**T.C.**
**BOLU ABANT İZZET BAYSAL ÜNİVERSİTESİ**
**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**


# MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖĞRENCİ MEZUNİYET NOTU TAHMİNİ

**YÜKSEK LİSANS TEZİ**

**SARP CİVELEK**

**TEZ DANIŞMANI**
**Doç. Dr. Murat BEKEN**

**BOLU, HAZİRAN - 2023**

# KABUL VE ONAY SAYFASI

**Sarp CİVELEK** tarafından hazırlanan "**MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖĞRENCİ MEZUNİYET NOTU TAHMİNİ**" adlı tez çalışması jürimiz tarafından Bilgisayar Mühendisliği Anabilim Dalı'nda Yüksek Lisans Tezi olarak oy birliği ile kabul edilmiştir. 13/06/2023

| **Jüri Üyeleri** | **İmza** |
|---|---|
| Danışman<br>Doç. Dr. Murat BEKEN<br>Bolu Abant İzzet Baysal Üniversitesi | ..... |
| Üye<br>Doç. Dr. Önder EYECİOĞLU<br>Bolu Abant İzzet Baysal Üniversitesi | ..... |
| Üye<br>Doç. Dr. Atınç YILMAZ<br>Beykent Üniversitesi | ..... |

**Lisansüstü Eğitim Enstitüsü Onayı**

**Prof. Dr. İbrahim KÜRTÜL**
**Lisansüstü Eğitim Enstitüsü Müdürü**

# ETİK BEYAN

Bolu Abant İzzet Baysal Üniversitesi, Lisansüstü Eğitim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu bildirir,

aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Teze ilişkin 13/06/2023 tarihinde Turnitin adlı intihal tespit programından enstitü müdürlüğünce belirlenen filtrelemeler uygulanarak alınmış olan benzerlik raporuna göre, tezin benzerlik oranı % 18 olarak tespit edilmiştir.

……..

**SARP CİVELEK**

# ÖZET


**MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖĞRENCİ MEZUNİYET NOTU TAHMİNİ**
**YÜKSEK LİSANS TEZİ**
**SARP CİVELEK**
**BOLU ABANT İZZET BAYSAL ÜNİVERSİTESİ**
**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**
**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**(TEZ DANIŞMANI: DOÇ. DR. MURAT BEKEN)**

**BOLU, HAZİRAN - 2023**
**XII + 51**

Kaliteli bir eğitim için Yükseköğretim kurumlarının yönetim ve eğitim alanlarında doğru ve güvenilir kararlar vermeleri önceliklidir. Yükseköğretim kurumlarının genel olarak yaşadığı problemlere örnek olarak hazırlanan akademik planlamada oluşabilecek eksiklikler veya yanlışlıklar, akademik olarak başarısız öğrenciler, mezun olacak öğrencilerin gelecekle ilgili yol haritaları gösterilebilir. Eğitim kalitesi açısından bu tarz problemlerin çözümü, tedbirlerin alınması çok önemlidir. Veri madenciliği ile yapay zeka yöntemlerinin gelişmesi sayesinde, yaşanan problemler üzerinde oransal olarak çok yüksek tahminler yapılabilmekte ve sonucunda çözüm odaklı anlamlı sonuçlar alınabilmektedir. Yüksek hızlı bilgisayarların hayatımızda daha fazla yer alması, geliştirilen algoritmalar ile Yapay Zeka teknikleri hızlı bir şekilde ilerlerken, hemen hemen her sektörde olduğu gibi eğitim alanında önemli gelişmelere yol açacak, akademik anlamda alınabilecek tedbirler için güçlü bir araç olma yolundadır. Tez çalışmasında, makine öğrenmesi yöntemlerinden Yapay Sinir Ağları, K-En Yakın Komşu Algoritması, Lineer Regresyon, Destek Vektör Makineleri ve Karar Ağaçları kullanılarak Bolu Abant İzzet Baysal Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Kamu Yönetimi Bölümü öğrencilerinin mezuniyet notlarının tahmin edilmesi işlemi yapılmıştır. Bu yüzden 2011-2018 yılları arasında kayıt yaptırıp mezun olmuş 832 Kamu Yönetimi Bölümü öğrencisinin akademik eğitimleri boyunca 1. ve 2. sınıfta almış oldukları toplam 31 dersin yılsonu notları kullanılmıştır. Yapılan çalışmada mezuniyet notunun tahmini için iki farklı senaryo oluşturulmuştur. İlk oluşturulan senaryoda öğrencilerin sadece birinci sınıfa ait derslerinin yılsonu notları ile mezuniyet not tahmin işlemi yapılırken, ikinci senaryoda öğrencilerin birinci ve ikinci sınıfa ait derslerinin yılsonu notları kullanılmıştır. Yapılan çalışmada Yapay Sinir Ağları ile oluşturulan modelin diğerlerine göre daha yüksek oranda başarılı tahminler yaptığı ve ikinci olarak oluşturulan senaryonun ilk senaryoya göre daha iyi tahmin sonuçları verdiği görülmüştür.



**ANAHTAR KELİMELER:** Yapay Zekâ, Makine Öğrenmesi, Yapay Sinir Ağları, Destek Vektör Makineleri, Lineer Regresyon, K-NN Algoritması, Karar Ağaçları

# ABSTRACT

**PREDICTION OF STUDENT GRADUATION GRADE WITH MACHINE LEARNING METHODS**
**MSC THESIS**
**SARP CİVELEK**
**BOLU ABANT IZZET BAYSAL UNIVERSITY**
**INSTITUTE OF GRADUATE STUDIES**
**DEPARTMENT OF COMPUTER ENGINEERING**
**(SUPERVISOR: ASSOC. DOC. MURAT BEKEN)**

**BOLU, JUNE 2023**
**XII + 51**

For a quality education, it is a priority for higher education institutions to make correct and reliable decisions in the fields of management and education. As an example of the problems faced by higher education institutions in general, deficiencies or mistakes that may occur in academic planning, academically unsuccessful students, road maps for the future of students who will graduate can be shown. In terms of the quality of education, it is very important to solve such problems and take precautions. Thanks to the development of data mining and artificial intelligence methods, very high estimates can be made proportionally on the problems experienced, and as a result, meaningful solution-oriented results can be obtained. The fact that high-speed computers take more place in our lives, while the developed algorithms and Artificial Intelligence techniques are progressing rapidly, it is on the way to become a powerful tool for measures that can be taken academically, which will lead to important developments in the field of education, as in almost every sector. In the thesis study, the graduation grades of Bolu Abant İzzet Baysal University, Faculty of Economics and Administrative Sciences, Department of Public Administration students were estimated using machine learning methods Artificial Neural Networks, K-Nearest Neighbor Algorithm, Linear Regression, Support Vector Machines and Decision Trees. For this reason, the year-end grades of a total of 31 courses taken in the 1st and 2nd grades of 832 Public Administration Department students who registered and graduated between 2011-2018 were used. In the study, two different scenarios were created for the estimation of the graduation grade. In the first scenario, only the year-end grades of the first-year courses of the students and the graduation grade estimation were made, while in the second scenario, the year-end grades of the first and second-year courses of the students were used. In the study, it was seen that the model created with Artificial Neural Networks made more successful predictions than the others and the second scenario gave better prediction results than the first scenario.

**KEYWORDS:** Artificial Intelligence, Machine Learning, Artificial Neural Networks, Support Vector Machines, Linear Regression, K-NN Algorithm, Decision Trees

# İÇİNDEKİLER

**Sayfa**

# ŞEKİL LİSTESİ

**Sayfa**

# TABLO LİSTESİ

**Sayfa**

# KISALTMA VE SEMBOLLER LİSTESİ

**DVM** **:** Destek Vektör Makineleri

**KA** **:** Karar Ağaçları

**K-NN** **:** K-En Yakın Komşu Algoritması

**LR** **:** Lineer Regresyon

**MAPE** **:** Ortalama Mutlak Yüzde Hatası

**MAE** **:** Ortalama Mutlak Hata

**MSE** **:** Ortalama Kare Hatası

**RMSE** **:** Ortalama Karekök Sapması

**YSA** **:** Yapay Sinir Ağları

**YZ** **:** Yapay Zekâ

# TEŞEKKÜR

Bu tez çalışması boyunca, ilgi ve yardımlarını esirgemeyen danışmanım Doç. Dr. Murat BEKEN'e, çalışmam boyunca desteklerini esirgemeyen Doç. Dr. Önder EYECİOĞLU ve Öğr. Gör. Yunus ÖZDEMİR'e, çalışmam boyunca beni özveri ve sabırla destekleyen eşim Aylin CİVELEK'e, insan ve bilime büyük önem veren rahmetli dayım Süleyman CİVELEK'e teşekkürlerimi ve şükranlarımı sunarım.

# 1. GİRİŞ

Veri madenciliği, geniş yer kaplayan veri yığınları ile çok büyük miktardaki verilerin faydalı bilgilere evrilmesine duyulan ihtiyaca binayen meydana gelmiştir (Han ve Kamber, 2006).

Veri madenciliği tıp, sağlık, mühendislik, finans, ekonomi, eğitim, bankacılık gibi alanlarda karar verme noktasında destek oluşturması, pazar stratejisi, yüksek oranlı tahminler gibi farklı birçok alanda kullanılabilir olması sebebiyle, son yıllarda, veritabanı kullanıcıları ve araştırmacıların büyük ölçüde dikkatini çekmektedir. Veri madenciliği; makine öğrenme, veri tabanları, istatistik gibi farklı alanlardaki teknikleri ortak noktada birleştirebilen bir yapıya sahip olması nedeni ile veri tabanlarında yer alan ham veriden yararlı ve değerli bilgiyi ortaya çıkarmamıza olanak sağlamaktadır (Ching, 2003).

Veri madenciliği için geliştirilen Yapay Zeka (YZ) yöntemlerinin alt dallarından olan makine öğrenmesi yöntemleri günümüzde sıklıkla kullanılır olmuştur. Yapay öğrenme olarak da adlandırılan makine öğrenmesi ile kompleks ve büyük veri kümelerinden bilgisayar yardımı ile bilgi elde edilmesini sağlayan, bünyesinde farklı istatiksel ve matematiksel tekniklere sahip ve hızlı bir şekilde çıkarımları tespit edebilen yapay zeka uzantısıdır. Geliştirlen algoritmalar sayesinde makine öğrenmesi bu çıkarımları yapmaktadır. Mevcut durum hakkındaki tutarlı ve anlamlı bilgi vermek ve gelceğe önelik bilinmeyene dair tahminde bulunmak bu algoritmalar sayesinde yapılabilmektedir. Dijital veri miktaraındaki artışla beraber bu verileri analiz edebilecek insan sayısındaki artışın yetersiz olması, analiz işlemi için veri madenciliği tekniklerinin kullanılmasına yönelmemizi sağlamıştır (Savaş, Topaloğlu ve Yılmaz, 2012).

Genel olarak iş, finans, sağlık, mühendislik alanlarında yaydın olarak kulanılan veri madenciliği teknikleri eğitim alanında da çeşitli uygulama alanlarına sahip olmaya başlamıştır. Eğitim alanında; bireylerin öğrenmesini keşfetmek, öğrenmeyi ve onu etkileyen faktörleri tahmin etmek, öğrenme metotları üzerine yorum yapmak gibi farklı alanlarda kullanılabilir veri madenciliği. Bu sayede eğitim sistmeinde kalitenin artırılması noktasında büyük faydalar sağlanabilir. Akılcı ve iyileştirilmiş öğrenme teknolojisi oluşturulma

yolunda fayda sağlayabilecek veri madenciliği ile hem eğitimciler hem de öğrenenler daha iyi bilgilendirilmek için kullanılabilir (Baker, 2014). Başka bir deyişle eğiticinin, öğretme ortamını tasarım ve geliştirme aşamasında vereceği kararlar için bir temel oluşturmasına yardımcı, yararlı bilgiler verebilir. Yapay zekâ teknolojisi olan makine öğrenmesi, yapay sinir ağları, derin öğrenme, uzman sistemler, genetik algoritmalar, bulanık mantık, robotik süreç otomasyonları gibi yapay zeka türleri ile daha etkili kararlar alınmasına olanak sağlamaktadır (İşler ve Kılıç, 2021). Makine öğrenmesi teknikleri ile insanlar tarafından uzun sürede yapabilecekleri hesaplamalar bilgisayarlar tarafından ise çok kısa sürede çok kolay şekilde yapılabilir.

Yükseköğretim kurumlarını oluşturan öğrenciler, akademik ve idari personeller, dersler, öğretim müfredatı, eğitim alanı, akademik takvim vb bileşenler ve veriler stratejik açıdan önemli verilerdir. Bu verilerin işlenerek faydalı ve anlamlı bilgilerin ortaya çıkarılması, kurumların bu doğrultuda grekli birtakım tedbirleri alarak eğitimdeki kaliteyi artırmasına olanak sağlayacaktır. Ham verinin işlenip anlamlı bilgileri ortaya çıkarmada her zaman istatistiki yöntemler işe yaramamakla birlikte makine öğrenmesi yöntemleri giderek artan şekilde kullanılmaktadır. Yükseköğretim kurumları da bünyesindeki aktif ve mezun öğrencilerinin gelecekle ilgili yol haritasını tahmin etme yönünde çalışmalar yapmaktadırlar (Davis vd., 2007). Kurumlar bünyesindeki öğrencilerden hangisinin mezun olmak için akademik tedbirler ve yönlendirme ihtiyacı duyabilecek durumda olduğunun, hangi öğrencilerin de okuldan ayrılma noktasında olduğu sorularına cevap bulma noktasında veri madenciliğinden faydalanmaktadırlar.

Ülkemizde genç nüfusun son yıllarda giderek artması ve buna paralel şekilde artan Yükseköğretim okul sayısı, öğrencilerin mezuniyet sonrası gelecek planlaması yapması noktasında daha fazla zorluk ve engel yaşamasına sebep olmaktadır. Akademisyen ihtiyacı, eğitim alanındaki veri büyüklüğü ve karmaşıklığı da gün geçtikçe artmakta, eğitmenlerin öğrencilerle ilgilenme süresi kısalmaktadır. Bu yüzden eğitim alanında keşfedilen bilgi yalnızca eğitimciler için değil aynı zamanda karşılıklı etkileşim içinde bulunulan öğrenciler tarafından kullanıbilir.

Veri tabanlarında öğrenciler hakkında yer alan büyük veri yığınları, öğrencilerin akademik başarılını yükseltmek için kullanılması, gelecek dönemlerde de faydalanılacak yararlı bilgilerin ortaya çıkarılmasında kullanılmaıdır. Eğitimden sorumlu yöneticiler bütçe, öğrenci başvurusu, ders kayıt ve yönetimi, tesis ve giderler gibi çeşitli alanlarda bu yöntemler sayesinde faydalı çıkarımlar yapabileceği gibi kişileştirilmiş öğretim imkanı, akadmeik başarının arttırılması, işbirlikçi öğrenme için akıllı destek sağlanması, öğreticiler için zaman kaybının önlenmesi, süreğen şekilde analiz ve geribildirim sağlanması, eğiticilerin vermiş oldukları dersleri tekrar düzenlemesine yardımcı olması, öğrencilerin akademik yetenek ve öğrenme seviyelerine uygun kişiselleştirilmiş ödev ve proje yapabilmesinin tespiti, derse kayıt yaptıracak öğrenci portföyünün tahmin edilmesi, çevrimiçi eğitim kaynaklarının geliştirilmesi, okulu bırakma riski olan öğrencilerin tahmini, genel olarak öğrenci performansının tahmin edilmesi gibi bilgilerin ve çıkarımların ortaya çıkarılmasına yönelik kullanılabilir.

Yükseköğretim okullarındaki öğrencilerin akademeik başarısı, kendisinin yanı sıra eğitmen, okul ve ailesi tarafından da önemsenmektedir. Yapılan araştırmalar sonucunda mezuniyet ortalamasına göre işsizlik sürelerinde ters orantı olduğu ve lisans eğitimi sonrası akademik olarak yoluna devam etmek isteyenler için mezuniyet notunun önemi gerçeği düşünüldüğünde, öğrenciler için akademik başarı tahmini analizi işe yarayabilir. Başarısız denilebilecek öğrenciler için gerekli rehberlik ve yönlendirme çalışmaları yapılabilir, eksik alanlarda destek oluşturabilecek kaynakların temin edilmesi sağlanabilir, ekstra ders programları düzenlenebilir. Bu doğrultuda "Makine Öğrenmesi Yöntemleri ile Öğrenci Performans Tahmini" araştırma konusu seçilmiştir.

### 1.1 Araştırmanın Amacı

Bu tezin amacı, farklı makine öğrenmesi yöntemleri ile Bolu Abant İzzet Baysal Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Kamu Yönetimi Bölümü öğrencilerinin mezuniyet notlarının tahmin edilmesini sağlamaktır. Bu yüzden 2011-2018 yılları arasında Kamu Yönetimi Bölümüne kayıt yaptırmış ve mezun olmuş 832 öğrencinin 4 yıllık eğitim sürecinde almış oldukları toplam 31 dersin yılsonu notları kullanışmıştır. Bu notlar Bolu Abant İzzet Baysal Üniversitesi, Bilgi İşlem Daire Başkanlığı veri tabanından gerekli izinler alındıktan sonra temin edilmiştir. İki farklı senaryo ile mezuniyet notu tahmini yapılmıştır. İlk senaryo ile

birinci sınıfa ait derslerin yılsonu notları ile mezuniyet notu tahmin edilirken, ikinci senaryoda birinci ve ikinci sınıfa ait derslerin yılsonu notları ile mezuniyet notu tahmin edilmiştir.

### 1.2 Araştırmanın Sınırlılıkları

Bu tez çalışması,

2011-2018 öğretim yılı ile sınırlıdır.

Çalışma Kamu Yönetimi öğrencilerini kapsamaktadır.

Tez çalışmasında makine öğrenmesi yöntemlerinden Yapay Sinir Ağları (YSA), Karar Ağaçları (KA), K-En Yakın Komşu Algoritması (K-NN), Lineer Regresyon (LR), Destek Vektör Makineleri (DVM) yöntemleri kullanılmıştır.

İlgili bütün yazılımlar Anaconda.Navigator içinde bulunan Orange.3 ortamı ile sınırlıdır.

### 1.3 Literatür ve İlgili Araştırmalar

Yükseköğretimde veri madenciliğinin farklı uygulamaları üzerine birçok araştırmacı ve yazar araştırma yapmıştır ve tartışmıştır. Yükseköğretimde veri madenciliği alanındaki uygulamaların önemi ve bu alandaki kaliteyi yükseltmek için yazarlar detaylı literatür taraması yapmaktadırlar. Bu bağlamda 2008 yılında Ranjan ve Khalil tarafından yapılan çalışmalarında eğitim yönetime yönelik bir veri madenciliği süreci önermiştir. Yapılan bu çalışmada yer alan örnekler veri madenciliği teknikleri ile hazırlanmış olarak yer almış, ilerisi için yöntem önerileri, dersler ve ders çalışmanın sınırlılıkları irdelenmiştir (Ranjan ve Khalil, 2008). Sembiring, Zarlis, Hartama, Ramliana ve Wani (2011) öğrenciler üzerinde yaptıkları çalışmada, davranış ve başarı analizi yapmışlar ve öğrencilerin performans tahmin modeli hazırlamak amacıyla veri madenciliği teknikleri kullanmışlardır. “FarthestFirst” ve “Weka J48” algoritmalarını veri kümeleme ve sınıflandırma yöntemleri için kullanarak öğrenci akademik başarı durumunu tespit amaçlı Bresfelean, Bresfelean ve Ghisoiu (2008) yılında çalışma yapmışlardır. Zhang, Oessena, Clark ve Kim (2010) tarafından veri madenciliği üzerine yapılan çalışmada ders ve modülün öğrenci üzerindeki uygunluğunun nasıl değerlendirilebileceğini ve sonuçların öğrenciler üzeirne nasıl uyarlanacağını konu alan bir çalışma yapmışlardır. Mardikyan ve Badur (2011) yılında regresyon

ve karar ağaçları yöntemleri ile öğretim üyelerinin öğretme performanslarını değerlendiren bir çalışma yapmışlardır. Gaafar ve Khamis (2009) yılında Kahire Amerikan Üniversitesinde makine mühendisliği bölümünde okuyan öğrenciler üzerine yaptıkları çalışmada, mezun olabilecek konumdaki öğrencilerin profillerini tespit edip belirleyebilecek yöntem önerisinde bulunmuşlardır. Veri madenciliğinin farklı yöntemlerini kullanmışlar, farklı birimlerden aldıkları veriler üzerinden bir veri tabanı meydana getirmişler ve öğrenciler üzerinde iki farklı model oluşturmuşlardır. Oluşturulan moderllerden birincisi mezun olabilecek yani başarılı öğrenciler iken ikinci öğrenci profili okulu bırakabilecek yani başarısız öğrencilerden oluşmaktadır. Erdoğan ve Timor (2005) tarafından yapılan çalışmada Maltepe Üniversitesindeki 722 öğrencinin bazı karakteristikleri K-ortalamalar kümeleme algoritması kullanılarak kümeleme işlemi yapılmıştır. 2007 yılında belirlenen bir dersi alan öğrenciler üzerinde yapılan ve öğretim kalitesinin artırılması üzerine yapılan çalışmada Vranic, Pintar ve Skocir (2007) yer almıştır. Minaei-Bidgoli, Kashy, Kortmeyer ve Punch (2003) yılında genetik algoritma kullanılarak web tabanlı eğitsel veri tabanında tutulan veriler ile öğrencilerin dinal sınavında alacakları notu tahmin eden bir çalışma yapmışlardır. Winnow, en yakın komşu ve artımsal Bayes algoritmalarını kullanan Kotsiantis, Patriarcheas ve Nikxenos (2010) yaptıkları çalışma ile birleşimsel olarak kullnılabilecek bir yapıyı önermişlerdir. Sen ve diğerleri (2012) yaptıkları çalışmada orta öğretim yerleştirme sınav sonuçlarının tahmin edilmesi amacıyla çeşitli veri madenciliği teknikleri kullanmışlardır. Çalışmada duyarlılık analizleri gerçekleştirerek yerleştirme sınavına etki eden parametreler arasından en çok etkili olanının belirlenmesi sağlanmıştır. Yapılan bu çalışmada C.5 karar ağacı yöntemi destek vektör makineleri, lojistik regresyon ve yapay sinir ağları yöntemlerine göre daha iyi başarım ürettiği tespit edilmiştir. Ben-Zadok, Hershkovitz, Mintz ve Nachmias (2007) yaptıkları çalışma ile final sınavı öncesi risk altında olabilecek öğrencilerin tespiti için öğrenci öğrenme davranışlarını veri madenciliği yöntemleri ile analiz etmişlerdir. Zaiane ve Luo (2001) yılında yaptıkları çalışmada eğitimciler ve öğrenme süreçlerinin en iyi şekilde değerlendirilmesi için web tabanlı öğrenme ortamı tasarımında veri madenciliği ile makine öğrenmesi yöntemlerinin kullanılabilirliği üzerinde tartışmıştır. Şengür (2013) çalışmasında mezun öğrencilerin mezuniyet not tahmini için yapay sinir ağları ve karar ağaçları yöntemlerini kullanmış ve doğruluk tahmin oranlarını karşılaştırmıştır. Aybek

(2016) yüksek lisans tezinde belirlenen bir ders üzerinden öğrencilerin dönem sonu sınav puanları ve dersten geçme-kalma durumlarını yapay sinir ağları tekniği tahmin edilmesi üzerine çalışma yapmıştır. Kılınç (2015) yapmış olduğu çalışmada sınıflandırma ve birliktelik kuralları algoritmalarını kulanarak üniversite öğrenci başarısı üzerine etki eden faktörlerin tahmini yapmıştır. Selvi (2020) Bilecik ili özelinde ilköğretimden liseye geçiş sınavlarındaki öğrenci başarı tahmini için makine öğrenmesi metodlarını kullanarak çalışma yapmıştır. Al-Khafaji (2021) e-öğrenme yönetim sistemini kullanan bir ortamda sınava giren öğrencilerin başarılarının tahmini üzerine hem yapay sinir ağı hem de bulanık mantık içeren yapay zekâ teknikleri kullanarak çalışma yapmıştır. Kaya (2022) makine öğrenmesi yöntemlerinden Rastgele Orman, Aşırı Gradyan Güçlendirme ve Destek Vektör Makineleri algoraitmalarını kullanarak öğrencilerin akademik başarısı üzerinde etkili olan faktörlerin tespit edilmesi üzerine çalışma yapmıştır. Aydoğan ve Zırhlıoğlu (2018) yaptıkları çalışmada yapay sinir ağlarını kullanarak öğrencilerin başarısının tahmini üzerine modelleme yapmışlardır. Özdemir vd. (2018) eğitim sisteminde veri madenciliği uygulamaları ve farkındalık üzerine durum çalışması yapmışlardır.

# 2. YAPAY ZEKA

İlk olarak 1950'li yılların ortalarında bir bilgisayar bilimi olarak ortaya çıkmış olan Yapay zekâ kavramını günümüzde oluşturuan yapay sinir ağları, makine öğrenmesi, uzman sistemler, bulanık mantık, doğal dil işleme ve görüntü işleme gibi farklı metotlar üzerinde yoğun ve hızlı şekilde çalışmalar geliştirilmeye devam etmektedir (Pham ve Pham, 1999). John McCarthy tarafından Amerika'da 1956 yılında Dartmouth Kolejinde düzenlenen Dartmouth Konferansı'nda "Yapay Zeka" terim olarak ilk defa önerilmiştir (Zhang ve Lu, 2021). Mind dergisinde 1950 yılında İngiliz matematikçi Alan Mathison Turing kaleme alınan "Computing Machinery and Intelligence" adlı makalesinde söz ettiği "Makineler Düşünebilir mi?" (Can Machine Think?) sorusu ile yapay zekâ tarihinin başladığı kabul edilir (Winston, 2017). Alan Turing bu makalesinde "Taklit Oyunu" (Imitation Game) adındaki oyunda 3 oyuncudan bahseder. Birisi erkek(A), birisi kadın(B) ve sorgulayıcı (C) erkek ya da kadın olarak oyuncular tanımlanmıştır. Sorgulayıcı pozisyonundaki oyuncu diğer iki oyuncunun önünde olup, bu iki oyuncuyu göremeyecek odada bulunur. A ve B'nin, sorgulayıcının sorularına verdiği cevaplar ses tonlarından dolayı belli olmasın diye yazılı olarak sunulur ve sorgulayıcı bu şekilde cevabın kime ait olduğunun tahminini yapmaya çalışır. Oyunda A'nın görevi verdiği cevaplar ile sorgulayıcıyı (C) yanıltmak, B'nin amacı ise sorulara verdiği cevaplar ile sorgulayıcıya (C) yardımcı olmaktır. Bu oyunda A okişisinin yerini bir makine alırsa ne olacağı sorusu "Makineler Düşünebilir mi?" sorusudur (Winston, 2017). Sorgulayıcı karşısındakilerden hangisinin erkek ya da kadın hangisinin makine olduğunu belirleyebilir mi? Makinelerin düşünmesi üzerine sunulan bu fikir yapay zekâ alanındaki çalışmalar için başlangıç olarak kabul edilir.

Yapay zekanın literatürde birçok farklı tanımı yapılmakla birlikte en genel anlamda insan zekasını taklit ederek topladıkları bilgilerle kendini iyileştirebilen sistem veya makine olarak adlandırılabilir. Hızla gelişen teknoloji ile paralel yapay zekâ tekniklerindeki ilerleme sayesinde insan hayatında da çok büyük kolaylıklar sağlamıştır. Sürücüsüz araçlar, sanal asistanlar, tercihlere göre öneri sunumları, siber güvenlik açıkları, hastalıkların tedavi süreçleri gibi farklı birçok alanda yapay zekâ sayesinde hızlı ve güvenilir çözümler üretilmektedir. Bununla

beraber finans, sağlık, ticaret, eğlence, eğitim, ulaşım, mühendislik gibi farklı alanlarda insanlığa yardımcı olmaktadır. Yapay zekayı kendi başına bir bilim dalı olarak düşünmekten ziyade matematik, fizik, kimya, bilgisayar bilimleri, biyoloji, felsefe, fizyoloji, elektronik gibi biçok alan ile ilişkilidir. Bu alanlarda yapar sinir ağları, makine öğrenmesi, genetik algoritmalar, bulanık mantık gibi yapay zekanın alt dallarında kullanılmaktadır.

## 2.1 Veri Madenciliği

Makine öğrenmesi ve veri madenciliği arasında yakın bir ilişki vardır. Veri madenciliği verilerden elde edilen bilgi ve bilginin değerlendirilmesi ile ilgilenirken makine öğrenmesi bilgiyi elde etmeyi sağlayan yöntemleri kullanan bilgisayar programlarının geliştirilmesi, öğrenme yöntemlerinin geliştirilerek tahminleri veya tanımları yüksek bir performans ile nasıl ortaya çıkarılabileceği ile ilgilenmektedir. Her ikisi de benzer süreçlerden geçmekle birlikte aralarında çok belirgin bir fark yoktur (Balaban ve Kartal, 2018). Değerli bilgileri elde etmek için büyük ölçekli verilerin kullanıldığı, veriler arasındaki ilişkilerin ortaya çıkarıldığı ve ilerisi için tahminde bulunma işlemi olarak açıklanabilir veri madenciliği (Özkan, 2013). Üzerinde çalışılan projenin hız, verim ve maliyet bakımından en yararlı şekilde yapılmasını sağlar veri madenciliği süreci. Çeşitli aşamalardan oluşan veri madenciliği süreci aşağıdaki gibidir.

**Problemin Tanımı:** Araştırma yapılacak konunun seçilmesi.

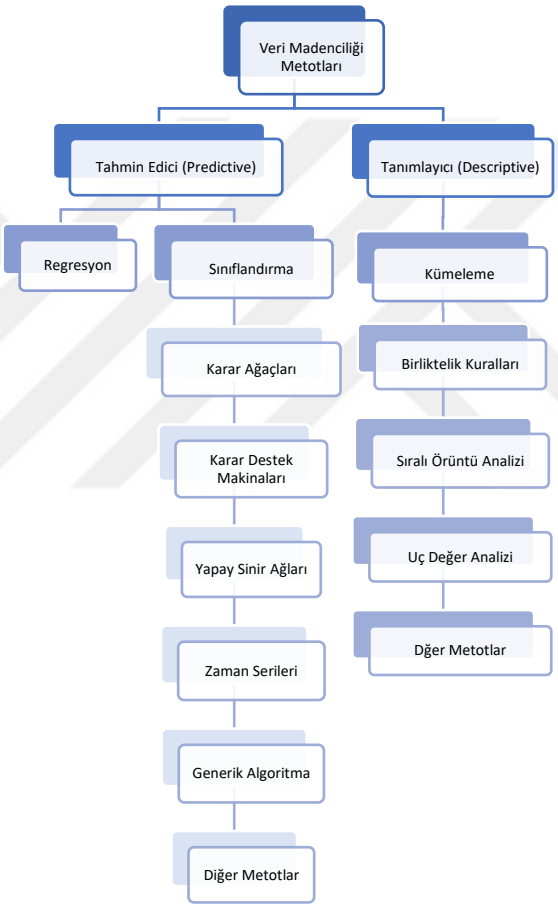
**Veri Entegrasyonu:** Birden fazla kaynaktan verinin toplanıp, seçilip tek kaynakta birleştirilmesi.

**Veri Ön İşleme:** Hatalı veya eksik olan verilerin temizlenmesi, dönüşümünün yapılması ve analize hazır biçime getirilmesidir.

**Model Kurma:** Veri madenciliği yöntemlerinin uygulandığı aşamadır. Doğru modelin kurulması için birçok farklı algoritma mevcuttur. En başarılı sonucu veren algoritma uygun olanlar arasından seçilir. Uygulanan yöntemler neticesinde elde edilen sonuçların performanslarına bakılır.

**Değerlendirme:** Performans olarak en iyi sonucu üreten yöntemin sonuçları değerlendirilir.

Veri madenciliğinde tanımlayıcı (descriptive) ve tahmin (predictive) olmak üzere iki farklı model kullanılmaktadır. Karar süreçlerine rehberlik etmede kullanılacak verilerdeki örüntülerin tanımlanması iin tanımlayıcı veri madenciliği modelleri kullanılırken, tanımlayıcı modellerde kümeleme ve ilişki analizi daha yaygın olarak kullanılmaktadır. Sonuçları önceden bilinen veri topluluklarından geliştirilen model ile sonuçları bilinmeyen veri kümeleri için model kurulması ile sonuçların değerlendirilmesi tahmin edici veri madenciliğinin özelliğidir. Genel olarak sınıflandırma ve regresyon tahmin edici modellerde kullanılmaktadır.




**Şekil 2.1.** Veri madenciliği yöntemleri

### 2.1.1 Veri Madenciliği Yöntemleri

Veri madenciği yöntemlerinden en çok kullanılanları üç ana başlık altında inceleyebiliriz (Çınar, 2019). Bunlar,

Sınıflandırma

Kümeleme

Birliktelik Analizi

#### 2.1.1.1. Sınıflandırma Yöntemi

Bu yöntemde veri kümesinde bulunan veriler ortak özelliklerine göre sınıflara ayrılır. Veri tabanlarındaki gizli örüntülerin meydana çıkarılması için kullanılmaktadır (Çınar, 2019). Bu yöntemde amaç kategorinin kestirilmesidir. Örneğin hava durumunun bulutlu, karlı, yağmurlu ya da güneşli sınıflarından hangisinde olacağının tahmininde bulunmasıdır. Yapay sinir ağları, destek vektör makineleri, regresyon analizi, karar ağaçları sık olarak kullanılan sınıflandırma teknikleridir. Karater-ses tanıma, finans alanında kredi başvurusu, sağlık hizmetlerinde hastalık teşhisi gibi farklı birçok alanda uygulanmaktadır.

#### 2.1.1.2. Kümeleme Yöntemi

Kümeleme yöntemi ile veriler arasındaki benzer özelliklerin değerlendirilerek gruplara ayrılması işlemi yapılır. Önceden belirli olan sınıfların olmaması nedeni ile kümeleme yöntemi ile sınıf oluşturma işlevi ön plana çıkmaktadır. Bir nevi sınıflandırma yaklaşımı da diyebileceğimiz küme analizinde baımlı değişken değerleri olmamasından dolayı nesnelerin sadece ortaya konan değişken değerlerine göre sınıflandırma işlemi yapmaktadır. Nesneler arasında değişken değerleri benzer olanlar bir araya getirilmek sureti ile kümeler meydana getirilmektedir. Kümeleme işlemi farklı uzaklık ölçüleri ile veri uzayı içinde yoğunlaştıkları noktalar gibi yaklaşımlar ile yapılmaktadır. Örnek olarak Mannattan uzaklığı, Minskowski uzaklığı verilebilir (Akpınar, 2014). Hiyerarşik ve hiyerarşik olmayan kümeleme yöntemleri olarak iki ana başlıkta incelenebilir. En yakın komşu algoritması ve en uzak komşu algoritması hiyerarşik kümeleme yöntemleri arasında yer alırken k-ortalamalar yöntemi hiyerarşik olmayan kümeleme yöntermleri için örnek verilebilir. Kümeleme yöntemi tıp, kimya, jeoloji, uzay araştırmaları gibi alanların yanı sıra segregation, antropometri gibi alanlarda da uygulamaktadır (Akpınar, 2014).

#### 2.1.1.3. Birliktelik Analizi

Veri setini oluşturan kayıtlar arasındaki ilişkileri inceleyerek olayların eş zamanlı olarak beraber gerçekleşme durumlarını ortaya koymaktadır. Bu işlemler için destek ve güven ölçütleri kullanılmaktadır (Özkan, 2013).

En yaygın olarak müşteri satın alma eğilimlerinin belirlenmesinde kullanılan birlitelik analizi yönteminin en bilinen algoritması Apriori algoritmasıdır.

## 2.2 Makine Öğrenmesi

Açık bir şekilde programlamaya gereksinim duyulmadan bir sistemin eldeki veriler üzerinden öğrenmesini sağlayan yapay zekanın alt dallarından biridir makine öğrenmesi. Makine öğrenmesi yöntemi ile veriler analiz edilerek geleceğe yönelik sonuçların tahmininde bulunulması işlemi gerçekleştirilir (Liu ve Lang, 2019). Sisteme girdi olarak adlandırılan veri üzerinde matematiksel formüller içeren algoritmaların ugulanması ile hedeflenen çıktının tahmin edilmesi amaçlanır.

Yapay zekâ alanının ve bilgisayar oyunlarının öncü isimlerinden Arthur Lee Samuel tarafından ilk defa makine öğrenmesi terimi kullanılmıştır (Jackson, 1988). 1959 yılında "IBM Journal of Research and Development" dergisinde yayınlanan "Some Studies in Machine Learning Using the Game Chekers" adlı makalesinde bu yaklaşımını açıklamıştır (Kirsch ve Hurwitz, 2018). Yayınlanan bu makalesinde dama oyununun özellekleri üzerinde makine öğrenmesi prosedürünü araştırmış, oyunun kurallarının verildiği zaman programın bunu kısa sürede öğrenme işlemini gerçekleştirdiğini ve oyunu hazırlayan programcıya nazaran bilgisayarın daha iyi oynadığını belirtmiştir (Samuel, 1988).

Makine öğrenmesi sürecinde birtakım faktörler etkilidir. İnsan hayatında deneyimin yaşama etkisi ne kadar fazla ise makine öğrenmesi sürecinde makineye deneyim kazanması için verilen veri setleri de o derece önem arz eder. Kazanılan deneyimin farklılığı ve fazlalığı sistem için öğrenme adına pozitif katkı sağlar (Balaban ve Kartal, 2018).

Deneyim kazanma adına elimizdeki değişkenlerin sistemin elde edeceği sonuca etkisi de fazladır. Örneğin hava kirliliğinin tespiti için kullanılan araç sayısı, yeşil alan, kullanılan yakıt, enerji türü, bir futbol maçının oynanıp oynanmayacağının tahmini için havanın ısı, nem, yağış, rüzgâr gibi farklı özelliklerine bakılır ve bu veriler değişken olarak dikkate alınır. Bu tarz özellikler makine öğrenmesinde nitelik (attribute) olarak adlandırılmaktadır.

Makine öğrenmesinde sistemi etkileyen bir diğer faktör de öğrenme stratejisinin belirlenmesidir. Bazı problemlerin çözümünde veri setinde çıktı değerlerine ihtiyaç duyulurken bazı problemlerin çözümünde çıktı değerlerine ihtiyaç duyulmaz. Çıktı değerlerine hedef nitelik (target attribute) olarak adlandırılır. Bir futbol müsabakasının oynanıp oynanmayacağı (Evet/Hayır) ile ilgili tahmin yapılırken ya da bir bankanın müşterileri için kredi risk gruplarına göre sınıflandırma yaparken risk grupları (Düşük/Orta/Yüksek) gibi örnekler hedef niteliğe ihtiyaç duyulan problemlerin çözümünde daha çok tahmin ve sınıflandırma üzerine yoğunlaşıldığı görülmektedir. Bu da bizlere makine öğrenmesinde eldeki problem için en uygun öğrenme stratejisi ve yolunun belirlenmesinin önemini göstermektedir.

Son olarak bir diğer fakör de makine öğrenmesi için kullanılan algoritmalar ve bunlara ait parametrelerin belirlenmesidir. Problemin çözümünde adım adım izlenmesi gereken yol algoritmadır. Sisteme girilen veri setinin makine tarafından öğrenilmesi için de adımlar mevcuttur. Bu amaçla K-en yakın komşu, destek verktör, K-ortalamalar gibi farklı algoritmalar geliştirilmiştir. Her bir algoritma öğrenme işlemi için kendine özgü tarzda çalışır. Kimisi k sayılarını analiz öncesi parametre olarak gerektirirken, kimi algoritmalar da ise k sayılarını parametre olarak kullanmazlar (Balaban ve Kartal, 2018).

Genel olarak makine öğrenmesinde sistem için gerekli veriler eğitim ve test aşamalarından geçirilirler. Model oluşturmak için eğitim sürecinde belirlenen oranda veri kullanılmaktadır. Kullanılan bu veriye eğitim verisi denilmektedir. Test sürecinde ise eğitim süreci ile meydana getirilen model üzerinde test işlemi yapılması için ayrılan ya da zamanla girilecek veriler sunulmaktadır.

Makine öğrenmesi sistemini kısaca özetlemeke gerekirse,

$$Y=f(x)+\in \tag{2.1}$$

Y : Hedeflenen sonuç

f(x) : Kullanılan fonksiyon

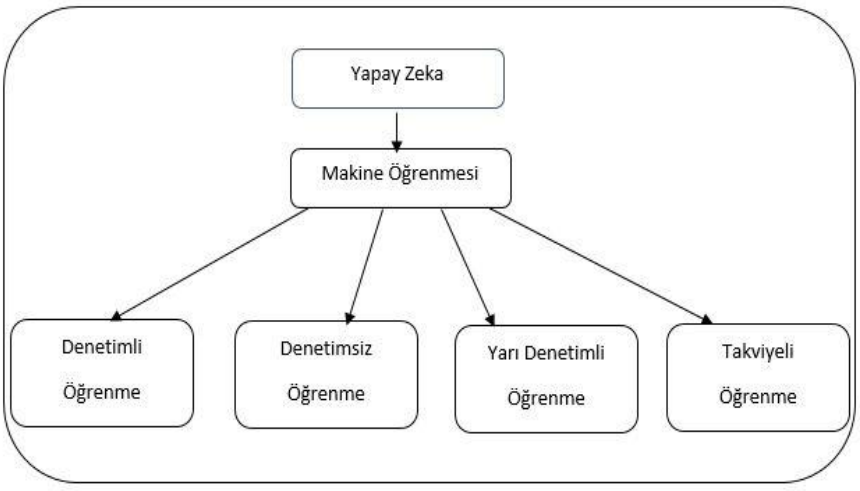
$\in$ : Hata

Makine sisteme girilen eğitim verilerinden f fonksiyonunu öğrenir ve bununla birlikte oluşturulan model sonucunda girilen test verilerinden Y'yi kestirir (Gürsakal, 2018)

Makine öğrenmesi günümüzde hastalıkların teşhisi, finans, biyoloji, ekoloji, tarım, sanayi, güç sistemleri ve daha birçok farklı alanda regresyon, sınıflandırma ve kümeleme problemlerinin çözümünde ve analizinde kullanılmaktadır. Dijital dünya teknolojisinin hızlı bir şekilde ilerlemesi ile dijital dönüşüm alanında da makine öğrenmesi kurumların müşterileri için değerli işler yapması noktasında yarar sağlamaktadır.

## 2.3 Makine Öğrenmesi Metotları

Makine öğrenmesinde genel olarak dört farklı metot kullanılmaktadır. Bunlar Denetimli Öğrenme (Supervised Learning), Denetimsiz Öğrenme (Unsupervised Learning), Yarı-Denetimli Öğrenme (Semi-Supervised Learning) ve Takviyeli Öğrenme (Reinforcement Learning) olarak tanımlanırlar. Bu tez çalışmasında denetimli öğrenmenin sınıflandırma algoritmaları ile veri seti üzerinde tahmin çalışması ve doğruluk oranları karşılaştırması yapılmıştır.

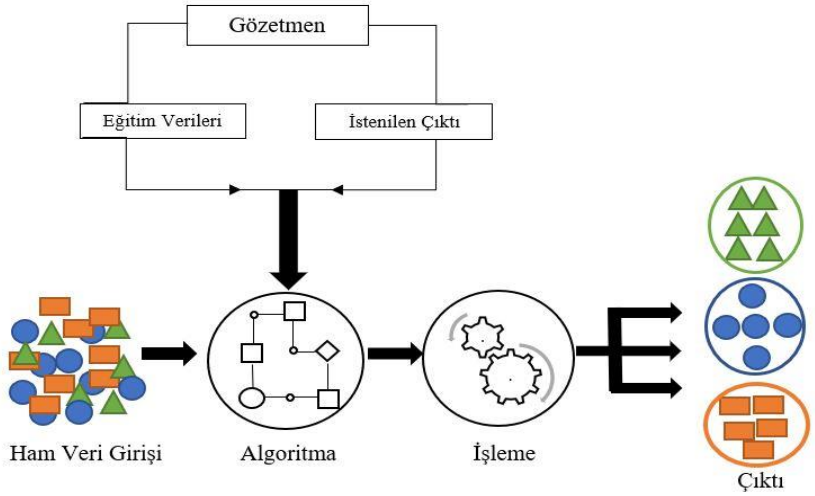



**Şekil 2.2.** Makine öğrenmesi metotları

### 2.3.1 Denetimli Öğrenme

Denetimli öğrenme metotunda algoritmaların uygulaması esnasında verilerin etiketli eğitim veri seti şeklinde olması gerekir. Veri setlerindeki öznitelik değerlerine ile bunlara karşılık gelen hedef değişkenlerine ait öznitelikler arasındaki ilişkinin keşfedilmesi üzerine çalışan denetimli öğrenmenin, regresyon

ve sınıflandırma olarak iki farklı modeli vardır (Elmas, 2011). Sınıflandırma modeli ile girdi değerleri önceden belli olan sınıflardan birine eşleme işlemi yapan bir sınıflandırıcı oluşturulur. Tez çalışmasında denetimli öğrenme yönteminin algoritmalarından olan K-en Yakın Komşu Algoritması, Karar Ağacı, Destek Vektör Makineleri, Lineer Regresyon ve Yapay Sinir Ağları yöntemleri kullanılmıştır.




**Şekil 2.3.** Denetimli makine öğrenmesi

#### 2.3.1.1. K-En Yakın Komşu Algoritması

Makine öğrenmesi sınıflandırma yöntemlerinden biri olan K-en yakın komşu yöntemi ilk olarak (Fix ve Hodges, 1952) tarafından parametrik olmayan bir yöntem olarak tanımlanmış olup örüntü tanımada kullanılmıştır. Daha sonra T. M. Cover ve P.E. Hart tarafından 1967 yılında geliştirilmiştir (Cover ve Hart, 1967). K-en yakın komşu algoritmasının (K-NN) çalışma prensibi temel oalrak test verisinin sınıflandırılması amacıyla öncelikle eğitim verilerine ait en yakın k adet komşularını bulmak ve kategori adaylarına özgün ağırlıklar vermek için en yakın komşu algoritmalarını kullanmaktır. K-NN algoritması sistem için uygulama yapılırken; öncelikle k-en yakın komşu sayısı belirlenir. Belirlenen noktaya en yakın mesafedeki komşuların belirlenmesi prosesi için söz konusu nokta ile geri kalan diğer noktalar arasındaki mesafeler tek tek hesaplanır (Özkan, 2013).

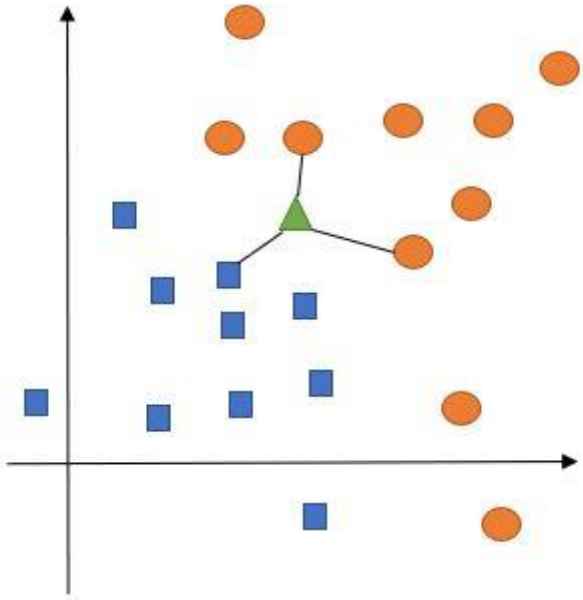
Mesafelerin hesaplanması esnasında kullanılan bazı fonksiyon ölçütleri aşağıdaki gibidir,

$$\text{Öklit Fonksiyonu} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{2.2}$$

$$\text{Manhattan Fonksiyonu} = \sum_{i=1}^{k}|x_i - y_i| \tag{2.3}$$

$$\text{Minkowski Fonksiyonu} = \left(\sum_{i=1}^{k}|x_i - y_i|^q\right)^{1/q} \tag{2.4}$$

Formüllerde yer alan x ve y değerleri aralarındaki yakınlık ölçülen iki farklı değişkeni temsil ederken, q ise değişken sayısını ifade etmektedir.



**Şekil 2.4.** K-en yakın komşu algoritması

K-NN algoritması uygulanırken bulunan uzaklıklar için sıralama yapılırken en küçük değerdeki k tanesinden en çok sayıdaki tekrarlanan kategori değeri seçilir (Özkan, 2013). Bu algoritma örnek tabanlı bir makine öğrenmesi algoritması olarak adlandırılabilir. Çalışma şekli örnek olarak verilen verileri ezberlemeye dayandığından özgün bir modele sahip değildir. Algoritmanın uygulanması ve anlaşılması oldukça kolaydır. K-NN algoritması sınıflandırma, regresyon problemlerinin çözümünde oldukça sık şekilde kullanılırken e-postaların spamlarının filtrelenmesi, müşterilerin kredi fizibilitesi, el yazısının

tanımlanması ve şahsa özel ilaç tespiti gibi farklı farklı alanlarda sıklıkla tercih edilen bir yöntemdir.

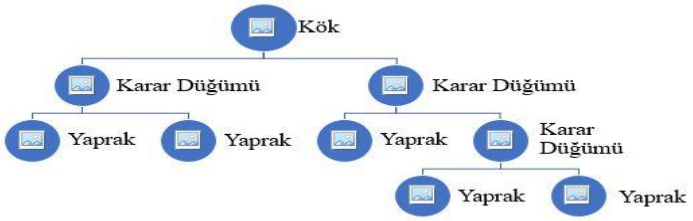
#### 2.3.1.2. Karar Ağacı

Veri setini oluşturan değişkenlerin bölümlenerek bir ağaç yapısını oluşturması mantığına dayandırılan bu sınıflandırma algoritması ilk olarak Bierman ve Friedman tarafından 1937 yılında önerilmiştir. Problemi meydana getiren girdi, çıktı ve olası durumlara ait olasılıkları gösteren, yapısı itibarı ile sade ve anlaşılması kolay, şekilsel gösterime dayalı bir algoritma türüdür. Öznitelik değerlerine göre iki veya daha fazla sayıda alt bölümlere ayrılabilen düğümlerden oluşur (Mashat vd., 2012). Ağacı meydana getiren her bir dal düğümdeki testin sonucunu gösterirken, yaprak olarak adlandırılan düğümler ise sınıf etiketini gösterir (Sharma ve Kumar, 2016). Genelleme hatasının minimize olarak amaçlandığı bu algoritmada hedef optimal karar ağacının bulunmasıdır. Şekilsel olarak seçim ve sonuçların bir ağaç olarak görselleştrildiği bu yapıyı oluşturan düğümler her bir seçimi gösterirken, kenarlar da koşulları ifade etmektedir. Genel olarak kök, iç düğüm (karar düğümü) ve yaprak bölümlerinden meydana gelir.

**Kök düğüm:** Veri setini oluşturan tüm örnekleri barındıran düğümdür.

**İç düğüm (karar düğümü):** Koşul ve kısıtlamaların kontrol edildiği düğümdür.

**Yaprak düğüm:** Yapılan sınıflandırmanın sonucunun yer aldığı düğümdür (Jackson, 1988).




**Şekil 2.5.** Karar ağacı algoritması genel yapısı

Karar ağaçları yönteminde en çok kullanılan algoritmalar ID3, C4.5 ve CART algoritmalarıdır (Liu ve Lang, 2019).

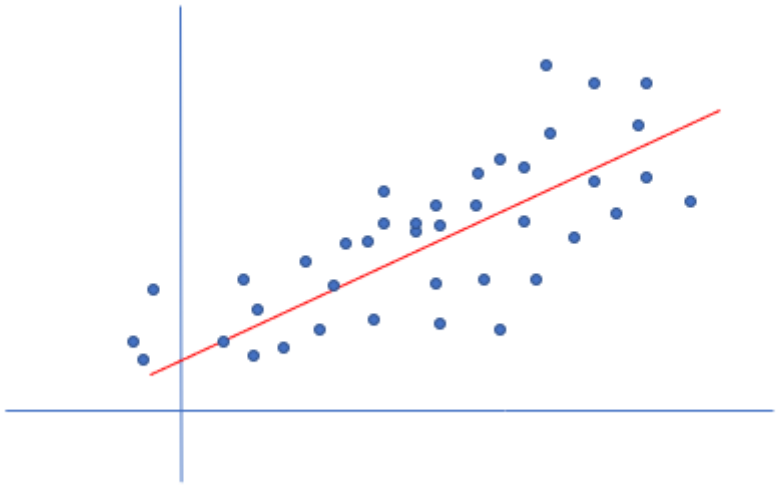
**ID3 (Iterative Dichotomiser 3):** Kategorik veriler ile kullanılan bu algoritma gürültülü verilere karşı direnç gösteremez.

**C4.5:** Kategorik ve sayısal veriler üzerinde kullanılabilen bu algoritma ID3 algoritmasının iyileştirilmiş halidir (Priyam vd., 2013).

**CART (Classification and Regression Trees):** Veri setinde oluşabilecek aykırı değerlerden etkilenmeyen bu algoritma ikili ağaçlar geliştirerek kategorik ve sayısal verileri kullanabilmektedir.

#### 2.3.1.3. Lineer Regresyon

Lineer regresyon (LR), makine öğrencmesi ve veri biliminde kullanılmakta olan kompleks yapıda olmayan X ve Y değişkenleri arasındaki ilişkiyi modellemede kullanılan bir doğrusal yaklaşım ve algoritmadır. Lineer regresyon algoritması yaygın olarak eğilimlerin tahmin edilmesi, oluşturulması gibi farklı analizlerin tahmin işlemlerinde kullanılan modellerden biridir. Başka bir ifade ile LR, en uygun çizgi olarak ifade edilebilecek düz çizgi anlamına gelen regresyon çizgisini referans alarak bir veya birden fazla olabilecek bağımsız X değişkeni ile bununla bağımlı olan Y değişkeni arasında ilişki kurulmasını sağlar.



**Şekil 2.6.** Lineer regresyon basit gösterim

Tek bir bağımsız değişken (X) ile buna bağlı değişken (Y) arasındaki ilişki basit lineer regresyon ile incelenilmektedir. X ve Y'den birinin değeri bilindiği durumda diğerinin değerinin hesaplanması sağlanır. Basit lineer regresyon formülü aşağıdaki gibidir (Kılıç, 2013).

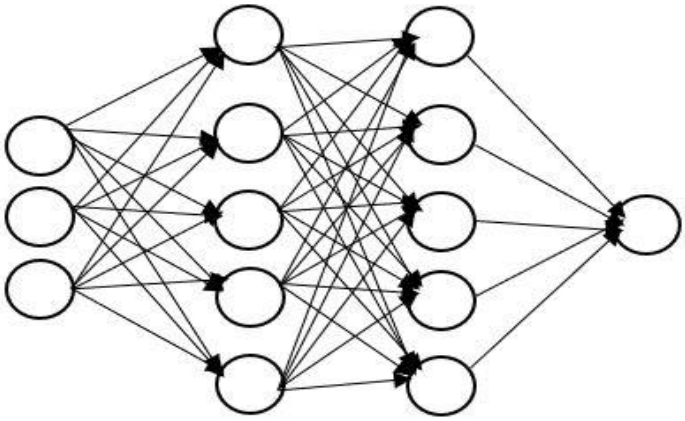
$$Y = \beta_0 + \beta_1 X \tag{2.5}$$

Bağımsız değişkenin birden fazla olduğu durumlarda bağımlı değişken ile arasında meydana gelen ilişki çoklu regresyon ile analiz edilir. Modelimizde yer alan bağımsız değişken X'in katsayılarını sabir değer olan $\beta$ gösterir. Denklemde yer alan $\in$ tesadüfi hata terimi olarak tanımlanır. Çoklu lineer regresyon formülü aşağıdaki gibidir. (Kılıç, 2013).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \in_i \tag{2.6}$$

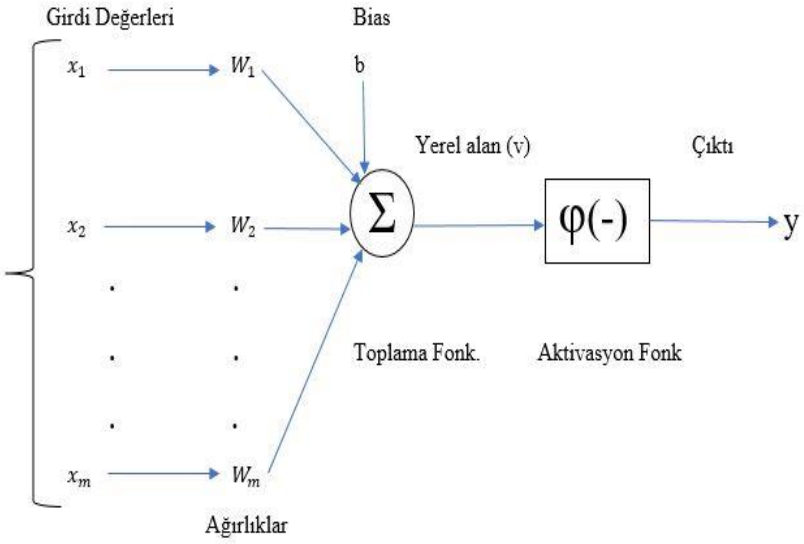
#### 2.3.1.4. Yapay Sinir Ağları

Öznitelik değişkenleri ile hedef arasındaki bağlantıyı analiz katmanları sayesinde işlemek için esnek bir tasarım sunan bir makine öğrenme algoritmasıdır. Yapay sinir ağlarının girdi, gizli (ara) ve çıktı olarak üç ana katmandan meydana gelen bir yapısı vardır. Mantıksal olarak biyolojik sinir ağlarının taklidi denilebilecek sentetik bir yapıya sahiptir (Eğrioğlu vd., 2009). Her bir katman bir input olarak değerlendirilir ve bir sonraki katmana output değerini iletir. Denetimli ve denetimsiz öğrenme yöntemleri için kullanılabilen YSA insan beyninin özelliklerine benzer şekilde paralel olarak işlem yapabilme kabiliyetine sahip, eksik ve gürültülü verilerin olduğu durumlarda da çalışabilmektedir. YSA'da farklı öğrenme kuralları mevcut olmasına rağmen genellikle denetimli öğrenme kurallarından biri olan geri yayılım (Back Propagation) algoritmasını tahmin ve sınıflandırma problemlerinde yüksek performans göstermesinden dolayı daha fazla tercih edilir. Sisteme tanıtılan örnekler üzerinden öğrenme işleminin gerçekleştiği geri yayılım yönteminde, değişkenler arasındaki matematiksel ilişkilere yer verilmez (Goh, 1995).



**Şekil 2.7.** Yapay sinir ağları genel yapısı

Yapay nöronlardan oluşan YSA girdi (input) değerleri ağırlık değerleri (W) ile çarpılır (multiplication). Daha sonra çarpılan bu değerler ve gerçek değer ile tahmin değeri arasındaki mesafeyi gösteren yanlılık (bias) değerleri toplanır (sum) ve bu toplama işleminin ardından elde edilen değerler etkinleştirme (activation ya da transfer function) fonksiyonundan geçerek çıktı meydana gelir.




**Şekil 2.8.** Yapay sinir ağları çalışma prensibi

YSA modelini oluşturan katman sayısı ve aktivasyon fonksiyonlarının modelin tahmin doğruluğu üzerinde etkisi oldukça fazladır. Bu fonksiyon ve

katmanlar girdi ile çıktı arasındaki lineer olmayan karışık eşleşmelerin öğrenilmesindeki rolleri çok önemlidir (Sharma ve Athaiya, 2017).

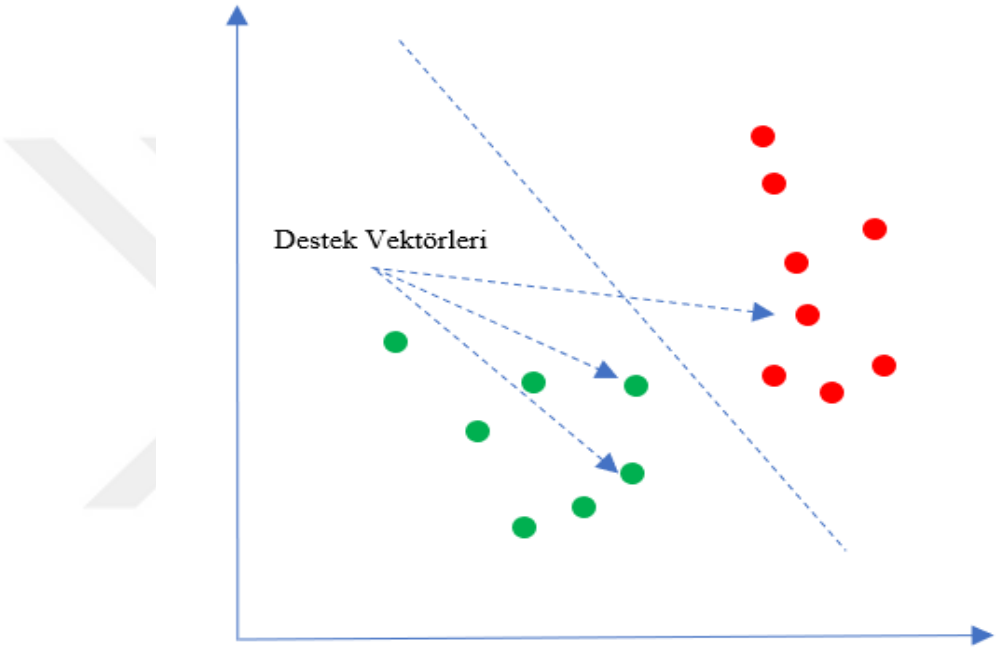
ReLU, Leaky ReLU, Sigmoid, Doğrusal, Swish, Softmax, Binary Step Function, Tanjant Hiperbolik gibi farklı birçok aktivasyon fonksiyonları arasında en çok tercih edileni, değerleri 0 ve 1 aralığına dönüştüren sigmoid aktivasyon fonksiyonudur (Sharma ve Athaiya, 2017). Sürekli ve türevi alınabilir bir fonksiyon olmasının yanında kayıp aktivasyon değeri oluşturmaz. Tanjant hiperbolik fonksiyonu genel olarak Sigmoid fonksiyonuna benzerlik gösteren bir yapısı vardır. Fonksiyon aralığı (-1,1) arasındadır. Bunun sayesinde daha çok sayıda değeri alması ile türevinin daha dik bir yapıda olması, böylece sınıflandırma ve öğrenme işini daha hızlı yapmasını sağlar. Sigmoid fonksiyonuna benzer şekilde gradyan kaybı vermektedir. (-1,1) aralık değerleridir. Doğrusal fonksiyon, doğrusal problemlerin çözümü için kullanılırken Sigmoid fonksiyonu gibi ikili değerler üretmez. Bunun sonucu olarak birden fazla çıkışa izin verir. Türevinin sabit olmasından dolayı model eğtimlerinde gerçekleştirilen geriye yayılıma (backpopagation) izin verir. Değer aralığı (-∞, ∞)'dur. ReLU (Rectified Linear Unit- Doğrultulmuş Lineer Birim) fonksiyonu yapı itibarı ile pozitif eksendeki doğrusal fonksiyona benzer özelliklere sahip gibi gözükse de aslında doğrusal olmayan bir yapıya sahiptir. Fonksiyonun değer aralığı nedeniyle negatif değerleri sıfıra çevirmesinden dolayı veri setlerinin eğitiminin azalmasına ve bundan dolayı da öğrenmenin zayıflamasına neden olabilmektedir. Tanjant Hiperbolik ve Sigmoid fonksiyonlarında olduğu gibi gradyan değer kaybı oluşturmaz. [0, ∞) aralık değerine sahiptir.

Biyolojik Sinir Sisteminin yapısını meydana getiren bileşenlerin Yapay Sinir Ağlarındaki karşılıkları: Nöron-İşlem elemanı, Dentrit-Toplama Fonksiyonu, Hücre Gövdesi-Ativasyon Fonksiyonu, Akson-Eleman Çıkışı, Sinaps- Ağırlıklar şeklinde terminolojik olarak adlandırılabilir (Öztürk ve Şahin,2018).

Yapay sinir ağları medikal uygulamalardan, bilgisayar ağları ve siber güvenliğe, finans, yönetim, pazarlama, üretim, ulaşım, mühendislik, tıp gibi birçok farklı alanda kullanılmaktadır (Choi ve Kim, 2021).

### 2.3.1.5. Destek Vektör Makineleri

Vaprik ve ekibi tarafından 1990'lı yıllarda meydana getirilen Destek Vektör Makineleri (DVM), makine öğrenmesi yöntemidir ve hem sınıflandırma hem de regresyon için kullanılmaktadır (Booser, Guyon ve Vapnik, 1992). DVM temelde matematiksel bir modeldir ve DVM'de temel amaç iki farklı sınıfı birbirinden ayıran en uygun karar verme fonksiyonunu (hiperdüzlemi) tahmin etmek ve hatanın minimize edilmesini sağlamaktır (Silahtaroğlu, 2020).




**Şekil 2.9.** Destek vektör makineleri basit gösterim

Sistemi meydana getiren iki sınıfın ayrımı için birçok farklı çizgi çizilebilir ancak optimum bir şekilde ayrım işlemi yapmak için aşağıdaki denklem kullanılabilir (Silahtaroğlu, 2020).

$$g(\vec{x}) = \overrightarrow{w}.\vec{x}+b \tag{2.7}$$

Formülde; b değeri iki sınıf arasındaki mesafeyi ifade ederken w olarak adlandırılan vektöre eklenecek olan b değeri ne kadar olursa diğer sınıfın etkileşim alanına girmesi sağlanmaktadır. Bu değer üzerinden yapılan hesaplamalarda bulunacak değer 0 olması halinde düzlem tam olarak ortada ve sınıfı da tam olarak ortadan ikiye ayırır denilmektedir. Formülde yer alan eşitliğin değeri 1'e eşit veya 1'den büyük bir değer olması veya -1'e eşit ya da küçük

olduğu durumlarda sistemi oluşturan sınıfların tahmini sağlanmaktadır (Silahtaroğlu, 2020). Formülü oluşturan x değeri girdi vektörünü ifade ederken g(x) fonksiyonu da karar fonksiyonudur. Doğrunun -1 ve +1 arasında oluşan bölgeye marjin denilmektedir.

$$g(\vec{x}) \geq 1 \tag{2.8}$$

$$g(\vec{x}) \leq -1 \tag{2.9}$$

$$z = \frac{|g(\vec{x})|}{||\vec{w}||} = \frac{1}{||\vec{w}||} \tag{2.10}$$

Bir sınıfın ayırıcı çizgi ile arasında olması gereken maksimum uzaklık, aşağıdaki gibi mesafenin toplamı ile bulunur.

$$\frac{1}{||\vec{w}||} + \frac{1}{||\vec{w}||} = \frac{2}{||\vec{w}||} \tag{2.11}$$

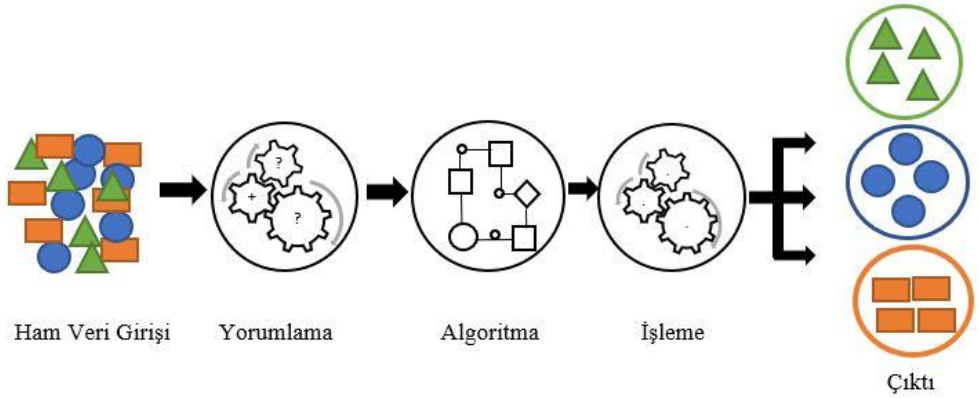
Destek vektörlerini hesaplamak için bu değerin minimum olarak belirlenmesi gerekmektedir. Sınıflar arasındaki mesafe son raddeye kadar ölçülerek arasındaki sınırın net bir şekilde belirlenmesi gerekmektedir. Bu işlemlerde wvektörü Langranj çarpanları yardımı ile maksimum seviyeye getirilmesi ile kolaylık sağlanır ve çözüm üretilir.

Diğer sınıflandırma yönetemlerine göre eğitim için geen sürenin daha uzun olmasına rağmen DVM, ezber öğrenmeye karşı mukavemet gücü ve lineer olmayan sınıflandırmadaki başarı seviyesi ile çokça tercih edilen bir yönetmdir (Akpınar, 2017). Makine öğrenmesinin sıklıkla kullanıldığı metin işleme, el yazısı tanımlama, görüntü işleme, jeoloji gibi farklı birçok alanda kullanılmaktadır.

### 2.3.2 Denetimsiz Öğrenme

Denetimsiz öğrenme metotunda veriler etiketsiz şekildedir. İlk olarak veriler üzerinde özniteliklerine göre gruplandırma işlemi yapılır sonrasında veriler üzerinde sınıflandırma işlemi gerçekleştirilir (Hurwitz ve Kirsch, 2018). Etiketlenmemiş olan yani sınıflandırılmamış, belli bir kategoriye ait olmayan veriler arasındaki ilişkilerin algoritmaların yardımı ile öğrenerek gruplara ayırma işlemi yapılır. Sıklıkla eğitim verisinin hazırlanması ve anlaşılmasının zor olduğu hallerde kullanılırlar. Denetimsiz algoritmalar genel olarak 3 başlıkta incelenirler.

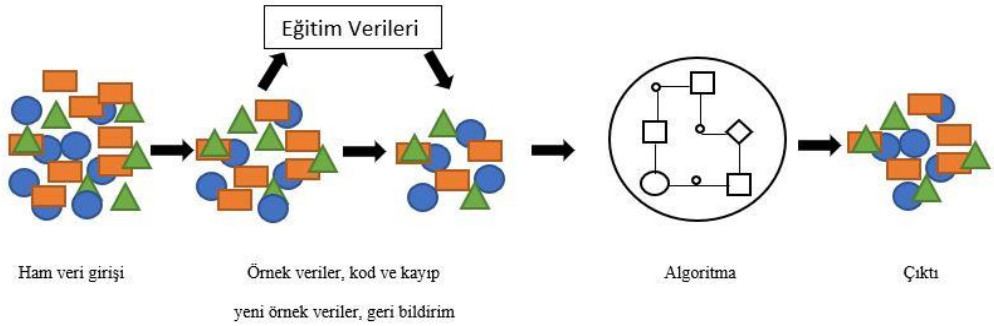
Bunlar: Kümeleme (Clustering), Birliktelik Kuralı (Asociation Rule Mining) ve Boyut Azaltma (Dimensionality Reduction) olarak adlandırılır.




**Şekil 2.10.** Denetimsiz öğrenme

### 2.3.3 Yarı Denetimli Öğrenme

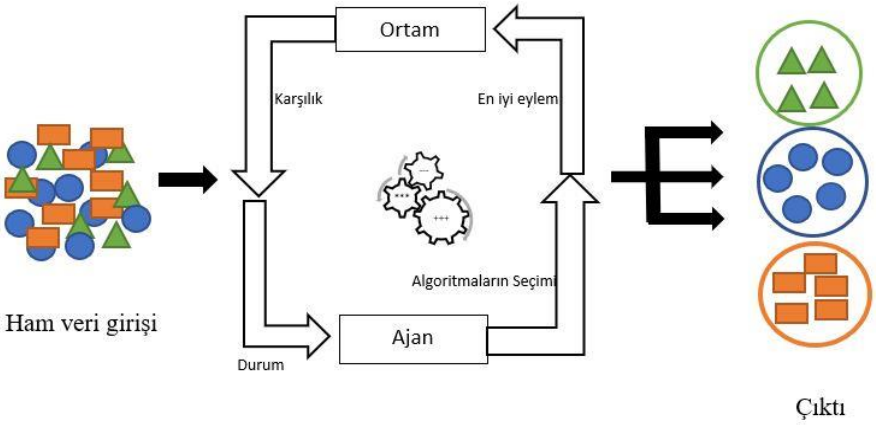
Yapı ve işleyiş olarak denetimli öğrenme ile benzer amacı olan yarı denetimli öğrenme metodunda, etiketli veri setleri birlikte genel olarak sayıca daha fazla olan etiketsiz veri setleri kullanılarak daha iyi bir model geliştirilmesi amaçlanır (Jackson, 1988).




**Şekil 2.11.** Yarı denetimli öğrenme

### 2.3.4 Takviyeli Öğrenme

Takviyeli öğrenme metodu ile kendi etkileşimlerinin sonuç değerlerini gözlemleyen bir sistemin önceki deneyimleri vasıtası ile öğrenme işlemini sağlaması amaçlanır (Maind ve Wankar, 2014). Takviyeli öğrenme algoritmaları denetimli öğrenmede olduğu gibi örnek veri seti üzerinden değil daha çok deneme yanılma tekniği ile öğrenme işlemini gerçekleştirir. Bu şekilde öğrenen sistem, verilerin analiz edilmesi ile aldığı bildirimler üzerinden kullanıcılara en doğru sonuca yönlendirme yapan davranışsal bir öğrenme tekniğidir. Verilerin analizi ile elde edilen geri bildirimlerin sayesinde yöntemini ödüllendirme veya cezalandırma olarak dereceleyen öğrenici, bir denge durumu oluşturmaya çaışır ve denge oluşana kadar parametrelerini ayarlar (Dongare, Kharde ve Kachare, 2012). Takviyeli öğrenme yöntemi ile öğrenme mantığı oluşturulan bir uygulama olarak Google tarafından 2017 yılında tasarlanan bir yapay zeka uygulaması olan AlphaGo, Go oyununda dünya şampiyonluğu bulunan Ke Jie'yi yenmiştir. Günümüzde Deep Learning algoritmaları da takviyeli öğrenme yöntemlerinin kullanılması ile geliştirilmiştir.




**Şekil 2.12.** Takviyeli öğrenme

## 2.4 Makine Öğrenmesinde Sınıflandırma Performans Ölçütleri

Makine öğrenmesi uygulamalarında karşılaşılan herhangi bir problemin çözümü için farklı modeller kurulabilmektedir. Bu modeller arasından en iyisinin seçimi için değerlendirme yapmakla mümkündür. Bunun için model

performansının değerlendirme yöntemleri ve değerlendirme ölçütleri geliştirilmiştir.

Problemin çözümü için oluşturulan modelin performansının bağlı olduğu birtakım parametreler vardır. Bunlar arasında öğrenme algoritması, eğitim ve test verilerinin büyüklüğü, sınıf dağılımı örnek olarak gösterilebilir (Raschka, 2018). Bu yöntemler arasında en sık olarak kullanılanları dışarda tutma (holdout), üçlü ayırma (three-way split) ve çapraz geçerleme (cross-validaiton) yöntemleri gösterilebilir.

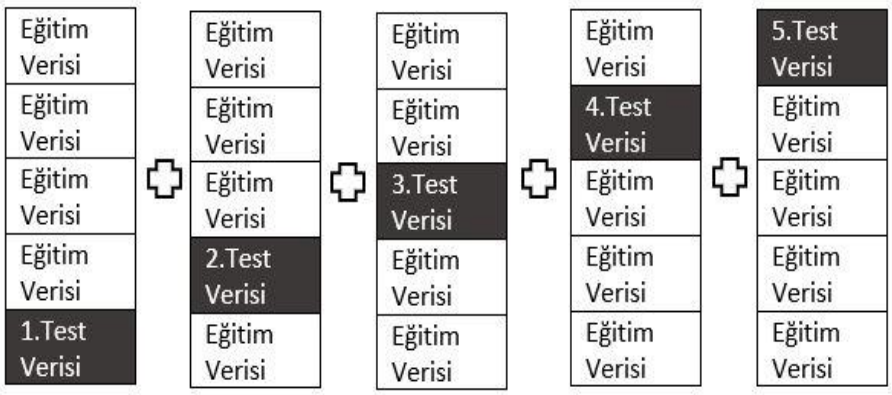
Dışarı tutma yönteminde veri seti eğitim ve test olarak iki bölüme ayrılarak oluşturulur. Öğrenme işlemi eğitim veri seti ile sağlanırken test veri seti ile öğrenme işleminin ne derece gerçekleştiğinin kontrolü sağlanmaktadır. Bu yöntemde test veri seti eğitim veri seti haricindeki verilerden oluşturulmaktadır. Gözlem sayısının düşük olması, eğitim ile test veri setlerinin sadece bir kez ayrılmaları bu yöntem açısından en dezavantajlı durumdur (Balaban ve Kartal, 2018).

Üçlü ayırma yönteminde, makine öğrenmesinde kullanılacak modelin seçilmesi ve performans tahmini eş zamanlı gerçekleşmektedir. Veri seti; eğitim, doğrulama ve test verisi olmak üzere üç sete ayrılır. Bu yöntemde doğrulama veri setindeki örnekler aracılığı ile öğrenme işleminde kullanılan algoritmaya ait parametrelerin ayarı yapılmaktadır. Modelin asli performansının değerlendirilmesi için test veri seti kullanılmaktadır (Balaban ve Kartal, 2018).

Çapraz geçerleme yönteminin ise en çok kullanılan iki türü vardır. Bunlar k-kat çapraz geçerleme (k-fold cross validation) ve birini dışarıda bırak çapraz geçerleme (leave one out cross validation) yöntemleridir. K-kat çapraz geçerleme yönteminde veri seti k eşit parçalara ayrılır, elde edilen her bir k paranın her biri bir kez test seti ve geri kalan k-1 parçası da eğitim seti olarak seçilmektedir. Bu şekilde k kere elde edilen sonuç performans ölçülerinin ortalaması alınarak nihai performans elde edilir. K-kat çapraz doğrulama ynteminde kullanılacakveri setinin kaç eşit parçaya bölüneceğini belirleyen k parametresinin öncelikle belirlenmesi gerekmektedir. Yapılan bazı çalışma ve araştırmalarda k değerimim 2,5 ile 10 olarak kullanılmakta olduğu görülmekle beraber çoğunlukla önerilen k değeri 10'dur. Diğer yöntem olan birini dışarıda bırak yönteminde ise veri

setinden sadece bir örnek her defa test seti olur, kalan n-1 örnek eğitim seti olarak kullanılmaktadır (Balaban ve Kartal, 2018).

Örnek olarak çapraz doğrulama değeri = 5 ise




**Şekil 2.13.** Çapraz doğrulama örneği

D.O = Doğruluk oranı

V.S. = Veri seti

$$Değerlendirme = \frac{1.V.S.D.O.+2.V.S.D.O.+3.V.S.D.O.+4.V.S.D.O.+5.V.S.D.O.}{5} \quad (2.12)$$

## 2.5 Makine Öğrenmesi Başarım Değerlendirmesi Yöntemleri

### 2.5.1 Doğruluk Oranı-Hata Oranı (Accuracy Rate-Error Rate)

Makine öğrenmesi sistemlerinde oluşturulan modelin başarımının ölçülmesi için en basit ve sık kullanılan yöntemlerin başında doğruluk oranı gelmektedir. Modelde oluşturulmuş olan doğru sınıflandırılmış örnek sayısının (TP+TN), modeldeki tüm örnek sayısına (TP+TN+FP+FN) oranı doğruluk oran sayısını vermektedir. Doğruluk oran değerini 1'e tamamlayan değer ise hata oranıdır. Başka bir şekilde ifade etmek gerekirse modelde oluşan yanlış sınıflandırılmış örnek sayısının (FP+FN), modeldeki tüm örnek sayısına (TP+TN+FP+FN) oranı hata oranını vermektedir (Nizam ve Akın, 2014).

Doğruluk oranı hesaplaması yapılırken aşağıdaki Confusion Matrix tablosunda yer alan bir sınıflandırma problemi için oluşturulan modelin gerçekleşen ve tahmin edilen değerleri üzerinden yapılır.

**Tablo 2.1.** Confusion Matrix

| | | Tahminlenen |
|---|---|---|
| Gerçekleşen | True Pozivities (TP) | False Negatives (FN) |
| | False Posivities (FP) | True Negatives (TN) |

TP: Gerçekte pozitif iken oluşturulan model tarafından da pozitif olarak sınıflandırılanlar.

FP: Gerçekte negatif ilen oluşturulan model tarafından da pozitif olarak sınıflandırılanlar.

TN: Gerçekte negatif iken model tarafından da negatif olarak sınıflandırılanlar.

FN: Gerçekte pozitif iken model tarafından negatif olarak sınıflandırılanlardır.

$$Doğruluk = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2.13)$$

$$Hata\ Oranı = \frac{(FP+FN)}{(TP+FP+FN+TN)} \quad (2.14)$$

### 2.5.2 Kesinlik (Precision)

Makine öğrenmesi sistemlerinde oluşturulan model ile sınıflandırma sonucunda sınıfı 1 olarak tahminlenmiş True Pozitif (TP) örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına (TP+FP) oranı kesinlik sayısını vermektedir. Kesinlik değeri False Positive tahminleme maliyetinin yüksek olduğu şartlarda daha fazla önem kazanır. Örnek olarak mail kutumuza gelmesi gereken mailleri eğer oluşturulan sınıflandırma modeli spam olarak etiketliyorsa (FP) bunun sonucu olarak almamız gereken önemli sayılabilecek maileri

görememiş oluruz ve bu durum kayıp oluşturur. Bundan dolayı Kesinlik değerinin yüksek olması sınıflandırma model seçiminde önemli bir yer almaktadır.

$$\text{Kesinlik} = \frac{TP}{(TP+FP)} \tag{2.15}$$

### 2.5.3 Duyarlılık (Recall)

Modelleme sonucunda doğru sınıflandırılmuş olan pozitif örnek sayısının (TP), toplamda elde edilen pozitif örnek sayılarının toplamına (TP+FN) oranı bize duyarlılık değerini verir. Başka bir deyişle gerçek pozitif sayısının ne kadarı doğru olarak tanımlandı sorunusunun cevabını verir.

$$\text{Duyarlılık} = \frac{TP}{(TP+FN)} \tag{2.16}$$

### 2.5.4 F-Ölçütü (F-Measure)

Kesinlik ve duyarlılık değerlerinin tek başlarına anlamlı ve yeterli bir model karşılaştıma sonucu elde etmemize yeterli olmadığı durumlardan dolayı, bu iki ölçütün birlikte değerlendirilmesi ile daha hassas ve doğru sonuçlar elde etmek mümkündür. F-ölçütü değeri kesinlik (K) ve duyarlılık (D) değerlerinin harmonik ortalaması alınarak hesaplanır (Nizam ve Akın, 2014). Harmonik ortalama ile sistem için oluşabilecek en uç durumların gözardı edilmemesi adına yapılmaktadır. Örnek olarak kesinlik değeri 1, duyarlılık değeri 0 olan bir modelin F-ölçüt değeri 0,5 olacaktır ve bu sistem için hassas bir değerlendirme yapılmasına engel oluşturabilecektir. F-ölçütünün doğruluk değeri yerine tercih edilmesinin en önemli sebeplerinden biri de eşit dağılım göstermeyen veri setlerinde hatalı bir model seçimi yapmamaktandır. Bununla beraber model değerlendirmesinde False Negative ve False Positive beraber tüm hata maliyetlerinin dahil olduğu değerlendirmeye ihtiyaç duyulduğundan F-Ölçütü değerlendirme için önemli bir yere sahiptir.

$$\text{F} = \frac{2*D*K}{(D+K)} \tag{2.17}$$

### 2.5.5 Kappa İstatiği

İki veya daha fazla sayıdaki gözlemcinin aynı modeli değerlendirdiği durumlarda gözlemciler arasındaki uyumun güvenilirliğini ölçmeye yarayan bir yöntemdir (Congalton ve Green, 1998). (-1,1) aralığında kappa katsayısı değişiklik göstermektedir. K=1 değeri tam uyumun söz konusu olduğu durumda gerçekleşir. Eğer ki gözlenen uyumun şansa bağlı uyuma eşit veya daha büyük olması durumunda K≥0; gözlenen uyumun şansa bağlı uyumdan daha küçük olması durumunda ise K<0 olmaktadır. Kappa katsayısı için yorumlanabilir aralık 0 ile +1 arasındadır ve negatif (K<0) değerlerinin güvenilirlik açısından anlam ifade etmemektedir (Landis ve Koch, 1977). Kappa değeri şu şekilde hesaplanmaktadır.

$$K= \frac{(P_0-P_c)}{(1-P_c)} \tag{2.18}$$

Po: Kabul edilen oran,

Pc: Beklenen oran

### 2.5.6 Ortalama Mutlak Yüzde Hatası (MAPE)

Oluşturulan model üzerine, gerçek değerler ile tahmin edilmiş değerler arasındaki hatanın yüzdelik olarak gösterildiği hata metriğidir. Mape performans metriği denklemi aşağıdaki gibidir:

$$MAPE= \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-y_i'}{y_i}\right| \tag{2.19}$$

$y_i$: Gerçek değer

$y_i'$: Tahmin edilen değer

n: Gözlem sayısı

Zaman serileri ve regresyon modellerinde doğruluk ölçeklendirmesi için ortalama mutlak hata sıklıkla kullanılmaktadır. Hata değerlerinin birim değerlerinin farklılık gösterdiği, örnek olarak bir tahmin modeli gerçek değerleri kullanırken diğer tahmin modeli doğal logaritması alınmış değerleri kullandığı durumlarda Mape yararlanılacak metriklerin başında gelmektedir. Gerçek değerler arasında "0" değeri varsa, bölünme işlemi 0 ile mümkün olmayacağı için MAPE'nin hesaplanması bu durumlarda mümkün değildir. Tahmin değerleri çok

düşük olan durumlarda yüzdelik hata %100'ü geçemez ancak gerçek değerlerden uzak tahminleme değerleri olduğu durumlarda yüzde hatasının üst sınırı yoktur (Chai & Draxler, 2014).

### 2.5.7 Ortalama Mutlak Hata (MAE)

Yorumu kolay olduğundan dolayı regresyon ve zaman serileri problemlerinin değerlendirilmesinde sıklıkla kullanılan, bir dizi tahmindeki hataların ortalama büyüklüğünün ölçülmesini sağlar. Mae performans metriği denklemi aşağıdaki gibidir.

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'| \tag{2.20}$$

$y_i$: Gerçek değer

$y_i'$: Tahmin edilen değer

n: Gözlem sayısı

Gerçekleşen ve tahmin edilen olmak üzere iki sürekli değişken arasındaki farkın ölçüsü olan MAE, bu değerler arasındaki farkın mutlak değerlerinin toplamıdır. Başka bir deyişle ortalama mutlak hata, tahmin edilen ile gözlemlenen arasındaki farkların mutlak değerlerinin doğrulama örneğinin ortalamasını göstermektedir.

### 2.5.8 Ortalama Karekök Sapması (RMSE)

Makine öğrenmesi modelinde, tahminleyicinin oluşturduğu tahmin değerleri ile gerçek değerleri arasındaki uzaklığı bulmaya yarayan, meydana gelen hatanın büyüklüğünü ölçen ikinci dereceden bir puanlama kuralıdır. RMSE denklemi aşağıdaki gibidir.

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \tag{2.21}$$

$y_i$: Gerçek değer

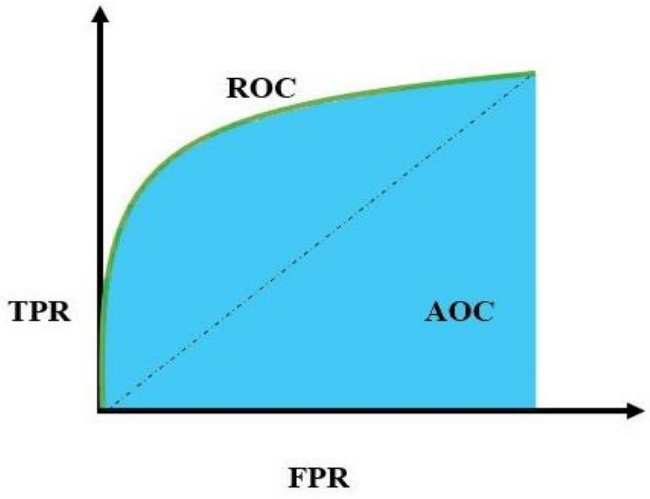
$y_i'$: Tahmin edilen değer

n: Gözlem sayısı

### 2.5.9 ROC/AUC Eğrisi

Makine öğrenmesi sınıflandırma problemlerinin peformans ölçümlerinde kullanılan bir diğer ve önemli ölçümlerden biri de Roc eğrisidir. Roc eğrisi olasılık eğrisi olarak tanımlanabilir ve bu eğrinin altında kalan alan olan Auc, ayrılabilirliğin ölçüsünü ve bununla birlikte derecesini de temsil etmektedir.

Roc eğrisini oluşturan eksenlerden x ekseni FPR (Yanlış Pozitif Oran) ve Y ekseninde TPR (Gerçek Pozitif Oran) yer almaktadır.




**Şekil 2.14.** Roc-Auc eğrisi

Roc eğrisinin Y eksenine yakınlaşması ve bunun sonucu olarak eğri altında kalan alanın artması ile sınıflar arasındaki ayırt etme performansı armaktadır.

### 2.5.10 Korelasyon Katsayısı (R)

Korelasyon katsayısı yöntemi, değişkenlerin birbirleri arasındaki ilişki ve bu ilişkiye dair şiddet ve hedefi ile ilgili bilgi vermeye yarayan istatistiksel bir yöntemdir. Bağımlı değişken ile bağımsız değişkenlerin birbirleri ile olan ilişkilerinin ne derece güçlü olduğunu gösteren bir katsayıdır. Bununla beraber korelasyon katsayısı dğeişkenlerin yönü ve bu değişkenlerin etkileşimlerinin nasıl olduğu konusunda bilgi vermektedir.

İki değişken arasındaki doğrusal olarak ifade edeceğimiz ilişkinin ölçüsü olan korelasyon katsayısı -1<R<1 arasında değerlerdedir. Değerin 0'a doğru yakınlaşması ilişkinin zayıflığını, 1'e doğru yaklaşması da güçlülüğünü gösterir. Değişkenler arasındaki ilişkinin pozitif yönde ilişki olması, değişkenlerin beraber artması ya da azalması; negatif yönde ilişki oluşması da değişkenlerin ters yönlü artıp azalması ile oluşmaktadır.

Değişkenlerden birinde tespit edilen değişikliğin diğer değişken tarafından ne kadarının açıklandığının yorumlaması için kullanılan ve korelasyon katsayısının karesine eşit olan $\boldsymbol{R^2}$ determinasyon katsayısı olarak ifade edilir. $\boldsymbol{R^2}$'nin 1 olması deneysel verilerin hatasız bir doğrusal eğri meydana getirdiğini göstermektedir. Örnek olarak $R^2$ =0.80 olduğunda, Y değişkenindeki toplam varyasyonun %80'i aıklanabilir, %20'si açıklanamaz.

### 2.5.11 Ortalama Kare Hatası (MSE)

Değişkenlerin gerçek değerleri ile model tarafından tahmin edilen değeri arasındaki meydana gelen farkın karesinin ortalamasının alınması ile elde edilen sonuçtur. Formüle edilmiş hali aşağıdaki gibidir.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2 \qquad (2.22)$$

$y_i$: Gerçek değer

$y_i'$: Tahmin edilen değer

n: Gözlem sayısı

# 3. MATERYAL VE YÖNTEM

## 3.1 Veriler ve Toplanması

Bu tez çalışmasında kullanmak amacıyla öğrencilerin kişisel bilgileri saklı tutularak, Bolu Abant İzzet Baysal Üniversitesi Bilgi İşlem Daire Başkanlığı veri tabanından elde edilmiştir. Veriler, Bolu Abant İzzet Baysal Üniversitesi, Kamu Yönetimi bölümünden 2011 yılı sonrasında mezun olan öğrencilerin dersleri ve yılsonu mezuniyet notlarından meydana gelmektedir. Veri kümesi meydana getirilirken 832 öğrencinin mezuniyet notları ele alınmıştır. Bir dersin sınavları 100 üzerinden değerlendirilirken, her bir derse ait ders notu sistemde harf ile gösterilir. Harfli notlar aşağıdaki tablodaki gibi dönüşüm işlemi uygulanarak derslere karşılık gelecek şekilde hesaplanır.

**Tablo 3.1.** Not dönüşüm tablosu

| Ders Notu | Başarı Katsayısı |
|---|---|
| AA | 4 |
| BA | 3.5 |
| BB | 3 |
| CB | 2.5 |
| CC | 2 |
| DC | 1.5 |
| DD | 1 |
| FF | 0 |

## 3.2 Uygulama

Bu tez çalışmasında, öğrencilerin eğitim süreçleri sonundaki mezuniyet not ortalamalarını tahmin edecek bir veri madenciliği uygulamasının makine öğrenmesi yöntemleri ile oluşturulması ve karşılaştırılmasıdır. Böylece ortalaması yeterli olmayan ya da mezun durumda olamayacak öğrenciler için erken uyarı niteliği taşıyacaktır. Bu bağlamda …. Oluşturulan veri seti kullanılarak iki ayrı senaryo hazırlanarak Anaconda Navigator Smile uygulaması ortamında uygulamalar gerçekleştirilmiştir. İlk oluşturulan senaryoda öğrencilerin ilk yıl aldıkları derslerin yılsonu notları değerlendirilmiştir. Böylece 15 adet derse ait

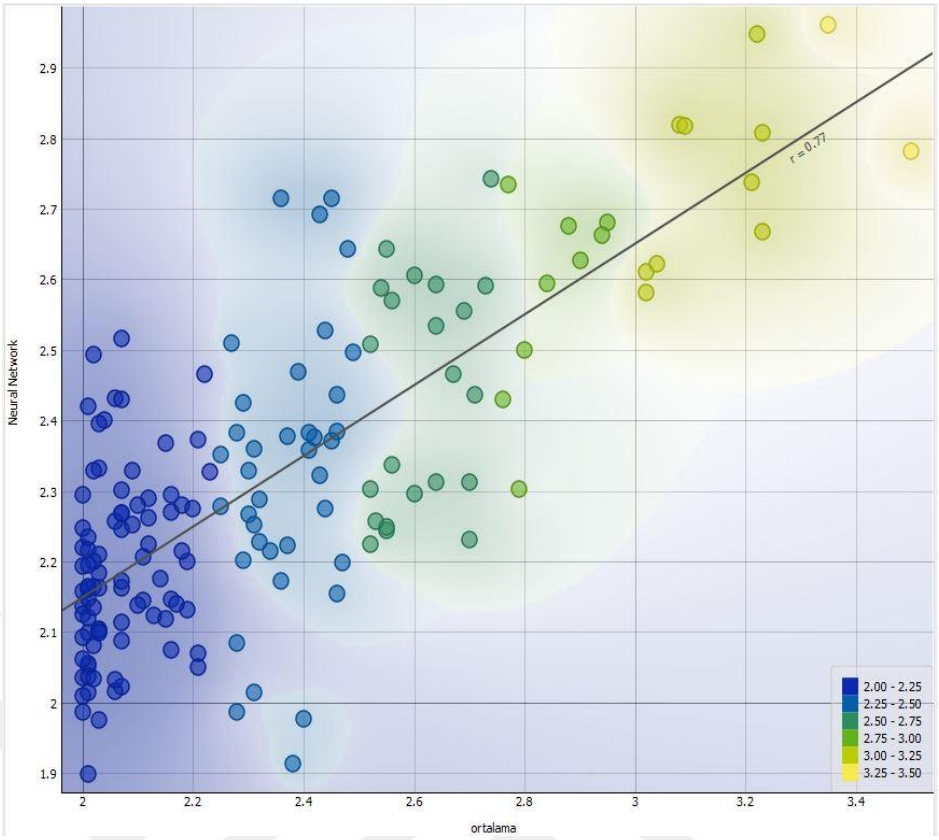
yılsonu notları ile öğrencilerin mezuniyet notlarının tahmin edilmesi test edilmiştir. İkinci hazırlanan senaryoda öğrencilere ait ilk iki yıl sonundaki aldıkları derslerin yılsonu notları değerlendirme için kullanılarak bu öğrencilere ait mezuniyet notlarının tahmin işlemi için test yapılmıştır. Öğrenciler ilk iki yıl sonunda toplamda 31 ders alarak tamamlamışlardır.

Daha önce de bahsettiğimiz gibi bu tez çalışmasında YSA, KA, DVM, KNN ve LR yöntemleri kullanılarak öğrencilerin mezuniyet not tahmin işlemi gerçekleştirilmiştir.YSA ile hazırladığımız modellerde logistic aktivasyon fonksiyonu ile gizli katmanda 30 adet nörondan oluşturulmuştur. Her iki senaryo için de modelin çıkış katmanında tek hücre yer almaktadır. Adam aktivasyon fonksiyonu çıkış katmanında kullanılmıştır. Hazırlanan veri setinin %80'i eğitim, %20'si test için kullanılmıştır. Böylece eğitim için 666 örnek ve test aşaması için 166 adet örnek kullanılmıştır. Aşağıdaki tablo 3.1'de eğitim-test sonucunda elde edilen sonuçlar verilmiştir.

**Tablo 3.2.** YSA ile birinci senaryo sonucunda elde edilen başarım değerleri

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.049 | 0.221 | 0.173 | 0.574 |

Tablo 3.2'de görüleceği üzere ilk senaryo sonucunda 0,049 MSE, 0.221 RMSE, 0.173 MAE ve 0.574 R2 değerleri elde edilmiştir. YSA modelimiz 200 iterasyon sonunda bu değerleri elde etmiştir.




**Şekil 3.1.** Birinci senaryo YSA ile mezuniyet not ortalaması dağılım grafiği

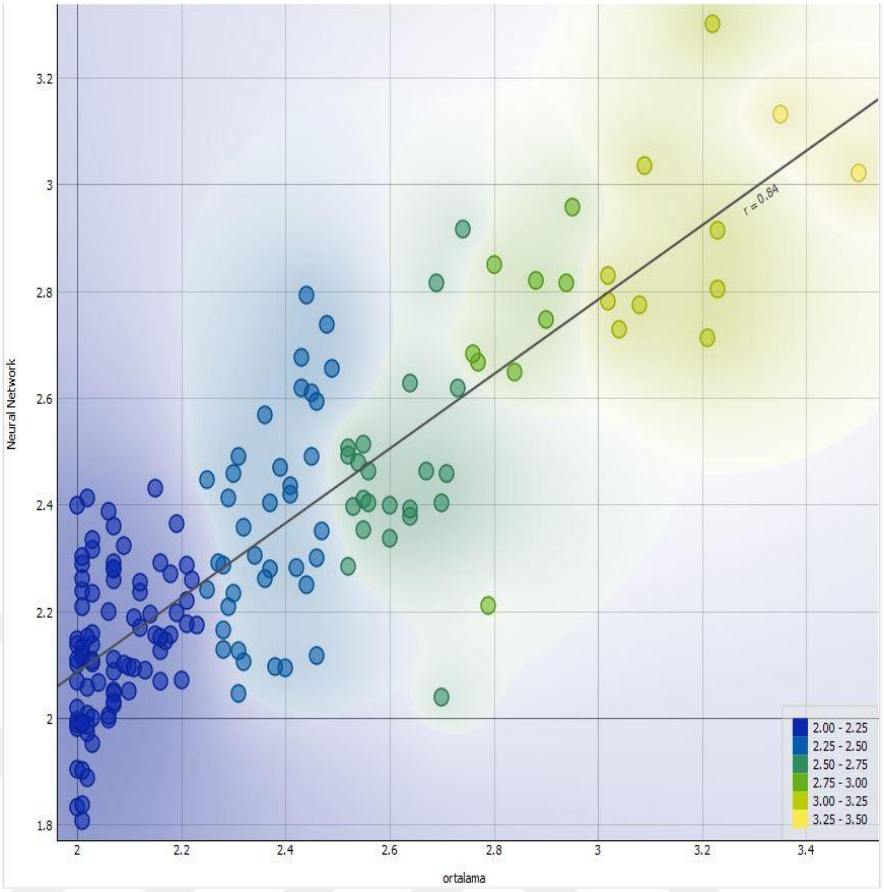
Şekil 3.1'de oluşturduğumuz YSA modelinin 132 örnek üzerinden elde ettiği tahmin sonuçları ve gerçek değerler gösterilmektedir. Korelasyon R=0,77 değeri ile ilk senaryodaki en yüksek sonuca YSA ile ulaşılmıştır.

Hazırlanan ikinci senaryo için oluşturulan başarım sonuçları Tablo 3.3'de yer almaktadır. 0.034 MSE, 0.184 RMSE, 0.140 MAE ve 0,705 R2 değerleri elde edilmiştir. Elde edilen bu değerler modellemenin ilk senaryoya göre daha başarılı ve gerçeğe yakın olduğunu ifade etmektedir.

**Tablo 3.3.** YSA ile ikinci senaryo sonucunda elde edilen başarım değerleri

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.034 | 0.184 | 0.140 | 0.705 |

İkinci senaryo için YSA, 96 iterasyon ile bu sonuca ulaşmıştır. Şekil 3.2'de YSA'nın gerçek ve tahmin değerlerini göstermektedir. Şekil 3.2 incelendiğinde ikinci senaryoda elde edilen tahminlerin daha yüksek oranda olduğu görülmektedir



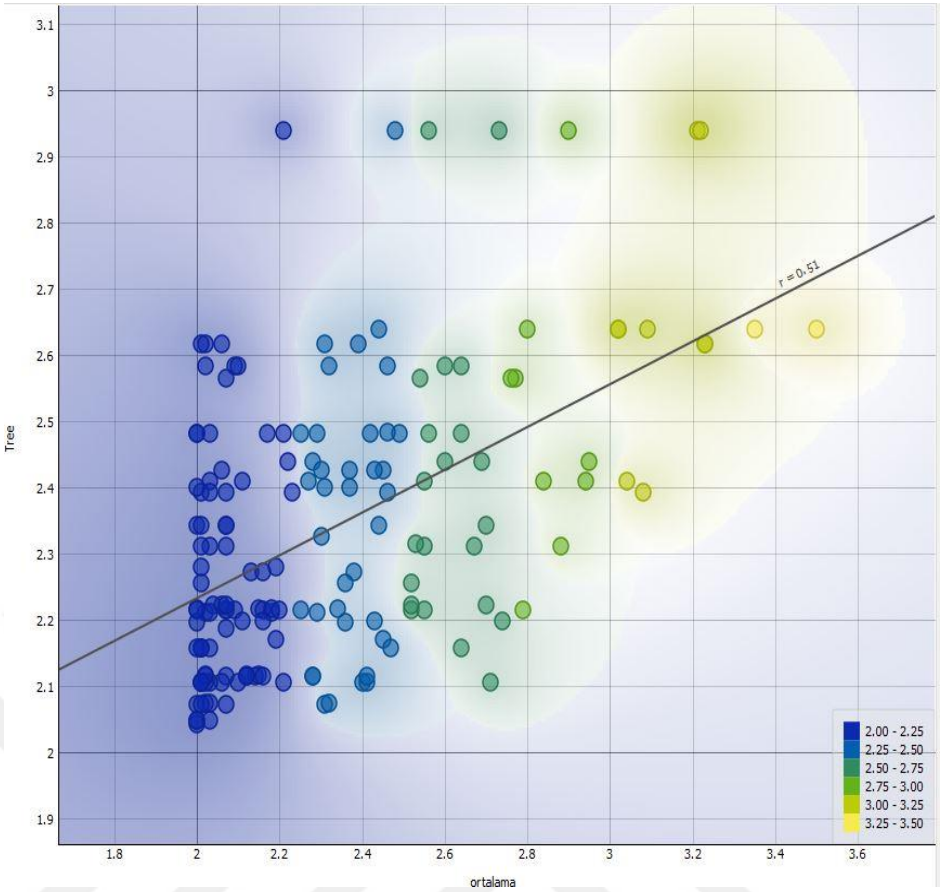

**Şekil 3.2.** İkinci senaryo YSA ile mezuniyet not ortalaması dağılım grafiği

Karar ağaçları ile uygulanan regresyon işlemi için oluşturulan iki senaryo için de aynı ağaç yapısı ile çalışılmıştır. Karar ağaçları modeli için yapraklardaki minimum örnek sayısı 14, 5'ten küçük alt kümelerin bölünmemesi ve maksimum ağaç derinliği 10 ile sınırlayacak şekilde modelleme oluşturulmuştur.

**Tablo 3.4.** KA ile birinci senaryo sonucunda elde edilen başarım değerleri

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.087 | 0.294 | 0.229 | 0.244 |

Tablo 3.4'te KA ile oluşturulan modelin başarım sonuçları 0.087 MSE, 0.294 RMSE, 0.229 MAE ve 0,244 R2 değerleri elde edilmiştir. Bununla beraber Şekil 3.3'te KA ile oluşturulan modele ait tahmin ve gerçek değerlerin karşılaştırılması gösterilmiştir.



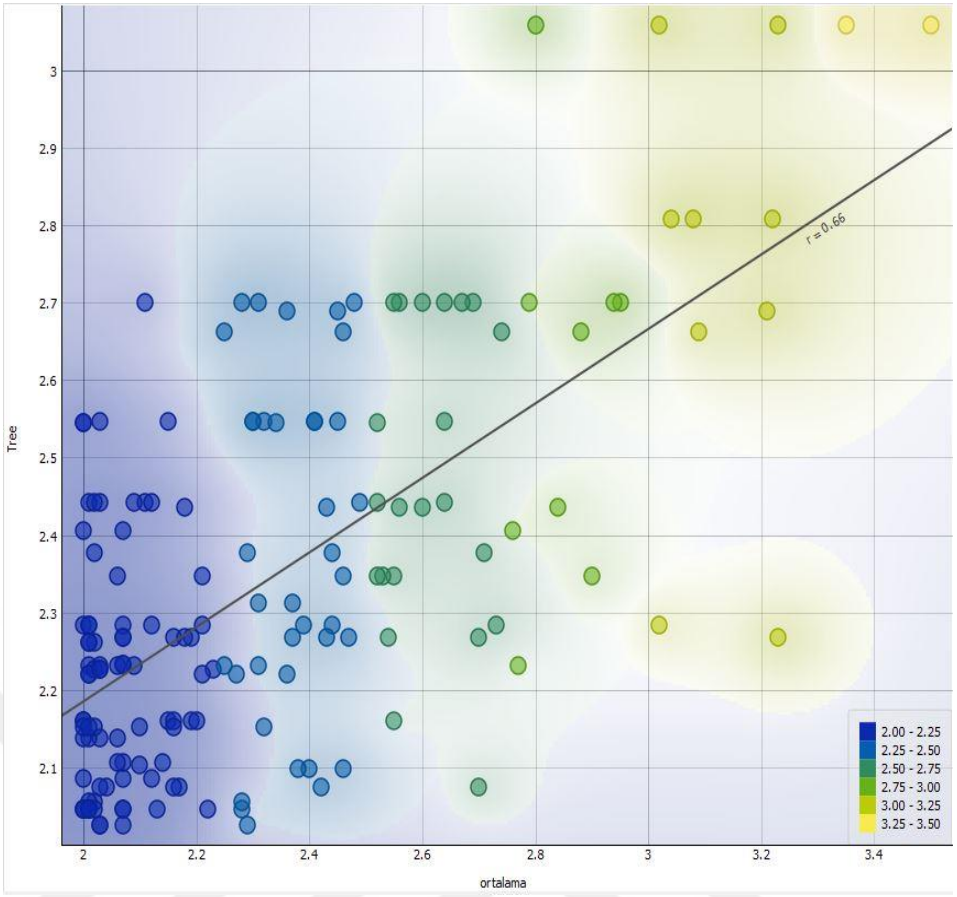

**Şekil 3.3.** Birinci senaryo KA ile mezuniyet not ortalaması dağılım grafiği

Tablo 3.5'te görüleceği üzere düzenlenen ikinci senaryo için KA modeline ait başarım sonuçları 0.066 MSE, 0.256 RMSE, 0.229 198 ve 0,426 R2 değerleri elde edilmiştir.

**Tablo 3.5.** KA ile ikinci senaryo sonucunda elde edilen başarım değerleri

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.066 | 0.256 | 0.198 | 0.426 |

Şekil 3.4'te KA modelinin hazırlanan ikinci senaryo için tahmin ve gerçek değerlerin karşılaştırıldığı değerler gösterilmektedir.




**Şekil 3.4.** İkinci senaryo KA ile mezuniyet not ortalaması dağılım grafiği

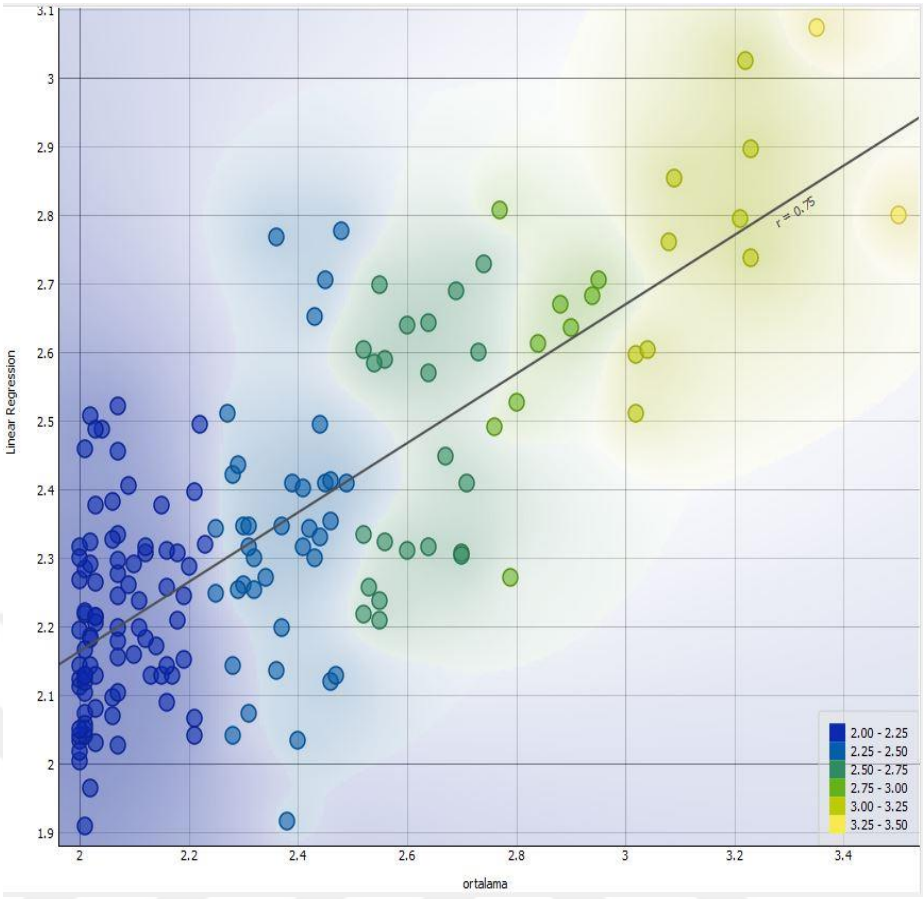
Şekil 3.4'te görüleceği üzere öznitelik artışı ile birlikte korelasyon değerindeki artış ile tahminlemedeki değerlerin gerçek değerlere yakınlığı artmıştır.

Lineer Regresyon modelimiz oluşturulurken her iki senaryo için de yapılan denemeler sonucunda Ridge, Lasso ya da Elastic net regresyon türleri arasından en iyi sonucu Ridge regresyon ile alınmıştır. Tablo 3.6'da görüleceği üzere oluşturulan ilk senaryo için oluşturulan modelde 0.051 MSE, 0.226 RMSE, 0.178 MAE ve 0.556 R2 değerleri elde edilmiştir.

**Tablo 3.6.** LR birinci senaryo metrik değerleri tablosu

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.051 | 0.226 | 0.178 | 0.556 |

Aşağıdaki Şekil 3.5'te ilk senaryo için LR modeli sonucunda meydana gelen mezuniyet notları ve tahminlere ait dağılım grafiği birlikte verilmiştir.



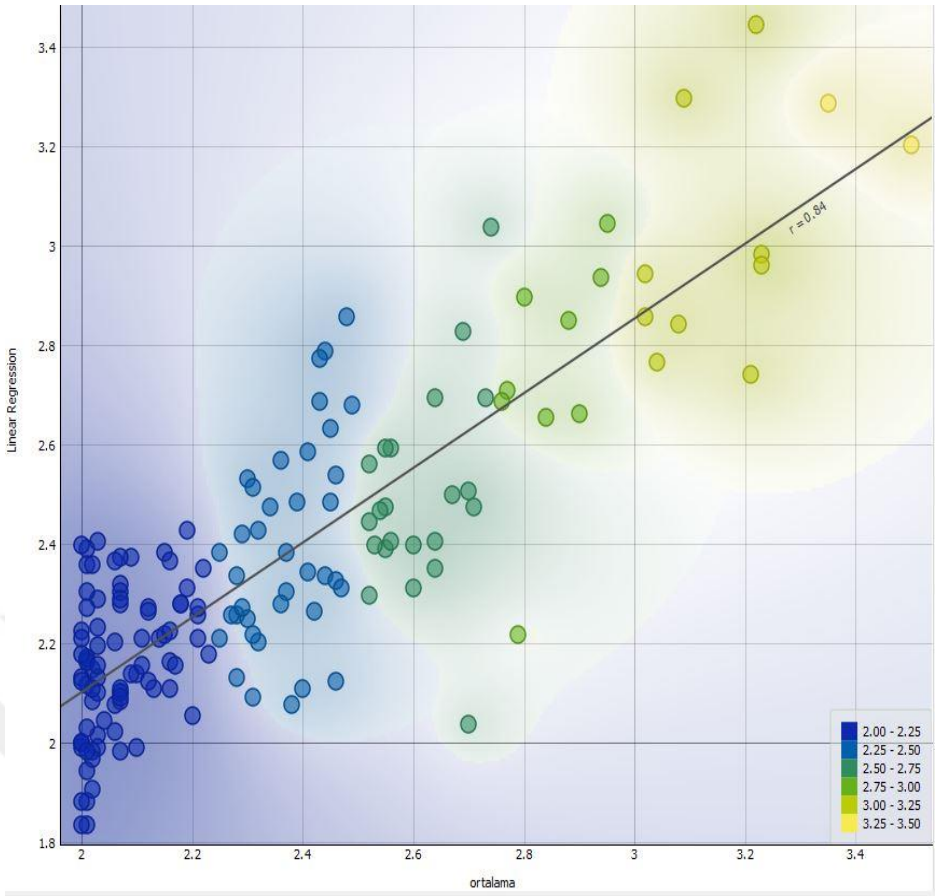

**Şekil 3.5.** Birinci senaryo LR ile mezuniyet not ortalaması dağılım grafiği

Lineer Regresyon modeli ile ikinci senaryo için elde edilen başarım değerleri Tablo 3.7'de de görüleceği üzere sırası ile 0.034 MSE, 0.186 RMSE, 0.146 MAE ve 0.699 R2 olarak elde edilmiştir. Öznitelik sayısı artışı ile birlikte R2 değerinde belirgin olarak artış gözlenmiştir.

**Tablo 3.7.** LR ikinci senaryo metrik değerleri tablosu

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.034 | 0.186 | 0.146 | 0.699 |

Şekil 3.6'da ikinci senaryoda LR modelimiz ile elde edilen tahmin ve gerçek mezuniyet notlarının yer aldığı dağılım grafiği verilmiştir.



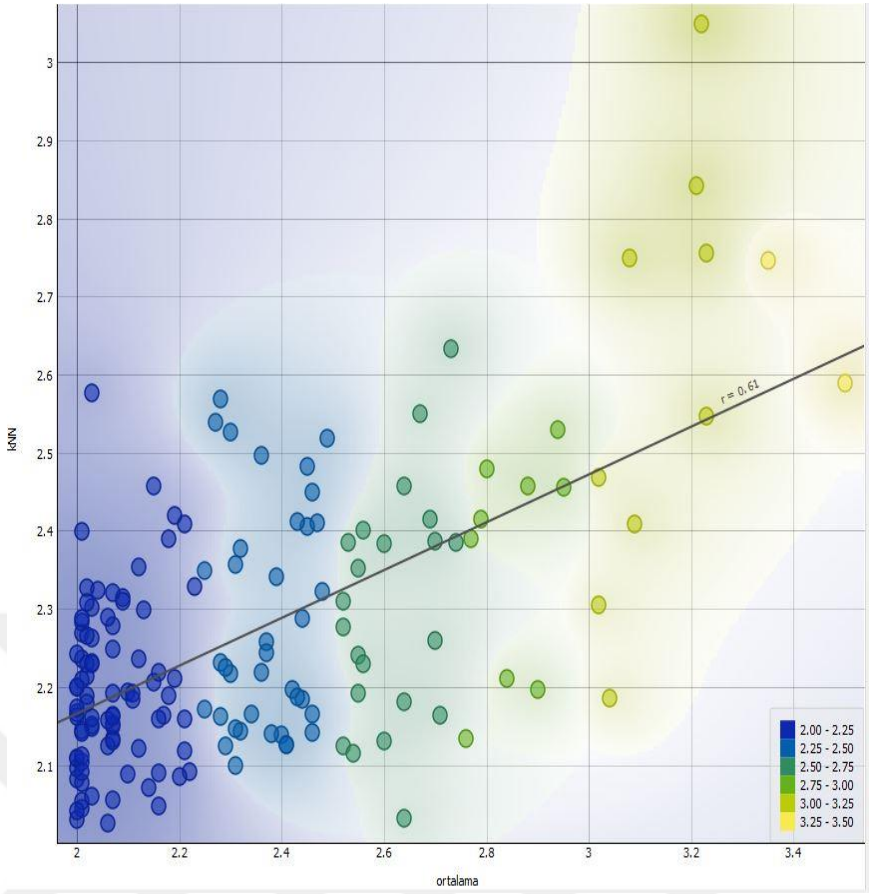

**Şekil 3.6.** İkinci senaryo LR ile mezuniyet not ortalaması dağılım grafiği

Mezuniyet not ortalaması tahmini için K-NN algoritması ile modelleme yaparken komşuluk sayısı 4 seçilirken, metrik olarak Manhattan mesafesi esas alınmıştır. Manhattan mesafesi iki vektörün mutlak olarak faklarının toplamıdır. Yani iki nokta (X1, Y1) ve (X2, Y2) olarak verildiğinde Manhattan mesafesi $|X1-X2|+|Y1-Y2|$ olarak hesaplanmaktadır. Birinci senaryo için K-NN algoritması ile elde edilen başarım değerleri aşağıdaki Tablo 3.8'de görüleceği üzere 0.077 MSE, 0.277 RMSE, 0.215 MAE ve 0.329 R2 olarak oluşmuştur.

**Tablo 3.8**. K-NN birinci senaryo metrik değerleri tablosu

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.077 | 0.277 | 0.215 | 0.329 |

Birinci senaryo ile elde edilen K-NN algoritması tahmin sonuçları ve mezuniyet not ortalamalarının grafiksel dağılımı Şekil 3.7'de gösterilmiştir.



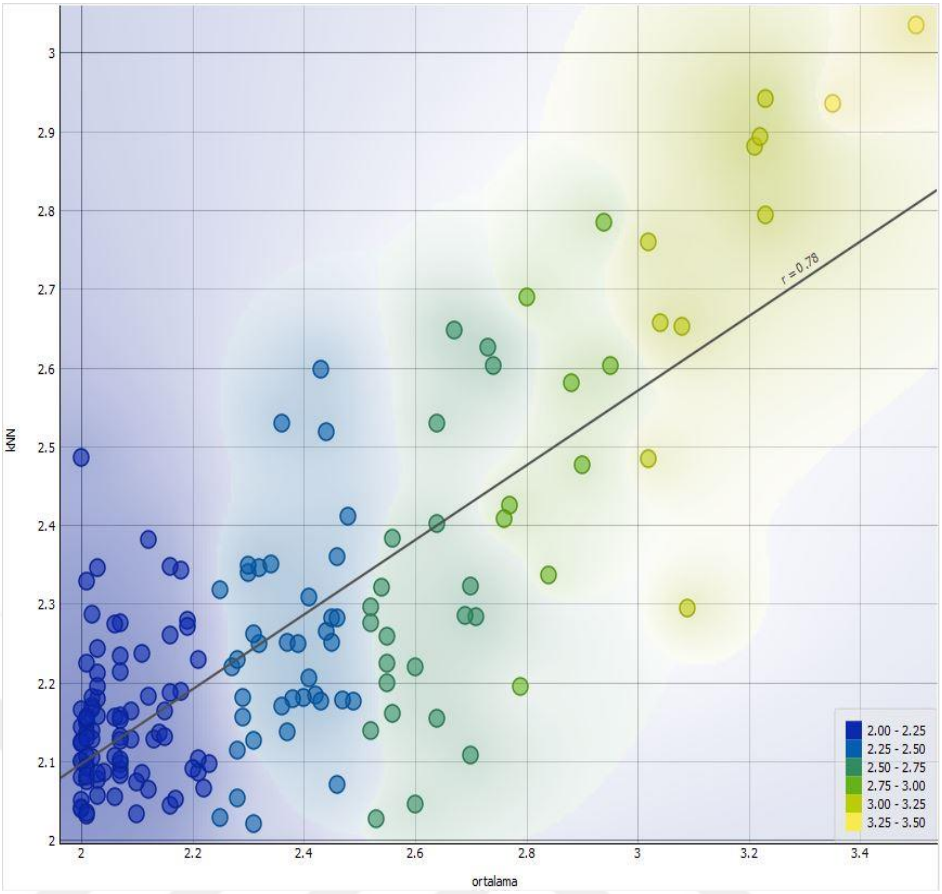

**Şekil 3.7.** Birinci senaryo K-NN ile mezuniyet not ortalaması dağılım grafiği

İkinci senaryomuzda giriş değerlerinin artması ile K-NN modelimizde de değerlerdeki değişim gözle görülür şekilde görülmektedir. Başarım değerleri sırası ile 0.054 MSE, 0.231 RMSE, 0.180 MAE ve 0.532 R2 olarak gerçekleşmiştir.

**Tablo 3.9.** KNN ikinci senaryo metrik değerleri tablosu

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.054 | 0.231 | 0.180 | 0.532 |

K-NN algoritması ile elde edilen tahmin sonuçları ve mezuniyet not ortalaması karşılaştırılmasının grafiksel gösterimi aşağıdaki Şekil 3.8'de yer almaktadır.



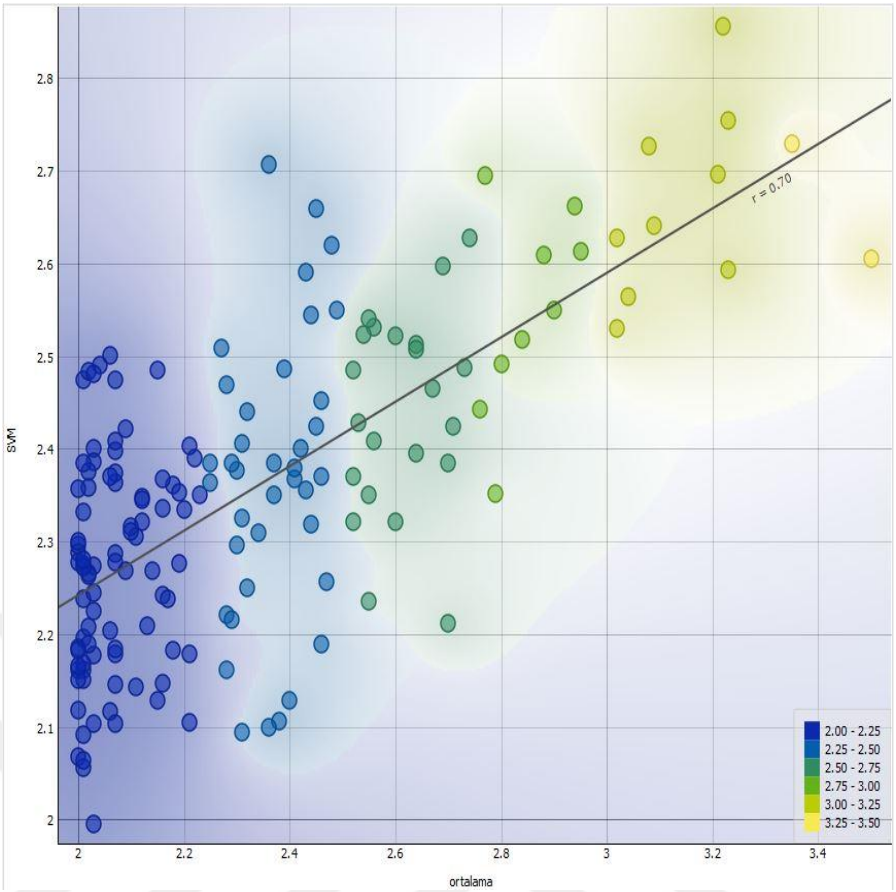

**Şekil 3.8.** İkinci senaryo K-NN ile mezuniyet not ortalaması dağılım grafiği

Çalışmamızda son olarak kullanılan DVM algoritması yöntemi için Cost değeri 1 ve Regression loss epsilon değeri 0,10 seçilip Kernel çekirdek yönteminde Radial Basis Function (RBF) seçilmiştir. Optimizasyon parametre değerleri seçilirken numerik tolerans değeri 0,0010 ve iterasyon limit değeri 100 olarak seçilmiştir. Yapılan çalışma ile ilk senaryo için başarım sonuçları 0.064 MSE, 0.253 RMSE, 0.207 MAE ve 0.442 R2 olarak gerçekleşmiştir.

**Tablo 3.10.** DVM birinci senaryo metrik değerleri tablosu

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.064 | 0.253 | 0.207 | 0.442 |

Birinci senaryo sonunda elde edilen DVM tahmin değerleri ve mezuniyet not ortalaması arasındaki dağılım aşağıdaki şekilde gösterilmiştir.



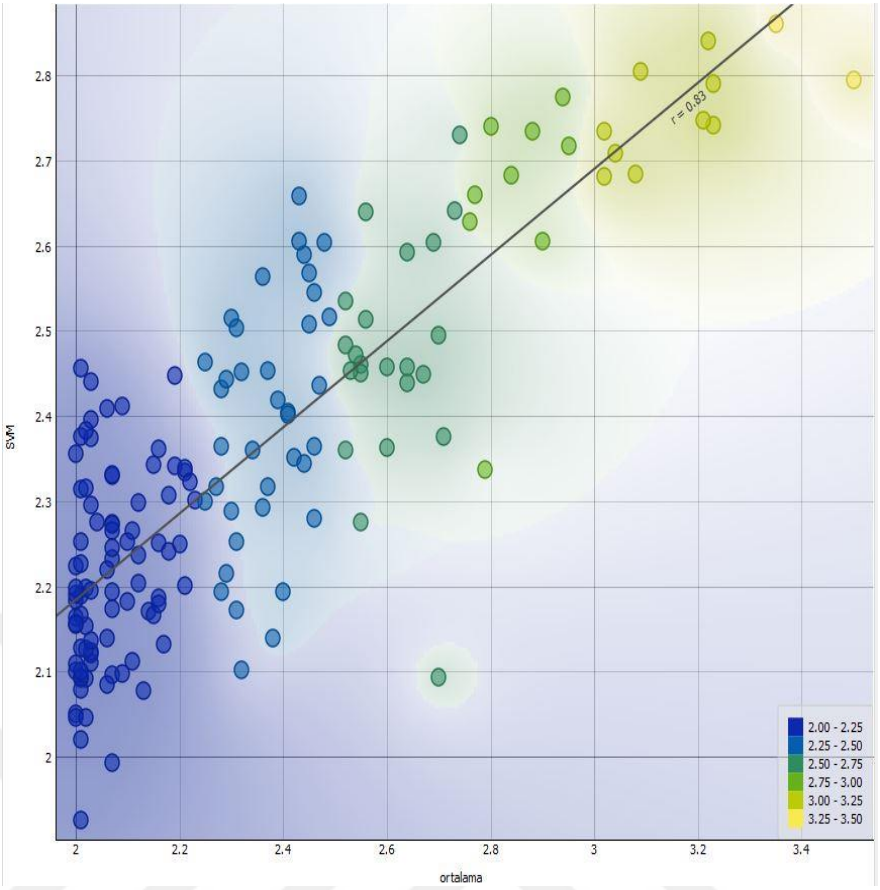

**Şekil 3.9.** Birinci senaryo DVM ile mezuniyet not ortalaması dağılım grafiği

İkinci senaryo için DVM algoritması ile oluşturulan modelimize ait başarım sonuçları 0.042 MSE, 0.204 RMSE, 0.162 MAE ve 0.635 R2 olarak gerçekleşmiştir. Öznitelik sayısının artması ile değerlerdeki olumlu yönde değişim görülmektedir.

**Tablo 3.11.** DVM ikinci senaryo metrik değerleri tablosu

| MSE | RMSE | MAE | R2 |
|---|---|---|---|
| 0.042 | 0.204 | 0.162 | 0.635 |

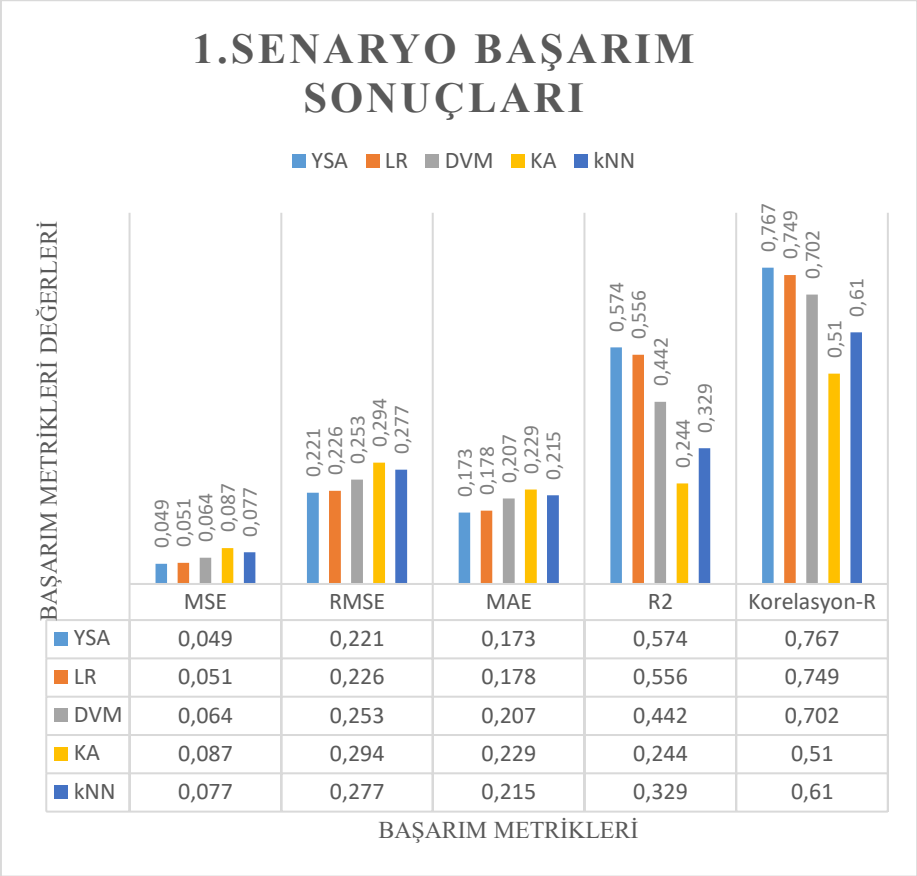
İkinci senaryo ile elde edilen DVM modeline ait tahmin sonuçlarının mezuniyet not ortalaması ile karşılaştırılması aşağıdaki Şekil 3.10'da gösterilmiştir.




**Şekil 3.10.** İkinci senaryo DVM ile mezuniyet not ortalaması dağılım grafiği

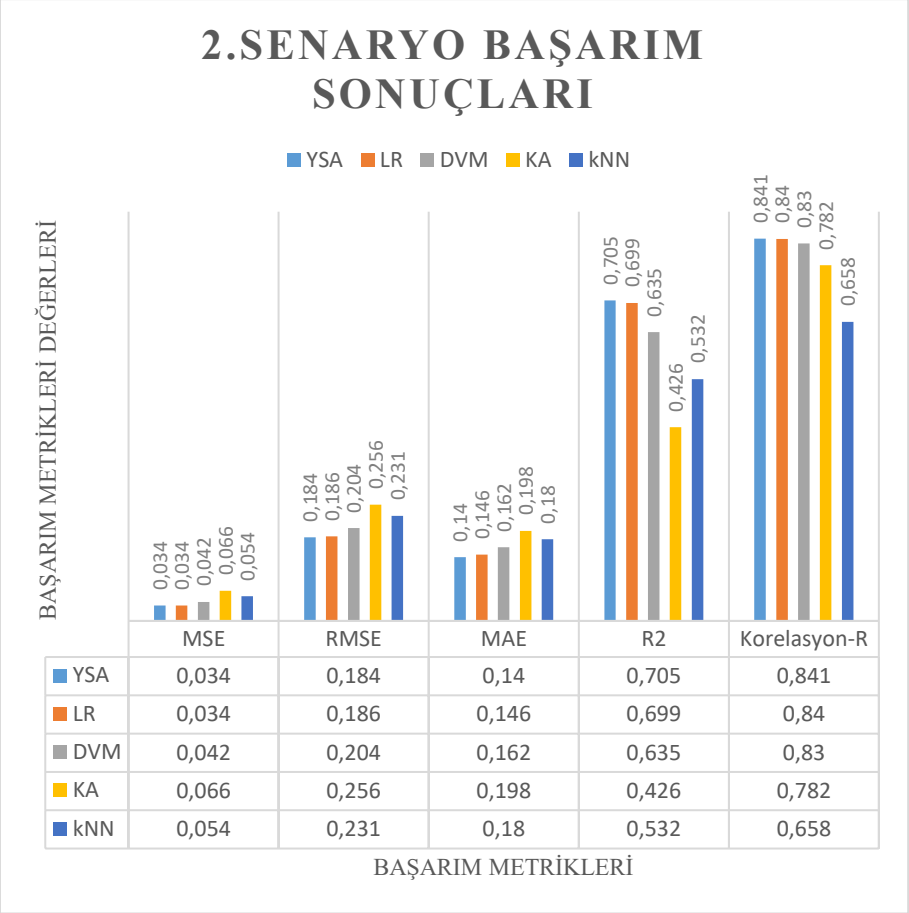
# 4. BULGULAR

Çalışmamızda uyguladığımız beş farklı yöntem ve iki farklı senaryo için elde edilen başarım değerleri grafikleri Şekil 4.1 ve Şekil 4.2'deki gibidir.




| | MSE | RMSE | MAE | R2 | Korelasyon-R |
|---|---|---|---|---|---|
| YSA | 0,049 | 0,221 | 0,173 | 0,574 | 0,767 |
| LR | 0,051 | 0,226 | 0,178 | 0,556 | 0,749 |
| DVM | 0,064 | 0,253 | 0,207 | 0,442 | 0,702 |
| KA | 0,087 | 0,294 | 0,229 | 0,244 | 0,51 |
| kNN | 0,077 | 0,277 | 0,215 | 0,329 | 0,61 |

**Şekil 4.1.** Birinci senaryo başarım metrikleri sonuçları grafiği




| | MSE | RMSE | MAE | R2 | Korelasyon-R |
|---|---|---|---|---|---|
| YSA | 0,034 | 0,184 | 0,14 | 0,705 | 0,841 |
| LR | 0,034 | 0,186 | 0,146 | 0,699 | 0,84 |
| DVM | 0,042 | 0,204 | 0,162 | 0,635 | 0,83 |
| KA | 0,066 | 0,256 | 0,198 | 0,426 | 0,782 |
| kNN | 0,054 | 0,231 | 0,18 | 0,532 | 0,658 |

**Şekil 4.2.** İkinci senaryo başarım metrikleri sonuçları grafiği

Makine öğrenmesi yöntemleri içinde ilk senaryo için R2 değerleri açısından KA, K-NN ve DVM başarı oranı çok düşük olmakla birlikte en başarılı olanlar LR ve YSA olmaktadır. MAE, MSE, RMSE değerleri açısından incelendiğinde 0'a en yakın elde edilen sonuçlar yine YSA ile gerçekleşmiştir. Bağımlı ve bağımsız değişkenlerin arasındaki ilişkiyi gösteren korelasyon R değeri en yüksek model yine YSA'dır.

Öznitelik sayısının arttığı ikinci senaryoda her beş model için de değerlerde gözle görülür iyileşme olmakla birlikte, R2 değerleri göz önüne alındığında YSA en iyi sonuca ulaşırken, K-NN sonuncu sırada yer almaktadır. MAE, MSE, RMSE değerleri açısından modeller karşılaştırıldığında yine en iyi sonuçlar YSA modeline ait olduğu görülmektedir. Korelasyon R değeri her bir model için artış gösterdiği ikinci senaryoda en yüksek değere YSA ulaşırken, LR modeli yine en başarılı ikinci sonucu elde etmiştir.

# 5. SONUÇ VE ÖNERİLER

Bu tezimizde temel hedefimiz, YSA, KA, LR, DVM ve K-NN gibi farklı veri madenciliği yöntemlerinin kullanılarak Bolu Abant İzzet Baysal Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Kamu Yönetimi bölümü öğrencilerine ait mezuniyet notlarının mezuniyet öncesinden erkenden tahmin edilmesi işleminin gerçekleştirilmesidir. Kullanılan farklı modeller ile notların erken tahmini sayesinde öğrenciler için erken uyarı sistemi oluşturulabilecektir. Mezuniyet not tahmini işlemi için iki farklı senaryo denemesi yapılmıştır. İlk denemede sadece birinci sınıf ders notları kullanılarak tahmin işlemi yapılırken oluşturulan ikinci senaryoda ise birinci ve ikinci sınıf ders notları kulanılmıştır. Tez çalışması sonucunda elde edilen çıkarımlar aşağıdaki gibidir,

1. Oluşturulan iki senaryo ve kullanılan makine öğrenmesi yöntemleri ile tahmin konusunda belirli bir başarım elde edilmiştir. Elde edilen sonuçlar görsel ve rakamsal olarak gösterilmiştir.
2. Yapılan çalışma ile YSA'nın her iki senaryo için diğer yöntemlere göre daha yüksek ve iyi tahmin başarımına ulaştığı görülmüştür.
3. Oluşturulan ikinci senaryonun birinci senaryoya göre daha iyi derecede tahmin sonucuna sahip olduğu tespit edilmiştir. Giriş değerleri sayısının artmasının bu noktada etkili olduğu görülmektedir.
4. Oluşturulan senaryolar için parametrelerin ayarlanması esnasında çok sayıda deneme yapılmıştır. Kullanılan tüm modeller için ayrı ayrı parametrik ayarlamalar yapılmıştır.

## 5.1 Öneriler ve Gelecek Çalışmalar

1. Bu tez çalışmasındaki başarımın daha da yükseltilmesi için farklı makine öğrenmesi yöntemleri kullanılabilir. Özellikle Bulanık mantık tabanlı regresyon yöntemleri ve bununla birlikte farklı bazı istatistiksel modeller de kullanılabilir.
2. Meydana getirilen senaryolar hem değiştirilebilir hem de farklı bölümler üzerinde de denenebilir.
3. Modeller için sisteme girilen giriş öznitelik vektörü üzerinde normalizasyon işlemi yapılarakta başarım değerlendirmesi yapılabilir.

# 6. KAYNAKLAR

(Bu tez çalışmasında APA atıf sistemi kullanılmıştır.)


Akpınar, H. (2014). *Data: Veri Madenciliği Veri Analizi*, 1. baskı. Papatya Yayıncılık Eğitim, İstanbul.

Al-Khafaji, M., & Eryilmaz, M. (2021, November). Using Artificial Intelligence Methods to Predict Student Academic Achievement. In *Proceedings of the Future Technologies Conference* (pp. 403-414). Springer, Cham.

Aybek, H. S. Y. (2018). *Öğrenci başarısının yapay sinir ağları ile kestirilmesi: Anadolu Üniversitesi Açıköğretim Sistemi örneği* (Doctoral dissertation, Anadolu University (Turkey)).

Aydoğan, İ., & Zırhlıoğlu, G. (2018). Öğrenci başarılarının yapay sinir ağları ile kestirilmesi. *Van Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi*, *15*(1), 577-610.

Baker, R. S. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems*, *29*(3), 78-82.

Balaban M., Kartal E. (2018). *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları.* 2. Baskı. İstanbul: Çağlayan Yayın Evi.

Ben-Zadok G., Hershkovitz, A., Mintz, R. and Nachmias, R. (2007). Examining online learning processes based on log files analysis: a case study. *Research, Reflection and Innovations in Integrating ICT in Education*.

Boser BE, Guyon IM, Vapnik, VN. A Training Algorithm for Optimal Margin Classifiers. Proceedingof the 5th Annual Workshop on Computational Learning Theory COLT'92(s.144- 152). Pittsburgh, PA,Usa: ACM Press, 1992 39.

Bresfelean P., Bresfelean M., Ghisoiu N. (2008). Determining Students' Academic Failure Profile Founded on Data Mining Methods, *Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces*, 23-26.

Ching, W. K., & Ng, M. K. (2003). *Advances in Data Mining and Modeling, Hong Kong, 27-28 June 2002*. World Scientific.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250.

Choi, S., & Kim, Y. J. (2021). Artificial neural network models for airport capacity prediction. *Journal of Air Transport Management*, *97*, 102146.

Congalton, R. G., & Green, K. (2019). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Çınar, A. (2019). Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi ve R Dili ile Bir Uygulama. *Öneri Dergisi*, *14*(51), 90-111.

Davis, C. M., Hardin, J. M., Bohannon, T., & Oglesby, J. (2007). Data mining applications in higher education. *Data Min. Methods Appl*, 123-148.

Dongare, A. D., Kharde, R. R., & Kachare, A. D. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189-194.

Egrioglu, E., Aladag, C. H., Yolcu, U., Uslu, V. R., & Basaran, M. A. (2009). A new approach based on artificial neural networks for high order multivariate fuzzy time series. *Expert Systems with Applications*, *36*(7), 10589-10594.

Elmas, Ç., (2011). *Yapay zeka uygulamaları*, 2. Baskı. Ankara: Seçkin Yayıncılık.

Erdoğan Ş., Timor M. (2005). A Data Mining Application in a Student Database, *Havacılık ve Uzay Dergisi*, 2(2), 57-64.

Gaafar L. and Khamis M, (2009), Applications of Data Mining for Educational Decision Support, *Proceedings of the 2009 Industrial Engineering Research Conference*, 228-233

Goh, A. T. (1995). Back-propagation neural networks for modeling complex systems. *Artificial intelligence in engineering*, *9*(3), 143-151.

Gürsakal, N. (2018). *Makine öğrenmesi*. Baskı, Bursa: Dora Basım Yayın Dağıtım Ltd. Şti.

Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. *Morgan Kaufmann*, *340*, 94104-3205.

Hurwitz, J., & Kirsch, D. (2018). Machine Learning IBM Limited Edition. *Retrieved March*, *22*, 2021.

İşler, B., & Kılıç, M. (2021). Eğitimde Yapay Zeka Kullanımı ve Gelişimi. *Yeni Medya Elektronik Dergisi*, *5*(1), 1-11.

Jackson, A. H. (1988). Machine learning: a probabilistic perspective.

Kaya, F. H. (2022). *Identifying the factors affecting students' academic achievement using machine learning algorithms* (Master's thesis, Konya Teknik Üniversitesi).

Kılıç, S. (2013). Doğrusal regresyon analizi. *Journal of Mood Disorders*, *3*(2), 90-92.

Kılınç, Ç. (2015). *Üniversite öğrenci başarısı üzerine etki eden faktörlerin veri madenciliği yöntemleri ile incelenmesi* (Master's thesis, ESOGÜ, Fen Bilimleri Enstitüsü).

Kotsiantis S.B., Patriarcheas K., NikXenos M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowl.-Based Syst*. 23(6), 529-535.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. *applied sciences*, *9*(20), 4396.

Maind, S. B., & Wankar, P. (2014). Research paper on basic of artificial neural network. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(1), 96-100.

Mardikyan, S., Badur B. (2011). Analyzing Teaching Performance of Instructors Using Data Mining Techniques, *Informatics in Education*, 10(2), 245–257.

Mashat, A. F., Fouad, M. M., Philip, S. Y., & Gharib, T. F. (2012). A decision tree classification model for university admission system. *International Journal of Advanced Computer Science and Applications*, *3*(10).

Minaei-Bidgoli, B., Kashy, D. A., Kortmeyer, G., and Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. *In The proceedings of the 33rd ASEE/IEEE frontiers in education conference*. Boulder, CO.

Nizam, H., & Akın, S. S. (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*, *1*(6).

Özdemir, A., Saylam, R., & Bilen, B. B. (2018). Eğitim Sisteminde Veri Madenciliği Uygulamaları ve Farkındalık Üzerine Bir Durum Çalışması. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *22*(Özel Sayı 2), 2159-2172.

Özkan Y. (2013). *Veri Madenciliği Yöntemleri.* 2. Baskı. İstanbul: Papatya Yayıncılık.

Öztürk, K., & Şahin, M. E. (2018). Yapay sinir ağları ve yapay zekâ'ya genel bir bakış. *Takvim-i Vekayi*, *6*(2), 25-36.

Pham, D. T., & Pham, P. T. N. (1999). Artificial intelligence in engineering. *International Journal of Machine Tools and Manufacture*, *39*(6), 937-949.

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, *3*(2), 334-337.

Ranjan, J., & Khalil, S. (2008). Conceptual framework of data mining process in management education in India: an institutional perspective. *Information Technology Journal*, *7*(1), 16-23.

Raschka, S. (2018). Model evaluation, model selection, and algortithm selection in maching learning. *asXiv preprint aeXiv:1811.12808*

Samuel, A. L. (1988). Some studies in machine learning using the game of checkers. II—recent progress. *Computer Games I*, 366-400.

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, *11*(21), 1-23.

Selvi, A. (2020). *Bilecik ilinde ilköğretimden liseye geçiş sınavlarında makine öğrenmesi yöntemleri ile öğrenci başarısının tahmini* (Master's thesis, Bilecik Şeyh Edebali Üniversitesi, Fen Bilimleri Enstitüsü).

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011, April). Prediction of student academic performance by an application of data mining techniques. In *International Conference on Management and Artificial Intelligence IPEDR* (Vol. 6, No. 1, pp. 110-114).

Sen, B., Ucar, E. and Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach, *Expert Systems with Applications*, 39, 9468–9476.

Silahtaroğlu G. (2020). *Bilgisayar Bilimleri Veri Madenciliği Kavram Algoritmaları*, Papatya Bilim Akademik Yayınevi 4. Basım

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *towards data science*, *6*(12), 310-316.

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4), 2094-2097.

Şengür, D. (2013). *Öğrencilerin Akademik Başarilarinin Veri Madenciliği Metotlari ile Tahmini, Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, Bilgisayar ve Öğretim Teknolojileri Eğitimi* (Doctoral dissertation, Yüksek Lisans Tezi).

Vranić M., Pintar D., Skoćır Z. (2007). The Use of Data Mining in Education Environment, *ConTEL 2007* Zagrep, 243-251.

Winston, P. H. (2017). On Computing Machinery and Intelligence. In *Philosophical Explorations of the Legacy of Alan Turing* (pp. 265-278). Springer, Cham.

Zaïane O. R., Luo J. (2001). Web usage mining for a better web-based learning environment, *Conference on Advanced Technology for Education*, 60-64

Zhang Y., Oussena S., Clark T. and Kim H. (2010). Use data mining to improve student retention in higher education - a case study. *In ICEIS 2010: Proceedings of the 12th International Conference on Enterprise Information Systems*, Volume 1: Databases and Information Systems Integration, pages 190-197. INSTICC, Funchal, Portugal.

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, *23*, 100224.