


# Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı

*Araştırma Makalesi/Research Article*

 Göksu AKGÜL,  Ali Akın ÇELİK,  Zeliha ERGÜL AYDIN,  Zehra KAMIŞLI ÖZTÜRK

Endüstri Mühendisliği, Eskişehir Teknik Üniversitesi, Eskişehir, Türkiye  
[goksuakgul@eskisehir.edu.tr](mailto:goksuakgul@eskisehir.edu.tr), [aliakincelik@eskisehir.edu.tr](mailto:aliakincelik@eskisehir.edu.tr), [zergul@eskisehir.edu.tr](mailto:zergul@eskisehir.edu.tr), [zkamisli@eskisehir.edu.tr](mailto:zkamisli@eskisehir.edu.tr)  
(Geliş/Received:28.03.2020; Kabul/Accepted:18.06.2020)

DOI: 10.17671/gazibtd.710728

**Özet**— Hastalık teşhisi, tıp alanında karşılaşılan en önemli problemlerden biridir. Belirli bir hastalığın farklı türlerinin ve diğer hastalıklarla benzer semptomlarının olması hastalığın teşhisini zorlaştırmaktadır. Tiroit hastalığı çeşitlerinden biri olan hipotiroidi de bu sebeplerle teşhisi geciken ve hastaların yaşam kalitesini düşüren bir hastalıktır. Bu çalışmanın amacı, tanı sürecinde hastalara sorulan soru ve uygulanan test sonuçlarını kullanarak hipotiroidi hastalığının doğru teşhis oranını arttıracak veri madenciliği temelli bir sistem önermektir. Diğer amaç ise dolaylı olarak teşhis için kullanılan girişimsel testlerden oluşabilecek komplikasyonları azaltmaktır. Bu amaçlar doğrultusunda UCI makine öğrenmesi veri tabanında yer alan ve 151 tanesi hipotiroidi geri kalanı hipotiroidi olmayan toplam 3163 örnekten oluşan veri seti kullanılarak yeni örneklerin hipotiroidi olup olmadığı tahmin edilmiştir. Veri setindeki dengesiz dağılımı ortadan kaldırmak için veri setine farklı örnekleme teknikleri uygulanarak Lojistik Regresyon, K En Yakın Komşu ve Destek Vektör Makinesi sınıflandırıcıları ile hipotiroidi hastalığını teşhis edecek modeller oluşturulmuştur. Bu yönüyle, çalışma örnekleme yöntemlerinin hipotiroidi hastalığı teşhisi üzerindeki etkisini göstermiştir. Geliştirilen modeller içinde en yüksek performansı, aşırı örnekleme teknikleri uygulanan veri seti ile eğitilen Lojistik Regresyon sınıflandırıcısı vermiştir. Bu sınıflandırıcı ile elde edilen en iyi sonuçlar; doğruluk oranı için %97.8, F-Skor değeri için %82.26, eğri altında kalan alan için %93.2 ve Matthews korelasyon katsayısı için de %81.8'dir.

**Anahtar Kelimeler**— Hastalık teşhisi, Hipotiroidi, Veri Madenciliği, Lojistik Regresyon, K En Yakın Komşu, Destek Vektör Makineleri

## Use of Classification Algorithms in Diagnosis of Hypothyroidism

**Abstract**— Disease diagnosis is one of the most important problems encountered in the medical field. Different types of a specific disease and similar symptoms with other diseases make the disease harder to diagnose. For these reasons Hypothyroidism, which is one of the types of thyroid disease, is a disease that decreases patient's quality of life due to the delay in its diagnosis. The purpose of this article is to propose a data mining-based system that will increase the correct diagnosis of hypothyroidism rate by using the question asked to the patients during the diagnosis process, and the test results applied. The other aim is to reduce the complications that may arise from interventional tests used indirectly for diagnosis. For these purposes, it was estimated whether new samples were hypothyroidism by using a data set consisting of 3163 samples in the UCI machine learning database, 151 of which were hypothyroid and the rest without hypothyroidism. In order to deal with the imbalanced class distribution in the data, different sampling techniques were applied to the data set and models to diagnose hypothyroidism with Logistic Regression, K Nearest Neighbor, and Support Vector Machine classifiers were created. With this aspect, the study demonstrated the effect of sampling methods on the diagnosis of hypothyroid disease. Among the developed models, the Logistics Regression classifier, which was trained with the data set applied to the oversampling techniques, gave the highest performance. The best results obtained with this classifier are 97.8% for accuracy rate, 82.26% for F-Score value, 93.2% for area under the curve and 81.8% for Matthews correlation coefficient.

**Keywords**— Disease diagnosis, Hypothyroidism, Data Mining, Logistic Regression, K Nearest Neighborhood, Support Vector Machine

## 1. GİRİŞ (INTRODUCTION)

Hastaların yaşam kalitesini oldukça düşüren ve genel olarak aşırı hormonal aktivite ya da düşük hormon üretimine bağlı semptomlarla seyreden tiroit hastalıkları, en sık karşılaşılan endokrin problemlerinden biridir. Tiroit hastalıkları nodül oluşumu veya tümör gelişimi ile de ortaya çıkabilir [1]. Tiroit bezi tarafından salgılanan hormonlar metabolik faaliyetlere, dokuların gelişip büyümesine, enerji sağlanması için besinlerin kullanılma hızına ve canlı ağırlık kazancına etkili olduklarından [2], tiroit hastalıklarının uygun tedavisi için doğru tanı büyük önem arz etmektedir. Ancak, tiroit hastalığının birçok çeşidinin olması ve bazı noktaların gözden kaçırılması doktorların hastalara doğru teşhis koymasını zorlaştırmaktadır.

Günümüzde sağlık verisini içeren büyük veri depolarının geliştirilmesi ve sürdürülmesi ile, sağlık kuruluşları, operasyonel ve klinik kararları iyileştirmek için bu yapılandırılmış ve yapılandırılmamış verilerde bulunan örüntüleri ve ilişkileri analiz etmek ve kullanmak üzere veri madenciliğini daha fazla kullanmaktadır. Makine öğrenmesi, istatistiksel analiz, modelleme teknikleri ve veri tabanı teknolojilerinde kullanılan farklı teknikleri tanımlayan bir terim olan veri madenciliği, verilerde farklı türde yapı ve ilişkileri bulmayı, kurallar elde etmeyi ve yeni durumlar için tahmini mümkün kılmaktadır. Sınıflandırma teknikleri de bir tür kategorik veya sayısal değer tahmin edilmesine yardımcı olabilecek veri madenciliği teknikleridir. Veri madenciliği, müşteri segmentasyonu, ürün tasarımı, müşterilerin tercihlerini anlama ve tahmin etme, risk yönetimi, hastalık teşhisi, öğrenci başarısı tahmini ve spor müsabakaları skor tahmini gibi birçok alanda kullanılmaktadır.

Hastalık teşhisi, veri madenciliği araçlarının başarılı sonuçlar verdiği uygulamalardan biridir. Doğru hastalık teşhisi, hastanın tıbbi öyküsünün alınması, muayene bulguları ve çeşitli biyobelirteçler (öznitelikler) ile elde edilebilir.

Bu çalışmada, hipotiroidi hastalığının doktorlar tarafından teşhis edilmesine yardımcı olmak amacıyla veri madenciliği tekniklerinden olan farklı sınıflandırma algoritmaları kullanılmıştır. UCI [3] veri tabanından alınan hipotiroidi veri setinde sağlıklı ve hastalıklı örneklerin dağılımındaki dengesizliği ortadan kaldırmak için de farklı örnekleme teknikleri uygulanarak sınıflandırıcı başarılarının artırılması hedeflenmiştir.

İzleyen bölümde sağlık alanında ve özellikle tiroit hastalığı teşhisinde yapılmış veri madenciliği çalışmaları incelenmiştir. Veri setine ait bilgiler üçüncü bölümde verilmiştir. Ardından, çalışmada kullanılan veri madenciliği yöntemlerinden bahsedilmiştir. Elde edilen deneysel sonuçlar dördüncü bölümde tartışılmıştır. Son bölümde ise çalışmanın katkılarına ve gelecek çalışmalara değinilmiştir.

## 2. YAZIN TARAMASI (LITERATURE REVIEW)

Kaya vd. [4] çalışmalarında, sağlık alanında kullanılan yapay zeka ve derin öğrenme yöntemlerini biyoinformatik uygulamaları, medikal görüntüleme uygulamaları [5] ve medikal bilişim [6] uygulamaları ana başlıkları altında incelemiştir. Buna göre; kanser teşhisi, gen seçimi ve ilaç tasarımı problemleri biyoinformatik uygulamalarına [7]; tümör tespiti, doku sınıflandırması ve organ bölümlenmesi medikal görüntüleme uygulamalarına; hastalık tahmini de medikal bilişim uygulamaları altında ele alınan problemlere örnek olarak verilebilir.

Hastalık teşhisinin erken olmasının yanı sıra yüksek doğrulukta olması da çok önemlidir. Yazında hastalık teşhisi konusunda yapılan veri madenciliği çalışmalarının sayısı, teşhis edilen hastalık türüne bağlı olarak çok sayıdadır. Örneğin, obstrüktif uyku apnesi tanıma için yapılan çalışmada [8], öncelikle öznitelik boyutu azaltılmış, ardından da Naive Bayes (NB), K En Yakın Komşu (KNN) ve Destek Vektör Makineleri (DVM) sınıflandırıcıları ile hastalığın teşhisi için bir sistem önerilmiştir. Benzer şekilde Alpaslan [9], meme kanseri teşhisi için önce öznitelik boyutunu azaltmış ardından da test görüntülerini sınıflandırmak için Aşırı Öğrenme Makinesi (AÖM) yöntemini kullanmıştır. Pala vd. [10] ise, meme kanseri teşhisi için 699 örnek içeren veri setine KNN ve Karar Ağaçları algoritmalarını uygulamışlardır. Hastalık teşhisi üzerine çalışılan bir başka hastalık da Parkinson hastalığıdır. Yılancıoğlu [2], önerdiği bir Yapay Sinir Ağı (YSA) modeli ile Parkinson hastalığı için tanısal bir öngörü modeli elde etmiştir. Bang vd. [11], demans hastalığı teşhisi için dört aşamalı bir veri madenciliği modeli önermiştir. Kalp hastalıkları da dünyanın önde gelen ölüm nedenlerindedir. Shouman vd. [12] de veri madenciliği tekniklerini kalp hastalığı tedavi verilerine uygulayarak, hastalık teşhisi için bir model önermiştir. Mello vd. [13] ise, Rio de Janeiro şehrinde halk sağlığı sisteminde görülen ayaktan hastalar arasında akciğer tüberkülozu olan SNPT hastalığının teşhisi için hem Lojistik Regresyon (LR) hem de Karar Ağaçları tekniklerini kullanarak bir tahmin modeli geliştirmiştir. Kılıçarslan vd. [14] çalışmalarında, Parçacık Sürü Optimizasyonu (PSO) ve Temel Bileşenler Analizi (TBA) yöntemleriyle birlikte KNN ve DVM kullanılarak prostat kanseri veri kümesinde öznitelik boyutunu azaltarak prostat kanserinde etkin olan genleri belirlemiştir. Diğer çalışmalardan farklı olarak metin verisi üzerine çalışan Yolcular vd. [15], kulak burun boğaz bölümüne ait taburcu notlarından oluşan bir veri setini kullanarak metin madenciliği ve birliktelik analizi teknikleri ile birliktelik kurallarını ortaya çıkarmışlardır.

Bu çalışmada, insan sağlığını etkileyen kritik hastalıklardan biri olan tiroit hastalığı teşhisi üzerine odaklanılmıştır. Dash vd. [16] çalışmalarında, tiroit verilerinin sınıflandırılması için uygulanan çeşitli teknikleri derlemiştir. Şengül ve Türkoğlu [17], biyokimya test sonuçlarından hipertiroidi ve hipotiroidi teşhisinde, doktorlara kolaylık sağlayacak karar ağaçları temelli bir karar destek sistemi tasarlanmıştır. Önerilen sistemde 120

hasta verisi değerlendirilmiştir. Yeh [18], UCI veritabanından tiroit bezi veri kümesini çıkarmak için geliştirilmiş, Basitleştirilmiş Sürü Optimizasyonu (SSO) kullanılarak yeni bir kural tabanlı sınıflandırıcı tasarım yöntemi kullanmıştır. Hipotiroidi teşhisi için Kaya [19] da UCI veritabanını kullanmış ve buradan alınan tiroit veri setine, hızlı öğrenilebilir tek gizli katmanlı ileri beslemeli bir YSA modeli olan AÖM yöntemi uygulanmıştır. Ele alınan veri seti için AÖM'ün, diğer makine öğrenmesi yöntemlerine göre hem %96,76'lık sınıflandırma başarısı hem de hız bakımından önemli avantajlar sağladığı görülmüştür.

Deepika ve Kalaiselvi [20], 776'sı tiroit verileri ve 6771'i tiroit dışı veriler olmak üzere toplam 7547 örnek üzerinde çalışmıştır. SVM, Karar Ağaçları ve YSA'nı kullanarak, belirledikleri 29 öznelik ile yeni bir hastanın tiroit hastası olup olmadığını tahmin etmişlerdir. Bu üç yöntemden, en yüksek doğruluk oranını YSA vermiştir. Sajadia vd. [21] ise çalışmalarında İran'ın İmam Humeyni Kliniği ve Shahid Beheshti Hastanesi'ne başvuran hastalardan elde edilen verileri kullanmıştır. Veriler üç sınıfta normal, sublinik hipotiroidizm ve hipotiroidizm olmak üzere 305 denek içermektedir. Tiroit bozukluklarının teşhisi için önerdikleri bulanık kural tabanlı sınıflandırıcının doğruluk oranı %97 olarak elde edilmiştir.

Yazın taraması özetlenecek olursa; makine öğrenmesi teknikleri kullanılarak yapılan hastalık teşhisi hakkında birçok çalışma yer almaktadır. Özellikle insanlarda yaygın olarak görülen kalp, diyabet, tiroit ve göğüs hastalıkları alanında hastalık teşhisi için çeşitli yöntemler denenmiştir. Tiroit hastalıklarının günümüzde artan seyir izlemesi bu çalışmada kullanılan veri seti olarak hipotiroidi hastalığının seçilmesinde önemli bir etken olmuştur. Bu çalışmada, hipotiroidi teşhisi için makine öğrenmesi algoritmalarından Lojistik Regresyon, K En Yakın Komşu ve Destek Vektör Makinesi kullanılmıştır.

### 3. MATERYAL ve METOT (MATERIAL AND METHOD)

#### 3.1. Materyal (Material)

Bu çalışmada kullanılan ve UCI [3] veri tabanından alınan hipotiroidi hastalığına ait veri setinin karakteristiği Tablo 1'de verilmiştir.

Tablo 1. Hipotiroidi veri seti karakteristiği  
(Hypothyroid data set characteristic)

Örnek Sayısı	3163
Öznelik Sayısı	25
Sınıf Sayısı	2
Öznelik Özellikleri	Kategorik, Nümerik
Eksik Veri	Evet

Veri seti 3163 örneğe ait veriden oluşmaktadır. Bu örneklerin 151 tanesi hipotiroidi tanısı konmuş hastaya aittir. Geriye kalan 3012 örnek ise sağlıklıdır. Her örnek 25 öznelik içermektedir. Bu öznelikler numune alınan

kişiye ait yaş, cinsiyet; tiroksin, antitiroit ilaç, tiroit, hipotiroidi, hipertiroidi, hamilelik, hastalık, tümör, lityum, guatr gibi durumların olup olmadığı; kan tahlilinde hipotiroidi hastalığı teşhisi için yapılan ölçümlerden (TSH, T3, TT4, T4U, FTI ve TBG) oluşmaktadır. Bu veri seti kullanılarak, tahlil sonuçları ile yeni gelen bir hastanın hipotiroidi olup olmadığı belirlenecektir. Nümerik değer alan özneliklere ait tanımlayıcı istatistikler Tablo 2'de, kategorik değer alan özneliklere ait tanımlayıcı istatistikler de Tablo 3'te özetlenmiştir.

Tablo 2. Nümerik özneliklere ait tanımlayıcı istatistikler

(Descriptive statistics of numerical attributes)						
Öznelik	Birim	Ortalama	Standart Sapma	En Küçük	En Büyük	Eksik veri (%)
Age	Yıl	51.1542	19.2944	1	98	14.1005
TSH	$\mu\text{IU} / \text{L}$	5.9232	23.8995	0	530	14.7961
T3	$\text{pg}/\text{mL}$	1.9397	0.9968	0	10.2	21.9728
TT4	$\text{nmol}/\text{L}$	108.8500	45.4854	2	450	7.8723
T4U	$\text{ng}/\text{dl}$	0.9782	0.2266	0	2.21	7.8407
FTI	$\mu\text{g}/\text{dL}$	115.3978	60.2396	0	881	7.8090
TBG	$\text{mg} / \text{dL}$	31.2831	19.2247	0	122	91.7800

Tablo 3. Kategorik özneliklere ait tanımlayıcı istatistikler

(Descriptive statistics for categorical attributes)				
Öznelik	Değer	Frekans	Yüzde	Mod
Sex	Kadın	2182	68.9852	Kadın
	Erkek	908	28.7069	
	Eksik Veri	73	2.3079	
on_thyroxine	TRUE	461	14.5748	FALSE
	FALSE	2702	85.4252	
	Eksik Veri	0	0	
query_on_thyroxine	TRUE	55	1.7389	FALSE
	FALSE	3108	98.2611	
	Eksik Veri	0	0	
on_antithyroid_medication	TRUE	42	1.3279	FALSE
	FALSE	3121	98.6721	
	Eksik Veri	0	0	
thyroid_surgery	TRUE	104	3.2880	FALSE
	FALSE	3059	96.7120	
	Eksik Veri	0	0	
query_hypothyroid	TRUE	241	7.6193	FALSE
	FALSE	2922	92.3807	
	Eksik Veri	0	0	
query_hyperthyroid	TRUE	243	7.6826	FALSE
	FALSE	2920	92.3174	
	Eksik Veri	0	0	
Pregnant	TRUE	63	1.9918	FALSE
	FALSE	3100	98.0082	
	Eksik Veri	0	0	

Tablo 3. Kategorik özniteliklere ait tanımlayıcı istatistikler devamı  
(Descriptive statistics for categorical attributes continued)

Öznitelik	Değer	Frekans	Yüzde	Mod
Sick	TRUE	99	3.1299	FALSE
	FALSE	3064	96.8701	
	Eksik Veri	0	0	
Tumor	TRUE	40	1.2646	FALSE
	FALSE	3123	98.7354	
	Eksik Veri	0	0	
Lithium	TRUE	2	0.0632	FALSE
	FALSE	3161	99.9368	
	Eksik Veri	0	0	
Goitre	TRUE	99	3.1299	FALSE
	FALSE	3064	96.8701	
	Eksik Veri	0	0	
TSH_measured	TRUE	468	14.7961	FALSE
	FALSE	2695	85.2039	
	Eksik Veri	0	0	
T3_measured	Yes	2468	78.2747	Yes
	No	685	21.7253	
	Eksik Veri	0	0	
TT4_measured	Yes	2914	92.1277	Yes
	No	249	7.8723	
	Eksik Veri	0	0	
T4U_measured	Yes	2915	92.1593	Yes
	No	248	7.8407	
	Eksik Veri	0	0	
FTI_measured	Yes	2916	92.1910	Yes
	No	247	7.8090	
	Eksik Veri	0	0	
TBG_measured	Yes	260	8.2200	No
	No	2903	91.7800	
	Eksik Veri	0	0	

### 3.2. Sınıflandırma Teknikleri (Classification Techniques)

Bu çalışmanın ana konusu olan veri madenciliği ile ilgili yazında birçok tanım yapılmıştır. Örneğin, Fayyad [22] veri madenciliğini bir veri tabanında saklanan verilerden örtük, daha önce bilinmeyen ve potansiyel olarak yararlı bilgilerin bir şekilde çıkarılması süreci olarak tanımlarken; Giudici [23] de eldeki büyük miktarlarda veriden net ve faydalı sonuçlar elde etmek amacıyla başlangıçta bilinmeyen düzenlilikleri veya ilişkileri keşfetmek için seçme, keşfetme ve modelleme süreci olarak tanımlamıştır [24].

Denetimli ve denetimsiz öğrenme, veri madenciliğinde kullanılan iki temel stratejidir. Denetimli öğrenmede, model parametrelerini öğrenmek için bir eğitim seti kullanılırken, denetimsiz öğrenmede herhangi bir eğitim veri seti kullanılmaz. Bu stratejiler temelinde geliştirilen

veri madenciliği teknikleri, modelleme hedefine bağlı olarak çeşitli uygulama alanlarında kullanılır. Sınıflandırma ve tahmin de bu alanda en sık karşımıza çıkan uygulama alanlarıdır.

Sınıflandırma problemleri genel olarak, yeni bir verinin hangi sınıfa ait olacağını bulmakla ilgilidir. Örneğin bir grup hastanın farklı test sonuçlarına (öznitelikler) göre, yeni bir hastanın ilgili test sonuçlarına bakarak bu kişinin hipotiroidi olup olmadığı (iki sınıf) belirlenebilir ya da yine bir grup hastanın test sonuçlarına bakarak, yeni bir hastanın hipotiroidi, hipertiroidi ya da guatr hastası olup olmadığı (çoklu sınıf) belirlenebilir.

Bu çalışmada, hipotiroidi hastalığının yeni hasta değerleriyle doğru olarak teşhis edilebilmesi adına, sınıflandırıcının bir veri seti yardımıyla hasta ve sağlıklı örneklere ait değerleri öğrenmesi gerektiğinden denetimli öğrenme tekniklerinden sınıflandırma kullanılmıştır. Ayrıca hastalık teşhisine yönelik çalışmalarda yüksek başarı oranları sağlayan sınıflandırma algoritmalarından Lojistik Regresyon (LR), K En Yakın Komşu (KNN) ve Destek Vektör Makinesi (DVM) ile yeni bir hastanın hipotiroidi olup olmadığının belirlenmesine çalışılmıştır. Lojistik Regresyon, bağımlı değişkenin iki veya daha fazla sınıfa sahip olduğu, bağımsız değişkenlerin ise sürekli veya kategorik bir yapıyı içerdiği durumlarda bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi araştıran bir yöntemdir [25].  $b$  kesişim noktası,  $w$  hiperdüzlemin normal vektörü ve  $z$   $x$ 'in hiperdüzlemdaki değeri olmak üzere;  $z = wx - b$  ve  $z$ 'nin lojistik fonksiyondaki değeri  $\phi(z) = \frac{1}{1+e^{-z}}$  şeklinde tanımlıdır. Bunlara bağlı olarak, Eşitlik (1)'deki log-benzerlik değerini en büyüleyecek  $w$  ve  $b$  parametresi araştırılır.

$$l(w) = \log \log L(w) = \sum_{i=1}^n \left[ y^i \log \log (\phi(z^i)) + (1 - y^i) \log \log (1 - \phi(z^i)) \right] \quad (1)$$

K En Yakın Komşu (KNN), denetimli öğrenme metodlarından en çok kullanılan sınıflandırma yöntemlerinden biridir. Parametrik olmayan tembel bir öğrenme algoritmasıdır. Sınıfı tahmin edilecek her bir örnek için, veri setindeki tüm örnekler arasında en yakın komşuluğu arar. Bu nedenle büyük miktarda bellek alanına ihtiyaç duyar. Veri seti büyüklüğüne bağlı olarak maliyet ve işlem yükünün artması dezavantajlarındandır. KNN algoritmasını adımları şu şekildedir:

1. Bir  $k$  komşuluk sayısı ve uzaklık metriği belirlenir.
2. Sınıflandırmak istenilen örneğin en yakın  $k$  komşuları bulunur.
3.  $k$  en yakın komşunun sınıfları arasında en yüksek frekansa sahip olan sınıfa atama yapılır.

KNN algoritmasında sürekli değişkenler için yakınlık hesabında kullanılan uzaklık ölçüleri Öklid, Manhattan ve Minkowski'dir. Öklid ve Manhattan uzaklık ölçüleri, Eşitlik (2)'de verilen Minkowski uzaklık ölçüsünün özel birer durumudur. Eşitlik (2)'deki ifade  $p=1$  için Manhattan uzaklık ölçüsüne karşılık gelirken,  $p=2$  için Öklid uzaklık ölçüsüne karşılık gelmektedir.

$$d(x, y) = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \quad (2)$$

Sınıflandırma problemlerinde yaygın olarak kullanılan bir diğer makine öğrenmesi algoritması da Destek Vektör Makineleridir (DVM). DVM istatistiksel öğrenme teorisine dayanan, dağılım hakkında önbilgisiz olarak çalışan bir denetimli öğrenme algoritmasıdır. Yüksek doğruluk oranına sahip olması, aşırı uygunluk (overfitting) sorununun olmaması avantajları arasında gösterilir. DVM doğrusal ve doğrusal olmayan olarak ikiye ayrılır. DVM iki amaçlı bir yapıya sahiptir. Bir taraftan sınıflandırma hatasını en küçüklemeye çalışırken, diğer taraftan da aşırı öğrenmeyi engellemek üzere marjı en büyükmeye imkan sağlar. Bu iki birbiri ile çelişen amaç,  $C$  parametresi ile kontrol edilir.  $C$  büyüdükçe sınıflandırma hatasına verilen önem artırılmış olur.  $x^i$  veri noktasını,  $y^i$  sınıf bilgisini,  $\xi^i$   $i$ . noktanın hiperdüzlemin yanlış tarafında kalması durumunda pozitif değer alacak aylak değişkeni,  $b$  kesişim noktasını ve  $w$  hiperdüzlemin normal vektörünü göstermek üzere, Eşitlik (3)'te ortaya çıkan matematiksel programlama modeli verilmiştir.

$$\text{minimize } \frac{1}{2} \|w\|^2 + C(\sum_i \xi^i) \quad (3)$$

s.t.

$$w_0 + w^T x^i - b \geq 1 - \xi^i, \text{ if } y^i = 1$$

$$w_0 + w^T x^i - b \leq -1 + \xi^i, \text{ if } y^i = -1$$

#### 4. HİPOTİROİDİ TEŞHİSİ (DIAGNOSIS OF HYPOTYROIDISM)

Hipotiroidi şüphesiyle gelen yeni bir hastanın çeşitli öznelikleri kullanılarak hipotiroidi olup olmadığının belirlendiği bu çalışmanın aşamaları izleyen alt bölümlerde detaylı olarak verilmiştir.

##### 4.1. Veri Önileme (Data Preprocessing)

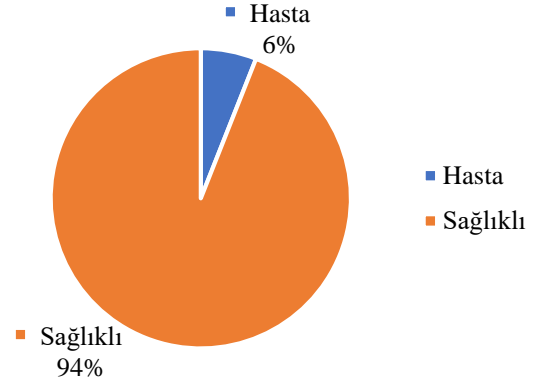
UCI'da yer alan hipotiroidi hastalığına ait Avustralya'dan alınan 3163 örnekten oluşan veri seti incelendiğinde, bazı örneklere ait öznelik verilerinin veri setinde yer almadığı görülmüştür. Bu durum veri madenciliğinde eksik veri problemi olarak tanımlanır. Eksik veri problemi iki şekilde çözülebilir; eksik veriler uygun yöntemlerle tamamlanabilir ya da eksik veriler veri setinden silinebilir. TBG değerinin ölçülüp kaydedildiği sadece 2 örnek bulunmaktadır. Bu nedenle "TBG\_measured" ve "TBG" öznelikleri veri setinden tamamen çıkarılmıştır. Bu işlemde sonra, geriye kalan öznelikler incelendiğinde eksik veri oranlarının öznelik bazında %30 'u geçmediği görülmüş ve eksik verilere sahip örnekler veri setinden tamamen çıkarılmıştır.

Eksik verilerin silinmesinden sonra, TSH, T3, TT4, T4U ve FTI değerlerinin ölçülüp ölçülmediğini bilgisini taşıyan "TSH\_measured", "T3\_measured", "TT4\_measured", "T4U\_measured" ve "FTI\_measured" öznelikleri sadece "evet" değerine sahip oldukları için bu öznelikler de veri setinden çıkarılmıştır.

Veri setindeki kategorik tipteki öznelikler, One-Hot Encoding yöntemi ile sayısal değerlere çevrilmiştir. Veriyi modele hazırlamak için Min-Max Normalizasyon tekniği

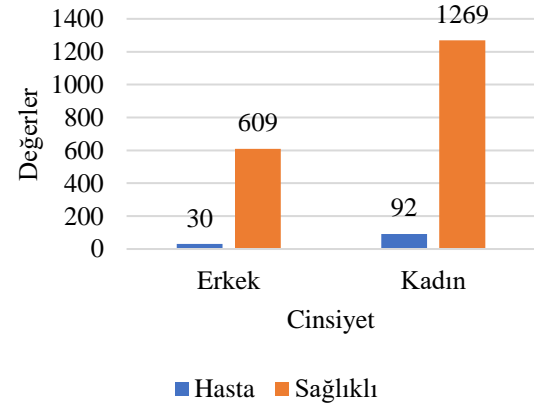
ile, en küçük değer 0 ve en büyük değer 1 olacak şekilde veri normalize edilmiştir.

Veri önileme sonucunda veri setinde kalan 2000 örneğin dağılımı Şekil 1'de gösterilmiştir. Bu örneklerin %6'sı hipotiroidi hastası ve %94 'ü sağlıklı olarak dağılmıştır. Şekil 1'den de açıkça anlaşılacağı gibi veri seti dengesiz bir dağılıma sahiptir.



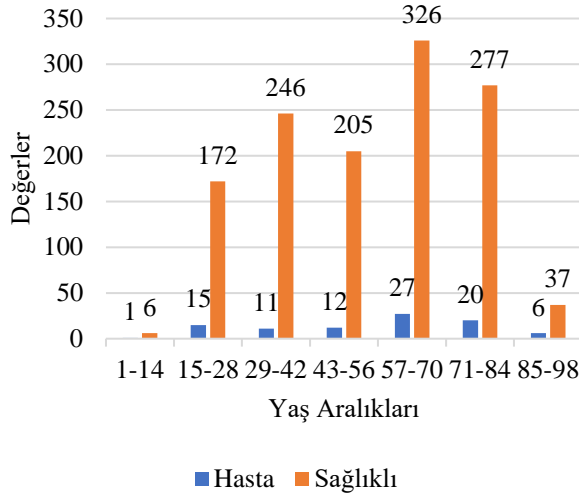
Şekil 1. Hasta/sağlıklı değerlere sahip verilerin dağılımı (Distribution of patient / healthy values data)

Şekil 2'de hasta/sağlıklı verilerin cinsiyet özneliğine göre dağılımı yer almaktadır. Kadın cinsiyetine sahip hasta ve sağlıklı verilerin erkek cinsiyetine göre oldukça fazla olduğu görülmektedir.

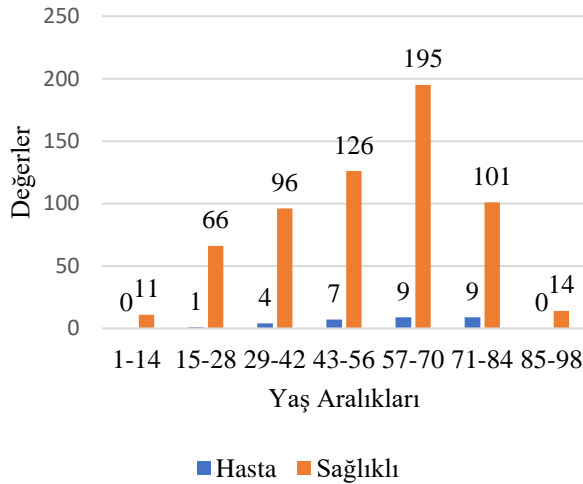


Şekil 2. Cinsiyet özneliğine göre hasta/sağlıklı verilerin dağılımı (Distribution of patient / healthy data by gender attribute)

Veri setinde yer alan cinsiyete göre hasta/sağlıklı kişilerin yaş dağılımları kadınlar için Şekil 3'te, erkekler için ise Şekil 4'te gösterilmektedir. 1 ile 98 yaş aralığında kişilere ait veri bulunmaktadır. Ele alınan veriye göre, kadınlarda en fazla hipotiroidi hastalığı 57-70, erkeklerde ise 57-70 ve 71-84 yaş gruplarında görülmektedir.



Şekil 3. Belirli yaş aralıklarına göre kadınların hasta/sağlıklı dağılımı  
(Patient / healthy distribution of women according to specific age ranges)



Şekil 4. Belirli yaş aralıklarına göre erkeklerin hasta/sağlıklı dağılımı  
(Patient / healthy distribution of men according to specific age ranges)

Veri setinde hastalıklı değerlere sahip veriler %6, sağlıklı değerlere sahip veriler ise %94 olarak dağıldığı için model, sağlıklı veriyi aşırı olarak öğrenmeye yatkın durumdadır. Bu durumdan kaçınmak için eğitim üzere kullanılan veri setine az örnekleme (undersampling) ve fazla örnekleme (oversampling) yöntemleri kullanılmıştır. Az örnekleme yöntemleri kullanılarak veri setindeki hipotiroidi olmayan örneklerin sayısı azaltılmış ve böylece veri setinin dengeli dağılımı sağlanmıştır. Fazla örnekleme yöntemleri kullanılarak da veri setindeki hipotiroidi olan örnekler sayısı artırılmış ve böylece veri setinin dengeli dağılımı sağlanmıştır. Python imblearn kütüphanesinin 0.5.0 versiyonu [26], kullanılarak uygulanan az ve fazla örnekleme yöntemleri Tablo 4'te verilmiştir.

Tablo 4. Az ve fazla örnekleme yöntemleri  
(Undersampling and oversampling methods)

Az Örnekleme Yöntemleri	Fazla Örnekleme Yöntemleri
CondensedNearestNeighbour	SMOTE
EditedNearestNeighbours	ADASYN
RepeatedEditedNearestNeighbours	KMeansSMOTE
AllKNN	RandomOverSampler
InstanceHardnessThreshold	SVMSMOTE
NearMiss	
NeighbourhoodCleaningRule	
OneSidedSelection	
RandomUnderSampler	
TomekLinks	
ClusterCentroids	

#### 4.2. Değerlendirme Kriterleri (Evaluation criteria)

Sınıflandırma yöntemleri ile elde edilen sonuçların başarısını ölçmek için kullanılan değerler Tablo 5'teki hata matrisi üzerinde gösterilmiştir. Doğru Pozitif (DP) ve Doğru Negatif (DN) modelin doğru olarak tahminlediği, Yanlış Pozitif (YP) ve Yanlış Negatif (YN) ise modelin yanlış olarak tahminlediği alanlardır. YP (Tip 1 Hata) sağlıklı olan örneğin hipotiroidi hastası olarak tahmin edilmesi iken, YN (Tip 2 Hata) hipotiroidi hastası olan örneğin sağlıklı olarak tahmin edilmesidir.

Tablo 5. Hata matrisi  
(Confusion matrix)

		Gerçek Sınıf	
		Var	Yok
Tahmin edilen Sınıf	Var	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Yok	Yanlış Negatif (YN)	Doğru Negatif (DN)

Bu çalışmada kullanılan sınıflandırıcıların başarısını ölçmek için kullanılan değerlendirme kriterleri Tablo 6'da görülmektedir. Doğruluk oranı modelde doğru tahmin edilen alanların toplam veri kümesine oranı ile hesaplanmaktadır. Kesinlik pozitif olarak tahminlediğimiz değerlerin gerçekten kaç adedinin pozitif olduğunu göstermektedir. Duyarlılık ise pozitif olarak tahmin edilmesi gereken işlemlerin ne kadarını pozitif olarak tahmin edildiğini gösteren bir metriktir. F-Skor değeri kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır ve [0,1] aralığında değer almaktadır. Tüm sınıfların doğru tahmin edildiği ideal durumda 1 değerini alır. Matthews korelasyon katsayısı tahmin edilen sınıflar ile gerçekte olan sınıflar arasındaki ilişkiyi gösterir ve [-1,1] aralığında değer alır. Matthews korelasyon katsayısının 1 olması sınıflandırıcının tüm tahminlerin doğru olduğunu, -1 olması ise tüm tahminlerin yanlış olduğunu gösterir. Alıcı işletim karakteristiği bir olasılık eğrisidir, bu eğri altında

kalan alan [0,1] aralığında değer alır ve modelin sınıfları ne kadar başarılı ayırt edebildiğini açıklar. Eğri altında kalan alan arttıkça model, hastaları hasta ve sağlıklıları sağlıklı olarak tahmin etmede daha iyidir. Ele alınan veri setindeki dengesiz yapı göz önüne alındığında, sınıflandırıcıların performansını değerlendirmek için doğruluk oranından daha fazla bilgi veren F-Skor, eğri altında kalan alan ve Matthews korelasyon katsayısı değerlerinin değerlendirme kriteri olarak kullanılmasına karar verilmiştir.

#### 4.3. Sayısal Sonuçlar (Numerical Results)

Hipotiroidi veri setinde belirtildiği üzere, sağlıklı ve hastalıklı örnekler için veriler dengesiz bir dağılım göstermektedir. Tablo 4'te verilen farklı örnekleme teknikleri ile dağılımındaki dengesizliği ortadan kaldırarak sınıflandırıcı performanslarının artırılması hedeflenmiştir.

Tablo 6. Değerlendirme kriterleri  
(Evaluation criteria)

Kriter	Açıklama
Doğruluk oranı	$\frac{DP}{DP + DN + YP + YN}$
Duyarlılık	$\frac{DP}{DP + YN}$
Kesinlik	$\frac{DP}{DP + YP}$
F-Skor	$2 * \frac{Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}$
Matthews Korelasyon Katsayısı	$\frac{DP * DN - YP * YN}{\sqrt{(DP + YP) * (DP + YN) * (DN + YP) * (DN + YN)}}$
Eğri Altında Kalan Alan	

Sınıflandırıcı hiperparametrelerinin belirlenmesi için 10 defa çapraz geçirme ile grid search yapılmıştır. Sınıflandırıcılara ait modellerin geliştirilmesinde Python Scikit-learn kütüphanesinin 0.22.0 versiyonu [27] kullanılmıştır.

Örnekleme tekniklerinin sınıflandırıcı performansı üzerindeki etkisini incelemek adına; KNN, LR ve DVM sınıflandırıcıların her biri için eğitim veri setine fazla örnekleme (Over Sampling) ve az örnekleme (Under Sampling) yöntemleri uygulanmış ve bu yöntemlerin uygulanmadığındaki durum ile elde edilen sonuçlar değerlendirme kriterleri bazında Lojistik Regresyon sınıflandırıcısı için Tablo 7'de, K En Yakın Komşu sınıflandırıcısı için Tablo 8'de ve Destek Vektör Makinesi sınıflandırıcısı için de Tablo 9'da raporlanmıştır. Kullanılan değerlendirme kriterleri; Doğruluk Oranı, F-Skor, Eğri Altındaki Alan ve Matthews Korelasyon Katsayısı şeklindedir. Sınıflandırıcı performanslarını genelleştirmek ve sınıflandırıcılar arasında doğru bir kıyaslama yapmak amacıyla 10 defa çapraz doğrulama kullanılmıştır. İzleyen tablolarda raporlanan her bir değer 10 defa çapraz doğrulama sonucunda elde edilen ortalamalardır. Sınıflandırıcıların örnekleme yöntemlerine göre sonuçlarının en iyi değerleriyle Tablo 10'daki gibi bir kıyaslama yapılmıştır.

Tablo 7'de verilen Lojistik Regresyon sınıflandırıcısı için elde edilen sonuçlara bakıldığında, veri setine uygulanan fazla örnekleme yöntemi, az örnekleme yöntemine ve mevcut duruma göre daha yüksek sonuçlar vermiştir. Lojistik Regresyon sınıflandırıcısı ile değerlendirme kriterlerince, %81.78'in üzerinde sonuçlar elde edilmiştir.

Tablo 8'de K En Yakın Komşu sınıflandırıcısında da Lojistik Regresyon'da olduğu gibi, fazla örnekleme yönteminde diğer durumlara göre başarılı sonuçlar elde edilmiştir. Fazla örnekleme yönteminin K En Yakın Komşu sınıflandırıcısı üzerindeki başarılı etkisiyle değerlendirme kriterlerince %74.08'in üzerinde sonuçlara ulaşılmıştır.

Tablo 9'da Destek Vektör Makinesi sınıflandırıcısında, diğer sınıflandırıcılara göre doğruluk oranı dışında diğer değerlendirme kriterleri için hiçbir yöntemin uygulanmadığı durumda daha iyi sonuçlar elde edilmiştir.

Tablo 7. Lojistik Regresyon sınıflandırıcısının örnekleme yöntemlerine göre sonuçları  
(Results of Logistic Regression classifier according to sampling methods)

Sınıflandırıcı	Örnekleme Tipi	Örnekleme Yöntemi	Doğruluk	F-Skor	Eğri Altında Kalan Alan	Matthews Korelasyon Katsayısı
LR	Fazla Örnekleme	SMOTE	<b>0.978</b>	<b>0.826</b>	0.9236	<b>0.8175</b>
		ADASYN	0.9728	0.796	0.9229	0.7878
		KMeansSMOTE	0.9752	0.81	0.9235	0.8025
		RandomOverSampler	0.9696	0.784	0.9269	0.7782
		SVMSMOTE	0.9704	0.79	<b>0.9319</b>	0.7842
	Az Örnekleme	ClusterCentroids	0.853	0.682	0.8763	0.6725
		CondensedNearestNeighbour	0.8685	0.69	0.8818	0.6811
		EditedNearestNeighbours	0.8824	0.709	0.8875	0.6999
		RepeatedEditedNearestNeighbours	0.8932	0.723	0.892	0.7142
		AllKNN	0.9019	0.734	0.8955	0.7255
		InstanceHardnessThreshold	0.9075	0.737	0.9005	0.7287
		NearMiss	0.9128	0.739	0.9005	0.7313
		NeighbourhoodCleaningRule	0.918	0.747	0.9026	0.7392
		OneSidedSelection	0.9222	0.751	0.9008	0.7431
		RandomUnderSampler	0.9224	0.742	0.9035	0.7359
		TomekLinks	0.9258	0.746	0.9016	0.7394
		-	Mevcut Durum	0.9289	0.748	0.8989

Tablo 8. K En Yakın Komşu sınıflandırıcısının örnekleme yöntemlerine göre sonuçları  
(Results of K Nearest Neighbour classifier according to sampling methods)

Sınıflandırıcı	Örnekleme Tipi	Örnekleme Yöntemi	Doğruluk	F-Skor	Eğri Altında Kalan Alan	Matthews Korelasyon Katsayısı
KNN	Fazla Örnekleme	SMOTE	0.9311	<b>0.748</b>	<b>0.8962</b>	<b>0.7408</b>
		ADASYN	0.933	0.746	0.8931	0.7392
		KMeansSMOTE	0.935	0.745	0.8897	0.7387
		RandomOverSampler	0.9361	0.743	0.8884	0.7358
		SVMSMOTE	<b>0.9373</b>	0.741	0.8868	0.7335



Tablo 8. K En Yakın Komşu sınıflandırıcısının örnekleme yöntemlerine göre sonuçları devamı  
(Results of K Nearest Neighbour classifier according to sampling methods continued)

Sınıflandırıcı	Örnekleme Tipi	Örnekleme Yöntemi	Doğruluk	F-Skor	Eğri Altında Kalan Alan	Matthews Korelasyon Katsayısı
KNN	Az Örnekleme	ClusterCentroids	0.9002	0.714	0.8704	0.7027
		CondensedNearestNeighbour	0.8962	0.699	0.8684	0.6888
		EditedNearestNeighbours	0.8992	0.7	0.8659	0.6903
		RepeatedEditedNearestNeighbours	0.902	0.701	0.8636	0.6917
		AllKNN	0.9046	0.702	0.8614	0.693
		InstanceHardnessThreshold	0.907	0.703	0.8605	0.6946
		NearMiss	0.9062	0.694	0.8596	0.6863
		NeighbourhoodCleaningRule	0.9083	0.695	0.8572	0.6869
		OneSidedSelection	0.9103	0.695	0.8551	0.6876
		RandomUnderSampler	0.9095	0.688	0.8562	0.682
	TomekLinks	0.9113	0.688	0.8539	0.6824	
	-	Mevcut Durum	0.913	0.688	0.8517	0.6828

Tablo 9. Destek Vektör Makinesi sınıflandırıcısının örnekleme yöntemlerine göre sonuçları  
(Results of Support Vector Machine classifier according to sampling methods)

Sınıflandırıcı	Örnekleme Tipi	Örnekleme Yöntemi	Doğruluk	F-Skor	Eğri Altında Kalan Alan	Matthews Korelasyon Katsayısı
DVM	Fazla Örnekleme	SMOTE	0.9149	0.6924	0.854	0.6871
		ADASYN	0.9164	0.6948	0.8562	0.6895
		KMeansSMOTE	0.9181	0.6987	0.8576	0.6933
		RandomOverSampler	0.9191	0.6994	0.8599	0.6941
		Over_SVMSMOTE	<b>0.9203</b>	0.701	0.8621	0.6958
	Az Örnekleme	ClusterCentroids	0.9024	0.6868	0.8548	0.6808
		CondensedNearestNeighbour	0.904	0.6892	0.8567	0.6832
		EditedNearestNeighbours	0.9058	0.6925	0.858	0.6865
		RepeatedEditedNearestNeighbours	0.9075	0.6957	0.8593	0.6896

Tablo 9. Destek Vektör Makinesi sınıflandırıcısının örnekleme yöntemlerine göre sonuçları devamı  
(Results of Support Vector Machine classifier according to sampling methods continued)

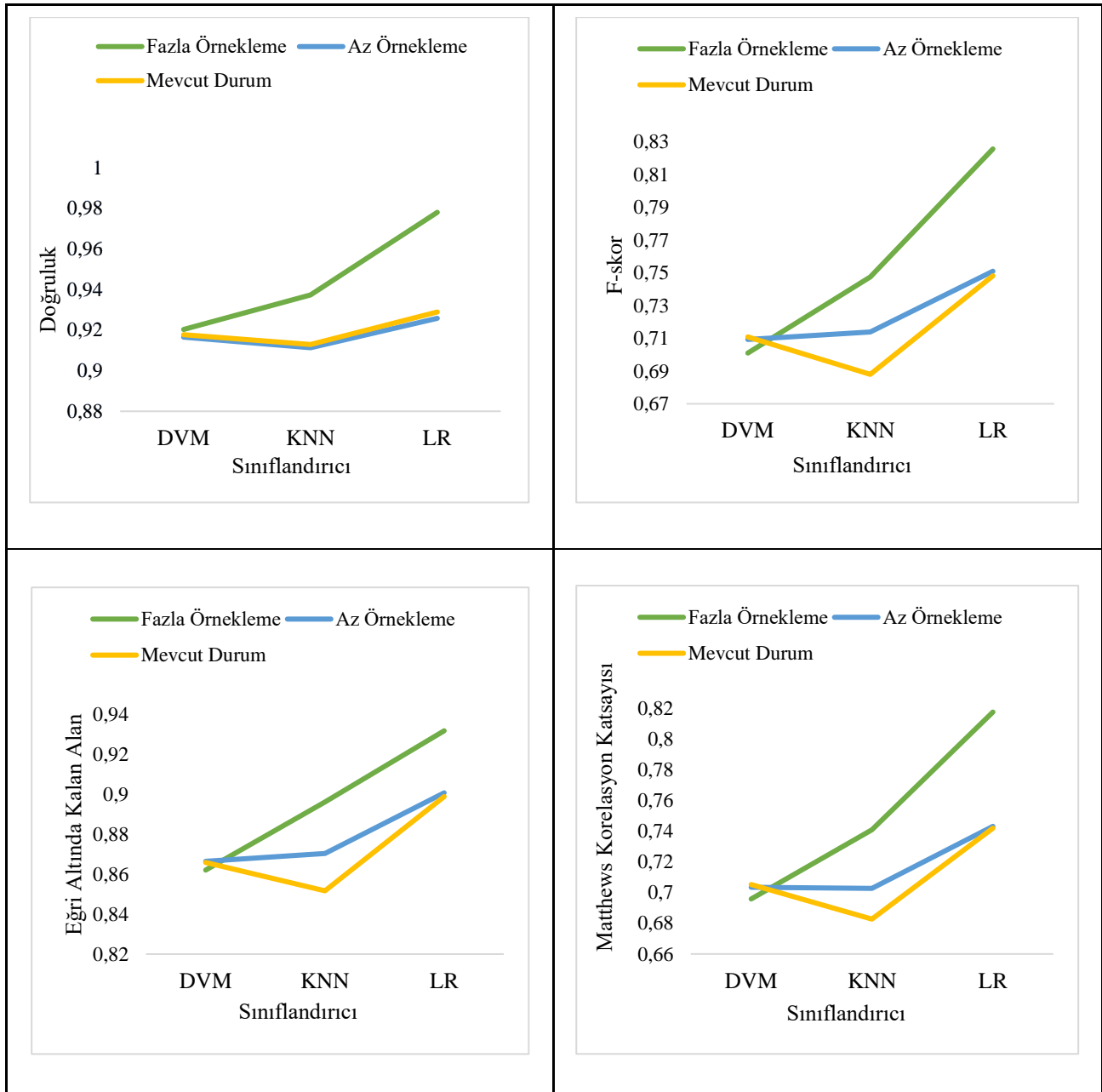
Sınıflandırıcı	Örnekleme Tipi	Örnekleme Yöntemi	Doğruluk	F-Skor	Eğri Altında Kalan Alan	Matthews Korelasyon Katsayısı
DVM	Az Örnekleme	AllKNN	0.9091	0.6989	0.8607	0.6927
		InstanceHardnessThreshold	0.9106	0.7013	0.8626	0.6951
		NearMiss	0.9121	0.7043	0.8639	0.6981
		NeighbourhoodCleaningRule	0.9135	0.7072	0.865	0.701
		OneSidedSelection	0.9149	0.7089	0.8648	0.7026
		RandomUnderSampler	0.9153	0.7076	<b>0.8665</b>	0.7018
		TomekLinks	0.9165	0.7093	0.8662	0.7035
	-	Mevcut Durum	0.9177	<b>0.7109</b>	0.8659	<b>0.7052</b>

Tablo 10'da verilen sınıflandırıcı başarılarına bakıldığında, LR ve fazla örnekleme (oversampling) tekniğinin beraber kullanıldığı durum tüm değerlendirme kriterleri için diğer sınıflandırıcılardan daha iyi sonuç vermiştir. LR ile elde edilen en iyi sonuçlar; doğruluk oranı için %97.8, F-Skor değeri

için %82.26, eğri altında kalan alan için %93.2 ve Matthews korelasyon katsayısı için %81.8'dir. Her bir değerlendirme kriteri için sınıflandırıcılar ve örnekleme yöntemlerine göre elde edilen en iyi değerler Şekil 5'te görselleştirilmiştir.

Tablo 10. Farklı sınıflandırıcı ve yöntemlerin en iyi sonuçları  
(Highest results of different classifiers and methods)

Sınıflandırıcı	Yöntem	En İyi Doğruluk Değeri	En İyi Fskor Değeri	Eğri Altında Kalan Alan	Matthews Korelasyon Katsayısı
Lojistik Regresyon	Fazla Örnekleme	0.978	0.8257	0.9319	0.8175
	Az Örnekleme	0.9258	0.751	0.9008	0.7431
	Mevcut Durum	0.9289	0.7482	0.8989	0.7419
K En Yakın Komşu	Fazla Örnekleme	0.9373	0.7475	0.8962	0.7408
	Az Örnekleme	0.9113	0.7139	0.8704	0.7027
	Mevcut Durum	0.9129	0.688	0.8517	0.6827
Destek Vektör Makinesi	Fazla Örnekleme	0.9203	0.701	0.8621	0.6958
	Az Örnekleme	0.9165	0.7093	0.8665	0.7035
	Mevcut Durum	0.9177	0.7109	0.8659	0.7052



Şekil 5. Değerlendirme kriterleri için sınıflandırıcılar ve örnekleme yöntemlerine göre en iyi değerler  
(Highest values according to classifiers and sampling methods for evaluation criteria)

Grafiklerden de anlaşıldığı üzere, KNN ve LR sınıflandırıcılarında veri setindeki dengesiz dağılımın fazla örnekleme yöntemiyle dengeli hale getirilmesi doğruluk oranı, F-Skor, eğri altında kalan alan ve Matthews korelasyon katsayısı değerlerinde ciddi bir artış sağlamaktadır. Sınıflandırıcılar bazında incelendiğinde, LR ve KNN için uygulanan fazla örnekleme yöntemlerinin, az örnekleme yöntemlerine ve hiçbir yöntem uygulanmadığı duruma göre oldukça başarılı olduğunu söylenebilir. DVM sınıflandırıcısında ise, doğruluk oranı hariç hiçbir örnekleme yönteminin uygulanmadığı durum fazla örnekleme ve az örnekleme yöntemlerine göre daha iyi sonuçlar vermiştir. DVM sınıflandırıcısında karşılaşılan bu durum, az örnekleme ve fazla örnekleme yöntemlerinin ayırma yüzeyini belirleyen

noktalar (destek vektörü) üzerindeki etkisi ile açıklanabilir [28].

Hastalık teşhisi için kullanılan diğer yöntemlerin literatürde sunulan doğruluk oranları ile çalışma kapsamında elde edilen en iyi doğruluk oranlarının karşılaştırılması Tablo 11'de verilmiştir. Çalışma kapsamında önerilen LR, KNN ve DVM sınıflandırıcılarına ait elde edilen doğruluk oranları, LVQ, geri yayımlı MLP, C4.5-1, MLP, AIRS, PNN ve RFRS yöntemlerinin doğruluk oranlarına göre daha yüksektir. C4.5, EBFN, LDA, KNN ve ANFIS yöntemleri ile de rekabet edebilir düzeyde başarı oranları elde edilmiştir.

Tablo 11. Çalışmada elde edilen sonuçların literatür ile karşılaştırılması  
(Comparison of obtained results with the literature)

Yöntem	k defa çapraz doğrulama	Doğruluk (%)
LVQ [31]	Test verisi	81.86
Geri yayımlı MLP [32]	3 defa çapraz doğrulama	86.33
C4.5-1 [33]	Test verisi	93.26
MLP [33]	Test verisi	96.24
AIRS [34]	10 defa çapraz doğrulama	81
PNN [35]	3 defa çapraz doğrulama	92.96
RFRS [36]	Test verisi	92.59
ANN [37]	Test verisi	97.8
C4.5, MLP, EBFN [38]	10 defa çapraz doğrulama	98.15
LDA-kNN-ANFIS [30]	10 defa çapraz doğrulama	98.5
Önerilen LR	10 defa çapraz doğrulama	97.8
Önerilen KNN	10 defa çapraz doğrulama	93.73
Önerilen DVM	10 defa çapraz doğrulama	92.03

## 5. SONUÇ VE ÖNERİLER (CONCLUSIONS AND RECOMMENDATIONS)

Veri madenciliğinin tıpta kullanılmasıyla beraber insan hayatının kalitesini artırmak için sürekli çalışmalar yapılmaktadır. Geline son noktada hastalık tanısında kullanılan teşhis sistemlerinin her geçen gün daha da önem kazandığını görmekteyiz. Veri madenciliğinin var olan bilgiyi yorumlama ve karar verme yeteneğiyle literatürde birçok hastalık için çalışmalar yapılmış ve başarılı sonuçlar elde edilmiştir. Elde edilen başarılı sonuçlar daha da iyileştirilmekte ve farklı yöntemlerle bu başarılar günden güne gelişme göstermektedir.

Hastanelerde ve çeşitli sağlık hizmetlerinde tıbbi işlemler/aşamalar arasındaki işlem ya da geri dönüş süresi, hizmet kalitesi, değişim yönetimi, maliyetlerin azaltılması

ve stratejik kararlar üzerinde de etkisi olan en önemli performans ölçütlerinden biri olarak kabul edilmektedir [29].

Hipotiroidi, doku düzeyinde tiroit hormonu yetersizliği veya nadiren etkisizliği sonucu ortaya çıkan, metabolik yavaşlama ile giden bir hastalıktır [13]. İnsan yaşamını olumsuz etkileyen bu hastalığın teşhis ve tedavi dönemlerinde hastayı daha az yormak ve hastayı yıpratmamak adına hastanın fazladan uygulanacak testlere maruz kalmaması ciddi bir şekilde önemlidir. Bu sebeple, yapılan bu çalışmada doktorlara yardımcı bir sistem tasarlanarak hem hipotiroidi hastalığının teşhis edilmesi sağlanacak hem de bu hastalığın teşhis edilmesine yönelik hastaya uygulanan testlerin ve tetkiklerin azaltılması yöntemlerine başvurulacaktır.

Bu çalışmada da farklı örnekleme teknikleri kullanılarak Lojistik Regresyon, K En Yakın Komşu ve Destek Vektör Makinesi Algoritmaları ile hipotiroidi hastalığı tanısında %92'in üzerinde başarı oranları elde edilmiştir.

Bu çalışmada ele alınan veri setindeki örneklerin Avustralya'ya ait olması ve veri seti büyüklüğü çalışmayı sınırlandırmıştır. Tahmin modellerinin eğitildiği örneklerin belirli bir ülkeye ait olması yanlış tahminlere neden olabilir. Veri seti büyüklüğünün sınırlı olması tahmin modellerinin başarısını etkileyebilir. Gelecek çalışmalar için farklı ülkelere ait daha büyük boyuttaki veri setleri ile çalışılması önerilmektedir.

Gelecek çalışmalar için bir diğer öneri de teşhis problemlerinde hataları en aza indirmek ve doktorlara kullanım kolaylığı sağlamak için daha doğru ve kesin sonuçlar elde etmek amacıyla tüm değerlendirme kriterlerinde en yüksek sonucu veren Lojistik Regresyon sınıflandırıcısı tabanlı bir klinik karar destek sisteminin geliştirilmesidir.

## KAYNAKÇA (REFERENCES)

- [1] B. Çakır, F. Sağlam, "Birinci Basamakta Tiroid Hastalıklarına Klinik Yaklaşım", *Ankara Medical Journal*, 12(3), 136-139, 2012.
- [2] K. Yılancıoğlu, "Vocal Cord Measures Based Artificial Neural Network Approach for Prediction of Parkinson's Disease Status", *SDÜ Sağlık Bilimleri Enstitüsü Dergisi*, 8(2), 8-11, 2017.
- [3] Internet: UCI Machine Learning Repository, <https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease>, 09.2019.
- [4] E. Kaya, M. Bulun, A. Arslan, "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları", *Akademik Bilişim 2003*, Adana, 2003.
- [5] Ö. Demir, B. Doğan, E. Ç. Bayezit, K. Yıldız, "Automatic Detection and Calculation of Drusen Areas in Retinal Fundus Fluorescein Angiography Images", *Marmara Fen Bilimleri Dergisi*, 2, 128-132, 2019.
- [6] A. Buldu, K. Yıldız, E. E. Ülkü, Ö. Demir, U. Kurgan, "Data Collection from Blood Glucose Meter and Anomaly Detection", *Karaelmas Fen ve Mühendislik Dergisi*, 7(2), 428-433, 2017.

- [7] Z. Chiara, "Data Mining in Bioinformatics", **Encyclopedia of Bioinformatics and Computational Biology**, 328-335, 2019.
- [8] M. Sert, "Feature Selection for Obstructive Sleep Apnea Recognition", *Bilişim Teknolojileri Dergisi*, 12(4), 333-342, 2019.
- [9] N. Alpaslan, "Meme Kanseri Tanısı için Derin Öznitelik Tabanlı Karar Destek Sistemi", *Selçuk Üniversitesi Mühendislik, Bilim Ve Teknoloji Dergisi*, 7(1), 213-227, 2019.
- [10] M. A. Pala, M. E. Çimen, Ö. F. Boyraz, M. Z. Yıldız, A. F. Boz, "Meme Kanserinin Teşhis Edilmesinde Karar Ağacı Ve KNN Algoritmalarının Karşılaştırmalı Başarım Analizi", **7th International Symposium on Innovative Technologies in Engineering and Science**, Şanlıurfa, 2019.
- [11] S. Bang, S. Son, H. Roh, J. Lee, S. Bae, K. Lee, C. Hong, H. Shin, "Quad-Phased Data Mining Modeling for Dementia Diagnosis", *BMC Medical Informatics and Decision Making*, 17(60), 2017.
- [12] M. Shouman, T. Turner, R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment", in **2012 Japan-Egypt Conference on Electronics, Communications and Computers**, Alexandria, 2012.
- [13] F. C. D. Q. Mello, L. G. d. V. Bastos, S. L. M. Soares, V. MC Rezende, M. B. Conde, R. E. Chaisson, A. L. Kritski, A. R. Netto, G. L. Werneck, "Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study", *BMC Public Health*, 6(43), 2006.
- [14] S. Kılıçarslan, K. Adem, O. Cömert, "Parçacık Sürü Optimizasyonu Kullanılarak Boyutu Azaltılmış Mikrodizi Verileri Üzerinde Makine Öğrenmesi Yöntemleri ile Prostat Kanseri Teşhisi", *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, cilt 7, 769-777, 2019.
- [15] B. O. Yolcular, U. Bilge, M. K. Samur, "Extracting Association Rules from Turkish Otorhinolaryngology Discharge Summaries", *Bilişim Teknolojileri Dergisi*, 11(1), 35-42, 2018.
- [16] S. Dash, M. N. Das, B. K. Mishra, "Implementation of an optimized classification model for prediction of hypothyroid disease risks", **2016 International Conference on Inventive Computation Technologies (ICICT)**, Coimbatore, 2016.
- [17] İ. Türkoğlu, Ş. Doğan, "Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees", *Doğu Anadolu Bölgesi Araştırmaları Dergisi*, 5(2), 163-169, 2007.
- [18] W.-C. Yeh, "Novel swarm optimization for mining classification rules on thyroid gland data", *Information Sciences*, 197, 65-76, 2012.
- [19] Y. Kaya, "Fast Intelligent Diagnosis System For Thyroid Disases Based On Extreme Learning Machine", *Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering*, 15(1), 41-49, 2014.
- [20] M. Deepika, K. Kalaiselvi, "A Empirical study on Disease Diagnosis using Data Mining Techniques", **2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)**, Coimbatore, 2018.
- [21] N.A. Sajadia, S. Borzouei, H. Mahjub, M. Farhadian, "Diagnosis of hypothyroidism using a fuzzy rule-based expert system", *Clinical Epidemiology and Global Health*, 7(4), 519-524, 2019.
- [22] U. Fayyad, "Data Mining and Knowledge Discovery in Databases: Implications for scientific databases", **Proc. of the 9 th Int Conf on Scientific and Statistical Database Management**, Olympia, Washington, USA, 1997.
- [23] P. Giudici, **Applied Data Mining: Statistical Methods for Business and Industry**, New York: John Wiley, 2003.
- [24] N. A. Sundar, P. P. Latha, M. R. Chandra, "Performance Analysis Of Classification Data Mining Techniques Over Heart Disease Data Base", *International Journal of Engineering Science & Advanced Technology*, 2(3), 470-478, 2012.
- [25] H. Bircan, "Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama", *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, cilt 2, 185-208, 2004.
- [26] Internet: Imbalanced-learn, <https://imbalancedlearn.readthedocs.io/en/stable/api.html>, 01.2020.
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Coumapeau, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, cilt 12, 2825-2830, 2011.
- [28] Y. Liu, X. Yu, J. X. Huang, A. An, "Combining Integrated Sampling with Svm Ensembles for Learning from Imbalanced Datasets", *Information Processing & Management*, 47(4), 617-631, 2011.
- [29] M. Eminağaoğlu, A. Vahaplar, "Turnaround Time Prediction for a Medical Laboratory Using Artificial Neural Networks", *Bilişim Teknolojileri Dergisi*, 11(4), 357-368, 2018.
- [30] W. Ahmad, A. Ahmad, C. Lu, B.A. Khoso, L. Huang, "A novel hybrid decision support system for thyroid disease forecasting", *Soft Computing*, 22, 5377-5383, 2018.
- [31] G. Serpen, H. Jiang, L. Allred, "Performance analysis of probabilistic potential function neural network classifier" **In: Proceedings of artificial neural networks in engineering conference**, St. Louis, MO, USA. Citeseer, 471-476, 1997.
- [32] L. Özyılmaz, T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods", **In: Proceedings of the 9th international conference on neural information processing, 2002. ICONIP'02 2002**. IEEE, 2033-2036, 2002.
- [33] L. Pasi, "Similarity classifier applied to medical data sets, 2004, 10 sivua, Fuzziness in Finland'04". **In: International conference on soft computing**, Helsinki, Finland & Gulf of Finland & Tallinn, Estonia, 2004.
- [34] K. Polat, S. Güneş, "A hybrid medical decision making system based on principles component analysis, k-NN based weighted pre-processing and adaptive neuro-fuzzy inference system", *Digit Signal Proc*, 16, 913-921, 2007.
- [35] F. Temurtas, "A comparative study on thyroid disease diagnosis using neural networks", *Expert Systems with Applications*, 36, 944-949, 2009.
- [36] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, Q. Wang, "A hybrid classification system for heart disease diagnosis based on the RFRS method", *Computational and Mathematical Methods in Medicine*, 2017, <https://doi.org/10.1155/2017/8272091>, 2017.

- [37] N.M. Sundaram, V. Renupriya, "Artificial neural network classifiers for diagnosis of thyroid abnormalities". In: International conference on systems, science, control, communication, engineering and technology, 206–211, 2016.
- [38] N. Rajkumar, J. Palanichamy J. "Optimized construction of various classification models for the diagnosis of thyroid problems in human beings", *Kuwait Journal of Science*, 42, 198–205, 2015.