

T.C.
BİLECİK ŐEYH EDEBALI ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĐİ ANABİLİM DALI

**BİLECİK İLİNDE İLKÖĐRETİMDEN LİSEYE GEÇİŐ SINAVLARINDA MAKİNE
ÖĐRENMESİ YÖNTEMLERİ İLE ÖĐRENCİ BAŐARISININ TAHMİNİ**

YÜKSEK LİSANS TEZİ

AYŐEGÜL SELVİ

TEZ DANIŐMANI

DR. ÖĐR. ÜYESİ SALİM CEYHAN

BİLECİK, 2020

10350023

T.C.
BİLECİK ŐEYH EDEBALI ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĐİ ANABİLİM DALI

**BİLECİK İLİNDE İLKÖĐRETİMDEN LİSEYE GEÇİŐ SINAVLARINDA MAKİNE
ÖĐRENMESİ YÖNTEMLERİ İLE ÖĐRENCİ BAŐARISININ TAHMİNİ**

YÜKSEK LİSANS TEZİ

AYŐEGÜL SELVİ

TEZ DANIŐMANI

DR. ÖĐR. ÜYESİ SALİM CEYHAN

BİLECİK, 2020

10350023

BEYAN

Bilecik ilinde ilköğretimden liseye geçiş sınavlarında makine öğrenmesi yöntemleri ile öğrenci başarısının tahmini adlı yüksek lisans tezinin hazırlık ve yazımı sırasında bilimsel ahlak kurallarına uyduğumu, başkalarının eserlerinden yararlandığım bölümlerde bilimsel kurallara uygun olarak atıfta bulunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, tezin herhangi bir kısmının Bilecik Şeyh Edebali Üniversitesi veya başka bir üniversitede başka bir tez çalışması olarak sunulmadığını beyan ederim.

Ayşegül SELVİ

06/08/2020

ÖN SÖZ

Bu tez çalışmasının yazılmasında, çalışmamı sahiplenerek takip eden, bilgi ve tecrübelerini paylaşan, hoşgörü ve anlayışla desteklerini esirgemeyen, sabırla katlanıp emeği geçen saygıdeğer danışmanım Dr. Öğr. Üyesi Salim CEYHAN'a katkı ve emekleri için, anket çalışmasına katılan öğrencilere, anketin uygulanması konusunda yardımlarını esirgemeyen kıymetli kurum yöneticilerine ve meslektaşlarıma, tezimi yazma aşamasında hep yanımda olan, maddi ve manevi desteğini esirgemeyen, beni her zaman cesaretlendiren sevgili hayat arkadaşım Öğr. Gör. Ali Osman SELVİ'ye ve biricik çocuklarım Ali ve Ayçıl SELVİ'ye teşekkür ederim.

Ayşegül SELVİ

06/08/2020



ÖZET

BİLECİK İLİNDE İLKÖĞRETİMDEN LİSEYE GEÇİŞ SINAVLARINDA MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE ÖĞRENCİ BAŞARISININ TAHMİNİ

İyi bir öğretim süreci öğrencilerin ihtiyaçlarının karşılanma gerekliliği ile doğru orantılıdır. Bu anlamda öğrencilerin akademik başarılarının tahmin edilmesi de eğitim kurumları için önem arz etmektedir. Son yıllarda ülkemizde birçok alanda olduğu gibi eğitim alanında da farklı projelerin hayata geçirilmesiyle öğrenci başarılarında artış gözlemlenmiştir. Özellikle Türkçe, Matematik ve Fen Bilgisi derslerindeki başarı artışı ele alındığında, eğitimin kalitesini korumak ve geliştirmek için öğrencilere ait önemli veri ve bilgiler düzenli olarak toplanıp çok sınıflı makine öğrenmesi yöntemleri kullanılarak çıkarılan sonuçlar okul yönetimlerince değerlendirilmelidir. Bu çalışmada öğrenci performansını etkilemesi beklenen değişkenler olan çeşitli demografik, sosyal / duygusal ve kapalı uçlu sorular içeren 32 soruluk anket, 2019 yılında sınav puanıyla öğrenci alan 4 farklı lise grubundaki 9.sınıf ve halen ortaokul 8.sınıfta okuyan öğrencilere uygulanmıştır. Kâğıt sayfalara dayanan öğrenci akademik bilgileri ile anket sonuçları üzerinden elde edilen veriler birleştirilip çok sınıflı makine öğrenmesi yöntemleri ile model oluşturularak başarı tahmini yapılmıştır. Öncelikle verilere öznitelik çıkarma yöntemi uygulanarak başlangıçtaki özniteliklerimizin en iyi bilgi veren bir alt kümesi elde edilmiştir. Bu filtrelenmiş özniteliklerle oluşturulan çok sınıflı makine öğrenmesi modelleri karşılaştırılarak, Random Forest yönteminin en iyi sonucu verdiği gözlemlenmiştir.

Anahtar Kelimeler: Eğitimsel Veri Madenciliği, Karar Ağaçları Algoritmaları, Öznitelik Seçimi, Sınıflandırma.

ABSTRACT

PREDICTING STUDENT ACHIEVEMENT WITH MACHINE LEARNING METHODS IN TRANSITION FROM PRIMARY TO HIGH SCHOOL EXAMS IN BİLECİK PROVINCE

A good teaching process is directly proportional to the needs of students. In this sense, estimating students' academic success is also important for educational institutions. In recent years, as in many fields in our country, an increase in student achievements has been observed with the implementation of different projects in the field of education. Especially considering the increasing success in Turkish, Mathematics and Science courses, in order to maintain and improve the quality of education, important data and information belonging to students should be collected regularly and the results obtained by using multi-class machine learning methods should be evaluated by school administrations. In this study, a 32-question questionnaire with various demographic, social / emotional, and closed-ended questions, which are the variables expected to affect student performance, was applied to students studying in the 9th grade and 8th grade in 4 high schools that received students with an exam score in 2019. Achievement estimation has been made by combining student academic information based on paper pages and data obtained from the survey results and by creating a model with multi-class machine learning methods. First of all, by applying the feature extraction method, a subset of our initial attributes giving the best information was obtained. By comparing multi-class machine learning models created with these filtered attributes, it was observed that Random Forest method gave the best result.

Key Words: Educational Data Mining, Decision Trees Algorithms, Feature Selection, Classification.

İÇİNDEKİLER

	Sayfa
ÖN SÖZ.....	i
ÖZET.....	ii
ABSTRACT	iii
İÇİNDEKİLER.....	iv
TABLolar LİSTESİ	vi
ŞEKİLLER LİSTESİ.....	vii
KISALTMALAR VE SİMGELER LİSTESİ	viii
1. GİRİŞ	1
1.1. Çalışmanın Amacı	3
1.2. Literatür Taraması	3
2. VERİ MADENCİLİĞİ.....	7
2.1. Veri Madenciliği Nedir?	7
2.2. Veri Madenciliği Süreci.....	8
2.2.1. Problemi tanımlama	9
2.2.2. Veri tanımlama ve toplama	9
2.2.3. Veri hazırlama	9
2.2.3.1. Veri temizleme	10
2.2.3.2. Veri bütünleştirme	10
2.2.3.3. Veri indirgeme	11
2.2.3.4. Veri dönüştürme.....	11
2.2.4. Veri madenciliği modelinin kurulması ve algoritmasının uygulanması	12
2.2.5. Sonuçların değerlendirilmesi.....	13
2.3. Veri Madenciliği Yöntemleri.....	13
2.3.1. Tahmin edici yöntemler	14
2.3.1.1. Sınıflandırma yöntemleri	15
2.3.1.2. Regresyon analizi.....	19
2.3.2. Tanımlayıcı yöntemler	19
2.3.2.1. Kümeleme yöntemi	19
2.3.2.2. Birliktelik kuralları.....	19
2.4. Veri Madenciliği Kullanım Alanları.....	20
2.4.1. Eğitim alanında veri madenciliği	20
2.4.2. Sağlık alanında veri madenciliği	20

2.4.3.	Bankacılık ve finans alanında veri madenciliği	21
2.4.4.	Telekomünikasyon alanında veri madenciliği.....	21
2.4.5.	İnşaat alanında veri madenciliği.....	21
2.4.6.	Mühendislik alanında veri madenciliği	21
2.4.7.	Endüstri alanında veri madenciliği.....	22
2.5.	Makine Öğrenmesi	22
3.	UYGULAMA	24
3.1.	Uygulamanın Amacı.....	24
3.2.	Veri Toplama Araçları.....	24
3.3.	Örnekleme Grubu	27
3.4.	Veri Temizleme	28
3.5.	Veri Dönüştürme	28
3.6.	Cinsiyet Ve Mezuniyet Değişkenlerine Ait Yorumlar	32
3.6.1.	Öğrencinin cinsiyeti	32
3.6.2.	Öğrencinin mezun olduğu ortaokul.....	32
3.7.	Model Oluşturma Ve Model Seçimi.....	33
3.7.1.	Uygulamada kullanılan veri madenciliği aracı.....	33
3.7.2.	Ön işleme süreci	34
3.7.3.	Model oluşturma	35
3.7.4.	Model performans ölçümlerinin değerlendirilmesi.....	36
3.7.4.1.	J48 modelinin başarı ölçütü	38
3.7.4.2.	Yapay sinir ağları (PNN) modelinin başarı ölçütü.....	38
3.7.4.3.	Random forest modelinin başarı ölçütü	39
3.7.4.4.	Decision tree modelinin başarı ölçütü.....	40
3.7.4.5.	RAPTree modelinin başarı ölçütü.....	40
3.7.4.6.	Hoeffding tree modelinin başarı ölçütü	41
3.7.5.	Modellerin karşılaştırılması.....	42
3.7.6.	Bağımlı değişkenlerin karşılaştırılması	43
4.	SONUÇLAR	44
	KAYNAKÇA	46
	ÖZ GEÇMİŞ	52

TABLULAR LİSTESİ

Sayfa

Tablo 2.1. Yapay Sinir İle Biyolojik Sinir Hücreleri Arasındaki Benzerlikler (Pala, 2013:13).	17
Tablo 3.1. Çalışmada Kullanılan Nitelikler Ve Alabileceği Değerler.	28
Tablo 3.2. Mezuniyet Değişkeninin Sayısallaştırılması.	30
Tablo 3.3. Cinsiyet Değişkeninin Sayısallaştırılması.	30
Tablo 3.4. Ders Çalışma Alışkanlığı Değişkeninin Sayısallaştırılması.	30
Tablo 3.5. Sosyal Etkinliklere Katılım Değişkeninin Sayısallaştırılması.	30
Tablo 3.6. Takviye Ders Alma Durumu Değişkeninin Sayısallaştırılması.	30
Tablo 3.7. Hafta İçi Günlük İnternet Kullanım Süresi Değişkeninin Sayısallaştırılması.	31
Tablo 3.8. Hafta İçi Günlük Televizyon İzleme Süresi Değişkeninin Sayısallaştırılması.	31
Tablo 3.9. Ders Çalışma Süresi Değişkeninin Sayısallaştırılması.	31
Tablo 3.10. Okula Ulaşım Değişkeninin Sayısallaştırılması.	31
Tablo 3.11. Evin Isınma Türü Değişkeninin Sayısallaştırılması.	31
Tablo 3.12. Anne Sağ Mı Değişkeninin Sayısallaştırılması.	31
Tablo 3.13. Baba Öz Mü Değişkeninin Sayısallaştırılması.	31
Tablo 3.14. Anne Öğrenim Durumu Değişkeninin Sayısallaştırılması.	31
Tablo 3.15. Aylık Gelir Değişkeninin Sayısallaştırılması.	32
Tablo 3.16. Öğrencilerin Cinsiyet Değişkenine Göre Dağılımı.	32
Tablo 3.17. Öğrencilerin Mezun Oldukları Ortaokul Değişkenine Göre Dağılımı.	32
Tablo 3.18. Hata Matrisi.	36
Tablo 3.19. Filtrelenmiş Özelliklerle Her Bir Algoritmanın Karşılaştırılması.	42

ŞEKİLLER LİSTESİ

	Sayfa
Şekil 2.1. Veri Madenciliği Süreci (Aydemir, 2017:12).	8
Şekil 2.2. Veri Hazırlama Süreci.....	10
Şekil 2.3. Veri Sınıflandırma İşlemi (Pala, 2013:51).	16
Şekil 2.4. Karar Ağaçları Yapısına Örnek Bir Dallanma (Gürel, 2019:22).	16
Şekil 2.5. Veri Madenciliği Ve Diğer Disiplinler (Özel ve Topsakal, 2014:45).	22
Şekil 3.1. Knime Programının Arayüzü.	33
Şekil 3.2. İleri Özellik Seçimi Filtreleme Yöntemi.....	34
Şekil 3.3. Özellik Seçimi Genel Akış Şeması.	35
Şekil 3.4. Model Düğüm Bağlantıları.	36
Şekil 3.5. J48 Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.	38
Şekil 3.6. J48 Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.	38
Şekil 3.7. J48 Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.	38
Şekil 3.8. PNN Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.	39
Şekil 3.9. PNN Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.....	39
Şekil 3.10. PNN Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.....	39
Şekil 3.11. Random Forest Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.....	39
Şekil 3.12. Random Forest Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.	39
Şekil 3.13. Random Forest Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.	39
Şekil 3.14. Decision Tree Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.....	40
Şekil 3.15. Decision Tree Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.	40
Şekil 3.16. Decision Tree Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.	40
Şekil 3.17. RAPTree Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.	41
Şekil 3.18. RAPTree Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.	41
Şekil 3.19. RAPTree Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.....	41
Şekil 3.20. Hoeffding Tree Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.....	41
Şekil 3.21. Hoeffding Tree Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.....	41
Şekil 3.22. Hoeffding Tree Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.	42

KISALTMALAR VE SİMGELER LİSTESİ

OECD	: Ekonomik İşbirliği ve Kalkınma Örgütü
PISA	: Programme for International Student Assessment
LGS	: Liselere Geçiş Sınavı
YSA	: Yapay Sinir Ağları
LR	: Lojistik Regresyon
MLR	: Çok Katmanlı Algılayıcı
ÖSS	: Öğrenci Seçme Sınavı
EM	: Expectation Maximization
DVM	: Destek vektör makineleri
KNN	: K-En Yakın Komşu
AYSE	: Araştırma, Yarışma ve Sosyal Etkinlik
KNIME	: Konstanz Information Miner
TP	: True Positive
TN	: True Negative
FN	: False Negative
FP	: False Positive
CCI	: Doğru Sınıflandırılmış Örnek Sayısı
ICI	: Yanlış Sınıflandırılmış Örnek Sayısı
ÖSYM	: Öğrenci Seçme ve Yerleştirme Merkezi

1. GİRİŞ

İçinde bulunduğumuz çağa verilen birçok isimden biri de 'Bilgi Çağı'dır. Her çağda üretilen bilginin yeri de günümüzde ayrı bir önem kazanmıştır. Bilgi ve iletişim teknolojilerinde meydana gelen gelişmeler ve yükselişler insanoğlunu yeni bilgi keşfetmeye ve üretmeye yöneltmiştir. Birçok alanda teknolojinin kullanılmasıyla toplumlar kendilerini geliştirmekte ve böylece hayatı daha yaşanılabilir halde kılmaktadır. Toplumların bu gelişime ayak uydurmaları için süreçte aktif olmaları gerekmektedir. Yani bilgiye hızlı bir şekilde ulaşarak bilgiyi kullanabilen ve aynı zamanda yeni bilgi üretebilen bireyler yetişmelidir.

Eğitim, bireylerin toplumla iç içe yaşamda yer edinebilmeleri için gerekli beceri, bilgi ve anlayışları kazanmalarına, okul dışında veya okul içinde, doğrudan veya dolaylı yardım ederek kişiliklerini geliştirme sürecidir (TDK, 2020).

Eğitimi geleceği kurmak, geleceğe yatırım yapmak olarak düşünürsek eğitim sisteminde oluşabilecek bir problem ileride oluşabilecek toplumsal bir problemin erken göstergesidir. Bu anlamda 2003 yılından beri Türkiye, Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından yapılan PISA (Programme for International Student Assessment) testlerine katılmaktadır. 15 yaş grubundaki öğrencilerin okuma becerileri, matematik ve fen okuryazarlığı alanlarında kazandıkları bilgi ve becerileri değerlendiren testin yedinci döngüsü olan PISA 2018 sonuçlarına göre, ülkemizin okuma becerileri alanındaki ortalama puanı 2015 verileri ile kıyaslandığında 38 puanlık artışla 466'ya, fen okuryazarlığı alanındaki ortalama puanı 43 puanlık artışla 468'e ve ortalama matematik puanı ise 34 puanlık artışla 454'e yükselmiştir (MEB, 2019).

Görülüyor ki araştırılan her alanda Türkiye'nin puanları bir önceki döngüye göre artış göstermiştir. PISA testindeki artışın en önemli sebebi öğrencilerdeki gelişimin belli bir plan dahilinde olmasıdır. Bu plana Eğitim-Öğretim Programı denilmektedir.

Eğitim programı, öğrencilere Milli Eğitim'in planladığı tüm hedeflerin gerçekleşmesine yönelik bir eğitim kurumunda aktarılmasıdır. Öğretim programı ise öğrencilerin herhangi bir disiplinde ulaşmaları gereken hedefleri, bu hedeflerin içerdiği davranışları, davranışlara sahip olabilmeleri için düzenlenen ve düzenlenecek eğitim durumlarını ve kazanımların ne ölçüde kazandırıldığını belirleyecek sınav durumlarını da içine alan, çok yönlü birbiri ile etkileşim halinde olan öğeler bütünüdür (Aydemir, 2017:1). Eğitim öğretim programının temelini kuşkusuz öğrenciler oluşturmaktadır. Düzenlenen eğitim öğretim programlarında öğrencilerin kazanım olarak ulaşmaları gereken hedef ve amaçlar vardır. Her bir disiplinde hedef ve amaçlara ulaşılması için eğitim öğretim etkinlikleri

düzenlenir. Bu etkinlikler sonucunda öğrencilerin ulaşması beklenen hedeflere ne kadarına ulaşıp ne kadarına ulaşılmadığını belirlemek için ölçme ve değerlendirme etkinlikleri yapılır.

Türk milli eğitim sistemi yaygın eğitim ve örgün eğitim olmak üzere iki bölümden oluşur. Örgün eğitim, mecburi olan ilköğrenim çağına henüz gelmemiş çocukların eğitimini kapsayan *okul öncesi*, 6 ile 14 yaş grubu arasındaki çocukların eğitimini kapsayan *ilköğretim*, ilköğretime dayalı dört yıl zorunlu yaygın veya örgün öğrenim veren mesleki, teknik ve genel eğitim öğretim kurumları ile meslek eğitimi merkezlerinin tamamını kapsayan *ortaöğretim* ve orta öğretime dayalı en az iki yıllık yüksek öğrenim veren eğitim kurumlarının tamamını kapsayan *yükseköğretim* kurumlarını içerir (MEVZUAT,2020).

Türk milli eğitim sisteminden de anlaşıldığı gibi ilköğretim kurumları ilköğretim ve ortaokulları kapsamaktadır. İlkokul kademesi 1., 2., 3. ve 4. sınıflardan oluşmaktadır. Bu kademe öğrencilerin eğitim temelinin atıldığı dönemdir. 5., 6., 7. ve 8. sınıfları içeren kısım ise ortaöğretime geçmeden önceki dönemdir. İlköğretim okullarında öğrenim gören 8. sınıf öğrencileri kısa adı LGS olan Liselere Geçiş Sistemine girmektedirler. LGS'ye giren 8. sınıf öğrencileri yüzde 30 diploma notu ve yüzde 70 sınav puanı ile sınav ile alım yapan liselere yerleşmektedir. Sınav puanı ile öğrenci alan bir liseye yerleşebilmek için LGS'de başarılı olmak gerektiği kadar diploma notlarının da etkisi büyük ölçüde önemlidir.

Teknolojinin gelişmesiyle bilgiye erişim daha kolay hale gelirken, arkadaş çevresinin yanlış tutumları, ailenin ekonomik durumu ve eğitim düzeyi, sosyal çevre gibi dış etkenlerin yanı sıra kardeş sayısı, derse karşı tutum, cinsiyet gibi etkenlere bağlı olarak öğrencilerin başarılarında azalmalar yaşanabilmektedir. Bununla birlikte öğretmenlerin de eğitim-öğretim üzerindeki etkisi büyük ölçüde önemlidir. Lisans eğitimi veya pedagojik formasyon dersleri öğretmenlik için sadece bir başlangıç adıdır. Öğretmenlerin derse karşı tutumu, yargısı, dersi işleyiş şekli, sınav türü öğrenci başarısını etkileyen diğer etmenlerdendir. Eğitim-öğretim sürecinde öğrencilere uygulanan sınavlar klasik, doğru-yanlış veya eşleştirme olurken, genel sınavlar test tekniği ile hazırlanmış sınavlar olduğundan farklı sınav türleri de öğrenci başarısını büyük oranda düşürmektedir. Öğrenci üzerinde büyük etkisi olan diğer bir etmende aileleridir. Sadece akademik başarıya odaklanmış olan aileler sebebiyle sınavda başarılı olabilecek öğrenciler stres altında kalarak olumsuz etkilenmektedirler. İfade edilen birçok olumsuz etken öğrencilerin performansını düşürebilmektedir. Bu yüzden öğrenciler disiplinler içindeki performanslarını artırmak ya da öğrenci performanslarında oluşabilecek kötü durumların önüne geçmek için ileriye dönük tahmin yapmak eğitimin kalitesini artırabilmektedir. Eğitimin kalitesini artırmak, ileriye dönük tahminde bulunabilmek için ise veri madenciliği yöntemleri etkili bir şekilde kullanılmaktadır. Son zamanlarda veri

madenciliği alanında yapılan birçok çalışma bulunmakla birlikte literatür taraması yapıldığında yirmi yıla yayılmış hem ulusal hem de uluslar arası birçok çalışmadan söz edilebilir. Çalışmalar incelendiğinde görülüyor ki; öğrencilerin gelecekteki başarılarını tahmin etmede öncü olan etmen geçmişteki akademik başarıları ve başarılarını etkileyen etmenlerin ortaya çıkarılmasıdır.

Veri madenciliği var olan verileri analiz ederek, ilişkiler çıkarma ve eldeki verilerden anlamlı bilgiler ortaya sunmak için kullanılan bir tekniktir (Tekin ve Öztekin, 2018:67). Eğitim alanında öğrenci başarılarını tahmin etmek için ise eğitsel veri madenciliği kullanılmaktadır. Yurt içinde veya yurt dışında bu yöntemleri kullanan çalışmalardan bazıları bölüm 1.2 literatür tarama kısmında incelenecektir.

1.1. Çalışmanın Amacı

Öğretim sürecinin kalitesi öğrencilerin ihtiyaçlarının karşılanma yeteneği olduğundan öğrenci başarısının tahmini eğitim kurumları için gerekli bir durumdur. Bu anlamda öğrencilere ait önemli veri ve bilgiler düzenli olarak toplanıp çok sınıflı makine öğrenmesi yöntemleri kullanılarak çıkarılan sonuçlar okul yönetimlerinde değerlendirilmeli ve eğitim kalitesini korumak ve geliştirmek için standartlar belirlenmelidir. Bu çalışmanın amacı Bilecik ilinde ilköğretimin ikinci kademesi olan ortaokul öğrencilerinin başarılarında etken olan üst seviye unsurları ortaya çıkarmaktır. Çalışma ile ortaöğretim 9. sınıf öğrencilerinin ilköğretim 8.sınıf temel dersleri olan Matematik, Türkçe ve Fen Bilgisi notları ile uygulanan ankete verdikleri cevaplar birlikte değerlendirilerek, hangi alanda başarılı olabileceklerine, başarılarını etkileyen unsurların belirlenmesine, başarı seviyelerine, başarısızlığa neden olan unsurların belirlenmesi sağlanacaktır. Sonuç olarak çalışma şu iki temel soruya cevap arayacaktır: Ülkemizde ilköğretim öğrencilerinin lise geçiş sınavlarındaki başarılarını etkileyen temel öznitelikler nelerdir? Lise geçiş sınavlarında öğrenci başarılarını tahmin etmek mümkün müdür?

1.2. Literatür Taraması

Murat GÖK 2017 yılında yapmış olduğu çalışmada, öğrencilerin sosyo-ekonomik koşullarının özellikle Matematik ve Türkçe derslerinin yanı sıra yıl sonu genel ortalamalarına etkilerini analiz etmek amacıyla, belli ölçütlere dayanarak hazırlanmış olduğu anketi ortaöğretimin tüm kademesindeki öğrencilere uygulamıştır. Anket sonucunda elde ettiği verilerden yararlanarak öğrencilerin dönem sonu genel başarı ortalamalarını sınıflandırma algoritmaları ve regresyon ile tahmin etmiştir. Öğrencilerin sınav notlarını Milli Eğitim Bakanlığı tarafından günümüzde kullanılan 5'li not ölçeğini esas alarak sınıflandırma

yöntemleri ile, 0-100 arasındaki sınav puanını ise regresyon yöntemleri ile belirlemiştir. 24 nitelik ve 1492 örnekten oluşan veri setini sosyo-ekonomik yönden farklı olan ortaokullardaki 6, 7 ve 8. sınıf öğrencilerine uyguladığı 24 soruluk anket ile elde etmiştir. Yazar bazı niteliklerin öğrenci başarısını doğrudan etkilediğini düşündüğünden, veri setindeki niteliklerin hepsini kullanmak yerine setini daha iyi temsil edecek nitelikleri korelasyon tabanlı öznelik alt kümesi seçme yöntemini kullanarak belirlemiştir. Verilerini, dokuz kümesi eğitim bir kümesi test kümesi olacak şekilde rastgele 10 kümeye ayırmıştır. Veriler arasında kaydırma yaparak 10 kez bu işlemi tekrarlayıp 10 tane başarı metriği elde etmiştir ve oluşan değerlerin ortalamasını almıştır. Bu deneysel sonuçları elde etmek için Weka yazılımını kullanmıştır. Sonuçlar karşılaştırıldığında Türkçe dersi için çoklu meta sınıflandırıcısı ile rastgele orman algoritması, Matematik dersi için KÖAK öznelik seçme ile DVM algoritması ve son olarak genel başarı ortalamasında yine KÖAK öznelik seçme ile lojistik algoritması en iyi sonucu vermiştir. (Gök, 2017:140-144).

Necdet GÜNER ve Emre ÇOMAK çalışmalarında Pamukkale Üniversitesi Mühendislik Fakültesi'ne (MF) kayıt yaptıran öğrencilerin bir kısmını riskli grup olarak belirlemiştir. Yazarların 2010 senesinde yapmış oldukları bu çalışmada riskli gruba aldıkları öğrenciler lise kademesindeki matematik bilgilerinin yeterli seviyede olmadığı düşünülen öğrencilerdir. Araştırmada 434 mühendislik öğrencisinin 20 özellik ile üniversiteye yerleştirmelerinde sınav sonuçlarına ait verileri ele alan yazarlar destek vektör makineleri yöntemi ile 145 öğrencinin verisini test için, 289 öğrencinin verisini ise makineyi eğitmek için kullanmışlardır. Yazarlar matematik I dersinde başarılı olabilecek öğrencileri %86,36 doğruluk oranı ile tahmin etmişlerdir. (Güner ve Çomak, 2011:89).

Gülçin ÇIRAK ve Ömay ÇOKLUK 2013 yılındaki araştırmalarında Ankara Üniversitesi eğitim bilimleri fakültesindeki 3. sınıfta olan 419 öğrenci ile çalışmışlardır. Yazarların farklı fakültelerin farklı programlarından seçtikleri öğrencilerden 115 tanesi erkek, 304 tanesi kız öğrencidir. Araştırmacılar veri setini oluşturan kişilere hangi liseden mezun oldukları ve mezuniyet ortalamaları, üniversite sınav sonuçları, yerleştirme puan türü gibi soruların yanı sıra fakültedeki sınavlarına hazırlık süreleri, düzenli çalışma alışkanlıkları, kardeş sayıları, ailelerinin ekonomik gelirleri, anne babalarının eğitim durumu vb. gibi sorulardan oluşan bir anket uygulamışlardır. Anket sonuçlarını yorumlamak için Yapay Sinir Ağları (YSA) Analizi ve Lojistik Regresyon (LR) Analizi kullanmışlardır. Ankette yer alan bazı nitelikler YSA Analizi için girdi değişkeni olurken, bazı nitelikler ise LR Analizi için bağımlı değişken olmuştur. YSA Analizinde kullanılan çıktı değişkeni ve LR Analizi için gerekli olan bağımlı değişken öğrencilerin lisedeki mezuniyet notları kabul edilmiştir. Aynı

zamanda çalışmada YSA için çok katmanlı algılayıcı (MLR), LR için ise adımsal yöntem olan ileriye doğru yöntemi kullanılmıştır. LR modeli kullanılarak yapılan sınıflandırma da başarı oranı %66.10'dur. YSA modeli ile yapılan eğitim setindeki sınıflandırma oranı %71.40 iken test setindeki bu oran %79.40'tır. Bu sonuçlara bakarak yazarlar ileriki yıllarda teste dahil edilebilecek yeni bir öğrencinin başarısını yapay sinir ağları ile daha doğru bir tahminde bulunabileceklerini %70.16'lık oran ile görmüşlerdir (Çırak ve Çokluk, 2013:74-76).

Ahmet Selman BOZKIR, Ebru SEZER ve Bilge GÖK 2009 yılında yapmış oldukları bildirimlerinde öğrenci seçme sınavı anketinden edindikleri veriler üzerinde çalışmışlardır. Bu veriler öğrenci seçme ve yerleştirme sınav merkezinin web sayfasından 2008 yılında uygulanan 80 soruluk ankete ait verilerdir. ÖSS'de öğrencilerin başarısını etkileyen etmenleri tespit ederken kümeleme ve sınıflandırma yöntemlerini kullanmışlardır. Bu yöntemlerin yanı sıra Expectation Maximization algoritmasına da yer vermişlerdir. Yazarların çalışmalarıyla amaçladıkları öğrencilerin ortaöğretim başarı puanı ve ağırlıklı ortaöğretim başarı puanı ile öğrenci seçme sınavındaki eşit ağırlık, sözel ve sayısal puanlarını etkileyen etmenleri aramaktır. 2008 yılında sınava giren on bin öğrenci veri setini (kümesini) oluşturmaktadır. Başarıyı etkileyen unsurları bulabilmek için ilk oluşturulan veri kümesine ortaöğretim başarı puanı, ağırlıklı ortaöğretim başarı puanı (SÖZEL), ağırlıklı ortaöğretim başarı puanı (SAYISAL), ağırlıklı ortaöğretim başarı puanı (EŞİT AĞIRLIK) puanlarının yanı sıra farklı puan çeşitleri ekleyerek sonuç veri seti oluşturmuşlardır. Girişteki veri kümesi ile sonuçtaki veri kümesi üzerinde birçok uygulama yapılmıştır. Çeşitli derslere duyulan ilginin, derslerde öğrencilerin kendilerini ne oranda başarılı bulduklarının, ödevlerine ayırdıkları zamanın sonuçları nasıl etkilediğini gözlemlenmişlerdir. Yazarlar bildirimdeki analizlerinin oluşturulma aşaması için karar ağacı ve kümeleme yöntemlerini kullanmışlardır. Karar ağacı için bir eğitim kümesi ve bu küme üzerinde 2 farklı modeli varsayılan olarak oluşturmuşlardır. 1. modelde öğrencilerin eşit ağırlık, sayısal ve sözel puanlarını girdi ve tahmin için kullanırken 2. modelde ise bu puanlar sadece tahmin için kullanılmıştır. Bu sebeple veri setindeki diğer niteliklere etkileri olmamıştır. Seçilen bu metod öğrenci başarılarına etki eden etmenleri tespit edebilirken, veri setindeki diğer farklı nitelikler için boş sonuç oluşturmuşlardır. Böyle bir netice için başka bir karar ağacı oluşturmanın şart olduğu görülmüş ve 2. karar ağacında oluşturulduktan sonra bu iki karar ağacına ait sonuçlar birleştirilmiştir. Aynı veriler üzerinde yazarlar kümeleme yöntemini de kullanmışlardır. Bu yöntemde öğrenciler yöntemin başarılı kümeler oluşturabilmesi için yazarlar küme sayısını belirlemeyip EM algoritması ile küme sayısının otomatik oluşturulmasını sağlamışlardır. Oluşan 7 küme içinde en başarılı küme 800 katılımcısı ile 6 numaralı küme olurken, en başarısız küme ise 2072 katılımcı sayısı ile 2

numaralı küme olmuştur. Çalışmanın sonucunda bütün öğrenciler için ortak ölçüt ortaöğretim başarı puanı olarak düşünülürse bu puana etki eden etmenler fen laboratuvarını kullanma oranı, matematik dersi için harcanan zaman (ödev yapma zamanı), okul türü, yaş, sosyal derlere alaka en etkin beş faktör olduğunu görmüşlerdir (Bozkır vd., 2009:2).

Burak AYDEMİR 2019 yılında yapmış olduğu çalışmasında Pamukkale Üniversitesi Meslek Yüksek Okul programlarına 2009-2012 yılları arasında kayıt yaptıran 1387 öğrencinin bilgileri ile akademik başarı tahmin etmede en iyi sonucu veren sınıflama algoritmasını seçmeye çalışmıştır. Bu yıllar arasında meslek yüksek okulları programlarına kayıt yaptıran öğrencilerin bilgileri ve bu öğrencilerin ailesel bazı bilgilerini edinebilmek için hazırlanan ankete verilen yanıtlar bilgi işlem merkezinden alınmıştır. Toplanan veriler veri madenciliği sürecinden geçirilerek çalışma için uygun hale getirilmiştir. Oluşturulan modeller sonucunda akademik ortalama bağımlı değişkenine göre J48 algoritması %59,55 doğruluk oranı ile ve SMO algoritması ise %59,98 doğruluk oranı ile en iyi sonucu veren algoritmalar olmuşlardır. Mezuniyet yılı bağımlı değişkenine göre ise J48 algoritması %81,326 doğruluk oranı ile ve NaiveBayes algoritması ise %81,11 doğruluk oranı ile en iyi başarıyı veren algoritmalar olduğu görülmüştür (Aydemir, 2017:41-45).

Alaa Khalaf Hamoud 2016 yılındaki çalışmasında, Portekizli bir öğrencinin Matematik kursu ve Portekizce dil kursuna ait verileri içererek 32 özellikten ve 1044 örnekten oluşan bir veri seti kullanmaktadır. Veri seti Minho Üniversitesi'ndeki Paulo Cortez ve Alice Silva tarafından toplanmış ve analiz edilmiştir. Yazar çalışmasında J48, RepTree ve Hoeffding Tree olmak üzere üç farklı karar ağacı algoritmasını WEKA 3.8.0 aracı ile uygulamış ve sonuçları listeleterek karşılaştırmıştır. Oluşturulan modeller sonucunda J48 algoritmasının öğrencinin eylemini tahmin etmekte ve sınıflandırmakta yol haritası olarak kullanılabilir en iyi karar ağacı algoritması olduğu gözlemlenmiştir (Hamoud, 2016:29).

Asiye Şengül AVŞAR ve Seher YALÇIN 2015 yılındaki çalışmalarında amaçladıkları PISA 2009 testinden seçtikleri bazı maddelerden temin ettikleri okuma okuryazarlığı başarı puanlarını açıklayan ailevi etmenleri belirlemektir. Veri setini Ankara'daki lise ikinci sınıfta öğrenim gören 170 öğrenci ve velilerine uygulanan anketlere verilen cevaplardan elde etmişlerdir. Karar ağacı algoritmalarından olan CHAID ile öğrencileri okuma becerilerini açığa çıkaran değişkenleri bulmuşlardır. Çalışmanın sonucunda okul öğretmenlerinin işini büyük bir özveri ile yapan ve mesleki alanında yeterli bilgiye sahip olduğunu düşünen velilerin çocuklarının diğer çocuklara oranla okuma başarılarının daha yüksek olduğu tespit edilmiştir (Avşar ve Yalçın, 2015:2).

2. VERİ MADENCİLİĞİ

2.1. Veri Madenciliği Nedir?

Günümüzde teknolojinin gelişmesiyle birlikte bilgiye verilen önem artmaktadır. Bu artış çok sayıda yeni veri üretilmesine ve üretilen verinin saklanmasına yol açmaktadır. Büyük oranda verinin artmasıyla birlikte veriden anlamlı sonuç çıkarma gereksinimi veri madenciliği (data mining) kavramını ortaya çıkarmıştır. Bilgisayar sistemleri veya el ile hazırlanarak yapılan anketlerden elde edilen veriler tek başına anlam ifade etmezler. Farklı yöntem ve metotlar ile işlenen veriler anlam kazanmaya başlarlar ve işlenen veriye bakarak konu ile ilgili tahminler yapılır. Buradan yola çıkarsak veri madenciliği elimizdeki verileri analiz ederek veriler arasında ilişki oluşturma ve nihayetinde anlamlı yeni bir bilgi ortaya çıkarma sürecidir.

Veri madenciliğini tanımlayabilmek için öncelikle kelime anlamlarına bakmak faydalı olacaktır. Madencilik kelime anlamı olarak yeryüzünün kıymetli ve gizli kaynaklarının ortaya çıkarılmasıdır. Madencilik kelimesi veri kelimesi ile ilişkilendirildiğinde ise veri kümeleri içinde modelleme yapılmadan önce fark edilemeyen/bilinmeyen anlamlı ve kıymetli bilgilerin bulunarak açığa çıkarılması fikrini doğurmaktadır (Irmak vd., 2012:102).

Veri madenciliği çok miktardaki veri yığını içinden anlamlı ve faydalı tahminler yapabilmek için bilgisayar programları aracılığıyla kurallar oluşturulması ve analiz edilmesi sürecidir. Bunun yanı sıra fazla miktardaki veriler arasındaki bağlantıları inceleyerek bu veriler arasındaki benzerlik ilişkilerini bulmaya çalışan bir veri analiz tekniği olarak da tanımlanmaktadır (Savaş vd., 2012:2).

Veri madenciliğini tek başına bir çözüm olarak değil veri analizi teknikleri bütünü olarak gören Aydemir (2017:2) veri madenciliğini büyük miktardaki verilerin içinden değerli bilgilerin ortaya çıkarılması ve bu bilgiler üzerinden karar alınarak ileriye dönük tahmin edilmesi süreci olarak tanımlamıştır.

Veri madenciliğini tanımlayan diğer çalışmalara bakacak olursak; Veri madenciliği, büyük miktarlardaki verinin gözlemlenerek analiz edilmesiyle, öngörülemez sonuç ve ilişkilerinin veri bilimcisine anlaşılır bir şekilde aktarılmasıdır (Gülçe, 2010:13).

Kaçmaz (2019:12) ise veri madenciliğini veri setleri içerisindeki öngörülemez net bir dille ifade edilememiş kullanılabilir ancak gizli ve anlamlı bilgilerin ortaya çıkarılması süreci olarak tanımlamıştır (Kaçmaz, 2019:12).

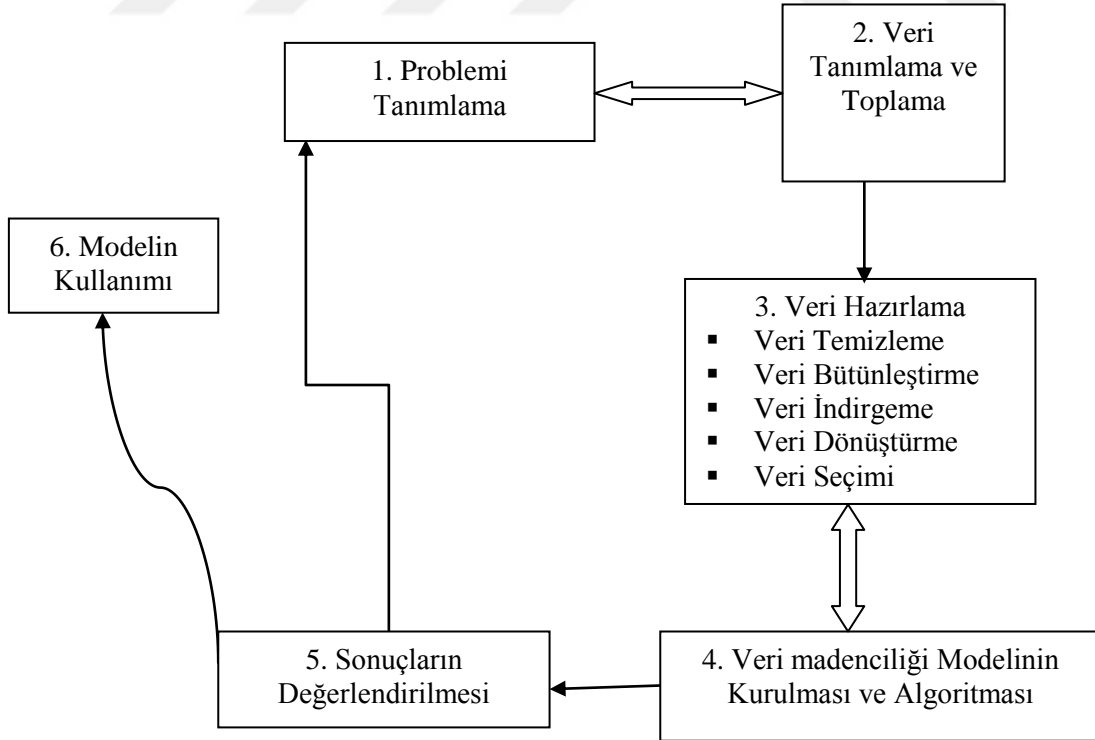
Gürel'e (2019:8) göre ise veri madenciliği; veri setlerini analiz etmek için istatistik ve yapay zeka araçları ile veri tabanı yöntemlerini birleştirerek, bilgi keşfi ve ilişkileri keşfetme süreci olarak ifade etmiştir.

Son olarak bazı anahtar kelimeleri ile veri madenciliğini 4 aşamalı ayrıntılı olarak tanımını şöyle tamamlayabiliriz;

- ✓ Veri madenciliği bir süreçtir.
- ✓ Veri madenciliği gizli kalmış bilgileri bulur.
- ✓ Veri madenciliği karar verme araçlarının niteliğini artırır.
- ✓ Veri madenciliği veri bilimcileri için kavrayış dağıtıcı bir sistemdir (Albayrak ve Yılmaz, 2009:36).

2.2. Veri Madenciliği Süreci

Veri madenciliğinin çeşitli tanımlarından da anlaşıldığı gibi veri madenciliği birbirini takip eden birçok adımdan oluşan bir süreçtir. Yapılacak tahminlerin geçerliliği için veri madenciliği yönteminin doğru sonuçlar vermesi gerekmektedir. Bu yüzden veri madenciliği sürecindeki aşamaları doğru bir şekilde yerine getirmek gerekmektedir. Veri madenciliği süreci Şekil 2.1' de gösterilmiştir.



Şekil 2.1. Veri Madenciliği Süreci (Aydemir, 2017:12).

2.2.1. Problemi tanımlama

Veri madenciliği sürecindeki ilk ve en önemli adım olan problemi tanımlanmanın amacı çalışma sürecinin planlanmasıdır. Çalışma hangi amaç için yapılacak açık bir şekilde tanımlanmalıdır ki böylece çalışma sonucunda elde edilen bilgi hangi amaç için kullanılacak belirlenmiş olacaktır. Aynı zamanda süreç içerisinde kullanılacak veriler neler, çalışmanın maliyeti ne kadar olacak, ne gibi riskler ile karşılaşılır gibi sorulara cevap aranır. Bu sorular göz önünde bulundurularak bir değerlendirme yapılır. Değerlendirme doğru bir şekilde yapılmazsa problemlere yol açar ve veri madenciliği çalışması amacına ulaşamaz.

2.2.2. Veri tanımlama ve toplama

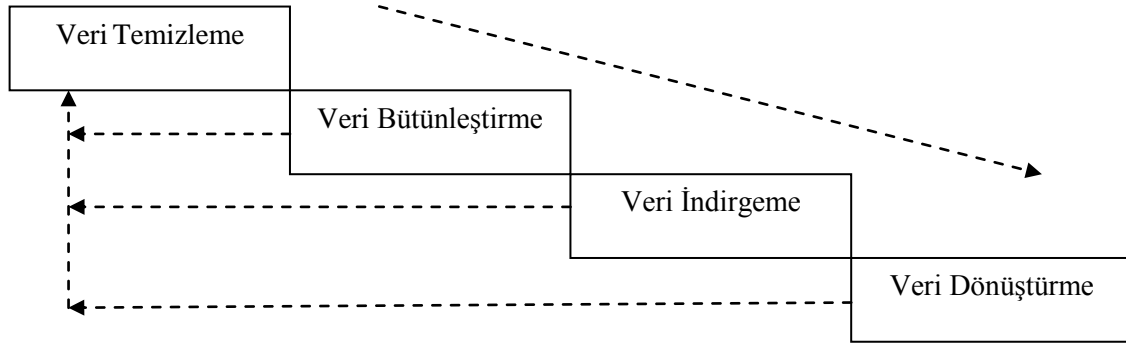
Bu aşamada verilerin nereden toplanacağı ve toplanan verinin çalışmanın amacına uygun olup olmadığı belirlenir. Daha sonra ise benzer veya aynı veriler bir araya getirilerek veri nitelikleri tanımlanır, veriler keşfedilir ve bilgiler sınıflandırılarak sürece bir sonraki adım ile devam edilir (Albayrak ve Yılmaz, 2009:36).

2.2.3. Veri hazırlama

Çalışma için toplanan verinin veri madenciliği modeline uygun hale dönüştürüldüğü aşama veri hazırlama aşamasıdır.

Veri madenciliği modelinin oluşturulması aşamasında ortaya çıkacak problemler, bu adıma tekrar tekrar geri dönülmesine ve veriler üzerinde yeniden düzenlenme yapılmasına neden olur. Böyle bir durumun oluşması veri bilimcisinin veri madenciliği süreci içerisindeki toplam zaman ve enerjisinin % 50 - % 85'lik kısmını harcamasına yol açar (Terzi vd., 2011:31).

Veri hazırlama süreci içerisindeki adımlar veri temizleme, bütünleştirme, indirgeme, dönüştürme, veri madenciliği algoritmasını uygulama ve sonuçları değerlendirip sunma olarak sıralanabilir (Özkan, 2013:40).



Şekil 2.2. Veri Hazırlama Süreci.

2.2.3.1. Veri temizleme

Veri setini oluşturan verilerin üzerinde işlem yapabilmek için istenilen özelliklere sahip olmadığı çalışmalar olabilir. Doğru verilerin yanı sıra hatalı veya eksik verilerde olabilir.

Veri hazırlama sürecinin ilk adımı olan veri temizleme, sonraki aşamalarda kullanılacak verinin veri madenciliği modelindeki kalitesini artırmak için yapılır (Zaki, 1996:3). Gürültü olarak tanımlanan eksik veya uygun olmayan veriler yapılacak analizlerde yanlış sonuçlar alınmasına neden olabileceği için sistemin bu verilerden temizlenmesi gerekmektedir. Eksik ve uygun olmayan veriler yerlerine konulacak yeni veriler ile değiştirilmelidir. Bunun için kullanılacak aşağıdaki birçok yöntemden biri veya birkaçı seçilebilir (Özkan, 2013:40).

- ✓ Eksik veya hatalı değer içeren kayıtlar veri setinden çıkarılabilir.
- ✓ Bütün kayıp değerler için aynı sabit değer olmak üzere genel bir sabit değer kullanılabilir.
- ✓ Niteliğin tüm değerleri kullanılarak ortalaması alınır ve eksik değer için bu ortalama değer kullanılabilir.
- ✓ Tüm değerlerin ortalaması yerine hatalı kaydın yapısına benzer örneklerin ortalaması alınarak bu sonuç eksik değer yerine kullanılabilir.
- ✓ Karar ağacı veya regresyon modeli oluşturularak verilere uygun bir tahmin yapılır ve bu tahmin eksik olan değer yerine kullanılabilir.

2.2.3.2. Veri bütünleştirme

Aynı veri kümesinde saklanan verilerin bir arada bulunması ile veriler bütünleştirilerek, veri analizi ve raporlama gibi işlemlerin yapılmasını sağlar (Köse, 2018).

Yani farklı kaynaklardan elde edilen farklı türdeki verilerin kullanılabilmesi için tek tipe dönüştürülmesi veri bütünleştirme olarak ifade edilir. Örneğin cinsiyet bilgilerinin alındığı bir nitelikte cevaplar "kadın" ve "erkek" olabileceği gibi "K" ifadesi kadını "E" ifadesi erkeği temsil ediyor olabilir. Ya da bu cevaplar 0 veya 1 değerleri ile de tutuluyor olabilir. Başka bir örnekte veri setini oluşturmak amacıyla öğrencilerin mezuniyet bilgileri için okul isimlerini kısa veya uzun, büyük veya küçük harf kullanılarak yazmış olabilirler. Bu gibi farklı kayıtlar veriler üzerinde çalışmayı imkansız hale getirebilir. Bu yüzden veri madenciliği modeli kurulmadan önce bu tip veriler ortak bir hale dönüştürülmelidir.

2.2.3.3. Veri indirgeme

Veri madenciliği uygulamalarında çözümlene işlemleri zaman zaman uzun sürebilir. Veri kümesi içerisinde aynı türde birçok kayıt varsa ise ve bu kayıtlardan bazılarının çalışmadan çıkarılması sonucu değiştirmeyeceğine inanılıyorsa, veri sayısı azaltılarak veri indirgeme işlemi yapılabilir.

Böyle bir durumda verilerin daha az yer kaplamaları için veri sıkıştırma işlemi yapılabilir. Diğer bir yöntem bazı öznelikleri ifade etmede büyük veri setleri yerine daha küçük veri kümeleri kullanılabilir veya benzer özelliğe sahip birçok nitelik yerine daha az nitelik kullanılarak işlemlerin daha hızlı yapılması sağlanabilir (Gülçe, 2010:17).

2.2.3.4. Veri dönüştürme

Veri madenciliği uygulamalarında bazen veri setindeki verileri aynen değerlendirmeye almak yani işleme tabi tutmak analiz için çok uygun olmayabilir. Bazı değerlerin ortalama, varyans ve standart sapmaları diğer değerlerden uzak olması durumlarında farka sebep olan büyük değerlerin diğer değerler üzerinde etkisi fazla olur ve değeri az olan değişkenlerin rollerini büyük oranda azaltırlar. Bu gibi durumlarda verilerin standartlaştırılması gerekir. Veri madenciliğinde veri standartlaştırmak için Z-score normalleştirme ve Min-Max normalleştirme yöntemleri kullanılır. Veri standartlaştırma yöntemleri yanı sıra değişkenlerin aralık değerleri göz önünde bulundurularak gruplama yapılabilir. Bu duruma örnek olarak milli eğitim bakanlığı not baremi olan 0-100 aralığındaki notlar 5'lik sisteme göre gruplandırılabilir. Böylece sistem analiz için daha uygun hale getirilmiş olur.

- **Min-Max normalleştirme**

Veri kümeleri içinde belli bir aralığın tamamının kendi içerisindeki oranları bozulmadan 0-1 arasındaki sayısal bir değerle ifade etmek için kullanılan veri dönüştürme yöntemidir. Bu yöntem için kullanılan denklem de veri içindeki aynı sütunda bulunan en

büyük değer (sayısal) ile en küçük değer belirlenerek diğer değerlerin buna uygun şekilde dönüştürülmesidir.

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2.1)$$

X^* : dönüştürülmüş değer

X : gözlem değeri

X_{min} : en küçük gözlem değeri

X_{max} : en büyük gözlem değeri (Budak, 2013).

- **Z-Score normalleştirme**

Z-Score normalleştirme, gözlemlerin aritmetik ortalamasının, gözlem değerinden çıkarılması sonucunda oluşan değer, ortalamaya en yakın, en doğru sonucu veren hesaplama yöntemi olan gözlem değerlerinin standart sapmasına bölünmesiyle hesaplanır. 0 ortalamalı normalleştirme olarak da bilinen Z-score normalleştirmede değişkenin herhangi bir y değeri, verilerin ortalaması ve standart sapmasına bağlı olarak normalleştirilir (Oğuzlar, 2003:73).

$$X^* = \frac{X - X_{ort}}{\sigma_x} \quad (2.2)$$

X^* : dönüştürülmüş değer

X : gözlem değeri

X_{ort} : gözlemlerin aritmetik ortalaması

σ_x : gözlem değerlerinin standart sapması

2.2.4. Veri madenciliği modelinin kurulması ve algoritmasının uygulanması

Veri madenciliği modelini kurabilmek ve algoritmanın uygulanabilmesi için yukarıda ifade edilen adımlardan gerekli görünenler yapılır. Eldeki problem için en uygun modelin bulunabilmesi, mümkün olduğunca çok sayıda modelin kurulup denenmesi ile mümkündür. Bu sebeple veri hazırlama ve model oluşturma aşamaları, en doğru ve en iyi olduğuna inanılan modele ulaşıncaya kadar süreç tekrarlanır (Terzi vd., 2011:32). Veri hazır hale geldikten sonra konuya uygun veri madenciliği algoritması belirlenir. Bu algoritmalar genellikle istatistiksel tabanlı olmakla birlikte, kümeleme, sınıflandırma ve birliktelik kuralları olarak sıralanabilir (Özkan, 2013:44).

2.2.5. Sonuçların değerlendirilmesi

Tüm aşamalar sonucunda elde edilen veriler ile oluşturulmuş model hayata geçirilmeden önce tüm yönleriyle bir kez daha değerlendirilir (Özçınar, 2006:9). Oluşturulan modelin uygun olup olmadığı hayata geçirilip geçirilmeyeceği yorumlanır. Problemin tüm yönleri için bir çözüm sağlayıp sağlamadığı kararına bağlanarak sonuçların doğruluğu kontrol edilir. Model oluşturma aşamasında kullanılan farklı teknikler arasında karşılaştırmalar yapılır. Model kullanılmaya uygun görülmezse problemin tanımlanması aşamasına geri dönülerek problem tekrar tanımlanır.

2.3. Veri Madenciliği Yöntemleri

Gün geçtikçe ilerleyen teknoloji verilerin kolay bir şekilde saklanabilmesini ve gerekli durumlarda erişimini kolaylaştırmakla kalmayıp veri işleme işlemlerinin her geçen gün daha ucuza yapılmasını sağlar. Birçok veri madenciliği yöntemleri, veri kümelerinden belli bir amaç doğrultusunda anlam yüklü sonuçlar çıkararak doğru kararlar alabilmek için geliştirilmektedir (Aydemir, 2017:17). Bölüm 2.1'de farklı yazarlara ait veri madenciliği tanımları verilmiştir. Tanımlardan yola çıkarsak veri madenciliğinin değişerek güncellenen ve değişmeyen taraflarından bahsedilebilir.

Veri madenciliği metotlarına her geçen gün yeni yöntem ve algoritmalar eklenmektedir. Bu metotların bir kısmı yıllardır kullanılan istatistiksel yöntemleri temel alan klasik teknikler iken, diğer bir kısmı ise istatistiği temel almakla birlikte makine öğrenimini de içeren ve yapay zeka destekli yeni nesil metotlardır (Savaş vd., 2012:8). Veri madenciliğinde kullanılan yöntemler, tanımlayıcı (Descriptive) ve tahmin edici (Predictive) olmak üzere iki başlık altında incelenmektedir. Tanımlayıcı modellerde karar vermeye öncülük olabilecek mevcut verilerdeki ilişkilerin tanımlanması sağlanır. Tahmin edici modellerde ise sonuçları var olan verilerden hareketle bir model oluşturulur. Oluşturulan modelden yararlanarak sonuçları kestirilemeyen/bilinmeyen veri kümesi içindeki verilerin sonuçlarının tahmin edilmesi amaçlanmaktadır (Ayık vd., 2017:444).

Veri madenciliği yöntemleri işlevlerine göre;

1. Sınıflama (Classification)
2. Regresyon (Regression),
3. Kümeleme (Clustering),
4. Birliktelik Kuralları (Association Rules),

olmak üzere 4 grupta incelenmektedir. Kümeleme ve Birliktelik Kuralları modelleri tanımlayıcı, Sınıflama ve Regresyon modelleri ise tahmin edici modellerdir (Savaş vd.,

2012:9). Ayrıca diğere veri madenciliđi yöntemlerinin bazıları da; Faktör Analizi, Bayesci Ağlar, Temel Bileşenler Analizi, Bulanık Mantığa Dayalı Yöntemler, Genetik Algoritmalar, Pürüzlü (Rough) Küme Teorisine Dayalı Yöntemler, Diskriminant Analizi, Kohonen Ağları olarak sıralanabilir. Sıralanan bu yöntemler dışında birden çok tekniđi içiren hibrid yöntemler ve zaman serilerine dayalı yöntemlerden de veri madenciliđi yöntemi olarak faydalanılmaktadır (Koyuncugil ve Özgülbaş 2009:26).

Veri madenciliđi temel olarak veriler üzerinde analizler yapmak ile ilgilenen bir süreçtir. Analiz yapabilmek için ise veriyi doğrulamaya dayalı yöntemler veya keşfetmeye dayalı yöntemlere ihtiyaç vardır. Yöntem ile kastedilen aranan bilgiye ulaşırken veriler üzerinden nasıl bir yol izleneceđini belirtmektir. Veri setleri içerisindeki veriler ile işlem yaparak yeni verilerin üretilmesini sağlayan yöntem keşfetmeye dayalı yöntemlerdir. Dolayısıyla verilerin keşfedilmesi yeni bilgilerin tahmin edilmesiyle veya tanımlanmasıyla mümkündür. Tanımlara bakıldığında veri madenciliđi yöntemleri “predictive” tahmin edici ve “descriptive” tanımlayıcı olarak incelenmektedir. Doğrulamaya dayalı yöntemler ise yeni bilgi üretmezler. Adından da anlaşıldığı üzere sorgu ve raporlamayı amaç olarak daha çok basit istatistiksel analizler yaparlar.

Keşfetmeye dayalı yöntemler içerisinde saydığımız tahmin edici yöntemler var olan veriler üzerinden bilinmeyen yeni bilgilerin tahmin edilmesinde kullanılırlar. Tahmin edici yöntemler olan “classification” sınıflama, “regression” regresyon ve “time series” zaman serileri gibi yöntemler ile ileriki bir durum tahmin edilebilir. Diğere bir yöntem olan tanımlayıcı yöntemler ise veri setleri içindeki saklı ortak özellikleri ve bu özellikler arasındaki ilişkileri açığa çıkarır. “Clustering” kümeleme, “summarization” özetleme ve “association rule mining” birliktelik kuralı gibi yöntemler de tanımlayıcı yöntemler içerisinde sayılırlar. Gerek tahmin edici yöntemler gerekse tanımlayıcı yöntemler olsun, her iki yöntem de veri madenciliđi teknikleriyle desteklenir (Gürel, 2019:18).

2.3.1. Tahmin edici yöntemler

Tahmin akla, mantığa veya bir takım verilere dayanarak bir durumu ya da gelecekteki bir olayı ön görmedir.

Tahmin edici veri madenciliđi yöntemlerinde sonucunun ne olduğu bilinen verilerden yola çıkarak bir model oluşturulur. Kurulan bu yeni model ile sonuç değerleri bilinmeyen veri setleri için yeni sonuç değerlerinin tahmin edilmesi çalışılır (Arslan, 2018:10). Örneğin ortaokul son sınıf öğrencilerinin temel dersleri olan Matematik, Türkçe ve Fen Bilgisi notlarına bakarak lise öğreniminde başarılarının tahmin edilmesi gibi. Tahmin edici yöntemler

karar verme süreçlerinde etkin bir rol oynarlar. Bu yöntemlerin amacı, sonuçları bilinen verilerden hareket ederek bir model oluşturmak ve bu model aracılığıyla sonuç değerleri bilinmeyen veri kümeleri için tahminde bulunmaktır. Sınıflandırma yöntemi ve regresyon analizi tahmin edici yöntemler olarak bilinmektedir.

2.3.1.1. Sınıflandırma yöntemleri

Sınıflandırma verileri önceden tanımlanmış uygun alanlara atamaya dayalı bir yapıdır. Daha farklı bir ifade ile sınıflandırma nitelik kümesi olan A'nın sınıf etiketi olan B ile eşleşmesi olarak ifade edilir (Özdemir, 2016:10). Sınıf olarak tanımlama yapabilmek için veri seti içindeki her bir verinin diğer veriler ile belirlenmiş ortak bir özelliği olmalıdır. Burada önemli bir nokta her bir veri sınıfının özellikleri daha önceden belirlenmelidir. Veri sınıflandırılması yapabilmek için belli bir sıra izlenmelidir. Mevcut veri setinin bir kısmı verinin eğitimi için kullanılır ve sınıflandırma kuralları oluşturulur. Oluşturulan bu kurallar ile gelecekte karşılaşılan yeni bir verinin hangi sınıfa dahil olacağı belirlenir. Sınıflandırma en çok bilinen ve kullanılan veri madenciliği yöntemlerinden birisidir.

Sınıflama modelinde kullanılan başlıca yöntemler şunlardır (Özdemir, 2016:12):

Karar Ağaçları (Decision Trees)

Yapay Sinir Ağları (Artificial Neural Networks)

Bayes Sınıflandırması

Destek Vektör Makineleri

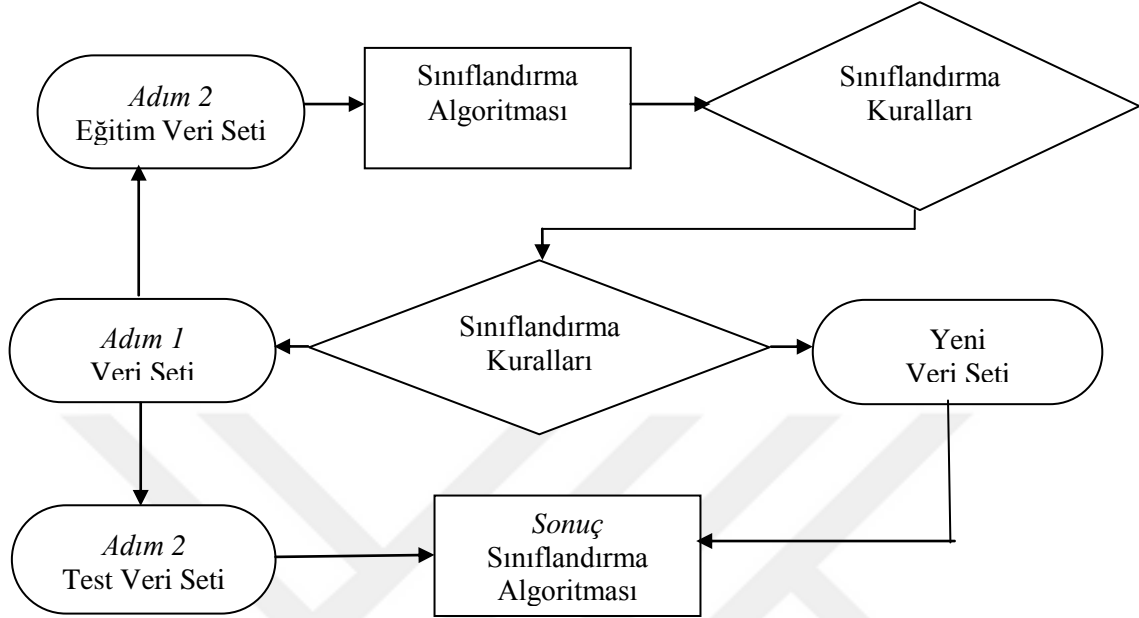
K-En Yakın Komşu (K-Nearest Neighbor)

Genetik Algoritmalar (Genetic Algorithms)

- **Karar ağaçları (decision trees)**

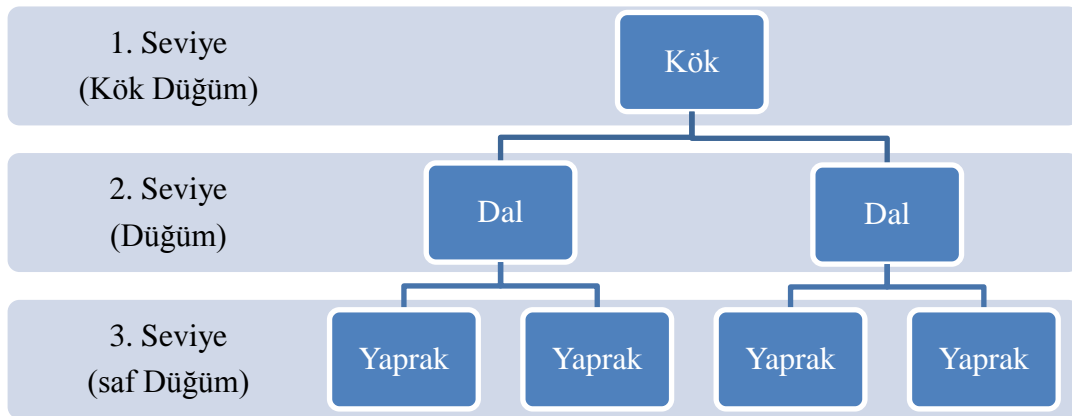
Veri madenciliği yöntemlerinden tahmin edici modeller arasında yer alan karar ağaçları (decision trees), anlaşılması kolay kurallar üretebilmesi ve ürettiği kuralları görselleştirebilen bir yapıya sahip olması sebebiyle en etkili olan ve en sık kullanılan sınıflandırma yöntemidir. Aynı zamanda karar ağaçları verilerin hızlı bir şekilde işlenmesine de olanak sağlarlar. Bu da karar ağaçlarını belki de en popüler sınıflandırma tekniği yapmaktadır. Karar ağaçları temel olarak veri setini eğitim (training) verisi ve test (test) verisi olmak üzere ikiye ayırır. Eğitim verisi için ayrılan kısım karar ağacını oluşturarak sınıf kurallarının belirlenmesi için kullanılır. Test verisi için ayrılan veriler ile de oluşturulmuş olan

sınıflandırma kuralları denenerek karar ağacının başarısı test edilir. Algoritma oluşturduğu sınıflandırmada başarılı olursa veri setine eklenen yeni verilerin algoritma sonucunda oluşan kurallara göre sınıflandırılması yapılır.



Şekil 2.3. Veri Sınıflandırma İşlemi (Pala, 2013:51).

Karar ağaçlarının yapısı kök düğüm, düğüm ve saf düğümlerden oluşur. Ağaç yapısı kökten dallara, dallardan yapraklara doğrudur. En üsteki düğüm kök düğümüdür. Ağacın dal ve yapraklarının her biri en uygun sınıflandırma sonucunu verebilmek için algoritmalar yardımıyla ayrılır (Pala, 2013:51). Şekil 2.4 bir karar ağacı yapısını göstermektedir.



Şekil 2.4. Karar Ağaçları Yapısına Örnek Bir Dallanma (Gürel, 2019:22).

Şekil 2.4'teki karar ağacı dallanması incelendiğinde yapıdaki ilk düğüm kök düğümüdür. Karar ağacı üzerinde sınıfı bilinmeyen bir verinin sınıfına karar vermek için karar kurallarına bakılır. Sınıfı bilinmeyen verinin değeri karar ağacı üzerinde teste tabi tutulur.

Şekilde de görüldüğü gibi oluşturulan ağaç yapısı üzerinde 1. seviyedeki kök düğümden saf düğümlere (yapraklara) doğru gidilerek verinin sınıfı kolayca bulunur (Pala, 2013:8).

Son zamanlarda çok sayıda farklı karar ağacı öğrenme teknikleri geliştirilmiştir. Bunlardan en popüler olanları ID3, C4.5, C5 ve J48 algoritmalarıdır (Sakarya, 2019:14).

- **Yapay sinir ağları (artificial neural networks)**

Yapay sinir ağları (YSA) insan beyninin bir özelliği olan öğrenme yoluyla yeni bilgi oluşturma, türetebilme ve keşfedebilme gibi özelliklerini yardım almadan gerçekleştirebilmek amacıyla geliştirilmiş sınıflandırma algoritmalarıdır (Selvi, 2010:28). Yapay sinir ağları aynı zamanda biyolojik sinir ağları mantığı ile çalışan yani onları taklit edebilen bilgisayar programıdır. Bu sayede bellek yönetimi, sınıflandırma, örüntü tanıma, kontrol ve optimizasyon gibi birçok alanda problemleri kolay ve etkili bir şekilde çözebilmektedir (Bilen, 2014:10).

Birbirine bağlı birçok sayıda hesaplama yapabilen sinir birimlerinden oluşan yapay sinir ağları insan beyninin bilişsel benzetimine dayalı bilgi işleme sistemleridir. Yapay sinir ağları olayların örneklerine bakarak, olaylar hakkında genelleme yapmakta, bilgi toplamakta ve sonrasında hiç karşılaşmadığı bir örnekle karşılaştığında eğitim sonucunda öğrendiği bilgiler ile kararlar verebilmektedir (Pala, 2013:13; Bilen, 2014:10).

Yapay sinir hücreleri ile biyolojik sinir hücreleri karşılaştırıldığında bu sistemler arasındaki benzerlikler Tablo 2.1'de verilmiştir.

Tablo 2.1. Yapay Sinir İle Biyolojik Sinir Hücresi Arasındaki Benzerlikler (Pala, 2013:13).

Biyolojik Sinir Sistemi	Yapay Sinir Ağı
Nöron	Algılayıcı
Dentrit	Toplama işlevi
Hücre gövdesi	Etkinleştirme işlevi
Aksonlar	Algılayıcı çıkışı
Sinapslar	Ağırlıklar

Normal bir sinir ağı giriş katmanı, saklı katman ve çıkış katmanı olmak üzere 3 tip katmandan oluşur. Burada dikkat edilmesi gereken yer, ağda farklı tipte üç katmanın olduğudur. Giriş katmanı ve çıkış katmanı sayısı sabit iken isteğe bağlı olarak saklı katman sayısı değiştirilebilir. Giriş katmanı giriş verilerini içerirken, işlemler yapıldıktan sonra saklı katmanlarda oluşturulan sonuçlar çıkış katmanında yer alır (Yu vd., 2010:315).

Yapay sinir ağları veri madenciliği yöntemleri içerisinde oldukça kullanışlı bir yöntemdir. Ancak bunun yanı sıra anlaşılması ve yorumlanması zor modeller ortaya çıkardığı

için uygulaması diğer sınıflandırma yöntemlerine göre daha uzun zaman alır. Bu duruma rağmen YSA'lar en az karar ağaçları kadar yaygın olarak kullanılmaktadır.

- **Bayes sınıflandırması**

Silahtaroglu'na (2008:50) göre bayes sınıflandırma yöntemi veri setine eklenen yeni bir verinin, eldeki mevcut sınıflara ayrılmış verilere bakarak hangi sınıfa ait olacağını olasılığını hesaplayan istatistiksel bir sınıflama yöntemidir. Örneğin boyu ve kilosunu verilen kişilerin cinsiyetinin tahmin edilmesi gibi.

Bayes teoremi şu şekilde formüle edilir (Bahadır, 2008:50):

$$p(A|B) = \frac{p(A) \times p(A|B)}{p(B)} \quad (2.3)$$

$p(A)$: A olayının olma olasılığı

$p(B)$: B olayının olma olasılığı

$p(A|B)$: B olayı olduğu zaman A olayının olma olasılığı

$p(B|A)$: A olayı olduğu zaman B olayının olma olasılığı

- **Destek vektör makineleri**

Destek vektör makineleri (DVM) yöntemi ile oluşturulan modeller yapay sinir ağları ile doğrudan ilişkilidir. DVM'ler sigmoid fonksiyonunu kullanan ileri beslemeli, iki katmanlı bir YSA ağına sahiptir. Destek vektör makinelerinin en önemli özelliği ise veri setlerindeki ortalama hata karesini en aza indirmektir (Kavzoğlu ve Çölkesen, 2010:75).

- **K-En yakın komşu (k-nearest neighbor)**

Veri madenciliği yöntemlerinden bir diğeri olan KNN algoritması diğer bir adıyla K-En yakın komşu algoritması örnek tabanlı öğrenme yöntemleri arasında yer almaktadır. Bu tür algoritmalarda yeni veriyi sınıflandırmayı öğrenme işlemi eğitim veri setinde yer alan veriler ile gerçekleşmektedir. Yani sınıfı bilinmeyen yeni bir örnek eğitim veri setindeki diğer veriler ile aralarındaki benzerliğe bakılarak en uygun şekilde sınıflandırılır (Taşcı ve Onan, 2016). Makine öğrenme algoritmaları içinde en çok kullanılan ve bilinen yöntemlerden biri olan KNN algoritmasında sınıflandırma yaparken kullanılan temel mantık verilerin özelliklerine göre hangi sınıfta yer alabileceğini bulmaktır (Kılınç vd., 2016:90).

- **Genetik algoritmalar (genetic algorithms)**

Doğal seçim ilkelerine dayanarak arama ve optimizasyon yapan genetik algoritmalar gün geçtikçe uygulama alanları genişerek artmakta olan bir araştırma tekniğidir (Emel ve Taşkın, 2002:130; Aydemir, 2017:21). Aynı zamanda genetik algoritmalar diğer yöntemler ile de bir arada kullanılarak hibrid sonuçlar doğmasına neden olmaktadır.

2.3.1.2. Regresyon analizi

Regresyon analizinde bağımlı ve bağımsız değişkenlerden söz edilmektedir. Bağımlı değişkenlerin birden çok bağımsız değişken ile fonksiyon halinde ilişki kurmasıyla ifade edilir. Kurulan bu fonksiyon ile bağımlı değişkenlerin ulaşabileceği değerler tahmin edilmeye çalışılır (Gürel, 2019:19).

2.3.2. Tanımlayıcı yöntemler

Tanımlayıcı veri madenciliği yöntemlerinde karar vermeye rehberlik edebilecek veri bağlantıları veya örüntüleri tanımlamada kullanılır. Örneğin bir bilgisayar kullanıcısının bir download sitesinden yeni bir program indirdiğinde, kullanıcıya o programa benzer başka programlarında önerilmesi tanımlayıcı yöntemlerinin veri madenciliğinde kullanıldığının bir örneğidir. Kümeleme yöntemi ve birliktelik kuralları en yaygın olarak kullanılan tanımlayıcı yöntemlerdir.

2.3.2.1. Kümeleme yöntemi

Kümeleme yöntemi veri madenciliğinde tanımlayıcı yöntemler içerisinde yer aldığı için sınıflandırma algoritmalarında olduğu gibi farklı farklı algoritmaları sonuç metrikleri ile karşılaştırarak en iyi sonucu veren algoritmayı seçmek gibi bir imkanı söz konusu değildir (Akçapınar vd., 2016:56). Kümeleme yöntemi adından da anlaşıldığı üzere bir veri seti içindeki örnekleri belli bir yakınlık derecelerine göre adına küme denilen gruplara ayırır (Sarıman, 2011:193).

Kümeleme metotunda gruplar oluşturulurken kümeler benzersiz, küme içindeki veriler birbiri ile en benzerleri seçilir. Hiyerarşik yöntemler, ızgara tabanlı yöntemler, bölümlenme yöntemleri, yoğunluk tabanlı yöntemler ve model tabanlı yöntemler en çok kullanılan kümeleme yöntemleridir (Özekes, 2003:71).

2.3.2.2. Birliktelik kuralları

Olayların bir arada gerçekleşme durumlarını inceleyen birliktelik kuralları eldeki mevcut kayıtlardan başka yeni bir kaydın var olma durumunu araştırır (Bozkır, 2009:62).

Birliktelik kurallarının en çok kullanıldığı uygulamalar perakende satışın söz konusu olduğu durumlarda müşterilerin satın alma eğilimlerini bulmaktır (Özkan, 2008:78). Bu sebeple literatürde geçen birçok çalışmada birliktelik kuralları "pazar sepeti" olarak da isimlendirilmektedir (Silahtaroglu, 2008:64).

2.4. Veri Madenciliği Kullanım Alanları

Veri madenciliği; eğitim, bilim ve mühendislik, zekâ, genetik, sağlık, bankacılık, sinyal işleme, borsa, sanayi, pazarlama yönetimi, telekomünikasyon, perakende satış, biyoloji, sigorta, elektronik ticaret, inşaat gibi birçok alanda kullanılmaktadır (Silahtaroglu, 2008:72; Özkan, 2008:46; Kaya ve Özel, 2014:48).

2.4.1. Eğitim alanında veri madenciliği

Birçok alanda olduğu gibi eğitim alanında da veri madenciliği kullanılarak oldukça fazla çalışma yapılmıştır. Literatürde yapılmış olan çalışmalarda genel amaçlar şöyle sıralanabilir: Eğitimin her kademesinde öğrencilerin sınıf geçme ve başarı/başarısızlık durumlarını belirlemek, öğrencilerin mezun oldukları lise türünün hangi üniversite bölümünü kazanması ile arasındaki ilişkiyi belirlemek, öğrencileri ailelerinin sosyo-ekonomik düzeyi ile öğrenme düzeyi arasındaki ilişkiyi incelemek, öğrencilerin lise veya üniversiteye geçişte tercih sıralarını belirlemek, bireylerin meslek seçiminde kişisel özelliklerinin etkisini belirlemek, rehberlik çalışmalarında rehberlik sürecini kaliteli bir şekilde yürütebilmek, öğrencilerin başarısızlığını önlemek için başarıyı etkileyen etmenleri bulmak, öğrencilerin sportif faaliyetlere/ders dışı etkinliklere katılmaları ile akademik başarıları arasındaki ilişkiyi belirlemek (Savaş vd., 2012:15; Gürel, 2019:11; Ayık vd., 2017:443; Kurt vd., 2012:112).

2.4.2. Sağlık alanında veri madenciliği

Veri madenciliği sağlık sektöründe yapılan birçok çalışma için oldukça önemlidir. Sağlık verileri üzerinde veri işleme çalışmalarının karmaşıklığını en aza indirmek için veri madenciliği yöntemleri yaygın bir şekilde kullanılmaktadır. Örneğin hastalıkların teşhis edilmesinde ve hastaya uygulanacak tedavi ve sürecinin belirlenmesinde, uygulanan tedavinin başarısının belirlenmesinde, hasta verilerini belli kriterlere göre sınıflandırılmasında, ürün geliştirme, ameliyatlarda veya uygulanan tedavilerde risk faktörlerinin belirlenmesinde veya yerleşim bölgelerine göre hastalık haritalarının çıkarılmasında (Aydın, 2014:32; Aydın vd.,2017:52; Durairaj vd., 2013:32).

2.4.3. Bankacılık ve finans alanında veri madenciliği

Bankacılık ve finans sektörlerinde veri madenciliği genellikle müşteri profillerini belirlemek, ne zaman ve niçin tercihte buldukları gibi soruların cevabını bulmak için kullanılmaktadır. Bunların yanı sıra veri madenciliğine hisse senedi fiyatlarını ve finansal tahminlerde bulunmak, yeni yatırımları yönetebilmek için risk analizi yapmak, pazarlama stratejilerini belirlemek, dolandırıcılığı önlemek için tahminlerde bulunmak, ücret yönetimi ve kart limitlerini belirlemek gibi uygulamalarda sıkça rastlanmaktadır. Türkiye'de de bazı bankalar müşteri davranışlarını modelleyebilmek için veri madenciliğine başvurmaktadır (Gürel, 2019:11).

2.4.4. Telekomünikasyon alanında veri madenciliği

Veri madenciliği telekomünikasyon alanında yani iletişim de mobil kullanıcılarının hareketlerini belirleyerek gelecekteki hareketleri tahmin etmek, oluşabilecek sahtekarlığı tahmin ederek önlemler almak, abonelik tespiti yapmak, analizlerde insan gücünü azaltmak, belli zaman aralığında daha fazla görüşme yapan müşterilerin hareketlerini gözlemlemek, hat yoğunluklarını tahmin etmek, en önemlisi de müşteri kaybının önüne geçmek için kullanılmaktadır (Gürel, 2019:12; Ünsal, 2011:23).

2.4.5. İnşaat alanında veri madenciliği

İnşaat sektöründe veri madenciliği, iş sağlığı ve güvenliği başta olmak üzere deprem analizi tahminleri ile zemin çalışmaları yapmak, proje ve yapı yönetimi, ulaşım alanlarındaki inşaat maliyetini tahmin etmek, yeni iş fırsatlarını belirlemek, olası iş kazalarını tahmin ederek önlem alma çalışmaları yapmak, iş ve işveren arasındaki motivasyonu artırmak, yatırım geleceğinde tahminlerde bulunmak gibi birçok konularda kullanılmaktadır (Keleş vd., 2014:833; Kaya Keleş, 2017:236; Kaya Keleş vd., 2017:130).

2.4.6. Mühendislik alanında veri madenciliği

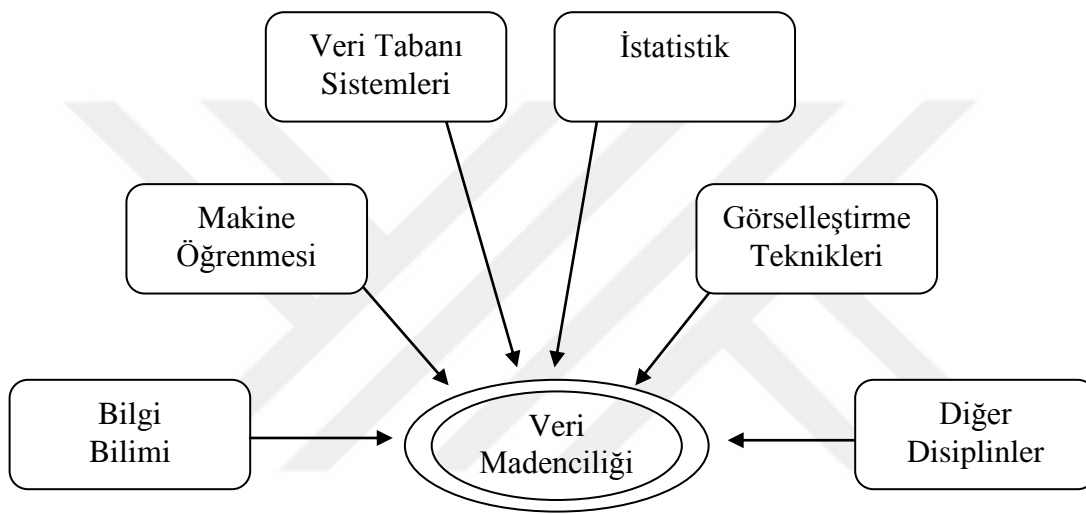
Bilgisayar, mühendislik, matematik, yazılım, yer bilimleri gibi birçok bilimsel alanlarda veriler toplanmaktadır. Bu yüzden birçok alanda olduğu gibi veri madenciliği mühendislik ve bilim alanında da sıkça kullanılmaktadır. Firmalar yazılım süreçlerini geliştirmek ve yönetmek, bu sürecin hızını artırarak zamandan tasarruf etmek, çeşitli tahmin analizleri yaparak rekabet ortamı sağlamak, üretim sürecini iyileştirerek kaliteyi artırmak, kalite kontrol uygulamaları yapmak, ekosistem modellemek ve nesne sınıflandırmak gibi birçok amaç için mühendislik ve bilim alanında veri madenciliği kullanılmaktadır (Gürel, 2019:14; Baykasoğlu, 2005:3).

2.4.7. Endüstri alanında veri madenciliği

Bilgisayar sistemlerinden elde edilen verilerin anlamlandırılarak sınıflandırılmasında, kalite kontrol uygulamalarında sistem performansını etkileyen etmenlerin bulunmasında, üretim sürecini kontrol altında tutulmasında gibi konularını temel alan endüstri alanında veri madenciliği kullanılmaktadır (Özbay, 2015:268; Shahbaz vd., 2008:789).

2.5. Makine Öğrenmesi

Veri madenciliği çok sayıda disiplin ile bir arada çalışan bir alandır. Bu alanlar makine öğrenmesi, yapay zeka, istatistik analizleri, örüntü tanıma, veri tabanı yönetimi vb. gibidir.



Şekil 2.5. Veri Madenciliği Ve Diğer Disiplinler (Özel ve Topsakal, 2014:45).

Günümüzde ilerleyen teknolojiler ve geliştirilen uygulamalar ile bilgisayarlar insanların iş gücünü büyük oranda hafifletmesine rağmen, beyin gücü hala istenilen seviyede değildir. Yapay zeka teknolojileri kullanılarak bilgisayar sistemlerinin bu konuda gelişmelerini sağlamak amaçlanmaktadır. Makine öğrenmesi ile mevcut verilerle sistemi eğitme yoluyla verileri analiz etmek ve ileride oluşabilecek durumlar hakkında tahminler yapması bilgisayarların karar verme sürecindeki rollerini artırılması ön görülebilir.

Makine öğrenmesi bilgisayarların geçmiş veri analizlerinden yararlanarak, gelecekteki bir durumu tahmin etme ve modelleme yapma imkanı sağlayan yapay zeka alanıdır. (Ünsal, 2011:30). Makine öğrenmesi adında da yer aldığı gibi aslında bir öğrenme sürecidir. Bu sebeple veri bilimcinin verileri nasıl topladığı ile veya verilerin hangi formatta kullanacağı gibi konular ile ilgilenmez. İlgilendiği nokta veriler üzerinde gelecekte nasıl doğru tahminler yapabileceğidir.

Dalyan'a (2006) göre makine öğrenmesi eldeki mevcut veriler ile geçmişteki deneyimlerden faydalanarak bilgisayarı en iyi performans ile programlamaktır. Bunu için de oluşturulan model aşağıdaki 3 amaçtan birine hizmet etmektedir:

- ✓ Tahmin edici model kullanılarak, ilgili konudaki verilerin geleceği ile ilgili tahminde bulunmak.
- ✓ Tanımlayıcı model kullanılarak, verilerin tanımlamasını yapmak.
- ✓ Her iki modeli de birlikte kullanmak.

Makine öğrenmesi çıkarımlar yaparak çalışmaktadır. Yapılan çıkarımlar ya tanım yapmak için ya da tahminde bulunmak için kullanılır. Eğer veriler tahmin için kullanılacaksa oluşturulan model tahmin edici modeldir ve eğitim aşaması ve test aşaması olmak üzere iki bölüme ayrılır. Veri setindeki verilerin büyük bir kısmı ayrılarak model oluşturulması eğitim aşamasında gerçekleştirilir. Model oluşturmada kullanılan bu veri eğitim verisi olarak isimlendirilir. Makineyi eğitmek için ayrılan veri kümesinin verisi ne kadar olmalı veya hangi yöntem ile seçilmeli bu veri bilimcisine bağlıdır. Bu kısımda dikkat edilmesi gereken konu model oluşturmada kullanılan veri seçilen örnek verileri değil veri setindeki tüm veriyi temsil etmelidir. Test aşaması için ayrılan veriler ise iki şekilde olabilir. Birincisi eğitilmiş modele, veri setinden eğitim aşaması için ayrılan verilerden geri kalanlar sunulur. İkincisi ise ileride toplanacak veriler sunulur. Eğitimden önce veri olarak isimlendirilen nitelikler, eğitimden sonra bilgi olarak ifade edilir. Bu bilgiler de tahmin için kullanılır.

Şekil 2.1'de veri madenciliği sürecinde de görüldüğü üzere, makine öğrenmesinin uygun metodlarından herhangi birini kullanmak sürecin bir adımıdır.

3. UYGULAMA

Çalışmanın uygulama kısmı iki aşamadan oluşmaktadır. Birinci aşamada Bilecik Şeyh Edebali Üniversitesinden alınan 15/11/2019 tarihli 7 numaralı Toplantı Etik Kurul Raporuyla birlikte Bilecik İl Milli Eğitim Müdürlüğü 27/11/2019 tarihli ve 21174680 sayılı valilik makam oluru ile anket çalışması yapabilmek için AYSE Projesi Yönetim Paneline kayıt olarak gerekli yazışmaları yapmaktır. İkinci aşamada ise alınan izinler doğrultusunda kişisel veriler içermeyecek şekilde veri setini oluşturup bu verileri veri madenciliği uygulama adımlarına uygun olarak veri madenciliği yöntemlerini kullanarak KNIME Analytics Platformunda incelemektir.

Bu bölümde sırasıyla uygulamanın amacı, veri toplama araçları, örneklem grubu, veri temizleme, veri dönüştürme, öğrencilerin demografik bilgilerine ilişkin bulgu ve yorumlar, model oluşturma ve model seçimi hakkında bilgiler verilecektir.

3.1. Uygulamanın Amacı

Bu çalışmanın amacını belirlemek için iki noktaya odaklanıldı.

- ✓ Ülkemizde ortaöğretim öğrencilerinin lise geçiş sınavlarındaki başarılarını etkileyen temel nitelikler nelerdir?
- ✓ Lise geçiş sınavlarında öğrenci başarısını tahmin etmek mümkün müdür?

Yukarıda yazılan amaçlar doğrultusunda veri madenciliği yöntemlerinin, veri setine uygunluğu göz önünde bulundurularak sınıflama algoritmaları ile Bilecik ilinde ilköğretimden liseye geçiş sınavlarında öğrenci başarısını tahmin etmede en iyi performansı verecek sınıflandırma algoritması bulunması amaçlanmaktadır.

3.2. Veri Toplama Araçları

Çalışmada kullanılan veri seti, Bilecik Şeyh Edebali Üniversitesi Bilgisayar Mühendisliği Bölümü ile Bilecik İl Milli Eğitim Müdürlüğü'nün ortak çalışması sonucu elde edilmiştir. Temel olarak iki kaynaktan oluşan veri setinde; birinci kaynak kâğıt sayfalara dayanan öğrenci akademik bilgileri (6. ve 7. sınıf Türkçe, Matematik, Fen Bilgisi ve genel başarı ortalamaları) iken ikinci kaynak Google formlar kullanılarak oluşturulmuş 32 soruluk anket çalışmasıdır.

- Birinci kaynağı oluşturmak için öğrencilerin kişisel bilgilerini kullanmamak adına sadece 6. ve 7. sınıf Türkçe, Matematik, Fen Bilgisi ve genel başarı ortalamalarını alabilmek için aşağıdaki makro komutları yazılmıştır ve Microsoft Excel programında çalıştırılmıştır.

Sub Runmacro()

Call aktar

Call sıfır_olan_satıları_sil

Call Kaydet

End Sub

Sub aktar()

Sheets("Sayfa1").Select

Range("A1:Z3000").Select

Selection.Copy

Sheets("Sayfa2").Select

Range("A1").Select

Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _

:=False, Transpose:=False

Sheets("Sayfa2").Select

Range("A1").Select

End Sub

Sub Kaydet()

Ad = CreateObject("wscript.Shell").SpecialFolders.Item("Desktop") & _

Application.PathSeparator & Format(Now, "dd_mm_yyyy_hh_mm_ss") & ".xls"

ActiveSheet.Copy

ActiveWorkbook.SaveAs Filename:=Ad

ActiveWorkbook.Close

End Sub

Sub sıfır_olan_satıları_sil()

Dim ilk As Date, son As Date, X As Byte

Application.ScreenUpdating = False

```

ilk = Time

Application.ScreenUpdating = False

Columns("A:A").Replace What:="0", Replacement:="", LookAt:=xlWhole

On Error Resume Next

For X = 1 To 3

    Range("A1:A3000").SpecialCells(xlCellTypeBlanks).EntireRow.Delete

    Next

On Error GoTo 0

son = Time

Application.ScreenUpdating = True

MsgBox "İşleminiz tamamlanmıştır." & Chr(10) & _
"İşleminizin Süresi ; " & Format((son - ilk), "hh:mm:ss")

End Sub

```

- İkinci kaynak ise aşağıdaki anket sorularından oluşmaktadır:
 1. Adınızı soyadınızı yazınız.
 2. Mezun olduğunuz ortaokulu yazınız.
 3. Cinsiyetinizi seçiniz.
 4. Yaşınızı yazınız.
 5. Şuan eğitim gördüğünüz lise türünü seçiniz.
 6. Kaç kardeşiniz var?
 7. Okuyan kardeş sayınız kaç?
 8. Kendinize ait odanız var mı?
 9. Düzenli ders çalışma alışkanlığınız var mı?
 10. Düzenli kütüphaneye gitme alışkanlığınız var mı?
 11. İlgilendiğiniz sosyal etkinlik var mı? Varsa aşağıdan seçim yapınız.
 12. Sanat derslerini (müzik, görsel sanatlar vb.) nasıl buluyorsunuz?
 13. Takviye ders (özel ders) alıyor veya dershaneye gidiyor musunuz?
 14. Hafta içi günlük yaklaşık uyku süresiniz ne kadar?
 15. Hafta içi günlük internet kullanım süresi ne kadar?
 16. Hafta içi günlük tv izleme süreniz ne kadar?

17. Günlük ders çalışma süreniz ne kadar?
18. Nerede ikamet ediyorsunuz?
19. Okula ulaşımı nasıl sağlıyorsunuz?
20. Yaşadığınız evin durumu nedir?
21. Yaşadığınız ev ne ile ısınıyor?
22. Anneniz sağ mı?
23. Babanız sağ mı?
24. Anneniz öz mü?
25. Babanız öz mü?
26. Anneniz çalışıyor mu?
27. Babanız çalışıyor mu?
28. Annenizin öğrenim durumu nedir?
29. Babanızın öğrenim durumu nedir?
30. Sürekli bir rahatsızlığınız var mı?
31. Ailenizin ortalama aylık geliri nedir?
32. 8. sınıftaki sınıf mevcudunuzu yazınız.

32 soruluk ankete verilen cevaplar ile öğrenci akademik bilgileri Microsoft Office Excel programında bir excel dosyası formatında birleştirilmiştir.

3.3. Örneklem Grubu

Çalışmanın ikinci kaynağı olan anket Bilecik ilinde sınav puanıyla öğrenci alan ve etiket kümesini oluşturacak 2019-2020 eğitim-öğretim yılında 4 farklı lise türünde okuyan 9. sınıf öğrencilerine uygulanmıştır. Bu liseler aşağıda sıralanmıştır:

- ✓ Refik Arslan Öztürk Fen Lisesi,
- ✓ Ertuğrulgazi Lisesi,
- ✓ Mesleki ve Teknik Anadolu Lisesi,
- ✓ Kız Anadolu İmam Hatip Lisesi.

Anket soruları içerisinde yer alan mezun olunan en çok 3 ortaokulu belirlemek için yapılan filtreleme sonucunda

- ✓ 78 öğrenci ile Edebalı Ortaokulu,
- ✓ 29 öğrenci ile Murat Hüdavendigâr Ortaokulu ve
- ✓ 24 öğrenci ile Dursun Fakih İmam Hatip Ortaokulu ilk üç sırada yer almaktadır.

3.4. Veri Temizleme

Veri setini oluşturmak için birinci kaynaktan elde edilen veriler ile ikinci kaynak olan ankete verilen cevaplar Microsoft Office Excel programında bir excel dosyası formatında birleştirilmiştir. Çalışmada akademik bilgi olarak 6. sınıf ve 7. sınıf bilgileri kullanılmıştır. 6. veya 7. sınıflardan birinde Bilecik ilinde olan öğrencilerin notlarının ortalaması alınarak veri setinden çıkarılması söz konusu olmamıştır. Ancak hem 6. sınıf hem de 7. sınıfın her ikisinde de Bilecik ilinde olmayan öğrencilerin akademik bilgilerine ulaşılamaması nedeniyle bu öğrenciler veri setinden çıkarılmıştır. Aynı zamanda ortaokul bazında sadece 3 ortaokul ile çalışıldığı için ilk durumda ankete cevap veren 300 öğrenci yukarıda sayılan sebeplerden dolayı veri temizleme işlemi yapıldıktan sonra 117 örneğe düşmüştür.

3.5. Veri Dönüştürme

Oluşturulan veri setini çalışmaya uygun hale getirebilmek için veriler üzerinde bazı dönüştürme işlemleri yapılmıştır. Veri setinde kullanılan 32 anket sorusu ve 7 öğrenci akademik başarısından oluşan 39 nitelik Tablo 3.1'de verilmiştir.

Tablo 3.1. Çalışmada Kullanılan Nitelikler Ve Alabileceği Değerler.

Öznitelik	Açıklama	Veri Tipi
Mezuniyet	Öğrencinin Mezun Olduğu Ortaokul	nominal ^a
Cinsiyet	Öğrencinin Cinsiyeti	binary: erkek veya kız
Yas	Öğrencinin Yaşı	nümerik: 13-16
K_Say	Kardeş Sayısı	nümerik:0-4
O_K_Say	Okuyan Kardeş Sayısı	nümerik:0-3
Oda_Sa	Kendisine Ait Oda	binary: yok veya var
Ders_Ca_Al	Düzenli Ders Çalışma Alışkanlığı	binary: yok veya var
Ku_Gi_Al	Düzenli Kütüphaneye Gitme Alışkanlığı	binary: yok veya var
Sos_Et	İlgilendiği Sosyal Etkinlik	nominal ^b
San_De	Sanat Derslerine Bakış Açısı	binary: gereksiz veya gerekli
Tak_De	Takviye Ders Alması	binary: evet veya hayır
Uy_Su	Hafta İçi Günlük Uyku Süresi	binary: 9 saatten fazla veya 9 saatten az
Int_Ku_Su	Hafta İçi Günlük İnternet Kullanım Süresi	binary: 1 saatten fazla veya 1 saat
Tv_Iz_Su	Hafta İçi Günlük Televizyon İzleme Süresi	nominal ^c
Ders_Ca_Su	Ders Çalışma Süresi	binary: 2 saatten az veya 2 saatten fazla
Oku_Ul	Okula Ulaşım Durumu	nominal ^d

Ev_Du	Yaşadığı Evin Durumu	binary: sahibi veya kiracı
Isi	Evin Isınma Türü	binary: soba veya Doğalgaz/Katı yakıt
A_S_M	Anne Sağ Mı?	binary: evet veya hayır
B_S_M	Baba Sağ Mı?	binary: evet veya hayır
A_O_M	Anne Öz Mü?	binary: evet veya hayır
B_O_M	Baba Öz Mü?	binary: evet veya hayır
An_Ca	Anne Çalışıyor Mu?	binary: evet veya hayır
Ba_Ca	Baba Çalışıyor Mu?	binary: evet veya hayır
A_O_D	Anne Öğrenim Durumu?	nominal ^e
B_O_D	Baba Öğrenim Durumu?	nominal ^f
Su_Ra	Sürekli Rahatsızlık Durumu	binary: evet veya hayır
Ay_Ge	Ailenin Aylık Ortalama Geliri	nominal ^g
O_Turk	6. ve 7. Sınıf Türkçe Ortalaması	nümerik: 0-100
O_Turk_Sis	6. ve 7. Sınıf Türkçe Ortalaması 5'lik Sistemdeki Notu	nümerik: 0-5
O_Mat	6. ve 7. Sınıf Matematik Ortalaması	nümerik: 0-100
O_Mat_Sis	6. ve 7. Sınıf Matematik Ortalaması 5'lik Sistemdeki Notu	nümerik: 0-5
O_Fen	6. ve 7. Sınıf Fen Bilgisi Ortalaması	nümerik: 0-100
O_Fen_Sis	6. ve 7. Sınıf Fen Bilgisi Ortalaması 5'lik Sistemdeki Notu	nümerik: 0-5
Ort_Bas	6. ve 7. Sınıf Yılısonu Genel Başarı Ortalaması	nümerik: 0-100
LiseTur	Kazanılan Lise Türü	nominal ^h

^a 0- Edebali Ortaokulu, 1 - Murat Hüdavendigâr Ortaokulu, 2 - Dursun Fakih İmam Hatip Ortaokulu

^b 0- Yüzme, 1- Voleybol, Basketbol veya Futbol, 2- Tekvando 3- Diğer, 4- Yok

^c 0- Hiç, 1- 1 saat, 2- Daha fazla

^d 0- Servis, 1- Dolmuş, 2- Yaya, 3- Özel Araç (Aile Bireyi Bırakıyor)

^e 0- İlkokul, 1- Ortaokul, 2- Lise, 3- Lisans, 4- Yüksek Lisans/Doktora

^f 0- İlkokul, 1- Ortaokul, 2- Lise, 3- Lisans, 4- Yüksek Lisans/Doktora

^g 0- 5.000,00 TL'den az, 1- 5.000,00 TL- 10.000,00 TL arası 2- 10.000,00 TL'den fazla

^h 0- Refik Arslan Öztürk Fen Lisesi, 1- Ertuğrulgazi Lisesi, 2- Mesleki ve Teknik Anadolu Lisesi, 3-Kız Anadolu İmam Hatip Lisesi.

Veri madenciliğinde kullanılan sınıflama tekniklerinin bir çoğunda nominal değerleri yerine nümerik değerler kullanılmaktadır. Bu çalışmada da sınıflandırma teknikleri kullanılacağından veri setini oluşturan birinci kaynaktan elde edilen veriler ile ikinci kaynak olan ankete verilen cevaplar Microsoft Office Excel programında bir excel dosyası formatında birleştirildikten sonra veri setindeki her bir niteliğin aldığı değerler değer sayısına göre sıfırdan başlayıp artarak sayısallaştırılmıştır. Diğer niteliklerinde alabilecekleri değerler göz önünde bulundurularak yine filtre yöntemi uygulanarak girilen değerlerin uygunluğu kontrol edilmiştir. Aynı zamanda veri setini daha doğru kılabilmek adına Mesleki ve Teknik Anadolu Lisesi ile Kız Anadolu İmam Hatip Lisesi Meslek Lisesi adı altında toplanmıştır.

Mapping adı verilen kelimelerin kategorikleştirilerek sayısallaştırılması işlemine tabi tutulması sonucu oluşan yeni değerler aşağıdaki tablolarda verilmiştir.

Tablo 3.2. Mezuniyet Değişkeninin Sayısallaştırılması.

Mezuniyet	Kategori Numarası
Fen Lisesi	0
Ertuğrulgazi Lisesi	1
Meslek Lisesi	2

Tablo 3.3. Cinsiyet Değişkeninin Sayısallaştırılması.

Cinsiyet	Kategori Numarası
Kız	0
Erkek	1

Tablo 3.4. Ders Çalışma Alışkanlığı Değişkeninin Sayısallaştırılması.

Ders_Ca_Al	Kategori Numarası
Hayır	0
Evet	1

Tablo 3.5. Sosyal Etkinliklere Katılım Değişkeninin Sayısallaştırılması.

Sos_Et	Kategori Numarası
Yüzme	0
Voleybol/Basketbol/Futbol	1
Taekwondo	2
Diğer	3
Yok	4

Tablo 3.6. Takviye Ders Alma Durumu Değişkeninin Sayısallaştırılması.

Tak_De	Kategori Numarası
Hayır	0
Evet	1

Tablo 3.7. Hafta İçi Günlük İnternet Kullanım Süresi Değişkeninin Sayısallaştırılması.

Int_Ku_Su	Kategori Numarası
1 saat	0
1 saatten fazla	1

Tablo 3.8. Hafta İçi Günlük Televizyon İzleme Süresi Değişkeninin Sayısallaştırılması.

Tv_Iz_Su	Kategori Numarası
Hiç	0
1 saat	1
Daha fazla	2

Tablo 3.9. Ders Çalışma Süresi Değişkeninin Sayısallaştırılması.

Ders_Ca_Su	Kategori Numarası
2 saatten az	0
2 saatten fazla	1

Tablo 3.10. Okula Ulaşım Değişkeninin Sayısallaştırılması.

Oku_UI	Kategori Numarası
Servis	0
Dolmuş	1
Yaya	2
Özel Araç	3

Tablo 3.11. Evin Isınma Türü Değişkeninin Sayısallaştırılması.

Isı	Kategori Numarası
Soba	0
Doğalgaz/Katı yakıt	1

Tablo 3.12. Anne Sağ Mı Değişkeninin Sayısallaştırılması.

A_S_M	Kategori Numarası
Hayır	0
Evet	1

Tablo 3.13. Baba Öz Mü Değişkeninin Sayısallaştırılması.

B_O_M	Kategori Numarası
Hayır	0
Evet	1

Tablo 3.14. Anne Öğrenim Durumu Değişkeninin Sayısallaştırılması.

A_O_D	Kategori Numarası
İlkokul	0
Ortaokul	1
Lise	2
Lisans	3
Yüksek Lisans/Doktora	4

Tablo 3.15. Aylık Gelir Değişkeninin Sayısallaştırılması.

Ay_Ge	Kategori Numarası
5.000,00 TL'den az	0
5.000,00 TL-10.000,00 TL arası	1
10.000,00 TL'den fazla	2

3.6. Cinsiyet Ve Mezuniyet Değişkenlerine Ait Yorumlar

3.6.1. Öğrencinin cinsiyeti

Cinsiyet değişkenine ait cevaplar analiz edildiğinde Tablo 3.16'daki veriler elde edilmiştir.

Tablo 3.16. Öğrencilerin Cinsiyet Değişkenine Göre Dağılımı.

Cinsiyet	Frekans	Yüzde
Kız	58	49,57
Erkek	59	50,43
Toplam	117	100

Ankete katılan öğrencilere ait veriler Tablo 3.16'ya göre incelendiğinde 117 kişinin %49,57'si kız öğrencilerden, %50,43'ü ise erkek öğrencilerden oluştuğu görülmektedir.

3.6.2. Öğrencinin mezun olduğu ortaokul

Mezun olunan ortaokul değişkenine ait cevaplar analiz edildiğinde Tablo 3.17'deki veriler elde edilmiştir.

Tablo 3.17. Öğrencilerin Mezun Oldukları Ortaokul Değişkenine Göre Dağılımı.

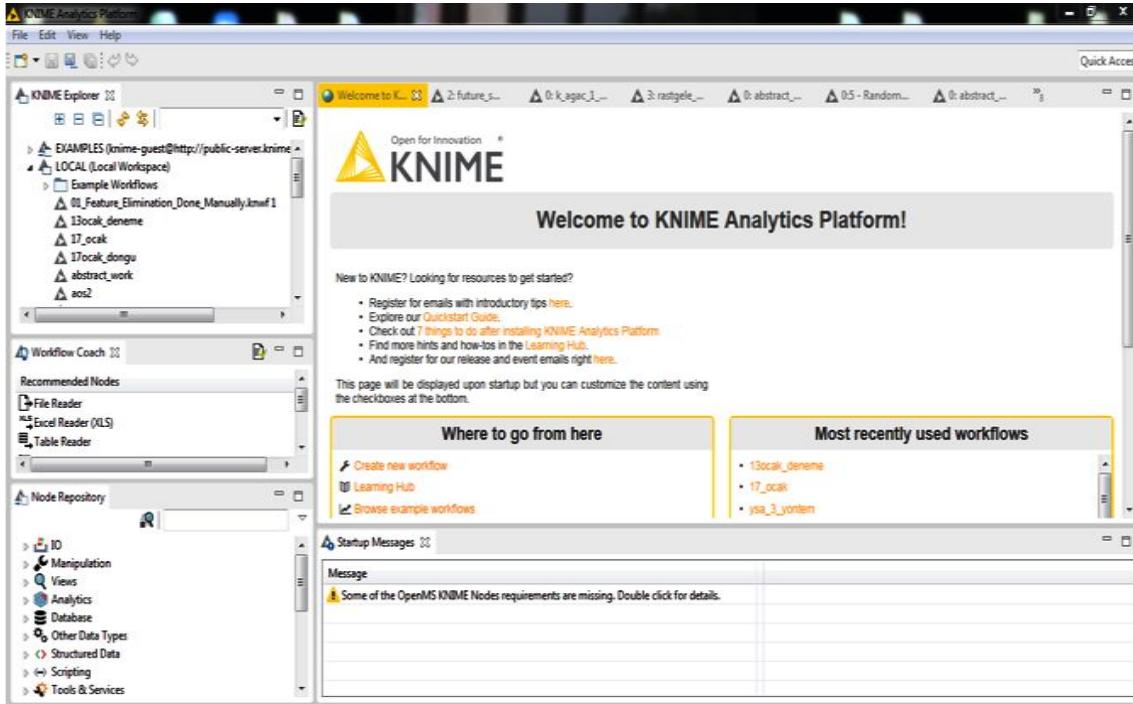
Mezuniyet	Frekans	Yüzde
Edebali Ortaokulu	73	62,40
Murat Hüdavendigâr Ortaokulu	20	17,09
Dursun Fakih İmam Hatip Ortaokulu	24	20,51
Toplam	117	100

Ankete katılan öğrencilere ait veriler Tablo 3.17'ye göre incelendiğinde 117 kişinin %62,40'ı Edebali Ortaokulundan, %17,09'u Murat Hüdavendigâr Ortaokulundan ve %20,51'i ise Dursun Fakih İmam Hatip Ortaokulundan mezun olan öğrencilerden oluştuğu görülmektedir.

Yukarıdaki iki tablodan da anlaşıldığı gibi veri seti oluşturulurken örneklem grubunun homojen olmasına dikkat edilmiştir. Bu durum bizlere en iyi performans ile çalışan modeli oluşturmaya çalışırken istatistiksel olarak normal dağılmış temel verilerin kullanıldığını göstermektedir.

3.7. Model Oluşturma Ve Model Seçimi

3.7.1. Uygulamada kullanılan veri madenciliği aracı



Şekil 3.1. Knime Programının Arayüzü.

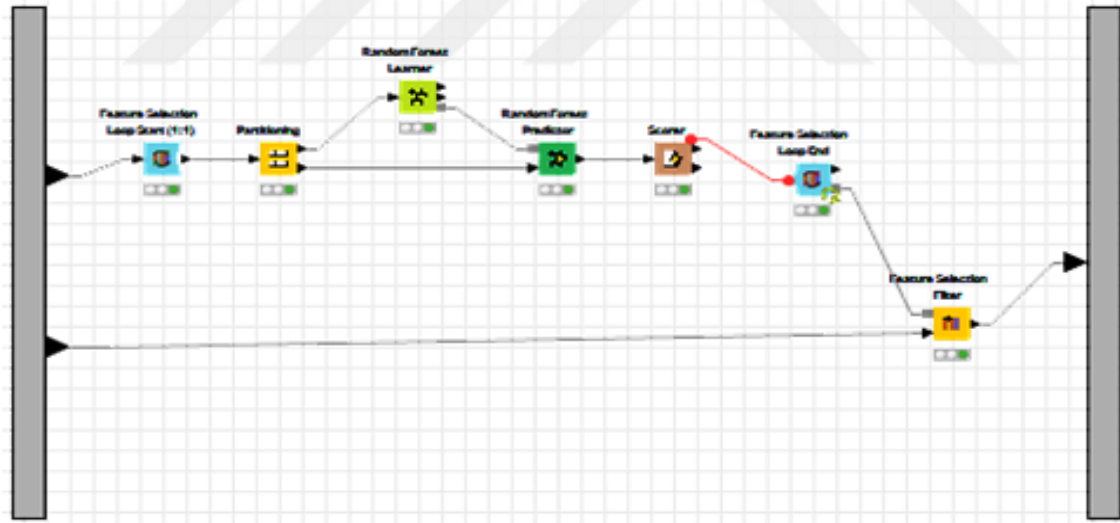
Üzerinde veri madenciliği algoritmalarını çalıştıran birçok yazılım mevcuttur. Bu yazılımlardan bir kısmı açık kaynak kodlu iken bir kısmı da ticari amaçlıdır. Açık kaynak yazılımlarına örnek Orange, RapidMiner, WEKA, R, Keel, Knime, Tanagra, Scriptella ETL, jHepWork ve Elki verilirken, ticari yazılımlara ise SPSS Modeler (Clementine), Excel, SPSS, SAS, Angoss, KXEN, MS SQL Server, MATLAB ve Oracle'ın bu amaçla geliştirdiği modüller örnek verilebilmektedir (Kaya ve Özel 2014:49).

Çalışmamızda model oluşturmak ve makine eğitimi yapmak için KNIME Analytics Platformu kullanılmıştır. Konstanz Information Miner (KNIME), Konstanz Üniversitesinin görsel veri madenciliği geliştirme araştırma grubu tarafından Eclipse Rich Client Platformu üzerinde geliştirilen bir yazılımdır (KNIME, 2020). Knime tamamen ücretsiz olan açık kaynak kodlu bir veri madenciliği yazılımıdır. (Chauhan ve Sehgal 2018:551).

Knime, kendi kategorisinde en zengin görsel araçlar sunan yazılımlarından biridir. Programda Node Repository içerisinde bulunan düğümler arasında ilişkilendirmeler yapılarak verinin işlenmesi, yorumlanması, görselleştirilmesi ve raporlanması gerçekleştirilir.

3.7.2. Ön işleme süreci

Veri seti öğrencilerin başarılarını etkileyen birçok nitelik göz önünde bulundurularak hazırlanmıştır. Ancak tahmin performansını iyileştirmek, daha hızlı sonuç almak ve sürecin daha iyi anlaşılmasını sağlamak için bazı niteliklerin veri setinden çıkarılması söz konusu olabilir. Bu yüzden oluşturulan modeller sonunda, öğrencilerin anne ve babasının sağ ve öz olması gibi bazı nitelikler neredeyse %100'nün evet değeri vermesinden dolayı veri setinden çıkarılmıştır. Bu işlem için denetimli öğrenmede daha yüksek sınıflandırma doğruluğu üreten bir öz nitelik alt kümesi bulmak amacıyla ileri özellik seçimi (forward feature selection) filtreleme yöntemi kullanılmıştır. Böylelikle veri setini oluşturan tüm nitelikleri kullanmak yerine veri setini daha güçlü kılarak daha iyi tahmin yapabilecek 15 nitelikli alt küme oluşturulmuştur. Bu 15 nitelik şunlardır; "Mezuniyet", "Cinsiyet", "Ders Çalışma Alışkanlığı", "Sosyal Etkinliklere Katılım", "Takviye Ders Alıp Almama Durumu", "Hafta İçi Günlük İnternet Kullanım Süresi", "Hafta İçi Günlük Televizyon İzleme Süresi", "Ders Çalışma Süresi", "Okula Ulaşım", "Evin Isınma Türü", "Anne Sağ Mı", "Baba Öz Mü", "Anne Öğrenim Durumu", "Aylık Gelir", "6. ve 7. Sınıf Matematik Ortalaması 5'lik Sistemdeki Notu" dur.

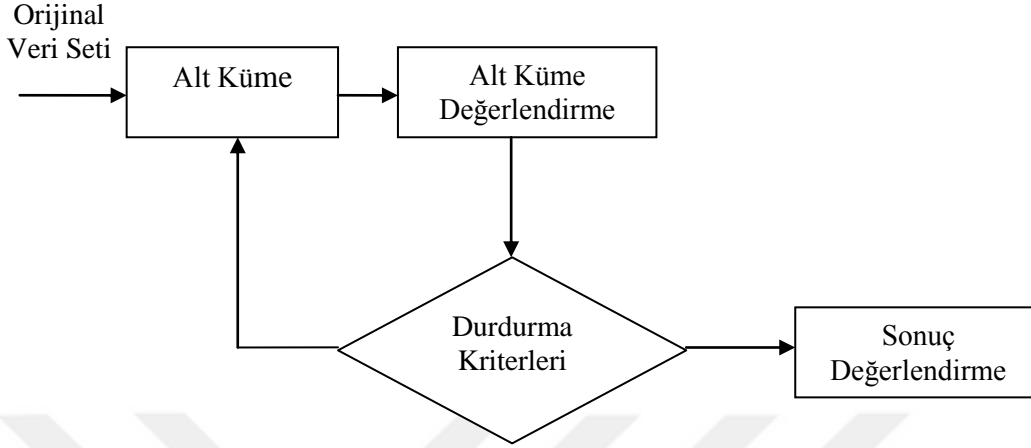


Şekil 3.2. İleri Özellik Seçimi Filtreleme Yöntemi.

İleri özellik seçimi

İleri özellik seçimi diğer adıyla değişken seçimi veya nitelik seçimi, mevcut veri setini temsil edecek en doğru öz alt kümeyi seçmektir. Özellik seçiminin amacı oluşturulacak model için veriye ait en önemli ve faydalı nitelikleri seçmek ve böylece küme eleman sayısını azaltarak algoritmanın hızlı çalışmasını sağlamaktır (Budak, 2018:21). Aynı zamanda özellik

seçimi ile veri kümesi daha anlaşılır hale getirildiği için veri kalitesi de artması mümkündür. Veri seti ile ilgili olmayan gürültülü verilerinde ortadan kaldırılmasıyla model başarımının artması da söz konusu olacaktır. Özellik seçimi Şekil 3.3'teki adımlardan oluşmaktadır.



Şekil 3.3. Özellik Seçimi Genel Akış Şeması.

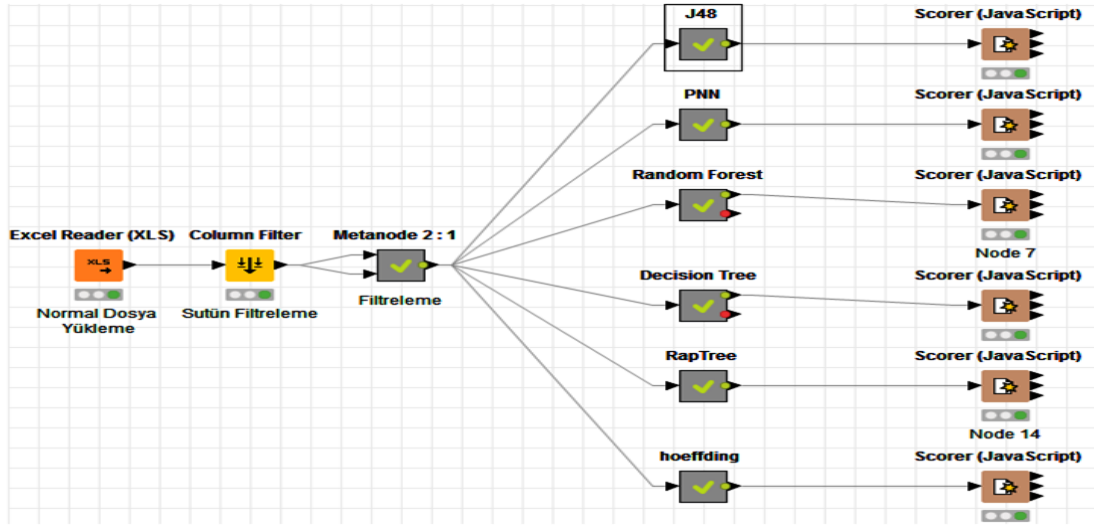
3.7.3. Model oluşturma

Model oluşturulurken mevcut veriler tahmin için kullanılacaksa oluşturulan model *tahmin edici modeldir* ve eğitim aşaması ve test aşaması olmak üzere iki bölümden oluşur. Model oluşturmada kullanılan veri eğitim verisi olarak isimlendirilir. Test aşaması için ayrılan veriler ise iki şekilde olabilir. İlk olarak eğitilmiş modele, veri setinden eğitim aşaması için ayrılan verilerden geri kalanlar sunulur. İkinci olarak ise ileride toplanacak veriler sunulur (Ünsal, 2011:22).

Çalışmada kullanılacak veri seti ön işleme aşamasından sonra 117 örnek ve 15 öznitelikten oluşmaktadır. Model oluştururken modele sunulan veriler eğitim kümesi ve test kümesi olarak ikiye ayrılmıştır. Eğitim seti modeli oluşturmak için kullanılırken, modelin doğruluğu test etmek için ise test seti kullanılmıştır.

Model oluşturulurken öncelikle Knime programında Excel reader (xls) düğümü ile veri seti sisteme yüklenmiştir. Excel dosyasında akademik bilgiler için excel formülleri kullanıldığından modele veri seti yüklendikten sonra ilgili sütunlar column filter düğümü ile filtrelenmiştir. Bölüm 3.7.2'deki filtreleme kısmını oluşturmak için iki girişli bir çıkışlı metanode düğümü içine, yapılan denemeler sonucunda alınan karar ile random forest modeli kullanılmıştır. Bu düğüm içinde bölümlenme yaparken linear sampling metodu seçilmiştir. Linear sampling metodu ile veriler atlamalı olarak eğitim ve test verisi şeklinde ayrılır. Bu metodun seçilmesinde ki amaç veri etiket kümesindeki oranın korunarak eğitim ve test

kümelerinde de uygulanmasıdır. Çalışmada eğitim kümesi için verinin %80'ni (93 örnek), test kümesi için %20'si (24 örnek) kullanılmıştır. Aynı model sayfası üzerinde metanode çıkışı sırası ile J48, PNN, Random Forest, Decision Tree, RapTree ve Hoeffding Tree algoritmalarına bağlanmıştır. Algoritmaları karşılaştırmak için ise her bir algoritmanın çıkışı scorer düğümüne gönderilmiştir. Oluşturulan model düğüm bağlantıları Şekil 3.4'te verilmiştir.



Şekil 3.4. Model Düğüm Bağlantıları.

3.7.4. Model performans ölçümlerinin değerlendirilmesi

Oluşturulan modelin başarısını değerlendirirken kullanılan temel kavramlar doğruluk oranı (CCI ve ICI), kappa istatistiği, kesinlik (Precision), duyarlılık (Recall) ve F-ölçütü (F-Measure) dür. Test işlemi sonucunda modelin başarısı doğru ve yanlış olarak sınıflandırılan değişkenlerin sayısı (yüzdeler oranı) ile hesaplanır. Hata matrisindeki satırlar test kümesindeki örneklere ait gerçek sayıları ifade ederken kolonlar ise modelin tahminini ifade eder. Tablo 3.18'de iki sınıflı veri kümesinde oluşturulmuş bir modelin hata matrisi verilmiştir.

Tablo 3.18. Hata Matrisi.

Gerçek Sınıf		Tahmin Edilen Sınıf	
		Sınıf = 1	Sınıf = 2
	Sınıf = 1	TP	FP
	Sınıf = 0	FN	TN

Tablo 3.18'de True Positive (TP) ve True Negative (TN) değerleri doğru sınıflandırılmış örnek sayılarıdır. FN (False Negative) pozitif yani 1 sınıflıyken negatif

yani 0 olarak tahmin edilmiş örneklerin sayısını ifade ederken, FP (False Positive) aslında negatif (0) sınıfındayken pozitif (1) olarak tahminlenmiş örneklerin sayısıdır. (Şık, 2014:88).

Doğruluk oranı

Model başarısının değerlendirilmesinde kullanılan en basit ve yaygın yöntem, doğruluk oranına bakmaktır. Doğru sınıflandırılmış örnek sayısının (True Positive + True Negative) toplam örnek sayısına (True Positive + True Negative + False Positive + False Negative) bölünmesiyle çıkan sonuçtur. Hata oranı ise yanlış sınıflandırılmış örnek sayısının toplam örnek sayısına bölünmesi ile bulunur. Diğer bir ifade ile hata oranı doğruluk oranını 1'e tamamlayan değerdir. (BİLGİN, 2017).

$$\text{doğruluk} = \frac{TN+TP}{FN+TN+TP+FP} \quad (3.1)$$

$$\text{hata oranı} = \frac{FP+FN}{FN+TN+TP+FP} \quad (3.2)$$

Kappa istatistiği

Oluşturulan model sonucundaki tahminin doğruluk ölçüsü olarak ifade edilir. Kappa istatistiği sonuçları [0,1] kapalı aralığında değişir. Değer 0,4 ile 0,6 aralığında ise orta seviyede uyum söz konusu iken, 0,6 ile 0,8 arasında ise iyi seviyede bir uyum vardır. 0,8 ile 1 aralığında çıkan sonuçlar için ise çok iyi seviyede bir uyum olduğunu göstermektedir. (Kaynar vd., 2016:236).

Kesinlik (precision)

Doğru sınıflandırılmış pozitif örneklem (TP) sayısının, sınıfı pozitif olarak tahmin edilmiş toplam örneklem sayısına bölünmesi ile bulunur. Sonuç [0,1] aralığında çıkar. (Bilgin ve Şentürk, 2019:9).

$$\text{kesinlik} = \frac{TP}{TP+FP} \quad (3.3)$$

Duyarlılık (recall)

Doğru sınıflandırılmış pozitif örneklem (TP) sayısının, toplam pozitif örneklem toplamına bölünmesi ile bulunur. Sonuç [0,1] aralığında çıkar (Bilgin ve Şentürk, 2019:9).

$$\text{duyarlılık} = \frac{TP}{TP+FN} \quad (3.4)$$

F-ölçütü (f- measure)

Doğru sınıflandırılmış pozitif örneklem (TP) sayısının, toplam pozitif örneklem toplamına bölünmesi ile bulunur. Sonuç [0,1] aralığında çıkar (EYÜPOĞLU, 2020:267).

$$f - \text{ölçütü} = \frac{2 \times \text{kesinlik} \times \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}} \quad (3.5)$$

3.7.4.1. J48 modelinin başarı ölçütü

Knime programında model oluşturabilmek için birçok algoritma mevcuttur. Bu çalışmada J48, PNN, Random Forest, Decision Tree, RapTree ve Hoeffding Tree çok sınıflı makine öğrenmesi algoritmaları ile oluşturulan modellerin başarıları; doğruluk oranı CCI (doğru sınıflandırılmış örnekler), ICI (yanlış sınıflandırılmış örnekler), kesinlik (precision), duyarlılık (recall), f-ölçütü (f-measure) ve kappa istatistiği (cohen's kappa) kavramları ile karşılaştırılmıştır. Model sonucunda J48 algoritmasından alınan sınıf istatistiği Şekil 3.5'te, karışıklık matrisi Şekil 3.6'da ve genel istatistik tablosu Şekil 3.7'de verilmiştir.

Okul Adı	Doğru Tahmin	Yanlış Tahmin	Duyarlılık	Kesinlik	F-Ölçütü
Ertuğrulgazi Lisesi	6	5	0.545	0.75	0.632
Fen Lisesi	3	2	0.6	0.5	0.545
Meslek Lisesi	8	0	1	0.8	0.889

Şekil 3.5. J48 Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.

Okul Adı	Ertuğrulgazi Lisesi	Fen Lisesi	Meslek Lisesi
Ertuğrulgazi Lisesi	6	3	2
Fen Lisesi	2	3	0
Meslek Lisesi	0	0	8

Şekil 3.6. J48 Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Doğru Sınıflandırma	Yanlış Sınıflandırma
0.708	0.292	0.556	17	7

Şekil 3.7. J48 Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.

J48 çok sınıflı makine öğrenmesi algoritması ile oluşturulan modele ait veriler %70,8 doğruluk oranı ile sınıflandırılmıştır. 117 örnek %80 bölümlenme sonucunda test kısmına ayrılan 24 örnekten 17'si doğru 7'si yanlış sınıflandırılmıştır.

3.7.4.2. Yapay sinir ağları (PNN) modelinin başarı ölçütü

Yapay sinir ağlarında sınıflandırma algoritmaları için Knime'da düğümler içinde yer alan MultiLayerPerceptron algoritması kullanılmıştır. Bu algortmadan alınan sınıf istatistiği Şekil 3.8'de, karışıklık matrisi Şekil 3.9'da ve genel istatistik tablosu Şekil 3.10'da verilmiştir.

Okul Adı	Doğru Tahmin	Yanlış Tahmin	Duyarlılık	Kesinlik	F-Ölçütü
Ertuğrulgazi Lisesi	7	4	0.636	0.778	0.7
Fen Lisesi	4	1	0.8	0.8	0.8
Meslek Lisesi	7	1	0.875	0.7	0.778

Şekil 3.8. PNN Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.

Okul Adı	Ertuğrulgazi Lisesi	Fen Lisesi	Meslek Lisesi
Ertuğrulgazi Lisesi	7	1	3
Fen Lisesi	1	4	0
Meslek Lisesi	1	0	7

Şekil 3.9. PNN Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Doğru Sınıflandırma	Yanlış Sınıflandırma
0.75	0.25	0.613	18	6

Şekil 3.10. PNN Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.

PNN çok sınıflı makine öğrenmesi algoritması ile oluşturulan modele ait veriler %75 doğruluk oranı ile sınıflandırılmıştır. 117 örnek %80 bölümlenme sonucunda test kısmına ayrılan 24 örnekten 18'si doğru 6'si yanlış sınıflandırılmıştır.

3.7.4.3. Random forest modelinin başarı ölçütü

Random Forest sınıflandırma algoritmaları için Knime'da düğümler içinde yer alan Random Forest Learner ve Random Forest Predictor bölümlenmesi kullanılmıştır. Random Forest algoritmadan alınan sınıf istatistiği Şekil 3.11'de, karışıklık matrisi Şekil 3.12'de ve genel istatistik tablosu Şekil 3.13'te verilmiştir.

Okul Adı	Doğru Tahmin	Yanlış Tahmin	Duyarlılık	Kesinlik	F-Ölçütü
Ertuğrulgazi Lisesi	7	4	0.636	1	0.778
Fen Lisesi	5	0	1	0.714	0.833
Meslek Lisesi	8	0	1	0.8	0.889

Şekil 3.11. Random Forest Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.

Okul Adı	Ertuğrulgazi Lisesi	Fen Lisesi	Meslek Lisesi
Ertuğrulgazi Lisesi	7	2	2
Fen Lisesi	0	5	0
Meslek Lisesi	0	0	8

Şekil 3.12. Random Forest Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Doğru Sınıflandırma	Yanlış Sınıflandırma
0.833	0.167	0.75	20	4

Şekil 3.13. Random Forest Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.

Random forest çok sınıflı makine öğrenmesi algoritması ile oluşturulan modele ait veriler %83,3 doğruluk oranı ile sınıflandırılmıştır. 117 örnek %80 bölümlenme sonucunda test kısmına ayrılan 24 örnekten 20'si doğru 4'si yanlış sınıflandırılmıştır.

3.7.4.4. Decision tree modelinin başarı ölçütü

Knime programında model oluştururken hali hazırda kullanılabilecek birçok karar ağacı mevcuttur. Çalışmada da karar ağaçları algoritmalarından olan birçok algoritma ile çalışılmıştır. Decision Tree algoritmadan alınan sınıf istatistiği Şekil 3.14'te, karışıklık matrisi Şekil 3.15'te ve genel istatistik tablosu Şekil 3.16'da verilmiştir.

Okul Adı	Doğru Tahmin	Yanlış Tahmin	Duyarlılık	Kesinlik	F-Ölçütü
Ertuğrulgazi Lisesi	6	5	0.545	0.75	0.632
Fen Lisesi	3	2	0.6	0.6	0.6
Meslek Lisesi	8	0	1	0.727	0.842

Şekil 3.14. Decision Tree Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.

Okul Adı	Ertuğrulgazi Lisesi	Fen Lisesi	Meslek Lisesi
Ertuğrulgazi Lisesi	6	2	3
Fen Lisesi	2	3	0
Meslek Lisesi	0	0	8

Şekil 3.15. Decision Tree Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Doğru Sınıflandırma	Yanlış Sınıflandırma
0.708	0.292	0.552	17	7

Şekil 3.16. Decision Tree Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.

Decision Tree çok sınıflı makine öğrenmesi algoritması ile oluşturulan modele ait veriler %70,8 doğruluk oranı ile sınıflandırılmıştır. 117 örnek %80 bölümlenme sonucunda test kısmına ayrılan 24 örnekten 17'si doğru 7'si yanlış sınıflandırılmıştır.

3.7.4.5. RAPTree modelinin başarı ölçütü

Rap tree sınıflandırma algoritmaları için Knime'da düğümler içinde yer alan REPTree düğümü kullanılmıştır. Bu düğümün test aşaması için REPTree predictor düğümü yerine weka predictor düğümü kullanılmıştır. Bu algoritmadan alınan sınıf istatistiği Şekil 3.17'de, karışıklık matrisi Şekil 3.18'de ve genel istatistik tablosu Şekil 3.19'da verilmiştir.

Okul Adı	Doğru Tahmin	Yanlış Tahmin	Duyarlılık	Kesinlik	F-Ölçütü
Ertuğrulgazi Lisesi	8	3	0.727	0.667	0.696
Fen Lisesi	3	2	0.6	0.75	0.667
Meslek Lisesi	6	2	0.75	0.75	0.75

Şekil 3.17. RAPTree Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.

Okul Adı	Ertuğrulgazi Lisesi	Fen Lisesi	Meslek Lisesi
Ertuğrulgazi Lisesi	8	1	2
Fen Lisesi	2	3	0
Meslek Lisesi	2	0	6

Şekil 3.18. RAPTree Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Doğru Sınıflandırma	Yanlış Sınıflandırma
0.708	0.292	0.533	17	7

Şekil 3.19. RAPTree Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.

Rap Tree çok sınıflı makine öğrenmesi algoritması ile oluşturulan modele ait veriler %70,8 doğruluk oranı ile sınıflandırılmıştır. 117 örnek %80 bölümlenme sonucunda test kısmına ayrılan 24 örnekten 17'si doğru 7'si yanlış sınıflandırılmıştır.

3.7.4.6. Hoeffding tree modelinin başarı ölçütü

Hoeffding Tree sınıflandırma algoritmaları için Knime'da düğümler içinde yer alan Hoeffding Tree düğümü kullanılmıştır. Bu düğümün test aşaması için Hoeffding Tree predictor düğümü yerine weka predictor düğümü kullanılmıştır. Bu algoritmadan alınan sınıf istatistiği Şekil 3.20'de, karışıklık matrisi Şekil 3.21'de ve genel istatistik tablosu Şekil 3.22'de verilmiştir.

Okul Adı	Doğru Tahmin	Yanlış Tahmin	Duyarlılık	Kesinlik	F-Ölçütü
Ertuğrulgazi Lisesi	11	0	1	0.458	0.629
Fen Lisesi	0	5	0	?	?
Meslek Lisesi	0	8	0	?	?

Şekil 3.20. Hoeffding Tree Makine Öğrenmesi Yöntemine Ait Sınıf İstatistiği.

Okul Adı	Ertuğrulgazi Lisesi	Fen Lisesi	Meslek Lisesi
Ertuğrulgazi Lisesi	11	0	0
Fen Lisesi	5	0	0
Meslek Lisesi	8	0	0

Şekil 3.21. Hoeffding Tree Makine Öğrenmesi Yöntemine Ait Karışıklık Matrisi.

Doğruluk Oranı	Hata Oranı	Kappa İstatistiği	Doğru Sınıflandırma	Yanlış Sınıflandırma
0.458	0.542	0	11	13

Şekil 3.22. Hoeffding Tree Makine Öğrenmesi Yöntemine Ait Genel İstatistik Tablosu.

Hoeffding Tree çok sınıflı makine öğrenmesi algoritması ile oluşturulan modele ait veriler %45,8 doğruluk oranı ile sınıflandırılmıştır. 117 örnek %80 bölümlene sonucunda test kısmına ayrılan 24 örnekten 11'ini doğru 13'ünü yanlış sınıflandırarak en kötü performansı veren algoritma olmuştur.

3.7.5. Modellerin karşılaştırılması

Veri kümesini oluşturan öğrencilerin hangi lise türünü kazanacağını tahmin ederken sonucu etkileyebileceğini düşündüğümüz ve en fazla bilgi verdiği görülen bölüm 3.7.2'de sayılan 15 öznelik ileri özellik seçimi filtreleme yöntemi ile seçildi. Bu özellikler ile veri setinin eğitimi için literatürde en çok kullanılan J48, PNN, Random Forest, Decision Tree, RAPTree ve Hoeffding Tree çok sınıflı makine öğrenmesi yöntemleri ile modeller oluşturuldu. Her bir modele ait eğitilmiş veri istatistikleri bir önceki bölümde verilmiştir. Her bir algoritma ile oluşturulan modellerin başarıları; doğruluk oranı CCI (doğru sınıflandırılmış örnekler), ICI (yanlış sınıflandırılmış örnekler), kesinlik (precision), duyarlılık (recall), f-ölçütü (f-measure) ve kappa istatistiği (cohen's kappa) kavramları ile karşılaştırıldı. Karşılaştırma sonucundaki tüm bu değerler tablo 3.19'da verilmiştir.

Tablo 3.19. Filtrelenmiş Özelliklerle Her Bir Algoritmanın Karşılaştırılması.

Algoritma	CCI	ICI	Duyarlılık	Kesinlik	F-Ölçütü	Kappa İstatistiği
J48	%70,8	%29,2	0,72	0,68	0,69	0,56
PNN	%75	%25	0,77	0,76	0,76	0,61
Random Forest	%83,3	%16,7	0,88	0,84	0,83	0,75
Decision Tree	%70,8	%29,2	0,72	0,69	0,69	0,55
RapTree	%70,8	%29,2	0,69	0,72	0,70	0,53
HoeffdingTree	%45,8	%54,2	0,33	0,15	0,21	0

Tablo 3.19 incelendiğinde söz konusu algoritmalarından Random Forest algoritmasının doğru sınıflandırılmış örneklem yüzdesi %83,3 olarak en yüksek iken yine Random Forest algoritmasının yanlış sınıflandırılmış örneklem yüzdesi ise %16,7 olarak en düşüktür. Kappa istatistiği 0,8 ile 1 arasında çıkan sonuçlar için yüksek seviye uyum olduğu bilindiğine göre tabloda bu değere en yakın sonuç yine Random Forest algoritması ile elde edilmiştir.

Dolayısıyla duyarlılık, kesinlik ve f-ölçütü değerlerinin de diğer tüm algoritmalara göre daha yüksek olduğu görülmektedir.

3.7.6. Bağımlı değişkenlerin karşılaştırılması

Çalışma da kullanılan algoritmalar ile oluşturulan tüm modeller de bağımlı değişken olarak Lise Tür kullanılmıştır. Ancak bağımlı değişken olarak birden fazla da değişken seçilebilir. Literatürde örneklerine sıkça rastlanmaktadır.

Örneğin GÖK "Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi" isimli makalesinde Milli Eğitim Bakanlığı puan ölçeğine göre bağımlı değişken olan öğrenci notlarını Naive Bayes, k-NN, doğrusal DVM, radyal tabanlı fonksiyon (RTF) DVM, rastgele orman ve lojistik yöntemleri ile tahmin etmiştir. Yazar lojistik sınıflandırma algoritmasının %62,80 doğruluk sonucu ile en iyi başarıyı gösterdiğini ortaya koymuştur. (Gök, 2017:140-144).

GÜNER ve ÇOMAK "Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi" isimli çalışmalarında matematik I dersinde başarılı olabilecek öğrencileri matematik I bağımlı değişkenine göre %86,36 doğruluk oranı ile tahmin etmişlerdir. (Güner ve Çomak, 2011:89).

ÇIRAK ve ÇOKLUK "Yükseköğretimde Öğrenci Başarılarının Sınıflandırılmasında Yapay Sinir Ağları ve Lojistik Regresyon Yöntemlerinin Kullanılması" isimli çalışmalarında YSA analizinde kullanılan çıktı değişkeni ve LR analizi için gerekli olan bağımlı değişken olarak öğrencilerin lisedeki mezuniyet notlarını kabul etmişlerdir. LR modeli kullanılarak yapılan sınıflandırma da başarı oranı %66.10 iken YSA modeli ile yapılan eğitim setindeki sınıflandırma oranı %71.40 iken test setindeki bu oran %79.40 olarak bulmuşlardır (Çırak ve Çokluk, 2013:74-76).

BOZKIR, SEZER ve GÖK "Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti" isimli çalışmalarında orta öğretim başarı puanı bağımlı değişkeni ile bu puana etki eden etmenleri, karar ağaçları kullanılarak fen laboratuvarını kullanma oranı, matematik dersi için harcanan zaman (ödev yapma zamanı), okul türü, yaş, sosyal derlere alaka en etkin beş faktör olduğunu görmüşlerdir (Bozkır vd., 2009:8). İncelenen bazı çalışmalarda bağımlı değişken ve doğruluk oranları Tablo 3.20'de verilmiştir.

4. SONUÇLAR

Ülkemizde eğitim süreçlerinde belirli kademelerde öğrenci seçme ve yerleştirme sınavları yapılmaktadır. Yükseköğretime devam etmek isteyen öğrenciler Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından yapılan Öğrenci Seçme Sınavlarına (ÖSS), liselere yerleşmek için ise öğrenciler Milli Eğitim Bakanlığı'nın yapmış olduğu Liselere Geçiş Sınavlarına (LGS) katılmaktadırlar. Eğitim sistemimizde zorunlu eğitimin 12 yıla çıkarılması ile birlikte liseler öğrencilerin eğitim hayatında büyük öneme sahip olmuşlardır. Öğrencilerin meslek edinmeden önce akademik hayatlarını belirleyecek yolda son temel eğitim bölümü olan liseler LGS sınavları başarımına göre öğrenci almaktadırlar.

Bu noktada öğrencilerin ortaokul seviyesindeki almış oldukları eğitim ve hazır bulunuşlukları büyük önem arz etmektedir. Bu kapsamda birçok ders, kurs, etüt ve rehberlik çalışmaları planlanmaktadır. Öğrencilerin sınavlardaki başarılarının tahmini onlara deneme sınavı uygulayarak test edilmektedir. Yapılan deneme sınavlarından elde edilen sonuçlara göre öğrencinin eksik müfredat ve bilgilerinin tamamlanmasına yönelik çalışmalar yapılabilmektedir. Tabi ki bu süreçte sadece akademik bilgi öğrenci başarısı ve başarısızlığında tek başına etken değildir. Başarı soyut bir kavram olduğundan başarıyı olumlu veya olumsuz etkileyen birçok etmen vardır. Öğrencilerin içerisinde bulunduğu sosyal ve aile ortamı, ders çalışma süreleri, bilgisayar veya tablet başında harcadıkları zaman, uyku süreleri ve benzeri etkenler belirleyici rol oynamaktadır.

Günümüzde bilgi ve teknoloji hayatımızı birçok yönden desteklemekte ve fayda sağlamaktadır. İçinde bulunduğumuz bilgi ve teknoloji çağının bizlere birçok alanda olduğu gibi bu alanda da destekleyici ve öngörü sağlayacak nitelikte olduğunu düşünerek öğrencilerin başarılarını sadece test sınavlarına verdikleri cevaplardan çıkarım yapmak yerine diğer etkenleri de hesaba katılarak tahmin edilmesinin faydalı olabileceği değerlendirilerek bu çalışma yapılmıştır. Gerçekleştirilen çalışma ile öğrencilerin eğitim süreçlerindeki ders notlarının yanı sıra öğrencinin okul dışındaki sosyal, ekonomik ve yaşam şeklindeki bileşenler birleştirilerek Lise Geçiş Sınavındaki Öğrenci yerleşim sonucunun tahmin edilmesi amaçlanmıştır. Bu amaç doğrultusunda veri madenciliği yöntemlerinin, veri setine uygunluğu göz önünde bulundurularak sınıflama algoritmaları ile Bilecik ilinde ilköğretimden liseye geçiş sınavlarında öğrenci başarısını tahmin etmede en iyi performansı verecek sınıflandırma algoritması seçilmiştir. Çalışmada oluşturulan modelin doğruluk payı için ise ileri de yapılacak çalışmaya öncü olması amacıyla uygulanan anket sonucunda en çok öğrencinin cevap verdiği Edebalı Ortaokulu'nda, Murat Hüdavendigâr Ortaokulu'nda ve Dursun Fakih

İmam Hatip Ortaokulu'nda 2019-2020 eğitim-öğretim yılında 8. sınıf kademesinde bulunan öğrencilere de aynı anket uygulanmıştır. Çalışmanın gerçekleştirildiği Bilecik ili merkez ilçe nüfusu kapasite ölçeğinde çalışmaya toplam 300 öğrenci alınabilmektedir. Bu öğrencilerden bazılarının farklı illerden nakil olma sebebiyle ortaokul notları temin edilememiştir. Aynı zamanda yapılan veri ön işleme adımlarında sadece bütün anket sorularına net cevap veren öğrencilerin değerlendirilmesiyle 117 öğrenci verisi ile çalışma gerçekleştirilmiştir. Ankete verilen cevaplar ile öğrenci akademik bilgileri excel dosyası formatında birleştirilerek Knime programında öğrenci başarıları tahmin edilmeye çalışılmıştır. En iyi tahmin sonucunu veren modeli bulmak için J48, PNN, Random Forest, Decision Tree, RapTree ve Hoeffding Tree çok sınıflı makine öğrenmesi algoritmaları kullanılmıştır. Bu algoritmalar ile oluşturulan modellerin başarıları; doğruluk oranı CCI (doğru sınıflandırılmış örnekler), ICI (yanlış sınıflandırılmış örnekler), kesinlik (precision), duyarlılık (recall), f-ölçütü (f-measure) ve kappa istatistiği (cohen's kappa) kavramları ile karşılaştırılmıştır. Lise Türü bağımlı değişkenine göre algoritmalar karşılaştırıldığında doğru sınıflama yüzdeleri J48, Decision Tree ve RAPTree karar ağacı algoritmaları %70,8 doğruluk oranını verirken, %83,3 doğruluk oranını ile Random Forest algoritması en iyi sonucu veren algoritma olmuştur.

Oluşturulan modeller sonucunda örnek sayısının artırılması ile model doğruluk oranının da artacağı ön görülmektedir. Bunun yanı sıra 2019-2020 eğitim-öğretim yılında 8. sınıf kademesinde bulunan öğrencilerin verileri de en iyi sonucu veren model üzerinde işlenerek sınav puanıyla alım yapan liselerden hangilerini kazanacakları tahmin edilebilir. Bu sebeple 20 Haziran 2020 tarihinde yapılan Liseye Geçiş Sınavı'nın (LGS) sonuçlarının açıklanmasıyla birlikte model başarılarını artırmak yeni çalışmalar da yapılabilir.

KAYNAKÇA

- Akçapınar, G., Altun, A., & Aşkar, P.** (2016). Çevrimiçi Öğrenme Ortamındaki Benzer Öğrenci Gruplarının Kümeleme Yöntemi İle Belirlenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 6(2), 46-64.
- Albayrak, A. S., & Yılmaz, S. K.** (2009). Veri Madenciliği: Karar Ağacı Algoritmaları Ve İmkb Verileri Üzerine Bir Uygulama. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, 14(1), 31-52.
- Arslan, H.** (2008). *Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği İle Analizi*. (Yayınlanmamış Yüksek Lisans Tezi). Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Sakarya.
- Avşar, A. Ş., & Yalçın, S.** (2015). Öğrencilerin Okuma Başarılarını Açıklayan Ailesel Değişkenlerin Chaid Analizi İle Belirlenmesi. *Eğitim ve Bilim*, 40(179), 1-9.
- Aydemir, B.** (2017). *Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Akademik Başarı Tahmini*. (Yüksek Lisans Tezi). Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli.
- Aydın, A., & Avşar Aydın, E.** (2017). Evaluation Of Limestone Layer's Effect For Uwb Microwave İmaging Of Breast Models Using Neural Network. *Technical Journal*, 11(1-2), 50-54.
- Aydın, E. A.** (2014). Meme Kanseri Tespitinde Mikrodalgaların Önemi Ve Kanserli/Sağlıklı Meme Dokularının Yapay Zeka Algoritmaları İle Tanımlanabilmesi. *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 29(2), 27-38.
- Ayık, Y. Z., Özdemir, A., & Yavuz, U.** (2007). Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 10(2), 441-454.
- Bahadır, İ.** (2008). *Bayes Teoremi Ve Yapay Sinir Ağları Modelleriyle Borsa Gelecek Değer Tahmini Uygulaması*. (Yüksek Lisans Tezi). TOBB Ekonomi ve Teknoloji Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
- Baykasoğlu, A.** (2005). Veri Madenciliği Ve Çimento Sektöründe Bir Uygulama. *Akademik Bilişim*, 1-14.
- Bilen, M.** (2014). *Yapay Sinir Ağları İçin Web Tabanlı Bir Eğitim Yazılımı Geliştirilmesi*. (Yayınlanmamış Yüksek Lisans Tezi). Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Isparta.

- Bilgin, M., & Şentürk, İ. F.** (2019). Danışmanlı Ve Yarı Danışmanlı Öğrenme Kullanarak Doküman Vektörleri Tabanlı Tweetlerin Duygu Analizi. *Balıkesir Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 21(2), 822-839.
- BİLGİN, M.**, (2017). Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi, *Breast*, 2(9), 683.
- Bozkır, A. S.** (2009). *OLAP Ve Veri Madenciliği Teknolojilerinden Yararlanılarak Web Tabanlı Bir Karar Destek Sisteminin Gerçekleştirilmesi*. (Yayınlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
- Bozkır, A. S., Sezer, E., & Gök, B.** (2009). Öğrenci Seçme Sınavında (Öss) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti. *5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09)*, 13-15 Mayıs 2009, Karabük, 1-9.
- Budak, E. Ç.** (2013). Üniversite Öğrencileri İçin Bilgisayar Okuryazarlığını Etkileyen Faktörlerin Etkisinin Veri Madenciliği İle Analizi. *AJIT-e: Online Academic Journal of Information Technology*, 4(11), 58-70.
- BUDAK, H.** (2018). Özellik Seçim Yöntemleri Ve Yeni Bir Yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*. (22) 21-31.
- Chauhan, C. ve Sehgal, S.** (2018). Knime Kullanarak Mobil İncelemeler İçin Duygu Sınıflandırması. In *Computing, Güç ve İletişim Teknolojileri (GUCON) hakkında 2018 Uluslararası Konferansı* (548-553). IEEE.
- Çırak, G., & Çokluk, Ö.** (2013). Yükseköğretimde Öğrenci Başarılarının Sınıflandırılmasında Yapay Sinir Ağları Ve Lojistik Regresyon Yöntemlerinin Kullanılması. *Mediterranean Journal of Humanities*, 3(2), 71-79.
- Dalyan, T.** (2006). *Makine Öğrenmesinde 1R Algoritması Ve İkinci Kuralın (2R) Oluşturulması*. (Yüksek Lisans Tezi). Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli.
- Durairaj, M., & Ranjani, V.** (2013). Data Mining Applications İn Healthcare Sector: A Study. *International journal of scientific & technology research*, 2(10), 29-35.
- Emel, G. G., & Taşkın, Ç.** (2002). Genetik Algoritmalar Ve Uygulama Alanları. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 21(1), 129-152.
- EYÜPOĞLU, C.** (2020). Korelasyon Temelli Özellik Seçimi, Genetik Arama Ve Rastgele Ormanlar Tekniklerine Dayanan Yeni Bir Rahim Ağzı Kanseri Teşhis Yöntemi. *Avrupa Bilim ve Teknoloji Dergisi*, (19), 263-271.
- Gök, M.** (2017). Makine Öğrenmesi Yöntemleri İle Akademik Başarımın Tahmin Edilmesi. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji*, 5(3), 139-148.

- Gülçe, G.** (2010). *Veri Ambarı Ve Veri Madenciliği Teknikleri Kullanılarak Öğrenci Karar Destek Sistemi Oluşturma*. (Yüksek Lisans Tezi). Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli.
- Güner, N., & Çomak, E.** (2011). Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 17(2), 87-96.
- Gürel, A. G.** (2019). *Üniversite Kütüphanesi Verileri Üzerinde Veri Madenciliği Yöntemlerinin Uygulanması*. (Yüksek Lisans Tezi). Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, Afyon.
- Hamoud, A.** (2016). Selection Of Best Decision Tree Algorithm For Prediction And Classification Of Students' Action. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, 16(1), 26-32.
- Irmak, S., Köksal, C. D., & Asilkan, Ö.** (2012). Hastanelerin Gelecekteki Hasta Yoğunluklarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi. *Journal of Alanya Faculty of Business/Alanya İşletme Fakültesi Dergisi*, 4(1), 101-114.
- Kaçmaz, A.** (2019). *Mesleki Ve Teknik Eğitim Öğrencilerinin Teknoloji Bağımlılıklarının Veri Madenciliği İle Analizi*. (Yüksek Lisans Tezi). Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.
- Kavzoğlu, T., & Çölkesen, İ.** (2010). Destek Vektör Makineleri İle Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. *Harita Dergisi*, 144(7), 73-82.
- Kaya Keleş, M.** (2017). An Overview: The Impact Of Data Mining Applications On Various Sectors. *Technical journal*, 11(3), 128-132.
- Kaya Keleş, M., & Keleş, A. E.** (2017). Veri Madenciliği Uygulamalarının Ve Sezgisel Optimizasyon Algoritmalarının Yapım Yönetimindeki Yeri. *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, 32(1), 235-242.
- Kaya, M., Özel, S. A.** (2014). Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması (Comparison of Open Source Data Mining Software). *Akademik Bilişim Konferansı (AB2014)*, AB2014 Proceedings, 47-53.
- Kaynar, O., Yıldız, M., Görmez, Y., & Albayrak, A.** (2016). Makine Öğrenmesi Yöntemleri İle Duygu Analizi. In *International Artificial Intelligence and Data Processing Symposium (IDAP'16)*, 234-241

Keleş, A. E., & Kaya, M. (2014). Duvar İnşa Edilmesinde Verimliliği Etkileyen Faktörlerin Apriori Veri Madenciliği Yöntemi Kullanılarak Analizi. *Akademik Bilişim Konferansı, AB2014 Bildiriler Kitapçığı*, 831-836.

Kılınc, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M., & Özçift, A. (2016). KNN Algoritması Ve R Dili İle Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi. *Marmara Fen Bilimleri Dergisi*, 28(3), 89-94.

Koyuncugil, A. S., & Özgülbaş, N. (2009). Veri Madenciliği: Tıp Ve Sağlık Hizmetlerinde Kullanımı Ve Uygulamaları. *International Journal Of Informatics Technologies*, 2(2), 21-32.

Köse, İ. (2018). *Veri Madenciliği Teori Uygulama ve Felsefesi*. Papatya, İstanbul.

Kurt, Ç., & Erdem, O. A. (2012). Öğrenci Başarısını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi. *Politeknik Dergisi*, 15(2), 111-116.

Milli Eğitim Bakanlığı [Erişim Tarihi: 28/042020, <https://www.mevzuat.gov.tr/MevzuatMetin/1.5.1739.pdf>

Milli Eğitim Bakanlığı, PISA 2018 Türkiye Ön Raporu, (2019, Aralık). [Erişim Tarihi: 01.04.2020,

http://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf.]

Oğuzlar, A. (2003). Veri Ön İşleme. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, (21), 67-76.

Open for Innovation KNIME (2020). [Erişim tarihi:13/07/2020, <https://www.knime.com/software-overview>]

Özbay, Ö. (2015). Veri Madenciliği Kavramı Ve Eğitimde Veri Madenciliği Uygulamaları. *Uluslararası Eğitim Bilimleri Dergisi*, (5), 262-272.

Özçınar, H. (2006). *KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi*. (Yüksek Lisans Tezi). Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Denizli.

Özdemir, Ş. (2016). *Eğitimde Veri Madenciliği Ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama*. (Doktora Tezi). İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Özekes, S. (2003). Veri Madenciliği Modelleri Ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi*, 65-82.

Özel, C., & Topsakal, A. (2014). Veri Madenciliği Kullanarak Beton Basınç Dayanımının Belirlenmesi. *Cumhuriyet Üniversitesi Fen-Edebiyat Fakültesi Fen Bilimleri Dergisi*, 35(1), 1-11.

Özkan, Y. (2013). *Veri Madenciliği Yöntemleri*. Papatya, İstanbul, 211s.

Özkan, Y., 2008, *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık, İstanbul, 216s.

- Pala, T.** (2013). *Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi*. (Yüksek Lisans Tezi). Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.
- Sakarya, M.** (2019). *Seçilen Siğil Tedavi Yönteminin Farklı Makine Öğrenmesi Algoritmaları İle Başarımının Tahmin Edilmesi*. (Yayınlanmamış Yüksek Lisans Tezi). Zonguldak Bülent Ecevit Üniversitesi, Fen Bilimleri Enstitüsü, Zonguldak.
- Sarıman, G.** (2011). Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means Ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 15(3), 192-202.
- 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. O.** (2010). *İnsan Nefesinden Kandaki Glikoz Ve Hba1c Değerlerinin Elektronik Burun Kullanılarak Belirlenmesi*, (Yüksek Lisans Tezi). Dumlupınar Üniversitesi, Fen Bilimleri Enstitüsü, Kütahya.
- Shahbaz, M., & Rahman, M.** (2008). Data Mining For Engineering Sector İn Pakistan: Issues And İmplications. *In Proceedings of the World Congress on Engineering and Computer Science*, 787-792.
- Silahtaroğlu, G.** 2008. *Kavram Ve Algoritmalarıyla Temel Veri Madenciliği*. Papatya Yayıncılık, İstanbul, 174.
- Şık, M. Ş.** (2014). *Veri Madenciliği Ve Kanser Erken Teşhisinde Kullanımı*. (Yüksek Lisans Tezi). İnönü Üniversitesi, Sosyal Bilimler Enstitüsü, Malatya.
- Taşcı, E., & Onan, A.** (2016). K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi. *Akademik Bilişim*. 1-8.
- Tekin, A., & Öztekin, Z.** (2018). Eğitsel Veri Madenciliği İle İlgili 2006-2016 Yılları Arasında Yapılan Çalışmaların İncelenmesi. *Eğitim Teknolojisi Kuram ve Uygulama*, 8(2), 108-124.
- Terzi, Ö., Küçüksille, E. U., Ergin, G., & İlker, A.** (2011). Veri Madenciliği Süreci Kullanılarak Güneş Işımının Tahmini. *Uluslararası Teknolojik Bilimler Dergisi*, 3(2), 29-37.
- Türk Dil Kurumu**, Eğitim nedir, [Erişim Tarihi: 27.03.2020, <https://sozluk.gov.tr/?kelime=e%C4%9Fitim%20kurumu>].
- Ünsal Ö.** (2011). *Mesleki Alan Seçimlerinin Makine Öğrenmesi Algoritması Kullanılarak Belirlenmesi*, (Yüksek Lisans Tezi). Gazi Üniversitesi, Bilişim Enstitüsü, Bilgisayar Eğitimi, Ankara.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A Data Mining Approach For Identifying Predictors Of Student Retention From Sophomore To Junior Year. *Journal of Data Science*, 8(2), 307-325.

Zaki, M., J. (1996). Sampling- An Efficient Solution For Data Mining Of Association Rules, *Department of Computer Science*, 14627(617), 1-12.



ÖZ GEÇMİŞ

Kişisel Bilgiler

Adı Soyadı : Ayşegül SELVİ
Doğum Yeri ve Tarihi : Havsa 01/01/1985



Eğitim Durumu

Lisans Öğrenimi : Muğla Sıtkı Koçman Üniversitesi
Teknik Eğitim Fakültesi/Elektronik Ve Bilgisayar
Eğitimi Bölümü/Elektronik Ve Bilgisayar Öğretmenliği
Pr. (4+1) (İngilizce Hazırlık) (2003-2008)

Bildiği Yabancı Diller : İngilizce

İş Deneyimi

Stajlar : Haziran 2007-Temmuz 2007
Muğla Üniversitesi Bilgi İşlem Dairesi Başkanlığı
Network Bölümü
Haziran 2006- Temmuz 2006
Devlet Su İşleri Müdürlüğü Bilgi İşlem Edirne
Temmuz 2005-Ağustos 2005
Türk Telekom İl Müdürlüğü Bilgi İşlem Edirne

Çalıştığı Kurumlar : 2016-2017 Bilecik Edebalı Ortaokulu
2015-2016 Bilecik Şeyh Edebalı Üniversitesi Bilgisayar
Programcılığı
2012-..... Bilecik Halk Eğitimi Merkezi Ve ASO
2012-2012 Bilecik Mesleki ve Teknik Anadolu Lisesi

2009-2012 Kütahya Anadolu Teknik, Teknik Ve
Endüstri Meslek Lisesi

2008-2009 Edirne Milli Piyango Anadolu Ticaret Meslek
Ve Ticaret Meslek Lisesi

İletişim

Adres : Beşiktaş Mah. Marmara Cd. 500 Evler Sitesi G/Blok D:11
Merkez /BİLECİK

E-Posta Adresi : aysegulselvi84@gmail.com

Akademik Çalışmaları

Emre DANDIL, Ayşegül SELVİ (2019)

Classification of Eye Blinks in EEG Signals Using Neural
Networks. 2nd International Conference on Data Science
And Applications (ICONDATA 2019) (Özet Bildiri/Sözlü
Sunum)

Ayşegül SELVİ, Salim CEYHAN (2020). Bilecik İlinde
Ortaöğretimden Liseye Geçiş Sınavlarında Makine
Öğrenmesi Yöntemleri İle Öğrenci Başarısının Tahmini.
3rd International Conference on Data Science And
Applications (ICONDATA 2020) (Tam Bildiri/Sözlü
Sunum)