KIRIKKALE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
YÜKSEK LİSANS TEZİ

ÖĞRENCİ PERFORMANSININ VERİ MADENCİLİĞİ İLE BELİRLENMESİ

Sevil ÖZARSLAN

TEMMUZ 2014

Bilgisayar Mühendisliği Anabilim Dalında Sevil ÖZARSLAN tarafından hazırlanan ÖĞRENCİ PERFORMANSININ VERİ MADENCİLİĞİ İLE BELİRLENMESİ adlı Yüksek Lisans Tezinin Anabilim Dalı standartlarına uygun olduğunu onaylarım.

Prof. Dr. Hasan ERBAY
Anabilim Dalı Başkanı

Bu tezi okuduğumu ve tezin **Yüksek Lisans Tezi** olarak bütün gereklilikleri yerine getirdiğini onaylarım.

Doç. Dr. Necaattin BARIŞÇI
Danışman

Jüri Üyeleri

Başkan	: Doç. Dr. Erdem Kamil YILDIRIM	_____
Üye (Danışman)	: Doç. Dr. Necaattin BARIŞÇI	_____
Üye	: Yrd. Doç. Dr. Taner TOPAL	_____

16/07/2014

Bu tez ile Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onaylamıştır.

Doç. Dr. Erdem Kamil YILDIRIM
Fen Bilimleri Enstitüsü Müdürü

ÖZET

ÖĞRENCİ PERFORMANSININ VERİ MADENCİLİĞİ İLE BELİRLENMESİ

ÖZARSLAN, Sevil
Kırıkkale Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi
Danışman: Doç. Dr. Necaattin BARIŞÇI
Temmuz 2014, 74 sayfa

Gelişen teknoloji ile birlikte yüz yüze eğitime alternatif olarak elektronik ortamlarda öğrenme giderek yaygınlaşmaktadır. Eğitim sektöründe çeşitli alanlarda Web'e dayalı öğrenme ortamları oluşturulmaktadır. Yükseköğretim kurumları da teknolojiyi yakından takip eden ve her türlü yeniliğe açık kurumlar olarak göze çarpmaktadır. Eğitim-öğretimde çok yeni olan Web'e dayalı uzaktan eğitim üniversitelerimizin çeşitli bölüm ve programlarında kullanılmaktadır. Tamamen uzaktan eğitim veren bölümler olduğu gibi sadece birkaç dersi uzaktan eğitim yolu ile veren bölümlerde bulunmaktadır.

Bu tez çalışmasında Kırıkkale Üniversitesinde okuyan, birinci sınıf öğrencilerinin ENF-101 kodlu Temel Bilgi Teknolojileri Kullanımı dersi için akademik performansları incelenmiştir. İnceleme dersi geleneksel bir yöntem olan yüz yüze eğitim ile alan öğrenciler ile yeni bir yöntem olan uzaktan eğitim ile alan 672 öğrenciye ait veriler veri madenciliği sınıflandırma algoritmaları ile incelenmiştir. Sonuçlara göre karar ağacı oluşturularak öğrenci başarısına etki eden faktörler belirlenmiştir.

Bu çalışma ile veri madenciliği teknikleri kullanılarak yükseköğretim kurumlarında eğitim yöntemlerinin başarıya olan etkisi hakkında hem üniversite yönetimine hem de öğrencilere faydalı bilgiler verebileceği ortaya konulmuştur.

Anahtar Kelimeler: Veri Madenciliği, Uzaktan Eğitim, Karar Tablosu, JRip, J48, Çok Katmanlı Algılayıcı

ABSTRACT

DETERMINATION OF STUDENTS PERFORMANCE WITH DATA MINING

ÖZARSLAN, Sevil
Kırıkkale University
Graduate School of Natural and Applied Sciences
Depertment of Computer Engineering, M.Sc. Thesis
Supervisor: Assoc. Prof. Dr. Necaattin BARIŞÇI
July 2014, 74 pages

With improving technology as an alternative to face to face education electronic learning environments is increasingly common. Education sector in various areas of the Web 'e-based learning environments are created. Higher education institutions also closely follow and all kinds of technology innovation is observed as public institutions. In education who are very new to the Web 'e-based distance education universities are used in various departments and programs. As part of providing distance education entirely in just a few courses through distance education department, which is located in.

In this thesis, studying in Kırıkkale University, of first class students ENF-101 coded courses Fundamentals of Information Technology Usage for academic performance were examined. Review of the course, which is a traditional method of face to face training and distance education students taking the field with the new method, the data of 672 students were examined by the data mining classification algorithms. According to the results of a decision tree forming factors have been identified that affect student achievement.

In this study, using data mining techniques to success in higher education institutions about the impact of the training methods and provide useful information to the university administration and the student was revealed.

Key Words : Data Mining, Distance Learning, Decision Table, JRip, J48 Algorithm, Multilayer Perceptron (MLP)

TEŞEKKÜR

Tezimin hazırlanması esnasında yardımlarını esirgemeyen tez danışmanım Sayın Doç. Dr. Necaattin BARIŞÇI'ya, tezimin birçok aşamasında yardım gördüğüm Okutman Volkan ATEŞ'e teşekkür ederim.

Doğumumdan bugüne bana desteklerini hiçbir zaman esirgemeyen canım annem ve babama, her zaman yanımda olan sevgili eşim ve çocuklarıma ayrıca teşekkür ederim.

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

ŞEKİL Sayfa

ÇİZELGELER DİZİNİ

ÇİZELGE Sayfa

SİMGELER VE KISALTMALAR DİZİNİ

SİMGELER DİZİNİ

f	Aktivasyon Fonksiyonu
Σ	Toplam Sembolü

KISALTMALAR DİZİNİ

ARFF	Attribute- Relation File Format
CRISP-DM	Cross- Industry Standard Process for Data Mining
ÇKA	Çok Katmanlı Algılayıcı
IREP	Incremental Reduced Error Pruning
NCR	National Cash Register
OLAP	On-Line Analytical Processing
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
SPSS	Statistical Package for the Social Sciences
TBTK	Temel Bilgi Teknolojileri Kullanımı Dersi
YSA	Yapay Sinir Ağları

1. GİRİŞ

Eğitim bir toplumun gelişimi için en önemli unsurlardan biridir. Öğrenmeyi en üst düzeye çıkartmak için yıllarca birçok yöntemler denenmiş ve bu yolda sürekli gelişmeler elde edilmiştir. Klasik eğitim yöntemlerine her defasında yenilikler katılmış ve günün teknolojik gelişimlerinden yararlanılmıştır.

İnternet her alanda olduğu gibi eğitim alanında da hayatımıza hızlı bir şekilde giriş yapmıştır. Günümüz şartlarına uygun olarak Web tabanlı yeni eğitim-öğretim, ölçme ve değerlendirme yöntemleri geliştirilmektedir. Bunlardan biri de uzaktan eğitim yöntemidir. Zaman ve mekân sorununu çözen bu yöntem giderek yaygınlaşmakta ve kullanımı zorunlu hale gelmektedir [1].

Uzaktan eğitim yöntemi bir nevi bilgisayar destekli eğitim şeklidir. Son birkaç yılda çok önem kazanmış ve internet üzerinden web kursları olarak yaygınlaşmaya başlamıştır. Fakat mevcut web tabanlı birçok derste kullanılan öğrenme materyalleri oluşturulurken öğrenci çeşitliliği dikkate alınmalıdır. Adaptif ve akıllı web tabanlı eğitim sitemleri zengin öğrenme ortamları için çözüm olarak görülmüştür [2].

Okulların otomasyon sistemlerinde çeşitli yazılımlarla öğrencilere ait birçok bilgi veri tabanlarında tutulmaktadır. Pek çok, tek başına anlamsız olan bu bilgilerden veri madenciliği teknikleri ile anlamlı sonuçlar alınabilmektedir. Böylece eğitim kurumları için önemli bilgilere ulaşılabilmektedir.

Eğitim alanında, öğrencilerin

- Başarı veya başarısızlık nedenlerinin bulunması,
- Öğrenci başarısının arttırılması için neler yapılabileceği,
- Üniversiteye yerleştirmede esas alınan giriş puanları ile öğrencinin okul başarısı arasında bir ilişkinin var olup olmadığı,
- Üniversiteye yerleştirmede esas alınan giriş puanları ile başarılı olduğu ders türleri ile arasında bir ilişkinin var olup olmadığı gibi soruların cevaplarının

araştırılmasında veri madenciliği yöntemleri kullanılarak, eğitimin kalitesi ve performansı arttırılabilir.

Günümüze kadar eğitim alanında yapılmış olan veri madenciliği çalışmaları aşağıda kısaca özetlenmiştir;

1995 yılında Sanjeev ve Zytkow tarafından yayınlanan çalışmada araştırmacılar bilgi keşfini "R aralığındaki veriler için P örüntüsü" şeklinde ifadeler halinde üniversite veri tabanından elde etmişlerdir. Sonuçlar kurumsal politikalarla ilgili stratejik kararların verilmesi için üniversite yönetimine sunulmuştur [3].

2002 yılında Jing Luan yükseköğretimde öğrencilerin belirleyici özelliklerinin kullanıldığı öğrenci memnuniyetini ölçmeye yönelik bir veri madenciliği uygulaması gerçekleştirmiştir. Bu çalışma sonucunda eğitim kurumlarının kaynak ve personel kullanımını daha verimli hale getirebilmeleri için C5.0 gibi tahmin edici denetimli öğrenme modelleri ve Kohonen ağları gibi kümeleyici denetimsiz öğrenme modellerini kullanmayı önermiştir [4].

2004 yılında Murat Karabatak ve Melih Cevdet İnce tarafından yapılan çalışmada Veri Madenciliği teknikleri kullanılarak Fırat Üniversitesi Teknik Eğitim Fakültesi Bilgisayar Eğitimi bölümü öğrencilerinin notları kullanılarak öğrenci başarılarının analizi yapılmıştır. Bu analizi yapmak için Veri Madenciliğinde, birliktelik kuralı çıkarım algoritmalarından biri olan Apriori algoritması kullanılmıştır [5].

2005 yılında Şenol Zafer Erdoğan ve Mehpare Timor tarafından gerçekleştirilen çalışmada Maltepe üniversitesi öğrencilerinin belirleyici özelliklerini "K-Means" algoritması kullanılarak kümelenmiştir. 2003 yılına ait 722 öğrenci verisini kullanıldığı çalışmada öğrencilerin üniversiteye giriş sınav sonuçları ile başarıları arasındaki ilişki kümeleme analizi ve K-Means algoritması teknikleri kullanılarak incelenmiştir [6].

2007 yılında Y. Ziya Ayık tarafından yapılan çalışmada, Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile

kazandıkları fakülteler arasındaki ilişki, veri madenciliği teknikleri kullanılarak incelenmiştir. Çalışma sonucunda, lise türünün arzu edilen bir fakültenin kazanılmasında çok büyük öneminin olduğu, yine lise başarısının da aynı derecede önemli olduğu tespit edilmiştir. Elde edilen sonuçlara göre, Atatürk Üniversitesi'ni sonraki yıllarda tercih edecek öğrenci profilinin belirlenmesine yardımcı olacağı sonucuna varılmıştır [7].

2010 yılında Yavuz Ünal, Ufuk Ekim ve Murat Köklü tarafından yapılan çalışmada veri madenciliği tekniklerinden K-Means kullanılarak 2009-2010 eğitim öğretim döneminde Selçuk Üniversitesinin 3 fakülte ve bir yüksekokulda okuyan öğrencilerin ortak zorunlu derslerdeki başarılarının analizi yapılmıştır. İnceleme sonucuna göre sayısal bölümlerden oluşan Mühendislik Mimarlık Fakültesi öğrencilerinin Atatürk İlkeleri ve İnkılâp Tarihi, Türk Dili ve Yabancı Dil gibi sözel derslerde diğer fakülte ve yüksekokul öğrencilerine göre daha başarılı oldukları görülmüştür. Üniversiteye giriş puan türüne göre sözel olan Sosyal Bilimler Meslek Yüksek Okulu öğrencilerinin, sayısal ağırlıklı olan Mühendislik Mimarlık Fakültesi ve Fen Fakültesi öğrencilerine göre başarı oranlarının düşük olduğu görülmüştür [8].

2012 yılında Mehmet Ali Alan tarafından yapılan çalışmada veri madenciliği yöntemiyle Cumhuriyet Üniversitesi Sosyal Bilimler Enstitüsü öğrencilerine ait veriler kullanılarak bir uygulama yapılmıştır. Lisansüstü öğrencilerine ait verilerden yararlanarak, hem bu verileri en başarılı sınıflandıran algoritma, hem de öğrencilerin programı, cinsiyeti, Sivas ilinden ya da başka bir ilden olması, kadrosunun araştırma görevlisi olup olmaması ve ders döneminin farklı olmasının notlarını etkileyip etkilemediği tespit edilmeye çalışılmıştır [9].

2012 yılında Baha Şen ve Emine Uçar tarafından yapılan diğer bir çalışmada veri madenciliği teknikleri kullanılarak Karabük Üniversitesi Bilgisayar Mühendisliği Bölümü öğrencilerinin başarılarını yaş, cinsiyet, lise mezuniyet türü, eğitimin uzaktan veya yüz yüze olması, dersin kültür dersi veya meslek dersi olması kriterlerine göre karşılaştırması yapılmıştır. Çalışmanın sonucunda başarı oranının öğrencinin yaşı ile ters orantılı olduğu, artan yaş ile başarının azaldığı, dersi yüz yüze

eğitim ile alan öğrencilerin başarılarının daha yüksek olduğu, öğrencilerin kültürel derslerde mesleki derslere göre daha başarılı olduğu sonuçlarına ulaşılmıştır [10].

Bu çalışmada 2012-2013 Eğitim-Öğretim yılı Kırıkkale Üniversitesinde Temel Bilgi Teknolojileri Kullanımı dersini alan öğrencilere ait veriler Veri Madenciliği yöntemleri ile incelenmiştir. Web tabanlı uzaktan eğitim ile klasik eğitim yöntemlerine göre öğrenci performansları değerlendirilmiştir.

2. MATERYAL VE YÖNTEM

Bu çalışmada materyal olarak Kırıkkale Üniversitesi'nin çeşitli bölümlerinde okuyan 672 adet öğrencinin ENF-101 kodlu Temel Bilgi Teknolojileri Kullanımı (TBTK) dersine ait başarı notları kullanılmıştır.

Öncelikle öğrencinin başarısına etkisi muhtemel faktörler; öğrencinin bölüme yerleştirmede esas alınan puan türü (sayısal, sözel, eşit ağırlık, yabancı dil, özel yetenek, sınavsız geçiş), öğrencinin eğitim gördüğü akademik birim (fakülte-yüksekokul), öğrencinin cinsiyeti (kız, erkek), öğrencinin başarı durumu (çok iyi, ortalama, başarısız), öğrencinin yaş aralıkları, öğrencinin dersi aldığı dönem (güz, bahar), dersin verildiği eğitim sistemi (yüz yüze eğitim, uzaktan eğitim) olarak belirlenmiştir.

Yapılan çalışma sonucunda öğrencinin başarısına etki eden faktörler kıyaslanarak öğrencilerin başarısızlıkları ve başarısızlıklarının nedenini bulup çözümlemek hedeflenmiştir. Uygulama WEKA 3.7. programı ile gerçekleştirilmiştir.

Bu bölümde veri madenciliğinin özellikleri ve önemi üzerinde durulmuştur.

2.1. Veri Madenciliğine Giriş

Geçmiş yıllarda insanlar bilgi ve tecrübelerini aktarmada kâğıt ortamlarını kullanmıştır. Zamanla bu durum hem iş yükünü arttırmış hem de bilgiye ulaşımı zorlaştırmıştır. Bu durum, insanların geleceğe yönelik farklı teknolojiler geliştirmeye yönelmesini sağlamıştır.

Dijital verilerin gün geçtikçe artış göstermesi ile birlikte bilgi miktarlarında büyük artışlar söz konusu olmaktadır. Bilgi teknolojilerinin çok hızlı ve sürekli gelişimi ve buna bağlı olarak daha ucuza teknolojiye sahip olunabildiğinden verilerin artması olağan bir durumdur. Günümüzde bilgi teknolojileri çok büyük miktardaki verilerin

toplanmasına, saklanmasına, işlenmesine ve tekrar bilgiye dönüştürülmesine olanak sağlamaktadır.

Boyutları gün geçtikçe artış gösteren veriler veri tabanlarında depolanmaktadır. Zamanla büyük miktardaki çeşitli veriler içinde sistemlerin ihtiyacı doğrultusunda anlamlı bilgilerin elde edilebilmesi gerekmektedir. Bundan dolayı büyük miktardaki verilerden anlamlı bilgilerin çıkartılması için veri inceleme ve analizi yapan çeşitli teknolojiler geliştirilmesine ihtiyaç duyulmuştur. Dolayısı ile veri tabanlarından bilgi keşfi yapacağımız bir süreç söz konusu olmuştur.

Depolanan bu veriler genelde tek başına bir anlam ifade etmemektedirler. Artık yetkililer veri tabanlarında bulunan verilerden anlamlı sonuçlar elde etmek istemektedirler. Büyük miktarda, tek başına anlamsız veri içerisinden anlamlı, gizli kalmış, kullanılabilir bilgileri çıkarmada Veri Madenciliği teknikleri önemli yer tutmaktadır.

2.2. Veri Madenciliğinin Tanımı

Literatürde Veri Madenciliği (Data Mining) ya da Veri Tabanlarında Bilgi Keşfi (Knowledge Discovery) olarak tanımlanmakta olan bu süreçte hedeflenen sonuçlar istatistik, veri tabanları, yapay öğrenme, modelleme, bilgisayar yazılımları kullanılarak elde edilmektedir.

Veri Madenciliği alanında çalışma yapan araştırmacılar tarafından pek çok tanım yapılmıştır. Bunlardan bazılarına aşağıda yer verilmiştir.

Veri Madenciliği; veri ambarlarındaki tutulan, çok çeşitli ve çok miktarda veriye dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarmak, bunları karar verme ve eylem planını gerçekleştirmek için kullanma sürecidir [11].

Veri madenciliği büyük miktarda veri içinden gelecekle ilgili tahmin yapmamızı sağlayacak bağıntı ve kuralların bilgisayar programları kullanarak aranmasıdır [12].

Veri madenciliği, hem duyarlı hem de anlaşılabilir verilerle, alışılmamış yollarla verileri özetleyen ve gizli ilişkileri ortaya koyan bir analiz yöntemidir [13].

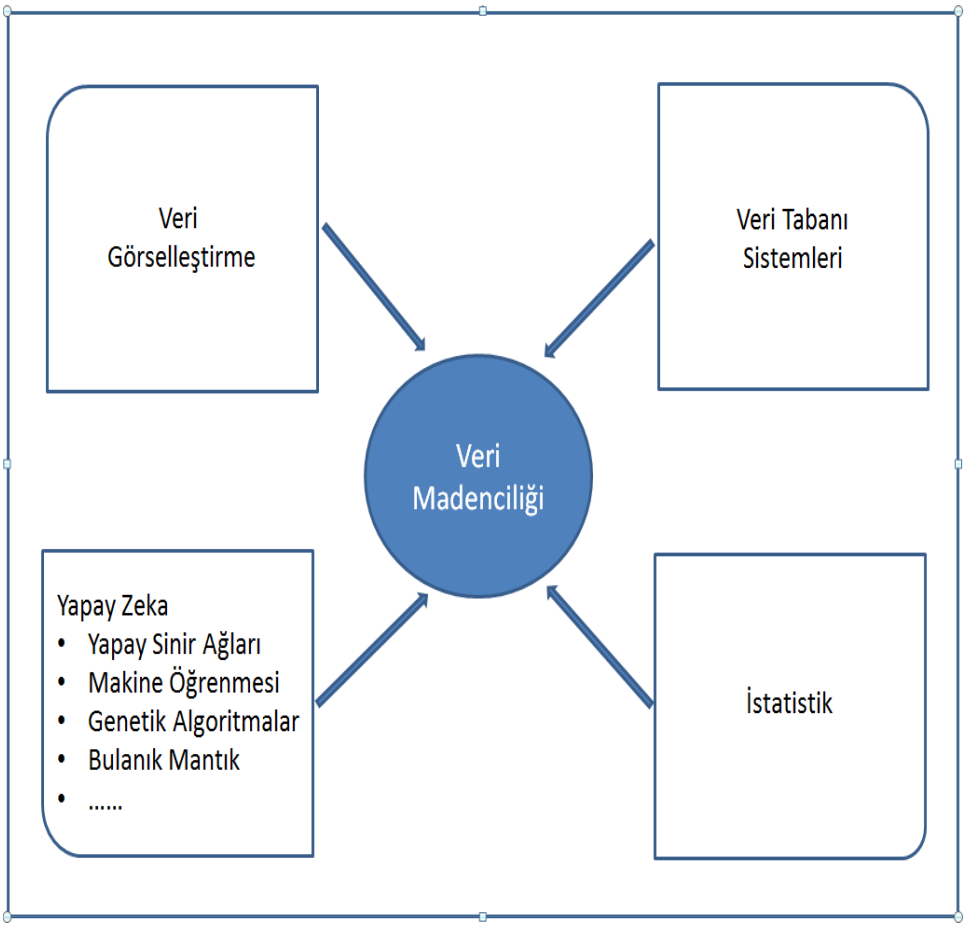
Veri madenciliği, önceden bilinmeyen, veri içinde gizli, anlamlı ve yararlı örüntülerin büyük ölçekli veri tabanlarından otomatik biçimde elde edilmesini sağlayan veri tabanlarındaki öz bilgi keşif analiz süreci içinde bir adımdır [14].

Veri Madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir [15].

Genel olarak veri madenciliği, eldeki yapılandırılmamış veriden, anlamlı ve kullanışlı bilgiyi çıkarmaya yarayacak tümevarım işlemlerini analiz etmeye ve uygulamaya yönelik çalışmaların bütününü içeren bir süreçtir. Geniş veri kümelerinden çeşitli desenleri, meydana gelen değişiklikleri, düzensizlikleri ve ilişkileri çıkarmakta kullanılmaktadır. Bu sayede, web üzerinde filtrelemeler, DNA sıraları içerisinde genlerin tespiti, ekonomideki eğilim ve düzensizliklerin tespiti, elektronik alışveriş yapan müşterilerin alışkanlıkları gibi karar verme mekanizmaları için önemli bulgular elde edebilmemize yardımcı olabilir [16].

Veri madenciliği ile büyük miktarda verilerden oluşan veri tabanları içerisinde gizli kalmış bilgilerin alınması sağlanır. Bu işlem, istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanı teknolojisi ve çeşitli bilgisayar yazılımları kullanılarak yapılır [7].

Veri madenciliği Şekil 2.1.'de görüldüğü gibi, veri görselleştirme, yapay zekâ, istatistik ve veri tabanları gibi alanlar ile yakından ilişkili disiplinler arası bir alandır.



Şekil 2.1. Veri madenciliğini oluşturan disiplinler

Veri madenciliği uygulamalarında modeller oluşturulurken, verideki gürültü ve eksik bilgiler giderilmekte ve istatistik bilimine dayalı tekniklerden faydalanılmaktadır. Verilerin depolanmasında veri tabanı sistemlerinden faydalanılmaktadır. Veri görselleştirme alanında ise verilerin tablo ve grafiklerle görüntülenmesi sağlanmaktadır.

Kullanılan veri madenciliği yaklaşımına bağlı olarak, yapay sinir ağları, bulanık mantık, genetik algoritmalar, mantıksal programlama ya da makine öğrenmesi gibi diğer teknikler ile kullanılabilir. Veri madenciliği sistemleri analiz türüne ve verinin içeriğine bağlı olarak uzaysal veri analizi (spatial data analysis), örüntü tanımlama (pattern recognition), görüntü analizi (image analysis), sinyal işleme (signal processing), bilgisayar grafikleri (computer graphics), web teknolojisi, ekonomi, iş

dünyası, biyoinformatik veya fizyoloji alanlarına ilişkin teknikler ile entegre olabilir [17].

Genel olarak veri madenciliğinde vurgulanan unsurların istatistiğin tanımı içinde yer aldığı görülmektedir. İstatistiksel uygulama aşamalarını, verilerin toplanması, sınıflandırılması, özetlenmesi, grafik ve tablolarla sunulması, analiz edilerek ana kütle hakkında anlamlı bilgiler elde edilmesi ve yorumlar yapılması olarak sıralayabiliriz. Veri madenciliğinde ulaşılmak istenen amaç ile istatistik biliminin amacı; verilerden bilgiyi keşfetmektir. Birçok tanımda veri madenciliğinde kullanılan temel aracın istatistiksel yöntemler olduğu belirtilmektedir. Her ikisinde de temel olan ögeler veri ve bilgidir. Bu nedenle birbiriyle oldukça örtüşen konulardır. İstatistiki açıdan bir tanım yapmak gerekirse Veri Madenciliği istatistik biliminin teknolojiyle bütünleşmesi sonucu oluşturulan bir araçtır [18].

İstatistiksel Analiz ile veri madenciliğinin karşılaştırması Çizelge 2.1.'de yer almaktadır [19].

Çizelge 2.1. İstatistiksel analiz ile veri madenciliği karşılaştırması

İstatistiksel Analiz	**Veri Madenciliği**
Genelde İstatistikçiler bir hipotezle başlar.	Veri madenciliğinde hipoteze gerek duyulmaz.
Hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorundadırlar.	Veri madenciliği algoritmaları, kendi eşitliklerini otomatik olarak geliştirir.
İstatistiksel analizler sadece sayısal verileri kullanmaktadır.	Veri madenciliği farklı tiplerde veriler kullanır. Sadece sayısal veri değil (metin, ses gibi) .
Kirli veriyi analizleri sırasında bulur ve filtre eder.	Veri Madenciliği tamamen temiz veriye dayanır.
İstatistikçiler kendi sonuçlarını yorumlar ve bu sonuçları yöneticilerine iletir.	Veri Madenciliği ile sonuçları yorumlamak kolay değildir. Sonuçlarını analiz etmede ve yorumlamada, bulguları yetkililere iletmede mutlaka bir istatistikçiye ihtiyaç duyulmaktadır.

2.3. Veri Madenciliği Uygulama Alanları

Veri madenciliği anlamsız veriden anlamlı bilgiler elde etmek için kullanılan yeni bir disiplin olmasına rağmen oldukça geniş bir kullanım alanına sahiptir. Veri madenciliği uygulama alanları gruplar halinde aşağıdaki gibi sınıflandırılabilir.

Bankacılık ve finans alanında;

- Kredi kartı dolandırıcılıklarının belirlenmesinde,

- Kredi kartı harcamalarına göre müşterilerin gruplandırılmasında,
- En iyi müşterinin tespitinde,
- Müşterilerin kredi taleplerinin değerlendirilmesinde,
- Vergi dolandırıcılıklarının tespitinde,
- Müşteri davranışlarına göre sınıflandırmada.

Sigortacılık alanında;

- Riskli müşterilerin davranışlarına göre tespit edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Poliçelerini yenilemeyecek müşterilerin tespitinde,
- Yeni poliçe alacak müşterilerin tahmininde,
- Riskli müşterilerin tahmininde.

Sağlık alanında;

- Tedavi sürecinin minimuma indirilmesinde,
- Tedavi sürecinin en aza indirilmesinde,
- İlaç kullanımında olası sahtekarlıkların belirlenmesinde,
- Tıbbi teşhis konulmasında,
- Tıbbi ürünlerin geliştirilmesinde,
- Test sonuçlarının tahmin edilmesinde,
- Hastalara ait tıbbi verilerden hastanın sağlık risklerinin tahmin edilmesinde.

Pazarlama alanında;

- Müşteri profillerinin belirlenmesinde,
- Müşteri ihtiyaçlarının belirlenmesinde,
- Kaybedilen müşterilerin benzer özelliklerinin belirlenmesinde,
- Müşterilerin elde tutulması için profillerinin belirlenmesinde,
- Müşteri davranışlarındaki özelliklerin sınıflandırılmasında,
- Çeşitli satış tahminlerinde (Sales Forecasting),
- Yapılacak satış miktarlarının tahmininde,
- Pazar sepeti analizi (Market Basket Analysis).

Mühendislik uygulamalarında;

- Örüntü tanımlama,
- Simülasyon,
- Sinyal işleme.

İnternet alanında;

- Web sayfalarında gezinen kullanıcıların profilinin belirlenmesinde,
- İnternet alışveriş siteleri kullanıcıların satın alma profillerinin belirlenmesinde,
- Web sayfalarını kullanan ziyaretçilerin sayfa içerisindeki davranışlarını analiz edilmesinde.

İmalat alanında;

- Etkin kaynak kullanımı,
- Araştırma ve geliştirme faaliyetlerinde,
- Ürün hatalarındaki sapmaların belirlenmesinde,
- Müşteri memnuniyet oranlarındaki sapmaların belirlenmesinde.

Telekomünikasyon alanında;

- Kaynak kullanımının iyileştirilmesinde,
- Geçmiş veriler kullanılarak dolandırıcılık yapan müşteriler için model oluşturma ve benzeri davranışları yapanları belirleme,
- Arama zamanı, mekânı, süresi, aranılan bilgiler gibi verilerden çeşitli örüntüleri tespit edilmesi,
- Kullanıcılara yönelik servis kalitesinin arttırılmasında.

Eğitim alanında;

- Öğrenci profillerine göre başarının tahmin edilmesinde,
- Benzer özellik gösteren öğrencilerin belirlenmesinde,
- Zeki ölçme ve değerlendirme sistemleri için bilgi geliştirmede,
- Öğrenme ortamlarının geliştirilmesine yönelik araştırma-geliştirme çalışmalarının yapılmasına,
- Başarılı e-öğrenme ortamlarının oluşturulabilmesi için çeşitli uygulamalar.

Biyomedikal ve DNA alanında;

- DNA dizilimindeki benzerliklerin karşılaştırılmasında,
- Zengin Genetik veri ambarlarının meydana getirilmesinde,
- Genler arasındaki ilişkilerin belirlenmesinde,
- Genlerin hastalıkların farklı seviyelerindeki etkilerinin belirlenmesinde,
- Biyomedikal verilerin anlaşılmasında görsel araçlardan faydalanılmasında.

2.4. Veri Madenciliğinin Tarihçesi

Veri madenciliği teriminin 1990'lı yıllardan itibaren tanıtılmasına rağmen geçmişi daha önceki yıllara dayanmaktadır. Veri madenciliği araştırmaları ve çalışmaları günümüze kadar çeşitli aşamalardan geçerek bugünkü haline gelmeyi başarmıştır.

Veri madenciliği teknikleri ile ilgili olarak ilk defa 1950'li yıllarda matematikçiler çalışmaya başlamışlardır. Mantık ve bilgisayar bilimleri alanlarında yapay zeka "artificial intelligence ve makine öğrenme "machine learning" konularını geliştirmişlerdir [20].

1960'lı yıllarda ise istatistikçiler yeni algoritmalar üzerinde çalışmışlardır. Örneğin regresyon analizi "regression analysis", en büyük olabilirlik kestirim "maximum likelihood estimates", sinir ağları "neural networks" gibi yöntemler başta gelmektedir. Bu yöntemler veri madenciliğinin ilk adımlarını oluşturmuştur [20].

1970, 1980, 1990'lı yıllarda yeni programlama dilleri ve bilgisayar tekniklerinin geliştirilmesi ile veri madenciliğindeki gelişim genetik algoritmalar "genetic algorithms", kümeleme yöntemleri "clustering methods", karar ağaçları "decision tree algorithms" gibi algoritmaları da içermiştir [20].

1990 yılının başlarından itibaren veri tabanlarından bilgi keşfinin ilk adımları atılmış ve büyük veri tabanları için veri ambarı veri tabanı "database warehouses" geliştirilmiştir. Ayrıca zaman içerisinde yeni teknolojilerle birlikte veri madenciliği değiştirilerek yaygın olarak kullanılarak standart bir işin parçası olmuştur [20].

Veri madenciliğinin tarihsel gelişimi Çizelge 2.2.'de gösterilmiştir.

Çizelge 2.2. Veri madenciliğinin tarihsel gelişimi

Zaman Aralıkları	**Gelişim Adımları**	**İşlem Soruları**	**Kullanılan Teknolojiler**
1960'lar	Veri Toplam	Son 3 yılda üniversitemizden mezun olan öğrenci sayısı nedir?	Bilgisayar, Diskler, Teypler.
1980'ler	Veri Erişim	Geçen yıl fakültelerden mezun olan öğrenci sayımız nedir?	İlişkisel veri tabanları SQL, ODBC.
1990'lar	Veri Ambarları ve Karar Destek Sistemleri	Geçen yıl fakültelerden mezun öğrenci sayısı nedir? Geçen yıl Yüksekokullardan mezun olan öğrenci sayıları ile karşılaştırmalı olarak.	OLAP, Çok boyutlu veri tabanı sistemleri ve veri ambarları.
1990'ların sonu ve bugün	Veri Madenciliği	Yüksekokullardan gelecek yıl mezun olabilecek öğrenci sayısı nedir? Ve neden?	Gelişmiş bilgisayar algoritmaları, Çok işlemcili bilgisayarlar, büyük veri tabanları.

Günümüzde bilgisayar teknolojilerinin hızla ilerlemesi ile birlikte veri miktarları ve bunların kullanımı artmakta ve vazgeçilmez bir ihtiyaç haline gelmektedir. Veri madenciliği çeşitli alanlarda farklı amaçlar için yaygın olarak kullanılmaktadır.

Faydalı sonuçlar alınmasından dolayı veri madenciliğine olan ilgi gün geçtikçe artış göstermektedir.

2.5. Veri Madenciliği Uygulama Adımları

Veri madenciliği kısaca gizli bilgilerin keşfedilmesi ile ilgili bir süreçtir. Birçok veri madenciliği yazılım geliştiricileri kullanıcılara yol göstermek amacı ile bir süreç modeli önerirler. Bu modeller ardışık aşamalardan oluşur. Her bir aşama bir önceki aşamanın sonuçlarına bağımlıdır.

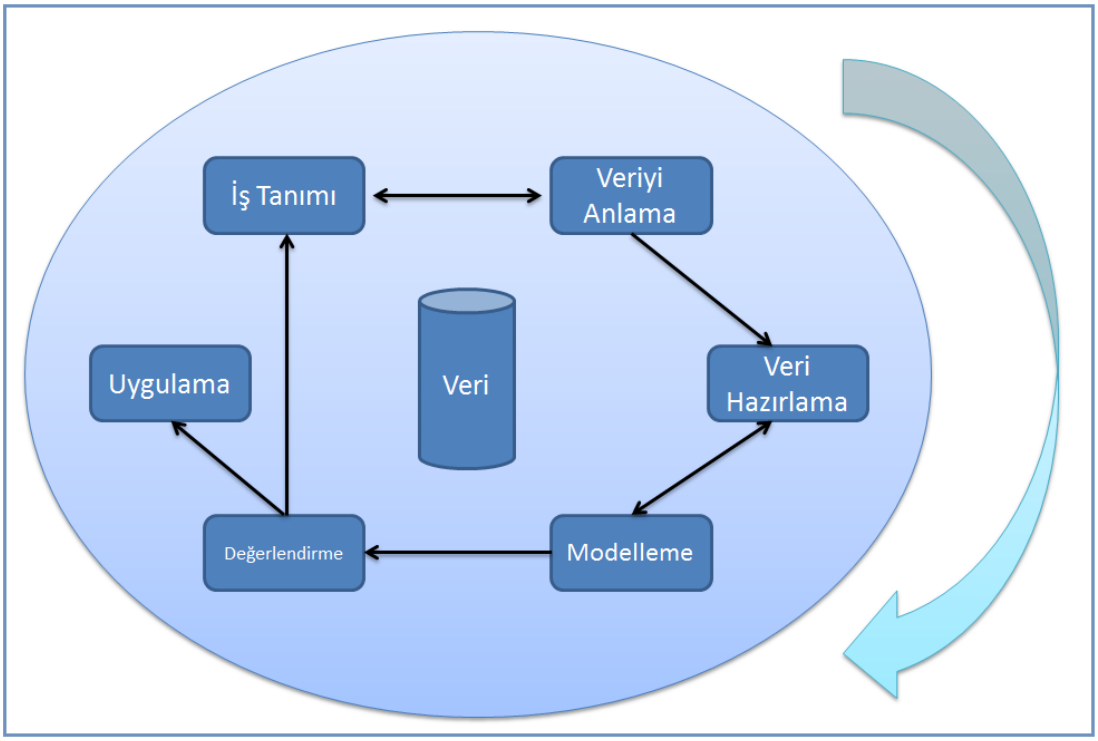
Veri madenciliği için belirlenen standart bir süreç söz konusudur. Bu standart süreç The Cross- Industry Standard Process for Data Mining (CRISP-DM) konsorsiyumu tarafından belirlenmiştir. CRISP-DM konsorsiyumu, 1996 yılının sonlarına doğru genç ve olgunlaşmamış veri madenciliği pazarında üç firma tarafından kurulmuştur [21].

Bu üç firmanın ilki olan Daimler Chrysler birçok endüstriyel ve ticari organizasyona, veri madenciliği tekniklerini uygulama konusunda öncü olmuştur. SPSS (Statistical Package for the Social Sciences) firması 1990 yılından beri veri madenciliği üzerine çeşitli hizmetler sağlamış ve ilk ticari veri madenciliği çalışma platformu olan Clementine‟i 1994 yılında harekete geçirmiştir. NCR (National Cash Register), müşterilerine değer katma içini sağlayabilmek ve alıcılarının ihtiyaçlarına hizmet edebilmek için birçok veri madenciliği danışmanlığı ve teknoloji uzmanlığı takımları kurmuştur [21].

Bu gelişmelerden bir yıl sonra, sözcüklerin baş harfleri “Cross- Industry Standard Process for Data Mining” açılımında olan CRISP-DM konsorsiyumu oluşturulmuş, Avrupa Komisyonundan fon elde edilmiş ve başlangıç fikirleri oluşturulmaya başlanmıştır [21].

CRISP-DM’in önerdiği sürecin ilk adımı “iş tanımı” adımıdır. Bu adımda çalışmanın amaçları ve ihtiyaçları belirlenir. Problem bu adımda tanımlanır. İkinci adım “veriyi

anlama" aşamasıdır. Bu adımda ilk adımda tanımlanan problemin çözümü için kullanılacak verilerin bir araya getirilmesi, verinin incelenmesi, veri kalite problemlerinin çözülmesi gibi işlem faaliyetlerini içermektedir.. Veri hazırlama aşamasında ise başlangıç veri kümesinden modelde kullanılacak veri kümesini oluşturmak için dönüşüm ve temizleme işlemleri uygulanır. Modelleme adımında problem ve veri özelliklerine uygun modelleme teknikleri seçilir ve model parametrelerinin en iyi değerleri belirlenir. Bu adımda uygulanan veri madenciliği teknikleri veri hazırlama adımına dönülmesini gerektirebilir. CRISP-DM uygulama sürecinin son iki adımında modelin değerlendirilmesi ve uygulamasına ilişkin görevler yer almaktadır [22]. CRISP-DM tarafından önerilen veri madenciliği adımları Şekil 2.2'de gösterilmiştir.



Şekil 2.2. Veri madenciliği süreci

Veri madenciliği sürecinde uygulama adımlarını aşağıdaki gibi sıralayabiliriz;

- Problemin tanımlanması,
- Veri tabanın oluşturulması,
- Verinin incelenmesi,
- Model için veri hazırlama,
- Modelin oluşturulması,
- Modelin değerlendirilmesi,
- Modelin uygulanması ve sonuçların izlenmesi [22].

2.5.1. Problemin Tanımlanması

Veri madenciliği uygulamalarında problemin tanımlanması adımı ilk adım olup en önemli aşamalarından biridir. Çalışmanın başarılı olabilmesi için işletme ya da organizasyonların amacı doğrultusunda problemin açık bir şekilde tanımlanması gerekmektedir. Problem ve amaçların açık olarak ifade edilmesi analizin doğru olarak yapılması için büyük önem taşımaktadır. Bu yüzden problemin tanımlanması adımı uygulama adımlarının arasında en zor olanıdır.

Problemin tanımlanması aşamasında, veri madenciliği uygulamasını yapacak olan kişi ilk olarak işletmenin geliştirmek istediği amacı dikkate almalıdır. Analizi yapan kişinin hedefi, veri madenciliği uygulamasının sonuçlarını etkileyebilecek önemli kriterleri ortaya çıkarmak olmalıdır. Veri madenciliği projesinin başarılı olması; projenin dikkatli bir şekilde planlanmış ve spesifik, gerçekleştirilebilir, ölçülebilir bir hedefin olmasına bağlıdır [21].

2.5.2. Veri Tabanın Oluşturulması

Bir diğer önemli aşama veri tabanının oluşturulması aşamasıdır. Veri madenciliği modeli oluşturma sürecinde ilk adım verilerin toplanmasıdır. Modelin kurulma aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır.

Bu aşamada elde olan veriler ve bunlara ek olarak toplanması gereken verilerin belirlenmesi gerekmektedir.

Veri tabanının oluşturulması aşaması, ilk adım olan problemin tanımlanması aşamasında tanımlanan problemin çözümünde ihtiyaç duyulan özellik ve nitelikteki verinin hazırlanması olarak ifade edilebilir. Bu aşamada veri kaynaklarının belirlenmesi, veri tanımlama, veri seçme, veri kalitesi ve ön hazırlık süreçleri, veri madenciliği veri tabanının yüklenmesi ve bakımı görevlerinin yerine getirilmesi ile tamamlanır. Bu adımları uygulamak zaman ve çaba açısından diğer tüm adımların uygulanmasından daha uzun zaman alır ve daha zordur. Veri hazırlama adımına, model geliştirme adımı gerçekleştirilirken geri dönmek gerekebilir. Bunun nedeni model oluşturma adımında modelden öğreneceğimiz herhangi bir enformasyonun veride değişiklik yapmamızı gerektirmesidir. Veri hazırlama adımları tüm bilgi keşfi süreci için harcanan zaman ve çabanın %50 ile %90 arası bir kısmını oluşturmaktadır [22].

2.5.2.1. Verinin Kaynaklarının Belirlenmesi

Bu aşamada tanımlanan problem için gerekli olduğu düşünülen veriler ve bu verilerin toplanacağı veri kaynakları belirlenir. Veriler birçok farklı kaynaktan elde edilebilir. Çeşitli kurumlar verilerini depolamakta farklı veri depolama ve yönetim sistemleri kullanabilmektedir.

Günümüzde veri depolama ve yönetim sistemlerinin uygulandığı yazılımların tablo olarak gösterimi Çizelge 2.3.'deki gibidir [22].

Çizelge 2.3. Veri depolama ve yönetim sistemlerinin uygulandığı yazılımlar

Kategori Adı	Yazılım Adı	Tanımı
Metin Editörleri	Note Pad	Basit metin çalışmaları için metin editörleri kullanılır. Notepad temel metin editörüdür, grafik ve OLE desteklemez. ".txt" uzantılı dosyaları açmak ve işlemekte, HTML yazmada Notepad kullanılır. Notepad *Ansi, Unicod*ve *UTF8* kodlarını destekler.
	Note Pad++	Notepad++, çok gelişmiş özelliklere sahip, standart bir notpad yazılımından defalarca büyük dosya açabilen ve yine defalarca kat hızlı işlem yapabilen ücretsiz bir yazılımdır.
Hesap Tablosu	Microsoft Excel	Microsoft Office yazılımında bu işi yapan program Excel adını taşır ve en çok kullanılan hesap tablosu yazılımıdır.
	Lotus 1-2-3	IBM, DB2 ve Oracle gibi veri tabanlarına öncülük eden, Excel ve Lotus Notes 'la uyumlu hesap tablosu yazılımıdır.
	Quatro Pro	Borland tarafından piyasa sürülmüştür. Daha sonra Corel yazılım şirketi tarafından satın alınmıştır. Windows tabanlı bir yazılımdır.
Veri tabanı	Microsoft SQL	Microsoft tarafından geliştirilmiş ilişkisel veri tabanı yönetim yazılımıdır. Zengin XML ve internet standartlarını destekleyen kullanıcılara bünyesindeki "stored procedureler" sayesinde XML formatındaki dosyalan kolayca depolama ve okuma olanağı tanır.
	Oracle	İlişkisel veri tabanı yönetim sistemi Oracle şirketinin ana ürünüdür. Bir istemci/sunucu veri tabanı yönetim yazılımıdır. Tam XML veri tabanı işlevi sağlar. Bünyesinde OLAP işlevlerini barındırır ve Windows ve Linux işletim sistemleri için oluşturulmuştur.
	IBM DB2	IBM tarafından geliştirilmiş ilişkisel veritabanı yönetim sistemidir. Unix başta olmak üzere Linux, IBM i, Z/OS ve Windows sunucularında çalışır.
OLAP	Microsoft OLAP	Microsoft SQL Server tarafından sağlanan analitik işleme, veri madenciliği ve raporlama aracıdır.
	Oracle Discover	Oracle'ın sağladığı OLAP çözümüdür. Sorgu, rapor, arama ve web yayını işlevlerini sağlamaktadır.
Veri Ambarı	SAP	Raporlama ve analiz için optimize edilmiş veri ambarı sistem yazılımıdır. Birçok ön tanımlı analiz modelini içerir. Raporlama aracı olarak Excel ve web sayfalarını kullanabilir. Yeni sorgular oluşturmada sürükle ve bırak teknolojisini kullanır.
	SAS	Artan veri yığını içinde değer yaratan çözümler sağlamayı amaçlayan, ileri yükleme, çıkarma ve dönüşüm tekniklerine sahip bir veri ambarı sistem yazılımıdır.

2.5.2.2. Veri Tanımlama

Bu aşamada veri madenciliği yapılacak verinin ayrıntıları tanımlanır. Veri kaynağında yer alan her tablo için raporlanması gereken bazı özellikler aşağıdaki gibi sıralanabilir [22].

- Tabloda yer alan sayısı
- Alan isimleri
- Veri Türü
- Açıklama
- Değer listesi
- Değer aralıkları

2.5.2.3. Veri Seçimi

Veri tanımlama aşamasından sonra veri seçimi aşamasına geçilir. Bu aşamada kurulacak model için veri seçimi yapılır. Model için gereksiz ve işlevsiz veri analiz dışı bırakılır. Örneğin isim, soy isim, kimlik numarası gibi model ile ilgili olmayan değişkenlerin modele girmesi algoritmaların yavaşlamasına, veriye ulaşma zamanlarının uzamasına neden olmaktadır.

2.5.2.4. Verilerin Birleştirilmesi ve Temizlemesi

Bu aşamada farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorun ve uyumsuzluklar mümkün olduğu ölçüde giderilerek, veriler tek bir veri tabanında toplanmaktadır. Eğer bu aşamada titiz davranılmazsa, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır [21].

Veri kalitesi problemlerinin farkına varılması ve doğrulanması veri temizleme olarak adlandırılır. Veri temizleme yoluyla eksik değerler tamamlanarak, gürültülü veri

düzeltilerek, aykırı değerler tanımlanarak veya çıkarılarak ve tutarsızlıklar giderilerek veri kalitesi arttırılmaya çalışılır [22].

2.5.3. Verinin İncelenmesi

Verinin incelenmesi kullanılacak verinin özelliklerinin daha iyi anlaşılmasını sağlar. Uygun veri analiz tekniğinin seçilmesine ve verinin kullanılacak model için hazırlanmasına yardımcı olur. Aynı zamanda veri madenciliği analizi tarafından cevaplanacak bazı sorulara ilişkin net ipuçları elde edilebilir. Örneğin, örüntüler görsel olarak verinin incelenmesi ile bulunabilir. Veri incelemesinde kullanılan görselleştirme gibi bazı teknikler de veri madenciliği sonuçlarını anlamada ve yorumlamada kullanılabilmektedir. Özet istatistikleri ve görselleştirme veri incelemesinde yaygın olarak kullanılan standart yöntemler arasındadır. Genellikle veri ambarlarında yer alan çok boyutlu verilerin incelenmesinde ise çok boyutlu veri analizinden faydalanılır. OLAP (On-Line Analytical Processing), verinin ve verideki önemli örüntülerin anlaşılması için kullanıcılara çok boyutlu veri tabanlarında inceleme yapmasına olanak sağlamaktadır.

OLAP görselleştirme gibi sadece veri madenciliği için tasarlanmış bir araç değildir. Çok boyutlu veri analizi, geçmişi çok gerilere dayanmayan çok boyutlu değerler dizilerini incelemek için kullanılan teknikler kümesidir [22].

2.5.4. Model Oluşturma

Veri madenciliğinde eldeki verilerden en fazla verimin alınabilmesi için model oluşturma aşaması büyük önem taşımaktadır. Veri madenciliği büyük boyutlardaki verilerin analiz edilerek en uygun hipotezlerin belirlenmesi ile ilgilenmektedir. Tahmin edici ve tanımlayıcı veri madenciliği görevlerinin başarılmasında istatistik disiplininden örnekleme, tahmin ve hipotez testlerinden faydalanırken yapay zeka, makine öğrenmesi, örüntü tanımlama disiplinlerinden de arama algoritmaları, modelleme teknikleri ve öğrenme teorileri kullanılmaktadır [22].

Veri madenciliği pek çok farklı algoritma kullanır. Bu algoritmalar veriyi inceler ve verinin özelliklerine en uygun modeli belirler. Verinin ve problemin özelliklerine göre uygulanabilecek birçok farklı algoritma sınıflama, kümeleme, birliktelik kuralları, örüntü tanımlama gibi görevlerin yerine getirilmesinde kullanılır [22].

2.5.5. Modelin Değerlendirilmesi

Çeşitli algoritmalar kullanılarak uygun model oluşturulduktan sonra sonuçların değerlendirilmesi ve bu sonuçların yorumlanması gerekmektedir. Model kurma aşaması uygun model bulunana kadar tekrar edilen bir süreçtir.

Model oluşturma sürecinde kullanılan modeller denetimli öğrenme ve denetimsiz öğrenme modelleri olarak farklılık göstermektedir.

Denetimli öğrenmede (örnekten öğrenme), bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı, verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir [21].

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanmakta ve yeni örneklerin hangi sınıfa ait olduğu, kurulan model tarafından belirlenmektedir [21].

Denetimsiz öğrenmede ise kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden yola çıkarak sınıfların tanımlanması hedeflenmektedir [21].

Denetimli öğrenmede seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı ile modelin öğrenilmesi, diğer kısmı ile de modelin geçerliliğinin test edilmesi için ayrılmaktadır. Modelin öğrenilmesi, öğrenim kümesi kullanılarak gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi belirlenmektedir [21].

Kurulan modelin doğruluk derecesi ne kadar yüksek olursa olsun, gerçek dünyayı tam anlamıyla modellediğini söylemek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olamamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir [21].

2.5.6. Modelin Uygulanması ve Sonuçlarının İzlenmesi

Bir veri madenciliği modeli oluşturulduktan sonra veri madenciliği iki şekilde uygulanabilir. Bunlardan ilki modelin sonuçlarına göre çeşitli faaliyetlerin önerilmesidir. Örneğin madencilik modelinin oluşturduğu kümelere veya modeli tanımlayan kurallara bakılarak faaliyet planları oluşturulabilir [15].
İkinci uygulama şekli ise elde edilen mevcut modelin kullanılan sistem içinde yerleştirilmesidir. Veri madenciliği modelleri genellikle risk analizi, kredi değerlendirme veya dolandırıcılık tespit süreçlerinde kullanılmaktadır. Bu durumlarda model bir yazılım haline getirilerek süreç içerisinde kullanılmaktadır. Örneğin tahmin edici bir model konut kredisi uygulaması ile birleştirilebilir. Bu durumda model, bir kredi uzmanının müşterisini değerlendirebileceği bir araç haline getirilebilir. Aynı şekilde model envanter sipariş gibi bir uygulama ile de birleştirilerek kullanılabilir. Sistem model sayesinde tahmini stok seviyeleri bir eşiğin altına düştüğünde otomatik olarak bir sipariş oluşturabilir. Buna benzer birçok iş sürecinde veri madenciliği yazılımları sistem yazılımlarına entegre edilerek uygulamalar gerçekleştirilebilmektedir [15].

Model uygulandıktan sonra sistemin sürekli olarak izlenmesi gerekir. Model ne kadar iyi çalışıyor olsa da zaman içerisinde tüm sistemlerin değişime uğrama ihtimali göz önünde bulundurulmalıdır. Değişen koşullara uyum sağlanması için modelin test edilmesi, tekrar eğitilmesi eğer gerekiyorsa yeniden oluşturulması gerekebilir. Tahmin edilen değerler ile gözlenen değerler arasındaki farklılıklar grafiksel model sonuçlarının takibi ile gözlemlenebilir. Hesaplamanın yoğun olmadığı bu tür

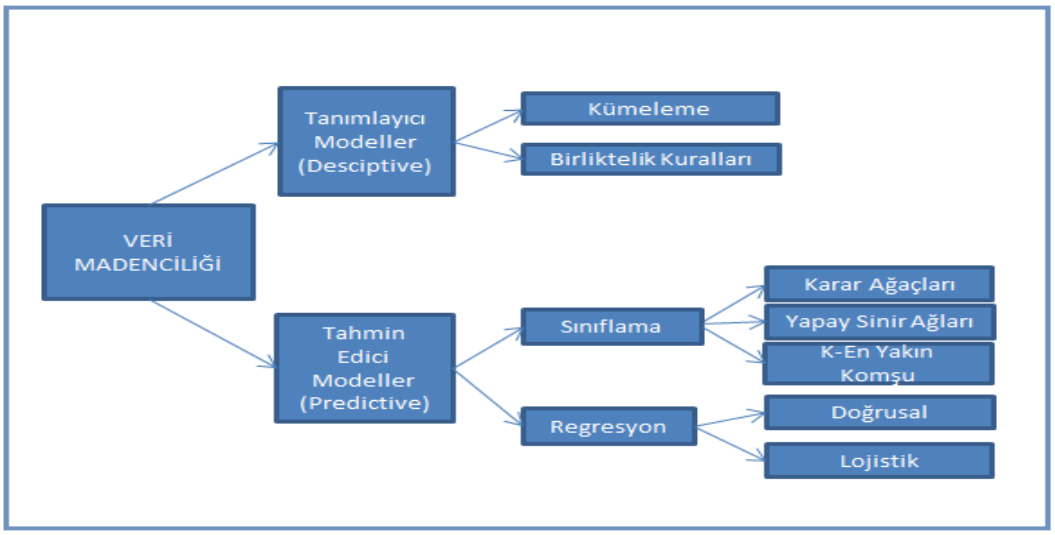
grafikleri kullanmak, anlamak kolaydır ve modeli uygulayan yazılımlar içine yerleştirilmesi ile sistemin kendini izlemesi sağlanabilir [22].

2.6. Veri Madenciliği Yöntemleri

Veri madenciliği büyük miktardaki verileri işleyebilen, bunlar arasında saklı bulunan örüntü ve eğilimleri keşfetme yeteneğine sahip bir süreçtir. Bu süreçte farklı görevleri yerine getirmek için farklı algoritmalar kullanılmaktadır. Bu algoritmaların amacı verilere en uygun modeli bulmaktır. Algoritmalar verileri inceler ve uygun modeli seçer.

Veri madenciliğinde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir [23].

Veri Madenciliği modelleri fonksiyonlarına göre sınıflandırma Şekil 2.3.'te görüldüğü gibi özetlenmiştir.



Şekil 2.3. Veri madenciliği modelleri

2.6.1. Tahmin Edici Modeller

Tahmin edici modellerde, sonuçları bilinen verileri kullanarak bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümelerinin sonuç değerlerinin tahmin edilmesi amaçlanmaktadır [23].

Örneğin, bir online alışveriş sitesindeki müşterilere ait veri setini düşünelim. Veri madenciliği teknikleri kullanarak müşterilere ait satın aldıkları ürün bilgileri ile ziyaret ettikleri ürünlerden elde edilmiş verilerinden bir tahmin modeli oluşturulabilir. Bu model sayesinde müşterilerin ne gibi ürünlere ilgi duyabileceği tahmin edilebilir ve site içerisinde yönlendirmeler yapılabilir.

2.6.1.1. Sınıflama

Veri madenciliği algoritmalarından ilki olan Sınıflama, veriler arasında önemli sınıflandırmaları tespit eden ve gelecek ile tahmin modelleri kurabilen bir veri analiz metodudur.

Sınıflama modelleri bir öğrenme algoritmasına dayanır. Veri tabanın bir kısmı örnek veri kümesi olarak belirlenir ve eğitim amacı ile kullanılır ve sınıflama kuralları oluşturulur. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir. Böylece hangi sınıfa ait olduğu bilinmeyen bir kayıt için bir sınıf belirlenebilir [24].

Sınıflama modelleri öğrenme verileri sayesinde oluşturulduğundan bir denetimli öğrenme olarak ifade edilebilir. Sınıflamada sınıf sayısı ve bir grup örneğin hangi sınıfa ait olduğunu bilinmektedir.

Örneğin bir sınıflama modeli banka-kredi uygulamalarında kredi kartı başvurularını düşük, orta ve yüksek risk gruplarına ayırmak amacı ile kurulabilir.

2.6.1.2. Karar Ağaçları

Karar ağaçları veri madenciliğinde akıllı veri analizi yapmak için kullanılan sezgisel ama güçlü bir araçtır. Karar ağaçları farklı değerli hedef fonksiyonlara yaklaşan bir yöntem olup burada öğrenilen işlevler, bir ağaç tarafından temsil edilmektedir [25].

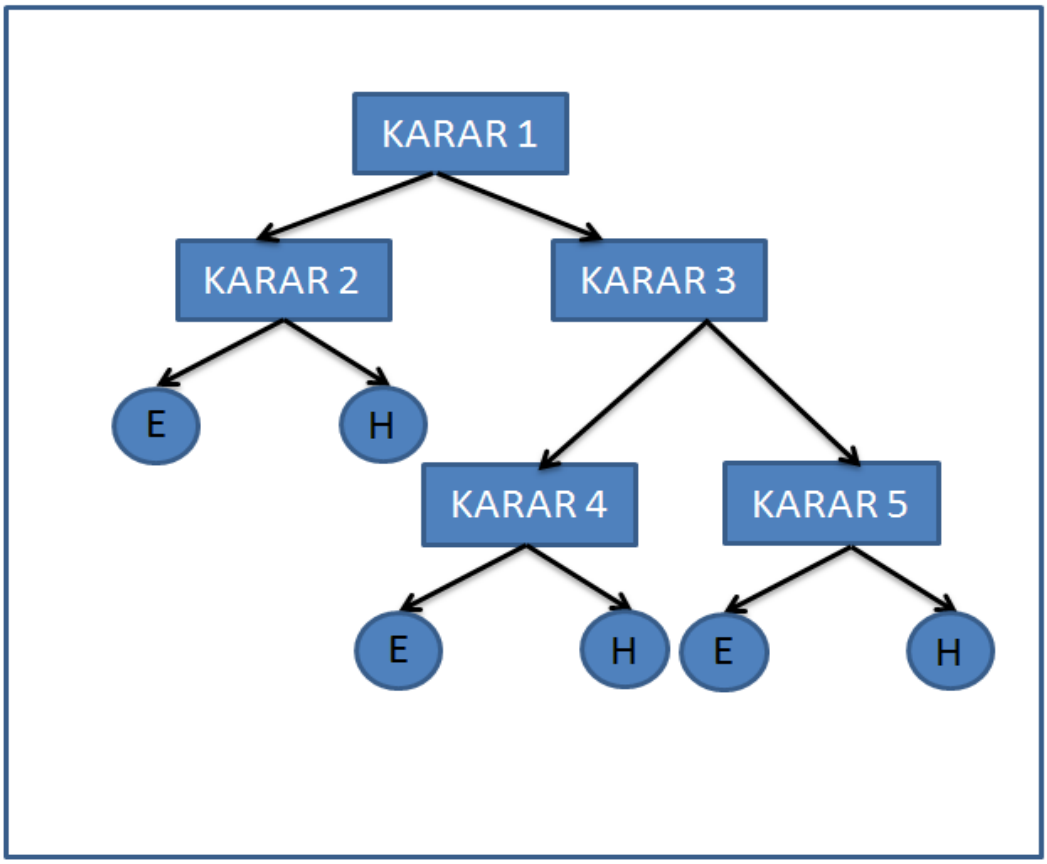
Ağaç yapısı sayesinde kolay anlaşılır kurallar üretebilen, fazla maliyet gerektirmeyen, yorumlanması kolay olan, veri tabanı sistemleri ile kolayca entegre olabilen bir tahmin edici bir tekniktir.

Karar ağaçlarının yapısı görünüm olarak bir ağaca benzemektedir. Kök, karar düğümleri, dallar ve yapraklardan oluşmaktadır. Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşemiyorsa, o düğüm sonucunda bir karar düğümü oluşur. Ancak düğüm sonunda belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden baslar ve yukarıdan aşağıya doğru yaprağa ulasana dek ardışık düğümleri takip ederek gerçekleşir [23].

Karar ağacı tekniğini kullanarak verinin sınıflanması iki basamaklı bir işlem ile gerçekleşir. Birinci basamak öğrenme basamağıdır. Öğrenme basamağında önceden bilinen bir eğitim verisi, model oluşturmak amacıyla sınıflama algoritması tarafından analiz edilir. Öğrenilen model, sınıflama kuralları veya karar ağacı olarak gösterilir. İkinci basamak ise sınıflama basamağıdır. Sınıflama basamağında test verisi, sınıflama kurallarının veya karar ağacının doğruluğunu belirlemek amacıyla kullanılır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflanması amacıyla kullanılır [23].

Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından

tahmin edilen sınıf ile karşılaştırılır. Eğer modelin doğruluğu kabul edilebilir bir değer ise model, sınıfı bilinmeyen yeni verileri sınıflama amacıyla kullanılabilir [18]. Şekil 2.4 te basit bir karar ağacı yapısı görülmektedir.



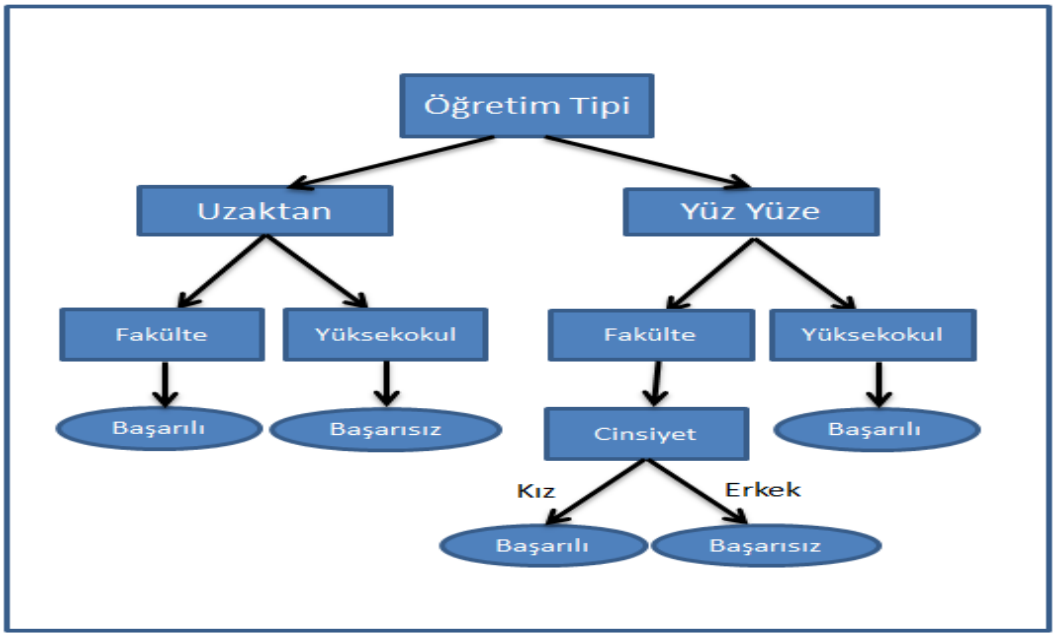
Şekil 2.4. Basit bir karar ağacı yapısı

Örneğin Çizelge 2.4.'de üniversitede eğitim gören öğrencilere ait küçük bir veri seti görülmektedir.

Çizelge 2.4. Üniversitede eğitim gören öğrencilere ait küçük bir veri seti

Öğretim Tipi	Birim	Cinsiyet	Başarı Durumu
Uzaktan	Fakülte	Erkek	Başarılı
Uzaktan	Fakülte	Erkek	Başarılı
Uzaktan	Fakülte	Kız	Başarılı
Yüz Yüze	Fakülte	Erkek	Başarısız
Yüz Yüze	Fakülte	Kız	Başarılı
Uzaktan	Yüksekokul	Kız	Başarılı
Yüz Yüze	Fakülte	Kız	Başarılı
Uzaktan	Yüksekokul	Erkek	Başarısız
Yüz Yüze	Yüksekokul	Erkek	Başarılı
Yüz Yüze	Yüksekokul	Kız	Başarılı

Veri setinde Öğretim tipi, birim ve cinsiyet olmak üzere üç adet tahmin edici değişken bulunmaktadır. Bu değişkenler yardımı ile öğrenci başarı durumunu belirlemek amacı ile Şekil 2.5.'te bir karar ağacı oluşturulmuştur.



Şekil 2.5. Çizelge 2.4.'den oluşturulan karar ağacı

Eğitim verisi incelenerek başarı durumu sınıfını tahmin edecek bir model oluşturulur. Bu modeli oluşturan bir sınıflama kuralı;

EĞER cinsiyet=Kız İSE VE Birim=Fakülte İSE VE ÖğretimTipi=YüzYüze İSE Başarı Durumu=Başarılı şeklindedir.

Bu kural gereğince fakülte öğrencisi olup dersi yüz yüze eğitim yöntemi ile alan kız öğrencilerin başarılı olduğu görülmektedir.

Oluşturulan model test verileri ile onaylandıktan sonra yeni verilere uygulanabilir ve sınıflama kuralı gereği yeni verinin sınıfı belirlenebilir.

Karar ağaçlarının bakımı ve anlaşılması verinin karmaşıklığının artmasıyla birlikte zorlaşır. Eksik verilerin olması durumunda bölünme, değişkenlerinin birisinin değeri bilinmiyorsa karara varılması mümkün değildir. Karar ağacı algoritması elimizdeki veriyi bölümlere ayırırken dikkat edilecek en önemli nokta, bağımlı değişkenin değerini en çok belirleyecek olan bağımsız değişkenleri ayırmaktır [26].

Algoritmaya ait adımlar aşağıdaki gibi sıralayabiliriz:

• Veri içinden ilgilendiğimiz bağımlı ve bağımsız değişkenlerin belirlenmesi,

• Hedef bağımlı değişkeni en çok etkileyecek olan bağımsız değişkenin bulunması, bu amaçla her değişkenin hedefi ne kadar etkilediğinin bulunması ve en çok etkileyen değişkenin seçilmesi (Burada amaç bölünmeden sonra kalan parçaların bölünme öncesine oranla daha sade olmasını sağlamaktır.).

En çok etkileyen değişkeni belirlemek için bilgi kazancını ölçümü yapılır. Entropi adı verilen bu yöntemle rastgelelik ve beklenmeyen durumların ortaya çıkma olasılığı hesaplanır.

Entropi matematiksel olarak aşağıdaki gibi tanımlanır:

$$\text{Entropi} = -\sum_{i=1}^{m} p_i \ log_2 (p_i) \quad (2.1)$$

Burada p_1 , p_2 p_3 ....p_m toplamları 1 olan olasılıklardır. Eğer örnekler aynı sınıfta ise Entropi değeri 0, aralarında eşit dağılmışlarsa Entropi değeri 1, rastgele dağılmışsa Entropi 0 ile 1 arasında bir değer alır.

Bölünme sonrasında kalan verilere aynı bölünme testlerinin yapılması ve daha sade gruplara ulaşılıncaya kadar bu işleme devam edilmesidir [26].

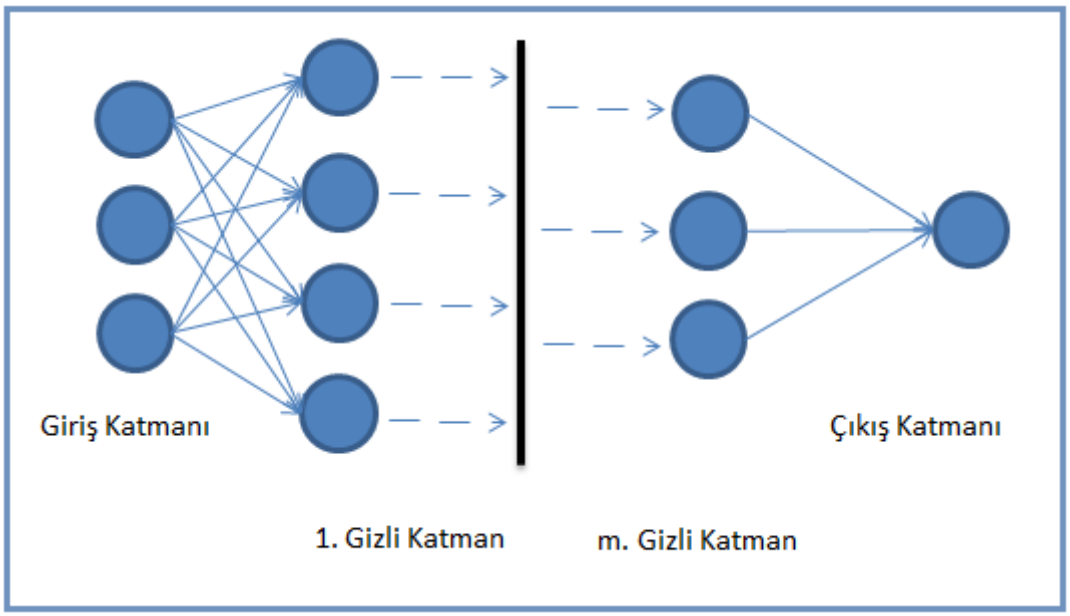
Risk grupları kategorileri oluşturmak, gelecekte olması muhtemel olaylar için tahmin kuralları oluşturmak, çeşitli kategorilerin birleştirilmesi, yeni bilinmeyen bir örneğin sınıflandırılması gibi durumlarda karar ağaçları kullanılmaktadır.

2.6.1.3. Yapay Sinir Ağları

Yapay sinir ağları (YSA), insan beyni örneklenerek geliştirilmiş bir teknolojidir. Öğrenme, hatırlama, düşünme gibi tüm insan davranışlarının temelinde sinir hücreleri bulunmaktadır. İnsan beyninde tahminen 10^{11} adet sinir hücresi olduğu düşünülmektedir ve bu sinir hücreleri arasında sonsuz diyebileceğimiz sayıda sinirler arası bağ vardır. Bu sayıdaki bir birleşimi gerçekleştirebilecek bir bilgisayar sisteminin dünya büyüklüğünde olması gerektiği söylenmektedir. İnsan beyninin bu karmaşıklığı göz önüne alındığında, günümüz teknolojisinin 1.5 kg'lık İnsan beynine oranla henüz çok geride olduğunu söylemek yanlış olmayacaktır [7].

Yapay sinir ağlarında amaç fonksiyon birbirine bağlı basit işlemci ünitelerinden oluşan bir ağ üzerine dağıtılmıştır. Yapay sinir ağlarında kullanılan öğrenme algoritmaları veriden üniteler arasındaki bağlantı ağırlıklarını hesaplar. YSA istatistiksel yöntemler gibi veri hakkında parametrik bir model varsaymaz yani uygulama alanı daha geniştir ve bellek tabanlı yöntemler kadar yüksek işlem ve bellek gerektirmemektedir [12].

YSA, öngörülen sayıda yapay sinir hücresinin, bazı amaçlarla belirli bir mimaride yapılandırılmasıyla ortaya çıkmaktadır. Bu mimari yapı, çok katmanlı bir yapı olmakla birlikte ilk katmanı, giriş katmanı olarak adlandırılmaktadır. Giriş katmanda herhangi bir işlem yapılmaz. Giriş katmanı ile çıkış katmanı arasında yer alan katmanlara ara katman denir. Bir ağ modelinde birden fazla ara katman bulunabilir. Çıkış katmanı aynı zamanda son katman olarak adlandırılır. Ara katmanların ortak adı ise gizli katmandır. Giriş-çıkış katmanları arasında bilgi aktarımı gizli katman üzerinden yapılmaktır. Çok katmanlı yapılarda herhangi bir katmanın çıkış sinyalleri bir sonraki katmanın giriş sinyalleri olarak kullanılmaktadır. Giriş katmanında k adet giriş nöronu, gizli katmanda h adet nöron ve çıkış katmanında q adet nöron bulunan tek katmanlı ileri beslemeli sinir ağı k-h-q ağı olarak bilinmektedir. Tam bağlantılı ağ yapısında her katmanda bulunan nöronlar bir sonraki katmanın tüm nöronlarına bağlıdır. Ayrıca, bir ağ modelinde sinaptik bağlantılardan bazıları eksik ise bu ağ, kısmi bağlantılı ağ adını almaktadır [27]. Şekil 2.6. da çok katmanlı bir YSA ağı gösterilmiştir.



Şekil 2.6. Çok katmanlı yapay sinir ağı

Yapay sinir ağları; halka arzlar, hisse senedi piyasaları tahmini, kredi değerlendirmesi, belirtilere göre hastalık tahmini, eğitim alanında veri setine göre öğrenci başarısını tahmini gibi alanlarında kullanılmaktadır.

2.6.1.4. k-En Yakın Komşu

Veri madenciliğinde sınıflama amacıyla kullanılan, denetimli öğrenme yöntemleri arasında yer alan, sınıflama problemlerini çözmeye yaran bir modeldir. Bu yöntemde, sınıflandırma yapılacak verilerin öğrenme kümesindeki normal veri kümelerine benzerlikleri hesaplanarak; en yakın olduğu düşünülen n tane verinin ortalamasının alınmasıyla elde edilen eşik değere göre sınıflandırma yapılmaktadır. Sınıflandırma yapılmadan önce, her bir sınıfın özelliklerinin önceden net bir şekilde belirtilmiş olması algoritmanın temelini oluşturmaktadır [28].

Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Her bir örnek n-boyutlu uzayda bir noktayı temsil eder. Bu şekilde tüm eğitim örnekleri n-boyutlu uzayda depolanır. Bilinmeyen bir örnek geldiğinde, bir k-en yakın komşu sınıflandırıcısı bilinmeyen örneğe en yakın k eğitim örneğini bulmak için bu örüntü uzayını tarar. K eğitim örnekleri bilinmeyen örneğin k-en yakın komşularıdır. Yakınlık Öklit mesafesi kullanılarak ölçülür.

Öklit mesafesi X = $(x_1; x_2; ; x_n)$ ve Y= = $(y_1; y_2; ; y_n)$ olarak adlandırılan iki nokta arasında;

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.2}$$

formülü ile bulunur [26].

k-en yakın komşu algoritmasını kısaca özetlemek gerekirse;

- Bütün örnekler n boyutlu uzayda bir nokta olarak alınır,
- Öklid mesafesi kullanılarak en yakın komşu belirlenir, $d(X_1 , Y_1)$

- Hangi sınıfa ait olduğu bilinmeyen X_i örneği, kendisine en yakın k örneğin sınıfına aittir denir [29].

2.6.1.5. Regresyon Analizi

Regresyon ile amaç, girdiler ile çıktıyı ilişkilendirecek modeli oluşturup, en iyi tahmine ulaşmaktır. Regresyon Analizi ile bir ya da daha çok değişkenin başka değişkenler cinsinden tahmin edilmesini sağlayacak ilişkiler belirlenir ve bunlar tanımlanır. Regresyon analizinin temelinde, gözlenen bir olayı değerlendirilirken, hangi olaylardan etkilendiğini belirlemek yatmaktadır. Bu olayların sayısı bir veya birden çok olabileceği gibi etki düzeyleri farklı seviyelerde de olabilir [30].

Regresyonda, verilerin matematiksel olarak, bir fonksiyon olarak tanımlanması gerekmektedir. Matematiksel modelde yer alan değişkenler bağımlı değişken ve bağımsız değişkenlerden oluşmaktadır. Değişkenler sayılabilir veya ölçülebilir niteliktedir. Örneğin; bir hissenin fiyatını ile ona dolaylı veya direkt etkili olan faiz oranları, enflasyon, vb. gibi değişkenler ile ilişkilendirmek mümkündür. Sadece faiz oranlarının etkisi ile ilgileniyorsak, tek değişkenli bir matematiksel model, faiz oranları ile birlikte enflasyon oranı ile de ilgileniyorsak, iki değişkenli bir matematiksel model kurulmalıdır [30].

Regresyon analizi iki değişken arasındaki ilişkiyi bulmak, ilişki varsa bu ilişkinin gücünü belirlemek, değişkenler arasındaki ilişkinin türünü belirlemek, ileriye dönük değerleri tahmin etmek gibi konularda kullanılır. Genel olarak araştırma, matematik, finans, ekonomi, tıp gibi bilim alanlarında yoğun olarak kullanılmaktadır. “Ev sahibi olan, evli, aynı iş yerinde 10 yıldan fazladır çalışan, geçmiş kredilerinde geç ödemesi bir ayı geçmemiş bir erkeğin kredi skoru 875dir.” sonucu bir regresyon ilişkisine örnek olarak verilebilir [30].

a) Doğrusal Regresyon

Bir bağımlı bir bağımsız değişkenden oluşan regresyon analizi "basit doğrusal regresyon", birden fazla bağımsız değişken içeren regresyon analizi de "çoklu regresyon analizi" olarak tanımlanmaktadır.

b) Lojistik Regresyon

Genelleştirilmiş doğrusal regresyon modelinin en yaygın tipi lojistik regresyon modelidir. Lojistik regresyon doğrusal regresyona çok benzer olmakla birlikte, lojistik regresyonda bağımlı değişkenin sürekli olmaması (kesikli veya kategorik) aralarındaki en önemli farklılıktır. Bu fark özellikle bir teklife yanıt veya bir secim yapmak gibi kesikli aksiyonları belirlemeye yönelik sınıflandırma modellerinde önem kazanmaktadır. Özellikle sınıflandırma analizlerinde doğrusal regresyonun kullanılması mümkün olmamaktadır. Lojistik regresyon, çok değişkenli normal dağılım varsayımına ihtiyaç göstermediğinden bu tür uygulamalarda avantaj sağlamaktadır [30].

Lojistik regresyon ile bağımsız değişkenleri kullanarak ikili çıktısı olan bağımlı değişkenin istenilen durumunun gerçekleşme olasılığını hesaplanır. Regresyon yapabilmek için bağımlı değişken sürekli değere dönüştürülür. Bu değer beklenen olayın olma olasılığıdır [30].

Örneğin; bir müşterinin kredibilitesinin iyi veya kötü olduğunu tahmin etmek için, lojistik yöntem iyi kredibilite olasılığını tahmin etmeye çalışır. Bağımlı değişkenin güncel durumu tahmin edilen olasılığa bakarak tespit edilir. Eğer tahmin edilen olasılık değeri 0.50 değerinden büyükse tahmin EVET (iyi kredibilite), diğer durumda ise HAYIR (kötü kredibilite) değerine yakındır. Bu yüzden lojistik regresyonda kredibilite "p" başarı olasılığı olarak adlandırılır. Diğer taraftan, girdilerin bazılarının sayısal olup olmaması lojistik regresyon modeli için önemli değildir. Bundan dolayı lojistik regresyon daha genel veri çeşitlerinin kullanımını desteklemektedir [26].

2.6.2. Tanımlayıcı Modeller

Tanımlayıcı modeller verilerdeki karar vermeye rehberlik etmede kullanılabilecek örüntüleri veya ilişkileri tanımlamaktadır. Kümeleme, özetleme, birliktelik kuralları, ardışık zamanlı örüntüler tanımlayıcı modeller olarak nitelendirilir.

2.6.2.1. Kümeleme

Kümeleme kısaca verileri sınıflara veya kümelere ayırma işlemine verilen addır. Aynı kümedeki elemanlar birbirleriyle benzerlik gösterirlerken, başka kümelerin elemanlarından farklıdırlar. Kümeleme veri madenciliği, istatistik, biyoloji, sosyoloji, arkeoloji, psikoloji ve makine öğrenimi gibi pek çok alanda kullanılan bir yöntemdir. Kümeleme modelinde, sınıflama modelinden farklı olarak veri sınıfları yoktur. Yani verilerin herhangi bir sınıfı bulunmamaktadır. Sınıflama modelinde, verilerin sınıfları bilinmekte ve yeni bir veri geldiğinde bu verinin hangi sınıftan olabileceği tahmin edilmektedir. Kümeleme modelinde ise sınıfları bulunmayan veriler gruplar halinde kümelere ayrılırlar. Bazı uygulamalarda kümeleme modeli, sınıflama modelinin bir önişlemi gibi görev alabilmektedir [23]. Sınıflama ile kümeleme arasındaki en önemli fark sınıflamada sınıfların önceden belli olmasıdır. Kümeleme analizinde ise eldeki veri setine göre kümeler oluşturulmaktadır [21]. Hangi nesnenin hangi sınıfa ait olduğu ve grup sayısı belirsizdir.

Örnek olarak; biyolojide bitki ve hayvan sınıflandırmaları ve işlevlerine göre benzer genlerin sınıflandırılması, marketlerde farklı müşteri gruplarının keşfedilmesi ve bu müşteri grupların alışveriş örüntülerinin belirlenmesi, şehir planlanmasında evlerin tiplerine, değerlerine ve coğrafik konumlarına göre gruplara ayrılması gibi uygulamalar verilebilir. Kümeleme aynı zamanda Web üzerinde bilgi keşfi için dokümanların sınıflanması amacıyla da kullanılmaktadır [23].

Örneğin öğrenci yaşlarını gösteren bir Yaş değişkeni olduğunu varsayalım.

Yaş= (18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 33, 37, 44, 45, 46)

Yukarıda belirtilen öğrenci yaş değerleri dikkate alındığında

1. Küme = (18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29)
2. Küme = (33, 37, 44, 45, 46)

İki küme olacak şekilde kümeleri oluştuğu görülmektedir.
Yaş veri seti (17, 18, 18, 19, 20, 20, 20, 20, 21, 22, 22, 22) değerlerinden oluştuğunu varsayarsak veri tabanını

1. Küme = (17, 18, 18, 19)
2. Küme = (20, 20, 20, 21, 22, 22, 22, 22)

Olarak 2'ye ayrıldığı görülmektedir. İlk ayrımda 20, 21, 22 yaş 1. Kümede iken ikinci ayrımda 2. Küme de yer almaktadır. Bunun nedeni olarak kümeler oluşturulurken veri tabanındaki benzerliğe göre kümeleme yapıldığı söylenebilir.

2.6.2.2. Birliktelik Kuralları

Büyük veri kümeleri arasında birliktelik ilişkileri bulmak için Birliktelik Kuralları kullanılır. Büyük miktardaki mesleki işlem kayıtlarından ilginç birliktelik ilişkilerini keşfedilmesi, şirketlerin karar alma sürecinde işlemlerini daha verimli hale getirmektedir. Veri madenciliği uygulamalarında en yaygın birliktelik kurallarının kullanıldığı işlem market sepeti analizi uygulamasıdır. Müşterilerin yapmış oldukları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını analiz eder. Bu tip birlikteliklerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarmaya yardımcı olur ve market yöneticileri de bu bilgi ışığında daha etkili satış stratejileri geliştirebilirler. Örneğin bir müşteri süt satın alıyorsa, aynı alışverişte sütün yanında ekmek alma olasılığı nedir? Bu tip bir bilgi ışığında rafları düzenleyen market yöneticileri ürünlerindeki satış oranını arttırabilirler. Örneğin bir marketin müşterilerinin süt ile birlikte ekmek satın alan oranı yüksekse, market yöneticileri süt ile ekmek raflarını yan yana koyarak ekmek satışlarını arttırabilirler [23].

Bu analizin sonucunda amaç, X hizmeti talebiyle Y hizmeti talebi arasında kuvvetli bir ilişki bulunması durumunda X hizmetini talep eden müşteriye Y hizmetini sunmaktır.

2.7. WEKA

WEKA, Yeni Zelanda'daki Waikato Üniversitesi tarafından geliştirilmiş, 1996 yılında ilk resmi sürümü yayınlanmış, makine öğrenimi algoritmalarının bir arada bulunduran, işlevsel bir grafik ara yüzüne sahip, açık kaynak kodlu bir veri madenciliği programıdır. Çeşitli akademik araştırmalarda, eğitim ve endüstriyel uygulamalarda geniş bir kullanım alanına sahiptir. WEKA çeşitli veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişkilendirme kuralları ve görselleştirme araçları içerir. Algoritmalar veri kümesine doğrudan veya Java kodundan çağrılarak uygulanabilir. Aynı zamanda yeni makine öğrenme algoritmaları geliştirmek için de uygundur. Veri analizi ve tahminleyici modelleme için geliştirilmiş algoritmalar içeren ve görsel raporlar sunabilen WEKA Programı çalıştırıldığında Şekil 2.7 'deki kullanıcı ara yüzü ekrana gelir [31].

Şekil 2.7. WEKA programı ara yüzü

2.8. Kullanılan Veri Madenciliği Sınıflama Algoritmaları

2.8.1. J48 Algoritması

J48 algoritması, C4.5 Karar Ağacı algoritmasını kullanan WEKA'nın sınıflandırma algoritmalarından biridir. Sayısal nitelikli karar ağaçlarının oluşturulmasına imkan tanımaktadır.

C4.5 algoritmasının genel özellikleri şu şekilde sıralanabilir:

• Eksik veri: Karar Ağacı kurulduğunda eksik veri basitçe ihmal edilmektedir. Yani kazanç oranı yalnızca, söz konusu parametre için bir değere sahip diğer kayıtlara bakılarak hesaplanmaktadır. Eksik bir parametre değeri olan bir kaydı sınıflandırmak için söz konusu kalemin değeri, diğer kayıtlar için parametre değerlerine ilişkin bilgiler kullanılarak tahmin edilebilir.

• Sürekli veri: Sayısal değerlere sahip değişken içerisinde uygun eşik değeri bulunduktan sonra ikili ya da daha çok bölünme ile veri kümesi bölünebilir.

• Budama: C4.5'te önerilmekte olan iki ana budama stratejisi mevcuttur:

a. 'Alt ağaç değiştirmeli' adı verilen stratejide değiştirme sonucunda, başlangıçtaki ağacınkine yakın hata oranı elde edilebiliyorsa alt ağaç, yaprak düğümle değiştirilir. Alt ağaç değiştirme, ağacın altından yukarıdaki köküne kadar uygulanır.

b. 'Alt ağaç yükseltme' adı verilen diğer budama stratejisiyle alt ağaç, en çok kullanılan kendi alt ağacıyla değiştirilir. Böylece alt ağaç, mevcut konumundan ağacın daha üst noktasındaki düğüme yükseltilir. Yine de bu değiştirme için hata oranındaki artışı tespit etmemiz gerekmektedir.

• Ayırma: Aşırı uyum (overfitting) sebebiyle oluşan hatalar, C4.5 tarafından geliştirilmekte olan bir yöntemle telafi edilmeye çalışılmaktadır [32].

Sayısal nitelikleri belirli aralıklara bölme konusunda bazı zorluklar görülebilir. Ancak en uygun t eşik değerini hesaplamak için çeşitli yöntemler bulunmaktadır. Nitelik değerleri sıralanır ve $\{V_1, V_2, \ldots, V_n\}$ şeklini alır. Nitelik değerler kümesi iki parçaya ayrılır ve Eşik değeri olarak $[V_i, V_{i+1}]$ Aralığının orta noktası olarak alınabilir [33]:

$$t_i = (V_i + V_{i+1})/2 \quad (2.3)$$

C4.5 algoritması sınıflandırmada en ayırıcı özelliğe sahip değişkeni bulurken Entropi kavramından yararlanır. Entropi kavramı eldeki verinin sayısallaştırılmasıdır. Entropi bir veri kümesi içindeki belirsizliği ve rastgeleliği ölçmek için kullanılır.

Veri tabanının tamamının entropisi hesaplanır; eğer veri tabanı farklı bölümlere ayrılıyorsa her bir alt bölümün de entropisinin hesaplanması gerekir. C4.5 algoritması kullanılarak ağaç elde edilirken her bir alt ağaçlar yapraklara dönüştürülür. Ağaç yapısı oluşturmak için, her bir alt ağacın yaprağa dönüşümü kazanım ve ayırma oranları ile gerçekleştirilir [33].

2.8.2. JRip Algoritması

JRip, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algoritmasının WEKA uygulamasıdır [34].

RIPPER Algoritması "if …. then…. Kural tablosunu kullanan doğrudan kural tabanlı bir sınıflandırma tekniğidir. Amacı gürültülü veriler üzerinde etkili kural geliştirmektir ve bu doğrultuda C4.5 algoritması ile rekabet etmektedir.

Öğrenme algoritması IREP'in (Incremental Reduced Error Pruning) gelişmiş bir versiyonu olan RIPPER kural öğrenme algoritması, tüm olumlu örneklerin

kapsamakta ve algoritmanın, gürültülü veri setleri üzerine etkili performans gösterdiği bir kural setinden oluşmaktadır [35].

Bir kural oluşturulmadan önce mevcut eğitim örnekleri seti iki alt sete ayrılır. Bunlardan biri gelişen kurallar listesi (genellikle 2/3) diğeri de budama listesi (genellikle 1/3) dir. Kural gelişen kurallar listesindeki örneklerden oluşturulur. Bütün olumsuz örnekler belirlenene kadar kurallar bu listeye kurallar ilave edilir [35].

Gelişen kurallar listesinde bir kural geliştirildikten sonra kural listesinin performansını arttırmak için budama (kural silme) yapılır. Bir kuralın budamasında RIPPER yalnızca bu kuraldan oluşan son koşulu göz önünde bulundurur [35].

Algoritmanın sonunda eğitim veri setine göre Eğer-O-Zaman kuralları listesi elde edildikten sonra yeni bir örneğin sınıflandırılmasında sırayla kullanılır. Eğer listedeki ilk kural örneği kapsamıyorsa, yani hem kural hem örnekteki nitelikler için eşleşen değerler yoksa o zaman bir sonraki kural denenir. Sırayla örnek bir kural tarafından sınıflandırılana kadar devam eder. Eğer kural hiçbir örneği kapsamıyorsa o zaman karar listesinin en altında varsayılan bir kural işletilir. Yani sınıflandırılamayan tüm örnekler bu sınıfta toplanır [36].

Karar setleri gürültülü eğitim veri setine uyma problemi ile karşı karşıyadır. Bu yüzden genellikle budama işlemi yapılır.

Bu tez çalışmasında JRip algoritması ile yapılan modellemede elde edilen kurallar Şekil 2.8'de görülmektedir.

```
JRIP rules:
===========

(OGRTIP = UZAKTAN) and (SINIF = FAKULTE) => SONUC=COKIYI (200.0/28.0)
(OGRTIP = UZAKTAN) and (YAS = 20_21_YAS) => SONUC=COKIYI (33.0/12.0)
 => SONUC=ORTALAMA (409.0/113.0)

Number of Rules : 3
```

Şekil 2.8. JRip algoritma kuralları

Şekil 2.8.'de yer alan 3 adet JRip kuralının, kodları açık hale getirildiğinde oluşan açıklamalar Çizelge 2.5.'de yer almaktadır.

Çizelge 2.5. JRip kural açıklamaları

Kural No	**Kural**	**Açıklama**
1	(OGRTIP = UZAKTAN) and (SINIF = FAKULTE) => SONUC=COKIYI	Dersi uzaktan eğitim ile alan fakülte öğrencilerinin başarı durumu çok iyidir.
2	(OGRTIP = UZAKTAN) and (YAS = 20_21_YAS) => SONUC=COKIYI	Dersi uzaktan eğitim ile alan 20-21 yaşlarındaki öğrencilerin başarı durumu çok iyidir.
3	=> SONUC=ORTALAMA	Yukarıdaki 2 kurala da uymayan öğrencilerin başarı durumu ortalamadır.

JRip algoritmasının avantajlı yönlerini aşağıdaki gibi sıralayabiliriz:

- Kural Kümesini yorumlamak diğerlerine göre daha kolaydır.
- Karar Ağacı öğrenmesine kıyasla daha iyi öğrenir.
- Birinci dereceden mantık gösterilemeyen uygulamalarda kolay uygulanabilir.

2.8.3. Çok Katmanlı Algılayıcı (Multilayer Perceptron) Algoritması

Çok katmanlı algılayıcılar (ÇKA) birçok problemi çözmede YSA modelleri içerisinde en çok kullanılan bir ağ tipidir. Bir girdi katmanı, en az bir tane gizli katman ve bir çıktı katmanına sahiptir. Girdi katmanından gelen sinyaller ileri doğrultulu olarak, katman katman ilerler. Genelde geri yayılım (Backpropogation) algoritması ile öğrenmektedir [37].

Çok katmanlı algılayıcıda gizli katmandaki her bir j nöronu, w_{ji} bağlantı ağırlığıyla giriş işaretlerinin çarpımlarının toplamını alır ve y_j çıkışını bu toplamın bir fonksiyonu olarak hesaplar:

$$y_j = f(\sum w_{ji}\ x_i) \tag{2.4}$$

Burada f bir nörona etki eden işaretlerin ağırlıklı toplamını çıkış değerine dönüştüren bir aktivasyon fonksiyonudur. Aktivasyon fonksiyonu basit bir eşik fonksiyonu, sigmoidal veya hiperbolik tanjant fonksiyonu olabilir. Çıkış nöronlarının hesaplanan ve istenen değerleri arasındaki karesel farkların toplamı aşağıdaki gibi tanımlanır:

$$e = \frac{1}{2}\ \sum_J\ (y_j^* - y_j)^2 \tag{2.5}$$

Burada y_j^* ve y_j sırasıyla j nci çıkış nöronunun hesaplanan ve istenen değerleridir. Her bir w_{ji} ağırlığı olabildiğince hızlı bir şekilde e değerini azaltmak için ayarlanır. w_{ji} değerinin nasıl ayarlanacağı eğitim algoritmalarına bağlıdır. [38].

Çok katmanlı algılayıcı ağında girdi sinyali katmanlar boyunca ileriye doğru işlenerek çıktı değerleri üretilmektedir. Bu ağ tipinde bilgi sırasıyla üç katmandan ileri doğru geçirilmektedir. İlk katman olan girdi katmanının görevi; bilgiyi bir sonraki katman olan gizli katmana iletmektir. Burada veri üzerinde herhangi bir işlemden yapılmamaktadır. Gizli katman olarak adlandırılan ikinci katmanda ise yer alan sinir hücreleri veriyi transfer fonksiyonundan geçirerek bir sonraki katman olan çıktı katmanına gönderirler. Çıktı katmanı kendisine gelen bu bilgiyi kendi transfer fonksiyonundan geçirerek çıktıyı üretmiş olur. Elde edilen çıktı istenilen çıktı değeriyle karşılaştırılır. Ortaya çıkan hata payı ağ boyunca ağırlıklara geri yayılır. Düzenlemeler yapılarak veri seti ağa tekrar sunulur, başta anlatılan süreç tekrarlanarak çıktı tekrar elde edilir ve istenilen çıktı değeriyle karşılaştırılır. Bu işlem istenilen çıktıya kabul edilir hata değerince yaklaşılana kadar sürdürülür. Bu işlem için kullanılan en yaygın kabul görmüş algoritma geri yayılım algoritmasıdır [32].

ÇKA ağları denetimli öğrenme stratejisine göre çalışırlar. Bu ağlara eğitim sırasında hem girdiler hem de o girdilere karşılık üretilmesi gereken (beklenen) çıktılar gösterilir. Ağın görevi her girdi için o girdiye karşılık gelen çıktıyı üretmektir. ÇKA ağının öğrenme kuralı en küçük kareler yöntemine dayalı Delta Öğrenme Kuralının genelleştirilmiş halidir. O nedenle öğrenme kuralına Genelleştirilmiş Delta Kuralı da denmektedir. Verilerin bir kısmı ağı eğitmek için kullanılırken, diğer kısmı test etmek için kullanılır. Delta Kuralına göre önce ağın çıktısı hesaplanır daha sonra da hata oranını min. yapacak şekilde ağırlıklar değiştirilir [32].

2.9. Sınıflandırma Modelini Değerlendirme

Sınıflandırma Metodu tarafından oluşturulan modelin başarısını ölçmek için; Doğruluk (Accuracy), Duyarlılık, Belirlilik, Hassaslık (Precision) gibi ölçüler kullanılır.

İki Sınıflı bir model için sınıflama matrisi Çizelge 2.6'da görülmektedir.

Çizelge 2.6. İki Sınıflı bir model için sınıflama matrisi

	Gerçek Sınıf (Actual Class)		
Öngörülen Sınıf (Predicted Class)		Pozitif (Positive)	Negatif (Negative)
	Pozitif (Positive)	Doğru Pozitif Sayısı (True positive) DP	Yanlış Pozitif Sayısı (False positive) YP
	Negatif (Negative)	Yanlış Negatif Sayısı (False negative) YN	Doğru Negatif Sayısı (True negative) DN

Modelde Bulunan Toplam Örnek Sayısı = DP + YP +YN + DN =N (2.6)

Modelin Doğruluk Oranı (Doğruluk) = (DP + DN) / N (2.7)

Doğru Pozitif Oranı(Duyarlılık) = DP / (DP + YN) (2.8)

Doğru Negatif Oranı (Belirlilik) = DN / (DN + YP) (2.9)

Hassaslık (Precision) = DP / (DP + DN) (2.10)

3. ARAŞTIRMA BULGULARI

Tez uygulaması Kırıkkale Üniversitesi öğrenci bilgi sisteminden alınan verilerden yararlanarak gerçekleştirilmiştir. Bu çalışmada ön lisans ve lisans öğrencilerine ait 672 adet veri kullanılmıştır. ENF-101 kodlu Temel Bilgi Teknolojileri Kullanımı (TBTK) dersi bazı bölümlerde geleneksel bir yöntem olan yüz yüze eğitim ile bazı bölümlerde ise yeni bir yöntem olan uzaktan eğitim yolu ile verilmektedir. Her iki eğitim sistemi için öğrencilerin akademik performansları araştırılmıştır. Farklı eğitim sistemlerinin öğrencilerin başarısı üzerine etkisi incelenmiştir. Öğrencilerin bölüme yerleştirmede esas alınan puan türünün TBTK dersindeki başarısına etkisinin olup olmadığı incelenmiştir. Öğrencileri ön lisans, lisans, kız, erkek, dersin alındığı dönemlere göre başarı notları kıyaslanmıştır. Yapılan çalışma sonucunda, farklı iki eğitim türünde eğitim gören öğrencilerin farklı kriterlere göre başarısızlıkları ve bu başarısızlıklarının nedenini bulup çözümlemek hedeflenmiştir. Farklı iki eğitim sistemi olan yüz yüze ve uzaktan eğitim sistemlerinin karşılaştırılması yapılmıştır. Üniversite bünyesinde verilmekte olan başka derslerde de uzaktan eğitim yönteminin kullanılabilirliği hakkında bilgi vermesi amaçlanmıştır. Uygulama WEKA 3.7 (with console) programı yardımı ile yapılmıştır.

3.1. Verinin Tanımlanması ve Hazırlanması

Öncelikle öğrencinin başarısına etkisi muhtemel kriterler belirlenmiştir. Bunlar; öğrencinin bölüme yerleştirmede esas alınan puan türü (sayısal, sözel, eşit ağırlık, yabancı dil, özel yetenek, sınavsız geçiş), öğrencinin eğitim gördüğü akademik birim (fakülte-yüksekokul), öğrencinin cinsiyeti (kız, erkek), öğrencinin başarı durumu (çok iyi, ortalama, başarısız), öğrencinin dersi aldığı dönem (güz, bahar), dersin verildiği eğitim sistemi (yüz yüze eğitim, uzaktan eğitim) ve öğrencinin yaşı olarak belirlenmiştir.

Öğrenci bilgi otomasyonundan elde edilen bilgiler Oracle veri tabanında yeni bir tablo oluşturularak kaydedilmiştir. Oluşturulan yeni veri tabanında gerekli incelemeler yapılmıştır.

Yapılan incelemeler sonucunda;

- Veri tabanından öğrencinin adı, soyadı, ara sınavlarda ve finallerde almış olduğu notlar gibi gereksiz alanlar temizlenmiştir.
- Verilerin rahat modellenebilmesi için bazı alanların yapısı değiştirilmiştir. Çizelge 3.1. de öğrencinin dersten aldığı notlara göre oluşturulmuş başarı durumları görülmektedir.

Çizelge 3.1. Başarı durumlarının gruplandırılması

Not Aralığı	Başarı Durumu
0-59	BAŞARISIZ
60-79	ORTALAMA
80-100	ÇOK İYİ

- Bölümler fakülte ve yüksekokul olarak gruplandırılmıştır.
- Yerleştirme puanlar türleri sayısal, sözel, eşit ağırlık, yabancı dil, özel yetenek, sınavsız geçiş olarak gruplandırılmıştır.
- TBTK dersine devam etmeyen ya da sınavlarına girmeyen öğrencilere ait bilgiler çalışmaya dâhil edilmemiştir.

Veri temizleme işlemi sonucunda 642 kayıt elde edilmiştir. Bu kayıtlara ait veri madenciliği uygulamasında kullanılacak veri seti üzerindeki veri dağılımı Çizelge 3.2.'de verilmiştir.

Çizelge 3.2. Veri madenciliği çalışması için kullanılacak verilerin dağılımı

Sınıf	**Sayı**
ÇOK İYİ	230
ORTALAMA	324
BAŞARISIZ	88

Verilere ait istatistiki bilgiler Çizelge 3.3.'de gösterilmiştir.

Çizelge 3.3. Veri tabanı istatistikleri

Niteliği	Özelliği	Kişi Sayısı
Cinsiyeti	Kız	342
	Erkek	300
Sınıfı	Yüksekokul	171
	Fakülte	471
Sonuç	Çok İyi	230
	Ortalama	324
	Başarısız	88
Dönem	Güz	360
	Bahar	282
Öğretim Tipi	Yüz Yüze	312
	Uzaktan	330

Verileri WEKA programında kullanabilmek için öncelikle Attribute- Relation File Format (ARFF) formatına dönüştürülmesi gerekmektedir. ARFF dosyaları, değişken tanımlamasına izin veren ASCII metin dosyalarıdır. ARFF dosyasının başlık kısmında, değişkenler(veri tabanındaki her bir kolonun ismi), bunlar arasındaki ilişkiler ve her bir değişkenin türü ve alacağı değerler bulunmaktadır.

ARFF dosyasını oluşturmak için Microsoft Excel 2010 programı kullanılmıştır. Çalışmada kullanılan verilere ilişkin tanımlanan değişkenler ve tipleri Şekil 3.1.'de gösterilmiştir.

```
@RELATION KKU_BASARI
@ATTRIBUTE SINIF {FAKULTE,YOKUL}
@ATTRIBUTE CINSIYET {KIZ,ERKEK}
@ATTRIBUTE DONEM {GUZ,BAHAR}
@ATTRIBUTE OGRTIP {YUZYUZE,UZAKTAN}
@ATTRIBUTE PTUR {SOZEL,SAYISAL,OYETENEK,YDIL,SINAVSIZ,EA}
@ATTRIBUTE YAS {18_19_YAS,20_21_YAS,21>YAS}
@ATTRIBUTE SONUC {COKIYI,ORTALAMA,BASARISIZ}
```

Şekil 3.1. Çalışmada oluşturulan ARFF dosyasının başlık kısmı

Verilerin bulunduğu kısım @DATA satırından sonra gelir. Şekil 3.2.de verilerin bulunduğu bölüm gösterilmektedir.

```
@DATA
YOKUL,KIZ,GUZ,UZAKTAN,SINAVSIZ,18_19_YAS,COKIYI
FAKULTE,ERKEK,GUZ,YUZYUZE,SAYISAL,20_21_YAS,BASARISIZ
FAKULTE,KIZ,BAHAR,YUZYUZE,YDIL,21>YAS,ORTALAMA
FAKULTE,KIZ,BAHAR,UZAKTAN,EA,21>YAS,ORTALAMA
FAKULTE,KIZ,BAHAR,UZAKTAN,SAYISAL,20_21_YAS,COKIYI
FAKULTE,KIZ,BAHAR,UZAKTAN,SAYISAL,20_21_YAS,COKIYI
FAKULTE,KIZ,BAHAR,UZAKTAN,EA,20_21_YAS,COKIYI
YOKUL,KIZ,GUZ,UZAKTAN,SINAVSIZ,20_21_YAS,COKIYI
FAKULTE,ERKEK,BAHAR,YUZYUZE,YDIL,20_21_YAS,BASARISIZ
FAKULTE,KIZ,GUZ,YUZYUZE,OYETENEK,20_21_YAS,BASARISIZ
FAKULTE,KIZ,BAHAR,YUZYUZE,SAYISAL,20_21_YAS,COKIYI
YOKUL,KIZ,GUZ,UZAKTAN,SINAVSIZ,18_19_YAS,ORTALAMA
FAKULTE,ERKEK,GUZ,YUZYUZE,SAYISAL,18_19_YAS,BASARISIZ
FAKULTE,KIZ,BAHAR,UZAKTAN,EA,20_21_YAS,BASARISIZ
YOKUL,ERKEK,GUZ,YUZYUZE,SOZEL,20_21_YAS,ORTALAMA
FAKULTE,ERKEK,BAHAR,UZAKTAN,EA,21>YAS,COKIYI
FAKULTE,ERKEK,GUZ,YUZYUZE,SAYISAL,20_21_YAS,ORTALAMA
YOKUL,KIZ,GUZ,UZAKTAN,SINAVSIZ,18_19_YAS,ORTALAMA
FAKULTE,ERKEK,BAHAR,YUZYUZE,SAYISAL,20_21_YAS,ORTALAMA
FAKULTE,ERKEK,BAHAR,UZAKTAN,SAYISAL,20_21_YAS,COKIYI
FAKULTE,ERKEK,GUZ,YUZYUZE,SAYISAL,18_19_YAS,ORTALAMA
FAKULTE,ERKEK,BAHAR,YUZYUZE,YDIL,20_21_YAS,ORTALAMA
FAKULTE,KIZ,BAHAR,UZAKTAN,EA,20_21_YAS,COKIYI
```

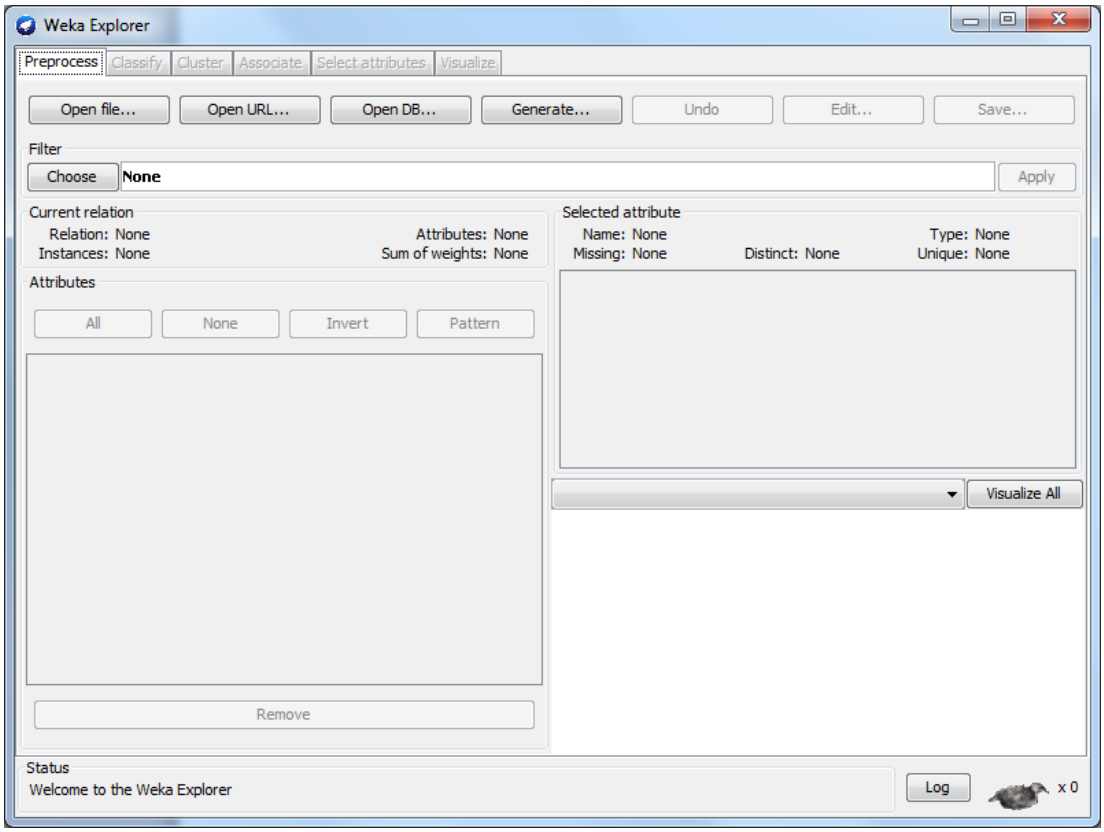
Şekil 3.2. Çalışmada oluşturulan ARFF dosyasında verilerin bulunduğu bölüm

3.2. Modelin Kurulması

ARFF formatındaki veri dosyası oluşturulduktan sonra WEKA programı çalıştırılmıştır. Şekil 3.3. de WEKA Programının Ara yüz ekranı görülmektedir.

Şekil 3.3. WEKA ara yüz görünümü

WEKA ara yüzünde bulunan menüde Explorer butonuna basılarak Explorer penceresi açılmıştır. Verileri ön işlemden geçirmek için Preprocess sayfasında Open File seçeneği yardımı ile oluşturmuş olduğumuz ARFF dosyası açılmıştır. Sekil 3.4.'de WEKA Explorer Ara Yüzü görülmektedir.



Şekil 3.4. WEKA Explorer ara yüzü

Model oluşturulurken verilerin bir kısmı eğitilmiş (training), diğer kısımda eğitilen verilerden oluşan örüntüler kullanılarak test edilmiştir. Bu işlemler WEKA sınıflandırma algoritmaları kullanılarak yapılmıştır. Verileri işlemede tahmin edilecek alan için doğruluk performansı yüksek olan algoritmalar ele alınmıştır.

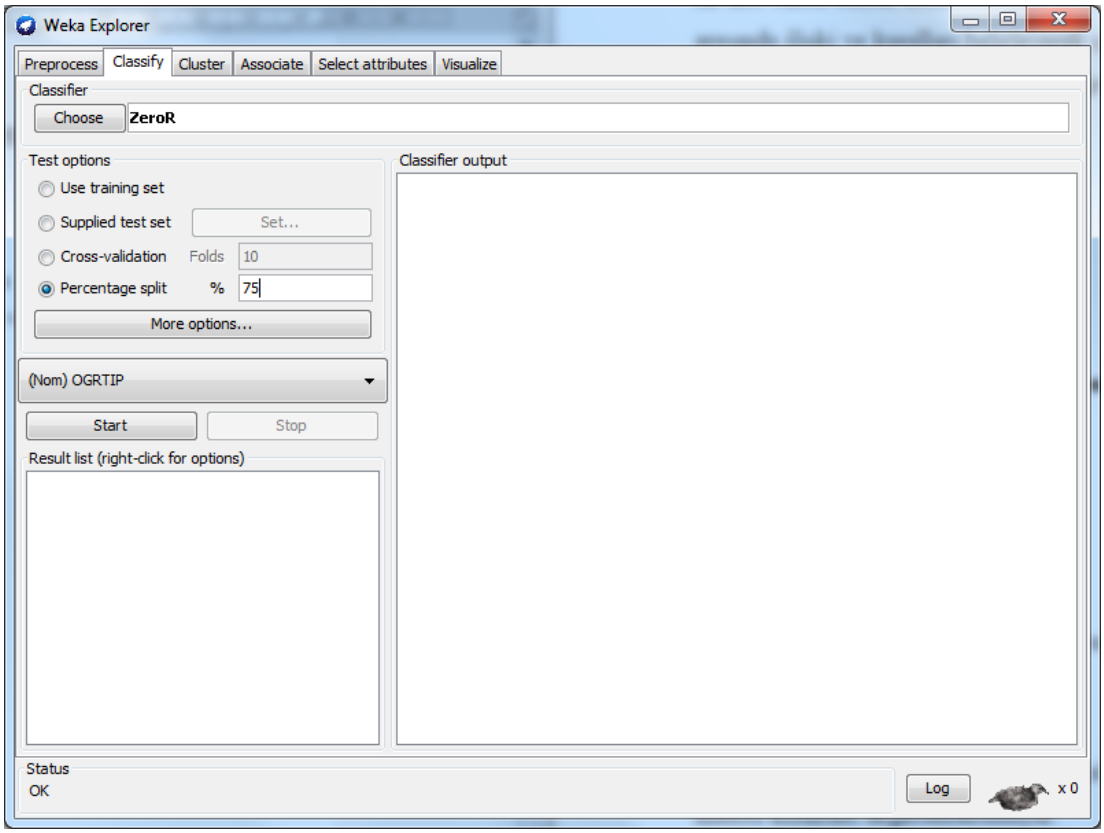
Sınıflandırma algoritmalarının performanslarını test etmek için WEKA'da ki sınıflandırma algoritmalarından J48 Algoritması, Çok Katmanlı Algılayıcı (Multilayer Perceptron) Algoritması ve JRip Algoritması kullanılmıştır.

Çalışmada WEKA 3.7 sürümü kullanılmıştır. Veriler tüm sınıflandırıcılar ile test edilmiştir. Bu işlem yapılırken yüzde ayırma (percentage split) yöntemi kullanılmıştır.

Verilerin %79'u yüzde ayırma yöntemi ile eğitim (training) için, diğer %21'i de test verisi olarak kullanılmıştır. Çeşitli algoritmalarla test verisi test edilerek algoritmaların doğruluğu belirlenmiştir.

3.3. Modelin Değerlendirilmesi

WEKA Programında Explorer penceresinde "Classify" sayfasında programda bulunan çeşitli algoritmalar seçilerek modeller oluşturulmuştur. Şekil 3.5'te WEKA Explorer Penceresinde Classify Sekmesi Ekranı görülmektedir. Uygulamada Doğruluk performansları en yüksek olan modeller ve test sonuçları bu bölümde anlatılacaktır.



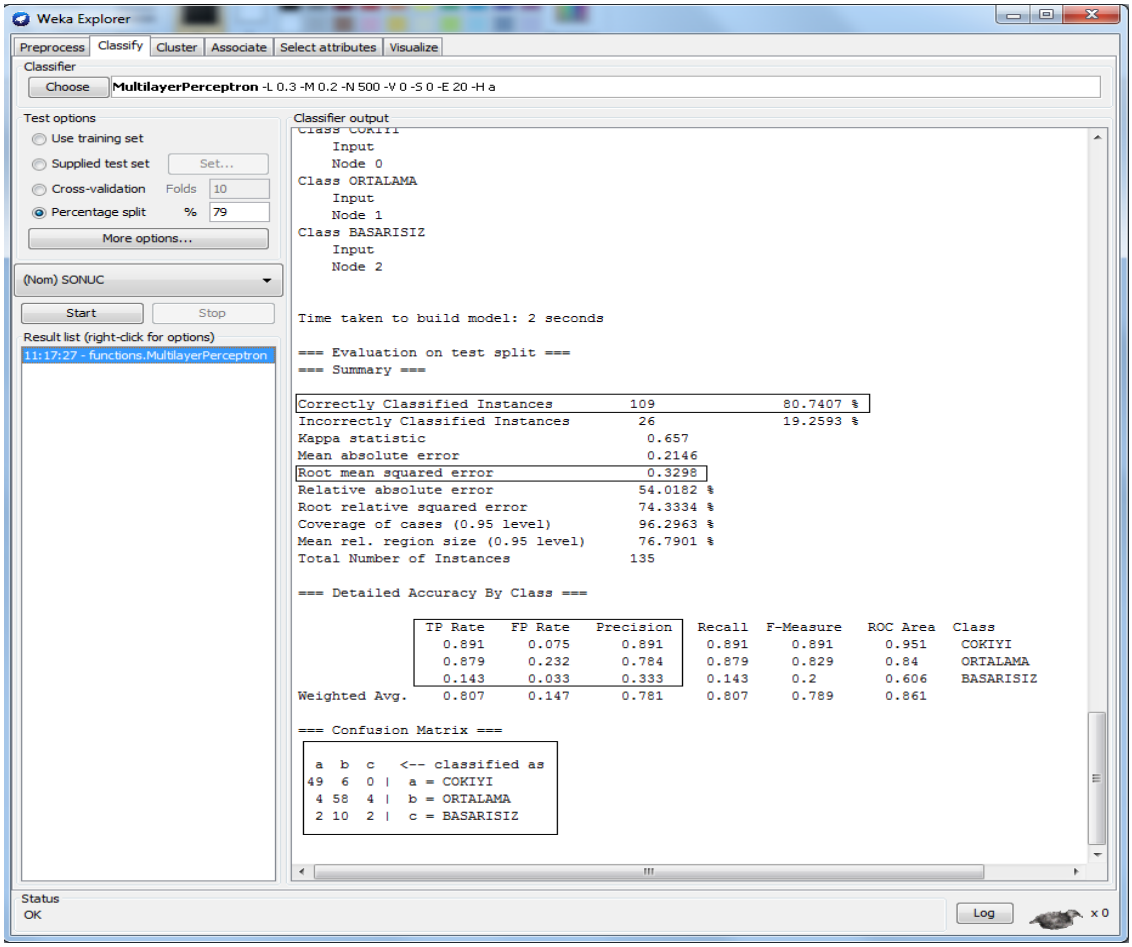
Şekil 3.5. WEKA Explorer penceresinde classify sekmesi ekranı

3.3.1. Multiplayer Perceptron Algoritması İle Oluşturulan Veri Modellemesi

Şekil 3.6.'da Uygulamada Multilayer Perceptron algoritması ile oluşturulmuş veri modellemesi ve sonuç ekranı görülmektedir.

Şekil 3.6.'da ki çıkış ekranı incelendiğinde;

642 kaydın %79'u olan 507 adet kayıt modelin eğitimi için kullanılmıştır. Geriye kalan 135 adet kayıt ise test amaçlı kullanılmıştır. Çıkış ekranında görülen Düzensizlik Matrisine göre (Confusion Matrix) 135 adet kayıta ait düzensizlik matrisi Çizelge 3.4.'de görüldüğü gibidir.



Şekil 3.6. Multilayer Perceptron algoritması ile oluşturulmuş modelin sonuç ekranı

Çizelge 3.4. Multilayer Perceptron algoritması için düzensizlik matrisi

a	b	c	
49	6	0	a=COKIYI
4	58	4	b=ORTALAMA
2	10	2	c=BASARISIZ

Çizelge 3.4.'e göre;
55 adet başarı durumu "COKIYI" değerli test verisinin 49 tanesi "COKIYI", 6 tanesi "ORTALAMA",

66 adet başarı durumu "ORTALAMA" değerli test verisinin 4 TANESİ "COKIYI" 58 tanesini "ORTALAMA" ve 4 tanesi "BASARISIZ",

14 adet başarı durumu "BASARISIZ" test verisinin 2 tanesini "COKIYI", 10 tanesini "ORTALAMA", 2 tanesini de "BASARISIZ" olarak tahmin edilmiştir.

Bu durumda 6+4+4+2+10= 26 adet veri Multilayer Perceptron Algoritması ile oluşturulan modele göre yanlış, 49+58+2=109 adet veri doğru sınıflandırılmıştır. Buna göre modelin doğruluk sınıflandırma yüzdesi aşağıdaki gibi hesaplanmıştır;

Doğruluk Yüzdesi= (109/135)*100=%80,74

Şekil 3.6.'da sınıflandırıcı çıkış ekranında görülen sınıflara göre 135 adet verinin detaylı olarak doğruluk tablosu Çizelge 3.5.'te verilmiştir.

Çizelge 3.5. Detaylı doğruluk tablosu

Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
0,891	0,075	0,891	a=COKIYI
0,879	0,232	0,784	b=ORTALAMA
0,143	0,033	0,333	c=BASARISIZ

Çizelge 3.4.'e göre, Çizelge 3.5.'de ki değerler aşağıdaki gibi hesaplanmıştır:

Duyarlılık (Doğru Pozitif Oranı)
a sınıfı için 49/55 = 0,891
b sınıfı için 58/66 = 0,879
c sınıfı için 2/14 = 0,143

Yanlış Hassaslık Oranı (Yanlış Pozitif Oranı)
a sınıfı için 6 / 80 = 0,075
b sınıfı için 16 / 69 = 0,074
c sınıfı için 4/121 =0,033

Hassaslık (Precision)
a sınıfı için 49 / 55 = 0,891
b sınıfı için 58 / 74 = 0,784
c sınıfı için 2 / 6 = 0,333

3.3.2. JRip Algoritması İle Oluşturulan Veri Modellemesi

Çizelge 3.6.'da JRip Algoritması için düzensizlik matrisi verilmiştir.

Çizelge 3.6. JRip Algoritması için düzensizlik matrisi

a	b	c	
49	6	0	a=COKIYI
3	60	3	b=ORTALAMA
2	11	1	c=BASARISIZ

Çizelge 3.6.'ya göre ;

55 adet başarı durumu "COKIYI" değerli test verisinin 49 tanesi "COKIYI", 6 tanesi "ORTALAMA",

66 adet başarı durumu "ORTALAMA" değerli test verisinin 3 TANESİ "COKIYI" 60 tanesini "ORTALAMA" ve 3 tanesi "BASARISIZ",

14 adet başarı durumu "BASARISIZ" test verisinin 2 tanesini "COKIYI", 11 tanesini "ORTALAMA", 1 tanesini de "BASARISIZ" olarak tahmin edilmiştir.
Çizelge 3.6.'da verilen Düzensizlik matrisine göre 135 adet verinin detaylandırılmış doğruluk tablosu aşağıda Çizelge 3.7.'de verilmiştir.

Çizelge 3.7. JRip Algoritması için detaylı doğruluk tablosu

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
81,4815%	0,891	0,063	0,907	a=COKIYI
	0,909	0,246	0,779	b=ORTALAMA
	0,071	0,025	0,25	C=BASARISIZ

Çizelge 3.7.'ye göre, JRip Algoritması için doğruluk yüzdesi, %81,4815 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, COKIYI, ORTALAMA VE BASARISIZ sınıfları için ayrı ayrı hesaplanmıştır.

3.3.3. J48 Algoritması İle Oluşturulan Veri Modellemesi

Çizelge 3.8'de J48 Algoritması için düzensizlik matrisi verilmiştir.

Çizelge 3.8. J48 Algoritması için düzensizlik matrisi

a	b	c	
49	6	0	a=COKIYI
4	62	0	b=ORTALAMA
2	12	0	c=BASARISIZ

Çizelge 3.8.'e göre;

55 adet başarı durumu "COKIYI" değerli test verisinin 49 tanesi "COKIYI", 6 tanesi "ORTALAMA",

66 adet başarı durumu "ORTALAMA" değerli test verisinin 4 TANESİ "COKIYI" 62 tanesini "ORTALAMA"

14 adet başarı durumu "BASARISIZ" test verisinin 2 tanesini "COKIYI", 12 tanesini "ORTALAMA" olarak tahmin edilmiştir.

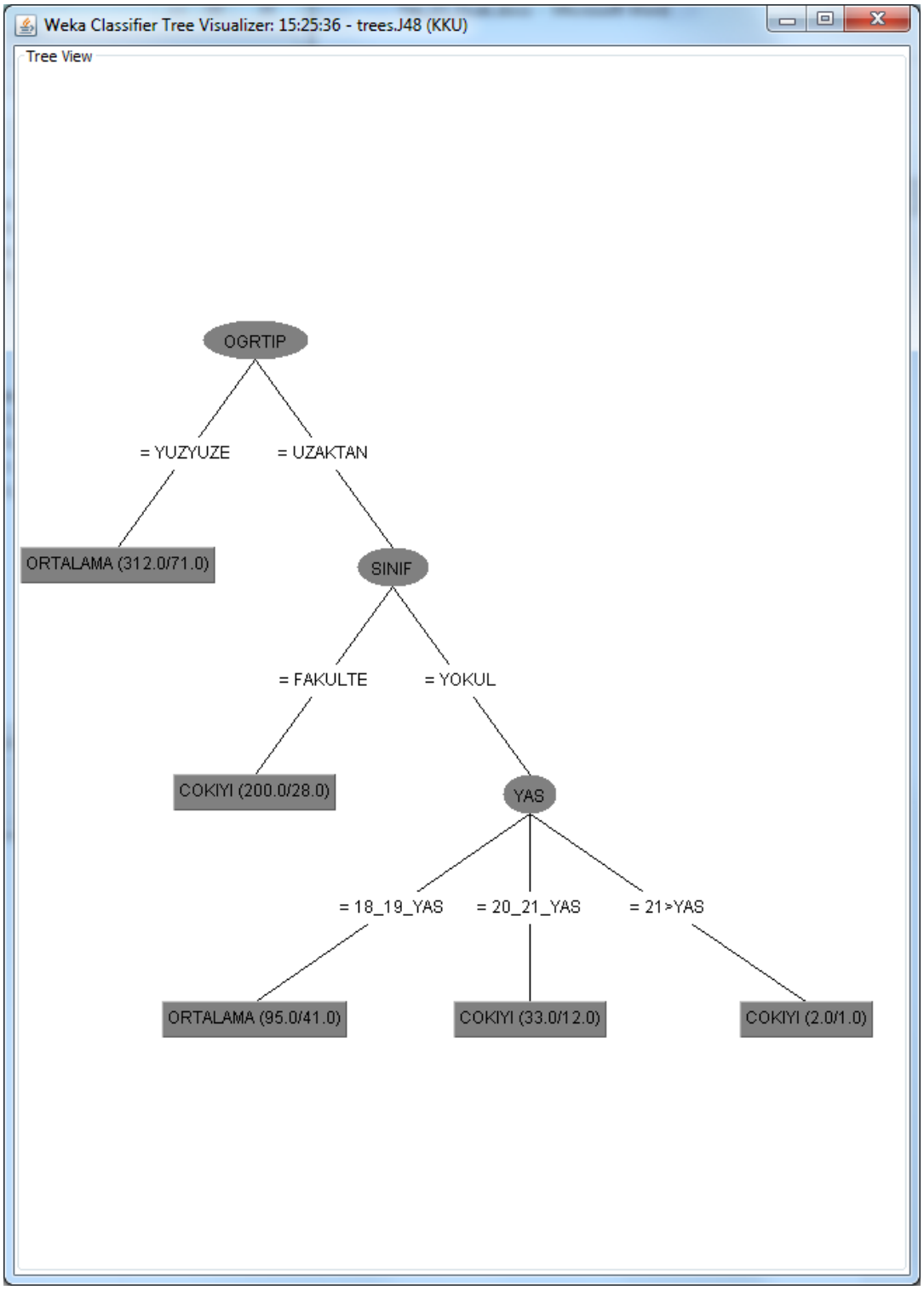
Çizelge 3.8.'de verilen düzensizlik matrisine göre 135 adet verinin detaylandırılmış doğruluk tablosu aşağıda Çizelge 3.9'da verilmiştir.

Çizelge 3.9. J48 Algoritması için detaylı doğruluk tablosu

Doğruluk Yüzdesi	Duyarlılık (DP Oranı)	Yanlış Hassaslık Oranı (YP Oranı)	Hassaslık (Precision)	Sınıf (Class)
82,2222%	0,891	0,075	0,891	a=COKIYI
	0,939	0,261	0,775	b=ORTALAMA
	0	0	0	C=BASARISIZ

Çizelge 3.9.'a göre, J48 Algoritması için doğruluk yüzdesi, %82,222 olarak hesaplanmıştır. Duyarlılık, Yanlış Hassaslık Oranı ve Hassaslık değerleri, COKIYI, ORTALAMA VE BASARISIZ sınıfları için ayrı ayrı hesaplanmıştır.

Şekil 3.7.'de J48 Algoritması için karar ağacı sonuç ekranı görülmektedir.



Şekil 3.7. J48 Algoritması için karar ağacı sonuç ekranı

Her satır, ağaçtaki bir düğümü, alt satırlar, ilk satırın çocuk düğümlerini; düğümlerde parantezin içindeki ilk sayı veri kümesindeki kaç durumun bu düğüm için doğru olarak sınıflandırıldığını; eğer varsa; parantezin içindeki ikinci sayı, düğüm tarafından yanlış olarak sınıflandırılan durumların sayısını göstermektedir.

Şekil 3.7.'de görüldüğü gibi kademeler en üstten aşağıya doğru öğretim tipi, öğrencinin birimi, öğrencinin yaşı şeklindedir.

J48 karar ağacı algoritmasının sonuçlarını aşağıdaki şekilde değerlendirmek mümkündür:

1. İlk dallanmada öğretim tipi yüz yüze olan öğrencilerin başarısının ortalama notlarda seyrettiği, uzaktan eğitim olan öğrencilerin başarı eğiliminin ise fakülte veya yüksekokul öğrencisi olması durumuna göre farklılık gösterdiği görülmektedir.
2. İkinci dallanmada uzaktan eğitim ile eğitim alan fakülte öğrencilerinin başarı notlarının daha yüksek olduğu, yüksekokulda eğitim gören öğrencilerin ise başarı notlarının öğrencinin yaşı ile bağlantılı olduğu görülmektedir.
3. Üçüncü dallanmada ise öğrencinin yaş kriterinin başarıda etkili olduğu, 20 yaş ve üzeri yaştaki öğrencilerin çok daha başarılı olduğu görülmektedir.

WEKA Programı ile veriler üzerinde çeşitli algoritmalar uygulanmış ve doğruluk yüzdeleri ayrı ayrı bulunarak sonuçlar Çizelge 3.10.'da gösterilmiştir.

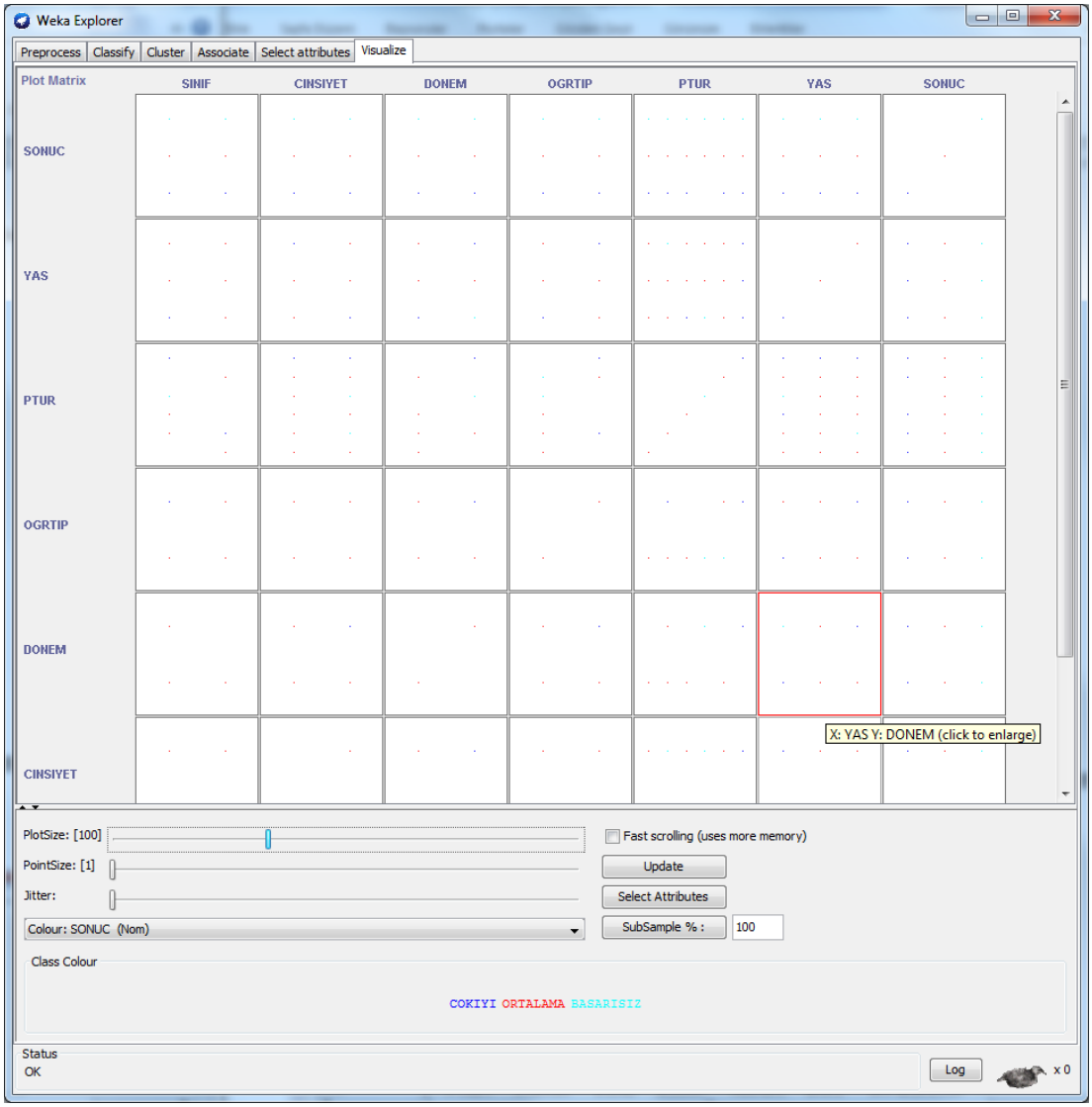
Çizelge 3.10. Seçilen sınıflandırma algoritmaları ve doğruluk yüzdeleri

Algoritmalar	Doğru Sınıflandırılan Örnek Sayısı	Yanlış Sınıflandırılan Örnek Sayısı	Doğruluk Yüzdesi
J48	111	24	82,2222
JRip	110	25	81,4815
Multilayer Perceptron	109	26	80,7407

3.3.4. WEKA Programı İle Elde Edilen Görsel Sonuçlar

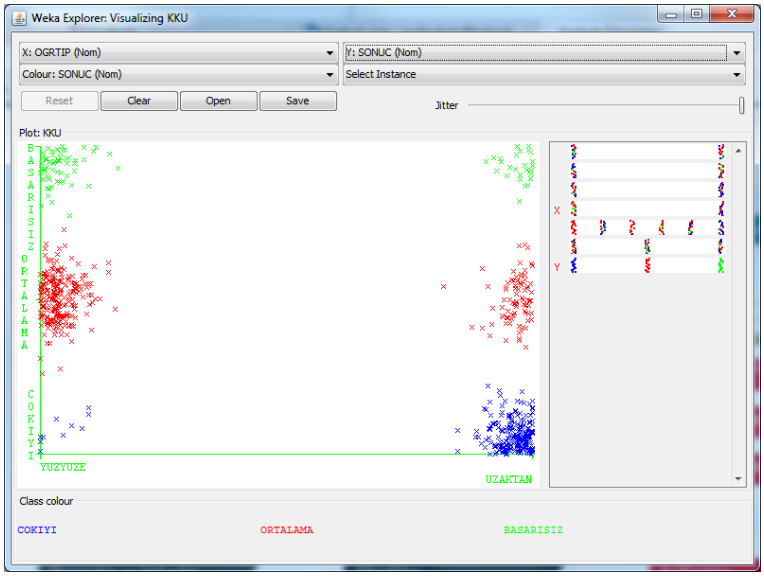
WEKA programı görsel olarak da sonuçlar üretebilmektedir. WEKA programı ile elde edilen görsel sonuçlar bu bölümde incelenecektir.

Şekil 3.8.'de görüldüğü gibi WEKA programı görselleştir (visualize) sayfasında farklı veri alanları özelliklerine göre üç boyutlu olarak analiz edilmesine olanak sağlamaktadır.



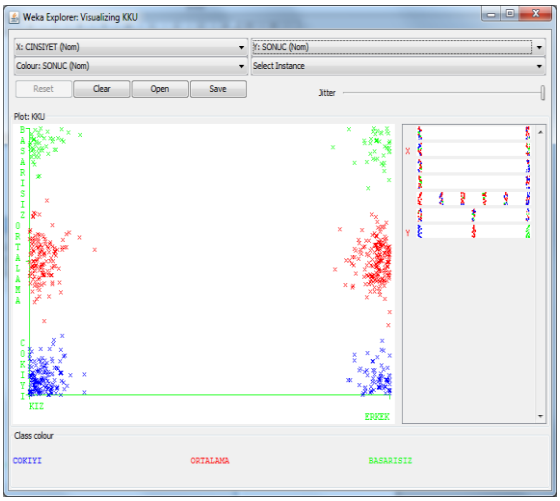
Şekil 3.8. WEKA programı grafiksel tahmin aracı

Şekil 3.9.'de TBTK dersini alan öğrencilerin eğitim tipleri ile başarı durumları arasındaki ilişkisel grafik raporlanmıştır. Bu rapora göre uzaktan eğitim yöntemi ile TBTK dersini alan öğrencilerin başarı durumlarının çok iyi, yüz yüze eğitim yöntemi ile alan öğrencilerin ise daha ortalama değerler olduğu görülmektedir.



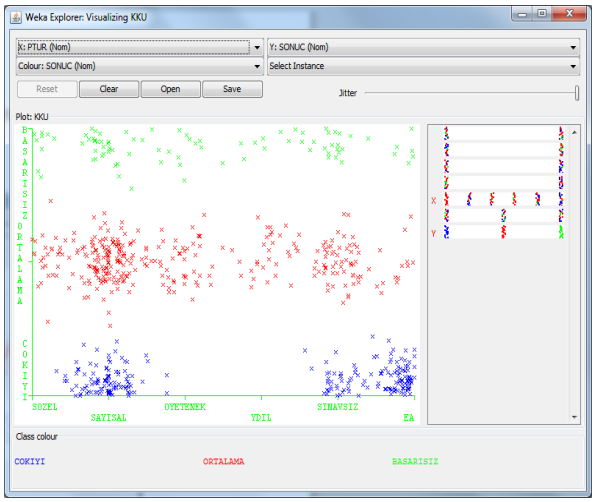
Şekil 3.9. Eğitim tiplerine göre başarı durumun dağılımı

Sekil 3.10.'da TBTK dersini alan öğrencilerin cinsiyetleri ile başarı durumları arasındaki ilişkisel grafik raporlanmıştır. Bu rapora göre kız öğrencilerin başarı durumlarının daha iyi olduğu görülmektedir.



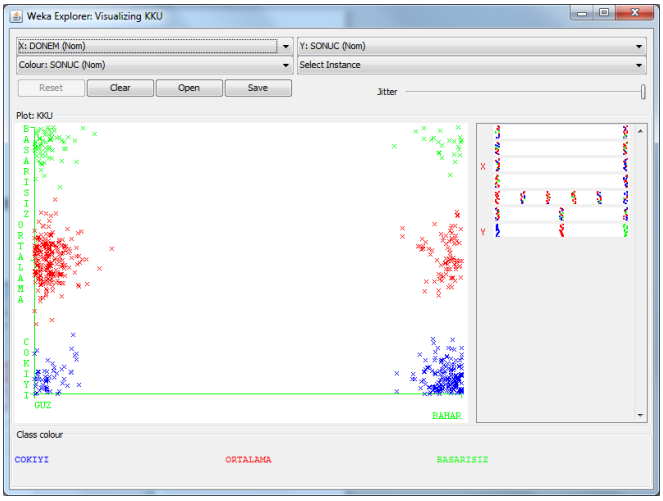
Şekil 3.10. Cinsiyetlere göre başarı durumlarının dağılımı

Şekil 3.11.'de TBTK dersini alan öğrencilerin üniversiteye yerleştirmede esas alınan puan türü ile başarı durumları arasındaki ilişkisel grafik raporlanmıştır. Bu rapora göre puan türü; sayısal, eşit ağırlık ve sınavsız geçişle ile yerleştirilen öğrencilerin daha başarılı oldukları görülmektedir. Puan türleri sözel, özel yetenek, yabancı dil olan öğrencilerin ortalamada kaldıkları ya da ortalamanın altına düştükleri görülmektedir.



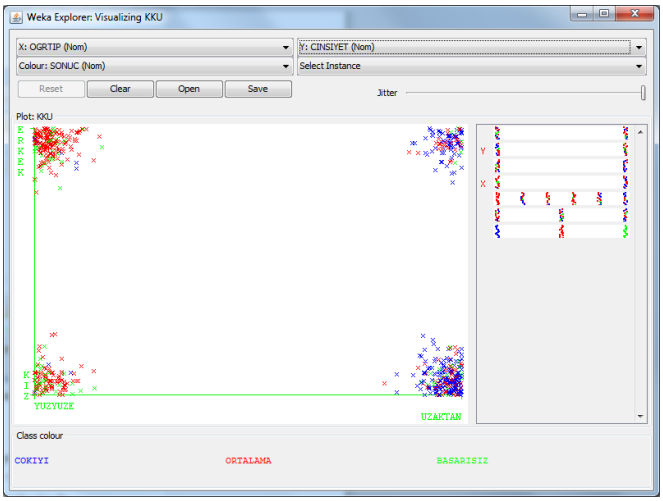
Şekil 3.11. Yerleştirme puan türüne göre notların dağılımı

Şekil 3.12.'de TBTK dersini alan öğrencilerin bu dersi aldıkları dönem ile başarı durumları arasındaki ilişkisel grafik raporlanmıştır. Bu rapora göre TBTK dersinin Bahar döneminde alınması durumunda başarının daha iyi olduğu güz döneminde ise daha ortalama sonuçlar alındığı görülmektedir.



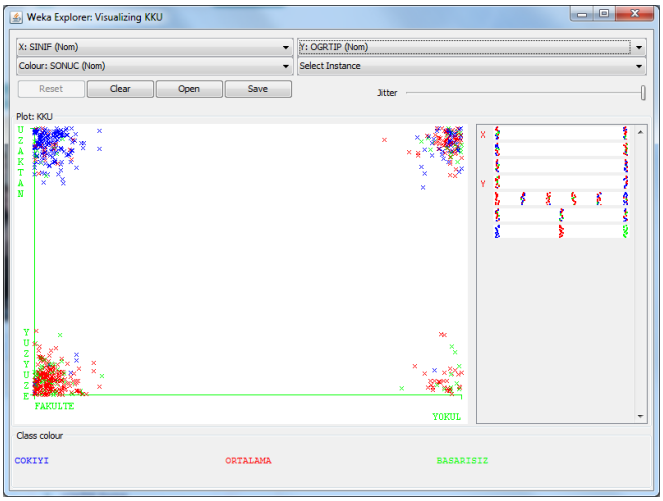
Şekil 3.12. Dersin alındığı döneme göre notların dağılımı

Şekil 3.13.'de TBTK dersini alan öğrencilerde cinsiyeti "Kız" olan öğrencilerin "uzaktan eğitim" ile alanların başarı durumlarının daha iyi olduğu açıkça görülmektedir. Genel olarak bakıldığında dersi yüz yüze eğitim ile alan öğrencilerin notlarının ortalama da kaldığı görülmektedir.



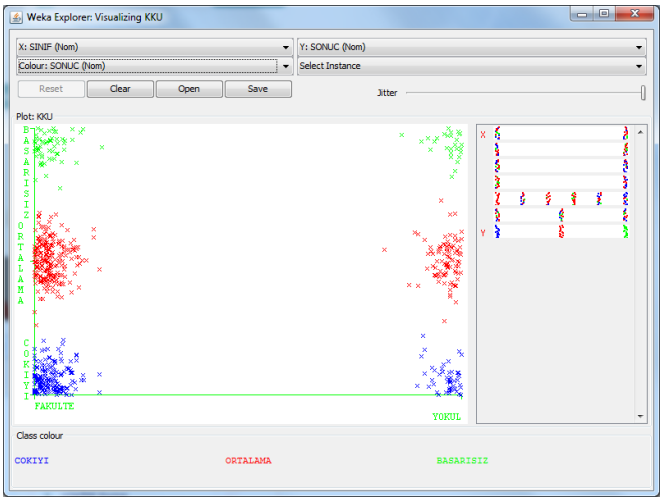
Şekil 3.13. Öğrenci cinsiyetleri ile eğitim tipleri arasındaki ilişkiye göre başarı

Şekil 3.14.'de TBTK dersini alan fakültede okuyan öğrencilerin uzaktan eğitim ile ders almaları durumunda başarı durumlarının çok daha iyi olduğu görülmektedir.



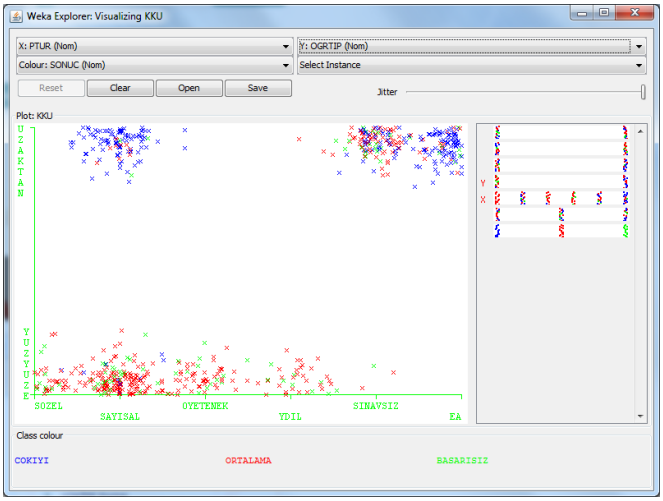
Şekil 3.14. Fakülte ve yüksekokul programları ile eğitim tipleri arasındaki ilişki

Şekil 3.15.'de TBTK dersini alan öğrencilerden fakültelerde eğitim alanların daha başarılı oldukları görülmektedir.



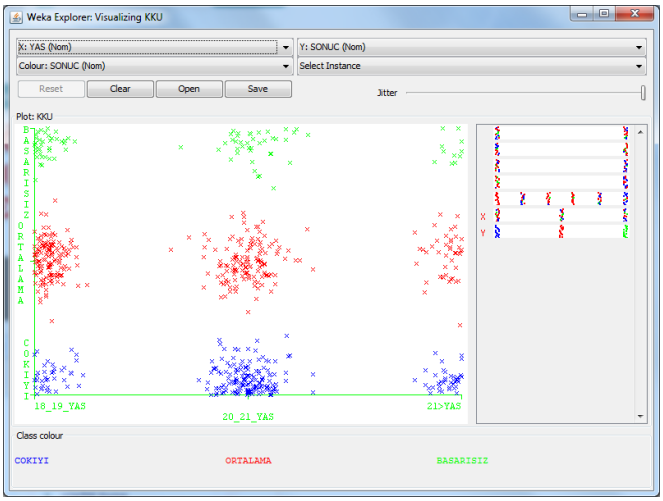
Şekil 3.15. Yüksekokul ve fakültelere göre notların dağılımı

Şekil 3.16. incelendiğinde üniversiteye sayısal puanla yerleştirile öğrencilerden dersi uzaktan eğitim yöntemi ile alan öğrencilerin daha başarılı olduğu görülmektedir.



Şekil 3.16. Yerleştirmede esas puan türleri ile eğitim tipi arasındaki ilişki

Şekil 3.17'de TBTK dersini alan öğrencilerden 20-21 yaşlarında olan öğrencilerin başarı durumlarının daha çok "çok iyi" sınıfında, 18-19 yaşında olan öğrencilerin ise "ortalama" sınıfında oldukları görülmektedir.



Şekil 3.17. Öğrencilerin yaşları ile başarı durumları arasındaki ilişki

4. TARTIŞMA VE SONUÇ

Günümüzde birçok kurum bilişim teknolojileri sayesinde depoladıkları verilerden anlamlı bilgiler elde etmek gayesindedir. Eğitim kurumları da kullanıcılara daha iyi hizmet verebilmek amacı ile bilişim teknolojilerine çok değer vermekte ve aktif olarak her aşamada kullanmaktadır. Artık üniversiteler veri tabanlarında saklanan verilerden kendi kurumlarına faydalı olabilecek bilgilere ulaşabileceklerin fark etmişlerdir.

Kırıkkale Üniversitesi öğrencilerine ait kimlik ve not bilgileri gibi pek çok bilgi otomasyon yazılımı ile sistemde saklanmaktadır. Bu çalışmada veri madenciliği teknikleri kullanılarak üniversitemiz için anlamlı sonuçlar alınması hedeflenmiştir.

İlk olarak eğitim alanında yapılmış çalışmalarla ile ilgili bir literatür taraması yapılmış daha sonra Kırıkkale Üniversitesi Öğrenci Otomasyonu veri tabanından bilgiler alınarak veri madenciliği uygulaması gerçekleştirilmiştir.

Yapılan literatür taramasında eğitim alanında veri madenciliği teknikleri kullanılarak üniversitelerin faydalı bilgilere ulaşabildiği ve bu yönde eğitim alanında iyileştirmelerin yapılabildiği görülmüştür. Yükseköğretim kurumlarında eğitim gören öğrencilerin başarısını etkileyen faktörler tespit edilerek, öğrenci başarısını nasıl arttırabiliriz sorusuna yanıt aranmıştır.

İkinci aşamada veri madenciliği üzerinde durulmuştur. Veri tabanlarından içinde gizli ve doğrudan erişilemeyen veriler mevcuttur. Faydalı bilgilere ulaşmada veriler içinden SQL sorgulama yapılması veya raporlama araçlarının kullanılması artık yetersiz kalmaktadır. Bu durum kişilerin veri madenciliği tekniklerine yönelmesine sebep olmuştur. Güncelliğini koruyan veri madenciliği her geçen gün daha da gelişerek hizmet vermeye devam etmektedir.

Tezin son bölümünde ise Kırıkkale Üniversitesi Öğrenci Bilgi Otomasyonu veri tabanından alına bilgiler çeşitli veri madenciliği yöntemleri ile incelenmiş ve

karşılaştırma yapılmıştır. Çalışmada ücretsiz bir veri madenciliği yazılımı olan WEKA programı kullanılmıştır. En iyi sonucu veren algoritmalar tespit edilmiştir. Bu çalışmada, yapılan karşılaştırma sonucunda, %82,2222'lik doğruluk yüzdesi ile C4.5 algoritmasının WEKA implementasyonu olan J48 karar ağacı algoritmasının, diğer algoritmalara göre daha başarılı olduğu Çizelge 4.1.'de görülmektedir [40] .

Çizelge 4.1 Seçilen sınıflandırma algoritmaları ve doğruluk yüzdeleri

Algoritmalar	Doğruluk Yüzdesi
J48	82,2222
JRip	81,4815
Multilayer Perceptron	80,7407

Sonuç olarak Veri Madenciliği yöntemlerinin eğitim alanında da başarılı sonuçlar verdiği görülmüştür. Kurum yöneticilerine geçmişle ilgili anlamlı bilgiler vererek gelecek ile ilgili kararlar alınması konusunda yardımcı olacağı tespit edilmiştir.

Öğrenci profillerine göre hangi öğrencilerin dersi uzaktan eğitim ya da yüz yüze eğitim yoluyla alması durumunda başarının daha yüksek olacağı belirlenmiştir. Öğrenci performansının etkileyen başlıca faktörlerin; eğitim türü, fakülte ya da yüksekokul öğrencisi olması, öğrencinin yaşı olduğu görülmüştür.

Genel olarak sonuçlar değerlendirildiğinde üniversitemizde TBTK dersini alan öğrencilerin başarı durumları incelendiğinde;

- Kız öğrencilerin,
- Bölüme yerleştirmede esas alınan puan türü sayısal, eşit ağırlık ve sınavsız geçiş olan öğrencilerin,
- Dersi bahar döneminde alan öğrencilerin
- Fakülte öğrencilerinin,

- Ders aldığı eğitim öğretim yılında 20-21 yaş aralığında olan öğrencilerin daha başarılı oldukları tespit edilmiştir.

Öğretim tipin öğrenci başarısı üzerine etkisi incelendiğinde;

- Kız öğrencilerin,
- Fakülte öğrencilerinin,
- Sayısal tabanlı öğrencilerin

Uzaktan eğitim ile TBTK dersini aldıklarında daha başarılı oldukları tespit edilmiştir.

Bu durumda TBTK dersinde öğrenci performansını arttırmak amacı ile üniversitemizde fakülte öğrencilerine uzaktan eğitim ile verilmesinin daha uygun olduğu söylenebilir. Aynı şekilde öğrenci başarısını artırmak amacı ile sayısal tabanlı öğrencilere TBTK dersinin uzaktan eğitim ile verilmesi daha uygun olacaktır.

KAYNAKLAR

[1] İ. Göksu, Web Tabanlı Öğrenme Ortamında Veri Madenciliğine Dayalı Öğrenci Değerlendirmesi. Yüksek Lisans Tezi, Fırat Üniversitesi, Elazığ, 2012.

[2] C. Romero ve S. Ventura, Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications,33:135-146, 2007.

[3] A. P. Sanjeev ve J. M. Zytkow, Discovering Enrollment Knowledge in University Database,KDD-95 Proceedings, aaai.org, 1995.

[4] J. Luan, Data Mining and Knowledge Management in Higher Education Potential Applications,Workshop Associate of institutional Research International Conference, Toronto, 2002.

[5] M. Karabatak ve M. C. İnce, Apriori Algoritması ile Öğrenci Başarısının Analizi, ELOCO International Conference, Bursa, 2004.

[6] Ş. Z. Erdoğan ve M. Timor, A Data Mining Application in a Student Database, Journale of Aeronautics and Space Technologies, 2(2) : 53-57, 2005.

[7] Y. Z. Ayık, A. Özdemir ve U. Yavuz, Lise Türü ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi, Sosyal Bilimler Enstitüsü Dergisi, 10(2): 441-454, 2007.

[8] Y. Ünal, U. Ekim ve M. Köklü, Üniversite Öğrencilerinin Ortak Zorunlu Derslerdeki Başarılarının K-Means Algoritması İle İncelenmesi, e-Journal of New World Sciences Academy Engineering Sciences, 6(1): 342-347, 2011.

[9] M. A. Alan, Veri Madenciliği ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama, Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, 33: 165-174,

2012.

[10] B. Şen ve E. Uçar, Evaluating the achievements of computer engineering department of distance education students with data mining methods, Procedia Technology, 1(2012): 262-267, 2012.

[11] M. Dener, M. Dörterler ve A. Orman, Açık Kaynak Kodlu Veri Madenciliği Programları: WEKA'da Örnek Uygulama, XI. Akademik Bilişim Konferansı Bildirileri, Şanlıurfa, s. 787-796, 2009.

[12] E. Alpaydın, Zeki Veri Madenciliği: Ham Veriden Altın Veriye Ulaşma Yöntemleri, Bilişim Eğitim Semineri, 2000.

[13] D. T. Larose, Data Mining Methods and Models, Wiley, 2006.

[14] H. Sever ve B. Oğuz, Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları ve Algoritmalar, Information World, 1(3): 173-204, 2002.

[15] Two Crows Corporation, http://www.twocrows.com/intro-dm.pdf (Erişim tarihi: 21 Mart 2014)

[16] A. Vahaplar ve M. M. İnceoğlu, Veri Madenciliği ve Elektronik Ticaret, Türkiye'de İnternet Konferansları, Harbiye, İstanbul (2001): 1-3, 2001.

[17] J. Han ve M. Kamber, Data Mining Concepts and Techniques. 29. Ed: Morgan Kaufmann, Multiscience Press San Francisco, 2006.

[18] Ş. Özmen, İş Hayatı Veri Madenciliği İle İstatistik Uygulamaları,V.Ulusal Ekonometri ve İstatistik Sempozyumu, Adana, 2001.

[19] A. S. Koyuncugil ve N. Özgülbaş, Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları, Bilişim Teknolojileri Dergisi, 2 (2): 21-32, 2009.

[20] H. Kaya ve K. Köymen, Veri Madenciliği Kavramı ve Uygulama Alanları, Doğu Anadolu Bölgesi Araştırmaları Dergisi, 6(2): 159-164, 2008.

[21] U. T. G. Şimşek, Veri Madenciliği ve Bilgi Keşfi, Pegem Akademi, Ankara, 2009.

[22] S. Aydın, Veri Madenciliği ve Anadolu Üniversitesi Uzaktan Eğitim Sisteminde Bir Uygulama, Doktora Tezi. Anadolu Üniversitesi, Eskişehir, 2007.

[23] S. ÖZEKES, Veri Madenciliği Modelleri ve Uygulama Alanları,İstanbul Ticaret Üniversitesi Dergisi, 3: 65-82, 2003.

[24] P. Tapkan, L. Özbakır ve A. Baykasoğlu, WEKA İle Veri Madencilği Süreci ve Örnek Uygulama, Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi, İzmir, 2011.

[25] L. Padua, H. Schulze, K. Matkovic ve C. Delrieux, Interactive Exploration of Parameter Space in Data Mining: Comprehending The Predictive Quality of Large Decision Tree Collections, Computers&Graphics, 41 (2014): 99-113, 2014.

[26] Y. Kökver, Veri Medenciliğinin Nefroloji Alanında Uygulanması, Yüksek Lisans Tezi. Kırıkkale Üniversitesi, Kırıkkale, 2012.

[27] D. Şengür ve A. Tekin, Öğrencilerin Mezuniyet Notlarının Veri Madenciliği, Bilişim Teknolojileri Dergisi, 6 (3): 7-16, 2013.

[28] E. Acar ve M. S. Özerdem, Laws Doku Enerji Ölçümü Tabanlı k-NN Sınıflandırıcı Modeli ile İris Tanıma Sistemi, Signal Processing and Communications Applications Conference, Kıbrıs, 2013.

[29] S. Albayrak, Sınıflama ve Kümeleme Yöntemleri, https://www.ce.yildiz.edu.tr/personal/songul/file/324/Veri+Madencili%C4%9Fi, Veri Madenciliği-SınıflamaKumeleme (Erişim Tarihi: 07.04.2014)

[30] Y. Argüden ve B. Erşahin, Veri Madenciliği Veriden Bilgiye, Masraftan Değere, ARGE Danışmanlık Yayınları, No: 10, 2008.

[31] Y. İşler ve A. Narin, WEKA Yazılımında k-Ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi,SDU Teknik Bilimler Dergisi, 2 (4): 21-39, 2012.

[32] D. Altaş ve V. Gülpınar, A COMPARISON OF CLASSIFICATION PERFORMANCES OF THE DECISION TREES AND THE ARTIFICIAL NEURAL NETWORKS: EUROPEAN UNION, Trakya Üniversitesi Sosyal Bilimler Dergisi, 14 (1): 1-22, 2012.

[33] Y. Özkan, Veri Madenciliği Yöntemleri, Papatya Yayıncılık, İstanbul, 2008.

[34] T. Tuncer ve Y. Tatar, Karar Ağacı Kullanılarak Saldırı Tespit Sistemlerinin Performans Değerlendirmesi, 4. İletişim Teknolojileri Ulusal Sempozyumu, Adana, 2009.

[35] Ö. Çöllüoğlu Gülen ve S. Özdemir, Analysis of Gifted Students' Interest Areas Using Data Mining Techniques, Journal of Gifted Education Research, 1 (3): 213-226, 2013.

[36] M. Sasaki ve K. Kıta, Rule-Based Text Categorization Using Hierarchical

Categories, 1998 IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, 1998.

[37] B. Alataş ve E. Akın, Sınıflandırma Kurallarının Karınca Koloni Algoritmasıyla Keşfi, ELECO`2004, Bursa, 2004.

[38] A. Uğur ve A. C. Kınacı, Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması, inet-tr'06 - XI. "Türkiye'de İnternet" Konferansı, Ankara, 2006.

[39] U. Orhan, M. Hekim ve M. Özer, Discretization Approach to EEG Signal Classification Using Multilayer Perceptron Neural Network Model, Biomedical Engineering Meeting (BIYOMUT), Antalya, 2010.

[40] S. Özarslan ve N. Barışçı, Öğrenci Performansının Veri Madenciliği İle Belirlenmesi, ISITES2014 , Karabük, 2014.