T.C.
RECEP TAYYİP ERDOĞAN ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
İŞLETME ANA BİLİM DALI

İŞLETME BÖLÜMÜ LİSANS DERSLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ: RECEP TAYYİP ERDOĞAN ÜNİVERSİTESİ ÖRNEĞİ

(Yüksek Lisans Tezi)

Sema PEÇE

Dr. Öğr. Üyesi Burcu KARTAL
Danışman

RİZE
2019

KABUL VE ONAY

Recep Tayyip Erdoğan Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Ana Bilim Dalında, Sema PEÇE tarafından hazırlanan *"İşletme Bölümü Lisans Derslerinin Veri Madenciliği Yöntemleri ile Analizi: Recep Tayyip Erdoğan Üniversitesi Örneği"* başlıklı bu çalışma 08/07/2019 tarihinde yapılan savunma sınavı sonucu oy birliği/oy çokluğuyla başarılı bulunarak jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Başkan: Prof. Dr. Abdullah NARALAN — Kabul/~~Red~~

Üye: Dr. Öğretim Üyesi Burcu KARTAL — Kabul/~~Red~~

Üye: Prof. Dr. Tuba YAKICI AYAN — Kabul/~~Red~~

6./8./.2019

Doç. Dr. Ahmet YANIK
Enstitü Müdürü

ETİK BEYAN

Bu tezdeki bütün bilgileri etik davranış ve akademik kurallar çerçevesinde elde ettiğimi ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yaptığımı bildiririm. İfade ettiklerimin aksi ortaya çıktığında ise her türlü yasal sonucu kabul ettiğimi beyan ederim. 08/07/2019

Sema PEÇE

ÖN SÖZ

Verinin anlamının her geçen gün arttığı günümüzde veriye ulaşma, veriyi ayıklama ve veriden anlamlı bilgiler elde etmek birçok açıdan değer görmektedir. Veri tabanlarında işlevsiz kalan verilerden bilgi kazanımı elde etme yöntemi olan veri madenciliği ise verinin anlamını pekiştirmede kullanılan bir yöntemdir. Pazarlamadan, tıbba, borsaya verinin depolanabildiği birçok alanda analiz imkanı sunan veri madenciliği bu çalışmada eğitim sektörüne uygulanmıştır. Öğrencinin ders ve diğer bilgileriyle elde edilen sonuçlarının, öğrencilere ve öğrencilerle ilgili karar merkezlerine karar vermede etkili sonuçlar bulunduğuna inanılarak bu çalışma yapılmıştır.

Çalışmanın başlangıç aşamasından, geliştirme aşamasına ve sonlandırma aşamasına kadar yanımda olan danışmanım Dr. Öğr. Üyesi Burcu Kartal'a ve fikirleriyle desteğini esirgemeyen Arş Gör. Mehmet Fatih Sert'e teşekkürlerimi sunarım. Verilerin elde edilme sürecinde yardımcı olan Recep Tayyip Erdoğan Üniversitesi Bilgi İşlem Merkezi Daire Başkanı Doç. Dr. Ömer Faruk Ursavaş'a ve daire çalışanlarına, İktisadi İdari Bilimler Fakültesi çalışanları Davut Mert ve Kemal Diler'e ayrıca teşekkürlerimi sunarım.

Son olarak ise her türlü desteğini esirgemeyen babama, anneme, kardeşime ve dedem İbrahim Erdoğan'a teşekkürlerimi sunarım.

Sema PEÇE

Rize 2019

İÇİNDEKİLER

İKİNCİ BÖLÜM

İŞLETME BÖLÜMÜ LİSANS DERSLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

Recep Tayyip Erdoğan Üniversitesi Sosyal Bilimler Enstitüsü
Ana Bilim Dalı: İşletme
Tez Türü: Yüksek Lisans Tezi
Danışman: Dr. Öğr. Üyesi Burcu KARTAL
Hazırlayan: Sema PEÇE
Yıl: 2019
Sayfa: 99

ÖZET

İŞLETME BÖLÜMÜ LİSANS DERSLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ: RECEP TAYYİP ERDOĞAN ÜNİVERSİTESİ ÖRNEĞİ

Teknoloji ve teknolojinin getirdiği yenilikler yaşamın her parçasında yerini hızla almaktadır. Bu yeniliklerin getirdiği külfeti taşımak ise veritabanlarının görevi olmuştur. Veritabanlarında anlamsız bir yük ve maliyet unsuru olan verileri değerlendirmek veri madenciliğinin ilgilenme alanıdır. Verinin depolandığı her alanda yararlanılabilen veri madenciliği bu tez çalışmasında eğitim alanında uygulanmıştır. Çalışmanın amacı lisans derslerinin analiz edilmesidir. Bu amaçla Recep Tayyip Erdoğan Üniversitesi, İktisadi İdari Bilimler Fakültesi, İşletme bölümüne 2009'dan itibaren kaydolan 731 öğrencinin ders ve öğrenci bilgileri incelenmiştir. Derslerin kendi aralarındaki not temelli ilişkileri Birliktelik Kuralları ile incelenmiştir. Ders notlarına göre mezuniyet puanları, Öğrenci Seçme Sınavı puanları, Öğrenci Seçme Sınavı yerleştirme sıralamaları, dersleri ilk alımda geçme durumları, sınıflandırma algoritmalarıyla incelenmiştir. Apriori, Naive Bayes, Sıralı Minimal Optimizasyon ve J48 algoritmaları analizde kullanılan veri madenciliği algoritmalarıdır.

Anahtar Kelimeler: Veri Madenciliği, Eğitim, Sınıflandırma, Birliktelik Kuralları

Recep Tayyip Erdogan University Graduate School of Social Sciences
Department: Business Administration
Thesis Type: Master Thesis
Supervisor: Dr. Öğr. Üyesi Burcu KARTAL
Author: Sema PEÇE
Year: 2019
Page: 99

ABSTRACT

ANALYSIS OF BUSINESS ADMINISTRATION UNDERGRADUATE COURSES WITH DATA MINING METHODS: RECEP TAYYİP ERDOĞAN UNIVERSITY CASE

Technology and the innovations brought by technology take their place in every part of life rapidly. It is the duty of the databases to carry the burden of these innovations. Evaluating data that is a meaningless burden and cost factor in databases is the area of interest of data mining. Data mining, which can be used in every field where data is stored, has been applied in the field of education in this thesis. The aim of the study is to analyze undergraduate courses. For this purpose, the course and student information of 731 students enrolled in Recep Tayyip Erdoğan University, Faculty of Economics and Administrative Sciences, Department of Business Administration since 2009 were examined. Grade-based relationships between the courses are examined with the Association Rules. According to the course grades, graduation scores, Student Selection Exam scores, Student Selection Exam placement rankings, passing the courses at first intake, with Classification algorithms were examined. Apriori, Naive Bayes, Sequential Minimal Optimization and J48 algorithms are the data mining algorithms used in the analysis.

Key Words: Data Mining, Education, Classification, Association Rules

KISALTMALAR

RTEÜ	: Recep Tayyip Erdoğan Üniversitesi
İİBF	: İktisadi İdari Bilimler Fakültesi
VM	: Veri Madenciliği
EVM	: Eğitsel Veri Madenciliği
SMO	: Sıralı Minimal Optimizasyon
CRISP-DM	: Cross Industry Standart Process for Data Mining
RFID	: Radio Frequency Identification
OLTP	: On Line Transactional Processing
OLAP	: On Line Analytical Processing
SLINK	: An Optimally Efficient Algorithm For The Single Link Cluster Method
CURE	: Clustering Using Reprisentatives
BIRCH	: Balanced Iterative Reducing And Clustering Using Hierarchies
CLUCDUH	: Clustering Categorical Data Using Hierarchies
PAM	: Partitioning Around Medoids
CLARA	: Clustering Large Application
CLARANS	: Clustering Large Applications Based on Randomized Search
DBSCAN	: Density Based Spatial Clustering of Applications with Noise
OPTICS	: Ordering Points to Identify the Clustering Structure
DENCLUE	: Density Based Clustering
STING	: Statistical Information Grid
CLIQUE	: Clustering in Quest
ECLAT	: Equivalence Class Transformation
FP-Growth	: Frequent Pattern Growth
CART	: Classification and Regression Tree
CHAID	: Chi-square Automatic Interaction Detector
KNN	: K-Nearest Neighborh
DVM	: Destek Vektör Makinesi
MOODLE	: Modular Object Oriented Dynamic Learning Environment

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Teknolojinin ve bilginin günden güne değer kazanması ve biriktirdiklerinin artması birçok kavramı, sistemi hayata katmıştır. Teknoloji sayesinde gelişim gören bilgisayar birçok alanda yerini almıştır. Bilgisayarın girdiği her alanda veri girişleri olmakta ve bu veriler yığınlar haline gelmektedir. Bakıldığında bir yük ve maliyet unsuru olarak görülebilen veri kavramı avantaj haline getirilebilir. Verilerin tutulduğu veritabanında, verilerin sistematik tutulması ve karar vermede etkili hale gelebilmesi için geliştirilen veri ambarlarından yararlanılması uzun zamandır süregelmektedir. Tutulan bu verilerin kendi içinde anlamlı hale gelebilmesi, aralardaki örüntülerin keşfedilmesi veri madenciliği sistematiğini oluşturmuştur. Veri ve veri yığınları işlem görmemiş fakat değeri bilinen kömür gibi düşünülebilir. İşlem gördüğünde elmasa dönüşen kömür veriye, işlem görme hali ise veri madenciliğine benzetilir.

Her gün milyonlarca insan alışveriş yapmaktadır. Yapılan bu alışveriş hem tüketici için hem de üretici için birikimler yaratmaktadır. Üreticinin yaptığı bu veri birikimi, verilerin depolanmasını, korunmasını, okunabilecek halde olmasını ve de tüm bu işlemler için bilgisayarın yanında kullanıcı desteğini gerektirmektedir. Tüm bu süreç üretici ya da kurum için ayrı bir maliyet unsurudur. Bu maliyet unsurunu avantaj haline getirmek veri madenciliğinden yararlanmayı popüler hale getirmektedir. Tüketicinin yaptığı alışveriş kayıtlarıyla alınan ürünler arasındaki ilişki keşfedilebilir. Milyonlarca kayıt arasından kolalı içecek alanların çekirdek de aldığı bilgisini keşfetmek pazarlama açısından bilgi kazanımı yaratabilir. Kolalı içecek ve çekirdeğin raf yerlerini birbirine yakın tutmak, kolalı içecek alan tüketicileri çekirdek almaya itebilir.

Pazarlamadan, tıbba, bankacılığa, borsaya, eğitime ve birçok veri yığınının olduğu alanda veri madenciliği çalışması yapılmaktadır. Eğitim alanında yapılan veri madenciliği çalışmaları Eğitsel Veri Madenciliği başlığı altında toplanmıştır. Eğitsel veri madenciliği; Eğitimde Yapay Zeka Dergisi (Journal of Artificial Intelligence in Education), Uluslararası Akıllı Ders Sistemi Konferansı (International Conference of Intelligent Tutoring System), Uluslararası Eğitsel Veri Madenciliği Konferansı (International Conference on Educational Data

Mining), Eğitsel Veri Madenciliği Dergisi (Journal of Educational Data Mining), Kullanıcı Modellemesi için Veri Madenciliği Uluslararası Konferansı (International Conference on Data Mining for User Modelling) gibi birçok konferansta, dergide, söyleşide yer almaya başlamış ve gelişme göstermiş, popülerliği artmakta olan bir çalışma alanıdır. (Barahate, 2012).

Veri madenciliğinin eğitim alanındaki uygulaması, eğitimcilere ve eğitim planlamacılarına farklı açılardan ışık tutarak eğitim stratejilerinin belirlenmesinde ihtiyaca yönelik çözümler sunmaktadır. Eğitsel veri madenciliği, eğitim ortamından elde edilen verilerin daha önceden bilinmeyen veya oluşturulmamış yapısını keşfetmek için yöntem geliştiren ve geliştirilen yöntemleri öğrencileri daha iyi tanımada kullanan gelişmekte olan bir disiplin olarak tanımlanır (Bilen, Hotaman, Aşkın, & Büyüklü, 2014).

Eğitim alanında yapılan çalışmalar ile sorgulanabilecek problemler şu şekildedir (Bhullar, 2012);

- Zayıf öğrenciler kim?
- Hangi öğrenciler en fazla kaç ders saati almıştır?
- Öğrencilerin ilgi duyduğu konular nelerdir?
- Ne tür dersler öğrencilerin daha çok ilgisini çekmektedir?
- Zayıf öğrencilere nasıl yardım edilebilir?
- Puan durumları nasıl değiştirilebilir?
- Öğrenci durumları tahmin edilebilir mi?

Romero & Ventura, eğitim alanında yapılan veri madenciliği çalışmalarını şu başlıklarda toplamıştır;

- Veri analizi ve görselleştirme
- Geri bildirim sağlama ve öneri sistemleri
- Öğrenci modelleme, öğrenci gruplama
- İstenmeyen öğrenci davranışlarını belirleme
- Sosyal ağ analizi
- Kavram haritası geliştirme
- Ders malzemelerini yapılandırma
- Planlama, çizelgeleme

- Öğrenci performansını tahminleme

Eğitim üzerine yapılmış veri madenciliği çalışmaları incelendiğinde başarı tahminleme, başarıyı etkileyen faktörleri gruplama, ders notlarına göre öğrenciye ait değişkenleri inceleme, eğitim ortamlarının başarı üzerine etkisini belirleme gibi çok faktörlü çalışmalar yapıldığı görülmüştür. Bu çalışmalardan bazıları aşağıda özetlenmiştir;

Kurt & Erdem (2012), öğrenci başarısını etkilediği düşünülen kişisel, sosyal, ekonomik ve barınma ile ilgili demografik özellikleri içeren likert tipinde bir ölçek hazırlamıştır. Anket, Gazi Üniversitesi, Teknik Eğitim Fakültesindeki 545 üniversite öğrencisine uygulanmıştır. Veriler, SPSS Clementine programında, CHAID, CRT, Yapay Sinir Ağları, Apriori, K Means algoritmaları kullanılarak analiz edilmiştir. Mezuniyet sonrası bölümle ilgili bir işte çalışıp çalışamama ihtimali, kişilik, bölümün istenerek okunup okunmadığı, lise başarı durumlarının başarıyı etkilediği gözlemlenmiştir. Kılınç, (2015:61-62)Osmangazi Üniversitesi, Bilgisayar Mühendisliği öğrencilerinin sağlık bilgileri, demografik bilgileri, ders bilgileri ve öğrenci bilgi sistemindeki bilgilerle veri seti oluşturmuştur. Öğrencilerin başarılarına etki eden faktörleri belirlemek ve 2011 yılında değişen atılma politikasının öğrenciler üzerinde etkisinin olup olmadığını öğrenmek tezin başlıca amaçlarıdır. Burs alıp almama, anne baba eğitim durumu, okuldan atılma durumunun artık ortadan kalkması öğrenci başarısını etkileyen etmenler olarak bulunmuştur. Analizleri WEKA'da KNN, J48, Apriori, Predictive Apriori ile yapmıştır. Özarslan (2014:67-69) Kırıkkale Üniversitesinde, Temel Bilgi Teknolojileri Kullanımı dersini alan öğrencilere ait verileri veri madenciliği yöntemleri ile incelemiştir. Web tabanlı uzaktan eğitim ile klasik eğitim yöntemlerine göre öğrenci performansları değerlendirmiştir. Öğrenci performansını etkileyen başlıca faktörler ise eğitim türü, fakülte ya da yüksekokul öğrencisi olması ve öğrenci yaşı bulunmuştur. Analizleri WEKA'da J48 ve JNP algoritmalarıyla yapmıştır. Vandamme, Meskens ve Superby, (2007) Mons Katolik Üniversitesi, Üretim Yönetimi bölümü, 533 tane birinci sınıf öğrencisinin ilk dönem notlarını, ekonomik durumları ve demografik bilgilerine göre düşük riskli, orta riskli, yüksek riskli olmak üzere üç kategoriye ayırıp erkenden önlem

almak istemiştir. Analiz SAS programında Karar Ağaçları ve Yapay Sinir Ağları ile gerçekleştirilmiştir.

Güner ve Çomak (2010) 2007 yılında Pamukkale Üniversitesi, Mühendislik Fakültesine başlayan 434 öğrencinin üniversiteye giriş sınavı sonuçlarına göre ve lisedeki ders bilgilerine göre Matematik1 dersi başarılarını incelemiştir. Çalışmada Destek Vektör Makineleri kullanılmıştır. Öğrencilerin Matematik, Fen Bilimleri, Türkçe testlerinin sonuçları ile lise mezuniyet başarı puanlarının, Matematik I dersindeki başarılarını tahminde önemli rol oynadığı bulunmuştur. Aydın ve Özkul (2015) öğrenci bilgi sistemi ve e-öğrenme sisteminden sağlanan verilerle öğrenci başarı durumunu tahmin etmeyi hedeflemiştir. SPSS Clementine programında C5.0, Logistic Regression, Neural Net, CART, CHAID ve QUEST algoritmaları kullanılmıştır. En iyi sonuçları C5.0 algoritması vermiştir. Özçınar (2006:48-49) Pamukkale Üniversitesi, Eğitim Fakültesi, Sınıf Öğretmenliği öğrencilerinin Kamu Personeli Seçme Sınavı'ndan aldıkları puanları, öğrencilerin lisans eğitimleri süresince bazı derslerden aldıkları geçme notları, genel not ortalamaları ve öğretim türleri tahmin edici değişkenler olarak kullanarak öngörmeye çalışmıştır. Yapay sinir ağları ve regresyonun kullanıldığı çalışmada, yapay sinir ağları gerçeğe daha yakın sonuçlar vermiştir. Calvo-Flores, Galindo, Jiménez, & Pérez (2006) öğrencilerin Moodle (Modular Object Oriented Dynamic Learning Environment) sistemine giriş bilgileriyle öğrenci notlarını tahmin etmiştir. Cordoba Üniversitesi, Metodoloji ve Programlama dersini alan 240 öğrencinin verileriyle yapılan çalışmada Yapay Sinir Ağları kullanılmıştır. Wang, Zhu, Huang, Li, & Ji, (2018) öğrenci sınav puanları ve kişisel özellikler yerine wi-fi erişim bilgileri, elektronik kart kullanım bilgileri ve kampüs internetine erişim bilgileri gibi geleneksel olmayan değişkenlerle üniversite öğrencilerinin başarısını tahmin etmeyi amaçlamıştır. Destek Vektör Makineleri, Lojistik Regresyon, Karar Ağaçları ve Naive Bayes ile yapılan çalışma geleneksel öğrenci bilgileriyle kıyaslandığında doğru sonuçlar vermiştir. Bresfelean (2007) Cluc Napoka Üniversitesi, Ekonomi ve İşletme Fakültesi öğrencilerinin demografik bilgileri, lise bilgileri, üniversite kabul puanı, lisans ders ve not bilgileri, aktiviteleri, katıldığı araştırmalar, ileri dönük fikirleri gibi bilgilerin yer aldığı bir veri seti hazırlamıştır. Amaç lisans öğrencilerinin

lisansüstü tercihlerini belirlemek ve tahmin etmektir. Weka'da J48 algoritması ile analiz yapılmıştır. Asif, Merceron, Ali ve Haider (2017) Pakistan Mühendislik Fakültesi, Bilgi Teknolojileri bölümü, 210 lisans öğrencisinin lisans öncesi eğitim bilgileri ve lisans dönemi ilk iki yıllık ders bilgileri ile mezuniyet notlarını tahmin etmek istemiştir. Analizde RapidMiner'da Karar ağaçları, Naive Bayes,Yapay Sinir Ağları, K En Yakın Komşu algoritmaları kullanılmıştır.

Can (2017:120-121) Gazi Üniversitesi, dönem sonu ders değerlendirme anketi verilerini analiz etmiştir. Üç farklı eğitmen için öğrenciler tarafından sağlanan toplam 5820 adet anket bilgisi çalışmada yer almıştır. Likert tipteki anket verileri 28 adet sorunun cevabı ve 3 adet nitelik bilgisini (öğrenci tarafından algılanan zorluk seviyesi, devam durumu ve ders tekrar sayıları) içermektedir. Veriler lojistik regresyon ve karar ağaçlarıyla analiz edilmiştir. Güven (2016:1-2) Türkiye'deki 88 üniversitenin, bilgisayar mühendisliği bölümü öğrencilerinin 2014 yılına ait üniversite giriş sınavı sonuç bilgileri, lisans ders içerikleri ve ders planlarını ayrı ayrı incelemiştir. KNIME'da karar ağaçları, kümeleme ve birliktelik analizi yapılmıştır. Yakupoğlu (2018:81-82) Milli Eğitim Bakanlığı'na bağlı bir özel okul öğrencilerinin, bir akademik yıl boyunca okulun bilgi yönetim sisteminde kayıt altına alınan verilerini kullanarak, öğrencilerin 5 ana ders üzerindeki a-, a, b, c ve c+ yetkinlik sınıflarının eğitimsel veri madenciliği teknikleriyle tahmin edilip edilemeyeceğini araştırmıştır. Çalışma WEKA'da C4.5, Çok Katmanlı Yapay Sinir Ağları, K En Yakın Komşu ve Destek Vektör Algoritmaları kullanılmıştır. Araştırma sonunda DVM algoritması dışında diğer algoritmalarda kayda değer bir başarı sağlanamamıştır. Ayık, Özdemir ve Yavuz (2007) çalışmalarında Atatürk Üniversitesinden 1976 yılından itibaren mezun olan ve okumakta olan öğrenci bilgilerinin bulunduğu veritabanında veri madenciliği teknikleri uygulamıştır. Purple Insight Miner"da sınıflandırma teknikleri kullanılmıştır. Analiz sonucunda, lise türünün istenen fakültenin kazanılmasında çok büyük öneminin olduğu, lise başarısının da aynı derecede önemli olduğu gözlemlenmiştir. Mamcenko, Sileikiene, Lieponiene ve Kulvietiene (2011) Vilnius Gediminas Teknik Üniversitesi, 19 alandaki 3167 öğrencinin C++ dersi elektronik sınavına dair davranışlarını keşfetmek istemiştir. Kümeleme için Kahonen Algoritması ve birliktelik kuralları için SIDE algortiması

kullanılmıştır. Dersin sınavına kaç kez girildiği, derse ait sınavda doğru yanlış sayıları, cevaplama süreleri gibi değişkenler analiz edilmiştir.

Tissera, Athauda ve Fernando (2006), Sri Lankadaki bir Bilgi ve İletişim Teknolojisi eğitim kurumunda, 2004-2005 yıllarındaki 3 anabilim dalında lisans derecesindeki 20 ders ve derse ait başarı durumlarını göz önüne alarak 3 alandaki derslerin kendi aralarındaki ilişkilerini incelemiştir. Bu şekilde ders programındaki dersleri organize etmek istemiştir. Olası kombinasyonları belirlemek için Birliktelik Kuralları kullanılmış. İlişkilerin gücünü belirlemek için ise Pearson korelasyon katsayısı uygulanmış. Çalışma sonucunda programlama ve matematik dersleri arasında ilişki bulunamamıştır.

İnce & Karabatak (2004), öğrencilerin genel kültür derslerinden aldıkları notları dikkate alarak bu notların nasıl bir dağılım gösterdiği, aralarında nasıl ilişkiler olduğu ve bu dersler arasında ne gibi kuralların bulunduğu tespit etmeye çalışmıştır. Apriori algoritmasıyla MATLAP'da yapılan analiz sonucunda ÖSS sayısal puan türü ile gelen ve yüksek puanlarla bölüme yerleşen öğrencilerin ilginç bir şekilde sayısal derslerden düşük notlarla veya zoraki geçtikleri sonucuna varılmıştır.

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİ

1. Veri Madenciliği

Teknoloji ve bilişimin getirdiği hızlı yenilikler, hayatı kolaylaştırırken ürünlerin ucuzlaması ve talebin artması kaçınılmazdır. Bu gelişmeler bazı sorunları peşinde getirmiştir. Yapılan tüm alışveriş bilgileri, müşteri bilgileri, satış bilgileri ve bilgisayarın girdiği her ortam veri yığınları haline gelmiştir. Veri ambarlarında işlevsiz kalan bu bilgilerden anlamlı cümleler üretmek ve yeni bilgiler elde etmek bilişimin görevi olmuştur (Özkan, 2016: 11). Bu aşamada veri madenciliği kavramı türemiştir. Bu kavram altın veya kömür madenciliğine benzetilir. Veri yığınlarından soyut kazılarla veriyi ortaya çıkarma işlemidir. Yapısız veriden, anlamlı ve işe yarar bilgiyi çıkarmayı sağlayacak işlemleri analiz etmektir.

Veri madenciliği 1990'lardan beri veri depolama araçları, barkod ve RFID teknolojilerine paralel olarak gelişmekte ve kullanım alanı yayılmakta olan bir konudur (Silahtaroğlu, 2016:12).

Gelişimi dinamik olan bir konunun net bir tanımını yapmak zor ve yetersiz kalmaktadır. Literatürde yapılan çeşitli Veri Madenciliği tanımları şu şekildedir:

VM, veri setleri arasındaki düzenin ya da desenin, verinin analizi ve yazılım tekniklerinin kullanılmasıyla alakalıdır. Veriler arasındaki ilişkiyi, kuralları ve özellikleri belirleyen bilgisayardır. Amaç, daha önceden görülmemiş veri desenlerini tespit etmektir (Vahaplar & İnceoğlu, 2002). Veri madenciliği klasik istatistik uygulamalarına çok benzemesine rağmen sağlam çizgilerle ayrılır. İstatistik uygulamaları düzenlenmiş ve genelde özet veriler üzerinde çalışır. Veri madenciliği milyonlarca veri ve çok daha fazla değişken ile uğraşır.

Dijital verilerin artması, veritabanlarında kayıtların çoğalması bu verilerden en verimli şekilde faydalanma ihtiyacını ortaya çıkarmıştır. Bu süreç Veri Tabanlarından Bilgi Keşfi adı altında ilerlemeler kaydetmiştir. Veri madenciliği bu süreçte en önemli yeri almıştır. Bu kavram büyük veri içinden

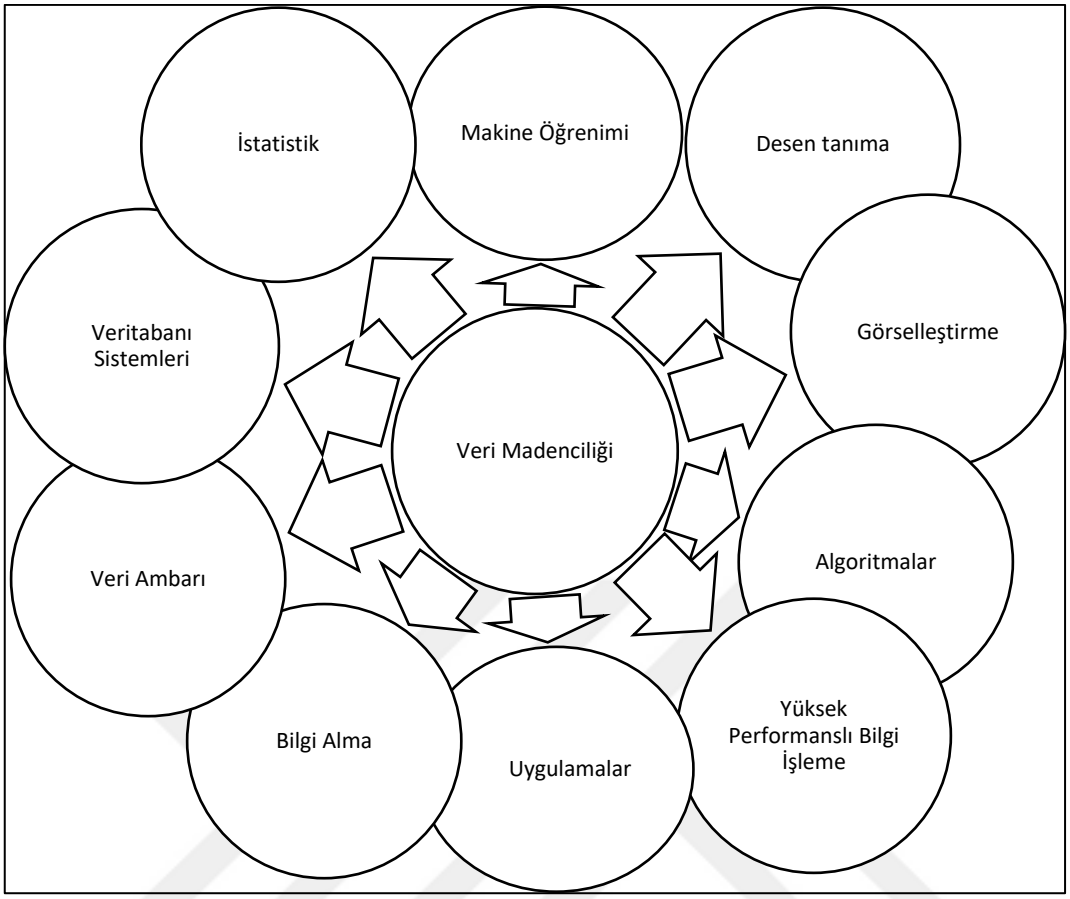
geleceğe dair tahminler yapmayı sağlayacak bağıntı ve kuralların bilgisayar programları vasıtasıyla aranmasıdır (Karabatak & Cevdet, 2012).

Veri madenciliği, büyük hacimdeki veri yığınlarından karar alabilmek için potansiyel olarak yarar sağlayacak, uygulanabilir ve anlamlı bilgilerin çıkarılmasına odaklanır. Tek başına bir çözüm değil veri analiz teknikleri bütünüdür (Altun, 2017).

Bilginin olağandışı artmasıyla kurum içi ve kurum dışı bilgilere ek olarak tahmini zor sorulara cevap bulan, ileri dönük sistemlere ihtiyaç duyulmuştur. Biriken bu veri içerisinden kurumlara yararlı bilgileri bulma ve ortaya çıkarma işine veri madenciliği denir (Kurt & Erdem, 2012). Günümüzde her kurum kendi alanında rekabet edebilmesi için iş süreçlerinde elde ettiği verilerden mutlaka anlamlı bilgiler çıkartmak ister. Kurumlar veriler üzerinde yaptıkları analiz çalışmaları sonucu elde ettiği anlamlı bilgiler ışığında gelecekteki politikalarını belirler. Bu bilgiler sayesinde belirlenen politikalar, kurumların ileriye yönelik düzeltici ve iyileştirici adımlar atmasına olanak tanır ve maliyet, zaman, işgücü anlamında kazanç sağlar (Ünsal, 2011:57-59).

Veri madenciliği, özetle büyük miktardaki verinin veri tabanlarında, veri ambarlarında veya diğer bilgi depolama alanlarında, farklı kalıpların keşfedilmesiyle alakalı bir süreçtir.

Yüksek uygulama odaklı bir alan olan veri madenciliği Şekil 1'de gösterildiği gibi istatistik, makine öğrenimi, desen tanıma, görselleştirme, algoritmalar, yüksek performanslı bilgi işleme, uygulamalar, bilgi alma, veri ambarı, veritabanı sistemleri gibi birçok alanı kapsamaktadır. Veri madenciliğinin disiplinlerarası niteliği veri madenciliğinin başarısına ve kapsamlı uygulamalarına önemli katkılar sağlamaktadır (Han, Pei, & Kamber, 2012: 23).



Şekil 1.Veri madenciliğinin benimsediği alanlar (Han, Pei ve Kamber, 2012:23)

İşlenmemiş veriden, son kullanıcının kolayca anlayıp karar alma sürecine dâhil edebileceği bilgiyi oluşturana kadar geçen tüm süreci kapsayan bir yöntem olmasından, hipotez doğrulamaya yönelik değil yeni, gizli örüntüler bulmaya yönelik bir alan olmasından ve çok çeşitli teknikleri aynı uygulama içinde kullanabilmeye olanak sağlamasından dolayı veri madenciliği kullanıcılarına kendisini oluşturan makine öğrenmesi, istatistik matematik gibi yöntemlerden daha farklı bir açı sunar (Özçınar, 2006: 6) .

Veri madenciliği; büyük miktarda ve hızlı toplanan verilerin, çeşitli analizler vasıtasıyla anlamlı bilgilere dönüştürülmesi noktasında devreye giren bir süreçtir. Veri madenciliği tanımlarında ortak olan unsurlardan ilki çok fazla miktarda verinin veri ambarında tutulması, ikincisi ise bu verilerden anlamlı bilgiler elde edilmesidir (Şimşek, 2006:5-6).

Veri madenciliği, araştırmacıya anlamlı ve yararlı olabilecek şekilde veri kümesindeki ilişkileri ortaya çıkarmak ve yeni veriyi özetlemek için veri

kümelerinin incelenmesidir (Bilgin, 2013:3). Hangi potansiyel bilginin veriler içinde olduğunu bulmak, problemi çözmek için olası yöntemlerin nasıl kullanılacağını anlamak için verilerle uğraşmanın diğer bir adıdır aslında veri madenciliği (Kırlıoğlu & Ceyhan, 2014).

Günlük hayatta tıp, eğitim, endüstri, otomotiv, pazarlama, bankacılık gibi sektörler mevcut faaliyetlerini devam ettirirken aynı zamanda büyük veriler yığınlamaktadır. Bu veri yığınları pek çok keşfedilmemiş bilgiyi içerirler. Veri madenciliği, teknolojide yaşanan hızlı değişikliklere paralel olarak her sektörde oluşan bu veri yığınları içerisinden değer yaratacak veriyi anlamlandırma sürecidir (Akbal, Doğan, & Varol, 2017).

Tarihsel olarak bakıldığında verilerden yararlı kalıplar bulma fikri çeşitli şekillerde isimlendirilmiştir; veri madenciliği, bilgi çıkarımı, bilgi keşfi, bilgi toplama, veri arkeolojisi ve veri modeli işleme gibi (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Aynı şekilde veriden bilgi madenciliği, bilgi çıkarma örüntü analizi, veri tarama gibi isimlerde literatürde geçmektedir (Han, Pei, & Kamber, 2012: 6). İlk olarak 1989 yılında bir atölye çalışmasında tanımlanan veri madenciliği kavramı, veri işleme sürecinde bilginin son ürün olduğunu belirtmek için Veri Tabanlarında Bilgi Keşfi olarak ifade edilmiştir (Bilgin, 2013).

İş yönetiminde, devlet yönetiminde, bilimde, mühendislikte ve diğer pek çok uygulamada milyonlarca veri tabanı kullanılmaktadır. Güçlü ve uygun fiyatlı veritabanı sistemleri sayesinde veritabanlarının sayısında artış olmuştur. Bu veri ve veri tabanlarındaki patlayıcı büyüme, bilgiye otomatik olarak dönüştürebilen yeni teknikler ve araçlar için acil bir ihtiyaç doğurmuştur. Sonuç olarak, veri madenciliği önemi artan bir araştırma alanı olmuştur (Chen, Han, & Yu, 1996).

Veri madenciliği temelde şu beş faktörden etkilenir: veri, donanım, bilgisayar ağları, bilimsel hesaplamalar, ticari eğilimler. Veri, veri madenciliğinin hammaddesidir. Donanım, gelişen bellek ve işlem hızı kapasitesiyle önceden yapılamayan madenciliğin günümüzde rahatça yapılabilmesini sağlar. Bilgisayar ağları sayesinde erişim hızı artmakta, zamandan tasarruf edilmekte ve dağınık verilere erişim sağlanmaktadır. Bilimsel hesaplamalar; teori, deney ve simülasyonu bağdaştırmada önemlidir. Ticari eğilimler ise veri madenciliğini

maliyet, pazarlama, hız bağlamında etkiler. İşletmeler bu açılardan avantaj sağlamak için veri madenciliği tekniklerini kullanır (Akpınar,2000).

Veri madenciliği, bilgisayar teknolojilerinin sağladığı hızlı veri işleme ve yüksek hacimli veri toplama imkanlarıyla ve farklı disiplinlerin katkısıyla sağlanan araçlarla, sahip olunan büyük hacimli veriden, karar vericinin etkin ve daha fazla bilgiye göre karar vermesinde kullanabilmesi için önceden bilinmeyen, gizli, örtük, klasik metotla ortaya çıkarılması zor, faydalı, ilginç, anlaşılır; ilişki, örüntü, bağıntı veya trendlerin otomatik ya da yarı otomatik şekilde ortaya çıkarılmasıdır (Şentürk, 2006:3-4).

Veri madenciliği ne değildir ve ne olmalıdır (Gorunescu, 2011):

Ne değildir?

- İnternetten ayrıntılı bilgi araştırmak,
- Aynı hastalığa sahip hasta kayıtlarını sorgulamak,
- Yer listesinden termal otellerin yerini sorgulamak,
- Şirketlerin finansal raporlarından tabloları analiz etmek

Ne olmalıdır?

- İnternette aynı içerikteki benzer bilgileri gruplamak,
- Benzer semptomlu hastaları gruplamak,
- Termal otelleri, ilgilendikleri hastalıklara göre gruplamak
- Şirketlerin satışa dair veri tabanlarından müşteri profillerini çıkartmak.

Veri madenciliği çalışmaları işletmelere veriyi değerlendirmede büyük yararlar sağlar. Atıl kalabilecek veriler bu sayede değer görür. Fakat veri madenciliğinin de kendi içerisinde maliyetleri ve yanılsamaları bulunmaktadır. Larose (2005), Nautulius Sistemlerin başkanı Jen Que Louie'nin ABD Temsilciler Meclisi Teknoloji, Bilgi Politikası, Hükümetlerarası İlişkiler ve Nüfus Sayımı Alt Komitesi'ndeki konuşmasında geçen Veri Madenciliğine ilişkin yanılsamaları kitabında şu şekilde ifade etmiştir.

Yanılsama 1: Veri havuzlarımızda serbest bırakabileceğimiz ve sorunlarımıza cevap bulmak için kullanabileceğimiz veri madenciliği araçları var.

Gerçek: Sorunlarınızı mekanik olarak "beklerken" çözecek otomatik veri madenciliği araçları yoktur.

Yanılsama 2: Veri madenciliği süreci özerktir, az veya hiç insan gözetimi gerektirmez.

Gerçek: Veri madenciliği süreci her aşamada önemli insan etkileşimi gerektirir. Model dağıtıldıktan sonra bile, yeni verilerin tanıtılması genellikle modelin güncellenmesini gerektirir. Sürekli kalite izleme ve diğer değerlendirme önlemleri, insan analistler tarafından değerlendirilmelidir.

Yanılsama 3: Veri madenciliği kısa sürede harcanan maliyeti karşılar

Gerçek: Karşılama durumu, başlangıç maliyetlerine, analiz personeli maliyetlerine, veri ambarına hazırlık maliyetlerine bağlı olarak değişir.

Yanılsama 4: Veri madenciliği yazılım paketleri sezgiseldir ve kullanımı kolaydır.

Gerçek: Kullanım kolaylığı değişmektedir. İşletme ihtiyaçlarına ve araştırma modeline göre süreç değişebilir.

Veri madenciliğinin benimsediği bir çok alandan önem gösteren dört alan istatistik, makine öğrenimi, veritabanı ve veri ambarı aşağıda tanımlanmıştır.

1.1 İstatistik

İstatistik, verilerin toplanması, analiz edilmesi, yorumlanması veya açıklanması ile sunumunu inceler. Veri madenciliği, istatistiklerle içsel bir bağlantıya sahiptir. İstatistiksel bir model, bir hedef sınıftaki nesnelerin rastgele değişkenler ve bunlarla ilişkili olasılık dağılımları açısından davranışını tanımlayan bir matematiksel fonksiyonlar kümesidir. İstatistiksel modeller, veri ve veri sınıflarını modellemek için yaygın olarak kullanılır. Örneğin, veri karakterizasyonu ve sınıflandırma gibi veri madenciliği görevlerinde hedef sınıfların istatistiksel modelleri oluşturulabilir. Başka bir deyişle, bu gibi istatistiksel modeller bir veri madenciliği görevinin sonucu olabilir. Alternatif olarak, veri madenciliği görevleri istatistiksel modellerin üzerine inşa edilebilir. Örneğin, gürültülü ve eksik veri değerlerini modellemek için istatistik kullanılabilir. Büyük bir veri kümesinde madencilik desenleri oluştururken, veri madenciliği işlemi verilerdeki gürültülü veya eksik değerleri tanımlamak ve işlemek için modeli kullanabilir (Han, Pei, & Kamber, 2012:23-24).

1.2. Makine Öğrenimi

Makine öğrenimi, bilgisayarların verilere dayanarak nasıl öğrenebileceklerini veya performanslarını artırabileceklerini araştırır. Ana araştırma alanı bilgisayar programlarının karmaşık kalıplarını tanımayı otomatik olarak öğrenmesi ve verilere dayalı akıllı kararlar almasıdır. Makine öğrenimindeki denetimli öğrenme, denetimsiz öğrenme ve yarı denetimli öğrenme ve aktif öğrenme veri madenciliğinin ilgilendiği kısımlardır (Han, Pei, & Kamber, 2012:25-26).

Denetimli öğrenme, veri madenciliği tekniklerinden olan sınıflandırmanın mantığını oluşturur. Öğrenmedeki denetim, eğitim veri setindeki etiketli örneklerden gelir. Örneğin, posta kodu tanıma probleminde, bir dizi el yazısı posta kodu görüntüsü ve bunlara karşılık gelen makineyle okunabilen çeviriler, sınıflandırma modelinin öğrenilmesini denetleyen eğitim örnekleri olarak kullanılır. Kısaca veri dizisinin önceden belirlenen sınıflara atanması işlemidir.

Denetimsiz öğrenme yine veri madenciliği tekniklerindeki kümelemenin mantığını oluşturur. Burada girdi örnekleri etiketli olmadığı için öğrenme süreci denetlenmez. Örneğin, denetlenmeyen bir öğrenme yöntemi, girdi olarak, el yazısı rakamların bir dizi görüntüsünü alabilir. 10 veri kümesi bulduğu varsayılsın. Bu kümeler, sırasıyla 0 - 9 arasındaki 10 farklı haneye karşılık gelebilir. Bununla birlikte, eğitim verileri etiketlenmediğinden, öğrenilen model bulunan kümelerin anlamını söyleyemez.

Yarı denetimli öğrenme bir model öğrenirken hem etiketli hem de etiketlenmemiş örneklerden faydalanan bir makine öğrenme teknikleri sınıfıdır. Bu yaklaşımda, etiketli örnekler sınıf modellerini öğrenmek için kullanılır, etiketlenmemiş örnekler, sınıfları sınırlandırmak için kullanılır.

Aktif öğrenme, kullanıcıların öğrenme sürecinde aktif bir rol oynamasını sağlayan bir makine öğrenme yaklaşımıdır. Aktif bir öğrenme yaklaşımı, kullanıcıdan bir etiketlenmemiş örnek kümesinden veya öğrenme programı tarafından sentezlenmiş bir örneği etiketlemesini isteyebilir.

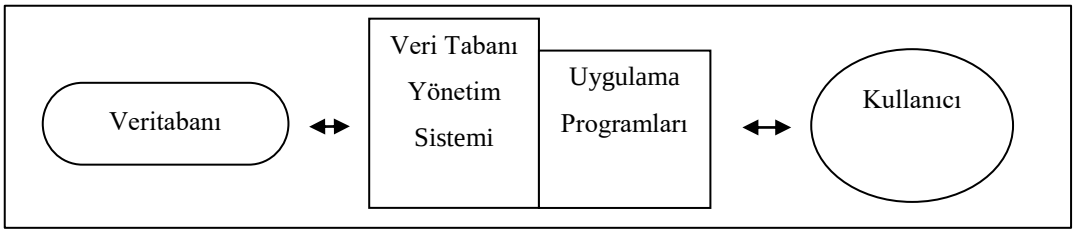
1.3. Veritabanı

Bir okul; öğretmen ve öğrencilerine ait kimlik bilgilerini, ders programlarını, sınav sonuçlarını, yoklama raporlarını saklama gereksinimi duyabilir. Ticari bir firma; personel özlük bilgilerini, depodaki malzemelere ait stok bilgilerini, müşterilerin ulaşım bilgilerini, gelen siparişleri, giden teslimatları, ürün fiyatlarını saklamak isteyebilir. Düzen içinde korunmak istenen bilgilerin içeriği ne kadar farklı olursa olsun, ortak bir konu veya belirli bir amaçla ilişkili bilgilerin oluşturduğu bütüne Veritabanı denir (Çayırlı, 2005).

Veritabanları sayısal, metinsel, tarihsel, mantıksal türlerde olabilir. Bu veritabanları yüksek miktarlarda veri saklamaya imkan tanırken, geliştirilen SQL ve benzeri sorgulama dilleri ile veriler üzerinde istatistiksel hesaplama işlemleri yapılabilmektedir.

Bakıldığında tüm bilgilerin kayıt altında olması güvenli gelmektedir. Fakat değerlendirilemedikten sonra veri tabanlarında tutulan veri, insanlar için sorun ve yük haline gelmektedir. Toplanan veri miktarı büyüdükçe ve toplanan verilerdeki karmaşıklık arttıkça, daha iyi çözümleme tekniklerine olan gereksinim de artmaktadır. Bu noktada Veri Madenciliği ve Veri Tabanlarında Bilgi Keşfi kavramları ortaya çıkmaktadır. Veri tabanlarında bilgi keşfi süreci başarılı olursa, keşfedilen bilgi, organizasyonların karar verme sürecini geliştirmede ve geleceğe yönelik fayda yaratmada etkili olabilir (Şimşek, 2006:2).

Karışık dosya yapıları, dosyalar arası ilişkilerin çok olması ve kullanıcıların dosyalara erişmesi gerektiği durumlarda geleneksel sistemin yetersiz kalması günümüzde karşılaşılan sorunlardandır. Bunun üzerine yeni yazılım teknolojilerine yönelme başlamıştır. Veritabanı sistemlerini oluşturmak ve bu sistemi yönetmek için Veritabanı Yönetim Sistemleri ortaya çıkmıştır. Veritabanı, Veri yönetim sistemi ve kullanıcı arasındaki ilişki şekil 2'de gösterilmiştir (Özkan, 2016: 13).



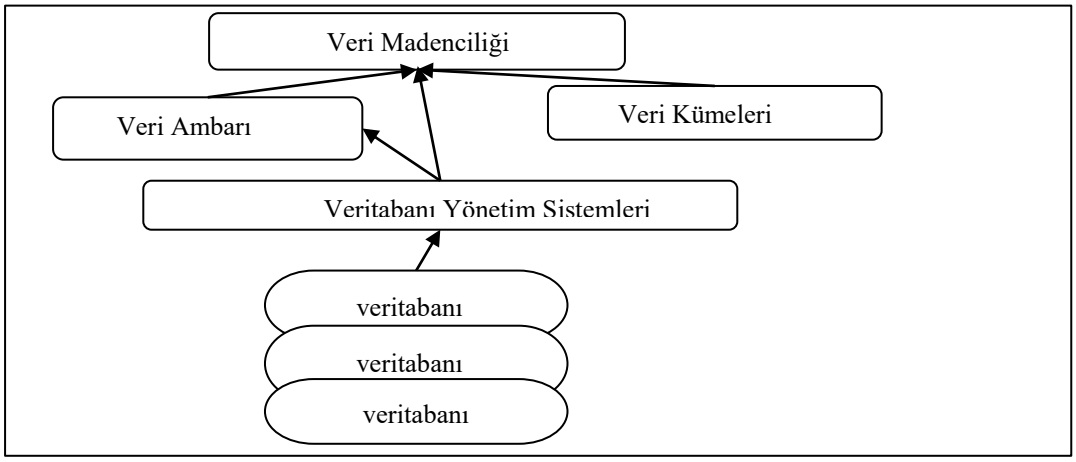
Şekil 2.Veritabanı ile kullanıcı arasındaki ilişki (Özkan, 2016:13)

1.4. Veri Ambarı

Veritabanı birçok avantaj ve kolaylık sağlamasına rağmen karar destek uygulamalarında yetersiz kalmıştır. Bu durum verinin farklı biçimde saklanmasının ve hızlı erişim elde etmenin çalışmalarını başlatmıştır. Geleneksel veritabanlarından elde edilemeyen büyük verinin kullanılabilmesi veri ambarı kavramını ortaya çıkarmıştır (Özkan, 2016:13-14).

İşletmelerde kullanılan işlemsel veritabanları veri madenciliği çalışmaları için doğrudan kullanılamaz. Bu amaç doğrultusunda veritabanındaki veriler uygun hale getirilmelidir. Belirli bir döneme ait, yapılacak çalışmaya göre düzenlenmiş, sabitlenmiş, birleştirilmiş veritabanlarına veri ambarı denir (Silahtaroğlu, 2016:17-18). Veritabanları, veri ambarı ve veri madenciliği arasındaki işlem akışı Şekil 3'te gösterilmiştir (Özkan, 2016:14).

Kurumlar günlük is akışları ve raporlama uygulamalarını Çevrimiçi Hareket İşleme (On Line Transactional Processing/OLTP) yöntemleri ile gerçekleştirmeye başlamışlardır. Veritabanı ve veritabanı sistemlerinin gelişmesi ve yaygın kullanılmasıyla depolanan veri miktarı artmış ve büyük veri ambarları ortaya çıkmıştır. Bu büyük veri ambarlarından karar destek sistemlerinde kullanılmak üzere bilgi elde etme sürecinde ise klasik OLTP sistemler yerine analiz çözümleri sunan Çevrimiçi Analiz İşleme (On Line Analytical Processing/OLAP) sistemleri kullanılarak ileri veri madenciliği uygulamalarına geçilmiştir. Veri sayısı çok olunca özel analiz algoritmaları geliştirilmiş,verinin saklandığı ortamlar da veri ambarı biçiminde yeniden düzenlenmiştir (Özkan, 2016:21-22).



Şekil 3. Veritabanı, veri ambarı ve veri madenciliği arasındaki ilişki (Özkan, 2016:14)

2. Veri Madenciliği Uygulama Alanları

Verinin yoğun olduğu her alanda veri madenciliği çalışmaları görülmektedir. Pazarlama, Bankacılık, Savunma Sistemleri, Borsa, Telekomünikasyon, Finans, Elektronik Ticaret, Sağlık, Sosyal Medya, Endüstri ve Eğitim veri madenciliğinin kullanılabildiği alanlardır.

Bir süpermarkette, müşterinin aldığı kalem sayısı o firma için sadece bir satış kaydı verisidir. O gün satılan toplam kalem sayısı, aylık ortalama kalem satış miktarı, kalem ile birlikte en çok hangi ürünün satıldığı bilgisi, gelecek ay bu ürünün ne kadar satılacağı gibi ifadeler anlamlı birer bilgidir. Ürünle birlikte en çok tercih edilen diğer bir ürünün bilgisi ürüne yönelik reklam ve satış politikalarında oldukça belirleyicidir. Kalem ile birlikte en çok kalem ucu ürününün satıldığı bilgisi varsayılırsa bu durumda bu süpermarket bu iki ürün reyonunu yan yana koyabilir veya çeşitli promosyon çalışmalarında bu iki ürünü bir arada müşterilerine sunarak bu ürünün satış oranlarını arttırabilir (Ünsal, 2011:18-19). Geçmiş yıllarda, aylarda, haftalarda hangi ürün hangi ürünle beraber satın alındı gibi soru kalıplarını anlamlandırmada ve ileriye dönük pazarlama çalışmaları yapmada veri madenciliği yöntemleri kullanarak analizler yapılabilmektedir.

Bankacılık alanında; müşteri profili belirlenir, kredibilite ve risk değerlendirmesi yapılır, hizmet planlaması ve seçilimi buna göre uygulanır. Bankalar veri madenciliği tekniklerini kredi kartı satışlarında, müşterilerin davranış ve güvenilirliklerini ölçmek ya da belirli bir müşterinin ödemelerini aksatma ihtimalini öngörmek amacıyla da kullanabilir (Altun, 2017).

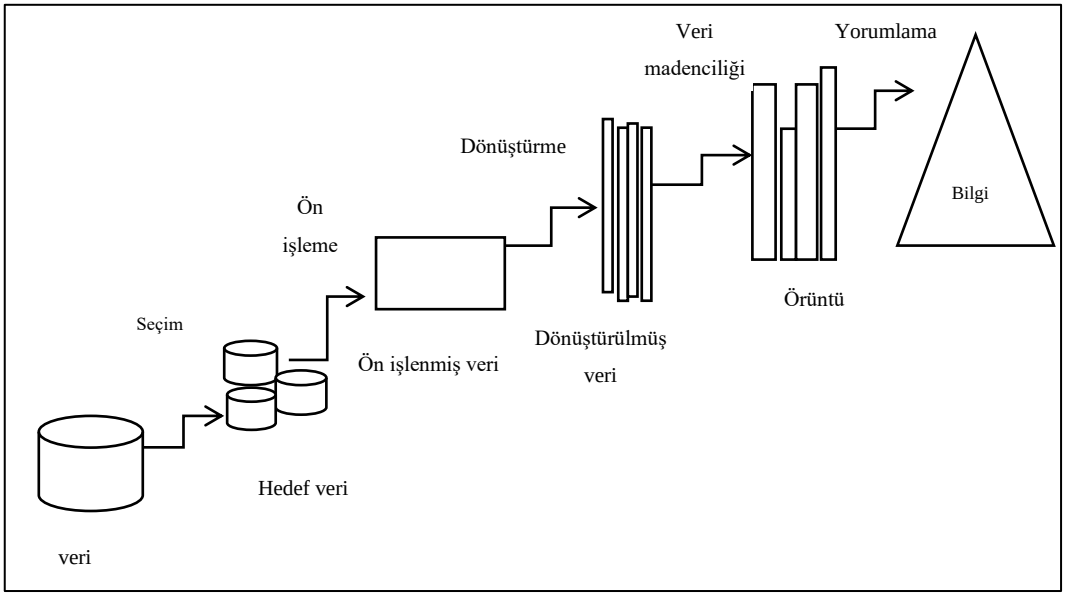
Eğitim alanında öğrencilerin başarısını, psikolojisini, geleceğe dair kariyer planlarını etkileyen faktörler tespit edilerek öğrenci, aile, okul ve öğretmenlere bilgiler verilebilir. Bu sayede öğrenciler eğitim hayatlarına dair sağlam temeller oluşturabilir, kurumlar durumlarını kontrol etmede fikirler edinebilir, öğretmenler rehber olma görevini analitik bir şekilde sürdürebilir. Tıp ve sağlık alanında yapılan veri madenciliği analizleriyle ise hastalıkların teşhisi hızlanmakta, ilaç yan etki analizleri kolaylaşmakta, tedavi süreçleri sağlam bir şekilde ilerleyebilmektedir.

3. Veri Madenciliği Süreci

Veri madenciliği, veri tabanlarından bilgi keşfi süreci içerisinde, modelin kurulması ve değerlendirilmesi aşamalarında görülen en önemli kısımdır. Farklı veri kaynaklarından verilerin toplanmasıyla başlangıç yapan veri tabanlarında bilgi keşfi süreci, toplanan verilerin analiz için uygun hale getirilmesi aşaması ile devam eder. Veri ambarlarına sahip olan kuruluşlarda, gerekli verilerin veri deposu olarak adlandırılan daha küçük veri ambarlarına aktarılması ile doğrudan veri madenciliği işlemlerine başlanması da mümkündür (Şimşek, 2006:4).

Veri madenciliği ve veri tabanlarından bilgi çıkarımı konusunda çeşitli süreçler bulunmaktadır. Fayyad, Piatetsky-Shapiro, & Smyth, 1996 yılında yayımladıkları makalede veri madenciliği sürecini akademik olarak açıklamıştır. Literatüre bakıldığında veri madenciliği sürecini açıklamak için bir çok çalışma yapılmıştır. Bunlar içinde en çok rağbet gören SPSS'in liderliğini yaptığı CRISP-DM ve SAS Institute Inc. tarafından Enterprise Miner için geliştirilen SEMMA'dır.

Fayyad, Piatetsky-Shapiro, & Smyth' in 1996 yılında Veri Madenciliğinden Veritabanlarında Bilgi Keşfine adlı makalede VM süreci Şekil 4'te gösterilmiştir;



Şekil 4. Veri madenciliği süreci (Fayyad,Piatetsky-Shapiro,Smyth, 1996)

Bu süreç 5 adımdan meydana gelir: Seçim, Ön İşleme, Dönüştürme, Veri Madenciliği, Yorumlama

Seçim: Önemli görülen veri seçilir veya oluşturulur. Sonraki adımda bu veri hedef veri olur.

Ön İşleme: Veri madenciliği ile yapılan analizlerde iyi bir sonuç alabilmek için başlangıçta hazırlanan verinin uygunluğuna ve kalitesine dikkat etmek gerekir. Veride ne kadar az parazit varsa sonuç kalitesi de o kadar iyi olacaktır. Sorun yaratan verilerin temizlenmesi, eksik verilerin tamamlanması ve yeni özniteliklerin oluşturulması bu aşamada önemlidir.

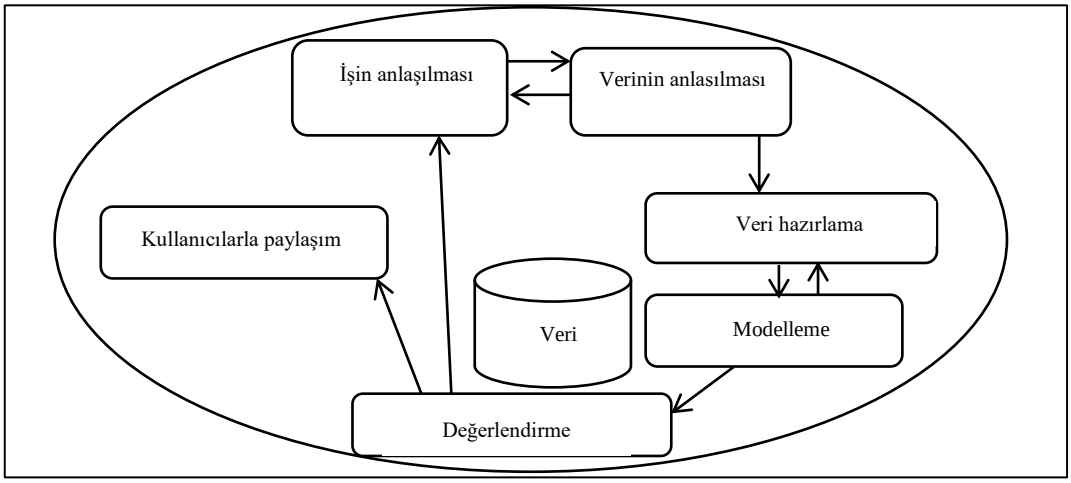
Dönüştürme: Verinin farklı veri madenciliği yöntemleriyle uygun forma dönüştürülmesidir.

Veri Madenciliği: Amaca uygun veri madenciliği yöntemlerinin kullanılması aşamasıdır. Bu aşama sonunda çeşitli örüntüler elde edilir.

Yorumlama: Veri madenciliği aşamasında belirlenen örüntülerin yeteri bilgiyi içerip içermediği araştırılır. Elde edilen örüntüler yeterli bulunmazsa önceki adımlar tekrar edilir.

3.1. CRISP-DM

CRISP-DM (Cross Industry Standart Process for Data Mining) süreci, veri madenciliği sürecini altı aşamaya ayırmıştır. Aşama sıraları kesin değildir ve farklı aşamalarda ileri geri geçişler olabilmektedir (SPSS, 2000). Şekil 5'te CRISP-DM'nin farklı aşamalarındaki ilişkileri gösterilmektedir.



Şekil 5: CRISP-DM süreci (SPSS, 2000)

İşin Anlaşılması: Bu ilk aşamada işletmenin bakış açısına uygun proje için amaç ve gereksinimlerin anlaşılmasına odaklanılır. Sonrasında ise amaç ve ihtiyaçlar VM problemi olarak tanımlanır ve amaca erişim için öncü plan tasarlanır.

Verinin Anlaşılması: Veri kalitesine ilişkin problemleri tanımlamak, veriye dair ilk izlenimleri edinmek, ilginç alt dizileri belirlemek ya da veri dizilerinde gizli kalan enformasyonun çıkarılması için hipotezlerin geliştirilmesi gibi amaca yönelik faaliyetleri içeren aşamadır.

Veri Hazırlama: Bu aşama ham veriden nihai veri elde edene kadar geçen tüm faaliyetleri içerir. Tablo, kayıt, öznitelik seçimi ve modelleme araçları için verinin temizlenmesi ve dönüştürülmesi gibi faaliyetleri içeren aşamadır.

Modelleme: Çeşitli modelleme yöntemleri seçilir ve uygulanır. Aynı VM problemi için birçok yöntem bulunmaktadır. Modellemede parametreleri optimal değerler elde edecek şekilde ayarlamak gerekir.

Değerleme: Modelleme aşamasında analiz için en iyi model ve modeller kurulmuştur. Bu aşamada kurulan model ya da modellerin en ince ayrıntısına kadar değerlendirmesi yapılır. Modelin işletme amaçlarını gerçekleştireceğine emin olunmalıdır. İncelenmemiş önemli işletme amaçları varsa bunları belirlemek önemli olacaktır. Bu aşama sonunda VM sonuçlarının kullanımına dair kararlar verilir.

Kullanıcılarla Paylaşım: Analiz sonucunda elde edilen bilgilerin son kullanıcının faydalanabileceği şekilde düzenlenmesi ve takdim edilmesi gerekir. Bu bir raporla yapılabileceği gibi sürecin tekrarlanması gibi karmaşık da olabilir.

3.2. SEMMA

SEMMA, SAS tarafından geliştirilen ve Sample, Explore, Modify, Model, Asses kelimlerinin ilk harfleriyle oluşmuş bir kısaltımdır. Bir VM yazılımı olan SAS Enterprise Miner için tasarlannıştır ve bir veri madenciliği projesinde modelleme yöntemleri üzerinde durulmuştur.

SEMMA beş aşamadan meydana gelmektedir.

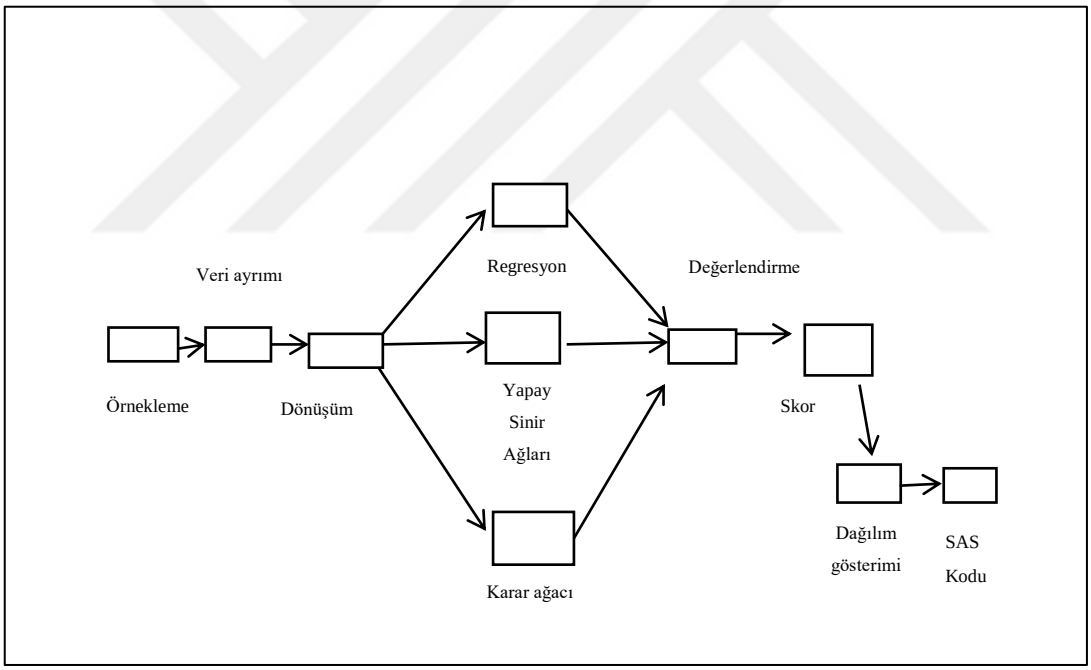
Örnekleme: Veri dizisi örüntü çıkarımı için yeterli enformasyonu içerecek kadar büyük, yöntemlerin etkili kullanımına olanak sağlayacak kadar küçük şekilde örneklenir.

İnceleme: Bu aşama öznitelikler arasındaki öngörülen veya öngörülmeyen ilişkilerin keşfi için, verinin anlaşılmasını ve veri görselleştirmeyle farklı değerlerin belirlenmesini içerir.

Dönüştürme: Özniteliklerin seçilmesi, oluşturulması ve dönüştürülmesi için kullanılan yöntemlerin belirlenmesi aşamasıdır.

Modelleme: İstenilen çıktıyı sağlayabilecek modellerin oluşturulabilmesi için, hazırlanan veri dizileri üzerinde çeşitli veri madenciliği tekniklerinin uygulandığı aşamadır.

Değerleme: Bu aşama oluşturulan modellerin geçerlilik ve etkinliliğinin değerlendiği aşamadır.



Şekil 6. SAS Enterprise Miner Semma süreci (Akpınar, 2014:76)

Şekil 6'da ilk olarak örnekleme işlemi yapılmakta, veri eğitim verisi ve test verisi olarak ayrılmakta, değişkenlerin dönüşümü yapılmakta. Modelleme aşamasında ise regresyon,yapay sinir ağları ve karar ağacı modelleri kullanılmakta. Modellemeden sonra ise değerlendirme, skorlama, dağılımın gösterimi ve sonuçların SAS kodunda elde edilmesi gösterilmiştir.

Yürütülen veri madenciliği analizlerinde öncelikle amaçların belirlendiği, problem odaklı, sonuçların nasıl ölçüleceğinin bilindiği, katlanılacak maliyetin hesaplandığı, doğru tahminlemelerle elde edilecek faydanın netleştirildiği kesin cümleler kullanmak gerekmektedir. Bu aşama genel olarak problemin tanımlandığı aşamadır.

Verinin anlaşılması ikinci aşamadır. Bu aşamada öznitelik değerlerinin dağılımı, sıra dışı değerlerin anlaşılması, hangi özniteliklerin projede önem taşıyacağı, veri kalitesinin belirlenmesi gibi kıstaslar yer alır.

Verinin hazırlanması üçüncü aşamayı oluşturur. Projenin yarısından fazla zamanı bu aşama alır. Verinin hazırlanması aşaması kendi içerisinde dörde ayrılır: veri entegrasyonu, veri temizleme, veri dönüştürme, veri indirgeme.

Veri Entegrasyonu: Probleme yönelik veri dizilerinin ve veri kaynaklarının belirlendiği adımdır.

Veri Temizleme: Çalışmada kullanılacak veri dizilerinin farklı kaynaklardan toplanması veriler arasında uyumsuzluğa neden olabilmektedir. Bu uyumsuzluk farklı zaman, farklı kodlama, faklı ölçü birimlerinden meydana gelebilmektedir. Önemli olan veri temizleme işini baştan savma bir şekilde yapmamaktır.

Veri Dönüştürme: Kullanılan algoritmalara göre verinin düzenlenmesidir.

Veri İndirgeme: Artan boyut sayısı ve toplam veri miktarları için çeşitli veri indirgeme yöntemleri geliştirilmektedir. Bu yöntemler boyut sayısının azaltılması, öznitelik alt dizisinin seçiminde makine öğrenme yöntemleri, faktör analizi, etkin örnekleme teknikleri gibi yöntemlerdir.

Problemin tanımlanması, verinin anlaşılması, verinin hazırlanması aşamalarından sonraki dördüncü aşama modelin kurulması aşamasıdır. Bu aşamada veri özelliklerine göre denetimli ya da denetimsiz yaklaşımların kullanımına karar verilir. Seçilen yaklaşıma uygun olarak yöntemler belirlenir. Modele uygun veri seçimi yapılır. Seçilen yöntemlere uygun olarak verinin dönüştürülmesi gerekebilir. Sıradışı değerler önemli bir bilgi sağlamıyorsa veri dizisinden çıkartılabilir. Verinin çok büyük olması durumunda veri içerisinden örnekleme yapılabilir ve bu örnekleme birçok model denenip bunlar içerisinde en güvenilir ve güçlü model seçilebilir.

Beşinci aşama kullanılacak yazılım veya yazılımların seçimidir. Altıncı aşama sonuçların geçerliliğinin sınandığı aşamadır. Yedinci aşama ise kurulan modelden elde edilen sonuçların değerlendirildiği başlangıç hedeflerine uygun olup olmadığının tartışıldığı aşamadır.

4. Veri Madenciliği Yöntemleri

Veri madenciliği yöntemlerinden kastedilen verinin veri madenciliği ile bilgiye nasıl dönüştürülebileceğidir. Veri madenciliği yöntemleri, bilgi keşfinin hedeflenen çıktılarına bağlı olarak çok farklı amaçlara sahip olabilirler. İstenen sonucu başarılı bir şekilde sağlamak amacıyla farklı amaçlara sahip birçok model, yöntem ve algoritma vardır. Kullanılan yöntemlerin birçoğu istatistiksel tabanlıyken bazıları ise yapay zeka tabanlıdır.

Veri madenciliği modelleri genel çerçevede tanımlayıcı modeller ve tahmin edici modeller olarak ikiye ayrılmıştır. Tanımlayıcı modeller; Kümeleme, Birliktelik Kuralları ve Ardışık Zamanlı Örüntüleri kapsar. Tahmin edici modellerde ise Sınıflandırma yer alır. Sınıflandırma, makine öğrenimi dilinde denetimli öğrenme içerisinde yer alırken, Kümeleme denetimsiz öğrenmede yer alır. Bu iki veri madenciliği stratejisi verinin modellenmesi aşamasında veriyi öğrenme, analiz etme ve modelleme metodolojilerini belirlemede esastır. Hedef kesin olduğunda denetimli öğrenme, elde edilmek istenen sonuç için net bir tanımlama yapılmamışsa denetimsiz öğrenmeden bahsedilir. Denetimli öğrenmede, veri örneklerinden hareket edilerek her bir sınıfa ilişkin özellikler bulunur ve bu özellikler kurallarla ifade edilir. Denetimsiz öğrenme, veriyi anlamaya, tanımaya, keşfetmeye yönelik kullanılır ve sonraki uygulanacak yöntemler için fikir verir. Denetimsiz öğrenmede hedef değişken yoktur. Ardışık Zamanlı Örüntüler ve Birliktelik Kuralları veri madenciliğinde, Pazar Sepet Analizi başlığında yaygın olarak kullanılmaktadır. Modeli kurmak için kullanılan tüm değişkenler açıklayıcı değişkendir.

4.1. Tanımlayıcı Modeller

Tanımlayıcı modellerde amaç, karar vermeye rehberlikte kullanılabilecek mevcut verilerdeki örüntülerin yani bağıntıların tanımlanmasını sağlamaktır

(Şimşek, 2011:5). Geliri 3000 ve 5000 para birimi aralığında olan, iki veya daha fazla arabası olan çocuklu aileler ve çocuğu olmayan, geliri 1500 ve 3000 para birimi aralığından düşük olan ailelerin satın alma örüntülerinin benzerlik gösterip göstermediğinin belirlenmesi tanımlayıcı modeli açıklamak için bir örnektir (Akpınar,2000). Kümeleme, Birliktelik Kuralları, Ardışık Zamanlı Örüntüler tanımlayıcı modele örnek yöntemlerdir.

4.1.1. Kümeleme

Kümeleme, verinin içerisindeki benzerlikler baz alınarak gruplandırılmasıdır. Amaç benzer özellikli nesneleri bir araya toplayarak farklı özellikli gruplar oluşturmaktır. Aynı küme içindeki nesnelerin benzer özellikleri fazla, ayrı kümelerdeki nesnelerin benzerlikleri ise az olmalıdır. Sınıflandırmada da amaç gruplama yapmaktır. Fakat sınıflandırmada gruplar önceden bellidir. Kümelemede ise verilerin hangi gruplara ayrılacağı belli değildir. Veriler benzerliklerine göre gruplara ayrılır (Aydemir, 2017:39). Makine öğrenimi bakımından, her bir küme gizli bir örüntüyü temsil eder ve uygulanan strateji denetimsiz öğrenmedir

Kümeleme süreci, her bir verinin farklı özelliklerini ifade eden değerlerinin, diğer bir veriden farklı ya da aynı olmasına göre işler. Farklılık ya da aynılık kümeleme algoritmaları için uzaklıkları ifade eder. Bu uzaklıklar Öklit Uzaklığı, Manhattan Uzaklığı, Minkowski Uzaklığıdır (Saygılı, 2013:27).

Kümeleme; istatistik, bilgisayar ve matematik gibi birçok bilim alanında kullanılmaktadır. İstatistikte çok değişkenli istatistiksel tahmin ve örüntü tanıma konuları kümeleme analizinden yararlanır (Arthur, Laırd, & Rubın, 1977).

Literatürde birçok kümeleme algoritması bulunmaktadır. Kümelemenin oluşturuluş şekline göre, veri türüne göre ve çalışmanın amacına göre bu algoritmalar çeşitlenmektedir. Genel olarak hiyerarşik ve bölümlemeli olarak ikiye ayrılan kümeleme algoritmaları literatürde daha alt bölümlere de ayrılmaktadır.

- Hiyerarşik Kümeleme Yöntemleri: SLINK, CURE, CHAMELEON, BIRCH, CLUCDUH

- Bölümlemeli Kümeleme Yöntemleri: K-Means, PAM, CLARA, CLARANS
- Yoğunluğa Dayalı Kümeleme Yöntemleri: DBSCAN, OPTICS, DENCLUDE
- Izgara Temelli Kümeleme Yöntemleri: STING, CLIQUE
- Genetik Algoritmalar

4.1.1.1. Hiyerarşik Yöntemler

Hiyerarşik kümeleme, kümelerin ana kümede ele alınması ve aşamalı olarak alt kümelere ayrılması ya da ayrı ayrı alınan kümelerin aşamalı olarak tek küme biçiminde birleştirilmesi esasına dayanır (Özkan, 2016: 11-12).

SLINK

SLINK (An Optimally Efficient Algorithm For The Single Link Cluster Method) algoritması tek bağlantı veya en yakın komşu tekniğini kullanır (Sibson, 1973). Kümeler arasındaki mesafe tespit edilirken iki küme içinde birbirine en yakın elemanların uzaklığı ya da iki kümeyi en yakın yapan elemanların mesafesi kümeler arası mesafeyi oluşturur. Slink algoritması k-en yakın komşu algoritmasının bir türevidir. Literatürde Slink, k-en yakın komşu algoritması olarak da bilinir (Dunham, 2003:142).

CURE

CURE (Clustering Using Reprisentatives) ile kümelemede oluşturulan kümelerin kalitesini etkileyen en önemli faktör, ana veride diğer verilerden uzakta olan ve sayıları az olup ve hiçbir kümeye ait olmaması gereken uç verilerdir. CURE algoritması bu uç verilerin kümelerin kalitesini etkilememesi için geliştirilmiştir (Silahtaroğlu, 2016:169).

CURE büyük veri tabanlarını hızlı bir şekilde işlemeye çalışmaktadır. Bunu yaparken tüm nesnelerin kümelenmesi yerine tesadüfi örnekleme yaklaşımıyla çekilen örneklem üzerinde kümeleme işlemi yapar (Akpınar, 2014:50).

CHAMELEON

İki küme arasındaki benzerliği dinamik modelle belirler. İki alt kümenin benzerliği, yakınlığı bu iki kümeden her birinin kendi iç benzerlikleri ve

yakınlıkları kıyaslanarak belirlenir Chameleon'da. Karşılaştırma sonucunda iki alt küme birbirine yakınsa birleştirilir. Benzerlik mesafe matrisinin kullanılabildiği tüm veri türleri ve de kümeleri için uygulanabilecek bir algoritmadır CHAMELEON. SLINK, CURE, DBSCAN ve K-Means gibi algoritmaların yanılsamaları varken CHAMELEON bu algoritmalardan daha iyi sonuçlar vermektedir (Silahtaroğlu, 2016:169-170).

BIRCH

BIRCH (Balanced Iterative Reducing And Clustering Using Hierarchies) algoritması büyük veritabanlarının kümelenme işlemi için tasarlanmış bir algoritmadır. Gürültülü verileri kontrol etmek için öne sürülen ilk algoritma özelliğini taşır. Kümelemenin yapılabilmesi için bir ağaç oluşturulur ağaç taranarak kümeleme yapılır. Sadece sayısal veriler üzerinde analiz yapılmaktadır (Silahtaroğlu, 2016: 175-176).

CLUCDUH

CLUCDUH (Clustering Categorical Data Using Hierarchies) BIRCH'in alternatifi olan bir algoritmadır. Kök düğümden başlayarak kümelenecek verileri bir ağaç haline dönüştürür ve kümelemeyi bu ağaç üzerinden yapar. Kategorik verilerle çalışır (Silahtaroğlu, 2016:176-177).

4.1.1.2. Bölümlemeli Yöntemler

Bölümlemeli yöntemlerde, n adet nokta önceden belirlenen k adet kümeye ayrılır. Hiyerarşik yöntemlerin tersine kullanıcı tarafından bazı kriterlere göre kümeler oluşturulur ve oluşturulacak küme sayısı önceden belirlenir. Kümeler arası min-max mesafeyi ve küme iç benzerlik kriterlerini de kullanıcı verir. Benzerlik- mesafe matrisi kullanılmaz (Giudici & Silvia, 2009:83). Büyük veri tabanlı çalışmalar için daha uygundur.

K-Means

K-Means algoritması kümelerin sürekli yenilendiği ve optimum çözüme ulaşana kadar devam ettiği döngüsel bir algoritmadır. Bölümlemeli algoritmaların birçoğu K-means'den esinlenerek ya da bu algoritmanın geliştirilmesiyle meydana gelmiştir (Han, Pei, & Kamber, 2012:108). K-means algoritması eldeki veriyi k adet kümeye ve küme ortalamalarına göre kümelere ayırır. Küme sayısı, k

kullanıcı tarafından verilir. Buradaki Means (ortalama) kavramı daha önce belirtilen küme merkezini ifade eder (Silahtaroğlu, 2016:186).

PAM

PAM (Partitioning Around Medoids), k adet kümeyi bulmak için seçili temsilcilerin etrafına ana kümedeki tüm elemanları toplar ve her defasında bu temsilcileri değiştirip kümeleme yapar. PAM'ın temsilci olarak seçtiği noktaya medoid denir. Bundan ötürü bu algoritma K-Medoid algoritması olarak da bilinir. Medoid seçimi, kümenin merkezine yakın olan noktanın belirlenmesidir. K adet küme için seçilen k adet temsilci seçiminden sonra veritabanından temsilci olmayan veriler kendilerine en çok benzerlik gösteren temsilcinin etrafında yerini alır (Silahtaroğlu, 2016: 188).

CLARA

CLARA (Clustering Large Application), bütün veritabanını tarayıp temsilcileri seçmek yerine rastgele örnek küme alıp PAM algoritmasını bu örnekleme uygular. Bu işlem sonunda oluşacak kümelerin temsilcisi belirlenir ve ana kümeyi oluşturan veritabanından bir örneklem daha seçilir. Bu aşamada ilk temsilcilerin rastgele seçilmesi yerine bir önceki aşamada halihazırda belirlenmiş temsilciler kullanılır. Böylece algoritma içi temsilci değişimi azalacak, algoritma hızlı çalışacak ve kaliteli sonuçlar verecektir. PAM'a göre CLARA daha geniş veritabanlarında güvenli sonuçlar vermiştir (Silahtaroğlu, 2016: 190).

CLARANS

CLARANS (Clustering Large Applications Based on Randomized Search) n adet nesnenin temsilcilerle bir şebeke diyagramından yararlanılarak k adet kümeye ayrılmasıdır. CLARANS da CLARA gibi tüm veritabanını taramaz fakat farklı olarak taramayı belirli bir alt şebeke ile sınırlandırmaz. CLARA tarama işleminin başında örnek düğümler seçer CLARANS ise her adımda örnek komşular seçer. PAM'dan farklı olarak bir düğümün komşuları taranma aşamasındayken örnekleme yapar ve bu örnekleme dinamiktir (Silahtaroğlu, 2016:190).

4.1.1.3. Yoğunluğa Dayalı Algoritmalar

Kümeleme algoritmaları sadece küresel biçimlerde dağılan veri noktaları için etkin sonuçlar vermektedir. Milyonlarca dünya olayı, mekânsal verinin yönetimini ve keyfi biçimde kümelerin keşfedilmesi gerekmektedir. Veri noktalarının belirli noktalarda yoğunlaşması üzerine odaklanan algoritmalar yoğunluğa dayalı algoritmaları meydana getirmiştir. Bu algoritmalar parametrik değildir (Akpınar, 2014:357).

DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise) veri uzayında yoğun bölgelerin bulunup kümelenmesi, seyrek bölgelerinse parazit olarak tanımlanması fikrine dayanır. Yoğun ve seyrek bölgelerin bulunmasında kullanıcının başlangıçta tanımladığı ve tüm kümeleme sürecinde aynı kalan Yarıçap ve MinNts parametreleri kullanılır. Yarıçap içerisinde kalan noktaların sayısı MinNts değerine eşit veya büyükse bu bölge yoğun olarak nitelenir ve başlangıç kümesini oluşturur (Akpınar, 2014:359).

OPTICS

OPTICS (Ordering Points to Identify the Clustering Structure) DBSCAN'da karşılaşılan parametre değerlerinin tanımlama zorluklarına ve tüm süreç için bu değerlerin genel olarak atanmasının yol açtığı sorunlara çözüm arar (Akpınar, 2014:365).

DENCLUE

DENCLUE (Density Based Clustering) yoğunluk temelli algoritmalarda sorun yoğunluğun ne olduğu ve nasıl ölçülebileceğidir. DBSCAN ve OPTİCS algoritmalarında yoğunluk, tanımlı yarıçap içinde yer alan nesnelerin sayımı ile ölçülür. Bu farklı yarıçap değerleri sonuçları fazlaca etkilemektedir. Bu problemi aşmak için DENCLUE algoritmasıyla çekirdek yoğunluğu tahmin yaklaşımı kullanılır (Akpınar,2014:370).

4.1.1.4. Izgara Temelli Algoritmalar

Izgara temelli kümelemede çalışma uzayı sonlu sayılı hücrelere bölünerek çok boyutlu ızgara yapısı oluşturulur.

STING

STING (Statistical Information Grid) algoritmasında ele alınan alan dikdörtgen hücrelere bölünür ve hiyerarşik bir yapı ortaya çıkar. Her üst katmandaki hücre alt katmanlarda daha küçük hücrelere bölünür. Diğer kümeleme algoritmalarından farklıdır. Kümeler oluşturmak yerine ızgara biçimli hücreler oluşturur (Akpınar, 2014:373)

CLIQUE

CLIQUE (Clustering in Quest) algoritması yoğunluğa dayalı ve ızgara temelli yöntemleri bir araya getiren algoritmadır. Çok büyük veri gruplarının kümelenmesi için geliştirilmiştir. Çok boyutlu veri uzayındaki alt uzaylarda çalışır ve böylece daha iyi kümeleme yapar (Silahtaroğlu, 2016:204).

4.1.1.5. Genetik Algoritmalar

Genetik algoritmalar evrim yasalarından veya bunların öne sürümlerinin esinlenimlerinden oluşmuştur. Veri madenciliği çalışmalarında kümeleme, sınıflandırma ve örüntü tanıma analizlerinde de kullanılmaktadır. Popülasyon olarak adlandırılan ve kromozomlar tarafından temsil edilen sonuçlarla işlemler başlar. Sonuçlar kullanılarak yeni bir nüfus sonuç değeri elde edilir. Her bir yeni nüfus sonuçta da bir öncekinden daha iyi sonuçlar hedeflenir. İstenilen durma kriterine kadar bu döngü devam eder (Silahtaroğlu, 2016:205).

4.1.2. Birliktelik Kuralları

Veritabanındaki kayıtların birbirleriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceğini ortaya koyan veri madenciliği yöntemine birliktelik kuralı denir. Pazarlama alanındaki pazar sepet analizi uygulaması birliktelik kuralları mantığıyla işleyen bir analizdir. Bu tür analizlerle müşteri alışveriş alışkanlıkları öğrenilmeye çalışılır (Özkan, 2016).

Pazar sepet analizinde ürünler arasındaki ilişkileri ortaya koymak için destek ve güven ölçütlerinden bahsedilir. Bu iki ölçütün hesaplanmasında destek sayısı değerinden yararlanılır. Kural destek ölçütü ise bir ilişkinin tüm alışverişler içinde ne oranda tekrarlandığını ifade eder. Yani, kural güven ölçütü A ürününden

alanların B ürününden alma durumunu irdeler. Destek ve güven ölçütlerinin yanı sıra bu değerleri karşılamak maksadıyla eşik değer belirlenir. Hesaplanan destek ve güven ölçütlerini destek(eşik) ve güven(eşik) değerlerinden büyük olması gerekir. Hesap edilen destek ve güven ölçütleri ne kadar büyükse birliktelik kuralı da o derece güçlüdür (Özkan, 2016).

Birliktelik Kurallarının matematiksel ifadesi şu şekildedir (Agrawal, Imielinski, & Swami, Mining Association Rules between Sets of Items in Large Databases, 1993):

$I = \{I_1 , I_2 , \dots , I_m \}$ nesne kümesi

$T = \{t_1, t_2 , \dots , t_n\}$ veritabanındaki işlemler (alışverişler)

t_k 'ların alacağı değer 0 veya 1'dir. t_k =0 olursa I_k satın alınmamış, t_k=1 olursa I_k satın alınmış demek olur. Her işlem için veritabanında ayrı kayıtlar bulunur. $X \subseteq I$ için X'teki her bir I_k'ya karşı gelen t_k değeri t_k=1 dir (Silahtaroğlu, 2016:139).

Yani;

$X \Rightarrow I_j$; X, I'nın bir alt kümesidir. I_j ise I içindeki herhangi bir elamandır ve X içinde yer almamaktadır. $X \Rightarrow I_j$ kuralının T için uygun olup olmadığını söylemek için güven seviyesini açıklamak gerekir. Yani T içindeki tüm X'lerin ne kadarının I_k'yı sağladığı %c değeriyle ifade edilmelidir. Birliktelik kuralını $0 \leq c \leq 1$ güven seviyesiyle ; $X \Rightarrow I_j | c$ açıklayabiliriz. Güven seviyesi birliktelik kuralının gücünü yansıtır (Silahtaroğlu, 2016:140).

Birliktelik kurallarında diğer önemli kavram destek seviyesidir. Bu kavram T içindeki işlemlerin ne kadarının X'i sağladığını ifade eder. Veri setinde analizin yapılıp kuralların ortaya çıkartılması, kullanıcının vereceği en küçük destek seviyesi ve en küçük güven seviyesinden daha büyük güven ve destek seviyelerine sahip kuralların tespit edilmesidir. En küçük destek seviyesini sağlayan kümelere geniş nesne kümesi, diğerlerine ise küçük nesne kümesi denir (Agrawal & Srikant, 1994).

Veri madenciliğinde birliktelik kurallarıyla çözümlenen birçok algoritma bulunmaktadır. Bunlar içinde en çok bilinenleri, Apriori, ECLAT, FP-Growth'tur.

Apriori

Apriori bağlantı analizlerinin yapıldığı ve bu bağlantılardan kurallar çıkartılmada en çok bilinen ve kullanılan algoritmadır. Apriori'nin aşamaları şu şekildedir (Özkan, 2016: 219):

- Destek ve güven ölçütlerini karşılaştırmak üzere eşik değerler belirlenir. Sonuçlar bu eşik değere eşit ya da büyük olmalıdır.
- Veritabanı taranarak çözüme dâhil olacak destek sayıları hesaplanır. Bu sayılar eşik sayısı ile karşılaştırılır. Eşik destek sayısından küçük değerler çözümlemeden çıkartılır uygun olanlar ise göz önüne alınır.
- Üst adımlardaki ürünler ikişerli gruplandırılır ve destek sayıları elde edilir. Eşik destek sayısından küçükler çıkartılır.
- Üçerli, dörderli vb. gruplandırmalarla aynı işlemler tekrarlanır.
- Ürün grubu belirlendikten sonra kural destek ölçütüne bakılır, birliktelik kuralları türetilir ve ilgili güven ölçütleri hesaplanır.

ECLAT

ECLAT (Equivalence Class Transformation) algoritması set kesişimini kullanan derinlemesine bir ilk arama algoritmasıdır. Sık ürün kümelerini bulmak için basit bir algoritmadır. Veritabanını tek tek taramaz bir kez tarar. Sadece destek değerini bulur güven değeriyle çalışmaz (Zerman, 2018:53).

FP-Growth

FP-Growth (Frequent Pattern Growth) algoritması birliktelik kurallarındaki diğer algoritmalardan daha yüksek performans göstermektedir. En önemli avantajı büyük veriler için hızlı çalışmasıdır ve sistem kaynaklarının verimli kullanmasıdır. Algoritma tüm verileri FP-tree denen sıkıştırılmış bir ağaç yapısında tutar. Diğer bir özelliği de veri tabanını sadece iki kez taramasıdır. İlk taramada tüm nesnelerin destek değerlerini bulunur, ikincide ise ağaç yapısı oluşturulur. (Kiraz & Deliismail, 2018).

Algoritma adımları şu şekildedir;

- Minimum destek değeri hesaplanır

- Sıklık değerleri bulunur
- Ürünler önem derecelerine göre sıralanır
- Siparişin içindeki ürünler önem derecelerine göre sıralanır.
- FP-Tree çizilir.

4.1.3. Ardışık Zamanlı Örüntüler

Bir müşterinin süt, peynir ve ekmek alması bir örüntü oluşturur. Müşterinin birinci gün A ürünü, ertesi günlerde B ürünü alması ve takip eden günlerde C ürünü alması da bir örüntü oluşturmaktadır. Farkı yaratan zaman içinde ardışık bir örüntünün oluşmasıdır (Silahtaroğlu, 2016:55).

Perakendeciliğin ve beraberinde barkod teknolojisinin gelişimi büyük miktarda ve sık aralıklarla verilerin depolanmasını sağlamıştır. İşlem tarihleri, satılan ürünler gibi bilgiler market veritabanlarından edinilir. Kayıtlar incelenerek ardışık zaman örüntüleri belirlenir. Çeşitli ürün yerleştirme ve pazarlama faaliyetleri bu örüntüler vasıtasıyla geliştirilebilir (Silahtaroğlu, 2016:55).

4.2. Tahminleyici Modeller

Tahminleyici modellerde sonuçları bilinen verilerden yola çıkılarak bir model geliştirilmesi ve kurulan bu modelden yararlanılarak, sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesi amaçlanır (Şimşek, 2011:5). Bir banka önceki dönemlerde vermiş olduğu kredilere ilişkin gerekli tüm verilere sahip olabilir. Bu verilerde bağımsız değişken kredi alan müşterinin özellikleri, bağımlı değişken değeri ise kredinin geri ödenip ödenmediğidir. Bu verilere uygun bir şekilde kurulan model, daha sonraki kredi taleplerinde müşteri özelliklerine göre verilecek olan kredinin geri ödenip ödenmeyeceğinin tahmininde kullanılabilir. Sınıflandırma algoritmaları tahminleyici model mantığıyla işler.

4.2.1. Sınıflandırma

Sınıflandırma bir verinin özelliklerinin incelenerek hangi sınıfa dâhil olduğunun belirlenmesidir. Bağımlı değişken sayısal değildir. Veri

madenciliğinde en çok bilinen teknik olan sınıflandırma; resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti gibi birçok alanda kullanılmaktadır (Silahtaroğlu, 2016:67).

Sınıflandırma iki temel adımda çalışır. İlk adımda hangi sınıfta olduğu belirli olan veri setiyle algoritma bir model oluşturur. Buna algoritmanın öğrenme süreci adı verilir. İkinci adımda ise girdileri hangi sınıfa ait olduğu bilinmeyen veri seti önceki adımda oluşturulan model yardımı ile sınıflandırılır. Bu adıma deney aşaması denir. Deney aşamasında öğrenme sürecinde oluşturulan model ile girdilerin hangi sınıflara ait olduğu belirlenir (Kılınç, 2015:18).

Sınıflandırma ile yapılan veri madenciliği çalışmaları genel olarak şu algoritmalardan oluşmaktadır;

- Karar Ağaçları: ID3, C4.5, CART
- İstatistiğe Dayalı Algoritmalar: Naive Bayes, Regresyon, CHAID
- K En Yakın Komşu
- Yapay Sinir Ağları
- Destek Vektör Makineleri

4.2.1.1. Karar Ağaçları

Karar ağaçlarında sınıflandırma yapmak için bir ağaç oluşturulur ve veritabanındaki her bir kayıt bu ağaca uygulanır, çıkan sonuca göre de bu kayıt sınıflandırılır. Karar ağaçlarının yapısı iki adımdan oluşur. İlki ağacın kurulması, ikincisi de verilerin tek tek ağaca uygulanarak sınıflandırılmasıdır (Silahtaroğlu, 2016:68). Karar ağaçlarından ID3, C4.5 ve CART aşağıda açıklanmıştır.

ID3

ID3 değişkenler içerisinde sınıflandırma için en ayırıcı özellikli değişkeni entropi ile bulur. Entropi bilginin sayısallaştırılmasını ve beklentisizliğin maksimumlaşmasını ifade eder. ID3 veritabanı bölünmeden gelen bilgiyle veritabanı bölündükten sonraki bilgi arasındaki farkı kullanarak öncelikli düğüme ve dallanmalara karar verir. Bu aradaki farka kazanım denir. Veritabanı bölündükçe yani dallanmalar oluştukça sınıflandırmaya dair bilgiler azalacaktır (Silahtaroğlu, 2016:74).

ID3 algoritması şu şekilde açıklanır;

Veritabanından eğitim için elde edilen kayıt kümesi ele alınsın. Eğitim kümesi sınıf niteliğinin alacağı değerlere göre $C_1, C_2, \dots, C_k$ olmak üzere k sınıfa ayrıldığı düşünülsün. Bu sınıflarla ilgili olarak ortalama bilgi miktarına ihtiyaç duyulur. Aşağıdaki formülde T, sınıf değerlerini içeren küme için P_t sınıfların olasılık dağılımıdır ve şöyle hesaplanır .

$$P_t = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \dots, \frac{|C_k|}{|T|}\right)$$

$|C_i|$ ifadesi, C_i kümesindeki elemanların sayısını vermektedir. Burada $P_1 = \frac{|C_1|}{|T|}$ olasılığını ifade eder. Bu durumda T için ortalama bilgi miktarı yani entropi şu şekilde ifade edilir;

$$H(T) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{1.1}$$

Her öznitelik için sınıf niteliği göz önüne alınarak ağırlıklı ortalamalar hesaplanır ve şu bağlantı kullanılır;

$$H(X,T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \mathrm{H}(T_i) \tag{1.2}$$

Her öznitelik için bilgi kazancı ise 1.3'teki forrmül ile hesaplanır.

$$Bilgi\ Kazancı(X,T) = H(T) - H(X,T) \tag{1.3}$$

Yüksek bilgi kazancı sağlayan nitelik ayrımın yapılacağı nitelik olur. Bu nitelik ağacın bir düğümüdür. Bir önceki düğümden çıkan her bir dal için veri kümesi tekrar düzenlenir ve önceki adıma dönülür. Her adımda verideki satırlar azalır. Satır kalmayınca işlem sona erer (Özkan, 2016:48)

C4.5

ID3 algoritmasında adımların benzeri şekilde çalışır. ID3 karar ağaçlarını oluştururken kayıp verileri hesaba katarken C4.5 kayıp verileri hesaba katmaz. C4.5 kazanım oranını hesaplarken verileri eksik olmayan diğer kayıtları kullanır.

Kazanım hesaplamasında, kayıp verileri diğer veri ve değişkenler vasıtasıyla öngörerek kullanır (Silahtaroğlu, 2016: 80). Daha anlamlı kurallar çıkartır böylece. Ayrıca ID3'ten farklı olarak sayısal değerlere sahip özniteliklerle de ağaç oluşturur. Her iki algoritma da entropi ile çalışır.

CART

CART (Classification and Regression Tree) da ID3 gibi entropi ile çalışır. Fakat dallara ayırma kriterini belirlemek için ID3 ve C4.5' ten farklı bir yol izler. CART, ikili ağaçlar üretir ve hangi düğümün kök ya da düğüm olacağının yanında belirlenen düğümün hangi noktasında ikiye ayrılması gerektiğini hesaplar (Silahtaroğlu, 2016: 82).

4.2.1.2. İstatistiğe Dayalı Algoritmalar

Naive Bayes

Bayesyen sınıflandırma elde var olan hazır sınıflanmış verileri kullanarak yeni bir verinin hali hazırdaki sınıflardan herhangi birine girme olasılığını hesaplar (Silahtaroğlu, 2016:97).

X kümesi sınıfı bilinmeyen veri kümesi olsun. H ise bu X veri örneğinin C sınıfına ait olup olmadığını irdeleyen hipotez olsun. H'nin C sınıfına ait olduğu düşüncesiyle $P(H|X)$ olasılığını hesaplamak gerekir (Özkan, 2016:157).

$$P(H|X) = \frac{P(X|H)\ \mathrm{P(H)}}{P(X)} \tag{1.4}$$

Naive Bayes ya da Bayes Sınıflandırıcı denilen kavram şu şekilde açıklanır (Han, Pei, & Kamber, 2012, s.350-353);

X kümesi sınıfı bilinmeyen veri kümesi olsun. $X = \{x_1, x_2 \dots x_n\}$ öznitelik değerleri olsun. $C_1, C_2 \dots C_n$ sınıf değerleri olsun. Sınıfı belirlenecek X veri kümesinin olasılıkları formül (1.5) ile bulunur.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{1.5}$$

İşlem yükünü azaltmak için $P(X|C_i)$ olasılığı için basitleştirme yapılır. x_i değerlerinin birbirinden bağımsız olduğu kabul edilerek şu bağıntı kullanılır;

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k\,|C_i)$$

(1.6)

Bilinmeyen örnek X'i sınıflandırmak için (1.5)'te $P(C_i|X)$ içinde yer alan paydalar birbirine eşit olduğuna göre sadece paylar karşılaştırılır. Bu değerler içinde en büyük olanı seçilir bilinmeyen örneğin bu sınıfa ait olduğu belirlenir.

$$arg\ \max_{C_i}\{\ P(X|C_i)P(C_i)\}$$

(1.7)

Sonrasal olasılığı kullanan (1.7)'deki ifade en büyük sonrasal sınıflandırma yöntemi olarak bilinir. En son durum olarak Bayes sınıflandırıcısı olarak şu ifade kullanılır;

$$C_{MAP} = \arg\max_{C_i} \prod_{k=1}^{n} P(x_k\,|C_i)\ P(C_i)$$

(1.8)

Regresyon

Herhangi bir değişkenin bir veya daha çok değişkenler arasındaki ilişkinin denklem şeklinde yazılmasıdır. Bir bağımlı değişken bir bağımsız değişkenle açıklanabiliyorsa basit regresyon olarak adlandırılır. Bir bağımlı değişken birden çok bağımsız değişkenle açıklanıyorsa çoklu regresyon adını alır. Oluşturulan denklemin türüne göre ise doğrusal ve doğrusal olmayan regresyon olarak ifade edilir. Regresyon iki temelde kullanılır; veriler sınıflara göre çeşitli bölgelere ayrılır, çıktı değerinin hesabı için formüller kullanılır (Silahtaroğlu, 2016:103-104).

CHAID

CHAID (Chi-square Automatic Interaction Detector) algoritması istatistikte kullanılan ki-kare test yöntemiyle ağaç oluşturur. Hem karar ağacı hem de istatistik temelli bir algoritmadır. CART gibi ikili ağaç üretir (Silahtaroğlu, 2016:104).

4.2.1.3. K En Yakın Komşu

K-Nearest Neighbor yani KNN olarak da ifade edilir. Mesafeye dayalı bir algoritmadır. Veriler arasındaki mesafe ölçülürken öklit mesafesi kullanılır. Sınıflandırma yapılırken her bir kayıt için diğer kayıtlarla uzaklık hesaplanır. Bir kayıt için diğer kayıtlardan k tanesi gözönünde bulundurulur. k adet kayıt, dikkate alınan noktaya diğerlerine göre en yakın olandır (Silahtaroğlu, 2016:118-119).

4.2.1.4. Yapay Sinir Ağları

Biyolojik sinir ağlarının yapısından yola çıkarak geliştirilmiş bilgi işleme sistemidir. Yapay sinir ağları sınıflandırmada olduğu gibi kümeleme ve örüntü tanımada da kullanılabilir. Sınıflandırma için kullanıldıklarında algoritma öğrenme sürecini sürdürür. Çıktı katmanına ulaşabilmek için w ağırlıklarını hesaplamaya çalışır. Öğrenme için var olan veri kümesi üzerinde ağırlıkları hesapladıktan sonra eldeki verilerin diğer kısımlarını kullanarak öğrenmenin ne kadar gerçekleştiğini bulmak adına ağırlıkları test eder. Test sonucundaki ağırlıkların etkinliği doğrulanırsa öğrenme işlemi tamamlanmış olur. Tersi durumda ise w ağırlıklarında düzeltme ya da yeniden hesaplama işlemi yapılır. Öğrenme işleminden sonra gelen yeni verinin eldeki ağırlıklardan hangi sınıfa ait olduğu hesaplanabilir (Akpınar, 2014:239-240).

4.2.1.5. Destek Vektör Makineleri

Destek vektör makineleri öğrenmeyi doğrusal ya da doğrusal olmayan bir fonksiyon ile gerçekleştirir. Veriyi birbirinden ayırmak için en uygun fonksiyonu tahmin etmeyi amaçlar (Özkan, 2016:169). Eğitim süresinin uzun olmasına rağmen yüksek güvenilirlik ve lineer olmayan sınıflandırma başarısıyla DVM tercih edilen bir yöntemdir (Akpınar, 2014, s. 268).

Destek vektör makinelerinde, bir düzlemde iki sınıf verisi olduğu düşünülürse, buradaki iki veri grubu bir sınır çizgisiyle birbirinden ayrılabilir. Sınır çizgisi, iki farklı gruptaki verilerin birbirine en yaklaştığı yerde, iki veri setine en uzak olan bir orta noktadan geçer. Destek vektör makineleri, bu sınırların en az hata ile en iyi şekilde nasıl çizileceğini belirler ve buna göre

sınıflandırma yapar (Erdal, 2011:45-46). DVM çekirdek tabanlı bir yöntemdir. DVM'lere "Çekirdek Makineleri" de denmektedir. "Kernel" denilen çekirdek fonksiyonlar DVM'nin en önemli öğrenme parametreleridir.

DVM, istatistiksel öğrenme teorisine dayanır. İstatistiksel yöntemlerden üstünlüğü ise sadece istatistiksel değerlendirmelerle sınıflandırma yapmak yerine düzlemsel ayrımı esas alan geometrik yaklaşımı kullanıyor olmasıdır. Bu şekilde DVM istatistiksel dağılım tahminlerine bağlılıktan kurtulur (Kartal, 2012:68). İki farklı sınıf arasında, sınırlarında birbirine en yakın iki örneğin arasındaki mesafenin en fazla olması amaçlanır ve bunu en iyi sağlayan ayırıcı düzlem tespit edilir. Bu düzleme en iyi ayıran üstün düzlem denilir ve bunu belirleyen ve düzlemin iki yanındaki en yakın vektörlere destek vektörleri denir (Erdal, 2011:47).

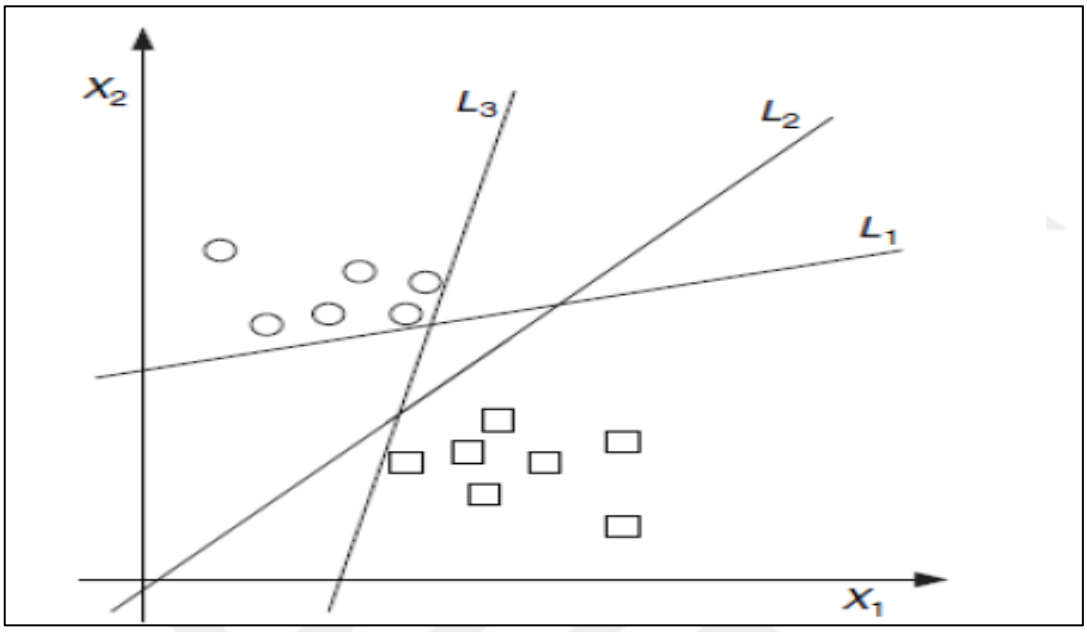
DVM'yi etkin kullanabilmek için nasıl çalıştığını iyi anlamak lazımdır. Nerede hangi çekirdek fonksiyonun kullanılması gerektiği, hangi parametrelerin kullanılması gerektiği ile ilgili doğru kararlar verilmelidir (Aydemir, 2017:28).

DVM 3 ayrı veri durumuna göre detaylandırılabilir.

- Verinin Doğrusal olarak ayrılabilme durumu
- Verinin doğrusal olarak ayrılamama durumu
- Doğrusal Olmayan Sınıflandırıcılar

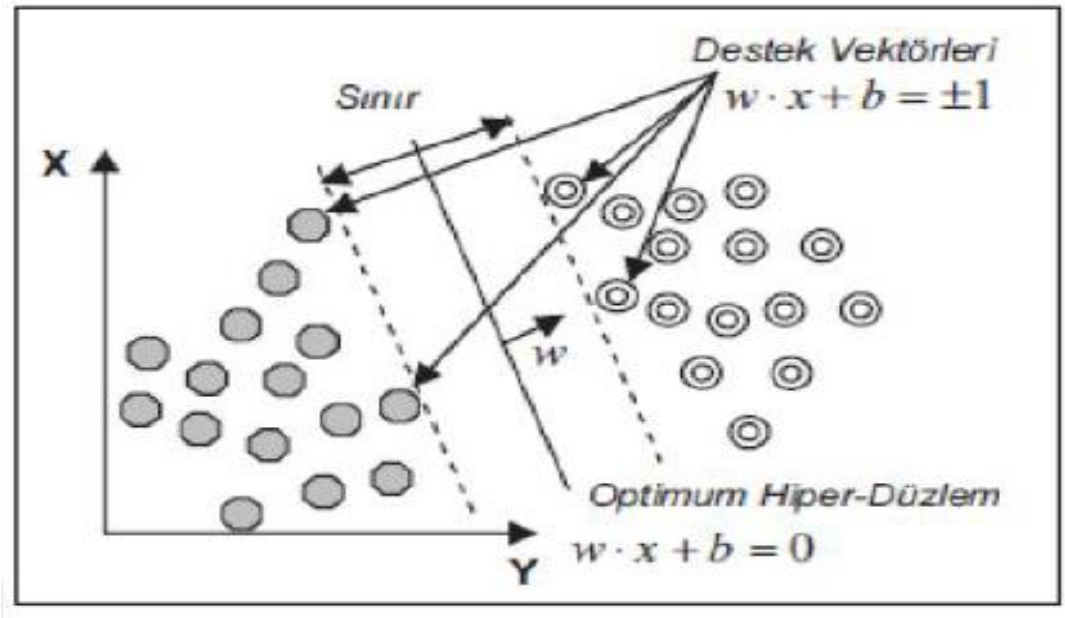
Verinin Doğrusal Olarak Ayrılabilme Durumu

DVM'nin en basit modeli doğrusal sınıflandırıcılardır.



Şekil 7.Doğrusal olarak ayrılabilen veriler (Olson & Delen, 2008:112)

Veri farklı biçimlerde doğrusal olarak ayrılmaktadır. Çok boyutlu uzayda bu doğruların yerini hiper düzlemler alır. Veriyi birbirinden ayıran bu düzlemlerden biri maksimum ayırma başarısına sahiptir. Bu başarı verisetindeki iki sınıfın birbirine en yakın noktalarını en iyi sınıflandıran ve en geniş aralığın seçilmesini ifade eder. Şekil 7'de L_2 en iyi sınıflandırmayı temsil eder.

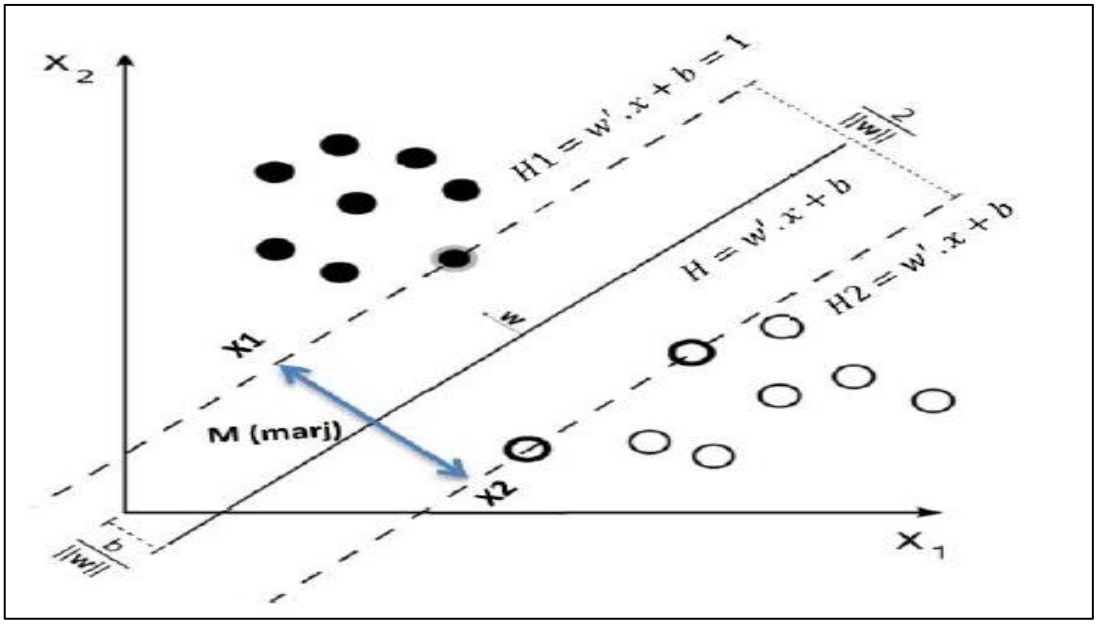


Şekil 8. Destek vektörler (Aydemir, 2017:30)

Şekil 8'de iki hiper düzlemin arasındaki orta düzlem iki sınıf veriyi ayırsan doğrusal hiper düzlemdir. Bu düzleme optimal ayırma hiper düzlemi denir. Bu düzlem şu şekilde ifade edilir (Özkan, 2016:170-172);

$$w.x + b = 0 \tag{1.9}$$

Formül (1.9)'da, x bir vektör noktası, w ağırlık vektörü ve b yan (bias) olmak üzere bir sabit sayıdır.



Şekil 9.Marj hesabı (Karakaynak, 2014)

Şekil 9'da ifade edilen örnekte yuvarlak içine alınmış gözlemler destek vektörlerdir. Bu vektörlerden geçen hiper düzlemler şu şekilde ifade edilir;

$$H_1:(w'.x) + b = +1 \tag{1.10}$$

$$H_2:(w'.x) + b = -1 \tag{1.11}$$

Bu iki düzlem arasındaki uzaklığı bulmak için düzlemler üzerinde x_1 ve x_2 noktaları alınır. Uzaklıklarını bulmak için ise şu denklemler kullanılır (Karakaynak, 2014);

$$M = H_1 - H_2 \tag{1.12}$$

$$(w'.(x_1 - x_2)) = 2 \tag{1.13}$$

$$M = \left(\frac{(w'.(x_1 - x_2))}{||w||}\right) = \frac{2}{||w||} \quad (1.14)$$

Amaç M'yi yani marjı maksimum yapmaktır. $\frac{2}{||w||}$ ifadesinin maksimize edilmesi gerekir. Bunun için ise $||w||$ ifadesinin minimum olması gerekir. $||w||$ minimum yaparken aynı zaman da veri noktalarının marjın içine düşmesini engellemek için (1.15) ve (1.16) 'da gösterilen kısıtlar da bu minimizasyon problemine eklenmelidir.

$$y = -1 \; için \; (w'.x) + b \leq -1 \quad (1.15)$$

$$y = +1 \; için \; (w'.x) + b \geq +1 \quad (1.16)$$

Şekil 9'da ayırıcı hiper düzlemin pozitif tarafında kalan ifadeler için (1.16) eşitsizliği, alt tarafında kalan ifadeler için ise (1.15) eşitsizliği kullanılabilir. Bu iki eşitsizlik (1.17)'deki gibi birleştirilebilir (Karakaynak, 2014:19-20);

$$y_i(w'.x + b) - 1 \geq 0 \quad (1.17)$$

Maksimum sınırın bulunması için;

$$min: \frac{1}{2}||w||^2 \quad (1.18)$$

$$y_i(w'.x_i + b) - 1 \geq 0 \; , \forall i \quad (1.19)$$

İşlemleri yapılır. (1.18) çözülecek problem (1.19) problemin çözümü sırasında kullanılan koşuldur. Ve bu ifade ikinci dereceden optimizasyon problemidir. Çözüm için Langrange formülasyonu yapılır. Pozitif lagrangian çarpanları a_i'ler kullanılarak dönüştürülen yeni optimizasyon problemi şu şekildedir (Karakaynak, 2014).

$$L_P = min\frac{1}{2}||w||^2 - \sum_{i=1}^{n} a_i \, y_i(w'.x_i + b) + \sum_{i=1}^{n} a_i \quad (1.20)$$

Bu formül birincil değişkenler w ve b bakımından minimize edilmeli, ikincil değişkenler bakımından maksimize edilmelidir. Bu da oldukça karışık bir çözümdür. Karush- Kuhn- Tucker koşulları ile çözüm bulmak için de formül (1.20)'nin w ve b'ye göre türevleri alınır (Karakaynak, 2014).

$$\frac{\partial L_p}{\partial_w} = 0 \rightarrow w = \sum_{i=1}^{n} a_i y_i x_i$$

(1.21)

$$\frac{\partial L_p}{\partial_b} = 0 \rightarrow w = \sum_{i=1}^{n} a_i y_i$$

(1.22)

Bu formüller (1.20)'de yerine yazıldığında;

$$L_P = \frac{1}{2}(w'w) - w' \sum_{i=1}^{n} a_i y_i x_i - b \sum_{i=1}^{n} a_i y_i + \sum_{i=1}^{n} a_i$$

(1.23)

$$L_P = -\frac{1}{2}(w'w) + \sum_{i=1}^{n} a_i$$

(1.24)

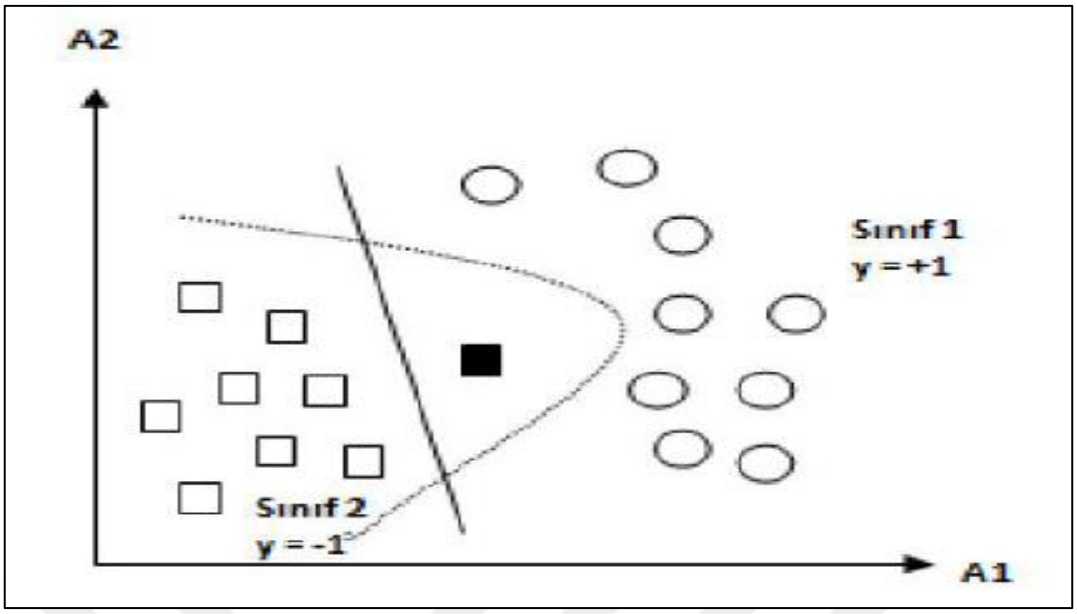
$$L_P = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j}^{n} a_i, a_j, y_i, y_j, x_i', x_j \quad a_i \geq 0 \ \forall_i$$

(1.25)

İfadesi elde edilir. Her eğitim örneğinde bir Lagrange çarpanı görülür. Çözümde elde edilen Lagrange çarpanlarının büyük çoğunluk değeri sıfır olacaktır. Kalan $a_i > 0$ değerli x_i örnekleri destek vektörlerdir. H_1 ve H_2 hiper düzlemlerinin arkasında kalırlar (Karakaynak, 2014:30-31).

Verinin Doğrusal Olarak Ayrılamama Durumu

Veriler her zaman doğrusal bir düzlem ile birbirinden ayrılmayabilir, veride gürültüler ve yanlış veri girişleri olabilir. Örnekler doğrusal olarak tamamen ayrılabilir durumda değilse problemin çözümü için pozitif zayıflık değişkenleri $\xi_i, i = 1,2, \dots, N$ kullanılır (Aydemir, 2017:33). Aşağıdaki Şekil 10'a göre denklemler zayıflık değişkenleri ile yeniden tanımlanarak oluşturulmuştur;



Şekil 10.Doğrusal olarak ayrılamama durumu (Özkan, 2016)

$$y_i = +1 \; için \; w^T . x_i + b \geq +1 - \xi_i \tag{1.26}$$

$$y_i = -1 \; için \; w^T . x_i + b \geq -1 + \xi_i \tag{1.27}$$

$$\xi_i \geq 0 \;\; , \forall_i \tag{1.28}$$

$\xi_i = 0$ olması durumunda x_i örneği doğru sınıflandırılmış, $0<\xi_i<1$ olması durumunda x_i örneği doğru sınıflandırılmış ancak marj bölgesi içerisinde yer alıyor, $\xi_i \geq 1$ ise yanlış sınıflandırılmış demektir.

Doğrusal olarak ayrılamama durumunda eğitim hatası için bir C üst sınırı eklenir. Bu üst sınır Lagrange çarpanlarının alabilecekleri maksimum değeri göstermektedir. Bu şekilde Lagrange çarpanlarının $0 \leq a_i \leq C$ aralığında kalması sağlanır. Böylece Lagrange formülü şu şekle dönüşür (Yakut, 2012:47);

$$L_P = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N} a_i\{y_i(w^T . x_i + b) - 1 + \xi_i\} - \sum_{i=1}^{N}\mu_i\,\xi_i \tag{1.29}$$

$\mu_i\,\xi_i$ pozitif olması için kullanılan Lagrange formülasyonunda çözülmesi zordur bu yüzden dual problemine dönüştürülür. Karush-Kuhn-Tucher şartları uygulanırsa (Yakut, 2012:48);

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{N} a_i\, y_i x_i = 0 \tag{1.30}$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{N} a_i\, y_i = 0 \tag{1.31}$$

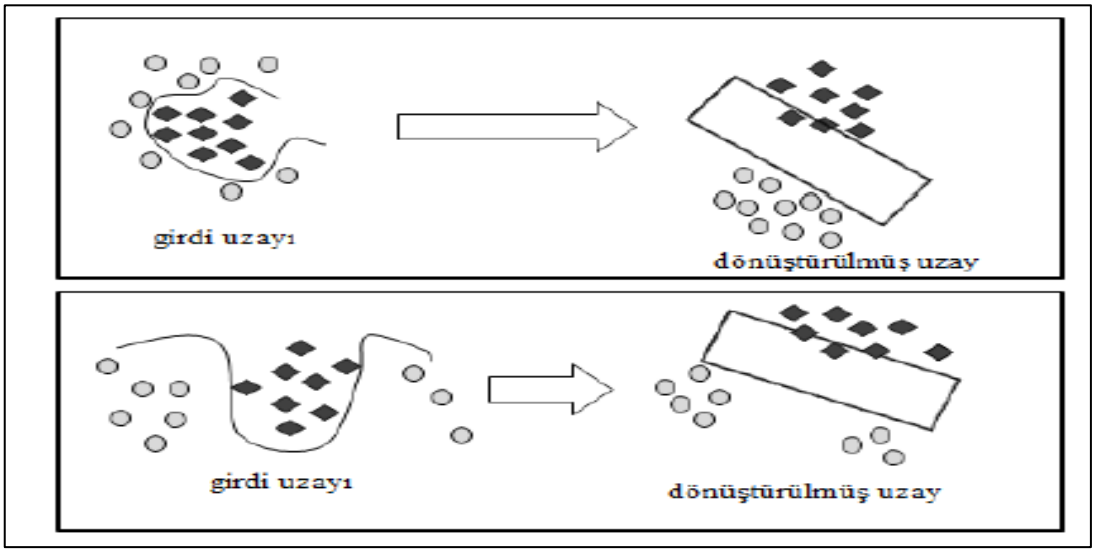
$$\frac{\partial L_p}{\partial \xi} = C - a_i - \mu_{\dot{1}} = 0 \tag{1.32}$$

$$L_p = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j}^{N} a_i, a_j, y_i, y_j, x_i^T, x_j \; 0 \le a_i \le C, \forall_i \tag{1.33}$$

İfadeler düzenlenirse problem çözümünde $0 \le a_i \le C$ aralığındaki Lagrange çarpanlarına karşılık gelen x_i değerleri destek vektörlerdir (Yakut, 2012:48).

Doğrusal Olmayan Sınıflandırıcılar

Uygulamada her zaman yukarıda anlatıldığı gibi verilerin doğrusal olarak ayrılabildiği durumlarla karşılaşılmaz. Şekil 11'deki gibi iki sınıf iç içe ya da veri grupları arasında kalmış gibi bir yapıda olabilir (Uçar, 2013:39).



Şekil 11.Doğrusal ayrılamayan veriler (Uçar, 2013:39)

Destek Vektör Makineleri böyle durumlarla karşılaştığında verileri doğrusal olarak ayırabileceği n boyutlu girdi uzayından daha yüksek boyuta sahip olay uzayına taşır.

$$x \in R^n \rightarrow \Phi(x) \epsilon R^f \tag{1.34}$$

Doğrusal olmayan DVM, verilerin taşındığı bu yeni boyutta doğrusal DVM gibi çalışarak verileri ayıracak optimum çoklu düzlem arar.

Dönüştürme işlemi için kullanılacak fonksiyon $\Phi(x)$, x'lerin $\Phi(x)$'e dönüşümünü sağlayan sabit bir fonksiyondur. Girdi uzayını oluşturan x vektörü x_i gözlemlerinden oluşurken, özellik uzayını oluşturulan $\Phi(x)$ vektörü $\Phi_i(x)$'lerden oluşmaktadır. Böylelikle Φ uzayında x görüntülerinin doğrusal olarak sınıflandırılabildiği bir ortam bulunması amaçlanmaktadır. Buradan hareketle dönüştürülmüş uzayda kullanılacak karar fonksiyonu (Uçar, 2013:40):

$$< w, \Phi(x) > + b = 0 \quad (1.35)$$

Şeklinde olacaktır. Destek vektörlerinin üzerinde yer aldığı ve ayırıcı çoklu düzleme paralel doğruların ayırdığı veriler aşağıdaki şekilde sınıflanır:

$$< w, \Phi(x) > + b \geq +1 \quad (1.36)$$

$$< w, \Phi(x) > + b \leq -1 \quad (1.37)$$

Aynı şekilde nesne fonksiyonu ve buna ilişkin formül şu şekilde olur:

$$min_{w,b} T(w) = \frac{1}{2}||w||^2 \quad (1.38)$$

$$y_i(< \Phi(\mathrm{x}).\mathrm{w} > + \mathrm{b}) - 1 \geq 0, \forall i \quad (1.39)$$

Burada iki sorun ortaya çıkmaktadır. İlk olarak dönüştürülmüş uzayda oluşturulacak doğrusal karar sınırı ile ilgili nasıl bir haritalama fonksiyonu kullanılacağı açık değildir. İkinci sorun ise uygulanan haritalama fonksiyonu biliniyorsa, kurulan optimizasyon probleminin yüksek boyutlu olay uzayında çözümü zor ve karmaşık hesaplamalar gerektirecektir. w ve b parametrelerini hesaplamak için:

$$w = \sum a_i y_i \Phi(x_i) \quad (1.40)$$

$$f(x) = (\sum a_i y_i \Phi(x_i).\Phi(\mathrm{x})) + b = 0 \quad (1.41)$$

Bu denklemler dönüştürülmüş uzaydaki iki vektörün iç çarpımını içermektedir. Boyut sorunundan dolayı bu iç çarpımların hesaplanması zordur. Bu sorunu önlemek amacıyla çekirdek düzenlenmesi denen Kernel Trick yöntemi önerilmiştir (Uçar, 2013:40-41).

Çekirdek Düzenlemesi ve Çekirdek Fonksiyonları

Çekirdek fonksiyonu kullanmanın avantajı, bütün değerlerin tekrar tekrar iç çarpım değerlerinin hesaplanarak bulunması yerine doğrudan çekirdek fonksiyonunda değerin yerine koyularak nitelik uzayındaki değerinin

bulunmasıdır. Bu sayede, son derece yüksek boyutlu bir nitelik uzayı ile uğraşma olasılığı kalmaz. Diğer avantajı ise, direk girdi uzayındaki veriler kullanılacağı için Φ haritalama fonksiyonun kesin olarak ne olduğunu bilmeye gerek yoktur (Uçar, 2013:41).

İç çarpımlar iki girdi vektörü arasındaki benzerliğin bir ölçüsüdür. Çekirdek fonksiyonu da veriyi kullanarak dönüştürülmüş uzayda bir benzerlik hesaplaması yapar. Dönüştürülmüş uzaydaki iki girdi vektörü u ve v için iç çarpımlar (Uçar, 2013:41):

$$\Phi(u)\Phi(v) = (u_1^2, u_2^2, \sqrt{2}\,u_1, \sqrt{2}\,u_2, 1)(v_1^2, v_2^2, \sqrt{2}\,v_1, \sqrt{2}\,v_2, 1)=(uv+1)^2 \quad (1.42)$$

Dönüştürülmüş uzaydaki iç çarpımlar orijinal nitelik verisinden hesaplandığı için bu benzerlik fonksiyonu K ile gösterilen "çekirdek fonksiyon" olarak adlandırılır (Uçar, 2013:41).

$$K(u,v) = \Phi(u)\Phi(v)=uv+1)^2 \quad (1.43)$$

DVM yaygın olarak kullanılan dört çekirdek fonksiyonu vardır. Bu fonksiyonlar: Doğrusal Fonksiyon, Polinomiyal Fonksiyon, Sigmoid Fonksiyon, Radyal Tabanlı Fonksiyon

Doğrusal Fonksiyon

Girdi uzayında veriler doğrusal olarak ayrılabiliyor ise veriyi yüksek boyuta taşımaksızın doğrusal çekirdek fonksiyonu yardımıyla sınıflama işlemi yapılır. Bu fonksiyon herhangi bir boyut değeri ya da katsayı içermemektedir (Uçar, 2013:42).

$$K(x_i, x_j) = x_i^T x_j \quad (1.44)$$

Polinomiyal Fonksiyon

Polinomiyal çekirdek fonksiyon, d gibi belirli bir derecede girdi vektörlerinin iç çarpımından oluşmaktadır. Fonksiyonun matematiksel gösterimi aşağıdaki şekildedir:

$$K(x_i, x_j) = (x_i, x_j)^d \quad (1.45)$$

d=1 olduğunda polinomiyal fonksiyon doğrusal fonksiyona dönüşmektedir (Uçar, 2013:42).

Sigmoid Fonksiyon

Sigmoid fonksiyon k ve δ gibi iki parametre içermektedir. Kaynaklar belirli parametreler için sigmoid radyal tabanlı fonksiyon çalıştığını göstermektedir (Uçar, 2013:42).

$$K(x_i, x_j) = \tan h(kx_1, x_j - \delta \quad (1.46)$$

Radyal Tabanlı Fonksiyon

Radyal tabanlı fonksiyon çekirdek fonksiyonlar arasında kullanımı en yaygın çekirdek fonksiyondur. R programında sistem standart çıktılarını radyal tabanlı fonksiyona göre vermektedir. yarıçap kontrolünü sağlayan parametredir (Uçar, 2013:42).

$$K(x_i, x_j) = \exp\left(-y\left||x_i - x_j\right|\right|^2\right),\ y > 0 \quad (1.47)$$

İKİNCİ BÖLÜM

İŞLETME BÖLÜMÜ LİSANS DERSLERİNİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ANALİZİ

2.1. Amaç

Araştırmanın Amacı ve Önemi

Bu çalışmanın amacı İşletme bölümü lisans dersleri arasındaki ilişkilerin öğrencilerin notları bakımından incelenmesidir. Bu amaçla öğrencilerin ders ve not ilişkileri incelenip analiz edilmiştir. Mezun öğrencilerin verileriyle yapılan bu çalışma sayesinde gelecekte mezun olacak öğrencilerin derslerine ve notlarına olumlu müdahaleler yapabilmek mümkün olacaktır. Bu şekilde hem öğrenci başarısı hem de okul başarısı olumlu yönde etkileşim gösterebilecektir.

İktisadi ve İdari Bilimler Fakültesi, İşletme lisans eğitiminin amacı, esnek bir düşünce yapısına sahip ve giderek artan küresel rekabete uyum sağlayabilen öğrenciler yetiştirmektir. Program, işletmecilik ve sosyal bilimlerle ilgili derslerin yanı sıra, öğrencilerin istatistik, sayısal yöntemler ve bilgisayar konularında da yetişmesini amaçlamaktadır. Programın amacı, öğrencileri kamu ve özel sektördeki iş hayatına uzman veya araştırmacı olarak hazırlamaktır. Lisans Programı; Muhasebe ve Finansman, Üretim Yönetimi ve Pazarlama, Yönetim ve Organizasyon, Sayısal Yöntemler, Ticaret Hukuku ve Kooperatifçilik Anabilim dalları dâhilinde açılan derslerden oluşmaktadır.

İşletme programında mevcut olan derslerin tümünü başarıyla tamamlamak ve 4.00 üzerinden en az 2.00 ağırlıklı not ortalamasına sahip olmak mezuniyet için gerekli yeterlilik koşuludur. Öğrenciler 8 yarıyılda 38 zorunlu, 12 seçmeli ders almaktadır. Dönemlik olarak gerçekleştirilen derslerde bir ara sınav ve bir genel sınav yapılmaktadır. Ara sınavının %40'ı ve genel sınavının %60'ı toplanarak öğrencinin başarı notu belirlenmektedir.

Öğrencilerin okula kabul edilmesi için Öğrenci Seçme Ve Yerleştirme Kurumunun yaptığı sınavlardan başarıyla geçmesi gerekmektedir. İşletme bölümü eşit ağırlık sınav puanıyla alım yapan bir bölümdür. Yani Türkçe ve Matematik derslerinden başarı isteyen bir bölümdür.

Araştırmanın Sınırlılıkları

Çalışmada sadece mezun olan öğrenciler örneklemi oluşturmuştur. Okuldan geçiş yapan, kayıt sildiren, okuldaki eğitimine devam eden öğrenciler örnekleme dâhil edilmemiştir.

Örneklem

Çalışmanın örneklemini Recep Tayyip Erdoğan Üniversitesi, İktisadi İdari Bilimler Fakültesi, İşletme bölümüne 2009'dan itibaren kayıt yaptıran ve 2018-2019 güz yarıyılına kadar mezun olan öğrenciler oluşturmuştur. Toplamda 731 öğrencinin bilgileri incelenmiştir. Veriler üniversite bilgi işlem merkezinden talep edilmiştir. Elde edilen veriler ön işlemlerden geçirilmiş, temizlenmiş ve çalışmaya uygun hale getirilmiştir.

Veri Toplama Aracı

Öğrencilere ait ders ve diğer bilgiler üniversite Bilgi İşlem Merkezi vasıtasıyla öğrenci veritabanından çekilmiştir.

2.2. Veri Toplama Süreci

Öğrencilerin ders bilgileri, mezuniyet bilgileri, birinci öğrenim mi ikinci öğrenim mi oldukları bilgisi, doğum yerleri, cinsiyetleri, okula kayıt tarihleri, mezuniyet tarihleri, bölümü kazanmalarında ölçüt olan Öğrenci Seçme Sınavı puan bilgileri ve bu sınavda Türkiye geneli sıralama bilgileri bilgi işlem merkezinden alınan bilgilerdir.

Öğrenci Verilerine Ait Tanımlayıcı Bilgiler

Tablo 1

Öğrencilerin Cinsiyet Dağılımları

Cinsiyet	Frekans (f)	Yüzde (%)
Erkek	290	39,7
Kadın	441	60,3
Toplam	731	100

Tablo 1'de örneklemi oluşturan öğrencilerin %60,3'ünün kadın olduğu, %39,7'inin erkek olduğu görülmektedir.

Tablo 2

Öğrencilerin Doğum Yerleri

Doğum Yeri	Frekans (f)	Yüzde (%)
Akdeniz	62	8,5
Doğu Anadolu	83	11,4
Ege	42	5,7
Güneydoğu Anadolu	15	2,1
İç Anadolu	66	9,0
Karadeniz	348	47,6
Marmara	111	15,2
Yurtdışı	4	0,5
Toplam	731	100

Tablo 2'de öğrencilerin nüfusa kayıtlı oldukları yerler görülmektedir. Okula en fazla gelenin olduğu bölge Karadeniz bölgesidir. Bunda okulun bu bölgede olması etkili olmuştur denebilir.

Tablo 3

Öğrencilerin Öğrenim Durumlarına Göre Dağılımlar

Öğrenimi	Frekans (f)	Yüzde (%)
Gece	339	46,4
Gündüz	392	53,6
Toplam	731	100

Tablo 3'te gündüz eğitim gören öğrenciler örneklemin %53,6'sını, gece eğitim gören öğrenciler ise %46,4'ünü oluşturmaktadır. Burada belirtilmesi gereken ise 2009 ilk kayıt yılında okulun gece eğitimi için bölüme öğrenci almadığıdır.

Tablo 4

Öğrencilerin Okula Kayıt Yıllarına Göre Dağılımları

Kayıt Yılı	Frekans(f)	Yüzde(%)
2009	41	5,6
2010	148	20,2
2011	166	22,7
2012	162	22,2
2013	119	16,3
2014	80	10,9
2015	7	1,0
2016	8	1,1
Toplam	731	100

Tablo 4'te yıllara göre okula kayıt yaptıran öğrenci sayıları görülmektedir. İşletme bölümü en çok öğrenci alımını 2011 yılında yapmıştır. Bunu 2012 yılı takip etmektedir. 2015 ve 2016 yıllarında az öğrenci görülmesi örneklem kapsamını sadece mezun öğrencilerin almasıdır. Bu yıllarda kayıt olan normal öğrencilerin mezuniyet tarihi 2019 ve 2020 bahar dönemidir.

Tablo 4'te görülen 2015 ve 2016 kayıtlı öğrenciler ise geçiş yapan öğrencileri ifade etmektedir.

Tablo 5

Öğrencilerin Mezuniyet Yıllarına Göre Dağılımları

Mezuniyet Yılı	Frekans(f)	Yüzde(%)
2013	34	4,7
2014	114	15,6
2015	172	23,5
2016	199	27,2
2017	101	13,8
2018	97	13,3
2019	14	1,9
Toplam	731	100

Tablo 5'te öğrencilerin mezuniyet tarihleri görülmektedir. En fazla mezun 2016 yılında verilmiştir. 2019 yılındaki mezunlar 2018-2019 eğitim yarıyılı güz döneminde mezun olan öğrencileri göstermektedir.

Tablo 6

Öğrenci Notlarının Harflik, Yüzlük ve Dörtlük Sistemde Karşılıkları

Harflik sistem	Yüzlük sistem	Dörtlük sistem
AA	90-100	4
BA	80-89	3,5
BB	70-79	3
CB	60-69	2,5
CC	50-59	2
DC	45-49	1,5

Tablo 6'da öğrencilerin derslerini geçmeleri ve mezun olabilmeleri için gereken ders notları görülmektedir. Analizlerde öğrencilerin notları harf notu cinsinden incelenmiştir. Tablodaki DC harf notu ortalama ile ders geçmeyi ifade eder. Yani öğrencinin ortalaması 2 ve üzerindeyse ve de herhangi bir dersten DC almışsa ortalama ile geçebilir ve mezun olabilir.

Öğrencilere ait ders ve mezuniyet bilgilerinde yer alan, öğrenci ders harf notları, dersi kaç kere aldığı bilgisi ile elde edilen bilgilerin özet halleri ' Öğrenci Ders Notlarının Frekans ve Yüzdelik Değerleri ve Öğrencilerin Ders Notları ve Dersi Kaç Kere Aldığı Bilgisiyle Ortaya Çıkan Grafiksel Özetler ' başlıklarında açıklanmıştır.

Öğrenci Ders Notlarının Frekans ve Yüzdelik Değerleri

Tablo 7

1.Sınıf 1.Dönem Dersleri

Harf Notları	Atatürk İlkeleri ve İnkılap Tarihi-I		Temel Bilgi Teknolojileri		İşletme Bilimine Giriş		Genel Muhasebe-I		Genel Matematik-I		Türk Dili-I		Yabancı Dil-I	
	(f)	%	(f)	%	(f)	%	(f)	%	(f)	%	(f)	%	(f)	%
AA	30	4,1	137	18,7	17	2,3	35	4,8	30	4,1	32	4,4	17	2,3
BA	76	10,4	50	6,8	42	5,7	50	6,8	37	5,1	49	6,7	41	5,6
BB	146	20,0	54	7,4	135	18,5	122	16,7	85	11,6	147	20,1	100	13,7
CB	230	31,5	161	22,0	244	33,4	189	25,9	190	26,0	280	38,3	216	29,5
CC	246	33,7	314	43,0	289	39,5	321	43,9	370	50,6	222	30,4	354	48,4
DC	3	0,4	15	2,1	4	0,5	14	1,9	19	2,6	1	0,1	3	0,4
Toplam	731	100	731	100	731	100	731	100	731	100	731	100	731	100

Tablo 7'de öğrencilerin birinci sınıfın birinci döneminde aldığı dersler görülmektedir. Türk Dili 1 dışındaki bütün derslerden en çok alınan not CC'dir. Sadece Türk Dili 1'den öğrencilerin çoğu CB ile derslerini geçmiştir.

Tablo 8

1.Sınıf 2.Dönem Dersleri

Harf Notları	Atatürk İlkeleri ve İnkılap Tarihi-II		Hukukun Temel Kavramları		Genel Muhasebe-II		Davranış Bilimleri		Genel Matematik-II		Türk Dili-II		Yabancı Dil-II	
	(f)	%	(f)	%	(f)	%	(f)	%	(f)	%	(f)	%	(f)	%
AA	30	4,1	13	1,8	13	1,8	11	1,5	66	9,0	41	5,6	32	4,4
BA	76	10,4	28	3,8	28	3,8	34	4,7	77	10,5	135	18,5	69	9,4
BB	146	20,0	71	9,7	76	10,4	114	15,6	112	15,3	230	31,5	145	19,8
CB	230	31,5	204	27,9	152	20,8	226	30,9	170	23,3	194	26,5	211	28,9
CC	246	33,7	404	55,3	443	60,6	330	45,1	293	40,1	131	17,9	269	36,8
DC	3	0,4	11	1,5	19	2,6	16	2,2	13	1,8	-	-	5	0,7
Toplam	731	100	731	100	731	100	731	100	731	100	731	100	731	100

Tablo 8'de öğrencilerin birinci sınıf ikinci dönem dersleri görülmektedir. Bu tabloda da Türk Dili 2 dışındaki bütün derslerden en fazla alınan not CC olmuştur. Türk Dili 1'den alınan en fazla not ise BB'dir. Ayrıca Türk Dili 2 dersinden DC ile mezun olan öğrenci olmamıştır.

Tablo 9

2.Sınıf 1.Dönem Dersleri

	Mikro İktisat		İşletme Yönetimi		İstatistik-I		Türk Vergi Sistemi	
	(f)	%	(f)	%	(f)	%	(f)	%
AA	8	1,1	16	2,2	13	1,8	18	2,5
BA	13	1,8	50	6,8	28	3,8	39	5,3
BB	50	6,8	145	19,8	98	13,4	101	13,8
CB	158	21,6	231	31,6	187	25,6	243	33,2
CC	458	62,7	286	39,1	379	51,8	326	44,6
DC	44	6,0	3	0,4	26	3,6	4	0,5
Toplam	731	100	731	100	731	100	731	100

Tablo 9'da öğrencilerin ikinci sınıf birinci dönemde aldığı dersler görülmektedir. Bu dönemde öğrenciler en fazla CC notu almışlardır.

Tablo 10

2.Sınıf 2.Dönem Dersleri

	Makro İktisat		Yönetim ve Organizasyon		Maliyet Muhasebesi		İstatistik-II	
	(f)	%	(f)	%	(f)	%	(f)	%
AA	13	1,8	16	2,2	32	4,4	19	2,6
BA	34	4,7	42	5,7	53	7,3	20	2,7
BB	77	10,5	116	15,9	82	11,2	76	10,4
CB	175	23,9	264	36,1	206	28,2	154	21,1
CC	412	56,4	290	39,7	341	46,6	417	57,0
DC	20	2,7	3	0,4	17	2,3	45	6,2
Toplam	731	100	731	100	731	100	731	100

Tablo 10'da öğrencilerin ikinci sınıf ikinci dönemde aldıkları dersler görülmektedir. Bu dönemde de en fazla alınan ders notu CC'dir.

Tablo 11

3.Sınıf 1.Dönem Dersleri

	Finansal Yönetim-I		Pazarlama İlkeleri		Üretim Yönetimi-I		Yöneylem Araştırması-I	
	(f)	%	(f)	%	(f)	%	(f)	%
AA	21	2,9	35	4,8	17	2,3	35	4,8
BA	50	6,8	97	13,3	30	4,1	49	6,7
BB	98	13,4	158	21,6	81	11,1	96	13,1
CB	155	21,2	212	29,0	192	26,3	188	25,7
CC	383	52,4	223	30,5	387	52,9	340	46,5
DC	24	3,3	6	0,8	24	3,3	23	3,1
Toplam	731	100	731	100	731	100	731	100

Tablo 11'de öğrencilerin üçüncü sınıf birinci dönemde aldıkları toplam ders notları görülmektedir. Bu dönemde alınan en fazla not CC'dir.

Tablo 12

3.Sınıf 2.Dönem Dersleri

	Finansal Yönetim-II		Pazarlama Stratejileri		Üretim Yönetimi-II		Yöneylem Araştırması-II	
	(f)	%	(f)	%	(f)	%	(f)	%
AA	30	4,1	39	5,3	13	1,8	51	7,0
BA	49	6,7	78	10,7	21	2,9	66	9,0
BB	141	19,3	138	18,9	62	8,5	94	12,9
CB	201	27,5	194	26,5	151	20,7	164	22,4
CC	277	37,9	274	37,5	445	60,9	338	46,2
DC	33	4,5	8	1,1	39	5,3	18	2,5
Toplam	731	100	731	100	731	100	731	100

Tablo 12'de öğrencilerin üçüncü sınıf ikinci dönemde aldıkları dersler görülmektedir. Bu dönemde en fazla alınan not CC'dir.

Tablo 13

4.Sınıf 1.Dönem Dersleri

	Finansal Tablolar Analizi		Muhasebe Denetimi		Pazarlama Araştırmaları		Araştırma Yöntemleri	
	(f)	%	(f)	%	(f)	%	(f)	%
AA	50	6,8	73	10,0	46	6,3	26	3,6
BA	67	9,2	140	19,2	63	8,6	63	8,6
BB	148	20,2	187	25,6	130	17,8	109	14,9
CB	193	26,4	190	26,0	210	28,7	220	30,1
CC	265	36,3	140	19,2	261	35,7	303	41,5
DC	8	1,1	1	0,1	21	2,9	10	1,4
Toplam	731	100	731	100	731	100	731	100

Tablo 13'te öğrencilerin dördüncü sınıf birinci dönemde aldıkları dersler görülmektedir. Bu dönemde Muhasebe Denetimi dışındaki tüm derslerden not çoğunluğu CC üzerinedir. Muhasebe denetiminde ise CB'li notlar en fazla alınan nottur.

Tablo 14

4.Sınıf 2.Dönem Dersleri

	Sermaye Piyasası Analizleri		Ticaret Hukuku		Stratejik Yönetim		İnsan Kaynakları Yönetimi	
	(f)	%	(f)	%	(f)	%	(f)	%
AA	22	3,0	94	12,9	64	8,8	23	3,1
BA	43	5,9	172	23,5	188	25,7	57	7,8
BB	113	15,5	191	26,1	225	30,8	124	17,0
CB	214	29,3	169	23,1	163	22,3	172	23,5
CC	333	45,6	101	13,8	85	11,6	338	46,2
DC	6	0,8	4	0,5	6	0,8	17	2,3
Toplam	731	100	731	100	731	100	731	100

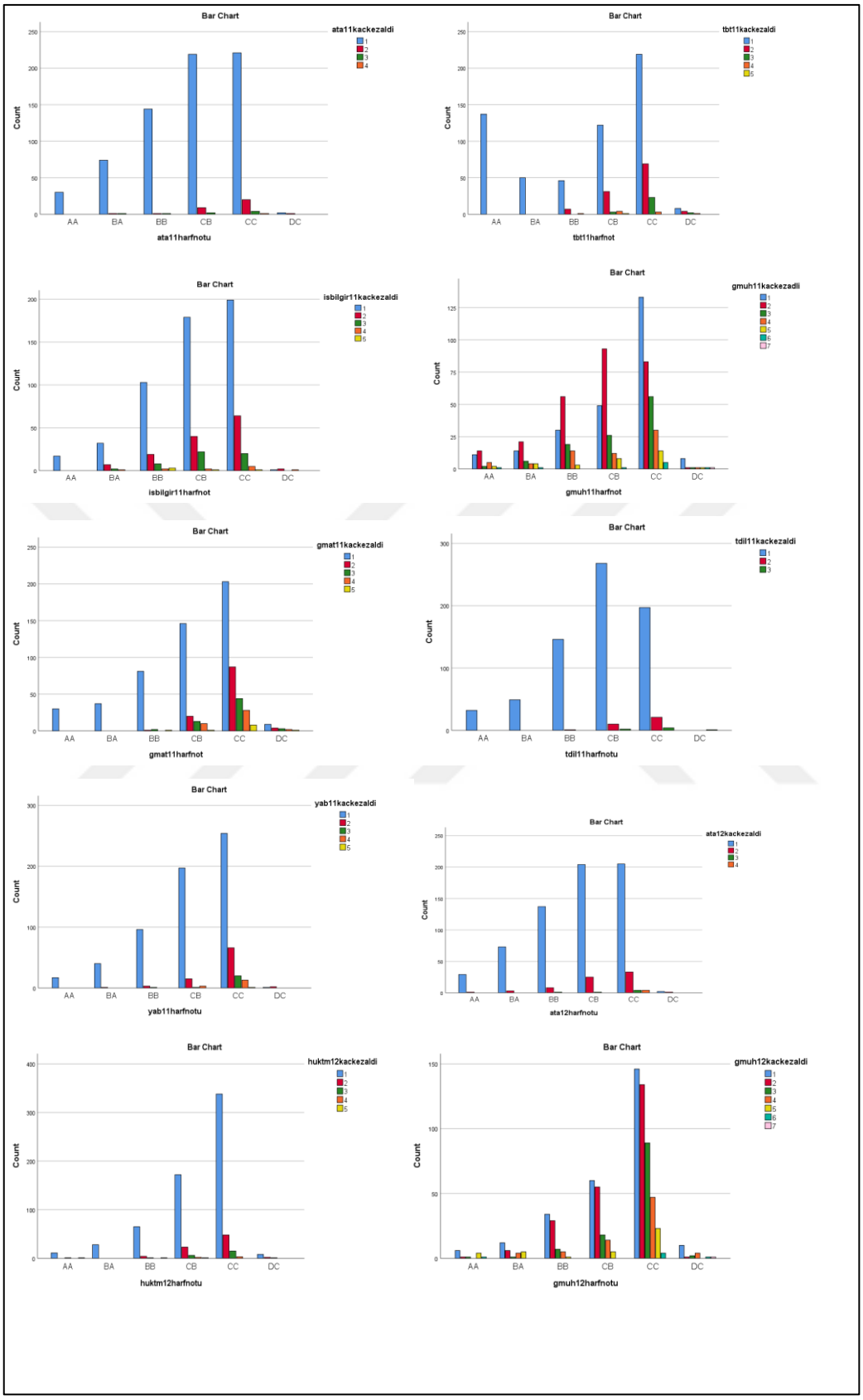
Tablo 14'te öğrencilerin dördüncü sınıf ikinci dönemde aldığı dersler görülmektedir. Bu dönemde Sermaye Piyasası Analizleri ve İnsan Kaynakları Yönetiminden en çok alınan ders notu CC'dir. Fakat Ticaret Hukuku ve Stratejik Yönetim derslerinden alınan en fazla not BB'dir.

Öğrencilerin Ders Notları ve Dersi Kaç Kere Aldığı Bilgisiyle Ortaya Çıkan Grafiksel Özetler

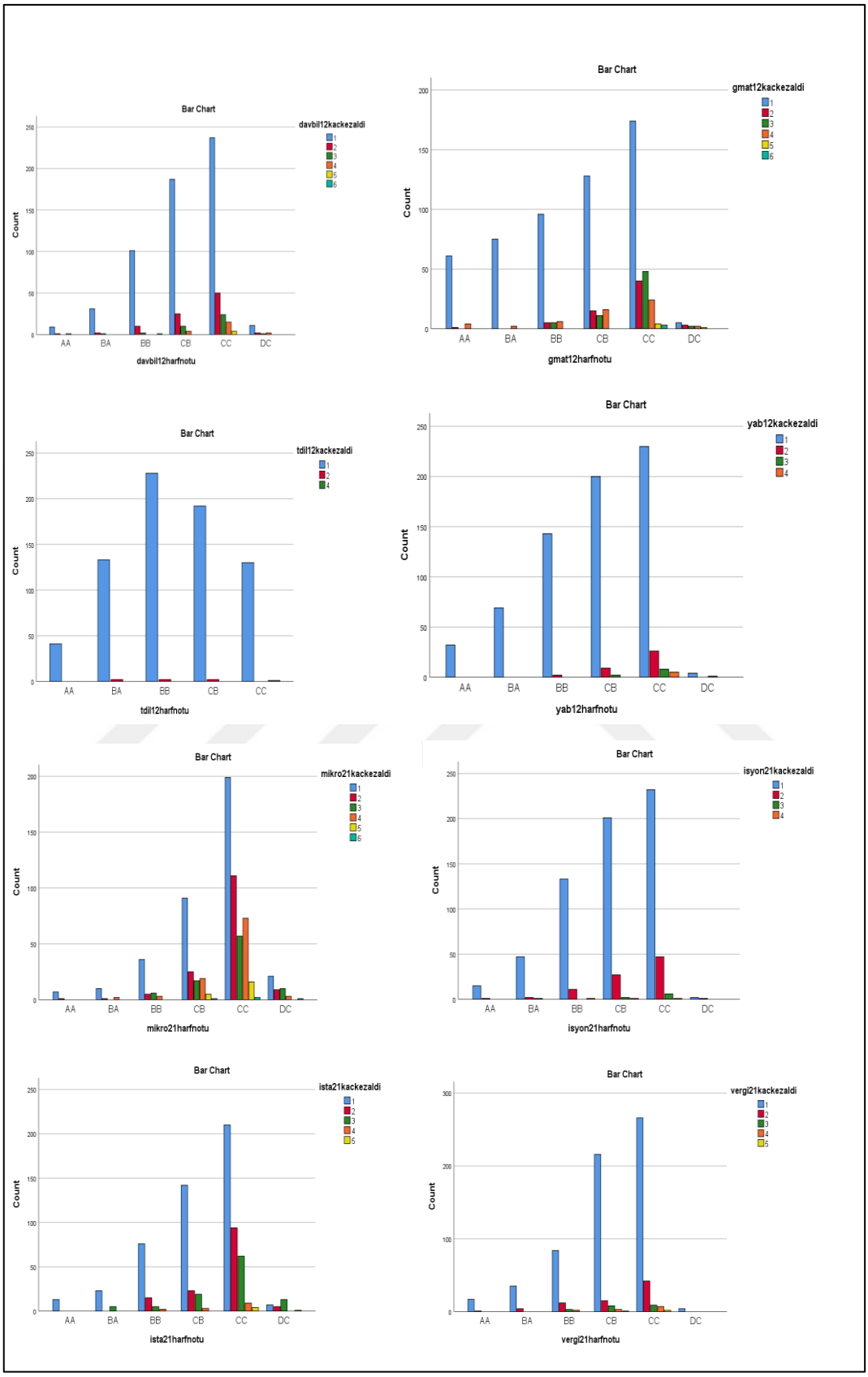
731 öğrencinin tümünün her bir dersten aldığı toplam harf notu değerleri ve dersi kaç kere aldığı bilgisine dayanarak SPSS veri analizi programının Çapraz Tablo özelliği kullanılarak Şekil 12,13,14,15'teki grafikler elde edilmiştir. Şekil 12' de en çok kez ders alımının Genel Muhasebe 1 ve 2 derslerinde olduğu görülmektedir. Düşük bir yoğunlukta olsa da öğrenciler arasından dersi 7 kez alan öğrenciler bulunmaktadır. Şekil 13'te ise Genel Matematik 2 İstatistik 1 ve Makro İktisat derslerinde öğrencilerin bu dersleri birçok defa aldığı görülmektedir. Şekil 14'teki derslerde göze çarpan derslerde genel olarak 1 veya 2 defa alımın yoğunlukta olmasıdır. Şekil 15'teki derslerde genel olarak 1 defa ders alımın yoğunluktadır.

Grafiklerdeki renkler dersin kaç kere alındığını ifade etmektedir. Mavi rengin yoğun olması öğrencilerin genelinin ilk defada derslerini geçtiklerini ifade etmektedir. Mavi rengin ardından görülen kırmızı, dersi ikinci defada geçenleri ifade etmektedir.

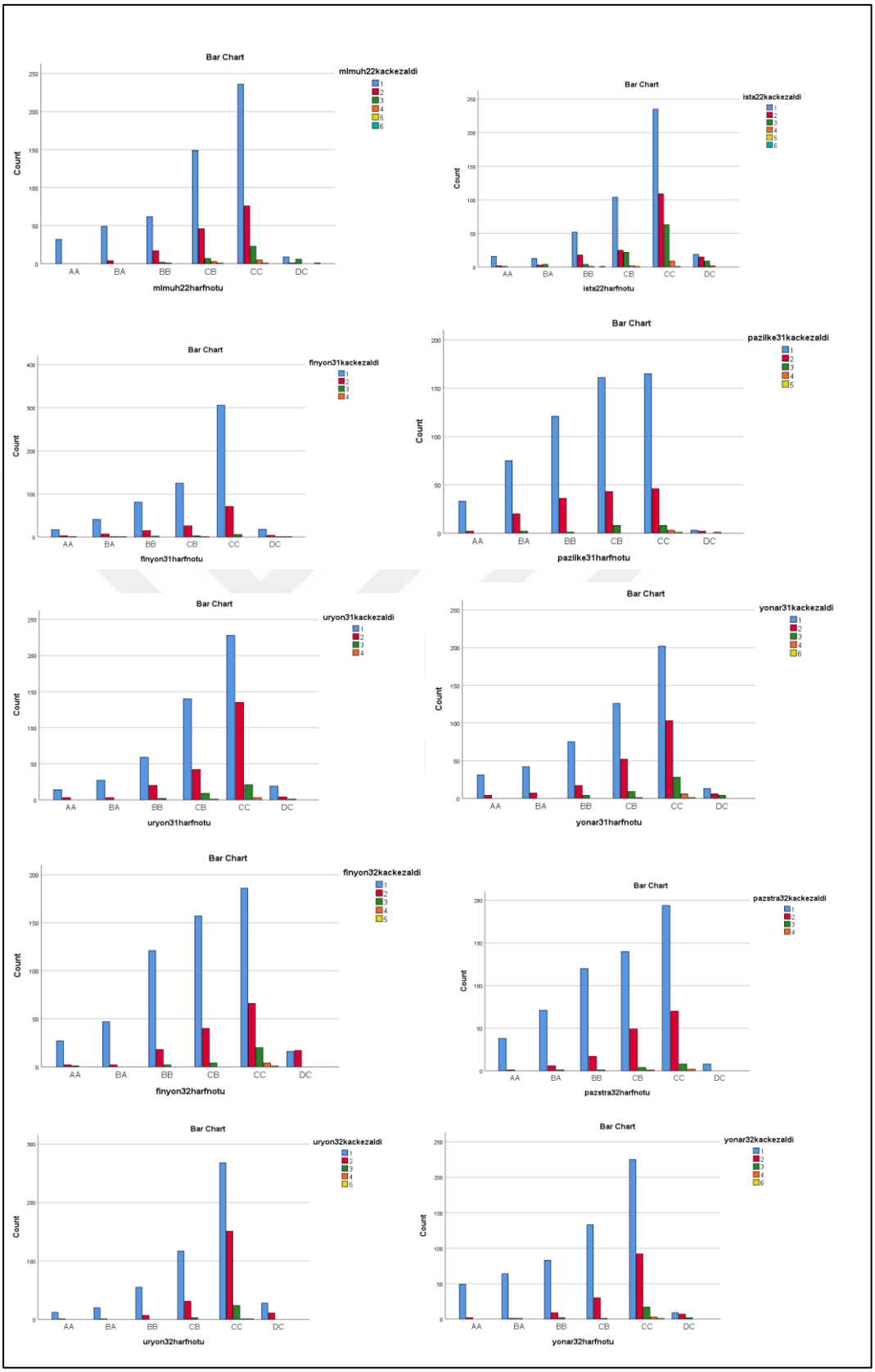
Grafiklerde görülmektedir ki öğrenciler en fazla muhasebe, iktisat, istatistik gibi sayısal ağırlıklı derslerde zorlanmaktadır. Bu dersleri birden çok kez alma eğilimindedirler. Genel olarak notların CC seyrinde olması dikkat çekmektedir. CC sayısal ifadede 50-59 arası notu ifade eder. 4lük sistemde ise bu 2.00'ı ifade eder. 2.00 dersleri geçmek için en son sınırdır.



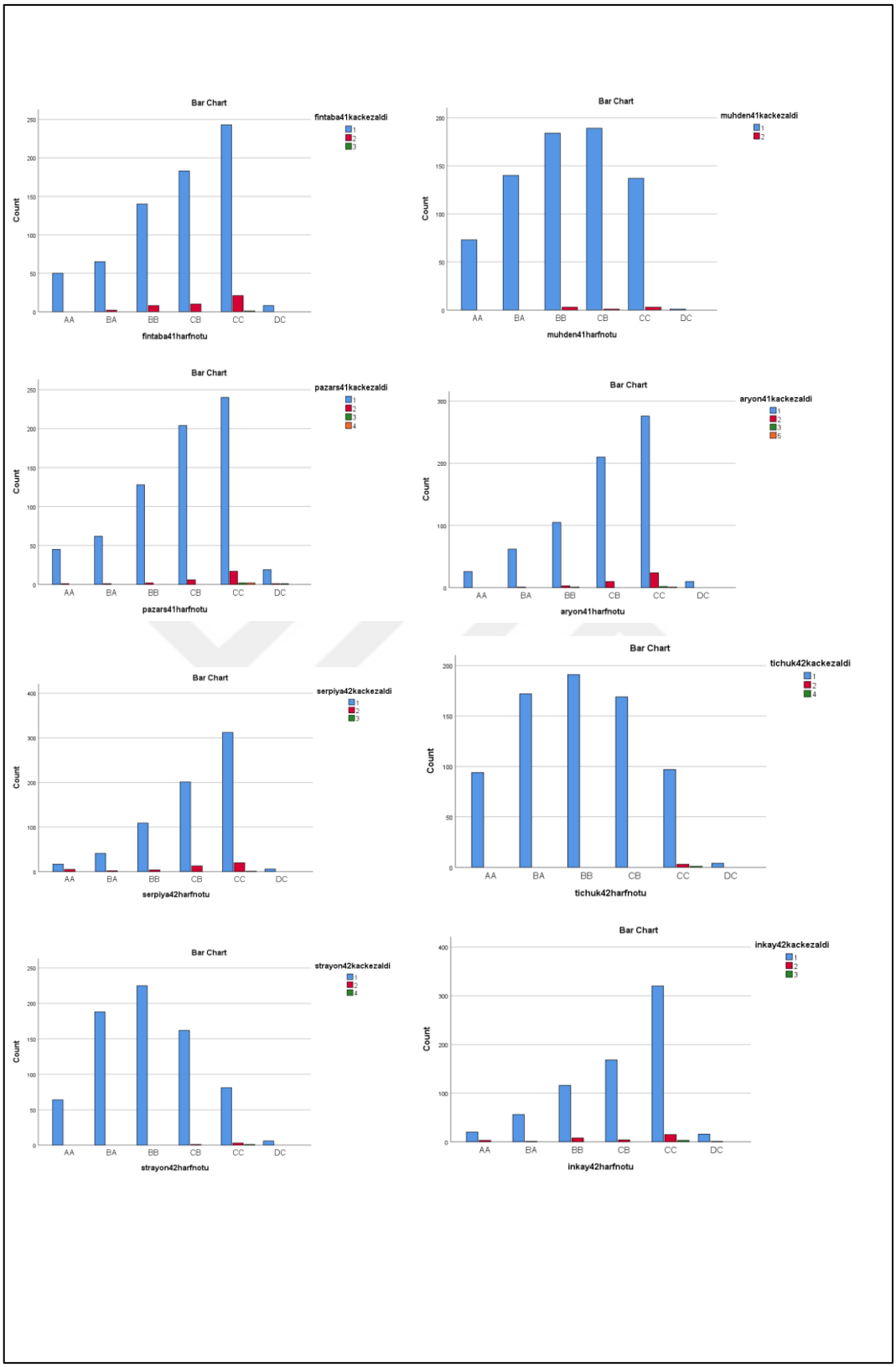
Şekil 12. Öğrenci not bilgisi ve dersi kaç kere aldığı bilgisinin çapraz tablosu



Şekil 13. Öğrenci not bilgisi ve dersi kaç kere aldığı bilgisinin çapraz tablosu (Devam)



Şekil 14. Öğrenci not bilgisi ve dersi kaç kere aldığı bilgisinin çapraz tablosu (Devam)



Şekil 15. Öğrenci not bilgisi ve dersi kaç kere aldığı bilgisinin çapraz tablosu (Devam)

2.3. Veri Temizleme

Bilgi işlem merkezinden talep edilen öğrenci bilgileri excel formatında verilmiştir. Bakıldığında bu veri setinde analizi yapabileceğimiz satır ve sütun bilgilerinin yerlerinin karışık olduğu gözlemlenmiştir. Veriler üzerinde gerekli düzenlemeler yapılıp analize uygun hale getirilmiştir. Bütün niteliklerin aynı düzende olması gerektiği için bu değerler sütunlar haline getirilmiştir.

Veri setinde okula direkt kayıt olan öğrenciler olduğu gibi, yatay geçişle gelen dikey geçişle gelen öğrenciler de bulunmaktadır. Bu öğrencilerin öğrenim süresi diğer öğrencilere göre farklılık göstermektedir. Ayrıca okulun sağladığı Erasmus programına dâhil olup birkaç eğitim dönemini yurt dışında tamamlayan öğrenciler de bulunmaktadır.

Elde edilen veri setinde dersi kaç kere aldığı ve önceden hangi harf notlarını aldığı bilgisi de bulunmaktadır. Önceki alımlarda dersin harf notu bilgisi mezuniyete bir etki etmediği için çalışmaya dâhil edilmemiştir. Öğrenci doğum tarihleri de çalışmaya dâhil edilmemiştir.

Öğrencilerin aldığı 38 tane zorunlu dersin hepsi analize dâhil edilmiştir. Fakat alınması gereken seçmeli 12 dersin toplamda 70 ve üstü farklı dersten oluşması, bazı derslerin sadece 1 kişi tarafından alınması gibi durumlar veri madenciliğinin gerektirdiği bütüncül yaklaşıma ters düşmektedir. Bu durumun oluşmasına bazı öğrencilerin Erasmus programında aldığı dersleri burada saydırması, iki yıllık programlardan gelen öğrencilerin oradaki derslerini burada saydırması neden olmuştur. Bu gibi nedenlerden ötürü seçmeli dersler analiz dışı tutulmuştur.

2.4. Veri Dönüştürme

Çalışmanın ilk amacı dersler ve notlar arasındaki ilişkileri bulmaktır. Bu amaçla Birliktelik Analizi yapılmıştır. Birliktelik analizinde derslerin harf notu değerleri girilerek analiz gerçekleştirilmiştir. Mezuniyet notu, ÖSS puanı, ÖSS sıralaması ve de derslerini ilk geçişte geçen öğrenci bilgilerine Sınıflandırma algoritmaları uygulanarak tahminleme yapılmıştır.

Veri setinde bütünlüğün sağlanabilmesi açısından bazı ders bilgilerinde değişiklikler yapılmıştır. Geçiş yapan öğrencilerin önceki ders bilgileri muaf

olarak görünmektedir. Bu dersler öğrencilerin mezuniyetine ortalama bir değer olarak yansımaktadır. Bütünlüğün sağlanması açısından M olarak görünen dersler ortalama bir harf notu olan CB olarak değiştirilmiştir. Yine bu şekilde bazı öğrencilerin not bilgisi G geçti olarak görünmekteydi bu değerler de CB olarak değiştirilmiştir.

Çalışmada mezuniyet notları ÖSS puanları ve ÖSS sıralamaları, ders notları kullanılarak tahmin edilmeye çalışılmıştır. Derslerini ilk defada geçen öğrencilerin ders bilgileri kullanılarak ise mezuniyet puanları tahmin edilmiştir.

Tablo 15

Sınıflandırma Algoritmalarında Kullanılan Sınıf Değişkenleri ve Alabilecekleri Değerler

Değişkenler	Açıklama	Alan
Mezuniyet Notları	3.00'ten düşük olanlar	düşük
	3 ve üstü olanlar	yüksek
ÖSS Puanları	200-299 arası olanlar	düşük
	300 ve üstü olanlar	yüksek
ÖSS Sıralamaları	ilk 200binde olanlar	düşük
	200binden yukarda olanlar	yüksek

Tablo 15'te sınıflandırma yani tahminleme yapılacak değişkenlerin alabilecekleri değerler görülmektedir. Bu değerler ders notları kullanılarak program vasıtasıyla tahmin edilmeye çalışılmıştır. 2.00-4.00 arasında değişen mezuniyet ortalamaları ikili sınıflara ayrılmıştır. Ortalaması 3'ten düşük olanlar düşük,3 ve üstü olanlar ise yüksek olarak ayrılmıştır. ÖSS puanları 200-299 arası olanlar düşük, 300 ve üstü olanlar yüksek puanlı olarak ayrım yapılmıştır. Türkiye geneli yapılan Öğrenci Seçme Sınavından ilk 200bin'e girenler yüksek dereceli, 200bin'den yukarda sıralamaları olan öğrenciler ise düşük dereceli olarak belirlenmiştir.

2.5. Modelin Oluşturulması

Bu araştırmanın amacı İşletme bölümü lisans derslerinin veri madenciliği yöntemleriyle analiz edilmesidir. RTEÜ, İİBF, İşletme bölümü öğrencilerinin lisans ders bilgileri ve diğer öğrenci bilgileri kullanılarak model oluşturulmuştur.

Modelin gerçekleştirilmesi aşamasında iki yöntemsel yaklaşımda bulunulmuştur. Birincisi Birliktelik Kuralları çıkarımı, ikincisi ise Sınıflandırma algoritmalarıdır. Birliktelik kurallarıyla dersler arasındaki ilişkilerin notlar bakımından tanımlamaları yapılırken, sınıflandırma algoritmalarıyla ders notlarına göre mezuniyet notları, ÖSS puanı, ÖSS sıralamaları tahmin edilmiştir. Ayrıca veritabanından elde edilen öğrencilerin her dersi kaç defada geçtiği bilgisine dayanarak, derslerini ilk defada geçen öğrencilerin dersleri ve mezuniyet notları arasındaki tahminleme işlemi Sınıflandırma algoritmalarıyla incelenmiştir.

Dersler arasındaki ilişkilerin belirlenmesi için her dersten alınan harf notları kullanılmıştır. Bu analiz Birliktelik Kuralları algoritması olan Apriori ile gerçekleştirilmiştir. Mezuniyet puanları, ÖSS puanları ve ÖSS yerleşme sıralamaları sınıf değişkeni yani hedef nitelik olarak seçilip ders notlarına bakılarak bu durumlar tahmin edilmiştir. Bunun için ise Sınıflandırma algoritmaları kullanılmıştır. Bu algoritmalar; Naive Bayes, J48 ve SMO algoritmalarıdır.

2.6. Modelin Değerlendirilmesi

Birliktelik Kurallarında, öğeler arasındaki ilişki, destek ve güven kriterleriyle ölçülür. Gerek görülürse kaldıraç değerine de bakılır.

Destek Değeri (Support): Veride öğeler arasındaki ilişkinin ne kadar sık olduğunu belirler.

$$Destek = \frac{n(A \cup B)}{N} \tag{2.1}$$

Formül (2.1), A ve B'nin birlikte olduğu işlem sayısının, toplam işlemlerin sayısına (N) bölünerek elde edildiğini gösterir (Demirok, 2018).

Güven Değeri (Confidence): A ve B'nin birlikte olduğu işlem sayısının, A'nın bulunduğu tüm işlem sayılarına bölünerek elde edildiğini gösterir. Güven değerinin 0 çıkması demek, A ürününün bulunduğu işlemlerin hiçbirinde B ürününün bulunmadığı anlamına gelir (Demirok, 2018).

$$Güven = \frac{n(A \cup B)}{N(A)} \tag{2.2}$$

Kaldıraç Değeri (Lift): Bir kuralın ne derece ilginç olduğunu belli eder. Destek ve güven değerinin yüksek olması her zaman önemi yüksek ve ilginç

kuralların elde edileceği anlamı taşımamaktadır. Böyle durumlarda lift değerine bakılır. Lift değerinin, 1 değerini alması ilginçliğin olmadığını 1'den büyük veya küçük değerler alması ise ilginçliğin arttığını göstermektedir (Ateş & Karabatak, 2017).

$$Kaldıraç = \frac{destek(A \cup B)}{deste(A).destek(B)} \quad (2.3)$$

Sınıflandırma algoritmalarında kullanılan değerlendirme kriterleri aşağıda açıklanmıştır.

Hata Matrisi (Confusion Matrix): Sınıflandırma modellerinin başarısı hata matrisindeki (confusion matrix) doğru sınıflandırılan değişkenler ve yanlış sınıflandırılan değişkenlerin sayısıyla hesaplanır. Satırlar test kümesindeki örneklerin gerçek sayılarını, sütunlar ise modelin tahminlemesini ifade eder. Tablo 16'da iki sınıflı bir veri kümesinde oluşturulmuş modelin hata matrisi verilmiştir. Bu matriste ana köşegen doğru tahminlenmiş örnek sayılarını, kalan matris elemanları ise hata sonuçlarını verir. TP (True Pozitif) ve TN (True Negatif) doğru sınıflandırılmış örnek sayısını verir, FP (False Pozitif) negatif sınıfındayken pozitif olarak tahminlenmiş örneklerin sayısıdır. FN (False Negatif) ise pozitif sınıfındayken negatif olarak tahminlenmiş örneklerin sayısını belirtir.

Tablo 16

Hata Matrisi (Confusion Matrix)

		Tahmin Edilen	
		pozitif	negatif
Gerçek Sınıf	pozitif	TP	FP
	negatif	FN	TN

Doğruluk Oranı: Doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır (Şık, 2014).

$$Doğruluk = \frac{TN+TP}{FN+TN+TP+FP} \quad (2.4)$$

Kesinlik (Precision): Doğru sınıflandırılmış pozitif örneklerin toplam pozitif tahmin edilen örneklere oranıdır (Balaban & Kartal, 2018).

$$Kesinlik = \frac{TP}{TP+FN}$$

(2.5)

Duyarlılık (Recall): Modelin hedef değişkeninin, pozitif sınıf etiketini tahmin etmedeki etkililiğidir. Duyarlılık (precision), Doğru Pozitiflik Oranı (True Positive Rate) olarak da adlandırılır (Balaban & Kartal, 2018).

$$Duyarlılık = \frac{TP}{TP+FP}$$

(2.6)

Yanlış Pozitiflik Oranı(False Positive Rate): Yanlış sınıflandırılmış pozitif örneklem sayısının gerçek sınıfı pozitif olan tüm örneklemlerin sayısına oranını ifade eder (Şık, 2014).

$$Yanlış\ Pozitiflik\ Oranı = \frac{FP}{TP+FP}$$

(2.7)

F Ölçütü (F Measure): Kesinlik ve Duyarlılık değerlendirme ölçütlerinin harmonik ortalamasıdır.

$$F\ Ölçütü = \frac{2*Duyarlılık*Kesinlik}{Duyarlılık+Kesinlik}$$

(2.8)

Kappa İstatistiği (Kappa Statistic): Yapılan tahminin doğruluk ölçüsüdür. Kappa değeri 0-1 arasında değişir. 0,4-0,6 arasındaysa orta, 0,6-0,8 arasındaysa iyi, 0,8-1 arasında ise çok iyi seviyede uyum vardır.

Modelin performansı kullanılan öğrenme algoritmasının yanı sıra sınıf dağılımı, hatalı sınıflandırma ya da eğitim ve test kümelerinin büyüklüğüne bağlıdır. Bu gibi durumları ortadan kaldırmak için çeşitli yöntemler vardır. Bu tez çalışmasında kullanılan yöntemler aşağıda açıklanmıştır.

Holdout: Holdout yönteminde veri, eğitim ve test veri seti olarak ikiye ayrılır. Eğitim veri seti, eğitim aşamasında model oluşturulması için seçilen sınıflandırıcının eğitiminde kullanılmaktadır. Model parametreleri için en iyi değerler ve uygun performans ölçüsü bu adımda belirlenir. Test veri seti ise oluşturulan modelin genel performansını bulabilmek için kullanılmaktadır. Bu yöntemin iki temel dezavantajı vardır. Bunlardan ilki elde az verinin olması durumunda test veri seti için yeteri kadar örneğin ayrılamayacak olmasıdır. Diğeri ise, eğitim ve test veri seti bir kereye mahsus olmak üzere birbirinden

ayrıldığından, elde edilen hata oranında bu ayrımdan kaynaklanan sorunlar yaşanabilir.

Tabakalı Örnekleme (Stratified Sampling): Çıktı değeri birer sınıf olan veri setlerinde, bazı sınıfların örnekleri çok az olabilir ya da hiç olmayabilir. Bu gibi durumlarda sınıf oranlarının korunması istenebilir. Bu nedenle tabakalı örnekleme tercih edilmektedir, fakat çıktı değeri bir sınıf yerine nümerik bir değer olduğunda bu yöntemden faydalanılamaz.

Çapraz Geçerleme (Cross Validation): Çapraz geçerleme, k-kat çapraz geçerleme ve birini dışarıda bırak çapraz geçerleme olmak üzere iki farklı biçimde uygulanabilmektedir. k-kat çapraz geçerlemede, veri seti k eşit parçaya ayrılır. k parçadan her seferinde bir tanesi test, k-1 tanesi de eğitim için kullanılmaktadır. Sonuçta k tane hata oranı elde edilmekte ve tüm tahmin hatasını hesaplayabilmek için hataların ortalaması alınmaktadır. Kullanılan k kat sayısının büyük olması halinde, tahminci daha doğru olacağından gerçek hata tahmincisinin sapması (bias) küçük; varyans ve hesaplama zamanı ise büyüktür. Bir başka deyişle, k kat sayısı küçük seçilirse, tahmincinin sapması küçük ve sapması gerçek hata tahmincisinden büyük olacaktır. Sapmadan kaynaklanan hata gerçek değer ile tahmin edilen değer arasındaki farktır, varyanstan kaynaklanan hata ise verilen bir nokta için model tahminindeki değişikliktir

2.7. Uygulama Sonuçları

Verilerin analizi için iki program kullanılmıştır. Birliktelik kuralları için R'dan, Sınıflandırma yapmak için ise WEKA'dan yararlanılmıştır.

2.7.1. Birliktelik Kurallarıyla Elde Edilen Sonuçlar

Apriori algoritmasıyla yapılan analizler 5 farklı şekilde incelenecektir. İlk adımda destek (support) değeri 0,1, güven (confidence) değeri 0,9 seçilerek tüm zorunlu dersler analiz edilmiştir. Programın sunduğu kurallardan anlam yaratan bilgiler Tablo 17'de gösterilmiştir. Diğer adımlarda ise kural çıkarımının sol kısmını oluşturan lhs değerleri değiştirilerek kurallar edilmiştir. Tablo 18, lhs değeri AA seçildiğinde çıkan kuralları, Tablo 19, lhs değeri BA seçildiğinde

çıkan kuralları, Tablo 20, lhs değeri BB seçildiğinde çıkan kuralları, Tablo 21, lhs değeri CB seçildiğinde çıkan kuralları göstermektedir.

Sonuçların kolay bir şekilde okunup yorumlanabilmesi amacıyla her dersin kısaltımı yapılıp, kaçıncı sınıfta, hangi dönemde alındığı yanlarında belirtilmiştir. Ders isimlerinin orijinal hallerine Tablo 7,8,9,10,11,12,13,14'ten bakılabilir.

Tablo 17

Support Değeri %10, Confidence Değeri %90 Seçildiğinde Çıkan Kurallar

lhs	=>	rhs	support	confidence	lift
{huktm12=CC,pazilke31=CC,yonar31=CC}	=>	{uryon32=CC}	11%	91%	1,49
{gmat11=CC,isyon21=CC,yonar31=CC}	=>	{uryon32=CC}	11%	90%	1,48
{gmuh12=CC,yonar31=CC,pazstra32=CC,yonar32=CC}	=>	{uryon32=CC}	10%	90%	1,49
{huktm12=CC,gmuh12=CC,isyon21=CC,ista21=CC}	=>	{mikro21=CC}	11%	92%	1,47
{yonorg22=CC,uryon31=CC,yonar31=CC,yonar32=CC}	=>	{uryon32=CC}	11%	91%	1,50
{tbt11=CC,huktm12=CC,gmuh12=CC,ista21=CC}	=>	{mikro21=CC}	11%	91%	1,45
{gmuh12=CC,davbil12=CC,ista21=CC,yonar31=CC}	=>	{uryon32=CC}	10%	90%	1,48
{gmat11=CC,davbil12=CC,ista21=CC,ista22=CC}	=>	{uryon32=CC}	10%	90%	1,48
{gmat11=CC,gmuh12=CC,davbil12=CC,ista22=CC}	=>	{uryon32=CC}	10%	93%	1,52
{gmuh12=CC,mlmuh22=CC,yonar31=CC,yonar32=CC}	=>	{uryon32=CC}	12%	91%	1,49
{gmat11=CC,gmuh12=CC,yonar31=CC,yonar32=CC}	=>	{uryon32=CC}	13%	90%	1,48
{gmat11=CC,gmuh12=CC,mlmuh22=CC,yonar31=CC}	=>	{uryon32=CC}	11%	91%	1,49

Tablo 17'de support değeri %10, confidence değeri %90 seçildiğinde çıkan kurallar görülmektedir. Burada bu seçilen değerler eşik değerlerdir. Program en az bu değerleri baz alarak ve daha yüksek değerlerde sonuçlar sunabilmektedir. İlk sıradaki kuralı yorumlarsak;

"1.sınıf 2.dönemde alınan Hukukun Temel Kavramlarından CC alanlar, 3.sınıf 1.dönemde alınan Pazarlama İlkelerinden CC alanlar, 3.sınıf 1.dönemde alınan Yöneylem Araştırmasından CC alanlar ve 3.sınıf 2.dönemde Üretim Yönetiminden CC alanlar tüm öğrencilerin % 11 ini oluşturmaktadır. "

"1.sınıf 2.dönemde alınan Hukukun Temel Kavramlarından CC alanlar, 3.sınıf 1.dönemde alınan Pazarlama İlkelerinden CC alanlar, 3.sınıf 1.dönemde alınan Yöneylem Araştırmasından CC alanlar, %91 ihtimalle 3.sınıf 2.dönemde Üretim Yönetiminden CC alacaklardır."

"1.sınıf 2.dönemde alınan Hukukun Temel Kavramlarından CC alındığında, 3.sınıf 1.dönemde alınan Pazarlama İlkelerinden CC alındığında,

3.sınıf 1.dönemde alınan Yöneylem Araştırmasından CC alındığında, 3.sınıf 2.dönemde Üretim Yönetiminden CC alma olasılığı sadece Üretim Yönetiminin CC olma olasılığından %49 daha fazladır.”

Tablo 17’deki tüm kurallara baktığımızda not değerlerinin hepsinin CC olduğu görülür. CC, 100’lük sistemde 50 ile 59 arası puana tekabül eder. 4’lük sistemde ise 2.00’ye denk gelir.

Tablo 18

Lhs Değeri AA Olarak Belirlendiğinde Elde Edilen Kurallar

Lhs	=>	rhs	support	confidence	lift
{tbt11=AA}	=>	{vergi21=CC}	12%	63%	1,41
{tbt11=AA}	=>	{huktm12=CC}	13%	67%	1,22

Tablo 18’de kuralların şart kısmını oluşturan lhs değeri AA seçildiğinde çıkan iki sonuçtan ilkini yorumlarsak;

“1.sınıf 1.dönemde alınan Temel Bilgi Teknolojileri dersinden AA alanlar, 2.sınıf 1.dönemde alınan Türk Vergi Sisteminden CC alanlar toplam öğrencilerin %12 sini tekabül etmektedir. Temel Bilgi Teknolojileri dersinden AA alanlar, %63 ihtimalle Türk Vergi Sisteminden CC alacaktır. Temel Bilgi Teknolojileri AA alındığında Türk Vergi Sisteminin CC olma olasılığı sadece Türk Vergi Sisteminin CC olma olasılığından %41 daha fazladır.

Tablo 19

Lhs Değeri BA Olarak Belirlendiğinde Elde Edilen Kurallar

Lhs	=>	rhs	support	confidence	lift
{tdil12=BA}	=>	{uryon31=CC}	11%	60%	1,13
{tdil12=BA}	=>	{uryon32=CC}	11%	61%	1,00

Lhs değeri BA olarak belirlendiğinde çıkan Tablo 19’daki sonuçlardan ilkini yorumlarsak;

“ 1.sınıf 2. dönemde alınan Türk Dili dersinden BA alanlar, 3.sınıf 1.dönemde alınan Üretim Yönetiminden CC alanlar toplam öğrencilerin %11 ine denk gelmektedir. Türk Dilinden BA alanlar, %60 ihtimalle Üretim Yönetiminden

CC alacaktır. Türk Dili BA olduğunda Üretim Yönetiminin CC olma olasılığı sadece Üretim Yönetiminin CC olma olasılığından %13 daha fazladır.

Tablo 20

Lhs Değeri BB Olarak Belirlendiğinde Elde Edilen Kurallar

Lhs	=>	rhs	support	confidence	lift
{davbil12=CC,tdil12=BB}	=>	{mikro21=CC}	10%	70%	1,11
{tdil12=BB,yonar31=CC}	=>	{uryon32=CC}	11%	73%	1,19
{tdil12=BB,mlmuh22=CC}	=>	{makro22=CC}	11%	72%	1,28
{tdil12=BB,mlmuh22=CC}	=>	{uryon32=CC}	10%	71%	1,17
{gmat11=CC,tdil12=BB}	=>	{uryon32=CC}	11%	72%	1,19
{tdil12=BB,ista21=CC}	=>	{uryon32=CC}	11%	70%	1,15
{tdil12=BB,ista21=CC}	=>	{mikro21=CC}	13%	79%	1,26
{huktm12=CC,tdil12=BB}	=>	{mikro21=CC}	13%	73%	1,17
{gmuh12=CC,tdil12=BB}	=>	{uryon32=CC}	14%	70%	1,14
{gmuh12=CC,tdil12=BB,mikro21=CC}	=>	{uryon32=CC}	10%	76%	1,24

Lhs değeri BB olarak belirlendiğinde çıkan 50'ye yakın anlamlı bilgiden confidence değeri %70 ve üzeri olanlar Tablo 20'de gösterilmiştir. İlk kuralı yorumlarsak;

"1.sınıf 2.dönem dersi Davranış Bilimlerinden CC alanlar, 1.sınıf 2.dönem dersi Türk Dilinden BB alanlar ve 2.sınıf 1.dönem dersi Mikro İktisattan CC alanlar tüm öğrencilerin %10'una denk gelmektedir. Davranış Bilimlerinden CC alanlar, Türk Dilinden BB alanlar %70 ihtimalle Mikro İktisattan CC alacaktır. Davranış Bilimlerinden CC alındığında, Türk Dilinden BB alındığında, Mikro İktisattan CC alınma olasılığı sadece Mikro İktisattan CC alınma olasılığından %11 daha fazladır."

Tablo 21

Lhs Değeri CB Olarak Belirlendiğinde Elde Edilen Kurallar

Lhs	=>	rhs	support	confidence	lift
{ista21=CC,vergi21=CB}	=>	{mikro21=CC}	13%	80%	1,28
{isbilgir11=CB,yonar32=CC}	=>	{uryon32=CC}	13%	84%	1,38

Lhs değeri CB olarak seçildiğinde çıkan 500'yakın anlamlı bilgiden Confidence değeri %80 ve üzeri olanlar Tablo 21'de gösterilmiştir.

2.7.2. Sınıflandırma Algoritmalarıyla Elde Edilen Sonuçlar

Çalışmada sınıflandırma metodları kullanılarak amaçlanan öğrenci bilgilerini tahmin etmektir. Bu işlem için çalışmada Naive Bayes, Sıralı Minimal Optimizasyon ve J48 algoritmaları seçilmiştir. SMO algoritması, destek vektör makinelerinin WEKA'ya uyarlanmış halidir. J48 ise ID3 ve ID3'ün bir sonraki uyarlaması olan C4.5'in WEKA'ya uyarlanmasıdır.

Çalışmada model performans değerleme ölçüsü olarak Doğru Sınıflandırma Oranı, Kappa İstatistiği, Duyarlılık, Kesinlik ve F Ölçütüne bakılmıştır.

Model performans değerlendirme yöntemleri olarak ise Tabakalı Çapraz Geçerleme ve Tabakalı Örnekleme ile Holdout yapılmıştır. Çapraz geçerlemede 2 kat, 4 kat, 5 kat ve 10 katlı çapraz geçerleme yapılmıştır. Bu geçerlemelerden 10 katlı geçerleme şu şekilde açıklanır. Veri kaynağı 10 bölüme ayrılır ve her bölüm bir kez test kümesi, kalan diğer 9 bölüm öğrenme kümesi olarak kullanılır. 2 kat, 4 kat ve 5 kat geçerlemeler de bu mantıkta çalışır. Holdoutta ise %50/%50, %75/%25, %80/%20, %90/%20 değerleriyle analiz yapılmıştır. Holdouttaki %50/%50 ifadesinde, veri setinin %50 sinin eğitim seti %50 sinin ise test seti olarak kullanılmasını ifade eder. Diğer oranlar da bu mantık ile çalışır. Performans değerlendirme yöntemlerinde 8 farklı ayrımın yapılması en iyi tahmin sonuçlarını elde edebilmek içindir.

Sınıflandırma çalışmalarında tahmin edilmek istenen değişken hedef(sınıf) değişken adını alır, tahminde yararlanılan değişkenler ise tanımlayıcı değişken olarak adlandırılır. Çalışmada hedef değişken olarak mezuniyet ortalamaları, ÖSS puanları ve ÖSS sıralamaları seçilmiştir. Tanımlayıcı değişken olarak ise öğrenci harf notları seçilmiştir. Her bir sınıflama işlemi için çıkan sonuçlar aşağıdaki tablolarda özetlenmiştir.

Hedef Değişken Mezuniyet Ortalamaları Olduğunda Elde Edilen Sonuçlar

Hedef Değişken: Mezuniyet Ortalamaları (Düşük;673kişi, Yüksek;58kişi)

Tanımlayıcı Değişkenler: Ders Harf Notları

Kullanılan Algoritmalar: Naive Bayes, SMO, J48

Performans Değerlendirme Yöntemleri:

Tabakalı Çapraz Geçerleme (2kat,4kat,5kat,10kat)

Tabakalı Örnekleme ile Holdout (%50/%50, %75/25 ,%80%20, %90/%10)

Tablo 22

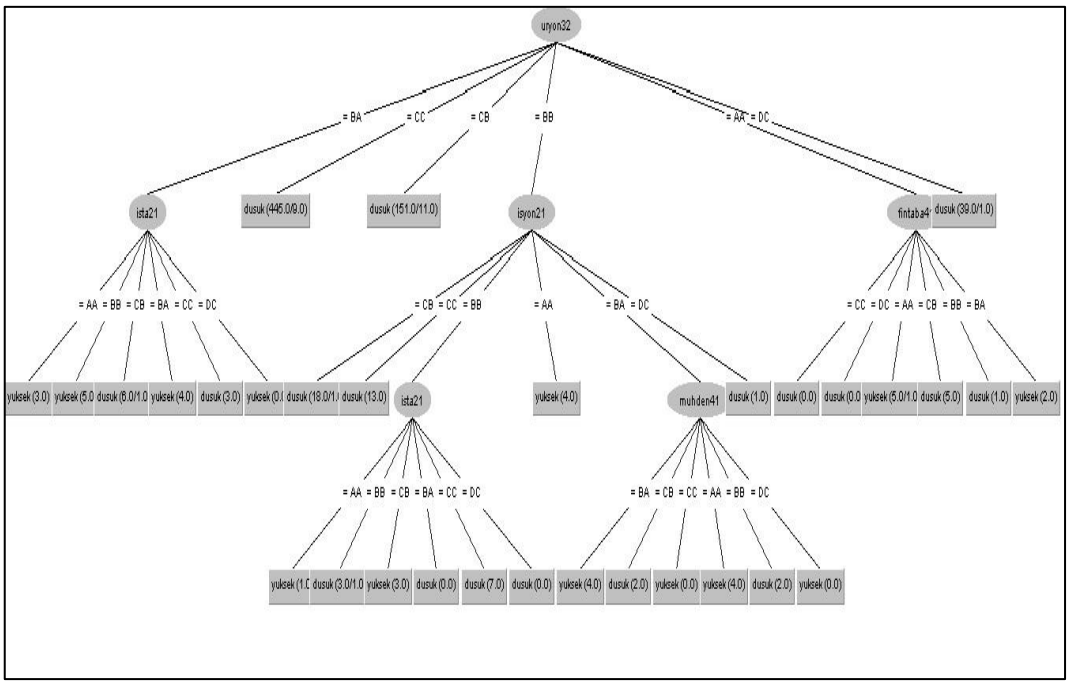
Hedef Değişken Mezuniyet Ortalamaları Olduğunda Elde Edilen Sonuçlar

Naive Bayes	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	94%	93,84%	95%	94%	95%	95%	95%	93%
Kappa İstatistiği	0,6882	0,6815	0,7137	0,6882	0,7189	0,6382	0,696	0,6685
Duyarlılık	0,962	0,963	0,966	0,962	0,963	0,96	0,973	0,947
Kesinlik	0,941	0,938	0,947	0,941	0,948	0,945	0,952	0,932
F Ölçütü	0,947	0,946	0,952	0,947	0,953	0,95	0,958	0,937
SMO	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	95%	96%	95%	95%	95%	95%	96%	95%
Kappa İstatistiği	0,6617	0,712	0,6658	0,6835	0,628	0,5541	0,6782	0,684
Duyarlılık	0,952	0,958	0,952	0,954	0,947	0,945	0,964	0,945
Kesinlik	0,955	0,958	0,953	0,955	0,951	0,945	0,959	0,945
F Ölçütü	0,953	0,958	0,952	0,954	0,947	0,945	0,961	0,945
J48	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	92%	92%	93%	92%	91%	93%	95%	93%
Kappa İstatistiği	0,3852	0,3406	0,3624	0,3221	0,3515	0,3441	0,4715	0,5114
Duyarlılık	0,911	0,911	0,915	0,906	0,903	0,921	0,94	0,924
Kesinlik	0,917	0,917	0,929	0,922	0,91	0,929	0,945	0,932
F Ölçütü	0,913	0,913	0,917	0,911	0,906	0,924	0,942	0,923

Tablo 22'de Naive Bayes algoritmasıyla doğru sınıflandırma oranı en yüksek sonuçlar 5 kat çapraz geçerleme, %50/%50 tabakalı holdout, %75/25 tabakalı holdout ve %80/%20 holdoutta %95 oranıyla bulunulmuştur. Kappa istatistiği en yüksek 0,7189 ile %50/%50 tabakalı holdout ile bulunulmuştur. Duyarlılık, Kesinlik ve F Ölçütü hepsinde 1'e yakın iyi sonuçlar vermiştir.

SMO algoritmasıyla en yüksek doğru sınıflandırmayı 4 kat çapraz geçerleme ve %80/%20 tabakalı holdout %96 oranıyla vermiştir. Kappa istatistiği en yüksek ise 0,712 ile 4 kat çapraz geçerlemede elde edilmiştir. Duyarlılık, Kesinlik ve F ölçütleri ise hepsinde 1'e yakın iyi sonuçlar vermiştir.

J48 algoritmasıyla en iyi doğru sınıflandırma %80/%20 tabakalı holdout ile %95 oranıyla elde edilmiştir. Kappa istatistiği en yüksek ise 0,4715 ile %80/20 ile tabakalı holdouttur. 0,4-0,6 arasındaki kappa ortalama bir değeri ifade eder. Duyarlılık, Kesinlik ve F ölçütleri ise hepsinde 1'e yakın iyi sonuçlar vermiştir.



Şekil 16. Mezuniyet ortalamaları için j48 ağaç yapısı

Şekil 16'da öğrencilerin mezuniyet notlarını tahmin etmede J48 algoritmasının çizdiği ağaç yapısı görülmektedir. Ağaçlandırma Üretim Yönetimi 2 dersi ile başlamıştır. Üretim Yönetimi 2'de BA dersi alındığında İstatistik 1 dersine gidilmiştir. İstatistik dersinde AA, BB, BA, DC alındığında mezuniyet notlarının yüksek sınıfta yer aldığı görülmektedir. Üretim Yönetimi 2'de BB alındığında İşletme Yönetimi dersine gidilmiştir. Bu dersten CB, CC, DC alındığında düşük sınıfa öğrenci ait olmuştur. AA alınca yüksek sınıfta yer alınmıştır. BB alındığında İstatistik 1 dersine gidilmiştir. BA alındığında ise Muhasebe Denetimine gidilmiştir. Üretim Yönetimi 2 dersinde AA alındığında

Finansal Tablolar Analizi dersine gidilmiştir. Üretim Yönetimi 2 dersinden DC alındığında ise mezuniyet ortalaması düşük sınıfta öğrenci yer almıştır.

Hedef Değişken ÖSS Puan Durumu Olduğunda Elde Edilen Sonuçlar

Hedef Değişken: ÖSS Puanları (Düşük(602kişi),Yüksek(129kişi))

Tanımlayıcı Değişkenler: Ders Harf Notları

Kullanılan Algoritmalar: Naive Bayes, SMO, J48

Performans Değerlendirme Yöntemleri:

Tabakalı Çapraz Geçerleme (2kat,4kat,5kat,10kat)

Tabakalı Örnekleme ile Holdout (%50/%50, %75/25 ,%80%20, %90/%10)

Tablo 23

Hedef Değişken ÖSS Puan Durumu Olduğunda Elde Edilen Sonuçlar

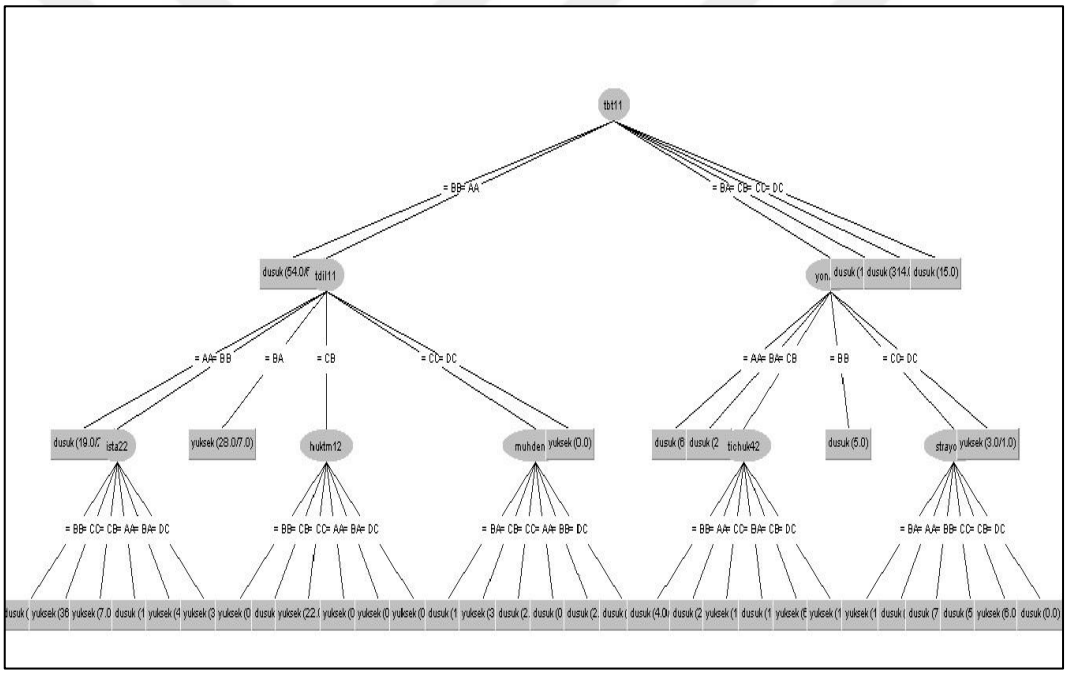
Naive Bayes	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	82%	83%	85%	84%	80%	80%	80%	79%
Kappa İstatistiği	0,4194	0,4341	0,5122	0,4854	0,3424	0,2312	0,2662	0,2474
Duyarlılık	0,832	0,836	0,859	0,851	0,813	0,801	0,82	0,88
Kesinlik	0,824	0,829	0,851	0,843	0,803	0,798	0,801	0,795
F Ölçütü	0,827	0,832	0,854	0,846	0,807	0,799	0,81	0,827
SMO	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	82%	84%	85%	84%	79%	81%	79%	81%
Kappa İstatistiği	0,4261	0,4454	0,4854	0,4606	0,3025	0,3036	0,2429	0,2003
Duyarlılık	0,834	0,839	0,851	0,843	0,801	0,82	0,815	0,866
Kesinlik	0,824	0,836	0,854	0,841	0,795	0,814	0,788	0,808
F Ölçütü	0,828	0,837	0,852	0,842	0,798	0,817	0,8	0,832
J48	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	86%	86%	87%	87%	85%	86%	86%	89%
Kappa İstatistiği	0,4465	0,5148	0,5643	0,5431	0,5407	0,5407	0,4686	0,4966
Duyarlılık	0,845	0,86	0,873	0,867	0,872	0,865	0,871	0,919
Kesinlik	0,858	0,865	0,874	0,87	0,855	0,863	0,856	0,89
F Ölçütü	0,848	0,862	0,874	0,869	0,861	0,864	0,862	0,901

Tablo 23'te Naive Bayes ile yapılan analizde ÖSS puanlarını tahmin etmede en iyi doğru sınıflandırma 5 kat çapraz geçerleme ile elde edilmiştir.

Kappa istatistiği en iyi sonucu 0,5122 değeriyle 5 kat çapraz geçerlemede vermiştir. Duyarlılık, Kesinlik ve F Ölçütü ise 0,7 ve 0,8 civarlarındadır.

SMO algoritmasında en yüksek oranda doğru tahminleme 5 kat çapraz geçerleme ile elde edilmiştir. Kappa değeri 0,4854 ile 5 kat çapraz geçerlemede en yüksek değeri almıştır. Duyarlılık, Kesinlik ve F Ölçütü ise 0,7 ve 0,8 civarlarındadır.

J48 algoritması sonuçlarında en yüksek doğru tahminleme %90/%10 holdout ile %89 oranında ortaya çıkmıştır. Diğerlerine göre en iyi Kappa değerini ise 0,5643 ile 5 kat çapraz geçerleme vermiştir. Duyarlılık, Kesinlik ve F Ölçütü ise 0,8 ve 0,9 civarlarındadır.



Şekil 17. ÖSS puan durumları için j48 ağaç yapısı

ÖSS puan durumlarını tahmin etmek için kullandığımız Karar Ağacı algoritması J48 ile Şekil 17'deki ağaç yapısı elde edilmiştir. Ağaçlandırma Temel Bilgi Teknolojileri dersi ile başlamıştır. Temel Bilgi Teknoloji dersinde BB, CB, CC, DC alındığında ÖSS puanı düşük olan sınıfta yer almıştır öğrenciler. Bu dersten AA alındığında program Türk Dili 1 dersine gitmiştir. Türk Dili 1 dersinden DC ve BA alındığında öğrenci ÖSS puanı yüksek sınıfta yer almıştır. AA alındığında düşük sınıfta yer alınmıştır. BB alındığında program İstatistik 2 dersine gitmiştir. CC alındığında Muhasebe Denetimi dersine gidilmiştir. Temel

Bilgi Teknolojileri dersinden BA alındığında Yöneylem Araştırması dersine gidilmiştir. Bu dersten AA, BA, BB, alındığında ÖSS puanlarına göre düşük sınıfta yer almıştır öğrenciler. Yöneylemden DC alan öğrenciler ise yüksek sınıfta yer almıştır. AA, BA, BB gibi notlar yüksek, DC düşük bir notken ÖSS puanlamasında öğrencilerin düşük sınıfta olması öğrencilerin üniversitedeki dersleri başarıyla tamamlayabilirken ÖSS sınavlarında düşük başarı gösterdiğini göstermektedir. Yöneylemden CB alındığında Ticaret Hukukuna, CC alındığında Stratejik Yönetim dersine gidilmiştir. Ticaret Hukuku dersinden BB, AA, BA alanlar düşük sınıfta, CC, CB, DC alanlar ise yüksek sınıfta yer almıştır. Stratejik Yönetim dersinde BA ve CB alanlar yüksek sınıfta yer alırken, AA, BB, CC, DC alanlar düşük sınıfta yer almıştır. Stratejik Yönetim dersinde çıkan bu sonuçlar anlamlı bilgiler olarak değerlendirilemezler. Düşük ve yüksek notlar arasında anlamlı bir sınıflandırma yapılmamıştır. Buradan yola çıkılarak öğrenci lisans ders notlarıyla ÖSS puanlarını tahmin etmenin önemli bir bilgi kazandırmadığı sonucuna varılır.

Hedef Değişken ÖSS Sıralaması olduğunda Elde Edilen Sonuçlar

Hedef Değişken: ÖSS Sıralamaları (Düşük(28kişi),Yüksek(708kişi))

Tanımlayıcı Değişkenler: Ders Harf Notları

Kullanılan Algoritmalar: Naive Bayes, SMO, J48

Performans Değerlendirme Yöntemleri:

Tabakalı Çapraz Geçerleme (2kat,4kat,5kat,10kat)

Tabakalı Örnekleme ile Holdout (%50/%50, %75/25 ,%80%20, %90/%10)

Tablo 24

Hedef Değişken ÖSS Sıralaması olduğunda Elde Edilen Sonuçlar

Naive Bayes	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	90%	89%	90%	893297%	92%	91%	91%	86%
Kappa İstatistiği	0,0791	0,0585	0,1085	0,1044	0,127	0,2299	0,1896	0,101
Duyarlılık	0,933	0,932	0,936	0,936	0,941	0,937	0,927	0,908
Kesinlik	0,904	0,888	0,896	0,893	0,918	0,913	0,911	0,863
F Ölçütü	0,918	0,908	0,914	0,913	0,929	0,924	0,918	0,884
SMO	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	95%	92%	93%	92%	96%	96%	95%	90%
Kappa İstatistiği	0,0649	0,0269	0,0908	0,055	0,0932	0,1858	0,2053	-0,0493
Duyarlılık	0,932	0,928	0,933	0,93	0,94	0,941	0,936	0,891
Kesinlik	0,945	0,923	0,926	0,922	0,956	0,956	0,952	0,904
F Ölçütü	0,938	0,926	0,93	0,926	0,947	0,944	0,939	0,898
J48	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	96%	96%	96%	96%	96%	96%	95%	95%
Kappa İstatistiği	0	0	0	0	0	0	0	0
Duyarlılık	-	-	-	-	-	-	-	-
Kesinlik	0,962	0,962	0,962	0,962	0,964	0,956	0,952	0,945
F Ölçütü	-	-	-	-	-	-	-	-

Öğrencilerin ders notlarıyla ÖSS puanlarını tahmin etmek için yapılan Tablo 24'teki analizlerde Naive Bayes ve SMO algoritması anlamlı sonuçlar verirken J48 anlamlı sonuçlar vermemiştir ve ağaçlandırmayı örneklemin dağılımına göre sadece düşük ve yüksek olarak iki sınıfa ayırmıştır. Burada

örneklemde yer alan 731 kişinin 28'inin düşük, 708'inin yüksek sınıfta yer alması ve dengesiz dağılması J48'in ağaçlandırma yapmamasına neden olduğu söylenir.

Derslerini İlk defada Geçen Öğrencilerin Mezuniyet Ortalaması Tahmini

Hedef Değişken: İlk alımda dersini geçen 43 öğrenci (23yüksek, 20düşük)

Tanımlayıcı Değişkenler: Ders Harf Notları

Kullanılan Algoritmalar: Naive Bayes, SMO, J48

Performans Değerlendirme Yöntemleri:

Tabakalı Çapraz Geçerleme (2kat,4kat,5kat,10kat)

Tabakalı Örnekleme ile Holdout (%50/%50, %75/25 ,%80%20, %90/%10)

Tablo 25

Derslerini İlk defada Geçen Öğrencilerin Mezuniyet Ortalaması Tahmini

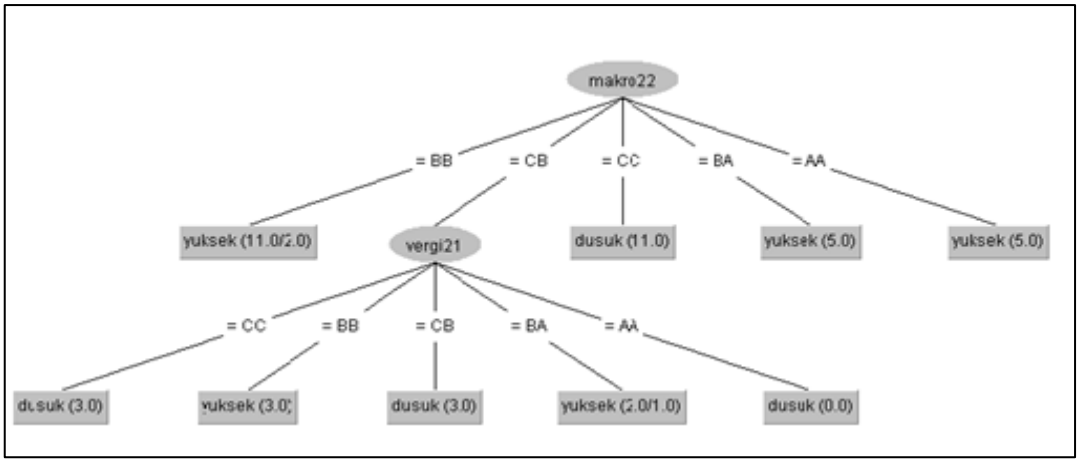
Naive Bayes	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	93%	95,35%	93%	93%	81%	100%	89%	100%
Kappa İstatistiği	0.8593	0.9059	0.8593	0.8593	0.6216	1	0.7692	1
Duyarlılık	0,931	0,957	0,931	0,931	0,826	1	0,907	1
Kesinlik	0,93	0,953	0,93	0,93	0,81	1	0,889	1
F Ölçütü	0,93	0,953	0,93	0,93	0,81	1	0,886	1
SMO	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	91%	88%	79%	86%	71%	64%	78%	100%
Kappa İstatistiği	0.813	0.764	0.5807	0.7177	0.4167	0.2414	0.55	1
Duyarlılık	0,907	0,89	0,792	0,863	0,714	0,644	0,778	1
Kesinlik	0,907	0,884	0,791	0,86	0,714	0,636	0,778	1
F Ölçütü	0,907	0,883	0,791	0,86	0,714	0,617	0,778	1
J48	2kat	4kat	5kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	70%	74%	72%	79%	71%	82%	78%	100%
Kappa İstatistiği	0.3698	0.4808	0.4391	0.5752	0.4167	0.6333	0.55	1
Duyarlılık	0,76	0,747	0,721	0,794	0,714	0,818	0,778	1
Kesinlik	0,698	0,744	0,721	0,791	0,714	0,818	0,778	1
F Ölçütü	0,67	0,742	0,721	0,789	0,714	0,818	0,778	1

Tablo 25'te aldığı 38 zorunlu dersi ilk alımında geçen 43 öğrenci ile yapılan analizlerde sonuçların 1'e çok yakın hatta 1 değerini almasında örneklemin birbirine yakın dağılması ekili olmuştur. Naive Bayes ile en doğru

sınıflandırma %100 değeriyle %75/%25 ve %90/%10 holdout ile elde edlmiştir. Kappa değeri 1 olarak en iyi sonucu da %75/%25 ve %90/%10 holdout vermiştir. Duyarlılık, Kesinlik ve F Ölçütü değeri ise 1 ve 1'e yakın iyi sonuçlar vermiştir.

SMO algoritmasında en iyi tahmini %100 ile %90%10 holdout, Kappa istatistiğini de en iyi 1 değeriyle %90/%10 holdout vermiştir. Duyarlılık, Kesinlik ve F Ölçütü değeri ise genel olarak 1'e yakın iyi sonuçlar vermiştir.

J48 algoritmasında da en iyi sınıflandırmayı %100 değeriyle %90/%10 holdout vermiştir. %90/%10 holdout vermiştir. Kappa istatistiğini de en iyi 1 değeriyle %90/%10 holdout vermiştir. Duyarlılık, Kesinlik ve F Ölçütü değeri ise genel olarak 1'e yakın iyi sonuçlar vermiştir.



Şekil 18. İlk defada derslerini geçenlerin mezuniyet ortalamaları tahmininde J48 ağaç yapısı

Şekil 18'de, J48 ile ilk defada derslerini geçenlerin mezuniyet ortalamalarının düşük mü yüksek mi olduğunu anlamada algoritma Makro İktisat 2 dersiyle ağaçlandırmaya başlamıştır. BB, BA, AA alan öğrencilerin mezuniyet ortalaması yüksek olan sınıfta yer almıştır. CC alanlar düşük sınıfta yer almıştır. CB alanlar için ise Türk Vergi Sistemi dersine gidilmiştir. Bu dersten CC, CB, AA alanlar düşük sınıfta yer alırken, BB ve BA alanlar yüksek sınıfta yer almıştır.

SONUÇ VE ÖNERİLER

Veri madenciliği yöntemleriyle, İşletme lisans derslerinin analizinin yapıldığı bu çalışmada Recep Tayyip Erdoğan Üniversitesi, İktisadi İdari Bilimler Fakültesi, İşletme bölümüne 2009 ilk kayıt yılından itibaren kaydolmuş ve 2019 bahar dönemine kadar mezun olmuş 731 öğrenci verisi kullanılmıştır. Üniversite Bilgi İşlem Merkezinden talep edilen veri setinde; öğrenci dersleri ve dersi geçme notları, mezuniyet ortalamaları, her dersi kaç defa aldığı bilgisi, Öğrenci Seçme Sınavı puanı, Türkiye geneli Öğrenci Seçme Sınavı yerleştirme sıralaması, cinsiyeti, doğum yeri, kayıt yılı, mezuniyet yılı, gece veya gündüz öğrenci olduğu bilgisi gibi veriler yer almıştır. Örneklemin çoğunluğunu kadın öğrenciler, gündüz öğrencileri ve Karadeniz bölgesi doğumlu öğrenciler oluşturmaktadır.

Öğrencilerin aldığı zorunlu derslerle yapılan analizde, dersler arasındaki harf notlarına dayalı ilişkiler Birliktelik Kurallarıyla çıkarılmıştır. Apriori algoritması ile yapılan sonuçlarda öğrenci notlarının CC ortalamasında olduğu görülmüştür. CC harf notu yüzlük sistemde 50 ile 59 puan aralığına denk gelmektedir.

Öğrencilerin ders notları kullanılarak mezuniyet notlarının tahmini, ÖSS puanının tahmini ve ÖSS sıralamasının tahmini Sınıflandırma algoritmalarıyla gerçekleştirilmiştir. Bu algoritmalar Naive Bayes, Sıralı Minimal Optimizasyon ve J48 algoritmasıdır. Ayrıca her öğrencinin aldığı dersleri kaç defada geçtiği bilgisi kullanılarak, derslerini ilk defada geçen öğrencilere bu Sınıflandırma algoritmaları mezuniyet ortalaması sınıf değişken seçilip uygulanmıştır. 731 öğrencinin ders notlarıyla mezuniyet puanlarının, ÖSS puanlarının ve ÖSS sıralamalarının düşük mü yüksek mi olduğu bilgisine erişilmek istenmiştir ve her üç algoritma için farklı sonuçlar elde edilmiştir. Örneklem bu 3 hedef değişken için farklı dağılımlar göstermiştir. Bu dağılım farklılığı elde edilen sonuçlarda da etkisini göstermiştir. ÖSS sıralamalarında 703 kişinin düşük, 28 kişinin yüksek sınıfta yer alıyor olması J48 algoritmasında programın ağaçlandırma yapmamasına neden olmuştur.

Öğrencilerin aldığı 38 tane zorunlu dersin genel olarak sayısal ağırlıkta dersler olduğu görülmektedir. ÖSS puan ve sıralamalarına bakıldığında çoğunlukta zayıf olan öğrenciler lisansta aldıkları sayısal derslerde de zorlanmaktadırlar. Öğrenciler aldıkları derslerden düşük puanlarla bile geçtiklerinde tekrar dersi alıp yükselme eğiliminde değildirler. Mezuniyet ortalamalarının yüksek olması için öğrencilerle ayrıca çalışma yapılması gelecek için önerilmektedir.

Öğrencilerin mezuniyet ortalamalarında belirleyici olan dersler; Üretim Yönetimi, İstatistik, İşletme Yönetimi, Finansal Tablolar Analizi, Muhasebe Denetimi gibi sayısal derslerdir. Öğrenciler arasında Genel Muhasebe 1-2 derslerini 7 kez alanlar olmuştur. Davranış Bilimleri, Genel Matematik 2, Mikro İktisat, Maliyet Muhasebesi, İstatistik 2, Yöneylem Araştırması 1-2 derslerini öğrenciler arasında 6 defa alanlar olmuştur.

Çalışmada öğrencilerin zorunlu dersleriyle analiz yapılmıştır. Alınması gereken 50 dersin 38'i zorunlu 12'si seçmelidir. Toplamda 70 ve üstü farklı seçmeli ders alınmıştır. Alınan seçmeli dersler öğrenciler arasında eşit değildir. Bu sebeple çalışmada seçmeli dersler analize dahil edilmemiştir. Fakat çeşitli dönüşüm işlemleriyle bu dersler arasında bütünlük sağlanıp gelecek çalışmalarda kullanılması önerilmektedir.

KAYNAKÇA

Abidin, D., Öztürk, Ö., & Öztürk, T. Ö. (2017). Klasik Türk Müziğinde Makam Tanıma İçin Veri Madenciliği Kullanımı. *Mühendislik Mimarlık Fakültesi Dergisi*, 1221-1232.

Agrawal , R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules.* 04 05, 2019 tarihinde http://www.vldb.org/conf/1994/P487.PDF adresinden alındı

Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining Association Rules between Sets of Items in Large Databases.* 04 02, 2019 tarihinde https://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf adresinden alındı

Akbal, E., Doğan, Ş., & Varol, N. (2017). Karar Ağaçları ile Telefon Dolandırıcılığı Verilerinin Analiz i. *Fırat Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, 171-177.

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İ.Ü.İşletme Fakültesi Dergisi*, 1-22.

Akpınar, H. (2014). DATA Veri Madenciliği Veri Analizi. İstanbul: Papatya Yayıncılık Eğitim.

Altaş, D., & Gülpınar, D. (2012). Karar Ağaçları ve Yapay Sinir Ağlarının Sınıflandırma Performanslarının Karşılaştırılması: Avrupa Birliği Örneği. *Trakya Üniversitesi Sosyal Bilimler Dergisi*, 1-22.

Altun, M. (2017). Veri Madenciliği ve Uygulama Alanları. 05 17, 2019 tarihinde https://docplayer.biz.tr/50639085-Veri-madenciligi-ve-uygulama-alanlari-doktora-semineri-raporu-egitim-bilimleri-bolumu-eytepe-abd-doktora-programi.html adresinden alındı

Arthur, D., Laırd, N. M., & Rubın, D. B. (1977). Maximum likelihood from ıncomplete data via the EM algorithm. *Royal Statistical Society*, 1-38.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Elsevier*, 177-194.

Ateş, Y., & Karabatak, M. (2017). Nicel birliktelik kuralları için çoklu minimum destek değeri. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 57-65.

Ay, D., & Çil, İ. (2008). Migros Türk A.Ş. de Birliktelik kurallarının yerleşim düzeni planlamada kullanılması. *Endüstri Mühendisliği Dergisi*, 14-29.

Aydemir, B. (2017). *Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Akademik Başarım Tahmini.* (Yüksek Lisans Tezi). https://tez.yok.gov.tr/UlusalTezMerkezi/.

Aydın, S., & Özkul, A. E. (2015). Veri madenciliği ve anadolu üniversitesi açıköğretim sisteminde bir uygulama. *Eğitim ve Öğretim Araştırmaları Dergisi*, 36-44.

Ayık, Y. Z., Özdemir, A., & Yavuz, U. (2007). Lise türü ve lise mezuniyet başarısının kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 441-454.

Balaban, M. E., & Kartal, E. (2018). *Veri madenciliği ve makine öğrenmesi: temel algoritmalar ve R dili ile uygulamaları.* İstanbul: Çağlayan.

Barahate, S. R. (2012). Educational data mining as a trend of data mining in educational system. *International Conference & Workshop on Recent Trends in Technology* (11-16). International Journal of Computer.

Bhullar, M. S. (2012). Use of data mining in education sector . *World Congress on Engineering and Computer Science* (24-26). San Francisco: International Association of Engineers.

Bilen, Ö., Hotaman, D., Aşkın, Ö. E., & Büyüklü, A. H. (2014). LYS başarılarına göre okul performanslarının eğitsel veri madenciliği teknikleriyle incelenmesi: 2011 istanbul örneği. *Eğitim ve Bilim*, 78-93.

Bilgin, Ş. (2013). *Banka Müşterilerinin İnternet Bankacılığına İlişkin Yaklaşımlarının Veri Madenciliği Teknikleri İle İncelenmesi.* (Yüksek Lisans Tezi). https://tez.yok.gov.tr/UlusalTezMerkezi/.

Bresfelean, V. P. (2007). *Analysis* and Predictions on Students' Behavior Using Decision Trees in Weka Environment. 29th International Conference on Information Technology Interfaces . https://www.ieee.org/

Calvo-Flores, M. D., Galindo, E. G., Jiménez, M. P., & Pérez, O. (2006). Predicting students' marks from Moodle logs using neural network models. *Current Developments in Technology-Assisted Education*. (586-590)

Can, Ş. (2017). *Veri madenciliği ve eğitim sektöründe bir uygulama.* (Yüksek Lisans Tezi) . https://tez.yok.gov.tr/UlusalTezMerkezi/.

Can, Ş. (2018). İşyerinde devamsızlık yapmaya etki eden demografik faktörlerin veri madenciliği teknikleriyle araştırılması. *Uluslararası Sosyal Araştırmalar Dergisi*, 891-900.

Chen, M. S., Han, J., & Yu, P. S. (1996, December). *Data mining: an overview from a database perspective.* IEEE Transactions on Knowledge and Data Engineering. https://www.ieee.org/

Çakmakoğlu, A. (2018). *Finans sektöründe veri madenciliği teknikleri kullanılarak kampanya modelleme (*Yüksek Lisans Tezi). https://tez.yok.gov.tr/UlusalTezMerkezi/.

Çayırlı, M. (2005). Veritabanı tasarımlarında karşılaşılan güçlükler ve çözüm önerileri. (Ders Notları) https://keciborlumyo.isparta.edu.tr/tr/dokumanlar

Çınar, A., & Silahtaroğlu, A. (2012). Veri madenciliği teknikleri ile müşteri memnuniyetine etki eden gizli etmenlerin keşfi. *Marmara Üniversitesi İİBF Dergisi*, 309-330.

Demirok, Y. (2018). *Birliktelik kuralı yöntemleri ile e-ticaret satışlarının analizi.* (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Dunham, M. H. (2003). *Data mining ıntroductory and advanced topics*. New Jersey: Perason Educaiton.

Erdal, H. İ. (2011). Destek vektör makineleri ile tahmine dayalı modelleme ve bir uygulama. (Doktora Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi.

Erpolat, S. (2012). Otomobil yetkili servislerinde birliktelik kurallarının belirlenmesinde Apriori ve FP-Growth algoritmalarının karşılaştırılması. *Anadolu Üniversitesi Soyal Bilimler Dergisi*, 137-146.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Al Magazine*, 37-54.

Giudici, P., & Silvia, F. (2009). *Applied data mining for business and ındustry*. Wiley.

Gorunescu, F. (2011). *Data mining concept, models and techniques.* Berlin: Springer.

Güner, N., & Çomak, E. (2010). Mühendislik öğrencilerinin matematik ı derslerindeki başarısının destek vektör makineleri kullanılarak tahmin edilmesi. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 87-96.

Güven, Z. B. (2016). *Türk üniversitelerindeki bilgisayar mühendisliği bölümleri müfredatları kullanılarak veri madenciliği uygulaması gerçekleştirilmesi*. (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Han, J., Pei, J., & Kamber, M. (2012). *Data mining: concepts and techniques*. Morgan Kaufman Publishers.

İnce, M. C., & Karabatak, M. (2004). Apriori algoritması ile öğrenci başarısı analizi. http://www.emo.org.tr/ekler/24f4c5eef7ec01c_ek.pdf.

Karaatlı, M., & Altıntaş, E. (2018). Borsa İstanbul İşletmelerinin Veri Madenciliği İle Kümelenmesi. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 871-886.

Karabatak, M., & Cevdet, M. (2012). *Apriori Algoritması İle Öğrenci Başarısı Analizi.* Nisan 17, 2019 tarihinde Elektrik Mühendisleri Odası: http://www.emo.org.tr/ekler/24f4c5eef7ec01c_ek.pdf adresinden alındı

Karakaynak, Z. (2014). *Destek vektör makineleri ile sınıflandırma ve görüntü tanıma üzerine bir uygulama*. (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi.

Kartal, H. B. (2012). *Envanter sınıflandırmada yapay öğrenme yöntemlerinin kullanımı ve destek vektör makineleri ile bir uygulama.* (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi.

Kılınç, Ç. (2015). *Üniversite öğrenci başarısı üzerine etki eden faktörlerin veri madenciliği yöntemleri ile incelenmesi*. (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Kiraz, A., & Deliismail, İ. (2018). İnternetten yapılan alışverişlerin veri madenciliği teknikleri ile analizi ve depo süreçlerinin iyileştirilmesi. *Bayburt Üniversitesi Fen Bilimleri Dergisi*, 28-41.

Kırlıoğlu, H., & Ceyhan, İ. F. (2014). Mali tablo denetiminde ön analitik inceleme tekniği olarak veri madenciliğinin kullanımı: borsa istanbul uygulaması. *Akademik Yaklaşımlar Dergisi*, 13-36.

Kurt, Ç., & Erdem, O. A. (2012). Öğrenci başarısını etkileyen faktörlerin veri madenciliği yöntemleriyle incelenmesi. *Politeknik Degisi,* 111-116.

Kuzey, C. (2012). Veri madenciliğinde destek vektör makinaları ve karar ağaçları yöntemlerini kullanarak bilgi çalışanlarının kurum performansı üzerine etkisinin ölçülmesi ve bir uygulama. (Doktora Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Larose, D. T. (2005). *Discovering knowledge in data*. Canada: John Wiley & Sons, Inc. Publication.

Mamcenko, J., Sileikiene, I., Lieponiene, J., & Kulvietiene, R. (2011). Analysis of e-exam data using data mining techniques. Nisan 22, 2019 tarihinde https://isd.ktu.lt/it2011/material/Proceedings/6_ITTL_3.pdf adresinden alındı

Natek, S., & Zwilling, M. (2014). Student Data Mining solution–knowledge management system related to higher education institutions. *Elsevier*, 6400–6407.

Olson, D. L., & Delen, D. (2008). *Advancent data mining techniques*. USA: Springer.

Özarslan, S. (2014). *Öğrenci performansının veri madenciliği ile belirlenmesi. (*Yüksek Lisans Tezi). https://tez.yok.gov.tr/UlusalTezMerkezi/.

Özcan, C. (2014). *Veri madenciliğinin güvenlik uygulama alanları ve veri madenciliği ile sahtekârlık analizi*. (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Özçınar, H. (2006). *KPSS sonuçlarının veri madenciliği yöntemleriyle tahmin edilmesi.*(Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Özkan, Y. (2016). *Veri madenciliği yöntemleri*. İstanbul: Papatya Yayıncılık Eğitim.

Romero , C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. ıeee transactions on systems, man, and cybernetics, part c (applications and reviews). https://ieeexplore.ieee.org

Safalı, Y., Avaroğlu, E., & Ergen, B. (2018). *Twitter verilerinden kullanıcıların siyasi eğilimlerinin veri madenciliği teknikleri ile kestirimi*. International Conference on Artificial Intelligence and Data Processing (IDAP). Malatya: IEEE.

Saygılı, A. (2013). *Veri madenciliği ile mühendislik fakültesi öğrencilerinin okul başarılarının analizi.* (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single link cluster method . *The Computer Journal*, 30–34.

Silahtaroğlu, G. (2016). *Veri madenciliği kavram ve algoritmaları*. İstanbul: Papatya Yayıncılık Eğitim.

SPSS. (2000). CRISP-DM 1.0. Nisan, 17 2019 tarihinde https://the-modeling-agency.com/crisp-dm.pdf adresinden alındı.

Şentürk, A. (2006). *Veri madenciliği kavram ve teknikler*. Bursa: Ekin Yayınevi.

Şık, M. Ş. (2014). *Veri madenciliği ve kanser erken teşhisinde kullanımı.* (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Şimşek, U. T. (2006). Veri madenciliği ve müşteri ilişkileri yönetiminde (crm) bir uygulama. (Doktora Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Şimşek, U. T. (2011). *Uygulamalı veri madenciliği-sektörel analizler.* Pegem Akademi.

Terzi, Ö., Küçüksille, E. U., Ergin, G., & İlker, A. (2011). Veri madenciliği süreci kullanılarak güneş ışınımı tahmini. *SDU International Technologic Science*, 29-37.

Tissera, W. R., Athauda, R. I., & Fernando, H. C. (2006). *Discovery of strongly related subjects in the undergraduate syllabi using data mining*. International Conference on Information and Automation . https://ieeexplore.ieee.org/Xplore/home.jsp

Uçar, E. (2013). Ortaöğretime geçiş sistemi (oges) yerleştime puanlarının uzman sistemler ile tahmini. (Doktora Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi.

Ünsal, Ö. (2011). *Mesleki alan seçimlerinin makine öğrenmesi algoritması kullanılarak belirlenmesi.* (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Vahaplar, A., & İnceoğlu, M. M. (2002). *Veri madenciliği ve elektronik ticaret. türkiye'de ınternet konferansları* VI. inet-tr.org.tr.

Vandamme, J. P., Meskens, N., & Superby, J. F. (2007). Predicting academic performance by data mining methods. *Education Economics*, 405-419.

Wang, Z., Zhu, X., Huang, J., Li, X., & Ji, Y. (2018). *Prediction of academic achievement based on digital campus.* International Conference on Educational Data Mining . https://eric.ed.gov/

Yakupoğlu, Y. (2018). *Eğitimsel veri madenciliği ve bir uygulaması.* (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

Yakut, E. (2012). Veri madenciliği tekniklerinden c5.0 algoritması ve destek vektör makineleri ile yapay sinir ağlarının sınıflandırma başarılarının karşılaştırılması:imalat sektöründe bir uygulama. (Doktora Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi.

Zerman, M. (2018). *Birliktelik kuralı algoritmaları ile büyük veriler üzerinde analitik analizler: havaalanı örneği*. (Yüksek Lisans Tezi) https://tez.yok.gov.tr/UlusalTezMerkezi/.

ÖZ GEÇMİŞ

Adı,Soyadı	Sema PEÇE			
Doğum Yeri ve Yılı	08.02.1993 / Rize			
Medeni Durumu	Bekar			
Bildiği Yabancı Diller ve Düzeyi	İngilizce/ B Düzeyi			
Öğrenim Durumu	Başlama-Bitirme Yılı		Kurum Adı	
Lisans	2012	2016	Atatürk Üniversitesi/İİBF	
Yüksek Lisans	2016	2019	Recep Tayyip Erdoğan Üniversitesi/SBE	
Doktora				
Çalıştığı Kurumlar			Başlama- Ayrılma Yılı	
1.	-			
2.	-			
3.	-			
Üye Olduğu Bilimsel ve Mesleki Kuruluşlar	-			
Katıldığı Proje ve Toplantılar	-			
Yayınlar	-			
Aldığı Ödüller	-			
İletişim(eposta)	semapece@gmail.com			