HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

# EĞİTSEL VERİLERDE WEKA VE ORANGE VERİ MADENCİLİĞİ YAZILIMLARINDAN ELDE EDİLEN ANALİZ SONUÇLARININ KARŞILAŞTIRILMASI

Semih Topuz

Yüksek Lisans Tezi

Ankara, 2021

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

# EĞİTSEL VERİLERDE WEKA VE ORANGE VERİ MADENCİLİĞİ YAZILIMLARINDAN ELDE EDİLEN ANALİZ SONUÇLARININ KARŞILAŞTIRILMASI

# COMPARISON OF ANALYSIS RESULTS OBTAINED FROM WEKA AND ORANGE DATA MINING SOFTWARE IN EDUCATIONAL DATA MINING

Semih Topuz

Yüksek Lisans Tezi

Ankara, 2021

## Öz

Bu çalışmada, veri madenciliğinde sıklıkla kullanılan WEKA ve Orange programları, eğitimde veri madenciliğinde kullanılabilecek sınıflama yöntemleri temel alınarak karşılaştırılmıştır. Araştırmada, Millî Eğitim Bakanlığı'nca yapılan ABİDE sınavına ilişkin sonuçlar kullanılmıştır. Analiz sürecinde, öğrencilerin Türkçe dersi puanlarına göre sınıflandırılmış verilerinin, demografik ve psiko-sosyal değişkenler kullanılarak tahmin edilmesi için k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları yöntemleri kullanılmıştır. Araştırmanın ikinci aşamasında; araştırma kapsamında belirlenen sınıflama algoritmalarıyla sınıflanmış öğrencilerden elde edilen ölçme sonuçlarının güvenirlik ve çapraz geçerlik değerleri incelenmiştir. Araştırma kapsamında elde edilen sonuçlara göre; k-en yakın komşu ve yapay sinir ağları algoritmalarında Orange, destek vektör makinesi ve naive bayes algoritmasında ise WEKA'nın daha yüksek doğru sınıflama oranına sahip olduğu görülmüştür. Rastgele orman algoritmasının ise ikili ve beşli sınıflamada sırasıyla, WEKA ve Orange paket programında daha yüksek doğru sınıflama oranına sahip olduğu tespit edilmiştir. Bunun yanında, ikili ve beşli olarak sınıflandırılmış puanlarda en yüksek doğru sınıflama oranını, yapay sinir ağları algoritması elde etmiştir.



**Anahtar sözcükler:** veri madenciliği, eğitsel veri madenciliği, Orange, WEKA, ABİDE.

**Abstract**

In this study, WEKA and Orange programs, which are frequently used in data mining, are compared based on the classification methods that can be used in data mining in education. The results of the ABIDE exam conducted by the Ministry of National Education were used in the study. During the analysis process, k-nearest neighbor, random forest, support vector machine, naive bayes and artificial neural networks methods were used to estimate the data of students classified according to their Turkish course scores by using demographic and psycho-social variables. In the second phase of the research, the reliability and cross validity values of the measurement results obtained from the students who were classified with the classification algorithms determined within the scope of the research were examined. According to the results obtained within the scope of the research; It was observed that Orange in k-nearest neighbor and artificial neural networks algorithms and WEKA in support vector machine and naive bayes algorithm have higher correct classification rates. Random forest algorithm has been found to have higher correct classification rate in binary and quinary classification in WEKA and Orange package program, respectively. In addition, the artificial neural networks algorithm obtained the highest correct classification rate in the scores classified as binary and quinary.

**Keywords**: data mining, educational data mining, Orange, WEKA, ABİDE

## Teşekkür

Tez sürecimdeki yardımlarından ötürü, başta Prof. Dr. Nuri Doğan hocam olmak üzere tüm Hacettepe Üniversitesi Eğitimde Ölçme ve Değerlendirme bölüm hocalarıma, ardından desteklerini hiçbir şekilde esirgemeyen aileme ve arkadaşlarıma teşekkür ederim. Ayrıca, talep ettiğim veri, bilgi ve belge istekler karşılandığı için Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü Veri Analizi, İzleme ve Değerlendirme Daire Başkanlığı'na teşekkür ederim.

## İçindekiler

## Tablolar Dizini

## Şekiller Dizini

## Simgeler ve Kısaltmalar Dizini

**ABİDE**: Akademik Becerilerin İzlenmesi ve Değerlendirilmesi Projesi

**DVM (SVM):** Destek Vektör Makinesi (Support Vector Machine)

**EVM**: Eğitimde Veri Madenciliği

**KYK:** k-en Yakın Komşu (k-nearest Neighbors)

**NB:** Naïve Bayes (Naive Bayes)

**PISA:** Programme for International Student Assessment (Uluslararası Öğrenci Değerlendirme Programı)

**RO:** Rastgele Orman (Random Forrest)

**VM**: Veri Madenciliği

**YSA (ANN):** Yapay Sinir Ağları (Artificial Neural Networks)

# Bölüm 1
# Giriş

Bu başlıkta problemin genel durumu, araştırmanın amacı ve önemi, araştırma soruları ve araştırmanın diğer özelliklerinden bahsedilmiştir.

## Problem Durumu

Dünyada, teknolojinin gelişimi ile birlikte veri üretimi muazzam bir miktara ulaşmıştır. Modern bilgisayar sistemleri hayal edilemeyecek büyüklükteki verileri biriktirmekte ve kullanmaktadır (Bramer, 2007). Büyük verilerin işlenmesi ve anlamlı hale getirilmesi veri miktarının büyümesiyle daha önemli hale gelmiştir. Veri madenciliği (VM), büyük veri kümelerindeki ilginç, beklenmedik veya değerli yapıların keşfedilmesini sağlamaktır (Hand, 2007). Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen, ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır (Witten, Frank, Hall, and Pal, 2005).

Veri madenciliği ülkemizde ve dünyada birçok alanda kullanılmaktadır. Bu alanlardan biri de eğitimdir. Eğitimde Veri Madenciliği ile her öğrenenin fayda sağlayabileceği bilgiler elde etmek temel amaç olmalıdır. Örneğin; PISA sınavından elde edilmiş duyuşsal ve sosyoekonomik verilerden, bilişsel verilere dönük başarılı tahminler yapmak, gelecekte öğrencilerin başarılarının daha yüksek düzeylere çıkması için hangi duyuşsal özelliklerle ilgilenilmesi gerektiğini bize gösterebilir. Böylelikle; düşük başarı düzeyinde olabilecek bir öğrencinin tespiti, sınav yapmadan sağlanabilecek ve daha yüksek bir başarıya ulaşması için nelere dikkat edilmesi gerektiği daha hızlı şekilde kararlaştırılacaktır. Eğitimde veri madenciliği yoluyla keşfedilebilecek veriler öğretmenlerin yanında, öğrenciler, veliler ve diğer paydaşlar için de yararlı olabilir. Yani eğitimde veri madenciliği uygulamaları, farklı bakış açıları ve farklı katılımcılarla uyarlanabilir.

Veri madenciliği kullanımı, büyük verilerin işlenmesi ve anlamlandırılması için zorunlu hale gelmektedir. Veri madenciliği kullanımına en uygun ortamın oluşabilmesi için veri işleme yöntemlerinin verimliliğinin sorgulanması ve geliştirilmesi gerekmektedir. Veri işleme için uygun tekniklerin seçimi ile daha büyük veriler, daha hızlı şekilde işlenebilecek ve değerlendirilebilecektir. Bahsedilen gelişmelerin sağlanması için ticari ve bilimsel çalışmalar devam etmektedir.

Günümüzde, etik sorunlara sebep olma ihtimali olsa da özellikle ticari reklamlar için kullanılan veri madenciliği değişkenler arasındaki örüntüleri tespit etmek ve bilgiyi keşfetmek için her alanda kullanılabilecek bir yöntemdir. Bilginin keşfedilmesi, veri önemsiz görülse bile, potansiyel olarak yararlı bilginin çıkarılabilmesidir (Brown, 2014).

Her alanda artan veri miktarı ile değişkenler arasındaki doğrudan gözlenemeyen ilişkileri barındırabilecek eğitsel veri tabanları da büyümektedir. Neredeyse her öğrencinin notlarının ve hatta sınavlarının çevrimiçi olduğu bu dönemde, büyük eğitimsel veri depoları oluşmaktadır. Bu veri depolarında yer alan ve akademik başarının geliştirilmesinde kullanılabilecek veriler de eğitimde veri madenciliği ile anlamlandırılarak işlenebilir. Örneğin; eğitimde veri madenciliği, eğitim sistemlerini geçerli hale getirmek ve değerlendirmek, eğitim kalitesini geliştirmek ve daha etkili öğrenme süreçlerini ortaya çıkarmak için kullanılabilir (Romero, Ventura, ve Bra, 2014). Bunların yanında eğitimde veri madenciliği, öğretmenlerin, sınıflarını yönetmelerinde yardımcı olmak amacıyla, öğrencilerinin öğrenme şekillerini tespit etmesi ve bununla birlikte öğrencilerin düşünmesini desteklemesi ve öğrencilere proaktif geribildirim sağlaması amacıyla da kullanılabilmektedir (Merceron ve Yacef, 2005). Büyük eğitim verilerini anlamlandırma işlemi için kullanılabilecek birçok yöntem ve bu yöntemleri uygulamak için farklı yazılımlar veya paket programlar bulunmaktadır. WEKA, Orange, Knime, R ve RapidMiner gibi paket programlar veri madenciliği için sıklıkla kullanılmakta ve bu programlarda karar ağaçları, regresyon, yapay sinir ağları, k-en yakın komşu, destek vektör makinesi ve naive bayes gibi birçok algoritma yer almaktadır. Paket programların farklı programlama dilleri kullanması, aynı algoritmalarda farklı sonuçlar almasına sebep olabilmektedir. Kullanılacak yöntemlerin hangisinin daha doğru (ya da daha az hatalı) hesaplamalar yaptığı; aynı analiz yöntemleri için benzer sonuçlar üretip üretmedikleri önemli bir sorundur. Ancak farklı paket programlarda kullanılan aynı analiz yöntemleri ve/veya aynı paket programda kullanılan farklı analiz yöntemleri, analize alınan örneklemin farklı özellikleri (büyüklük, veri yapısı vb.) farklı sonuçlar elde edilmesine neden olabilmektedir. Örneğin; Naik ve Samant (2016)'ın yaptığı çalışmada karar ağacı algoritması Orange'da %66 doğrulukla tahmin yaparken Knime'da %95 doğruluk oranına çıkmaktadır. Aynı çalışmada WEKA, naive bayes algoritmasında %54

doğrulukla tahmin yaparken, k-en yakın komşu algoritmasında bu oran %99'a çıkmaktadır. Özetle; sağlık alanında yapılan bu çalışma, bir açıdan da veri seti sabit kaldığında algoritma ve paket programların tahminlemelerinin nasıl değiştiğini göstermektedir. Bu nedenlerle veri madenciliği yöntemlerinin hangilerinin eğitimde uygulanmasının daha isabetli sonuçlar ürettiğini saptamak için, veri madenciliği paket programlarının ve yöntemlerinin karşılaştırılması yararlı sonuçlar verebilir. Verimli yöntem ve programların tespit edilmesi, eğitimde veri madenciliği ile hızlı ve ekonomik değerlendirme sürecine katkı sağlayabilecektir.

## Araştırmanın Amacı ve Önemi

Bu tezin temel amacı, veri madenciliği paket programları WEKA ve Orange'ın ve bu paket programlarda yer alan algoritmaların, eğitimde veri madenciliği için kullanılması amacıyla karşılaştırılmasıdır. Bu karşılaştırmada örneklem olarak Türkiye'de uygulanan ABİDE sınavının verilerinin kullanılması ise Türkiye'de eğitimde veri madenciliği uygulamalarının gelişmesi ve yaygınlaşması; Türkiye'de uygulanan büyük ölçekli sınavların değerlendirilmesinde alternatif yollar üretmesi bakımından önemlidir. Diğer yandan, araştırmada bağımlı değişkenin hem iki kategorili hem de çok kategorili olarak ele alınmasının, paket program ve ele alınan kestirim yöntemlerinin farklı kategori düzeyleri için performanslarının karşılaştırılması bakımından da alanyazına katkı sağlaması beklenmektedir. Türkiye'de, eğitimde veri madenciliği için ABİDE sınavının örneklem olarak kullanıldığı bir çalışma, bu çalışmaya dek yayımlanmamıştır. Alanyazın incelendiğinde birçok veri madenciliği paket programının karşılaştırması bulunmaktadır. Naik ve Samant'ın (2016)'da yaptıkları çalışmada sağlık sektöründen elde edilmiş verileri karşılaştırdığı; Hussain ve diğerlerinin (2018) çalışmasında ise sınıflama işlemleri yalnızca WEKA üzerinde yapıldığı görülmektedir. Bu iki paket programın sınıflama yöntemleri açısından eğitimde veri madenciliği verilerinde karşılaştırması da henüz yapılmamıştır. Belirtilen nedenlerle çalışmanın eğitim bilimleri ve eğitsel veri madenciliği alanyazınına katkı getireceği düşünülmektedir.

## Araştırma Problemi

ABİDE sınavında ölçülen demografik ve psiko-sosyal değişkenlerden yararlanılarak, Türkçe alt testi başarısını tahmin etmede kullanılan Veri Madenciliği yöntemlerinden k-en yakın komşu (k-nearest neighbors - KNN), rastgele orman (random forest), destek vektör makinesi (support vector machine - SVM), naive bayes ve yapay sinir ağları (artificial neural network - ANN) ile elde edilen analiz sonuçları, WEKA ve Orange programlarında farklılaşmakta mıdır?

### Alt problemler.

A) ABİDE sınavında Türkçe alt testi puanları beşli sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak;

1) K-en yakın komşu
2) Rastgele orman
3) Destek vektör makinesi
4) Naive bayes
5) Yapay sinir ağları

yöntemleri ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve eğri (ROC) altında kalan alan değerleri, WEKA ve Orange programlarında farklılaşmakta mıdır?

B) ABİDE sınavında Türkçe alt testi puanları ikili sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak;

1) K-en yakın komşu
2) Rastgele orman
3) Destek vektör makinesi
4) Naive bayes
5) Yapay sinir ağları

yöntemleri ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve eğri (ROC) altında kalan alan değerleri, WEKA ve Orange programlarında farklılaşmakta mıdır?

C) ABİDE sınavında Türkçe alt testi puanları ikili ve beşli sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak, WEKA programında;

1) K-en yakın komşu
2) Rastgele orman
3) Destek vektör makinesi
4) Naive bayes
5) Yapay sinir ağları

yöntemleri ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve eğri (ROC) altında kalan alan değerleri farklılaşmakta mıdır?

D) ABİDE sınavında Türkçe alt testi puanları ikili ve beşli sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak, Orange programında;

1) K-en yakın komşu
2) Rastgele orman
3) Destek vektör makinesi
4) Naive bayes
5) Yapay sinir ağları

yöntemleri ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve eğri (ROC) altında kalan alan değerleri farklılaşmakta mıdır?

**Sınırlılıklar**

1) Çalışma kapsamında kullanılan sınıflama ve tahmin modelleri, WEKA ve Orange programında var olan k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları algoritmaları ile sınırlıdır.

2) Başarının tahmin edilmesinde kullanılacak değişkenler, ABİDE(2018) öğrenci anketinde yer alan alt ölçekler ve sosyodemografik özellikler ile sınırlıdır.

**Tanımlar**

ABİDE (Akademik Becerilerin İzlenmesi ve Değerlendirilmesi): Millî Eğitim Bakanlığı tarafından geliştirilen ABİDE Projesi kullanılarak Türkiye'deki öğrencilerin üst düzey bilişsel özelliklere sahip olma durumlarına yönelik ölçümler yapmak amaçlanmaktadır. ABİDE projesinin sonucunda ise ulusal bir izleme ve değerlendirme sistemi kurulması planlanmaktadır.

# Bölüm 2
# Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Bu kısımda araştırmanın kuramsal temelinde neler olduğundan ve ilgili araştırmalardan bahsedilecektir.

## Veri Madenciliği

Teknolojinin getirdiği bilgi ulaşılabilirliği ve bilgi paylaşma özgürlüğü, akla gelebilecek her alanda verinin toplanmasını ve depolanmasını sağlamaktadır. Kredi kartı kullanımından uydu bağlantısı verilerine, sosyal medya paylaşımlarından gen haritalarına, dünyanın her noktasında, her an, veri oluşmaktadır (Bramer, 2007). Üretilen verinin, farklı şirketlerin sahip olduğu bilgisayarlarda toplanması ve işlenmesi ile şirketler kullanıcıların alışkanlıklarını ve davranışlarını tahmin edebilmektedir. Veri madenciliği (VM), kullanıcılara ait kalıpları bulma ve yorumlama konusunda şirketlere bilgi sağlamaktadır. Şirketler bu yolla kullanıcılara özel reklam ve öneriler sunmakta ve bir yandan hayatlarımızı kolaylaştırırken, bir yandan da kişisel verilerin gizliliği ile ilgili soru işaretlerine sebep olmaktadır (North, 2012). Makine öğrenmesi değişkenler arasındaki ilişkilerin keşfi için kullanılabilecek birçok yol sunmaktadır. Bu yollardan her biri veriden elde edilecek çıktıdan anlam çıkarmak için kullanılabilecek tekniğin türünü belirlemektedir. Kullanılacak tekniğin seçimi ve uygulanması veriden veriye değişirken, elde edilen çıktıdan yapılacak yorumlar, teknikten tekniğe değişiklik göstermektedir (Aksu, 2018).

Veri madenciliğinin literatürde birçok farklı tanımı bulunmaktadır. Bunlardan bazılarında veri madenciliği şöyle tanımlanmıştır:

- Geçerli tahminler yapmak için, kullanılan verilerdeki ilişkiyi ve örüntüleri açığa çıkarmak amacıyla çeşitli veri analizi tekniklerinin kullanılmasıdır (Two Crows Corp., 1999),
- Büyük veri tabanlarında bulunan veri yığınlarından yararlanarak, gizil ilişkilerin ve genel örüntülerin ortaya çıkarılmasıdır (Holsheimer ve Siebes, 1994),

- Büyük veri yığınlarında bulunan verilerden anlamlı örüntülerin otomatik veya yarı otomatik olarak keşfedilme sürecidir (Witten I. H., Frank, Hall, ve Pal, 2005),
- Veri tabanında yer alan verilerden bilginin otomatik olarak çıkarılması ve analiz edilmesinde bir veya daha fazla bilgisayar öğrenme tekniklerinin uygulanması sürecidir (Roiger, 2017).

Veri madenciliği, bir şirketin çalışma sistemini geliştirmeye de yardımcı olabilir. İşletmenin farklı bileşenlerinin birbiriyle daha uyumlu olabilmesi için verebileceği ipuçları sağlanabilmektedir. Nerede maliyetin düşürülebileceği, en iyi gelirin nasıl sağlanacağı gibi çıktılar kurumun büyümesi ve gelişmesi için yararlı olabilmektedir. Bilgi keşfi, önemsiz görülen bir veriden potansiyel olarak yararlı bilginin çıkarımıdır (Brown, 2014).

Veri madenciliği birçok alanda olduğu gibi eğitimde de faydalı bilgiler üretebilmektedir. Eğitimde, mevcut veri tabanlarının veri madenciliğine uyarlanması süreci, öğrenmenin hızlanması ve verimlileşmesi için önemli olacaktır. Veri madenciliği yoluyla kestirilen örüntüler, daha sonra modellerle bireysel davranış ve bilişsel özellikleri kestirmede kullanılabilmektedir (Aydın, 2007).

Eğitim sistemlerinde öğrenmenin gelişmesi için veri madenciliği uygulamalarının yapılması, biçimlendirici ölçme ve değerlendirmeye (*formative assessment*) örnek gösterilebilir. Bütünsel bir yaklaşımla, eldeki tüm öğrenci verilerini kullanarak ölçme ve değerlendirmenin yapılması için veri madenciliği teknikleri büyük bir avantaj olacaktır. Ancak, öncelikle verilerin paylaşılması ve işlenmesi ile ilgili problemler ortadan kaldırılmalı, uygun tekniklerin seçimi için çalışmalar yapılmalı, uygun teknikle işlenmiş verilerin geçerlik ve güvenirliğinin tespiti yapılmalı ve yöntem bütünüyle sisteme sunulmalıdır. Bireylerin doğru ve hızlı şekilde değerlendirilmesi birçok açıdan yararlı olacaktır. Örneğin; ekonomiklik ve güven sağlayacaktır.

Veri madenciliği, farklı görevleri yerine getirmek için farklı algoritmalar kullanır. Algoritmaların tümünde veriler incelenir ve incelenen veri ile oluşturulan model, verinin özelliklerine göre belirlenir. Verinin ve problemin özelliklerine göre farklı algoritmalar, sınıflama, kümeleme, örüntü tanımlama gibi görevleri yerine getirir (Aydın, 2007). Veri madenciliği modelleri 3 grupta toplanabilir:

- Kümeleme *(clustering),*
- Birliktelik kuralları ve ardışık zamanlı örüntüler *(association rules and sequential patterns),*
- Sınıflama ve regresyon *(classification)*

**Kümeleme.** Benzer özelliklere sahip olan nesnelerin gruplara ayrılma sürecidir. Bölümleme olarak da adlandırılır. Veri seti içerisindeki kategoriler önceden bilinmediğinde özellikle yararlıdır. Çok boyutlu ortamlarda kendine has özellikler sergileyen veri gruplarının ortaya konması ya da örüntü parçalama aşamalarında kullanılır (Amershi ve Conati, 2006).

Kümeleme analizinin amacı, veri setinde doğal olarak oluşan altsınıfları bulmaktır. Sınılama analizinden farkı denetimsiz olmasıdır. Ayrılmak istenen küme sayısının bilinmemesi sebebiyle iki aşamalı şekilde gerçekleştirilir. Mevcut küme sayısı üzerinde dıştan bir döngü ve belirli sayıdaki en iyi kümeleme için içsel bir döngü ile yapılır (Fayyad ve Stolorz, 1997).

**Birliktelik kuralları ve ardışık zamanlı örüntüler.** Birliktelik kuralları, veri kümeleri arasında birliktelik ilişkilerini bulurlar. Veri seti büyüdükçe birliktelik kurallarının ortaya çıkarılması gerekir. Daha etkili ve verimli bir karar için birliktelik kurallarının keşfi, zorunlu hale gelir. Ardışık zamanlı örüntüler, birbirleriyle ilişkisi olan ve ardışık zamanlı gerçekleşen ilişkileri tanılamak için kullanılmaktadır (Hark, 2013).

Birliktelik kuralları yöntemlerinde dikkat edilmesi gereken iki hususdan bahsedilebilir. Birinci husus, maliyettir. Çok büyük veri setleri, örüntü çıkarılmasını, bilgisayar kaynakları ve hesaplamalar açısından zaman alıcı ve maliyetli hale getirir. İkinci husus ise, bazı örüntülerin tesadüfi ve sahte olmasıdır. Bu durumu çözmek için örüntüler ayrıca incelenmelidir. Birliktelik kurallarının gücünü belirlemek için destek (support), güven (confidence), ve yükselme oranı (lift ratio) olarak isimlendirilen ölçütler kullanılmaktadır (Irmak, 2009).

**Sınıflama.** Sınıflama, yeni bir nesneyi, daha önceden tanımlanmış sınıf setlerinden birine atama işlemine denir. Sınıflama yapabilmek için, sınıflandırılmamış veriye uygun bir model seçilmelidir. Böylelikle daha hızlı ve güvenli bir sınıflama yapılabilecektir (Taşdemir, 2012). Sınıflama işleminde test verileri ile denetimli bir öğrenme sağlanır. Yani; eldeki veri seti farklı yöntemlerle

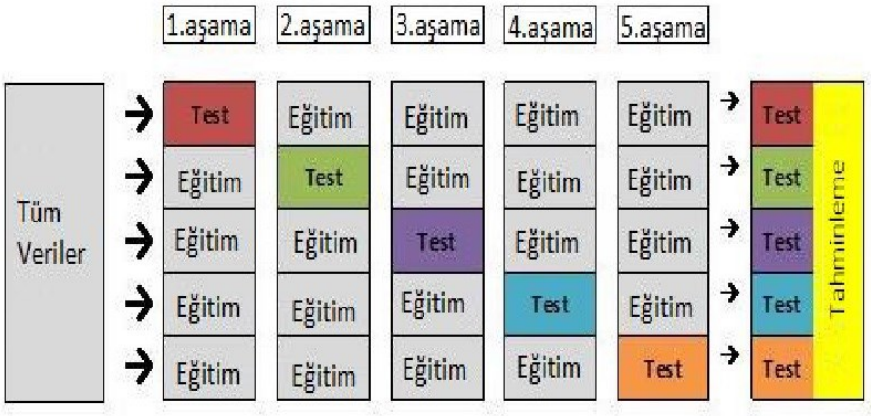
ikiye bölünür ve ilk veri eğitim verisi olarak kabul edilir. Bu ilk parça veri seti, modelin kurulması için kullanılır. Bu denetimli bir öğrenme sağlar. Ardından ikinci veri ile ilk veriden elde edilen model test edilir. Bu veri ise test verisidir.

Sınıflama teknikleri veri madenciliğinde yoğun biçimde kullanılır. Temel birkaç algoritma şöyledir (Şeker, 2013):

- Destekçi Vektör Makinesi (Support Vector Machine)
- Doğrusal Olmayan Destekçi Vektör Makinesi (Non-Linear SVM)
- Veri Akış Madenciliği (Data Stream Mining)
- Naive Bayes Sınıflandırıcısı (Naive Bayes Classifier)
- Naive Bayes ile Metin Sınıflandırma (Naive Bayes Text Classification)
- Karar Ağacı Öğrenmesi (Decision Tree Learning)
- K-En Yakın Komşu (KNN, K-nearest neighborhood)

Verilerle algoritmanın eğitilmesinin ardından veriler test edilmektedir. Bunun için eldeki veri setinin kullanılması yaygın olarak görülmektedir. Eldeki veri seti, farklı yöntemlerle ikiye bölünerek bir bölümü eğitim verisi haline getirilmektedir. Pratik olarak verinin üçte ikisinin (2/3) eğitim için kullanılması ve üçte birinin (1/3) test için kullanılması da yaygın olarak uygulanmaktadır. Bu yöntem veri setinin yeterli sayıda olmasını gerektirmektedir. Bu yüzden veri setinin yetersiz olması durumlarında kullanılabilecek yöntemler de vardır.

Veri setinde az sayıda ve sınırlı veri var ise çapraz geçerlik yöntemi uygulanabilir. Çapraz geçerlikte veri seti *k* eşit parçaya bölünmekte ve bir parça eğitim verisi olarak kullanılarak model oluşturulmaktadır. Ardından *k-1* parça ile bu model test edilmektedir. Daha sonra her *k* parça için ayrı ayrı aynı işlem tekrarlanmakta ve en iyi sonucu veren parçanın modeli doğru model olarak kullanılmaktadır. Böylelikle modelin eğitimi için kullanılan veriler test aşamasında kullanılmamaktadır. 5 katlı çapraz geçerleme sürecinin nasıl işlediği Şekil 1'de görsel olarak gösterilmektedir (Aksu ve Doğan, 2018).




*Şekil 1.* 5 katlı çapraz geçerleme süreci örneği (Aksu ve Doğan, 2018).

Tahminin istatistiksel değerlendirilmesi için kullanılan bir diğer yöntem *bootstrap*'tır. Bu yöntemde ise rastgele oluşturulmuş örneklemde yer alan birimler her bir adımda yerleri değiştirilerek yeniden örnekleme dahil edilmektedir.

Yöntemlerin doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği, karekök hata ve göreceli karekök hata değerleri hesaplanabilmektedir. Bu değerler modelin test veri setini ne doğrulukla kestirdiğini bize bildirmektedir. Bu değerlerin incelenmesi sonucunda hangi modelin kullanılmasının daha iyi sonuçlar verdiği ortaya çıkarılabilmektedir.

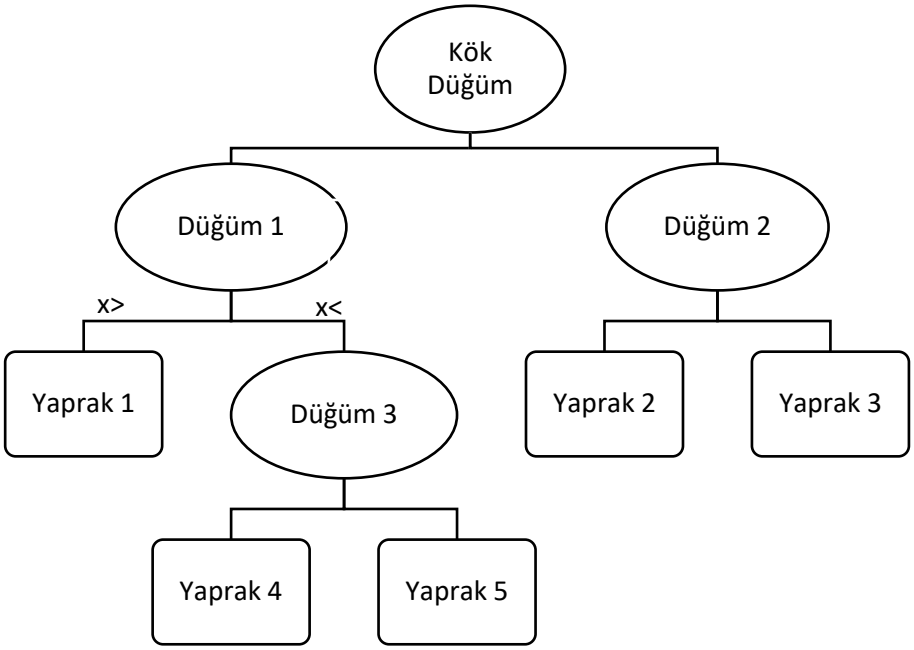
Sınıflama analizinin verdiği sonuçlar, risk matrisi, birikimli kazanç eğrisi ve kaldıraç grafiği, alıcı çalışma karakteristik grafiği (ROC) gibi yöntemlerle değerlendirilmektedir. Bu değerler ise modelin ne kadar etkili olacağını, yani modelin kullanılmasının risk ve yararlarını göstermektedir (Aydın, 2007).

## Sınıflama Algoritmaları ve Sınıflandırma Kalitesine İlişkin Ölçütler

**Karar ağaçları.** Karar ağaçları sınıflandırma algoritmaları içerisinde en sık kullanılan algoritmalardan biridir. Sınıflandırma algoritmaları içerisinde nispeten en kolay oluşturulabilen ve yorumlanabilen algoritmalardan biri olması bunun temel sebebidir (Friedl ve Brodley, 1997). Ayrıca, karar ağaçları doğrusal olmayan özellik ve sınıf ilişkilerini ele alabilir, kayıp verilere izin verir ve hem kategorik hem de

sayısal verilerle çalışmaya izin verir (Fayyad ve Irani, 1992). Karar ağaçları, sınıflandırmada iki aşamalı olarak kullanılır. İlk adımda karar ağacı kurulur. İkinci aşamada ise sınıflandırma yapılır.

Veri setinin belirlenen kısmındaki (eğitim verisi) verilerin, farklı yöntemlerle analiz edilmesi sonucunda ortaya çıkan değerlere bağlı olarak, karar ağaçlarının yapısı oluşturulur. Karar ağaçlarının yapısı *düğüm, dal* ve *yapraklardan* oluşmaktadır. *Kök düğüm* ilk dallandırmanın yapıldığı yerdir. Dallandırmanın sona ermesi durumunda ortaya yapraklar çıkar. Yapraklar saf düğüm olarak da adlandırılır (Altunkaynak, 2017).




*Şekil 2.* Örnek karar ağacı yapısı

Karar ağaçlarında bulunan düğümler ve yapraklar içerisine genellikle bağımlı değişkenin dağılım bilgisi yazılmaktadır. Dallar üzerine ise genellikle dallandırma yapılan bağımsız değişkenin sınıfı ya da sınıflandırma kuralı yazılmaktadır.

Karar ağaçlarının oluşturulmasının ardından sınıflama kuralları belirlenmiş olur. Bu kurallar algoritmanın eğitimini sağlamaktadır ve bunlara *karar kuralları (decision rules)* denir. Bu kurallarla veriler sınıflanabilir.

Bu bölümde, sıklıkla kullanılan birkaç karar ağacı algoritmasından bahsedilecektir.

***ID3.*** ID3 algoritması, Quinlan (1986) tarafından geliştirilmiştir. Adımsal ikiye bölme (Iterative Dichotomiser 3; ID3) algoritması olarak adlandırılan ID3, kategorik değişkenler için geliştirilmiştir. Daha sonrasında geliştirilen birçok algoritmaya da temel olan bu algoritma C4.5 algoritmasının da öncülü olarak görülmektedir (Akpınar, 2014).

Algoritmada, bölen özniteliklerin belirlenmesi için entropi ve kazanım (gain) değeri kullanılır. ID3 algoritmasının ilk adımı genel entropi değerinin hesaplanmasıdır. Entropi; bir değişkendeki belirsizliğin ölçülmesi için kullanılmaktadır. Eğer değişkenin tüm değerleri eşitse entropi (belirsizlik) yoktur ve entropi değeri sıfır olacaktır. Entropinin hesaplanması için özniteliğin ortaya çıkma olasılığı bilinmelidir.

$$H(Y) = \sum_{j=1}^{k} \left( p_j \log_b \left( \frac{1}{p_j} \right) \right)$$

Formülde yer alan $k$ değeri herhangi *Y* değişkeninin düzey sayısıdır. $p_j$ ise $j$. düzeyin ortaya çıkma olasılığıdır. $b$ değeri logaritma tabanıdır. $b = 2$ olarak alındığında elde edilen entropi değeri *Shannon entropisi* olarak adlandırılır. Shannon entropisinin değeri ise *bit*tir. Shannon entropisi genellikle değişken iki düzeye sahip ($k = 2$) ise tercih edilir. İkiden fazla düzeye sahip değişkenlerde ise logaritma değeri 10 alınarak ($b = 10$) elde edilen *Hartley entropisi* kullanılır. Entropi değerinin en küçük değeri sıfır, en büyük değeri ise değişkenin düzey sayısının, formülde kullanılan logaritma tabanıyla yazımıdır ($\log_b k$ ) (Quinlan, 1986; Altunkaynak, 2017).

Algoritma için entropi değerleri hesaplandıktan sonra, her bir bağımsız değişkenin kazanım (gain) değeri hesaplanmalıdır. Kazanım değerinin hesaplanması için, bağımsız değişkene ait entropi değerinin, bağımlı değişkenin entropi değerinden çıkarılması gerekir.

$$Gain(X_i) = H(Y) - H(Y, X_i)$$

Her bir bağımsız değişkenin kazanım değeri hesaplandıktan sonra, en yüksek kazanım değerine sahip olan bağımsız değişkende dallandırma yapılır. Kazanım değerinin büyüklüğü, bağımlı değişken üzerindeki etkiyle doğru orantılıdır. Gelinen noktada, saf düğüm olmayan her bir düğüm için veriler indirgenerek aynı

işlemler yapılmalıdır. Bu işlem, tüm düğümler saf düğüm haline gelene kadar yapılmalıdır (Altunkaynak, 2017).

***C4.5.*** C4.5, ID3 algoritmasında, bağımsız değişkenin kategori sayısının artması ile ortaya çıkan, kazanım değerinin artması probleminin önüne geçilmesi amacıyla geliştirilmiştir. ID3 algoritması, sadece kategorik değişkenler için kullanılmaktadır. C4.5 algoritması, bu sorunları aşması amacıyla Quinlan tarafından 1993'te önerilmiştir. Hem kategorik hem de sürekli değişkenle çalışabilmesi ve veri setindeki kayıp verileri hesaplamaya katmaması en önemli geliştirmeleri olarak görülmektedir (Akpınar, 2014).

C4.5 algoritması, ID3 algoritmasında yer alan kazanım değeri yerine *kazanım oranı*'nı kullanmaktadır. Kazanım oranı, kazanım değerinin entropiye bölünmesi ile elde edilmektedir. Bu oranlama, değerin kategori sayısına göre düzeltilmesini sağlamaktadır. Kazanım oranının büyüklüğü, kazanım değerinde olduğu gibi bağımlı değişkende en fazla etki anlamına gelmekte ve dallandırma için kullanılmaktadır.

$$GainRatio(X_i) = \frac{Gain(X_i)}{H(X_i)}$$

Kazanım oranının hesaplanması sırasında, ID3 algoritmasında yer alan işlemler aynı sırayla yapılır. Yalnızca dallandırma işlemi için kazanım değeri değil, kazanım oranı kullanılır. C4.5 algoritmasının lisansı, Quinlan tarafından alındığı için telif ücretine tabidir. Bu yüzden, algoritmanın ücretsiz versiyonu olan J48 algoritması kullanılmaktadır. Ayrıca birçok yazılımda C5.0 ismiyle de yer almaktadır (Altunkaynak, 2017).

***CART.*** CART (*Classification and Regression Trees*) algoritması, karar ağaçları içerisinde önemli bir yere sahiptir. CART, sürekli ve kategorik verilerle çalışabilmektedir. Breiman ve arkadaşları tarafından 1984'te geliştirilmiş olan algoritma, her bir düğümde ikili dallandırma yapmaktadır. Bu ikili dallandırma sebebiyle ikili ağaç (*binary tree*) yapısına sahiptir. Yani her bir dallandırma işleminde, bağımsız değişken üzerinden, sorulara evet veya hayır şeklinde cevaplar verilir. Bu ikili dallandırma, bir sonraki dallandırma için daha homojen hale getirilmiş düğümler elde etmeyi sağlamaktadır. Bağımlı değişkenin kategorik olması durumunda genellikle *twoing kriteri* ve *gini ölçüsü* kullanılmaktadır. Bağımlı

değişken sürekli ise *hata kareler toplamı* kullanılmaktadır (Crawford, 1989; Wu ve Kumar, 2009).

Twoing kriterinde dallandırma yapılırken, her bir bağımsız değişkene bağlı, bütün ikili ayrıştırmalar önemlidir. Tüm ayrıştırmalar belirlendikten sonra, twoing kriterleri hesaplanır.

$$\phi(D_i) = 2P_L P_R \sum_{j=1}^{k} \left[ \left| P\left(Y_j \,/\, L\right) - P\left(Y_j \,/\, R\right) \right| \right]$$

Burada; $\phi(D_i)$, $i.$ bölünme için twoing kriterini; $P_L$ ve $P_R$, bu bölünme için gözlemin ağacın solunda ve sağında kalma olasılığını; $P\left(Y_j \,/\, L\right)$ ve $P\left(Y_j \,/\, R\right)$ ise ağacın solunda ve sağında bağımlı değişkenin $j.$ düzeyinin olasılığıdır. Twoing kriterinin büyüklüğü, bağımsız değişkenin, bağımlı değişken üzerindeki etkisinin büyüklüğü ile açıklanır. En büyük kriter değerine sahip bağımlı değişken dallandırma için seçilir. Dallandırmanın ardından, geriye kalan mümkün bölünmeler için twiong kriteri hesaplanır. Bu işlemler mümkün bölünme kalmayana dek yapılır (Altunkaynak, 2017; Wu ve Kumar, 2009).

Gini ölçüsü ile dallandırma ise bağımsız değişkenlerin *gini ayırma değerine* dayalıdır.

$$Gini_{split}(D_i) = P(L_i)Gini(L_i) + P(R_i)Gini(R_i)$$

Burada, $P(L_i)$ ve $P(R_i)$ değerleri, $i.$ bölünme durumunda bir kaydın ağacın solunda ve sağında olma olasılıklarını göstermektedir. $Gini(L_i)$ ve $Gini(R_i)$ değerlerinin hesaplanması için ise şu formüllere başvurulur.

$$Gini(L_i) = 1 - \sum_{j=1}^{k} \left[ P(Y_j \,/\, L_i) \right]^2$$

$$Gini(R_i) = 1 - \sum_{j=1}^{k} \left[ P(Y_j \,/\, R_i) \right]^2$$

Burada; $k$, bağımlı değişkenin düzey sayısını; $P(Y_j \,/\, L_i)$ ve $P(Y_j \,/\, R_i)$ ise $i.$ bağımsız değişkenin, sol ve sağ dallandırmasında $j.$ düzeyin görülme olasılığıdır (Roe, ve diğerleri, 2005).

Gini ayırma değeri en küçük olan bağımsız değişken, bağımlı değişken üzerinde en fazla belirleyiciliğe sahiptir. Dallandırma en küçük değere sahip değişkenden başlar. Bu dallandırmanın ardından, geriye kalan mümkün bölünmeler için gini ayırma değeri hesaplanır. Mümkün olan bir bölünme kalmayıncaya dek bu işlem devam ettirilir (Altunkaynak, 2017).

***CHAID.*** CHAID (CHi-squared Automatic Interaction Detection) algoritması, Automatic Interaction Detection (AID: Otomatik Etkileşim Algılama) isimli, daha eski bir tekniğin üzerine, Kass (1980) tarafından geliştirilmiştir. Otomatik etkileşim algılama isimli teknik, aralıklı ölçeklendirilmiş bir bağımlı değişken üzerinde çalışır ve her bir ikiye bölmede gruplar arası kareler toplamını (temel olarak F-istatistiği) maksimize eder. Buna karşılık, CHAID, nominal ölçeklendirilmiş bağımlı değişken üzerinde çalışır, her bölmede ki-kare istatistiğinin anlamlılığını maksimize eder. AID, verilerin doğasındaki örnekleme değişkenliğini hiçbir zaman dikkate almaz. Bu yüzden ciddi sınırlamalara sahip olduğu eleştirisi kabul edilebilir olmaktadır. CHAID, bu sorunu, bölümleme problemini bir anlamlılık testi çerçevesine dahil ederek ele almaktadır. Böylelikle CHAID, AID'de yer alan fazla kategoriye sahip değişkenlerle çalışma problemini, geçersiz kılabilmektedir (Kass, 1980).

CHAID, her hiyerarşik seviyede ikiden fazla dala ayrılabilmektedir. Kayıp verileri ayrı bir kategoride analiz eder ve CHAID'de nesnelere ağırlık ve frekans değerleri atanabilmektedir. Ayrıca verilerde normal dağılım aranmaması önemli bir avantajıdır. CHAID, bağımlı değişken kategorik ise $\chi^2$ testini, sürekli ise F testini kullanmaktadır. Bağımsız değişkenler ise analizde otomatik olarak kategorik değişkene dönüştürülmektedir. Bu dönüştürme sırasında çıkabilecek sorunlara karşılık, Biggs, De Ville ve Suen (1991) tarafından *Exhaustive CHAID* geliştirilmiştir. Bağımsız değişkenlerin kategorik hale getirilmesi sırasında, her bir değişken için mümkün bölen noktaları hesaplanmalıdır. Mümkün bölen noktasının bulunması sırasında çıkabilecek sorunlar, Exhaustive CHAID tarafından, sadece iki süper kategori kalıncaya dek bağımsız değişkenin kategorilerini birleştirmeyi sürdürerek çözülmeye çalışılmıştır. Bu yöntem bağımsız değişken birleşmeleri içerisindeki, bağımlı değişkenle en kuvvetli birlikteliğe sahip kategoriler dizisini bulmaya çalışır (Akpınar, 2014).

***Rastgele orman.*** Sınıflandırma işlemlerindeki gelişim, *ensemble learning (çoklu öğrenme)* ile sağlanmıştır. Ensamble learning, sınıflandırma atamalarında farklı algoritmaların bir arada kullanılması olarak tanımlanabilir. Rastgele orman (random forrest – RF) da ensamble learning kullanan algoritmalara bir örnek olarak verilebilir. Breiman (2001) tarafından geliştirilen rastgele orman, ensamble learning metodunun *oylama* yaklaşımını kullanır. Oylama yaklaşımı, sınıflandırıcıları birleştirmek için olasılık değerlerini veya sayısal tahminleri kullanır. Rastgele orman algoritmasında ise birçok karar ağacının bir araya gelmesi sonucunda oluşan yapıda, bireysel ağaçların oylaması ile kazanan sınıf belirlenir (Korkem, 2013).

Rastgele orman algoritması için veri seti içerisinden *bootstrap* yöntemi ile *n* tane örnekleme yapılmalıdır. Burada *n* kullanılacak karar ağacı algoritması sayısıdır. Ardından her örneklemin üçte ikisinin eğitim verisi olarak bölünmesi gerekir. Daha sonra her örnekleme için ayrı şekilde karar ağacı oluşturulmalıdır. Bu ağaçlar oluşturulurken dikkat edilmesi gereken bir nokta da her düğümde tahmin değişkenleri içerisinde en iyi değişken yerine, rastgele *m* tane tahmin değişkeni seçilmesi gerekliliğidir. Bu *m* tane değişken içerisinden, en fazla bilgi kazancı (*information gain*) sağlayacak olan değişken, dallandırma için seçilmelidir. Burada belirlenecek olan *m* değerinin regresyon ağaçlarında $p/3$, sınıflama ağaçlarında ise $\sqrt{p}$ olarak önerilmiştir. Buradaki $p$ değeri ise bağımsız değişken sayısıdır. Tüm bu işlemlerin ardından sınıflama ağaçları için en çok oyu alan tahmin, regresyon ağaçları içinse yapılan oylamanın ortalaması alınarak nihai tahmin yapılır (Breiman, 2001).

Rastgele orman algoritması hem kategorik ve sürekli hem de ikisinin birlikte bulunduğu veri setlerinde rahatlıkla kullanılabilir. Kayıp verilerin olduğu bir veri setinde başarılı bir sınıflama yapabilmektedir. Değişkenler arasındaki ilişkinin ve mesafenin hesaplanabilmesi de algoritmanın bir avantajıdır. Ancak oluşan sonuç için bir güven aralığının belirlenemeyişi algoritmanın dezavantajı olarak ele alınabilir. Birden fazla karar ağacı oluşturulması ve bu ağaçların bellekte yaratacağı yük sebebiyle, daha üstün belleğe sahip bilgisayarlarda uygulanması mümkün olabilmektedir (Akman, Genç, ve Ankaralı, 2011).

**k-en yakın komşu.** k-en yakın komşu (k-nearest neighbor - KNN) algoritması, veri madenciliği ve makine öğrenmesindeki en kolay yaklaşımlardan biridir. KNN'de bağımsız değişkenlerin sayısal olması gerekmektedir. Bu algoritmada, sınıflandırılmak istenen veri kendisine en yakın $k$ komşunun sınıfına uygun olarak atanır. Bu $k$ değerinin belirlenmesi için bir ön çalışma yapılması ve çapraz geçerliliklerin incelenmesi önerilir. Bunun yanında veri setinin büyük olduğu durumlarda bu sayının olabildiğince yüksek olması tavsiye edilmektedir (Wu ve Kumar, 2009; Dudani, 1976).

KNN algoritması için öncelikle $k$ değerinin belirlenmesi gerekir. Uygun koşullara göre $k$ değeri belirlendikten sonra, sınıflandırılmak istenen verinin gözlemlenen değerlere uzaklıklarının hesaplanması gerekir. Ardından en yakın $k$ komşunun sınıflandırmasında en fazla tekrar eden sınıfa atama yapılır. Matematiksel olarak, $k$ sayısının tek sayı olarak belirlenmesi yararlı olmaktadır. Çünkü; olası bir komşu sayısı eşitliği, sınıflandırma atamasını yapmaya engel olacaktır. KNN algoritmasında, gözlemlenen değerlere uzaklığın hesaplanmasında genellikle *öklid ölçüsü* kullanılmaktadır. Öklid ölçüsü, hipotenüsle hesaplama yapar. Eğer özel bir uzaklık ölçüsünden bahsedilmediyse öklid uzaklığına bakılacaktır. Ancak bazı durumlarda *manhattan uzaklığı* ve *mahalanobis uzaklığı* gibi hesaplama yöntemleri de kullanılabilir (Şeker, 2013; Akpınar, 2014).

Komşuların uzaklığına göre sınıflandırma etkilerinin değişmesi için uygulanan yaklaşımlar da vardır. Bu işlem için 1'in, hesaplanan uzaklık değerinin karesine bölünmesi gerekir. Yani ağırlık için $w$, uzaklık için $d$ kullanılırsa;

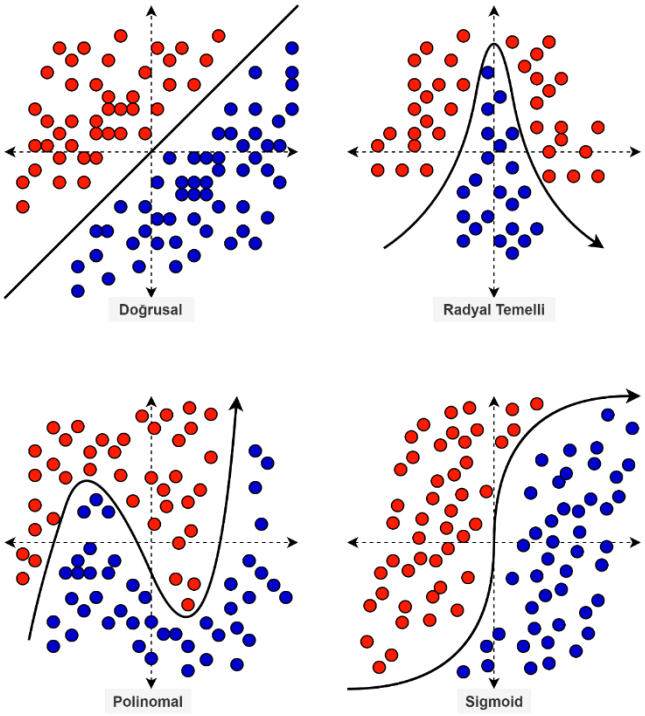
$$w = {}^{1}/_{d^2}$$

eşitliğinden bahsedilebilir. Buradaki ağırlık değeri, sınıflandırma ataması için kullanılarak daha hassas bir atama yapılabilir (Altunkaynak, 2017).

**Destek vektör makineleri.** Destek vektör makinesi (Support vector machines - SVM) algoritması, destek vektör sınıflayıcısı ve destek vektör regresyoncusunu içerir. İstatistiksel öğrenme teorisine dayanan bu algoritmanın temelleri 1960'lı yıllarda Vapnik – Chervonenskis teorisi ile atılmış olsa da algoritma 1992 yılında yine Vapnik, Boser ve Guyon tarafından geliştirilmiştir. Temel olarak DVM algoritmasında, iki grubun düzlemde gösterilmesi durumunda, iki grubu ayıracak bir çizgi çizilmesi ve gelecek verilerin bu çizgiye göre sınıflanması

gerekmektedir. Algoritma temel olarak bu çizginin nereye çizileceğini belirler. DVM algoritmasını iki boyutlu ve doğrusal düzlemde anlatmak ve hayal etmek kolay olsa da esasen çok boyutlu ve doğrusal olmayan sınıflandırma için tercih edilmekte ve kullanılmaktadır (Akpınar, 2014; Şeker, 2013; Wu ve Kumar, 2009).

Algoritmada, doğrusal olarak ayrılabilecek verilerin sınır çizgisinin sonsuz sayıda çizilmesi mümkündür. Ancak karar doğrusunun en iyi sınıflamayı yapabilmesi, iki sınıftaki nokta veya noktalara, olabilecek en uzak noktada olması ile mümkün olabilmektedir. Karar sınırına en yakın olan bağımsız değişken verileri, destek noktası/noktaları olarak adlandırılmaktadır. Bazı durumlarda karar sınırının çizilmesi *aylak değişken (slack veriable)* ile mümkün olmaktadır. Aylak değişken; doğrusal karar çizgisi çekilebilmesi için herhangi veri noktasının, sınırın diğer tarafında kalması ile oluşur. Aylak değişkenin olduğu durumlarda, bu değişkenin karar çizgisi çekilmesini en az şekilde etkilemesi için optimizasyon yapılmalıdır. Algoritmanın çizeceği karar sınırının belirlenmesi ise *çekirdek fonksiyonuna (kernel)* bağlı bir sınır belirleme yapılabilmektedir. Çekirdek fonksiyonu; sınırın çizileceği biçimi açıklamaktadır. Çekirdek fonksiyonlarına örnek olarak doğrusal, polinomal, radyal ve sigmoid verilebilir (Haykin, 2004; Wu ve Kumar, 2009).




*Şekil 3.* Çekirdek fonksiyonu örnekleri

**Naive bayes.** Naive bayes algoritması, 18. yüzyılda yaşamış Thomas Bayes'e ait olan Bayes teoremine dayanan bir sınıflayıcıdır. Bayes teoremi basitçe, herhangi B olayına koşullu bir A olayı ile A olayına koşullu bir B olayının olasılıklarının birbirinden farklılığının ilişkisi olarak tanımlanabilir.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

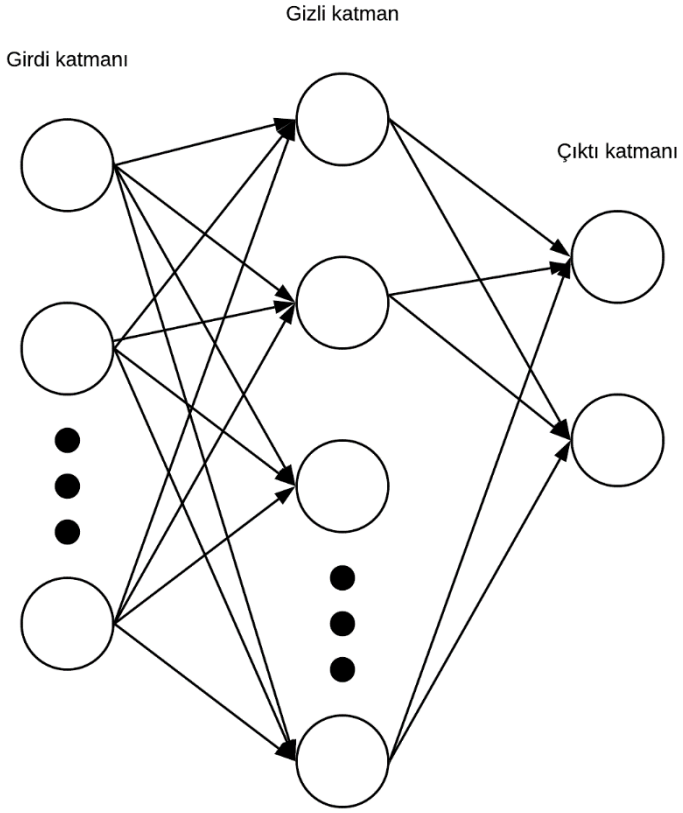
Burada; $P(A \mid B)$ A için B'nin koşullu olasılığı, $P(A)$ A'nın önsel(marjinal) olasılığı, $P(B)$ B'nin önsel(marjinal) olasılığı, $P(B \mid A)$ ise B için A'nın koşullu olasılığıdır. Bu formül aracılığıyla, gözlemlenmiş bir B olayının olması durumunda, A için B'nin koşullu olasılığının bulunması mümkündür. Koşullu bağımsızlık şartının gerçek dünyada karşılanması çok zordur. Ama bu durum aynı zamanda algoritmanın öğrenme sürecini hızlandırmaktadır (Dimitoglou, Adams, ve Jim, 2012; Patil ve Sherekar, 2013).

Naive bayes sınıflamasında, eğitim verisinden elde edilen olasılıklara göre bir algoritma oluşturulmalıdır. Bu algoritma, test verisi ve yeni gelen verideki sınıflanacak özelliği olasılıklarına göre sınıflandıracaktır. Sınıflandırmanın amacı, değeri bilinen bir bağımsız değişken vektörü ile bağımlı değişkenin değerinin tahminin yapılmasıdır. Bağımlı değişkenin değerinin tahmin edilmesi amacıyla bayes olasılığı en büyük değere sahip sınıf seçilir. Yukarıdaki formülde yer alan A değerinin, bağımsız değişken kabul edilmesi durumunda, daha önce gözlemlenmiş olan B olayında gerçekleşme durumu maksimize edilmelidir. Bu değerin maksimize edilebilmesi, yine formüle dayanarak, B olayının A'da gerçekleşme olasılığının yani $P(B \mid A)$'nın maksimizasyonuna bağlıdır. Dolayısıyla sınıflandırma için maksimizasyonu sağlayacak olan *en büyük sonsal (Maximum A Posteriori, MAP)* seçilmelidir. En büyük sonsal değer

$$\arg\max\{P(B \mid A)P(A)\}$$

şeklinde gösterilir (Altunkaynak, 2017).

**Yapay sinir ağları.** Yapay sinir ağları (artificial neural networks - ANN), insan beyninin fizyolojisinden ilham alarak geliştirilmiş bir modeldir. Biyolojik sinir hücrelerinin, bir diğer hücreyi tetiklemesi, sinirsel aktarımı devam ettirmesi için bir eşik değer kadar uyarılması gerekir. Yapay sinir hücresinin uyarımı da aynı şekilde gerçekleşmektedir. Genel olarak bir yapay sinir ağı modeli, n adet *katman (layer)* ve her katmanda değişik sayılarda olabilen hesaplama elemanlarından oluşur. Bir yapay sinir ağı modelinde bulunan hesaplama elemanları ise, *yapay sinir hücresi (artificial neuron), düğüm (node), birim (unit) veya işlem elemanı (processing element)* olarak adlandırılır (Akpınar, 2014).



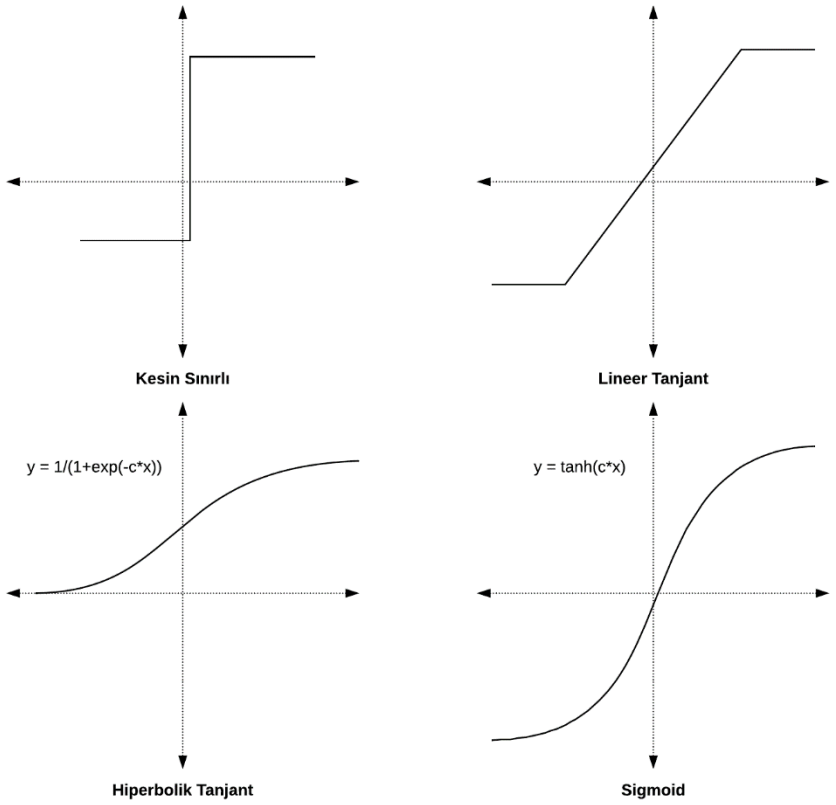

*Şekil 4.* Örnek bir YSA modeli

Her düğümün, birden fazla girdisi ve bir çıktısı vardır. Bu çıktı, diğer düğümlere iletilmelidir. Önceki düğümlerden iletilen veriler, düğümün hesaplama ve sonuç gönderimi yapmasına olanak sağlar. Hesaplama işlemi, düğüme gelen girdilerin ağırlıklandırılması ile olur. Engelleyici ve uyarıcı bağlantıların, negatif ve pozitif çarpanları olabilir. Düğümlerde, negatif ve pozitif çarpanlara sahip uyarıcı ve engelleyici bağlantıların yanında, *kazandırıcı (gain), söndürücü (quenching), ve*

*uyandırıcı (nonspesific arousal)* özel bağlantılar da olabilir. Ağırlık değerleriyle çarpılmış girdi değerlerinin toplanması ile herhangi düğümün *net girdi (net input)* değeri oluşmaktadır. Net girdi değerinin hesaplanmasının ardından, düğüm için *faaliyet değeri (activation value)* hesaplanmalıdır. Bu hesaplama için, herhangi $t$. zamandaki faaliyet değeri için, $t-1$ zamandaki faaliyet değeri ve $t$. zamandaki net girdi değerinin bir fonksiyonu hesaplanmalıdır (Hsu, Gupta, ve Sorooshian, 1995).

$$x_i(t) = F_i\big(x_i(t-1), net_i(t)\big)$$

Burada; $x_i(t)$, $t$. zamandaki faaliyet değerini; $x_i(t-1)$, $t-1$ zamandaki faaliyet değerini; $net_i(t)$ ise $t$. zamandaki net girdi değerini göstermektedir. Faaliyet değeri, birçok modelde net girdi değerine eşdeğer olarak kullanılır. Yapay sinir ağı modeline göre çıktı değerinin elde edilmesinde, doğrusal olmayan fonksiyonlar kullanılabilmektedir. Bu fonksiyonlara örnek olarak, *kesin sınırlı (hard limiters)*, *doğrusal tanjant, hiperbolik tanjant ve sigmoid* verilebilir (Lippmann, 1987).




*Şekil 5.* YSA için örnek eşik fonksiyonları

**Sınıflandırma kalitesi ölçütleri.** Veri madenciliğinde, sınıflandırma algoritmalarının kalitesinin ölçülmesi, hangi yöntemin daha iyi olduğunu belirlemek için önemlidir. Sınıflandırmada kullanılan temel ölçüt, doğru ve yanlış sınıflandırılan gözlemlerin sayısı *karışıklık matrisi (confusion matris)* karşılaştırılması ile yapılmaktadır (Altunkaynak, 2017).

Tablo 1

*Karışıklık Matrisi Örneği*

| | | Gerçek durum | |
|---|---|---|---|
| | | *Doğru (+)* | *Yanlış (-)* |
| Yöntemin tahmini | Doğru (+) | Doğru pozitif (DP) | Yanlış pozitif (YP) |
| | Yanlış (-) | Yanlış negatif (YN) | Doğru negatif (DN) |

Karışıklık matrisinden elde edilecek DP, YP, YN ve DN değerleri farklı formüllerle farklı doğruluk ölçütlerine dönüştürebilir. Bunlardan ilki ve en basiti *doğruluk (accuracy)* oranıdır.

$$Doğruluk = DP + DN/N$$

Doğru sınıflanmış değerlerin sayısının, tüm değerlere bölünmesi ile elde edilen doğruluk oranı, algoritmanın doğru sınıflama sayısı ve kalitesi hakkında önemli bir bilgi verir. Ayrıca, doğruluk oranını 1'e tamamlayan değer de hata oranı olarak kullanılır ve algoritmanın yanlış sınıflamaları hakkında fikir verir. Bu değerlerin yanında, karışıklık matrisinden elde edilen değerlerle hesaplanan, *duyarlık (sensitivity), özgüllük (specificity), kesinlik (precision), F ölçüsü ve Matthews korelasyon katsayısı* da kullanılmaktadır (Coşkun ve Baykal, 2011).

$$Duyarlık = DP/DP + YN$$

$$Özgüllük = DN/YP + DN$$

$$Kesinlik = DP/DP + YP$$

$$F\ değeri = 2 * duyarlık * özgüllük/duyarlık + kesinlik$$

$$Matthews\ korelason\ katsayısı = \frac{DP * DN - YP * YN}{\sqrt{(DP + YP)(DP + YN)(DN + YP)(DN + YN)}}$$

Sınıflama performansının ölçümünde, kappa istatistiği de kullanılmaktadır. Kappa istatistiği, nominal ölçeklerde, iki gözlemci arasındaki değerlendirme

uyumunu ölçmek için Cohen(1960) tarafından geliştirilmiş bir yöntemdir. Cohen'in kappası, iki gözlemci arasındaki uyumu ölçerken, Fleiss'in kappası ise gözlemci sayısının ikiden fazla olması durumundaki uyumu ölçer. Gözlemcilerin uyumsuzluklarının eşit kabul edilmesi durumunda kappa ölçüme uygundur. Eğer farklı uyumsuzluklara farklı ölçeklendirmeler yapılabilecekse, ağırlıklı kappa ölçüme uygun olacaktır (Fleiss, Cohen, ve Everitt, 1969). Hesaplama ise, gerçekte ne kadar uyum olduğu ve ne kadar uyum beklediğinin değerleri arasındaki fark üzerinden hesaplanmaktadır. Ayrıca, gözlemciler arasındaki uyumun şans eseri olabileceği durumların dikkate alınması, istatistiği daha az yanıltıcı yapmaktadır. Kappa istatistiği, 0 ile 1 arasında değer alabilmektedir. Kappa istatistiği, 0 ile .20 arasında değer aldığında, değer *zayıf* olarak nitelendirilir. Aynı şekilde, .20 ile .40 arası *makul*, .40 ile .60 arası *ortalama*, .60 ile .80 arası *güçlü*, .80 ve üzeri olması durumunda ise *neredeyse mükemmel* olarak nitelendirilir (Viera ve Garrett, 2005). Kappa değerinde yer alan iki gözlemci, veri madenciliğindeki gözlenen ve beklenen değerle değiştirilir. Yani iki gözlemcinin değerlerinin uyumu yerine, gözlenen ve beklenen değerin uyumu kontrol edilir (Aksu, 2018).

$$kappa(k) = \frac{gözlenen\ doğruluk - beklenen\ doğruluk}{1 - beklenen\ doğruluk}$$

Bir diğer performans ölçümü ise karekök hata değeri ile yapılmaktadır. Karekök hata ve göreceli karekök hata değerleri, farklı büyüklükteki hataları ortaya çıkardığı için doğru-yanlış sınıflaması dışında, hataların büyüklüğü ile ilgili fikir vermektedir. Hata kare ortalaması (mean-squared error), gözlenen ve beklenen değerlerin farkının karelerinin toplanması ile elde edilen sayının örnek sayısına bölünmesi ile elde edilir (Witten I. H., Frank, Hall, ve Pal, 2005).

$$MSE(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

Basit bir şekilde hesaplanabilmesi, hataları ikili (1-0) şekilde ele almaması, optimizasyon kolaylığı hata kare ortalamasının olumlu özellikleridir. Hata kare ortalamasının karekökü ise karekök hata değerini verir (Wang ve Bovik, 2009).

$$RMSE(x, y)\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}$$

RMSE, kestirilen değer ile gerçek değerin arasındaki mutlak farklılığa ilişkin istatistiktir. Karekök hata değeri sıfıra yaklaştıkça, modelin uyumu yükselmektedir (Eroğlu ve Kelecioğlu, 2015).

## İlgili Çalışmalar

Aksu (2018), Türkiye veri setinde yer alan 5895 öğrencinin PISA (2015) fen okuryazarlığı başarısını tahmin etmede Decision Stump, Hoeffding Tree, J.48, Lojistik Model, RepTree, Rastgele Orman, Rastgele Ağaç ve Ridge Lojistik Regresyon yöntemlerini kullanmıştır. Kullanılacak veri setini test ve eğitim verisi olarak ayırmada farklı yöntemler kullanılmıştır. 10 katlı çapraz geçerleme, veri setini belirli bir yüzde ile (%5, %10, %15, %20, %25, %30 ve %35) bölme ve tüm veri setini hem eğitim hem test verisi olarak kullanma yöntemleri kullanılmıştır. Elde edilen sonuçlara göre verilerin eğitim ve test verisi olarak bölünmesi sırasında veri setinin en az %15'inin test verisi olarak kullanılması önerilmektedir. Daha kararlı ve daha doğru kestirimler için ise çapraz geçerleme kullanılması ve katman sayısının en az 10 olması önerilmektedir.

Şengür (2013)'de yaptığı çalışmada, Fırat Üniversitesi Eğitim Fakültesi Bilgisayar ve Öğretim Teknolojileri bölümünden mezun olmuş 127 öğrencinin yılsonu notları kullanılarak mezuniyet notları yapay sinir ağları ve karar ağacı algoritmalarıyla tahmin edilmeye çalışılmıştır. Bu tahmin için iki senaryo oluşturulmuştur. İlk senaryoda öğrencilerin bir ve ikinci sınıfa ait yılsonu notları, ikinci senaryoda öğrencilerin bir, iki ve üçüncü sınıf yılsonu notları kullanılmıştır. Yapılan analizler sonuçlarında yapay sinir ağları algoritmasının her iki senaryoda da karar ağaçlarından daha başarılı olduğu tespit edilmiştir. Ayrıca ikinci senaryonun iki algoritma için de daha başarılı sonuçlar verdiği görülmüştür.

Şengül (2011)'de yaptığı çalışmada Türk öğrencilerin PISA 2009 öğrenci anketine verdikleri yanıtlarla okuma becerilerini açıklayan değişkenleri belirlemek amacıyla CHAID analizi yapmıştır. Analizde PISA 2009'a katılan 4996 Türk öğrenciye ait veriler kullanılmıştır. Araştırma sonucunda Türk öğrencilerin okuma becerilerini açıklayan en önemli değişkenin "edebi eserlere sahip olma" olduğu

görülmüştür. Ayrıca okul türü, sınıf düzeyi, cinsiyet, öğrencilerin çalışma alışkanlıklarıyla ilgili bir madde olan “çalışırken anlamadığım bir şey olursa, bunu açıklığa kavuşturmak için ek bilgi ararım” maddesi, öğrencilerin okul ders saati dışında çalışmaya ayırdıkları zaman değişkenlerinden Matematik, Türkçe, Fen ve bu dersler dışındaki diğer derslere çalışma maddesi diğer açıklayıcı değişkenler olarak belirtilmiştir.

Aksu ve Güzeller’in (2016) çalışmasında ise PISA 2012’ye katılan 15 yaşındaki 1391 Türk öğrenciden elde edilen verilerle, matematik okuryazarlığı bakımından başarılı ve başarısız öğrencileri derse ilişkin ilgi, tutum, motivasyon, algı, öz yeterlik, kaygı ve çalışma disiplini değişkenlerine göre CHAID analizi ile sınıflandırmak ve bu değişkenlerin sınıflandırmadaki etkisi ortaya konulmak istenmiştir. Araştırma sonucunda elde edilen bulgulara dayalı olarak, matematik okuryazarlığında, Türkiye örneklemi için özellikle öz yeterlik algısı, derse ilişkin tutum ve kaygı durumları ile çalışma disiplini konuları üzerinde durulması önerilmiştir.

Naik ve Samant (2016)’da yaptıkları çalışmada, Hindistan Karaciğer Hastaları Veri Seti’nde yer alan 583 veriyle, Decision Tree, Navie Bayes ve K en yakın komşu algoritmalarını kullanarak, açık kaynak kodlu Rapid Miner, WEKA, Tanagra, Orange ve Knime programlarından elde edilen sonuçları karşılaştırmıştır. Elde edilen sonuçlara göre, Decision Tree algoritmasında en yüksek yüzdeli doğruluk Knime’da, en düşük yüzdeli doğruluk Orange’ta, Naive Bayes algoritmasında en yüksek yüzdeli doğruluk Knime’da, en düşük yüzdeli doğruluk WEKA’da, K en yakın komşu algoritmasında en yüksek yüzdeli doğruluk WEKA’da, en düşük yüzdeli doğruluk Orange’ta çıkmıştır.

Hussain ve diğerlerinin (2018) yaptığı çalışmada ise Dibrugarh Üniversitesi’ne giriş için kullanılan Ortak Giriş Sınavı (CEE)’ndan elde edilmiş 666 veriden yararlanılmıştır. Veri seti kullanılarak, Orange, WEKA ve R Studio programlarıyla çeşitli kümeleme ve sınıflandırma yöntemleri karşılaştırılmıştır. Sınıflama algoritmalarından olan Yapay Sinir Ağları, bu veri setinde en yüksek oranda doğruluğu vermiştir. Ayrıca kümeleme araçları içerisinde K-means kümeleme yönteminin başarımının daha iyi olduğu görülmüştür.

Devasia, Vinushree ve Hedge'nin (2016)'da yaptığı çalışmada Mysuru'daki Amrita Vishwa Vidyapeetham Üniversitesi'ndeki 700 öğrenciden elde edilmiş web tabanlı veriler kullanılmıştır. Veriler Naive Bayes, Regresyon, Decision Tree ve Yapay Sinir Ağları kullanılarak sınıflandırılmıştır. Öğrencilerin yeni dönemdeki, durumlarını tahmin etmede en iyi algoritma Naive Bayes olarak bulunmuştur.

İlgili araştırmalar ve alanyazın incelendiğinde, paket programların sıklıkla kullanıldığı görülmektedir. Ancak veri madenciliği analizi yapan paket programların çıktı sonuçlarının karşılaştırmasına ilişkin yeterli sonuç bulunmamaktadır. Özellikle de eğitimde veri madenciliği söz konusu olduğunda paket program çıktılarının karşılaştırılmasına ilişkin çalışmaya rastlanılmamıştır. Alayazın taramasında paket programların karşılaştırılması ile ilgili çalışma olarak Naik ve Samant (2016)'ın 583 veriyle Hindistan Karaciğer Hastaları Veri Seti'ndeki çalışmasına rastlanmıştır. Ancak sözü edilen çalışma oldukça kapsamlı olmasına karşın sağlık verileri üzerine yapılmış bir çalışmadır. Bu çalışmada ise Türkiye'de elde edilmiş eğitim verileri ve çok daha büyük bir veri seti ile (11264 kişilik örneklem) paket programlar karşılaştırılmıştır. Eğitim verilerinde karşılaştırma yapılmış olması, gelecekte yapılacak çalışmaların da yardımıyla, Türkiye'de eğitimde veri madenciliği alanına katkı sağlaması açısından önemli olacaktır.

# Bölüm 3
# Yöntem

Çalışmada ABİDE (2018) verileri kullanılarak, WEKA ve Orange programlarında bulunan k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları yöntemleri ile elde edilen, doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve ROC eğrisi altında kalan alan değerleri karşılaştırılmıştır. ABİDE'de yer alan Türkçe testinden elde edilen Türkçe puanlarının tahmin edilmesi için öğrencilerin anket sorularına verdiği cevaplar kullanılmıştır. Verilerin ön işlemesi yapıldıktan sonra programların kabul ettiği formatlara dönüştürülmüş ve analizlere başlanmıştır. Analizlerin ardından her bir paket programda elde edilen sonuçlar hem kendi içerisinde yöntemler bağlamında hem de iki paket programdaki aynı yöntemlerin sonuçları dikkate alınarak karşılaştırmalar yapılmıştır.

Bu kapsamda, söz konusu modelleri ve programları test etme amacını taşımasından dolayı çalışma betimsel araştırma niteliğindedir.

## Araştırmanın Evreni ve Örneklemi

Bu çalışmanın evrenini, ABİDE (2018) sınavına girmiş tüm öğrenciler oluşturmaktadır. Çalışma örneklemi ise, üç kitapçıktan birinde (A Kitapçığı) yer alan tüm öğrencilere ait veriden oluşmaktadır.

## Verilerin Elde Edilme Süreci

Çalışmada ABİDE (2018) sınavı ile elde edilmiş veriler kullanılmıştır. Bunun sebebi; geçerlik ve güvenirliği yeterli düzeyde olan bir test kullanılmak istenilmesidir.

ABİDE (2018) sınavını oluşturan üç kitapçıktan birinde yer alan 11264 kişilik örneklem, Millî Eğitim Bakanlığı'ndan istenmiş ve veriler temizlenerek işleme koyulmuştur.

## Veri Toplama Araçları

ABİDE (2018) sınavında yer alan Türkçe alt testi puanı bağımlı değişken, öğrenci anketindeki sosyoekonomik düzey, okula yönelik tutum, akran zorbalığına maruz kalma indisi, aile ilgisi, aile baskısı, her bir derse yönelik tutum, her bir derse verilen değer ve her bir derse ilişkin öz-yeterlik algısı puanları ise bağımsız değişken olarak ele alınmıştır. Dolayısıyla veri toplama araçları, ABİDE (2018) Türkçe başarı testi ve öğrenci anketi olarak ifade edilebilir. Başarı testinin yaklaşık yarısı çoktan seçmeli diğer yarısı ise açık uçlu sorulardan oluşmaktadır. Başarı testinin geçerlik ve güvenirliğine ilişkin hesaplanan değerlerin, testin geçerli ve güvenilir, puanlayıcılar arasındaki tutarlılığın yüksek olduğunu gösterdiği belirtilmiştir (MEB, 2019). Bahsi geçen öğrenci anketi ve başarı testlerinin yanında okul ve yönetici anketi de uygulanmıştır.

## Verilerin Analizi

Verilerin analize uygun hale getirilmesi için öncelikle alt testlere verilen cevapların yer aldığı sütunlar silinmiştir. Ardından, kategorik halde yer alan sütunlar düzenlenmiştir. Bu işlem sırasında; 4 alt test için ".... dersi için bu öğretim yılı içerisinde okulunuzda açılan Destekleme ve Yetiştirme Kurslarına katıldınız mı?" sorusuna verilen yanıtlar “Hiç katılmadım./ Birinci dönem katıldım./ İkinci dönem katıldım./ Her iki dönemde de katıldım.” şeklinde kategorilendirilmişken, bu kategorilendirmede yer alan birinci ve ikinci dönemde katılmanın, bu çalışma için anlamsız olması sebebiyle iki seçenek birleştirilmiş ve “Hiç katılmadım./Bir dönem katıldım./İki dönemde de katıldım.” şeklinde düzenlenmiştir. Ardından, veride hedef değişken olarak kullanılacak değişkenlerin oluşturulması için ilk olarak Türkçe alt testinin puanlarına göre ABİDE (2018) raporunda yer alan 5’li sınıflandırma kriterleri gözetilerek puanlar sınıflandırılmıştır. Bir diğer hedef değişken olan 2’li sınıflandırılmış puanlar için ise, Türkçe alt testinden elde edilmiş puanların ortalaması, ayırıcı olarak kullanılmıştır.

Verilerin analizi için veri setinin temizlenmiş ve uygulamaya hazır hale getirilmiştir. Bu süreçte kayıp veri dağılımının tamamen rastgeleliği Little's MCAR testi ile test edilmiş ve kayıp verilerin tamamen rastgele dağıldığı görülmüştür

($\chi 2$=4653,305, sd=4569, p=,188). Verilerin çift bazında *(pairwise)* silinmesi uygun görülmüştür. Ardından WEKA ve Orange programlarında k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları yöntemleri ile analizler yapılmıştır.

Ardından, araştırma kapsamında belirlenen sınıflama algoritmalarıyla, temel altı, temel, orta, orta üstü ve ileri olarak sınıflanan öğrencilerden elde edilen ölçme sonuçlarının güvenirlik ve geçerlik değerleri incelenmiştir. Ardından öğrenciler ortalama puanlara göre başarılı ve başarısız olarak ikili (binary) şekilde sınıflandırılarak ve algoritmaların sınıflama performans değerleri tekrar incelenmiştir. Analizler sonrasında tüm yöntemlerin doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve ROC eğrisi altında kalan alan değerleri karşılaştırılmıştır.

# Bölüm 4

# Bulgular ve Yorumlar

Araştırmada öncelikle, verilerin veri madenciliğine hazırlanması amacıyla gereken ön işlemler yapılmıştır. Ön işleme işleminin sonucunda elde edilen veri, Türkçe alt test puanı ile birlikte 35 değişkenden oluşmaktadır. Bu değişkenlerden 32'si bağımsız değişken olarak ele alınmış ve öğrencilerin Türkçe alt testi puanlarından ikili ve beşli sınıflandırma ile elde edilen puanları yordamak için kullanılmıştır.

Tablo 2

*Kullanılan Veri Kodları ve Kısa Açıklamaları*

| Kod | Kısa açıklama | Kod | Kısa açıklama |
|---|---|---|---|
| TUR_500 | Öğrencilerin Türkçe alt testinden elde ettiği puan | okultutum | Okula karşı tutum |
| anneegitim | Annenizin eğitim durumu nedir | zorbalikk | Zorbalığa maruz kalma düzeyi |
| egitimhedefi | Eğitiminizdeki hedefiniz nedir | aileilgisi | Ailenin ilgi düzeyi |
| kitapsayi | Öğrencilerin evlerindeki kitap sayısı | ailebaskisi | Ailenin baskı düzeyi |
| turkurs | Türkçe dersi Destekleme ve Yetiştirme Kurslarına katılım | matsevgisi | Matematik dersi sevgi düzeyi |
| matkurs | Matematik dersi Destekleme ve Yetiştirme Kurslarına katılım | matozyet | Matematik dersi özyeterlilik düzeyi |
| fenkurs | Fen ve Teknoloji dersi Destekleme ve Yetiştirme Kurslarına katılım | matverdeger | Matematik dersine verilen değer |
| soskurs | Sosyal Bilgiler dersi Destekleme ve Yetiştirme Kurslarına katılım | fensev | Fen Bilgisi dersi sevgi düzeyi |
| turodevverme | Türkçe-Ödev Verme Sıklığı | fenozyet | Fen Bilgisi dersi özyeterlilik düzeyi |
| matodevverme | Matematik-Ödev Verme Sıklığı | fenverdeger | Fen Bilgisi dersine verilen değer |
| fenodevverme | Fen ve Teknoloji-Ödev Verme Sıklığı | tursevgisi | Türkçe dersi sevgi düzeyi |
| sosodevverme | Sosyal Bilgiler-Ödev Verme Sıklığı | turozyet | Türkçe dersi öz yeterlilik düzeyi |
| turodevsure | Türkçe-Ödev Yapmak İçin Harcanan Süre | turverdeger | Türkçe dersine verilen değer |
| matodevsure | Matematik-Ödev Yapmak İçin Harcanan Süre | sossevgisi | Sosyal Bilgiler dersi sevgi düzeyi |
| fenodevsure | Fen ve Teknoloji-Ödev Yapmak İçin Harcanan Süre | sosozyet | Sosyal Bilgiler dersi özyeterlilik düzeyi |
| sosodevsure | Sosyal Bilgiler-Ödev Yapmak İçin Harcanan Süre | sosverdeger | Sosyal Bilgiler dersine verilen değer |
| sosyoekonomik | Sosyoekonomik düzey | TUR_2 | Türkçe alt testi puanından elde edilmiş ikili sınıflandırma değeri |
| | | TUR_5 | Türkçe alt testi puanından elde edilmiş beşli sınıflandırma değeri |

İlk olarak kayıp verinin durumunun tespiti için kayıp veriler silinerek ve silinmeden analizler gerçekleştirilmiştir. Analizlerin sonucunda kayıp verilerin silinmesinin algoritmaları etkilemediği görülmüştür.

## Birinci Alt Problem: ABİDE 2018 Türkçe Alt Testi Puanlarının 5'li Sınıflandığı Duruma İlişkin Bulgular

*ABİDE sınavında Türkçe alt testi puanları beşli sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak; WEKA ve Orange programlarında k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları algoritmaları ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve* ROC eğrisi altında kalan alan değerler *karşılaştırılmıştır.*

Çalışmanın birinci alt probleminde 10 katlı çapraz geçerleme yöntemiyle her iki programda da analizler gerçekleştirilmiştir.

*K-en yakın komşu algoritmasına ilişkin sonuçlar*

K-en yakın komşu algoritmasında belirlenen $k$ değeri 51'dir. Öklidyen metrikle birlikte uzaklık değerine göre ağırlıklandırma kullanılmıştır. K-en yakın komşu algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 3'te verilmiştir.

Tablo 3

*K-En Yakın Komşu Algoritmasına Ait 5'li Sınıflama Sonuçları*

| *k-en yakın komşu* | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *4913* | *5041* |
| Doğru sınıflama oranı | *.436* | *.448* |
| Oranın güven aralığı (%95) | *.428 - .446* | *.439 - .458* |
| Kappa istatistiği | *.119* | *.141* |
| ROC eğrisi altında kalan alan | *.629* | *.646* |

Tablo 3 incelendiğinde k-en yakın komşu algoritmasının, WEKA'da, örnekleminin %43,6'sını doğru sınıfladığı, aynı koşullarda Orange'ın ise %44,8'ini

doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesiştiği görülmektedir. WEKA 11264 kişilik örneklemden 4913'ünü doğru sınıflarken, Orange 5041 doğru sınıflama yapmıştır. WEKA'da kappa istatistiği .119 iken Orange'ta .141 olmuştur. ROC eğrisi altında kalan alan WEKA'da .629 iken Orange'ta .646 olmuştur.

Programların k-en yakın komşu algoritmasında sınıflandırma sonuçlarının karışıklık matrisleri Tablo 4'te verilmiştir.

Tablo 4

*K-En Yakın Komşu Algoritmasının 5'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| *Beklenen* | 1 | *0* | *2* | *35* | *12* | *2* |
| | 2 | *0* | *19* | *604* | *180* | *1* |
| | 3 | *0* | *17* | *2311* | *1593* | *43* |
| | 4 | *0* | *1* | *1836* | *2346* | *147* |
| | 5 | *0* | *0* | *640* | *1238* | *237* |
| *Orange* | | *Gözlenen* | | | | |
| *Beklenen* | 1 | *0* | *0* | *32* | *14* | *5* |
| | 2 | *0* | *10* | *614* | *179* | *1* |
| | 3 | *0* | *8* | *2448* | *1467* | *41* |
| | 4 | *0* | *4* | *1832* | *2253* | *241* |
| | 5 | *0* | *0* | *570* | *1715* | *330* |

Tablo 4'teki karışıklık matrisleri incelendiğinde beşli sınıflandırmada 1 olarak sınıflandırılan ve temelaltı olarak nitelenen öğrencilerin temel, orta, ortaüstü ve ileri olarak (2, 3, 4, 5) sınıflandığı görülmüştür. Orta olarak sınıflanması beklenen 3 şeklinde kodlanan öğrencilerin büyük bir kısmı ise, 4 yani ortaüstü olarak sınıflandırılmıştır. Aynı şekilde ortaüstü öğrencilerin, orta olarak sınıflandırıldığı görülmüştür. İleri olarak sınıflandırılması beklenen ve 5 olarak kodlanan öğrencilerin ortaüstü yani 4 olarak sınıflandırıldığı tespit edilmiştir.

WEKA, 2 ve 4 olarak kodlanmış öğrencilerde, sırasıyla 9 ve 93 fazla doğru sınıflama yapmışken; Orange, 3 ve 5 olarak kodlanmış öğrencilerde, sırasıyla 137 ve 93 fazla doğru sınıflama yapmıştır. 1 olarak kodlanmış öğrenciler ise her iki algoritmada da doğru sınıflanamamıştır.

*Rastgele orman algoritmasına ilişkin sonuçlar*

Rastgele orman algoritmasında kullanılacak ağaç sayısı 100 olarak belirlenmiştir. Tekil ağaçların derinliği ise 7 ile sınırlandırılmıştır. Rastgele orman algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 5'te verilmiştir.

Tablo 5

*Rastgele Orman Algoritmasına Ait 5'li Sınıflama Sonuçları*

| *Rastgele orman* | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *5297* | *5349* |
| Doğru sınıflama oranı | *.470* | *.475* |
| Oranın güven aralığı (%95) | *.461 - .480* | *.466 - .484* |
| Kappa istatistiği | *.173* | *.175* |
| ROC eğrisi altında kalan alan | *.673* | *.685* |

Tablo 5 incelendiğinde rastgele orman algoritmasının, WEKA'da, örnekleminin %47'sini doğru sınıfladığı, aynı koşullarda Orange'ın ise %47,5'ini doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesiştiği görülmektedir. WEKA 11264 kişilik örneklemden 5297'ünü doğru sınıflarken, Orange 5349 doğru sınıflama yapmıştır. WEKA'da kappa istatistiği .173 iken Orange'ta .175 olmuştur. ROC eğrisi altında kalan alan WEKA'da .673 iken Orange'ta .685 olmuştur.

Programların rastgele orman algoritmasında sınıflandırma sonuçlarının karışıklık matrisleri Tablo 6'da verilmiştir.

Tablo 6

*Rastgele Orman Algoritmasının 5'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| *Beklenen* | 1 | *0* | *1* | *35* | *11* | *4* |
| | 2 | *0* | *23* | *600* | *181* | *0* |
| | 3 | *0* | *13* | *2062* | *1836* | *53* |
| | 4 | *0* | *1* | *1256* | *2827* | *246* |
| | 5 | *0* | *0* | *260* | *1470* | *385* |
| *Orange* | | *Gözlenen* | | | | |
| *Beklenen* | 1 | *0* | *0* | *34* | *14* | *3* |
| | 2 | *0* | *2* | *628* | *174* | *0* |
| | 3 | *0* | *2* | *2071* | *1870* | *21* |
| | 4 | *0* | *0* | *1190* | *2960* | *180* |
| | 5 | *0* | *0* | *222* | *1577* | *316* |

Tablo 6'daki karışıklık matrisleri incelendiğinde beşli sınıflandırmada 1 olarak sınıflandırılan ve temelaltı olarak nitelenen öğrencilerin temel, orta, ortaüstü ve ileri olarak (2, 3, 4, 5) sınıflandığı görülmüştür. Orta olarak sınıflanması beklenen 3 şeklinde kodlanan öğrencilerin büyük bir kısmı ise, 4 yani ortaüstü olarak sınıflandırılmıştır. Aynı şekilde ortaüstü öğrencilerin, orta olarak sınıflandırıldığı görülmüştür. İleri olarak sınıflandırılması beklenen ve 5 olarak kodlanan öğrencilerin ortaüstü yani 4 olarak sınıflandırıldığı tespit edilmiştir.

WEKA, 2 ve 5 olarak kodlanmış öğrencilerde, sırasıyla 21 ve 69 fazla doğru sınıflama yapmışken; Orange, 3 ve 4 olarak kodlanmış öğrencilerde, sırasıyla 9 ve 133 fazla doğru sınıflama yapmıştır. 1 olarak kodlanmış öğrenciler ise her iki algoritmada da doğru sınıflanamamıştır.

*Destek vektör makinesi algoritmasına ilişkin sonuçlar*

Destek vektör makinesi algoritması için model uyumuna sağladığı katkı dolayısıyla radyal tabanlı çekirdek fonksiyonu kullanılmıştır. Destek vektör makinesi algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 7'de verilmiştir.

Tablo 7

*Destek vektör makinesi algoritmasına ait 5'li sınıflama sonuçları*

| *Destek vektör makinesi* | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *5409* | *5335* |
| Doğru sınıflama oranı | *.480* | *.474* |
| Oranın güven aralığı (%95) | *.471 - .490* | *.465 - .483* |
| Kappa istatistiği | *.185* | *.185* |
| ROC eğrisi altında kalan alan | *.589* | *.687* |

Tablo 7 incelendiğinde destek vektör makinesi algoritmasının, WEKA'da, örnekleminin %48'ini doğru sınıfladığı, aynı koşullarda Orange'ın ise %47,4'ünü doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesiştiği görülmektedir. WEKA 11264 kişilik örneklemden 5409'unu doğru sınıflarken, Orange 5335 doğru sınıflama yapmıştır. kappa istatistiği WEKA'da ve Orange'ta .185 olmuştur. ROC eğrisi altında kalan alan WEKA'da .589 iken Orange'ta .687 olmuştur.

Tablo 8

*Destek Vektör Makinesi Algoritmasının 5'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| *Beklenen* | 1 | *0* | *0* | *35* | *12* | *4* |
| | 2 | *0* | *0* | *643* | *161* | *0* |
| | 3 | *0* | *0* | *2142* | *1801* | *21* |
| | 4 | *0* | *0* | *1218* | *2925* | *187* |
| | 5 | *0* | *0* | *239* | *1534* | *342* |
| *Orange* | | *Gözlenen* | | | | |
| *Beklenen* | 1 | *0* | *0* | *36* | *10* | *5* |
| | 2 | *0* | *19* | *628* | *156* | *1* |
| | 3 | *0* | *16* | *2188* | *1698* | *62* |
| | 4 | *0* | *5* | *1394* | *2609* | *322* |
| | 5 | *0* | *0* | *303* | *1293* | *519* |

Tablo 8'deki karışıklık matrisleri incelendiğinde beşli sınıflandırmada 1 olarak sınıflandırılan ve temelaltı olarak nitelenen öğrencilerin orta, ortaüstü ve ileri olarak (3, 4, 5) sınıflandığı görülmüştür. Temel olarak sınıflandırılması beklenen ve 2 şeklinde kodlanan öğrencilerin, çoğunlukla 3 yani orta olarak sınıflandırıldığı görülmüştür. Orta olarak sınıflanması beklenen 3 şeklinde kodlanan öğrencilerin büyük bir kısmı ise, 4 yani ortaüstü olarak sınıflandırılmıştır. Aynı şekilde ortaüstü yani 4 olarak kodlanan öğrencilerin, orta yani 3 olarak sınıflandırıldığı görülmüştür. İleri olarak sınıflandırılması beklenen ve 5 olarak kodlanan öğrencilerin ortaüstü yani 4 olarak sınıflandırıldığı tespit edilmiştir.

WEKA, 4 olarak kodlanmış öğrencilerde, 1316 fazla doğru sınıflama yapmışken; Orange, 2, 3 ve 5 olarak kodlanmış öğrencilerde, sırasıyla 19, 46 ve 177 fazla doğru sınıflama yapmıştır. 1 olarak kodlanmış öğrenciler ise her iki algoritmada da doğru sınıflanamamıştır.

*Naive bayes algoritmasına ilişkin sonuçlar*

Naive bayes algoritması için programlarda bulunan temel algoritmalar kullanılmıştır. Naive bayes algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 9'da verilmiştir.

Tablo 9

*Naive Bayes Algoritmasına Ait 5'li Sınıflama Sonuçları*

| *Naive bayes* | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *4809* | *4127* |
| Doğru sınıflama oranı | *.426* | *.366* |
| Oranın güven aralığı (%95) | *.418 - .436* | *.358 - .376* |
| Kappa istatistiği | *.190* | *.145* |
| ROC eğrisi altında kalan alan | *.651* | *.638* |

Tablo 9 incelendiğinde naive bayes algoritmasının, WEKA'da, örnekleminin %42,6'sını doğru sınıfladığı, aynı koşullarda Orange'ın ise %36,6'sını doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesişmediği görülmektedir. WEKA 11264 kişilik örneklemden 4809'unu doğru sınıflarken, Orange 4127 doğru sınıflama yapmıştır. kappa istatistiği WEKA'da .190 Orange'ta ise .145 olmuştur. ROC eğrisi altında kalan alan WEKA'da .651 iken Orange'ta .638 olmuştur.

Naive bayes algoritmasına ilişkin sonuçların karışıklık matrisi Tablo 10'da verilmiştir.

Tablo 10

*Naive Bayes Algoritmasının 5'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| *Beklenen* | 1 | *0* | *18* | *18* | *6* | *9* |
| | 2 | *16* | *278* | *371* | *125* | *14* |
| | 3 | *46* | *577* | *1706* | *1285* | *530* |
| | 4 | *72* | *292* | *1200* | *1808* | *958* |
| | 5 | *36* | *78* | *272* | *712* | *1017* |
| *Orange* | | *Gözlenen* | | | | |
| *Beklenen* | 1 | *8* | *15* | *9* | *7* | *12* |
| | 2 | *149* | *234* | *221* | *134* | *66* |
| | 3 | *463* | *524* | *1116* | *1303* | *558* |
| | 4 | *406* | *239* | *842* | *1746* | *1097* |
| | 5 | *172* | *55* | *198* | *667* | *1023* |

Tablo 10'daki karışıklık matrisleri incelendiğinde, 2 olarak kodlanmış temel düzeydeki öğrencilerin sınıflandırılmasında ise WEKA'nın Orange'a göre 44 fazladan doğru sınıflama yaptığı görülmektedir. 3 olarak kodlanan orta düzeyde bulunan 590 öğrenciyi WEKA'nın Orange'a göre doğru sınıfladığı görülmektedir. 4 olarak kodlanmış ortaüstü düzeyde bulunan 62 öğrenciyi WEKA'nın Orange'a göre doğru sınıfladığı görülmektedir. 1 ve 5 olarak kodlanmış temelaltı ve ileri düzeyde bulunan öğrencilerin sınıflamasında ise Orange lehine sırasıyla 8 ve 6 fazladan doğru sınıflama yapıldığı görülmektedir.

*Yapay sinir ağları algoritmasına ilişkin sonuçlar*

Yapay sinir ağları algoritması için programlarda bulunan temel algoritmaların üzerine, gizli katmanda 25 sinir ağı ve 20 tekerrür sınırlaması ile oluşturulmuştur. Yapay sinir ağları algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 11'de verilmiştir.

Tablo 11

*Yapay Sinir Ağları Algoritmasına Ait 5'li Sınıflama Sonuçları*

| *Yapay sinir ağları* | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *5141* | *5419* |
| Doğru sınıflama oranı | *.456* | *.481* |
| Oranın güven aralığı (%95) | *.447 - .466* | *.472 - .491* |
| Kappa istatistiği | *.175* | *.204* |
| ROC eğrisi altında kalan alan | *.649* | *.693* |

Tablo 11 incelendiğinde yapay sinir ağları algoritmasının, WEKA'da, örnekleminin %45,6'sını doğru sınıfladığı, aynı koşullarda Orange'ın ise %48,1'ini doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesişmediği görülmektedir. WEKA 11264 kişilik örneklemden 5141'ini doğru sınıflarken, Orange 5419 doğru sınıflama yapmıştır. kappa istatistiği WEKA'da .175 Orange'ta ise .204 olmuştur. ROC eğrisi altında kalan alan WEKA'da .649 iken Orange'ta .693 olmuştur.

Yapay sinir ağları algoritmasına ilişkin sonuçların karışıklık matrisi Tablo 12'de verilmiştir.

Tablo 12

*Yapay Sinir Ağları Algoritmasının 5'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| *Beklenen* | 1 | *0* | *3* | *32* | *10* | *6* |
| | 2 | *0* | *88* | *557* | *150* | *9* |
| | 3 | *0* | *118* | *2244* | *1480* | *122* |
| | 4 | *0* | *31* | *1655* | *2168* | *476* |
| | 5 | *0* | *7* | *429* | *1038* | *641* |
| *Orange* | | *Gözlenen* | | | | |
| *Beklenen* | 1 | *0* | *3* | *32* | *9* | *7* |
| | 2 | *0* | *34* | *618* | *151* | *101* |
| | 3 | *0* | *25* | *2195* | *1643* | *101* |
| | 4 | *0* | *1* | *1358* | *2510* | *461* |
| | 5 | *0* | *0* | *279* | *1156* | *680* |

Tablo 12'deki karışıklık matrisleri incelendiğinde beşli sınıflandırmada 1 olarak kodlanan ve temelaltı olarak nitelenen öğrencilerin temel, orta, ortaüstü ve ileri olarak (2, 3, 4, 5) sınıflandığı görülmüştür. Temel olarak sınıflandırılması beklenen ve 2 şeklinde kodlanan öğrencilerin, çoğunlukla 3 ve 4 yani orta ortaüstü olarak sınıflandırıldığı görülmüştür. Orta olarak sınıflanması beklenen 3 şeklinde kodlanan öğrencilerin büyük bir kısmı ise, 4 yani ortaüstü olarak sınıflandırılmıştır. Aynı şekilde ortaüstü yani 4 olarak kodlanan öğrencilerin, orta yani 3 olarak sınıflandırıldığı görülmüştür. İleri olarak sınıflandırılması beklenen ve 5 olarak kodlanan öğrencilerin çoğunlukla ortaüstü yani 4 olarak sınıflandırıldığı tespit edilmiştir.

WEKA, 2 ve 3 olarak kodlanmış öğrencilerde, sırasıyla 54 ve 49 fazla doğru sınıflama yapmışken; Orange, 4 ve 5 olarak kodlanmış öğrencilerde, sırasıyla 342 ve 39 fazla doğru sınıflama yapmıştır. 1 olarak kodlanmış öğrenciler ise her iki algoritmada da doğru sınıflanamamıştır.

## İkinci Alt Problem: ABİDE 2018 Türkçe Alt Testi Puanlarının 2'li Sınıflandığı Duruma İlişkin Bulgular

*ABİDE sınavında Türkçe alt testi puanları ikili sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak; WEKA ve Orange programlarında k-*

*en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları algoritmaları ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve* ROC eğrisi altında kalan alan değerler *karşılaştırılmıştır.*

Çalışmanın ikinci alt probleminde 10 katlı çapraz geçerleme yöntemiyle her iki programda da analizler gerçekleştirilmiştir.

*K-en yakın komşu algoritmasına ilişkin sonuçlar*

K-en yakın komşu algoritmasında belirlenen $k$ değeri 39'dur. Öklidyen metrikle birlikte uzaklık değerine göre ağırlıklandırma kullanılmıştır. K-en yakın komşu algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 13'te verilmiştir.

Tablo 13

*K-En Yakın Komşu Algoritmasına Ait 2'li Sınıflama Sonuçları*

| k-en yakın komşu | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *7188* | *7324* |
| Doğru sınıflama oranı | *.638* | *.650* |
| Oranın güven aralığı (%95) | *.630 - .648* | *.642 - .660* |
| Kappa istatistiği | *.278* | *.301* |
| ROC eğrisi altında kalan alan | *.698* | *.713* |

Tablo 13 incelendiğinde k-en yakın komşu algoritmasının, WEKA'da, örnekleminin %63,8'ini doğru sınıfladığı, aynı koşullarda Orange'ın ise %65'ini doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesiştiği görülmektedir. WEKA 11264 kişilik örneklemden 7188'ini doğru sınıflarken, Orange 7324 doğru sınıflama yapmıştır. kappa istatistiği WEKA'da .278 Orange'ta ise .301 olmuştur. ROC eğrisi altında kalan alan WEKA'da .698 iken Orange'ta .713 olmuştur.

Tablo 14

*K-En Yakın Komşu Algoritmasının 2'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | |
|---|---|---|---|
| | | 1 | 2 |
| *Beklenen* | 1 | *3666* | *1782* |
| | 2 | *2294* | *3522* |
| *Orange* | | *Gözlenen* | |
| | | 1 | 2 |
| *Beklenen* | 1 | *3699* | *1749* |
| | 2 | *2191* | *3625* |

Tablo 14 incelendiğinde, 2'li sınıflanmış verilerle k-en yakın komşu algoritmasında, Orange'ın 1 olarak kodlanmış öğrencilerde 33 fazla doğru sınıflama yaptığı; WEKA'nın, 2 olarak kodlanmış öğrencilerde 103 fazla doğru sınıflama yaptığı görülmektedir.

*Rastgele orman algoritmasına ilişkin sonuçlar*

Rastgele orman algoritmasında kullanılacak ağaç sayısı 15 olarak belirlenmiştir. Tekil ağaçların derinliği ise 7 ile sınırlandırılmıştır. Rastgele orman algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 15'te verilmiştir.

Tablo 15

*Rastgele Orman Algoritmasına Ait 2'li Sınıflama Sonuçları*

| Rastgele orman | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *7671* | *7620* |
| Doğru sınıflama oranı | *.681* | *.676* |
| Oranın güven aralığı (%95) | *.673 - .690* | *.668 - .686* |
| Kappa istatistiği | *.360* | *.351* |
| ROC eğrisi altında kalan alan | *.751* | *.744* |

Tablo 15 incelendiğinde rastgele orman algoritmasının, WEKA'da, örnekleminin %68,1'ini doğru sınıfladığı, aynı koşullarda Orange'ın ise %67,6'sını doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesiştiği görülmektedir. WEKA 11264 kişilik örneklemden 7671'ini doğru sınıflarken, Orange 7620 doğru sınıflama yapmıştır. kappa istatistiği WEKA'da .360 Orange'ta ise .351 olmuştur. ROC eğrisi altında kalan alan WEKA'da .751 iken Orange'ta .744 olmuştur.

Rastgele orman algoritmasına göre elde edilen sonuçların karışıklık matrisleri Tablo 16'da verilmiştir.

Tablo 16

*Rastgele Orman Algoritmasının 2'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | |
|---|---|---|---|
| | | 1 | 2 |
| *Beklenen* | 1 | *3469* | *1979* |
| | 2 | *1614* | *4202* |
| *Orange* | | *Gözlenen* | |
| | | 1 | 2 |
| *Beklenen* | 1 | *3478* | *1970* |
| | 2 | *1674* | *4142* |

Tablo 16 incelendiğinde, 2'li sınıflanmış verilerle rastgele orman algoritmasında, Orange'ın 1 olarak kodlanmış öğrencilerde 9 fazla doğru sınıflama yaptığı; WEKA'nın, 2 olarak kodlanmış öğrencilerde 60 fazla doğru sınıflama yaptığı görülmektedir.

*Destek vektör makinesi algoritmasına ilişkin sonuçlar*

Destek vektör makinesi algoritması için sigmoid çekirdek fonksiyonu kullanılmıştır. Destek vektör makinesi algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 17'de verilmiştir.

Tablo 17

*Destek Vektör Makinesi Algoritmasına Ait 2'li Sınıflama Sonuçları*

| Destek vektör makinesi | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *7374* | *7069* |
| Doğru sınıflama oranı | *.655* | *.628* |
| Oranın güven aralığı (%95) | *.646 - .664* | *.619 - .637* |
| Kappa istatistiği | *.308* | *.254* |
| ROC eğrisi altında kalan alan | *.654* | *.663* |

Tablo 17 incelendiğinde rastgele orman algoritmasının, WEKA'da, örnekleminin %65,5'ini doğru sınıfladığı, aynı koşullarda Orange'ın ise %62,8'ini doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesişmediği görülmektedir. WEKA 11264 kişilik örneklemden 7374'ünü doğru sınıflarken, Orange 7069 doğru sınıflama yapmıştır. Kappa istatistiği WEKA'da .308 Orange'ta ise .254 olmuştur. ROC eğrisi altında kalan alan WEKA'da .654 iken Orange'ta .663 olmuştur.

Destek vektör makinesi algoritmasına göre elde edilen sonuçların karışıklık matrisler Tablo 18'de verilmiştir.

Tablo 18

*Destek Vektör Makinesi Algoritmasının 2'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | |
|---|---|---|---|
| | | 1 | 2 |
| *Beklenen* | 1 | *3397* | *2051* |
| | 2 | *1839* | *3977* |
| *Orange* | | *Gözlenen* | |
| | | 1 | 2 |
| *Beklenen* | 1 | *3269* | *2179* |
| | 2 | *2016* | *3800* |

Tablo 18 incelendiğinde, 2'li sınıflanmış verilerle destek vektör makinesi algoritmasında, WEKA'nın, 1 olarak kodlanmış öğrencilerde 128, 2 olarak kodlanmış öğrencilerde ise 177 fazla doğru sınıflama yaptığı görülmektedir.

*Naive bayes algoritmasına ilişkin sonuçlar*

Naive bayes algoritması için programlarda bulunan temel algoritmalar kullanılmıştır. Naive bayes algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 19'da verilmiştir.

Tablo 19

*Naive bayes algoritmasına ait 2'li sınıflama sonuçları*

| Naive bayes | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *7449* | *7287* |
| Doğru sınıflama oranı | *.661* | *.647* |
| Oranın güven aralığı (%95) | *.653 - .671* | *.639 - .656* |
| Kappa istatistiği | *.321* | *.291* |
| ROC eğrisi altında kalan alan | *.723* | *.702* |

Tablo 19 incelendiğinde naive bayes algoritmasının, WEKA'da, örnekleminin %66,1'ini doğru sınıfladığı, aynı koşullarda Orange'ın ise %64,7'sini doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesiştiği görülmektedir. WEKA 11264 kişilik örneklemden 7449'unu doğru sınıflarken, Orange 7287 doğru sınıflama yapmıştır. Kappa istatistiği WEKA'da .321 Orange'ta ise .291 olmuştur. ROC eğrisi altında kalan alan WEKA'da .723 iken Orange'ta .702 olmuştur.

Naive bayes algoritmasına göre elde edilen sonuçların karışıklık matrisleri Tablo 20'de verilmiştir.

Tablo 20

*Naive Bayes Algoritmasının 2'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | |
|---|---|---|---|
| | | 1 | 2 |
| *Beklenen* | 1 | *3409* | *2039* |
| | 2 | *1776* | *4040* |
| *Orange* | | *Gözlenen* | |
| | | 1 | 2 |
| *Beklenen* | 1 | *3151* | *2297* |
| | 2 | *1680* | *4136* |

Tablo 20 incelendiğinde, 2'li sınıflanmış verilerle naive bayes algoritmasında, WEKA'nın 1 olarak kodlanmış öğrencilerde 258 fazla doğru sınıflama yaptığı; Orange'ın 2 olarak kodlanmış öğrencilerde 96 fazla doğru sınıflama yaptığı görülmektedir.

*Yapay sinir ağları algoritmasına ilişkin sonuçlar*

Yapay sinir ağları algoritması için programlarda bulunan temel algoritmaların üzerine, gizli katmanda 25 sinir ağı ve 20 tekerrür sınırlaması ile oluşturulmuştur. Yapay sinir ağları algoritmasına göre elde edilen sonuçların güvenirlik değerleri Tablo 21'de verilmiştir.

Tablo 21

*Yapay Sinir Ağları Algoritmasına Ait 2'li Sınıflama Sonuçları*

| Yapay sinir ağları | WEKA | Orange |
|---|---|---|
| Doğru sınıflama sayısı | *7312* | *7708* |
| Doğru sınıflama oranı | *.649* | *.684* |
| Oranın güven aralığı (%95) | *.641 - .659* | *.676 - .694* |
| Kappa istatistiği | *.297* | *.367* |
| ROC eğrisi altında kalan alan | *.709* | *.756* |

Tablo 21 incelendiğinde yapay sinir ağları algoritmasının, WEKA'da, örnekleminin %64,9'unu doğru sınıfladığı, aynı koşullarda Orange'ın ise %68,4'ünü doğru sınıfladığı görülmektedir. Doğru sınıflama oranının güven aralıklarının kesişmediği görülmektedir. WEKA 11264 kişilik örneklemden 7312'sini doğru sınıflarken, Orange 7708 doğru sınıflama yapmıştır. Kappa istatistiği WEKA'da .297 Orange'ta ise .367 olmuştur. ROC eğrisi altında kalan alan WEKA'da .709 iken Orange'ta .756 olmuştur.

Yapay sinir ağları algoritmasına göre elde edilen sonuçların karışıklık matrisleri Tablo 22'de verilmiştir.

Tablo 22

*Yapay Sinir Ağları Algoritmasının 2'li Karışıklık Matrisi (WEKA, Orange)*

| *WEKA* | | *Gözlenen* | |
|---|---|---|---|
| | | 1 | 2 |
| *Beklenen* | 1 | *3374* | *2074* |
| | 2 | *1878* | *3938* |
| *Orange* | | *Gözlenen* | |
| | | 1 | 2 |
| *Beklenen* | 1 | *3590* | *1858* |
| | 2 | *1698* | *4118* |

Tablo 22 incelendiğinde, 2'li sınıflanmış verilerle yapay sinir ağları algoritmasında, Orange'ın, 1 olarak kodlanmış öğrencilerde 216, 2 olarak kodlanmış öğrencilerde ise 180 fazla doğru sınıflama yaptığı görülmektedir.

### Üçüncü Alt Problem: WEKA'dan Elde Edilen Değerlerin Karşılaştırılmasına İlişkin Bulgular

Bu bölümde, WEKA paket programından elde edilen güvenirlik değerleri karşılaştırılmıştır.

WEKA paket programından elde edilen güvenirlik değerleri Tablo 23'te verilmiştir.

Tablo 23

*WEKA'da elde edilen güvenirlik değerleri*

| *WEKA* | Doğru sınıflama oranı | | Güven aralığı %95 | | Kappa istatistiği | | ROC altında kalan | |
|---|---|---|---|---|---|---|---|---|
| | 5'li | 2'li | 5'li | 2'li | 5'li | 2'li | 5'li | 2'li |
| KYK | *.436* | *.638* | *.428 - .446* | *.630 - .648* | *.119* | *.278* | *.629* | *.698* |
| RO | *.470* | *.681* | *.461 - .480* | *.673 - .690* | *.173* | *.360* | *.673* | *.751* |
| DVM | *.480* | *.655* | *.471 - .490* | *.646 - .664* | *.185* | *.308* | *.589* | *.654* |
| NB | *.426* | *.661* | *.418 - .436* | *.653 - .671* | *.190* | *.321* | *.651* | *.723* |
| YSA | *.456* | *.649* | *.447 - .466* | *.641 - .659* | *.175* | *.297* | *.649* | *.709* |

Tablo 23 incelendiğinde, 5'li sınıflamada en yüksek doğru sınıflama oranına sahip algoritmanın destek vektör makinesi olduğu görülmektedir. Ardından sırasıyla rastgele orman, yapay sinir ağları, k-en yakın komşu ve naive bayes algoritmaları gelmektedir. 5'li sınıflama verilerinde, en düşük doğru sınıflama oranına sahip olsa da en yüksek kappa istatistiği değeri naive bayes algoritmasındadır. Ardından sırasıyla destek vektör makinesi, yapay sinir ağları, rastgele orman ve k-en yakın komşu gelmektedir. Yine 5'li sınıflamada en yüksek ROC altında kalan alan değerine sahip algoritma rastgele orman algoritmasıdır. Ardından ise sırasıyla naive bayes, yapay sinir ağları, k-en yakın komşu ve destek vektör makinesi algoritmaları yer almaktadır.

Tablo 23'te yer alan WEKA'nın 2'li sınıflamasına ait güvenirlik değerleri incelendiğinde, en yüksek doğru sınıflama oranına sahip olan algoritmanın rastgele orman olduğu görülmektedir. Ardından sırasıyla naive bayes, destek vektör makinesi, yapay sinir ağları ve k-en yakın komşu gelmektedir. Kappa istatistikleri incelendiğinde doğru sınıflama oranları ile elde edilen sıralamanın değişmediği görülmektedir. ROC altında kalan alan değerleri incelendiğinde en yüksek değere rastgele orman algoritmasının sahip olduğu görülmektedir. Ardından sırasıyla naive bayes, yapay sinir ağları, k-en yakın komşu ve destek vektör makinesi yer almaktadır.

**Dördüncü Alt Problem: Orange'tan Elde Edilen Değerlerin Karşılaştırılmasına İlişkin Bulgular**

Bu bölümde, Orange paket programından elde edilen güvenirlik değerleri karşılaştırılmıştır.

Orange paket programından elde edilen güvenirlik değerleri Tablo 24'te verilmiştir.

Tablo 24

*Orange'ta elde edilen güvenirlik değerleri*

| *Orange* | Doğru sınıflama oranı | | Güven aralığı %95 | | Kappa istatistiği | | ROC altında kalan | |
|---|---|---|---|---|---|---|---|---|
| | 5'li | 2'li | 5'li | 2'li | 5'li | 2'li | 5'li | 2'li |
| KYK | *.448* | *.650* | *.439 - .458* | *.642 - .660* | *.141* | *.301* | *.646* | *.713* |
| RO | *.475* | *.676* | *.466 - .484* | *.668 - .686* | *.175* | *.351* | *.685* | *.744* |
| DVM | *.474* | *.628* | *.465 - .483* | *.619 - .637* | *.185* | *.254* | *.687* | *.663* |
| NB | *.366* | *.647* | *.358 - .376* | *.639 - .656* | *.145* | *.291* | *.638* | *.702* |
| YSA | *.481* | *.684* | *.472 - .491* | *.676 - .694* | *.204* | *.367* | *.693* | *.756* |

Tablo 24 incelendiğinde 5'li sınıflama değerlerinde Orange ile elde edilmiş en yüksek doğru sınıflama oranına sahip algoritmanın yapay sinir ağları olduğu görülmektedir. Ardından, sırasıyla rastgele orman, destek vektör makinesi, k-en yakın komşu ve naive bayes algoritmaları gelmektedir. 5'li sınıflamada en yüksek kappa istatistiğine sahip algoritma yapay sinir ağları olmuştur. Ardından, sırasıyla destek vektör makinesi, rasgele orman, naive bayes ve k-en yakın komşu algoritmaları gelmektedir. ROC altında kalan alan değerleri incelendiğinde en yüksek değere sahip algoritmanın yapay sinir ağları olduğu görülmektedir. Ardından, sırasıyla destek vektör makinesi, rastgele orman, k-en yakın komşu ve naive bayes algoritmaları gelmektedir.

Tablo 24'te yer alan Orange'ın 2'li sınıflamasına ait güvenirlik değerleri incelendiğinde, en yüksek doğru sınıflama oranına sahip olan algoritmanın yapay sinir ağları olduğu görülmektedir. Ardından, sırasıyla rastgele orman, k-en yakın komşu, naive bayes ve destek vektör makinesi algoritmaları yer almaktadır. 2'li sınıflamada en yüksek kappa istatistiğine sahip algoritmanın yapay sinir ağları

olduğu görülmektedir. Ardından, sırasıyla rastgele orman, k-en yakın komşu, naive bayes ve destek vektör makinesi gelmektedir. ROC altında kalan alan değerinde en yüksek değer yapay sinir ağları algoritması ile elde edilmiştir. Ardından, sırasıyla rastgele orman, k-en yakın komşu, naive bayes ve destek vektör makinesi gelmektedir.

# Bölüm 5
# Sonuç, Tartışma ve Öneriler

Bu çalışmada, popüler veri madenciliği araçlarından ikisinin, eğitim alanındaki verilerle yaptığı sınıflamalar karşılaştırılmıştır. Bu bölümde ise; çalışma kapsamında elde edilen araştırma sonuçları, alan yazında yapılan çalışmalar da göz önünde bulundurularak, değerlendirilmiş ve ileride yapılabilecek araştırmalara yönelik önerilerde bulunulmuştur.

## Sonuçlar

Araştırmanın birinci alt problemine ilişkin elde edilen sonuçların güvenirlik değerleri incelendiğinde doğru sınıflama oranı açısından en yüksek değeri, yapay sinir ağlarının, Orange'ta aldığı görülmüştür. Ardından sırasıyla, destek vektör makinesi (WEKA) ve rastgele orman (Orange ve WEKA) gelmektedir.

Orange'ın k-en yakın komşu, rastgele orman ve yapay sinir ağları algoritmalarında WEKA'dan daha yüksek doğru sınıflama oranına sahip olduğu görülmektedir. WEKA ise destek vektör makinesi ve naive bayes algoritmalarında daha yüksek doğru sınıflama oranına sahiptir. Orange'ın, doğru sınıflama oranı açısından, üstün olduğu algoritmalarda farklılaşma sırasıyla %1,2 %0,5 ve %2,5'tir. WEKA'nın doğru sınıflama oranı açısından, üstün olduğu algoritmalarda farklılaşma ise sırasıyla %0,6 ve %6'dır.

Destek vektör makinesi dışında, doğru sınıflama oranı yüksek olan algoritmaların, kappa istatistiklerinin daha yüksek olduğu görülmektedir. Destek vektör makinesinde ise kappa istatistikleri, doğru sınıflama oranındaki WEKA üstünlüğüne rağmen, eşit çıkmıştır.

ROC altında kalan değerlerde ise Orange, naive bayes dışında, daha yüksek değerlere sahiptir. Destek vektör makinesinde, doğru sınıflama oranı olarak WEKA değeri yüksek iken, ROC altında kalan alan değerinde Orange'ın .098'lik bir üstünlüğü bulunmaktadır.

Araştırmanın ikinci alt problemine ilişkin elde edilen sonuçların güvenirlik değerleri incelendiğinde, doğru sınıflama oranı açısından en yüksek değeri, yapay sinir ağlarının, Orange'ta aldığı görülmüştür. Ardında sırasıyla, rastgele orman (WEKA, Orange) ve naive bayes (WEKA) gelmektedir.

k-en yakın komşu ve yapay sinir ağları algoritmasında Orange'ın daha yüksek doğru sınıflama oranlarına sahip olduğu görülmektedir. Rastgele orman, destek vektör makinesi ve naive bayes algoritmasında ise WEKA daha yüksek doğru sınıflama oranına sahiptir. Orange'ın, doğru sınıflama oranı açısından, üstün olduğu algoritmalarda farklılaşma sırasıyla %1,2 ve %3,5'tir. WEKA'nın doğru sınıflama oranı açısından, üstün olduğu algoritmalarda farklılaşma ise sırasıyla %0,5 %2,7 ve %1,4'tür.

Kappa istatistiği değerlerinin, doğru sınıflama oranı değerleriyle aynı şekilde, WEKA'da naive bayes, rastgele orman ve destek vektör makinesi, Orange'ta ise k-en yakın komşu ve yapay sinir ağları algoritmalarında daha yüksek değere sahip olduğu görülmektedir. Bunun yanında kappa değerleri 5'li sınıflama için genellikle zayıf, 2'li sınıflama için ise makul düzeyde uyuma işaret etmektedir.

ROC altında kalan alan değerinde destek vektör makinesi, daha düşük doğru sınıflama oranına sahip olmasına rağmen, Orange lehine yüksektir. Destek vektör makinesi dışındaki ROC altında kalan alan değerleri ise, doğru sınıflama oranıyla aynı şekilde, k-en yakın komşu ve yapay sinir ağları algoritmalarında Orange'ta, rastgele orman ve naive bayes algoritmalarında ise WEKA'da yüksektir.

Üçüncü ve dördüncü alt probleme ilişkin sonuçlar incelendiğinde (Tablo 23 ve Tablo 24) ise 5'li sınıflama yapılan durumda WEKA'da en yüksek doğru sınıflama oranına sahip algoritmanın destek vektör makinesi, Orange'ta ise yapay sinir ağları olduğu görülmektedir. 2'li sınıflama yapılan durumda ise en yüksek doğru sınıflama oranına WEKA'da rastgele orman, Orange'ta ise yapay sinir ağları algoritmasının sahip olduğu görülmektedir.

Tüm alt problemlerde; doğru sınıflama oranlarının güven aralıkları göz önüne alındığında yaklaşık yarısında kesişme olduğu görülmektedir. Ancak farklılaşmaların istatistiksel olarak anlamlı olmadığı sonuçlardaki farklar da göz önünde bulundurulmalıdır. Öneriler bu farklar gözetilerek yazılmıştır.

## Tartışma

Yapay sinir ağları algoritmasının, ikili ve beşli sınıflandırmada, en yüksek doğru sınıflama oranına sahip olduğu görülmektedir. Bu durumun literatürde de örnekleri vardır (Hussain, Atallah, Kamsin, ve Hazarika, 2018; Şengür, 2013).

Alanyazın ve bu çalışma birlikte ele alındığında, daha tutarlı sınıflamalar için, yapay sinir ağları algoritmasının kullanılması uygun görünmektedir.

Naik ve Samant'ın (2016) Hindistan Karaciğer Hastaları Veri Seti'nde yer alan 583 veriyle yaptığı çalışmada, naive bayes algoritmasında Orange'ın WEKA'ya %13,2'lük üstünlüğü varken, bu çalışmada WEKA'nın beşli sınıflamada %6, ikili sınıflamada %1,4'lük üstünlüğü vardır. Yine aynı çalışmada, k-en yakın komşu algoritmasında WEKA'nın Orange'a %34,5'lik üstünlüğü varken, bu çalışmada Orange'ın ikili ve beşli sınıflamada %1,2'lik üstünlüğü vardır. Bunun nedeni çalışmalarda kullanılan örneklem büyüklükleri ve tipleri arasında önemli fark bulunması olabilir. Alanyazın ve bu çalışma ele alındığında, eğitimsel verilerde daha tutarlı sınıflamalar için Orange'ın kullanılması uygun görünmektedir. Ancak eğitimsel verilerle daha fazla çalışma yapılması bu sonucu pekiştirecektir.

Bu çalışmada, beşli sınıflamada Orange'ın k-en yakın komşu, rastgele orman ve yapay sinir ağları algoritmalarında WEKA'dan daha yüksek doğru sınıflama oranına sahip olduğu görülmektedir. WEKA ise destek vektör makinesi ve naive bayes algoritmalarında daha yüksek doğru sınıflama oranına sahip olduğu görülmektedir. Ancak doğru sınıflama oranlarının güven aralıkları incelendiğinde 5'li sınıflamada WEKA ve Orange karşılaştırmasında naive bayes ve yapay sinir ağları algoritmalarındaki değişimlerin istatistiksel olarak anlamlı olduğu söylenebilir. İkili sınıflamada ise, k-en yakın komşu ve yapay sinir ağları algoritmasında Orange'ın daha yüksek doğru sınıflama oranlarına sahip olduğu görülmektedir. Rastgele orman, destek vektör makinesi ve naive bayes algoritmasında ise WEKA daha yüksek doğru sınıflama oranına sahiptir. Ancak yine doğru sınıflama oranlarının güven aralıkları incelendiğinde 2'li sınıflamada WEKA ve Orange karşılaştırmasında destek vektör makinesi ve yapay sinir ağları algoritmalarındaki değişimlerin istatistiksel olarak anlamlı olduğu söylenebilir.

Çalışmadan elde edilen kappa istatistikleri ve eğri altında kalan alan değerleri göz önünde bulundurulduğunda doğru sınıflamanın yeterli düzeyde olmadığı görülmektedir. Bu durumun sebebi olarak, özellikle 5'li sınıflamada hücrelere düşen verilerin azlığı gösterilebilir. 2'li sınıflamada değerler 5'li sınıflamaya göre karışıklık matrisine daha dengeli dağılmaktadır.

Özetle; k-en yakın komşu ve yapay sinir ağları algoritmalarında Orange, destek vektör makinesi ve naive bayes algoritmasında ise WEKA'nın daha doğru sınıflama yaptığı görülmektedir. Rastgele orman algoritması ise beşli ve ikili sınıflamada Orange'tan WEKA'ya geçmiştir. Araştırmanın alt problemlerinde yer aldığı şekliyle ifade etmek gerekirse; ABİDE sınavında Türkçe alt testi puanları beşli/ikili sınıflandırıldığında, demografik ve psiko-sosyal değişkenlerden yararlanarak; WEKA ve Orange programlarında k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları algoritmaları ile elde edilen doğru sınıflama sayısı, doğru sınıflama oranı, kappa istatistiği ve ROC eğrisi altında kalan alan değerler farklılaşmaktadır.

Veri madenciliğinde doğru algoritmanın seçimi, yani en doğru sınıflamayı yapacak yöntemin bulunması, verinin niteliklerine bağlı olarak değişim göstermektedir (Aksu, 2018). Her algoritma ve her program, kendi çözümlerini ve sorunlarını ortaya çıkarmaktadır. Bu çalışmada, daha sonraki araştırmalarda kullanılabilecek program ve algoritmayı belirlemeye yardımcı olması amacıyla karşılaştırmalar yapılmıştır.

## Öneriler

Bu kısımda yer alan öneriler, olası uygulayıcılar ve daha sonra bu konuyu çalışacak araştırmacılar için iki ayrı başlıkta düzenlenmiştir.

### Uygulayıcılara Yönelik Öneriler

Elde edilen sonuçlar incelendiğinde, Türkçe olarak elde edilmiş eğitsel verilerde yapılacak bir sınıflama işlemi için, yapay sinir ağları algoritmasının kullanılmasının daha yüksek doğru sınıflama oranı sağlaması mümkün görünmektedir. Ayrıca, ikili sınıflandırılmış puanların tahmininin daha yüksek doğru sınıflama oranı sağladığı görülmüştür.

- K-en yakın komşu ve yapay sinir ağları algoritmalarının kullanımına karar verilmiş bir uygulamada, Orange paket programının,
- Destek vektör makinesi ve naive bayes algoritmalarının kullanımına karar verilmiş bir uygulamada, WEKA paket programının,

- Rastgele orman algoritmasının kullanımına karar verilmiş bir uygulamada ise ikili ve beşli sınıflandırma için farklı sonuçlar elde edildiği için beşli sınıflandırmada WEKA, ikili sınıflandırmada Orange paket programının, kullanılması önerilmektedir.

Bulunan doğru sınıflama oranı güven aralıklarının da göz önünde bulunması önerilmektedir. Bununla birlikte, veri madenciliği kullanarak yapılması planlanan bir eğitimsel sınıflama için birçok çalışmanın birleştirilmesi ve değerlendirilmesi gerekecektir. Yalnızca bu çalışmadan elde edilen sonuçlara göre sınıflama yöntemi ve paket program seçmek yerine, bu çalışma ile birlikte gelecekte yapılabilecek araştırmaların bir bütün olarak ele alınması daha doğru olacaktır.

**Daha Sonraki Araştırmacılar İçin Öneriler**

- Çalışmanın sonuçlarının daha anlamlı ve güvenilir olabilmesi için, farklı eğitim verileriyle tekrar edilmesi yararlı olacaktır.

Veri madenciliğinde, kullanılan her veri çeşidinin, sınıflamanın yapılacağı algoritmaya etkisi farklı olacaktır. Bu açıdan, çalışmanın, Türkiye'de elde edilmiş bir eğitim verisiyle tekrar edilmesi daha anlamlı sonuçlara ulaşılmasını sağlamak için önemlidir.

- Yine çalışmanın geliştirilmesi için, farklı popüler veri madenciliği programlarının karşılaştırmada kullanılması da yararlı olacaktır.

Böylelikle, daha sonra Türkiye'de hayata geçebilecek, öğrenci sınıflayan bir veri madenciliği projesinde kullanılabilecek algoritma ve programların belirlenmesi kolaylaşacaktır. Bu çalışmanın sınırlılığı olan, WEKA ve Orange paket programlarının kullanılması, farklı paket programlar eklenerek bu sınırlılık azaltılmalıdır (Rapidminer, R, Knime vb.).

- Çalışmada yer alan k-en yakın komşu, rastgele orman, destek vektör makinesi, naive bayes ve yapay sinir ağları algoritmalarının sayısı artırılmalı ve algoritmalar çeşitlendirilmelidir.

Tüm bunların yanında, ABİDE projesi ile elde edilmiş veriler gibi, Türkiye'den elde edilmiş verilerin kullanılması, Türkiye içerisinde yapılacak bir "öğrenci sınıflama projesi" için yararlı olacaktır.

## Kaynaklar


Akman, M., Genç, Y., & Ankaralı, H. (2011). Random forests yöntemi ve sağlık alanında bir uygulama. *Türkiye Klinikleri Journal of Biostatistic, 3*(1), 36-48.

Akpınar, H. (2014). *Data veri madenciliği - veri analizi.* İstanbul: Papatya Yayıncılık.

Aksu, G. (2018). PISA başarısını tahmin etmede kullanılan veri madenciliği yöntemlerinin incelenmesi. *Doktora Tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü*.

Aksu, G. (2018). PISA başarısını tahmin etmede kullanılan veri madenciliği yöntemlerinin incelenmesi. *(Doktora tezi), Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü*.

Aksu, G., & Doğan, N. (2018). Veri madenciliğinde kullanılan öğrenme yöntemlerinin farklı koşullar altında karşılaştırılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 51*(3), s. 71-100.

Aksu, G., & Güzeller, C. O. (2016). PISA 2012 Matematik okuryazarlığı puanlarının karar ağacı yöntemiyle sınıflandırılması: Türkiye örneklemi. *Eğitim ve Bilim, 41*(185), 101-122.

Altunkaynak, B. (2017). *Veri madenciliği yöntemleri ve R uygulamaları.* Ankara: Seçkin Yayıncılık.

Amershi, S., & Conati, C. (2006). Automatic Recognition of Learner Groups in Exploratory Learning Environments. *Intelligent Tutoring Systems (ITS).* Berlin: Springer.

Aydın, S. (2007). Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama. *(Doktora tezi), Anadolu Üniversitesi Sosyal Bilimler Enstitüsü*.

Biggs, D., De Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics, 18*(1), 49-62.

Bramer, M. (2007). *Principles of data mining.* Londra: Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Brown, M. (2014). *Data mining for dummies.* New Jersey: Wiley & Sons.

Coşkun, C., & Baykal, A. (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması. (s. 1-8). Malatya: Akademik Bilişim.

Crawford, S. L. (1989). Extensions to the CART algorithm. *International Journal of Man-Machine Studies, 31*(2), 197-217.

Devasia, T., Vinushree, T., & Hedge, V. (2016). Prediction of students performance using educational data mining. *International Conference on Data Mining and Advanced Computing (SAPIENCE)* (s. 91-95). Ernakulam: SAPIENCE.

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4.5 and a Naivebayes classifier for the prediction of lung cancer survivability. *Journal of Computing, 4*(8).

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*(4), 325-327.

Eroğlu, M. G., & Kelecioğlu, H. (2015). Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 28*(1), 31-52.

Fayyad, U. M., & Irani, K. B. (1992). On the handling of continuous-valued attributes in decision tree generation. *Machine Learning, 8*(1), 87-102.

Fayyad, U., & Stolorz, P. (1997, 11). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems, 13*(2-3), s. 99-115.

Fleiss, J., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychologlcal Bulletin, 72*(5), 323-327.

Friedl, M., & Brodley, C. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment, 61*(3), s. 399-409.

Hand, D. J. (2007). Principles of Data Mining. *Drug Safety*, *30*, s. 621-622. doi:10.2165/00002018-200730070-00010

Hark, C. (2013). Öğrencilerin akıllı tahtaya ilişkin tutumlarının incelenmesine yönelik bir veri madenciliği uygulaması. *(Yüksek lisans tezi), Fırat Üniversitesi Eğitim Bilimleri Enstitüsü*.

Haykin, S. (2004). *Neural networks: A comprehensive foundation.* Upper Saddle River, N.J: Prentice Hall.

Holsheimer, M., & Siebes, A. (1994). *Data mining: The search for knowledge in databases.* Amsterdam: Centrum voor Wiskunde en Informatica.

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res., 31 (10)*, 2517– 2530. doi:10.1029/95WR01955

Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018). *Classification, clustering and association rule mining in educational datasets using data mining tools: A case study.* Cybernetics and Algorithms in Intelligent Systems. CSOC2018 2018. Advances in Intelligent Systems and Computing. doi:10.1007/978-3-319-91192-2_21

Irmak, S. (2009). Veri madenciliği yöntemleri ile sağlık sektörü veri tabanlarında bilgi keşfi: Tanımlayıcı ve kestirimci model uygulamaları. *(Doktora tezi), Akdeniz Üniversitesi Sosyal Bilimler Enstitüsü*.

Kass, G. V. (1980). An exploratory technique for ınvestigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 29*(2), s. 119-127.

Korkem, E. (2013). Mikroarray gen ekspresyon veri setlerinde random forest ve naıve bayes sınıflama yöntemleri yaklaşımı. *(Yüksek lisans tezi), Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü*.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine, 4*(2), 4-22.

MEB. (2019). *ABİDE 2018, 8. sınıflar özet rapor.* Ankara: MEB.

Merceron, A., & Yacef, K. (2005). Educational data mining: a case study. *AIED* (s. 467-474). Amsterdam: IOS Press.

Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science, 85*, 662-668.

North, M. A. (2012). *Data Mining for the Masses.* Atina: Global Text Project Book.

Patil, T. R., & Sherekar, S. S. (2013, Nisan). Performance analysis of naive bayes and J48 classification algorithm for data classification. *International Journal Of Computer Science And Applications, 6*(2), s. 256-261.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106.

Roe, B. P., Yang, H.-J., Z. J., Liu, Y., Stancu, I., & McGregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 543*(2-3), 577–584.

Roiger, R. (2017). *Data mining: A Tutorial-Based Primer (second edition).* CRC Press.

Romero, C., Ventura, S., & Bra, P. D. (2014). Knowledge discovery with genetic programming for providing feedback to courseware authors. *User Modeling and User-Adapted Interaction, 14*(5), 425-464.

Şeker, S. E. (2013). *WEKA ile veri madenciliği.* İstanbul: Bilgisayar Kavramları Yayınları.

Şengül, A. (2011). Türk öğrencilerin PISA 2009 okuma becerilerini açıklayan değişkenlerin CHAID analizi ile belirlenmesi. *(Yüksek lisans tezi), Ankara Üniversitesi Eğitim Bilimleri Enstitüsü*.

Şengür, D. (2013). Öğrencilerin akademik başarılarının veri madenciliği metotları ile tahmini. *(Yüksek lisans tezi), Fırat Üniversitesi Eğitim Bilimleri Enstitüsü*.

Taşdemir, M. (2012). Veri madenciliği (Öğrenci başarısına etki eden faktörmlerin regresyon analizi ile tespiti). *(Yüksek lisans tezi), Dicle Üniversitesi Sosyal Bilimler Enstitüsü*.

Two Crows Corp. (1999). *Introduction to data mining and knowledge discovery (Third edition).* Maryland: Two Crows.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med, 37*(5), 360-363.

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures. *IEEE Signal Processing Magazine, 26*(1), 98–117. doi:10.1109/msp.2008.930649

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. I. (2005). *Data mining: Practical machine learning tools and techniques.* Cambridge: Elsevier Inc.

Wu, X., & Kumar, V. (2009). *The top ten algorithms in data mining.* Boca Raton: CRC Press.

## EK-A: Etik Komisyonu Onay Bildirimi

Tarih: 24/10/2019
Sayı: 35853172-300-E.00000828656

T.C.
HACETTEPE ÜNİVERSİTESİ
Rektörlük

Sayı : 35853172-300
Konu : Semih TOPUZ (Etik Komisyon İzni)

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

İlgi : 09.10.2019 tarihli ve 51944218-300/00000810503 sayılı yazı.

Enstitünüz Eğitim Bilimleri Anabilim Eğitimde Ölçme ve Değerlendirme Bilim Dalı yüksek lisans programı öğrencilerinden **Semih TOPUZ**'un **Prof. Dr. Nuri DOĞAN** danışmanlığında yürüttüğü "**Eğitimsel Verilerde WEKA ve Orange Veri Madenciliği Yazılımından Elde Edilen Analiz Sonuçlarının Karşılaştırılması**" başlıklı tez çalışması Üniversitemiz Senatosu Etik Komisyonunun **15 Ekim 2019** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

Bilgilerinizi ve gereğini saygılarımla rica ederim.

e-imzalıdır
Prof. Dr. Rahime Meral NOHUTCU
Rektör Yardımcısı

Evrakın elektronik imzalı suretine https://belgedogrulama.hacettepe.edu.tr adresinden 525e7ef7-08ce-4803-9c06-7a450d76ab4d kodu ile erişebilirsiniz.
Bu belge 5070 sayılı Elektronik İmza Kanunu'na uygun olarak Güvenli Elektronik İmza ile imzalanmıştır.

Hacettepe Üniversitesi Rektörlük 06100 Sıhhiye-Ankara
Telefon:0 (312) 305 3001-3002 Faks:0 (312) 311 9992 E-posta:yazimd@hacettepe.edu.tr İnternet Adresi: www.hacettepe.edu.tr

Sevda TOPAT

## EK-B: Etik Beyanı

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı bütün bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin bütününü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

…...../……/…….

(İmza)
Semih Topuz

## EK-C: Yüksek Lisans/Doktora Tez Çalışması Orijinallik Raporu

……/……./.....

HACETTEPE ÜNİVERSİTESİ
Eğitim Bilimleri Enstitüsü
Eğitim Bilimleri Ana Bilim Dalı Başkanlığına,

Tez Başlığı: Eğitsel Verilerde WEKA ve Orange Veri Madenciliği Yazılımlarından Elde Edilen Analiz Sonuçlarının Karşılaştırılması

Yukarıda başlığı verilen tez çalışmamın tamamı (kapak sayfası, özetler, ana bölümler, kaynakça) aşağıdaki filtreler kullanılarak **Turnitin** adlı intihal programı aracılığı ile kontrol edilmiştir. Kontrol sonucunda aşağıdaki veriler elde edilmiştir:

| Rapor Tarihi | Sayfa Sayısı | Karakter Sayısı | Savunma Tarihi | Benzerlik Oranı | Gönderim Numarası |
|---|---|---|---|---|---|
| 31/05/2021 | 57 | 11278 | 28/06/2021 | %10 | 1597825966 |

Uygulanan filtreler:

1. Kaynaklar hariç
2. Alıntılar dâhil
3. 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan eder, gereğini saygılarımla arz ederim.

**Ad Soyadı:** Semih Topuz
**Öğrenci No.:** N18133131
**Ana Bilim Dalı:** Eğitim Bilimleri
**Programı:** Eğitimde Ölçme ve Değerlendirme
**Statüsü:** ☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

İmza

## DANIŞMAN ONAYI

UYGUNDUR.
(Prof. Dr. Nuri Doğan, İmza)

## EK-Ç: Thesis/Dissertation Originality Report

……/……./……

HACETTEPE UNIVERSITY
Graduate School of Educational Sciences
To The Department of Educational Sciences

Thesis Title: Comparison of Analysis Results Obtained from WEKA and Orange Data Mining Software in Educational Data Mining

The whole thesis that includes the *title page, introduction, main chapters, conclusions and bibliography section* is checked by using **Turnitin** plagiarism detection software take into the consideration requested filtering options. According to the originality report obtained data are as below.

| Time Submitted | Page Count | Character Count | Date of Thesis Defense | Similarity Index | Submission ID |
|---|---|---|---|---|---|
| 31/05/2021 | 57 | 11278 | 28/06/2021 | 10% | 1597825966 |

Filtering options applied:
1. Bibliography excluded
2. Quotes included
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Educational Sciences Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

**Name Lastname:** Semih Topuz
**Student No.:** N18133131
**Department:** Educational Sciences
**Program:** Educational Measurement and Evaluation
**Status:** ☒ Masters ☐ Ph.D. ☐ Integrated Ph.D.

Signature

## ADVISOR APPROVAL

APPROVED
(Prof. Dr. Nuri Doğan, Signature)

## EK-D: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kâğıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım haklan bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan **"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına ilişkin Yönerge"** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- o Enstitü/Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir. (1)
- o Enstitü/Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren … ay ertelenmiştir. (2)
- o Tezimle ilgili gizlilik kararı verilmiştir. (3)

..... /..... /.....

(imza)

Semih Topuz

---

*"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

*(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent* alma *sürecinin devam etmesi durumunda, tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü Üzerine enstitü veya fakülte yönetim kurulu iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

*(2) Madde 6.* 2. *Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç; imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez danışmanın önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulunun gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

*(3) Madde* 7. *1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir*. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
*Madde* 7.2. *Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir*

** Tez danışmanının önerisi ve enstitü anabilim dalının uygun görüşü üzerine enstitü veya fakülte yönetim kurulu tarafından karar verilir.*