T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA

EĞİTİMDE VERİ MADENCİLİĞİ VE ÖĞRENCİ AKADEMİK BAŞARI ÖNGÖRÜSÜNE İLİŞKİN BİR UYGULAMA

Şebnem ÖZDEMİR

Enformatik Anabilim Dalı

Enformatik Programı

Danışman

Prof. Dr. M. Erdal BALABAN

Ocak, 2016

İSTANBUL

Bu çalışma 07/01/2016 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı Doktora programında Doktora Tezi olarak kabul edilmiştir.

Tez Jürisi:

Prof. Dr. Erdal BALABAN(Danışman)
İstanbul Üniversitesi
İşletme Fakültesi

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul Üniversitesi
Enformatik Bölümü

Prof. Dr. Zuhal TANRIKULU
Boğaziçi Üniversitesi
Yönetim Bilişim Bölümü

Prof. Dr. Adem KARAHOCA
Bahçeşehir Üniversitesi
Mühendislik Fakültesi

Doç. Dr. Dilek ÇAĞIRGAN GÜLTEN
İstanbul Üniversitesi
Hasan Ali Yücel Eğitim Fakültesi

ÖNSÖZ

Hayat; doğru yol göstericilerle buluştuğunuzda pek çok iyi özelliği kazanabildiğimiz ve sağlam adımlar atabildiğimiz bir labirenti temsil etmektedir. Bu labirent içinde donanımlı, ilkeli ve etik bir biçimde hareket edebilmek, bireyin aile unsuru dışında eğitim öğretim sürecindeki kazanımları ile geçirdiği değişim ve gelişime bağlıdır. Bireysel hareketlerin kitlesel sonuçları doğurduğu çağımızda, bilişim ve matematik avantajlarının eğitim-öğretim faaliyetleri gibi kritik öneme sahip bir alanda daha etkin ve daha özgün bir biçimde kullanılması kaçınılmazdır. Özellikle "Ben olacağımı hayal ettiğim şeyim" kimlik duygusu ile eğitim-öğretim sürecine girmesi beklenen bireyin, "bana öğretilenlerin tümüyüm" kimlik duygusu ile gelişirken "ben kimim" sorusuna yanıt arayışını tamamlamadan önce öngörüler oluşturularak desteklenmesi daha evrensel düşünebilen bireyin oluşumuna katkı sağlayacaktır. Bu katkı, teknoloji çağının içine doğmamış, sonradan entegre olmuş olan tez yazarı/araştırmacı açısından düşünüldüğünde, doğru yol göstericilerle buluşmuş olması şansına sahip olduğunu söylemek mümkündür. Bu şansın ilk temsilcilerinden olan, tanıştığımız andan itibaren vizyonerliği, öğrenme merakı, öğretme hevesi, bilimsel titizliği ile yol göstericim olan, saygıdeğer danışmanım Prof. Dr. M. Erdal BALABAN'a en derin saygı ve sevgilerimle teşekkür ederim.

Öğrenmekten ve öğretmekten vazgeçmediğim hayatıma, öğrenci sıfatı kazanarak yeniden başlamama yardımcı olan, her iki yüksek lisansımın danışmanlığını yürütme yüce gönüllüğü ve sabrını göstermiş değerli öğretim üyesi, manevi annem, Yrd. Doç. Dr. Zerrin AYVAZ REİS'e teşekkür ederim. Yılmadan, usanmadan beni "açınsama düzeyine eriştirmek için" emek sarfeden, rol-model öğretmenim, akıl hocam, Yrd. Doç. Dr. Zekeriya KARADAĞ'a teşekkür ederim.

Lisanüstü eğitim sürecinde gerek idari gerek akademik desteklerinden ötürü Enformatik Bölümü Başkanı Prof. Dr.Sevinç GÜLSEÇEN'e teşekkür ederim.

Tezde ihtiyaç duyulan verinin alınabilmesi için gerekli izinleri sağlayan İstanbul İl Milli Eğitim Müdürlüğü ve İstanbul Valiliği'ne, çalışma yaptığım okullarda yardımlarını eksik etmeyen, okul müdürleri, rehber öğretmenler, öğretmenler ve en önemlisi gönüllü öğrencilere teşekkürlerimi sunarım.

Bu tez çalışmasında, annesinden mahrum kalmaya razı olma büyüklüğünü gösteren, 3 yaşından beri akademik hayatın ve öğrenciliğin getirdiği her duruma, yaşından beklenmeyecek bir olgunlukla sabreden kızıma, her daim desteklerini hissettiren eşim ve aileme teşekkür ederim.

Öğrenmek ve öğretmek ideasıyla hazırlanmış bu tezin, aynı idea ile yürüyecek kişilere yol gösterici olmasını dilerim.

Ocak, 2016 Şebnem ÖZDEMİR

Bu tez çalışması Yahya ÖZDEMİR'in aziz hatırasına ithaf edilmiştir.

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **i**
İÇİNDEKİLER ..... **iii**
ŞEKİL LİSTESİ ..... **v**
TABLO LİSTESİ ..... **viii**
SİMGE VE KISALTMA LİSTESİ ..... **ix**
ÖZET ..... **x**
SUMMARY ..... **xii**
1. GİRİŞ **1**
2. GENEL KISIMLAR **5**
2.1. VERİ MADENCİLİĞİ 5
2.2. VERİ MADENCİLİĞİ MODELLERİ 9
2.2.1. Sınıflandırma 10
2.3. VERİ MADENCİLİĞİ SÜRECİ 12
2.3.1. CRISP-DM (The Cross-Industry Standard Process for Data Mining) 15
2.3.2. CRISP-EDM (The Cross-Industry Standard Process for Educational Data Mining) Modeli Önerisi 18
2.4. EĞİTİM KAVRAMI VE ÖNEMİ 21
2.4.1. Türkiye'de Eğitimin Genel Durumu 23
2.4.1.1. PISA *25*
2.4.1.2. TIMSS *28*
2.5. AKADEMİK BAŞARIYI ETKİLEYEN FAKTÖRLER 30
2.6. EĞİTİMDE BİLGİNİN KEŞFİ VE VERİ MADENCİLİĞİ 32
2.6.1. Dünyada Yapılan Çalışmalar 35
2.6.2. Ülkemizde Yapılan Çalışmalar 39
3. MALZEME VE YÖNTEM **41**
3.1. ÖRNEKLEM GRUBU VE SEÇİMİ 41
3.2. VERİ TOPLAMA ARAÇLARI VE İÇERİKLERİ 41
3.2.1. Kişisel Bilgi Formu (KBF) 43
3.2.2. Durumluk Sürekli Kaygı Ölçeği (DSKÖ) 43

3.2.3. Akademik Güdülenme Ölçeği (AGÖ) 44
3.2.4. Maslach Tükenmişlik Envanteri (MTE) 45
3.2.5. Beck Derpresyon Envanteri (BDI) 46
3.3. ANALİZ ARACI 47
3.4. SÜREÇLER 49
3.4.1. Problemi/Hedefi Tanımlama 49
3.4.2. Uygulama Adımlarını Planlama 50
3.4.3. Verileri Derleme ve Ön İnceleme 51
3.4.4. Veriyi Anlama ve Hazırlama 51
3.4.5. Modelleme 54
3.4.5.1. k-En Yakın Komşu Algoritması 55
3.4.5.2. Karar Ağacı Algoritmaları 56
3.4.5.3. Naive Bayes Sınıflandırıcı 59
3.4.5.4. Logistik Regresyon 60
3.4.5.5. Destek Vektör Makineleri (Support Vector Machine-SVM) 62
3.4.6. Modelleri Değerlendirme ve Seçme 64
3.4.7. Seçilen Modeli Uygulama 64
3.4.8. Sonucu Karar/Eylem/Yeni Girdi Haline Dönüştürme 65
4. BULGULAR 66
4.1. VERİ SETİNİN TANIMAYA YÖNELİK TEMEL İSTATİSTİK BULGULAR 66
4.2. K-EN YAKIN KOMŞU ALGORTİMASINDAN ELDE EDİLEN BULGULAR 73
4.3. NAIVE BAYES SINIFLANDIRICISINDAN ELDE EDİLEN BULGULAR 78
4.4. KARAR AĞACI ALGORİTMALARINDAN ELDE EDİLEN BULGULAR 83
4.5. LOGISTIK REGRESYON ANALİZİNDEN ELDE EDİLEN BULGULAR 92
4.6. DESTEK VEKTÖR SINIFLANDIRICI İLE ELDE EDİLEN BULGULAR 95
4.7. MODEL PERFORMANSLARININ KARŞILAŞTIRILMASI 98
4.8. SHINY 100
5. TARTIŞMA VE SONUÇ 102
KAYNAKLAR 108
EKLER 118
ÖZGEÇMİŞ 174

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1: Sınıflandırma Problemi 10

Şekil 2.2: Sınıflandırma Süreci 11

Şekil 2.3: Bilginin Keşfedilmesi Süreci 13

Şekil 2.4: Veri Ön İşleme Aşamaları 14

Şekil 2.5: CRISP Modelinin Adımları 15

Şekil 2.6: CRISP Modeli Ana Aşamalar Ve Bu Aşamalara İlişkin Eylemler 17

Şekil 2.7: CRISP-EDM Modelinin Adımları 21

Şekil 2.8: Ülkemiz Derse Geç Kalma, Ders Kırma Veya Okulu Asma Oranları 24

Şekil 2.9: Bölgeler Bazında Alanlardaki Ortalamalar 26

Şekil 2.10: Türkiye'nin Okul Ortamındaki Değişim Durumu 27

Şekil 2.11: Okul İkilimi 2003-2012 Yılları Arasında OECD-Türkiye Karşılaştırılması 27

Şekil 2.12: Yıllara Göre Türkiye TIMSS Katılım Durumu 28

Şekil 2.13:TIMSS 2011 8. Sınıfların Türkiye Ortalamasına Göre Durumları 29

Şekil 2.14: Bir Döngü Olarak EVM 34

Şekil 3.1: EDM Excel Görüntüsü 41

Şekil 3.2: Rstudio Ekran Görüntüsü 48

Şekil 3.3: İlçelere Göre Analizde Kullanılan Verilerin Dağılım 54

Şekil 3.4: Eğitim Ve Test Kümlerinin Ayrımına İlişkin Beklenti 54

Şekil 3.5: k=5 Değeri İçin En Yakın Komşu Sınıflandırıcı 55

Şekil 3.6: x ve y Nitelikleri Üzerinden Oluşturulan Basit Bir Karar Ağacı Yapısı 57

Şekil 3.7: Yüksek Boyutlarda Doğrusal Ayrılma 63

Şekil 3.8: DVM Sınıflandırıcımın Yapısı 63

Şekil 4.1: EDM Veri Setinin İlk Halinin Nitelikler Bazında Özeti 67

Şekil 4.2: DUK, SUK Ve BDI Niteliklerinin Dağılmını Gösteren Histogram Grafiği 69

Şekil 4.3: Annenin Eğitim Durumu (AB_EGIT), Babanın Eğitim Durumu (BA_EGT), Öğrencinin Ders Çalışma Ortamını (D_ORTAM), Anne İle Algılanan Bağlanma Düzeyini Gösteren (AN_BAG) Histogram Grafikleri 69

Şekil 4.4: Baba İle Algılanan Bağlanma Durumunu (BA_BAG), Tükenmşlik Puanlarının (T), Duyarsızlaşma Düzeyinin (D) Ve Akademik Güdülenme Puanlarının (AG) Dağılımını Gösteren Histogram Grafiği 70

Şekil 4.5: EDM Veri Setindeki Niteliklerin Cinsiyete Göre Dağılımını Gösteren Boxplot Grafikleri 71

Şekil 4.6: EDM Veri Setindeki Niteliklerin Histogram Yoğunluk Dağılımını Gösteren Grafikler.....72

Şekil 4.7: Okul Başarısı Ve Anne-Babanın Birliktelik Durumlarının Sınıf Mevcuduna Göre Gruplayan, Cinsiyete Göre De Dağılımını Gösteren Xyplot Grafiği 72

Şekil 4.8: Tükenmişlik Puanının (T) Okul Başarısı Kategorilerine (OKT_MUL) Göre Yoğunluk Grafiği.....73

Şekil 4.9: K-NN Algoritmasından K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Grafiği 74

Şekil 4.10: K-NN Algoritması K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Grafiği.....75

Şekil 4.11: K-NN Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Grafiği 76

Şekil 4.12: K-NN Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Grafiği 76

Şekil 4.13: K-NN Algoritması %90/%10 Tabakalı Holdout ROC Eğrisi 78

Şekil 4.14: Naive Bayes Algoritmasından K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Değerleri Grafiği 79

Şekil 4.15: Naive Bayes Algoritması K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Değerleri Grafiği 80

Şekil 4.16: Naive Bayes Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği.....81

Şekil 4.17: Naive Bayes Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği.....81

Şekil 4.18: Naive Bayes Sınıflandırıcı %90/%10 Tabakalı Holdout ROC Eğrisi.....82

Şekil 4.19: Karar Ağacı Algoritmasından K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Değerleri Grafiği 83

Şekil 4.20: Karar Ağacı Algoritması K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Değerleri Grafiği 84

Şekil 4.21: Karar Ağacı Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği 85

Şekil 4.22: Karar Ağacı Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği 85

Şekil 4.23: Karar Ağacı Algoritması %90/%10 Tabakalı Holdout ROC Eğrisi 86

Şekil 4.24: Kurulan Modelin 90/10 Ayırımda Başarısını Gösteren Rstudio Ekran Görüntüsü 87

Şekil 4.25: Kurulan Modelin 10 Kat Çapraz Geçerlemede Başarısını Gösteren Rstudio Ekran Görüntüsü 87

Şekil 4.26: Karar Ağacı Algoritmasında EDM Veri Setinin Bilgi Kazanç Değerleri 90

Şekil 4.27: EDM Veri Setindeki Nümerik Değerler İçin Pearson R Korelasyon Katsayılarını İçeren R Ekran Görüntüsü 92

Şekil 4.28: Logistik Regresyon Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği 93

Şekil 4.29: Logistik Regresyon Analizinde Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği 94

Şekil 4.30: Logistik Regresyon Analizi %90/%10 Tabakalı Holdout ROC Eğrisi 95

Şekil 4.31: Destek Vektör Sınıflandırıcı K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Değerleri Grafiği 96

Şekil 4.32: Destek Vektör Sınıflandırıcı K-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Değerleri Grafiği 96

Şekil 4.33: Destek Vektör Sınıflandırıcı Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği 97

Şekil 4.34: Destek Vektör Sınıflandırıcı Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği 98

Şekil 4.35: Karar Ağacının Shiny İle Webe Aktarılması Sürecinde Rstudio Ekran Görüntüleri 100

Şekil 4.36: Shinyapps.İo Kullanıcı -Uygulama Arayüzü 101

Şekil 4.37: Publish Edilen Karar Ağacı Modelinin Shiny Arayüzü 101

TABLO LİSTESİ

Sayfa No

Tablo 2.1: Tanımlayıcı Ve Tahminleyici Metotlar 7

Tablo 2.2: Türkiye Geneli Net Okullaşma Oranları.....24

Tablo 3.1: Cronbach Alfa Güvenirlik Katsayısı Yorum Aralıkları.....42

Tablo 3.2: AGÖ'nün Alt Faktörleri Ve Madde Numaraları.....45

Tablo 3.3: MTE-ÖĞR'nin Alt Faktörleri Ve Madde Numaraları.....46

Tablo 3.4: BDI Puanlarına İlişkin Aralıklar Ve Yorumları 47

Tablo 3.5: EDM Veri Setine İlişkin Tüm Değişkenler, Gösterim Biçimleri Ve Türleri.....51

Tablo 4.1: Okul Başarı Puanın Veri Setindek Karşılıkları.....68

Tablo 4.2: Kullanılan Algoritma/Sınıflandırıcıların Tabakalı 90/10 Tabakalı Hold Out Yöntemine Göre Genel Performanslarının Karşılaştırılması.....99

Tablo 4.3: Kullanılan Algoritma/Sınıflandırıcıların Tabakalı 10-Kat Çapraz Geçerleme Yöntemine Göre Genel Performanslarının Karşılaştırılması.....99

SİMGE VE KISALTMA LİSTESİ

Simgeler **Açıklama**

C : M Boyutlu Sınıflar Kümesi
D : Üzerinde Çalışılacak Olan Veri Setinden Elde Edilen Eğitim Veri Seti
d : Olayın Görülüş Sıklığına Göre Yapılmak İstenen + Sapma
N : Çalışma Evrenindeki Birey Sayısı
n : Veri Setindeki Nitelik Sayısı
n^1 : Örnekleme Alınacak Birey Sayısı
m : Sınıf Kümesinin Boyutu
p : İncelenecek Olayın Görülüş Sıklığı (Olasılığı)
q : İncelenecek Olayın Görülmeyiş Sıklığı (1-p)
t : Belirli Serbestlik Derecesinde Ve Saptanan Yanılma Düzeyinde t Tablosunda Bulunan Teorik Değer
X : N Boyutlu Nitelik Vektörü

Kısaltmalar **Açıklama**

AGÖ : Akademik Güdülenme Ölçeği
AÖYÖ : Akademik Öz-Yeterlilik Ölçeği
BDI : Beck Depresyon Envanteri
BİT : Bilgi ve iletişim teknolojileri
BK : Bilgiyi Kullanma
CRISP-DM : Cross Industry Standard Process for Data Mining
CRISP-EDM : Cross Industry Standard Process for Educational Data Mining
DKÖ : Durumluk Kaygı Ölçeği
DOR : Tanısal Üstünlük Değeri
DSKÖ : Durumluk Sürekli Kaygı Ölçeği
DVM : Destek Vektör Makineleri
EVM : Eğitimde veri madenciliği
K : Keşif
KA : Kendini Aşma
KBF-ÖGR : Kişisel Bilgi Formu-Öğrenci
KBF-ÖGRT : Kişisel Bilgi Formu-Öğretmen
KDD : Knowledge Discovery in Databases
MTE-ÖĞR : Maslach Tükenmişlik Envanteri-Öğrenci Formu
SKÖ : Sürekli Kaygı Ölçeği

ÖZET

DOKTORA TEZİ

EĞİTİMDE VERİ MADENCİLİĞİ VE ÖĞRENCİ AKADEMİK BAŞARI ÖNGÖRÜSÜNE İLİŞKİN BİR UYGULAMA

Şebnem ÖZDEMİR

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Prof. Dr. M. Erdal BALABAN

Eğitim-öğretim süreci ve bu sürece ilişkin tüm faaliyetler toplumların geleceğine yön verebilme gücüne sahiptir. Bu açıdan değerlendirildiğinde sürecin girdi, çıktı ve diğer süreç elemanları bakımından sıklıkla analiz edilmesi gerektiğini söylemek mümkündür. Her ne kadar bu analiz; mikro ve makro düzeyde başarı ölçme sınavları ile gerçekleştirilmekte olsa da, elde edilen başarının "istenilen başarıya olan yakınsaklığı", girdi sayısı göz önüne alındığında tartışmalıdır. Bu nedenle çeşitli faktörlerle girdinin istenen başarı düzeyine sahip çıktıya dönüşüm sürecinin önceden kestirilmesi; süreçteki aksaklıklara müdahale edilmesi gereken durumların farkındalığının oluşturulması açısından önemlidir. Tezin en genel biçimde amacı; klasik eğitim-öğretim sürecindeki, öğrenci girdisinin başarılı öğrenciye dönüşüm sürecinde literatürde yer verilen faktörlerin etkisi ışığında, başarı anlamında nasıl bir çıktı oluşturacağının öngörülmesine dayanmaktadır. Bu öngörünün oluşturulmasında, günümüzde verinin analizi açısından yararı kanıtlanmış veri madenciliği yöntemlerinden sınıflandırma teknikleri kullanılmıştır. Tezin sınırları daha özelleştirilmiş amacı ise; lise düzeyindeki öğrencilerin klasik eğitim ortamına ait akademik başarılarının, sınıflandırma teknikleri kullanılarak belirlenebilmesidir. Akademik başarıyı etkilen faktörler olarak sosyo-demografik değişkenler (yaş, cinsiyet, İstanbul'da ikamet süresi, anne-baba birlikteliği, annenin eğitim durumu, babanın eğitim durumu, annenin çalışma durumu, babanın çalışma durumu, algılanan maddi gelir düzey, günlük ortalama ders çalışma süresi, günlük ortalama internet kullanım süresi, günlük ortalama televizyon izleme süresi, eğitim hayatında sınıf tekrarı yapmış olma durumu, yükseköğretime devam etme isteği,

örnek aldığı bir rol modelin varlığı, anne ile ilişki düzeyi, baba ile ilişki düzeyi vb.) ile kaygı, tükenme, akademik güdülenme, iletişimde olduğu öğretmenlerin depresyon düzeyi gibi faktörler ele alınmıştır. Bu faktörlere ek olarak okul idaresi aracılığıyla öğrencinin yılsonu başarı ortalaması ve devamsızlık bilgisi de ele alınmıştır. İfade edilen faktörlerin tespitinde araştırmacı tarafından geliştirilen bilgi formu, izinleri alınmış/satın alınmış ölçek ve envanterler kullanılmıştır. Araştırmada kullanılan veri seti, İstanbul il sınırlarında bulunan sosyo-demografik açısından farklılara sahip ilçelerdeki lise düzeyi okullardan derlenmiştir. 2371 öğrenciden derlenen veriler değerlendirilmiş, 887'si erkek ve 819'u kadın olmak üzere 1706 öğrencinin verisinin kullanılabilir olduğu anlaşılmıştır. Tez çalışması klasik eğitim ortamından derlenen verilere sınıflandırma teknikleri uygulanması açısından ülkemizde bir ilki temsil etmektedir. Bu nedenle araştırma boyunca gerek veri toplama sürecinde gerekse analizlerde birbirinden farklı pek çok sorun ile karşılaşılmıştır. Ancak başlangıç sorunu olması nedeniyle kritik öneme sahip olan "educational data mining" ifadesinin dilimize doğru bir biçimde çevrilmesidir. Doktora çalışmasına başlandığı süreçte kavramın karşılığının henüz literatüre girmemiş oluşu nedeniyle "eğitsel veri madenciliği, eğitimsel veri madenciliği ve eğitimde veri madenciliği" ifadelerinden hangisinin daha uygun olduğu, alanında uzman matematik eğitimcisi, dilbilimci ve veri madencileri ile görüşmeler yapılarak karara bağlanmıştır. Verilerin analizi ve sınıflandırma işlemlerinin gerçekleştirilmesinde CRISP-DM (Cross Industry Standard Process for Data Mining) süreci baz alınarak geliştirilen CRISP-EDM (Cross Industry Standard Process for Educational Data Mining) süreç modeli önerisi kullanılmıştır. Sınıflandırma tekniklerinden k-En Yakın Komşu Algoritması, Naive Bayes Sınıflandırıcı, C4.5 Karar Ağacı Algoritması, Logistik Regresyon Analizi ve Destek Vektör Makineleri kullanılarak farklı modeller oluşturulmuştur. Modellerin performansları tabakalı k-kat çapraz geçerleme ve hold out yöntemleri ile kontrol edilmiş, belirli kriterler ışığında kıyaslanmıştır. Modellerin oluşturulmasında araştırmacı, R dilinde kodlar yazmış ve yine bu dilde yazılmış hazır paketleri kullanmıştır. Kodların gerçekleştirilmesinde geliştirme aracı olarak RStudio ortamından yararlanılmıştır. Yapılan analizler sonucunda C4.5 Karar Ağacı Algoritmasının akademik başarının öngörülmesine ilişkin daha başarılı sonuçlar ürettiği anlaşılmıştır. Kurulan model, tezin topluma katkı sağlaması beklentisiyle Shiny paket ve shinyappsio aracılığıyla web ortamına aktarılmıştır.

Ocak 2016, 191 sayfa.

Anahtar kelimeler: Eğitimde Veri Madenciliği, Sınıflandırma Teknikleri, Akademik Başarı, R, Shiny

SUMMARY

Ph.D THESIS

EDUCATIONAL DATA MINING AND AN APPLICATION RELATED TO PREDICTION OF STUDENT ACADEMIC SUCCESS

Şebnem ÖZDEMİR

Istanbul University

Institute of Graduate Studies in Science and Engineering

Informatics Department

Supervisor : Prof. Dr. M. Erdal BALABAN

Educational process has very important role on shaping future of society. Because of that role, the whole process with inputs, outputs and other elements should be frequently evaluated. In spite of macro and micro level achievement tests, the expected academic success can be accepted as controversial, because of the number of inputs, students in Turkey. So the prediction of academic success, in educational process, can give a crucial opportunity for prevansion and early intervention on the effect of factors. The general goal of thesis was based on predict how an output is formed as a meaning of success in the light of factors effect which included in literature during the process of the transforming student input to successful student in classic educational process. In order to build that prediction, the classification methods in data mining were used. The specific goal of that thesis based on predicting academic success of high school students in face to face educational environment by using some classification methods. The factors, effecting academic success, were specified via literature review. Those factors were socio-demographic variables (age, gender, educational status of mother, educational status of father, daily duration of studying, daily duration of TV watching, daily duration of internet surfing, grade repetition, wiling to continue for undergraduate, perceived relationship with parents, parental status and etc.), anxiety, exhaustion, academic motivation, and depression level of teachers. In addition to those factors, student's average of year-end school success and absenteeism were handled via school administration. In the determination of expressed factors, information form improved by researcher, gotten permissions/purchased scale and inventories were used. Data set used in the research, were collected from high-schools which have differences in terms of socio-demographic and settled in İstanbul provincial border. Compiled data with 2371

students were evaluated and only 1706 of them (887 male and 819 female) were used for data mining process.The Thesis has an importance because of applying the classification methods to data collected from classic education environment in our country. Therefore during the research, it was faced with many different problems about not only collecting data but also analysis. But in consequence of initial problem, the thing had critical importance was to be translated “educational data mining” in the correct way to our language. During to start doctorate study, as a consequence the term has not took part yet in our literature, it was arrived at a decision about which expression was more proper by meeting with expert math educator, linguist and data miner. In the data analysis and classification process actualization, CRISP-EDM (Cross Industry Standard Process for Educational Data Mining) which developed to base on CRISP-DM (Cross Industry Standard Process for Data Mining), process model was used. Different models were created by using classification techniques, such as K-Nearest Neighborhood, Naive Bayes Classifier, C 4.5 Decision Tree, Logistic Regression and Support Vector Machine. Performance of models, depend on choosing training and set sets, were controlled with k-fold cross validation and hold out methods and compared in the light of specific criteria. In the consist of models, researcher coded in R language and used R-packages. R studio was used for coding environment. As a result of conducted analysis, it was understood that decision tree algorithm produces more successful results which related predicting academic achievement. The model was transferred to web environment via Shiny package and shinyappsio in the expectation that thesis contributes to society.

January 2016, 191 pages.

Keywords: Educational Data Mining, Classification Techniques, Academic Succes, R, Shiny

1. GİRİŞ

Günümüz dünyası; teknolojideki pek çok değişimin insan hayatını avantajlı ve dezavantajlı durumlar oluşturacak şekilde, dolaylı ve/veya doğrudan etkilediği bir tabloyu sergilemektedir. Bu tablo içinde en kritik eylemler; doğru bilgiye, doğru kanallar kullanılarak, ihtiyaç süresi içinde erişilmesi ve kar zarar dengesi içinde en optimumum kararın üretilmesi olarak sayılabilmektedir. Birey, işletme, kurum, toplum, devlet gibi kavramlar düşünüldüğünde, mikro ve makro düzeylerde tüm bu eylemlerin sağlıklı bir biçimde gerçekleştirilmesi, kavramın etki alanına bağlı olarak kritik sonuçlara ulaşılmasına neden olmaktadır. Karara/sonuca varma yolculuğunda, "bilgi" kavramının analiz ve sentezinin yapılması önemli bir başlangıç gibi gözükse de, temelde "ham verinin (data)" işlenerek, daha değerli bir yapıya, "enformasyona (information)", dönüştürülmesi işlemleri yer almaktadır. Verinin hacmindeki büyüme, artışındaki hız, çeşidindeki farklılaşma, değeri ve doğruluğundaki belirsizlik; veriden enformasyona, enformasyondan bilgiye geçiş sürecinde, insan beyninin algılama ve örüntü keşfedebilme yeteneğinin oldukça üstünde bir yığınla karşı karşıya kalınmasına neden olmuştur. Bu yığının süzgeçten geçirilerek değerli/işe yarar verinin çıkarılması, bu veriye bağlı olarak bilginin elde edilmesi, kararlar oluşturulması, tanımlamalar, tahminler yapılması, gizli kalmış kurallara, ilişkilere erişilmesi, veri madenciliği yöntem ve tekniklerinin kullanılması ile mümkün olabilmektedir.

Veri madenciliği; istatiksel ve matematik yöntemleri kullanarak, bilgisayarın veri işleme gücünü ve insanın örüntü keşfedebilme yeteneğini bir araya getiren etkili bir yoldur. Veri madenciliği yöntem ve teknikleri; laboratuvar araştırmalarından, klinik denemelere, risk analizlerine, son derece küçük olanlardan (genomics) oldukça büyük olanlara (astrofizik), en genelden (CRM), en özelleştirilmiş olanlara (otomatik pilot), en açık olanlardan (e-ticaret) en gizli olanlara (terörün engellenmesi, banka kartı uygulamaları, cep telefonlarında dolandırıcılık tespitine), en pratik olanlardan (kalite kontrol, tedarik yönetimi) en teorik olanlara (beşeri bilimler, biyoloji, tıp ve ilaç), gereksinimlerden (tarım ve besin bilimi) eğlence sektörüne (izlenme oranlarının tahmini) kadar pek çok araştırma ve inceleme alanında uygulanmaktadır (Tufferry,

2011). Bu geniş yelpazeye bakıldığında, sektördeki farklı uygulama ve çalışma alanlarına (Harding ve diğ., 2006; Ngai ve Xiu, 2009; Yen ve diğ., 2013; Maab ve diğ., 2014) karşın, eğitim alanında veri madenciliği yöntemlerini içeren çalışmaların azlığı (Faulkner ve diğ., 2010) dikkat çekmektedir. Oysa eğitim gibi, toplumların geleceğine yön verebilme gücüne sahip bir alandan derlenen verilerin incelenmesi; mevcut sorunlara çözümler üretilebilmesi ve ileriye dönük eylem planları oluşturulmasınakatkı sağlayacaktır. Eğitim-öğretim süreci içinde irdelenmesi gereken sorunlardan bazıları aşağıdaki şekilde sıralanabilmektedir:

- Bilgideki aşırı artış nedeniyle, kime, neyi, nasıl ve hangi bilgi teknolojisi aracıyla sunulması sorunu, bilginin takip edilebilmesindeki zorluk ve öğrenenin işleme kapasitesinin üzerinde bilgiye maruz kalması: *Aşırı Bilgi Artışı, Aşırı Bilgilenme* (Meyer, 1998; Bawden ve diğ., 1999; Taslitz, 2013; Özdemir, 2015).
- Kendisinden öncekilere göre farklı öğrenme ihtiyaçları ve beklentileri olan yeni nesil öğrenci profilinin öğrenme süreci içinde tanımlanması ve anlaşılması ile öğrenme ikliminin yeniden organize edilmesi: *Dijital Yerliler, İnternet Kuşağı, Milenyumlular* (Tapscott, 1998; Howe & Strauss, 2000; Prensky, 2001; Oblinger ve Oblinger, 2005; Lancaster & Stillman, 2010).
- Gerek sosyo-ekonomik koşullar, gerekse coğrafi faktörlere bağlı olarak, bilgi ve iletişim teknolojilerine (BİT) erişimde yaşanan sorunlar nedeniyle teknoloji destekli eğitimde fırsat eşitsizliği yaşanması: *Sayısal Uçurum* (Özcivelek ve diğ., 2000; OECD, 2001; Uçkan, 2009).
- Eğitim öğretim süreci içinde kazanımlara dönüşmesi beklenen bilginin eşit biçimde dağılamaması: *Bilgi Uçurumu* (Kargbo, 2002; Akkoyunlu ve Soylu, 2010).
- Genç nüfus ve her yıl eğitim öğretim sürecine katılan bireylerden kaynaklı oluşan veri yığınları: *Büyük Veri* (Ohlhorst, 2013; Iafrate, 2014).

Eğitimde veri madenciliği (Educational Data Mining-EDM, EVM); eğitim sürecinin ham/belli miktarlarda işlenmiş çıktılarına, verilere, (öğrencinin sosyo-demografik bilgileri, başarı durumu, devamlılığı, hazırbulunuşluk düzeyi, güdülenmesi, vb), pedogojik unsurlar göz ardı edilmeksizin, veri madenciliği metotlarının uygulanması, elde edilen bilginin yine eğitim-öğretim sürecinin bir girdisi haline dönüştürülmesi ve

gerekiyorsa tekrar veri madenciliği yöntem ve teknikleri ile analiz edilmesini ifade etmektedir. Bu şekildeki iteratif yapı; mikro düzeyde; birey gibi dinamik bir unsurun daha etkili bir biçimde değerlendirilmesi ve eğitim-öğretim sürecine daha hızlı entegre edilmesini sağlamaktadır. Makro düzeyde ise eğitim-öğretim sürecini etkileyen değişkenlerin tespiti, mevcut eğitim politikalarının incelenmesi ve ihtiyaç halinde güncellenmesine imkan tanımaktadır (Streifer, 2002; Popham, 2003; Siemens ve Long, 2011). Eğitim sistemlerinin; bireyi içinde bulunduğu topluma faydalı olacak şekilde donanımlı hale getirmeyi hedeflediği ve bu hedefi eğitim politikaları üzerinden, kurumlar aracılığıyla gerçekleştirmeye çalıştığı düşünüldüğünde, EVM çalışmalarının gerekliliği daha net fark edilmektedir. Ülkemizde eğitim öğretim sürecinin başarısı; bireylerin akademik başarıları, başarı ölçme temelli sınavlar (YGS, LYS vb) gibi iç ölçümlerle, eğitim öğretim sürecine dahil olan ve bir sonraki akademik yıla başlayan öğrenci sayıları, genç nüfus ve okullaşma oranı gibi istatistiki bilgilerle, TIMSS ve PISA gibi daha küresel alanda yapılan değerlendirmelerle genel bir perspektifte yorumlanabilmektedir. Ancak bu yorumlar; neden sonuç ilişkisinin daha derin analiz edilmesi, sorunun kaynağının daha detaylı tespiti anlamında yeterli olamamaktadır. 2014 yılı verilerine göre, yaklaşık 22 milyon 838 bin 482 çocuğun %27,7'si 5-9 yaş, %27,4'ü 10-14 yaş ve %17,4'ü ise 15-17 yaş grubuna mensup çocuklardan oluşmaktadır (TUIK, 2015). Bu açıdan değerlendirildiğinde, ülke nüfusunun neredeyse çeyreğinin eğitim öğretim sürecine dahil olduğunu söylemek mümkündür. Bu önemli potansiyelin varlığına rağmen, mevcut eğitim sistemimizin kalitesi diğer OECD ülkelerinin çoğunda görülenden daha düşük olup, bu kalite sorunu farklı türlerdeki okullarda kendini daha netlikle ortaya koymaktadır (World Bank, 2011). MEB Türkiye Geneli Okullaşma (2014) bilgilerine göre, ilköğretim kademesinden ortaöğretim kademesine gidildikçe okullaşma oranları açısından ciddi bir düşüş yaşanmaktadır. Dış değerlendirme ölçütleri olarak PISA ve TIMSS sınavlarına bakıldığında, genel ortalamanın altında bir düzeye sahip olduğumuz, hatta öğrencilerin neredeyse yarısının asgari performans düzeyini dahi sağlayamadıkları görülmektedir (Yıldırım ve diğ., 2013) (Büyüköztürk ve diğ., 2014). 2015 yılı Yükseköğretime Geçiş Sınavı (YGS) istatistiklerine bakıldığında, 2 milyon 126 bin 684 adayın sınava girmek üzere başvuru yaptığı, 59.227 adayın başvurduğu halde girmediği, sınava giren adayların ise ancak 1.369.147'sinin herhangi bir puan türünde barajı aşabildiği görülmektedir (OSYM, 2015). Bu sayılar YGS 2014 değerleri ile kıyaslandığında, sınava başvuran aday

sayısındaki artışa rağmen, herhangi bir puan türünde barajı aşan aday sayısında ciddi bir düşüş olduğu fark edilmektedir.

Türk Eğitim sistemi başarı odaklı bir yapıya sahiptir (Yıldırım, 2006). En genel biçimde eğitim sisteminin görevi; öğrencinin akademik başarılarını ölçerek, elde edilen performans sonuçlarına göre başarılı olacakları alanlara yönelmelerini sağlamaktır (Silah, 2003). Gerek ülke genelinde düzenlenen başarı ölçme sınavları, gerekse OECD kıyaslamaları, mevcut akademik başarısının beklenenin altında bir tablo çizdiğini göstermektedir. Bu nedenle akademik başarıyı etkileyen faktörlerin sınıflandırmaya dayalı veri madenciliği yöntemleri kullanılarak incelenmesinin, başta bireyler bazındaki akademik başarı düşüşünün öngörülebilmesi olmak üzere pek çok avantajı da beraberinde getireceğine inanılmaktadır.

Bu tezin amacı; akademik başarının, eğitim öğretim sürecinde etkili olan faktörler göz önüne alınarak, sınıflandırmaya dayalı veri madenciliği yöntemleri ile belirlenebilmesidir. Bu amaç doğrultusunda, tezin kavramsal temellerini oluşturmak için GENEL KISIMLAR bölümünde, VERİ MADENCİLİĞİ, VERİ MADENCİLİĞİ MODELLERİ, VERİ MADENCİLİĞİ SÜRECİ, EĞİTİM KAVRAMI VE ÖNEMİ, AKADEMİK BAŞARIYI ETKİLEYEN FAKTÖRLER ve EĞİTİMDE BİLGİNİN KEŞFİ VE VERİ MADENCİLİĞİ'ne değinilmiştir. Eğitimde veri madenciliği için Çapraz Endüstri Standard Süreç Modeli'nin (Cross Industry Standard Process for Data Mining- CRISP-DM) eğitim süreci baz alınarak yeni bir model önerisi geliştirilmiş ve bu modelin adımları takip edilmiştir. Ayrıca tez çalışmasında kullanılan algoritmalara ilişkin olarak Veri Madenciliği Modelleri Bölümü'nde temel bilgiler sunulmuş, detaylı bilgiler ise Malzeme ve Yöntem Bölümü'nün altında verilmiştir.

MALZEME ve YÖNTEM bölümünde, tez çalışmasında geliştirilen akademik başarı değerlendirmesine ilişkin uygulama, CRISP-EDM model önerisine ait süreç adımları takip edilerek açıklanmıştır. Veri analizine ve analizler sonucunda elde edilen modellerin uygulanabilirliğine ilişkin sonuçlara BULGULAR'da yer verilmiştir. Yine aynı bölüm içinde sınıflandırma algoritmaları birbirleriyle kıyaslanarak, en iyi tahminleyici model seçilmiştir. Son olarak TARTIŞMA ve SONUÇ'ta tezin bütününü kapsayan bir değerlendirme yapılarak, ileride yapılabilecek araştırmalara ışık tutması için önerilerde bulunulmuştur.

2. GENEL KISIMLAR

2.1. VERİ MADENCİLİĞİ

Örüntülerin keşfedilmesi, trendlerin ve anomalilerin yoğun miktarlardaki veriler içerisinde fark edilebilmesi, bilgi çağının büyük zorlukları arasındadır. İnsan beyninin bilgiyi algılama ve işleme kapasitesindeki sınırlılık düşünüldüğünde, büyük miktarlardaki verinin izlenmesi, analiz edilerek amaca yönelik yorumlar oluşturulmasının pek de mümkün olmadığı söylenebilmektedir. Veri madenciliği kavramı; bu zorlukların aşılmasında önemli bir avantaj sağlamaktadır. Gupta (2014) göre; veri madenciliğinin giderek daha popüler hale gelmesinde, verideki büyüme, veri işleme sürecindeki maliyetin düşmesi, veri depolama kapasitelerindeki genişleme, rekabet ortamı ve veri madenciliği yazılımlarının varlığı önemli rol oynamaktadır.

Veri madenciliği; laboratuvar araştırmalarından, klinik denemelere, risk analizleri gibi farklı alanlarda varlığını sürdürmekte, son derece küçük olanlardan (genomics) oldukça büyük olanlara (astrofizik), en genelden (CRM), en özelleştirilmiş olanlara (otomatik pilot), en açık olanlardan (e-ticaret) en gizli olanlara (terörün engellenmesi, banka kartı uygulamaları, cep telefonlarında dolandırıcılık tespiti), en pratik olandan (kalite kontrol, tedarik yönetimi) en teorik olana (beşeri bilimler, biyoloji, tıp ve ilaç), gereksinimlerden (tarım ve besin bilimi) eğlence sektörüne (izlenme oranlarının tahmini) kadar farklı araştırma ve inceleme alanlarına doğru genişlemektedir (Özkan, 2008; Silahtaroğlu, 2008; Tufferry, 2011; Akpınar, 2014). Veri elde edilebilecek kaynaklar düşünüldüğünde, veri madenciliğinin oldukça geniş bir hareket alanı olduğunu söylemek mümkündür. Bu açıdan değerlendirildiğinde, veri madenciliği kavramının daha net anlaşılabilmesi için nasıl tanımlandığının da irdelenmesi gerekmektedir.

- Veri madenciliği; verideki örüntülerin geçerli, özgün, kullanışlı ve oldukça anlaşılır biçimde tanımlanması sürecidir (Fayyad ve diğ., 1996)
- Veri madenciliği; büyük veri setlerinden bilginin elde edilmesi sorununun çözülebilmesi için makine öğrenmesi, örüntü tanıma, istatistik, veri tabanları ve

görselleştirme gibi çeşitli teknikleri bir araya getiren dişilinler arası bir çalışma sahasıdır (Cabena ve diğ., 1998).

- Veri Tabanlarında Bilginin Keşfi (Knowledge Discovery in Databases -KDD) olarak da adlandırılan veri madenciliği, çeşitli problemlerin çözümünde depolanmış verilerin analiz edilmesini içermektedir (Witten ve Frank, 2000).
- Veri madenciliği; veriyi işleyen için anlaşılır ve faydalı olacak şekilde varlığı bilinmeyen ilişkilerin ortaya çıkarılması ve verinin özgün bir şekilde özetlenmesi için büyük, gözleme dayalı elde edilen veri setlerinin analiz edilmesidir (Hand ve diğ., 2001)
- Veri Madenciliği; keşfetme alarak adlandırılan bir gelişim süreci içinde gerçekleşen iteratif bir süreçtir (Kantardzic, 2011)
- Veri madenciliği; istatiksel ve matematik yöntemlerin kullanılarak, anlamlı yeni korelasyonların, örüntülerin ve trendlerin depolanmış olan büyük miktarlardaki veriden elenerek keşfedilmesi sürecidir (Gartner Group, 2013)
- Veri madenciliği veya KDD; analitik metotlara ve araçlara dayalı keşfetme teknikleri ile büyük yığınlar halindeki bilgi ile baş edilmesidir (Gupta, 2014).

Veri madenciliği tanımları incelendiğinde, kullanılan yöntemden bağımsız olarak yapılan işlemin; karar verme sürecine yardımcı bilgi üretilmesi ve bu sayede gerçek hayatın anlaşılabilir düzeyde bir modelinin oluşturulmasını içerdiği söylenebilir.

Kullanılan yöntemler açısından incelendiğinde ise, veri madenciliğinin en temel düzeydeki amacının; tanımlama ve tahmin yapmak olduğu ifade edilebilir. Tanımlama; yorumlanabilir veriyi içeren örüntüleri bulmayı, tahmin ise; veri setindeki bazı değişkenleri ya da alanları kullanarak, diğer değişkenlerin bilinmeyen/ geleceğe ilişkin değerlerini tespit etmeyi içermektedir.

Veri madenciliği yöntemlerinden tanımlamaya dayalı yöntemler; baş edilmesi mümkün olmayan büyüklükteki verilerde bugüne ilişkin gömülü bilginin çıkarılmasına olanak sağlamaktadırlar (Tufferry, 2011). Bu sayede mevcut veri setine dayalı yeni ve çözülmesi kolay olamayan bilginin üretilmesi mümkün olabilmektedir (Kantardzic, 2011). Tahmine dayalı olanlar ise; bugünün bilgisine/verisine bağlı olarak ileriye dönük, nitel ya da nicel biçimdeki yeni bilginin oluşturulmasına imkan tanımaktadırlar (Tufferry, 2011). Diğer bir deyişle veri setleri ile tanımlanmış sisteme yönelik, model

üretilmesini sağlamaktadırlar (Kantardzic, 2011). İster tanımlama ister tahmine dayalı yöntemler kullanılsın, veri madenciliği yöntem ve teknikleri; bilgisayarın veri işleme gücünü kullanarak, büyük yığınlar halindeki veriyi, insan zihninin örüntü/ilişki keşfetme yeteneği için kullanılabilir/işlenebilir hale indirgemektedirler.

Veri madenciliğinde kullanılan metotlar; temel düzeydeki tanımlayıcı istatistikten çok daha karmaşık yapıyı barındırmaktadırlar (Tufferry, 2011). Bu metotların en genel haliyle tanımlamaya ve tahmine dayalı yöntemler olduğundan daha önce bahsedilmişti. Tablo 2.1'de bu metotların içerikleri, kullandıkları algoritmalar sunulmaktadır.

Tablo 2.1: Tanımlayıcı ve Tahminleyici Metotlar (Tufferry, 2011)

Tür	Aile	Alt-Aile	Algoritma
Tanımlamaya Dayalı Yöntemler	Geometrik Modeller	Faktör Analizleri (Düşük Boyuttaki Uzayda Yansıtma ve Görselleştirme)	Principal Component Analysis (PCA) (Sürekli Değişkenler) Correspondence Analysis (CA) (Nitel ve İkili Değişkenler) Multiple Correspondence Analysis (CA) (nitel ve ikili değişkenler)
		Kümeleme Analizleri (Uzayın tamamında homojen kümeler içinde gruplama)	Partitioning Methods (moving centres, k-means, dynamic clouds, k-medoids, vb.) Hiyerarşik Metotlar (Agglomerative, Divisive)
		Kümeleme Analizlieri+ Dimension Reduction	Neural Clustering (Kohonen maps)
	Combinatory Models		Clusteering by aggregation of similarities (nitel

			değişkenler)
	Mantıksal Kural Tabanlı Modeller (logical rule-based models)	Link detection	Birliktelik kurallarının araştırılması Benzer dizilimlerin araştırılması (search for similar sequences)
	Mantıksal Kural Tabanlı Modeller (logical rule-based models)	Karar Ağaçları	Karar Ağaçları (bağımlı değişkenler nümerik veya nicel)
Tahmine Dayalı Yöntemler	Matematiksel Fonksiyonlara Dayalı Modeller	Parametrik ya da Yarı-Parametrik Modeller	Lineer Regresyon, ANOVA, MANOVA, ANCOVA, MANCOVA, Genelleştirilmiş Lineer Model (GLM), PLS Regresyon (sürekli bağımlı değişkenler) Fisher Diskirminant Analizi, Logistic Regresyon, PLS Logistic Regresyon (nicel bağımlı değişken) Logaritmik Lineer Model (bağımlı değişken) Genelleştirilmiş Lineer Model (GLM), Generalized Additive Model (GAM)
	Modelsiz Öngörü (Predction Without Model)	Probablistik Analiz	K Yakınsak Komşular

Tablo 2.1. incelendiğinde veri madenciliğine yönelik pek çok metot ve algoritma bulunduğu görülmektedir. Ancak bu algoritmalar/metotlardan hangisinin seçilmesi gerektiği, yapılacak analizdeki ana amaç ile ortaya çıkabilecektir. Bu ana amacın belirlenmesinden sonra mevcut veri madenciliği modellerinin irdelenmesi, metot ve algoritma konusunda daha sağlıklı karar verilmesini sağlayacaktır. Bölüm 2.2'de veri madenciliği modelleri ele alınarak, tez kapsamında kullanılan sınıflandırma yöntemi hakkında detaylı bilgiler sunulacaktır.

2.2. VERİ MADENCİLİĞİ MODELLERİ

Bölüm 1'de farklı yıllara ait veri madenciliği tanımları verilerek, kavramın yolculuğu, güncellenen ve değişmeyen tarafları paylaşılmıştı. Tanımlardan da genel olarak çıkarılabilen veri madenciliği sürecinin faaliyetleri ve faaliyet detayları, seçilecek modele göre daha netlik kazanacaktır.

Veri madenciliği modellerini en genel hali ile üç kategoride değerlendirmek mümkündür. Bunlar; birliktelik kuralları, kümeleme ve sınıflandırma şeklinde ifade edilebilir.

Birliktelik kuralları; veritabanı içindeki kayıtların birbirleriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerinin ortaya çıkarılmasına dayanmaktadır (Özkan, 2008). Özellikle pazarlama alanında kullanılan birliktelik kuralları, pazar sepet analizi gibi uygulamalarla tüketim alışkanlıklarının tespit edilmesi sağlamaktadır.

Kümeleme modeli; sınıf içi benzerliğin maksimum, sınıflar arası benzerliğin minimum olması prensibine göre nesneler kümelenmekte veya gruplandırılmaktadır (Han ve diğ., 2012). Veriler arasındaki ayırım; küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesi ile gerçekleştirilmektedir. Başlangıç aşamasında veri tabanındaki kayıtların hangi kümelere ayrılacağı (kaç farklı kümeden oluşması gerektiği), veya kümelemenin hangi değişken özelliklerine göre yapılacağı netlikle bilinemeyebilir. Kümeleme işlemlerinde verilerin doğru biçimde kümelenmesi birkaç iterasyon sonrasında oluşabilir.

Bu tez çalışmasında birliktelik kuralları ve kümeleme dışında kalan sınıflandırmaya dayalı algoritmalar kullanılacaktır. Bu nedenle sınıflandırma modeli 2.2.1. Sınıflandırma kısmında ayrıntılı olarak anlatılmaktadır.

2.2.1. Sınıflandırma

Sınıflandırma; nesneleri önceden tanımlı kategorilere atamaya dayalı, çok çeşitli uygulamaları barındırabilecek nitelikteki geniş ölçekli bir problemdir (Tan ve diğ., 2006). Daha matematiksel olarak sınıflandırma; nitelik kümesi *A*'nın, sınıf etiketi *B*'ye eşlenmesi olarak ifade edilebilir. Şekil 2.1'de sınıflandırma probleminin genel yapısı verilmektedir.

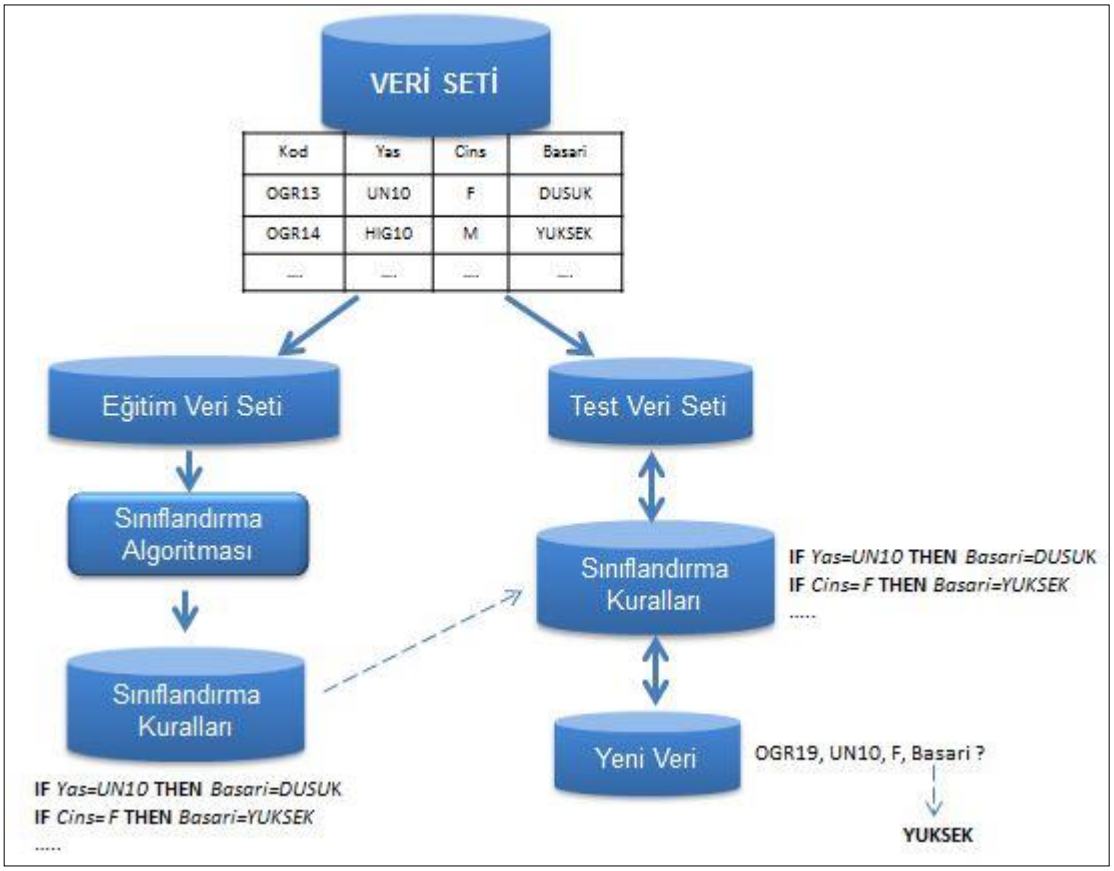
Şekil 2.1: SınıflandırmaProblemi (Tan ve diğ., 2006).

Sınıflandırma; veriyi, sınıflara ya da kavramlara tanımlayacak ve ayırabilecek bir modelin ya da fonksiyonun bulunması sürecidir (Han ve diğ., 2011). Tan ve diğ. (2006) göre ise sınıflandırma; her nitelik kümesi *x*'ileri, önceden tanımlı sınıf etiketleri *y*'lere eşleyen *f* hedef fonksiyonunun öğrenmesi görevidir. *f* hedef fonksiyonu sınıflandırma modeli olarak da adlandırılabilmektedir.

Sınıflandırma süreci; iki temel adımı içermektedir:

- Öğrenme adımı (sınıflandırma modelinin yapılandırıldığı adım)
- Sınıflandırma adımı (test verisi üzerinde sınıf etiketlerinin, sınıflandırma modeli kullanılarak tahmin edildiği adım)

Şekil 2.2'de Han ve diğ. (2011)'den esinlenerek oluşturulmuş olan sınıflandırma süreci gösterilmiştir.



Şekil 2.2: Sınıflandırma Süreci.

A1, A2, …, An ile tanımlanmış niteliklere sahip veritabanından, eğitim kümesi oluşturmak üzere seçilenler, ***X, n*** boyutlu bir nitelik vektörünü oluşturmaktadırlar.

$$X = (x_1, x_2, \dots, x_n)$$

Eğitim kümesinin seçiminden sonra bu veri sınıfları kümesine dayalı bir sınıflandırıcı **(classifier)** oluşturulur. Öğrenme adımında **(learning step or training phase)**; sınıflandırma algoritması, eğitim kümesini analiz ederek, ondan "öğrenerek", bir sınıflandırıcı kurgular (Han ve diğ., 2012).

Veri madenciliğinde sınıflandırma yöntemi, mevcut veri dizisinin istatistik, makine öğrenmesi gibi yöntemler kullanılarak daha önce belirlenmiş olan sınıflara atanması işlemini temsil etmektedir (Akpınar, 2014). Bu yöntem veritabanındaki, gizli örüntüleri ortaya çıkarma amacıyla kullanılmaktadır (Özkan, 2008). Bu örüntülerin keşfedilmesinde; ağaç sınıflandırıcıları, kural tabanlı sınıflandırıcıları (rule-based classifiers), yapay sinir ağları, destek vektör makinaları, naive bayes sınıflandırıcılar

kullanılabilir. Bu tekniklerin hepsi, nitelik kümesi ile girdi verisinin sınıf etiketi arasındaki ilişkiye en iyi uyan modeli tanımlayan bir öğrenme algoritması çalıştırmaktadırlar.

Sınıflandırma yöntemlerinden istatistik temelli olanlar; lineer regresyon analizi, lojistik regresyon analizi, diskriminant analizi, bayes sınıflandırma yöntemleridir. Makine öğrenmesine dayalı olarak ise karar ağaçları, en yakın komşu yöntemi, yapay sinir ağları, destek vektör makineleri şeklindedir. Bu tez çalışmasında kullanılacak olan modeller, Bölüm 3.4.5'te özetlenmektedir.

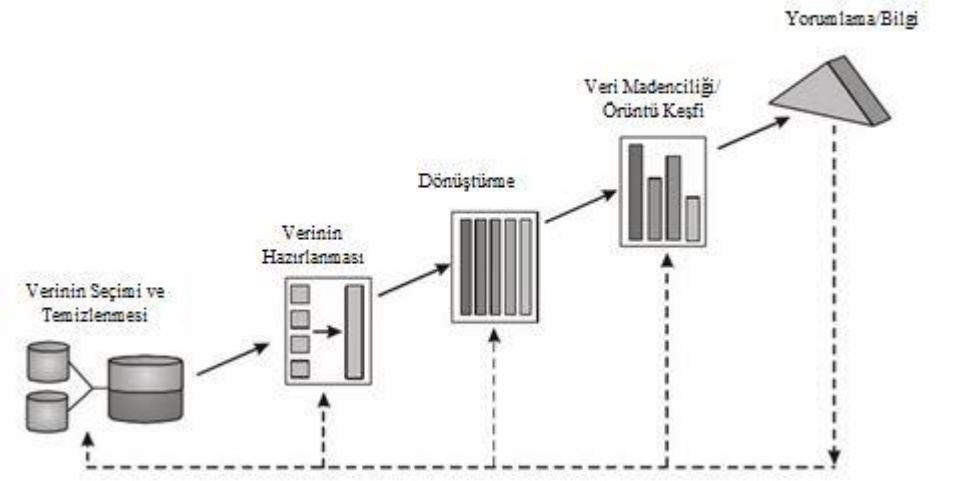
2.3. VERİ MADENCİLİĞİ SÜRECİ

Veri Madenciliği çalışmaları, bazı ortak özellikleri olan farklı projeler gibidir (Gupta, 2014). Bu projelerin başarısında izlenmesi gereken birtakım temel adımlar bulunmaktadır (Lavrac ve diğ., 2004; Tufferry, 2011):

- Hedeflerin tanımlanması
- Verinin toplanması/mevcut verilerin listelenmesi
- Verinin keşfedilmesi ve hazırlanması
- Modellerin uygulanması
- Performansların karşılaştırılarak en uygun modelin seçilmesi
- Seçilen model ışığında sonuçlar/yorumlar üretilmesi
- Sonuçların eylemlere dönüştürülmesi

Bu temel adımlar içerisinde en kritik olan kısım; verinin analize uygun hale getirilme süreci ile alakalı olandır.

Pujari (2001) veri madenciliğini; bilginin keşfedilmesi olarak adlandırılan geniş bir sürecin bir bileşeni olarak kabul etmektedir. Şekil 2.3'de veri bilginin keşfedilmesi süreci adımları verilmektedir.

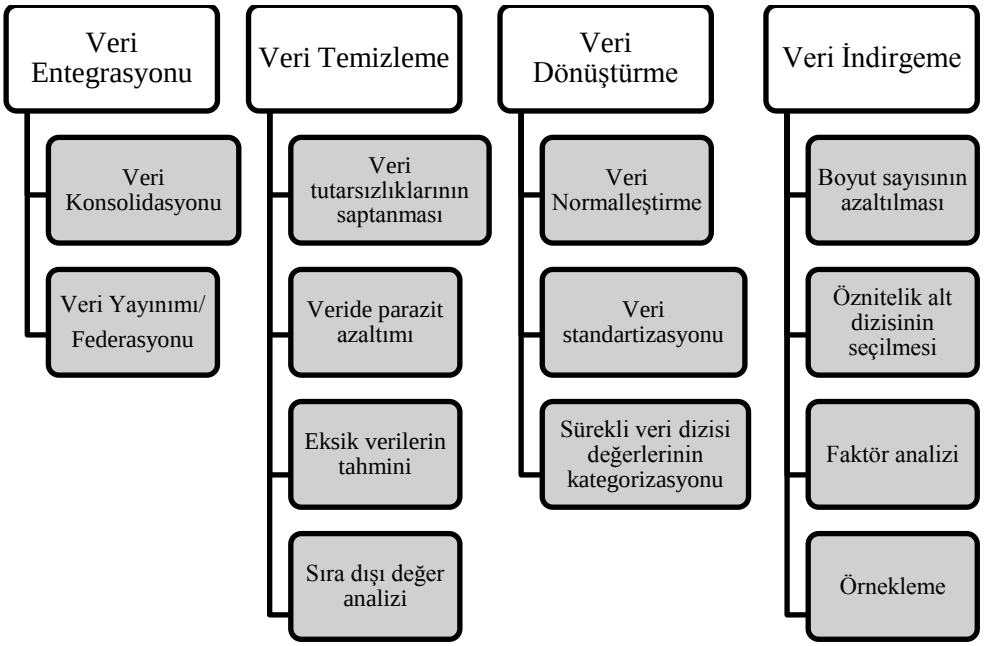


Şekil 2.3: Bilginin Keşfedilmesi Süreci (Pujari, 2001).

Han ve Kamber (2006) bilgi keşfini, ardışık bir takım adımların tekrarlanması olarak niteleyerek, veri madenciliğini bu adımlardan biri olarak sunmuşlardır (Han ve diğ., 2012):

- Gürültülü ve tutarsız verilerin kaldırılması ile elde edilen verilerin temizlenmesi
- Farklı türdeki kaynaklardan elde edilen verilen bir araya getirilmesi ile verinin bütünleştirilmesi
- Bilginin keşfi sürecinde konulan hedefe uymayan/gereksiz olan verilerin veritabanından çıkarılması
- Yine hedefe uygun olacak şekilde verinin uygun biçimlere dönüştürülmesi
- Gizli bilginin/örüntünün çıkarılabilmesi için matematik ve bilişim tabanlı sistemlerin kullanılması (veri madenciliği)
- Elde edilen desenlerin değerlendirilmesi ile asıl ilgi çeken desenin tanımlanması
- Keşfedilen bilginin, çeşitli görselleştirme teknikleri ve bilgi sunuş türleri kullanılarak gösterilmesi

Bilgiyi keşfetme sürecinde en kritik kısım verinin hazırlanması ile ilgili olandır. Bu süreçte, toplanan veri, veri madenciliği teknikleri uygulanmadan önce ön işlemeye tabi tutulmaktadır. Şekil 2.4'de veri ön işlemenin aşamaları verilmektedir.



Şekil 2.4: Veri Ön İşleme Aşamaları (Akpınar, 2014).

Veri ön işleme adımında ölçme ve/veya kodlamaya bağlı hatalar sonucu oluşan uç değerlerin, gözlemde oluşan anormal değerlerin ve nitelikleri eksik verilerin tespiti söz konusudur (Feelders ve diğ., 2000). Veri tabanındaki tutarsız ve hatalı veriler gürültülü veri olarak adlandırılmaktadır. Uç değerler, gürültülü veriler, veri ön işleme adımında analizi yapılacak veri setinde tespit edilmelidir. Eksik kalan veriler için aşağıda verilen yöntemlerin kullanılması mümkündür (Özkan, 2008) ;

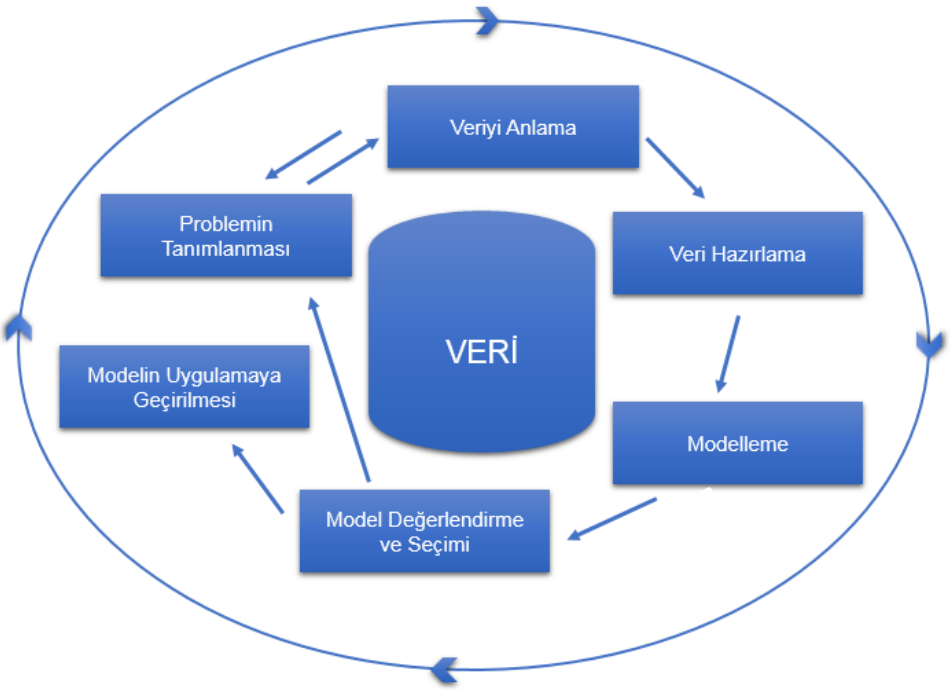
- Eksik değer içeren kayıtların üzerinde çalışılacak veri kümesinden atılması
- Kayıp değerler yerine aynı sabit değerin/ifadenin kullanılması
- Değişkene ait verilerin ortalaması alınarak, eksik değerler yerine bu ortalama değerin atanması
- Sadece bir sınıfa ait örneklerin değişken ortalaması alınarak eksik değer yerine kullanılması
- Regresyon, jack knife veya karar ağacı modeli kurularak eksik değerlerin tahmin edilmesi ve bu tahmin değerinin/değerlerinin atanması

Literatür incelendiğinde; veri madenciliğinde izlenmesi gereken belli başlı adımların neler olduğunu keşfedebilmek mümkündür. Bu adımların belli bir koordinasyonla uygulanması, veriden bilgiye uzanan yolculuğun daha nitelikli ilerlemesini sağlayacaktır. Veri madenciliği sürecine yönelik olarak SPSS liderliğinde geliştirilen

CRISP-DM ve SAS Institute Inc. tarafından Enterprise Miner için geliştirilen SEMMA'dan bahsedebilmek mümkündür. Bu tez çalışmasında sınıflandırmaya dayalı eğitimde veri madenciliği uygulaması gerçekleştirilirken Bölüm 2.3.1.'de verilmekte olan CRISP-DM modeli baz alınarak geliştirilen, Bölüm 2.3.2'de sunulmakta olan CRISP-EDM model önerisine ait adımlar izlenmiştir.

2.3.1. CRISP-DM (The Cross-Industry Standard Process for Data Mining)

CRISP-DM 1996 yılında geliştirilmiş ve Daimler Chrysler (SPSS Inc.) ve NCR Corporation tarafından tanıtılmıştır. CRISP-DM iteratif ve adaptif bir süreçtir. Bu sürece göre bir veri madenciliği projesi altı aşamadan oluşan bir yaşam döngüsü barındırmaktadır. Bu aşamaların sırası kesin olmamakla birlikte, her bir aşamanın çıktısı bir sonraki aşamanın ne olacağı ya da aşamaya ilişkin hangi görevin yerine getirileceğini belirler (Chapman ve diğ., 2000; Akpınar, 2014). Şekil 2.5'de CRISP modelinin adımları verilmektedir.



Şekil 2.5: CRISP Modelinin Adımları (Chapman ve diğ., 2000; Balaban ve Kartal, 2015).

CRISP modeli; ilk olarak problemin tanımlanması ile başlar. İşletmeler açısından bakıldığından bu adım çeşitli kaynaklarda işin/araştırmanın anlaşılması (business or research understanding phase) olarak ifade edilmektedir (Chapman ve diğ., 2000;

Akpınar, 2014). Bu aşama; ilk etapta bir projedeki gibi çıktıların ve gerekliliklerin belirlenmesi, problemin tanımlanması gibi çeşitli eylemleri barındırmaktadır. Bu eylemlerin netlikle ortaya konulmasından sonra bunların bir veri madenciliği problemi gibi tanımlanması söz konusudur. Problemin başlangıç adımında ortaya konulması, veri madenciliği yapılacak alana/probleme özgü bir yol haritası çizilmesini sağlayacaktır (Balaban ve Kartal, 2015).

Problemin tanımlanması aşamasından sonraki iki aşama; verinin anlaşılması ve hazırlanması aşamaları, ham verinin analize uygun hale getirilmesine yöneliktir (Balaban ve Kartal, 2015). Bu adımlarda, eldeki veri setinin hangi kaynaklardan derlendiği ve ne tür verilerden oluştuğunun belirlenmesi/netleştirilmesine ek olarak, verideki uç değerler, tekrar eden gözlemler, eksik/hatalı değerlerin tespit edilmesi, gerekli ise normalize edilmesi bulunmaktadır. Veri madenciliğinin herhangi bir problemi otomatik olarak çözen bir algoritma ya da araç olduğu düşünülmemelidir. Çözümün elde edilmesinde verinin kaliteli hale getirilmesi gerekmektedir. Verinin anlaşılması ve hazırlanması sürecinin atlanması ve/veya özenli yapılmaması ile "körlemesine veri madenciliği yazılımlarının kullanılması söz konusu olacaktır. Bu durum yanlış türdeki veriye uygulanmış yanlış sorulardan elde edilen yanlış cevapların eldesine neden olacaktır (Larose ve Larose, 2014)

Modelleme (Modelling) aşaması; çeşitli modelleme tekniklerinin seçimi, uygulanması ve parametrelerin optimal değerlere göre ayarlanmasını içermektedir. Aynı veri madenciliği problemi için farklı teknikler kullanılması mümkündür. Ancak bazı teknikler verinin formuna yönelik kısıtlara sahip olabilirler. Bu nedenle verinin hazırlanması aşamasına sıklıkla dönülebilmektedir (Chapman ve diğ., 2000).

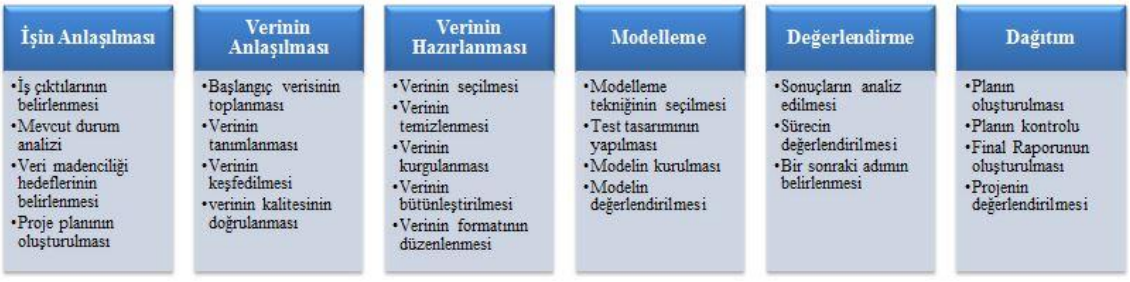
Model değerlendirme ve seçim (evaluation) aşamasında veri analizi bakış açısına göre yüksek kalitede model/modeller kurulduğu varsayılmaktadır. Bir sonraki aşamaya geçilmeden modelin isteneni verebilecek düzeyde olup olmadığı detaylı şekilde incelenmelidir. Bu incelemede esas; gözden kaçırılmış ya da üzerinde yeterli derecede düşünülmemiş noktalar olup olmadığı belirleyebilmektir. Bu incelemenin yanı sıra kurgulanmış modelin/modellerin performans açısından değerlendirmesi de yine bu adımda yapılmaktadır. Literatürde model performansına yönelik olarak holdout (dışarıda bırak), tabakalı örnekleme (stratified sampling), üçlü ayırma (three-way split),

çapraz geçerleme (cross validation), bootstrap örnekleme ve random subsampling (tesadüfi alt kümeleme) gibi yöntemlere yer verilmektedir (Tan ve diğ., 2006; Balaban ve Kartal, 2015)

Modelin performansının değerlendirilmesinde, özellikle sınıflandırma modellerinde, hedef niteliğin gerçek değerinin ve sınıflayıcı ile üretilen tahmin değerlerinin birlikte sunulduğu kontenjans tablosuna (confusion matrix) dayalı ölçütler kullanılabilmektedir. Bu ölçütler arasında, doğruluk (ACC), duyarlılık (TPR), belirleyicilik (TNR/SPC), yanlış pozitif oranı (FPR), yanlış negatif oranı(FNR), kesinlik(PPV), negatif öngörü değeri(NPV), F-ölçütü, pozitif olabilirlik oranı(LR+), negatif olabilirlik oranı(LR-), tanısal üstünlük değeri (DOR) ve ROC eğrileri sayılabilir (Balaban ve Kartal, 2015).

Modelin/modellerin performansının değerlendirilmesinden sonra artık belirlenen modelin uygulanması ile sonuçlara yönelik karar üretilmesi ya da yetersiz performans değerleri ışığında problemin yeniden tanımlanması aşamasına geçilebilir.

CRISP modeli, işletmeler açısından ele alındığında son aşama olarak kullanıcılarla paylaşım (deployment) ile sunulmaktadır. Modelin oluşturulması ve bazı kararlara varılması projenin sonlandığı anlamına gelmemektedir (Chapman ve diğ., 2000). Modelin temel amacı veriden bilgi üretimi olsa bile elde edilen bu bilginin organize edilmesi ve anlaşılır bir biçime getirilerek sunulması gerekmektedir. Bu aşama ihtiyaçlara bağlı olarak sadece bir rapor oluşturulması gibi basit bir eylemi içerebileceği gibi, veri madenciliği sürecinin tekrarlanması gibi daha karmaşık eylemler dizisi barındırabilir.



Şekil 2.6: CRISP Modeli Ana Aşamalar Ve Bu Aşamalara İlişkin Eylemler.

CRISP modeli kanıt madenciliği (Venter ve diğ., 2007); banka dolandırıcılıkları (Da Rocha ve de Sousa Júnior, 2010) ve pazarlama kampanyaları (Moro ve diğ., 2011) olmak üzere çeşitli çalışmalarda kullanılmıştır.

2.3.2. CRISP-EDM (The Cross-Industry Standard Process for Educational Data Mining) Modeli Önerisi

CRISP-EDM; CRISP-DM sürecinden esinlenerek oluşturulmuş, eğitim sürecinin getirdiği bir takım farklılıkları içeren, veri madenciliği sürecine yönelik bir model önerisidir. CRISP-EDM'de izlenecek adımların aşağıdaki gibi olması önerilmektedir:

- Problemi/hedefi tanımlama
- Uygulama adımlarını planlama
- Verileri derleme ve ön inceleme
- Verileri anlama ve hazırlama
- Modelleme
- Model değerlendirme ve seçme
- Seçilen modeli uygulama
- Sonucu karar/eylem/yeni girdi haline dönüştürme

Problemi/hedefi tanımlama aşaması; CRISP-DM'de olduğu gibi eğitimde veri madenciliği yöntem ve tekniklerinin uygulanmasını gerektirecek bir problemin varlığı/hedefin kurgulanması ve bu problemin/hedefin net bir biçimde tanımlanmasını içermektedir. Bu sayede araştırmacı yapacağı çalışmanın sınırlarını ve sınırlılıklarını daha doğru bir biçimde görecektir.

CRISP-EDM adımlarındaki farklılık *uygulama adımlarını planlama* kısmı ile başlamaktadır. Bu aşamada, tespit edilen probleme ve/veya belirlenen hedefe göre örneklemin tanımlanması, veri derleme yöntemleri ve araçlarının netleştirilmesi, veri kaynaklarının çeşitliliğinin belirlenmesi işlemleri gerçekleştirilmelidir. Örneklemin özellikle tanımlanan problemi/hedefi karşılayacak şekilde tasarlanması gerekmektedir. Yanlış, eksik niteliğe sahip ya da yetersiz kurgulanan örneklem; derlenen verinin kalitesinden bağımsız olarak, doğru olmayan, manipülatif yargılara varılmasına neden olacaktır.

Uygulama adımlarını planlama aşamasında, örneklemin seçiminden sonra, bu örneklemden derlenecek verilere ilişkin bir takım kararlara varılmalıdır. Örneğin bir A dersini bırakacak öğrencilerin tahmin edilmesinde LCMS'de (Learning Content Management System) hali hazırda kayıtlı veriler/veritabanları/veri ambarları

kullanılacak ise (sistemde bağlantılı kalma süresi, ara sınav notları, etkileşim sayısı vb.), çok ayrıntılı uygulama adımları planlamaya gerek olmadan *verileri anlama ve hazırlama aşamasına* geçilmesi uygun olacaktır. Ancak klasik eğitim ortamından veri toplanması söz konusu ise, veri toplama araçlarının doğru analiz edilerek yine hedef çerçevesinde uygun/uygulanabilir anket, ölçek ve envanter seçimi yapılmalıdır. Araştırmacının böyle bir durumda pedagojik unsurları göz ardı etmeksizin, uygulanacak anket/ölçek/envanterin niteliğine göre uygulama sırası düzenlemesi beklenmektedir. Örneğin toplamda 100 soruluk bir ölçek ve envanter takımının aynı zaman dilimi içerisinde uygulanması, cevaplayıcının yaş ve motivasyonuna göre gerçeklikten uzak verilerin eldesine neden olabilir. Bu noktada hangi ölçeğin/envanterin ne zaman uygulanacağı, ne sıklıklarla çalışmalar yapılacağı doğru biçimde planlanmalıdır.

Verileri derleme ve ön inceleme aşaması; veri toplama işlemi tamamlanmadan başlayan ve tamamlanana kadar devam eden bir aşamadır. Araştırmacı özellikle klasik ortamdan derlemekte olduğu verileri "eksiklik içeren", "eksik sayıda veri toplanmaması" adına periyodik olarak denetlemelidir. Eğitim-öğretim süreci psikolojik değişkenleri barındırmaktadır. Bu nedenle, uygulanan ölçeğin/envanterin eksik cevaplanması durumunda, standart veri tamamlama/kayıp değer tahmini yöntemleri ile gerçeğe yakın cevaplar üretilmesi soru işaretleri barındırması nedeniyle tercih edilmeyecektir.

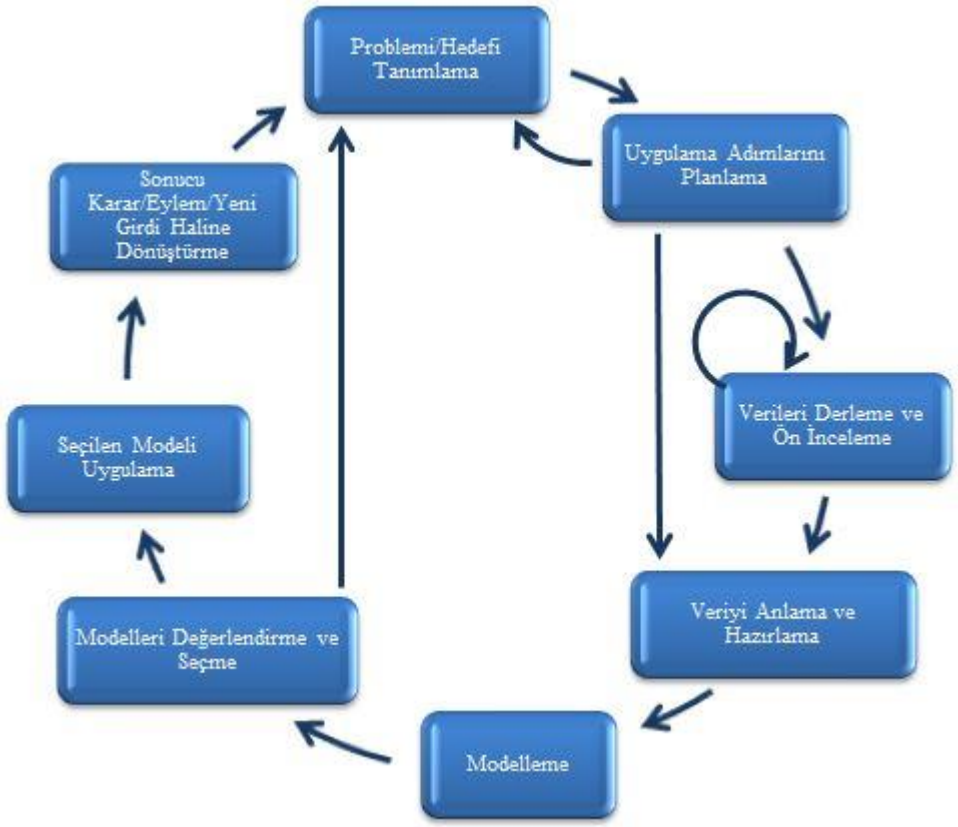
Verileri anlama ve hazırlama aşaması; derlenen verilerin uygun biçimsellikte analiz ortamına aktarılması için yapılacak çalışmaları içermektedir. Araştırmacı dosya uzantılarından, niteliklerinin nümerik, kategorik olma durumlarına kadar detaylı bir inceleme yapmalıdır. Örneğin annenin eğitim durumunun sayılar cinsinden ifadesi (1-ilkokul, 3-lise gibi), sayısal bir değer gibi girilse de gerçek bir sayı değerini ifade etmeyecek, o nitelik için bir durum sınıflaması yapmış olacaktır. Bu aşamada veri toplama işlemi de tamamlandığından, ölçek/envanterler kullanımı söz konusu ise gerekli güvenirlik çalışmaları yapılmalıdır. Cevaplayıcıların gerçekçiliği ve verilerin kullanımına yönelik şüpheler bu şekilde giderilmiş olacaktır.

Modelleme aşaması; CRISP-DM'de olduğu gibi çeşitli modelleme tekniklerinin seçimi ve uygulanmasını barındırmaktadır. Tek bir veri madenciliği yöntemi kullanılması gibi bir kısıt söz konusu değildir.

Model değerlendirme ve seçme aşaması; çeşitli performans yöntemleri kullanarak uygulanan yöntem ve teknikleri ile "eldeki veri setinin doğru bir biçimde madenciliğinin yapılıp yapılmadığının" denetlendiği aşamadır. Bu aşamada çeşitli performans ölçütlerinin yanı sıra model ile elde edilen enformasyonun eğitim-öğretim süreci içindeki anlamlılık düzeyi de irdelenmelidir.

Seçilen modeli uygulama aşaması, mevcut modeller içinde problem/hedef bazında en uygun olduğu düşünülen modelin, ürettiği sonuçların değerlendirilmesi ve eğitim süreci açısından daha detaylı sonuçlar üretebilecek şekilde yeniden yorumlanmasını barındırmaktadır.

Sonucu karar/eylem/yeni girdi haline dönüştürme aşaması ise, kullanılan modelin yorumlanarak sürece fayda sağlayacak çıktılar oluşturulmasını içermektedir. Bu çıktılar kurallar bütünü olabileceği gibi, daha önce etkisi fark edilmemiş ilişkiler yumağı da olabilir. Eğitim-öğretim süreci aktif işleyen bir yapıya sahiptir. Bu nedenle uygulanan modelin çıktıları, yeni girdilere dönüşebilir. Şekil 2.7.'de CRISP-EDM model önerisinin adımlarının şematize edilmiş hali verilmektedir. Bu şekil; Chapman ve arkadaşları (2000) tarafından ortaya konulmuş ve Balaban ve Kartal (2015) tarafından dilimize uyarlanmış CRISP modeli üzerinden geliştirilerek tasarlanmıştır.



Şekil 2.7: CRISP-EDM Modelinin Adımları.

Günümüz dünyasının veri üretme hızı düşünüldüğünde, farklı alanlarda çözülmek üzere bekleyen veri madenciliği problemlerinin olduğunu söylemek mümkündür. Bu tez çalışması kapsamında ele alınan veri madenciliği problemi eğitim sürecinin içinden elde edilmiştir. Problemin daha net kavranabilmesi için eğitim kavramı ve ülkemizdeki eğitimin en genel biçimde sunumu Bölüm 2.4.'de verilmektedir.

2.4. EĞİTİM KAVRAMI VE ÖNEMİ

Eğitim; bireyin davranışlarında kasıtlı olarak ve kendi yaşantıları yoluyla istendik davranış değişikliği meydana getirme sürecidir (Ertürk, 1973)Ertürk tarafından yapılan tanım incelendiğinde eğitim süreci ile ilgili aşağıdaki yorumlar yapılmaktadır (Ayas, 2013):

- Eğitim bir süreçtir: Bir başlangıç aşamasından itibaren zamanla değişim gösterir, durağan değil dinamik bir süreçtir

- Eğitim planlı ise kasıtlıdır: Belirli bir amaç uğruna yapılır. Diğer bir deyişe bir şeyi gerçekleştirme hedefi ile yürütülen bir çabadır.
- Eğitim sonunda bireyde davranış değişikliği meydana gelir: Eğitim alan bireyin eğitim alamdan önceki bilgi, beceri ve tutumları eğitim aldıktan sonra farklılaşır.
- Davranış değişikliği bireyin yaşantısı yoluyla olur: Birey eğitimin konusunu bizzat yaşamalıdır. Yani birey; yapma, yapılanı izleme-seyretme, okuma veya görenin (yaşayanın) anlattıklarını dinleme eylemlerinden en az birini gerçekleştirmelidir.

Ertürk tarafından yapılan bu tanım, eğitim felsefelerindeki değişim ve teknolojinin eğitim-öğretim faaliyetlerine entegrasyonu ile daha bireyselleştirilmiş bir tanıma doğru kaymıştır. Eğitimin yeni tanımı ve çehresinde; öğrencinin; vatandaşlık algısına uygun, nitelikli iş gücü oluşturabilecek, kültürel okur-yazarlık sahibi, kritik düşünme becerisine sahip, küresel anlamda katma değer oluşturabilecek şekilde yetiştirilmesi yer almaktadır (Jones, 2012). Alkan (2011)'e göre eğitim; kişinin bireysel, çevresel ve sosyal yönden başarılı olması ve barış, özgürlük, sosyal adalet ve evrensel bütünlük ideallerine ulaşmasında temel bir araç, toplumsal anlamda da ekonomik kalkınmanın itici gücüdür. Çınar (2014) göre ise eğitim; çağımızdaki değişim hızı düşünüldüğünde belli kalıplara oturtulamayacak kadar dinamik bir yapı kazanmıştır. Bu nedenle eğitim tanımlanmak yerine betimlenmeli ve betimi; "*bireyin doğuştan getirdiği bilişsel, duyuşsal, devinişsel ve toplumsal gizilgüçlerinin geliştirilmesi, yeni beceriler kazandırılması, düşünsel kalıpları aşması ve kendini gerçekleştirebilmesi için, bireye göre yapılan, planlı çalışmaları kapsayan ve içeriği toplumun maddi, manevi ve teknik temeline göre belirlenen etkinliklerin tümü*" şeklinde olmalıdır (Çınar, 2014).

Sanayi toplumuna uygun olarak kurgulanan eğitim-öğretim ortamları, bireye "istenilen davranış ve kalıpları yine istenilen şekilde enjekte etme" felsefesini barındırmaktadır. Bu nedenle üretkenlik, mevcut devinimi bozmayacak öğrencilerin sistematik olarak topluma kazandırılması algısına dayalıdır. Oysa bilgi toplumunda daha bireyselleştirilmiş, yaratıcılığın ve girişimciliğin desteklendiği, fiziksel eğitim ortamlarının çok üzerinde bir yapı tasvirlenmektedir. Bu yapıda öğrencinin iç dünyasında kendi bilgi örüntülerini oluşturması beklenmektedir. Bu örüntülerin

oluşması sürecinde, öğrenciye ve eğitim sürecine dair her verinin derinlemesine incelenmesi, analiz edilmesi daha başarılı sonuçlar elde edilmesinde katkı sahibidir.

Makro düzeyde eğitim sürecinden elde edilen veriler; ülkenin üniversite sınavına giren öğrenci sayısı, okullaşma oranı, belirli yaş gruplarındaki öğrenci sayısı, belirli yaş gruplarındaki nüfus değeri gibi çok farklı alanlara dayanmaktadır. Daha küçük boyutta, mikro düzeyde ele alındığında ise; bu tür veriler aşağıda sayılmakta olan kaynaklar ve/veya eylemler aracılığıyla elde edilebilmektedir (Mostow ve diğ., 2005):

- Yüz yüze yapılan eğitimlerde, öğrenci başarısına ilişkin yapılan sınavlar, teslim alınan ödevler/projeler, eğitmenin sürece ilişkin gözlemleri,
- CMS, LMS ve LCMS sistemleri ile yapılan eğitimlerde ise öğrencinin sistemde kalış süresi, ders materyalini okuma süresi, yapılan testlerdeki cevaplama, durma, araştırma, görevleri gerçekleştirme, diğer öğrencilerle iletişim kurma eylemleri.

2.4.1. Türkiye'de Eğitimin Genel Durumu

Ülkemizde öğrenci nüfusuna yönelik çeşitli veri türleri kayıt altına alınmaktadır. Bunlardan en çok bilinenleri;

- E-okul sistemine kayıt edilen; klasik sınıf ortamlarına ilişkin olarak öğrenci devam/devamsızlık, sınav ve proje sonuç bilgileri
- YÖK ve MEB veritabanlarındaki öğrenci sayısı, mezun sayısı, lise türlerine göre öğrenci sayısı, liselere ve üniversitelere giriş başarı düzeyleri, net sayıları ve Türkiye ortalaması,
- TÜİK tarafından derlenen genç nüfus eğitim değerlendirmeleri
- Yurtdışı kaynaklı yapılan eğitim sistemi durum değerlendirme araştırmaları şeklindedir.

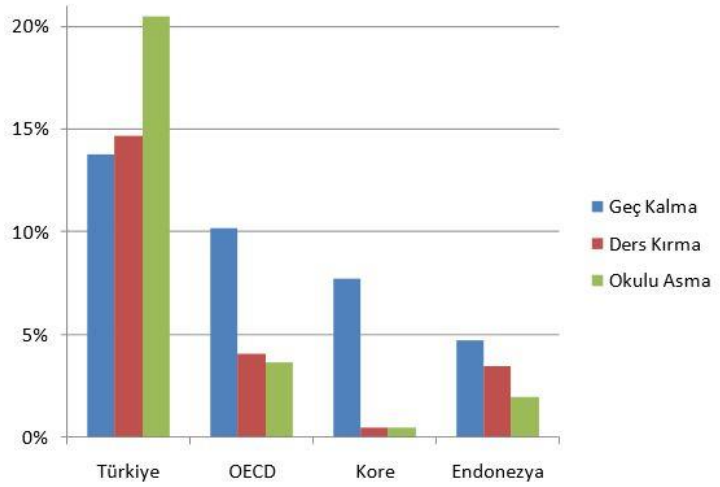
Okullaşma oranından, genç nüfus ve öğrenci nüfusu karşılaştırmalarına, LGS, LYS netlerinden, eğitimde teknolojinin kullanımını ifade eden istatistiki değerlere kadar pek çok veri, mevcut eğitim sisteminin durumunun anlaşılmasında yardımcı niteliktedir. Tablo 2.2'de verilen okullaşma yüzdesi sadece belirli bir yıla ait durumu genel olarak göstermektedir.

Tablo 2.2: Türkiye Geneli Net Okullaşma Oranları (MEB, 2014).

Grup	Yüzde
Ortaöğretim (Mesleki)	%29,30
Ortaöğretim (Genel)	%35,66
Ortaöğretim (14-17 yaş)	%64,95
İlköğretim (6-13 yaş)	%98,17
Okul Öncesi (3-5 yaş)	%26,92

Tablo 2.2. incelendiğinde ilköğretimden daha üst kademelere gidildikçe okullaşma yüzdesinin düştüğü görülmektedir.

Türkiye'deki eğitim sisteminin kalitesi diğer OECD ülkelerinin çoğunda görülenden daha düşüktür ve bu kalite sorunu farklı türlerdeki okullarda kendini daha netlikle ortaya koymaktadır (World Bank, 2011). Bu sorun mevcut eğitim sistemi içerisinde ders sürecinin aksamasına neden olabilecek faaliyetlerin gözlenebilirliği ile de anlaşılmaktadır. Yıldırım, ve diğ. (2013) tarafından hazırlanan raporda ders düzenini bozabilecek davranışlar ve okulu gelmeme bakımından OECD ile kıyaslamalar verilmiştir. Şekil 2.8'de Ülkemiz derse geç kalma, ders kırma veya okulu asma oranları sunulmaktadır.



Şekil 2.8: Ülkemiz Derse Geç Kalma, Ders Kırma Veya Okulu Asma Oranları.

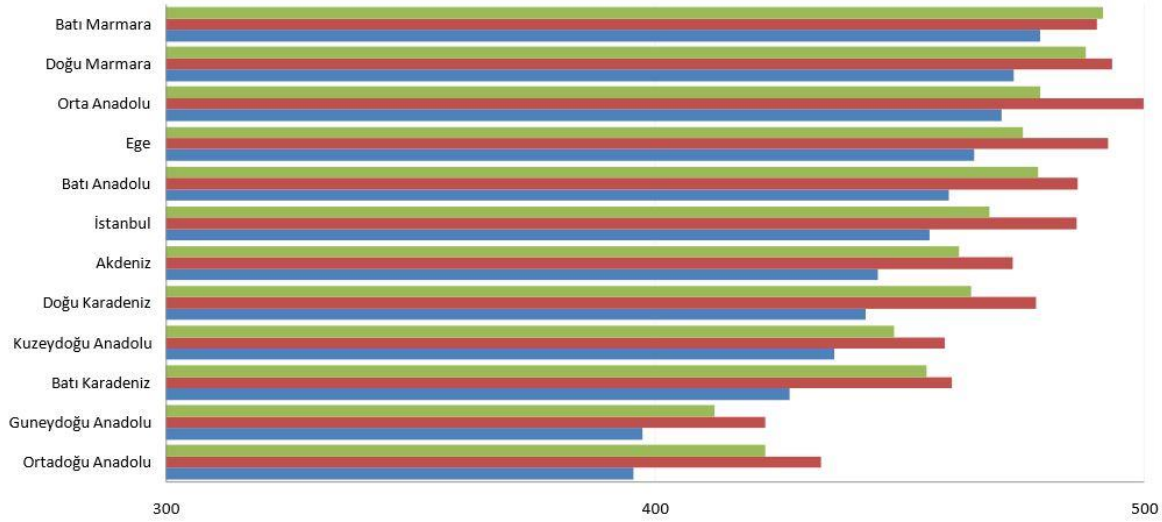
Şekil 2.8 incelendiğinde belirtilen kriterlere ilişkin oranlar bakımından OECD ortalamasının oldukça üzerinde bir durum ile karşı karşıya kalındığı söylenebilmektedir.

2.4.1.1. PISA

PISA (Programme for International Student Assessment)- Uluslararası Öğrenci Değerlendirme Programı; Ekonomik İşbirliği ve Kalkınma Teşkilatı (OECD) tarafından 2000 yılından itibaren başlatılan en kapsamlı eğitim araştırmasıdır (Yıldırım ve diğ., 2013). Araştırma kapsamında matematik, fen ve okuma becerileri alanları üç yılda bir döngüsel şekilde incelenmektedir. Bu inceleme her üç alanda da okuryazar olabilme kavramı çerçevesinde değerlendirilmektedir.

Türkiye bu sınava 2003 yılından beri katılmaktadır. 34'ü OECD ülkesi olmak üzere yaklaşık 70 ülke çalışmaya katılmaktadır. Çalışma kapsamında 15 yaş grubu öğrencilerin örgün eğitimde matematik, fen ve okuma becerileri alanında kazanmış oldukları bilgileri günlük yaşamdan ne ölçüde kullandıkları ölçülmekte, ayrıca eğitim hakkındaki kişisel görüşleri, kendileri ve aileleri hakkında bilgiler de değerlendirilmektedir.

Ülkemiz 2003-2012 yıllarında düzenlenen uygulamaların hepsinde belli bir gelişim düzeyi gösterse bile OECD ortalamasının altında kalmaktadır (Yıldırım ve diğ., 2013). 65 ülkenin katıldığı PISA 2012 yılı değerlendirme sonuçlarına göre Türkiye; matematik sıralamasında 44. okuma sıralamasında 42. ve Fen sıralamasında 43. ülke olarak yer almıştır. Şekil 2.9'da bölgeler bazında alanlardaki ortalamalar verilmektedir.



	Ortadoğu Anadolu	Guneydoğu Anadolu	Batı Karadeniz	Kuzeydoğu Anadolu	Doğu Karadeniz	Akdeniz	İstanbul	Batı Anadolu	Ege	Orta Anadolu	Doğu Marmara	Batı Marmara
Fen	423	412	456	449	465	462	468	478	475	479	488	492
Okuma	434	423	461	459	478	473	486	486	493	503	494	490
Matematik	395	397	428	437	443	446	456	460	465	471	473	479

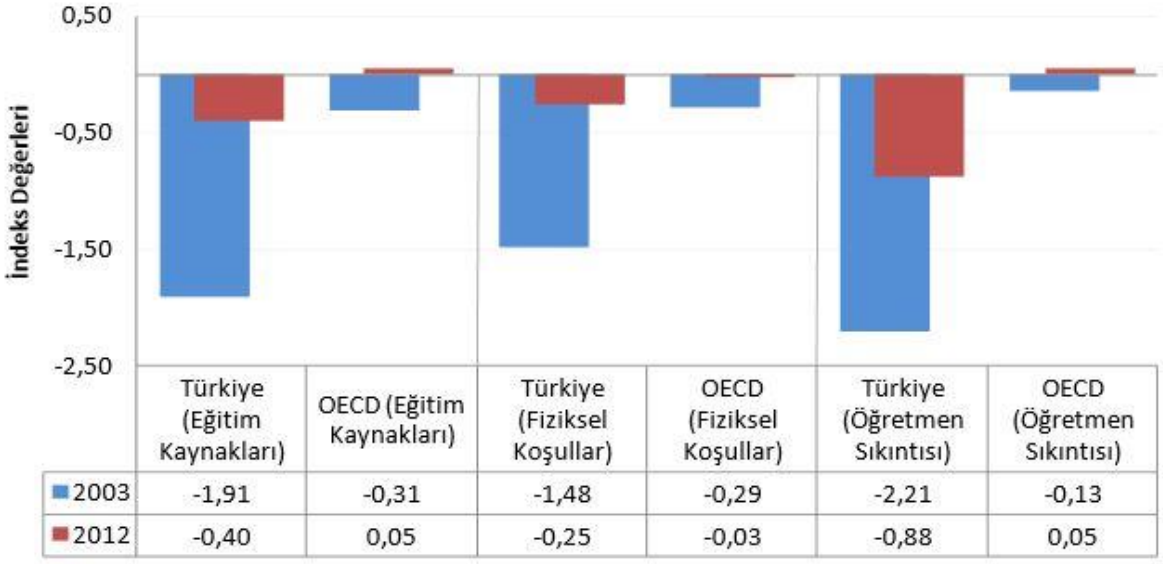
Şekil 2.9: Bölgeler Bazında Alanlardaki Ortalamalar.

Şekil 2.9 incelendiğinde genel olarak doğu bölgelerinin ortalamasının daha düşük olduğu görülmektedir.

Türkiye'de öğrencilerin %42'si matematik, %61,7si fen, %52,5'i okuma alanında 2. yeterlilik (asgari performans düzeyi) düzeyine ulaşamamıştır. PISA'da her alan için 6 yeterlik düzeyi tanımlanmış olup, bilgi çağında öğrencilerin büyük bir kısmının (tercihen hepsinin) en azından 2. yeterlik düzeyine ulaşmış olması gerektiği kabul edilmektedir.

2003-2012 yılları arasındaki PISA matematik alanı sonuçlarına göre puan farklılıkları okullar arasındaki farklılıktan kaynaklanmaktadır (Yıldırım ve diğ., 2013). Bu durum özellikle matematik eğitimi alanında ülke genelinde öğrenciler arasında okul eğitimi açısından ciddi bir imkan farklılığı olduğunu göstermektedir.

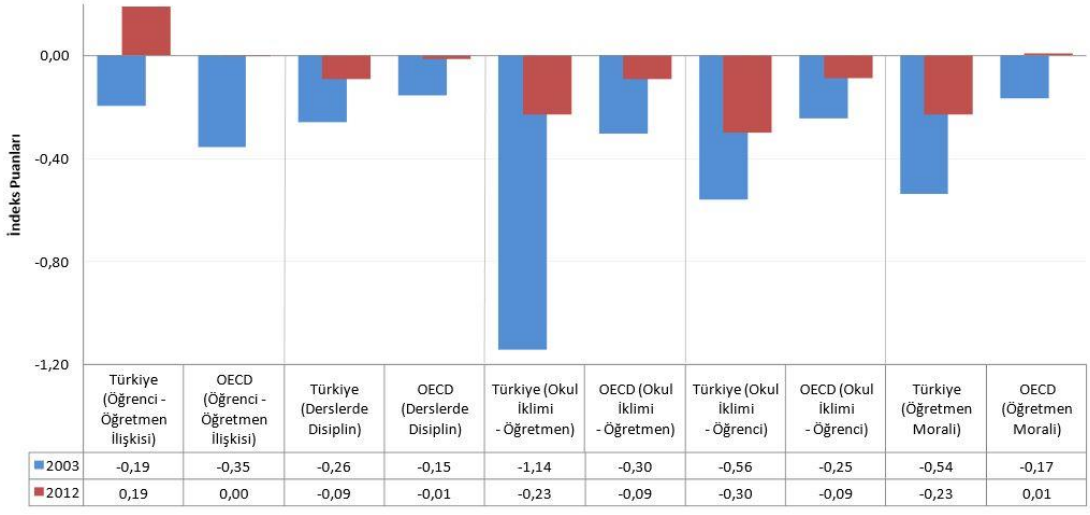
Türkiye'nin okul ortamındaki değişim durumu (eğitim kaynakları, fiziksel koşullar ve öğretmen ihtiyacı) okul yöneticilerinin alandaki sorulara verdikleri cevaplar üzerinden hesaplanmaktadır. Şekil 2.10'da bu durumun alanlar bazındaki indeks değerleri verilmektedir.



	Türkiye (Eğitim Kaynakları)	OECD (Eğitim Kaynakları)	Türkiye (Fiziksel Koşullar)	OECD (Fiziksel Koşullar)	Türkiye (Öğretmen Sıkıntısı)	OECD (Öğretmen Sıkıntısı)
2003	-1,91	-0,31	-1,48	-0,29	-2,21	-0,13
2012	-0,40	0,05	-0,25	-0,03	-0,88	0,05

Şekil 2.10: Türkiye'nin Okul Ortamındaki Değişim Durumu.

Şekil 2.10'da verilen indeks değerleri incelendiğinde, ülkemizin 2003-2012 yılları arasında genel bir gelişim gösterdiği ancak beklenen düzeyde olmadığı söylenebilmektedir. Benzer şekilde okul ikilimi açısından 2003-2012 yılları arasında OECD ve Türkiye karşılaştırılması Şekil 2.11'de verilmektedir.



	Türkiye (Öğrenci - Öğretmen İlişkisi)	OECD (Öğrenci - Öğretmen İlişkisi)	Türkiye (Derslerde Disiplin)	OECD (Derslerde Disiplin)	Türkiye (Okul İklimi - Öğretmen)	OECD (Okul İklimi - Öğretmen)	Türkiye (Okul İklimi - Öğrenci)	OECD (Okul İklimi - Öğrenci)	Türkiye (Öğretmen Morali)	OECD (Öğretmen Morali)
2003	-0,19	-0,35	-0,26	-0,15	-1,14	-0,30	-0,56	-0,25	-0,54	-0,17
2012	0,19	0,00	-0,09	-0,01	-0,23	-0,09	-0,30	-0,09	-0,23	0,01

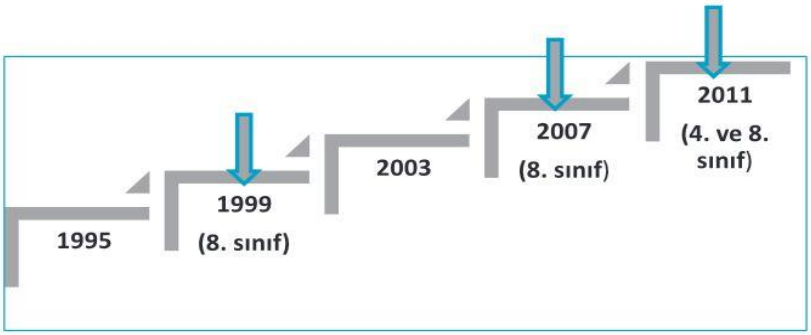
Şekil 2.11: Okul İkilimi Açısından 2003-2012 Yılları Arasında OECD Ve Türkiye Karşılaştırılması.

Şekil 2.11 incelendiğinde derslerde disiplin, öğrenci, öğretmen ve öğretmen morali açısından geçen zaman içerisinde gelişme gösterilmiş olmasına rağmen ortalamanın altında kalındığı görülmektedir.

2.4.1.2. TIMSS

Uluslararası Matematik ve Fen Eğilimleri Araştırması olan TIMSS, Uluslararası Eğitim Başarılarını Değerlendirme Kuruluşu- IEA'nın (International Association for the Evaluation of Educational Achievement) dört yıllık aralıklarla düzenlediği bir tarama araştırmasıdır (Büyüköztürk ve diğ., 2014). Bu tarama araştırması öğrencilerin matematik ve fen alanlarındaki başarılarını ölçme ve değerlendirmenin yanı sıra, bu alanlardaki öğrenim ve öğretimin okullarda nasıl gerçekleştiğini belirlemek, ulusal sistemler arasındaki farklılıkları dünya çapında ölçmek için tasarlanmıştır.

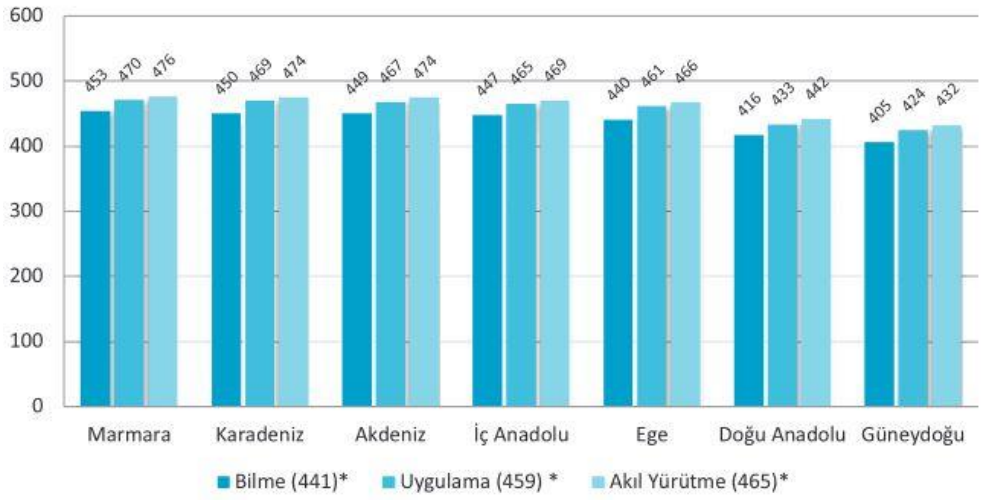
Dünyada 1995 yılından beri 60'dan fazla ülkenin katılımıyla gerçekleştirilen bu tarama araştırmasına Türkiye ilk defa 1999 yılında 8. sınıflar düzeyinde katılmıştır. Şekil 2.12'de ülkemizin yıllara göre TIMSS değerlendirmesine katılım durumu verilmektedir.



Şekil 2.12: Yıllara Göre Türkiye TIMSS Katılım Durumu.

TIMSS 2011 çalışmasına 4. sınıf düzeyinde 50 ülke katılmış, Türkiye 35. sırada yer almıştır. Öğrencilerin %50'si ortalama 469 başarı puanı ile TIMSS dört yeterlilik düzeyi arasından, alt düzey ve alt düzey altına yerleşmiştir (Büyüköztürk ve diğ., 2014). TIMSS 2011 çalışmasına 8. sınıf düzeyinde ise 42 ülke katılmış, Türkiye 24. sırada yer almıştır. Öğrencilerin %60'ı ortalama 452 başarı puanı ile TIMSS dört yeterlilik düzeyi arasından, alt düzey ve alt düzey altına yerleşmiştir. 1999 yılında öğrencilerin %35'i, 2007 yılında %41'i ve 2011 yılında %33'ü TIMSS değerlendirmesine dahil olamayan

grup olarak alt düzey altına yerleşmiştir (Büyüköztürk ve diğ., 2014) . Şekil 2.13'te TIMSS 2011 8. sınıfların bölgeler bazında Türkiye ortalamasına göre durumları bilme, uygulama ve akıl yürütme bilişsel düzeyleri verilmektedir.



Şekil 2.13: TIMSS 2011 8. Sınıfların Bölgeler Bazında Türkiye Ortalamasına Göre Durumları Bilme, Uygulama Ve Akıl Yürütme Bilişsel Düzeyleri.

TIMSS taramasında öğretmenlerin çalışma koşulları ile ilgili yaşadıkları problem türleri ile öğrenci başarısı arasındaki ilişkiye de bakılmaktadır. Türkiye'de TIMSS 2011 çalışmasına 4. sınıf düzeyinde katılan öğrencilerin %26'sının öğretmenlerinin problem yaşamadığı, %47'sinin küçük problemi olduğu ve %27'sinin çalışma koşullarına yönelik orta düzeyde problem yaşadığı belirlenmiştir. Bu durumun ortalama öğrenci başarısına yansıması 53 puanlık (hiç problem yaşamayan öğrencilerin başarı puanları ile orta düzey problem yaşayan öğrencilerin ortalama başarı puanları arasındaki fark) bir fark şeklinde gözlenmiştir. Okul iklimi açısından yapılan değerlendirmelerde de olumlu okul iklimine sahip okullarda başarı puanlarının daha yüksek olduğu söylenebilmektedir (Büyüköztürk ve diğ., 2014)

TIMSS 2011 8. sınıflar incelemesinde öğrencilerin geleceğe ilişkin eğitim beklentilerine bakıldığında %28'inin lisansüstü eğitim yapmak istediği (matematik başarı ortalamaları: 532), %44'ünün üniversite mezunu, %5'inin meslek yüksekokulu mezunu ve %16'sının

lise mezunu olmayı hedefledikleri %7'sinin ise eğitim durumları hakkında kararsız olduğu tespit edilmiştir. (Büyüköztürk ve diğ., 2014).

Eğitimin başarısı; ulusal ve uluslararası boyutta yapılan ölçümler, mikro düzeyde semtler ve ilçelerdeki okul içi/ilçe içi değerlendirmeler; makro düzeyde iller ve bölgelere dair düzenlenen genel sınavlar ile akademik başarı biçiminde ortaya konabilmektedir. TIMSS ve PISA genel değerlendirmeleri ve literatürde ülkemize özgü yapılan çalışmalar; akademik başarının şimdiye kadar istatistiki yöntemlerle sıklıkla değerlendirildiğini göstermektedir. Bu değerlendirmenin veri madenciliği yöntem ve teknikleri ile gerçekleştirilmesi, eğitim gibi değerli bir veri kaynağı barındıran bir alan açısından yeni bir çözüm/inceleme sürecinin başlatılabilmesine olanak tanıyacaktır. Bölüm 2.5.'te bu tez çalışması kapsamında araştırılan veri madenciliği probleminin literatür alt yapısı verilecektir.

2.5. AKADEMİK BAŞARIYI ETKİLEYEN FAKTÖRLER

Akademik başarı; öğrencilerin okul yaşamında amaçlanmış olan davranışlara ulaşma düzeyidir (Silah, 2003). Yıldırım (2006)'a göre Türk Eğitim sistemi başarı odaklı olmakla birlikte eğitim sürecine dahil olan bireyler açısından en önemli hedef üniversiteye girebilmek haline dönüşmüştür.

En genel biçimde eğitim sisteminin görevi; öğrencinin akademik başarılarını ölçerek, elde edilen performans sonuçlarına göre başarılı olacakları alanlara yönelmelerini sağlamaktır (Silah, 2003). Ancak bu yönelmenin sağlanabilmesi için bireylerin başarılı olma süreçlerinin çeşitli açılardan ele alınması gerekmektedir. Öğrencinin akademik başarısını etkilen fiziksel, psikolojik ve toplumsal farklı nitelikte ve çeşitlilikte pek çok faktör bulunmaktadır (Türnüklü ve diğ., 2001; Silah, 2003). Bu faktörlerin oluşumunu sağlayan kaynaklar aşağıda verilmektedir:

- Öğrenenin bireysel nitelikleri (Harris, 1940; Özgüven, 1974; Ponzetti & Gate, 1981; Can, 1992; De Raad ve Schouwenburg, 1996; Busato ve diğ., 2000; Diamond ve diğ., 2004; Keser, 2007).
- Öğrencinin ailesi (Özgüven, 1974; Maton ve Hrabowski III, 1998; Gonzales-Pienda, ve diğ., Zellman, 1998; 2002; Gonzales-Pienda, ve diğerleri, 2002; Bean ve diğ., 2003; Yıldırım, 2006; Epstein, 2008).

- Öğretmenin nitelikleri (Ekici, 2002; Gordon ve diğ., 2006; Şama ve Tarım, 2007).
- Okul yönetimi (Özgüven, 1974; Günçer & Köse, 1993; Tural, 2002).
- Uygulanan eğitim sistemi (Stein & Wang, 1988; Guskey, 2002)

Değişen teknoloji ve bu değişimden olumlu etkilenmesi beklenen eğitim-öğretim sürecinde öğrencilerin akademik performanslarında düşüş ve okula devamlarındaki problemler devam etmektedir (Tinto, 1994; Lloyd ve diğ., 2001). Bu durum birey özelinde ve toplum genelinde mevcut potansiyelin tam olarak değerlendirilemediği, yapılan yatırımların beklenen karşılığı getirmediğini göstermektedir.

Öğrencinin okul başarısını etkilen faktörler arasında doğrudan etkili olmamasına karşın okul devamsızlığı da yer almaktadır (Altınkurt, 2008). Okula/derslere devam edememe davranışı, farklı sorunlardan kaynaklı istenmeyen öğrenci davranışıdır (Başar, 2001; Külahoğlu, 2001). Okula devam sorunu dışında karşılaşılan bir diğer sorun öğrencinin eğitim-öğretim sürecinde yaşadığı tükenmişliktir.

Tükenmişlik olgusu; bireyin yaptığı meslekle ilişkilendirilmesine rağmen, öğrenciler arasında da sıklıkla gözlenebilmektedir (Yang, 2004; Schaufeli ve Salanova, 2007). Literatürde farklı yıllara ait çalışmaların varlığı değişen toplum ve birey profiline rağmen tükenmişliğin yaygın bir problem olduğunu göstermektedir (Fimian, 1988; Balogun ve diğ., 1999; Jacobs ve Dodd, 2003; Santen ve diğ., 2010) Tükenmişlik öğrencinin okul çalışmalarına ilgisizleşmesi, kendisinden beklenen çalışmalara yönelik yaşadığı bezginlik ve akademik anlamda yetersizlik hissi yaşaması olarak tanımlanmaktadır (Schaufeli ve diğ., 2002). Bu hissi yaşayan öğrenci; devamsızlık gösterebilmekte, eğitim sürecine ilişkin motivasyon düşüklüğü yaşamakta ve okulu bırakma eğilimi gösterebilmektedir. Akademik başarı ve tükenmişlik düzeyi arasındaki ilişkiyi inceleyen çalışmalar; öğrencinin yaşamakta olduğu tükenmişliğin, akademik başarısını belirgin şekilde olumsuz etkilediğini ortaya koymuştur (McCarthy ve diğ., 1990; Jacobs ve Dodd, 2003; Çapulcuoğlu ve Gündüz, 2013). Balkıs ve diğ. (2011) yaptıkları çalışmada tükenmişliğin akademik başarı ile anlamlı düzeyde negatif ilişkili olduğunu ortaya koymuşlardır.

2.6. EĞİTİMDE BİLGİNİN KEŞFİ VE VERİ MADENCİLİĞİ

Çeşitli eğitim ortamlarından elde edilen veriler; öğrenen sayısının az olduğu durumlarda doğru organize edilerek, eğitim sürecinin analizine olanak tanımaktadır. Bu sayede sürece ilişkin tahminlerde bulunmak, olası durumlara karşı önlemler alabilmek mümkündür. Ancak eğitimin sadece yüz yüze biçimde ve sınırlı sayıda öğrenci ile yapılmadığı durumlarda, veri kaynaklarının kontrolü, verilerin derlenmesi ve elde edilen verilerin yorumlanması oldukça zorlaşmakta, hatta etkili ve verimli bir şekilde yapılamamaktadır. Gerçekten de Cromey ve Hanson (2000); çeşitli değerlendirme sistemlerinden elde edilen verinin yönetilmesi, sentezlenmesi, yorumlanması ve öğrenci için bir değerlendirme oluşturulmasının eğitimciler için oldukça zor olduğunu belirtmişlerdir.

Yüz yüze yapılan eğitim dışında önemli bir eğitim alternatifi olan e-öğrenme, CMS, LMS ve LCMS ile öğrenenin davranışlarını incelemek amacıyla kullanılabilecek büyük miktarlardaki veriyi barındırmaktadırlar (Mostow ve Beck, 2006). Sistemdeki her hareketin veritabanlarına kaydedilmesi, öğrenciler hakkında çeşitli çıkarımlarda bulunmaya yarayacak büyük miktarlarda veri yığınları oluşmasına neden olmaktadır. Özellikle öğrenci sayısının çok olduğu durumlarda, bu yığınlara bakarak karar vermek ya da öğrenciler hakkında çeşitli çıkarımlarda bulunmak oldukça zorlaşmaktadır (Dringus ve Ellis, 2005). Bazı e-öğrenme ortamlarda bu zorluğun aşılması adına bir takım yardımcı araçlar bulunmaktadır. Ancak bu araçlara rağmen, kursun yapısı ve içeriği ile öğrenme sürecindeki etkinliğinin tüm öğrenciler açısından değerlendirilebilmesi pek de mümkün değildir (Zorrilla ve diğ., 2005).

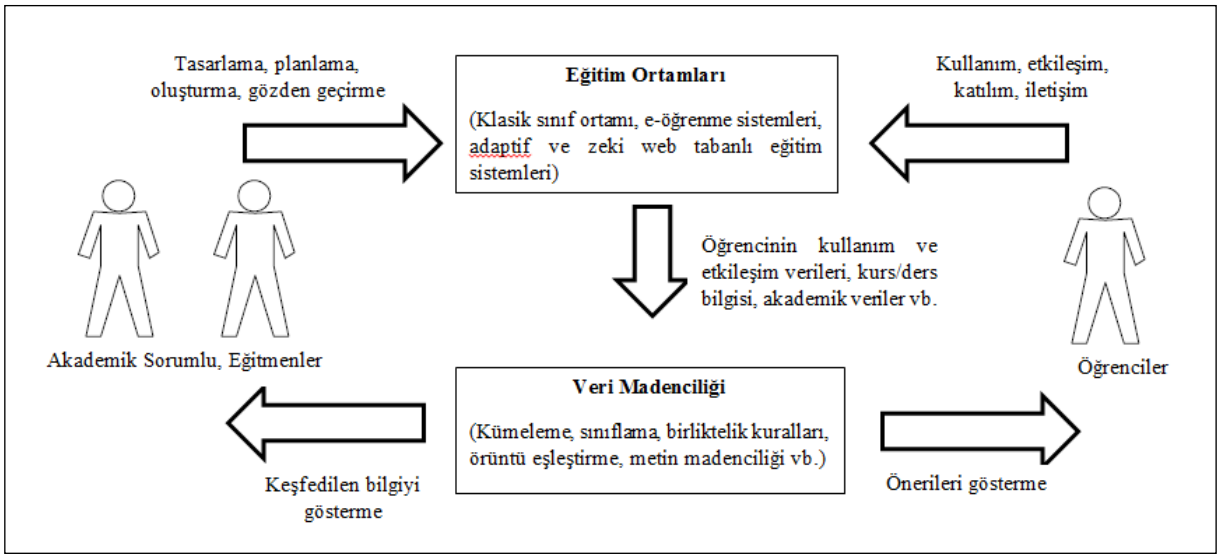
Hangi kanaldan eğitim-öğretim metodundan elde edilirse edilsin, süreçteki öğrenci sayısı ve gerçekleştirilen planlı/plansız faaliyetler düşünüldüğünde; büyük miktarlarda değerli verinin üretildiği açıktır. Bu nedenle eğitim öğretim süreci; veri madenciliği açısından oldukça önemli bir çalışma/araştırma alanını barındırmaktadırlar.

Eğitimde Veri Madenciliği (EVM); büyük bir hızla çoğalan depolanmış verinin analiz edilerek, makro yapıdan mikro yapıya kadar faydalı hale getirilmesine odaklanmaktadır (Kurniawan ve Halim, 2013).

Veri madenciliği teknikleri; işlenmemiş veriyi; karar verirken kullanılabilecek türden bilgiye dönüştüren bilgi keşif araçları olduğundan daha önce bahsedilmişti. Bu araçların; özellikle iş dünyasında ileriye dönük eğilimlerin, tüketici davranışlarının tahmininde sıklıkla kullanıldığı literatürdeki pek çok çalışmada görebilmek mümkündür (Harding, Shahbaz, & Srinivas, 2006; Ngai & Xiu, 2009; Yen, Gu, & Lee, 2013; Maab, Spruit, & Waal, 2014). Veri madenciliği yöntemlerinin sektörel bazda yoğun kullanımına karşın, eğitim alanında genellikle çok sınırlı çalışmalar gerçekleştirilmiştir (Faulkner ve diğ., 2010). Oysa eğitim alanında derlenen verilerin kullanılması, bu alanda başarı elde edilebilmesi (Streifer, 2002) ve öğrenci başarısının arttırılmasında (Popham, 2003) merkezi bir öneme sahiptir. Siemens ve Long (2011); özellikle eğitim alanındaki verilerin değerlendirilmesi ile aşağıda belirtilen hususlarda önemli iyileştirme ve önleme çalışmalarının yapılabileceğini öne sürmüşlerdir:

- Yönetsel anlamda karar-vermenin iyileştirilmesi,
- Eğitim kurumların gerçek anlamda başarıları ve karşılaştıkları zorlukların anlaşılabilmesi,
- Çeşitli öğrenme zorlukları yaşayan bireylerin tespit edilmesi, nasıl desteklenecekleriinin belirlenmesi ve bu sayede başarıda artışın sağlanması,
- LMS, CMS, LCMS ortamlarındaki çeşitli verilerin analiz edilmesi ile dersi/kursu bırakma riski olan bireylerin tespiti,
- Bireylerin öğrenme alışkanlıklarının tespit edilmesi ve bu alışkanlıkların iyileştirilmesi.

Eğitimde veri madenciliği (educational data mining); veri madenciliği metotlarının, çeşitli eğitim ortamlarından/portallarından gelen verilerden bilginin çıkarılmasında kullanılan bir disiplindir (Baker ve Yacef, 2009). Romero ve Ventura (2007); EVM'yi; veri madenciliğinin eğitim ortamlarında uygulandığı bir döngü olarak ifade etmişlerdir (Şekil 2.14).



Şekil 2.14: Bir Döngü Olarak EVM (Romero ve Ventura, 2007).

EVM bir model olarak düşünüldüğünde aşağıda verilen adımları içermekte olduğu söylenebilir:

- Bir veritabanına kayıtlı verilerin "veri çıkarımı (data extraction)",
- Verinin temizlenmesi ve davranışların tanımlanması (veri ön işleme)
- Veri türlerine (ID, öğrenme davranışları, süre-harcanan zaman, bağlı değişkeler vs.) ve düzeylerine (kurs, süre, tekil kayıtlar vs.) ayrılması (veri ön işleme)
- Veri madenciliği tekniklerinin uygulanması (veri görselleştirme, kümeleme, ilişkilerin araştırılması, tahmin)
- Bilginin eldesi
- Elde edilen bilginin gerekiyorsa tekrar veri madenciliği yöntemlerine tabi tutulması adımlarını içermektedir (Hung ve diğ., 2012).

EVM sadece eğitim ortamlarından alınan verinin analizi ile bir takım anlamlı sonuçlar çıkarılmasını içermemektedir. Bu disiplin ile eğitmenlere; öğrenme verilerinin etkisi ve nasıl kullanılabileceği gösterilebileceği gibi, makro boyutta eğitim sürecine yardımcı karar destek ve tavsiye sistemlerinin geliştirilmesi (Huebner, 2013); eğitim politikalarında gözden geçirilmesi ve iyileştirilmesi (Peña-Ayala ve Cárdenas, 2014) sağlanabilir.

2.6.1. Dünyada Yapılan Çalışmalar

Gray ve diğerleri (2014) tarafından yapılan çalışmada sınıflama modelleri kullanılarak, öğrencilerin eğitimlerinin ilk yıllarında başarısız olma risklerinin tanımlanması üzerinde durulmuştur. 2010-2012 yılları arasında eğitim gören ve aynı özelliklere sahip 1074 öğrenciden derlenen verilerle yapılan çalışmada, psikometrik indikatörler de göz önünde bulundurulmuştur. Davranış/tutum, kişilik, motivasyon ve öğrenme stratejileri akademik performans ile doğrudan ve dolaylı olarak ilişkili kabul edilerek incelemeler yapılmıştır. Çalışma kapsamında birinci sınıf öğrencilerine online bir anket uygulanmıştır. Elde edilen veriler RapidMiner V5.3 ile altı farklı sınıflama algoritması kullanılarak analiz edilmiştir. Bu algoritmalardan üçü lineer sınıflandırıcılar olan Naive Bayes (NB), Karar Ağaçları (unpruned Decision Tree) ve Logistic Regresyon; diğer üçü ise lineer olmayan sınıflandırıcılar olan Anova kernal fonksiyonu (Anova kernal function) kullanan destek vektör makinası, Yapay Sinir Ağları ve k-yakınsak Komşu (k-Nearest Neighbour) şeklinde seçilmiştir. Modeller eldeki tüm veri setleri ile eğitilmiş ve yaş baz alınarak iki gruba ayrılmıştır. Farklı yaş guruplarında yapılan entropi hesabı sonucu, 21 yaş optimal ayrılma noktası olarak belirlenmiştir. Araştırma sonucunda genç grubun (21 yaş altı) modellenmesinde, çalışma kapsamında kullanılan tüm algoritmaların daha iyi tahminleme eğilimi gösterdiği fark edilmiştir. 21 yaş üstü grupta k-NN hariç kullanılan Naive Bayes ve Lojistik regresyon analizi tahminleme başarısı açısından zayıf kalmıştır. Sonuç olarak çalışmada; 21 yaş üstü öğrencilerin modellenmesinin 21 yaş altına göre daha karmaşık bir yapıya sahip olduğu görüşüne ulaşılmıştır.

Márquez-Vera ve diğerleri (2013); yaptıkları çalışmada veri madenciliği teknikleri kullanılarak eğitim sürecinde başarısızlığa uğrayacak ve okulu bırakacak öğrencilere yönelik tahminleme çalışmaları yapmışlardır. Zacatecas Meksika'daki 670 orta eğitim öğrencisine ait veriler baz alınarak yapılan araştırmada beyaz kutu sınıflama (white-box classification) metodu, indüksiyon kuralları ve karar ağaçları kullanılmıştır. Veri toplama aşamasında; öğrenci performansını etkilediği düşünülen tüm faktörler tanımlanmış ve farklı kaynaklardan bu faktörleri elde etmeye yönelik veriler derlenmiştir. Oluşturulan veri setlerine standart ön işleme işlemleri uygulanmıştır. Veri madenciliği aşamasında, kurallara dayalı sınıflama algoritmaları ve karar verme ağaçları kullanılmıştır. Beyaz kutu teknikleri olarak adlandırılan bu yöntemlerle yorumlanabilir

modeller oluşturulması sağlamıştır. Ayrıca dengelenmemiş veri sorununu (imbalanced data problem) çözmek için maliyete duyarlı sınıflama yaklaşımı (cost sensitive classification approach) da kullanmışlardır. Kullanılan tüm bu yaklaşımlar en iyi sonucu sunan algoritmayı belirleyebilmek için karşılaştırılmıştır. Yorumlama aşamasında oluşturulan modeller öğrencinin başarısızlığını belirleyebilmek için analiz edilmiştir. Çalışmada Weka programı yardımıyla on adet sınıflama algoritması kullanılmıştır. Bunlarda beşi indüksiyon kural algoritmalar (rule induction algorithms) olan JRip (propositional rule learner), NNge (nearest neighborlike algorithm), OneR (minimum-error attribute for class prediction), Prism (algorithm for inducing modular rules), Ridor (an implementation of the Ripple-Down Rule learner) şeklindedir. Diğer beşi ise karar ağacı algoritmaları olan J48 (algorithm for generating apruned or unpruned), C4.5 karar ağacı, Simple Cart (minimal cost-complexity pruning), ADTree (an alternating decision tree), RandomTree (K randomly chosen attributes at each node of the tree), REPTree (a fast decision tree learner) şeklindedir. Bu çalışma ile öğrencinin başarısını tahminlemeye yönelik on sınıflama algoritması karşılaştırılarak; kullanılabilecek algoritmaları ve performansları, tahminleme işlemleri için akademik perforansı etkileyen faktörlerin neler olabileceği ortaya konulmuştur. Sonuç olarak Prism, OneR ve ADTree'nin en iyi sonuçları veren algoritmalar olduğu belirtilmiştir (Márquez-Vera ve diğ., 2013)

Gamulin ve diğerleri (2013) hibrit öğrenmede veri madenciliği tekniklerinin kullanımıyla final sınav notlarının tahmin edilebilmesi üzerine araştırma yapmışlardır. Araştırma kapsamında biyomedikal programının fizik dersine kayıtlı 302 öğrencinin sınıf içi ve online ortamlardaki hareket/aktivite, sınav ve değerlendirme sonuçları baz alınmıştır. Tahmini değişkenler arasında çoklu doğrusal bağlantı (multi-colineraity) olabilmesi ve boyut küçültme ihtiyacına dayanılarak tahmin etme işlemi için Temel Bileşenler Regresyon Analizi (Principal Component Regression) ve Kısmi En Küçük Kareler Regresyon Analizi (Partial Least Square regression) seçilmiştir. Araştırma kapsamında bu analizleri gerçekleştirmek için Mathlab (PLS Toolbox eklentisi) programı kullanılmıştır. Online ve klasik yapılan sınav sonuçları, Moodle öğrenme materyalleri ve/veya genel kurs bilgi sayfasına giriş sayısı gibi çeşitli bilgiler bağımlı ve bağımsız değişkenler olacak şekilde düzenlenmiştir. Kurulan modelin başarısını ölçmede iki tür parametre kullanılmıştır. Bunlar; kullanılan veri setinin modele

uygunluğu hakkında bilgi veren Ortalama Hata Karekök Yaklaşımı (root-mean square error-RMSE) ve ölçülen değerler arasındaki korelasyon katsayısı R2 olarak alınmıştır. Sonuç olarak hibrit öğrenme ortamları ve farklı türlerdeki veri kaynaklarına dayanarak final sınav sonuçlarının kesinliği düşükte olsa tahmin edilebileceği ifade edilmiştir. Benzer şekilde R2, RMSE ve bileşen sayısı gibi parametrelere bağlı olarak PLS modelinin PCR modele göre daha tutarlı sonuçlar ürettiği belirtilmiştir (Gamulin ve diğ., 2013).

Bydzovska ve Popelinsky (2013) tarafından yapılan çalışmada; öğrenci başarı durumuna göre öğrencilere bir kursu seçip seçmemeleri yönünde öneri sunabilecek bir "tavsiye sistemi"nin gerekliliği üzerinde durulmuştur. Çalışma kapsamında öğrenci notlarına göre her kurs için ortalama zorluk düzeyi belirlenmiştir. Benzer şekilde bir öğrencinin potansiyeli de onun notları ve aldığı kurların zorluk düzeyine göre oluşturulmuştur. Masaryk Üniversitesi bilgi sisteminden elde edilen verileri ile yapılan çalışmada sadece veri madenciliği yöntemleri değil aynı zamanda sosyal ağ analizi yöntemleri de kullanılmıştır. Özellikle öğrencinin mesajlaşmaları, yorumları, kişisel sayfalara yaptığı ziyaretler vb. online-sosyal hareketlerine dayanılarak bir sosyogram oluşturulmuştur. Weka programı üzerinde Naive Bayes, destek vektör makinesi, sınıflama kuralları (PART), OneR ve J48 algoritmaları çalıştırılmıştır. Beş kurs üzerinden yapılan çalışmada kursu alan kişi sayısına göre başarılı olabilecek yaklaşık öğrenci sayıları tahmin edilmiştir. Benzer şekilde kursun ortalama başarısı ile öğrenci potansiyelini kıyaslayan bir yapı oluşturulmuştur.

Kurniawan ve Halim (2013); yaptıkları çalışmada eğitim kurumlarında depolanan bilginin, eğitim süreci açısından ne kadar önemli olduğuna değinmişlerdir. Bu noktadan yola çıkarak veri ambarlarından bilginin doğru bir şekilde çekilebilmesi ve öğrencilerin akademik performanslarının tahmin edilmesine yönelik veri madenciliği tekniklerinin kullanılmasında uygulanabilecek bir model önerisinde bulunmuşlardır. İlk olarak "analiz yöntemi" altında genel bir literatür tarama, veri ambarlarının kullanımı ve veri madenciliği tekniklerinin karşılaştırılması üzerine çalışmaların incelenmesi, etkin bir veri tabanının nasıl olmasına yönelik bir tasarım yapabilmek için paydaşlarla yapılan görüşmeler ve anket uygulamaları yapılmıştır. İkinci aşamada ise veri ambarı tasarımın nasıl yapılması gerektiği üzerinde durulmuş ve Ralph Kimball tarafından yapılan

metodolojik yaklaşım kullanılmıştır. Çalışma sonunda oluşturulan model önerisi ile eğitim kurumlarında doğru tasarlanan veri ambarlarının, veri madenciliği teknikleri kullanılarak öğrenci performanslarının tahmininde daha başarılı olunacağı öne sürülmüştür (Kurniawan ve Halim, 2013).

Hung ve diğerleri (2012) yaptıkları çalışmada eğitimde veri madenciliğinin analitik sürecine yardımcı olması adına kişiselleştirilmiş bir model öneri sunmuşlardır. Bu model önerisi ile eğitmenlere EVM modelini online eğitim öğretim sürecinde bilginin üretilmesine nasıl destek olabileceğini göstermeye çalışmışlardır. Çalışma kapsamında sunulan model sadece LMS aracılığıyla derlenecek verilere uygulanacak şekilde tasarlanmıştır (Hung ve diğ., 2012).

Abdous ve diğerleri (2012); EVM'nin yükseköğretimde öğrencilerin öğrenme deneyimlerinin anlaşılması açısından önemli bir güç olduğunu belirtmişlerdir. Özellikle EVM'nin bir analiz ve karar verme aracı olarak, öğrenci bilgi sistemlerine (student information system- SIS) ve LMS entegrasyonunun pek çok yeni fırsatı sağlayacağına inandıklarını vurgulamışlardır. Yaptıkları çalışmada; öğrencilere ilişkin video kayıtlarını analiz edebilmek için EVM ve regresyon analizi kullanan bir hibrit yaklaşım geliştirmişlerdir. Bu yaklaşım ile öğrencilerin online öğrenme davranışları ve aldıkları derslerdeki performanslarını analiz etmeye çalışmışlardır. Araştırma; Orta-Atlanta bölgesinde uzaktan eğitimde lider kabul edilen, 17.000 lisans ve 6.000 yüksek lisans eğitimi veren bir devlet üniversitesinde gerçekleştirilmiştir. Araştırma kapsamında 2009 yılında 138 farklı derse kayıtlı toplam 1144 öğrencinin derse katılımları, sisteme girme sıklıkları, mesajlaşma sayıları, eğitmenlerine yönelttikleri soru sayıları ve final notlarına ilişkin veriler derlenmiştir. Veri ön işlemesi sonrasında 298 öğrencinin verileri ile veri analiz süreci devam ettirilmiştir. Araştırmada sunulan analitik yaklaşım üç temel safhada yürütülmüştür:

- Ön-işleme safhası: Ham verinin, temizlenme, niteliklerin (attributes) atanması ve verilerin entegre edilmesi ile kullanılabilir formata dönüştürülebilmesi
- İkinci safha: Sınıflama, kümeleme ve görselleştirme gibi veri madenciliği stratejilerinin kullanılması
- Son işleme safhası: Yorumlama ve elde edilen bilginin yeniden düşünülme ve karar vermede kullanımı.

Araştırma kapsamında tüm öğrenci verileri MSQL server veri tabanında derlenmiş, kümeleme analizinde verinin işlenmesi için Php programlama dili ile bir program yazılmıştır. Ayrıca öğrencilerin soruları ve cevapları için otomatik kodlama yapılabilmesi için NVivo 9 yazılımı kullanılmıştır. Öğrencilerin kısa mesajlaşmalarından elde edilen veriyi SPSS Clementine metin madenciliği aracılığıyla analiz etmişlerdir. Araştırma sonucunda EDM'nin öğrencinin öğrenme kapasitesinin belirlenmesinde önemli bir güç olduğu, özellikle bu hibrit yaklaşım kullanılarak öğrenme deneyimlerinin nasıl derinleştirilebileceği ve yüksek eğitimin yeniden şekillendirilebilmesinin mümkün olabileceğini vurgulamışlardır.

Osmanbegovic ve Suljic (2012); yükseköğretimin kaliteli insan kaynağının oluşturulmasında önemli bir basamak olduğu düşüncesinden yola çıkarak bir EVM çalışması yapmışlardır. Bu çalışmayı Bosna-Herzigova Tuzla Üniversitesi, Ekonomi Fakültesi 2010-2011 eğitim yılında, öğrenime başlayan birinci sınıf öğrencilerinden alınan verileri kullanarak gerçekleştirmişlerdir. Çalışma kapsamında klasik eğitime tabi 257 öğrenciye ait sosyo-demografik değişkenler, lise ve üniversite giriş sınavı sonuçları, ders çalışmaya yönelik tutumları incelenmiştir. Öğrenci başarısında kriter olarak yıl sonu sınavları seçilmiştir. Elde edilen veriler üzerinde bayezyen sınıflama, yapay sinir ağları ve karar ağaçları gibi farklı sınıflama teknikleri uygulanmıştır. Çalışmada WEKA yazılımından yararlanılmış, girdi değişkenlerinin etkisinin öğrenci başarısının tahmin edilmesinde daha iyi analiz edilmesi için ki-kare testi, one R-testi, Info Gain test ve Gain Ratio testi kullanılmıştır. Çalışma sonucunda Naif Bayez sınıflayıcının, yapay sinir ağları ve karar ağaçlarına göre daha başarılı bir tahmin sağladığı sonucuna varılmıştır.

2.6.2. Ülkemizde Yapılan Çalışmalar

Ülkemizde veri madenciliği yöntem ve tekniklerinin eğitim alanında uygulandığı, eğitimde veri madenciliği çalışmaları oldukça sınırlı düzeydedir. Yapılan çalışmaların neredeyse tamamına yakını veri tabanlarında bulunan kayıtlı veriler üzerinden yürütülmüştür.

Çifti (2006) tarafından yapılan çalışmada uzaktan eğitim verileri baz alınarak, ders içi etkinliklerin değerlendirilmesine yönelik anketler uygulanmıştır. Bu anketler üzerinden veri madenciliği çalışmaalrı yapılmıştır.

Gülen ve Özdemir (2013) yaptıkları çalışmada 113 öğrenciden elde edilen verileri kullanarak, üstün yetenekli öğrencilerin tespitine yönelik veri madenciliği çalışması yapmışlardır. Bu çalışmada apriori algoritması kullanılmıştır (Gülen ve Özdemir, 2013)

Şengür ve Tekin (2013); yapay sinir ağları ve karar ağaçları yöntemlerini kullanarak Fırat Üniversitesi, Eğitim Fakültesi, Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü (BÖTE) öğrencilerinin (n=127) mezuniyet notlarını tahminlemeye yönelik çalışmalar yapmışlardır.

Yıldız (2014); Bilgisayar Bilimleri Uzaktan Eğitim Programına kayıtlı 218 öğrenciye ait verileri kullanarak öğrenci performansını makine öğrenmesi yöntemleri ile tahminlemeye çalışmıştır.

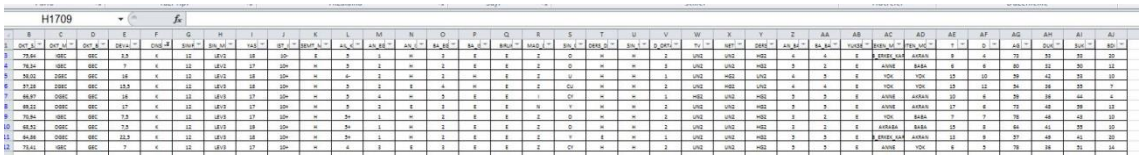
3. MALZEME VE YÖNTEM

Bu bölümde sırasıyla, örneklem grubu, veri toplama araçları ve içerikleri ile uygulanan süreçler, kullanılan programlama dili ve analiz ortamı hakkında bilgiler verilmektedir. Süreçler bölümünün altındaki adımlar, tez kapsamında geliştirilen ve Bölüm 2.3.2.'de sunulan model önerisi ışığında düzenlenmiştir.

3.1. ÖRNEKLEM GRUBU VE SEÇİMİ

Araştırmanın örneklem grubunu; İstanbul ilinde bulunan Fatih, Gaziosmanpaşa, Beşiktaş, Etiler, Levent, Şişli, Kağıthane, Bakırköy, Bağcılar, Bahçelievler, K. Çekmece, Sarıyer, Eyüp, Kadıköy ve Üsküdar ilçelerindeki liselerde 2014-2015 ve 2015-2016 eğitim öğretim yıllarında eğitim görmekte olan lise öğrencileri oluşturmaktadır.

Örneklem grubu olasılıksız örnekleme yöntemlerinden amaçlı örnekleme yöntemi baz alınarak oluşturulmuştur. İlçelerdeki genel eğitim düzeyini temsil eden her liseden, okulun en başarılı ve en başarısızları olarak nitelenen toplam 2371 öğrenci yer almaktadır. Veri ön işleme sonrasında EVM çalışması kapsamında, 887'si erkek ve 819'u kadın olmak üzere 1706 öğrencinin verisinin kullanılabilir durumda olduğu anlaşılmıştır (Şekil 3.1).



Şekil 3.1: EDM Excel Görüntüsü.

3.2. VERİ TOPLAMA ARAÇLARI VE İÇERİKLERİ

Bu araştırmada, akademik başarı üzerinde etkisi olan değişkenlerin tespiti için araştırmacı tarafından oluşturulan "kişisel bilgi formu" ve literatürde geçerlilik

güvenirlik çalışmaları yapılmış, kabul görmüş çeşitli ölçekler, veri toplama araçları olarak kullanılmıştır.

Araştırmada soru formları ve ölçekler öğretmenler ve öğrenciler için ayrı ayrı hazırlanmış ve uygulanmıştır:

- Öğrenciler:
  - Kişisel Bilgi Formu (KBF-ÖGR),
  - Maslach Tükenmişlik Envanteri (MTE-ÖĞR),
  - Akademik Güdülenme Ölçeği (AGÖ),
  - Akademik Öz-Yeterlilik Ölçeği (AÖYÖ)
  - Durumluk Sürekli Kaygı Ölçeği (DSKÖ)
- Öğretmeler:
  - Kişisel Bilgi Formu (KBF-ÖGRT),
  - Beck Depresyon Envanteri (BDI)

Ölçekler; psikolog/psikiyatr gibi özel bir uzmanın eşliğinde uygulanmasını ve yorumlanmasını gerektirmeyecek şekilde seçilmiştir. Ölçeklerden elde edilen değerlerin/ölçümlerin güvenirliği, cronbach alfa güvenirlik katsayısı hesaplanarak incelenmiştir. Araştırmacı; güvenirlik katsayısı için excelde formül oluşturarak herbir ölçek ve envanter için güvenirlik analizi yapmıştır. Bu katsayı için yorum aralıkları Tablo 3.1'de verilmektedir.

Tablo 3.1: Cronbach Alfa Güvenirlik Katsayısı Yorum Aralıkları.

Güvenirlik	Yorum
$1> \alpha \geq 0{,}9$	Yapılan ölçüm yüksek derecede güvenilir
$0.9 > \alpha \geq 0.8$	Yapılan ölçüm oldukça güvenilir
$0.8 > \alpha \geq 0.7$	Yapılan ölçüm güvenilir
$0.7 > \alpha \geq 0.6$	Yapılan ölçüm kabul edilebilir derecede güvenilir
$0.6 > \alpha \geq 0.5$	Yapılan ölçüm zayıf derecede güvenilir
$0.5 > \alpha$	Yapılan ölçüm güvenilir değil

Veri toplama araçlarının geliştiricileri, uyarlayıcıları, cevaplanma biçimi, aracın uygulanmasındaki temel prensip/sağlayacağı değişken, puanlanma biçimi ve güvenirlik katsayıları bundan sonraki bölümlerde detaylı olarak açıklanacaktır.

3.2.1.Kişisel Bilgi Formu (KBF)

Araştırmacı tarafından geliştirilen Kişisel Bilgi Formu (KBF); öğrenciler ve öğretmenler için ayrı ayrı iki form olacak şekilde tasarlanmıştır. KBF'ler; çoktan seçmeli, 5'li likert tipi ve açık uçlu olmak üzere toplam 28 sorudan oluşmaktadır.

Öğrencilere uygulanan Kişisel Bilgi Formu (KBF-ÖGR); cinsiyet, doğum tarihi, doğum yeri, oturduğu semt, ilde ikamet süresi, semtte ikamet süresi, semtten memnuniyet, ailedeki birey sayısı, kardeş sayısı, annenin doğduğu şehir, eğitim durumu, çalışma durumu; babanın doğduğu şehir, eğitim durumu, çalışma durumu; anne ve babanın birlikteliği, ailenin ortalama aylık geliri, öğrencinin sınıfta oturduğu yer (öğretmen masası ve tahtaya olan mesafe), okul dışı ders desteği alması, evde ders çalışma ortamı, evdeki teknolojik imkanlar, cep telefonu sahipliği, okuldan sonraki zamanın değerlendirilme biçimi, günlük TV izleme, internet kullanma ve ders çalışma süresi, rol model olarak aldığı kişi belirlemeyi amaçlayan soruları içermektedir.

Öğretmenlere uygulanan Kişisel Bilgi Formu (KBF-ÖGRT); cinsiyet, doğum tarihi, doğum yeri, oturduğu semt, medeni durum, üniversiteden mezuniyet yılı, mezun olunan üniversite, branş, KPSS'ye girme sayısı, mesleki deneyim, okuldaki görev süresi, ilde ikamet süresi, semtte ikamet süresi, semtten memnuniyet, ailedeki birey sayısı, annenin doğduğu şehir, eğitim durumu, çalışma durumu; babanın; doğduğu şehir, eğitim durumu, çalışma durumu; evdeki teknolojik imkanlar, cep telefonu sahipliği, okuldan sonraki zamanın değerlendirilme biçimi, günlük TV izleme, internet kullanma süresini belirlemeyi amaçlayan soruları içermektedir.

3.2.2.Durumluk Sürekli Kaygı Ölçeği (DSKÖ)

Ölçek; Spielberg ve diğerleri (1970) tarafından geliştirilmeye başlanmış, kaygının anlık/durumluk ve sürekli olmak üzere iki seviyede incelenmesi gerektiği ortaya konulmuştur. Öner (1978) ölçeği Türkçe'ye uyarlamış ve standardizasyonunu gerçekleştirmiştir. Kısa ifadelerden oluşan ölçek; başlangıçta normal yetişkinlerdeki kaygının incelenmesi amacıyla geliştirilmiş olup, 10 yıllık denemeler sonrasında

psikiyatrik bozuklukları ve fiziksel hastalıkları olan bireylere, gençlere ve yetişkinlere de uygulanabileceği sonucuna ulaşılmıştır.

DSKÖ; toplamda 40 maddeyi içeren, durumluk kaygı ve sürekli kaygıyı ölçmeyi amaçlayan iki ayrı bölümden oluşmaktadır. Ölçekte düz/doğrudan ve ters/tersine (reverse) dönmüş ifadeler yer almaktadır. Bu ölçek aracılığıyla bireyler için biri durumluk kaygıyı, diğeri sürekli kaygıyı gösteren iki farklı değer elde edilmektedir.

Durumluk Kaygı Ölçeği (DKÖ) ; 1 (hiç), 2 (biraz), 3 (çok) ve 4(tamamıyla) olmak üzere 4 puanlı bir Likert Ölçeği biçimine sahiptir. Bu bölümde yer alan maddeler; belirtilen duygu ya da davranışların kişi tarafından şiddetinin değerlendirilmesini amaçlamaktadır.

Sürekli Kaygı Ölçeği (SKÖ) de 1 (hemen hiçbir zaman), 2 (bazen), 3 (çok zaman) ve 4 (hemen her zaman) olmak üzere 4 puanlı bir Likert Ölçeği'ni içermektedir. Bu bölüm; maddelerde bulunan duygu ve davranışların kişinin hayatındaki sıklık derecesini incelemektedir.

DSKÖ'den elde edilen cevapların puanlaması yapılırken düz/doğrudan ve ters/tersine dönmüş ifadelere dikkat edilmesi gerekmektedir. Durumluk kaygı bölümünde 1, 2, 5, 8, 11, 15, 16, 19 ve 20. maddeler, sürekli kaygı bölümünde ise 21, 26, 27, 30, 33, 36 ve 39. maddeler; ters/tersine dönmüş ifadeleri içermektedir. Puanlamada doğrudan ifadelerin toplam ağırlıklı puanından, tersine dönmüş ifadelerin toplam ağırlıklı puanı çıkarılmaktadır. Elde edilen değerlere ölçeğin değişmeyen değerleri; durumluk kaygı için 50, sürekli kaygı için 35 puan eklenmektedir. Öner (1978)'e göre 36-41 puan arası lise ve üniversite öğrencileri için normal kabul edilebilecek bir aralığı temsil etmektedir.

Araştırmada DSKÖ yardımıyla elde edilen ölçüm; DKÖ için 0,93, SKÖ için 0,95 cronbach alfa güvenirlik değeri ile yüksek derecede güvenilirdir.

3.2.3.Akademik Güdülenme Ölçeği (AGÖ)

Akademik Güdülenme Ölçeği (AGÖ); Bozanoğlu (2004) tarafından geliştirilmiştir. Ölçek; Kendini Aşma (KA), Bilgiyi Kullanma (BK), Keşif (K) olmak üzere üç faktörden oluşmaktadır. Her faktöre yönelik ayrı ayrı puanlar elde edilebildiği gibi,

toplam bir puanda verilebilmektedir. Tablo 3.2'de AGÖ'nün alt faktörleri ve madde numaraları verilmektedir.

Tablo 3.2: AGÖ'nün Alt Faktörleri Ve Madde Numaraları.

Faktörler	Maddeler
Kendini Aşma (KA)	16, 8, 10, 9, 7, 6, 2
Bilgiyi Kullanma (BK)	15, 1, 12, 18, 5, 14
Keşif (K)	11, 13, 4, 3, 17, 19, 20

KA; öğrencinin ders ve ders dışı konuları öğrenme ve ödev hazırlığında kendisinden beklenenden daha iyisini yapma isteğini temsil etmektedir. Diğer bir deyişle bu faktör; bireyin kendini akademik olarak geliştirme isteği ile ilgilidir.

BK faktörü; öğrencinin öğrenmedeki heves ve heyecanını ifade etmektedir. Aynı zamanda öğrendiklerini okul da ve okul dışında kullanabilme isteği ile ilgilidir.

K ise; ödül beklemeksizin merak ettiği ve daha iyisini yapmak istediği için öğrenme isteğini temsil etmektedir.

AGÖ; 20 maddeden oluşmakta olup, beşli Likert tipinde cevaplandırılma özelliğine sahiptir. Ölçekte 4. madde hariç diğer maddeler doğrudan/düz olarak puanlanmaktadır. Alınabilecek minimum puan 20, maksimum puan ise 100'dür (Bozanoğlu, 2004).

Araştırmada AGÖ yardımıyla elde edilen ölçüm; 0,82 cronbach alfa güvenirlik değeri ile oldukça güvenilirdir.

3.2.4. Maslach Tükenmişlik Envanteri (MTE)

Maslach Tükenmişlik Envanteri (MTE); Maslach ve Jackson (1981) tarafından geliştirilmiş olup, uyarlaması Çapri (2006) tarafından gerçekleştirilmiştir. Ölçeğin öğrenci formu (MTE-ÖĞR) 16 maddeden oluşmakta ve üç farklı faktör içermektedir. Ölçekte 6, 12 ve 15. maddeler puan değeri olarak hesaplanmamaktadır. Tablo 3.3'de MTE-ÖĞR'nin alt faktörleri ve madde numaraları verilmektedir.

Tablo 3.3: MTE-ÖĞR'nin Alt Faktörleri Ve Madde Numaraları.

Faktörler	Maddeler
Tükenme (T)	1, 4, 7, 10,13
Duyarsızlaşma (DY)	2, 5, 8, 11
Yetkinlik (Y)	3, 9, 14,16

MTE-ÖĞR'de üç farklı puan türü elde edilmektedir, bunların tek bir puan biçiminde yorumlanması mümkün değildir. T faktörü; uğraşılan iş/meslek/uğraşı anlamında bireyin tüketilmiş ve kendisine aşırı yüklenilmiş olması duygularını tanımlamaktadır. D faktörü; bireyin aynı uğraşıyı paylaştığı kişilere ve kendisine karşı duygudan yoksun bir şekilde davranmasını ifade etmektedir. Y faktörü ise; yeterlilik ve başarı ile üstesinden gelme duygularını tanımlamaktadır. Ölçekte T ve D'ye ait yüksek puanlar, Y'ye ait düşük puan tükenmişliği göstermektedir.

Araştırmada MTE-ÖĞR yardımıyla elde edilen ölçüm; T için 0,65, DY için 0,69 ve Y için 0,63 cronbach alfa güvenirlik değeri ile kabul edilebilir düzeyde güvenilirdir.

3.2.5.Beck Derpresyon Envanteri (BDI)

Beck Depresyon Envanteri (BDI); 1978 yılında Beck tarafından geliştirilmiş kendini değerlendirme türünde olan ve grup uygulaması da yapılabilecek bir yapıya sahiptir. Uyarlanması, geçerlilik ve güvenirlik çalışmaları Hisli (1989) tarafından yapılmıştır.

BDI'nın amacı; depresyon tanısı koymak değil, depresyona ilişkin belirtilerin derecesini objektif sayılar olarak sunmaktır (Hisli, 1989). Envanterdeki her madde depresyona özgü bir davranışsal örüntüyü belirlemeyi hedefleyen toplan 21 değerlendirme cümlesinden oluşmaktadır. BDI'dan elde edilen cevaplar puanlanırken a=0, b=1, c=2 ve d=3 kullanılır. Dolayısıyla envanterden alınabilecek en düşük puan 0, en yüksek puan ise 63'dür. Puanın yüksek olması depresyon düzeyinin yüksek olmasını göstermektedir. Tablo 3.4'de BDI puanlarına ilişkin aralıklar ve yorumları verilmektedir.

Tablo 3.4: BDI Puanlarına İlişkin Aralıklar Ve Yorumları.

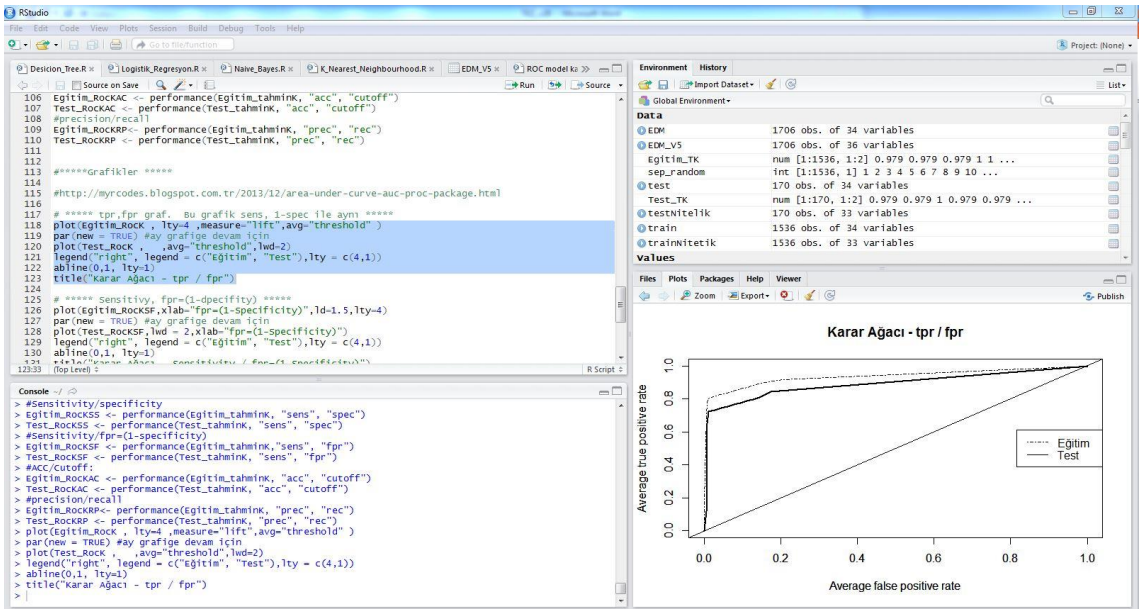
Puan Aralıkları	Yorum
$0 \leq BDI < 10$	Minimal düzeyde depresif belirtiler
$10 \leq BDI < 17$	Hafif düzeyde depresif belirtiler
$17 \leq BDI < 30$	Orta düzeyde depresif belirtiler
$30 \leq BDI \leq 63$	Şiddetli düzeyde depresif belirtiler

Araştırmada BDI sadece öğretmenlere uygulanmıştır. Elde edilen ölçüm; 0,69 cronbach alfa güvenirlik değeri ile kabul edilebilir düzeyde güvenilirdir.

3.3. ANALİZ ARACI

EVM ilişkin çalışmalar ve sınıflandırma algoritmalarının uygulanarak gerekli analizlerin yapılması işlemlerinde R dili kullanılmıştır. Analiz sürecinde gerekli kodların yazılması için daha kullanıcı dostu ara yüze sahip RStudio ortamı tercih edilmiştir.

R dili istatistik analiz ve görselleştirme amacıyla kullanılabilen açık kaynak kodlu ve ücretsiz bir yazılımdır. Bu yazılım Bell laboratuvarlarında geliştirilmiş olup, istatistik ve ekonometride kullanılan S dilinin bir uzantısıdır (Becker, Chambers, & Wilks, 1988) (Grunsky, 2002). İnternet üzerinden erişime açık farklı istatistik yöntemler ve farklı veri türlerine yönelik analiz yöntemleri üzerine pek çok paketi bulunmaktadır. 1 Ağustos 2012'de web katılan Comprehensive R Archive Network (CRAN) paket deposu ile bireyin analiz yapmak istediği yönteme ya da veriye uygun paketi indirmesi, kurması mümkündür. Tüm bunların yanı sıra kullanıcın kendi paketini oluşturmasına olanak sağladığından geliştirilebilir niteliktedir. Şekil 3.2'de RStudio ekran görüntüsü verilmektedir.



Şekil 3.2: RStudio Ekran Görüntüsü.

R dili; http://www.r-project.org/ adresinden Linux, Mac ve Windows gibi farklı işletim sistemleri üzerinde kolaylıkla çalıştırılabilecek seçenekleri ile sunulmaktadır (CRAN, 1999).

R; istatiksel programlama ve grafikleştirme sistemidir. Bu sistem; istatiksel analiz için kullanılabilirliğinin yanı sıra bir programlama dili, yüksek düzeyli grafikleştirme imkanı, diğer dillerle uyumlu arayüzler ve hata yakalama özellikleri de sunmaktadır.

R dili akademik çalışmalarda ve özel sektöre yönelik analizlerde de kullanılmaktadır. Dil, C programlama dilinin komutlarına benzer şekilde kodlar ile çalıştırılabilmektedir. Sadece çevrimdışı ortamlardaki verilerle değil, .csv uzantılı ve webte kullanıma hazır olarak sunulmuş veri setlerinin de içe aktarılarak analiz edilmesi mümkündür.

R dilinin kullanımı; nesne yönelimli programlamadaki gibi veri yapısının nasıl tanımlandığına dayanmaktadır. Böylede R nesneleri sınıfları ve metotları oluşturabilmektedir. R dilinin sağladığı bu çalışma kolaylığı ile, işlenen verinin türüne yönelik fonksiyonun çalıştırılmasına dayalı daha yüksek seviyeli programlama yapılabilmektedir (Grunsky, 2002).

R matematiksel ve istatiksel programlama yapılmasına, veri yönetim fonksiyonları üretilmesine uygun bir ortama sahiptir. Sadece basit düzeyde ya da sadece gelişmiş düzeyde programlama yapılması ile sınırlı değildir. Programlama ortamında kullanıcının bireysel yetkinliği ile farklı düzeylerde nesne eğilimli programları kullanabilmesi hatta kendi programını/fonksiyonunu geliştirilebilmesi mümkündür.

3.4. SÜREÇLER

Bu bölümde uygulamanın planlanmasından hayata geçirilmesine, verilerin derlenmesinden, analiz edilerek sonuçlar üretilmesine kadar geçen aşamalar; tez kapsamında önerilen CRISP-EDM adımları çerçevesinde sunulmaktadır.

3.4.1.Problemi/Hedefi Tanımlama

Ülkemizde eğitim-öğretim sürecinde akademik başarıya dayalı öğrenci performansı ölçülmektedir. Bu ölçüt doğrultusunda öğrenci o akademik yılı başarılı ya da başarısız olarak tamamlamaktadır. Öğrencinin akademik başarısı, onun bir sonraki kademeye geçişini sağlamakta ya da mevcut kademeyi tekrar etmesine neden olmaktadır. Eğitim sistemimiz içindeki akademik başarı; belirli bir müfredat çerçevesinde işlenen derslerin yıl içi ve yılsonu sınavlarındaki başarıların ortalamaları uyarınca hesaplanmaktadır.

Milli Eğitim Bakanlığı 2015-2019 Stratejik Planı'nda öğrencinin başarısı; stratejik planının mimarisindeki önemli ayaklardan biridir (MEB, 2015). Eğitimde kalite başlığı altında sunulan öğrenci başarısı; ülke genelinde düzenlenen, diğer ülkelerle kıyaslanma noktasında ölçümler ortaya koyan sınavlar açısından da parlak bir durum sergilememektedir. Nitekim PISA 2009 sonuçlarına göre ülkemizden katılan 15 yaş grubu öğrencilerin %42'sinin basit matematiksel problemleri çözebilecek düzeyde değildir, %25'inin okuduğunu anlamamakta, %30'unun ise günlük hayatta karşılaşılabilecek fen ve teknoloji problemlerini çözememektedir (ERG, 2011).

Düşük okullaşma oranı, eğitimde bilgiye erişim ve fırsat anlamındaki eşitsizliklerin yanı sıra, devamlılıkta öğrencinin akademik başarısını etkilemektedir. 2015-2019 Stratejik Planı'nda "ortaöğretimde devamsızlık, sınıf tekrarı ve okuldan erken ayrılma nedenlerinin tespiti için araştırmalar yapılması, devamsızlık, sınıf tekrarı ve okul terkinin azaltılması amacıyla çeşitli uyum programlarının yaygınlaştırılması, öğrenci devamsızlıklarını izleme ve önleme mekanizmalarının geliştirilmesi" stratejilerine yer

verilmektedir. Gerçekten de Ülkemizde 2008-2009 yılları arasında 360.000, 2009-2010 yılları arasında 295.000 ortaöğretim öğrencisi, diploma almaksızın okulu bırakmıştır (ERG, 2011). 2014 yılında ise eğitim ve öğretimden erken ayrılma oranı %38,2'dir. Orgün eğitimde 20 gün ve üzeri devamsız öğrenci oranı; ortaöğretim düzeyinde 2014 yılı itibariyle %34,8'dir (MEB, 2015). Okula kayıt olmuş, eğitim-öğretim sürecine dahil olmayı seçmiş bu kadar çok sayıdaki öğrencinin eğitim-öğretim süreci içinde kalamaması, hem sürece, hem de akademik başarıya özgü bir takım sıkıntılara işaret etmektedir.

Öğrencinin akademik başarısını etkileyen faktörler üzerine literatürde çok çeşitli istatistik temelli çalışmalar bulunmaktadır. Ülkemizde yapılan çalışmalarda, 6 ya da 7 faktör üzerinden başarıyı yordama, başarı ile ilişki durumları incelenmiş, faktörlerin başarıyı öngörmesi noktasında bir çalışmaya rastlanmamıştır.

Bu tez çalışmasında "veri madenciliği tekniklerinden sınıflandırma algoritmaları kullanılarak, eğitim sürecinden derlenen veriler ışığında öğrencinin yılsonu akademik başarısının belirlenmesi problemi" ele alınmıştır.

3.4.2. Uygulama Adımlarını Planlama

Bu araştırma için, ilçeler bazında kaymakamlıklar ve ilçe milli eğitim müdürlüklerinden, il bazında İstanbul Valiliği ve İl Milli Eğitim Müdürlüğü'nden gerekli izinler alınmıştır. Uygulama 2014-2015 ve 2015-2016 eğitim öğretim yıllarında, Fatih, Gaziosmanpaşa, Beşiktaş, Etiler, Levent, Şişli, Kağıthane, Bakırköy, Bağcılar, Bahçelievler, K. Çekmece, Sarıyer, Eyüp, Kadıköy ve Üsküdar ilçelerindeki liselerde yapılmıştır. Uygulamanın ilk aşamasında ilgili makamlarla görüşülerek çalışmanın amacı ve kapsamı hakkında detaylı bilgiler verilmiş, kullanılacak veri toplama araçları incelemeye sunulmuştur. Sonraki aşamada ilçe, il milli eğitim müdürlükleri ile görüşülerek ilçeyi genel olarak temsil edebilecek okulların isimleri belirlenmiştir. Belirlenen okulların müdürleri, rehber öğretmenler ve sınıf öğretmenleri ile görüşülerek her sınıf seviyesinden başarılı en az 30 öğrenci ve başarısız en az 30 öğrenci olacak şekilde listeler oluşturulmuştur. Öğrencilerin araştırmaya katılım süreçlerinde gönüllülük ilkesi ve ailelerinin onayının alınmasına dikkat edilmiştir.

Okul yöneticileri ve rehber öğretmenler ile birlikte yapılan planlamalar ışığında okullara gidilerek, zaman kısıtlaması olmaksızın ve gruplar halinde veri toplama araçları uygulanmıştır.

Veri toplama araçlarına ek olarak; öğrencilerin dönem sonundaki başarıları, devamsızlık durumları, okul idaresinin izni dahilinde alınmıştır.

Uygulama süreci; fazla sayıda soru formu ve ölçekten oluştuğundan, tek defada uygulanmamıştır. Özellikle DSKÖ uygulanırken "kendini nasıl hissedersin" sorunun cevabını aramaya yönelik olduğu söylenmiş ve kaygı sözcüğü kullanılmamıştır.

3.4.3.Verileri Derleme ve Ön İnceleme

Tez çalışması kapsamında derlenen veri; veri madenciliği yöntem ve teknikleri uygulanmadan önce ön işleme sürecinden geçirilmiştir. Bu sayede eksik veriler belirlenmiş, istenilen düzeyde örneklemin oluşup oluşmadığı kontrol edilmiştir.

3.4.4.Veriyi Anlama ve Hazırlama

EDM olarak isimlendirilen veri seti ham halinde 61 farklı nitelik ve 2371 öğrenci verisi içermektedir. Tablo 3.5'de EDM veri setine ilişkin tüm değişkenler, gösterim biçimleri ve türleri verilmektedir.

Tablo 3.5: EDM Veri Setine İlişkin Tüm Değişkenler, Gösterim Biçimleri Ve Türleri.

Açıklama	Gösterim	Tür
Adı Soyadı	AS_KOD	Text
Okul Ortalaması (Hedef Nitelik)	OKT	Tamsayı
Toplam Devamsızlık Sayısı	DEVAM	Tamsayı
Cinsiyeti	CINS	Kategorik
Okul Türü	OK_TUR	Kategorik
Sınıfı	SINIF	Sayısal
Yaşı	YAS	Tamsayı
Doğum Yeri	DOGUM_YER	Text
Oturduğu Semt	SEMT	Text
Anne Memleket	AN_DY	Text
Baba Memleket	BA_DY	Text
İstanbul'da İkamet	IST_IK	Sayısal
Semtte İkamet	SEMT_IK	Sayısal
Semtten Memnuniyet	SEMT_MEM	Kategorik
Aynı Evde Yaşayan Kişi Sayısı	AİL_KS	Tamsayı
Kardeş Sayısı	KS	Tamsayı
Kaçıncı Çocuk	KC	Sayısal

Anne Eğitim Durumu	AN_EGIT	Sayısal
Anne Çalışma Durumu	AN_IS	Kategorik
Babanın Eğitim Durumu	BA_EGIT	Sayısal
Babanın Çalışma Durumu	BA_IS	Kategorik
Kardes Egitim Durumu	KA_EGIT	Sayısal
Anne Babanın Birlikteliği	BIRLIK	Kategorik
Ailenin Aylık Maddi Durumu	MAD_DU	Sayısal
Aile Geçimine Destek	MAD_DE	Kategorik
9. Sınıfa Bu Okulda Başlama	OKUL	Kategorik
Sıra Öğretmen Masası Ve Tahta Mesafesi	SIN_O	Sayısal
Ders Desteği Alması	DERS_DEST	Kategorik
En Sevdiği Ders	ESEV_DERS	Text
Nedeni	SEV_N	Text
En Az Sevdiği Ders	EASEV_DER	Text
Nedeni	SEVM_N	Text
En Çok Sevdiği Öğretmenin Branşı	ESEV_D_OGRT	Text
Nedeni	SEV_OGRT_N	Text
En Az Sevdiği Öğretmenin Branşı	ESEVM_D_OGRT	Text
Nedeni	SEVM_OGRT_N	Text
Sınıf Tekrarı	SIN_T	Kategorik
Ders Çalışma Ortamı	D_ORTAM	Sayısal
Evin Teknolojik Cihazlar Bakımından Durumu	TEK_D	Sayısal
Cep Telefonu Sahipliği	CEP	Kategorik
Kullanım Süresi	CEP_SUR	Sayısal
Telefonun Türü	CEP_TUR	Sayısal
Televizyon İzleme Süresi	TV	Sayısal
İnternet Kullanım Süresi	NET	Sayısal
Ders Çalışma Süresi	DERS	Sayısal
Anne İle İlişki	AN_BAG	Sayısal
Baba İle İlişki	BA_BAG	Sayısal
Yüksek Öğrenime Devam Etme İsteği	YUKSEK	Kategorik
Hayır İse Nedeni	YUKSEK_N	Text
Kim Gibi Olmak İster	CEKEN_MODEL	Text
Kim Gibi Olmak İstemez	ITEN_MODEL	Text
En Çok Yapmak İstediği Meslek	IST_MES	Text
Asla Yapmak İstemediği Meslek	ISTM_MES	Text
Olumu Yanı	POZ_OZ	Text
Olumsuz Yanı	NEG_OZ	Text
Tükenme	T	Tamsayı
Duyarsızlaşma	D	Tamsayı
Akademik Güdülenme	AG	Tamsayı
Durumluk Kaygı	DUK	Tamsayı
Sürekli Kaygı	SUK	Tamsayı
Öğretmen Depresyon Durumu	OGRT_BDI	Tamsayı

Analiz sürecinde dahil edilecek verinin, örneklem olarak mevcut evreni temsil durumu Denklem (3.1) ‘de verilen formül ile hesaplanmıştır.

N: evrendeki birey saysıs

n: örnekleme alınacak birey sayısı

p: İncelenecek olayın görülüş sıklığı (olasılığı)

q: İncelecek olayın görülmeyiş sıklığı (1-p)

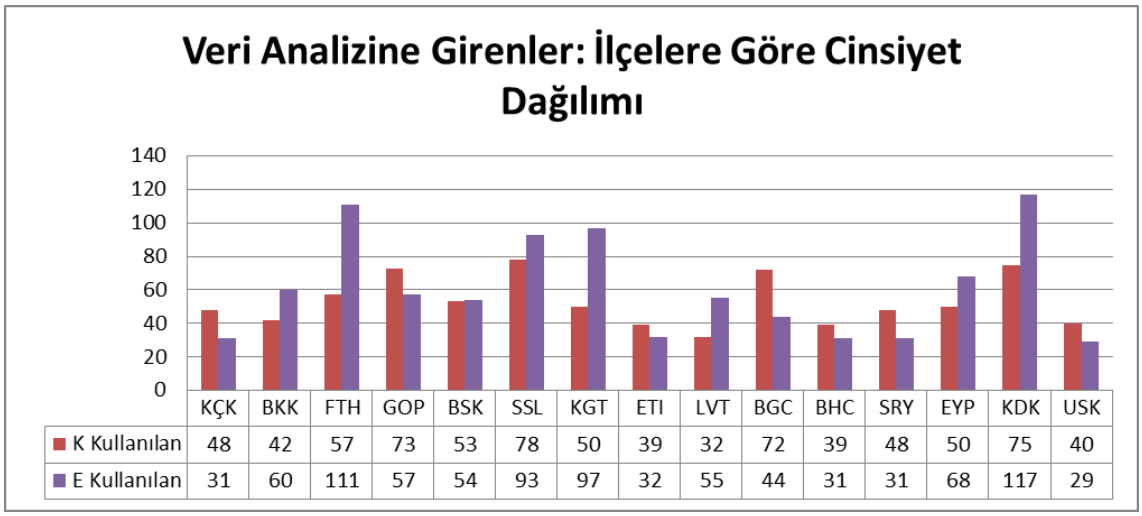
t: Belirli serbestlik derecesinde ve saptanan yanılma düzeyinde t tablosunda bulunan teorik değer

d: Olayın görülüş sıklığına göre yapılmak istenen ± sapma olmak üzere;

$$n^{\iota} = \frac{N\ t^2 p\ q}{t^2(N-1) + \ t^2\ p\ q} \qquad (3.1)$$

Evren büyüklüğünün 500 bin ve üstü olduğu durumlarda %95'lik güven düzeyi üzerinden incelendiğinde, örneklem büyüklüğünün 384 (%5) ile 9423 (%1) arasında olması beklenmektedir (Yazıcıoğlu ve Erdoğan, 2004). Bu açıdan değerlendirildiğinde analize dahil edilmesi planlanan öğrenci verisinin mevcut evreni temsil edebilecek sayıda olduğu söylenebilmektedir.

Veri ön işleme sürecinde ana hedef doğrultusunda 34 değişken ile devam edilmesine karar verilmiştir. Elde edilen veriler incelendiğinde eksik verilerden kaynaklı olarak analiz sürecine 1706 öğrenci verisinin dahil edilmesine karar verilmiştir (Şekil 3.3). Eksik değerlerin "kayıp değerlerin analizi ve tahminine" yönelik yöntemler kullanılarak bulunmamasının ana nedeni; öğrencinin ruh hali ve davranışlarına yönelik olan envanterlerdeki eksikliklerin, "öğrenci profilinin canlandırılarak, o soruya cevap verme biçiminin uzmanlar tarafından tahmin edilmesi" ya da "bizzat öğrencinin kendisinden alınması" gerekliliğinden kaynaklanmaktadır.



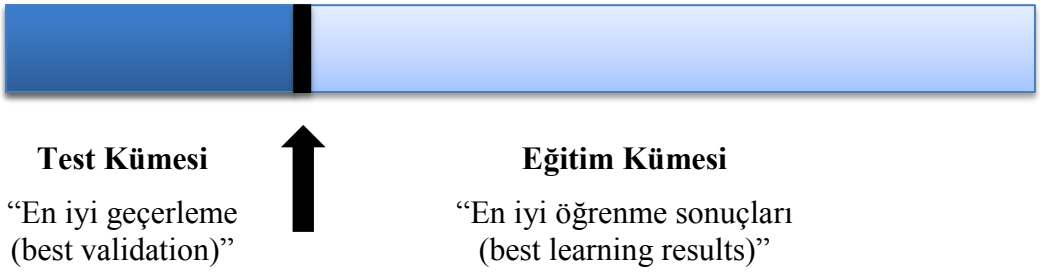
	KÇK	BKK	FTH	GOP	BSK	SSL	KGT	ETI	LVT	BGC	BHC	SRY	EYP	KDK	USK
K Kullanılan	48	42	57	73	53	78	50	39	32	72	39	48	50	75	40
E Kullanılan	31	60	111	57	54	93	97	32	55	44	31	31	68	117	29

Şekil 3.3: İlçelere Göre Analizde Kullanılan Verilerin Dağılım.

3.4.5.Modelleme

Bu tez çalışmasında, öğrencinin akademik başarısının sınıflandırılması için alternatif modeller oluşturulmuştur. Bu modellerin kurulmasında bu bölüm içinde detayları verilen *Logistik Regresyon Analizi, ID3 ve C4.5 Karar Ağacı Algoritmaları, Naive Bayes Sınıflandırıcı ve Destek Vektör Sınıflandırıcı*'dan faydalanılmıştır. Kullanılan algoritmaların kıyaslanması için, model performans değerlendirme yöntemlerinden tabakalı örnekleme ve çapraz geçerleme yöntemleri seçilmiştir.

Belirli bir amaç doğrultusunda eldeki veri setine, veri madenciliği algoritmaları uygulanırken, eğitim ve test verilerine ayrımında geçerlilik ve aşırı öğrenmeye kaçmadan en iyi biçimde öğrenmenin sağlanması hedeflenmektedir (Şekil 3.4).



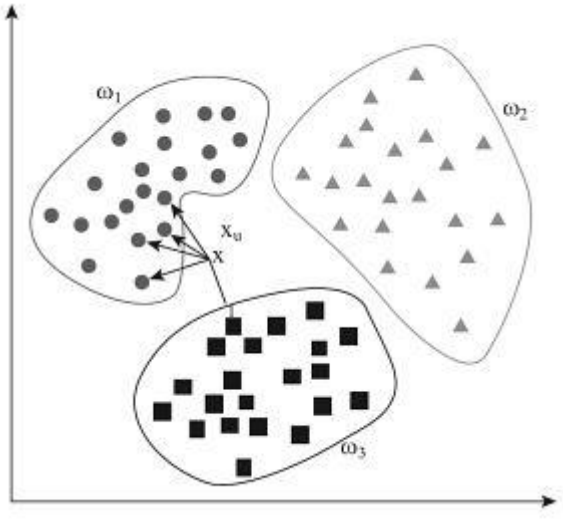
Şekil 3.4: Eğitim Ve Test Kümlerinin Ayrımına İlişkin Beklenti.

Tabakalı k-kat çapraz geçerlemede algoritmaların performansları 2-kat, 4-kat, 5-kat ve 10-kat çapraz geçerleme ile kıyaslanmıştır. Benzer şekilde hold-out yöntemi ile mevcut veri seti %90/%10, %80/%20, %75/%25, %65%35, %50/%50, %35/%65, %25/%75 oranlarında eğitim ve test veri setleri olacak şekilde ayrılmıştır. Bu yöntemler sayesinde, elde edilen performans ölçümlerinin doğruluk ve tutarlılık açısından karşılaştırılması mümkün olabilecektir.

3.4.5.1. k-En Yakın Komşu Algoritması

K-En yakın komşu (K-Nearest Neighbourhood, KNN) algoritmasına dayalı sınıflandırma işlemi, diğer yöntemlere göre daha basit bir yapı göstermektedir. Yakın komşuluk sınıflandırıcıları; benzeşim yoluyla öğrenmeye dayalıdırlar (Han ve diğ., 2012).

En yakın komşu sınıflandırıcı; bir eğitim kümesine ait üç sınıfı (w_1, w_2 ve w_3) kullanarak X_u gibi bir test örneklemi için bir sınıf etiketi bulmayı hedefler. Bu hedefe ulaşırken belli bir uzaklık formunu ve eşik değeri olarak belirlenmiş bir k değerini kullanır. Şekil 2.5'de k=5 değeri için en yakın komşu sınıflandırıcının çalışma biçimi verilmektedir.



Şekil 3.5: k=5 Değeri İçin En Yakın Komşu Sınıflandırıcı (Kantardzic, 2011).

Şekil 3 incelendiğinde sınıf etiketi bulunmaya çalışılan X_u'nun en yakın beş komşusundan dördünün w_1sınıfına ait olduğu görülmektedir. Bu durumda X_u, w_1sınıfının komşuluktaki baskınlığından dolayı bu sınıfa atanır.

En yakın komşu sınıflandırıcının çalışma prensibine göre; n niteliğe sahip bir eğitim veri seti için, veri setindeki her bir örnek n boyutlu uzayda bir nokta oluşturur. Buna bağlı olarak eğitim veri setinin n boyutlu bir örüntü uzayını barındırdığı söylenebilir. Sınıf değeri bilinmeyen yeni örnek verildiğinde, k-en yakın komşu algoritması, eğitim veri setinde bilinmeyen örneğe en yakın olanı bulmak için bu örüntü uzayını tarar (Han ve diğ., 2012). Buradaki yakınlık kavramı, Öklid uzaklığı gibi metrik bir uzaklığı ifade etmektedir. İki nokta veya iki örnek arasındaki Öklid uzaklığı Denklem 2.1'de verildiği gibi hesaplanır.

$X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ ve $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$ olmak üzere

$$dist(X_1, X_2) = \sqrt{\sum_{i}^{n} (x_{1i}, x_{2i})^2} \tag{2.1}$$

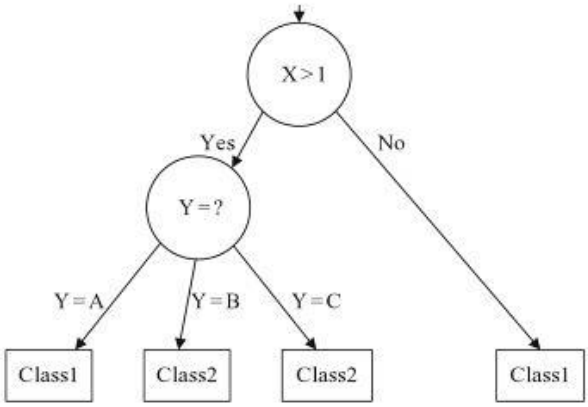
k-en yakın komşu sınıflandırıcı; n boyutlu uzayda çalışırken, bir k parametresi, eğitim veri seti ve bir uzaklık ölçütüne ihtiyaç duyar. kNN ile sınıflandırma sürecinde genellikle aşağıdaki adımlar izlenir (Kantardzic, 2011):

- En yakın komşuluk sayısı için k parametresi belirlenir
- Her test örneklemi ile tüm eğitim örneklemleri arasındaki uzaklık hesaplanır
- Uzaklıkların sıralanması ve k. eşik değerine göre en yakın komşuluklar tespit edilir
- En yakın komşulukların her biri için sınıfın/kategorinin belirlenir
- En yakın komşuların kategorilerinin dağılımdaki çoğunluklarına/en sık tekrar etme durumlarına göre sınıf tahmini, tahmini değeri olarak alınır.

3.4.5.2. Karar Ağacı Algoritmaları

Karar ağaçları, sınıflandırma problemlerinde en çok kullanılan algoritmalar arasında yer almaktadır. Bu algoritmalar; yukarıdan aşağıya doğru ilerleyen bir strateji ile çözüm elde etmek için araştırma uzayını taramaya dayalı çalışmaktadırlar (Kantardzic, 2011).

Bir karar ağacı; her düğümün nitelik değeri üzerindeki ölçümünü, her dalın o ölçüme ait bir çıktıyı ve ağaç yapraklarının da sınıfı ya da sınıf dağılımlarını gösterdiği ağaca benzeyen bir görünüme sahip bir akış diyagramıdır (Han ve diğ., 2012). Bu akış diyagramında tıpkı bir ağaç yapısında olduğu gibi kök, dal ve yapraklar bulunur. Kök düğüm ile başlayan yapı, veri dizisinin belli kriterlere göre ara düğümlere ayrılması ile devam eder ve yaprak düğümlerle sonlanır. Karar ağacında bulunan düğümler; birbirleri ile seviyelerine göre ebeveyn ve çocuk düğüm şeklinde isimlendirilir. Şekil 3.6'da iki nitelik için oluşturulmuş olan basit bir karar ağacı yapısı verilmektedir.



Şekil 3.6: X ve Y Nitelikleri Üzerinden Oluşturulan Basit Bir Karar Ağacı Yapısı.

Ağaç dallanması ile ilerleyen yapı; daha sağlıklı ilerlemenin sağlanması, hatta ezbere öğrenme sorununun ortadan kaldırılması için budama işlemine tabi tutulabilir (Akpınar, 2014). Budama işlemi çok az sayıda nesneyi barındıran yaprak düğümlerin ağaç yapısından çıkarılmasını içermektedir. Ancak bu işlemin fazlaca gerçekleştirilmesi, ağaç yapısında aşırı bir küçülmeye ve dolayısıyla örnek uzayı hakkında yeterli bilginin (information) elde edilememesine neden olabilir.

ID3 algoritması; ağacın kök düğümünde bütün eğitim veri seti ile başlar. Bu veri setini bölmek için bir nitelik seçilir. Niteliğin her bir değeri için bir dal oluşturulur. Dallar ile belirlenmiş nitelik değerlerine sahip alt örneklem kümeleri, yeni oluşturulan çocuk düğüme taşınır. Karar ağacında yaprağa giden her yol, bir sınıflandırma kuralı sunar (Kantardzic, 2011). C4.5 algoritması; ID3 algoritmasının daha genişletilmiş bir

versiyonudur. Bu algoritma hem kategorik, hem de sürekli değişkenleri işleyebilir (Akpınar, 2014):

ID3 ve C4.5 algoritmaları, bir düğümdeki örneklere uygulanan bilgi entropisi (information entropy) ölçümünün minimize edilmesine dayalı çalışırlar.

Bir karar ağacında en iyi bölen niteliğin seçilmesi için çeşitli yöntemler bulunmaktadır. Kategorik değişkenler için Entropi Endeksi, Gini Endeksi, Sınıflandırma Hatası Endeksi, Twoing veya Ordered Twoing; sürekli değişkenler için ise En Küçük Kareler Sapması bu yöntemlerden en başlıca bilinenleridir (Akpınar, 2014). ID3 algoritması için en iyi bölen niteliğin seçiminde entropi ve bilgi kazancı (information gain) yöntemleri kullanılırken, C4.5 algoritmasında kazanç oranı (gain ratio) yöntemi uygulanmaktadır. Denklem (2.2)'de bir T kümesi için entropi denklemi verilmektedir.

$$Info(T) = -\sum_{i=1}^{k}\left(\left(freq(C_i,T)\Big/_{|T|}\right).log_2\left(freq(C_i,T)\Big/_{|T|}\right)\right) \tag{2.2}$$

Bu T kümesi; X niteliğindeki n farklı değeri için n parçaya ayrılabilir. Bu durumda beklenen bilgi ihtiyacı, alt kümelerdeki entropilerin ağırlıklı toplamı şeklinde hesaplanabilir (Denklem 2.3) (Kantardzic, 2011).

$$Info_x(T) = \sum_{i=1}^{n}\left(\left(|T_i|\Big/_{|T|}\right).Info(T_i)\right) \tag{2.3}$$

n parçaya ayrılmış T kümesi için bilgi kazancı Denklem (1.2) ve Denklem (1.3) farkı ile elde edilir (Denklem 1.4).

$$Gain(X) = Info(T) - Info_x(T) \tag{2.4}$$

Bu sayede kazanç kriteri $Gain(X)$'i maksimize eden niteliği belirlemiş olacaktır.

Kazanç kriteri, sıkıştırılmış bir ağaç yapısı içerisinde iyi sonuçlar üretmektedir. Ancak pek çok değeri olan niteliklere doğru kuvvetli bir sapma gösterebilmesinden dolayı

performansı olumsuz etkilemektedir Bu olumsuzluğun giderilmesinde bir çeşit normalizasyon uygulanmaktadır (Kantardzic, 2011).

$$Split - info(T) = -\sum_{i=1}^{n}\left(\left({|T_i|}/{|T|}\right).log_2\left({|T_i|}/{|T|}\right)\right) \quad (2.5)$$

Denklem (1.5) T kümesini n adet T_i alt kümesine bölünmesiyle üretilen potansiyel bilgiyi sunmaktadır. Bu durumda yeni kazanç ölçütü Denklem (2.6)'da verilmektedir.

$$Gain - ratio(X) = gain(X)/Split - info(X) \quad (2.6)$$

Bu yeni kazanç ölçütü bölünme sonucu elde edilen üretilen bilginin oranını ifade etmektedir. $Gain - ratio(X)$ kriteri, daha önce sunulan kazanç kriterine göre daha sağlam ve daha tutarlı bir seçim yapılmasına olanak tanır.

3.4.5.3. Naive Bayes Sınıflandırıcı

Naive Bayes Sınıflandırıcı (Naive Bayes Classifier); herhangi bir sınıflandırma probleminde olduğu gibi, birden fazla özelliğe sahip bir vektörü kullanarak, öğrenme gerçekleştirmek ve bu öğrenmeye dayanarak yeni gelen verileri hedef nitelik bazında doğru bir şekilde sınıflandırmaktadır. Yöntem; herbir niteliğin sonuca olan etkilerinin olasılık olarak hesaplanmasına dayanmaktadır. Bayes teoreminden oluşturulmuş olan Naive Bayes algoritması, sınıflandırma probleminde sıklıkla kullanılan etkili ve basit bir yöntemdir (Soria ve diğ. , 2011)

Bir naive bayes sınıflandırıcı aşağıdaki şekilde çalışır (Han ve diğ., 2012):

- Üzerinde çalışılacak olan veri setinden elde edilen eğitim veri seti olarak D belirlenmiş olsun. $\boldsymbol{X}$, $\boldsymbol{n}$ boyutlu bir nitelik vektörü olmak üzere $X = (x_1, x_2, \dots, x_n)$ şeklinde tanımlansın.
- C m boyutlu bir sınıflar kümesi olmak üzere $C = (C_1, C_2, \dots, C_m)$ şeklinde tanımlansın.

- Sınıflandırıcı, X'in sonsal olasığı (posterior probability) en yüksek olan sınıfa ait olduğunu tahminleyecektir. Diğer bir deyişle naive bayes sınıflandırıcı X'in C_i sınıflarından hangisine ait olduğunu ancak ve ancak $P(C_i|X) > P(C_j|X)$ $(1 \leq j \leq m, j \neq i)$ olacak şekilde tahmin eder.
- Böylece $P(C_i|X)$ maksimize edilmiş olur. Bayes teoremine ait temel denklem düzenlenerek Denklem (1.7) elde edilmiş olur:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (2.7)$$

- P(X) tüm sınıflar için sabitken, sadece $P(X|C_i)P(C_i)$ maksimize edilmesi gerekmektedir. Sınıf öncelik olasılıklarının (class prior probabilities) bilinmediği durumlarda sınıf olasılıkları eşit kabul edilir. Bundan dolayı da $P(X|C_i)$ maksimize edilir (aksi taktirde $P(X|C_i)P(C_i)$ maksimize edilirdi).
- Pek çok niteliğe sahip veri kümelerinde $P(X|C_i)$ olasılığını hesaplamak oldukça zahmetli olabilir. Bu tür durumlarda değerin hesaplanmasının daha kolaylaştırma adına sınıf koşullu bağımsızlık için naive varsayımı (naive assumption of class-conditional independence) bulunmaktadır. Bu varsayım niteliklerin değerlerinin birbirlerinden koşullu bağımsız olduğu savına dayanmaktadır. Buna bağlı olarak $P(X|C_i)$ olasılık değeri her bir x_j için Denklem 2.8 kullanılarak hesaplanır.

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \qquad (2.8)$$

$$= P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

X'in sınıf etiketini tahmin edebilmek için, her C_i sınıfı için $P(X|C_i)\,P(C_i)$ hesaplanmalıdır.

3.4.5.4. Logistik Regresyon

Logistik Regresyon; x'ler ile ikili değerlere sahip (dichotomous) bağımsız değişkenler arasındaki ilişkiyi tanımlayan matematiksel modelleme yaklaşımıdır (Gail, Krickeberg, Samet, Tsiatis, & Wong, 2010). Analizde temel amaç, bağımlı ve bağımsız değişkenler

arasındaki ilişkiyi en az değişken ile ne iyi uyuma sahip olacak biçimde tanımlayan, kabul edilebilir nitelikte bir model kurmaktır (Atasoy, 2001; Çokluk, 2010).

Logistik Regresyon analizinde kategorik olan hedef niteliğin diğer nitelikler, bağımsız değişkenler ile arasındaki ilişki tespit edilmeye çalışılır. Bu tespitte hedef nitelik iki kategoriden oluşuyor ise ikili logistik regresyon, ikiden fazla ve sıralı kategorili ise sıralı logistik regresyon, ikiden fazla ve sırasız kategoriye sahip ise de çok kategorili logistik regresyon analiz uygulanmalıdır.

X veri serindeki her x_i için y_i çıktı değerleri 1 yada 0 değerini almaktadır. Burada y_i=1 değeri pozitif sınıfa aitliği, y_i=0 ise negatif sınıfa aitliği göstermektedir. Sınıflandırma prensibi gereği kurulacak olan logistik regresyon modelinin girdilere bakarak çıktıların pozitif ya da negatif sınıflardan hangisine ait olduğunun belirlemesi beklenmektedir. En genel haliyle logistik regresyon modeli Denklem 2.9'daki haliyle sunulmaktadır:

$$log\frac{p(x)}{1-p(x)} = \beta_0 + x.\beta \qquad (2.9)$$

Buradaki p değerinin hesaplanmasında Denklem 2.10'dan faydalanılmaktadır:

$$p(x;b,w) = \frac{e^{\beta_0+x.\beta}}{1+\ e^{\beta_0+x.\beta}} = \frac{1}{1+\ e-(\beta_0+x.\beta)} \qquad (2.10)$$

Logistik regresyon analizinde olasılık, üstünlük (odds) ve üstünlüğün logaritmasına dayanmaktadır. Üstünlük; bir olayın olma olasılığının o olayın olmamam olasılığına bölümü ile tanımlanmaktadır (Mertler ve Vannatta, 2005)

X olayının olma olasılığı p(x) ile gösterilsin, olmama olasılığı 1-p(x) ile ifade edilecektir. Bu durumda üstünlük değeri Denklem 2.10 gösterilen şekilde hesaplanacaktır.

$$\text{Üstünlük} = \frac{p(x)}{1 - p(x)} \qquad (2.11)$$

Logistik regresyon analizinde kurulacak modelin kestiriminde en çok olabilirlik analizi (maximum likelihood) yöntemi kullanılmaktadır. Yöntem logistik regresyon sapmaların karesini (en küçük kareler) "en az" yapmak yerine, bir olayın olma olasılığını "en çok" yapmak ile ilgilenmektedir (Hair ve diğ., 2006; Çokluk, 2010).

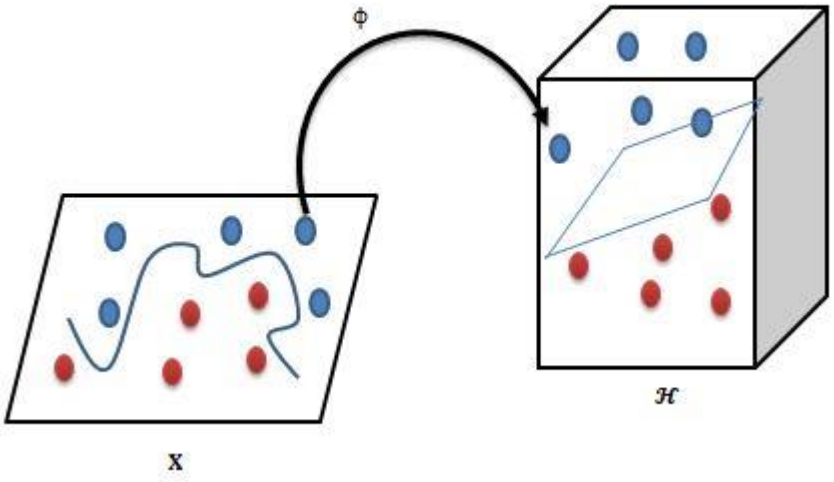
3.4.5.5. Destek Vektör Makineleri (Support Vector Machine-SVM)

Destek Vektör Makineleri (DVM); sınıflandırma teknikleri arasında popülaritesi gittikçe artan, güçlü istatistik teoriler üzerine inşaa edilmiş bir makine öğrenmesi yöntemi olarak kabul edilmektedir (Çomak ve Güner, 2011).

DVM; n boyutlu x girdi vektörünün bir k-boyutlu özellik uzayına $\phi(x)$ gibi bir eşleme fonksiyonunun kullanılarak taşınması ve doğrusal ilişkilerin bu yeni uzayda aranması felsefesine dayanmaktadır. Bu arama işleminde, öklidyen uzay yerine Hilbert Uzayı'nda çalışılmakta, benzerliklerin yakalanması adına iç çarpım fonksiyonları kullanılmaktadır. x, özellik vektörü Denklem 2.12'de verilen şekliyle tanımlansın.

$$\phi: \mathrm{X} \longrightarrow \mathcal{H} \ (\phi: x \longrightarrow \mathbf{x}, \boldsymbol{x} \in \boldsymbol{X} \subseteq R^d, \mathcal{H} \subseteq R^m \ ve \ d \ll m) \qquad (2.12)$$

$\phi(x)$; eşleme/dönüşüm (mapping) ya da öz nitelik vektörü olarak tanımlanmaktadır. $\phi(x)$'e ait her $\phi(x)_i$ baz fonksiyonları olarak ifade edilen ve doğrusal olmayan fonksiyonlardır (Bishop, 1995).



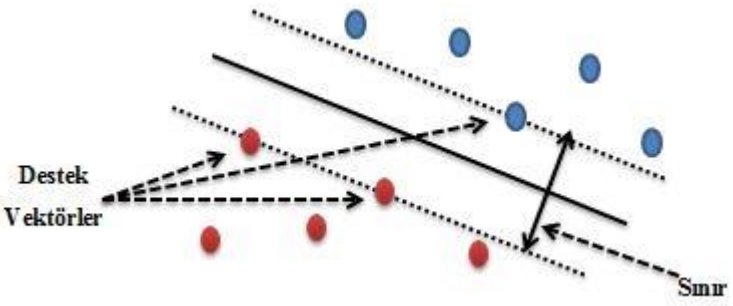
Şekil 3.7: Yüksek Boyutlarda Doğrusal Ayrılma (Gürsoy, 2013).

$\mathcal{H}$ uzayında benzerlikleri yakalamak için kernel (çekirdek) fonksiyonları kullanılır. Bu fonksiyonlar iç çarpımlar üzerinden tanımlanmaktadır (Denklem 2.13).

$$\mathrm{k}(x, x') = < \boldsymbol{x}, \mathbf{x}' > = < \phi(x), \phi(\mathbf{x}') > \qquad (2.13)$$

Kerneller sayesinde doğrusal olmayan algoritmaların doğrusal algoritmalara indirgenmesi mümkün olabilmektedir. Ancak bu indirgeme işlemi, girdi uzayı ile doğrusal bir ilişkisi bulunmayan öznitelik uzayının oluşturulması ile mümkündür.

Bir DVM sınıflandırıcı, sınıflama işlemini, verilen sınıflar arasında maksimum aralığa sahip ayırıcı bir düzlem bularak gerçekleştirmektedir (Şekil 3.8).



Şekil 3.8: DVM Sınıflandırıcımın Yapısı.

Veri Madenciliği algoritmalarının matematiksel olarak incelenmesi, kullanılan aracın/paketin aslında nasıl bir iç isleyişe sahip olduğunun, "öğrenmesinin" nasıl gerçekleştiğinin anlaşılması açısından oldukça önemlidir. Ancak algoritmanın/algoritmaların anlaşılması, çözümü araştırılan veri madenciliği problemi için yeterli değildir. Veri madenciliği, problemden bağımsız bir süreç olarak alındığında belli bir standarda sahip olması gerektiği açıktır. Bölüm 2.3.'de veri madenciliğinin bir süreç olarak ilerleyişinin nasıl olması gerektiğine yer verilmiştir.

3.4.6.Modelleri Değerlendirme ve Seçme

Veri madenciliği yöntemlerinde temel amacın, en genel biçimde gizlenmiş örüntülerin tespit edilmesi, karara destek olacak yorumların/kuralların üretilmesi olabileceği başlangıç bölümlerinde anlatılmıştı. Ancak işe yarar bilginin/kuralın/yorumun çıkarılması tekniklerin uygulanması ile sınırlandırılmamalıdır. Özellikle tahmine dayalı çalışmalarda, yetersiz sınıflamalar, yetersiz/yetkinlik içermeyen sonuçlara/kararlara götürebilir (Kumar, 2011). Böylesine istenmeyen bir durumun önüne geçilmesi, kurgulanan modellerin çeşitli performans değerlendirme yöntemleri kullanılarak, kıyaslama ölçütleri üzerinden karşılaştırılmaları ile mümkün olabilecektir. Veri madenciliği algoritmaları yardımıyla kurulan modellerde, eğitim ve test veri kümelerinin ayrımına, büyüklüğüne bağlı olarak çeşitli performans değerlendirme yöntemleri bulunmaktadır. Bu tez çalışmasında performans değerlendirme yöntemlerinden tabakalı holdout ve k-kat çapraz geçerleme yöntemleri seçilmiştir.

Hold-out yönteminde veri seti farklı oranlarda eğitim ve test kümelerine ayrılmaktadır (Blum ve diğ., 1999). Bu ayrıma bağlı olarak doğruluk, hata, tanısal üstünlük değeri (DOR), F-ölçütü gibi kriterlerle performans değerleri karşılaştırılmaktadır. k-Kat çapraz geçerleme yönteminde ise veri seti k eşit (olabildiği kadar eşit/yakın) parçaya ayrılmakta ve k-1 tanesi eğitim olacak şekilde analize dahil edilmektedir (Kohavi, 1995). Bölüm 4'de kullanılan kıyaslama ölçütleri ve elde edilen bulgular detaylı bir biçimde sunulmaktadır.

3.4.7.Seçilen Modeli Uygulama

Bölüm 4.7'de performans değerlendirme ölçütlerinin kıyaslanması sonucu kurulan modellerin başarısına ve hangi modelin seçildiğine yer verilmektedir.

3.4.8.Sonucu Karar/Eylem/Yeni Girdi Haline Dönüştürme

Bu adımın CRISP-EDM içindeki hem nihai hem de başlangıç adımı niteliği taşıdığından daha önce bahsedilmişti. Seçilen model ile elde edilen sonuçların; karara destek amaçlı bir yapı olarak nasıl kurgulandığı, Bölüm 4.8'de anlatılmaktadır. Elde edilen sonuçların yeni bir veri madenciliği çalışmasının başlangıcı olabilmesi yönündeki görüşlere ise Bölüm 5'te yer verilmiştir.

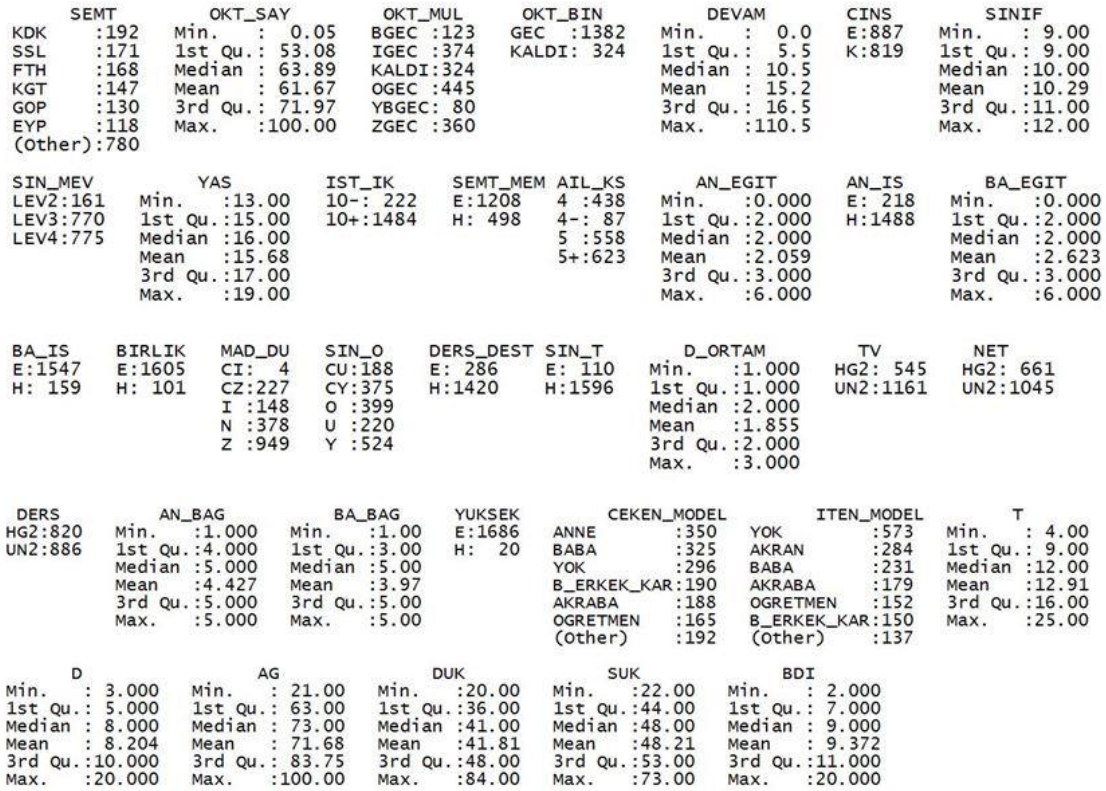
4. BULGULAR

Tezin bu bölümünde *Logistik Regresyon Analizi, ID3 ve C4.5 Karar Ağacı Algoritmaları, Naive Bayes Sınıflandırıcı ve DVM Sınıflandırıcı*'nın derlenen verilerden oluşturulan veri setine uygulanması ve bu analizlerin sonucunda elde edilen bulgulara yer verilmiştir. Her analiz için analizde kullanılan veri seti, tahmin edici nitelikler, hedef nitelik, kullanılan R paketleri, performans değerlendirme yöntemine ilişkin temel bilgiler sunulmuştur.

EDM veri seti üzerinden yapılan analizlerde, Bölüm 3.4.5'de anlatılmış olan sınıflandırıcılar ve algoritmalar kullanılmıştır. Kullanılan sınıflandırıcının başarısında önemli bir etkiye sahip eğitim ve kümlerinin ayrımında tabakalı Hold-out yöntemi ve k-kat çapraz geçerleme yöntemleri kullanılmıştır. Bu yöntemlerle oluşturulan her eğitim-test kümeleri için performans değerleri hesaplanmıştır. Bu değerler ile oluşturulan ortalama doğruluk, duyarlılık, belirleyicilik, kesinlik, negatif öngörü, f-ölçütü, pozitifi olabilirlik oranı gibi maksimum değer alması; otalama hata, yanlış pozitif oranı, yanlış negatif oranı, negatif olabilirlik oranı gibi minimum değer alması beklenmektedir.

4.1. VERİ SETİNİN TANIMAYA YÖNELİK TEMEL İSTATİSTİK BULGULAR

Veri setinin, tez çalışması kapsamında kullanılacak sınıflandırma teknikleri uygulanmadan önce, temel istatistiki yöntemlerle genel görünümüne bakılmalıdır. Bu bölümde RStudio ile oluşturulan EDM veri setine ait özet tablolar ve grafikler verilmektedir. Şekil 4.1'de EDM veri setinin en temel istatistiksel değerleri RStudio'da hesaplanmış ekran görüntüsü verilmektedir.



```
     SEMT         OKT_SAY          OKT_MUL      OKT_BIN        DEVAM          CINS        SINIF
KDK     :192   Min.    :  0.05   BGEC :123   GEC  :1382   Min.    :  0.0   E:887   Min.    : 9.00
SSL     :171   1st Qu.: 53.08   IGEC :374   KALDI: 324   1st Qu.:  5.5   K:819   1st Qu.: 9.00
FTH     :168   Median : 63.89   KALDI:324                Median : 10.5           Median :10.00
KGT     :147   Mean   : 61.67   OGEC :445                Mean   : 15.2           Mean   :10.29
GOP     :130   3rd Qu.: 71.97   YBGEC: 80                3rd Qu.: 16.5           3rd Qu.:11.00
EYP     :118   Max.   :100.00   ZGEC :360                Max.   :110.5           Max.   :12.00
(Other):780

SIN_MEV         YAS          IST_IK     SEMT_MEM AIL_KS      AN_EGIT        AN_IS       BA_EGIT
LEV2:161   Min.   :13.00   10-: 222   E:1208    4 :438   Min.   :0.000   E: 218   Min.   :0.000
LEV3:770   1st Qu.:15.00   10+:1484   H: 498    4-: 87   1st Qu.:2.000   H:1488   1st Qu.:2.000
LEV4:775   Median :16.00               5 :558    Median :2.000                     Median :2.000
           Mean   :15.68               5+:623    Mean   :2.059                     Mean   :2.623
           3rd Qu.:17.00               3rd Qu.:3.000                              3rd Qu.:3.000
           Max.   :19.00               Max.   :6.000                              Max.   :6.000

BA_IS    BIRLIK    MAD_DU    SIN_O    DERS_DEST SIN_T      D_ORTAM         TV         NET
E:1547   E:1605   CI:  4    CU:188   E: 286    E: 110   Min.   :1.000   HG2: 545   HG2: 661
H: 159   H: 101   CZ:227    CY:375   H:1420    H:1596   1st Qu.:1.000   UN2:1161   UN2:1045
                  I :148    O :399                      Median :2.000
                  N :378    U :220                      Mean   :1.855
                  Z :949    Y :524                      3rd Qu.:2.000
                                                        Max.   :3.000

 DERS         AN_BAG          BA_BAG      YUKSEK        CEKEN_MODEL          ITEN_MODEL          T
HG2:820   Min.   :1.000   Min.   :1.00   E:1686   ANNE         :350   YOK          :573   Min.   : 4.00
UN2:886   1st Qu.:4.000   1st Qu.:3.00   H:  20   BABA         :325   AKRAN        :284   1st Qu.: 9.00
          Median :5.000   Median :5.00            YOK          :296   BABA         :231   Median :12.00
          Mean   :4.427   Mean   :3.97            B_ERKEK_KAR:190   AKRABA       :179   Mean   :12.91
          3rd Qu.:5.000   3rd Qu.:5.00            AKRABA       :188   OGRETMEN     :152   3rd Qu.:16.00
          Max.   :5.000   Max.   :5.00            OGRETMEN     :165   B_ERKEK_KAR:150   Max.   :25.00
                                                  (Other)      :192   (Other)      :137

       D                AG              DUK             SUK             BDI
Min.   : 3.000   Min.   : 21.00   Min.   :20.00   Min.   :22.00   Min.   : 2.000
1st Qu.: 5.000   1st Qu.: 63.00   1st Qu.:36.00   1st Qu.:44.00   1st Qu.: 7.000
Median : 8.000   Median : 73.00   Median :41.00   Median :48.00   Median : 9.000
Mean   : 8.204   Mean   : 71.68   Mean   :41.81   Mean   :48.21   Mean   : 9.372
3rd Qu.:10.000   3rd Qu.: 83.75   3rd Qu.:48.00   3rd Qu.:53.00   3rd Qu.:11.000
Max.   :20.000   Max.   :100.00   Max.   :84.00   Max.   :73.00   Max.   :20.000
```

Şekil 4.1: EDM Veri Setinin İlk Halinin Nitelikler Bazında Özeti.

Şekil 4.1 incelendiğinde; SINIF, AN_EGIT, BA_EGIT, D_ORTAM, AN_BAG, BA_BAG niteliklerinin analizin yapılacağı program tarafından sayısal değerler olarak algılandığı görülmektedir. Dolayısıyla sınıflandırma teknikleri uygulanmadan önce kategorik olarak değerlendirilmesi gereken bu niteliklerin düzeltilmesi gerekmektedir. EDM verisetindeki kayıp değer analizleri CRISP-EDM sürecinin *Verileri derleme ve ön inceleme aşamasın*da yapılmaktadır.

EDM veri setinde yer alan OKT_MUL ve OKT_BIN, OKT_SAY niteliğine bağlı olarak üretilmiştir. OKT_MUL ve OKT_BIN öğrencinin yıl sonu başarı ortalamasının aralıklar cinsinden kategorilere ayrılmış halini içermektedir.

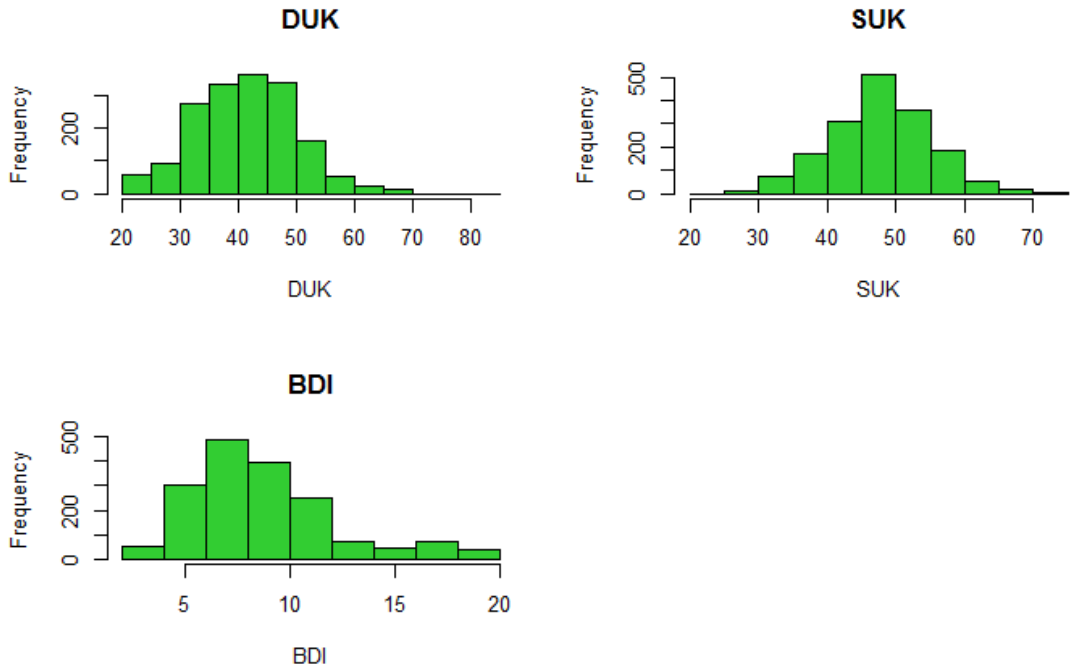
x, öğrencinin yılsonu ağırlıklı not ortalaması olmak üzere Tablo 4.1'de bu değerin niteliklerdeki karşılıkları verilmektedir.

Tablo 4.1: Okul Başarı Puanın Veri Setindek Karşılıkları.

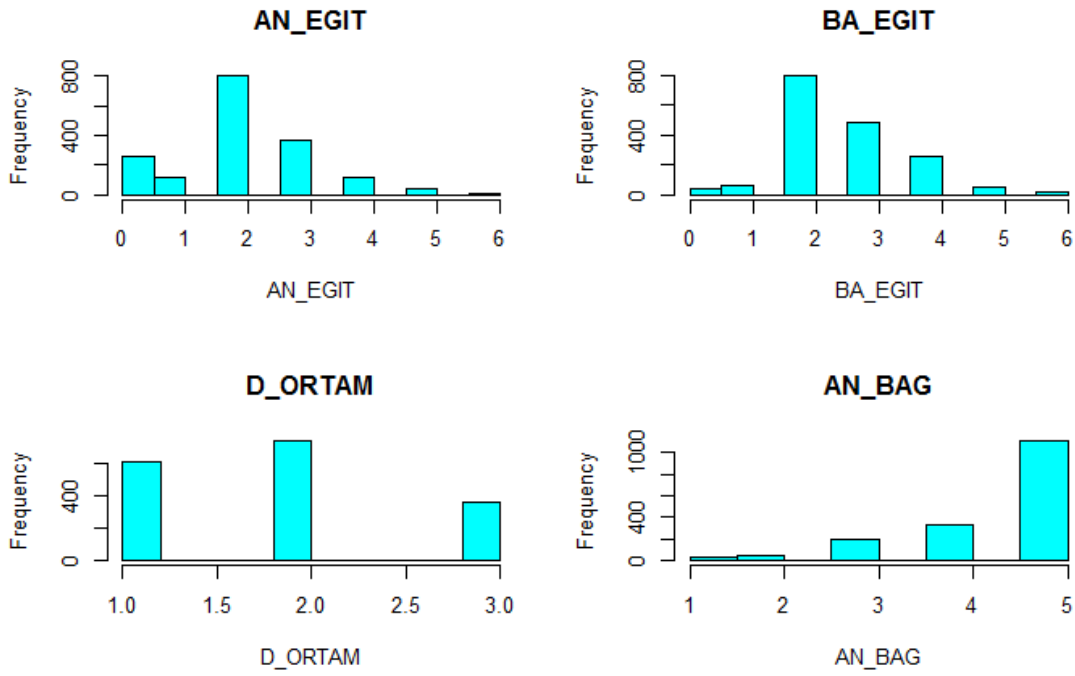
Puan Aralığı	OKT_MUL Niteliğindeki Kategori	OKT_BIN Niteliğindeki Kategori
$0 \leq x < 50$	KALDI	KALDI
$50 \leq x < 60$	ZGEC	GEC
$60 \leq x < 70$	OGEC	GEC
$70 \leq x < 80$	IGEC	GEC
$80 \leq x < 90$	BGEC	GEC
$90 \leq x < 100$	YBGEC	GEC

Analizlerde öğrenci yılsonu başarı ortalamalarının OKT_BIN niteliğine çevrilmiş halleri kullanılmıştır. Altı kategoriye sahip olam OKT_MUL niteliği yapılan analizlerde performans olarak düşük değerler ürettiğinden bu şekildeki kategorizasyon tercih edilmemiştir.

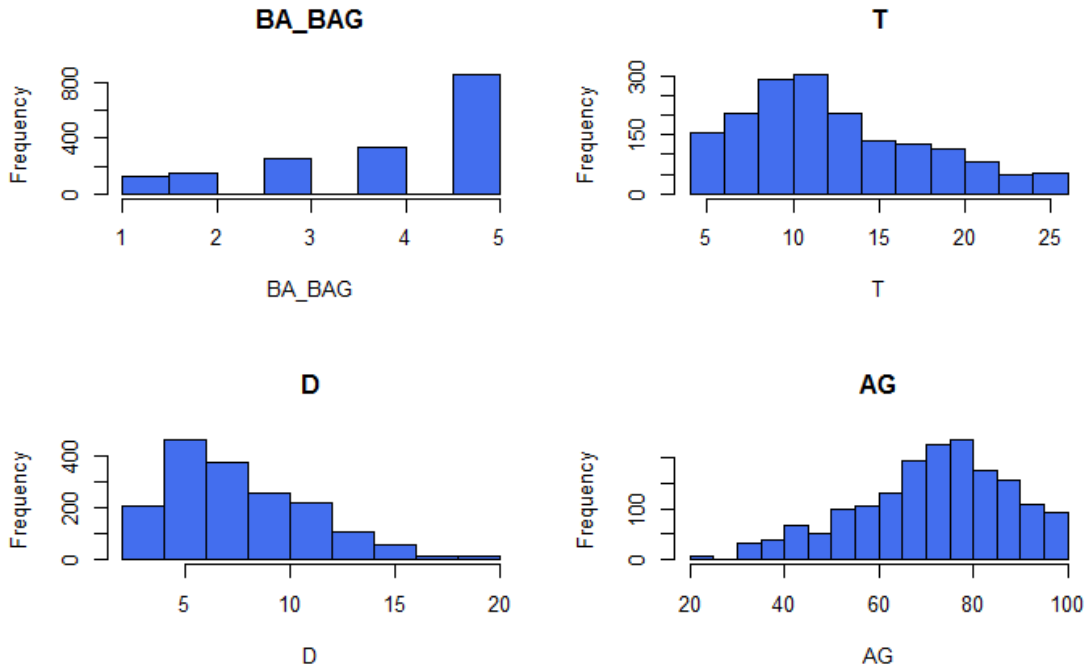
Verinin grafikler biçiminden sunulması; anlaşılması ve yapılacak analizlerde dikkat edilmesi gereken noktalar açısından kolaylık sağlayacaktır. Şekil 4.2, 4.3, 4.4, 4.5, 4.6, 4.7 ve 4.8'de EDM veri setinde yer alan niteliklerin dağılımı çeşitli karşılaştırmalar ve farklı grafik türleri üzerinden sunulmaktadır.



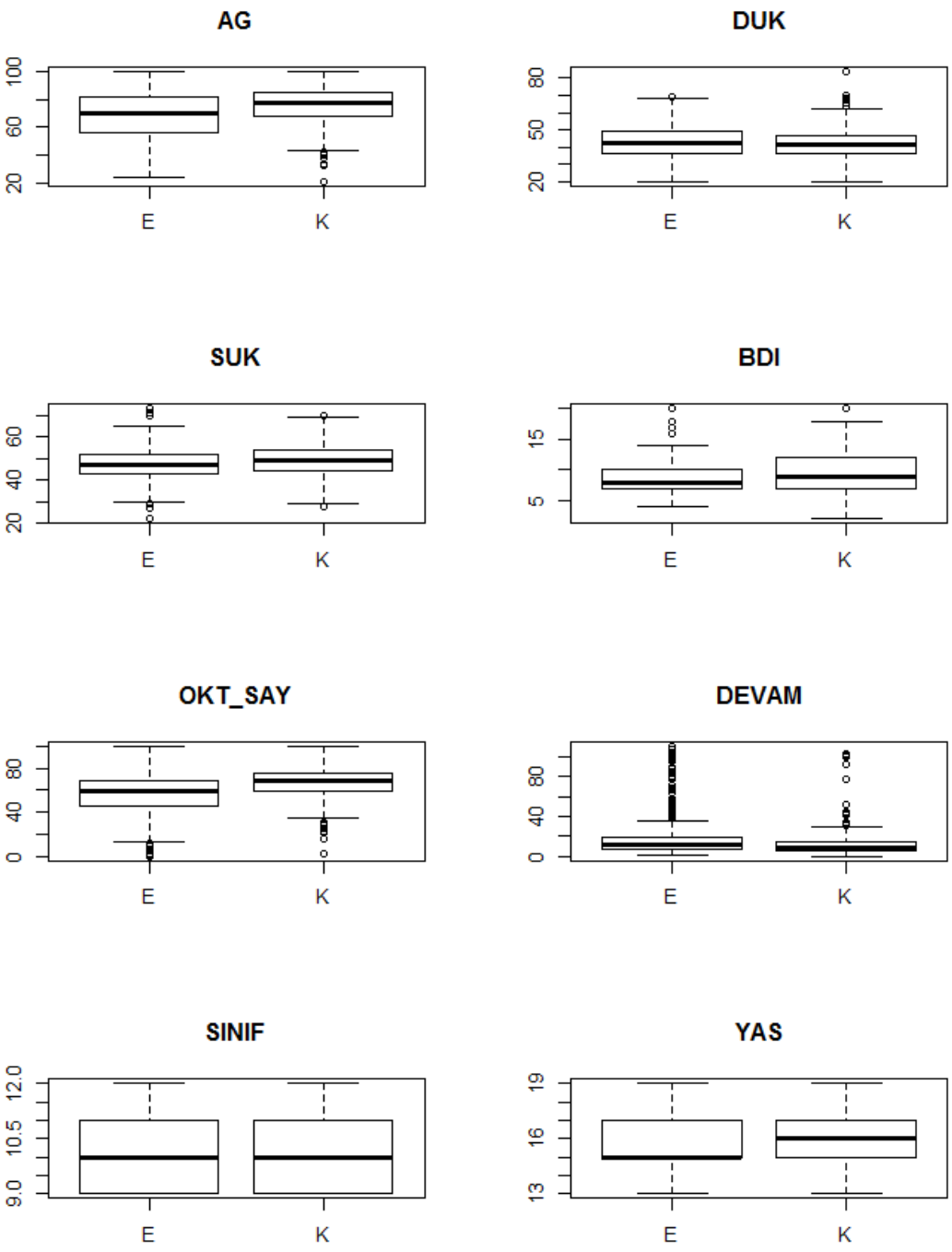
Şekil 4.2: DUK, SUK ve BDI Niteliklerinin Dağılmını Gösteren Histogram Grafiği.



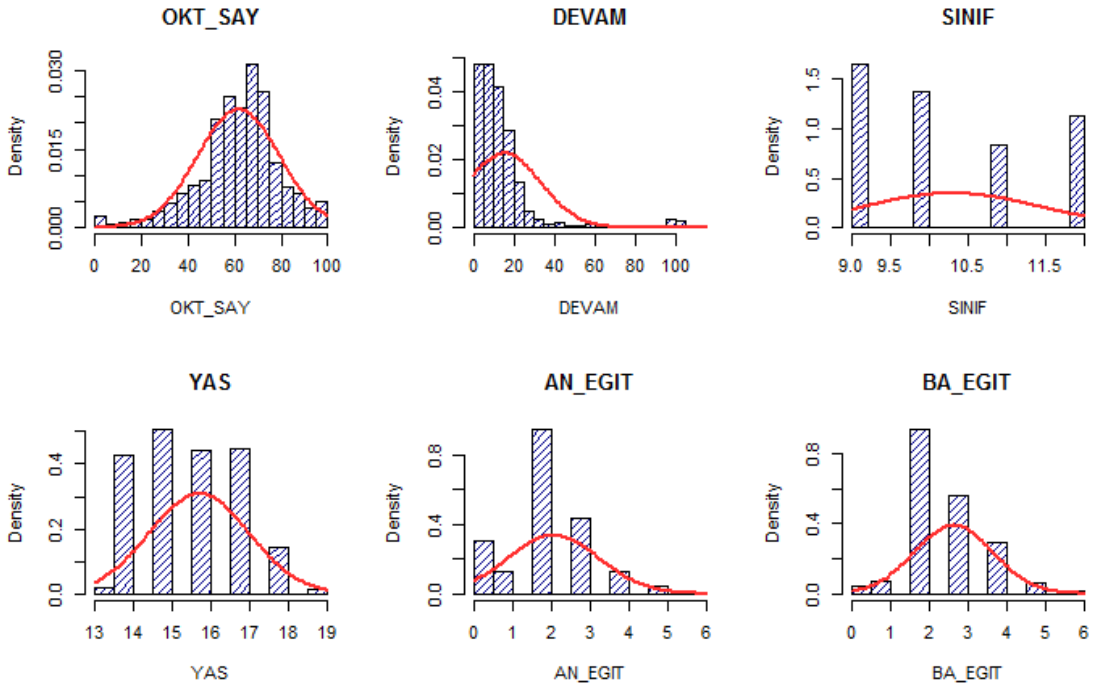
Şekil 4.3: Annenin Eğitim Durumu (AB_EGIT), Babanın Eğitim Durumu (BA_EGT), Öğrencinin Ders Çalışma Ortamını (D_ORTAM), Anne İle Algılanan Bağlanma Düzeyini Gösteren (AN_BAG) Histogram Grafikleri.



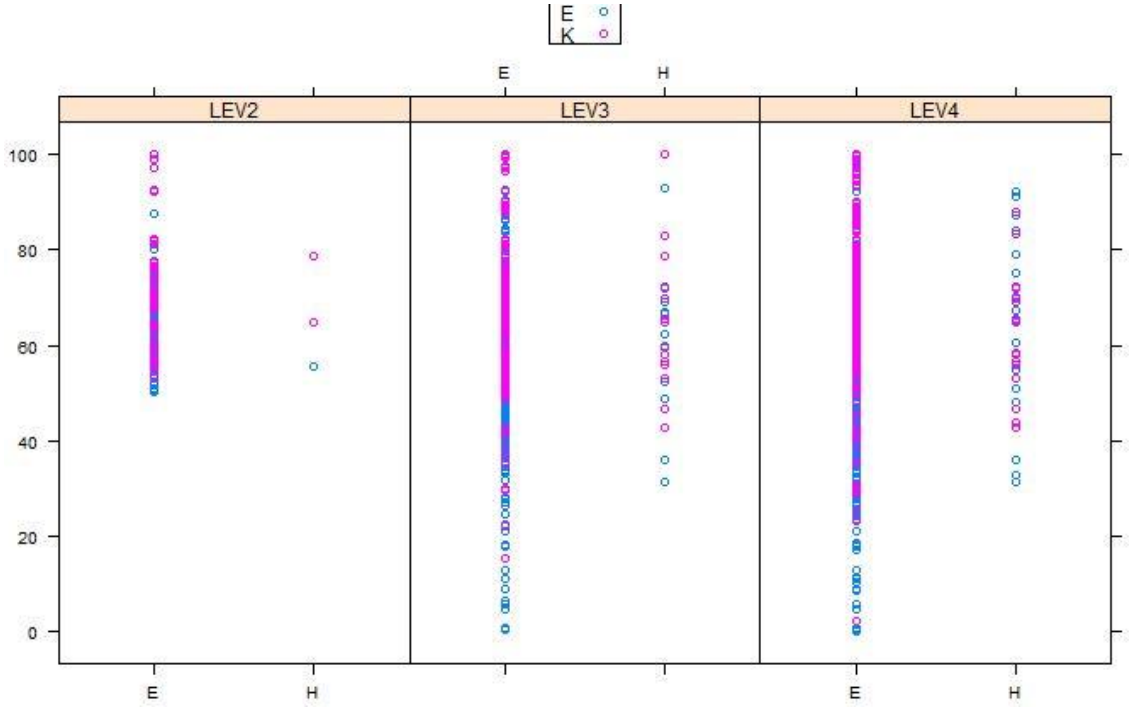
Şekil 4.4: Baba İle Algılanan Bağlanma Durumunu (BA_BAG), Tükenmşlik Puanlarının (T), Duyarsızlaşma Düzeyinin (D) Ve Akademik Güdülenme Puanlarının (AG) Dağılımını Gösteren Histogram Grafiği.



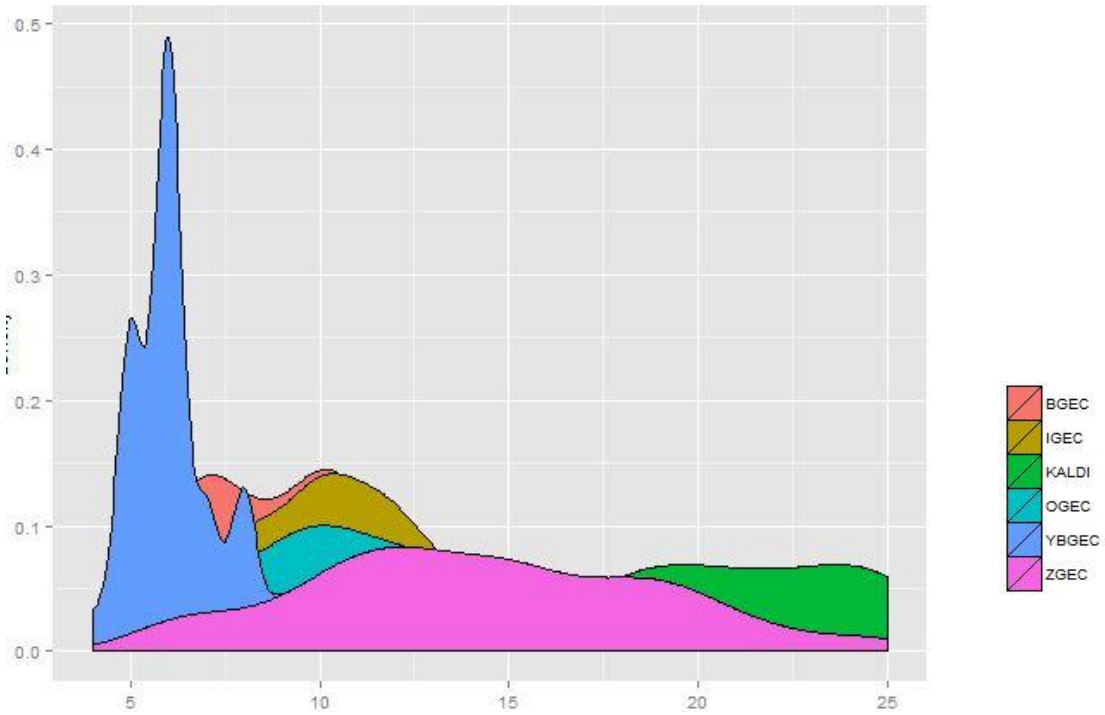
Şekil 4.5: EDM Veri Setindeki Niteliklerin Cinsiyete Göre Dağılımını Gösteren Boxplot Grafikleri.



Şekil 4.6: EDM Veri Setindeki Niteliklerin Histogram Yoğunluk Dağılımını Gösteren Grafikler.



Şekil 4.7: Okul Başarısı Ve Anne-Babanın Birliktelik Durumlarının Sınıf Mevcuduna Göre Gruplayan, Cinsiyete Göre De Dağılımını Gösteren Xyplot Grafiği.



Şekil 4.8: Tükenmişlik Puanının (T) Okul Başarısı Kategorilerine (OKT_MUL) Göre Yoğunluk Grafiği.

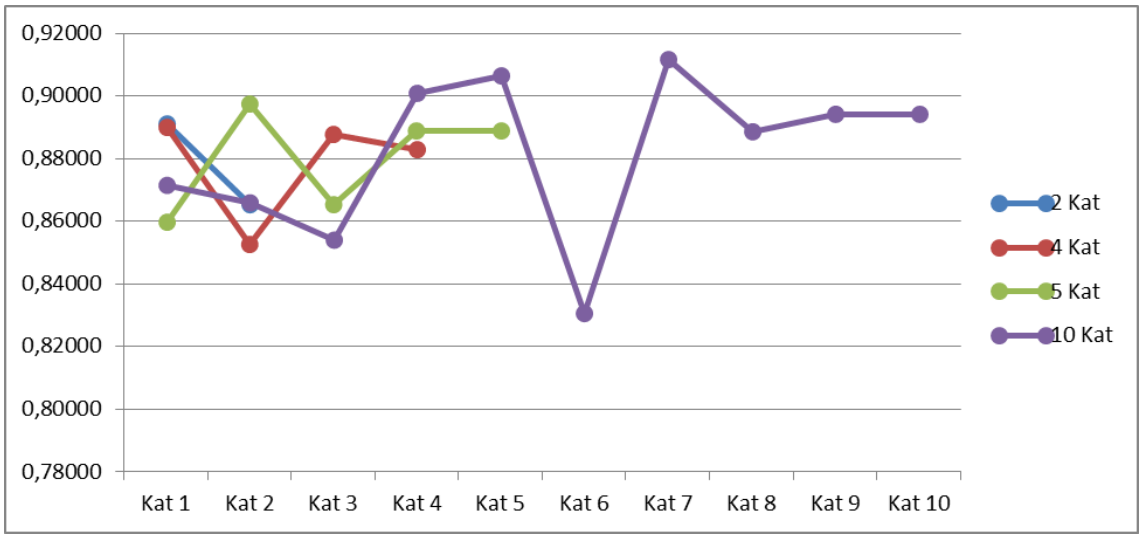
4.2. K-EN YAKIN KOMŞU ALGORTİMASINDAN ELDE EDİLEN BULGULAR

EDM veri seti üzerinde k-En Yakın Komşu (k-NN) algoritması ile yapılan analizlerde, OKT_BIN hedef niteliğinin tahmininde, Class (Ripley ve Venables, 2015), Caret (Kuhn, ve diğ., 2015) ve TunePareto (Müssel ve diğ., 2012) paketleri kullanılmıştır. Sınıflandırıcının performansı; 2-kat, 4-kat, 5-kat ve 10-kat çapraz geçerleme ve e %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında hold-out oluşturulan eğitim ve test kümeleri üzerinden çeşitli performans kriterleri kullanılarak hesaplanmıştır. Ek 1, Ek 2 ve Ek 3'de performans tabloları ayrıntılı şekilde sunulmaktadır.

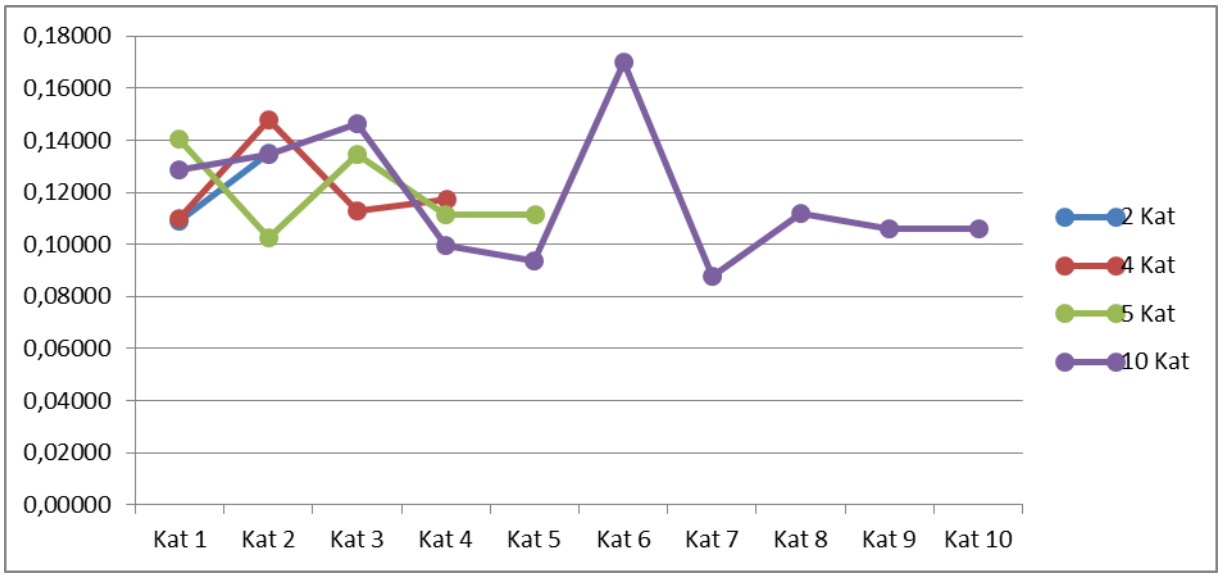
k-NN algoritmasında performansın seçilmesinde en iyi değeri veren k değerinin hesaplanması gerekmektedir. Bu parametrenin eldesi için aynı model üzerinde farklı k değerleri uygulanarak, elde edilen doğruluk değerleri kıyaslanmalıdır. Enas ve Choi (1986)'ye göre en iyi k parametresinin bulunması, veri setinin boyutuna, gözlem

değerlerinin sayısına bağlıdır. k-NN sınıflandırıcının komşuluk sayısının büyük seçildiği durumlarda daha stabil bir yapı göstermekte ve daha başarılı sonuçlar elde edilmektedir (Yang & Liu, 1999). Bu tez çalışmasında eldeki veri seti 1706 gözlem değerine sahiptir. Bu durumda 41 değeri için k'nın performansının istenilen düzeyde olması beklenmektedir. Ancak bu değerlendirme biçimi kesin olmayıp, kabul edilebilir düzeyde olduğundan k'nın en ideal performanslı değerinin bulunması için 1 ile 82 aralığına bakılmıştır. En uygun değerin seçimi için bu aralıktaki arasındaki tüm k değerlerine göre doğruluk değerleri hesaplanmıştır. Bu hesaplama sonucunda k=12, 13 ve 14 değerleri için maksimum doğruluk değerinin elde edildiği görülmüştür (Ek 4 ve Ek 5). Bu nedenle bundan sonraki performans karşılaştırması için kullanılacak yöntemlerde k=12 değerine göre hesaplamalar yapılacaktır.

Şekil 4.9'da k-NN algoritmasının EDM veri setine 2-kat, 4-kat, 5-kat, 10-kat çapraz geçerleme kullanılması sonucu elde edilen performans değerlerinden doğruluk (ACC) değerine ilişkin grafik verilmektedir. Doğruluğa bağlı olarak hesaplanabilen hata değeri, Şekil 4.10'da sunulmaktadır.



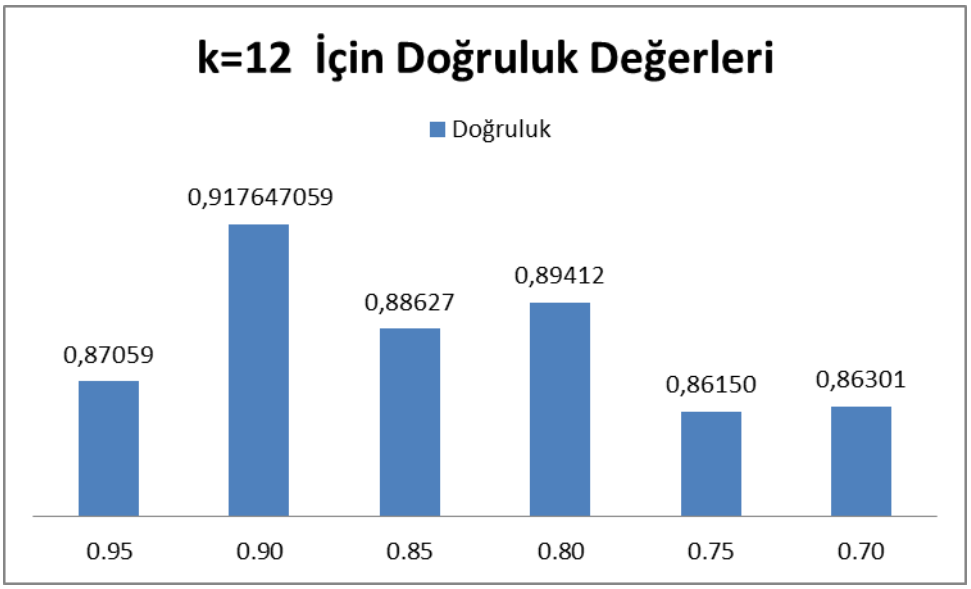
Şekil 4.9: k-NN Algoritmasından k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Grafiği.



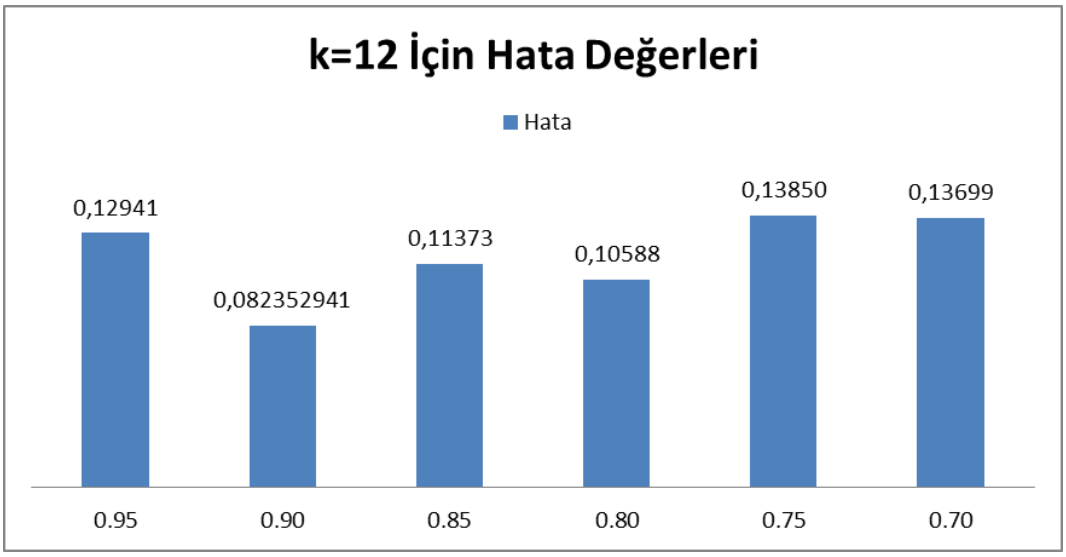
Şekil 4.10: k-NN Algoritması k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Grafiği.

Şekil 4.10 ve 4.11 incelendiğinde k kat çapraz geçerlemelerle elde edilen her test ve eğitim küme ikilisi için doğruluk ve hata değerlerinin farklılaştığı görülmektedir. Doğruluk değerleri, 0,82 ile 0,92 arasında değişim göstermektedir. Modelin performansının doğru bir biçimde irdelenebilmesi için her kat için ortalama performans değerleri oluşturulması gerekmektedir. Ek 2'de k=12 değeri için ortalama performans değerleri sunulmaktadır.

k-kat çapraz geçerleme dışında, EDM veri setinde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında tabakalı hold-out örnekleme uygulanmıştır. k-NN algoritmasından tabakalı hold-out örnekleme yöntemiyle elde edilen doğruluk grafiği Şekil 4.11'de, hata grafiği ise Şekil 4.12'de verilmektedir.



Şekil 4.11: k-NN Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Grafiği.



Şekil 4.12: k-NN Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Grafiği.

Şekil 4.12 ve 4.13 incelendiğinde test kümesine giren veri seti arttıkça doğruluk değerinde düşüşler/hata değerinde yükselişler görülmektedir. Bu durum literatürde modelin sağlıklı bir biçimde seçilmesinde, algortimanın öğrenmesinin sağlanması için yeterli sayıda gözlem değerine sahip eğitim kümesi oluşturulması gerektiğine işaret

etmektedir. Holdout yöntemiyle yapılan analizlerde doğruluk değeri yaklaşık olarak 0,86 ile 0,91 arasında değişim göstermektedir.

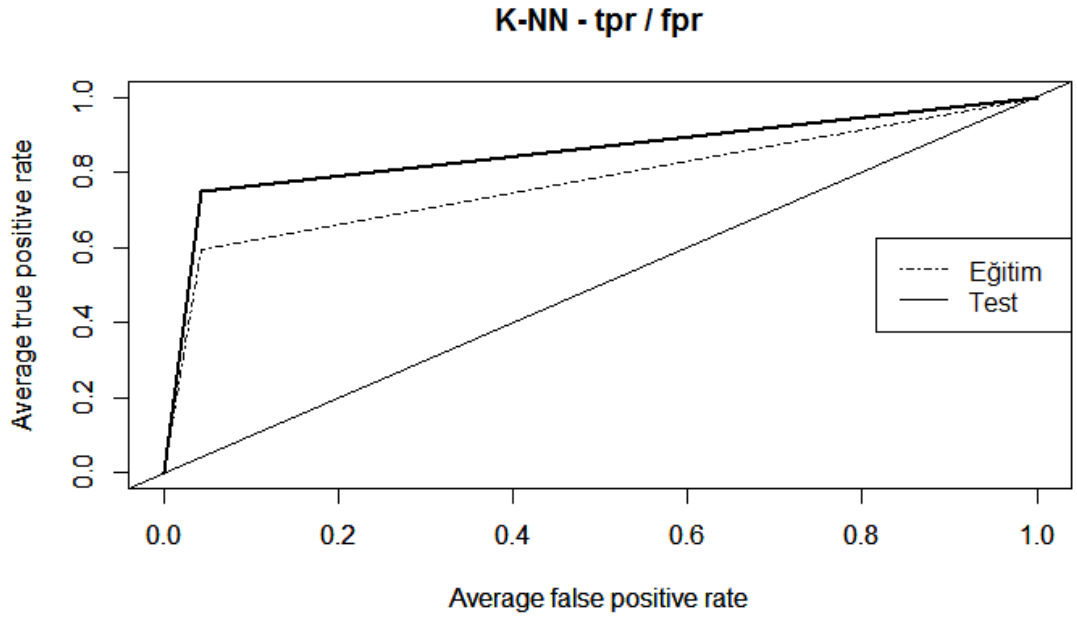
Yapılan analizlerde sadece doğruluk ve/veya hataya bakılarak performans konusunda net bir karar varmak yeterli olmayacaktır. Bu nedenle "keskinlik", "duyarlılık" değerlerine bakılması, hatta bu değerler baz alınarak hesaplanan "F ölçümü"nün incelenmesi ile daha kapsamlı bir performans değerlendirmesi yapılabilmesi mümkün olacaktır. Bu performans kriterlerinin dışında, pozitif olabilirlik oranı, negatif olabilirlik oranı, bunların yardımıyla hesaplanan ve daha kapsamlı olan tanısal üstünlük değerinin de incelenmesi kıyaslanabilirlik açısından önemlidir. Çalışmada bundan sonraki algoritmalarda, doğruluk ve hata değişim grafikleri sunulacak, ancak nihai performans değerlendirme kriteri olarak tanısal üstünlük değeri (DOR) ve F değeri baz alınacaktır.

k-NN algoritmasında, k parametresinin seçimi dışında, hangi uzaklık ölçüsünün kullanılacağı da doğru bir biçimde belirlenmelidir. Bu çalışmada veri setindeki nümerik nitelikler üzerinde Öklid uzaklık formulune bağlı olarak hesaplama yapılmıştır. k-NN algoritması ile yapılan analizlerde, ortalama doğruluk değeri, holdout (dışarıda bırak) yöntemi için 0,88219; k-kat çapraz geçerleme de ise 0,87941 değerleri elde edilmiştir (Ek 2).

Hold out yönteminde en yüksek doğruluk derecesi 0,917647059 değeri ile %90/%10 ayırımda elde edilmiştir. k-Kat çapraz geçerleme yönteminde ise en yüksek doğruluk derecesi 0,88163 ise 10-kat ayırımda bulunmuştur. 10-kat ayırımın 7. katında doğruluk en yüksek değerine 0,91176 ulaşmıştır. (Ek 1 ve Ek 3).

k-NN algoritmasında k= 12 için k-kat çapraz geçerlemede ortalama F ölçütü 0,92761, ortalama tanısal üstünlük değeri ise 157,21994'dir. Hold out yönteminde ise ortalama F ölçütü 0,92958, ortalama tanısal üstünlük değeri ise 180,92707'dir.

F değeri ve tanısal üstünlük değeri dışında, model performansının ölçümünde ROC eğrilerinin çizimi ve eğri altında kalan alanın (AUC- area under curve) değerlendirilmesi de kullanılmaktadır. Şekil 4.13'de k-NN algoritması ile elde edilen modelin %90/%10 ayırımına ilişkin ROC eğrisi verilmektedir.



Şekil 4.13: k-NN Algoritması %90/%10 Tabakalı Holdout ROC Eğrisi.

%90/%10 ayırımla elde edilen ROC eğrisinde test kümesi için eğrinin altında kalan alan (AUC) 0,85326 olarak bulunmuştur (Ek 3). Diğer ayırımlar için AUC değeri hesaplanmış ve ortalama AUC değeri 0,75765 olarak elde edilmiştir(Ek 3).

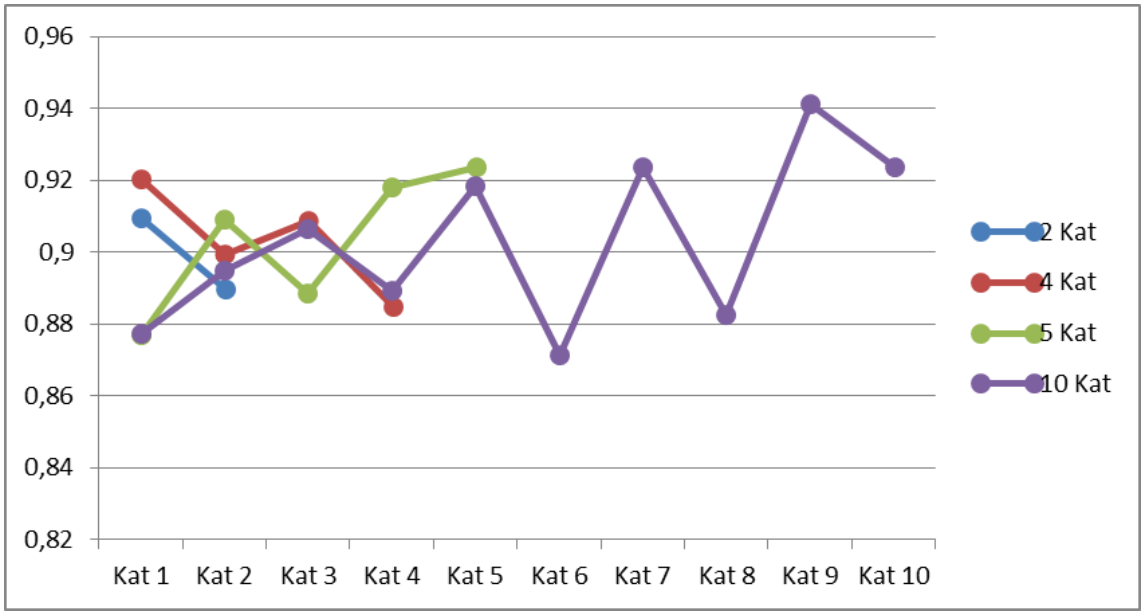
k-NN algoritmasında oluşturulan modelin k=12 değeri için yapılan hesaplamaların hepsinde 0,75'in üzerinde performans değeri elde edilmiştir. Bu durum EDM veri seti için k-NN algoritmasında kurgulanan modelin kullanılabileceği ve karşılaştırması yapılabilecek modeller arasına alınabileceğini göstermektedir.

4.3. NAIVE BAYES SINIFLANDIRICISINDAN ELDE EDİLEN BULGULAR

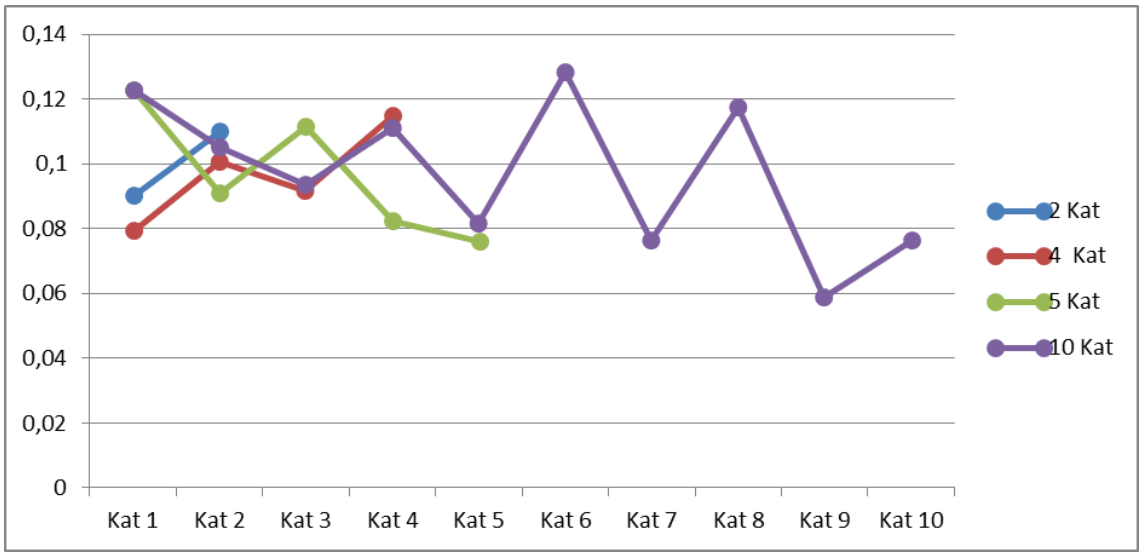
EDM veri seti üzerinde Naive Bayes sınıflandırıcı ile yapılan analizlerde, OKT_BIN hedef niteliğinin tahmininde, e1071 (Meyer ve diğ., 2014), Caret (Kuhn ve diğ., 2015) ve TunePareto (Müssel ve diğ., 2012) paketleri kullanılmıştır. Sınıflandırıcının performansı; 2-kat, 4-kat, 5-kat ve 10-kat çapraz geçerleme ve %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında hold-out yöntemleri ile oluşturulan eğitim ve test kümeleri üzerinden çeşitli performans kriterleri kullanılarak

hesaplanmıştır. Ek 6, Ek 7 ve Ek 8'de performans tabloları ayrıntılı şekilde sunulmaktadır.

Şekil 4.15'de Naive Bayes algoritmasının EDM veri setine 2-kat, 4-kat, 5-kat, 10-kat çapraz geçerleme kullanılması sonucu elde edilen performans değerlerinden doğruluk (ACC) değerine ilişkin grafik verilmektedir. Doğruluğa bağlı olarak hesaplanabilen hata değeri ise Şekil 4.14'de sunulmaktadır.



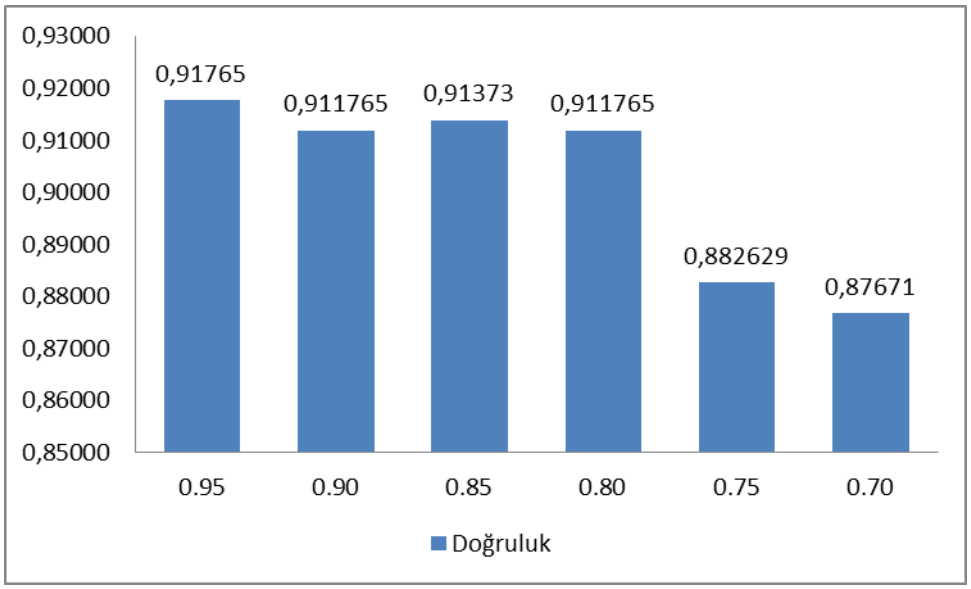
Şekil 4.14: Naive Bayes Algoritmasından k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Değerleri Grafiği.



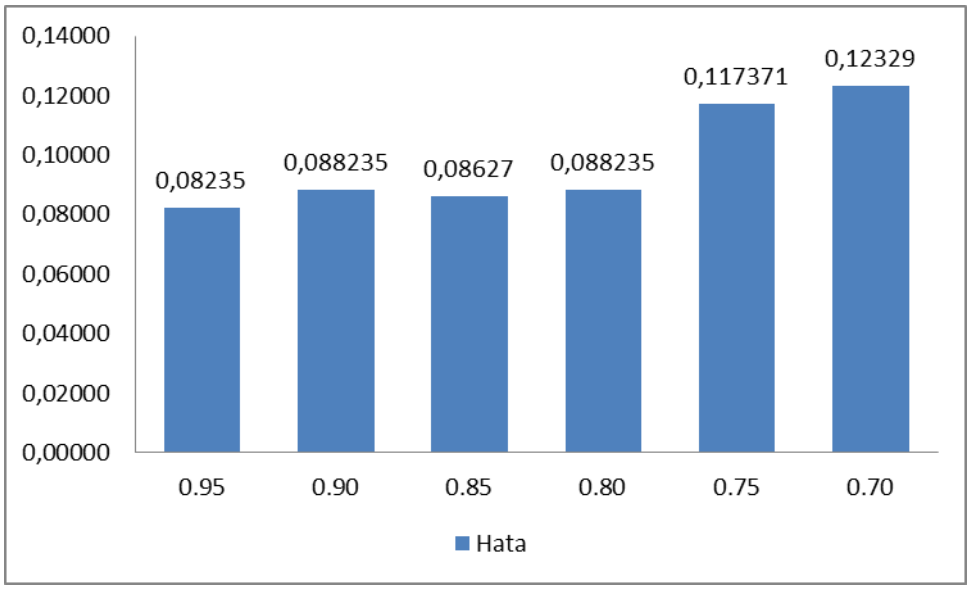
Şekil 4.15: Naive Bayes Algoritması k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Değerleri Grafiği.

Şekil 4.14 ve 4.15 incelendiğinde k kat çapraz geçerlemelerle elde edilen her test ve eğitim küme ikilisi için doğruluk ve hata değerlerinin farklılaştığı görülmektedir. Doğruluk değerleri, 0,86 ile 0,96 arasında değişim göstermektedir. Modelin performansının doğru bir biçimde irdelenebilmesi için her kat için ortalama performans değerleri oluşturulması gerekmektedir. Ek 7'de ortalama performans değerleri sunulmaktadır.

k-kat çapraz geçerleme dışında, EDM veri setinde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında tabakalı hold-out örnekleme uygulanmıştır. Naive Bayes algoritmasından tabakalı hold-out örnekleme yöntemiyle elde edilen ortalama doğruluk grafiği Şekil 4.16'de, hata grafiği ise Şekil 4.17'da verilmektedir.



Şekil 4.16: Naive Bayes Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği.

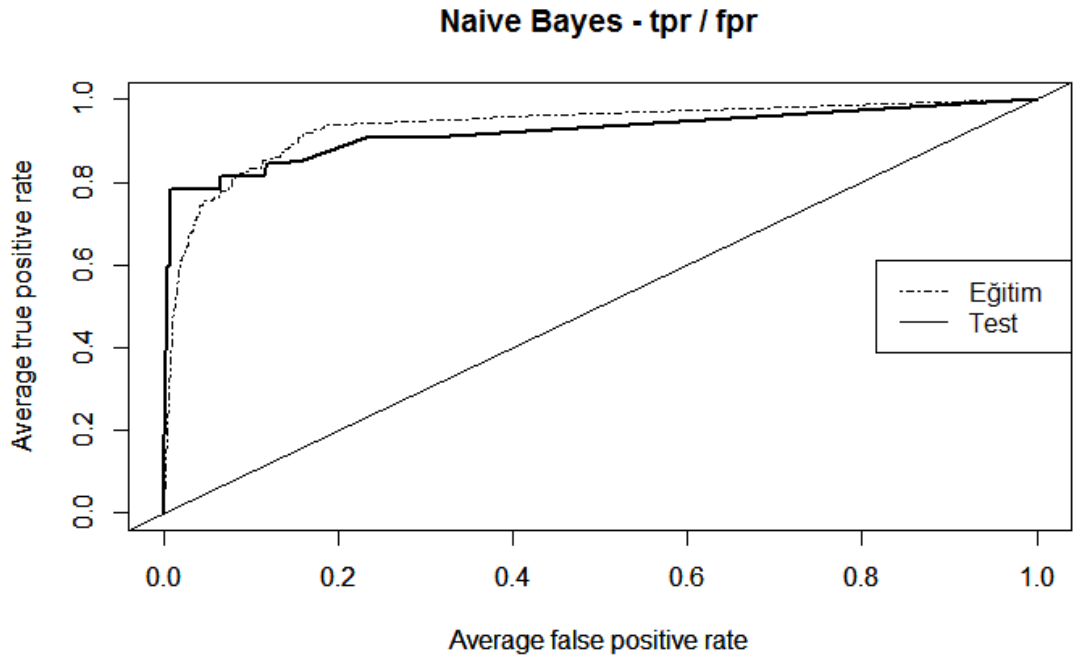


Şekil 4.17: Naive Bayes Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği.

Şekil 4.16 ve 4.17 incelendiğinde holdout yöntemiyle yapılan analizlerde doğruluk değerinin yaklaşık olarak 0,87 ile 0,92 arasında değiştiği görülmektedir. k-NN

algoritmasında en yüksek doğruluk performansı %90/%10 ayırımda elde edilirken, naive bayes sınıflandırıcıda %95/%5 ayırımda elde edilmektedir.

Naive Bayes algoritmasında k-kat çapraz geçerleme ile elde edilen sonuçlar incelendiğinde ortalama doğruluk değeri 0,90227, ortalama F değeri 0,94005 ve ortalama DOR değeri 283,63136 bulunmaktadır. Benzer şekilde holdout yönteminde doğruluk değeri 0,90237, ortalama F değeri 0,93946 ve ortalama DOR değeri 269,44636 olarak bulunmuştur (Ek 7 ve Ek 8). Tüm bu değerler ışığında, EDM veri seti üzerinde naive bayes sınıflandırıcınında iyi bir performans gösterdiğini söylemek mümkündür. Bu değerlere ek olarak, Naive Bayes sınıflandırıcı %90/%10 tabakalı holdout için ROC eğrisi çizdirilmiştir (Şekil 4.18).



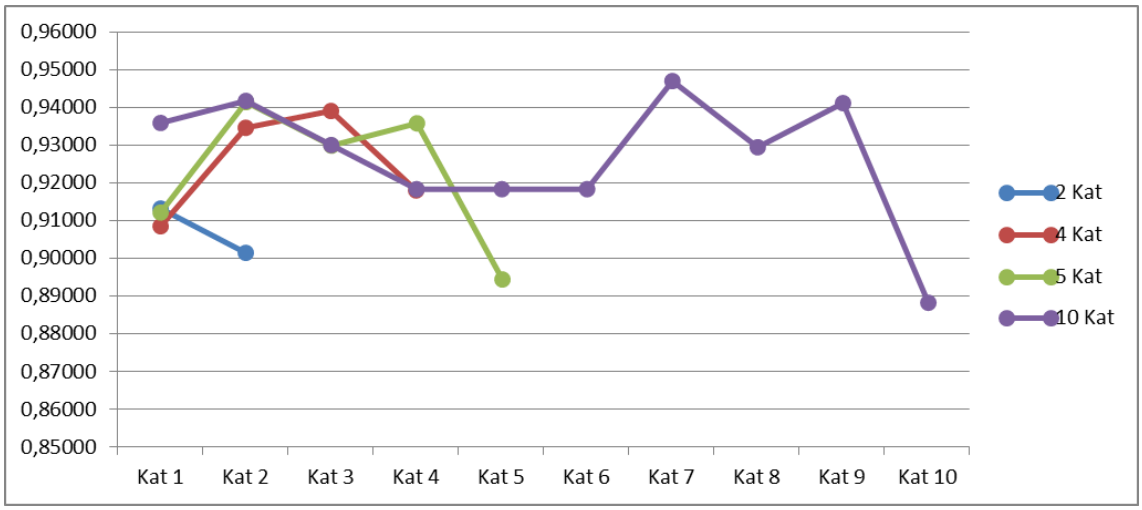
Şekil 4.18: Naive Bayes Sınıflandırıcı %90/%10 Tabakalı Holdout ROC Eğrisi.

%90/%10 ayırımla elde edilen ROC eğrisinde test kümesi için eğrinin altında kalan alan (AUC) 0,94248 olarak bulunmuştur (Ek 8). Diğer ayırımlar için AUC değeri hesaplanmış ve ortalama AUC değeri 0,94562 olarak elde edilmiştir (Ek 8).

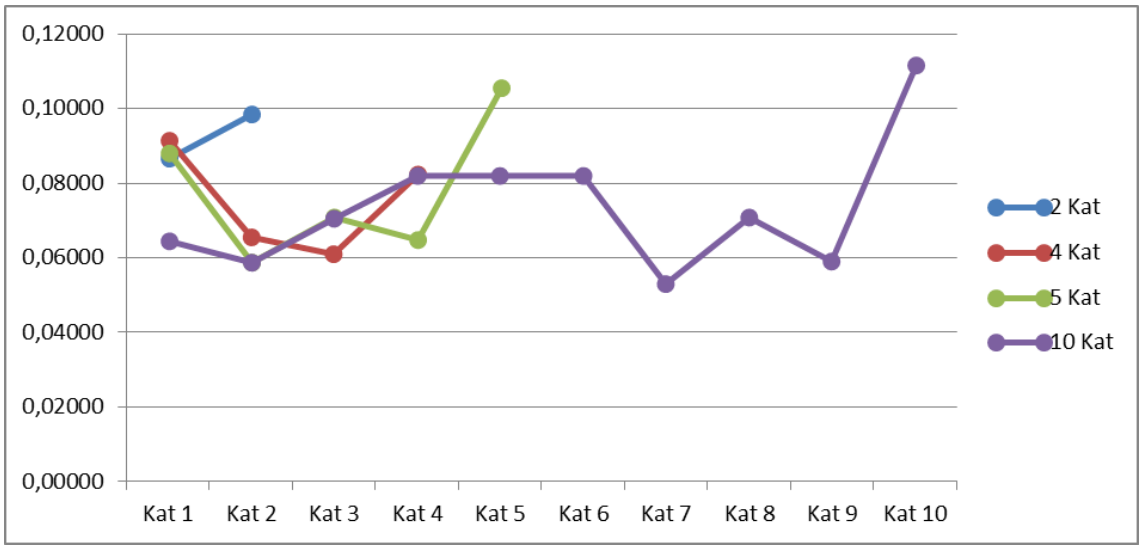
4.4. KARAR AĞACI ALGORİTMALARINDAN ELDE EDİLEN BULGULAR

EDM veri seti üzerinde Karar Ağacı algoritması ile yapılan analizlerde, OKT_BIN hedef niteliğinin tahmininde, Rpart (Therneau ve diğ., 2015), RWeka (Hornik ve diğ., 2015), Partykit, (Hothorn ve Zeileis, 2015), Caret (Kuhn ve diğ., 2015) ve TunePareto (Müssel ve diğ., 2012), FSelector (Romanski ve Kothoff, 2015) paketleri kullanılmıştır. Sınıflandırıcının performansı; 2-kat, 4-kat, 5-kat ve 10-kat çapraz geçerleme ve %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında hold-out ile oluşturulan eğitim ve test kümeleri üzerinden çeşitli performans kriterleri kullanılarak hesaplanmıştır. Ek 9, Ek 10 ve Ek 11'de performans tabloları ayrıntılı şekilde sunulmaktadır.

Şekil 4.19'da Karar Ağacı algoritmasının EDM veri setine 2-kat, 4-kat, 5-kat, 10-kat çapraz geçerleme kullanılması sonucu elde edilen performans değerlerinden doğruluk (ACC) değerine ilişkin grafik verilmektedir. Doğruluğa bağlı olarak hesaplanabilen hata değeri, Şekil 4.20'de sunulmaktadır.



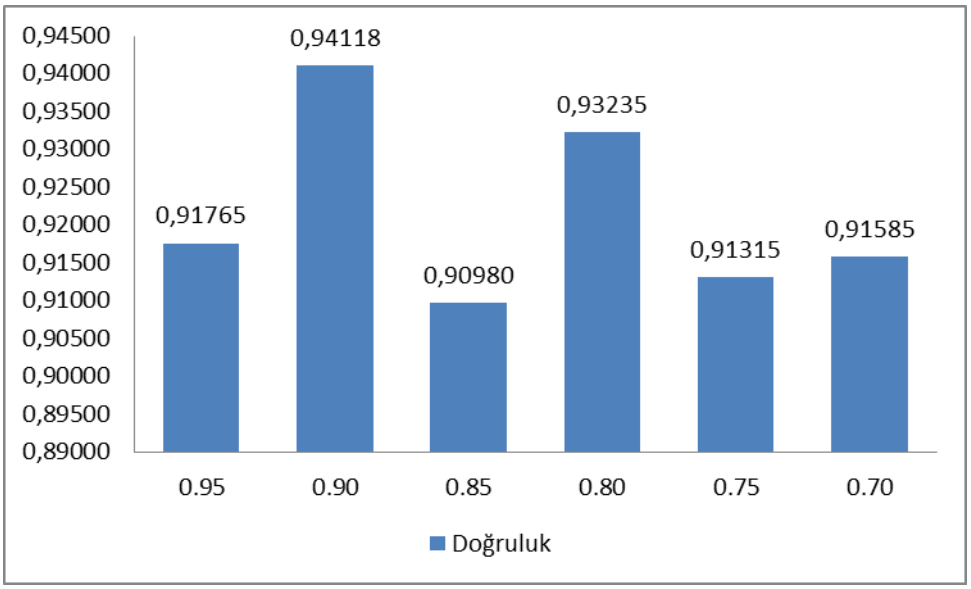
Şekil 4.19: Karar Ağacı Algoritmasından k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Değerleri Grafiği.



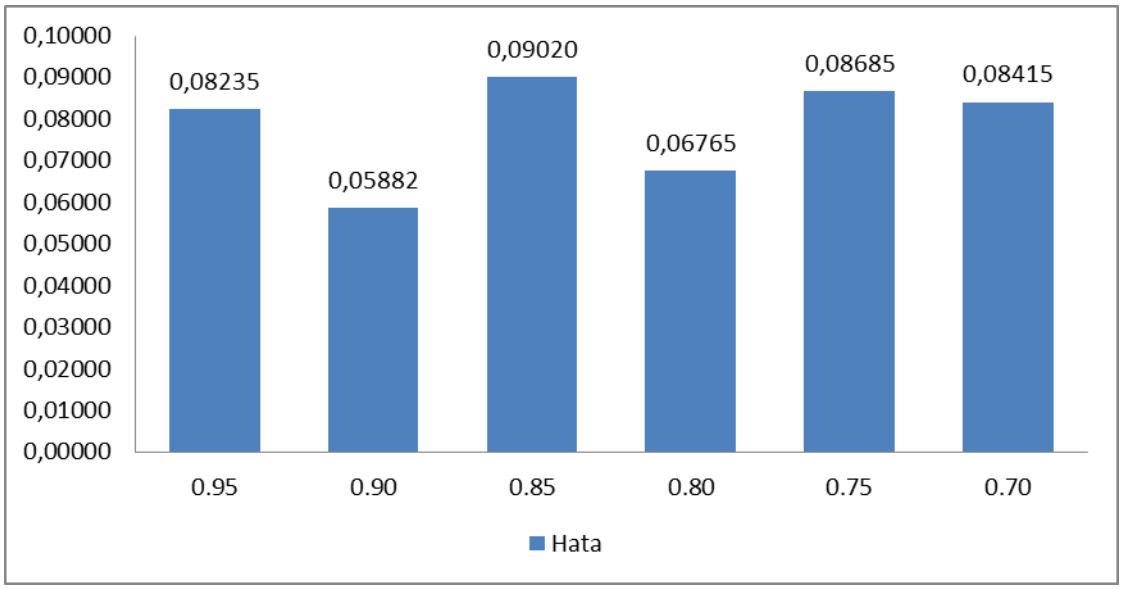
Şekil 4.20: Karar Ağacı Algoritması k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Değerleri Grafiği.

Şekil 4.19 ve 4.20 incelendiğinde k kat çapraz geçerlemelerle elde edilen her test ve eğitim küme ikilisi için doğruluk ve hata değerlerinin değişkenlik gösterdiği görülmektedir. Doğruluk değerleri, 0,88 ile 0,95 arasında değişim göstermektedir. Modelin performansının doğru bir biçimde irdelenebilmesi için her kat için ortalama performans değerleri oluşturulması gerekmektedir. Ek 10'da ortalama performans değerleri sunulmaktadır.

k-kat çapraz geçerleme dışında, EDM veri setinde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında tabakalı hold-out örnekleme uygulanmıştır. Karar Ağacı algoritmasından tabakalı hold-out örnekleme yöntemiyle elde edilen ortalama doğruluk grafiği Şekil 4.21'de, hata grafiği ise Şekil 4.22'de verilmektedir.



Şekil 4.21: Karar Ağacı Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği.

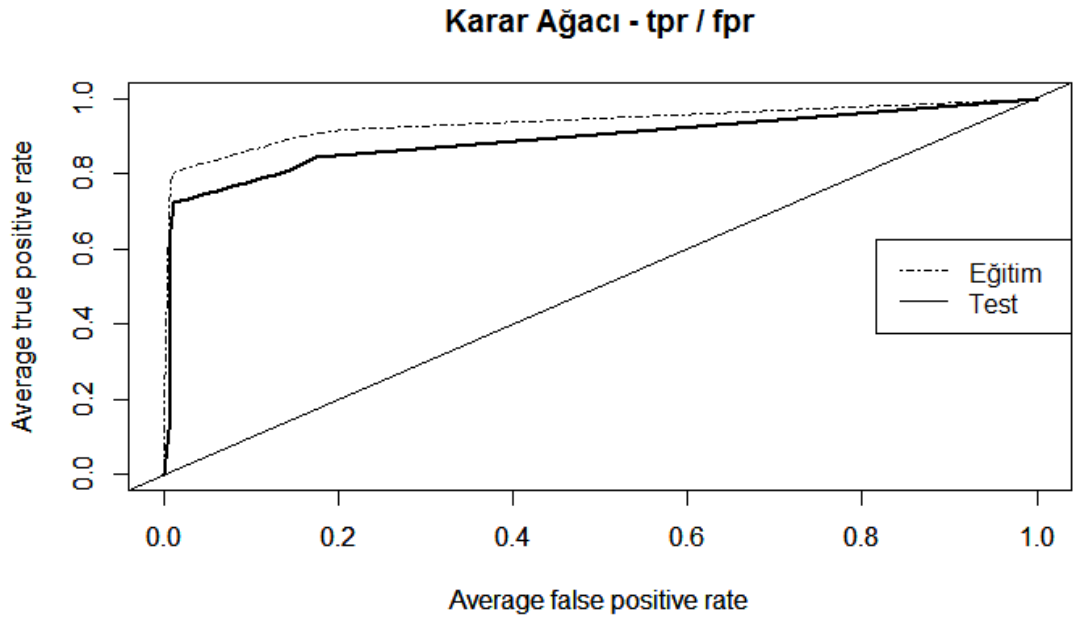


Şekil 4.22: Karar Ağacı Algoritmasından Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği.

Şekil 4.21 ve 4.22 incelendiğinde holdout yöntemiyle yapılan analizlerde doğruluk değerinin yaklaşık olarak 0,90 ile 0,95 arasında değiştiği görülmektedir. Karar ağacı

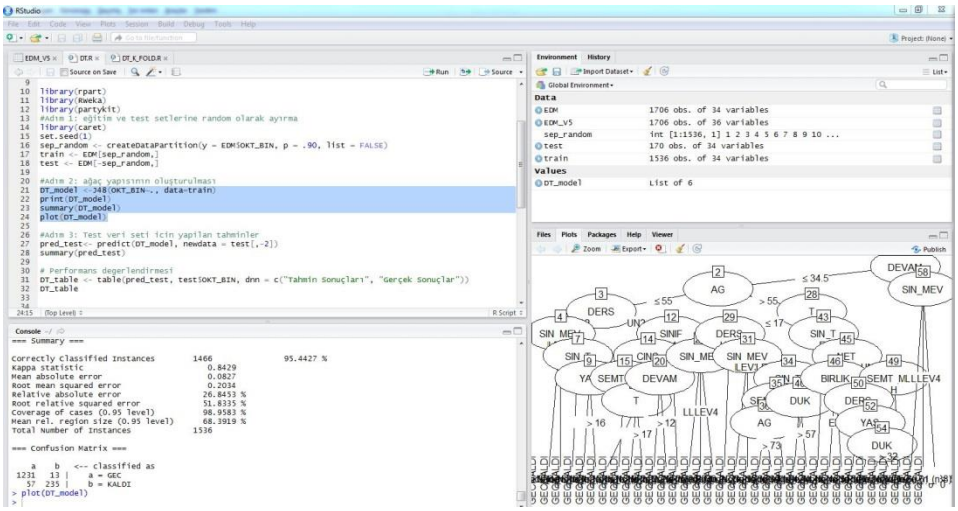
algortimasında en yüksek doğruluk değeri, k-NN algoritmasında olduğu gibi %90/%10 ayırımda elde edilmektedir.

Karar ağacı algoritmasında k-kat çapraz geçerleme ile elde edilen sonuçlar incelendiğinde ortalama doğruluk değeri 0,92043, ortalama F değeri 0,95118 ve ortalama DOR değeri 471,83494 bulunmaktadır. Benzer şekilde holdout yönteminde doğruluk değeri 0,92166, ortalama F değeri 0,95255 ve ortalama DOR değeri 59011,57905 olarak bulunmuştur (Ek 10 ve Ek 11). Tüm bu değerler ışığında, EDM veri seti üzerinde karar ağacı algortimasınında iyi bir performans gösterdiğini söylemek mümkündür. Bu değerlere ek olarak, %90/%10 tabakalı holdout için ROC eğrisi çizdirilmiştir (Şekil 4.23).

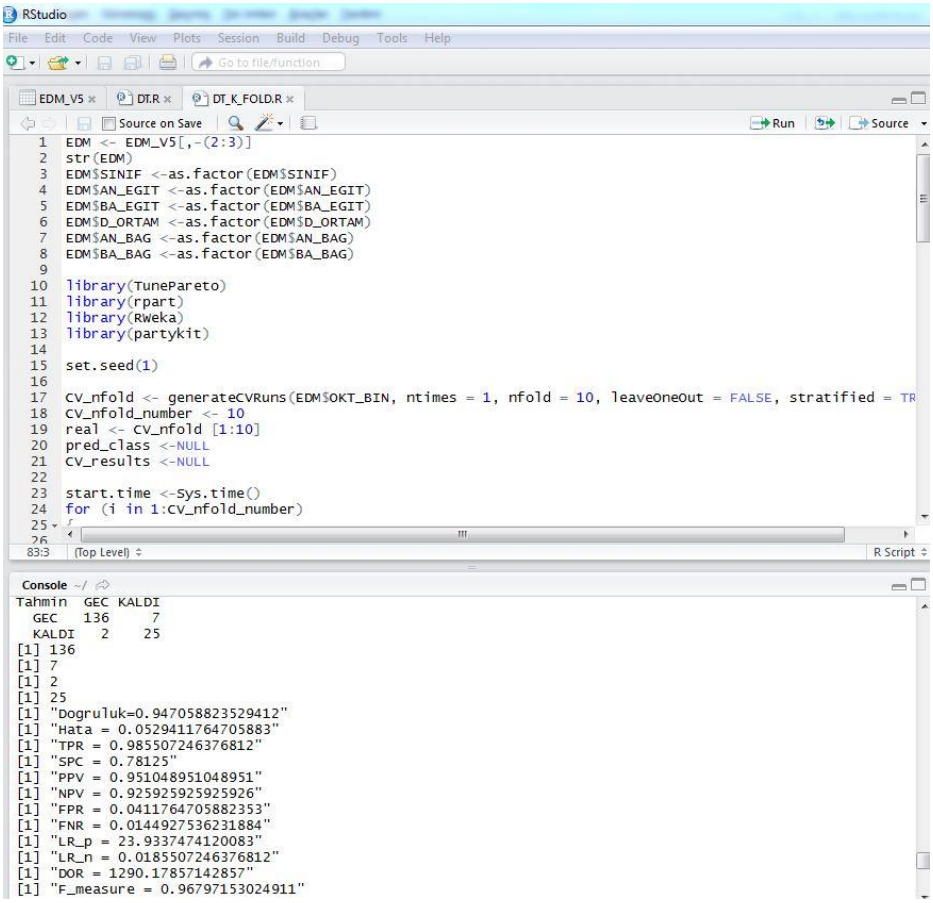


Şekil 4.23: Karar ağacı Algoritması %90/%10 Tabakalı Holdout ROC Eğrisi.

Karar ağacı algoritmasında %90/%10 ayırımla elde edilen ROC eğrisinde test kümesi için eğrinin altında kalan alan (AUC) 0,8591486 olarak bulunmuştur. Diğer ayırımlar için AUC değeri hesaplanmış ve ortalama AUC değeri 0,87683 olarak elde edilmiştir (Ek 11). RStudio ortamında kurulan karar ağacı modelinin, 90/10 ayırımda başarısını gösteren ekran görüntüsü Şekil 4.24, 10 kat çapraz geçerlemedeki gösteren ekran görüntüsü Şekil 4.25'te verilmektedir.



Şekil 4.24: Kurulan Modelin 90/10 Ayırımda Başarısını Gösteren Rstudio Ekran Görüntüsü.



Şekil 4.25: Kurulan Modelin 10 Kat Çapraz Geçerlemede Başarısını Gösteren Rstudio Ekran Görüntüsü.

EDM veri seti üzerinde karar ağacının çizimi ve kural çıkarımı üzerinden derinlemesine analiz gerçekleştirildiğinde 10-kat çapraz geçerleme kullanılarak elde edilen sonuçlar göz önünde bulundurulmuştur. 10-kat çapraz geçerlemenin tüm katları incelendiğinde en yüksek doğruluk değerinin, tanısal üstünlük değeri (DOR) ve F değerinin 7. katta elde edildiği görülmüştür (Ek 9).

DEVAM <= 34.5
| AG >55
| | DERS = HG2
| | | SIN_MEV = LEV2: GEC (1.0)
| | | SIN_MEV = LEV3: GEC (22.0)
| | | SIN_MEV = LEV4
| | | | SIN_T = E: KALDI (2.0)
| | | | SIN_T = H
| | | | | YAS <= 16: GEC (10.0)
| | | | | YAS > 16: KALDI (2.0)
| | DERS = UN2
| | | SINIF = 9: GEC (16.0)
| | | SINIF = 10
| | | | CINS = E
| | | | | SEMT_MEM = E: GEC (53.0/1.0)
| | | | | SEMT_MEM = H
| | | | | | T <= 17: GEC (5.0)
| | | | | | T > 17: KALDI (14.0)
| | | | CINS = K
| | | | | T <= 20: GEC (5.0)
| | | | | T > 20)
| | | SINIF = 11: GEC (22.0)
| | | SINIF = 12
| | | | | DUK <= 45: GEC (2.0)
| | | | | DUK > 45: KALDI (7.0)
| | | | SIN_MEV = LEV2: GEC (2.0)
| | | | SIN_MEV = LEV3: GEC (6.0)
| | | | SIN_MEV = LEV4: KALDI (30.0/1.0)
| AG <=55
| | T <= 17
| | | DERS = HG2: GEC (658.0/14.0)
| | | DERS = UN2
| | | | SIN_MEV = LEV2: GEC (53.0)
| | | | SIN_MEV = LEV3: GEC (161.0/9.0)
| | | | SIN_MEV = LEV4
| | | | | SIN_T = E
| | | | | | SEMT_MEM = H
| | | | | | | D <= 14: GEC (75.0)
| | | | | | | D > 14: KALDI (14.0/4.0)
| | | | | | SEMT_MEM = E: GEC (2.0)

| | | | | SIN_T = H
| | | | | | DUK <= 57: GEC (204.0/27.0)
| | | | | | DUK > 57: KALDI (8.0/2.0)
| | T > 17
| | | SIN_T = H: GEC (22.0)
| | | SIN_T = E
| | | | NET = HG2
| | | | | BIRLIK = E: GEC (58.0/4.0)
| | | | | BIRLIK = H: KALDI (4.0/1.0)
| | | | NET = UN2
| | | | | SEMT_MEM = E
| | | | | | DERS = HG2: GEC (8.0/1.0)
| | | | | | DERS = UN2
| | | | | | | YAS <= 14: GEC (2.0)
| | | | | | | YAS > 14
| | | | | | | | DUK <= 41: GEC (2.0)
| | | | | | | | DUK > 41: KALDI (27.0)
| | | | | SEMT_MEM = H: KALDI (22.0/2.0)
DEVAM > 34.5
| BDI <= 7: GEC (18.0)
| BDI > 7
| | AN_EGIT = 0: KALDI (2.0)
| | AN_EGIT = 1: GEC (2.0)
| | AN_EGIT = 2: GEC (8.0/1.0)
| | AN_EGIT = 3: KALDI (3.0)
| | AN_EGIT = 4: GEC (6.0)
| | AN_EGIT = 5: GEC (5.0)
| | AN_EGIT = 6: GEC (15.0)

Number of Leaves : 42

Size of the tree : 70

Karar ağacının düğümleri niteliğin sağladığı bilgi kazancı (information gain) doğrultusunda oluşmaktadır. Ağaç yapısı incelendiğinde dallanmanın okula devam/devamsızlık durumunu belirten nitelik *DEVAM* ile başladığı görülmektedir. Bu durumda bu niteliğin diğer niteliklere kıyasla en fazla bilgiyi sağlayan düğüm, bilgi kazanç değeri en yüksek nitelik olduğu söylenebilecektir. Şekil 4.24'te EDM veri setinin bilgi kazanç değerleri verilmektedir.

DEVAM	9.599054e-01
CINS	5.320650e-02
SINIF	4.487834e-02
SIN_MEV	2.556130e-02
YAS	1.804137e-02
IST_IK	1.309445e-03
SEMT_MEM	9.196310e-03
AIL_KS	1.377677e-04
AN_EGIT	2.183424e-03
AN_IS	1.954492e-03
BA_EGIT	2.142135e-03
BA_IS	5.833629e-04
BIRLIK	9.819331e-05
MAD_DU	7.656387e-04
SIN_O	6.895833e-03
DERS_DEST	7.214810e-04
SIN_T	6.933526e-04
D_ORTAM	6.372435e-04
TV	4.328613e-03
NET	1.448082e-04
DERS	7.915102e-02
AN_BAG	1.189269e-03
BA_BAG	1.518335e-03
YUKSEK	1.296571e-04
CEKEN_MODEL	1.311820e-02
ITEN_MODEL	7.796008e-03
T	1.204190e-01
D	4.020646e-02
AG	1.457087e-01
DUK	2.451854e-02
SUK	6.461967e-03
BDI	6.804434e-02

Şekil 4.26: Karar Ağacı algoritmasında EDM veri setinin bilgi kazanç değerleri.

Karar Ağacı yapısı, analiz yapılan veri seti üzerinden bir takım kurallar çıkarılmasına olanak sağlamaktadır. Aşağıda EDM veri setine ait karar ağacı yapısından çıkarılabilen kuralların bazıları verilmektedir:

KURAL 1: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5) v**e akademik güdülenmesi orta düzeyin üzerinde **(AG >55)** ve günlük ders çalışma süresi en az 2 saat **(DERS = HG2)** ve sınıf mevcudu 35 kişiden az **(SIN_MEV = LEV2, LEV3)** ise yılsonu akademik başarı ortalaması geçme düzeyinde olur **(OKT_BIN = GEÇ).**

KURAL 2: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5) v**e akademik güdülenmesi orta düzeyin üzerinde **(AG >55)** ve günlük ders çalışma süresi en az 2 saat **(DERS = HG2)** ve sınıf mevcudu 35 kişiden fazla **(SIN_MEV = LEV4)**

ve sınıf tekrarı yapmış **(SIN_T = E)** ise yılsonu akademik başarı ortalaması kalma düzeyinde olur **(OKT_BIN = KALDI).**

KURAL 3: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5)** ve akademik güdülenmesi orta düzeyin üzerinde **(AG >55)** ve günlük ders çalışma süresi en az 2 saat **(DERS = HG2)** ve sınıf mevcudu 35 kişiden fazla **(SIN_MEV = LEV4)** ve sınıf tekrarı yapmamış **(SIN_T = H)** ve yaşı en fazla 16 **(YAS <= 16)** ise yılsonu akademik başarı ortalaması geçme düzeyinde olur **(OKT_BIN = GEÇ).**

KURAL 4: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5)** ve akademik güdülenmesi orta düzeyin üzerinde **(AG >55)** ve günlük ders çalışma süresi 2 saatin altında **(DERS = UN2)** ve cinsiyeti erkek **(CINS = E)** ve yaşadığı semtten memnun değilse **(SEMT_MEM = H)** ve tükenmişlik düzeyi orta düzey ve altında ise (**T <= 17)** ise yılsonu akademik başarı ortalaması geçme düzeyinde olur **(OKT_BIN = GEÇ).**

KURAL 5: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5)** ve akademik güdülenmesi orta düzeyin üzerinde **(AG >55)** ve günlük ders çalışma süresi 2 saatin altında **(DERS = UN2)** ve cinsiyeti kadın **(CINS = K)** ve tükenmişlik düzeyi yüksek-kabul düzey ve altında (**T <= 20)** ise yılsonu akademik başarı ortalaması geçme düzeyinde olur **(OKT_BIN = GEÇ).**

KURAL 6: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5)** ve akademik güdülenmesi orta düzeyin üzerinde **(AG >55)** ve günlük ders çalışma süresi 2 saatin altında **(DERS = UN2)** ve cinsiyeti kadın **(CINS = K)** ve tükenmişlik düzeyi yüksek düzeyde (**T > 20)** ve 12. sınıfta okuyorsa ve anlık kaygı düzeyi kabul edilebilir düzeyde **(DUK <= 45)** ise yılsonu akademik başarı ortalaması geçme düzeyinde olur **(OKT_BIN = GEÇ).**

KURAL 7: EĞER öğrencinin devamsızlığı 34,5 gün ve altı **(DEVAM ≤ 34,5)** ve akademik güdülenmesi orta düzeyin altında **(AG ≤55)** ve tükenmişlik düzeyi orta düzey ve üstünde (**T >17)** ve günlük ders çalışma süresi 2 saatin altında **(DERS = UN2)** ve sınıf tekrarı yapmış ve günlük internet kullanımı 2 saatten fazla **(NET = HG2)** ve anne-babası birlikte **(BIRLIK = E)** ise yılsonu akademik başarı ortalaması geçme düzeyinde olur **(OKT_BIN = GEÇ).**

4.5. LOGISTIK REGRESYON ANALİZİNDEN ELDE EDİLEN BULGULAR

EDM veri seti üzerinde Logistik Regresyon Analizi'nde, OKT_BIN hedef niteliğinin tahmininde, Caret (Kuhn ve diğ., 2015) ve TunePareto (Müssel ve diğ., 2012), FSelector (Romanski ve Kothoff, 2015) paketleri kullanılmıştır. Sınıflandırıcının performansı; %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında hold-out ile oluşturulan eğitim ve test kümeleri üzerinden çeşitli performans kriterleri kullanılarak hesaplanmıştır. Ek 12'de performans tabloları ayrıntılı şekilde sunulmaktadır.

Hedef nitelik OKT_BIN geçme ve kalma üzerine kurulu olduğundan ikili logistik regresyon analizi yapılmıştır. Kullanılan veri setindeki değişkenlerin birbirinden bağımsız olup olmadıklarına dair ilişkiler nümerik nitelikler için Pearson r korelasyon katsayısı kullanılarak incelenmiştir (Şekil 4.27).

P

	DEVAM	YAS	T	D	AG	DUK	SUK	BDI
DEVAM		0.0000	0.0000	0.0000	0.0000	0.2626	0.0065	0.1698
YAS	0.0000		0.0000	0.0000	0.0002	0.1916	0.5295	0.0000
T	0.0000	0.0000		0.0000	0.0000	0.0187	0.0144	0.0293
D	0.0000	0.0000	0.0000		0.0000	0.0000	0.0009	0.3319
AG	0.0000	0.0002	0.0000	0.0000		0.0000	0.3454	0.0225
DUK	0.2626	0.1916	0.0187	0.0000	0.0000		0.0000	0.8615
SUK	0.0065	0.5295	0.0144	0.0009	0.3454	0.0000		0.1850
BDI	0.1698	0.0000	0.0293	0.3319	0.0225	0.8615	0.1850	

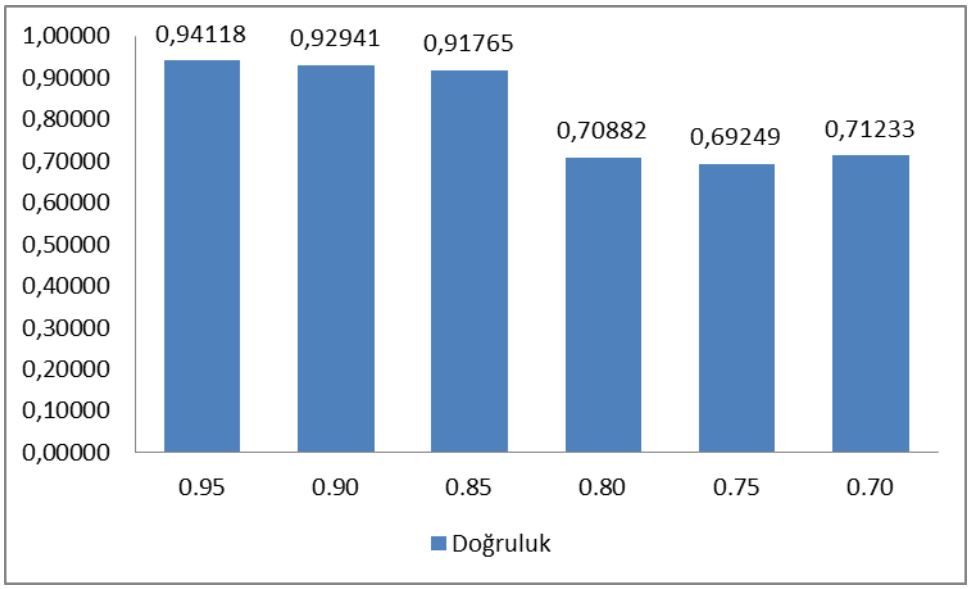
Şekil 4.27:EDM Veri Setindeki Nümerik Değerler İçin Pearson R Korelasyon Katsayılarını İçeren R Ekran Görüntüsü.

Şekil 4.25 incelendiğinde SUK ile YAS değişkeni arasında 0.5 üzerinde bir korelasyon olduğu görülmektedir. Bu durumda SUK değişkeninin analizden çıkarılarak YAS değişkeni ile devam edilmesi uygun görülmüştür. Nitekim SUK değişkeni gibi kaygıyı temsil eden DUK değişkeninin varlığı "kaygının" başarı üzerindeki etkisinin bu analizde görülebilmesi açısından önemlidir.

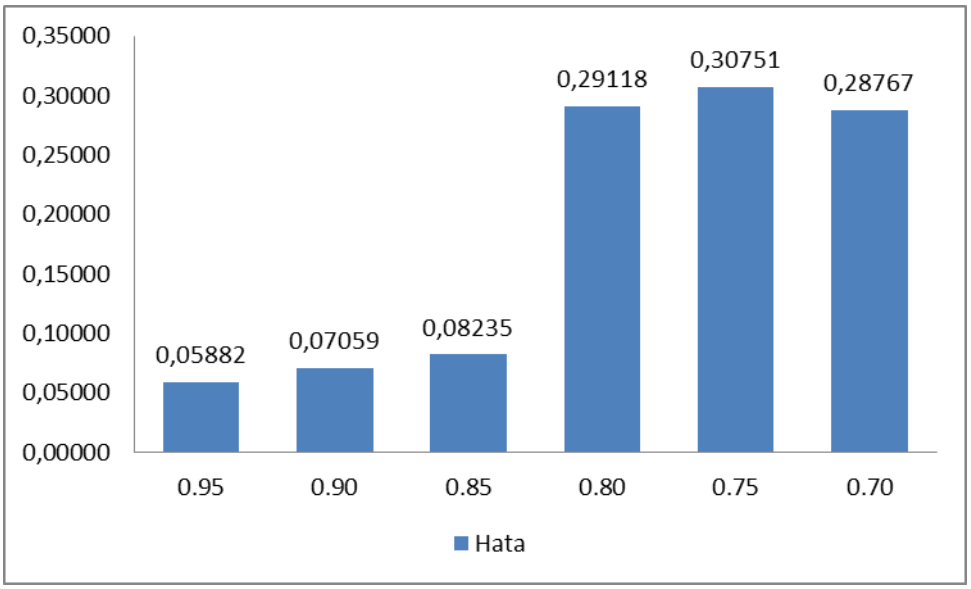
EDM veri setindeki iki değerli kategorik değişkenler arasındaki ilişki için phi korelasyon (Cramér's V) katsayıları hesaplanmıştır. R hazır paketleri arasında bu hesabı yapabilen bir pakete rastlanmadığından bu korelasyon katsayısı hesabı için fonksiyon

yazılmıştır. EDM veri setinden elde edilen phi korelasyon değerleri Ek 13'de verilmektedir.

Korelasyon hesaplamalarından sonra tekrarlı holdout yöntemi ile analizlere devam edilmiştir. Logistik Regresyon Analizine ait R ekran görüntüleri Ek 14'te verilmektedir. Yapılan analizlerde modelin verileri yorumlama gücü Nagelkerke katsayısı hesaplanarak araştırılmış ve 0.7862636 olarak elde edilmiştir. EDM veri setinde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında tabakalı hold-out örnekleme uygulanmıştır. Logistik Regresyon Analizi tabakalı hold-out örnekleme yöntemiyle elde edilen ortalama doğruluk grafiği Şekil 4.28.'de, hata grafiği ise Şekil 4.29'da verilmektedir.



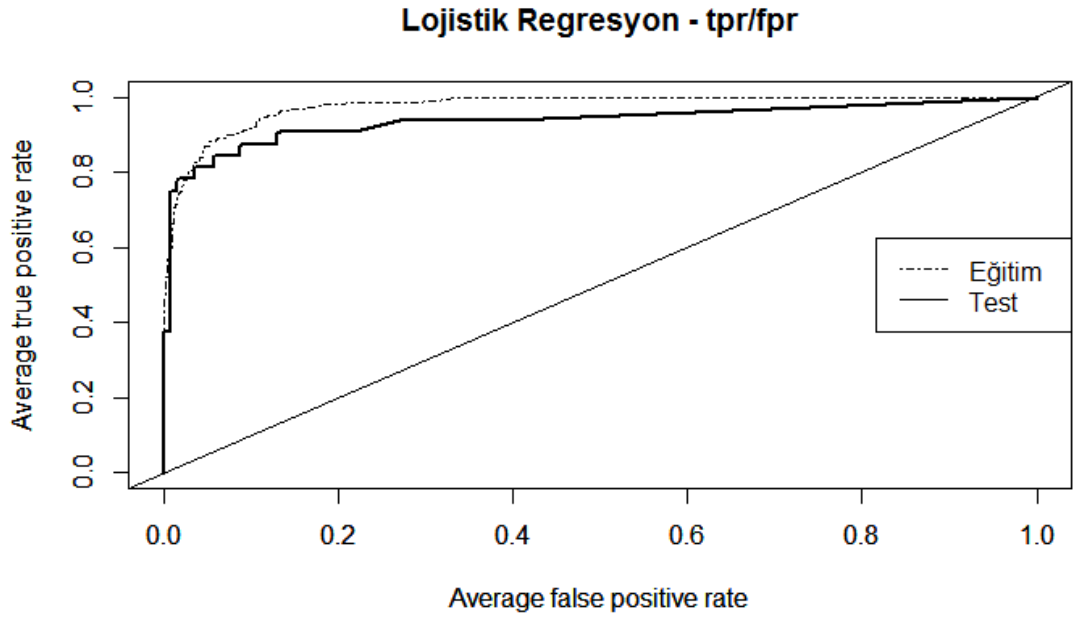
Şekil 4.28: Logistik Regresyon Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği.



Şekil 4.29: Logistik Regresyon Analizinde Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği.

Şekil 4.29 ve 4.30 incelendiğinde holdout yöntemiyle yapılan analizlerde doğruluk değerinin yaklaşık olarak 0,70 ile 0,95 arasında değiştiği görülmektedir. Logistik regresyon analizinde en yüksek doğruluk değeri, %95/%5 ayırımda elde edilmektedir.

Logistik regresyon analizinde k-kat çapraz geçerleme ile elde edilen sonuçlar incelendiğinde ortalama doğruluk değeri 0,91176, ortalama F değeri 0,93878 ve ortalama DOR değeri 925,02223 bulunmaktadır. Benzer şekilde holdout yönteminde doğruluk değeri 0,81698, ortalama F değeri 0,88898 ve ortalama DOR değeri 274,20213 olarak bulunmuştur (Ek 10 ve Ek 11). Tüm bu değerler ışığında, EDM veri seti üzerinde karar ağacı algortimasınında iyi bir performans gösterdiğini söylemek mümkündür. Bu değerlere ek olarak, %90/%10 tabakalı holdout için ROC eğrisi çizdirilmiştir (Şekil 4.30).



Şekil 4.30: Logistik Regresyon Analizi %90/%10 Tabakalı Holdout ROC Eğrisi.

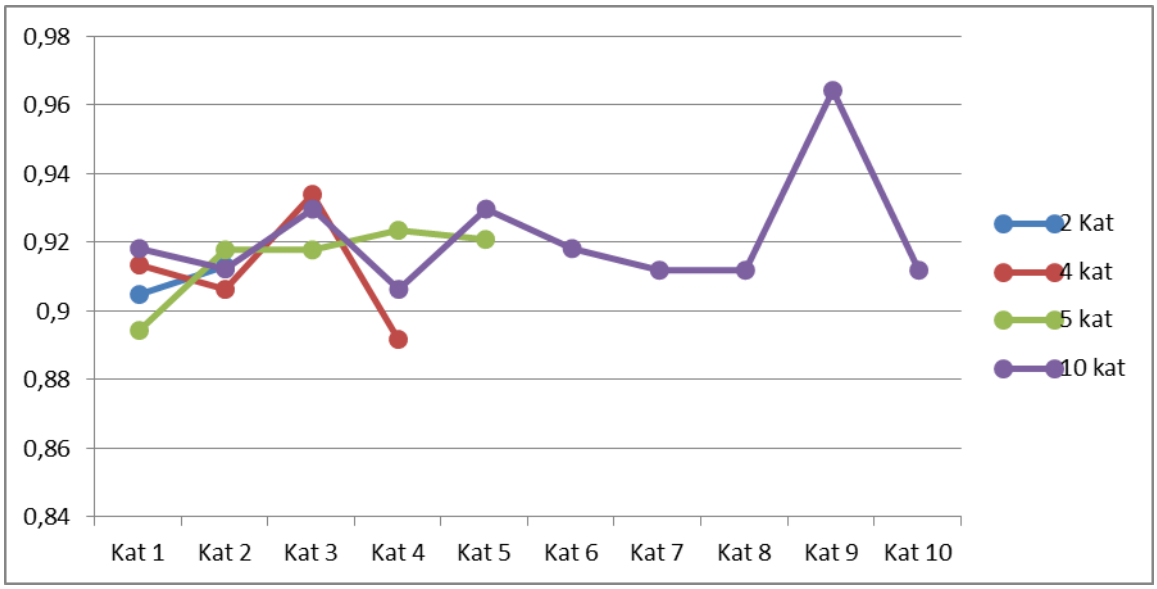
Karar ağacı algoritmasında %90/%10 ayırımla elde edilen ROC eğrisinde test kümesi için eğrinin altında kalan alan (AUC) 0,95199 olarak bulunmuştur. Diğer ayırımlar için AUC değeri hesaplanmış ve ortalama AUC değeri 0,96229 olarak elde edilmiştir (Ek 12).

4.6. DESTEK VEKTÖR SINIFLANDIRICI İLE ELDE EDİLEN BULGULAR

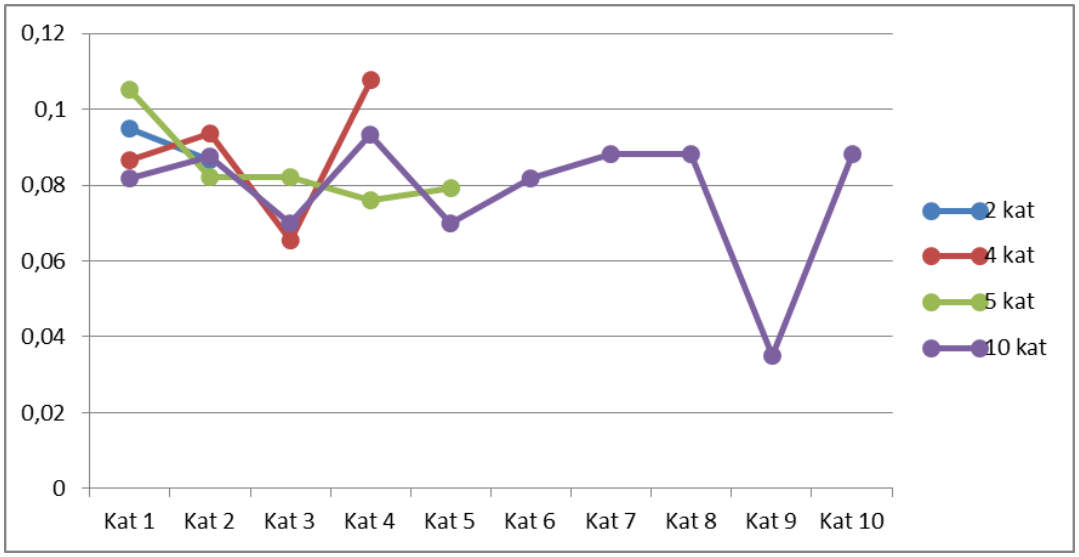
EDM veri seti üzerinde Destek Vektör Sınıflandırıcı ile, OKT_BIN hedef niteliğinin tahmininde, Caret (Kuhn ve diğ., 2015), e1071 (Meyer ve diğ., 2014) paketleri kullanılmıştır. Sınıflandırıcının performansı; %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında hold-out ve 10, 5, 4 ve 2 kat çapraz geçerleme ile oluşturulan eğitim ve test kümeleri üzerinden çeşitli performans kriterleri kullanılarak hesaplanmıştır. Ek 15, Ek 16 ve Ek 17'de performans tabloları ayrıntılı şekilde sunulmaktadır.

Şekil 4.31.'de DV sınıflandırıcı ile EDM veri setinde 2-kat, 4-kat, 5-kat, 10-kat çapraz geçerleme kullanılması sonucu elde edilen performans değerlerinden ortalama doğruluk

(ACC) değerine ilişkin grafik verilmektedir. Doğruluğa bağlı olarak hesaplanabilen hata değeri, Şekil 4.32'de sunulmaktadır.



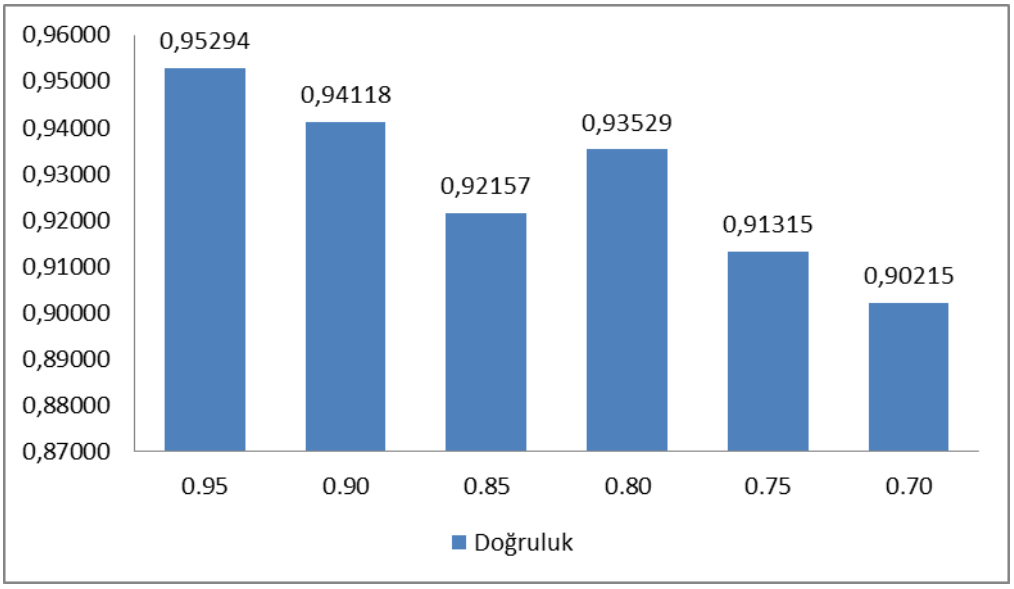
Şekil 4.31: Destek Vektör Sınıflandırıcı k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Doğruluk Değerleri Grafiği.



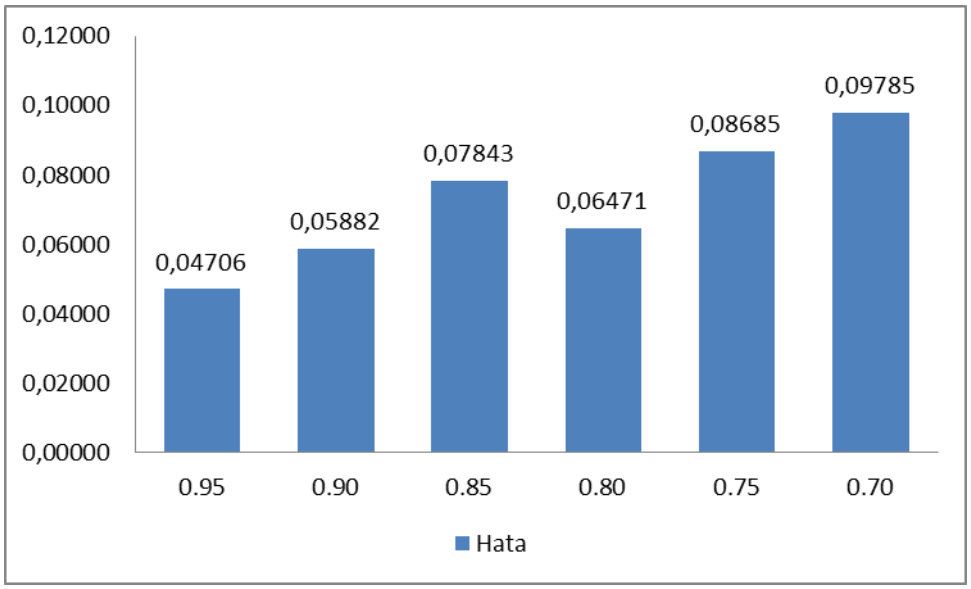
Şekil 4.32: Destek Vektör Sınıflandırıcı k-Kat Çapraz Geçerleme Kullanılarak Elde Edilen Hata Değerleri Grafiği.

Şekil 4.31 ve 4.32 incelendiğinde k kat çapraz geçerlemelerle elde edilen her test ve eğitim küme ikilisi için doğruluk değerlerinin 0,88 ile 0,98 arasında değişim göstermektedir. Modelin performansının doğru bir biçimde irdelenebilmesi için her kat için ortalama performans değerleri oluşturulması gerekmektedir. Ek 16'da ortalama performans değerleri sunulmaktadır.

k-kat çapraz geçerleme dışında, EDM veri setinde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 oranlarında tabakalı hold-out örnekleme uygulanmıştır. DV sınıflandırıcı tabakalı hold-out örnekleme yöntemiyle elde edilen doğruluk grafiği Şekil 4.33'de, hata grafiği ise Şekil 4.34'te verilmektedir.



Şekil 4.33: Destek Vektör Sınıflandırıcı Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Doğruluk Değerleri Grafiği.



Şekil 4.34: Destek Vektör Sınıflandırıcı Tabakalı Hold-Out Örnekleme Yöntemiyle Elde Edilen Hata Değerleri Grafiği.

Şekil 4.33 ve 4.34 incelendiğinde holdout yöntemiyle yapılan analizlerde doğruluk değerinin yaklaşık olarak 0,90 ile 0,96 arasında değiştiği görülmektedir. DV sınıflandırıcıda en yüksek doğruluk değeri, %95/%5 ayırımda elde edilmektedir.

Sınıflandırıcının k-kat çapraz geçerleme ile elde edilen sonuçlar incelendiğinde ortalama doğruluk değeri 0,91428, ortalama F değeri 0,94752 ve ortalama DOR değeri 401,20089 bulunmaktadır. Benzer şekilde holdout yönteminde doğruluk değeri 0,92771, ortalama F değeri 0,95577 ve ortalama DOR değeri 596,72980 olarak bulunmuştur (Ek 14 ve Ek 15). Tüm bu değerler ışığında, EDM veri seti üzerinde DV sınıflandırıcının iyi bir performans gösterdiğini söylemek mümkündür.

4.7. MODEL PERFORMANSLARININ KARŞILAŞTIRILMASI

EDM veri seti üzerinde uygulanan modellerin başarısı bölümün başında belirtilen performans kriterleri açısından kıyaslanacaktır. Bu sayede eldeki veri setinde en iyi sonuç üreten algoritmanın belirlenmesi söz konusu olacaktır. Tez çalışması kapsamında kullanılan sınıflandırma yöntemlerinin performanslarının genel olarak karşılaştırması Tablo 4.2 ve Tablo 4.3'de verilmektedir.

Tablo 4.2: Kullanılan Algoritma/Sınıflandırıcıların Tabakalı 90/10 Tabakalı Hold Out Yöntemine Göre Genel Performanslarının Karşılaştırılması.

	Ortalama Doğruluk (ACC)	Ortalama Hata	Tanısal Üstünlük Değeri (DOR)	F
k-NN Algoritması	0,88219	0,11781	180,92707	0,92958
Naive Bayes Sınıflandırıcı	0,90237	0,09763	269,44636	0,93946
Karar Ağacı Algoritması	0,92166	0,07834	**632,51655**	0,95255
Logistik Regresyon Algoritması	0,81698	0,18302	274,20213	0,88898
Destek Vektör Sınıflandırcı	0,92771	0,07229	596,72980	**0,95577**

Tablo 4.2 incelendiğinde; 90/10 tabakalı hold out yönteminde, EDM veri seti üzerinde akademik başarının tahmin edilmesine yönelik en iyi performans; ortalama doğruluk değeri üzerinden bakıldığında Karar Ağacı Algoritması ve Destek Vektör Sınıflandırıcı ile elde edildiği görülmektedir. Performans kıyasında daha hassas bilgiler sunan tanısal üstünlük değeri açısından incelendiğinde Karar Ağacı Algoritması'nın daha başarılı sonuçlar ürettiğini söylemek mümkündür.

Tablo 4.3: Kullanılan Algoritma/Sınıflandırıcıların Tabakalı 10-Kat Çapraz Geçerleme Yöntemine Göre Genel Performanslarının Karşılaştırılması.

	Ortalama Doğruluk (ACC)	Tanısal Üstünlük Değeri (DOR)	F	Ortalama Hata
k-NN Algoritması	0,90588	258,98438	0,94203	0,09412
Naive Bayes Sınıflandırıcı	0,94118	961,05655	0,96324	0,05882
Karar Ağacı Algoritması	0,94706	**1.290,17857**	**0,96797**	0,05294
Logistik Regresyon Algoritması	0,91176	925,02223	0,93878	0,08824
Destek Vektör Sınıflandırcı	0,91176	527,98077	0,94774	0,08824

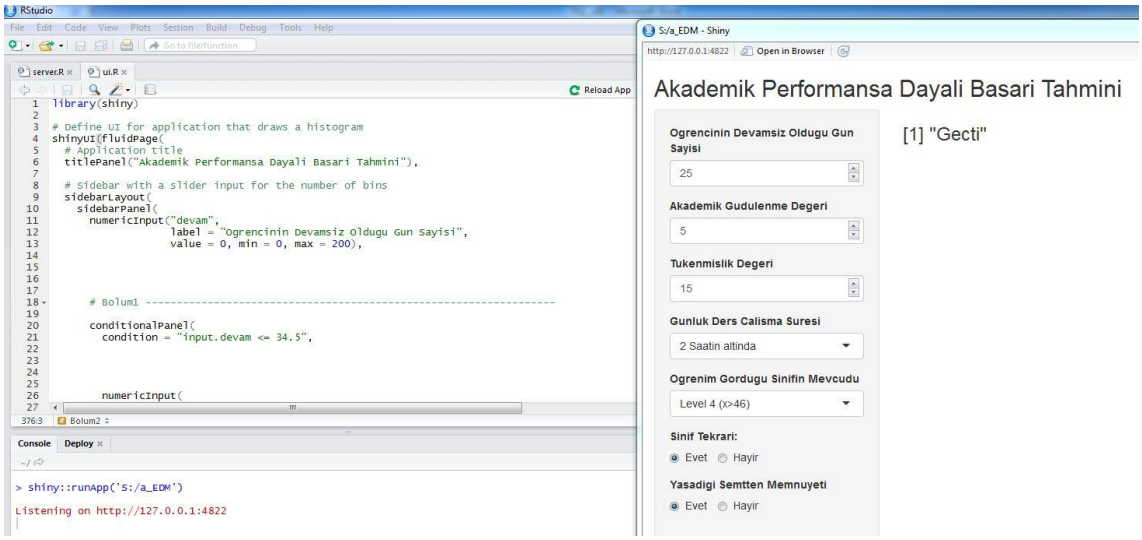
Tablo 4.3. incelendiğinde 10-kat çapraz geçerleme yöntemine göre DOR ve F değerleri ile en iyi performansın karar ağacı algoritması tarafından sergilendiği görülmektedir. Bu algoritmayı, tanısal üstünlük değeri bakımından kıyaslandığında Naive Bayes Sınıflandırıcı, Logistik Regresyon analizi, Destek Vektör Sınıflanrıcı ve k-NN analizi izlemektedir.

Tüm bu değerler ışığında, EDM veri seti üzerinde karar ağacı algoritmasının, diğer modellere göre daha iyi bir performans gösterdiğini söylemek mümkündür. Bu

değerlere ek olarak, %90/%10 tabakalı holdout için modeller içinROC eğrisi çizdirilmiştir (Ek 16).

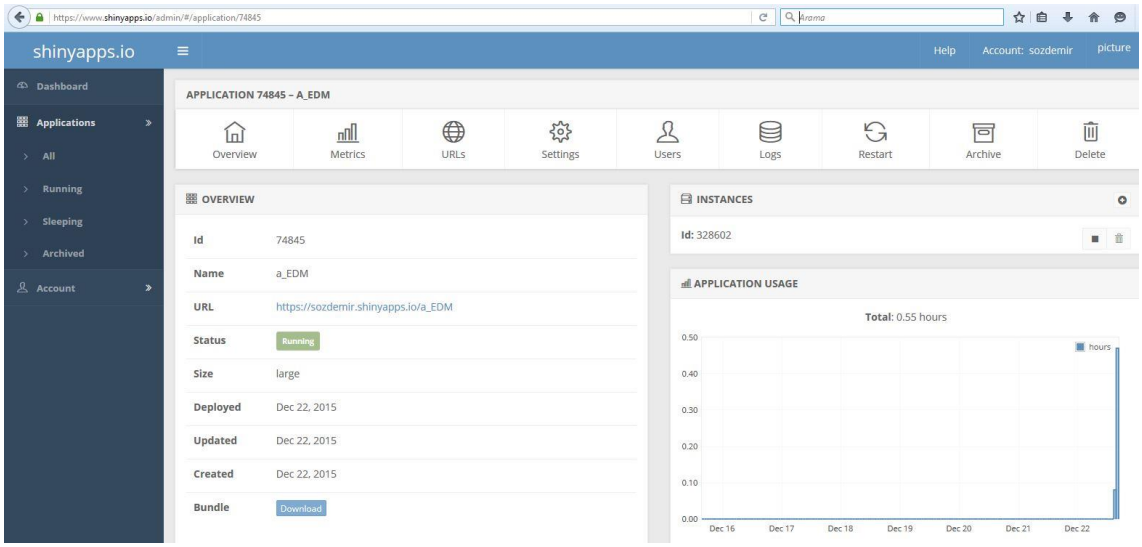
4.8. SHINY

Shiny; R dili ile yapılan analizlerin/çalışmaların web üzerinden paylaşıma açılmasını sağlayarak, aktif kullanılabilir yapılar haline dönüştürülmesine imkan tanımaktadır. Bir RStudio projesi olan yapı, shinyapps.io'da ücretsiz hesap açılması ile paylaştırılmaktadır. Kullanıcının web programlama konusunda temel düzeyde bilgi sahibi olması bile çalışmasını shiny ile paylaşıma açmasında yeterlidir. Bu tez çalışması kapsamında shinyapps.io'da açılan hesap ile karar ağacı modeli paylaşılmıştır (Şekil 4.35).

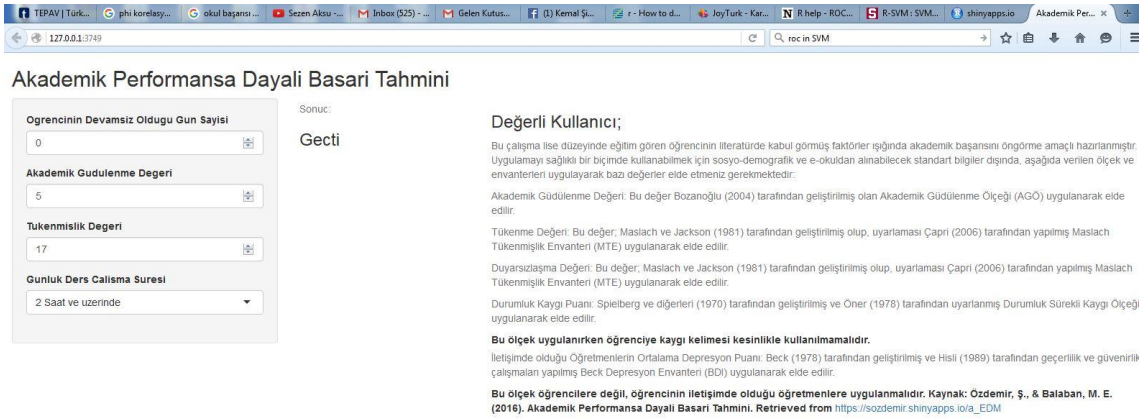


Şekil 4.35: Karar Ağacının Shiny İle Webe Aktarılması Sürecinde Rstudio Ekran Görüntüleri.

Ui ve server üzerinden hazırlanan karar ağacını içeren uygulama "publish" komutunun seçilmesi ile shinyapps.io'da açılan hesapta aktif hale gelmektedir. Kullanıcı shinyapps.io'ya aktardığı uygulamayı aktifleştirme, pasifleştirme ve arşivleme seçeneklerine sahiptir. Ayrıca hazırlanan uygulamanın kullanım durumuna yönelik grafik yine bu arayüzde sunulmaktadır (Şekil 4.36 ve 4.37).



Şekil 4.36: Shinyapps.io Kullanıcı -Uygulama Arayüzü.



Şekil 4.37: Publish Edilen Karar Ağacı Modelinin Shiny Arayüzü.

5. TARTIŞMA VE SONUÇ

Mevcut eğitim sistemimize her yıl 1 milyonu aşkın öğrenci dahil olmaktadır. Bu sistemin çıktıları bakımından değerlendirilmesi yapıldığında, ülkemiz eğitim hedef ve politikalarına göre nitelikli/kaliteli bir düzeyin henüz elde edilemediğini söylemek mümkündür. Mezun olmak için asgari düzeyde ortalama şartının sağlanması ile bir sonraki kademeye geçmesi beklenen öğrenci; eğitim-öğretim süreci içinde içsel ve/veya dışsal faktörlere bağlı olarak, kimi zaman fiili kimi zaman ruhen okuldan kopmalar yaşamaktadır. Bu kopmalara bağlı olarak, öğrenci; içselleştirilmiş bir başarı isteğinden uzak, geçme koşulunu sağlayacak kadar performans sergilemektedir.

Bireylerin gelecekteki başarıları için belirli bir akademik başarının sağlanarak bir sonraki aşamaya geçmeleri önemli bir kriter ise, bu kriterin eğitim-öğretim süreci içinde gizlenen başarı/başarısızlık örüntülerini ortaya çıkaran ilişkiler üzerinden keşfedilmesi, sürecin paydaşları, özellikle eğitimciler, için hayati derecede bir bilgi sağlayacaktır.

Bu tez çalışmasında lise öğrencilerinin akademik başarıları, sosyo-demografik değişkenler, başarıyı dolaylı ve/veya doğrudan etkileyen kavramlar baz alınarak sınıflandırma yöntemleri ile belirlenmeye çalışılmıştır. Öğrencinin akademik başarısı üzerinde etkisi olabilecek faktörler; literatür taramaları ile belirlenmiştir. Veri seti; İstanbul'da bulunan sosyo-kültürel ve ekonomik anlamda farklılıklar gösteren, bu farklılıkları ile İstanbul'un kozmopoliti yapısını temsil edebilecek nitelikteki ilçelerde bulunan liselerde yapılan uygulamalarla oluşturulmuştur. Bu çalışma; literatürde eğitimde veri madenciliği anlamında yapılan çalışmalardan aşağıda belirtilen yönleri ile farklılık göstermektedir:

- Çalışmada farklı akademik araştırmalarda ilişkileri sınırlı olarak kontrol edilen değişkenler (başarı-akademik güdülenme ilişkisi, başarı- tükenmişlik ilişkisi, başarı-aile ortamı ilişkisi vb.) bir bütün şeklinde ele alınmıştır.
- Eğitim alanında yapılan çalışmalarda, bağımlı ve bağımsız değişkenler arasındaki ilişki incelenmekte ve sonuç olarak ilişkinin varlığı, yokluğu, negatif

veya pozitif yönde oluşu ile ilgili çalışmalar yapılmaktadır. Bu tez çalışmasında ilişkinin varlığı, yokluğu, yönü bir adım ileri taşınarak, hedef nitelik üzerindeki etkisi ile, niteliğin değeri tahminlenmiştir.

- Ülkemiz literatüründe başarının irdelendiği pek çok çalıma bulunmakta olup, bunların büyük bir bölümü çeşitli seviyelerdeki istatistik analizlere dayanmaktadır. Bu tez çalışması ile istatistik ve matematik yöntemlerin bir arada kullanıldığı veri madenciliği yöntem ve teknikleri ile analiz yapılmıştır.
- Modellere ilişkin performanslar tabakalı çapraz geçerleme, tabakalı hold out yöntemleri ile analiz edilmiş, elde edilen sonuçlar üzerinden performanslar yorumlanmıştır.
- Modellerin performanslarının kıyaslanmasında sadece doğruluk, sadece hata gibi tekil ölçütler yerine; bütünleşik bir yapıya sahip, farklı performans ölçütlerinden oluşturulmuş olan f-ölçüsü, tanısal üstünlük oranı değerleri baz alınmıştır.

Tez çalışması kapsamında öğrencinin akademik başarısının tahmin edebilmek için, k-NN Algoritması, Naive Bayes Sınıflandırıcı, Karar Ağacı Algoritması (C4.5), Logistik Regresyon Analizi ve Destek Vektör Sınıflandırıcı'dan faydalanılmıştır. Çalışmada en yüksek performans karar ağacı algoritmasından elde edilmiştir (Tablo 4.2 ve 4.3).

EDM veri setinde okul başarısı ikili hale dönüştürülerek (geçti/kaldı) incelenmiştir. Okul başarısının kategorilere ayrılması (Tablo 3.7) ile yapılan analizlerde oldukça düşük performanslar elde edilmiştir.

Kullanılan sınıflandırma yöntemlerinin hata oranlarına bakıldığında 0,08-0,05 aralığında değişmeler olduğu görülmektedir (Tablo 4.2 ve 4.3). Uygulama alanı olarak seçilen eğitim-öğretim süreci için bu aralıkta değişen hatalar ihmal edilebilir düzeyde kabul edilmektedir.

Tez kapsamında yapılan tüm analizler R programlama dili ile gerçekleştirilmiş olup, programlama ortamı olarak RStudio, web ortamı olarak Shiny kullanılmıştır. R; mevcut programlama dilleri arasında popülaritesi giderek artan bir programlama dilidir (IEEE, 2015). Ayrıca bilimsel hesaplama, istatistik, matematiksel hesaplama, veri madenciliği alanlarında ağırlıklı olarak kullanılmaktadır. Alternatiflerine kıyasla açık kaynak kodlu

oluşu; paketleri, aktif iletişim platformları, toplulukları ile daha zengin bir eko-sistemi sunması, R dilinin daha öne çıkmasını sağlamaktadır (Theuwissen, 2015).

EDM veri seti 1706 öğrenci verisi barındırmakta olup, tabakalı 10-kat çapraz geçerleme kullanıldığında yaklaşık 170 gözlem her denemede test veri seti olarak ayrılmıştır. Tabakalı hold out yöntemi ile de %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 ayrımlarda eğitim ve test veri setleri oluşturulmuştur.

Bu tez çalışmasının gerek örneklem grubu gerekse kullanılan yöntemler açısından bazı sınırlılıkları bulunmaktadır:

- Örneklem grubu sadece İstanbul ilini temsil edecek şekilde seçilen ilçelerden alınan lise düzeyinde eğitim gören öğrencilerle sınırlıdır.
- Veri madenciliği yöntemlerinden sadece sınıflandırmaya dayalı tekniklerden seçilmiş olan karar ağacı, naive bayes, k-nn, logistik regresyon, destek vektör makineleri modeller ışığında veriler analiz edilmiştir. Bu nedenle tez çalışması belirtilen modellerin kullanımı ile sınırlıdır.

Bu tez çalışmasının klasik eğitim ortamından derlenen verilere, sınıflandırma teknikleri uygulanması ile akademik başarının tahmin edilmesi öne çıkan özelliğidir. Bu özelliğin "doktora çalışmalarında bilime, insanlığa katkı yapması felsefesi" ile Shinny projesi kullanılarak, herhangi bir eğitimcinin bile kendi öğrencileri için analiz yapabileceği hale getirilerek desteklenmesi, bir diğer öne çıkan özelliktir. Tez kapsamında geliştirilen ve sunulan CRISP-EDM süreç modeli önerisi de eğitm alanındaki veri madenciliği problemlerinin çözümünde izlenecek adımları sunması açısından öne çıkan bir başka özelliğidir.

EDM veri seti üzerinde yapılan çalışmalar sonucunda, akademik başarının aşağıda önem derecelerine göre sıralanmış faktörler üzerinden yorumlanabileceği söylenebilmektedir:

- Akademik başarıyı en çok etkileyen faktör "okula devamlı" olmaktır. Bir yılda ortalama 185 iş günü eğitim-öğretim faaliyetleri sürdürülmektedir. Kurulan karar ağacı modeli 34,5 iş günü devamsızlığı kritik bir ayırım noktası kabul etmektedir.

- İkinci dereceden önemli olan faktör ise Akademik Güdülenme düzeyidir. Nitekim güdülenmenin eğitimdeki yeri ve önemi düşünüldüğünde akademik anlamda güdülenmiş öğrencilerin daha başarılı olacakları beklenmektedir. Akademik güdülenmenin akadamik başarıyı yordayabilen bir faktör olarak yer alması, Boyd (2002), Goldberg ve Yumuşak (2006), Yılmaz (2007) tarafından yapılan çalışmalarla tutarlılık göstermektedir.
- Akademik Güdülenme düzeyini öğrencinin günlük ders çalışma süresi ve eğitim gördüğü sınıfın mevcudu takip etmektedir.
- Akademik başarıda cinsiyetin etkinliği, bir faktör olarak değerlendirilebilmesi, Tükenme ve Duyarsızlaşma puanları üzerinden yapılabilmektedir. Ağaç ayırımı incelendiğinde kadınların, tükenmişlik düzeylerini erkeklere göre daha iyi tolere edebildiklerini söylemek mümkündür.
- Tükenmişlik düzeyi yüksek olan öğrencilerde sınıf tekrarı yapılmış olması başarıyı olumsuz etkilemektedir. Bu durum sıkça kullanılan "öğrenilmiş çaresizliğe" vurgu yapmaktadır.
- Devamsızlığın 34,5 iş gününden fazla olduğu durumlarda, öğrencinin ağırlıklı olarak iletişim kurduğu öğretmenlerin depresyon durumları belirleyici rol oynamaktadır. Öğretmendeki depresyon, gerek sınıf içi gerekse sınıf dışı tutum ve davranışlarına yansıyabilecek bir durumdur. Literatürde öğretmenin tutumunun öğrenci başarısını etkilediğine ilişkin olarak Razon (1987), Demir (2009), Aysan ve diğ. (1996) tarafından yapılan çalışmalar, karar ağacındaki bu dallanmanın oluşmasının beklenen bir sonuç olduğunu göstermektedir.
- Depresyon düzeyinin yüksek olması durumunda, annenin eğitm durumu akademik başarı açısından belirleyici rol oynamaktadır. Ailenin destekleyiciliğinin okul başarısı üzerindeki etkisine ilişkin olarak, Sama ve Tarım (2007), Çelenk (2003)'e ait çalışmalar, karar ağacında oluşan ve bir kural biçiminde yazılabilen bu durumu destekler niteliktedir.

Yukarıda sayılan bu faktörler dışında karar ağacında, kaygı faktörü de kuralların oluşturulmasında rol oynamaktadır. Nitekim literatürde kaygının akademik başarı üzerindeki etkisi hakkında Pintrich ve Groot (1990), Başarır (1990), Cassady ve Johnson (2002), Yıldırım ve Ergene (2003), Chapell ve diğ. (2005) tarafından yapılan çalışmalar

ağaç dallanmasınnda ve kural oluşmunda bu katörün yer almasının gerekliliği ile uyum göstermektedir.

Veri madenciliğinin en temel felsefesi yığınlar halindeki verilerden tanımlama ya da tahminleme yapılmasıdır. Sınıflandırma teknikleri, tüketici davranışlarından, sağlık sektörüne kadar pek çok alanda sıklıkla kullanılmakta ve elde edilen sonuçlar doğrultusunda stratejiler geliştirilmekte, eylem planları hazırlanmaktadır. Eğitim gibi kritik bir alanda yapılabilecek uygulamaların, ülkemizde CMS, LMS, LCMS web-logları ile yapılan çalışmalarla sınırlı kalması önemli bir açığı oluşturmaktadır. Bu açığın kapatılması adına ileride yapılması planlanan diğer çalışmalar aşağıda sunulmaktadır:

- Bu tez çalışmasında yapılan tahminler öğrencinin genel başarısını dayalı olarak oluşturulmuştur. Ders bazında başarıyı etkileyen faktörlerin belirlenerek – özellikle matematik dersi-, modellerin kıyaslanması ve shiny projesinin daha detaylı hale getirilmesi,
- Liseler ile sınırlandırılmış bu çalışmanın farklı eğitim-öğretim düzeylerinde de tekrarlanarak, Milli Eğitim Bakanlığı'nın eğitim-öğretim sürecinin kontrol edebileceği, tedbirler alması noktasında uyarılar üretebilecek bir yapının oluşturulması,
- Tez çalışması kapsamında kullanılamayan diğer sınıflandırma yöntemlerinin de uygulanarak en uygun modelin belirlenmesi ve shiny ile geliştirilen yapının daha performanslı modeller ile desteklenmesi,
- Literatüre bağlı olarak oluşturulan akademik başarıyı etkileyen değişkenlerin, bölgesel bazda ve sosyo-kültürel yapı ele alınarak değerlendirilmesi ile daha özerk değişkenler ışığında akademik başarının tahminlenmesi çalışmalarının yapılması,
- Shiny ile webe aktarılan modellerin hassasiyet ve tahmin kuvvvetlerinin arttırılarak, okullar, ilçe milli eğitim müdürlükleri ve il milli eğitim müdürlüğü tarafından aktif olarak kullanılmasının sağlanması.

Sonuç olarak; bilginin büyük bir rekabet avantajı sağladığı çağımızda "bilgi ile ilgili her şeyi ve her süreci" içine alan Enformatik Bilimi'nin eğitim sektöründe uygulanması ile klasik eğitim ortamından derlenen veri seti üzerinde, öğrencinin akademik başarısının

tahminine dayalı sınıflandırıcılar yardımıyla öğrenen ve destekleyici nitelikteki kararlar üretilmiştir. Kullanılan tüm yöntemler ile mikro düzeyde birey, makro düzeyde ulusların, toplumların geleceğini etkileme gücü olan eğitim sürecinin akademik başarı açısından daha kaliteli ve izlenebilir, önlem alınabilir hale getirilmesi sağlanarak katkı verilmesi amaçlanmıştır. Veri madenciliği yöntem ve tekniklerinin, pedagojik unsurlar göz ardı edilmeksizin, eğitim-öğretim sürecinin farklı kademelerine, farklı disiplinlerdeki akademik başarı üzerinden, farklı sınıflandırma teknikleri ile uygulanarak devam ettirilmesi, önce içinde yaşadığımız topluma sonra da insanlığa faydalı olacaktır.

KAYNAKLAR

Akpınar, H., 2014, *Data: Veri Madenciliği, Veri Analizi.* İstanbul: Papatya Yayıncılık.

Alkan, C., 2001, *Türk Milli Eğitim Sisteminin 2000'li yıllarda Yeniden Yapılanmasının Temel Esasları Eğitimde Yansımalar VI.* Ankara: H.H. Tekışık Eğitim Araştırma Geliştirme Merkezi .

Altınkurt, Y., 2008, Öğrenci Devamsızlıklarının Nedenleri ve Devamsızlığın AKademik Başarıya olan Etkisi. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi,* 129-142.

Atasoy, D., 2001, Lojistik Regresyon Analizinin İncelenmesi Ve Bir Uygulaması. *Yayınlanmamış Yüksek Lisans Tezi*. Cumhuriyet Universitesi, Sosyal Bilimler Enstitüsü.

Ayas, A., 2013, Eğitimle İlgili Temel Kavramlar. H. Özmen, & D. Ekiz içinde, *Eğitim Bilimine Giriş* (s. 1-12). Ankara: Pegem Akademi.

Aysan, F., Tanrıöğen, G., & Tanrıöğen, A., 1996, Perceived Causes of Academic Failure Among The Students at the Faculty of Education at Buca. *Teacher Training for The Twenty First Century*.

Baker, R., & Yacef, K., 2009, The State Of Educational Data Mining in 2009: A review and future visions. *Journal of Educational Data Mining,* 3-17.

Balaban, M. E., & Kartal, E., 2015, *Veri Madenciliği ve Makine Öğrenmesi: Temel Uygulamaları ve R Dili ile Uygulamaları.* İstanbul: Çağlayan Kitabevi.

Balkıs, M., Duru, E., Buluş, M., & Duru, S., 2011, Tükenmişliğin Öğretmen Adayları Arasındaki yaygınlığı, Demografik Değişkenler ve Akademik Başarı ile İlişkisi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 151-165.

Balogun, J., Hoeberlein, T., J., K., & Schneider, E., 1999, Pattern of physical therapist students' burnout within an academic semester. *Journal of Physical Therapy Education, 13*(1), 12-17.

Başar, H., 2001, *Sınıf Yönetimi.* Ankara: PegemA Yayınları.

Başarır, D., 1990, Ortaokul Son Sınıf Öğrencilerinde Sınav Kaygısı, Durumluk Kaygı, Akademik Başarı ve Sınav Başarısı Arasındaki İlişkiler. *Yayımlanmamış yüksek lisans tezi*. Ankara: Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü.

Bean, R., Bush, K., McKenry, P., & Wilson, S., 2003, The İmpact Of Parental, Support, Behavioralcontrol And Psychological Control On The Academic Achievement

And Self-Esteem Of African-American Andeuropean-American Adolescents. *Journal of Adolescent Research, 18*(5), 523–541.

Becker, R., Chambers, J. M., & Wilks, A., 1988, *The (new) S language: A programming Environment For Data Analysis And Graphics.* Wadsworth & Brooks/Cole.

Bishop, C. M.,1995, *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press.

Blum, A., Kalai, A., & Langford, J., 1999, Beating the hold-out: Bounds for k-fold and progressive cross-validation. *In Proceedings of the twelfth annual conference on Computational learning theory*, (s. 203-208).

Boyd, F. B., 2002, Motivation to Continue: Enhancing Literacy Learning For Struggling Readers And Writers. *Reading and Writing Quarterly: Overcoming Learning Difficulties, 18*, 257-277.

Bozanoğlu, İ., 2004, Akademik Güdülenme Ölçeği: Geliştirmesi, Geçerliği, Güvenirliği. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 37*(2), 83-89.

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C., 2000, Intellectual ability, Learning Style, Personality, Achievement Motivation And Academic Success Of Psychology Students İn Higher Education. *Personality and Individual differences, 29*(6), 1057-1068.

Büyüköztürk, Ş., Çakan, M., Tan, Ş., & Atar, H. Y., 2014, *TIMSS 2011 Ulusal Matematik ve Fen Raporu: 4. Sınıflar.* TC MEB YEĞİTEK Genel Müdürlüğü: Anakra.

Büyüköztürk, Ş., Çakan, M., Tan, Ş., & Atar, H. Y., 2014, *TIMSS 2011 Ulusal Matematik ve Fen Raporu: 8. Sınıflar* . TC MEB YEĞİTEK Genel Müdürlüğü: Ankara.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A., 1998, *Discovering Data Mining: From Concepts to Implementation.* Upper Saddle River, NJ: Prentice Hall.

Cassady, J. C., & Johnson, R. E., 2002, Cognitive Test Anxiety And Academic Performance. *Contemporary Educational Psychology, 27*, 270-295.

Chapell, M., Blanding, Z., Silverstein, M., Takahashi, M., Newman, B., Gubi, A., & McCann, N., 2005, Test Anxiety And Academic Performance İn Undergraduate And Graduate Students. *Journal of Educational Psychology, 97*, 268-274.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., 2000, *CRISP-DM 1.0 Step-by-step data mining guide.* SPSS.

CRAN., 1999, *The Comprehensive R Network.* Mayıs 5, 2015 tarihinde CRAN: http://cran.r-project.org adresinden alındı

Çapri, B., 2006, Tükenmişlik Ölçeğinin Türkçe Uyarlaması: Geçerlik Ve Güvenirlik Çalışması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 2*(1), 62-78.

Çapulcuoğlu, U., & Gündüz, B., 2013, Lise Öğrencilerinde Tükenmişliğin Cinsiyet, Sınıf Düzeyi, Okul Türü Ve Algılanan Akademik Başarı Değişkenlerine Göre İncelenmesi. *Trakya Üniversitesi Eğitim Fakültesi Dergisi, 3*(1), 12-24.

Çelenk, S., 2003, Okul Aile İşbirliği ve Okuduğunu Anlama Başarısı Arasındaki İlişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*(24), 33-39.

Çınar, İ., 2014, Eğitim ve Otoriteye Bağlılık. *Eğitişim Dergisi*(42).

Çokluk, Ö., 2010, Lojistik Regresyon Analizi: Kavram Ve Uygulama. *Kuram ve Uygulamada Eğitim Bilimleri*, 1357-1407.

Çomak, E., & Güner, N., 2011, Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi. *Pamukkale University Journal of Engineering Sciences, 17*(2), 87-96.

Da Rocha, B. C., & de Sousa Júnior, R. T., 2010, Identifying Bank Frauds Using Crisp-DM And Decision Trees. *International journal of computer science & information Technology*, 162-169.

De Raad, B., & Schouwenburg, H., 1996, Personality in Learning And Education: A Review. *European Journal of Personality*, 303-336.

Demir, C., 2009, Factors Influencing The Academic Achievement of The Turkish Urban Poor. *International Journal of Educational Development*, 17-29.

Diamond, J., Randolph, A., & Spillane, J., 2004, Teachers' Expectations And Sense Of Responsibility For Students Learning: The İmportance Of Race, Class And Organizational Habitus. *Anthropology & Education Quarterly, 35*(1), 75-98.

Ekici, G., 2002, Öğrencilerin Öğrenme Stillerine Dayalı Eğitim-Öğretim Nedir? Ve Bu Uygulamada Öğretmenlere Düşen Görevler Nelerdir? *TSE Standard Ekonomik ve Teknik Dergi, 481*, 21-25.

Enas, G. G., & Choi, S. C., 1986, Choice Of The Smoothing Parameter And Efficiency Of K-Nearest Neighbor Classification. *Computers & Mathematics with Applications, 12*(2), 235-244.

Epstein, J. L., 2008, Improving family And Community İnvolvement İn Secondary Schools. *Principal Leadership, 8*(2), 16-22.

ERG., 2011, Aralık 6, 2015 tarihinde Türkiye Eğitim Sisteminin Öncelikli Sorunları ve İlinizde Eğitimin Durumu: http://erg.sabanciuniv.edu/sites/erg.sabanciuniv.edu/files/InfografikDosya.16.10.11.pdf adresinden alındı

Ertürk, S., 1973, *Eğitimde Program Geliştirme.* Ankara: Yelkentepe Yayınları.

Faulkner, R., Davidson, J. W., & McPherson, G. E., 2010, The Value Of Data Mining İn Music Education Research And Some Findings From İts Application To A Study Of İnstrumental Learning During Childhood. *International Journal of Music Education*, 212-230.

Fayyad, U., Pitatesky-Shapiro, G., Smyth, P., & Uthurasamy, R., 1996, *Advances in Knowledge Discovery and Data Mining.* A A A I/M IT Press.

Fimian, M. J., 1988, Predictors of Classroom Stress And Burnout Experienced By Gifted And Talented Students. *Psychology In the Schools, 5*(4), 392-405.

Gail, M., Krickeberg, K., Samet, J. M., Tsiatis, A., & Wong, W., 2010, *Statistics for Biology and Health.* London: Springer (3rd Editon).

Gartner Group, 2013, *IT Glossary*. Nisan 8, 2015 tarihinde Gartner: http://www.gartner.com/it-glossary/data-mining adresinden alındı

Goldberg, M. D., & Cornell, D. G., 1998, The Influence Of Intrinsic Motivation And Self-Concept On Academic Achievement In Second- And Third-Grade Students. *Journal of Education of the Gifted, 21*, 179-205.

Gonzales-Pienda, J. A., Nunez, J. C., Gonzales-Pumariega, S., Alvarez, L., Roces, C., & Pat Garcia, M., 2002, A Structural Equation Model of Parental Involvement, Motivational and Aptitudinal Characteristics, and Academic Achievement. *The Journal of Experimental Education, 70*(3), 257-287.

Gordon, R., Kane, T., & Staiger, D., 2006, *Hamilton: The Project, İdentifying Effective Teachers Using Performance On The Job.* Brookings Institute.

Grunsky, E. C., 2002, R: A Data Analysis And Statistical Programming Environment An Emerging Tool For The Geosciences. *Computers & Geosciences* , 1219-1222.

Gupta, G. K., 2014, *Introduction to Data Mining with Case Studies.* Delhi: PHI Learning Pvt. Ltd.

Guskey, T., 2002, Professional Development And Teacher Change. *Teachers and Teaching: Theory and Practice, 8*((3-4)), 381-391.

Gülen, Ö., & Özdemir, S., 2013, Veri Madenciliği Teknikleri İle Üstün Yetenekli Öğrencilerin İlgi Alanlarının Analizi. *stün Yetenekliler Eğitimi Araştırmaları Dergisi-Journal of Gifted Education Research.*

Günçer, B., & Köse, R., 1993, Effects Of Family And School On Turkish Students' Academic Performance. *Education and Society, 11*(1), 51-63.

Gürsoy, M., 2013, Destek Vektör Makineleri ile Hibrid Modelleme:Menkul Kıymet Getirilerindeki Volatilitenin Tahminlenmesi. *Yayınlanmamış Doktora Tezi*. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.

Hair, J., Black, W. C., Babin, B., Anderson, R. E., & Tatham, R. L., 2006, *Multivariate Data Analysis.* Upper Saddle River, NJ: Prentice Hall.

Han, J., Kamber, M., & Pei, J., 2012, *Data Mining Concepts and Techniques* (3nd Edition b.). Amsterdam: Elsevier; Morgan Kauffman.

Hand, D., Mannila, H., & Smyth, P., 2001, *Principles o f Data Mining.* Cambridge: MA: MIT Press.

Harris, D., 1940, Factors Affecting College Grades: A Review Of The Literature, 1930–1937. *Psychological Bulletin*, 125-166.

Hen, L., & Lee, S., 2008, Performance Analysis Of Data Mining Tools Cumulating With A Proposed Data Mining Middleware. *Journal of Computer Science*, 826-833.

Hisli, N., 1989, Beck Depresyon Envanterinin Üniversite Öğrencileri İçin Geçerliği ve Güvenirliği. *Psikoloji Dergisi, 7*, 3-13.

Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Meyer, D., & Zeleis, A., 2015, *Package 'RWeka.* 03 03, 2015 tarihinde CRAN: https://cran.r-project.org/web/packages/RWeka/RWeka.pdf adresinden alındı

Hothorn, T., & Zeileis, A., 2015, *Partykit: A Modular Toolkit for Recursive Partytioning in R.* 10 1, 2015 tarihinde CRAN: https://cran.r-project.org/web/packages/partykit/partykit.pdf adresinden alındı

Hung, J. L., Rice, K., & Saba, A., 2012, An Educational Data Mining Model For Online Teaching And Learning. *Journal of Educational Technology Development and Exchange, 5*(2), 77-94.

IEEE, 2015, *IEEE Top Programming Languages: Design, Methods, and Data Sources*. Kasım 4, 2015 tarihinde IEEE: http://spectrum.ieee.org/ns/IEEE_TPL_2015/methods.html adresinden alındı

İnan, A., 2003, Privacy Preserving Distributed Spatio-Temporal Data Mining. *Yüksek Lisans Tezi*. Sabancı University, Computer Science and Engineering.

Jacobs, S., & Dodd, D., 2003, Student Burnout As A Function Of Personality, Social Support, And Workload. *Journal of College Students Development, 44*(3), 291-303.

Jones, K., 2012, *What Is The Purpose Of Education?* Aralık 5, 2015 tarihinde Forbes: http://www.forbes.com/sites/sap/2012/08/15/what-is-the-purpose-of-education/ adresinden alındı

Kantardzic, M., 2011, *Data Mining: Concepts, Models, Methods, and Algorithms.* New Jersey: John Wiley & Sons.

Kemer, G., 2006, Yayımlanmamış yüksek lisans tezi. *Öz-Yeterlilik, Umut ve Kaygının Onbirinci Sınıf Öğrencilerinin Üniversite Giriş Sınavı Puanlarını Yordamadaki Rolü*. Ankara: Orta Doğu Teknik Üniversitesi.

Keser, İ., 2007, İzmir'deki Özel Ve Devlet Üniversitelerindeki Öğrencilerin Başarılarını Etkileyen Faktörlerin Belirlenmesi Ve Karşılaştırılması. *Muğla Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 18*, 39-48.

Kohavi, R., 1995, A Study Of Cross-Validation And Bootstrap For Accuracy Estimation And Model Selection. *Ijcai*, 1137-1145.

Kovacic, Z., 2010, Early prediction of student success: Mining Students' Enrolment Data. *nforming Science + Information Technology Education Joint Conference*, (s. 647-665). Cassino, Italy.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Candan, C., 2015, *Caret: Classification and Regression Training.* 11 26, 2015 tarihinde CRAN: https://cran.r-project.org/web/packages/caret/index.html adresinden alındı

Kumar, D., 2011, Performance Analysis Of Vairous Data Minin Algortihms: A Review. *International Journal of Computer Applications* , 9-15.

Kurniawan, Y., & Halim, E., 2013, Data Warehouse and Data Mining to Predict Student Academic Performance in Schools: A Case Study. *Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference* (s. 98-103). IEEE.

Külahoğlu, Ş. Ö., 2001, Öğrenci Davranışını Etkileyen Sosyal ve Psikolojik Faktörler. L. Küçükahmet içinde, *Sınıf Yönetiminde Yeni Yaklaşımlar.* Ankara: Nobel Yayınları.

Larose, D. T., & Larose, C. D., 2014, *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley&Sons.

Lavrac, N., Motoda, H., Fawcett, T., Holte, R., Langley, P., & Adriaans, P., 2004, Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving. *Machine Learning*, 13-34.

Lloyd, K. M., Tienda, M., & Zajacova, A., 2001, Trends in educational achievement of minority students since Brown v. Board of Educatio. C. (. Snow içinde, *Achieving High Educational Standards for All: Conference Summary* (s. 147-182). Washington DC: National Academy Press.

Maslach, C., & Jackson, S., 1981, The Measurement Of Experienced Burnout. *Journal of Occupational Behavior, 2*, 99-113.

Maton, K., & Hrabowski III, F., 1998, Preparing the way: A Qualitative Study Of High Achieving African American Males And The Role Of The Family. *American Journal of Community Psychology, 26*(4), 639-668.

McCarthy, M., Pretty, G., & Catano, V., 1990, Psychological sense Of Community And Student Burnou. *Journal of College Student Development, 31*, 211-216.

MEB, 2015, *TC Milli Eğitim Bakanlığı 2015-2019 Stratejik Planı.* 12 09, 2015 tarihinde T.C. Millî Eğitim Bakanlığı Strateji Geliştirme Başkanlığı: http://sgb.meb.gov.tr/meb_iys_dosyalar/2015_09/10052958_10.09.2015sp17.15imzasz.pdf adresinden alındı

Mertler, C. A., & Vannatta, R. A., 2005, *Advanced and Multivariate Statistical Methods: Practical Application And İnterpretation.* Glendale, CA: Pyrczak Publishing.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F., 2014, *E1071: Misc Functions of the Department of Statistics (e1071).* Haziran 23, 2015 tarihinde e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien: http://CRAN.R-project.org/package=e1071 adresinden alındı

Moro, S., Laureano, R., & Cortez, P., 2011, Using Data Mining For Bank Direct Marketing: An Application Of The CRISP-DM Methodology. *Proceedings of European Simulation and Modelling Conference-ESM'2011* (s. 117-121). Eurosis.

Müssel, C., Lausser, L., Maucher, M., & Kestler H. A., 2012, Multi-Objective Parameter Selection for Classifiers. *Journal of Statistical Software, 46*(5), 1-27.

Owen, S., 1988, Development of A College Academic Self-Efficacy Scale. *Annual Meeting of the National Council on Measurement In Education.* New Orleans, LA: ERIC.

Öner, N., 1978, Türkçe'yeUyarlanmış Bir Kaygı Envanterinin Geçerlik Denemesi; Bir Araştırma Özeti. *Psikoloji Dergisi, 1*(1), 15.

Özgüven, İ. E., 1974, *Akademik Başarıyı Etkileyen Zihinsel Olmayan Faktörler.* Ankara.

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

Ponzetti, J., & Gate, R. M., 1981, Sex Differences in The Relationship Between Loneliness And Academic Performance. *Psychological Reports, 48*, 759-768.

Popham, j., 2003, The Seductive Allure Of Data. *Educational Leadership, 5*(60), 48-51.

Razon, N., 1987, Öğrenme Olgusu ve Okul Başarısını Etkileyen Faktörler. *Eğitim Bilimleri Dergisi, 11*(63), 17.

Ripley, B., & Venables, W., 2015, *Class: Functions for Classification.* 09 7, 2015 tarihinde CRAN: https://cran.r-project.org/web/packages/class/index.html adresinden alındı

Romanski, P., & Kothoff, L., 2015, *Package 'FSelector'.* Nisan 21, 2015 tarihinde CRAN: https://cran.r-project.org/web/packages/FSelector/FSelector.pdf adresinden alındı

Sama, E., & Tarım, K., 2007, Öğretmenlerin Başarısız Olarak Algıladıkları Öğrencilere Yönelik Tutum ve Davranışları. *Türk Eğitim Bilimleri Dergisi, 5*(1), 135-154.

Santen, S. A., Holt, D. B., Kemp, J., & Hemphill, R. R., 2010, Burnout in Medical Students: Examining the Prevalence and Associated Factors. *Southern Medical Journal, 103*(8), 758-763.

Schaufeli, W. B., & Salanova, M., 2007, Efficacy or Inefficacy, That's The Question: Burnout And Engament, And Their Relationships With Efficacy Beliefs. *Anxiety, Coping & Stress*, 177-196.

Schaufeli, W. B., Martinez, I., Marques-Pinto, A., Salanova, M., & Bakker, A., 2002, Burnout and Engagement in University Students. *Jurnal of Cross-Cultural Psychology, 33*(5), 464-481.

Silah, M., 2003, Üniversite Öğrencilerinin Akademik Başarılarını Etkişeyen Çeşitli Nedenler Arasından Süreksiz Durumluk Kaygının Yeri ve Önemi. *Eğitim Araştırmaları Dergisi*, 102-115.

Silahtaroğlu, G., 2008, *Kavram ve Algoritmalarıyla Temel Veri Madenciliği.* İstanbul: Papatya Yayıncılık.

Soria, D., Garibaldi, J. M., Ambrogi, F., Baiganzoli, E. M., & Ellis, I. O., 2011, A Non-Parametric Version of The Naive Bayes Classifier. *Knowledge Based Systems, 24*(6), 775-784.

Stein, M., & Wang, M., 1988, Teacher Development And School Improvement: The Process Of Teacher Change. *Teaching & Teacher Education, 4*(2), 171-187.

Streifer, P., 2002, *Using Data To Make Better Educational Decisions.* Lanham, MD: Scarecrow Press.

Şama, E., & Tarım, K., 2007, Öğretmenlerin Başarısız Olarak Algıladıkları Öğrencilere Yönelik Tutum Ve Davranışları. *Türk Eğitim Bilimleri Dergisi, 5*(1), 135-154.

Şengür, D., & Tekin, A., 2013, Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini. *International Journal Of Informatics Technologies,*, 7-16.

Tan, P., Steinbach, M., & Kumar, V., 2006, *Introduction to Data Mining.* Boston: Pearson Education.

Therneau, T., Atkinson, B., & Ripley, B., 2015, *Package 'rpart'.* 08 12, 2015 tarihinde CRAN: https://cran.r-project.org/web/packages/rpart/rpart.pdf adresinden alındı

Theuwissen, M., 2015, *R vs Python for Data Science*. Kasım 29, 2015 tarihinde KDnuggets: http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html adresinden alındı

Tinto, V., 1994, *Leaving College: Rethinking the Causes and Cures of Student Attrition.* Chicago: University of Chicago Press.

Tufferry, S., 2011, *Data Mining and Statistics for Desicion Making.* John Wiley & Sons.

Tural, N., 2002, *Eğitim finansmanı.* Ankara: Anı Yayıncılık.

Türnüklü, A., Zoroğlu, Y., & Gemici, Y., 2001, İlköğretim Okullarında Okul Yönetimine Yansıyan Disiplin Sorunları. *Kuram ve Uygulamada Eğitim Yönetim Dergisi*, 417-441.

Venter, J., de Waal, A., & Willers, C., 2007, Specializing CRISP-DM for evidence mining. J. Venter, A. de Waal, & C. Willers içinde, *Advances in Digital Forensics* (s. 303-315). New York: Springer.

Witten, I. H., & Frank, E., 2000, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Diego: Morgan Kaufmann Publishers.

World Bank, 2011, *Improving th Quality and Equity of Basic Education in Turkey: Challenges and Options.* Washington DC: World Bank.

Yang, H. J., 2004, Factors Affecting Student Burnout and Academic Achievement in Multiple Enrollment Programs in Taiwan's Technical-Vocationa Colleges. *Internationaş Journal of Educational Development*, 283-301.

Yang, Y., & Liu, X., 1999, A Re-Examination Of Text Categorization Methods. *Proceedings of SIGIR-99* (s. 42-49). Berkeley: 22nd ACM International Conference on Research and Development in Information Retrieval.

Yazıcıoğlu, Y., & Erdoğan, S., 2004, *SPSS Uygulamalı Bilimsel Araştırma Yöntemleri.* Ankara: Detay.

Yıldırım, H. H., Yıldırım, S., Yetişir, M. İ., & Ceylan, E., 2013, *PISA 2012 Ulusal Ön Raporu.* TC MEB YEĞİTEK Genel Müdürlüğü: Ankara.

Yıldırım, İ., 2006, Akademik Başarının Yordayıcısı Olarak Gündelik Sıkıntılar Ve Sosyal Destek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi,, 30*, 258-267.

Yıldırım, İ., & Ergene, T., 2003, Lise Son Sınıf Öğrencilerinin Akademik Başarılarının Yordayıcısı Olarak Sınav Kaygısı, Boyun Eğici Davranışlar Ve Sosyal Destek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 25*, 224-234.

Yıldız, O., 2014, Makine Öğrenmesi İle Uzaktan Eğitim Öğrencilerinin Performanlarının Değerlendirilmesi. *Yayınlamamış Doktora Tezi*. İstanbul Üniversitesi Fen Bilimleri Enstitüsü.

Yılmaz, E., 2007, Ortaöğretimde İngilizce Derslerinde Öğrenci Başarısında Motivasyonun Rolü: Bartın İli Örneği. *Yayımlanmamış yüksek lisans Tezi*. Zonguldak: Zonguldak Karaelmas Üniversitesi.

Yumuşak, N., 2006, Predicting Academic Achievement With Cognitive and Motivational Variables. *Unpublished master thesis, .* Ankara.: METU.

Zellman, G. L., 1998, Understanding the Impact of Parental School Involvement on Children's Educational Outcomes. *Journal of Educational Research, 91*(6), 370-381.

EKLER

Ek 1: k=12 için k-kat çapraz geçerleme yönteminde elde edilen performans değerleri

	KNN	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	Kat 1	0,89097	0,10903	0,96671	0,56790	0,90515	0,80000	0,08206	0,03329	11,78011	0,05861	200,98907	0,93492
	Kat 2	0,86518	0,13482	0,94211	0,53704	0,89669	0,68504	0,08792	0,05789	10,71496	0,10779	99,40609	0,91884
4 kat	Kat 1	0,88993	0,11007	0,95954	0,59259	0,90959	0,77419	0,07728	0,04046	12,41583	0,06828	181,83614	0,93390
	Kat 2	0,85246	0,14754	0,94798	0,44444	0,87936	0,66667	0,10539	0,05202	8,99525	0,11705	76,84829	0,91238
	Kat 3	0,88732	0,11268	0,95652	0,59259	0,90909	0,76190	0,07746	0,04348	12,34783	0,07337	168,29630	0,93220
	Kat 4	0,88263	0,11737	0,95942	0,55556	0,90191	0,76271	0,08451	0,04058	11,35314	0,07304	155,42989	0,92978
5 kat	Kat 1	0,85965	0,14035	0,94946	0,47692	0,88552	0,68889	0,09942	0,05054	9,55044	0,10597	90,12043	0,91638
	Kat 2	0,89736	0,10264	0,95307	0,65625	0,92308	0,76364	0,06452	0,04693	14,77256	0,07151	206,56731	0,93783
	Kat 3	0,86510	0,13490	0,94203	0,53846	0,89655	0,68627	0,08798	0,05797	10,70773	0,10766	99,45833	0,91873
	Kat 4	0,88856	0,11144	0,95652	0,60000	0,91034	0,76471	0,07625	0,04348	12,54515	0,07246	173,12308	0,93286
	Kat 5	0,88856	0,11144	0,94928	0,63077	0,91608	0,74545	0,07038	0,05072	13,48762	0,08042	167,72079	0,93238
10 kat	Kat 1	0,87135	0,12865	0,95683	0,50000	0,89262	0,72727	0,09357	0,04317	10,22617	0,08633	118,45313	0,92361
	Kat 2	0,86550	0,13450	0,93525	0,56250	0,90278	0,66667	0,08187	0,06475	11,42343	0,11511	99,24107	0,91873
	Kat 3	0,85380	0,14620	0,92754	0,54545	0,89510	0,64286	0,08772	0,07246	10,57391	0,13285	79,59273	0,91103
	Kat 4	0,90058	0,09942	0,97826	0,57576	0,90604	0,86364	0,08187	0,02174	11,94876	0,03776	316,46104	0,94077
	Kat 5	0,90643	0,09357	0,96377	0,66667	0,92361	0,81481	0,06433	0,03623	14,98221	0,05435	275,67273	0,94326
	Kat 6	0,83041	0,16959	0,93478	0,39394	0,86577	0,59091	0,11696	0,06522	7,99239	0,16555	48,27727	0,89895
	Kat 7	0,91176	0,08824	0,95652	0,71875	0,93617	0,79310	0,05294	0,04348	18,06763	0,06049	298,68056	0,94624
	Kat 8	0,88824	0,11176	0,97101	0,53125	0,89933	0,80952	0,08824	0,02899	11,00483	0,05456	201,69792	0,93380
	Kat 9	0,89412	0,10588	0,94203	0,68750	0,92857	0,73333	0,05882	0,05797	16,01449	0,08432	189,92188	0,93525

118

	Kat 10	0,89412	0,10588	0,97101	0,56250	0,90541	0,81818	0,08235	0,02899	11,79089	0,05153	228,81696	0,93706

Ek 2: k=12 için k-kat çapraz geçerleme yönteminde 2,4, 5 ve 10 katlarda ortalama performans değerleri

119

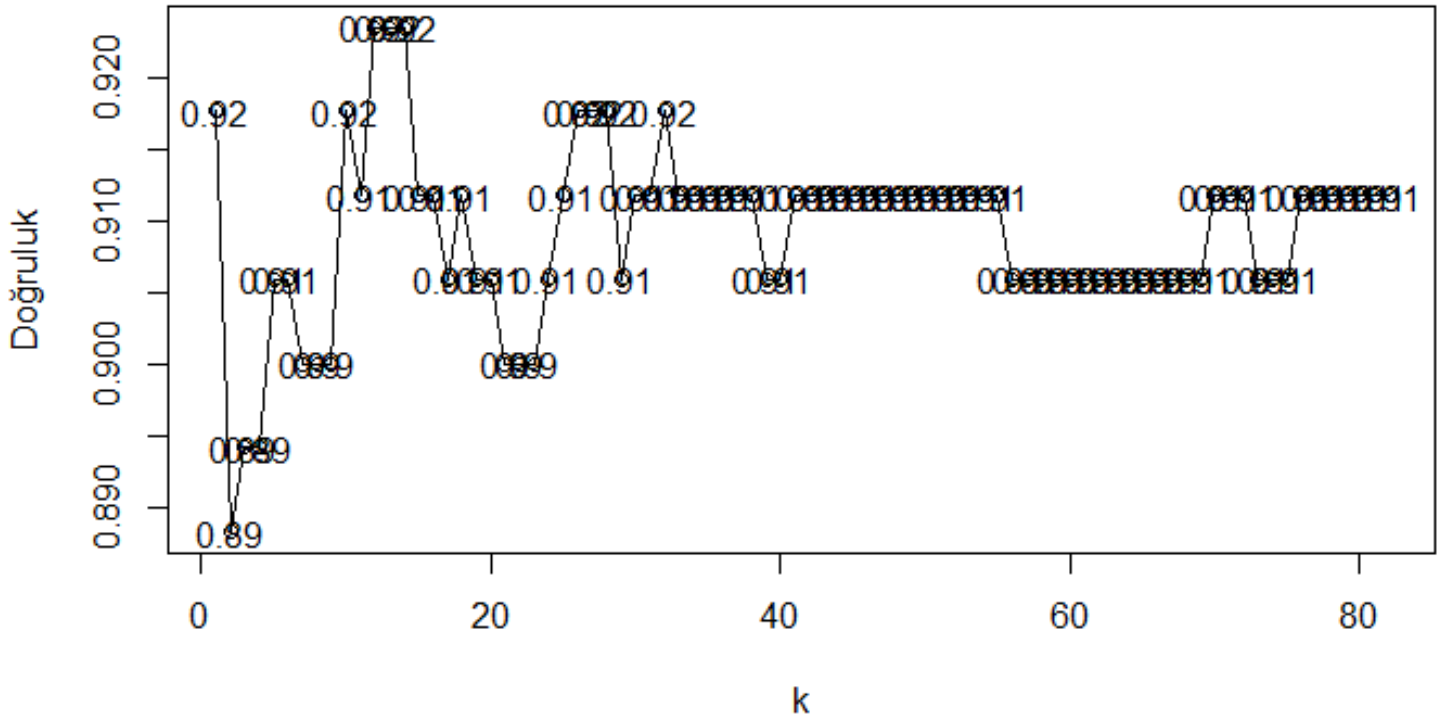
2 kat	0,87808	0,12192	0,95441	0,55247	0,90092	0,74252	0,08499	0,04559	11,24754	0,08320	150,19758	0,92688
4 kat	0,87809	0,12191	0,95586	0,54630	0,89999	0,74137	0,08616	0,04414	11,27801	0,08294	145,60265	0,92706
5 kat	0,87985	0,12015	0,95007	0,58048	0,90632	0,72979	0,07971	0,04993	12,21270	0,08761	147,39799	0,92764
10 kat	0,88163	0,11837	0,95370	0,57443	0,90554	0,74603	0,08087	0,04630	12,40247	0,08428	185,68153	0,92887
kFold ort	**0,87941**	**0,12059**	**0,95351**	**0,56342**	**0,90319**	**0,73993**	**0,08293**	**0,04649**	**11,78518**	**0,08451**	**157,21994**	**0,92761**

Ek 3: k=12 için holdout(dışarıda bırak) yönteminde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 ayırımlarda performans değerleri

k=12	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure	AUC
0.95	0,87059	0,12941	0,97101	0,43750	0,88158	0,77778	0,10588	0,02899	9,17069	0,06625	138,42014	0,92414	0,70426
0.90	0,91765	0,08235	0,95652	0,75000	0,94286	0,80000	0,04706	0,04348	20,32609	0,05797	350,62500	0,94964	0,85326
0.85	0,88627	0,11373	0,97101	0,52083	0,89732	0,80645	0,09020	0,02899	10,76560	0,05565	193,44429	0,93271	0,74592
0.80	0,89412	0,10588	0,96739	0,57813	0,90816	0,80435	0,07941	0,03261	12,18196	0,05640	215,97608	0,93684	0,77276
0.75	0,86150	0,13850	0,93623	0,54321	0,89722	0,66667	0,08685	0,06377	10,77932	0,11739	91,82382	0,91631	0,73972
0.70	0,86301	0,13699	0,94444	0,51546	0,89269	0,68493	0,09198	0,05556	10,26832	0,10778	95,27309	0,91784	0,72995
ORT.	**0,88219**	**0,11781**	**0,95777**	**0,55752**	**0,90331**	**0,75670**	**0,08356**	**0,04223**	**12,24866**	**0,07691**	**180,92707**	**0,92958**	**0,75765**

120

Ek 4: Farklı k değerleri için doğruluk değerleri



121

Ek 5: 1 ile 82 Arasındaki Tüm k Değerleri İçin Doğruluk Değerleri Tablosu

[1] "2'deki dogruluk değeri: 0.8882"	[1] "42'deki dogruluk değeri: 0.9118"
[1] "3'deki dogruluk değeri: 0.8941"	[1] "43'deki dogruluk değeri: 0.9118"
[1] "4'deki dogruluk değeri: 0.8941"	[1] "44'deki dogruluk değeri: 0.9118"
[1] "5'deki dogruluk değeri: 0.9059"	[1] "45'deki dogruluk değeri: 0.9118"
[1] "6'deki dogruluk değeri: 0.9059"	[1] "46'deki dogruluk değeri: 0.9118"
[1] "7'deki dogruluk değeri: 0.9"	[1] "47'deki dogruluk değeri: 0.9118"
[1] "8'deki dogruluk değeri: 0.9"	[1] "48'deki dogruluk değeri: 0.9118"
[1] "9'deki dogruluk değeri: 0.9"	[1] "49'deki dogruluk değeri: 0.9118"
[1] "10'deki dogruluk değeri: 0.9176"	[1] "50'deki dogruluk değeri: 0.9118"
[1] "11'deki dogruluk değeri: 0.9118"	[1] "51'deki dogruluk değeri: 0.9118"
[1] "12'deki dogruluk değeri: 0.9235"	[1] "52'deki dogruluk değeri: 0.9118"
[1] "13'deki dogruluk değeri: 0.9235"	[1] "53'deki dogruluk değeri: 0.9118"
[1] "14'deki dogruluk değeri: 0.9235"	[1] "54'deki dogruluk değeri: 0.9118"
[1] "15'deki dogruluk değeri: 0.9118"	[1] "55'deki dogruluk değeri: 0.9118"
[1] "16'deki dogruluk değeri: 0.9118"	[1] "56'deki dogruluk değeri: 0.9059"
[1] "17'deki dogruluk değeri: 0.9059"	[1] "57'deki dogruluk değeri: 0.9059"
[1] "18'deki dogruluk değeri: 0.9118"	[1] "58'deki dogruluk değeri: 0.9059"
[1] "19'deki dogruluk değeri: 0.9059"	[1] "59'deki dogruluk değeri: 0.9059"
[1] "20'deki dogruluk değeri: 0.9059"	[1] "60'deki dogruluk değeri: 0.9059"
[1] "21'deki dogruluk değeri: 0.9"	[1] "61'deki dogruluk değeri: 0.9059"
[1] "22'deki dogruluk değeri: 0.9"	[1] "62'deki dogruluk değeri: 0.9059"
[1] "23'deki dogruluk değeri: 0.9"	[1] "63'deki dogruluk değeri: 0.9059"
[1] "24'deki dogruluk değeri: 0.9059"	[1] "64'deki dogruluk değeri: 0.9059"
[1] "25'deki dogruluk değeri: 0.9118"	[1] "65'deki dogruluk değeri: 0.9059"
[1] "26'deki dogruluk değeri: 0.9176"	[1] "66'deki dogruluk değeri: 0.9059"
[1] "27'deki dogruluk değeri: 0.9176"	[1] "67'deki dogruluk değeri: 0.9059"
[1] "28'deki dogruluk değeri: 0.9176"	[1] "68'deki dogruluk değeri: 0.9059"
[1] "29'deki dogruluk değeri: 0.9059"	[1] "69'deki dogruluk değeri: 0.9059"
[1] "30'deki dogruluk değeri: 0.9118"	[1] "70'deki dogruluk değeri: 0.9118"
[1] "31'deki dogruluk değeri: 0.9118"	[1] "71'deki dogruluk değeri: 0.9118"

[1] "32'deki dogruluk değeri: 0.9176"	[1] "72'deki dogruluk değeri: 0.9118"
[1] "33'deki dogruluk değeri: 0.9118"	[1] "73'deki dogruluk değeri: 0.9059"
[1] "34'deki dogruluk değeri: 0.9118"	[1] "74'deki dogruluk değeri: 0.9059"
[1] "35'deki dogruluk değeri: 0.9118"	[1] "75'deki dogruluk değeri: 0.9059"
[1] "36'deki dogruluk değeri: 0.9118"	[1] "76'deki dogruluk değeri: 0.9118"
[1] "37'deki dogruluk değeri: 0.9118"	[1] "77'deki dogruluk değeri: 0.9118"
[1] "38'deki dogruluk değeri: 0.9118"	[1] "78'deki dogruluk değeri: 0.9118"
[1] "39'deki dogruluk değeri: 0.9059"	[1] "79'deki dogruluk değeri: 0.9118"
[1] "40'deki dogruluk değeri: 0.9059"	[1] "80'deki dogruluk değeri: 0.9118"
[1] "41'deki dogruluk değeri: 0.9118"	[1] "81'deki dogruluk değeri: 0.9118"
	[1] "82'deki dogruluk değeri: 0.9118"

Ek 6: Naive (Basit) Bayes Sınıflandırıcıda k-kat çapraz geçerleme yönteminde elde edilen performans değerleri

	NB	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	Kat 1	0,90973	0,09027	0,93488	0,80247	0,9528	0,74286	0,03751	0,06512	24,92031	0,08115	307,0771	0,94375
	Kat 2	0,8898	0,1102	0,93632	0,69136	0,92826	0,71795	0,05862	0,06368	15,97369	0,0921	173,4338	0,93228
4 kat	Kat 1	0,92037	0,07963	0,95376	0,77778	0,94828	0,79747	0,04215	0,04624	22,62524	0,05945	380,544	0,95101
	Kat 2	0,8993	0,1007	0,94509	0,7037	0,93162	0,75	0,05621	0,05491	16,81467	0,07803	215,4769	0,93831
	Kat 3	0,90845	0,09155	0,92754	0,82716	0,95808	0,72826	0,03286	0,07246	28,2236	0,08761	322,1672	0,94256
	Kat 4	0,88498	0,11502	0,92464	0,71605	0,93275	0,69048	0,05399	0,07536	17,1259	0,10525	162,7204	0,92868
5 kat	Kat 1	0,87719	0,12281	0,92058	0,69231	0,92727	0,67164	0,05848	0,07942	15,74188	0,11472	137,2185	0,92391
	Kat 2	0,90909	0,09091	0,94585	0,75	0,94245	0,7619	0,04692	0,05415	20,15839	0,0722	279,1938	0,94414
	Kat 3	0,88856	0,11144	0,92754	0,72308	0,93431	0,70149	0,05279	0,07246	17,57166	0,10022	175,3381	0,93091
	Kat 4	0,91789	0,08211	0,9529	0,76923	0,94604	0,79365	0,04399	0,0471	21,66256	0,06123	353,7791	0,94946
	Kat 5	0,92375	0,07625	0,94203	0,84615	0,96296	0,77465	0,02933	0,05797	32,12319	0,06851	468,875	0,95238
10 kat	Kat 1	0,87719	0,12281	0,90647	0.75	0,9403	0,64865	0,04678	0,09353	19,3759	0,1247	155,3798	0,92308
	Kat 2	0,89474	0,10526	0,92806	0,75	0,94161	0,70588	0,04678	0,07194	19,83723	0,09592	206,8031	0,93478
	Kat 3	0,90643	0,09357	0,94928	0,72727	0,93571	0,77419	0,05263	0,05072	18,03623	0,06975	258,5974	0,94245
	Kat 4	0,88889	0,11111	0,95652	0,60606	0,91034	0,76923	0,07602	0,04348	12,58194	0,07174	175,3846	0,93286
	Kat 5	0,91813	0,08187	0,92754	0,87879	0,9697	0,74359	0,02339	0,07246	39,65217	0,08246	480,8727	0,94815
	Kat 6	0,87135	0,12865	0,92754	0,63636	0,91429	0,67742	0,07018	0,07246	13,21739	0,11387	116,0727	0 ,92086
	Kat 7	0,92353	0,07647	0,96377	0,75	0,94326	0,82759	0,04706	0,03623	20,48007	0,04831	423,9375	0,95341
	Kat 8	0,88235	0,11765	0,91304	0,75	0,9403	0,66667	0,04706	0,08696	19,40217	0,11594	167,3438	0,92647
	Kat 9	0,94118	0,05882	0,94928	0.90625	0,97761	0,80556	0,01765	0,05072	53,79227	0,05597	961,0566	0,96324
	Kat 10	0,92353	0,07647	0,94203	0,84375	0,96296	0,77143	0,02941	0,05797	32,02899	0,06871	466,1719	0,95238

124

Ek 7: Naive(Basit) Bayes Sınıflandırıcıda k-kat çapraz geçerleme yönteminde 2,4, 5 ve 10 katlarda ortalama performans değerleri

	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	0,89977	0,10024	0,93560	0,74692	0,94053	0,73041	0,04807	0,06440	20,44700	0,08663	240,25542	0,93802
4 kat	0,90328	0,09673	0,93776	0,75617	0,94268	0,74155	0,04630	0,06224	21,19735	0,08259	270,22712	0,94014
5 kat	0,90330	0,09670	0,93778	0,75615	0,94261	0,74067	0,04630	0,06222	21,45154	0,08338	282,88090	0,94016
10 kat	0,90273	0,09727	0,93635	0,74278	0,94361	0,73902	0,04570	0,06365	24,84044	0,08474	341,16201	0,94187
kFold ort	**0,90227**	**0,09773**	**0,93687**	**0,75051**	**0,94236**	**0,73791**	**0,04659**	**0,06313**	**21,98408**	**0,08433**	**283,63136**	**0,94005**

125

Ek 8: Naive(Basit) Bayes Sınıflandırıcıda holdout(dışarıda bırak) yönteminde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 ayırımlarda performans değerleri

NB	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure	AUC
0.95	0,91765	0,08235	0,94203	0,81250	0,95588	0,76471	0,03529	0,05797	26,69082	0,07135	374,08854	0,94891	0,97645
0.90	0,911765	0,088235	0,942029	0,78125	0,948905	0,757576	0,041176	0,057971	22,87785	0,074203	308,3147	0,945455	0,94248
0.85	0,91373	0,08627	0,95169	0,75000	0,94258	0,78261	0,04706	0,04831	20,22343	0,06441	313,96875	0,94712	0,93841
0.80	0,911765	0,088235	0,938406	0,796875	0,952206	0,75	0,038235	0,061594	24,54292	0,077295	317,524	0,945255	0,96677
0.75	0,882629	0,117371	0,921739	0,716049	0,932551	0,682353	0,053991	0,078261	17,07221	0,109295	156,2025	0,927114	0,91791
0.70	0,87671	0,12329	0,91063	0,73196	0,93548	0,65741	0,05088	0,08937	17,89734	0,12210	146,57969	0,92289	0,93169
ort	**0,90237**	**0,09763**	**0,93442**	**0,76477**	**0,94460**	**0,73244**	**0,04444**	**0,06558**	**21,55076**	**0,08644**	**269,44636**	**0,93946**	**0,94562**

126

Ek 9: Karar Ağacı Algoritmasında k-kat çapraz geçerleme yönteminde elde edilen performans değerleri

	DT	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	Kat 1	0,91325	0,08675	0,96093	0,70988	0,93390	0,80986	0,05510	0,03907	17,43979	0,05504	316,83877	0,94722
	Kat 2	0,90152	0,09848	0,94211	0,72840	0,93669	0,74684	0,05158	0,05789	18,26414	0,07947	229,81816	0,93939
4 kat	Kat 1	0,90867	0,09133	0,95954	0,69136	0,92997	0,80000	0,05855	0,04046	16,38890	0,05853	280,02765	0,94452
	Kat 2	0,93443	0,06557	0,95954	0,82716	0,95954	0,82716	0,03279	0,04046	29,26590	0,04892	598,27337	0,95954
	Kat 3	0,93897	0,06103	0,96522	0,82716	0,95965	0,84810	0,03286	0,03478	29,37019	0,04205	698,44841	0,96243
	Kat 4	0,91784	0,08216	0,95072	0,77778	0,94798	0,78750	0,04225	0,04928	22,50048	0,06335	355,15468	0,94935
5 kat	Kat 1	0,91228	0,08772	0,95307	0,73846	0,93950	0,78689	0,04971	0,04693	19,17350	0,06355	301,69328	0,94624
	Kat 2	0,94135	0,05865	0,97834	0,78125	0,95088	0,89286	0,04106	0,02166	23,82955	0,02773	859,47731	0,96441
	Kat 3	0,92962	0,07038	0,95652	0,81538	0,95652	0,81538	0,03519	0,04348	27,18116	0,05332	509,75128	0,95652
	Kat 4	0,93548	0,06452	0,97464	0,76923	0,94718	0,87719	0,04399	0,02536	22,15676	0,03297	672,00733	0,96071
	Kat 5	0,89443	0,10557	0,92754	0,75385	0,94118	0,71014	0,04692	0,07246	19,76812	0,09613	205,64923	0,93431
10 kat	Kat 1	0,93567	0,06433	0,94245	0,90625	0,97761	0,78378	0,01754	0,05755	53,71942	0,06351	845,87109	0,95971
	Kat 2	0,94152	0,05848	0,95683	0,87500	0,97080	0,82353	0,02339	0,04317	40,90468	0,04933	829,17188	0,96377
	Kat 3	0,92982	0,07018	0,97826	0,72727	0,93750	0,88889	0,05263	0,02174	18,58696	0,02989	621,81818	0,95745
	Kat 4	0,91813	0,08187	0,96377	0,72727	0,93662	0,82759	0,05263	0,03623	18,31159	0,04982	367,56364	0,95000
	Kat 5	0,91813	0,08187	0,93478	0,84848	0,96269	0,75676	0,02924	0,06522	31,96957	0,07686	415,92727	0,94853
	Kat 6	0,91813	0,08187	0,97101	0,69697	0,93056	0,85185	0,05848	0,02899	16,60435	0,04159	399,25909	0,95035
	Kat 7	0,94706	0,05294	0,98551	0,78125	0,95105	0,92593	0,04118	0,01449	23,93375	0,01855	1290,17857	0,96797
	Kat 8	0,92941	0,07059	0,96377	0,78125	0,95000	0,83333	0,04118	0,03623	23,40580	0,04638	504,68750	0,95683
	Kat 9	0,94118	0,05882	0,97101	0,81250	0,95714	0,86667	0,03529	0,02899	27,51208	0,03567	771,19792	0,96403
	Kat 10	0,88824	0,11176	0,93478	0,68750	0,92806	0,70968	0,05882	0,06522	15,89130	0,09486	167,52083	0,93141

127

Ek 10: Karar Ağacı Algoritmasında k-kat çapraz geçerleme yönteminde 2,4, 5 ve 10 katlarda ortalama performans değerleri

	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	0,90739	0,09261	0,95152	0,71914	0,93529	0,77835	0,05334	0,04848	17,85197	0,06726	273,32846	0,94331
4 kat	0,92497	0,07503	0,95875	0,78086	0,94929	0,81569	0,04161	0,04125	24,38137	0,05321	482,97603	0,95396
5 kat	0,92263	0,07737	0,95802	0,77163	0,94705	0,81649	0,04337	0,04198	22,42182	0,05474	509,71569	0,95244
10 kat	0,92673	0,07327	0,96022	0,78438	0,95020	0,82680	0,04104	0,03978	27,08395	0,05065	621,31960	0,95500
kFold ort	**0,92043**	**0,07957**	**0,95713**	**0,76400**	**0,94546**	**0,80933**	**0,04484**	**0,04287**	**22,93477**	**0,05646**	**471,83494**	**0,95118**

128

Ek 11: Karar Ağacı Algoritmasında holdout(dışarıda bırak) yönteminde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 ayırımlarda performans değerleri

DT	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure	AUC
0.95	0,91765	0,08235	0,95652	0,75000	0,94286	0,80000	0,04706	0,04348	20,32609	0,05797	350625,00000	0,94964	0,9080616
0.90	0,94118	0,05882	0,99275	0,71875	0,93836	0,95833	0,05294	0,00725	18,75201	0,01008	1859,96528	0,96479	0,8591486
0.85	0,90980	0,09020	0,97101	0,64583	0,92202	0,83784	0,06667	0,02899	14,56522	0,04488	324,53125	0,94588	0,852808
0.80	0,93235	0,06765	0,96377	0,79688	0,95341	0,83607	0,03824	0,03623	25,20624	0,04547	554,37981	0,95856	0,9127604
0.75	0,91315	0,08685	0,95942	0,71605	0,93503	0,80556	0,05399	0,04058	17,77013	0,05667	313,56292	0,94707	0,8831812
0.70	0,91585	0,08415	0,97343	0,67010	0,92644	0,85526	0,06262	0,02657	15,54446	0,03965	392,03506	0,94935	0,8450247
ort	**0,92166**	**0,07834**	**0,96948**	**0,71627**	**0,93635**	**0,84884**	**0,05359**	**0,03052**	**18,69403**	**0,04245**	**59011,57905**	**0,95255**	**0,87683**

129

Ek 12: Logistik Regresyon analizi holdout(dışarıda bırak) yönteminde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 ayırımlarda performans değerleri

LR	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure	AUC
0.95	0,94118	0,05882	0,97101	0,81250	0,95714	0,86667	0,03529	0,02899	27,51208	0,03567	771,19792	0,96403	0,93931
0.90	0,92941	0,07059	0,96377	0,78125	0,95000	0,83333	0,04118	0,03623	23,40580	0,04638	504,68750	0,95683	0,95199
0.85	0,91765	0,08235	0,96135	0,72917	0,93868	0,81395	0,05098	0,03865	18,85730	0,05300	355,78425	0,94988	0,97061
0.80	0,70882	0,29118	0,84058	0,14063	0,80836	0,16981	0,16176	0,15942	5,19631	1,13366	4,58368	0,82416	0,97061
0.75	0,69249	0,30751	0,83188	0,09877	0,79722	0,12121	0,17136	0,16812	4,85456	1,70217	2,85197	0,81418	0,97061
0.70	0,71233	0,28767	0,83575	0,18557	0,81412	0,20930	0,15460	0,16425	5,40592	0,88513	6,10748	0,82479	0,97061
ort	**0,81698**	**0,18302**	**0,90072**	**0,45798**	**0,87759**	**0,50238**	**0,10253**	**0,09928**	**14,20533**	**0,64267**	**274,20213**	**0,88898**	**0,96229**

130

EK 13: EDM veri seti Phi Korelasyon değerleri

	OKT_BIN	CINS	IST_IK	SEMT_ME	AN_IS	BA_IS	BIRLIK	DERS_DEST	SIN_T	TV	NET	DERS	YUKSEK
OKT_BIN	1	0,3127108	0,052583	0,130098	0,0599924	0,0349713	0,013815	0,0372717	0,0358449	0,0945328	0,0169813	0,3730861	0,0166834
CINS	0,3127108	1	0,0647776	0,0772255	0,0047256	0,0188425	0,0323781	0,0009426	0,0420753	0,0921927	0,0995279	0,3084297	0,0283557
IST_IK	0,052583	0,0647776	1	0,0337387	0,0749818	0,0737704	0,0284731	0,0243315	0,0376972	0,0370677	0,096613	0,0045132	0,0421256
SEMT_ME	0,130098	0,0772255	0,0337387	1	0,0670558	0,0292092	0,0465285	0,0984133	0,1306591	0,0218682	0,0157549	0,0842102	0,0100402
AN_IS	0,0599924	0,0047256	0,0749818	0,0670558	1	0,026077	0,1197318	0,0213688	0,0147	0,0626689	0,0379671	0,0624933	0,0253763
BA_IS	0,0349713	0,0188425	0,0737704	0,0292092	0,026077	1	0,2698622	0,0342466	0,0020692	0,0181871	0,0190614	0,0103942	0,0349172
BIRLIK	0,013815	0,0323781	0,0284731	0,0465285	0,1197318	0,2698622	1	0,0194948	0,0150446	0,0227217	0,0044199	0,0126581	0,0188277
DERS_DEST	0,0372717	0,0009426	0,0243315	0,0984133	0,0213688	0,0342466	0,0194948	1	0,0092054	0,0045964	0,0219429	0,0707673	0,0197226
SIN_T	0,0358449	0,0420753	0,0376972	0,1306591	0,0147	0,0020692	0,0150446	0,0092054	1	0,0709404	0,0606488	0,018497	0,0197226
TV	0,0945328	0,0921927	0,0370677	0,0218682	0,0626689	0,0181871	0,0227217	0,0045964	0,0709404	1	0,1544087	0,1634452	0,018813
NET	0,0169813	0,0995279	0,096613	0,0157549	0,0379671	0,0190614	0,0044199	0,0219429	0,0606488	0,1544087	1	0,1100893	0,0083739
DERS	0,3730861	0,3084297	0,0045132	0,0842102	0,0624933	0,0103942	0,0126581	0,0707673	0,018497	0,1634452	0,1100893	1	0,0284821
YUKSEK	0,0166834	0,0283557	0,0421256	0,0100402	0,0253763	0,0349172	0,0188277	0,0197226	0,0197226	0,018813	0,0083739	0,0284821	1

131

Ek 14: Logistik Regresyon Analizine ait R ekran görüntüleri

```
Call:
glm(formula = OKT_BIN ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7200  -0.1692  -0.0196   0.0000   3.6479

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.755e+01  1.159e+04  -0.005  0.99604
SEMTBHC           -1.783e+01  1.784e+03  -0.010  0.99202
SEMTBKK            3.056e-03  6.960e-01   0.004  0.99650
SEMTBSK           -1.261e+00  9.010e-01  -1.400  0.16157
SEMTETI           -1.014e+00  1.078e+00  -0.941  0.34691
SEMTEYP            1.575e-01  6.824e-01   0.231  0.81742
SEMTFTH            3.095e+00  6.823e-01   4.536 5.73e-06 ***
SEMTGOP           -1.566e+00  8.936e-01  -1.752  0.07978 .
SEMTKÇK           -1.617e+01  1.627e+03  -0.010  0.99207
SEMTKDK            1.242e+00  6.582e-01   1.887  0.05912 .
SEMTKGT            2.734e+00  6.607e-01   4.138 3.50e-05 ***
SEMTLVT           -1.724e+01  1.639e+03  -0.011  0.99161
SEMTSRY            1.586e-01  9.322e-01   0.170  0.86487
SEMTSSL            2.361e-01  7.194e-01   0.328  0.74281
SEMTUSK           -1.852e+01  1.746e+03  -0.011  0.99154
DEVAM              4.066e-02  7.533e-03   5.397 6.77e-08 ***
CINSK             -1.364e+00  3.377e-01  -4.037 5.41e-05 ***
SINIF10            7.774e-01  4.793e-01   1.622  0.10482
SINIF11           -4.267e-01  6.577e-01  -0.649  0.51655
SINIF12           -1.262e+00  8.570e-01  -1.472  0.14100
SIN_MEVLEV3        1.664e+01  1.120e+03   0.015  0.98815
SIN_MEVLEV4        1.867e+01  1.120e+03   0.017  0.98670
YAS                1.897e-01  2.546e-01   0.745  0.45616
IST_IK10+         -1.176e-01  4.021e-01  -0.292  0.76997
SEMT_MEMH         -1.089e+00  3.587e-01  -3.036  0.00240 **
AIL_KS4-          -2.373e-01  7.318e-01  -0.324  0.74575
AIL_KS5            4.947e-01  3.604e-01   1.373  0.16985
AIL_KS5+          -4.254e-01  3.908e-01  -1.088  0.27638
AN_EGIT1          -1.457e+00  8.155e-01  -1.787  0.07394 .
AN_EGIT2          -9.538e-01  4.660e-01  -2.047  0.04068 *
AN_EGIT3          -1.025e+00  5.453e-01  -1.879  0.06020 .
```

AN_EGIT4	-1.660e+00	8.138e-01	-2.039	0.04142	*
AN_EGIT5	-2.973e+00	1.325e+00	-2.244	0.02485	*
AN_EGIT6	-1.191e+01	5.905e+03	-0.002	0.99839	
AN_ISH	8.393e-01	4.344e-01	1.932	0.05333	.
BA_EGIT1	1.579e+00	1.393e+00	1.134	0.25699	
BA_EGIT2	6.700e-01	1.031e+00	0.650	0.51593	
BA_EGIT3	6.198e-01	1.029e+00	0.603	0.54679	
BA_EGIT4	9.324e-01	1.094e+00	0.852	0.39424	
BA_EGIT5	2.625e+00	1.428e+00	1.838	0.06600	.
BA_EGIT6	4.228e+00	2.063e+00	2.049	0.04045	*
BA_ISH	4.107e-02	5.108e-01	0.080	0.93592	
BIRLIKH	-2.597e-01	7.140e-01	-0.364	0.71605	
MAD_DUCZ	1.799e+01	8.541e+03	0.002	0.99832	
MAD_DUI	1.718e+01	8.541e+03	0.002	0.99840	
MAD_DUN	1.646e+01	8.541e+03	0.002	0.99846	
MAD_DUZ	1.746e+01	8.541e+03	0.002	0.99837	
SIN_OCY	-2.737e-01	4.658e-01	-0.588	0.55680	
SIN_OO	-6.761e-01	4.467e-01	-1.513	0.13017	
SIN_OU	-8.554e-01	4.993e-01	-1.713	0.08671	.
SIN_OY	-1.165e+00	4.517e-01	-2.579	0.00992	**
DERS_DESTH	-3.904e-01	4.134e-01	-0.944	0.34500	
SIN_TH	-1.318e+00	5.408e-01	-2.437	0.01482	*
D_ORTAM2	3.511e-01	3.358e-01	1.046	0.29574	
D_ORTAM3	-2.031e-01	4.147e-01	-0.490	0.62429	
TVUN2	-4.493e-01	2.999e-01	-1.498	0.13411	
NETUN2	2.843e-02	2.873e-01	0.099	0.92115	
DERSUN2	1.628e+00	3.407e-01	4.777	1.78e-06	***
AN_BAG2	1.481e+00	1.890e+00	0.783	0.43335	
AN_BAG3	1.538e+00	1.574e+00	0.977	0.32852	
AN_BAG4	1.029e+00	1.491e+00	0.690	0.49006	
AN_BAG5	1.224e+00	1.502e+00	0.815	0.41496	
BA_BAG2	3.389e-02	7.390e-01	0.046	0.96343	
BA_BAG3	-4.458e-01	7.623e-01	-0.585	0.55864	
BA_BAG4	-4.524e-01	8.285e-01	-0.546	0.58507	
BA_BAG5	1.098e-01	7.661e-01	0.143	0.88598	
YUKSEKH	3.539e-01	1.007e+00	0.352	0.72514	
CEKEN_MODELAILE_KAD_MOD	8.642e+00	9.392e+03	0.001	0.99927	
CEKEN_MODELAKRABA	1.787e+01	7.756e+03	0.002	0.99816	
CEKEN_MODELAKRAN	1.725e+01	7.756e+03	0.002	0.99823	

```
CEKEN_MODELAKRAN          1.725e+01  7.756e+03   0.002  0.99823
CEKEN_MODELANNE           1.838e+01  7.756e+03   0.002  0.99811
CEKEN_MODELB_ERKEK_KAR    1.711e+01  7.756e+03   0.002  0.99824
CEKEN_MODELB_KIZ_KAR      1.542e+01  7.756e+03   0.002  0.99841
CEKEN_MODELBABA           1.759e+01  7.756e+03   0.002  0.99819
CEKEN_MODELBUYUK_KAR     -6.041e-01  1.935e+04   0.000  0.99998
CEKEN_MODELEBEVYN         1.446e+00  1.139e+04   0.000  0.99990
CEKEN_MODELOGRETMEN       1.737e+01  7.756e+03   0.002  0.99821
CEKEN_MODELYOK            1.677e+01  7.756e+03   0.002  0.99828
ITEN_MODELAKRAN          -3.424e-02  5.454e-01  -0.063  0.94995
ITEN_MODELANNE            4.738e-01  8.900e-01   0.532  0.59449
ITEN_MODELB_ERKEK_KAR    -2.334e+00  8.463e-01  -2.758  0.00582 **
ITEN_MODELB_KIZ_KAR      -1.390e+00  1.206e+00  -1.153  0.24890
ITEN_MODELBABA           -1.149e+00  8.464e-01  -1.358  0.17456
ITEN_MODELEBEVYN          4.682e-01  9.964e+00   0.047  0.96252
ITEN_MODELOGRETMEN       -1.248e+00  6.981e-01  -1.788  0.07385 .
ITEN_MODELYOK             2.756e-01  5.336e-01   0.516  0.60556
T                         2.145e-01  4.416e-02   4.857 1.19e-06 ***
D                        -1.042e-01  5.464e-02  -1.907  0.05654 .
AG                       -5.284e-02  1.172e-02  -4.507 6.57e-06 ***
DUK                       2.042e-02  1.774e-02   1.151  0.24969
SUK                      -1.129e-02  2.062e-02  -0.547  0.58405
BDI                       8.903e-02  5.195e-02   1.714  0.08656 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1494.14  on 1535  degrees of freedom
Residual deviance:  462.83  on 1444  degrees of freedom
AIC: 646.83

Number of Fisher Scoring iterations: 19
```

Ek 15: SVM Sınıflandırıcıda k-kat çapraz geçerleme yönteminde elde edilen performans değerleri

	SVM	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	Kat 1	0,90504	0,09496	0,9479	0,72222	0,93571	0,76471	0,05275	0,0521	17,968	0,07214	249,0842	0,94177
	Kat 2	0,91325	0,08675	0,95514	0,73457	0,93883	0,79333	0,05041	0,04486	18,9473	0,06107	310,2378	0,94692
4 kat	Kat 1	0,91335	0,08665	0,95954	0,71605	0,93521	0,80556	0,05386	0,04046	17,814	0,05651	315,2485	0,94722
	Kat 2	0,90632	0,09368	0,9422	0,75309	0,9422	0,75309	0,04684	0,0578	20,1159	0,07676	262,0778	0,9422
	Kat 3	0,93427	0,06573	0,96522	0,80247	0,95415	0,84416	0,03756	0,03478	25,6989	0,04334	592,8993	0,95965
	Kat 4	0,89202	0,10798	0,95072	0,64198	0,91877	0,75362	0,06808	0,04928	13,9658	0,07676	181,9512	0,93447
5 kat	Kat 1	0,89474	0,10526	0,94224	0,69231	0,92883	0,7377	0,05848	0,05776	16,1123	0,08343	193,1149	0,93548
	Kat 2	0,91789	0,08211	0,96751	0,70313	0,9338	0,83333	0,05572	0,03249	17,3642	0,04621	375,773	0,95035
	Kat 3	0,91789	0,08211	0,9529	0,76923	0,94604	0,79365	0,04399	0,0471	21,6626	0,06123	353,7791	0,94946
	Kat 4	0,92375	0,07625	0,96377	0,75385	0,94326	0,83051	0,04692	0,03623	20,5403	0,04806	427,3648	0,95341
	Kat 5	0,92082	0,07918	0,93841	0,84615	0,96283	0,76389	0,02933	0,06159	31,9996	0,07279	439,5968	0,95046
10 kat	Kat 1	0,91813	0,08187	0,96403	0,71875	0,93706	0,82143	0,05263	0,03597	18,3166	0,05005	365,9875	0,95035
	Kat 2	0,91228	0,08772	0,94964	0,75	0,94286	0,77419	0,04678	0,05036	20,2986	0,06715	302,3036	0,94624
	Kat 3	0,92982	0,07018	0,97101	0,75758	0,94366	0,86207	0,04678	0,02899	20,7554	0,03826	542,4716	0,95714
	Kat 4	0,90643	0,09357	0,96377	0,66667	0,92361	0,81481	0,06433	0,03623	14,9822	0,05435	275,6727	0,94326
	Kat 5	0,92982	0,07018	0,93478	0,90909	0,97727	0,76923	0,01754	0,06522	53,2826	0,07174	742,7273	0,95556
	Kat 6	0,91813	0,08187	0,97101	0,69697	0,93056	0,85185	0,05848	0,02899	16,6044	0,04159	399,2591	0,95035
	Kat 7	0,91176	0,08824	0,98551	0,59375	0,91275	0,90476	0,07647	0,01449	12,8874	0,02441	527,9808	0,94774
	Kat 8	0,91176	0,08824	0,94928	0,75	0,94245	0,77419	0,04706	0,05072	20,1721	0,06763	298,2589	0,94585
	Kat 9	0,96471	0,03529	0,98551	0,875	0,97143	0,93333	0,02353	0,01449	41,8841	0,01656	2528,75	0,97842
	Kat 10	0,91176	0,08824	0,94203	0,78125	0,94891	0,75758	0,04118	0,05797	22,8779	0,0742	308,3147	0,94545

135

Ek 16: SVM Sınıflandırıcıda k-kat çapraz geçerleme yönteminde 2,4, 5 ve 10 katlarda ortalama performans değerleri

	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
2 kat	0,90915	0,09086	0,95152	0,72840	0,93727	0,77902	0,05158	0,04848	18,45763	0,06661	279,66102	0,94435
4 kat	0,91149	0,08851	0,95442	0,72840	0,93758	0,78911	0,05159	0,04558	19,39866	0,06334	338,04421	0,94589
5 kat	0,91502	0,08498	0,95297	0,75293	0,94295	0,79182	0,04689	0,04703	21,53580	0,06234	357,92573	0,94783
10 kat	0,92146	0,07854	0,96166	0,74991	0,94306	0,82634	0,04748	0,03834	24,20611	0,05059	629,17262	0,95204
kFold ort	**0,91428**	**0,08572**	**0,95514**	**0,73991**	**0,94022**	**0,79657**	**0,04938**	**0,04486**	**20,89955**	**0,06072**	**401,20089**	**0,94752**

136

Ek 17: SVM Sınıflandırıcıda holdout(dışarıda bırak) yönteminde %95/%5, %90/%10, %85/%15, %80%20, %75/%25, %70/%30 ayırımlarda performans değerleri

SVM	Dogruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	LR_p	LR_n	DOR	F_measure
0.95	0,95294	0,04706	0,97101	0,87500	0,97101	0,87500	0,02353	0,02899	41,26812	0,03313	1245,78125	0,97101
0.90	0,94118	0,05882	0,97101	0,8125	0,95714	0,86667	0,03529	0,02899	27,5121	0,03567	771,1979	0,96403
0.85	0,92157	0,07843	0,96135	0,75000	0,94313	0,81818	0,04706	0,03865	20,42874	0,05153	396,44531	0,95215
0.80	0,93529	0,06471	0,96014	0,82813	0,96014	0,82813	0,03235	0,03986	29,6772	0,04813	616,6451	0,96014
0.75	0,91315	0,08685	0,96232	0,7037	0,93258	0,81429	0,05634	0,03768	17,0812	0,05355	318,9943	0,94722
0.70	0,90215	0,09785	0,94686	0,71134	0,93333	0,75824	0,05479	0,05314	17,28019	0,07470	231,31490	0,94005
ort	**0,92771**	**0,07229**	**0,96212**	**0,78011**	**0,94956**	**0,82675**	**0,04156**	**0,03788**	**25,54125**	**0,04945**	**596,72980**	**0,95577**

137

EK 18 Tez Çalışmasında Kullanılan R Kodları

DESICION TREE

```
EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)

library(rpart)
library(RWeka)
library(partykit)
#Adım 1: eğitim ve test setlerine random olarak ayırma
library(caret)
set.seed(1)
sep_random <- createDataPartition(y = EDM$OKT_BIN, p = .90, list = FALSE)
train <- EDM[sep_random,]
test <- EDM[-sep_random,]
testNitelik <- test[, -2]
testHedefNit <- test[[2]]
trainNitetik <- train[, -2]
trainHedefNit <- train[[2]]

DT_model <-J48(OKT_BIN~., data=train)
print(DT_model)
summary(DT_model)
plot(DT_model)

pred_test<- predict(DT_model, newdata = test[,-2])
summary(pred_test)
Egitim_TK  <- predict(DT_model, train, type = "prob")
Test_TK  <- predict(DT_model, test, type = "prob")

DT_table <- table(pred_test, test$OKT_BIN, dnn = c("Tahmin ", "Gerçek Sonuç"))
DT_table

 (tp <- DT_table[1])
(fp <- DT_table[3])
(fn <- DT_table[2])
(tn <- DT_table[4])

paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(DT_table)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
```

```
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/sum(DT_table)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))

library(FSelector)
information.gain(OKT_BIN~., data = EDM)

Egitim_tahminK <- prediction(Egitim_TK[,2], train$OKT_BIN)
Test_tahminK <- prediction(Test_TK[,2],test$OKT_BIN)
K_auce = performance(Egitim_tahminK, "auc")
K_auct = performance(Test_tahminK, "auc")
K_auce
K_auct
#*****Grafik için Hazırlık *****
Egitim_RocK <- performance(Egitim_tahminK, "tpr", "fpr")
Test_RocK <- performance(Test_tahminK, "tpr", "fpr")
Test_RocK

#Sensitivity/specificity
Egitim_RocKSS <- performance(Egitim_tahminK, "sens", "spec")
Test_RocKSS <- performance(Test_tahminK, "sens", "spec")
#Sensitivity/fpr=(1-specificity)
Egitim_RocKSF <- performance(Egitim_tahminK,"sens", "fpr")
Test_RocKSF <- performance(Test_tahminK, "sens", "fpr")
#ACC/Cutoff:
Egitim_RocKAC <- performance(Egitim_tahminK, "acc", "cutoff")
Test_RocKAC <- performance(Test_tahminK, "acc", "cutoff")
#precision/recall
Egitim_RocKRP<- performance(Egitim_tahminK, "prec", "rec")
Test_RocKRP <- performance(Test_tahminK, "prec", "rec")

# ***** tpr,fpr *****
plot(Egitim_RocK , lty=4 ,measure="lift",avg="threshold" )
par(new = TRUE) #ay grafige devam için
plot(Test_RocK ,   ,avg="threshold",lwd=2)
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("Karar Ağacı - tpr / fpr")

# ***** Sensitivy, fpr=(1-dpecifity) *****
plot(Egitim_RocKSF,xlab="fpr=(1-Specificity)",ld=1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKSF,lwd = 2,xlab="fpr=(1-Specificity)")
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
```

```
abline(0,1, lty=1)
title("Karar Ağacı - Sensitivity / fpr=(1-Specificity)")

# ***** precision/recall curve *****
plot(Egitim_RocKRP,   xlim=c(0,1),lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKRP   , lwd = 2, xlim=c(0,1))
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("Karar Ağacı - precision / recall(Sensitivity)")

# ***** acc/cutoff curve *****
plot(Egitim_RocKAC,   ylim=c(0.45,1),xlim=c(0,1), lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKAC    ,ylim=c(0.45,1),xlim=c(0,1),lwd = 2)
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("Karar Ağacı - Accuracy / CuOff")
```

DESICION TREE k-Fold

```
EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)

library(TunePareto)
library(rpart)
library(RWeka)
library(partykit)

set.seed(1)

CV_nfold <- generateCVRuns(EDM$OKT_BIN, ntimes = 1, nfold = 10, leaveOneOut
      = FALSE, stratified = TRUE)
CV_nfold_number <- 10
real <- CV_nfold [1:10]
pred_class <-NULL
CV_results <-NULL

start.time <-Sys.time()
for (i in 1:CV_nfold_number)
{

 print(paste("*** cross validation-k =", i))
 test_k <- CV_nfold$Run[[i]]
 print("test data indis")
 print(CV_nfold$Run[[i]])
 test <- EDM [test_k,]
 train <-EDM [-test_k,]
 print("Egitim veri kümesine ait sınıf degerleri:")
 print(table(train [,2]))
 print("Test veri kümesine ait sınıf degerleri:")
 print(table(test [,2]))
 test_Nit <-test[,-2]
 test_HNit <- test[[2]]
 real[[i]] <- test_HNit
 train_Nit <- train[,-2]
 train_HNit<- train[[2]]

 #DT ALGORITMASI
 EDM_DT<-J48(OKT_BIN~., data=train)
 pred_class[[i]]<- predict(EDM_DT, newdata= test_Nit)
 end.time <- Sys.time()
```

```
  taken.time <- end.time - start.time
  print(paste0("fold ", i, " tamamlanma suresi:", taken.time))
}
print("************************************************")
print("Gercek Sinif Degerleri:")
print(real)
print("Tahmin Edilen Sinif Degerleri:")
print(pred_class)
summary(pred_class)
print("************************************************")

for(j in 1:CV_nfold_number)
{
  print(paste0("Kontanjans Tablosu", "Fold=", j))
  DT_T<- table(factor(pred_class[[j]], levels = c("GEC","KALDI")), real[[j]], dnn =
        c("Tahmin","Gercek"))
  print( DT_T)
  (tp <-  DT_T[1])
  print(tp)
  (fp <-  DT_T[3])
  print( fp)
  (fn <-  DT_T[2])
  print(fn)
  (tn <-  DT_T[4])
  print( tn)
  print(paste0("Dogruluk=", dogruluk <- ((DT_T[1]+ DT_T[4])/sum( DT_T))) )
  print(paste0("Hata = ",(hata <- 1-dogruluk)) )
  print(paste0("TPR = ",(TPR <- (DT_T[1]/(DT_T[1]+DT_T[2])))))
  print(paste0("SPC = ",(SPC <- (DT_T[4]/(DT_T[3]+DT_T[4])))))
  print(paste0("PPV = ",(PPV <- (DT_T[1]/(DT_T[1]+DT_T[3])))))
  print(paste0("NPV = ",(NPV <- (DT_T[4]/(DT_T[4]+DT_T[2])))))
  print(paste0("FPR = ",(FPR <- (DT_T[3]/sum(DT_T)))))
  print(paste0("FNR = ",(FNR <- (DT_T[2]/(DT_T[2]+DT_T[1])))))
  print(paste0("LR_p = ",(LR_p <- TPR/FPR)))
  print(paste0("LR_n = ",(LR_n <- FNR/SPC)))
  print(paste0("DOR = ",(DOR <- LR_p/LR_n)))
  print(paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR))))

  print("************************************************")
}
```

K-NEAREST NEIGHBOURHOOD

```
library(class)
EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)

##KNN için sadece nümerik değerleri içerecek alt kümenin oluşturulması
EDM_NUM <-EDM[,c(2, ,7,29,30,31,32,33,34)]
str(EDM_NUM)
library(caret)
set.seed(1)
sep_random <- createDataPartition(y = EDM_NUM$OKT_BIN, p = .25, list = FALSE)
train <- EDM_NUM[sep_random,]
test <- EDM_NUM[-sep_random,]
testNit <- test[, -1]
testHNit<- test[[1]]
trainNit <- train[, -1]
trainHNit<- train[[1]]
sqrt(1706)
k_value=41
library(class)
set.seed(1)
predS_KNN = knn(trainNit, testNit, trainHNit, k = k_value)
predS_KNN
testHNit

KNN_tablo <- table(predS_KNN, testHNit, dnn = c("Tahmini Sonuclar", "Gercek
      Sonuclar"))
KNN_tablo
tp <- KNN_tablo[1]
fp <- KNN_tablo[3]
fn <- KNN_tablo[2]
tn <- KNN_tablo[4]

paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(KNN_tablo)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/sum(KNN_tablo)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
```

```
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))

confusionMatrix(data = predS_KNN, reference = testHNit)

# k'nin farkli degerleri icin dogruluk hesaplanirsa
dogruluk <- NULL

for(i in 1:k_value)
{
  set.seed(1)
  (predS_KNN = knn(trainNit, testNit, trainHNit, k = i))
  dogruluk[i] <- mean(predS_KNN == testHNit)
}

plot(seq(1:k_value), dogruluk, type="o", xlab="k", ylab="Doğruluk")
text(x = dogruluk, labels=round(dogruluk,2))
train.probsKnn <- attr(KNN_model, "prob")
probEKnn <- ifelse(KNN_modelE == "KALDI", 1, 0)
probTKnn <- ifelse(KNN_model == "KALDI", 1, 0)
probEKnn
probTKnn
Egitim_tahminKNN <- prediction(probEKnn, train$OKT_BIN)
Test_tahminKNN <- prediction(probTKnn,test$OKT_BIN)

#***** AUC *****
KNN_auce = performance(Egitim_tahminKNN, "auc")
KNN_auct = performance(Test_tahminKNN, "auc")
KNN_auce
KNN_auct


#tpr,fpr graf
Egitim_RocKnn <- performance(Egitim_tahminKNN, "tpr", "fpr")
Test_RocKnn <- performance(Test_tahminKNN , "tpr", "fpr")
#Sensitivity/specificity
Egitim_RocKnnSS <- performance(Egitim_tahminKNN, "sens", "spec")
Test_RocKnnSS <- performance(Test_tahminKNN, "sens", "spec")
#sens / fpr=1-spec
Egitim_RocKnnSF <- performance(Egitim_tahminKNN, "sens", "fpr")
Test_RocKnnSF <- performance(Test_tahminKNN, "sens", "fpr")
#ACC/Cutoff:
Egitim_RocKnnAC <- performance(Egitim_tahminKNN, "acc", "cutoff")
Test_RocKnnAC <- performance(Test_tahminKNN, "acc", "cutoff")
#precision/recall:
Egitim_RocKnnRP <- performance(Egitim_tahminKNN, "prec", "rec")
Test_RocKnnRP <- performance(Test_tahminKNN, "prec", "rec")
```

```
# ***** tpr,fpr *****
plot(Egitim_RocKnn , lty=4 ,measure="lift",avg="threshold" )
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnn ,   ,avg="threshold",lwd=2)
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("K-NN - tpr / fpr")

# ***** Sensitivy, fpr=(1-dpecifity) *****
plot(Egitim_RocKnnSF,xlab="fpr=(1-Specificity)",ld=1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnnSF,lwd = 2,xlab="fpr=(1-Specificity)")
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("K-NN - Sensitivity / fpr=(1-Specificity)")

# ***** precision/recall *****
plot(Egitim_RocKnnRP,   ylim=c(0.50,1),xlim=c(0.5,1),lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnnRP   , lwd = 2, xlim=c(0.5,1),ylim=c(0.50,1))
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("K-NN - precision / recall(Sensitivity)")

# *****acc/cutoff *****
plot(Egitim_RocKnnAC,   ylim=c(0.45,1),xlim=c(0,1), lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnnAC    ,ylim=c(0.45,1),xlim=c(0,1),lwd = 2)
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)

title("K-NN - Accuracy / CuOff")
```

K-NEAREST NEIGHBOURHOOD K-FOLD

```
library(class)
library(TunePareto)

EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)

EDM_NUM <-EDM[,c(2,3,7,29,30,31,32,33,34)]
str(EDM_NUM)
set.seed(1)

CV_nfold <- generateCVRuns(EDM_NUM$OKT_BIN, ntimes = 1, nfold = 2,
      leaveOneOut = FALSE, stratified = TRUE)
CV_nfold_number <- 2
real <- CV_nfold [1:2]
pred_class <-NULL
CV_results <-NULL

k_value=41

start.time <-Sys.time()

for (i in 1:CV_nfold_number)
{

 print(paste("*** cross validation-k =", i))
 test_k <- CV_nfold$Run[[i]]
 print("test data indis")
 print(CV_nfold$Run[[i]])
 test <- EDM_NUM [test_k,]
 train <-EDM_NUM [-test_k,]
 print("Egitim veri kümesine ait sınıf degerleri:")
 print(table(train [,1]))
 print("Test veri kümesine ait sınıf degerleri:")
 print(table(test [,1]))
 test_Nit <-test[,-1]
 test_HNit <- test[[1]]
 real[[i]] <- test_HNit
 train_Nit <- train[,-1]
 train_HNit<- train[[1]]

 #KNN ALGORITMASI
```

```
  EDM_KNN <- knn(train_Nit,test_Nit, train_HNit, k=k_value)
  pred_class[[i]]<- EDM_KNN
  test_Nit
  end.time <- Sys.time()
  taken.time <- end.time - start.time
  print(paste0("fold ", i, " tamamlanma suresi:", taken.time))

}

print("***********************************************")
print("Gercek Sinif Degerleri:")
print(real)
print("Tahmin Edilen Sinif Degerleri:")
print(pred_class)
summary(pred_class)
print("***********************************************")

for(j in 1:CV_nfold_number)
{
  print(paste0("Kontanjans Tablosu", "Fold=", j))
  KNN_T<- table(factor(pred_class[[j]], levels = c("GEC","KALDI")), real[[j]], dnn =
        c("Tahmin","Gercek"))
  print(KNN_T)
  (tp <- KNN_T[1])
  print(tp)
  (fp <- KNN_T[3])
  print( fp)
  (fn <- KNN_T[2])
  print(fn)
  (tn <- KNN_T[4])
  print( tn)
  print("***********************************************")
  print(paste0("Dogruluk=", dogruluk <- ((KNN_T[1]+KNN_T[4])/sum(KNN_T))) )
  print(paste0("Hata = ",(hata <- 1-dogruluk)) )
  print(paste0("TPR = ",(TPR <- (KNN_T[1]/(KNN_T[1]+KNN_T[2])))))
  print(paste0("SPC = ",(SPC <- (KNN_T[4]/(KNN_T[3]+KNN_T[4])))))
  print(paste0("PPV = ",(PPV <- (KNN_T[1]/(KNN_T[1]+KNN_T[3])))))
  print(paste0("NPV = ",(NPV <- (KNN_T[4]/(KNN_T[4]+KNN_T[2])))))
  print(paste0("FPR = ",(FPR <- (KNN_T[3]/sum(KNN_T)))))
  print(paste0("FNR = ",(FNR <- (KNN_T[2]/(KNN_T[2]+KNN_T[1])))))
  print(paste0("LR_p = ",(LR_p <- TPR/FPR)))
  print(paste0("LR_n = ",(LR_n <- FNR/SPC)))
  print(paste0("DOR = ",(DOR <- LR_p/LR_n)))
  print(paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR))))

  print("***********************************************")
}
```

LOGISTIK REGRESYON

```
library(ROCR)
library(caret)
EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)
set.seed(1)

sep_random <- createDataPartition(y = EDM$OKT_BIN, p = .90, list = FALSE)
train <- EDM[sep_random,]
test <- EDM[-sep_random,]
testNitelik <- test[, -2]
testHedefNit <- test[[2]]
trainNitetik <- train[, -2]
trainHedefNit <- train[[2]]

LR_model<- glm(OKT_BIN~., data=train, family=binomial)
summary(LR_model)
print(LR_model$coefficients)
(confint.default(LR_model))
print(exp(LR_model$coefficients))

train.probsL <-predict(LR_model, train, type='response')
pred.logitL <- rep('KALDI',length(train.probsL))
pred.logitL[train.probsL>0.5] <- 'GEC'

confusionMatrix(trainHedefNit, pred.logitL)

test.probsLT <-predict(LR_model, test, type='response')
pred.logitLT <- rep("GEC", length(test$OKT_BIN))
pred.logitLT[test.probsLT>0.5] <- 'KALDI'

confusionMatrix(test$OKT_BIN, pred.logitLT)
LR_tablo <- table(pred.logitLT, relevel(test$OKT_BIN, ref="GEC"), dnn = c("Tahmini
      Sonuçlar", "Gercek Sonuclar"))
LR_tablo

# #******LR Model performans kriterleri **********************
# library(fmsb)
# NagelkerkeR2(LR_model)
# library(survey)
# regTermTest(LR_model, "OKT_BIN")
# varImp(LR_model)
```

```
##doğruluk değeri hesaplama
(tp <- LR_tablo[1])
(fp <- LR_tablo[3])
(fn <- LR_tablo[2])
(tn <- LR_tablo[4])

paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(LR_tablo)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/sum(LR_tablo)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))

# Hosmer-Lemeshow Testi-Goodness of Fit
# install.packages("ResourceSelection")
library(ResourceSelection)
hoslem.test(LR_model$y, fitted(LR_model), g=10)

#ROC için hesaplar
Egitim_TL  <- predict(LR_model, train, type = "response")
Test_TL  <- predict(LR_model, test, type = "response")
Egitim_tahminL <- prediction(Egitim_TL, train$OKT_BIN)
Test_tahminL <- prediction(Test_TL,test$OKT_BIN)

#***** AUC *****
L_auce = performance(Egitim_tahminL, "auc")
L_auct = performance(Test_tahminL, "auc")
L_auce
L_auct

#tpr,fpr
Egitim_RocL <- performance(Egitim_tahminL, "tpr", "fpr")
Test_RocL <- performance(Test_tahminL, "tpr", "fpr")
#Sensitivity/specificity
Egitim_RocLSS <- performance(Egitim_tahminL, "sens", "spec")
Test_RocLSS <- performance(Test_tahminL, "sens", "spec")
#Sensitivity/fpr=(1-specificity)
Egitim_RocLSF <- performance(Egitim_tahminL, "sens", "fpr")
Test_RocLSF <- performance(Test_tahminL, "sens", "fpr")
#ACC/Cutoff:
Egitim_RocLAC <- performance(Egitim_tahminL, "acc", "cutoff")
Test_RocLAC <- performance(Test_tahminL, "acc", "cutoff")
```

```
#precision/recal
Egitim_RocLRP <- performance(Egitim_tahminL, "prec", "rec")
Test_RocLRP <- performance(Test_tahminL, "prec", "rec")

#***** tpr,fpr *****
plot(Egitim_RocL , lty=4 ,measure="lift",avg="threshold" )
par(new = TRUE) #ay grafige devam için
plot(Test_RocL ,   ,avg="threshold",lwd=2)
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("Lojistik Regresyon - tpr/fpr")

#***** Sensitivy, fpr=(1-dpecifity) *****
plot(Egitim_RocLSF,xlab="fpr=(1-Specificity)",ld=1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocLSF,lwd = 2,xlab="fpr=(1-Specificity)")
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("Lojistik Regresyon - Sensitivity / fpr=(1-Specificity)")

# ***** precision/recall *****
plot(Egitim_RocLRP,   xlim=c(0,1),lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocLRP   , lwd = 2, xlim=c(0,1))
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("Lojistik Regresyon - precision / recall(Sensitivity)")

# *****acc/cutoff  *****
plot(Egitim_RocLAC,   ylim=c(0.45,1),xlim=c(0,1), lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocLAC    ,ylim=c(0.45,1),xlim=c(0,1),lwd = 2)
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("Lojistik Regresyon - Accuracy / CuOff")
```

NAIVE BAYES

```
library(e1071)
library(caret)
EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)
set.seed(1)
sep_random <- createDataPartition(y = EDM$OKT_BIN, p = .75, list = FALSE)
train <- EDM[sep_random,]
test <- EDM[-sep_random,]

testNit <- test[, -2]
testHNit <- test[[2]]
trainNit <- train[, -2]
trainHNit <- train[[2]]

NB_model <- naiveBayes(trainNit, trainHNit)
NB_model
predS <- predict(NB_model, testNit)
summary(predS)

NB_table <- table(predS, testHNit, dnn = c("Tahmini Sonuçlar", "Gercek Sonuçlar"))
NB_table
confusionMatrix(data = predS, reference = testHNit)


##doğruluk değeri hesaplama
(tp <- NB_table[1])
(fp <- NB_table[3])
(fn <- NB_table[2])
(tn <- NB_table[4])

paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(NB_table)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/sum(NB_table)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
```

```
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))

#ROC için hesaplar
Egitim_TN  <- predict(NB_model, train,  "raw")
Test_TN  <- predict(NB_model, test, "raw")
Egitim_tahminN <- prediction(Egitim_TN[,2], train$OKT_BIN)
Test_tahminN <- prediction(Test_TN[,2],test$OKT_BIN)

#***** AUC *****
N_auce = performance(Egitim_tahminN, "auc")
N_auct = performance(Test_tahminN, "auc")
N_auce
N_auct

#tpr,fpr graf
Egitim_RocN <- performance(Egitim_tahminN, "tpr", "fpr")
Test_RocN <- performance(Test_tahminN , "tpr", "fpr")
#Sensitivity/specificity
Egitim_RocNSS <- performance(Egitim_tahminN, "sens", "spec")
Test_RocNSS <- performance(Test_tahminN, "sens", "spec")
#sens / fpr=1-spec
Egitim_RocNSF <- performance(Egitim_tahminN, "sens", "fpr")
Test_RocNSF <- performance(Test_tahminN, "sens", "fpr")
#ACC/Cutoff:
Egitim_RocNAC <- performance(Egitim_tahminN, "acc", "cutoff")
Test_RocNAC <- performance(Test_tahminN, "acc", "cutoff")
#precision/recall:
Egitim_RocNRP <- performance(Egitim_tahminN, "prec", "rec")
Test_RocNRP <- performance(Test_tahminN, "prec", "rec")

# ***** tpr,fpr *****
plot(Egitim_RocN , lty=4 ,measure="lift",avg="threshold" )
par(new = TRUE) #ay grafige devam için
plot(Test_RocN ,   ,avg="threshold",lwd=2)
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("Naive Bayes - tpr / fpr")

# ***** Sensitivy, fpr=(1-dpecifity) *****
plot(Egitim_RocNSF,xlab="fpr=(1-Specificity)",ld=1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocNSF,lwd = 2,xlab="fpr=(1-Specificity)")
legend("right", legend = c("Eğitim", "Test"),lty = c(4,1))
abline(0,1, lty=1)
title("Naive Bayes - Sensitivity - fpr=(1-Specificity)")

# ***** precision/recall *****
plot(Egitim_RocNRP,   xlim=c(0,1),lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
```

```
plot(Test_RocNRP   , lwd = 2, xlim=c(0,1))
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("Naive Bayes - precision / recall(Sensitivity)")

# *****acc/cutoff *****
plot(Egitim_RocNAC,   ylim=c(0.45,1),xlim=c(0,1), lwd = 1.5,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocNAC    ,ylim=c(0.45,1),xlim=c(0,1),lwd = 2)
legend("bottom", legend = c("Eğitim", "Test") ,lty = c(4,1))
abline(0,1, lty=1)
title("Naive Bayes - Accuracy / CuOff")
```

NAIVE BAYES K-FOLD

```
library(e1071)
library(TunePareto)

EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)
set.seed(1)

CV_nfold <- generateCVRuns(EDM$OKT_BIN, ntimes = 1, nfold = 10, leaveOneOut
      = FALSE, stratified = TRUE)
CV_nfold_number <- 10
real <- CV_nfold [1:10]
pred_class <-NULL
CV_results <-NULL

start.time <-Sys.time()
for (i in 1:CV_nfold_number)
{

 print(paste("*** cross validation-k =", i))
 test_k <- CV_nfold$Run[[i]]
 print("test data indis")
 print(CV_nfold$Run[[i]])
 test <- EDM [test_k,]
 train <-EDM [-test_k,]
 print("Egitim veri kümesine ait sınıf degerleri:")
 print(table(train [,2]))
 print("Test veri kümesine ait sınıf degerleri:")
 print(table(test [,2]))
 test_Nit <-test[,-2]
 test_HNit <- test[[2]]
 real[[i]] <- test_HNit
 train_Nit <- train[,-2]
 train_HNit<- train[[2]]
 #NB ALGORITMASI
 EDM_NB <- naiveBayes(train_Nit, train_HNit)
 pred_class[[i]]<- predict(EDM_NB, test_Nit)
 end.time <- Sys.time()
 taken.time <- end.time - start.time
 print(paste0("fold ", i, " tamamlanma suresi:", taken.time))
}
print("**********************************************")
```

```
print("Gercek Sinif Degerleri:")
print(real)
print("Tahmin Edilen Sinif Degerleri:")
print(pred_class)
summary(pred_class)
print("************************************************")

for(j in 1:CV_nfold_number)
{
  print(paste0("Kontanjans Tablosu", "Fold=", j))
 NB_T<- table(factor(pred_class[[j]], levels = c("GEC","KALDI")), real[[j]], dnn =
        c("Tahmin","Gercek"))
 print(NB_T)
 (tp <- NB_T[1])
print(tp)
 (fp <- NB_T[3])
print( fp)
 (fn <- NB_T[2])
 print(fn)
 (tn <- NB_T[4])
 print( tn)
 print("************************************************")
print(paste0("Dogruluk=", dogruluk <- ((NB_T[1]+NB_T[4])/sum(NB_T))) )
print(paste0("Hata = ",(hata <- 1-dogruluk)) )
print(paste0("TPR = ",(TPR <- (NB_T[1]/(NB_T[1]+NB_T[2])))))
print(paste0("SPC = ",(SPC <- (NB_T[4]/(NB_T[3]+NB_T[4])))))
print(paste0("PPV = ",(PPV <- (NB_T[1]/(NB_T[1]+NB_T[3])))))
print(paste0("NPV = ",(NPV <- (NB_T[4]/(NB_T[4]+NB_T[2])))))
print(paste0("FPR = ",(FPR <- (NB_T[3]/sum(NB_T)))))
print(paste0("FNR = ",(FNR <- (NB_T[2]/(NB_T[2]+NB_T[1])))))
print(paste0("LR_p = ",(LR_p <- TPR/FPR)))
print(paste0("LR_n = ",(LR_n <- FNR/SPC)))
print(paste0("DOR = ",(DOR <- LR_p/LR_n)))
print(paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR))))

 print("************************************************")
  }
```

ROC KARŞILAŞTIRMALARI

```
# ***** tpr,fpr *****
plot(Test_RocK, col = "darkblue", lwd = 2,lty=1)
par(new = TRUE) #ay grafige devam için
plot(Test_RocL , col = "red", lwd = 2,lty=3)
par(new = TRUE) #ay grafige devam için
plot(Test_RocN , col = "darkgreen", lwd = 2,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnn , col = "purple", lwd = 2,lty=6)
abline(0,1, lty=1)
title("Test Verileri -tpr - fpr")
legend("right","bottom", legend = c("Karar agaç", "Log.Reg.",
                              "Naive Bayes","K-NN"), col = c("blue",
                                                    "red","darkgreen","purple"), lty =
       c(1,3,4,6),lwd=2)


# ***** Sensitivity, fpr=(1-Specificity) *****
plot(Test_RocKSF, col = "darkblue", xlab="fpr=(1-Specificity)",
    lwd = 2,lty=1)
par(new = TRUE) #ay grafige devam için
plot(Test_RocLSF , col = "red", xlab="fpr=(1-Specificity)",
    lwd = 2,lty=3)
par(new = TRUE) #ay grafige devam için
plot(Test_RocNSF , col = "darkgreen",xlab="fpr=(1-Specificity)",
    lwd = 2,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnnSF , col = "purple", xlab ="fpr=(1-Specificity)",
    lwd = 2,lty=6)
abline(0,1, lty=1)
title ("Test Verileri - Sensitivity - fpr=(1-Specificity)")
legend("right","bottom", legend = c("Karar agaç", "Log.Reg.",
                              "Naive Bayes","K-NN"), col = c("blue",
                                                    "red","darkgreen","purple"), lty =
       c(1,3,4,6),lwd=2)


# ***** Recall,Prec. *****
plot(Test_RocKRP, col = "blue", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=1)
par(new = TRUE) #ay grafige devam için
plot(Test_RocLRP , col = "red", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=3)
par(new = TRUE) #ay grafige devam için
plot(Test_RocNRP , col = "green", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=4)
par(new = TRUE) #ay grafige devam için
plot(Test_RocKnnRP , col = "purple", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
```

```
    lty=6)
par(new = TRUE) #ay grafige devam için

abline(0,1, lty=1)
title("Test Verileri Precision - Recall(Sevsitivity)")
legend("bottom", legend = c("Karar agaç", "Log.Reg.",
                      "Naive Bayes","K-NN"), col = c("blue",
                                          "red","green","purple"), lty = c(1,3,4,6),lwd=2)

# ***** ACC-Cutoff *****
plot(Test_RocKAC, col = "blue", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=1)

par(new = TRUE) #ay grafige devam için
plot(Test_RocLAC , col = "red", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=3)

par(new = TRUE) #ay grafige devam için
plot(Test_RocNAC , col = "green", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=4)

par(new = TRUE) #ay grafige devam için
plot(Test_RocKnnAC , col = "purple", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=6)

abline(0,1, lty=1)
title("Test Verileri - Accuracy / CuOff")
legend("bottom", legend = c("Karar agaç", "Log.Reg.",
                      "Naive Bayes","K-NN"), col = c("blue",
                                          "red","green","purple"), lty = c(1,3,4,6),lwd=2)

# Modellerin Karşılaştırılması (Eğitim Verileri)

#***** tpr, fpr *****
plot(Egitim_RocK, col = "darkblue", lwd = 2,lty=1)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocL , col = "red", lwd = 2,lty=3)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocN , col = "darkgreen", lwd = 2,lty=4)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocKnn , col = "purple", lwd = 2,lty=6)
abline(0,1, lty=1)
title("Egitim dataları -tpr - fpr")
legend("right","bottom", legend = c("Karar agaç", "Log.Reg.",
                             "Naive Bayes","K-NN"), col = c("blue",
                                                 "red","darkgreen","purple"), lty =
       c(1,3,4,6),lwd=2)
```

```
#***** Sensitivity, fpr=(1-Specificity) *****
plot(Egitim_RocKSF, col = "darkblue", xlab="fpr=(1-Specificity)",lwd = 2,
    lty=1)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocLSF , col = "red", xlab="fpr=(1-Specificity)",lwd = 2,
    lty=3)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocNSF , col = "darkgreen",xlab="fpr=(1-Specificity)", lwd = 2,
    lty=4)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocKnnSF , col = "purple", xlab ="fpr=(1-Specificity)",lwd = 2,
    lty=6)
abline(0,1, lty=1)
title ("Egitim Verileri Sensitivity - fpr=(1-Specificity)")
legend("right","bottom", legend = c("Karar agaç", "Log.Reg.",
                            "Naive Bayes","K-NN"), col = c("blue",
                                                  "red","darkgreen","purple"), lty =
       c(1,3,4,6),lwd=2)

# ***** Recall,Prec. ******
plot(Egitim_RocKRP, col = "blue", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=1)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocLRP , col = "red", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=3)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocNRP , col = "green", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=4)
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocKnnRP , col = "purple", lwd = 2,xlim=c(0,1),ylim=c(0.5,1),
    lty=6)
par(new = TRUE) #ay grafige devam için

abline(0,1, lty=1)
title("Egitim Verileri Precision - Recall(Sevsitivity)")
legend("bottom", legend = c("Karar agaç", "Log.Reg.",
                      "Naive Bayes","K-NN"), col = c("blue",
                                            "red","green","purple"), lty = c(1,3,4,6),lwd=2)


# ***** ACC-Cutof  *****
plot(Egitim_RocKAC, col = "blue", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=1)

par(new = TRUE) #ay grafige devam için
plot(Egitim_RocLAC , col = "red", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=3)
```

```
par(new = TRUE) #ay grafige devam için
plot(Egitim_RocNAC , col = "green", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=4)

par(new = TRUE) #ay grafige devam için
plot(Egitim_RocKnnAC , col = "purple", lwd = 2,xlim=c(0,1),ylim=c(0.3,1),
    lty=6)

abline(0,1, lty=1)
title("Egitim Verileri - Accuracy / CuOff")
legend("bottom", legend = c("Karar agaç", "Log.Reg.",
                        "Naive Bayes","K-NN"), col = c("blue",
                                              "red","green","purple"), lty = c(1,3,4,6),lwd=2)
```

DESTEK VEKTÖR MAKİNELERİ

```
library(e1071)
library(caret)
EDM <- EDM_V5[,-(2:3)]
str(EDM)
EDM$SINIF <-as.factor(EDM$SINIF)
EDM$AN_EGIT <-as.factor(EDM$AN_EGIT)
EDM$BA_EGIT <-as.factor(EDM$BA_EGIT)
EDM$D_ORTAM <-as.factor(EDM$D_ORTAM)
EDM$AN_BAG <-as.factor(EDM$AN_BAG)
EDM$BA_BAG <-as.factor(EDM$BA_BAG)
set.seed(1)
sep_random <- createDataPartition(y = EDM$OKT_BIN, p = .90, list = FALSE)
train <- EDM[sep_random,]
test <- EDM[-sep_random,]

SVM_model=svm(OKT_BIN~., data=train, kernel = "linear", cost=10)
summary (SVM_model)
SVM_model$index

predictionSVM= predict(SVM_model, test[,-2])
predictionSVM
SVM_table <- table(predict=predictionSVM, truth=test$OKT_BIN)
SVM_table
SVM_table
(tp <- SVM_table[1])
(fp <- SVM_table[3])
(fn <- SVM_table[2])
(tn <- SVM_table[4])

paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(SVM_table)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/sum(SVM_table)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))

#plot(SVM_model, train$CINS)
#lower cost wider margines
#cross validation
set.seed(1)
```

```
tune.out=tune(svm, OKT_BIN~., data=train, kernel = "linear",ranges= list(cost= c(.001,
      0.01, 0.1, 1,5,10,100)))
summary(tune.out)
bestmodel= tune.out$best.model
summary(bestmodel)

predictionSVM= predict(bestmodel, test[,-2])
predictionSVM
SVM_table <- table(predict=predictionSVM, truth=test$OKT_BIN)
SVM_table
(tp <- SVM_table[1])
(fp <- SVM_table[3])
(fn <- SVM_table[2])
(tn <- SVM_table[4])

paste0("Dogruluk = ",(dogruluk <- (tp+tn)/sum(SVM_table)))
paste0("Hata = ",(hata <- 1-dogruluk))
paste0("TPR = ",(TPR <- tp/(tp+fn)))
paste0("SPC = ",(SPC <- tn/(fp+tn)))
paste0("PPV = ",(PPV <- tp/(tp+fp)))
paste0("NPV = ",(NPV <- tn/(tn+fn)))
paste0("FPR = ",(FPR <- fp/sum(SVM_table)))
paste0("FNR = ",(FNR <- fn/(fn+tp)))
paste0("LR_p = ",(LR_p <- TPR/FPR))
paste0("LR_n = ",(LR_n <- FNR/SPC))
paste0("DOR = ",(DOR <- LR_p/LR_n))
paste0("F_measure = ",(F_measure <- (2*PPV*TPR)/(PPV+TPR)))
#plot(predictionSVM, train, train$DEVAM ~ train$T)
#plot(predictionSVM, test$OKT_BIN)
#plot(predictionSVM, test$OKT_BIN, col = c(gray(0.2), gray(0.8)))
#abline(predictionSVM)
#points(test$OKT_BIN,predictionSVM, col= "blue")
#plot(SVM_model, EDM$OKT_BIN, EDM$DEVAM ~ EDM$DUK, fill = TRUE, grid
      = 50, slice = list(AN_IS=1, CINS= "K")
```

SHINY UYGULAMASI

Server

```
library(shiny)

# Define server logic required to draw a histogram
shinyServer(function(input, output) {
  # Expression that generates a histogram. The expression is
  # wrapped in a call to renderPlot to indicate that:
  #
  #  1) It is "reactive" and therefore should re-execute automatically
  #     when inputs change
  #  2) Its output type is a plot

  output$sonuc <- renderPrint({

    if (input$devam <= 34.5) {

      if (input$akademik_gudulenme_degeri > 55) {

        if (input$gunluk_ders_calisma_suresi == 2) {

          switch (input$sinif_mevcudu,
            "1" = "Gecti",
            "2" = "Gecti",
            "3" = "Gecti",
            "4" =  if(input$sinif_tekrari == 1) "Kaldi"
                   else
                   {
                     if (input$yas <= 16) {
                       "Gecti"
                     }
                     else{
                       "Kaldi"
                     }

                   }


          )

        }
        else
        {
          if(input$sinifi == 9){
            "Gecti"
          }
```

```
# sinif=10
else if (input$sinifi == 10)
{
 if (input$cinsiyeti == 1) {
  if (input$semt_memnuniyeti == 1) {
   "Gecti"
  }
  else
  {
   if (input$tukenmislik_degeri <= 17) {
    "Gecti"

   }
   else
   {
    "Kaldi"
   }

  }

 }
 #Kadin ise
 else
 {
  if (input$tukenmislik_degeri <= 20) {
   "Gecti"
  }
  else
  {
   "Kaldi"
  }

 }


}
#sinif=11
else if (input$sinifi == 11)
{
 "Gecti"
}

#sinif=12
else if (input$sinifi == 12)
{
 if (input$durumluk_kaygi <= 45) {
  "Gecti"
 }
 else
```

```
          {
            "Kaldi"
          }
        }

      }



    }
    else
    {
      if (input$tukenmislik_degeri1 <= 17) {
        if (input$gunluk_ders_calisma_suresi1 == 2) {
          "Gecti"
        }
        else
        {
          if (input$sinif_mevcudu1 == 1 || input$sinif_mevcudu1 == 2 ||
input$sinif_mevcudu1 == 3) {
            "Gecti"
          }
          else if (input$sinif_mevcudu1 == 4)
          {
            if (input$sinif_tekrari1 == 1) {

              if (input$semt_memnuniyeti1 == 2) {
                if (input$duyarsizlasma <= 14) {
                  "Gecti"
                }
                else
                {
                  "Kaldi"
                }

              }
              else
              {
                "Gecti"
              }


            }
            else
            {
              if (input$durumluk_kaygi1 <= 57)
              {
                "Gecti"
              }
```

```
        else
        {
          "Kaldi"
        }

      }

    }

  }
}
else
{
  if (input$sinif_tekrari2 == 2) {
    "Gecti"
  }
  else
  {
    if (input$gunluk_internet_kullanimi == 2) {
      if (input$anne_baba_birlikteligi == 1) {
        "Gecti"
      }
      else
      {
        "Kaldi"
      }

    }
    else
    {
      if (input$semt_memnuniyeti2 == 1) {

        if (input$gunluk_ders_calisma_suresi2 == 2 ) {
          "Gecti"
        }
        else
        {
          if (input$yas2 <=14 ) {
            "Gecti"
          }
          else
          {
            if (input$durumluk_kaygi2 <= 41) {
              "Gecti"
            }
            else
            {
              "Kaldi"
            }
```

```
            }

           }

          }
          else
          {
           "Kaldi"
          }

         }
        }
       }
      }
     }
     else
     {
      if (input$ogretmen_depresyon_degeri <= 7) {
         "Gecti"
      }
      else
      {
       if (input$annenin_egitim_duzeyi == 1) {
        "Kaldi"
       }
       else if(input$annenin_egitim_duzeyi == 2 || input$annenin_egitim_duzeyi == 3
|| input$annenin_egitim_duzeyi == 4)
       {
        if (input$sinif_mevcudu2 == 3 || input$sinif_mevcudu2 == 4) {
         "Kaldi"
        }
        else {
         "Gecti"
        }
       }
       else
       {
        "Gecti"
       }
      }
     }
    })
    })
```

ui

```
library(shiny)

# Define UI for application that draws a histogram
shinyUI(fluidPage(
  # Application title
  titlePanel("Akademik Performansa Dayali Basari Tahmini"),

  # Sidebar with a slider input for the number of bins
  sidebarLayout(
    sidebarPanel(
      numericInput("devam",
                   label = "Devam degeri",
                   value = 0),


        # Bolum1 -------------------------------------------------------------

      conditionalPanel(
        condition = "input.devam <= 34.5",




        numericInput(
          "akademik_gudulenme_degeri",
          label = "Akademik gudulenme degeri",
          value = 5
        ),


        conditionalPanel(
          condition = "input.akademik_gudulenme_degeri > 55",


          selectInput(
            "gunluk_ders_calisma_suresi",
            label = "Gunluk ders calisma suresi",
            choices = c("2 Saatin altinda" = 1, "2 Saat ve uzerinde" =
                          2),
            selected = 2
          ),

          conditionalPanel(
            condition = "input.gunluk_ders_calisma_suresi == 1",

            selectInput(
              "sinifi",
              label = "Ogrencinin sinifi",
```

```
    choices = c(9, 10, 11, 12),
    selected = 9
  )
),


conditionalPanel(
  condition = "input.sinifi == 10",

  radioButtons(
    "cinsiyeti",
    "Cinsiyeti",
    c("Erkek" = 1, "Kadin" = 2),
    selected = 1,
    inline = TRUE
  )
),

conditionalPanel(
  condition = "input.cinsiyeti == 1",

  radioButtons(
    "semt_memnuniyeti",
    "Semtinden Memnun mu?",
    c("Evet" = 1, "Hayir" = 2),
    selected = 1,
    inline = TRUE
  )
),

conditionalPanel(
  condition = "input.semt_memnuniyeti == 2",

  numericInput("tukenmislik_degeri",
               label = "Tukenmislik degeri",
               value = 17)

),

conditionalPanel(
  condition = "input.sinifi == 12",

  numericInput("durumluk_kaygi",
               label = "Durumluk kaygi",
               value = 45)
),

selectInput(
  "sinif_mevcudu",
```

```
    label = "Sinif Mevcudu",
    choices = c(
      "Level 1 (x<25)" = 1, "Level 2 (25<x<35)" = 2,
      "Level 3 (36<x<45)" = 3, "Level 4 (x>46)" =   4
    ),
    selected = 4
  ),

  radioButtons(
    "sinif_tekrari", "Sinif Tekrari:",
    c("Evet" = 1,
      "Hayir" = 2),
    selected = 1,
    inline = TRUE

  ),

  conditionalPanel(
    condition = "input.sinif_tekrari == 2",

    numericInput("yas",
                 label = "Ogrencinin yasi",
                 value = 0, min = 12, max = 19)

  )
)
,

#akademik_gudulenme_degeri <= 55
conditionalPanel(
  condition = "input.akademik_gudulenme_degeri <= 55",

  numericInput("tukenmislik_degeri1",
               label = "Tukenmislik degeri",
               value = 17),


  conditionalPanel(
    condition = "input.tukenmislik_degeri1 <= 17",

    selectInput(
      "gunluk_ders_calisma_suresi1",
      label = "Gunluk ders calisma suresi",
      choices = c("2 Saatin altinda" = 1, "2 Saat ve uzerinde" =
                    2),
      selected = 2
    ),

    conditionalPanel(
```

```
condition = "input.gunluk_ders_calisma_suresi1 == 1",

selectInput(
  "sinif_mevcudu1",
  label = "Sinif Mevcudu",
  choices = c(
    "Level 1 (x<25)" = 1, "Level 2 (25<x<35)" = 2,
    "Level 3 (36<x<45)" = 3, "Level 4 (x>46)" =   4
  ),
  selected = 4
),
conditionalPanel(
  condition = "input.sinif_mevcudu1 == 4",

  radioButtons(
    "sinif_tekrari1", "Sinif Tekrari:",
    c("Evet" = 1,
      "Hayir" = 2),
    selected = 1,
    inline = TRUE

  ),

  conditionalPanel(
    condition = "input.sinif_tekrari1 == 1",

    radioButtons(
      "semt_memnuniyeti1",
      "Semtinden Memnun mu?",
      c("Evet" = 1, "Hayir" = 2),
      selected = 1,
      inline = TRUE
    ),

    conditionalPanel(
      condition = "input.semt_memnuniyeti1 == 2",

      numericInput("duyarsizlasma",
                   label = "Duyarsizlasma degeri",
                   value = 14, min = 4, max = 20)

    )

  ),

  conditionalPanel(
    condition = "input.sinif_tekrari1 == 2",

    numericInput("durumluk_kaygi1",
```

```
                label = "Durumluk kaygi",
                value = 45)
    )

  )


 )
),

#input.tukenmislik_degeri1 > 17
conditionalPanel(
 condition = "input.tukenmislik_degeri1 > 17",

 radioButtons(
  "sinif_tekrari2", "Sinif Tekrari:",
  c("Evet" = 1,
   "Hayir" = 2),
  selected = 1,
  inline = TRUE

 ),

 conditionalPanel(
  condition = "input.sinif_tekrari2 == 1",

  selectInput(
   "gunluk_internet_kullanimi",
   label = "Gunluk internet kullanimi",
   choices = c("2 Saatin altinda" = 1, "2 Saat ve uzerinde" =
               2),
   selected = 2
  ),

  conditionalPanel(
   condition = "input.gunluk_internet_kullanimi == 2",

   radioButtons(
    "anne_baba_birlikteligi", "Ebeveynleri birlikte mi?",
    c("Evet" = 1,
     "HayÄ±r" = 2),
    selected = 1,
    inline = TRUE

   )
  ),

  conditionalPanel(
   condition = "input.gunluk_internet_kullanimi == 1",
```

```
        radioButtons(
          "semt_memnuniyeti2",
          "Semtinden Memnun mu?",
          c("Evet" = 1, "Hayir" = 2),
          selected = 0,
          inline = TRUE
        ),

        conditionalPanel(
          condition = "input.semt_memnuniyeti2 == 1",


          selectInput(
            "gunluk_ders_calisma_suresi2",
            label = "Gunluk ders calisma suresi",
            choices = c("2 Saatin altinda" = 1, "2 Saat ve uzerinde" =
                        2),
            selected = 2
          ) ,

          conditionalPanel(
            condition = "input.gunluk_ders_calisma_suresi2 == 1",

            numericInput("yas2",
                         label = "Ogrencinin yasi",
                         value = 0, min = 12, max = 19),

            conditionalPanel(
              condition = "input.yas2 > 14",

              numericInput("durumluk_kaygi2",
                           label = "Durumluk kaygi",
                           value = 45)
            )

          )

        )

      )

    )

  )

),
```

```
    # Bolum2 ------------------------------------------------------------------

    conditionalPanel(
      condition = "input.devam > 34.5",


      numericInput(
        "ogretmen_depresyon_degeri",
        label = "Ogretmen depresyon puani",
        value = 7
      ),

      conditionalPanel(
        condition = "input.ogretmen_depresyon_degeri > 7",
        selectInput(
          "annenin_egitim_duzeyi",
          label = "Annenin egitim duzeyi",
          choices = c("Okuma yazma bilmiyor" = 1, "Sadece okuma yazma biliyor"
=2, "Ilkokul mezunu"=3, "Ortaokul mezunu"=4,
                      "Lise Mezunu"=5, "Universite Mezunu" = 6, "Lisansustu" = 7),
          selected = 1
        ) ,
        conditionalPanel(
          condition = "input.annenin_egitim_duzeyi > 1 &&
input.annenin_egitim_duzeyi < 6",
          selectInput(
            "sinif_mevcudu2",
            label = "Sinif Mevcudu",
            choices = c(
              "Level 1 (x<25)" = 1, "Level 2 (25<x<35)" = 2,
              "Level 3 (36<x<45)" = 3, "Level 4 (x>46)" =   4
            ),
            selected = 4
          )
        )
    )
    )
  ),
  # Show a plot of the generated distribution
  mainPanel(h3(textOutput("sonuc"))) #plotOutput("distPlot"))
 )
))
```

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı Soyadı	Şebnem ÖZDEMİR
Uyruğu	TC
Doğum tarihi, Yeri	06.03.1981, Elazığ
Telefon	212 440 00 00-10037
E-mail	sebnemozde@gmail.com

Eğitim

Derece	Kurum/Anabilim Dalı/Programı	Yılı
Doktora	İ.Ü. Fen Bilimleri Enstitüsü/ Enformatik / Enformatik	2016
Yüksek Lisans	İ.Ü. Fen Bilimleri Enstitüsü/ Enformatik / Enformatik	2012
Lisans	Yıldız Teknik Üniv/Fen Fakültesi/Matematik	2004
Lise	Küçükçekmece Anadolu Lisesi	1998

Makaleler / Bildiriler

Ayvaz Reis Z.; **Özdemir Ş.** (2014). The Effect of DIMLE on Computer Literacy Level of Pre-Service Teachers. Athens Journal of Technology Engineering. Vol1. No.2 pp. 95-102.

Özdemir Ş., Ayvaz Reis Z. (2012). The Effect Of Dynamic And Interactive Mathematics Learning Environments (Dimle), Supporting Multiple Representations, On Perceptions Of Elementary Mathematics Pre-Service Teachers In Problem Solving Process, Mevlana International Journal of Education (MIJE), vol.3, pp.85-94.

Selçukcan Erol, Ç.; **Özdemir, Ş.;** Özen, Z.; Akadal, E.; Ayvaz Reis, Z. (2012). Fibonacci Spiral in Sunflower with Geogebra. Journal Of Education and Instructional Studies In The World, 2 (25), ISNN: 2146-7463 pp.190-201

Ayvaz Reis Z., **Özdemir Ş.,** (2010). Education And Intelligent Tutor System In Turkey, International Journal on New Trends in Education and Their Implications (IJONTE), Vol.1 Issue 3, ISSN: 1309-6249

Gülseçen S., **Özdemir Ş**., Gezer M., Akadal E., (2015) "The Good Reader Of Digital World, Digital Natives: Are They Good Writer Of That World?", Procedia-Social and Behavioral Sciences, 191, Elsevier, 2396-2401

Ayvaz Reis, Z. **; Özdemir, Ş.**;.Akadal, E. (2015). Awareness Research: Do We Know The New Generation Students?. ERPA International Congresses on Education 2015, ERPA Congresses 2015, 4-7 June 2015 (Baskıda)

Ayvaz Reis, Z. ;.Akadal, E.; **Özdemir, Ş.** (2015). Mobile guidance in touchscreen era: YARDIM@. ERPA International Congresses on Education 2015, ERPA Congresses 2015, 4-7 June 2015 (Baskıda)

Özdemir Ş. (2014) Conrtibutions of Universities to Children Informatics Education in Turkey. Local Proceedings of 7th International Conference on Informatics in Schoolc: Situation, Evolution and Perspectives (ISSEP 2014). 22-25 Eylül 2014, Istanbul, Türkiye 119-121. (Poster)

Ak O., Özdemir Ş., Ayvaz Reis Z. (2014)."Determining The Hierarchy Of Education Criteria In Educational Games By Using Ahp ", 5th International Future-Learning Conference on Innovations in Learning for the Future 2014: e-Learning, İSTANBUL, TÜRKİYE, 5-7 Mayıs 2014, pp.1-1 (Özet olarak basıldı)

Akadal, E.; **Özdemir, Ş.**; Ayvaz Reis, Z. (2012). The Analysis of Category of Educational Applications in Online Mobile Application Markets. Future Learning 2012 4th International Future Learning Conference on Innovations in Learning for the Future: e-Learning, 14-16 November 2012, Istanbul University, Istanbul, Turkey.

Özdemir, Ş., Ayvaz Reis, Z., Karadağ, Z., (2012) Exploring Elementary Mathematics Pre-Service Teachers' Perception To Use Multiple Representations In Problem Solving, I. International Dynamic Explorative Active Learning (IDEAL) 2012 Conference, 2-5 July 2012, Bayburt University, Conference Proceedings, p: 241-281

Ayvaz Reis, Z., Kara Öztürk, E., **Özdemir, Ş**. (2012). Investigating the Web Based Distance Education Modules among Teacher Candidates in Special Education. SITE 2012--Society for Information Technology & Teacher Education International Conference, March 5-9, 2012, Austin, Texas, USA, Paul Resta (Ed.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2012*, ISBN 1-880094-92-4, (pp. 103-108). Chesapeake, VA: AACE., Retrieved from http://www.editlib.org/p/39545 .

Özdemir, Ş. (2011). Oyun Tabanlı Öğrenmede Geogebra Kullanımı: Köklü Sayılar Keşif Oyunu. 5th International Computer & Instructional Technologies Symposium, 22-24 September 2011, Fırat University, ELAZIĞ- TURKEY

Ayvaz Reis, Z., **Özdemir, Ş.** (2011). Using Geogebra As An Information Technology Tool: Parabola Teaching, "World Conference on Learning, Teaching and Administration-2010", Cairo-Egypt, 29-31 October 2010 Conference Proceedings (included in the international ISI proceedings database), Volume 9, p.565-572

Özdemir Ş.; Akadal, E.; Ayvaz Reis, Z. (2016). "An Analysis of the Education Category in App Markets" Human-Computer Interaction: Concepts, Methodologies, Tools, and Applications. Information Science Reference. ISBN: 978-1-4666-8790-5, 1270-1282.

Yarman B.S.B., Zaim Gökbay İ., Özdemir Ş. (2015). "Hayata Bir Çocuk Bir Çocuğa Hayat: Suça Karışmanın Erken Yaşta Önlenmesi", Nobel Yayın Dağıtım, İSTANBUL

Özdemir Ş., "Aşırı Bilgi Artışı"(2015). Bilgi Yönetimi: Bilgi Türeticileri, Büyük Veri, İnovasyon,

Kurumsal Zeka. Gülseçen, S.(Ed), Papatya Yayıncılık, İstanbul, ss.14-24, 2015

Gülseçen S., Çelik S., Özdemir Ş., Uğraş T., Özcan M. (2013). "Education In Smart Cities", in: New Challenges in Education, Gallova, M., Guncaga, J., Chanasova, Z., Chovancova, M., Eds., VERBUM – vydavatelstvo Katolíckej univerzity v Ruzomberku, Ruzomberok, pp.118-139-, 2013 (Atıflı)

Özdemir Ş.; Gülseçen, S. (2015). "Aşırı Bilgi Artışının Bilgiye Erişim Sürecindeki Etkileri: İstanbul Üniversitesi Enformatik Bölümü Örneği" Eğitim ve Öğretim Araştırmaları Dergisi 4(3) 334-344.

Özdemir Ş., Lanpir E., Atasoy Y., Gülseçen S. (2015). "New Technology Experience In Turkey: The Case Of Bitcoin ", 5 Istanbul Journal of Innovation in Education. Vol.1, Issue 2, pp.47-52

Bello M., **Özdemir Ş**., Gülseçen S. (2015). "Cultural E-Learning Through Erasmus Experience Management ", Istanbul Journal of Innovation in Education. Vol.1, Issue 2, pp.25-36

Ayvaz Reis, Z.; Baktır, H. Ö.; Çelik, B.; Erkoç, M. F.; Özçakır, F. C.; **Özdemir, Ş**.; Şahin, K. (2012). Açık Kaynak Kodlu Öğrenme Yönetim Sistemleri Üzerine Bir Karşılaştırma Çalışması. Eğitim ve Öğretim Araştırmaları Dergisi (Journal of Research in Education and Teaching), Mayıs, Haziran, Temmuz 2012 Cilt 1 Sayı 2 ISSN: 2146-9199, Sayfa: 42-58

Gülseçen, S.; **Özdemir, Ş**.; Çelik, S.; Uğraş, T.; Özcan, M. (2014). Dijital Dünyadan Yansımalar: Bilgide ve Vatandaşlıkta Değişim. XVIII. Türkiye'de İnternet Konferansı Bildiri Kitapçığı (İNET-TR'13). 223-227. ISBN:978-605-85087-1-2 (ATIFLI)

Özdemir, Ş.; Akadal, E.; Ayvaz Reis, Z. (2013). Uygulama Marketlerinin Eğitim Kategorisi Altındaki Uygulamalarının İncelenmesi. Akademik Bilişim 2013, X. Akademik Bilişim Konferansı Bildirileri, 23-25 Ocak 2013, Akdeniz Üniversitesi, Antalya.

Akadal, E.; **Özdemir, Ş**.; Ayvaz Reis, Z. (2013). GNU Özgür Belgeleme Lisansı (GFDL) Kapsamındaki Dokümanlar İçin Bir Çevrimiçi Arşiv Geliştirilmesi. Akademik Bilişim 2013, X. Akademik Bilişim Konferansı Bildirileri, 23-25 Ocak 2013, Akdeniz Üniversitesi, Antalya.

Balaban, E.; Çelik, S.; **Özdemir, Ş**. (2012). Yaşanabilir Şehirler Modelinin Kurulması ve Bilişim-İletişim Teknolojilerinin Bu Model Üzerinde Etkisi. VI. İstanbul Bilişim Kongresi, 7-8 Kasım 2012, Bahçeşehir Üniversitesi, İstanbul. (ATIFLI)

Özdemir, Ş., Ayvaz Reis, Z., Erol, Ç. (2011). Yazılım Ürünü Geliştirme Sürecinin Örneklenmesi. Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri, 1-5 Şubat 2011 İnönü Üniversitesi, Malatya, Baskıda.