T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SINIF DENGESİZLİĞİ SORUNUNU ÇÖZMEK İÇİN KULLANILAN ALGORİTMALARIN FARKLI SINIFLANDIRMA YÖNTEMLERİNDE PERFORMANSLARININ KARŞILAŞTIRILMASI

Duygu AYDIN HAKLI

Biyoistatistik Programı
BÜTÜNLEŞİK DOKTORA TEZİ

ANKARA
2018

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

SINIF DENGESİZLİĞİ SORUNUNU ÇÖZMEK İÇİN KULLANILAN ALGORİTMALARIN FARKLI SINIFLANDIRMA YÖNTEMLERİNDE PERFORMANSLARININ KARŞILAŞTIRILMASI

Duygu AYDIN HAKLI

Biyoistatistik Programı
BÜTÜNLEŞİK DOKTORA TEZİ

TEZ DANIŞMANI
Prof. Dr. Erdem KARABULUT

ANKARA
2018

SINIF DENGESİZLİĞİ SORUNUNU ÇÖZMEK İÇİN KULLANILAN ALGORİTMALARIN FARKLI SINIFLANDIRMA YÖNTEMLERİNDE PERFORMANSLARININ KARŞILAŞTIRILMASI

Duygu AYDIN HAKLI

Danışman: Prof. Dr. Erdem KARABULUT

Bu tez çalışması 27/06/2018 tarihinde jürimiz tarafından "Biyoistatistik Programı" nda bütünleşik doktora tezi olarak kabul edilmiştir.

Jüri Başkanı:	*Prof. Dr. A. Ergun KARAAĞAOĞLU* *(Hacettepe Üniversitesi)*
Üye:	*Prof. Dr. Ersin ÖĞÜŞ* *(Başkent Üniversitesi)*
Üye:	*Prof. Dr. Atilla Halil ELHAN* *(Ankara Üniversitesi)*
Üye:	*Prof. Dr. Pınar ÖZDEMİR* *(Hacettepe Üniversitesi)*
Üye:	*Doç. Dr. Jale KARAKAYA KARABULUT* *(Hacettepe Üniversitesi)*

Bu tez, Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

11 Temmuz 2018

Prof. Dr. Diclehan ORHAN

Enstitü Müdürü

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ..... ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

26/07/2018

Duygu AYDIN HAKLI

i

---

[i] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

*(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

*(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

*(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

** Tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Prof.Dr. Erdem KARABULUT danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

26/07/2018

Duygu AYDIN HAKLI

TEŞEKKÜR

Tezin planlanmasında, düzenlenmesinde, sonuçların yorumlanmasında ve eğitimim süresince desteklerini ve bilgilerini esirgemeyen tez danışmanım Prof.Dr. Erdem KARABULUT'a,

Eğitimim boyunca yanımda desteklerini esirgemeyen arkadaşlarım Araş. Gör. Dinçer GÖKSÜLÜK, Araş. Gör. Merve BAŞOL ve Araş Gör. Ebru ÖZTÜRK'e,

Tezimin aşamalarında Tez İzleme Komitesinde beni dinleyen ve yönlendiren Prof. Dr. Atilla H. ELHAN ve Prof. Dr. A. Ergun KARAAĞAOĞLU'na,

Biyoistatistik Anabilim Dalı'nda tüm değerli hocalarıma ve idari çalışanlarına,

Tez çalışmalarım boyunca her zaman yanımda olan ve beni destekleyen eşime sabırları ve sevgileri için çok teşekkür ederim.

ÖZET

Aydın Haklı, D. Sınıf dengesizliği sorununu çözmek için kullanılan algoritmaların farklı sınıflandırma yöntemlerinde performanslarının karşılaştırılması. Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı Bütünleşik Doktora Tezi, Ankara, 2018. Veri setlerinde sınıf dengesizliği problemi bir gruptaki gözlem sayısının diğer gruptaki gözlem sayısınından küçük olması olarak tanımlanmaktadır. Dengesiz veri setlerini makine öğrenme yöntemleri ile analiz etmek son yıllarda yaygın ve dikkate değer bir araştırma alanı konumuna gelmiştir. Ancak bu problemden dolayı model performanslarında bir azalma olmaktadır. Bunun yanı sıra, verinin dağılımı ve verinin yapısı sınıflama için model seçimi, en uygun (optimum) model parametrelerinin elde edilmesi, modelin geçerliğinin etkileyebilmektedir. SMOTE, SMOTEBoost, RUSBoost, MWMOTE, EasyEnsemble, SMOTEBagging ve UnderBagging gibi algoritmalar sınıf dengesizliği probleminin etkisini azaltmak için önerilmiştir. Tez çalışmasında gerçek veri setleri ile birlikte kapsamlı bir benzetim çalışması ile elde edilen veri setleri kullanılarak sınıflama yöntemlerinin performanslarını değerlendirildi. Farklı sınıflama yöntemleri, farklı sınıf dengesizlik algoritmaları, farklı örneklem genişlikleri, farklı korelasyon yapıları ve farklı dengesizlik oranlarını kapsayacak bir benzetim çalışması gerçekleştirildi. Her senaryo 1000 kez tekrarlandı ve 5-kat çapraz geçerlik kullanılarak model doğruluğu sağlandı. Benzetim çalışmasındaki kurulan modellerin performanslarının, örneklem genişliği ve bağımlı-bağımsız değişkenler arasındaki ilişki ile arttığı görüldü. Korelasyon sıfıra yaklaştığında ve dengesizlik çok olduğunda, RUSBoost algoritması diğer algoritmalara göre sonuçlar üzerinde daha etkili bulundu. Veri setleri dengeli hale geldikçe yedi (7) algoritma örneklem genişliğinden ve korelasyon yapısından bağımsız olarak benzer sonuçlar verdi. Genel olarak benzetim çalışması sonucunda, RUSBoost tüm örneklem genişliklerinde, EasyEnsemble ise küçük örneklem genişliklerinde daha iyi sonuç verdi.

Anahtar Kelimeler: Sınıf dengesizliği, Korelasyon yapıları, Sınıflama yöntemleri, Alt örnekleme, Aşırı örnekleme.

ABSTRACT

Aydın Haklı, D. Comparing the performance of the algorithms used to solve class imbalance problem in different methods of classification. Hacettepe University Institute of Health Sciences, PhD. Dissertation in Biostatistics, Ankara, 2018. Class imbalance, for a given dataset, occurs when there are relatively small observations in one or more groups comparing to other groups. Analyzing imbalanced data sets via machine learning algorithms has become a common and remarkable research area in recent years. However, this problem leads to a decrease in the model performance. Besides that, selection of the model for classification, optimizing model parameters, validating the fitted model, underlying distribution and data structure may also affect model performance. Furthermore, several data balancing algorithms were proposed to overcome class imbalance problem such as SMOTE, SMOTEBoost, RUSBoost, MWMOTE, EasyEnsemble, SMOTEBagging and UnderBagging. In this study, we evaluated model performances using a comprehensive simulation study along with real data examples. We conducted a simulation study under different classification models, class imbalance algorithms, sample sizes, correlation structures and class imbalance ratios. Each scenario was repeated 1000 times and the fitted models were optimized using 5-folds cross-validation. Simulation study showed that the model performances increase with sample size and correlation among dependent and independent variables. When the correlation approaches zero and classes are highly imbalanced, RUSBoost outperforms other algorithms. As data become more balanced, the seven algorithms gave similar results independently from sample size and correlation structure. Overall simulation results, RUSBoost algorithm provided better result for all sample sizes and EasyEnsemble for small sample size the most of the simulation combinations.

Key Words: Class imbalance, Correlation structure, Classification methods, Undersampling, Oversampling.

İÇİNDEKİLER

SİMGELER ve KISALTMALAR

CART	Sınıflama ve Regresyon Ağaçları
CHAID	Otomatik Ki-Kare Etkileşim Belirleyicisi
DDO	Düzeltilmiş Doğruluk Oranı
DN	Doğru Negatif
DO	Dengesizlik oranı
DP	Doğru Pozitif
DUY	Duyarlılık
DVM	Destek vektör makineleri
EAA	Eğri altında kalan alan
GDO	Genel doğruluk oranı
k	Değişken sayısı
n	Örneklem genişliği
NKD	Negatif kestirim değeri
OOB	Ağaç oluşturmayacak veri
PKD	Pozitif kestirim değeri
r	Korelasyon katsayısı
RF	Random forest
ROS	Rastgele aşırı örnekleme
RUS	Rastgele alt örnekleme
SEÇ	Seçicilik
SH	Standart hata
SMOTE	Sentetik azınlık aşırı örnekleme
x_i	Bağımsız değişkenler
y	Bağımlı değişken
YN	Yanlış Negatif
YP	Yanlış Pozitif

ŞEKİLLER

TABLOLAR

1 GİRİŞ

Son yıllarda, dengesiz veri setlerinin makine öğrenmesi yöntemleri kullanılarak analiz edilmesi dikkate değer bir araştırma alanı haline gelmiştir. Çünkü, gerçek yaşamda dengesiz veri setleri ile karşılaşılması daha olasıdır. İki sınıflı bir veri setinde sınıflardan birinde diğerine göre daha az sayıda gözlem varsa, sınıf dengesizliği problemi ortaya çıkmaktadır (1–8). Bu problem ile Sağlık Bilimlerindeki nadir görülen hastalıklar, kanser araştırmaları, vb. sıklıkla karşlaşılır. Örneğin kesitsel araştırmalarda dengesizlik oranı (olgu grubundaki gözlemlerin sayısının toplam gözlem sayısına oranı) %5, %10 veya %15 olabilmektedir (3, 9). Daha yüksek gözlem sayısına ait sınıf "çoğunluk sınıfı", diğer sınıf ise "azınlık sınıfı" olarak adlandırılmaktadır (9). Azınlık sınıfındaki gözlem sayısı azaldıkça dengesizlik artmaktadır.

Dengesiz veri setleri ile kurulan bazı makine öğrenmesi yöntemlerinde düşük model doğrulukları ve aşırı uyum (veya yetersiz uyum) gibi çeşitli problemler ortaya çıkmaktadır (4, 9–12). Çoğunluk sınıfındaki gözlem sayısı azınlık sınıfı gözlem sayısından fazla olduğu için kurulan modelde çoğunluk sınıfı gözlemlerinin etkisi daha fazla olmaktadır. Bu nedenle, azınlık sınıfına ait bazı gözlemler yanlış sınıflandırılır (1, 4, 6, 8). Azınlık sınıfına ait gözlemlerin yanlış sınıflandırılması modelin doğruluğunu düşürmektedir. Sınıflama yöntemleri sonucunda model performansını ölçmek için çeşitli ölçüler geliştirilmiştir. Genel doğruluk oranı, duyarlılık, seçicilik, eğri altında kalan alan, F-ölçüsü, vb. bu ölçülerden bazılarıdır. Dengesizlik çok yüksekse (%5-%10) bazı performans ölçüleri çok küçük çıkma eğilimindedir. Dengesiz verilerin doğrudan kullanılması sınıflama yöntemlerinin doğruluğunu azaltır. Alanyazında (literatürde), dengesizliği azaltacak veya etkisini ortadan kaldıracak birkaç yaklaşım vardır. Bunlardan üçü,

- Sentetik (yapay) veri üretme,
- Az örnekleme (undersampling)
- Aşırı örnekleme (oversampling)

Birinci yaklaşımda eldeki veriler kullanılarak yeni gözlemler üretilmektedir (1, 5). İkinci yaklaşımda, çoğunluk sınıfından rastgele gözlemler çıkarılarak sınıflardaki gözlem sayıları dengelenmektedir (4, 6, 7, 13). Son yaklaşımda ise, azınlık sınıfındaki gözlemler yeniden örnekleme ile rastgele birden fazla kere seçilerek sınıflardaki gözlem sayıları dengelenmektedir (1, 4, 6, 7, 13). Alanyazında, sınıf dengesizliği problemini aşabilmek için yukarıda ki yaklaşımlardan yararlanılarak çeşitli algoritmalar önerilmiştir.

Bu problemin üstesinden gelebilmek için Chawla ve diğ. 2002 yılında sentetik gözlemler üreten sentetik azınlık aşırı örnekleme (Synthetic Minority Oversampling Technique-SMOTE) algoritmasını önermiştir (1). Önerilmiş başka bir algoritma SMOTEBoost ise SMOTE algoritması ve boosting yöntemlerinin birleştirilmesi ile oluşmaktadır (5). Boosting yöntemi Freud (1990) tarafından önerilen bir yöntemdir (14). Boosting yönteminde, gözlemlere başlangıçta eşit ağırlık verilmekte, sonra yanlış sınıflanan gözlemlere daha fazla ağırlık verilmektedir. SMOTEBoost algoritması, her gözlemle ilişkili ağırlıkları güncelleyerek eğitim verilerinin dağılımını değiştirmek yerine SMOTE algoritmasını kullanarak yeni azınlık sınıfı gözlemleri ekleyerek dağılımı değiştirir. Ancak bu yöntem zaman alıcı bir yöntemdir. Seiffert ve diğ. Rastgele Az Örnekleme (RUS) ve boosting algoritmalarının birleştirilmesi ile oluşan RUSBoost algoritmasını önermişlerdir (6). Hem RUSBoost hem de SMOTEBoost algoritmalarının temelinde boosting vardır. RUSBoost algoritmasında modelin eğitim süresi daha kısa olduğu için SMOTE ve SMOTEBoost'a göre daha fazla tercih edilmektedir. Barandela ve diğ. (2003) tarafından yazılan makalede UnderBagging algoritması tanıtılmıştır (15). Bu algoritma rastgele az örnekleme ile bagging algoritmalarını birleştirmektedir. DataBoost-IM algoritmasını öne süren Hongyu ve diğ. (2004) sentetik aşırı örnekleme ile boosting yöntemini birleştirmişlerdir (16). Bu yöntemde, hem azınlık hem de çoğunluk sınıfı için yeni sentetik gözlemler oluşturulurken, azınlık sınıfı için daha fazla gözlemler oluşturulur. Han ve diğ. (2005) iki sınıf arasındaki ayrımda sınıra yakın gözlemlerin yanlış sınıflanması üzerine çalışmışlar ve Borderline-SMOTE algoritmasının bu soruna çözüm olabileceğini belirtmişlerdir (17). 2008 yılında He ve diğ. (2008), azınlık sınıfının dağılımından yararlanarak örnekler üretebilmek için ADASYN (Adaptive Synthetic Sampling Approach) algoritmasını oluşturmuşlardır (18). Liu ve diğ. (2009) iki yeni algoritma önermişlerdir (4). Bunlar Easy Ensemble ve Balance Cascade algoritmalarıdır. İlk yaklaşımda, çoğunluk sınıfından birden fazla alt küme örnekleri oluşturulup, her biri kullanılarak öğrenici eğitilir ve bu öğrenicilerin çıktıları birleştirilir. Diğer yöntemde ise, eğitim setinde, her adımda, çoğunluk sınıf örneklerini doğru bir şekilde sınıflandıran eğitilmiş öğreniciler değerlendirmeden çıkarılır ve işlemler tekrar eder. Bagging içeren başka bir algoritmada Wang ve diğ. (2009) tarafından önerilen OverBagging algoritmasıdır. Bu algoritma rastgele aşırı örnekleme ile bagging algoritmalarını birleştirmektedir (19). Chen ve diğ. (2010) eğitim verisindeki azınlık sınıfı gözlemlerini çevreleyen en yakın komşularını dikkate alan sentetik gözlemler üretebilecek RAMOBoost algoritmasını önermişlerdir (20). Barua ve diğ. (2012) tarafından önerilen MWMOTE algoritması azınlık sınıfından

seçilen örneklere ağırlıklar atayıp sentetik veri üretmektedir (21). Hanifah ve diğ. (2015) modelin performans başarısını arttırmak için SMOTEBagging algoritması önermişlerdir (22). Bu yöntemde SMOTE algoritması ile bagging yöntemlerini birleştirmişlerdir.

Bu tez çalışmasında, sınıf dengesizliğinin etkisini kaldırmak için yedi algoritma kullanılmış ve daha sonra makine öğrenmesi yöntemlerinin performansları karşılaştırılmıştır. Bu sınıflandırma yöntemleri Destek Vektör Makinaları (DVM), Sınıflama ve Regresyon Ağaçları (CART) ve Random Forest (RF) (10, 23, 24). Alanyazında, DVM doğrusal olarak ayrılamayan sınıfları ayırmak için önerilmiş iyi bilinen yöntemlerden biridir (25). Sınıflandırma, makine öğrenmesinin önemli bir parçasıdır. Sınıflandırmada veriler eğitim seti ve test seti olmak üzere iki parçaya ayrılır. Model kurma işlemi eğitim setinde yapılır ve çapraz geçerlilik, bootstrap vb. gibi tekniklerle model sonuçlarının genelleştirilmesi sağlanır. Sonra test seti yardımıyla kurulan modelin performansı bulunur (10).

Bu tez kapsamında, SMOTE (1), SMOTEBoost (SMOTE + boosting) (5), RUSBoost (Rastgele Az Örnekleme + boosting) (6), MWMOTE (21), EasyEnsemble (4), SMOTEBagging (SMOTE + bagging) (26) ve UnderBagging (Az örnekleme + Bagging) (15) algoritmaları ele alınmıştır.

Alanyazına bakıldığında birçok çalışma yeni algoritma önermektedir. Ancak önerilen bu algoritmaların çoğunluğunun başarısı gerçek veri setleri kullanılarak değerlendirilmiştir. Bunların daha farklı bir yapıya sahip veri setlerinde kullanılıp kullanılamayacağı konusu üstünde çalışma bulunmamaktadır.

Bu tez kapsamında, iki sınıflı verilerin dengesiz olduğu durumlarda model performansını etkileyebilecek çeşitli faktörler incelenmiştir. Bu faktörler, örneklem genişliği, farklı korelasyon yapıları ve farklı dengesizlik oranlarıdır. Üç faktörün yer aldığı bir benzetim (simülasyon) çalışması ile sınıflama yöntemlerinin performansları karşılaştırılmıştır.

1.1 Amaç

Bu tez çalışmasının amacı, sınıf dengesizliği sorunu olduğunda, çözüm algoritmalarının veri madenciliği yöntemlerinin performansları üzerine etkisini incelemektir. Ayrıca, bu algoritmaların performansı üzerinde dengesizlik oranının, bağımlı değişkenle bağımsız değişkenler arasındaki korelasyon miktarının ve gözlem sayısının etkisini incelemektir.

Tez çalışmamızın hipotezleri aşağıda sıralanmıştır:

- Sınıf dengesizliği olduğunda, kullanılan çözüm algoritmaları sınıflama yöntemlerinin başarısını arttırır.

- Dengesizlik oranına göre çözüm algoritmalarının performansları farklıdır.
- Korelasyon yapılarına göre çözüm algoritmalarının performansları farklıdır.

1.2 Tez organizasyonu

Tezin genel organizasyonu:

Bu bölümde, sınıf dengesizliği problemi ve alanyazında bu problem için yapılmış çalışmalardan bahsedilmiştir.

Genel Bilgiler: Alanyazında önerilen yöntemler bu bölümde ayrıntılı olarak verilmiştir. Ayrıca, sınıflama yöntemleri ile de ilgili bilgi verilmiştir.

Gereç ve Yöntem: Bu bölümde kapsamlı bir benzetim çalışması tasarlanmıştır. Ayrıca, gerçek veri setleri de kullanılmıştır. Bu veri setlerinin özellikleri incelenerek analizler yapılmıştır.

Bulgular: Benzetim ve gerçek veri setlerine ait sonuçlar, grafikler ve tablolar ile birlikte bu bölümde verilmiştir.

Tartışma: Elde edilen sonuçlarla alanyazında bulunan diğer çalışmalar tartışılmaktadır.

Sonuç: Tezin önemli genel sonuçları verilmiştir.

2 GENEL BİLGİLER

2.1 Veri Madenciliği

Dünyada birçok yöntemlerle elde edilen veriler hızlı bir şekilde artmakta ve bu verilerin toplanıp saklanması gibi bazı sorunlar ortaya çıkmaktadır. Günümüzde çok kısa sürede büyük miktarlara ulaşan veriler herhangi bir araç kullanılmadan etkin bir biçimde analiz edilememektedir. Toplanan veri miktarları arttıkça verileri daha iyi inceleyecek analiz yöntemlerine gereksinim duyulmaktadır. Örneğin bir hastanede tutulan hasta kayıtları, yeni hastalar için tanı konması, hastanede kaç gün yatacağı veya aynı tanıdan dolayı tekrar hastaneye gelip gelmeyeceği konusunda bilgi sahibi olunmasına veri madenciliği yöntemleri kullanarak ulaşmak olanaklıdır. Veri madenciliği, veri tabanlarında bilgi keşfi sürecinde modelin kurulması ve değerlendirmesini kapsamaktadır. Birçok kaynaktan elde edilen verilerin toplanması, verilerin temizlenmesi, bütünleştirilmesi, veri madenciliği veri tabanlarında bilgi keşfi sürecinin parçalarını oluşturmaktadır (27–29). Veri madenciliği aşamaları:

- *Problemin Tanımlanması:* Problemin tanımlanması veri madenciliği aşamalarının başında gelmektedir. Probleme uygun hangi analizin yapılacağının anlaşılması gerekmektedir. Eğer problem doğru bir şekilde ortaya konmazsa yapılan diğer aşamalar yanlışlıklar üzerine kurulmuş olacak ve elde edilen sonuçlar yanıltıcı olacaktır.

- *Veriyi anlama:* Verinin toplanması ile başlayan bu aşama, veride hangi değişkenlerin olduğu, bu değişkenlerin neleri ifade ettiklerini anlamayı barındırmaktadır.

- *Veri Hazırlama:* Veri hazırlama aşaması, toplanan veriler içerisinden değişken seçimi, veri temizliği, yeni değişkenler oluşturma, gerekiyorsa uygun dönüşümler yapma gibi işlemlerden oluşmaktadır.

- *Verinin Modellenmesi:* Uygun model seçimi analiz sonucunda elde edilecek sonuçları etkilemektedir. Danışmanlı ve danışmansız öğrenmenin kullanımına göre sonuçlar farklılık göstermektedir.

- *Değerlendirme:* Modelin uygulama aşamasına geçmeden bu aşamada kurulan modelin uygunluğu değerlendirilmektedir.

- *Uygulama:* Bu aşamada kurulan ve geçerliliği kabul edilen modelin kullanılması aşamasıdır. Kurulan modelin zaman içerisinde izlenilmesi ve ortaya çıkan değişiklikler karşısında modelin güncellenmesi gerekmektedir.

Veri madenciliğinde yöntemler tanımlayıcı ve kestirici olmak üzere iki gruba ayrılmaktadır. Tanımlayıcı modeller, veri setindeki değişkenler arasındaki ilişkileri özetlemektedir. Kestirici modelleri ise çeşitli yöntemler ile elde edilen modeller ve bu modeller yardımıyla kestirimler yapmayı amaçlamaktadır.

2.2 Makine Öğrenmesi

Veri madenciliğinde kullanılan yöntemlerin içerisinde öğrenme yer almaktadır. Genellikle veri kullanılarak olayların girdi ve çıktıları arasında ilişkiler öğrenilir. Öğrenilen bilgiler gelecekteki benzer olaylar için kullanılır ve karar verilir. Öğrenme temel olarak iki şekilde olabilir: Danışmanlı ve danışmansız öğrenme (25, 27, 29).

2.2.1 Danışmansız Öğrenme

Danışmansız öğrenme yöntemlerinde, sınıf bilgisine gerek yoktur. Bir veri kümesindeki yapıları bulmayı hedefler. Bu yöntemlere kendi kendine öğrenebilen modeller de denilmektedir. Kümeleme Analizi, birliktelik analizi, faktör analizi vb. yöntemler danışmansız öğrenme yöntemlerindendir (25, 27, 29).

2.2.2 Danışmanlı Öğrenme

Danışmanlı öğrenmede ise, veri setindeki gözlemlerin sınıfı önceden belirlidir. Yeni gelen verinin, eğitim setinde kurulan model yardımı ile sınıfı belirlenir. Danışmanlı öğrenmede, sınıflama ve regresyon yöntemleri bulunmaktadır. Sınıflama yöntemleri, gözlemlerin ait oldukları sınıfları belirlemede kullanılırken, regresyon belirli değerlere ilişkin kestirim yapılmasında kullanılır. Sınıflama yöntemleri; karar ağaçları (CART, CHAID, vb.), random forest, naive bayes, k-en yakın komşu, yapay sinir ağları, genetik algoritmalar, destek vektör makineleridir (25, 27, 29). Bu tez çalışmasında, danışmanlı öğrenme yöntemlerinden Destek Vektör Makineleri (DVM), Sınıflama ve Regresyon Ağaçları (CART) ve Random Forest (RF) yöntemleri kullanılmıştır.

2.3 Veri Madenciliği Yöntemleri

2.3.1 Tanımlayıcı Yöntemler

Tanımlayıcı modellerde karar vermeye yardımcı olabilecek bilgiler elde edilir.

Bunlardan bazıları aşağıdaki gibidir:

- Birliktelik kuralları
- Kümeleme
- Uç/aykırı değer analizi
- Tanımlayıcı istatistikler

2.3.2 Kestirici Yöntemleri

Sonucu bilinen gözlemlerden yararlanılarak oluşturulan model ve kurulan bu model yardımı ile sonuçları bilinmeyen verilere ilişkin kestirim yapılması amaçlanmaktadır. Kestirici modeller, sonuç değişkenine bağlı olarak sınıflama ve regresyon olmak üzere ikiye ayrılmaktadır. Bazı kestirici yöntemler aşağıdaki gibidir:

- Karar ağaçları
- Random forest
- Yapay sinir ağları
- Naïve-Bayes
- K-En yakın komşuluk
- Genetik algoritmalar
- Destek vektör makineleri

Bu tez kapsamında kestirici yöntemlerden sınıflama yöntemleri olarak karar ağaçları, random forest ve destek vektör makineleri seçilmiş ve bu yöntemler sonraki bölümde detaylı olarak ele alınmıştır.

2.4 Sınıflama Yöntemleri

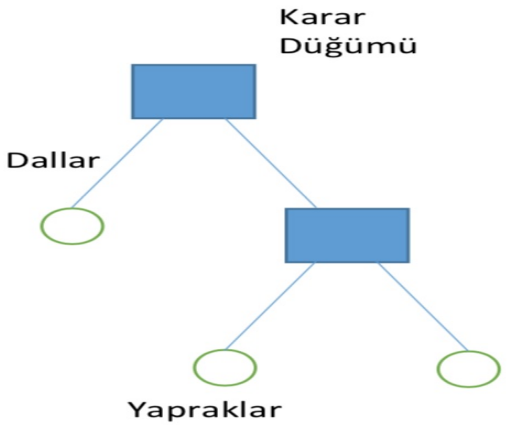
2.4.1 Karar Ağaçları

Karar ağaçları, kolay oluşturulabilen ve kolay anlaşılır olan ağaç görünümünde bir kestirim yöntemidir. Aşırı değerlere karşı dayanaklı olan bu yöntem çok sayıda gözlem ve çok sayıda değişkene sahip veri setlerinde uygulanabilir bir yöntemdir. Karar ağaçları;

- Karar düğümü,

- Dallar,
- Yapraklardan

oluşur. Karar düğümleri yapılacak olan testi, dallar testteki değerleri, yapraklar ise sınıfı belirtir. Karar düğümleri ile belirtilen testin sonucu ağacın veri kaybetmeden dallara ayrılmasını sağlar. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer dalın bir ucunda sınıflama işlemi gerçekleşemiyorsa, o dalın sonucunda bir karar düğümü oluşur. Ancak dalın sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi, kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana kadar ardışık düğümleri takip ederek gerçekleşir. Şekil 2.1'de basit bir karar ağacı örneği yer almaktadır (25, 27, 29, 30).



Şekil 2.1. Karar ağacı şeması

Birçok sınıflama yönteminde ortaya çıkan aşırı uyum (overfitting) durumu, karar ağaçlarında da ortaya çıkmaktadır. Aşırı uyum, eğitim veri seti yardımı ile elde edilen modelin yeni bir veri seti ile karştılaştığında aynı performansı sergilememesi durumudur. Bu durumda budama (pruning) yaparak modelin veri setini ezberlemesinin önüne geçilebilmektedir. Budama, ön budama ve sonradan budama olmak üzere iki şekilde yapılabilmektedir. Ön budama, dallarda en az kaç gözlem olacağı veya ne kadar dallanacağı, vb. gibi sınırlamalar ile ağacın oluşmasına önceden müdahale etme olarak adlandırılmaktadır. Sonradan budama ise, ağaç tamamen oluştuktan sonra planlı bir şekilde modele katkı sağlamayan dalların belirlenerek modelden çıkarılması işlemidir (24, 25, 29, 31).

Karar ağacı algoritmalarında bölünmenin başlayacağı değişken önemlidir.

Bilgi kazancı (information gain), kazanç oranı (gain ratio) ve gini indeksi (gini index) seçim kriterleri ile bölünmenin başlayacağı değişken seçilmektedir (25, 29). *Bilgi Kazancı (Information gain):* Beklenen bilgi (entropi) bilgi kuramı içerisinde yer alan temel kavramlardan biridir ve bir rastgele değişkenin entropisi, rastgele değişkenin belirsizliğinin ölçüsü olarak tanımlanır. p_i, D veri setindeki bir bireyin C_i sınıfına ait olma olasılığı ise, beklenen bilgi Eşitlik 2.1 ile hesaplanır. Beklenen bilgi 0 ve 1 aralığında değerler almaktadır. 1 değerine yaklaştıkça belirsizlik artar. Yüksek beklenen bilgi değerine sahip değişken daha çok bilgi içerir.

$$Bilgi(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{2.1}$$

D veri setinde A değişkenine göre v parçaya bölündükten sonra elde edilen bilgi Eşitlik 2.2 ile elde edilir.

$$Bilgi_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Bilgi(D_j) \tag{2.2}$$

A değişkenine göre elde edilen bilgi kazancı ise Eşitlik 2.3 ile elde edilir ve bölünme, en yüksek bilgi kazancı olan değişkenden başlar (25).

$$\text{Bilgi kazancı}(A) = Bilgi(D) - Bilgi_A(D) \tag{2.3}$$

Kazanç Oranı (Gain ratio): Bilgi kazancı, çok fazla değişkenin olduğu durumlarda yanlı bölünme yapabilmektedir. Bunun için kazanç oranı önerilmektedir. D_j, değişkene ait değerlerin tekrarlanma sayısı ve D ele alınan olay sayısı ise Bölünme bilgisi Eşitlik 2.4'e göre bulunmaktadır.

$$\text{Bölünme Bilgisi}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right) \tag{2.4}$$

Bölünme bilgisinden yararlanılarak kazanç oranı Eşitlik 2.5 ile elde edilir. Bölünme, en yüksek kazanç oranına sahip değişkenden bölünme başlamaktadır (25).

$$\text{Kazanç oranı} = \frac{Bilgi(A)}{\text{Bölünme Bilgisi}(A)} \tag{2.5}$$

Gini İndeksi (Gini index): D veri seti n sınıf içeriyorsa, p_i, i sınıfının D veri setinde

görülme sıklığı ise gini indeksi Eşitlik 2.6 ile hesaplanır (25).

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.6}$$

D veri seti A değişkenine göre D_1 ve D_2 olmak üzere ikiye bölünüyorsa, A değişkenine göre gini indeksi Eşitlik 2.7 ile hesaplanır.

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{2.7}$$

A değişkenine göre yapılan ikili bölünme sonrası belirsizlikteki indirgeme Eşitlik 2.8 ile hesaplanır.

$$\triangle Gini(A) = Gini(D) - Gini_A(D) \tag{2.8}$$

Bölünme, en büyük gini indeksi değerine sahip değişkenden başlamaktadır.

Karar ağaçlarında kullanılan bazı yöntemler CART, CHAID ve C4.5'tir. Kullanılan yönteme göre ağacın şekli değişebilmektedir.

CART (Sınıflama ve Regresyon Ağaçları-Classification and Regression Tree)

Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında geliştirilmiştir. CART nicel veya nitel bağımlı değişkeni, nicel veya nitel bağımsız değişkenler yardımı ile kestirim yapmaya yarayan bir yöntemdir. Kullandığı bilgi ölçütü "gini indeksi" dir. Girdi değişkenler ağacın bölünme aşamasında sadece 2'ye bölünebilir. Bu nedenle fazla sayıda kategori içeren girdi değişken varlığında dezavantaja sahiptir (25, 29).

Bu tez kapsamında Karar ağaçlarından CART yöntemi kullanılmıştır.

CHAID (Otomatik Ki-Kare Etkileşim Belirleme-Chi-squared Automatic Interaction Detection)

CHAID yöntemi, 1980 yılında G.V. Kass tarafından geliştirilmiştir. Bağımlı değişkeni en fazla etkileyen bağımsız değişken seçiminde; bağımlı değişkenin kategorik olması durumunda Ki-Kare testi kullanılırken, sürekli olması durumunda F testi kullanılır. CHAID yöntemi, kestirim için değişkenlerin tüm değerlerini dikkate almaktadır. Tüm girdi ve sonuç değişkenle çapraz tablolar oluşturulur ve en fazla anlamlı olan değişkenden en az anlamlı olan değişkene doğru ağaç bölünmeye başlar. Bu yöntem yardımı ile iki veya daha fazla bölünme yapı-

labilmektedir (25, 29).

C4.5

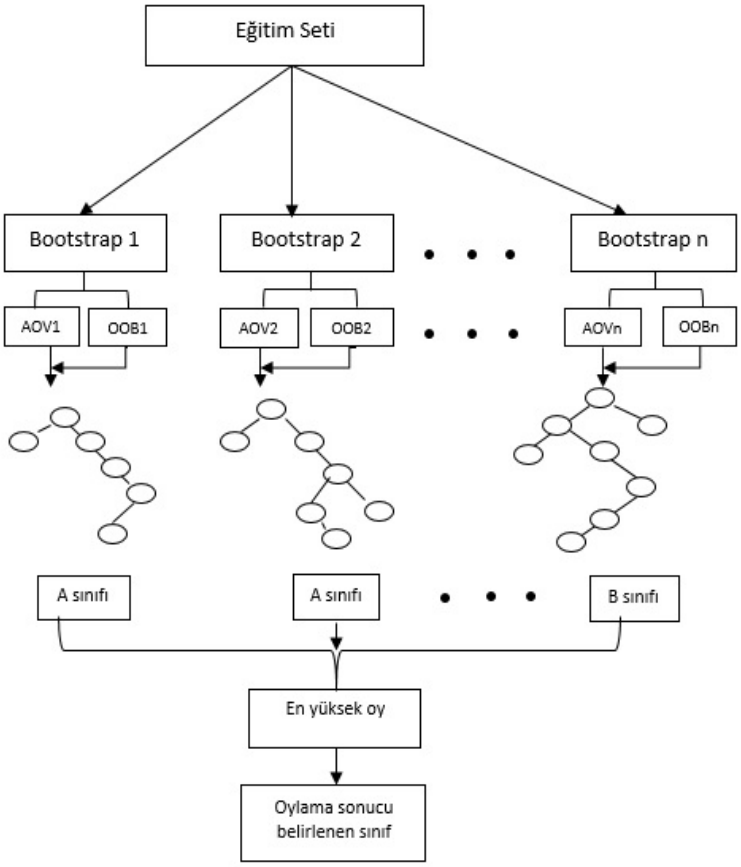
ID3 yöntemi Quinlan tarafından 1986 yılında geliştirilmiştir. ID3 yöntemi, bilgi kazancından yararlanarak önemli gördüğü değişkeni belirler. Eğitim veri setindeki değişkenlerin bilgi kazancı değerleri hesaplanır ve en yüksek bilgi kazancı olan değişkenden bölünme yapılır. C4.5, ID3 yönteminin gelişmiş hali olduğunu söylebiliriz. 1993 yılında Quinlan tarafından geliştirilmiş olan C4.5 yönteminde kayıp veriler analize alınmamaktadır. Kazanç oranı (gain ratio) kullanılarak ağaç dallanmaktadır. En yüksek kazanç oranına sahip değişkenden dallanma başlamaktadır (25, 29).

2.4.2 RF (Random Forest)

RF, çok sayıda karar ağacının bir araya gelmesiyle oluşan bir yöntemdir. Bu yöntemle hem regresyon hem de sınıflama yapılabilmektedir. Veri setindeki bağımlı değişken nitel ise sınıflama, nicel ise regresyon ağaçları kurulmaktadır. Her bir karar ağacı bootstrap tekniği ile hem veri setinden örneklem çekilmesi hem de her bir karar ağacı her karar düğümünde tüm değişkenler içinden belirlenen sayıda rastgele değişkenin seçilmesi ile oluşturulmaktadır (32).

RF yönteminde, ağaçlar CART yöntemi ile oluşturulur ve budama yapılmaz. Tüm veri eğitim ve test seti olarak ikiye ayrılır. Bootstrap tekniği ile ağaç oluşturacak veri (AOV-inbag) ve ağaç oluşturmayacak veri (out of bag-OOB) olmak üzere eğitim veri setinden örneklemler seçilir. Eğitim setinin 2/3'ü ağaç oluşturacak veri ve 1/3'ü ağaç oluşturmayacak veri olarak ayrılır. Ağaç oluşturulacak veri ile tüm değişkenler içinden her bir düğümde m tane değişken seçilir ve gini indeksi kullanılarak en iyi ayrılmalar sağlanır. Ağaç oluşturulmayan veri yardımı ile kurulan modelde kestirimler yapılır ve kestirim hataları hesaplanır. Her bir ağacın yaptığı OOB kestirimleri birleştirilerek karar ağacının hatası kestirilir. Her bir ağaca OOB hata oranına göre bir ağırlık verilir ve en düşük hata oranına sahip ağaç en yüksek ağırlığı alırken en yüksek hata oranına sahip ağaç en düşük ağırlığı almaktadır (25, 32, 33).

Sınıflama yapan her bir karar ağacı bireysel oy almakta ve işlem sonunda en yüksek oyu alan karar ağacının yaptığı sınıflama kullanılmaktadır. Her bir karar ağacı eğitildiği veri grubundan farklı bir veri grubuyla karşılaştığında aynı performansı gösteremeyeceği için, yöntem çok sayıda karar ağacını birleştirmekte ve bu sayede sınıflama performansını ve doğru sınıflama oranını artırmaktadır. RF yönteminde ağaç, bütün verinin oluşturduğu tek bir düğümle başlamakta, eğer



Şekil 2.2. Random forest şeması

örneklerin hepsi aynı sınıfa ait ise düğüm, yaprak olarak sonlanmakta ve sınıf etiketi verilmektedir. Eğer örnekler aynı sınıfa dahil değilse, örnekleri sınıflara en iyi bölecek olan özellik seçilmektedir. Şekil 2.2'de RF yöntemine ait basit bir gösterim yer almaktadır.

RF yönteminin avantajlarından birisi de değişkenler arasında önem derecesini belirlemesidir. Önem derecesi belirlenirken aşağıdaki adımları gerçekleştirilmektedir:

- Karar ağacı oluşturulduktan sonra OOB'ye göre kestirilen sınıflar yukarıdan aşağıya doğru yerleştirilir ve doğru sınıflama sayısı kaydedilir.
- OOB'deki m. değişkenin değerlerinin yeri değiştirilir ve artık değiştirilmiş OOB olur.
- Değiştirilmiş OOB değerleri daha önceden oluşturulan karar ağacı üzerinde yukarıdan aşağıya doğru yerleştirilir ve doğru sınıflama sayısı hesaplanır.
- OOB doğru sınıflama sayısından değiştirilmiş OOB doğru sınıflama sayısı çıkarılır ve d_i hesaplanmış olur.

- Yukarıda sayılan tüm işlemler her bir karar ağacı için yapılır ve elde edilen d_i'lerin ortalaması alınır ve $\bar{d}$ ile ifade edilir.

- Bulunan d_i değerlerinin standart hatası (SH) hesaplanır ve aşağıdaki Eşitlik 2.9 yardımı ile önem derece skoru elde edilmiş olur.

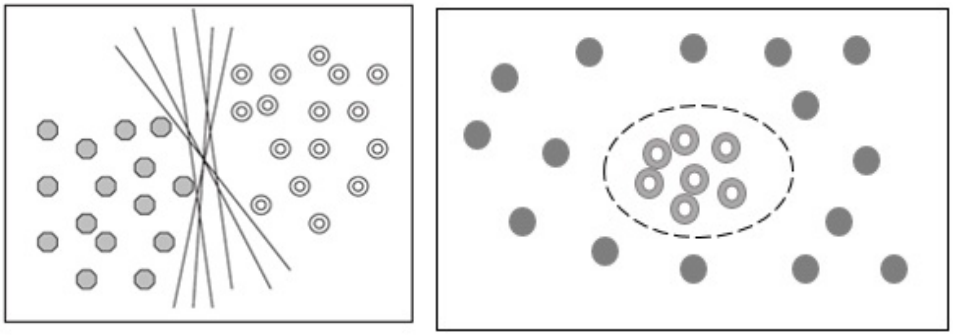
$$\text{Önem Derece Skoru} = \frac{\bar{d}}{SH_{d_i}} \tag{2.9}$$

Anlatılan adımlar her bir değişken için tek tek uygulanır. Böylelikle her bir değişkene ait önem derece skorları hesaplanmış olunur.

Değişken önem dereceleri için diğer bir yöntem ise gini değerlerinin yardımı ile hesaplanmaktadır. m. değişkenden dallara ayırma gerçekleşmeden önceki gini indeksi değerleri ile dallara ayrım yapıldıktan sonraki gini değerleri arasındaki fark alınır. Tüm ağaçlar için bu değer hesaplanır ve elde edilen değerler toplanır. Tüm değişkenler için bu değer hesaplanır ve buradan önem dereceleri hesaplanmış olunur.

2.4.3 DVM (Destek Vektör Makineleri-Support Vector Machine)

İlk kez 1960'lı yılların sonunda Vladimir Vapnik ve Alexey Chervoenkis tarafından geliştirilmiştir. Ancak bu yöntemin kullanımı 1992 yılında yapılan yayında sınıflama başarısının yüksek olduğu gösterildikten sonra yaygınlaşmıştır. Güçlü bir kuramsal temele sahip olması, büyük veri setleri üzerinde çalışabilmesi, çekirdek fonksiyonlar ile esnek bir yapıda olması ve sonuçların yüksek doğrulukta olması bakımından diğer yöntemlere göre daha fazla tercih edilmektedir. DVM yöntemi, son yıllarda özellikle veri madenciliğinde değişkenler arasındaki örüntülerin bilinmediği veri setlerindeki sınıflama problemleri için sıklıkla kullanılmaktadır. Sınıfları birbirinden ayıran marjini en büyük, doğrusal bir ayrıt edici fonksiyon bulmayı amaçlamaktadır. Bu yöntem, temelde iki sınıflı problemlerin çözümünde doğrusal bir sınıflayıcı olarak düşünülmüş, daha sonra doğrusal olarak ayrılamayan veya çok sınıflı sınıflama problemlerinin çözümüne de genelleştirilmiştir (23, 30). Şekil 2.3a'de doğrusal olarak ayrılabilen ve Şekil 2.3b'de doğrusal olarak ayrılamayan örnekler görülmektedir.



(a) Doğrusal olarak ayrılabilen durum **(b)** Doğrusal olarak ayrılamayan durum

Şekil 2.3. DVM için doğrusal ayrılabilen ve ayrılamayan durumlar

İki sınıfı ayırmak için sonsuz sayıda doğru çizilebilir. DVM yönteminin amacı, iki sınıfı birbirinden ayırırken sınıflama hatasını en küçük yapacak hiperdüzlemi seçmektir. Bu hiperdüzlem maksimum marjinli hiperdüzlem tekniği ile bulunmaktadır. Eşitlik 2.10 'de hiperdüzleme ait denklem yer almaktadır.

$$f(x) =< w, x > +b \tag{2.10}$$

w hiperdüzlemin ağırlık vektörü ve b yanlılığı göstermektedir. Eğitim verileri n boyutlu x vektörleriyle temsil edilmektedir. $< w, x > +b = 0$ hiperdüzlemi üzerindeki noktalar x vektörüne aittir. $< w, x >$ ifadesi iç çarpımı göstermektedir. Hiperdüzlem ile marjin üzerindeki herhangi bir nokta arasındaki uzaklık $\frac{1}{\|w\|}$ ile ifade edilmektedir. Marjini en büyük yapmak için uygun w ve b değerleri seçilir.

$$\begin{aligned} w.x + b &= +1 \qquad y_i = +1 \\ w.x + b &= -1 \qquad y_i = -1 \end{aligned} \tag{2.11}$$

Eşitlik 2.11 bir optimizayon problemidir ve w değerinin mutlak değerine bağlıdır. Bu durum konveks olmayan optimizasyon olarak adlandırılmaktadır. Çözümü kolaylaştırmak için $\|w\|$ yerine $\frac{1}{2}\|w\|^2$ kullanılabilmektedir. ξ_i ve ξ_i^* esnek (slack) değişkenler eklenerek optimizasyonda zorluk çıkaran kısıtlar da dikkate alınmaktadır. Optimizasyon Eşitlik 2.12 ve kısıtlar (Eşitlik 2.13) ile birlikte kuadratik bir optimizasyon problemine dönüşmektedir.

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.12}$$

$$
\begin{aligned}
& y_i - < w, x > -b \leq \epsilon + \xi_i \\
& < w, x > +b - y_i \leq \epsilon + \xi_i^* \\
& xi_i, \xi_i^* \geq 0
\end{aligned} \tag{2.13}
$$

Doğrusal olarak ayrılamayan verileri sınıflandırmak için, veri çeşitli yollarla çok boyutlu uzaya taşınır ve burada en iyi ayırıcı hiperdüzlem bulunarak veri sınıflandırılır. Farklı çekirdek fonksiyonlar yardımı ile çok boyutlu uzayda sınıflama yapılmaktadır. Tablo 2.1'de bazı çekirdek fonksiyonlar ve matematiksel ifadeleri yer almaktadır (23).

Tablo 2.1. DVM yöntemine ait bazı çekirdek fonksiyonlar

Çekirdek Fonksiyonlar	
Doğrusal	$K\left(x_i, x_j\right) = x_i^T x_j$
Polinomiyal	$K\left(x_i, x_j\right) = \left(1 + x_i^T x_j\right)^d$
Radyal Tabanlı	$K\left(x_i, x_j\right) = exp\left(\frac{\|x_i - x_j\|^2}{2^\gamma}\right)$
Sigmoid	$K\left(x_i, x_j\right) = tanh\left(kx_i^T x_j - \delta\right)$

2.5 Sınıf Dengesizliği Problemi

Bir veri setinde sınıf değişkeninde gruplardan birinde bulunan gözlem sayısı diğer grupta bulunan gözlem sayısından önemli ölçüde fazla ya da az olduğu zaman sınıf dengesizliği problemi ortaya çıkmaktadır. Gözlem sayısının az olduğu gruba "azınlık sınıfı" denirken diğer gruba "çoğunluk sınıfı" denilmektedir.

Sınıf dengesizliği probleminin çözümü için alanyazında önerilmiş birçok yöntem vardır. Bu yöntemlerin temel aldığı yapılar aşağıda sıralanmıştır.

- Sentetik (yapay) veri oluşturma
- Çoğunluk sınıfından azınlık sınıfı kadar örnekler seçme (az örnekleme - undersampling)
- Azınlık sınıfından yeni azınlık sınıfı (yerine koyarak) oluşturma (aşırı örnekleme - oversampling)

Tez çalışmamızda SMOTE (1), SMOTEBoost (5), RUSBoost (6), MWMOTE (21), EasyEnsemble (4), SMOTEBagging (26) ve UnderBagging (15) yöntemleri kullanılarak sınıf dengesizliği problemini ortadan kaldırmak veya etkisinin azaltılması planlanmıştır.

2.5.1 Temel Kavramlar

Bootstrap

Bootstrap yeniden örnekleme yöntemlerinden biridir. 1979'da Efron tarafından önerilen bu yöntem küçük örneklem genişliğine sahip örneklemler için kullanılmaktadır (34).

n genişlikte bir veride, yerine koyarak n örneklemin rastgele seçilmesi ile eğitim veri seti oluşturulmaktadır. Bir eğitim setinde aynı örnekten birden fazla olabilir ya da hiç olmayabilir. Her örneğin seçilme olasılığı eşittir ve gözlemlerin seçilme olasılığı $\frac{1}{n}$'dir. Örnekleme seçilmemiş olan gözlemlerin seçilmeme olasılıkları ise $\left(1-\frac{1}{n}\right)$ olarak hesaplanır. Bu işlemi N kez tekrarladığımızda, seçilmemiş olanların seçilmeme olasılıkları $\left(1-\frac{1}{n}\right)^N = e^{-1} \approx 0.368$ ve %36.8'i test setini oluşturduğu söylenebilir. Eğitim seti ise %63.2 olarak hesaplanır.

Bagging

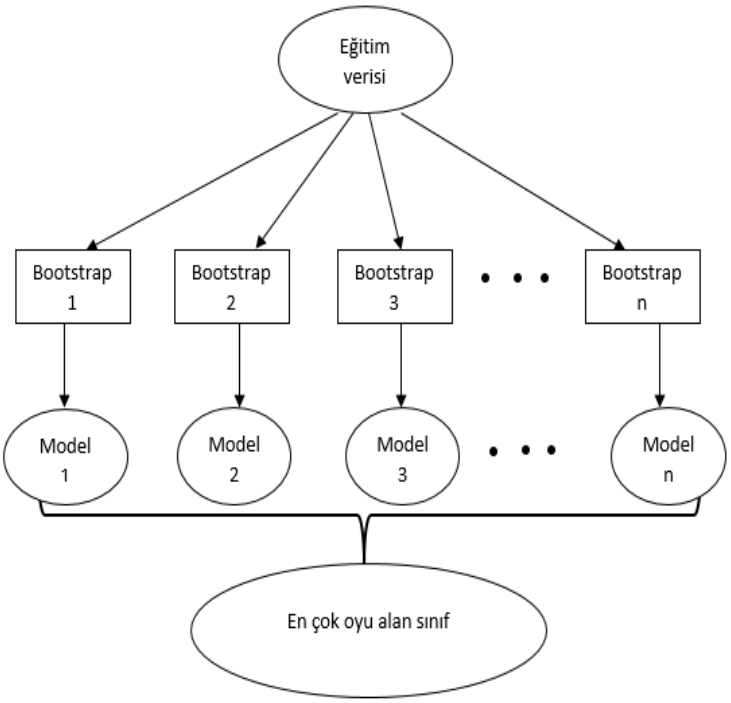
Bagging (Bootstrap Aggregating) bir topluluk (ensemble) öğrenme yöntemidir. Leo Breiman tarafından 1996 yılında önerilen bu yöntem kestirimlerin doğruluğunu arttırmak için kullanılmaktadır (35). Doğru sınıflama oranını arttırmakta ve varyansı düşürmektedir. Bagging yöntemi veri setinden sınıf yapısını bozmayacak şekilde rastgele örnekler seçerek (bootstrap) eğitim setini oluşturur ve sınıf kestirimleri oylanarak en çok oyu alan sınıfı sonuç sınıfı olarak belirleyen öğrenme yöntemidir. Şekil 2.4'de basit bir bagging şeması görülmektedir.

Bagging algoritmasının aşamaları aşağıdaki gibidir:

- Eğitim veri setinden n tane eğitim seti oluşturulur. Eğitim setleri oluşturulurken her bir gözlemin seçilme olasılıkları eşittir ve yerine koyarak seçilmektedir. Bu aşama Bootstrap aşamasıdır.
- Oluşturulan n tane eğitim veri seti ile n tane model kurulmaktadır.
- Kurulan her model sonucunda sınıflar belirlenir.
- Sınıflar arasında en çok oyu alan sınıf kestirim yapılan sınıf olarak adlandırılmaktadır.

Boosting

Boosting, 1990 yılında Freund tarafından bir topluluk öğrenme yöntemi olarak önerilmiştir (14). Başlangıçta eğitim veri setindeki her bir gözleme eşit



Şekil 2.4. Bagging şeması

ağırlık verilmektedir. Önceki eğitim yineleme adımlarında yanlış öğrenilen örneklere daha fazla önem verilerek ağırlıklandırma yapılmaktadır. Yinelemeli olarak birkaç temel sınıflayıcı oluşturulur ve ağırlıklı oy ile gelecek kestirimlerde kullanılmak üzere birleştirilir. Boosting aşamalı bir yöntemdir.

Boosting algoritmasının aşamaları aşağıdaki gibidir:

- Bagging algoritmasında olduğu gibi eğitim veri setinden n tane eğitim seti oluşturulur.
- Eğitim veri setindeki gözlemlere başlangıçta eşit ağırlık verilir.
- Bootstrap ile çekilen n gözlem içeren eğitim veri seti ile model kurulmaktadır.
- Kurulan model sonucunda sınıflar belirlenir.
- Yanlış sınıflandırılan gözlemlerin ağırlıkları arttırılır ve doğru sınıflandırılan gözlemlerin ağırlıkları azaltılır.
- Tekrar model kurulur ve tekrar doğru sınıflanan ve yanlış sınıflanan gözlemler üzerinde ağırlıklar değiştirilir. Bu işlem n kez tekrarlanır.

- Sınıflar arasında en çok oyu alan sınıf kestirilen sınıf olarak adlandırılmaktadır.

Literatürde boosting algoritmasının çeşitleri mevcuttur. AdaBoost (Adaptive Boosting), Gradient Boosting ve XGBoost (Extreme Gradient Boosting) algoritmaları bunlardan bazılarıdır. Bunların içinden en basit hali olan AdaBoost algoritmasının adımları aşağıda sıralanmıştır (25).

AdaBoost Algoritması:

- Adım 1. D sınıf etiketli eğitim seti sınıfları, k döngü sayısı (döngü başına bir sınıflandırıcı oluşturur), M_i hata miktarı

$$error(M_i) = \sum_{j=1}^{d} w_j \cdot err(x_j) \tag{2.14}$$

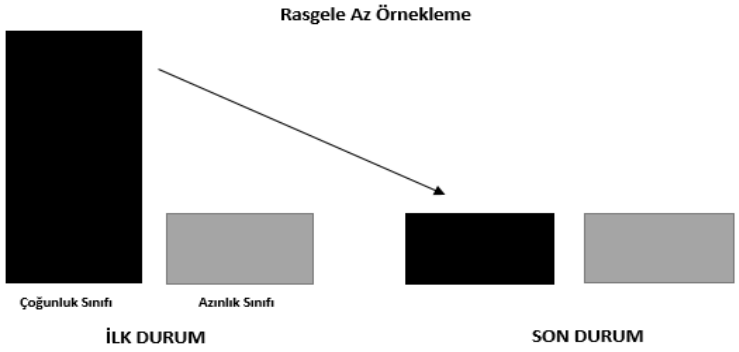
- Adım 2. Her bir gözlemin ağırlığı 1/d olarak belirlenir.
- Adım 3. i=1,2,. . . ,k
  - D'den D_i elde etmek için değişkenlerin ağırlıklarını değiştirme.
  - D_i eğitim setini kullanarak model kurulur.
  - Hata (M_i) ve hata oranı hesaplanır.

  o Eğer hata oranı 0,5'den büyükse, Adım 3'e dönülür.
  - Her bir gözlemin sınıfı bulunur.
  - Kestirilen sınıf yardımı ile ağırlıklar yenilenir.
- Adım 4. En çok oyu alan sınıf kestirilen sınıf olarak adlandırılmaktadır.

2.5.2 Sınıf Dengesizliği Problemi İçin Kullanılan Algoritmalar

RUS (Rastgele az örnekleme)

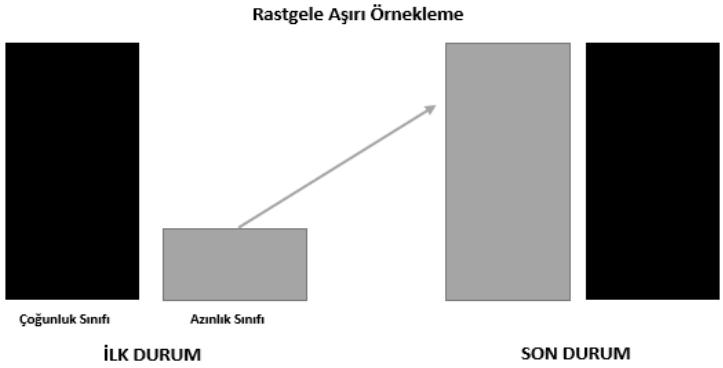
Çoğunluk sınıfı gözlemlerinin azınlık sınıfı gözlem sayısına eşit olacak şekilde çoğunluk sınıfından rastgele gözlemlerin çıkarılmasına rastgele az örnekleme denmektedir (25). Şekil 2.5'de rastgele az örneklemenin basit bir şekli yer almaktadır. İlk durumda çoğunluk sınıfı azınlık sınıfından çok fazla gözleme sahip iken son durumda azınlık ile çoğunluk sınıfı gözlem sayıları birbirine eşittir. Bu işlemin yapılması, sınıflama yapılırken hem süreyi azaltmakta hem de sınıflama performanslarının doğruluğunu arttırmaktadır.



Şekil 2.5. Rastgele az örnekleme

ROS (Rastgele aşırı örnekleme)

Bu yöntem, azınlık sınıfında bulunan gözlemler arasından rastgele gözlemleri tekrar azınlık sınıfına ekleyerek gözlem sayısının artmasını sağlayan örnekleme yöntemidir (36). Şekil 2.6'de rastgele aşırı örneklemenin basit bir şekli yer almaktadır. İlk durumda çoğunluk sınıfı azınlık sınıfından çok fazla gözleme sahip iken son durumda azınlık sınıfı gözlem sayısı artmaktadır. Bu yöntemin dezavantajlarından birisi sınıflama yapılırken süreyi arttırmasıdır.



Şekil 2.6. Rastgele aşırı örnekleme

SMOTE (Sentetik Azınlık Aşırı Örnekleme)

SMOTE algoritması, sentetik azınlık örnekler üretmeye dayanan bir aşırı örnekleme yöntemidir (1). Arka plandaki algoritma kısaca şöyle tanımlanabilir:

- Adım 1: Azınlık sınıfına ait her gözlemin k yakın komşusu aranır,
- Adım 2: Azınlık sınıfına ait gözlem ile k yakın komşusu (kNN) olan gözlem arasındaki fark alınır,

- Adım 3: (0,1) arasında rastgele bir sayı (α) seçilir, Adım 2'de bulunan fark ile bu sayı çarpılır.
- Adım 4: Eşitlik 2.15 kullanılarak yeni sentetik gözlem elde edilir.

$$x_{yeni} = x_i + (x_j - x_i) * \alpha \tag{2.15}$$

- Adım 5: İstenen sayıda sentetik gözlem oluşturmak için Adım 1-4 yinelenir.

SMOTEBoost (SMOTE + Boosting)

SMOTEBoost algoritması SMOTE ve boosting algoritmalarının birleştirilmesi ile oluşturulmuştur (5). Bu yöntemin SMOTE'a göre iki ayrı avantajı vardır.

- Birincisi, standart boosting uygulamasıdır yani yanlış sınıflandırılmış tüm örneklere eşit ağırlıklar verilirken, SMOTEBoost algoritmasında, yeni sentetik gözlemler üretildiği için ağırlıklar dolaylı olarak değişmektedir. Böylelikle dengesiz dağılım dengelenmiş olmaktadır.

- İkincisi, doğru sınıflama oranını artırır, bu nedenle modelin doğruluğu artmış olmaktadır.

Algoritma adımları aşağıdaki gibidir:
P azınlık sınıfı gözlemleri, N çoğunluk sınıfı gözlemleri olsun (|P|<|N|).

- Adım 1. i=1,2,..,T

  - SMOTE algoritması kullanılarak P azınlık sınıfı gözlemlerinden N kadar sentetik gözlemler üretilir.

  - Eğitim seti ile eğitilir.

  - Test seti ile test edilir.

  - Kestirilen sınıf yardımı ile ağırlıklar yenilenir.

- Adım 2. 1. adım T kez tekrarlanır.
- Adım 3. Ağırlıklandırılmış oy ile kestirilen sınıf bulunur.

RUSBoost (Rastgele az örnekleme + Boosting)

RUSBoost, RUS ve standart boosting algoritmasını birleştirir (6). Algoritma aşağıdaki gibidir:
P azınlık sınıfı gözlemleri, N çoğunluk sınıfı gözlemleri olsun (|P|<|N|).
m eğitim setindeki toplam gözlem sayısı
S geçici eğitim seti

- Adım 1. Her gözlemin ağırlığı D_t(i)= 1/m olarak ayarlanır.
- Adım 2. i=1,2,..,T
  - Geçici eğitim seti rasgele az örnekleme ile oluşturulur.
  - S ile ağırlıklar kullanılarak model elde edilir.
  - Test seti ile test edilir.
  - Kestirilen sınıf yardımı ile ağırlıklar yenilenir.
- Adım 3. 2. adım T kez tekrarlanır.
- Adım 4. Ağırlıklandırılmış oy ile kestirilen sınıf bulunur.

EasyEnsemble

EasyEnsemble algoritması bir topluluk öğrenme algoritmasıdır (4). Bu algoritma hem boosting hem de bagging algoritmalarının kombinasyonundan yararlanmaktadır. Algoritma aşağıdaki gibidir:
P azınlık sınıfı gözlemleri, N çoğunluk sınıfı gözlemleri olsun (|P|<|N|).
T çoğunluk sınıfından elde edilen alt örnekleme sayısı

- Adım 1. i=0,1,. . . ,T
- Adım 2. N'den rastgele gözlemler çekilerek $|N_i|$=|P| olacak şekilde alt kümeler oluşturulur.
- Adım3. P ve N_i kullanılarak eğitim seti oluşturulup eğitilir.
- Adım 4. 2. ve 3. adımlar T kez tekrarlanır.
- Adım 5. Test seti ile test edilir.
- Adım 6. Sınıflar arasından en çok oyu alan sınıf, kestirim yapılan sınıf olarak adlandırılır.

MWMOTE

MWMOTE algoritmasında temel amaç; dengesiz öğrenme problemlerini hafifletme ve faydalı sentetik azınlık sınıfı örneklerini oluşturmaktır (21).

Adımları şöyledir:

P azınlık sınıfı gözlemleri, N çoğunluk sınıfı gözlemleri olsun (|P|<|N|).

k_1 gürültülü azınlık sınıfı gözlemlerini kestirmek için kullanılan azınlık sınıfı komşu sayısı

k_2 bilgilendirici azınlık sınıfı gözlemlerini kestirmek için kullanılan çoğunluk sınıfı komşu sayısı

k_3 bilgilendirici azınlık sınıfı gözlemlerini kestirmek için kullanılan azınlık sınıfı komşu sayısı

- Adım 1. Her bir azınlık sınıfı gözlemlerinin k_1 komşularını bulmak için Öklid uzaklıkları hesaplanır.
- Adım 2. Azınlık sınıfı gözlemlerinden k_1 olanlar çıkarılır.
- Adım 3. k_2 belirlenerek çoğunluk sınıfına ait sınır oluşturulur.
- Adım 4. k_3 belirlenir.
- Adım 5. Her bir çoğunluk ve azınlık sınıfına ait gözlemlerin bilgi ağırlıkları hesaplanır.
- Adım 6. Her bir gözlemin kümesi (sınıfı) bulunur.
- Adım 7. SMOTE algoritması yardımı ile sentetik gözlemler üretilir.

SMOTEBagging (SMOTE + Bagging)

SMOTEBagging algoritması, SMOTE ve Bagging algortimalarının birleştirilmesinden oluşmaktadır (19) . Bu yöntem rastgele aşırı örnekleme (oversampling) yönteminden farklılık göstermektedir. Baggingleri yaratma şekli farklıdır. Rastgele aşırı örnekleme yerine SMOTE kullanılmaktadır. Eğitimde azınlık-çoğunluk sınıfı aynı sayıda olmaktadır. Yeniden örnekleme (resampling) oranı (a) yardımıyla, SMOTE ile oluşturulacak azınlık sınıfı oranı belirlenir.

Algoritmanın adımları aşağıdaki gibidir:

- Adım 1. D eğitim seti olsun.

- Adım 2. Tüm sınıflardan aynı sayıya sahip gözlemleri içeren D_k kümesi oluşturulur.

  - %100 yerine koyarak yeniden örnekleme yapılır.

  - Her bir sınıf için %b oranında yerine koyarak alt kümeler oluşturulur.

  - SMOTE algoritması kullanılarak yeni sentetik gözlemler üretilir.

- Adım 3. Her bir alt küme sınıflama yöntemleri ile modellenir.

- Adım 4. b oranı değiştirilir.

- Adım 5. 2. , 3. ve 4. adımlar T kez tekrarlanır.

- Adım 6. Test seti ile test edilir.

- Adım 7. Sınıflar arasından en çok oyu alan sınıf, kestirim yapılan sınıf olarak adlandırılır.

UnderBagging (Az örnekleme + Bagging)

Bu yöntem, Bagging ile rastgele az örnekleme (undersampling) yöntemini kullanır (15). Çoğunluk sınıfındaki gözlemleri seçerken rastgele az örnekleme yapmaktadır.
Algoritmanın adımları şöyledir:

- Adım 1. D eğitim seti olsun.

- Adım 2. Tüm sınıflardan aynı sayıya sahip gözlemleri içeren D_k kümesi oluşturulur.

  - Yeniden örnekleme oranı %a olarak ayarlanır.

  - Her bir sınıf için %a oranında yerine koyarak alt kümeler oluşturulur.

- Adım 3. Her bir alt küme sınıflama yöntemleri ile modellenir.

- Adım 4. 2. ve 3. adımlar T kez tekrarlanır.

- Adım 5. Test seti ile test edilir.

- Adım 6. Sınıflar arasından en çok oyu alan sınıf, kestirim yapılan sınıf olarak adlandırılır.

2.6 Sınıflama Yöntemleri İçin Performans Ölçüleri

Makine öğrenmesinde, sınıflandırma yöntemlerinde kurulan modelin performansını ölçen çeşitli ölçüler vardır. Bunlardan bazıları aşağıdaki gibidir:

- Genel doğruluk oranı (overall accuracy)
- Duyarlılık (Sensitivity)
- Seçicilik (Specificity)
- Pozitif kestirim değeri (Positive Predictive Value)
- Negatif kestirim değeri (Negative Predictive Value)
- Eğri altında kalan alan (Area Under Curve)
- Düzeltilmiş doğruluk oranı (Balanced Accuracy)
- F-Ölçüsü (F-Measure)

Yukarıda yer alan model performans ölçüleri, ikili sınıflama durumu için aşağıda verilen 2x2'lik sınıflama tablosundan yararlanılarak hesaplanmaktadır.

Tablo 2.2. İki sınıflı veri için 2x2 çapraz tablo

	GERÇEK DURUM		**TOPLAM**
	Pozitif	**Negatif**	
KESTİRİM **Pozitif**	Doğru Pozitif (DP)	Yanlış Pozitif (YP)	DP+YP
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)	YN+DN
TOPLAM	DP+YN	DN+YP	N

2.6.1 Genel Doğruluk Oranı (GDO)

Doğru pozitif ve doğru negatif sonuçların toplamının çalışmada bulunan toplam gözlem sayısına oranıdır. Literatürde sıklıkla kullanılan genel doğruluk oranının sınıf dengesizliğinde kullanılmaması gerektiği (çeşitli makale ve kitaplarda) vurgulanmıştır (25, 37–39). GDO değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\text{GDO} = \frac{DP + DN}{N} \tag{2.16}$$

2.6.2 Duyarlılık (DUY)

Gerçek durumda pozitif olduğu bilinen bir gözlemin kestirim sonucunun pozitif çıkma olasılığıdır (25, 39, 40). Veri madenciliğinde sınıflama performanslarını gösteren sonuçlarda "geri çağırma (recall)" diye adlandırılmaktadır. DUY değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\text{Duyarlılık} = \frac{DP}{DP + YN} \tag{2.17}$$

2.6.3 Seçicilik (SEÇ)

Gerçek durumda negatif olduğu bilinen bir gözlemin kestirim sonucunda da negatif çıkması olasılığıdır(25, 39, 40). SEÇ değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\text{Seçicilik} = \frac{DN}{YP + DN} \tag{2.18}$$

2.6.4 Pozitif Kestirim Değeri (PKD)

Kestirim değeri pozitif olan bir gözlemin gerçek durumda da pozitif olma olasılığıdır (39). Veri madenciliğinde sınıflama performanslarını gösteren sonuçlarda "kesinlik (precision)" olarak da adlandırılmaktadır. PKD değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\text{PKD} = \frac{DP}{DP + YP} \tag{2.19}$$

Prevelansın etkisinin dahil edilmesi ile Eşitlik 2.20 elde edilmektedir.

$$\text{PKD} = \frac{DUY.Prevelans}{DUY.Prevelans + (1 - \text{SEÇ}).(1 - Prevelans)} \tag{2.20}$$

2.6.5 Negatif Kestirim Değeri (NKD)

Kestirim değeri negatif olan bir gözlemin gerçek durumda da negatif olma olasılığıdır (39). NKD değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yak-

laşması performansın yüksek olduğunu göstermektedir.

$$\text{NKD} = \frac{DN}{DN + YN} \tag{2.21}$$

Prevelansın etkisinin dahil edilmesi ile Eşitlik 2.22 elde edilmektedir.

$$\text{NKD} = \frac{\text{SEÇ.}(1 - Prevelans)}{\text{SEÇ.}(1 - Prevelans) + (1 - DUY).Prevelans} \tag{2.22}$$

2.6.6 Eğri Altında Kalan Alan (EAA)

Farklı kesim noktalarında farklı duyarlılık ve seçicilik değerleri ile elde edilen ROC eğrisi altında kalan alandır (25, 39). EAA değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\widehat{EAA} = \int_0^1 \widehat{ROC}(t)dt \tag{2.23}$$

2.6.7 Düzeltilmiş Doğruluk Oranı (DDO)

Duyarlılık ve seçicilik değerlerinin ortalamasıdır (41). DDO değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\text{DDO} = \frac{1}{2} * (\text{Duyarlılık} + \text{Seçicilik}) \tag{2.24}$$

2.6.8 F-ölçüsü

Duyarlılık ve seçicilik ölçülerinin tek başına yeterli olmadıkları düşünülerek elde edilmiştir. Pozitif kestirim değeri ve duyarlılık değerlerinin harmonik ortalamasıdır (42). F-ölçüsü değeri, 0 ile 1 arasında değişmektedir. Bu değerin 1'e yaklaşması performansın yüksek olduğunu göstermektedir.

$$\text{F-ölçüsü} = 2 * \frac{\text{Pozitif Kestirim Değeri} * \text{Duyarlılık}}{\text{Pozitif Kestirim Değeri} + \text{Duyarlılık}} \tag{2.25}$$

3 GEREÇ ve YÖNTEM

Tez çalışmamızda, Bölüm 3.1'de benzetim çalışması ve Bölüm 3.2'de gerçek veri setlerine ait detaylı bilgiler yer almaktadır.

3.1 Benzetim Çalışması

Alanyazında, sınıf dengesizliği problemi ile ilgili birçok çalışma bulunmaktadır. Bu çalışmaların çoğunluğu gerçek veri setleri kullanılarak değerlendirilmiştir. Benzetim çalışması yapılmış çalışma çok azdır ve bu çalışmalarda farklı korelasyon yapılarında inceleme yapılmamıştır. Bu nedenle, bu tez kapsamında farklı korelasyon yapılarını da dikkate alarak algoritmaların geçerliği gösterilmeye çalışılmıştır. Sınıflandırma yöntemlerinin performansları, örneklem büyüklüklerinden, korelasyon yapılarından ve sınıf dengesizlik oranlarından etkilenebilir. Bu nedenle, benzetim senaryoları bu etkiler dikkate alınarak hazırlanmıştır. Benzetim çalışması aşağıdaki maddelerin tüm olası kombinasyonları içeren senaryolardan oluşmaktadır:

1) Düşük, orta, yüksek ve gerçek (yani gerçek bir veri kümesinden elde edilen korelasyon) korelasyon yapıları

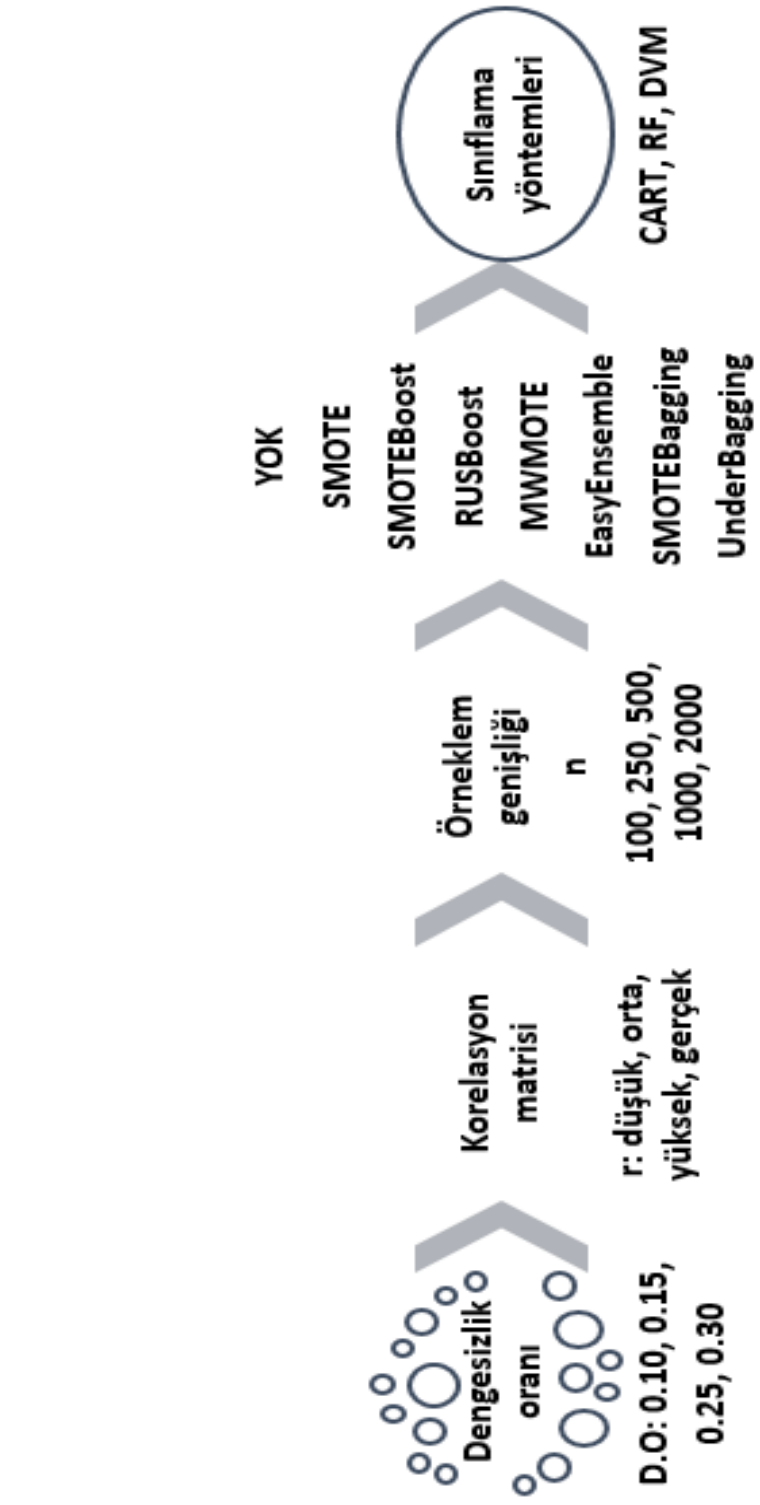
2) Örneklem genişlikleri 100, 250, 500, 1000 ve 2000

3) Dengesizlik oranları %10, %15, %25 ve %30

Bağımlı ve bağımsız değişkenler arasındaki korelasyon düşük, orta ve yüksek korelasyonlu yapılar için sırasıyla [0,00-0,29], [0,30-0,60] ve [0,61-0,90] aralıklarında tanımlanmıştır. Bağımsız değişkenler arasında zayıf veya hiç ilişki olmadığı varsayılmıştır. Sınıf değişkeniyle birlikte on beş değişken (5 ikili, 10 sayısal) farklı korelasyon yapıları dikkate alınarak R programında (3.2.3 versiyonu) "BinNor" (43) paketi kullanılarak türetildi. Ayrıca, nitel değişkenler iki kategorili şekilde üretilmiştir. İki kategorili değişkenler arasındaki ilişki Phi korelasyon katsayısı, iki kategorili değişkenler ile nicel değişkenler arası ilişki Nokta çift serili korelasyon katsayısı ve nicel değişkenler arası ilişki Pearson korelasyon katsayısı dikkate alınarak oluşturulmuştur. "BinNor" paketi çok değişkenli normal dağılım varsayımı altında veri türetmekte olup her bir korelasyon yapısı elle oluşturulup veri türetme aşamasında tanımlanmıştır. 15 değişkenden sadece 5 bağımsız değişken bağımlı değişken ile ilişki tanımlanmış ve diğer değişkenler ilişkisiz olarak tanımlanmıştır. Tüm senaryolar (yani 80 farklı kombinasyon) 1000 kez tekrarlandı ve optimum model parametrelerini bulmak için 5 katlı çapraz geçerlik uygulandı. SMOTE, SMOTEBoost, RUSBoost, EasyEnsemble, MWMOTE, SMOTEBagging ve UnderBagging algoritmaları sınıflama yapılamadan önce verilerin denge-

sizliğini azaltmak ya da tamamen ortadan kaldırmak için kullanıldı. Benzetim ile elde edilen veriler ve gerçek veri setleri CART, DVM ve RF yöntemleri kullanılarak sınıflama yapıldı. Sınıflarda dengesizlik varken sadece sınıflama yapılarak elde edilen sonuçlar diğer algoritmaların sonuçları ile karşılaştırıldı. Herhangi bir algoritma kullanılmadan sınıflama yöntemleri uygulanan durum, tablolarda ve grafiklerde "YOK" olarak ifade edildi. Tablo ve grafiklerde 1000 tekrarın ortalaması olarak sunulmuştur.

R programından (versiyon 3.2.3) "caret" (44), "ROSE" (36), "caretEnsemble" (45), "imbalance" (46), "ebmc"(47) ve "ggplot"(48) paketleri kullanıldı. Benzetim çalışmasının iş akışı Şekil 3.1'dedir.



Şekil 3.1. Benzetim çalışması özeti

Korelasyon yapılarına ait matrisler Tablo 3.1, 3.2, 3.3, 3.4'de sunulmuştur. Tablolarda örnek olarak sadece beş değişkene ait yapı gösterilmiştir. Tablo 3.1'de hem bağımsız değişkenler arasında ilişki hem de bağımlı değişken ile bağımsız değişkenler arasındaki ilişki çok zayıf olarak oluşturuldu ve düşük korelasyonlu yapı olarak kabul edildi.

Tablo 3.1. Düşük düzey korelasyon yapısı

$$\begin{bmatrix} & y & x_1 & x_2 & x_3 & x_4 & \dots & x_{15} \\ y & 1,000 & 0,100 & 0,100 & 0,100 & 0,250 & \dots & 0,188 \\ x_1 & 0,100 & 1,000 & 0,060 & 0,100 & 0,178 & \dots & 0,029 \\ x_2 & 0,100 & 0,060 & 1,000 & 0,100 & 0,100 & \dots & 0,100 \\ x_3 & 0,100 & 0,100 & 0,100 & 1,000 & 0,100 & \dots & 0,100 \\ x_4 & 0,250 & 0,178 & 0,100 & 0,100 & 1,000 & \dots & 0,100 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{15} & 0,188 & 0,029 & 0,100 & 0,100 & 0,100 & \dots & 1,000 \end{bmatrix}$$

Tablo 3.2 'de bağımsız değişkenler arasındaki ilişkinin çok düşük, fakat bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin orta düzey olduğu durum orta düzey korelasyonlu yapı olarak dikkate alındı.

Tablo 3.2. Orta düzey korelasyon yapısı

$$\begin{bmatrix} & y & x_1 & x_2 & x_3 & x_4 & \dots & x_{15} \\ y & 1,000 & 0,500 & -0,480 & 0,460 & -0,450 & \dots & 0,150 \\ x_1 & 0,500 & 1,000 & -0,157 & -0,071 & 0,158 & \dots & 0,150 \\ x_2 & -0,480 & -0,157 & 1,000 & -0,166 & -0,040 & \dots & 0,150 \\ x_3 & 0,460 & -0,071 & -0,166 & 1,000 & 0,147 & \dots & 0,150 \\ x_4 & -0,450 & 0,158 & -0,040 & 0,147 & 1,000 & \dots & 0,150 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{15} & 0,150 & 0,150 & 0,150 & 0,150 & 0,150 & \dots & 1,000 \end{bmatrix}$$

Tablo 3.3 'de bağımsız değişkenler arasındaki ilişkinin çok düşük ya da hiç ilişki yok, fakat bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin yüksek düzey olduğu durum yüksek korelasyonlu yapı olarak dikkate alındı.

Tablo 3.3. Yüksek düzey korelasyon yapısı

$$\begin{bmatrix} & y & x_1 & x_2 & x_3 & x_4 & \dots & x_{15} \\ y & 1,000 & 0,692 & -0,790 & 0,000 & 0,158 & \dots & -0,780 \\ x_1 & 0,692 & 1,000 & -0,103 & 0,000 & 0,000 & \dots & 0,000 \\ x_2 & -0,790 & -0,103 & 1,000 & 0,000 & 0,000 & \dots & 0,000 \\ x_3 & 0,000 & 0,000 & 0,000 & 1,000 & 0,000 & \dots & 0,000 \\ x_4 & 0,000 & 0,158 & 0,000 & 0,000 & 1,000 & \dots & 0,000 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{15} & -0,780 & 0,000 & 0,000 & 0,000 & 0,000 & \dots & 1,000 \end{bmatrix}$$

Bu tez kapsamında gerçek veriden yararlanılarak elde edilen korelasyon yapısı, gerçek korelasyon yapısı olarak kullanılmış ve korelasyon matrisi Tablo 3.4'de verilmiştir.

Tablo 3.4. Gerçek korelasyon yapısı

$$\begin{bmatrix} & y & x_1 & x_2 & x_3 & x_4 & \dots & x_{15} \\ y & 1,000 & 0,378 & -0,383 & 0,322 & -0,283 & \dots & 0,115 \\ x_1 & 0,378 & 1,000 & 0,281 & 0,330 & 0,265 & \dots & 0,109 \\ x_2 & -0,383 & 0,281 & 1,000 & -0,335 & 0,348 & \dots & 0,064 \\ x_3 & 0,322 & 0,330 & -0,335 & 1,000 & -0,369 & \dots & 0,172 \\ x_4 & -0,283 & 0,265 & 0,348 & -0,369 & 1,000 & \dots & 0,075 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{15} & 0,115 & 0,109 & 0,064 & 0,172 & 0,075 & \dots & 1,000 \end{bmatrix}$$

3.2 Gerçek Veri Setleri

Tez çalışmamızda, yedi adet gerçek veri seti ile çalışıldı. Bu verilere ilişkin örneklem genişlikleri, azınlık-çoğunluk sınıfı yüzdeleri ve bağımsız değişken sayıları Tablo 3.5'te özetlenmiştir. Altı veri seti açık erişim olan UC Irvine Machine Learning Repository web sayfasından alınmıştır. Bir tanesi ise tablonun altında yer verilen makaleden alınmıştır.

Tablo 3.5. Gerçek veri setlerine ait özet tablo

Veri setleri	Örneklem genişliği	Azınlık sınıfı-Çoğunluk sınıfı	Bağımsız Değişken sayısı
Deniz Kabukları	731	%5,74 - %94,26	8
Doğurganlık	100	%12 - %88	9
Göğüs Cerrahisi	470	%15 - %85	16
Hepatit	155	%20 - %80	19
Kan Nakli	748	%23 - %77	4
Alzheimer*	70	%30 - %70	25
Diyabet	768	%34,9 - %66,1	8

* A blood based 12-miRNA signature of Alzheimer disease patients, (2013), Leidinger P et al.

Tablo 3.6'da veri setlerine ait bağımlı değişken ve bağımsız değişkenlerin bilgileri yer almaktadır.

Tablo 3.6. Gerçek veri setlerinin değişkenlerine ait özet tablo

Veri Setleri	Bağımsız Değişken Sayısı	Bağımsız Değişkenler	Sınıf Değişkeni
Deniz Kabukları	1 nitel, 7 nicel	cinsiyet, uzunluk, çap, yükseklik, tüm ağırlık, soyulmuş ağırlık, iç organ ağırlığı, kabuk ağırlığı	pozitif, negatif
Doğurganlık	6 nitel, 3 nicel	yaş, mevsim, travma, çocukluk hastalığı, cerrahi müdahale, ateş, alkol tüketim sıklığı, sigara alışkanlığı, günlük oturarak geçirilen saat	normal, diğer
Göğüs Cerrahisi	13 nitel, 3 nicel	tanı, zorlu viral kapasite (FVC), Zorlu ekspirasyonun birinci saniyesinde atılan volüm (FEV1), Zubrod ölçeği, ameliyattan önce ağrı, ameliyattan önce hemoptizi, ameliyattan önce solunum güçlüğü, ameliyattan önce öksürük, ameliyattan önce halsizlik, tümör boyutu, tip2 diyabet, 6 ay içinde miyokard infarktüsü, periferik arter hastalıkları, sigara alışkanlığı, astım, yaş	yaşıyor, ölü
Hepatit	13 nitel, 3 nicel	yaş, cinsiyet, steroid, antiviral, yorgunluk, halsizlik, iştahsızlık, karaciğer büyüklüğü, karaciğer sağlamlığı, dalak belirginliği, karında su toplanması, örümcek ısırığı, varis, bilirubin, alkil fosfat, sgot, albümin, prothrombin time, histoloji	yaşıyor, ölü
Kan Nakli	0 nitel, 3 nicel	son bağıştan bu zamana kadar geçen aylar, toplam bağış sayısı, toplam bağışlanan kan ml, ilk bağıştan bu zamana kadar geçen aylar	kan bağışı evet,kan bağışı hayır
Alzheimer	0 nitel, 25 nicel	miR.168, miR.2175, miR.1311, miR.1988, miR.2205, miR.1425, miR.639, miR.1695, miR.2551, miR.2750, miR.1119, miR.1472, miR.2279, miR.481, miR.2668, miR.2696, miR.1268, miR.657, miR.2344, miR.372, miR.2576, miR.2729, miR.2538, miR.1818, miR.1679	hasta, sağlam
Diyabet	0 nitel, 8 nicel	hamile kalma sayısı, glukoz, diyastolik kan basıncı, triseps cilt kat kalınlığı, insülin, beden kütle indeksi, diyabet pedigree fonksiyonu, yaş	hasta, sağlam

4 BULGULAR

Gereç ve Yöntem bölümünde, ayrıntılı olarak verilmiş olan benzetim çalışmasına ve gerçek veri setlerine ait sonuçlar, Bölüm 4.1 ve Bölüm 4.2'de yer almaktadır. Bölüm 4.1'de dört farklı korelasyon yapısına ait sonuçlar ayrı ayrı incelenmiş ve sonuçların değerlendirilmesi F-ölçüsü dikkate alınarak yapılmıştır.

Bölüm 4.2'de veri setlerine ait sonuçlar tek bir tabloda sunulmuştur.

4.1 Benzetim Çalışması Sonuçları

Benzetim çalışmasına ait sonuçlar alt başlıklar halinde sunulmuştur. Yöntemlerin karşılaştırmalarının daha net görülebilmesi için grafikler verilmiştir. Karşılaştırma yapılan algoritmalar ile ilgili yorumlar her bir korelasyon yapısına ait sonuçlar ile birlikte verilmiştir.

3 farklı sınıflama yöntemi ve 8 farklı algoritma sonuçları 24 grafiği bir arada bulunduran grafikler halinde sunulmuştur. Grafiklerin dikey ekseninde F-Ölçüsü değerleri yatay eksende ise örneklem genişlikleri bulunmaktadır. Grafik içerisinde dengesizlik oranları farklı renklerle ifade edilmiştir. Algoritmaların adları ve dengesizlik oranları grafiklerin içerisinde belirtilmiştir. F-ölçüsü değerleri 1000 tekrarın ortalama değerleridir.

Grafikler R programında "ggplot" paketi yardımı ile çizilmiştir.

Tablolar ise her bir korelasyon düzeyinde ayrı ayrı verilmiş olup her bir dengesizlik oranına göre de ayrı ayrı verilmiştir. Örneğin düşük düzey korelasyon ve 0,10 dengesizlik durumu veya yüksek düzey korelasyon ve 0,30 dengesizlik durumu olacak şekilde sunulmuştur. Grafiklerde olduğu gibi tablolarda da F-ölçüsü dikkate alınarak sonuçlar yorumlanmıştır.

4.1.1 Düşük Düzey Korelasyona Ait Sonuçlar

Grafik 4.1'de Düşük düzey korelasyon tanımlanarak üretilmiş verilerden elde edilen sınıflama sonuçları gösterilmektedir.

Örneğin, Grafikte (1,1) hücresi herhangi bir algoritma olmadan ve sınıflama yöntemi DVM olduğu durumdaki farklı örneklem genişliklerinde dengesizlik oranlarını göstermektedir.

Grafiğe bakıldığında, net bir şekilde görülen algoritma kullanılmadan uygulanan sınıflama yöntemlerinin performansıdır. Dengesizlik azalsa da algoritma kullanılmadan elde edilen performans sonuçları örneklem genişliği arttıkça azalmıştır. Her bir algoritma tek tek değerlendirilmiştir. SMOTE algoritması sonuçları (1,2), (2,2) ve (3,2) hücrelerinde yer almaktadır. Dengesizliğin 0,10 ve 0,15

olduğunda her bir sınıflama yönteminde performans sonuçları benzer ve örneklem genişliği arttıkça performanslar düşmektedir. Dengesizlik azaldıkça yani 0,25 ve 0,30 olduğu durumda SMOTE algoritmasının sınıflama performansları üzerindeki etkisi net bir şekilde görülmektedir. SMOTEBoost algoritmasının etkisi SMOTE algoritmasının etkisi ile benzer bulunmuştur.

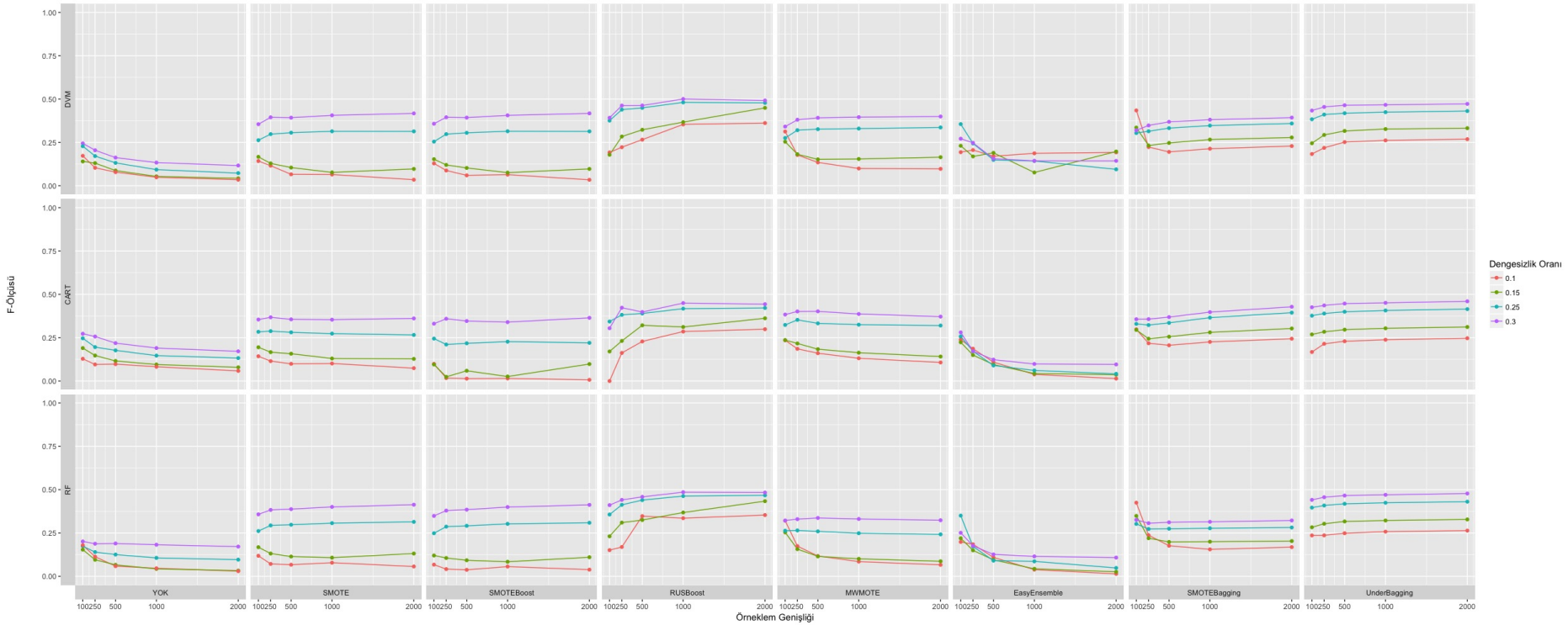
RUSBoost algoritmasının etkileri grafikte (1,4), (2,4) ve (3,4) hücrelerinde yer almaktadır. Örnekleme genişliği arttıkça her bir dengesizlik oranında sınıflama performansları da artmaktadır. UnderBagging algoritmasının da sınıflama yöntemleri üzerindeki etki RUSBoost algoritmasının etkisine benzerken EasyEnsemble algoritmasının etkisi tam tersi yöndedir. Yani örneklem genişliği arttıkça sınıflama yöntemlerindeki performanslar düşmektedir.

MWMOTE algoritmasının performansı dengesizlik 0,10 ve 0,15 olduğunda örneklem genişliği arttıkça düşmektedir. Dengesizliğin 0,25 ve 0,30 olduğu durumda sınıflama yöntemleri üzerindeki etki artmaktadır.

0,10 ve 0,15 dengesizlik durumlarında RUSBoost, SMOTEBagging ve UnderBagging algoritmalarının sınıflama üzerindeki etkisi algoritma olmadan uygulanan sınıflama üzerindekine göre yüksek bulunmuştur. Sınıflama yöntemlerindeki performansı arttırdığı görülmüştür.

Düşük korelasyon yapısının etkisi performans sonuçlarına bakıldığında net bir şekilde görülmektedir.

Dört farklı dengesizlik durumuna ait sonuçlar Tablo 4.1, Tablo 4.2, Tablo 4.3 ve Tablo 4.4'de verilmiştir.



Şekil 4.1. Düşük düzey korelasyon sonuçları

Tablo 4.1. Düşük düzey korelasyon ve 0,10 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0.907	0,995	0,015	0,505	0,173	0,903	0,997	0,008	0,502	0,104	0,902	0,999	0,004	0,501	0,079	0,901	0,999	0,003	0,501	0,049	0,900	0,999	0,003	0,099	0,035
	SMOTE	0,904	0,993	0,018	0,506	0,143	0,904	0,994	0,018	0,506	0,116	0,901	0,998	0,008	0,503	0,066	0,899	0,996	0,020	0,508	0,065	0,899	0,998	0,007	0,503	0,035
	SMOTEBoost	0,903	0,992	0,019	0,506	0,129	0,904	0,995	0,015	0,505	0,088	0,901	0,998	0,008	0,503	0,060	0,899	0,996	0,020	0,508	0,064	0,899	0,998	0,008	0,503	0,034
	RUSBoost	0,567	0,566	0,577	0,572	0,192	0,505	0,485	0,704	0,595	0,222	0,589	0,579	0,681	0,630	0,266	0,613	0,606	0,681	0,643	0,354	0,614	0,603	0,712	0,657	0,362
	MWMOTE	0,883	0,968	0,066	0,517	0,313	0,889	0,980	0,048	0,514	0,178	0,887	0,978	0,059	0,518	0,135	0,891	0,985	0,044	0,514	0,100	0,890	0,983	0,054	0,519	0,097
	EasyEnsemble	0,726	0,778	0,241	0,509	0,194	0,899	0,995	0,009	0,502	0,206	0,900	0,998	0,004	0,501	0,170	0,897	0,995	0,013	0,504	0,187	0,894	0,990	0,021	0,506	0,193
	SMOTEBagging	0,861	0,938	0,099	0,518	0,435	0,845	0,917	0,156	0,536	0,223	0,816	0,880	0,221	0,551	0,195	0,787	0,841	0,295	0,568	0,214	0,760	0,804	0,360	0,582	0,229
	UnderBagging	0,463	0,458	0,516	0,487	0,183	0,550	0,541	0,641	0,591	0,219	0,613	0,607	0,663	0,635	0,252	0,629	0,624	0,668	0,646	0,262	0,637	0,634	0,671	0,652	0,269
CART	YOK	0,909	0,998	0,003	0,500	0,128	0,900	0,993	0,011	0,502	0,096	0,899	0,994	0,011	0,503	0,098	0,899	0,996	0,010	0,503	0,082	0,900	0,999	0,003	0,099	0,058
	SMOTE	0,877	0,958	0,063	0,510	0,143	0,892	0,980	0,031	0,506	0,115	0,897	0,992	0,015	0,503	0,099	0,895	0,991	0,021	0,506	0,101	0,897	0,996	0,011	0,503	0,074
	SMOTEBoost	0,881	0,964	0,052	0,508	0,099	0,904	0,995	0,015	0,505	0,017	0,901	0,997	0,005	0,501	0,014	0,899	0,997	0,006	0,502	0,015	0,899	0,999	0,003	0,501	0,007
	RUSBoost	0,909	1,000	0,000	0,500	0,000	0,471	0,452	0,658	0,555	0,162	0,491	0,476	0,637	0,557	0,228	0,525	0,515	0,618	0,566	0,286	0,539	0,529	0,626	0,577	0,299
	MWMOTE	0,824	0,894	0,157	0,525	0,235	0,861	0,943	0,099	0,521	0,186	0,872	0,958	0,079	0,519	0,160	0,888	0,981	0,038	0,510	0,131	0,896	0,994	0,016	0,505	0,107
	EasyEnsemble	0,833	0,910	0,128	0,519	0,239	0,902	1,000	0,001	0,500	0,186	0,901	1,000	0,000	0,500	0,109	0,901	1,000	0,000	0,500	0,038	0,900	1,000	0,000	0,500	0,014
	SMOTEBagging	0,754	0,801	0,284	0,543	0,298	0,785	0,838	0,274	0,556	0,218	0,781	0,834	0,292	0,563	0,206	0,754	0,797	0,365	0,581	0,226	0,703	0,727	0,483	0,605	0,244
	UnderBagging	0,492	1,000	0,100	0,500	0,167	0,569	0,564	0,619	0,591	0,215	0,589	0,585	0,625	0,605	0,229	0,599	0,594	0,637	0,616	0,238	0,607	0,603	0,649	0,626	0,247
RF	YOK	0,908	0,996	0,014	0,505	0,179	0,903	0,996	0,015	0,506	0,113	0,901	0,997	0,009	0,503	0,059	0,901	0,998	0,009	0,504	0,046	0,900	0,998	0,008	0,099	0,030
	SMOTE	0,897	0,983	0,033	0,508	0,118	0,898	0,988	0,021	0,504	0,071	0,897	0,991	0,024	0,508	0,067	0,894	0,988	0,043	0,515	0,078	0,895	0,991	0,031	0,511	0,057
	SMOTEBoost	0,901	0,988	0,025	0,506	0,067	0,902	0,992	0,018	0,505	0,042	0,900	0,995	0,016	0,506	0,037	0,897	0,992	0,032	0,512	0,056	0,898	0,995	0,022	0,508	0,038
	RUSBoost	0,570	0,572	0,552	0,562	0,151	0,555	0,547	0,634	0,590	0,169	0,595	0,591	0,635	0,613	0,347	0,612	0,608	0,646	0,627	0,335	0,617	0,610	0,674	0,642	0,353
	MWMOTE	0,892	0,980	0,047	0,513	0,321	0,893	0,985	0,044	0,514	0,174	0,892	0,984	0,046	0,515	0,117	0,893	0,987	0,039	0,513	0,085	0,894	0,990	0,035	0,512	0,066
	EasyEnsemble	0,767	0,835	0,147	0,491	0,198	0,895	0,991	0,013	0,502	0,185	0,900	0,998	0,002	0,500	0,109	0,900	0,999	0,000	0,500	0,038	0,896	0,995	0,000	0,498	0,014
	SMOTEBagging	0,858	0,932	0,127	0,529	0,425	0,867	0,945	0,116	0,531	0,237	0,863	0,942	0,128	0,535	0,176	0,861	0,940	0,129	0,535	0,156	0,858	0,936	0,147	0,541	0,168
	UnderBagging	0,559	0,548	0,654	0,601	0,236	0,594	0,587	0,661	0,624	0,236	0,610	0,605	0,657	0,631	0,249	0,617	0,611	0,674	0,643	0,258	0,624	0,618	0,678	0,648	0,264

Tablo 4.1 düşük düzey korelasyon ve 0,10 dengesizlik durumunu göstermektedir.

Tablo 4.1'e göre GDO değerlerine bakıldığında, UnderBagging algoritması hariç YOK ve diğer algoritmalarda yüksek bulunmuştur. SEÇ değerleri çok yüksek çıkmış iken DUY değerlerine bakıldığında çok düşük çıkmasından dolayı GDO değerine bakmak yanıltıcı olabilmektedir. Birçok çalışmada performans değerlendirme ölçüsü olarak DDO değerleri kullanılmaktadır. DDO değerleri SEÇ ve DUY değerlerinden hesaplanmakta ve DDO değerleri etkilenmektedir. YOK ve tüm algoritmalarda DDO değerleri yaklaşık %50 ve üzeri civarındadır. Bunun aksine F-ölçüsü DUY değerlerinden etkilendiği için 0,10 dengesizlik durumunda çok düşük değerlere sahiptir.

Örneklem genişliği arttıkça YOK, SMOTE, SMOTEBoost ve MWMOTE algoritmalarında F-ölçüsü değerleri düşmektedir. EasyEnsemble algoritması DVM yönteminin sınıflama performansı üzerinde etkili olmuş iken diğer sınıflama yöntemlerinde örneklem genişliği arttıkça etkisi olmamıştır.

Örneklem genişliği arttıkça RUSBoost algoritması sınıflama yöntemlerinin performanslarını arttırmaktadır. 2000 örneklem genişliğinde %30'un üzerine çıkarmıştır. Bu rakam çok düşük olmasına rağmen YOK ve diğer algoritmaların etkisine göre daha iyi sonuç verdiği söylenebilir.

Tablo 4.2. Düşük düzey korelasyon ve 0,15 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,856	0,990	0,016	0,503	0,140	0,852	0,995	0,012	0,504	0,131	0,852	0,998	0,007	0,502	0,088	0,851	0,999	0,005	0,502	0,054	0,850	0,999	0,005	0,502	0,043
	SMOTE	0,839	0,960	0,079	0,520	0,167	0,845	0,979	0,054	0,517	0,129	0,840	0,975	0,057	0,516	0,105	0,844	0,984	0,040	0,512	0,077	0,843	0,982	0,058	0,520	0,097
	SMOTEBoost	0,837	0,958	0,079	0,519	0,153	0,845	0,979	0,054	0,516	0,120	0,839	0,975	0,057	0,516	0,103	0,844	0,984	0,040	0,512	0,076	0,843	0,982	0,058	0,520	0,097
	RUSBoost	0,702	0,756	0,363	0,559	0,179	0,649	0,667	0,535	0,601	0,284	0,564	0,549	0,655	0,602	0,323	0,619	0,617	0,626	0,622	0,367	0,584	0,564	0,698	0,631	0,449
	MWMOTE	0,822	0,933	0,099	0,516	0,254	0,810	0,929	0,127	0,528	0,182	0,824	0,942	0,108	0,525	0,152	0,824	0,946	0,113	0,529	0,155	0,821	0,945	0,120	0,533	0,165
	EasyEnsemble	0,819	0,951	0,057	0,504	0,231	0,850	0,997	0,006	0,502	0,169	0,846	0,992	0,014	0,503	0,190	0,850	0,999	0,002	0,501	0,076	0,845	0,989	0,023	0,506	0,198
	SMOTEBagging	0,797	0,901	0,151	0,526	0,336	0,765	0,857	0,224	0,540	0,232	0,740	0,818	0,289	0,554	0,247	0,721	0,787	0,342	0,564	0,266	0,703	0,758	0,385	0,572	0,278
	UnderBagging	0,481	0,469	0,557	0,513	0,245	0,552	0,539	0,631	0,585	0,293	0,587	0,578	0,643	0,610	0,316	0,603	0,596	0,647	0,621	0,327	0,612	0,605	0,647	0,626	0,332
CART	YOK	0,855	0,990	0,015	0,502	0,190	0,845	0,986	0,024	0,505	0,147	0,843	0,985	0,025	0,505	0,116	0,845	0,990	0,019	0,505	0,095	0,848	0,995	0,010	0,503	0,079
	SMOTE	0,817	0,931	0,103	0,517	0,195	0,835	0,967	0,059	0,513	0,167	0,830	0,963	0,063	0,513	0,158	0,839	0,978	0,040	0,509	0,130	0,830	0,963	0,076	0,519	0,128
	SMOTEBoost	0,817	0,933	0,093	0,513	0,095	0,845	0,982	0,030	0,506	0,024	0,840	0,980	0,033	0,507	0,059	0,847	0,992	0,014	0,503	0,026	0,835	0,971	0,059	0,515	0,098
	RUSBoost	0,618	0,652	0,406	0,529	0,171	0,634	0,670	0,413	0,541	0,231	0,484	0,459	0,631	0,545	0,322	0,557	0,557	0,559	0,558	0,312	0,518	0,497	0,632	0,565	0,362
	MWMOTE	0,749	0,832	0,209	0,521	0,236	0,769	0,870	0,190	0,530	0,217	0,816	0,934	0,102	0,518	0,184	0,834	0,966	0,060	0,513	0,163	0,843	0,988	0,024	0,506	0,141
	EasyEnsemble	0,837	0,979	0,023	0,501	0,224	0,852	1,000	0,000	0,500	0,150	0,851	1,000	0,000	0,500	0,095	0,851	1,000	0,000	0,500	0,043	0,850	1,000	0,000	0,500	0,036
	SMOTEBagging	0,711	0,776	0,307	0,541	0,295	0,727	0,800	0,297	0,548	0,244	0,725	0,796	0,321	0,558	0,256	0,698	0,751	0,396	0,574	0,281	0,660	0,690	0,495	0,592	0,303
	UnderBagging	0,527	0,516	0,600	0,558	0,269	0,552	0,542	0,610	0,576	0,284	0,565	0,556	0,617	0,587	0,296	0,572	0,563	0,627	0,595	0,304	0,584	0,576	0,631	0,603	0,312
RF	YOK	0,855	0,988	0,029	0,508	0,154	0,850	0,991	0,024	0,508	0,095	0,850	0,994	0,020	0,507	0,066	0,849	0,995	0,016	0,506	0,043	0,849	0,996	0,014	0,505	0,033
	SMOTE	0,833	0,951	0,097	0,524	0,168	0,839	0,968	0,080	0,524	0,131	0,833	0,965	0,075	0,520	0,114	0,837	0,971	0,069	0,520	0,107	0,837	0,969	0,085	0,527	0,131
	SMOTEBoost	0,839	0,960	0,078	0,519	0,120	0,845	0,977	0,067	0,522	0,106	0,839	0,974	0,062	0,518	0,093	0,842	0,980	0,053	0,516	0,084	0,841	0,977	0,069	0,523	0,110
	RUSBoost	0,592	0,606	0,506	0,556	0,231	0,604	0,609	0,566	0,588	0,310	0,566	0,559	0,605	0,582	0,325	0,607	0,608	0,598	0,603	0,368	0,586	0,573	0,658	0,615	0,434
	MWMOTE	0,837	0,955	0,068	0,511	0,254	0,831	0,962	0,078	0,520	0,157	0,844	0,973	0,060	0,516	0,115	0,842	0,976	0,060	0,518	0,101	0,841	0,980	0,051	0,516	0,087
	EasyEnsemble	0,790	0,908	0,111	0,509	0,220	0,850	0,997	0,008	0,503	0,150	0,850	0,998	0,002	0,500	0,095	0,850	0,998	0,001	0,500	0,043	0,848	0,997	0,001	0,499	0,026
	SMOTEBagging	0,806	0,910	0,152	0,531	0,348	0,803	0,914	0,152	0,533	0,220	0,801	0,911	0,165	0,538	0,198	0,801	0,911	0,169	0,540	0,200	0,797	0,907	0,175	0,541	0,203
	UnderBagging	0,542	0,529	0,625	0,577	0,283	0,565	0,551	0,647	0,599	0,303	0,578	0,564	0,658	0,611	0,316	0,585	0,571	0,662	0,616	0,322	0,591	0,578	0,668	0,623	0,328

Tablo 4.2 düşük düzey korelasyon ve 0,15 dengesizlik durumunu göstermektedir.

Tablo 4.1'de olduğu gibi GDO ve SEÇ değerleri yüksek bulunmuş iken DUY ve F-ölçüsü değerleri düşük olduğu görülmektedir. DDO değerleri ise %50 ve üzerindedir.

Örneklem genişliği arttıkça YOK, SMOTE, SMOTEBoost, MWMOTE ve EasyEnsemble algoritmalarında F-ölçüsü değerleri düşmektedir. UnderBagging algoritmasının sınıflama yöntemleri üzerinde etkisi örneklem genişliği arttıkça benzerdir.

Örneklem genişliği arttıkça RUSBoost algoritması sınıflama yöntemlerinin performanslarını arttırmaktadır. CART yönteminin performansı diğer sınıflama yöntemlerinin performansından daha küçük çıkmaktadır. RUSBoost algoritması örneklem genişliği arttıkça performansları %35'in üzerine çıkarmaktadır. Tablo 4.1'e benzer bir sonuç çıkmaktadır.

Tablo 4.3. Düşük düzey korelasyon ve 0,25 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,748	0,971	0,049	0,510	0,228	0,748	0,983	0,032	0,507	0,171	0,749	0,989	0,023	0,506	0,132	0,750	0,993	0,019	0,506	0,093	0,750	0,993	0,017	0,250	0,073
	SMOTE	0,679	0,813	0,257	0,535	0,263	0,687	0,820	0,281	0,550	0,298	0,695	0,833	0,279	0,556	0,306	0,702	0,842	0,279	0,561	0,314	0,708	0,853	0,271	0,562	0,314
	SMOTEBoost	0,681	0,816	0,256	0,536	0,255	0,685	0,816	0,285	0,551	0,298	0,695	0,833	0,279	0,556	0,305	0,702	0,842	0,280	0,561	0,314	0,708	0,853	0,271	0,562	0,314
	RUSBoost	0,465	0,380	0,733	0,556	0,376	0,522	0,472	0,672	0,572	0,439	0,578	0,562	0,628	0,595	0,449	0,554	0,508	0,693	0,600	0,481	0,584	0,564	0,647	0,605	0,478
	MWMOTE	0,679	0,825	0,238	0,531	0,275	0,653	0,757	0,336	0,547	0,321	0,656	0,761	0,337	0,549	0,326	0,658	0,764	0,339	0,552	0,330	0,659	0,764	0,346	0,555	0,336
	EasyEnsemble	0,717	0,912	0,128	0,520	0,356	0,757	0,953	0,062	0,508	0,244	0,748	0,991	0,018	0,505	0,149	0,740	0,969	0,053	0,511	0,143	0,747	0,983	0,039	0,511	0,095
	SMOTEBagging	0,689	0,832	0,237	0,535	0,304	0,666	0,780	0,316	0,548	0,315	0,652	0,750	0,353	0,552	0,333	0,644	0,730	0,382	0,556	0,347	0,640	0,719	0,405	0,562	0,359
	UnderBagging	0,502	0,458	0,639	0,549	0,384	0,548	0,518	0,640	0,579	0,411	0,565	0,543	0,633	0,588	0,419	0,577	0,561	0,626	0,594	0,424	0,588	0,575	0,627	0,601	0,431
CART	YOK	0,733	0,941	0,080	0,511	0,247	0,739	0,966	0,048	0,507	0,196	0,743	0,975	0,039	0,507	0,177	0,744	0,983	0,028	0,505	0,147	0,749	0,994	0,011	0,250	0,133
	SMOTE	0,653	0,773	0,276	0,524	0,284	0,665	0,798	0,260	0,529	0,288	0,680	0,827	0,239	0,533	0,281	0,688	0,839	0,231	0,535	0,274	0,700	0,862	0,211	0,537	0,266
	SMOTEBoost	0,652	0,769	0,283	0,526	0,245	0,678	0,827	0,222	0,525	0,211	0,689	0,847	0,211	0,529	0,218	0,693	0,854	0,210	0,532	0,227	0,706	0,877	0,189	0,533	0,220
	RUSBoost	0,457	0,389	0,669	0,529	0,343	0,471	0,411	0,652	0,532	0,382	0,524	0,503	0,586	0,545	0,390	0,503	0,458	0,637	0,547	0,417	0,536	0,519	0,586	0,552	0,422
	MWMOTE	0,607	0,684	0,372	0,528	0,323	0,584	0,622	0,470	0,546	0,353	0,639	0,731	0,361	0,546	0,333	0,683	0,830	0,241	0,536	0,325	0,702	0,876	0,179	0,527	0,320
	EasyEnsemble	0,750	0,996	0,005	0,500	0,257	0,769	0,988	0,020	0,504	0,172	0,750	0,999	0,001	0,500	0,090	0,751	0,998	0,011	0,504	0,061	0,750	0,997	0,011	0,504	0,041
	SMOTEBagging	0,635	0,726	0,350	0,538	0,330	0,644	0,741	0,349	0,545	0,323	0,647	0,740	0,363	0,551	0,336	0,631	0,698	0,429	0,564	0,366	0,617	0,656	0,501	0,578	0,394
	UnderBagging	0,522	0,497	0,599	0,548	0,377	0,533	0,508	0,606	0,557	0,390	0,539	0,512	0,620	0,566	0,400	0,548	0,523	0,623	0,573	0,407	0,564	0,546	0,620	0,583	0,415
RF	YOK	0,740	0,946	0,094	0,520	0,174	0,743	0,961	0,081	0,521	0,140	0,745	0,966	0,076	0,521	0,125	0,746	0,973	0,062	0,518	0,106	0,748	0,978	0,055	0,250	0,096
	SMOTE	0,678	0,813	0,254	0,534	0,261	0,683	0,816	0,278	0,547	0,294	0,698	0,842	0,263	0,553	0,297	0,702	0,845	0,269	0,557	0,307	0,707	0,851	0,272	0,562	0,314
	SMOTEBoost	0,681	0,819	0,249	0,534	0,248	0,691	0,829	0,270	0,550	0,287	0,703	0,851	0,254	0,553	0,291	0,706	0,853	0,263	0,558	0,302	0,710	0,858	0,264	0,561	0,309
	RUSBoost	0,529	0,505	0,605	0,555	0,356	0,544	0,528	0,592	0,560	0,412	0,571	0,564	0,592	0,578	0,440	0,559	0,532	0,638	0,585	0,463	0,578	0,567	0,613	0,590	0,467
	MWMOTE	0,695	0,857	0,205	0,531	0,263	0,697	0,853	0,225	0,539	0,265	0,708	0,874	0,209	0,541	0,259	0,715	0,889	0,192	0,540	0,249	0,721	0,901	0,180	0,541	0,242
	EasyEnsemble	0,737	0,956	0,073	0,514	0,350	0,765	0,978	0,035	0,507	0,172	0,748	0,992	0,013	0,503	0,091	0,748	0,991	0,021	0,506	0,086	0,747	0,991	0,016	0,503	0,048
	SMOTEBagging	0,699	0,852	0,218	0,535	0,302	0,698	0,851	0,232	0,542	0,273	0,699	0,853	0,234	0,543	0,275	0,698	0,851	0,235	0,543	0,277	0,698	0,851	0,239	0,545	0,282
	UnderBagging	0,524	0,485	0,646	0,566	0,396	0,534	0,496	0,652	0,574	0,409	0,543	0,503	0,664	0,583	0,418	0,548	0,508	0,668	0,588	0,424	0,555	0,515	0,676	0,595	0,431

Tablo 4.3 düşük düzey korelasyon ve 0,25 dengesizlik durumunu göstermektedir.

GDO değerleri Tablo 4.1 ve Tablo 4.2'ye göre düşük iken DUY ve F-ölçüsü değerlerinin arttığı görülmektedir.

Örneklem genişliği arttıkça YOK ve EasyEnsemble algoritmalarında F-ölçüsü değerleri düşmektedir. SMOTE, SMOTEBoost, RUSBoost, MWMOTE, SMOTEBagging ve UnderBagging algoritmalarının sınıflama yöntemleri üzerinde etkileri örneklem genişliği arttıkça artmaktadır.

Sınıflama yöntemleri üzerinde en etkili RUSBoost algoritması görülmektedir. 0,10 ve 0,15 dengesizlik durumlarına göre sonuçları %5 arttırmıştır. Örneklem genişliği arttıkça RUSBoost algoritması sınıflama yöntemlerinin performanslarını arttırmaktadır. CART yönteminin performansı diğer sınıflama yöntemlerinin performansından daha küçük çıkmaktadır.

Tablo 4.4. Düşük düzey korelasyon ve 0,30 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,691	0,952	0,068	0,510	0,245	0,697	0,969	0,053	0,511	0,205	0,699	0,976	0,047	0,512	0,163	0,700	0,980	0,045	0,512	0,134	0,701	0,982	0,045	0,513	0,117
	SMOTE	0,602	0,688	0,395	0,542	0,355	0,608	0,677	0,443	0,560	0,395	0,631	0,726	0,408	0,567	0,393	0,637	0,731	0,418	0,575	0,406	0,638	0,724	0,435	0,580	0,417
	SMOTEBoost	0,600	0,682	0,404	0,543	0,358	0,605	0,673	0,444	0,558	0,395	0,630	0,725	0,409	0,567	0,393	0,637	0,730	0,418	0,574	0,406	0,638	0,725	0,436	0,580	0,418
	RUSBoost	0,568	0,581	0,542	0,561	0,392	0,537	0,490	0,650	0,570	0,463	0,591	0,594	0,584	0,589	0,463	0,567	0,521	0,674	0,597	0,500	0,589	0,572	0,631	0,601	0,492
	MWMOTE	0,616	0,729	0,345	0,537	0,341	0,611	0,699	0,405	0,552	0,381	0,605	0,680	0,428	0,554	0,392	0,608	0,683	0,431	0,557	0,396	0,607	0,681	0,436	0,558	0,399
	EasyEnsemble	0,687	0,950	0,064	0,507	0,271	0,676	0,916	0,109	0,513	0,248	0,699	0,980	0,039	0,510	0,156	0,699	0,971	0,066	0,518	0,144	0,700	0,963	0,085	0,524	0,144
	SMOTEBagging	0,644	0,802	0,265	0,533	0,320	0,634	0,761	0,334	0,547	0,348	0,624	0,733	0,369	0,551	0,369	0,620	0,717	0,392	0,555	0,381	0,620	0,708	0,412	0,560	0,393
	UnderBagging	0,512	0,456	0,644	0,550	0,434	0,546	0,507	0,639	0,573	0,455	0,561	0,529	0,637	0,583	0,464	0,572	0,549	0,627	0,588	0,467	0,582	0,564	0,625	0,594	0,472
CART	YOK	0,679	0,924	0,092	0,508	0,273	0,681	0,926	0,101	0,513	0,257	0,685	0,939	0,087	0,513	0,219	0,692	0,963	0,057	0,510	0,190	0,696	0,978	0,037	0,507	0,171
	SMOTE	0,585	0,667	0,387	0,527	0,355	0,579	0,641	0,434	0,537	0,368	0,611	0,714	0,370	0,542	0,356	0,617	0,727	0,361	0,544	0,354	0,618	0,724	0,372	0,548	0,361
	SMOTEBoost	0,576	0,645	0,410	0,527	0,331	0,578	0,636	0,439	0,538	0,359	0,609	0,705	0,382	0,544	0,346	0,618	0,729	0,361	0,545	0,340	0,615	0,714	0,384	0,549	0,365
	RUSBoost	0,567	0,619	0,441	0,530	0,305	0,489	0,427	0,636	0,532	0,423	0,556	0,576	0,510	0,543	0,398	0,512	0,464	0,625	0,544	0,450	0,536	0,519	0,577	0,548	0,443
	MWMOTE	0,563	0,609	0,452	0,531	0,384	0,553	0,570	0,514	0,542	0,402	0,573	0,603	0,500	0,552	0,402	0,599	0,668	0,438	0,553	0,387	0,616	0,712	0,392	0,552	0,372
	EasyEnsemble	0,685	0,949	0,060	0,504	0,281	0,696	0,977	0,038	0,508	0,172	0,700	0,990	0,020	0,505	0,123	0,700	0,984	0,036	0,510	0,098	0,702	0,981	0,050	0,516	0,096
	SMOTEBagging	0,607	0,708	0,366	0,537	0,356	0,617	0,724	0,364	0,544	0,357	0,618	0,721	0,377	0,549	0,369	0,611	0,688	0,430	0,559	0,398	0,605	0,652	0,495	0,573	0,428
	UnderBagging	0,520	0,483	0,610	0,546	0,425	0,531	0,497	0,613	0,555	0,436	0,537	0,498	0,627	0,563	0,447	0,543	0,506	0,628	0,567	0,451	0,564	0,539	0,620	0,580	0,460
RF	YOK	0,681	0,909	0,137	0,523	0,201	0,688	0,925	0,127	0,526	0,188	0,692	0,933	0,125	0,529	0,189	0,696	0,943	0,116	0,530	0,182	0,699	0,953	0,106	0,529	0,171
	SMOTE	0,608	0,699	0,387	0,543	0,357	0,607	0,686	0,419	0,553	0,383	0,634	0,737	0,392	0,565	0,388	0,637	0,735	0,407	0,571	0,400	0,638	0,728	0,426	0,577	0,413
	SMOTEBoost	0,608	0,704	0,379	0,541	0,348	0,611	0,695	0,413	0,554	0,379	0,636	0,742	0,388	0,565	0,384	0,639	0,738	0,406	0,572	0,399	0,640	0,732	0,424	0,578	0,412
	RUSBoost	0,556	0,553	0,562	0,558	0,411	0,548	0,531	0,587	0,559	0,441	0,574	0,577	0,567	0,572	0,458	0,562	0,532	0,630	0,581	0,485	0,579	0,568	0,606	0,587	0,483
	MWMOTE	0,637	0,781	0,294	0,538	0,322	0,644	0,792	0,298	0,545	0,330	0,651	0,802	0,299	0,550	0,337	0,660	0,822	0,282	0,552	0,331	0,666	0,837	0,267	0,552	0,323
	EasyEnsemble	0,686	0,938	0,090	0,514	0,251	0,692	0,962	0,059	0,510	0,176	0,700	0,982	0,039	0,510	0,126	0,699	0,978	0,051	0,514	0,115	0,702	0,978	0,059	0,518	0,108
	SMOTEBagging	0,652	0,816	0,259	0,537	0,326	0,657	0,826	0,259	0,542	0,306	0,657	0,825	0,264	0,544	0,312	0,657	0,825	0,265	0,545	0,314	0,659	0,825	0,271	0,548	0,322
	UnderBagging	0,517	0,460	0,652	0,556	0,441	0,530	0,472	0,666	0,569	0,456	0,534	0,472	0,679	0,576	0,466	0,540	0,479	0,683	0,581	0,470	0,547	0,486	0,691	0,588	0,478

Tablo 4.4 düşük düzey korelasyon ve 0,30 dengesizlik durumunu göstermektedir.

GDO değerleri ve SEÇ değerleri düşerken DUY ve F-ölçüsü değerlerinin arttığı görülmektedir.

Örneklem genişliği arttıkça YOK ve EasyEnsemble algoritmalarında F-ölçüsü değerleri düşmektedir. SMOTE, SMOTEBoost, RUSBoost, MWMOTE, SMOTEBagging ve UnderBagging algoritmalarının sınıflama yöntemleri üzerinde etkileri örneklem genişliği arttıkça artmaktadır. SMOTE, SMOTEBoost ve RUSBoost algoritmalarının DVM ve RF sınıflama yöntemlerindeki etkileri CART yöntemindeki göre biraz daha yüksek görülmektedir.

0,30 durumu dengesizliğin azaldığı bir durum olmasına rağmen düşük korelasyon yapısına sahip verilerden kaynaklı olarak örneklem genişliği arttıkça RUSBoost sınıflama performanslarını %45'in üzerinde çıkarmaktadır.

4.1.2 Orta Düzey Korelasyona Ait Sonuçlar

Orta düzey korelasyon tanımlanarak üretilmiş verilerden elde edilen sınıflama sonuçları Grafik 4.2'de gösterilmektedir.

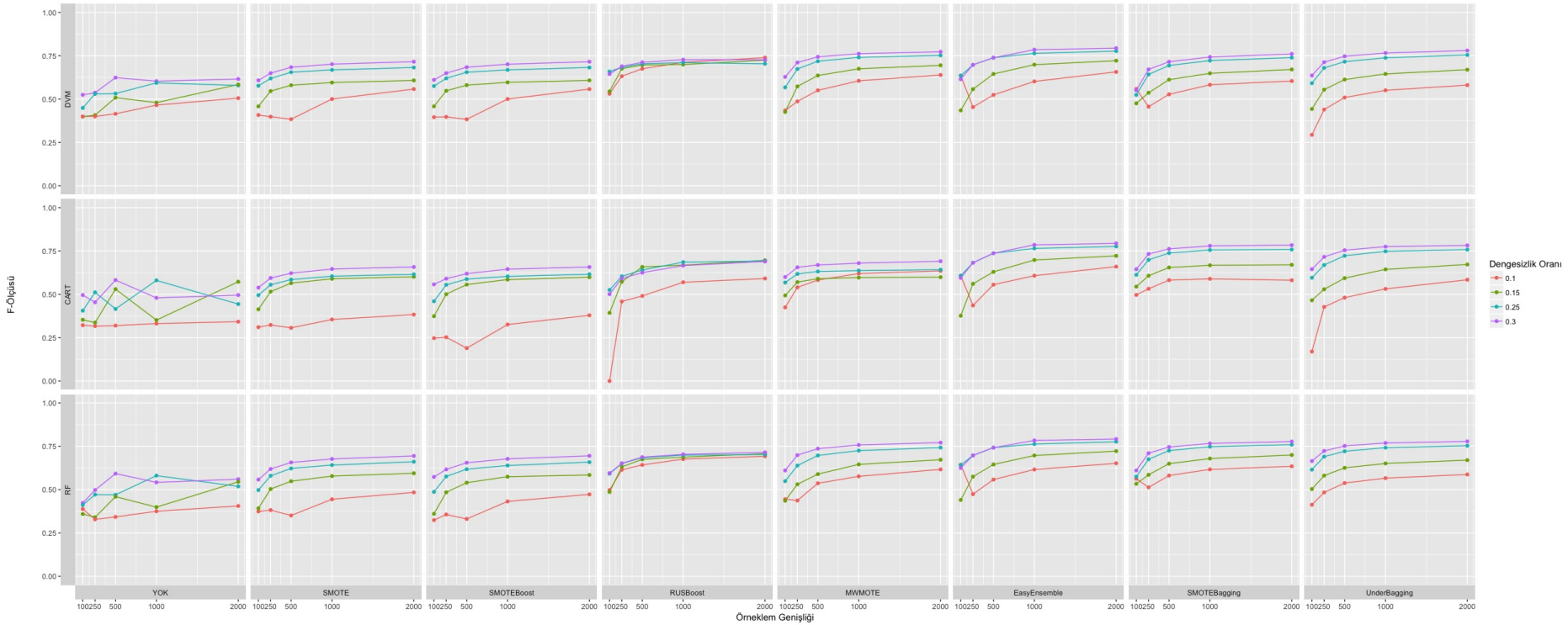
Grafiğe bakıldığında, verideki korelasyon düzeyinin artmasından dolayı algoritmaların farklı dengesizlik oranlarında sınıflama yöntemlerindeki performans sonuçlarına yansımıştır.

4 farklı dengesizlik oranında örneklem genişliği arttıkça sınıflama yöntemlerinde performanslar artmaktadır. Her bir algoritma tek tek değerlendirilmiştir. Algoritma olmadan uygulanan sınıflama yöntemlerinin performansları değişkenlik göstermektedir. SMOTE algoritmasının 0,10 dengesizlik durumunda sınıflama performans sonuçlarına etkisi görülmezken diğer dengesizlik durumlarında örneklem genişliği arttıkça sınıflama performanslarını arttırdığı görülmektektedir. SMOTEBoost algoritması SMOTE algoritmasına benzer sonuçlar vermektedir. CART sınıflama yöntemi (2,4) ve (2,8) hücrelerinde görüldüğü gibi 0,10 dengesizlik durumunda örneklem genişliği 100 olduğunda diğer sınıflama yöntemlerine göre performansı düşük iken örneklem genişliği arttıkça performansı artmıştır.

RUSBoost algoritmasında 0,10 dengesizlik hariç diğer dengesizlik oranlarında sınıflama yöntemlerinde performanslarını arttırmakta ve performanslar benzer bulunmaktadır.

EasyEnsemble, MWMOTE, SMOTEBagging ve UnderBagging algoritmaları DVM yönteminde performanslar üzerinde benzer etkiye sahip iken CART yönteminde MWMOTE yönteminin etkisi diğer algoritmalara göre düşük bulunmuştur.

Dört farklı dengesizlik durumuna ait sonuçlar Tablo 4.5, Tablo 4.6, Tablo 4.7 ve Tablo 4.8'de verilmiştir.



Şekil 4.2. Orta düzey korelasyon sonuçları

Tablo 4.5. Orta düzey korelasyon ve 0,10 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,910	0,985	0,138	0,562	0,400	0,910	0,980	0,244	0,612	0,400	0,912	0,977	0,309	0,643	0,415	0,919	0,979	0,371	0,099	0,465	0,922	0,978	0,409	0,694	0,506
	SMOTE	0,910	0,980	0,215	0,598	0,408	0,907	0,967	0,331	0,649	0,398	0,909	0,973	0,310	0,641	0,384	0,914	0,966	0,443	0,705	0,500	0,920	0,965	0,512	0,739	0,558
	SMOTEBoost	0,910	0,979	0,218	0,599	0,396	0,907	0,966	0,335	0,651	0,397	0,909	0,973	0,311	0,642	0,384	0,914	0,966	0,443	0,705	0,500	0,920	0,965	0,512	0,739	0,558
	RUSBoost	0,733	0,729	0,775	0,752	0,531	0,766	0,757	0,861	0,809	0,632	0,786	0,776	0,876	0,826	0,675	0,806	0,796	0,899	0,847	0,712	0,817	0,806	0,919	0,863	0,739
	MWMOTE	0,911	0,983	0,220	0,602	0,434	0,916	0,970	0,418	0,694	0,487	0,924	0,972	0,484	0,728	0,550	0,927	0,967	0,566	0,767	0,605	0,930	0,963	0,626	0,795	0,639
	EasyEnsemble	0,791	0,860	0,625	0,743	0,635	0,896	0,949	0,401	0,675	0,454	0,915	0,963	0,477	0,720	0,524	0,923	0,962	0,574	0,768	0,602	0,931	0,960	0,663	0,812	0,657
	SMOTEBagging	0,912	0,973	0,315	0,644	0,558	0,915	0,972	0,369	0,670	0,456	0,915	0,961	0,494	0,728	0,528	0,913	0,945	0,618	0,782	0,583	0,909	0,932	0,701	0,816	0,604
	UnderBagging	0,542	0,525	0,710	0,617	0,294	0,789	0,782	0,854	0,818	0,440	0,830	0,823	0,891	0,857	0,509	0,853	0,848	0,905	0,876	0,551	0,868	0,864	0,913	0,888	0,581
CART	YOK	0,903	0,988	0,050	0,519	0,322	0,897	0,976	0,144	0,560	0,317	0,900	0,978	0,176	0,577	0,320	0,902	0,978	0,203	0,099	0,332	0,904	0,978	0,232	0,605	0,342
	SMOTE	0,883	0,953	0,189	0,571	0,311	0,883	0,946	0,274	0,610	0,324	0,891	0,964	0,204	0,584	0,307	0,895	0,961	0,287	0,624	0,355	0,896	0,959	0,326	0,643	0,383
	SMOTEBoost	0,881	0,950	0,193	0,571	0,247	0,885	0,952	0,237	0,595	0,253	0,894	0,973	0,152	0,563	0,190	0,895	0,962	0,280	0,621	0,326	0,897	0,959	0,330	0,644	0,379
	RUSBoost	0,909	1,000	0,000	0,500	0,000	0,636	0,624	0,753	0,689	0,459	0,685	0,678	0,742	0,710	0,491	0,716	0,711	0,764	0,738	0,570	0,725	0,719	0,782	0,750	0,592
	MWMOTE	0,865	0,918	0,356	0,637	0,425	0,900	0,934	0,586	0,760	0,540	0,920	0,959	0,564	0,762	0,583	0,930	0,969	0,579	0,774	0,620	0,935	0,974	0,574	0,774	0,635
	EasyEnsemble	0,769	0,858	0,558	0,708	0,607	0,879	0,945	0,276	0,610	0,436	0,909	0,975	0,299	0,637	0,556	0,921	0,968	0,496	0,732	0,608	0,932	0,960	0,670	0,815	0,660
	SMOTEBagging	0,852	0,873	0,652	0,762	0,497	0,883	0,902	0,693	0,798	0,532	0,890	0,903	0,771	0,837	0,582	0,883	0,887	0,848	0,867	0,590	0,871	0,868	0,894	0,881	0,581
	UnderBagging	0,493	1,000	0,100	0,500	0,170	0,780	0,775	0,826	0,800	0,427	0,816	0,813	0,838	0,826	0,481	0,843	0,838	0,887	0,863	0,532	0,870	0,866	0,907	0,887	0,584
RF	YOK	0,912	0,987	0,144	0,566	0,388	0,912	0,984	0,215	0,600	0,329	0,912	0,984	0,244	0,614	0,342	0,914	0,984	0,272	0,099	0,375	0,915	0,983	0,299	0,641	0,406
	SMOTE	0,907	0,978	0,201	0,590	0,374	0,908	0,970	0,312	0,641	0,382	0,907	0,975	0,276	0,625	0,351	0,910	0,969	0,373	0,671	0,445	0,912	0,966	0,421	0,693	0,484
	SMOTEBoost	0,908	0,979	0,202	0,591	0,324	0,909	0,973	0,289	0,631	0,357	0,908	0,978	0,256	0,617	0,331	0,911	0,971	0,357	0,664	0,432	0,913	0,969	0,404	0,686	0,473
	RUSBoost	0,742	0,744	0,727	0,735	0,497	0,762	0,756	0,820	0,788	0,614	0,779	0,775	0,818	0,796	0,643	0,796	0,792	0,835	0,813	0,676	0,806	0,801	0,855	0,828	0,692
	MWMOTE	0,911	0,985	0,194	0,590	0,445	0,921	0,985	0,328	0,657	0,437	0,929	0,981	0,453	0,717	0,537	0,932	0,981	0,489	0,735	0,577	0,935	0,979	0,538	0,759	0,617
	EasyEnsemble	0,797	0,867	0,629	0,708	0,644	0,903	0,953	0,442	0,610	0,474	0,921	0,965	0,517	0,637	0,558	0,929	0,967	0,583	0,732	0,616	0,932	0,964	0,640	0,815	0,652
	SMOTEBagging	0,913	0,965	0,402	0,684	0,563	0,920	0,969	0,452	0,710	0,513	0,923	0,962	0,559	0,761	0,581	0,922	0,953	0,641	0,797	0,617	0,921	0,946	0,691	0,818	0,634
	UnderBagging	0,752	0,740	0,865	0,803	0,413	0,819	0,814	0,870	0,842	0,484	0,846	0,840	0,901	0,871	0,538	0,861	0,855	0,913	0,884	0,566	0,871	0,865	0,921	0,893	0,587

Tablo 4.5 orta düzey korelasyon ve 0,10 dengesizlik durumunu göstermektedir.

Tablo 4.5'e göre GDO değerlerine bakıldığında, UnderBagging algoritması hariç YOK ve diğer algoritmalarda yüksek bulunmuştur, SEÇ değerleri çok yüksek çıkmış iken DUY değerlerine bakıldığında çok düşük çıkmasından dolayı GDO değerine bakmak yanıltıcı olabilmektedir. YOK ve tüm algoritmalarda DDO değerleri yaklaşık %50 ve üzeri civarındadır. Bunun aksine F-ölçüsü DUY değerlerinden etkilendiği için 0,10 dengesizlik durumunda çok düşük değerlere sahiptir.

Örneklem genişliği arttıkça YOK ve algoritmaların F-ölçüsü değerleri arttırdığı görülmektedir. Örneklem genişliği 100 iken EasyEnsemble algoritması DVM, CART ve RF yöntemlerinin sınıflama performansları üzerinde etkili olmuş iken örneklem genişliği arttıkça RUSBoost algoritmasının diğerlerine göre daha etkilidir.

2000 örneklem genişliğinde RUSBoost algoritması F-ölçüsü değerini DVM yönteminde %70'in üzerine çıkarmıştır. Verilerin orta düzey korelasyona sahip olması performansları yükseltmiştir.

Tablo 4.6. Orta düzey korelasyon ve 0,15 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,862	0,969	0,192	0,580	0,399	0,864	0,965	0,270	0,617	0,407	0,884	0,966	0,411	0,148	0,508	0,879	0,965	0,386	0,676	0,480	0,898	0,971	0,483	0,150	0,584
	SMOTE	0,863	0,933	0,422	0,677	0,458	0,877	0,937	0,523	0,730	0,546	0,877	0,929	0,577	0,753	0,580	0,878	0,925	0,608	0,766	0,595	0,884	0,934	0,602	0,768	0,608
	SMOTEBoost	0,864	0,934	0,426	0,680	0,458	0,877	0,937	0,526	0,731	0,548	0,877	0,929	0,578	0,754	0,581	0,878	0,925	0,610	0,767	0,596	0,884	0,934	0,602	0,768	0,608
	RUSBoost	0,751	0,761	0,689	0,725	0,545	0,769	0,764	0,796	0,780	0,675	0,759	0,743	0,848	0,796	0,697	0,790	0,778	0,859	0,818	0,699	0,797	0,783	0,874	0,829	0,725
	MWMOTE	0,874	0,976	0,284	0,630	0,426	0,892	0,959	0,503	0,731	0,573	0,899	0,952	0,596	0,774	0,636	0,904	0,946	0,666	0,806	0,675	0,908	0,944	0,703	0,823	0,695
	EasyEnsemble	0,810	0,890	0,359	0,625	0,435	0,879	0,942	0,515	0,729	0,557	0,893	0,936	0,649	0,793	0,645	0,909	0,945	0,703	0,824	0,699	0,914	0,944	0,744	0,844	0,721
	SMOTEBagging	0,876	0,963	0,333	0,648	0,475	0,883	0,951	0,481	0,716	0,536	0,888	0,938	0,601	0,769	0,612	0,889	0,924	0,688	0,806	0,648	0,889	0,913	0,755	0,834	0,671
	UnderBagging	0,686	0,661	0,838	0,750	0,443	0,796	0,786	0,854	0,820	0,554	0,833	0,824	0,884	0,854	0,613	0,853	0,846	0,893	0,869	0,645	0,866	0,860	0,904	0,882	0,670
CART	YOK	0,849	0,967	0,113	0,540	0,353	0,848	0,964	0,169	0,566	0,337	0,879	0,955	0,443	0,148	0,530	0,855	0,962	0,242	0,602	0,352	0,893	0,965	0,484	0,150	0,573
	SMOTE	0,813	0,874	0,434	0,654	0,414	0,858	0,917	0,512	0,714	0,516	0,872	0,926	0,564	0,745	0,565	0,876	0,924	0,604	0,764	0,590	0,881	0,929	0,606	0,767	0,601
	SMOTEBoost	0,815	0,878	0,424	0,651	0,374	0,858	0,918	0,505	0,711	0,501	0,873	0,928	0,554	0,741	0,556	0,878	0,927	0,594	0,761	0,585	0,881	0,931	0,600	0,765	0,599
	RUSBoost	0,712	0,734	0,574	0,654	0,393	0,735	0,741	0,698	0,720	0,573	0,730	0,723	0,772	0,748	0,658	0,762	0,758	0,784	0,771	0,668	0,767	0,758	0,824	0,791	0,696
	MWMOTE	0,828	0,891	0,463	0,677	0,493	0,872	0,924	0,573	0,749	0,571	0,879	0,930	0,586	0,758	0,590	0,879	0,927	0,601	0,764	0,597	0,879	0,927	0,606	0,766	0,599
	EasyEnsemble	0,830	0,953	0,134	0,544	0,377	0,869	0,941	0,460	0,701	0,561	0,891	0,938	0,625	0,781	0,630	0,908	0,943	0,709	0,826	0,698	0,913	0,942	0,752	0,847	0,722
	SMOTEBagging	0,837	0,865	0,660	0,763	0,545	0,863	0,886	0,725	0,805	0,607	0,874	0,887	0,801	0,844	0,655	0,873	0,876	0,855	0,865	0,668	0,868	0,865	0,890	0,877	0,670
	UnderBagging	0,737	0,729	0,784	0,757	0,466	0,780	0,773	0,820	0,796	0,529	0,821	0,814	0,862	0,838	0,594	0,852	0,846	0,888	0,867	0,644	0,867	0,860	0,906	0,883	0,672
RF	YOK	0,866	0,972	0,203	0,587	0,359	0,866	0,972	0,245	0,609	0,341	0,879	0,967	0,370	0,148	0,459	0,870	0,970	0,297	0,634	0,399	0,891	0,970	0,442	0,150	0,545
	SMOTE	0,852	0,934	0,340	0,637	0,392	0,869	0,936	0,478	0,707	0,504	0,872	0,931	0,533	0,732	0,548	0,875	0,927	0,580	0,754	0,578	0,879	0,929	0,596	0,762	0,594
	SMOTEBoost	0,854	0,939	0,323	0,631	0,361	0,871	0,942	0,454	0,698	0,485	0,874	0,936	0,519	0,727	0,540	0,877	0,931	0,571	0,751	0,575	0,883	0,937	0,575	0,756	0,584
	RUSBoost	0,708	0,718	0,642	0,680	0,486	0,750	0,753	0,733	0,743	0,631	0,756	0,748	0,803	0,775	0,675	0,791	0,785	0,825	0,805	0,688	0,793	0,781	0,861	0,821	0,707
	MWMOTE	0,872	0,975	0,275	0,625	0,436	0,894	0,974	0,436	0,705	0,531	0,903	0,976	0,489	0,733	0,589	0,911	0,972	0,563	0,767	0,646	0,914	0,970	0,600	0,785	0,672
	EasyEnsemble	0,832	0,917	0,352	0,635	0,440	0,885	0,945	0,539	0,742	0,575	0,898	0,945	0,628	0,787	0,645	0,910	0,950	0,687	0,819	0,697	0,915	0,947	0,736	0,841	0,722
	SMOTEBagging	0,883	0,956	0,429	0,692	0,533	0,892	0,952	0,540	0,746	0,586	0,898	0,943	0,641	0,792	0,650	0,901	0,936	0,703	0,820	0,679	0,904	0,931	0,749	0,840	0,700
	UnderBagging	0,757	0,743	0,844	0,793	0,503	0,812	0,802	0,872	0,837	0,581	0,838	0,827	0,902	0,864	0,626	0,854	0,845	0,908	0,876	0,651	0,865	0,856	0,915	0,885	0,670

Tablo 4.6 orta düzey korelasyon ve 0,15 dengesizlik durumunu göstermektedir.

Tablo 4.6'de GDO değerlerine bakıldığında, düşük düzey korelasyon yapısındaki sonuçlara göre düşüş olmuştur. SEÇ değerleri yüksek iken DUY değerlerine bakıldığında düşük çıkmasından dolayı GDO değerine bakmak yanıltıcı olabilmektedir. Örneklem genişliği arttıkça YOK ve tüm algoritmalarda DDO değerleri yaklaşık %50 ve üzeri civarındadır. F-ölçüsü değerleri YOK'a göre diğer algoritmaların sınıflama üzerine etkisi daha fazladır.

Örneklem genişliği 100 iken CART ve RF yöntemlerinde en fazla katkı SMOTEBagging algoritması tarafından sağlanırken, DVM yönteminde RUSBoost algoritmasının katkısı diğerlerine göre fazladır. Örneklem genişliği arttıkça EasyEnsemble algoritması F-ölçüsü değerlerini %70'in üzerine çıkarmaktadır.

Genel olarak, orta düzey korelasyon yapısına sahip veri türetildiği için düşük korelasyon yapısına sahip verilerin sonuçlarına göre algoritmaların sınıflama yöntemleri performansları üzerinde daha etkili olduğu görülmektedir.

Tablo 4.7. Orta düzey korelasyon ve 0,25 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,776	0,923	0,314	0,619	0,449	0,809	0,927	0,449	0,247	0,530	0,805	0,921	0,455	0,688	0,532	0,827	0,929	0,517	0,249	0,593	0,820	0,927	0,498	0,712	0,579
	SMOTE	0,800	0,868	0,583	0,726	0,577	0,804	0,856	0,648	0,752	0,619	0,818	0,858	0,698	0,778	0,655	0,820	0,850	0,729	0,790	0,668	0,824	0,845	0,761	0,803	0,682
	SMOTEBoost	0,798	0,866	0,584	0,725	0,575	0,804	0,855	0,649	0,752	0,620	0,819	0,858	0,698	0,778	0,655	0,820	0,851	0,729	0,790	0,668	0,824	0,845	0,761	0,803	0,682
	RUSBoost	0,692	0,669	0,765	0,717	0,659	0,733	0,711	0,798	0,755	0,684	0,763	0,746	0,814	0,780	0,704	0,766	0,744	0,832	0,788	0,710	0,805	0,802	0,815	0,808	0,704
	MWMOTE	0,818	0,921	0,504	0,713	0,568	0,847	0,914	0,644	0,779	0,674	0,859	0,906	0,720	0,813	0,718	0,869	0,907	0,752	0,830	0,741	0,872	0,905	0,775	0,840	0,752
	EasyEnsemble	0,791	0,860	0,625	0,743	0,635	0,850	0,900	0,700	0,800	0,698	0,868	0,908	0,750	0,829	0,739	0,880	0,914	0,777	0,846	0,764	0,887	0,919	0,791	0,855	0,777
	SMOTEBagging	0,811	0,929	0,442	0,685	0,524	0,837	0,915	0,599	0,757	0,642	0,850	0,905	0,685	0,795	0,694	0,858	0,895	0,743	0,819	0,722	0,862	0,889	0,783	0,836	0,739
	UnderBagging	0,721	0,689	0,822	0,756	0,592	0,798	0,778	0,859	0,819	0,680	0,827	0,812	0,874	0,843	0,716	0,844	0,831	0,881	0,856	0,738	0,856	0,846	0,886	0,866	0,755
CART	YOK	0,748	0,907	0,251	0,579	0,406	0,797	0,925	0,405	0,247	0,512	0,763	0,910	0,321	0,615	0,416	0,827	0,938	0,491	0,249	0,581	0,779	0,919	0,355	0,637	0,444
	SMOTE	0,745	0,802	0,565	0,683	0,495	0,760	0,808	0,616	0,712	0,555	0,783	0,835	0,624	0,730	0,585	0,795	0,847	0,638	0,742	0,605	0,793	0,832	0,675	0,753	0,616
	SMOTEBoost	0,744	0,808	0,543	0,676	0,460	0,760	0,809	0,612	0,710	0,555	0,780	0,831	0,626	0,729	0,588	0,795	0,849	0,635	0,742	0,604	0,793	0,831	0,676	0,754	0,616
	RUSBoost	0,617	0,592	0,694	0,643	0,526	0,703	0,700	0,714	0,707	0,604	0,721	0,720	0,726	0,723	0,640	0,712	0,683	0,799	0,741	0,686	0,720	0,682	0,835	0,758	0,693
	MWMOTE	0,776	0,837	0,593	0,715	0,568	0,813	0,879	0,614	0,746	0,618	0,824	0,894	0,612	0,753	0,632	0,833	0,912	0,593	0,753	0,637	0,836	0,918	0,592	0,755	0,642
	EasyEnsemble	0,769	0,858	0,558	0,708	0,607	0,841	0,890	0,693	0,791	0,682	0,866	0,903	0,754	0,829	0,737	0,880	0,912	0,782	0,847	0,765	0,886	0,917	0,794	0,856	0,777
	SMOTEBagging	0,797	0,839	0,667	0,753	0,613	0,837	0,861	0,765	0,813	0,699	0,856	0,870	0,815	0,842	0,738	0,862	0,864	0,855	0,860	0,756	0,861	0,857	0,874	0,865	0,759
	UnderBagging	0,741	0,728	0,783	0,755	0,596	0,794	0,781	0,833	0,807	0,669	0,833	0,822	0,868	0,845	0,723	0,849	0,835	0,894	0,864	0,748	0,857	0,845	0,893	0,869	0,758
RF	YOK	0,781	0,925	0,328	0,627	0,411	0,800	0,938	0,380	0,247	0,471	0,792	0,930	0,378	0,654	0,470	0,822	0,929	0,501	0,249	0,581	0,803	0,928	0,428	0,678	0,519
	SMOTE	0,774	0,865	0,489	0,677	0,497	0,789	0,852	0,596	0,724	0,580	0,808	0,864	0,640	0,752	0,622	0,813	0,859	0,675	0,767	0,642	0,819	0,856	0,709	0,782	0,661
	SMOTEBoost	0,775	0,870	0,478	0,674	0,487	0,792	0,860	0,588	0,724	0,576	0,811	0,870	0,630	0,750	0,618	0,816	0,865	0,669	0,767	0,639	0,822	0,861	0,704	0,782	0,659
	RUSBoost	0,658	0,647	0,695	0,671	0,592	0,711	0,696	0,754	0,725	0,651	0,745	0,733	0,781	0,757	0,684	0,757	0,739	0,811	0,775	0,699	0,786	0,778	0,810	0,794	0,702
	MWMOTE	0,820	0,938	0,463	0,700	0,549	0,849	0,945	0,556	0,751	0,638	0,865	0,943	0,631	0,787	0,697	0,876	0,946	0,664	0,805	0,725	0,881	0,945	0,689	0,817	0,743
	EasyEnsemble	0,795	0,865	0,628	0,746	0,642	0,850	0,902	0,695	0,798	0,697	0,871	0,911	0,750	0,830	0,742	0,878	0,914	0,770	0,842	0,763	0,886	0,919	0,788	0,854	0,777
	SMOTEBagging	0,823	0,923	0,512	0,717	0,575	0,849	0,917	0,642	0,780	0,675	0,865	0,914	0,718	0,816	0,725	0,872	0,910	0,759	0,835	0,748	0,877	0,908	0,781	0,845	0,759
	UnderBagging	0,746	0,719	0,830	0,774	0,616	0,803	0,779	0,878	0,828	0,690	0,827	0,805	0,894	0,850	0,721	0,843	0,822	0,905	0,864	0,742	0,852	0,833	0,906	0,870	0,753

Tablo 4.7 orta düzey korelasyon ve 0,25 dengesizlik durumunu göstermektedir.

Tablo 4.7'e bakıldığında GDO değerleri %70 ve civarında iken DUY ve DDO değerleri artmıştır. DDO değerleri %70 civarındadır.

F-ölçüsü değerleri, örneklem genişliği 100 iken DVM yönteminde RUSBoost ve EasyEnsemble algoritmaları, CART yönteminde EasyEnsemble ve SMOTEBagging algoritmaları, RF yönteminde ise EasyEnsemble ve UnderBagging algoritmaları %60 ve üzerinde çıkmıştır. Örneklem genişliği arttıkça EasyEnsemble algoritması üç sınıflama yöntemindeki performansları %70'in üzerine çıkarmıştır.

Dengesizlik azaldıkça ve korelasyon yapısının orta düzey olmasından dolayı performanslar yükselmiştir. YOK'a göre algoritmaların sınıflama yöntemleri performanslarını net bir şekilde arttırdığı görülmektedir.

Tablo 4.8. Orta düzey korelasyon ve 0,30 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,760	0,895	0,438	0,295	0,524	0,763	0,886	0,471	0,679	0,537	0,801	0,901	0,564	0,299	0,624	0,789	0,895	0,540	0,718	0,603	0,795	0,899	0,550	0,725	0,616
	SMOTE	0,764	0,820	0,633	0,726	0,608	0,777	0,811	0,697	0,754	0,650	0,800	0,832	0,725	0,778	0,684	0,810	0,836	0,747	0,792	0,701	0,819	0,845	0,759	0,802	0,716
	SMOTEBoost	0,764	0,817	0,640	0,728	0,611	0,777	0,811	0,697	0,754	0,650	0,801	0,833	0,725	0,779	0,684	0,810	0,837	0,746	0,791	0,701	0,819	0,845	0,759	0,802	0,716
	RUSBoost	0,726	0,728	0,721	0,725	0,645	0,733	0,709	0,790	0,750	0,689	0,765	0,754	0,790	0,772	0,712	0,764	0,745	0,808	0,776	0,727	0,802	0,809	0,785	0,797	0,727
	MWMOTE	0,801	0,893	0,583	0,738	0,628	0,828	0,875	0,715	0,795	0,711	0,844	0,880	0,759	0,819	0,744	0,852	0,880	0,789	0,834	0,762	0,858	0,880	0,806	0,843	0,773
	EasyEnsemble	0,785	0,868	0,589	0,728	0,615	0,700	0,991	0,017	0,504	0,203	0,699	0,979	0,042	0,511	0,156	0,870	0,903	0,793	0,848	0,785	0,875	0,906	0,803	0,854	0,794
	SMOTEBagging	0,781	0,909	0,476	0,693	0,551	0,819	0,899	0,631	0,765	0,671	0,834	0,890	0,703	0,796	0,715	0,845	0,885	0,750	0,817	0,743	0,852	0,881	0,783	0,832	0,760
	UnderBagging	0,716	0,669	0,829	0,749	0,636	0,793	0,764	0,860	0,812	0,713	0,823	0,803	0,871	0,837	0,747	0,840	0,825	0,874	0,849	0,766	0,852	0,840	0,878	0,859	0,780
CART	YOK	0,726	0,871	0,379	0,295	0,496	0,717	0,864	0,368	0,616	0,454	0,787	0,904	0,512	0,299	0,583	0,738	0,877	0,412	0,644	0,480	0,745	0,884	0,422	0,653	0,496
	SMOTE	0,710	0,752	0,609	0,680	0,539	0,722	0,732	0,697	0,714	0,594	0,739	0,745	0,725	0,735	0,622	0,752	0,750	0,756	0,753	0,646	0,763	0,764	0,760	0,762	0,658
	SMOTEBoost	0,711	0,761	0,591	0,676	0,558	0,723	0,739	0,686	0,712	0,591	0,740	0,749	0,718	0,733	0,619	0,751	0,751	0,754	0,752	0,646	0,764	0,766	0,759	0,762	0,658
	RUSBoost	0,675	0,717	0,575	0,646	0,501	0,708	0,717	0,687	0,702	0,592	0,722	0,721	0,723	0,722	0,625	0,712	0,673	0,802	0,737	0,666	0,715	0,662	0,839	0,750	0,689
	MWMOTE	0,754	0,810	0,620	0,715	0,600	0,790	0,836	0,680	0,758	0,656	0,799	0,845	0,691	0,768	0,669	0,800	0,834	0,720	0,777	0,680	0,804	0,829	0,745	0,787	0,691
	EasyEnsemble	0,762	0,857	0,536	0,696	0,595	0,700	0,994	0,010	0,502	0,178	0,700	0,989	0,024	0,506	0,121	0,870	0,901	0,796	0,849	0,786	0,875	0,905	0,804	0,855	0,794
	SMOTEBagging	0,781	0,823	0,683	0,753	0,645	0,832	0,855	0,777	0,816	0,733	0,849	0,863	0,814	0,839	0,762	0,858	0,865	0,840	0,853	0,780	0,858	0,856	0,862	0,859	0,784
	UnderBagging	0,742	0,721	0,791	0,756	0,645	0,801	0,785	0,839	0,812	0,716	0,830	0,812	0,872	0,842	0,755	0,846	0,831	0,883	0,857	0,775	0,853	0,841	0,881	0,861	0,782
RF	YOK	0,741	0,909	0,339	0,295	0,423	0,755	0,898	0,419	0,658	0,497	0,791	0,907	0,518	0,299	0,593	0,769	0,901	0,460	0,680	0,542	0,775	0,901	0,480	0,691	0,560
	SMOTE	0,740	0,812	0,570	0,691	0,558	0,759	0,800	0,662	0,731	0,619	0,784	0,823	0,693	0,758	0,657	0,791	0,818	0,729	0,774	0,677	0,804	0,831	0,743	0,787	0,695
	SMOTEBoost	0,741	0,814	0,565	0,690	0,573	0,763	0,808	0,657	0,733	0,617	0,788	0,830	0,689	0,759	0,655	0,795	0,823	0,731	0,777	0,677	0,809	0,837	0,744	0,791	0,695
	RUSBoost	0,678	0,688	0,654	0,671	0,595	0,713	0,702	0,739	0,721	0,652	0,747	0,740	0,762	0,751	0,687	0,752	0,734	0,794	0,764	0,704	0,780	0,776	0,791	0,783	0,715
	MWMOTE	0,802	0,911	0,543	0,727	0,611	0,836	0,916	0,647	0,782	0,699	0,853	0,922	0,692	0,807	0,736	0,862	0,922	0,723	0,822	0,758	0,869	0,923	0,740	0,832	0,771
	EasyEnsemble	0,785	0,860	0,609	0,734	0,625	0,699	0,986	0,025	0,506	0,125	0,698	0,978	0,041	0,510	0,129	0,868	0,901	0,791	0,846	0,784	0,873	0,904	0,799	0,852	0,792
	SMOTEBagging	0,801	0,906	0,553	0,729	0,611	0,837	0,904	0,680	0,792	0,710	0,851	0,900	0,736	0,818	0,746	0,860	0,899	0,768	0,834	0,766	0,866	0,900	0,784	0,842	0,777
	UnderBagging	0,746	0,706	0,840	0,773	0,665	0,798	0,763	0,882	0,822	0,723	0,823	0,792	0,896	0,844	0,752	0,838	0,811	0,901	0,856	0,769	0,846	0,822	0,902	0,862	0,778

Tablo 4.8 orta düzey korelasyon ve 0,30 dengesizlik durumunu göstermektedir.

Tablo 4.8'de örneklem genişliği arttıkça performanslarda da artış olmaktadır. F-ölçüsü değerlerine bakıldığında EasyEnsemble algoritmasının sınıflama yöntemlerindeki performansı 2000 örneklem genişliğinde %75 ve üzerindedir.

Dengesizlik azaldıkça DUY, DDO ve F-ölçüsü değerlerinde yükselme olmuştur, YOK, SMOTE, SMOTEBoost algoritmalarının etkisi diğerlerine göre daha düşüktür.

4.1.3 Yüksek Düzey Korelasyona Ait Sonuçlar

Grafik 4.3'de Yüksek düzey korelasyon tanımlanarak üretilmiş verilerden elde edilen sınıflama sonuçları gösterilmektedir.

Grafiğe bakıldığında, verinin yüksek düzey korelasyona sahip olmasından dolayı algoritmaların sınıflama yöntemlerindeki performans sonuçlarına etkisi net bir şekilde görülmektedir.

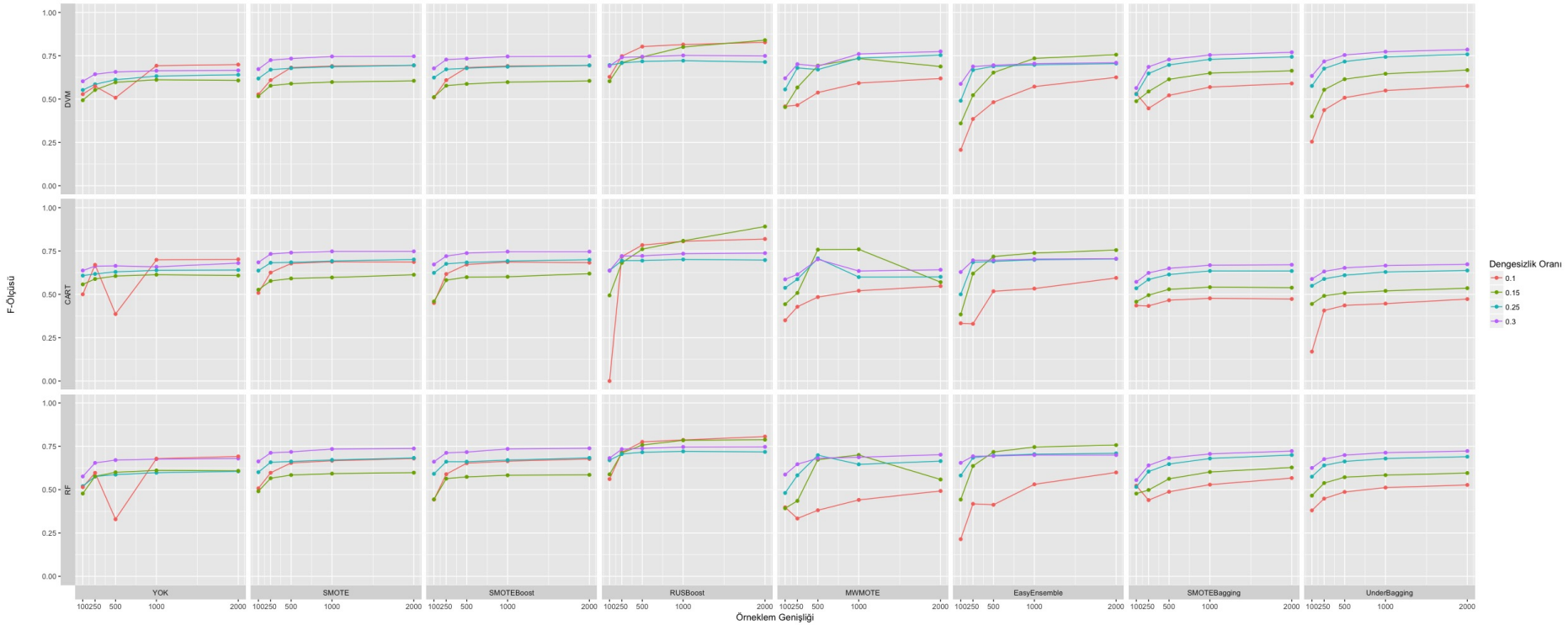
0,10 dengesizlik durumunda SMOTE, SMOTEBoost ve RUSBoost algoritmalarının etkileri diğer algoritmalara göre daha fazla bulunmuştur. Dengesizlik oranının azalmasıyla performans sonuçlarında da bir artış görülmektedir.

Örneklem genişliğinin 500'den 2000'e artması performans sonuçlarını çok değiştirmezken 100'den 250'ye artması sonuçlar üzerinde etkili olmuştur. CART sınıflama yöntemi orta düzey korelasyonda olduğu gibi (2,4) hücresinde 0,10 dengesizlik durumunda örneklem genişliği 100 olduğunda diğer sınıflama yöntemlerine göre performansı düşük iken örneklem genişliği arttıkça performansı artmıştır.

RUSBoost algoritmasında 0,25 ve 0,30 dengesizlik oranlarında sınıflama yöntemlerindeki performanslar benzerken 0,10 ve 0,15 dengesizliklerde sınıflama yöntemlerindeki performansları arttırmaktadır.

MWMOTE algoritmasının CART ve RF yöntemlerinde performanslar üzerinde değişkenlik göstermektedir. Ama DVM yönteminde örneklem genişliği arttıkça etkisi de artmaktadır.

Dört farklı dengesizlik durumuna ait sonuçlar Tablo 4.9, Tablo 4.10, Tablo 4.11 ve Tablo 4.12'de verilmiştir.



Şekil 4.3. Yüksek düzey korelasyon sonuçları

Tablo 4.9. Yüksek düzey korelasyon ve 0,10 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,919	0,987	0,239	0,091	0,528	0,896	0,960	0,508	0,734	0,573	0,922	0,975	0,429	0,702	0,508	0,948	0,987	0,598	0,792	0,692	0,949	0,987	0,603	0,100	0,699
	SMOTE	0,921	0,980	0,326	0,653	0,527	0,935	0,974	0,559	0,766	0,609	0,946	0,983	0,606	0,794	0,682	0,947	0,984	0,609	0,796	0,691	0,948	0,986	0,601	0,794	0,694
	SMOTEBoost	0,920	0,980	0,324	0,652	0,510	0,935	0,974	0,559	0,767	0,609	0,946	0,983	0,606	0,794	0,682	0,947	0,984	0,609	0,796	0,691	0,948	0,986	0,601	0,794	0,694
	RUSBoost	0,732	0,733	0,722	0,727	0,628	0,771	0,766	0,824	0,795	0,748	0,798	0,796	0,815	0,806	0,803	0,795	0,792	0,827	0,809	0,815	0,787	0,782	0,833	0,807	0,828
	MWMOTE	0,907	0,978	0,236	0,607	0,458	0,910	0,964	0,412	0,688	0,465	0,918	0,964	0,493	0,729	0,538	0,922	0,960	0,573	0,767	0,592	0,924	0,958	0,619	0,789	0,619
	EasyEnsemble	0,806	0,881	0,150	0,515	0,207	0,868	0,926	0,346	0,636	0,385	0,909	0,963	0,417	0,690	0,482	0,919	0,960	0,542	0,751	0,572	0,925	0,958	0,624	0,791	0,625
	SMOTEBagging	0,910	0,972	0,298	0,635	0,531	0,913	0,971	0,359	0,665	0,447	0,913	0,957	0,498	0,727	0,521	0,909	0,941	0,612	0,777	0,569	0,903	0,926	0,700	0,813	0,590
	UnderBagging	0,501	0,487	0,635	0,561	0,254	0,786	0,780	0,852	0,816	0,437	0,830	0,824	0,887	0,855	0,508	0,853	0,848	0,902	0,875	0,549	0,866	0,861	0,911	0,886	0,576
CART	YOK	0,907	0,994	0,036	0,091	0,500	0,860	0,950	0,309	0,629	0,464	0,899	0,973	0,209	0,591	0,387	0,949	0,987	0,607	0,797	0,699	0,949	0,987	0,608	0,100	0,702
	SMOTE	0,888	0,929	0,480	0,705	0,508	0,928	0,961	0,606	0,784	0,626	0,942	0,977	0,619	0,798	0,678	0,945	0,982	0,612	0,797	0,688	0,945	0,983	0,604	0,794	0,687
	SMOTEBoost	0,889	0,931	0,472	0,702	0,450	0,925	0,958	0,609	0,783	0,617	0,943	0,979	0,614	0,796	0,672	0,946	0,983	0,609	0,796	0,686	0,946	0,984	0,600	0,792	0,683
	RUSBoost	0,910	1,000	0,000	0,500	0,000	0,778	0,775	0,805	0,790	0,718	0,802	0,801	0,811	0,806	0,784	0,796	0,793	0,823	0,808	0,806	0,783	0,776	0,844	0,810	0,819
	MWMOTE	0,842	0,900	0,286	0,593	0,350	0,864	0,901	0,525	0,713	0,428	0,884	0,919	0,564	0,741	0,484	0,889	0,919	0,618	0,769	0,521	0,893	0,919	0,656	0,787	0,548
	EasyEnsemble	0,862	0,943	0,150	0,547	0,333	0,871	0,949	0,176	0,562	0,330	0,901	0,996	0,039	0,517	0,518	0,908	0,984	0,214	0,599	0,533	0,926	0,968	0,543	0,755	0,595
	SMOTEBagging	0,822	0,846	0,587	0,717	0,436	0,848	0,874	0,606	0,740	0,433	0,850	0,869	0,669	0,769	0,466	0,834	0,842	0,765	0,803	0,477	0,811	0,806	0,851	0,829	0,473
	UnderBagging	0,493	1,000	0,100	0,500	0,170	0,758	0,749	0,850	0,799	0,407	0,781	0,774	0,845	0,809	0,436	0,787	0,779	0,855	0,817	0,447	0,808	0,802	0,859	0,831	0,473
RF	YOK	0,915	0,993	0,132	0,091	0,514	0,879	0,973	0,305	0,639	0,422	0,911	0,987	0,207	0,597	0,329	0,946	0,986	0,585	0,785	0,679	0,948	0,987	0,595	0,100	0,692
	SMOTE	0,912	0,964	0,384	0,674	0,507	0,930	0,968	0,565	0,767	0,597	0,940	0,977	0,597	0,787	0,654	0,940	0,976	0,609	0,793	0,666	0,944	0,981	0,604	0,792	0,680
	SMOTEBoost	0,912	0,967	0,358	0,663	0,443	0,930	0,969	0,562	0,765	0,589	0,940	0,977	0,595	0,786	0,652	0,942	0,979	0,604	0,792	0,664	0,946	0,984	0,599	0,792	0,677
	RUSBoost	0,734	0,736	0,716	0,726	0,561	0,771	0,767	0,805	0,786	0,715	0,793	0,790	0,820	0,805	0,775	0,789	0,783	0,839	0,811	0,786	0,784	0,779	0,835	0,807	0,806
	MWMOTE	0,902	0,982	0,133	0,557	0,398	0,907	0,981	0,231	0,606	0,334	0,913	0,981	0,293	0,637	0,381	0,918	0,982	0,337	0,659	0,441	0,921	0,981	0,387	0,684	0,492
	EasyEnsemble	0,806	0,881	0,150	0,515	0,214	0,877	0,935	0,367	0,651	0,418	0,908	0,968	0,359	0,664	0,413	0,919	0,967	0,483	0,725	0,531	0,925	0,963	0,574	0,769	0,599
	SMOTEBagging	0,904	0,960	0,360	0,660	0,523	0,908	0,964	0,365	0,665	0,440	0,908	0,957	0,457	0,707	0,488	0,909	0,951	0,525	0,738	0,529	0,910	0,946	0,591	0,769	0,567
	UnderBagging	0,725	0,714	0,836	0,775	0,380	0,795	0,789	0,857	0,823	0,448	0,817	0,811	0,875	0,843	0,487	0,832	0,827	0,885	0,856	0,512	0,840	0,835	0,890	0,863	0,527

Tablo 4.9 yüksek düzey korelasyon ve 0,10 dengesizlik durumunu göstermektedir.

Tablo 4.9'ya genel olarak bakıldığında, sınıflama performanslarında yükselme olmuştur. Korelasyon yapısının yüksek olması sonuçlara da yansımıştır. GDO ve SEÇ değerleri tüm örneklem genişliklerinde, tüm algoritmalarda ve tüm sınıflama yöntemlerinde yüksek çıkmıştır.

Örneklem genişliği arttıkça YOK hariç algoritmalarla yapılan sınıflama yöntemlerinin F-ölçüsü değerleri artmaktadır. RUSBoost algoritması F-ölçüsü değerlerinin 2000 örneklem genişliğinde %80'nin üzerine çıkarmıştır. 2000 örneklem genişliğinde, diğer algoritmalardan sonra uygulanan sınıflama yöntemlerinin performansı RUSBoost algoritması hariç YOK'tan daha düşük sonuçlar vermiştir.

Tablo 4.10. Yüksek düzey korelasyon ve 0,15 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,882	0,973	0,314	0,643	0,493	0,902	0,981	0,436	0,709	0,553	0,913	0,993	0,449	0,721	0,597	0,916	0,998	0,449	0,724	0,611	0,915	0,999	0,441	0,720	0.607
	SMOTE	0,881	0,951	0,445	0,698	0,517	0,889	0,949	0,534	0,742	0,577	0,889	0,950	0,540	0,745	0,589	0,900	0,968	0,508	0,738	0,598	0,901	0,969	0,512	0,741	0,605
	SMOTEBoost	0,881	0,951	0,444	0,697	0,511	0,889	0,950	0,533	0,741	0,577	0,889	0,950	0,537	0,744	0,588	0,900	0,968	0,507	0,738	0,598	0,901	0,969	0,512	0,741	0,605
	RUSBoost	0,778	0,799	0,644	0,722	0,603	0,799	0,804	0,770	0,787	0,710	0,779	0,770	0,830	0,800	0,743	0,778	0,764	0,858	0,811	0,800	0,740	0,704	0,946	0,825	0,840
	MWMOTE	0,870	0,960	0,350	0,655	0,454	0,885	0,948	0,520	0,734	0,567	0,911	0,953	0,675	0,814	0,693	0,921	0,955	0,729	0,842	0,734	0,907	0,946	0,686	0,816	0,688
	EasyEnsemble	0,777	0,864	0,279	0,752	0,360	0,869	0,943	0,443	0,693	0,522	0,900	0,943	0,657	0,800	0,653	0,919	0,950	0,742	0,846	0,735	0,925	0,952	0,771	0,862	0,756
	SMOTEBagging	0,874	0,960	0,338	0,649	0,488	0,884	0,952	0,489	0,720	0,544	0,888	0,937	0,608	0,772	0,614	0,888	0,922	0,696	0,809	0,650	0,886	0,908	0,756	0,832	0,663
	UnderBagging	0,645	0,622	0,787	0,704	0,400	0,794	0,783	0,859	0,821	0,554	0,834	0,825	0,887	0,856	0,615	0,853	0,846	0,895	0,870	0,646	0,865	0,858	0,903	0,881	0,667
CART	YOK	0,859	0,952	0,277	0,615	0,558	0,907	0,986	0,440	0,713	0,589	0,915	0,995	0,455	0,725	0,606	0,917	0,998	0,451	0,724	0,613	0,916	0,999	0,443	0,721	0,609
	SMOTE	0,850	0,899	0,541	0,720	0,527	0,877	0,929	0,573	0,751	0,577	0,888	0,947	0,548	0,747	0,592	0,898	0,966	0,511	0,738	0,597	0,906	0,977	0,500	0,738	0,613
	SMOTEBoost	0,848	0,905	0,493	0,699	0,459	0,874	0,924	0,582	0,753	0,582	0,886	0,944	0,555	0,749	0,599	0,897	0,964	0,517	0,740	0,601	0,904	0,973	0,507	0,740	0,620
	RUSBoost	0,789	0,818	0,606	0,712	0,494	0,801	0,807	0,770	0,788	0,682	0,779	0,769	0,840	0,804	0,761	0,769	0,751	0,875	0,813	0,809	0,740	0,704	0,945	0,824	0,891
	MWMOTE	0,799	0,846	0,524	0,685	0,443	0,832	0,873	0,592	0,733	0,508	0,926	0,952	0,776	0,864	0,758	0,927	0,954	0,774	0,864	0,760	0,864	0,909	0,606	0,758	0,570
	EasyEnsemble	0,809	0,915	0,200	0,558	0,384	0,859	0,941	0,387	0,664	0,620	0,909	0,947	0,689	0,818	0,718	0,917	0,943	0,772	0,857	0,738	0,925	0,951	0,775	0,863	0,755
	SMOTEBagging	0,797	0,834	0,564	0,699	0,458	0,815	0,848	0,623	0,735	0,495	0,820	0,844	0,683	0,764	0,529	0,808	0,817	0,758	0,787	0,541	0,790	0,785	0,817	0,801	0,538
	UnderBagging	0,713	0,703	0,774	0,739	0,445	0,754	0,748	0,791	0,770	0,492	0,763	0,753	0,816	0,785	0,508	0,769	0,758	0,833	0,796	0,520	0,781	0,770	0,843	0,807	0,536
RF	YOK	0,879	0,978	0,257	0,617	0,478	0,906	0,983	0,456	0,720	0,577	0,912	0,991	0,457	0,724	0,600	0,915	0,996	0,454	0,725	0,611	0,915	0,998	0,445	0,721	0.608
	SMOTE	0,864	0,932	0,440	0,686	0,490	0,884	0,944	0,532	0,738	0,566	0,889	0,950	0,532	0,741	0,584	0,896	0,962	0,514	0,738	0,592	0,895	0,959	0,527	0,743	0,598
	SMOTEBoost	0,865	0,938	0,413	0,675	0,444	0,887	0,949	0,522	0,736	0,563	0,895	0,962	0,512	0,737	0,574	0,902	0,972	0,497	0,735	0,583	0,902	0,971	0,506	0,739	0,585
	RUSBoost	0,755	0,766	0,685	0,726	0,588	0,786	0,786	0,789	0,787	0,711	0,768	0,751	0,866	0,808	0,757	0,773	0,757	0,861	0,809	0,784	0,760	0,736	0,895	0,816	0,788
	MWMOTE	0,864	0,972	0,239	0,605	0,392	0,878	0,972	0,335	0,654	0,435	0,911	0,960	0,632	0,796	0,673	0,915	0,959	0,669	0,814	0,700	0,895	0,973	0,449	0,711	0,559
	EasyEnsemble	0,818	0,896	0,367	0,632	0,443	0,897	0,945	0,620	0,783	0,637	0,919	0,958	0,701	0,829	0,717	0,923	0,954	0,752	0,853	0,746	0,926	0,954	0,767	0,861	0,757
	SMOTEBagging	0,866	0,948	0,354	0,651	0,477	0,874	0,948	0,442	0,695	0,498	0,880	0,941	0,529	0,735	0,562	0,884	0,935	0,591	0,763	0,602	0,888	0,933	0,634	0,783	0,628
	UnderBagging	0,728	0,714	0,818	0,766	0,465	0,783	0,773	0,844	0,808	0,538	0,807	0,796	0,866	0,831	0,572	0,813	0,803	0,874	0,838	0,584	0,821	0,810	0,881	0,845	0,596

Tablo 4.10 yüksek düzey korelasyon ve 0,15 dengesizlik durumunu göstermektedir.

Tablo 4.10'da GDO değerlerine bakıldığında, Tablo 4.9'e göre düşmektedir. EasyEnsemble algoritması düşük örneklem genişliğinde (n=100) sınıflama yöntemlerinin performansı üzerine etkisi diğer algoritmalara göre düşük iken örneklem genişliği arttıkça performanslara olan etki de artmaktadır.

Örneklem genişliği arttıkça tüm algoritmaların F-ölçüsü değerleri üzerindeki etkileri artmaktadır. RUSBoost algoritması DVM ve CART yöntemlerinin sınıflama performansı %80'nin üzerine çıkarmaktadır.

Tablo 4.11. Yüksek düzey korelasyon ve 0,25 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,813	0,923	0,468	0,241	0,552	0,823	0,921	0,522	0,722	0,585	0,829	0,922	0,549	0,249	0,612	0,836	0,925	0,569	0,747	0,632	0,839	0,926	0,576	0,250	0,640
	SMOTE	0,817	0,874	0,637	0,756	0,619	0,814	0,828	0,770	0,799	0,669	0,809	0,807	0,815	0,811	0,678	0,807	0,793	0,849	0,821	0,687	0,804	0,772	0,897	0,835	0,695
	SMOTEBoost	0,817	0,871	0,646	0,759	0,624	0,813	0,825	0,776	0,800	0,672	0,809	0,806	0,815	0,811	0,678	0,807	0,793	0,848	0,821	0,687	0,804	0,773	0,896	0,835	0,695
	RUSBoost	0,734	0,715	0,794	0,754	0,696	0,795	0,773	0,864	0,818	0,708	0,798	0,757	0,923	0,840	0,717	0,796	0,744	0,954	0,849	0,722	0,797	0,743	0,959	0,851	0,714
	MWMOTE	0,817	0,926	0,485	0,706	0,556	0,851	0,918	0,646	0,782	0,680	0,840	0,902	0,653	0,777	0,670	0,867	0,908	0,745	0,826	0,736	0,873	0,906	0,775	0,840	0,753
	EasyEnsemble	0,735	0,835	0,434	0,635	0,490	0,839	0,899	0,656	0,777	0,667	0,850	0,910	0,667	0,789	0,689	0,854	0,914	0,673	0,794	0,698	0,860	0,925	0,667	0,796	0,705
	SMOTEBagging	0,815	0,935	0,438	0,686	0,528	0,841	0,920	0,601	0,760	0,648	0,852	0,905	0,690	0,798	0,697	0,862	0,899	0,749	0,824	0,729	0,865	0,890	0,788	0,839	0,744
	UnderBagging	0,703	0,666	0,820	0,743	0,576	0,796	0,776	0,855	0,816	0,676	0,827	0,812	0,874	0,843	0,716	0,848	0,837	0,882	0,859	0,743	0,859	0,850	0,888	0,869	0,759
CART	YOK	0,820	0,914	0,523	0,241	0,608	0,826	0,910	0,570	0,740	0,618	0,832	0,916	0,578	0,249	0,630	0,837	0,922	0,581	0,751	0,638	0,837	0,922	0,583	0,250	0,640
	SMOTE	0,805	0,832	0,722	0,777	0,637	0,808	0,797	0,842	0,820	0,682	0,802	0,777	0,877	0,827	0,685	0,803	0,774	0,890	0,832	0,692	0,798	0,746	0,952	0,849	0,701
	SMOTEBoost	0,805	0,837	0,706	0,771	0,624	0,809	0,804	0,826	0,815	0,676	0,801	0,777	0,874	0,825	0,684	0,804	0,775	0,889	0,832	0,692	0,798	0,748	0,948	0,848	0,700
	RUSBoost	0,728	0,728	0,728	0,728	0,637	0,801	0,779	0,871	0,825	0,695	0,795	0,761	0,897	0,829	0,695	0,798	0,759	0,916	0,838	0,701	0,797	0,750	0,941	0,845	0,698
	MWMOTE	0,760	0,828	0,554	0,691	0,538	0,786	0,841	0,618	0,729	0,587	0,871	0,954	0,622	0,788	0,708	0,808	0,883	0,581	0,732	0,600	0,811	0,890	0,574	0,732	0,601
	EasyEnsemble	0,730	0,841	0,393	0,617	0,500	0,842	0,891	0,693	0,792	0,685	0,848	0,904	0,679	0,792	0,690	0,856	0,920	0,667	0,793	0,699	0,863	0,932	0,654	0,793	0,705
	SMOTEBagging	0,754	0,806	0,591	0,699	0,535	0,776	0,819	0,646	0,732	0,586	0,786	0,818	0,689	0,753	0,615	0,787	0,802	0,743	0,773	0,635	0,775	0,772	0,784	0,778	0,635
	UnderBagging	0,709	0,705	0,719	0,712	0,549	0,735	0,727	0,762	0,744	0,589	0,749	0,737	0,788	0,762	0,610	0,762	0,746	0,809	0,778	0,629	0,768	0,751	0,818	0,785	0,638
RF	YOK	0,811	0,928	0,443	0,241	0,520	0,820	0,917	0,523	0,720	0,576	0,821	0,918	0,529	0,249	0,587	0,824	0,921	0,533	0,727	0,598	0,827	0,924	0,537	0,250	0,606
	SMOTE	0,810	0,869	0,625	0,747	0,601	0,811	0,834	0,742	0,788	0,658	0,812	0,833	0,747	0,790	0,662	0,812	0,826	0,771	0,798	0,671	0,810	0,807	0,820	0,813	0,683
	SMOTEBoost	0,809	0,870	0,616	0,743	0,591	0,811	0,831	0,751	0,791	0,661	0,813	0,836	0,744	0,790	0,661	0,813	0,828	0,768	0,798	0,670	0,810	0,807	0,821	0,814	0,683
	RUSBoost	0,734	0,725	0,765	0,745	0,670	0,794	0,773	0,859	0,816	0,705	0,798	0,769	0,888	0,828	0,715	0,800	0,763	0,909	0,836	0,721	0,800	0,759	0,926	0,842	0,718
	MWMOTE	0,801	0,935	0,393	0,664	0,481	0,831	0,943	0,489	0,716	0,583	0,874	0,968	0,590	0,779	0,699	0,847	0,943	0,560	0,752	0,646	0,853	0,941	0,586	0,764	0,665
	EasyEnsemble	0,785	0,852	0,584	0,718	0,581	0,850	0,915	0,654	0,784	0,684	0,860	0,932	0,645	0,788	0,697	0,867	0,943	0,637	0,790	0,705	0,871	0,953	0,628	0,790	0,709
	SMOTEBagging	0,804	0,917	0,448	0,683	0,516	0,825	0,914	0,554	0,734	0,604	0,836	0,909	0,612	0,761	0,647	0,846	0,908	0,658	0,783	0,680	0,852	0,904	0,694	0,799	0,700
	UnderBagging	0,711	0,684	0,796	0,740	0,575	0,767	0,745	0,832	0,789	0,640	0,786	0,767	0,843	0,805	0,663	0,799	0,780	0,855	0,817	0,679	0,806	0,787	0,863	0,825	0,690

Tablo 4.11 yüksek düzey korelasyon ve 0,25 dengesizlik durumunu göstermektedir.

Tablo 4.11'de DUY ve F-ölçüsü değerlerinde Tablo 4.10'e göre yükselme görülmekte iken GDO değerlerinde düşüş olmaktadır. MWMOTE, EasyEnsemble, SMOTEBagging ve UnderBagging algoritmalarının etkisi YOK'a göre daha düşük F-ölçüsü değerlerine sahiptir.

Örneklem genişliği arttıkça DUY ve F-ölçüsü değerleri artmaktadır. Düşük ve orta düzey korelasyon yapılarındaki gibi yüksek bir artış bulunmamaktadır.

2000 örneklem genişliğinde DVM yönteminde UnderBagging ve MWMOTE algoritmalarının, CART yönteminde SMOTE, SMOTEBoost ve EasyEnsemble algoritmalarının ve RF yönteminde ise RUSBoost, SMOTEBagging ve EasyEnsemble algoritmalarının katkısı diğer algoritmalara göre fazladır. Bu algoritmalar performansları yaklaşık %70 civarına çıkarmaktadır.

Tablo 4.12. Yüksek düzey korelasyon ve 0,30 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,796	0,906	0,532	0,295	0,603	0,810	0,903	0,591	0,747	0,644	0,816	0,906	0,602	0,299	0,657	0,818	0,907	0,610	0,299	0,663	0,819	0,908	0,610	0,759	0.666
	SMOTE	0,804	0,846	0,704	0,775	0,673	0,819	0,821	0,812	0,817	0,725	0,813	0,791	0,863	0,827	0,734	0,816	0,781	0,900	0,840	0,746	0,815	0,775	0,908	0,300	0,747
	SMOTEBoost	0,805	0,843	0,713	0,778	0,677	0,818	0,818	0,818	0,818	0,728	0,813	0,791	0,862	0,827	0,733	0,816	0,781	0,899	0,840	0,746	0,815	0,776	0,908	0,300	0,747
	RUSBoost	0,773	0,788	0,739	0,764	0,691	0,810	0,796	0,845	0,820	0,740	0,813	0,779	0,892	0,836	0,745	0,816	0,774	0,916	0,845	0,752	0,815	0,773	0,915	0,300	0,749
	MWMOTE	0,787	0,865	0,600	0,733	0,620	0,823	0,876	0,699	0,788	0,702	0,801	0,827	0,739	0,783	0,690	0,854	0,888	0,775	0,831	0,760	0,861	0,887	0,799	0,843	0,775
	EasyEnsemble	0,772	0,869	0,543	0,706	0,588	0,815	0,870	0,685	0,778	0,688	0,819	0,874	0,691	0,782	0,695	0,827	0,884	0,692	0,788	0,705	0,830	0,889	0,692	0,791	0,710
	SMOTEBagging	0,789	0,918	0,482	0,700	0,564	0,828	0,909	0,639	0,774	0,686	0,842	0,898	0,710	0,804	0,728	0,851	0,890	0,762	0,826	0,754	0,858	0,886	0,793	0,840	0,770
	UnderBagging	0,714	0,666	0,827	0,746	0,633	0,796	0,769	0,862	0,815	0,717	0,829	0,811	0,873	0,842	0,754	0,845	0,831	0,879	0,855	0,773	0,856	0,845	0,881	0,863	0,786
CART	YOK	0,806	0,905	0,569	0,295	0,637	0,814	0,891	0,632	0,761	0,663	0,818	0,904	0,617	0,299	0,664	0,818	0,912	0,599	0,299	0,659	0,819	0,890	0,653	0,771	0,680
	SMOTE	0,802	0,820	0,758	0,789	0,685	0,816	0,796	0,862	0,829	0,733	0,813	0,776	0,898	0,837	0,741	0,816	0,776	0,910	0,843	0,748	0,815	0,773	0,915	0,300	0,748
	SMOTEBoost	0,800	0,826	0,739	0,782	0,673	0,814	0,806	0,835	0,820	0,720	0,811	0,778	0,891	0,834	0,738	0,816	0,777	0,907	0,842	0,747	0,815	0,774	0,911	0,300	0,747
	RUSBoost	0,775	0,816	0,677	0,746	0,637	0,815	0,806	0,834	0,820	0,721	0,810	0,792	0,850	0,821	0,722	0,815	0,788	0,879	0,834	0,735	0,815	0,782	0,892	0,300	0,738
	MWMOTE	0,729	0,765	0,643	0,704	0,586	0,761	0,810	0,644	0,727	0,615	0,823	0,875	0,702	0,788	0,702	0,774	0,823	0,660	0,741	0,635	0,775	0,819	0,674	0,746	0,642
	EasyEnsemble	0,772	0,861	0,559	0,710	0,629	0,818	0,867	0,702	0,784	0,696	0,819	0,872	0,695	0,784	0,697	0,826	0,883	0,691	0,787	0,703	0,829	0,891	0,684	0,788	0,706
	SMOTEBagging	0,738	0,796	0,601	0,699	0,572	0,765	0,810	0,660	0,735	0,624	0,778	0,816	0,691	0,753	0,650	0,782	0,804	0,732	0,768	0,668	0,775	0,778	0,766	0,772	0,671
	UnderBagging	0,700	0,692	0,717	0,705	0,589	0,737	0,727	0,758	0,743	0,632	0,751	0,738	0,782	0,760	0,653	0,760	0,743	0,798	0,771	0,666	0,766	0,750	0,805	0,777	0,674
RF	YOK	0,794	0,910	0,516	0,295	0,576	0,806	0,879	0,633	0,756	0,654	0,813	0,882	0,649	0,299	0,671	0,813	0,879	0,658	0,299	0,676	0,815	0,882	0,658	0,770	0.680
	SMOTE	0,796	0,837	0,700	0,768	0,663	0,813	0,826	0,783	0,805	0,712	0,811	0,814	0,804	0,809	0,718	0,816	0,801	0,850	0,826	0,734	0,815	0,794	0,865	0,300	0,737
	SMOTEBoost	0,797	0,838	0,699	0,768	0,661	0,814	0,827	0,784	0,806	0,712	0,812	0,815	0,802	0,809	0,717	0,816	0,801	0,852	0,826	0,735	0,815	0,793	0,869	0,300	0,738
	RUSBoost	0,754	0,760	0,741	0,750	0,682	0,800	0,786	0,831	0,809	0,733	0,808	0,788	0,855	0,821	0,738	0,814	0,785	0,882	0,834	0,746	0,815	0,782	0,892	0,300	0,747
	MWMOTE	0,780	0,881	0,541	0,711	0,587	0,811	0,906	0,586	0,746	0,646	0,827	0,916	0,619	0,768	0,681	0,830	0,919	0,623	0,771	0,687	0,837	0,920	0,642	0,7818	0,702
	EasyEnsemble	0,803	0,866	0,654	0,760	0,655	0,819	0,873	0,691	0,782	0,693	0,821	0,882	0,679	0,781	0,694	0,828	0,896	0,670	0,783	0,700	0,830	0,904	0,659	0,781	0,699
	SMOTEBagging	0,778	0,900	0,488	0,694	0,555	0,807	0,900	0,586	0,743	0,640	0,823	0,900	0,641	0,771	0,682	0,831	0,897	0,678	0,787	0,706	0,838	0,895	0,704	0,800	0,723
	UnderBagging	0,711	0,672	0,804	0,738	0,625	0,762	0,732	0,831	0,782	0,676	0,783	0,757	0,843	0,800	0,699	0,795	0,771	0,850	0,810	0,713	0,802	0,779	0,857	0,818	0,722

Tablo 4.12 yüksek düzey korelasyon ve 0,30 dengesizlik durumunu göstermektedir.

Tablo 4.12 GDO değerlerine bakıldığında, Tablo 4.11'ye göre düşüş gözlemlenmektedir. Korelasyon yapısının yüksek olması ve dengesizliğin azalması performans değerlerinin genelinde bir artışa neden olmaktadır.

Örneklem genişliği 100 olduğu durumda; F-ölçüsüne bakıldığında, DVM yönteminde EasyEnsemble, SMOTEBagging ve UnderBagging algoritmaları, CART yönteminde MWMOTE, SMOTEBagging ve UnderBagging algoritmaları, RF yönteminde ise sadece SMOTEBagging algoritmasının katkısı YOK'a göre daha düşük bulunmaktadır. Örneklem genişliği arttıkça tüm algoritmaların F-ölçüsü üzerindeki etkileri YOK'a göre %5'in üzerindedir.

2000 örneklem genişliğinde, F-ölçüsü değerleri her bir sınıflama yönteminde %70'in üzerinde bulunmuştur.

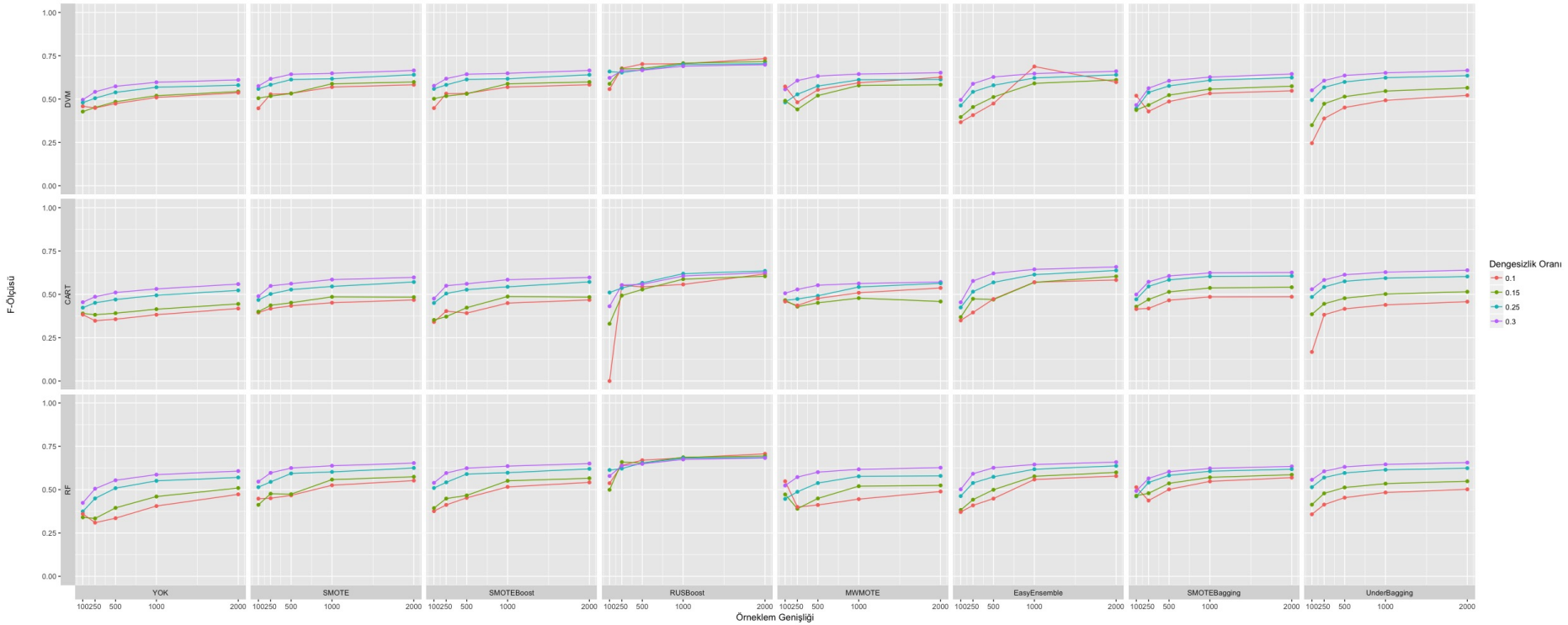
4.1.4 Gerçek Korelasyona Ait Sonuçlar

Grafik 4.4'de gerçek veri setinden yararlanılarak üretilmiş verilerden elde edilen sınıflama sonuçları gösterilmektedir. Gerçek veri setinde bağımlı değişken ile bağımsız değişkenler arasında orta düzey bir ilişki vardır.

Grafiğe genel olarak bakıldığında, YOK ile algoritmaların kullanıldığı sınıflama yöntemlerindeki performansları örneklem genişliği arttıkça artmaktadır. Dengesizlik azaldıkça performanslar artmaktadır.

SMOTE, SMOTEBoost, MWMOTE, EasyEnsemble, SMOTEBagging ve UnderBagging algoritmlarında sonuçlar benzerken, RUSBoost algoritmasında dengesizlik oranlarının hepsinde performanslar daha yüksek bulunmuştur. CART sınıflama yöntemi (2,4) ve (2,8) hücrelerinde görüldüğü gibi 0,10 dengesizlik durumunda örneklem genişliği 100 olduğunda diğer sınıflama yöntemlerine göre performansı düşük iken örneklem genişliği arttıkça performansı artmıştır.

Dört farklı dengesizlik durumuna ait sonuçlar Tablo 4.13, Tablo 4.14, Tablo 4.15 ve Tablo 4.16'de verilmiştir.



Şekil 4.4. Gerçek korelasyon sonuçları

Tablo 4.13. Gerçek korelasyon ve 0,10 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,913	0,987	0,170	0,578	0,458	0,917	0,980	0,311	0,646	0,449	0,921	0,980	0,373	0,676	0,474	0,925	0,981	0,410	0,696	0,509	0,928	0,982	0,430	0,706	0,537
	SMOTE	0,909	0,968	0,320	0,644	0,447	0,919	0,961	0,512	0,736	0,528	0,917	0,962	0,499	0,731	0,533	0,917	0,957	0,557	0,757	0,569	0,922	0,963	0,550	0,757	0,583
	SMOTEBoost	0,909	0,967	0,331	0,649	0,448	0,919	0,961	0,518	0,739	0,532	0,917	0,962	0,500	0,731	0,533	0,917	0,957	0,556	0,757	0,569	0,922	0,963	0,550	0,757	0,583
	RUSBoost	0,767	0,766	0,783	0,774	0,558	0,785	0,783	0,807	0,795	0,677	0,802	0,796	0,858	0,827	0,702	0,814	0,808	0,873	0,841	0,704	0,826	0,819	0,889	0,854	0,733
	MWMOTE	0,914	0,982	0,219	0,601	0,572	0,919	0,973	0,395	0,684	0,481	0,923	0,968	0,503	0,735	0,553	0,925	0,965	0,560	0,762	0,594	0,928	0,963	0,612	0,787	0,626
	EasyEnsemble	0,778	0,821	0,370	0,595	0,367	0,899	0,965	0,286	0,625	0,408	0,910	0,966	0,397	0,681	0,474	0,907	0,946	0,686	0,816	0,688	0,924	0,963	0,570	0,766	0,598
	SMOTEBagging	0,899	0,962	0,260	0,611	0,519	0,904	0,959	0,373	0,666	0,429	0,898	0,941	0,500	0,720	0,486	0,895	0,926	0,610	0,768	0,533	0,887	0,910	0,686	0,798	0,548
	UnderBagging	0,506	0,494	0,632	0,563	0,246	0,756	0,751	0,802	0,776	0,388	0,797	0,791	0,844	0,818	0,451	0,824	0,820	0,862	0,841	0,493	0,840	0,837	0,873	0,855	0,522
CART	YOK	0,906	0,994	0,028	0,511	0,383	0,897	0,972	0,184	0,578	0,348	0,901	0,976	0,204	0,590	0,357	0,905	0,975	0,256	0,616	0,382	0,910	0,976	0,313	0,644	0,418
	SMOTE	0,880	0,940	0,299	0,620	0,395	0,889	0,938	0,420	0,679	0,418	0,889	0,937	0,440	0,689	0,434	0,887	0,933	0,476	0,704	0,452	0,902	0,953	0,438	0,695	0,468
	SMOTEBoost	0,879	0,931	0,359	0,645	0,341	0,895	0,944	0,422	0,683	0,403	0,888	0,939	0,411	0,675	0,392	0,887	0,932	0,474	0,703	0,450	0,902	0,953	0,439	0,696	0,468
	RUSBoost	0,908	1,000	0,000	0,500	0,000	0,764	0,768	0,730	0,749	0,552	0,744	0,745	0,733	0,739	0,543	0,777	0,780	0,745	0,763	0,557	0,791	0,795	0,759	0,777	0,618
	MWMOTE	0,873	0,931	0,281	0,606	0,459	0,885	0,931	0,442	0,686	0,436	0,891	0,931	0,520	0,726	0,476	0,894	0,929	0,575	0,752	0,509	0,896	0,925	0,626	0,776	0,537
	EasyEnsemble	0,830	0,883	0,323	0,603	0,350	0,892	0,977	0,115	0,546	0,395	0,901	0,986	0,127	0,556	0,473	0,864	0,909	0,606	0,758	0,570	0,923	0,964	0,547	0,756	0,583
	SMOTEBagging	0,822	0,854	0,505	0,679	0,415	0,849	0,877	0,573	0,725	0,419	0,855	0,878	0,644	0,761	0,466	0,846	0,859	0,731	0,795	0,485	0,831	0,835	0,798	0,817	0,486
	UnderBagging	0,492	1,000	0,100	0,500	0,168	0,757	0,758	0,754	0,756	0,382	0,779	0,778	0,787	0,782	0,417	0,793	0,791	0,814	0,802	0,439	0,800	0,795	0,845	0,820	0,458
RF	YOK	0,909	0,993	0,070	0,531	0,360	0,911	0,990	0,156	0,573	0,309	0,914	0,988	0,227	0,608	0,335	0,919	0,987	0,293	0,640	0,405	0,923	0,986	0,354	0,670	0,473
	SMOTE	0,906	0,968	0,293	0,631	0,448	0,911	0,963	0,404	0,684	0,450	0,910	0,963	0,422	0,692	0,467	0,912	0,957	0,498	0,728	0,525	0,916	0,960	0,523	0,741	0,552
	SMOTEBoost	0,908	0,971	0,284	0,627	0,376	0,905	0,960	0,378	0,669	0,412	0,911	0,966	0,401	0,684	0,453	0,913	0,961	0,481	0,721	0,516	0,918	0,963	0,504	0,734	0,542
	RUSBoost	0,740	0,745	0,685	0,715	0,537	0,742	0,736	0,793	0,765	0,640	0,788	0,783	0,829	0,806	0,670	0,805	0,800	0,854	0,827	0,683	0,820	0,815	0,870	0,842	0,707
	MWMOTE	0,908	0,982	0,152	0,567	0,548	0,911	0,979	0,257	0,618	0,400	0,915	0,978	0,327	0,652	0,411	0,918	0,981	0,350	0,665	0,445	0,922	0,981	0,384	0,683	0,489
	EasyEnsemble	0,792	0,831	0,422	0,627	0,371	0,901	0,965	0,319	0,642	0,409	0,912	0,971	0,372	0,672	0,448	0,895	0,973	0,449	0,711	0,559	0,923	0,966	0,533	0,749	0,578
	SMOTEBagging	0,895	0,952	0,335	0,643	0,514	0,904	0,957	0,392	0,675	0,437	0,905	0,949	0,497	0,723	0,501	0,907	0,944	0,571	0,758	0,548	0,907	0,939	0,619	0,779	0,569
	UnderBagging	0,715	0,706	0,806	0,756	0,358	0,775	0,770	0,817	0,794	0,414	0,797	0,791	0,849	0,820	0,454	0,816	0,810	0,867	0,839	0,483	0,825	0,819	0,882	0,850	0,502

Tablo 4.13 gerçek korelasyon ve 0,10 dengesizlik durumunu göstermektedir.

Gerçek korelasyon yapısı, gerçek veri setinden yararlanılarak elde edilen bir korelasyon yapısıdır. Gerçek veri seti korelasyon yapısı orta düzey korelasyon yapısına benzemektedir. 4.1'de korelasyon yapısı hakkında bilgi verilmektedir.

Tablo 4.13'e göre örneklem genişliği 100 olduğunda GDO ve SEÇ değerleri yüksek iken DUY ve F-ölçüsü değerleri düşük çıkmaktadır. F-ölçüsü değerlerine bakıldığında, DVM ve RF yöntemlerinde RUSBoost, MWMOTE ve SMOTEbagging algoritmaları, CART yönteminde ise SMOTEBagging algoritması YOK'a göre daha etkili görülmektedir.

Dengesizliğin yüksek olmasına rağmen korelasyonun orta düzey olması örneklem genişliği arttıkça F-ölçüsü değerlerini yükseltmektedir. Sınıflama yöntemlerinde en yüksek örneklem genişliğinde RUSBoost algoritması diğerlerine göre daha etkili olmuştur. DVM ve RF'de performansları %70'in, CART yönteminde ise %60'ın üzerine çıkarmaktadır.

Tablo 4.14. Gerçek korelasyon ve 0,15 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,868	0,970	0,227	0,598	0,428	0,876	0,966	0,341	0,654	0,452	0,882	0,969	0,384	0,676	0,485	0,888	0,971	0,415	0,693	0,519	0,892	0,972	0,435	0,704	0,544
	SMOTE	0,871	0,938	0,456	0,697	0,505	0,874	0,936	0,502	0,719	0,517	0,872	0,936	0,499	0,717	0,532	0,882	0,937	0,569	0,753	0,588	0,883	0,935	0,585	0,760	0,598
	SMOTEBoost	0,871	0,938	0,456	0,697	0,502	0,873	0,936	0,501	0,718	0,517	0,871	0,935	0,495	0,715	0,530	0,882	0,937	0,569	0,753	0,589	0,883	0,935	0,585	0,760	0,599
	RUSBoost	0,775	0,786	0,707	0,747	0,588	0,766	0,754	0,833	0,794	0,674	0,752	0,741	0,818	0,779	0,676	0,798	0,792	0,832	0,812	0,707	0,790	0,778	0,859	0,819	0,717
	MWMOTE	0,874	0,969	0,279	0,624	0,490	0,881	0,977	0,318	0,647	0,441	0,884	0,961	0,436	0,699	0,520	0,883	0,942	0,543	0,743	0,578	0,890	0,956	0,516	0,736	0,583
	EasyEnsemble	0,830	0,932	0,212	0,572	0,396	0,864	0,956	0,335	0,645	0,454	0,870	0,947	0,428	0,688	0,512	0,886	0,948	0,538	0,743	0,591	0,891	0,947	0,574	0,760	0,611
	SMOTEBagging	0,849	0,938	0,296	0,617	0,437	0,852	0,921	0,448	0,685	0,466	0,851	0,903	0,554	0,728	0,523	0,850	0,887	0,636	0,762	0,558	0,848	0,876	0,686	0,781	0,574
	UnderBagging	0,590	0,572	0,707	0,639	0,350	0,737	0,727	0,796	0,762	0,473	0,770	0,762	0,816	0,789	0,514	0,794	0,787	0,831	0,809	0,546	0,807	0,802	0,837	0,819	0,565
CART	YOK	0,852	0,971	0,105	0,538	0,390	0,854	0,964	0,211	0,587	0,382	0,857	0,962	0,258	0,610	0,391	0,864	0,960	0,312	0,636	0,414	0,871	0,962	0,348	0,655	0,445
	SMOTE	0,811	0,874	0,417	0,645	0,400	0,836	0,902	0,445	0,673	0,437	0,834	0,896	0,472	0,684	0,452	0,850	0,914	0,485	0,699	0,486	0,851	0,917	0,474	0,695	0,484
	SMOTEBoost	0,810	0,878	0,389	0,633	0,352	0,843	0,915	0,409	0,662	0,372	0,843	0,914	0,430	0,672	0,423	0,848	0,911	0,490	0,700	0,488	0,851	0,917	0,474	0,695	0,484
	RUSBoost	0,689	0,717	0,513	0,615	0,330	0,711	0,728	0,603	0,666	0,493	0,680	0,681	0,672	0,676	0,528	0,746	0,752	0,712	0,732	0,588	0,729	0,725	0,753	0,739	0,605
	MWMOTE	0,836	0,911	0,367	0,639	0,466	0,854	0,946	0,309	0,628	0,429	0,853	0,930	0,413	0,671	0,451	0,853	0,920	0,464	0,692	0,478	0,858	0,937	0,409	0,673	0,459
	EasyEnsemble	0,842	0,965	0,096	0,530	0,369	0,854	0,987	0,090	0,538	0,474	0,865	0,965	0,291	0,628	0,470	0,884	0,949	0,518	0,733	0,569	0,890	0,947	0,565	0,756	0,604
	SMOTEBagging	0,784	0,829	0,497	0,663	0,429	0,813	0,854	0,573	0,713	0,470	0,820	0,851	0,643	0,747	0,514	0,818	0,837	0,709	0,773	0,537	0,810	0,822	0,745	0,784	0,541
	UnderBagging	0,679	0,681	0,667	0,674	0,385	0,725	0,722	0,741	0,732	0,445	0,749	0,745	0,768	0,757	0,478	0,762	0,754	0,804	0,779	0,502	0,768	0,759	0,821	0,790	0,515
RF	YOK	0,865	0,982	0,131	0,556	0,340	0,869	0,979	0,219	0,599	0,334	0,875	0,977	0,289	0,633	0,395	0,882	0,976	0,347	0,661	0,460	0,888	0,975	0,394	0,685	0,509
	SMOTE	0,861	0,945	0,335	0,640	0,412	0,870	0,944	0,428	0,686	0,476	0,864	0,939	0,424	0,682	0,474	0,875	0,935	0,533	0,734	0,558	0,876	0,931	0,562	0,747	0,574
	SMOTEBoost	0,862	0,946	0,342	0,644	0,393	0,876	0,957	0,389	0,673	0,448	0,866	0,943	0,413	0,678	0,467	0,879	0,941	0,520	0,731	0,551	0,879	0,937	0,546	0,741	0,566
	RUSBoost	0,713	0,720	0,666	0,693	0,499	0,748	0,743	0,776	0,760	0,659	0,744	0,735	0,800	0,768	0,654	0,791	0,788	0,812	0,800	0,687	0,785	0,775	0,840	0,807	0,694
	MWMOTE	0,870	0,971	0,239	0,605	0,472	0,872	0,973	0,276	0,625	0,390	0,880	0,973	0,344	0,659	0,449	0,887	0,969	0,418	0,693	0,520	0,889	0,971	0,417	0,694	0,524
	EasyEnsemble	0,823	0,918	0,236	0,577	0,383	0,868	0,958	0,350	0,654	0,442	0,878	0,957	0,425	0,691	0,499	0,886	0,950	0,523	0,737	0,577	0,890	0,950	0,553	0,751	0,599
	SMOTEBagging	0,852	0,932	0,351	0,641	0,465	0,860	0,931	0,448	0,689	0,479	0,866	0,924	0,529	0,727	0,536	0,869	0,918	0,587	0,753	0,571	0,870	0,914	0,616	0,765	0,586
	UnderBagging	0,691	0,680	0,758	0,719	0,413	0,743	0,734	0,795	0,764	0,478	0,767	0,757	0,821	0,789	0,512	0,782	0,772	0,839	0,806	0,535	0,790	0,780	0,848	0,814	0,548

Tablo 4.14 gerçek korelasyon ve 0,15 dengesizlik durumunu göstermektedir.

Tablo 4.14'de GDO değerlerine bakıldığında Tablo 4.13'e göre bir düşüş gözükmektedir.

DUY ve F-ölçüsü değerleri örneklem genişliği 100 olduğu durumda çok düşük çıkmaktadır. Özellikle UnderBagging algoritması diğerlerine göre daha az etkili olmuştur. Örneklem genişliği arttıkça değerlerde de bir artış olmaktadır, 2000 örneklem genişliğinde RUSBoost algoritması performansları YOK'a göre yaklaşık %10 arttırmaktadır.

Örneklem genişliği arttıkça F-ölçüsü değerleri üzerinde en çok RUSBoost etkili görünmesinin yanı sıra EasyEnsemble ve SMOTEBagging algoritmaları da diğer algoritmalara göre performansları arttırmaktadır.

Tablo 4.15. Gerçek korelasyon ve 0,25 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,790	0,927	0,357	0,642	0,480	0,804	0,930	0,418	0,674	0,505	0,816	0,939	0,443	0,691	0,539	0,824	0,941	0,470	0,706	0,568	0,828	0,945	0,478	0,711	0,580
	SMOTE	0,793	0,865	0,564	0,715	0,559	0,788	0,847	0,606	0,727	0,583	0,797	0,847	0,647	0,747	0,613	0,808	0,869	0,624	0,746	0,617	0,815	0,866	0,662	0,764	0,640
	SMOTEBoost	0,791	0,862	0,569	0,716	0,559	0,787	0,847	0,606	0,726	0,582	0,796	0,845	0,648	0,747	0,613	0,808	0,869	0,624	0,746	0,617	0,815	0,866	0,662	0,764	0,640
	RUSBoost	0,714	0,697	0,769	0,733	0,659	0,729	0,713	0,777	0,745	0,653	0,748	0,737	0,781	0,759	0,669	0,751	0,730	0,813	0,772	0,700	0,767	0,752	0,813	0,783	0,704
	MWMOTE	0,796	0,933	0,365	0,649	0,481	0,808	0,926	0,447	0,686	0,527	0,817	0,920	0,505	0,713	0,575	0,816	0,894	0,581	0,738	0,611	0,828	0,922	0,547	0,735	0,613
	EasyEnsemble	0,756	0,892	0,343	0,618	0,463	0,798	0,899	0,495	0,697	0,542	0,811	0,905	0,528	0,717	0,579	0,821	0,898	0,590	0,744	0,622	0,830	0,904	0,606	0,755	0,640
	SMOTEBagging	0,762	0,882	0,384	0,633	0,445	0,781	0,864	0,526	0,695	0,539	0,786	0,853	0,586	0,719	0,576	0,796	0,848	0,639	0,743	0,609	0,800	0,844	0,666	0,755	0,624
	UnderBagging	0,626	0,587	0,748	0,667	0,495	0,706	0,684	0,775	0,730	0,568	0,739	0,724	0,784	0,754	0,599	0,760	0,747	0,797	0,772	0,623	0,770	0,760	0,801	0,780	0,635
CART	YOK	0,756	0,916	0,255	0,585	0,423	0,774	0,914	0,349	0,631	0,451	0,785	0,918	0,384	0,651	0,470	0,793	0,919	0,413	0,666	0,495	0,803	0,924	0,439	0,681	0,523
	SMOTE	0,720	0,785	0,514	0,649	0,468	0,734	0,798	0,539	0,669	0,502	0,754	0,819	0,559	0,689	0,528	0,773	0,846	0,554	0,700	0,546	0,782	0,846	0,587	0,717	0,571
	SMOTEBoost	0,721	0,789	0,508	0,649	0,450	0,735	0,793	0,556	0,675	0,505	0,757	0,824	0,557	0,690	0,527	0,773	0,848	0,549	0,698	0,544	0,782	0,846	0,588	0,717	0,572
	RUSBoost	0,581	0,553	0,666	0,610	0,511	0,667	0,681	0,623	0,652	0,536	0,704	0,723	0,644	0,684	0,566	0,691	0,683	0,714	0,699	0,619	0,699	0,688	0,732	0,710	0,635
	MWMOTE	0,757	0,881	0,365	0,623	0,465	0,774	0,897	0,401	0,649	0,473	0,787	0,906	0,428	0,667	0,492	0,794	0,891	0,503	0,697	0,542	0,799	0,889	0,527	0,708	0,564
	EasyEnsemble	0,728	0,863	0,321	0,592	0,424	0,787	0,897	0,458	0,677	0,515	0,807	0,904	0,516	0,710	0,569	0,820	0,901	0,577	0,739	0,614	0,830	0,906	0,601	0,753	0,638
	SMOTEBagging	0,725	0,792	0,511	0,652	0,471	0,759	0,815	0,590	0,702	0,546	0,777	0,826	0,630	0,728	0,584	0,783	0,822	0,665	0,743	0,604	0,785	0,826	0,662	0,744	0,606
	UnderBagging	0,654	0,648	0,674	0,661	0,485	0,693	0,680	0,736	0,708	0,543	0,719	0,704	0,764	0,734	0,575	0,729	0,709	0,791	0,750	0,593	0,737	0,716	0,800	0,758	0,603
RF	YOK	0,781	0,943	0,270	0,607	0,374	0,796	0,941	0,353	0,647	0,449	0,809	0,942	0,407	0,674	0,508	0,819	0,941	0,451	0,696	0,551	0,824	0,941	0,471	0,706	0,570
	SMOTE	0,774	0,859	0,508	0,684	0,514	0,772	0,842	0,558	0,700	0,545	0,790	0,848	0,615	0,732	0,593	0,800	0,864	0,609	0,736	0,602	0,807	0,860	0,647	0,753	0,625
	SMOTEBoost	0,772	0,858	0,501	0,679	0,509	0,779	0,853	0,552	0,703	0,542	0,796	0,859	0,608	0,733	0,590	0,805	0,873	0,598	0,736	0,598	0,811	0,869	0,636	0,752	0,620
	RUSBoost	0,680	0,669	0,715	0,692	0,613	0,701	0,691	0,730	0,710	0,621	0,741	0,736	0,758	0,747	0,654	0,747	0,733	0,790	0,761	0,681	0,765	0,758	0,789	0,773	0,685
	MWMOTE	0,788	0,931	0,337	0,634	0,446	0,801	0,934	0,396	0,665	0,488	0,814	0,936	0,444	0,690	0,538	0,822	0,932	0,490	0,711	0,577	0,826	0,940	0,483	0,711	0,579
	EasyEnsemble	0,768	0,894	0,386	0,640	0,463	0,800	0,906	0,483	0,694	0,538	0,812	0,911	0,514	0,712	0,574	0,822	0,904	0,577	0,740	0,618	0,831	0,909	0,595	0,752	0,637
	SMOTEBagging	0,766	0,879	0,411	0,645	0,461	0,790	0,883	0,509	0,696	0,541	0,803	0,885	0,556	0,720	0,582	0,809	0,882	0,591	0,736	0,606	0,813	0,881	0,608	0,744	0,618
	UnderBagging	0,658	0,631	0,742	0,687	0,514	0,704	0,677	0,787	0,732	0,569	0,730	0,707	0,799	0,753	0,596	0,744	0,720	0,817	0,769	0,615	0,752	0,728	0,824	0,776	0,624

Tablo 4.15 gerçek korelasyon ve 0,25 dengesizlik durumunu göstermektedir.

Tablo 4.15'de GDO değerlerindeki düşüş devam etmekte, DUY ve F-ölçüsü değerlerinde ise yükselme gözlemlenmektedir.

Örneklem genişliği 100 olduğu durumda, F-ölçüsü değerlerinde en fazla katkıyı RUSBoost algoritması sağlamaktadır. Algoritmaların katkıları YOK'a göre daha fazladır.

Örneklem genişliği arttıkça performanslarda yükselme olmaktadır. SVM ve RF yöntemlerinde en fazla katkıyı RUSBoost algoritması yapmakta iken CART yönteminde EasyEnsemble algoritması daha fazla katkı yapmaktadır. Bu algoritmaların yanı sıra sınıflama performansları üzerinde diğer algoritmaların katkısıda bulunmaktadır.

Örneklem genişliği 100'den 2000'e çıktığında F-ölçüsü değerlerinde yaklaşık %10 artış olmaktadır.

Tablo 4.16. Gerçek korelasyon ve 0,30 dengesizlik durumu

Sınıflama Yöntemleri	Örneklem Genişlikleri	100					250					500					1000					2000				
	Algoritmalar	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü	GDO	SEÇ	DUY	DDO	F-ölçüsü
DVM	YOK	0,751	0,903	0,388	0,646	0,496	0,775	0,907	0,464	0,685	0,542	0,787	0,914	0,488	0,701	0,574	0,796	0,918	0,509	0,714	0,597	0,803	0,925	0,518	0,721	0,611
	SMOTE	0,755	0,824	0,588	0,706	0,575	0,762	0,808	0,651	0,730	0,617	0,767	0,794	0,704	0,749	0,643	0,773	0,806	0,698	0,752	0,649	0,782	0,807	0,723	0,765	0,665
	SMOTEBoost	0,749	0,814	0,591	0,702	0,577	0,762	0,807	0,653	0,730	0,618	0,768	0,794	0,705	0,750	0,644	0,773	0,806	0,698	0,752	0,649	0,782	0,807	0,723	0,765	0,665
	RUSBoost	0,728	0,739	0,701	0,720	0,622	0,733	0,723	0,758	0,741	0,663	0,750	0,746	0,760	0,753	0,666	0,747	0,725	0,799	0,762	0,690	0,760	0,741	0,804	0,773	0,698
	MWMOTE	0,757	0,850	0,535	0,693	0,557	0,770	0,841	0,602	0,722	0,607	0,778	0,835	0,643	0,739	0,633	0,784	0,839	0,656	0,747	0,644	0,787	0,839	0,666	0,753	0,652
	EasyEnsemble	0,733	0,877	0,393	0,635	0,495	0,774	0,868	0,552	0,710	0,588	0,790	0,873	0,595	0,734	0,628	0,799	0,878	0,616	0,747	0,647	0,806	0,881	0,630	0,756	0,660
	SMOTEBagging	0,727	0,858	0,413	0,636	0,466	0,755	0,846	0,539	0,692	0,563	0,769	0,843	0,597	0,720	0,606	0,775	0,837	0,631	0,734	0,627	0,783	0,837	0,657	0,747	0,645
	UnderBagging	0,636	0,587	0,753	0,670	0,550	0,702	0,674	0,770	0,722	0,606	0,732	0,711	0,782	0,747	0,636	0,748	0,733	0,784	0,759	0,651	0,761	0,748	0,792	0,770	0,665
CART	YOK	0,715	0,892	0,291	0,592	0,455	0,741	0,886	0,399	0,643	0,486	0,756	0,896	0,429	0,662	0,511	0,764	0,897	0,454	0,675	0,531	0,774	0,898	0,484	0,691	0,559
	SMOTE	0,686	0,756	0,517	0,636	0,489	0,705	0,744	0,613	0,679	0,549	0,703	0,727	0,647	0,687	0,562	0,723	0,752	0,656	0,704	0,585	0,733	0,762	0,664	0,713	0,598
	SMOTEBoost	0,691	0,759	0,523	0,641	0,476	0,704	0,742	0,614	0,678	0,550	0,705	0,731	0,644	0,687	0,561	0,723	0,752	0,655	0,704	0,585	0,733	0,763	0,663	0,713	0,598
	RUSBoost	0,646	0,708	0,499	0,603	0,431	0,667	0,684	0,626	0,655	0,553	0,690	0,713	0,637	0,675	0,558	0,694	0,689	0,705	0,697	0,606	0,694	0,681	0,726	0,703	0,626
	MWMOTE	0,711	0,801	0,497	0,649	0,506	0,735	0,830	0,511	0,670	0,529	0,746	0,835	0,536	0,685	0,553	0,757	0,853	0,530	0,692	0,563	0,760	0,853	0,540	0,697	0,571
	EasyEnsemble	0,708	0,861	0,346	0,604	0,454	0,766	0,860	0,546	0,703	0,577	0,787	0,873	0,587	0,730	0,621	0,798	0,879	0,610	0,745	0,644	0,806	0,882	0,626	0,754	0,658
	SMOTEBagging	0,699	0,776	0,515	0,646	0,499	0,740	0,804	0,589	0,697	0,572	0,760	0,820	0,620	0,720	0,606	0,771	0,827	0,638	0,732	0,624	0,774	0,835	0,631	0,733	0,626
	UnderBagging	0,642	0,623	0,685	0,654	0,529	0,691	0,676	0,728	0,702	0,583	0,712	0,690	0,765	0,727	0,614	0,723	0,699	0,780	0,740	0,628	0,732	0,708	0,790	0,749	0,639
RF	YOK	0,740	0,909	0,338	0,623	0,424	0,767	0,917	0,414	0,666	0,505	0,781	0,917	0,463	0,690	0,554	0,791	0,916	0,499	0,707	0,587	0,799	0,917	0,521	0,719	0,607
	SMOTE	0,730	0,794	0,576	0,685	0,546	0,749	0,800	0,629	0,715	0,597	0,758	0,794	0,675	0,734	0,624	0,767	0,801	0,686	0,744	0,638	0,775	0,804	0,708	0,756	0,653
	SMOTEBoost	0,734	0,804	0,564	0,684	0,539	0,754	0,807	0,627	0,717	0,596	0,762	0,801	0,672	0,737	0,623	0,770	0,809	0,680	0,745	0,635	0,779	0,813	0,701	0,757	0,651
	RUSBoost	0,676	0,682	0,661	0,672	0,579	0,712	0,707	0,724	0,715	0,638	0,739	0,743	0,733	0,738	0,649	0,742	0,728	0,774	0,751	0,674	0,758	0,749	0,779	0,764	0,682
	MWMOTE	0,756	0,876	0,469	0,672	0,524	0,779	0,892	0,508	0,700	0,572	0,788	0,893	0,540	0,716	0,601	0,795	0,898	0,554	0,726	0,617	0,799	0,899	0,565	0,732	0,627
	EasyEnsemble	0,741	0,871	0,435	0,653	0,502	0,776	0,871	0,553	0,712	0,592	0,792	0,880	0,587	0,733	0,626	0,800	0,884	0,606	0,745	0,645	0,807	0,886	0,622	0,754	0,659
	SMOTEBagging	0,734	0,857	0,439	0,648	0,492	0,764	0,866	0,523	0,695	0,565	0,780	0,871	0,565	0,718	0,604	0,787	0,871	0,589	0,730	0,623	0,791	0,871	0,604	0,738	0,634
	UnderBagging	0,649	0,609	0,744	0,677	0,557	0,696	0,658	0,784	0,721	0,606	0,721	0,687	0,800	0,743	0,631	0,734	0,701	0,810	0,756	0,646	0,742	0,708	0,820	0,764	0,656

Tablo 4.16 gerçek korelasyon ve 0,30 dengesizlik durumunu göstermektedir.

Dengesizliğin azalması ve gerçek korelasyon yapısından dolayı DUY ve F-ölçüsü değerlerinde artış olmaktadır.

Tablo 4.16'e göre örneklem genişliği 100 olduğunda dengesizliğin azalması ve gerçek korelasyon yapısından dolayı DUY ve F-ölçüsü değerlerinde artış olmaktadır. F-ölçüsü değerlerine bakıldığında, DVM ve RF yöntemlerinde RUSBoost algoritması, CART yönteminde ise UnderBagging algoritması en fazla katkıyı sağlamaktadır.

Örneklem genişliği arttıkça DUY değerlerindeki artış F-ölçüsü değerlerine de yansımaktadır. F-ölçüsü değerleri üzerinde YOK hariç diğer algoritmaların etkileri görülmektedir. Ayrıca, örneklem genişliği arttıkça DVM ve RF yöntemlerinde RUSBoost etkili iken CART yönteminde EasyEnsemble algoritmasının etkisi gözükmektedir.

4.2 Gerçek Veri Setlerine Ait Sonuçlar

Tablo 4.17'de gerçek veri setlerine ait sınıflama performans sonuçları yer almaktadır, n örneklem genişliği, k değişken sayısını, r korelasyon düzeyini ve DO dengesizlik oranını temsil etmektedir, Sonuçlar Düzeltilmiş doğruluk oranı ve F-ölçüsü üzerinden sunulmuştur.

- ***Deniz Kabukları (Abalone)***

  Düşük korelasyon yapısına sahip ve dengesizliğin %5,74 olduğu veri setidir.

  DDO değerlerine bakıldığında, DVM ve CART yöntemlerinin performansları üzerinde en çok etkili RUSBoost algoritması görülmekte iken, RF yönteminde en çok etkili UnderBagging algoritması görülmektedir.

  F-ölçülerine bakıldığında bütün algoritmalar için sonuçlar çok düşük çıkmıştır. Bu algoritmaların düşük korelasyon yapısı ve dengesizliğin yüksek olduğu durumlarda sonuçlar üzerinde etkileri görülmemektedir.

- ***Doğurganlık (Fertility)***

  Düşük korelasyon yapısına sahip ve dengesizliğin %12 olduğu bir veri setidir.

  DDO değerlerine bakıldığında, sınıflama yöntemlerinin performansları üzerinde en çok etkili SMOTEBagging ve UnderBagging algoritmaları görülmektedir. UnderBagging algoritması DVM yönteminin performansı %93'e çıkarmıştır.

  UnderBagging ve SMOTEBagging algoritmalarının DVM yönteminin F-ölçüsü değerleri üzerinde etkisi net bir şekilde görülmektedir. F-ölçüsü değerleri %80 üzerine çıkmaktadır. CART ve RF yöntemlerinin üzerinde etkileri daha az olmuştur.

- ***Göğüs Cerrahisi (Thoracic Surgery)***

  Düşük korelasyon yapısına sahip ve dengesizliğin %15 olduğu bir veri setidir.

  DVM, CART ve RF yöntemlerinin performanslarını DDO değerlerine bakarak değerlendirdiğimizde en çok etkili SMOTEBagging algoritması görülmektedir.

  SMOTEBagging algoritmasının DVM, CART ve RF yöntemlerinin F-ölçüsü değerleri üzerinde etkisi diğer algoritmalarının etkisine göre daha iyidir. Düşük korelasyon yapısının etkisi sonuçlara yansıdığı görülmektedir.

- ***Hepatit***

  Orta düzey korelasyon yapısına sahip ve dengesizliğin %20 olduğu bir veri setidir.

  DVM, CART ve RF yöntemlerinin performanslarını DDO değerlerine bakarak değerlendirdiğimizde SMOTE ve SMOTEBoost algoritmalarının DVM yönteminin performansını %83'e çıkardığı görülmektedir. Ayrıca RF sınıflama yöntemindeki performanslar (%90) diğer sınıflama yöntemlerine göre daha yüksek bulunmuştur.

  F-ölçüsü değerlerine bakıldığında en iyi sınıflama performans DVM yöntemine ait olup SMOTE, SMOTEBoost ve RUSBoost algoritmaları performansı %80'e çıkarmıştır. RF yönteminin performansı algoritma YOK olduğu durum ile SMOTE, SMOTEBoost ve RUSBoost algoritmalarının kullanıldığı durumlarda benzer bulunmuştur. En kötü performansa sahip yöntem CART yöntemidir. Buna göre, Orta düzey korelasyonun ve dengesizliğin azalmış olmasının etkisi sınıflama performansları üzerinde artışa neden olmuştur.

- ***Kan Nakli (Blood Transfusion)***

  Düşük düzey korelasyon yapısına sahip ve dengesizliğin %23 olduğu bir veri setidir.

  DDO değerlerine bakıldığında RF sınıflama yönteminde MWMOTE algoritmasını performansı %87,5 yaparken, DVM ve CART yöntemlerinde RUSBoost algoritmasının etkisi görülmektedir.

  DVM, CART ve RF yöntemlerinde F-ölçüsü değerlerine bakıldığında sınıflama performansları %50 civarına kadar çıkmaktadır. Dengesizlik azaldığı halde ilişkinin Düşük düzey korelasyon sahip olması sınıflama performanslarını etkilemiştir.

- ***Alzheimer***

  Düşük düzey korelasyon yapısına sahip, dengesizliğin %30 olduğu ve veriler arasında en fazla değişkeni olan bir veri setidir.

  Hem DDO hem de F-ölçüsü değerlerine bakıldığında EasyEnsemble algoritması hariç diğer algoritmalar sınıflama performanslarını arttırmaktadır. Ayrıca MWMOTE algoritmasının her iki ölçüde de sınıflama yöntemleri üzerinde etkili olduğu görülmektedir.

DDO değerlerine baktığımızda sınıflama performansları %70'in üzerinde bulunmaktadır. Sınıflama yöntemlerinde ise RF yönteminin performansı diğerlerine göre yaklaşık %3 fazladır.

F-ölçüsü değerlerine bakıldığında MWMOTE algoritması DVM ve RF yöntemleri üzerinde etkili iken, CART üzerinde UnderBagging algoritması daha etkili bulunmaktadır. Sınıflama yöntemlerinin performansları yaklaşık %70 civarındadır. Dengesizliğin azalmış olmasına rağmen verinin düşük korelasyon yapısına sahip olması sınıflama performansları çok yüksek değildir.

- ***Diyabet***

  Orta düzey korelasyon yapısına sahip ve dengesizliğin %34,9 olduğu bir veri setidir.

  Hem DDO hem de F-ölçüsü değerlerine bakıldığında Alzaheirmer verisinde olduğu gibi EasyEnsemble algoritması hariç diğer algoritmalar sınıflama performanslarını arttırmaktadır. Bunun yanısıra, RUSBoost algoritmasının etkisinin diğerlerine göre daha fazla olduğu (yaklaşık %5) her iki performans ölçüsünde de görülmektedir.

  DDO değerlerine baktığımızda sınıflama performansları %70'in üzerinde bulunmaktadır. RF yönteminin sınıflama performansı yaklaşık %3 diğer yöntemlere göre fazladır.

  F-ölçüsü değerlerine bakıldığında SMOTE, SMOTEBoost ve RUSboost algoritmalarının sınıflama performansları üzerindeki etkileri benzer bulunmakta ve yaklaşık %60'ın üzerindedir. Sınıflama yöntemleri içerisinde en iyi performans RF yöntemine aittir. Dengesizlik azalmış ve orta düzey korelasyon yapısından dolayı sonuçlar %60'ın üzerinde çıkmaktadır.

Tablo 4.17. Gerçek veri setlerine ait sınıflama performansı sonuçları

Veri Setleri	Algoritmalar	DVM		CART		RF	
		DDO	F-Ölçüsü	DDO	F-Ölçüsü	DDO	F-Ölçüsü
Deniz Kabukları n=731 k=9 r=düşük düzey DO=%5,74	**YOK**	0,500	NA	0,581	0,267	0,623	0,375
	SMOTE	0,583	0,286	0,578	0,250	0,576	0,375
	SMOTEBoost	0,583	0,286	0,655	0,250	0,662	0,235
	RUSBoost	0,812	0,286	0,709	0,250	0,682	0,235
	MWMOTE	0,583	0,286	0,606	0,261	0,620	0,353
	EasyEnsemble	0,166	0,022	0,225	0,033	0,185	0,022
	SMOTEBagging	0,694	0,293	0,640	0,222	0,598	0,231
	UnderBagging	0,753	0,302	0,699	0,246	0,815	0,313
Doğurganlık n=100 k=5 r=düşük düzey DO=%12	**YOK**	0,500	NA	0,500	NA	0,667	0,500
	SMOTE	0,423	NA	0,404	NA	0,404	NA
	SMOTEBoost	0,647	0,400	0,500	0,182	0,647	0,500
	RUSBoost	0,500	0,400	0,500	0,182	0,622	0,500
	MWMOTE	0,631	0,444	0,500	NA	0,631	0,444
	EasyEnsemble	0,036	NA	0,667	0,500	0,071	NA
	SMOTEBagging	0,881	0,833	0,726	0,615	0,726	0,615
	UnderBagging	0,929	0,857	0,738	0,625	0,738	0,625
Göğüs Cerrahisi n=470 k=16 r=düşük düzey DO=%15	**YOK**	0,500	NA	0,500	NA	0,500	NA
	SMOTE	0,451	0,113	0,547	0,268	0,514	0,248
	SMOTEBoost	0,416	0,113	0,520	0,268	0,533	0,269
	RUSBoost	0,539	0,113	0,551	0,261	0,548	0,269
	MWMOTE	0,462	NA	0,500	NA	0,423	NA
	EasyEnsemble	0,423	NA	0,500	NA	0,615	0,231
	SMOTEBagging	0,667	0,500	0,756	0,444	0,667	0,500
	UnderBagging	0,468	0,167	0,500	0,188	0,660	0,286

Table 4.17. Gerçek veri setlerine ait sınıflama performansı sonuçları (devamı)

Veri Setleri	Algoritmalar	DVM		CART		RF	
		DDO	F-Ölçüsü	DDO	F-Ölçüsü	DDO	F-Ölçüsü
Hepatit n=155 k=19 r=orta düzey DO=%20	**YOK**	0,500	NA	0,500	NA	0,767	0,667
	SMOTE	0,833	0,800	0,642	0,400	0,895	0,667
	SMOTEBoost	0,833	0,800	0,642	0,400	0,921	0,667
	RUSBoost	0,733	0,800	0,658	0,400	0,796	0,600
	MWMOTE	0,500	NA	0,506	0,121	0,521	0,167
	EasyEnsemble	0,470	0,216	0,501	0,197	0,392	0,175
	SMOTEBagging	0,464	0,093	0,533	0,200	0,531	0,148
	UnderBagging	0,448	0,211	0,579	0,289	0,661	0,381
Kan Nakli n=748 k=5 r=düşük düzey DO=%23	**YOK**	0,625	0,410	0,500	NA	0,615	0,404
	SMOTE	0,553	0,247	0,540	0,219	0,593	0,366
	SMOTEBoost	0,553	0,247	0,540	0,219	0,597	0,366
	RUSBoost	0,730	0,477	0,666	0,505	0,621	0,582
	MWMOTE	0,642	0,400	0,592	0,286	0,875	0,545
	EasyEnsemble	0,292	0,105	0,500	0,231	0,317	0,111
	SMOTEBagging	0,667	0,500	0,592	0,286	0,642	0,400
	UnderBagging	0,617	0,333	0,492	0,182	0,542	0,222
Alzheimer n=70 k=26 r=düşük düzey DO=%30	**YOK**	0,560	0,364	0,476	0,200	0,595	0,400
	SMOTE	0,679	0,571	0,679	0,571	0,679	0,400
	SMOTEBoost	0,679	0,571	0,679	0,571	0,679	0,571
	RUSBoost	0,631	0,571	0,536	0,571	0,488	0,571
	MWMOTE	0,776	0,693	0,756	0,673	0,808	0,733
	EasyEnsemble	0,212	0,083	0,276	0,241	0,244	0,155
	SMOTEBagging	0,705	0,612	0,769	0,692	0,673	0,560
	UnderBagging	0,737	0,652	0,788	0,713	0,788	0,710

Table 4.17. Gerçek veri setlerine ait sınıflama performansı sonuçları (devamı)

Veri Setleri	Algoritmalar	DVM		CART		RF	
		DDO	F-Ölçüsü	DDO	F-Ölçüsü	DDO	F-Ölçüsü
Diyabet n=768 k=9 r=orta düzey DO=%34,9	**YOK**	0,667	0,517	0,673	0,551	0,679	0,548
	SMOTE	0,731	0,641	0,712	0,613	0,763	0,683
	SMOTEBoost	0,731	0,641	0,673	0,613	0,750	0,683
	RUSBoost	0,744	0,641	0,712	0,613	0,788	0,683
	MWMOTE	0,683	0,521	0,690	0,531	0,653	0,469
	EasyEnsemble	0,266	0,096	0,310	0,186	0,304	0,177
	SMOTEBagging	0,712	0,532	0,698	0,520	0,637	0,453
	UnderBagging	0,726	0,563	0,690	0,521	0,653	0,471

5 TARTIŞMA

Bu tez kapsamında, benzetim çalışması ile elde edilen veri setleri ve gerçek veri setleri üzerinde çalışılmıştır. İki sınıflı veri setlerinde farklı sınıflama yöntemlerinin sonuçları performans ölçüsü olarak kullanılan F-ölçüsü üzerinden değerlendirilmiştir. Literatürde çoğu çalışmada algoritmalar, gerçek veri setleri üzerinde değerlendirilmiştir. Bundan dolayı, bu tez kapsamında ele alınan senaryoların benzeri bir çalışma literatürde bulunmamaktadır.

Literatürde yapılan çalışmalarda farklı algoritmalar uygulayarak çok çeşitli sonuçlara ulaşılmaktadır. Quinlan (49) 1996 yılındaki çalışmasında topluluk öğrenme yöntemlerle yapılan sınıflamaların iyi sonuç verdiğini göstermektedir.

Chawla ve diğ. (1) 2002 yılında SMOTE algoritması ile farklı bir yaklaşım izleyerek azınlık sınıfı üzerinden yeni gözlemler (yapay) üreterek azınlık sınıfı gözlem sayısını arttırmaya çalışmışlardır. Farklı bir yaklaşım olmasına rağmen algoritmada yapay gözlem üretme arka planda uzun sürebilmektedir. Diğer taraftan Seiffert ve diğ. (6) 2010 yılında RUSBoost algoritması ile çoğunluk sınıfından azınlık sınıfı kadar örneklemler çekmenin daha iyi sonuçlar verdiğini göstermiştir. Benzetim çalışmamızda RUSBoost algoritması 4 farklı korelasyon yapısında iyi sonuç veren algoritmalar arasındadır. Ayrıca, Seiffert ve diğ. (6) RUSBoost'un SMOTE ve SMOTEBoost algoritmalarına göre arka planda kısa sürede sonuçlar verdiğini göstermişlerdir.

Liu ve diğ. (4) 2009 yılındaki makalelerinde 15 farklı algoritmayı gerçek veri setlerini kullanarak karşılaştırmıştır. Ayrıca, makalelerinde EasyEnsemble ve BalanceCascade adını verdikleri iki yeni algoritma önermişlerdir. Bu algoritmaların arka planda çalışma sürelerini azalttıklarını bunun yanı sıra, çoğunluk grubundan her gözlemin kullanılmasının gerekli olmadığını söylemişlerdir. Adaboost ve Bagging yöntemlerinin karar ağaçlarının performanslarını arttırdıklarını göstermişlerdir. Tez çalışmamızda, Bagging ve Boosting tabanlı algoritmaların performansları arttırdığını benzetim çalışmasında gördük.

Qing ve diğ. (50) makalelerinde 11 farklı algoritmayı gerçek veri setlerinde uygulamışlardır. EasyEnsemble algoritmasının veri setlerinde performansları arttırmadığı RUSBoost ve UnderBagging algoritmalarının performanslarının EasyEnsemble algoritmasından daha iyi olduğunu göstermişlerdir. Benzetim çalışmamızda ve gerçek veri setlerinde RUSBoost algoritmasının EasyEnsemble algoritmasına göre performanslar üzerinde etkisi daha fazla olarak bulmuştuk.

6 SONUÇ VE ÖNERİLER

Sınıf dengesizliği problemi, makine öğrenmede önemli bir konudur. Bu konu ile ilgili literatürde birçok çalışma vardır. Gün geçtikçe araştırma sayısı artmakta ve yeni algoritmalar önerilmektedir. Bu algoritmaların ortak özellikleri vardır. Ortak olarak temel aldıkları,

- sentetik veri üretme
- az örnekleme
- aşırı örnekleme

gibi bazı yaklaşımlardır. Algoritmalar çeşitli kodlama içeren bazı programlarda kullanılmaktadır. Bu algoritmaların süre açısından farklılıkları vardır. Aşırı örnekleme uygulayan algoritmaların süreleri diğerlerine göre daha uzun sürebilmektedir. Bu da aşırı örnekleme uygulayan algoritmaların dezavantajlarından biridir.

Literatürde yapılan çalışmaların birkaçında hem gerçek veri setleri hem de benzetim çalışması uygulanmıştır; ancak benzetim çalışmaları farklı korelasyon yapıları altında incelenmemiştir. Bu çalışmalardan farklı olarak sonuçları etkileyebilecek üç etkinin olduğu benzetim çalışması uygulanmıştır. Farklı korelasyon yapıları, farklı örneklem genişlikleri ve farklı dengesizlik oranları sınıflama performanslarını etkilemektedir. Bu üç etkinin olduğu 80 farklı senaryo ile algoritmaların geçerliği gösterilmeye çalışılmıştır.

Benzetim çalışmasının sonuçları Bölüm 4.1'de her bir korelasyon düzeyi için ayrı ayrı verilmiştir. Genel olarak GDO, SEÇ ve DDO değerleri yüksek çıkmakta iken DUY ve F-ölçüsü değerleri daha düşük çıkmaktadır. SEÇ değerleri çok yüksek, DUY değerleri ise düşük olduğunda GDO değerlerine bakarak yorum yapmak yanıltıcı olacaktır. Aynı şekilde DDO değerleri de DUY ve SEÇ değerlerinden hesaplandığı için DDO değerleri etkilenmektedir. Sayılan bu sebeplerden dolayı sınıf dengesizliği problemi olduğunda F-ölçüsü üzerinden performanslar değerlendirilmiştir.

Düşük düzey korelasyon yapısında, 0.10 ve 0.15 dengesizlik oranlarında RUSBoost algoritması örneklem genişliğinin artması ile sonuçları %30 ve üzerine çıkarmıştır. 0.25 ve 0.30 dengesizlik oranlarında SMOTE, SMOTEBoost, RUSBoost, MWMOTE, SMOTEBagging ve UnderBagging algoritmalarının etkileri örneklem genişliği arttıkça artmaktadır. EasyEnsemble algoritması düşük düzey korelasyon yapısında sınıflama performansları üzerinde etkili olamamıştır.

Verilerin orta düzey korelasyona sahip olmasından dolayı, düşük korelasyon yapısındaki performans sonuçlarına göre daha yüksek performans sonuçları elde edilmiştir. 0.10 dengesizlikte küçük örneklemde (n=100) EasyEnsemble etkili bir algoritma iken örneklem genişliği arttıkça RUSBoost algoritmasının etkisi artmıştır. 0.15 dengesizlikte EasyEnsemble'ın etkisi örneklem genişliği arttıkça artmaktadır. Küçük örneklem genişliğinde SMOTEBagging ve RUSBoost'un etkisi vardır. 0.25 dengesizlikte örneklem genişliği arttıkça EasyEnsemble algoritması performansları %70'in üzerine çıkarmaktadır. Küçük örneklem genişliğinde DVM yönteminde RUSBoost ve EasyEnsemble algoritmaları, CART yönteminde EasyEnsemble ve SMOTEBagging algoritmaları, RF yönteminde ise EasyEnsemble ve UnderBagging algoritmaları %60 ve üzerinde çıkmıştır. YOK'a göre algoritmaların sınıflama yöntemleri performanslarını net bir şekilde arttırdığı görülmektedir. Dengesizlik 0.30 olduğunda EasyEnsemble algoritması daha etkili görülmektedir.

Yüksek düzey korelasyonda, korelasyon yapısının yüksek olması her bir algoritmada sınıflama performanslarını yükseltmektedir. 0.10 dengesizlik durumunda, YOK hariç algoritmalarde örneklem genişliği arttıkça F-ölçüsü değerlerinde de artış olmaktadır. RUSBoost algoritması sonuçlar üzerinde diğer algoritmalara göre daha etkilidir. Düşük ve orta korelasyon yapılarındaki gibi örneklem genişliği arttıkça performanslarda çok yüksek bir artış olmamaktadır. 0.15 dengesizlikte, 2000 örneklem genişliğinde DVM yönteminde UnderBagging ve MWMOTE algoritmalarının, CART yönteminde SMOTE, SMOTEBoost ve EasyEnsemble algoritmalarının ve RF yönteminde ise RUSBoost, SMOTEBagging ve EasyEnsemble algoritmalarının katkısı diğer algoritmalara göre fazladır. 0.25 ve 0.30 dengesizlikte tüm algoritmaların etkileri örneklem genişliği arttıkça artmaktadır.

Gerçek korelasyonda, 0.10 dengesizlikte F-ölçüsü değerleri üzerinde, DVM ve RF yöntemlerinde RUSBoost, MWMOTE ve SMOTEbagging algoritmaları, CART yönteminde ise SMOTEBagging algoritması YOK'a göre daha etkili görülmektedir. 2000 örneklem genişliğinde RUSBoost, DVM ve RF'de performansları %70'in, CART yönteminde ise %60'ın üzerine çıkarmaktadır. 0.15 dengesizlikte örneklem genişliği arttıkça F-ölçüsü değerleri üzerinde en çok RUSBoost etkili görünmesinin yanı sıra EasyEnsemble ve SMOTEBagging algoritmaları da diğer algoritmalara göre performansları arttırmaktadır. Dengesizlik 0.25 ve 0.30 olduğunda SVM ve RF yöntemlerinde en fazla katkıyı RUSBoost algoritması yapmakta iken CART yönteminde EasyEnsemble algoritması daha fazla katkı yapmaktadır. Bu algoritmaların yanı sıra sınıflama performansları üzerinde diğer

algoritmaların katkısı da bulunmaktadır.

Genel sonuç olarak sınıflama yöntemlerini uygulamadan önce aşağıdaki etkiler göz önünde bulundurulmalıdır.

- Verideki dengesizlik oranına,
- Korelasyon yapısına,
- Örneklem genişliğine

Çözüm algortimalarının dengesizlik durumunda (%30'un altında) kullanılmasının sınıflama performansları üzerinde etkili olduğu görülmektedir.

Gelecekte bu alanda yapılabilecek çalışmalar;

Benzetim çalışmamızda dikkate alınmayan bağımsız değişkenler arasında ilişkininde olduğu durumlar dikkate alınarak algoritmalar uygulanabilir.

Dikkate almadığımız farklı dengesizlik oranları incelenebilir.

Gerçek veri setlerinde yüksek korelasyona sahip veri yapıları da çalışmaya dahil edilebilir.

Çok sınıflı veri setleri için çalışmalar genişletilebilir.

Algoritmalar gen veri setlerinde de uygulanabilir.

7. KAYNAKLAR

1. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–357.
2. He H, Ma Y. Imbalance Learning: Foundation, Algorithms, and Application. Hoboken,New Jersey: Wiley; 2013.
3. Galar M, Fernández A, Tartas EB, Sola HB, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Trans Systems, Man, and Cybernetics, Part C. 2012;42(4):463–484.
4. Liu XY, Wu J, Zhou ZH. Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans Systems, Man, and Cybernetics, Part B. 2009;39(2):539–550.
5. Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. In: Lavrac N, Gamberger D, Blockeel H, Todorovski L, editors. PKDD, vol. 2838 of Lecture Notes in Computer Science Springer; 2003. p. 107–119.
6. Seiffert C, Khoshgoftaar TM, Hulse JV, Napolitano A. RUSBoost: Improving classification performance when training data is skewed. In: ICPR IEEE Computer Society; 2008. p. 1–4.
7. Estabrooks A, Jo T, Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence. 2004;20(1):18–36.
8. KrishnaVeni C, Sobha Rani T. On the Classification of Imbalanced Datasets . International Journal of Computer Science and Technology. 2011;2(1):145–148.
9. Chawla NV, Japkowicz N, Kotcz A. Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explor Newsl. 2004 Jun;6(1):1–6.
10. Zhang J, Mani I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets; 2003. p. 42–48.
11. Cao P, Zhao D, Zaiane O. An Optimized Cost-Sensitive SVM for Imbalanced Data Learning. In: Advances in Knowledge Discovery and Data Mining Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 280–292.
12. Cieslak DA, Chawla NV. Learning Decision Trees for Unbalanced Data. In: Daelemans W, Goethals B, Morik K, editors. ECML/PKDD (1), vol. 5211 of Lecture Notes in Computer Science Springer; 2008. p. 241–256.
13. Drummond C, Holte RC. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. In: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets II Washington DC; 2003. p. 1–8.

14. Freund Y. Boosting a Weak Learning Algorithm by Majority. Inf Comput. 1995 Sep;121(2):256–285.

15. Barandela R, Sanchez J, Valdovinos R. New Applications of Ensembles of Classifiers. Pattern Analysis and Applications. 2003;6(3):245–256.

16. Guo H. Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. SIGKDD Explorations. 2004;6:30–39.

17. Han H, Wang WY, Mao BH. Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning. In: Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part I ICIC'05, Berlin, Heidelberg: Springer-Verlag; 2005. p. 878–887.

18. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IJCNN; 2008. p. 1322–1328.

19. Wang S, Yao X. Diversity analysis on imbalanced data sets by using ensemble models. In: CIDM IEEE; 2009. p. 324–331.

20. Chen S, He H, Garcia EA. RAMOBoost: Ranked Minority Oversampling in Boosting. IEEE Transactions on Neural Networks. 2010;21:1624–1642.

21. Barua S, Islam MM, Yao X, Murase K. MWMOTE-Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. IEEE Trans Knowl Data Eng. 2014;26(2):405–425.

22. Hanifah FS. SMOTE Bagging Algorithm for Imbalanced Dataset in Logistic Regression Analysis (Case: Credit of Bank X). vol. 9; 2015. p. 6857–6865.

23. Vapnik VN. The nature of statistical learning theory. New York, NY, USA: Springer-Verlag New York, Inc.; 1995.

24. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. 3 ed. Morgan Kaufmann Series in Data Management Systems, Amsterdam: Morgan Kaufmann; 2011.

25. Han J, Kamber M. Data mining: concepts and techniques. San Francisco:CA: Morgan Kaufmann; 2006.

26. Shuo W, Xin Y. Diversity analysis on imbalanced data sets by using ensemble models; 2009. p. 324–331.

27. Şimşek Gürsoy T. Veri Madenciliği ve Bilgi Keşfi. Pegem Akademi; 2009.

28. Rochana L, Comparison of Data Mining and Statistical Tech- niques for Classification Model; 2006.

29. Larose DT. Discovering Knowledge in Data: An Introduction to Data Mining. New York, NY, USA: Wiley-Interscience; 2004.

30. Olson DL, Delen D. Advanced Data Mining Techniques. 1st ed. Springer Publishing Company, Incorporated; 2008.

31. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Us ed. Addison Wesley; 2005.

32. Ho TK. Random Decision Forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1 ICDAR '95, Washington, DC, USA: IEEE Computer Society; 1995. .

33. Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32.

34. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. No. 57 in Monographs on Statistics and Applied Probability, Boca Raton, Florida, USA: Chapman & Hall/CRC; 1993.

35. Breiman L. Bagging Predictors. Mach Learn. 1996 Aug;24(2):123–140.

36. Lunardon N, Menardi G, Torelli N. ROSE: a Package for Binary Imbalanced Learning. R Journal. 2014;6(1):82–92.

37. He H, Garcia EA. Learning from Imbalanced Data. Knowledge and Data Engineering, IEEE Transactions on. 2009 Sept;21(9):1263–1284.

38. Menardi G, Torelli N. Training and Assessing Classification Rules with Imbalanced Data. Data Min Knowl Discov. 2014 Jan;28(1):92–122. http://dx.doi.org/10.1007/s10618-012-0295-5.

39. Karaağaoğlu E, Karakaya J, Kılıçkap M. Tanı Testlerinin Değerlendirilmesinde İstatistiksel Yöntemler. Detay Yayıncılık; 2016.

40. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Sciences Series; 2003.

41. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. In: Proceedings of the 2010 20th International Conference on Pattern Recognition ICPR '10, Washington, DC, USA: IEEE Computer Society; 2010. p. 3121–3124. https://doi.org/10.1109/ICPR.2010.764.

42. Powers DMW. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Adelaide, Australia: School of Informatics and Engineering, Flinders University; 2007.

43. Demirtas H, Amatya A, Doganay B. BinNor: An R Package for Concurrent Generation of Binary and Normal Data. Communications in Statistics - Simulation and Computation. 2014;43(3):569–579.

44. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Articles. 2008;28(5):1–26. https://www.jstatsoft.org/v028/i05.

45. Deane-Mayer ZA, Knowles JE. caretEnsemble: Ensembles of Caret Models; 2016, https://CRAN.R-project.org/package=caretEnsemble, r package version 2.0.0.

46. Cordón I. imbalance: Preprocessing Algorithms for Imbalanced Datasets; 2017, https://CRAN.R-project.org/package=imbalance, r package ver-

sion 0.1.1.

47. Hao H, Chen. ebmc: Ensemble-Based Methods for Class Imbalance Problem; 2017, `https://CRAN.R-project.org/package=ebmc`, r package version 1.0.0.

48. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2009.

49. Quinlan JR. Bagging, Boosting, and C4.S. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1 AAAI Press; 1996. p. 725–730.

50. Yin QY, Zhang JS, Zhang CX, Ji NN. A Novel Selective Ensemble Algorithm for Imbalanced Data Classification Based on Exploratory Undersampling. Mathematical Problems in Engineering. 2014;.

ÖZGEÇMİŞ

Duygu AYDIN HAKLI

Contact Information

Hacettepe University, Faculty of Medicine, Department of Biostatistics Ankara, Turkey

+90 5303722446
d_aydn24@hotmail.com

Education

Marmara University, Ankara

M.S., Operation Research, Aug 2012

- Topic: *Demand forecasting analysis with artificial neural networks and an application in maritime transport sector*
- Advisor: Habip KOCAK, Assistant Professor

Mimar Sinan Fine Art University, Istanbul,Turkey

B.S., Statistics , May 2008

Research Experience

Research Assistant — June 2012 to present
Department of Biostatistics,
University of Hacettepe

Research Assistant — September 2011 to June 2012
Department of Biostatistics,
University of Duzce

Research Assistant — January 2011 to Aug 2011
Department of Statistics,
University of Artvin Coruh

Refereed Journal Publications

1. Beksac Mehmet Sinan, Tanacan Atakan, **Aydin Hakli Duygu**, Ozyuncu Ozgur. "Use of the 50g glucose challenge test to predict excess delivery weight." *International Journal of Gynecology & Obstetrics*, 2018.
2. Samadi Afshin, **Aydin Hakli Duygu**, Lay Incilay. "Effects of haloperidol and clozapine treatment on plasma concentration of thyroid hormones in Rats." *The FEBS Journal*, 283 127–427, 2016.
3. Aksehirli Ozge, Ankarali Handan, **Aydin Hakli Duygu**, Baltaci Davut. "A Comparison of Error Correcting Output Coding Methods for Multiclass Classification by Using Support Vector Machine The Prediction of Self Monitoring of Blood Sugar." *International Journal of Statistics in Medical Research*, 2(2):123–134, 2013.
4. Aksehirli Ozge, Ankarali Handan, Kizilay Munevver, Arslanoglu Ilknur **Aydin Hakli Duygu**. "The Use of Nonparametric Quantile Regression and Least Median of Squares Regression for Construction of Growth Curves of Weight." *Turkish Journal of Medical Science*, 33(3):692–701, 2013.
5. Ankarali Handan, **Aydin Hakli Duygu**, Aksehirli Ozge. "Correlation Structure Between Items and Sample Size Effects on Factoring: A Simulation Study." *Turkish Journal of Medical Science*, 33(3):751–761, 2013.
6. Aksehirli Ozge, **Aydin Hakli Duygu**, Ankarali Handan, Sezgin Melek. "Knee Osteoarthritis Diagnosis Using Support Vector Machine and Probabilistic Neural Network." *International Journal of Computer Science*, 10(3):283–291, 2013.
7. Aksehirli Ozge, Ankarali Handan, **Aydin Hakli Duygu**, Saracli Ozge. "An Alternative Approach in Medical Diagnosis: Support Vector Machine." *Turkish Journal of Biostatistics*, 5(1):19–28, 2013.
8. Ankarali Handan, Aksehirli Ozge,**Aydin Hakli Duygu**. "Comparison of Paired T Test and Wilcoxon Signed Rank Test forVarious Change Measures in Pre Post Designs A Simulation Study." *Turkish Journal of Biostatistics*, 4(2) 55-59 2012.

Submitted Journal Publications

1. "Gestational outcomes of pregnant women following invasive prenatal testing for spinal muscular atrophy. " 2017. Submitted to *Prenatal Diagnosis*.
2. "Polypharmacy and Drug Related Problems Among Turkish HIV Infected Patients: A Single Center Experience. " 2017. Submitted to *HIV Medicine*.
3. "Changing Rates of the Modes of Delivery Over the Decades (1976, 1986, 1996, 2006, and 2016) Based on the Robson-10 Group Classification System in a Single Tertiary Health Care" 2018. Submitted to *Birth*.
4. "Evaluation of the factors affecting classification performance in class imbalance problem" 2018. Submitted to *Data & Knowledge Engineering*.

Symposium, Congress, Conference

Statistical Meetings

- 10th International Statistics Congress, Ankara, Turkey — December 2017
- User2017, Brussels, Belgium — July 2017
- 9th Conference Of The Eastern Mediterranean Region and The Italian Region Of The International Biometric Society, Selanik, Greece — May 2017
- XVIII. National Biostatistics Congress, Antalya, Turkey — October 2016
- 4th Big Data Spain, Madrid, Spain — October 2015
- XVII. National Biostatistics Congress, Girne, KKTC — October 2015
- 8th conference of the Eastern Mediterranean Region of the International Biometric Society, Cappadocia, Turkey — May 2015
- 4th International Interdisciplinary Choas Symposium on Choas and Complex Systems, Antalya, Turkey — May 2012