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

VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARI İLE YUMURTALIK KANSERİ HASTALARININ İNCELENMESİ: RECEP TAYYİP ERDOĞAN ÜNİVERSİTESİ EĞİTİM VE ARAŞTIRMA HASTANESİ ÖRNEĞİ

(Yüksek Lisans Tezi)

İlyas YILMAZ

Dr. Öğr. Üyesi Burcu KARTAL
Danışman

RİZE
2020

KABUL VE ONAY

Recep Tayyip Erdoğan Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Ana Bilim Dalında, İlyas YILMAZ tarafından hazırlanan "*Veri Madenciliği Sınıflandırma Algoritmaları ile Yumurtalık Kanseri Hastalarının İncelenmesi: Recep Tayyip Erdoğan Üniversitesi (RTEÜ) Eğitim ve Araştırma Hastanesi Örneği*" başlıklı bu çalışma 14/07/2020 tarihinde yapılan savunma sınavı sonucu oy birliği ile başarılı bulunarak jürimiz tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Başkan: Prof. Dr. Abdullah NARALAN Kabul/Red

Üye: Prof. Dr. Birdoğan BAKİ Kabul/Red

Üye: Dr. Öğr. Üyesi Burcu KARTAL Kabul/Red

…/…/…

Doç. Dr. Ahmet YANIK

Enstitü Müdürü

ETİK BEYAN

Tezli yüksek lisans programından mezun olmak üzere teslim ettiğim "*Veri Madenciliği Sınıflandırma Algoritmaları ile Yumurtalık Kanseri Hastalarının İncelenmesi: Recep Tayyip Erdoğan Üniversitesi (RTEÜ) Eğitim ve Araştırma Hastanesi Örneği*" konulu tezimi, bilim ve araştırma etiği prensiplerine riayet edilerek tarafımdan yazılmıştır.

Eserimde, başka kaynaklardan aktarılan bütün bilgi ve alıntılar, Enstitünüz Tez Yazım Kılavuzuna uygun olarak açıkça gösterilmiştir. Kaynağı gösterilenler dışında kalan bütün bilgiler uygun araştırma yöntemi kullanılarak tarafımdan edinilmiş ve esere bu şekilde yansıtılmıştır. Şahsıma ait olmayan hiçbir bilgi, kasıt veya kusurlar, şahsıma aitmiş gibi gösterilmemiştir. İnternet kaynakları dahil, sahibine / kaynağına atıf yapılmaksızın hiçbir bilgi kullanılmamıştır.

Aksinin ortaya çıkması halinde doğacak bütün hukuki, idari, akademik ve etik sorumluluk tarafıma ait olacaktır. Eserin tesliminden sonra herhangi bir zamanda, bilim etiğine aykırılık tespit edilmesi ve / veya eserimle ilgili intihal veya intihal şeklinde anlaşılacak bir durumun ortaya çıkması halinde; Üniversiteniz ve eğitim kadronuzun hiç bir şekilde sorumlu tutulmayacağını hür irademle kabul, beyan ve taahhüt ederim. **14/07/2020**

İlyas YILMAZ

ÖN SÖZ

Son zamanlarda oldukça önem kazanmış konuların başında yer alan veri madenciliği, veri tabanlarında anlamsız bir şekilde bulunan bu verilerden anlamlı bilgiler keşfetmemize olanak sağlamaktadır. Borsa, Bankacılık, Pazarlama, Sigortacılık ve Endüstri gibi veri tabanlarında milyonlarca verinin depolanabildiği birçok alanda kullanılan veri madenciliği, bu çalışmada yumurtalık kanseri teşhisi konulan hastalara ait gerçek verilerin analiz edilmesi için sağlık alanında kullanılmıştır. Bu çalışmada, yumurtalık kanseri için belirlenen hedef değişkenlerin sınıflandırılmasına en çok katkı sağlayan tanımlayıcı değişkenlerin belirlenmesi ve sınıflandırma algoritmalarına ait performans ölçütlerinin karşılaştırılması amaçlanmıştır. Belirlenen amaç doğrultusunda temel sınıflandırma algoritmaları kullanılmıştır. Analiz sonucunda elde edilen bulguların yumurtalık kanseri ile ilgili çalışma yapacak olan akademisyenlere ve sağlık çalışanlarına fayda sağlayacağı düşünülmüştür.

Eğitim öğretim hayatıma önemli katkılarda bulunan danışmanım Sayın Dr. Öğr. Üyesi Burcu KARTAL'a ve çalışmam boyunca bilgi ve önerilerinden faydalandığım Sayın Arş. Gör. Mehmet Fatih SERT'e teşekkürlerimi sunarım.

Verilerin elde edilmesi sürecinde yardımcı olan RTEÜ Eğitim ve Araştırma Hastanesi İdari ve Mali İşler Müdür Yardımcısı Sayın Vedat TANIŞ'a, verilerin düzenlenmesi ve değişkenlerin belirlenmesi kısmında değerli vaktini bizim için harcayan Sayın Uzm. Dr. Elanur KARAMAN'a teşekkür ederim.

Son olarak öğrenim hayatım boyunca hiçbir fedakarlıktan kaçınmayan annem Rahşan YILMAZ ve babam Mustafa YILMAZ başta olmak üzere abim Yunus Emre YILMAZ ve kardeşim Muhammet Uğur YILMAZ'a teşekkürlerimi sunmayı bir borç bilirim.

İlyas YILMAZ

RİZE 2020

İÇİNDEKİLER

ÜÇÜNCÜ BÖLÜM

MATERYAL VE YÖNTEM

DÖRDÜNCÜ BÖLÜM

YUMURTALIK KANSERİ VERİ SETİNE VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARININ UYGULANMASI

Recep Tayyip Erdoğan Üniversitesi Sosyal Bilimler Enstitüsü
Ana Bilim Dalı: İşletme
Tez Türü: Yüksek Lisans Tezi
Danışman: Dr. Öğretim Üyesi Burcu KARTAL
Hazırlayan: İlyas YILMAZ
Yıl: 2020
Sayfa Sayısı: 120

ÖZET

VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARI İLE YUMURTALIK KANSERİ HASTALARININ İNCELENMESİ: RTEÜ EĞİTİM VE ARAŞTIRMA HASTANESİ ÖRNEĞİ

Bu tezin amacı, yumurtalık kanseri hastalarından elde edilen veri setine veri madenciliği sınıflandırma algoritmalarının uygulanması ve bu algoritmaların performanslarının karşılaştırılması olarak belirlenmiştir. Tez için kullanılan veri seti Recep Tayyip Erdoğan Üniversitesi Eğitim ve Araştırma Hastanesi Onkoloji bölümünden elde edilmiştir. Çalışmamızda uzman görüşüne başvurularak Tedaviye Yanıt, Sağ Kalım ve Tümörün Lokalizasyonu olmak üzere üç hedef değişken belirlenmiştir. Hedef değişkenlerin sınıflandırılmasında Naive Bayes, Destek Vektör Makinesi (DVM), Lojistik Regresyon, J48 Karar Ağacı ve k-En Yakın Komşu algoritmaları kullanılmıştır. Algoritmaların uygulanması aşamasında Weka paket programından yararlanılmıştır. Çalışma sonucunda, veri madenciliği algoritmalarının elde ettiği sınıflandırma performansları neticesinde oluşan sıralama her hedef değişkeni için farklılık göstermiştir. Ayrıca lokalizasyon hedef değişkeninde elde edilen sınıflandırma performanslarının oldukça düşük olduğu görülmüştür.

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma, Yumurtalık Kanseri, Naïve Bayes, Destek Vektör Makinesi (DVM), Lojistik Regresyon, 148 Karar Ağacı, k-En Yakın Komşu, Weka.

Recep Tayyip Erdogan University Graduate School of Social Sciences

Department: Business Administration

Thesis Type: Master Thesis

Supervisor: Dr. Burcu KARTAL

Author: İlyas YILMAZ

Year: 2020

Pages: 120

ABSTRACT

INVESTIGATION OF PATIENTS OF OVERIAN CANCER WITH DATA MINING CLASSIFICATION ALGORITHMS: RTEU EDUCATION AND RESEARCH HOSPITAL SAMPLE

The purpose of this thesis is to apply, data mining classification algorithms to the data set obtained from ovarian cancer patients and to compare the performance of these algorthms. The data set used was taken from Recep Tayyip Erdoğan University Education and Research Hospital Oncology Department. In our study, three target variables, namely Response to Treatment, Survival and Localization of the Tumor, were determined by seeking expert opinion. Naive Bayes, Support Vector Machine (DVM), Logistic Regression, J48 Decision Tree and k-Nearest Neighbor algorithms were used to classify target variables. Weka package program was used during the implementation of the algorithms. As a result of the study, the ordering resulting from the classification performances obtained by data mining algorithms have been differed for each target variable. In addition, the classification performances obtained in the localization target variable have been seen to be quite low.

Key Words: Data Mining, Classification, Ovarian cancer; Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, 148 Decision Tree, k-Nearest Neighbor, Weka.

KISALTMALAR DİZİNİ

AUC	Area Under the ROC Curve
ANN	Artificial Neural Network
CART	Classification and Regression Trees
DVM	Destek Vektör Makinesi
EM	Expectation Maximization
FS	Feature Selection
K-NN	K-Nearest Neighbor
NB	Naive Bayes
RF	Random Forest
ROC	Receiver Operrating Characterstic
RTEÜ	Recep Tayyip Erdoğan Üniversitesi
SMO	Sıralı Minimal Optimizasyon
TP	True Positive
TN	True Negative
UCI	University of California at Irvine
VM	Veri Madenciliği
VTBK	Veri Tabanlarında Bilgi keşfi
VTYS	Veri Tabanı Yönetim Sistemleri
WEKA	Waikato Environment for Knowledge Analysis

TABLOLAR LİSTESİ

ŞEKİLLER TABLOSU

GİRİŞ

Teknoloji alanında yaşanan gelişmeler ve teknolojinin kolay ulaşılabilir olması yaşantımızın her alanında bilgisayar, tablet, telefon ve internet kullanımını oldukça yaygınlaştırmıştır. Yaygınlaşan teknoloji kullanımı mevcut veri ve gizli kalmış bilgi oranında yüksek bir artış yaşanmasına ve veri yığınlarının oluşmasına neden olmuştur. Oluşan veri yığınları veriyi depolama, saklama ve koruma maliyetlerini de beraberinde getirmiştir. Dezavantaj gibi görünen bu durumun, veri yığınlarından anlamlı bilgilerin elde edilmesi ve veriler arasında gizli kalmış ilişkilerin ortaya çıkarılması ile kişi, kurum ve kuruluşlara birçok konuda avantaj sağlayacağı anlaşılmıştır. Ancak veri yığınlarının oluşması ile birlikte, yüzyıllardır kullanılan geleneksel analiz yöntemlerinin veriyi birçok açıdan değerlendirme ve inceleme konusunda yetersiz kaldığı görülmüştür. Bu durum verilerin analizinde yeni bir yöntemin kullanılması ihtiyacını ortaya çıkarmıştır.

Günümüzde sağlık, bankacılık, borsa, pazarlama ve endüstri gibi birçok alanın veri tabanlarında milyonlarca veri bulunmaktadır. Bu verilerin analizinde veri madenciliği algoritmalarının kullanılması, verilerin farklı açılardan incelenmesine, analiz edilmesine ve değerlendirilmesine olanak sunmuştur. Örneğin; pazarlama alanında müşteri satın alma alışkanlıklarının belirlenmesinden, bankacılık alanında kredi kartı dolandırıcılıklarının belirlenmesine kadar birçok alanda veri madenciliği algoritmaları kullanılabilmektedir. Bu sayede veri tabanlarında işlevsiz ve anlamsız bir biçimde bulunan milyonlarca veriden anlamlı bilgiler keşfedilmektedir. Özellikle sağlık alanında keşfedilen anlamlı bilgiler ile sağlık çalışanlarının karar verme sürecine katkıda bulunmak insan hayatı için büyük bir önem arz etmektedir. Bu nedenle sağlık alanında veri madenciliği algoritmalarının kullanılması oldukça önemli ve kıymetlidir.

Çalışmada, yumurtalık kanseri teşhisi konulan hastalara ait gerçek veriler incelenmiştir. Yaygın bir şekilde görülen kadınsal hastalıklardan birisi olan yumurtalık kanserinin tedavi süreci oldukça uzun ve zordur. (Ergün vd., 1996:

274). Buna rağmen yumurtalık kanserinin, diğer kanser türleri gibi erken teşhis edilmesi ile tedavi sürecinde başarı sağladığı görülmüştür. Ancak bazı koşullarda yumurtalık kanseri belirtileri, hastalık ilerleyene kadar hiçbir rahatsızlığa ve şikayete yol açmamaktadır. Bu nedenle hastalığın erken teşhis edilmesinde bu belirtilerin kullanılması oldukça zordur Hastalık, ancak hastalardan alınan doku hücrelerinin incelenmesi ile ortaya çıkmaktadır. (Kurman vd., 2008: 351). Bu yüzden hastalığın daha fazla ilerlememesi için teşhisten sonra tedavilerin aksatılmadan uygulanması, hastaların tedavi süreçlerinin takip edilmesi, yumurtalık kanseri ameliyatı olan ve kemoterapi gören hastaların nüks olma ihtimaline karşı kontrollerinin düzenli olarak yapılması gerekmektedir. Bu çalışma ile öncelikle, tedaviye yanıt, sağ kalım ve lokalizasyon hedef değişkenlerinin sınıflara ayrılmasında, en çok etkiye sahip tanımlayıcı değişkenlerin belirlenmesi amaçlanmaktadır. Ayrıca, kullanılan veri madenciliği sınıflandırma algoritmaları ile oluşturulan modellerin performanslarının karşılaştırılması sağlanacaktır.

Çalışmasının birinci bölümünde, veri madenciliği tekniklerinin kanser araştırmalarında kullanılmasına yönelik literatür çalışmalarından bahsedilecektir.

Çalışmanın ikinci bölümünde, veri madenciliği tarihçesi, veri madenciliği tanımı ve veri madenciliğinin önemi üzerinde durulmuş, daha sonra veri madenciliği kullanım alanları, veri madenciliğinde karşılaşılan problemler, veri madenciliği süreci, veri madenciliği yöntemleri ve çalışmada kullanılan veri madenciliği sınıflandırma algoritmaları hakkında genel bilgiler verilmiştir.

Çalışmanın üçüncü bölümünde araştırmanın amacına ve araştırmanın önemine değinilmiştir. Daha sonra veri toplama süreci, veriyi anlama, veri temizleme, model oluşturma ve model değerlendirme süreçleri ele alınmıştır.

Çalışmanın son bölümünde ise; yumurtalık kanseri teşhisi konulan hastalara ait veri setine Naive Bayes, Destek Vektör Makinesi (DVM), Lojistik Regresyon, J48 Karar Ağacı ve k-En Yakın Komşu algoritmalarının uygulanması ile elde edilen bulgular ve sınıflandırma performanslarının karşılaştırılmasına ilişkin değerlendirmeler yer almaktadır.

BİRİNCİ BÖLÜM

LİTERATÜR TARAMASI

Sağlık alanında yaşanan teknolojik yenilikler ve bu yeniliklerin sağlık personellerine doğru ve hızlı bir şekilde ulaştırılması oldukça önemlidir. Çünkü yaşanan teknolojik gelişmeler ile bazı hastalıkların erken teşhis edilmesi, hastalıklar üzerinde etkili olabilecek özelliklerin belirlenmesi ve tedavilerinin gerçekleştirilmesi mümkün olabilmektedir. Özellikle bu durum kanser hastalığı gibi erken teşhis edilmesi durumunda tedavi şansını yükseltebilecek hastalıklar için hayati değer taşımaktadır. Bu nedenle kanser hastalıklarının veri madenciliği algoritmaları ile incelenmesi ve son yıllarda bu konu ile ilgili literatür çalışmalarında artış yaşanması oldukça doğaldır. Literatür taraması sonucunda kanser hastaları üzerinde yapılan veri madenciliği çalışmaları, kanser hastalıklarının erken teşhis edilmesi, kanser hastalarının tahmin edilmesi ve hastalık üzerinde etkili olan özelliklerin belirlenmesi konuları üzerinde yoğunlaşmaktadır. Konu ile ilgili yapılan çalışmaların birçoğunda hazır veri kaynağının kullanılması ile benzer sonuçlar elde edilmektedir. Ancak gerçek hasta verilerinin ve farklı yöntemlerin kullanıldığı çalışmalar ile farklı sonuçların elde edildiği çalışmalarda azımsanmayacak kadar fazladır.

Geliştirilen birçok açık kaynak kodlu program ile verilerin analizinde veri madenciliği algoritmalarının kolayca uygulanabilir olması ve sağlık alanında elde edilebilecek her bilginin hayati önem taşıması yerli literatürde veri madenciliği çalışma sayısının önemli ölçüde artmasına neden olmuştur. Bu çalışmaların çoğunlukla son yirmi yıl içerisinde yapılmış olması da dikkatlerden kaçmamaktadır.

Bu bölümde, farklı kanser türlerini inceleyen, çalışmanın farklı boyutlarını ele alan ve anlamsal olarak katkı sağlayabilecek veri madenciliği sınıflandırma algoritmalarının kullanıldığı çalışmalar detaylı bir şekilde incelenmiştir. İlgili çalışmalar, temelde "kanser, sağlık, veri madenciliği, cancer, health ve data mining" anahtar kelimelerinin kullanılması ile Google Scholar üzerinden taranmıştır. Çalışmalara ilişkin özet tablo, Tablo 1'de sunulmuştur.

Tablo 1. Literatür Özet Tablosu.

Yazar- Yıl	Kanser Türü	Performans Değerlendirme Ölçütü	Kullanılan Yöntemler
Akgöbek ve Kaya (2011)	Göğüs Kanseri	Doğruluk Oranı Standart Sapma Değeri	C4.5, NavieBayes, PART, CORE, Ant-Miner, CN2, GA-SVM
Coşkun ve Baykal (2011)	Göğüs Kanseri	Hata Oranı Kesinlik Duyarlılık F-ölçütü	J48, Naive Bayes, lojistik regresyon, Kstar
Riberio vd., (2011)	Karaciğer Kanseri	Doğruluk Oranı	Destek Vektör Makinesi, Bayesian ve k-En Yakın Komşu
Şentürk (2011)	(Bütün Kanser Türleri)	Güvenilirlik Oranı	k-En Yakın Komşu Algoritması, Birliktelik kuralı, Diskriminant Analizi, Karar Ağacı
Poyraz (2012)	Göğüs Kanseri	Doğruluk Kesinlik Duyarlılık F-ölçütü	Naive Bayes, J48, Lojistik Regresyon ve K-Star
Şık (2014)	Göğüs Kanseri	Doğruluk Kesinlik Duyarlılık F-ölçütü ROC alanı	Bayes Ağı, Naive Bayes, Çok Katmanlı Algılayıcı, Basit Lojistik, SGD, SMO, IB1, KStar, Lojistik Model Ağları, Rassal Orman
Demircioğlu ve Bilge (2015)	Yumurtalık Kanseri	Doğruluk Oranı	Destek Vektör Makinesi ve k-En Yakın Komşu
Bektaş ve Batur (2016)	Göğüs Kanseri	Doğruluk Duyarlılık Kesinlik MCC AUC	DVM, K-Star, Rastgele Orman Algoritması ve Seçimli Algılayıcı Sinir Ağı Yöntemi
Hambali ve Gbolagade (2016)	Yumurtalık Kanseri	Doğruluk TP FP Kesinlik Duyarlılık F-Ölçütü	RBF, Çok Katmanlı Algılayıcı (LMP), SMOTE+ LMP ve SMOTE + SMOTE

Tablo 2. (Devam) Literatür Özet Tablosu.

Yazar- Yıl	Kanser Türü	Performans Değerlendirme Ölçütü	Kullanılan Yöntemler
Osmanovic vd. (2017)	Yumurtalık Kanseri	Doğruluk oranı	J48, LMT ve Çok Katmanlı Algılayıcı
Alaybeyoğlu ve Mulayim (2018)	Karaciğer Kanseri	Doğruluk Duyarlılık Özgüllük Pozitif Yordama Değer Negatif Yordama Değer	Destek Vektör Makinesi (DVM)
Nabawy vd. (2018)	Yumurtalık Kanseri	Doğruluk Kesinlik Duyarlılık Özgünlük F-Ölçütü AUC	Destek vektör Makinesi, Rastgele Orman, Lojistik Regresyon, Boosting
Sebik ve Bülbül (2018)	Akciğer Kanseri	Doğruluk Duyarlılık Kesinlik F-ölçütü	Naive Bayes, Bayes Net, Lojistik Regresyon K-Star, Bagging, OneR, ZeroR, J48 ve Random Tree
Yasodha ve Ananthanarayanan (2018)	Yumurtalık Kanseri	Doğruluk Ortalama Kare Hata Değeri	Destek Vektör Makinesi (SVM), Çok Katmanlı Algılayıcı (MLP), İleri Beslemeli Sinir Ağı (FFNN). Kombine SOMICS ve GENN algoritmaları
Yücebaş (2018)	Melonom ve Prostat Kanseri	Kesinlik Duyarlılık ROC Alanı	Karar Ağacı, Naive Bayes, Destek Vektör Makinesi (SVM)
Nuhić vd. (2019)	Yumurtalık Kanseri	Doğruluk	Naive Bayes, Çok Katmanlı Algılayıcı, Basit Lojistik, En yakın komşu, AdaBoost, Rastgele Komite, PART, LMT ve Rastgele Orman

Tablo 3. (Devam) Literatür Özet Tablosu.

Yazar- Yıl	Kanser Türü	Performans Değerlendirme Ölçütü	Kullanılan Yöntemler
Sardouk (2019)	Göğüs Kanseri	TP Oranı FP-Oranı Duyarlılık Kesinlik F-Ölçütü ROC alanı	AdaBoost M1, Classification Via Rregression, Rastgele Orman, Jrip, RBF-NN ve J48
Sevli (2019)	Göğüs Kanseri	Doğruluk Duyarlılık Belirleyicilik Kesinlik	Naïve Bayes, Rastgele Orman, k-En Yakın Komşu ve Lojistik Regresyon
Taşdelen (2019)	Göğüs Kanseri	Doğruluk Duyarlılık Kesinlik F-Ölçütü	Perceptron Öğrenme Algoritması, k-En Yakın Komşu (KNN) ve Derin Öğrenme Algoritması
Thulasiraman ve Kavitha (2019)	Yumurtalık Kanseri	Doğruluk Kesinlik Duyarlılık Özgünlük	Destek Vektör Makinesi ve Çok Katmanlı Algılayıcı
Yalçın (2019)	Göğüs Kanseri	Doğruluk Duyarlılık Kesinlik F-Ölçütü ROC Alanı MCC PRC	ID3, C4.5, Rassal Orman, CART, LMT, Naive Bayes, Lojistik Regresyon, KNN, SVM ve ANN
Yetginler (2019)	Rahim Ağzı Kanseri	Doğruluk Oranı Duyarlılık	Naive Bayes, J48 Karar Ağacı ve Destek Vektör Makinesi (SVM)

Medikal verilerden bilgi keşfetmeyi amaçlayan Akgöbek ve Kaya (2011) Wisconsin Breast Cancer, Ljubljana Breast Cancer, Dermatology, Hepatitis ve Diabetes veri setleri ile çalışmayı yürütmüştür. Veri setlerine ait örneklem sayısı sırasıyla 699, 286, 366, 155 ve 768 ‘dir. Her veri seti kendi içerisinde bulunan özelliklere göre değerlendirilmiştir. Elde edilen veri setlerinin analizinde C4.5,

NavieBayes, PART, CORE, Ant-Miner, CN2, GA-SVM algoritmaları kullanılmıştır. Algoritmaların performansını karşılaştırmak için ise ortalama doğruluk oranı, en iyi doğruluk oranı ve standart sapma değerleri incelenmiştir. Çalışma sonucunda REX-1 algoritmasının medikal veri setlerini sınıflandırmada elde ettiği performanslar üzerinde durulmuştur.

Coşkun ve Baykal (2011) çalışmalarında, veri madenciliği algoritmalarını SEER veri seti üzerinde uygulamış, veri seti üzerinde algoritmaların elde ettiği başarım ölçütlerini karşılaştırmışlardır. Çalışmada kullanılan veri seti National Cancer Institute (NCI)'den elde edilmiştir. Yıllık olarak düzenlenen SEER veri kaynağından alınan 2008 yılına ait göğüs kanseri hastaları çalışmanın örneklemini oluşturmuştur. Ayrıca veri setine ait 118 nitelik çalışma kapsamında incelenmiştir. Elde edilen veri setinin analizinde ise J48, Naive Bayes, lojistik regresyon, Kstar algoritmaları kullanılmıştır. Algoritmaların kullanılmasında WEKA paket programından yaranılmıştır. Oluşturulan modellerin performanslarını karşılaştırmak için hata oranı, kesinlik, duyarlılık ve F-ölçütü kullanılmıştır. Çalışma sonucunda algoritmalar, her başarım ölçütü için elde ettikleri performansa göre sıralanmıştır. Elde edilen sonuçlara göre doğruluk derecesinde oluşan sıralama ile F- ölçütü sıralamasının aynı olduğu, kesinlik ve duyarlılık ölçütlerinde elde edilen sıralamanın ise birbirleri ile zıt olduğu vurgulanmıştır. Doğruluk ve F-ölçütüne göre en iyi performansın J48 algoritması ile elde edildiğinin altı çizilmiştir.

Riberio vd. (2011), karaciğer kanserinin tanımlanmasına yardımcı olmak için bu çalışmayı yürütmüşlerdir. Çalışmanın örneklemini karaciğer kanserinin beş farklı evresinde bulunan 88 hasta oluşturmuştur. Bu hastalardan elde edilen veri seti içerisinde bulunan hastalık özellikleri, klinik ve laboratuar bilgileri hastalığın evresini tahmin etmek için kullanılmıştır. Çalışmada Destek Vektör Makinesi, Bayesian ve k-En Yakın Komşu algoritmalarının performansları karşılaştırılmıştır. Karşılaştırma sonucunda en iyi sonuç %80,68 doğrulukla k-En Yakın Komşu algoritması ile elde edilmiştir. En iyi performansa sahip algoritma yardımı ile karaciğer kanseri olan hastaların evresinin tahmin edilebileceği vurgulanmıştır.

Hastalara kanser teşhisi koymak için veri madenliği algoritmalarını kullanan Şentürk (2011) bu çalışma ile sağlık sektörüne katkı sağlamayı amaçlamıştır. Bu amaç doğrultusunda Düzce Üniversitesi Eğitim ve Araştırma Hastanesi'nden alınan verilerin analize hazırlanma sürecinde PHP programlama dili ile MySQL veri tabanı yönetim sistemi kullanılmıştır. Ayrıca çalışmada kullanılan eğitim verisi kanser teşhisi konulmuş hasta verileri olarak belirlenirken, henüz kanser teşhisi konulmamış hasta verileri de test verisi olarak belirlenmiştir. Kanser teşhisi konulmayan hastalara ait veriler, k-En Yakın komşu algoritması, Birliktelik kuralı, Diskriminant analizi, Karar Ağacı ile analiz edilmiştir. Algoritmaların uygulanmasında ise Rapid Miner 5.0 paket programı kullanılmıştır. Çalışma sonuncunda, henüz kanser tanısı konulmamış hastaların tahmininde en doğru sonuçlara k-en yakın komşu algoritması ile ulaşılırken, en yanlış sonuçlar ise karar ağaçları yöntemi ile elde edilmiştir. Poyraz (2012), Wisconsin veri tabanından elde ettiği 1991 yılına ait meme kanseri verilerini veri madenciliği algoritmaları ile analiz etmiştir. Çalışmada sınıflandırma algoritmaları kullanılarak meme kanserinin iyi/kötü huylu şeklinde sınıflandırması ve algoritmaların başarım ölçütlerinin karşılaştırılması sağlanmıştır. Veri setinde bulunan 683 örnek çalışmanın örneklemini oluşturmuştur. Verilerin analizi Naive Bayes, J48, Lojistik Regresyon ve K-Star algoritmaları ile gerçekleştirilmiştir. Algoritmaların uygulanmasında ise açık kaynak kodlu WEKA paket programından yararlanılmıştır. Çalışmada performans değerlendirme ölçütü olarak doğruluk, duyarlılık, kesinlik ve F-ölçütü kullanılmıştır. Elde edilen bulgular sonucunda, algoritmalar arasında doğruluk ölçütüne göre en iyi performans lojistik regresyon algoritması ile %96.92 olarak elde edilmiştir. Kesinlik ve duyarlılık ölçütlerine göre oluşan başarı sıralamasının birbiri ile zıt olduğu da çalışma sonucunda gözlenmiştir.

Poyraz (2012), Wisconsin veri tabanından elde ettiği 1991 yılına ait meme kanseri verilerini veri madenciliği algoritmaları ile analiz etmiştir. Çalışmada sınıflandırma algoritmaları kullanılarak meme kanserinin iyi/kötü huylu şeklinde sınıflandırması ve algoritmaların başarım ölçütlerinin karşılaştırılması sağlanmıştır. Veri setinde bulunan 683 örnek çalışmanın örneklemini oluşturmuştur. Verilerin analizi Naive Bayes, J48, Lojistik Regresyon ve K-Star

algoritmaları ile gerçekleştirilmiştir. Algoritmaların uygulanmasında ise açık kaynak kodlu WEKA paket programından yararlanılmıştır. Çalışmada performans değerlendirme ölçütü olarak doğruluk, duyarlılık, kesinlik ve F-ölçütü kullanılmıştır. Elde edilen bulgular sonucunda, algoritmalar arasında doğruluk ölçütüne göre en iyi performans lojistik regresyon algoritması ile %96.92 olarak elde edilmiştir. Kesinlik ve duyarlılık ölçütlerine göre oluşan başarı sıralamasının birbiri ile zıt olduğu da çalışma sonucunda gözlenmiştir.

Göğüs kanserinin erken teşhis edilmesini amaçlayan Şık (2014) UCI veri tabanından elde edilen Wisconsin göğüs kanseri verilerini incelemiştir. Veri setinde bulunan 357'si iyi huylu tümör teşhisi konulan hasta ve 212'si kötü huylu tümör teşhisi konulan hasta olmak üzere toplam 569 hasta çalışmanın örneklemini oluşturmuştur. Veriler, çeşitli özellik değerlendiriciler kullanılarak Bayes Ağı, Naive Bayes, Çok Katmanlı Algılayıcı, Basit Lojistik, SGD, SMO, IB1, KStar, Lojistik Model Ağları, Rassal Orman algoritmaları ile analiz edilmiştir. Çalışmada performans değerlendirme yöntemi olarak 10-kat çapraz doğrulama kullanılmıştır. Ayrıca algoritmaların performanslarını değerlendirmek için ise kappa istatistiği, doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC alanı tercih edilmiştir. Çalışma sonucunda, basit lojistik algoritması ile doğruluk oranı %97,40 olarak elde edilmiştir. Elde edilen başarılı sonuçlar neticesinde meme kanserinde erken teşhisi sağlayabilmek için veri madenciliğinden yararlanılabileceği vurgulanmıştır.

Demircioğlu ve Bilge (2015) çalışmalarında, veri madenciliği algoritmaları kullanılarak halka açık bir veri kümesinden elde edilen yumurtalık kanseri hastalarına ait genlerin analizini gerçekleştirmişlerdir. Bu kapsamda veri setinde 91 tanesi sağlıklı, 162 tanesi hasta olmak üzere toplam 253 örnek ve 15154 adet gen incelenmiştir. Fisher korelasyon skorlama ve Weltch t testi kullanılarak veri setinde bulunan genler ilgililik neticesine göre sıralandırılmıştır. Daha sonra veriler, farklı parametreler kullanılarak Destek Vektör Makinesi ve k-En Yakın Komşu algoritmaları ile analiz edilmiştir. Algoritmaların performansları ise doğruluk oranı neticesinde karşılaştırılmıştır. Çalışma sonucunda, Destek Vektör Makinesi ve k-En Yakın Komşu algoritmalarına ait en iyi sınıflandırma performanslarının genel itibari ile ilk 100 gen için elde edildiği gözlenmiştir. Böylece yumurtalık kanseri veri setinin incelenmesinde binlerce gen yerine

yalnızca 100 genin kullanılmasının yeterli olabileceği yazarlar tarafından belirtilmiştir.

Makine öğrenme teknikleri kullanarak hastaların kalp ameliyatı sırasında veya sonrasında hayati risklerini belirlemeyi amaçlayan Kartal (2015) çalışmasında, Acıbadem Maslak Hastanesi'nden elde ettiği veri setini kullanmıştır. EuroSCORE risk faktörleri kullanılarak hastaların ölüm riski tahmin edilmeye çalışılmıştır. Çalışmada performans değerlendirme yöntemi olarak Tabakalı Çapraz Geçerleme ve Tabakalı Hold-Out kullanılmıştır. Veriler, k-En Yakın Komşu Algoritması, Logistik Regresyon Analizi, ID3 ve C4.5 Karar Ağacı Algoritmaları kullanılarak analiz edilmiştir. Algoritmaların performanslarını değerlendirmek için Kesinlik, Duyarlılık, F-Ölçüsü Pozitif Olabilirlik Oranı, Negatif Olabilirlik Oranı ve Tanısal Üstünlük Oranı kullanılmıştır. Çalışma sonucunda, en iyi performans C4.5 algoritması ile elde edilmiştir. C4.5 algoritması ve Lojistik Regresyon analizi sonucunda oluşan modeller için Shiny uygulaması geliştirilmiş ve shinyapps.io aracılığıyla web üzerinden sunulmuştur.

Bektaş ve Batur (2016), meme kanserinin teşhisi için birçok veri madenciliği algoritmasını kullanılarak performanslarını karşılaştırdıkları çalışmada *Kent Ridge 2* mikrodizi veri setini kullanmışlardır. Çalışmada kullanılan veri setinde 97 meme kanseri hastası ve bu hastalara ait 24482 özniteliğin bulunduğu görülmüştür. Kullanılan algoritmaların başarısını ve sınıflandırma performansını artırmak için veri setinde yer alan ilgisiz nitelikler veri setinden çıkarılmıştır. Bunun gerçekleştirilmesi için Correlation-based Feature Subset Selection öznitelik seçme yöntemi kullanılmıştır. Öznitelik seçimi sonucunda hastalık üzerinde yüksek vektörlere sahip 139 öznitelik belirlenerek çalışmada kullanılacak yeni veri seti elde edilmiştir. Veri seti için önişlemlerin tamamlanmasının ardından veriler, DVM, K-Yıldız, Rastgele Orman Algoritması ve Seçimli Algılayıcı Sinir Ağı Yöntemi ile analiz edilmiştir. Algoritmaların performanslarının değerlendirilmesi için ise Doğruluk, Duyarlılık, Kesinlik, MCC ve AUC ölçütleri kullanılmıştır. Çalışma sonucunda, 139 öznitelik ile elde edilen doğruluk oranında en iyi performans %90,72 ile Rastgele Orman Algoritmasında elde edilmiştir. Ayrıca çalışmada öznitelik seçimi yapılmadan önce 24482 özniteliğin yer aldığı veri seti de analiz edilmiştir. Öznitelik seçiminden sonra

oluşan 139 özniteliğin bulunduğu veri seti için elde edilen doğruluk oranlarının, seçim yapılmadan önceki veri seti için elde edilen doğruluk oranlarından daha yüksek olduğu vurgulanmıştır.

Hambali & Gbolagade (2016), halka açık veri tabanından elde ettikleri yumurtalık kanseri veri seti üzerinde SMOTE önişleme algoritmasını kullanarak algoritmaların sınıflandırma performansını arttırmayı amaçlamışlardır. Algoritmalar, veri seti üzerinde Sentetik Azınlık Aşırı Örnekleme Tekniği uygulanmadan önce ve uygulandıktan sonra sınıflandırmada kullanılmış ve iki durum için karşılaştırma yapılmıştır. Algoritmaların performanslarını karşılaştırmak için Doğruluk, TP, FP Kesinlik, Duyarlılık, F-Ölçütü ve ROC alanı kullanılmıştır. Çalışma sonucunda, veri seti üzerine SMOTE önişleme algoritmasının uygulanması ile elde edilen doğru sınıflandırma oranına katkı sağladığı görülmüştür. SMOTE uygulanmadan önce Çok Katmanlı Algılayıcı algoritması ile elde edilen doğru sınıflandırma oranı %95,3 iken, SMOTE uygulandıktan sonra %96,8 olarak gerçekleşmiştir. RBF algoritması için SMOTE uygulanmadan önce doğru sınıflandırma oranı %83,4 iken, uygulandıktan sonra %88,4 olarak elde edilmiştir.

Osmanovic vd. (2017) çalışmalarında, yumurtalık kanseri hastalarının hayatta kalıp kalamayacaklarını tahmin etmek için bu çalışmayı gerçekleştirmişlerdir. Danimarka kanser kayıtlarından elde edilen 318 hastaya ait veri seti J48, LMT ve Çok Katmanlı Algılayıcı algoritmaları ile WEKA paket programı kullanılarak analiz edilmiştir. Elde edilen veri setinde; kanserin büyüklüğü, kanser hareketliliği, hastanın yaşı, kanserin yüzeyi, kanserin tutarlılığı olmak üzere 5 öznitelik yer almıştır. Çalışmada öncelikle 318 hastanın bulunduğu veri seti sınıflandırılmaya tabi tutulmuş, daha sonra veri setinde yer alan eksik değerler veri setinden çıkarılarak sınıflandırma algoritmaları uygulanmıştır. Eksik verilerin kaldırılmasından sonra algoritmaların elde ettiği sınıflandırma performanslarında ciddi bir artış yaşandığı gözlenmiştir. Ayrıca veri setinde bulunan kanserin hareketliliği, kanserin yüzeyi ve kanserin tutarlılığı özniteliklerinin sınıflandırmada elde edilen sonuçlara önemli ölçüde katkı sağladığı görülmüştür. Kanserin büyüklüğü ve hastanın yaşı özniteliklerinin ise

sınıflandırmada elde edilen sonuçlar üzerinde pek bir etki yaratmadığı gözlenmiştir.

Tseng vd. (2017), yumurtalık kanseri olan hastaların nüks olma durumları üzerinde etkili olabilecek faktörleri belirlemek ve nüks durumunu tahmin edebilmek için veri madenciliği algoritmalarını kullanmışlardır. Chung Shan Tıp Üniversitesi Hastanesi Tümör Sicilinden elde edilen hastalara ait tıbbi bilgiler ve hastanın patoloji durumuna ilişkin veriler, Destek Vektör Makinesi, C5.0, Aşırı Öğrenme Makinesi, Çok Değişkenli Uyarlanabilir Regresyon Spline ve Rastgele Orman (RF) algoritmaları ile analiz edilmiştir. Veri setinde bulunan 987 hasta ve 13 özellik bu çalışma kapsamında incelenmiştir. Çalışmada, eğitim verisi rastgele seçilmiş 687 hastadan oluşurken, geriye kalan 300 hasta ise test verisi olarak ayrılmıştır. Bu durum, 10 defa oluşturulan bağımsız eğitim ve test verisi için tekrar edilerek 10 defa uygulanmıştır. Çalışma sonucunda, FIGO, Patolojik M, Yaş ve Patolojik T risk faktörlerinin yumurtalık kanseri nüksü için en etkili faktörler olduğu tespit edilmiştir. Risk faktörleri belirlendikten sonra algoritmaların sınıflandırma performanslarının arttığı görülmüştür. En yüksek sınıflandırma performansı ise C5.0 algoritması ile %90 olarak elde edilmiştir.

Karaciğer kanserinin teşhisi için uzman sistem tasarımı geliştirmeyi amaçlayan Alaybeyoğlu ve Mulayim (2018) çalışmalarında Indian Liver Patient Dataset (ILPD) veri tabanından elde edilen veri setini kullanmışlardır. Çalışmada, veri setinde bulunan 583 karaciğer hastası ve bu hastalara ait 8 özellik incelenmiştir. Veri setinde yer alan bilgilerin 167 tanesi karaciğer kanseri olmayan hastalardan elde edilirken, 416 tanesi karaciğer kanseri olan hastalardan elde edilmiştir. Elde edilen verilerin analizinde Destek Vektör Makinesi kullanılmıştır. Algoritmanın performansını değerlendirmek için ise Doğruluk, Duyarlılık, Özgüllük, Pozitif Yordama Değer ve Negatif Yordama Değerlerinden faydalanılmıştır. Elde edilen bulgular neticesinde önerilen sistemin sağlık personellerine ve uzmanlara hastalığın teşhis edilmesinde yardımcı olabileceği düşünülmektedir.

Nabawy vd. (2018), bünyesinde iki alt tür bulunduran yumurtalık kanserine ait verileri Destek vektör Makinesi, Rastgele Orman, Lojistik Regresyon, Boosting algoritmaları ile R yazılımı kullanılarak analiz etmişlerdir.

Veri setinde bulunan 194 hasta ve bu hastalara ait 32.898 gen ve 12 özellik çalışma kapsamında incelenmiştir. Öncelikle kanser alt tipini sınıflandırmak için veri setinde bulunan gen ekspresyonu boyutsal olarak azaltılmış ve klinik veriler ile birleştirilmiştir. Daha sonra önerilen bu yaklaşım farklı sınıflandırma algoritmaları ile test edilmiştir. Kullanılan algoritmaların performansları ise Doğruluk, Kesinlik, Duyarlılık, Özgünlük, F-Ölçütü ve AUC değerleri kullanılarak karşılaştırılmıştır. Çalışma sonucunda, tüm karşılaştırma ölçekleri için en iyi sonuçlar Boosting algoritması ile edilmiştir.

Sebik ve Bülbül (2018), sağlık bakanlığı veri tabanında bulunan daha önce hastalık teşhisi konulmuş kimliği belli olmayan hasta verilerini veri madenciliği sınıflandırma algoritmaları ile analiz etmişlerdir. Yazarların amacı, akciğer kanserine veri madenciliği sınıflandırma algoritmaları uygulayarak hangi algoritmanın daha iyi performans elde ettiğini ortaya çıkarmaktadır. Bu doğrultuda elde edilen veriler çalışmada kullanılması için önişlemlerden geçirilmiştir. Önişlemlerin tamamlanmasının ardından 404 adet hastaya ait 9 adet öznitelik üzerinden çalışma yürütülmüştür. Çalışmada akciğer kanseri hastalarının sınıflandırılması için Naive Bayes, Bayes Net, Lojistik Regresyon K-Star, Bagging, OneR, ZeroR, J48 ve Random Tree algoritmaları kullanılmıştır. Sınıflandırma algoritmalarının uygulanması aşamasında açık kaynak kodlu WEKA paket programından yararlanılmıştır. Performans değerlendirme ölçütleri olarak Doğruluk, Kesinlik, Duyarlılık ve F-Ölçütü incelenmiştir. Çalışma sonucunda, en iyi performans Naive Bayes algoritması ile elde edilirken, en düşük performansın ise Zeror algoritmasında gerçekleştiği sonucuna varılmıştır.

Yasodha ve Ananthanarayanan (2018), yumurtalık kanseri veri seti üzerinde Kendi Kendini Düzenleyen Haritalar Bağışıklık Klon Seçimi (SOMICS) ve Dilbilgisel Evrim Sinir Ağları (GENN) yöntemini uygulayarak yeni bir yaklaşım sunmuşlardır. Ayrıca, yumurtalık kanseri olan 2970 hastadan elde edilen bilgiler, Destek Vektör Makinesi, Çok Katmanlı Algılayıcı, İleri Beslemeli Sinir Ağı. Kombine SOMICS ve GENN algoritmaları ile analiz edilmiş ve algoritmalar elde ettiği sonuçlara göre birbiri ile kıyaslanmıştır. Elde edilen bulgular sonucunda, %98,23 ile en yüksek doğruluk oranı ve 0,0021 ile en düşük ortalama kare hata değeri Kombine SOMICS ve GENN yöntemi ile üretilmiştir.

Yücebaş (2018) çalışmasında, veri madenciliği algoritmaları ile melonom ve prostat kanser hastalarına ait verileri incelemiştir. Çalışmada kullanılan veri setleri dbGaP veri tabanından elde edilmiştir. Ayrıca prostat kanseri veri setinin Afro Amerikan, Japon ve Latin birçok etnik kökene sahip hastalardan oluştuğu vurgulanmıştır. Daha sonra veri seti önişlemlerden geçirilerek çalışmada kullanılacak son halini almıştır. Çalışmanın analizinde ise Karar Ağacı, Naive Bayes, Destek Vektör Makinesi algoritmaları kullanılmıştır. Algoritmaların performanslarını karşılaştırmak için kesinlik, duyarlılık ve ROC eğrisinden faydalanılmıştır. Çalışma sonucunda algoritmaların performanslarının karşılaştırılması neticesinde her iki veri seti içinde en yüksek kesinliğin Destek Vektör Makinesi ile elde edildiği gözlemlenmiştir.

Yumurtalık kanserinin teşhisini amaçlayan Nuhić vd. (2019) çalışmalarında, veri madenciliği sınıflandırma algoritmalarını kullanmışlardır. Gene Expression Omnibus (GEO) veritabanından elde edilen 91 tanesi yumurtalık kanseri ve 57 tanesi yumurtalık kanseri olmayan toplam 148 örnek çalışma kapsamında incelenmiştir. Elde edilen veriler, Naive Bayes, Çok Katmanlı Algılayıcı, Basit Lojistik, En yakın komşu, AdaBoost, Rastgele Komite, PART, LMT ve Rastgele Orman algoritmaları ile analiz edilmiş ve algoritmalar doğru sınıflandırma oranına göre birbirleri ile kıyaslanmıştır. Çalışma sonucunda, en yüksek doğru sınıflandırma oranı %96,78 olarak LMT algoritması ile elde edilmiştir. Diğer algoritmalar da elde edilen doğru sınıflandırma oranlarının birbirine yakın olduğu gözlenmiştir. Ayrıca elde edilen doğru sınıflandırma oranları literatürde benzerlik gösteren çalışma sonuçlar ile karşılaştırılmıştır.

Sardouk (2019), öznitelikler yardımı ile hastaların meme kanserine yakalanıp yakalanmadığını tahmin etmek amacıyla bu çalışmayı gerçekleştirmiştir. Çalışmada, Coimbra veri tabanında elde edilen meme kanseri hastaları ve bu hastalara ait öznitelikler, veri madenciliği algoritmaları kullanılarak incelenmiştir. Veriler, WEKA paket programı kullanılarak AdaBoost M1, Classification Via Rregression, Rastgele Orman, Jrip, RBFNN, J48 algoritmaları ile analiz edilmiştir. Analizlerin performanslarını karşılaştırmak için TP Oranı, FP-Oranı, Duyarlılık, Kesinlik, F-Ölçütü ve ROC alanı kullanılmıştır.

Sevli (2019) çalışmasında, Wisconsin Üniversitesinden elde ettiği göğüs kanseri hastalarına ait veri setini sınıflandırmak için Destek vektör makinesi, Naive Bayes, Rastgele Orman, k-En Yakın Komşu ve Lojistik Regresyon algoritmalarını kullanmış ve bu algoritmaların performanslarını karşılaştırmıştır. Çalışmanın örneklemini Wisconsin üniversitesinden elde edilen 569 göğüs kanseri hastası oluşturmuştur. Bu veri setinin %80'i eğitim veri seti olarak kullanılırken %20'si test veri seti olacak biçimde rastgele seçilmiştir. Ayrıca çalışmada bu veri setine ait 10 nitelik incelenmiştir. Algoritmaların performansını denetlemek için ise Doğruluk, Duyarlılık, Belirleyicilik ve Kesinlik ölçütleri kullanılmıştır. Wisconsin üniversitesinden alınan göğüs kanseri veri setinin sınıflandırılmasında en iyi performansın Lojistik Regresyon ile elde edildiği görülmüştür.

Taşdelen (2019) çalışmasında, UCI Makine Öğrenme Deposunda yer alan çeşitli hastalıklara ait verileri perceptron öğrenme algoritması, KNN algoritması ve derin öğrenme algoritmaları ile analiz etmiştir. Çalışmada, göğüs kanseri, pima yerlilerine ait diyabet, mamografik kitle ve bupa karaciğer hastalık verileri %70'i eğitim veri seti ve %30'u test veri seti olacak şekilde ikiye ayrılarak incelenmiştir. Kullanılan algoritmaların başarımlarını karşılaştırmak için doğruluk, hassasiyet, kesinlik ve F-ölçüsü değerleri kullanılmıştır. Verilerin analiz edilmesi sonucunda ise, kullanılan algoritmalar doğruluk oranları baz alınarak karşılaştırmış, tüm veri seti üzerinde en yüksek doğruluk oranlarının derin öğrenme algoritması ile elde edildiği görülmüştür. Çalışmacı tarafından önerilen algoritma ise derin öğrenme algoritmasına yakın sonuçlar üretmiştir.

Yalçın (2019) tez çalışmasında, biyopsi öncesinde meme kanseri hastalığını tespit etmek amacı ile verilerin analiz sürecinde ID3, C4.5, Random Forest, CART, LMT, Naive Bayes, Logistic Regresyon, KNN, SVM ve ANN algoritmalarını kullanmıştır. Çalışma kapsamında, Winconsin Üniversite Hastanesi'nden 1989-1991 yılları arasında elde edilen 669 hasta ve bu hastalara ait 10 özellik incelenmiştir. Algoritmaların uygulanmasında ise WEKA paket programından faydalanılmıştır. Algoritmaların birbirleri ile performanslarını karşılaştırmak için doğruluk, duyarlılık, kesinlik, F-ölçütü, ROC alanı, MCC ve PRC değerleri kullanılmıştır. Çalışma sonucunda, en iyi doğruluk oranı SVM ve

Naive Bayes algoritmaları ile %97 civarında elde edilmiştir. En düşük doğruluk oranı ise %92,6 ile Lojistik Regresyon algoritmasında gerçekleşmiştir.

Kanser riskini tahmin edebilecek bir sistem oluşturmayı hedefleyen Thulasiraman ve Kavitha (2019) çalışmalarında, veri madenciliği sınıflandırma ve kümeleme yöntemlerini kullanarak yumurtalık kanseri hastalarına ait verileri incelemişlerdir. Çalışmanın örneklemini 121'i yumurtalık kanser hastası ve 95'i normal hasta olmak üzere toplam 216 hasta oluşturmuştur. Veri setinde bulunan 113 hasta eğitim verisi olarak ayrılırken, 103'ü test verisi olarak ayrılmıştır. Verilerin analizinde ise Destek Vektör Makinesi ve Çok Katmanlı Algılayıcı algoritmaları kullanılmıştır. Algoritmaların performanslarını karşılaştırmak için Doğruluk, Kesinlik, Duyarlılık ve Özgünlük değerleri kullanılmıştır. Uygulama sonucunda, karşılaştırmada kullanılan tüm ölçütler için en iyi sonuçlar Çok Katmanlı Algılayıcı algoritması ile elde edilmiştir. Ayrıca çalışma sonucunda araştırmacılar tarafından kanser risk tahmin sistemi önerilmiştir.

Veri madenciliği algoritmaları kullanarak rahim ağzı kanserinin erken teşhisini amaçlayan Yetginler (2019) çalışmasında, California University Irvine (UCI) veri tabanından elde edilen rahim ağzı kanser veri setini kullanmıştır. Kullanılan veri setinde 858 kadın ve dört tanesi hedef değişken (Hinselmann, Schiller, Sitoloji ve Biyopsi) olmak üzere toplam 36 özellik incelenmiştir. Veri setine ait dengesizliği gidermek için SMOTE yöntemi tercih edilmiştir. Önişlemlerin tamamlanmasının ardından son halini alan veri setine Naive Bayes, J48 Karar Ağacı ve Destek Vektör Makinesi (DVM) algoritmaları uygulanmıştır. Algoritmaların performansları ise doğrululuk ve duyarlılık ölçütleri ile karşılaştırılmıştır. Çalışma sonucunda, rahim ağzı kanseri Biopsy hedef değişkeni ile diğer hedef değişkenlere göre daha başarılı bir şekilde tespit edilmiştir. Destek Vektör Makinesinin diğer algoritmalara göre daha iyi sonuçlar elde ettiği gözlenmiştir. Ayrıca J48 algoritmasının hasta kayıtlarını doğru tahmin etmede başarısız olduğu saptanmıştır.

Bu çalışma yumurtalık kanseri ile ilgili literatürde bulunan diğer çalışmalardan aşağıdaki yönleri ile farklılaşmaktadır.

- Çalışmada kullanılan hedef ve tanımlayıcı değişkenler uzman görüşü ile belirlenmiş, Tedaviye Yanıt, Sağ Kalım ve Lokalizasyon hedef değişkenleri çalışmamız kapsamında incelenmiştir.
- Verilerin analizinde kullanılan algoritmalar için performans değerlendirme yöntemi tabakalı 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme ve %50/%50, %75/%25, %80/%20 ve %90/%10 oranlarında Hold-Out olarak tercih edilmiştir.
- K-En Yakın Komşu algoritması için diğer algoritmalarda en yüksek doğru sınıflandırma performansının elde edildiği tabakalı çapraz geçerleme performans değerlendirme yöntemi olarak belirlenmiştir.
- J48 Karar Ağaç algoritması ile hedef değişkenlerin karar ağacı çizilmiş ve sınıflandırmaya en çok katkı sağlayan tanımlayıcı değişkenler belirlenmiştir.
- Performans değerlendirme ölçüsü olarak birbirinden farklı birçok ölçüt kullanılmış; ancak doğru sınıflandırma oranı ve F-Ölçüsü değerlendirmelerde yer almıştır.

Ayrıca literatür taraması sonucunda, yumurtalık kanseri hastalarının veri madenciliği sınıflandırma algoritmaları ile incelenmesi üzerine uluslararası birçok kaynak yer alırken, ulusal kaynakta bu konu ile ilgili yalnızca bir çalışma bulunmaktadır. Bu çalışmanın içeriği ise çalışmamızla benzerlik taşımamaktadır.

İKİNCİ BÖLÜM

VERİ MADENCİLİĞİ

2.1. VERİ MADENCİLİĞİ HAKKINDA GENEL BİLGİLER

2.1.1. Veri Madenciliği Kavramı

Bilgisayarların kısa zamanda istenilen bilgilere ulaşmak için yeterli olmaması 1960'lı yıllarda bu problemin çözümü için veri madenciliği yöntemini ortaya çıkarmıştır. Bu yönteme daha önceleri veri yakalanması (data fishing), veri taraması (data dredging) gibi bazı isimler verilse de bilgisayar mühendisleri tarafından 1990'lı yıllarda veri madenciliği ismi verilmiştir (Dönmez, 2008: 18). Kavramsal açıdan bakıldığında analizler sonucunda milyonlarca veri arasından değerli bilgiyi ortaya çıkarma işlemini gerçekleştiren veri madenciliği bu yönüyle, değerli madenlerin çıkarılması için uygulanan madencilik çalışmalarına benzetilmekte ve veri madenciliği ismi buradan gelmektedir (Zaiane, 1999'dan aktaran: Yetginler, 2019: 8).

2.1.2. Veri Madenciliğinin Tanımı

Veri madenciliğinin kesin bir tanımı olmamakla beraber en basit şekli ile veri madenciliği, büyük ölçekli verilerin içerisinden faydalı bilgiye erişme, bilgiyi madenleme işi olarak tanımlanır (Wikipedia, 2020). Bu tanım dışında araştırmacılar tarafından veri madenciliği ile ilgili birçok tanım daha yapılmış ve çalışmamızda bunlardan bazılarına yer verilmiştir. Bu tanımlar şöyledir;

- Berry ve Linoff'a (2004) göre: veri madenciliği, çok büyük miktardaki verinin içerisinde gizli kalmış anlamlı kurallar, yollar ve modeller keşfetme işlemidir.
- Fayyad ve arkadaşlarının (1996) yaptıkları çalışmaya göre: veri madenciliği, kabul görülebilir sayısal verimlilik neticesinde, veri analizi ve keşif algoritmalarının uygulanması ile veri içerisinde bulunan ilişki ve örüntüleri ortaya çıkaran Bilgi Keşfi Sürecinin bir adımıdır.

- Frawley ve arkadaşlarının (1992) yaptıkları çalışmada veri madenciliği, yararlı olma potansiyeli bulunan verilerin keşfedilmesi olarak tanımlanmıştır.
- Özbay (2015) çalışmasında, milyonlarca verinin arasından "değeri olan" anlamlı bilgilerin ortaya çıkarılması işlemini veri madenciliği olarak tanımlar.
- Adriaabs ve Zantinge'ne (1996) göre: veri madenciliği, kurallar keşfetmek ve anlamlı bilgiler elde etmek amacı ile büyük miktardaki verilerin analiz edilmesini sağlayan bir yaklaşımdır (Dönmez, 2008: 17).
- Lakshmanan ve arkadaşları (2015) veri madenciliğini, veri tabanlarında bulunan milyonlarca anlamsız ve düzensiz veriler arasından, gizli kalmış yararlı ve anlamlı bilgileri keşfetmek için kullanılan bir teknoloji olarak tanımlamaktadırlar.

Yapılan veri madenciliği tanımlarından yola çıkarak veri madenciliği, milyonlarca veri arasında gizli kalmış bilgiler içerisinden isteğimiz doğrultusunda yararlı ve anlamlı olabilecek bilgilerin elde edilmesi işlemi olarak tanımlanabilir.

2.1.3.Veri Madenciliği Tarihçesi

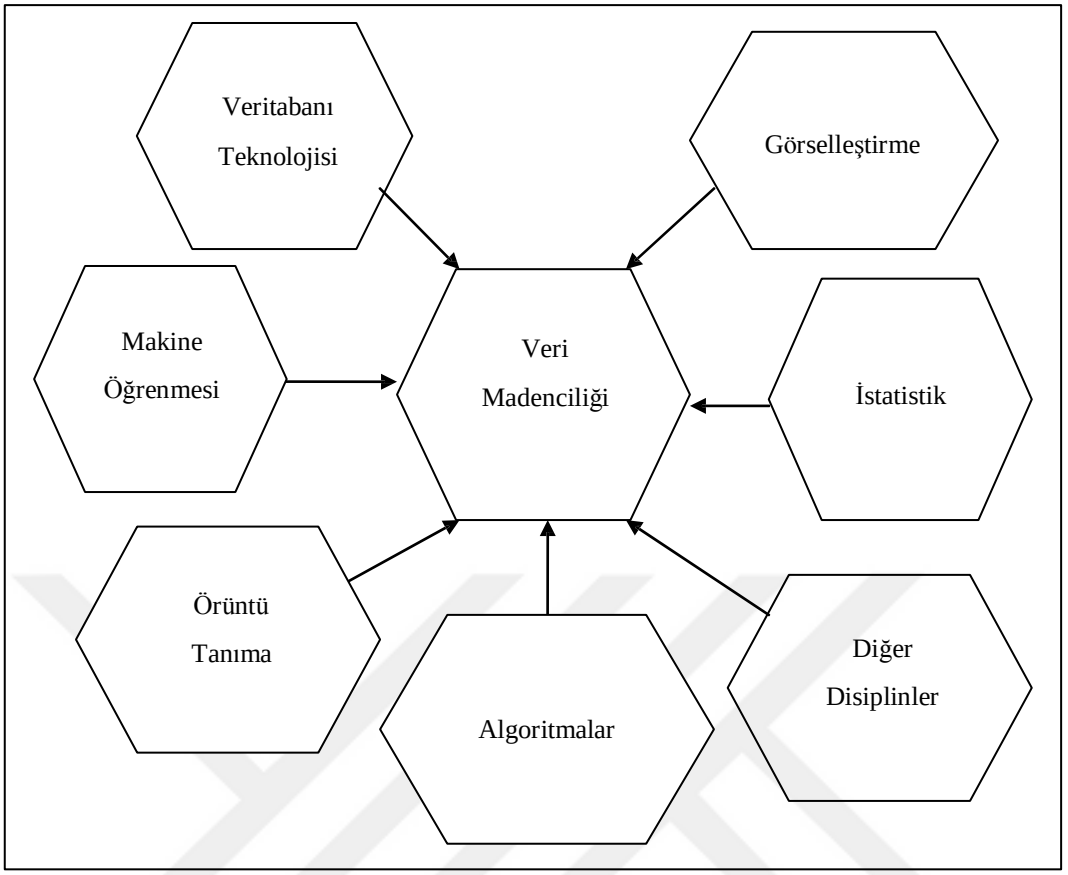
Uzun yıllar süren bir araştırma ve ürün geliştirme sonucunda veri madenciliği ortaya çıkmıştır (Küçüksille, 2009: 30). Veri madenciliği gelişiminin başlangıç serüveni 1950'li yıllarda bilgisayarların sayımlar için kullanılması ile başlamıştır (Savaş, 2012: 5). Daha sonra 1950'li yıllarda matematikçiler veri madenciliği teknikleri konusu üzerinde çalışmalar yapmış, istatistikçiler ise 1960'lı yıllarda regresyon analizi, sinir ağları ve en büyük olabilirlik kestirim algoritmaları gibi bazı algoritmalar üzerinde çalışmalar gerçekleştirmişlerdir (Can vd., 2012: 2). Veri tabanı yönetim sistemleri 1980'li yıllarda yaygınlaşmış, bu yıllarda şirketler müşteri, ürün ve rakip firmalardan elde ettikleri verilerinden veri tabanları oluşturmuş ve bu veri tabanlarına erişebilmek için SQL sorgulama dili kullanmışlardır (Savaş, 2012: 5). 1990'lı yıllara gelindiğinde ise veri tabanında bilgi keşfinin başlangıç aşamaları oluşturulmuş ve veri ambarları geliştirilmiştir (Can vd., 2012: 2). Yine 1990 yılından beri veri depolama araçları, barkod ve

RFID teknolojileri ile birlikte gelişen veri madenciliğinin kullanım alanları oldukça yaygınlaşmıştır. (Silahtaroğlu, 2016: 12). 2000'li yıllara gelindiğinde ise veri madenciliği tüm alanlar için kullanılmaya başlanmıştır. Tablo 4.'de veri madenciliğinin tarihsel gelişimi verilmiştir.

Tablo 4. Veri madenciliğinin Tarihsel Gelişimi (Savaş, 2012: 6)

1950'ler	• İlk Bilgisayarlar (Sayım İçin)
1960'lar	• Veri Tabanı ve verilerin depolanması • Perseptonlar
1970'ler	• İlişkisel Veri Tabanı Yönetim Sitemleri • Basit kurallara dayanan uzman sistemler ve Makine öğrenimi
1980'ler	• Büyük miktarda veri içeren veri tabanları • SQL sorgu dili •
1990'lar	• Veri Tabanlarında Bilgi Keşfi Çalışma Grubu ve Sonuç Bildirgesi • Veri madenciliği için ilk yazılım
2000'ler	• Tüm alanlar için veri madenciliği uygulamaları

Günümüzde ise, veri madenciliği algoritmaları veritabanlarında veri depolanabilen birçok alanda (Sağlık, Pazarlama, Borsa vb.) farklı amaçlar için yaygın olarak kullanılmış ve uygulama alanı oldukça genişlemiştir. Şekil 1.'de yer alan disiplinlerden esinlenilerek veri madenciliği için çeşitli yaklaşımlar geliştirilmiş (Akpınar, 2004'den aktaran: Darakçı, 2011: 18), böylece verilerden elde edilen bilgilere ulaşmak daha kolay bir hal almıştır.



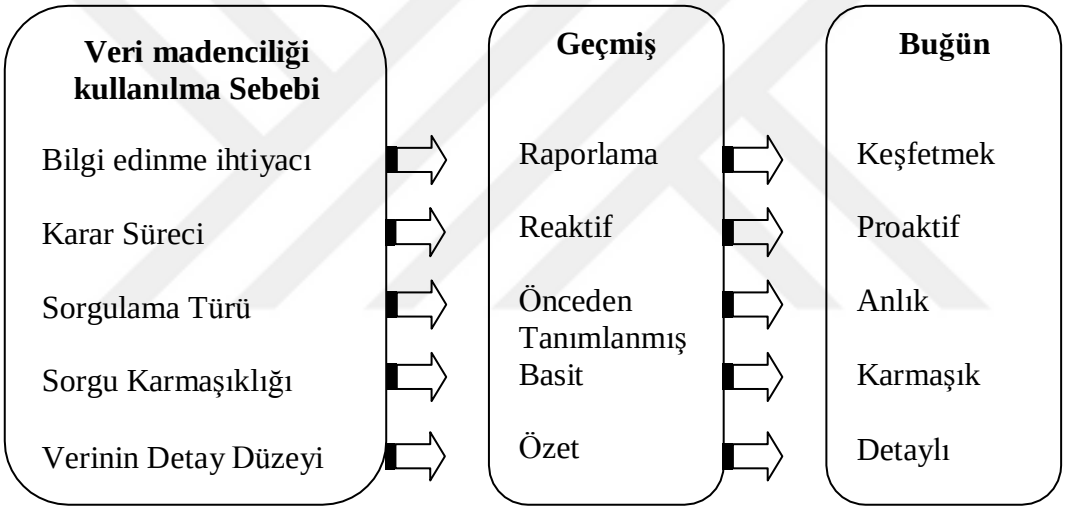
Şekil 1. Veri Madenciliğine Katkıda Bulunan Disiplinler (Dönmez, 2008: 19)

Hızlı değişen ekonomik ve yaşam koşulları nedeniyle deneyim ve önsezilere dayanarak alınan kararların yanlış olma ihtimali oldukça yüksektir. Bu ihtimali minimuma indirmek için bilgiye dayalı yönetimi öngören karar destek sistemlerini kullanmak gerekir. Bu noktada veri madenciliği teknikleri ile karar destek sistemi oluşturulurken bilgi teknolojilerinden yararlanmak olmazsa olmazlar arasındadır.

2.1.4 Veri Madenciliğinin Önemi

Son zamanlarda teknolojinin gelişmesi ile bilgiye ulaşmak daha kolay ve ucuz bir hal almış ve veriler günden güne giderek artmıştır. Bu durum dolayısıyla veri tabanlarında depolama alanlarının artmasına sebep olmuştur. Depolama alanlarının sürekli artması verilerin daha kapsamlı bir şekilde tutulmasına ve saklanmasına olanak sağlamıştır. Ancak geçmişte depolama alanlarının bu kadar

büyük olmamasından dolayı veriler özet şeklinde tutulduğundan veriler arasındaki bağıntılar basit bir şekilde keşfedebilmekte ve sonuçlardan çıkarımlar yapılabilmekteydi. Günümüzde ise kapsamlı bir şekilde veri tabanlarında tutulan büyük hacimli veriler arasından gizli kalmış yararlı bilgilerin elde edilmesi geçmişe kıyasla değişkenlerin daha fazla olması sebebiyle oldukça karmaşıklaşmıştır. Bu durumda veriler arasındaki ilişki ve örüntülerin ortaya çıkarılması için bilgisayar algoritmalarının kullanılması ihtiyacı doğmuştur. Burada büyük veriler arasından yararlı bilgileri elde etme süreci olarak nitelendirilen veri madenciliği algoritmaları devreye girmiştir. Geçmişten günümüze ihtiyaç duyulan bilginin önemine göre veri madenciliği çıktılarının yapısı da değişiklik göstermiştir (Dönmez, 2008: 20; Şekeroğlu, 2010: 16).



Şekil 2. Veri Madenciliğinin Önemi (Şekeroğlu, 2010: 16)

Şekil 2'de görüldüğü üzere veri madenciliğinin geçmişte elde ettiği bilgiler bir olay veya durum sonucunda gerçekleşirken, günümüze gelindiğinde ise bilgiler olay veya durum gerçekleşmeden önce elde edilmektedir. Bu durum ilgili kişi, kurum ve kuruluşların olay gerçekleşmeden önce risk karşısında önlem alabilmesine veya farklı hamleler yapabilmesine olanak sunmaktadır. Bu durumun birçok alanda ve konuda avantaj sağlayacağı oldukça açıktır. Örneğin; Şirketlerin, rakiplerine karşı rekabet üstünlüğü sağlayabileceği gibi hastalıkların erken teşhisi ile tedavi imkanı sağlanabilir. Ancak bu durumun doğru bir şekilde ve sürekli olarak gerçekleştirilebilmesi için bilgi akışının devamlı olarak

sağlanabilmesi ve teknolojik gelişmelere adapte olunması gerekmektedir (Şekeroğlu, 2010: 16).

2.1.5. Veri Madenciliği Kullanım Alanları

Veri madenciliği, birçok probleme uygun çözümler sunabilecek yöntemleri içerisinde barındırdığından neredeyse tüm alanlarda yaygın bir şekilde kullanılmaktadır. Günümüzde geniş bir şekilde veri madenciliği algoritmalarının kullanıldığı alanlar ve bu alanlardaki kullanım amaçları şöyledir (Cihan, 2009: 6-8).

Pazarlama alanında

- Müşteri segmentasyonu
- Mevcut müşterilerin elde tutulması
- Yeni müşterilerin kazanılması
- Pazar sepeti analizi
- Çapraz satış analizleri ve satış tahminleri
- Müşteri değerlendirme ve müşteri ilişkileri yönetimi

Bankacılık alanında

- Farklı finansal göstergeler arasındaki gizli ilişkilerin bulunması
- Kredi kartı dolandırıcılıklarının tespiti
- Kredi taleplerinin değerlendirilmesi
- Usulsüzlük tespiti
- Riskli müşterilerin belirlenmesi
- Kart harcamalarına göre müşteri gruplarının belirlenmesi

Sigortacılık alanında

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespiti
- Riskli müşteri tipinin belirlenmesi

Perakendecilik alanında

- Satış noktası veri analizleri

- Alış-veriş sepeti analizleri
- Tedarik ve mağaza yerleşim optimizasyonu

Borsa alanında

- Hisse senedi fiyat tahmini
- Genel piyasa analizleri
- Alım-satım stratejilerinin optimizasyonu

Telekomünikasyon alanında

- Kalite ve iyileştirme analizleri
- Abonelik tespitleri
- Hatların yoğunluk tahminleri

Sağlık ve Tıp alanında

- Hastalığın erken teşhisi
- Tedavi sürecinin belirlenmesi
- Test sonuçlarının tahmini
- Hasta ve ilaç kullanımın profilinin belirlenmesi
- Ürün geliştirme

Endüstri alanında

- Kalite kontrol analizleri
- Lojistik
- Üretim süreçlerinin optimizasyonu

Şekil 5.'de toplam 552 kişinin katıldığı 2016 yılında veri madenciliğinin sektörel bazda kullanılmasına ilişkin bir araştırmanın sonuçları bulunmaktadır (Anonim, 2016'dan aktaran: Yetginler, 2019: 10).

Tablo 5. Veri madenciliğinin uygulandığı alanlar (Yeginler, 2019: 11)

Sektör	**Sektör Bazında Veri Madenciliği Kullanım Yüzdesi**
Finans	%15
Bankacılık	%13,4
Biyoteknoloji/ Genom	%5,8
Reklamcılık	%12
CRM	%16,3
Sigorta	%9,2
Sağlık Hizmetleri	%12
E-Ticaret	%8,9
Bilim	%12
Perakende	%10,3
Dolandırıcılık Algılama	%11,1
Telekomünikasyon	%8,3
BT/ Ağ Altyapısı	%7.2
Tıp/ İlaç	%6,5
Devlet/ Askeri	%5,6
Önemsiz E-Mail/ Anti Spam	%1,1
Doğrudan Pazarlama	%4,3
Sosyal Medya	%8,3
Diğer	%6,3

2.1.6. Veri Madenciliğinde Karşılaşılan Problemler

Veritabanlarından elde edilen ham veriyi veri madenciliği girdi olarak kullanır. Bu durumda veri tabanlarının eksiksiz, dinamik, geniş ve net veri içermemesinden dolayı veri madenciliğinde problemler yaşanmaktadır (Aydoğan, 2003'den aktaran: Cihan, 2009: 8-10). Veri madenciliğinde karşılaşılan problemler genel hatları ile 2.1.6.1, 2.1.6.2, 2.1.6.3 ve 2.1.6.4. başlıkları altında yer verilmiştir.

2.1.6.1. Sınırlı Bilgi

Genel olarak veri tabanları belirli özellik ve nitelikleri sunmak gibi amaçlarla geliştirilmişlerdir. Bu nedenle, çalışmada öğrenme ve çıkarım işlemlerinin daha kolay bir şekilde gerçekleştirilmesini sağlayabilecek özellikler veri tabanlarında yer almayabilir.

2.1.6.2. Veritabanı Boyutu

Son zamanlarda artan veriler ile birlikte veri tabanı boyutları da hızlı bir şekilde artmıştır. Bu çok büyük boyuttaki verileri işleyebilmek için veri tabanı algoritmaları geliştirilmiş ve veriler ayrı ayrı incelenebilmiştir. Araştırmacılar tarafından çalışmada kullanılacak veri miktarının büyük çapta olması, tahminlerin doğruluğu bakımından avantaj sağlasa da algoritma hatalarına neden olabilir. Bu yüzden algoritmaların büyük ölçekli veriler için kullanımı aşamasında gerekli önem verilmeli ve dikkatli olunmalıdır.

2.1.6.3. Aykırı veri

Veri madenciliğinde yaşanan bir diğer problem de gürültülü verilerden kaynaklanmaktadır. Kullanıcıların yanlış veri girişi yapması sonucunda veya girilen değerlerin hatalı ölçülmesinden dolayı ortaya çıkmış olabilir. Sonuçların güvenilir olması için gürültülü verilerin az olması istenir. Ayrıca gürültülü veriler geleceğe yönelik yapılan tahminlerin veya çıkarımların doğruluğunun azalmasına yol açar. Bu nedenle gürültülü verilerin teşhis edilmesi ve bu problemin ortadan kaldırılması gerekmektedir. Bu amaçla gürültülerin teşhisinde anormali tespiti metotları, kümeleme analizi, regresyon ve histogram yöntemleri

kullanılabilmektedir. Gürültülü verilerin belirlenmesinden sonra bu veriler yerine anlamlı ve aşırı uç noktalarda olmayan veriler kullanılmalıdır.

2.1.6.4. Eksik Veri

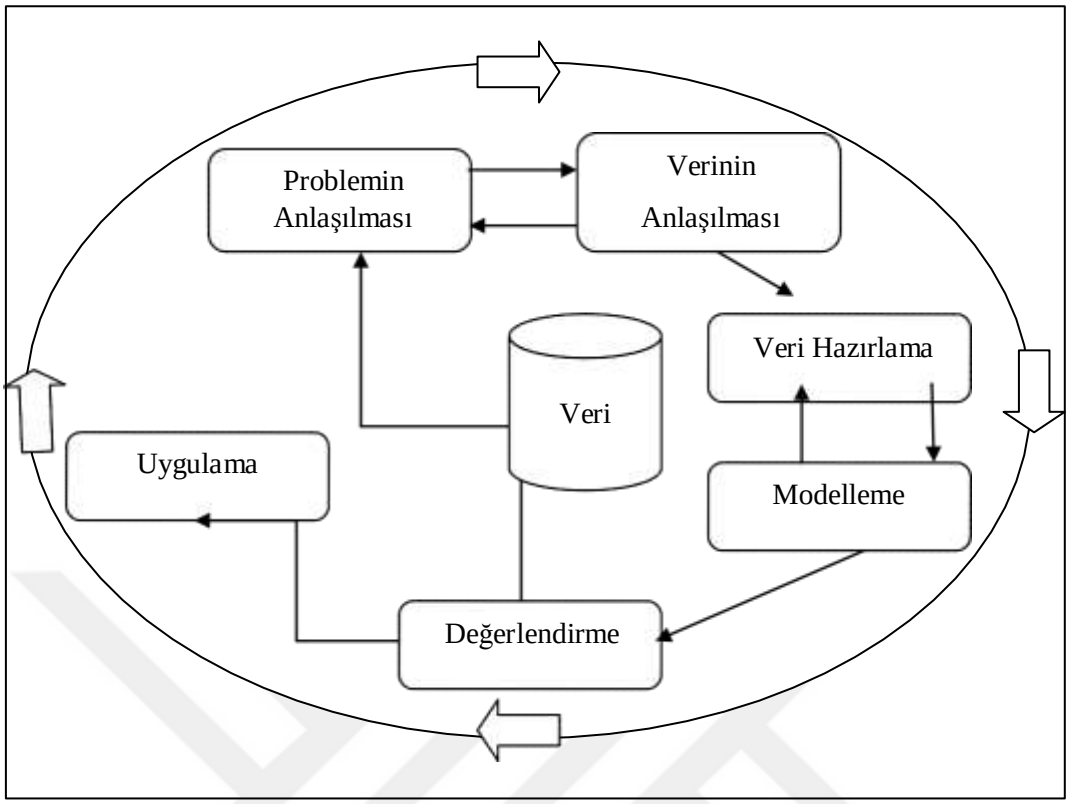
Veriler kayıt edilirken gerekli olabilecek bazı veriler kayıt edilmemiş olabilir olabilir veya veriye bazı bilgiler mevcut olmayabilir. Bu gibi eksik veri içeren kayıtların bulunması durumda iki yol izlenebilir. Birincisi, eksik veri içeren kayıtlar çalışmadan çıkarılabilir. İkincisi ise değişkenlerde yer alan verilerde yer alan değerlerin ortalaması eksik verinin yerine kullanılabilir.

2.2. VERİ MADENCİLİĞİ SÜRECİ

Veri madenciliği, çeşitli veritabanlarında bulunan milyonlarca düzensiz veri arasından anlamlı bilgilerin keşfedilmesi ve veriler arasında gizli kalmış ilişki ve örüntülerin ortaya çıkarılması sürecidir

Veri madenciliğine gelişi güzel yaklaşmak ve onu bir süreç halinde değerlendirmemek işlemlerin sürekli tekrar etmesine neden olabilir. Bu nedenle veri madenciliği sürecini ifade etmek için standart bir süreç modeli kullanmak gerekmektedir. Bunun için DaimlerChrysler, SPSS, NCR olmak üzere üç firmayı temsil eden analistler tarafından Çapraz Endüstri Standardı Süreci (The Cross Industry Standard Process for Data Mining) olarak adlandırılan CRISP-DM Veri Madenciliği için 1996' da geliştirilmiştir (Larose, 2005: 5).

Projenin yaşam döngüsüne genel anlamda bakış açısı sunmayı amaçlayan CRSP-DM modeli veri madenciliği projesinin yaşam döngüsünü altı aşamaya ayrılmıştır. Bu aşamaların sıraları sabit olmamakla beraber aşamalar arasında ileri geri hareketler yaşanabilmektedir. Ayrıca her aşamanın kendi içerisindeki görevinin yerine getirilmesi gerekmektedir (SPSS, 2000: 10). Şekil 1.'de CRISP-DM sürecinde bulunan aşamalar ve aşamalar arasındaki ilişkiler gösterilmiştir.



Şekil 3. CRISP-DM Modeli (Shearer, 2000'den aktaran Kartal, 2015: 14)

2.2.1 Problemin Anlaşılması

Veri madenciliği çalışmalarında gerekli olan ilk adım problemin iyi tanımlanmasıdır. Bu aşamada projenin amaç ve ihtiyaçlarının üzerine iyice odaklanılmalı ve açıkça ifade edilmelidir. Bu aşama çalışmanın beklenti ve standartlarının belirlendiği aşamadır. Bu yüzden çalışmanın başarılı sonuçlar elde etmesi için projenin dikkatle planlanması gerekmektedir. Ayrıca belirlenen hedeflerin ölçülebilir olması da çalışmanın sonucu üzerinde oldukça etkilidir. Çalışmanın amaç, ihtiyaç ve hedeflerin iyi anlaşılmasına rağmen belirsiz ve kabul edilemeyen sonuçlar elde ediliyorsa işin tanımı yerine veri kalitesi ve verinin tanımı üzerinde durulması gerekmektedir.

2.2.2 Verinin Anlaşılması

Verinin anlaşılması aşamasında ilk adım çalışmada kullanılacak verilerin elde edileceği kaynağın belirlenmesi ve toplanmasıdır. Eğer araştırmacı kullanacağı veriler hakkında yeterince bilgiye ve donanıma sahip değilse konu ile

ilgili uzman kişi tarafından veriler hakkında gerekli bilgileri edinmelidir. Elde edilen veriler doğru bir şekilde anlaşılmamışsa doğru modellerin kurulması da mümkün olmayacaktır. Buna bağlı olarak elde edilen sonuçlarda oldukça düşük gerçekleşecek ve proje başarısız bulunacaktır. Bu yüzden elde edilen veriler hakkında bilgi sahibi olmak ve verileri tanımak çalışmanın seyri için oldukça önemlidir. Ayrıca bu aşama; verinin kalitesine ilişkin sorunların tanımlanması, veriye dair ilk izlenimlerin edinilmesi, hipotezlerin geliştirilmesi gibi amaca hizmet eden faaliyetleri içermektedir. Sürecin bu aşamasında gerçekleşebilecek bir hata ya da problem sürecin diğer aşamaların tekrar edilmesine neden olabilir.

2.2.3. Veri Hazırlama

Veri tabanlarında milyonlarca hatta milyarlarca veri bulunmakta, bunların veri tabanına nasıl aktarıldığı çoğu zaman araştırmacılar tarafından bilinmemektedir. Bu nedenle veri tabanlarında kontrolsüz bir şekilde bulunan bu verilerden analiz yoluyla elde edilen sonuçların doğruluğu şüphe altındadır. Çünkü veri tabanına bilgi girişi ya da aktarımı sırasında eksik, fazla ve tekrar edilen veriler girilmiş olabilir. Ayrıca ilişkili veri tabanlarında bulunan aynı kayıtlar farklı değişken isimleri altında kaydedilebilir. Dolayısıyla veri hazırlama aşaması çok önemli bir ön hazırlık aşamasıdır. Araştırmacı zamanının neredeyse %60 -%75'ini bu aşamada harcamaktadır (Şentürk, 2006: 11).

Çalışmamızda veri hazırlama aşaması veri temizleme, veri birleştirme ve veri dönüştürme olmak üzere üç başlık altında incelenmektedir.

2.2.3.1. Veri Temizleme

Bu aşamada, veri setinde bulunan eksik, gürültülü ve tutarsız verilerin düzeltme işlemleri gerçekleştirilmektedir. Bunu sağlamak için bir takım yöntemlerle eksik veriler tamamlanmaya çalışılır. Gürültüye neden olan veriler belirlenir ve gürültünün en aza indirilmesi için çaba gösterilir. Diğer yandan tutarsız verilerin düzeltilmesi veya kaldırmasıyla veri kalitesinin artırılması hedeflenir.

Veri setinde bulunan bazı değişkenlerin değerleri yok ise eksik veri söz konusudur. Bu noktada eksik değerin yer aldığı sınıf için aşağıdaki işlemler uygulanabilir.

1. Eksik değer içeren kayıt veya kayıtlar veri setinden çıkarılabilir.
2. İlgili değişkenin ortalaması eksik değer için tercih kullanılabilir
3. Aynı sınıfta yer alan tüm örnekler için değişkenin ortalama kullanılabilir.
4. Eksik değerin yerine en olası değer kullanılabilir (Han ve Kamber, 2001: 109)

Hatalı veri toplama araçları, veri girişi problemi, bir değerin hatalı olarak ölçülmesi, hatalı nitelik değerleri verilerde gürültünün kaynağını oluşturabilir. Bu probleminin giderilmesinde ise, demetleme yöntemi, kümeleme yöntemi ve regresyon kullanılmaktadır (Şentürk, 2006: 12).

2.2.3.2. Veri Birleştirme

Çalışmamızda kullanılacak verilerin çeşitli veri tabanlarından elde edilerek bir çatı altında toplanması işlemi birleştirme olarak adlandırılır. Bu işlem sonucunda farklı veri tabanlarından elde edilen veriler arasında uyumsuzluk olabilir. Bu tür uyumsuzluklara, verilerin farklı zamanlarda toplanması, değişkenlerin farklı biçimlerde kodlanması (Bir veri tabanında Nüks değişkeni Pozitif/ Negatif şeklinde kodlanırken, başka bir veri tabanında ise 1/0 şeklinde kodlanması), veri formatların farklı olması, güncelleme problemleri neden olabilir. Dolayısıyla farklı veri tabanlarından elde edilen veriler arasında bu tür uyumsuzluklar söz konusu ise veriler mümkün olduğunca tek bir formatta olacak şekilde düzenlenmelidir.

2.2.3.3. Veri Dönüştürme

Kullanılan veri madenciliği algoritmaları, teknikleri veya modelleri belirli tipteki verilerle çalışıp, bazıları ile çalışamamaktadır (Şekeroğlu, 2010: 16) Bu yüzden veri setinde bulunan veriler, veri dönüştürme işlemleri kullanılarak veri madenciliği algoritmaları için uygun biçime getirilmelidir. Veri dönüştürme ile verilerin kalitesini arttırmak için veriler, hedefe uygun olarak seçilir, veri

yapılandırma işlemleri gerçekleştirilir ve temizleme işlemleri yapılır (Yetginler, 2019: 16) Bu işlemler için dönüştürme türlerinden düzeltme, birleştirme genelleştirme ve normalleştirme sıklıkla kullanılmaktadır. Veri madenciliği için veriler uygun bir hale dönüştürülürken bu türlerden bir veyahut birkaçı uygulanabilmektedir (Şentürk, 2006: 15).

2.2.4. Modelleme

Veri setinin ön işlemlerden geçirilmesi ile son halini alan veri setine çeşitli veri madenciliği modelleri uygulanır. Çalışma da kurabilecek kadar çok sayıda model kurulup denenerek en iyi veri madenciliği modeli bulunabilir. Verilerin, bazı modeller için sürekli farklı işlemlerden geçirilmesi gerekebilir. Dolayısıyla en iyi veri madenciliği modelini bulabilmek için veri hazırlama ve model kurma aşamaları sürekli tekrar edilebilir (Şekeroğlu, 2010: 16). En uygun modele ulaşana kadar süreci dört kısımda incelemek mümkündür.

1. Model tekniği seçmek: Çalışmada uygulanacak algoritma/fonksiyonların belirlendiği, modele yönelik fikrin oluştuğu kısımdır.

2. Modelin test tasarımını yapmak: Modelin kalite ve geçerliliğinin model uygulanmadan önce test edildiği kısımdır. Çalışmada kullanılan fonksiyonun türüne göre doğruluk ve hata oranı gibi çeşitli kalite göstergeleri tercih edilebilir. Olabildiğince çok modeli deneyerek kalite göstergelerinin karşılaştırılması ile çalışmanın hedeflerine ulaştıracak en uygun model belirlenebilir.

3. Modeli kurmak: Veri üzerinde seçilen algoritmanın/fonksiyonun uygulandığı kısımdır.

4. Modeli değerlendirmek: Teknik değerlendirmelerin hedef ve tecrübe gibi başarı kriterlerine dayanarak yapıldığı kısımdır (Argüden ve Erşahin 2002'den aktaran: Yetginler, 2019: 16-17)

2.2.5 Değerlendirme

En iyi model ve modeller analiz için modelleme aşamasında kurulmuştur. Değerlendirme aşamasında kurulan modellerin çalışmanın amaç ve hedeflerine uygun olup olmadığı incelenir, gerçekleştirilen adımlar tekrardan gözden geçirilir. Her model için geliştirilen birçok performans değerlendirme yöntemi ve

performans değerlendirme ölçütleri bulunmaktadır. Çalışmada kullanılan bu ölçüt ve yöntemlerle bu aşamada kurulan model ya da modeller en ufak ayrıntısına kadar değerlendirilir. Son olarak çalışmanın ulaştığı noktanın yeterliliği incelenerek ek çalışmalara karar verilir.

2.2.6 Uygulama

Kurulan veri madenciliği modellerinin elde ettiği çıktılarla süreç sonlanmamaktadır. Çünkü ilerleyen zamanla birlikte sistemlerde değişiklikler oluşabilir. Bu durum kurulan modelle elde edilen çıktıların geçersiz olmasına yol açabilir. Bu nedenle kurulan modeller sürekli takip edilmeli ve bu modeller değişikler göz önünde bulundurularak güncel tutulmalıdır. Son olarak analiz sonucunda elde edilen çıktılar, kazanımlar, projenin değerlendirilmesi, hedefler ve sonuçların kıyaslanması kullanıcıların yararlanabileceği bir şekilde rapor halinde sunulması gerekmektedir.

2.3. VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri tabanlarında bulunan büyük ölçekteki verilerin bilgiye dönüştürülmesi aşamasında veri madenciliği teknikleri kullanılır. Çalışmada kullanılan teknikler araştırmada hedeflenen çıktılara bağlı olarak değişiklik gösterebilir. Hedeflenen çıktılara ulaşılabilmesi için veri madenciliğinde birçok algoritma ve yöntem bulunmaktadır. Kullanılan yöntemlerin bir bölümü uzun süredir kullanılan klasik istatistiksel yöntemlerdir. Diğer bir bölümü ise yapay zeka ve makine öğrenme destekli yöntemlerdir. Bu yöntemler, araştırmacılar tarafından çeşitli başlıklar altında incelenmiştir.

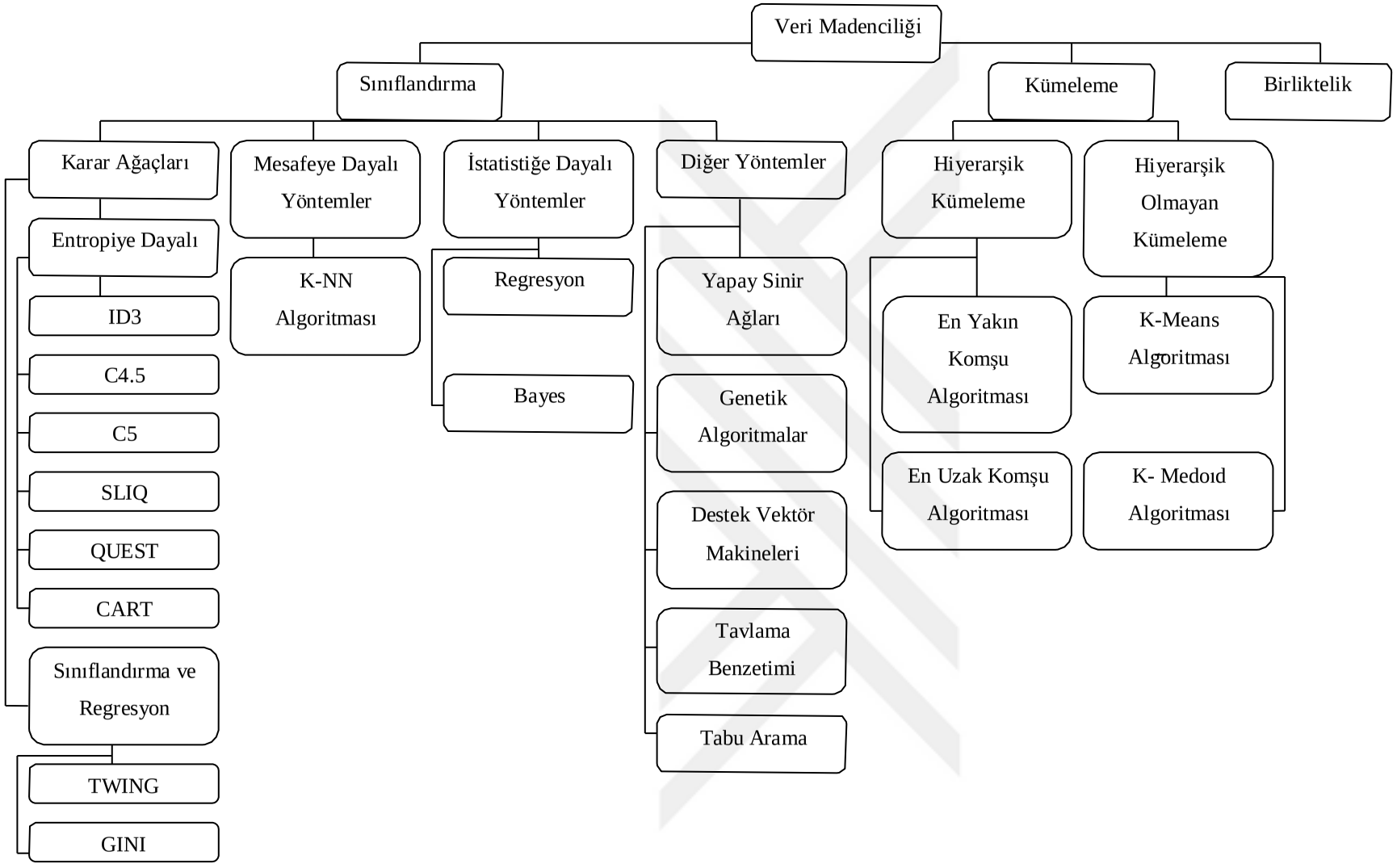
Özkeleş (2003) yaptığı çalışmada veri madenciliği yöntemlerini üç ana başlık altında ifade etmiştir.(Savaş, vd., 2012: 14)

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

Berkey ve Linoff, 2004 yılındaki çalışmalarında ve Larose, 2014 yılındaki çalışmasında veri madenciliği altı başlıkta incelemişlerdir.

1. Tanımlayıcı (descriptive)
2. Tahmin (estimation)
3. Öngörü (prediction)
4. Sınıflandırma (classification)
5. Kümeleme (clustering)
6. Birliktelik (association)

Ayrıca Gürbüz ve arkadaşları (2008) veri madenciliği yöntemlerini tanımlayıcı ve tamamlayıcı olmak üzere iki başlık altında ifade edilmiştir. Çalışmamızda ise veri madenciliği yöntemleri Sınıflama, Kümeleme ve Birliktelik olmak üzere üç başlık altında incelenmiş, bu yöntemler hakkında kısaca bilgiler verilmiştir.



Şekil 4. Veri Madenciliği Yöntemleri (Terzi, 2019: 3)

Şekil 3'de görülen bir çok veri madenciliği algoritmaları, çeşitli alanlarda bulunan verilerden faydalanarak sonuçlar elde etmektedir. Veri madenciliği algoritmalarının bu verilerin üzerinde kullanılması için verilerin belirli bir yapı da olması gerekmektedir. Bir çok veri gibi çalışmamızda kullanılan veriler de genellikle Tablo 6'daki bir yapıda yer almaktadır.

Tablo 6. Veri Setinin yapısı.

	a_1	a_2	a_3	.	.	a_p
Örnek1: X_1	x_{11}	x_{12}	.	.	.	x_{1p}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.		.	.	.	.
.	.	.	.	.	.	.
Örnek n: X_n	x_{n1}	x_{n2}	.	.	.	x_{np}

V={x1 , x2 ,……., xn} : Veri setinde bulunan örnekleri,

A={ a1 , a2 ,….., ap } : Veri setindeki özellikleri (öznitelik ve değişkenleri) ifade etmektedir.

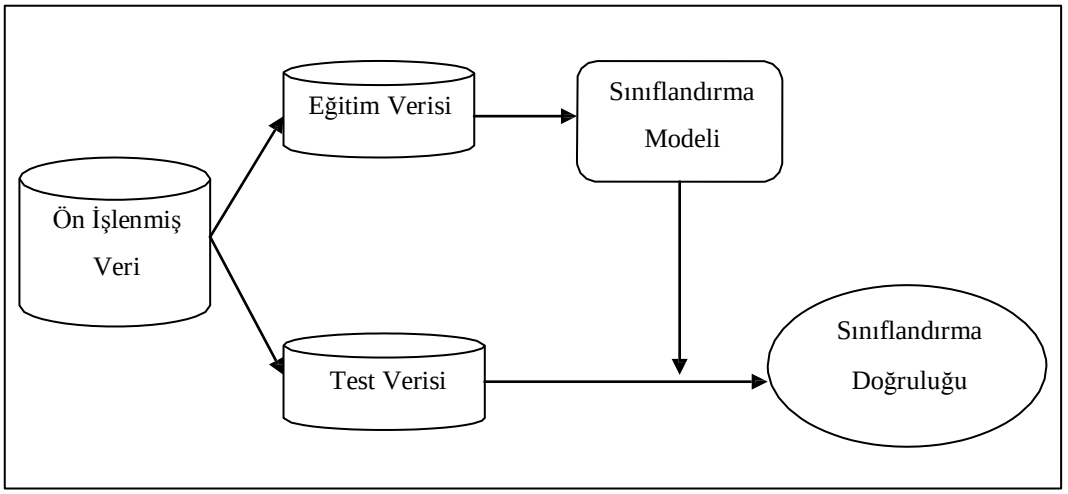
2.3.1. Sınıflandırma Yöntemi

Sınıflandırma yöntemi, yeni bir kaydın veya nesnenin daha önce belirlenmiş kurallar ile nasıl sınıflandırması gerektiğini belirler. Bunun için sınıflandırma kuralını oluşturan üç adımın takip edilmesi gerekir (Yılmaz, 2006'dan aktaran: Küçüksille, 2009: 38). Bu adımlar şöyledir:

Adım 1: Veri seti üzerinden rastgele seçilen bir kısım veri, eğitim verisi olarak belirlenir.

Adım 2: Sınıflandırma için eğitim veri seti sonucunda elde edilen model kullanılır. Modelin tahmin gücü bulunan modele dayanarak belirlenir. Test veri seti de yine tüm veri seti içerisinden rastgele seçilir.

Adım 3: Eğer modelin doğruluğu test veri seti üzerinde gerçekleştirilen denemeler neticesinde yeterli bulunursa, model yeni sonuçları bulmak için kullanılır.



Şekil 5. Sınıflandırma Yöntemlerindeki Akış (Olson & Delen, 2008'den aktaran: Balaban & Kartal, 2019: 142)

Denetimli öğrenme yöntemlerinin uygulanabilmesi için veriler uygun bir yapıda olmalı ve hedef değişken sınıfı belirlenmelidir. Denetimli öğrenme yöntemlerinin uygulanabileceği bir veri seti yapısı Tablo 7'de gösterilmiştir.

Tablo 7. Denetimli Öğrenme İçin Veri Setinin Yapısı

	a_1	a_2	a_3	.	.	a_p	Sınıf
	x_{11}	x_{12}	.	.	.	x_{1p}	y_1
Örnek1: X_1	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.		.	.	.	.	.
.	.	.	.	.	.	.	.
Örnek n: Xn	x_{n1}	x_{n2}	.	.	.	x_{np}	y_n

veri setinde bulunan örnekleri (Kayıtları ve Nesneleri),

veri setinde yer alan özellikleri (Öznitelik ve Değişkenleri),

veri setinde belirlenen hedef değişkeni için sınıflar kümesini ifade etmektedir.

Denetimli öğrenme işlemleri yapılırken $yi \in C = \{c1 , c2 , ..., ck\}$ olmak üzere, her yeni kayıt için atanacak olan sınıf kategorisi $c \in C$ belirlenir.

2.3.2 Kümeleme Yöntemi

Verinin içerisindeki benzer özellikler baz alınarak grupların oluşturulması işlemine kümeleme denir. Bu işlemle benzer özellikli nesneler bir araya getirilerek farklı özelliklerde grupların oluşturulması amaçlanmıştır. Oluşturulan her bir kümede yer alan veriler, içinde bulunduğu kümenin verileriyle daha fazla benzer özellik taşırken, ayrı kümelerdeki verilerle benzerlikleri daha az olmalıdır (Aydemir, 2017: 39).

Kümeleme ve sınıflandırma yöntemleri verileri gruplara ayırmak için kullanılmaktadır. Ancak kümeleme yöntemi ile sınıflandırma yöntemi arasında önemli bir fark bulunmaktadır. Sınıflandırma yönteminde gruplar daha önceden bilinmektedir. Kümeleme yönteminde ise gruplar belli olmamakla birlikte veriler benzerlikleri neticesinde gruplara ayrılmaktadır (Aydemir, 2017: 39; Yetginler, 2019: 21). Hatta Ramkumar Swami'nin (1998) dediği gibi bazı çalışmalarda sınıflandırma modelinin önişleminde kümeleme modeli kullanılabilmektedir.

Kümeleme yöntemi son yıllarda veri tabanlarında depolanan veri miktarında ciddi bir artışın yaşanması ile birlikte aktif olarak kullanılan bir yöntem haline gelmiştir (Uyan ve Çay, 2008'den aktaran: Küçüksille, 2009: 36). Bu yöntem, marketler de farklı müşteri gruplarının belirlenmesinden şehir planlaması için ev tiplerinin gruplandırılmasına kadar birçok alanda farklı konular için kullanılabilmektedir.

2.3.3 Birliktelik Kuralı

Birliktelik kuralı, veri madenciliği analizlerinde kullanılan önemli yöntemlerden birisidir. Veri ambarlarındaki değişkenler ve veriler arasındaki gizli

kalmış ilişkilerin ortaya çıkarılması ile sık görülen kayıtların belirlenmesi bu yöntemin temelini oluşturmaktadır. Bu yöntem pazar sepeti analizi uygulamalarında yaygın olarak kullanılmaktadır (Şentürk, 2006: 19).

Birliktelik kuralında, güven ve destek kriterleri ile veri veya öğeler arasındaki bağıntılar hesaplanmaktadır. Destek kriteri, veri setinde bulunan nesne veya öğeler arasındaki bağıntının ne kadar sık olduğunu ifade etmektedir. Güven kriteri ise A nesnesi veya öğesinin hangi olasılıkla B nesnesi veya öğesi ile beraber olduğunu gösterir. Veri setinde bulunan nesne veya öğelerin birlikteliğinin önem arz etmesi için destek ve güven kriterlerinin yüksek olması beklenmektedir (Şentürk, 2006: 19).

Destek ve güven değerleri A ön koşul, B bağlı koşul olmak üzere aşağıdaki şekilde ifade edilir:

Destek $(A \rightarrow B) = P (A \cup B)$

Güven $(A \rightarrow B) = P (A \cap B)$

Örneğin yerel bir market zinciri, şarküteri ürünleri için müşteri satın alma davranışlarını belirlemek istiyor. Buna göre olası birliktelik kuralları aşağıdaki şekilde oluşuyor;

1. Müşteri yumurta satın alır ise sucuk da satın almaktadır.
2. Müşteri beyaz peynir ve zeytin satın alır ise vişne reçeli de satın almaktadır.

Birinci birliktelik kuralında müşteri yumurta satın aldığında sucukta satın almaktadır. Bu birliktelik için güven değeri yumurta satın alındığında sucuk almanın koşullu olasılığına bağlıdır. Bu örnek için destek ve güven değeri eğer yüksek olarak gerçekleşmiş olursa işletme müşterin bu satın alma davranışlarına göre raf düzenlemelerini tekrar gözden geçirebilir ve müşterilerin ürünlere kolayca ulaşabilmesini sağlayabilir.

2.4. KULLANILAN SINIFLANDIRMA ALGORİTMALARI

Bölüm 1.2'de yapılan literatür araştırması sonucunda en çok tercih edilen ve popüler olan sınıflandırma algoritmaları çalışmada kullanılmıştır. Bunlar; Naive Bayes, Destek Vektör Makinesi, Lojistik Regesyon, J48 ve K-NN algoritmalarıdır.

2.4.1. Naive Bayes Algoritması

İstatiksel bir sınıflandırıcı olan Naive Bayes, elde bulunan sınıflandırılmış verilerden yola çıkarak yeni bir verinin, hali hazırdaki sınıflardan hangisinde yer alacağının olasılıklarını öngören bir algoritmadır. Ayrıca belirsizliğin olduğu durumlarda karar verebilmek için Bayes algoritması oldukça kullanışlıdır (Han vd., 2011). Naive Bayes veya Bayes sınıflandırıcı kavramı aşağıdaki şekilde ifade edilir. (Han, vd., 2012: 350-353);

X kümesini sınıfını bilmediğimiz bir veri kümesi olarak varsayalım. X= $\{x_1, x_2 \ldots x_n\}$ öznitelik değerleri olsun. $C_1, C_2 \ldots C_n$ sınıf değerleri olarak kabul edelim. Sınıfı belirlenecek olan X veri kümesinin olasılıkları formül (2.1) ile bulunur.

——————— (2.1)

$P(X\backslash C_i)$ olasılığında işlem yükünü azaltmak için basitleştirme yapılır. değerlerinin birbirinden bağımsız olduğu varsayılarak aşağıdaki bağıntı kullanılır;

(2.2)

Bilinmeyen *X'i* sınıflandırmak için (2.1)'de $P(C_i\backslash X)$ içinde bulunan paydalar birbirine eşit olduğu için yalnızca paylar karşılaştırılır. Bu değerler içinde yer alan en büyük değer seçilir ve sınıfı belirli olmayan örneğin bu sınıfa ait olduğu belirlenir.

(2.3)

Sonrasal olasılığı kullanan (2.3)'deki ifade en büyük sonrasal sınıflandırma yöntemi olarak bilinir. En son durum da Naive Bayes olarak şu ifade kullanılır;

(2.4)

2.4.2. J48 (C4.5) Karar Ağacı Algoritması

C4.5 karar ağacı algoritması, WEKA paket programında J48 algoritması olarak kullanılmaktadır. En çok bilinen karar ağaç modellerinden birisi olan C4.5 algoritması, ID3 algoritmasının temeline dayanmaktadır. Ancak bu algoritmanın eksiklerini gidermek için geliştirilmiştir. Bu yöntemde bilgi kazancı öznitelik seçim ölçütü olarak kullanılmakta, en yüksek bilgi kazancına sahip öznitelik, her bir kayıt için seçilmektedir (Quinlan, 1986'dan aktaran: Onan, 2014: 12).

C4.5 algoritması, ID3 algoritmasında izlenen adımlara benzer bir şekilde çalışır. Ancak karar ağaçlarının oluşturulması sırasında ID3 algoritması kayıp verileri hesaba katarken C4.5 algoritması bu verileri hesaba katmaz. C4.5 algoritması verileri eksik olmayan kayıtları kazanım oranını hesaplarken kullanılır. Kayıp verileri diğer veriler ve değişkenler sayesinde öğrenerek kazanım hesaplamasında kullanır. C4.5 algoritmasının, ID3 algoritmasından diğer bir farkı da sayısal değer bulunduran özniteliklerle karar ağacı oluşturabilir (Silahtaroğlu, 2016: 80).

Eğitim için veritabanından elde edilen kayıt kümesinin incelendiğini varsayalım. Sınıf niteliğinin alacağı değerlere göre eğitim kümesinin
olmak üzere k sınıfa ayrıldığını düşünelim. Bu durumda sınıfların ortalama bilgi miktarına ihtiyaç duyulacaktır. Aşağıda yer alan formül *T*, sınıf değerlerini içeren küme için sınıfların olasılık dağılımını ifade eder ve hesaplanış şekli şöyledir;

— — — (2.5)

kümesindeki elemanların sayısını | | ifadesi vermektedir. — burada olasılığı ifade eder. Bu durumda T için ortalama entropi 2.6'da bulunan formül ile bulunur.

(2.6)

Sınıf niteliği göz önüde bulundurularak her öznitelik için ağırlıklı ortalamalar hesaplanır ve bu hesaplamada 2.7'de bulunan formül kullanılır.

(2.7)

—

Aşağıda yer alan formül ile her özbitelik için bilgi kazancı hesaplanır.

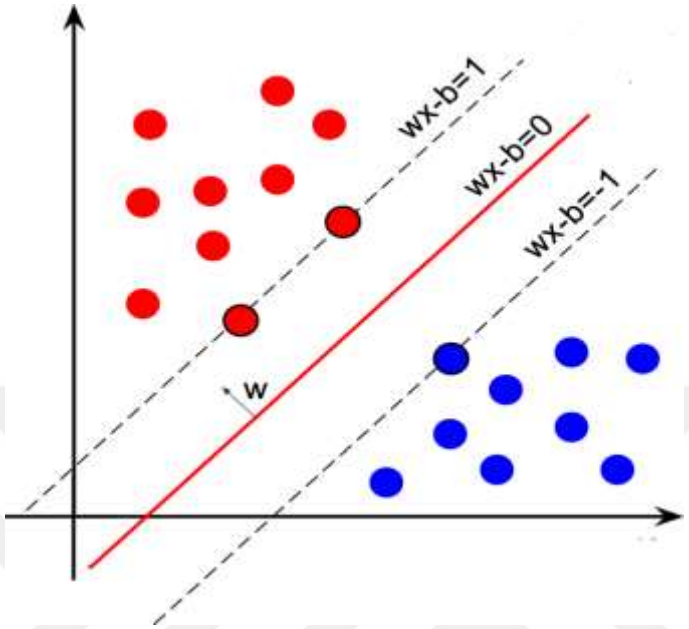
Bilgi Kazancı$(X,T) = H(T) - H(X,T)$ (2.8)

Nitelik ayrımı yüksek bilgi kazancı sağlayan nitelikte gerçekleşir. Böylece karar ağacının ilk düğümü oluşur. Bundan sonra tüm dallar için bir önceki düğümden sonra veri kümesi tekrar düzenlenir ve işleme tabi tutulur. Her işlem sonucunda verideki satır sayısında azalma yaşanır ve satır bitinceye kadar işlem devam eder (Özkan, 2016: 48-49).

2.4.3. Destek Vektör Makinesi

Vapnik ve ekibi tarafından 1990'lı yıllarda geliştirilen Destek Vektör Makinesi, iki temel sınıfa ait olan verileri birbirinden ayırmak için uygulanan, sınıflandırma ve regresyon işlemleri için kullanılabilen, istatistiksel öğrenme teorisine dayalı gözetimli makine öğrenme algoritmasıdır (Cortes & Vapnik 1995'den aktaran: Sevli, 2019: 178). Destek Vektör Makinesi, eğitim verisinde bulunan farklı sınıflara ait üyeleri birbirinden ayırmaya ve bu ayrım sonucunda

oluşan sınıflara en uzak mesafede yer alacak hiper düzlemi bulmaya çalışır. Bu hiper düzlem aşağıdaki gibi tanımlanır (Yücebaş, 2018: 19).



Şekil 6. Destek Vektör Makinesi

D: Veri Kümesi, *:* Sınıf belirteci ve *:* Veri noktasını gösteren p boyutlu reel vektör iken verilerin her bir noktası Formül 2.9'da yer alan biçimde tanımlanır.

∈ ∈ (2.9)

Bu durumda bir özellik uzatımdaki hiper düzlem şu şekilde ifade edilebilir:

(2.10)

(2.11)

İki düzlem arasındaki uzaklığın hesaplanmasında;

—— (2.12)

Burada amaç marjı yani ── ifadesini maksimum yapmaktır. Bunun için ifadesinin minimize edilmesi gerekmektedir. Eldeki sınıflama probleminin doğrusal olarak sınıflara ayrılamadığı durumda elde bulunan öznitelik uzayı daha yüksek boyutlara çıkarılmalıdır. Çekirdek yöntemi bu durumda kullanılır (Baudat vd., 2001'den aktaran: Yücebaş, 2018: 19). Böylece X' in n boyutlu bir uzayda bir vektör ϕ(.)'nin ise girdi uzayından yüksek boyutlu bir uzaya eşleme gerçekleştiren doğrusal olmayan bir fonksiyon olduğu kabul gördüğünde sınıflandırma kararını çizecek hiper düzlem şu şekilde tanımlanır (Yücebaş, 2018: 19).

(2.13)

Yukarıda yer alan eşitlikte eğitim verisini yüksek boyutlu uzaya eşleyecek ağırlık katsayısı vektörünü w temsil etmektedir. Nokta çarpımı yerine çekirdek fonksiyonu kullanılabilir, böylece girdi vektörünün yüksek boyutlu uzaya eşlenmesi durumu ortadan kaldırılabilir (Yücebaş, 2018: 19). Dönüştürülmüş uzaydaki iç çarpımlar orijinal nitelik verisinden yararlanılarak hesaplandığı için bu benzerlik fonksiyonu K ile gösterilen "çekirdek fonksiyon" olarak isimlendirilir (Uçar, 2013: 41).

(2.14)

Destek Vektör makinesinde dört çekirdek fonksiyonu yaygın olarak kullanılmaktadır. Bunlar: Doğrusal Fonksiyon, Polinomiyal Fonksiyon, Sigmoid Fonksiyon ve Radyal Tabanlı Fonksiyondur. Aşağıda bu çekirdek fonksiyonların açıklamaları ve formülleri ile birlikte yer almıştır.

Doğrusal Fonksiyon: Veriler girdi uzayında doğrusal olarak ayrılabiliyor ise veriyi yüksek boyuta taşımadan doğrusal çekirdek fonksiyonu vasıtasıyla sınıflama işlemi gerçekleşir. Herhangi bir boyut değeri ya da katsayı bu fonksiyonda yer almamaktadır (Uçar, 2013: 42).

(2.15)

Polinomiyal Fonksiyon: Belirli bir derecede girdi vektörlerinin iç çarpımından polinomiyal çekirdek fonksiyon oluşmaktadır. Fonksiyonun matematiksel olarak ifade ediliş şekli aşağıdaki gibidir:

(2.16)

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

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

(2.17)

Radyal Temelli Fonksiyon (RBF): Çekirdek fonksiyonlar arasında kullanımı en yaygın fonksiyon Radyal tabanlı fonksiyondur. Radyal tabanlı fonksiyona göre R programında sistem standart çıktıları vermektedir. Formülasyonu:

(2.18)

2.4.4. Lojistik Regresyon

Son zamanlarda ekonomi, tıp, biyoloji, tarım gibi birçok alanda yaygın olarak kullanılan Lojistik Regresyon analizi veri setinde bulunan grup sayısı ve verilerden yararlanarak ayrımsama modeli elde etmektedir. Daha sonra veri kümesine alınan yeni gözlemlerin gruplara ataması kurulan bu model yardımı ile yapılmaktadır. İstatistikte kullanılan model yapılandırma teknikleri gibi Lojistik Regresyon analizi de en az değişken kullanarak en iyi uyuma sahip olacak biçimde bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen bir model kurmayı amaçlamaktadır (Bircan, 2004: 186).

Lojistik regresyon analizi, hedef değişkenin ölçüldüğü ölçek türüne ve hedef değişkenin seçenek sayısına göre üçe ayrılmaktadır. Hedef değişken iki seçenekli bir kategorik değişkene sahip ise "İkili Lojistik Regresyon Analizi (Binary Logistic Regression Analysis)" adını alır. Eğer hedef değişken ikiden daha fazla kategorili sınıflı bir değişkene sahip ise "Çok Kategorili/Düzeyli İsimsel Lojistik Regresyon Analizi (Multinominal Logistic Regression Analysis)" adını alır. Eğer hedef değişken sıralama ölçeğiyle elde edilmiş ise, bu durumda da "Sıralı Lojistik Regresyon Analizi (Ordinal Logistic Regression Analysis)" şeklinde tanımlanır (Çokluk, 2004: 1363).

Çok kategorili lojistik regresyonda gerçekleşen işlemlerle ikili lojistik regresyondaki işlemler benzerlik göstermektedir. Bu nedenle hesaplamalara temel oluşturması ve daha iyi kavranabilmesi için tek bağımsız değişken içeren ikili lojistik regresyona ait bazı gösterimlere aşağıda yer verilmiştir (Karabulut & Alpar, 2011'den aktaran: Kartal, 2015: 23):

olasılığı ile bağımsız değişkenleri arasındaki ilişki aşağıda yer alan lojistik fonksiyon ile gösterilmektedir.

(2.19)

Üstünlük (*odds*), Bir olayın olma olasılığının, olmama olasılığına oranı şeklinde ifade edilmektedir.

(2.20)

Ulutürk Akman'ın (2004) değindiği gibi hedef değişken yalnızca iki değer aldığından parametrelerin tahminleri en küçük kareler yöntemi ile gerçekleştirilememektedir. Bu sebeble modeli doğrusal hale getirmek için (2.21)'de bulunan formül ile logit dönüşüm uygulanarak tahminde kolaylık sağlanmaktadır.

_____ (2.21)

Hosmer, ve arkadaşları (2013) tarafından çok kategorili lojistik regresyon şu şekilde özetlenmiştir:

Gösterimlerin basit ve anlaşılabilir olabilmesi için hedef değişkenin =0, =1 ve =2 gibi üç farklı kategoriden oluştuğunu varsayalım. adet bağımsız değişkenimiz olsun. İkili lojistik regresyon analizinde tek bir logit fonksiyon kullanılırken, bu durumda çok kategorili lojistik regresyonda iki logit fonksiyona ihtiyaç vardır. Hedef değişkenin kategorilerinden birisi referans kategorisi olarak seçilmekte ve diğer kategoriler ile kıyaslaması yapılmaktadır.

Bu durumda logit fonksiyonlar;

_____ (2.22)

_____ (2.23)

Her kategoriye ait koşullu olasılık değerleri;

_____ (2.24)

_____ (2.25)

_____ (2.26)

Hedef değişkenin değeri yerine 0, 1 ve 2 olmak üzere üç adet ikili nitelik biçiminde ifade edilerek olabilirlik fonksiyonu oluşturulmaktadır. ’dir.

Bu durumda adet gözlemden oluşan bir örneklemin *koşullu olabilirlik fonksiyonu*;

(2.27)

Σ 2 =0=1 fonksiyonunun doğal logaritmasının alınmasıyla elde edilen *log-olabilirlik fonksiyonu;*

(2.28)

Olabilirlik fonksiyonları ( )’nin birinci kısmi türevleri alınarak bulunmaktadır. =1,2; =0,1,2,…, ve olmak üzere:

(2.29)

Bu denklemler sıfıra eşitlenerek çözülmesiyle maksimum olabilirlik tahmincisi elde edilmektedir. Eğer ki çözüm bulunuyorsa bulunan her çözüm maksimum ya da minimum olabilecek bir kritik noktayı göstermektedir. Eğer kritik nokta maksimumsa ikinci kısmi türevler negatif tanımlıdır. (Czepiel, 2002: 6-7). Aynı zamanda bu matrisin diğer bir özelliği de tahminlerin varyans-kovaryans matrisini oluşturur.

(2.30)

Logistik regresyon modelinin uyum iyiliğini belirlemek için geliştirilmiş yöntemlerden biri olan *Hosmer-Lemeshow* ile uyum iyiliği testi onlu risk grupları ve sabit kesim noktası yöntemine göre hesaplanmaktadır (Bircan, 2004: 198) Modelin uyum iyiliği onlu risk grupları ile hesaplanmak istendiğinde *Hosmer-Lemeshow* istatistiği kullanılmaktadır.

(2.31)

Kestirilen modelin uyum iyiliği testi sabit kesim noktası yöntemiyle hesaplanmak istendiğinde ise, Hosmer-Lemeshow istatistiği kullanılmaktadır.

(2.32)

2.4.5. K-En Yakın Komşu Algoritması

Doğrusal denetimli örüntü tanımaya dayalı K-NN algoritması, 1951 yılında Fix ve Hodges tarafından tanımlanmıştır (Lu ve Zhu, 2014: 276). K-NN algoritması 1960'lı yıllara kadar büyük ölçekli veri setlerinde yavaş çalışması nedeniyle popülerlik kazanamamıştır. 1960'lı yıllara gelindiğinde ise bilgisayarların işlem yeteneğinin artması ile birlikte o yıllardan günümüze gelene kadar sıklıkla kullanılan bir algoritma olmuştur (Witten vd., 2011'den aktaran: Balaban & Kartal, 2019: 14). Bu algoritmayı kavramsal olarak Han ve arkadaşları (2012) şu şekilde aktarmıştır:

- Veri setinde yer alan örnekler *n* sayıda nitelik tarafından tanımlanır.
- Eğitim verisi *n* boyutlu örnek uzayında yer alır. Her bir örnek bu *n* boyutlu uzayda bir noktayı temsil eder.

- K-NN algoritması, sınıfı bilinmeyen bir örnek ile karşılaştığında, ilgili örneğe en yakın *k* tane eğitim örneğini belirleyerek en yakın örüntü uzayını arar.

Veri setinde bulunan satırlar örnekleri, sütunlar da nitelikleri ifade etmektedir. Buna göre K-NN tanımı kapsamında, her bir örnek ile n sayıdaki nitelik aşağıdaki gibi ifade edilebilir.

(2.33)

(2.34)

Sınıfı belirli olmayan örnek için, belirlenen k tane komşuya olan uzaklığın hesaplanmasında çeşitli uzaklık ölçütleri bulunmaktadır. Uzaklıkların kullanımı verinin sayısal veya kategorik olma durumuna göre değişiklik gösterir. Sayısal veriler için Öklid, Minkowski, Manhattan gibi uzaklık hesaplamaları kullanılırken, kategorik veriler için Gower uzaklığından faydalanılır (Kartal, 2015: 21)

Sınıflandırma ve kümeleme algoritmalarında en çok tercih edilen uzaklıklardan birisi olan Öklid uzaklığı Denklem 2.35'e göre hesaplanır (Balaban & Kartal, 2019: 145)

(2.35)

Minkowski uzaklığı Denklem (2.36)'e göre hesaplanır (Kreesse & Danko, 2012: 133)

(2.36)

Manhattan Uzaklığı ise Denklem (2.37)'e göre hesaplanır (Kreesse & Danko, 2012: 133).

(2.37)

Gower benzerlik ölçüsü iki gözlem arasındaki benzerliği bulmak için geliştirilmiştir. Benzerlik ölçüsünün hesabı formül (2.38) ile yapılmaktadır.

(2.38)

Kaufman ve Rouseeuw'a (1990) göre Benzerlik ölçüsü 1'den çıkarılarak Gower uzaklık ölçüsü formül (2.39) ile bulunabilir.

(2.39)

ÜÇÜNCÜ BÖLÜM

MATERYAL VE YÖNTEM

3.1. ARAŞTIRMANIN AMACI VE ÖNEMİ

3.1.1. Araştırmanın Amacı

Bu çalışma ile yumurtalık kanser teşhisi konulan hastalara ait veri seti için belirlenen tedaviye yanıt, sağ kalım ve lokalizasyon hedef değişkenlerinin sınıflandırılmasına en çok katkı sağlayan tanımlayıcı değişkenlerin belirlenmesi ve hedef değişkenlerine uygulanan veri madenciliği algoritmalarına ait sınıflandırma performanslarının karşılaştırılması amaçlanmıştır.

3.1.2.Araştırmanın Önemi

Sağlık alanında yapılan çalışmalar çoğunlukla tanımsal istatistikler üzerinden yürütülmektedir. Bu durum elde edilen bilgilerin sınırlı bir düzeyde kalmasına sebep olmaktadır. Ancak elde edilecek her anlamlı ve doğru bilgi, insan hayatı için oldukça önemli olup, bu verilerin birçok yönden incelenmesi gerekmektedir. Son yıllarda veri madenciliği yöntemlerinin sağlık alanında kullanılması ile bu alanda bulunan veriler farklı açılardan birçok yöntemle incelenerek; hastalıkların tahmin edilmesi, tedavi süreçlerinin belirlenmesi ve hastalığa yol açan önemli özelliklerin ortaya çıkarılması sağlanmıştır. Bu açıdan bakıldığında veri madenciliği sınıflandırma algoritmaları kullanılarak yumurtalık kanseri hastalarından elde edilen veri setinin farklı açılardan birçok yöntemle analiz edilmesi çalışmamızın önemini göstermektedir. Ayrıca yumurtalık kanseri hastalarına ait değişkenlerin analiz edilmesi sonucunda elde edilen bulguların, sağlık çalışanlarına ve bu alanda çalışma yapacak olan akademisyenlere katkı sağlayacağı düşünülmektedir.

3.2. ARAŞTIRMA HAKKINDA GENEL BİLGİLER

3.2.1. Araştırmanın Örneklemi

RTEÜ Eğitim ve Araştırma Hastanesi bilgi işlem merkezinden talep edilen yumurtalık kanseri hastalarına ait veriler gerekli izinlerin alınması ile Excel formatında tarafımıza teslim edilmiştir. Ancak bilgi işlem merkezinde yaşanan sistem değişikliği nedeni ile hastanede yumurtalık kanseri tanısı konulan tüm hastalara ulaşılamamıştır. Bu nedenle Onkoloji bölümünde bulunan arşivler uzman yardımıyla taranmış ve yumurtalık kanseri tanısı konulan hasta dosyaları incelenmiştir. Yaklaşık bir hafta boyunca incelenen hasta dosyaları bilgi işlem merkezinden elde edilen veriler ile birleştirilmiştir. Bu sayede çalışmada kullanılan ana kütle sayısına ulaşılmıştır. Veri seti hastanenin kuruluş tarihi olan 2005 yılından verilerin elde edildiği 2019 yılına kadar olan hastaları kapsamaktadır. Çalışmada bu yıllar içerisinde yumurtalık kanseri tanısı konulan toplam 90 hasta incelenmiştir.

3.2.2. Araştırma Sınırlılıkları

- Çalışmanın Recep Tayyip Erdoğan Üniversitesi Eğitim ve Araştırma Hastanesi'nden elde edilen veriler ile yürütülmesi,
- Çalışmada kanser türlerinden yalnızca yumurtalık kanseri hastalarına ait verilerin incelenmesi,
- Veri setinde bulunan örnek sayısının az olması,
- Hedef değişkeni olarak yalnızca tedaviye yanıt, sağ kalım ve lokalizasyon değişkenlerinin incelenmesi,
- Hastalardan elde edilen verilerin analizinde ise sadece veri madenciliği sınıflandırma yönteminin kullanılması çalışmanın sınırlılıklarını oluşturmaktadır.

3.2.3 Etik Kurul izni

Bu çalışmada kullanılan veri setinin elde edilmesi için etik kurul izni, Recep Tayyip Erdoğan Üniversitesi Bilimsel Araştırmalar Etik Kurulundan alınmıştır (27.02.2019-31.12.2019).

3.2.4. Veri Toplama Aracı

Hastalara ait bilgiler Recep Tayyip Erdoğan Üniversitesi Eğitim ve Araştırma hastanesi bilgi işlem merkezi veri tabanı ve arşiv taraması sonucunda elde edilmiştir.

3.3.VERİ TOPLAMA VE HAZIRLAMA SÜRECİ

Çalışmada kullanılmak üzere oluşturulan veri seti bilgi işlem merkezi ve arşiv taraması ile elde edilmiştir. Bilgi işlem merkezinden talep edilen yumurtalık kanseri hastalarına ait veriler gerekli izinlerin alınması itibari ile Excel formatında tarafımıza teslim edilmiştir. Ancak bilgi işlem merkezinden alınan veri setinde satır ve sütun bilgilerinin karmaşık olduğu, analize uygun bir hale getirilmesi için düzenlemeler yapılması gerektiği görülmüştür. Ayrıca bilgi işlem merkezinde yaşanan sistem değişikliği nedeni ile hastanede yumurtalık tanısı konulan tüm hastalara ulaşılamamıştır. Bu nedenle Onkoloji bölümünde bulunan yumurtalık kanseri hastalarına ait dosyalarının taranması gerekmiştir. Ancak arşiv dosyaları taranmadan önce bilgi işlem merkezinden elde edilen veri seti değişkenler sütunda, hastalara ait bilgiler ise satırda yer alacak şekilde düzenlenmiştir. Daha sonra arşivde bulunan hasta dosyaları, veri seti bütünlüğünü sağlamak için düzenlenen Excel dosyası formatında taranmıştır. Hasta dosyalarının taranması sonucunda elde edilen veriler Bölüm 2.3.3.2'de bahsedildiği üzere bilgi işlem merkezi veri tabanından alınan veriler ile birleştirilmiştir.

Yumurtalık tanısı konulan kadın hastalara ait veriler 2005 yılından 2019 yılına kadar olan hastaları kapsamaktadır. Veri setinde, başka hastanede yumurtalık kanseri tanısı konulan ancak tedavisine RTEÜ Eğitim ve Araştırma hastanesinde devam etmekte olan hastalar da bulunmaktadır. Ancak tedavisine başka bir hastanede devam etmek üzere dosyasını alan hastalar bilgilerine ulaşılamaması sebebi ile çalışmaya dahil edilememiştir. Ayrıca yeni tanı konulan hasta dosyaların da çoğu bilginin yer almaması sebebi ile bu hastalara çalışmada yer verilmemiştir.

Kullanılan veri setinde, hasta kimliğinin açığa çıkmasına neden olabilecek bilgiler (tc, ad, soyad, yaş, iletişim bilgileri vb.) çalışma dışında bırakılmıştır.

Ayrıca veri setinde tanı ca, son ca, hastaların nüks sonrası kemoterapi alıp almadığı ve alınan kemoterapi tedavilerine ait bilgiler de bulunmaktadır. Ancak bu bilgiler, çok sayıda eksik veri barındırması nedeni ile çalışmada yer almamıştır. Sağ kalım bilgilerinin elde edilmesinde ise veri setinde bulunan tanı tarihi, ex tarihi ve son kontrol tarihleri kullanılmıştır.

Ayrıca veri setimize ait tanımlayıcı değişkenlerde 5 adet kayıp değer gözlenmiş, bu kayıp değerler Bölüm 2.3.3.1'deki bilgilerden yola çıkarak bulundukları özellik grubuna göre en çok tekrar eden değerlerle tamamlanmıştır. Değişkenlerimizin kategorik olması sebebi ile kayıp verilerin tamamlanmasında en çok tekrar eden değerler kullanılmıştır.

3.4. VERİYİ ANLAMA

Çalışmada kullanılan veri seti RTEÜ Eğitim ve Araştırma Hastanesi'nde bulunan yumurtalık kanseri hastalarından elde edilmiştir. Veri tabanı ve arşiv dosyalarının taranması sonucu elde edilen düzensiz veriler Bölüm 2.2'de bahsedildiği üzere uzman görüşüne başvurularak düzenlenmiştir. Düzenlenen veri setine ait bilgiler doğrultusunda yumurtalık kanseri üzerinde etkili olabileceği düşülen değişkenler de yine uzman görüşüne başvurularak belirlenmiştir. Bu işlemlerin ardından çalışmada kullanılacak veri seti son halini almıştır.

3.4.1. Yumurtalık Kanseri Veri Seti İçin Kullanılan Değişkenler

Bu çalışmada kullanılan yumurtalık kanseri veri seti uzman görüşüne başvurularak oluşturulmuş, veri hakkında gerekli bilgiler alınmıştır. Ayrıca yumurtalık kanseri ile ilgili çalışmalar da incelenmiştir. Böylelikle disiplinler arasındaki farklılıklardan oluşabilecek bilgi eksikliği giderilmeye çalışılmıştır.

Uzman görüşüne başvurularak yumurtalık kanseri üzerinde etkili olabileceği düşünülen Tanı yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümörün Patolojisi, Kemoterapi, Nüks, Tedaviye Yanıt, Sağ Kalım ve Tümörün Lokalizasyonu olmak üzeri 10 değişken belirlenmiştir. Tedaviye Yanıt, Sağ Kalım ve Tümörün Lokalizasyonu hedef değişken olarak seçilmiştir. Diğer değişkenler ise hedef değişkenlerin sınıflandırılmasında kullanılacak tanımlayıcı değişkenler

olarak belirlenmiştir. Aşağıda, çalışmamızda kullanılan veri setine ait, değişkenler hakkında kısaca açıklamalar yapılmıştır.

- *Tanı Yaşı:* Hastanın yumurtalık kanseri teşhisi konulduğu tarihteki yaşı.
- *Başvuru Sebebi:* Hastalık tanısı konulmadan önce hastaların acil servislere başvuruda bulundukları şikayetler.
- *Debulking:* Hastaların tanı sonrası cerrahi bir operasyon geçirip, geçirmediğini ifade eder.
- *Kemoterapi:* Hastaların tanı sonrası kemoterapi tedavisi alıp, almadığını gösterir.
- *Tümör Patolojisi:* Yumurtalık kanseri tanısındaki tümör tiplerini ifade eder.
- *Tümörün Evresi:* Hastaların yumurtalık kanserinde kaçıncı evrede olduğunu belirtir.
- *Nüks:* Tedavide başarı gösteren hastalarda, hastalığın yeniden ortaya çıkıp çıkmadığını ifade eder.
- *Tedaviye Yanıt:* Hastaların uygulanan tedavi sonucunda tedaviye verdikleri yanıtı ifade eder.
- *Sağ Kalım:* Tanı tarihinden, hastanın ölüm tarihi ya da son muayene tarihine kadar geçen süreyi belirtir.
- *Tümörün Lokalizasyonu:* Tümörün hangi lokalizasyonda yer aldığını ifade eder.

3.4.2. Değişkenlerin Frekans Değerleri ve Yüzdeleri

Aşağıda yer alan tablolarda çalışmamızda kullanılan veri setine ait değişkenlerin frekans değerleri ve yüzdeleri verilmiştir.

Tablo 8. Hastaların Tanı Yaşına Göre Dağılımları

Tanı Yaşı	Frekans (f)	Yüzde (%)
51 yaş ve altı	29	32
52-62 yaş	29	32
63 yaş ve üstü	32	36
Toplam	90	100

Tablo 8'de yumurtalık kanseri teşhisi konulan hastaların tanı yaşları bulunmaktadır. Buna göre hastaların %32'si 51 yaş ve altında, %32'si 52-62 yaşları arasında ve %36'sı ise 63 yaş ve üstü yaşlarda bu hastalığa yakalanmıştır.

Tablo 9. Hastaların Başvuru Sebeplerine Göre Dağılımları

Başvuru Sebebi	Frekans (f)	Yüzde (%)
Karında Şişlik	29	32
Karın Ağrısı	23	26
Karında Şişlik ve Karın Ağrısı	9	10
Diğer	29	32
Toplam	90	100

Tablo 9'da hastaların teşhis konulmadan önce hastaneye başvuru sebepleri görülmektedir. Buna göre hastaların %32'si karında şişkinlik, %26'sı karın ağrısı, %10'u karında şişkinlik ve karın ağrısı ve %32'si diğer sebeplerden dolayı hastaneye müracaat etmişlerdir.

Tablo 10. Hastaların Debulking Dağılımları

Debulking	Frekans (f)	Yüzde (%)
Pozitif	69	77
Negatif	21	23
Toplam	90	100

Tablo 10'a bakıldığında hastaların %77'sinde debulking pozitif, %23'ünde ise debulking negatif olarak görülmektedir. Veri seti incelendiğinde debulking negatif olan hastaların hepsinde tümör evresinin Evre IV olduğu görülmüştür. Evre IV olan hastalarda kanser neredeyse vücudun tüm organlarına yayılmaktadır. Bu nedenle bazı durumlarda Evre IV hastaları için debulking yapılması mümkün olmamaktadır.

Tablo 11. Hastaların Kemoterapi Tedavisine Göre Dağılımları

Kemoterapi	Frekans (f)	Yüzde (%)
Aldı	79	88
Almadı	11	12
Toplam	90	100

Tablo 11. İncelendiğinde hastaların %88'inin kemoterapi tedavisi aldığı, %12'sinin ise kemoterapi tedavisi almadığı görülmektedir. Veri seti incelendiğinde Evre II ve Evre III hastaların tamamı kemoterapi tedavisi almaktadır. Kemoterapi almayan hastalar Evre I ve Evre IV hastalarından oluşmaktadır. Bunun nedeni olarak erken evre olan Evre I hastalar için kemoterapiye gereksinim duyulmaması, bazı Evre IV hastaları için kemoterapinin fayda sağlanmayacağının düşülmesi ve hastanın isteği doğrultusunda kemoterapi tedavisinin uygulanmaması gösterilebilir.

Tablo 12. Hastaların Tümörün Patolojisine Göre Dağılımı

Tümörün Patolojisi	Frekans (f)	Yüzde (%)
Seroz Ca	44	49
Musinöz Ca	5	6
Clear Cell	3	3
Diğer	38	42
Toplam	90	100

Tablo 12'de çalışmanın ana kütlesini oluşturan hastalara ait tümör patolojileri görülmektedir. Hastalara ait tümör patolojilerinin neredeyse yarısını Seroz Ca oluşturmaktadır. Gürdal vd. (2007) ve Güzel vd. (2019) çalışmalarında inceledikleri yumurtalık kanserine ait veri seti içerisinde Seroz Ca daha yaygın olarak görülmektedir. Bu durumda mevcut veri setimize ait tümör patoloji dağılımının literatürle uyumlu olduğu görülmektedir.

Tablo 13. Hastaların Tümör Evrelerine Göre Dağılımı

Tümörün Evresi	Frekans (f)	Yüzde (%)
Birinci Evre	27	30
İkinci Evre	2	2
Üçüncü Evre	19	21
Dördüncü Evre	42	47
Toplam	90	100

Tablo 13. İncelendiğinde hastaların %30'u yumurtalık kanserinin birinci evresinde, %2'si kanserin ikinci evresindedir. Yumurtalık kanseri üçüncü evresinde olan hastalar veri setinin %21'ini oluştururken, hastalığın dördüncü evresinde olanlar ise veri setinin %47'sini oluşturmaktadır.

Tablo 14. Hastaların Nüks Olma Durumlarına Göre Dağılımı

Nüks	Frekans (f)	Yüzde (%)
Pozitif	48	53
Negatif	32	36
Progresyon	10	11
Toplam	90	100

Tablo 14'te veri setimizi oluşturan hastaların %53'ünde nüks pozitif olarak gerçekleşmektedir. Geriye kalan hastaların %36'sında nüks negatif, %11'inde ise nüks progresyon olarak görülmektedir.

Tablo 15. Hastaların Tedaviye Yanıtlarına Göre Dağılımı

Tedaviye Yanıt	Frekans (f)	Yüzde (%)
Tam Yanıt	33	37
Kısmi Yanıt	13	14
Progresyon	44	49
Toplam	90	100

Tablo 15'te görüldüğü üzere hastaların tedaviye vermiş oldukları yanıtın %37'si tam yanıt, %14'ü kısmi yanıt, %49'u ise progresyon olarak gerçekleşmiştir. Progresyon oranının yüksek olması başlangıç aşamasında olan Evre I hastaların tedavi sürecinin henüz tamamlanmamış olmasından ve bazı Evre IV hastalar da ise kanserin tedaviye rağmen hızla yayılmaya ve büyümeye devam etmesinden kaynaklandığı düşünülmektedir.

Tablo 16. Hastaların Sağ Kalıma Göre Dağılımı

Sağ Kalım	Frekans (f)	Yüzde (%)
1080 gün ve altı	58	64
1081 gün ve üstü	32	36
Toplam	90	100

Tablo 16'da veri setimizi oluşturan hastaların %64'ü 1080 gün ve altı, %36'sı ise 1081 gün ve üstü sağ kalıma sahiptir.

Tablo 17. Hastaların Tümör Lokalizasyonuna Göre Dağılımı

Tümörün Lokalizasyonu	Frekans (f)	Yüzde (%)
Sol	25	28
Sağ	32	36
Sol ve Sağ	33	36
Toplam	90	100

Hastaların lokalizasyon oranlarının bulunduğu Tablo 17'de hastaların %28'inde tümör lokalizasyonu sol, % 36'sında tümör lokalizasyonu sağ ve geriye kalan %36'lık kısımda ise tümör lokalizasyonu hem sol hem de sağ olarak görülmektedir.

3.5. MODELİN OLUŞTURULMASI

Ön işlemlerinin gerçekleştirilmesinin ardından son halini alan yumurtalık kanseri hastalarına ait veri setine Naive Bayes, Destek Vektör Makinesi (DVM), Lojistik Regresyon Analizi, J48 Karar Ağacı ve k-En Yakın Komşu algoritmaları uygulanmıştır. Çalışmada kullanılan veri madenciliği sınıflandırma algoritmaları tercih edilirken konuyla ilgili çalışmalar incelenmiştir.

Veri setinin analizinde kullanılmak için belirlenen sınıflandırma algoritmalarının performanslarını karşılaştırmak için performans değerlendirme yöntemi olarak 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme ve tabakalı %50/%50, %75/%25, %80/%20 ve %90/%10 Hold-Out kullanılmıştır. K-En Yakın komşu algoritmasının diğer algoritmalarla birlikte performansının karşılaştırılması için diğer algoritmalarda elde edilen en yüksek sınıflandırma performansları dikkate alınmıştır. Bunun yanı sıra K-En Yakın Komşu algoritması için parametre seçiminde k komşu sayısı 1'den 10'a kadar belirlenmiştir.

Böylelikle hedef değişkenlerin sınıflandırılmasında etkili olacağı düşünülen tanımlayıcı değişkenlerin analiz edilmesi ile hem algoritmalarının yumurtalık kanseri üzerindeki sınıflandırma performansları karşılaştırılmış hem de hedef değişkenler sınıflandırma performanslarına göre kıyaslanmıştır. Ayrıca hedef değişkenlerin sınıflandırılmasına en çok katkı sağlayan tanımlayıcı değişkenler J48 karar ağaç algoritması ile çizilen karar ağacı ile belirlenmiştir.

3.6. PERFORMANS DEĞERLENDİRME YÖNTEMİ VE PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ

3.6.1. Performans Değerlendirme Yöntemleri

Modellerin performansını değerlendirmek için birçok yöntem bulunmaktadır. Ancak bu bölümde çalışmada kullanılan performans değerlendirme yöntemleri üzerinde durulmuştur. Bunlar;

Hold-Out: Holdout performans değerlendirme yönteminde elde edilen veri seti ikiye ayrılır. Veri Setinin bir kısmını eğitim veri seti oluştururken diğer kısmını ise test veri seti oluşturmaktadır. Eğitim aşamasında model oluşturulması için eğitim seti, belirlenen sınıflandırıcının eğitiminde kullanılmaktadır. Bu adımda model parametreleri için en iyi değerler ve uygun performans ölçüsü tespit edilir. Oluşturulan modelin genel performansını bulabilmek için ise test veri seti kullanılmaktadır. Kullanılan bu performans değerlendirme yöntemin iki temel dezavantajı bulunmaktadır. Birincisi veri setinde az verinin bulunması ile test veri seti için yeteri kadar örneğin ayrılamayacak olmasıdır. İkinci dezavantaj ise, eğitim ve test veri seti bir defaya özgü olmak üzere birbirinden ayrıldığından, gerçekleşen hata oranında bu ayrımdan ortaya çıkan sorunlar yaşanabilir (Kartal, 2015: 30).

Tabakalı Örnekleme (Stratified Sampling): Bazı veri setlerinde çıktı değeri birer sınıf olabilir bu durumda bazı sınıflara ait örnekler çok az veyahut hiç olmayabilir. Sınıf oranlarının korunması bu gibi durumlarda beklenebilir. Bu sebeple tabakalı örnekleme tercih edilmektedir, ancak çıktı değeri bir sınıf yerine nümerik bir değer olduğunda bu yöntemden yararlanılamaz (Kartal, 2015: 30).

Çapraz Geçerleme (Cross Validation): Çapraz geçerleme iki farklı şekilde uygulanabilmektedir. Bunlardan birincisi k-kat çapraz geçerleme, ikincisi ise birini dışarıda bırak çapraz geçerlemedir. Veri seti k-kat çapraz geçerleme ile k eşit parçaya ayrılır. Ayrılan k parçadan her defasında bir tanesi test için kullanılırken, k-1 tane parça ise eğitim için kullanılmaktadır. Sonuçta ayrılan k parça sayısı kadar hata oranı elde edilmekte ve tüm tahmin hatasını hesaplayabilmek için hataların ortalaması alınmaktadır. Birini dışarıda bırak çapraz geçerleme yönteminde ise, veri setinde her defasında sadece 1 örnek test veri seti olarak kullanılırken, geriye kalan örnekler ise eğitim veri seti olarak kullanılmaktadır. Özellikle veri setindeki örnek sayısının çok az olduğu durumlarda birini dışarıda bırak yönteminden faydalanılmaktadır.

3.6.2. Performans Değerlendirme Ölçüleri

Saptona ve arkadaşları'nın (2018) değindiği gibi model performans değerlendirme ölçütleri, veri madenciliği algoritmalarının oluşturduğu modellerin ne kadar doğru sonuçlar elde ettiğini anlayabilmek için kullanılmaktadır. En iyi performansı sergileyen algoritmaları bulabilmek ve algoritmaların performanslarını karşılaştırmak için de yine bu ölçütlerden yararlanılmaktadır. Model için elde edilen sonuçların başarısının ölçülmesinde ve performanslarının karşılaştırılmasında en sık kullanılan ve kullanılması oldukça basit olan ölçütler doğruluk değeri ve hata oranıdır. Bu ölçütlerin hesaplanması için karmaşıklık matrisi (confusion matrix) kullanılmaktadır. Tablo 1.' de iki sınıflı veri setine ait modelin karmaşıklık matrisi gösterilmiştir.

Tablo 18. Karmaşıklık Matrisi (Confusion Matrix)

	Gerçek Sınıf Değeri		
Tahmin Edilen Sınıf Değeri		Pozitif	Negatif
	Pozitif	Doğru Pozitif (TP)	Yanlış Pozitif (FP)
	Negatif	Yanlış Negatif (FN)	Doğru Negatif (TN)

Doğru Pozitif *(TP):* Gerçek sınıf değeri pozitif olan model tarafından sınıf değeri de pozitif olarak tahmin edilen doğru sınıflandırılmış örneklerin sayısıdır.

Yanlış Pozitif *(FP):* Gerçek sınıf değeri negatif olan model tarafından sınıf değeri pozitif olarak tahmin edilen yanlış sınıflandırılmış örneklerin sayısıdır.

Doğru Negatif *(TN):* Gerçek sınıf değeri negatif olan model tarafından sınıf değeri de negatif olarak tahmin edilen doğru sınıflandırılmış örneklerin sayısıdır.

Yanlış Negatif *(FN):* Gerçek değeri pozitif olan model tarafından sınıf değeri negatif olarak tahmin edilen yanlış sınıflandırılmış örneklerin sayısıdır.

Algoritmaların oluşturduğu modellerin performanslarını değerlendirmek ve karşılaştırmak için doğruluk oranı ve hata oranı dışında çeşitli ölçütlerde bulunmaktadır. Bu ölçütlerin hesaplanmasında da yine karmaşıklık matrisi üzerindeki değerler kullanılmaktadır. Bu ölçütlerin bazılarına aşağıda açıklamaları ile birlikte yer verilmiştir.

Kappa İstatistiği (*Kappa Statistics*)**:** İki kümenin birbiriyle ne derecede uyumlu olduğunu gösterir. Kappa istatistiği sonucunda elde edilen değerler (0-1) arasında yer alır. Kappa istatistiği= 0 ise sınıflandırılmış iki küme arasında hiç uyum bulunmamaktadır. Kappa istatistiği= (0,40-0,59) arasında ise iki küme arasında orta seviyede bir uyum bulunmaktadır. Kappa istatistiği= (0,60-0,79] ise iki küme arasında iyi seviyede bir uyum vardır ve Kappa istatistiği = (0,80-0,99) ise iki küme arasında çok iyi seviyede bir uyum bulunmaktadır. Eğer Kappa istatistiği= 1 ise iki küme birbirleriyle mükemmel bir uyum içerisindedir. (Landis & Koch, 1977).

Doğruluk Oranı: Doğru sınıflandırılmış örnek sayısının (TP+TN) toplam örnek sayısına (TP+FP+TN+FN) oranıdır (Nizam&Akın, 2014: 4).

$$\text{Doğruluk} = TP + TN \,/\, TP + TN + FP + FN \qquad (2.1)$$

Hata Oranı: Yanlış sınıflandırılmış örnek sayısının (FP+FN) toplam örnek sayısına (TP+FP+TN+FN) oranıdır (Nizam&Akın, 2014: 4).

$$\text{Hata Oranı} = FP + FN \,/\, TP + TN + FP + FN \qquad (2.2)$$

Duyarlılık *(Recall):* Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır (Nizam ve Akın, 2014: 4).

Duyarlılık = TP / TP + FP (2.3)

Kesinlik *(Precision):* Doğru sınıflandırılmış pozitif örnek sayısının(TP) sınıfı pozitif olarak tahmin edilmiş toplam pozitif örneklem sayısına (TP+FN) oranıdır. (Şık, 2014: 104).

Kesinlik= TP / TP + FN (2.4)

Yanlış Pozitiflik Oranı (*False Positive Rate*)**:** Yanlış olarak sınıflandırılmış pozitif örneklerin gerçek sınıfı pozitif olan tüm örneklere oranına yanlış pozitiflik oranı denir (Şık, 2014: 105).

Yanlış pozitif oranı = FP / FP + FN =1- Belirleyicilik (2.5)

Yanlış Negatiflik Oranı (*False Negative Rate*)**:** Yanlış olarak sınıflandırılmış negatif örnek sayısının gerçek sınıfı pozitif olan tüm örnek sayısına oranıdır (Kartal, 2015: 34).

Yanlış negatif oranı = FN / FN + TP = 1- Duyarlılık (2.6)

F Ölçütü *(F Measure):* Performansların karşılaştırmasında kesinlik ve duyarlılık ölçütleri tek başına yeterli değildir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar elde etmemizi sağlar. Bu nedenle her iki ölçütün birlikte değerlendirmek için kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması olan F-Ölçütü kullanılır (Nizam ve Akın, 2014: 4).

F-Ölçütü = (2 * Kesinlik * Duyarlılık) / (Kesinlik * Duyarlılık) (2.7)

3.7. MODELİN UYGULAMA GEÇİRİLMESİ

Hedef değişkenlerin sınıflandırılması için veri seti hakkında gerekli bilgiler edinilmiş ve ilgili literatür taramaları yapılmıştır. Daha sonra hedef değişkenlerin sınıflandırılmasında kullanılacak olan tanımlayıcı değişkenler

belirlenmiştir. Tanımlayıcı değişkenlerde eksik veriye rastlanmış, bu eksik veriler yine o değişken sınıfına ait en çok tekrar eden değerler ile tamamlanmıştır. Uzman görüşü ile elde edilen veri seti yine uzman görüşüne başvurularak analize hazır hale getirilmiştir. Bu aşamaların tamamlanmasının ardından hiçbir eksikliğin veya sorunun ortaya çıkmamasıyla modellerin başarım ve performanslarının karşılaştırılması için kullanılacak yöntem ve ölçütler seçilmiştir.

Veri madenciliği sınıflandırma algoritmalarında oluşturulan modellerin analiz sürecinde kullanılması için Bölüm 3.5.1'de bulunan performans değerlendirme yöntemleri ve Bölüm 3.5.2'de yer alan performans değerlendirme ölçütleri belirlenmiştir. Hedef değişkenlere veri madenciliği sınıflandırma algoritmalarının uygulanması ve elde edilen bulgulara Bölüm 4'de yer verilmiştir.

DÖRDÜNCÜ BÖLÜM

YUMURTALIK KANSERİ VERİ SETİNE VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARININ UYGULANMASI

Bu bölümde, yumurtalık kanseri hastalarından elde edilen veri setine Bölüm 2.5' te bahsi geçen sınıflandırma algoritmaları uygulanmış ve her hedef değişken için karar ağacı çizilmiştir. Ayrıca Bölüm 3.5.2'de yer alan performans değerlendirme ölçütleri ile modellerin performans değerleri hesaplanmış ve algoritmalar birbiri ile kıyaslanmıştır. Öncelikle her bir hedef değişkenine, tüm algoritmalar uygulanarak algoritmaların performansları karşılaştırılmıştır. Her bir hedef değişkeni için kullanılan algoritmalar, veri seti, tanımlayıcı değişkenler, performans değerlendirme yöntemi ve parametre seçim bilgisi özet bir tablo biçiminde sunulmuştur. Daha sonra uygulanan her sınıflandırma algoritması için tüm hedef değişkenlerine ait doğru sınıflandırma oranları karşılaştırılmıştır. Sınıflandırma algoritmaları için kullanılan veri seti, hedef değişkeni, tanımlayıcı değişken, performans değerlendirme yöntemi ve parametre seçimin bilgisi özet bir tablo halinde ortaya konulmuştur.

4.1. TEDAVİYE YANIT HEDEF DEĞİŞKENİ İÇİN ELDE EDİLEN BULGULAR

Veri Seti	Overca_csv				
Hedef Değişken	Tedaviye Yanıt (Tam Yanıt/Kısmi Yanıt/Progresyon)				
Tanımlayıcı Değişken	Tanı Yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümör Patolojisi, Kemoterapi, Nüks, Tümörün Lokalizasyonu, Sağkalım				
Kullanılan Algoritmalar	Naive Bayes	DVM	Lojistik	J48	K-En Yakın Komşu Algoritması
Performans Değerlendirme Yöntemi	• Tabakalı 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme • Tabakalı örnekleme kullanılarak %50/%50, %75/%25, %80/%20 ve %90/%10 oranlarında Hold-Out				• Tabakalı 3-kat, 7-kat ve 9-kat çapraz geçerleme
Parametre Seçimi	Yok				K komşu sayısı 1'den 10'a kadar belirlenmiştir.

Tablo 19. Tedaviye Yanıt Hedef Değişkeni Analiz Özeti

Tablo 20. Tedaviye Yanıt Hedef Değişkeni Analizinde Kullanılan Algoritmaların Performans Ölçüleri

Naive Bayes	2kat	3kat	4kat	5kat	6kat	7kat	8kat	9kat	10kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma	75,6%	76,7%	70,0%	73,3%	75,6%	76,7%	75,6%	75,6%	75,6%	77,8%	81,8%	77,8%	88,9%
Kappa İstatistiği	0,58	0,61	0,50	0,55	0,59	0,60	0,58	0,59	0,59	0,62	0,68	0,62	0,79
Duyarlılık	0,76	0,77	0,70	0,73	0,76	0,77	0,76	0,76	0,76	0,78	0,82	0,78	0,89
Kesinlik	0,72	0,76	0,68	0,73	0,75	0,75	0,73	0,75	0,75	0,82	0,81	0,77	1
F Ölçütü	0,74	0,76	0,69	0,73	0,75	0,76	0,74	0,75	0,75	0,79	0,80	0,76	0,94
ROC Alanı	0,86	0,90	0,89	0,88	0,90	0,90	0,89	0,89	0,88	0,92	0,93	0,93	0,96
MAE	0,22	0,19	0,21	0,21	0,20	0,20	0,21	0,20	0,21	0,19	0,18	0,19	0,13
RMSE	0,36	0,33	0,35	0,34	0,33	0,33	0,34	0,34	0,34	0,31	0,31	0,33	0,24
RAE	54,9%	48,0%	52,3%	51,9%	49,5%	49,1%	51,4%	49,9%	51,2%	45,4%	43,4%	46,2%	32,7%
RRSE	81,1%	73,0%	77,0%	76,6%	74,3%	74,0%	75,7%	75,3%	76,4%	45,4%	68,7%	71,5%	52,5%
DVM	**2kat**	**3kat**	**4kat**	**5kat**	**6kat**	**7kat**	**8kat**	**9kat**	**10kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
Doğru Sınıflandırma	72,2%	78,9%	73,3%	77,8%	72,2%	77,8%	75,6%	75,6%	77,8%	80,0%	77,3%	72,2%	77,8%
Kappa İstatistiği	0,53	0,65	0,55	0,62	0,54	0,63	0,59	0,59	0,62	0,64	0,59	0,51	0,54
Duyarlılık	0,72	0,79	0,73	0,78	0,72	0,78	0,76	0,76	0,78	0,80	0,77	0,72	0,78
Kesinlik	0,71	0,79	0,71	0,75	0,72	0,78	0,75	0,76	0,76	0,83	0,73	0,68	0,89
F Ölçütü	0,71	0,79	0,72	0,76	0,72	0,78	0,75	0,76	0,77	0,80	0,74	0,68	0,82
ROC Alanı	0,80	0,85	0,81	0,81	0,83	0,86	0,82	0,82	0,83	0,82	0,86	0,82	0,82
MAE	0,31	0,29	0,30	0,30	0,30	0,28	0,30	0,30	0,29	0,29	0,27	0,28	0,27
RMSE	0,40	0,38	0,39	0,38	0,38	0,37	0,39	0,39	0,38	0,37	0,35	0,37	0,35
RAE	76,5%	71,8%	74,3%	73,6%	73,6%	70,0%	73,7%	74,3%	71,8%	69,9%	66,5%	68,9%	66,7%
RRSE	88,6%	83,4%	86,6%	85,6%	85,6%	81,4%	85,8%	86,5%	83,9%	82,9%	77,6%	80,4%	77,9%

Tablo 21. (Devam) Tedaviye Yanıt Hedef Değişkeni Analizinde Kullanılan Algoritmaların Performans Ölçüleri

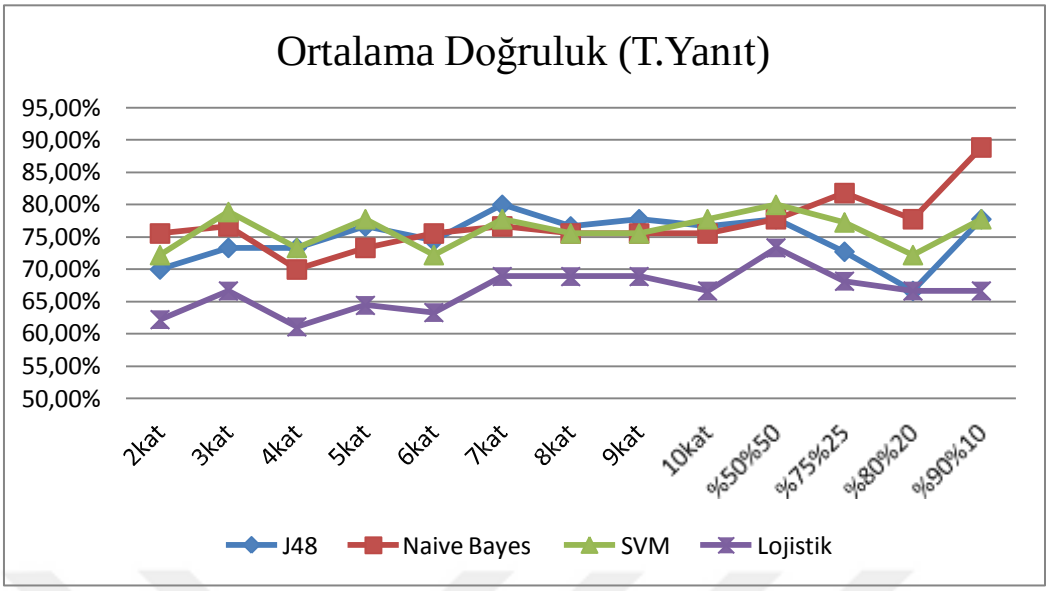
Lojistik	2-kat	3-kat	4-kat	5-kat	6-kat	7-kat	8-kat	9-kat	10-kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma	62,2%	66,7%	61,1%	64,4%	63,3%	68,9%	68,9%	68,9%	66,7%	73,3%	68,2%	66,7%	66,7%
Kappa İstatistiği	0,39	0,47	0,38	0,42	0,42	0,49	0,49	0,49	0,45	0,54	0,46	0,43	0,44
Duyarlılık	0,62	0,67	0,61	0,64	0,63	0,69	0,69	0,69	0,67	0,73	0,68	0,67	0,67
Kesinlik	0,66	0,70	0,66	0,67	0,67	0,69	0,71	0,69	0,68	0,75	0,71	0,66	0,89
F Ölçütü	0,64	0,68	0,63	0,65	0,65	0,69	0,70	0,69	0,67	0,74	0,68	0,64	0,76
ROC Alanı	0,78	0,79	0,79	0,78	0,80	0,83	0,77	0,78	0,78	0,86	0,81	0,74	0,76
MAE	0,24	0,22	0,26	0,24	0,24	0,21	0,21	0,21	0,22	0,18	0,21	0,22	0,22
RMSE	0,49	0,47	0,51	0,48	0,49	0,45	0,45	0,45	0,47	0,42	0,46	0,47	0,47
RAE	60,2%	55,4%	63,9%	58,2%	59,7%	51,1%	51,4%	51,1%	54,6%	43,4%	51,1%	53,9%	54,7%
RRSE	107,9%	103,9%	112,7%	107,6%	108,7%	101,0%	101,1%	100,9%	103,6%	94,2%	100,0%	102,8%	104,5%
J48	**2-kat**	**3-kat**	**4-kat**	**5-kat**	**6-kat**	**7-kat**	**8-kat**	**9-kat**	**10-kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
Doğru Sınıflandırma	70,0%	73,3%	73,3%	76,7%	74,4%	80,0%	76,7%	77,8%	76,7%	77,8%	72,7%	66,7%	77,8%
Kappa İstatistiği	0,52	0,54	0,54	0,59	0,55	0,65	0,60	0,61	0,59	0,59	0,51	0,43	0,57
Duyarlılık	0,70	0,73	0,73	0,77	0,74	0,80	0,77	0,78	0,77	0,78	0,73	0,67	0,78
Kesinlik	0,74	0,73	0,71	0,78	0,72	0,80	0,76	0,81	0,76	0,80	0,75	0,71	0,87
F Ölçütü	0,72	0,72	0,71	0,75	0,72	0,78	0,75	0,75	0,75	0,77	0,70	0,63	0,78
ROC Alanı	0,83	0,80	0,82	0,84	0,80	0,84	0,85	0,82	0,82	0,84	0,85	0,78	0,89
MAE	0,21	0,23	0,23	0,20	0,22	0,19	0,20	0,21	0,20	0,24	0,23	0,26	0,19
RMSE	0,39	0,37	0,38	0,35	0,37	0,35	0,35	0,35	0,36	0,33	0,36	0,40	0,32
RAE	51,4%	57,7%	57,8%	48,6%	54,4%	47,7%	49,7%	51,1%	49,4%	59,2%	55,5%	64,0%	46,3%
RRSE	85,8%	82,7%	83,9%	78,1%	82,2%	76,9%	78,5%	78,4%	79,3%	75,4%	79,4%	86,8%	71,5%

Naive Bayes Algoritması ile elde edilen en yüksek doğru sınıflandırma performansı %90%10 tabakalı Hold-outta %88,9 olarak bulunmuştur. Kappa istatistiği en yüksek 0,79 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. En yüksek kesinlik değeri %90%10 tabakalı Hold-Out ile 1,00 olarak bulunmuştur. En yüksek duyarlılık değeri ise 0,89 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. F-ölçütü incelendiğinde ise en yüksek performans %90%10 tabakalı Hold-Out ile 0,93 olarak bulunmuştur.

DVM algoritmasıyla elde edilen en yüksek doğru sınıflandırma %50%50 tabakalı Hold-Out ile %80 olarak elde edilmiştir. Kappa istatistiği en yüksek 0,65 olarak 3-kat çapraz geçerleme ile bulunmuştur. En yüksek kesinlik değeri 0,83 ile %50%50 tabakalı Hold-Outta gerçekleşmiştir. Duyarlılık değeri en yüksek 0,80 ile %50%50 tabakalı Hold-Outta görülmüştür. En yüksek F Ölçütü ise %90%10 tabakalı Hold-Outt ile 0,82 olarak elde edilmiştir.

Lojistik Regresyon ile elde edilen en yüksek doğru sınıflandırma performansı %50%50 tabakalı Hold-Out ile %73,3 olarak bulunmuştur. Kappa istatistiği en yüksek 0,54 olarak %50%50 tabakalı Hold-Out ile elde edilmiştir. En yüksek Kesinlik değeri 0,89 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. Duyarlılık değeri en yüksek 0,73 ile %50%50 tabakalı Hold-Outta görülmüştür. En yüksek F Ölçütü ise 0,76 ile %90%10 tabakalı Hold-Outta bulunmuştur.

J48 algoritmasıyla en yüksek doğru sınıflandırma performansı %80 olarak 7-kat çapraz geçerleme ile elde edilmiştir. Kappa istatistiği en yüksek 0,65 ile 7-kat çapraz geçerlemede bulunmuştur. En yüksek kesinlik değeri 0,87 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. Duyarlılık değeri en yüksek 7-kat çapraz geçerleme ile %80 olarak görülmüştür. En yüksek F Ölçütü ise 0,78 ile 7-kat çapraz geçerleme ve %90%10 tabakalı Hold-Outta bulunmuştur.



Şekil 7. Tedaviye Yanıt Hedef Değişkenine Ait Ortalama Doğruluk

Tedaviye Yanıt hedef değişkenine tüm algoritmaların uygulanması sonucu elde edilen ortalama doğruluk oranları karşılaştırılmıştır. Naive Bayes algoritması için elde edilen en yüksek sınıflandırma performansı %88,9 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. DVM algoritması için en yüksek sınıflandırma performansı %80 ile %50%50 tabakalı Hold-Outta, Lojistik Regresyon için %73,3 ile 50%50 tabakalı Hold-Outta ve J48 algoritması için ise %80 ile 7-kat çapraz geçerlemede görülmüştür. Bu sonuçlar doğrultusunda, tedaviye yanıt hedef değişkeni için doğru sınıflandırma oranı açısından en iyi performansı Naive Bayes algoritması sunmuştur. En düşük performans ise Lojistik Regresyon algoritması ile gerçekleşmiştir.

Tedaviye yanıt hedef değişkenine K-En Yakın Komşu algoritmasının uygulanması aşamasında, en yüksek k parametresini elde etmek için performans değerlendirme yöntemi olarak 3-kat, 7-kat ve 9-kat çapraz geçerleme kullanılmıştır. Ayrıca k'nın en yüksek değere ulaşabilmesi için 1'den 10'a kadar ayrı ayrı uygulanmıştır.

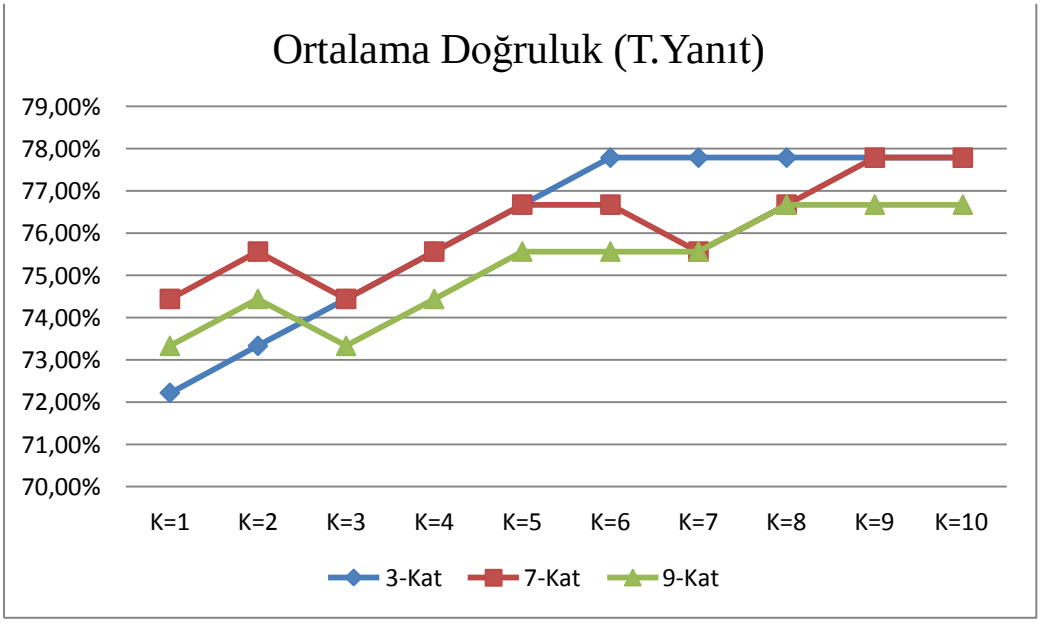
Tablo 22. Tedaviye Yanıt Hedef Değişkeni Analizinde Kullanılan K-En Yakın Komşu Algoritmasına Ait Performans Ölçüleri.

	Tabakalı 3-Kat Çapraz Geçerleme									
K	**K=1**	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**	**K=10**
Doğru Sınıflandırma Oranı	72,2%	73,3%	74,4%	75,6%	76,7%	77,8%	77,8%	77,8%	77,8%	77,8%
Kappa İstatistiği	0,53	0,54	0,55	0,57	0,58	0,60	0,60	0,60	0,60	0,60
Duyarlılık	0,72	0,73	0,74	0,76	0,77	0,78	0,78	0,78	0,78	0,78
Kesinlik	0,71	0,71	0,72	0,74	0,77	0,76	0,76	0,76	0,76	0,69
F Ölçütü	0,72	0,72	0,72	0,73	0,74	0,74	0,74	0,74	0,74	0,72
ROC Alanı	0,85	0,88	0,87	0,86	0,86	0,87	0,87	0,87	0,87	0,87
MAE	0,20	0,21	0,22	0,23	0,24	0,25	0,26	0,26	0,27	0,28
RMSE	0,38	0,34	0,35	0,35	0,35	0,34	0,35	0,35	0,36	0,36
RAE	49,2%	51,1%	54,4%	57,5%	60,0%	61,1%	63,7%	64,7%	66,6%	68,1%
RRSE	83,8%	75,9%	77,0%	77,0%	77,5%	76,4%	77,5%	78,3%	79,0%	79,2%
	Tabakalı 7-Kat Çapraz Geçerleme									
K	**K=1**	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**	**K=10**
Doğru Sınıflandırma Oranı	74,4%	75,6%	74,4%	75,6%	76,7%	76,7%	75,6%	76,7%	77,8%	77,8%
Kappa İstatistiği	0,57	0,58	0,55	0,57	0,58	0,58	0,56	0,58	0,60	0,60
Duyarlılık	0,74	0,76	0,74	0,76	0,77	0,77	0,76	0,77	0,78	0,78
Kesinlik	0,73	0,73	0,71	0,72	0,78	0,78	0,75	0,76	0,76	0,76
F Ölçütü	0,73	0,74	0,71	0,72	0,74	0,74	0,72	0,73	0,74	0,74
ROC Alanı	0,81	0,87	0,87	0,88	0,89	0,89	0,89	0,89	0,90	0,89
MAE	0,21	0,21	0,22	0,23	0,22	0,23	0,24	0,24	0,25	0,26
RMSE	0,39	0,36	0,35	0,35	0,34	0,34	0,34	0,34	0,34	0,34
RAE	50,9%	52,4%	54,4%	56,1%	55,4%	57,7%	59,6%	60,2%	61,7%	63,2%
RRSE	87,3%	80,4%	77,7%	77,4%	75,6%	75,4%	75,8%	76,0%	75,6%	76,3%
	Tabakalı 9-Kat Çapraz Geçerleme									
K	**K=1**	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**	**K=10**
Doğru Sınıflandırma Oranı	73,3%	74,4%	73,3%	74,4%	75,6%	75,6%	75,6%	76,7%	76,7%	76,7%
Kappa İstatistiği	0,54	0,55	0,52	0,54	0,56	0,56	0,56	0,58	0,58	0,58
Duyarlılık	0,73	0,74	0,73	0,74	0,76	0,76	0,76	0,77	0,77	0,77
Kesinlik	0,70	0,74	0,64	0,66	0,68	0,68	0,68	0,68	0,68	0,68
F Ölçütü	0,71	0,71	0,68	0,69	0,70	0,70	0,70	0,71	0,71	0,71
ROC Alanı	0,84	0,87	0,87	0,88	0,88	0,88	0,88	0,89	0,89	0,89
MAE	0,20	0,20	0,22	0,22	0,56	0,24	0,24	0,25	0,25	0,26
RMSE	0,38	0,34	0,35	0,34	0,34	0,34	0,34	0,34	0,34	0,35
RAE	48,1%	49,6%	53,5%	53,7%	55,5%	58,1%	59,3%	61,3%	62,7%	64,0%
RRSE	83,8%	76,7%	77,3%	75,7%	75,4%	76,0%	76,4%	75,8%	76,4%	77,1%

Tedaviye yanıt hedef değişkenine K-En yakın komşu algoritmasının uygulanması ile elde edilen sonuçlara göre 3-Kat çapraz geçerleme için en yüksek sınıflandırma k=6, 7, 8, 9 ve 10 iken % 77,8 olarak gerçekleşmiştir. Kappa istatistiği en yüksek k=6, 7, 8 ve 9 iken 0,60 olarak bulunmuştur. Duyarlılık değeri ise en yüksek k=6, 7, 8, 9 ve 10 iken 0,78 olarak elde edilmiştir. En yüksek kesinlik değeri k=5 iken 0,77 olarak görülmüştür. F Ölçütü ise en yüksek k=5, 6, 7, 8 ve 9 iken 0,74 olarak bulunmuştur.

Tabakalı 7–kat çapraz geçerleme için en yüksek sınıflandırma k=9 ve 10 iken %77,8 olarak görülmüştür. Kappa istatistiği en yüksek k=9 ve 10 iken 0,60 olarak elde edilmiştir. Duyarlılık değeri en yüksek k=9 ve 10 iken 0,78 olarak bulunmuştur. En yüksek kesinlik değeri k=5 ve 6 iken 0,78 olarak gerçekleşmiştir. F Ölçütü ise en yüksek k=5, 6, 9 ve 10 iken 0,74 olarak elde edilmiştir.

Tabakalı 9-kat çapraz geçerleme için en yüksek sınıflandırma performansı k=8, 9 10 iken %76,7 olarak gerçekleşmiştir. Kappa istatistiği en yüksek k=8, 9 ve 10 iken 0,58 olarak görülmüştür. Duyarlılık en yüksek k=8, 9 ve 10 iken 0,77 olarak bulunmuştur. Kesinlik k=2 iken 0,74 olarak elde edilmiştir. F Ölçütü ise en yüksek k=2, 8, 9 ve 10 iken 0,71 olarak gerçekleşmiştir.



Şekil 8. Tedaviye Yanıt Hedef Değişkenine Ait K-En Algoritması ile Elde Edilen Ortalama Doğruluk

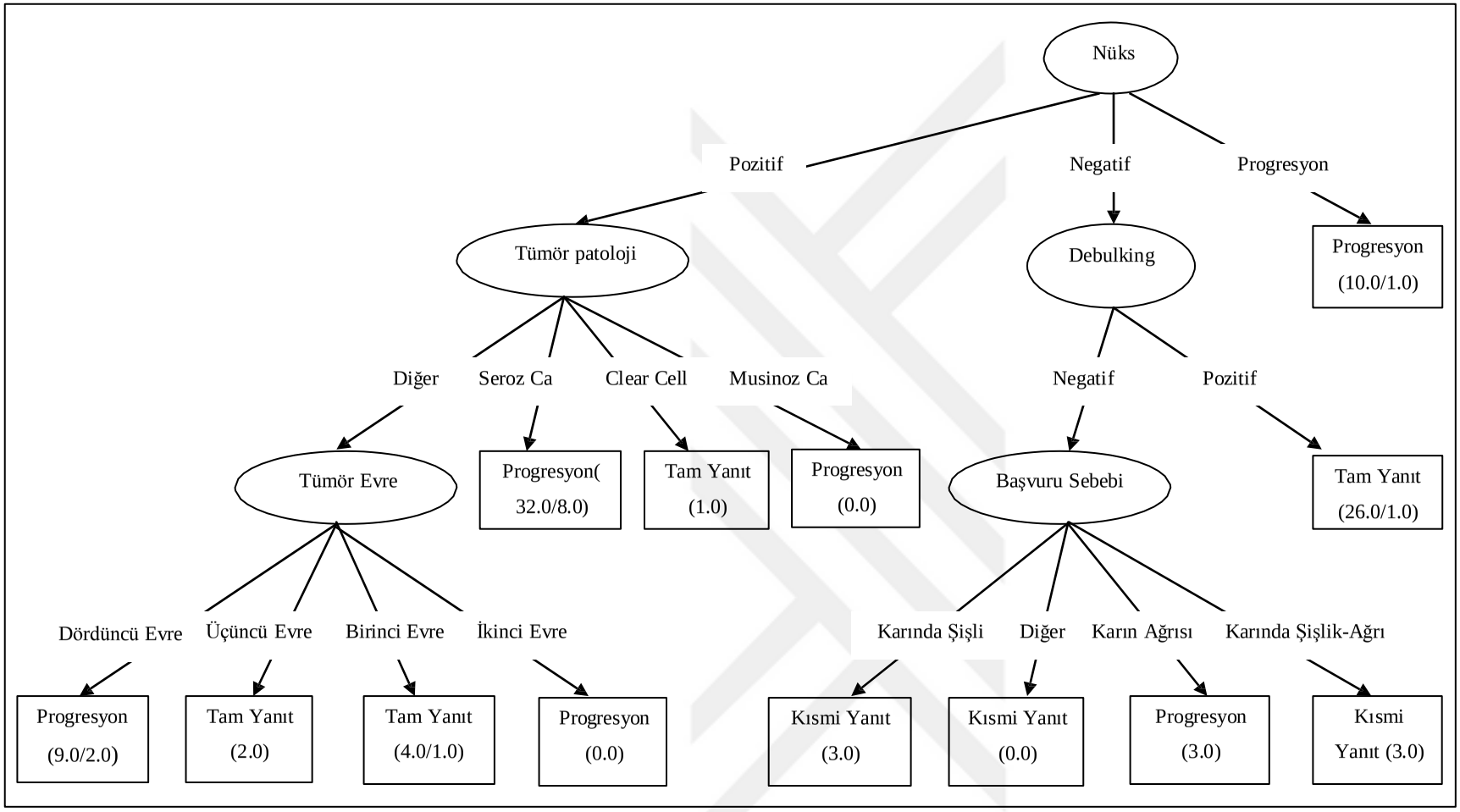
Şekil 8'de yer alan grafik incelendiğinde, k-En Yakın Komşu algoritmasının tedaviye yanıt hedef değişkenine uygulanması sonucunda, en yüksek sınıflandırma performansı 3-kat çapraz geçerleme ve 7-kat çapraz geçerleme ile %77,78 olarak elde edilmiştir. Ayrıca elde edilen diğer sınıflandırma oranları da %70'in üzerinde gerçekleşmiştir.

Tablo 23. Tedaviye Yanıt Hedef Değişkeni İçin Elde Edilen Ölçütlerin Karşılaştırılması

Algoritma / Ölçüt	Naive Bayes	DVM	Lojistik Regresyon	J48 Karar Ağacı	K-En Yakın Komşu Algoritması
Doğru Sınıflandırma Oranı	%88,9	%80	%73,3	%80	%77,78
F-Ölçütü	0,93	0,80	0,76	0,78	0,74

Tablo 23'te bulunan sonuçlar incelendiğinde doğru sınıflandırma ölçütüne göre en iyi performans Naive Bayes algoritmasının %88,9 doğruluk derecesiyle elde edilmiştir. Bu ölçüte göre J48 algoritmasından sonra en iyi performans aynı sınıflandırma oranına sahip DVM ve J48 Karar Ağacı ile elde edilmiştir. Bu algoritmaları sırası ile K-En Yakın Komşu ve Lojistik Regresyon algoritmaları takip etmektedir.

F-Ölçütünün incelenmesi, bu ölçütün hesaplanmasında kullanılan kesinlik ve duyarlılık ölçütlerinin birlikte incelenmesine olanak sağlayacaktır. Bu yüzden F-Ölçütünün karşılaştırma ölçütü olarak kullanılmasının daha kapsayıcı olacağı düşünülmüştür. F-Ölçütüne baktığımızda en yüksek sonuç Naive Bayes algoritması ile gerçekleşmiştir. En düşük sonuç ise Lojistik regresyon algoritması ile elde edilmiştir. F-Ölçütü için en yüksek ve en düşük sonuçlara sahip algoritmaların doğru sınıflandırma oranında da aynı performans sıralamasına sahip oldukları görülmüştür. Şekil 9'da tedaviye yanıt hedef değişkenine J48 algoritmasının uygulanması ile elde edilen en yüksek doğru sınıflandırma oranına sahip 7-kat çapraz geçerleme için çizilen karar ağaç yapısı verilmiştir.



Şekil 9. Tedaviye Yanıt Hedef Değişkenine Ait J48 Ağaç Yapısı

Şekil 9'da hastaların tedaviye vermiş oldukları yanıtların tahmin edilmesinde J48 algoritmasının çizmiş olduğu ağaç yapısı görülmektedir. Ağaçlandırma tanımlayıcı değişkenlerden nüks değişkeni ile başlamıştır. Nüks eğer progresyon ise tedaviye verilen yanıtın da progresyon olduğu görülmektedir. Nüks eğer negatif ise debulkinge bakılmaktadır. Debulking pozitif ise tedaviye verilen yanıt tam yanıt olarak gerçekleşmektedir. Debulking eğer negatif ise başvuru sebebine bakılır. Başvuru sebebi karında şişlik, diğer ve karında şişlik-ağrı ise tedaviye yanıt kismi yanıt olarak gerçekleşmektedir. Karın ağrısı ise tedaviye yanıt, progresyon olarak gerçekleşmektedir. Nüks eğer pozitif ise tümörün patolojine gidilir. Tümörün patolojisi Seroz Ca ve Musinoz Ca ise tedaviye verilen yanıtın progresyon, Clear Cell ise tedaviye verilen yanıtın pozitif olduğu görülmektedir. Tümörün patolojisi diğer ise tümörün evresine gidilir. Tümör ikinci ve dördüncü evre ise tedaviye yanıt progresyon olarak gerçekleşmektedir. Birinci ve üçüncü evre ise tedaviye verilen yanıtın tam yanıt olduğu görülmektedir.

4.2. SAĞ KALIM HEDEF DEĞİŞKENİ İÇİN ELDE EDİLEN BULGULAR

Tablo 24. Sağ Kalım Hedef Değişkeni Analiz Özeti.

Veri Seti	Overca_csv				
Hedef Değişken	Sağkalım (1080 gün ve altı/1081 gün ve üstü)				
Tanımlayıcı Değişken	Tanı Yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümör Patolojisi, Kemoterapi, Nüks, Tedaviye Yanıt, Tümörün Lokalizasyonu				
Kullanılan Algoritmalar	Naive Bayes	DVM	Lojistik	J48	K-En Yakın Komşu Algoritması
Performans Değerlendirme Yöntemi	• Tabakalı 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme • Tabakalı örnekleme kullanılarak %50/%50, %75/%25, %80/%20 ve %90/%10 oranlarında Hold-Out				• Tabakalı 3-kat, 8-kat ve 9-kat çapraz geçerleme
Parametre Seçimi	Yok				K komşu sayısı 1'den 10'a kadar belirlenmiştir.

Tablo 25. Sağ Kalım Hedef Değişkeni Analizinde Kullanılan Algoritmaların Performans Ölçüleri

Naive Bayes	2-kat	3-kat	4-kat	5-kat	6-kat	7-kat	8-kat	9-kat	10-kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma	65,6%	73,3%	67,8%	67,8%	68,9%	67,8%	68,9%	67,8%	70,0%	66,7%	50,0%	55,6%	66,7%
Kappa İstatistiği	0,27	0,43	0,31	0,30	0,34	0,30	0,31	0,31	0,35	0,29	0,17	0,20	0,18
Duyarlılık	0,66	0,73	0,68	0,68	0,69	0,68	0,69	0,68	0,70	0,67	0,50	0,56	0,67
Kesinlik	0,67	0,74	0,69	0,68	0,70	0,68	0,69	0,69	0,70	0,66	0,79	0,69	0,72
F Ölçütü	0,66	0,74	0,68	0,68	0,69	0,68	0,69	0,68	0,70	0,66	0,44	0,56	0,69
ROC Alanı	0,74	0,77	0,73	0,73	0,73	0,72	0,75	0,74	0,74	0,74	0,88	0,78	0,79
MAE	0,36	0,34	0,36	0,36	0,36	0,37	0,35	0,35	0,35	0,38	0,42	0,41	0,37
RMSE	0,46	0,44	0,46	0,46	0,46	0,47	0,45	0,46	0,46	0,48	0,50	0,49	0,42
RAE	78,1%	73,9%	78,1%	77,8%	78,2%	81,2%	76,4%	77,0%	76,7%	81,9%	74,6%	71,6%	61,1%
RRSE	95,5%	92,2%	96,0%	96,8%	96,2%	99,1%	94,7%	95,2%	95,3%	93,6%	82,6%	80,5%	68,1%
DVM	**2-kat**	**3-kat**	**4-kat**	**5-kat**	**6-kat**	**7-kat**	**8-kat**	**9-kat**	**10-kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
Doğru Sınıflandırma Oranı	67,8%	68,9%	70,0%	65,6%	63,3%	66,7%	71,1%	74,4%	68,9%	60,0%	50,0%	44,4%	66,7%
Kappa İstatistiği	0,29	0,33	0,36	0,22	0,22	0,28	0,36	0,45	0,31	0,08	0,17	0,12	0,37
Duyarlılık	0,68	0,69	0,70	0,66	0,63	0,67	0,71	0,74	0,69	0,60	0,50	0,44	0,67
Kesinlik	0,68	0,69	0,71	0,65	0,64	0,67	0,71	0,75	0,69	0,63	0,79	0,79	0,87
F Ölçütü	0,68	0,69	0,70	0,65	0,64	0,67	0,71	0,75	0,69	0,50	0,44	0,37	0,69
ROC Alanı	0,65	0,67	0,68	0,61	0,61	0,64	0,68	0,73	0,65	0,53	0,61	0,58	0,79
MAE	0,32	0,31	0,30	0,34	0,37	0,33	0,29	0,26	0,31	0,40	0,50	0,56	0,33
RMSE	0,57	0,56	0,55	0,59	0,61	0,58	0,54	0,51	0,56	0,63	0,71	0,75	0,58
RAE	70,0%	67,7%	65,3%	74,9%	79,8%	72,5%	62,9%	55,6%	67,7%	85,4%	88,9%	97,1%	55,2%
RRSE	118,6%	116,4%	114,4%	122,5%	126,4%	120,5%	112,3%	105,5%	116,4%	124,2%	117,1%	122,7%	92,6%

Tablo 26. (Devam) Sağ Kalım Hedef Değişkeni Analizinde Kullanılan Algoritmaların Performans Ölçüleri

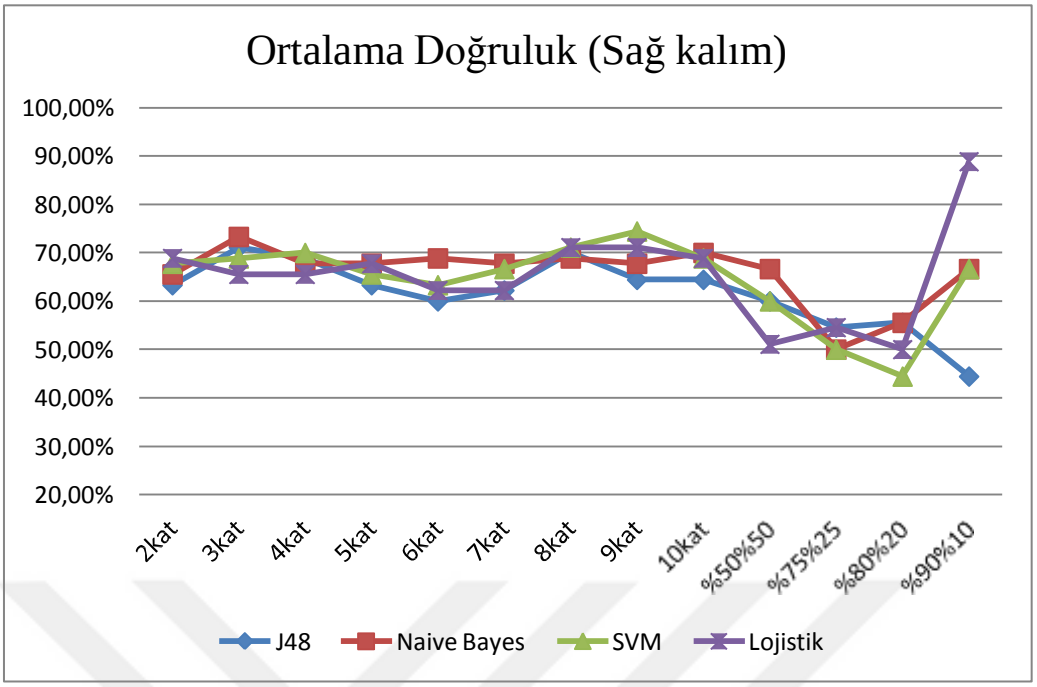
Lojistik	2-kat	3-kat	4-kat	5-kat	6-kat	7-kat	8-kat	9-kat	10-kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma Oranı	68,9%	65,6%	65,6%	67,8%	62,2%	62,2%	71,1%	71,1%	68,9%	51,1%	54,6%	50,0%	88,9%
Kappa İstatistiği	0,32	0,28	0,22	0,26	0,21	0,18	0,36	0,38	0,31	0,00	0,23	0,13	0,73
Duyarlılık	0,69	0,66	0,66	0,68	0,62	0,62	0,71	0,71	0,69	0,51	0,55	0,50	0,89
Kesinlik	0,69	0,68	0,65	0,67	0,64	0,62	0,71	0,72	0,69	0,51	0,80	0,66	0,93
F Ölçütü	0,69	0,66	0,65	0,67	0,63	0,62	0,71	0,71	0,69	0,51	0,51	0,49	0,90
ROC Alanı	0,74	0,70	0,69	0,68	0,69	0,65	0,76	0,75	0,73	0,60	0,79	0,78	1,00
MAE	0,29	0,34	0,35	0,35	0,37	0,37	0,32	0,33	0,35	0,49	0,44	0,42	0,28
RMSE	0,52	0,54	0,52	0,52	0,54	0,54	0,47	0,48	0,50	0,68	0,55	0,49	0,34
RAE	63,7%	73,9%	75,6%	76,5%	81,4%	81,5%	69,0%	71,4%	75,2%	103,6%	77,7%	72,7%	46,0%
RRSE	109,4%	113,0%	108,3%	108,3%	111,7%	112,6%	97,8%	100,3%	103,8%	133,3%	91,7%	80,6%	53,7%
J48	**2-kat**	**3-kat**	**4-kat**	**5-kat**	**6-kat**	**7-kat**	**8-kat**	**9-kat**	**10-kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
Doğru Sınıflandırma Oranı	63,3%	71,1%	68,9%	63,3%	60,0%	62,2%	70,0%	64,4%	64,4%	60,0%	54,6%	55,6%	44,4%
Kappa İstatistiği	0,24	0,35	0,31	0,23	0,09	0,13	0,30	0,18	0,20	0,10	0,23	0,25	0,15
Duyarlılık	0,63	0,71	0,69	0,63	0,60	0,62	0,70	0,64	0,64	0,60	0,55	0,56	0,44
Kesinlik	0,65	0,70	0,69	0,65	0,58	0,60	0,69	0,63	0,64	0,59	0,80	0,81	0,84
F Ölçütü	0,64	0,71	0,69	0,64	0,59	0,61	0,69	0,63	0,64	0,55	0,51	0,53	0,44
ROC Alanı	0,67	0,74	0,70	0,62	0,63	0,60	0,65	0,64	0,61	0,65	0,83	0,81	0,75
MAE	0,38	0,35	0,36	0,41	0,41	0,41	0,38	0,40	0,41	0,42	0,45	0,44	0,49
RMSE	0,49	0,46	0,48	0,54	0,53	0,53	0,49	0,52	0,53	0,55	0,56	0,53	0,55
RAE	83,7%	77,1%	79,3%	89,7%	88,7%	89,3%	82,4%	86,0%	88,5%	89,9%	80,8%	76,8%	81,8%
RRSE	101,8%	96,0%	100,5%	113,2%	110,4%	110,3%	102,5%	109,3%	109,8%	108,3%	92,1%	87,7%	88,3%

Tablo 25'te Naive Bayes Algoritması için en yüksek sınıflandırma 3-kat çapraz geçerleme ile %73,3 olarak bulunmuştur. Kappa istatistiği en yüksek 0,43 olarak 3-kat çapraz geçerlemede gerçekleşmiştir. Kesinlik değeri en yüksek %75%25 tabakalı Hold-Outta 0,79 olarak bulunmuştur. En yüksek duyarlılık değeri 0,73 ile 3-kat geçerlemede görülmüştür. F ölçütü incelendiğinde ise en yüksek performans 0,74 olarak 3-kat çapraz geçerleme ile elde edilmiştir.

DVM algoritmasıyla elde edilen en yüksek sınıflandırmayı 9-kat çapraz geçerleme %74,4 oranıyla vermiştir. Kappa istatistiği en yüksek 0,45 olarak 9-kat çapraz geçerleme ile elde edilmiştir. En yüksek kesinlik değeri 0,87 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. Duyarlılık değeri 0,74 ile 9-kat çapraz geçerlemede görülmüştür. F Ölçütü ise en yüksek 0,75 ile 9-kat çapraz geçerlemede bulunmuştur.

Lojistik Regresyon ile elde edilen sonuçlara göre en yüksek sınıflandırma performansı %90%10 tabakalı Hold-Outta %88,9 ile gerçekleşmiştir. Kappa istatistiği en yüksek 0,73 ile %90%10 tabakalı Hold-Outta elde edilmiştir. En yüksek kesinlik değeri 0,93 olarak %90%10 tabakalı Hold-Out ile bulunmuştur. Duyarlılık değeri en yüksek 0,89 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. F Ölçütü ise en yüksek 0,90 ile %90%10 tabakalı Hold-Outta bulunmuştur.

J48 algoritmasıyla en yüksek sınıflandırma performansı %71,1 ile 3-kat çapraz geçerlemeden elde edilmiştir. Kappa istatistiği en yüksek 0,35 ile 3-kat çapraz geçerleme ile gerçekleşmiştir. En yüksek kesinlik değeri 0,84 ile %90%10 tabakalı Hold-Outta bulunmuştur. Duyarlılık değeri en yüksek 0,71 ile 3-kat çapraz geçerlemede görülmüştür. F Ölçütü ise en yüksek 0,71 ile 3-kat çapraz geçerlemede bulunmuştur. Şekil 4.8'de sağ kalım hedef değişkenine J48 algoritmasının uygulanması ile elde edilen en yüksek doğru sınıflandırma oranına sahip 3-kat çapraz geçerleme için çizilen karar ağaç yapısı verilmektedir.



Şekil 10. Sağ Kalım Hedef Değişkenine Ait Ortalama Doğruluk

Şekil 10'da bulunan grafik incelendiğinde Naive Bayes algoritması için elde edilen en yüksek sınıflandırma performansı 3-kat çapraz geçerleme ile %73,3 olarak elde edilmiştir. DVM algoritması için en yüksek doğru sınıflandırma %74,4 ile 9-kat çapraz geçerlemede, Lojistik Regresyon için %88,9 ile %90%10 tabakalı Hold-Outta ve J48 algoritması için ise %71,1 ile 3-kat çapraz geçerlemede gerçekleşmiştir. Bu sonuçlar doğrultusunda, sağ kalım hedef değişkeni için en iyi sınıflandırma performansını Lojistik Regresyon algoritması sağlamıştır. En düşük sınıflandırma performansı ise J48 algoritması ile gerçekleşmiştir.

Sağ kalım hedef değişkeni için K-En Yakın Komşu algoritmasının uygulanması aşamasında, en yüksek k parametresini elde etmek için performans değerlendirme yöntemi olarak 3-kat, 8-kat ve 9-kat çapraz geçerleme kullanılmıştır. Ayrıca k'nın en yüksek değere ulaşabilmesi için 1'den 10'a kadar ayrı ayrı uygulanmıştır.

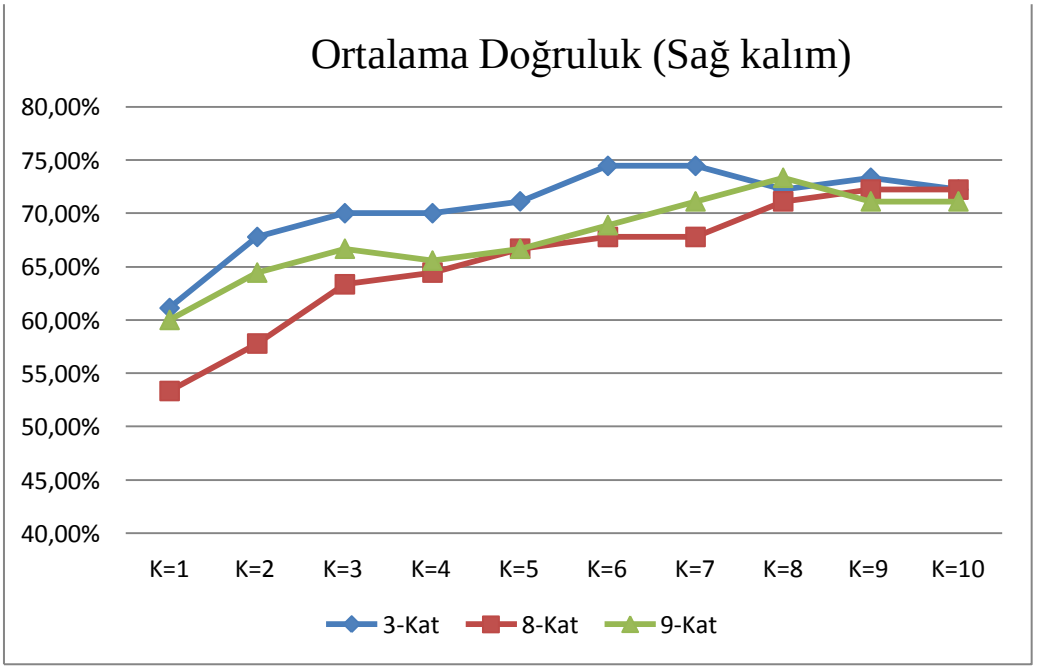
Tablo 27. Sağ Kalım Hedef Değişkeni Analizinde Kullanılan K-En Yakın Komşu Algoritmasına Ait Performans Ölçüleri

	Tabakalı 3-Kat Çapraz Geçerleme									
K-NN	**K=1**	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**	**K=10**
Doğru Sınıflandırma Oranı	61,1%	67,8%	70,0%	70,0%	71,1%	74,4%	74,4%	72,2%	73,3%	72,2%
Kappa İstatistiği	0,18	0,31	0,35	0,35	0,36	0,43	0,42	0,38	0,39	0,36
Duyarlılık	0,61	0,68	0,70	0,70	0,71	0,74	0,74	0,72	0,73	0,72
Kesinlik	0,63	0,69	0,70	0,70	0,71	0,74	0,74	0,72	0,73	0,71
F Ölçütü	0,62	0,68	0,70	0,70	0,71	0,74	0,74	0,72	0,73	0,71
ROC Alanı	0,65	0,73	0,72	0,72	0,75	0,78	0,77	0,78	0,79	0,79
MAE	0,38	0,36	0,38	0,39	0,38	0,38	0,39	0,39	0,39	0,40
RMSE	0,55	0,45	0,45	0,45	0,44	0,43	0,43	0,43	0,43	0,43
RAE	83,4%	78,1%	82,9%	85,2%	83,5%	82,7%	83,9%	85,2%	85,0%	86,4%
RRSE	114,9%	94,4%	94,2%	94,2%	91,7%	89,5%	89,8%	90,0%	89,6%	90,1%
	Tabakalı 8-Kat Çapraz Geçerleme									
K-NN	**K=1**	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**	**K=10**
Doğru Sınıflandırma Oranı	53,3%	57,8%	63,3%	64,4%	66,7%	67,8%	67,8%	71,1%	72,2%	72,2%
Kappa İstatistiği	0,02	0,12	0,21	0,21	0,24	0,26	0,26	0,34	0,35	0,35
Duyarlılık	0,53	0,58	0,63	0,64	0,67	0,68	0,68	0,71	0,72	0,72
Kesinlik	0,55	0,60	0,64	0,64	0,66	0,67	0,67	0,70	0,71	0,71
F Ölçütü	0,54	0,58	0,64	0,64	0,66	0,67	0,67	0,70	0,71	0,71
ROC Alanı	0,58	0,65	0,69	0,70	0,71	0,71	0,72	0,75	0,74	0,74
MAE	0,40	0,39	0,37	0,39	0,39	0,40	0,40	0,40	0,40	0,40
RMSE	0,56	0,49	0,46	0,46	0,45	0,45	0,45	0,44	0,44	0,44
RAE	86,3%	84,0%	80,3%	83,9%	84,6%	87,3%	87,4%	86,3%	87,0%	86,9%
RRSE	117,9%	102,6%	96,9%	95,3%	94,2%	94,2%	93,8%	92,0%	92,9%	92,5%
	Tabakalı 9-Kat Çapraz Geçerleme									
K-NN	**K=1**	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**	**K=7**	**K=8**	**K=9**	**K=10**
Doğru Sınıflandırma Oranı	60,0%	64,4%	66,7%	65,6%	66,7%	68,9%	71,1%	73,3%	71,1%	71,1%
Kappa İstatistiği	0,16	0,25	0,28	0,25	0,27	0,32	0,35	0,40	0,34	0,34
Duyarlılık	0,60	0,64	0,67	0,66	0,67	0,69	0,71	0,73	0,71	0,71
Kesinlik	0,62	0,66	0,67	0,66	0,67	0,69	0,70	0,73	0,70	0,70
F Ölçütü	0,61	0,65	0,67	0,66	0,67	0,69	0,71	0,73	0,70	0,70
ROC Alanı	0,60	0,66	0,69	0,72	0,71	0,73	0,74	0,74	0,74	0,73
MAE	0,39	0,39	0,38	0,38	0,39	0,39	0,39	0,39	0,39	0,40
RMSE	0,57	0,49	0,47	0,45	0,45	0,44	0,44	0,44	0,44	0,44
RAE	85,3%	83,8%	83,6%	83,4%	85,6%	85,5%	85,2%	85,7%	85,7%	86,7%
RRSE	118,7%	101,6%	97,6%	94,2%	94,4%	92,8%	92,1%	92,0%	91,1%	92,2%

Sağ kalım hedef değişkenine K-En yakın komşu algoritmasının uygulanması ile elde edilen sonuçlara göre 3-Kat çapraz geçerleme için en yüksek sınıflandırma performansı k=6 ve 7 iken % 74,4 olarak gerçekleşmiştir. Kappa istatistiği en yüksek k=6 iken 0,43 olarak görülmüştür. Duyarlılık değeri en yüksek k=6 ve 7 iken 0,74 olarak bulunmuştur. Kesinlik değeri en yüksek k=6 iken 0,74 olarak gerçekleşmiştir. F Ölçütü ise en yüksek 0,74 olarak k=6 ile elde edilmiştir.

Tabakalı 8–kat çapraz geçerleme için en yüksek sınıflandırma performansı k=9 ve 10 iken %72,2 olarak görülmüştür. Kappa istatistiği en yüksek k=9 ve 10 iken 0,35 olarak gerçekleşmiştir. Duyarlılık ise en yüksek 0,72 olarak k=9 ve 10 ile elde edilmiştir. Kesinlik değeri en yüksek k=9 ve 10 iken 0,71 olarak gerçekleşmiştir. En yüksek F-Ölçütü ise k=9 ve 10 ile 0,74 olarak görülmüştür.

Tabakalı 9-kat çapraz geçerleme için elde edilen en yüksek sınıflandırma performansı k=8 iken %73,3 olarak gerçekleşmiştir. Kappa istatistiği en yüksek k=8 ile 0,40 olarak görülmüştür. Duyarlılık değeri en yüksek 0,73 olarak k=8 ile elde edilmiştir. En yüksek kesinlik değeri k=2 ile 0,73 olarak bulunmuştur. F Ölçütü ise en yüksek 0,73 olarak k=8 ile gerçekleşmiştir.



Şekil 11. Sağ Kalım Hedef Değişkenine Ait K-En Algoritması ile Elde Edilen Ortalama Doğruluk

K-En Yakın Komşu algoritmasının sağ kalım hedef değişkenine uygulanması ile elde edilen Şekil 11'de bulunan grafik incelendiğinde, en yüksek sınıflandırma performansı 3-kat çapraz geçerleme ile 0,74 olarak gerçekleşmiştir. Genel olarak incelendiğinde 3- kat çapraz geçerleme ile elde edilen sonuçların daha yüksek olduğu görülmüştür.

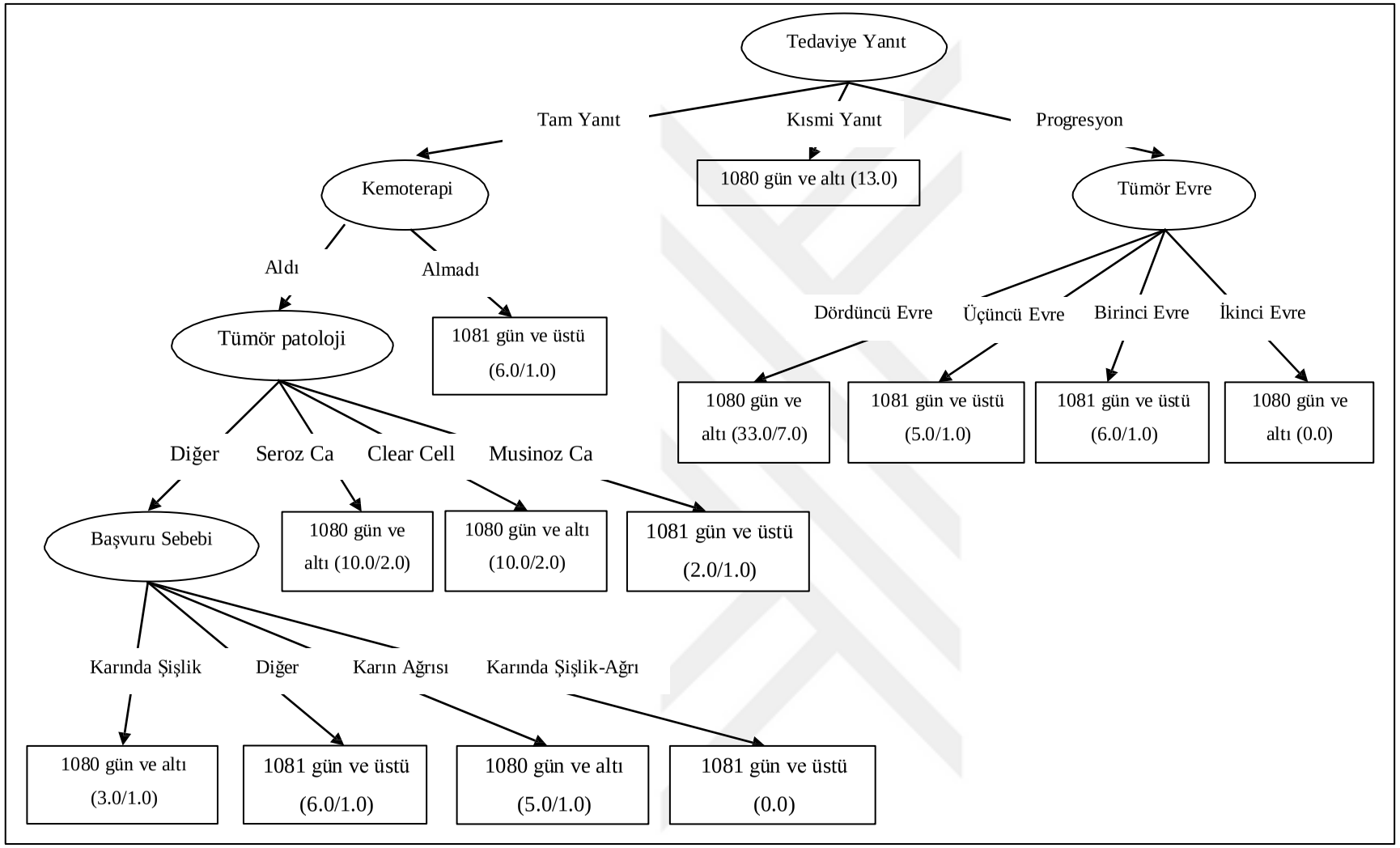
Tablo 28. Sağ Kalım hedef değişkeni için elde edilen ölçütlerin karşılaştırılması

Algoritma / Ölçüt	Naive Bayes	DVM	Lojistik Regresyon	J48 Karar Ağacı	K-En Yakın Komşu Algoritması
Doğru Sınıflandırma Oranı	%73,3	%74,4	%88,9	%71,1	0,74
F-Ölçütü	0,74	0,75	0,90	0,71	0,74

Tablo 28'de bulunan sonuçlar incelendiğinde doğru sınıflandırma ölçütüne göre en iyi performans Lojistik Regresyon analizinde %88,9 doğruluk oranı ile gerçekleşmiştir. Bu ölçüte göre Lojistik Regresyon analizini sırasıyla DVM, K-En yakın Komşu, Naive Bayes ve J48 karar ağacı algoritmaları izlemiştir.

F-Ölçütünün incelenmesi, bu ölçütün hesaplanmasında kullanılan kesinlik ve duyarlılık ölçütlerinin birlikte incelenmesine olanak sağlayacaktır. Bu yüzden F-Ölçütünün karşılaştırma ölçütü olarak kullanılmasının daha kapsayıcı olacağı düşünülmüştür. F-Ölçütüne baktığımızda en yüksek sonuç Lojistik Regresyon analizi ile gerçekleşmiştir. En düşük sonuç ise J48 Karar ağacı algoritması ile elde edilmiştir. F-Ölçütü için en yüksek ve en düşük sonuçlara sahip algoritmaların doğru sınıflandırma oranında da aynı sıralama performansına sahip oldukları görülmüştür.

Şekil 12'de tedaviye yanıt hedef değişkenine J48 algoritmasının uygulanması ile elde edilen en yüksek doğru sınıflandırma oranına sahip 3-kat çapraz geçerleme için çizilen karar ağaç yapısı verilmiştir.



Şekil 12. Sağ Kalım Hedef Değişkenine Ait J48 Ağaç Yapısı

Şekil 12'de yer alan ağaçlandırma tanımlayıcı değişkenlerden tedaviye yanıt değişkeni ile başlamıştır. Tedaviye yanıt eğer kısmi yanıt olarak gerçekleşir ise sağ kalımın 1080 gün ve altı olduğu görülmektedir. Tedaviye yanıt eğer progresyon ise tümörün evresine bakılmaktadır. Tümör ikinci ve dördüncü evre ise sağ kalım 1080 gün ve altı olarak gerçekleşmektedir. Eğer tümör birinci ve üçüncü evre ise sağ kalım 1081 gün ve üstü olarak görülmektedir. Hastaların tedaviye tam yanıt vermesi durumunda Kemoterapi'ye bakılır. Kemoterapi almamış ise sağ kalım 1081 gün ve üstü olarak gerçekleşmektedir. Eğer Kemoterapi almış ise tümörün patolojisine bakılır. Tümörün patolojisi Seroz Ca ise sağ kalım 1080 gün ve altı olarak gerçekleşmektedir. Clear Cell ve Musinoz Ca ise sağ kalımın 1081 gün ve üstü olduğu görülmektedir. Tümörün patolojisi diğer ise başvuru sebebine gidilir..

4.3. LOKALİZASYON HEDEF DEĞİŞKENİ İÇİN ELDE EDİLEN BULGULAR

Tablo 29. Lokalizasyon Hedef Değişkeni Analiz Özeti

Veri Seti	Overca_csv				
Hedef Değişken	Lokalizasyon (Sol/Sağ/ Sol ve Sağ)				
Tanımlayıcı Değişken	Tanı Yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümör Patolojisi, Kemoterapi, Nüks, Tedaviye Yanıt, Sağkalım				
Kullanılan Algoritmalar	Naive Bayes	DVM	Lojistik	J48	K-En Yakın Komşu Algoritması
Performans Değerlendirme Yöntemi	• Tabakalı 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme • Tabakalı örnekleme kullanılarak %50/%50, %75/%25, %80/%20 ve %90/%10 oranlarında Hold-Out				• Tabakalı 5-kat, 7-kat ve 10-kat çapraz geçerleme
Parametre Seçimi	Yok				K komşu sayısı 1'den 10'a kadar belirlenmiştir.

Tablo 30. Lokalizasyon Hedef Değişkeni Analizinde Kullanılan Algoritmaların Performans Ölçüleri

Naive Bayes	2-kat	3-kat	4-kat	5-kat	6-kat	7-kat	8-kat	9-kat	10-kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma	27,8%	40,0%	36,7%	46,7%	40,0%	43,3%	38,9%	43,3%	45,6%	31,1%	54,6%	66,7%	55,6%
Kappa İstatistiği	-0,10	0,09	0,03	0,19	0,08	0,14	0,07	0,13	0,17	0,01	0,34	0,52	0,40
Duyarlılık	0,28	0,40	0,37	0,47	0,40	0,43	0,39	0,43	0,46	0,31	0,55	0,67	0,56
Kesinlik	0,27	0,39	0,36	0,46	0,39	0,42	0,37	0,42	0,45	0,38	0,64	0,87	0,85
F Ölçütü	0,27	0,39	0,36	0,45	0,39	0,42	0,37	0,41	0,44	0,30	0,55	0,68	0,52
ROC Alanı	0,45	0,56	0,46	0,59	0,52	0,52	0,50	0,54	0,55	0,59	0,74	0,80	0,84
MAE	0,46	0,42	0,45	0,40	0,43	0,42	0,43	0,42	0,41	0,44	0,39	0,37	0,36
RMSE	0,56	0,50	0,54	0,48	0,51	0,50	0,51	0,50	0,49	0,51	0,45	0,42	0,44
RAE	105,5%	94,2%	101,0%	90,5%	96,5%	95,7%	97,4%	94,0%	92,9%	97,0%	87,2%	83,8%	78,0%
RRSE	120,1%	106,6%	114,7%	102,5%	108,4%	106,7%	108,9%	106,7%	104,8%	104,7%	94,4%	89,0%	88,1%
DVM	**2-kat**	**3-kat**	**4-kat**	**5-kat**	**6-kat**	**7-kat**	**8-kat**	**9-kat**	**10-kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
Doğru Sınıflandırma	34,4%	36,7%	31,1%	44,4%	37,8%	47,8%	35,6%	37,8%	34,4%	35,6%	40,9%	50,0%	55,6%
Kappa İstatistiği	0,01	0,05	-0,05	0,16	0,06	0,21	0,03	0,06	0,01	0,09	0,15	0,27	0,33
Duyarlılık	0,34	0,37	0,31	0,44	0,38	0,48	0,36	0,38	0,34	0,36	0,41	0,50	0,56
Kesinlik	0,35	0,37	0,31	0,45	0,38	0,48	0,35	0,38	0,35	0,46	0,52	0,64	0,68
F Ölçütü	0,34	0,36	0,31	0,45	0,38	0,48	0,35	0,38	0,35	0,33	0,42	0,52	0,55
ROC Alanı	0,49	0,54	0,49	0,57	0,49	0,59	0,50	0,53	0,51	0,55	0,70	0,72	0,75
MAE	0,44	0,42	0,44	0,41	0,45	0,40	0,44	0,43	0,44	0,44	0,38	0,37	0,37
RMSE	0,54	0,52	0,54	0,51	0,54	0,50	0,54	0,53	0,53	0,53	0,48	0,46	0,46
RAE	100,6%	96,2%	100,7%	93,4%	101,2%	91,7%	100,1%	97,8%	99,4%	9816,0%	85,7%	82,8%	80,0%
RRSE	115,3%	111,1%	115,4%	108,7%	115,8%	106,7%	114,6%	112,8%	113,6%	109,8%	99,4%	96,0%	93,6%

Tablo 31. (Devam) Lokalizasyon Hedef Değişkeni Analizinde Kullanılan Algoritmaların Performans Ölçüleri

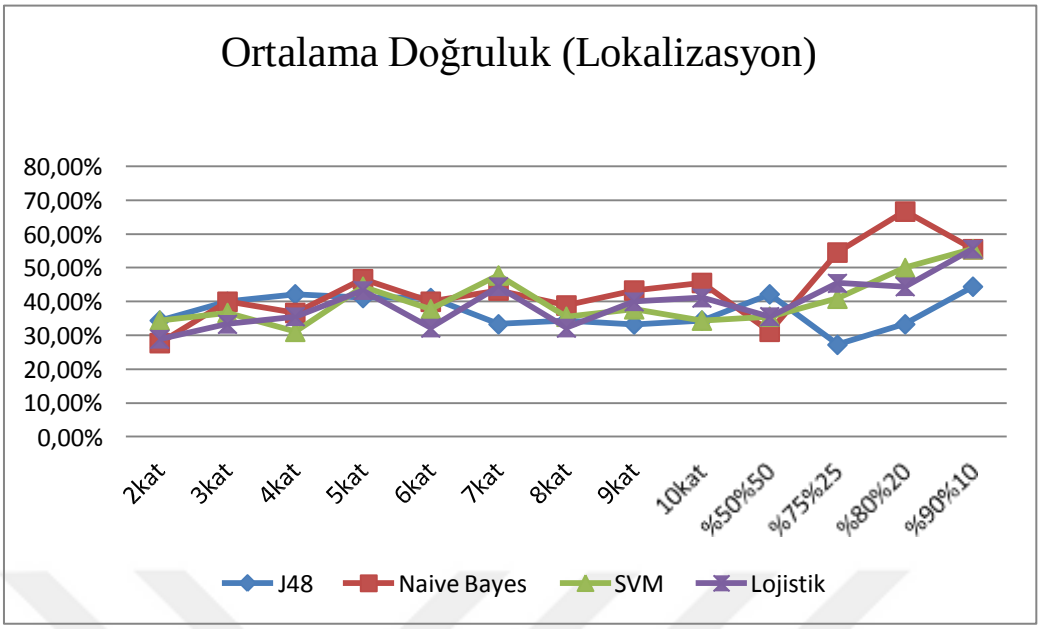
Lojistik	2-kat	3-kat	4-kat	5-kat	6-kat	7-kat	8-kat	9-kat	10-kat	%50%50	%75%25	%80%20	%90%10
Doğru Sınıflandırma	28,9%	33,3%	35,6%	43,3%	32,2%	44,4%	32,2%	40,0%	41,1%	35,6%	45,5%	44,4%	55,6%
Kappa İstatistiği	-0,08	-0,01	0,02	0,14	-0,03	0,15	-0,02	0,09	0,11	0,05	0,18	0,17	0,40
Duyarlılık	0,29	0,33	0,36	0,43	0,32	0,44	0,32	0,40	0,41	0,36	0,46	0,44	0,56
Kesinlik	0,29	0,33	0,35	0,43	0,31	0,44	0,32	0,40	0,41	0,37	0,50	0,53	0,79
F Ölçütü	0,28	0,33	0,35	0,43	0,32	0,44	0,32	0,40	0,41	0,35	0,47	0,47	0,49
ROC Alanı	0,46	0,51	0,50	0,56	0,46	0,55	0,51	0,55	0,54	0,53	0,68	0,64	0,67
MAE	0,47	0,44	0,43	0,40	0,45	0,41	0,43	0,42	0,42	0,45	0,37	0,38	0,34
RMSE	0,68	0,60	0,58	0,52	0,58	0,53	0,55	0,54	0,53	0,64	0,49	0,50	0,47
RAE	106,7%	99,0%	98,1%	90,8%	102,3%	93,4%	98,4%	94,0%	94,0%	98,3%	8340,0%	86,0%	72,9%
RRSE	145,4%	128,1%	123,4%	111,6%	123,0%	112,3%	117,7%	114,5%	112,4%	132,7%	101,6%	104,5%	94,1%
J48	**2-kat**	**3-kat**	**4kat**	**5-kat**	**6-kat**	**7-kat**	**8-kat**	**9-kat**	**10-kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
Doğru Sınıflandırma	34,4%	40,0%	42,2%	41,1%	41,1%	33,3%	34,4%	33,3%	34,4%	42,2%	27,3%	33,3%	44,4%
Kappa İstatistiği	0,01	0,08	0,12	0,10	0,11	-0,02	0,00	-0,02	0,00	0,15	0,01	0,09	0,29
Duyarlılık	0,34	0,40	0,42	0,41	0,41	0,33	0,34	0,33	0,34	0,42	0,27	0,33	0,44
Kesinlik	0,35	0,38	0,41	0,38	0,41	0,31	0,34	0,31	0,34	0,47	0,52	0,51	0,29
F Ölçütü	0,35	0,39	0,41	0,39	0,41	0,32	0,34	0,32	0,34	0,42	0,24	0,28	0,32
ROC Alanı	0,51	0,54	0,52	0,55	0,55	0,49	0,51	0,49	0,51	0,57	0,53	0,52	0,74
MAE	0,44	0,42	0,43	0,41	0,42	0,45	0,44	0,44	0,43	0,40	0,44	0,44	0,38
RMSE	0,57	0,56	0,54	0,52	0,53	0,53	0,55	0,54	0,52	0,57	0,51	0,54	0,44
RAE	98,8%	96,1%	96,8%	93,3%	94,7%	101,7%	99,4%	99,8%	96,9%	88,9%	97,2%	98,7%	82,7%
RRSE	121,3%	118,5%	115,2%	110,3%	113,4%	111,9%	116,6%	113,9%	111,4%	117,0%	107,5%	113,3%	89,5%

Lokalizasyon hedef değişkenine uygulanan Naive Bayes Algoritması için elde edilen en yüksek sınıflandırma %80%20 tabakalı Hold-Outta %66,7 olarak bulunmuştur. Kappa istatistiği en yüksek 0,52 olarak %80%20 tabakalı Hold-Out ile gerçekleşmiştir. Elde ettiğimiz sonuçlar incelendiğinde en yüksek kesinlik değeri %80%20 tabakalı Hold-Out ile 0,87 olarak bulunmuştur. Duyarlılık değeri en yüksek 0,67 ile %80%20 tabakalı Hold-Outta görülmüştür. En yüksek F-ölçütü ise 0,68 olarak %80%20 tabakalı Hold-Out ile elde edilmiştir.

DVM algoritması ile elde edilen en yüksek doğru sınıflandırma %90%10 tabakalı Hold-Out ile %55,6 olarak gerçekleşmiştir. Kappa istatistiği en yüksek 0,33 olarak %90%10 tabakalı Hold-Out ile bulunmuştur. Kesinlik değeri en yüksek 0,68 ile %90%10 tabakalı Hold-Outta görülmüştür. Duyarlılık değeri en yüksek 0,56 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. En yüksek F-Ölçütü ise 0,55 olarak %90%10 tabakalı Hold-Out ile elde edilmiştir.

Lojistik Regresyon ile elde edilen en yüksek sınıflandırma %90%10 tabakalı Hold-Out ile %55,6 olarak gerçekleşmiştir. Kappa istatistiği en yüksek 0,4 olarak %90%10 tabakalı Hold-Out ile elde edilmiştir. Kesinlik değeri en yüksek 0,79 ile %90%10 tabakalı Hold-Outta bulunmuştur. Duyarlılık değeri 0,56 olarak %90%10 tabakalı Hold-Outta görülmüştür. En yüksek F Ölçütü ise 0,49 ile %90%10 tabakalı Hold-Outta bulunmuştur.

J48 algoritması ile elde edilen en yüksek doğru sınıflandırma %90%10 tabakalı Hold-Out ile %44,4 olarak bulunmuştur. Kappa istatistiği en yüksek ise 0,29 olarak %90%10 tabakalı Hold-Out ile gerçekleşmiştir. Kesinlik değeri en yüksek 0,52 ile %75%25 tabakalı Hold-Outta görülmüştür. Duyarlılık değeri en yüksek 0,44 ile %90%10 tabakalı Hold-Outta bulunmuştur. En yüksek F-Ölçütü ise 0,42 olarak %50%50 tabakalı Hold-Out ile elde edilmiştir. Şekil 4.11'de lokalizasyon hedef değişkenine J48 algoritmasının uygulanması ile elde edilen en yüksek doğru sınıflandırma oranına sahip %90%10 tabakalı Hold-Out için oluşturulan karar ağaç yapısı verilmektedir.



Şekil 13. Lokalizasyon Hedef Değişkenine Ait Ortalama Doğruluk

Lokalizasyon hedef değişkenine tüm algoritmaların uygulanması sonucu elde edilen ortalama doğruluk oranları karşılaştırılmıştır. Naive Bayes algoritması için elde edilen en yüksek sınıflandırma performansı %66,7 ile %80%20 tabakalı Hold-Out ile elde edilmiştir. DVM algoritması için en yüksek sınıflandırma %55,6 ile %90%10 tabakalı Hold-Outta, Lojistik Regresyon için %55,6 ile %90%10 tabakalı Hold-Outta ve J48 algoritması için %44,4 ile %90%10 tabakalı Hold-Outta gerçekleşmiştir. Bu sonuçlar doğrultusunda, lokalizasyon hedef değişkeni için en yüksek sınıflandırma performansını Naive Bayes algoritması sağlamıştır. En düşük sınıflandırma performansı ise J48 algoritması ile elde edilmiştir. Sonuçların genel seyrine bakıldığında Lokalizasyon hedef değişkenine ait doğru sınıflandırma oranlarının oldukça düşük olduğu, tanımlayıcı değişkenlerin hedef değişkeni sınıflandırmak için yeterli olmadığı görülmüştür.

Lokalizasyon hedef değişkeni için K-En Yakın Komşu algoritmasının uygulanması aşamasında, en yüksek k parametresini elde etmek için performans değerlendirme yöntemi olarak 5-kat, 7-kat ve 10-kat çapraz geçerleme kullanılmıştır. Ayrıca k'nın en yüksek değere ulaşabilmesi için 1'den 10'a kadar ayrı ayrı uygulanmıştır.

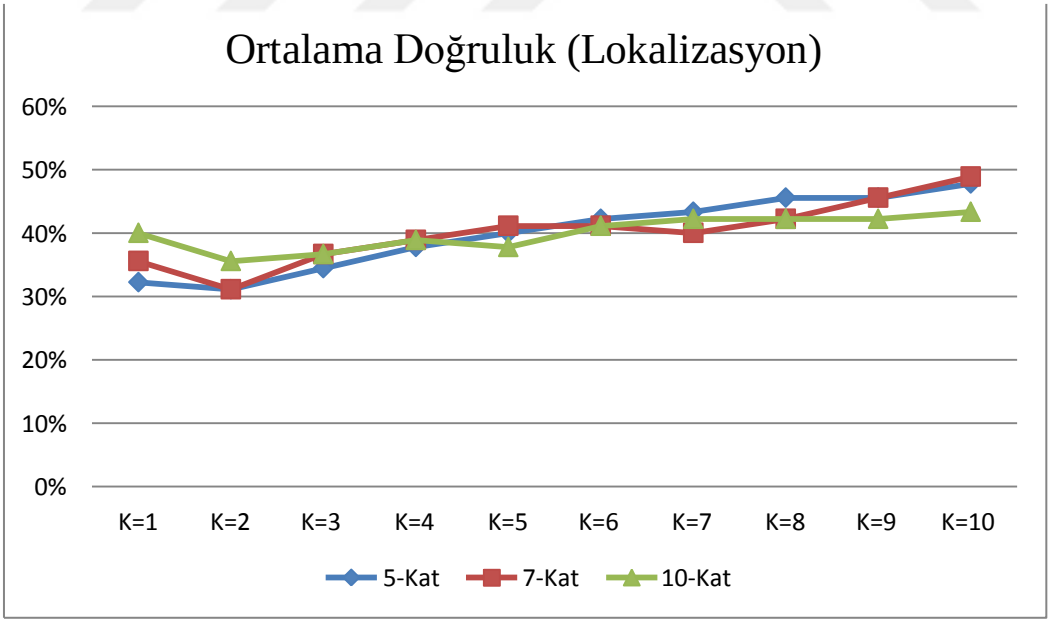
Tablo 32. Lokalizasyon Hedef Değişkeni Analizinde Kullanılan K-En Algoritmasına Ait Performans Ölçüleri

Tabakalı 5-Kat Çapraz Geçerleme										
i	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Doğru Sınıflandırma	32,2%	31,1%	34,4%	37,8%	40,0%	42,2%	43,3%	45,6%	45,6%	47,8 %
Kappa İstatistiği	-0,02	-0,04	0,01	0,06	0,09	0,13	0,15	0,18	0,18	0,21
Duyarlılık	0,32	0,31	0,34	0,38	0,40	0,42	0,43	0,46	0,46	0,48
Kesinlik	0,33	0,31	0,32	0,36	0,39	0,42	0,43	0,45	0,45	0,47
F Ölçütü	0,32	0,31	0,33	0,37	0,39	0,42	0,43	0,45	0,45	0,47
ROC Area	0,54	0,56	0,55	0,55	0,56	0,58	0,57	0,60	0,60	0,62
MAE	0,42	0,41	0,42	0,42	0,42	0,42	0,42	0,42	0,42	0,41
RMSE	0,59	0,52	0,50	0,49	0,48	0,48	0,48	0,47	0,47	0,46
RAE	95,1%	93,7%	94,8%	95,2%	94,6%	94,5%	95,9%	94,7%	94,8%	93,7 %
RRSE	124,6 %	110,2 %	105,4 %	103,9 %	102,8 %	101,1 %	101,7 %	94,7%	99,0%	97,6 %
Tabakalı 7-Kat Çapraz Geçerleme										
K	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Doğru Sınıflandırma	35,6%	31,1%	36,7%	38,9%	41,1%	41,1%	40,0%	42,2%	45,6%	48,9 %
Kappa İstatistiği	0,03	-0,03	0,04	0,08	0,11	0,11	0,09	0,13	0,17	0,22
Duyarlılık	0,36	0,31	0,37	0,39	0,41	0,41	0,40	0,42	0,46	0,49
Kesinlik	0,37	0,32	0,37	0,39	0,41	0,41	0,40	0,42	0,45	0,48
F Ölçütü	0,36	0,31	0,37	0,39	0,41	0,41	0,40	0,42	0,45	0,48
ROC Area	0,52	0,49	0,53	0,55	0,55	0,55	0,57	0,57	0,57	0,60
MAE	0,43	0,44	0,43	0,43	0,43	0,43	0,43	0,43	0,43	0,42
RMSE	0,58	0,54	0,50	0,49	0,49	0,49	0,48	0,48	0,47	0,46
RAE	96,3%	100,0 %	97,3%	96,5%	96,6%	96,7%	96,7%	96,7%	96,3%	95,1 %
RRSE	123,0 %	113,8 %	107,1 %	104,7 %	103,9 %	103,6 %	102,3 %	101,9 %	100,9 %	98,9 %
Tabakalı 10-Kat Çapraz Geçerleme										
K	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Doğru Sınıflandırma	40,0%	35,6%	36,7%	38,9%	37,8%	41,1%	42,2%	42,2%	42,2%	43,3 %
Kappa İstatistiği	0,10	0,03	0,04	0,08	0,06	0,11	0,13	0,13	0,13	0,14
Duyarlılık	0,40	0,36	0,37	0,39	0,38	0,41	0,42	0,42	0,42	0,43
Kesinlik	0,41	0,36	0,36	0,39	0,37	0,41	0,42	0,42	0,42	0,43
F Ölçütü	0,40	0,36	0,37	0,39	0,38	0,41	0,42	0,42	0,42	0,43
ROC Area	0,55	0,54	0,53	0,54	0,53	0,56	0,56	0,56	0,58	0,60
MAE	0,41	0,42	0,42	0,42	0,43	0,42	0,42	0,43	0,42	0,42
RMSE	0,57	0,51	0,50	0,49	0,49	0,48	0,48	0,48	0,47	0,46
RAE	92,5%	94,7%	95,9%	95,3%	96,4%	95,6%	96,1%	96,6%	95,5%	94,7 %
RRSE	120,2 %	109,4 %	105,9 %	103,7 %	104,3 %	102,1 %	101,9 %	101,8 %	100,3 %	98,8 %

Lokalizasyon hedef değişkenine K-En yakın komşu algoritmasının uygulanması ile elde edilen sonuçlara bakıldığında 5-Kat çapraz geçerleme için en yüksek sınıflandırma k=10 iken % 47,8 olarak gerçekleşmiştir. Kappa istatistiği en yüksek k=10 için 0,21 olarak görülmüştür. Duyarlılık değeri en yüksek k=10 iken 0,48 olarak bulunmuştur. Kesinlik değeri en yüksek k=10 için 0,47 olarak gerçekleşmiştir. En yüksek F- Ölçütü ise k=10 iken 0,47 olarak elde edilmiştir.

Tabakalı 7–kat çapraz geçerleme için en yüksek sınıflandırma k=10 iken %48,9 olarak görülmüştür. Kappa istatistiği en yüksek k=10 iken 0,22 olarak gerçekleşmiştir. Duyarlılık değeri en yüksek k=10 için 0,49 olarak bulunmuştur. Kesinlik değeri en yüksek k=10 iken 0,48 olarak saptanmıştır. En yüksek F-Ölçütü ise k=10 iken 0,48 olarak elde edilmiştir.

Tabakalı 10-kat çapraz geçerleme için en yüksek sınıflandırma performansı k=10 iken %43,3 olarak gerçekleşmiştir. Kappa istatistiği en yüksek k=10 için 0,14 olarak saptanmıştır. Duyarlılık değeri en yüksek k=10 iken 0,43 olarak bulunmuştur. Kesinlik değeri en yüksek k=10 için 0,432 olarak görülmüştür. En yüksek F-Ölçütü ise k=10 iken 0,43 olarak elde edilmiştir.



Şekil 14. Sağ Kalım Hedef Değişkenine Ait K-En Algoritması ile Elde Edilen Ortalama Doğruluk

Şekil 14'de yer alan grafik incelendiğinde, en yüksek doğru sınıflandırma performansı 7-kat çapraz geçerleme ile % 48,9 olarak görülmüştür. Elde edilen doğru sınıflandırma performanslarının tümü %50'nin altında gerçekleşmiştir.

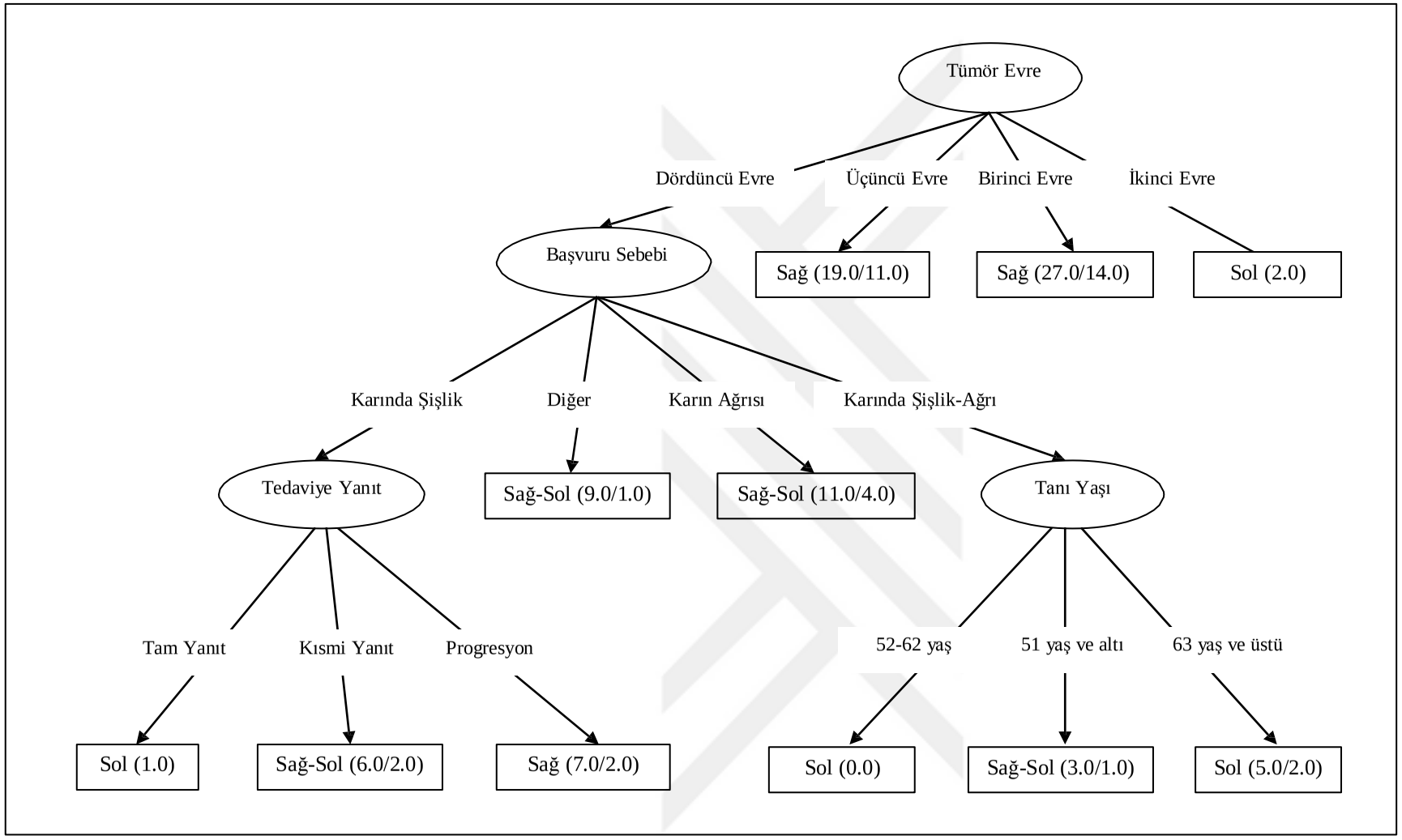
Tablo 33. Lokalizasyon Hedef Değişkeni İçin Elde Edilen Ölçütlerin Karşılaştırılması

Algoritma / Ölçüt	Naive Bayes	DVM	Lojistik Regresyon	J48 Karar Ağacı	k-En Yakın Komşu
Doğru Sınıflandırma Oranı	66,7%	%55,6	%55,6	%44,4	%48,9
F-Ölçütü	0,68	0,55	0,49	0,42	0,48

Tablo 33'de yer alan sonuçlar incelendiğinde doğru sınıflandırma ölçütüne göre en iyi performans Naive Bayes algoritmasında %72,2 doğruluk oranı ile gerçekleşmiştir. Bu ölçüte göre Naive Bayes algoritmasından sonra en iyi sınıflandırma oranını DVM ve Lojistik Regresyon algoritmaları sunmaktadır. Daha sonra sırası ile K-En Yakın Komşu ve J48 karar ağacı algoritmaları gelmektedir.

F-Ölçütünün incelenmesi, bu ölçütün hesaplanmasında kullanılan kesinlik ve duyarlılık ölçütlerinin birlikte incelenmesine olanak sağlayacaktır. Bu yüzden F-Ölçütünün karşılaştırma ölçütü olarak kullanılmasının daha kapsayıcı olacağı düşünülmüştür. F-Ölçütüne baktığımızda en yüksek sonuç Naive Bayes algoritması ile gerçekleşmiştir. F-Ölçütüne göre Naive Bayes algoritmasını sırasıyla, DVM, Lojistik Regresyon, K-En Yakın Komşu ve J48 Karar ağaç algoritmaları takip etmektedir.

Şekil 15'te tedaviye yanıt hedef değişkenine J48 algoritmasının uygulanması ile elde edilen en yüksek doğru sınıflandırma oranına sahip 3-kat çapraz geçerleme için çizilen karar ağaç yapısı verilmiştir.



Şekil 15. Lokalizasyon Hedef Değişkenine Ait J48 Ağaç Yapısı

Tümörün Lokalizasyonunu tahmin etmek için j48 algoritmasının çizmiş olduğu ağaç yapısı Şekil 4.12'de görülmektedir. Ağaçlandırma tanımlayıcı değişkenlerden tümörün evresi ile başlamıştır. Tümör birinci ve üçüncü evre ise tümörün lokalizasyonu sağ, ikinci evre ise tümörün lokalizasyonu sol olarak görülmektedir. Eğer tümör dördüncü evrede ise başvuru sebebine bakılır. Başvuru sebebi eğer karın ağrısı ve diğer ise lokalizasyon sağ-sol olarak gerçekleşmektedir. Başvuru sebebi eğer karında şişlik-ağrı ise tanı yaşına, sadece karında şişlik olarak gerçekleşir ise tedaviye vermiş olduğu yanıta bakılır.

4.4. VERİ MADENCİLİĞİ SINIFLANDIRMA ALGORİTMALARININ ELDE ETTİĞİ BULGULAR

Tablo 34. Yöntemlerin Analiz Özeti

Veri Seti	Overca_csv		
Hedef Değişken	Tedaviye Yanıt (Tam Yanıt/Kısmi Yanıt/Progresyon)	Sağkalım (1080 gün ve altı/1081 gün ve üstü)	Tümörün Lokalizasyonu (Sol/Sağ/ Sol ve Sağ)
Tanımlayıcı Değişken	Tanı Yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümör Patolojisi, Kemoterapi, Nüks, Tümörün Lokalizasyonu, Sağkalım	Tanı Yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümör Patolojisi, Kemoterapi, Nüks, Tedaviye Yanıt, Tümörün Lokalizasyonu	Tanı Yaşı, Başvuru Sebebi, Debulking, Tümörün Evresi, Tümör Patolojisi, Kemoterapi, Nüks, Tedaviye Yanıt, Sağkalım
Performans Değerlendirme Yöntemi	• Tabakalı 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme • Tabakalı örnekleme kullanılarak %50/%50, %75/%25, %80/%20 ve %90/%10 oranlarında Hold-out		
Paremetre Seçimi	Yok		

Tablo 35. Hedef Değişkenleri İçin Elde Edilen Doğru Sınıflandırma Oranları

	Ortalama Doğruluk (Naive Bayes)												
Hedef Değişken	**2kat**	**3kat**	**4kat**	**5kat**	**6kat**	**7kat**	**8kat**	**9kat**	**10kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
T,Yanıt	75,6%	76,7%	70,0%	73,3%	75,6%	76,7%	75,6%	75,6%	75,6%	77,8%	81,8%	77,8%	88,9%
Sağkalım	65,6%	73,3%	67,8%	67,8%	68,9%	67,8%	68,9%	67,8%	70,0%	66,7%	50,0%	55,6%	66,7%
Lokalizasyon	27,80%	40,00%	36,70%	46,70%	40,00%	43,30%	38,90%	43,30%	45,60%	31,10%	54,60%	66,70%	55,60%
	Ortalama Doğruluk (SVM)												
Hedef Değişken	**2kat**	**3kat**	**4kat**	**5kat**	**6kat**	**7kat**	**8kat**	**9kat**	**10kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
T,Yanıt	72,2%	78,9%	73,3%	77,8%	72,2%	77,8%	75,6%	75,6%	77,8%	80,0%	77,3%	72,2%	77,8%
Sağkalım	67,8%	68,9%	70,0%	65,6%	63,3%	66,7%	71,1%	74,4%	68,9%	60,0%	50,0%	44,4%	66,7%
Lokalizasyon	34,40%	36,70%	31,10%	44,40%	37,80%	47,80%	35,60%	37,80%	34,40%	35,60%	40,90%	50,00%	55,60%
	Ortalama Doğruluk (Lojistik)												
Hedef Değişken	**2kat**	**3kat**	**4kat**	**5kat**	**6kat**	**7kat**	**8kat**	**9kat**	**10kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
T,Yanıt	62,2%	66,7%	61,1%	64,4%	63,3%	68,9%	68,9%	68,9%	66,7%	73,3%	68,2%	66,7%	66,7%
Sağkalım	68,9%	65,6%	65,6%	67,8%	62,2%	62,2%	71,1%	71,1%	68,9%	51,1%	54,6%	50,0%	88,9%
Lokalizasyon	28,90%	33,30%	35,60%	43,30%	32,20%	44,40%	32,20%	40,00%	41,10%	35,60%	45,50%	44,40%	55,60%
	Ortalama Doğruluk (J48)												
Hedef Değişken	**2kat**	**3kat**	**4kat**	**5kat**	**6kat**	**7kat**	**8kat**	**9kat**	**10kat**	**%50%50**	**%75%25**	**%80%20**	**%90%10**
T,Yanıt	70,0%	73,3%	73,3%	76,7%	74,4%	80,0%	76,7%	77,8%	76,7%	77,8%	72,7%	66,7%	77,8%
Sağkalım	63,3%	71,1%	68,9%	63,3%	60,0%	62,2%	70,0%	64,4%	64,4%	60,0%	54,6%	55,6%	44,4%
Lokalizasyon	34,40%	40,00%	42,20%	41,10%	41,10%	33,30%	34,40%	333,30%	34,40%	42,20%	27,30%	33,30%	44,40%

Tablo 35’te bulunan doğru sınıflandırma performanslarının karşılaştırılması sonucunda, Naive Bayes Algoritması için en yüksek performans %88,9 ile tedaviye yanıt hedef değişkeninde bulunmuştur. DVM algoritması için en yüksek performans %80 ile tedaviye yanıt hedef değişkeninde elde edilmiştir. Lojistik Regresyon için elde edilen en yüksek performans %88,9 ile sağ kalım hedef değişkeninde gerçekleşmiştir. J48 Algoritması için en yüksek sınıflandırma performansı ise %80 ile tedaviye yanıt hedef değişkeninde bulunmuştur. Tüm algoritmalar için en düşük sınıflandırma performansı ise lokalizasyon hedef değişkeninde gerçekleşmiştir.

SONUÇ VE ÖNERİLER

Bu tez çalışmasında, yumurtalık kanseri teşhisi konulan hastalara ait değişkenler, veri madenciliği sınıflandırma algoritmaları kullanılarak analiz edilmiştir. Veri seti Recep Tayyip Erdoğan Üniversitesi Eğitim ve Araştırma Hastanesinden elde edilmiştir. Çalışmada kullanılan veri setine ait hedef ve tanımlayıcı değişkenler uzman görüşü ile belirlenmiştir. Hedef değişkenlerine uygulanan veri madenciliği algoritmalarının performans karşılaştırmaları ise doğru sınıflandırma oranı ve F-Ölçütü ile gerçekleştirilmiştir.

Tedaviye yanıt, sağ kalım, lokalizasyon hedef değişkenlerinin analizi için Naive Bayes, DVM, Lojistik Regresyon, J48 ve K-En Yakın Komşu algoritmalarından faydalanılmıştır. Performans değerlendirme yöntemi olarak 2-kat, 3-kat, 4-kat, 5-kat, 6-kat, 7-kat, 8-kat, 9-kat ve 10-kat çapraz geçerleme ve tabakalı %50/%50, %75/%25, %80/%20 ve %90/%10 Hold-Out kullanılmıştır. K-En Yakın Komşu algoritmasının uygulanması aşamasında k'nın en yüksek değer alabilmesi için 1'den 10'a kadar ayrı ayrı uygulanmıştır.

Tedaviye yanıt hedef değişkeni için en yüksek sınıflandırma performansı Naive Bayes algoritması ile %88,9 olarak gerçekleşmiştir. Naive Bayes algoritması için en yüksek F-ölçütü ise 0,79 olarak bulunmuştur. En düşük sınıflandırma performansı ise Lojistik Regresyon algoritması ile elde edilmiştir. DVM ve J48 algoritması ise aynı performansı sergileyerek, Naive Bayes algoritmasından sonra en iyi performansı sunmuşlardır. J48 algoritmasında, en yüksek performansın sağlandığı 7-kat çapraz geçerleme için karar ağacı çizilmiştir. Elde edilen karar ağacı modeli ile tedaviye yanıt hedef değişkeni için en ayırt edici değişken ağaçlandırmanın kök düğümü olan nüks değişkeni olarak belirlenmiştir. Bunun dışında tümörün patolojisi, debulking, tümörün evresi ve başvuru sebebi değişkenleri oluşan karar ağaç modeline en çok katkı sağlayan diğer değişkenler olarak görülmüştür.

Sağ kalım hedef değişkeni için en yüksek sınıflandırma performansı Lojistik Regresyon algoritması ile %88,9 olarak gerçekleşmiştir. Lojistik Regresyon için F-Ölçütü ise 0,90 olarak bulunmuştur. Doğru sınıflandırma performansına göre Lojistik Regresyon algoritmasını sırasıyla DVM, K-En Yakın Komşu, Naive Bayes, J48 algoritmaları takip etmektedir. J48 algoritmasında, en yüksek performansın sağlandığı 3-kat çapraz geçerleme için karar ağacı oluşturulmuştur. Elde edilen karar ağacı modeli ile sağ kalım hedef değişkeni için en ayırt edici değişken ağaçlandırmanın kök düğümü olan tedaviye yanıt değişkeni olarak saptanmıştır. Ayrıca tümörün evresi, kemoterapi, tümörün patolojisi ve başvuru sebebi değişkenleri de oluşan karar ağaç modeli için doğrudan katkı sağlayan diğer değişkenler olarak tespit edilmiştir.

Lokalizasyon hedef değişkeni için en yüksek sınıflandırma performansı Naive Bayes algoritması ile % 66,7 olarak saptanmıştır. En düşük sınıflandırma performansı ise K-En Yakın Komşu algoritması ile elde edilmiştir. Lokalizasyon hedef değişkeni için elde edilen doğru sınıflandırma performanslarının oldukça düşük seyrettiği, tanımlayıcı değişkenlerin lokalizasyon hedef değişkenini sınıflandırmak için yeterli olmadığı görülmüştür.

Çalışmada, lokalizasyon hedef değişkeni haricinde diğer hedef değişkenler için elde edilen doğru sınıflandırma oranlarının kabul edilebilir bir başarıya sahip olduğu ortaya çıkmaktadır. Tedaviye yanıt ve Lokalizasyon hedef değişkenleri için en yüksek doğru sınıflandırma oranı Naive Bayes Algoritması ile gerçekleşmiştir. Sağ kalım hedef değişkeni için en yüksek doğru sınıflandırma oranı ise Lojistik regresyon ile elde edilmiştir. Ayrıca uygulanan her sınıflandırma algoritması için hedef değişkenlerine ait doğru sınıflandırma oranları karşılaştırılmıştır. Buna göre; tüm sınıflandırma algoritmalarında en yüksek sınıflandırma performansı tedaviye yanıt hedef değişkeninde görülmüştür. Tüm algoritmaların en düşük sınıflandırma performansları ise lokalizasyon hedef değişkeninde gerçekleşmiştir.

Yumurtalık kanseri ile ilgili yapılan çalışmalara bakıldığında, nüks değişkeninin hedef değişken olarak belirlendiği Tseng vd. (2017)'in yaptığı çalışmada FIGO, Patolojik M, Yaş ve Patolojik T faktörlerinin etkin olduğunu, bu tez çalışmasında ise nüks değişkeninin sınıflandırılmasında, kullanılan diğer özniteliklerin etkin olmadığı görülmüştür. Bu durumda nüks değişkeninin sınıflandırılması için daha ileri tıbbi değişkenlere ihtiyaç olduğu söylenebilir. Osmanovic vd. (2017)'nin çalışmalarında, tümörün patolojisi (kanserin hareketliliği, kanserin yüzeyi ve kanserin tutarlılığı) özniteliklerinin, yumurtalık kanseri hastalarının hayatta kalıp kalamayacaklarını tahmin etmekte önemli ölçüde katkı sağlaması bu çalışma ile benzerlik göstermektedir.

Daha sonra yapılacak olan çalışmalarda, örneklem genişletilerek farklı hastanelerden elde edilen veriler ile bölgeler kıyaslanabilir. Çalışmada kullanılan nüks, tedaviye yanıt, sağ kalım, lokalizasyon hedef değişkenleri için farklı tanımlayıcı değişkenler kullanılarak doğru sınıflandırma oranları yeniden incelenebilir. Uygulanan veri madenciliği algoritmaları dışında veri madenciliği algoritmalarının hibrit biçimde uygulanması ile daha yüksek doğru sınıflandırma oranları elde edilebilir. Ayrıca çalışmada diğer veri madenciliği yöntemleri olan kümeleme ve birliktelik analizleri de kullanılabilir.

KAYNAKÇA

Akgöbek, Ö., & Kaya S., Veri Madenciligi Teknikleri İle Veri Kümelerinden Bilgi Keşfi: Medikal Veri Madenciliği Uygulaması. *Engineering Sciences*, *6*(1), 237-245.

Alaybeyoglu, A., & Mulayim, N. Karaciğer Kanseri Teşhisinde Destek Vektör Makinesi Tabanlı Uzman Sistem Tasarımı Support Vector Machine Based Expert System Design For Liver Disorder Diagnosis.

Aydemir, B. (2017). *Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Akademik Başarım Tahmini.* (Yüksek Lisans Tezi). Https://Tez.Yok.Gov.Tr/Ulusaltezmerkezi/ adresinden edinilmiştir.

Balaban, M. E. ve Kartal, E. (2018). *Veri Madenciliği Ve Makine Öğrenmesi: Temel Algoritmalar Ve R Dili İle Uygulamaları.* İstanbul: Çağlayan.

Balaban, E., Kartal, E. (2019) Veri Madenciliği Ve Makine Öğrenmesi: Temel Kavramlar, Algoritmlar, Uygulamalar.

Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management*. John Wiley & Sons.

Bircan, H. (2004). Lojistik Regresyon Analizi: Tıp Verileri Üzerine Bir Uygulama.

Can, M. B., Eren, Ç., Koru, M., Özkan, Ö., & Rzayeva, Z. (2012). Veri Kümelerinden Bilgi Keşfi: Veri Madenciliği. *Başkent Üniversitesi Tıp Fakültesi Xıv. Öğrenci Sempozyumu, Ankara*.

Cihan, A. (2009). *Pozitif Ve Negatif İlişkilerin Veri Madenciliğiyle Belirlenmesine Yönelik Bir Model* (Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü).

Coşkun, C. ve Baykal, A. (2011). Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. *Akademik Bilişim*, 1-8.

Czepiel, S. A. (2002). Maximum Likelihood Estimation Of Logistic Regression Models: Theory And İmplementation. *Available At Czep. Net/Stat/Mlelr. Pdf*, 1825252548-1564645290.

Çokluk, Ö. (2010). Lojistik Regresyon Analizi: Kavram Ve Uygulama. *Kuram Ve Uygulamada Eğitim Bilimleri*, *10*(3), 1357-1407.

Darakçı, H. Ç. (2011). *Kümeleme Analizi Kullanılarak Benzin İstasyonlarının Operasyonel Değerlendirilmesi* (Yüksek Lisans Tezi, Maltepe Üniversitesi, Fen Bilimleri Enstitüsü).

Demircioğlu, H. Ve Bilge, H. (2015). Yumurtalık Kanseri Veri Kümesindeki Gen İfadelerinin Veri Madenciliği İle Analizi. *Marmara Fen Bilimleri Dergisi*, *27*(4), 125-134.

Dönmez, Z. S. (2008). *Bayi Performans Değerlendirmesinde Bir Veri Madenciliği Uygulaması* (Doktora Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü).

El-Bendary, N., & Belal, N. A. (2018). Epithelial Ovarian Cancer Stage Subtype Classification Using Clinical And Gene Expression Integrative Approach. *Procedia Computer Science*, *131*, 23-30.

Ergün, A., Dilek, S., & Atay, V. (1996). Over Kanseri Taraması. *Journal Of Clinical Obstetrics & Gynecology*, *6*(3), 274-276.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining To Knowledge Discovery İn Databases. *Al Magazine*, 37-54.

Frawley W. J., Piatetsky-Shapiro G., Matheus C. J., "Knowledge Discovery İn Databases : An Overview", 13(3): 57–70, 1992.

Güldal, D., Soylu, F., Çamlı, L., & Edirne, T. (2007). Over Tümörlerinin Epidemiyolojik Özellikleri. *Türkiye Aile Hekimliği Dergisi*, *6*(1), 24-28.

Güzel, D., Yıldırım, N., Besler, A., Akman, L., Özdemir, N., Zekioğlu, O., & Terek, M. C. (2019). Over Kanserinin Epidemiyolojisi Ve Genel Sağ Kalım Özellikleri. *Ege Tıp Dergisi*, *58*, 44-49.

Hambali, M., & D Gbolagade, M. (2016). Ovarian Cancer Classification Using Hybrid Synthetic Minority Over-Sampling Technique And Neural Network. *Journal Of Advances In Computer Research*, *7*(4), 109-124.

Han, J., Kamber, M., ve Pei, J., (2011) "Data Transformation By Normalization".

Kartal, E., Programı, E., & Balaban, M. E. Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri Ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama.

Kaufman, L. Ve Rousseeuw, P. J., (1990), *Finding Groups In Data: An Introduction To Cluster Analysis*, John Wiley & Sons, Inc. Hoboken, New Jersey, Isbn: 0-471-73578-7.

Kurman, R. J., Visvanathan, K., Roden, R., Wu, T. C., & Shih, I. M. (2008). Early Detection And Treatment Of Ovarian Cancer: Shifting From Early Stage To Minimal Volume Of Disease Based On A New Model Of Carcinogenesis. *American Journal Of Obstetrics And Gynecology*, *198*(4), 351-356.

Küçüksille, E. (2009). *Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İmkb Hisse Senetleri Piyasasında Bir Uygulama* (Doktora Tezi, Sosyal Bilimler).

Kresse, W., & Danko, D. M. (Eds.). (2012). *Springer Handbook Of Geographic İnformation*. Springer Science & Business Media.

Lakshmanan B. C., Srinivasan, V., Ponnuraja C., "Data Mining With Decision Tree To Evaluate The Pattern On Effectiveness Of Treatment For Pulmonary Tuberculosis : A Clustering And Classification Techniques", 3(6): 43–48, 2015.

Landis, J.Richard, G.G. Koch (1977). "The Measurement Of Observer Agreement For Categorical Data", *Biometrics*, C. 33, 59–174.

Larose, D. T. (2005). An Introduction To Data Mining. *Traduction Et Adaptation De Thierry Vallaud*.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge In Data: An İntroduction To Data Mining* (Vol. 4). John Wiley & Sons.

Lu, L., & Zhu, Z. (2014). Prediction Model For Eating Property Of İndica Rice. *Journal Of Food Quality*, *37*(4), 274-280.

Nizam, H. ve Akın, S. S. (2014). *Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması*, XIX. Türkiye'de İnternet Konferansı, İzmir.

Nuhić, J., Spahić, L., Ćordić, S., & Kevrić, J. (2019). Comparative Study On Different Classification Techniques For Ovarian Cancer Detection. In *International Conference On Medical And Biological Engineering* (Pp. 511-518). Springer, Cham.

Onan, A. (2015). Şirket İflaslarının Tahmin Edilmesinde Karar Ağacı Algoritmalarının Karşılaştırmalı Başarım Analizi. *Bilişim Teknolojileri Dergisi*, *8*(1), 9-19.

Özbay Ö. (2015). "Veri Madenciliği Kavramı Ve Eğitimde Veri Madenciliği Uygulamaları", 262–272,.

Özkan, Y. (2016). *Veri Madenciliği Yöntemleri*. İstanbul: Papatya Yayıncılık Eğitim.

Poyraz, O. (2012). Tıpta Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi.

Ramkumar, G. D., Swami, A. N., & America, H. (1998). Clustering Data Without Distance Functions. *Ieee Data Eng. Bull.*, *21*(1), 9-14.

Ribeiro R., Marinho R. , Velosa J. , Ramalho F., Sanches J.M., (2011) " Diffuse Liver Disease Classification From Ultrasound Surface Characterization, Clinical And Laboratorial Data", *Pattern Recognition And Image Analysis* .167-175

Saptono, R., Winarno E. ve Drajati, D. P. (2018), "Comparison Of Under-Sampling And Oversampling Techniques İn Diabetic Mellitus (Dm) Patient Data Classification By Using Naive Bayes Classifier (Nbc)", *Journal Of Telecommun. Electron. Comput. Eng.*, 10: 145–148,.

Sardouk, F. (2018). Usıng data mınıng for the classıfıcatıon of breast canser (Yüksek Lisans Tezi). https://tez.yok.gov.tr/Ulu-salTezMerkezi adresinden edinilmiştir.

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri.

Sebik, N. B., & Bülbül, H. İ. (2018). Veri Madenciliği Modellerinin Akciğer Kanseri Veri Seti Üzerinde Başarılarının İncelenmesi. *Tübav Bilim Dergisi*, *11*(3), 1-7.

Sevli, O. (2019). Göğüs Kanseri Teşhisinde Farklı Makine Öğrenmesi Tekniklerinin Performans Karşılaştırması. *Avrupa Bilim Ve Teknoloji Dergisi*, (16), 176-185.

Silahtaroğlu, G. (2016). *Veri Madenciliği Kavram Ve Algoritmaları*. İstanbul: Papatya Yayıncılık Eğitim.

Spss. (2000). Crısp-Dm 1.0. Haziran, 5 2019 Tarihinde Https://The-Modeling-Agency.Com/Crisp-Dm.Pdf Adresinden Alındı.

Şekeroğlu, S. (2010). *Hizmet Sektöründe Bir Veri Madenciliği Uygulaması* (Doctoral Dissertation, Fen Bilimleri Enstitüsü).

Şentürk, Z.K. (2011). Veri madenciliği ile kanser tanısı (Yüksek Lisans Tezi). https://tez.yok.gov.tr/Ulu-salTezMerkezi adresinden edinilmiştir.

Şentürk, A. (2006). *Veri Madenciliği: Kavram Ve Teknikler*. Ekin Yayınevi.

Şık, M. Ş. (2014). *Veri Madenciliği Ve Kanser Erken Teşhisinde Kullanımı.* (Yüksek Lisans Tezi) https://Tez.Yok.Gov.Tr/Ulusaltezmerkezi/ adresinden edinilmiştir.

Terzi, M. (2019). Data Mining Applications In Health Sector İn Turkey. *Black Sea Journal Of Health Science*, *2*(2), 45-48.

Thulasiraman, V., & Kavitha, S. (2019). Risk Prediction System Using Data Mining Techniques In Gynecological Ovarian Cancer. *ICTACT Journal On Soft Computıng*, ,9(4), 1993-1998.

Tseng, C. J., Lu, C. J., Chang, C. C., Chen, G. D., & Cheewakriangkrai, C. (2017). Integration Of Data Mining Classification Techniques And Ensemble Learning To İdentify Risk Factors And Diagnose Ovarian Cancer Recurrence. *Artificial Intelligence İn Medicine*, *78*, 47-54.

Uçar, E. (2013). Ortaöğretime Geçiş Sistemi (Oges) Yerleştime Puanlarının Uzman Sistemler İle Tahmini. (Doktora Tezi) https://Tez.Yok.Gov.Tr/Ulusaltezmerkezi adresinden edinilmiştir.

Yalçın, L. (2019). *Sağlık sektöründe veri madenciliği* (Yüksek Lisans Tezi). https://tez.yok.gov.tr/Ulu-salTezMerkezi adresinden edinilmiştir.

Yasodha, P., & Ananthanarayanan, N. R. (2018). Detecting The Ovarian Cancer Using Big Data Analysis With Effective Model. *Biomedical Research*, 29, 309-315.

Yetginler, B. (2019). *Rahim ağzı kanserinin veri madenciliği yöntemleri ile sınıflandırılması* (Yüksek Lisans Tezi). https://tez.yok.gov.tr/Ulu-salTezMerkezi adresinden edinilmiştir.

Yıldırım, P., Uludağ, M., & Görür, A. (2008). Hastane Bilgi Sistemlerinde Veri Madenciliği. *Çanakkale On Sekiz Mart Üniversitesi Akademik Bilişim*.

Yücebaş, S. (2018). Karmaşık Hastalıkların Teşhisinde Veri Madenciliği Yöntemlerinin Başarım Karşılaştırması. *Çanakkale Onsekiz Mart Üniversitesi Fen Bilimleri Enstitüsü Dergisi* , 4 (1) , 14-27 . DOI: 10.28979/comufbed.395117

Westpal, C. ve Blaxton T. (1998). *Data Mining Solutions Methods And Tools For Solving Real-World Problems*, Wiley Computer Publishing, John Wiley & Sons, Inc., , S. 14.

Zaïane, O. (1999). "Chapter I: Introduction To Data Mining", *Princ. Knowl. Discov. Databases*, Sayı June, 1–15.

ÖZ GEÇMİŞ

Adı,Soyadı	İlyas YILMAZ			
Doğum Yeri ve Yılı	01.10.1994 / Gaziantep			
Medeni Durumu	Bekar			
Bildiği Yabancı Diller ve Düzeyi	İngilizce/ A2 Düzeyi			
Öğrenim Durumu	Başlama-Bitirme Yılı		Kurum Adı	
Lisans	2013	2017	Recep Tayyip Erdogan Üniversitesi/İİBF	
Yüksek Lisans	2017	2020	Recep Tayyip ErdoğanÜniversitesi/SBE	
Doktora				
Çalıştığı Kurumlar			Başlama- Ayrılma Yılı	
1. -				
2. -				
3. -				
Üye Olduğu Bilimsel ve Mesleki Kuruluşlar	-			
Katıldığı Proje ve Toplantılar	-			
Yayınlar	-			
Aldığı Ödüller	-			
İletişim(eposta)	ilyasylmz1@gmail.com			