T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

MEDİKAL VERİ SINIFLANDIRMADA SENTETİK VERİ ÜRETİMİNİN ETKİSİ

Mehmet ŞEYHANLI

ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

ŞANLIURFA
2021

T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

MEDİKAL VERİ SINIFLANDIRMADA SENTETİK VERİ ÜRETİMİNİN ETKİSİ

Mehmet ŞEYHANLI

ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

ŞANLIURFA
2021

İÇİNDEKİLER

Sayfa No

ÖZET i
ABSTRACT ii
TEŞEKKÜR iii
ŞEKİLLER DİZİNİ iv
ÇİZELGELER DİZİNİ v
SİMGELER ve KISALTMALAR DİZİNİ vi
1. GİRİŞ 1
2. ÖNCEKİ ÇALIŞMALAR 4
3. MATERYAL ve YÖNTEM 12
3.1. Medikal Veri Setleri 12
3.1.1. Ortopedi Veri Seti 12
3.1.2. Meme Kanseri Veri Seti 16
3.3.2. Ameliyat Sonrası Bakım Veri Seti 19
3.2. Sentetik Veri Oluşturma Yöntemleri 22
3.2.1. Sentetik Azınlık Aşırı Örnekleme Tekniği (SAAÖT) 24
3.2.2. Uyarlanabilir Sentetik Örnekleme (USÖ) 26
3.3. Sınıflandırma Yöntemleri 28
3.3.2. Başarı Ölçütleri 31
4. ARAŞTIRMA BULGULARI ve TARTIŞMA 34
4.1. Araştırma Bulguları 34
4.1.1. Ortopedi Veri Seti 34
4.1.2. Meme Kanseri Veri Seti 37
4.1.3. Ameliyat Sonrası Bakım Veri Seti 40
5. SONUÇLAR ve ÖNERİLER 44
KAYNAKLAR 46
ÖZGEÇMİŞ 49

ÖZET

Yüksek Lisans Tezi

MEDİKAL VERİ SINIFLANDIRMADA
SENTETİK VERİ ÜRETİMİNİN ETKİSİ

Mehmet ŞEYHANLI

Harran Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik Elektronik Mühendisliği

Danışman: Dr. Öğr. Üyesi Abdülkadir GÜMÜŞÇÜ
YIL: 2021, Sayfa: 49

Veri madenciliği, son yıllarda birçok alanda yaygın olarak kullanılmaktadır. Veri sınıflandırmalarında, örnek sayıların sınıflara göre eşit dağılımı çok önemlidir. Ancak, bazı durumlarda veri setinde yer alan sınıflar eşit dağılıma sahip olmayabilir. Özellikle veri oluşturma aşamasında zorluk yaşanan medikal veri setleri bu probleme verilecek örnekler arasında yer almaktadır. Bu gibi durumlarda, sınıflandırma başarı parametreleri olumsuz yönde etkilenir. Bu soruna bir çözüm olarak, yeni sentetik veri örnekleri üretilerek veri seti dengeli bir hale getirilebilir. Bu çalışmada tıbbi rahatsızlıklara ait nümerik, kategorik ve hem nümerik hem de kategorik olan veri setlerinin sınıflandırılmasında dengesiz veriyi oluşturma yöntemlerinin etkisi tartışılmıştır. Bu veri setlerine ait dengesizlik problemleri iki farklı sentetik örnekleme yaklaşımı yöntemi uygulanarak çözümlenmiştir. Veri setine, sentetik örnekleme yaklaşımı yöntemleri uygulandıktan sonra, Karar Ağacı (KA) ve K- En Yakın Komşuluk (KEYK) sınıflandırma yöntemleri uygulanmıştır. Böylelikle, sentetik veri üretiminin farklı veri türleri üzerindeki etkisi gözlemlenmiştir. Sonuçlar incelendiğinde; sadece kategorik ve hem kategorik hem sayısal veri tipine sahip veri setlerinde en başarılı sonuçların Sentetik Azınlık Aşırı Örnekleme Tekniği (SAAÖT) algoritmasında KA sınıflandırılmasında, sadece numerik veri setinde en başarılı sonucun ise Uyarlamalı Sentetik Örnekleme (USÖ) algoritmasında, k en yakın komşuluk sınıflandırılmasında elde edildiği gözlemlenmiştir.

ANAHTAR KELİMELER: SMOTE, Sentetik Veri, Sınıflandırma, Medikal

ABSTRACT

MSc Thesis

EFFECT OF SYNTHETIC GENERATION ON MEDİCAL DATA CLASSIFICATION

Mehmet ŞEYHANLI

Harran University
Graduate School of Natural and Applied Sciences
Department of Electrical and Electronic Engineering

Supervisor: Assist. Prof. Dr. Abdülkadir GÜMÜŞÇÜ
Year: 2021, Page: 49

Data mining has been widely used in many areas in recent years. In data classification, equal distribution of sample numbers according to classes is very important. However, in some cases, the classes in the data set may not be equally distributed. Medical datasets, which have difficulties especially in data generation, are among the examples to be given to this problem. In such cases, the classification success parameters are adversely affected. As a solution to this problem, the data set can be balanced by producing new synthetic data samples. In this study, the effect of synthetıc generation on medical data classification of numerical, categorical and both numerical and categorical data sets of medical diseases are discussed. The imbalance problems of these data sets have been solved by applying two different adaptive synthetic sampling approach methods. After the adaptive synthetic sampling approach methods were applied to the data set, Decision Tree (DT) and K-Nearest Neighbors (KNN) classification methods were applied. Thus, the effect of synthetic data generation on different data types has been observed. When the results are examined; It was observed that the most successful results were obtained in the decision tree classification in the Synthetic Minority Oversampling Technique (SMOTE) algorithm in the data sets with only categorical and both categorical and numerical types, and the most successful result in only the numerical data set was obtained in the Adaptive Synthetic Sampling Method (ADASYN) algorithm and k nearest neighbors classification.

KEY WORDS: SMOTE, Synthetic Data, Classification, Medical

TEŞEKKÜR

Bu tez çalışmasının hazırlanmasında bana her türlü desteği veren, bilgi ve deneyimlerini titizlikle aktaran değerli danışman hocam Dr. Öğr. Üyesi Abdülkadir GÜMÜŞÇÜ'ye, maddi manevi destekleri ile her zaman yanımda olan aileme ve arkadaşlarıma sonsuz teşekkürlerimi sunarım.

ŞEKİLLER DİZİNİ

Sayfa No

Şekil 3.1. Pelvik insidens açısı, FB: Femur başı (Karademir ve ark., 2014) 13
Şekil 3.2. Lomber Lordoz Açısı (LLA) (Esmailiejah ve ark., 2017) 14
Şekil 3.3. Orijinal veri seri ve USÖ uygulanmış veri setine ait dağılım 27
Şekil 3.4. K en yakın komşuluk tespiti 29
Şekil 4.1. Sentetik Veri Oluşturma 34

ÇİZELGELER DİZİNİ

Sayfa No

Çizelge 3.1. Ortopedi veri seti öznitelikleri 12
Çizelge 3.2. Meme kanseri veri seti öznitelikleri 16
Çizelge 3.3. Ameliyat sonrası bakım veri seti öznitelikleri 20
Çizelge 4.1. Ortopedi Veri Seti SAAÖT Sonrası Sınıf Dağılımları 35
Çizelge 4.2. Ortopedi Veri Seti USÖ Sonrası Sınıf Dağılımları 35
Çizelge 4.3. Ortopedi Veri Seti Veri Seti Sınıflandırma Tablosu 35
Çizelge 4.4. Ortopedi Veri Seti KEYK Başarı Ölçütleri 36
Çizelge 4.5. Ortopedi Veri Seti KA Başarı Ölçütleri 36
Çizelge 4.6. Meme Kanseri Veri Seti SAAÖT Sonrası Sınıf Dağılımları 37
Çizelge 4.7. Meme Kanseri Veri Seti USÖ Sonrası Sınıf Dağılımları 38
Çizelge 4.8. Meme Kanseri Veri Seti Sınıflandırma Tablosu 38
Çizelge 4.9. Meme Kanseri Veri Seti KEYK Başarı Ölçütleri 39
Çizelge 4.10. Meme Kanseri Veri Seti KA Başarı Ölçütleri 39
Çizelge 4.11. Ameliyat Sonrası Bakım Veri Seti SAAÖT Sonrası Sınıf Dağılımları 40
Çizelge 4.12. Ameliyat Sonrası Bakım Veri Seti USÖ Sonrası Sınıf Dağılımları 41
Çizelge 4.13. Ameliyat Sonrası Bakım Veri Seti Sınıflandırma Tablosu 41
Çizelge 4.14. Ameliyat Sonrası Bakım Veri Seti KEYK Başarı Ölçütleri 42
Çizelge 4.15. Ameliyat Sonrası Bakım Veri Seti KA Başarı Ölçütleri 42

SİMGELER ve KISALTMALAR DİZİNİ

AB	AdaBoost
BDT	Bilgisayar Destekli Tasarım
BT	Bilgisayarlı Tomografi
DSÇ	Dinamik Süper Çözünürlük
DVM	Destek Vektör Makinası
EMBSO	En az Mutlak Büzülme ve Seçim Operatörü
EMG	Elektromiyografi
GBS	Geçiş Başlatma Siteleri
KA	Karar Ağacı
KEYK	K En Yakın Komşuluk
MR	Manyetik Rezonans
MRG	Manyetik Rezonans Görüntüleme
NB	Naif Bayes
OKK	Ortalama Kareli Kayma
RAİS	Resim Arşivleme İletişim Sistemleri
RAY	Rastgele Ağaçlar Yöntemi
SAAÖT	Sentetik Azınlık Aşırı Örnekleme Tekniği
TIS	Transition Initiation Sites
TSA	Tamamlayıcı Sinir Ağı
UG	Uyarlamalı Güçlendirme
YSA	Yapay Sinir Ağı

1. GİRİŞ

Son dönemlerde yapılan çalışmalar incelendiğinde yapay zekânın tıbbi uygulamalar üzerinde kullanımının yaygınlaştığı görülmektedir (Akben, 2016). Yapay zekânın tıpla entegrasyonu sonucunda hastalıkların tanı ve teşhisinde, tıbbi görüntülerin yorumlanmasında, hasta izleme ve kontrolünde, ilaç üretiminde ve birçok tıp alanında gelişmeler sağladığı bilinmektedir.

Özellikle medikal veri setlerinin toplanması süreci oldukça sıkıntılı ve zor bir süreçtir (Torkaman, 2015). Hastaların kişisel sağlık verileri, hastalığın seyrine göre hastaya ait sağlık değerlerinin kaydedilmesiyle elde edilir. Aynı süreç sağlıklı bireylerde daha farklı seyretmektedir. Ayrıca elde edilecek olan verilerde hastaların yaş, ırk, cinsiyet vb. özelliklerine bağlı olarak süreç değişkenlik gösterebilir. Veri setleri oluşturulurken yaşanan tüm bu sıkıntılar medikal veri setlerine ait örnek sayılarının, az sayıda olmasına neden olmaktadır. Çünkü incelenen kişilerin hastalıklı olma olasılığı, çoğunlukla düşük ihtimallidir. Bu nedenle genellikle medikal veri setlerinde, sınıflar arasında dengesiz bir dağılım söz konusu olmaktadır.

Bu çalışmayla yapay veri üretiminin veri tipleriyle ilişkisi gözlemlenmiştir. Bu amaçla kategorik, sayısal ve hem kategorik hem sayısal veri olmak üzeri üç farklı tip veri seti ele alınmıştır. Öncelikle dengesiz sınıf dağılımına sahip olan ortopedi veri seti, kanser veri seti ve ameliyat sonrası bakım veri setlerinden Uyarlamalı Sentetik Örnekleme (USÖ) ve Sentetik Azınlık Aşırı Örnekleme Tekniği (SAAÖT) yöntemleri kullanılarak yapay veriler üretilecek ve veri setleri yapay verilerle güncellenecektir. Daha sonra ham veri seti ve yapay veriden oluşan veri setleri makine öğrenmesi yöntemlerinden K En Yakın Komşuluk (KEYK) ve Karar Ağacı (KA) kullanılarak sınıflandırılacak ve başarı oranları kaydedilecektir.

Son yıllarda yapılan çalışmalarda KA ve KEYK gibi makine öğrenmesi yöntemlerinin hastalıkların tanı ve tedavisinde önemli çıktılarının olduğu gözlemlenmiştir. Dengesiz veri seti kullanılarak yapılan analizlerin sonuçlarında elde edilen başarı oranları yanıltıcı olabilmekte buna bağlı olarak elde edilecek yüksek başarımlı sonuçlar modelin başarılı olduğu anlamına gelmemektedir (Nizam ve Akın, 2014). Çünkü azınlıkta olan sınıfa ait örneklerin yetersizliği modelin tam anlamıyla başarılı sonuçlar oluşturmasını engellemektedir. Bu durumu ortadan kaldırmak amacıyla, sentetik veri üretimi gerçekleştirilerek, azınlık sınıfı ile çoğunluk sınıfına ait örnekler arasındaki dengesizliğin ortadan kaldırılması amaçlanmaktadır. Bu bağlamda bu tez çalışmasında Uyarlanabilir Sentetik Örnekleme (USÖ) ve Sentetik Azınlık Aşırı Örnekleme Tekniği (SAAÖT) yöntemi kullanılmaktadır. Genel olarak bu iki algoritmasının amacı, mevcuttaki azınlık sınıfından yeni sentetik örnekler oluşturarak sınıf dengesini iyileştirmektir.

Bu çalışmada sayısal veri tipinde ortopedi veri seti ele alınmış ve ortopedik hastalıklara ait çalışmalar incelenmiştir. Çalışmalar incelendiğinde yaşamları boyuna birçok insanın fiziksel, psikolojik ve fizyolojik olarak yaralanmalara maruz kaldığı görülmektedir. Bu yaralanmaların birçoğu kas-iskelet sisteminin zarar görmesine neden olacak şekilde gerçekleşmektedir.

Kategorik veri tipinde ise meme kanseri veri seti kullanılmıştır. Bu bağlamda literatürde yapılan çalışmalar incelendiğinde; kanser olarak tabir edilen hastalıklarla dünya üzerinde birçok insanın mücadele ettiği görülmektedir. Öldürücülüğü fazla olduğu için bu tip hastalıklarda tanı ve tedavi süresi hayati önem arz etmektedir.

Son olarak hem kategorik hem de sayısal veri tipinde ameliyat sonrası bakım veri seti kullanılmıştır. Her iki rahatsızlık türünde de hasta belirli zaman aralıklarında bir veya birden fazla cerrahi müdahale ve ameliyatla karşı karşıya kalabilmektedir. Ameliyat süreci tamamlansa bile hastalığın seyri ve takibinin sağlanması önem arz etmektedir. Bu nedenle hasta izlem süreci, hastanın sağlık durumunu etkileyen önemli unsurlar arasında yer almaktadır (Tighe ve ark., 2015). Literatürde ameliyat sonrası bakım olarak adlandırılan bu süreçte hastanın tüm hayati bulguları belirli periyotlarla

ölçülmekte ve durum analizi yapılmaktadır. Hasta bu analiz sonucunda ya yoğun bakım destek ünitesi bağlanır ya normal servise alınır ya da taburcu edilir. Bu süreç sayesinde hastanın hastanede kalış durumu ve tedavi akışı belirlenir. Tüm bu tıbbi süreçlerin kolay yürütülmesi ve yorumlanması, değerlendirilen veri setiyle doğrudan ilişkilidir (Fritz ve ark., 2017).

2. ÖNCEKİ ÇALIŞMALAR

Bu çalışmada sentetik veri üretiminin farklı veri tipleri üzerindeki etkileri tartışılmaktadır. Hatice Nizam ve Saliha Sıla Akın (2014), yapmış oldukları duygu analizi çalışmasında twitterdan kendi yöntemleriyle elde ettikleri gıda ürünlerine ait yorumlardaki kelime analizi ile negatif, pozitif ve nötr sınıfına ait veriler oluşturmuşlardır. Oluşturulan veri setinde, sınıflar arasında dengesiz bir dağılım olduğu görülmüştür. Sentetik veri üretimiyle oluşturulan yeni veri seti, dengesiz veri setinden daha yüksek başarılı sonuçlar oluşturmuştur.

Sultan TURHAN ve ark. (2008), sağlık verilerinde çokça görülen dengesizlik problemini diyabet hastalığı üzerinden ele almışlardır. Dengesiz dağılıma sahip veri setinden elde edilen başarı oranlarının düşük sonuçlar oluşturulmasıyla mevcut veri seti alt örnekleme, aşırı örnekleme ve DSÇ'li SAAÖT yöntemleriyle çoğaltılarak dengesizlik sorunu çözülmüştür. Böylelikle oluşturulan veri setinden yüksek başarılı sonuçlar elde edilmiştir. Ancak bu çalışmada kullanılan DTE-SBD yöntemi tüm diğer yöntemlere göre daha başarılı sonuçlar üretmiştir.

Sentetik veri üretiminin önem kazandığı sektörlerden biri de bankacılık sektörüdür. Jie sun ve ark. (2018), çalışmalarında kredi değerlendirme sisteminin çoğunlukla dengesiz veriler içerdiğini ve bu noktada SAAÖT kullanarak sentetik veri üretilmesi sonucunda, yüksek başarılı sonuçlar elde edildiğini belirtmişlerdir.

Günümüzde Bilgisayar Destekli Tanı (BDT) sistemi birçok tıbbi çalışmalarda etkin olarak kullanılmaktadır. Bel ağrısı, ABD'de 10 milyonun üzerinde Amerikalıda intervertebral disk hastalığı olarak en sık görülen nörolojik hastalıkların başında gelmektedir (Alomari ve ark., 2011). Dolayısıyla hasta sayısının fazla, hastayla ilgilenecek radyolog sayısının yetersizliği bu sistemi gerekli kılmaktadır. Bu sistem Resim Arşivleme İletişim Sistemleri (RAİS) sistemine göre daha yenilikçi ve güncel bir yapıya sahiptir. BDT 'nin hastalığın tanı aşamasında yardımcı rol oynaması bu sistemin birçok önemli araştırmaların konusu olmasına neden olmaktadır. BDT radyologların dikkat etmediği veyahut gözden kaçırmış olduğu tüm radyolojik

bulguların tespitini sağlayarak radyoloğa yardımcı olur. Daha önce yapılan bir çalışmada Çok Katmanlı Algılayıcılar (ÇKA) ve Naif Bayes (NB) yöntemleri uygulanarak vertebral kolon veri seti sınıflandırılmış, bu sınıflandırma ise BDT sistemi ile otomatik hesaplanabilir hale getirilmiştir. (Yavuz Ünal ve ark., 2013)

Bunun yanı sıra Bağcı ve ark. (2011), solunum yolu enfeksiyonlarında akciğer ve BT (Bilgisayarlı Tomografi) görüntülerini kullanarak bir çalışma yapmışlar, BDT yönteminin akciğer nodülünün erken teşhisini sağladığı sonucuna varmışlardır.

Benzer bir şekilde Liu ve ark. (2013), böbrek lezyonu bulunan 167 hastayı incelemişler, bu incelemede BDT sisteminin yararlı olduğu sonucuna varmışlardır.

Raja' S. Alomari ve ark. (2010), ise vertebral kolon veri seti üzerinde gerçekleştirecekleri vertebral disk dejenerasyonu teşhisinde iki seviyeli bir olasılık modeli kullanarak BDT sistemini daha etkin bir yapıya kavuşturmayı hedeflemişlerdir.

Bu hastalıkların tespitinde olasılık temelli bir çözüm sunarak intervertebral disk anormalliklerinin tespiti için olasılıksal bir model kullanmışlardır. Görüntü, parametrelerin bulunduğu konum ve bağlam gibi anormal olan diskleri etiketlemek için üç özellik kullanılmış. Anormal disk görünümünü bir gauss modeliyle, konumunu 2B gauss modeliyle ve bağlamı yine aynı şekilde gauss modeliyle anormal diskler arasındaki mesafeye göre modellemişlerdir. Bu sayede çarpraz doğrulama ile %91'lik başarı oranına sahip bir sonuç elde etmişlerdir.

Jason J. Corso ve ark. (2011), Disk herniasyonunun teşhisi için Gibbs dağılımına sahip bir Bayes tabanlı sınıflandırıcı tasarlayıp uygulamıştır. Çarpraz doğrulama deneyinde ortalama %92.5'lik bir başarı oranı elde etmiştir.

Eklem hastalığına neden olan kolon parametreleri çoğu zaman radyolog tarafından veri seti şeklinde oluşturulmaz. Bu noktada görüntü işleme metotları ön özümleme aşaması olarak kullanılmaktadır. Görüntü işleme metotları sayesinde hastaların Manyetik Rezonans Görüntüleme (MRG) çıktıları anlamlandırılır ve sayısal

bir veriye dönüştürülür. Bu sayede değerlendirme sayısal bir veri seti üzerinden yapılır (Özer ve ark., 2014).

Hasan Erdinc Kocer ve ark. (2011), hastadan alınan saggital MR görüntülerini öncelikle gri seviyeye dönüştürme yöntemi ve Ortalama Mutlak Sapma (OMS) yöntemleri kullanılarak özümlemiş, oluşturulan ve içerisinde özellik vektörlerini bulunduran veri kümesi, ÇKA ile yapay sinir ağları kullanılarak sınıflandırılmıştır.

Margarita Kotti ve ark. (2012), yapmış oldukları çalışmada Ortalama Kareli Kayma (OKK) olarak adlandırılan gevşeklik değişkenlerinin enerjisinin en aza indirilmesine dayanan bir ikili sınıflandırma algoritması sunmuştur. Bu algoritmanın temelinde Destek vektör makinası (DVM) yöntemi yer almaktadır. DVM konuşma ve konuşmacı tanıma, duygu sınıflandırması, veri tabanı pazarlaması, coğrafi ve çevre bilimleri, e-öğrenme açısından çeşitli alanlarda kullanılmaktadır. Dolayısıyla makine öğrenmesi metotlarının hastalık teşhisi aşamasında da önemli bir yere kavuşması kaçınılmaz olacaktır.

Bazı bölgelerde imkânların yetersizliğinden dolayı eklem hastalıklarından veri seti üretebilecek veri elde edilememektedir. Bu bölgelerden biri Endonezyadır. Genel olarak, Endonezya'nın devlet hastanelerinde çok sayıda ortopedi hastası mevcuttur. Bu tip hastaların yoğun olduğu bölgelerde sade ve kolay uygulanabilir nitelikte olan Karar Ağacı (KA) yöntemi kullanılmaktadır. Bu yöntem sayesinde yaklaşık %81 ile %85 arasında başarım oranları elde edilmektedir. (Indriana Hidayah ve ark., 2014)

Imran Shafi ve ark. (2013), elde edilen başarı oranları kullanılan sınıflandırma metoduna göre değişkenlik göstermektedir. Yapmış oldukları çalışmada ileri geri besleme yayılım sinir ağı, genelleştirilmiş regresyon sinir ağı ve DVM gibi makine öğrenme sınıflandırma metotlarını kullanarak vertebral kolon bozukluklarının teşhisini ve sınıflandırılmasını önermektedir. Bu bağlamda yapmış oldukları çalışmada birçok sınıflandırma yöntemini kullanmış ve performanslarını değerlendirmişlerdir. Kullanılan yöntemler arasında maksimum başarı oranını %86 ile DVM yönteminde yakalamışlardır.

Sınıflandırma işlemleri yapılırken dikkat edilmesi gereken hususlardan bir tanesi de kullanılan veri setinin dengesiz bir yapıya sahip olup olmamasıdır. Çünkü eğitim verilerinde yer alan sınıflar arasındaki dağılımın eşit olmadığı durumda, çoğunluk sınıfına ait özellikler öğrenme algoritması içerisinde belirgin bir yapıya sahip olmasına neden olacaktır.

Yapılan çalışmalar incelendiğinde sağlık ve makine öğrenmesi yöntemlerinin bir arada kullanıldığı birçok çalışmanın mevcut olduğu görülmektedir. Bunlardan biri de Geçiş Başlatma Siteleri (GBS) tahminidir. Bu uygulama gen yapısı tahminindeki önemli yöntemlerden biridir. Bu tahmin genel olarak bir sınıf dengesizliği problemi olarak karşımıza çıkmaktadır. Literatürde TIS tahmini olarak geçmektedir ve Yapay Sinir Ağı (YSA), DVM vb. gibi çeşitli makine öğrenmesi teknikleri ile yapılmaktadır. Muhammad Hossain ve ark. (2011), yaptıkları çalışma ile bu yöntemlerden Kaçıran Ağ (KA) algoritmasını kullanarak %93 doğruluğa sahip bir başarı oranı elde etmiştir.

Nicolás García-Pedrajas ve ark. (2011), yılında yapmış oldukları çalışma ile sağlık ve makine öğrenimi yöntemlerini bir arada kullanarak, GBS'nin bir sınıf dengesizliği problemi olduğunu bu nedenle, bu dengesiz veri kümeleriyle başa çıkmak için geliştirilen farklı yöntemleri uygulamayı amaçlamışlardır. Bu yaklaşımın avantajlarından biri makine öğrenmesinin sağlık verileri üzerinde standart sınıflandırma yöntemleri kullanıldığında etkisini anlamaktır. Bunun yanı sıra veri seti üzerinde var olan dengesizlik problemi giderilerek uygulanabilecek algoritmaların sayısını arttırmak amaçlanmıştır. Bu sayede, eğitim için gerekli olan süre kısalmış olacaktır.

Sonuçlar, sınıf dengesizliği göz önüne alınmadan yapılan sınıflandırmalar ile, sınıf dengesizliği giderilen veri seti üzerinde yapılan sınıflandırmalar arasındaki farkı göstermiştir. Bu durumda azınlık sınıfına ait özelliklerin algoritma tarafından tam anlamıyla tanınması zorlaşmaktadır. Bu sorunu ortadan kaldırmak ve azınlık sınıfları arasındaki düşük başarı oranlarını iyileştirmek üzeri birçok çalışma yapılmıştır.

Piyasak Jeatrakul ve ark. (2010), yapmış oldukları çalışmada önermiş olduğu yöntem, dengesiz verilerin sınıflandırılması problemini düzeltmek için SAAÖT ve Tamamlayıcı Sinir Ağı (TSA)'yı birleştirmişlerdir. Önerilen tekniğin dengesiz verilerin sınıflandırılmasına yardımcı olabileceğini göstermek için, çeşitli sınıflandırma algoritmaları kullanılmışlardır. Bunlar YSA, KEYK ve DVM'dir. Sonuçlar, sınıf dengesizliği problemi için performansı artırabileceğini göstermektedir.

Tingting Pan ve ark. (2019) 'nın yaptıkları çalışmada, gauss aşırı örnekleme yöntemini yani boyut küçültmeyi gauss dağılımı ile birleştirerek aşırı dengesiz sınıflandırmalar üzerinde daha başarılı sonuçlar elde ettiklerinden bahsetmişlerdir. Aynı çalışmada SAAÖT yöntemi de kullanılmış ve SAAÖT' den, mevcut diğer örnekleme yöntemlerinden daha yüksek verimli sonuçlar elde edildiğini ifade etmişlerdir.

SAAÖT yönteminin birçok uzantısı bulunmaktadır. USÖ yöntemi bu uzantılar arasında en çok tercih edilen yöntemlerin başında gelmektedir. Bunun nedeni adaptif bir algoritma mantığıyla çalışıyor olmasıdır.

Yulia Ery Kurniawati ve ark. (2017), yapmış oldukları çalışma ile Pap testi sonuçlarının sınıflandırılmasında kullandıkları verileri ele almıştır. Bu veri seti üzerinde dengesizliği gidermek amacıyla bir USÖ algoritması geliştirmişlerdir. Çalışmada, algoritma olarak dengesizliği ortadan kaldırmak amacıyla SAAÖT-N algoritması kullanılmıştır. Sonuç olarak, USÖ-KEYK ile sonucu "0" olan verilerin %95,38'i doğru tahmin edilmiştir. Yine Aynı şekilde sonucu "1" olan verilerin ise %99.183'u USÖ-KEYK ile doğru olarak tahmin edilmiştir.

Kanser araştırmalarında uzun zamandır makine öğrenmesi yöntemleri kullanılmaktadır. Özellikle KA ve YSA yöntemleri 20 yıldan fazla bir zamandır kanser teşhislerinde önemli rol oynamaktadır. (Simes 1985; Maclin vd. 1991; Ciccheti 1992).

McCarthy ve ark. (2004), makine öğrenmesi yöntemleri birçok çalışmada kanser teşhisi ve tespitini sağlamak amacıyla kullanılmıştır. Son PudMed verileri incelendiğinde bu çalışma biçimine ait 1500'den fazla makale yazıldığı görülmektedir.

Sung-Bae Cho ve ark. (2003), 3 farklı kanser türüne ait teşhisine ait gen verilerini kullanarak, lösemi,kolon ve lenfoma kanser türlerini ÇKA, KEYK, DVM ve yapı uyarlamalı kendi kendini organize eden harita yöntemlerini kullanarak başarılı tanıma oranını üretmişlerdir.

Bilinen kanser türleri içinde en yaygın görülen meme kanseridir ve literatürde bu alanda yapılan çok sayıda çalışma mevcuttur. Hiba Asri ve ark. (2016), meme kanserine ait parametrelerden oluşan bir veri setini birçok sınıflandırma yöntemiyle analiz etmiş ve performans karşılaştırması yapmıştır. Bu yöntemler arasında en yüksek başarım oranını % 97.13 lük bir oran ile DVM yönteminin verdiğini belirtmiştir. Bu alanda yapılan çalışmalardan bir diğeri ise Ahmad LG ve ark. (2013), yapmış oldukları çalışmada kanser veri setini makine öğrenmesi yöntemleri ile değerlendirmiş ve bu değerlendirmelerde en yüksek başarı oranını %95.7 ile DVM yönteminde elde etmiştir.

Öte yandan bilinen sınıflandırma yöntemlerinin yanı sıra, bu yöntemlere ait alt sınıflandırma yöntemlerinin de değerlendirildiği birçok çalışma mevcuttur. Mitra Montazeri ve ark. (2016), yapmış oldukları araştırmada meme kanserinin erken teşhisini sağlamak amacıyla farklı makine öğrenim tekniklerini birleştirmeye çalışmıştır. NB, Rastgele Ağaçlar Yöntemi (RAY), KEYK, Uyarlamalı Güçlendirme (UG), DVM, RBF Ağı (RBFN) gibi birçok makine öğrenmesi yöntemlerini kullanmış ve RAY'da en yüksek başarımı elde edildiğini kanıtlamışlardır.

Kanser hastalığına ait veriler çoğu zaman radyolojik görüntülerden sağlanmaktadır. Bu görüntüler içerisindeki parametreleri anlamlandırmak ve daha sonrasında değerlendirmek için makine öğrenmesi yöntemleri sıklıkla kullanılmaktadır. Vishwa S.Parekh ve ark. (2017), MR görüntülemesi yapılan 124 hastanın radyolojik görüntülerini incelemiş ve bunu sayısal verilere

dönüştürmüşlerdir. Bu veri setini makine öğrenmesinin çoklu görünümlü DVM yöntemini kullanarak çözümlemişler ve yüksek başarı oranı elde etmişlerdir. Bu durum görüntülü verilerin dönüşümünde de DVM'nin başarısını bir kez daha göz önüne sermiştir.

Hastaların ameliyat sonrası durumuna bağlı olarak taburcu edilmesi, yoğun bakıma alınması veya servise çıkarılması söz konusu olabilmektedir. Bu durumda hastanın ameliyat sonrası durumunu anlamak için hastadan elde edilen veriler değerlendirilebilir ve hastanın ameliyat sonrası taburcu olup olmamasına karar verilebilmektedir.

Bradley A Fritz ve ark. (2012), yapmış oldukları çalışma ile 4 yıllık bir süre boyunca ABD'de bulunan bir hastanede ameliyat olan tüm yetişkin hastaların değerlendirilmesi sağlanmıştır. Hasta öyküsü, laboratuvar değerleri, dakika dakika intraoperatif vital bulgular ve uygulanan ilaçlar kayıt altına tutulmuş ve veri seti oluşturulmuştur. Bu veri seti üzerinden makine öğrenmesi yöntemleri kullanılarak hastaların durumları anlık olarak görüntülenmiştir.

Öte yandan ameliyat sonrası hastalarda meydana gelebilecek ağrının boyutları da makine öğrenme yöntemleri ile önceden tahmin edilebilmektedir. Patrick J. Tighe ve ark. (2015), yaptık oldukları çalışma ile bunu kanıtlar nitelikte sonuçlar elde etmişlerdir. 8701 hastaya ait 796 klinik değişkenin incelendiği çalışmada diğer çalışmalarda nadiren karşılaştığımız En az Mutlak Büzülme ve Seçim Operatörü (EMBSO) algoritması kullanılmış ve %72.7'lik bir başarı oranı elde edilmiştir. Fakat veri kümesinin temelinde yer alan sınıf dağılımları orantısız olması aslında bu başarı oranının adaptif veri üretimiyle daha yüksek oranlara çekilebileceğini göstermektedir.

Yapılan araştırmalar gösteriyor ki ameliyat sonrası hastalarda meydana gelebilecek komplikasyonların ön görülmesi imkânlı hale gelebiliyor. Stefanos Voglis ve ark. (2020), yapmış oldukları çalışmayla hipofiz ameliyatı sonrası hastada meydana gelebilecek olan sıvı fazlalığına ilişkin bir öngörü metodu üretmişlerdir. Bu çalışma

sonucunda ÇKA yönteminin öngörüde başarı oranını en çok etkileyen yöntem olduğu savunulmuştur.

Dengesiz veri dağılıma sahip medikal veri setlerinde sentetik veri üretimiyle ile ilgili çalışmalar ele alınmış ve bu alanda kullanılan yöntemlere ait literatür çalışmaları özetlenmiştir. Görülmüştür ki medikal veri setlerinde sentetik veri üretimi olumlu etki yaratmaktadır.

3. MATERYAL ve YÖNTEM

3.1. Medikal Veri Setleri

Bu çalışmada sentetik veri oluşturma yöntemlerinin dengesiz veri setleri üzerindeki etkisi incelenmiştir. Bu amaçla farklı veri tiplerine sahip numerik, kategorik, numerik-kategorik veri setleri ele alınmıştır. Bu veri setleri:

1. Ortopedi Veri Seti
2. Meme Kanseri Veri Seti
3. Ameliyat Sonrası Bakım Veri Seti

3.1.1. Ortopedi Veri Seti

Bu çalışmada kullanmış olduğumuz ilk veri seti sayısal veri tipine sahip ve Henrique da Mota tarafından, Fransa'da tedavi gören hastalardan elde edilmiştir (Barreto ve Neto, 2009). Veri setinde hastalar iki kategori halinde sınıflandırılmıştır. Bu sınıflandırma grupları normal ve anormaldir. Normal sınıfına ait 102 örnek, anormal sınıfına ait 208 örnek bulunmaktadır. Veri setinde ayrıca 6 adet öznitelik yer almaktadır. Bu öznitelikler aşağıdaki tabloda belirtilmiştir.

Çizelge 3.1. Ortopedi veri seti öznitelikleri

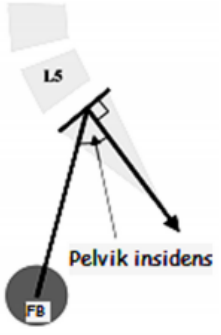
Pelvik İnsidens
Pelvik Eğim Numarası
Lomber Lordoz Açısı
Sakral Slop
Pelvik Açısı
Spondilolistezis Derecesi

Ortopedi kelime anlamı itibariyle orthos ve paedia sözcüklerinin birleşiminden oluşmuştur. Latincede düzgün (orthos), çocuk (paedia) anlamına gelir. Yani ortopedi kelimesi tam anlamıyla düzgün çocuk demektir. Temel anlamda kas-iskelet sisteminin

tümüyle ya da bölgesel değerlendirilmesini sağlayan, vücut morfolojisinin işlevlerini araştıran, doğumla birlikte gelen, sonradan edinilmiş veya travma sonrası meydana gelen lezyonları inceleyen, önlemler alınmasını sağlayan bir tıp bilimi alanıdır.

Başlarda ortopedinin çalışma alanları çocuklarla ilgili kas-iskelet sorunları olsa da birinci dünya savaşı itibariyle travmaya bağlı olarak gerçekleşen sakatlık ve şekilsel bozuklukları gidermek amacıyla yeni yöntemler geliştirilmiştir (Angay, 2009). Ortopedi alanındaki tedavi yöntemleri günümüzde her yaş için uygulanmaktadır. Bitişik ayak ile doğmuş bir bebekten, kapalı ameliyat ihtiyacı duyan bir genç atlete ya da eklem ağrıları yaşayan ileri yaştaki bir insanı ilgilendirebilecek popülasyonda tedavi yöntemleri uygulanmaktadır. Bu tedavilerin amacı genel olarak vertebral kolonun hasar gören kısımlarını tedavi etmektir (Unal ve ark., 2011).

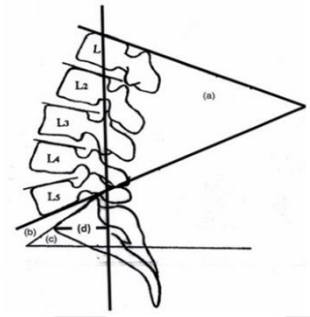
Vertebral kolon, genel anlamda vücudun gövde kısmının kemik sütunudur. Bu bütünlüğün en önemli parçası omurgadır. Omurgada dengesizliğe en çok yol açan neden bel kısmında meydana gelen bozulmalar ve sakatlanmalardır ki bel omurganın en fazla yük taşıyan hareketli kısmıdır (Ozer ve ark., 2014). Bu denge belirli parametrelerden oluşan sistematik bir oluşumdur. Bu çalışmada bu parametrelerin oluşturduğu veri setinden yararlanılacağından aşağıda bu parametrelerden bahsedilecektir. Pelvik insidens, üst sakral son plağın tam orta noktasından geçen dikey çizgi ile femur başı eksenini bu orta noktaya birleştiren çizgi arasındaki açı olarak tanımlanır.



Şekil 3.1. Pelvik insidens açısı, FB: Femur başı (Karademir ve ark., 2014)

Pelvik eğim, bel bölgesinde bulunan kasların esnekliğini ifade eder. Ayrıca hareket esnasında, karın kaslarının kasılma ve gevşemesini de belirtir. Pelvik eğim, pelvisin pozisyonuna göre değişkenlik gösteren parametredir.

Lomber lordoz açısı, omurlar arasındaki toplam açıyı belirtir (Şekil 3.4).



Şekil 3.2. Lomber Lordoz Açısı (LLA) (Esmailiejah ve ark., 2017)

Sakral slop, üst sakral son plaktan gelen çizgi ile üst sakral son plak orta noktasından gelen yatay çizgi arasında kalan açıdır. Pelvik açısı, pelviğin yarıçapını ifade eder. Yatay olarak ölçülür.

Spondilolistezis; çoğunlukla beşinci, bazı durumlarda dördüncü lumbar vertebranın interartiküler bölgesinde bir birleşmeme sorunu meydana gelmektedir.

Bir vertebranın herhangi bir nedenle üzerinde bulunduğu vertebral kolonla birlikte altında bulunan vertebra üzerinde öne doğru kaymasına denir. Bunun risk düzeyinin belirlenmesi bu parametre ile gösterilir.

3.1.1.1. Ortopedik Hastalıkların Teşhis Yöntemleri

Öncelikle hastaya bel fıtığı teşhisi konmadan önce hekim tarafından hastanın öyküsü alınır, daha sonra tıbbi muayenesi yapılır. Daha sonra hastanın nörolojik muayenesi yapılarak kaslarında, güç kaybı, hissizlik gibi semptomların olup olmadığının tespiti yapılır (Tighe ve ark., 2015).

Bu noktadan sonra hastanın yüksek çözünürlüklü sonuç üreten tanı cihazlarıyla MR veya BT taraması gerçekleştirilir. EMG (elektromiyogram) cihazı kullanılarak hastanın vücudunda fıtığa sebep olan ve zarar görmüş sinir kökleri belirlenir ve bu fıtığı temsil eden parametrelerin belirlenmesi sağlanır. Bu amaçla destekleyici ve yardımcı bir sistem olan BDT sistemleri kullanılır.

Bilgisayarlar artan teknolojiyle beraber görüntülerin içeriğini hafızasında tutacak kadar gelişmiş bir yapıya dönüşmüştür (Alomari ve ark., 2011). Araştırmacılar görüntü verileri üzerinde analiz yapmak üzere bilgisayar programları oluşturarak bu veriler üzerinden yorumlama yapmaya yönelik çalışmalar yapmaya başlamışlardır.

Tıpta, insan anatomisinin görüntüleri üzerinden hastalığın teşhisi ve beraberinde tedavisine yardımcı olmak amacıyla sıkça kullanılmaya başlanmıştır. Özellikle radyoloji alanında tıbbı görüntülerin yorumlanması önem arz etmektedir.

Hastaların görüntülerine bakarak hastalık tanımlaması yapmak zor bir süreçtir. Bu süreci kolaylaştırmak ve başarılı sonuçlar elde edebilmek için bilgisayarlara ihtiyaç duyulmaktadır (Alomari ve ark., 2011). Bu araştırma alanı BDT olarak bilinmektedir.

BDT sistemi bilgisayarlarda kullanılan yazım denetleyicisi gibi çalışmaktadır. Bilgisayarda metin yazılırken eğer bir yazım hatası yapılmışsa sözcüklerin alt kısmında bir işaretleyici belirir ve hatanın türü belirtilir. BDT sistemi de bu şekilde bir çalışma mekanizmasına sahiptir ancak daha kapsamlı ve ileri seviyede bir denetleme sağlar. Benzer şekilde, tıbbi görüntüler yorumlanırken radyolog bilgisayar üzerine bir işaretleyici yerleştirerek görüntülerde şüpheli olduğunu düşündüğü alanların yorumlanmasını sağlar.

1998'de ilk BDT ürünü ABD Gıda ve İlaç İdaresi'nin onayını aldı. California merkezli R2 şirketi tarafından geliştirilen bu sistem mamogramlardaki mikro kalsifikasyonları ve kitleleri tespit etmek için kullanılmıştır. R2 şirketi kendi üretmiş olduğu BDT sisteminin 2002 yılında bir süre “Ay’ın Vakası” olarak nitelendirilen 61 yaşındaki bir kadının göğüs kanseri teşhisini sağladığından bahsetmiştir. Bu vaka ile

kadından alınan mamografi görüntülerini inceleyen radyolog kadının göğsünde kansere rastlamıştır. Bu sebeple R2 şirketine ait "Malignite de R2 ImageChecker" üzerinden takip başlatmıştır. 2002 yılında satışı başlayan ImageChecker ürününde, radyolog bir düğmeye bastığında BDT sisteminin işaretlediği şüpheli noktalar bir monitörde hastaya ait olan görüntüler başka bir monitörde görüntülenme şansı sunmaktaydı. Bu sayede radyologun önce görüntüleri incelemesi ve bundan sonra BDT tarafından ikinci görüş alınması sağlanmaktaydı.

3.1.2. Meme Kanseri Veri Seti

Bu çalışmada kullanmış olduğumuz diğer veri seti kategorik veri tipine sahiptir. Bu veri seti fizikçi Matjaz Zwitter & Milan Soklic tarafından elde edilmiştir. Veri setinde hastalar iki kategori halinde sınıflandırılmıştır (Zwitter ve Soklic, 1988). Bu sınıflandırma grupları tekrarlayan ve tekrarlanmayandır. Tekrarlamayan sınıfına ait 201 örnek, tekrarlayan sınıfına ait 85 örnek bulunmaktadır. Veri setinde 9 adet öznitelik yer almaktadır. Bu öznitelikler aşağıdaki tabloda belirtilmiştir.

Çizelge 3.2. Meme kanseri veri seti öznitelikleri

Yaş
Menopoz
Lenf Düğümleri
Düğüm Kapakları
Kötü Huyluluk Derecesi
Meme
Meme Kadranı
Tümör Boyutu
Işınlama

Onkoloji; vücutta yer alan urların (tümör) oluşumu, oluşma nedenleri, tanı konulma sürecini, tedavisini ve kalıtımsal bağlantılarını inceleyen bilim dalıdır. Onkoloji sadece kanserli hücreleri incelemez, kanserli olmayan hücreleri de kendi içerisinde inceler. Bu sebeple ilgili olduğu alt dallar genişlemiş olur. Özellikle kanser hastalarının tedavisinde uygulanan kemoretapi uygulamalarını inceleyen bu tıbbi

onkoloji beraberinde kanser taraması ve kanserden korunma ile hastalığın ön tanısını sağlamaya yardımcı olan uygulamaları da inceler.

Onkoloji daha çok kanserli hücreleri inceler. Kanser immünolojisi bu bağlamda onkoloji hastalıkları arasında en önemli olanıdır. Ardından akciğer kanseri bulunmaktadır. Kanser, bağlantılı olduğu tüm hastalıklara verilen ad olarak ifade edilebilir. Kanserli hücre tüm kanser türlerinde anormal bir hızla bölünür ve çoğalırlar. Çoğalan kanserli hücreler çevre dokulara yayılmaya başlarlar İnsan vücudu kendi içerisinde trilyonlarca hücre barındırır ve bu sebeple kanser bu hücrelerden herhangi birinde başlayabilir. Normal şartlarda vücudun ihtiyaç duyduğu anda yeni hücreler üretilir ve çoğalırlar. Var olan hücreler zamanla yaşlanır ve ölürler. Ölen hücrelerin yerine yeni hücreler meydana gelir.

Bununla beraber, kanser büyümeye devam ettiğinde, kusursuz olarak nitelendirilebilecek olan bu süreç bozulmaya uğrar. Hücreler normalliğini yitirir, hasar görmüş ve yaşlı olan hücreler ölmeleri gerekirken yaşamaya devam ederler. Gerek duyulmamasına rağmen yeni hücrelerin meydana geldiği görülür. Bu oluşan yeni hücreler sürekli bölünebilir ve bu bölünmeler tümör olarak adlandırılan yumruları oluştururlar (Asri ve ark., 2016). Kanser doku kütlesini büyük tümörler haline getirir ve bu tümörler katı tümör olarak adlandırılır.

Kötü huylu tümörlerden olan kanseröz tümörler, çevresinde bulunan diğer dokulara yayılabilmektedirler. Ayrıca, bazı kanser hücrelerinin kanserli bölgeden ayrılarak dolaşım veya kan yoluyla vücudun diğer kısımlarına giderek başka hücrelerde kanserli tümörler meydana getirebilirler. Öte yandan iyi huylu tümörler kötü huylar gibi hareket etmezler ve çevredeki dokular üzerinde yayılım göstermezler. İyi huylu tümörler normalin üzerinde büyüklükte de bulunabilirler. Cerrahi işlemle alındıktan sonra yeniden gelişim göstermezler (Cruz ve ark., 2006). Fakat kötü huylu tümörler alındıktan sonra büyüyerek tekrar ortaya çıkabilirler. Bu sebeple kötü huylu tümörler yaşamı tehdit eder nitelikte olmaktadır.

Normal vücut hücrelerinin yenilenmesi kontrollü bir şekilde gerçekleşirken kanser hücreleri kontrolsüz bir bölünme ve çoğalma özelliği göstermektedir (Lg ve ark., 2013). Kanser hücreleri bulundukları yere göre isimlendirilirler Bilinen 100 ayrı çeşit kanser türü mevcuttur. Bunların başında; mesane kanseri, meme kanseri, cilt kanseri, mide kanseri yer almaktadır.

Onkolojik rahatsızlıkların belirlenmesinde ilk olarak kanserli hücrenin oluşum nedenlerine bakılır. Bu aşamadan sonra tanı ve tedavi evlerine geçilerek kemoterapi uygulamalarına başlanır (Parekh ve ark., 2017).

3.1.1.2. Meme Kanseri

Meme kanseri, memeyi meydana getiren dokularda yer alan hücre gruplarının herhangi birinde meydana gelen değişimle birlikte kontrolsüz çoğalması sonucu meydana gelen tümörle gelen bir hastalıktır. Kanserli hücre, etrafında bulunan diğer hücrelere ardından meme etrafında bulunan lenf bezlerine ulaşarak yayılım göstermektedir. Özellikle erken tanının yapılması bu süreçte yayılımın önüne geçilmesine yardımcı olabilmektedir.

Günümüzde meme kanserinin görülme sıklığı oldukça artmıştır. ABD verilerine bakıldığında bir kadının hayatı boyunca meme kanserine yakalanma olasılığının %12.5 olduğu görülmektedir (Cruz ve ark., 2006).

3.1.1.3. Meme Kanseri Risk Faktörleri

Bu çalışmada kullanılan veri setinde belirtilen parametrelerin tanımlanması ve ifade edilmesi için, veri setinde yer alan ifadelerin tam olarak anlaşılması gerekir. Bu bağlamda kansere neden olan risk faktörlerinden bahsedecek olursak;

50 ile 70 yaş arasında menopoz geçirmiş olmak, kadın olmak, daha önce meme kanserine yakalanmış kan bağı olan kişilerin olması, doğum yapmamış olmak, daha önce meme kanserini geçirmiş olmak, adete başlama döneminin erken, menopoz

başlama döneminin geç olduğu durumlar, ilk çocuğunu 30 yaş sonrasında dünyaya getirmek, çok uzun zaman süren hormon tedavilerine maruz kalmak, büyük şehirde yaşıyor olmak, sigara ve alkol kullanmak, aşırı şişmanlık, zararlı yiyecekler tüketmek, fiziksel aktiviteleri az yapmak gibi unsurlar kanserin seyrini ve dolayısıyla tedavi sürecini doğrudan etkileyebilmektedir.

3.1.1.4. Tanı Yöntemleri

Hekim tarafından öncelikle hastanın ve yakınlarının tıbbi öyküsü dinlenir. Daha sonra fiziki tıbbi muayene yapılarak hastanın mamografisi veya meme ultrasonu gibi tıbbi görüntüleri istenir. Tüm bu süreçlerden önce erken tanının sağlanması için kişinin alması gereken bazı tedbirler mevcuttur. Bunlar;

Meme kanseri hastaları genel olarak (yaklaşık %70) kanserli bölgeyi kendi memelerini incelerken fark ediyorlar. Bu sebeple 20 yaşını geçmiş kadınların ayda bir defa kendi memelerini ayna karşısında gözlemlemelidirler. Herhangi bir değişim veya anormallik fark edildiğinde ise direkt olarak hekime başvurmalıdırlar.

Kadınların memeleriyle ilgili bir şikâyeti olmasa dahi 20 ila 40 yaş aralığında ise 3 yılda bir, 40 yaşından büyük ise yılda bir kez hekim muayenesini yapmalıdır.

Günümüzde teknolojinin gelişmesiyle birlikte tanıda en belirgin ve erken sonuçlar meme görüntülemesi yöntemiyle sağlanmaktadır. Bu bağlamda mamografi ve özellikle tarama mamografisi kanser vakalarında olmazsa olmaz yöntemlerin başında gelmektedir (Asri ve ark., 2016). Bu sayede hastanın elle fark edemediği tümörler bile ortaya çıkmaktadır.

3.3.2. Ameliyat Sonrası Bakım Veri Seti

Bu çalışmada kullanmış olduğumuz son veri seti hem sayısal (numerik) hem de kategorik veri tipine sahip ve Sharon Summers tarafından hazırlanmış veri setidir (Summers ve ark., 1993). Veri setinde hastalar I (hasta yoğun bakım ünitesine

gönderildi), S (hasta eve gitmeye hazır) ve A (hasta genel hastane katına gönderildi) olmak üzere üç kategori halinde sınıflandırılmıştır. S sınıfına ait 24 örnek, A sınıfına ait 63 örnek, I sınıfına ait 2 örnek bulunmaktadır. Veri setinde ayrıca 8 adet öznitelik yer almaktadır. Bu öznitelikler aşağıda yer alan tabloda belirtilmiştir.

Çizelge 3.3. Ameliyat sonrası bakım veri seti öznitelikleri

L-CORE
L-SURF
L-O2
L-BP
SURF-STBL
CORE-STBL
BP-STBL
COMFORT

Ameliyat sonrası bakım süreci hemşirelerin hasta bakımını, hasta ameliyathaneden çıktıktan taburcu olana kadar yapması gerekli olan tedavi ve bakımların uygulanma sürecini belirtmektedir. Cerrahi müdahale sonrası hasta ameliyatın hemen ardından ayılma ünitesine aktarılmaktadır.

Hasta, ameliyat sonrası anestezinin etkisinden kurtuluncaya kadar, yaşam kalitesinin düzenli olarak kontrol edildiği, yaklaşık 2 saat süren bir ünite olan ayılma ünitesine bağlanır. Hastanın ameliyat sonrası ihtiyaçlarını karşılamak üzeri, ameliyat öncesi ayılma ünitesi hazırlanmaktadır. Böylelikle hasta ameliyattan çıktıktan sonra direkt olarak ayılma ünitesine alınır ve tetkik ile uygulamalarda meydana gelebilecek gecikmelerin önüne geçilmiş olur. Ayılma ünitesine alınan hastaya öncelikle en uygun pozisyon verilir. Yani hastanın başı yan tarafa döndürülüp baş kısmı mandibuladan ileriye doğru itilir. Böylelikle hastanın dilinin geriye kayarak solunumunu engellemesinin önüne geçilmektedir.

Hastaya uygun olan pozisyon verildikten hemen sonra hemşire; hastanın cilt rengini, yapılan pansumanların durumunu, hastada herhangi bir anormal durumun olup olmadığını, hastaya ait yaşam fonksiyonu değerlerini, yapılan drenlerinin durumunu, hastanın mevcut bilinç durumunu kontrol etmelidir ve tüm bunlara ek olarak; hastanın

ayılma ünitesine hangi saatte alındığı, hastanın bilinç durumu ve düzeyi, hastanın uyarıları tepki verip vermediğine hemşire tarafından bakılmalıdır. Ayılma ünitesine bağlı hastanın yaşamsal fonksiyonları stabil olana kadar her 15 dakikada bir hastanın yaşam bulguları ölçülmelidir. Hastanın şok durumuna girip girmediği kontrol edilmelidir. Bu sebeple cilt rengi ve nem oranı sürekli ölçülmelidir. Hastanın cildi soluk, soğuk ve nemli ise şok müdahalesi uygulanmalıdır. Ayrıca ayılma ünitesinde hastanın pansumanları da kontrol edilmelidir. Herhangi bir şekilde ıslaklık veya kanama gözlemlenirse hastaya müdahale edilmelidir. Hastanın vücut ısısı belirli aralıklarla ölçülmelidir.

Hastaya varsa verilen damar içi sıvıların durumu kontrol altında tutulmalıdır. Verilen sıvıların miktarı, cinsi ve gidiş hızı gözlemlenmelidir. Hastaya kan veriliyorsa eğer kanın vücuda geliş hızı kontrol edilmelidir. Hastanın ayılma ünitesinden ayrılması, hastanın genel durumuna bağlı olarak değişkenlik göstermektedir. Hasta bu aşamadan sonra ya yoğun bakım ünitesine aktarılır ya da servise iletilir. Hastanın ayılma ünitesinden ayrılarak servise iletilmesi için hastanın bazı kriterleri sağlamış olması gerekmektedir. Bunlar; ameliyattan sonra hasta genel anestezinin etkilerinden kurtulmuş olmalıdır, hastanın yaşam fonksiyonları stabil hale gelmiş olmalıdır ve hastanın bilinci açık olmalıdır.

Hasta, servise alındıktan sonra; hastanın yaşam bulguları, hastanın drenaj ölçümleri, hastanın bilinç durumu, eğer damar içi sıvı iletiliyorsa miktarı ve gidiş hızı, cilt rengi ve hastanın ısısı kontrol edilmelidir. Hastanın ayılma ünitesine alınması sürecinden, hastanın taburcu olmasına kadar geçen süreçteki temel bakımın başlıca amacı, cerrahi müdahale sonucu hastada bozulmaya uğrayan homeostatik dengenin tekrar düzenli bir hale gelmesini sağlamaktır.

Ameliyat sonrası bakım sürecinin temel amacı; hastada meydana gelebilecek herhangi bir komplikasyonu daha karşılaşmadan düzelterek hastanın normal fonksiyonlarına kavuşabilmesini sağlamaktır (Fritz ve ark., 2017).

3.2. Sentetik Veri Oluşturma Yöntemleri

Veriler sınıflandırılırken veri seti üzerinde bulunan sınıf değerleri eşit dağılıma sahip olmayabilir. Bu durum ağırlıklı olarak bulunan sınıfın tüm veri seti üzerinde baskın bir yapıya sahip olmasına neden olmaktadır (Turhan ve ark., 2020). Dolayısıyla verinin karakteristik eğrisinde azınlıkta bulunan sınıfa ait sonuçların başarılı ve gerçekçi sonuçlar elde edilmemesine neden olmaktadır.

Bu sorunu ortadan kaldırmak amacıyla alt örnekleme ve aşırı örnekleme olmak üzere iki farklı sentetik örnekleme yöntemi kullanılmaktadır. Alt örnekleme, çoğunluk sınıfı örnek sayısını, azınlık sınıfı örnek sayısına eşit oluncaya kadar azaltırken, aşırı örnekleme ise çoğunluk sınıf oranını takiben azınlık sınıfı örnek sayısını yeniden örnekler. Örneğin hasta ve sağlıklı sınıf etiketine sahip bir veri setinde, sınıfı hasta olan örnek sayısının 200, sınıfı sağlıklı olan örnek sayısının ise 30 olduğunu varsayalım. Bu durumda hasta sınıfının örnek sayısı veri setinin yaklaşık %87 sini oluştururken, sağlıklı sınıfının örnek sayısı ise veri setinin yaklaşık %13 ünü oluşturmaktadır. Bu durum sağlıklı sınıfından yetersiz sonuçlar elde edilmesine neden olacaktır.

Bir insan nasıl ki üzerinde çok çalıştığı, araştırma yaptığı konuyu daha iyi anlayıp kavrıyorsa makine öğrenmesi de algoritmaları sayesinde benzer bir şekilde öğrenme gerçekleştirmektedir. Bu durum değerlendirmenin kaliteli sonuçlar doğuramayacağı anlamına gelmektedir.

Dengesiz sınıflandırma, eğitim veri kümesinde sınıfların eşit olmayan dağılımı olduğunda sınıflandırma sorunudur (Nizam ve ark., 2014). Sınıf dağılımındaki dengesizlik değişebilir, ancak ciddi bir dengesizlik problemini değiştirmek zordur ve özel teknikler gerektirebilir. Gerçek dünyadaki sınıflandırma sorunlarının birçoğu (dolandırıcılık tespiti, spam tespiti ve kayıp tahmini gibi) dengesiz bir sınıf dağılımı vardır.

Dengesiz bir yapıya sahip sınıflama probleminde sınıf dağılımı üzerindeki dengesizliğin birçok sebebi olabilir. Veri seti üzerindeki bu problemin iki ana nedeni bulunduğu söylenebilir. Bunlar, veri örnekleme ve verinin toplanıldığı alanın özellikleridir. Sınıflar için belirtilen örneklerdeki dengesizlik, verideki örneklerin problem alanından toplanmasından veya örneklenmesinden kaynaklanabilir. Bu durum, verinin toplanması esnasında meydana gelen önyargıları ve veri toplama esnasında yapılan hataları barındırabilir.

Bu hataların en başında ölüm hataları gelmektedir. Örneğin, örneklerin bütünü dar bir zaman diliminden veya dar bir coğrafi alandan toplanmıştır. Bu durumda sınıfların dağılımı farklı olabilir, hatta belki de farklı bir biçimde toplanabilir.

Gözlemlerin toplanması aşamasında hatalar yapılmış olabilir. Oluşan bir hata biçimi çoğu örneğe yanlış sınıf etiketleri uyguluyor olabilir. Alternatif olarak, örneklerin toplandığı işlemler veya sistemler, dengesizliğe neden olacak şekilde hasar görmüş veya bozulmuş olabilir.

Dengesizliğin ölçüm hatası dolayısıyla meydana geldiği durumlarda, ortaya çıkan dengesizlik problemi oluşan ölçüm hatası düzeltilerek veya örnekleme yöntemleri kullanılarak giderilebilir. Bunun sebebi, eğitim veri kümesinin bahsi geçen sorunun etki alanını adil bir temsilcisi olmamasıdır. Dengesizlik problemi, sorunlu alana ait bir özellik olabilir.

Bir sınıf kümesinin varlığı ve doğal meydana gelişi diğer sınıflara egemen olabilmektedir. Sınıf dağılımı üzerindeki dengesizlik, oluşan probleme bağlı olarak değişecektir. Yine, sınıflandırma sorunu, belirli bir eğitime ait veri seti için bir sınıf üzerinde çok sayıda örnek ve başka bir sınıfta daha az örnek olabilen önemli bir dengesizliğe sahip olabilir.

Deneme yapılmak için kullanılan veri setinde yer alan örneklere ait sınıf etiketleri arasındaki eşitsizliğin az olduğu sınıflandırma problemi hafif dengesizlik olarak ifade edilir. (örn. 5 Hasta, 8 Sağlıklı).

Öte veri setinde yer alan örneklere ait sınıf etiketleri arasındaki eşitsizliğin az olduğu sınıflandırma problemi hafif dengesizlik olarak adlandırılır (örn. 5 Hasta, 125 Sağlıklı).

Dengesiz bir dağılıma sahip bir veri kümesinde, örneği diğerine nazaran daha fazla olan sınıflar çoğunluk sınıfı ya da majör sınıf olarak adlandırılır. Az sayıda örneğe ve dengesiz bir dağılıma sahip bir veri kümesinde, örneği diğerine nazaran daha az olan sınıflar ise azınlık sınıfı ya da minör sınıf olarak adlandırılır.

Dengesiz bir veriye sahip sınıflandırma problemi incelenirken, minör sınıf tipik olarak oldukça dikkat çekicidir. Dolayısıyla modelin azınlık sınıfı olasılığını veya sınıf etiketini doğru bir şekilde tahmin etmesinin çoğunluk sınıfından daha önemli olduğunu gösterir. Birçok sınıflandırma problemi sınıf dağılımında ciddi dengesizliğe sahip olabilir; yine de doğal olarak dengesiz olan ortak sorun alanlarına bakmak, sınıf dengesizliğinin fikirlerini ve zorluklarını somutlaştıracaktır (Coşgun, 2019).

Talep Tahmini, aykırı algılama, izinsiz giriş tespiti, sahtekarlık tespiti, dönüşüm tahmini, anomali tespiti, varsayılan tahmin, spam algılama gibi sorunların olduğu alanların her biri, o alandaki birebir sorunların çalışma alanını temsil etmektedir. Makine öğrenmesi bu spesifik sorun alanına ön görücü bir etkiyle, sorununun çözümüne kolaylaştırıcı bir ışık tutar (Zhou ve ark., 2020).

3.2.1. Sentetik Azınlık Aşırı Örnekleme Tekniği (SAAÖT)

Dengesiz sınıflandırma, sınıf dengesizliğinin çok fazla olduğu sınıflandırma veri kümelerinin üzerinde veri hakkında tahminde bulunmaya dayalı modeller geliştirmeyi çalışır. Veride azınlık sınıfı çoğunlukla uygulanan makine öğrenmesi metotları tarafından göz ardı edilebilmektedir. Bu da dengesiz veri setleri ile çalışmayı zorlaştırmaktadır. Bu nedenle veri kümesinden elde edilecek sonuçlar düşük performans gösterirler. Bu düşük performans çoğunlukla azınlık sınıfından elde edilen performanstan kaynaklanmaktadır (Nguyen ve ark., 2020).

Verinin dengesiz olduğu durumlarda azınlık sınıfına ait örneklerin aşırı örneklenmesi sağlanarak dengesizlik problemi giderilebilmektedir. Bu yaklaşımla veri kümesi içerisinde yer alan azınlık sınıfı örnekleri çoğaltılır. Bu sayede örnek sayısı arttırılmış olur ancak bu veriler modele yeni bilgi eklemezler, zaten var olan örnekleri çoğaltırlar. Dolayısıyla bu bir tür veri arttırma yöntemidir ve bu artırım azınlık sınıfı üzerinden gerçekleştirilmelidir. Bu yöntem SAAÖT olarak adlandırılır

SAAÖT'de azınlık sınıfının her örneği alınır ve alınan örneğe ait komşunun rastgele herhangi birine veya tamamına bakılarak sentetik örnekler oluşturulur. Bu sayede azınlık sınıfına ait örnekler arttırılmış ve veri aşırı örneklenmiş olur (Chawla, Bowyer, Hall ve Kegelmeyer, 2002). Diğer yöntemlere bakıldığında SAAÖT'ü diğerlerinden ayıran özelliği, azınlık sınıfındaki örnekler direkt kopyalanarak arttırılmaz. Bunun yerine azınlık örneğinin yakınında bulunan komşularına bakılır ve o komşular üzerinden sentetik örnekler üretilir.

Dengesiz sınıflandırma ile ilgili bir diğer problem, modelin oluşturacağı karar sınırını etkili bir şekilde öğrenmesi için azınlık sınıfının çok az örneğinin olmasıdır (Dong ve ark., 2020). Bu sorunu çözmenin bir yolu, azınlık sınıfındaki örnekleri aşırı örneklemektir. Bu, bir model yerleştirilmeden önce eğitim veri kümesindeki azınlık sınıfından örnekler kopyalanarak elde edilebilir. Bu sınıf dağılımını dengeleyebilir ancak modele herhangi bir ek bilgi sağlamaz. Bu teknik Nitesh Chawla ve ark. (2002) tarihli *"SAAÖT: Sentetik Azınlık Aşırı Örnekleme Tekniği"* makalelerinde bahsetmişlerdir.

SAAÖT, ilk olarak özellik yüzeyine yakın olan örnekleri belirler, daha sonra özellik yüzeyindeki örnekler arasına bir çizgi çizerek, oluşturulan çizgi üzerindeki herhangi bir noktada yeni bir örnek oluşturur. Alınacak olan örneğin azınlık sınıfına ait olması gerekir. Daha sonra k değerine üzerinden en yakın komşuların tespiti yapılır (k değeri çoğunlukla 5 olarak alınır) bu k değeri üzerinden rastgele bir komşu (b olarak ifade edebiliriz) belirlenir ve özellik yüzeyindeki iki örnek arasında çizilen yüzeyde rastgele sentetik bir örnek üretilir. Sentetik örnekler, seçilen iki a ve b örneğinin dışbükey bir kombinasyonu olarak üretilir.

$$x' = x + rand(0,1) * |x - x_k| \quad (3.1)$$

x': Yeni oluşturulacak veri
rand(0,1): rastgele üretilecek sınıf değeri

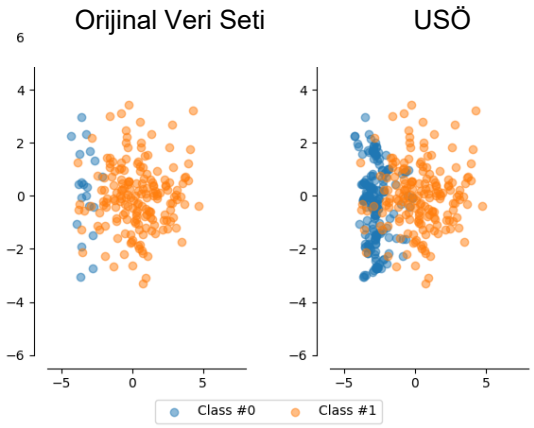
Bu yöntem, sadece azınlık sınıfına sentetik veri üretmek için kullanılmaz. Genel anlamda veri seti üzerinde örnek sayıların arttırılmak istendiğinde de kullanılabilir. Sentetik aşırı örnekleme yöntemi, özellikle sınıflandırıcıya yakın bulunan azınlık sınıf noktalarını bulunduran büyük karar bölgelerinin oluşturmasını sağlar. Böylelikle veri kümesi genişlemiş olur.

3.2.2. Uyarlanabilir Sentetik Örnekleme (USÖ)

Doğal dengesiz sınıf dağılımını barındıran dengesiz veri kümesiyle uğraşmak istatiksel modellemedeki en büyük problemlerden biridir (Zhang ve ark., 2020). Bir makine öğrenimi algoritmasının birincil amacı, bu nadir görülen problemleri çözerek örnekleri kendi sınıflarına başarılı bir şekilde sınıflandırmaktır.

USÖ SAAÖT'un geliştirilmiş bir versiyonudur. Bu örneği oluşturduktan sonra, noktalara rastgele küçük değerler ekleyerek daha gerçekçi hale getirir. Diğer bir deyişle, eğitim verisinin tamamını doğrudan kullanmak yerine parçalı halde kullanır.

USÖ algoritmasının amacı, var olan azınlık sınıfı örnekleri arasındaki doğrusal enterpolasyon yoluyla azınlık sınıfından sentetik olarak yeni örnekler oluşturarak sınıf dengesini geliştirmektir. Bu yaklaşım tek başına SAAÖT yöntemi olarak bilinir. USÖ, SAAÖT'nin bir uzantısıdır ve iki sınıf arasındaki sınır çevresinde azınlık sınıfının iç mekânından daha fazla örnek oluşturur.



Şekil 3.3. Orijinal veri seri ve USÖ uygulanmış veri setine ait dağılım

USÖ algoritması öncelikle eğitim veri kümesini oluşturur. Daha sonras M_r ve m_x sırasıyla azınlık ve çoğunluk sınıfı örneklem sayısı olarak kabul edilir. $m_r \preceq m_x$ ve $m_r + m_x = m$ şartı doğrultusunda işleme devam edilir. Daha sonra aşağıdaki formül aracılığıyla dengesizlik derecesi hesaplanır.

$$d = m_r / m_x \quad (3.2)$$

Daha sonra azınlık sınıfından üretilecek sentetik numune sayısı aşağıdaki formül aracılığıyla hesaplanır. hesaplanır

$$G = (m_x - m_r) \times \beta \quad (3.3)$$

β üretilen sentetik numunelerin denge seviyesidir. $\beta = 1$, iki sınıf arasında maksimum bir denge olduğu anlamına gelir. Ardından her $x_i \in$ azınlık örneği için, öklid mesafesine göre k en yakın komşuları bulunur.

$$r_i, r_i = \Delta_i / K \quad (3.4)$$

Tüm bu hesaplamaların ardından yoğunluk dağılımı hesaplanır. Ki bu hesaplama adımı USÖ'yü SAAÖT'ten ayıran en önemli adımdır. Aşağıda yer alan formül yardımıyla r_x normalize edilir.

$$r_x \leftarrow r_i / \sum r_i \quad (3.5)$$

Bir sonraki adımda her bir azınlık veri noktası için oluşturulan sentetik veri aşağıdaki formülle hesaplanır.

$$g_i = r_x \times G \quad (3.6)$$

G, yukarıda bahsedilen denklemde tanımlandığı gibi azınlık sınıfı için oluşturulması gereken sentetik veri örneklerinin toplam sayısıdır.

Son olarak her bir azınlık sınıfı veri örneği x_i için, sentetik veri örnekleri oluşturulur:

$$s_i = x_i + (x_u - x_i) \times \lambda \quad (3.7)$$

burada $(x_u - x_i)$ n boyutlu uzaylarda fark vektörü ve λ rastgele bir sayıdır: $\lambda \in [0, 1]$.

SAAÖT ve USÖ arasındaki en büyük fark, azınlık veri noktaları için sentetik numune noktalarının oluşturulmasındaki farktır. USÖ'de, belirli bir nokta için üretilecek sentetik örneklerin sayısına karar veren yoğunluk dağılımını r_x dikkate alırız, oysa SAAÖT'de tüm azınlık noktaları için tek tip bir ağırlık vardır.

3.3. Sınıflandırma Yöntemleri

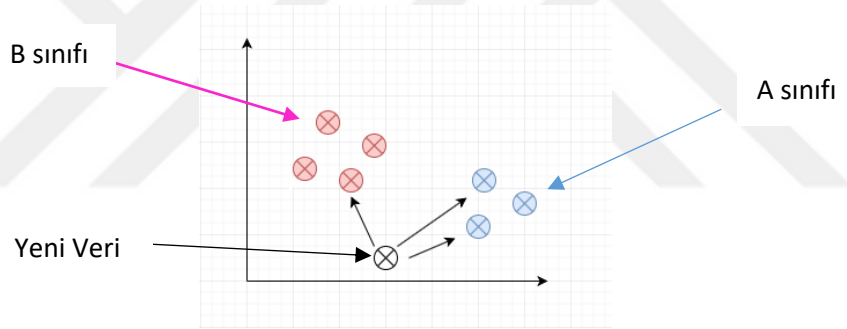
Sınıflandırma, genel olarak bilgisayar programının dış ortamdan girilen verileri öğrenmesi, anlamlandırması ve bu verileri yorumlaması yöntemidir. Bu öğrenme adımından sonra yeni girilen verilerin yorumlarının doğru tahmin edilmesi beklenmektedir. Bu bir denetimli öğrenme yaklaşımı modelidir. Bu veri setleri tek sınıfa sahip olduğu gibi birden çok sınıfa da sahip olabilir.

Makine öğrenmesinde sık kullanılan sınıflandırma algoritmaları türlerimiz lojistik regresyon, NB, DVM , KA ve KEYK'dir.

3.3.1.1. k- En Yakın Komşuluk Sınıflandırma Yöntemi (KEYK)

KEYK algoritması, uygulaması oldukça kolay olan gözetimli öğrenme algoritmalarından biridir. Sınıflandırma problemlerinin yanı sıra regresyon problemlerinin çözümünde de kullanılmaktadır. Endüstride genellikle sınıflandırma problemlerinin çözümlenmesinde kullanılmaktadır (Liao ve ark., 2020).

KEYK algoritmaları, ilk olarak 1967 yılında T. M. Cover ve P. E. Hart tarafından önerilmiştir. Algoritmada sınıfları bilinen bir örnek kümesinde yer alan verilerden yararlanmaktadır. Örnek olarak verilen veri setine yeni eklenecek olan verilerin, mevcut verilere göre uzaklığı hesaplanıp, k ile nitelendirilen sayıda yakın komşuluğuna bakılır. Uzaklık hesapları için genelde euclidean uzaklık, manhattan uzaklık ve minkowski uzaklığı olmak üzere 3 tip uzaklık fonksiyonu kullanılmaktadır.



Şekil 3.4. K en yakın komşuluk tespiti

KEYK; basit, eski ve gürültülü eğitim verilerine karşı dirençli olması sebebiyle en çok kullanılan ve bilinen makine öğrenme algoritmalarının başında yer almaktadır (Wang ve ark., 2020). Fakat bunun yanında olumsuz yönleri de mevcuttur. Örneğin, uzaklık hesabını yaparken mevcut bütün olasılıkları kendi içerisinde sakladığından, daha büyük veri kümelerinde kullanıldığında oldukça fazla bellek kullanılmasına neden olmaktadır.

3.3.1.2. Karar Ağaçları (KA) Sınıflandırma Yöntemi

En sık kullanılan öğrenme algoritması ağaç tabanlı olanlardır (Ranzato ve ark., 2020). KA, incelenmesi gereken bütün problemler için çözüm arayışına cevap

vermektedir. Bu sınıflandırma algoritmalarından biri de KA algoritmasıdır. KA algoritması önceden tanımlanmış olan bir hedef değişken barındırır. Tümden gelim yöntemiyle en baştan başlayarak sona doğru inen bir strateji oluşturur. Karar ağaçlarının kullanılma sebebi içerisinde birden fazla veri barındıran bir kümeyi belli kararlar altında daha küçük kümelere bölmektir. Büyük yapıdaki kayıtları küçük küçük gruplar haline getirir.

Bir KA, fazla sayıda veriyi barındıran veri kümesini kendine ait kurallar çerçevesinde indirgeyerek küçük veri kümelerine bölmeyi sağlamaktadır. Kısacası kolay ve anlaşılabilir karar verme adımları oluşturup bunu uygular.

KA yöntemi, kolayca kullanılıp yorumlandırılır. Yapıları görselleştirerek kolaylık sağlarlar. Çok fazla veri hazırlığına ihtiyaç duymaz. Ağacın maliyeti, ağacı oluşturmak için kullanılan veri noktalarının sayısıyla ilişkilidir. Sayısal ve kategorik veriler aynı anda işlenebilir. Birden fazla çıktısı olan problemleri işleyebilir. KA vasıtasıyla modellerin doğrulanması olanaklıdır.

KA yönteminde, aşırı karmaşık ağaçların üretilmesi durumunda verilerin açıklanması zorlaşabilir (Frosst ve ark., 2017). Ağaç dallarının takibinde problem yaşanabilir. Öğrenmekten ziyade ezberletmeyi hedef alabilir. Bu durumun yaşanmaması için budama yöntemi kullanmak gerekmektedir. Budama yapılırken az nesnesi olan yaprak düğümleri modelden çıkarılarak sadeleştirme anlamlıdır. Bölünmenin nasıl yapılacağına karar vermek ağacın doğruluğunu etkiler. Birden fazla algoritma kullanılan yöntemde her düğümü bölmeye karar vermek için en az iki alt düğüm oluşturulur. Alt düğümlerin meydana gelmesi homojeniteyi arttıran bir unsurdur.

Hedef değişkenin tipine göre algoritma seçilir. Gini, entropi ve en küçük kareler sıklıkla kullanılan yöntemlerdendir. Verilerdeki bilinmezliğin ölçüsü entropi olarak adlandırılır. Veri kümesine ait yalnızca bir etiket varsa daha düşük entropiye sahiptir denilebilir. Bu sebeple entropiyi en yalın haline indirgeyerek bölmemiz gerekmektedir. Bölünmelerin sağlıklı olması tahminimizi de o denli güçlendirecektir.

$$H = -\Sigma\, p(x) \log p(x) \quad (3.8)$$

(Entropiye ait hesaplama denklemi, Claude Shannon, 1948.)

Bu denklemde, p (x) belirlenmiş bir sınıfa ait grubun yüzde oranını, H ise entropiyi ifade etmektedir.

Karar ağacına ait entropi değerini minimuma indirgemeye sağlayan parçalanmalar oluşturmasını isteriz. En iyi parçalanmayı nasıl yapabileceğimizi bu bilgi kazancı değeriyle belirleriz. Bilgi kazancı aşağıda yer aldığı şekliyle hesaplanır.

$$Gain(S,D)=H(S)-\sum_{v\in D}\frac{|v|}{|S|}\cdot H(V) \quad (3.9)$$

Orijinal veri kümemiz S olarak adlandırılırken, kümenin parçalanmış kısmı ise D olarak nitelendirilir. V, ayrık bir kümedir ve S, V'yi kapsar. Buna bağlı olarak bilgi kazancımız, bölünme olmadan önceki veri setimize ait entropi değeri ile her bir özniteliğe ait entropi değeri arasındaki fark bilgi kazancı olarak ifade edilebilir.

Bu algoritmaya göre mevcut sınıflandırma problemlerinde veri seti iki ana parça üzerinden değerlendirilecektir. Bu parçalardan bizi eğitim diğeri ise testtir. Algoritma eğitim verilerinden elde ettiği bilgiler doğrultusunda oluşmaktadır. Meydana gelen bu model ile test verilerindeki değerlerinden anlamlandırılması ve yorumlandırılması sağlanmaktadır. Bu model bize başarı oranlarını vermektedir. ID3, C4.5, CHAID, CART en çok kullanılan KA algoritmalarıdır.

3.3.2. Başarı Ölçütleri

Özellikle hastalık tespiti ve teşhisinde başarı parametrelerin önemi büyüktür. Bu parametrelerin sonuçları ilgili sınıflandırma yönteminin başarısına göre değişiklik göstermektedir. Sağlık alanında veri madenciliği yöntemlerinin tam anlamıyla kullanılabilmesi için bu parametrelere ait terimlerin incelenmesi gerekmektedir.

Duyarlılık, veri setinde bulunan hastaların doğru bir şekilde tespitinin yapılabilmesini sağlayan yetenektir. Hasta olunmadan önce hastalığı ne kadar oranla doğru teşhis ettiğimizi gösterir. Matematiksel olarak ifade edersek:

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{3.10}$$

TP = Hasta teşhisi konulup gerçekten hasta olanlar (doğru teşhis)

TN = Hasta teşhisi konulup hata yapılan ve aslında sağlıklı olanlar

FP = Sağlıklı teşhisi konulup gerçekten sağlıklı olanlar

FN= Sağlıklı teşhisi konulup hata yapılan ve aslında hasta olanlar

Özgüllük, hasta olduğu bilinen kişiler ile sağlıklı hastaların oranıdır. Matematiksel olarak ifade edersek:

$$\text{Özgüllük} = \frac{TN}{TN+FP} \tag{3.11}$$

Hassasiyet, sınıflandırma sisteminin "Hasta" olarak sınıfladığı verilerin kaçının gerçekten hasta olduğunu gösterir. Matematiksel olarak ifade edersek:

$$\text{Hassasiyet} = \frac{TP}{TP+FP} \tag{3.12}$$

Geri Dönüt, hasta olanların doğru tespit edilme oranını belirtir. Matematiksel olarak ifade edersek:

$$\text{Geri Dönüt} = \frac{TP}{TP+FN} \tag{3.13}$$

F ölçütü, hassasiyet ve geri dönüt iki önemli ölçümdür. Aralarında bir bağlantı mevcuttur. Bu sorunu ortadan kaldırabilmek için ise F1 Measure (f1-skoru) kullanılmaktadır. Matematiksel olarak ifade edersek:

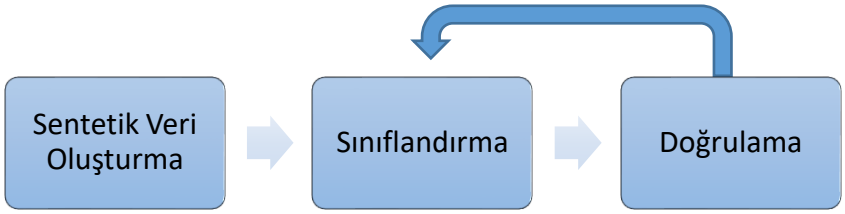
$$F1=2*\frac{\text{Hassasiyet}*\text{G.Dönüt}}{\text{Hassasiyet}+\text{G.Dönüt}} \quad (3.14)$$

G ortalama, değişken bir oran biçimidir.

$$\text{G Ortalama}=\sqrt{\left(\frac{TP}{p}\right)*\left(\frac{TN}{n}\right)} \quad (3.15)$$

4. ARAŞTIRMA BULGULARI ve TARTIŞMA

4.1. Araştırma Bulguları



Şekil 4.1. Sentetik Veri Oluşturma

Bu çalışmada dengesiz veri setleri üzerinden öncelikle sentetik verilerin oluşturulması sağlanacaktır. Daha sonra oluşturulan yeni veri setleri üzerinde sınıflandırma metotları uygulanarak başarı oranları hesaplanacaktır. Tüm bu işlemler ortopedi, meme kanseri ve ameliyat sonrası bakım veri setleri için tekrarlanacaktır.

Bu çalışmada belirlenen sonuçlar intel core i7 işlemcili, 8 GB RAM'e sahip 64 bit işletim sistemli bir bilgisayar üzerinde gerçekleştirilmiştir. Sınıflandırma ve önişleme adımları simülasyon programları üzerinden gerçekleştirilmiştir.

4.1.1. Ortopedi Veri Seti

Veri setimizde 102'si normal sınıfına ait, 208'i ise anormal sınıfına ait toplam 310 tane verimiz mevcuttur. Sınıflar bazında veriler yüzde olarak ifade edilirse; Normal sınıfına ait veriler tüm veri setinin %32,9'unu oluştururken, anormal sınıfına ait veriler ise tüm veri setinin %67.1'ini oluşturmaktadır. Bu durum sınıflara ait verilerin dengesizliği sorununu meydana getirmektir. Bu sorunu ortadan kaldırmak üzere veri setine SAAÖT ve USÖ metotları uygulandı.

Çizelge 4.1. Ortopedi Veri Seti SAAÖT Sonrası Sınıf Dağılımları

Ham Veri		***SAAÖT***	
Normal	Anormal	Normal	Anormal
102	208	204	208

Çizelge 4.2. Ortopedi Veri Seti USÖ Sonrası Sınıf Dağılımları

Ham Veri		***USÖ***	
Normal	Anormal	Normal	Anormal
102	208	225	208

Bu sayede azınlık sınıfına (normal) ait örneklerde artış sağlanarak iki sınıf arasındaki örnek veri sayıları eşitlenmiş olmaktadır. Uygulanan her iki yöntem incelendiğinde azınlık sınıfı için üretilen veri sayıları arasında az da olsa bir fark olduğu gözlemlenmiştir. Bu iki yöntem arasında USÖ, SAAÖT'e nazaran daha fazla sentetik veri üretmiştir. Fakat USÖ iki sınıfa ait örnekler arasındaki eşitliği SAAÖT'e nazaran daha az sağlamış olmaktadır.

Çizelge 4.3. Ortopedi Veri Seti Veri Seti Sınıflandırma Tablosu

SINIFLANDIRMA YÖNTEMİ	***Ham Veri***	***USÖ***	***SAAÖT***
KEYK	84.16	87.99	88.56
Karar Ağacı	80.65	84.99	84.18

Ham veri üzerinde uyguladığımız SAAÖT ve USÖ yöntemleri sonunda oluşturulan sentetik veriler ile yeni veri setleri oluşturuldu ve bu veri setleri sınıflandırma yöntemlerinden KEYK ve KA ile çözümlendi. Sınıflandırma işlemi simülasyon programları üzerinden gerçekleştirilerek veri analizi yapıldı.

Çizelge 4.3. incelendiğinde ham veri setinde en başarılı sonucun KEYK yönteminde elde edildiği görülmektedir. USÖ ve SAAÖT yöntemiyle elde edilen yeni veri setleri incelendiğinde ise en yüksek başarı oranının yine KEYK yönteminde elde edildiği görülmektedir. Sentetik veri oluşturulmadan önce ve oluşturulduktan sonrasına ait başarı ölçütleri aşağıda yer almaktadır.

Çizelge 4.4. Ortopedi Veri Seti KEYK Başarı Ölçütleri

	HAM	***USÖ***	***SAAÖT***
Doğruluk	0.8416	0,8799	0,8856
Duyarlılık	0.7800	0,893	0,9461
Özgüllük	0,8714	0,8654	0,8261
Hassasiyet	0,7429	0,8777	0,8428
Geri Dönüt	0,7800	0,8933	0,9461
F Ölçütü	0,7610	0,8855	0,8915
G Ortalama	0,8244	0,8792	0,8840

Çizelge 4.5. Ortopedi Veri Seti KA Başarı Ölçütleri

	HAM	***USÖ***	***SAAÖT***
Doğruluk	0,8065	0,8499	0,8418
Duyarlılık	0,7100	0,8622	0,8186
Özgüllük	0,8524	0,8365	0,8647
Hassasiyet	0,6961	0,8509	0,8564
Geri Dönüt	0,7100	0,8622	0,8186
F Ölçütü	0,7030	0,8565	0,8371
G Ortalama	0,7779	0,8493	0,8414

Öncelikle duyarlılık sonuçlarını inceleyecek olursak, SAAÖT sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında en yüksek duyarlılık sonucu elde edilmiştir. Öte yandan en düşük hassasiyet oranı ise ham veri setinin KA algoritması ile sınıflandırılmasında görülmüştür. Genel olarak KEYK sınıflandırması bu veri seti üzerinde başarılı hassasiyet değerleri üretmiştir.

Özgüllük değerine ait sonuçları incelersek, USÖ sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırmasında en yüksek özgüllük değeri elde edilmiştir. Öte yandan en düşük özgüllük oranı ise SAAÖT sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında görülmüştür. Genel olarak KA sınıflandırılması bu veri seti üzerinde başarılı özgüllük değerleri üretmiştir.

Hassasiyet değerlerinde ise USÖ sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında en yüksek sonuçlar elde edilmiştir. En düşük sonuçlar ise ham

veri setimizin KA algoritması ile sınıflandırılmasında elde edilmiştir. Genel olarak KEYK sınıflandırması bu veri seti üzerinde başarılı hassasiyet değerleri üretmiştir.

Geri Dönüt parametresine ait sonuçlara baktığımızda; en yüksek sonuca SAAÖT sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında varılmıştır. En düşük sonuç ise ham veri setinin KA algoritması ile sınıflandırılmasında elde edilmiştir. Genel olarak KEYK sınıflandırması bu veri seti üzerinde başarılı sonuçlar elde etmiştir.

F ölçütü sonuçlarında, en yüksek oran SAAÖT sonrası elde edilen veri setinin KEYK algoritmasında, en düşük oran ise ham veri setinin KA algoritmasında elde edilmiştir. KEYK algoritması genel olarak başarılı sonuçlar üretmiştir.

G ortalama değeri USÖ sonrası elde edilen veri setinin KEYK ile sınıflandırılmasında en yüksek, ham verinin KA algoritmasıyla ile sınıflandırılmasında ise en düşük sonuçları üretmiştir. KEYK algoritması genel olarak G ortalama değeri için başarılı olduğu sonuçlar üretmiştir.

4.1.2. Meme Kanseri Veri Seti

Veri setimizde 201'i tekrarlamayan sınıfına ait, 85'i ise tekrarlayan sınıfına ait toplam 286 tane veri mevcuttur. Sınıflar bazında veriler yüzde olarak ifade edilirse; tekrarlamayan sınıfına ait veriler tüm veri setinin %70,28'ini oluştururken, tekrarlayan sınıfına ait veriler ise tüm veri setinin %29.72'sini oluşturmaktadır. Bu durum sınıflara ait verilerin dengesizliği sorununu meydana getirmektir. Bu sorunu ortadan kaldırmak üzere veri setine SAAÖT ve USÖ metotları uygulanmıştır.

Çizelge 4.6. Meme Kanseri Veri Seti SAAÖT Sonrası Sınıf Dağılımları

Ham Veri		***SAAÖT***	
Tekrarlayan	Tekrarlamayan	Tekrarlayan	Tekrarlamayan
85	201	170	201

Çizelge 4.7. Meme Kanseri Veri Seti USÖ Sonrası Sınıf Dağılımları

Ham Veri		***USÖ***	
Tekrarlayan	Tekrarlamayan	Tekrarlayan	Tekrarlamayan
85	201	190	201

Bu sayede azınlık sınıfına ait örneklerde artış sağlanarak iki sınıf arasındaki örnek veri sayıları eşitlenmiş olmaktadır. Uygulanan her iki yöntem incelendiğinde azınlık sınıfı için üretilen veri sayıları arasında az da olsa bir fark olduğu gözlemlenmiştir. Bu iki yöntem arasında USÖ, SAAÖT'e nazaran daha fazla sentetik veri üretmiştir.

Çizelge 4.8. Meme Kanseri Veri Seti Sınıflandırma Tablosu

SINIFLANDIRMA YÖNTEMİ	***Ham Veri***	***USÖ***	***SAAÖT***
KEYK	69.93	78.21	56.87
Karar Ağacı	68.53	72.82	78.44

Ham veri seti üzerinde uyguladığımız SAAÖT ve USÖ yöntemleri sonunda oluşturulan sentetik veriler ile yeni veri setleri oluşturuldu ve bu veri setleri sınıflandırma yöntemlerinden KEYK ve KA ile çözümlendi. Sınıflandırma işlemi simülasyon programları üzerinden gerçekleştirilerek veri analizi yapıldı.

Çizelge 4.8. incelendiğinde ham veri setinde en başarılı sonucun KEYK yönteminde elde edildiği görülmektedir. USÖ yöntemiyle elde edilen yeni veri seti incelendiğinde en yüksek başarı oranının yine KEYK yönteminde elde edildiği görülmektedir. Öte yandan SAAÖT yöntemiyle elde edilen veri setimiz incelendiğinde ise en yüksek başarı oranının KA yönteminde elde edildiği anlaşılmaktadır.

Sentetik veri oluşturulmadan önce ve oluşturulduktan sonrasına ait başarı ölçütleri aşağıda yer almaktadır.

Çizelge 4.9. Meme Kanseri Veri Seti KEYK Başarı Ölçütleri

	Ham Veri	***USÖ***	***SAAÖT***
Doğruluk	0,6993	0,7821	0,5687
Duyarlılık	0,2706	0,8307	0,2000
Özgüllük	0,8805	0,7363	0,8806
Hassasiyet	0,4894	0,7476	0,5862
Geri Dönüt	0,2706	0,8307	0,2000
F Ölçütü	0,3485	0,7870	0,2982
G Ortalama	0,4881	0,7821	0,4197

Çizelge 4.10. Meme Kanseri Veri Seti KA Başarı Ölçütleri

	Ham Veri	***USÖ***	***SAAÖT***
Doğruluk	0,6853	0,7282	0,7844
Duyarlılık	0,4471	0,7037	0,7647
Özgüllük	0,7861	0,7512	0,8010
Hassasiyet	0,4691	0,7268	0,7647
Geri Dönüt	0,4471	0,7037	0,7647
F Ölçütü	0,4578	0,7151	0,7647
G Ortalama	0,5928	0,7271	0,7826

Öncelikle duyarlılık değeri sonuçlarını inceleyecek olursak, USÖ ile elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında en yüksek hassasiyet sonucu elde edilmiştir. Öte yandan en düşük hassasiyet oranı ise SAAÖT sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında görülmüştür. Genel olarak KA sınıflandırması bu veri seti üzerinde başarılı hassasiyet değerleri üretmiştir.

Özgüllük değerine ait sonuçları incelersek, SAAÖT sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırmasında en yüksek özgüllük değeri elde edilmiştir. Öte yandan en düşük özgüllük oranı ise USÖ sonrası elde edilen veri setinin KA algoritması ile sınıflandırılmasında görülmüştür. Genel olarak KEYK sınıflandırılması bu veri seti üzerinde optimum özgüllük değerleri üretmiştir.

Hassasiyet değerlerinde ise SAAÖT sonrası elde edilen veri setinin KA algoritması ile sınıflandırılmasında en yüksek sonuçlar elde edilmiştir. En düşük sonuçlar ise ilk veri setimizin KA algoritması ile sınıflandırılmasında elde edilmiştir.

Genel olarak KA sınıflandırması bu veri seti üzerinde başarılı hassasiyet değerleri üretmiştir.

Geri Dönüt parametresine ait sonuçlara baktığımızda; en yüksek sonuca USÖ sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında varılmıştır. En düşük sonuç ise SAAÖT sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında elde edilmiştir. Genel olarak KA sınıflandırması bu veri seti üzerinde iyi sonuçlar elde etmiştir.

F ölçütü sonuçlarında, en yüksek oran USÖ sonrası elde edilen veri setinin KEYK algoritmasında, en düşük oran ise SAAÖT sonrası elde edilen veri setinin KEYK algoritmasında elde edilmiştir. KA algoritması genel olarak başarılı sonuçlar üretmiştir.

G ortalama SAAÖT sonrası elde edilen veri setinin KA ile sınıflandırılmasında en yüksek, SAAÖT sonrası elde edilen veri setinin KEYK algoritmasıyla ile sınıflandırılmasından en düşük sonuçlarını almıştır. KA genel olarak G ortalama değerinin optimum olduğu sonuçlar üretmiştir.

4.1.3. Ameliyat Sonrası Bakım Veri Seti

Veri setimizde A sınıfına ait 63 örnek, S sınıfına ait 24 örnek ve I sınıfına ait ise 2 örnek bulunmaktadır. Toplamda 89 tane verimiz mevcuttur. Sınıflar bazında veriler yüzde olarak ifade edilirse; A sınıfına ait veriler tüm veri setinin %70,78'ini, S sınıfına ait veriler %26,96'sını I sınıfına ait veriler ise tüm veri setinin %2.25'ini oluşturmaktadır. Bu durum sınıflara ait verilerin dengesizliği sorununu meydana getirmektir. Bu sorunu ortadan kaldırmak üzere veri setine SAAÖT ve USÖ metotları uygulandı.

Çizelge 4.11. Ameliyat Sonrası Bakım Veri Seti SAAÖT Sonrası Sınıf Dağılımları

Ham Veri			***SAAÖT***		
A	S	I	A	S	I
63	24	2	63	48	64

Çizelge 4.12. Ameliyat Sonrası Bakım Veri Seti USÖ Sonrası Sınıf Dağılımları

Ham Veri			***USÖ***		
A	S	I	A	S	I
63	24	2	63	64	64

Bu sayede azınlık sınıfına (tekrarlayan) ait örneklerde artış sağlanarak üç sınıf arasındaki örnek veri sayıları eşitlenmiş olmaktadır. Uygulanan her iki yöntem incelendiğinde azınlık sınıfı için üretilen veri sayıları arasında az da olsa bir fark olduğu gözlemlenmiştir. Bu iki yöntem arasında USÖ, SAAÖT'e nazaran daha fazla sentetik veri üretmiştir.

Çizelge 4.13. Ameliyat Sonrası Bakım Veri Seti Sınıflandırma Tablosu

SINIFLANDIRMA YÖNTEMİ	***Ham Veri***	***USÖ***	***SAAÖT***
KEYK	55.05	78.01	77.59
Karar Ağacı	62.92	79.06	83.33

İlk veri seti üzerinde uyguladığımız SAAÖT ve USÖ yöntemleri sonunda oluşturulan sentetik veriler ile yeni veri setleri oluşturuldu ve bu veri setleri sınıflandırma yöntemlerinden KEYK ve KA ile çözümlendi. Sınıflandırma işlemi simülasyon programları üzerinden gerçekleştirilerek veri analizi yapıldı.

Çizelge 4.13. incelendiğinde ham veri setinde en başarılı sonucun KA yönteminde elde edildiği görülmektedir. USÖ yöntemiyle elde edilen yeni veri seti incelendiğinde ise en yüksek başarı oranının yine KA yönteminde elde edildiği görülmektedir.

Çizelge 4.13. 'de SAAÖT yöntemiyle elde edilen yeni veri setimiz incelendiğinde ise en yüksek başarı oranının yine KA yönteminde elde edildiği görülmektedir.

Sentetik veri oluşturulmadan önce ve oluşturulduktan sonrasına ait başarı ölçütleri aşağıda yer almaktadır.

Çizelge 4.14. Ameliyat Sonrası Bakım Veri Seti KEYK Başarı Ölçütleri

	HAM	***USÖ***	***SAAÖT***
Doğruluk	0,5505	0,7801	0,7759
Duyarlılık	0,0417	0,7031	0,5319
Özgüllük	0,7385	0,8189	0,8661
Hassasiyet	0,0555	0,6618	0,5952
Geri Dönüt	0,0417	0,7031	0,5319
F Ölçütü	0,0476	0,6818	0,5618
G Ortalama	0,1754	0,7588	0,6788

Çizelge 4.15. Ameliyat Sonrası Bakım Veri Seti KA Başarı Ölçütleri

	HAM	***USÖ***	***SAAÖT***
Doğruluk	0,6292	0,7906	0,8333
Duyarlılık	0,3333	0,6406	0,6170
Özgüllük	0,7385	0,8661	0,9134
Hassasiyet	0,3200	0,7069	0,7250
Geri Dönüt	0,3333	0,6406	0,6170
F Ölçütü	0,3265	0,6721	0,6667
G Ortalama	0,4961	0,7449	0,7507

Öncelikle duyarlılık değeri sonuçları inceleyecek olursak, USÖ sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında en yüksek hassasiyet sonucu elde edilmiştir. Öte yandan en düşük hassasiyet oranı ise ham veri setinin KEYK algoritması ile sınıflandırılmasında görülmüştür. Genel olarak KA sınıflandırması bu veri seti üzerinde başarılı hassasiyet değerleri üretmiştir.

Özgüllük değerine ait sonuçları incelersek, SAAÖT sonrası elde edilen veri setinin KA algoritması ile sınıflandırmasında en yüksek özgüllük değeri elde edilmiştir. Öte yandan en düşük özgüllük oranı ise ham veri setinin KA algoritması ile sınıflandırılmasında görülmüştür. Genel olarak KA sınıflandırılması bu veri seti üzerinde başarılı özgüllük değerleri üretmiştir.

Hassasiyet değerlerinde ise SAAÖT sonrası elde edilen veri setinin KA algoritması ile sınıflandırılmasında en yüksek sonuçlar elde edilmiştir. En düşük

sonuçlar ise ham veri seti KEYK algoritması ile sınıflandırılmasında elde edilmiştir. Genel olarak KA sınıflandırması bu veri seti üzerinde başarılı hassasiyet değerleri üretmiştir.

Geri Dönüt parametresine ait sonuçlara baktığımızda; en yüksek sonuca USÖ sonrası elde edilen veri setinin KEYK algoritması ile sınıflandırılmasında varılmıştır. En düşük sonuç ise ham veri setinin KEYK algoritması ile sınıflandırılmasında elde edilmiştir. Genel olarak KA sınıflandırması bu veri seti üzerinde iyi sonuçlar elde etmiştir.

F ölçütü sonuçlarında, en yüksek oran USÖ sonrası elde edilen veri setinin KEYK algoritmasında, en düşük oran ise ham veri setinin KEYK algoritmasında elde edilmiştir. DT algoritması genel olarak optimum sonuçlar üretmiştir.

G ortalama USÖ sonrası elde edilen veri setinin KEYK ile sınıflandırılmasında en yüksek, ham veri setinin KEYK algoritmasıyla ile sınıflandırılmasından en düşük sonuçlarını almıştır. KA genel olarak G ortalama değeri için başarılı sonuçlar üretmiştir.

5. SONUÇLAR ve ÖNERİLER

Bu çalışmada 3 farklı veri tipinde veri seti kullanılmıştır. Bu veri setlerinden ortopedi veri seti numerik, kanser veri seti kategorik ve son olarak ameliyat sonrası bakım veri seti ise hem kategorik hem de numerik yapıya sahipti.

Bu veri setleri tümüyle dengesiz veri seti türünde olan veri dağılımlarına sahip olduğu için dengesizlik giderme yöntemleri uygulanmış ve aşağıda yer alan sonuçlar gözlemlenmiştir. Kanser veri seti üzerinde sentetik veri üretilmediği durumda yüksek başarı oranı % 68.53 ile KA yönteminde elde edilmiştir.

SAAÖT ile daha fazla yapay veri üretimi sağlanmış ve sınıflar arasındaki dengeden kaynaklı olarak karar ağacı yöntemi sayesinde hesaplanan entropi değerleri de bu bağlamda yakın sonuçların oluşmasını sağlamıştır. Bu sayede KA yönteminin kategorik veride optimum olarak çalıştığı görülmektedir.

KA yöntemi hem kategorik hem de sayısal verilerin sınıflandırılmasında kullanılır. İlgili veri setimiz her iki veriyi de kendi içerinde barındırmaktadır. Bu sayede KA yönteminin hem kategorik hem sayısal verinin bir arada bulunduğu veri setinde optimum olarak çalıştığı görülmektedir.

Genel olarak, kullanılan veri setlerinden elde edilen sonuçlara göre; Yapay veri üretiminde sınıflar arasındaki dengesizliği en aza indirgeyen yöntem SAAÖT'tir. USÖ yöntemiyle veri üretiminde veri setinin kategorik ve sayısal veriyi bir arada içerdiği durumlarda daha başarılı sonuçlar elde etmiştir. SAAÖT yöntemi sadece kategorik veya sadece sayısal veriye sahip veri setlerinde başarılı sonuçlar elde etmiştir.

Sayısal medikal veri setinde, KEYK yöntemi ile sınıflandırması yapılırken sentetik veri üretiminde USÖ yönteminin başarılı sonuçlar verdiği gözlemlenmiştir.

Sayısal medikal veri setinde, KA yöntemi ile sınıflandırması yapılırken sentetik veri üretiminde SAAÖT yönteminin başarılı sonuçlar verdiği gözlemlenmiştir.

Kategorik medikal veri setinde, KEYK yöntemi ile sınıflandırması yapılırken sentetik veri üretiminde USÖ yönteminin başarılı sonuçlar verdiği gözlemlenmiştir.

Kategorik medikal veri setinde, KA yöntemi ile sınıflandırması yapılırken sentetik veri üretiminde SAAÖT yönteminin başarılı sonuçlar verdiği gözlemlenmiştir.

Hem sayısal hem kategorik olan medikal veri setinde, KA yöntemi ile sınıflandırması yapılırken sentetik veri üretiminde USÖ yönteminin başarılı sonuçlar verdiği gözlemlenmiştir.

Hem sayısal hem kategorik olan medikal veri setinde, KA yöntemi ile sınıflandırması yapılırken sentetik veri üretiminde SAAÖT yönteminin başarılı sonuçlar verdiği gözlemlenmiştir.

İlerleyen çalışmalarda benzer yöntemler kullanılarak yeni bir aşırı örnekleme yöntemi önerilmesi konusu ele alınacaktır. Bunun yanında sentetik veri üretme yerine veri azaltmanın sınıflandırma üzerindeki etkisi karşılaştırılacaktır.

KAYNAKLAR

AHMAD, L. G., ESHLAGHY, A. T., POOREBRAHIMI, A., EBRAHIMI, M. and RAZAVI, A. R., 2013. Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(124), 3.

AKBEN, S. B., 2016. Importance of the shape and orientation of the spine and pelvis for the vertebral column pathologies diagnosis with using machine learning methods.

ALOMARI, R. S., CHAUDHARY, V. and Dhillon, G., 2011. Computer aided diagnosis system for lumbar spine. In Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies, pp.1-5.

ANSARI, S., SAJJAD, F., NAVEED, N. and Shafi, I. ,2013. Diagnosis of vertebral column disorders using machine learning classifiers. In 2013 International Conference on Information Science and Applications (ICISA), pp. 1-6.

ANSARI, S., SAJJAD, F., NAVEED, N. and SHAFI, I., 2013. Diagnosis of vertebral column disorders using machine learning classifiers. In 2013 International Conference on Information Science and Applications, pp. 1-6.

ASRI, H., MOUSANNIF, H., AL MOATASSIME, H. and NOEL, T., 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064-1069.

BULUT, F. Dengesiz Veri Setlerinde Denetimli Öğrenicilerin Başarım Değerlendirmesi/Performance Evaluations of Supervised Learners on Imbalanced Datasets.

BULUT, F., 2016. Sınıflandırıcı Topluluklarının Dengesiz Veri Kümeleri Üzerindeki Performans Analizleri. Bilişim Teknolojileri Dergisi, 9(2):153.

CAMPBELL, M.P., BECHTOLD, J.E. and ERDMAN, A.G., 1995. Implantable limb lengthening nail driven by a shape memory alloy. U.S. Patent No 5,415,660.

ÇELİK, S., BOZKURT, Ö. Ç. ve ÇEŞMELİ, M. Ş., 2018. İnsan Omurgası Veri Setinin Sinir-Bulanık Sınıflayıcı İle Öznitelik Tespiti ve Sınıflandırılması. Yönetim Bilişim Sistemleri Dergisi, 4(1), 39-52.

CHAWLA, N. V., 2009. Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, pp.875-886.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O. and KEGELMEYER, W. P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321-357.

CHO, S. B. and WON, H. H., 2003. Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003, 19 :189-198.

CRUZ, J. A. and WISHART, D. S., 2006. Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2p.

DA ROCHA NETO, A. R., SOUSA, R., BARRETO, G. D. A. and CARDOSO, J. S., 2011. Diagnostic of pathology on the vertebral column with embedded reject option. In Iberian Conference on Pattern Recognition and Image Analysis, pp.588-595

DOĞAN, Ş., 2007. Veri madenciliği kullanarak biyokimya verilerinden hastalık teşhisi/An analyse of a deseases with biochemical data by using data mining.

ESRA, Ö., AKÇAPINAR SEZER, E. and SEVER, H., 2019. Sınıflandırmada Küçük ve Dengesiz Veri Kümesi Problemi.

FRITZ, B. A., CHEN, Y., MURRAY-TORRES, T. M., GREGORY, S., ABDALLAH, A. B., KRONZER, A. and AVIDAN, M. S., 2018. Using machine learning techniques to develop forecasting algorithms for postoperative complications: protocol for a retrospective study. BMJ open, 8(4).

GÜVEN, Y. and BARKANA, D. E., 2010. Medical user interface for orthopedical surgical robotic system. In 2010 15th National Biomedical Engineering Meeting, pp.1-5.

HIDAYAH, I., ADHISTYA, E. P. and KRISTY, M. A., 2014. Application of J48 and bagging for classification of vertebral column pathologies. In Proceedings of the 6th International Conference on Information Technology and Multimedia, pp. 314-317

JAPKOWICZ, N., 2000. Learning from imbalanced data sets: a comparison of various strategies. In AAAI workshop on learning from imbalanced data sets, 68:10-15.

JEATRAKUL, P., WONG, K. W. and FUNG, C. C., 2010. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In International Conference on Neural Information Processing, pp.152-159.

KAYA, M. Evaluation Of Sagittal Spinopelvic Parameters Before And After Lumbar Stabilization Surgeries. 57:11-5.

KOTTI, M. and DIAMANTARAS, K. I., 2012. Towards minimizing the energy of slack variables for binary classification. In 2012 Proceedings of the 20th European Signal Processing Conference, pp.644-648.

MANGIONE, P., SENEGAS, J., 1997, Sagittal balance of the spine. Revue de chirurgie orthopedique et reparatrice de l'appareil moteur, 83.1: 22-32.

NIZAM, H. ve AKIN, S., 2014. Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. XIX. Türkiye'de İnternet Konferansı, s.1-6.

PAREKH, V. S. and JACOBS, M. A., 2017. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. NPJ breast cancer, 3(1):1-9.

PATIL, S. and JAYAKUMAR, T., 2019. Spino-pelvic radiological parameters in patients presenting with low back pain in a tertiary care centre in India. International Journal of Orthopaedics, 5(2):172-176.

PROVOST, F., 2000. Machine learning from imbalanced data sets 101. In Proceedings of the AAAI'2000 workshop on imbalanced data sets, No. 2000, 1-3

RAJA'S, A., CORSO, J. J., CHAUDHARY, V. and DHILLON, G., 2011. Toward a clinical lumbar CAD: herniation diagnosis. International journal of computer assisted radiology and surgery, 6(1):119-126.

REDDY, S. K., KODALI, S. R. and GUNDABATHINA, J. L., 2012. Classification of vertebral column using naïve bayes technique. International Journal of Computer Applications, 58:7.

SERHATLIOĞLU, S., 2016, Tıbbi Bilişim ve Bilgisayar Destekli Tanı. Fırat Üniversitesi Sağlık Bilimleri Tıp Dergisi, 30.1: 39-42.

SUN, J., LANG, J., FUJITA, H. and LI, H., 2018. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 425:76-91.

TIGHE, P. J., HARLE, C. A., HURLEY, R. W., AYTUG, H., BOEZAART, A. P. and FILLINGIM, R. B., 2015. Teaching a machine to feel postoperative pain: combining high-dimensional clinical data with machine learning algorithms to forecast acute postoperative pain. Pain Medicine, 16(7): 1386-1401.

UNAL, Y. and KOCER, H. E., 2013. Diagnosis of pathology on the vertebral column with backpropagation and Naïve Bayes classifier. In 2013 the international conference on technological advances in electrical, electronics and computer engineering (TAEECE), pp. 276-279.

UNAL, Y., & KOCER, H. E., 2013. Diagnosis of pathology on the vertebral column with backpropagation and Naïve Bayes classifier. In 2013 the international conference on technological advances in electrical, electronics and computer engineering (TAEECE), pp. 276-279.

UNAL, Y., POLAT, K. and KOCER, H.E., 2014. Pairwise FCM based feature weighting for improved classification of vertebral column disorders, Computers in biology and medicine, 46:61-70.

UTKAN, A., DAYICAN, A., TOYRAN, A. ve TÜMÖZ, M. A., 2006. Ortopedi ve travmatoloji hastalarında hepatit B, hepatit C ve HIV seroprevalans. Acta Orthop Traumatol Turc, 40(5):367-370.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı **:** Mehmet ŞEYHANLI

EĞİTİM

Derece	**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	: Urfa Anadolu Lisesi, Şanlıurfa	2013
Üniversite	: Harran Üniversitesi Mühendislik Fakültesi Elektrik-Elektronik Mühendisliği Bölümü, Şanlıurfa	2017
Yüksek Lisans	: Harran Üniversitesi Fen Bilimleri Enstitüsü Elektrik-Elektronik Mühendisliği Anabilim Dalı, Şanlıurfa	2021

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2018-2018	Harran Üniversitesi	Proje Araştırmacısı
2018-2019	Şanlıurfa Teknokent	Proje Yöneticisi
2019-Devam Ediyor	Türkiye Kalkınma Vakfı	Koordinatör

YAYINLAR

GÜMÜŞÇÜ, A. ve ŞEYHANLI, M., 2018. Sosyal Medya Ortamında Kullanılan Enerji Kavramlarının Karşılaştırılması. GAPYENEV-2018 International Renewable Energy and Energy Efficiency Congress, 10-12 May, Şanlıurfa, 1(1):357-357.

GÜMÜŞÇÜ, A. ve ŞEYHANLI, M., 2019. Adaptive Synthetic sampling approach for orthopedic diseases classification. 5TH Internatıonal Conference On Engineering and natural sciences,12-16 june, Prague

GÜMÜŞÇÜ, A., ŞEYHANLI, M. ve GÖNCÜ, Z.T., 2020. Yenilenebilir Enerji Santrali Elemanlarının AR ile Eğitimde Kullanılması. GAPYENEV-2020 II. Uluslararası Gap Yenilenebilir Enerji ve Enerji Verimliliği Kongresi, 5-6 Kasım, Şanlıurfa.

YAPRAK, M., GÜMÜŞÇÜ, A., ŞEYHANLI, M. ve ALTUNDAĞ, C., 2020. Akıllı Şehir Uygulamaları için İklim Verileri ile Enerji Talebi Sınıflandırılması. GAPYENEV-2020 II. Uluslararası Gap Yenilenebilir Enerji ve Enerji Verimliliği Kongresi, 5-6 Kasım, Şanlıurfa.