T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE FARKLI KARAR AĞAÇLARI VE K-EN YAKIN KOMŞULUK YÖNTEMLERİNİN İNCELENMESİ: KADIN HASTALIKLARI VE DOĞUM VERİSİNDE BİR UYGULAMA

Öğr. Gör. Sadi ELASAN
BİYOİSTATİSTİK ANABİLİM DALI
(Tıp Programı)

DOKTORA TEZİ

DANIŞMAN
Prof. Dr. Sıddık KESKİN

VAN – 2019

T.C.
YÜZÜNCÜ YIL ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE FARKLI KARAR AĞAÇLARI VE K-EN YAKIN KOMŞULUK YÖNTEMLERİNİN İNCELENMESİ: KADIN HASTALIKLARI VE DOĞUM VERİSİNDE BİR UYGULAMA

Öğr. Gör. Sadi ELASAN
BİYOİSTATİSTİK ANABİLİM DALI
(Tıp Programı)
DOKTORA TEZİ

DANIŞMAN
Prof. Dr. Sıddık KESKİN

VAN – 2019

KABUL VE ONAY

Van Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalında Sadi ELASAN tarafından hazırlanan *"Veri Madenciliğinde Farklı Karar Ağaçları ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları ve Doğum Verisinde Bir Uygulama"* adlı tez çalışması aşağıdaki jüri tarafından DOKTORA TEZİ olarak OY BİRLİĞİ/OY ÇOKLUĞU ile KABUL/RET edilmiştir.

Tez Savunma Tarihi: 09.04.2019

Prof. Dr. Sıddık KESKİN
Van Yüzüncü Yıl Üniversitesi
Jüri Başkanı

Prof. Dr. Abdullah YEŞİLOVA
Van Yüzüncü Yıl Üniversitesi
Jüri Üyesi

Doç. Dr. Gazel SER
Van Yüzüncü Yıl Üniversitesi
Jüri Üyesi

Doç. Dr. Serdal Kenan KÖSE
Ankara Üniversitesi
Jüri Üyesi

Doç. Dr. Mustafa Agah TEKİNDAL
Selçuk Üniversitesi
Jüri Üyesi

Tez hakkında alınan jüri kararı, Van Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu tarafından onaylanmıştır.

Prof. Dr. Semiha DEDE
Sağlık Bilimleri Enstitüsü Müdürü

ETİK BEYAN

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE

Doktora tezi olarak hazırlayıp sunduğum *"Veri Madenciliğinde Farklı Karar Ağaçları ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları ve Doğum Verisinde Bir Uygulama"* başlıklı tezim; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan çalışma/araştırma tarafımdan yapılmış olup tüm cümleler ve yorumlar bana aittir. Bu tezdeki bütün bilgiler akademik kurallara ve etik ilkelere uygun olarak hazırlanıp bu kural ve ilkeler gereği, çalışmada bana ait olmayan tüm veri, düşünce ve sonuçlara atıf yapılmış ve kaynak gösterilmiştir.

Yukarıda belirtilen hususların doğruluğunu beyan ederim.

Sadi ELASAN
09.04.2019

TEŞEKKÜR

Tez çalışmamasına olan değerli katkılarından dolayı; Tıp Fakültesi Biyoistatistik Anabilim Dalı Başkanı ve tez danışmanım Prof. Dr. Sıddık KESKİN'e, tez izleme komitesi öğretim üyeleri Prof. Dr. Abdullah YEŞİLOVA'ya ve Doç. Dr. Gazel SER'e teşekkürlerimi sunarım. Çalışmamın uygulama materyali için Uz. Dr. Orkun ÇETİN'e, desteğini benden esirgemeyen Dr. Öğr. Üyesi Can ATEŞ'e ve tezde bana yardımcı olan Arkadaşlarıma teşekkürlerimi sunarım. Ayrıca beni her konuda destekleyen ve yanımda olan değerli eşim Feyza'ya ve kızım Selen'e sonsuz sevgilerimi sunarım.

ÖZET

Elasan S, Veri Madenciliğinde Farklı Karar Ağaçları ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları ve Doğum Verisinde Bir Uygulama. Van Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Doktora Tezi, Van, 2019. Veri madenciliğinde, sınıflandırma amacıyla kullanılan algoritmalar genel olarak; "denetimsiz (unsupervised)" ve "denetimli (supervised)" olmak üzere iki başlık altında incelenebilir. Denetimli veri madenciliğinde "karar ağaçları (decision trees)" ve "k-en yakın komşu (k-nearest neighbor | KNN)" algoritmaları; parametrik olmayan yöntemler arasında olup, tahmin edici özelliğe sahiptir. Sınıflandırma amacıyla uygulanan bu algoritmalarla, çalışmadaki cevap değişkeni (bebeklerin doğum ağırlığı) üzerine etkili olan açıklayıcı değişkenler belirlenmiştir. Karar ağaçlarından; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5" algoritmaları kullanılmıştır. K-en yakın komşu algoritmasında; "Öklid" ve "Manhattan (City block)" uzaklık ölçüleri kullanılarak uygulama yapılmıştır. Sınıflandırma performansları göz önüne alınarak, en iyi tahmin değerini veren algoritmalar belirlenmeye çalışılmıştır. Bu sonuçlara göre; Duyarlık (Sensitivity) ölçütü bakımından en yüksek tahmin oranı %88.4 ile "CART" algoritmasında gözlenmiştir. Özgüllük (Specificity) ölçütü bakımından en yüksek tahmin oranı %98.2 ile "Rasgele Orman" algoritmasında görülmüştür. Genel doğruluk ölçütü bakımından ise en yüksek tahmin oranı %94.5 ile "C4.5" algoritmasında gözlenmiştir. Risk (hata) tahmin ölçütü bakımından en düşük algoritma, %5.6 ile "C4.5" algoritması olmuştur. Genel olarak sonuçlar incelendiğinde; tüm algoritmaların "iyi sınıflandırma, yüksek tahmin ve düşük hata oranı" ile çalıştığı söylenebilir. Ayrıca bu çalışma, yeni doğacak bebeklerin doğum ağırlığının, düşük doğum ağırlığında olup olmayacağına erken karar verme ve böylece koruyucu tedbirlerin alınması açısından araştırmacılara katkı sağlayabilir.

Anahtar kelimeler: Çapraz Geçerlik, Denetimli Yöntemler, Öklid Uzaklığı, Risk Tahmini, Sınıflama

ABSTRACT

Elasan S, Investigation of Different Decision Trees and K-Nearest Neighbor Methods in Data Mining: An Application on Gynecology and Birth Data. Van Yuzuncu Yil University, Institute of Health Sciences, Ph.D. Thesis in Department of Biostatistics, Van, 2019. In data mining, the algorithms used for classification can generally be examined under two headings as "unsupervised" and "supervised". "Decision trees" and "k-nearest neighbor (KNN)" algorithms in supervised data mining; nonparametric methods and has predictive feature. With these algorithms applied for classification purposes, explanatory variables which are most effective on the answer variable in the study (birth weight of babies) have been determined. From decision trees; "CART, CHAID, exhaustive CHAID, QUEST, Random Forest and C4.5" algorithms have been used. In k-nearest neighbor algorithm; "Euclidean" and "Manhattan (City block)" distance measurements have been applied. Considering the classification performances, it has been tried to determine optimal estimation algorithms. According to these results; the highest estimation rate in terms of sensitivity has been observed in the "CART" algorithm with 88.4%. The highest estimation rate in terms of specificity criterion has been seen 98.2% in the "Random Forest" algorithm. The highest estimation rate in terms of accuracy criterion has been seen 94.5% in the "C4.5" algorithm. The lowest rate in terms of the risk estimate has been observed in the "C4.5" of 5.6%. When the results are examined in general; it can be said that all algorithms work with "good classification, high estimation and low error rate". In addition, this study may contribute to early investigations of the birth weight of newborn babies, whether it is low birth weight or not, and thus taking preventive measures.

Keywords: Cross Validation, Supervised Methods, Euclidean Distance, Risk Estimation, Classification

İÇİNDEKİLER

SİMGELER

x	:	Açıklayıcı (bağımsız) değişken
y	:	Cevap (bağımlı) değişkeni
n	:	Gözlem sayısı
N	:	Toplam gözlem sayısı
V	:	($1, 2, ..., V$) tamsayı değerleri
$\varepsilon\ (e)$	:	Hata
x_i	:	i. açıklayıcı (bağımsız) değişkendeki gözlemin değeri
y_i	:	i. cevap değişkendeki gözlemin değeri
p	:	Önemlilik (significantly) değeri/derecesi
P *veya* Pr	:	Olasılık (probability)
$\bar{x}$	:	Açıklayıcı değişkenin ortalama
$\bar{y}$	:	Cevap değişkeninin ortalaması
d	:	Sapma, uzaklık, veri seti
$E(D)$	:	D veri setinin bölünmeden önceki entropisi
T	:	Öğrenme veri setindeki kayıt sayısı
B	:	Bölünecek aday düğümü
χ^2	:	Ki-kare test istatistiği
$\hbar$	:	Öğrenme örneği
$\hbar(t)$	:	Düğüm t 'ye ait öğrenme örneği
W_n	:	n gözlem ile ilişkili vaka ağırlığı
F_n	:	Gözlemlerin frekans ağırlığı
$p(t)$	:	Düğüm t sınıfındaki olasılık
$\wedge$	:	"*Ve*" koşul sembolü
$\mid$	:	Ayrım çizgisi
Φ	:	Fi (phi φ)

KISALTMALAR

AID	**:**	Automatic Interaction Detector (otomatik etkileşim belirleme)
ART	**:**	Yardımcı Üreme Teknikleri
C4.5	**:**	Successor of ID3 (ID3'ün halefi | yerine geliştirilmiş)
CART	**:**	Classification & Regression Trees (sınıflandırma ve regresyon ağaçları)
CHAID	**:**	Chi-squared automatic interaction detector (Ki-kare otomatik etkileşim)
GDM	**:**	Gestasyonel Diyabet
ID3	**:**	Iterative Dichotomiser3 (yinelemeli ikiye ayırıcı)
KA	**:**	Karar Ağaçları
KNN	**:**	K-Nearest Neighbor (k-en yakın komşu)
MARS	**:**	Multivariate adaptive regression splines (çok değişkenli uyarlamalı reg.)
QUEST	**:**	Quick Unbiased Efficient Statistical Trees (hızlı tarafsız verimli ağaçlar)
SGA	**:**	Anne Karnında Bebeğin Gelişim Geriliği
SLIQ	**:**	Supervised Learning in Quest (Quest için denetimli öğrenme)
SPRINT	**:**	Scalable Parallelizable Induction of Decision Trees (Karar ağaçlarının ölçeklenebilir ve paralelleştirilebilir indüksiyonu)
VM	**:**	Veri Madenciliği

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

1. GİRİŞ

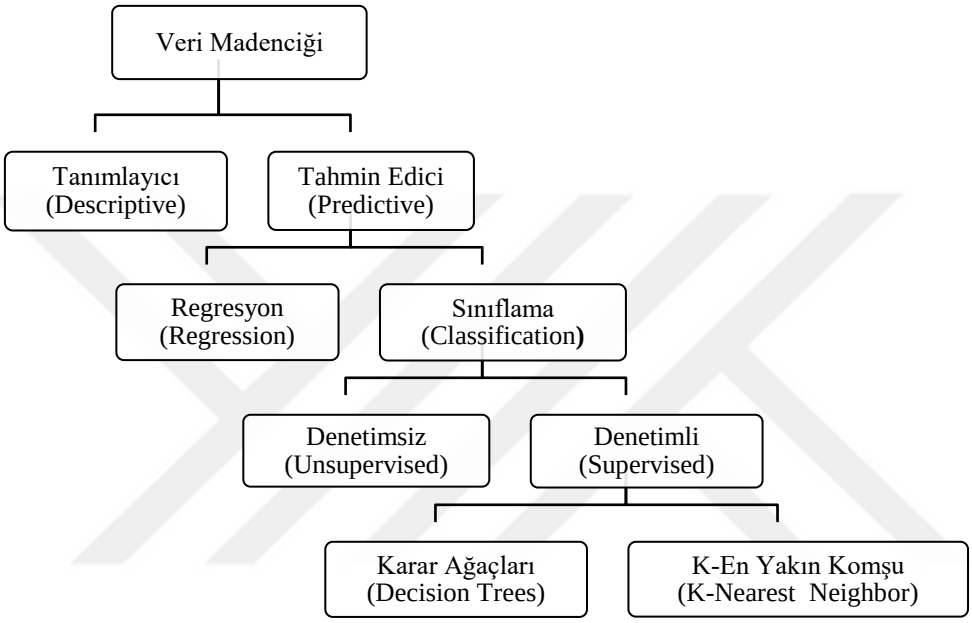
İstatistik, var olan sınırlı veriyi kullanarak doğru sonuçlara ulaştıran, sürekli gelişen ve her alana uygulanabilen disiplinler arası bir bilim dalıdır. Veri analizinde istatistik önemli bir araç olmakla birlikte, bazı durumlarda kullanımı sınırlı olabilmektedir. Değişen ve gelişen çevre koşulları, teknolojinin hızla gelişmesi ve bilgi saklama kapasitesinin artması; verinin değil bilginin önemini ön plana çıkarmaktadır. Bu gibi durumlarda "veri madenciliğine" gereksinim ortaya çıkmaktadır. Veri madenciliği, bilgiye hızlı ve güvenilir şekilde ulaşmak üzere, araştırmacılara optimum çözümler sağlayan bir karar destek aracıdır.

Genel anlamıyla veri madenciliği; büyük veri popülasyonları arasından amaca uygun, doğru güncel bilgiye ulaşmak üzere, objektif ve en uygun çözümleri kullanmayı sağlayan bir karar destek aracıdır. Günümüzde veri madenciliğinin kullanım alanlarının artması ile birlikte, tüm dünyada olduğu gibi ülkemizde de veri madenciliğine gösterilen ilgi gün geçtikçe artmaktadır. Literatürde; tıp, eczacılık, biyoloji, eğitim, elektronik, pazarlama, bankacılık, borsa, mühendislik ve haberleşme gibi birçok alanda veri madenciliği ile ilgili çalışmalar yer almaktadır.

Veri madenciliği, daha çok elektronik alanda yaygın olarak kullanılmakla birlikte, tıp alanında da özellikle karar verme süreçlerinde geniş kullanım alanı bulmuştur. Bu alandaki çalışmalar; halk sağlığı politikalarının belirlenmesi, ilaçların geliştirilmesi ve etkilerinin tespiti, hasta test sonuçlarının tahmini, hastalıkların teşhis ve tedavisi, hastane maliyetlerinin tahmini, genetik bozuklukların tespiti, ölüm oranları ve salgın hastalıkların tahmini açısından önemli bir yere sahiptir. Hasta bilgi modülü veya hastane kayıt sistemleri gibi veri toplama araçlarından elde edilen ham verilere uygulanan veri madenciliği çalışmaları; sağlık uzmanları, hastane yönetimi ve hastalar için büyük faydalar sağlamaktadır (Koyuncugil ve Özgülbaş, 2009).

Veri madenciliği, geniş yöntem ve kavramlar içermektedir. Veri madenciliği için kullanılan modeller, amaca göre; tahmin edici (predictive) ve tanımlayıcı (descriptive) modeller şeklinde iki başlıkta incelenebilir (Şekil 1). Tahmin edici modeller ise genel olarak, "regresyon (regression)" ve "sınıflama (classification)" olmak üzere iki başlık altında incelenebilir. Sınıflama yöntemleri, bilimsel çalışmalarda çözüme yönelik sıkça

başvurulan yöntemlerden birisidir. Veri madenciliği yöntemleri içerisinde sınıflandırma amacıyla kullanılan pek çok algoritma bulunmaktadır. Sınıflama yöntemleri genel olarak; "denetimsiz (unsupervised)" ve "denetimli (supervised)" yöntemler olmak üzere iki başlık altında incelenebilir. Çalışmada incelenen "Karar Ağaçları (decision trees)" ve "K-En Yakın Komşu (k-nearest neighbor | KNN)" algoritmaları, sınıflandırma amacıyla kullanılan denetimli veri madenciliği yöntemleri içerisinde yer almaktadır (Bigus, 1996).



Şekil 1. Sınıflandırma amaçlı kullanılan veri madenciliği şeması

İlk olarak Breiman ve ark. (1984) tarafından önerilen karar ağaçları, parametrik olmayan tahmin edici özelliğe sahiptir. Büyük veri tabanlarıyla kolayca entegrasyonu, güvenilirliğinin yüksek, maliyetinin düşük, sonuçlarının görsel ve kolay yorumlanabilir olması gibi nedenlerle karar ağaçları, sınıflama yöntemleri içerisinde sıkça kullanılan algoritmalardan birisidir. K-en yakın komşu yöntemi ise ilk olarak Cover ve Hart (1967) tarafından önerilmiş olup, belirlenen veri noktasının yer aldığı sınıfın veya en yakın komşunun, *k*-değerine göre belirlendiği bir sınıflandırma metodudur. Gözlemler veya nesneler arası uzaklığa dayalı sınıflandırma yapan denetimli veri madenciliği algoritmalarından birisidir. Parametrik olmayan yöntemler arasında yer almakta olup, kolay yorumlanabilir özelliğe sahiptir. Karar ağaçları ve k-en yakın komşu algoritmaları; örüntü (model, pattern) tanıma, veri madenciliği, istatistik, bilişsel

psikoloji, yapay zeka, biyoinformatik ve tıp gibi birçok alanda kullanılmaktadır (Hall ve Holmes, 2003; Mao ve ark., 2015).

İstatistik analizlerde, hem tahmin yapma hem de sınıflama problemlerinin çözümü için geliştirilen algoritma veya analiz yöntemlerinin yüksek doğruluk derecesine sahip olması istenir. Veri madenciliğinde sınıflandırma amacıyla kullanılan CART, CHAID, Ayrıntılı CHAID, QUEST ve K-En Yakın Komşu yöntemleri gibi çok sayıda yöntem bulunmaktadır. Bu yöntemlerin sınıflama problemlerindeki başarısı, genel olarak birbirine yakın (benzer) olmakla birlikte, örneklem büyüklüğüne ve değişken tipine göre farklılık gösterebilmektedir. Diğer yandan bu algoritmaların her birinin farklı avantajları ve uygulama kriterleri bulunmaktadır.

Bu algoritmalar; uygulanma aşamasından önce veri temizleme, veri dönüştürme ve indirgeme işlemlerini yönetebilmekle birlikte, bölme, durdurma, birleştirme ve budama gibi işlemleri yapabilmektedir. Ayrıca bu algoritmalar; tahminleyici ve denetimli olmaları ile birlikte, çapraz geçerlik testlerini kullanabilmektedir. Yapılan literatür incelemesinde, bu algoritmaların hepsinin birlikte incelenmesi ve aynı veri setinde performanslarının değerlendirilmesine ilişkin sınırlı sayıda çalışma olduğu görülmüştür. Böylece tez çalışmasında, veri madenciliğinde sınıflandırma amacıyla kullanılan karar ağacı yöntemlerinden; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman, C4.5" ve "K-En Yakın Komşu" yönteminin incelenmesi amaçlanmıştır. Ayrıca, bebeklerde doğum ağırlığına etkili olan faktörlerin erken ve yüksek doğrulukla belirlenmesi ve düşük doğum ağırlığını sınıflamada algoritmaların performanslarının değerlendirilmesi amacıyla Kadın Hastalıkları ve Doğum alanından elde edilen veri seti ile uygulama yapılması amaçlanmıştır.

2. GENEL BİLGİLER

Veri madenciliği yöntemleri üzerinde matematikçiler, 1950'lerden itibaren çalışmaya başlamış; makine öğrenme (machine learning), mantık, yapay zeka (artificial intelligence) ve bilgisayar yazılımları alanlarında çalışmışlardır. Daha sonraki yıllarda ise istatistikçiler, k-en yakın komşu gibi yeni bazı algoritmalar üzerinde çalışmışlardır. 1970'li yıllardan itibaren yeni yazılımlar, genetik algoritmalar, kümeleme yöntemleri ve karar ağaçları gibi algoritmaları da içermiştir. Veri madenciliğinin, özellikle sınıflandırma algoritmalarının; tıp, sağlık, eczacılık, biyoloji, genetik, pazarlama, bankacılık, sigortacılık, borsa, telekomünikasyon, mühendislik, endüstri ve istihbarat gibi birçok alanda uygulamaları görülmektedir. Veri madenciliğinin bu alanlardaki gelişiminin yanında, bilimsel anlamda akademik amaçlı veri madenciliği yazılımları da geliştirilmiştir (Brown ve ark., 1992; Jacobs, 1999; Witten ve Frank, 2000).

Karar ağacı uygulamaları ile ilgili literatür bildirişleri aşağıda özetlenmiştir:

Temelleri AID (automatic interaction detector) algoritmasına dayanan karar ağacı modelleri daha sonra yeni algoritmalar eklenerek geliştirilmiştir. Morgan ve Sonquist (1963) tarafından önerilen AID algoritması**,** karar ağacı tabanlı ilk algoritma olarak bilinmektedir. Bilgisayar bilimlerinde, veri grupları bir karar ağacı yöntemi ile tanımlanmasına rağmen bu yöntem istatistik uygulama düzeyinde uzun yıllar ilgi görmemiştir. Breiman ve ark. (1984) tarafından yayınlanan "Classification and Regression Trees" adlı kitapta, karar ağacı yöntemi CART algoritması ilk defa tanıtılmıştır. Bu çalışma, CART yönteminin istatistik biliminde gelişimine önemli katkı sağlamıştır. Quinlan (1986), karar ağaçları literatürüne ID3 algoritması olarak bilinen yeni bir algoritma eklemiştir. Yine aynı araştırmacı, C4.5 adlı karar ağacı algoritmasını "Programs for Machine Learning" adlı kitabında tanıtmıştır (Witten ve Frank, 2000).

Lopez ve ark. (1999) tarafından, sosyoekonomik açıdan orta düzeydeki 10-89 yaş aralığında 360 bireyin, Hepatit A antikorunun yaygınlığıyla ilişkili faktörler CHAID algoritması kullanılarak (%90.5 doğrulukla) incelenmiştir. Zekic Susac ve ark. (2004) tarafından, bir bankadan elde edilen veri seti kullanılarak, küçük işletmelerin kredi dereceleri modellenmiştir. Modelleme için; CART karar ağacı, yapay sinir ağları ve logistik regresyon olmak üzere üç farklı model kullanılmış ve sonuçları

karşılaştırılmıştır. Chan ve ark. (2006) tarafından, ortopedik engellilerin istihdam durumunu etkileyen faktörler CHAID algoritması (%92.6 doğruluk) ile incelenmiştir. McCarty ve Hastak (2007) tarafından yapılan çalışmada, doğrudan pazarlama performansının incelenmesi amacıyla iki farklı veri seti kullanılarak, CHAID ve logistik regresyon yöntemleri kullanılmış ve sonuçlar karşılaştırılmıştır.

Albayrak ve Yılmaz (2009) tarafından, sanayi ve hizmet sektörlerinde faaliyet gösteren şirketlere ait finansal göstergelere CHAID modeli uygulanmıştır. Bu modelin, logistik regresyon yöntemlerine göre daha etkili olduğu belirtilmiştir. Kavzoğlu ve Çölkesen (2010) tarafından, uzaktan algılanan uydu görüntülerinin sınıflandırılmasında kullanılan CART, CHAID ve QUEST karar ağaçlarının performansları, Z-oran testi ile analiz edilmiştir. Olson ve ark. (2012) tarafından yapılan bir çalışmada, CHAID karar ağacı yöntemi kullanılarak, bazı şirketlerin iflas verileri karşılaştırılmıştır. CHAID yönteminin, destek vektör makineleri ve yapay sinir ağları yöntemlerine oranla daha iyi performans gösterdiği belirtilmiştir. Doolatabadi ve ark. (2013) tarafından, menkul kıymetler borsasından elde edilen veriler kullanılarak bazı şirket iflaslarının tahmininde; logistik regresyon analizi, faktör analizi ve CHAID algoritmasının etkinlikleri incelenmiştir. Çalışmada, CHAID ile %77.6 doğru sınıflandırma oranı elde edilmiştir.

K-en yakın komşu uygulamaları ile ilgili literatür bildirişleri aşağıda özetlenmiştir:

Fix ve Hodges (1951) tarafından, örüntü tanımada kullanılmak üzere, k-en yakın komşu algoritması tanıtılmıştır. Daha sonra bu algoritma Cover ve Hart (1967) tarafından geliştirilmiş ve bu algoritma ile Bayes hata oranı geliştirme (Fukunaga ve Hostetler, 1975), uzaklık ağırlıklı yaklaşımlar (Dudani, 1976; Bailey ve Jain, 1978), bulanık mantık yöntemleri (Jozwik, 1983; Keller ve ark., 1985) ve esnek hesaplama yöntemleri (Bermejo ve Cabestany, 2000) de dahil olmak üzere pek çok yeni algoritma türetilmiştir.

Bhatt (2004) tarafından yapılan bir çalışmada; semptomlar, ilaçlar, hastalıklar ve kimyasallar arasındaki ilişkiler, k-en yakın komşu ve karar ağaçları ile incelenerek, magnezyum eksikliğinin migren ağrılarına neden olabileceği vurgulanmıştır. Pehlivan ve Apaydın (2005) tarafından, k-en yakın komşu ve radyal temelli fonksiyon ağı bulanıklaştırılarak, bulanık k-en yakın komşu ve bulanık radyal tabanlı fonksiyon ağları

tahmin edicileri oluşturulmuştur. Her iki tahmin edici, aynı deneyden alınan gerçek verilere uygulanmış ve elde edilen sonuçlar karşılaştırılmıştır. Santos ve ark. (2006) tarafından yapılan bir çalışmada, şirket iflaslarının tahmin edilmesinde, k-en yakın komşu, karar ağaçları ve diğer farklı veri madenciliği modellerini içeren bir sonuç sunulmuştur. Bu yöntemlerin %86-99 aralığında doğru sınıflandırma yaptığı ve şirketlerin finansal performansının tahmin edilmesinde uygun bir araç olarak kullanılabileceği vurgulanmıştır.

Çalışkan ve Soğukpınar (2008) tarafından yapılan bir çalışmada, internet ağ güvenliğinin tespiti için; k-en yakın komşu, kümeleme, k-ortalamalar ve denetimli-denetimsiz öğrenme yöntemlerinin bir arada kullanılabildiği hibrit bir yapı geliştirilmiştir. Bu uygulamada, en hızlı sonuç veren k-en yakın komşu yöntemi, test veri seti küçük alt kümelere ayrılarak diğer yöntemlerin zaman ve bellek dezavantajlarını gidermiştir. Karakış ve ark. (2011) tarafından yapılan bir çalışmada, meme kanserli hastaların patolojik verileri, hastaların koltukaltı lenf nodunun durum tespitinde k-en yakın komşu yöntemi kullanılmıştır. Köktürk (2012) tarafından, erken ve zamanında doğum yapan kadınlardan elde edilen verilere; yapay sinir ağları, karar ağacı ve k-en yakın komşu yöntemleri uygulanmış, yöntemlerin sınıflandırma performansları karşılaştırılmıştır.

Ertan ve ark. (2012) tarafından, Alzheimer gibi bilişsel rahatsızlıkları olan kişilerden kablosuz algılayıcılarla toplanan veriler; Naive Bayes ve k-en yakın komşu gibi yöntemlerle sınıflandırılmış ve k-en yakın komşu yönteminde %95 oranında doğru sınıflandırma ile yüksek performans gözlenmiştir. Küçük ve ark. (2013) tarafından yapılan bir çalışmada, ALS hastalarına ait verilerin sınıflandırılması için, destek vektör makinesi ve k-en yakın komşu yöntemleri tercih edilmiştir. Çalışma sonucunda, k-en yakın komşu algoritmasıyla başarılı bir sınıflandırma performansı tespit edilmiştir. Acar ve Özerdem (2013) tarafından, göz imgelerinin içerdiği iris doku özellikleriyle, kişilerin tanınması amaçlanmış ve komşu sayısı (k) farklı değerlerde alınarak test edilen sonuçlar sınıflandırılmıştır. Çalışma sonucunda en yüksek performans, k=2 olduğunda gözlenmiş ve %80.7 oranında sınıflandırma performansı elde edilmiştir.

Maleki ve ark. (2013) tarafından yapılan bir çalışmada, k-en yakın komşu yönteminde optimum k değerinin hesaplanması için yeni bir yöntem önerilmiş, alt

örnekleme ve *n*-kat çapraz geçerlik testinin kullanıldığı sınıflandırma algoritmaları kullanılmıştır. Çalışma sonucunda, önerilen algoritmanın, klasik algoritmaya göre daha iyi bir performans gösterdiği ortaya konulmuştur. Sezgin ve Çelik (2013) tarafından yapılan bir çalışmada, veri madenciliği yöntemleri içinde kayıp gözlemlerin tahmininde kullanılan yöntemler karşılaştırılmıştır. Çalışmada, eksik gözlemler belirlenirken, benzerlik tahmininde bulunmak için eksik gözlem satırı ile doldurulan satır arasındaki uzaklık hesabında k-en yakın komşu yönteminin kullanılabileceği belirtilmiştir.

Çalış ve ark. (2013) tarafından, metin temelli veri madenciliği algoritmaları ele alınarak reklam e-postalarının tespiti amaçlanmıştır. Çalışmada k-en yakın komşu, Naive Bayes ve destek vektör makinesi algoritmaları kullanılmış ve çalışma sonucunda, k-en yakın komşu algoritması ile %97 oranında doğru sınıflama performansı elde edildiği bildirilmiştir. Daş ve Türkoğlu (2014) tarafından yapılan bir çalışmada, DNA dizilimlerini sınıflandırmada; destek vektör makineleri, k-en yakın komşu ve yapay sinir ağı yöntemleri kullanılarak sonuçlar karşılaştırılmış ve k-en yakın komşu yönteminin diğer iki yöntemden daha yüksek sınıflandırma performansı verdiği görülmüştür.

2.1. Veri Madenciliği

Veri madenciliği; bilginin büyük veri popülasyonlarından elde edilmesi işlemlerini kapsar. Çok sayıda analiz yöntemiyle, verideki ilişkileri bularak, bunları geçerli tahmin ve/veya sınıflama yapmak için kullanan bir analiz sürecidir. Veri madenciliğinin temel amacı, önceki bilgilerin analizini yaparak, geleceğe yönelik tahmin ve karar verme modelleri geliştirmektir (Bhatt, 2004).

Veri madenciliği, büyük ham veri içerisinden araştırılan konuya ulaşma modelleri ortaya çıkarma işlemleri bütünüdür. Diğer bir ifadeyle, karmaşık veri setlerindeki örüntülerin ortaya çıkarılıp, karar vermek üzere kullanıma sunma sürecidir. Literatürde, veri madenciliğinin farklı tanımıyla karşılaşılmıştır. Buna göre veri madenciliğinin farklı tanımları şu şekilde özetlenebilir: Veri madenciliği, istatistik ve makine öğrenmeyle etkileşimli yeni bir disiplin ve geniş veri tabanlarından önceden bilinmeyen ilişkilerin analizidir (Han ve Kamber 2000). Veri madenciliği yüksek tahmin performansıyla, cevap değişkeninin büyük veri popülasyonlarından ayrılmasını sağlama yeteneğidir (Kitler ve Wang, 1998). Veri madenciliği, ham veriden tek başına elde edilemeyen bilginin ortaya çıkarılmasını sağlayan analiz sürecidir (Jacobs, 1999). Veri madenciliği, özet olarak büyük ham veri içerisinden tahminlerde bulunulabilmesini sağlayacak ilişkilerin incelenmesi sürecidir.

Veri madenciliği yöntemleri, amaca göre; tahmin edici ve tanımlayıcı olmak üzere iki başlıkta incelenebilirken, işlevine göre sınıflama, kümeleme, regresyon ve birliktelik kuralları şeklinde üç başlıkta incelenebilir. Sınıflama ve regresyon modelleri "tahmin edici"; kümeleme ve birliktelik kuralları ise "tanımlayıcı" yöntemlerdir (Berson ve ark., 2000; Hastie ve ark., 2001; Kovalerchuk ve Vityaev 2002). Tahmin edici sınıflama modelleri, denetimsiz (unsupervised) ve denetimli (supervised) olmak üzere iki başlık altında incelenebilir. Denetimsiz modeller; veriyi tanımaya ve keşfetmeye yönelik kullanılır ve daha sonra uygulanacak modeller için bilgi vermeyi amaçlar. Kümeleme yönteminde olduğu şekilde, gözlemler arasındaki benzerliklerden hareketle yeni sınıfların belirlenmesi amaçlanmaktadır. Denetimli sınıflama modelleri ise ham veriden yeni bilgiler ve karar vermede kullanılacak sonuçlar çıkarmak amacıyla kullanılmaktadır. (Hastie ve ark., 2001).

2.1.1. Veri Madenciliğinde Sınıflandırma (Classification)

Sınıflandırma, en temel veri madenciliği fonksiyonlarından biri olarak, cevap değişkeninin kategorik olması durumunda kullanılır. Burada amaç, cevap değişkeninin kategorilerinin bilinmesi veya bilinmemesi durumunda, nesneleri doğru sınıfa atayabilen sınıflama modelinin kurulmasıdır. Sınıflandırma ile yeni bir gözlemin özellikleri belirlendikten sonra, bu gözlemi bilinen bir sınıfa atanmaktadır (Harrington, 2012). Sınıflandırmada özellikler elde edildikten sonra, algoritmaya önceden tanıtılmamış örnekleri en yüksek doğrulukla atayacak sınıflandırıcı model geliştirilir. Böylece sonradan girilecek gözlemlerin, tanımlanmış sınıflardan belirli özellikler bakımından uygun olan sınıfa atanır (Fawcett, 2006). Sınıflandırma modelinin ilk aşamasında, her nesneye bir sınıf etiketi tanımlanır. İkinci aşamada ise bu model, mevcut verilerle uygulanmaya başlanır. Test veri seti rastgele ve öğrenme veri setinden bağımsız belirlenir. Daha sonra, test ve öğrenme veri setleri karşılaştırılarak (sınıflandırılmış test veri seti örneklerinin, toplam test veri seti örneklerine oranıyla) modelin doğruluğu belirlenir. Son aşamada, modelin kullanımından sonra yeni "veri sınıf etiketi" tahmini yapılır (Eui-Hong ve ark., 1996).

Modelin oluşumunda kullanılacak veri setine, öğrenme veri seti denilmektedir. Tablo 1'de üç değişkenli ve iki sınıflı küçük bir öğrenme veri seti ve bu veri setinin kullanılmasıyla, bebek doğum ağırlığının (BDA) sınıf dağılım bilgisi verilmiştir.

Tablo 1. Bir öğrenme veri seti ile sınıflandırma örneği

Anne Yaşı	İlk Gebelik mi?	Bebeğin Cinsiyeti	Bebek Doğum Ağırlığı (BDA)*
24	Hayır	Erkek	Normal
27	Evet	Kız	Düşük
29	Hayır	Kız	Normal
18	Evet	Kız	Düşük
21	Evet	Erkek	Düşük
22	Hayır	Kız	Normal
35	Hayır	Erkek	Düşük
17	Evet	Kız	Düşük
27	Evet	Kız	Normal
22	Hayır	Erkek	Normal
18	Evet	Kız	Düşük
32	Hayır	Erkek	Normal
17	Evet	Erkek	Düşük
21	Evet	Erkek	Normal
17	Evet	Kız	Düşük

* BDA≤2500g olanlar düşük doğum ağırlığına sahiptir

Tablo 2. Kategorik değişkenler için sınıf dağılımı

		Sınıf (BDA)	
		Düşük	**Normal**
İlk gebelik mi?	Evet	7	8
	Hayır	1	14
Bebeğin cinsiyeti	Erkek	2	13
	Kız	5	10

Kategorik değişkenler için her bir nesnenin sınıf dağılım bilgisi Tablo 2'de gösterilmiştir. Sürekli bir değişken için, sınıf kategorisinin tüm farklı değerlerini içeren ikili (binary) test dikkate alınır. Anne yaşının ikili yapıdaki sınıf dağılımı Tablo 3'te gösterilmiştir.

Tablo 3. Anne Yaşı değişkeninin ikili (binary) yapıdaki sınıf dağılımı

Anne yaşı	**İkili Test**	**Sınıf (BDA)**	
		Düşük	**Normal**
≤18	**1**	9	0
>18	**2**	15	21
≤21	1	6	1
>21	2	2	6
≤22	**1**	6	3
>22	**2**	2	4
≤24	1	6	4
>24	2	2	3
≤27	**1**	7	5
>27	**2**	1	2
≤29	1	7	6
>29	2	1	1
≤32	**1**	7	7
>32	**2**	1	0
≤35	1	8	7
>35	2	0	0

En iyi sınıflama başarısı için entropiden faydalanılır. Entropi, doğru sınıflama olasılıklarıyla ilgili olarak, belirsizliğin ölçülmesinde kullanılan bir kriterdir ve kategorik değişken seçimi, özelliklerin entropi kazanımlarına dayanır. Bir niteliğin entropisi sınıf dağılımı bilgilerinden elde edilir. Bütün değişkenlerin sınıf dağılımı bilgileri toplandığında; entropiye dayalı olan, bilgi kazanç oranı ya da Gini kriteri hesaplanır. En yüksek entropi değerine sahip olan değişken, düğüm genişletmede ayırma kriteri olarak seçilir (Eui-Hong ve ark., 1996).

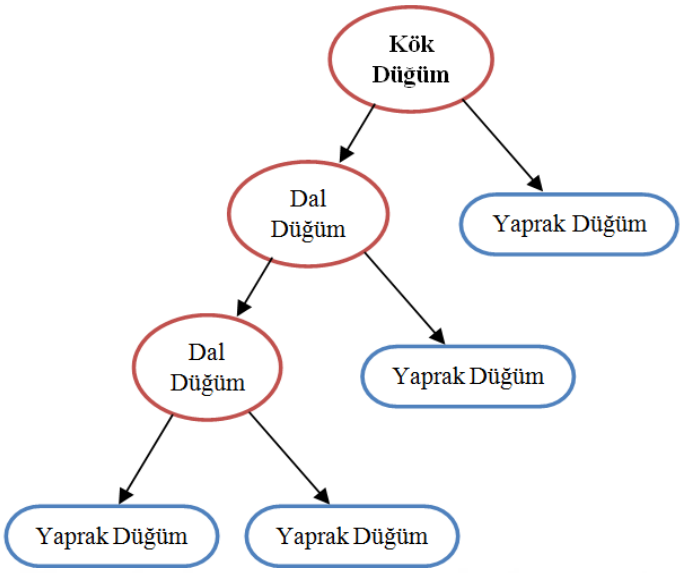
2.2. Karar Ağaçları Yöntemi (Decision Trees)

İlk olarak Breiman ve ark. (1984), tarafından önerilen karar ağaçları, güvenilir sınıflandırma yöntemlerinden birisi olarak bilinmektedir. Genel bir ifade ile karar ağaçları, kural çıkarma algoritmalarıdır. Karar ağaçlarını oluşturacak değişkenler kategorik veya sürekli özellikte olabilir. Karar ağaçları; cevap değişkeninin sürekli olması durumunda regresyon ağacı (regression tree), kategorik olması durumunda ise sınıflama ağacı (classification tree) olarak adlandırılmaktadır. Bu farklılığa rağmen karar ağaçları, iki durum için de benzer şekilde oluşturulmaktadır (Omitaomu, 2006). Klasik istatistik yöntemlerde, veriden bir fonksiyon elde edildikten sonra bu fonksiyonun anlaşılabilir bir kural olarak yorumlanması zordur. Oysaki karar ağaçları oluşturulduktan sonra kök düğümden yaprak düğümlere doğru inilerek, her dal bir kural oluşturacak şekilde fonksiyon yazılabilir (Luan, 2002).

Karar ağaçları yöntemi kullanılarak verinin sınıflandırılması işlemi iki aşamalıdır. Birinci aşama olan öğrenme aşamasında, bilinen bir öğrenme veri seti, model oluşturmak amacıyla sınıflandırma algoritması ile belirlenir. Öğrenilen bu model, sınıflandırma kurallarını oluşturur ve karar ağacı şeklinde ifade edilir. İkinci aşamada ise test verisi, sınıflandırma kurallarının doğruluğunu belirlemek amacıyla kullanılır. Sınıflandırma kurallarının doğruluğu kabul edilebilir düzeyde ise elde edilen kurallar yeni verilerin sınıflaması için kullanılabilir (Han ve Kamber 2000; Lewis 2000).

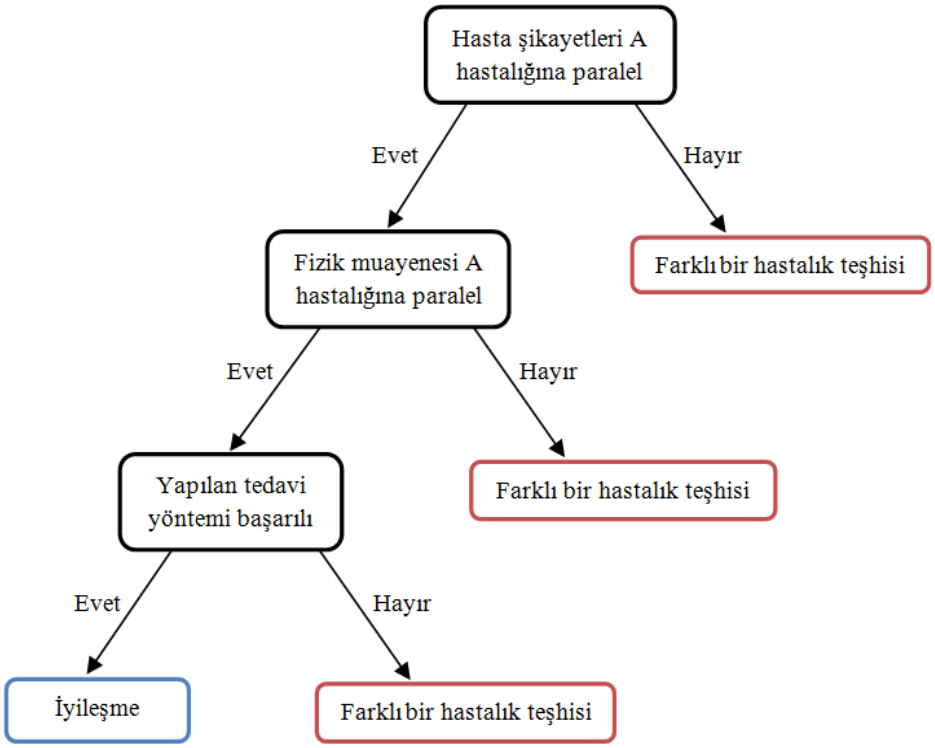
2.2.1. Karar ağacının yapısı

Karar ağacı genel olarak; "düğüm, dal ve yaprak" şeklinde üç kısımdan oluşur. Ağaç yapısındaki her bir değişken, başka düğüm tarafından temsil edilir. Karar ağacındaki en üst kısım kök düğüm, en alt kısım ise yaprak düğüm olarak adlandırılmaktadır. Kök ve yaprak düğümler arasında kalan kısımlar ise dal olarak ifade edilir. Diğer bir ifadeyle ağacın yapısı; verileri temsil eden bir kök düğüm, homojen olmayan dal düğümler ve homojen olan yaprak (terminal) düğümlerden oluşur. Karar ağacı, kök düğümden başlayarak devam eden ve deney ünitelerine uygulanan evet-hayır gibi cevaplara göre oluşan yollardan oluşur (Han ve Kamber 2000; Lewis 2000). Bir karar ağacının genel yapısı Şekil 2'deki gibi gösterilebilir.



Şekil 2. Karar ağacının genel yapısı

Ağaç yapısında bulunan her dal kural oluşturan bir fonksiyondur. Yeni dal oluşumu sırasında sınıflama tamamlanmamış ise yeni karar düğümü oluşturulur. Dallanma sonucunda oluşan karar düğümlerinin sayısı derinlik olarak ifade edilir. Derinlik sayısı, veri setinin büyüklüğüne ve homojenliğine bağlıdır (Carvalho ve Freitas, 2004). Herhangi bir hastalığının teşhisine ait karar ağacı Şekil 3'teki gibi gösterilebilir.



Şekil 3. Karar vermede kullanılan basit bir karar ağacı örneği

Karar ağacı düğümleri, birçok durumda bilgiyi daha anlaşılır gösteren test koşulu (if-then) kuralları setine dönüştürülebilir. Dal düğümden yaprak düğüme geçerken bir test koşulu kuralı oluşturulur. Şekil 3'teki örnek için, Eşitlik (1)'deki test koşulu kuralları, hastalığın teşhisinde bir kural seti verir. Bu test koşuluna göre; muayene A hastalığına paralel ve tedavi başarılı ise "doğru teşhis konulmuş ve iyileşme gözlenmiş" şeklinde yorum yapılacaktır.

$$Teşhis = \text{“muayene A hastalığına paralel” ve “tedavi başarılı”} \Rightarrow \text{“iyileşme”} \quad (1)$$

2.2.2. Karar ağacı oluşturma

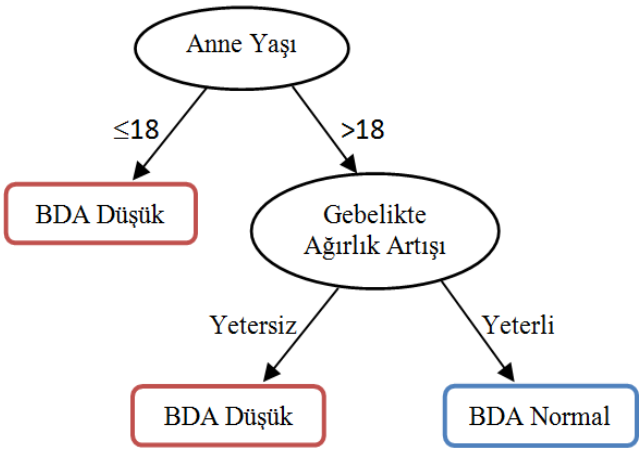
Kullanılan algoritmaya göre oluşturulan ağacın şekli değişebileceğinden karar ağaçları oluşturulurken kullanılan algoritma önemlidir. İlk düğüm olan kök düğümün yapısı, uçtaki yapraklara ulaşırken izlenecek yolu ve sınıflamayı etkilemektedir. Hem kök düğümün hem de sonraki her bir düğümün belirlenmesindeki temel kriter, düğümün bulunduğu noktadan itibaren dallara ayrıldığında kalan verilerin benzer büyüklükte dallara ayrılıp ayrılmadığıdır. Veri setindeki cevap değişkeni; evet-hayır ise düğüm iki eşit parçaya bölünürken; evet, hayır ve kararsız gibi üç kategorili ise düğüm üç parçaya bölünür (Lewis 2000; Han ve Kamber 2000).

Tablo 4'te bir öğrenme veri setine ait sınıf dağılımı ve bu sınıf dağılımının kullanılmasıyla Şekil 4'te bebek doğum ağırlığını (BDA) tahmin etmede kullanılan bir karar ağacı örneği verilmiştir.

Tablo 4. Bebek doğum ağırlığına (BDA) ait sınıf dağılımı

		Sınıf (BDA)	
		Düşük	**Normal**
Anne yaşı	≤18	4	0
	>18	0	6
Gebelikte Ağırlık Artışı	Yetersiz	3	2
	Yeterli	1	4

Tablodaki öğrenme veri seti örneği kullanılarak elde edilen bebek doğum ağırlığı karar ağacı aşağıdaki (Şekil 4) gibidir.



Şekil 4. Bebek doğum ağırlığına (BDA) ait karar ağacı

Test verisini sınıflandırmak amacıyla oluşturulan karar ağacı modelinde kök düğümden başlanarak, veriye bu test koşulu uygulanır ve uygun dallar oluşturulur. Bu dallardan, başka bir dal düğüme veya yaprak düğüme ulaşılarak işlem tamamlanır. Böylece, bu test verisinin hangi sınıfa dahil olacağı, sonlandığı yaprak düğüme bakılarak belirlenmiş olur.

Karar ağaçlarının oluşturulmasındaki en önemli adım, veri setindeki değişkenlerin sınıflamasını sağlayacak dallanmanın hangi kritere veya hangi değişkene göre yapılacağının belirlenmesidir. Bu aşamada, belirsizliği en yüksek olan değişken belirlenerek ağacın kök düğümünde test için kullanılır. Bunu belirlemeye yönelik geliştirilmiş literatürde farklı yaklaşımlar vardır. Bunlardan en önemli olan yaklaşımlar; Entropiye dayalı olan, bilgi kazancı (information gain) ve bilgi kazanç oranı (Quinlan, 1987; 1990; 2014) Twoing kuralı (Breiman ve ark., 1984), Gini kriteri (Breiman ve ark., 1984) ve Ki-kare olasılık (Mingers, 1989) tablo istatistiğidir (Han ve Kamber 2000).

CART, ID3 ve C4.5 algoritmalarında kullanılan Entropi sonucu, 0 ile 1 arasında değerler alır ve 1'e yaklaştıkça belirsizliğin arttığını, sıfır (0) değerine yaklaştıkça ise belirsizliğin azaldığını gösterir. Üzerinde çalışılan veri setindeki gözlemlerin tümü sadece bir tek sınıfa aitse, belirsizlik yoktur ve entropi sonucu 0 olacaktır. Bir durumun olma olasılığı $p=1$ değerine yakınsa, bu durumun gerçekleşmesinde güçlü bir belirsizlik olduğu sonucu çıkar. Karar ağacı algoritmasında amaç, gözlemlerin tümünün tek sınıf (homojen) olan yaprak düğümün entropisini sıfıra düşürmeye çalışmaktır (Breiman ve ark., 1984; Quinlan, 1987; 1990).

D, sınıf olasılık dağılımı *P* (p_1, p_2, ..., p_i) olan veri seti olmak üzere; P_i, *D* veri setindeki *i.* sınıfın olasılığı olup bu değer, *i.* sınıfa düşen örnek sayısının tüm veri setindeki toplam örnek sayısına bölünmesi ile elde edilir. Bu durumda *D*'nin entropisi Eşitlik (2)'deki gibidir (Quinlan, 1987; 1990).

$$Ent(D) = -\sum_{i=1}^{m} p_i \, log_2 \, (p_i) \tag{2}$$

Basit Entropi hesaplamaya aşağıdaki örnek (Dunham, 2003) verilebilir.

D={□, □, Δ, Δ, Δ, Δ, Δ, Δ} elemanlı bir veri seti olsun;

2 tane □ ve 6 tane Δ var.

Olasılıklar hesaplanırsa; □ için 0.25 (2/8) ve Δ için 0.75 (6/8) bulunur.

Bu örneğe ait entropi hesabı aşağıdaki gibidir.

$$Ent(D) = I(2.6) = -\frac{2}{8} \, log_2 \left(\frac{2}{8}\right) - \frac{6}{8} \, log_2 \left(\frac{6}{8}\right) = 0.1887 \tag{3}$$

Entropi 0'a yakın bulunduğundan, olayın gerçekleşmesinde belirsizlik düşüktür.

Karar ağaçları; tek cevap değişken ve çok sayıda açıklayıcı değişken içermesi bakımından regresyon modellerine benzer olup, cevap değişkeninin kategorik olduğu durumlarda logistik regresyon analizine alternatif olabilecek kullanışlı örüntüler keşfeder (Han ve Kamber, 2006). Parametrik olmayan yöntemler arasında olan karar ağaçlarında, girdiye göre bir model oluşturulur ve modelleme için büyük veri seti gerekir (Baykal, 2003; Köktürk, 2012). Sürekli ya da kesikli veriler ile çalışabilmeleri ve eksik ya da hatalı veriler ile tahminleme yapabiliyor olmaları karar ağaçlarının avantajlarındandır (Oğuzlar, 2004). Karar ağacındaki sürekli veriler, birden fazla katmanda (alt düğümde) yer alarak farklı kesim (binary) değerlerine göre dallanma oluşturabilirler. Bunun için belli bir sınırlama yoktur. Kategorik değişkenler de yine ağacın herhangi bir dalında daha ileri seviyelere bölmek için önceden kullanılan özelliği kullanmaz. Ancak burada dikkat edilmesi gereken en önemli şey, aynı özelliğin ağacın farklı bir dalında farklı bir seviyede kullanılabileceğidir (Stackexchange, 2018).

2.3. Karar Ağacı Algoritmaları

Karar ağacı yöntemleri içinde; genel anlamda "cevap değişkenini açıklayıcı değişkenlere göre sınıflandırma" mantığına dayanan farklı algoritmalar olsa da bu yöntemlerin her biri farklı amaçlara hizmet etmektedir. Ağaç tabanlı yöntemlerin temelini oluşturan ilk uygulamalar, AID (automatic interaction detector; Morgan ve Sonquist, 1963) algoritması ile yapılmış olup sonrasında ilave algoritmalar geliştirilmiştir. Geliştirilen bu algoritmalar aşağıdaki gibi sıralanabilir.

- CART (classification and regression trees; Breiman ve ark., 1984)
- CHAID (Chi-squared automatic interaction detector; Kass, 1980)
- Ayrıntılı (exhaustive) CHAID (Bigss ve ark., 1991)
- QUEST (quick, unbiased, efficient statistical tree; Loh ve Shih, 1997)
- Rastgele Orman (Random Forest; Breiman, 2001)
- C4.5 (successor of ID3) karar ağacı (Quinlan, 1993),
- C5 (successor of ID3) karar ağacı (Quinlan, 1993),
- ID3 ağacı (iterative dichotomiser 3; Quinlan, 1986)
- J48 ağacı (C4.5'in Weka'daki uygulaması; Witten and Frank 2005)
- CAL5 karar ağacı algoritması (Michie ve ark., 1994)
- Döndürme Ağacı (rotation forest; Rodriguez ve ark., 2006)
- MARS (multivariate adaptive regression splies; Friedman, 1991)
- Artırılmış Ağaçlar (boosted trees; Friedman, 2002)
- SLIQ (supervised learning in QUEST, Shafer ve ark., 1996)
- SPRINT (scalable parallelizable induction of decision trees; Shafer ve ark., 1996)

Yukarıda sayılan bu algoritmalardan; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5"e bu tez çalışmasında değinilmiştir.

2.3.1. CART (classification and regression trees) algoritması

CART (CRT, C&RT | sınıflandırma ve regresyon ağaçları), Breiman ve ark. (1984), tarafından tanıtılmış ve geliştirilmiştir. CART, bir değişkenin iki öğrenme basamağını içeren kök düğümden başlayarak, art arda iki düğüme bölünmesiyle

(homojenlik kriterine ulaşılana kadar süreç tekrarlanarak) oluşturulmuş ikili karar ağacıdır. Bu algoritma, sürekli ve kategorik verileri kullanabilmektedir. CART, ağaç modelinde ayırma kriterini hesaplama sırasında eksik gözlemleri önemsemeyen bir algoritmadır. Cevap değişkeni kategorik yapıda olduğunda yöntem "sınıflandırma ağaçları (classification trees | CT)", cevap değişkeni sürekli yapıda olması durumunda ise "regresyon ağaçları (regression trees | RT)" adını almaktadır (Breiman ve ark., 1984).

CART karar ağacı; düğüm, dal ve yaprak olmak üzere üç kısımdan oluşmaktadır. CART ağaç yapısında her bir değişken bir düğüm tarafından temsil edilir ve anlaşılması kolaydır. Diğer bir ifadeyle ağaç yapısı; verileri içeren bir kök düğümü, homojen olmayan dal düğümler ve homojen olan yaprak düğümlerden oluşur. Karar ağacı, kök düğümden başlayarak devam eden deney ünitelerine uygulanan evet-hayır gibi cevaplara göre geliştirilen dallar içerir. Her düğümde sorulan sorulara ayıraç ve bu işleme de ayırma denir (Breiman ve ark., 1984; Han ve Kamber, 2000).

CART karar ağacının oluşturulmasındaki en önemli adım, veri setindeki değişkenlerin sınıflandırılmasını sağlayacak dallanmanın hangi kritere veya hangi değişkene göre yapılacağının belirlenmesidir. Bu aşamada, belirsizlik oranı en düşük olan değişken işleme alınır ve kök düğümde test için kullanılır. Bunu belirlemeye yönelik, literatürde geliştirilmiş farklı yaklaşımlar vardır. Ancak, hangi değişkene göre dallanmanın yapılacağını belirlemede, yaygın olarak "entropi" sürecini içeren "bilgi kazancı" ve "bilgi kazanç oranı" kullanılmaktadır (Breiman ve ark., 1984).

Bilgi kazancı en yüksek olan değişkende dallanmaya başlanır. Entropi (Eşitlik 2) hesaplamasındaki *D* veri seti, *X* sınıf değerine göre alt kümelere (*n* tane alt bölüme) ayrıldığı varsayılsın. *D* veri seti kullanılarak, *X* sınıf değerine bölünmesiyle oluşan kazanç, Eşitlik (4)'teki gibi hesaplanır (Breiman ve ark., 1984; Quinlan, 1987, 1990).

$$Bilgi\ Kazancı\ (D,X) = Ent(D) - \sum_{i=1}^{n} p(D_i)\ Ent(D_i) \qquad (4)$$

E(D), veri setinin *X*'e göre bölünmeden önce hesaplanan entropi değeri

$E(D_i)$, *i* alt bölümünün *X*'e göre bölünmeden sonra elde edilen entropi değeri ve

$p(D_i)$, *i* alt bölümünün *X*'e göre bölünmeden sonra elde edilen olasılık ($\frac{|D_i|}{|D|}$) değeridir.

Eşitlik (4)'teki D veri seti için, uygun X sınıf değerinin belirlenmesinde Eşitlik (5)'te verilen "bölünme bilgisi" kullanılır

$$Bölünme\ Bilgisi\ (X) = -\sum_{i=1}^{k} \frac{|D_i|}{|D|} log_2\left(\frac{|D_i|}{|D|}\right) \quad (5)$$

Buradan, entropinin azalışını gösteren bilgi kazanç oranı, Eşitlikteki (6) gibi hesaplanır (Quinlan, 1987; 1990).

$$Bilgi\ Kazanç\ Oranı\ = \frac{Bilgi\ Kazancı\ (D,X)}{Bölünme\ Bilgisi\ (X)} \quad (6)$$

Eşitlikteki (6) bilgi kazanç oranı; sınıflandırma işleminde kullanılır. Bilgi kazanç oranı ölçütü ile karar ağacının düğümlerindeki her bir kazanç oranı en yüksek olacak şekilde, D öğrenme veri seti tekrarlı olarak bölünür. Her bir yaprak düğümü sadece bir sınıfa karşılık gelene kadar işlem tekrar edilir (Quinlan, 1987; 1990).

Diğer sınıflama ağaçlarında olduğu gibi CART'ta da tahmin için üç doğruluk testi bulunmaktadır. Bunlar; "yeniden yerine koyma", "test örneği tahmini" ve "çapraz geçerlik" testleridir. Model geçerlik testi için sık kullanılan "n-kat çapraz geçerlik" testinde, veri seti rastgele n gruba ayrılır. Bu testte n değeri genellikle 10 olarak seçilir. Bu gruplardan biri dışarıda bırakılarak, geriye kalan kısım ile model oluşturulur. Oluşturulan modelin sınıflama performansı, dışarıda bırakılan veriler üzerinde test edilir. Bu işlem sırayla diğer gruplar için de tekrarlanır. Son olarak elde edilen k adet doğru sınıflama oranının ortalaması alınarak modelin genel performansı değerlendirilir (Breiman ve ark., 1984).

CART algoritmasında, ağaç büyütme sürecinin sonlanıp sonlanmayacağı "durdurma kuralları" ile kontrol edilir. Aşağıdaki durumlarda, durdurma kuralları uygulanır ve düğüm bölünmesi durdurulur (Breiman ve ark., 1984).

- Bir düğümdeki bütün gözlemler, cevap değişkeni için özdeş değerler veriyorsa,
- Düğüm ve dal sayısı, kullanıcının belirlediği minimum sayıdan küçükse,
- Mevcut ağaç derinliği, kullanıcının belirlediği maksimum ağaç derinliği sınırına ulaşıyorsa düğüm bölünmez.

CART algoritması diğer ağaç temelli sınıflandırma yöntemleriyle karşılaştırıldığında bazı avantajlara sahiptir. Bu avantajlardan birisi, parametrik olmayışı ve böylece açıklayıcı değişken değerlerine ilişkin varsayımları gerektirmemesidir. Bu sebeple CART algoritmasında düzensiz sürekli değişkenler olabileceği gibi sınıflayıcı ve sıralı yapıda kategorik değişkenler yer alabilir. Bu durum, normallik testi ve verilen dönüşümü gibi işlemleri gerektirmemekte ve araştırıcıya zaman tasarrufu sağlamaktadır. CART, yüzlerce açıklayıcı değişkene sahip olsa bile, bütün bu değişkenleri inceleme gücüne sahiptir. CART algoritması, cevap değişkeninde eksik gözlemler olması durumunda kullanılabilir. Benzer şekilde, açıklayıcı değişkenlerde sıfır veya negatif değerler bulunması halinde de tercih edilebilir (Breiman ve ark., 1984; Loh ve Shih, 1997).

2.3.2. CHAID (chi-squared automatic interaction detector) algoritması

CHAID (Ki-kare otomatik etkileşim belirleme) algoritması ilk olarak Kass (1980) tarafından önerilmiş olup, optimal bölünmelerin (dallanmaların) belirlenmesinde Ki-kare test istatistiğini kullanır. Ki-kare adını almasının nedeni, algoritmasında çok sayıda çapraz (kontenjans) tablonun yer almasıdır. CHAID algoritması diğer algoritmalardan ayıran en belirgin farklardan birisi; "ID3, C4.5 ve CART algoritmaları" ikili dallanma üretirken, CHAID algoritması çoklu dallanma üretebilmektedir. Bu algoritmada cevap değişkeni kategoriktir (niteliksel, sıralı, sınıflayıcı). Açıklayıcı değişken sürekli olduğunda ise bu değişken sıralı değişkene dönüştürülür (Kass, 1980).

CHAID algoritmasında en uygun bölünmeleri belirlemek için CART'ta kullanılan Entropi kriterleri yerine Ki-kare testi tercih edilmektedir. Algoritmada, optimum bölünmeyi sağlamak amacıyla her bir açıklayıcı değişken için kategorilerin anlamlı bir şekilde birleştirilmesinden sonra, cevap değişkenine göre çapraz olasılık tabloları oluşturularak, Ki-kare istatistikleri ve Bonferroni *p*-değerleri hesaplanır. Düzeltilen *p*-değeri, bir Bonferroni çarpanı olarak bulunur. Açıklayıcı değişkenler birbirleriyle karşılaştırılıp, en küçük Bonferroni *p*-değerine sahip olan açıklayıcı değişkenin kategorilerine göre veriler alt gruplara ayrılır (Kass, 1980; Mingers, 1989).

Ki-kare olasılık tablo istatistiği, iki değişken arasındaki ilişkinin ölçüldüğü standart Ki-kare istatistiğine dayanmaktadır. Elde edilen yüksek değerler, büyük ilişkiyi ifade eder. Bu fonksiyona ait temel Eşitlik (7) aşağıdaki gibidir (Mingers, 1989).

$$x^2 = \sum_j \sum_i \frac{(x_{ij} - E_{ij})^2}{E_{ij}} \tag{7}$$

Eşitlik (7)'de E_{ij}: ($E_{ij}=x_i x_j/N$), olasılık tablosundaki her hücrenin beklenen değeridir ve bu değer, karar ağacı sınıflandırmasının son aşamasında ayırmanın nerede bitirilmesinin gerektiğini belirlemede kullanılır (Mingers, 1989).

CHAID algoritması, bir düğümün birden fazla dallanmasına veya bölünmesine izin verir ve kök düğümden başlayarak her düğümde aşağıdaki "üç adımı" tekrarlayarak büyür (Kass, 1980):

Birleştirme: Her açıklayıcı değişken (X) için hesaplanan Ki-kare istatistikleri dikkate alınarak anlamlı olmayan kategoriler birleştirilir. Bu işlem, açıklayıcı değişkenin kendi içindeki birleşmeleri anlamsız oluncaya kadar devam eder. Birleştirme basamağı, bölme sırasında kullanılacak düzeltilmiş p-değerini de hesaplar.

Bölme: Üç ya da daha fazla orijinal kategori içeren her tahmin için, en iyi bölme derecesine sahip açıklayıcı değişken ve düzeltilmiş p-değeri birleştirme aşamasında bulunur. Eğer orijinal kontenjans tablosunda herhangi bir indirgenme yoksa Ki-kare testi kullanılabilir. Kesin sonuçları bilinmiyorsa ya da orijinal olasılık tablosu indirgenmemiş ise Bonferroni düzeltmesi sonuçlarının kullanılması tercih edilir. Bölme basamağı, düğümü en iyi bölme için hangi tahmin değişkenini kullanacağını seçer. Seçim, her cevap değişkeni ile ilişkili düzeltilmiş p-değerleri karşılaştırılarak gerçekleştirilir.

Durdurma: CHAID'de, CART algoritmasında olduğu gibi bazı şartlar sağlandığı durumlarda ağaç büyütme süreci sonlandırılır (durdurulur) (Kass, 1980; Bigss ve ark., 1991). CHAID algoritması, cevap değişkeninde eksik gözlemler olması durumunda kullanılabilir. Benzer şekilde, açıklayıcı değişkenlerde sıfır veya negatif değerler bulunması halinde de kullanılabilir (Breiman ve ark., 1984; Loh ve Shih, 1997).

2.3.3. Ayrıntılı (exhaustive) CHAID algoritması

Ayrıntılı CHAID, her açıklayıcı değişken için olası tüm bölünmeleri inceleyen değiştirilmiş bir CHAID analizidir. Bu algoritma, CHAID'in zayıf yönlerini gidermek amacıyla geliştirilmiştir. Ayrıntılı CHAID algoritması, kullandığı istatistik testleri ve kayıp değerleri değerlendirmesi açısından CHAID analizine benzerdir. Zaman sorunu bulunmadığı durumlarda, Ayrıntılı CHAID algoritmasının tercih edilmesi daha yararlı olmaktadır ve kullanışlı bölünmeler bulabilmektedir (Bigss ve ark., 1991).

İlk defa Bigss ve ark. (1991), tarafından önerilen Ayrıntılı CHAID algoritması, açıklayıcı değişkenin benzer kategori çiftlerini 1'e indirmek üzere sürekli birleştirerek en iyi bölünmenin bulunmasını hedeflemektedir. CHAID, bir cevap değişkeni için en uygun bölünmeyi bulamayabilir. Bu durumda tüm kalan kategoriler istatistik olarak farklı bulunduğunda, CHAID algoritması kategorileri birleştirmeyi durdurur. Buna çözüm olarak, ayrıntılı CHAID yalnızca iki kategori kalana kadar açıklayıcı değişkenin kategorilerini birleştirmeyi sürdürür. Her bir açıklayıcı değişken için, kendi içinde kategoriler en anlamlı şekilde birleştirilip en iyi bölünme bulunduktan sonra, cevap değişkenine göre olasılık tablosu oluşturulur. Açıklayıcı değişkenler birbiri ile karşılaştırılıp, veriler en küçük *p*-değerine sahip olan açıklayıcı değişkenin kategorilerine göre alt gruplara ayrılır. Daha sonra, her bir açıklayıcı değişken için Ki-kare ve Bonferroni *p*-değeri hesaplanır. Düzeltilen *p*-değeri, bir Bonferroni çarpanı olarak bulunur. Açıklayıcı değişkenler birbirleriyle karşılaştırılıp en küçük Bonferroni *p*-değerine sahip olan açıklayıcı değişkenin kategorilerine göre veriler alt gruplara ayrılır (Goodman, 1979; Kass, 1980; Mingers, 1989; Bigss ve ark., 1991).

Ayrıntılı CHAID algoritmasında; bölme, birleştirme ve durdurma adımları, CHAID algoritmasındaki ile aynıdır. CHAID algoritmasında olduğu gibi açıklayıcı değişken olarak, kategorik (niteliksel, sıralı, sınıflayıcı) değişken tipine izin verilir. Benzer şekilde açıklayıcı değişkenler sürekli yapıda olduğunda, algoritmayı kullanmadan önce bu değişkenler sıralı kategorik değişkene dönüştürülür (Kass, 1980; Bigss ve ark., 1991).

Ayrıntılı CHAID algoritması, cevap değişkeninde eksik gözlemler olması durumunda kullanılabilir. Benzer şekilde, açıklayıcı değişkenlerde sıfır veya negatif

değerler bulunması halinde de tercih edilebilir. Eksik gözlemler için tahmini gözlemler türetilerek bu gözlemler işleme alınabilir (Kass, 1980; Breiman ve ark., 1984; Bigss ve ark., 1991).

2.3.4. QUEST (quick, unbiased, efficient statistical tree) algoritması

İlk defa Loh ve Shih (1997), tarafından önerilen QUEST (hızlı, tarafsız, etkili istatistik ağacı); bir değişkenin iki öğrenme basamağını içeren kök düğümden başlayarak art arda iki düğüme bölünmesiyle oluşturulmuş, ikili karar ağacı sağlayan bir sınıflandırma algoritmasıdır. Bu algoritma CART'a göre daha yüksek bir verimlilik sağladığı ve CART'ın dezavantajlarını ortaya çıkarttığı gözlenmiştir. Bu algoritma, ikili karar ağaçlarında olduğu gibi; bölme, durdurma ve budama gibi işlemlere izin verir (Loh ve Shih, 1997; Lim ve ark., 2000).

QUEST algoritmasında; cevap değişkeni kategorik iken, açıklayıcı değişkenler niteliksel, sıralı veya sürekli yapıda olabilir. Bu algoritma; ağacın ikili bölünmeyle sınırlandırılması, yansız ağaç tahmininin önemli olması, hesaplama maliyetinin düşürülmek istenmesi ve büyük veri setiyle çalışılması gibi durumlarda tercih edilir. Ayrıca QUEST algoritmasının diğer özellikleri arasında; varsayılan olarak yansız değişken seçme yöntemini kullanması ve niteliksel değişkenler için karesel diskriminant analizini (QDA) kullanarak bölünmeler sağlaması sayılabilir (Lim ve ark., 2000).

QUEST algoritması, karar ağacı yapımında kullanılacak değişkenleri seçerken; Ki-kare, Anova F veya Anova z_n istatistiğini gösteren ve $z_n = |x_n - \bar{x}^{(yn)}(t)|$ dönüşümüyle elde edilen Levene F-test istatistiklerini kullanır. Sürekli açıklayıcı değişken X için hesaplanan Levene F-istatistiği, Anova F-test istatistiğindeki Z_n değerine eşittir. (Loh ve Shih, 1997; Lim ve ark., 2000).

ANOVA F-test istatistiği: Bir t düğümü için, cevap değişkeni Y'nin J_t sınıfa sahip olduğu varsayılsın. Açıklayıcı değişken (X) sürekli yapıdayken, X'in farklı değerleri için Y tahminlerinin aynı olup olmadığını test eden bir Anova F-testi ve p-değeri aşağıdaki gibi hesaplanır (Loh ve Shih, 1997; Lim ve ark., 2000).

Sürekli açıklayıcı değişkenler için *F*-istatistiği;

$$F^{x} = \frac{\sum_{j=1}^{J_t} N_{f,j}\,(t)\left(\bar{x}^{(j)}(t) - \bar{x}(t)\right)^2 / (J_t - 1)}{\sum_{n\in\hbar(t)} \int_n \left(x_n - \bar{x}^{(y_n)}(t)\right)^2 / (N_f(t) - j_t)} \tag{8}$$

olarak hesaplanır. Eşitlik (8)'de,

$$\bar{x}^{(j)}(t) = \frac{\sum_{n\in\hbar(t)} \int_n x_n I(y_n = j)}{N_{f,j}(t)} \tag{9}$$

$$\bar{x}(t) = \frac{\sum_{n\in\hbar(t)} \int_n x_n}{N_f(t)} \tag{10}$$

p-değeri Eşitlik (11)'deki gibidir.

$$p_x = P_r\left(F_{(J_t - 1),\ N_f(t) - J_t)} > F_x\right) \tag{11}$$

Eşitlik (11)'deki *F, $(J_t - 1)$ ve $(N_f(t) - J_t)$* serbestlik dereceli F-dağılımı gösterir.

QUEST algoritmasında; bölme, birleştirme ve durdurma adımları, CART algoritması ile aynıdır (Loh ve Shih, 1997; Lim ve ark., 2000).

Eksik gözlem varlığı: QUEST algoritması, cevap değişkeninde eksik gözlemler olması durumunda da kullanılabilir. Benzer şekilde, sürekli, sıralı ve kesikli yapıdaki açıklayıcı değişkenlerde, sıfır veya negatif değerler bulunması halinde de bu algoritma kullanılabilir. Eksik gözlemler için tahmini gözlemler türetilerek bu gözlemler işleme alınabilir (Loh ve Shih, 1997; Lim ve ark., 2000).

2.3.5. Rastgele Orman (Random Forest) Algoritması

Rastgele Orman algoritmasında ağaçları bir dizi orman oluşturur. Bu ormandaki ağaç sayısı arttıkça yüksek doğrulukta sonuçlar elde edilir. Kısaca, Rastgele Orman Algoritması sınıflandırma işlemi sırasında birden fazla karar ağacı kullanılarak sınıflandırma doğruluğunun yükseltilmesi hedeflenir. Breiman (2001) tarafından geliştirilen algoritmada amaç tek bir karar ağacı üretmek yerine her biri farklı öğrenilmiş veri setinden oluşan, fazla sayıda ağacın kararlarını birleştirmektir. Farklı eğitim kümeleri oluştururken önyükleme (bootstrap) ve rastgele özellik seçimi kullanılır. Rastgele Orman, *p*'nin büyük ve *n*'nin küçük olduğu durumlarda geçerli olan ve özellikler arasında etkileşimlerin yanı sıra korelasyonu hesaba katabilen, yaygın bir ağaç tabanlı veri madenciliği algoritmasıdır. Rastgele Orman, özellikle büyük boyutlu veri analizleri için de tercih edilebilir (Breiman, 2001).

Rastgele Orman, bir düğümü parçalara ayırırken en önemli özelliği aramak yerine, rastgele bir özellik alt kümesi arasında en iyi özelliği arar. Bu genellikle daha iyi bir modelle sonuçlanan geniş bir çeşitlilik sağlamaktadır. Genel olarak, Rastgele Orman, sınırlamaları olmasına rağmen (çoğunlukla) hızlı, basit ve esnek bir araçtır (İnternet, 2018).

Rastgele Orman algoritması, karar ağacı oluşumunda kullanılacak değişkenleri seçerken, entropiye dayalı bilgi kazancı kriterlerinden Gini değerlerini hesaplar (Breiman ve ark., 1984; Breiman, 2001).

Gini algoritmasında niteliksel değerler iki parçaya ayrılarak bölünme yapılır. Her bölüm için $Gini_{sol}$ ve $Gini_{sağ}$ değerleri (Eşitlik 12) hesaplanır (Breiman ve ark., 1984).

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{|Tsınıf_i|}{|B_{sol}|} \right)^2 \quad ve \quad Gini_{sağ} = 1 - \sum_{i=1}^{k} \left(\frac{|Tsınıf_i|}{|B_{sağ}|} \right)^2 \qquad (12)$$

Eşitlik (12)'de; $Tsınıf_i$, sağ ve sol bölümlerdeki her bir sınıf değerini, $|B_{sol}|$ sol bölümdeki tüm değerlerin sayısını, $|B_{sağ}|$ sağ bölümdeki tüm değerlerin sayısını gösterir. Her bir *j* değişkeni için *n*, öğrenme veri setindeki örnek sayısı olmak üzere, aşağıdaki eşitlik hesaplanır (Eşitlik 13).

$$Gini_j = \frac{1}{n}[(|B_{sol}|Gini_{sol}) + (|B_{sağ}|Gini_{sağ})] \quad (13)$$

Eşitlik 13'te, her *j* değişkeni için hesaplanan $Gini_j$ ifadelerinden en küçük olanı seçilir ve bölünme bu değişken üzerinden yapılır.

Rastgele Orman algoritmasının özellikleri

- Büyük veri tabanlarında etkin bir şekilde çalışır.
- Orman oluştururken genel hata oranını iyi tahmin eder.
- Kayıp / eksik veri varlığında model doğruluğunu korur.
- Rastgele Orman, aşırıya kaçmadan istenildiği kadar ağaç taşıyabilir.
- Rastgele Orman hızlı çalışan bir algoritmadır.
- Algoritmada aşırı uyum (overfitting) problemi ortaya çıkmaz (Breiman, 2001).

2.3.6. C4.5 (successor of ID3) karar ağacı algoritması

C4.5 algoritması, ID3'ün tasarımcısı olan Quinlan (1993), tarafından geliştirilmiştir. ID3 algoritmasından farklı olarak sürekli veriler kategorik hale dönüştürülebilir. Ayrıca ağaç üzerinde erişim sıklıklarına göre alt ağaçların farklı seviyelere taşınması da mümkündür. ID3 ağacının yaklaşımından farklı olarak C4.5 ağacında budama işlemi yapılmaktadır. ID3 ağacı üzerinde entropi hesabı yapılır ve bu değere göre karar noktaları belirlenir. ID3 algoritmasının değişkenleri birçok alt bölüme ayırması sırasında yaşanılan aşırı öğrenme durumunun önüne geçilebilmesi için C4.5 algoritmasında ID3'ten daha farklı bir entropiye dayalı kriterlerden olan Gini kazanç oranı (Eşitlik 12-13) kullanılmaktadır (Quinlan, 1993; Breiman, 2001).

C4.5 algoritmasının, diğer karar ağaçlarından olan en büyük farkı normalizasyon yapılmasıdır (Eşitlik 14). Algoritmada her özelliğin normalize edilmiş bilgi kazancı kullanılır. En iyi bilgi kazancını veren özellik karar ağacına eklenir ve bütün yollar için bu adımlar tekrar edilir. C4.5 karar ağacında ön budama ve son buda işlemleri yapılabilir (Quinlan, 1993; Kantardzic, 2011).

Normalize edilmiş bilgi kazancı aşağıdaki Eşitlik (14) yardımıyla elde edilir.

$$B\ddot{o}l\ddot{u}nmeBilgisi_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|}) \quad (14)$$

Eşitlik (14)'te elde edilen bilgi kullanılarak aşağıdaki gibi formülize edilir:

"Normalize edilmiş Bilgi Oranı(A)=Kazanç(A)/Bölünme Bilgisi(A)"

Eşitlik sonucunda, en yüksek kazanım oranına sahip nitelik seçilir

C4.5 algoritmasında, ağaç oluşturma sırasında en iyi bilgi kazancını sağlayacak bir eşik değer belirlenir. Bunun için tüm değerler küçükten büyüğe doğru $\{v_1, v_2, ..., v_m\}$ şeklinde sıralanır. Bu durumda seçilen eşik değeri v_i ile v_{i+1} arasında olursa $\{v_1, v_2, ..., v_i\}$ ile $\{v_{i+1}, v_{i+2}, ..., v_m\}$ gibi iki grup ortaya çıkar. Buradan da görüldüğü üzere *m-1* adet eşik değeri seçilebilir. Bu seçim işlemi için olası bütün eşik değerleri aşağıdaki Eşitlik (15) ile hesaplanır (Quinlan, 1993; Breiman, 2001).

$$t_i = \frac{v_i + v_{i+1}}{2} \quad (15)$$

Bu şekilde nominal değerlere uygulanan bilgi oranı eşitliği, tüm eşik değerler için uygulanır ve bilgi kazanımı en iyi olan eşik değer söz konusu özelliğin eşiği olarak kabul edilir. En iyi eşik değer *e* gibi bir değer ise, bu kümedeki $v_i < e$ koşulunu sağlayan elemanlar küçük kategorisine ve $v_i > e$ koşulunu sağlayan elamanlar büyük kategorisine dahil edilir (Quinlan, 1993; Breiman, 2001). Bu algoritma hem kategorik hem de sürekli veriler üzerinde çalışabilir ve eksik verileri işleyebilir. Bazı programlarda J48 adıyla da bilinen C4.5 algoritması, öğrenme verisi çok küçük olduğunda iyi çalışmaz. Ayrıca "önce derinlik" (depth first) ilkesine göre çalışan C4.5 algoritması, aynı anda birden fazla yaprak oluşumunu da sağlayamaz (İnternet, 2012; 2017).

2.4. K-En Yakın Komşu Yöntemi (K-Nearest Neighbor)

Veri madenciliği sınıflandırma yöntemlerinden olan k-en yakın komşu yöntemi, ilk defa Fix ve Hodges (1951) tarafından, örüntü (model) tanımada kullanılmak amacıyla, parametrik olmayan bir yöntem olarak tanıtılmış ve daha sonra Cover ve Hart (1967) tarafından geliştirilmiştir. K-en yakın komşu yöntemi, gözlemlerin yer alacağı sınıfı ve en yakın komşuyu, *k*-değerine göre belirleyen bir sınıflama yöntemidir. Gözlemler veya nesneler arası uzaklığa dayalı sınıflandırma yapan denetimli veri madenciliği algoritmalarından biridir. Örüntü tanıma, yapay zeka, veri madenciliği, istatistik, bilişsel psikoloji, tıp ve biyoinformatik gibi birçok alanda kullanılmaktadır (Fix ve Hodges, 1951; Cover ve Hart, 1967).

K-en yakın komşu algoritması, uzaklık veya yakınlık hesaplaması yardımıyla sınıflandırma yapar. Özetle bu sınıflama algoritmasının temelinde, "örnek uzayında birbirine yakın olan nesneler muhtemelen aynı kategoriye aittir" düşüncesi yer alır. Algoritmanın amacı, bireyleri ya da nesneleri, bu nesnelere ait özelliklerden yararlanarak, önceden belirlenene sınıflara veya gruplara en doğru şekilde atamaktır. Yöntem ayrıca yeni bir gözlemin de sınıflamasını sağlar. Sınıflandırılmak istenen gözlem, öğrenme veri seti yardımıyla, en yakınında bulunan *k* tane gözlemden en fazla benzer olanlarla aynı veri setinde sınıflandırılması yapılır. Bir modelin oluşumunda kullanılacak olan verilerin oluşturduğu veri setine öğrenme veri seti denilmektedir (Fix ve Hodges, 1951; Cover ve Hart, 1967; Harrington, 2012).

K-en yakın komşu yöntemi; anlaşılır ve etkili sonuçlar vermesi, sürekli değişkenlerdeki eksik gözlemleri göz ardı edebilmesi, kategorik değişkende eksik gözlemleri değerlendirme seçeneğinin bulunması, cevap değişkeninin kategorik, sürekli veya ikisinin bir birleşimi olabilmesi ve parametrik olmayan yöntemlerden olması nedeniyle varsayımlarının az olması gibi birçok avantaja sahiptir. Ayrıca; en yakın komşuların sayısını veren k sayısının gerekli olması, seçilen uzaklık ölçüsünden etkilenmesi, hangi uzaklık ölçüsünün kullanılacağına dair bir kesinliğin olmaması gibi unsurlar algoritmanın dezavantajlarındandır (Cunningham ve Delany, 2007). K-en yakın komşu algoritması; veri madenciliği, istatistik, örüntü tanıma, yapay zeka, görüntü işleme, bilişsel psikoloji, biyoinformatik ve tıp (daha çok teşhise karar verme süreçleri) gibi pek çok alanda kullanılmaktadır (Mao ve ark., 2015).

2.4.1. K-en yakın komşu algoritmasının adımları

K-en yakın komşu algoritması, gözlemleri diğer olgulara benzerliklerine göre sınıflandırmak için kullanılır. Öğrenilen kalıplara veya modellere tam eşleşme yapılmadan, veri modellerini tanımanın bir yolu olarak geliştirilmiştir. Benzer gözlemler birbirine yakındır (komşudur) ve benzer olmayan gözlemler birbirinden uzaktır. Dolayısıyla iki gözlem arasındaki uzaklık, birbirine benzemezliği belirleyen bir kriterdir. Yeni bir gözlemin modeldeki gözlemlerden uzaklıkları hesaplanır. Bu gözlem, en fazla tekrar eden/benzer kategoriye atanır (Fix ve Hodges, 1951; Cover ve Hart, 1967).

Bu yöntemi uygularken (Cover ve Hart; 1967; Bridge, 2013);

- Yeni gözlemin, veri setindeki bütün gözlemlere uzaklığı hesaplanır,
- Bulunan bu uzaklık değerleri sıralanır,
- *k* adet en küçük uzaklığa sahip gözlem seçilir,
- *k* gözlemde en fazla tekrarlayan (majority voting) kategori, sınıf değeri olur.

K gözlemin içinde en fazla tekrarlayan kategoriyle sınıf değeri elde edilmesinin yanında, ağırlıklı komşu seçimi (weighted voting) de kullanılabilir. Eşitlik (16)'da görüldüğü üzere, uzaklıkların tersi ya da tersinin karesi ağırlık olarak kullanılır. Her bir sınıf için hesaplanan ağırlıklar yardımıyla, en fazla ağırlığa sahip kategori sınıf değeri olarak seçilmektedir. Öğrenme veri setindeki gözlemler (x_i) ve test örneğindeki gözlemler (x_q) olmak üzere, *k*-en yakın gözlemlerin uzaklığını artırarak, Eşitlik (15) yardımıyla ağırlıklı k-en yakın komşu haline getirmek mümkündür (Cover ve Hart; 1967; Zhou ve ark., 2009; Bridge, 2013).

$$W \equiv \frac{1}{d(x_q,\ x_i)^2} \tag{16}$$

Sınıflandırma işleminin başında, veriler sayısal değerlere dönüştürülür ve kaç tane en yakın komşuya (*k*) bakılacağı belirlenir. Test örneğindeki gözlemlerin sınıfı belirlenirken, her gözlemin öğrenme veri setindeki gözlemlere olan uzaklıkları hesaplanır ve en yakın *k* tane gözlem seçilir. Uzaklık hesaplanırken; Öklid, Manhattan (City Block), Minkowski, Chebyshev, Dilca gibi farklı uzaklık ölçülerinden yararlanılabilir (Fix ve Hodges, 1951; Cover ve Hart, 1967; Bridge, 2013).

Verilerin ön işleme süreci

K-en yakın komşu algoritması, isteğe bağlı olarak veriyi; "öğrenme" (training) ve "test" (holdout) veri seti olmak üzere ikiye ayırır. Öğrenme veri seti, modelin oluşumunda kullanılır. Test veri seti ise modelin bağımsız olarak değerlendirilmesinde kullanılır. Sürekli değişkenlerdeki eksik gözlemler göz ardı edilebilir. Kategorik değişkenlerde ise eksik gözlemlerin değerlendirilmesi seçeneği bulunmaktadır. Benzer kategoriler birleştirilerek veya model uygulanmadan önce az gözlenen kategoriler çıkarılarak kategori sayısı azaltılabilir. Ayrıca, aykırı gözlemler de modelden çıkarılabilir (Fix ve Hodges, 1951; Cover ve Hart, 1967).

K-en yakın komşu algoritması, sürekli yapıdaki gözlemleri sınıflandırmak için kullanıldığında, en yakın komşuların yaklaşık bir (ortalama veya medyan) değeri yeni bir gözlemin sınıflandırma tahmini için kullanılır. K-en yakın komşu algoritması isteğe bağlı olarak değişkenleri yeniden ölçeklendirerek (normalize ederek) modeli eğitmeden önce ölçek süreklilik tahmini yapar. Sürekli gözlemlerin normalleştirilme işlemi Eşitlikte (17) verilmiştir (Fix ve Hodges, 1951; Cover ve Hart, 1967; Zhou ve ark., 2009).

$$x_{pn} = \frac{2(x_{pn}^{0} - min(x_{p}^{0}))}{max(x_{p}^{0}) - min(x_{p}^{0})} \qquad (17)$$

Eşitlik (17)'de; x_{pn}, n gözlemin normalize edilmiş değeri, x^0_p, n gözlemin orijinal değeri, $min(x^0_p)$, tüm öğrenme durumları için gözlemin minimum değeri ve $max(x^0_p)$, tüm öğrenme durumları için gözlemin maksimum değeridir. Değişkendeki gözlemler kategorik yapıda ise model bu gözlemleri en iyi kategoriye sınıflamak/atamak için kullanır. Modelde, kategorik özellikler binary (ikili) olarak da kodlanabilir ve bu şekilde kullanılabilir (Fix ve Hodges, 1951; Cover ve Hart, 1967).

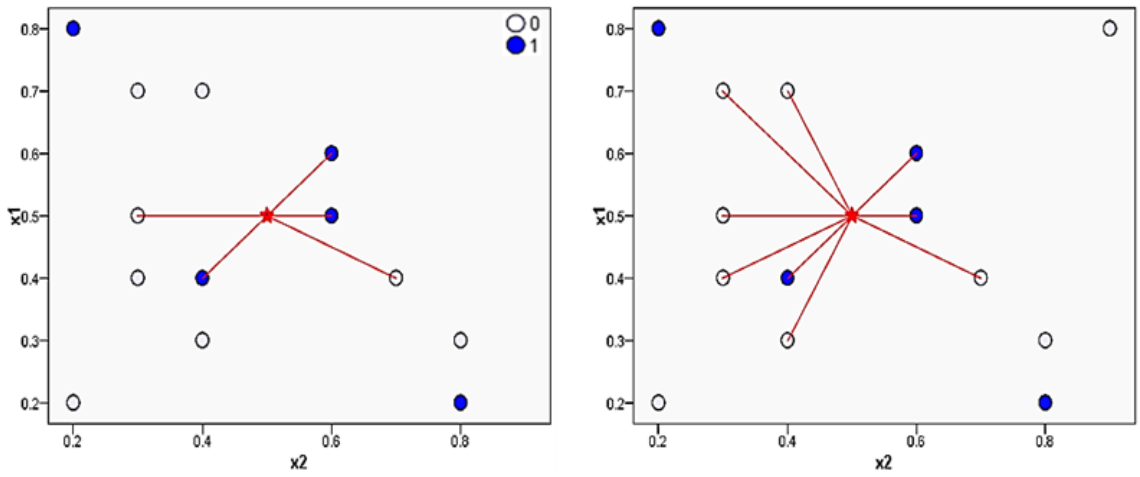
K-en yakın komşu algoritmasında algoritmaya değişken seçiminde ileriye doğru (forward) seçim yöntemi kullanılmaktadır. Değişkenler sırayla seçilir ve her adımda seçilen değişken, hata oranının veya hata kareler toplamının minimum olmasını sağlayan değişkendir (Cover ve Hart, 1967; Cunningham ve Delany, 2007).

Modele yeni bir değişken eklenmesiyle modelin daha fazla geliştirilemeyeceği anlaşıldığı durumda algoritma durur (Cover ve Hart, 1967).

2.4.2. K sayısı ve sınıflandırmaya etkisi

K-en yakın komşu algoritması, *n* boyutlu özellik uzayında gözlemleri sınıflandırmak ya da tahmin etmek için en yakın komşu örneklerini kullanır. K-en yakın komşu algoritmasında, yeni gözlemi sınıflandırabilmek için kaç adet en yakın komşu sayısının dikkate alınacağı, *k* gibi bir pozitif tam sayı ile belirtilir. *k*=1 olması durumunda, sınıflandırılmaya çalışılan yeni gözlem, en yakın komşusunun bulunduğu sınıfa dahil olacaktır. Bu yöntem kestirim için de kullanılmaktadır. *K* sayısı örnek sayısına (*N*) yaklaştıkça, komşu olan nesnelerden fazla olanların bulunduğu kategoriye atama (sınıflama) yapılır. Veri setinde yer alan tüm veriler dikkate alındığında ise en fazla tekrar eden kategoriye atama yapılmaktadır. Kısaca *k*, yeni bir gözlemin sınıflandırılmasında, kaç adet en yakın komşunun dikkate alınacağını belirten sayıdır (Cover ve Hart, 1967).

Yeni bir gözlemin iki farklı *k* değeri kullanılarak nasıl sınıflandırılacağı Şekil 5'te gösterilmektedir. *k*=5 olduğunda (Şekil 5a), en yakın komşuların çoğunluğu kategori 1'e ait olduğu için yeni gözlem kategori 1'e atanır. *k*=9 olduğunda (Şekil 5b) ise en yakın komşuların çoğunluğu 0 kategorisine ait olduğundan, yeni gözlem bu kategoriye atanır.



(a) *k*=5 alındığında (b) *k*=9 alındığında

Şekil 5. İki farklı *k* sayısı için sınıflandırma (atama)

K'nın seçiminde çapraz geçerlik (cross validation) testi

Bir modelin doğruluğunun test edilmesinde sık kullanılan yöntemlerden birisi de çapraz geçerlik testidir. Bu yöntemde veri seti rastgele iki veya daha fazla gruba ayrılır. İlk aşamada gruplardan biri üzerinde model öğrenmesi, diğeri üzerinde ise test işlemi yapılır. İkinci aşamada model ve test grupları yer değiştirir ve elde edilen hata oranlarının ortalaması kullanılır. Birkaç bin veya daha az örneklem genişliğine sahip veri tabanlarında da verilerin *n* gruba ayrıldığı *n*-kat çapraz geçerlik (*n*-fold cross validation) testi tercih edilebilir (Cover ve Hart, 1967; Berson ve ark., 2000).

Çapraz geçerlik testi için literatürde en çok tercih edilen *n* (kat) değerinin 10 olduğu gözlenmektedir. Verilerin 10 gruba ayrılması durumunda, ilk aşamada birinci grup test, diğer gruplar öğrenim için kullanılır. Bu süreç her defasında bir grubun test, diğer grupların öğrenme amaçlı kullanılması ile sürdürülür. Sonuçta elde edilen 10 hata oranının ortalaması, kurulan modelin tahmini hata oranı olacaktır (Cover ve Hart, 1967; Berson ve ark., 2000).

Modelde *n*-kat çapraz geçerlikte, en yakın komşu sayısı otomatik olarak belirlenirken; hata oranı, k_{min} ile k_{max} arasında değerler alır. Öğrenme veri setinin (*1, 2, ..., n*) tamsayı değerlerine sahip değişken olduğu varsayılırsa, çapraz geçerlik testi (CV_k) Eşitlik (18)'deki gibi olacaktır. (Cover ve Hart, 1967).

$$CV_k = \sum_{i=1}^{n} e_n/n \tag{18}$$

2.4.3. Benzerlik, uzaklık ve yakınlık (proximity) ölçüleri

Benzerlik, iki özellik veya nesne arasındaki ilişkinin gücünü yansıtan sayısal bir büyüklüktür ve bunun ölçümü oldukça zordur. Bu büyüklük genelde ±1 aralığındadır ve normalize edilerek 0 ile +1 aralığına çekilebilir. Uzaklık ise benzemezliği ölçer. Benzemezlik, iki nesne arasındaki uyumsuzluğun bir ölçüsü olarak da düşünülebilir. Bu ölçüler nesne için özellikler uzayında koordinat değerleri olarak da kullanılabilir. Sınıflandırma işleminin başında, veriler sayısal değerlere dönüştürülür ve kaç tane en yakın komşuya (*k*) bakılacağı belirlenir. Test örneğindeki gözlemlerin sınıfı belirlenirken, öğrenme veri setindeki gözlemlere olan uzaklıkları hesaplanır ve en yakın

k tane gözlem seçilir (Cover ve Hart; 1967; Teknomo, 2006; Cunningham ve Delany, 2007). Uzaklık hesaplanırken; Öklid, Manhattan (City Block), Minkowski, Chebyshev, Dilca gibi farklı uzaklık ölçülerinden yararlanılabilir (Fix ve Hodges, 1951; Cover ve Hart, 1967; Bridge, 2013).

Uzaklık ölçüleri; verilerin sürekli ve/veya kategorik değişkenler içermesi durumunda farklılık göstermektedir. Değişkenlerin sürekli ve kategorik olması durumunda, gözlemler arası uzaklık için sık kullanılan ölçüler Öklid ve Manhattan'dır. Veri seti, tüm boyutlar bakımından sürekli değişkenler içermesi durumunda Öklid uzaklık fonksiyonun kullanılması, kategorik değişkenler içermesi durumunda ise Manhattan uzaklık fonksiyonunun kullanılması önerilmektedir (Dekhtyar, 2009). Bu bağlamda, çalışmada gözlemler arası uzaklıkların hesaplanmasında, veri tipi (kategorik ve sürekli) göz önüne alınarak Öklid ve Manhattan (City Block) ölçüleri kullanılmıştır.

Öklid uzaklığı, *x* ve *y* gibi iki nokta arasındaki doğrusal uzaklıktır. Gözlem değerleri arasındaki ağırlıklı karesel farkların, tüm boyutlar üzerinden toplamının karekökü olarak hesaplanır. $d_{(x,\,y)}$, *x* ve *y* noktaları arasındaki uzaklığı göstermek üzere; Öklid uzaklık ölçüsüne ait Eşitlik (19) aşağıdaki gibidir (Dekhtyar, 2009).

$$d(x,y) = \sqrt{\left(\sum_{j=1}^{N}(x_j - y_j)^2\right)} \tag{19}$$

Manhattan uzaklığı, gözlem boyutları arasındaki ağırlıklı mutlak farkların tüm boyutları üzerinden toplamı (hipotenüse göre iki dik noktanın birleştiği mesafe) şeklindedir. $P=(x_1, x_2, \ldots, x_n)$ ve $Q=(y_1, y_2, \ldots, y_n)$ olmak üzere P ve Q değişkenlerinin noktaları arasındaki Manhattan uzaklığı Eşitlik (20)'deki gibi hesaplanır (Kresse ve Danko, 2012).

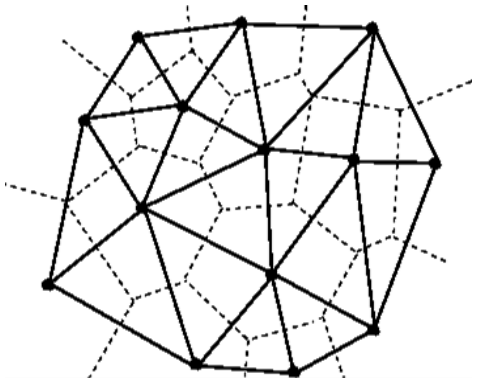
$$Manhattan_{PQ} = \sum_{p=1}^{P} w_{(p)}(x_{(p)P} - y_{(p)Q})^2 \tag{20}$$

Bu Eşitlik (20), aşağıdaki Eşitlikteki (21) gibi de ifade edilebilir;

$$d(i,j) = (|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|) \tag{21}$$

Nesneler arası yakınlık, herhangi bir uzaklık ölçüsü olabilir. Yakınlık bulmada ilk akla gelen problem, verilen bir noktalar kümesinde birbirine en yakın noktaların bulunmasıdır. Bu ölçüler için en çok tercih edilen modellemeler, Voronoi diyagramı ve Delaunay üçgenlemesidir. Birçok alanda uygulama alanı bulmuş olan bu modellemeler, genel olarak tüm noktaların bir noktaya olan yakınlık alanlarının bulunmasıyla ilgilidir (McAllister ve Snoeyink, 2000).

Voronoi diyagramı ve Delaunay üçgenleri (üçgenlemesi) yaklaşımı: Noktalar kümesindeki verilerin mozaiklere bölünmesi olarak sunulan ve adını aldığı Voronoi (1903) tarafından önerilmiş bir veri sınıflandırma yöntemidir. Diğer bir ifadeyle, bir noktaya en yakın diğer noktaların bölgelere bölümlenmesi ile elde edilen bu yapıya Voronoi diyagramı denilmektedir. Nokta kümeleri arasındaki uzaklıklar, Öklid veya Mahalanobis uzaklığı gibi farklı ölçüler kullanılarak hesaplanabilir (Fan ve Zhang, 2004; Maynard, 2005). Değişken setini oluşturan verilerin, noktalara en yakın bölgelere bölümlenmesi ve ardından noktaları birleştiren doğru parçalarının ortalarından geçen dikmeler yardımıyla elde edilen Voronoi diyagramı ile bu diyagramdaki üç noktadan geçecek daireler çizilerek elde edilen noktalardan geçen Delaunay üçgenleri örneği aşağıda (Şekil 6) verilmiştir (McAllister ve Snoeyink, 2000; Zhao, 2005).



Şekil 6. Voronoi diyagramı ve Delaunay üçgenleri (Voronoi 1903; Delaunay, 1934)

Voronoi diyagramının tamamlayıcısı olan Delaunay üçgenleri modellemesi, Delaunay (1934) tarafından önerilmiştir. Delaunay üçgenleri oluşturulurken, üç noktadan geçecek ve içerisinde başka herhangi bir nokta bulunmayacak şekilde daireler çizilir. Bir dairenin üzerindeki noktalara ait üçgenlerin her biri birer Delaunay üçgenidir (Şekil 6). Mühendislikte ve üç boyutlu alanlarda Delaunay üçgenleri sıkça kullanılmaktadır (Lawson, 1997; Sibson, 1997; McAllister ve Snoeyink, 2000).

3. GEREÇ VE YÖNTEM

3.1. Gereç

Çalışmanın uygulama materyali olarak, Çetin ve ark. (2015), tarafından elde edilen 910 bireye ait veri seti kullanılmıştır. Bu veri setinden; 4'ü sürekli yapıda olmak üzere toplam 34 adet değişken seçilmiştir. Veri seti, yaşlarına göre 3 gruptan (Kontrol n=301 | $\bar{x}$=27, Adolesan n=306 | $\bar{x}$=16, İleri yaş n=303 | $\bar{x}$=41) oluşmaktadır. Bebeklerin %31.2'sinin doğum ağırlığı "düşük (≤2500 g)" kategoride yer alırken, %68.8'inin doğum ağırlığı "normal (>2500 g)" kategoride yer almıştır. Çalışmada kullanılan sürekli ve kategorik değişkenler ile bunlara ait tanımlayıcı istatistikler (ortalama, standart sapma, minimum, maksimum, sayı ve yüzde) Tablo 5'te özetlenmiştir.

Tablo 5. Çalışmada ele alınan değişkenler ve tanımlayıcı istatistikler

		n	Ortalama	Std. Sapma	Minimum	Maksimum
Anne Yaşı	Kontrol	301	26.60	4.20	20.0	39.0
	Adolesan	306	15.78	0.89	13.0	17.0
	İleri Yaş	303	40.93	0.95	40.0	44.0
	Genel	910	27.73	10.6	13.0	44.0
Doğum Haftası		910	35.62	3.58	26.0	40.0
Gravida		910	2.69	1.60	1.00	7.00
Parite		910	1.18	1.23	0.00	6.00

		n	%			n	%			n	%
Bebek Doğ. Ağır.	Düşük	284	31.2	Polihidromnios	Yok	873	95.9	SGA	Yok	739	81.2
	Normal	626	68.8		Var	37	4.1		Var	171	18.8
Doğum Şekli	Normal	700	76.9	Oligohidramnio	Yok	824	90.5	Gestasyonel Diyabet	Yok	812	89.2
	Sezaryen	210	23.1		Var	86	9.5		Var	98	10.8
Preeklampsi	Yok	861	94.6	Fetal Ölüm	Yok	868	95.4	PROM	Yok	861	94.6
	Var	49	5.4		Var	42	4.6		Var	49	5.4
Eklampsi	Yok	897	98.6	Erken Doğum	Yok	877	96.4	Koryoamniyonit	Yok	888	97.6
	Var	13	1.4		Var	33	3.6		Var	22	2.4
Anne karnında oksijensizlik	Yok	863	94.8	Amniyosentez	Yok	885	97.3	Urinenfe	Yok	801	88.0
	Var	47	5.2		Var	25	2.7		Var	109	12.0
Acil Sezaryen	Yok	851	93.5	ART	Yok	885	97.3	Fetolum	Yok	900	98.9
	Var	59	6.5		Var	25	2.7		Var	10	1.1
Oksitosin Takılma	Yok	898	98.7	Sigara	Yok	816	89.7	Antehemo	Yok	810	89.0
	Var	12	1.3		Var	94	10.3		Var	100	11.0
Prespont	Yok	809	88.9	Anemi	Yok	729	80.1	Posthemo	Yok	835	91.8
	Var	101	11.1		Var	181	19.9		Var	75	8.2
Previa	Yok	889	97.7	Çoklu Gebelik	Yok	866	95.2	Nonvertp	Yok	861	94.6
	Var	21	2.3		Var	44	4.8		Var	49	5.4
Nullipar (ilk doğumu)	Yok	564	62.0	Multiparite (çoklu doğum)	Yok	346	38.0	Gestasyonel Hipertansiyon	Yok	837	92.0
	Var	346	38.0		Var	564	62.0		Var	73	8.0

ART: Yardımcı üreme teknikleri, SGA: Bebeğin anne karnında gelişimi, Fetal Distress: Koryoamniyonit:Uterus enfeks. Anne karnında oksijensizlik, Prom: Erken su gelme, Nonvertp: Geliş anomalisi, Prespont: Kend. erken doğum, Previa: Plasentanın rahim ağzına yapışması, Eklampsi: Gebede nöbet geçirme, Polihidromnios: Su fazlalığı, Oligohidromnio: Su azlığı, Posthemo: Kanama, Multiple: Çoklu gebelik, Amniyosentez: Sıvı alınması

3.2. Yöntem

Çalışmada, "veri madenciliğinde farklı karar ağaçları ve k-en yakın komşu yöntemlerinin incelenmesi" amacıyla tanımlayıcı istatistikleri Tablo 6'da verilen veri seti kullanılmıştır. Çalışmada "Bebek Doğum Ağırlığı (normal>2500g | düşük≤2500g)" cevap değişkeni olarak alınmış ve bu değişken ile açıklayıcı değişkenler arası ilişkileri belirlemek üzere, "karar ağaçları" yöntemlerinden; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5" algoritmaları kullanılmıştır. Yine bu verilere "k-en yakın komşu" algoritması da uygulanarak, bu yöntemlerin performansları incelenmiştir. Verilerin analizi için "Python (3.5)", "Weka (3.9) ve "SPSS (25)" istatistik paket programları kullanılmıştır.

Karar ağacı analizlerinde minimum hata oranlarına ulaşabilmek amacıyla yapılan ön deneme sonuçlarına göre 10-kat çapraz geçerlik testi tercih edilmiştir. Benzer şekilde, elde edilen ön deneme sonuçlarına göre; CART, CHAID, Ayrıntılı CHAID ve QUEST için "minimum dal düğüm sayısı" 25 ve "minimum yaprak düğüm sayısı 5, olarak belirlenmiştir. Rastgele Orman ve C4.5 için "minimum dal düğüm sayısı" 25 ve "minimum yaprak düğüm sayısı 10, olarak belirlenmiştir. CART, QUEST, Rastgele Orman ve C4.5 algoritmalarında "maksimum ağaç derinliği" (otomatik seçimle) 5 olarak alınmıştır. CHAID ve Ayrıntılı CHAID algoritmalarında ise ağaç derinliği 3 olarak belirlenmiştir. Karar Ağaçları analizi için "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5"e ait ön deneme ile sınıflandırma performansı sonuçları elde edilmiştir. Bu algoritmaların performansını belirlemede, gerçek ve tahmin değerlerine ait; "Risk Katsayısı, Duyarlık (sensitivity), Özgüllük (specificity) ve Genel Doğruluk Oranı (accuracy)" ölçütleri kullanılmıştır. Ayrıca her bir algoritmaya ait "karar ağacı diyagramı", "karar ağacı tahmin sonucu", "sınıflandırma sonucu", "risk oranı" ve "ortalama mutlak hata (MAPE)" verilmiştir.

"K-en yakın komşu (KNN)" yönteminin doğruluğunun test edilmesinde minimum hata oranına sahip *k* değerine ulaşılmaya çalışılmıştır. K-en yakın komşu yönteminin performansını açıklayabilmek için ön deneme ile sınıflandırma performansı sonuçları elde edilmiştir. Bu yöntemin performansını belirlemede, (öğrenme ve test verisi için ayrı ayrı) gerçek ve tahmin değerlerine ait; "Duyarlık (sensitivity), Özgüllük

(specificity) ve Genel Doğruluk Oranı (accuracy)" ölçütleri kullanılmıştır. Ayrıca, öğrenme ve test verisi için ayrı ayrı; *k* seçiminde yanlış sınıflandırma (hata/risk) oranları (%) tespit edilmiştir. K'nın seçiminde çapraz doğrulama olarak, "10-kat çapraz geçerlik testi" kullanılmıştır. Bu analizde "komşu sayısı" (*k*) 5 olarak alınmış ve bölümlere atama için "öğrenme (training)" verisi %70 oranında belirlenerek işlem yapılmıştır. Çıkan sonuçlara göre açıklayıcı değişkenlerin cevap değişkeni (bebek doğum ağırlığı) üzerindeki önem sıralaması verilmiştir. K-en yakın komşu yöntemi uygulamasında, gözlemler arası uzaklıkları hesaplamada, veri tipi (kategorik, sürekli ve sıralı) göz önüne alınarak iki farklı uzaklık ölçüsü (Öklid ve Manhattan/City Block) kullanılmıştır.

Karar ağaçları ve k-en yakın komşu algoritmalarının performansını belirlemede, gerçek ve tahmin değerlerine ait; "Risk Katsayısı, Duyarlık (sensitivity), Özgüllük (specificity), Genel Doğruluk Oranı (accuracy)" ve "MAPE katsayısı" ölçütleri Tablo 6'daki gibi hesaplanmıştır.

Tablo 6. Performans ölçülerinin (tanı testlerinin) hesaplanması

Performans Ölçütü	Açıklama	Hesaplama
Duyarlık (Sensitivity)	Gerçekte "Pozitif" olanlar içinden, "Pozitif" olarak tahmin edilenlerin oranı	GP/(GP+YN)
Özgüllük (Specificity)	Gerçekte "Negatif" olanlar içinden, "Negatif" olarak tahmin edilenlerin oranı	GN/(GN+YP)
G. Doğr. Oranı (Accuracy)	Gerçekte "Pozitif" ve "Negatif" olanların toplam içindeki oranı	(GP+GN)/(GP+YP+YN+GN)
MAPE	Ortalama mutlak yüzde hata (mean absolute percentage error)	| Gerçek-Tahmin | / Gerçek * 100

GP: Gerçek Pozitif, GN: Gerçek Negatif, YP: Yanlış Pozitif, YN: Yanlış Negatif

Düşük doğum ağırlığı, doğum ağırlığının 2500 gramdan az (≤2500g) olması durumudur. Bu çalışmada, "düşük doğum ağırlığındaki bebekler", tanı testi bakımından pozitifliği göstermektedir. Buna göre; "gerçekte pozitif (düşük doğum ağırlıklı) olanlar içinden, "Pozitif" olarak tahmin edilenlerin oranı "Duyarlığı (sensitivity)" vermektedir. Dolayısıyla elde edilen uygulama sonuçlarına göre "Duyarlığı yüksek" bulunan algoritmaların performansının, klinik bakımdan önemli olacağı bildirilmektedir (Singh ve ark., 2009; Zenciroğlu ve ark., 2009).

4. BULGULAR

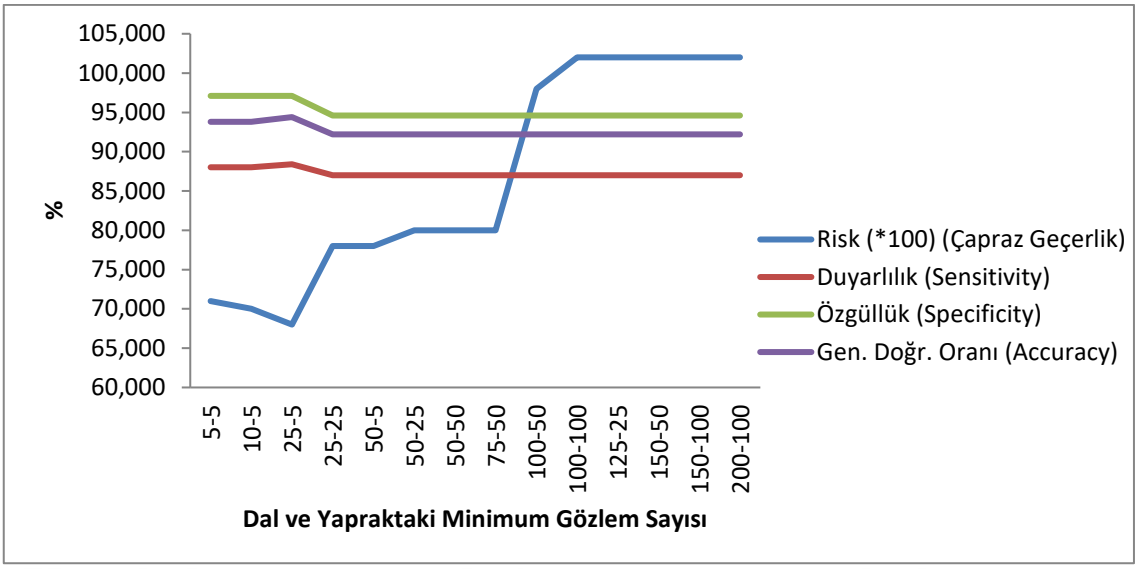
4.1. Karar Ağaçları Analizi Sonuçları

Uygulamanın ilk aşamasında, karar ağacı yöntemlerinden CART, CHAID, Ayrıntılı CHAID ve QUEST'e ait ön deneme sonuçları elde edilmiştir. Buna göre CART algoritmasına ait performans sonuçları Tablo 7'de verilmiştir.

Tablo 7. CART'a ait ön deneme (performans) sonuçları

CART					
K-kat Çap. Geç.	Min. Gözlem (Dal-Yaprak) Sayısı	Risk/Hata (Çapraz Geçerlik)	Duyarlık (Sensitivity)	Özgüllük (Specificity)	Gen. Doğr. Oranı (Accuracy)
10	5-5	0.071	88.00	97.10	93.80
10	10-5	0.070	88.00	97.10	93.80
10	25-5	0.068	88.40	97.10	94.40
10	25-25	0.078	87.00	94.60	92.20
10	50-5	0.078	87.00	94.60	92.20
10	50-25	0.078	87.00	94.60	92.20
10	50-50	0.078	87.00	94.60	92.20
10	75-50	0.078	87.00	94.60	92.20
10	100-50	0.098	87.00	94.60	92.20
10	100-100	0.102	87.00	94.60	92.20
10	125-25	0.102	87.00	94.60	92.20
10	150-50	0.102	87.00	94.60	92.20
10	150-100	0.102	87.00	94.60	92.20
10	200-100	0.102	87.00	94.60	92.20

CART (classification and regression trees) algoritmasının performansını belirlemek üzere yapılan ön deneme işlem seçenekleri olarak; "10-kat çapraz geçerlilik", "5 ile 200 aralığında dal düğüm sayısı" ve "5 ile 100 aralığında yaprak düğüm sayısı" kullanılarak alternatif sonuçlar elde edilmiştir. Buna göre; bu algoritmanın performansını gösteren ölçütlerinden, gerçek ve tahmin değerlerine ait; "risk/hata katsayısı, duyarlık (sensitivity), özgüllük (specificity) ve genel doğruluk oranı (accuracy)" ölçütleri verilmiştir (Tablo 7 ve Şekil 7). Tahmin değerlerine ait performans sonuçları bakımından; "en düşük risk/hata oranı" ve "optimum; duyarlık, özgüllük, genel doğruluk oranı" göz önüne alındığında, dal-yaprak düğüm sayısının 25-5 olması bu ver seti için en iyi sonuçları vermektedir. Diğer alternatif denemelerin sonuçları da Tablo 7 ve Şekil 7'de verilmiştir.



Şekil 7. CART'a ait ön deneme (performans) sonuçları

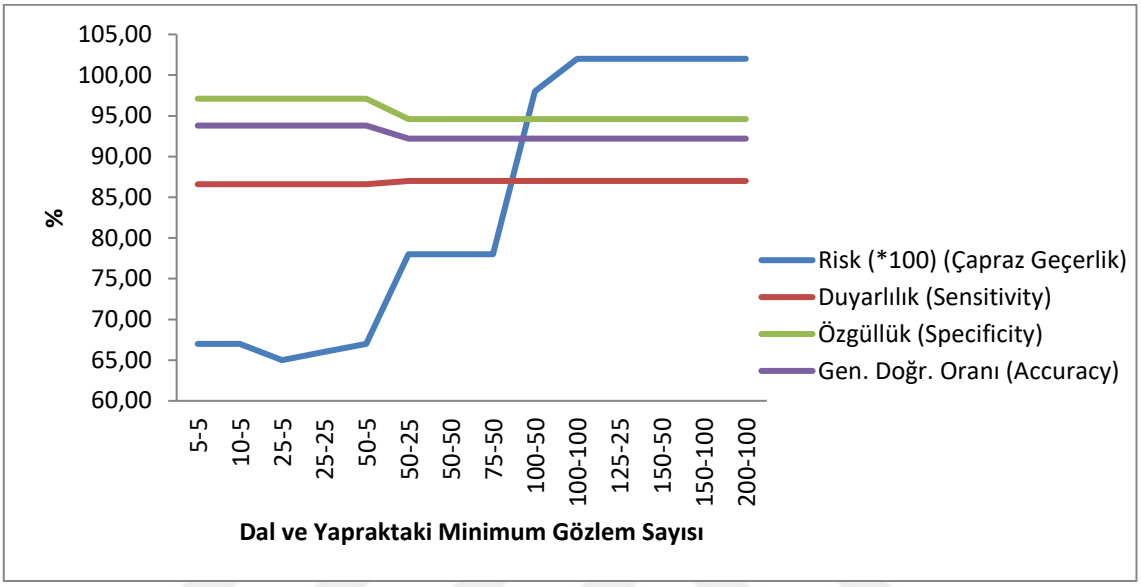
Karar ağacı yöntemlerinden CHAID (chi-squared automatic interaction detector) algoritmasına ait ön deneme sonuçları Tablo 8'de verilmiştir.

Tablo 8. CHAID'e ait ön deneme (performans) sonuçları

CHAID					
K-kat Çap. Geç.	Min. Gözlem (Dal-Yaprak) Sayısı	Risk/Hata (Çapraz Geçerlik)	Duyarlık (Sensitivity)	Özgüllük (Specificity)	Gen. Doğr. Oranı (Accuracy)
10	5-5	0.067	86.60	97.10	93.80
10	10-5	0.067	86.60	97.10	93.80
10	25-5	0.065	86.60	97.10	93.80
10	25-25	0.066	86.60	97.10	93.80
10	50-5	0.067	86.60	97.10	93.80
10	50-25	0.078	87.00	94.60	92.20
10	50-50	0.078	87.00	94.60	92.20
10	75-50	0.078	87.00	94.60	92.20
10	100-50	0.098	87.00	94.60	92.20
10	100-100	0.102	87.00	94.60	92.20
10	125-25	0.102	87.00	94.60	92.20
10	150-50	0.102	87.00	94.60	92.20
10	150-100	0.102	87.00	94.60	92.20
10	200-100	0.102	87.00	94.60	92.20

CHAID algoritmasının performansını belirlemek üzere yapılan ön deneme, işlem seçenekleri olarak; "10-kat çapraz geçerlilik", "5 ile 200 aralığında dal düğüm sayısı" ve "5 ile 100 aralığında yaprak düğüm sayısı" kullanılarak alternatif sonuçlar elde edilmiştir. Buna göre; bu algoritmanın performansını gösteren ölçütlerden gerçek ve tahmin değerlerine ait; "risk/hata katsayısı, duyarlık (sensitivity), özgüllük (specificity) ve genel doğruluk oranı (accuracy)" ölçütleri verilmiştir (Tablo 8 ve Şekil 8). Tahmin

değerlerine ait performans sonuçları bakımından; "en düşük risk/hata oranı" ve "optimum; duyarlık, özgüllük, genel doğruluk oranı" göz önüne alındığında, dal-yaprak düğüm sayısının sırasıyla 25 ve 5 olması bu ver seti için en iyi sonucu vermektedir. Diğer alternatif denemelerin sonuçları da Tablo 9 ve Şekil 8'de verilmiştir.



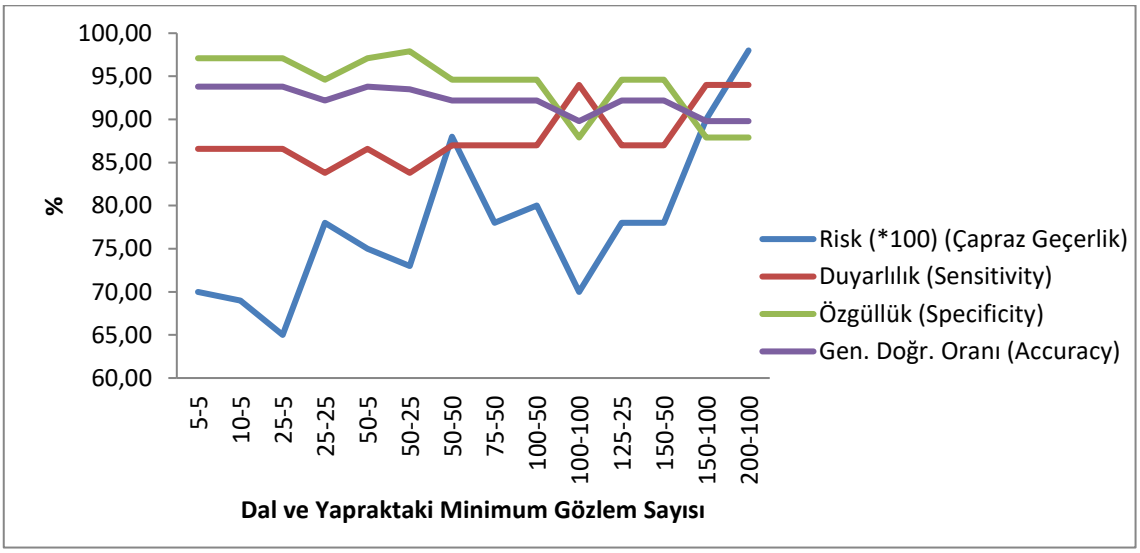
Şekil 8. CHAID'e ait ön deneme (performans) sonuçları

Karar ağacı yöntemlerinden Ayrıntılı (exhaustive) CHAID algoritmasına ait ön deneme sonuçları Tablo 9'da verilmiştir.

Tablo 9. Ayrıntılı CHAID'e ait ön deneme (performans) sonuçları

Ayrıntılı CHAID					
K-kat Çap. Geç.	Min. Gözlem (Dal-Yaprak) Sayısı	Risk/Hata (Çapraz Geçerlik)	Duyarlık (Sensitivity)	Özgüllük (Specificity)	Gen. Doğr. Oranı (Accuracy)
10	5-5	0.070	86.60	97.10	93.80
10	10-5	0.069	86.60	97.10	93.80
10	25-5	0.065	86.60	97.10	93.80
10	25-25	0.078	83.80	94.60	92.20
10	50-5	0.075	86.60	97.10	93.80
10	50-25	0.073	83.80	97.90	93.50
10	50-50	0.088	87.00	94.60	92.20
10	75-50	0.078	87.00	94.60	92.20
10	100-50	0.080	87.00	94.60	92.20
10	100-100	0.070	94.00	87.90	89.80
10	125-25	0.078	87.00	94.60	92.20
10	150-50	0.078	87.00	94.60	92.20
10	150-100	0.090	94.00	87.90	89.80
10	200-100	0.098	94.00	87.90	89.80

Ayrıntılı CHAID algoritmasının performansını belirlemek üzere yapılan ön deneme işlem seçenekleri olarak; "10-kat çapraz geçerlilik", "5 ile 200 aralığında dal düğüm sayısı" ve "5 ile 100 aralığında yaprak düğüm sayısı" kullanılarak alternatif sonuçlar elde edilmiştir. Buna göre; bu algoritmanın performansını gösteren ölçütlerden gerçek ve tahmin değerlerine ait; "risk/hata katsayısı, duyarlık (sensitivity), özgüllük (specificity) ve genel doğruluk oranı (accuracy)" ölçütleri verilmiştir (Tablo 9 ve Şekil 9). Tahmin değerlerine ait performans sonuçları bakımından; "en düşük risk/hata oranı" ve "optimum; duyarlık, özgüllük, genel doğruluk oranı" göz önüne alındığında, dal-yaprak düğüm sayısının sırasıyla 25 ve 5 olması bu ver seti için en iyi sonucu vermektedir. Diğer alternatif denemelerin sonuçları da Tablo 9 ve Şekil 9'da verilmiştir.

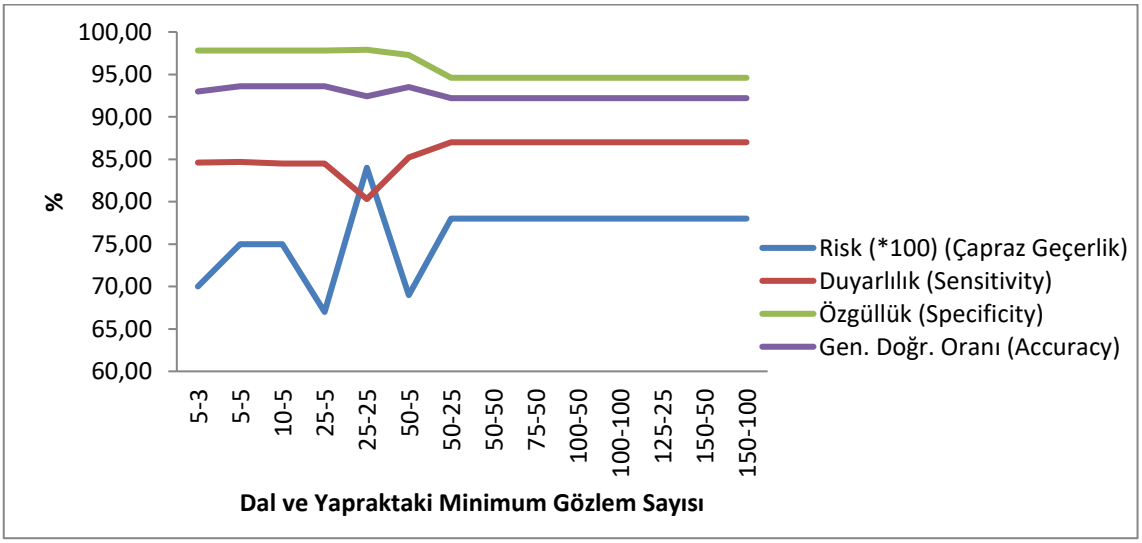


Şekil 9. Ayrıntılı CHAID'e ait ön deneme (performans) sonuçları

Tablo 10. QUEST'e ait ön deneme (performans) sonuçları

QUEST					
K-kat Çap. Geç.	Min. Gözlem (Dal-Yaprak) Sayısı	Risk/Hata (Çapraz Geçerlik)	Duyarlık (Sensitivity)	Özgüllük (Specificity)	Gen. Doğr. Oranı (Accuracy)
10	5-3	0.070	84.60	97.80	93.00
10	5-5	0.075	84.70	97.80	93.60
10	10-5	0.075	84.50	97.80	93.60
10	25-5	0.067	84.50	97.80	93.60
10	25-25	0.084	80.30	97.90	92.40
10	50-5	0.069	85.20	97.30	93.50
10	50-25	0.078	87.00	94.60	92.20
10	50-50	0.078	87.00	94.60	92.20
10	75-50	0.078	87.00	94.60	92.20
10	100-50	0.078	87.00	94.60	92.20
10	100-100	0.078	87.00	94.60	92.20
10	125-25	0.078	87.00	94.60	92.20
10	150-50	0.078	87.00	94.60	92.20
10	150-100	0.078	87.00	94.60	92.20

QUEST'in performansını belirlemek amacıyla yapılan ön deneme işlem seçenekleri olarak; "10-kat çapraz geçerlilik", "5 ile 200 aralığında dal düğüm sayısı" ve "5 ile 100 aralığında yaprak düğüm sayısı" kullanılarak alternatif sonuçlar elde edilmiştir. Buna göre; bu algoritmanın performansını gösteren ölçütlerden gerçek ve tahmin değerlerine ait; "risk/hata katsayısı, duyarlık (sensitivity), özgüllük (specificity) ve genel doğruluk oranı (accuracy)" ölçütleri verilmiştir (Tablo 10 ve Şekil 10). Tahmin değerlerine ait performans sonuçları bakımından; "en düşük risk/hata oranı" ve "optimum; duyarlık, özgüllük, genel doğruluk oranı" göz önüne alındığında, dal-yaprak düğüm sayısının sırasıyla 25 ve 5 olması bu ver seti için en iyi sonucu vermektedir. Diğer alternatif denemelerin sonuçları da Tablo 10 ve Şekil 10'da verilmiştir.



Şekil 10. QUEST'e ait ön deneme (performans) sonuçları

Karar ağacı algoritmalarının bu ver seti üzerindeki ön denemesinde, tahmin değerlerine ait en iyi performansı gösteren ölçütler (en düşük risk/hata oranı ve optimum duyarlık, özgüllük, genel doğruluk oranı) göz önüne alınarak işlem seçenekleri belirlemiştir. Bu algoritmalar için belirlenen işlem seçenekleri Tablo 11'de verilmiştir. Buna göre, CART ve QUEST için; cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği (otomatik seçim) "5"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "5" olarak işleme alınmıştır. Benzer şekilde, CHAID ve Ayrıntılı CHAID için; cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği (otomatik seçim) "3"; minimum dal düğüm sayısı "25 ve minimum yaprak düğüm sayısı "5" olarak işleme alınmıştır. Rastgele Orman ve C4.5 algoritmalarının uygulamasında işlem seçenekleri olarak, çapraz geçerlik "10-kat"; maksimum ağaç derinliği "5"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "5" olarak alınmıştır.

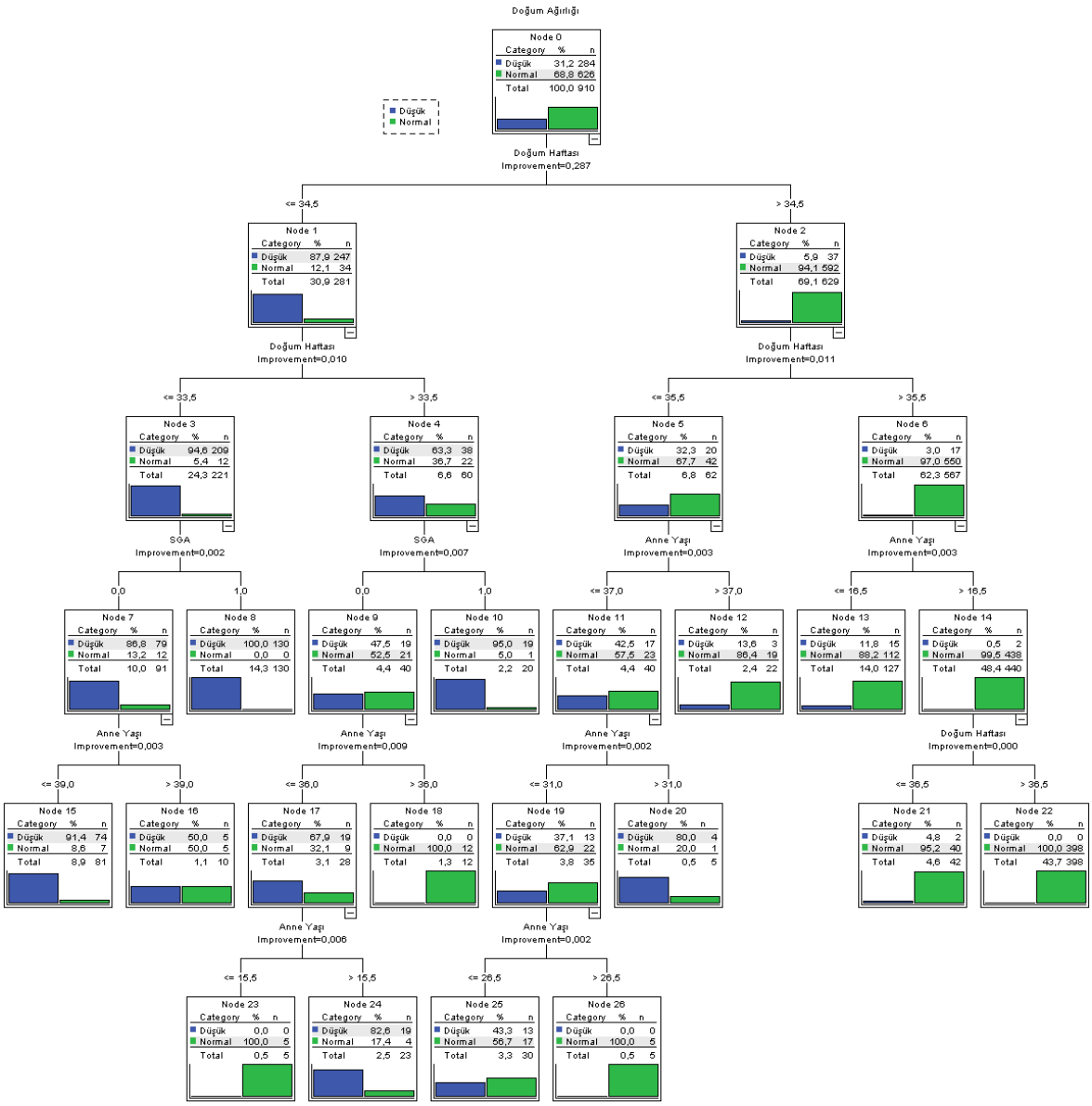
Tablo 11. Karar ağacı algoritmalarında işlem seçenekleri

Seçenekler	**CART**	**CHAID**	**Ayrıntılı CHAID**	**QUEST**	**Rastgele Orman**	**C4.5**
Çapraz Geçerlik	10-kat	10-kat	10-kat	10-kat	10-kat	10-kat
Maks. Ağaç Derinliği (Oto. Seç.)	5	3	3	5	5	5
Min. Dal Düğüm Sayısı	25	25	25	25	25	25
Min. Yaprak Düğüm Sayısı	5	5	5	5	10	10
Cevap değişkeni	Bebek Doğum Ağırlığı (Normal/Düşük)					

Karar ağacı algoritmalarında işlem seçenekleri (Tablo 11), dört ayrı karar ağacı algoritmasında uygulanmış ve analiz sonuçları aşağıdaki gibi elde edilmiştir.

4.1.1. CART algoritmasına ait analiz sonuçları

CART algoritmasının uygulamasında işlem seçenekleri olarak, cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği (otomatik seçim) "5"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "5" alınmıştır. Buna göre CART algoritmasına ait karar ağacı diyagramı ve analiz sonuçları Şekil 11 ve Tablo 12'de verilmiştir.



(*SGA: Bebeğin anne karnında gelişimi)*

Şekil 11. CART'a ait karar ağacı diyagramı

CART algoritmasına ait karar ağacı diyagramı (Şekil 11) ve tahmin sonuçları incelendiğinde (Tablo 12), cevap değişkeni (bebek doğum ağırlığı) üzerine etkili olan açıklayıcı değişkenler "maksimum bölme kriterine (improvement: belirsizliği en çok azaltan bilgi kazancına)" göre verilmiştir.

Tablo 12. CART'a ait karar ağacı tahmin sonuçları

Düğüm	Düşük		Normal		Genel		Tahmin Kategori	Dal Düğüm	Açıklayıcı Değişken		
	n	%	n	%	n	%			Değişken	Bölme Kriteri/ Bilgi Kazancı	Kesim Değeri
0	284	31.2	626	68.8	910	100.0	Normal				
1	247	87.9	34	12.1	281	30.9	Düşük	0	Doğ. Haftası	.287	≤34.5
2	37	5.9	592	94.1	629	69.1	Normal	0	Doğ. Haftası	.287	>34.5
3	209	94.6	12	5.4	221	24.3	Düşük	1	Doğ. Haftası	.010	≤33.5
4	38	63.3	22	36.7	60	6.6	Düşük	1	Doğ. Haftası	.010	>33.5
5	20	32.3	42	67.7	62	6.8	Normal	2	Doğ. Haftası	.011	≤35.5
6	17	3.0	550	97.0	567	62.3	Normal	2	Doğ. Haftası	.011	>35.5
7	79	86.8	12	13.2	91	10.0	Düşük	3	SGA	.002	Yok
8	130	100.0	0	0.0	130	14.3	Düşük	3	SGA	.002	Var
9	19	47.5	21	52.5	40	4.4	Normal	4	SGA	.007	Yok
10	19	95.0	1	5.0	20	2.2	Düşük	4	SGA	.007	Var
11	17	42.5	23	57.5	40	4.4	Normal	5	Anne Yaşı	.003	≤37.0
12	3	13.6	19	86.4	22	2.4	Normal	5	Anne Yaşı	.003	>37.0
13	15	11.8	112	88.2	127	14.0	Normal	6	Anne Yaşı	.003	≤16.5
14	2	0.5	438	99.5	440	48.4	Normal	6	Anne Yaşı	.003	>16.5
15	74	91.4	7	8.6	81	8.9	Düşük	7	Anne Yaşı	.003	≤39.0
16	5	50.0	5	50.0	10	1.1	Düşük	7	Anne Yaşı	.003	>39.0
17	19	67.9	9	32.1	28	3.1	Düşük	9	Anne Yaşı	.009	≤36.0
18	0	0.0	12	100.0	12	1.3	Normal	9	Anne Yaşı	.009	>36.0
19	13	37.1	22	62.9	35	3.8	Normal	11	Anne Yaşı	.002	≤31.0
20	4	80.0	1	20.0	5	0.5	Düşük	11	Anne Yaşı	.002	>31.0
21	2	4.8	40	95.2	42	4.6	Normal	14	Doğ. Haftası	.001	≤36.5
22	0	0.0	398	100.0	398	43.7	Normal	14	Doğ. Haftası	.001	>36.5
23	0	0.0	5	100.0	5	0.5	Normal	17	Anne Yaşı	.006	≤15.5
24	19	82.6	4	17.4	23	2.5	Düşük	17	Anne Yaşı	.006	>15.5
25	13	43.3	17	56.7	30	3.3	Normal	19	Anne Yaşı	.002	≤26.5
26	0	0.0	5	100.0	5	0.5	Normal	19	Anne Yaşı	.002	> 26.5

Tablo 12'de; "bebek doğum ağırlığına" en fazla etki eden açıklayıcı değişkenlerin "bilgi kazancı" ile "kesim değerleri" verilmiştir. Bunun yanında etkili açıklayıcı değişkelerin bulunduğu "dal düğüm numaraları" ile bu düğümlerde "bebek doğum ağırlığı (normal, düşük, genel)" bakımından "tahmin değerleri (sayı ve yüzde)" gösterilmiştir. Buna göre, bebeklerin doğum ağırlığına en fazla etki eden değişkenlerin sırasıyla; "Doğum haftası", "SGA" ve "Anne yaşı" olduğu tespit edilmiştir.

Bölme kriteri bakımından en yüksek değere sahip olan "doğum haftası" dal düğümde en iyi tahmini sağlayan değişken olmuştur. Bu düğümde; "doğum haftası

≤34.3” olan bebeklerin 247’si (%87.9) düşük doğum ağırlığı ve %28.7 verimlilik ile tahmin edildiği gözlenmiştir. Bu ilk düğümdeki bebeklerin sadece %12.1’inin normal ağırlıkta doğduğu tahmin edilmiştir. Takip eden diğer düğümlerde de benzer şekilde tahmin değerleri verilmiştir (Tablo 12).

Tablo 13. CART’a ait sınıflandırma ve risk oranları

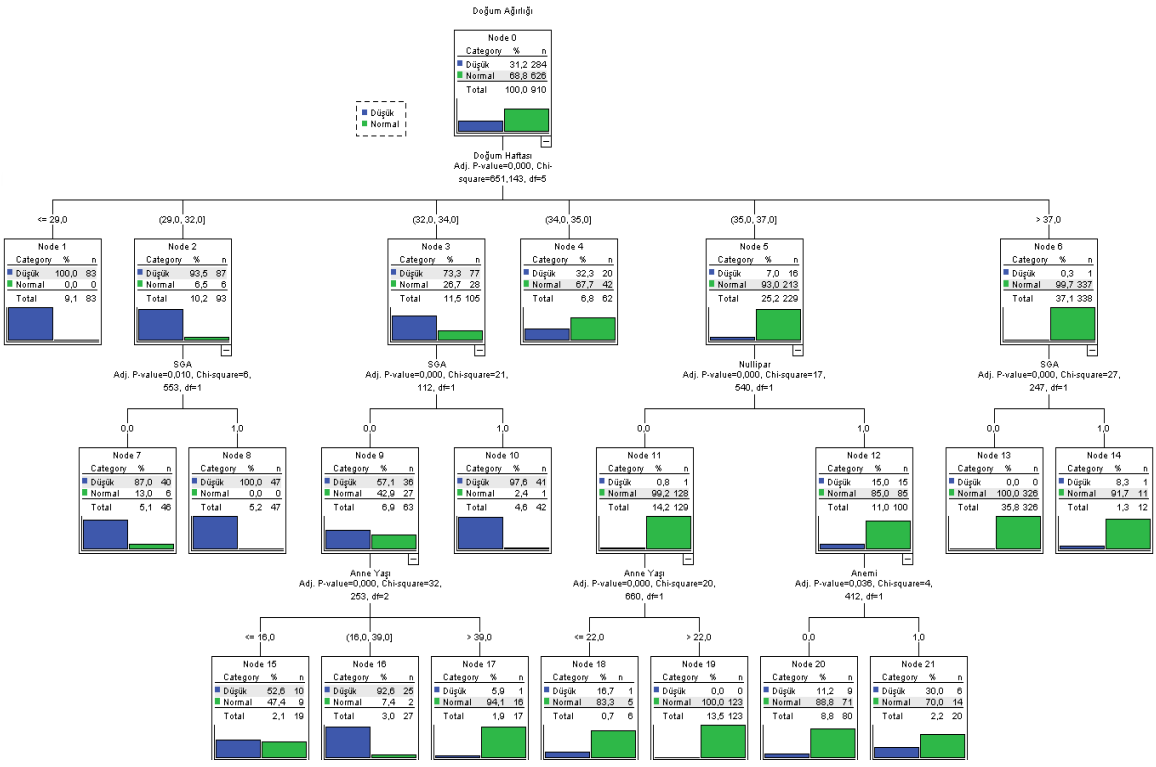
	Tahmin Değeri		
Gerçek Değer	**Düşük**	**Normal**	**Performans Ölçüleri (%)**
Düşük	251	33	Duyarlık (Sensitivity): 88.4
Normal	18	608	Özgüllük (Specificity): 97.1
Genel (%)	29.6	70.4	Gn. Doğr. Oranı (Accuracy): 94.4
Risk	**Tahmin**	**Std. Hata**	**MAPE**
Çapraz Geçerlik	0.068	0.008	0.116

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

CART algoritmasına ait sınıflandırma sonucu ve risk oranları Tablo 13’te verilmiştir. Bu algoritmada, en iyi performansı gösteren ölçütler kullanılarak tahmin değerleri elde edilmiştir. Buna göre sonuçlar incelendiğinde; 251 bebeğin “düşük ağırlıkta”, 608 bebeğin ise “normal ağırlıkta” olduğu doğru tahmin edilmiştir. Hesaplanan algoritmada duyarlık %88.4, özgüllük %97.1 ve genel doğruluk oranı %94.4 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.8 olduğu gözlenmiştir.

4.1.2. CHAID algoritmasına ait analiz sonuçları

CHAID algoritmasının uygulamasında işlem seçenekleri olarak, cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği (otomatik seçim) "3"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "5" alınmıştır. Buna göre CHAID algoritmasına ait karar ağacı diyagramı ve analiz sonuçları Şekil 12 ve Tablo 14'te verilmiştir.



(*SGA: Bebeğin anne karnında gelişimi)*

Şekil 12. CHAID'e ait karar ağacı diyagramı

CHAID algoritmasına ait karar ağacı diyagramı (Şekil 12) ve tahmin sonuçları incelendiğinde (Tablo 14), cevap değişkeni (bebek doğum ağırlığı) üzerine etkili olan açıklayıcı değişkenler, "maksimum bölme kriterine (Ki-kare ve Bonferroni p-değerine)" göre verilmiştir.

Tablo 14. CHAID'e ait karar ağacı tahmin sonuçları

Düğüm	Düşük		Normal		Genel		Tahmin	Dal	Açıklayıcı Değişken			
	n	%	n	%	n	%	Kategori	Düğüm	Değişken	p.[a]	Ki-kare	Kesim Değeri
0	284	31.2	626	68.8	910	100.0	Normal					
1	83	100.0	0	0.0	83	9.1	Düşük	0	Doğ. Haftası	.001	651.14	≤29.0
2	87	93.5	6	6.5	93	10.2	Düşük	0	Doğ. Haftası	.001	651.14	(29.0, 32.0]
3	77	73.3	28	26.7	105	11.5	Düşük	0	Doğ. Haftası	.001	651.14	(32.0, 34.0]
4	20	32.3	42	67.7	62	6.8	Normal	0	Doğ. Haftası	.001	651.14	(34.0, 35.0]
5	16	7.0	213	93.0	229	25.2	Normal	0	Doğ. Haftası	.001	651.14	(35.0, 37.0]
6	1	0.3	337	99.7	338	37.1	Normal	0	Doğ. Haftası	.001	651.14	>37.0
7	40	87.0	6	13.0	46	5.1	Düşük	2	SGA	.010	6.55	Yok
8	47	100.0	0	0.0	47	5.2	Düşük	2	SGA	.010	6.55	Var
9	36	57.1	27	42.9	63	6.9	Düşük	3	SGA	.001	21.11	Yok
10	41	97.6	1	2.4	42	4.6	Düşük	3	SGA	.001	21.11	Var
11	1	0.8	128	99.2	129	14.2	Normal	5	Nullipar	.001	17.54	Yok
12	15	15.0	85	85.0	100	11.0	Normal	5	Nullipar	.001	17.54	Var
13	0	0.0	326	100.0	326	35.8	Normal	6	SGA	.001	27.25	Yok
14	1	8.3	11	91.7	12	1.3	Normal	6	SGA	.001	27.25	Var
15	10	52.6	9	47.4	19	2.1	Düşük	9	Anne Yaşı	.001	32.25	≤16.0
16	25	92.6	2	7.4	27	3.0	Düşük	9	Anne Yaşı	.001	32.25	(16.0, 39.0]
17	1	5.9	16	94.1	17	1.9	Normal	9	Anne Yaşı	.001	32.25	>39.0
18	1	16.7	5	83.3	6	0.7	Normal	11	Anne Yaşı	.001	20.66	≤22.0
19	0	0.0	123	100.0	123	13.5	Normal	11	Anne Yaşı	.001	20.66	>22.0
20	9	11.3	71	88.8	80	8.8	Normal	12	Anemi	.036	4.41	Yok
21	6	30.0	14	70.0	20	2.2	Normal	12	Anemi	.036	4.41	Var

a. Bonferroni düzeltmeli en küçük p-değeri

Tablo 14'te. "bebek doğum ağırlığı" üzerine en fazla etkili olan açıklayıcı değişkenler ve bunların "kesim değerleri" verilmiştir. Bunun yanında, her bir açıklayıcı değişken için en iyi bölünmeyi sağlayan Bonferroni düzeltmeli en küçük p-değeri ve Ki-kare istatistiği verilmiştir. Ayrıca etkili açıklayıcı değişkelerin bulunduğu "dal düğüm numaraları" ile bu düğümlerde "bebek doğum ağırlığı (normal, düşük, genel)" bakımından "tahmin değerleri (sayı ve yüzde)" gösterilmiştir. Buna göre, bebeklerin doğum ağırlığına en fazla etki eden değişkenler sırasıyla; "Doğum haftası", "SGA", "Nullipar", "Anne yaşı" ve "Anemi" olduğu tespit edilmiştir.

En iyi tahmin değerini veren birinci dal düğümde, "doğum haftası ≤29.0" olan bebeklerin 83'ünün (%100) düşük doğum ağırlığına sahip olduğu ve anlamlı ($p<0.05$) düzeyde tahmin edildiği gözlenmiştir. Takip eden diğer düğümlerde de benzer şekilde tahmin değerleri verilmiştir (Tablo 14).

Tablo 15. CHAID'e ait sınıflandırma ve risk oranları

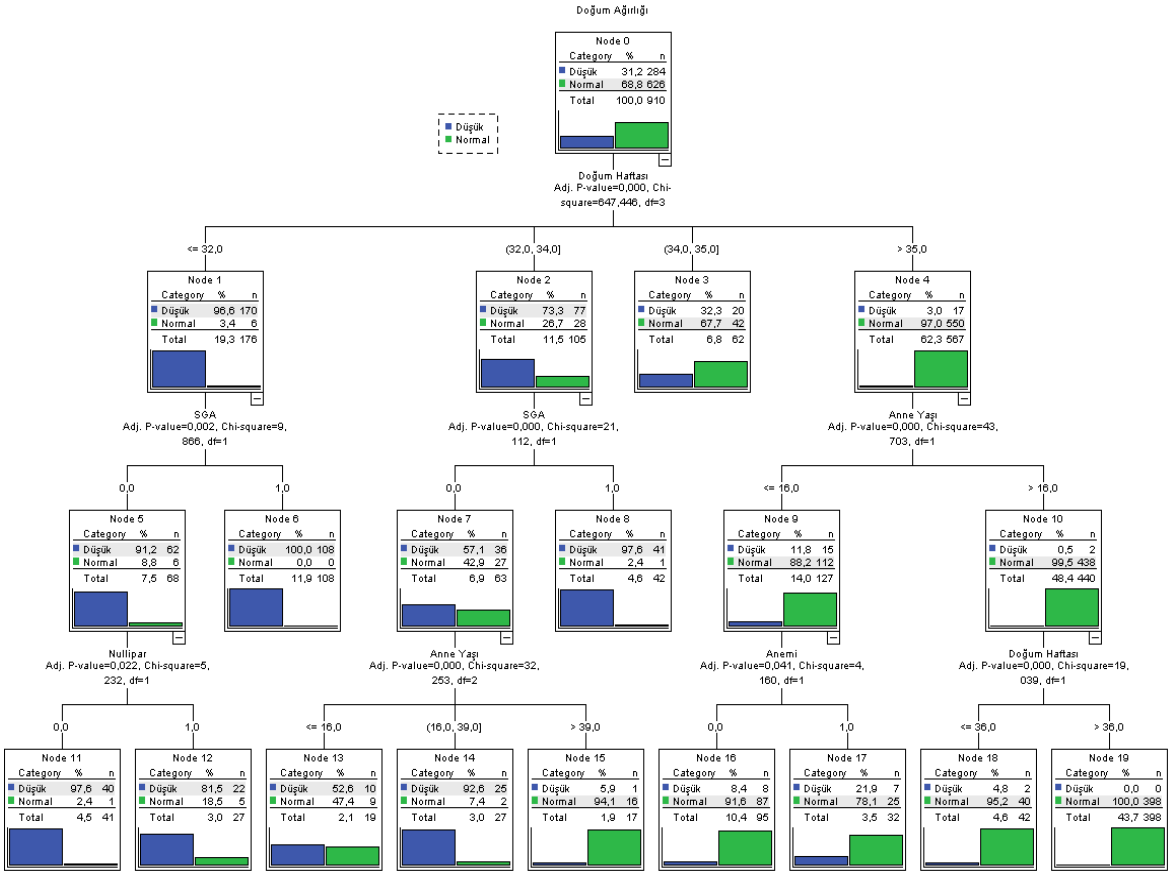
	Tahmin Değeri		
Gerçek Değer	**Düşük**	**Normal**	**Performans Ölçüleri (%)**
Düşük	246	38	Duyarlık (Sensitivity): 86.4
Normal	18	608	Özgüllük (Specificity): 97.1
Genel (%)	29.0	71.0	Gn. Doğr. Oranı (Accuracy): 93.6
Risk	**Tahmin**	**Std. Hata**	**MAPE**
Çapraz Geçerlik	0.066	0.008	0.136

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

CHAID algoritmasına ait sınıflandırma sonucu ve risk oranları Tablo 15'te verilmiştir. Bu algoritmada, en iyi performansı gösteren ölçütler kullanılarak tahmin değerleri elde edilmiştir. Buna göre sonuçlar incelendiğinde; 246 bebeğin "düşük ağırlıkta", 608 bebeğin ise "normal ağırlıkta" doğdukları doğru tahmin edilmiştir. Hesaplanan algoritmada duyarlık %86.4, özgüllük %97.1 ve genel doğruluk oranı %93.6 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.6 olduğu gözlenmiştir.

4.1.3. Ayrıntılı (Exhaustive) CHAID algoritmasına ait analiz sonuçları

Ayrıntılı (Exhaustive) CHAID algoritmasının uygulamasında işlem seçenekleri olarak, cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği (otomatik seçim) "3"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "5" alınmıştır. Buna göre Ayrıntılı CHAID algoritmasına ait karar ağacı diyagramı ve analiz sonuçları Şekil 13 ve Tablo 16'da verilmiştir.



(*SGA: Bebeğin anne karnında gelişimi)*

Şekil 13. Ayrıntılı CHAID'e ait karar ağacı diyagramı

Ayrıntılı CHAID algoritmasına ait karar ağacı diyagramı (Şekil 13) ve tahmin sonuçları (Tablo 16) incelendiğinde; cevap değişkeni (bebek doğum ağırlığı) üzerine en fazla etkili olan açıklayıcı değişkenler, "maksimum bölme kriterine (Ki-kare ve Bonferroni p-değerine)" göre verilmiştir.

Tablo 16. Ayrıntılı CHAID'e ait karar ağacı tahmin sonuçları

Düğüm	Düşük		Normal		Genel		Tahmin	Dal	Açıklayıcı Değişken			
	n	%	n	%	n	%	Kategori	Düğüm	Değişken	p.[a]	Ki-kare	Kesim Değeri
0	284	31.2	626	68.8	910	100.0	Normal					
1	170	96.6	6	3.4	176	19.3	Düşük	0	Doğ. Haftası	.001	647.446	≤32.0
2	77	73.3	28	26.7	105	11.5	Düşük	0	Doğ. Haftası	.001	647.446	(32.0, 34.0]
3	20	32.3	42	67.7	62	6.8	Normal	0	Doğ. Haftası	.001	647.446	(34.0, 35.0]
4	17	3.0	550	97.0	567	62.3	Normal	0	Doğ. Haftası	.001	647.446	>35.0
5	62	91.2	6	8.8	68	7.5	Düşük	1	SGA	.002	9.866	Yok
6	108	100.0	0	0.0	108	11.9	Düşük	1	SGA	.002	9.866	Var
7	36	57.1	27	42.9	63	6.9	Düşük	2	SGA	.001	21.112	Yok
8	41	97.6	1	2.4	42	4.6	Düşük	2	SGA	.001	21.112	Var
9	15	11.8	112	88.2	127	14.0	Normal	4	Anne Yaşı	.001	43.703	≤16.0
10	2	0.5	438	99.5	440	48.4	Normal	4	Anne Yaşı	.001	43.703	>16.0
11	40	97.6	1	2.4	41	4.5	Düşük	5	Nullipar	.022	5.232	Yok
12	22	81.5	5	18.5	27	3.0	Düşük	5	Nullipar	.022	5.232	Var
13	10	52.6	9	47.4	19	2.1	Düşük	7	Anne Yaşı	.001	32.253	≤16.0
14	25	92.6	2	7.4	27	3.0	Düşük	7	Anne Yaşı	.001	32.253	(16.0, 39.0]
15	1	5.9	16	94.1	17	1.9	Normal	7	Anne Yaşı	.001	32.253	>39.0
16	8	8.4	87	91.6	95	10.4	Normal	9	Anemi	.041	4.160	Yok
17	7	21.9	25	78.1	32	3.5	Normal	9	Anemi	.041	4.160	Var
18	2	4.8	40	95.2	42	4.6	Normal	10	Doğ. Haftası	.001	19.039	≤36.0
19	0	0.0	398	100.0	398	43.7	Normal	10	Doğ. Haftası	.001	19.039	>36.0

a. Bonferroni düzeltmeli en küçük p-değeri

Tablo 16'da "bebek doğum ağırlığına" en fazla etki eden açıklayıcı değişkenler ve bunların "kesim değerleri" verilmiştir. Bunun yanında, her bir açıklayıcı değişken için en iyi bölünmeyi sağlayan Bonferroni düzeltmeli en küçük p-değeri ve Ki-kare istatistiği verilmiştir. Ayrıca etkili açıklayıcı değişkelerin bulunduğu "dal düğüm numaraları" ile bu düğümlerde "bebek doğum ağırlığı (normal, düşük, genel)" bakımından "tahmin değerleri (sayı ve yüzde)" gösterilmiştir. Buna göre, bebeklerin doğum ağırlığına en fazla etki eden değişkenlerin sırasıyla; "Doğum haftası", "SGA", "Anne yaşı", "Nullipar" ve "Anemi" olduğu tespit edilmiştir.

En iyi tahmin değerini veren birinci dal düğümde, "doğum haftası ≤32.00" olan bebeklerin 170'inin (%96.6) düşük doğum ağırlığına sahip olduğu ve anlamlı ($p<0.05$) düzeyde tahmin edildiği gözlenmiştir. Bu ilk düğümdeki bebeklerin sadece %3.4'ünün normal ağırlıkta doğduğu tahmin edilmiştir. Takip eden diğer düğümlerde de benzer tahmin değerleri elde edilmiştir (Tablo 16).

Tablo 17. Ayrıntılı CHAID'e ait sınıflandırma ve risk oranları

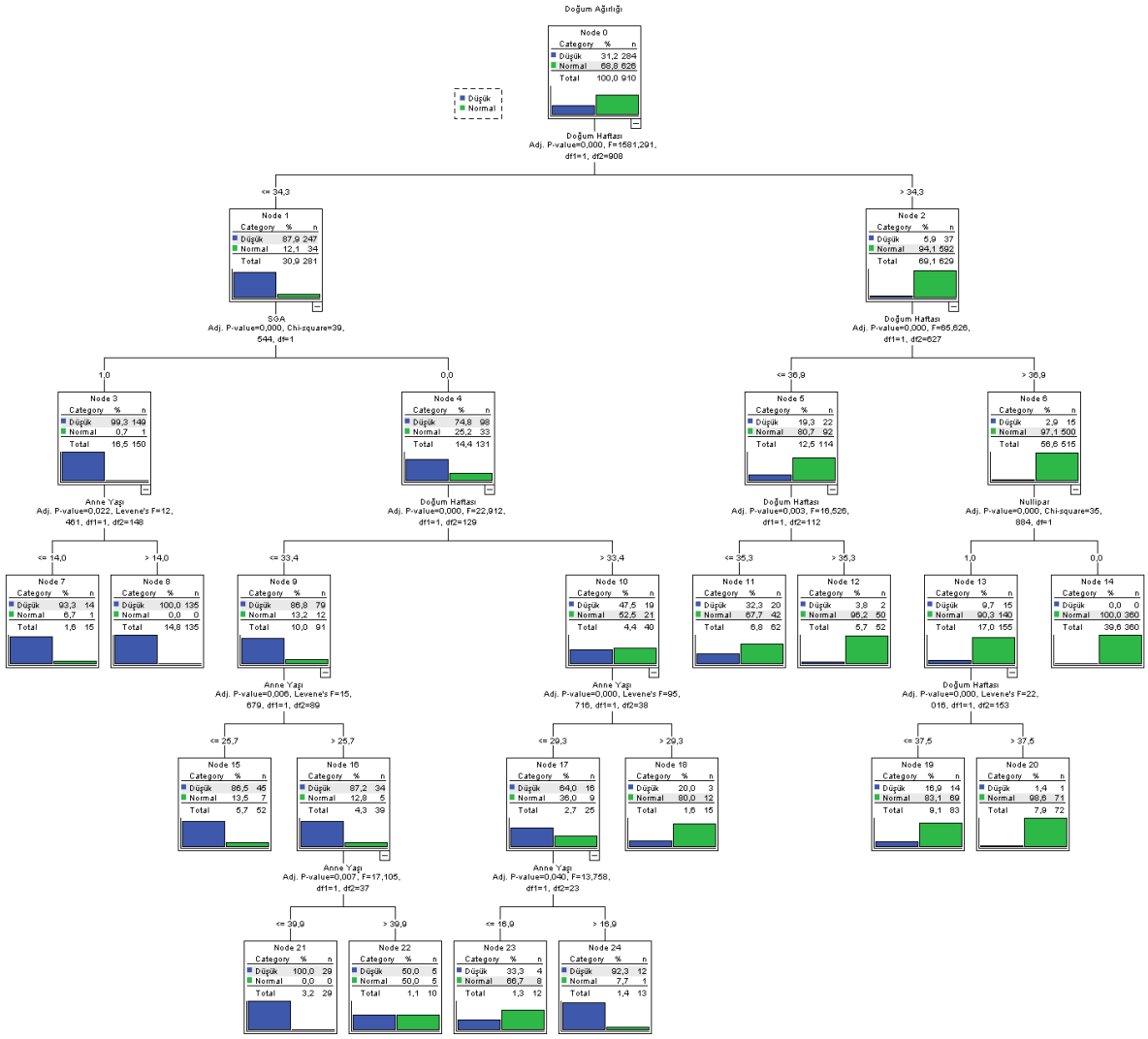
	Tahmin Değeri		
Gerçek Değer	**Düşük**	**Düşük**	**Performans Ölçüleri (%)**
Düşük	246	38	Duyarlık (Sensitivity): 86.6
Normal	18	608	Özgüllük (Specificity): 97.0
Genel (%)	29.0	71.0	Gn. Doğr. Oranı (Accuracy): 93.7
Risk	**Tahmin**	**Std. Hata**	**MAPE**
Çapraz Geçerlik	0.067	0.007	0.134

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

Ayrıntılı CHAID algoritmasına ait sınıflandırma sonucu ve risk oranları Tablo 17'de verilmiştir. Bu algoritmada, en iyi performansı gösteren ölçütler kullanılarak tahmin değerleri elde edilmiştir. Buna göre sonuçlar incelendiğinde; 246 bebeğin "düşük ağırlıkta", 608 bebeğin ise "normal ağırlıkta" doğduğu doğru tahmin edilmiştir. Hesaplanan algoritmada duyarlık %86.6, özgüllük %97.0 ve genel doğruluk oranı %93.7 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.7 olduğu gözlenmiştir.

4.1.4. QUEST algoritmasına ait analiz sonuçları

QUEST algoritmasının uygulamasında işlem seçenekleri olarak, cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği (otomatik seçim) "5"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "5" alınmıştır. Buna göre QUEST algoritmasına ait karar ağacı diyagramı ve analiz sonuçları Şekil 14 ve Tablo 18'de verilmiştir.



(*SGA: Bebeğin anne karnında gelişimi. GDM: Gestasyonel diyabet)*

Şekil 14. QUEST'e ait karar ağacı diyagramı

Ayrıntılı CHAID algoritmasına ait karar ağacı diyagramı (Şekil 14) ve tahmin sonuçları incelendiğinde (Tablo 18), cevap değişkeni (bebek doğum ağırlığı) üzerine

etkili olan açıklayıcı değişkenler "maksimum bölme kriterleri olan Ki-kare (kategorik veri için) veya F-istatistiğine (sürekli veri için)" göre verilmiştir.

Tablo 18. QUEST'e ait karar ağacı tahmin sonuçları

Düğüm	Düşük		Normal		Genel		Tahmin Kategori	Dal Düğüm	Açıklayıcı Değişken				
	n	%	n	%	n	%			Değişken	p.	Ki-kare	F	Kesim Değeri
0	284	31.2	626	68.8	910	100.0	Normal						
1	247	87.9	34	12.1	281	30.9	Düşük	0	D.Haftası	.001		1581.29	≤34.3
2	37	5.9	592	94.1	629	69.1	Normal	0	D.Haftası	.001		1581.29	>34.3
3	149	99.3	1	0.7	150	16.5	Düşük	1	SGA	.001	39.544		Yok
4	98	74.8	33	25.2	131	14.4	Düşük	1	SGA	.001	39.544		Var
5	22	19.3	92	80.7	114	12.5	Normal	2	D.Haftası	.001		65.63	≤36.9
6	15	2.9	500	97.1	515	56.6	Normal	2	D.Haftası	.001		65.63	>36.9
7	14	93.3	1	6.7	15	1.6	Düşük	3	An. Yaşı	.022		12.46	≤14.0
8	135	100.0	0	0.0	135	14.8	Düşük	3	An. Yaşı	.022		12.46	>14.0
9	79	86.8	12	13.2	91	10.0	Düşük	4	D.Haftası	.001		22.91	≤33.4
10	19	47.5	21	52.5	40	4.4	Normal	4	D.Haftası	.001		22.91	>33.4
11	20	32.3	42	67.7	62	6.8	Normal	5	D.Haftası	.003		16.53	≤35.3
12	2	3.8	50	96.2	52	5.7	Normal	5	D.Haftası	.003		16.53	>35.3
13	15	9.7	140	90.3	155	17.0	Normal	6	Nullipar	.001	35.884		Yok
14	0	0.0	360	100.0	360	39.6	Normal	6	Nullipar	.001	35.884		Var
15	45	86.5	7	13.5	52	5.7	Düşük	9	An. Yaşı	.006		15.68	≤25.7
16	34	87.2	5	12.8	39	4.3	Düşük	9	An. Yaşı	.006		15.68	>25.7
17	16	64.0	9	36.0	25	2.7	Düşük	10	An. Yaşı	.001		95.71	≤29.3
18	3	20.0	12	80.0	15	1.6	Normal	10	An. Yaşı	.001		95.71	>29.3
19	14	16.9	69	83.1	83	9.1	Normal	13	D.Haftası	.001		22.01	≤37.5
20	1	1.4	71	98.6	72	7.9	Normal	13	D.Haftası	.001		22.01	>37.5
21	29	100.0	0	0.0	29	3.2	Düşük	16	An. Yaşı	.007		17.10	≤39.9
22	5	50.0	5	50.0	10	1.1	Düşük	16	An. Yaşı	.007		17.10	>39.9
23	4	33.3	8	66.7	12	1.3	Normal	17	An. Yaşı	.040		13.75	≤16.9
24	12	92.3	1	7.7	13	1.4	Düşük	17	An. Yaşı	.040		13.75	>16.9

Tablo 18'de, "bebek doğum ağırlığına" en fazla etkili olan açıklayıcı değişkenler ve bunların "kesim değerleri" verilmiştir. Bunun yanında etkili açıklayıcı değişkelerin bulunduğu "dal düğüm numaraları" ile bu düğümlerde "bebek doğum ağırlığı (normal, düşük, genel)" bakımından "tahmin değerleri (sayı ve yüzde)" gösterilmiştir. Buna göre, bebeklerin doğum ağırlığına en fazla etki eden değişkenlerin sırasıyla; "Doğum haftası", "SGA", "Anne yaşı" ve "Nullipar" olduğu tespit edilmiştir.

En iyi tahmin değerini veren birinci dal düğümde, "doğum haftası ≤34.3" olan bebeklerin 247'sinin (%87.9) düşük doğum ağırlığına sahip olduğu ve anlamlı ($p<0.05$) düzeyde tahmin edildiği gözlenmiştir. Bu ilk düğümdeki bebeklerin yalnızca %12.1'inin normal ağırlıkta olduğu tahmin edilmiştir. Takip eden diğer düğümlerde de benzer tahmin değerleri elde edilmiştir (Tablo 19).

Tablo 19. QUEST'e ait sınıflandırma ve risk oranları

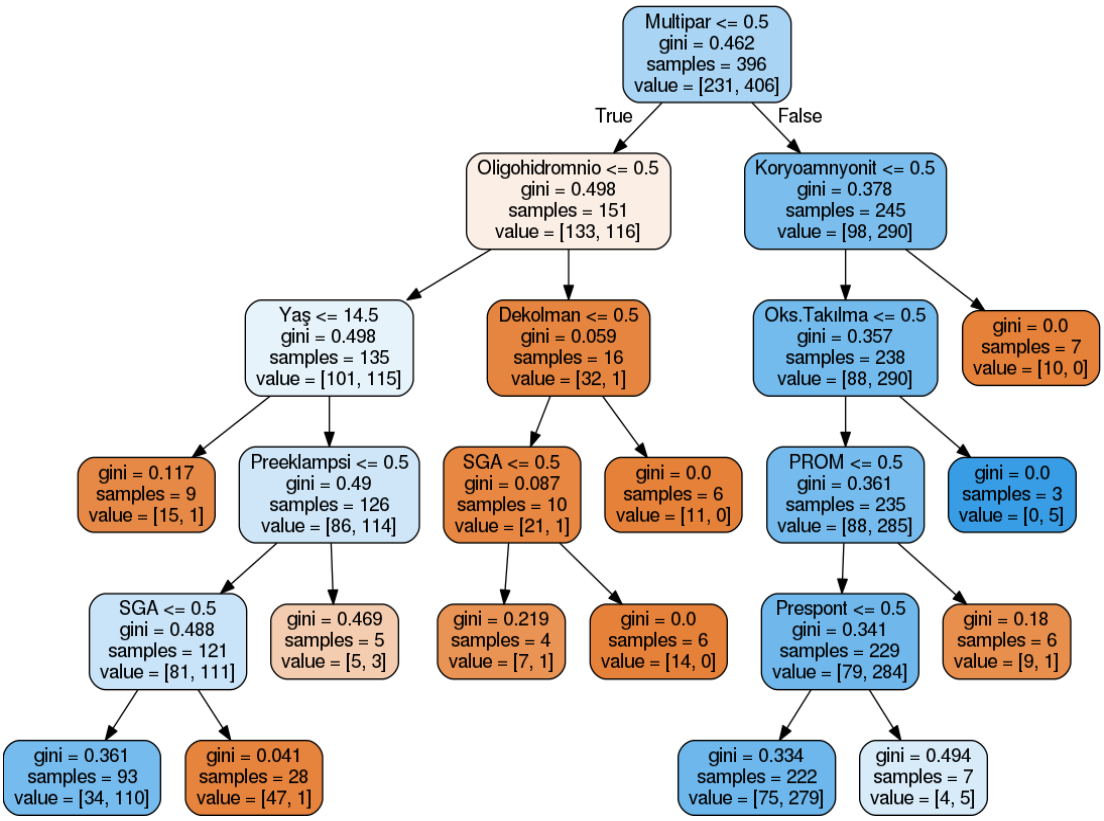
	Tahmin Değeri		
Gerçek Değer	**Düşük**	**Normal**	**Performans Ölçüleri (%)**
Düşük	240	44	Duyarlık (Sensitivity): 84.5
Normal	14	612	Özgüllük (Specificity): 97.8
Genel (%)	27.9	72.1	Gn. Doğr. Oranı (Accuracy): 93.6
Risk	**Tahmin**	**Std. Hata**	**MAPE**
Çapraz Geçerlik	0.067	0.008	0.155

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

QUEST algoritmasına ait sınıflandırma sonucu ve risk oranları Tablo 19'da verilmiştir. Bu algoritmada, en iyi performansı gösteren ölçütler kullanılarak tahmin değerleri elde edilmiştir. Bu sonuçlara göre; 240 bebeğin "düşük ağırlıkta", 612 bebeğin ise "normal ağırlıkta" doğduğu doğru tahmin edilmiştir. Hesaplanan algoritmada duyarlık %84.5, özgüllük %97.8 ve genel doğruluk oranı %93.6 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.7 olduğu gözlenmiştir.

4.1.5. Rastgele Orman (Random Forest) algoritmasına ait analiz sonuçları

Rastgele Orman algoritmasının uygulamasında işlem seçenekleri olarak, cevap değişkeni "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği "5"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "10" alınmıştır (Tablo 20). Rastgele Orman algoritmasına ait karar ağacı diyagramında (Şekil 15), cevap değişkeni (bebek doğum ağırlığı) üzerine etkili olan açıklayıcı değişkenler ve bölme kriterleri (entropiye dayalı bilgi kazanç oranlarından Gini değeri) verilmiştir.



Multiparite: Önceden doğum yapmış, Oligohidroamniyoz: Anne karnındaki amniyotik sıvının eksik olması, Preeklampsi: Yüksek tansiyon ve organ hasarı belirtileri, SGA: Bebeğin gelişim geriliği, Prom: Erken su gelme

Şekil 15. Rastgele Orman algoritmasına ait karar ağacı diyagramı

Şekil 15 incelendiğinde, "bebek doğum ağırlığı" üzerine en fazla etkili olan açıklayıcı değişkenlerin sırasıyla; "Multiparite", "Oligohidroamniyoz", "Anne yaşı", "Preeklampsi" ve "SGA" olduğu tespit edilmiştir. Tablo 20'de Rastgele Orman algoritmasının uygulamasında kullanılan işlem seçenekleri yer almaktadır.

Tablo 20. Rastgele Orman algoritmasına ait işlem seçenekleri

Seçenekler	Rastgele Orman
Çapraz Geçerlik	10-kat
Öğrenme veri seti	%70
Test veri seti	%30
Bölme kriteri	Gini
Maksimum ağaç derinliği	5
Min. Dal Düğüm Sayısı	25
Min. Yaprak Düğüm Sayısı	10
Cevap değişkeni	Bebek doğum ağırlığı (normal/düşük)

Tablo 20'deki işlem seçenekleri dikkate alınarak yapılan Rastgele Orman algoritmasına ait sınıflandırma ve risk oranları aşağıdaki gibidir.

Tablo 21. Rastgele Orman algoritmasına ait sınıflandırma ve risk oranları

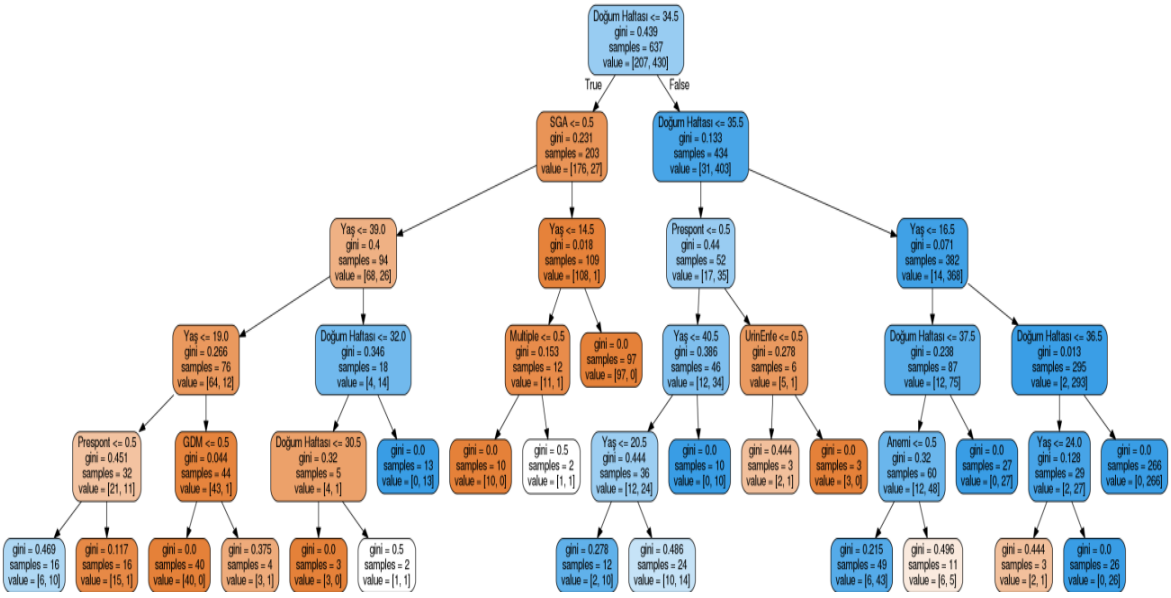
	Performans Ölçüleri (%)		
Duyarlık (Sensitivity)	87.1		
Özgüllük (Specificity)	98.2		
Gn. Doğr. Oranı (Accuracy)	93.8		
Risk	**Tahmin**	**Std. Hata**	**MAPE**
Hata kareler ortalaması	0.062	0.002	0.185

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

Rastgele Orman algoritmasına ait hesaplanan sonucunda, duyarlık %87.1, özgüllük %98.2 ve genel doğruluk oranı %93.8 olarak bulunmuştur. Hata kareler ortalaması risk/tahmin değerinin ise %6.2 olduğu gözlenmiştir (Tablo 21).

4.1.6. C4.5 karar ağacı algoritmasına ait analiz sonuçları

C4.5 karar ağacı algoritmasının uygulamasında işlem seçenekleri olarak, cevap değişken "bebek doğum ağırlığı"; çapraz geçerlik "10-kat"; maksimum ağaç derinliği "5"; minimum dal düğüm sayısı "25" ve minimum yaprak düğüm sayısı "10" alınmıştır. Uygulamada J48 adıyla da bilinen C4.5 karar ağacı algoritmasına ait karar ağacı diyagramında (Şekil 16), cevap değişkeni (bebek doğum ağırlığı) üzerine etkili olan açıklayıcı değişkenler ve bölme kriterleri (entropiye dayalı bilgi kazanç oranlarından Gini değeri) verilmiştir. Uygulamada işlem seçenekleri kullanılarak hesaplanan bu algoritmaya ait karar ağacı diyagramı aşağıdaki gibidir.



Prespont: kendiliğinden erken doğum, SGA: anne karnında gelişim geriliği, GDM: gestasyonel diyabet

Şekil 16. C4.5 algoritmasına ait karar ağacı diyagramı

Şekil 16 incelendiğinde, "bebek doğum ağırlığı" üzerine en fazla etkili olan açıklayıcı değişkenler sırasıyla; "Doğum haftası", "SGA", "Anne Yaşı", "Prespont" ve "GDM" olduğu tespit edilmiştir. Tablo 22'de C4.5 karar ağacı algoritmasının uygulamasında kullanılan işlem seçenekleri yer almaktadır.

Tablo 22. C4.5 karar ağacı algoritmasına ait işlem seçenekleri

Seçenekler	C4.5 Karar Ağacı
Çapraz Geçerlik	10-kat
Öğrenme veri seti	%70
Test veri seti	%30
Bölme kriteri	Gini
Maksimum ağaç derinliği	5
Min. Dal Düğüm Sayısı	25
Min. Yaprak Düğüm Sayısı	10
Cevap değişkeni	Bebek doğum ağırlığı (normal/düşük)

Tablo 22'deki işlem seçenekleri dikkate alınarak yapılan C4.5 karar ağacı algoritmasına ait sınıflandırma ve risk oranları aşağıdaki gibidir.

Tablo 23. C4.5 karar ağacı algoritmasına ait sınıflandırma ve risk oranları

	Performans Ölçüleri (%)		
Duyarlık (Sensitivity)	87.5		
Özgüllük (Specificity)	97.3		
Gn. Doğr. Oranı (Accuracy)	94.5		
Risk	**Tahmin**	**Std. Hata**	**MAPE**
Hata kareler ortalaması	.056	.001	0.109

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

C4.5 karar ağacı algoritmasına ait hesaplanan sonucunda, duyarlık %87.5, özgüllük %97.3 ve genel doğruluk oranı %94.5 olarak bulunmuştur. Hata kareler ortalaması risk/tahmin değerinin ise %5.6 olduğu gözlenmiştir (Tablo 23).

4.2. K-En Yakın Komşu Analizi Sonuçları

Uygulamanın ilk aşamasında, k-en yakın komşu (KNN) algoritmasına ait ön deneme (performans) sonuçları elde edilmiştir. Buna göre k-en yakın komşu algoritmasına ait performans sonuçları Tablo 24'teki gibidir.

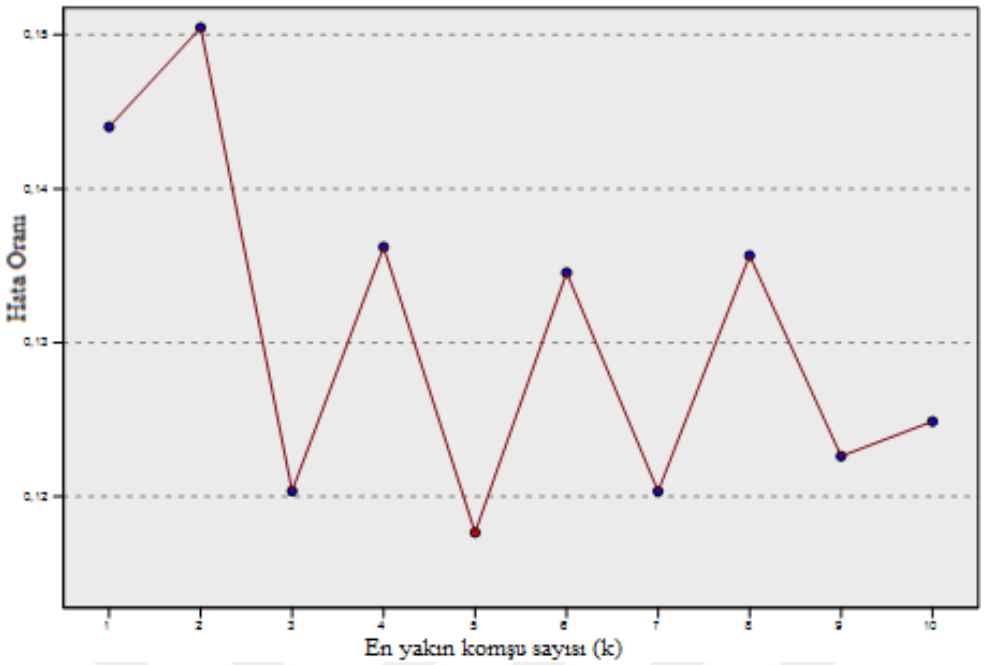
Tablo 24. K-en yakın komşu algoritmasına ait ön deneme (performans) sonuçları

K-kat Çapraz Geçerlik	Komşu Sayısı (KNN)	K seçiminde Yanlış Sınıflandırma (Hata/Risk %'si)	Bölümlere Atama Yüzdesi (%)	Duyarlık (Sensitivity) (%)	Özgüllük (Specificity) (%)	Gen. Doğr. Oranı (Accuracy) (%)
10	1	10.6	32.1	78.6	93.8	89.4
10	2	12.6	28.8	65.5	97.8	87.4
10	3	10.8	27.6	75.6	96.4	89.2
10	4	9.7	28.4	72.8	98.3	90.3
10	**5**	**7.9**	**29.1**	**78.4**	**97.4**	**92.2**
10	6	8.2	29.2	77.4	97.4	90.9
10	7	9.4	28.7	77.1	96.7	90.6
10	8	10.4	29.2	75.6	96.2	89.6
10	9	11.6	28.4	69.0	98.7	88.4
10	10	12.1	28.1	67.8	97.2	87.9

K seçiminde yanlış sınıflandırma/hata (K selection Error)

K-en yakın komşu algoritmasının performansını belirlemek üzere yapılan ön deneme, işlem seçenekleri olarak; "10-kat çapraz geçerlilik" ve "1 ile 10 arası k-en yakın komşu sayısı" kullanılarak alternatif sonuçlar elde edilmiştir. Buna göre, bu algoritmanın performansını gösteren ölçütlerden, "Test (Holdout)" veri setlerine ait; "*k* seçiminde yanlış sınıflandırma (hata/risk) oranı, duyarlık (sensitivity), özgüllük (specificity) ve genel doğruluk oranı (accuracy)" ölçütleri verilmiştir (Tablo 24).

"Test" veri setlerinin tahmin değerlerine ait performans sonuçları bakımından; "en düşük risk/hata oranı" ve "optimum duyarlık, özgüllük, genel doğruluk oranı" göz önüne alındığında, *k* komşu sayısının 5 olması en iyi sonucu vermektedir. Diğer alternatif denemelerin sonucunda elde edilen "hata katsayıları (oranları)" Şekil 17'de verilmiştir.



Şekil 17. K sayısına göre hata (risk) oranları (%)

Elde edilen bu sonuçlara göre; "Öklid" ve "Manhattan (City block)" uzaklık ölçüleri ile diğer işlem seçenekleri (Tablo 24) kullanılarak KNN algoritması uygulanmış ve aşağıdaki analiz sonuçları elde edilmiştir.

4.2.1. Öklid uzaklığı kullanıldığında k-en yakın komşu analizi sonuçları

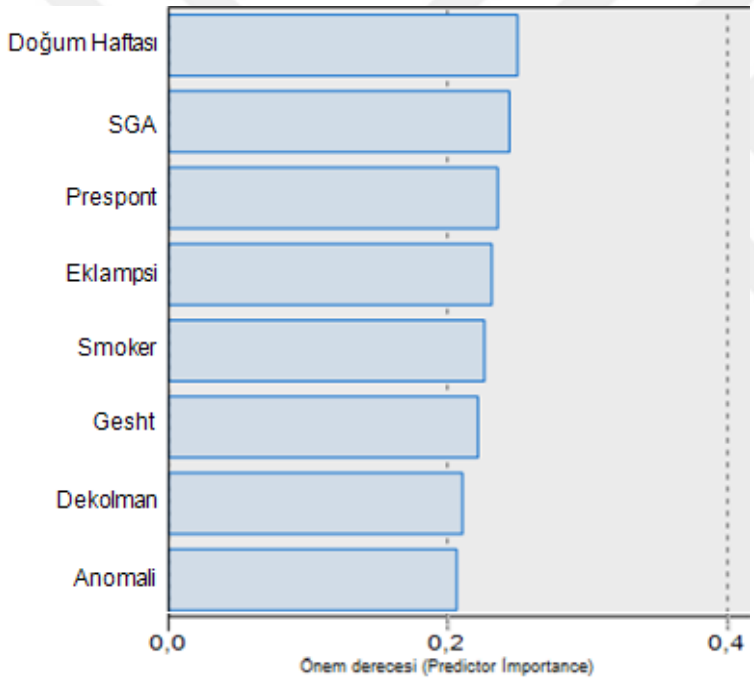
K-en yakın komşu algoritmasının bu uygulamasında, bebek doğum ağırlığı cevap değişkeni için işlem seçenekleri olarak; "10 kat çapraz geçerlik", "5-komşu sayısı", "öğrenme veri seti %70", "test veri seti %30" ve "Öklid uzaklığı" seçenekleri kullanılarak (Tablo 25) analiz yapılmıştır.

Tablo 25. Öklid uzaklığı kullanıldığında KNN'de işlem seçenekleri

Seçenekler	KNN
Çapraz Geçerlik	10-kat
Komşu sayısı (KNN)	*k*=5
Uzaklık ölçüsü	Öklid
Öğrenme veri seti	%70
Test veri seti	%30
Cevap değişkeni	Bebek doğum ağırlığı (normal/düşük)

Tablo 25'teki seçeneklere göre yapılan analiz sonucunda, KNN'de bölümlere atama yüzdeleri program tarafından; "öğrenme verisi %70.9 (n=645)" ve "test verisi %29.1 (n=265)" olarak belirlenmiştir.

Şekil 16'da açıklayıcı değişkenlerin "bebek doğum ağırlığı" üzerindeki önem sıralaması verilmiştir. Ön denemeyle belirlenen işlem seçenekleri kullanılarak elde edilen sonuçlara göre cevap değişkenini en fazla etkileyen ilk 10 açıklayıcı değişkenin sırasıyla; "Doğum haftası, SGA, Prespont, Eklampsi, Sigara, Gestasyonel hipertansiyon, Dekolman ve Anomali" olduğu gözlenmiştir (Şekil 18). Bu algoritmada, ön denemeyle belirlenen işlem seçenekleri kullanılarak bulunan sonuçlara göre en iyi performansı gösteren "öğrenme ve test" veri setlerine ait tahmin değerleri elde edilmiştir (Tablo 26).



Şekil 18. Öklid uzaklığı kullanıldığında açıklayıcı değişkenlerin önem sıralaması

K-en yakın komşu algoritmasına ait sınıflandırma sonucu ve yanlış sınıflandırma (hata/risk) oranları Tablo 26'da verilmiştir. Bu algoritmada, ön denemeyle belirlenen işlem seçenekleri kullanılarak bulunan sonuçlara göre, en iyi performansı gösteren tahmin değerleri elde edilmiştir.

Tablo 26. Öklid uzaklığı kullanıldığında KNN'ye ait sınıflandırma ve risk oranları

	Tahmin Değeri		
Gerçek Değer	**Düşük**	**Normal**	**Performans Ölçüleri (%)**
Düşük	58	16	Duyarlık (Sensitivity): 78.4
Normal	5	186	Özgüllük (Specificity): 97.4
Genel (%)	23.8	76.2	Gn. Doğr. Oranı (Accuracy): 92.1
Risk	**Yanlış Sınıflandırma**		**MAPE**
Çapraz Geçerlik	0.079		0.157

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

Tablo 26'daki sonuçlar incelendiğinde; veri setinde bulunan 265 bebekten 58'si "düşük ağırlıkta", 186'sı "normal ağırlıkta" olarak tahmin edilmiştir. Benzer şekilde; duyarlık %78.4, özgüllük %97.4 ve genel doğruluk oranı %92.1 olarak bulunmuştur. Algoritma hesaplaması sonucunda, yanlış sınıflandırma (hata/risk) oranları ise; %7.9 olarak gözlenmiştir.

4.2.2. Manhattan uzaklığı kullanıldığında k-en yakın komşu analizi sonuçları

K-en yakın komşu algoritmasının bu uygulamasında, bebek doğum ağırlığı cevap değişkeni için işlem seçenekleri olarak; "10 kat çapraz geçerlik", "5-komşu sayısı", "öğrenme veri seti %70", "test veri seti %30" ve "Manhattan uzaklığı" seçenekleri kullanılarak (Tablo 27) analiz yapılmıştır.

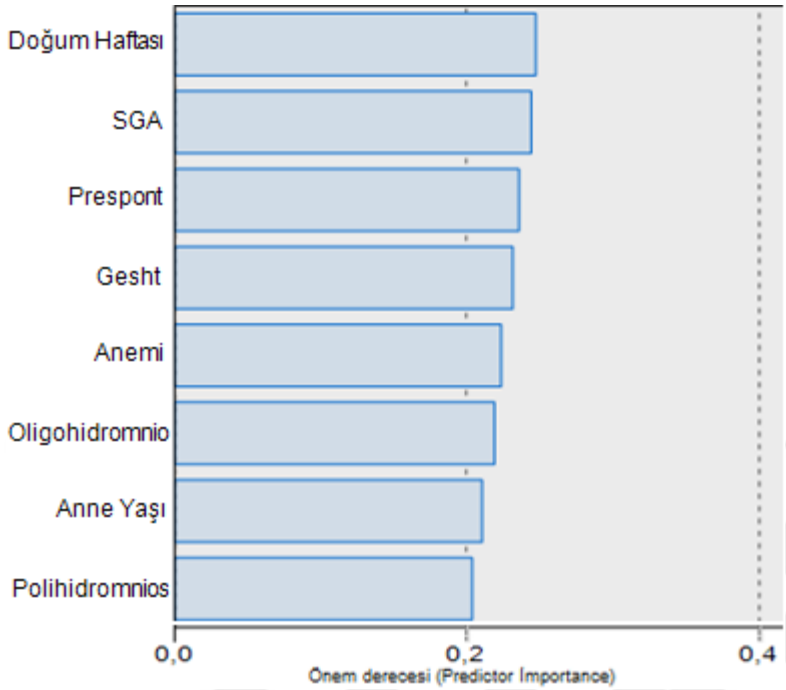
Tablo 27. Manhattan uzaklığı kullanıldığında KNN'de işlem seçenekleri

Seçenekler	**KNN**
Çapraz Geçerlik	10-kat
Komşu sayısı (KNN)	*k*=5
Uzaklık ölçüsü	Manhattan (City block)
Öğrenme veri seti	%70
Test veri seti	%30
Cevap değişkeni	Bebek doğum ağırlığı (normal/düşük)

Tablo 27'deki seçeneklere göre yapılan analiz sonucunda. KNN'de bölümlere atama yüzdeleri program tarafından; "öğrenme verisi %70.9 (n=645)" ve "test verisi %29.1 (n=265)" olarak belirlenmiştir.

Açıklayıcı değişkenlerin "bebek doğum ağırlığı" üzerindeki önem sıralaması Şekil 17'de verilmiştir. Ön deneme ile belirlenen işlem seçenekleri kullanılarak elde edilen sonuçlara göre cevap değişkenini en fazla etkileyen ilk 10 açıklayıcı değişkenin

sırasıyla; "Doğum haftası, SGA, Prespont, Gestasyonel hipertansiyon, Anemi, Oligohidromnio, Anne yaş, Polihidromnios" olduğu gözlenmiştir (Şekil 19).



Şekil 19. Manhattan uzaklığı kullanıldığında açıklayıcı değişkenlerin önem sıralaması

K-en yakın komşu algoritmasına ait sınıflandırma sonucu ve yanlış sınıflandırma (hata/risk) oranları Tablo 28'de verilmiştir. Bu algoritmada, ön denemeyle belirlenen işlem seçenekleri kullanılarak bulunan sonuçlara göre en iyi performansı gösteren tahmin değerleri elde edilmiştir.

Tablo 28. Manhattan uzaklığı kullanıldığında KNN'ye ait sınıflandırma ve risk oranları

	Tahmin Değeri		
Gerçek Değer	**Düşük**	**Normal**	**Performans Ölçüleri (%)**
Düşük	59	18	Duyarlık (Sensitivity): 76.6
Normal	5	188	Özgüllük (Specificity): 97.4
Genel (%)	23.7	76.3	Gn. Doğr. Oranı (Accuracy): 91.5
Risk	**Yanlış Sınıflandırma**		**MAPE**
Çapraz Geçerlik	0.08		0.162

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata)

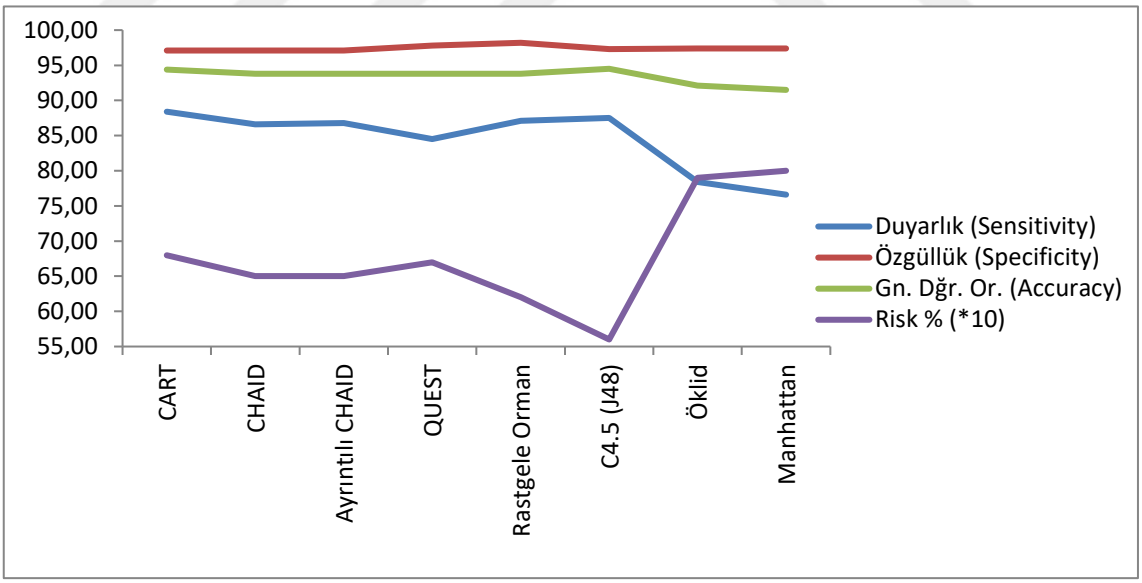
Tablo 28'deki sonuçlar incelendiğinde; veri setinde bulunan 270 bebekten 59'u "düşük ağırlıkta", 188'i ise "normal ağırlıkta" tahmin edilmiştir. Benzer şekilde; duyarlık %76.6, özgüllük %97.4 ve genel doğruluk oranı %91.5 olarak bulunmuştur. Algoritma hesaplaması sonucunda, yanlış sınıflandırma (hata/risk) oranları ise; %8 olarak gözlenmiştir.

4.3. Algoritmaların sınıflandırma performanslarının incelenmesi

Tablo 29. Kullanılan algoritmaların sınıflama performanslarının incelenmesi

		Karar Ağaçları						K-En Yakın Komşu	
Performans Ölçütü		**CART**	**CHAID**	**Ayrıntılı CHAID**	**QUEST**	**Rastgele Orman**	**C4.5**	**Öklid**	**Manhattan**
Duyarlık (Sensitivity %)		88.4	86.4	86.6	84.5	87.1	87.5	78.4	76.6
Özgüllük (Specificity %)		97.1	97.1	97.0	97.8	98.2	97.3	97.4	97.4
Gn. Doğ. Oranı (Accuracy %)		94.4	93.6	93.7	93.8	93.8	94.5	92.1	91.5
Risk tahmini (%)		6.8	6.6	6.7	6.7	6.2	5.6	7.9	8.0
MAPE		11.6	13.6	13.4	15.5	16.5	10.9	15.7	16.2
Etkili Değişken	1	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası	Multiparite	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası
	2	SGA	SGA	SGA	SGA	Oligo-hidroamniyoz	SGA	SGA	SGA
	3	Anne Yaşı	Nullipar	Anne Yaşı	Anne Yaşı	Anne Yaşı	Anne Yaşı	Prespont	Prespont
	4		Anne Yaşı	Nullipar	Nullipar	Preeklampsi	Prespont	Eklampsi	Gesth
	5		Anemi	Anemi		SGA	Gestasyonel Diyabet	Smoker	Oligo-hidroamniyoz

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata), SGA: bebeğin anne karnında gelişimi, Prespont: kendiliğinden erken doğum, Nullipar: hast. ilk doğ., Multiparite: önceden doğum yapmış, Oligohidroamniyoz: anne karnındaki amniyotik sıvının eksik olması, Preeklampsi: yüksek tansiyon ve organ hasarı



Şekil 20. Kullanılan algoritmaların sınıflandırma performanslarının karşılaştırılması

Çalışmada ele alınan algoritmaların sınıflandırma performansları Tablo 29'da ve Şekil 20'de gösterilmiştir. Buna göre, kullanılan algoritmaların sınıflandırma

performansları genel olarak birbirine yakın bulunmuştur. Ancak bu performans ölçülerine göre en iyi tahmin değerini veren yöntemler aşağıdaki gibi özetlenmiştir.

Duyarlık (Sensitivity): Bu tahmin ölçütü bakımından en yüksek tahmin oranı %88.4 ile CART algoritmasında gözlenmiştir.

Özgüllük (Specificity): Bu tahmin ölçütü bakımından en yüksek tahmin oranı %98.2 ile Rastgele Orman algoritmasında gözlenmiştir.

Genel Doğruluk (Accuracy): Bu tahmin ölçütü bakımından en yüksek tahmin oranının %94.5 ile "C4.5" algoritmasında olduğu tespit edilmiştir.

Risk (hata) tahmini: Bu tahmin ölçütü bakımından en düşük oran %5.6 ile "C4.5" algoritmasında gözlenmiştir.

MAPE: (Ortalama mutlak yüzde hata): Bu tahmin ölçütü bakımından en düşük oran %10.9 ile "C4.5" algoritmasında gözlenmiştir.

Modele giren ve cevap değişkeni (bebek doğum ağırlığı) üzerinde etkili olan açıklayıcı değişkenler incelendiğinde; Rastgele Orman algoritması haricinde (Multiparite), tüm algoritmalarda "doğum haftasının" en etkili değişken olduğu gözlenmiştir. Benzer şekilde, ikinci sırada en etkili açıklayıcı değişkenler algoritmalarda genel olarak "SGA" olurken, Rastgele Orman algoritmasında "Oligohidroamniyoz (anne karnındaki amniyotik sıvının eksik olması)" yer almıştır. Genel anlamda algoritmaların performansları bakımından az değişken ile yüksek performans gösteren algoritmaların tercih edildiği göz önüne alındığında; CART'ın 3 ve CHAID'in 4 adet açıklayıcı değişken ile model oluşturması, diğer değişkenlere (5 adet değişken) göre daha başarılı olduğu söylenebilir.

Weka analiz programında algoritması bulunan "Rastgele Orman, C4.5 ve K-En Yakın Komşu" yöntemleri için sonuçlar bu programda analiz edilmiş ve elde edilen sonuçlar Tablo 30'da verilmiştir.

Tablo 30. Weka ile algoritmaların sınıflandırma performanslarının incelenmesi

Performans Ölçütü		**Rastgele Orman**	**C4.5**	**K-En Yakın Komşu**
[1] Doğru sınıflandırma (%)		93.5	92.4	87.6
[2] Kappa istatistiği (%)		84.6	81.8	69.3
[3] Ortalama mutlak yüzde hata (%)		16.5	10.9	15.7
[4] Kök ortalama karesi hatası		25.5	25.0	30.1
[5] Göreceli mutlak hata (%)		43.1	25.4	36.6
[6] Kök bağıl kare hatası (%)		55.0	53.9	64.9
[7] Ortalama doğruluk % (GP)		93.5	92.4	87.6
Etkili Değişkenler	1	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası
	2	Prespont	SGA	SGA
	3	Oligohidroamniyoz	Anne Yaşı	Prespont
	4	Anemi	Prespont	Gesth
	5	Parite	Sigara	Oligohidroamniyoz

1: Correctly Classified Instances, 2: Kappa statistic, 3: Mean absolute error, 4: Root mean squared error, 5: Relative absolute error, 6: Root relative squared error, 7: Average accuracy (Gerçek Pozitif /True Positive)

Weka ile elde edilen sonuçlara göre, "doğru sınıflandırma oranı" bakımından en başarılı algoritmanın "Rastgele Orman" olduğu gözlenmiştir. Sınıflandırma hataları bakımından ise en iyi performans "C4.5" algoritmasında yer aldığı görülmektedir. Rasgele Orman algoritması yüksek doğruluk derecesi ile tahmin yapabilmekte ancak yanlış sınıflandırma hata oranları bakımından C4.5 daha başarılı olduğu gözlenmiştir. Bu durum, öğrenme veri setinde "aşırı öğrenme (overfitting)" olduğunun belirtisi niteliğindedir. Sonuç olarak, Python ve SPSS ile yapılan analizde olduğu gibi, Weka ile yapılan analiz sonucunda C4.5 algoritmasının daha iyi performans gösterdiği gözlenmiştir.

Bu çalışmada, "düşük doğum ağırlığındaki bebekler", tanı testi bakımından Pozitifliği göstermektedir. Elde edilen uygulama sonuçlarına göre "Duyarlığı yüksek" bulunan algoritmaların performansının, klinik bakımdan önemli olacağı bildirilmektedir (Neyzi, 2002; Singh ve ark., 2009). Buna göre; "CART" algoritmasının en yüksek Duyarlık (Sensitivity, %88.4) değeriyle klinik bakımdan en başarılı algoritma olduğu, uygulama sonuçları genel olarak incelendiğinde ise tüm algoritmaların "iyi sınıflandırma, yüksek tahmin ve düşük hata oranı" ile çalıştığı söylenebilir. Ancak, optimum performans bakımından, "C4.5" algoritmasının diğer algoritmalara göre bir miktar daha iyi sınıflandırma sağladığı ifade edilebilir.

5. TARTIŞMA VE SONUÇ

Çalışmada, veri madenciliği içerisinde sınıflandırma amacıyla kullanılan yöntemlerden olan; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman, C4.5 karar ağaçları" ve "K-En Yakın Komşu" yöntemi incelenerek bir veri seti üzerine uygulamalar yapılmıştır. Bu veri setinden elde edilen uygulama sonuçlarına göre söz konusu yöntemlerin performansları incelenmiştir (Tablo 29).

CART algoritması için belirlenen işlem seçenekleri kullanılarak elde edilen sonuçlara göre; bebeklerin doğum ağırlığına en fazla etki eden değişkenlerin sırasıyla; "Doğum haftası", "SGA" ve "Anne yaşı" olduğu tespit edilmiştir. Hesaplanan algoritmada duyarlık %88.4, özgüllük %97.1 ve genel doğruluk oranı %94.4 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.8 olduğu gözlenmiştir.

CHAID algoritması sonucu; bebeklerin doğum ağırlığına en fazla etki eden değişkenlerin sırasıyla; "Doğum haftası", "SGA", "Nullipar", "Anne yaşı" ve "Anemi" olduğu tespit edilmiştir (Tablo 29). Hesaplanan algoritmada duyarlık %86.4, özgüllük %97.1 ve genel doğruluk oranı %93.6 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.6 olduğu gözlenmiştir.

Ayrıntılı CHAID algoritması sonucu; bebeklerin doğum ağırlığı üzerine en fazla etkili olan değişkenlerin sırasıyla; "Doğum haftası", "SGA", "Anne yaşı" ve "Nullipar" olduğu tespit edilmiştir. Hesaplanan algoritmada duyarlık %86.6. özgüllük %97.0 ve genel doğruluk oranı %93.7 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.7 olduğu gözlenmiştir.

QUEST algoritması sonucu; bebeklerin doğum ağırlığı üzerine en fazla etkili olan değişkenlerin sırasıyla; "Doğum haftası", "SGA", "Anne yaş grubu", "Prespont" ve "GDM" olduğu tespit edilmiştir. Hesaplanan algoritmada duyarlık %84.5, özgüllük %97.8 ve genel doğruluk oranı %93.6 olarak bulunmuştur. Çapraz geçerlik risk/tahmin değerinin ise %6.7 olduğu gözlenmiştir.

Rastgele Orman algoritması sonucu; bebeklerin doğum ağırlığı üzerine en fazla etkili olan değişkenlerin sırasıyla; "Multiparite", "Oligohidroamniyoz", "Anne yaşı", "Preeklampsi" ve "SGA" olduğu tespit edilmiştir (Tablo 29). Hesaplanan algoritmada duyarlık %87.1, özgüllük %98.2 ve genel doğruluk oranı %93.8 olarak bulunmuştur. Hata kareler ortalaması risk/tahmin değerinin ise %6.2 olduğu gözlenmiştir.

C4.5 karar ağacı algoritması sonucu; bebeklerin doğum ağırlığı üzerine en fazla etkili olan değişkenlerin sırasıyla; "Doğum haftası", "SGA", "Anne Yaşı", "Prespont" ve "GDM" olduğu tespit edilmiştir. Hesaplanan algoritmada; duyarlık %87.5, özgüllük %97.3 ve genel doğruluk oranı %94.5 olarak bulunmuştur. Hata kareler ortalaması risk/tahmin değerinin ise %5.6 olduğu gözlenmiştir.

K-en yakın komşu algoritmasında, belirlenen işlem seçenekleriyle birlikte "Öklid" ve "Manhattan (City block)" uzaklık ölçüleri kullanılarak uygulama yapılmıştır. Buna göre, Öklid uzaklığı ile hesaplanan k-en yakın komşu algoritması sonucu, bebeklerin doğum ağırlığı üzerine en fazla etkili olan değişkenlerin sırasıyla; "Doğum haftası, SGA, Prespont, Eklampsi, Sigara, Gestasyonel hipertansiyon, Dekolman ve Anomali" olduğu tespit edilmiştir. Hesaplanan algoritmada; veri setinde bulunan 265 bebekten 58'si "düşük ağırlıkta", 186'sı "normal ağırlıkta" tahmin edilmiştir. Benzer şekilde; duyarlık %78.4, özgüllük %97.4 ve genel doğruluk oranı %92.1 olarak bulunmuştur. Algoritma hesaplaması sonucunda, yanlış sınıflandırma (hata/risk) oranı ise %7.9 olarak gözlenmiştir (Tablo 29).

Manhattan uzaklığı ile hesaplanan k-en yakın komşu algoritması sonucu, bebeklerin doğum ağırlığı üzerine en fazla etkili olan değişkenlerin sırasıyla; "Doğum haftası, SGA, Prespont, Gestasyonel hipertansiyon, Anemi, Oligohidromnio, Anne yaşı ve Polihidromnios" olduğu tespit edilmiştir (Tablo 29). Hesaplanan algoritmada; veri setinde bulunan 270 bebekten 59'u "düşük ağırlıkta", 188'i "normal ağırlıkta" tahmin edilmiştir. Benzer şekilde; duyarlık %76.6, özgüllük %97.4 ve genel doğruluk oranı %91.5 olarak bulunmuştur. Algoritma hesaplaması sonucunda, yanlış sınıflandırma (hata/risk) oranları ise %8 olarak gözlenmiştir.

Sınıflandırma amacıyla uygulanan bu algoritmalar ile cevap değişkeni (bebeklerin doğum ağırlığı) üzerine en fazla etkili olan açıklayıcı değişkenler

belirlenmiştir. Buna göre, her algoritmada bir miktar değişiklik olmakla birlikte, benzer değişkenler model girmiştir. Kullanılan algoritmalarda genel olarak ilk sırada "Doğum haftası" açıklayıcı değişkeni yer alırken, Rastgele Orman algoritmasında "Multiparite" değişkeni yer almıştır (Tablo 29).

Literatürde, birçok alanda bu algoritmalar ve bu algoritmaların sınıflandırmadaki tahmin başarıları ile ilgili çalışmalar bulunmaktadır. Kusiak ve ark. (2000) tarafından akciğerdeki bir tümörün çeşidini belirlemede karar destek amaçlı yapılan çalışmada, CART ve CHAID veri madenciliği yöntemlerinin, hastalığı doğru teşhis etmede %96-98 aralığında doğru tahminler gösterdiği belirtilmiştir. Türe ve ark. (2009) tarafından, meme kanserli hastaların sağkalımını etkileyen faktörlerin belirlenmesi amacıyla; karar ağacı algoritmalarından CART, QUEST, C4.5, CHAID, ID3 ile Kaplan-Meier sağkalım analizleri birlikte kullanılmıştır. Bozkır ve Sezer (2009) tarafından, Hacettepe Üniversitesi öğrencileri ve çalışanların beslenme tiplerinin incelenmesi amacıyla, Birliktelik Kuralları ile Ayrıntılı CHAID karar ağacı yöntemleri uygulanmış, %80 performansla gıda tüketim deseni açıklanabilmiştir.

Genel olarak, bu çalışmada kullanılan algoritmalara ait analiz sonuçları incelendiğinde; tüm algoritmaların "yüksek tahmin ve düşük hata oranı" ile çalıştığı söylenebilir. Ancak, "C4.5" algoritmasının diğer algoritmalara göre daha iyi performansla sınıflandırma sağladığı gözlenmiştir.

Karar verme yöntemlerinden olan veri madenciliğinin sağlık sektöründe kullanımı, sağlık hizmetlerinin daha etkin sunumu ve kaynakların daha verimli kullanılması açısından önemlidir (Koyuncugil ve Özgülbaş, 2009). Tıpta daha çok teşhise karar verme amacıyla kullanılan "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5 karar ağaçları" ve "K-En Yakın Komşu" yöntemleri ile "bebek doğum ağırlığını" etkileyen faktörler veri seti üzerinde incelenmiştir.

Literatürde, birçok alanda veri madenciliği yöntemleri uygulanmış çalışmalar bulunmakta ve kullanılan yöntemlerin sınıflandırma başarıları karşılaştırılmaktadır. Bu yöntemlerin karşılaştırılması akademik çalışmalara ve güncel uygulamalara fayda sağlamaktadır (Coşkun ve Baykal. 2011). Bu amaçla çalışmada, "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman, C4.5 karar ağaçları" ve "K-En Yakın

Komşu" yöntemleri ile ilgili tanıtıcı bilgiler verilmiş ve "bebek doğum ağırlığı" üzerine etkili olabilecek değişkenleri belirlemek amacıyla uygulamalar yapılmıştır. Bu yöntemler ile elde edilen sonuçlar yorumlanmıştır. Kullanılan yöntemlere ilişkin özet bilgiler sunularak, bebek doğum ağılığı üzerine etkili olabilecek değişkenlerin tespitine (en doğru tahmin ile sınıflandırılmasına) yönelik analizler yapılmıştır.

Dünyada bebeklerin %16'sı düşük doğum ağırlığı (≤2500g) ile doğmaktadır. Türkiye'de ise bu oran ortalama %10-12 arasındandır. Giderek artan bu oranlar toplum sağlığı açısından olumsuz sonuçlara yol açmakta ve bebeklerde gelişimsel geriliğe yol açan biyolojik etkenlerden biri olarak görülmektedir (Sağlık Bakanlığı, 2017). Bebek doğum ağırlığı birçok faktörden etkilenmektedir. Dolayısıyla doğum ağırlığını olumlu yada olumsuz yönde etkileyen faktörlerin ve bu faktörlerin sınıflandırmadaki başarılarının belirlenmesi toplum sağlığı açısından önemlidir. Dolayısıyla bu çalışma, yeni doğacak bebeklerin doğum ağırlığının, düşük doğum ağırlığında olup olmayacağına erken karar verme ve böylece koruyucu tedbirlerin alınması açısından araştırmacılara katkı sağlayabilir.

KAYNAKLAR

Acar E ve Özerdem MS. An iris recognition system by laws texture energy measure based KNN classifier. In: 2013 21st Signal Processing and Communications Applications Conference (SIU). IEEE, 2013:1-4.

Albayrak AS ve Yılmaz ŞK. Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama. SDÜ, İİBF Derg, 2009;14(1):31-52.

Bailey T ve Jain AK. A Note on distance-weighted k-nearest neighbor rules. IEEE Trans Syst Man Cybern Part B Cybern. SMC-8. 1978;4:311-3.

Bayır A, Özdemir Ş, Gülseçen S. Türkiye'deki seçmen eğilimlerinin C4.5 karar ağacı algoritması ile belirlenmesi. J Manage Inf Syst: JMIS, 2016;1(3):223-33.

Baykal N. Veri Tabanı ve Veri Madenciliği. Tıp Bilişimi Güz Okulu Ders Notları. 2003:102-15.

Bermejo S ve Cabestany J. Adaptive soft k-nearest-neighbour classifiers. J Pattern Recognit. 2000;33(12):199-205.

Berson A ve Smith S. Thearling K. Buildind Data Mining Applications for CRM. Mcgraw-Hill. USA. 2000;510-39.

Bhatt C. Mining the Medical Literature [Internet]. 2004 [ET: 08.01.2018]. Erişim adresi: https://slideplayer.com/slide/5334684/

Bigss D, Ville B, Suen E. A method of choosing multiway partitions for classification and decision trees. J Appl Stat. 1991;18(1):49-62.

Bigus JP. Data mining with neural networks: Solving business problems from application development to decision support. McGraw-Hill. 1996:403-12.

Bozkır AS, Sezer E, Gök B. Öğrenci seçme sınavında öğrenci başarısını etkileyen faktörlerin veri madenciliği yöntemleriyle tespiti. V. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09). 13-15 Mayıs, Karabük Üniversitesi. 2009:37-43.

Breiman L, Friedman J, Stone CJ. Olshen RA. Classification and Regression Trees. Taylor and Francis, Chapman&Hall/CRC. 1984.

Breiman L. Random Forests. J Mach Learn. 2001;45(1):5-32.

Bridge D. Classification: K-nearest neighbours. Online Courses [Internet]. 2013 [ET: 26.03.2018] Erişim adresi: www.cs.ucc.ie/~dgb/ courses/tai/notes/handout4.pdf

Brown P, Pietra F, DeSouza P, Lai V. Class-based n-gram models of natural language. J Comput Ling. 1992;18:467-79.

Carvalho DR ve Freitas AA. A hybrid decision tree/genetic algorithm method for Data mining. J Inf Sci. 2004;163(1-3):13-35.

Chan F, Cheing G, Chan JYC, Rosenthal DA, Chronister J. Predicting employment outcomes of rehabilitation clients with orthopedic disabilities: A CHAID analysis. Disabil Rehabil. 2006;28(5):257-70.

Coşkun C ve Baykal A. Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karsılaştırılması. 13.Akademik Bilişim Konf. 2 Şubat-4 Mayıs. Malatya. 2011.

Cover TM ve Hart PE. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.

Cunningham P ve Delany SJ. K-neighbor classifiers. J Multiple Classifier Syst. 2007;34(8):1-17.

Çalış K, Gazdağı O, Yıldız. Reklam içerikli e-postaların metin madenciliği yöntemleri ile otomatik tespiti. Bilişim Teknolojileri Derg. 2013;6(1):1-7.

Çalışkan SK ve Soğukpınar İ. K*x*Knn: K-means ve K-en yakın komşu yöntemleri ile ağlarda nüfuz tespiti. Elektrik Mühendisleri Odası (EMO) Yayınları. 2008:120-4.

Çetin O, Verit FF, Zebitay AG, Aydın Z, Kurdoğlu Z, Yücel O. Neither early nor late for becoming pregnant: Comparison of the perinatal outcomes of adolescent. reproductive age and advanced maternal age pregnancies. Clinical Investigation. Turk J Obstet Gynecol. 2015;12(3):151-7.

Daş B ve Türkoğlu İ. Biyomedikal sinyallerde sınıflandırma uygulamaları. DNA dizilimlerindeki nükleotit çiftlerinin frekans değerlerine göre farklı sınıflandırma yöntemleri ile karşılaştırılması. TIPTEKNO. 25-27 Eylül 2014. Kapadokya.

Dekhtyar A. Knowledge discovery from data. Distance/Similarity Measures. 2009 [Internet]. users.csc.calpoly.edu/~dekhtyar/560-Fall2009/lectures/lec09.466.pdf 2009 [ET: 19.02.2018].

Delaunay B. Sur la sphere vide. Izv. Akad. Nauk SSSR. Otdelenie Matematicheskii Estestvennyka Nauk. 1934;7(793-800):1-2.

Doolatabadi HR, Hoseini SM, Tahmasebi R. Using decision tree model and logistic regression to predict companies financial bankruptcy in Tehran stock exchanges. Int J Emerging Res Manage Technol. 2013;2(9):7-16.

Dudani SA. The distance-weighted k-nearest-neighbor rule. IEEE Trans Syst Man Cybern. 1976;4:325-7.

Dunham MH. Data Mining: Introductory and advanced topics. Prentice-Hall. Upper Saddle River. NJ USA. 2003.

Ertan H, Alemdar H, İncel ÖD. Ersoy C. Bilişsel rahatsızlıkları olan kişilerin yaşam kalitesini artırmak için akıllı bir koltuk tasarımı. Signal Process Commun Appl Conf (SIU-20). IEEE. 2012:1-4.

Eui-Hong SH, Srivastava A, Kumar V. Parallel formulations of inductive classification learning algorithm. Department of Computer Science Technical Report. 1996:96-9.

Fan G ve Zhang J. A novel geometric diagram and its applications in wireless networks. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM. 7-11 March. 2004;1:681-3.

Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861-74.

Fix E ve Hodges JL. Discriminatory analysis-nonparametric discrimination: consistency properties. California Univ. Berkeley.1951.

Friedman JH. Multivariate adaptive regression splines. Ann Stat. 1991;13(2):50-67.

Friedman JH. Stochastic gradient boosting. Comput StatData Anal. 2002;38(4):367-78.

Fukunaga K, Hostetler L. K-nearest neighbor Bayes-risk estimation. IEEE Trans Inf Theory. 1975;21(3):285-293.

Goodman LA. Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories. J Am Stat Assoc. 1979;74:537-52.

Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng. 2003;15(6):1437-47.

Han DJ, Kamber M. Data mining: concepts and techniques. Morgan Kaufmann Publishers. San Francisco; 2000:332-6.

Han DJ, Kamber M. Data mining: concepts and techniques. 2nd ed.. Morgan Kaufmann Publishers: ISBN: 978-1-55860-901-3; 2006.

Harrington P. Machine learning in action. Newyork: Manning Publications Shelter Island; 2012.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning; data mining. inference and prediction. Springer Ser Stat. 2001:533-4.

Neyzi O, Ertuğrul T. Pretermlerin fizik özellikleri. Pediatri, 3.Baskı, Nobel Tıp Kitapevi, İstanbul, 2002:326-7.

İnternet. C4.5 Karar Ağacı [İnternet]. 2012 [ET: 07.03.2019]. Erişim adresi: http://bilgisayarkavramlari.sadievrenseker.com/2012/11/13/c4-5-agaci-c4-5-tree/

İnternet. Karar Ağacı Öğrenmesi. C4.5 algoritması [İnternet]. 2017 [ET: 08.03.2019]. Erişim adresi: http://yazilimagiris.com/2017/11/karar-agaci/

İnternet. The random forest algorithm [İnternet]. 2018 [ET: 10.03.2019]. Erişim adresi: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

İnternet. Selcuk University Digital Archive Systems [İnternet]. 2018 [ET: 06.01.2019]. Erişim adresi: acikerisim.selcuk.edu.tr:8080

Jacobs P. Data Mining: What general managers need to know. Harv Manage Update. 1999;4(10):8-9.

Jozwik A. A learning scheme for a fuzzy k-nn rule. Pattern Recognit Lett. 1983;1(5-6):287-9.

Kantardzic M. Data mining: concepts. models. methods. and algorithms. John Wiley & Sons. 2011.

Karakış R, Tez M, Güler İ. Classification the axillary lymph node status of breast cancer patients with the analysis of pattern recognition. Signal Process Commun Appl. SIU, IEEE'19 Conf. 2011:988-91.

Kass GV. An exploratory technique for investigating large quantities of categorical data. J Appl Stat. 1980;20(2):119-27.

Kavzoğlu T, Çölkesen İ. Karar ağaçları ile uydu görüntülerinin sınıflandırılması: Kocaeli örneği. Harita Teknolojileri Elektronik Derg. 2010;2(1):36-45.

Kitler R, Wang W. The Emerging Role of Data Mining. Solid State Technol. 1998;42(11):45-7.

Kovalerchuk B, Vityaev E. Data mining in finance: advances in relational and hybrid methods. Kluwer Academic Publisher: USA; 2002:308-9.

Koyuncugil AS, Özgülbaş N. Veri madenciliğinin tıp ve sağlık alanında kullanımı. Bilişim Teknolojileri Derg. 2009;2(2):21-32.

Köktürk F. K-en yakın komşuluk. yapay sinir ağları ve karar ağaçları yöntemlerinin sınıflandırma başarılarının karsılaştırılması [Doktora Tezi]. Bülent Ecevit Ünv.; 2012.

Kresse W, Danko DM. Springer handbook of geographic ınformation. Springer-Verlag: Berlin; 2012

Kusiak A, Kernstine KH, Kern JA, McLaughlin KA, Tseng TL. Medical and Engineering Case Studies. In: Proceedings of the Industrial Engineering Res Conf Cleveland; Ohio, May 2000:1-7.

Küçük H, Tepe C, Eminoğlu I. Classification of EMG signals by k-nearest neighbor algorithm and support vector machine methods. 21st Signal Processing and Commun Appl Conf (SIU) IEEE; 2013.

Lawson CL. Software for C1 surface ınterpolation. Rice J Ed. In Mathematical Software III: Academic Press; New York. 1977:161-94.

Lewis RJ. An introduction to classification and regression tree (CART) analysis. In Annual Meeting of the Society for Academic Emergency Medicine. San Francisco. California. 2000.

Lim TS, Loh WY, Shih YS. A comparison of prediction accuracy. complexity. and training time of thirty-three old and new classification algorithms. Mach Learn. 2000;40(3):203-28.

Loh WY, Shih YS. Split selection methods for classification trees. Stat Sinica.1997;7:815-40.

Lopez H, Zitto T, Bare P, Vidal G, Vukasovic J, Gomez R. Prevalence of anti-hepatitis a antibodies in an urban middle class area of argentina: some associated factors. Int J Infect Dis. 1999;4:34-7.

Luan J. Data mining and its applications in higher education. In Knowledge Management: Building a Competitive Advantage in Higher Education (Edited by A. Serban and J. Luan). Jossey Bass: San Francisco. USA; 2002:17-36.

Maleki M, Eroğlu K, Aydemir Ö, Manshoori N, Kayıkçıoğlu T. A new method for selection optimum k value in k-nn classification algorithm. In Signal Processing and Communications Applications Conference (SIU). 21st IEEE; 2013.

Mao C, Hu B, Wang M, Moore P. Learning from neighborhood for classification with local distribution characteristics. IEEE International Joint Conference on Neural Networks (IJCNN); 2015.

Maynard DN, Hibbeler JD. Measurement and reduction of critical area using Voronoi diagrams. Advanced Semiconductor Manufacturing Conference and Workshop. IEEE/SEMI; 2005:243-9.

McAllister M, Snoeyink L. Medial axis generalization of river networks. Cartography Geog Inf Sci. 2000;27(2):129-38.

McCarty J, Hastak M. Segmentation approaches in data-mining: A comparison of RFM. CHAID and logistic regression. J Buss Res. 2007;60(6):656-62.

Michie D, Spiegelhalter DJ, Taylor CC. Machine learning, neural and statistical classification. Prentice Hall: New York, USA; 1994:131-74.

Mingers J. An empirical comparison of selection measures for decision-tree induction. March. J Mach Learn. 1989;3(4):227-43.

Morgan JN, Sonquist JA. Problems in the analysis of survey data and a proposal. J Am Stat Assoc. 1963;58(302):415-34.

Oğuzlar A. CART analizi ile hane halkı iş gücü anketi sonuçlarının özetlenmesi. Atatürk Üniv. İktisadi ve İdari Bilimler Fakültesi Derg. 2004;18(3-4):79-90.

Olson DL, Delen D, Meng Y. Comparative analysis of data mining methods for bankruptcy prediction. Decis Support Syst. 2012;52(2):464-73.

Omitaomu OA. Decision Trees. In Michael W. Berry and Murray Browne (Eds). Lecture Notes in Data Mining. World Scientific Publishing of Hackensack; New Jersey. USA: 2006:39-51.

Pehlivan NY, Apaydın A. Bulanık k-en yakın komşuluk tahmin edicisi ve bulanık radyal tabanlı fonksiyon ağları. Selçuk Üniversitesi, Fen Fakültesi, Fen Derg, 2005;1(26):19-32.

Quinlan JR. Induction of decision trees. International J Mach Learn. 1986;1(1):81-106.

Quinlan JR. Simplifying decision trees. Int J Man-Mach Stud. 1987;27(3):221-34.

Quinlan JR. Decision trees and decision-making. IEEE Trans Syst Man Cybern. 1990;20(2):339-46.

Quinlan JR. C4.5: Programs for Machine Learning. Morgan Kaufmann: Sn Mateo. CA; 1993.

Quinlan JR. C4.5: programs for machine learning. Elsevier; 2014.

Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell. 2006;28(10):1619-30.

Sağlık Bakanlığı. Halk Sağlığı Genel Müdürlüğü [internet] 2017. [ET: 27.02.2019] https://hsgm.saglik.gov.tr/tr/ beslenme/gebelik-doneminde-beslenme.html.

Santos MF, Cortez P, Pereira J, Quintela H. Corporate bankruptcy prediction using data mining techniques. WIT Trans Inf Commun Technol. 2006;37:349-57.

Sezgin E, Çelik Y. Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması, XV. Akademik Bilişim Konferansı Bildirileri 23-25 Ocak. Akdeniz Üniversitesi. Antalya; 2013:23-25.

Shafer JC, Agrawal R, Mehta M. SPRINT: A scalable parallel classifier for data mining. In Proc 22. Int Conf Very Large Databases. Morgan Kaufmann. 1996;3(6):544-55.

Sibson R. Localy Equingular Triangulations. J Comput. 1997;21:243-5.

Singh G, Chouhan C, Sidhu MK. Maternal Factors for Low Birth Weight Babies. Med J Malaysia. 2009;65:10-12.

Stackexchance. [internet] 2018 [ET: 28.02.2019]. Erişim adresi: https://stats. stackexchange.com/questions.

Teknomo K. K-means clustering tutorial. J Med. 2006;100(4):3-7.

Türe M, Tokatlı F, Kurt İ. Using Kaplan-Meier analysis together with decision tree methods (C&RT. CHAID. QUEST. C4.5. ID3) in determining recurrence-free survival of breast cancer patients. Expert Syst Appl. 2009;36(2):2017-26.

Voronoi G. Sur un problème du calcul des fonctions asymptotiques. Journal für die reine und angewandte Mathematik. 1903;126:241-82.

Witten IH, Frank E. Data Mining: Practical Machine learning tools with Java implementations; SanFrancisco. California: 2000:371-2.

Witten IH, Frank E, Morgan K. Mining: Practical machine learning tools and techniques 2nd edition, Morgan Kaufmann, San Francisco. 2005.

Zekic-Susac M, Sarlija N, Bensic M. Small business credit scoring: a comparison of logistic regression. Neural network and decision tree models. Information Technology Interfaces. 26th International Conference on IEEE; 2004:265-70.

Zenciroğlu A, Gündüz R, Onat R, Dilli D ve ark. Türkiye Halk Sağlığı Kurumu. Temel Yenidoğan Bakımı Çocuk ve Ergen Sağlığı Daire Başkanlığı. 2015 Ankara 4.

Zhao G, Joshi RP, Lakdawala VK. Percolative breakdown model for ceramics based on a random grain-boundary network. Electrical Insulation and Dielectric Phenomena CEIDP'05. Annual Report Conference on IEEE; 2005:544-7.

Zhou Y, Li Y, Xia S. An improved KNN text classification algorithm based on clustering. J Comp. 2009;4(3):230-7.

ÖZGEÇMİŞ

Sadi ELASAN; 1984 yılında Van/İpekyolu'da doğdu. 2002 yılında Milli Piyango Anadolu Lisesinden mezun oldu. Aynı yıl Ankara Üniversitesi Sağlık Bilimleri Fakültesi'ne başladı ve 2006 yılında mezun oldu. Üniversiteden mezun olduğu yıl KPSS ile merkezi ataması yapıldı. 2007 yılında Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü Tıp Biyoistatistik Anabilim Dalında Yüksek Lisans öğrenimine başladı. 2013 yılında Yüzüncü Yıl Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalında Öğretim Görevlisi kadrosuna atandı ve ardından Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü Tıp Biyoistatistik Anabilim Dalında Doktora öğrenimine başladı. Evli ve bir çocuk babasıdır.

EKLER

EK 1. Etik Kurul Raporu

T.C.
YÜZÜNCÜ YIL ÜNİVERSİTESİ
Girişimsel Olmayan Klinik Araştırmalar
Etik Kurul Başkanlığı

Sayı : B.30.2.YYU.0.01.00.00/25 Tarih: 21/02/2018

Konu :Etik Kurul Başvurunuz

Sn. Prof.Dr. Sıddık KESKİN

İlgi: 08.02.2018 tarih ve bila sayılı yazınız.

İlgi yazı ile Etik Kurulumuza sunulan **"Veri Madenciliğinde Karar Ağaçları ve K-En Yakın Komşu Yöntemlerinin İncelenmesi "** isimli proje özetiniz 16.02.2018 tarihinde yapılan Etik Kurul toplantımızda görüşüldü. Yapılan görüşmede çalışmanın proje özetinin içeriği tamamiyle istatiksel analiz değerlendirmesi olduğundan adı geçen çalışmanın etik kurulla doğrudan ilişkisi saptanmamıştır.

Bilgilerinize rica ederim.

Prof. Dr. Oğuz TUNCER
Girişimsel Olmayan
Klinik Araştırmalar
Etik Kurul Başkanı

Adres: 65080 Zeve Kampüsü - VAN Tel : (0432) 2150470 E-Posta: etikkurull@gmail.com
Fax : (0432) 2168352 Web: http://www.yyu.edu.tr

EK 2. Veri Kullanım İzin Belgesi

VERİ KULLANIM İZİN BELGESİ

Perinatal ve erken neonatal sonuçların karşılaştırılması amacıyla, İstanbul Süleymaniye Eğitim ve Araştırma Hastanesi'nde 1 Ocak 2007 ve 31 Ocak 2015 tarihleri arasında doğum yapan 910 kadından elde edilmiş ve bilimsel makalemizde (*Çetin ve ark. 2015)* yayınlanmış verilerin; uygulanacak istatistik yöntemlerin performanslarının değerlendirilmesi amacıyla; Öğr. Gör. Sadi ELASAN'ın *"Veri Madenciliğinde Karar Ağaçları ve K-En Yakın Komşu Yöntemlerinin İncelenmesi"* adlı Doktora Tez Çalışmasında kullanılmasında tarafımızca herhangi bir sakınca bulunmamaktadır.

Bilgilerinize arz olunur. 01.03.2018

Sorumlu (Correspondence) Yazar
Doç. Dr. Orkun ÇETİN

ÇİĞLİ BÖLGE EĞİTİM HASTANESİ
Doç. Dr. Orkun ÇETİN
Dip. No: 45972 Diploma Tescil No: 125102
Perinatoloji Uzmanı

EK 3. Tez Orjinallik Raporu

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
Sağlık Bilimleri Enstitüsü

DOKTORA TEZİ ORİJİNALLİK RAPORU

Tarih: 07/05/2019

Tez Başlığı/Konusu:
Veri Madenciliğinde Farklı Karar Ağaçları ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları ve Doğum Verisinde Bir Uygulama

Yukarıda başlığı/konusu belirlenen tez çalışmamın Kapak sayfası, Giriş, Ana bölümler ve Sonuç bölümlerinden oluşan toplam **77 sayfalık** kısmına ilişkin. 07/05/2019 tarihinde şahsım tarafından **Turnitin** intihal tespit programından aşağıda belirtilen filtreleme uygulanarak alınmış olan orijinallik raporuna göre tezimin **benzerlik oranı %7 (yedi)** dir.

Uygulanan filtreler aşağıda verilmiştir:
- Kabul ve onay sayfası hariç
- Teşekkür hariç
- İçindekiler hariç
- Simge ve kısaltmalar hariç
- Gereç ve yöntemler hariç
- Kaynakça hariç
- Alıntılar hariç
- Tezden çıkan yayınlar hariç
- 7 kelimeden daha az örtüşme içeren metin kısımları hariç (Limit match size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

Sadi ELASAN

Öğrencinin Adı Soyadı	:	Sadi ELASAN
Anabilim Dalı	:	Biyoistatistik AD
Öğrenci No	:	149302057
Programı	:	☐ Yüksek Lisans ☒ Doktora

DANIŞMAN ONAYI
UYGUNDUR
Prof. Dr. Sıddık KESKİN

ENSTİTÜ ONAYI
UYGUNDUR
Hacer ŞAHİN AYDINYURT

EK 4. Python kodları

```
##Random Forest
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier(n_estimators=10)
clf = clf.fit(X_train, y_train)

# In[58]:
y_pred_random = clf.predict(X_test)
y_pred_random

# In[59]:
print("Train-Acurracy",metrics.accuracy_score(y_train,clf.predict(X_train)))
print("Train-Confusion Matrix",metrics.confusion_matrix(y_train,clf.predict(X_train)))
print("Train-Classification_report",metrics.classification_report(y_train,clf.predict(X_train)))
print("Train-RMSE:",metrics.mean_squared_error(y_train,clf.predict(X_train)))

# In[60]:
print("Test-Acurracy",metrics.accuracy_score(y_test,clf.predict(X_test)))
print("Test-Confusion Matrix",metrics.confusion_matrix(y_test,clf.predict(X_test)))
print("Test-Classification_report",metrics.classification_report(y_test,clf.predict(X_test)))
print("Test-RMSE:",metrics.mean_squared_error(y_test,clf.predict(X_test)))

# In[61]:
print ("Accuracy is ", accuracy_score(y_test,y_pred_random)*100)

# In[62]:
from IPython.display import Image
from sklearn.externals.six import StringIO
import pydotplus as pydot
forest_clf = RandomForestClassifier(criterion = "gini", random_state = 100, max_depth=5,
min_samples_leaf=3)
forest_clf.fit(X_train, y_train)
dot_data = StringIO()
tree.export_graphviz(forest_clf.estimators_[2], out_file=dot_data,
feature_names=features,filled=True,rounded=True)
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
Image(graph.create_png())

# In[63]:
y_pred_random = forest_clf.predict(X_test)
y_pred_random

# In[64]:
print("Train-Acurracy",metrics.accuracy_score(y_train,forest_clf.predict(X_train)))
print("Train-Confusion Matrix",metrics.confusion_matrix(y_train,forest_clf.predict(X_train)))
print("Train-
Classification_report",metrics.classification_report(y_train,forest_clf.predict(X_train)))
print("Train-RMSE:",metrics.mean_squared_error(y_train,forest_clf.predict(X_train)))

# In[65]:
print("Test-Acurracy",metrics.accuracy_score(y_test,forest_clf.predict(X_test)))
```

```
print("Test-Confusion Matrix",metrics.confusion_matrix(y_test,forest_clf.predict(X_test)))
print("Test-Classification_report",metrics.classification_report(y_test,forest_clf.predict(X_test)))
print("Test-RMSE:",mean_squared_error(y_test,forest_clf.predict(X_test)))

# In[66]:
print ("Accuracy is ", accuracy_score(y_test,y_pred_random)*100)
```

##C4.5

```
from sklearn.ensemble import C4.5Classifier

#!/usr/bin/env python
# coding: utf-8

# In[41]:
import numpy as np
import pandas as pd
from sklearn import tree
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.tree import C4.5DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import mean_squared_error

input_file = "/home/dl4cvubuntu/DataScience/DataScience-Python3/dogumson.xlsx"
df = pd.read_excel(input_file)

# In[42]:
df.head()

# In[43]:
features = list(df.columns[1:])
features

# In[44]:
y = df["Dogum Agirligi"]
X = df[features]

# In[45]:
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size = 0.3, random_state = 100)

# In[46]:
clf_gini = DecisionTreeClassifier(criterion = "gini", random_state = 100, max_depth=5,
min_samples_leaf=2)
clf_gini.fit(X_train, y_train)

# In[47]:
from IPython.display import Image
from sklearn.externals.six import StringIO
import pydotplus
```

```
dot_data = StringIO()
tree.export_graphviz(clf_gini, out_file=dot_data, feature_names=features,
filled=True,rounded=True)
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
Image(graph.create_png())

# In[48]:

print("Train-Acurracy",metrics.accuracy_score(y_train,clf_gini.predict(X_train)))
print("Train-Confusion Matrix",metrics.confusion_matrix(y_train,clf_gini.predict(X_train)))
print("Train-Classification_report",metrics.classification_report(y_train,clf_gini.predict(X_train)))
print("Train-RMSE:",metrics.mean_squared_error(y_train,clf_gini.predict(X_train)))

# In[49]:
print("Test-Acurracy",metrics.accuracy_score(y_test,clf_gini.predict(X_test)))
print("Test-Confusion Matrix",metrics.confusion_matrix(y_test,clf_gini.predict(X_test)))
print("Test-Classification_report",metrics.classification_report(y_test,clf_gini.predict(X_test)))
print("Test-RMSE:",metrics.mean_squared_error(y_test,clf_gini.predict(X_test)))

# In[50]:
#entropy
clf_entropy = DecisionTreeClassifier(criterion = "entropy", random_state = 100,
max_depth=5, min_samples_leaf=5)
clf_entropy.fit(X_train, y_train)

# In[51]:
print("Train-Acurracy",metrics.accuracy_score(y_train,clf_entropy.predict(X_train)))
print("Train-Confusion Matrix",metrics.confusion_matrix(y_train,clf_entropy.predict(X_train)))
print("Train-
Classification_report",metrics.classification_report(y_train,clf_entropy.predict(X_train)))
print("Train-RMSE:",metrics.mean_squared_error(y_train,clf_entropy.predict(X_train)))

# In[52]:
print("Test-Acurracy",metrics.accuracy_score(y_test,clf_entropy.predict(X_test)))
print("Test-Confusion Matrix",metrics.confusion_matrix(y_test,clf_entropy.predict(X_test)))
print("Test-Classification_report",metrics.classification_report(y_test,clf_entropy.predict(X_test)))
print("Test-RMSE:",metrics.mean_squared_error(y_test,clf_entropy.predict(X_test)))

# In[53]:
y_pred = clf_gini.predict(X_test)
y_pred

# In[54]:
y_pred_en = clf_entropy.predict(X_test)
y_pred_en

# In[55]:
print ("Accuracy is ", accuracy_score(y_test,y_pred)*100)

# In[56]:
print ("Accuracy is ", accuracy_score(y_test,y_pred_en)*100)
```