# Yapay Sinir Ağları ve K-En Yakın Komşu Algoritmalarının Birlikte Çalışma Tekniği (Ensemble) ile Metin Türü Tanıma

**Mehmet Ali KUTLUGÜN[1], Mert Yılmaz ÇAKIR[2], Yrd.Doç.Dr. Farzad KIANI[3]**

[1,2] İstanbul Sabahattin Zaim Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul

[3] İstanbul Sabahattin Zaim Üniversitesi, Bilgisayar Mühendisliği Bölümü Bölüm Başkan Yardımcısı, İstanbul

mehmet.kutlugun@std.izu.edu.tr, mert.cakir@std.izu.edu.tr, farzad.kiani@izu.edu.tr

**Özet:** Günümüzde bazı problemlerin oldukça karmaşık olması, sıradan algoritmalarla çözümlenemeyecek derecelere ulaşabilmektedir. Bu sebeple daha zeki algoritmalara ihtiyaç duyulmaktadır. Metin tanıma, karakter tanıma, yüz tanıma, parmak izi tanıma gibi problemlerin çözümlerinde yapay sinir ağları sıklıkla kullanılmaktadır. Bu çalışmada metin türü belirleme işlemi için seçilen veri kümesi üzerinde önce k-en yakın komşu algoritması uygulanmış, elde edilen hatalı durumlar yapay sinir ağları ile tekrar ele alınarak birlikte çalışma tekniği (ensemble) uygulanmıştır. Bu sayede başarımı artıracak bir çözüm önerisi sunularak problem üzerinde değerlendirmeye gidilmiştir.

**Anahtar Sözcükler:** Yapay Sinir Ağları, Metin Tanıma, Yapay Zeka, K-En Yakın Komşu Algoritması, Sınıflandırma, Örüntü Tanıma

**Abstract:** Nowadays, some problems are so complicated that it may not be solved with ordinary algorithms. For this reason, more intelligent algorithms are needed. Artificial neural networks are frequently used in the solution of problems such as text recognition, character recognition, face recognition, fingerprint recognition. In this study, the k-nearest neighbors algorithm was applied first on the selected dataset for text type determination, and ensemble technique was applied re-handling the obtained errors with artificial neural networks. At this point, a solution proposal is presented that will increase performance and evaluated on the problem.

## 1. Giriş

Günümüz dünyasında, yapay zeka yaklaşımları en çok çalışılan konular haline gelmiştir. Bunun başlıca nedeni, problemlerin büyümesiyle birlikte karmaşıklıklarının ve beraberinde ortaya çıkan hatalar ile bu hataları düzeltme maliyetlerinin artmasıdır. [1][2].

Bazı projelerde ortaya çıkan hataların önceden tespit edilip düzeltilmesi, öngörülen maliyeti ve proje süresini aşma risklerini azaltır. Ortaya çıkması muhtemel hataları mümkün olduğu kadar erken tespit edebilmek için, verimli ve etkili bir test planının uygulanması gerekir. [3].

Karar verme amacıyla bilgisayarda geliştirilen algoritmalardan çevremizde gelişen tüm olaylara insan gibi cevap vermesini bekleyemeyiz. Örüntüleri tanımada, eğitim ve tanıma işlemine başlamadan önce örüntülerin doğasına bağlı olarak çevremizde gelişen olayları gruplandırmamız gerekir. Ses işaretinden konuşmacıları tanıma, konuşulan kelimeleri tanıma, kamera karşısındaki objeleri tanıma işlemlerinin her biri ayrı bir konu olarak ele alınmalıdır[4].

Bu çalışmada metin türü tanıma konusunda başarımı artırmak amacıyla farklı algoritmaların birlikte çalışma tekniği (ensemble) ile hataların en aza indirilmesi için bir çözüm önerisi sunulmuştur.

## 2. Literatür Bilgisi

### Yapay Zeka

Yapay zekâ, 1950'li yıllarda ortaya çıkmış, genel bir ifadeyle insan davranış biçiminden esinlenerek insan gibi davranan makine sistemlerinin modellenmesidir.

Yapay Zeka;

- İnsan gibi davranma: Turing test
- İnsan gibi düşünme: Bilişsel modelleme
- Rasyonel düşünme: Mantık
- Rasyonel davranma: İnsanlar gibi düşünen sistemler yapmak

çerçevesinde disiplinler arası bir kavram olarak ele alınır[5].

### Makine Öğrenmesi

Yapay Zekânın alt bir dalı olan Makine Öğrenmesi (Machine Learning), bilgisayarların "öğrenme" işlemini sağlayacak algoritma ve tekniklerin gelişimi ile ilgili bir çalışma alanıdır.

Makine öğrenmesi; Doğal Dil işleme, Konuşma ve El Yazısı Tanıma, Nesne Tanıma, Bilgisayar Oyunları, Robot Hareketleri, Arama Motorları ve Tıbbi Teşhis gibi birçok alanda kullanılır.

Makine öğrenmesi üç önemli asamadan oluşur:

- Dokümanların Hazırlanması
- Öğrenme Metotların Uygulanması
- Öğrenmenin Performansının Değerlendirilmesi

Makine öğrenmesinde öncelikle öğrenme yapılacak veri setinin uygulanacak öğrenme metoduna uygun bir şekilde hazırlanması gerekir. Öğrenme metodunda istatistiksel yöntemler kullanılır. Geliştirilen yeni metotlar da istatistiksel temelli metotlardır. Yeni bir metot bulunduktan sonra bu metodun performansı ölçülür ve diğer metotlarla karşılaştırılması yapılır[6].

### Gözetimsiz Öğrenme

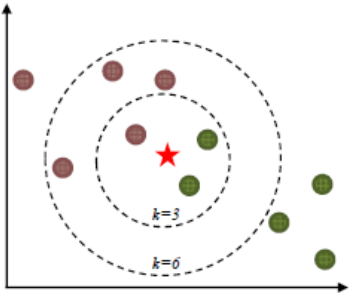
Gözetimsiz öğrenme (Unsupervised Learning) modeli gözlemlere bağlı bir makine öğrenmesi tekniğidir. Başka bir deyişle yöntem çıktı verilerinin kullanmadan sadece girdiler üzerinden öğrenme işlemini gerçekleştirmeye çalışır. Bu yöntem özellikle veri nesneleri kümesini toplamada kullanılır[7].

### Gözetimli Öğrenme

Gözetimli öğrenme (Supervised Learning) eğitim verileri üzerinden bir fonksiyon üreten bir makine öğrenmesi tekniğidir. Başka bir deyişle, bu öğrenme tekniğinde algoritma girdilerle (etiketlenmemiş veri) çıktılar (etiketlenmiş veri) arasında esleme yapan bir fonksiyon üretir[8].

### K-En Yakın Komşu Algoritması

Sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacı ile K-En Yakın Komşu algoritması (K-Nearest Neighbors Algorithm) kullanılmaktadır. Bu yöntem, örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip k sayıda gözlemin bulunduğu sınıfın seçilmesi esasına dayanmaktadır[9],[10],[11].

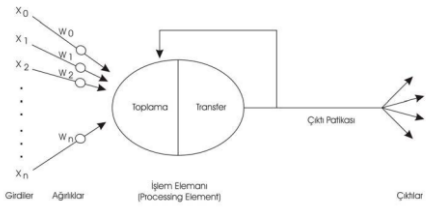



**Şekil 1.** K-En Yakın Komşu sınıflandırması

## Yapay Sinir Ağları

Yapay zeka çalışmaları kapsamında ortaya çıkan ve bir noktada yapay zeka çalışmalarına destek sağlamakta olan farklı alanlardan bir tanesi de Yapay Sinir Ağları teknolojisidir. Dolayısıyla, yapay zeka alanının bir alt dalını oluşturan YSA teknolojisi öğrenebilen sistemlerin temelini oluşturmaktadır[12].

Aşağıdaki şekilde yapay nöronun genel yapısı gözükmektedir[5].




**Şekil 2.** Yapay Nöronun Genel Yapısı.

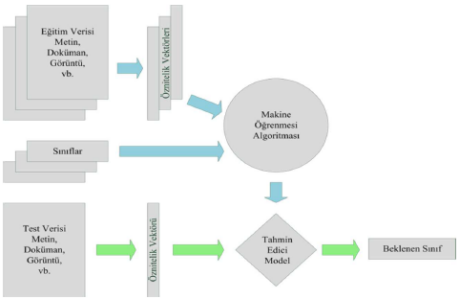
## Eğitim Kümesi

Sistemin etiketli veriler kullanılarak eğitilmesi ile öğrenmenin sağlanmasıdır. Sistem eğitilirken veri setinde bulunan her bir örneğe ait giriş ve çıkışlar verilir. Metin Sınıflandırma çalışmalarında giriş metnin içeriğini, çıkış ise kategorisini temsil eder[13].

## Test Kümesi

Test veri seti ise sistemin doğrulanması amacıyla kullanılır. Sistemin doğrulanması aşamasında öğrenme algoritması kategorisi bilinmeyen bir test verisine, eğitim verisinde bulunan çıkışlardan herhangi birini atar[13].

Öğrenme modeli süreci Şekil 3'te verildiği gibi gerçekleşmektedir[3],[14].




**Şekil 3.** Öğrenme Modeli.

## Birlikte Çalışma (Ensemble)

Birlikte çalışmanın temel amacı, daha önceden farklı sınıflandırıcılar tarafından elde edilen değerlerin bir araya getirilmesi ile bir sonuç üretilmesidir. En büyük avantajı diğer yöntemlerin verilerini bir arada kullandığı için daha iyi değerler elde edilebilmesidir[15]. Kolektif Sınıflandırma içerisinde, yerine koyarak örnekleme (bagging), hızlandırma (boosting), rotasyon ormanı (rotation forest) ve rastgele orman (random forest) gibi çeşitli algoritmalar bulunmaktadır[3].

## Hızlandırma (Boosting)

Bu yöntem sayesinde sınıflandırıcının bulmuş olduğu doğruluk değeri arttırılabilir. Boosting yönteminde veriye ait bir önceki sınıflandırıcının doğru olarak belirleyemediği veriler kullanılır[16]. Hatalı veriler sonradan kullanılacak eğitim seti içerisine tekrardan eklenerek daha doğru tahmin yapılmaya çalışılır[3],[17].

# 3. Uygulama

## Amaç

Bu çalışmada seçilen veri kümesi üzerinde iki farklı algoritmanın önce ayrı ayrı çalıştırılarak başarım oranlarının ölçülmesi, daha sonra ise ilk olarak çalıştırılan algoritmanın çıktısından elde edilen hatalı durumların ikinci algoritmaya girdi olarak verilmesi ile hata oranının daha da azaltılması hedeflemiştir.

## Veri Seti

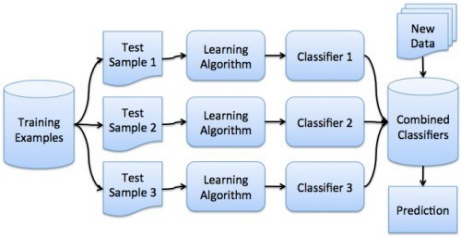
Veri seti için akışkanlar dinamiği, bilimsel endeksler ve tıbbi konulardaki 3 farklı türe ait toplamda 3891 adet makale içeren, literatürde "classic3" veri seti olarak adlandırılan veri kümesi kullanılmıştır.

| Tipi | Adet | Konusu |
|---|---|---|
| CISI | 1460 | Akışkanlar Dinamiği |
| CRAN | 1398 | Bilimsel Endeksler |
| MED | 1033 | Tıbbi |

**Tablo 1.** Veri Kümesi Tablosu

## Sistem Modeli

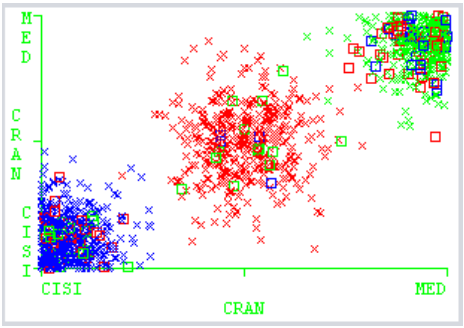
Sistem modeli aşağıdaki gibi tasarlanmıştır.



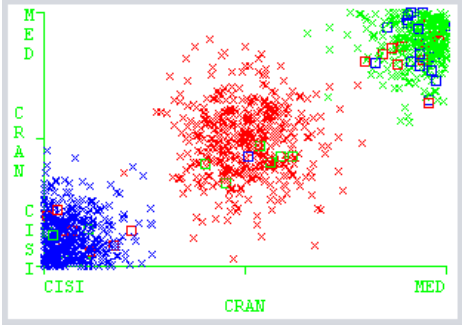

**Şekil 4.** Sistem Modeli.

## Uygulamanın Gerçekleştirilmesi

İlk olarak K-nn ve ANN algoritmaları ayrı ayrı çalıştırılmış ve görsel olarak metin türlerinin aşağıdaki gibi sınıflandırıldığı görülmüştür.




**Şekil 5.** K-En Yakın Komşu Algoritması ile Sınıflandırma.




**Şekil 6.** Yapay Sinir Ağları ile Sınıflandırma.

Şekillerden görüldüğü gibi bu veri kümesi için Yapay Sinir Ağları ile yapılan sınıflandırmanın K-En Yakın Komşu Algoritmasına göre daha başarılı olduğu tespit edilmiştir. Ancak her iki algoritmada da hatalı sonuçlar ile karşılaşılmıştır. Bu hataları en aza indirmek için iki algoritma birlikte çalıştırılarak sonuca etkisi değerlendirilmiştir.

Bunun için öncelikle K-En Yakın Komşu Algoritması uygulanmış, bunun çıktısında hata tespit edilen durumlar incelenmiştir. Bu hatalı sınıflandırmalar ayrıca Yapay Sinir Ağlarına eğitim kümesi olarak tekrar girdi olarak verilmiştir. Yapay Sinir Ağlarının çıktısı için eşik değeri olarak, K-En Yakın Komşu Algoritmasındaki yüzdece fazla olan 10 öznitelik değerleri baz alınmıştır. Bu özniteliklere yakınlığa göre çıktı kümesi elde edilerek metin türleri yeniden belirlenmiştir.

| **Algoritma Türü** | **Doğruluk (%)** | **Hata Oranı (%)** |
|---|---|---|
| Knn | 88,326 | 11,674 |
| ANN | 89,164 | 10,835 |
| Ensemble Tekniği ile | 95,776 | **4,224** |

**Tablo 2.** Doğruluk Tablosu

## 4. Sonuçlar ve Değerlendirme

Yapılan çalışma sonunda problemin sadece bir algoritma ile ele alınması yerine birlikte çalışma tekniği uygulanarak başarımın daha da arttığı, hata oranının gözle görülür biçimde azaldığı sonucuna varılmıştır. Bu yöntem uygulanırken problemin türüne ve büyüklüğüne göre en uygun algoritmalara karar verilmelidir. Birlikte çalışma tekniğinin ideal bir biçimde uygulanması durumunda hata oranını azaltıcı etkisi olduğu görülebilecektir.

## Kaynaklar

[1] Song, Q., Sheppard, M., Cartwright, M., and Mair, C.: Software Defect Association Min-ing and Defect Correction Effort Prediction. In: IEEE Transactions on Software Engineer-ing, Vol.32, No.2, pp. 69-82 (2006).

[2] Fenton, N., and Ohlsson, N.: Quantitative Analysis of Faults and Failures in a Complex Software System. In IEEE Transactions on

Software Engineering, Vol.26, No.8, pp. 797-814 (2000).

[3] Kılınç D. ve arkadaşları, “Yazılım Hata Kestiriminde Kolektif Sınıflandırma Modellerinin Etkisi”, Ulusal Yazılım Mühendisliği Sempozyomu (UYMS), İzmir, 2015.

[4] Ölmez T. Ve Dokur Z., ”Uzman Sistemlerde Örüntü Tanıma”, İTÜ Elektrik. Elektronik Fakültesi Elektronik ve Haberleşme Mühendisliği Bölümü, İstanbul, 2009.

[5] Tektaş, M., Tektaş, N., Onat, N., Gökmen, G., Koçyiğit, G. ve Akıncı, T. Ç., “Web Tabanlı Yapay Zeka Teknikleri Eğitim Simülatörlerinin Hazırlanması”, Marmara Üniversitesi, (Proje No: FEN-E-050608-138), 2010.

[6] Alpaydın E., "Introduction to Machine Learning", The MIT Press, Syf: 3-6, 2004.

[7] Hinton G. ve Sejnowski T.J., "Unsupervised Learning and Map Formation: Foundations of Neural Computation", MIT Press, ISBN 0-262-58168-X, 1999

[8] Uzun E., “İnternet Tabanlı Bilgi Erişimi Destekli Bir Otomatik Öğrenme Sistemi” Doktora Tezi, Trakya Üniversitesi, Edirne, 2007

[9] Cover, T.M., and Hart, P.E., "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, 13:21-27, (1967).

[10] Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y., "An optimal algorithm for approximate nearest neighbor searching in fixed dimensions", Journal of the ACM,45:891-923, (1998).

[11] Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C., Stamatopoulos, P., "Ling-spam - from a memory-based approach to anti-spam filtering for mailing lists", Information Retrieval, 6:49-73, (2003).

[12] Yurtoğlu, Hasan.,“Yapay Sinir Ağları Metodolojisi ile Öngörü Modellemesi: Bazı Makroekonomik Değişkenler İçin Türkiye Örneği”, Ekonomik Modeller ve Stratejik Araştırmalar Genel Müdürlüğü, Yayın No: DPT 2683, 2005.

[13] Kotsiantis, S. B., Zaharakis, I., Pintelas, P., “Supervised machine learning: A review of classification techniques, 2007.

[14] Afrin, F., Nahar, I., “Incremental learning based intelligent job search system” Doctoral dissertation, BRAC University, 2015.

[15] Augusty, S. M., Izudheen, S.: Ensemble Classifiers A Survey: Evaluation of Ensemble Classifiers and Data Level Methods to Deal with Imbalanced Data Problem in Protein-Protein Interactions. Review of Bioinformatics and Biometrics, Volume 2 Issue 1, 2013.

[16] Schapire, R. E.: Theoretical Views of Boosting and Applications. In: Proceedings of the 10th International Conference on Algorithmic Learning Theory (1999).

17. Schapire, R. E.: A Brief Introduction to Boosting. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (1999).