YAPAY ZEKA TEKNİKLERİ İLE AÇIK ÖĞRETİM LİSESİ ÖĞRENCİLERİNİN MEZUNİYET TAHMİNİ

Mirhaç SULAK

2021
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ

Tez Danışmanı
Dr. Öğr. Üyesi Yüksel ÇELİK

YAPAY ZEKA TEKNİKLERİ İLE AÇIK ÖĞRETİM LİSESİ ÖĞRENCİLERİNİN MEZUNİYET TAHMİNİ

Mirhaç SULAK

T.C.
Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalında
Yüksek Lisans Tezi
Olarak Hazırlanmıştır

Tez Danışmanı
Dr. Öğr. Üyesi Yüksel ÇELİK

KARABÜK
Ocak 2021

Mirhaç SULAK tarafından hazırlanan "YAPAY ZEKA TEKNİKLERİ İLE AÇIK ÖĞRETİM LİSESİ ÖĞRENCİLERİNİN MEZUNİYET TAHMİNİ" başlıklı bu tezin Yüksek Lisans Tezi olarak uygun olduğunu onaylarım.

Dr. Öğr. Üyesi Yüksel ÇELİK .....
Tez Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

Dr. Öğr. Üyesi Erdal BAŞARAN .....
İkinci Danışman, Ağrı İbrahim Çeçen Üniversitesi MYO Bilgisayar Teknolojileri Bölümü

Bu çalışma, jürimiz tarafından Oy Birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir. 29/06/2021

Ünvanı, Adı SOYADI (Kurumu)	İmzası
Başkan : Prof. Dr. Oğuz FINDIK (KBÜ)	.....
Üye : Dr. Yüksel ÇELİK (KBÜ)	.....
Üye : Dr. Öğr. Üyesi Abidin ÇALIŞKAN (BTÜ)	.....

KBÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulu, bu tez ile, Yüksek Lisans derecesini onamıştır.

Prof. Dr. Hasan SOLMAZ .....
Lisansüstü Eğitim Enstitüsü Müdürü

"Bu tezdeki tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim."

Mirhaç SULAK

ÖZET

Yüksek Lisans Tezi

YAPAY ZEKA TEKNİKLERİ İLE AÇIK ÖĞRETİM LİSESİ ÖĞRENCİLERİNİN MEZUNİYET TAHMİNİ

Mirhaç SULAK

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Yüksel ÇELİK
Dr. Öğr. Üyesi Erdal BAŞARAN
Haziran 2021, 51 sayfa

Açık Öğretim Lisesine kayıt yaptıran öğrencilerin normal eğitim süresi içerisinde mezun olup olmayacaklarının erken tahmini, Millî Eğitim Bakanlığının gerekli eğitim politikalarını geliştirebilmesi bakımından önemlidir. Milli Eğitim Bakanlığı Hayat Boyu Öğrenme Genel Müdürlüğü'ne bağlı Açık Öğretim Lisesi örgün eğitim dışında kalan bütün halkı kapsayan bir okul türüdür. Bu liselere kayıt yaptıran öğrenciler için en kısa sürede lise diplomalarını alıp hayatlarında yeni yollara yönelmek çok önemlidir. Bu çalışmada, öğrencilerin başarılı oldukları ders kredilerinin, devam ettikleri dönem sayısının, yaşadıkları il ve ilçelerin, yaşlarının ve cinsiyetlerinin mezuniyet sürelerine etkilerini analiz etmek amacıyla yapay zeka algoritmalarının uygulamaları hakkında bilgi verilmektedir. Bu amaçla 2010-2012 yılları arasında Açık Öğretim Lisesine kayıt yaptıran öğrencilerden 142.714

öğrencinin kayıtları kullanılarak Karar Ağaçları (KA), K-En Yakın Komşuluk (KNN) Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA) algoritmaları uygulanmıştır. Etkili bir tahmin sonucu için öğrencilerin farklı özellikleri dikkate alınarak en uygun giriş parametreleri belirlenmiştir. Bu parametreler; öğrencilerin yaşadıkları şehirlerin sosyo-ekonomik gelişmişlik seviyeleri ve skorları, kayıt esnasındaki yaşları, cinsiyetleri, başarılı oldukları toplam ders krediler, öğrenim gördükleri toplam dönemler, özel durumları ve engel durumlarıdır.

Önerilen yöntemlerin zamanında mezun olma/olamama durumu hakkında tatmin edici tahminler yaptığı gözlenmiştir. Bu sistem, Millî Eğitim Bakanlığının mevcut AOL Bilgi Sistemine entegre edilerek öğrencilerin en kısa sürede mezun olabilecek başarıyı göstermelerine yardımcı olabileceğini öngörüyoruz.

Anahtar Sözcükler : Mezuniyet Tahmini, Karar Ağacı, K-En Yakın Komşuluk, Destek Vektör Makineleri, Yapay Sinir Ağları.

Bilim Kodu : 92518

ABSTRACT

M. Sc. Thesis

PREDICTING GRADUATION OF OPEN EDUCATION HIGH SCHOOL STUDENTS WITH ARTIFICIAL INTELLIGENCE TECHNICS

Mirhaç SULAK

Karabük University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Asst. Prof. Yüksel ÇELİK
Asst. Prof. Erdal BAŞARAN
June 2021, 51 pages

An early estimate of whether students enrolled in an open Education High School will graduate during the normal education period is important for the Ministry of Education to develop the necessary education policies. Open Education High School, which is affiliated to the General Directorate of lifelong learning of the Ministry of Education, is a type of school that covers all people outside of formal education. For students who enroll in these high schools, it is very important to get their high school diplomas as soon as possible and move on to new paths in their lives. In this study, it is informed about the applications of artificial intelligence algorithms to analyze the effects of the course credits which the students are successful, number of semesters they joined, provincial and districts they live in, their genders and registration ages to graduation periods. For this purpose, 142.714 records belong to students registered in

years between 2010-2012 have been used for training and validating decision trees (DT), k-nearest Neighbour (KNN), support vector machines (SVM) and artificial neural networks (ANN) algorithms. The optimal input parameters are determined by taking into account the different features of students for an effective estimate result. These parameters; Socio-economic development levels and scores of the cities where students live in, the ages during registration total number of semesters which they joined, their special situations and disabilities.

It has been observed that the proposed methods make satisfactory estimates of the status of graduation/failure on time. We anticipate that this system can be integrated into the existing AOL Information System of the Ministry of Education, helping students demonstrate success that can be graduated as soon as possible.

Key Word : Graduation Prediction, Decision Tree, K-Nearest Neighborhood, Support Vector Machines, Artificial Neural Networks.

Science Code : 92518

TEŞEKKÜR

Bu tez çalışmasında geniş bilgi ve tecrübe birikimini esirgemeyen ve bilimsel temeller üzerine yönlendirmeleriyle daha iyiye yönelten sayın hocam Dr. Öğr. Üyesi Yüksel ÇELİK'e, çalışmalarımda yardımlarını esirgemeyen sayın hocam Ağrı Üniversitesi Dr. Öğr. Üyesi Erdal BAŞARAN'a sonsuz teşekkürlerimi sunarım.

Ayrıca çalaışmalarımda daha iyiye yönlendiren Osmaniye Korkut Ata Üniversitesi Dr. Öğr. Üyesi Haydar TUNA'ya teşekkür ederim.

Manevi hiçbir yardımını esirgemeyen iş arkadaşlarım ve sevgili dostum Raşit DEMİREL ve saygı ve sevgisini hiç bir zaman esirgemeyen sevgili eşime tüm kalbimle teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

ANO	: Ağırlıklı Not Ortalaması
AÖL	: Açık Öğretim Lisesi
CART	: Sınıflandırma ve Regresyon Ağacı
DVM	: Destek Vektör Makineleri
FN	: Yanlış Negatif
FP	: Yanlış Pozitif
GBM	: Gradyan Arttırma Makinesi
HBÖGM	: Hayatboyu Öğrenme Genel Müdürlüğü
KA	: Karar Ağacı
KNN	: K-En Yakın Komşuluk
LR	: Lojistik Regresyon
MEB	: Millî Eğitim Bakanlığı
NB	: Naive Bayes
NBC	: Naive Bayes Sınıflandırıcı
ROC	: Alıcı İşlem Karakteristik
TN	: True Negative
TP	: Doğru Pozitif
VM	: Veri Madenciliği
YSA	: Yapay Sinir Ağları
YZ	: Yapay Zeka

BÖLÜM 1

GİRİŞ

Günümüzde birçok kurum bilişim teknolojileri sayesinde depoladıkları verilerden anlamlı bilgiler elde etmeyi amaçlamaktadır. Eğitim kurumları da kullanıcılarına daha iyi hizmet verebilmek amacı ile bilişim teknolojilerine önem vermekte ve bu teknolojileri aktif olarak her aşamada kullanmaktadır. Artık kurumlar veri tabanlarında saklanan verilerden kendi kurumlarına faydalı olabilecek bilgilere ulaşabileceklerini fark etmişlerdir.

Bu kapsamda Millî Eğitim Bakanlığına (MEB) bağlı okullarda öğrencilerin başarılarını tahmin eden sistemler öğrencilerin başarısız olma risklerinin önceden tespiti ve önlenmesi, başarılarını arttıracak çözümlerin belirlenmesi konusunda fikir sahibi olmamızı sağlayabilir ve bakanlık tarafından bu çözümlerin bir bütün olarak eğitim politikalarına yansıtılarak eğitim maliyetlerinin düşürülmesine katkı sağlayabilir. Mezuniyet tahmin sistemi, öğrenci kayıt yaptırdıktan bir süre sonra öğrencinin mezuniyet durumu hakkında erkenden bilgi veren bir tahmin sistemidir denilebilir.

Veri madenciliği (VM), büyük miktardaki veri kümlerinin işlenerek ilk bakışta anlaşılamayan, göz önünde olmayan kullanışlı bilgilerin elde edilmesidir [1]. Bu kavramın daha iyi anlaşılabilmesi için veritabanlarında bilgi madenciliği [2], bilgi çıkarımı [3], veri ve örüntü analizi [4] gibi farklı isimlerle de anılmaktadır. Günümüzde logaritmik bir şekilde artan veri kümelerinde kullanıcılar için pek bir anlam ifade etmeyen kayıtlı veriler, belirli bir hedefte sistematik bir şekilde işlenirse bu değersiz veri yığınlarından çok kıymetli bilgiler elde edilebilir [5]. Veri madenciliğinde kullanılan bazı teknikler; Sınıflandırma (Classification), Kümeleme (Clustering), Regresyon (Regression), Birliktelik Kuralları (Association Rules) [5], Aykırılık Algılama (Outlier Detection) [6], Sıralı Örüntü (Sequential Patterns) [7] gibi tekniklerdir.

Son zamanlarda eğitim alanındaki veri madenciliğinde öğrencilerin başarı performansları analiz edilerek başarısızlık nedenleri hakkında fikir edinilmekte ve bu bilgiler ışığında daha isabetli eğitim stratejileri belirlenebilmektedir. Bu amaçla yapay zekâ araçlarını kurum sistemlerine engetre etme yönelimi vardır [8].

Bu çalışmada, MEB Hayat Boyu Öğrenme Genel Müdürlüğü (HBOGM) Açık Öğretim Lisesi (AÖL) öğrencilerinden 142.714 öğrencinin dönem ders başarıları üzerinde deneyler yapıp öğrencilerin öngörülen mezuniyet süresi içerisinde mezun olup olmama durumlarının tahmin edilmesi amaçlanmıştır. Bunun sonucunda edinilen bilgiler ışığında ve risk faktörleri belirlenip gerekli önlemler alınabilir.

Verileri alınan bu 142.714 öğrenci, AÖL'ye 2010-2012 yılları arasında kayıt yaptırmış öğrencilerdir. Bu öğrencilerin bir kısmı 2020 yılına kadar mezun olmuş bir kısmı ise halen mezun olamamış ve öğrenimine devam etmektedir. Ayrıca bu öğrenciler AÖL'ye ortaokul diplomaları ile kayıt yaptırmış, herhangi bir yurt dışı okuldan denklikleri veya aynı düzeyde başka bir örgün öğretim lise seviyesi bir okuldan getirdikleri geçmiş kredileri bulunmamaktadır.

AÖL tüm Türkiye'de aynı koşullarda eğitim verdiği için ülke genelinde yaygındır. Bu okula, ortaokul mezunu olarak diploma ile kayıt ve arasınıftan tasdikname ile kayıt olmak üzere iki türlü kayıt yapılmaktadır. Ara sınıftan tasdikname ile kayıt yaptıran çok fazla öğrenci olması ve bu kayıtlardaki ders ve kredi bilgilerinin tamamının kontrol edilip doğruluğunun onaylanmamış olması nedeniyle yalnızca ortaokul mezunu olarak diploma ile kayıt yaptırmış olan öğrenciler seçilmiştir. Ayrıca her öğrenci farklı il ve ilçelerde yaşadığı için yaşadıkları şehirlerin sosyo-ekonomik gelişmişlikleri farklıdır. Sosyo-ekonomik durum ise mezuniyete etki eden ciddi bir faktördür. Bu nedenle ülke geneli her şehrin sosyo-ekonomik gelişmişlik endeks skorları belirlenmiştir [9]. Çok fazla veri bulunduğu için ülke genelini temsil edebilecek, tüm gelişmişlik seviyelerini barındıran, 7 bölgeden 14 il, ilçeleriyle birlikte seçilmiştir. Bu il ve ilçelerden ortaokul mezunu olarak 2010-2012 yılları arasında kayıt yaptırmış tüm öğrenciler alınmıştır. Ayrıca her öğrencinin yaşadığı şehirlerin sosyo-ekonomik gelişmişlik endeksleri de alınmıştır.

Literatürde yapay zekâ teknikleri kullanarak öğrencilerin kayıt yaptırdıktan bir süre sonra mezuniyet durumlarının erkenden tahminine yönelik bir çalışma olmaması, öğrencilerin mezuniyetlerinin tahmini için farklı kriterlerin uygulanıyor olması çalışmayı özgün kılmaktadır. Yapay zekâ (YZ) algoritmalarından elde edilen sonuçların karşılaştırmalı analizlerinin yapılmış olması da çalışmayı öne çıkarmaktadır. Bu amaçla algoritmalar da kendi aralarında kıyaslanmıştır.

BÖLÜM 2

LİTERATÜR ÖZETİ

Literatür araştırmasında öğrencilerin başarısızlık risklerini, başarısızlık nedenlerini, ders bırakma durumlarının ve performanslarının tespitine, yönelik birçok çalışma yapıldığı görülmüştür. Bu çalışmalarda öğrenci notları, devamsızlık bilgileri, seçtikleri ve bıraktıkları dersler, ödev notları gibi birçok bilgileri kullanılmış ve bu bilgileri incelemek adına çok çeşitli veri madenciliği teknikleri geliştirildiği görülmüştür.

Öğrencilerin eğitim performanslarının tahminlerine yönelik çok boyutlu araştırmalar yapılmıştır. Devam etmekte olunan bir dersi bırakmanın ve seçilen derslerin iptalinin erken tahmini [10], öğrencilerin performanslarını etkileyen iç faktörlerin analizi [11] gibi farklı perspektiflerle yaklaşılmış çalışmalar vardır. Ayrıca okulu bırakanları ve yavaş öğrenenleri değerlendirmek için eğitim veri setlerinde farklı veri madenciliği teknikleri kullanıldığı çalışmaların da yapıldığı görülmüştür [12].

Erken tahmin, bu alandaki yeni bir olgudur. Uygun önleyici stratejiler önerir ve öğrencilerin başarısızlıklarının zamanında önüne geçme adına alınması gereken önlemleri erkenden sunar [13].

Bunların yanı sıra, Ağırlıklı Not Ortalaması (ANO) ve ödev, sınav notları gibi değerlendirmeleri temel parametreler olarak alan öğrenci performansı değerlendirmeleri de vardır [14]. Birkaç çalışmada da başarısızlık riskine sahip öğrencilerin davranış analizlerini yapmak için makine öğrenmesi tekniklerinin kullanıldığı görülmüştür [15–17].

Amerika Washington Üniversitesi 1998 – 2006 yılları arasında kayıt yaptıran 69.116 öğrenci verisi üzerinde okulu bırakmalarının erken tahmini üzerine çalışma yapılmıştır. Veri setinde öğrencilerin ırk, cinsiyet, doğum tarihi, ikamet durumu,

lisans öncesi Scholastic Aptitude Test (SAT) ve American College Testing (ACT) puanları, aldıkları dersler ve ders notları parametre olarak alınmıştır. Veri setinin %70'i eğitim, %30 test verisi olacak şekilde rasgele bölünmüştür. Lojik Regresyon, Random Forest, K-En Yakın Komşuluk (KNN) algoritmaları ile deneyler yapılmıştır. Tahmin sonuçların doğruluk oranları sırasıyla %66.59, %62.24, %64.60 olarak bulunmuştur [13].

Bir diğer çalışmada 2013 yılında uzaktan eğitim alan 262 lisans öğrencisinin verileri ve 2014 yılında yüzyüze eğitim alan 161 öğrencinin verilerileri üzerinde öğrencilerin başarısızlıklarının tahmini üzerine çalışılmıştır. İki adet veri seti üzerinde Destek Vektör Makineleri (DVM), Karar Ağacı (KA), Yapay Sinir Ağları (YSA), Naive Bayes (NB) metotları ugulanmıştır. Sonuç olarak uzaktan eğitimde %50 - %82 arası başarı doğruluk sağlanırken yüzyüze eğitimde %50 - %79 arası bir doğruluk değeri elde edilmiştir. Çalışma sonucunda en başarılı metot olarak uzaktan eğitimde %82 ve yüzyüze eğitimde %79 doğruluk oranı ile KA metodu bulunmuştur [15].

Kolombiya Ulusal Üniversitesi lisans öğrencilerinde akademik statü kaybını modellemek için akademik ve akademik olmayan verileri analiz etmek amacıyla iki farklı veri madencilik modeli tanımlanmıştır. Modellerde, ilk kayıt esnasında dönem kaybı ve düşük akademik başarı nedeniyle dönem kaybının tespitinde veri kalitesini değerlendirmek için NB ve KA sınıflandırma teknikleri uygulanmıştır. Deneysel çalışmalar sonucunda ilk kayıt verileri üzerindeki çalışmalada KA ile %51-52 arası, NB %54-57 arası değerler elde edilmiştir. Öğrencilerin ağırlıklı diğer bilgilerini de dâhil ettikten sonra tatmin edici sonuçlar elde edilmiştir. NB ile %75 üzerinde, %85'e varan doğruluğa ulaşılmıştır. Sonuç olarak, mühendislik programlarını tercih eden öğrencilerin dönem kayıplarında kabul testi puanlarından matematik ve fen bilimleri puanlarının daha etkili olduğu, kayıt esnasında daha genç yaşta olanların, özellikle 23-28 yaş aralığındaki öğrencilerin daha yüksek riskli olduğu görülmüştür [18].

Eğitimsel veri madenciliği modeli kullanılarak öğrencilerin akademik performanslarını etkileyen kritik dersleri tahmin etmek amacıyla bir çalışma yapılmıştır. Bu çalışmada akademik performans tahmin modelini geliştirmek için

ID3 Karar Ağacı İndüksiyon Algoritması kullanılmıştır. Modelin geliştirilmesinde Riyad Kral Suud Üniversitesi Lisans programlarından Bilgi Teknolojileri Bölümünden 2013-2014 eğitim öğretim yılında mezun olan 100 adet kız öğrenciye ait mezuniyet puanı, lise mezuniyet puanı, genel yetenek test puanı, eğitimsel erişim test puanı ve her öğrencinin seçtikleri ders bilgileri alınmıştır. 100 veriden 75 tanesi eğitim veri seti, 25 tanesi de test verisi olarak kullanılmıştır. her yıl için ayrı ayrı elde edilen modellerin performans değerlendirmelerinde ilk yıl için %68, ikinci yıl için %80, üçüncü yıl için %76 doğruluk değerlerine ulaşılmıştır. Sonuçlar, öğrencilerin ikinci yıl derslerindeki performansları üzerinde daha başarılı tahminler alınabildiğini göstermiştir [19].

Brezilya federal devlet liselerinde öğrenim gören öğrencilerin akademik performans öngürüsü üzerine bir analiz yapılmıştır. Bu çalışmada yine eğitimsel veri madenciliği teknikleri kullanılmıştır. Öncelikle öğrenci verilerinin genel kapsamını anlayabilmek için tanımlayıcı istatistiksel analiz yapılmıştır. Bu analiz sonucunda, öğrencilerin okula başlamadan önceki bilgileri ve okula başladıktan iki ay sonra toplanan akademik veriler olmak üzere iki grup parametre elde edilmiştir. Çalışmada, 2015-2016 yılı lise öğrencilerinin bilgileri alınmıştır. Alınan bilgilerden okulun bulunduğu şehir, okul adı, dönem, sınıftaki özel ihtiyaçların varlığı, sınıf türü, öğrenci adı, cinsiyeti, yaşı, devlet desteği, öğrencinin yaşadığı şehir ve mahalle, varsa özel ihtiyaçlarıı okuduğu bölüm, ilk iki ayluık notları, devamsızlığı, sınıf terar durumu gibi 17 farklı parametre elde edilmiştir. Öğretim yılı sonundaki öğrenci başarısının tahimini için Gradyan Arttırma Makinesi (Gradient Boosting Machine - GBM) tabanlı sınıflandırma modelleri oluşturulmuştur. Deneyler sonunda sınıf tekrarına kalma oranları 2015 yılı için %12.51, 2016 yılı için %13.08, her iki yıl için ortalama ise %12.80 çıkmıştır. Sonuçlar, öğrenci notları ve devamsızlık bilgilerinin en etkili parametreler olduğunu gösterse de öğrencilerin akademik başarı veya başarısızlıklarında mahalle, okul, yaş gibi parametrelerin de etkili olduğu gözlenmiştir [20].

Standartlara dayalı sınıflandırmayı kullanan bir okulda başarısızlık riski altındaki öğrencileri tanımlamak için kullanılan tahmin yöntemleri karşılaştırılmıştır. Tahmin metotlarında, ders öğretmenlerinin verdiği dönem içi sınav notları kullanılmıştır.

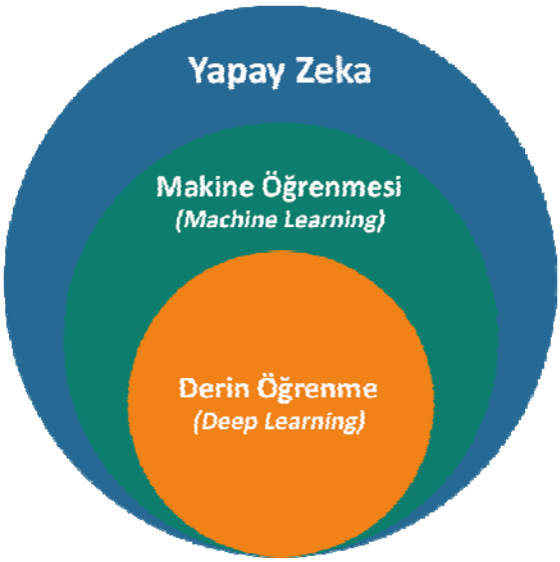
Modellerde hatalı negatif sonuçları minimize ederken hatalı pozitiflerin de artmamasına dikkat edilmiştir. Karşılaştırılan yedi farklı modelden en iyi sonuçları veren modeller sırasıyla DVM, KNN ve Naive Bayes Sınıflandırıcı (NBC) modelleri olmuştur. Çalışmada ABD orta batı eyaletlerindeki büyük üniversitelerin mühendislik programlarına 2013 bahar döneminde kayıt yaptıran 1650 birinci sınıf öğrencilerinin ve 2014 bahar döneminde kayıt yaptıran 1.413 birinci sınıf öğrencilerinin ortak zorunlu derse ait haftalık quiz, ödev, sınav ve proje notu bilgileri toplanmıştır. 2013 yılına ait veri seti rasgele olacak şekilde %50 eğitim, %25 metotları karşılaştırma ve iyileştirme, %25 ikincil testler için olmak üzere üç gruba ayrılmıştır. 2014 yılına ait veri setinin tamamı modellerin son testleri için kullanılmıştır. Çalışmada eğitimsel veri madenciliğinde genel olarak kullanılan Lojistik Regresyon (LR), DVM, KA, Çok Katmanlı Algılayıcılar (MLP), NBC, KNN olmak üzere altı farklı metot risk altındaki öğrencilerin tespiti için kullanılmıştır. Bütün metotlarda eğitim, doğrulama ve test işlemleri için aynı veriler kullanılmıştır. Veriler içerisinde derste başarısız olan öğrencileri bütün modeller doğru bir şekilde tespit etmiştir. Sonuç olarak NBC, DVM ve KNN modellerinden oluşturulan topluluk modeli %85 ile en yüksek doğruluk oranını vermiştir [21].

BÖLÜM 3

YÖNTEM VE ARAÇLAR

3.1. YAPAY ZEKA

Yapay zekâ ve Makine Öğrenmesi günümüz dünyasında gittikçe daha fazla yer almakta ve hayatımıza daha çok girmektedir. Harita uygulamalarından sağlık alanına ve savunma teknolojilerine kadar pek çok alanda giderek artan bir kullanıma sahiptir.



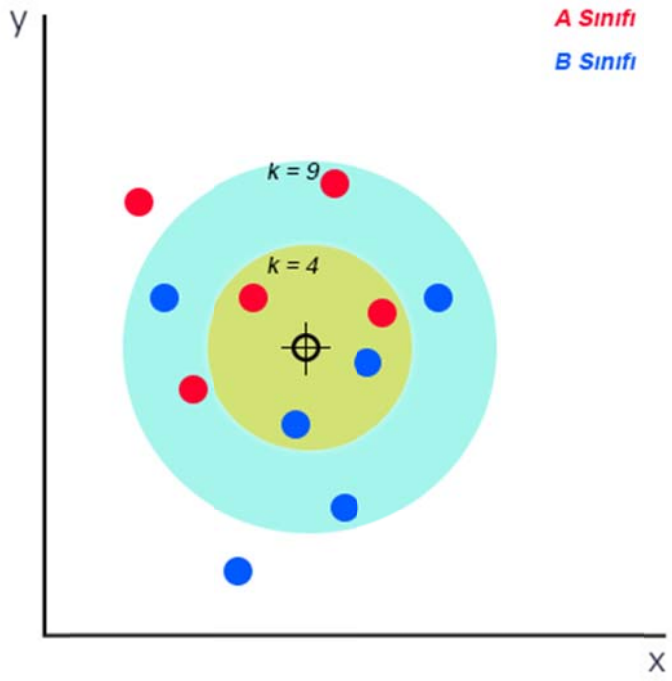
Şekil 4.1. Yapay zeka ve alt birimleri.

3.2. MAKİNE ÖĞRENMESİ

Makine Öğrenmesi, yapay zekânın bir alt birimidir (Şekil 3.1). Bilgisayar sistemlerinin yüksek miktarda verileri matematiksel ve istatistiksel yöntemlerle işleyerek kendi başlarına öğrenmelerini sağlar. Gelişen teknolojiyle daha büyük işlem gücüne ulaşan günümüz bilgisayarlarının bu gücünden yararlanarak yüksek miktarda işlemler daha kısa sürede çözüme ulaştırılabilmektedir. Makine öğrenmesinden beklenen doğru tahminler yapmasıdır [22].

3.2.1. K-En Yakın Komşu

Makine öğrenmelerinden En Yakın Komşu kuralı, sınıfı bilinen bir noktadan en yakın komşu noktasına hareketle bilinmeyen verilerin sınıflandırılması anlamına gelir. Bu kural genellikle örüntü belirlenmesi [23], metin sınıflandırma [24, 25], sıralama modelleri [26], olay tanıma [27] ve nesne tanıma [28] uygulamalarında kullanılır. KNN algoritması ise bir verinin sınıfının belirlenmesinde şekil 3.2'de gösterildiği gibi, kaç tane en yakın komşunun dikkate alınacağını belirten k değerine bağlı olarak hesaplanır [29]. Bu algoritma çok fazla bellek tüketmesi ile bilinir.



Şekil 4.2. KNN sınıflandırma örneği.

KNN algoritması sınıflandırma işleminde başlangıç noktasına en benzer k örnekleri arasındaki metrik mesafe hesaplanır. Bu hesaplamada Öklid Mesafesi (Formül 3.1), Manhattan Mesafesi (Formül 3.2) ve Minkovski Mesafesi (Formül 3.3) olmak üzere genelde üç tür uzaklık fonksiyonu kullanılır.

$$\sum_{i=1}^{n} |x_i - y_i| \tag{4.1}$$

$$\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.2}$$

$$\left(\sum_{i=1}^{n} (|x_i - y_i|)^q \right)^{1/q} \tag{4.3}$$

En yakın komşu kuralını kullanan KNN algoritması büyük veri örneklerini hedef alır. Bu algoritmanın bazı avantajları ve dezavantajları vardır [30].

Avantajları:

1. Eğitim çok hızlıdır,
2. Basit ve kolay öğrenilir,
3. Gürültülü eğitim verilerine karşı dirençlidir,
4. Eğitim verileri büyükse etkilidir.

Dezavantajları ise;

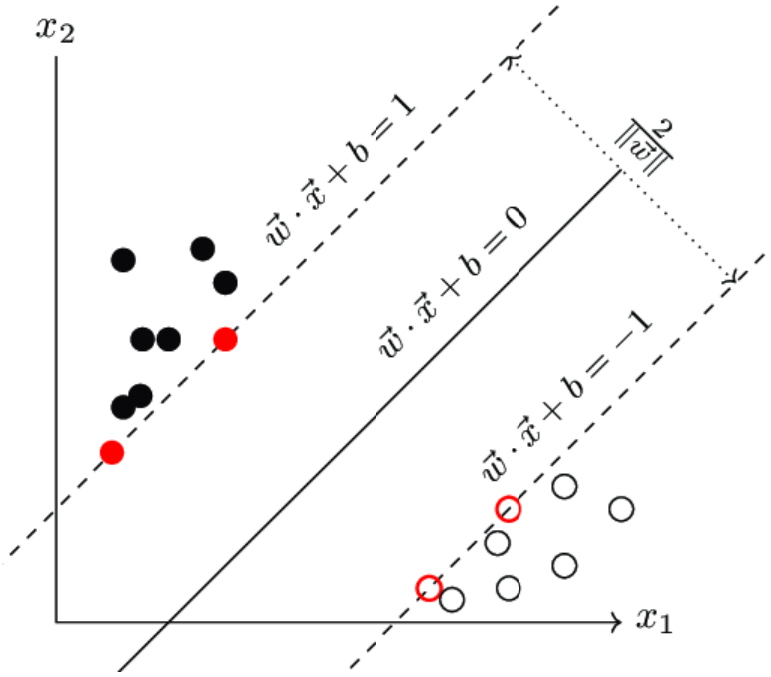
1. K değerine bağımlıdır,
2. Hesaplamalar karmaşıktır,
3. Bellek sıkıntısı yaşanabilir,
4. Çok yavaştır,
5. Alakasız parametre değerlerinde kolayca yanlış sonuç verebilir.

KNN algoritmasında dört adımda sınıflandırma yapılır.

1. Veriler incelenir, k parametresi belirlenir.
2. Uzaklık fonksiyonuna göre hesaplanır (Öklid, Manhattan, Minkowski),
3. Sınıflandırılacak veriye en yakın k adet komşu bulunur,
4. Sınıflandırılacak veri, K adet komşusunun uzaklıkları toplamı en büyük olanın sınıfına dahil edilir.

3.2.2. Destek Vektör Makineleri

Destek Vektör Makineleri Corinna Cortes ve Vladimir Vapnik [31] tarafından önerilmiştir. Makine öğrenmesi algoritmaları arasında çok kullanılan ve kernel tabanlı, denetimli öğrenme kullanan bir algoritmadır. Bu algoritmada amaç verileri iki sınıfa ayırmaktır. Bir düzlem üzerine yerleştirilmiş olan iki sınıfa da ait verileri bir hiperdüzlem ile ayırır. Bu hiperdüzlem, şekil 3.3'te gösterildiği gibi, her iki sınıfın da verilerine maksimum uzaklıkta olacak şekilde geçer.



Şekil 4.3. Destek vektör makineleri sınıflandırma örneği [32].

Şekil 4.3'te gösterilen kırmızı noktalardan geçen doğrular, formül 3.4'te gösterilen +1 ve -1 destek hiperdüzlemleridir ve bu iki destek hiperdüzlemi arasında kalan kısım ise margin olarak adlandırılır [32].

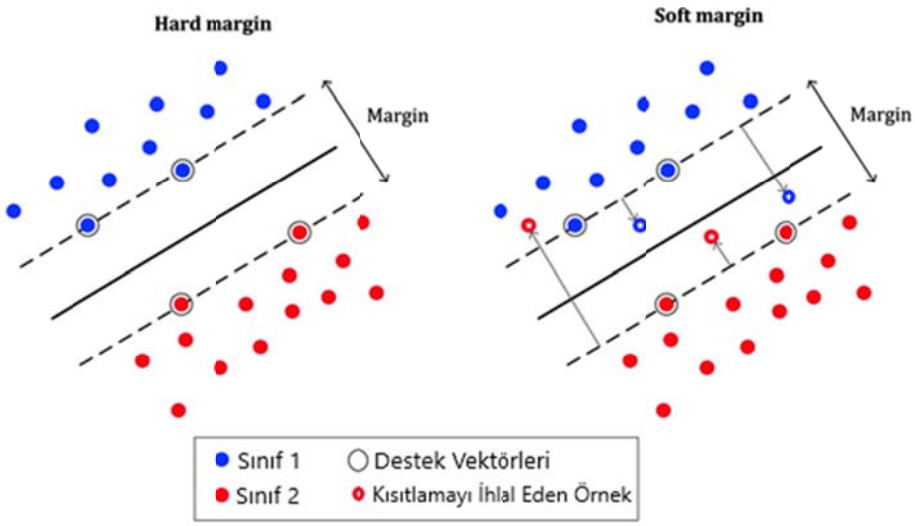
$$\begin{cases} y_i = +1, & \vec{w}\cdot\vec{x_i}+b \geqslant +1 \\ y_i = -1, & \vec{w}\cdot\vec{x_i}+b \leqslant -1 \end{cases} \tag{4.4}$$

DVM'ler Hard Margin ve Soft Margin olarak ikiye ayrılır. Hard Margin DVM'de verilerin, marginleri ihlal etmeden doğru bir şekilde sınıflandırılması gerekmektedir [33]. Başka bir ifadeyle, hard margin DVM'de örneklerin tamamının hiperdüzlemin

doğru tarafında olması ve formül 3.5'te ifade edildiği gibi, bu hiperdüzleme uzaklıklarının margin genişliğine eşit ya da büyük olması gerekmektedir.

$$\begin{aligned} \max \quad & \frac{s}{\sqrt{w^T w}} \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq s, \ i \in I \\ & w \in \mathbb{R}^n, \ b \in \mathbb{R}. \end{aligned} \tag{4.5}$$

Fakat bu durum şekil 3.4'te gösterildiği gibi, yalnızca örneklerin lineer bir şekilde ayrılabildiği durumlarda mümkündür. Bunun mümkün olmadığı durumlarda şekil 3.4'te gösterildiği gibi diğer bir DVM türü olan Soft Margin kullanılabilir.



Şekil 4.4. Hard Margin ve Soft Margin örnekleri.

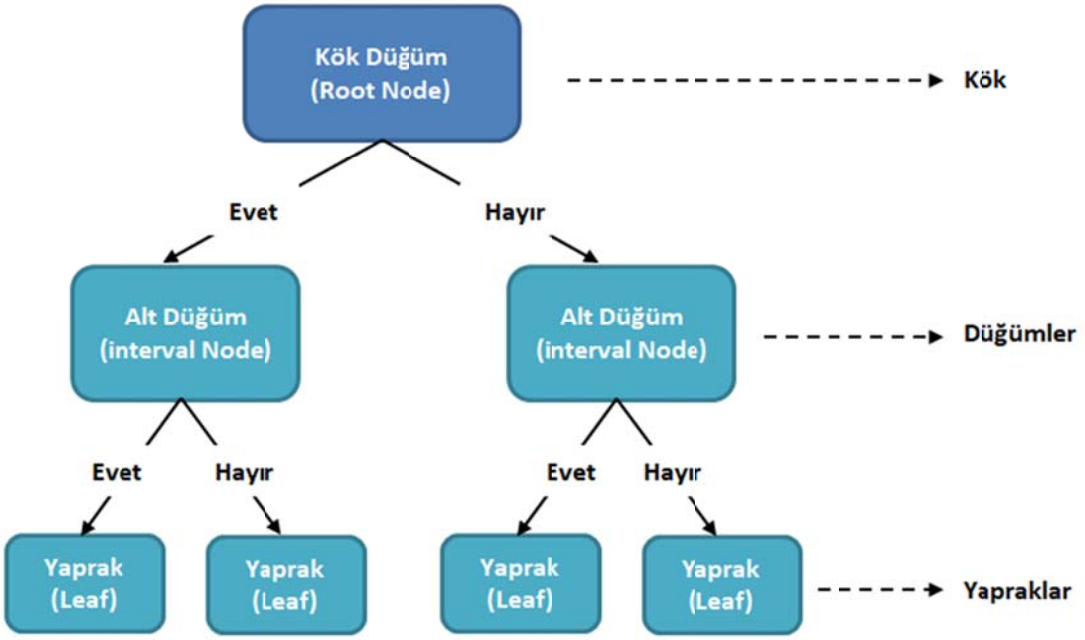
$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C \sum_{i \in I} \xi_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \ i \in I \\ & w \in \mathbb{R}^n, \ b \in \mathbb{R}, \ \xi \in \mathbb{R}^l_+. \end{aligned} \tag{4.6}$$

Soft Margin, formül 3.6'da gösterildiği gibi bir C değerine sahiptir. C değerinin küçük olması hiperdüzlemin margin genişliğini arttırır. Fakat C değerinin çok küçük olması, eğitim veri setinin sınıflandırıcıyı yeterince eğitmediği anlamına gelen

underfit sorununa yol açar. C değerinin çok büyük seçilmesi ise, sınıflandırıcının eğitim veri setini ezberlediği ve yeni gelen verileri sınıflandırmada başarısız olacağı anlamına gelen overfitting sorununa yol açacaktır. Sonuç olarak C parametresi sınıflandırma tahmininde overfitting ve underfitting sorunlarını önlemede ciddi bir etkiye sahiptir.

3.2.3. Karar Ağacı

Karar Ağaçları, denetimli öğrenme modellerinden biridir. Sınıflandırma, regresyon problemleri gibi birçok problemin çözümünde kullanılabilen bir modeldir. Veri setinden tahmin modelleri üretmek için kullanılır. Gürültülü verilere karşı dirençlidir. Kolay kullanımı ve etkili bir metot olması en önemli avantajlarındandır. KA ile üretilen kuralların yorumlanması ve anlaşılması kolaydır. Bu algoritma çoklu parametreye sahip bir veri setinden bir ağaç üretir. Bu ağaç, kök düğüm olarak adlandırılan bir değişkenden başlayarak dallara ve başka düğümlere ayrılarak yaprak adı verilen sonuçlara ulaşmayı hedefleyen ağaca benzer bir şekilde oluşturulmuş hiyerarşik bir ilişkiler yapısıdır.



Şekil 4.5. Örnek KA yapısı.

Karar ağacında en üstte kök düğüm adı verilen bir değişken vardır. Bu kök düğüm, birden fazla dala ayrılarak yeni düğümler oluşturur. Her yeni düğümde bir soru sorulur ve bu sorunun yanıtına göre alt düğümler oluşur. Evet/hayır şeklinde iki yanıtı olan sorularda düğüm iki dala ayrılır (Şekil 3.5). Her bir dal diğer değişkenlerin sınıflarına veya aralıklarına göre bölünerek yeni dallar oluşturur. Her dalda bölünen düğüme ana düğüm, oluşan yeni düğümlere alt düğüm adı verilir. Bu bölünmeler, kesme kuralı gerçekleşene kadar devam eder [34].

Bu mantıktan yola çıkılarak, Iterative Dichotomiser 3 (ID3) isimli bir karar ağacı geliştirilmiş ve daha sonra bu yaklaşım daha da geliştirilerek C4.5 versiyonu oluşturulmuştur [35]. Algoritma, eğitim verilerinin nitelikleri arasında arama yaparak örnekleri en iyi şekilde bölen ayırım niteliğini bulur ve eğer bu nitelik eğitim setini en iyi şekilde sınıflandırıyorsa algoritma durur.

Karar ağaçlarında Classification and Regression Tree (CART), ID3 ve C4.5 adında üç tür algoritma kullanılır.

CART algoritmasında kök düğümden bölünen her bir alt düğüm için "Gini" adı verilen bir saflık değeri hesaplanır (Formül 3.7).

$$Gini = 1 - \sum_{j} p_j^2 \tag{4.7}$$

Formül 3.12'de, p^2_j, j sınıfının gerçekleşme olasılığıdır. Her bir j sınıfı için p gerçekleşme olasılığı hesaplanır. Hesaplama sonuçların kareleri toplanarak 1 den çıkarılır. Bu değer 0-1 aralığındadır ve sıfıra ne kadar yakında o derece iyi ayrım yapılmış demektir.

ID3 algoritmasında, Entropy hesaplaması kullanılır (Formül 3.8). Entropy hesaplanırken $\log_2 n$ formülü kullanılır. N değeri olası farklı ihtimallerin sayısını ifade eder.

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j) \tag{4.8}$$

C4.5 algoritmasında, ID3 algoritmasına bölünme bilgisi eklenmiştir (Formül 3.9). Burada D hedef değişkeni temsil etmektedir. V ise tahmin edici değişkenin alabileceği değerlerin sayısını ifade eder.

$$-\sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ Info(D_j) \tag{4.9}$$

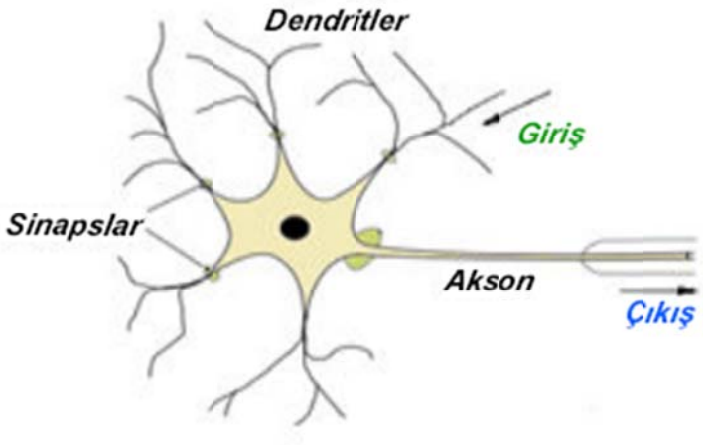
Karar ağaçlarında kök düğümden itibaren her bölünmeyle oluşan yeni düğüm ile birlikte yeni bir seviye oluşmaktadır. Her seviyeye derinlik adı verilir. Kök düğümün derinlik değeri "0" sıfırdır. Her düğümün kök düğüme olan uzaklığı derinliğini verir. Karar ağaçları en üstte bulunan kök düğümden aşağı doğru inmektedir [36].

Eğitim veri setini en iyi şekilde sınıflandırma için her dalı derinleştirilir. Her dal saf yapraklar elde edilene kadar derinleştirilirse, doğruluk oranı grafiğinde plato oluşmaya başlar ve aşırı uyum (overfitting) durumu ortaya çıkar. Bu durumun oluşmasında etken verilerde gürültü olmasıdır. Bu nedenle verilerde gürültünün azaltılması gerekir. Bunun için Budama (Pruning) işlemi yapılır.

Budama işleminde düşün önemdeki niteliklere sahip dallar kaldırılarak ağacın karmaşıklığı azaltılarak aşırı uyum düşürülür ve daha düşük maliyetle daha isabetli tahminlerde bulunmaya başlar. Budama işlemi ID3 algoritmasında yapılmazken C4.5 ve CART algoritmalarında sırasıyla Erken Budama (Pre-prunning) [37] ve Geç Budama (Post-prunning) [38] olmak üzere iki şekilde yapılır. Erken budamada verideki gürültü eğitim esnasında yapılırken geç budamada ağaç tamamlandıktan sonra aşırı uyum tespit edilirse yapılır [39] ve daha doğru bir yaklaşımdır.

3.2.4. Yapay Sinir Ağları

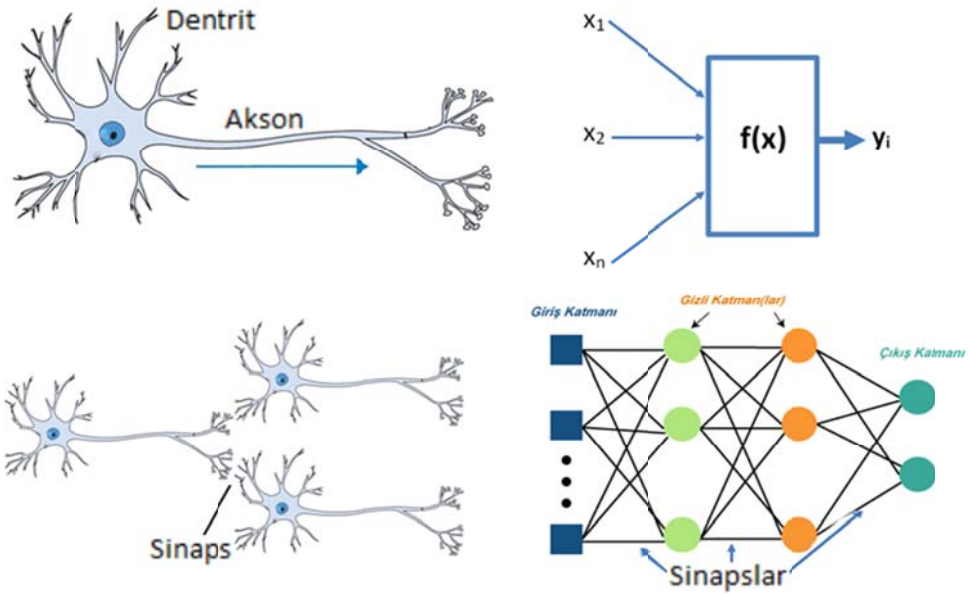
Yapay Sinir Ağları, biyolojik sinir ağları örnek alınarak matematiksel mantıkla oluşturulmuş yazılımsal yapılardır [40]. YSA'lar bilgisayar ya da bilgisayar destekli bir makine tarafından insanların kullandığı çözüm bulma, anlama, anlam çıkarma, geçmiş tecrübelerden öğrenme gibi bazı yüksek mantık gerektiren organizasyonel ilkelere dair görevleri yerine getirme kabiliyetidir [41]. Yapay nöron, Şekil 3.6'da gösterilen doğal nöronlardan esinlenen bir modeldir. Yapay nöronlardan oluşturulan bu nöron ağlarına yapay sinir ağları denir.



Şekil 4.6. Biyolojik nöron.

YSA'ların temel amacı bilgiyi işlemektir. Bu nedenle bilgi işleme alanlarının tamamında kullanılırlar. Gerçek sinir ağlarını modelleyerek örüntü analizi, tahmin ve veri sıkıştırma gibi mühendislik amaçları için kullanılan farklı YSA'lar oluşturulmuştur.

YSA'lar, şekil 3.7'de gösterildiği gibi doğal nöronların birbiri ile sinaptik bağlanması gibi birbirine hiyerarşik olarak bağlı ve paralel olarak çalışabilen yapay nöronlardan oluşmaktadır [42].



Şekil 4.7. Biyolojik sinir hücresi ve yapay sinir ağı.

Biyolojik sinir sistemi ile yapay sinir ağı arasındaki benzerlikler tablo 3.1'de gösterilmiştir.

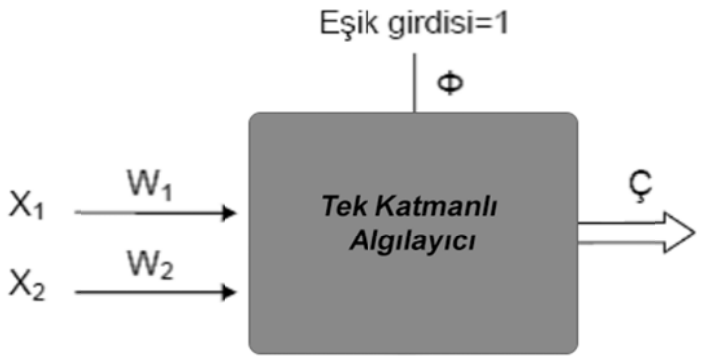
Çizelge 4.1. Biyolojik sinir sistemi ve yapay sinir ağı karşılıkları [42].

Biyolojik Sinir Sistemi	Yapay Sinir Ağları
Nöron	İşlemci Elemanı
Hücre Gövdesi	Toplama Fonksiyonu
Dentrit	Transfer Fonksiyonu
Akson	Yapay Nöron Çıkışı
Sinapslar	Ağırlıklar

YSA'lar tek katmanlı algılayıcılar, çok katmanlı algılayıcılar, ileri ve geri beslemeli YSA'lar olarak dört kategoride incelenebilir.

3.2.4.1. Tek Katmanlı Algılayıcılar

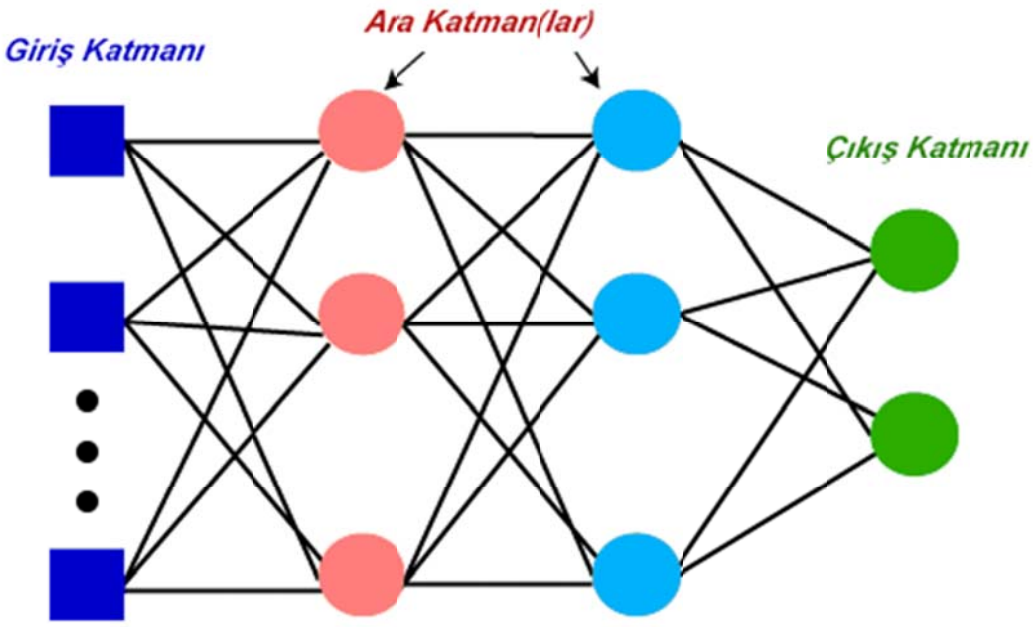
Tek katmanlı YSA'lar Şekil 3.8'de gösterildiği gibi, yalnızca girişten ve doğrusal bir fonksiyon ile çıktı veren bir çıkıştan oluşan sinir ağı türüdür.



Şekil 4.8. Tek katmanlı algılayıcı örneği

3.2.4.2. Çok Katmanlı Algılayıcılar

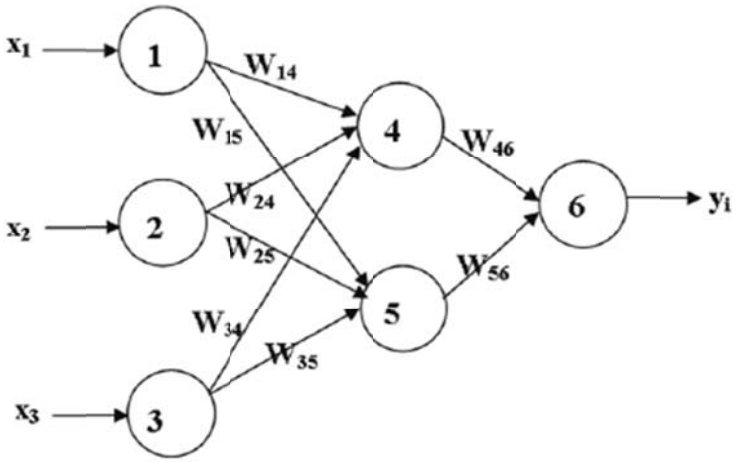
Doğrusal olmayan aktivasyon fonksiyonu kullanırlar, bir çok nöronun belli bir hiyerarşide birbirine bağlanmasıyla oluşurlar (Şekil 3.9).



Şekil 4.9. Çok katmanlı algılayıcı örneği.

3.2.4.3. İleri Beslemeli Yapay Sinir Ağları

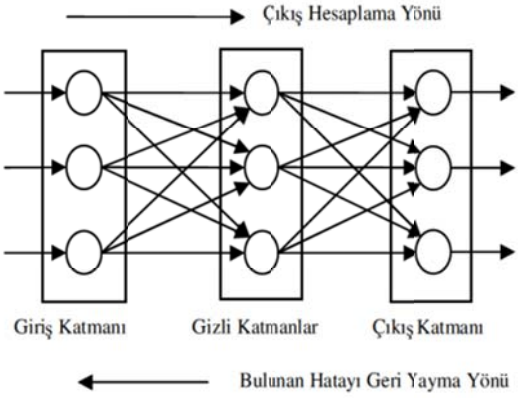
Bu tür YSA'larda nöronlar giriş yönünden çıkış yönüne doğru katmanlar halinde bağlanırlar. Şekil 3.10'da görüldüğü gibi, bir katmandan yalnızca bir sonraki katmana doğru bağlantı vardır. Girişe uygulanan veriler ara katmanlardan çıkış katmanına doğru işlenerek aktarılır.



Şekil 4.10. İleri beslemeli yapay sinir ağı örneği.

3.2.4.4. Geri Beslemeli Yapay Sinir Ağları

Bu tür ağlarda bir nöron çıkışı kendinden önceki, sonraki veya kendi katmanındaki başka bir nörona giriş olarak bağlanabilir (Şekil 3.11). Bu nedenle doğrusal olmayan dinamik bir davranış modeline sahiptir. Geri besleme bağlantı şekillerine göre aynı YSA ile farklı yapıda geri beslemeli YSA'lar elde edilebilir.



Şekil 4.11. Geri beslemeli ağ modeli.

YSA'lar günümüzün teknolojik gelişimine paralel olarak hayatımızın her alanına girmektedir. Sağlık, elektronik, meteoroloji, ekonomi ve finans sektörlerinin yanı sıra askeri teknolojilerde de oldukça yaygın kullanılmaktadır [43].

3.2.5. Sınıflandırıcı Performansının Değerlendirilmesi

Sınıflandırma sonucunun ve algoritma performansının belirlenmesinde çeşitli yaklaşımlar kullanılmaktadır [44]. Tüm sınıflandırma algoritmalarında bir algoritma diğer bir sınıflandırma algoritmasından daha iyi veya daha kötü denilemez. Bir veri seti üzerinde hangi algoritmanın daha iyi olduğunu belirleyen değerlendirme ölçütleri bulunmaktadır.

Bir sınıflandırma modelinde başarıyı değerlendirmek ve diğer modellerle karşılaştırmak için aşağıdaki ölçütler en sık kullanılan ölçütlerdir:

- Doğruluk / Hata Oranı
- Kesinlik
- Hassasiyet
- F-Skoru
- Özgüllük
- ROC Eğrisi

Bu değerlerin hesaplamasında karışıklık matrisi (Confusion Matrix) kullanılmaktadır. Karışıklık Matrisi, bir sınıflandırma sistemi tarafından yapılan gerçek ve tahmin edilen sınıflandırmalardan oluşmaktadır. Sınıflandırmanın performansı matristeki veriler kullanılarak değerlendirilmektedir [45].

Karışıklık matrisinde $CM_{i,j}$ değeri sınıflandırma algoritması tarafından "j" sınıfına dahil edilmiş "i" sınıfına ait örneklerin sayısını göstermektedir. İki tür sınıf değeri olan bir sınıflandırma probleminde karışıklık matrisi Tablo 3.1'deki gibi olacaktır.

Çizelge 4.2. Karışıklık matrisinin yapısı.

Gerçek Sınıf / Tahmin Edilen Sınıf	S_1	$-S_1$
S_1	**A: True Positive (TP)**	**B: False Negative (FN)**
$-S_1$	**C: False Positive (FP)**	**D: True Negative (TN)**

Karışıklık matrisinde A alanı, test verisinde S_1 sınıfında olup, kullanılan model tarafından da S_1 olarak tahmin edilen örnek sayısını gösterir. Bu alan, TP olarak adlandırılmaktadır.

Karışıklık matrisinde B alanı, test verisinde S_1 sınıfında olup kullanılan model tarafından S_1 olarak tahmin edilmeyen örnek sayısını gösterir. Bu alan, FN olarak adlandırılır.

Karışıklık matrisinde C alanı, test verisinde S_1 sınıfında olmayan, fakat kullanılan model tarafından S_1 sınıfı olarak tahmin edilen örnek sayısını gösterir. Bu alan, FP olarak adlandırılır.

Karışıklık matrisinde D alanı, test verisinde S_1 sınıfında olmayan ve kullanılan model tarafından da S_1 sınıfı olarak tahmin edilmeyen örnek sayısını gösterir. Bu alan, TN olarak adlandırılır.

3.2.5.1. Doğruluk

Doğruluk bir modelin başarısını vermektedir. Doğru tahminlerin tüm verilere oranlanması ile bulunur (Formül 3.10). Hata oranı ise yanlış tahminlerin tüm verilere oranı ile hesaplanır (Formül 3.11).

$$Doğruluk\ Oranı = \frac{TP + TN}{TP\ +\ FN\ +\ FP\ +\ TN} \tag{4.10}$$

$$Hata\ Oranı = \frac{FN + FP}{TP\ +\ FN\ +\ FP\ +\ TN} \tag{4.11}$$

3.2.5.2. Kesinlik

Kullanılan modelin doğru olarak sınıflandırdığı pozitif örnek sayısının, toplam pozitif örnek sayısına oranı ile hesaplanır. Başka bir deyişle ile tahminde S sınıfında çıkan örneklerin gerçekte S sınıfında olma oranıdır. Kesinlik, kullanılan modelin, FP'leri eleme yeteneğini gösterir ve Formül 3.12'de gösterildiği gibi hesaplanır.

$$Kesinlik = \frac{TP}{TP \ + \ FP} \tag{4.12}$$

3.2.5.3. Hassasiyet

Bu değer, kullanılan modelin isabetli bir şekilde sınıflandırdığı pozitif örneklerin oranını ölçer. Başka bir deyişle S sınıfında olarak isabetli tahmin edilmiş pozitif örneklerin, S sınıfındaki toplam örnek sayısına bölünmesiyle bulunur. Anma değeri kullanılan modelin, FN'leri eleme yeteneğini ölçer ve Formül 3.13'de gösterildiği gibi hesaplanır.

$$Anma = \frac{TP}{TP \ + \ FN} \tag{4.13}$$

Kesinlik ve Anma skor değerlerinde en iyi değer 1'dir.

3.2.5.4. F-Skor

Kullanılan modelin kesinlik ve anma değerlerinin harmonik ortalaması alınarak Formül 3.14'te gösterildiği hesaplanmaktadır.

$$F = \frac{2 \ x \ Kesinlik \ x \ Anma}{Kesinlik \ + \ Anma} \tag{4.14}$$

Buradaki kesinlik değerine anma değerlerine göre β kat ağırlık verilmek istenirse F-Skor'una ait eşitlik, Formül 3.15'teki gibi şekillenecektir.

$$F = \frac{(1 + \beta^2)\ x\ Kesinlik\ x\ Anma}{\beta^2\ x\ Kesinlik\ +\ Anma} \tag{4.15}$$

3.2.5.5. Özgüllük

Bu değer, veri setinde S sınıfına ait olmayan örneklerin, S sınıfına ait olmadığının isabetli bir şekilde tahmin edilme oranıdır. Bu değer, true negative sınıflandırma başarısını gösterir (Formül 3.16).

$$Spesifiklik = \frac{TN}{FP\ +\ TN} \tag{4.16}$$

Buraya kadar açıklanan "Doğruluk", "Kesinlik", "Anma", "Spesifiklik" ve "F-Skor" değerleri yüksek olan model, bu değerleri düşük olan diğer modellere göre daha başarılıdır. Yeni örneklerin tahmini için bu model tercih edilir.

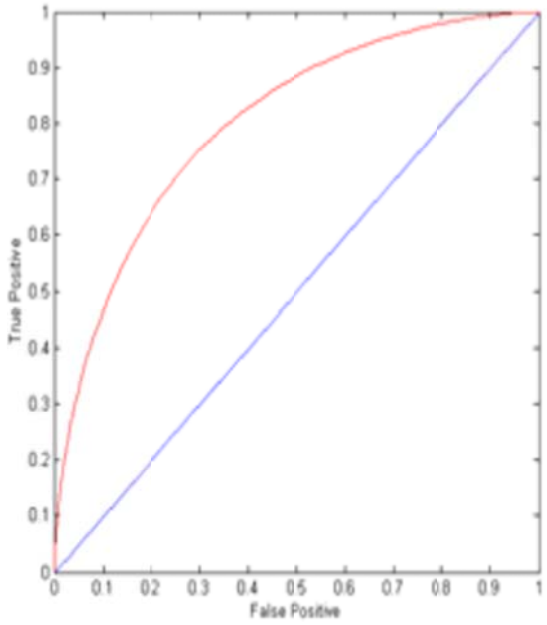
3.2.5.6. ROC Eğrisi

Kullanılan sınıflandırma modellerinin karşılaştırılması amacıyla kullanılır. ROC eğrisinin altında kalan alanın büyüklüğü, kullanılan modelin başarısını göstermektedir.

1967 yılında Lusted tarafından önerilen ROC eğrisi, 1969 yılında medikal görüntüleme alanında kullanılmaya başlanmıştır ve kullanımın yaygınlaşmasıyla birlikte tıp alanında hastalık teşhislerinde ROC eğrisi analizinin kullanımı artış göstermektedir [46]. ROC eğrisi y eksenindeki TP (hassasiyet) ve x eksenindeki FP (özgüllük) oranının çizilmesi ile elde edilmektedir. Kısacası, TP değerinin FP değerine oranı ROC eğri grafiğini vermektedir. Şekil 3.12'de grafik üzerinde ROC eğrisinin altında kalan alan (AUC) değeri 0'dan 1'e doğru yaklaşması pozitif değerlerin negatif değerlerden başarılı bir şekilde ayrıldığını göstermektedir ve tanı değeri yükselmektedir [47].

Bu eğriye göre, uç değerler aşağıdaki şekilde ifade edilebilir:

- (0,0): Bütün örneklerin negatif sınıflandırılması.
- (1,1): Bütün örneklerin pozitif sınıflandırılması.
- (0,1): İdeal durum.
- (1,0): tüm durumların hatalı tahmin edilmesi.



Şekil 4.12. Örnek ROC eğrisi.

3.2.6. Çapraz Doğrulama

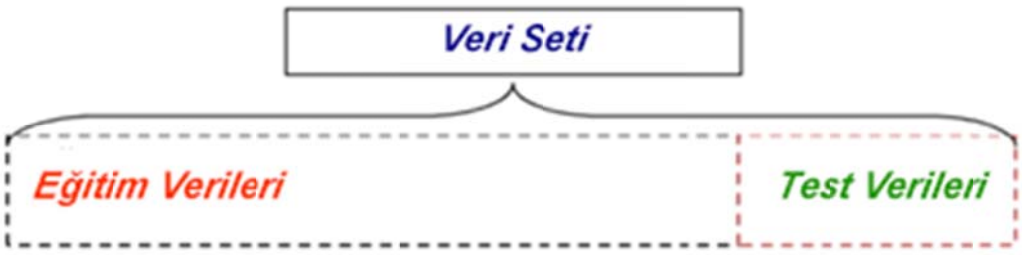
Veri setinin eğitim ve test verisi olarak bölünmesi kullanılacak olan modelin başarısına doğrudan etki eden bir işlemdir. Bir model değerlendirme yöntemi olan çapraz doğrulamada, öğreticinin hiç görmediği veriler hakkında tahmin yapması istenir ve onun bu tahmini ne kadar iyi yapacağını gösterir [48].

Kullanılan modelin eğitiminde veri setinin tamamı kullanılmaz. Veri seti eğitim verisi ve test verisi olarak ikiye ayrılır. Model öğrenmesi ve parametrelerin tahmini için eğitim verisi kullanılır. Kullanılan modelin etkinliğini doğrulamak için ise test verisi kullanılır [49]. Literatürde çapraz doğrulama için farklı yöntemler önerilmiştir [50–53], önerilen yöntemlerin büyük bir kısmının temeli benzerdir [54]. Çapraz doğrulama yönteminin temelinde, bütün veri seti eşit sayıda örneğe sahip “k” adet

gruba ayrılır. Bu grupların k-1 adedi rastgele seçilir ve kullanılan modelin eğitiminde kullanılır. Kalan veri grubu ise modelin tahmin isabetini ölçmede kullanılır. Holdout yöntemi, k kat çapraz doğrulama, biri hariç çapraz doğrulama ve rastgele alt-örnekleme yöntemleri, alanda en çok kullanılan dört yöntemdir.

3.2.6.1. Holdout Yöntemi

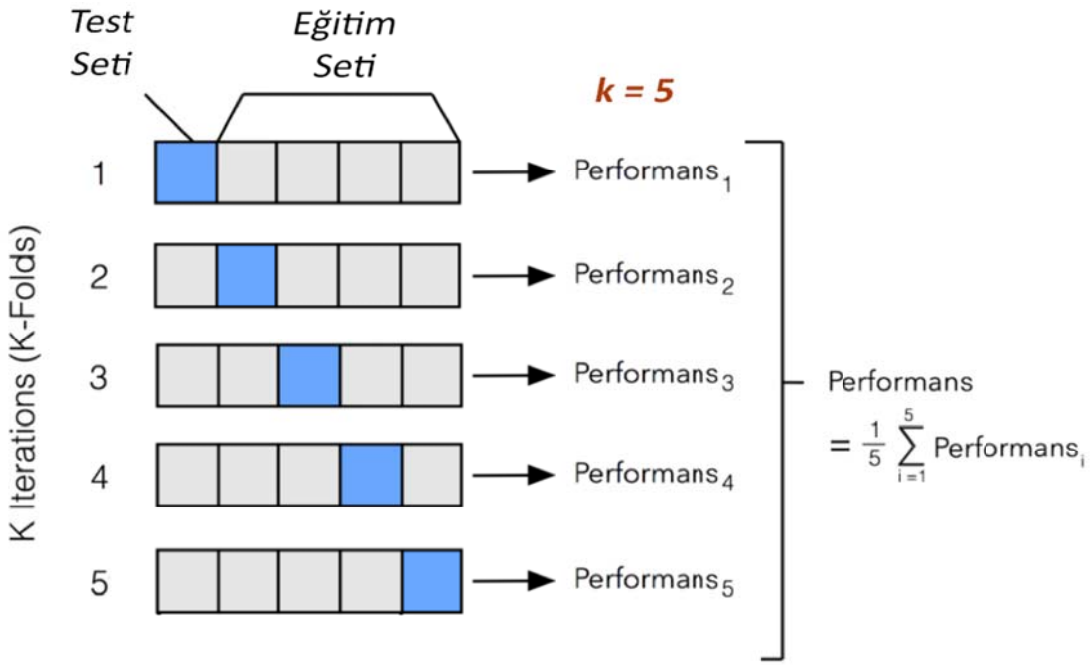
Çapraz doğrulamanın en temel şeklidir. Bir kısım veri doğrulama için kullanılır, kalan veriler eğitim için kullanılır. Şekil 3.13'te görüldüğü gibi, veri setinin 2/3'ü eğitim verisi, 1/3'ü doğrulama verisi olarak ayrılır. Hesaplama yönünden fazla yük getirmemektedir [55].



Şekil 4.13. Holdout yöntemi ile veri setinin bölünmesi örneği.

3.2.6.1. K-Kat Çapraz Doğrulama

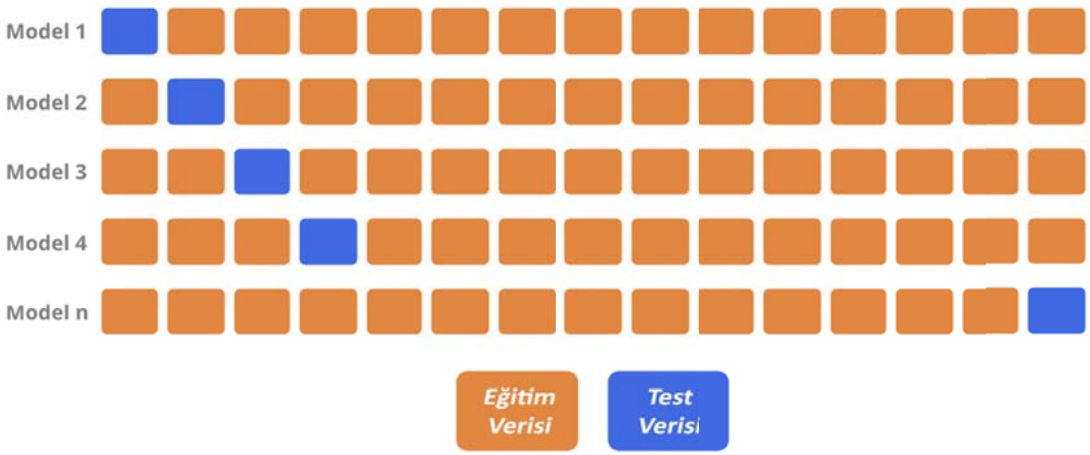
Holdout yönteminin geliştirilmesiyle oluşturulan k-fold yönteminde veri seti, "k" adet alt kümeye ayrılır ve holdout yöntemi "k" defa tekrarlanır. Her tekrarda bir küme test veri seti, kalan k-1 adet küme de eğitim veri seti olarak tekrar birleştirilir. "k" defa tekrarlanan her eğitimin için hata değerleri hesaplanır. Hesaplanan değerlerin ortalaması alınarak bir hata değeri bulunur. Bu işlem şekil 3.14'te gösterildiği gibi, k defa tekrarlandığı için oldukça maliyetli bir işlemdir. Veri seti alt kümlerinin boyutu ve tekrarlama değeri tercihe bağlıdır. Bu da yönteme bir avantaj getirmektedir [56].



Şekil 4.14. K-kat çapraz doğrulama.

3.2.6.1. Biri Hariç Çapraz Doğrulama Yöntemi

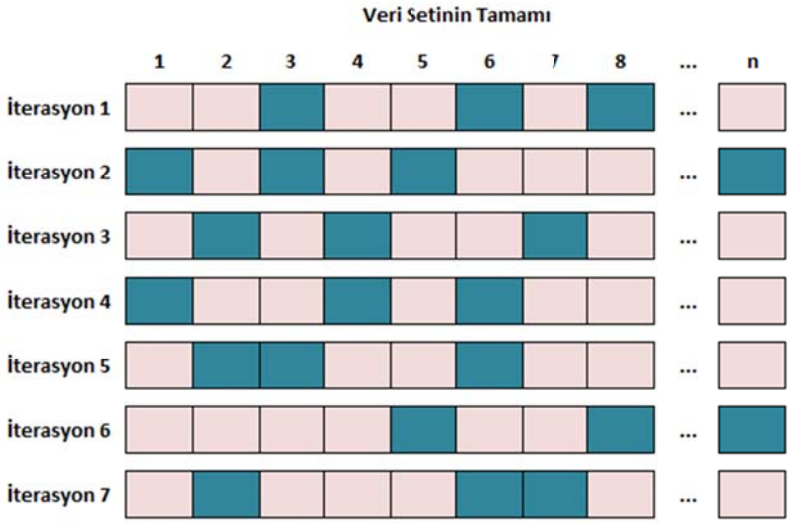
K-kat çapraz doğrulama yönteminin özelleştirilmiş şeklidir. "n" adet verinin biri test diğerleri eğitim için kullanılır. Bu işlem şekil 3.15'te gösterildiği gibi, her veri test için 1 kez kullanılmak üzere "n" defa tekrarlanır. Her iterasyon için bir hata değeri hesaplanır. İşlem sonunda hataların ortalaması metodun nihai hata değerini verir.



Şekil 4.15. Biri hariç çapraz doğrulama yöntemi örneği.

3.2.6.1. Rastgele Alt-Örnekleme Yöntemi

Rastgele alt-örneklemede, belli sayıda veriler rasgele olarak eğitim ve test için seçilir. Aynı veri hem eğitim hem de test veri setinde bulunamaz. Şekil 3.16'da da görüldüğü gibi veriler rastgele seçilir ve her iterasyon için ayrı ayrı hata değeri hesaplanır.



Şekil 4.16. Rastgele alt-örnekleme yöntemi örneği.

BÖLÜM 4

VERİ TOPLAMA VE ÖN HAZIRLIK

Bu tez çalışmasında, MEB HBÖGM bağlı AÖL öğrencilerinin normal eğitim süresi içerisinde mezun olma ve olamama durumlarını tahmin etmede yapay zekâ tekniklerinin etkinlikleri incelenmiştir. Bu amaçla, yapay zekâ tekniklerinden karar ağaçları (KA), k-en yakın komşuluk (KNN), destek vektör makineleri (DVM), yapay sinir ağları (YSA) modellerinin yetenekleri test edilmiştir.

4.1. ÇALIŞMANIN SINIRLILIKLARI VE EVRENİ

Öğrencilerin tamamı AÖL'ye ortaokul mezunu olarak kayıt yaptıran öğrencilerdir. Tasdikname ile ara sınıftan kayıt yaptıran öğrenciler verilerinin doğruluğu onaylanmış olmadığından dolayı alınmamıştır. Belge ve bilgileri onaylanmış ve mezun olan sayısının miktarı açısından yalnızca 2010-2012 yılları arası kayıt yaptırmış olan öğrenciler seçilmiştir. Ayrıca verilerin büyüklüğü nedeniyle Türkiye geneli 81 il ve ilçelerinin tamamı alınmamıştır. Bunun yerine ülke genelini temsil edebilecek, ülkenin 7 bölgesinden ikişer tane il ve ilçeleriyle birlikte tüm sosyo-ekonomik seviyeleri barındıran 14 il seçilmiştir.

Bu sınırlılıklar çerçevesinde, Milli Eğitim Bakanlığı, Bilgi İşlem Genel Müdürlüğü'nden öğrencilerin aşağıdaki bilgilerini içeren 142.714 öğrenciye ait veriler Excel ortamında alınmıştır;

1. Cinsiyet,
2. Doğum yılı,
3. Kayıt Yılı,
4. Kayıt İl ve İlçesi,
5. Başarılı olduğu toplam ders kredisi,

6. Devam ettiği toplam dönem sayısı,
7. Özel durumu,
8. Engel grubu,
9. Devam ettiği her bir dönem için seçtiği kredi miktarı
10. Devam ettiği her dönem başarılı olduğu kredi miktarı
11. Mezuniyet durumu

4.2. VERİ SETİNİN ÖN İŞLEMESİ

Alınan bu bilgiler, Oracle veri tabanı programına aktarılarak ön işlemeye tabi tutulmuş ve metinsel bilgiler sayısallaştırılmıştır. İşlenmiş veriler Oracle veri tabanından tekrar Excel ortamına export edilerek Matlab programında deneysel çalışmalar yapılmıştır. İşlem adımları şekil 4.9'da gösterilmiştir.

Şekil 5.1. Veri ön-işleme işlem adımları.

4.3. VERİ SETİNİN ÖZELLİKLERİ

AÖL öğrencilerinin yaşadıkları ilçelerin sosyo-ekonomik seviyeleri ve sosyo-ekonomik skorları öğrencilerin verilerine eklenmiştir. Veri setindeki parametrelerden sosyo-ekonomik seviye ve sosyo-ekonomik skor bilgileri normalize edilmiştir, diğer veriler kategorik veri olarak bırakılmıştır. Normalizasyon, verileri 0-1 aralığında yeniden ölçeklendirmektir [57].

Özellikler belirlenirken, daha önceden bu alanda yapılmış çalışmalar incelenerek referans alınmıştır. Ham değerleri sayısal verilere dönüştürmek için kapsamlı ön

işleme yapılmıştır. Metinsel bilgi olan özel durum ve engel durumu verilerinin sayısal karşılıkları üretilmiştir.

Toplanan özellikler şu şekilde işlenmiştir:

- Demografik bilgiler: Kayıt esnasındaki yaş, cinsiyeti, yaşadığı il ve ilçe.
- Yaşadığı ilçenin sosyo-ekonomik seviyesi ve sosyo-ekonomik skoru.
- Akademik: Başarılı olduğu toplam ders kredisi ve devam ettiği toplam dönem sayısı, devam ettiği dönemlerde seçtiği kredi bilgisi, bu dönemlerde başardığı kredi bilgisi
- Öğrencinin özel durumu ve engel durumu.

Veri setinde K (Kadın), E (Erkek) olarak gelen öğrencinin cinsiyeti erkek için 1 ve kadın için 2 olarak sayısallaştırılmıştır.

Kullanılan veri setinde özel durumlar metinsel bilgi olarak geldiğinden Tablo 4.1'deki sayısal karşılıklarına dönüştürülmüştür.

Çizelge 5.1. Özel durum bilgisinin sayısal karşılıkları.

Kod	Özel Durum
1	Yok
2	Denetimli Serbestlik Kapsamında
3	Gazi / Gazi Eş, Çocuk, Anne veya Babası
4	İl/İlçe Özel Eğitim Hizmetleri Kurul Kararı Var
5	Sosyal Hizmetler Ve Çocuk Esirgeme Kurumunda Kalıyor
6	Şehit Eş, Çocuk, Anne, Baba veya Kardeşi
7	Tutuklu Veya Hükümlü
8	%40 ve Üzeri Engelli
9	5395 Sayılı Çocuk Koruma Kanunu Kapsamında

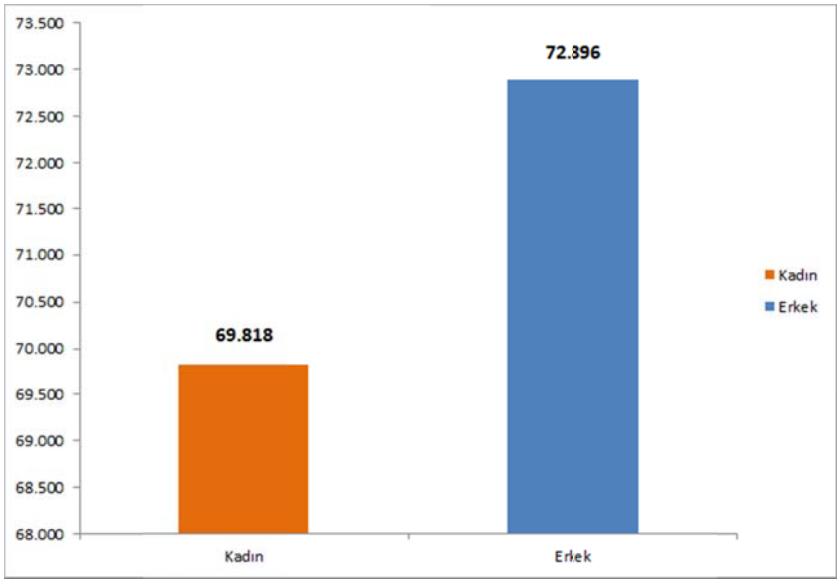
Engel durumları da yine metinsel veri olarak geldiğinden, Tablo 4.2'deki sayısal karşılıklarına dönüştürülmüştür.

Çizelge 5.2. Engel durumu bilgisinin sayısal karşılıkları.

Kod	Engel Türü
1	Herhangi Bir Özrü Yok
2	Görme Engelli
3	Hafif Zihinsel Engelli
4	İşitme Engelli
5	Konuşma Engelli
6	Ortopedik Engelli - Üst Beden Kullanamıyor
7	Ortopedik Engelli - Alt Beden Kullanamıyor
8	Ortopedik Engelli - Alt ve Üst Beden Kullanamıyor
9	Ruhsal ve Duygusal Bozukluk
10	Yatarak Tedavi Görüyor

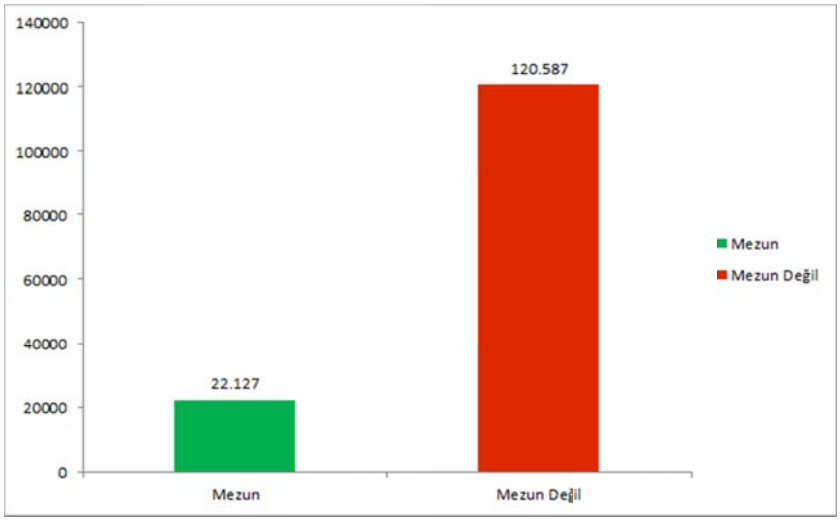
4.4. VERİ SETİNİN ANALİZİ

Mezuniyet tahmini, 69.818 kadın ve 72.896 erkek olmak üzere toplam 142.714 öğrenciden oluşturulmuştur (Şekil 4.2)



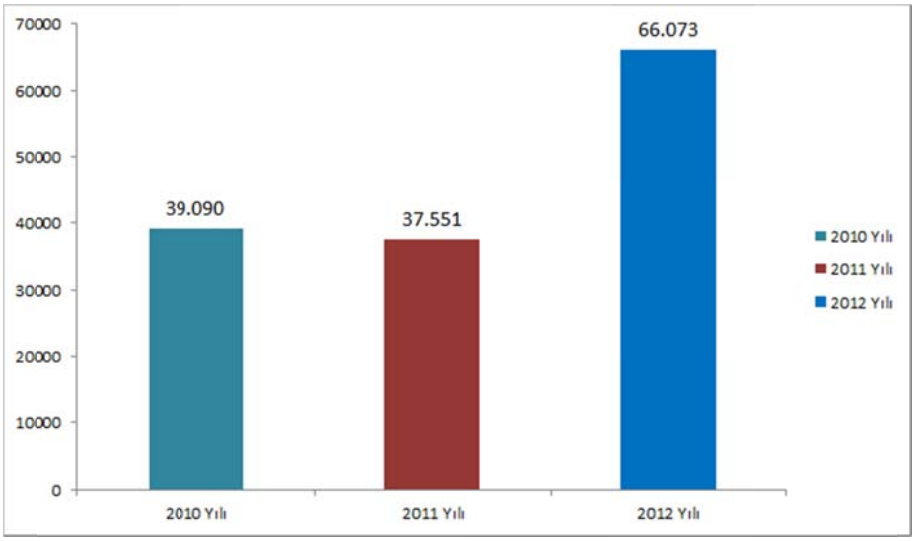
Şekil 5.2. Cinsiyete göre öğrenci dağılımı.

Veri setinde 22.127 öğrenci mezun, 120.587 öğrenci ise mezun olamamış durumdadır (Şekil 4.3).



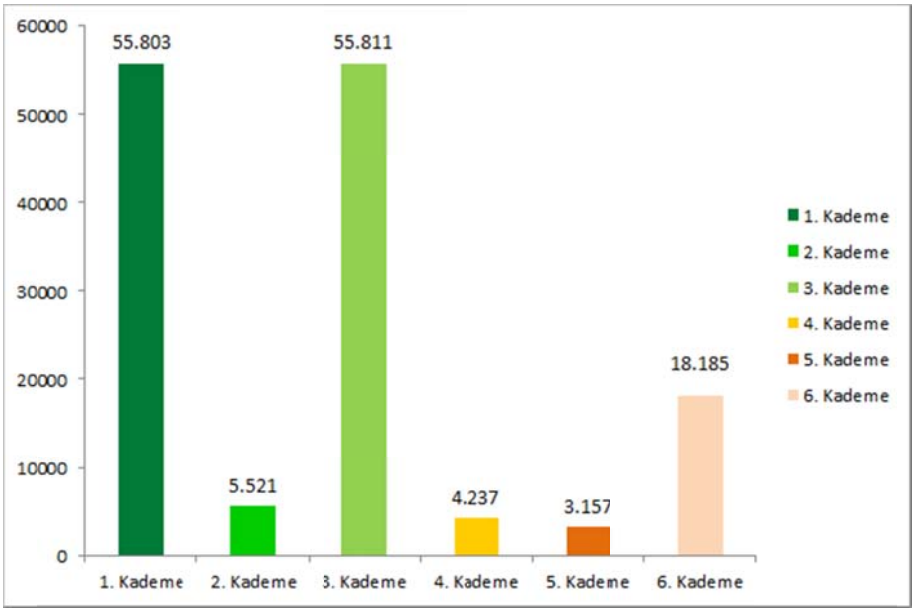
Şekil 5.3. Mezuniyet durumlarına göre öğrencilerin dağılımı.

Veri setinde 39.090 öğrenci 2010 yılında, 37.551 öğrenci 2011 yılında, 66.073 öğrenci 2012 yılında kayıt yaptırmıştır (Şekil 4.4).



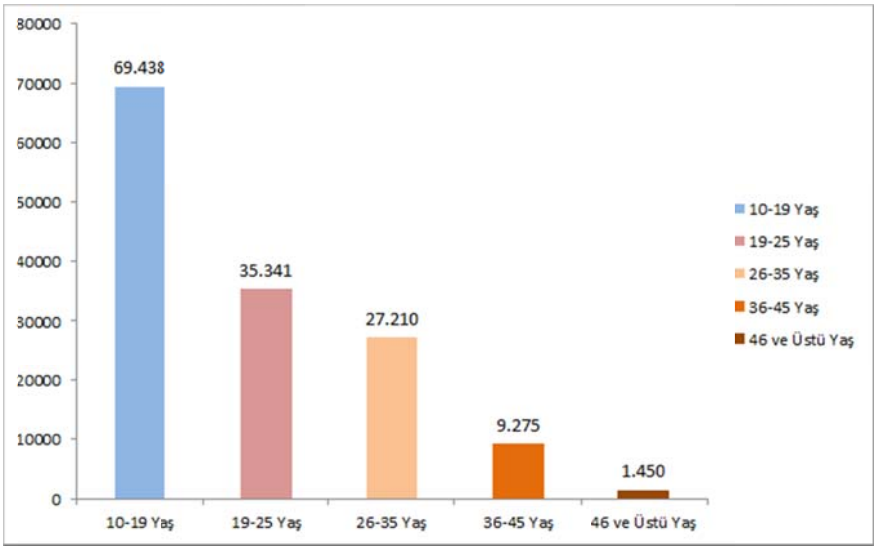
Şekil 5.4. Kayıt yılına göre öğrencilerin dağılımı.

Sosyo-ekonomik seviyelere göre 55.803 öğrenci 1. Kademe, 5.521 öğrenci 2. Kademe, 55.811 öğrenci 3. Kademe, 4.237 öğrenci 4. Kademe, 3.157 öğrenci 5. Kademe, 18.185 öğrenci 6. Kademe şehirlerde yaşamaktadır (Şekil 4.5).



Şekil 5.5. Sosyo-ekonomik seviyelerine göre öğrencilerin dağılımı.

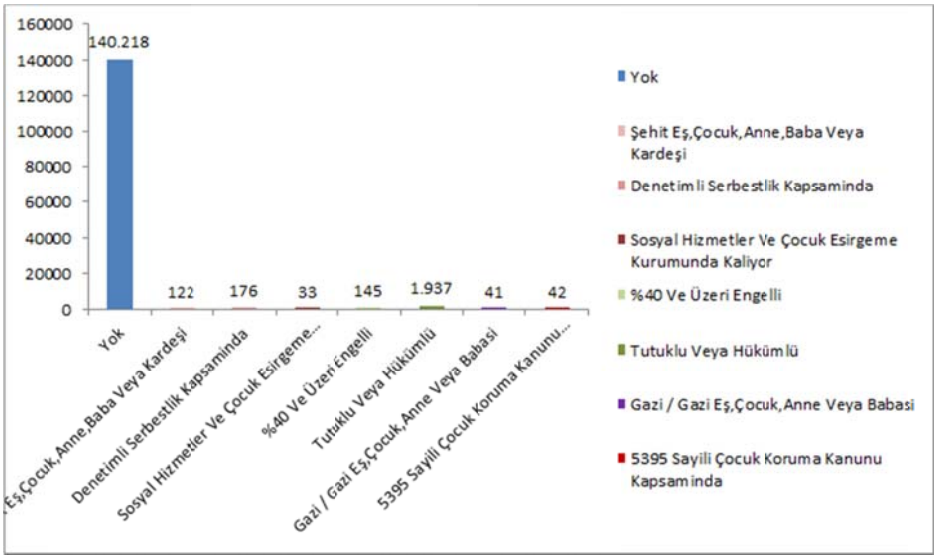
Yaş gruplarına göre 69.438 öğrenci 10-19 yaş, 5.521 öğrenci 19-25 yaş, 55.811 öğrenci 26-35 yaş, 4.237 öğrenci 36-45 yaş, 3.157 öğrenci 46 ve üzeri yaş grubuna dağıldıkları görülmüştür (Şekil 4.6).



Şekil 5.6. Yaş gruplarına göre öğrencilerin dağılımı.

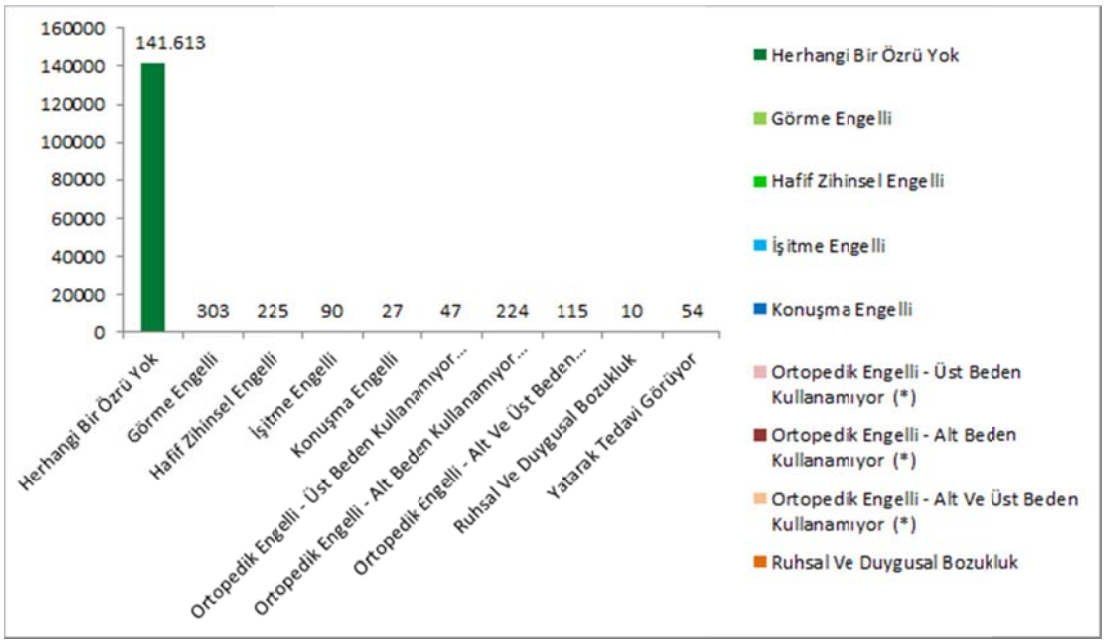
Özel durumlarına göre 140.218 öğrencinin herhangi bir özel durumu yokken, 122 öğrenci şehit yakını, 176 öğrenci denetimli serbestlikte, 33 öğrenci çocuk esirgeme kurumunda kalmakta, 145 öğrenci %40 ve üzeri engelli, 1.937 öğrenci tutuklu ve

hükümlü, 41 öğrenci gazi veya yakını, 42 öğrenci ise 5395 sayılı çocuk koruma kanunu kapsamındadır (Şekil 4.7).



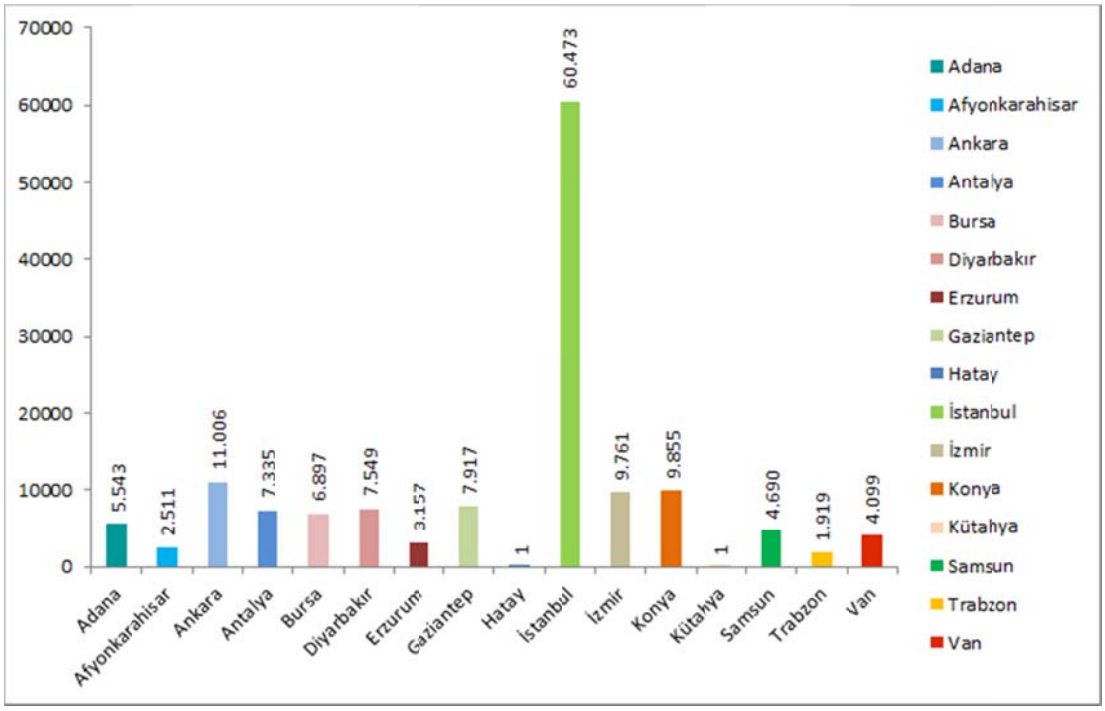
Şekil 5.7. Özel durumlarına göre öğrencilerin dağılımı.

Engel durumlarına göre 141.613 öğrencinin herhangi bir engeli yokken 1.095 öğrenci ise engelli öğrencidir (Şekil 4.8).



Şekil 5.8. Engel grubuna göre öğrencilerin dağılımı.

Yaşadıkları illere göre 5.543 öğrenci Adana'da, 2.511 öğrenci Afyonkarahisar'da, 11.006 öğrenci Ankara'da, 7.335 öğrenci Antalya'da, 6.897 öğrenci Bursa'da, 7.549 öğrenci Diyarbakır'da, 3.157 öğrenci Erzurum'da, 7.917 öğrenci Gaziantep'te, 1 öğrenci Hatay'da, 60.473 öğrenci İstanbul'da, 9.761 öğrenci İzmir'de, 9.855 öğrenci Konya'da, 1 öğrenci Kütahya'da, 4.690 öğrenci Samsun'da, 1.919 öğrenci Trabzon'da, 4.099 öğrenci de Van'da yaşamaktadır (Şekil 4.9).



Şekil 5.9. Yaşadıkları illere göre öğrencilerin dağılımı.

4.5. DENEYSEL TEST ORTAMI

Çalışmamızda veri setimiz üzerinde üç farklı deneysel test yapılmıştır. Bu üç deneysel testin tamamında aşağıdaki dört farklı yapay zeka teknikleri kullanılmıştır.

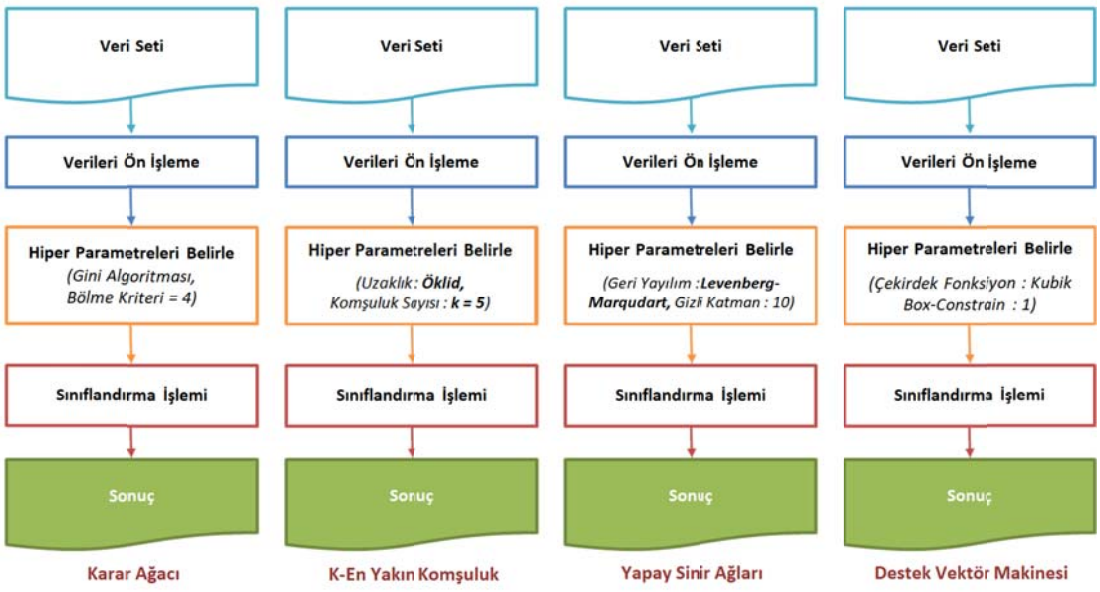
1. Karar Ağacı.
2. K-En Yakın Komşuluk.
3. Destek Vektör Makineleri.
4. Yapay Sinir Ağları.

4.5.1. Sınıflandırma İşlemleri

Yapılan ilk deneysel çalışma olarak sınıflandırma işleminde, öğrencilerin normal süresi içerisinde mezun olup olamama durumlarının tahmin etmek amaçlanmıştır. Bu amaçla veri setinde bulunan öğrencilerin yaşı, cinsiyeti, sosyo-ekonomik seviyesi, sosyo-ekonomik skoru, başarılı olduğu toplam ders kredisi, devam ettiği toplam dönem sayısı, özel durumu, engel durumu, mezuniyet durumu bilgileri kullanılmıştır. Veri setindeki 142.714 adet verinin %70'i eğitim %30'u test verisi olacak şekilde rastgele bölünerek ayrılmıştır.

KA algoritması için bölünme kriteri olarak Gini algoritması, maksimum bölme kriteri olarak 4 değeri seçilmiştir. KNN algoritması için uzaklık hesaplama fonksiyonu olarak Öklid Fonksiyonu, komşuluk değeri olarak k=5 değeri seçilmiştir. YSA algoritması için geri yayılım algoritması olarak Lavenberg-Marqudart, gizli katman sayısı olarak 5 değeri belirlenmiştir. DVM algoritması için Kubik çekirdek fonksiyonu kullanılmış, box-constrain değeri olarak 1 değeri belirlenmiştir.

Yapılan deneysel çalışmaya ait işlem adımları ve seçilen hiper parametreler Şekil 4.10'da gösterilmiştir.



Şekil 5.10. Sınıflandırma işlem adımları.

4.5.2. Model Hatalarının Tespiti

Yapılan ikinci deneysel çalışmada model hatalarının tespiti amacıyla birinci deneysel çalışma aynı hiper parametreler kullanılarak tekrar edilmiştir. Bu çalışmada farklı olarak k-katlı çapraz doğrulama yöntemi kullanılmıştır. k katlı çapraz doğrulama yönteminde k = 10 olarak belirlenmiştir. Yapılan deneysel çalışmaya ait işlem adımları Şekil 4.11'de gösterilmiştir.

Şekil 5.11. K-katlı çapraz doğrulama işlem adımları.

4.5.3. Mezun Olunabilecek Dönem Tahmini

Üçüncü deneysel çalışmada ise, yalnızca mezun olan öğrenciler çalışmaya dahil edilmiştir. Veri setinde bulunan 22.128 öğrencinin yaşı, cinsiyeti, sosyo-ekonomik seviyesi, sosyo-ekonomik skoru, başarılı olduğu toplam ders kredisi, devam ettiği toplam dönem sayısı, özel durumu, engel durumu, mezuniyet durumu bilgileri yanında devam ettikleri dönemlerde seçtikleri kredi sayısı ve başarılı oldukları kredi sayısı da veri setine dahil edilmiş ve öğrencilerin kaç dönemde mezun olabileceklerinin tahmini için çalışmalar yapılmıştır. Bu deneysel çalışmada yine birinci deneysel çalışmadaki aynı yapay zeka teknikleri yine aynı hiper parametrelerle kullanılmıştır.

BÖLÜM 5

BULGULAR VE TARTIŞMA

Yapılan ilk iki deneysel çalışmada, AÖL öğrencilerine ait yaş, cinsiyet, yaşadıkları il, ilçe, yaşadıkları şehrin sosyo-ekonomik seviye, sosyo-ekonomik skor, özel durum, engel durumu, bugüne kadar toplam devam ettikleri dönem, başarılı oldukları toplam kredi ve mezuniyet durumu bilgilerini içeren veri seti alınmıştır. Veri setinde bulunan 142.714 adet verinin %70'i eğitim ve %30'u test verisi olacak şekilde rasgele bölünerek ayrılmıştır.

Birinci deneysel çalışmada, bu veriler üzerine öğrencilerin normal süre içerisinde mezun oluş olamayacaklarına dair tahminde bulunan modeller test edilmiştir. İkinci deneysel çalışmada aynı şartlar altında 10 katlı çapraz doğrulama yöntemi ile sınıflandırma sonuçlarının hataları test edilmiştir.

Üçüncü deneysel çalışmada ise yanı veri setinde bulunan ve yalnızca mezun olmuş olan 22.128 adet öğrenci ele alınmıştır. Veri setine, ilk iki deneyde kullanılan özelliklere ek olarak devam ettikleri dönemlerde seçtikleri kredi ve başarılı oldukları kredi bilgileri de eklenmiştir. 22.128 adet verinin %70'i eğitim ve %30'u test verisi olacak şekilde rastgele bölünmüştür. Öğrencilerin kaç dönemde mezun olabileceklerinin tahminine yönelik üçüncü bir çalışma yapılmıştır.

Deneysel çalışmaları gerçekleştirmek için Matlab 2021a programı ve Matlab Classification Learner Tool kullanılmıştır. Etkili bir sınıflandırma işlemi gerçekleştirmek amacıyla makine öğrenme algoritmaları için en etkili hiper parametreler Classification Learner Tool ile belirlenmiştir. Hiper-parametreler, veri setine göre değişiklik gösterirler [58].

Yapılan deneysel çalışmalarda elde edilen sonuçlar ayrı ayrı sunulmuştur.

5.1. SINIFLANDIRMA TESTİ

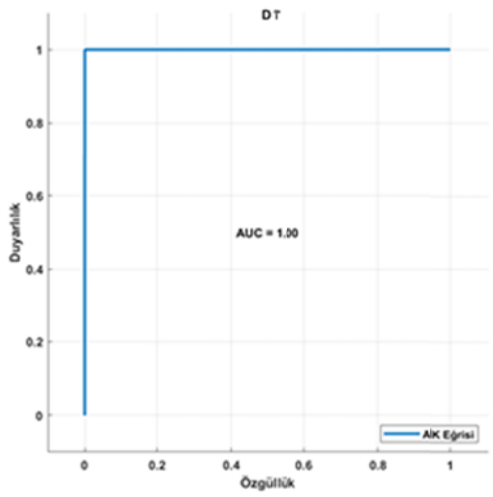
Sınıflandırma işlemi sonucunda en iyi sonuçlar Karar ağacı algoritması ile belirlenmiştir. KA sınıflandırma oranı %99.99 olarak belirlenirken, Hassaslık, Özgüllük, Kesinlik ve F-Skor değerleri sırasıyla %99.99, %99.98, %100 ve %99.99 olarak belirlenmiştir. Deneysel çalışmalar sonucunda makine öğrenme algoritmalarından elde dilen bulgular Tablo 5.1'de verilmiştir.

Çizelge 6.1. %70 Eğitim, %30 Test.

	Doğruluk	**Hassasiyet**	**Özgüllük**	**Kesinlik**	**F-Skor**
DT	**99.99**	**99.99**	**99.98**	**100**	**99.99**
KNN	99.56	99.65	99.02	99.82	98.74
YSA	99.97	99.98	99.94	99.99	99.98
DVM	99.90	99.95	99.64	99.93	99.63

KA algoritmasına ait karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.1'de verilmiştir.

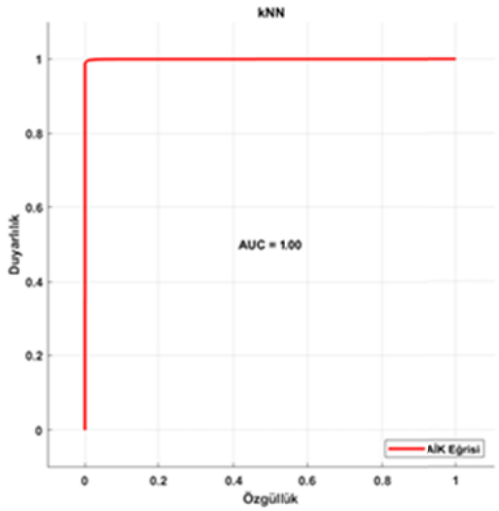
	Mezun Değil	**Mezun**
Mezun Değil	36.171	5
Mezun	1	6.637



Şekil 6.1. KA karmaşıklık matrisi ve ROC eğrisi

KNN algoritmasında %99.56 doğruluk oranı bulunmuştur. KNN karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.2'de verilmiştir.

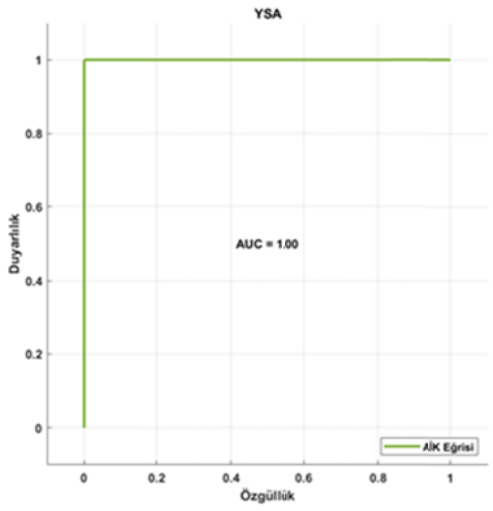
	Mezun Değil	**Mezun**
Mezun Değil	36.051	125
Mezun	65	6.573



Şekil 6.2. KNN karmaşıklık matrisi ve ROC eğrisi

YSA algoritmasında %99.97 doğruluk oranı bulunmuştur. YSA karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.3'te verilmiştir.

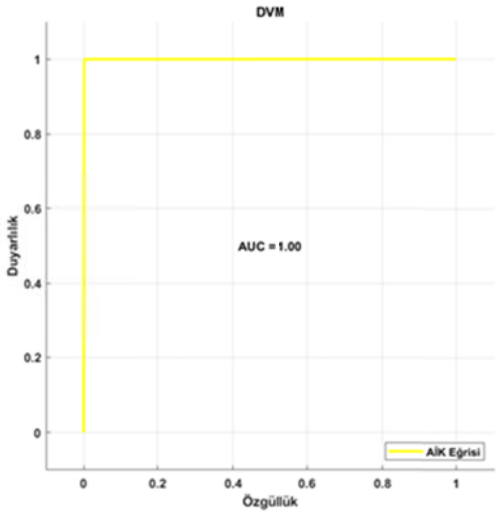
	Mezun Değil	**Mezun**
Mezun Değil	36.171	5
Mezun	1	6.637



Şekil 6.3. YSA karmaşıklık matrisi ve ROC eğrisi

DVM algoritmasında %99.90 doğruluk oranı bulunmuştur. DVM karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.2'de verilmiştir.

	Mezun Değil	Mezun
Mezun Değil	36.051	125
Mezun	65	6.573



Şekil 6.4. DVM karmaşıklık matrisi ve ROC eğrisi

5.2. SINIFLANDIRMA HATALARININ TESPİTİ

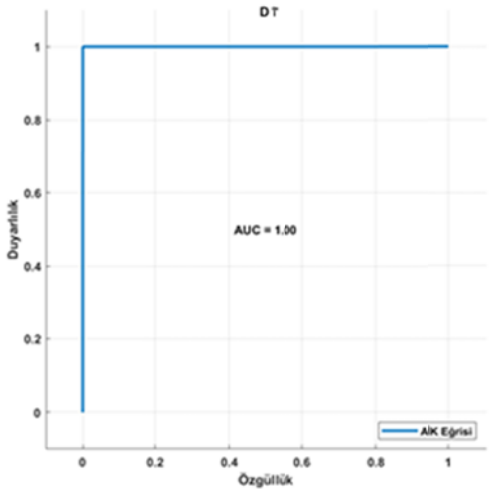
İkinci deneysel çalışmada kullanılan makine öğrenme algoritmalarının model hatalarının daha iyi tespit edilebilmesi için çapraz doğrulama yöntemi kullanılmıştır. Tüm makine öğrenme algoritmaları aynı hiper parametreler kullanılarak k = 10 alınarak k-kat çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama yöntemi sonucunda elde edilen bulgulara bakıldığı zaman en iyi sonuçların KA algoritması ile elde edildiği gözlemlenmiştir. Doğruluk oranı %99.98 olarak belirlenmişken, hassasiyet, özgüllük, kesinlik ve F-skor değerleri sırasıyla % 99.98, % 100, %100 ve %99. 99 olarak belirlenmiştir. Çapraz doğrulama sonucu elde dilen bulgular tablo 5.2'de verilmiştir.

Çizelge 6.2. 10-katlı çapraz doğrulama %70 Eğitim, %30 Test.

	Doğruluk	Hassasiyet	Özgüllük	Kesinlik	F-Skor
DT	99.98	99.98	100	100	99.99
KNN	99.54	99.65	98.97	99.81	99.73
YSA	99.98	99.98	99.96	99.99	99.99
DVM	99.92	99.96	99.74	99.95	99.95

KA algoritmasına ait 10 katlı çapraz doğrulama karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.5'te verilmiştir.

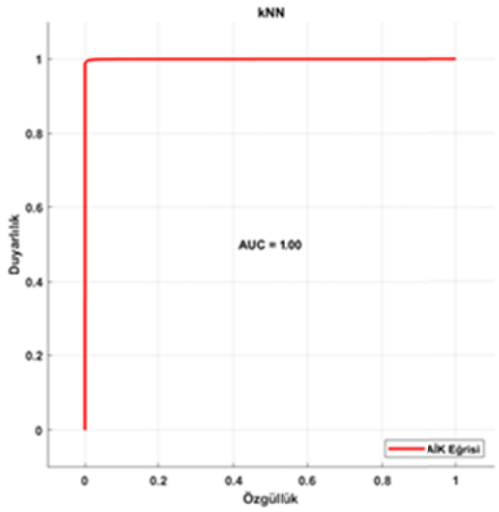
	Mezun Değil	**Mezun**
Mezun Değil	120.565	22
Mezun	1	22.126



Şekil 6.5. KA karmaşıklık matrisi ve ROC eğrisi

KNN algoritmasında %99.56 doğruluk oranı bulunmuştur. KNN 10 katlı çapraz doğrulama karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.6'da verilmiştir.

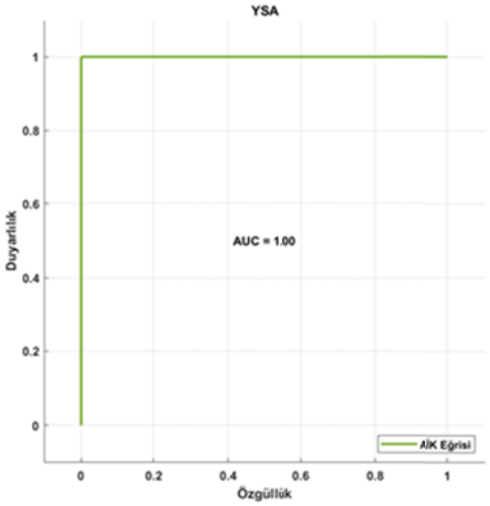
	Mezun Değil	**Mezun**
Mezun Değil	120.160	427
Mezun	227	21.900



Şekil 6.6. KNN karmaşıklık matrisi ve ROC eğrisi

YSA algoritmasında %99.97 doğruluk oranı bulunmuştur. YSA 10 katlı çapraz doğrulama karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.7'de verilmiştir.

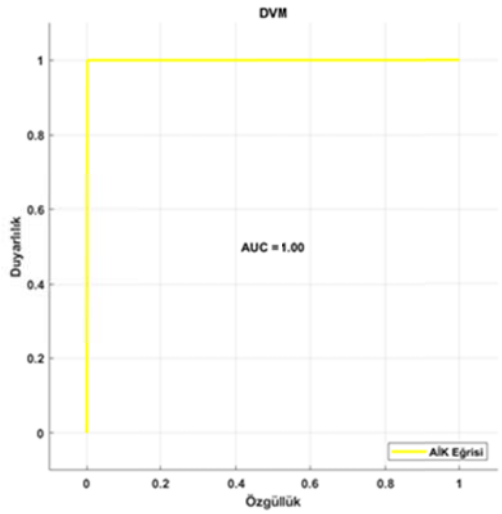
	Mezun Değil	**Mezun**
Mezun Değil	120.560	27
Mezun	8	22.119



Şekil 6.7. YSA karmaşıklık matrisi ve ROC eğrisi

DVM algoritmasında %99.90 doğruluk oranı bulunmuştur. DVM 10 katlı çapraz doğrulama karmaşıklık matrisi ve ROC eğrisi grafiği şekil 5.8'de verilmiştir.

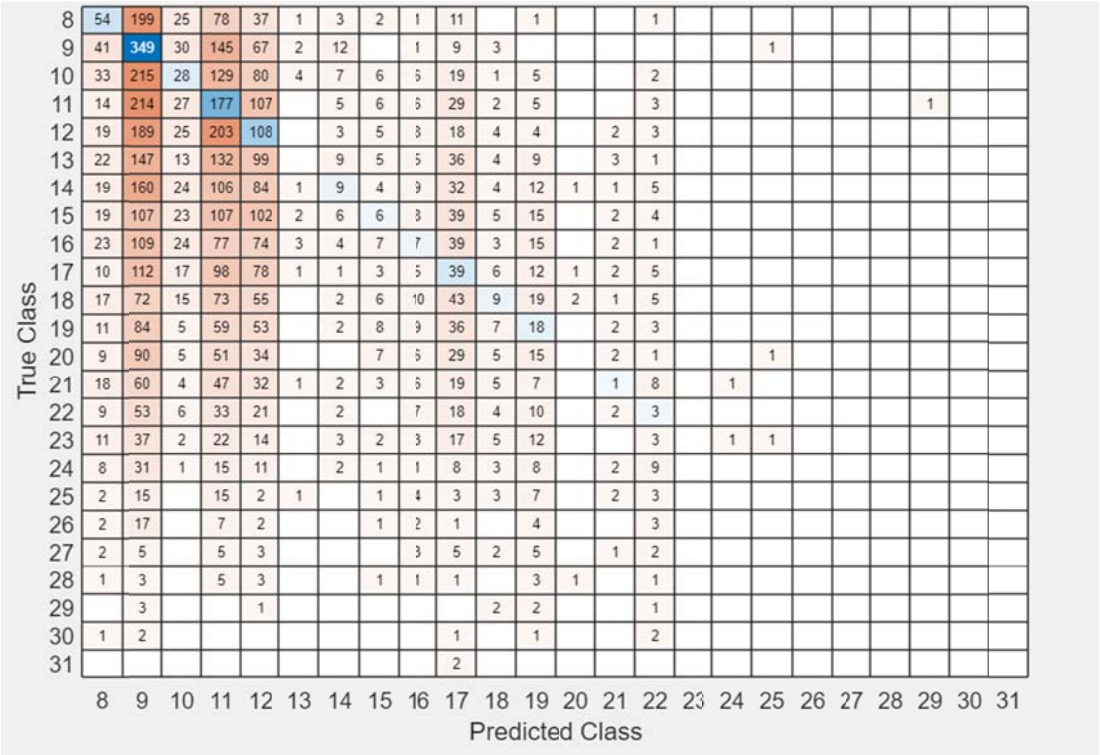
	Mezun Değil	**Mezun**
Mezun Değil	120.536	51
Mezun	58	22.069



Şekil 6.8. DVM karmaşıklık matrisi ve ROC eğrisi

5.3. KAÇ DÖNEMDE MEZUN OLUNABİLİR TESTİ

Yapılan üçüncü deneysel çalışmada ise ilk iki deneyde kullanılan aynı yapay zeka teknikleri kullanılmıştır. Mezun olan 22.128 öğrencinin kaç dönemde mezun olacaklarının tespitine dair deneysel çalışmalar yapılmıştır. Bu deneysel çalışmalar sonucunda en iyi sonuçlar YSA yöntemi ile elde edilmiştir. YSA ile %12.2 doğruluk oranı elde edilirken DT, KNN ve SVM, yöntemleri ile sırasıyla % 11.6, %9.6 ,%11.1 doğruluk oranları elde edilmiştir. En iyi sonuçların elde edildiği YSA yöntemine ait karmaşıklık matrisi şekil 5.9'da verilmiştir.



Şekil 6.9. YSA karmaşıklık matrisi.

BÖLÜM 6

SONUÇ VE ÖNERİLER

Bu çalışmada, AÖL öğrencilerinin ele alınan özellikleri kullanılarak normal sürede mezun olup olamayacaklarını ve kaç dönemde mezun olabileceklerini tahmin etmek için yapay zeka metotlarından KA, KNN, YSA ve DVM algoritmalarının tahmin yetenekleri incelenmiştir. Bu amaçla AÖL öğrencilerinden 142.714 öğrencinin verileri kullanılmıştır.

Yapılan ilk deney sonuçlarında KA algoritmasının KNN, YSA ve DVM algoritmalarına oranla biraz daha iyi sonuç verdiği görülmüştür. Yapılan ilk deneylerde %99.99 başarı oranı yakalanırken k-katlı çapraz doğrulama ile yapılan ikinci deneylerin sonucunda ise %99.98 başarı elde edilmiştir.

Öğrencilerin kaç dönemde mezun olacaklarına dair yapılan üçüncü deneyde ise YSA algoritması ile %12.20 doğruluk oranı ele edilmiştir. Aynı verilere öğrencilerin devam ettikleri dönemlerde seçtikleri krediler ve başarılı oldukları krediler eklense dahi kaç dönemde mezun olabileceklerinin tahmini için tatmin edici sonuçlar elde edilememiştir.

Bu çalışmada öğrencilerin ele alınan özellikleri üzerinden normal süre içerisinde mezun olup olamayacakları % 99.99 oranında doğru tahmin edilmektedir. Elde edilen deneysel sonuçlara göre karar ağaçları modeli önerilmektedir.

Yapılan deneyler, eldeki öğrenci verileri ile öğrencilerin zamanında mezun olup olamayacağının tahmin edilebileceği görülmüştür. Bu çalışmanın sonuçlarına göre öğrenciler zamanında mezun olup olamayacaklarının farkında olabileceklerdir. Gelecekteki çalışmalarda daha farklı özellikler kullanılarak derin öğrenme teknikleri uygulanabilir.

KAYNAKLAR

[1] Şengür D., Tekin A. ve Hayrettin B., "Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini", (2014).

[2] Fayyad U., Piatetsky-Shapiro G. ve Smyth P., "The KDD Process for Extracting Useful Knowledge from Volumes of Data", ***Commun. ACM***, 39(2): 27-34, (1996).

[3] Metan G., Sabuncuoglu I. ve Pierreval H., "Real time selection of scheduling rules and knowledge extraction via dynamically controlled data mining", ***Int. J. Prod. Res.***, 48(23): 6909-6938, (2010).

[4] Gutierrez-Osuna R., "Pattern analysis for machine olfaction: A review", ***IEEE Sensors Journal***, 2(3): 189-202, (2002).

[5] Özekes S., "Veri madenciliği modelleri ve uygulama alanları", ***İstanbul Ticaret Üniversitesi Derg.***, (2003).

[6] Ben-Gal I., "Outlier detection", ***Data Mining and Knowledge Discovery Handbook***, Boston, MA: Springer US, 131-146 (2005).

[7] Agrawal R. ve Srikant R., "Mining sequential patterns", ***Proceedings-International Conference on Data Engineering***, 3-14 (1995).

[8] Özdemir A., Saylam B. ve Bilen B. B., "Eğitim Sisteminde Veri Madenciliği Uygulamaları Ve Farkındalık Üzerine Bir Durum Çalışması", ***Atatürk Üniversitesi Sos. Bilim. Enstitüsü Derg.***, 22(2): 2159-2172 (2018).

[9] Yılmaz F., Acar S., Gültekin L. ve Meydan, M.C., "İlçelerin Sosyo-Ekonomik Gelişmişlik Sıralaması Araştırması SEGE″. ***Kalkınma Ajansları Genel Müdürlüğü Yayını***, (2019).

[10] Abu-Oda G. S. ve El-Halees A. M., "Data mining in higher education: University student dropout case study", ***Int. J. Data Min. Knowl. Manag. Process***, 5(1): 15 (2015).

[11] Hardman J., Paucar-Caceres A. ve Fielding A., "Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm", ***Syst. Res. Behav. Sci.***, 30(2): 194-203 (2013).

[12] Kaur P., Singh M. ve Josan G. S., "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector", ***Procedia Computer Science***, 57: 500–508 (2015).

[13] L. Aulck L., Velagapudi N., Blumenstock J. ve West J., "Predicting student dropout in higher education", ***arXiv preprint arXiv:1606.06364***, (2016).

[14] Papamitsiou Z. K., V. Terzis V. ve Economides A. A., "Temporal learning analytics for computer based testing", ***ACM International Conference Proceeding Series***, 31-35 (2014).

[15] E. B. Costa E. B., Fonseca B., Santana M. A., de Araújo F. F. ve Rego J., "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", ***Comput. Human Behav.***, 73: 247-256 (2017).

[16] Hassan S.-U., Waheed H., Aljohani N. R., Ali M., Ventura S. ve Herrera F., "Virtual learning environment to predict withdrawal by leveraging deep learning", ***Int. J. Intell. Syst.***, 34(8): 1935-1952 (2019).

[17] Wasif M., Waheed H., Aljohani N. R., ve Saeed-Ul H., "Understanding Student Learning Behavior and Predicting Their Performance", ***Cognitive Computing in Technology-Enhanced Learning***, 1-28 (2019).

[18] Lopez Guarin C. E., Guzman E. L., ve Gonzalez F. A., "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining", ***Rev. Iberoam. Tecnol. del Aprendiz.***, 10(3): 119-125 (2015).

[19] Y. Altujjar, W. Altamimi, I. Al-Turaiki, ve M. Al-Razgan, "Predicting Critical Courses Affecting Students Performance: A Case Study", ***Procedia Computer Science***, 82: 65-71 (2016).

[20] Fernandes E., Holanda M., Victorino M., Borges V., Carvalho R. ve G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil", ***arXiv preprint arXiv:1606.06364***, 94: 335-343 (2019).

[21] Marbouti F., Diefes-Dux H. A. ve Madhavan K., "Models for early prediction of at-risk students in a course using standards-based grading", ***Comput. Educ.***, 103: 1-15 (2016).

[22] Akay E. Ç., "Ekonometri'de Yeni Bir Ufuk: Büyük Veri ve Makine Öğrenmesi", ***Sos. Bilim. Araştırması Derg.***, 7(2): 41-53 (2018).

[23] Vaidehi V. ve Vasuhi S., "Person authentication using face recognition,", ***Proc. World Congress on Engineering and Computer Science***, (2008).

[24] Toker G. ve Kirmemis O., "Text categorization using k nearest neighbor classification", ***Surv. Pap. Middle East Tech. Univ.***, (2013).

[25] Liao Y. ve Vemuri V. R., "Using text categorization techniques for intrusion detection", ***USENIX Security Symposium***, 12: 51-59 (2002).

[26] Geng X., Liu T.-Y., Qin T., Arnold A., Li H., ve Shum H.-Y., "Query dependent ranking using k-nearest neighbor", ***Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval***, 115-122 (2008).

[27] Yang Y., Ault T., Pierce T. ve Lattimer C. W., "Improving text categorization methods for event tracking", ***Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval***, 65-72 (2000).

[28] Bajramovic F., Mattern F., Butko N., ve Denzler J., "A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition", ***Advanced Concepts for Intelligent Vision Systems***, 1186-1197 (2006).

[29] Cover T. M. ve Hart P. E., "Nearest Neighbor Pattern Classification," ***IEEE Trans. Inf. Theory***, 13(1): 21-27 (1967).

[30] Bhatia N., "Survey of Nearest Neighbor Techniques", ***International Journal of Computer Science and Information Security***, 8(2) (2010).

[31] Cortes C. ve Vapnik V., "Support-vector networks", ***Mach. Learn.***, 20(3): 273-297 (1995).

[32] Yang J., Awan A. J., ve Vall-Llosera G., "Support Vector Machines on Noisy Intermediate Scale Quantum Computers", ***arXiv preprint arXiv:1909.11988***, (2019).

[33] Boser B. E., Guyon I. M., ve Vapnik V. N., "Training algorithm for optimal margin classifiers", ***Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory***, 144-152 (1992).

[34] Nisbet R., Miner G., ve Yale K., "Handbook of statistical analysis and data mining applications", ***Academic Press***, (2017).

[35] Quinlan J. R., "Introduction of Decision Trees Machine learning", ***Kluwer Academic Publishers***, 1(1) 1986.

[36] Kantardzic M., "Data mining: concepts, models, methods, and algorithms", ***John Wiley & Sons***, (2011).

[37] Angluin D. ve Laird P., "Learning from noisy examples", ***Machine Learning***, 2(4): 343-370 (1988).

[38] Breiman L., Friedman J., Stone C. J., ve Olshen R. A., "Classification and regression trees", ***CRC press***, (1984).

[39] Fürnkranz J., "Pruning Algorithms for Rule Learning", ***Machine Learning***, 27(2): 139-172 (1997).

[40] Egrioglu E., Aladag C. H., Yolcu U., Uslu V. R., ve Basaran M. A., "A new approach based on artificial neural networks for high order multivariate fuzzy time series," ***Expert Syst. Appl.***, 36(7): 10589-10594 (2009).

[41] Nabiyev V., "Yapay Zeka", ***Seckin Yayincilik***, (2012).

[42] Ozturk K. ve Sahin M. E., "Yapay Sinir Aglari ve Yapay Zeka'ya Genel Bir Bakis", ***Tak. Vekayi***, 6(2): 25-36 (2018).

[43] Kaya Ü., Oğuz Y., ve Şenol Ü., "An Assessment of Energy Production Capacity of Amasra Town Using Artificial Neural Networks," ***Turkish J. Electromechanics Energy***, 3(1): 22-26 (2018).

[44] Masetic Z. ve Subasi A., "Congestive heart failure detection using random forest classifier," ***Comput. Methods Programs Biomed.***, 130: 54-64 (2016).

[45] Mursalin M., Zhang Y., Chen Y., ve Chawla N. V., "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier," ***Neurocomputing***, 241: 204-214 (2017).

[46] Kartal E., "Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama", ***İstanbul Üniversitesi***, (2015).

[47] Chaovalitwongse W. A., Fan Y. J., ve Sachdeo R. C., "On the time series K-nearest neighbor classification of abnormal brain activity", ***IEEE Trans. Syst. Man, Cybern. Part ASystems Humans***, 37(6): 1005-1016 (2007).

[48] Najah A. A., El-Shafie A., Karim O. A., ve Jaafar O., "Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation", ***Neural Comput. Appl.***, 21(5): 833-841 (2012).

[49] Zhang G., Hu M. Y., Patuwo B. E., ve Indro D. C., "Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis", ***Eur. J. Oper. Res.***, 116(1): 16-32 (1999).

[50] N. R. Draper ve H. Smith, "Applied regression analysis", ***John Wiley \& Sons***, (1998).

[51] Efron B., "Estimating the error rate of a prediction rule: improvement on cross-validation", ***J. Am. Stat. Assoc.***, 78(382): 316-331 (1983).

[52] Burman P., Chow E., ve Nolan D., "A cross-validatory method for dependent data," ***Biometrika***, 81(2): 351-358 (1994).

[53] Hall P., "Large sample optimality of least squares cross-validation in density estimation," ***The Annals of Statistics***, 11(4): 1156-1174 (1983).

[54] Noureldin A., El-Shafie A., ve Taha M. R., "Optimizing neuro-fuzzy modules for data fusion of vehicular navigation systems using temporal cross-validation", ***Eng. Appl. Artif. Intell.***, 20(1): 49-61 (2007).

[55] Albayrak A. S. ve Yilmaz S. K., "Veri Madenciliği: Karar Ağacı Algoritmalari ve İmkb Verileri Üzerine Bir Uygulama", ***Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences***, 14(1), (2009).

[56] Wiens T. S., Dale B. C., Boyce M. S., ve Kershaw G. P., "Three way k-fold cross-validation of resource selection functions", ***Ecological Modelling***, 212(3–4): 244-255 (2008).

[57] Al Shalabi L. ve Shaaban Z., "Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix", ***Proceedings of International Conference on Dependability of Computer Systems, DepCoS-RELCOMEX 2006***, 207-214 (2006).

[58] Çarkacı N., "Derin Öğrenme Uygulamalarında En Sık kullanılan Hiper-parametreler," **https://medium.com/deep-learning-turkiye/derin-ogrenme-uygulamalarinda-en-sik-kullanilan-hiper-parametreler-ece8e9125c4** (2018).

[59] Amrieh E. A., Hamtini T., ve Aljarah I., "Mining educational data to predict student's academic performance using ensemble methods", ***International Journal of Database Theory and Application***, 9(8): 119-136 (2016).

ÖZGEÇMİŞ

Mirhaç SULAK, ilk öğrenimini Seydişehir'de, ortaöğrenimini Antalya'da tamamladı. Antalya Anadolu Teknik Lisesi Bilgisayar Bölümünden mezun oldu. 1998 yılında Gazi Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Sistemleri Öğretmenliği Bölümü'nde öğrenime başlayıp 2002 yılında iyi derece ile mezun oldu. 2002 yılında Karaman Temizel-Ünlü Bilgisayar Anadolu Teknik Lisesi'nde öğretmen olarak göreve başladı. 2009 yılında Millî Eğitim Bakanlığı Bilgi İşlem Genel Müdürlüğünde göreve başladı. 2016 yılına kadar bu birimde SoruBankası (Basılı sınav kitapçığı ve E-Sınav Soru Bankası), REBUS (Yüksek Öğretim Resmî Burslu Öğrenci Sistemi), Hayat Boyu Öğrenme Genel Müdürlüğü E-Yaygın projesi gibi birçok projeye tamamladı. 2017 yılında Karabük Üniversite Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği A.B.D. yüksek lisans eğitimine başladı. 2018 yılında Hayat Boyu Öğrenme Genel Müdürlüğü Açık Öğretim Lisesi bünyesinde görevlendirildi. Halen bu birimde Açık Öğretim Lisesi öğrencilerinin iş ve işlemlerinin elektronik ortamda yürütülmesi üzerinde çalışmaktadır.