500	2.500	3.000	1.000
V2	3.500	6.000	8.000	5.000	7.000	2.000	3.500	1.000
V3	6.000	5.500	7.500	4.500	6.500	2.000	2.000	2.000
V4	2.000	7.000	7.500	5.000	6.500	2.500	3.500	2.000
V5	4.000	6.500	8.000	5.000	6.500	3.000	1.500	1.500
V6	2.000	6.500	8.000	5.500	6.000	2.500	3.500	2.000
V7	5.500	5.500	6.500	4.000	5.000	1.500	2.500	5.500
V8	5.500	7.000	5.500	3.500	7.000	1.500	2.500	3.500
V9	6.000	5.000	7.000	5.000	7.000	2.000	2.500	1.500
V10	2.000	3.500	8.000	3.500	4.000	6.500	5.000	3.500
V11	1.500	5.000	8.000	6.000	7.000	2.500	4.000	2.000
V12	4.000	3.500	8.000	5.000	7.000	2.500	1.500	4.500
V13	2.500	5.500	8.000	6.500	6.000	2.500	2.500	2.500
V14	4.500	4.500	8.000	2.500	6.000	6.500	1.000	3.000
V15	3.500	5.500	8.000	5.500	4.000	3.000	4.000	2.500
V16	1.000	3.500	8.000	6.000	7.000	4.500	3.500	2.500
V17	5.000	3.000	8.000	5.500	5.500	4.000	4.000	1.000
V18	3.500	3.500	8.000	4.000	5.500	2.000	5.500	4.000
V19	2.500	2.500	8.000	6.500	4.000	4.500	5.500	2.500
V20	5.000	4.000	8.000	5.500	6.500	2.500	2.000	2.500

V21	2.000	5.000	8.000	6.500	6.500	1.500	4.000	2.500
V22	2.500	5.500	8.000	5.000	7.000	1.250	2.250	4.500
V23	4.000	5.000	8.000	5.000	6.000	4.000	2.000	2.000
V24	1.500	6.000	8.000	4.500	7.000	4.000	1.500	3.500
V25	3.500	4.500	8.000	6.000	7.000	3.000	1.000	3.000
V26	7.000	5.500	7.500	3.500	3.500	4.000	1.000	4.000
V27	4.500	7.500	6.000	4.000	6.000	3.000	1.500	3.500
V28	2.000	6.000	8.000	2.000	5.500	2.000	6.000	4.500
V29	3.500	6.000	4.500	4.000	4.000	2.500	6.000	5.500
V30	2.000	2.000	8.000	3.000	5.500	6.500	3.500	5.500
V31	3.000	3.000	8.000	4.250	4.000	4.500	4.500	4.750
V32	6.250	3.750	8.000	2.000	5.250	4.000	4.000	2.750
V33	5.000	3.000	7.000	7.500	3.000	3.000	4.500	3.000
V34	2.250	4.000	8.000	5.000	2.500	5.000	3.250	6.000
V35	6.500	1.500	8.000	3.000	5.000	4.500	4.500	3.000
V36	1.000	5.500	8.000	5.000	6.500	2.500	2.500	5.000
V37	7.000	5.500	8.000	4.000	5.500	2.000	2.500	1.500
V38	7.000	3.500	8.000	2.250	4.000	3.000	3.750	4.500
V39	1.000	6.000	8.000	2.500	7.000	4.750	2.500	4.250
V40	4.500	6.500	8.000	3.500	1.000	4.500	6.000	2.000
V41	4.250	3.750	8.000	3.250	7.000	2.250	3.750	3.750
V42	5.500	1.500	8.000	2.000	6.000	4.000	4.500	4.500
V43	7.000	4.000	8.000	2.500	3.750	2.500	5.000	3.250
V44	1.000	6.500	5.000	6.500	4.500	4.000	5.000	3.500
V45	1.000	6.500	6.500	3.500	5.000	6.000	4.500	3.000
V46	1.000	3.500	7.000	3.500	5.000	5.500	4.500	6.000
V47	4.000	4.000	8.000	4.000	4.000	4.000	4.000	4.000
V48	1.500	5.500	8.000	6.500	6.000	3.500	1.500	3.500
V49	4.000	4.000	8.000	2.500	4.500	1.500	5.000	6.500
V50	3.500	5.000	8.000	3.000	5.000	2.500	4.500	4.500
V51	7.000	3.000	8.000	3.000	2.500	2.500	5.500	4.500
V52	2.500	3.500	8.000	4.500	7.000	4.500	1.500	4.500
V53	7.000	4.000	8.000	1.500	6.000	3.000	3.000	3.500
V54	6.500	2.500	8.000	4.500	2.500	2.500	3.000	6.500
V55	6.000	3.000	8.000	1.500	5.500	2.000	6.000	4.000
V56	7.000	2.500	8.000	1.500	5.000	4.000	3.500	4.500
V57	7.000	3.000	8.000	2.000	4.000	3.500	5.000	3.500
V58	6.500	4.000	8.000	1.000	4.500	4.000	4.500	3.500
V59	6.500	3.000	8.000	4.500	1.000	3.000	6.500	3.500
V60	6.500	3.500	8.000	6.000	5.500	1.000	2.500	3.000
V61	4.000	6.500	8.000	5.000	4.000	1.500	3.000	4.000
V62	6.500	2.000	8.000	2.000	5.000	3.500	5.500	3.500
V63	4.000	1.000	8.000	5.500	7.000	3.000	4.000	3.500
V64	4.000	6.000	8.000	7.000	3.000	1.500	4.000	2.500
V65	3.000	4.500	8.000	5.500	7.000	2.000	3.000	3.000
V66	3.500	6.000	8.000	7.000	2.500	4.000	2.500	2.500
V67	4.000	5.000	8.000	6.000	7.000	2.500	2.500	1.000
V68	2.500	6.000	8.000	6.000	6.000	2.000	3.000	2.500

V69	3.500	4.500	8.000	6.000	7.000	2.000	2.000	3.000
V70	3.000	4.000	8.000	5.000	5.500	3.000	3.500	4.000
V71	2.000	6.500	8.000	5.000	6.500	2.500	1.500	4.000
V72	4.500	5.500	8.000	5.000	7.000	3.000	1.000	2.000
V73	1.500	5.000	8.000	4.000	5.500	4.000	3.000	5.000
V74	5.000	4.500	8.000	6.000	6.000	2.500	1.000	3.000
Ortalama	4.024	4.568	7.750	4.355	5.331	3.145	3.412	3.416

Ek 9. İki sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.057	0.134	-0.144	0.127	0.146	0.189	0.249	0.122
V2	-0.020	-0.019	-0.321	-0.039	-0.060	-0.001	0.018	0.056
V3	0.083	0.125	-0.190	0.106	0.134	0.130	0.291	0.195
V4	-0.008	-0.025	-0.363	-0.044	-0.098	-0.002	0.005	0.031
V5	-0.009	0.020	-0.160	0.031	0.015	0.124	0.168	0.158
V6	0.165	0.192	-0.140	0.167	0.168	0.242	0.274	0.246
V7	0.643	0.565	0.352	0.494	0.436	0.666	0.671	0.675
V8	0.050	0.083	0.022	0.198	0.083	0.310	0.294	0.294
V9	0.463	0.401	0.312	0.483	0.447	0.504	0.495	0.447
V10	0.601	0.578	0.369	0.626	0.588	0.605	0.586	0.614
V11	0.474	0.464	0.081	0.428	0.397	0.490	0.489	0.535
V12	0.541	0.562	0.146	0.533	0.465	0.565	0.578	0.574
V13	0.605	0.530	0.297	0.545	0.467	0.641	0.631	0.624
V14	0.618	0.629	0.383	0.657	0.632	0.600	0.613	0.638
V15	0.596	0.620	0.430	0.614	0.617	0.614	0.614	0.624
V16	0.661	0.568	0.326	0.512	0.482	0.654	0.653	0.665
V17	0.589	0.568	0.407	0.608	0.596	0.594	0.582	0.604
V18	0.535	0.552	0.215	0.483	0.467	0.535	0.533	0.536
V19	0.592	0.616	0.427	0.639	0.604	0.611	0.610	0.613
V20	0.591	0.618	0.410	0.599	0.585	0.606	0.600	0.612
V21	0.669	0.570	0.374	0.509	0.498	0.671	0.666	0.665
V22	0.680	0.636	0.444	0.598	0.525	0.666	0.670	0.664
V23	0.639	0.667	0.495	0.664	0.646	0.664	0.690	0.682
V24	0.449	0.419	0.231	0.442	0.387	0.422	0.458	0.431
V25	0.649	0.677	0.327	0.669	0.644	0.680	0.690	0.687
V26	0.022	0.110	-0.002	0.083	0.103	0.162	0.150	0.140
V27	0.148	0.150	0.073	0.158	0.160	0.215	0.199	0.200
V28	0.004	0.130	-0.210	0.181	0.177	0.148	0.129	0.139
V29	0.002	0.127	-0.256	0.094	0.068	0.128	0.118	0.159
V30	0.039	0.127	-0.215	0.185	0.085	0.136	0.127	0.150
V31	0.763	0.740	0.603	0.707	0.718	0.780	0.797	0.785
V32	0.992	0.985	0.717	0.992	0.992	0.992	0.985	0.985
V33	0.983	0.975	0.931	0.961	0.975	0.978	0.974	0.978
V34	0.925	0.886	0.582	0.890	0.918	0.943	0.906	0.905
V35	0.875	0.798	0.350	0.821	0.826	0.910	0.832	0.853
V36	0.675	0.586	0.236	0.628	0.682	0.658	0.641	0.621
V37	0.007	0.001	-0.252	0.075	0.049	0.093	0.093	0.097
V38	0.831	0.781	0.221	0.768	0.839	0.851	0.921	0.919
V39	0.696	0.456	0.312	0.650	0.704	0.598	0.644	0.625
V40	0.804	0.851	0.716	0.849	0.834	0.828	0.826	0.825
V41	0.897	0.951	0.453	0.937	0.839	0.880	0.897	0.879

V42	-0.012	0.069	-0.341	0.064	0.064	0.092	0.067	0.112
V43	-0.004	0.262	-0.348	0.397	0.733	0.299	0.476	0.652
V44	0.820	0.750	0.692	0.716	0.721	0.809	0.804	0.802
V45	0.445	0.386	0.202	0.409	0.398	0.406	0.384	0.412
V46	0.737	0.710	0.544	0.729	0.716	0.731	0.720	0.734
V47	0.990	0.996	0.762	1.000	1.000	1.000	0.995	1.000
V48	-0.061	-0.085	-0.090	-0.076	-0.059	-0.055	-0.040	-0.061
V49	0.890	0.859	0.581	0.860	0.840	0.885	0.881	0.873
V50	0.756	0.685	0.310	0.695	0.654	0.762	0.716	0.744
V51	0.680	0.735	0.313	0.719	0.684	0.764	0.711	0.760
V52	0.587	0.613	0.150	0.627	0.570	0.665	0.647	0.662
V53	0.520	0.646	0.137	0.630	0.633	0.674	0.672	0.696
V54	0.678	0.688	0.303	0.686	0.671	0.734	0.719	0.745
V55	0.520	0.560	0.163	0.569	0.502	0.635	0.584	0.608
V56	0.382	0.500	0.057	0.505	0.456	0.539	0.502	0.535
V57	0.440	0.634	0.192	0.638	0.635	0.661	0.657	0.684
V58	0.740	0.739	0.404	0.745	0.724	0.771	0.764	0.780
V59	0.453	0.587	0.090	0.571	0.536	0.612	0.593	0.617
V60	0.004	0.064	-0.304	0.026	0.015	0.051	0.131	0.171
V61	-0.028	-0.008	-0.296	-0.021	0.014	0.008	0.055	0.098
V62	0.027	0.144	-0.186	0.083	0.038	0.003	0.058	0.034
V63	-0.021	0.061	-0.282	0.024	-0.057	0.002	0.040	0.019
V64	0.453	0.324	0.207	0.315	0.384	0.418	0.427	0.503
V65	0.712	0.610	0.457	0.606	0.677	0.649	0.735	0.724
V66	0.153	0.148	-0.106	0.138	0.194	0.204	0.252	0.206
V67	0.155	0.056	-0.206	0.001	0.094	0.240	0.252	0.213
V68	-0.043	-0.073	-0.294	-0.097	-0.080	-0.033	-0.043	-0.057
V69	0.210	0.204	-0.127	0.141	0.131	0.214	0.283	0.277
V70	0.614	0.556	0.484	0.591	0.540	0.615	0.594	0.577
V71	0.706	0.656	0.519	0.637	0.650	0.703	0.692	0.685
V72	0.196	0.143	-0.039	0.099	0.062	0.265	0.325	0.298
V73	0.659	0.566	0.244	0.578	0.554	0.619	0.575	0.590
V74	0.306	0.256	0.012	0.251	0.227	0.419	0.427	0.444
Ortalama	0.449	0.452	0.178	0.452	0.443	0.492	0.500	0.505

Ek 10. İki sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	6.000	4.250	8.000	4.750	4.500	2.750	1.500	4.250
V2	4.750	4.000	8.000	4.500	5.000	5.000	2.750	2.000
V3	5.250	5.250	8.000	4.750	4.500	4.250	1.500	2.500
V4	3.500	3.750	8.000	4.500	7.000	3.875	3.125	2.250
V5	5.750	4.500	8.000	5.250	5.750	2.750	2.000	2.000
V6	5.250	4.500	8.000	5.750	5.000	2.500	2.000	3.000
V7	3.500	4.500	8.000	6.000	7.000	3.000	2.250	1.750
V8	6.750	5.500	7.000	4.250	5.750	1.500	3.000	2.250
V9	4.500	5.750	7.250	3.500	4.875	2.500	3.000	4.625
V10	4.500	5.750	8.000	2.000	4.500	3.750	4.500	3.000
V11	4.750	4.750	8.000	4.500	6.500	3.500	2.750	1.250
V12	4.500	3.250	8.000	5.250	6.750	2.750	2.750	2.750
V13	4.250	4.750	8.000	5.250	7.000	2.250	2.000	2.500
V14	5.000	3.750	8.000	2.250	4.000	4.750	5.250	3.000

V15	5.500	3.250	7.750	3.750	4.500	4.500	3.750	3.000
V16	2.000	5.000	7.500	6.500	7.000	2.750	3.250	2.000
V17	4.000	6.000	8.000	2.750	3.000	4.250	5.000	3.000
V18	3.250	2.500	8.000	5.750	7.000	3.250	3.500	2.750
V19	6.000	4.250	8.000	1.750	4.000	4.125	3.875	4.000
V20	5.625	2.500	8.000	5.250	4.500	3.125	4.375	2.625
V21	2.250	5.000	7.250	6.500	7.000	2.000	2.750	3.250
V22	3.500	4.000	7.250	5.750	7.000	2.500	2.750	3.250
V23	5.000	4.500	8.000	4.250	5.750	4.500	2.000	2.000
V24	2.750	5.000	8.000	3.000	6.250	5.000	2.500	3.500
V25	5.250	4.250	8.000	4.250	5.750	3.750	2.500	2.250
V26	7.000	4.750	7.750	5.500	4.500	1.500	1.750	3.250
V27	4.750	5.500	7.750	6.000	4.250	1.250	3.000	3.500
V28	6.000	3.500	8.000	3.000	3.250	3.750	4.750	3.750
V29	5.750	3.500	8.000	4.000	5.000	4.000	3.250	2.500
V30	5.250	3.750	8.000	2.750	4.250	4.250	4.500	3.250
V31	3.500	4.375	7.500	6.500	6.000	3.375	1.875	2.875
V32	3.750	4.500	7.125	3.750	3.750	3.750	4.875	4.500
V33	3.250	4.625	3.875	5.750	4.875	4.375	4.875	4.375
V34	5.000	4.250	8.000	3.500	4.500	3.000	4.000	3.750
V35	4.375	5.125	8.000	4.750	4.875	1.375	4.500	3.000
V36	3.250	6.750	8.000	4.500	2.000	3.000	4.000	4.500
V37	6.000	5.500	8.000	3.750	5.250	2.500	2.750	2.250
V38	4.875	4.500	8.000	5.125	4.125	3.500	3.000	2.875
V39	2.625	6.750	8.000	4.250	2.125	4.875	3.500	3.875
V40	4.750	3.000	7.750	3.250	3.500	4.750	4.250	4.750
V41	3.625	2.500	8.000	3.625	5.750	5.000	3.000	4.500
V42	6.125	4.250	8.000	3.750	4.500	3.000	3.625	2.750
V43	6.500	5.750	8.000	4.000	1.750	5.000	2.250	2.750
V44	3.000	5.000	7.500	6.000	6.500	2.500	2.250	3.250
V45	3.750	4.250	8.000	3.000	3.500	3.750	5.500	4.250
V46	4.000	5.500	8.000	3.750	4.000	3.250	4.500	3.000
V47	4.875	4.375	7.125	3.750	3.750	3.750	4.625	3.750
V48	3.750	6.750	7.250	5.750	3.500	3.000	1.500	4.500
V49	3.000	4.250	8.000	4.000	5.500	3.250	3.250	4.750
V50	2.750	5.000	8.000	4.000	6.500	1.500	4.500	3.750
V51	5.250	3.750	8.000	4.500	5.250	1.500	5.250	2.500
V52	4.000	4.750	8.000	4.250	6.000	2.750	3.500	2.750
V53	5.750	4.000	8.000	5.000	4.250	3.500	3.500	2.000
V54	4.000	5.000	8.000	5.000	5.250	2.750	4.250	1.750
V55	5.000	4.250	8.000	3.750	6.250	2.000	4.500	2.250
V56	5.250	4.000	8.000	3.750	5.750	2.500	4.250	2.500
V57	7.000	4.500	8.000	4.000	3.750	3.250	3.250	2.250
V58	5.500	4.000	8.000	3.750	5.000	3.500	3.750	2.500
V59	5.500	4.000	8.000	4.500	5.500	2.500	3.500	2.500
V60	6.000	4.000	8.000	4.750	4.750	4.750	2.500	1.250
V61	5.000	4.750	8.000	4.750	4.000	5.000	2.750	1.750
V62	4.000	2.000	8.000	2.750	5.000	6.250	3.500	4.500
V63	4.625	2.250	8.000	4.000	5.750	4.500	2.375	4.500
V64	3.500	6.750	7.750	6.250	3.250	3.500	3.000	2.000
V65	2.750	6.000	8.000	5.250	3.750	5.500	2.250	2.500
V66	5.250	5.250	8.000	5.250	4.250	3.250	1.500	3.250
V67	4.500	5.250	8.000	6.750	5.000	2.500	1.250	2.750
V68	3.000	4.500	7.750	5.500	4.000	2.750	3.250	5.250
V69	4.250	4.750	8.000	5.500	5.750	4.500	1.750	1.500

V70	2.250	6.250	8.000	3.500	5.500	2.000	4.000	4.500
V71	2.250	4.750	8.000	5.750	5.500	2.500	3.250	4.000
V72	4.750	4.500	8.000	6.000	6.500	3.250	1.250	1.750
V73	2.500	4.250	8.000	4.750	4.500	3.000	4.500	4.500
V74	4.500	5.500	8.000	5.500	5.500	2.750	1.750	2.500
Ortalama	4.476	4.561	7.826	4.524	4.965	3.356	3.231	3.061

Ek 11. İki sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.125	-0.092	-0.161	-0.125	-0.103	0.022	-0.003	-0.004
V2	-0.020	-0.232	-0.373	-0.253	-0.248	-0.063	-0.067	-0.063
V3	0.173	-0.031	-0.176	-0.090	-0.082	0.177	0.193	0.195
V4	-0.016	-0.222	-0.357	-0.272	-0.305	-0.034	-0.034	-0.034
V5	0.052	-0.174	-0.204	-0.191	-0.183	0.004	-0.039	-0.031
V6	0.219	0.092	-0.049	0.045	0.091	0.198	0.188	0.161
V7	-0.075	-0.196	-0.194	-0.185	-0.185	-0.209	-0.205	-0.204
V8	-0.040	-0.027	-0.041	-0.019	-0.035	0.018	0.050	0.085
V9	0.217	0.014	0.011	0.057	0.038	0.143	0.105	0.103
V10	0.634	0.462	0.407	0.601	0.513	0.661	0.658	0.660
V11	0.450	0.241	0.119	0.306	0.205	0.378	0.404	0.375
V12	0.474	0.320	0.215	0.384	0.278	0.489	0.503	0.490
V13	0.620	0.450	0.310	0.539	0.478	0.634	0.634	0.630
V14	0.714	0.562	0.481	0.665	0.643	0.722	0.718	0.715
V15	0.654	0.555	0.476	0.674	0.637	0.694	0.693	0.696
V16	0.519	0.234	0.144	0.271	0.200	0.496	0.509	0.489
V17	0.605	0.511	0.396	0.612	0.566	0.643	0.643	0.647
V18	0.487	0.353	0.257	0.446	0.353	0.512	0.513	0.511
V19	0.661	0.545	0.487	0.671	0.637	0.688	0.692	0.689
V20	0.627	0.537	0.473	0.623	0.585	0.668	0.667	0.669
V21	0.583	0.278	0.196	0.310	0.234	0.537	0.544	0.536
V22	0.619	0.363	0.222	0.446	0.318	0.658	0.655	0.656
V23	0.649	0.549	0.562	0.629	0.605	0.655	0.661	0.661
V24	0.391	0.303	0.282	0.334	0.284	0.419	0.416	0.417
V25	0.693	0.457	0.429	0.604	0.549	0.693	0.700	0.693
V26	0.036	-0.244	-0.265	-0.206	-0.216	-0.040	-0.052	-0.002
V27	-0.028	-0.191	-0.229	-0.185	-0.160	0.002	0.034	0.006
V28	-0.067	-0.064	-0.207	-0.052	-0.225	0.015	0.007	-0.003
V29	-0.161	-0.110	-0.320	-0.130	-0.279	-0.076	-0.083	-0.081
V30	-0.174	-0.098	-0.317	-0.117	-0.276	-0.092	-0.063	-0.076
V31	0.698	0.643	0.608	0.644	0.655	0.700	0.698	0.693
V32	0.381	0.219	-0.070	0.228	0.471	0.561	0.452	0.759
V33	0.605	0.505	0.322	0.512	0.551	0.822	0.822	0.853
V34	0.595	0.478	0.431	0.547	0.576	0.618	0.648	0.528
V35	0.254	0.165	0.149	0.208	0.258	0.391	0.427	0.299
V36	0.254	0.109	0.099	0.160	0.167	0.194	0.220	0.181
V37	-0.090	-0.205	-0.317	-0.184	-0.150	-0.174	-0.182	-0.162
V38	0.429	0.233	0.141	0.286	0.333	0.306	0.414	0.399
V39	0.503	0.257	0.217	0.316	0.399	0.422	0.425	0.430
V40	0.718	0.697	0.699	0.762	0.753	0.841	0.844	0.846
V41	0.557	0.499	0.235	0.516	0.441	0.552	0.542	0.560
V42	-0.012	-0.283	-0.458	-0.183	-0.233	-0.082	-0.069	-0.090

V43	-0.017	-0.246	-0.373	-0.236	0.030	-0.205	-0.254	-0.009
V44	0.692	0.693	0.666	0.701	0.689	0.717	0.723	0.713
V45	0.062	-0.050	-0.063	0.004	-0.022	0.078	0.062	0.067
V46	0.227	0.060	0.042	0.121	0.125	0.174	0.110	0.165
V47	0.860	0.661	0.543	0.707	0.704	0.917	0.922	0.907
V48	-0.136	-0.197	-0.306	-0.203	-0.100	-0.253	-0.245	-0.149
V49	0.709	0.571	0.503	0.593	0.584	0.643	0.684	0.720
V50	0.578	0.397	0.321	0.407	0.364	0.491	0.486	0.530
V51	0.486	0.382	0.332	0.444	0.401	0.429	0.481	0.476
V52	0.294	0.223	0.143	0.238	0.195	0.286	0.271	0.280
V53	0.303	0.213	0.099	0.228	0.170	0.244	0.237	0.247
V54	0.547	0.365	0.311	0.423	0.374	0.393	0.431	0.438
V55	0.353	0.159	0.085	0.200	0.145	0.173	0.228	0.208
V56	0.282	0.082	-0.034	0.109	0.050	0.117	0.149	0.131
V57	0.257	0.126	0.047	0.169	0.118	0.160	0.195	0.188
V58	0.561	0.367	0.374	0.445	0.406	0.433	0.481	0.491
V59	0.243	0.232	0.162	0.295	0.234	0.226	0.299	0.259
V60	0.168	-0.069	-0.229	-0.107	-0.051	0.149	0.154	0.153
V61	0.047	-0.117	-0.231	-0.184	-0.112	0.000	-0.002	0.008
V62	0.015	-0.042	-0.158	-0.043	-0.079	0.028	0.064	0.066
V63	-0.001	-0.124	-0.248	-0.131	-0.179	-0.004	-0.016	-0.014
V64	0.285	0.403	0.148	0.418	0.439	0.406	0.394	0.425
V65	0.622	0.550	0.221	0.559	0.574	0.595	0.591	0.624
V66	0.103	0.078	0.067	0.100	0.080	0.083	0.113	0.136
V67	0.131	-0.250	-0.354	-0.340	-0.209	0.095	0.097	0.125
V68	-0.043	-0.254	-0.321	-0.272	-0.257	-0.143	-0.148	-0.140
V69	0.209	-0.095	-0.203	-0.099	-0.062	0.104	0.122	0.129
V70	0.631	0.468	0.415	0.509	0.496	0.571	0.596	0.604
V71	0.569	0.428	0.468	0.485	0.478	0.589	0.601	0.597
V72	0.178	0.006	-0.006	0.022	0.014	0.173	0.187	0.178
V73	0.475	0.277	0.195	0.317	0.276	0.395	0.398	0.389
V74	0.291	0.094	0.020	0.141	0.084	0.265	0.320	0.301
Ortalama	0.331	0.187	0.098	0.219	0.204	0.312	0.319	0.326

Ek 12. İki sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	1.250	5.250	7.500	6.750	6.000	2.750	3.750	2.750
V2	2.500	6.250	7.500	6.250	6.000	2.500	2.250	2.750
V3	3.000	5.500	7.250	6.250	6.750	3.000	2.250	2.000
V4	2.125	6.000	8.000	5.250	6.750	2.375	3.125	2.375
V5	1.250	6.000	7.250	6.750	6.000	2.500	3.000	3.250
V6	2.500	5.750	7.750	7.000	5.500	2.625	2.375	2.500
V7	2.750	5.500	5.000	3.750	3.500	6.000	5.000	4.500
V8	6.250	5.500	6.125	4.500	5.625	3.250	2.500	2.250
V9	4.250	6.250	6.000	5.000	4.875	2.500	3.750	3.375
V10	2.125	7.000	7.750	5.000	6.250	2.375	2.875	2.625
V11	1.750	5.750	8.000	5.250	7.000	2.500	2.250	3.500
V12	2.750	6.000	8.000	5.000	7.000	2.500	1.500	3.250
V13	2.500	6.250	8.000	5.000	6.750	2.250	2.250	3.000
V14	3.000	7.250	7.750	5.250	5.750	1.875	2.500	2.625
V15	4.000	7.250	7.750	2.750	5.000	3.500	3.125	2.625

V16	1.875	6.250	7.500	5.250	7.000	2.750	2.250	3.125
V17	3.375	7.000	7.750	3.750	6.250	2.875	2.625	2.375
V18	3.125	6.250	8.000	4.750	6.250	2.625	2.375	2.625
V19	3.875	7.000	7.750	4.000	5.500	2.875	2.375	2.625
V20	3.625	6.250	7.750	5.000	5.500	2.625	2.875	2.375
V21	1.375	6.250	7.250	5.500	7.000	2.625	2.875	3.125
V22	3.375	6.000	8.000	5.000	7.000	1.750	2.750	2.125
V23	2.750	6.000	7.250	4.250	6.500	3.250	2.750	3.250
V24	2.875	6.000	7.250	5.250	7.500	2.125	2.625	2.375
V25	2.625	7.000	7.750	5.000	6.250	2.875	2.125	2.375
V26	3.000	7.250	6.750	5.500	5.000	2.750	3.250	2.500
V27	2.250	7.000	6.750	6.000	5.250	2.750	3.000	3.000
V28	4.750	4.750	7.250	4.000	7.000	2.500	2.750	3.000
V29	5.250	4.750	6.750	4.250	7.000	2.875	2.625	2.500
V30	5.500	4.750	7.750	3.500	6.750	3.625	1.500	2.625
V31	3.000	5.375	6.500	6.875	5.250	2.625	3.000	3.375
V32	3.875	5.250	8.000	4.750	2.750	3.125	4.875	3.375
V33	4.125	5.375	8.000	4.750	3.375	3.375	3.625	3.375
V34	3.000	5.750	7.250	3.875	4.375	4.000	2.750	5.000
V35	3.750	5.250	6.750	6.250	4.750	3.250	2.250	3.750
V36	3.750	5.500	6.500	3.750	4.750	4.000	3.000	4.750
V37	2.250	5.500	8.000	3.750	4.000	3.500	4.500	4.500
V38	2.875	5.500	7.250	4.625	5.000	3.875	2.750	4.125
V39	2.500	5.500	7.500	5.000	5.500	3.375	3.125	3.500
V40	4.375	7.250	6.750	5.250	5.750	2.750	2.250	1.625
V41	3.000	5.250	7.750	4.750	5.500	3.500	3.500	2.750
V42	2.375	6.750	8.000	5.250	5.625	2.750	1.750	3.500
V43	2.875	5.125	7.375	5.125	2.750	3.625	5.625	3.500
V44	2.750	5.500	7.500	4.000	6.750	3.250	2.250	4.000
V45	3.750	6.250	6.750	4.000	5.500	3.250	2.750	3.750
V46	2.500	6.250	7.500	4.000	4.500	3.250	4.250	3.750
V47	3.625	6.500	7.250	6.000	5.500	2.375	2.125	2.625
V48	3.750	4.250	6.500	4.750	4.500	3.500	5.250	3.500
V49	2.500	6.250	7.750	5.000	5.750	3.500	3.000	2.250
V50	1.750	5.500	8.000	5.750	6.250	3.000	3.000	2.750
V51	3.500	6.250	7.750	3.500	5.750	4.750	2.000	2.500
V52	3.000	5.500	7.500	5.000	6.250	2.250	3.000	3.500
V53	2.750	4.750	7.750	4.500	6.750	3.000	3.750	2.750
V54	1.250	6.750	8.000	4.250	5.750	4.250	2.750	3.000
V55	1.500	5.750	8.000	4.000	6.750	4.500	2.000	3.500
V56	1.000	6.000	8.000	4.750	6.750	4.000	2.000	3.500
V57	1.500	5.250	8.000	4.500	5.250	4.750	3.250	3.500
V58	1.250	6.000	6.750	5.500	6.500	3.500	3.500	3.000
V59	4.000	5.500	7.750	2.500	5.500	5.250	1.500	4.000
V60	2.125	6.000	7.750	6.750	5.500	2.875	2.375	2.625
V61	2.500	5.500	7.750	7.250	5.500	2.250	3.000	2.250
V62	3.500	5.250	8.000	5.500	6.250	3.375	2.375	1.750
V63	2.250	4.500	8.000	5.500	7.000	2.125	3.375	3.250
V64	6.500	4.000	7.500	3.500	2.250	4.250	4.500	3.500
V65	3.250	5.750	7.250	5.500	5.000	3.250	4.000	2.000
V66	3.750	4.250	5.000	3.750	4.500	6.000	4.250	4.500
V67	2.500	6.000	6.750	7.500	5.750	2.750	2.500	2.250
V68	1.250	5.750	7.250	6.500	6.500	2.500	3.250	3.000
V69	1.750	6.000	8.000	6.000	6.000	3.000	2.500	2.750
V70	3.000	5.500	7.500	5.000	6.500	3.250	2.500	2.750

V71	3.250	7.250	6.250	5.750	5.750	3.000	2.500	2.250
V72	3.000	6.000	7.750	5.500	6.750	2.750	1.750	2.500
V73	1.500	6.500	7.750	5.250	6.000	3.125	3.375	2.500
V74	2.250	6.500	8.000	4.750	6.500	3.750	2.000	2.250
Ortalama	2.921	5.870	7.405	5.010	5.725	3.159	2.904	3.007

Ek 13. İki sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.081	0.216	-0.080	0.117	0.153	0.227	0.251	0.221
V2	-0.020	0.022	-0.369	-0.057	-0.053	0.048	0.057	0.039
V3	0.155	0.158	-0.110	0.106	0.062	0.189	0.324	0.155
V4	-0.008	0.002	-0.338	-0.039	-0.071	-0.008	-0.008	0.017
V5	0.001	0.062	-0.144	0.021	0.017	0.116	0.175	0.121
V6	0.090	0.152	0.026	0.150	0.201	0.236	0.257	0.243
V7	0.676	0.377	0.416	0.475	0.411	0.686	0.693	0.690
V8	0.426	0.108	-0.009	0.357	0.452	0.307	0.303	0.304
V9	0.358	0.076	0.198	0.197	0.208	0.449	0.484	0.476
V10	0.746	0.594	0.430	0.678	0.608	0.731	0.737	0.731
V11	0.550	0.184	0.266	0.386	0.313	0.478	0.548	0.564
V12	0.655	0.321	0.260	0.509	0.416	0.586	0.631	0.599
V13	0.679	0.309	0.584	0.448	0.432	0.651	0.676	0.672
V14	0.784	0.583	0.607	0.690	0.672	0.784	0.784	0.784
V15	0.772	0.607	0.543	0.692	0.691	0.772	0.772	0.772
V16	0.506	0.405	0.418	0.481	0.465	0.532	0.561	0.572
V17	0.687	0.624	0.482	0.665	0.655	0.651	0.640	0.655
V18	0.614	0.372	0.301	0.513	0.513	0.592	0.644	0.628
V19	0.782	0.639	0.580	0.701	0.713	0.782	0.782	0.782
V20	0.707	0.674	0.506	0.720	0.686	0.686	0.701	0.696
V21	0.531	0.381	0.469	0.461	0.457	0.465	0.477	0.507
V22	0.562	0.400	0.476	0.544	0.462	0.569	0.596	0.589
V23	0.718	0.433	0.489	0.591	0.547	0.684	0.709	0.728
V24	0.488	0.299	0.219	0.398	0.358	0.432	0.467	0.443
V25	0.798	0.587	0.412	0.628	0.589	0.786	0.792	0.794
V26	0.055	0.028	-0.045	0.093	0.115	0.135	0.159	0.138
V27	0.221	0.027	0.055	0.111	0.114	0.162	0.185	0.194
V28	0.026	0.187	-0.393	0.208	0.082	0.277	0.225	0.318
V29	0.015	-0.007	-0.472	0.179	0.063	0.397	0.326	0.360
V30	0.023	0.102	-0.397	0.228	0.159	0.229	0.347	0.391
V31	0.746	0.702	0.754	0.763	0.763	0.746	0.746	0.746
V32	0.865	0.966	0.915	1.000	0.966	0.966	1.000	0.966
V33	0.947	0.964	0.947	0.964	0.964	0.947	0.947	0.947
V34	0.915	0.863	0.294	0.872	0.872	0.902	0.902	0.902
V35	0.714	0.660	-0.049	0.718	0.786	0.832	0.774	0.865
V36	0.595	0.413	0.029	0.501	0.481	0.557	0.554	0.558
V37	0.117	-0.023	-0.347	0.313	0.169	0.106	0.103	0.108
V38	0.797	0.881	0.041	0.966	0.975	0.966	0.966	0.966
V39	0.690	0.481	0.415	0.553	0.513	0.460	0.552	0.505
V40	0.830	0.845	0.868	0.900	0.881	0.843	0.907	0.903
V41	0.813	0.764	0.215	0.829	0.794	0.857	0.893	0.874
V42	-0.012	0.154	-0.429	0.072	0.048	0.072	0.174	0.099
V43	0.373	0.484	-0.493	0.414	0.294	0.574	0.727	0.541

V44	0.814	0.580	0.668	0.616	0.665	0.830	0.841	0.840
V45	0.632	0.319	0.171	0.320	0.418	0.509	0.565	0.545
V46	0.851	0.718	0.633	0.756	0.780	0.798	0.839	0.832
V47	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
V48	-0.060	-0.115	-0.100	0.016	-0.023	-0.007	-0.005	-0.019
V49	0.955	0.893	0.345	0.969	0.961	0.977	0.991	0.991
V50	0.944	0.779	0.366	0.945	0.950	0.936	0.949	0.961
V51	0.923	0.798	0.269	0.918	0.897	0.935	0.956	0.966
V52	0.885	0.563	0.049	0.850	0.830	0.870	0.893	0.930
V53	0.926	0.649	0.018	0.859	0.850	0.949	0.955	0.962
V54	0.898	0.728	0.137	0.834	0.822	0.862	0.861	0.894
V55	0.768	0.429	-0.015	0.679	0.652	0.688	0.780	0.802
V56	0.705	0.278	-0.192	0.638	0.629	0.600	0.743	0.740
V57	0.852	0.488	0.025	0.764	0.746	0.818	0.872	0.910
V58	0.892	0.618	0.175	0.810	0.781	0.855	0.868	0.897
V59	0.759	0.439	-0.103	0.736	0.729	0.736	0.834	0.843
V60	0.017	0.087	-0.051	-0.046	0.035	0.085	0.051	0.050
V61	-0.033	0.011	-0.200	-0.057	-0.015	-0.013	0.005	-0.013
V62	-0.028	0.048	-0.223	-0.002	0.005	-0.028	-0.031	-0.031
V63	-0.021	0.000	-0.332	-0.065	-0.084	-0.021	-0.026	-0.026
V64	0.475	0.089	0.201	0.181	0.282	0.436	0.446	0.483
V65	0.754	0.357	0.503	0.438	0.458	0.668	0.676	0.686
V66	0.162	-0.020	-0.018	0.078	0.143	0.218	0.265	0.273
V67	0.140	-0.019	-0.311	-0.084	-0.140	0.209	0.224	0.243
V68	-0.043	-0.115	-0.346	-0.184	-0.228	-0.041	-0.061	-0.044
V69	0.041	0.053	-0.155	0.064	0.057	0.241	0.286	0.306
V70	0.679	0.507	0.524	0.605	0.599	0.624	0.659	0.664
V71	0.665	0.413	0.452	0.514	0.461	0.587	0.638	0.634
V72	0.210	0.131	-0.019	0.081	0.059	0.217	0.265	0.194
V73	0.654	0.492	0.205	0.490	0.476	0.570	0.576	0.523
V74	0.381	0.178	0.060	0.114	0.111	0.363	0.419	0.337
Ortalama	0.512	0.385	0.166	0.459	0.445	0.526	0.552	0.548

Ek 14. İki sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	7.000	4.000	8.000	6.000	5.000	2.000	1.000	3.000
V2	5.000	4.000	8.000	7.000	6.000	2.000	1.000	3.000
V3	4.000	3.000	8.000	6.000	7.000	2.000	1.000	5.000
V4	3.000	2.000	8.000	6.000	7.000	4.000	5.000	1.000
V5	7.000	4.000	8.000	5.000	6.000	3.000	1.000	2.000
V6	7.000	5.000	8.000	6.000	4.000	3.000	1.000	2.000
V7	4.000	8.000	6.000	5.000	7.000	3.000	1.000	2.000
V8	2.000	7.000	8.000	3.000	1.000	4.000	6.000	5.000
V9	4.000	8.000	6.000	7.000	5.000	3.000	1.000	2.000
V10	1.000	7.000	8.000	5.000	6.000	3.500	2.000	3.500
V11	2.000	8.000	7.000	5.000	6.000	4.000	3.000	1.000
V12	1.000	7.000	8.000	5.000	6.000	4.000	2.000	3.000
V13	1.000	8.000	5.000	6.000	7.000	4.000	2.000	3.000
V14	2.500	8.000	7.000	5.000	6.000	2.500	2.500	2.500
V15	2.500	7.000	8.000	5.000	6.000	2.500	2.500	2.500
V16	4.000	8.000	7.000	5.000	6.000	3.000	2.000	1.000

V17	1.000	7.000	8.000	2.000	3.000	5.000	6.000	4.000
V18	3.000	7.000	8.000	6.000	5.000	4.000	1.000	2.000
V19	2.500	7.000	8.000	6.000	5.000	2.500	2.500	2.500
V20	2.000	7.000	8.000	1.000	6.000	5.000	3.000	4.000
V21	1.000	8.000	4.000	6.000	7.000	5.000	3.000	2.000
V22	4.000	8.000	6.000	5.000	7.000	3.000	1.000	2.000
V23	2.000	8.000	7.000	5.000	6.000	4.000	3.000	1.000
V24	1.000	7.000	8.000	5.000	6.000	4.000	2.000	3.000
V25	1.000	7.000	8.000	5.000	6.000	4.000	3.000	2.000
V26	6.000	7.000	8.000	5.000	4.000	3.000	1.000	2.000
V27	1.000	8.000	7.000	6.000	5.000	4.000	3.000	2.000
V28	7.000	5.000	8.000	4.000	6.000	2.000	3.000	1.000
V29	6.000	7.000	8.000	4.000	5.000	1.000	3.000	2.000
V30	7.000	6.000	8.000	4.000	5.000	3.000	2.000	1.000
V31	5.500	8.000	3.000	1.500	1.500	5.500	5.500	5.500
V32	8.000	4.500	7.000	1.500	4.500	4.500	1.500	4.500
V33	5.500	2.000	8.000	2.000	2.000	5.500	5.500	5.500
V34	1.000	7.000	8.000	5.500	5.500	3.000	3.000	3.000
V35	6.000	7.000	8.000	5.000	3.000	2.000	4.000	1.000
V36	1.000	7.000	8.000	5.000	6.000	3.000	4.000	2.000
V37	3.000	7.000	8.000	1.000	2.000	5.000	6.000	4.000
V38	7.000	6.000	8.000	5.000	1.000	3.000	3.000	3.000
V39	1.000	6.000	8.000	2.000	4.000	7.000	3.000	5.000
V40	8.000	6.000	5.000	3.000	4.000	7.000	1.000	2.000
V41	5.000	7.000	8.000	4.000	6.000	3.000	1.000	2.000
V42	7.000	2.000	8.000	4.000	6.000	5.000	1.000	3.000
V43	6.000	4.000	8.000	5.000	7.000	2.000	1.000	3.000
V44	4.000	8.000	5.000	7.000	6.000	3.000	1.000	2.000
V45	1.000	7.000	8.000	6.000	5.000	4.000	2.000	3.000
V46	1.000	7.000	8.000	6.000	5.000	4.000	2.000	3.000
V47	4.500	4.500	4.500	4.500	4.500	4.500	4.500	4.500
V48	6.000	8.000	7.000	1.000	5.000	3.000	2.000	4.000
V49	6.000	7.000	8.000	4.000	5.000	3.000	1.000	2.000
V50	5.000	7.000	8.000	4.000	2.000	6.000	3.000	1.000
V51	4.000	7.000	8.000	5.000	6.000	3.000	2.000	1.000
V52	3.000	7.000	8.000	5.000	6.000	4.000	2.000	1.000
V53	4.000	7.000	8.000	5.000	6.000	3.000	2.000	1.000
V54	1.000	7.000	8.000	5.000	6.000	3.000	4.000	2.000
V55	3.000	7.000	8.000	5.000	6.000	4.000	2.000	1.000
V56	3.000	7.000	8.000	4.000	5.000	6.000	1.000	2.000
V57	3.000	7.000	8.000	5.000	6.000	4.000	2.000	1.000
V58	2.000	7.000	8.000	5.000	6.000	4.000	3.000	1.000
V59	3.000	7.000	8.000	4.000	6.000	5.000	2.000	1.000
V60	6.000	1.000	8.000	7.000	5.000	2.000	3.000	4.000
V61	6.000	1.000	8.000	7.000	5.000	3.000	2.000	4.000
V62	4.000	1.000	8.000	3.000	2.000	5.000	6.000	7.000
V63	2.000	1.000	8.000	6.000	7.000	3.000	4.000	5.000
V64	2.000	8.000	6.000	7.000	5.000	4.000	3.000	1.000
V65	1.000	8.000	5.000	7.000	6.000	4.000	3.000	2.000
V66	4.000	8.000	7.000	6.000	5.000	3.000	2.000	1.000
V67	4.000	5.000	8.000	6.000	7.000	3.000	2.000	1.000
V68	2.000	5.000	8.000	6.000	7.000	1.000	4.000	3.000
V69	7.000	6.000	8.000	4.000	5.000	3.000	2.000	1.000
V70	1.000	8.000	7.000	5.000	6.000	4.000	3.000	2.000
V71	1.000	8.000	7.000	5.000	6.000	4.000	2.000	3.000

V72	3.000	5.000	8.000	6.000	7.000	2.000	1.000	4.000
V73	1.000	5.000	8.000	6.000	7.000	3.000	2.000	4.000
V74	2.000	5.000	8.000	6.000	7.000	3.000	1.000	4.000
Ortalama	3.581	6.176	7.426	4.838	5.297	3.581	2.507	2.595

Ek 15. İki sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	0.001	0.096	-0.105	0.107	0.153	0.197	0.221	0.204
V2	-0.020	-0.077	-0.287	-0.099	-0.141	0.027	-0.019	0.061
V3	0.037	0.165	-0.091	0.137	0.150	0.316	0.394	0.320
V4	-0.008	-0.137	-0.306	-0.119	-0.175	-0.034	-0.022	-0.005
V5	-0.006	-0.049	-0.151	-0.036	-0.049	0.128	0.153	0.126
V6	0.198	0.188	0.015	0.162	0.213	0.310	0.256	0.260
V7	0.401	0.502	0.387	0.505	0.459	0.619	0.576	0.560
V8	0.117	0.108	0.086	0.291	0.282	0.326	0.317	0.320
V9	0.455	0.440	0.410	0.490	0.496	0.542	0.518	0.544
V10	0.699	0.701	0.478	0.727	0.719	0.667	0.696	0.701
V11	0.484	0.509	0.145	0.460	0.415	0.561	0.532	0.585
V12	0.624	0.616	0.253	0.608	0.506	0.637	0.641	0.625
V13	0.583	0.591	0.281	0.580	0.527	0.668	0.628	0.665
V14	0.719	0.729	0.498	0.734	0.683	0.708	0.730	0.736
V15	0.735	0.705	0.514	0.730	0.705	0.723	0.709	0.727
V16	0.655	0.602	0.251	0.587	0.535	0.629	0.632	0.650
V17	0.669	0.704	0.484	0.713	0.678	0.693	0.669	0.715
V18	0.599	0.583	0.335	0.567	0.530	0.615	0.586	0.621
V19	0.697	0.758	0.558	0.732	0.702	0.715	0.717	0.721
V20	0.698	0.728	0.485	0.725	0.693	0.729	0.704	0.714
V21	0.658	0.604	0.353	0.597	0.565	0.677	0.657	0.665
V22	0.705	0.706	0.393	0.686	0.621	0.728	0.731	0.692
V23	0.688	0.705	0.548	0.694	0.676	0.698	0.723	0.709
V24	0.513	0.409	0.292	0.428	0.364	0.477	0.510	0.446
V25	0.741	0.720	0.480	0.702	0.652	0.736	0.754	0.727
V26	-0.008	0.034	-0.001	0.049	0.050	0.069	0.094	0.065
V27	0.083	0.105	0.096	0.133	0.122	0.181	0.175	0.162
V28	0.078	0.107	-0.083	0.225	0.150	0.221	0.184	0.216
V29	0.075	0.137	0.053	0.160	0.188	0.191	0.117	0.217
V30	0.100	0.185	-0.089	0.292	0.207	0.256	0.207	0.213
V31	0.804	0.793	0.644	0.716	0.783	0.783	0.824	0.770
V32	0.949	0.953	0.577	0.983	0.970	0.983	0.970	0.970
V33	0.982	0.974	0.910	0.918	0.974	0.972	0.956	0.972
V34	0.933	0.952	0.704	0.948	0.947	0.959	0.952	0.959
V35	0.884	0.912	0.397	0.884	0.892	0.930	0.915	0.923
V36	0.697	0.568	0.265	0.564	0.596	0.629	0.652	0.607
V37	-0.018	0.011	-0.178	0.037	0.128	0.092	0.079	0.093
V38	0.806	0.969	0.215	0.889	0.905	0.980	1.000	0.936
V39	0.738	0.501	0.329	0.651	0.594	0.645	0.677	0.645
V40	0.896	0.902	0.821	0.892	0.899	0.895	0.892	0.893
V41	0.933	0.940	0.569	0.941	0.841	0.964	0.973	0.918
V42	-0.012	0.200	-0.325	0.150	0.039	0.133	0.044	0.126
V43	0.280	0.809	-0.224	0.834	0.919	0.752	0.813	0.839
V44	0.826	0.575	0.730	0.541	0.643	0.625	0.465	0.680
V45	0.433	0.261	0.211	0.301	0.279	0.279	0.269	0.340
V46	0.731	0.640	0.558	0.665	0.637	0.641	0.625	0.669

V47	0.980	0.992	0.826	1.000	1.000	1.000	0.990	1.000
V48	-0.061	-0.099	-0.100	-0.102	-0.084	-0.074	-0.049	-0.072
V49	0.930	0.918	0.614	0.932	0.913	0.945	0.938	0.932
V50	0.832	0.772	0.511	0.778	0.748	0.815	0.782	0.811
V51	0.755	0.819	0.464	0.810	0.801	0.851	0.807	0.839
V52	0.600	0.598	0.296	0.638	0.557	0.697	0.699	0.668
V53	0.622	0.738	0.307	0.749	0.706	0.781	0.762	0.761
V54	0.745	0.782	0.518	0.782	0.764	0.808	0.809	0.800
V55	0.604	0.638	0.269	0.662	0.599	0.730	0.675	0.671
V56	0.474	0.598	0.194	0.605	0.547	0.633	0.610	0.625
V57	0.574	0.721	0.314	0.728	0.683	0.761	0.729	0.752
V58	0.821	0.817	0.469	0.825	0.785	0.851	0.840	0.858
V59	0.555	0.657	0.241	0.639	0.635	0.700	0.664	0.703
V60	-0.017	0.080	-0.307	0.010	0.002	0.195	0.145	0.185
V61	-0.033	0.012	-0.261	-0.037	-0.024	0.050	0.063	0.063
V62	-0.002	0.159	-0.115	0.074	0.026	0.052	0.050	0.022
V63	-0.021	0.051	-0.217	-0.005	-0.080	0.024	0.048	0.043
V64	0.470	0.385	0.319	0.396	0.513	0.484	0.460	0.477
V65	0.729	0.629	0.503	0.644	0.655	0.688	0.729	0.701
V66	0.152	0.166	-0.045	0.128	0.313	0.259	0.295	0.269
V67	0.054	-0.083	-0.277	-0.135	-0.127	0.224	0.249	0.225
V68	-0.043	-0.138	-0.323	-0.193	-0.189	-0.038	-0.072	-0.048
V69	0.183	0.196	-0.151	0.061	0.059	0.222	0.251	0.237
V70	0.649	0.576	0.446	0.595	0.576	0.674	0.651	0.647
V71	0.712	0.560	0.496	0.616	0.550	0.682	0.687	0.664
V72	0.094	0.064	-0.056	0.045	0.007	0.277	0.312	0.293
V73	0.651	0.521	0.274	0.561	0.525	0.606	0.559	0.529
V74	0.196	0.230	0.010	0.178	0.177	0.422	0.484	0.473
Ortalama	0.473	0.480	0.239	0.479	0.464	0.535	0.528	0.533

Ek 16. İki sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V1	7.000	5.500	8.000	5.000	4.000	2.000	2.000	2.500
V2	3.500	5.500	8.000	5.500	7.000	2.000	3.500	1.000
V3	7.500	5.000	7.500	5.000	4.500	2.000	2.000	2.500
V4	1.500	6.500	7.500	5.500	6.500	3.000	3.000	2.500
V5	4.500	5.500	8.000	5.500	6.500	2.000	2.000	2.000
V6	5.000	5.000	8.000	6.000	5.000	1.000	3.000	3.000
V7	6.000	4.000	7.000	4.000	6.000	1.500	3.500	4.000
V8	7.000	7.000	6.500	3.500	3.500	1.500	4.000	3.000
V9	6.000	6.000	7.000	5.000	5.000	1.500	4.000	1.500
V10	4.000	4.000	8.000	2.500	1.500	7.000	5.000	4.000
V11	4.000	4.500	8.000	5.500	7.000	2.500	3.500	1.000
V12	4.000	5.000	8.000	5.000	7.000	1.500	1.500	4.000
V13	5.000	4.500	8.000	5.500	7.000	1.500	3.000	1.500
V14	4.000	3.500	8.000	2.500	7.000	6.000	2.500	2.500
V15	3.000	5.000	8.000	2.500	6.000	4.000	4.500	3.000
V16	1.000	5.000	8.000	6.000	7.000	3.500	3.500	2.000
V17	6.000	3.500	8.000	2.500	5.000	3.500	6.000	1.500
V18	3.000	4.500	8.000	6.000	7.000	2.000	4.000	1.500
V19	5.000	1.500	8.000	3.500	5.000	5.000	4.500	3.500

V20	6.000	3.500	8.000	3.000	5.500	2.000	4.500	3.500
V21	2.500	5.000	8.000	6.000	7.000	1.000	3.500	3.000
V22	4.000	4.000	8.000	5.000	7.000	1.750	1.250	5.000
V23	5.500	3.500	8.000	4.500	6.500	4.500	1.000	2.500
V24	1.500	5.500	8.000	5.500	7.000	3.000	1.500	4.000
V25	2.500	5.000	8.000	6.000	7.000	3.500	1.000	3.000
V26	7.000	5.500	7.500	3.500	3.500	4.000	1.000	4.000
V27	6.000	7.000	7.000	4.500	5.000	1.500	2.000	3.000
V28	5.000	6.000	8.000	2.000	6.000	2.000	4.500	2.500
V29	5.500	5.000	4.500	5.000	3.500	3.500	5.500	3.500
V30	5.000	4.000	8.000	1.000	5.000	4.000	4.500	4.500
V31	3.500	3.500	8.000	6.500	4.000	4.000	2.000	4.500
V32	4.750	5.750	8.000	2.500	3.750	2.500	4.750	4.000
V33	2.000	3.000	5.000	7.250	3.500	4.250	6.750	4.250
V34	4.500	3.500	8.000	4.500	4.500	3.000	4.000	4.000
V35	5.500	3.000	8.000	6.000	6.000	2.500	2.500	2.500
V36	1.500	5.500	8.000	6.000	4.000	3.000	3.500	4.500
V37	7.000	5.500	8.000	4.500	3.500	2.000	3.500	2.000
V38	6.500	3.250	8.000	4.250	5.000	2.750	2.000	4.250
V39	3.000	6.500	8.000	4.500	5.500	3.000	2.500	3.000
V40	4.000	3.500	8.000	4.500	2.500	4.500	4.500	4.500
V41	4.250	3.750	8.000	4.500	7.000	3.000	1.500	4.000
V42	6.500	1.500	8.000	2.500	5.500	3.000	6.000	3.000
V43	7.000	4.500	8.000	3.000	2.250	4.500	3.500	3.250
V44	1.000	6.500	5.000	6.000	5.000	4.000	5.000	3.500
V45	1.000	6.000	7.000	3.500	5.000	5.500	5.500	2.500
V46	2.000	4.500	7.000	3.500	5.500	5.000	5.500	3.000
V47	5.500	4.500	8.000	3.250	3.250	3.250	5.000	3.250
V48	3.500	6.000	7.000	7.000	5.000	3.000	1.000	3.500
V49	5.500	4.500	8.000	3.500	6.000	1.000	3.000	4.500
V50	1.500	6.000	8.000	4.500	7.000	2.000	4.500	2.500
V51	7.000	4.000	8.000	4.500	4.000	1.000	5.000	2.500
V52	4.500	5.500	8.000	4.500	6.500	2.000	2.000	3.000
V53	7.000	5.000	8.000	3.000	6.000	2.500	1.500	3.000
V54	7.000	4.500	8.000	4.500	4.500	2.000	2.000	3.500
V55	6.000	4.500	8.000	3.000	6.500	1.000	4.000	3.000
V56	7.000	4.000	8.000	3.000	6.000	3.000	2.000	3.000
V57	7.000	4.500	8.000	3.000	6.000	1.500	3.500	2.500
V58	5.000	4.500	8.000	3.500	7.000	3.500	3.000	1.500
V59	7.000	3.500	8.000	5.000	3.500	2.500	4.500	2.000
V60	7.000	3.500	8.000	5.500	5.500	2.500	2.500	1.500
V61	5.500	5.500	8.000	5.000	5.500	2.500	1.500	2.500
V62	5.500	1.000	8.000	2.500	5.500	4.000	3.500	6.000
V63	5.500	1.000	8.000	5.500	7.000	3.500	3.000	2.500
V64	4.500	7.000	7.500	6.500	1.000	2.500	3.500	3.500
V65	1.500	6.500	8.000	5.500	5.500	4.000	1.500	3.500
V66	6.000	5.500	8.000	6.500	2.000	3.500	1.500	3.000
V67	4.000	5.000	8.000	6.500	6.500	2.500	1.500	2.000
V68	2.500	5.000	8.000	6.500	6.500	2.000	3.000	2.500
V69	3.500	4.000	8.000	6.500	6.500	2.500	2.000	3.000
V70	3.000	6.500	8.000	5.500	5.500	1.500	3.000	3.000
V71	1.000	6.500	8.000	5.000	6.500	3.000	3.000	3.000
V72	4.500	5.500	8.000	5.000	7.000	3.000	1.000	2.000
V73	1.000	6.000	8.000	3.500	5.500	2.500	4.000	5.500
V74	5.500	4.500	8.000	6.000	6.000	2.500	1.000	2.500

Ortalama	4.534	4.713	7.743	4.591	5.355	2.845	3.172	3.047
----------	-------	-------	-------	-------	-------	--------------	--------------	--------------

Ek 17. İkiiden fazla sınıflı veri setlerinde genel sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.064	-0.106	-0.098	-0.081	-0.081	-0.074	-0.078	-0.079
V76	0.079	0.086	0.006	0.091	0.097	0.155	0.176	0.168
V77	-0.347	-0.311	-0.348	-0.306	-0.285	-0.300	-0.279	-0.284
V78	0.382	0.409	0.367	0.393	0.402	0.397	0.421	0.395
V79	0.592	0.588	0.594	0.592	0.592	0.595	0.613	0.595
V80	0.081	0.082	0.000	0.089	0.049	0.155	0.159	0.122
V81	0.382	0.376	0.112	0.415	0.279	0.448	0.442	0.416
V82	0.888	0.869	0.860	0.869	0.862	0.887	0.885	0.882
Ortalama	0.249	0.249	0.187	0.258	0.239	0.283	0.292	0.277

Ek 18. İkiiden fazla sınıflı veri setlerinde genel sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	3.643	6.214	6.000	4.429	3.929	3.714	4.000	4.071
V76	5.714	5.500	7.786	5.000	4.429	2.500	2.357	2.714
V77	6.214	4.143	6.643	4.071	3.929	4.286	3.000	3.714
V78	5.286	4.071	4.357	5.107	4.321	4.679	4.000	4.179
V79	4.786	5.500	4.857	4.643	4.500	4.214	3.214	4.286
V80	4.500	5.357	6.643	4.929	5.000	3.429	2.500	3.643
V81	3.857	4.179	7.286	3.679	5.929	3.357	3.714	4.000
V82	3.714	4.571	5.643	5.321	5.286	4.071	3.714	3.679
Ortalama	4.714	4.942	6.152	4.647	4.665	3.781	3.313	3.786

Ek 19. İkiiden fazla sınıflı veri setlerinde lineer sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.113	-0.120	-0.125	-0.113	-0.117	-0.106	-0.121	-0.112
V76	0.100	0.111	0.075	0.121	0.110	0.229	0.206	0.219
V77	-0.353	-0.300	-0.324	-0.283	-0.211	-0.286	-0.282	-0.221
V78	0.073	0.092	0.117	0.078	0.081	0.058	0.100	0.072
V79	0.348	0.346	0.367	0.355	0.364	0.373	0.415	0.382
V80	-0.055	-0.056	-0.049	-0.044	-0.050	-0.032	-0.029	-0.042
V81	-0.056	0.268	-0.110	0.296	0.171	0.295	0.301	0.288
V82	0.929	0.912	0.889	0.913	0.897	0.923	0.927	0.929
Ortalama	0.109	0.157	0.105	0.165	0.156	0.182	0.190	0.190

Ek 20. İkiiden fazla sınıflı veri setlerinde lineer sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	4.667	5.000	6.333	4.000	4.333	2.667	5.667	3.333
V76	6.333	5.667	7.667	5.000	5.333	1.000	2.667	2.333
V77	7.333	3.333	6.333	3.000	2.667	5.333	3.667	4.333
V78	5.667	4.000	3.000	4.333	4.333	6.000	3.333	5.333
V79	6.667	7.000	4.667	5.667	4.000	3.667	1.667	2.667
V80	3.333	6.000	4.333	4.333	4.667	4.000	3.333	6.000
V81	7.333	3.667	7.667	3.667	5.667	3.000	2.333	2.667
V82	2.667	5.667	7.333	4.333	6.000	4.000	3.333	2.667
Ortalama	5.500	5.042	5.917	4.292	4.625	3.708	3.250	3.667

Ek 21. İkiiden fazla sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.050	-0.103	-0.091	-0.072	-0.072	-0.065	-0.066	-0.070
V76	0.073	0.079	-0.013	0.083	0.094	0.134	0.168	0.155
V77	-0.345	-0.314	-0.355	-0.313	-0.305	-0.304	-0.278	-0.301
V78	0.466	0.495	0.435	0.478	0.490	0.489	0.509	0.484
V79	0.659	0.654	0.656	0.656	0.654	0.655	0.667	0.653
V80	0.118	0.119	0.013	0.125	0.076	0.206	0.211	0.167
V81	0.502	0.405	0.173	0.448	0.308	0.490	0.480	0.451
V82	0.876	0.858	0.852	0.857	0.852	0.878	0.873	0.869
Ortalama	0.287	0.274	0.209	0.283	0.262	0.310	0.320	0.301

Ek 22. İkiiden fazla sınıflı veri setlerinde non-lineer sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	3.364	6.545	5.909	4.545	3.818	4.000	3.545	4.273
V76	5.545	5.455	7.818	5.000	4.182	2.909	2.273	2.818
V77	5.909	4.364	6.727	4.364	4.273	4.000	2.818	3.545
V78	5.182	4.091	4.727	5.318	4.318	4.318	4.182	3.864
V79	4.273	5.091	4.909	4.364	4.636	4.364	3.636	4.727
V80	4.818	5.182	7.273	5.091	5.091	3.273	2.273	3.000
V81	2.909	4.318	7.182	3.682	6.000	3.455	4.091	4.364
V82	4.000	4.273	5.182	5.591	5.091	4.091	3.818	3.955
Ortalama	4.500	4.915	6.216	4.744	4.676	3.801	3.330	3.818

Ek 23. İkiiden fazla sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.035	-0.059	-0.039	-0.018	-0.047	-0.017	-0.041	-0.016
V76	0.049	0.111	0.012	0.129	0.121	0.192	0.202	0.175
V77	-0.330	-0.277	-0.276	-0.267	-0.278	-0.412	-0.380	-0.409
V78	0.512	0.523	0.547	0.477	0.493	0.509	0.458	0.508
V79	0.878	0.885	0.870	0.876	0.869	0.842	0.863	0.871
V80	0.178	0.159	0.084	0.141	0.146	0.197	0.208	0.223
V81	0.644	0.648	0.247	0.613	0.480	0.609	0.590	0.626
V82	0.924	0.955	0.924	0.934	0.935	0.931	0.934	0.954
Ortalama	0.352	0.368	0.296	0.361	0.340	0.356	0.354	0.367

Ek 24. İkiiden fazla sınıflı veri setlerinde sinir ağı sınıflayıcıları için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	5.500	6.500	4.500	4.000	5.500	3.500	4.000	2.500
V76	7.500	5.500	7.500	4.000	5.500	1.500	1.500	3.000
V77	5.000	2.500	2.500	2.500	2.500	7.500	6.000	7.500
V78	4.000	3.000	2.000	7.000	4.500	4.000	7.000	4.500
V79	3.000	2.500	5.000	4.500	5.000	6.500	5.500	4.000
V80	4.500	4.500	7.500	6.000	6.000	3.500	2.500	1.500
V81	3.000	2.500	8.000	3.500	6.000	4.000	5.500	3.500
V82	5.500	1.250	6.250	5.250	4.500	7.000	4.500	1.750
Ortalama	4.750	3.531	5.406	4.594	4.938	4.688	4.563	3.531

Ek 25. İkiiden fazla sınıflı veri setlerinde ağaç yapılı sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.009	-0.042	-0.047	-0.014	-0.021	-0.050	-0.036	-0.061
V76	0.095	0.133	0.018	0.106	0.133	0.118	0.182	0.173
V77	-0.276	-0.241	-0.289	-0.250	-0.246	-0.253	-0.228	-0.249
V78	0.502	0.600	0.439	0.563	0.568	0.464	0.553	0.473
V79	0.702	0.699	0.697	0.701	0.704	0.723	0.730	0.746
V80	0.233	0.268	0.111	0.251	0.214	0.261	0.280	0.267
V81	0.750	0.716	0.453	0.699	0.452	0.728	0.702	0.661
V82	0.835	0.837	0.784	0.819	0.808	0.831	0.814	0.806
Ortalama	0.354	0.371	0.271	0.359	0.327	0.353	0.375	0.352

Ek 26. İkidem fazla sınıflı veri setlerinde ağaç yapıllı sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	2.750	5.500	6.000	2.500	2.250	6.250	4.000	6.750
V76	5.250	4.500	7.750	5.750	3.000	4.750	2.500	2.500
V77	6.250	3.000	7.500	4.500	4.250	4.500	2.250	3.750
V78	5.250	2.750	5.000	4.000	3.500	6.875	3.500	5.125
V79	5.750	6.000	6.000	5.750	5.000	3.500	2.500	1.500
V80	4.750	4.500	7.250	5.250	4.000	4.250	2.250	3.750
V81	2.250	3.125	7.250	3.875	7.500	3.250	4.000	4.750
V82	3.750	3.500	6.750	4.625	5.375	3.000	4.000	5.000
Ortalama	4.500	4.109	6.688	4.531	4.359	4.547	3.125	4.141

Ek 27. İkidem fazla sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.111	-0.184	-0.161	-0.152	-0.143	-0.114	-0.125	-0.120
V76	0.047	0.007	-0.039	0.020	0.019	0.135	0.151	0.146
V77	-0.514	-0.485	-0.508	-0.481	-0.385	-0.377	-0.349	-0.325
V78	0.292	0.276	0.236	0.289	0.291	0.374	0.403	0.365
V79	0.641	0.637	0.639	0.646	0.647	0.646	0.650	0.622
V80	-0.154	-0.121	-0.137	-0.121	-0.206	0.073	0.047	-0.061
V81	0.145	-0.090	-0.057	0.000	0.014	0.160	0.191	0.136
V82	0.920	0.887	0.922	0.908	0.908	0.920	0.920	0.920
Ortalama	0.158	0.116	0.112	0.138	0.143	0.227	0.236	0.210

Ek 28. İkidem fazla sınıflı veri setlerinde yoğunluk temelli sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	3.500	7.250	6.000	5.250	4.750	2.000	4.000	3.250
V76	5.500	6.000	7.750	5.500	5.000	1.500	2.250	2.500
V77	6.750	6.250	7.000	5.500	3.750	2.750	2.000	2.000
V78	5.375	5.625	6.500	5.250	5.000	3.250	2.000	3.000
V79	4.250	5.500	4.500	3.750	3.750	4.250	3.750	6.250
V80	5.750	5.750	6.000	5.500	6.250	1.250	2.750	2.750
V81	3.750	6.500	6.500	5.500	4.750	3.250	2.500	3.250
V82	3.875	5.625	3.375	6.625	4.875	3.875	3.875	3.875
Ortalama	4.844	6.063	5.953	5.359	4.766	2.766	2.891	3.359

Ek 29. İkiden fazla sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	-0.070	-0.205	-0.174	-0.164	-0.161	-0.126	-0.121	-0.145
V76	0.055	0.017	-0.038	0.046	0.041	0.057	0.086	0.066
V77	-0.230	-0.271	-0.290	-0.255	-0.257	-0.202	-0.209	-0.201
V78	0.411	0.453	0.365	0.405	0.480	0.423	0.423	0.423
V79	0.887	0.885	0.883	0.896	0.895	0.913	0.890	0.885
V80	0.261	0.173	-0.063	0.306	0.272	0.266	0.279	0.268
V81	0.130	0.275	-0.089	0.340	0.255	0.164	0.109	0.090
V82	0.877	0.860	0.800	0.855	0.816	0.935	0.916	0.890
Ortalama	0.290	0.273	0.174	0.304	0.293	0.304	0.296	0.285

Ek 30. İkiden fazla sınıflı veri setlerinde komşuluk temelli sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	1.000	8.000	7.000	6.000	5.000	3.000	2.000	4.000
V76	4.000	7.000	8.000	5.000	6.000	3.000	1.000	2.000
V77	4.000	7.000	8.000	5.000	6.000	2.000	3.000	1.000
V78	6.000	2.000	8.000	7.000	1.000	4.000	4.000	4.000
V79	5.000	6.000	8.000	2.000	3.000	1.000	4.000	7.000
V80	6.000	7.000	8.000	1.000	3.000	5.000	2.000	4.000
V81	5.000	2.000	8.000	1.000	3.000	4.000	6.000	7.000
V82	4.000	5.000	8.000	6.000	7.000	1.000	2.000	3.000
Ortalama	4.375	5.500	7.875	4.125	4.250	2.875	3.000	4.000

Ek 31. İkiden fazla sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO-ROS	UGO-SMOTE	UGO-RWO
V75	0.010	-0.041	-0.039	0.004	-0.019	-0.040	-0.028	-0.015
V76	0.097	0.130	0.053	0.131	0.134	0.186	0.244	0.221
V77	-0.293	-0.245	-0.270	-0.250	-0.234	-0.334	-0.313	-0.336
V78	0.511	0.576	0.562	0.511	0.575	0.502	0.508	0.531
V79	0.810	0.833	0.812	0.833	0.822	0.809	0.822	0.835
V80	0.149	0.191	0.097	0.173	0.208	0.219	0.243	0.233
V81	0.562	0.649	0.378	0.570	0.400	0.628	0.623	0.549
V82	0.926	0.909	0.915	0.913	0.906	0.924	0.925	0.936
Ortalama	0.347	0.375	0.313	0.361	0.349	0.362	0.378	0.369

Ek 32. İkiiden fazla sınıflı veri setlerinde Bayesci sınıflayıcılar için IAM sıra ortalamaları

Veri seti	NORES	ROS	RUS	SMOTE	RWO	UGO- ROS	UGO- SMOTE	UGO- RWO
V75	2.000	6.000	6.000	2.000	3.500	6.500	5.000	5.000
V76	6.500	5.500	7.500	5.500	5.000	2.500	1.000	2.500
V77	6.000	3.000	5.500	3.500	3.000	5.000	3.500	6.500
V78	5.500	2.500	2.000	6.500	2.000	6.500	6.500	4.500
V79	6.500	3.000	5.500	3.000	4.500	5.000	5.000	3.500
V80	7.500	4.500	7.500	6.000	3.500	3.500	1.500	2.000
V81	4.500	1.500	7.000	4.500	7.500	2.500	3.500	5.000
V82	3.500	4.500	5.250	6.500	5.500	5.000	3.750	2.000
Ortalama	5.250	3.813	5.781	4.688	4.313	4.563	3.719	3.875

ÖZ GEÇMİŞ

Fatih SAĞLAM, Zile Anadolu Öğretmen Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi, İstatistik bölümünden 28.06.2020 tarihinde mezun oldu. 2020 yılında OMÜ LEE İstatistik Ana Bilim Dalı Yüksek Lisans programını bitirdi ve hemen ardından OMÜ İstatistik Ana Bilim Dalı doktora programına başladı. Mezuniyetinden bu yana OMÜ İstatistik Bölümünde araştırma görevlisi olarak görev yapan Fatih Sağlam iyi derecede İngilizce bilmektedir (YDS: 87.5).

İletişim Bilgileri

ORCID ID : 0000-0002-2084-2008.

Yayınlar:

SCI endeksli Makaleler:

- Sağlam, F., & Cengiz, M. A. (2024). Local resampling for locally weighted Naïve Bayes in imbalanced data. *Computing*, 106(1), 185-200.
- Sağlam, F., Şanlı, T., Cengiz, M. A., & Terzi, Y. (2023). Alternative expectation approaches for expectation-maximization missing data imputations in cox regression. *Communications in Statistics-Simulation and Computation*, 52(12), 5966-5974.
- Sağlam, F., & Cengiz, M. A. (2022). A novel SMOTE-based resampling technique through noise detection and the boosting procedure. *Expert Systems with Applications*, 200, 117023.
- Sağlam, F., Yıldırım, E., & Cengiz, M. A. (2022). Clustered Bayesian classification for within-class separation. *Expert Systems with Applications*, 208, 118152.
- Terzi, E., Sarıbacak, B., Sağlam, F., & Cengiz, M. A. (2022). A Novel Expert System for Diagnosis of Iron Deficiency Anemia. *Computational and Mathematical Methods in Medicine*, 2022.
- Öztürk, E., Yıldızlı, K., & Sağlam, F. (2021). Investigation on an innovative internally cooled smart cutting tool with the built-in cooling-control system. *Arabian Journal for Science and Engineering*, 46, 2397-2411.

ESCI endeksli makaleler:

- Sağlam, F., Dünder, E., & Cengiz, M. A. (2023). An Enhanced Classification System Based On Kernel Principal Component Analysis And Data Complexity Measures. *Journal of Science and Arts*, 23(2), 447-458.

Öztürk, E., Kaygusuz, S., Sağlam, F., & Yıldızlı, K. (2022). Park halindeki araçta unutulmuş canlıları korumak için araç camlarını kontrol eden sensör tasarımı ve üretimi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 28(1), 26-36.

TR Dizin endeksli makaleler:

Sağlam, F., Sözen, M., & Cengiz, M. A. (2021). Optimization Based Undersampling for Imbalanced Classes. *Adıyaman University Journal of Science*, 11(2), 385-409.

Diğer endeksli makaleler:

Sağlam, F., & Cengiz, M. A. (2023). Sosyal Bilimlerde Veri Bilimi Kullanımı. *Avrasya Sosyal ve Ekonomi Araştırmaları Dergisi*, 10(4), 1175-1207.

Pala, M., Sağlam, F., & Sözen, Ç. (2019). Kapula Modelinin Belirlenmesinde AIC Değerinin Hatalı Seçimi. *The Journal of International Scientific Researches*, 4(3), 335-343.

Pala, M., & Sağlam, F. (2019). PISA fen, matematik ve okuma puanları arasındaki bağımlılık yapısının kapula ile modellenmesi. *Selçuk Üniversitesi Fen Fakültesi Fen Dergisi*, 45(2), 149-162.

Uluslararası kongre tam metinler:

Sağlam, F., & Cengiz, M. A. (2019, September). Noise detection in imbalanced classes using adaptive boosting. In *2019 4th International Conference on Computer Science and Engineering (UBMK)* (pp. 449-452). IEEE.

Sağlam, F., Turan, S. C., & Cengiz, M. A. (2019). Boosting İle Gözlemlerin Doğru Tahmin Edilebilirlik Seviyesi Tespiti Ve Dengesiz Sınıf Probleminde Sentetik Veri Üretiminde Kullanımı. Presented at the Uluslararası 19 Mayıs Multidisipliner Çalışmalar Kongresi, SAMSUN.

Bildiriler:

Kındap Tepe, Y., Koçoğlu, F., & Sağlam, F. (2023). Annenin Psikolojik Kontrol ve Özerklik Desteği Ebeveynlik Uygulamalarının Öncülleri Mükemmeliyetçilik ve Psikolojik Dayanıklılık Olabilir Mi . Presented at the X. Prof. Dr. Işık Savaşır Klinik Psikoloji Sempozyumu: Dijital Dünyada Klinik Psikoloji, İZMİR.

Şanlı, T. & Sağlam, F. (2022). R-Shiny Application of Double-Censored Parametric Survival Data. Presented at 5th International Conference on Data Science and Applications, Fethiye, MUĞLA.

Kazanılan Ödüller, Teşvikler ve Burslar

1. 2023 yılı Ondokuz Mayıs Üniversitesi Akademik Teşviği
2. “A novel SMOTE-based resampling technique through noise detection and the boosting procedure” makalesi için TÜBİTAK Uluslararası Bilimsel Yayın Teşviği
3. “Clustered Bayesian classification for within-class separation” makalesi için TÜBİTAK Uluslararası Bilimsel Yayın Teşviği

4. “A Novel Expert System for Diagnosis of Iron Deficiency Anemia” makalesi için
TÜBİTAK Uluslararası Bilimsel Yayın Teşviği

