KORONER ARTER HASTALARINDA HİPERTANSİYONUN SINIFLANDIRILMASI İÇİN DENGESİZ SINIF PROBLEMİNİN TIBBİ BİLGİ KEŞFİ SÜRECİ İLE GİDERİLMESİ

Ahmet Kadir ARSLAN

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

Tez Danışmanı
Prof. Dr. Cemil ÇOLAK

Yüksek Lisans Tezi – 2018

T.C.

İNÖNÜ ÜNİVERSİTESİ

SAĞLIK BİLİMLERİ ENSTİTÜSÜ

KORONER ARTER HASTALARINDA HİPERTANSİYONUN SINIFLANDIRILMASI İÇİN DENGESİZ SINIF PROBLEMİNİN TIBBİ BİLGİ KEŞFİ SÜRECİ İLE GİDERİLMESİ

Ahmet Kadir ARSLAN

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı

Prof. Dr. Cemil ÇOLAK

MALATYA

2018

KABUL VE ONAY SAYFASI

İnönü Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalı Yüksek Lisans Programı çerçevesinde yürütülmüş olan; **Ahmet Kadir ARSLAN'ın "Koroner Arter Hastalarında Hipertansiyonun Sınıflandırılması İçin Dengesiz Sınıf Probleminin Tıbbi Bilgi Keşfi Süreci İle Giderilmesi "** konulu bu çalışması, aşağıdaki jüri tarafından Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 04/01/2018

Prof. Dr. Saim YOLOĞLU
İnönü Üniversitesi
Jüri Başkanı

Prof. Dr. Cemil ÇOLAK
İnönü Üniversitesi
Tez Danışmanı
Üye

Yrd. Doç. Dr. Mehmet Onur KAYA
Fırat Üniversitesi
Üye

ONAY

Bu tez, İnönü Üniversitesi Lisansüstü Eğitim-Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından kabul edilmiş ve Enstitü Yönetim Kurulu'nun ...../...../2018 tarih ve 2018/.....sayılı Kararıyla da uygun görülmüştür.

Prof. Dr. Yusuf TÜRKÖZ
Enstitü Müdürü

İÇİNDEKİLER

TEŞEKKÜR

Akademik eğitimim ve çalışmalarımın yanı sıra günlük yaşantımda bilgi, birikim ve deneyimleri ile bana yol gösteren ve destek olan değerli danışman hocam Sayın Prof. Dr. Cemil ÇOLAK'a, eğitimim boyunca desteğini esirgemeyen ve önerileriyle bana ışık tutan Sayın Prof. Dr. Saim YOLOĞLU'na sonsuz teşekkür ve saygılarımı sunarım. Bu süreçte yardımını hiç esirgemeyen, destekleriyle beni hiçbir zaman yalnız bırakmayan eşime sonsuz teşekkür ederim.

Arş. Gör. Ahmet Kadir ARSLAN

ÖZET

Koroner Arter Hastalarında Hipertansiyonun Sınıflandırılması için Dengesiz Sınıf Probleminin Tıbbi Bilgi Keşfi Süreci ile Giderilmesi

Amaç: Bu çalışmanın birinci temel amacı, Koroner arter hastalarında mortalite ve morbiditenin artma nedenlerinden biri olan hipertansiyonun, çeşitli risk faktörleri yardımıyla Tıbbi Bilgi Keşfi Süreci uygulanması suretiyle tahmin edilmesi (sınıflandırılması) dir. Çalışmada kullanılan veri setinin bağımlı değişkeni olan hipertansiyonun sınıf dengesizliği probleminin olması nedeniyle, sınıflandırma işlemi yapılmadan önce bu probleminin giderilmesi için çeşitli yaklaşımları kullanan ve ara yüzü Türkçe olan bir web-tabanlı yazılımın geliştirilmesi bu çalışmanın ikinci temel amacıdır.

Materyal ve Metot: Çalışmada kullanılan veri seti, 149'u (%16) hipertansiyonu bulunan, 780'i (%84) hipertansiyonu bulunmayan toplam 929 koroner arter hastası kayıtlarından oluşmaktadır. Koroner arter hastalarında hipertansiyonun sınıflandırılması 8 adet bağımsız değişkene dayalı olarak yapılmıştır. Dengesiz sınıf problemini gidermek için çeşitli alt örnekleme, üst örnekleme ve hem alt hem de üst örnekleme yöntemleri kullanılmıştır. Sınıflandırma yöntemleri olarak Çok Katmanlı Algılayıcı, Aşırı Öğrenme Makinesi ve Destek Vektör Makineleri modelleri uygulanmıştır.

Bulgular: En iyi sınıflandırma performansının, DBSMOTE sınıf dengeleme yöntemi uygulandıktan sonra Destek Vektör Makinesi modeli ile elde edildiği görülmüştür. İlgili modelin, doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama değerleri sırasıyla, 0.99, 0.99, 0.99, 0.95, 0.97 ve 0.97 olarak hesaplanmıştır.

Sonuç: Çalışmada uygulanan üst örnekleme yöntemlerinin, alt örnekleme yöntemlerine göre modellerin sınıflandırma performanslarına belirgin bir şekilde olumlu katkı yaptığı görülmüştür. Bu çalışma kapsamında yer almayan ancak ilerleyen çalışmalarda ele alınacak olan; hibrit yöntemleri, Maliyet-Duyarlı Öğrenme Tabanlı Yöntemler, Topluluk Öğrenme Tabanlı Yöntemler, Öznitelik Seçimi Tabanlı Yöntemler, daha sağlam ve tutarlı sonuçlar elde edilmesi açısından okuyuculara önerilebilir.

Anahtar Kelimeler: Dengesiz sınıf problemi, alt-üst örnekleme yöntemleri, hipertansiyon, tıbbi bilgi keşfi süreci, koroner arter hastalığı.

ABSTRACT

Handling Imbalanced Class Problem for the Classification of Hypertension in the Coronary Artery Disease Patients by Using Medical Knowledge Discovery Process

Aim: The primary aim of this study is to estimate (classify) hypertension, one of the causes of mortality and morbidity increase in coronary artery disease patients, by applying Medical Knowledge Discovery Process with various risk factors. Because of the class imbalance problem of hypertension, which is dependent variable of the dataset used in the study, the development of a web-based software which uses various approaches to resolve this problem before the classification process and whose interface is Turkish is the second main aim of this study.

Material and Method: The dataset used in the study consisted of records of 929 coronary artery patients with 149 (16%) hypertension and 780 (84%) non-hypertension. Classification of hypertension in coronary artery patients was done based on 8 independent variables. Various over-under sampling and both over and under sampling methods was used to handle the imbalanced class problem. As the classification methods, Multilayer Perceptron, Extreme Learning Machine and Support Vector Machine models were performed.

Results: The best classification performance was obtained by the Support Vector Machine model after applying the DBSMOTE class balancing method. The accuracy, sensitivity, specificity, precision, f-measure and g-mean metrics of the relevant model were calculated as 0.99, 0.99, 0.99, 0.95, 0.97 and 0.97, respectively.

Conclusion: Compared to the undersampling methods, the oversampling methods used in the study showed a positive contribution to the classification performance of the models. Hybrid Methods, Cost-Sensitive Learning Based Methods, Ensemble Learning Based Methods, Feature Selection Based Methods, which aren't included in the scope of this study but will be discussed in further studies, can be suggested to readers for more robust and consistent results.

Key Words: Imbalanced class problem, over and under sampling methods, hypertension, medical knowledge discovery process, coronary artery disease.

SİMGELER VE KISALTMALAR DİZİNİ

KAH	: Koroner arter hastalığı
HT	: Hipertansiyon
VTBK	: Veri tabanlarında bilgi keşfi süreci
TBKS	: Tıbbi bilgi keşfi süreci
RAÖ	: Rastgele alt örnekleme
CNN	: Condensed nearest neighbor
OSS	: One-sided selection
ENN	: Edited nearest neighbor
NCR	: Neighborhood cleaning rule
RÜÖ	: Rastgele üst örnekleme
SMOTE	: Synthetic minority over-sampling technique
BLSMOTE	: Borderline synthetic minority over-sampling technique
SLSMOTE	: Safe-level synthetic minority over-sampling technique
DBSMOTE	: Density-based synthetic minority over-sampling technique
ADASYN	: Adaptive synthetic sampling
PREE	: Prediction risk based feature selection for easy ensemble
ROC	: Receiver operating characteristic
PR	: Precision-Recall
AUC	: Area under curve
kNN	: k-en yakın komşu algoritması
ÇKA	: Çok katmanlı algılayıcı
AÖM	: Aşırı öğrenme makinesi
DVM	: Destek vektör makinesi

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

TABLOLAR DİZİNİ

Tablo No **Sayfa No**

1. GİRİŞ

Koroner arter hastalığı (KAH) gelişmiş ve gelişmekte olan ülkeler için önemli halk sağlığı problemleri arasında bulunmaktadır. Ayrıca, KAH dünya genelinde mortalite ile morbiditenin başlıca nedenlerindendir. KAH; cinsiyet, yaş, sigara, hipertansiyon, diyabetes mellitus ve düşük yoğunluklu lipoprotein (LDL) gibi pek çok geleneksel ve yeni risk etkenleri kullanılarak tahmin edilebilmektedir (1-4).

KAH teşhisi konulmuş veya gelişme potansiyeli yüksek bireylerin tedavilerinin önemli aşamalarından biri, hastalığın risk faktörlerinin belirlenmesidir (1). Risk faktörlerinin doğru şekilde belirlenmesi, KAH'ın oluşumunu engellemek veya gelişimini kontrol altında tutarak mortalite ve morbiditenin azaltılmasında önemli rol oynamaktadır.

Hipertansiyon (HT), tüm yaş / ırk / cinsiyet gruplarında KAH için bağımsız önemli bir risk faktörüdür. HT, ateroskleroz gelişimini hızlandırır, sürekli kan basıncının yükselmesi vasküler lezyonları dengesiz hale getirir ve akut koroner olayları hızlandırır (2, 3). Ayrıca kardiyovasküler morbidite ve mortaliteyi ciddi şekilde arttırmaktadır (4).

KAH'lı hastalarda HT'nin tedavisi önem taşımaktadır. KAH'lı hastalarda HT tedavisinde, kan basıncını düşürmek, iskemiyi azaltmak ve kardiyovasküler olayları ve ölümü önlemek amaçlanır. Bu hedeflere ulaşmak için hem farmakolojik olmayan müdahaleler hem de farmakolojik tedaviler önerilmektedir (2).

Gelişen teknoloji ile paralel olarak üretilen veri sayısı her geçen gün artmaktadır. Üretilen muazzam büyüklükteki verinin depolanması, yönetilmesi ve kullanışlı hale getirilmesi büyük önem arz etmektedir. Verinin kullanışlı hale getirilmesi ile ilgili olarak; yığın veri içindeki gizli kalmış ve faydalı olma ihtimali yüksek bilgilerin çıkarılması, ileriye yönelik çıkarsama ve karar verme aşamalarında kullanılmasına yönelik süreçler ortaya konulmuştur. Bunlardan Veri Tabanlarında Bilgi Keşfi (VTBK) süreci, araştırmada kullanılacak ham veri setinin seçilmesi, seçilen veri seti üzerinde önişleme ve dönüşüm aşamalarının gerçekleştirilmesi, veri seti üzerinde veri madenciliği/makine öğrenmesi modellerinin uygulanması (modelleme) ve modelleme çıktılarının çeşitli performans ölçütleri ile değerlendirilmesi ve yorumlanması aşamalarını kapsar (5).

Günümüzde; sağlık, bankacılık, telekomünikasyon ve birçok alanda üretilen yığın veri setlerindeki örüntülerin veri madenciliği/makine öğrenmesi teknikleri ile çıkarılarak,

tahmin ve karar destek bileşenleri olarak kullanılması büyük önem taşımaktadır. Bu kapsamda sıklıkla uygulanan veri madenciliği/makine öğrenmesi konularından biri sınıflandırmadır. Sınıflandırma, veri setini oluşturan gözlemlerin, daha önce belirlenmiş sınıflara, belirli kurallar çerçevesinde atamalarını gerçekleştiren bir tahmin sürecidir. Günümüzde özellikle sağlık alanında, hastalıkların, bu hastalıklarla ilişkisi olduğu düşünülen risk faktörleri yardımıyla veri madenciliği yöntemleri kullanılarak sınıflandırılması ve bu risk faktörlerinin ilgili hastalığa katkı düzeylerinin belirlenmesi yaygın hale gelmiştir. Fakat klasik istatistiksel testlerde olduğu gibi varsayımlara dayanmayan ve çıktılarının yorumlanması kolay olduğu için rağbet gören geleneksel sınıflandırma algoritmaları bazı durumlarda hatalı ve yanlı sonuçlar verebilmektedir. Bu durumun başlıca nedenlerinden biri sınıflandırmada referans alınan bağımlı (hedef) değişkeni oluşturan sınıflar arasında sayıca çarpık bir dağılımın varlığıdır. Literatürde dengesiz sınıf problemi olarak adlandırılan bu durumun varlığında geleneksel sınıflandırma algoritmaları çoğunluk sınıf tarafından etkilenmekte ve böylece sonuçlarda yanlılık (bias) problemi ortaya çıkmaktadır (6). Bu problemin etkisinin azaltılması/giderilmesi günümüze revaçta olan bir konu olup, bu konu ile ilgili birçok çalışma yapılmış ve hala da yapılmaya ve yeni yaklaşımlar ortaya konulmaya devam edilmektedir.

Bu çalışmanın birinci temel amacı, KAH'lı hastalarda mortalite ve morbiditenin artma nedenlerinden biri olan HT'nin çeşitli risk faktörleri yardımıyla VTBK süreçlerinin tıp alanında uygulaması olan Tıbbi Bilgi Keşfi Süreci (TBKS) nin uygulanması suretiyle tahmin edilmesi (sınıflandırılması) dir. Çalışmada kullanılan veri setinin bağımlı değişkeni olan HT'de sınıf dengesizliği probleminin olması nedeniyle, sınıflandırma işlemi yapılmadan önce sınıf dengesizliği probleminin giderilmesi için çeşitli yaklaşımları kullanan ve ara yüzü Türkçe olan bir web-tabanlı yazılımın geliştirilmesi bu çalışmanın ikinci temel amacıdır.

2. GENEL BİLGİLER

2.1. Veri Tabanlarında Bilgi Keşfi Süreci (VTBK) ve Tıbbi Bilgi Keşfi Süreci (TBKS)

Veri tabanlarında bilgi keşfi (VTBK), üzerinde çalışılacak veri setinde saklı olarak bulunan bilgi, örüntü ve özellikleri ortaya çıkarmak için uygulanan ve 5 adımdan oluşan bir süreçtir. VTBK süreci, analize konu veri setinin ilgili veri tabanlarından temin edilmesi, veri önişlemesi (pre-processing), veri dönüşümü (transformation) ve indirgemesi (feature selection/reduction), veri madenciliği (data mining) ve veri madenciliği çıktılarının değerlendirilmesi ve yorumlanması (evaluation and interpretation) safhalarından oluşur (5). VTBK süreçlerinin tıp alanında kullanılması Tıbbi Bilgi Keşfi Süreci (TBKS) olarak adlandırılmaktadır (7). VTBK süreci Şekil 2.1'de gösterilmiştir.

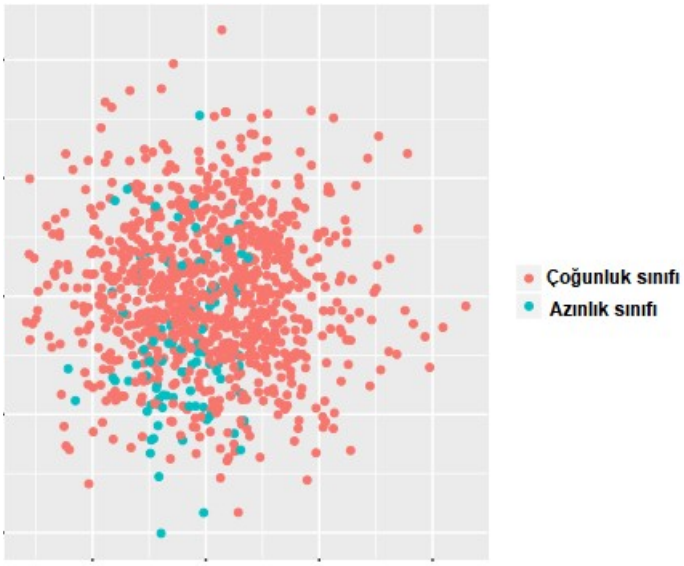
Şekil 2.1: VTBK süreci

VTBK sürecinin ilk aşamasında, analize konu veri setinin ilgili veri tabanlarından temin edilir. Veri Önişleme aşamasında, veri setindeki kayıp değerlere çeşitli yöntemler kullanılarak değer ataması yapılır, çeşitli istatistik ve/veya makine öğrenmesi teknikleri kullanılarak aşırı/aykırı değerlerin tespiti ve veri setinden çıkarılması işlemi gerçekleştirilir. Veri Dönüşümü aşamasında veri setine standardizasyon (z-dönüşüm),

normalizasyon, Box-Cox vb. dönüşümler uygulanarak veri setini oluşturulan gözlemler belirli bir standarda sokulur. Veri İndirgeme aşamasında, uygulanacak sınıflandırma veya regresyon algoritmalarına bağlı (wrapper) veya bunlardan bağımsız (filter) yaklaşımlar kullanılarak veri setindeki değişken sayısında indirgeme yapılır. Veri Madenciliği aşamasında ise çeşitli sınıflandırma, regresyon ve kümeleme analizleri kullanılarak veri setinde bilgi ve örüntü çıkarma işlemi yapılır. Yorumlama ve Değerlendirme aşamasında ise veri madenciliği çıktıları yorumlanır ve modelin performansı ölçülür.

2.2. Dengesiz Sınıf Problemi ve Dengesiz Sınıflarda Öğrenme

Dengesiz sınıf problemi, makine öğrenmesi alanındaki önemli konulardan biridir. Üzerinde çalışılan veri setinde, gözlemlerin oluşturduğu sınıflardan birinin diğer gözlemlerin oluşturduğu sınıf veya sınıflara göre sayıca yüksek derecede fazla olması dengesiz sınıf problemini ortaya çıkarır. Sınıflar arası dağılımda dengesizlik bulunan iki sınıflı bir veri setinin grafiksel gösterimi Şekil 2.2'de verilmiştir. Örneğin, vaka-kontrol araştırmalarında kullanılan veri setlerinde hasta sınıfına ilişkin gözlem sayısı, hasta olmayan sınıfa göre azdır (özellikle nadir görülen hastalıklarda). Bu durumda hasta ve hasta olmayan bireylerin sınıflandırılmasında yanlılık (bias) sorunu ortaya çıkmaktadır. Çünkü sınıflandırma problemlerinde kullanılan ve dengesiz sınıf dağılımlarına karşı duyarlı olan temel makine öğrenmesi modelleri büyük sınıf(lar)ın etkisi altında kalmaktadır ve küçük sınıfların varlığı ortadan kaybolmaktadır (8-10).



Şekil 2.2: Dengesiz sınıf dağılımı gösteren bir veri seti

Literatürde bu problemin çözümüne yönelik birçok yaklaşım önerilmiştir. Genel olarak bu yaklaşımlar 4 ana başlık altında incelenir. Bu başlıklar ve bu başlıklar altında kullanılan yaklaşımlar Tablo 2.1'de verilmiştir (11).

Tablo 2.1: Dengesiz sınıf probleminin çözümü için geliştirilen yaklaşımlar (11)

Örnekleme Tabanlı Yöntemler	**Topluluk Öğrenme (Ensemble Learning) Tabanlı Yöntemler**
1. Temel (Basit) Örnekleme Yöntemleri a) Rastgele Alt Örnekleme b) Rastgele Üst Örnekleme **2. İleri Alt Örnekleme Yöntemleri** a) Tomek Link b) Condensed Nearest Neighbor (CNN) c) Neighborhood Cleaning Rule (NCL) d) One-Sided Selection (OSS) e) Edited Nearest Neighbor Rule (ENN) f) Neighborhood Cleaning Rule (NCR) **3. İleri Üst Örnekleme Yöntemleri** a) Synthetic Minority Over-sampling Technique (SMOTE) b) Borderline-SMOTE c) Safe-level SMOTE d) Density-based SMOTE (DBSMOTE) e) Adaptive Synthetic Sampling (ADASYN) **4. Hibrit Örnekleme (Hem alt hem üst örnekleme) Yöntemleri**	**1. Bagging** a) Asymmetric bagging, SMOTE Bagging b) Over Bagging, Under Bagging c) Roughly balanced bagging , Lazy Bagging d) Random features selection **2. Boosting** a) Adaboost, SMOTEBoost, DataBoost-IM **3. Rastgele Orman (Random Forest)** a) Dengelenmiş Rastgele Orman (Balanced Random Forest) b) Ağırlıklandırılmış Rastgele Orman (Weighted Random Forest)
Maliyet-Duyarlı Öğrenme (Cost-Sensitive Learning) Tabanlı Yöntemler	**Öznitelik Seçimi (Feature Selection) Tabanlı Yöntemler**
a) Direkt Maliyet-Duyarlı Öğrenme b) Maliyet-Duyarlı Meta Öğrenme c) Maliyet-Duyarlı Meta Öğrenme Eşikleme Yöntemleri (MetCost) d) Maliyet Duyarlı Meta Öğrenme Örnekleme Yöntemleri	a) Wrapper b) PREE (Prediction Risk based feature selection for Easy Ensemble)

2.2.1. Örnekleme Tabanlı Yöntemler

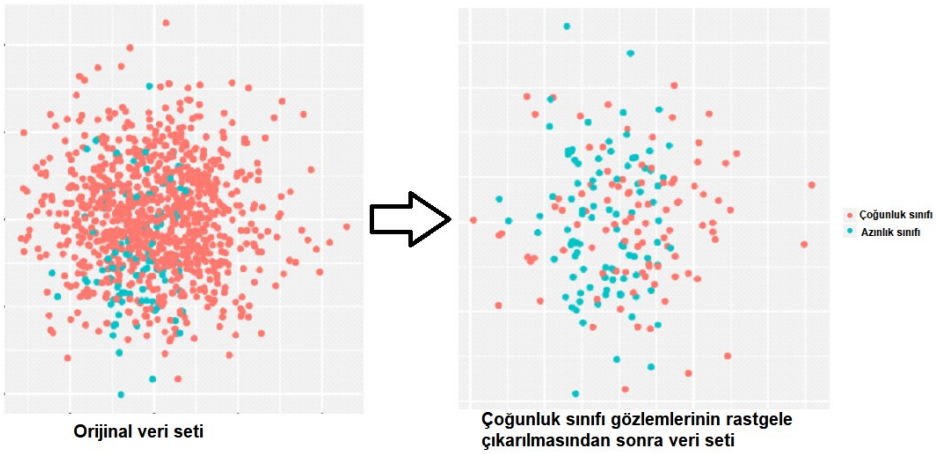
Örnekleme tabanlı yöntemler, sınıf dengesizliği bulunan verilerle başa çıkmada yaygınlıkla kullanılan yaklaşımlardır. Temel fikir, sınıflar arasındaki farkları en aza indirgemek için eğitim (training) veri setini ön işlemeye tabi tutmaktır. Başka bir deyişle, örnekleme tabanlı yöntemler, her sınıfta daha dengeli sayıda gözlem elde etmek için eğitim veri setindeki azınlık ve/veya çoğunluk sınıfının dağılımını değiştirir (11, 12). Örnekleme tabanlı yöntemler alt ve üst örnekleme yöntemleri olmak üzere 2 kısımda incelenirler. Azınlık sınıf gözlemlerinden örnekleme yöntemleri yardımıyla sentetik veriler türetilerek çoğunluk sınıfın sayısına eşit veya hemen hemen eşit duruma getirilir. Azınlık sınıf kullanılarak yapılan bu işleme üst örnekleme (oversampling) denir. Diğer yandan çoğunluk sınıf verilerinin sayısı alt örnekleme (undersampling) yöntemi kullanılarak azınlık sınıfın sayısına çekilebilmekte ve veri seti dengeli hale getirilebilmektedir. Ayrıca veri setini oluşturan sınıfların dengelenmesi için üst örnekleme ve alt örnekleme yöntemleri aynı anda da kullanılabilmektedir (hibrit örnekleme yöntemleri).

2.2.1.1. Alt Örnekleme Yöntemleri

Alt örnekleme yöntemleri, dengesiz sınıf probleminin giderilmesinde kullanılan en yaygın ve basit stratejilerden biridir. Çoğunluk sınıfa ait gözlemler veri setinden rastgele veya analitik bir yöntem vasıtasıyla çıkarılarak sınıflar arası denge sağlanmaya çalışılır.

2.2.1.1.1. Rastgele Alt Örnekleme (RAÖ)

Bu yöntemde çoğunluk sınıfına ait gözlemler, azınlık sınıfına ait gözlem sayısına eşit olacak şekilde rastgele olarak seçilerek veri setinden çıkarılır ve veri seti dengeli hale getirilir. Şekil 2.3'de RAÖ yöntemi ile dengeli hale getirilen veri setinin temel bileşen grafikleri verilmiştir.



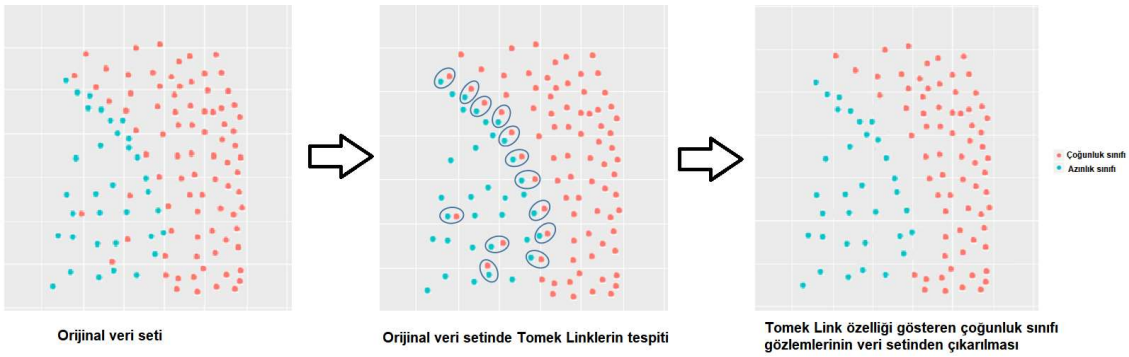
Şekil 2.3: Rastgele Alt Örnekleme ile veri setinin dengeli hale getirilmesi

2.2.1.1.2. Tomek Link

a ve b biri azınlık diğeri çoğunluk sınıfından olmak üzere iki gözlem olsun. $d(\mathrm{a},\mathrm{b})$, a ile b arasındaki uzaklığı tanımlayan bir metrik (ör. Öklidyen) olmak üzere, eğer;

$$d(\mathrm{a},\mathrm{c}) < d(\mathrm{a},\mathrm{b}) \text{ veya } d(\mathrm{b},\mathrm{c}) < d(\mathrm{a},\mathrm{c})$$

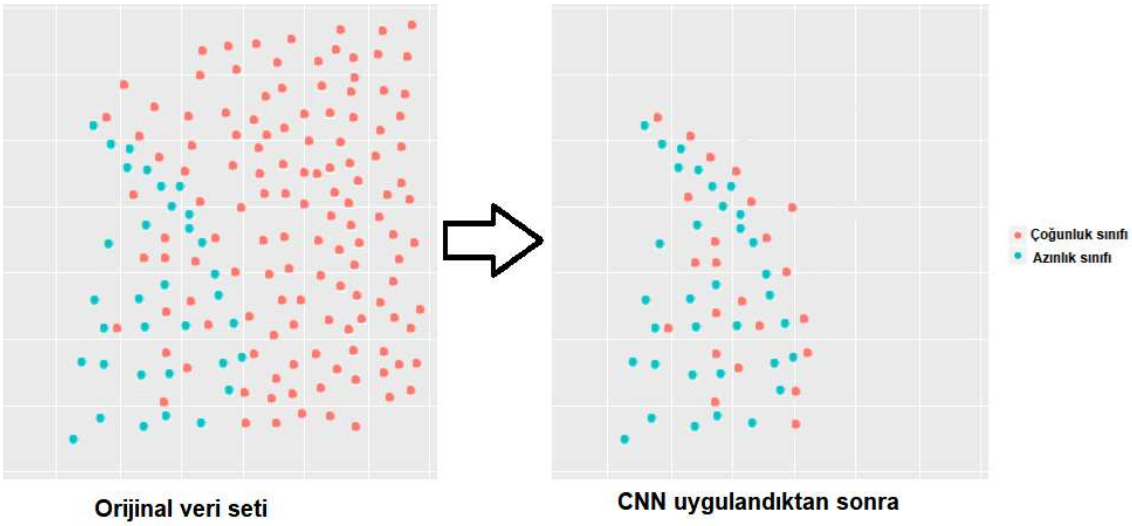
olacak şekilde bir c gözlemi yoksa, a ve b gözlem çifti bir Tomek link belirtir. Daha açık ifade etmek gerekirse, a ve b gözlem çiftinin bir Tomek Link belirtmesi için, a gözleminin en yakın komşusunun b, b gözleminin en yakın komşusunun ise a olması gerekir. Bu tanımlamalardan a ve b gözlemlerinin sınır (borderline) gözlemleri veya gürültülü (noisy) gözlemler olduğu görülebilir. Bunun nedeni, yalnızca sınır gözlemlerinin veya gürültülü gözlemlerin, karşı sınıftan en yakın komşulara sahip olmalarıdır. Üzerinde çalışılan veri setinde, yukarıdaki ölçütlere uyan gözlem çiftleri tespit edilerek Tomek Link belirten çoğunluk sınıfına ait gözlem veri setinden kaldırabilir (12, 13). Şekil 2.4'de sınıf dengesizliği olan veri setinde öncelikle Tomek Link belirten gözlem çiftlerinin tespiti ve bu gözlem çiftlerinden çoğunluk sınıfı gözlemi olanların veri setinden çıkarılması durumu temel bileşen grafikleri ile gösterilmiştir.



Şekil 2.4: Tomek Link yaklaşımı

2.2.1.1.3. Condensed Nearest Neighbor (CNN)

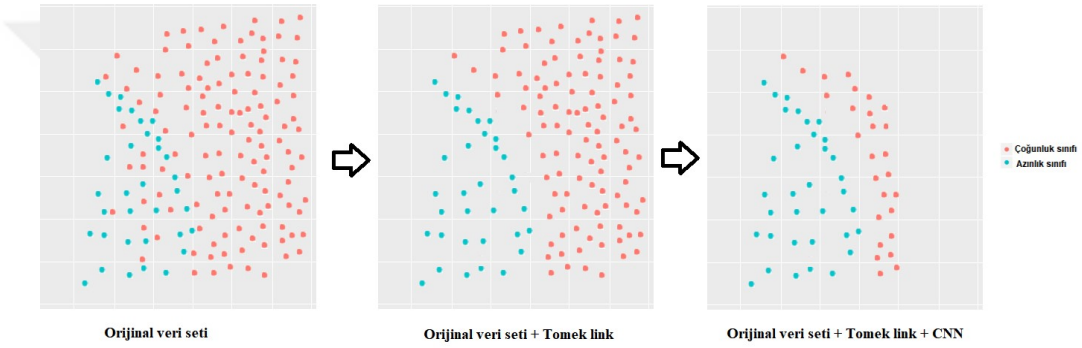
Condensed Nearest Neighbor (CNN) (14) yaklaşımında amaç, sınıf dengesizliği bulunan bir veri setinde, çoğunluk sınıfından, karar sınırından uzak olan ve makine öğrenmesi modelinin öğrenme (training) süreci için gereksiz olduğu düşünülen (redundant) gözlemleri çıkarmaktır (15). Bunu gerçekleştirmek için 1-En Yakın Komşu yöntemi kullanılarak orijinal veri setinin tutarlı bir alt kümesi oluşturulur. Şekil 2.5'de sınıf dengesizliği olan veri setine önce CNN algoritması uygulanmadan önceki ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



Şekil 2.5: CNN yaklaşımı

2.2.1.1.4. One-Sided Selection (OSS)

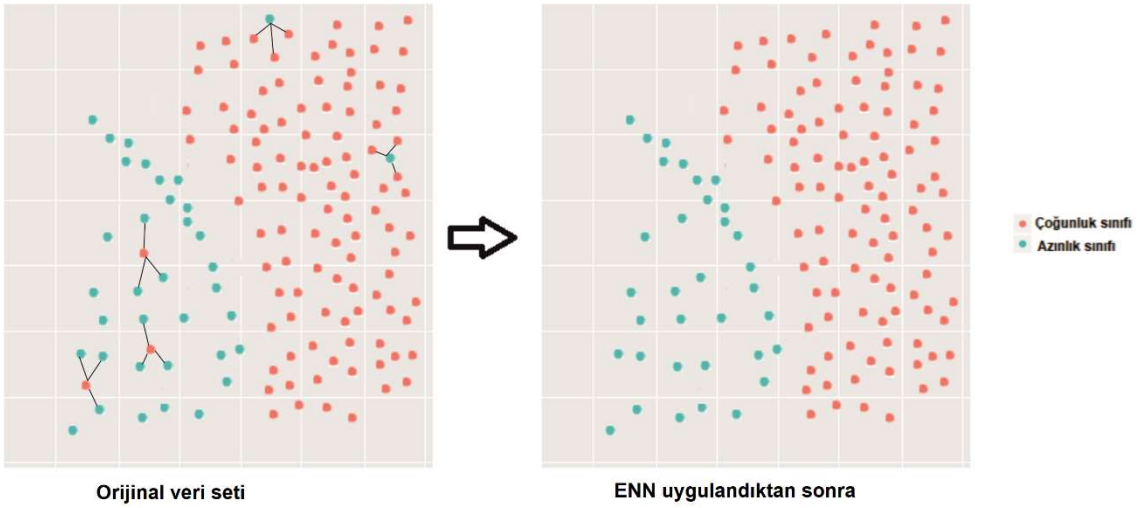
One-Sided Selection (OSS) (16), Tomek Link ve CNN yöntemlerinin sırayla uygulanmasından elde edilen bir alt örnekleme yöntemidir. Tomek Link yöntemi ile öncelikle çoğunluk sınıfına ait gürültülü ve sınır gözlemleri veri setinden çıkarılır. Ardından CNN yöntemi uygulanarak makine öğrenmesi modelinin öğrenme süreci için gereksiz olduğu düşünülen çoğunluk sınıfına ait gözlemleri veri setinden çıkarır (15, 17). Şekil 2.6'de sınıf dengesizliği olan veri setine önce Tomek Link, daha sonra hem Tomek Link hem de CNN algoritmaları uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



Şekil 2.6: OSS yaklaşımı uygulanarak veri setinin dengeli hale getirilmesi

2.2.1.1.5. Edited Nearest Neighbor (ENN)

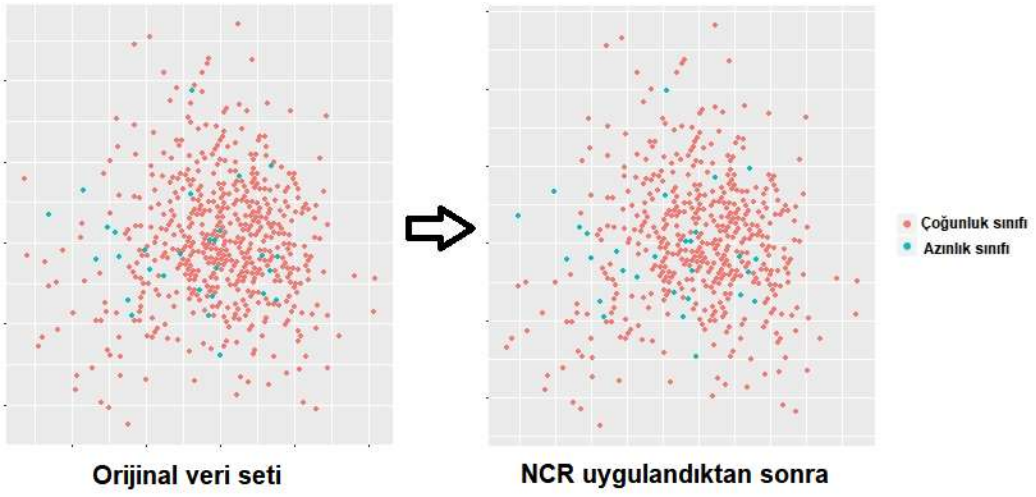
Edited Nearest Neighbor (ENN) (18) yönteminde, bir k-en yakın komşu sınıflandırıcısı tarafından yanlış sınıflandırılmış tüm gözlemler silinir. Burada k-parametresi genellikle k=3 olarak seçilir (19). ENN algoritması hem çoğunluk hem de azınlık sınıfından gözlemleri veri setinden çıkarabilmektedir. Şekil 2.7'de sınıf dengesizliği olan veri setine ENN algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



Şekil 2.7: Sınıf dengesizliği olan veri setine ENN algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri

2.2.1.1.6. Neighborhood Cleaning Rule (NCR)

Neighborhood Cleaning Rule (NCR) (20), veri kümesinden kaldırılacak çoğunluk sınıf örneklerini seçmek için yukarıda bahsedilen ENN (18) yöntemini kullanır. NCR'de, veri kümesindeki her bir a gözlemi için, en yakın üç komşu gözlemi tespit edilir. Eğer a bir çoğunluk örneği ise ve üç en yakın komşusu tarafından yanlış sınıflandırılmışsa, veri setinden çıkarılır. Alternatif olarak, eğer bir azınlık sınıfında bir örnek ise ve en yakın üç komşusu tarafından yanlış sınıflandırılmışsa, o zaman komşular arasındaki çoğunluk sınıfı örnekleri kaldırılır. Şekil 2.8'de bir veri setinin NCR yöntemi uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



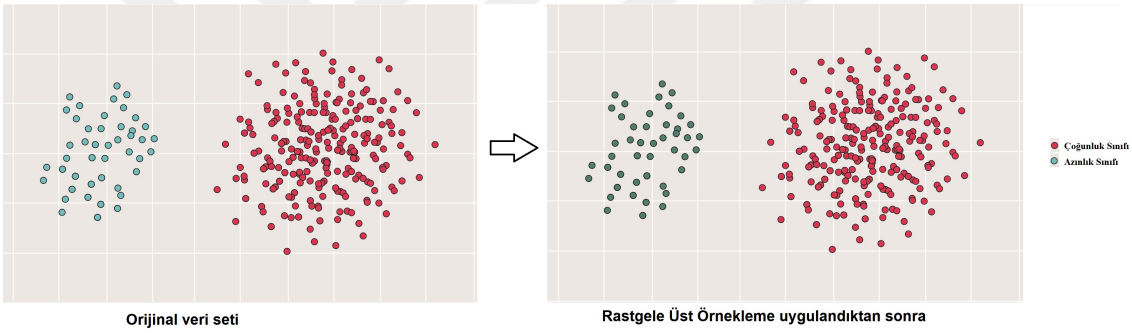
Şekil 2.8: Bir veri setinin NCR yöntemi uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri

2.2.1.2. Üst Örnekleme Yöntemleri

Üst örnekleme yöntemlerinde, alt örneklem yöntemlerinden farklı olarak, azınlık sınıfa ait gözlemlerin sayısı, rastgele veya analitik bir yöntem vasıtasıyla çoğunluk sınıf gözlem sayısına eşit veya yakın bir sayıya çıkarılarak sınıflar arası denge sağlanmaya çalışılır.

2.2.1.2.1. Rastgele Üst Örnekleme (RÜÖ)

Bu yöntemde azınlık sınıfına ait bazı gözlemler rastgele seçilip tekrar edilerek çoğaltılır ve bu işlem çoğunluk sınıfına ait gözlem sayısına eşit olana dek devam eder. Sonuç olarak iki sınıfta da gözlem sayıları eşit olan yeni bir veri seti elde edilir. Şekil 2.9'de sınıf dengesizliği olan veri setine RÜÖ uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



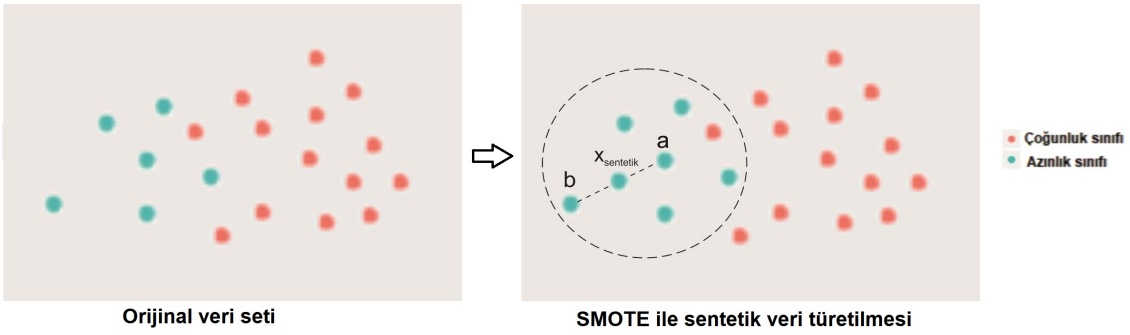
Şekil 2.9: Sınıf dengesizliği olan veri setine Rastgele Üst Örnekleme uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri

2.2.1.2.2. Synthetic Minority Over-sampling Technique (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) (21), farklı disiplinlerde kendisine uygulama alanı bulan güçlü bir algoritmadır. SMOTE algoritması mevcut azınlık gözlemleri arasındaki özellik uzayı (feature space) benzerliklerine dayalı sentetik/yapay veriler oluşturur (22). Yeni sentetik azınlık sınıfı gözlemlerini oluşturmak için, SMOTE önce bir azınlık sınıf gözlemini rastgele seçer (a) ve onun k-en yakın azınlık sınıfı komşularını bulur. Daha sonra k-en yakın komşu elemanlarından biri (b) rastgele seçilir ve sentetik gözlem, özellik uzayında a'yı b'ye bağlayan bir doğru parçası oluşturarak türetilir. Sentetik gözlemler, seçilen iki a ve b gözleminin konveks birleşimi olarak oluşturulur (12). Sentetik gözlemin türetilmesinde

$$x_{sentetik} = a + (b - a).\delta$$

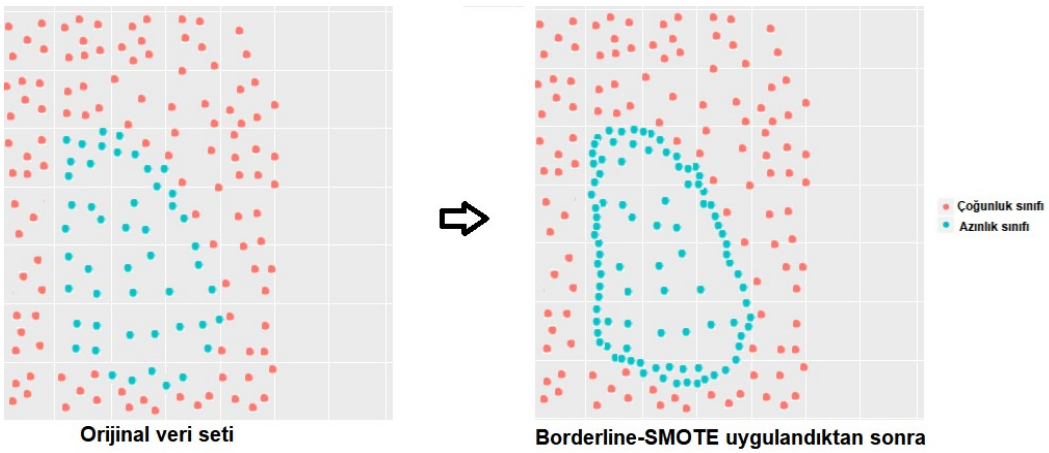
formülü kullanılır (23). Burada δ, [0,1] aralığında rastgele seçilen bir sayıdır. Şekil 2.10'da SMOTE algoritmasının çalışma prensibi 5-en yakın komşuluk algoritması kullanılarak gösterilmiştir.



Şekil 2.10: SMOTE algoritmasının çalışma prensibi

2.2.1.2.3. Borderline-SMOTE

Borderline-SMOTE (24), SMOTE tarafından türetilen gözlemlerin kalitesini doğrudan geliştirmeye çalışan algoritmalardan biridir (25). Bu algoritma, sınırda ve sınır yakınlarında yer alan azınlık sınıfı gözlemlerinin SMOTE algoritması kullanılarak çoğaltılması mantığına dayanır. Şekil 2.11'de sınıf dengesizliği olan veri setine Borderline-SMOTE algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



Şekil 2.11: Sınıf dengesizliği olan veri setine Borderline-SMOTE algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri

2.2.1.2.4. Safe-level SMOTE

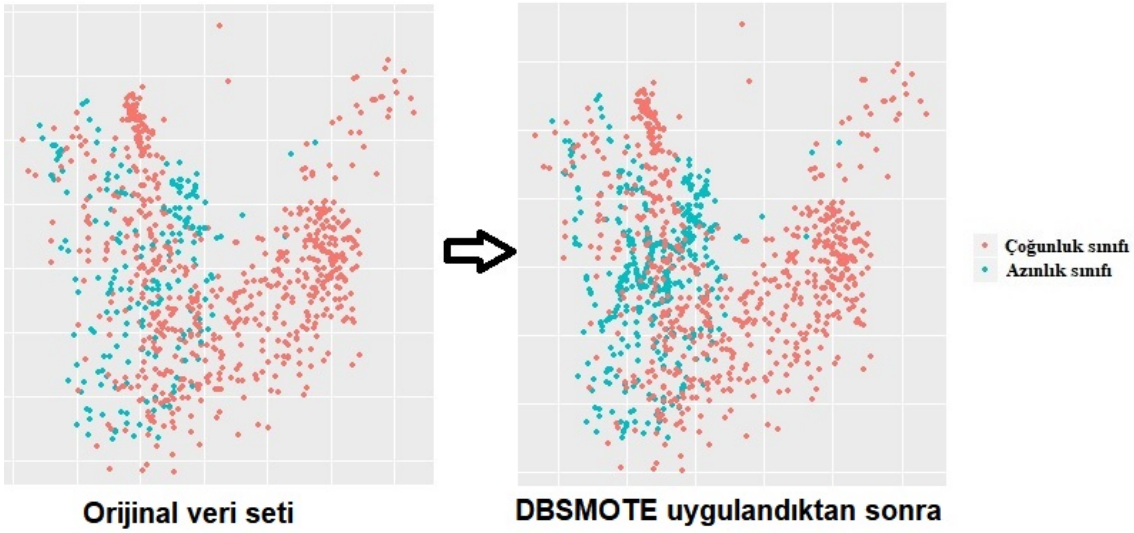
Safe-level SMOTE (26), her bir azınlık sınıfı gözlemine, sentetik gözlemler türetmeden önce bir güven seviyesi atar. Sentetik gözlemler SMOTE ile türetilirken, bu gözlemler en güvenli düzeye sahip (sınırda bulunan azınlık sınıfı gözlemlerine uzak) olan gözleme daha yakın konumlandırılır, böylece sentetik gözlemler yalnızca güvenli bölgelerde oluşturulur (25). Şekil 2.12'de bu algoritmanın çalışması görselleştirilmiştir.

Şekil 2.12: Safe-level SMOTE algoritmasının çalışma prensibi

2.2.1.2.5. Density-based SMOTE (DBSMOTE)

Density-based SMOTE (DBSMOTE) (27) algoritması yoğunluk temelli bir küme kavramına dayanır. Bu algoritma, pozitif (azınlık) sınıf kümesine (cluster) DBSCAN (Density-Based Spatial Clustering of Applications with Noise) (28) algoritmasını uygular. DBSCAN tarafından keşfedilen, keyfi biçimde şekillendirilmiş bir kümenin üst örneklenmesi için tasarlanmıştır. DBSMOTE, her azınlık sınıfı gözleminden, azınlık sınıfı kümesinin sahte-merkez (pseudo-centroid) noktasına kadar en kısa yol boyunca sentetik gözlemler türetir. Bu durum, elde edilen sentetik veri kümesinin, orijinal pozitif gözlem grubunun merkezinin çevresinde yoğun olmasını sağlar (29).

Şekil 2.13'de sınıf dengesizliği olan veri setine DBSMOTE algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



Şekil 2.13: Sınıf dengesizliği olan veri setine DBSMOTE algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri

2.2.1.2.6. Adaptive Synthetic Sampling (ADASYN)

Adaptive Synthetic Sampling (ADASYN) (30) algoritmasının amacı, varolan azınlık sınıfı gözlemleri arasında doğrusal interpolasyon yoluyla azınlık sınıfından yeni gözlemler türetmek suretiyle sınıf dengesi geliştirmektir. ADASYN, SMOTE algoritmasının bir uzantısıdır ve sentetik gözlemleri azınlık sınıfının içinden ziyade azınlık ve çoğunluk sınıfların sınır çevresinde oluşturmaktadır (31). Çalışma prensibi şöyledir:

$S_{azınlık}$ ve $S_{çoğunluk}$, azınlık ve çoğunluk sınıflarına ait veri setlerini ifade etmek üzere, öncelikle sınıf dengesinin sağlanması için türetilmesi gereken sentetik azınlık sınıf gözlem sayısı hesaplanır,

$$G = \left(\left|S_{çoğunluk}\right| - \left|S_{azınlık}\right|\right).\beta$$

burada $\beta \in [0,1]$, sentetik gözlem oluşturma sürecinden sonra sınıflar arası istenilen denge düzeyini belirlemeye yönelik bir parametredir. Her bir $x_i \in S_{azınlık}$ için, öklidyen uzaklık metriğine göre k-en yakın komşular tespit edilir ve aşağıda tanımlanan Γ_i oranı hesaplanır:

$$\Gamma_i = \frac{\triangle_i / k}{Z}, i = 1, \ldots, \left|S_{azınlık}\right|$$

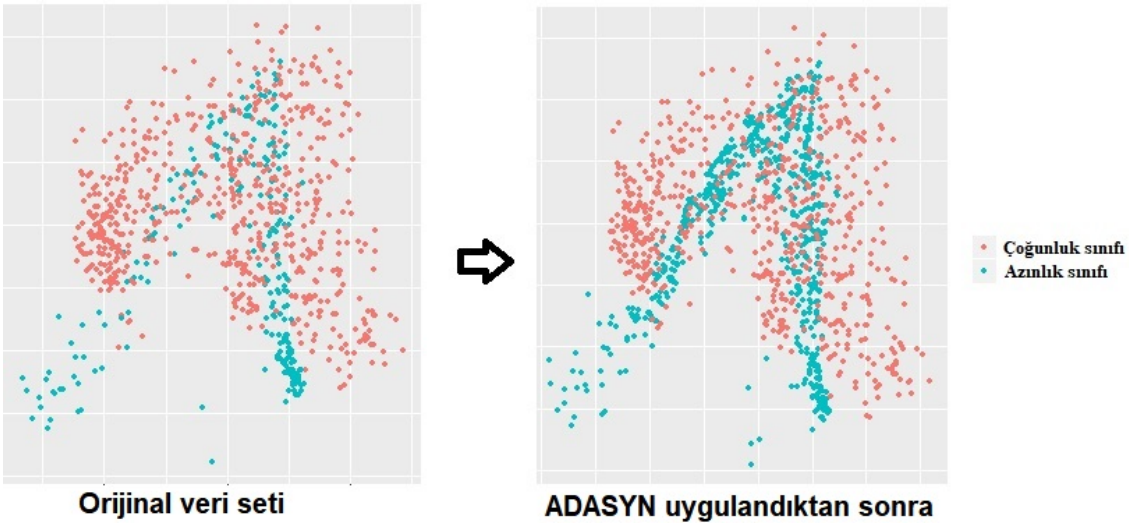
burada $\triangle_i$, x_i gözleminin k-en yakın komşuları içerisinde çoğunluk sınıfı gözlemi sayısı, Z ise normalizasyon sabitidir. Böylece Γ_i, $\sum \Gamma_i = 1$ olan bir kesikli dağılım fonksiyonu olur. Daha sonra her bir $x_i \in S_{azınlık}$ için türetilecek sentetik gözlem sayısı belirlenir:

$$g_i = \Gamma_i . G$$

Son olarak her bir $x_i \in S_{azınlık}$ için türetilecek g_i adet sentetik gözlem, SMOTE algoritmasında sentetik gözlem türetmek için kullanılan,

$$x_{sentetik} = a + (b - a) . \delta$$

formülü kullanılarak oluşturulur (12, 30). Şekil 2.14'de sınıf dengesizliği olan veri setine ADASYN algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri verilmiştir.



Şekil 2.14: Sınıf dengesizliği olan veri setine ADASYN algoritması uygulanmadan önce ve uygulandıktan sonraki temel bileşen grafikleri

2.2.1.3. Hibrit Örnekleme Yöntemleri

Alt örnekleme ve üst örneklemeye yöntemlerine ek olarak, her ikisinin bir kombinasyonunu gerçekleştiren teknikler geliştirilmiştir. Alt örnekleme yöntemlerini, üst örnekleme yöntemleri ile birleştirerek, veri setinde, hem fazla bilgi kaybı (yani çok fazla çoğunluk sınıfı gözlemini veri setinden çıkarılması) hem de aşırı uyum (overfitting) (diğer bir deyişle azınlık sınıfı gözlemlerini aşırı derecede örneklenmesi) problemleri dengelenebilir (trade-off). Rastgele alt-üst örnekleme yöntemlerinin birlikte

uygulanması, SMOTE + Tomek Link ve SMOTE + ENN geliştirilen hibrit tekniklerine örnek olarak verilebilir. Analitik olmayan rastgele alt-üst örnekleme yöntemleri, veri setinden çoğunluk sınıfı gözlemlerin çıkarılmasını rastgele olarak, azınlık sınıfı gözlemlerini ise diğer azınlık sınıfı gözlemlerinden faydalanarak rastgele yer değiştirme ile çoğaltarak sınıflar arası dengeyi sağlamaya çalışır. SMOTE + Tomek Link ve SMOTE + ENN analitik yöntemlerinde ise SMOTE, azınlık sınıfını gözlemlerine üst örnekleme uygularken, Tomek Link ve ENN çoğunluk sınıfının gözlemlerini indirgemekte kullanılır (12).

Örnekleme Tabanlı Yöntemler, dengesiz sınıf probleminin gözlendiği veri setinde çeşitli alt-üst veya her ikisinin kullanıldığı örnekleme yöntemlerini uygulayarak, sınıflandırma öncesi veri setini dengeli hale getirmeye çalışır. Bu yöntemlerde veri setini dengeleme işlemi sınıflandırma algoritmasından bağımsız olarak yapılmaktadır. Maliyet-Duyarlı Öğrenme (Cost-Sensitive Learning) Tabanlı Yöntemler, Topluluk Öğrenme (Ensemble Learning) Tabanlı Yöntemler ve Öznitelik Seçimi (Feature Selection) Tabanlı Yöntemler de ise dengesiz sınıf problemini bir bütün olarak yani sınıflandırma algoritması ile birlikte ele alır. Bunu da sadece dengesiz veri seti üzerinde oynama yaparak değil, dengesiz veri seti üzerinde alt veri seti kümeleri seçerek, alt-üst örnekleme yöntemleri ile birlikte kullanılarak veya sınıflandırma algoritması üzerinde çeşitli değişikler yaparak gerçekleştirirler. Sınıflandırma algoritmasının eğitim (training) safhası dengesiz veri seti üzerinde gerçekleştiği için bu yöntemlerde dengesiz sınıf problemi yerine dengesiz sınıflarda öğrenme problemi (imbalanced class learning problem) ifadesi kullanılmıştır.

2.2.2. Maliyet-Duyarlı Öğrenme (Cost-Sensitive Learning) Tabanlı Yöntemler

Maliyet-duyarlı öğrenme yöntemleri, belirli gözlem verilerinin yanlış sınıflandırılmasına yönelik maliyetleri tanımlar ve farklı maliyet matrisleri kullanarak dengesiz sınıflarda öğrenme problemini gidermeyi amaçlar. Geçmişte yapılan araştırmalar, maliyet-duyarlı öğrenme ve dengesiz sınıflarda öğrenme arasında güçlü bir bağlantı olduğunu göstermektedir. Genel olarak, dengesiz veriler için maliyet duyarlı öğrenmeyi uygulanmasına yönelik üç yaklaşım kategorisi vardır. Birinci kategorideki teknikler, yanlış sınıflandırma maliyetlerini, ağırlıklandırma biçimi olarak veri setine uygular. Bu teknikler, aslında yanlış sınıflandırma maliyetlerinin en iyi eğitim dağılımını seçmek için kullanıldığı maliyet duyarlı önyükleme örnekleme yaklaşımlarıdır. İkinci

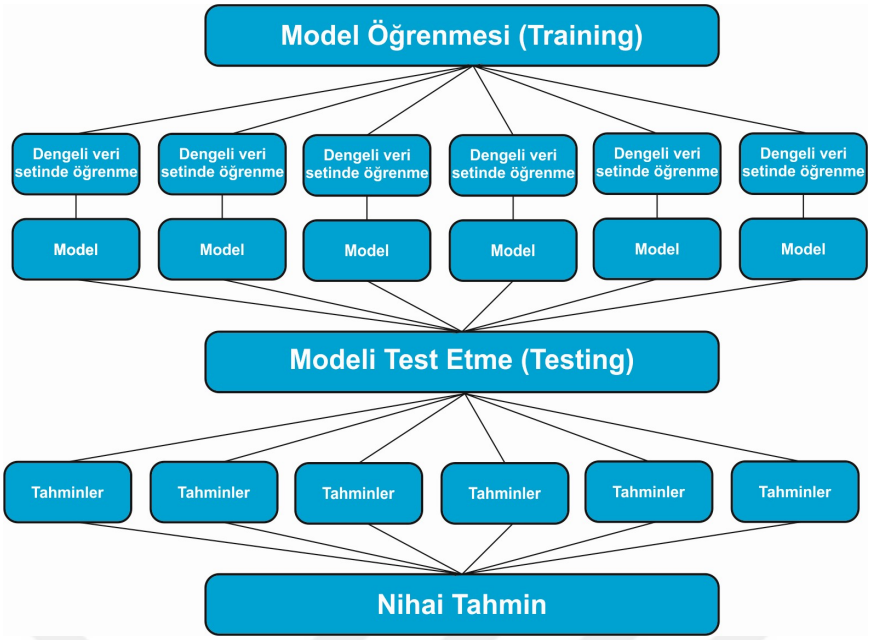
kategorideki teknikler, maliyeti minimize eden teknikleri topluluk yöntemlerinin kombinasyon şemalarına uygular. Üçüncü kategorideki teknikler, maliyet duyarlı karar ağaçları, maliyet duyarlı sinir ağları, maliyete duyarlı bayesyen sınıflandırıcıları ve maliyet duyarlı destek vektör makinelerinde (DVM) olduğu gibi, temel olarak maliyet duyarlı çerçeveyi bu sınıflandırıcılara uygun hale getirmek için, maliyet duyarlı fonksiyonları veya özellikleri doğrudan sınıflandırma algoritmalarına dâhil eder (12).

2.2.3. Topluluk Öğrenme (Ensemble Learning) Tabanlı Yöntemler

Topluluk öğrenme yöntemleri, tahmin yapmak için bir dizi sınıflayıcı kullanır. Bir topluluğun genelleme kabiliyeti genellikle topluluğun bireysel üyelerininkinden daha güçlüdür. Topluluk öğrenimi, makine öğrenmede ana öğrenme yaklaşımlarından biridir ve nesne algılama, yüz tanıma, karar destek sistemleri, tıbbi teşhis, metin sınıflandırması gibi hemen hemen her yerde öğrenme yöntemlerinin uygulanmasında büyük başarı elde etmiştir. Dengesiz sınıflarda öğrenme probleminde, mevcut yöntemleri daha da iyileştirmek veya yeni yöntemleri tasarlamak için topluluk yöntemleri yaygın olarak kullanılır. Örneğin, alt ve üst örnekleme yöntemlerini iyileştirmek için topluluk yöntemleri kullanılmış ve sınıf-dengesiz verileri işlemek için bir dizi boosting tabanlı yöntemler geliştirilmiştir (12, 22).

2.2.4. Öznitelik Seçimi (Feature Selection) Tabanlı Yöntemler

Büyük boyutlu veri kümeleri araştırması için, özellikle yüksek boyutlu veri kümeleri ile kabul edilen, uygun bir yaklaşımdır. Dengeli olmayan veri setleri bağlamında, öznitelik seçimi, sınıflar arasında daha fazla ayrılabilirliğe yol açan nitelikleri seçmek için kullanılır. Kohavi (32) tarafından önerilen Wrapper yöntemi, dengesiz veri setlerindeki ilk somut özellik seçimi uygulamalarından biridir. Bu yöntemde, öğrenme algoritması farklı alt veri seti kümelerini kullanarak veri setinden ayrı bir bölüm üzerinde yinelenerek (recursive) yürütülür; böylece en iyi performans değerlendirmesine sahip alt-sınıflandırma, tüm öğrenme seti üzerinde nihai sınıflandırıcı oluşturmak için nihai bir veri seti olarak kullanılır. PREE (Prediction Risk based feature selection for Easy Ensemble) yönteminde ise dengesiz sınıf öğrenme yaklaşımlarından olan Easy Ensemble yöntemi ile öznitelik seçimi yöntemleri birleştirilerek sınıflandırma gerçekleştirilir (11). Easy Ensemble yönteminin çalışma prensibi Şekil 2.15'de verilmiştir.



Şekil 2.15: Easy Ensemble yönteminin çalışma prensibi (15)

3. MATERYAL VE METOT

3.1. Çalışma İzni

Bu çalışma, Malatya Klinik Araştırmalar Etik Kurulu'nun 2016/162 protokol numaralı izni ile onaylanmıştır (Ek-1).

3.2. Çalışmada Kullanılan Veri Seti

Çalışmada, İnönü Üniversitesi Turgut Özal Tıp Merkezi Kalp ve Damar Cerrahisi Anabilim Dalı hasta kayıtlarından geriye dönük (retrospektif) olarak elde edilen veri seti kullanılmıştır. İlgili veri seti, 149'u (%16) hipertansiyonu bulunan, 780'i (%84) hipertansiyonu bulunmayan toplam 929 koroner arter hastası kayıtlarından oluşmaktadır. Koroner arter hastalarında hipertansiyonun sınıflandırılması aşağıda ayrıntıları belirtilen 8 adet bağımsız değişkene dayalı olarak yapılmıştır. Bu değişkenlere ilişkin tanımlayıcı tablo Tablo 3.1'de verilmiştir.

Tablo 3.1: Çalışmada kullanılan değişkenlere ilişkin tanımlayıcı tablo

Değişkenler	**Değişken Tipi**	**Açıklama**	**Değişken Rolü**
Hipertansiyon (HT)	Kategorik	Var/Yok	Bağımlı
Yaş	Nümerik	Doğal sayı	Bağımsız
Vücut Yüzey Alanı (VYA)	Nümerik	Pozitif reel sayı	Bağımsız
Vücut Kitle İndeksi (VKİ)	Nümerik	Pozitif reel sayı	Bağımsız
Kan Üre Azotu (BUN)	Nümerik	Pozitif reel sayı	Bağımsız
Kreatinin (KR)	Nümerik	Pozitif reel sayı	Bağımsız
Hemoglobin (HB)	Nümerik	Pozitif reel sayı	Bağımsız
Hematokrit (HCT)	Nümerik	Pozitif reel sayı	Bağımsız
Platelet (PLT)	Nümerik	Pozitif reel sayı	Bağımsız

3.3. Örneklem Büyüklüğü

Bazı kaynaklarda (33, 34), çok değişkenli istatistiksel modellerde bağımsız değişken sayısı 6 ya da daha büyük olduğunda, uygun örneklem büyüklüğünün belirlenmesinde,

$$n > 104 + k\ (k\text{: bağımsız değişken sayısı})$$

şartının kullanılabileceği belirtilmiştir. Çalışmada kullanılan veri setinde, 149'u HT'li, 780'i HT'li olmayan toplam 929 koroner arter hastası kayıtlarından oluşmaktadır. k=8 olduğu için, hem HT'li olan ve olmayan hasta sayısında, hem de toplam hasta sayısında yukarıda belirtilen şartın sağlandığı görülmüştür.

3.4. Geliştirilen Web-Tabanlı Yazılım

Literatürde dengesiz sınıf probleminin çözümüne yönelik birçok yaklaşım önerilmiştir. Fakat bu yaklaşımlar çeşitli programlama dillerinde kod olarak çalıştırıldığı ve kullanıcı ara yüzüne sahip olmadıkları için kullanıcı dostu değillerdir. Bu sebeple dengesiz sınıf probleminin giderilmesi ve tutarlı sınıflandırma ölçütleri elde etmek için örnekleme tabanlı yöntemleri kullanan, dili Türkçe olan ve kullanıcı dostu bir yazılım geliştirilmiştir.

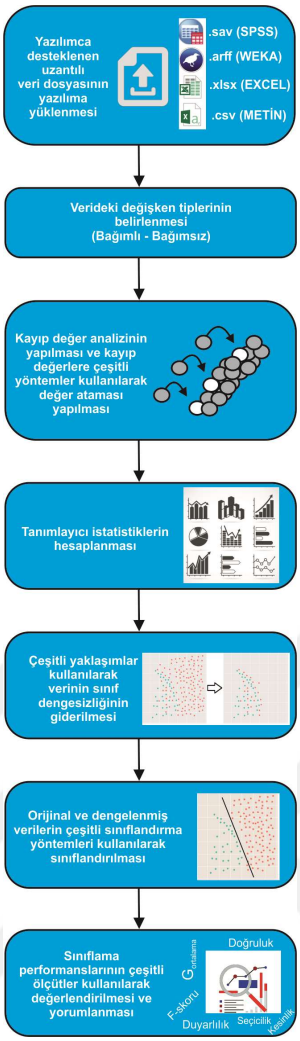
3.4.1. Yazılımın Arka Planı ve Kullanıcı Ara Yüzü

Yazılımın arka planındaki istatistiksel hesaplamalar için R programlama dili ile yazılmış çeşitli kütüphaneler kullanılmıştır. Yazılımdaki sınıflandırma işlemleri için caret (35), dengeleme analizleri için ise smotefamily (36), unbalanced (37) ve ROSE (38) kütüphaneler kullanılmıştır. Bu kütüphanelere ek olarak ihtiyaç duyulan yerlerde fonksiyonlar oluşturulmuştur. Yazılımdaki tüm betikler (script) RStudio Version 1.1.383 derleyicisi kullanılarak oluşturulmuştur.

Yazılımın kullanıcı ara yüzü ise Shiny kütüphanesi ile oluşturulmuştur. Shiny, R programlama dilinde oluşturulan uygulamaların birer web uygulamasına dönüşmesini sağlayan bir RStudio projesidir (39, 40). Shiny; web geliştiriciliği konusunda hiç deneyimi olmayan R kullanıcılarına yöneliktir. HTML, CSS, PHP ve JavaScript bilgisi gerektirmeyen Shiny kütüphanesi sayesinde kullanıcılar kendi web-tabanlı uygulamalarını kolayca oluşturabilmektedirler.

Geliştirilen yazılımdaki uygulamalar, TBKS'ye göre oluşturulmuştur. İlk olarak yazılıma, analize konu veri setini içeren dosya yüklenir. Dosya yüklemesi, veri

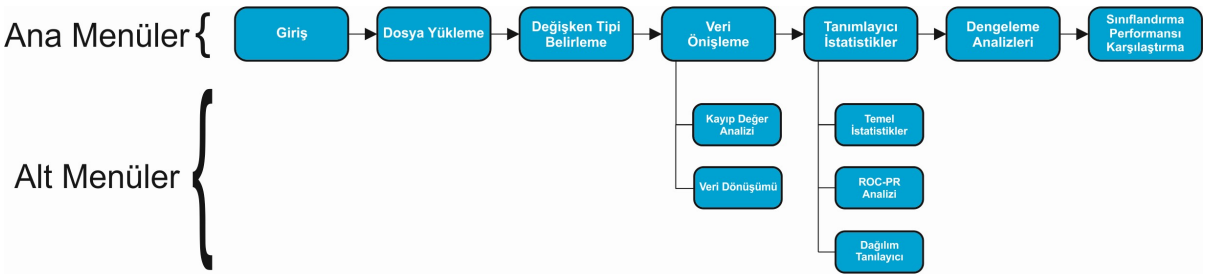
analitiğinde sıkça kullanılan dört farklı dosya türünde yapılabilmektedir. İkinci aşamada yüklenen veri setinde değişkenlerin tipleri belirlenmektedir. Değişkenler, bağımlı değişken ve bağımsız değişkenler olmak üzere iki kategoride yazılıma tanıtılır. Üçüncü aşamada, eğer veri setinde kayıp değerler mevcutsa bu problem çeşitli analitik ve analitik olmayan yöntemler kullanılarak giderilir. Ayrıca bu aşamada isteğe bağlı olarak çeşitli veri dönüşümü teknikleri uygulanabilir. Dördüncü aşamada, veri setinin tanımlayıcı istatistikleri çıkarılır. Bu aşamada veri setinin temel istatistikleri, ROC (Receiver operating characteristic) ve PR (Precision-Recall) eğrilerinin grafikleri ve bu eğriler altında kalan alan (AUC) değerleri ile değişkenlerin dağılım tanılamaları yapılır. Beşinci aşamada, dengesiz sınıf problemi olan veri setinin çeşitli sınıf dengeleme analizleri uygulanarak, sınıflar arası denge sağlanmaya çalışılır. Altıncı ve son aşamada, çeşitli sınıf dengeleme analizleri uygulanan veri setlerine yazılımda tanımlanmış olan sınıflandırma algoritmaları uygulanır ve her bir veri seti için elde edilen sınıflandırma performans ölçütleri tablolarda belirtilir. Geliştirilen yazılımın çalışma prensibi Şekil 3.1'de belirtilmiştir.



Şekil 3.1: Geliştirilen yazılımın çalışma prensibi

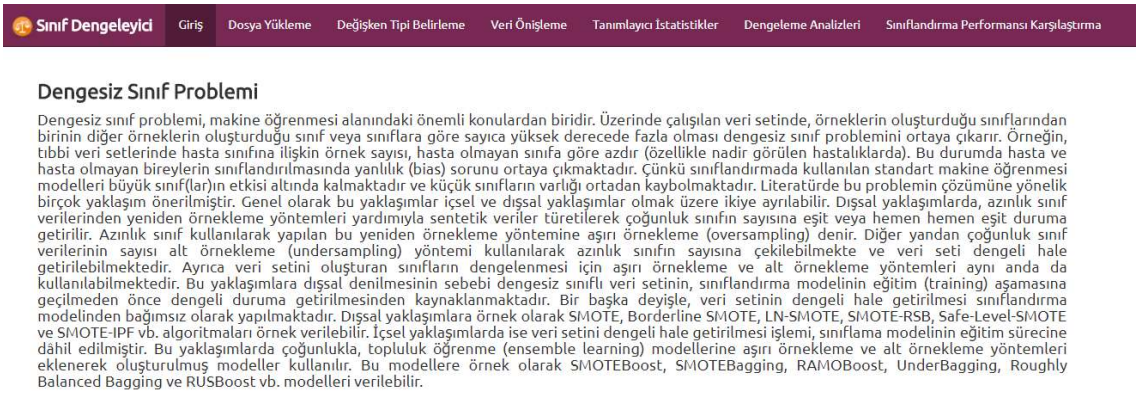
3.4.2. Menüler

Geliştirilen yazılım, 7 ana menü ve 3 alt menüden oluşmaktadır. Menü dizilimi ve adları Şekil 3.2'de verilmiştir.



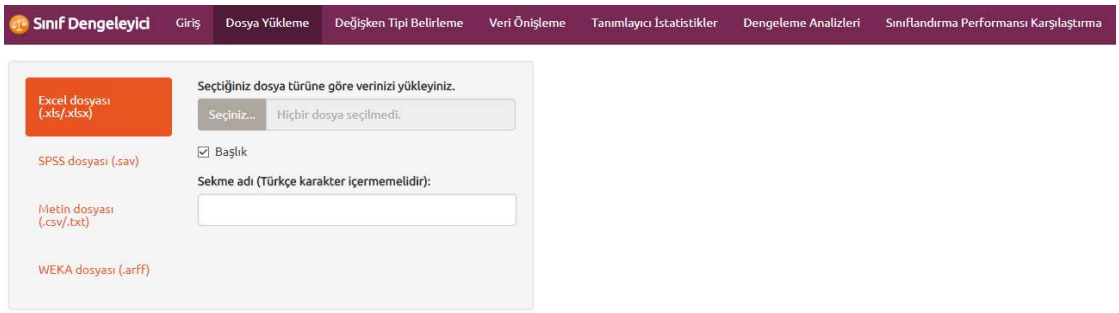
Şekil 3.2: Yazılımdaki menü dizilimi ve adları

1. Giriş: "Giriş" menüsünde dengesiz sınıf probleminden kısaca bahsedilip, bu problemin giderilmesinde kullanılan yaklaşımlar isim olarak tanıtılmıştır. Ayrıca yazılımın nasıl kullanılacağına ilişkin araştırıcıya kısa bir yönerge verilmiştir. "Giriş" ana menüsünün ekran görüntüsü Şekil 3.3'de belirtilmiştir.



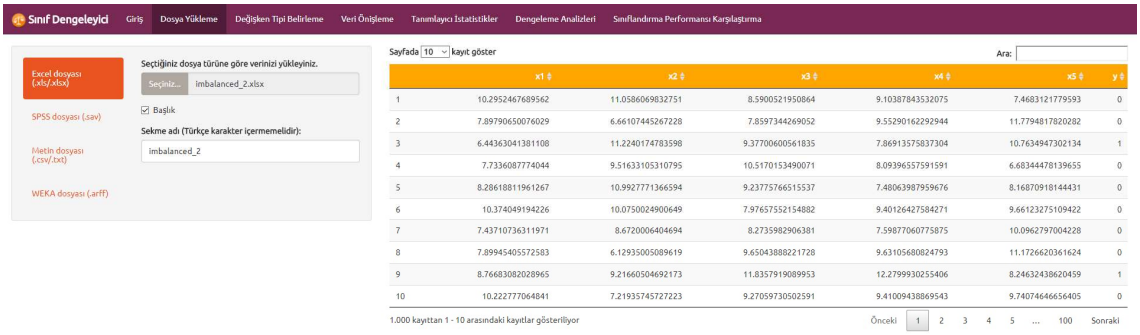
Şekil 3.3: Yazılımın "Giriş" ana menüsünün görünümü

2. Dosya Yükleme: Bu menüde analiz edilecek veri setinin bulunduğu dosya yazılıma yüklenir. Yazılıma veri analizlerinde en çok kullanılan 4 farklı uzantılı dosya türü yüklenebilmektedir. Bunlar MS Excel (.xls/.xlsx), SPSS (.sav), WEKA (.arff) ve metin (.csv/.txt) dosya türleridir. Araştırıcı yüklemek istediği dosya türünü seçtikten sonra yazılımda tanımlanmış ilgili dosya türüne ait parametreleri girmek suretiyle dosyayı yazılıma yükleyebilmektedir. Şekil 3.4'de "Dosya Yükleme" menüsünün ekran görüntüsü görülmektedir.



Şekil 3.4: Yazılımın "Dosya Yükleme" ana menüsünün görünümü

Dosya yazılıma yüklendikten sonra veri setinin ana paneldeki görünümü Şekil 3.5'de verilmiştir.

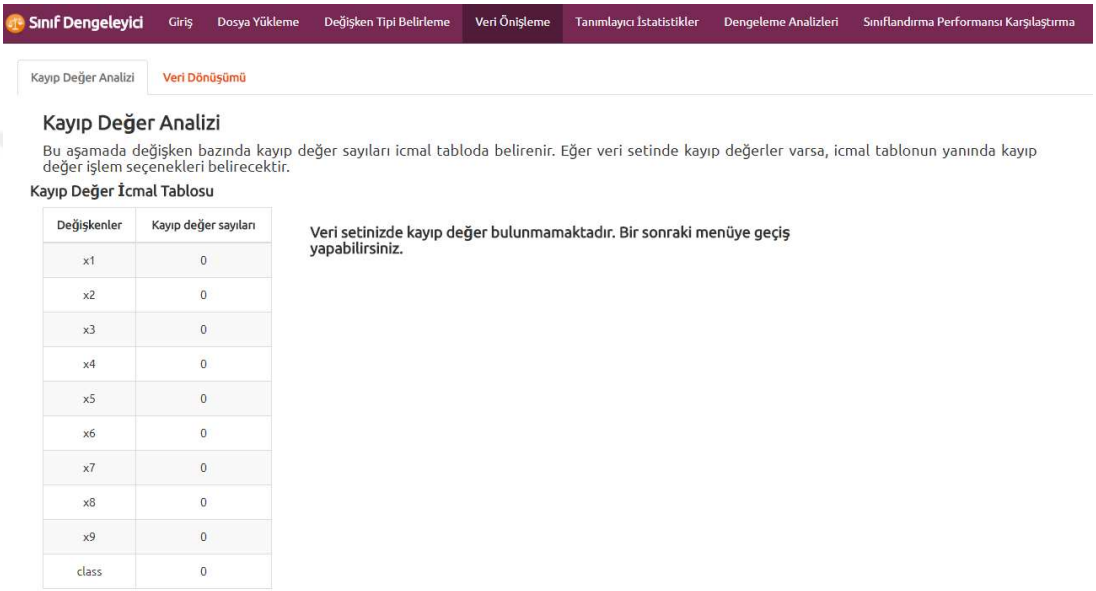


Şekil 3.5: Yüklenen dosyadaki veri setinin görünümü

3. Değişken Tipleri: Bu menüde analizi yapılacak veri setinin bağımlı (yanıt) ve bağımsız değişkenleri sisteme tanıtılır. Dosya yazılıma yüklendikten sonra her bir sütunda bulunan değişkenlerin isimleri otomatik olarak "Değişken Tipleri" menüsünde görüntülenir. Bu aşamada verilerin değişken tipleri belirlenmesinin yanı sıra, analize dâhil edilmekten vazgeçilen değişkenler de yazılımdan çıkarılabilir. Değişken isimleri uygun kutucuklarda belirtildikten sonra "Uygula" butonuna basılır ve değişken tipleri sisteme tanıtılmış olur. Değişken tipleri sisteme tanıtılmadan diğer menülere geçilemeyecektir. "Değişken Tipleri" ana menüsünün ekran görüntüsü Şekil 3.6'da verilmiştir.

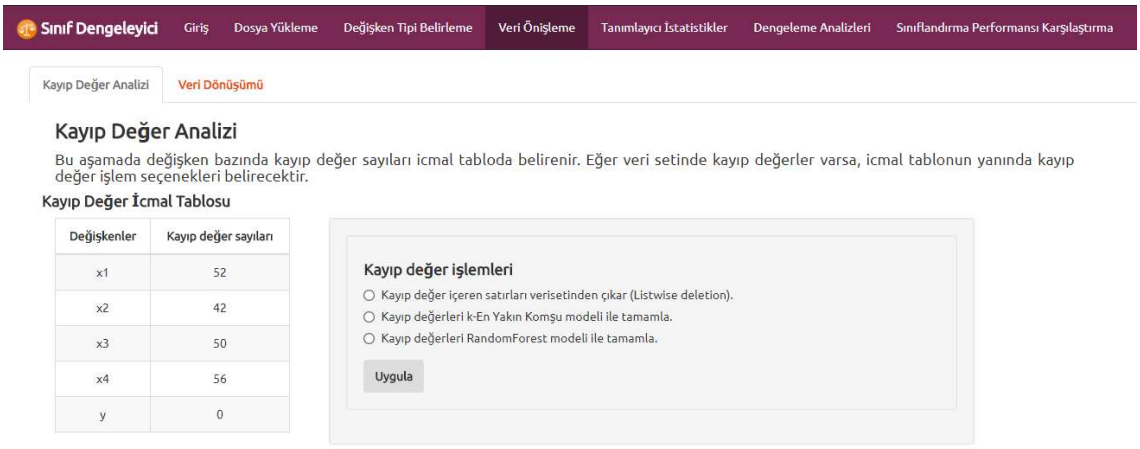
Şekil 3.6: "Değişken Tipi Belirleme" ana menüsünün görünümü

4. Veri Önişleme: Bu aşamada, ilgili veri setinin kayıp değer analizi ve veri dönüşümü işlemleri yapılmaktadır. Sınıf dengeleme analizlerinin yapılabilmesi için veri setinde kayıp değerlerin bulunmaması gerekmektedir. Menü içerisinde değişken bazında kayıp değer sayılarının yer aldığı bir icmal tablo görülmektedir. Eğer veri setinde kayıp değer bulunmuyorsa "Veri setinizde kayıp değer bulunmamaktadır. Bir sonraki menüye geçiş yapabilirsiniz." uyarısı ana panelde yer alacaktır. İlgili alt menünün ekran görüntüsü Şekil 3.7'de belirtilmiştir.



Şekil 3.7: Veri setinde kayıp değer/değerler olmadığı durumda "Kayıp Değer Analizi" alt menüsünün görünümü

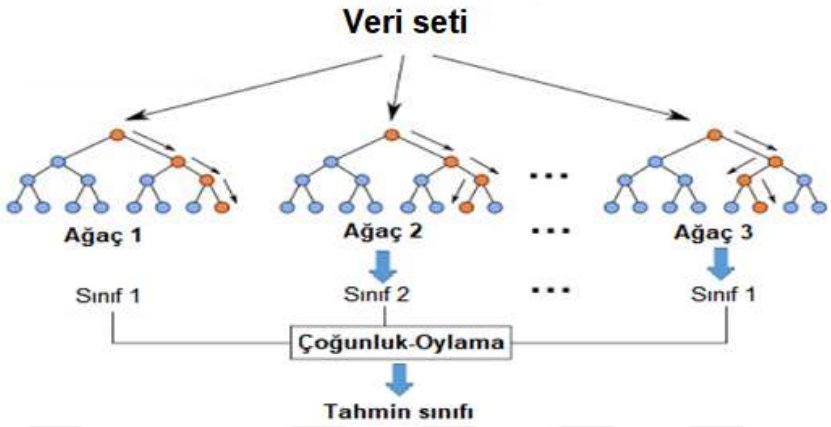
Şekil 3.8'de görüleceği üzere, eğer veri setinde kayıp değer/değerler mevcut ise, bu sorunu bertaraf etmek için çeşitli seçeneklerin bulunduğu bir panel belirecektir.



Şekil 3.8: Veri setinde kayıp değer/değerler olduğu durumda "Kayıp Değer Analizi" alt menüsünün görünümü

Bu panelde, kayıp değer sorununun giderilmesi için 3 seçenek çıkmaktadır. Birinci seçenekte, kayıp değer içeren satırın tamamı veri setinden çıkarılmaktadır. İkinci seçenekte ise kayıp değer içeren hücrelere k-En Yakın Komşu algoritması (kNN) yardımıyla değer ataması (imputasyon) yapılmaktadır. kNN tabanlı atama, kayıp değerli gözleme benzer (benzerlik ölçütü olarak genellikle Öklidyen uzaklık seçilir.) diğer gözlem değerlerinden faydalanılarak yapılır. Eğer kayıp değer sürekli bir sayısal değişken ise, k tane en yakın gözlemin ağırlıklı ortalamalarıyla değiştirilir. Burada ağırlık değerleri, Öklidyen uzaklık değerlerinin tersidir (41, 42). Son seçenekte ise Rastgele Orman (Random Forest) modeli ile kayıp değer ataması yapılmaktadır. Rastgele Orman (43) modeli en bilinen ağaç tabanlı topluluk öğrenme (ensemble learning) yöntemlerinden biridir. Rastgele Orman modelinin amacı, zayıf sınıflandırıcıları (örneğin tek bir sınıflandırma ağacı) güçlü bir sınıflandırıcıda toplamaktır. M adet bağımsız değişken, N adet de gözlem içeren bir veri seti düşünelim. Süreç ilk olarak bootstrap örnekleme ile başlar. Burada ormanı oluşturacak her bir sınıflandırma ağacı için yerine koyma yöntemi ile rastgele N gözlemli bootstrap örneklemleri eğitim (training) veri seti oluşturulur. Daha sonra, her düğümde süreç, önce rastgele m adet bağımsız değişken ($m < M$) seçer ve m değişkenler arasında en iyi bölünmeyi sağlayan tahmin edici değişkeni bulur. Böylece süreçte budama yapmadan ağaç büyür. Sonuçta, her bir ağacın bir örneğinin tahmin sonucu elde edilir. Sonuç olarak, çıktının ortalama veya ağırlıklı ortalaması (regresyon için) veya çoğunluk oyu (majority vote) kullanarak (sınıflandırma için) tahmin sonucu

bulunur (44). Şekil 3.9'da bir Rastgele Orman modelinin sınıflandırma problemlerinde basitleştirilmiş çalışma prensibi verilmiştir.



Şekil 3.9: Rastgele Orman modelinin basitleştirilmiş çalışma prensibi

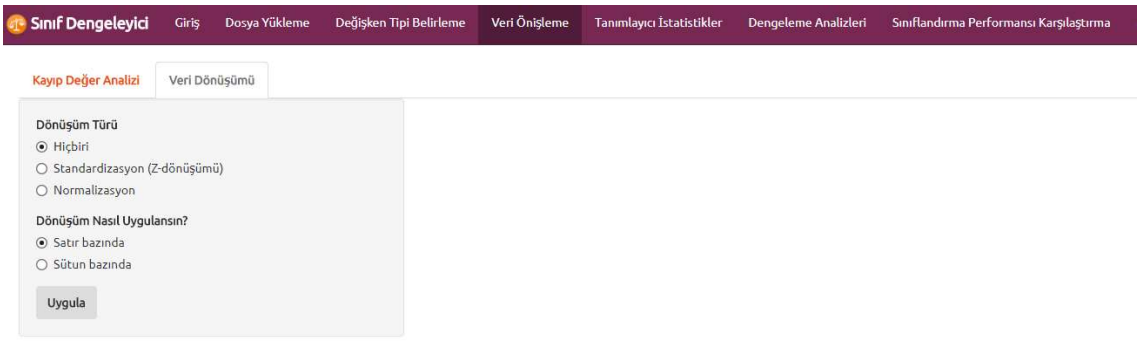
Rastgele Orman kayıp değer atama algoritmaları kayıp değerlerle başa çıkmak için etkili yaklaşımlardır. Karışık veri türlerini işleyebilecekleri, etkileşimlere (interactions) ve doğrusal olmamaya (non-linearity) uygun oldukları için büyük verilerde tercih sebebidir (45). Ayrıca Rastgele Orman kayıp değer atama algoritmaları veride değişkenliği koruyarak, atanan değerlerde yanlılığı azaltmaktadır (46).

"Veri Dönüşümü" alt menüsünde ise isteğe bağlı olarak, veri setine en çok kullanılan veri dönüştürme tekniklerinden olan standardizasyon (z-dönüşümü) ve normalizasyon teknikleri uygulanır. Standardizasyon ve normalizasyon dönüşüm tekniklerine ilişkin formüller aşağıda verilmiştir:

$$Standardizasyon\ (z-dönüşümü)\ = \frac{x-\bar{x}}{s}$$

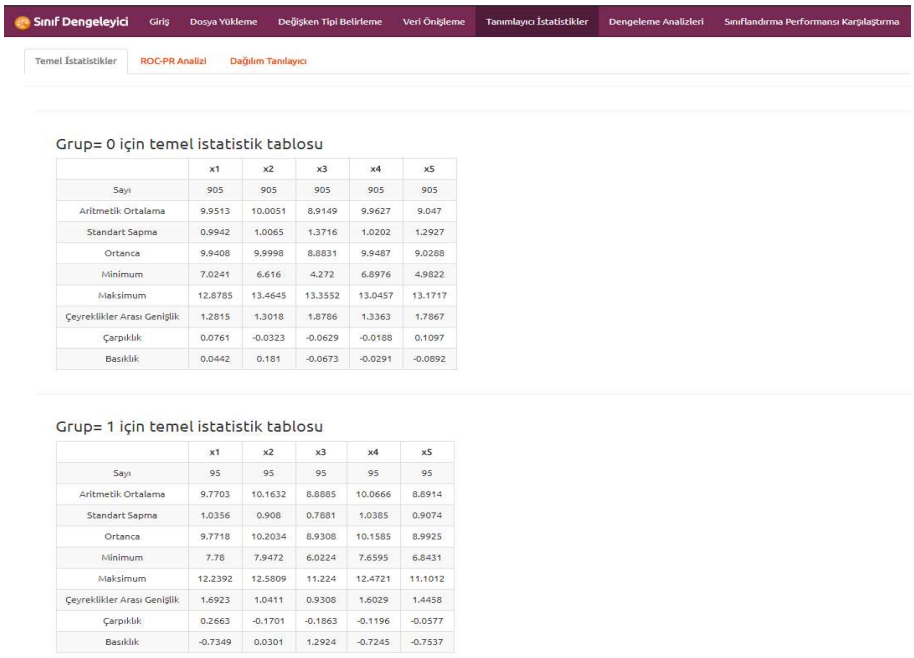
$$Normalizasyon = \frac{x-x_{min}}{x_{maks}-x_{min}}$$

Burada; x ilgili değişkendeki her bir gözlem değerini, $\bar{x}$ örneklem ortalamasını ve s ise örneklem standart sapmasını temsil eder. x_{min} ve x_{maks} değerleri ise ilgili değişkendeki sırasıyla en küçük ve en büyük gözlem değerlerini ifade etmektedir. Şekil 3.10'da "Veri Dönüşümü" alt menüsünün genel görünümü verilmiştir.



Şekil 3.10: "Veri Dönüşümü" alt menüsünün genel görünümü

5. Tanımlayıcı İstatistikler: Yüklenen veri setine dair tanımlayıcı istatistiklerin hesaplandığı bu ana menüde 3 adet alt menü bulunmaktadır. Bunlar; "Temel İstatistikler", "ROC-PR Analizi" ve "Dağılım Tanılayıcı" dır. Temel İstatistikler alt menüsünde veri setinin grup bazında; gözlem sayısı, aritmetik ortalama, standart sapma, ortanca, minimum ve maksimum değerler, çeyreklikler arası genişlik, çarpıklık ve basıklık gibi temel istatistikleri hesaplanır ve tablo olarak verilir. İlgili alt menünün ekran görüntüsü Şekil 3.11'da verilmiştir.



Grup= 0 için temel istatistik tablosu

	x1	x2	x3	x4	x5
Sayı	905	905	905	905	905
Aritmetik Ortalama	9.9513	10.0051	8.9149	9.9627	9.047
Standart Sapma	0.9942	1.0065	1.3716	1.0202	1.2927
Ortanca	9.9408	9.9998	8.8831	9.9487	9.0288
Minimum	7.0241	6.616	4.272	6.8976	4.9822
Maksimum	12.8785	13.4645	13.3552	13.0457	13.1717
Çeyreklikler Arası Genişlik	1.2815	1.3018	1.8786	1.3363	1.7867
Çarpıklık	0.0761	-0.0323	-0.0629	-0.0188	0.1097
Basıklık	0.0442	0.181	-0.0673	-0.0291	-0.0892

Grup= 1 için temel istatistik tablosu

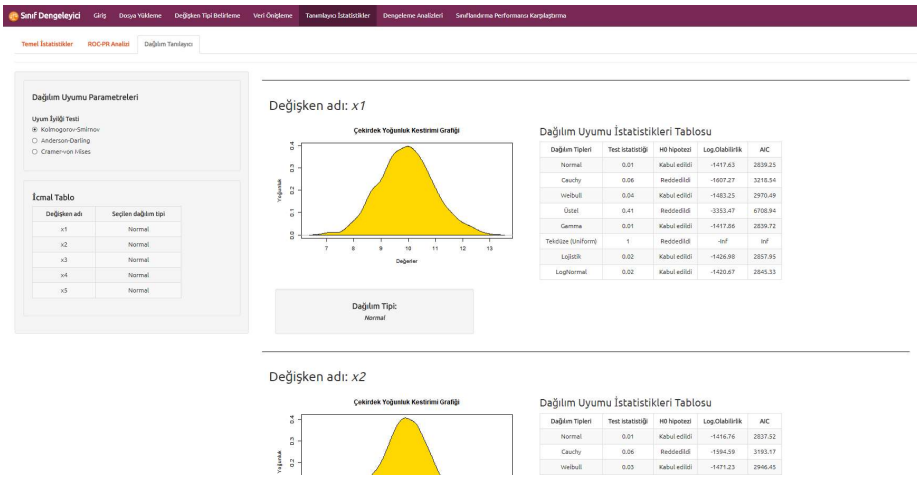
	x1	x2	x3	x4	x5
Sayı	95	95	95	95	95
Aritmetik Ortalama	9.7703	10.1632	8.8885	10.0666	8.8914
Standart Sapma	1.0356	0.908	0.7881	1.0385	0.9074
Ortanca	9.7718	10.2034	8.9308	10.1585	8.9925
Minimum	7.78	7.9472	6.0224	7.6595	6.8431
Maksimum	12.2392	12.5809	11.224	12.4721	11.1012
Çeyreklikler Arası Genişlik	1.6923	1.0411	0.9308	1.6029	1.4458
Çarpıklık	0.2663	-0.1701	-0.1863	-0.1196	-0.0577
Basıklık	-0.7349	0.0301	1.2924	-0.7245	-0.7537

Şekil 3.11: "Temel İstatistikler" alt menüsünü görünümü

ROC-PR Analizi alt menüsünde, yüklenen iki sınıflı veri seti için ROC (Receiver operating characteristic) ve PR (Precision-Recall) eğrilerinin grafikleri istenilen değişkenler için çizdirilir ve bu eğriler altında kalan alan (AUC) değerleri tablo olarak grafiklerin altında sunulur. İlgili alt menünün ekran görüntüsü Şekil 3.12'de verilmiştir.

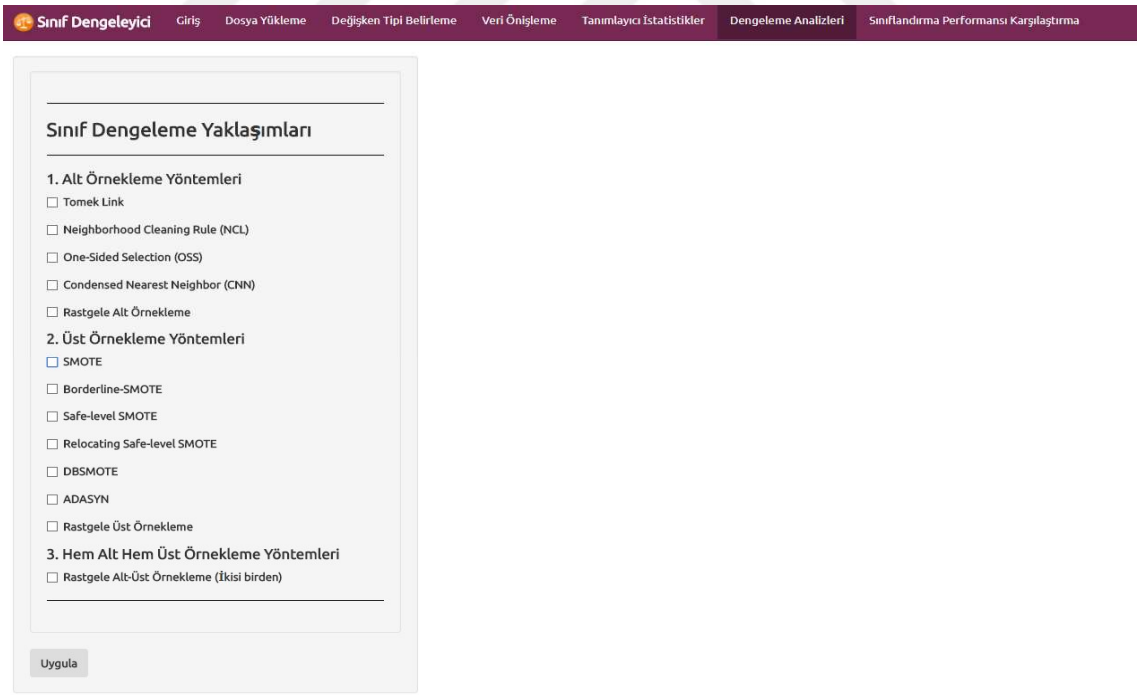
Şekil 3.12: "ROC-PR Analizi" alt menüsünü görünümü

Dağılım Tanılayıcı alt menüsünde, veri setindeki değişkenlerin, normal, cauchy, weibull, üstel, gamma, tekdüze (uniform), lojistik ve logaritmik normal teorik istatistiksel dağılım türlerine olan uyumları araştırılır. İlgili değişkenler parametre kestirimleri maksimum uyum iyiliği (Maximum Goodness of Fit) (47) yöntemi ile yapılmaktadır. Bu teorik dağılımlara olan uyumun iyiliği, Kolmogorov-Smirnov, Anderson-Darling ve Cramer von Mises testleri ile incelenmektedir. Uygulanan testler sonucunda ilgili değişkenin birkaç teorik dağılıma uygunluğu tespit edilirse, hangi teorik dağılımın değişkeni daha iyi tanımlandığı Akaike Bilgi Ölçütü (Akaike Information Criteria, AIC) kullanılarak belirlenmektedir. Bu işlemler değişken sayısı kadar yapılmakta ve sonuçlar ana panelde alt alta verilmektedir. Son olarak yan panelde ise tüm değişkenler için belirlenen teorik istatistiksel dağılım türleri "İcmal Tablo" da gösterilmektedir. Şekil 3.13'de Dağılım Tanılayıcı alt menüsünün görünümü verilmiştir.



Şekil 3.13: "Dağılım Tanılayıcı" alt menüsünü görünümü

6. Dengeleme Analizleri: Bu menüde bağımlı değişkeninin sınıfları arasında sayıca dengesizlik bulunan veri setinde, daha önce Genel Bilgiler kısmında ayrıntıları bahsedilen sınıf dengeleme yöntemlerinin bazıları kullanılarak sınıflar arası denge kurulmaya çalışılır. İlgili menünün görünümü Şekil 3.14'de belirtilmiştir.



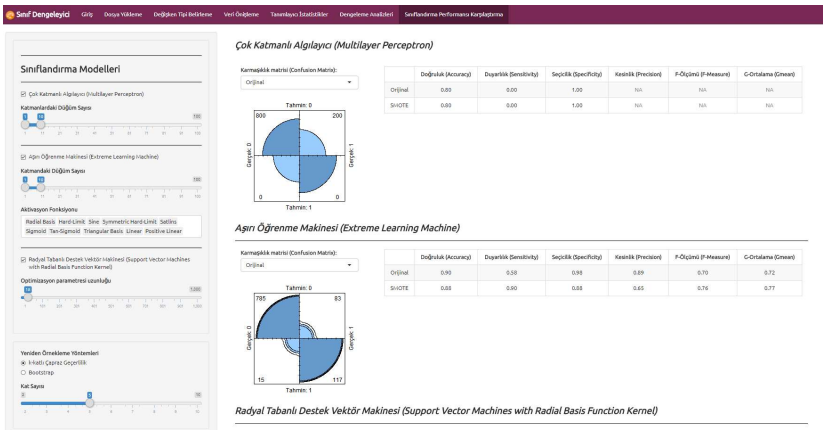
Şekil 3.14: "Dengeleme Analizleri" ana menüsünün görünümü

Yan panelde bulunan sınıf dengeleme yaklaşımlarından kullanılmak istenileni, yanlarında bulunan kutucuklar işaretlenerek seçilmektedir. Kutucuklar işaretlendiğinde,

ilgili sınıf dengeleme yaklaşımının parametresi mevcutsa bu parametre ve ayarlama (tuning) aralığı otomatik olarak belirecektir. İlgili yaklaşım seçilip, "Uygula" butonuna basıldıktan sonra ana panelde veri setinin ilgili yaklaşım uygulanmadan önce ve uyguladıktan sonraki temel bileşen (principal component) grafikleri ile bağımlı değişken sınıflarının dağılım tabloları verilmektedir. Dağılım tablolarını, sınıf bazında gözlem sayıları ve sınıf yüzdeleri oluşturmaktadır. Birden çok yaklaşım seçildiğinde sonuçlar ana panelde alt alta verilir. Şekil 3.15'de SMOTE ve ADASYN yaklaşımları uygulandığında ana panelde oluşturulan temel bileşen grafikleri ve dağılım tabloları görülmektedir.

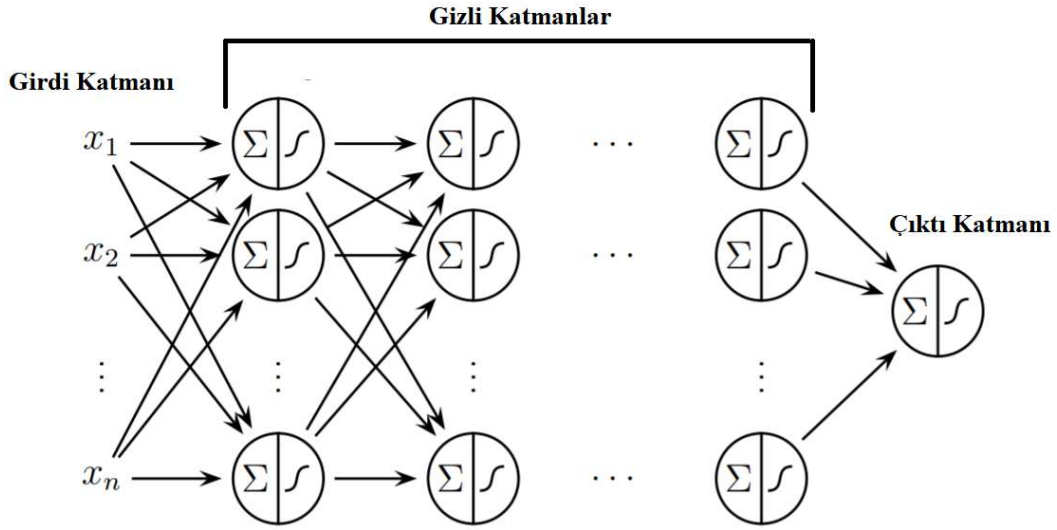
Şekil 3.15: SMOTE ve ADASYN sınıf dengeleme yaklaşımlarının uygulanması ve çıktılar

7. Sınıflandırma Performansı Karşılaştırma: Bu ana menüde, sınıf dengeleme yaklaşımı/yaklaşımları uygulanan veri setinin bağımlı değişkene göre çeşitli yöntemler kullanılarak sınıflandırılması yapılır. Daha sonra, uygulanan her bir sınıf dengeleme yaklaşımı ve sınıflandırma modeli bazında sınıflandırma performans sonuçları tablo ve grafikler ile gösterilir. Sınıflandırma yöntemi olarak üç model yazılıma eklenmiştir. Bunlar; Çok Katmanlı Algılayıcı (Multilayer Perceptron), Aşırı Öğrenme Makinesi (Extreme Learning Machine) ve Radyal Tabanlı Destek Vektör Makinesi (Support Vector Machines with Radial Basis Function Kernel)'dir. Şekil 3.16'da Sınıflandırma Performansı Karşılaştırma ana menüsünün içeriği görülmektedir.



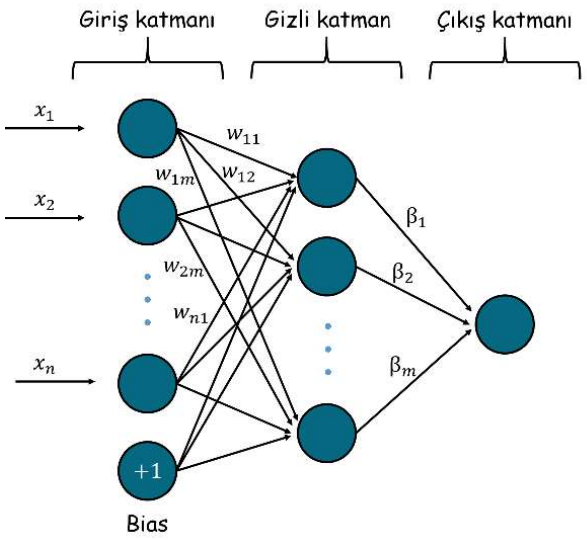
Şekil 3.16: "Sınıflandırma Performansı Karşılaştırma" ana menüsünün içeriği

Çok Katmanlı Algılayıcı (ÇKA), birçok disiplinde kendisine uygulama alanı bulan bir ileri beslemeli yapay sinir ağı türüdür. ÇKA mimarisinde, 1 giriş katmanı, 1 veya daha fazla gizli katman ve 1 çıkış katmanı bulunur. Her katman birkaç düğümden (node) oluşur. Birinci katman düğümlerindeki girdiler ağırlıklandırılır ve gizli katman olarak bilinen, nöron benzeri düğümlerden oluşan ikinci bir tabakaya aynı anda gönderilir. Düğümlerin her biri, önceki katmanın çıktılarının ağırlıklı toplamını girdi olarak alır ve çıktısını belirlemek için doğrusal olmayan bir aktivasyon fonksiyonu uygular. Aktivasyon fonksiyonu genellikle çıktıyı 0 ile 1 arasında bir sayıya çeviren bir sigmoid fonksiyondur. Eğer sistem tek gizli katmandan oluşuyorsa gizli katmanın çıktı değerleri son çıktı katmanına gönderilir ve tahmin sınıfı belirlenir. Eğer sistem birden fazla gizli katmandan oluşuyorsa ilk gizli katmanın çıktıları diğer gizli katmana girişleri olarak gönderilir (48). Bir sigmoid aktivasyon fonksiyonlu bir ÇKA yapısı Şekil 3.17'da görselleştirilmiştir.



Şekil 3.17: Bir sigmoid aktivasyon fonksiyonlu bir ÇKA yapısı (49)

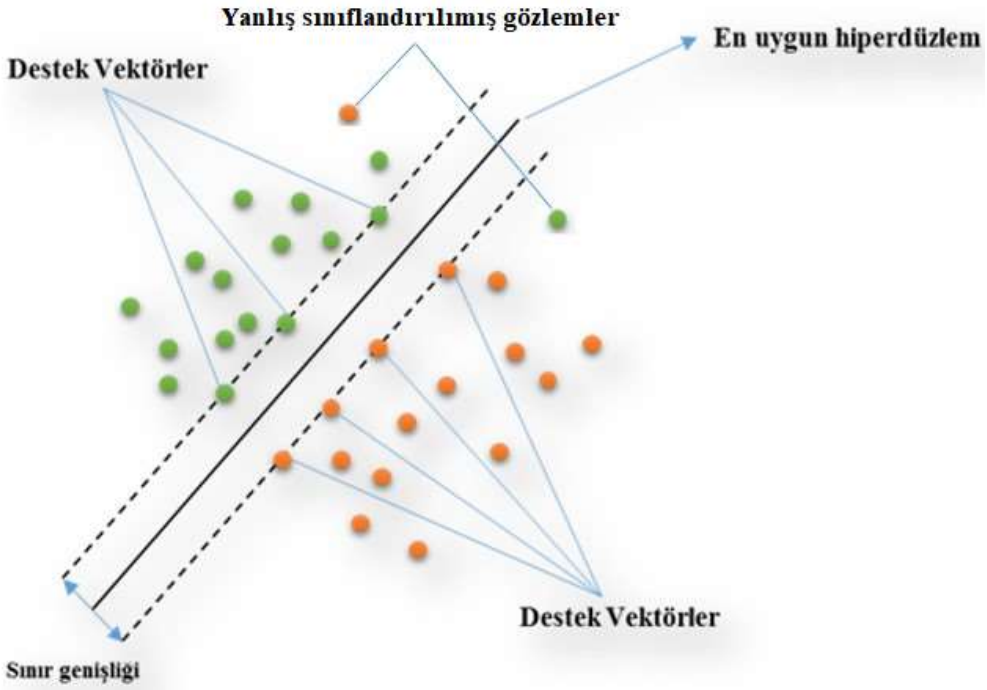
Aşırı Öğrenme Makinesi (AÖM) modeli, Huang ve arkadaşları (50) tarafından 2006 yılında geliştirilen AÖM, girdi ağırlıklarının rastgele, çıktı ağırlıklarının analitik olarak hesaplandığı, tek gizli katmanlı ve ileri beslemeli bir yapay sinir ağları modelidir. Standart yapay sinir ağları algoritmalarında giriş ağırlıkları en iyi hata oranını elde edinceye kadar denenir ve her bir düğümdeki en uygun (optimal) ağırlıkları tespit edebilmek için tüm parametrelerin (momentum, öğrenme katsayısı) iteratif yöntemlerle ayarlanması gerekir. Bu da modelin öğrenmesi için çok uzun bir zamana ihtiyacı olduğunu gösterir. AÖM'de standart yapay sinir ağları algoritmalarından farklı olarak, giriş ağırlıkları ve biaslar rastgele seçilir. Buna karşın gizli katmandan çıkış ağırlıkları ise analitik olarak belirlenir (Moore–Penrose pseudo-inverse yöntemi) (50-52). Geleneksel öğrenme algoritmalarından farklı olarak AÖM, sadece daha küçük öğrenme hatası sağlamakla kalmaz ayrıca yüksek öğrenme hızı sayesinde daha iyi performans da sağlar. AÖM'nin mimarisi Şekil 3.18'de verilmiştir.



Şekil 3.18: AÖM modelinin mimarisi

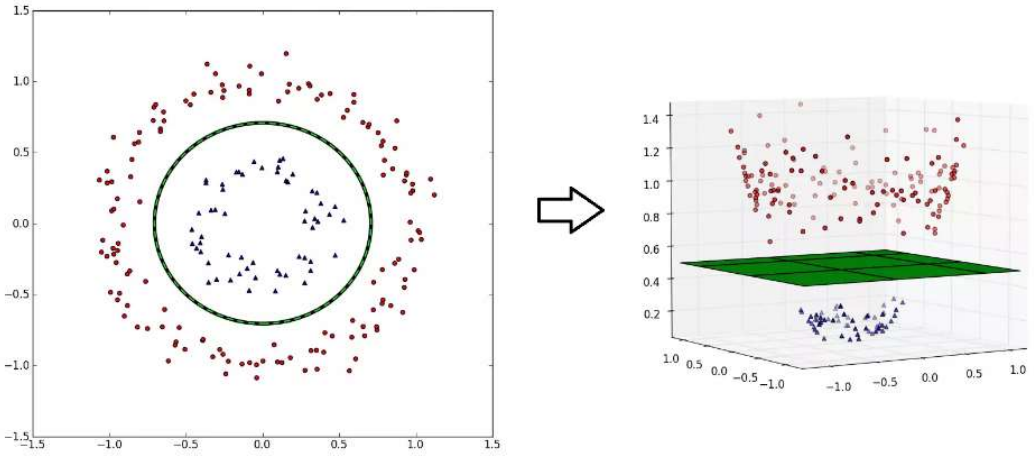
Yüksek sınıflama doğruluğu ve hesaplama hızı nedeniyle yazılımda AÖM modeli kullanılmıştır.

Destek Vektör Makineleri (DVM) (53), sınıflandırma ve regresyon problemlerinde kullanılan, yüksek tahmin başarısı gösteren bir modeldir. Genellikle sınıflandırma modellerinde kullanılan DVM, çok genel bir sınıflandırma metodu grubunu içerir ve bunlar hem doğrusal hem de doğrusal olmayan sınıflandırıcıları kapsar (48). Sınıflandırma problemlerinde DVM'nin amacı, bağımlı değişkenin sınıflarını birbirinden ayıracak en uygun hiperdüzlemi tespit etmektir (54). DVM bunu, hiperdüzlemler üzerindeki en büyük kenar boşluklarına sahip veri noktalarını tanımlayan "destek vektörleri" bularak başarır. Şekil 3.19'da görüleceği üzere, bağımlı değişken sınıflarının doğrusal olarak sınıflanabildiği durumlarda en uygun hiperdüzlemi bulmak kolaydır.



Şekil 3.19: Doğrusal ayrılma durumu

Fakat sınıfların doğrusal olarak ayrılamadığı durumlarda süreç daha karmaşıklaşır. Bu durumda çekirdek fonksiyonları yardımıyla veri doğrusal ayrımın mümkün olduğu daha yüksek boyutlu bir uzaya dönüştürülür ve sınıflandırma işlemi gerçekleştirilir (48). Çekirdek fonksiyon yardımıyla doğrusal ayırmanın sağlanması Şekil 3.20'de gösterilmiştir.



Şekil 3.20: Çekirdek fonksiyon yardımıyla doğrusal ayırmanın sağlanması (55)

Gerek sınıflandırma problemlerinde gösterdiği başarısı, gerekse hesaplama hızının yüksekliği nedeniyle yazılımda radyal tabanlı çekirdek fonksiyonu ile oluşturulmuş DVM kullanılmıştır.

Yazılımda, model performansını değerlendirmek, yansız çıktılar elde etmek ve aşırı uyum problemini ortadan kaldırmak için k-katlı çapraz geçerlilik ve bootstrap örnekleme yöntemleri kullanılmıştır. Yazılımda kullanılan sınıflandırma modellerinin optimizasyon parametreleri için optimizasyon aralıkları, ilgili modellerin altında belirtilmiştir. Sınıflandırma modellerinin optimizasyon parametreleri ızgara arama (grid search) algoritması ile optimize edilmiştir. Bu çalışma için modellerin optimizasyon parametreleri ve seçilen optimizasyon aralıkları Tablo 3.2'de verilmiştir.

Tablo 3.2: Modellerin optimizasyon parametreleri ve seçilen optimizasyon aralıkları

Model	**Optimizasyon Parametresi**	**Optimizasyon Parametresi Aralığı**
ÇKA	• Gizli katmandaki düğüm sayısı	• 1-10
AÖM	• Gizli katmandaki düğüm sayısı • Aktivasyon Fonksiyonu	• 1-10 • Sigmoid, Sine, Radial Basis, Hard-Limit, Symmetric Hard-Limit, Sigmoid, Triangular Basis, Satlins, Positive Linear, Linear
DVM	• Maliyet (C) parametresi	• 2^{-2} - 2^{50}

Modellerin sınıflandırma performansları çeşitli ölçütler kullanılarak hesaplanmıştır. Bu ölçütler; doğruluk (accuracy), duyarlılık (sensitivity), seçicilik

(specificity), kesinlik (precision), F-ölçümü (F-measure) , $G_{Ortalama}$ (G_{mean}) dir. İlgili ölçütler Tablo 3.3'de verilen sınıflandırma matrisi (confusion matrix) değerleri ile hesaplanmaktadır.

Tablo 3.3: Gerçek durum ile model çıktıları ile oluşturulan sınıflandırma matrisi

Hipertansiyon varlığı		**Gerçek durum**		
		Var	**Yok**	**Toplam**
Tahmin edilen durum	**Var**	DP	YP	DP+YP
	Yok	YN	DN	YN+DN
	Toplam	DP+YN	YP+DN	DN+YP+YN+DP

DN: Doğru Negatif, YP: Yanlış Pozitif, YN: Yanlış Negatif, DP: Doğru Pozitif

Yukarıda bahsedilen performans ölçütleri Tablo 3.3'de bahsedilen sınıflandırma matrisi değerleri kullanılarak şöyle hesaplanır:

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

$$Duyarlılık = \frac{DP}{DP + YN}$$

$$Seçicilik = \frac{DN}{DN + YP}$$

$$Kesinlik = \frac{DP}{DP + YP}$$

$$F - \text{ölçümü} = \frac{2 * Duyarlılık * Kesinlik}{Duyarlılık + Kesinlik}$$

$$G_{ortalama} = \sqrt{Duyarlılık * Kesinlik}$$

4. BULGULAR

Çalışmada kullanılan veri seti 1 bağımlı ve 8 bağımsız değişken olmak üzere toplam 9 değişkenden oluşmaktadır. Bağımlı değişken olan hipertansiyon değişkeninin dağılım tablosu ve sınıflar arası dengesizlik oranı Tablo 4.1'de verilmiştir.

Tablo 4.1: Hipertansiyon değişkeninin dağılım tablosu ve sınıflar arası dengesizlik oranı

Hipertansiyon varlığı				**Sınıflar arası**
Yok		**Var**		**Dengesizlik**
Sayı	**Yüzde**	**Sayı**	**Yüzde**	**Oranı**
780	%84.0	149	%16.0	5.235

Burada sınıflar arası dengesizlik oranı, HT'li olmayan hasta sayısının HT'li olan hasta sayısına oranlanmasıyla elde edilmiştir. Tablo 4.2'de görüleceği üzere, bağımsız değişkenlerin 5 tanesinde toplam 18 tane kayıp değer içeren gözlem bulunmaktaydı.

Tablo 4.2: Değişken bazında kayıp değer sayıları

Değişkenler	**Yaş**	**VYA**	**VKİ**	**BUN**	**KR**	**HB**	**HCT**	**PLT**
Kayıp değer sayıları	0	0	0	1	1	5	5	6

Kayıp değer içeren gözlemlere, geliştirilen yazılımdaki Rastgele Orman (Random Forest) algoritması kullanılarak değer ataması yapılmıştır. Kayıp değer ataması yapıldıktan sonra 8 adet sayısal bağımsız değişkene ait ayrıntılı tanımlayıcı istatistik tablosu Tablo 4.3'de verilmiştir.

Tablo 4.3: Veri setindeki sayısal bağımsız değişkenlere ilişkin ayrıntılı tanımlayıcı istatistik tablosu

Değişkenler	Hipertansiyon durumu: Yok (*n*=780)								Hipertansiyon durumu: Var (*n*=149)							
						Çeyreklikler								Çeyreklikler		
	Aritmetik Ortalama	Standart Sapma	Ortanca	Minimum	Maksimum	Arası Genişlik	Çarpıklık	Basıklık	Aritmetik Ortalama	Standart Sapma	Ortanca	Minimum	Maksimum	Arası Genişlik	Çarpıklık	Basıklık
Yaş	59.81	9.88	59	24	82	14	-0.13	-0.43	63.01	8.97	65	40	82	14	-0.37	-0.56
VYA	179.58	18.97	181	1.64	234	20	-3.20	28.06	174.97	16.97	175	137	236	21	0.18	0.37
VKİ	26.08	3.64	26	16	44	4	0.77	2.45	27.15	4.12	26	18	41	5	0.70	0.72
BUN	18.71	7.31	17	6	77	7	2.35	10.61	20.75	11.06	18	7	86	8	2.90	11.23
KR	1.01	0.60	0.95	0.1	13.3	0.26	15.68	294.21	1.08	0.57	1	0.39	6	0.38	5.34	39.07
HB	14.13	1.92	14.2	9.2	39	2.1	3.57	44.17	13.41	1.65	13.6	8.5	19.4	2.1	0.05	1.75
HCT	41.69	4.93	42	11.1	54	6	-1.09	4.40	40.09	4.93	40.7	26	57.4	6.1	0.02	1.05
PLT	257.39	74.22	248	105	794	83.25	1.48	5.73	270.93	68.45	262	108	507	83	0.57	0.62

Veri setindeki kayıp değer sorunu giderildikten sonra, standardizasyon (z-dönüşümü) tekniği uygulanmıştır. Daha sonra, alt ve üst örnekleme yöntemleri tek tek uygulanarak sınıflar arası dengesizlik giderilmeye çalışılmıştır. Bu kapsamda geliştirilen yazılımda tanımlı olan Tomek Link, NCL, OSS, CNN, Rastgele Alt Örnekleme (RAÖ), SMOTE, Borderline-SMOTE (BLSMOTE), Safe-level SMOTE (SLSMOTE), DBSMOTE, ADASYN, Rastgele Üst Örnekleme (RÜÖ) ve Rastgele Alt-Üst Örnekleme (RAÜÖ) yöntemleri uygulandıktan sonra çıkarılan/eklenen gözlem sayısı ve yeni oluşan sınıflar arası dengesizlik oranı Tablo 4.4'de verilmiştir.

Tablo 4.4: Sınıf dengeleme yöntemleri uygulandıktan sonra çıkarılan/eklenen gözlem sayısı ve yeni oluşan sınıflar arası dengesizlik oranları

Yöntem	**Orijinal veri seti sınıf dağılımı**		**Orijinal veri seti sınıflar arası dengesizlik oranı**	**İlgili yöntem uygulandıktan sonra sınıf dağılımı**		**İlgili yöntem uygulandıktan sonra sınıflar arası dengesizlik oranı**
	HT: Yok	**HT: Var**		**HT: Yok**	**HT: Var**	
Tomek Link	780	149	5.235	672	149	4.510
NCL	780	149	5.235	575	149	3.859
OSS	780	149	5.235	665	149	4.463
CNN	780	149	5.235	771	149	5.174
RAÖ	780	149	5.235	149	149	1.000
SMOTE	780	149	5.235	780	745	1.047
BLSMOTE	780	149	5.235	780	772	1.010
SLSMOTE	780	149	5.235	780	557	1.400
DBSMOTE	780	149	5.235	780	665	1.173
ADASYN	780	149	5.235	780	761	1.025
RÜÖ	780	149	5.235	780	780	1.000
RAÜÖ	780	149	5.235	462	467	0.989

Dengeleme analizleri uygulandıktan sonra, her bir yaklaşımla elde edilen veri setleri ve orijinal (hiçbir sınıf dengeleme yaklaşımı uygulanmamış) veri setine sınıflandırma modelleri uygulanmıştır. İlgili sınıflandırma modellerinin öğrenme performansı 5-kat çapraz geçerlilik yöntemi ile test edilmiştir. Modellerin sınıflandırma

performansları, tüm sınıf dengeleme yaklaşımları bazında Tablo 4.5, Tablo 4.6 ve Tablo 4.7'de verilmiştir.

Tablo 4.5: Tüm sınıf dengeleme yaklaşımları bazında ÇKA modelinin sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-Ölçümü	G-Ortalama
Hiçbiri	0.84	0.00	1.00	-	-	-
Tomek Link	0.84	0.00	1.00	-	-	-
NCL	0.42	0.77	0.35	0.18	0.30	0.38
OSS	0.84	0.00	1.00	-	-	-
CNN	0.84	0.00	1.00	-	-	-
RAÖ	0.18	0.99	0.03	0.16	0.28	0.40
SMOTE	0.82	0.08	0.96	0.29	0.13	0.15
BLSMOTE	0.84	0.00	1.00	-	-	-
SLSMOTE	0.84	0.00	1.00	-	-	-
DBSMOTE	0.80	0.10	0.94	0.24	0.14	0.15
ADASYN	0.84	0.00	1.00	-	-	-
RÜÖ	0.83	0.01	0.99	0.15	0.02	0.05
RAÜÖ	0.65	0.52	0.68	0.24	0.32	0.35

Tablo 4.5'e göre ÇKA modelinin, orijinal veri seti üzerinde gösterdiği sınıflandırma performans değerleri ile alt örnekleme yaklaşımlarından Tomek Link, OSS ve CNN ile üst örnekleme yaklaşımlarından BLSMOTE, SLSMOTE ve ADASYN uygulandıktan sonraki sınıflandırma performans değerlerinin aynı olduğu görülmektedir. NCL yaklaşımı uygulandıktan sonra ÇKA modelinin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.42, 0.77, 0.35, 0.18, 0.30 ve 0.38 olarak bulunmuştur. Bir diğer alt örnekleme yöntemi olan RAÖ uygulandıktan sonra ÇKA modelinin sınıflandırma performans

değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.18, 0.99, 0.03, 0.16, 0.28 ve 0.40 olarak bulunmuştur. SMOTE, DBSMOTE, ADASYN ve RÜÖ yöntemlerine göre ÇKA sınıflandırma performansları birbirlerine benzer niteliktedir. RAÜÖ yöntemi uygulandıktan sonra ÇKA modelinin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.65, 0.52, 0.68, 0.24, 0.32 ve 0.35 olarak bulunmuştur.

Tablo 4.6: Tüm sınıf dengeleme yaklaşımları bazında AÖM modelinin sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-Ölçümü	G-Ortalama
Hiçbiri	0.84	0.00	1.00	-	-	-
Tomek Link	0.84	0.00	1.00	-	-	-
NCL	0.84	0.00	1.00	-	-	-
OSS	0.84	0.00	1.00	-	-	-
CNN	0.84	0.00	1.00	-	-	-
RAÖ	0.66	0.64	0.66	0.26	0.37	0.41
SMOTE	0.68	0.66	0.69	0.29	0.40	0.43
BLSMOTE	0.68	0.66	0.68	0.28	0.40	0.43
SLSMOTE	0.75	0.45	0.80	0.30	0.36	0.37
DBSMOTE	0.71	0.54	0.75	0.29	0.37	0.39
ADASYN	0.67	0.62	0.68	0.27	0.38	0.41
RÜÖ	0.70	0.63	0.71	0.29	0.40	0.43
RAÜÖ	0.69	0.62	0.70	0.28	0.39	0.42

Tablo 4.6'ya göre AÖM modelinin, orijinal veri seti üzerinde gösterdiği sınıflandırma performans değerleri ile alt örnekleme yaklaşımlarından Tomek Link,

NCL, OSS ve CNN uygulandıktan sonraki sınıflandırma performans değerlerinin aynı olduğu görülmektedir. Duyarlılık ölçütü değeri sıfır olduğundan, bir başka deyişle model HT olan hastaları yanlış sınıflandırdığından dolayı kesinlik, f-ölçümü ve g-ortalama değerleri hesaplanamamıştır. RAÖ alt örnekleme yaklaşımı uygulandıktan sonra AÖM modelinin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.66, 0.64, 0.66, 0.26, 0.37 ve 0.41 olarak bulunmuştur. Üst örnekleme yaklaşımları uygulandıktan sonra AÖM modelinin sınıflandırma performansı incelendiğinde, en yüksek doğruluk, seçicilik ve kesinlik değerlerinin SLSMOTE, en yüksek duyarlılık değerinin BLSMOTE uygulandıktan sonra elde edilmiştir. En yüksek f-ölçümü ve g-ortalama değerleri SMOTE, BLSMOTE ve RÜÖ yaklaşımları uygulandıktan sonra elde edilmiştir. Rastgele hem alt hem üst örnekleme yönteminin uygulandığı RAÜÖ yaklaşımında ise AÖM modelinin performans ölçüt değerlerinin, üst örnekleme yöntemlerinin değerleri ile paralel olduğu görülmektedir.

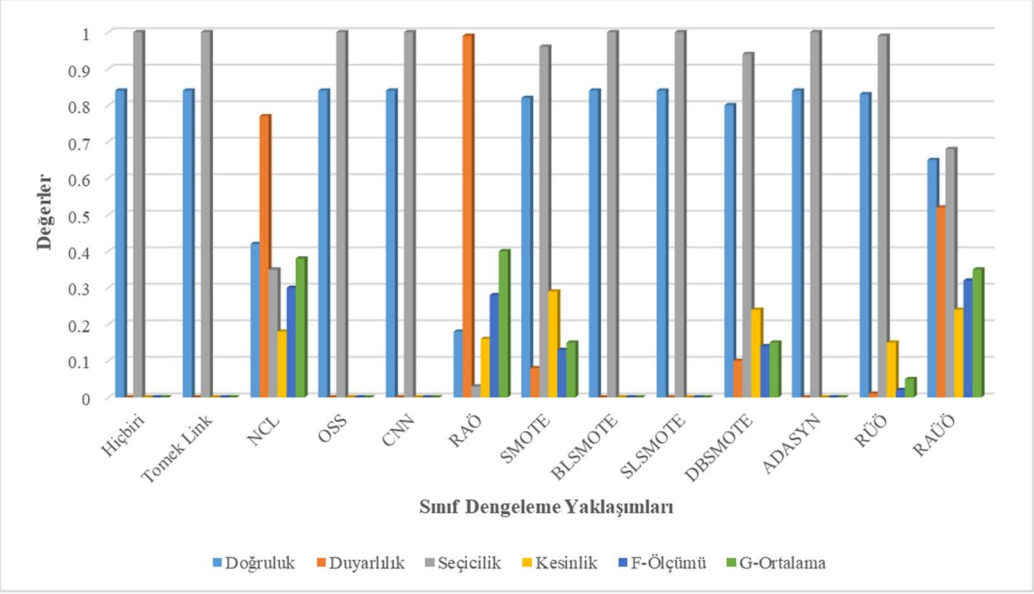
Tablo 4.7: Tüm sınıf dengeleme yaklaşımları bazında DVM modelinin sınıflandırma performansı

Yöntem	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-Ölçümü	G-Ortalama
Hiçbiri	0.84	0.00	1.00	-	-	-
Tomek Link	0.84	0.00	1.00	-	-	-
NCL	0.84	0.00	1.00	-	-	-
OSS	0.92	0.54	0.99	0.92	0.68	0.71
CNN	0.84	0.00	1.00	-	-	-
RAÖ	0.70	1.00	0.64	0.35	0.52	0.59
SMOTE	0.99	1.00	0.99	0.94	0.97	0.97
BLSMOTE	0.97	0.94	0.98	0.89	0.92	0.92
SLSMOTE	0.98	0.91	0.99	0.94	0.92	0.92
DBSMOTE	0.99	0.99	0.99	0.95	0.97	0.97
ADASYN	0.98	1.00	0.97	0.88	0.94	0.94

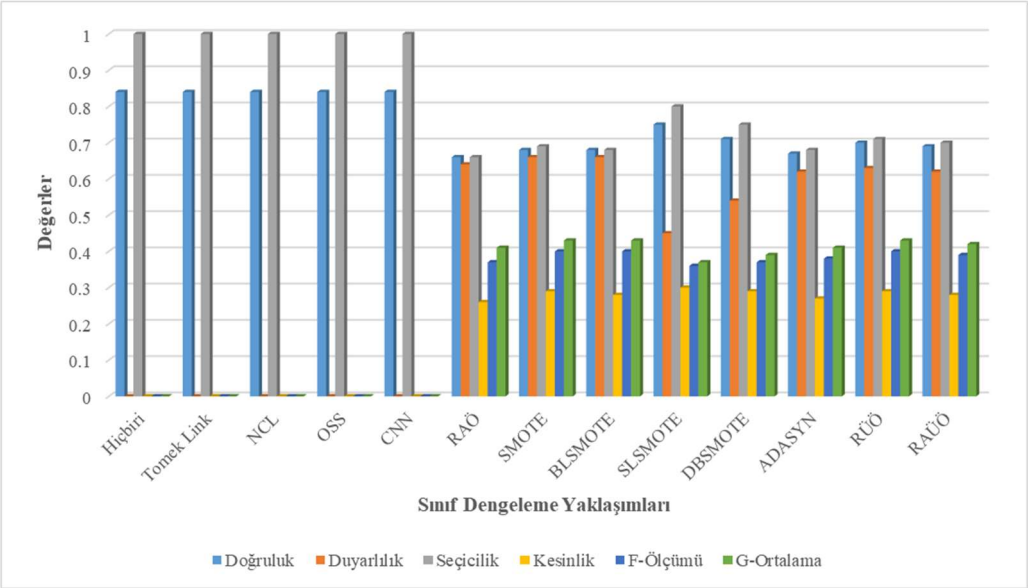
RÜÖ	0.92	0.87	0.93	0.70	0.78	0.78
RAÜÖ	0.92	0.96	0.92	0.69	0.80	0.81

Tablo 4.7'ye göre DVM modelinin, orijinal veri seti üzerinde gösterdiği sınıflandırma performans değerleri ile alt örnekleme yaklaşımlarından Tomek Link, NCL ve CNN uygulandıktan sonraki sınıflandırma performans değerlerinin aynı olduğu görülmektedir. Bunlardan farklı olarak OSS yaklaşımından sonra elde edilen DVM modelinin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.92, 0.54, 0.99, 0.92, 0.68 ve 0.71 olarak hesaplanmıştır. Yine alt örnekleme yaklaşımlarından olan RAÖ uygulandıktan sonra elde edilen DVM sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.70, 1.00, 0.64, 0.35, 0.52 ve 0.59 olarak elde edilmiştir. Üst örnekleme yaklaşımları uygulandıktan sonra DVM'nin sınıflandırma performansı tüm ölçütlerde 0.70'in üzerinde elde edilmiştir. RAÜÖ yaklaşımında ise DVM modelinin sınıflandırma performans değerleri doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçümü ve g-ortalama ölçütleri açısından sırasıyla 0.92, 0.96, 0.92, 0.69, 0.80 ve 0.81 olarak hesaplanmıştır.

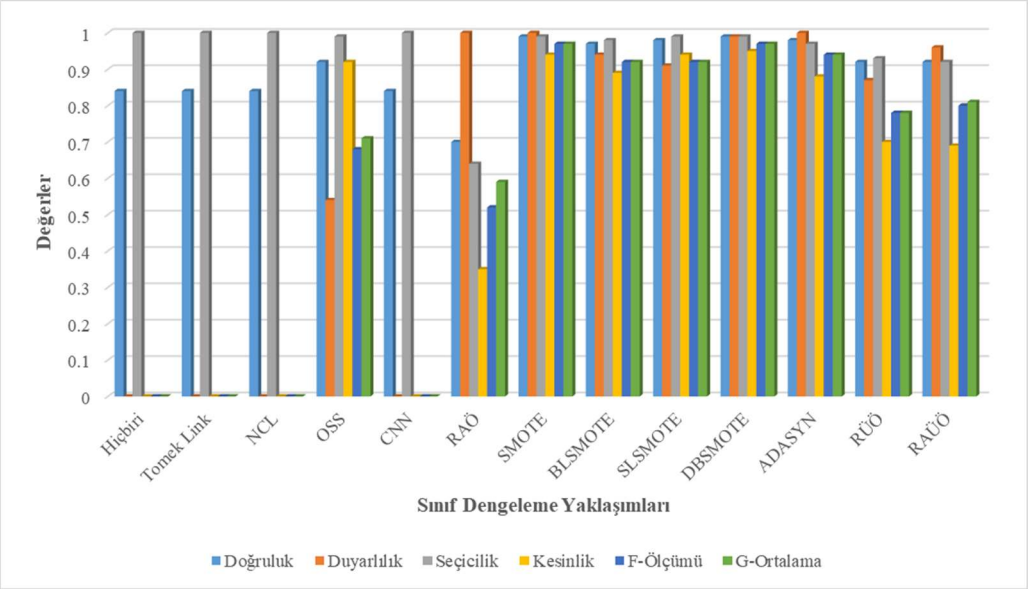
Tüm sınıf dengeleme yaklaşımları bazında ÇKA, AÖM ve DVM modellerinin sınıflandırma performansları sırasıyla Şekil 4.1, Şekil 4.2 ve Şekil 4.3'de verilmiştir.



Şekil 4.1: Tüm sınıf dengeleme yaklaşımları bazında ÇKA modelinin sınıflandırma performansı



Şekil 4.2: Tüm sınıf dengeleme yaklaşımları bazında AÖM modelinin sınıflandırma performansı



Şekil 4.3: Tüm sınıf dengeleme yaklaşımları bazında DVM modelinin sınıflandırma performansı

5. TARTIŞMA

Çalışmada kullanılan veri seti, HT bağımlı değişkeni açısından dengesiz sınıf problemi göstermekteydi. (HT'li hasta sayısı: 149 (%16), HT'li olmayan hasta sayısı: 780 (%84)) Bu durum, sınıflandırma algoritmalarının yanlı sonuç vermesine ve sonuçların hatalı yorumlanmasına neden olacağı için, örnekleme tabanlı sınıf dengeleme yöntemleri kullanılarak sınıf dengesizliği problemi giderilmeye çalışılmıştır. İlgili sınıf dengeleme yöntemleri uygulandıktan sonra ÇKA, AÖM ve DVM sınıflandırma modelleri uygulanarak, hem modellerin kendi aralarındaki sınıflandırma performansları incelenmiş, hem de uygulanan sınıf dengeleme yöntemlerinin, sınıflandırma modellerinin performansına katkısı irdelenmiştir. Hiçbir sınıf dengeleme yönteminin kullanılmadığı durumda, diğer bir deyişle sınıf dengesizliği bulunan orijinal veri setine sınıflandırma modelleri uygulandığı zaman, her 3 modelde de doğruluk, duyarlılık, seçicilik değerlerinin sırasıyla 0.84, 0.00, 1.00 olduğu, kesinlik, f-ölçümü ve g-ortalama değerlerinin ise hesaplanamadığı görülmüştür. Duyarlılık ölçütü değerlerinin 0.00 olarak hesaplanması, HT'li hastaların hiçbirinin söz konusu modeller tarafından doğru olarak tahmin edilemediği (sınıflandırılamadığı) anlamına gelmektedir. Modellerin doğruluk ölçütü değerlerine (0.84) bakıldığında ilgili modellerin başarılı bir sonuç ortaya koyduğu algısı oluşsa da, aslında bu doğruluk değerlerinin sadece HT'li olmayan hastaların doğru sınıflandırıldığından ileri geldiği görülmektedir. Bir başka deyişle, HT'li olmayan hastaların, HT'li olan hastalara göre sayıca üstün olması nedeniyle doğruluk ölçütü (dolayısıyla sınıflandırma modelleri) çoğunluk sınıfın etkisi altında kalmıştır ve her bir modelde yanlı sonuçlar vermiştir. Bu durum literatürde Doğruluk Çelişkisi (Accuracy Paradox) (56) olarak adlandırılmaktadır. Bu kapsamda, dengesiz sınıf problemi bulunan veri setlerinde, sınıflandırma performans ölçütü olarak doğruluğun tek başına ele alınmasının sakıncalı olduğu ortaya konulmuştur.

Alt örnekleme yöntemleri uygulandıktan sonra modellerin sınıflandırma performansları incelendiğinde, Tomek Link ve CNN yöntemleriyle veri setinden çıkarılan çoğunluk sınıf gözlemlerinin, üç modelinin de sınıflandırma performanslarına katkıda bulunmadığı ve sonuçların, hiçbir yöntem uygulanmadığında elde edilen sonuçlarla aynı olduğu görülmüştür. Aynı durum NCL yöntemi için AÖM ve DVM modelleri, OSS yöntemi için ÇKA ve AÖM modellerinde de görülmüştür. ÇKA modelinde NCL yöntemi uygulandıktan sonra sınıflandırma performans ölçüt değerlerinin, duyarlılık ölçüt değeri

(0.77) haricinde istenilen seviyede olmadığı tespit edilmiştir. RAÖ yöntemi uygulandıktan sonra elde edilen model sınıflandırma performansları incelendiğinde, bu yöntemin ÇKA model performansında ters etki yaptığı, duyarlılık değerini yükseltirken diğer ölçüt değerlerini düşürdüğü görülmüştür. AÖM ve DVM modellerinde RAÖ uygulandıktan sonra elde edilen sonuçların ÇKA modeline göre daha tutarlı olduğu fakat yine de istenilen seviyede olmadığı tespit edilmiştir. OSS modeli uygulandıktan sonra DVM modelinin sınıflandırma performans ölçüt değerleri, doğruluk, duyarlılık, seçicilik kesinlik, f-ölçümü ve g-ortalama için sırasıyla, 0.92, 0.54, 0.99, 0.92, 0.68 ve 0.71 olarak hesaplanmıştır. Her üç modelde de alt örnekleme yöntemleri ve performans ölçütleri bir bütün olarak ele alındığında, OSS yöntemi uygulandıktan sonra DVM modelinin en iyi sınıflandırma performansını gösterdiği görülmektedir. Bu sonuç, alt örnekleme yöntemleri içerisinde elde edilen en iyi sınıflandırma performansıdır.

Üst örnekleme yöntemleri açısından modellerin sınıflandırma performansları incelendiğinde, BLSMOTE, SLSMOTE ve ADASYN yöntemleri uygulandıktan sonra ÇKA modelinin performans çıktılarının, hiçbir yöntem uygulanmadığında elde edilen çıktılarla aynı olduğu ve modelin sınıflandırma performansına olumlu katkı yapmadıkları görülmüştür. SMOTE, DBSMOTE ve RÜÖ yöntemleri uygulandıktan sonra elde edilen sonuçların da yukarıda bahsedilen üst örnekleme yöntemlerinden elde edilen sonuçlarla küçük farklılıklar dışında benzer olduğu ve yine modelin sınıflandırma performansına olumlu katkı yapmadıkları tespit edilmiştir. AÖM modelinin sınıflandırma performansına bakıldığında, modelin tüm üst örnekleme yöntemlerinde ÇKA modeline göre daha kararlı sonuçlar verdiği görülmüştür. Fakat AÖM modelinin tüm üst örnekleme yöntemleri için elde ettiği en yüksek kesinlik, f-ölçümü ve g-ortalama değerlerinin sırasıyla 0.30, 0.40 ve 0.43'ü olması nedeniyle AÖM modelinin üst örnekleme yöntemleri uygulandıktan sınıflandırma performansının yeterli olmadığı görülmüştür. DVM modeli ele alındığında, RÜÖ yöntemi hariç, diğer tüm yöntemler uygulandıktan sonra elde edilen DVM sınıflandırma performans ölçüt değerlerinin tümünde 0.88 ve üzerinde olduğu tespit edilmiştir. RÜÖ yönteminde ise, RÜÖ uygulandıktan sonra elde edilen DVM performans ölçütlerinin tamamında elde edilen en küçük değerin 0.70 olduğu görülmüştür. DVM modelinin sınıflandırma performansı açısından performans metrikleri ve üst örnekleme yöntemleri bir bütün olarak ele alındığında DBSMOTE yönteminin diğer üst örnekleme yöntemlerine göre (performans ölçütleri açısından aralarında fazla farklılık olmamasıyla birlikte) daha iyi sonuçlar verdiği görülmüştür.

Hem alt hem de üst örnekleme yöntemi olan RAÜÖ yöntemi uygulandıktan sonra ÇKA ve AÖM modelleri, tüm performans ölçütleri ele alındığında benzer sınıflandırma performansı gösterdiği görülmektedir. DVM modelinin sınıflandırma performans ölçüt değerleri, doğruluk, duyarlılık, seçicilik kesinlik, f-ölçümü ve g-ortalama için sırasıyla, 0.92, 0.96, 0.92, 0.69, 0.8 ve 0.81 olarak hesaplandığından dolayı DVM, RAÜÖ yöntemi için en başarılı sınıflandırma modeli olmuştur.

Genel olarak, alt örnekleme yöntemleri uygulandığında zaman üç sınıflandırma modelinin de iyi sınıflandırma performansı gösteremedikleri, buna karşın üst örnekleme yöntemleri uygulandığında DVM modelinin diğer modellere göre bariz üstünlük gösterdiği görülmüştür. Bu durum hem alt hem de üst örnekleme yöntemi olan RAÜÖ yöntemi için de geçerlidir.

KAH'lı hastalarda mortalite ve morbiditenin en önemli sebeplerinden olan HT'nin erken teşhisi ve tedavisi büyük önem arz etmektedir. Bu kapsamda, bu çalışmada HT'nin tahmini belirli risk faktörleri açısından TBKS uygulanarak yapılmıştır. Çalışma kapsamında, üzerinde çalışılan veri setinde sınıf dengesizliği probleminin bulunması, bu problem giderilmeden uygulanacak TBKS çıktılarının yanıltıcı olacağı durumunu oraya çıkarmıştır. Bu sebeple sınıf dengesizliği probleminin giderilmesi için hem sınıf dengeleme analizlerini yapan hem de TBKS'ni uygulayan web-tabanlı ve dili Türkçe olan bir yazılım geliştirilmiştir. Bunlara ek olarak;

- Literatürde yapılan taramalarda sınıf dengeleme analizlerinin kapsamlı bir şekilde TBKS uygulanarak yapıldığı Türkçe kullanıcı ara yüzüne sahip bir yazılıma rastlanamamıştır.
- Yazılımda kullanılan sınıf dengeleme analizleri, çeşitli programlama dillerinde kod yazılarak uygulandığı ve ilgili programlama dillerinde belirli bir bilgi birikimine sahip olmayı gerektirdiği için kullanıcı dostu değillerdir.
- Sınıf dengesizliği problemi klinik veri setlerinde sıklıkla karşılaşılan bir durumdur ve geliştirilen yazılımın sadece bu çalışmada kullanılan veri setinde değil, sınıf dengesizliği probleminin görüldüğü tüm klinik veri setlerinde uygulanabilmesi mümkündür.

Tüm bu gerekçeler birlikte düşünüldüğünde, geliştirilen yazılımın ve bu yazılım vasıtasıyla elde edilen sonuçların literatüre katkı sağlayacağı düşünülmektedir.

6. SONUÇ VE ÖNERİLER

Çalışmadaki dengesiz sınıf problemini gidermek için uygulanan üst örnekleme yöntemlerinin, alt örnekleme yöntemlerine göre modellerin (özellikle DVM modelinin) sınıflandırma performanslarına belirgin bir şekilde olumlu katkı yaptığı tespit edilmiştir. Elde edilen sınıflandırma performans çıktıları ışığında, hem gösterdiği düşük sınıflandırma performansı hem de veri setinden gözlem çıkararak bilgi kaybına yol açması nedeniyle, alt örnekleme yöntemlerinin klinik araştırmalarda uygulanması önerilmemektedir. Üst örnekleme yöntemlerinde özellikle SMOTE ve SMOTE tabanlı yöntemlerin (BLSMOTE, SLSMOTE, DBSMOTE, ADASYN, vb.), hem analitik yöntemler olması hem de sınıflandırma performanslarına yaptıkları katkı nedeniyle kullanımı önerilmektedir. Fakat bazı çalışmalarda (57, 58), SMOTE yönteminin çoğunluk sınıfı ile azınlık sınıfı kümeleri arasındaki ayrım çoğu zaman net olmadığı için gürültülü veriler türetilebileceği yönünde dezavantajlara sahip olduğu, ayrıca sentetik gözlemler türetirken sadece azınlık sınıfı gözlemlerini dikkate aldığı için aşırı uyum ve aşırı genelleme sorunu ortaya çıkarabileceği belirtilmiştir. Bu nedenle SMOTE ve SMOTE türevlerinin uygulanmadan önce yukarıda belirtilen durumların göz önünde bulundurulması gerekmektedir.

Bu çalışma kapsamında yer almayan ancak ilerleyen çalışmalarda ele alınacak olan;

- Hem alt hem de üst örnekleme yöntemlerini uygulayarak bilgi kaybı ve aşırı uyum sorunlarının dengelediği hibrit yöntemleri,
- Dengesiz sınıf probleminin sınıflandırma modeliyle birlikte ele alındığı Maliyet-Duyarlı Öğrenme (Cost-Sensitive Learning) Tabanlı Yöntemler, Topluluk Öğrenme (Ensemble Learning) Tabanlı Yöntemler, Öznitelik Seçimi (Feature Selection) Tabanlı Yöntemler,

daha sağlam ve tutarlı sonuçlar elde edilmesi açısından okuyuculara önerilebilir.

KAYNAKLAR

1. Çolak C, Çolak MC, Orman MN. Koroner arter hastalığının tahmininde lojistik regresyon modeli seçim yöntemlerinin karşılaştırılması. *Anatolian J Cardiol* 2007, 7(1).

2. Olafiranye O, Zizi F, Brimah P, Jean-louis G, Makaryus AN, McFarlane S, Ogedegbe G. Management of hypertension among patients with coronary heart disease. *Int J Hypertens* 2011, 2011.

3. Rosendorff C, Lackland DT, Allison M, Aronow WS, Black HR, Blumenthal RS, Cannon CP, De Lemos JA, Elliott WJ, Findeiss L. Treatment of hypertension in patients with coronary artery disease. *Hypertension* 2015, 65(6): 1372-407.

4. Taner A. Koroner Arter Hastalarında Hipertansiyonun İnflamatuar Reaksiyona Additif Etkisi. *Kocatepe Tıp Dergisi* 2013, 14(3).

5. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI mag* 1996, 17(3): 37.

6. Sarmanova A. Veri madenciliğindeki sınıf dengesizliği sorununun giderilmesi. Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı. Yüksek Lisans Tezi, İstanbul: Yıldız Teknik Üniversitesi, 2013.

7. Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. *Comput Methods Programs Biomed* 2016, 13087-92.

8. Nanni L, Fantozzi C, Lazzarini N. Coupling different methods for overcoming the class imbalance problem. *Neurocomputing* 2015, 15848-61.

9. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explor Newsl* 2004, 6(1): 1-6.

10. Sarmanova A, Albayrak S, editors. Alleviating class imbalance problem in data mining. Signal Process Commun Appl Conf (SIU); 2013: IEEE.

11. Bekkar M, Alitouche TA. Imbalanced data learning approaches review. *Int J Data Min Knowl Manag Process* 2013, 3(4): 15.

12. He H, Ma Y. *Imbalanced learning: foundations, algorithms, and applications*, John Wiley & Sons 2013.

13. Tomek I. An experiment with the edited nearest-neighbor rule. *IEEE Trans syst Man Cybern* 1976, (6): 448-52.

14. Hart P. The condensed nearest neighbor rule. *IEEE Trans Inf Theory* 1968, 14(3): 515-6.

15. Dal Pozzolo A, Caelen O, Waterschoot S, Bontempi G, editors. Racing for unbalanced methods selection. Int Conf Intell Data Eng Autom Learn; 2013: Springer.

16. Kubat M, Matwin S, editors. Addressing the curse of imbalanced training sets: one-sided selection. ICML; 1997: Nashville, USA.

17. Durahim AO. Comparison of Sampling Techniques for Imbalanced Learning. *Yönetim Bilişim Sistemleri Dergisi* 2016, 1(3): 181-91.

18. Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans syst Man Cybern* 1972, 2(3): 408-21.

19. García-Borroto M, Villuendas-Rey Y, Carrasco-Ochoa JA, Martínez-Trinidad JF, editors. Using maximum similarity graphs to edit nearest neighbor classifiers. Iberoamerican Congr Pattern Recognit; 2009: Springer.

20. Laurikkala J. Improving identification of difficult small classes by balancing class distribution. *Artif Intell Med* 2001, 63-6.

21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002, 16321-57.

22. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans knowl data eng* 2009, 21(9): 1263-84.

23. Zheng Z, Cai Y, Li Y. Oversampling method for imbalanced classification. *Comput Inform* 2016, 34(5): 1017-37.

24. Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. *Adv intell comput* 2005, 878-87.

25. Verbiest N, Ramentol E, Cornelis C, Herrera F. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. *Appl Soft Comput* 2014, 22511-7.

26. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. *Adv knowl discov data min* 2009, 475-82.

27. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. DBSMOTE: Density-Based Synthetic Minority Over-sampling TEchnique. *Appl Intell* 2012, 36(3): 664-84.

28. Ester M, Kriegel H-P, Sander J, Xu X, editors. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd; 1996.

29. Siriseriwan W, Sinapiromsaran K. Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling. *Songklanakarin J Sci Technol* 2017, 39(5).

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

31. Siedhoff D. ADASYN (improves class balance, extension of SMOTE). MATLAB Central; 2015.

32. Kohavi R, John GH. The wrapper approach. Feature extraction, construction and selection: Springer; 1998. p. 33-50.

33. Alpar CR. *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. Ankara, Detay Yayıncılık 2013.

34. Sümbüloğlu V, Sümbüloğlu K. *Klinik saha araştırmalarında örnekleme yöntemleri ve örneklem büyüklüğü*. Ankara, Hatiboğlu Yayınevi 2005.

35. Kuhn M. Caret package. *Journal of Statistical Software* 2008, 28(5): 1-26.

36. Siriseriwan W. smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE. 2016.

37. Dal Pozzolo A, Caelen O, Waterschoot S, Bontempi G, editors. Racing for unbalanced methods selection. International Conference on Intelligent Data Engineering and Automated Learning; 2013: Springer.

38. Lunardon N, Menardi G, Torelli N. ROSE: A Package for Binary Imbalanced Learning. *R Journal* 2014, 6(1).

39. Kartal E, Balaban ME. Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama.

40. https://shiny.rstudio.com 02.12.2017

41. Silva JdA, Hruschka ER. An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. *Data Knowl Eng* 2013, 84(Supplement C): 47-58.

42. Yılmaz H. Random Forests yönteminde kayıp veri probleminin incelenmesi ve sağlık alanında bir uygulama. Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı. Yüksek Lisans Tezi, Eskişehir: Osmangazi Üniversitesi, 2014.

43. Breiman L. Random forests. *Mach learn* 2001, 45(1): 5-32.

44. Yu-Wei CDC. *Machine learning with R cookbook*, Packt Publishing Ltd 2015.

45. Tang F, Ishwaran H. Random forest missing data algorithms. *Stat Anal Data Min: ASA Data Sci J* 2017.

46. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. Random forests for classification in ecology. *Ecology* 2007, 88(11): 2783-92.

47. Luceño A. Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators. *Comput Stat Data An* 2006, 51(2): 904-17.

48. Ledolter J. *Data mining and business analytics with R*, John Wiley & Sons 2013.

49. Riedmiller M, Lernen AM. Multi Layer Perceptron. *Machine Learning Lab Special Lecture, University of Freiburg* 2014.

50. Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. *Neurocomputing* 2006, 70(1): 489-501.

51. Kaya Y, Tekin R. Epileptik Nöbetlerin Tespiti için Aşırı Öğrenme Makinesi Tabanlı Uzman Bir Sistem. *Int J Inform Technol* 2012, 5(2).

52. Akın M, Ceylan M, editors. Comparison of artificial neural network and extreme learning machine in benign liver lesions classification. Med Technol Natl Conf (TIPTEKNO); 2015: IEEE.

53. Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995, 20(3): 273-97.

54. Ayhan S, Erdoğmuş Ş. Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi* 2014, 9(1).

55. http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html 27.11.2017

56. Russell J, Cohn R. *Accuracy Paradox*, Book on Demand 2012.

57. Douzas G, Bacao F. Geometric SMOTE: Effective oversampling for imbalanced learning through a geometric extension of SMOTE. *arXiv preprint arXiv:170907377* 2017.

58. Theeramunkong T, Kijsirikul B, Cercone N, Ho T-B. *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings*, Springer 2009.

EKLER

EK-1. ÖZGEÇMİŞ

ÖZGEÇMİŞ VE ESERLER LİSTESİ

ÖZGEÇMİŞ

Adı Soyadı: Ahmet Kadir ARSLAN

Doğum Tarihi: 1988

Öğrenim Durumu: Yüksek Lisans

Derece	Bölüm/Program	Üniversite	Yıl
Lisans	Matematik	Afyon Kocatepe Üniversitesi	2010
Y. Lisans	Biyoistatistik ve Tıp Bilişimi AD	İnönü Üniversitesi	2018

Yüksek Lisans Tez Başlığı ve Tez Danışman(lar)ı:

Koroner Arter Hastalarında Hipertansiyonun Sınıflandırılması İçin Dengesiz Sınıf Probleminin Tıbbi Bilgi Keşfi Süreci ile Giderilmesi, Prof. Dr. Cemil ÇOLAK

Görevler:

Görev Unvanı	Görev Yeri	Yıl
Arş. Gör.	İnönü Üniversitesi	2014 – Devam ediyor
Uzman Yardımcısı	Türkiye Halk Bankası A.Ş.	2012 - 2014

Projelerde Yaptığı Görevler:

Obez Bireylerde Atriyal Fibrilasyonu Etkileyebilecek Faktörlerinin Belirlenmesi, Yükseköğretim Kurumları tarafından destekli bilimsel araştırma projesi (BAP), Araştırmacı, 15.02.2016 - 30.12.2016

Ödüller:

Poster birinciliği ödülü, XVIII. Ulusal ve I. Uluslararası Biyoistatistik Kongresi, Biyoistatistik Kongresi-Afyon Kocatepe Üniversitesi, 2016

ESERLER

A. Uluslararası hakemli dergilerde yayımlanan makaleler :

A1. GÜLDOĞAN EMEK, ARSLAN AHMET KADİR, ÇOLAK MEHMET CENGİZ, ÇOLAK CEMİL, ERDİL NEVZAT (2017). An intelligent system for the classification of postoperative pleural effusion between 4 and 30 days using medical knowledge discovery. BIOMEDICAL RESEARCH-INDIA, 28(4), 1553-1556.

A2. ÇOLAK MEHMET CENGİZ, KARAASLAN EROL, ÇOLAK CEMİL, ARSLAN AHMET KADİR, ERDİL NEVZAT (2017). Handling imbalanced class problem for the prediction of atrial fibrillation in obese patient. BIOMEDICAL RESEARCH-INDIA, 28(7), 3293-3299.

A3. ÖZDEMİR RAMAZAN, YAĞMUR JÜLİDE, AÇIKGÖZ NUSRET, CANSEL MEHMET, KARINCAOĞLU YELDA, ERMİŞ NECİP, PEKDEMİR HASAN, ARSLAN AHMET KADİR (2017). Relationship between serum homocysteine levels and structural-functional carotid arterial abnormalities in inactive Behçet's disease. Kardiologia Polska (Polish Heart Journal), Doi: 10.5603/KP.a2017.0227

A4. ARSLAN AHMET KADİR, ÇOLAK CEMİL, SARIHAN MEHMET EDİZ (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer Methods and Programs in Biomedicine, 130, 87-92., Doi: 10.1016/j.cmpb.2016.03.022

A5. FIRAT FEYZA, ARSLAN AHMET KADİR, ÇOLAK CEMİL, HARPUTLUOĞLU HAKAN (2016). Estimation of risk factors associated with colorectal cancer an application of knowledge discovery in databases. Kuwait Journal of Science, 43(2), 151-161.

A6. ÇOLAK CEMİL, AYDOĞAN MUSTAFA SAİD, ARSLAN AHMET KADİR, YÜCEL AYTAÇ (2015). Application of Medical Data Mining on the Prediction of APACHE II Score. Medicine Science | International Medical Journal, 1, Doi: 10.5455/medscience.2015.04.8274

B. Uluslararası bilimsel toplantılarda sunulan ve bildiri kitabında (*Proceedings*) basılan bildiriler :

B1. ARSLAN AHMET KADİR, ÇOLAK CEMİL (2017). Sınıf Dengeleyici: İki Sınıflı Verilerde Sınıf Dengesizliği Problemini Gidermek İçin Web Tabanlı Bir Yazılım. XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi (Sözlü Sunum)

B2. ARSLAN AHMET KADİR, ÇUĞLAN SONGÜL, KÖSE EVREN, ÇOLAK CEMİL (2017). KRONİK OBSTRUKTİF AKCİĞER HASTALIĞININ ÇEŞİTLİ MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANILARAK SINIFLANDIRILMASI. XIX. Ulusal ve II. Uluslararası Biyoistatistik Kongresi (Kısa Sözlü Sunum)

B3. ARSLAN AHMET KADİR, GÜLDOĞAN EMEK, ÇOLAK CEMİL (2017). KDAY: Kayıp Değer Analizinde Kullanılan Çeşitli Tekniklerin Performansını Karşılaştıran Web Tabanlı Bir Yazılım. XVIII. Uluslararası Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu (Sözlü Sunum)

B4. GÜLDOĞAN EMEK, ARSLAN AHMET KADİR, ÇOLAK CEMİL, YAĞMUR JÜLİDE (2017). Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama. XVIII. Uluslararası Ekonometri, Yöneylem Araştırması ve İstatistik Sempozyumu (Sözlü Sunum)

B5. ÇOLAK MEHMET CENGİZ, KARAASLAN EROL, ÇOLAK CEMİL, ARSLAN AHMET KADİR, ERDİL NEVZAT (2016). Handling Imbalanced Class Problem for the Prediction of Atrial Fibrillation in Obese Patients. I. International Biostatistics Congress (Poster Sunum)

B5. GÜLDOĞAN EMEK, ARSLAN AHMET KADİR, ÇOLAK MEHMET CENGİZ, ÇOLAK CEMİL, ERDİL NEVZAT (2016). An Intelligent System for the Classification of Postoperative Pleural Effusion between 4 and 30 days using Medical Knowledge Discovery. I. International Biostatistics Congress (Poster Sunum)

B6. ARSLAN AHMET KADİR, GÜLDOĞAN EMEK, ÇOLAK CEMİL (2016). Makine Öğrenmesi Yaklaşımlarından Aşırı Öğrenme Makinesinin Sınıflandırma Performansının Değerlendirilmesi: Bir Simülasyon Çalışması. I. International Biostatistics Congress (Poster Sunum)

B7. ARSLAN AHMET KADİR, ÇOLAK CEMİL, SARIHAN MEHMET EDİZ (2016). Different medical data mining approaches based prediction of ischemic stroke. 17. International Symposium on Econometrics, Operations Research and Statistics (Poster Sunum)

B8. FIRAT FEYZA, ARSLAN AHMET KADİR, ÇOLAK CEMİL, HARPUTLUOĞLU HAKAN (2016). Estimation of risk factors associated with colorectal cancer an application of knowledge discovery in databases. 17. International Symposium on Econometrics, Operations Research and Statistics (Poster Sunum)

B9. ÇOLAK MEHMET CENGİZ, ÇOLAK CEMİL, ERDİL NEVZAT, ARSLAN AHMET KADİR (2016). Investigatıng Optimal Number of Cross Validation on The Predictıon of Postoperatıve Atrial Fibrillation By Voting Ensemble Strategy. 17. International Symposium on Econometrics, Operations Research and Statistics (Poster Sunum)

D. Ulusal hakemli dergilerde yayımlanan makaleler:

D1. GÜLDOĞAN EMEK, ARSLAN AHMET KADİR, YAĞMUR JÜLİDE (2017). Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama. Fırat Tıp Dergisi, 22(3), 136-142.

D2. ÇOLAK MEHMET CENGİZ, ÇOLAK CEMİL, ERDİL NEVZAT, ARSLAN AHMET KADİR (2016). Investigating Optimal Number of Cross Validation on the Prediction of Postoperative Atrial Fibrillation by Voting Ensemble Strategy. Turkiye Klinikleri Journal of Biostatistics, 8(1), 30-35., Doi: 10.5336/biostatic.2016-50382

E. Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler:

E1. ÇOLAK CEMİL, AYDOĞAN MUSTAFA SAİD, ARSLAN AHMET KADİR, YÜCEL AYTAÇ (2015). APACHE II skorunun tahmininde bir Tıbbi Veri Madenciliği Uygulaması. XVII. Ulusal Biyoistatistik Kongresi (Poster Sunum)

EK-2. ETİK KURUL ONAY FORMU

KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	Koroner arter hastalarında hipertansiyonun sınıflandırılması için dengesiz sınıf probleminin tıbbi bilgi keşfi süreci ile giderilmesi.
VARSA ARAŞTIRMANIN PROTOKOL KODU	2016/162

ETİK KURUL BİLGİLERİ	ETİK KURULUN ADI	MALATYA KLİNİK ARAŞTIRMALAR ETİK KURULU
	AÇIK ADRESİ:	İnönü Üniversitesi Merkez Kampüsü, 44280, Malatya, Türkiye
	TELEFON	+90 422 341 06 60 / 1219
	FAKS	+90 422 341 00 36
	E-POSTA	inu.dhek@inonu.edu.tr

BAŞVURU BİLGİLERİ	KOORDİNATÖR/SORUMLU ARAŞTIRMACI UNVANI/ADI/SOYADI	Doç. Dr. Cemil Çolak			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACININ UZMANLIK ALANI	İnönü Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıp Bilişimi AD			
	KOORDİNATÖR/SORUMLU ARAŞTIRMACININ BULUNDUĞU MERKEZ	MALATYA			
	VARSA İDARİ SORUMLU UNVANI/ADI/SOYADI				
	DESTEKLEYİCİ				
	PROJE YÜRÜTÜCÜSÜ UNVANI/ADI/SOYADI (TÜBİTAK vb. gibi kaynaklardan destek alanlar için)				
	DESTEKLEYİCİNİN YASAL TEMSİLCİSİ				
	ARAŞTIRMANIN FAZI VE TÜRÜ	FAZ 1	☐		
		FAZ 2	☐		
		FAZ 3	☐		
		FAZ 4	☐		
		Gözlemsel ilaç çalışması	☐		
		Tıbbi cihaz klinik araştırması	☐		
		İn vitro tıbbi tanı cihazları ile yapılan performans değerlendirme çalışmaları	☐		
		İlaç dışı klinik araştırma	☐		
		Diğer ise belirtiniz			
	ARAŞTIRMAYA KATILAN MERKEZLER	TEK MERKEZ ☐	ÇOK MERKEZLİ ☐	ULUSAL ☐	ULUSLARARASI ☐

Etik Kurul Başkanının
Unvanı/Adı/Soyadı: Prof. Dr. Rıfat KARLIDAĞ
İmza:

Not: Etik kurul başkanının her sayfada imzasının olması gerekmektedir.

KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	Koroner arter hastalarında hipertansiyonun sınıflandırılması için dengesiz sınıf probleminin tıbbi bilgi keşfi süreci ile giderilmesi.
VARSA ARAŞTIRMANIN PROTOKOL KODU	2016/162

	Belge Adı	Tarihi	Versiyon Numarası	Dili
DEĞERLENDİRİLEN BELGELER	ARAŞTIRMA PROTOKOLÜ			Türkçe ☐ İngilizce ☐ Diğer ☐
	BİLGİLENDİRİLMİŞ GÖNÜLLÜ OLUR FORMU			Türkçe ☐ İngilizce ☐ Diğer ☐
	OLGU RAPOR FORMU			Türkçe ☐ İngilizce ☐ Diğer ☐
	ARAŞTIRMA BROŞÜRÜ			Türkçe ☐ İngilizce ☐ Diğer ☐

	Belge Adı		Açıklama
DEĞERLENDİRİLEN DİĞER BELGELER	SİGORTA	☐	
	ARAŞTIRMA BÜTÇESİ	☐	
	BİYOLOJİK MATERYEL TRANSFER FORMU	☐	
	İLAN	☐	
	YILLIK BİLDİRİM	☐	
	SONUÇ RAPORU	☐	
	GÜVENLİLİK BİLDİRİMLERİ	☐	
	DİĞER:	☐	

KARAR BİLGİLERİ	**Karar No:2016/162**	**Tarih:10.08.2016**
	Yukarıda bilgileri verilen başvuru dosyası ile ilgili belgeler araştırmanın/çalışmanın gerekçe, amaç, yaklaşım ve yöntemleri dikkate alınarak incelenmiş ve uygun bulunmuş olup araştırmanın/çalışmanın başvuru dosyasında belirtilen merkezlerde gerçekleştirilmesinde etik ve bilimsel sakınca bulunmadığına toplantıya katılan etik kurul üye tam sayısının salt çoğunluğu ile karar verilmiştir. İlaç ve Biyolojik Ürünlerin Klinik Araştırmaları Hakkında Yönetmelik kapsamında yer alan araştırmalar/çalışmalar için Türkiye İlaç ve Tıbbi Cihaz Kurumu'ndan izin alınması gerekmektedir.	
KLİNİK ARAŞTIRMALAR ETİK KURULU		
ETİK KURULUN ÇALIŞMA ESASI	İlaç ve Biyolojik Ürünlerin Klinik Araştırmaları Hakkında Yönetmelik, İyi Klinik Uygulamaları Kılavuzu	
BAŞKANIN UNVANI / ADI / SOYADI:	**Prof. Dr. Rıfat KARLIDAĞ**	

Unvanı/Adı/Soyadı	Uzmanlık Alanı	Kurumu	Cinsiyet		Araştırma ile ilişki		Katılım *		İmza
Prof. Dr. Rıfat KARLIDAĞ	Psikiyatri	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☐	E ☐	H ☐	
Prof. Dr. Metin GENÇ	Halk Sağlığı	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☐	E ☐	H ☐	
Prof. Dr. Saim YOLOĞLU	Biyoistatistik	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☐	E ☐	H ☐	
Prof. Dr. Türkan TOĞAL	Anesteziyoloji ve Rea.	İnönü Üniversitesi Tıp Fakültesi	E ☐	K ☒	E ☐	H ☐	E ☐	H ☐	
Prof. Dr. İbrahim ŞAHİN	İç Hastalıkları	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☐	E ☐	H ☐	
Prof. Dr. Sedat YILDIZ	Fizyoloji	İnönü Üniversitesi Tıp Fakültesi	E ☒	K ☐	E ☐	H ☐	E ☐	H ☐	
Doç. Dr. Seda TAŞDEMİR	Tıbbi Farmakoloji	İnönü Üniversitesi Tıp Fakültesi	E ☐	K ☒	E ☐	H ☐	E ☐	H ☐	

Etik Kurul Başkanının
Unvanı/Adı/Soyadı: Prof. Dr. Rıfat KARLIDAĞ
İmza:

Not: Etik kurul başkanının her sayfada imzasının olması gerekmektedir.

KLİNİK ARAŞTIRMALAR ETİK KURULU KARAR FORMU

ARAŞTIRMANIN AÇIK ADI	Koroner arter hastalarında hipertansiyonun sınıflandırılması için dengesiz sınıf probleminin tıbbi bilgi keşfi süreci ile giderilmesi.
VARSA ARAŞTIRMANIN PROTOKOL KODU	2016/162

Doç. Dr. Derya DOĞAN	Çocuk Sağlığı ve Hast.	İnönü Üniversitesi Tıp Fakültesi	E☐	K☒	E☐	H☐	E☐	H☐	Katılmadı
Doç. Dr. Özden KAMIŞLI	Nöroloji	İnönü Üniversitesi Tıp Fakültesi	E☐	K☒	E☐	H☒	E☒	H☐	
Doç. Dr. Hakan HARPUTLUOĞLU	Onkoloji	İnönü Üniversitesi Tıp Fakültesi	E☒	K☐	E☐	H☐	E☐	H☐	
Yrd. Doç. Dr. Mehmet KARATAŞ	Tıp Tarihi ve Etik	İnönü Üniversitesi Tıp Fakültesi	E☒	K☐	E☐	H☐	E☐	H☐	Katılmadı
Dr. Mahmut Bakın AKGÜL	Tıp Doktoru	Halk Sağlığı Müdürlüğü	E☒	K☐	E☐	H☒	E☒	H☐	
Metin TAY	Eczacı	Serbest Eczacı	E☒	K☐	E☐	H☐	E☐	H☐	Katılmadı
Zafer ERGÜZEL	Hukuk	İnönü Üniversitesi	E☒	K☐	E☐	H☐	E☐	H☐	Katılmadı
Hasan KONAN	Sivil Üye	MSD Ltd. Şti.	E☒	K☐	E☐	H☐	E☐	H☐	

Etik Kurul Başkanının
Unvanı/Adı/Soyadı: Prof. Dr. Rıfat KARLIDAĞ
İmza:

Not: Etik kurul başkanının her sayfada imzasının olması gerekmektedir.