

Bagging Sınıflandırıcısı ile Kalp Krizi Riski Belirleme

Heart Attack Risk Detection Using Bagging Classifier

Faruk BULUT

Bilgisayar Mühendisliği Bölümü, Gediz Üniversitesi, İzmir, Türkiye
faruk.bulut@ikc.edu.tr

Özetçe—Kalp ve Damar hastalıkları dünyada en çok ölüme neden olan hastalık gurubudur. Çalışmamız, Makine Öğrenmesi kolektif sınıflandırma algoritmalarından Bagging yöntemini kullanarak bireyde kalp krizi risk durumunu belirlemeyi amaçlamaktadır. Devlet hastanelerinden kalp krizi geçirmiş hastalara ait tıbbi veriler resmi izinlerle elde edilmiştir. Bu veriler sınıflandırıcı algoritmalar için veri setine dönüştürülmüştür. Yapılan uygulamalarda yeni ve başarı oranı yüksek sınıflandırıcılar ile bireyde oluşabilecek kalp krizi riski erkenden tespit edilmeye çalışılmıştır. Ayrıca çapraz geçerleme işlemlerinde önerilen modelin diğer tekil sınıflandırıcılara göre yüksek performans gösterdiği tespit edilmiştir. Tasarlanan Klinik Karar Destek Sistemi (KKDS) ile bu hastalığa karşı önceden önlem alınabilmesi sağlanmaktadır.

Anahtar Kelimeler — Kalp krizi tespiti, Bagging, sınıflandırma, KKDS.

Abstract—Cardiovascular diseases in the world are the most common cause of death. Our study aims to predict the rate of heart attack risk for individuals using the Bagging Method, an ensemble Machine Learning classification algorithm. For this reason, a questionnaire has been prepared to collect the relevant data. After obtaining the official permissions, the questionnaires are applied to the patients who have had a heart attack. By this way a predefined dataset is created to be used in the classification algorithms. In the applications, heart attack risk can be detected for an individual by using powerful ensemble classifiers. Additionally, in cross validation process the proposed model shows a high-performance in regression. Therefore, this suggested Clinical Decision Support System (CDSS) enables to take some precautions before a heart attack.

Keywords — Heart attack detection, Bagging, classification, CDSS.

I. GİRİŞ

Kalp krizi olarak tabir edilen akut miyokardiyal enfarktüs (AMI veya MI), kalbe ulaşan damarlardan birinde tıkanıklığa ve bunun sonucunda kalp dokusunda oluşan hasara denir. Dünya Sağlık Örgütü (WHO) istatistiklerine göre dünyada ölüm nedenleri arasında birinci sırada olan Kalp ve Damar Hastalıkları (KDH) tüm ölümlerin %30'unu oluştururken Türkiye'de bu oran % 47 civarındadır. Türkiye'de yetişkin nüfusun yarıya yakını kalp damar hastalıkları riski altındadır. Bununla birlikte kalp krizinden ölen erkeklerin oranı daha fazladır [1]. Kalp damar hastalıklarının başlıca sebepleri hipertansiyon, yüksek kötü kolesterol (LDL ve trigliserit), düşük iyi kolesterol (HDL), damarın tıkanıklığı, damar hastalıkları, sigara tiryakiliği ve alkol bağımlılığıdır [2].

Günümüzde kalp hastalıkları tıbbi tahlillerle sadece doktorlar tarafından teşhis edilmektedir. Doktorların üzerindeki bu yükün bir kısmı Klinik Karar Destek Sistemleri (KKDS) ile kaldırılabilir. Yapay tahmin yöntemleri %100 doğru sonuçlar vermese de sağlık çalışanlarına olan desteği bu türdeki çalışmaları önemli kılmaktadır. Son yıllarda, veri madenciliği, şekil tanıma ve yapay öğrenme, tıp bilimi dâhil olmak üzere birçok alana uygulanmıştır. Sağlık sektöründe doktorlar tarafından yapılan teşhis ve tedavilerde daha doğru sonuçlar elde etmek, insan kaynaklı hataları engellemek ve doktorun kararına yardımcı olmak amacıyla KKDS modelleri kullanılmaktadır.

Bu alanda yapılan bir takım bilimsel çalışmalar literatürde yer almıştır. Bir çalışmada Fuzzy sistemler ile kalp krizi riski tespit edilmeye çalışılmış [3], başka bir çalışmada [4] ise veri madenciliği yöntemleri kullanılarak yapılan çalışmaların analizi yapılmıştır. Bu alanda doğrusal olmayan optimizasyon metotları da kullanılarak çalışma zenginleştirilmiştir.

Chadha ve arkadaşları tarafından 2016 yılında yapılan kapsamlı bir çalışmada [5] o ana dek kalp krizi seziimi ve tespiti konusunda yapılmış çalışmalar özetlenerek incelenmiş, farklılıkları ortaya konmuştur. Aynı çalışma içerisinde sağlık kuruluşlarının elinde bulunan çok sayıdaki hasta kayıtlarının efektif bir şekilde kullanılabilmesi için veri madenciliğinde bulunan birçok tekniğin nasıl uygulanması gerektiğinden bahsedilmiştir. Daha önce yapılan çalışmalardaki teknikler bir araya getirilip her bir tekniğin avantaj ve dezavantajları listelenmiştir. Ayrıca sınıflandırıcıların da karşılaştırılmasının yapıldığı çalışmada Naive Bayes ve Kara Ağaçlarına göre Çok Katmanlı Algılayıcıların daha yüksek performans gösterdiği belirtilmiştir.

Bu çalışmada bireyde kalp krizi riskinin varlığı önerilen bir KKDS ile tespit edilmesi amaçlanmıştır. Bu nedenle çalışmada kullanılacak olan verilerin hastanelerde kalp rahatsızlığı geçirmiş hastalardan elde edilmesi için gerekli olan kuruluşlara resmi olarak müracaatlar yapılmıştır. Alınan resmi izinlerle üniversite ve devlet hastanelerinde tedavi gören ve kalp krizi geçirmiş haftalara bir takım anketler uygulanmıştır. Elde edilen veriler sayısal değerlere dönüştürülerek bir veri seti oluşturulmuştur. Oluşturulan veri seti yeni ve gelişmiş kolektif öğrenme algoritmalarına aktararak sistem eğitilmiştir. Yapılan deneysel çalışmalarda ve çapraz geçirme işlemlerinde doğru ve güvenilir sonuçların elde edildiği görülmüştür.

II. MODEL OLUŞTURMA VE VERİ TOPLAMA

KKDS modeli için en uygun verilerin toplanması amacıyla özgün bir veri toplama anketi oluşturulmuştur. Bu yüzden kalp krizine neden olan faktörler ankette 6 ana gruba ayrılmıştır. Bunlar bireydeki demografik özellikler, yaşamsal alışkanlıklar, sağlık durumu, spor ve hareketli yaşam, beslenme alışkanlıkları ve psikolojik durumdur.

Demografik özellikler bölümünde bireye ait kilo, boy, yaş, cinsiyet ve medeni durum istenmektedir. Boy, kilo ve cinsiyeti elde edildiği için vücut kitle indeksi (VKİ) ile obezite tespit edilebilir. Obezite kalp krizi için önemli bir ölçüttür.

Yaşamsal alışkanlıklar bölümünde bireyin sigara kullanması, uyku düzeni ve süresi, alkol tüketim sıklığı gibi sorular ile kişinin yaşamsal alışkanlıkları incelenmiştir. Bu sorular bireyin kalp krizi geçirmesini tetikleyen en önemli faktörlerdendir [2].

Sağlık durumu bölümünde nabız değeri, menopoz durumu, doğum kontrol hapı kullanımı durumu, ailede kalp hastalığı geçiren kişinin varlığı, iyi huylu ve kötü huylu kolesterol seviyesi, diyabet durumu ve genel tansiyon değerleri sorulmuştur. Araştırmalar, nabızdaki değişkenliğin, kolesterol ve tansiyondaki anormalliğin bir atardamar hastalığının nedeni olabileceğini ve bunun da kalp krizini tetikleyici bir etken olabileceğini göstermektedir [2]. Sigara bağımlılığı, yüksek kan basıncı ve yüksek kolesterol seviyeleri kalp damar hastalarının %80'inde görülmektedir. Bu risk faktörleri, erken yaşta

koroner kalp hastalığı gelişmesinde büyük rol oynamaktadır.

Spor ve hareketlilik bölümündeki sorular, kişinin hareketliliğine göre kalp hastalığına yatkınlığını incelemek içindir. Yapılan çalışmalarda bireyin düzenli olarak spor yapması ve sıkça yürümesi kalp krizi riskini düşürdüğü gözlemlenmiştir [2], [6]. Bu bölümde bireyin günlük olarak ne kadar yürüdüğü ve spor yaptığı, ne kadar oturduğu sorulmaktadır.

Beslenme alışkanlıkları bölümünde tüketilen yumurta, kırmızı et, sakatat, balık, sebze, meyve, kızartma, kuruyemiş, yağ türü ve yapılan diyet türü sorulmaktadır. Haftada kaç öğün balık tüketiyorsunuz?" sorusu, kalp krizi riskini azaltmaya yarayan ω -3 yağ asitlerinin hangi periyotta alındığını tespit etmek amacıyla sorulmuştur. "Yemeklerde en çok ne tür yağ kullanıyorsunuz?" sorusu, margarinler, süt ve vitaminlerle zenginleştirilmelerine karşın içerdikleri trans yağ asidi miktarının bilinmesinin önemine bağlı olarak sorulmuştur. Trans çoklu doymamış yağ asitleri, LDL kolesterolünü artırdığı ve HDL kolesterolünü azalttığı için kalp damar hastalıklarının oluşumunda önemli bir faktördür [7]. Ayrıca "Besinleri pişirirken kızartma ve kavurma yöntemlerini kullanıyor musunuz?" sorusu ile besinlerin kızartma yerine ızgara, buhar, fırınlama, haşlama şeklinde pişirilmesinin kalp krizini önleyebileceği ifade edilmektedir [8].

Kalp hastalıkları psikososyal durum ve psikiyatrik patoloji ile iki yönlü bir ilişkiye sahiptir. Ani ölüm ya da hastalığın getirdiği tehdit yaygın olarak anksiyete, disfori ve depresyona neden olur. Bunun tersi de doğrudur. Buna ek olarak, aynı belirtiler birincil olarak psikiyatrik ya da kardiyolojik durumların habercisi olabilir. Kalp rahatsızlıklarını etkileyen psikolojik faktörler bazı sorularla belirlenmeye çalışılmıştır. Psikolojik bir rahatsızlığınız oldu mu, sizi çok üzen bir olay yaşadınız mı, ailevi sorunlarınız var mı, sık sık öfkelenir misiniz, maddi durumunuzdan memnun musunuz ve size mobbing uygulandı mı gibi sorular sorulmuştur.

Yapılan müracaatlar neticesinde alınan resmi izinler ile Ege Üniversitesi Hastanesi ve Manisa Merkezefendi Devlet Hastanesi bünyesindeki Kardiyoloji bölümünde tedavi gören toplamda 62 hastaya hazırlanan anketler uygulanmıştır. Kalp krizi geçirmemiş değişik yaş ve cinsiyet grubundaki 47 sağlıklı bireye de aynı anket uygulanmıştır. Kalp krizi geçirmiş ve geçirmemiş kişiler veri setine nominal etiketleme ile yerleştirilmiştir.

Anketteki sorulara verilen cevaplar sayısal değerler şeklinde eğitim setine aktarılmıştır. Sorulara ait cevap "evet" ise veri setine 1 olarak; "hayır" ise 0 şeklinde aktarılmıştır. Ayrıca çoktan seçmeli sorular katar şeklinde kaydedilmiş, bazı sorulardaki a, b, c ve d şıkları ise sırasıyla 1, 2, 3 ve 4 şeklinde sayısal değerler şeklinde kaydedilmiştir. Proje üzerinde çalışma yapılırken bazı faktörlerin kalp krizine olan etkisinin az olduğu gözlemlenmiştir. Ayrıca anket çalışmasına katılan hastalar,

bazı soruları da cevaplamamıştır. Bu nedenlerden ötürü bazı sorular eğitim seti oluşturulurken listeden çıkarılmıştır.

Anket soruları hasta ile yüz yüze görüşülerek yapıldığı için güvenilir bir yapıdadır ve geçerlilik analizi testinin yapılmasına gerek yoktur. Doğru, hassas ve net sonuçlar veren güvenilir bir çalışma amaçladığı için bu zorluklara katlanılmıştır.

III. BAGGING YÖNTEMİ

Denetimli öğrenme, temel olarak sınıflandırma işlemiyle yapılır. Denetimli öğrenme, etiketi bilinen örneklerden yararlanarak etiketsiz verileri etiketleme yöntemidir. Bu işlem için önceden tanımlanan ve etiketli verilerden oluşan bir eğitim ölçüt kümesine ihtiyaç vardır. Sınıflandırma işlemi için yapay öğrenici kendini, hazırlanmış olan eğitim seti ile eğitir.

Bu çalışmada temel öğrenici olarak Karar Ağaçları tercih edilmiştir. Genelde kolektif sınıflandırma yöntemlerinde temel öğrenici karar ağaçlarıdır. Karar ağaçlarının oluşturulması ve kullanımı kolaydır, verdikleri kararlar hızlıdır ve ayrıca yaygın bir kullanım alanına sahiptir. Karar ağacı algoritmasında, bir ağaç yapısı oluşturulur. Ağacın yaprakları üzerinde sınıflar ve gövdeden yapraklara giden düğümler üzerinde de karşılaştırmalar yer almaktadır. Ağacın oluşturulması sırasında, veri setindeki her bir özneliliğin entropi değerlerine göre bölümlenme yapılır. Bu işlem, özyinelemeli bir şekilde tekrarlanır ve tekrarlama işlemi önemini yitirene kadar sürer. ID3, C4.5 ve CART algoritmaları en bilinenlerindendir.

Tekil sınıflandırıcılara göre yüksek performans sunan kolektif metotlar değişik alanlarda sıklıkla kullanılmaktadır. Bagging, Boosting, Oyama, Rastgele Ormanlar, Rotasyon Ormanları, Uzman Karışımları bunlardan bir kaçıdır [9]. Bagging (Bootstrapping Aggregating) metodu, var olan bir eğitim setinden yeni eğitim setleri oluşturarak temel sınıflandırıcıyı tekraren eğiten bir yöntemdir. Her bir veri seti için eğitilen temel öğrenicinin belirli bir test örneği üzerindeki kararları ortalama alınarak hesaplanır. Bagging’de amaç yeni veri setleri türeterek farklılıkları (diversity) oluşturmak ve bu sayede toplam sınıflandırma başarısını artırmaktır.

Bu yöntemde eğitim setinden yaklaşık olarak %63,2 kadar orijinal örnek rastgele alınır ve alınan örneklerden bazıları çoğaltılarak (resampling) eğitim seti %100’e tamamlanır. Bu yöntemle birbirinden farklı bir miktar eğitim seti elde edilir. Her eğitim seti aynı temel öğreniciye uygulanır ve alınan kararlar ağırlıklı oylama yöntemiyle birleştirilir [9]. Eğitim setinden %63,2 kadar örnek seçilmesindeki neden şu formülle açıklanır:

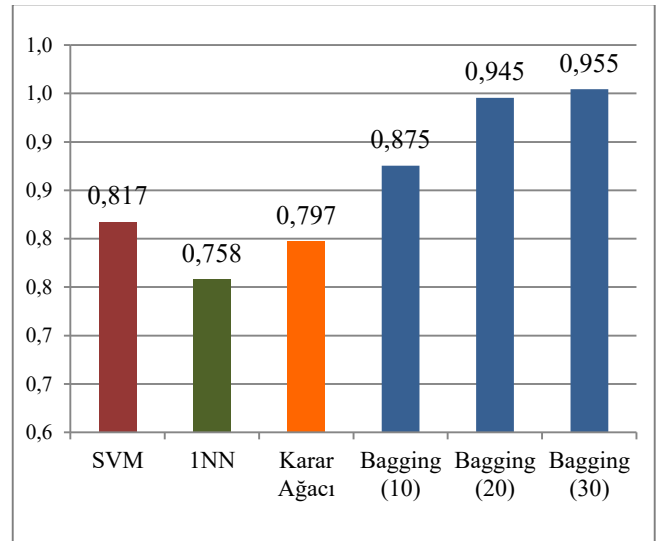
$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0,368 \quad (1)$$

Burada n işlem sayısını veya eleman sayısını gösterir. n değeri sonsuza giderken doğal logaritma tabanı olan e sayısının tersi elde edilmiş olur. n sonsuza giderken bu yöntemle orijinal eğitim kümesinden yaklaşık olarak %36’sı büyük ihtimalle hiçbir zaman seçilmemiş olacaktır. Bagging yönteminde seçilmeyenlerin seçilmesini sağlamak için eğitim setinden %63,2 kadar örnek $(1-0,368=0.632)$ rastgele seçilir ve yeni bir eğitim seti oluşturulur.

IV. DENEYSEL SONUÇLAR

Sınıflandırıcıların sınıflandırma performanslarını test etmek amacıyla çapraz geçerleme işlemi kullanılmıştır. Literatürde çeşitli çapraz geçerleme yöntemleri vardır. Birini devre dışı bırak (Leave One Out Cross Validation, LOO-CV), 10-Fold ve 5x2 çapraz geçerleme bunlardan bazılarıdır. Etiketli tıbbi verilerin elde edilmesi zordur ve bu nedenle veri setinde genellikle fazla sayıda örnek yoktur. Seyrek veri setleri için en uygun çapraz geçerleme yöntemi LOO-CV olarak literatürde tavsiye edilmektedir [10]. LOO-CV, veri seti üzerindeki sınıflandırıcının başarısını diğer bir deyişle hata oranını hesaplamak için kullanılır. LOO-CV yönteminde veri setinden veriler teker teker alınıp test noktası olarak belirlenir. Daha sonra geride kalan noktalar ise eğitim seti olarak belirlenir. Test noktaları tek tek eğitim setine sorulur. Eğitim setinin verdiği cevapların doğruluğu sayılır ve genel başarı ortalaması bulunur [11].

MATLAB ortamında bazı tekil sınıflandırıcılar ile Karar Ağacı ve Bagging yöntemi hazırlanan veri seti üzerinde denenmiştir. Sınıflandırma başarıları Şekil 1’de görüldüğü gibidir.



Şekil 1. Başarım değerlendirme

Bilinen ve yaygın olarak kullanılan iki farklı tekil sınıflandırıcı da çalışmaya ilave edilerek performans karşılaştırması yapılmıştır. Destek Vektör Makineleri (SVM) ve En Yakın Komşu Sınıflandırıcısı (1NN) aynı veri seti üzerinde koşturulmuş ve %81,7'lik ve %75,8'lik doğruluk oranı elde etmişlerdir. Bagging'de temel öğrenici olarak kullanılan Karar Ağacı ise tek başına %79,7'lik bir başarı elde edebilmiştir. Görüldüğü üzere tekil sınıflandırıcıların başarı oranları birbirine yakın ve aynı zamanda kullanılan kolektif öğreniciye göre başarı oranı oldukça düşüktür.

Karar Ağaçları ile oluşturulan Bagging yönteminde ise daha yüksek doğruluk oranları elde edilmiştir. Oluşturulan 10, 20 ve 30 veri seti ile Bagging algoritması çalıştırılarak performans analizi yapılmıştır. 20 adet veri seti ile elde edilen başarı değerinden sonra 30 veri seti ile elde edilen başarı değeri kıyaslandığında fazla bir artışın olmadığı gözlemlenmiştir. 20 adet veri seti ile de yeterince farklılık elde edilmiş ve yüksek bir oranda sınıflandırma başarısı elde edilmiştir. Bu durum türetilen veri seti sayısının artması ile başarı oranının doğru orantılı bir şekilde artabileceği düşüncesini çürütmektedir.

V. İLERİ ÇALIŞMALAR VE DEĞERLENDİRME

Projenin geliştirilmesi adına yapılabilecek bazı çalışmalar olabilir. Kolektif öğrenme yöntemlerinde tekil öğrenici olarak genellikle Karar Ağaçları tercih edilir. Fakat destek vektör makineleri, örnek tabanlı öğreniciler, yapay sinir ağları gibi başka tekil öğreniciler de tercih edilebilir. Ayrıca bireyde kalp rahatsızlıklarını tetikleyen başka etmenler de incelenip ankete alınabilir. Kişilerin kan grupları, protein yapıları, genetik hastalıkları, kan ve idrar tahlilleri gibi bazı laboratuvar sonuçları da kalp krizini etkileyebilecek olası etkenler olduğu için ankete eklenebilir.

Daha fazla hastadan elde edilen verilerle sistemin kalp rahatsızlıkları ile ilgili muhtemel her durumuna karşı eğitilmesi sağlanmış olabilir. Bu amaçla veri zengini hastanelerin kayıtları kullanılabilir ve veri seti seyrek bir yapıdan yoğun bir yapıya dönüştürülmüş olur. Bu sayede kalp krizine neden olan hemen hemen tüm faktörler veri setine dâhil edilebileceği için daha doğru ve güvenilir tahminler yapılabilir. Ayrıca veri madenciliği algoritmaları kullanılarak bilgi keşfi gibi analizler de yapılabilir.

Tıp ve bilişim bilim dallarını kapsayan multidisipliner çalışmamızda, yüksek başarı oranı sağlayan kolektif sınıflandırıcılar ile kalp krizi riskinin belirlenmesine çalışılmıştır. Özgün, güvenilir ve gerçek bir eğitim seti oluşturulmuş ve kolektif öğreniciler kullanarak da çağın büyük bir sorununa karşı yapay bir klinik karar destek sistemi önerilmiştir. Oluşturulan bu erken uyarı sistemi sayesinde bireyin kalp krizi riskinin varlığı önceden tahmin edilebilmiştir. Elde edilen deneysel sonuçlara göre yüksek risk yüzdesine sahip kişiler haberdar edilip gerekli tedbirlerin alınması sağlanabilir.

Bu çalışmamızdaki amaç oluşturulmuş bir veri seti üzerinde, var olan ve bilinen sınıflandırma yöntemlerinin uygulanması ve performanslarının karşılaştırılması değildir. Amaç daha ileri ve detaylı bir klinik karar destek isteminin oluşturulması için ilk sınıflandırma örneğinin sunulmasıdır. Önerilen bu erken uyarı sistemi insan sağlığını ve hayatını korumaya yönelik bir çalışma olması nedeniyle daha da ayrı bir öneme sahiptir.

TEŞEKKÜRLER

Tıbbi verilerin elde edilmesinde resmi izni veren, destekleyen, hoş gören Manisa İl Sağlık Müdürlüğü'ne ve Ege Üniversitesi Rektörlüğü'ne teşekkür ederiz. Manisa Merkez Efendi Devlet Hastanesi Yöneticilerinden Uzm. Dr. Abdullah Korkut KAFESÇİLER'e, Baştabip Yardımcısı Uzm. Dr. Emine KOCAAYAN'a ve Ege Üniversitesi personeline desteklerinden ötürü şükranlarımızı sunarız. Verilerin toplanmasında ve sayısallaştırılmasına olağanüstü bir gayret sergileyen sevgili Mustafa Ege ŞEKER ve Yusuf Miraç UYAR'a da özellikle takdirlerimizi ve teşekkürlerimizi sunarız.

KAYNAKÇA

- [1] Haney, M. Ö. ve Bahar. Z., “Çocuk Kalp Sağlığını Geliştirme Tutum Ölçeği'nin Geçerlik ve Güvenirliği” Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi, cilt 2, no. 7, 2014.
- [2] Güleç, S., “Kalp damar hastalıklarında global risk ve hedefler”, Türk Kardiyol Dern Arş, 2(37): 1-10, 2009.
- [3] ANOOJ, P. K. “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules”, Journal of King Saud University-Computer and Information Sciences, 24(1): 27-40, 2012.
- [4] THANIGAIVEL, R.; KUMAR, K. Ramesh. “Review on Heart Disease Prediction System using Data Mining Techniques”, Asian Journal of Computer Science and Technology (AJCST) Vol, 3: 68-74, 2015.
- [5] CHADHA, Ritika, et al. “Application of Data Mining Techniques on Heart Disease Prediction: A Survey”, In: Emerging Research in Computing, Information, Communication and Applications. Springer India, 413-426, 2016.
- [6] Öztürk, S. “Kardiyovasküler Risk Faktörü Olarak Dislipidilere Yaklaşım” Abant Medical Journal, 2(1): 89-93, 2012.
- [7] Çakmakçı, S. ve Kahyaoğlu, D., “Yağ asitlerinin sağlık ve beslenme üzerine etkileri”, Türk Bilimsel Derlemeler Dergisi, 5(2): 133-137, 2012.
- [8] Nehir, S. ve Çam, O., «Miyokard İnfarktüsü Geçiren Hastalarda Psikososyal Sağlık ve Hastalık Uyumu», Ege Üniversitesi Hemşirelik Yüksek Okulu Dergisi. 26 (3): 73-84, 2010.
- [9] Zhou Zhi-Hua., Ensemble Methods: Foundations and Algorithms, CRC Press, ISBN: 978-1439830031, 2012.
- [10] E. Alpaydın, Yapay Öğrenme, s. 417, İstanbul: Boğaziçi Üniversitesi Yayınevi, ISBN: 9786054238491, 2010.
- [11] REFAELZADEH, Payam; TANG, Lei; LIU, Huan. Cross-validation. In: Encyclopedia of database systems. Springer US, 532-538, 2009.