Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 258-264, Ağustos 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 258-264, August 2020

Research Article

Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması

Mustafa Yavaş[1*], Aysun Güran[2], Mitat Uysal[3]

[1] Doğuş Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-9111-9095)
[2] Doğuş Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0001-7066-0635)
[3] Doğuş Üniversitesi, Mühendislik Fakültesi, Yazılım Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0001-9713-2525)

(Bu yayın 26-27 Haziran 2020 tarihinde HORA-2020 kongresinde sözlü olarak sunulmuştur.)

(**DOI:** 10.31590/ejosat.779952)

ATIF/REFERENCE: Yavaş, M., Güran, A. & Uysal, M. (2020). Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (Special Issue), 258-264.

Öz

Son yıllarda dengesiz tıbbi veri kümeleri üzerinde gerçekleştirilen öğrenme problemine verilen önem artmaktadır. Çünkü gerçek yaşamda karşılaşılan tıbbi veri kümeleri sıklıkla dengesiz veri kümeleridir. Sınıflandırıcıların dengesiz ortamdaki davranışlarını inceleyen pek çok çalışma, başarım değerlerindeki önemli kaybın veri kümelerinde oluşan çarpık sınıf dağılımından kaynaklandığını vurgulamıştır. Literatürde, bu çarpıklık sorununu çözmek için Sentetik Azınlık Örneklem Arttırma Yöntemi (SMOTE) algoritması önerilmiştir. Bu çalışmada, hastanelere yapılan şüpheli bir Covid-19 vaka başvurusunda, yaygın olarak toplanan laboratuvar test sonuçlarına dayanarak, SARS-Cov-2 test sonucu negatif veya pozitif sınıfa sahip hastaları SMOTE ve YSA modeli kullanarak daha yüksek oranla tahmin etmeye yönelik deneysel çalışma yapılmıştır. Orijinal veri kümesinin YSA ile sınıflandırılması sonucunda doğruluk değeri 0.86, F-ölçüm değeri 0.48 bulunmuş olup, SMOTE ile dengelenen veri kümesinin yine YSA ile sınıflandırılması sonucunda doğruluk değeri 0.90, F-ölçüm değeri 0.68 bulunmuştur. Bu nedenle SMOTE ile dengelenmiş Covid-19 veri kümesinin YSA ile sınıflandırılması sonucunda daha başarılı sonuçlar bulunmuştur. Çalışmamızın sonunda orijinal ve SMOTE ile dengelenen veri kümesi arasında karşılaştırma yapılmış olup, sınıflandırıcının diğer başarım değerlerini de arttırdığı görülmüştür.

Anahtar Kelimeler: Covid-19, Dengesiz tıbbi veri kümesi, SMOTE, Yapay sinir ağları.

Classification of Covid-19 Dataset by Applying Smote-based Sampling Technique

Abstract

In recent years, the importance given to the learning problem performed on unbalanced medical datasets has been increasing. Because real life medical datasets are often unbalanced datasets. Many studies examining the behavior of classifiers in an unstable environment have emphasized that the significant loss in performance values is due to the distorted class distribution in datasets. In the literature, the Synthetic Minority Sampling Method (SMOTE) algorithm has been proposed to solve this distortion problem. In this study, an experimental study was conducted in a suspected Covid-19 case application to predict patients with a negative or positive class with a higher rate of SARS-Cov-2 test results based on commonly collected laboratory test results. As a result of the classification of the original dataset with Artificial neural network (ANN), the accuracy value was found to be 0.86, the F-measure value was 0.48, and the dataset balanced with SMOTE was again classified by ANN, and the accuracy value was found to be 0.90 and the F-measure value was 0.68. For this reason, Covid-19 dataset balanced with SMOTE was classified with ANN and more successful results were found. At the

[*] Sorumlu Yazar: Doğuş Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, ORCID: 0000-0002-9111-9095, 20172104001@dogus.edu.tr

end of our study, a comparison was made between the original and SMOTE balanced dataset, and it was seen that the classifier also increased other performance values.

Keywords: Covid-19, Unbalanced medical dataset, SMOTE, Artificial neural networks.

1. Giriş

Son yıllarda, teknolojinin hızla ilerlemesiyle birlikte makine öğrenmesi algoritmalarını temel alan ve tıbbi veri kümelerinin sınıflandırılmasına yoğunluk veren çalışmaların sayısında bir artış söz konusu olmuştur. Yapılan çalışmalar, özellikle dengesiz tıbbi veri kümelerinin sınıflandırılması amacıyla uygulanmıştır. Dengesiz veri kümeleri söz konusu olduğunda sınıflandırıcıların başarım değerlerinde bir düşüş görülmektedir. Bu nedenle sınıflandırma aşamasına geçilmeden önce, dengesiz veri kümelerini dengeli hale getiren veri örnekleme yöntemlerine başvurulması, sınıflandırıcıların başarım değerlerinin arttırılması adına önemli bir aşama olarak karşımıza çıkmaktadır.

Literatürde gerek tıbbi, gerekse diğer alanlarda karşılaşılan dengesiz veri kümelerinin dengeli hale getirilmesi ile ilgilenen birçok çalışma yapılmıştır (Chawla ve ark., 2004; Oğul ve ark., 2019). Bu tür dengesiz veri kümelerinde öğrenme aşamasında örneklem sayısının fazla olduğu sınıflar baskın gelmekte ve azınlık sınıflara ait olan gözlem değerleri sınıflandırılırken bazı dengesizlikler görülebilmektedir. Tıbbi verilerde yapılan çalışmalar genellikle dengesiz veri kümesi problemleriyle karşılaşılmaktadır. Hastalık teşhisi ve maliyet gerektiren tanı testleri, hastalık yaygınlığının düşük olması veya hastalıkla ilgili verilerde örnek sayısının az olması, verilerin oluşturulması ve toplanmasının bazen yüksek maliyet gerektirmesi, numerik ve görüntü tabanlı tıbbi veri kümelerini kısıtlamaktadır (Berner, 2006). Görüntü tabanlı yapılan bir çalışmada, göğüs röntgeni görüntülerinde Covid-19 ve Pnömoni hastaları için suni örnekleme yöntemleri ile makine öğrenmesi ve derin öğrenme algoritmaları kullanılarak, hastalık yüksek oranda tahmin edilmeye çalışılmıştır (Kumar, 2020). 235 yetişkin hastadan alınan sayısal örnekler, makine öğrenmesi algoritmaları ile Covid-19 pozitif tanı riski taşıyan hastalar tahmin edilmeye çalışılmıştır (De Moraes Batista ve ark., 2020). Koronavirüs hastalığı için klinik öngörücü modellerin sistematik bir çalışmasıyla, klinik yolları tahmin etmek, bakım için bilgilendirmeye ve kaynakları önceliklendirmeye yardımcı olabileceğini göstermiştir (Schwab, 2020). Tıbbi veri kümesi hepatit virüs tanısı için YSA'ya dayalı bir yaklaşım sunarak, hastaların performansını etkileyebilecek birçok faktör özetlenmiştir. Bu faktörler YSA modelinde girdi değişkeni olarak kullanılıp, YSA modelinin olası hastaların %93'ünden fazlasının tanısını doğru bir şekilde tahmin edebildiğini göstermiştir (AbuSharekh, 2018). Dengesiz verilerin özellikle tıbbi bilişimde sınıflandırılması zordur ve yeniden dengeleme algoritması kullanarak bir sınıflandırıcı geliştirme ihtiyacı duyulmuştur. İki aşamalı bir sınıflandırma modeli kullanılarak, birinci adımda, literatürde önerilen SMOTE modeli ile veriler ön işlemden geçirilir ve ikinci adımda makine öğrenmesi ve derin öğrenme algoritmaları ile diyabet ve diğer tıbbi verilerin tahmininde dengeli veri kümesinin en iyi sınıflandırıcısını seçmek amaçlanmıştır (Shuja, 2020). Tıbbi alanda yapılan başka bir çalışmada, SMOTE başka tekniklerle Tomek (Tomek,1976) bağlantılar tekniği birleştirilerek güçlü bir ön işleme yöntemi önerilmiş ve diyabet, parkinson, vertebral kolon dengesiz tıbbi veri kümelerine uygulanarak daha yüksek başarım değerleri elde etmiştir (Zeng, 2016). Yapılan tıbbi çalışmalarda nümerik ve görüntü tabanlı kullanılan verilerin çoğunda hastalığa ait gözlem sayısı azdır. Veri kümesinde bulunan bu gerçek gözlemlere bakılarak suni örnekleme yöntemleri ile sınıflandırıcıların model performansını arttırmak için veri örnekleme yaklaşımı önerilmiştir. Chawla ve arkadaşları tarafından önerilen SMOTE algoritması dengesiz veri kümeleri ile mücadele eden en iyi algoritmalar arasındadır (Chawla ve ark., 2002). Önerilen bu yöntem çalışmada kullanılan tıbbi veriler yanında birçok alanda; biyomedikal bilişim, yüz tanıma, kayıp müşteri tespiti, robotik, örüntü tanıma, dolandırıcılık tespiti, doğal dil işleme gibi benzeri birçok karmaşık ve gerçek problemleri çözmek amacıyla geniş olarak incelenmiş ve birçok alanda uygulanmıştır.

Bu çalışmada hastanelere yapılan şüpheli bir Covid-19 vaka başvurusunda, yaygın olarak toplanan laboratuvar test sonuçlarına dayanarak, SARS-Cov-2 test sonucu negatif veya pozitif sınıfa sahip hastaları SMOTE ve YSA modeli kullanarak daha yüksek oranla tahmin etmeye yönelik deneysel çalışma yapılmıştır. Çalışmamızda incelediğimiz veri kümesinin dengesiz olma sebebi veri kümesindeki bir sınıfın çok sayıda örnekleme sahipken, diğer sınıfın daha az sayıda örneklem içermesidir. Çalışmamızda gerçekleştirdiğimiz deneyler sonucunda dengesiz veri kümelerinin dengeli hale getirilmesi için literatürde önerilen SMOTE tabanlı örneklem arttırma yöntemi Covid-19 veri kümesini dengeli hale getirdiği ve YSA algoritmasının başarım değerlerinin arttırdığı gösterilmiştir. Başarım değeri olarak, doğruluk (accuracy), kesinlik (precision-P), duyarlılık (recall-R), F-ölçüm, Aritmetik ortalama ve Ağırlıklı ortalama değerleri dikkate alınmıştır.

Çalışmamızın diğer bölümleri şu şekildedir: İkinci bölümde materyal ve metot, çalışmada kullanılan veri kümesi, örnekleme yöntemi, sınıflandırma algoritması ve değerlendirme metriklerinden bahsedilmiştir; Üçüncü bölüm araştırma sonuçları ve tartışma kısmında, sınıflandırma sonucu başarım değerleri ve kıyaslama sonuçları paylaşılmış ve yorumlarda bulunulmuştur. Son bölümde ise çalışmanın sonuç bilgileri paylaşılmıştır.

2. Materyal ve Metot

2.1. Çalışmada Kullanılan Veri Kümesi

Çalışmada kullanılan Covid-19 veri kümesi Kaggle veri bilimci ve makine öğrenme çalışmalarında bulunan çevrimiçi bir topluluk olan platformdan alınmıştır. 2020 yılının ilk aylarında Brezilya'nın São Paulo şehrinde Israelita Albert Einstein Hastanesi'nde görülen ve hastaneye ziyareti sırasında SARS-CoV-2 RT-PCR ve ek laboratuvar testleri yapmak için toplanan 5644 hastadan alınan anonim verilerilerdir (Kaggle, 2020). Tüm klinik veriler ortalama sıfır ve birim standart sapmaya sahip olacak şekilde paylaşılan platform

tarafından standartlaştırılmıştır. Veri kümesinin kullanılma amacı, hastanelerin acil servisine yapılan şüpheli bir Covid-19 vaka başvurusunda, yaygın olarak toplanan laboratuvar test sonuçlarına dayanarak, SARS-Cov-2 (pozitif / negatif) sonucunu tahmin etmektir. Pozitif sonuçlanan vakalar hem hasta hemde hastanelerin yoğunluğu açısından büyük önem taşıdığından, çalışmamızda pozitif vakaları yani gerçekte hasta olan kişileri yüksek oranda tespit etmek amaçlanmıştır. Veri kümesinde 5644 örnek bulunmakta olup, her bir örnek 111 niteliğe sahiptir. 5644 örnekten 5086'sı sağlıklı (test sonucu negatif) sınıfta ve 558'i hasta (test sonucu pozitif) sınıfına aittir. Bu sayılar ve orana bakıldığında pozitif sınıfa ait verilerin azınlıkta olduğu ve veri kümesinin dengesiz bir dağılıma sahip olduğu görülmektedir.Veri kümesi yüksek oranda boş değer içermekte olup, sınıflandırıcılarda eğitilmeden önce, ön işlem aşamasında çoğunluğunu boş değer içeren nitelikler ihmal edilmiştir. Bu ihmal işlemleri ve veri kümesi ön işlem adımları aşağıda açıklanmıştır. Herhangi bir ön işleme tabi tutulmamış orijinal veri kümesinin bir bölümü Şekil 1'de gösterilmiştir.

Patient ID	Patient age quantile	SARS-Cov-2 exam result	Patient addmited to regular ward (1=yes, 0=no)	Patient addmited to semi-intensive unit (1=yes, 0=no)	Patient addmited to intensive care unit (1=yes, 0=no)	Hematocrit	Hemoglobin	Platelets	Mean platelet volume	...
44477f75e8169d2	13	negative	0	0	0	NaN	NaN	NaN	NaN	...
126e9dd13932f68	17	negative	0	0	0	0.236515	-0.02234	-0.517413	0.010677	...
a46b4402a0e5696	8	negative	0	0	0	NaN	NaN	NaN	NaN	...
f7d619a94f97c45	5	negative	0	0	0	NaN	NaN	NaN	NaN	...
d9e41465789c2b5	15	negative	0	0	0	NaN	NaN	NaN	NaN	...

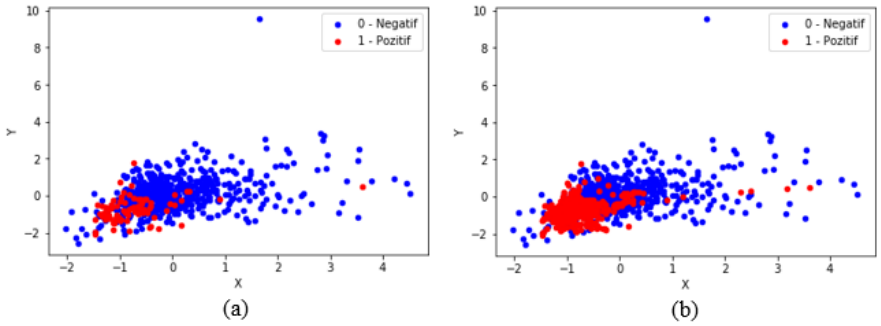
Şekil 1 Covid-19 Orijinal veri kümesi

Öncelikle tamamen boş değer içeren (6 nitelik), en fazla 3 değer içeren (3 nitelik) ve SARS-Cov-2 analiz sonucu üzerinde etkisi olmayan (Hasta ID niteliği) ve hastanın hangi servise (normal servis, yarı yoğun ve yoğun bakım) yönlendirildiği ile ilgili (3 nitelik) ihmal edilerek 98 niteliğe düşürülmüştür. Kalan niteliklerden %90 ve üzeri boş değer içeren 65 nitelik olup toplam 5079 boş değer içermektedir. Bu nitelikler eğitim aşamasında algoritmanın performasını olumsuz etkileyeceğinden ihmal edilerek yeni nitelik sayısı 33'e düşürülmüştür. Kalan niteliklerin 20'si yine çoğunlukta boş değer içeren, genelde hastalara ait daha önce geçirdiği herhangi bir rahatsızlığı olup olmadığı ile ilgili kategorik (tespit edildi-1, tespit edilmedi-0) bilgilerdir. Bu niteliklerden herhangi birine sahip olan hastaların nitelik değerini (1 ve 0) olacak şekilde tek bir nitelik altında toplanmıştır. Nitelik sayısı en optimum düzeye getirilerek sınıflandırma öncesi 14 niteliğe düşürülmüştür. Son ihmal işlemi ise, 5644 hasta örneği, kalan niteliklerin %80'ninde (11 nitelikte) boş değer içermektedir. Bu örneklemlerin yine sınıflandırıcı algoritmanın performasını olumsuz etkileyeceğinden, 5644 örnekten 5042'si ihmal edilerek 602'ye düşürülmüştür. Kalan 14 nitelik ve 602 örnek tamamen dolu değerlere sahip olup YSA'da eğitilmek üzere hazır hale getirilmiştir. Bu niteliklerin tamamı ve örneklemlerin bir kısmı Şekil 2'de gösterilmiştir.

Hasta yaş kantili	SARS-Cov-2 sonucu	Trombositler	Ortalama trombosit hacmi	Kırmızı kan hücreleri	Lenfositler	Ortalama korpüsküler hemoglobin konsantrasyonu (MCHC)	Lökositler	Bazofil	Eozinofiller	Ortalama korpüsküler hacim (MCV)	Monocytes	Kırmızı kan hücresi dağılım genişliği (RDW)	Diğer test sonuçları
17	0	-0.517413	0.010677	0.102004	0.318366	-0.950790	-0.094610	-0.223767	1.482158	0.166192	0.357547	-0.625073	1
1	0	1.429667	-1.672222	-0.850035	-0.005738	3.331071	0.364550	-0.223767	1.018625	-1.336024	0.068652	-0.978899	0
9	0	-0.429480	-0.213711	-1.361315	-1.114514	0.542882	-0.884923	0.081693	-0.666950	1.668409	1.276759	-1.067355	1
11	0	0.072992	-0.550290	0.542763	0.045436	-0.452899	-0.211488	-0.834685	-0.709090	0.606842	-0.220244	0.171035	1
9	0	-0.668155	1.020415	-0.127191	0.002791	-1.249524	-1.132592	0.387152	-0.709090	0.566783	2.012129	0.613318	0

Şekil 2 Ön işlem sonrası Covid-19 veri kümesi

Ön işlem sonrası veri kümesinde toplam 602 örnek olup, bunlardan 519'unun test sonucu negatif sınıfa ait çoğunluk grubunda, 83'ünün test sonucu pozitif sınıfa ait olup azınlık grubundadır. Sınıf dağılım sayısına bakıldığında pozitif sınıfa ait verilerin azınlıkta olduğu ve veri kümesinin dengesiz bir dağılıma sahip olduğu görülmektedir. Veri kümesi %80 eğitim ve kalan kısmı test verisi olarak ayrılmış olup, YSA'da eğitilmiştir. Eğitim kısmı için ayrılan negatif sınıftan 415, pozitif sınıftan 66 örnek, SMOTE örnekleme algoritması ile (negatif 415, pozitif 415) dengelenerek tekrar YSA ile eğitilmiştir. Dengelenen veri kümesinin sınıf dağılımını görselleştirmek için, Lökositler (x) ve Trombositler (y) nitelikleri baz alınarak Şekil 3'te SMOTE öncesi ve sonrası iki boyutlu düzlemde gösterilmiştir.

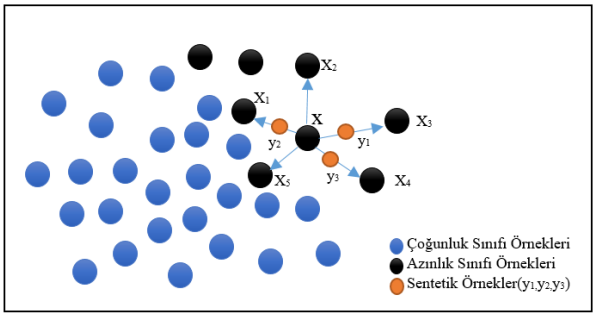


Şekil 3 SMOTE öncesi(a) ve SMOTE sonrası(b) veri dağılımı

Şekil 3'ten anlaşılacağı üzere azınlık gruptaki pozitif sınıfların çoğunluk gruba yaklaştırıldığı ve yoğunlukta pozitif örneklerin bulunduğu bölgede en yakın k-komşu değerlerin örneklendiğini görülmektedir. Sınıflandırma performansını arttırmak için, azınlık gruba ait örnekler, çalışmada önerilen örnekleme yöntemi ile suni olarak çoğaltılıp veri kümesi dengeli hale getirilmiştir. Çalışmamız kapsamında kullanmış olduğumuz veri kümesi ve algoritmalar Python kütüphaneleri kullanılarak kodlanmıştır.

2.2. Sentetik Azınlık Örneklem Arttırma Yöntemi (SMOTE)

SMOTE yöntemi, en yaygın kullanılan ve çoğu zaman en başarılı örnekleme yöntemi olarak bilinmektedir. 2002 yılında geliştirilen algoritma birçok dengesiz veri kümesi problemine uygulanmıştır (Chawla ve ark., 2002). Rassal örnekleme yöntemlerinden farkı, azınlık sınıfı verileri kopyalamak yerine, incelenen örneklerin en yakın k komşusunu baz almak suretiyle yapay örnekler üretmesidir. Şekil 4'te algoritmanın çalışma sistemi gösterilmiştir.



Şekil 4. SMOTE algoritması çalışma sistemi

Algoritmanın çalışma adımları aşağıda belirtilen şekilde özetlenebilir:

- Adım-1: Azınlık sınıfına ait her gözlemin k yakın komşusu aranır,
- Adım-2: Azınlık sınıfına ait gözlem ile k yakın komşusu (kNN) olan gözlem arasındaki fark alınır,
- Adım-3: (0,1) arasında rastgele bir sayı (α) seçilir, Adım 2'de bulunan fark ile bu sayı çarpılır,
- Adım-4: Eşitlik 1 kullanılarak yeni sentetik gözlem elde edilir.

$$x_{yeni} = x_i + (x_j - x_i) \ * \alpha \qquad (1)$$

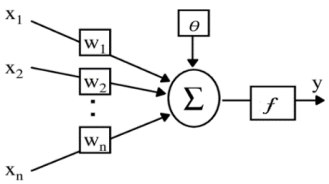
- Adım-5: İstenen sayıda sentetik gözlem oluşturmak için Adım 1-4 yinelenir.

2.3. Yapay Sinir Ağları (YSA) Algoritması

Yapay sinir ağları, insan beyni gibi sinir sistemlerinin bilgi işleme yöntemlerinden esinlenerek geliştirilmiş matematiksel modellerdir (Mitchell, 2009). Bilimsel açıdan insan beynindeki nöron hücre yapısını model olarak ortaya konulmasıyla günümüzde YSA geliştirilerek birçok alanda kullanılmaya başlanmıştır. YSA farklı topolojik ve sezgisel yapılarından dolayı pek çok mühendislik ve tıp alanlarda sınıflandırma, öğrenme, ilişkilendirme, optimizasyon ve özellik belirlemede kullanılmaktadır. YSA'nın biyolojik yapısında sinir ağları sistemi 10 milyardan fazla sinir hücresinin birbirleri ile bağlanmasından meydana gelmektedir. Bir sinir hücresi, soma (hücre gövdesi), akson ve dendrit yapılardan oluşmaktadır. Temel olarak soma, hücre çekirdeği, çekirdekçik ve sitoplazma sıvısından oluşmaktadır. Akson ise lifli yapıdan meydana gelmiştir. Görevi ise somadan gelen sinyalleri uçları yardımıyla diğer sinir hücrelerine taşımaktır. Dendritleri yapılarında bulunan lifler ile diğer sinir hücrelerin dendritleri ile bağ kurmaktadır. Bu yapı sayesinde nöronlardaki sinyaller somaya iletilmektedir. Snapslar olarak adlandırılan bağlantı elemanları sayesinde sinir hücreleri birbirleri ile iletişim kurmaktadır (Erkaymaz, 2014).

Yapısal olarak YSA, yapay nöronların kendi aralarında kurulan bağlantılardır. Nöronlar arasındaki bağlantı şekilleri ağ topolojisi modelini oluşturmaktadır. Yapay nöronların davranışları topolojik model ile matematiksel olarak ifade edilmektedir. Yapay nöron hücrelerinin davranışlarına benzer şekilde, kendine gelen bilgileri analog toplayıcı benzeri bir yöntem ile toplar. Elde edilen toplam veri belirlenen eşik değerine göre işlem yaparak iletimine karar veren paralel işlem birimidir (Du ve ark., 2002).

McCulloch ve Pitts tarafından 1943 'te önerilen yapay nöron modeli yapay sinir hücresinden diğerine sinapslar yoluyla sinyal göndererek, sinyali alan hücre gövdesinin elektrik potansiyel değerinin yükseltilmesi veya düşürülmesi yoluyla gerçekleşir. Bu potansiyelin bir eşik değerine varması halinde nöron ateşlenir (McCulloch ve ark.,1943).



Şekil 5. McCulloch-Pitts Yapay sinir hücresi hesaplama modeli

Şekil 5'te verilen modelin x_1, x_2, x_n ile gösterilen n tane giriş verisi vardır. Bu giriş verileri w_1, w_2, w_n ağırlıklarına bağlıdır. Ağırlıklar sinaptik bağlantılara karşılık gelmektedir. Yapay sinir hücresinin aktivasyon değeri, her bir giriş verisinin kendisine karşılık gelen ağırlıkla çarpılıp toplanmasından sonra θ eşik değerinin eklenmesiyle hesaplanır:

$$a = \sum_{j=1}^{n} x_j w_j + \theta \qquad (2)$$

Eşik değeri, sanal bir giriş değeri olarak gösterilip $x_o = 1$ ve $w_0 = \theta$ olarak tanımlandığında, aktivasyon formülü Eş.3 olur.

$$a = \sum_{j=1}^{n} x_j w_j \qquad (3)$$

Nöronun çıkış değeri, biyolojik sinir hücresinin ateşleme frekansına benzer şekilde, aktivasyonun bir fonksiyonudur:

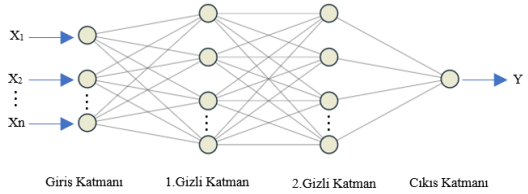
$$y = f(a) \qquad (4)$$

Doğrusal olmayan problemlerin analizinde beklenen çıkışa yakınsama yapabilmek için aktivasyon fonksiyonları kullanılmaktadır. Bu yüzden problem türüne göre aktivasyon fonksiyonu seçmek önemli bir parametredir. YSA modelinde genelde sigmoid fonksiyonu, birim basamak fonksiyonu, doğrusal fonksiyon, rampa fonksiyonu ve hiperbolik tanjant fonksiyonları aktivasyon fonksiyonları kullanılmaktadır (Gürsoy, 2018). Genelde doğrusal olmayan problem analizlerinde tercih edilen Sigmoid fonksiyonu, çalışmamızda kullanılmıştır. Sigmoid fonksiyonu Eş. 2 ile hesaplanmaktadır.

$$f(x) = \frac{1}{1+e^{-x}} \qquad (5)$$

YSA sistem mimarisi girdi katmanı, gizli katmanlar ve çıktı katmanı olmak üzere 3 katmandan oluşur. Her katmanda belirli sayıda yapay sinir hücreleri bulunmaktadır. Gizli katmanlarda sinir hücrelerine ek olarak toplama ve aktivasyon fonksiyonları bulunur. Sinir hücreleri arasında bulunan ağlara başlangıçta sistem tarafından rastgele bir ağırlık değeri atanır. Bu değerler sonradan gelen veriler ile güncellenir ve yeni verilerin sistem üzerindeki etkisini hesaplamada kullanılır.

YSA nın giriş katmanı, verilerin sisteme alındığı ilk katmandır. Bu katmanda alınan verileri bir eşik değeri eklenerek ve ağırlık katsayılarıyla ile çarpılarak bir sonraki gizli katmanlara gönderilir. Gizli katman bir ya da daha fazla katmandan oluşabilir. Bu katmanda, giriş katmanından gelen veriler toplanır ve aktivasyon fonksiyonu uygulanır. Çıkış katmanında işlenen veriler değerlendirilerek çıktı elde edilir (Öztemel, 2003). YSA çoklu katman mimari yapısı Şekil 5'te gösterilmiştir.



Şekil 5. Yapay sinir ağları örnek

Çalışmamız Şekil 5'te verildiği gibi çok katmanlı YSA olup; giriş katmanı, 2 gizli katman ve çıkış katmanından oluşan çok katmanlı modeldir. SARS-Cov-2 teşhisinde kullanılmak üzere 13 adet nitelik bulunmaktadır. Bu yüzden giriş katmanında 13 nöron, 1. gizli katmanda 26 nöron, 2. gizli katmanda 39 nöron ve çıkış katmanında 1 nöron optimize edilerek kullanılmıştır. Veri kümesinin %80'i ile yapay sinir ağı eğitilmiş ve eğitim sırasında epoch(döngü) sayısı 60 olarak seçilmiştir. Kalan %20'lik kısmı ile eğitilen yapay sinir ağının başarısı ölçülmüştür. Başarım değerleri için çıktı katmanında, Eş. 5'te verilen aktivasyon fonksiyonundan çıkan değer hücrenin çıktı değeri olmaktadır.

2.4. Değerlendirme Metrikleri

Dengesiz veri kümeleri üzerinde uygulanan sınıflandırma metotlarının başarım değerlerinin kıyaslanması için doğruluk değerinin dışında, kesinlik (precision-P), duyarlılık (recall-R) ve F-ölçüm gibi değerlendirme kriterlerinin kullanılması daha güvenilir bir ortam yaratacaktır. Sadece doğruluk değerinin kullanılmaması sınıflardaki dengesiz dağılımın dikkate alınmamasından ötürü oluşabilecek güvensiz koşullardır. Tablo I ile verilen karmaşıklık matrisi baz alınarak doğruluk, P, R ve F-ölçüm değerleri Eş.6-9 ile belirtilen şekilde tanımlanmaktadır.

Tablo 1. Karmaşıklık Matrisi

		Tahmin Edilen Değerler	
		Negatif	*Pozitif*
Gerçek Değerler	*Negatif*	TN	FP
	Pozitif	FN	TP

Karmaşıklık matrisinde belirtilen ifadelerin kullanılmasıyla sınıflandırıcıların doğruluk değeri Eş. 6 ile belirtilen şekilde hesaplanır:

$$\text{Doğ}ruluk = (TP + TN)/(TP + FP + TN + FN) \qquad (6)$$

İncelediğimiz veri kümesi iki sınıflı bir veri kümesi olduğundan pozitif ve negatif durumlar için eşitlikler ayrı ayrı verilmiştir. Pozitif sınıf için hesaplamalar Eş. 7-9 ile gösterildiği gibidir:

$$P^+ = TP/(TP + FP) \tag{7}$$

$$R^+ = TP/(TP + FN) \tag{8}$$

$$F - \text{ölçüm} = 2 * P * R/(P + R) \tag{9}$$

Negatif sınıf için hesaplamalar ise Eş. 10-11 ile belirtilmiştir. Negatif sınıf için F-ölçüm değeri pozitif sınıf için hesaplanan F-Ölçüm değeri formülü ile aynıdır (Eş. 9)

$$P^- = TN/(TN + FN) \tag{10}$$

$$R^- = TN/(TN + FP) \tag{11}$$

Genel durumu ifade etmek için algoritmaların pozitif ve negatif sınıflar üzerindeki başarım değerlerinin ortalaması alınmış ve ortalama alınırken hem aritmetik ortalama hem de veri kümesindeki her sınıfın oranını (ağırlığını) dikkate alan ağırlıklı ortalama değerleri elde edilmiştir. Bu durumlar Eş.12-13 ile belirtilmiştir:

Aritmetik Ortalama: $P = (P^+ + P^-)/2,\ R = (R^+ + R^-)/2,\ F = (F^+ + F^-)/2$ (12)

Ağırlıklı Ortalama: $P = (W_1P^+ + W_2P^-)/(W_1 + W_2),\ R = (W_1R^+ + W_2R^-)/(W_1 + W_2),\ F = (W_1F^+ + W_2F^-)/(W_1 + W_2)$ (13)

3. Araştırma Sonuçları ve Tartışma

Çalışmamızda ön işlem aşamalarından geçirilen ve SMOTE örnekleme metodunun kullanılması ile dengeli hale getirilen Covid-19 veri kümesinin YSA ile sınıflandırılması sonucu elde edilen azınlık sınıfı (hastalığı pozitif olanlar) ve çoğunluk sınıfı (hastalık teşhisi negatif olanlar) örneklemlerin doğruluk, P, R ve F-ölçüm değerleri Tablo 2-6 ile verilmiştir. Deneysel çalışmamız sonucunda veri kümesinde bulunan örneklemlerin laboratuvar testlerine bakarak, SARS-Cov-2 test sonucu negatif ya da pozitif tahmin edilmiştir.

Daha önce belirtildiği gibi ön işlem aşamasından geçirildikten sonra elde edilen veri kümesinde pozitif teşhisi konulmuş hasta sayıları, negatif teşhis konulmuş hasta sayılarından daha az sayıdadırlar. Bu nedenle ele aldığımız problemde pozitif sınıf azınlık sınıftır. Pozitif sonuçlanan vakalar hem hasta hemde hastanelerin yoğunluğu açısından büyük önem taşıdığından, çalışmamızda pozitif vakaları yani gerçekte hasta olanlar tespit edilmeye çalışılmıştır. Tablo 2'de görüldüğü gibi SMOTE ile dengelenen pozitif sınıfa ait örneklemler, orijinal veri kümesi örneklemlere daha başarılı sonuçlar elde edilmiştir. Bu sonuçlar Tablo 2'de P, R ve F-ölçüm değerleriyle verilmiştir.

Tablo 2. SARS-Cov-2 Test Sonucu Pozitif Sınıfa Ait Başarım Değerleri

Örnekleme Yöntemi	**P**	**R**	**F**
Orijinal	0.50	0.47	0.48
SMOTE	0.62	0.76	0.68

Çoğunluk sınıfına ait örneklemler, tüm veri kümesinde yüksek bir dağılıma sahip olduğundan herhangi bir dengelemeye tabi tutulmadan başarılı bir şekilde tahmin edilmiştir. Ancak çalışmada uyguladığımız SMOTE yöntemi çoğunluk sınıfına ait başarım değerleri az da olsa arttırmıştır. Bu sonuçlar Tablo 3'te P, R ve F-ölçüm değerleriyle verilmiştir.

Tablo 3. SARS-Cov-2 Test Sonucu Negatif Sınıfa Ait Başarım Değerleri

Örnekleme Yöntemi	**P**	**R**	**F**
Orijinal	0.91	0.92	0.92
SMOTE	0.96	0.92	0.94

Çalışmamızda uygulanan yöntem, pozitif ve negatif sınıfların genel başarım sonucu olan Doğruluk değerini arttırmış olup, bu değerler Tablo 4'te gösterilmiştir.

Tablo 4. Pozitif ve Negatif Sınıfların Doğruluk Oranları

Örnekleme Yöntemi	**Doğruluk**
Orijinal	0.86
SMOTE	0.90

Tablo 2-4'te verilen ölçümlere ek olarak, uygulanan yöntemin genel performansı için, pozitif ve negatif sınıfların başarım değerlerinin hem aritmetik ortalama hem de veri kümesindeki her sınıfın oranını (ağırlığını) dikkate alan ağırlıklı ortalama değerleri elde edilmiştir. Aritmetik ve ağırlıklı ortalama değerleri Tablo 5-6'da gösterilmiştir.

Tablo 5. Pozitif ve Negatif Sınıfların Aritmetik Ortalama Başarım Değerleri

Örnekleme Yöntemi	**P**	**R**	**F**
Orijinal	0.71	0.70	0.70
SMOTE	0.79	0.84	0.81

Tablo 6. Pozitif ve Negatif Sınıfların Ağırlıklı Ortalama Başarım Değerleri

Örnekleme Yöntemi	**P**	**R**	**F**
Orijinal	0.86	0.86	0.86
SMOTE	0.91	0.90	0.91

4. Sonuç

Çalışma kapsamında, orijinal tıbbi veri kümesi SMOTE tabanlı örnekleme yöntemiyle dengeli hale getirilerek, azınlık örnekler suni olarak çoğaltılıp YSA algoritmasıyla sınıflandırılarak orijinal veri kümesi ile kıyaslanmıştır. Yapılan deneysel çalışma sonucunda, veri kümesinin dengesiz olan ilk durumuna kıyasla SMOTE örnekleme yönteminim ürettiği Doğruluk, P ve R değerlerinin harmonik ortalaması olan F-ölçüm değerlerinin her iki sınıfta da yükseldiği ve YSA algoritmasının başarım değerlerini arttırdığı görülmüştür. Tablo 2-4'da verilen sonuçlara göre uygulanan yöntemlerin, sınıflandırma doğruluğunu arttırdığı açıktır. Elde edilen doğruluk oranı 0.86'dan 0.90'a, pozitif sınıfa ait F-ölçüm değeri 0.48'den 0.68'e ve negatif sınıfa ait F-ölçüm değeri ise 0.92'den 0.94'e yükseltilmiştir. Tablo 5-6'da verilen pozitif ve negatif sınıfların aritmetik ve ağırlıklı ortalama başarım değerlerinin orijinal veri kümesine göre uygulanan SMOTE yöntemiyle YSA'nın performansında genelde artış olduğu görünmektedir.

Çalışmamızın genel sonucunda, elde edilen 0.90 doğruluk oranına göre hastanelerin acil servisine yapılan şüpheli bir Covid-19 vaka başvurusunda, hastanelerde yaygın olarak toplanan laboratuvar test sonuçlarına dayanarak, SARS-Cov-2 sonucu tespit edilebildiğini göstermektedir.

Kaynakça

Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz. "Special issue on learning from imbalanced data sets." ACM Sigkdd Explorations Newsletter 6.1 (2004): 1-6.

Oğul, H. A., & Güran, A. (2019, September). "Imbalanced Dataset Problem in Sentiment Analysis." In 2019 4th International Conference on Computer Science and Engineering (UBMK) (pp. 313-317). IEEE.

Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

Berner E, "Clinical Desicion Support Systems",Department of Health Services Administration University of Alabama at Birmingham, USA, Springer, ISBN -10: 0-387-33914-0, 2006.

Kumar, R., Arora, R., Bansal, V., Sahayasheela, V. J., Buckchash, H., Imran, J., ... & Raman, B. (2020). Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers. medRxiv.

De Moraes Batista, A. F., Miraglia, J. L., Donato, T. H. R., & Chiavegatto Filho, A. D. P. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv.

Schwab, P., Schütte, A. D., Dietz, B., & Bauer, S. (2020). predCOVID-19: A Systematic Study of Clinical Predictive Models for Coronavirus Disease 2019. arXiv preprint arXiv:2005.08302.

AbuSharekh, E. K., & Abu-Naser, S. S. (2018). Diagnosis of hepatitis virus using artificial neural network.

Shuja, M., Mittal, S., & Zaman, M. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. In Advances in Computing and Intelligent Systems (pp. 195-211). Springer, Singapore.

Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016, May). Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS) (pp. 225-228). IEEE.

I. Tomek, "Two modifications of CNN", IEEE Transactions on Systems Man and Cybernetics, vol. 6, pp. 769-772, 1976

Kaggle veri bilimci ve makine öğrenme çevrimiçi topluluğu, https://www.kaggle.com/dataset/e626783d4672f182e7870b1bbe75fae66bdfb232289da0a61f08c2ceb01cab01?select=dataset.xlsx,04.05.2020.

T. M. Mitchell, Machine Learning.(2009) http://profsite.um.ac.ir/~monsefi/machine-learning/pdf/Machine-Learning-Tom-Mitchell.pdf,15.05.2020.

Erkaymaz, H., 2014. Elektrookulogram (EOG) Sinyallerinin İncelenmesi ve Yapay Zaka Teknikleri ile Modellenmesi. Doktora Tezi. Bülent Ecevit Üniversitesi Fen Bilimleri Enstitüsü. Zonguldak. 126s.

Du, K. L., Lai, A.K.Y., Cheng, K.K.M., Swamy, M.N.S., 2002. Neural Methods for Antenna Array Signal Processing: A Review, Elsevier Signal Processing 82 : 547-561

Gürsoy, Mİ., 2018. Alçak Gerilim Şebekeleri İçin Durağan ve Durağan Olmayan Güç Kalitesi Olaylarının Tespiti ve Sınıflandırılması için Yeni Bir Yaklaşım. Doktora Tezi. Kahramanmaraş Sütçü İmam Üniversitesi Fen Bilimleri Enstitüsü. Kahramanmaraş. 114s.

McCulloch, W.S. ve PITTS, W., A Logical Cafeulus of the Ideas Immane nt in Nervous Activity, Bulletin of Mathematical Biophysics, volume 5 (1943).

Öztemel, E., 2006. Yapay Sinir Ağları. Papatya Yayıncılık Eğitim. İstanbul, Türkiye. 231s.