Journal of the Faculty of Engineering and Architecture of Gazi University 32:1 (2017) 65-76

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Örnek tabanlı sınıflandırıcı topluluklarıyla yeni bir klinik karar destek sistemi

Faruk Bulut*
İstanbul Rumeli Üniversitesi, Mühendislik - Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Silivri, 34570, İstanbul

Ö N E Ç I K A N L A R

- Yeni bir sınıflandırma yönteminin farklı bir alana uygulanması
- Makine öğrenmesi ile bir tahmin modelinin geliştirilmesi
- Yeni bir klinik karar destek sistem modeli

Makale Bilgileri
Geliş: 14.10.2015
Kabul: 01.11.2016

DOI:
10.17341/gazimmfd.300595

Anahtar Kelimeler:
Makine öğrenmesi, kolektif öğrenme, sınıflandırma

ÖZET

Çocukluk yıllarındaki beslenme ve yaşam alışkanlıkları ileri yaşlarda ortaya çıkabilecek obezite hastalığının nedenini oluşturur. Bu çalışma çocuklarda obeziteye yakalanma riskini hesaplayan bir erken uyarı sisteminin geliştirilmesi üzerinedir. Makine öğrenmesi kolektif öğrenme algoritmaları kullanılarak yapay ve özgün bir klinik karar destek sistemi (KKDS) geliştirilmiştir. Obeziteye neden olan faktörler hazırlanan anket içerisine yerleştirmiştir. Devlet hastanelerinden ve okullarından alınan resmi izinlerle anketler çocuklara uygulanmış ve elde edilen verilerle güvenilir bir eğitim seti oluşturulmuştur. k En Yakın Komşuluk algoritmasının geliştirilmiş versiyonları Oylama, Bagging, Boosting ve Rastsal Altuzaylar yöntemlerinde tekil öğrenici olarak kullanılmıştır. Eğitim seti üzerinde yapılan öğrenme ve çapraz geçerleme işlemlerinde algoritmalara ait yüksek doğruluk oranları elde edilmiş ve en başarılı yöntemin 0,839'lık MCC (Matthews Correlation Coefficient) değeriyle Rastsal Altuzaylar olduğu görülmüştür. Çağın önemli bir sorununa karşı önerilen bu model sayesinde, ileri yaşlarda oluşabilecek obezite riski önceden tespit edilebilmektedir. Ayrıca ilgili kişiler tarafından gerekli önlemlerin zamanında alınabilmesi sağlanmaktadır.

A new clinical decision support system with instance based ensemble classifiers

H I G H L I G H T S

- A new classification technique applied to a different field
- A new developed prediction model using Machine Learning
- A new model of clinical decision support system

Article Info
Received: 14.10.2015
Accepted: 01.11.2016

DOI:
10.17341/gazimmfd.300595

Keywords:
Machine learning, ensemble methods, classification

ABSTRACT

The main reason of obesity occurring in the future years is strongly related with the lifestyle and eating habits in childhood. This study focuses on developing an urgent precaution system which calculates the obesity risks. An original clinical decision support system (CDSS) has been developed by using Ensemble Classification methods in Machine Learnings. A questionnaire has been prepared and applied to the patients in the hospitals and to the elementary school students with official permissions in order to construct an original and reliable dataset. Extended versions of k Nearest Neighbors methods are used in Voting, Bagging, Boosting and Random Subspaces algorithms as base learners. During the experimental studies in the applications of cross validation procedures, successful results have been computed and Random Subspaces has been chosen as the most successful algorithm with 0.839 MCC (Matthews Correlation Coefficient) scores. With the help of the suggested model to a worldly wide health problem, the future probability of obesity risk for a child might be easily determined. Additionally, it has been enabled that some precautions can be taken by responsible people if there is a computed high risk for this child.

* Sorumlu Yazar/Corresponding author: faruk.bulut@rumeli.edu.tr / Tel: 444 2 917

1. GİRİŞ (INTRODUCTION)

Dünyadaki en yaygın hastalıklardan biri olan obezite, besinlerle alınan enerjinin harcanan enerjiden fazla olması ve fazla enerjinin vücutta yağ olarak depolanmasıdır. Dünya Sağlık Örgütü tarafından obezite, sağlığı bozacak ölçüde vücutta aşırı yağ birikmesi olarak tanımlanmıştır. Klinik olarak obeziteyi tanımlamak için ise vücut kitle indeksi (VKİ) ölçüt olarak kullanılır ve Eş. 1 ile hesaplanır. VKİ sonucuna göre bireyin obez olup olmadığı Tablo 1'deki verilere göre kategorilere ayrılır.

$$VKI = \frac{Ağırlık}{boy^2} \quad (1)$$

Tablo 1. VKİ'ye göre obezite (Obesity according to BMI)

Vücut Kitle İndeksi (VKİ)	Sonuç
18,5 kg/m²'den düşük	Zayıf
18,5-24,9 kg/m² arasında	Normal kilolu
25-29,9 kg/m² arasında	Fazla kilolu
30-39,9 kg/m² arasında	Obez *(şişman)*
40 kg/m²'den büyük	İleri derecede obez

Çocukluk çağında başlayan obezitenin erişkin dönemde de devam etmesi ve sağlık için risk oluşturması söz konusudur. Çocuklukta obezite, 5–6 yaş arası ve gelişim döneminde artış göstermektedir. Obez çocukların yaklaşık %35'i, obez ergenlerin ise %80'i erişkin yaşa ulaştıklarında da obez kalmaktadırlar. Diğer yandan erişkin yaşlarda görülen obezite vakalarının %30 kadarında obezite başlangıcının çocukluk çağlarına dayandığı bilinmektedir. Bu nedenle çocukluk ve ergenlik döneminde obeziteden korunma ve hastalığın tedavisi giderek önem kazanmaktadır. Gerekli tedbirler erken yaşta alınmadığı takdirde hastalığın ileri yaşlardaki tedavisi daha da zorlaşmaktadır [1]. Çocuklarda obezite teşhisi okullarda yapılan sağlık taramalarıyla ya da bilinçli ailelerin VKİ değeri yüksek olan çocuklar için sağlık kurumlarına müracaat etmeleriyle yapılabilmektedir. Bu süreçlerde çocuklardaki obezite riskinin erkenden tespit ve tedavi edilmesi daha da önemli hale gelmektedir. Çalışmamız erken teşhis koyan bir klinik karar destek sistemi (KKDS) üzerinedir. Erken tespit ile alınacak önlemler ve yapılacak tedaviler sonucunda çocuğun bir ömür boyu sağlıklı bir hayat geçirmesi amaçlanmıştır. Tıbbi alanda yapılan bilimsel çalışmalar ise obezitenin neden ve sonuçları üzerine yoğunlaşmıştır. Önerilen bu çalışma, retrospektif (meydana gelen olayların gerisine bakma) nedenlerden yola çıkarak bir yapay tahmin modeli öne sürmesi bakımından özgün bir yapıdadır. Ayrıca önerilen bu modelde, sınıflandırma ve tahmin başarısının artırılması için tekil sınıflandırıcıların yerine kolektif modeller (Ensemble Methods) tercih edilmiştir. Kullanılan yöntemler, üzerinde çalışılan konunun özellikleri göz önüne alınarak uygun bir şekilde özgünleştirilerek daha güvenilir sonuçların elde edilmesi sağlanmıştır. Çalışmanın geriye kalan kısmında beş bölüm daha vardır. İkinci bölümde obezite alanında yapılan bilimsel çalışmalara, üçüncü bölümde önerilen yapay karar destek modelinin açıklamasına ve veri setinin nasıl oluşturulduğuna, dördüncü bölümde kullanılan yapay öğrenme metotlarına, beşinci bölümde elde edilen deneysel sonuçlara ve analiz yöntemlerine yer verilmiştir. Son bölüm ise yapılabilecek ileri uygulamalara ve değerlendirmelere ayrılmıştır.

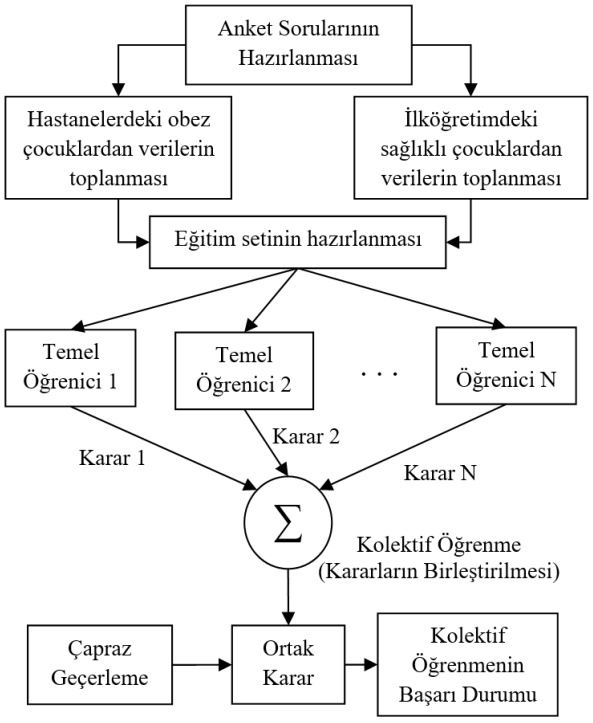
2. KAYNAK TARAMASI (SURVEY)

Bilişim alanındaki gelişmeler birçok alana uygulandığı gibi sağlık alanına da uygulanmaktadır. Çocuklarda obezite oluşumunun neden ve sonuçlarını tespit etmeye yönelik tıp alanında birçok çalışma bulunmaktadır. İstatistik bilim dalında da bir takım çalışmalar yapılarak çeşitli analizler ortaya konmuştur. Makine öğrenmesi ve veri madenciliği alanında konu ile alakalı yapılmış bir takım çalışmalar vardır. 2012 yılında yayımlanan bir çalışmada Adnan, Husain ve Rashid isimli araştırmacılar o yıla dek yapılmış temel çalışmaların neler olduğu göstermişlerdir. Çocuklarda obezite riskinin tespit edilmesi ile ilgili veri madenciliği sınıflandırma ve regresyon algoritmalarıyla yapılan çalışmalar listelenmiştir. Kullanılan yöntemlerin güçlü ve zayıf yönlerinden bahsedilmiştir. Ayrıca çalışmalarında NBTree (Naive Bayes Tree) sınıflandırıcı algoritmasını kullanarak 5-7 yaş arasındaki 200 kadar çocuktan elde ettikleri bilgilerle bir çalışma yapmışlardır. Çalışmalarında obeziteyi tetikleyen temel faktörler baz alınarak bir denetimli öğrenme çalışması yapılmıştır. Naive Bayes sınıflandırıcısı ile Genetik Algoritma (GA) kullanarak obezite alanında hibrid bir tahmin modeli önermişlerdir. Hastalığa neden olan 19 faktörün seçildiği çalışmalarında, GA yöntemi ile sınıflandırmada diğer algoritmalara göre %75 daha fazla doğruluk elde ettiklerini göstermişlerdir [2]. Heydari ve arkadaşları tarafından yapılan bir çalışmada Yapay Sinir Ağları (YSA) algoritması ile Logistic Regression yönteminin obezite tespiti üzerindeki performansları karşılaştırılmıştır [3]. 400 kadar askeriyede çalışan personelden alınan verilerle bir veri seti oluşturulmuştur. Oluşturulan veri seti kullanılarak her iki algoritmanın sınıflandırmadaki başarı oranları ROC (Receiver-Operating Characteristic) eğrileri kullanılarak analiz edilmeye çalışılmıştır. En yüksek doğruluk oranını veren parametrelerle oluşturulmuş YSA modelinin diğer yönteme göre daha yüksek performansta çalıştığı belirtilmiştir. Başka bir çalışmada [4] çocuklarda obezite hastalığına etki eden faktörlerin analizi için Fuzzy Signatures (FS) yöntemi [5] kullanılmış ve bazı tespitler yapılmıştır. İstatistiksel yöntemlerle bilinen sınıflandırıcı algoritmalarının obezite rahatsızlığının teşhisinde yetersiz kaldığından bahsedilmiş ve FS modeliyle hem sınıflandırma yapıldığı hem de en etkili risk faktörlerinin analiz edildiği vurgulanmıştır. 2015 yılında yayınlanan bir çalışmada ise Makine Öğrenmesi disiplininde kullanılan farklı yapıdaki algoritmalar yardımıyla çocuklarda obezite tespiti yapılmaya çalışılmıştır. RandomTree, RandomForest, J48, ID3, Naïve Bayes ve Bayes trained CDSS gibi altı farklı algoritma ile belirli veri setleri üzerinde performans testleri ve bazı kıyaslamalar yapılmıştır [6].

3. MODEL OLUŞTURMA VE VERİ TOPLAMA (CREATING MODEL AND DATA COLLECTION)

3.1. KKDS Modeli (CDSS Model)

Çalışmada önerilen KKDS (Klinik Karar Destek Sistemi) modelin genel işleyiş yapısı ve aşamaları Şekil 1'de görülmektedir.



Şekil 1. Önerilen KKDS Modeli (Suggested CDSS Model)

İlk aşamada KKDS'nin en iyi sonucu verebilmesi için obezite hastalığını tetikleyen en etken faktörler soru olarak ankete yerleştirilmiştir. Ardından çalışmanın güvenilirliğini sağlamak için anket soruları hastanelerdeki ve okullardaki öğrencilere uygulanmış ve gerekli veriler toplanmıştır. Elde edilen veriler karar destek sistemine aktarılarak çocuklardaki risk yüzdesi hesaplanmıştır. Ayrıca temel öğrenicilerin kararları değişik algoritmalarla birleştirilerek alınan komite kararının doğruluğu çapraz geçerleme işlemleriyle test edilmiştir. Önerilen KKDS modeli sayesinde obezite riski önceden tespit edilebilmektedir.

3.2. Anket Oluşturma ve Veri Seti Hazırlama (Creating Questionnaire and Preparing Dataset)

Bireyde obeziteye neden olan faktörler bu alanda yapılmış olan çeşitli bilimsel çalışmalar taranarak belirlenmiştir. Obeziteye neden olan birçok etmen vardır. Ancak bu çalışma için hazırlanan ankette, obeziteyi tetikleyen 6 ana faktöre yer verilmiştir [7]: 1. Kişinin temel özellikleri, 2. Psikolojik durumu, 3. Aile bilgileri, 4. Haftalık aktivite sıklığı, 5. Günlük öğün sıklığı, 6. Bazı besin gruplarını tüketim sıklığı. Çalışmada veri toplamak için kullanılan anket sorularının tamamı web adresinde [8] bulunmaktadır. Ancak burada belirtilen kategorilerin dışında obeziteye neden olan bazı başka faktörler de vardır. Örneğin sakatlık neticesi hareketsiz yaşantı ve kortizon tedavisi ile istemsiz kilo alımı gibi birçok etmen çalışmamızda ele alınmamıştır. Bu çalışmadaki amaç, daha önce de belirtildiği üzere günlük yaşamsal faaliyetlerin ve psikolojik durumun obezite üzerindeki etkilerini tespit eden bir erken uyarı sistemi önermektir. Altı ana faktör şu şekildedir:

3.2.1. A Grubu: Temel özellikler faktörü (Group A: Basic features factor)

Anketin bu bölümünde bireyin cinsiyeti, yaşı, kilosu ve boyu sorulmuştur. Bu değerler ile VKİ hesaplanarak tek bir özniteliğe (attribute) dönüştürülmüştür.

3.2.2. B Grubu: Psikolojik özellikler faktörü (Group B: Psychological features factor)

Psikolojik sorunların yemek yeme alışkanlığını tetiklediği bilimsel çalışmalarda yer almaktadır [7]. Bundan dolayı bu bölümde 3 adet soru sorularak bireyin psikolojik durumu değerlendirilmeye çalışılmıştır. Ankete katılan kişiye psikolojik bir rahatsızlığının olup olmadığı, üzücü bir olay yaşayıp yaşamadığı ve ailevi sorunlarının olup olmadığı sorulmaktadır.

3.2.3. C Grubu: Ailevi özellikler faktörü (Group C: Family features factor)

Bu bölümdeki sorular, çocukların genetik olarak obeziteye yatkın olup olmadığını incelemek amacıyla bilimsel çalışmalardan yararlanılarak [9] hazırlanmıştır. Aile bireylerinde obezite ve şeker hastalığının bulunması obeziteye etki etki eden genetik faktörlerden olduğu için ankete konulmuştur. Ayrıca ailede sigara içen birinin varlığı obeziteye yatkınlığı arttırdığı için ankete alınmıştır.

3.2.4. D Grubu: Aktivite sıklığı faktörü (Group D: Activity frequency factor)

Hareketsiz yaşam ve elektronik cihaz karşısında geçirilen fazla süre obezite riskini arttırmaktadır [7]. Bireyin aktivite sıklığını incelemek amacıyla sportif ve bedensel aktivitelere ayrılan süre ile bilgisayar ve televizyon karşısında geçirilen süre sorulmuştur.

3.2.5. E Grubu: Öğün alışkanlıkları faktörü (Group E: Meal habits factor)

Öğünlerin sıklık durumu, ara öğünlerde abur cubur yeme alışkanlığı, fastfood tarzı yiyeceklerin tüketim sıklığı obeziteye neden olan önemli faktörler arasındadır [9]. Bu nedenle bu bölümdeki sorular kişinin öğün alışkanlıklarını belirlemeye yöneliktir.

3.2.6. F Grubu: Besin gruplarının tüketim sıklığı faktörü (Group F: Consumption of food frequency factor)

Anketin bu bölümüne kalorisi yüksek besinler yerleştirilmiştir. Bu kategorideki besinler, tüketilme sıklığına göre obezite riskini arttırır [9]. Bu nedenle kişinin bazı besin gruplarını tüketme sıklığı bu bölümde incelenmiştir.

3.3. Veri Toplama ve Eğitim Seti Oluşturma (Collecting Data and Creating Dataset)

Anket çalışmaları, etik kurul kararı alınarak resmi izin belgeleriyle İstanbul Üniversitesi Çapa Eğitim ve Araştırma Hastanesi ile Şişli Etfal Eğitim ve Araştırma Hastanesi bünyesindeki Çocuk Sağlığı ve Anabilim dalı Büyüme-Gelişme ve Pediatrik Endokrinoloji bölümlerinde yapılmıştır. 7-15 yaş aralığındaki 61 kadar obez hasta ile yüz yüze yapılan görüşmelerden toplanan veriler eğitim setine eklenmiştir. Ayrıca Büyükçekmece İlçe Milli Eğitim Müdürlüğünden alınan diğer bir resmi izinle Adem Çelik İlk Okulu ve Orta Okulunda, 7–15 yaş aralığındaki sağlıklı 42 öğrenciye anketler uygulanarak eğitim seti güçlendirilmiştir. Böylece 103 veriden oluşan özgün bir eğitim seti hazır hale getirilmiştir. Anket çalışmaları yüz yüze yapıldığı için elde edilen veriler güvenilir bir yapıdadır. Çalışmada kullanılan ankette [8] de görüldüğü üzere 6 kategori altında toplam 29 soru vardır. Anket içerisine yerleştirilen soruların bir kısmına anket çalışmaları esnasında cevap alınamamıştır ya da alınan cevaplar azınlıkta kalmıştır. Bu nedenle bazı sorular anketten çıkarılmıştır. Her bir kategorideki sorular kendi grupları içerisinde puanlandırılarak sayısal değerlere dönüştürülmüş ve veri setine aktarılmıştır. Sorulara verilen cevapların nasıl sayısal değere dönüştürüldüğü ilgili web sitesinde [8] detaylı olarak açıklanmaktadır. Her bir kategori eğitim seti için tek bir öznitelik olarak düşünülmüştür. Bu sayede veri seti 6 boyutlu bir veri uzayı haline dönüşmüştür. Oluşturulan eğitim setindeki sayıların tamamı tam sayı şeklindedir. Tüm sayısal değerler normalize edilerek [0,1] aralığına çekilmiştir. Çünkü her bir boyutun aldığı sayısal değer aralıkları farklı farklıdır. Verilerdeki düzensiz dağılımları ve oluşabilecek sorunları önlemek için Min-Max yöntemi ile veriler Eş. 2 ile normalize edilmiştir:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Eş. 2'de x_n normalleştirilmiş değer, x gözlemlenen değer, x_{min} ve x_{max} da sırayla veri ilgili öznitelik grubu içerisinde bulunan en küçük ve en büyük değerlerdir. Obeziteye etki eden faktörler 6 ana grupta incelendiği için oluşturulan eğitim seti 6 boyutludur ve 103 adet örnek içermektedir. 61 obez 1 değeriyle, 42 sağlıklı birey de 0 değeriyle etiketlenmiştir. Doğal olarak veri seti üzerinde tekil öğreniciler ve kolektif sınıflandırıcılar ikili sınıflandırma (binary classification) yapmaktadır. Bilindiği üzere çok boyutlu veri setleri genel olarak seyrek bir yapıdadır. Bu durum sınıflandırma başarısı için olumsuz bir durumdur. Bu gibi durumlarda PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis) ve faktör analizi gibi algoritmalar kullanılarak boyut sayısında azaltma yapılmaktadır. Çalışmamızda kullanılan veri seti için bu tarz bir işleme gerek kalmamıştır.

4. YÖNTEMLER (METHODS)

Sınıflandırma yöntemlerinden biri olan kolektif öğrenme metotları tekil öğrenicilerin (base learner) öğrenme başarısını artırmak için kullanılmaktadır. Kolektif öğrenme yöntemlerinde temel öğrenici olarak genelde bilinen sınıflandırma algoritmaları tercih edilir. Destek Vektör Makineleri (Support Vector Machines), Çok Katmanlı Algılayıcılar (Multi-Layer Perceptron), Naive Bayes ve En yakın k komşulukları (k Nearest Neighbors, k-NN) kullanılan yöntemlerden bazılarıdır. Çalışmamızda temel öğrenici olarak basitliğinden, esnekliğinden, uygulanabilirliğinden, şeffaflığından ve bazı durumlarda başka algoritmalara göre yüksek doğruluk oranı vermesinden dolayı örnek tabanlı sınıflandırıcılar tercih edilmiştir. Ayrıca *k*-NN sınıflandırıcısının öğrenici topluluklarında kullanımı ile ilgili yapılmış bazı çalışmalar vardır [10]. Diğer çalışmalardan farklı olarak bu uygulamada literatürde önerilmiş performansı yüksek *k*-NN versiyonları kullanılmıştır.

4.1. Tekil Öğreniciler (Base Learners)

Çalışmamızda *k*-NN sınıflandırıcısının üç farklı türü temel öğrenici olarak kullanılmıştır:

1. *k*-NN
2. Uzaklık Ağırlıklı *k*-NN
3. En Yakın Küme Sınıflandırıcısı (1NC)

4.1.1. k-NN (k Nearest Neighbors)

Bu yöntemde test örneğine en yakın k adet örneğin sınıf etiketlerinin ortalaması ile sınıflandırma işlemi yapılır ve Eş. 3 ile bulunur:

$$f(x) = \text{argmax}_{c \in C} \sum_{i=1}^{k} w_i \delta\big(c, f(x_i)\big) \tag{3}$$

$f: R \to C$ olan $f(x)$ fonksiyonu x test noktasının sınıfını belirler. $f(x_i)$ ise veri uzayındaki x_i noktasının etiketini verir. Eğer a ve b birbirine eşit ise $\delta(a,b) = 1$ diğer türlü $\delta(a,b) = 0$ olmaktadır. C eğitim setinde bulunan sınıf etiketleri kümesidir. w_i ise her bir örneğin hesaplamadaki ağırlığıdır ve bu teknikte 1 olarak alınmıştır.

Veri uzayındaki örnekler arası uzaklığı hesaplamak için ise Öklid yöntemi kullanılmıştır. Öklid uzaklık ölçüsü kullanılarak D boyutlu (öznitelikli) uzayda x ve y noktaları arasındaki uzaklık Eş. 4 ile hesaplanır:

$$d(x,y) = \Big(\sum_{i=1}^{D} |a_i(x) - a_i(y)|^2 \Big)^{1/2} \tag{4}$$

x noktasının özellik vektörü $\langle a_1(x), a_2(x), a(x), \dots, a_D(x) \rangle$ şeklinde tanımlandığında $a_i(x)$, x noktasının i. inci boyutundaki sayısal değerine karşılık gelir. Her hangi bir x test noktasının sınıflandırılmasında tüm veri setinde bulunan örneklerin x noktasına olan uzaklıkları tek tek hesaplanır ve en yakın k tanesi hesaplamaya katılır. Hesaplama süresi açısından her bir test noktası için bu işlemin tekrar tekrar yapılması Tam Kapsamlı Arama (Exhaustive Search) olarak isimlendirilir ve istenmeyen bir durumdur. Bu arama yönteminin zaman karmaşıklığı oldukça yüksektir. N eleman sayısı olmak üzere

algoritmanın Big-O notasyonuna göre zaman karmaşıklığı O(k*D*N)'dir. Çalışmamızda kD-Tree (k Dimensional Tree) veri yapısı kullanılarak zaman karmaşıklığı ve hesaplama süresi azaltılmaya çalışılmıştır. BSP (Binary Space Partitioning) yöntemlerinden biri olan kD-Tree, ikili arama ağacı olan BST (Binary Search Tree) veri yapısının çok boyutlu türüdür. Big-O notasyonuna göre çok boyutlu kD ağacının kurulumu ile ilgili zaman karmaşıklığı $O(D * N * \log N)$'dir. kD ağacı bir kez oluşturulduktan sonra sınıflandırma işlemleri boyunca kullanılmaktadır ve bu da hesaplama süresini azaltmaktadır. Ağaçtaki bir örneği aramanın maliyeti $O(D * \log N)$'dir. Veri setinin kD ağacına aktarıldıktan sonra arama işlemlerinde kullanılması görüldüğü üzere oldukça hızlıdır. Bu nedenle tüm uygulamalarımızda kD ağacı kullanılan tüm tekniklerin temel veri yapısı olarak tercih edilmiştir [11].

4.1.2. Uzaklık Ağırlıklı k-NN (Distance Weighted k-NN)

$1/d^2$ uzaklık ağırlıklı *k*-NN yöntemi, örneklerin test örneğine olan uzaklıklarına bağlı olarak hesaplamadaki ağırlıklarını dikkate alarak sınıflandırma yapar. Bu teknikte yakındaki noktaların hesaplamadaki ağırlığı fazla; uzaktakilerin ağırlığı ise azdır [12]. Her bir kaydın hesaplamadaki ağırlığı (w) Eş. 5 ile bulunur:

$$w_i = \frac{1}{d_i^2} \tag{5}$$

d, x sorgu noktası ile eğitim setinde bulunan x_i noktası arasındaki uzaklıktır. x test örneğinin sınıfını bulan $f(x)$ fonksiyonu Eş. 6'da verilmektedir.

$$f(x) = \sum_{i=1}^{k} w_i f(x_i) / \sum_{i=1}^{k} w_i \tag{6}$$

Burada $f(x_i)$ hesaplamaya katılan k adet örnekten her birinin sınıf etiketi değerini göstermektedir. *k*-NN sınıflandırıcısında en uygun k değerinin kullanıcı tarafından belirlenmesinde bir takım zorluklar olduğu bir gerçektir. Fakat diğer sınıflandırıcıların da kullanıcılar tarafından belirlenen bir takım parametrelere gereksinim duyduğu bilinmektedir. Bu çalışmada k parametresi öğrenicilerin farklı kararlar vermesini sağlamak için kullanılan yöntemlerden biridir.

4.1.3. 1NC: En Yakın Küme Sınıflandırıcısı
(1NC: Nearest Cluster Classifier)

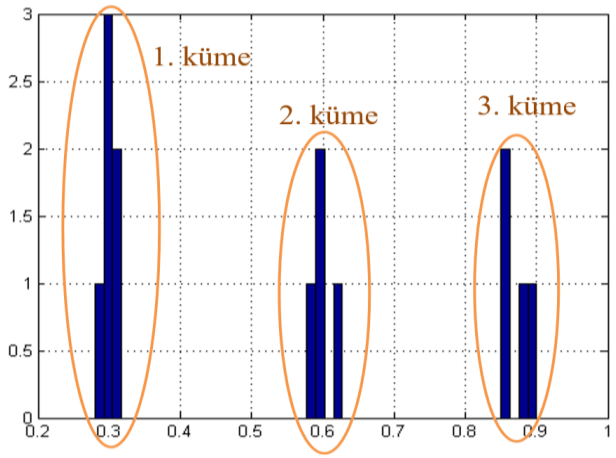
En uygun k parametresi, kullanıcı tarafından genellikle çapraz geçerleme yapılarak yani deneme-yanılma yöntemiyle belirlenmektedir. Bununla birlikte, bir veri setinde her bir test örneği için aynı k parametresinin kullanılması genel sınıflandırma başarısını olumsuz etkileyebilir. Örnek tabanlı sınıflandırıcılarda kümeleme yöntemiyle performans artırımı üzerine yapılan bir çalışmada [13] her bir test örneği için en uygun k değerini bulan bir yöntem sunulmuştur. Çalışmada her bir test örneği için en uygun k parametresini kümeleme yöntemiyle bulan ve bu sayede genel sınıflandırma başarısını artıran bir yöntem üzerinde çalışılmış ve başarılı sonuçlar elde edilmiştir. En yakın küme sınıflandırıcısı (One Nearest Cluster, 1NC) diye isimlendirilen bu öğrenici, test noktasına en yakın ilk kümedeki örneklerin tamamını hesaplamaya katmaktadır. 1NC tekniğinin işlem basamakları şu şekildedir:

1. Veri setindeki tüm değerleri normalize et,
2. x test noktasına en yakın M adet örneği al,
3. M örneği, test noktasına olan uzaklıklarına göre bir boyutlu bir eksen üzerine yerleştir.
4. Eksen üzerindeki noktaların uzaklık değerlerini baz alarak k-means ile l adet kümeye böl,
5. Test noktasının en yakınındaki kümenin tüm elemanlarını k-NN yöntemiyle sınıflandırma işlemine al.

Şekil 2'teki örnek senaryoda test noktana değişik uzaklıkta olan noktaların histogram grafiği görülmektedir. Bu noktalar 3 adet kümeye bölünmüştür. 1NC sınıflandırıcısı için en yakında bulunan küme içerisindeki 6 adet örnek hesaplamaya katılacaktır. Yani bu test noktası için k-NN'de k=6 alınacaktır. Test noktasına en yakın M adet örneğin l adet kümeye bölünmesi ve sadece en yakındaki ilk kümenin işleme dâhil edilmesi dinamik bir yapıyı oluşturmaktadır. l adet kümenin (k-means'deki k ifadesinin k-NN'deki k ile karıştırılmaması için l ifadesi tercih edilmiştir) her birinde yaklaşık olarak M/l adet eleman olduğu düşünülebilir. Normalde k-NN sınıflandırıcısı için kullanıcı tarafından seçilen k parametresi ile bizim yöntemimizdeki M/l kombinasyonuna denk olması Eş. 7 ile açıklanabilir:

$$k\text{NN}'deki\ k \cong \frac{M}{l} \tag{7}$$

Bu sayede M/l ikilisiyle her bir test örneğinin sınıflandırılması için uygun bir k değeri hesaplanmış olmaktadır.



Şekil 2. Kümeleyerek sınıflandırma (1NC) örneği
(An example of classification with clustering)

4.2. Örnek Tabanlı Kolektif Öğrenme
(Sample Based Ensemble Learning)

Tekil öğrenicilerin bir araya gelerek öğrenici topluluğu (ensemble) oluşturmalarına kolektif öğrenme denir. Birçok uygulamada kolektif öğrenme başarısı tekil öğrenmeye göre

yüksek çıkmaktadır [14]. Kolektif öğrenmede de sınıflandırma başarısının yüksek olabilmesi için öğrenicilerin farklı kararlarına (diversity) ihtiyaç vardır. Bu farklılıklar değişik şekillerde elde edilebilir:

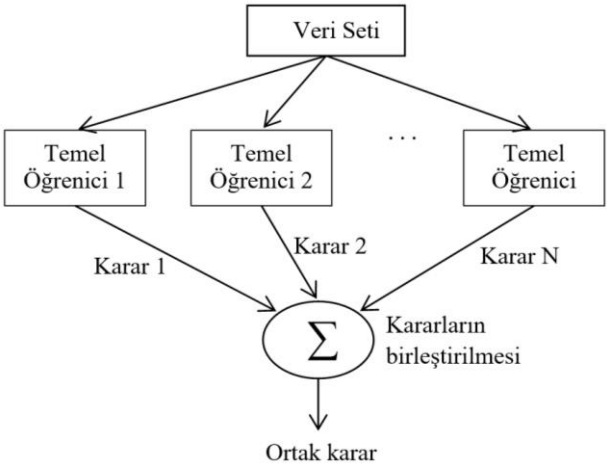
1. DT, SVM, *k*-NN gibi farklı temel öğreniciler kullanarak öğrenici topluluğu oluşturmak.
2. Aynı temel öğrenicinin farklı parametreleri (*k*-NN'deki *k* değerinin değiştirilmesi gibi) ile öğrenici topluluğu oluşturmak.
3. Aynı öğreniciyi aynı eğitim setinin değişik formları ile eğitmek. Değişik formlar elde edebilmek için eğitim setindeki örneklerin bir bölümü ile yeni veri setleri türetmek ya da eğitim setindeki özniteliklerin bazılarıyla yeni veri seti kombinasyonları türetmek olabilir.

Bu çalışmada aşağıdaki kolektif öğrenme metotları örnek tabanlı tekil öğreniciler ile kullanılabileceği için tercih edilmiştir:

1. Oylama (Voting) 2. Yerine koyarak örnekleme (Bagging) 3. Ardışık topluluklarla öğrenme (Boosting) 4. Rastsal alt uzaylar (Random Subspaces)

4.2.1. Oylama (Voting)

Güncel hayatta alınan ortak kararların genelde doğru sonuçlar verdiği ve güvenilir olduğu bilinmektedir. Temel kolektif öğrenme metotlarından olan oylama tekniğinde, en çok oy olan sınıf etiketi test noktasının sınıf etiketi olarak atanır. Basit oylamada tüm sınıflandırıcıların ağırlığı birbirine eşittir. Her bir sınıflandırıcının tüm sınıf etiketleri hakkında verdikleri kararlar birleştirilir ve ortalamaları alınır. En yüksek oranın çıktığı sınıf etiketi test örneğine atanır. Şekil 3'te algoritmanın çalışma stili görülmektedir [14].



Şekil 3. Oylama metodu (Voting Method)

Bu çalışmada ikili sınıflandırma yani {0,1} sınıfları vardır. Basit oylamada tüm modellerin ağırlığı birbirine eşittir. Oylama tekniğinde sınıf etiketi şu Eş. 8 ile bulunur:

$$\sum_{l=1}^{L} w_i d_i, j = \max(j = 1 \dots c) \sum_{l=1}^{L} w_i d_{i,j} \quad (8)$$

L adet öğrenicinin bulunduğu komitede her bir i.inci öğrenicinin T eğitim seti içinde bulunan bir test örneği için verdiği ikili sınıflandırma kararı $[d_{i,0}, d_{i,1}]^T \in \{0,1\}^c$ şeklinde tanımlansın. c sınıf etiketidir. Bazı öğrenicilerin hesaplamadaki ağırlığı artırılmak istenirse w_i değeri üzerinde değişiklik yapılabilir.

4.2.2. Yerine koyarak örnekleme (Bagging)

Bagging (Bootstrapping Aggregating) metodu, var olan bir eğitim setinden yeni eğitim setleri türeterek temel öğreniciyi yeniden 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 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 [15]. Eğitim setinden %63,2 kadar örnek seçilmesindeki neden Eş. 9 ile açıklanır:

$$\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0{,}368 \quad (9)$$

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.

4.2.3. Ardışık Topluluklar (AdaBoost)

Ardışık topluluklarla öğrenme yönteminde (Booting) en çok kullanılan AdaBoost yöntemi ilk olarak Freund ve Schapire tarafından önerilmiştir [16]. Diğer AdaBoost yöntemleriyle kıyaslandığında tahmin hızı yüksektir, daha az hafıza kullanır ve uygulanabilirliği kolaydır [17]. Bu yöntemde bir sınıflandırıcı için rastgele bir örnek seçilirken daha önce aynı sınıflandırıcının hata yaptığı örneklere öncelik verilir. Bagging'in her bir iterasyonunda tüm örneklerin eğitim kümesine seçilme olasılıkları aynıdır. Fakat Boosting'in her bir çevriminde örneklere ait seçilme olasılıkları güncellenmektedir. Bu da sistemin doğru verilen kararlardan çok, yanlış verilen kararlar üzerine odaklanılmasını sağlamaktadır. Toplam sınıflandırma başarısı seçilme olasılıklarının güncellenmesiyle sağlamaktadır [18]. AdaBoost metodunun işleyişi şu şekildedir [19]:

Adım 1: Veri setindeki n adet örnek $\{(x_1,y_1), \dots, (x_n,y_n)\}$ şeklinde verilmiş olsun. $y_i \in \{-1,+1\}, x_i \in X$ tanımlamasında x_i, her bir örneğin sınıf etiketi; y_i ise regresyon algoritmasının verdiği karardır. Pozitif örnekler için $y_i = +1$, negatifler için $y_i = -1$ olsun.

Adım 2: a pozitif örneklerin, *b* de negatif örneklerin sayısı olmak üzere *n*=*a*+*b*'dir. Ağırlıklar $w_{1,i} = 1/2b, 1/2l$ olacak şekilde her $y_i \in \{0, +1\}$ için ilklenir.

Adım 3: I iterasyon sayısı olmak üzere, her bir t=1, … ,*I* için:

a. Ağırlıklar normalize edilir: $w_{1,i} \leftarrow \frac{w_{1,i}}{\sum_{j=1}^{n} t,j}$

b. Her bir *j* özniteliği için, sadece bu *j* özniteliğini kullanan her bir h_j sınıflandırıcısı eğitilir. Hata oranı w_t ağırlığına göre şu şekilde ölçülür: $\varepsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$

c. En az ε_j hatasına sahip h_t sınıflandırıcısı seçilir.

d. Ağırlıklar şu işlem ile güncellenir: $w_{t+1,i} = w_{t,i}\beta_t^{1-e_i}$

Burada x_i doğru olarak sınıflandırma yaptıysa $e_i = 0$, aksi halde $e_i = 1$ olur. $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$ olarak hesaplanır.

Adım 4: $\alpha_t = \log 1/1 - \beta_t$ olarak alındığında $h(x)$ sınıflandırıcısının son durumu Eş. 10 gibi olur:

$$h(x) = \begin{cases} 1, & \sum_{t=1}^{I} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{I} \alpha_t \\ 0, & diğer\ durmlar \end{cases} \quad (10)$$

AdaBoost tekniğinde, en güçlü zayıf sınıflandırıcıların bir araya getirilmesiyle güçlü bir sınıflandırıcının oluşturulması hedeflenmektedir. Bunun için bu teknikte işlemler her bir eğitim örneği için eşit bir *D* dağılımıyla başlar. Her çevrimde sınıflama performansına bağlı olarak en iyi zayıf sınıflandırıcı tespit edilir ve ağırlıklar güncellenerek bir olasılık dağılım fonksiyonu elde edilir. İlerleyen adımlarda da bu işlemler tekrarlanır. Belirlenmiş bir iterasyon sonucunda en güçlü zayıf sınıflandırıcıların bir araya getirilmesiyle yüksek performanslı bir sınıflandırıcı elde edilmiş olur.

4.2.4. Rastsal Altuzaylar (Random Subspaces)

Rastsal Altuzaylar tekniğinde, belirli bir öğrenici ve kullanılan eğitim setinin farklı öznitelikleriyle oluşturulmuş yeni veri setleri vardır. Burada eğitim setinin bazı boyutları silinerek yeni eğitim setleri türetilir. Diğer bir deyişle *m* özniteliği bulunan bir veri setinden rastgele *n* adet özniteliği bulunan (*n*<*m*) yeni veri setleri türetilir. Daha sonra belirlenen temel öğrenici, türetilen eğitim setleri ile eğitilir. Öğrenicinin bir test noktasına ait kararı, türetilen veri setleri ile elde edilen öğrenicilerin verdikleri kararlar birleştirilerek hesaplanır [20].

Örnek tabanlı sınıflandırıcıların temel öğrenici seçildiği kolektif metotlarda *k* parametresinin değiştirilmesi farklılıkların oluşumunu yeterince sağlamamaktadır. Bilindiği üzere farklılıkların az olması, kolektif öğrenmede toplam sınıflandırma başarısını düşürmektedir. Diğer kolektif yöntemlerle kıyaslandığında Rastsal Altuzaylar tekniğinde daha fazla farklılıklar oluşmaktadır. Oluşan farklılıklar bu yöntemde toplam başarının artmasını sağlamaktadır [21].

5. SONUÇLAR VE TARTIŞMALAR (RESULTS AND DISCUSSIONS)

6 özniteliğe sahip veri setinde ikili sınıflandırma testleri C ve MATLAB programlama dillerinde yazılan kodlarla ve Weka yazılımı ile yapılmıştır. Elde edilen deneysel sonuçların güvenilirliği ve doğruluğu üç ana ölçüt kullanılarak incelenmiştir. Bunlar öğrenicilerin MCC (Matthews Correlation Coefficient) performans değerleri, özniteliklerin sınıflandırmadaki bilgi kazançları (Information Gain) ve özniteliklere ait korelasyon matrisidir.

5.1. MCC (Matthews Correlation Coefficient)

İkili sınıflandırma modelleri için doğruluk (Accuracy) oranı değerine göre MCC ölçütü daha güvenilir ve doğru sonuçlar vermektedir. Dengesiz bir dağılıma sahip veri setinde doğruluk oranı ister istemez yüksek çıkacaktır. Bu durumda her bir sınıfın Netlik (Precision), Hassasiyet (Recall) ve F1-Score değerlerine bakmak gerekecektir. MCC ölçütü netlik, hassasiyet ve F1-Score metriklerine bakılmaksızın iki sınıflandırma işlemlerinde en güvenilir sonucu veren performans metriğidir. MCC ölçütünün hesaplanmasında Karmaşıklık Matrisi (Confusion Matrix) kullanılmaktadır. "Hata Matrisi" olarak da isimlendirilebilecek bu yöntemde iki sınıflı bir veri setinde belirli bir sınıflandırıcının verdiği tahmin sonuçlarını barındırır.

Tablo 2. Karmaşıklık Matrisi (Confusion Matrix)

		Tahmin Edilen Sınıf	
		C_1	C_2
Gerçek Sınıf	C_1	TP	FN
	C_2	FP	TN

Tablo 2'de görüldüğü üzere C_1 ve C_2 sınıf türleri olmak üzere TP (True Positive) doğru sınıflandırılmış pozitif örnek sayısını, TN (True Negative) doğru sınıflandırılmış negatif örnek sayısını, FP (False Positive) yanlış sınıflandırılmış pozitif örnek sayısını, FN (False Negative) de yanlış sınıflandırılmış negatif örnek sayısını göstermektedir. Doğruluk (Accuracy), Hata (Error), Netlik (Precision), Hassasiyet (Recall), F1-Score ve MCC gibi başarı ölçütleri bu tablodan yararlanılarak bulunmaktadır. Karmaşıklık matrisinin oluşturulabilmesi için uygun bir çapraz geçerleme (Cross Validation) yöntemine ihtiyaç vardır. K-Fold ve Birini Devre Dışı Bırak (Leave One Out, LOO) bunlardan bir kaçıdır. Çalışmamızda veri setine en uygun çapraz geçerleme yöntemi LOO'dur. Tıbbi veriler gibi etiketli verilerin elde edilmesinin zor olduğu durumlarda veri seti seyrek (sparse) bir yapıda olmaktadır. Bu durumda literatürde en uygun çapraz geçerleme yöntemi olarak LOO tavsiye edilmektedir [22]. LOO yönteminde veri setindeki her bir örnek sırasıyla test noktası, geride kalanlar ise eğitim seti olarak belirlenir. Sınıflandırıcı geride kalan eğitim seti ile eğitilerek test noktasının sınıfını

tahmin etmeye çalışır. Tahmin değerlerinin doğru olup olmamasına göre sonuçlar karmaşıklık matrisine yerleştirilir. MCC, sınıf dağılımlarının dengesiz olduğu durumlarda bile en iyi sonucu vermektedir. MCC'de çıktı değeri -1 ile +1 arasında değişmektedir. 0 değeri rastgele sınıflandırma durumunu, -1 değeri, var olan gerçek değerler ile sınıflandırıcının verdiği kararların tamamen birbirinden zıt olduğunu göstermektedir. +1 ise sınıflandırma başarısının tam doğru olduğunu göstermektedir. MCC değeri Eş. 11 ile hesaplanır [23]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (11)$$

Tekil ve çoğul sınıflandırıcılarla yapılan deneysel işlemlerde elde edilen TP, TN, FP, FN ve MCC değerlerinin sonuçları Tablo 3 ve Şekil 4'de görüldüğü gibidir. Tüm sınıflandırma işlemlerinde MCC değerlerinin 0'dan büyük çıkması başarılı bir sınıflandırma işlemi yapıldığını göstermektedir. Çalışmada elde edilen deneysel sonuçları aynı zamanda kural tabanlı bir sınıflandırıcı olan Karar Ağacı (Decision Tree) algoritması ile de kıyaslanmaktadır. Şekilde de görüldüğü üzere yeşil renkteki örnek tabanlı tekil sınıflandırıcılar Karar Ağacı sınıflandırıcısına göre daha düşük oranda bir sınıflandırma başarısı göstermektedir. Fakat örnek tabanlı sınıflandırıcılarla oluşturulan kolektif yöntemler (Oylama tekniği hariç) hem örnek tabanlı tekil öğrenicilere göre hem de kural tabanlı bir sınıflandırıcıya göre daha yüksek performans göstermektedirler. Bu sonuçların elde edilebilmesi için kullanılan parametreler ve öğrenicilerin nasıl tasarlandığı ile ilgili açıklamalar aşağıdaki alt başlıklarda verilmiştir.

5.1.1. Tekil öğreniciler (Base Learners)

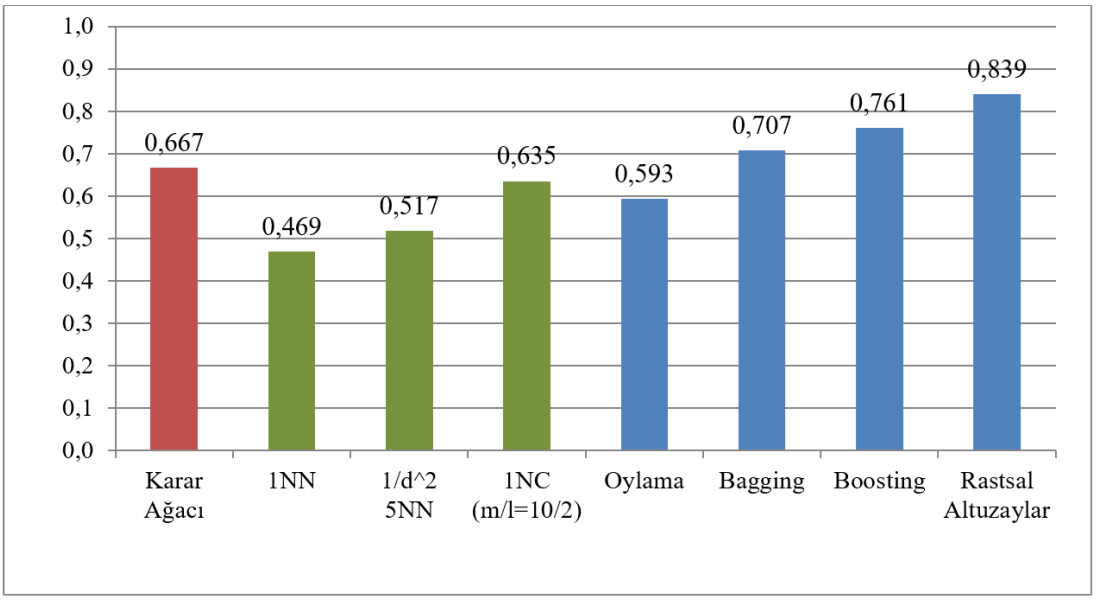
Tekil öğrenici olarak k-NN, $1/d^2$ uzaklık ağırlıklı k-NN ve 1NC sınıflandırıcıları seçilmiştir. Parametre olarak k-NN için k=1, uzaklık ağırlıklı k-NN için k=5 ve 1NC için de m/l oranı 10/2 seklinde belirlenmiştir. 1NN sınıflandırıcısı, sınıfların karar sınırlarını (decision boundary) belirlemekte ve veri setine ait Varanoi diyagramının çizilmesinde kullanıldığı için yaygın bir kullanıma sahiptir. Diğer tekil öğreniciler daha yüksek bir performans sağlarken 1NN sınıflandırıcısı çoğul öğreniciler için farklılığın sağlanmasında önemli bir katkıya sahiptir.

5.1.2. Oylama (Voting)

Bu uygulamada, *k*-NN, uzaklık ağırlıklı *k*-NN ve 1NC yöntemleri için değişik *k* parametreleriyle toplam 15 adet birbirinden farklı temel öğreniciler oluşturulmuştur. *k*-NN

Tablo 3. Metotların TP, TN, FP, FN ve MCC değerleri (TP, TN, FP, FN, and MCC values of methods)

	Karar Ağacı	1NN	1/d^2 5NN	1NC	Oylama	Bagging	Boosting	Rastsal Altuzaylar
TP	50	45	49	53	54	51	54	59
TN	36	31	30	32	29	37	37	36
FP	6	11	12	10	13	5	5	6
FN	1	16	12	8	7	10	7	2
MCC	0,667	0,469	0,517	0,635	0,593	0,707	0,761	0,869



Şekil 4. Sınıflandırıcıların MCC Performansları (MCC Performances of classifiers)

Tablo 4. Metotlarda kullanılan parametreler (Parameters used in the methods)

Yöntem	Veri Seti Sayısı	Temel Öğrenici(ler)	Temel Öğrenici(ler) için *k* veya *m*/*l* parametreleri
Oylama	1	*k*-NN	1, 3, 5, 7, 9
		Uz.Ağr. *k*-NN	1, 3, 5, 7, 9
		1NC	5/2, 10/2, 15/3, 20/4, 30/5
Bagging	15	1NC	10/2
AdaBoost	15	1NC	10/2
Rastsal Altuzaylar	15	*k*-NN	1

ve uzaklık ağırlıklı *k*-NN sınıflandırıcıları için *k* değeri 1, 3, 5, 7, 9 olarak belirlenmiştir. 1NC için ise *m*/*l* oranı 5/2, 10/2, 15/3, 20/4 ve 30/5 olacak şekilde seçilmiştir. Oylama neticesinde elde edilen toplam MCC değeri temel öğrenicilerin sınıflandırma başarısına çok yakındır. Kolektif sınıflandırmada performans artırımı için temel sınıflandırıcıların kararlarında farklılıklar olması zorunludur. Örnek tabanlı sınıflandırıcılarda sınıflandırma başarısını artırabilme için *k* parametresinin değiştirilmesi veya farklı *k*-NN öğrenicilerinin kullanılması yeterli olmadığı bu uygulama ile anlaşılmış olmaktadır. Çünkü oylama tekniğinde elde edilen başarı oranı diğer üç tekil sınıflandırıcıya ait başarı oranlarının ortalaması gibidir.

5.1.3. Bagging (Bootstrapping Aggregating)

Bagging yöntemi için orijinal eğitim setinden 15 adet yeni eğitim seti türetilmiştir. 4.2.2. nolu başlık altında da anlatıldığı gibi Bootstrapping yöntemiyle rastgele olarak yeni veri setleri türetilerek farklılıklar oluşturulmaya çalışılmıştır. Oluşturulan veri setleri yukarıda bahsedilen 3 farklı temel öğreniciye uygulanmış ve en yüksek başarı 1NC ile elde edilmiştir. Tablo 4'de de görüldüğü üzere 1NC için parametre (*m*/*l*=10/2) alınmıştır. Şekil 4'te görüldüğü üzere yerine koyarak örnekleme (Bagging) yöntemlerinde MCC metriğine göre 0,707 değerinde bir başarı elde edilmiştir. Bu durum doğruluk açısından bir miktar gelişme sağlandığını göstermektedir.

5.1.4. Ardışık Topluluklar (AdaBoost)

Bu yöntemde Bagging'de olduğu gibi 1NC tekil sınıflandırıcı kullanılmıştır. Bagging ile kıyaslandığında AdaBoost'a ait MCC değeri biraz daha yüksek çıkmaktadır. Bunun nedeni Boosting yönteminin hata yapılan örnekler üzerinde sistemin kendini yeniden eğitmesi olarak açıklanabilir. Çoğul sınıflandırıcılar için kullanılan tekil öğreniciler ve parametreler Tablo 3'de görüldüğü gibidir.

5.1.5. Rastsal Altuzaylar (Random Subspaces)

Bu yöntemde, *n* adet özniteliğe sahip bir veri setinden *m* adet (*m*<*n*) özniteliğe sahip rastgele kombinasyonlar türetilerek yeni veri setleri oluşturulur. Uygulamada 6 öznitelikli veri setinden en fazla 4 öznitelikli 15 rastgele kombinasyon Tablo 5'de görüldüğü gibi oluşturulmuştur. A'dan F'ye kadar isimlendirilmiş olan öznitelikler sırayla VKİ, Kişisel ve psikolojik bilgiler, Aile özellikleri, Aktiviteler, Öğün tüketimleri ve Besin tüketim sıklığıdır. Yıldız "*" ile işaretlenenler yeni eğitim setine seçilen özniteliklerdir. Bu yöntemde kullanılan tekil öğrenici 1NN'dir. 1NC'ye göre daha düşük bir başarı oranına sahip olmasına karşın yapılan uygulamalarda 1NN'li Rastsal Altuzaylar, 1NC'li Rastsal Altuzaylara göre daha yüksek başarı vermiştir. Bunun nedenini karar sınırında olan bir örnek için diğer tekil sınıflandırıcılara göre 1NN'in oluşturulan farklılıklar yardımıyla doğru karar vermesi olarak açıklanabilir.

Tablo 5. Oluşturulan eğitim setlerindeki öznitelikler (Attributes of created datasets)

No	Öz-A	Öz-B	Öz-C	Öz-D	Öz-E	Öz-F
1	*	*	*	*		
2	*		*		*	
3	*	*	*			*
4		*		*	*	*
5	*	*		*		*
6		*	*	*		*
7	*			*	*	*
8	*		*		*	*
9	*	*			*	*
10	*		*	*		*
11	*		*	*	*	
12	*			*		*
13		*	*		*	*
14	*	*			*	*
15			*		*	*

Tekil öğrenici olarak seçilen örnek tabanlı öğrenicilerin kolektif öğrenmede kullanılmasının başarıyı istenilen düzeyde artırmadığı saptanmıştır. Rastsal alt uzaylarda 0,839 gibi yüksek oranda başarı elde edilmesi türetilen yeni eğitim setlerinde oluşan farklılıklar sonucu oluştuğu görülmüştür. Kısaca türetilen yeni veri setleri çoğul öğrenme başarısını artırmıştır. Eğitim setindeki VKİ ve beslenme alışkanlıklarının bulunduğu öznitelikler rastsal alt uzaylar uygulamasında oluşturulan yeni eğitim setlerinde sıklıkla seçilmiştir. Bunun nedeni bu özniteliklerin obezite riskinin belirlenmesinde büyük bir öneme sahip olmasıdır. Bu özniteliklerin değişik kombinasyonlarda bulunmaması doğruluk oranının düşük çıkmasına neden olmaktadır.

Örnek tabanlı öğreniciler için en iyi farklılık oluşturma yöntemi boyutlar üzerinde yapılan değişikliklerle sağlanabildiği bu uygulama sayesinde gözlemlenmiştir. Bu uygulamada *k*-NN versiyonları içerisinde en yüksek performansı 1NC ile kıyaslandığında az bir farkla 1NN sınıflandırıcısı vermiştir.

5.2. Bilgi Kazancı (Information Gain)

Bilgi kazancı (IG), karar ağaçlarının inşa edilmesinde kullanılır ve kabaca belirsizlik ölçüsü olan entropi değerinin tersi olarak tanımlanır. Retrospektif bir yaklaşımla obeziteyi etkileyen risk faktörlerinin analizi yapılmak istendiğinde bilgi kazancı tablosu kullanılabilir. Diğer bir deyişle öznitelikler içerisinde obeziteyi en çok etkileyen faktör bu yöntem yardımıyla bulunur. IG değerlerinin yüksek olması ilgili özniteliğin sınıf etiketlerini ayrıştırmada etken bir rol oynadığını göstermektedir [24]. Her bir sınıf etiketi üzerindeki IG değerlerinin hesaplanabilmesi için entropi değerlerinin hesaplanması gerekir. Entropi, sistemdeki belirsizlik ölçüsüdür ve Eş. 12 ile hesaplanır:

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (12)$$

n tane sınıf etiketinin bulunduğu bir veri setinde her bir sınıf etiketinin bulunma olasılığı p_i'dir. 0 ile 1 arasında değişen Bilgi Kazancı Eş. 13 ile bulunur:

$$H(x,T) = H(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i) \quad (13)$$

T veri kümesi, *x* ise hesaplanması istenen sınıf türüdür. Tablo 6'da özniteliklerin bilgi kazançları sıralı bir şekilde verilmektedir.

Tablo 6. Özniteliklerin bilgi kazançları (IG of each attribute)

Sıra	Öznitelik	IG
1	Öz-F. Bazı besin gruplarını tüketim sıklığı	0,7166
2	Öz-E. Günlük öğün sıklığı	0,2001
3	Öz-D. Aktivite	0,1933
4	Öz-A. VKİ	0,0810
5	Öz-B. Psikolojik durumu	0,0164
6	Öz-C. Aile bilgileri	0,0001

Çalışmamızın bu bölümünde obeziteye neden olan faktörler IG değerlerine bakarak sırasıyla belirlenmiştir. Sonuç olarak obeziteye neden olan faktörler sırasıyla bazı besin gruplarını tüketim sıklığı, günlük öğün sıklığı, bedensel aktivite, VKİ değeri ve bireyin psikolojik durumudur. 0,7166'lık bir bilgi kazancı değeriyle en önemli etkenin bazı besin türlerinin sıklıkla tüketilmesi olduğu sonucuna varılabilir. VKİ bilindiği üzere obezitede bir neden değil bir sonuçtur. Bu yüzden bir faktör olarak değerlendirilmemelidir. Görüldüğü üzere günlük öğün tüketim sıklığı ve bedensel aktivite durumları obeziteyi tetikleyen en etkili ikinci ve üçüncü faktörlerdir. Bu iki etmenin IG değerleri birbirine yakın olduğu için aynı etki derecesinde olduğu düşünülebilir. Anket sorularında da anlaşılacağı üzere bireyin gece yatarken yemesi, düzensiz öğün alışkanlıkları ve kahvaltının yapılmaması gibi durumların en etkili faktörler olduğu görülmektedir. Ayrıca hareketsiz bir yaşantı ve elektronik cihazlar karşısında geçirilen sürenin fazla oluşu gibi durumlar da obeziteye tetikleyen etkenlerdendir. Tablo 6'da aileye ait özelliklerin IG değeri sıfıra çok yakın çıkmıştır. Aile bireylerinde kalıtsal olabilen obezite ve şeker hastalığı, ebeveynlerin sigara ve alkol bağımlıkları, ailenin ekonomik düzeyi gibi faktörler sınıflandırmada neredeyse hiçbir etkiye sahip değildir. Fakat bilindiği üzere obezite hastalığının ana nedenlerinden biri kalıtsal özelliklerdir [25]. Bu çalışmada IG değerinin beklenilenin altında çıkması kullanılan veri setine bağlı olabilir. Veri setindeki örneklerin yetersizliği IG değerinin düşük çıkmasına neden olmuş olabilir.

5.3. Korelasyon Matrisi (Correlation Matrix)

Korelasyon matrisi öznitelik çiftleri arasındaki doğru veya ters orantılı istatistiksel ilişkiyi analiz etmek için kullanılan bir yöntemdir. Eğitim setinde bulunan 6 özniteliğin ve obezite durumunu belirten sınıf türünün birbirleri aralarındaki korelasyon katsayıları Tablo 7'de verilmiştir. Katsayıların hesaplanmasında Pearson yöntemi kullanılmıştır [26]. Bilindiği üzere -1 değerlikli korelasyon, katsayısı iki değişken arasında tam ters ilişkiyi; 0, ilişkisizliği; +1 ise tam lineer düz ilişkiyi göstermektedir. Genel olarak +1'e ve -1'e yakın değerler anlamlı bir korelasyonun göstergesidir. [0,3 ve 0,7] aralığındaki korelasyon katsayısı ise orta düzey bir ilişkiyi göstermektedir [27].

Tablo 7. Özniteliklerin korelasyon matrisi
(Correlation matrix of attributes)

	Öz-A	Öz-B	Öz-C	Öz-D	Öz-E	Öz-F	Sınıf
Öz-A	1	0,17	0,08	0,05	0,08	0,39	0,67
Öz-B		1	0,03	0,26	0,20	0,32	0,46
Öz-C			1	0,14	0,06	0,08	0,21
Öz-D				1	0,13	0,29	0,39
Öz-E					1	0,41	0,48
Öz-F						1	0,80
Sınıf							1

Öz-F (bazı besin gruplarını tüketim sıklığı) ile sınıf etiketi arasındaki 0,80'lik değer güçlü bir korelasyonun varlığını göstermektedir. Öz-F, obeziteyi tetiklediği bilinen bir takım gıdaların tüketim sıklığını inceleyen soruları içermektedir. Yani ankete katılan öğrencilerin gün içerisinde tükettikleri besin türleri ile obezite durumu arasında doğru orantılı bir ilişki vardır. Bu nedenle çocuklarda obeziteye etki eden en önemli faktörün bazı besinlerin tüketimi ile alakalı olduğu ortaya çıkmaktadır. Öz-A (VKİ) – Obezite arasında yapılan korelasyon hesaplamasında ise 0,67 gibi yüksek sayılabilecek bir ilişki katsayısı bulunmuştur. Elde edilen bu değere göre obezitenin ikinci nedeni olarak VKİ görülebilir. Fakat VKİ, obeziteyi tanımlayan bir faktör olduğu için bir neden olarak görülmemeli; bir sonuç olarak değerlendirilmelidir. Öz-E (günlük öğün tüketim durumu) ile obezite arasında 0,48'lik bir katsayı çıkmıştır ki bu da

aradaki korelasyonun orta düzeyde olduğunu gösterir. Bireyin içinde bulunduğu psikolojik durumu irdeleyen Öz-B niteliği ile sınıf etiketi arasında 0,46'lık bir katsayı elde edilmiştir. Ayrıca haftalık aktivite durumunu içeren Öz-D özniteliği ile obezite durumu arasında 0,39'luk orta düzey bir korelasyon vardır. Bu iki katsayı ile obeziteyi tetikleyen iki etken faktör tespit edilebilmektedir. Öz-E ile Öz-F arasındaki 0,41'lik korelasyon katsayısı ile Öz-A ile Öz-F arasındaki 0,39'luk katsayılar bu ikililer arasında orta düzeyde bir ilişki olduğunu göstermektedir. Mantıksal açıdan bu ikililer arasındaki korelasyon değerleri incelendiğinde böyle bir ilişkinin olması normal karşılanmalıdır. Ayrıca diğer öznitelik ikilileri arasında anlamlı bir katsayının çıkmaması ilgili soruların birbirlerinden bağımsız olduğunu göstermektedir.

5.4. İleri Uygulamalar ve Değerlendirme
(Further Applications and Evaluations)

Bilindiği üzere insana ait kişisel özelliklerin, psikolojik ve sağlık durumunun, yaşamsal alışkanlıkların sayısallaştırılması oldukça zor bir iştir. Çalışmada bu tarz zorlukların üstesinden gelmek için özgün yapay bir tahmin modeli öne sürülmüştür. Elbette KKDS modelinin ileri aşamalarında yapılabilecek değişik uygulamaları olabilir.
Öncelikle anket soruları ve cevaplarının elde edilmesi ile ilgili daha farklı çalışmalar yapılabilir. Çocuklarda obeziteyi tetikleyen faktörlerden 30 kadarı bu çalışmada ankete alınmıştır. Daha fazla faktör anket içerisine alınarak çalışma detaylandırılabilir. Anket soruları, uzman doktorların, diyetisyenlerin ve bilişim uzmanlarının bulunduğu bir komisyon tarafından titizlikle hazırlanabilir. Kişilerin kan grupları, protein yapıları, genetik hastalıkları, daha önce geçirdikleri hastalıklar, düzenli olarak kullandıkları ilaçlar ve laboratuarlardan elde edilen idrar ve kan tahlili gibi birçok faktör ele alınabilir. Bu sayede olası her bir faktörün hesaplamaya katılmasıyla daha hassas, net ve güvenilir sonuçlar elde edilebilir. Ayrıca daha fazla denekten elde edilen verilerle daha güvenilir bir veri seti oluşturulabilir. Veri setinin güçlendirilmesi için hastanelerde obezite tedavisi gören hasta kayıtları kullanılabilir. Veri zengini ve bilgi yoksunu olarak değerlendirilen hastanelerin hasta kayıtları bu şekilde değerlendirilmiş olacaktır. Projenin geliştirilmesi adına farklı sınıflandırma modelleri çalışmada kullanılabilir. Öncelikle Destek Vektör Makineleri, Karar Ağaçları, Yapay Sinir Ağları, Naive Bayes sınıflandırıcısı ve değişik lineer modeller kullanılabilir [28]. En başarılı sonucu veren sınıflandırıcı, kolektif metotlarda tekil öğrenici olarak kullanılabilir. Bu çalışmadaki amaç farklı sınıflandırma modellerinin aynı veri seti üzerindeki performanslarını karşılaştırmak değildir. Amacımız daha önce belirtildiği üzere özgün bir KKDS modelinin tasarlanmasıdır. Hazırlanan veri setindeki obez hastalar 1; sağlıklı çocuklar ise 0 sayısal değeriyle etiketlendiği için bu veri seti regresyon çalışmaları için de uygundur. Lineer regresyon ve CART (Classification and Regression Tree) gibi algoritmalarla bireye ait [0, 1] aralığındaki yüzdelik obezite riski hesaplanabilir. KKDS sistemiyle elde edilen bulanık değerler sayesinde daha gerçekçi yorumlar elde edilebilir ve gerekli tedbirlerin erkenden alınması sağlanabilir.

6. SONUÇLAR (CONCLUSIONS)

Çalışmamızda makine öğrenmesi kolektif sınıflandırıcı yöntemleri farklı bir alana uygulanmıştır. Orijinal, güvenilir ve gerçek eğitim seti kullanılarak kolektif öğreniciler ile çağın büyük bir sorununa karşı hem erken tanı sistemi hem de klinik karar destek sistemi geliştirilmiştir. Oluşturulan erken uyarı sistemi sayesinde büyümekte olan çocuğun obezite riski çocuktan alınan kişisel, psikolojik ve yaşamsal alışkanlıklarına ait verilerle tahmin edilebilir. Yapılan geçerleme işlemlerinde uygulanan değişik metotlarda yüksek başarılı tahminlerin yapıldığı gözlemlenmiştir. Çalışmamızdaki amaç, örnek tabanlı sınıflandırıcıların kolektif yöntemlerde kullanılmasını test etmek değil sağlık bilişimi alanında kullanılabilir, faydalı ve güvenilir sonuçlar verebilen özgün bir tahmin modeli inşa edilmesidir. Sonuç olarak amaçlanan erken tanı sistemi oluşturulabilmiş ve çağımızın ciddi bir hastalığı olan obezite için gerekli önlemlerin önceden alınabilmesi mümkün hale gelmiştir.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Verilerin toplanabilmesi için bize resmi izin veren, bizi destekleyen ve hoş gören İstanbul İl Milli Eğitim Müdürlüğüne, İstanbul İl Sağlık Müdürlüğüne, Çapa Eğitim ve Araştırma Hastanesi yönetimine ve Pediatrik Endokrinoloji bilim dalı Öğretim Üyesi Prof. Dr. Nurçin SAKA'ya, Şişli Etfal Eğitim ve Araştırma Hastanesine ve Çocuk Endokrinoloji Doktoru Uz. Dr. Mehmet BOYRAZ'a, Büyükçekmece Adem Çelik İlkokulu yönetici ve öğretmenlerine teşekkürlerimizi bir borç bilir saygılarımızı sunarız. Ayrıca bu çalışmaya büyük katkı saylayan, özveri ile hastalara anket yapan ve verileri bilgisayara aktaran değerli bilim insanları Rukiye Dilruba KÖSE'ye ve Beyza Nur KÖKCÜ'ye sonsuz teşekkürlerimizi arz ederiz.

KAYNAKLAR (REFERENCES)

1. Ulutaş A.P., Atla P., Say Z.A., Sarı E., Okul Çağındaki 6-18 Yaş Arasi Obez Çocuklarda Obezite Oluşumunu Etkileyen Faktörlerin Araştırılması, Zeynep Kamil Tip Bülteni, 45 (4), 192-196, 2014.
2. Adnan M.H.M., Husain W., & Rashid N.A.A., A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction, IEEE International Conference Computer & Information Science (ICCIS), 1, 281-285, Temmuz 2012.
3. Heydari S.T., Ayatollahi S.M.T., Zare N., Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity, Journal of medical systems, 36 (4), 2449-2454, 2012.
4. Manna S., Jewkes A.M., Understanding early childhood obesity risks: An empirical study using fuzzy signatures, IEEE International Conference In Fuzzy Systems (FUZZ-IEEE), 1333-1339, 2014.

5. Kóczy L.T., Vámos T., Biró G, Fuzzy Signatures, 2nd International Conference on Soft and Intelligent Computing EUROPUSE-SIC, Budapeşte, 210–217, 1999.

6. Dugan T.M., Mukhopadhyay S., Carroll A., Downs S., Machine Learning Techniques for Prediction of Early Childhood Obesity, Applied clinical informatics, 6 (3), 506-520, 2015.

7. Uskun E., Öztürk M., Kişioğlu A.N., Kırbıyık S., Deömirel, R., İlköğretim Öğrencilerinde Obezite Gelişimini Etkileyen Risk Faktörleri, S.D.Ü. Tıp Fakültesi Dergisi, 12 (2), 19-25, 2005.

8. Bulut F. Study of Obesity: Identifying Risk Rate of Obesity with Instance Based Ensemble Classifiers. https://sites.google.com/site/bulutfaruk/study-of-obesity. Erişim tarihi Ocak 11, 2017.

9. Yılmaz A.A., Özaydın E., Demirel F., Kös, G. Obez Adölesanlarda Obezite Gelişimini Belirleyen Faktörlerin ve Metabolik Sendrom Varlığının Retrospektif Olarak Değerlendirilmesi, Türkiye Çocuk Hastalıkları Dergisi, 1-5, 2015.

10. Biudnik M., Pozniak I., Koszalka L., The Usage of the k-Nearest Neighbour Classifier with Classifier Ensemble, 12th International Conference on Computational Science and Its Applications (ICCSA), 170-173, Haziran 2012.

11. Chen P., Wang Y., Optimized KD Tree Application in Instance-Based Learning, Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'08), Shandong, 187-191, 2008.

12. Bulut F., Amasyalı M.F., Classification in mixture of experts using hard clustering and a new gate function, Journal of the Faculty of Engineering and Architecture of Gazi University, 31 (4), 1017-1025, 2016.

13. Bulut F., Amasyali M.F., Locally adaptive k parameter selection for nearest neighbor classifier: one nearest cluster, Pattern Analysis and Applications, Springer London, DOI 10.1007/s10044-015-0504-0, 1-11, 2015.

14. Zhou Z.H., Machine Learning Pattern Recognition Series, Ensemble Methods: Foundations and Algorithms, CRC Press, ISBN: 978-1439830031, 2012.

15. Zhou Z. H., Ensemble learning, Encyclopedia of Biometrics, Berlin Springer, 411-416, 2015.

16. Freund Y., A short introduction to boosting, Journal-Japanese Society For Artificial Intelligence, 14 (1), 771-780, 1999.

17. Mayr A., Binder H., Gefeller O., & Schmid M., The evolution of boosting algorithms, Methods of information in medicine, 53 (6), 419-427, 2014.

18. Breiman L., Bias, Variance, and Arcing Classifiers, Technical Report, Statistics Department, University Of California, Berkeley, 1996.

19. Tetik Y.E., Gürültülü Ortamlarda Konuşma Tespiti İçin Yenibir Öznitelik Çıkarım Yöntemi, Elektrik-Elektronik ve Bilgisayar Sempozyumu, Fırat Üniversitesi-Elazığ, 86-89, 2011.

20. Ho T.K., The Random Subspace Method for Constructing Decision, IEEE Transactions on Pattern Analysis and Machine Intelligence, Lucent Tech no l., AT&T Bell Labs., Murray Hill, 20 (8), 832-844, 1998.

21. Ho T.K., Lecture Notes in Computer Science, Nearest neighbors in random subspaces, Advances in Pattern Recognition, Springer Berlin Heidelberg, 640-648, 1998.

22. Alpaydın E., Yapay Öğrenme Kitabı, Boğaziçi Yayınları, Birinci Basım, ISBN: 9786054238491, 416-417, Mart 2011.

23. Liu Y., Cheng J., Yan,C., Wu X., Chen F., Research on the Matthews Correlation Coefficients Metrics of Personalized Recommendation Algorithm Evaluation, International Journal of Hybrid Information Technology, 8 (1), 163-172, 2015.

24. Akben S.B., Alkan A., Density-based feature extraction to improve the classification performance in the datasets having low correlation between attributes, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (4), 597-603, 2015.

25. Blakemore A.I.F., Buxton J.L., Obesity, Genetic Risk, and Environment, BMJ-British Medical Journal, doi:10.1136/bmj.g1900, 348, 2014.

26. Puth M.T., Neuhäuser M., Ruxton G.D., Effective use of Pearson's product–moment correlation coefficient, Animal Behaviour, 93, 183-189, 2014.

27. Asuero A.G., Sayago A., Gonzalez A.G., The Correlation Coefficient: An Overview, Critical Reviews in Analytical Chemistry, 36 (1), 41-59, 2006.

28. Haltaş A., Alkan A., Karabulut M., Performance analysis of heuristic search algorithms in text classification, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (3), 417-427, 2015.