# BANKA TELEPAZARLAMA BAŞARISININ TAHMİNİ İÇİN BİR BİRLEŞİK MAKİNE ÖĞRENME TABANLI KARAR DESTEK MODELİ

Ömer ALGORABİ[1], Ersin NAMLI[2]

***ÖZET***

***Amaç:*** Günümüzde bankacılık sektöründeişlem verilerinin yakalanmasını sağlayan elektronik bankacılık daha çok benimsenmeye başlanmış ve bu tür verilerin miktarı önemli ölçüde artmıştır. Bu verileri analiz etmek için veri madenciliğine dayalı teknikler benimsenmiştir. Bu çalışmada müşterilerin vadeli mevduat uygunluk durumlarına göre sınıflandırılması amaçlanmıştır.

***Yöntem:*** Bu çalışmada, kullanılan veri seti Portekiz Bankacılık Kurumu'nun müşterilerinden telefon ile iletişim yoluyla elde ettiği pazarlama kampanyaları verilerinden oluşmaktadır. Veriler C4.5, Naive Bayes, Bayes Ağları, k-En Yakın Komşu ve Sıralı Minimal Optimizasyon (SMO) sınıflandırma algoritmaları kullanılarak sınıflandırılmıştır. Sınıflandırma modelleri Sentez indeks (SI) değerlerine göre karşılaştırılmıştır.

***Bulgular:*** Elde edilen sonuçlara göre basit C4.5, en iyi sınıflandırma modeli olarak bulunmuştur. Önerilen model, literatürdeki diğer çalışmaların aynı veri seti üzerinde uyguladığı yöntemlerden daha üstün bulunmuştur.

***Özgünlük:*** Literatürdeki mevcut çalışmalardan farklı olarak bu çalışmada, topluluk öğrenme yöntemleri ile farklı sınıflandırma modelleri oluşturulmuş ve sentez indeks olarak yeni bir performans ölçütü geliştirilmiştir.

***Anahtar Kelimeler:*** Bankacılık, Tele Pazarlama, Makine Öğrenme, Sınıflandırma.

***JEL Kodları:*** C38, E50, M31.

# AN INTEGRATED MACHINE LEARNING BASED DECISION SUPPORT MODEL FOR PREDICTION OF BANK TELEMARKETING SUCCESS

***ABSTRACT***

***Purpose:*** Today, electronic banking, which enables the capture of transaction data, has started to be adopted more and the amount of such data has increased significantly. Data mining-based techniques have been adopted to analyze this data. In this study, it is aimed to classify customers according to their time deposit eligibility status. The bank usually needs to make more than one phone connection to the same customer to understand whether a time deposit product can be offered.

***Methodology:*** In this study, the data set used consists of the marketing campaigns data obtained by the Portuguese Banking Agency from its customers via telephone communication. Data is classified with C4.5, Naive Bayes, Bayes Networks, k-Nearest Neighbor and Sequential Minimal Optimization (SMO) classification algorithms. Classification models are compared according to synthesis index (SI) values.

***Findings:*** According to the results, simple C4.5 was found to be the best classification model. The proposed model was found to be superior to the methods applied by other studies in the literature on the same data set.

***Originality:*** Different from the existing studies in the literature, in this study, different classification models were created with ensemble learning methods and a new performance criterion was developed as a synthesis index.

***Keywords:*** Banking, Telemarketing, Machine Learning, Classification.

***JEL Codes:*** C38, E50, M31.

[1] Arş. Gör., İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, İstanbul, Türkiye, omer.algorabi@iuc.edu.tr, ORCID: 0000-0002-2016-8674.

[2] Dr. Öğr.Üyesi, İstanbul Üniversitesi-Cerrahpaşa, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, İstanbul, Türkiye, enamli@iuc.edu.tr, ORCID: 0000-0001-5980-9152 (Sorumlu Yazar-Corresponding Author).

DOI: 10.51551/verimlilik.748616

Araştırma Makalesi/ Research Article | Geliş Tarihi/Submitted Date: 05.06.2020 | Kabul Tarihi/Accepted Date: 22.03.2021

## 1. GİRİŞ

Telefonla pazarlama, bir satış temsilcisinin müşterileriyle telefon veya ürün satışı yapmak için iletişim kurduğu doğrudan pazarlama türüdür. Potansiyel müşterilere ait veriler doğrudan pazarlama veri tabanından gelir ve çoğunlukla iletişim, reklam ve analiz için kullanılmaktadır. Telefonla pazarlama ile başarıyı elde etmek için, şirketin, satmaya çalıştığı ürünü veya hizmeti kullanma olasılığı daha yüksek olan müşteri kitlesini tahmin ederek, potansiyel müşterilere odaklanmalıdır (Palaniappan ve diğerleri, 2017).

Dünya çapında bankacılık sektörü, işin yönetilme şekliyle ilgili çarpıcı değişiklikler yaşamıştır. Günümüzde işlem verilerinin yakalanmasını sağlayan elektronik bankacılık daha fazla benimsenmeye başlanmıştır ve bu tür verilerin miktarı önemli ölçüde artmıştır. İnsan aklının böylesine büyük miktarda ham veriyi analiz etmesi ve verileri bir organizasyonun yararına yararlı bilgiye dönüştürmesi mümkün değildir (Keles ve Keles, 2015).

İnsanın böylesine büyük ham veriyi analiz edemeyecek olduğu durumda, müşteri davranışlarını anlamak için, birçok firma, özel hizmetler sunmadan önce müşteri verilerini işleyerek müşterileri sınıflandıran veri madenciliğine dayalı teknikleri kullanmaktadır (Vajiramedhin ve Suebsing, 2014) . Veri madenciliği, veri kayıtlarında yer alan özellikler arasındaki gizli ve bilinmeyen ilişkileri keşfedebilen, ilginç olayları ve gömülü kalıpları tespit edebilen ve bilgi alanının özünü özetleyebilen, çok sayıda veri kümesinden yeni ve yenilikçi bilgileri izleme süreci olarak bilinir. Bu doğrultuda, veri madenciliği, bankacılık sektöründeki karar vericilere, banka tahliyesine neden olan riskli işlemlerden kaçınarak banka gelirlerini yükseltmek ve müşteri tutma teşviklerini artırarak ekonomik hilelere karşı koymaya yardımcı olmak için kullanılabilmektedir (Abbas, 2015).

Bu çalışmada, Portekiz Bankacılık Kurumu'nun müşterilerinden telefon ile iletişim yoluyla elde ettiği pazarlama kampanyaları verileri kullanılarak müşterilerin vadeli mevduat uygunluk durumları sınıflandırılmıştır. Basit ve entegre sınıflandırma yöntemleri kullanılarak karşılaştırmalar yapılmıştır. Bu karşılaştırmalar geliştirilen sentez indekse göre yapılmıştır. En son çalışmada üstün bulunan model, farklı çalışmaların aynı veri seti üzerinde kullandıkları sınıflandırma modelleri ile karşılaştırılılmış ve bu çalışmanın verimliliği ortaya konmuştur.

Çalışmanın geri kalanı şu şekilde organize edilmiştir. İkinci bölümde, literatür taraması yapılmıştır. Üçüncü bölümde, kullanılan veri seti, metodolojilerin matematiksel temelleri ve performans ölçütleri ayrıntılı olarak açıklanmıştır. Dördüncü bölümde, sınıflandırma sonucu elde edilen bulgular verilmiştir. Son bölümde yer alan sonuç bölümü ile çalışma sonlandırılmıştır.

## 2. LİTERATÜR TARAMASI

Literatürde “Bank Marketing” veri setinin birçok çalışmada kullanıldığı görülmüştür. Giriş verilerinin özelliğinin azalmasına ve bankanın tahmin oranının yükselmesine yardımcı olacak bir tahmin modeli öneren Vajiramedhin ve Suebsing (2014) çalışmalarında, korelasyon tabanlı öznitelik seçimi algoritması ve sınıflandırma için C4.5 metodu kullanılmıştır. Önerdikleri modelde doğru pozitif (TP) oranı %92,14 bulunmuştur. Bir diğer çalışmada ise, CRM veri madenciliği çerçevesi ile Çok Katmanlı Algılayıcı Sinir Ağı (MLPNN) ve Naïve Bayes (NB) sınıflandırıcıları kullanılmıştır (Bahari ve Elayidom, 2015). Doğruluk oranları MLPNN için %88,63 bulunurken NB için %87,97 olarak bulunmuştur. Popelka ve diğerleri (2016) çalışmalarında, veri setini sınıflandırmak için ADTree, Random Forest, BFTree, C4.5, Radial Basis Function, MLP ve SVM Algoritmaları kullanılmıştır. Elde edilen sonuçlara göre en iyi sınıflandırıcı %90,20 doğruluk oranıyla C4.5 bulunmuştur. Bir diğer çalışmada C4.5 kullanılarak %90,68 doğruluk oranı bulunmuştur (Palaniappan ve diğerleri, 2017). Bir başka çalışmada ise bankacılık kampanyaları için hedef müşterilerin tanımını desteklemek üzere rastgele ormanlar tarafından desteklenen bir veri madenciliği yanıt modeli Migueis ve diğerleri (2017) tarafından önerilmiştir. Bir alt-örnekleme yönteminin performansı, en uygun spesifikasyonun belirlenmesi için bir aşırı örnekleme yöntemiyle karşılaştırılmıştır. Özellikle, ayrımcı performans demografik bilgilerin, iletişim detaylarının ve sosyo ekonomik özelliklerin dahil edilmesiyle artırılmıştır. Bir alt-örnekleme algoritması tarafından desteklenen rastgele ormanlar, keşfedilen diğer tekniklerden daha iyi performans gösteren çok yüksek tahmin performansı sunmuştur. Bir başka çalışmada Yapay Sinir Ağları kullanılarak sınıflandırma oranı %84,4 elde edilmiştir (Koç ve Yeniay, 2013). Pradap ve Kamaludeen’in 2019’daki çalışmalarında ise farklı sınıflandırma yöntemleri içerisinde Rastegele Orman %85,76 ile en iyi doğru sınıflandırma oranını vermiştir (Pradap ve Kamaludeen, 2019). Derin öğrenme yöntemleri kullanılarak da veri seti sınıflandırılmıştır. Kim ve diğerleri (2015) çalışmalarında Evrişimsel Sinir Ağları kullanılarak %76,70 doğru sınıflandırma oranı elde edilmiştir. Türkmen (2021) çalışmasında ise Uzun Kısa Vadeli Bellek (LSTM), Kapılı Tekrarlayan Birim (GRU) ve Basit Tekrarlayan Sinir Ağı (SimpleRNN)) kullanarak telepazarlama verisini sınıflandırmıştır. Bir başka çalışmada sınıflandırmanın etkinliğini artırmak için karar ağaçları sınıflandırıcıları yanında metasezgisel yöntemlerden biri olan karınca kolonisi optimizasyonu kullanılmıştır (Kozak ve Juszczuk, 2018)

Literatürdeki mevcut çalışmalardan farklı olarak bu çalışmada, yığılma ve oylama topluluk öğrenme yöntemleri ile K-En Yakın Komşu, Bayes Ağları, Karar Ağacı ve Naïve Bayes sınıflandırma modelleri geliştirilmiştir. Doğruluk oranı, kappa istatistiği, hata değerleri gibi farklı performans ölçütlerinden performans sentez indeks ve hata sentez indeks olarak yeni performans ölçütleri geliştirilmiş ve bu ölçütlere göre karşılaştırmalar yapılmıştır.

## 3. YÖNTEM

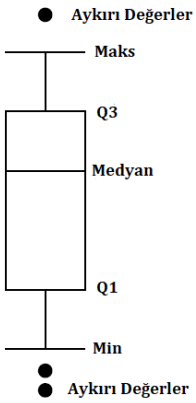
### 3.1. Kutu Grafiği

Kutu grafiği sürekli tek modlu verilerinin dağılımını görselleştirmek için çok popüler bir grafik aracıdır. Verilerin konumu, yayılımı, çarpıklığı ve kuyrukları ile ilgili bilgileri göstermektedir. Bununla birlikte, veriler çarpık olduğunda, genellikle birçok nokta bıyığı aşar ve aykırı değer olarak tespit edilir (Hubert ve Vandervieren, 2008). Bir kutu grafiği modeli Şekil 1'deki gibidir.

$X_n = \{x_1, x_2, \ldots, x_n\}$, n değişkenli bir veri kümesidir, kutu grafiği şu şekilde oluşturulur:

- Örnek medyan Q2'nin yüksekliğine bir çizgi koyulur.
- Birinci çeyrek Q1'den üçüncü çeyrek Q3'e bir kutu çizilir.
- Aralık (çit) dışındaki tüm noktaları aykırı değer olarak sınıflandırılır ve grafik üzerinde işaretlenir. Aralik dışındaki değerler Eşitlik 1 kullanılarak belirlenir.
  $[Q1 - 1.5\,IQR;\ Q3 + 1.5\,IQR]$ (1)
- Bıyıklar çizilir.

Böylece, kutu grafiği, medyan ve çeyrekler arası aralık aracılığıyla verilerin yeri ve yayılımı hakkında bilgi gösterir. Kutunun her iki tarafındaki bıçağın uzunluğu ve kutu içindeki medyanın konumu, verilerdeki olası çarpıklıkların tespit edilmesinde yardımcı olur. Son olarak, çitler dışında kalan gözlemler aykırı olarak belirlenir (Hubert ve Vandervieren, 2008).




**Şekil 1. Kutu grafiği**

### 3.2. Naïve Bayes Sınıflandırıcısı

Naive Bayes algoritması, belirli bir veri kümesindeki değerlerin frekanslarını ve kombinasyonlarını sayarak bir olasılık kümesini hesaplayan basit bir olasılıksal sınıflandırıcıdır. Algoritma, Bayes teoremini kullanmakadır ve sınıf değişkeninin değeri göz önüne alındığında tüm özniteliklerin bağımsız olduğunu varsaymaktadır (Patil, 2013). Basitliği ve gerçekçi olmayan bağımsızlık varsayımına rağmen, NB sınıflandırıcısının performansı uygulamalarda oldukça başarılıdır (Bermejo ve diğerleri, 2014).

D, sınıf etiketlerini de içeren bir eğitim kümesi ve her bir örnek n nitelikli bir vektör ile sunulmaktadır $(x_1, x_2, \ldots x_n)$ $C_1, C_2 \ldots . C_m$ olmak üzere m sınıf olduğunu varsayılır. Sınıflandırmada amaç, maksimum$P(C_i|X)$ 'i elde etmektir. Bu bayes teoreminden türetilir (Eşitlik 2) (Han ve diğerleri, 2011).

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)} \quad (2)$$

$P(X)$ tüm sınıflar için sabit olduğundan, yalnızca $P(C_i|X) = P(X|C_i) \times P(C_i)$ ifadesinin en büyüklenmesi gerekmektedir. Nitelikler birbirinden bağımsız olmasından dolayı Eşitlik 3 kullanılabilir (Han ve diğerleri, 2012).

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (3)$$

Bu denklem, hesaplama maliyetini büyük oranda düşürür. Yalnızca, sınıf dağlımı sayılır. Eğer $A_k$ kategorisel ise, $P(x_k|C_i)$ $A_k$ için $x_k$ değerine sahip $C_i$ sınıfındaki örneklerin sayısının eğitim kümesi D'deki $C_i$ sınıfının örnek sayısına bölünmesiyle elde edilir. Eğer $A_k$ sürekli bir değer ise, $P(x_k|C_i)$ genellikle ortalaması μ ve standart sapması σ olan Gaus dağılımı temelinde hesaplanmaktadır (Han ve diğerleri, 2012).

### 3.3. C4.5 Sınıflandırıcısı

C4.5 algoritması, çok basit bir karar ağacı algoritması olan ID3 algoritmasına dayanmaktadır. Bu algoritma karar ağacından geçerek, her düğümü ele alarak en uygun bölünmeyi seçmektedir (Masetic ve diğerleri, 2016).

*C*, sınıf sayısını gösterdiğinde. *p (S, j)*, *S* niteliğindeki *j*. sınıfa atanan örneklerin oranıdır. Bu nedenle, *S* niteliği entropisi Eşitlik 4'teki gibi hesaplanır (Dai ve Ji, 2014).

$$Entropi(S) = -\sum_{j=1}^{C} p(S,j) \times log\, p(S,j) \quad (4)$$

Buna göre, bir eğitim veri seti T'nin bilgi kazancı Eşitlik 5'teki gibi tanımlanır (Dai ve Ji, 2014).

$$Kazanç(S,T) = Entropi(S) - \sum_{v\in(T_s)} \frac{|T_{S,v}|}{|T_S|} Entropi(S_v) \quad (5)$$

Burada Değerler ($T_s$), T'deki S'nin değerlerinin kümesidir, $T_s$, S tarafından uyarılmış T'nin alt kümesidir ve $T_{s,v}$ ,S'nin bir v değerine sahip olduğu T'nin alt kümesidir (Dai ve Ji, 2014). Bu nedenle, S niteliğinin bilgi kazancı oranı Eşitlik 6'daki tanımlanır (Dai ve Ji, 2014).

$$KazançOranı(S,T) = \frac{Kazanç(S,T)}{BölünmeBilgisi(S,T)} \quad (6)$$

Bölünme Bilgisi $(S,T)$ Eşitlik 7'deki gibi hesaplanır (Dai ve Ji, 2014). Bilgi Kazancı Oranı büyük olan değer seçilip bölünme gerçekleşir.

$$Bölünme\ Bilgisi(S,T) = -\sum_{v\in(T_S)} \frac{|T_{S,v}|}{|T_S|} \times log \frac{|T_{S,v}|}{|T_S|} \quad (7)$$

### 3.4. K-En Yakın Komşu Sınıflandırıcısı (K-NN)

K-en yakın komşu sınıflandırıcısı, her nesnenin C sınıfı, sınıflardan bir c grubuna ait olduğu bilinen bir eğitim veri matrisi olan X'i kullanır. Bu sınıflandırıcı, bilinmeyen bir xt nesnesini, k-en yakın komşularının ait olduğu sınıfa atar. Bu komşular uygun bir metrik, genellikle Öklid mesafesine (Eşitlik 8) göre bulunur (Medina ve diğerleri, 2009). Burada $d(i,j)$, öklid uzaklığı ve $n$, örneklerin nitelik sayısıdır.

$$d(i,j) = \sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2 \quad (8)$$

Algoritmadaki k değeri önceden seçilir; değerinin yüksek olması birbirlerine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyse birbirine benzediği yani aynı sınıfın noktalan oldukları halde, bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur. Tipik k değerleri 3, 5 ve 7'dir (Silahtaroğlu, 2008: 118).

### 3.5. Sıralı Minimal Optimizasyon (SMO)

Sıralı Minimal Optimizasyon (SMO), Destek Vektör Makineleri (SVM'ler) için yeni bir algoritmadır. John Platt tarafından 1998 yılında önerilen Sıralı Minimal Optimizasyon (SMO) algoritması, SVM eğitimi için basit ve hızlı bir yöntemdir. Ana fikir, her iterasyonda iki eleman dahil olmak üzere minimal alt kümeyi optimize ederek çift kuadratik optimizasyon problemini çözmekten türetilmiştir. SMO'nun avantajı, basit ve analitik olarak uygulanabilmesidir (Chaurasia ve Pal, 2017).

İki sınıfa ait eğitim vektörlerinin setini ayırma problemini göz önünde bulundurun: $D = \{(x_i, y_i)\} l_i = 1$, burada $l$, eğitim örneklerinin sayısıdır, $x_i \in R$ d, $i.$ eğitim örneği ve $y \in \{+1, -1\}$, xi sınıfının etiketidir. SVM, Eşitlik 9-11'deki optimizasyon sorununun çözümünü gerektirir (Zhang ve diğerleri, 2018).

$$min\, w(\alpha) = \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right) - \sum_{i=1}^{l} \alpha_i \quad (9)$$

$$s.t \qquad \sum_{i=1}^{l} y_i \alpha_i = 0 \quad (10)$$

$$0 \le \alpha_i \le C,\ i = 1, \dots, l \quad (11)$$

Sıralı Minimal Optimizasyon (SMO), SVM'yi hızlıca çözebilen basit bir algoritmadır. Yakınsama sağlamak için büyük QP problemini bir dizi mümkün olan en küçük QP alt problemine böler. Her adımda, SMO, ortaklaşa optimize etmek için iki Lagrange çarpanını seçer, bu çarpanlar için en uygun değerleri belirler ve yeni optimal değerleri yeniden tanımlamak için SVM'yi günceller. İki seçilmiş değişkeninin $\alpha_i$, $\alpha_j$ olduğunu varsayalım, sonra yukarıdaki matematiksel model Eşitlik 12-14'teki gibi yazılabilir (Zhang ve

diğerleri, 2018). Orijin çözümü ($\alpha_1^0$, $\alpha_2^0$) ise, en uygun çözüm Eşitlik 15'teki gibi uygulanabilir (Zhang ve diğerleri, 2018).

$$min\, w(\alpha_i, \alpha_j) = \frac{1}{2}K_{11}\alpha_1^2 + \frac{1}{2}K_{22}\alpha_2^2 + y_1y_2K_12\alpha_1\alpha_2 - (\alpha_1 + \alpha_2) + y_1\alpha_1\sum_{i=3}^{l} y_i\alpha_1K_{i1} + y_2\alpha_2\sum_{i=3}^{l} y_i\alpha_1K_{i2} \quad (12)$$

$$s.t \quad y_1\alpha_1 + y_2\alpha_2 = -\sum_{i=3}^{l} y_i\alpha_i = c \quad (13)$$

$$0 \le \alpha_i \le C,\ i = 1, \dots, l \quad (14)$$

$$\begin{cases} \alpha_2^n = \alpha_2^0 + \frac{y_1y_2\nabla w(\alpha)_1 - w(\alpha)_2}{K_{11}+K_{22}-2K_{12}} \quad (U \le \alpha_2^n \le V) \\ \alpha_1^n = \alpha_1^0 + y_1y_2(\alpha_2^0 - \alpha_2^n) \\ if\ y_1 \ne y_2: \\ U = maks(0, \alpha_2^0 - \alpha_1^0),\ V = maks(C, C + \alpha_2^0 - \alpha_1^0) \\ if\ y_1 = y_2: \\ U = maks(0, \alpha_2^0 + \alpha_1^0 - C),\ V = maks(C, \alpha_2^0 + \alpha_1^0) \end{cases} \quad (15)$$

### 3.6. Bayes Ağları Sınıflandırıcısı

Bayes ağı, bir dizi rasgele değişkenleri temsil eden olasılıksal bir grafik modelidir. Bir Bayes ağı, yeni olaylarda akıl yürütme yaparken karar tablolarında bilgi akışını modeller. Gerekli olasılıklar, karar değerinin döngü içinde yer almasıyla doğrudan eğitim verilerinden hesaplanabilir. Bu sınıflandırıcı, bir istatistiksel öğrenme algoritmasının en iyi bilinen temsilidir (Narudin ve diğerleri, 2016).

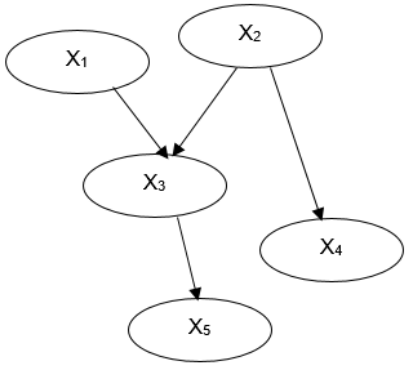
Şekil 2'de, bir Yönlü Düz Ağaçlar (DAG) kullanılarak resmedilen basit bir örnek gösterilmiştir. Şekil 1'de, düğümler ($X_1, X_2, X_3, X_4, X_5$) rasgele değişkenleri temsil eder ve yönlendirilmiş bağlantılar (yaylar), aralarında doğrudan olasılıksal bağımlılıkları temsil eder. Her bir yay bir ana düğümden başlar ve $X_2$ gibi bir çocuk düğümde biter. Şekil 2'de herhangi bir giriş yayı olmayan düğüm, örneğin $X_1$ ve $X_2$, kök düğümü olarak adlandırılır. Parametreler, düğümler arasındaki kantitatif olasılıksal ilişkileri temsil etmek için kullanır. Bayes ağının parametreleri, her bir durum için her bir kök düğümün önceki olasılığını ve ebeveynlik durumları verilen her bir çocuk düğümünün koşullu olasılık tablosunu (CPT) içerir. Gerekli olan parametre sayısının, düğüm noktalarının sayısıyla üssel olarak büyümesi sonucu oluşan karmaşıklığın üstesinden gelmek için, üç bağımsızlık varsayımı getirilmiştir: (i) tüm kök düğümler birbirinden bağımsızdır, (ii) iki düğüm ortak ana düğümlere sahipse ve aralarında hiçbir yönsel yay bulunmadıklarında, birbirlerine yakın ebeveynlerinin durumları göz önüne alındığında şartlı olarak birbirlerinden bağımsızdırlar; (iii) herhangi bir kök dışı düğüm için, tüm ana orta düğümlerin durumları göz önünde bulundurulduğunda bu, doğrudan kendi ana düğümlerinde koşullu olarak bağımsızdır. Bu bağımsız değerlendirmelere dayanarak, ortak olasılık dağılımı Eşitlik 16'daki gibidir (Wang ve diğerleri, 2017). Burada Pa$_i$, düğüm $X_i$'nin hemen tüm ana düğümleridir ve $k$, düğümlerin sayısıdır (Wang ve diğerleri, 2017).

$$P(X_1, X_2 \dots X_k) = \prod_{i-1}^{k} P(X_i|Pa_i) \quad (16)$$

Bayes ağına bir bulgu girildiğinde, bulgularla verilen bilgiler, bilgiyi güncellemek ve gözlemlenmeyen durumların bir test sonrası olasılıkları elde etmek için ağda yayılır. Bu sürece çıkarım denir. Bayes ağında çıkarım, Bayes teoremine dayanmaktadır. Bayes formülü Eşitlik 17'deki formda yazılabilir (Wang ve diğerleri, 2017):

$$P(B_i \backslash A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(B_i)P(A \backslash B_i)}{\sum_{i-1}^{k} P(B_i)P(A \backslash B_i)} \quad (17)$$

Eşitlik 13'te sağ tarafındaki öğeler test öncesi olasılığı ve soldaki öge test sonrası olasılığıdır. Bayes teoremi, önceki olasılıklardan sonraki olasılığın hesaplama yöntemini sağlar. Bayes çıkarımı hataları teşhis etmek için kullanıldığında, B$_i$ bir hatayı ve A bir hata bulgusunu temsil eder. $B_i(P(B_i))$ hatasının önceki olasılığı ve verilen bir $B_i$ ($(P(A \backslash B_i)$ bulgusunun koşullu olasılığı elde edildikten sonra, test sonrası olasılık $P(B_i \backslash A)$ Eşitlik 13'teki gibi hesaplanabilir (Wang ve diğerleri, 2017).




**Şekil 2. Bayes ağları**

### 3.7. Karışıklık Matrisi

Karışıklık matrisi, bir sınıflandırma sistemi tarafından yapılan gerçek ve tahmin edilen sınıflandırmalar hakkında bilgi içeren makine öğreniminden bir kavramdır. Bir karışıklık matrisi iki boyutlu olup; bir boyutu bir nesnenin gerçek sınıfı tarafından indekslenirken, diğeri sınıflandırıcının öngördüğü sınıf tarafından indekslenir. Tablo 1, $A_1$, $A_2$ ve $A_n$ sınıfları ile çok sınıflı bir sınıflandırma işlemi için temel karışıklık matrisini göstermektedir. Karışıklık matrisinde $N_{ij}$, gerçekte $A_i$ sınıfına ait olan ancak $A_j$ sınıfı olarak sınıflandırılan örneklerin sayısını temsil eder (Deng ve diğerleri, 2016).

**Tablo 1.** Karışıklık matrisi

| | | Tahmin Edilen | | |
|---|---|---|---|---|
| | | $A_1$ | … $A_j$ … | $A_n$ |
| Gerçekleşen | $A_1$ | $N_{11}$ | $N_{1j}$ | $N_{1n}$ |
| | . | | . | |
| | . | | . | |
| | . | | . | |
| | $A_i$ | | … $N_{ij}$ … | $N_{in}$ |
| | . | | . | |
| | . | | . | |
| | . | | . | |
| | $A_n$ | $N_{n1}$ | $N_{nj}$ | $N_{nn}$ |

### 3.8. Performans Ölçütleri

Tablo 2'deki gibi iki sınıfın tahmin sonuçlarına ait bir karışıklık matrisimiz olsun. Karışıklık matrisine dayanan bir dizi sınıflandırma performansı tanımlanabilir. Bazı yaygın ölçütler Eşitlik 18-22'deki gibi verilmiştir. Bu eşitliklerde; TP doğru pozitifi, FN, yanlış negatifi, FP yanlış pozitifi, TN doğru negatifi, P' pozitif değerleri ve N' negatif değerleri temsil etmektedir.

**Tablo 2. İki sınıflı karışıklık matrisi**

| | | *Tahminlenen sınıf* | |
|---|---|---|---|
| | | *Evet* | *Hayır* |
| Gerçek Sınıf | Evet | TP | FN |
| | Hayır | FP | TN |
| | Toplam | P' | N' |

$$Doğruluk = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (18)$$

$$Kesinlik = \frac{TP}{TP+FP} \quad (19)$$

$$Anma = \frac{TP}{TP+FN} \quad (20)$$

$$F\ ölçütü = \frac{(1+\beta^2)\times Anma\times kesinlik}{\beta^2\times Anma\times kesinlik} \quad (21)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (22)$$

Bir başka sıklıkla kullanılan performans ölçütü ise Kappa istatistiğidir. Sınıflama düzeyinde puanlama yapan iki puanlayıcı arasındaki uyumun derecesini belirlemek için geliştirilmiştir (Bilgen ve Doğan, 2017). $P_o$ kabul edilen oran, $P_c$ kabul edilmesi beklenen oran olmak üzere Kappa değeri Eşitlik 23'teki gibi hesaplanır (Nizam ve Akın, 2014):

$$K = \frac{P_o - P_c}{1 - P_c} \quad (23)$$

Bazı araştırmacılar, dengesiz veri kümeleri üzerinde oluşturulan tahmin sistemlerinin değerlendirilmesi için G-ortalama1 ve G-ortalama2 kullanmaktadır. Eşitlik 24 ve 25, sırasıyla bu ölçümlerin nasıl hesaplanacağını gösterir (Catal, 2012).

$$G\ ortalama1 = \sqrt{(Anma \times Kesinlik)} \quad (24)$$

$$G\ ortalama2 = \sqrt{(Anma * TNR)} \quad (25)$$

p gerçek değer ve a tahmin değeri olsun. Ortalama Mutlak Hata (MAE), Ortalama Hata Kareleri Kökü (RMSE), Bağıl Ortalama Hata (RAE) ve Kök Bağıl Hata Kareleri (RRSE) ölçütleri Eşitlik 26-29'daki gibi tanımlanmıştır;

$$MAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \quad (26)$$

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \quad (27)$$

$$RAE = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots |a_n - \bar{a}|} \quad (28)$$

$$RRSE = \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}} \quad (29)$$

Sınıflandırma modellerinin karşılaştırılması için sentez indeks (SI) değerleri hesaplanmıştır. SI değeri Eşitlik 30'daki gibi hesaplanabilir.

$$\sum SI = (1 - Hata\ SI) + Performans\ SI \quad (30)$$

Hata SI ve Performans SI, her bir ölçüt için normalizasyon işlemi yapılarak hesaplanır. Burada Hata SI değeri 4 hata ölçütü; MAE, RMSE, RAE ve RRSE'nin normalize değerlerinin ortalaması alınarak bulunur. Daha sonra modellerin Hata SI değerleri 0-1 arasında normalize edilir. Performans SI ise Doğruluk, Kappa istatistiği, TP oranı, FP oranı, Kesinlik, Anma, F-ölçütü, MCC, ROC Alanı, PRC Alanı, G-ortalama1 ve G-ortalama2 ölçütlerinin normalize değerlerinin ortalaması alınarak bulunur. Daha sonra modellerin Performans SI değerleri 0-1 arasında normalize edilir. Normalizasyon işlemi Eşitlik 31'deki gibidir. Burada $m$, ölçüt sayısı iken; $P_i$, $P$. modelin I ölçüt değeri, $P_{min,i}$ modeller arasında $i$.ölçütün en düşük değeri ve $P_{maks,i}$ modeller arasında $i$.ölçütün en yüksek değeridir.

$$SI = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{P_i - P_{min,i}}{P_{maks,i} - P_{min,i}} \right) \quad (31)$$

### 3.9. k-Kat Çapraz Doğrulama

k-kat çapraz doğrulama yöntemi, toplam n örneğin bulunduğu bir veri setinde, her birinde $n/k$ örneğin bulunduğu $k$ adet ayrık parçaya ayrılır. Her seferinde farklı bir veri seti kümesi test için ayrılarak kalan diğer $k-1$ veri seti eğitim için kullanılır. Her defasında test kümesi değiştirilerek sınıflandırıcı $k$ defa eğitilir. Bu şekilde elde edilen $k$ adet hatanın ortalaması ile sınıflandırıcı performansı tahmin edilmiş olur (Narin ve diğerleri, 2014).

### 3.10. Yığma Tekniği

Yığma tekniği, karar ağaçları, örnek tabanlı öğrenenler ve bunun gibi farklı makine öğrenimlerinin sınıflandırıcılarını birleştirmek için en çok kullanılan tekniklerden biridir. Her biri farklı bilgi gösterimi ve farklı öğrenme eğilimleri kullandığından, hipotez alanı farklı bir şekilde araştırılacak ve farklı sınıflandırıcılar elde edilecektir. Dolayısıyla, hatalarının ilişkilendirilmemesi ve sınıflandırıcıların kombinasyonunun temel sınıflandırıcılardan daha iyi performans göstermesi beklenmektedir. Yığma konusundaki erken araştırmalar, doğru sınıflandırıcıları, parametrelerini ve meta sınıflandırıcıları seçmenin ana darboğaz olduğunu gösterdi. Bu konuyla ilgili araştırmaların çoğu, sınıflandırıcıların ve parametrelerinin doğru kombinasyonunu elde etmek üzerine yapılmıştır (Ledezma ve diğerleri, 2004).

### 3.11. Oylama Tekniği

Oylama tekniği, çoklu sınıflandırıcı kararlarını birleştirmek için kullanılan bir birleştirme tekniğidir. Çoğulluğa veya oy çoğunluğuna dayanan en basit şekliyle, her bir sınıflandırıcı tek bir oylamaya katkıda bulunur. Toplam tahmini, oyların çoğunluğu tarafından belirlenir, yani, en çok oyu alan sınıf kesin tahmin olarak belirlenir (Paris ve diğerleri, 2010).

### 3.12. Veri Seti Tanımı

Kaliforniya Üniversitesi makine öğrenmesi veri setlerinden Bank Marketing verisi uygulama için kullanılmıştır. Bu veriler Portekiz Bankacılık Kurumu'nun Mayıs 2008 ile Kasım 2010 arasında elde ettiği mobil ve sabit telefonlara dayalı doğrudan pazarlama kampanyaları ile ilgilidir. Bu verilerin sınıflandırılmasında hedef müşterinin vadeli mevduat uygunluğunun tahmin edilmesidir. Banka vadeli mevduat ürünü verilip verilemeyeceğinin anlaşılması için genellikle aynı müşteriyle birden fazla telefon bağlantısı gerçekleştirmiştir. Veri seti 20 adet girdi ve 1 adet çıktı verisi bulunmaktadır. Sınıf etiketi müşteriye vadeli mevduat verilip verilmediği bilgisinden oluşmaktadır. Toplamda 41188 müşteri verisi kaydedilmiştir. Elde edilen verilere göre sınıf etiketi bilgisi 12217 ‘evet’ ve 2783 ‘hayır’ şeklindedir. Veri setindeki öznitelikler Tablo 3’te gösterilmiştir.

**Tablo 3. Veri seti**

| *Öznitelik Adı* | *Tanım* | *Veri tipi* | *Veriler* |
|---|---|---|---|
| Age | Müşterinin yaşı | Nümerik | |
| Job | Müşterinin işi | Kategorik | Yönetici, Mavi Yakalı, Girişimci, Serbest Meslek Sahibi, Hizmette Görevli, Öğrenci, Hizmetçi, Yönetimde Görevli, Emekli, Teknisyen, İşsiz, Belirsiz |
| Marital | Medeni hal | Kategorik | Boşanmış, Bekar, Evli, Belirsiz |
| Education | Eğitim seviyesi | Kategorik | Temel 4y, Temel 6y, Temel 9y, Lise, Okuma Yazma Bilmeyen, Profesyonel Kurs, Üniversite Diploması, Belirsiz |
| Default | Ödenmemiş kredisi var mı? | Kategorik | Hayır, Evet, Belirsiz |
| Housing | Konut kredisi var mı? | Kategorik | Hayır, Evet, Belirsiz |
| Loan | Bireysel kredisi var mı? | Kategorik | Hayır, Evet, Belirsiz |
| Contact | Müşteriyle temas şekli | Kategorik | Cep Telefonu, Sabit Telefon |
| Month | Müşteriyle kurulan son temas (ay) | Kategorik | Ocak, Şubat, Mart, Nisan, Mayıs, Haziran, Temmuz, Ağustos, Eylül, Ekim, Kasım, Aralık |
| Day Of Week | Kurulan son temasın günü | Kategorik | Pazartesi, Salı, Çarşamba, Perşembe, Cuma |
| Duration | Kurulan temas süresi | Nümerik | |
| Campaign | Müşteriye mevcut kampanya boyunca yapılan temas sayısı | Nümerik | |
| Pdays | Müşteriye mevcut kampanyadan önce yapılan temas sayısı | Nümerik | |
| Previous | Bir önceki pazarlama kampanyasının sonucu | Nümerik | |
| Poutcome | Toplam güne tanımlanan oran günleri | Kategorik | Başarısız, Başarılı, Yok |
| Empvarrate | İstihdam değişim oranı (çeyreklik gösterge) | Nümerik | |
| Conspriceidx | Tüketici fiyat endeksi (aylık gösterge) | Nümerik | |
| Consconfidx | Tüketici güven endeksi (aylık gösterge) | Nümerik | |
| Euribor3m | Euribor 3 aylık oranı (günlük göstergesi) | Nümerik | |
| Nremployed | Çalışan sayısı (üç aylık gösterge) | Nümerik | |

## 4. BULGULAR

Vadeli mevduat ürünü verilip verilemeyeceğinin tahminlenmesi için öncelikle veri setine kutu grafiği yöntemi ile aykırı değer analizi uygulanmıştır. Kutu grafiği yöntemiyle elde edilen sonuçlara göre, 41188 veriden 7612'si aykırı değer olarak tespit edilmiştir. Bu aykırı verilerin çoğu, mevcut kampanyadan önce temasın yapıldığı müşterilerden oluşmaktadır. Daha sonra C4.5, Naive Bayes, Bayes ağları, k-en yakın komşu ve Sıralı Minimal Optimizasyon (SMO) sınıflandırma algoritmaları kullanılarak aykırı değerden arındırılımış banka verileri sınıflandırılmıştır. Sınıflandırma için 10 kat çapraz doğrulama yöntemi kullanılmıştır. Sınıflandırma sonucu elde edilen karışıklık matrisi Tablo 4'teki gibidir.

Tablo 4'teki gibi elde edilen karışıklık matrisi sonuçlarına göre Tablo 5'teki belirtilen performans ölçütleri sonuçları elde edilmiştir. Doğru sınıflandırma oranı en yüksek C4.5 algoirtması ile bulunmuştur. Diğer yandan sınıflandırma güvenini ve sınıflandırılan verilerin gerçekle ne kadar örtüştüğünü ifade eden Kappa istatistiği ile de aynı durum söz konusudur. SMO algoritması ile kurulan modelde daha düşük ortalama mutlak hata değeri elde edilirken ortalama hata kareleri kökü ölçütü açısından C4.5 algoritması ile daha iyi sonuç elde edilmiştir. Bu bulgular doğrultusunda basit C4.5 ile kurulan sınıflandırma modelinin diğer basit modellere göre daha yüksek bir başarı sağladığı görülmektedir.

Daha sonra çoklu tahmin modellerinden gelen bilgileri birleştiren Oylama ve Yığma model oluşturma teknikleri uygulanarak yeni sınıflandırma modelleri oluşturulmuştur. Bu modeller, Tablo 5'teki performans ölçütlerine bakılarak seçilmiştir. Yığma ve Oylama teknikleri ile elde edilen karışıklık matrisi sonuçları sırasıyla Tablo 6 ve 7'deki gibidir.

**Tablo 4. Karışıklık matrisi (M1-M5)**

| *Sınıflandırma Modelleri* | *M1* | | *M2* | | *M3* | | *M4* | | *M5* | |
|---|---|---|---|---|---|---|---|---|---|---|
| Algoritmalar | Naive Bayes | | Bayes Ağları | | C4.5 | | k-NN | | SMO | |
| Sınıflar | a | b | a | b | a | b | a | b | a | b |
| a=hayır | 26518 | 4488 | 27799 | 3207 | 30218 | 788 | 29717 | 1289 | 30998 | 8 |
| b=evet | 792 | 1778 | 989 | 1581 | 1590 | 980 | 1857 | 713 | 2570 | 0 |

**Tablo 5. Performans ölçütleri (M1-M5)**

| *Performans Ölçütleri \ Sınıflandırma Modeli* | *M1* | *M2* | *M3* | *M4* | *M5* |
|---|---|---|---|---|---|
| Doğru Sınıflandırma Oranı (%) | 84,27 | 87,50 | 92,91 | 90,63 | 92,32 |
| Kappa İstatistiği | 0,3297 | 0,3666 | 0,4153 | 0,2625 | -0,0005 |
| MAE | 0,1582 | 0,1311 | 0,0916 | 0,0937 | 0,0768 |
| RMSE | 0,351 | 0,3363 | 0,2318 | 0,3061 | 0,2771 |
| RAE (%) | 111,89 | 92,75 | 64,79 | 66,28 | 54,30 |
| RRSE (%) | 132,02 | 126,48 | 87,16 | 155,13 | 104,22 |
| TP Oranı | 0,843 | 0,875 | 0,929 | 0,906 | 0,923 |
| FP Oranı | 0,294 | 0,363 | 0,573 | 0,67 | 0,923 |
| Kesinlik | 0,918 | 0,917 | 0,92 | 0,896 | 0,853 |
| Anma | 0,843 | 0,875 | 0,929 | 0,906 | 0,923 |
| F-ölçütü | 0,871 | 0,892 | 0,923 | 0,901 | 0,887 |
| MCC | 0,373 | 0,389 | 0,424 | 0,265 | -0,004 |
| ROC Alanı | 0,884 | 0,867 | 0,873 | 0,618 | 0,500 |
| PRC Alanı | 0,941 | 0,940 | 0,935 | 0,881 | 0,859 |

Tablo 8'de, Tablo 6 ve Tablo 7'deki karşılaştırma matrislerinden hesaplanan performans ölçütleri verilmiştir. Toplamda 19 sınıflandırma modeli ile veriler sınıflandırılmıştır. Karışıklık matrisi tabloları incelendiğinde 'no' sınıfını en iyi tahminleyen modellerin M14 ve M15 olduğu görülüp; Şekil 3'teki gibi gösterilmiştir. Diğer bir sınıf etiketi olan 'yes' sınıfını en iyi tahminleyen model ise M1 modeli olup; Şekil 4'te gösterilmiştir.

**Tablo 6. Karışıklık matrisi (M6-M13)**

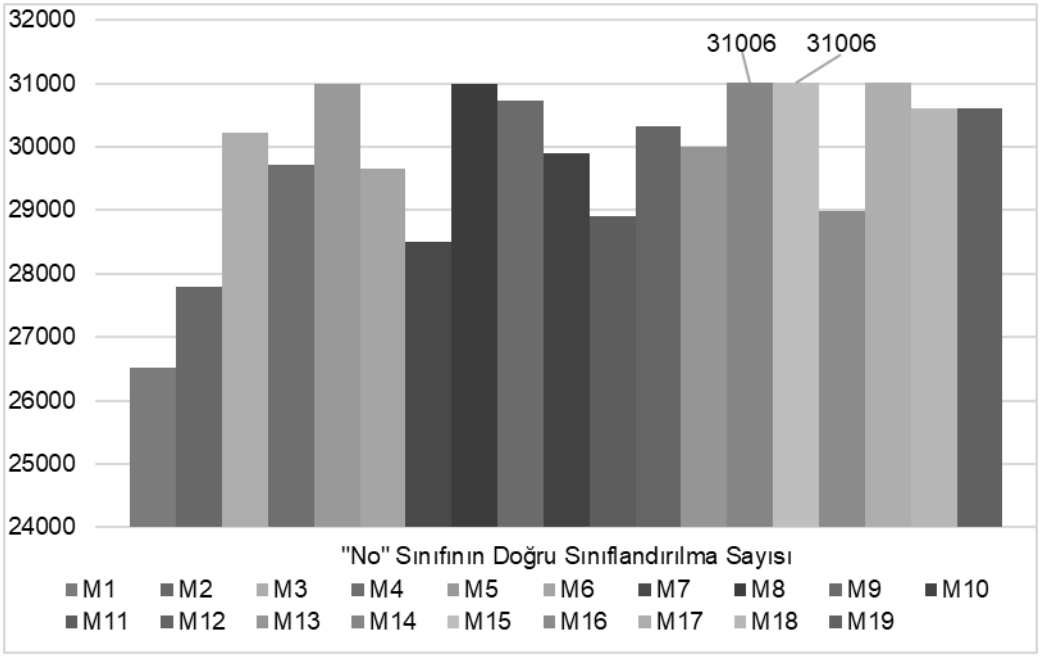
| *Sınıflandırma Modeli* | *M6* | | *M7* | | *M8* | | *M9* | | *M10* | | *M11* | | *M12* | | *M13* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Oylama | C4.5+k-NN | | C4.5+Bayes Ağ | | C4.5+SMO | | C4.5+k-NN+SMO | | C4.5 + k-NN + Bayes Ağ | | C4.5 + k-NN + Naive Bayes + Bayes Ağ | | C4.5 + k-NN + Bayes Ağ + SMO | | C4.5 + k-NN + Naive Bayes + Bayes Ağ + SMO | |
| Sınıflar | a | b | a | b | a | b | a | b | a | b | a | b | a | b | a | b |
| a=no | 29653 | 1353 | 28504 | 2502 | 30999 | 7 | 30732 | 274 | 29905 | 1101 | 28916 | 2090 | 30319 | 687 | 29996 | 1010 |
| b=yes | 1791 | 779 | 1089 | 1481 | 2570 | 0 | 2177 | 393 | 1545 | 1025 | 1173 | 1397 | 1986 | 584 | 1595 | 975 |

**Tablo 7. Karışıklık matrisi (M14-M19)**

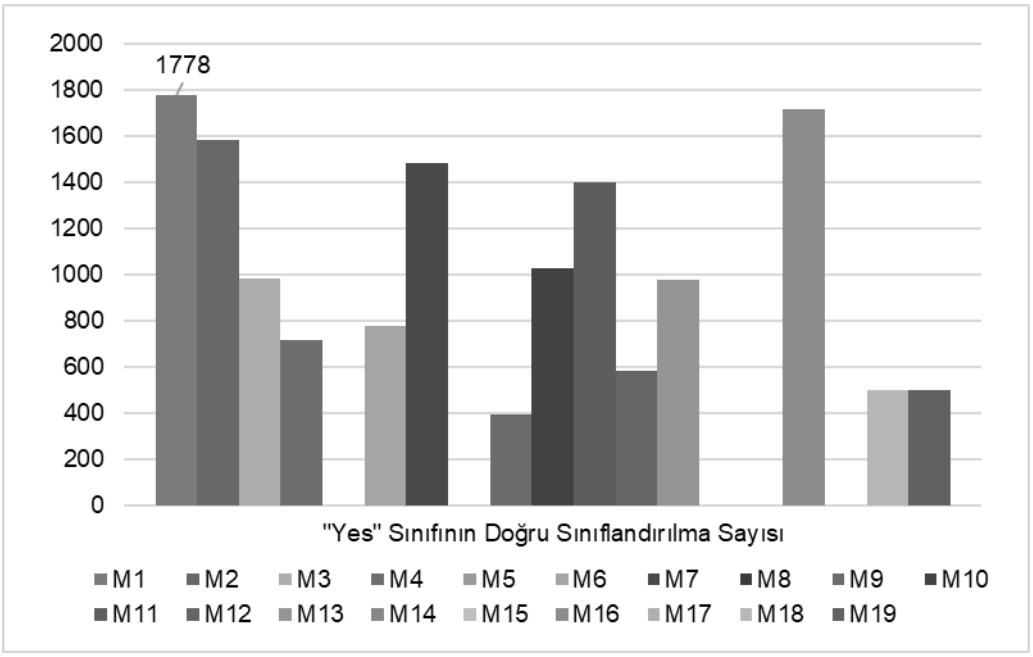
| *Sınıflandırma Modeli* | *M14* | | *M15* | | *M16* | | *M17* | | *M18* | | *M19* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Yığma | Ana Sınıflandırıcı: C4.5 Sınıflandırıcı:K-NN | | Ana Sınıflandırıcı: C4.5 Sınıflandırıcı: Bayes Ağ | | Ana Sınıflandırıcı: Bayes Ağ Sınıflandırıcı: C4.5 | | Ana Sınıflandırıcı: C4.5 Sınıflandırıcı:SMO | | Ana Sınıflandırıcı: C4.5 Sınıflandırıcı:K-NN + Naive Bayes + Bayes Ağ | | Ana Sınıflandırıcı: C4.5 Sınıflandırıcı:K-NN + Naive Bayes + Bayes Ağ + SMO | |
| Sınıflar | a | b | a | b | a | b | a | b | A | b | a | b |
| a=no | 31006 | 0 | 31006 | 0 | 28990 | 2016 | 31006 | 0 | 30601 | 405 | 30601 | 405 |
| b=yes | 2570 | 0 | 2570 | 0 | 854 | 1716 | 2570 | 0 | 2071 | 499 | 2071 | 499 |

**Tablo 8. Performans ölçütleri (M6-M19)**

| *Performan Ölçütleri\ Sınıflandırma Modeli* | *M6* | *M7* | *M8* | *M9* | *M10* | *M11* | *M12* | *M13* | *M14* | *M15* | *M16* | *M17* | *M18* | *M19* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Doğru Sınıflandırma Oranı (%) | 90,63 | 89,30 | 92,32 | 92,70 | 92,11 | 90,28 | 92,03 | 92,24 | 92,34 | 92,34 | 91,45 | 92,34 | 92,62 | 92,62 |
| Kappa İstatistiği | 0,2815 | 0,3958 | -0,0004 | 0,2182 | 0,3946 | 0,4092 | 0,267 | 0,3872 | 0 | 0 | 0,4992 | 0 | 0,2577 | 0,2577 |
| MAE | 0,0927 | 0,1114 | 0,0842 | 0,0874 | 0,1055 | 0,1187 | 0,0983 | 0,1103 | 0,1414 | 0,1414 | 0,1093 | 0,1414 | 0,1081 | 0,1081 |
| RMSE | 0,2441 | 0,2524 | 0,2361 | 0,24 | 0,2437 | 0,2562 | 0,2338 | 0,24 | 0,2659 | 0,2659 | 0,2664 | 0,2659 | 0,2334 | 0,2334 |
| RAE (%) | 65,54 | 78,77 | 59,54 | 61,79 | 74,61 | 83,93 | 69,53 | 78 | 100 | 99,98 | 77,31 | 99,98 | 76,45 | 76,45 |
| RRSE (%) | 91,8 | 94,9 | 88,7 | 90,2 | 91,6 | 96,3 | 87,9 | 90,3 | 100 | 100 | 100,2 | 100 | 87,8 | 87,8 |
| TP Oranı | 0,906 | 0,893 | 0,923 | 0,927 | 0,921 | 0,903 | 0,92 | 0,922 | 0,923 | 0,923 | 0,915 | 0,923 | 0,926 | 0,926 |
| FP Oranı | 0,647 | 0,397 | 0,923 | 0,783 | 0,558 | 0,427 | 0,715 | 0,576 | 0,923 | 0,923 | 0,312 | 0,923 | 0,745 | 0,745 |
| Kesinlik | 0,899 | 0,918 | 0,853 | 0,907 | 0,915 | 0,918 | 0,902 | 0,914 | 0,853 | 0,853 | 0,932 | 0,853 | 0,907 | 0,907 |
| Anma | 0,906 | 0,893 | 0,923 | 0,927 | 0,921 | 0,903 | 0,92 | 0,922 | 0,923 | 0,923 | 0,915 | 0,923 | 0,926 | 0,926 |
| F-ölçütü | 0,902 | 0,903 | 0,887 | 0,907 | 0,918 | 0,909 | 0,908 | 0,918 | 0,887 | 0,887 | 0,922 | 0,887 | 0,91 | 0,91 |
| MCC | 0,283 | 0,407 | -0,004 | 0,275 | 0,397 | 0,415 | 0,286 | 0,391 | 0,00 | 0,00 | 0,51 | 0,00 | 0,297 | 0,297 |
| ROC Alanı | 0,874 | 0,909 | 0,873 | 0,874 | 0,906 | 0,903 | 0,906 | 0,903 | 0,50 | 0,50 | 0,922 | 0,50 | 0,879 | 0,879 |
| PRC Alanı | 0,933 | 0,947 | 0,934 | 0,933 | 0,946 | 0,947 | 0,945 | 0,947 | 0,859 | 0,859 | 0,951 | 0,859 | 0,94 | 0,94 |

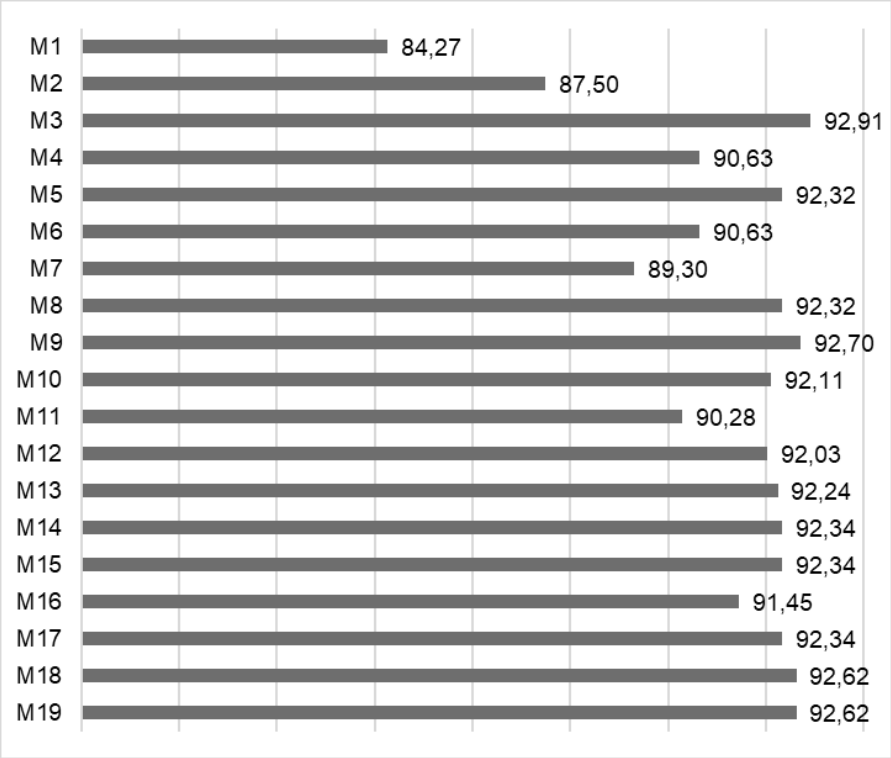



**Şekil 3. Modellerin "No" sınıfını sınıflandırma sayıları**



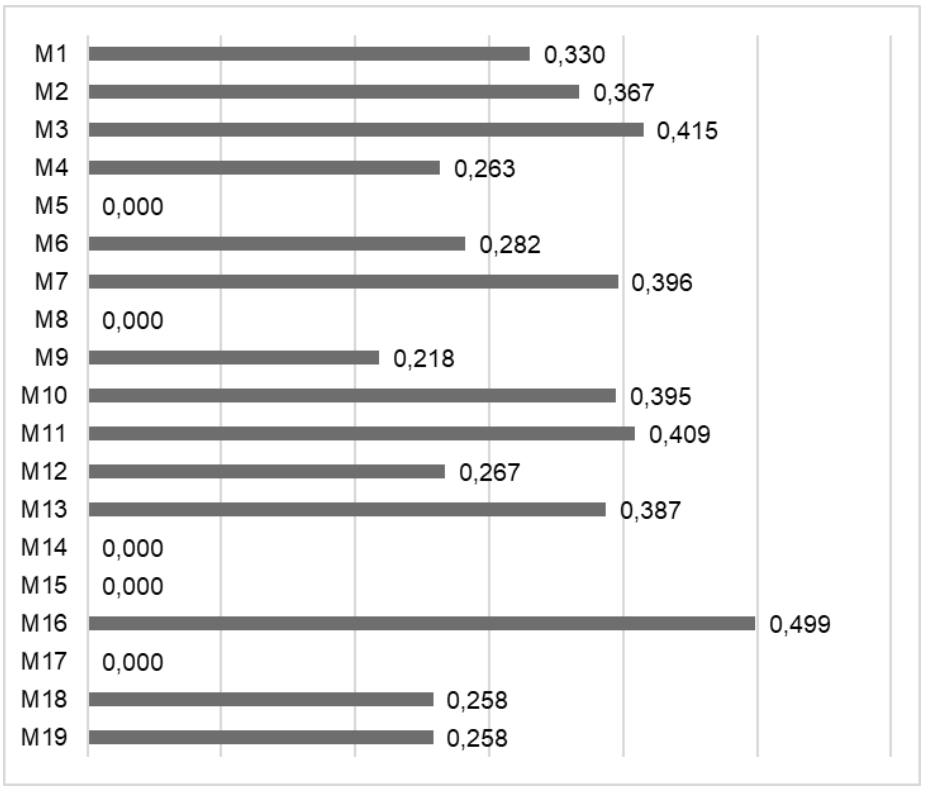

**Şekil 4. Modellerin "Yes" sınıfını sınıflandırma sayıları**

19 sınıflandırma modeli incelendiğinde en iyi doğruluk oranını veren modelin M3 modeli olduğu görülmektedir. Basit C4.5 sınıflandırma modeli dışındaki diğer basit modellerin oylama ve yığma tekniklerine göre daha düşük doğru sınıflandırma sonuçları verdiği belirlenmiştir. En düşük doğruluk oranını veren modelin basit Naïve Bayes olduğu bulunmuştur. Bütün modellerin doğruluk oranları Şekil 5'te gösterilmiştir.




**Şekil 5. Modellerin doğruluk oranları (%)**

İki puanlayıcı arasındaki uyumun derecesini belirleyen kappa istatistiği sonuçları incelendiğinde; M16 sınıflandırma modelinin 0,499 ile en iyi kappa istatistiğine sahip olduğu tespit edilmişitir. M16 modeli ile 'Yes' ve 'No' sınıf etiketlerinin doğru tahmin edilme oranları, diğer modellere göre daha dengelidir. M3 modeli ise 0,415 kappa istatistiği değeri ile ikinici en iyi model olduğu bulunmuştur. M5, M8, M14, M15 ve M17 modelleri ise 'Yes' sınıf etiketini doğru tahmin etme oranları 0 olmasından dolayı; kappa istatistikleri 0 olarak bulunmuştur. Bu modellerin sınıflandırma açısından bu veri seti için uygun olduğu söylenemez. Bütün modellerin kappa istatistiği değerleri Şekil 6'da gösterilmiştir.




**Şekil 6. Modellerin Kappa istatistiği**

**Tablo 9. Modellerin SI değerleri**

| *Sınıflandırma Modeli* | *Performans SI* | *Hata SI* | *Toplam SI* |
|---|---|---|---|
| M1 | 0,12 | 1,00 | 0,12 |
| M2 | 0,42 | 0,75 | 0,68 |
| M3 | 0,93 | 0,04 | 1,89 |
| M4 | 0,33 | 0,53 | 0,80 |
| M5 | 0,00 | 0,11 | 0,89 |
| M6 | 0,52 | 0,09 | 1,43 |
| M7 | 0,64 | 0,26 | 1,37 |
| M8 | 0,24 | 0,00 | 1,24 |
| M9 | 0,68 | 0,04 | 1,64 |
| M10 | 0,85 | 0,18 | 1,67 |
| M11 | 0,73 | 0,33 | 1,40 |
| M12 | 0,67 | 0,09 | 1,58 |
| M13 | 0,85 | 0,20 | 1,65 |
| M14 | 0,00 | 0,53 | 0,47 |
| M15 | 0,00 | 0,53 | 0,47 |
| M16 | 1,00 | 0,30 | 1,70 |
| M17 | 0,00 | 0,53 | 0,47 |
| M18 | 0,72 | 0,16 | 1,56 |
| M19 | 0,72 | 0,16 | 1,56 |

Modelleri karşılaştırmak için Performans SI ve Hata SI değerleri hesaplanmıştır (Tablo 9). Veri setinde her bir sınıf etiketinin sayısı birbirinden farklı olduğunda SI ölçütü, sınıflandırma modellerini karşılaştırmak için çok önemlidir. En iyi Performans SI değerine sahip model M16 olarak bulunurken, en iyi Hata SI değerine sahip model M8 olarak bulunmuştur. En iyi toplam SI değerinde sahip model M3 olarak bulunmuştur. En kötü toplam SI değerine sahip model ise M1 olarak bulunmuştur. Basit C4.5 sınıflandırma modeli veri setini sınıflandırmada daha iyi sonuçlar verdiği tespit edilmiştir.

Bu çalışmada elde edilen en iyi doğruluk oranı ile diğer çalışmalarda aynı veri seti kullanılarak elde edilen sonuçlar Tablo 10'daki gibi karşılaştırılmıştır. Çalışmada ortaya konan modelin diğer çalışmalardaki modellerden daha üstün olduğu gözlemlenmiştir.

**Tablo 10. Doğruluk oranlarının karşılaştırılması**

| *Karşılaştırma Kriterleri* | *Önerilen Yaklaşım* | *Koç ve Yeniay (2013)* | *Pradap ve Kemaludeen (2019)* | *Kim ve diğerleri (2015)* | *Palaniappan ve diğerleri (2017)* | *Bahari ve Elayidom (2015)* | *Popelka ve diğerleri (2012)* |
|---|---|---|---|---|---|---|---|
| Doğruluk Oranı (%) | 92,91 | 84,40 | 85,76 | 76,70 | 90,68 | 88,63 | 90,20 |
| En İyi Algoritma | Karar Ağacı C4.5 | Yapay Sinir Ağları | Rastgele Orman | Evrişimsel Sinir Ağları | Karar Ağacı C4.5 | Çok Katmanlı Algılayıcı Sinir Ağı | Karar Ağacı C4.5 |

## 5. SONUÇ

Teknolojinin gelişmesi ve bankacılık sektöründeki çarpıcı teknolojik değişiklikler, bankacılık işlem verilerinin önemini daha çok artırmıştır. Birçok banka, özel hizmetler sunmadan önce müşteri davranış ve taleplerini anlamak için müşterileri profillerine göre sınıflandırabilen yapay zekâ tekniklerini kullanmayı benimsemiştir. Çalışmada kullanılan veriler Portekiz Bankacılık Kurumu'nun Mayıs 2008 ile Kasım 2010 arasında elde ettiği mobil ve sabit telefonlara dayalı doğrudan pazarlama kampanyaları ile ilgilidir. Banka, vadeli mevduat ürünü verilip verilemeyeceğinin anlaması için genellikle aynı müşteriye birden fazla telefon bağlantısı yapması gerekmektedir. Bu verilerin sınıflandırılmasındaki hedef müşterinin vadeli mevduat uygunluğunun tahmin edilmesidir.

Literatürde farklı makine öğrenmesi algoritmaları kullanılarak veri seti üzerinde çalışmalar yapıldığı gözlemlenmiştir. Literatürde ilk defa bu çalışmada topluluk öğrenmesi yöntemleri kullanılarak vadeli mevduat hesabı için onay verecek olan ve onay vermeyecek olan müşterileri sınıflandırabilen bir model

geliştirilmiştir. Performans SI ve Hata SI gibi daha güvenilir performans ölçütleri kullanılarak makine öğrenmesi ve birleşik makine öğrenmesi yöntemleri karşılaştırılmıştır.

Çalışmada C4.5, Naive Bayes, Bayes ağları, k-en yakın komşu ve Sıralı Minimal Optimizasyon (SMO) sınıflandırma algoritmaları kullanılarak veriler sınıflandırılmıştır. Buna ek olarak oylama ve yığma topluluk öğrenmesi yöntemleriyle yeni hibrit modeler oluşturulmuştur. Toplamda 19 farklı sınıflandırma modeli ile veriler sınıflandırılmıştır. En iyi sınıflandırma modelini belirlemek için karşılaştırma ölçütü olarak Performans SI ve Hata SI değerleri hesaplanmıştır. Veri setinde her bir sınıf etiketinin sayısı birbirinden farklı olduğunda SI ölçütü, sınıflandırma modellerini karşılaştırmak için çok önemli bir kıyazlama ölçütüdür. Doğruluk oranı, tek başına sınıflandırma modelini değerlendirmek için yetersiz kalabilmektedir. Basit C4.5 sınıflandırma modeli, en iyi SI değerine sahip sınıflandırma modeli olarak bulunmuştur. Basit C4.5 sınıflandırma modeli ile verileri doğru sınıflandırma oranı %92,91 olarak elde edilmiştir, en kötü SI değerine sahip sınıflandırma modeli ise Basit Naïve Bayes sınıflandırma modelinin olduğu gözlemlenmiştir.

Bu çalışmada elde edilen en iyi doğruluk oranı ile diğer çalışmalarda aynı veri seti kullanılarak elde edilen sonuçlar karşılaştırılmıştır. Çalışmada ortaya konan modelin diğer çalışmalardaki modellerden daha üstün olduğu gözlemlenmiştir. Bu bağlamda bu çalışmada ortaya konan modellerin performanslarının iyi olduğu anlaşılmaktadır.

Literatürdeki mevcut çalışmalardan farklı olarak bu çalışmada, topluluk öğrenme yöntemleri ile farklı sınıflandırma modelleri oluşturulmuş ve sentez indeks olarak yeni bir performans ölçütü geliştirilmiştir. Geleceğe yönelik çalışmalarda derin öğrenme gibi daha farklı yöntemler geliştirilerek sınıflandırma performansı artırılabilir. Parametre optimizasyonu ve başlangıç çözümleri için hibrid çözümler ortaya konabilir. Bunun dışında bankacılık sektörü ile elde edilen verilerde farklı öznitelikler eklenip vadeli mevduat durumu sınıflandırılabilir.

## KAYNAKÇA


Abbas, S. (2015). "Deposit Subscribe Prediction Using Data Mining Techniques Based Real Marketing Dataset", arXiv preprint arXiv:1503.04344.

Bahari, T.F. ve Elayidom, M.S. (2015). "An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour", *Procedia Computer Science*, 46, 725-731.

Bermejo, P., Gámez, J.A. ve Puerta, J.M. (2014). "Speeding Up Incremental Wrapper Feature Subset Selection with Naive Bayes Classifier", *Knowledge-Based Systems*, 55, 140-147.

Bilgen, Ö.B. ve Doğan, N. (2017). "Puanlayıcılar Arası Güvenirlik Belirleme Tekniklerinin Karşılaştırılması", *Journal of Measurement and Evaluation in Education and Psychology*, 8(1), 63-78.

Catal, C. (2012). "Performance Evaluation Metrics for Software Fault Prediction Studies", *Acta Polytechnica Hungarica*, 9(4), 193-206.

Chaurasia, V. ve Pal, S. (2017). "A Novel Approach for Breast Cancer Detection Using Data Mining Techniques", *International Journal of Innovative Research in Computer and Communication Engineering,* 2(1), Ocak 2014.

Dai, W. ve Ji, W. (2014). "A Mapreduce Implementation of C4. 5 Decision Tree Algorithm", *International Journal of Database Theory and Application*, 7(1), 49-60.

Deng, X., Liu, Q., Deng, Y. ve Mahadevan, S. (2016). "An Improved Method to Construct Basic Probability Assignment Based on the Confusion Matrix for Classification Problem", *Information Sciences*, 340, 250-261.

Han, J., Pei, J. ve Kamber, M. (2011). Data Mining: Concepts and Techniques, Elsevier, DOI: 10.1016/B978-0-12-381479-1.00021-6.

Hubert, M. ve Vandervieren, E. (2008). "An Adjusted Boxplot for Skewed Distributions", *Computational Statistics and Data Analysis*, 52(12), 5186-5201.

Keles, A. ve Keles, A. (2015). "IBMMS Decision Support Tool for Management of Bank Telemarketing Campaigns", arXiv preprint arXiv:1511.03532.

Kim, K.H., Lee, C.S., Jo, S.M. ve Cho, S.B. (2015). "Predicting the Success of Bank Telemarketing Using Deep Convolutional Neural Network", *7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 314-317.

Koç, A.A. ve Yeniay, Ö. (2013). "A Comparative Study of Artificial Neural Networks and Logistic Regression for Classification of Marketing Campaign Results", *Mathematical and Computational Applications*, 18(3), 392-398.

Kozak, J. ve Juszczuk, P. (2018). "Ant Colony Optimization Algorithms in the Problem of Predicting the Efficiency of the Bank Telemarketing Campaign", *International Conference on Computational Collective Intelligence,* 335-344, Springer, Cham.

Ledezma, A., Aler, R., Sanchis, A. ve Borrajo, D, (2004). "Empirical Evaluation of Optimized Stacking Configurations", *16th IEEE International Conference on Tools with Artificial Intelligence,* 49-55.

Mašetic, Z., Subasi, A. ve Azemovic, J. (2016). "Malicious Web Sites Detection Using C4. 5 Decision Tree", *Southeast Europe Journal of Soft Computing*, 5(1), 68-72.

Medina, J.L.V., Boqué, R. ve Ferré, J. (2009). "Bagged K-Nearest Neighbours Classification with Uncertainty in the Variables", *Analytica Chimica Acta*, 646(1-2), 62-68.

Miguéis, V.L., Camanho, A.S. ve Borges, J. (2017). "Predicting Direct Marketing Response in Banking: Comparison of Class Imbalance Methods", *Service Business*, 11(4), 831-849.

Narin, A., İşler, Y. ve Mahmut, Ö. (2014). "Konjestif Kalp Yetmezliği Teşhisinde Kullanılan Çapraz Doğrulama Yöntemlerinin Sınıflandırıcı Performanslarının Belirlenmesine Olan Etkilerinin Karşılaştırılması", *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 16(48), 1-8.

Narudin, F.A., Feizollah, A., Anuar, N.B. ve Gani, A. (2016). "Evaluation of Machine Learning Classifiers for Mobile Malware Detection", *Soft Computing*, 20(1), 343-357.

Nizam, H. ve Akın, S.S. (2014). "Sosyal Medyada Makine Öğrenmesi ile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması", *XIX. Türkiye'de İnternet Konferansı*, 1-6, İzmir.

Palaniappan, S., Mustapha, A., Foozy, C.F.M. ve Atan, R. (2017). "Customer Profiling Using Classification Approach for Bank Telemarketing", *JOIV: International Journal on Informatics Visualization*, 1(4-2), 214-217.

Paris, I.H.M., Affendey, L.S. ve Mustapha, N. (2010). "Improving Academic Performance Prediction Using Voting Technique in Data Mining", *World Academy of Science, Engineering and Technology*, 62, 820-823.

Patil, T.R. (2013). "MSSS Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", *International Journal of Computer Science and Applications*, 6(2), 256-261.

Popelka, O., Hrebicek, J., Stencl, M. ve Trenz, O. (2012). "Comparison of Different Non-Statistical Classification Methods", *30th International Conference Mathematical Methods in Economics,* 727-732.

Pradap, R. ve Kamaludeen, P. (2019). "Machine Learning Modelsfor Bank Telemarketing Classification and Prediction", *The International Journal of Analytical and Experimental Modal Analysis*, 11(12), 962-967.

Silahtaroğlu, G. (2008). "Veri Madenciliği", Papatya Yayınları, İstanbul.

Türkmen, E. (2021). "Deep Learning Based Methods for Processing Data in Telemarketing-Success Prediction", *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV),* 1161-1166.

Vajiramedhin, C. ve Suebsing, A. (2014). "Feature Selection with Data Balancing for Prediction of Bank Telemarketing", *Applied Mathematical Sciences*, 8(114), 5667-5672.

Wang, Z., Wang, Z., He, S., Gu, X. ve Yan, Z.F. (2017). "Fault Detection and Diagnosis of Chillers Using Bayesian Network Merged Distance Rejection and Multi-Source Non-Sensor Information", *Applied Energy*, 188, 200-214.

Zhang, Q., Wang, J., Lu, A., Wang, S. ve Ma, J. (2018). "An Improved SMO Algorithm for Financial Credit Risk Assessment-Evidence from China's Banking", *Neurocomputing*, 272, 314-325.