T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİYLE HAYVAN HASTALIKLARINDA TEŞHİS, PROGNOZ VE RİSK FAKTÖRLERİNİN BELİRLENMESİ

PINAR CİHAN

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ PROGRAMI

DANIŞMAN
PROF. DR. OYA KALIPSIZ

İSTANBUL, 2018

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ YÖNTEMLERİYLE HAYVAN HASTALIKLARINDA TEŞHİS, PROGNOZ VE RİSK FAKTÖRLERİNİN BELİRLENMESİ

Pınar CİHAN tarafından hazırlanan tez çalışması 26.03.2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof. Dr. Oya KALIPSIZ

Yıldız Teknik Üniversitesi

Eş Danışman

Doç. Dr. Erhan GÖKÇE

Kafkas Üniversitesi

Jüri Üyeleri

Prof. Dr. Oya KALIPSIZ

Yıldız Teknik Üniversitesi _____

Prof. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi _____

Prof. Dr. Selim AKYOKUŞ

Doğuş Üniversitesi _____

Doç. Dr. Mehmet Sıddık AKTAŞ

Yıldız Teknik Üniversitesi _____

Doç. Dr. Erdinç UZUN

Namık Kemal Üniversitesi _____

ÖNSÖZ

Veri madenciliği yöntemleriyle hayvan hastalıklarında teşhis, prognoz ve risk faktörlerini belirlemeyi amaçlayan bu tez çalışması Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü Doktora programında yürütülmüştür.

Bu tezin hazırlanması süresi boyunca, bilgisi, deneyimi, görüş ve önerileriyle bana her zaman destek olan değerli tez danışmanım Sayın Prof. Dr. Oya KALIPSIZ, eş danışmanım Sayın Doç. Dr. Erhan GÖKÇE, Yrd. Doç. Dr. Heysem KAYA, Arş. Grv. Zeynep Banu ÖZGER ve çalışmamı bitirmem için moral, motivasyonumu üst düzeyde tutmamı sağlayan eşim Yrd. Doç. Dr. Mehmet Timur CİHAN'a teşekkürü bir borç bilirim.

Bu tezi, hayatımın anlamları olan oğullarıma ithaf ediyorum.

Mart, 2018

Pınar CİHAN

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

x'	Normalize edilmiş veri
x_i	Girdi değeri
x_{max}	Girdideki en küçük sayı
x_{min}	Girdideki en büyük sayı
μ_i	Girdi setinin ortalaması
σ_i	Girdi setinin standart sapması
e	Doğal logaritma değeri

KISALTMA LİSTESİ

ALB	Albümin
ALA	Alimentary lymphoma
ANN	Artificial Neural Networks
AUC	ROC Eğrisinin Altında Kalan Alan (Area Under a Curve, AUC)
BA	Bayes Ağları
ÇFA	Çok Faktörlü Analiz
ÇKA	Çok Katmanlı Algılayıcı
DT	Decision Trees
DVM	Destek Vektör Makinaları
EAKA	Eğri Altında Kalan Alan
ELISA	Enzyme-Linked Immunosorbent Assay
FA	Faktör Analizi
GLM	Genelleştirilmiş Lineer Modeller
GGT	Gamma-Glutamyl transferase
IBD	Inflammatory bowel disease
IgG	İmmunoglobulin G
KA	Karar Ağacı
KEYK	K-En Yakın Komşu
KEKK	Kısmi En Küçük Kareler Regresyonu
KEKKDA	Kısmi En Küçük Kareler Diskriminant Analizi
KNN	K-Nearest Neighbors
LDA	Lineer Diskriminant Analizi
LinR	Lineer Regresyon
LogR	Lojistik Regresyon
LT	Lactoferrin
MAR	Rasgele Olarak Kayıp (Missing at Random, MAR)
MCAR	Tamamen Rasgele Olarak Kayıp (Missing Completely at Random)
MNAR	Rasgele Olmayan Kayıp (Missing Not at Random, MNAR)
NB	Naive Bayes
OKHK	Ortalama Karesel Hatanın Karekökü
RF	Random Forest
RMSE	Root Mean Square Error
RO	Rastgele Ormanlar

ROC	Receiver Operating Characteristic
RotO	Rotasyon Ormanları
RR	Relative Risk
RTF	Radyal Tabanlı Fonksiyon
RTFA	Radyal Tabanlı Fonksiyon Ağları
SB	Saf Bayes
SMOreg	Support Vector Machine for Regression
TBA	Temel Bileşen Analizi
TP	Total Protein
VTBK	Veri Tabanı Bilgi Keşfi
YAK	Yapay Arı Koloni
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

VERİ MADENCİLİĞİ YÖNTEMLERİYLE HAYVAN HASTALIKLARINDA TEŞHİS, PROGNOZ VE RİSK FAKTÖRLERİNİN BELİRLENMESİ

Pınar CİHAN

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

Tez Danışmanı: Prof. Dr. Oya KALIPSIZ

Eş Danışman: Doç. Dr. Erhan GÖKÇE

İstatistik bilimi veri analizinde yüzyıllardan beri kullanılmaktadır. Ancak veri miktarındaki devasa artış, geçmiş veri içerisinden ilgi çekici (önemsiz olmayan, gizli, önceden bilinmeyen, potansiyel olarak kullanışlı) bilginin gelecekteki eğilimini kestirmek ya da sonraki aşamalarda analiz etme ihtiyacı, temeli istatistiğe dayanan veri madenciliği kavramını ortaya çıkarmıştır. Veterinerlik alanındaki çalışmalarda hayvanlardan elde edilen veri setleri genellikle istatistiksel yöntemlerle analiz ediliyor olsa da veri madenciliği, veri analizinde gün geçtikçe popülerliğini ve işlevini arttıran bir alan olarak karşımıza çıkmaktadır. Veri madenciliği, bilgilerin analiz edilmesi ve yorumlanacak bilgiler edinmeyi sağlayan bir süreçtir. Veri yığınları içinde açık olmayan fakat anlamlı gizli örüntüleri ve işe yarar bilgileri bulmak bu yöntemler ile gerçekleştirilir.

Bu tez çalışmasında, veri madenciliği yöntemleriyle hayvan hastalıklarında teşhis, prognoz ve risk faktörlerinin belirlenmesi amaçlanmaktadır. Veri setindeki eksik değerleri tamamlamak için en başarılı eksik değer tamamlama yöntemi belirlenmiştir. Bunun için ortalama, ortanca, k en yakın komşu, mice, missforest ve geliştirilen yapay arı koloni (YAK) yöntemleri ortalama karesel hatanın karekökü (OKHK) sonuçlarına göre karşılaştırılmıştır. Karşılaştırma sonucunda en başarılı yöntem YAK olarak belirlenmiştir. Verilerin normalizasyonu aşamasında; minimum-maksimum, ondalık ölçeklendirme, z-

değeri ve sigmoid normalizasyon yöntemleri karşılaştırılmıştır. K-ortalama kümeleme sonucunda 0.735 saflık ve 0.86 entropi ile en başarılı yöntemin sigmoid olduğu tespit edilmiştir. Verilerin sınıflandırılması aşamasında; karar ağaçları (KA), saf bayes (SB), k-en yakın komşu (KEYK), yapay sinir ağları (YSA) ve rastgele orman (RO) algoritmaları karşılaştırılmıştır. Doğruluk=0.8427, dengeli doğruluk=0.7132, seçicilik=0.91, duyarlılık=0.5164, kappa=0.4304 sonuçlarıyla en başarılı yöntemin Saf bayes olduğu belirlenmiştir. Ayrıca 0.765 eğri altında kalan alan (EAKA) değeriyle yine en başarılı yöntemin saf beyes olduğu görülmüştür. Bilgi kazancı yöntemi ile özellik seçimi yapıldıktan sonra, özellik sayısı 14'den 4'e düşürüldüğünde sınıflandırma başarısının %4 yükseldiği görülmüştür. Ortak bilgi yöntemine göre neonatal dönemde ölümler için eşik seviyesi immunoglobulin-G (IgG) < 500, Gamma-Glutamyl transferase (GGT) < 500, Lactoferrin (LT) 1201-1600, Total Protein (TP) 31-40 ve Albümin (ALB) < 35 olarak belirlenmiştir. Ayrıca veteriner hekime yardımcı mobil ve masaüstü uygulama geliştirilmiştir.

Anahtar Kelimeler: Veri madenciliği, veterinerlik, bilgisayar destekli tanı, sınıflandırma algoritmaları, eksik değer tamamlama yöntemleri

ABSTRACT

DETERMINATION OF DIAGNOSIS, PROGNOSIS AND RISK FACTORS IN ANIMAL DISEASES USING BY DATA MINING METHODS

Pınar CİHAN

Department of Computer Engineering

PhD. Thesis

Adviser: Prof. Dr. Oya KALIPSIZ

Co-Adviser: Doç. Dr. Erhan GÖKÇE

Statistics has been used for centuries in data analysis. But the increase in the amount of data reveals, to predict the future trend of interesting information (Insignificant, hidden, unknown, potentially useful) from past data or to analyze at a later stage, the concept of data mining which is based on statistics, was found. In the field of veterinary research, data sets obtained from animals are often analyzed using statistical methods, regardless of data mining field's day by day increasing popularity and function in data analysis. Data mining is a process that allows information to be analyzed and acquired. Finding the hidden secret patterns and information which are not clear in data stacks is carried out using these methods.

The aim of this thesis is to determine the diagnosis, prognosis and risk factors in animal diseases using data mining methods. In order to complete missing values in the data set, the most successful missing value imputation method has been determined. For this purpose; mean, median, nearest neighbors, mice, missForest and developed artificial bee colony (ABC) imputation methods were compared according to the root mean square error (RMSE). According to the conducted comparison results, ABC method with the lowest RMSE was determined as the most successful method. During the normalization of the data; min-max, decimal scaling, z-values, and sigmoid normalization methods are compared. It is determined that the most successful method

is sigmoid normalization method with 0.735 purity and 0.86 entropy. In the process of classifying the data; decision trees (DT), naive bayes (NB), k-nearest neighbors (KNN), artificial neural networks (ANN) and random forest (RF) algorithms are compared. It was determined that the most successful method was NB with 0.8427 accuracy, 0.7132 balanced accuracy, 0.91 specificity, 0.5164 sensitivity, 0.5226 f-measure and a 0.4304 for kappa. It was also found that the most successful method with a value of 0.765 AUC is naive bayes. After conducting feature selection using information gain method, the classification accuracy increased when the number of features was reduced from 14 to 4. According to the Mutual Information method, the threshold level for deaths in the neonatal period was determined as immunoglobulin-G (IgG) < 500, Gamma-Glutamyl transferase (GGT) < 500, Lactoferrin (LT) 1201-1600, Total Protein (TP) 31-40 and Albumin (ALB) < 35. In addition, an application for both mobile and desktop platforms have been developed for veterinary medicine.

Keywords: Data mining, veterinary, computer aided diagnosis, classification algorithms, missing value imputation methods

BÖLÜM 1

GİRİŞ

Veri madenciliği, bir veri analizi tekniği olup büyük miktardaki verinin analiz edilerek gizli kalmış bilgiye ulaşılmasını sağlamaktadır. Veri madenciliği, yeni bilgiler elde etme amacıyla çeşitli metotlardan oluşmakta olup çok farklı disiplinlerde kullanılmaktadır. Bilginin işlenerek verinin anlamlanması, yorumlanması ve ileriye dönük tahminler yapılması noktasında veri madenciliği yöntemleri fayda sağlamaktadır [1].

Özellikle mühendislik alanında yoğun olarak kullanılan veri madenciliği yöntemleri, uzun yıllardır insan sağlığı alanında da kullanılmaktadır. Hastaların elektronik tıbbi kayıtları, laboratuvar test sonuçları ve hastanın demografik bilgileri gibi verileri kullanılarak birçok veri madenciliği araştırması yapılmış ve insan sağlığı çalışanlarına büyük katkılar sağlamıştır [2].

Hayvan sağlığı çalışmalarında, hayvanlardan elde edilen veri kümeleri uzun yıllardan beri istatistiksel yöntemlerle analiz edilmekte olup veri madenciliği uygulamaları sınırlı sayıda ve genellikle alan dışı araştırmacılar tarafından gerçekleştirilmiştir [3]. Ayrıca Türkiye'de veri madenciliği konusunda birçok tez yapılmıştır ancak hayvan hastalıkları konusunda yapılmış bir tez ile karşılaşılmamıştır. Oysa veri madenciliği analizlerinin uygulama alanı bulabileceği alanlardan birinin de veterinerlik/hayvancılık alanı olduğu bilinmektedir [3].

Hayvancılık alanında veri madenciliği yöntemleri; hastalık tahmini ile erken teşhis koyma, hastalıkların tespiti, hastalıkların gruplandırılması, teşhis koymada doktorlara destek sağlayacak karar destek sistemlerinin geliştirilmesi, hastalıkların tespit edilmesi gibi daha birçok konuda farklı şekillerde hayvancılık alanına dolayısıyla da Türkiye ekonomisine katkı sağlayabilir.

Cumhuriyetin ilk yıllarında koyun varlığı, hayvan varlığımızın yaklaşık %53'ünü oluştururken, son yıllarda bu oran %80'lere çıkmış, daha sonra tekrar %50'lere düşmüştür [4]. Ülkemizde küçükbaş hayvan sayısının azalmasına paralel olarak, hayvansal üretimde de önemli azalmalar görülmektedir. Hayvansal üretimdeki azalma; hayvan sayısında ciddi azalma, yetersiz bakım-beslenme, uygun olmayan yetiştiricilikle birlikte hastalıklara bağlı hayvan başına düşen verim kayıpları ile açıklanabilir. Ülkemizdeki koyun yetiştiriciliğinin istihdamdaki yeri dikkate alındığında, hayvan varlığı sayısı ve hayvansal ürünlerdeki azalma özellikle kırsal kesimlerin daha da yoksullaşmasına neden olmaktadır [5]. Hayvancılık ve hayvansal ürünlerin üretimi ve pazarlanması, ülkemizin hem en önemli gıda kaynağı hem de en büyük ekonomik girdisini oluşturmaktadır. Kuzu hastalıkları ve ölümleri, koyunculuk işletmelerinde ciddi ekonomik kayıplara yol açmaktadır. Bu ekonomik kazancın daha da artırılması hem üreticinin refahına hem de ülke ekonomisine büyük katkı sağlayacaktır [5].

Bu tez çalışmasındaki amaç, veri madenciliği yöntemleriyle hayvan hastalıklarında teşhis, prognoz ve risk faktörlerinin belirlemesidir. Bu amaç doğrultusunda, Bölüm 2'de karar destek sistemleri ve ülkemizde hayvancılığın önemine, Bölüm 3'te veri madenciliği süreci, tez kapsamında kullanılan veri madenciliği yöntemleri ve model değerlendirme ölçütlerine değinilmiştir. Bölüm 4'te kullanılan veri seti ve tez kapsamında izlenen süreç hakkında bilgi verilmiştir. Bölüm 5'de R programlama dili ve RStudio kullanılarak veri önişlem adımları uygulanmış, sınıflandırma algoritmaları kullanılarak yapılan analizlerin performans değerlendirmeleri karşılaştırmalı olarak verilmiş, kuzu ölümlerinde eşik kan değeri belirlenmiş ve geliştirilen mobil ve masaüstü arayüz/veri tabanı sunulmuştur. Son olarak Bölüm 6'da bulgular tartışılmış ve gelecek araştırmalara yönelik öneriler ortaya konulmuştur.

1.1 Literatür Özeti

Veri madenciliği yöntemleri insan sağlığında teşhis, tedavi veya hastalıkların risk faktörlerinin tespitinde hekimlere yardımcı bir araç olarak sıklıkla kullanılmaktadır. Ancak hayvan sağlığında bu başarılı yöntemden yararlanma son yıllarda ivme kazanmış [3] olup, literatürde veri madenciliği yöntemleri kullanılarak hayvanlarda teşhis, prognoz ve risk faktörlerinin belirlendiği çalışmalarla karşılaşılmamıştır.

Literatürde hayvancılık alanında veri madenciliği / makine öğrenmesi uygulamaları sınırlı sayıda olup bu çalışmalardan bazılarına bu bölümde değinilmiştir.

Akıllı vd. [6] bulanık mantık tabanlı bir karar destek sistemi tasarlayarak uzman kararı ile arasındaki uyumu belirlemeye çalışmışlardır. Çalışmada Siyah Alaca ırkı süt sığırlarına ait üreme ve süt verimi kayıtları kullanılarak tasarlanan karar destek sisteminin %92.6 başarı ile doğru sınıflandırma yaptığı ve bulanık mantık tabanlı karar destek sistemlerinin hayvancılık alanında başarılı olacağı bildirilmiştir.

Hempstalk vd. [7] Karar Ağaçları (KA), Saf Bayes (SB), Bayes Ağları (BA), Lojistik Regresyon (LogR), Destek Vektör Makinaları (DVM), Kısmi En Küçük Kareler Regresyonu (KEKK), Rastgele Ormanlar (RO) ve Rotasyon Ormanları (RotO) yöntemlerini kullanarak sütçü sığırlarda tohumlama ile gebeliğin başarısını tahmin etmeye çalışmışlardır. Çalışmada 8 farklı yapay öğrenme yöntemi kullanılmış olup genel olarak lojistik regresyon yönteminin en iyi performans gösterdiği bildirilmiştir.

Küçükönder vd. [8] Japon bıldırcını yumurtalarının döllülük üzerine etkisi olan mevsim, seleksiyon ve yerleşim sıklığı faktörlerine göre sınıflandırmayı ve bu faktörlerin etkisini belirlemeye çalışmışlardır. Çalışmada Yapay Sinir Ağları (YSA), Radyal tabanlı fonksiyon ağları, SB, KStar ve Ridor yöntemleri kullanılmış olup, çalışma sonucunda %99.73 doğru sınıflandırma başarısı ile bıldırcın yumurtalarının genel olarak %85'inin döllü, %15'nin ise üreme kapasitelerinin düşük olduğu tespit edilmiştir. Yumurtaların döllülük özelliğine göre (döllü-dölsüz) sınıflandırılması için Ridor algoritmasının en az hata ile daha başarılı sonuçlar ürettiği görülmüş olup döllülük oranı %85.71 olarak belirlenmiştir.

Lewis vd. [9] karmaşık hayvan sağlığı verilerinde BA yönteminin analitik bir yöntem olduğunu göstermeye çalışmışlardır. Çalışma sonucunda Bayes Ağ modelinin istatistiksel çıkarımının herhangi bir standart istatistik teknikten daha zengin bir analitik araç sunduğu bildirilmiştir.

Karabağ vd. [10] Sinir ağları yöntemini kullanarak kınalı kekliklerde çıkış gücüne etki eden bazı dış yumurta özelliklerinin etkilerini belirlemeye çalışmışlardır. Kuluçkadan çıkış %56.2, döllülük %79.2 ve çıkış gücü %71.0 olarak tespit edilmiş olup, dış yumurta özelliklerinden; yumurta ağırlığı, hacmi, uzunluğu ve genişliğinin çıkış gücü üzerinde

önemli etkiye sahip olduğu sınıflandırma ağacı yöntemi ile %75.6 doğrulukla tahmin edildiği bildirilmiştir.

Pelaez ve Pfeiffer [11], Lineer Regresyon (LinR), SA, FA yöntemlerini kullanarak sığır sürülerini bulaşıcı hastalık varlığına göre sınıflandırmaya çalışmışlardır. Çalışmada yüksek yoğunluklu sığır bölgeleri ile kapalı sık hareketli birçok Galler bölgesinde bulaşıcı hastalık riskinin yüksek olduğu bildirilmiştir.

Gökçe vd. [12] Basit ve çoklu regresyon yöntemlerini kullanarak kuzularda serum laktoferrin konsantrasyonları ile serum IgG konsantrasyonları arasındaki ilişkiyi analiz etmeye çalışmışlardır. Kuzularda neonatal periyodun farklı günlerinde (1, 2, 4, 7, 14 ve 28) serum laktoferrin konsantrasyonlarının pasif immun durumun ve serum IgG konsantrasyonları arasında önemli bir lineer korelasyon (R^2=0.375) olduğunu, fakat bunun IgG konsantrasyonu hesaplamada yetersiz olduğu bildirilmiştir.

Teke vd. [13] Siyah Alaca (Holstein Friesian) ırkının vücut ölçüleri kullanılarak canlı ağırlıklarını modellemeye çalışmışlardır. Çalışmada LinR, Çok Katmanlı Algılayıcı (ÇKA) ve SMOreg yöntemlerini kullanılmıştır. Canlı ağırlığını tahmin etmek için uygulanan sınıflandırma yöntemlerinden LinR modeli %97.94, ÇKA %97.72 ve SMOreg %99.17 başarı göstermiş olup veri madenciliği süresince canlı ağırlığının yüksek güvenirlikle tahmin edilebileceği bildirilmiştir.

Ghotoorlar vd. [14] Yapay Sinir Ağları (YSA) yöntemini kullanarak öznel olarak hesaplanan topallama puanına uygun bir otomatik puan hesaplama sistemi geliştirmeye çalışmışlardır. Serbest gezen 105 süt ineğinden elde edilen 23 özellik kullanılmış olup inekler lokomosyon skorlama sistemi ile yürürken hareket puanına göre 5 gruba ayrılmıştır. Çalışma sonucunda en yüksek duyarlılık ve özgüllük değerine grup 1 ve 4'ün sahip olduğu bildirilmiştir.

Takma vd. [15] Çoklu regresyon ve YSA yöntemlerini kullanarak ineklerin laktasyon süt verimleri üzerine laktasyon süresi, buzağılama yılı ve servis periyodunun etkisini araştırmaya ve uyum yeteneklerini karşılaştırmaya çalışmışlardır. Siyah Alaca ineklerin süt verimlerinin tahmin edilmesinde YSA modelinin çoklu doğrusal regresyon modelinden daha iyi uyum sağladığı ve daha az hataya sahip sonuçlar verdiği için YSA'nın, regresyon analizine alternatif bir metot olabileceği bildirilmiştir.

Hermann-Bank vd. [16] Yeni neonatal domuz ishalinin, bağırsak mikrobiyotunun bileşimi ile ilişkili olup olmadığını incelemeyi amaçlamışlardır. Yapılan çalışmada analizler 50 kontrol ve 52 hasta domuzdan oluşan veri seti üzerinde gerçekleştirilmiştir. Lineer diskriminant analizi (LDA) göre; hasta deneklerin %70'i, kontrol deneklerinin ise %83'ü doğru sınıflandırılmıştır. Ayrıca farklı yaşlarda (üç ile yedi günlük) domuz yavrularındaki mikrobiyotik bileşime bakıldığında hasta ve kontrol gruplarının yaşla birlikte daha ayrıştığı gözlemlenmiştir. TBA göre ise, yeni neonatal domuz ishalinin bakteri bileşimi ile ilişkili olduğu ve ishal domuz yavruları ile aralarında büyük varyasyon olduğu bildirilmiştir.

Küçükönder vd. [8] Siyah alaca ırkı süt sığırlarına ait toplam on üç farklı ölçüt kullanılarak Bulanık c-ortalama yöntemi ile sığırları sınıflandırmaya çalışmışlardır. Çalışmada inekler %97.5 doğru sınıflandırma oranı ile 2 ayrı kümeye ayrıldığında bulanıklık düzeyinin minimum olduğu görülmüştür. Kümelere göre süt bileşenlerinin değişimi değerlendirildiğinde ise somatik hücre sayısı, süt yağı, kuru madde, süt yağı ve süt yoğunluğunun kümeler arası önemli bir farklılık gösterdiği, diğer parametrelerin ise istatistiksel açıdan önemli bir farklılık göstermediği bildirilmiştir.

Dupuy vd. [17] Sığırlara ait karkas ve sağlık verileri kullanılarak sığırları gruplara ayırmaya çalışmışlardır. Çalışmada k-ortalama, hiyerarşik kümeleme, Temel Bileşen Analizi (TBA) ve Çok Faktörlü Analiz (ÇFA) yöntemleri kullanılmıştır. Çalışma sonucunda kesim ve mezbaha yılana göre kararlı 12 küme elde edilmiş olup, kümeleme yöntemlerinin birden fazla faktör analizi ile kombinasyonunun büyük ve karmaşık mezbaha verileri için uygun olduğu bildirilmiştir.

Kılıç vd. [18] Bulanık kümeleme yöntemi ile Karakaya ve Bafra koyunlarını vücut ölçülerine göre sınıflandırmaya çalışmışlardır. Karayaka koyunları %75.4'lük doğru sınıflandırılma oranı ile 2 kümeye ve Bafra koyunları ise %77.1'lik doğru sınıflandırılma oranı ile 4 kümeye ayrıldığında bulanıklık düzeyi minimum olduğu görülmüş olup Bafra koyunlarının Karayaka koyunlarına göre vücut özellikleri bakımından daha heterojen bir yapıya sahip olduğu bildirilmiştir.

Petit vd. [19] Yaban hayvanlarına ait otopsi verilerini kullanarak hayvanları k-ortalama yöntemi ile sendromlara göre gruplamaya çalışmışlardır. Çalışmada sonucunda 9 küme

elde edilmiş olup kümelerin en belirgin ve en sık hastalıkları yansıttığı, k-ortalama yönteminin yararlı bir araç olduğu bildirilmiştir.

Gürcan vd. [20] Alman et merinosu ve Karacabey merinos genotiplerini canlı ağırlık, vücut ölçüleri ve yapağı inceliği değerlerine göre sınıflandırılmaya çalışılmıştır. Çalışmada Hiyerarşik kümeleme yöntemi kullanılmış olup, kümeleme analizi sonucunda iki genotipin vücut ölçüleri arasında benzerlik olduğu, her iki genotipde 1.5 ve 2.5 yaşlıların oluşturduğu alt grupta heterojenlik, 3.5, 4.5 ve 5.5 yaşlıların oluşturduğu alt grupta homojenlik olduğu görülmüştür. Ayrıca her iki genotip grubu birlikte değerlendirildiğinde sürünün %98.9'unun bir kümede toplandığı, iki genotip arasında vücut ölçüleri bakımından önemli bir farkın bulunmadığı bildirilmiştir.

Bozkurt vd. [21] Esmer İsviçre ve siyah alaca ırkı hayvanların, besi performansı ve karkas özelliklerini belirlenmeye çalışmışlardır. Sayısal görüntü analizi ve YSA'dan elde edilen modeller canlı ağırlığın tahmin edilmesinde vücut uzunluğu ve göğüs çevresinin en iyi tahmin edici değişken olduğunu; sıcak karkas ağırlığı tahmininde ise en iyi tahmin edici değişkenin karkas uzunluğu olduğu bildirilmiştir.

Mcevoy vd. [22] YSA ve Kısmi En Küçük Kareler Diskriminant Analizi (KEKKDA) yöntemlerini kullanarak köpeklere ait radyografi görüntüleri üzerinde kalça eklemi içeren bölgeyi belirlemeye çalışmışlardır. Çalışmada model eğitimi için 120 kalça görüntüsü, 80 başka bölgeye ait görüntü kullanılmıştır. Model başarısını test etmek için 36 kalça görüntüsü, 20 başka bölgeye ait görüntü kullanılmıştır. KEKKDA yöntemine göre sınıflandırma hatası, duyarlılık ve seçicilik değerleri sırasıyla %6.7, %100 ve %89 olup YSA yöntemine göre %8.9, %86 ve %100'dir. Çalışmada veterinerlik görüntülerini sınıflandırma ve gruplandırmada eğitim için kullanılabileceği potansiyeline sahip olduğu bildirilmiştir.

Slósarz vd. [23] YSA yöntemi ile kuzularda kas bölümünde bulunan yağ içeriğini tahmin etmeye çalışmışlardır. Kesimden önceki vücut ağırlığı ile kas yağ içeriği arasında düşük ilişki olmasına rağmen kuzuların yaşı ile anlamlı ilişki olduğu bildirilmiştir.

Saidani vd. [24] Sığırlarda warble sineği (büvelek) istilası üzerine iklimin etkisini ortaya koymayı amaçlamışlardır. Çalışmada Radyal Tabanlı Fonksiyon (RTF), DVM, ardışık minimal optimizasyon (sequential minimal optimization, SMO), KA ve LogR yöntemleri

kullanılmıştır. Çalışmada yarı kurak iklime sahip bölgelerdeki yaygınlık ve istila yoğunluğu (%38.23; 21.57±11.98) nemli iklime sahip olanlara göre anlamlı derecede daha yüksek bulunmuştur (%20.74; 14.84±7.86). CHAID algoritmasına göre, başlıca faktörün iklim diğer faktörlerin ise yetiştirme sistemi ve ırk olduğu gözlemlenmiştir. Lojistik regresyon ve çok değişkenli ANOVA' ya göre ise, iklime ek olarak, diğer iç (yaş, cinsiyet, ırk) ve çalışma kapsamındaki dış faktörlerin (yetiştiricilik sistemi, tedavi) hem yaygınlık hem de istila yoğunluğu ile ilişkili olduğu gözlemlenmiştir. Çalışma sonucunda Hypoderma spp. yaşam döngüsünün serbest aşamalarının (pupa ve ergin sinek) gelişimi için yarı kurak alanların, nemli alanlara göre daha uygun olduğu bildirilmiştir.

Awaysheh vd. [25] Serum kimyası değişkenleri ve tam kan hücresi sayısı üzerinde inflammatory bowel disease (IBD) ve alimentary lymphoma (ALA) hastalıklarının etkisini modellemek ve sınıflandırma algoritmaları yardımıyla iki hastalık arasında ayrım yapmayı amaçlamışlardır. Analizler 40 normal, 40 IBD ve 40 ALA yani toplamda 120 kediden toplanan tam kan sayımı ve serum kimyası sonuçları üzerinde gerçekleştirilmiştir. Çalışmada SB ve YSA sırasıyla %70.8 ve %69.2 duyarlılık oranıyla, %62 duyarlılık oranına sahip karar ağacına göre daha yüksek başarıyla sınıflandırma yaptığı gözlemlenmiştir. Ayrıca normal, IBD ve ALA durumlarına göre sınıflandırma yapıldığında ROC eğrisi altında kalan alanın SB tarafından %83, KA tarafından %79 ve YSA tarafından %82 olduğu gözlemlenmiştir. Kedileri 3 duruma göre sınıflandırıldığında 10 değişken kullanan SB ve 4 değişken kullanan YSA'nın, 5 değişken kullanan KA'dan daha iyi performans gösterdiği gözlemlenmiştir. Çalışma sonucunda, bu alanda tahmin modelleri oluşturmak için hem SB hem de YSA modellerinin iyi bir algoritma seçeneği olduğu bildirilmiştir.

Boujenane vd. [26] Holstein ineklerinde klinik masistis görülme sıklığını ve risk faktörlerini analiz etmeyi amaçlamıştır. LogR analizine göre; doğum sayısı 2 olan ineklerin %65, 3 olanların %88, 4 olanların ise %115 masistis riski taşıdığı ve bu riskin ilk doğuma göre daha yüksek olduğu sonucu elde edilmiştir. Çalışma sonucunda mastitisin önlenmesi ve yoğunluk oranının düşürülmesi için ineklerin ilk iki hafta incelenmesi gerektiği bildirilmiştir.

Amrine vd. [27] Hastalığın ilk teşhis ve tedavisinden elde edilen verilere dayalı olarak bireysel buzağıların tedavi sonrası sonuçlarını tahmin etmede KA, BA, YSA, LogR sınıflandırma algoritmaları karşılaştırılmıştır. Sınıflandırıcı performanslarının veri kümesine göre %63 ile %95 arasında değiştiği gözlemlenmiştir. Çalışma sonucunda doğru sınıflandırıcıyı mevcut verilerle eşleştirerek, besiyeri (feedlot) yöneticilerinin doğru tahminde bulunabileceği bildirilmiştir.

Piwczyński vd. [28] Sınıflandırma ağacı ve LogR yöntemlerini kullanarak kuzu ölümlerinden sorumlu olan risk faktörlerini belirleyebilmeyi amaçlamışlardır. Çalışmada hem sınıflandırma ağacı hem de LogR yönteminde sürü, kuzulama yılı, anne yaşı, anne vücut ağırlığı ve doğum türü faktörlerinin kuzu ölümünde etkili olduğu gözlemlenmiştir. Çalışmada sonucunda bilgiyi bir ağaç yapısı şeklinde sunarak karmaşık yapıların bile basit bir şekilde anlaşılmasına olanak sağlamasından dolayı sınıflandırma ağaçlarının kullanılmasının avantajlı olacağı bildirilmiştir.

Sandholm vd. [29] At koliği hastalığı tanısı koymada ve ölüm tahmininde sınıflandırıcı algoritmalarını karşılaştırmayı amaçlamışlardır. Çalışmada LinR, KA, KEYK, SA ve LogR yöntemleri kullanılmıştır. Yapılan çalışmada hastalık tanısını koymanın tüm yöntemler için kolay olduğu ve yöntemlerin ortalama doğruluğunun %94.7 ile 99.3 arasında değiştiği bildirilmiştir. Yöntemlerin doğruluk oranları arasında az farklar olduğu ve LogR ile SA yöntemlerinin en yüksek doğruluk oranlarına sahip olduğu gözlemlenmiştir. Ölüm tahmininin ise tüm yöntemler için zor olduğu ve ortalama doğruluğun %62 ile %72 arasında değiştiği bildirilmiştir. Ayrıca özellik sayısının azaltılması LogR ile hastalık sınıflandırmasında doğruluğu azaltırken ölüm tahmininde %65'den %73'e yükselttiği bildirilmiştir.

1.2 Tezin Amacı

Literatürde veterinerlik alanında veri madenciliği ile ilgili yapılmış çalışmalar olsa da kuzularda hastalığı konu alan ve veri madenciliği yöntemlerinin kullanıldığı bir çalışma ile karşılaşılmamıştır. Hayvan sağlığında veri madenciliği yöntemleri kullanılarak hayvan sağlığını konu alan çalışmalar sınırlı ve yetersizdir. Ayrıca kuzularda veri madenciliği yöntemleri kullanılarak bilgisayar destekli tanı çalışmaları bulunmamaktadır. Bu tez çalışmasının amacı genelde; veterinerlik alanında veri madenciliği yöntemlerinin

kullanılabilirliğini ve veri madenciliği sürecinin işleyişini araştırıcılara sunmaktır. Özelde ise; kuzularda bilgisayar destekli tanı ile veteriner hekime yardımcı olmaktır. Çalışma hem Türkiye ekonomisi üzerinde etkili olan hayvan sağlığının ele alınması hem de veterinerlik alanında veri madenciliği yöntemlerinin kullanılabilirliğini göstermesi açısından önem arz etmektedir.

1.3 Hipotez

Dünyada birçok üretici kuzu hastalık ve ölümlerini azaltmak için etkili ve uygun sistemlere sahip olmasına rağmen, mevcut programlarda hastalıkların önceden tahmin edilerek uygun tedavi ile ve ölümlerin önüne geçmenin yeterince düşünülemediği görülmektedir. İnsan sağlığında karşılaşılan problemlerin çözümünde veri madenciliği yöntemlerinden sıklıkla yararlanılırken en az insan sağlığı kadar önemli olan hayvan sağlığında veri madenciliği yöntemlerinden yeterince yararlanılmadığı görülmektedir. Hayvan sağlığı çalışmalarında, hayvanlardan elde edilen veri setleri genellikle istatistiksel yöntemlerle analiz edilmektedir. Veri madenciliği yöntemleri ile kuzularda bilgisayar destekli tanı sayesinde, veteriner hekime erken teşhis ve tedavi konusunda yardımcı olunacağı düşünülmektedir.

Araştırma Soruları

Soru 1. Veri setindeki eksik değerler neden ve nasıl tamamlanmıştır?

Soru 2. Veri setindeki veri dağılımı normal midir? Veri normal dağılım göstermiyorsa hangi normalizasyon yöntemi kullanılarak veriler normalize edilebilir?

Soru 3. Hasta kuzuları sınıflandırmada en başarılı yöntem hangisidir?

Soru 4. Daha az özellik kullanılarak daha başarılı sınıflandırma yapmak mümkün müdür? Daha az özellik kullanmak neden önemlidir? Hasta ve sağlıklı kuzuları ayırmada önemli rol oynayan özellikler hangileridir?

Soru 5. Kuzu ölümleri ile kan seviyeleri arasındaki ilişkiye bakıldığında ölümlerin en fazla gerçekleştiği kan değer aralıkları nelerdir?

Soru 6. Veteriner hekime yardımcı masaüstü ve mobil uygulama neden gereklidir?

BÖLÜM 2

KARAR DESTEK SİSTEMLERİ VE ÜLKEMİZDE HAYVANCILIĞIN ÖNEMİ

2.1 Karar Destek Sistemleri

Karar destek sistemleri, karar verme sürecince kullanıcıya destek sağlamak amacıyla bilgi üretimi ve bu bilginin sunulması için yazılım ve donanım araçlarının kullanılmasıyla oluşturulmuş sistemlerdir [30]. Tıbbi verinin her geçen gün hızla artması nedeniyle, bilginin yönetiminde zorlanan hekimlere karar verebilme konusunda destek olarak verinin daha iyi analiz edilmesi, yorumlanması ve etkinliğinin arttırılması konusunda yardımcı olurlar [31]. Kısacası veri madenciliği yardımıyla birlikte, geçmiş verilerden ya da diğer bir ifadeyle deneyimlerden sonuç çıkararak hekime karar konusunda yardımcı olurlar. Veri madenciliği yöntemleri hastalara teşhisin konulması, hastalık gruplarının belirlenmesi, tedavi sürelerinin kısaltılması, tedavi yöntemlerinin geliştirilmesi gibi alanlarda sağlık kurumlarına önemli katkılar sağlamaktadır.

Bilim dünyasında insan sağlığı son derece büyük öneme sahip olup bilimsel araştırmaların önemli bir bölümünün insan sağlığı alanında yapıldığı görülmektedir. Ekonomiye ve insan sağlığına büyük katkı sağlayan, havanların sağlığı göz önüne alındığında ise uzman ve otomatik hastalık teşhis sistemlerinin, henüz veterinerlik dünyasında hak ettikleri popülerliği yakalayamamış oldukları görülmektedir. Hayvan sağlığında uzman sistemlerin kullanımının, son yıllarda bilimsel çalışmalarda artan bir ivmeyle yer aldıkları ve geliştirilme noktasında önlerinin açık olduğu anlaşılmaktadır [3].

Tıpkı insan sağlığı alanında olduğu gibi hayvan sağlığında da hasta ve hastalıkla ilgili bilgilerin doğru bir şekilde elde edilmesi, depolanması, işlenmesi, analizinin yapılması,

değerlendirilmesi, sunulması önemlidir. Ancak günümüzde sağlık alanında hekime yardımcı veri tabanları için geliştirilen birçok yazılım uygulamaları olduğu halde, veterinerlik alanında veteriner hekime yardımcı ücretsiz bir masaüstü ve mobil uygulama ile karşılaşılmamıştır.

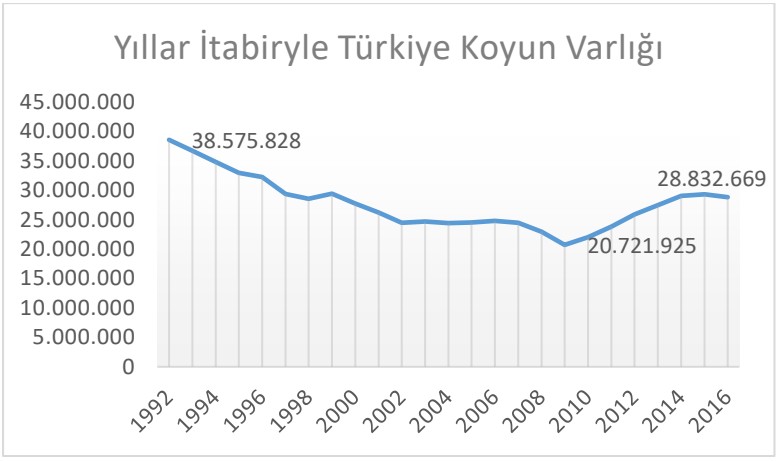
2.2 Türkiye'de Hayvancılık ve Koyun Yetiştiriciliği

Hayvancılık sektörü, insanlık tarihinde her zaman önemli bir yere sahip olmuş ve bu önemini artırarak günümüze kadar sürdürmüştür. Hayvancılık sektörünün en önemli kollarından biri ise koyun yetiştiriciliğidir [5].

Tüm dünya ülkeleri ve Türkiye'de koyun ve koyundan elde edilen ürünler ekonomik bakımdan büyük değer taşırlar. Ülkemizde koyunculuk, tarih boyunca çiftçilerin uğraştığı en önemli hayvan yetiştiriciliğinden biridir. Koyun yetiştiriciliği et, süt, deri ve yün üretimi bakımından ülke ekonomisinde önemli yer tutmaktadır. Kuzu eti, yüksek fiyatlara satılan ve sevilen değerli bir üründür. Koyun sütü, çok değerli olup daima inek sütünün iki katı yüksek fiyata satılır. Koyun peyniri en değerli besin maddesi olup, koyun sütünden yapılan yoğurt ve peynir her zaman alıcı bulmaktadır. Fransa'da koyun sütünden iki yüz çeşit peynir imal edilmekte ve konu ekonomik bakımdan büyük bir değer taşımaktadır. Türkiye'de koyunların süt verimleri düşük olmasına rağmen köylü işletmesinin ekonomik yapısında koyun sütü önemli bir dayanaktır [4]. Ayrıca ülkemizde sadece Kurban Bayramı nedeniyle bile her yıl 2 milyon koyunun kesilmesi koyunculuğun önemini ortaya koymaktadır [32]. Koyun yetiştiriciliği köyden kente göçün önlenmesi, işsizlik ve ekonomik krizden çıkışa katkısı açısından da oldukça önemlidir [4].

Son yıllarda ülkemizdeki küçükbaş hayvan sayısındaki azalmaya paralel olarak koyunculuğun payı da giderek azalmaktadır. Cumhuriyetin ilk yıllarında koyun varlığı, hayvan varlığının yaklaşık %53'ünü oluştururken, son yıllarda bu oran %80'lere çıkmış, daha sonra tekrar %50'lere düşmüştür [4].

Türkiye İstatistik Kurumu verilerine göre Türkiye'deki koyun varlığı oranı Şekil 2.1'de sunulmuştur.



Şekil 2. 1 Yıllara göre Türkiye'deki koyun varlığı

Türkiye'de, 1992'de yaklaşık 40 milyon olan koyun sayısı, 2009 yılında yaklaşık 21 milyona düşmüş, 2016 yılında ise yaklaşık 29 milyona yükselmesine karşın, 24 yılda koyun sayısında yaklaşık %30 oranında azalma olduğu bilinmektedir [33].

Doğu ve Güney Doğu Anadolu Bölgeleri'nde yaşayan halkımızın geçim kaynağının önemli bir kısmını koyunculuğun oluşturduğu bilinmektedir. Ancak, 1980 yılından sonra gereken destek sağlanmadığı için koyun sayısında önemli bir azalma görülmüştür. Son yıllarda koyunculuğa sağlanan devlet destekleri sayesinde koyun yetiştiriciliği tekrar önem kazanmış ve koyun sayısı artmış olsa da henüz istenilen seviyeye gelememiştir [32]. Türkiye İstatistik Kurumu verilerine göre Kars ilindeki koyun varlığı oranı Şekil 2.2'de sunulmuştur.

Şekil 2. 2 Yıllara göre Kars ilindeki koyun varlığı

Kars yöresinde 1998'de yaklaşık 630.000 olan sayı, 2010'da yaklaşık 200.000'e düşmüştür. Bu sayı sağlanan desteklerle yaklaşık 530.000'e yükselmiş olsa da hala 24 yıl önceki koyun sayının gerisindedir.

Ülkemizde Akkaraman, Morkaraman, Dağlıç, Kıvırcık, Sakız, Merinos, Karayaka, Karagül, İvesi, Malya, Tahirova, Herik ırkı koyunlar bulunmaktadır [4]. Türkiye'nin orta Anadolu ve doğu Anadolu bölgesinde yayın olarak görülen yağlı kuyruklu Akkaraman koyunu, sert çevre ve yönetim koşullarına, kötü beslenmeye ve hastalıklara en iyi adapte olabilen türdür. Verimliliği nedeniyle değerli olduğu düşünülen Akkaraman ırkı koyun türü, Türkiye'deki koyun nüfusunun yaklaşık %50'sini (yaklaşık 21 milyon baş) oluşturmaktadır [34]. Kars ili de dahil olmak üzere koyun yetiştiriciliği, Anadolu'da büyük bir ekonomik faaliyet olduğu için hayvan hastalıklarının önüne geçmek önem taşımaktadır.

2.3 Koyunlarda Neonatal Dönemin Önemi

Canlı vücudu bağışıklık sistemi sayesinde enfeksiyonlara bağlı oluşabilecek hasarların önüne geçme yeteneğine sahiptir [35]. İnsanlarda intrauterin hayatta anneden yavruya bağışıklık maddelerinin geçmesi yoluyla direnç kazanılırken, kuzular hastalıklara karşı yok denecek kadar zayıf bağışıklıkla dünyaya gelmektedir. Ruminantlarda gebelik esnasında anneden yavruya antikor geçişi olmadığı için, bağışıklık gelişimi için tek gıda kolostrumdur [36].

Halk arasında ağız sütü denilen kolostrum, memeli canlılar doğum yaptıktan sonra salgılanan; rengi, tadı, kokusu, bileşimi sütten oldukça farklı olan; yüksek besleyici değere sahip kompleks yapılı bir sıvıdır. Sütten en önemli farkı ise bileşimidir [37]. Yeni doğan ruminantların hayatta kalmaları için son derece önemli olan kolostrum, hayvanların büyüme ve gelişmelerinde etkili olan, enerji ve immunoglobulin yönünden zengin bir besin kaynağıdır [38], [37].

Yeni doğmuş kuzularda kolostrum alınması ve emilmesi, hastalıklara karşı kandaki antikorların titresini artırır. Kolostrum verilmesinde geç kalınan ve/veya tüketimi yetersiz olan kuzularda hastalığa yakalanma ihtimali ve ölüm oranı artar [39], [40], [41], [42].

Koyunculuk işletmelerindeki kayıpların daha çok neonatal (yaşamın 0-28. günleri) dönemde gözlendiği bildirilmektedir. Neonatal dönem yeni doğan kuzuların çevresel değişikliklere ve infeksiyonlara daha duyarlı olduğu dönem olduğundan hastalanma riski yüksektir. Neonatal dönemin ilk haftasında, kolostrumun yeterince veya hiç alınamaması, annelik içgüdüsünün gelişmemesi/yetersizliği, çevre koşullarının iyi olmaması gibi nedenler kuzuların hastalanma ve ölme riskini arttırır. Yani neonatal dönemde fiziksel ve immun direncinin gelişim safhasında olması, kuzuların infeksiyöz ve noninfeksiyöz etkenlere maruz kalarak ölmelerine neden olabilir. Bundan dolayı, neonatal dönem morbidite ve mortalite nedenleri ve risk faktörlerinin belirlenmesi ve önceden koruma ve kontrol stratejilerinin geliştirilip uygulanması, kuzuların yaşama şanslarını %50'lere varan oranlarda artırabileceği bildirilmiştir [5].

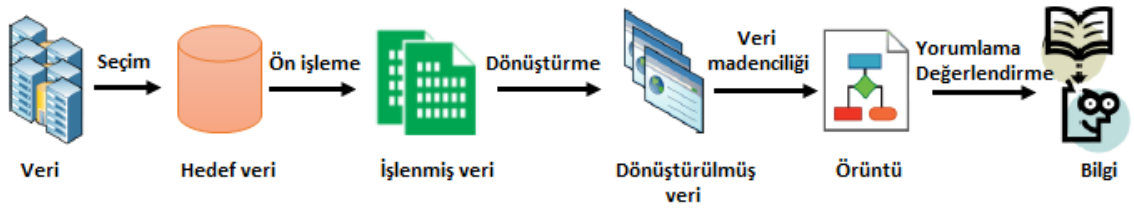
Koyunculuk işletmelerindeki kayıpların %35'i neonatal dönemde gerçekleşmekte ve bu ölümlerin de %75'i yaşamın ilk haftasında meydana gelmektedir. Özellikle yaşamın ilk 15 günü, hastalıkların ve ölüm oranının en yüksek olduğu dönemdir [5]. Neonatal kuzu kayıpları nedeniyle ülkemizde her yıl önemli kayıplar meydana gelmekte ve bu kayıplara bağlı olarak ülke ekonomisi ve hayvancılık olumsuz etkilenmektedir [43]. Modern yöntemlerle yapılan koyun yetiştiriciliğinde dahi yeni doğan kuzuların ve genç hayvanların ölüm oranları yüksektir ve bu durum büyük oranda ekonomik kayıplara neden olmaktadır.

İstatistiksel analizler bu noktada yetersiz kalmakta olup, hayvanlarda bilgisayar destekli tanı ile hastalıkların önceden tahmin edilmesinde veri madenciliği yöntemlerine ihtiyaç duyulmaktadır. Hayvan sağlığı alanında veri madenciliği yöntemlerinin kullanılması veteriner hekimlerin en uygun kararları almasına yardımcı olarak bilime katkı sağlayacaktır.

BÖLÜM 3

VERİ MADENCİLİĞİ

Veri madenciliği büyük veri setlerinden çeşitli algoritmalar ve veri analiz araçları kullanılarak anlamlı, işe yarar bilgilerin keşfedilmesi olarak tanımlanmaktadır. Veri madenciliği disiplinler arası bir çalışma alanı olup tıp, biyomedikal, genetik, bankacılık, eğitim gibi birçok alanda kullanılmaktadır. Fayyad ve diğ. tarafından veri madenciliği süreci Şekil 3.1'deki gibi ortaya konmuştur [44]. Buna göre veriden bilgiye giden süreç beş adımdan oluşmaktadır ve veri madenciliğinin bu sürecin adımlarından biri olduğu görülmektedir.



Şekil 3. 1 Fayyad ve diğ. göre veri madenciliği süreci

Şekil 3.1 inceldiğinde süreç adımları şu şekildedir;

1. Seçim: Verinin seçilmesi aşamasıdır.
2. Ön işleme: Madencilik için ham verinin analize hazır hale getirilmesi aşamadır.
3. Dönüştürme: Analiz için verinin uygun formata dönüştürüldüğü aşamadır.
4. Veri madenciliği: Çalışmanın amacına uygun olan veri madenciliği yöntemlerinin veriye uygulandığı aşamadır.
5. Yorumlama/Değerlendirme: Veri madenciliği ile elde edilen sonuçların yorumlandığı ve değerlendirildiği aşamadır.

3.1 Veri Ön İşleme

Verinin kalitesi veri madenciliği sonuçlarını etkilemektedir. Verinin kalitesini ve dolayısıyla madencilik sonuçlarının iyileştirilmesine yardımcı olmak için ham veriler önceden işlenir. Çünkü ham veriler işlenmemiş olup eksik ve gürültü içermektedir. Bir projede sürecin % 60 - % 90'ı verinin anlaşılması ve hazırlaması için harcanmaktadır. Bu da veri önişleme aşamasının önemini göstermektedir [45].

Literatürde çok sayıda veri ön işleme tekniği mevcuttur. Ancak veri ön işleme tekniklerine başlamadan ve modelin kurulmasından önce mevcut verinin iyi anlaşılması ve analiz edilmesi gerekmektedir. Veri seti hakkında ön fikir edinilmesi için bazı basit istatistiksel hesaplamalar yapılabilir ve grafikler çizilebilir. Niteliklerin/özelliklerin kategorik ya da nümerik olmasına göre maksimum, minimum, mod, ortanca, ortalama ya da kartil hesaplamaları yapılabilir, kutu grafiği (box-plot), histogram, sütun ve pasta grafikleri ile görselleştirilebilir. Tüm bu işlemler, veri ön işleme sürecinde hangi analizlerin gerçekleştirilmesi gerektiği hakkında bilgi vermektedir.

Kutu grafiği; sürekli değişkenin hem büyüklüğünü, hem de dağılımını gösteren bir grafik türüdür. Merkezsel konum, çarpıklık, yayılma ve basıklık yönünden verileri özetlemek ve aykırı (outlier) değerleri tanımlamak için kullanılır [46].

Histogram grafiği; histogram verinin frekans dağılımın hakkında kaba bilgi veren bir yaklaşımdır. X ekseninde sürekli bir değişkene ait aralıkları, Y ekseninde ise o aralıktaki örnek sayısını gösteren grafik türüdür [47].

Veri ön işleme sürecinde gerçekleştirilecek işlemlerin herhangi bir standardı olmayıp, kullanılan veri setine göre değişmektedir. Veri ön işleme tekniklerinden *Veri temizleme* veri setindeki eksik değerlerin ve verideki gürültünün giderilmesi ve için uygulanır. *Veri birleştirme* de farklı kaynaklı veriyi uygun bir veri tabanında birleştirir. *Veri dönüştürme* yöntemlerinde normalleştirme gibi işlemler uygulanabilir. *Veri indirgeme* de ise fazla olan bazı değişkenlerin atılması ve birleştirilmesi yolu ile veri büyüklüğünün azaltılması işlemleri gerçekleştirilir [46].

3.1.1 Veri İndirgeme

Veri madenciliği analizlerinde kullanılan veri setleri genellikle hem kayıt (örnek/gözlem) sayısı hem de nitelik/özellik sayısı açısından büyük, tekrarlayan veya ilgisiz olabilmektedir. Bu niteliklerin veya örneklerin veri setlerinden çıkarılması bu adımda gerçekleştirilmektedir.

- Nitelik sayısının azaltılması; bu aşamada tekrarlayan veya ilgisiz olan nitelik alanlarının veri setinden çıkarılabilir veya nitelik alanları birleştirilerek, yeni nitelik alanları oluşturulabilir [46].
- Gözlem sayısının azaltılması; kümeleme, parametrik veya parametrik olmayan yöntemlere göre veri setini temsil eden örneklem seçilere bunlar ile çalışabilir [46].

3.1.2 Veri Temizleme

Veri temizleme; kayıp/eksik değerlerin tamamlanması, kirli verinin temizlenmesi, uç değerlerin tespiti ve tutarsızlıkların giderilmesi gibi işlemleri içerir [46].

3.1.2.1 Eksik/Kayıp Değerlerin Tamamlanması

Bilimsel araştırmalar kapsamında üzerinde çalışılmak istenilen veriler hastalık, ölüm, analizin hatalı yapılması, ölçümü yapılan örneğin uygun olmaması gibi nedenlerle istenildiği gibi eksiksiz bir şekilde toplanılamayabilir. Bu niteliklerin veri kümesine kayıt edilememesinin yanı sıra gürültülü, aykırı veya tutarsız değerlerin tespit edilip veri kümesinden silinmesi sonucunda da eksik değerler oluşabilmektedir. Veri setlerindeki bu eksik değerler analizler için kullanılacak olan klasik ve modern istatistiksel yöntemlerin hemen hemen hepsi için önemli bir sorun oluşturur. Çünkü tüm yöntemler veri setinin eksiksiz olduğu varsayımı altında geliştirilmiştir [48], [49], [50].

Eksik değerler için çözüm ve atama yöntemlerinin doğru kullanımı, doğru ve geçerli sonuçlara ulaşabilmek açısından önemlidir. Eksik veri probleminin üstesinden gelmek için yapılacak en önemli adım eksik değere hangi mekanizmanın sebep olduğunu bulmaktır [51]. Little ve Rubin'e göre 3 farklı eksik değer yapısı vardır. Bunlar; Tamamen rasgele olarak kayıp (Missing Completely at Random, MCAR), Rasgele olarak kayıp (Missing at Random, MAR) ve Rasgele olmayan kayıp (Missing Not at Random, MNAR).

Kayıp veri mekanizmasının belirlenmesi, kayıp verilere uygulanacak uygun yöntemin seçilmesine, uygun yöntemin seçilmesi de analizler sonucunda doğru sonuca ulaşılmasını sağlar [52].

Tamamen Rasgele Olarak Kayıp (MissingCompletely at Random, MCAR): Bu yapıdaki eksik veriler, rasgeleliğin en yüksek olduğu durumdur. Burada verinin eksik olma olasılığı, veri setinde bulunan diğer değişkenlere ve değişkenin kendisine bağlı değildir [52], [53], [54], [55]. Yani verilerin kayıp olma durumu şans faktörüne bağlıdır [56]. MCAR varsayımının karşılanması durumunda kayıp veri mekanizması ihmal edilebilirdir (ignorible). İhmal edilebilirlik, kayıp değerlerle, seçkisiz olarak belirlenen gözlenen verilerin örtüşeceği varsayımını destekler. Bu durumda kestirim sürecinin bir parçası olarak kayıp veri mekanizmasının modellenmesine gerek olmadığı söylenebilir [57], [58]. Anket çalışmasında tesadüf olarak bazı soruların eksik olması ya da cevaplayan kişilerin soruyu görmemesinden kaynaklanan eksik değerler bu yapıya örnek olarak gösterilebilir.

Rasgele olarak kayıp (Missing at Random, MAR): Rasgele olarak kayıp, tamamıyla tesadüfi kayıp mekanizmasına göre daha zayıf bir varsayımdır. Bu kayıp veri mekanizmasında 'Rasgele' ifadesi geçse de aslında diğer değişkenlere bağlı olan sistematik bir eksik veri mekanizması vardır. Eksik veri mekanizması, eksik verilere değil gözlenen verilere bağlıdır. Örneğin, cinsiyet ve kilo gibi iki değişken olsun. Kadınlar, erkeklere göre kilolarını saklama eğiliminde oldukları için kilo değişkenindeki eksik veriler cinsiyet değişkenine bağlı olarak değişebilecektir. Bu durumda kilo değişkeninde bulunan eksik veriler rasgele olarak eksiktir [52]. MAR varsayımının karşılanması durumunda kayıp veri mekanizması ihmal edilebilirdir.

Rasgele olmayan kayıp (Missing Not at Random, MNAR): Verilerin eksik olma olasılığı gözlenemeyen yani eksik verinin yer aldığı değişkenlerle ilgilidir. Bu durumda olan eksik veriler ihmal edilemez (non-ignorable) ve eksik veri mekanizmasının, iyi bir parametre tahmini yapabilmek için modellenmesi gerekir. Eksik veri yine sistematiktir ve verinin eksik olma nedeni eksik verinin kendisiyle ilişkilidir [52]. Örneğin öğrencilere verilen bir okuma testinde, okuma güçlüğü çeken öğrencilerin bazı test sorularını boş bıraktığı görülebilir. Bu durumda eksik değerler öğrencilerin okuma güçlüğü ile ilişkilendirilebilir.

Eksik verilerin rastgeleleğinin incelenmesinde farklı yöntemler kullanılmakta olup bu yöntemler içerisinde en sık kullanılanı Little's MCAR testidir. Bu yöntem 3 aşamada gerçekleştirilmektedir. İlk aşamada beklenti maksimizasyonu yöntemiyle tahmini olarak ortalama ve varyans kovaryans matrisleri belirlenir. İkinci aşamada kayıp veri örüntüsüne göre gözlemler gruplandırılır ve her grup için gözlemlerin ortalaması alınır. Son aşamada ise gözlenen ve tahmin edilen ortalama değerlerin farkı alınarak, değerler varyans kovaryans matrisi ve gruplardaki gözlem sayısına göre ağırlıklandırılır ve elde edilen istatistiğe göre varsayım test edilir [59].

3.1.2.2 Kayıp Veri Analizinde Kullanılan Başlıca Yöntemler

Günümüzün güncel problemlerinden biri olan kayıp veri probleminin üstesinden gelmek için çeşitli çözüm ve veri atama yöntemleri geliştirilmiştir. Literatür incelendiğinde kayıp değerleri tamamlama problemlerinin çözümü için kullanılan yöntemler 4'e ayrılmaktadır.

i. **Eksik Değerli Kayıtları Yok Sayma (Ignoring Missing Records):** Eksik değerli gözlemlerin veri setinden çıkartılması ile gözlem sayısında ciddi bir azalmaya yol açabilir ve yeterli sayıda oluşturulmuş bir örneklem yetersiz sayıdaki bir örnekleme dönüşebilir. Bu durum yapılacak olan istatistiksel analizlerin gücünün azalmasına neden olabilir [60], [61]. Ayrıca kayıp verilerin analize dahil edilen başka değişkenlerle ilişkili olduğu durumlarda yapılacak olan silme işlemi önemli bir yanlılığa yol açabilir [48], [62], [63]. Bu nedenle kayıp veriler yerine yaklaşık değer atama yöntemleri, toplanılan verinin korunabilmesini sağlayacak bir yoldur. Bu nedenle eksik değerli gözlemlerin veri setinden çıkarılması araştırmacılar tarafından pek tercih edilmemiş ve kayıp değerleri tamamlamak için farklı çözümler aranmıştır.

ii. **Tek Değer Atama (Single Value Substitution):** Bu yöntemde eksik değerlerin tümüne tek bir değer ataması yapılır. Sıfır (zero), varsayılan (default), rastgele (random) veya ortalama/ortanca (mean/mode) değer atama bu kategoride sıklıkla kullanılan yöntemlerdir. Bu yöntemlerin uygulanması basittir ancak doğruluk söz konusu olduğunda araştırmacılar tarafından genellikle tercih edilmezler [64]. Çünkü gerçek değerler farklı ortamlarda farklı olur ve bu sapma değerlerin varyansında bozulmaya neden olur. Bu nedenle de eksik değerleri tamamlamak için

araştırmacılar daha farklı algoritmalar geliştirme yoluna gitmişler ve istatistiksel yöntemler ortaya çıkmıştır.

iii. **İstatistiksel Yöntemler (Statistical Methods):** İstatistikçiler genellikle eksik veri setlerinden kararlar vermek yerine eksik değerleri işlerler. Eksik değerleri doldurduktan sonra tam veri kümesi için kullanılan standart teknikler kullanılarak, atanan veri kümesi üzerinde analiz yapılır. Ancak herhangi bir analiz yapılırken, atanan değerlerle ilişkili olarak her zaman belli bir belirsizlik derecesi olduğu ve standart veri analiz yöntemlerinde iyileştirmeler yapılması gerekmektedir. Ayrıca öznitelik türüne bağlı olarak, gerekli teknik farklılık gösterebilir. Örneğin ayrık (discrete) durum için uygulanan teknik, sürekli (continuous) durum için kolayca uygulanamaz [65]. Bu kategoride sıklıkla kullanılan yöntemler şunlardır:

- **K-En Yakın Komşu ile Kayıp Değer Atama Yöntemi (kNN imputation):** En yakın komşu yöntemi aslında bir sınıflandırma yöntemi olup kayıp değerlerin çözümlenmesinde bu sınıflandırma mantığı kullanılmaktadır. K-en yakın komşu algoritması ile kayıp değer atama yöntemi, gözlemlerin birbirlerine olan yakınlıkları üzerine kuruludur. Eksik değer içermeyen değişkenler arasında mesafe ölçümü yaparak 'k' en yakın gözlemin ortalaması ile tamamlanır. Yani eksik olmayan özelliklere en çok benzeyen k tane özellik seçilir [66]. Bu yöntemi uygulamak için R'da 'VIM' paketi kullanılmıştır.
- **Rastgele Orman (Random Forest; RF) Yöntemi ile Kayıp Değer Atama:** Stekhoven ve Buhlmann tarafından önerilen missForest yöntemi, eksik veriyi tamamlarken Rastgele orman yöntemini kullanır. Veri kümesindeki değişkenlerin geri kalanını kullanarak her değişken için rastgele bir orman modeli oluşturur ve bu değişkenin eksik değerlerini tahmin etmek için onu kullanır [67]. Bu yöntemi uygulamak için R'da 'missForest' paketi kullanılmıştır.
- **Zincir Denklemleri ile Çoklu Değer Atama (Multivariate Imputation by Chained Equations; MICE):** Van Buuren ve arkadaşları tarafından önerilen bu yöntem diğer değişkenleri tahmin edici olarak kullanarak, her değişken için bir koşullu model belirlenmesini gerektirir. Bu yöntemi uygulamak için R'da 'mice' paketi kullanılmıştır [68].

iv. **Evrimsel Yöntemler (Evolutionary Methods):** Günümüzde Genetik Algoritma, Karınca Kolonisi Optimizasyonu, Parçacık Sürü Optimizasyonu, Yapay Arı Koloni Optimizasyonu gibi evrimsel algoritmalar birçok veri madenciliği probleminde en uygun çözümü bulmak için başarıyla uygulanmıştır. Son günlerde, evrimsel algoritmalar eksik veri tamamlamada popülerlik kazanmaktadırlar. Kayıp değer tamamlama yöntemleriyle ilgili çalışmalar incelendiğinde yapay arı kolonisi algoritması ile eksik değerlerin tamamlanmadığı gözlemlenmiştir. Bu nedenle tez çalışmasında kayıp değerleri tamamlamak için Yapay Arı Koloni algoritması geliştirilmiştir.

3.1.3 Veri Dönüştürme

Veri dönüştürme sürecinde, veri setlerine ayrıklaştırma ve normalizasyon işlemleri uygulanabilmektedir. Veri setindeki nümerik verinin kategorik veriye dönüştürülmesi işlemine ayrıklaştırma adı verilmektedir [69]. Veri setindeki nümerik verinin aldığı değer aralıklarının belli bir standart değere çekilmesine ise normalleştirme/normalizasyon denmektedir.

3.1.3.1 Veri Ayrıklaştırma

Veri önişleme süreci içerisinde yer alan adımlardan birisi de veri ayrıklaştırmasıdır. Veri ayrıklaştırma işlemi için farklı yöntemler kullanılmaktadır. Ayrıklaştırma (discretization) numerik verilerin kategorik karşılıklarına dönüştürülmesi işlemine verilen addır [70]. Ortak bilgiye dayalı en popüler öznitelik seçim yöntemleri, özniteliklerin ayrıklaştırılmasına başvurur. Ayrıca bu işlem verideki gürültüyü ve doğrusalsızlığı (non-linearity) azaltarak tahminleme modelinin kesinliğini artırabilir. Ayrıca bu yöntem aykırı gözlemlerin tespitini, geçersiz veya eksik numerik değerlerin tespitini kolaylaştırır. Numerik değişkenler genellikle sıklık dağılımlarına göre ayrıklaştırılırlar.

3.1.3.2 Veri Normalleştirme

Veri madenciliği uygulanırken ham veriyi kullanmak bazen uygun olmayabilir. Veri setindeki değişkenlerin aldığı değer aralıklarının birbirinden farklı olduğu durumlarda normalizasyon yöntemlerinden yararlanılır. Çünkü değişkenlerin ortalama ve

varyansların birbirlerinden önemli ölçüde farklı olmaları, büyük ortalama ve varyansa sahip değişkenlerin diğer değişkenler üzerindeki baskısını artırarak doğruluk ve performansı etkileyebilmektedir. Bu nedenle veri madenciliğinde her bir değişkenin sonuçlar üzerindeki etkisini standartlaştırmak için nümerik veri normalizasyon yöntemleri ile normalize edilir [71], [72].

Tez kapsamında normalizasyon için aşağıdaki yöntemlerden yararlanılmış olup bu yöntemlerin performansları karşılaştırılmıştır.

Minimum-maksimum normalizasyonu: Bu yöntem, verileri doğrusal olarak normalize etmektedir. Yani iki değer arasındaki değerlerin büyüklüğü ya da arasındaki fark değişmez. Normalizasyon sonrası değerler genellikle 0-1 aralığında olup; minimum bir verinin alabileceği en düşük değeri, maksimum ise verinin alabileceği en yüksek değeri temsil eder [73]. Minimum-maksimum modelinde değerler Eşitlik 3.1'den hesaplanır. Minimum-maksimum normalizasyonu için R'da 'mmnorm' yöntemi kullanılmıştır.

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3.1)$$

x' : $Normalize\ edilmiş\ veri$

x_i : $Girdi\ değeri$

x_{max} : $Girdideki\ en\ k$üçü$k\ sayı$

x_{min} : $Girdideki\ en\ b$üyü$k\ sayı$

Ondalık ölçekleme normalizasyonu: Normalleştirme işlemi gerçekleştirilirken ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Normalize edilen değerlerin aralığı -1 ve +1 arasındadır. Ondalık ölçekleme modelinde değerler Eşitlik 3.2'den hesaplanır. Ondalık ölçekleme normalizasyonu için R'da 'dscale' yöntemi kullanılmıştır.

$$x' = \frac{x_i}{10^j} \quad (3.2)$$

j: Max(|x^'|)<1 olacak şekildeki en küçük tam sayı

Z-değeri (Z-skor) normalizasyonu: Normalleştirme, değişkenin ortalaması ve standart sapmasına bağlı olarak bilinen Z dönüşümü ile gerçekleştirilir. Z-değeri, her ne kadar eksi sonsuz ile artı sonsuz arasında tanımlı olsa da dağılım genellikle -1.5 ile +1.5 arasında

gerçekleşir [73]. Z-skor modelinde değerler Eşitlik 3.3'den hesaplanır. Z-skor normalizasyonu için R'da 'znorm' yöntemi kullanılmıştır.

$$x' = \frac{x_i - \mu_i}{\sigma_i} \quad (3.3)$$

x' : $Normalize\ edilmiş\ veri$

x_i : $Girdi\ değeri$

μ_i : $Girdi\ setinin\ ortalaması$

σ_i : $Girdi\ setinin\ standart\ sapması$

Sigmoid normalizasyonu: Bir kaç tane sigmoid fonksiyon çeşidi olup lojistik ve hiperbolik tanjant fonksiyonları en yaygın olarak kullanılanlarıdır. Lojistik sigmoid normalizasyonu değerleri 0 ile 1 arasına dönüştürür ve değerler Eşitlik 3.4'den hesaplanır. Hiperbolik tanjant ise değerleri -1 ile 1 arasına dönüştürür ve değerler Eşitlik 3.5'den hesaplanır [74], [75]. Sigmoid normalizasyonu için R'da 'signorm' yöntemi kullanılmıştır.

$$x' = \frac{1}{1+e^{-x_i}} \quad (3.4)$$

$$x' = \frac{e^{xi} - e^{-x_i}}{e^{xi} + e^{-x_i}} \quad (3.5)$$

x' : $Normalize\ edilmiş\ veri$

x_i : $Girdi\ değeri$

e : $Doğal\ logaritma\ değeri$

3.2 Veri Madenciliği Yöntemleri

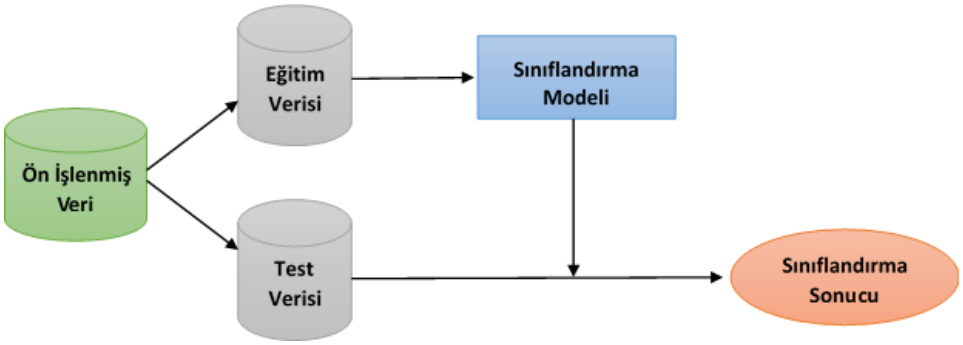
Veri madenciliğinde benzer problemlerin çözümü için birden fazla yöntem mevcuttur. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar tekrar eden bir süreçtir [47]. Bu süreçteki aşamalar şu şekildedir;

- *Model tekniğini seçmek:* Kullanılacak olan veri madenciliği yöntemi ve algoritması belirlenir.
- *Model Test Tasarımı Yapma:* Model işlenip sonuçlar elde edilmeye başlanmadan önce, modelin kalitesi ve geçerliliğinin test edilmesi gerekir. Veriler hazırlandıktan sonra, verilerin bir kısmı modelin eğitilmesinde, diğer kısmıysa kullanılacak modelin geçerliliğinin test edilebilmesinde kullanılır.

- *Model Kurma:* Model için kullanılacak algoritmanın, yöntemin hazırlanan veri üzerinde çalıştırılması aşamasıdır.
- *Model Değerlendirme:* Test sonuçlarının değerlendirildiği süreçtir.

Veri madenciliği alanında pek çok geliştirilmiş yöntem bulunmaktadır. Bu yöntemler genel olarak sınıflandırma ve regresyon, kümeleme, birliktelik kuralı şeklinde gruplandırılmaktadır. Tez çalışması kapsamında sınıflandırma yöntemleri kullanıldığı için sınıflandırma yöntemleri detaylı olarak ele alınmıştır. Ayrıca analizler, R programlama dilinde yazılan kodlar ile RStudio programı kullanılarak yapılmıştır.

Sınıflandırma, çeşitli sınıflandırma yöntemlerinin kullanılarak sınıf etiketi bilinmeyen örneklerin hangi sınıfa ait olduğunu belirlemeyi sağlar. Hangi sınıflandırma yöntemi kullanılırsa kullanılsın temel mantık aynı olup, verinin bir kısmı model eğitiminde, geri kalan kısmı ise test edilmesinde kullanılır. Model eğitilerek sınıflandırma kuralları ortaya çıkarılır ve verinin hangi sınıfa ait olacağı bu kurallar çerçevesinde tahmin edilir. Dolayısıyla yeni bir kayıt bir veri setine eklendiğinde veya hangi sınıfa ait olduğu bilinmeyen bir kayıt varsa modelin oluşturduğu kurallara göre sınıfın tahmin edilmesi sağlanır [76]. Sınıflandırma yöntemlerinin çalışma mantığı Şekil 3.2' de gösterilmiştir.



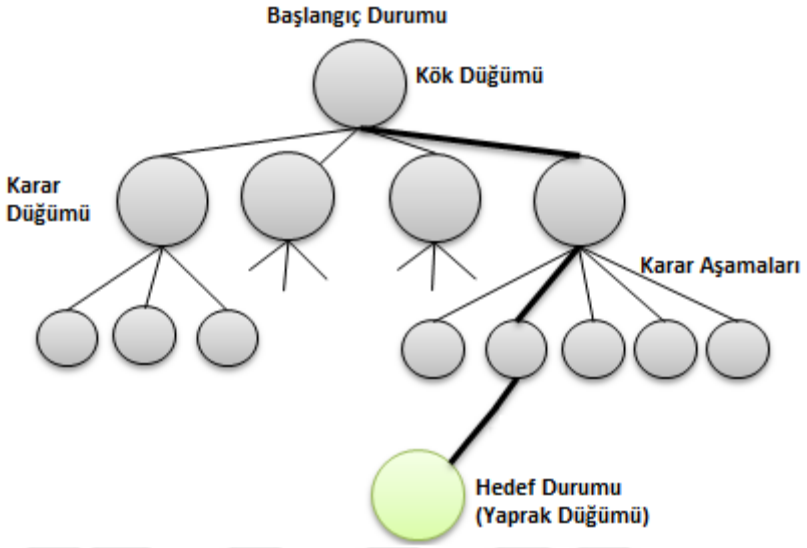
Şekil 3. 2 Sınıflandırma yöntemlerinin iş akışı

Bu tez çalışması kapsamında karar ağacı (J48), saf bayes, k-en yakın komşu, yapay sinir ağı ve rastgele orman algoritmaları kullanılmıştır.

3.2.1 Karar Ağaçları Algoritması

Karar ağaçları, geçmiş veriye dayanarak yeni verilerin hangi sınıfa ait olduğuna karar vermektedir. Ağaca benzer hiyerarşik bir yapısı vardır. Kolay anlaşılabilir olması,

yorumlanmasının kolay olması, gerçek hayattaki problemlere uyarlanabilmesi, hem nümerik hem de kategorik veri ile çalışma imkânı sağlaması gibi nedenlerden dolayı en yaygın kullanılan yöntemlerin başında gelmektedir [76]. Karar ağaçlarının yapısı Şekil 3.3'de gösterilmiştir.



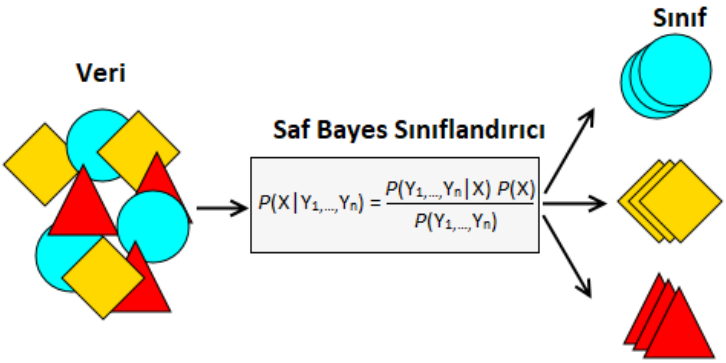
Şekil 3. 3 Karar ağaçları yönteminin genel yapısı

Karar ağacının en tepesindeki düğüm kök düğüm olup sınıflamada en fazla ağırlığı olan özelliği temsil etmektedir. Bu düğümün dallanması ile oluşan farklı düğümler dalları temsil eder. En son da yaprak adı verilen yapılar ile karar ağaçları sonlandırılır.

Karar ağacı yapılarında her düğüm bir özellik için yapılan testi, her dal bu testin sonucunu ve her yaprak düğüm ise sınıfları ifade etmektedir. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'J48' fonksiyonu kullanılmıştır [77].

3.2.2 Saf (Naive) Bayes Algoritması

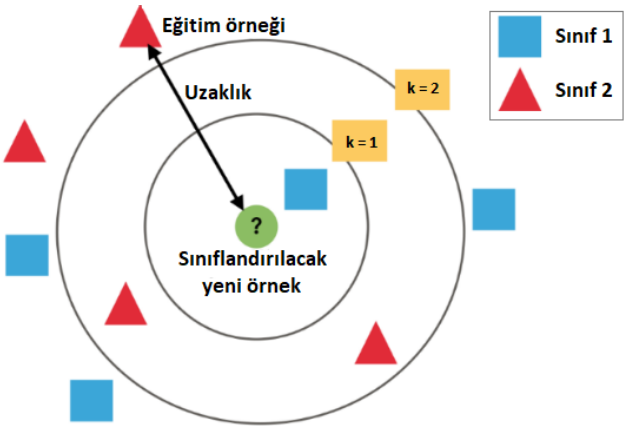
Bayes sınıflandırıcı olarak da adlandırılan bu algoritma istatistikteki bayes teoremine dayanmaktadır. Şekil 3.4'de görüldüğü gibi, sınıflandırılmış verileri kullanarak sınıfı belli olmayan verilerin hangi sınıfa ait olabileceğini istatistiksel olarak belirler. Saf bayes algoritmasında, kullanılan bütün nitelikler birbiriyle eş değer derecede önemli, birbirinden bağımsız ve birbirileri hakkında bilgi içermedikleri kabul edilir. Saf Bayes algoritması uygulanmasının kolay olması ve başarılı sonuçlar vermesi nedeniyle sınıflandırma modellerinde sıkça tercih edilir [78].



Şekil 3. 4 Saf bayes yönteminin genel yapısı

3.2.3 K - En Yakın Komşu Algoritması

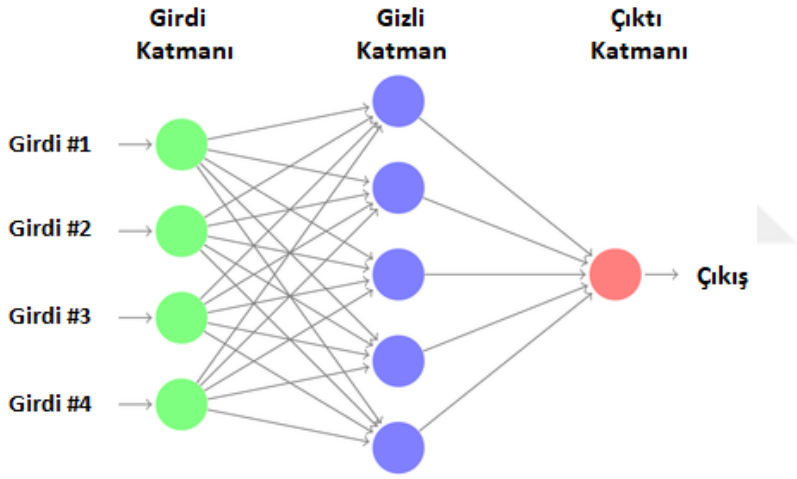
K-en yakın komşu yöntemi mesafeye dayalı olarak sınıflandırma yapan bir algoritmadır. Bu yöntem, hangi sınıfa ait olduğu bilinmeyen örneğe sınıf etiketi vermek için, eğitim kümesindeki örneklerin bu örneğe olan uzaklık ölçüsü hesaplanır. Kendisine en yakın örnekler (mesafe ölçüsü en küçük olan örnekler) seçilerek bu örneğin sınıf bilgisi yeni gelen örneğe verilir. Buradaki "k" değeri en yakın kaç komşuya bakılacağını yani komşu sayısını belirtmektedir. Sınıf etiketinin çoğunluk seçimiyle belirlenmesinden dolayı "k" değeri genellikle 3, 5 veya 7 gibi tek sayıda örnek seçilir. K-EYK algoritmasında komşular arasındaki uzaklık genellikle öklid uzaklığı ile bulunmakla birlikte mahalanobis, hamming, manhattan gibi uzaklık ölçütleri kullanılabilir. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'knn' fonksiyonu kullanılmıştır [77]. Şekil 3.5'de k-en yakın komşu yönteminin temel mantığı görülmektedir.



Şekil 3. 5 K - en yakın komşu yönteminin genel yapısı

3.2.4 Yapay Sinir Ağları Algoritması

Diğer adıyla sinir ağları, beyin hücreleri olan nöronların çalışma prensibini modelleyen, öğrenebilen algoritmalardır. İleriye yönelik ya da geri besleme alabilen yenilemeli ağlar olmak üzere iki tür yapısı vardır. YSA, ağırlıklı bağlantılar denilen tek yönlü iletişim kanalları sayesinde birbiriyle haberleşen, her biri kendi hafızasına sahip birçok nörondan oluşan paralel ve dağıtık bilgi işleme yapılarıdır. Yapı kurulduktan sonra sinir ağacı eğitilir. Giriş verilerine karşılık çıkış verileri alınır. Bu değer gerçek değerlerle karşılaştırılır ve ağın içerisindeki nöron fonksiyonlarının bu sonuçtaki hata miktarına göre ayarlanması sağlanır. Bu şekilde birçok değer ağa verilir ve ağın eldeki verinin yapısının öğrenilmesi sağlanır. Öğrenme işlemi tamamlandıktan sonra sinir ağı kullanıma hazır hale gelir [79]. Yapay sinir ağı yönteminin genel yapısı Şekil 3.6'da gösterilmiştir. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'nn' fonksiyonu kullanılmıştır [77].

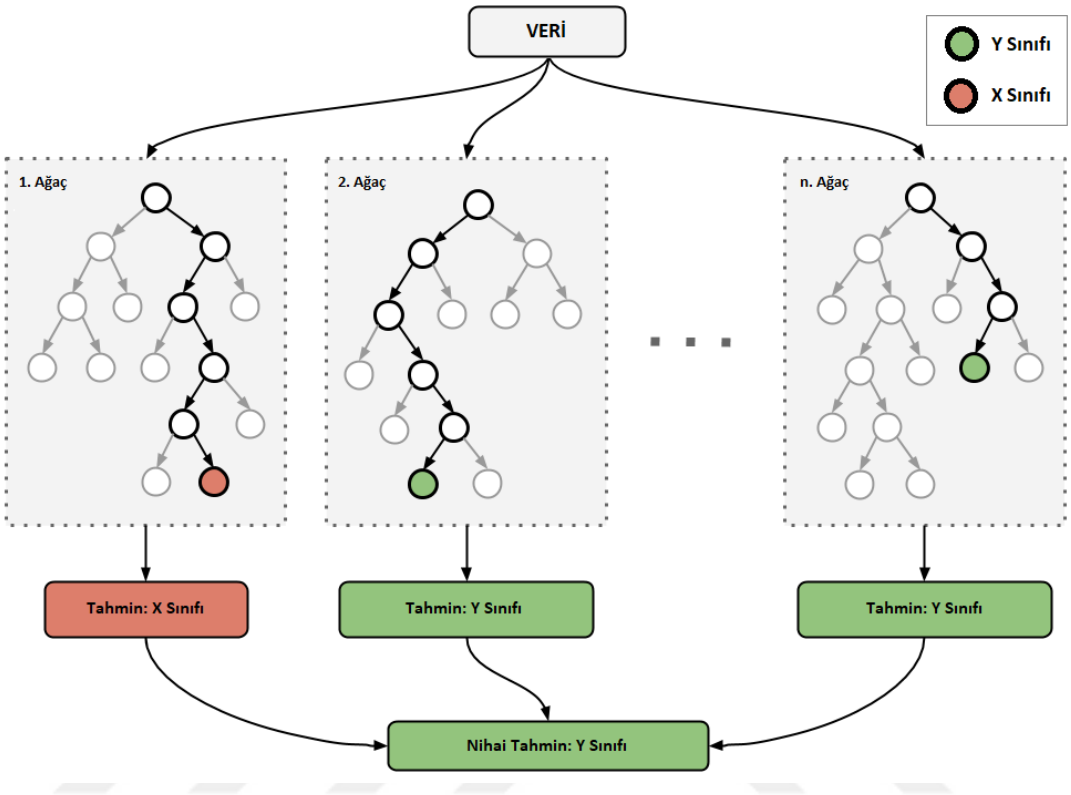


Şekil 3. 6 Yapay sinir ağı yönteminin genel yapısı

3.2.5 Rastgele Orman Sınıflandırma Algoritması

Birçok karar ağacının birleştirilmesi ile oluşan bir topluluk yöntemidir [1]. Topluluk öğrenme yöntemlerinde (ensemble learning) birden çok sınıflayıcının ortaya koyduğu sonuçlar bir araya getirilerek, topluluk adına tek bir karar verilmektedir. Ormandaki her karar ağacı, orijinal veri setinden bootstrap (yeniden yerleştirilerek örneklenen) tekniği ile farklı örneklemler seçilerek oluşturulur ve yine rastgele torbalama (bagging) mekanizması ile seçilen bir öznitelik kümesi ile eğitilir [80]. Daha sonra, birbirinden farklı çok sayıda bireysel ağaç tarafından verilen kararlar oylamaya tabi tutar ve oylama

sonucunda en çok oyu alan sınıfı topluluğun (komitenin) sınıf tahmini olarak sunar. Rastgele orman yönteminin genel yapısı Şekil 3.7'de gösterilmiştir. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'rf' fonksiyonu kullanılmıştır [77].



Şekil 3. 7 Rastgele orman yönteminin genel yapısı

3.3 Model Değerlendirme ve Seçimi

Bilgisayar destekli tanıda, sınıflandırma algoritmalarından elde edilen sonuçların kıyaslanabilmesi için nesnel değerlendirme ölçütlerine ihtiyaç duyulmaktadır. Bu ölçütler modelin ne derece başarılı olduğunu değerlendirmek açısından önemlidir. Sınıflandırma sonucu elde edilen sonuçlar karışıklık matrisi (confusion matrix) ile ifade edilmektedir. Bu matris etiketli verilerin sınıflandırılması sonucunda verilerin öngörülen sınıflarını ve gerçek sınıflarını içerir. Matrisinin sütunları gerçek değerleri, satırları ise sınıflandırma sonucu elde edilen sonuçlara karşılık gelmektedir. Çizelge 3.1'de iki sınıfa ait karışıklık matrisi sunulmuştur.

Çizelge 3. 1 İki sınıf için oluşturulmuş örnek bir karışıklık matrisi

Karışıklık Matrisi		**Verinin Tahmin Edilen Sınıfı**	
		Pozitif (Hasta)	Negatif (Sağlıklı)
Verinin Gerçek Sınıfı	Pozitif (Hasta)	True Positive (TP)	False Negative (FN)
	Negatif (Sağlıklı)	False Positive (FP)	True Negative (TN)

Karışıklık matrisi tablosunda yer alan TP gerçekte hasta alan ve sınıflandırma sonucunda da hasta olarak etiketlenen örnek sayısını, FN gerçekte hasta olup sınıflandırma sonucunda sağlıklı olarak etiketlenmiş örnek sayısını göstermektedir. FP gerçekte sağlıklı olup sınıflandırma sonucunda hasta olarak etiketlenen örnek sayısını, TN gerçekte sağlıklı alan ve sınıflandırma sonucunda da sağlıklı olarak etiketlenen örnek sayısını gösterir.

Bu tez çalışması kapsamında sınıflandırma başarıları değerlendirilirken doğruluk (accuracy), dengeli doğruluk (balanced accuracy), seçicilik (specifity), duyarlılık (sensitivity / recall), kesinlik (precision), F-ölçütü (F-measure), kappa değeri ve ROC eğrisi altında kalan alan (AUC) ölçütleri kullanılmıştır.

3.3.1 Doğruluk

Doğruluk, modelin genel doğru sınıflandırma başarısı hakkında bilgi vermektedir. Doğru tahmin edilen örneklerin, tüm örneklerin sayısına oranıyla elde edilir ve 1'e yakın olması model başarımını ideale yaklaştırır [81]. Doğruluk, aynı duyarlılık ve seçicilik değerlerinde dahi sınıf içindeki örnek sayısının değişmesinden etkilenir [82]. Doğruluk oranı Eşitlik 3.6'daki gibi hesaplanır.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.6}$$

3.3.2 Dengeli Doğruluk

Literatürde sıklıkla doğruluk ölçütü kullanılmasına karşın dengeli doğruluk ölçütünün kullanılması daha tarafsız bir yaklaşımdır. Çünkü geleneksel doğruluk ölçütünde sadece tek bir sınıfa (hasta veya sağlıklı) ait doğruluk değeri ele alınırken dengeli doğrulukta her

iki sınıftan (hasta ve sağlıklı) elde edilen doğruluğun ortalaması ele alınmaktadır. Yani hastaların doğru tahmin edilmesinin yanında sağlıklıların da doğru tahmin edilmesi gerekmektedir. Sınıflandırıcı her iki sınıfta da eşit derecede iyi performans gösteriyorsa, bu terim geleneksel doğruluk değerine göre daha düşük bir orana sahip olur [83]. Dengeli doğruluk denklemi Eşitlik 3.7'deki gibi hesaplanır.

$$Dengeli\ Doğruluk = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) = \frac{1}{2}(\text{Duyarlılık } + \text{Seçicilik}) \quad (3.7)$$

3.3.3 Seçicilik

Testin, gerçek sağlamlar içinden sağlamları ayırma yeteneğidir. Yani gerçekte sağlıklı olan kuzuların ne kadarının sistem tarafından doğru tahmin edildiğini verir [71]. Seçicilik denklemi Eşitlik 3.8'deki gibi hesaplanır.

$$Seçicilik = \frac{TN}{TN+FP} \quad (3.8)$$

3.3.4 Duyarlılık

Testin, gerçek hastalar içinden hastaları ayırma yeteneğidir. Yani gerçekte hasta olan kuzuların ne kadarının sistem tarafından doğru tahmin edildiğini verir [71]. Duyarlılık denklemi Eşitlik 3.9'deki gibi hesaplanır.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (3.9)$$

3.3.5 Kesinlik

Kesinlik, model tarafından pozitif olarak atanan örneklerin hangi oranda doğru sınıfa atandığını verir. Hasta olarak doğru sınıflandırılmış örnek sayısının, model tarafından hasta olarak sınıflandırılan toplam örnek sayısına oranıdır [71]. Kesinlik denklemi Eşitlik 3.10'deki gibi hesaplanır.

$$Kesinlik = \frac{TP}{TP+FP} \quad (3.10)$$

3.3.6 F-Ölçütü

Başarım kıstasları hesaplanırken tek başına kesinlik veya duyarlılığın hesaplanması sistemin başarısını değerlendirmede eksik kalabilmektedir [71]. F-Ölçütü değeri, kesinlik

ve duyarlılık değerlerinin harmonik ortalaması alınarak hesaplanır. F-ölçütü Eşitlik 3.11'deki gibi hesaplanır.

$$F - \text{Ölçütü} = 2 * \frac{\text{Duyarlılık} * Kesinlik}{\text{Duyarlılık} + Kesinlik} \quad (3.11)$$

3.3.7 Kappa Değeri

Özellikle sağlık alanında yaygın olarak kullanılan kappa değeri, gerçek uyumluluk ölçütü olup rastgele beklenenin dışındaki uyumluluğu ölçen bir istatistik yöntemidir [84]. Araştırmacıların hastalık veya başka bir faktörün varlığını veya yokluğunu belirlemek için isimsel ölçekte uyumluluğun değerlendirilmesinde Kappa istatistiği çok sık kullanılmaktadır [85]. Kappa değeri sınıflandırıcı modelin doğru sınıflandırma başarımı hakkında bilgi verir. Landis ve Koch; kappa değerinin <0.00 olmasını zayıf, 0.00-0.20 aralığını önemsiz, 0.21-0.40 aralığını orta, 0.41-0.60 aralığını makul, 0.61-0.80 aralığını önemli ve 0.81-1.00 aralığını neredeyse mükemmel olarak yorumlamıştır [86].

3.3.8 ROC Eğrisinin Altında Kalan Alan (AUC)

ROC eğrisi, tanı testlerinin ayrım gücü yönünden başarım değerlendirmesinde yaygın olarak kullanılmaktadır. ROC eğrisi pozitif ya da negatif sınıf için belirlenen farklı eşik değerlerine göre bulunan 1-seçicilik ve duyarlılık değerlerinin sırasıyla x ve y koordinatlarına yerleştirilmesiyle elde edilir ve bu değerlerinin ilişkisi ROC eğrisi ile grafiksel olarak gösterilir.

ROC analizi özellikle tıp, veteriner hekimliği, radyoloji, psikoloji, makine öğrenme teknikleri ve veri madenciliği alanlarında geniş bir uygulama alanına sahiptir.

ROC eğrisinin altında kalan alan, ROC [87] eğrisini özetleyen ortalama bir performans değeri verir. AUC, testin hastalar ile hasta olmayan bireyleri ayırmadaki doğruluk oranını belirler. 0 - 1 arasında değer alan eğri altındaki alan boyutu 1'e yaklaştıkça sınıflandırıcı model başarımı da o derece yüksek olacaktır. 0.5 değerinin altında elde edilen sonuç modelin başarısız olduğunu gösterir [82], [88]. Bu yöntemi uygulamak için R'da 'pRoc' paketindeki 'roc' fonksiyonu kullanılmıştır [89].

BÖLÜM 4

VERİ VE YÖNTEM

4.1 Hayvan Materyali

4.1.1 Hayvanlar, Veri Toplama ve Çiftlik Yönetimi

Veri seti 2009 yılında Kuzeydoğu Anadolu bölgesinde Kars ilinde bulunan iki koyun çiftliğinden ilgili alandaki araştırıcılar tarafından TÜBİTAK projesi (Proje Kodu: TOVAG 108 O 847) kapsamında toplanmıştır [90], [91], [92]. Kuzeydoğu Anadolu bölgesindeki sürülerde kuzulama sezonu tipik olarak kış (Aralık-Şubat) veya bahar (Mart-Mayıs)'dır. Tez çalışmada kullanılan veri seti 301 Akkaraman melezi koyun ve bunlardan doğan 347 kuzuya aittir. Tüm koyun ve kuzular aynı beslenme şekli ve işletme koşullarında tutulmuştur. Doğumda anne ve kuzulara kulak küpesi uygulanarak hastalık ve diğer bilgiler bu numara üzerinden takip edilmiştir. Doğumda kuzuların cinsiyet, doğum tarihi, doğum ağırlığı, anne doğum şekli ve doğan kuzu sayısı (ikiz/tek) kayıt altına alınmıştır. Kuzular doğumda (kolostrum almadan önce) bir baskül yardımıyla tartılmıştır. Bu işlemden sonra bir hafta boyunca kuzuların, annelerini doğal yolla emmelerine izin verilmiştir. Bu süre zarfında kuzular ilave kolostrum ile beslenmemiştir. Bu dönemden sonra kuzular ayrı bir ağıla sevk edilmiş olup üç ay boyunca günde iki kere (sabah ve akşam) annelerinin emzirmesine izin verilmiştir.

4.1.2 Klinik Muayene

Kuzuların sağlık muayeneleri neonatal periyotta günlük, post-neonatal dönemde (yaşamının 5 ila 12 haftalık dönemini kapsayan süre) 2 günde bir yapılan ziyaretlerle ilgili alandaki araştırıcılar [90], [91], [92] tarafından yapılarak hastalıklar kayıt altına alınmıştır. Araştırma süresi boyunca hastalık (mastitis, pnömoni, enterit, gebelik toksoemisi vb.) olduğu tespit edilen koyunlar hasta olarak sınıflandırılmış ve kulak etiketi numarası ile kaydedilmiştir.

4.1.3 Kan Örneklerinin Toplanması

Kan örnekleri doğumdan sonra 24±1 saat sonra alınmıştır. Kan örnekleri 3000 rpm'de 5 dakika santrifüj edildikten sonra serum örnekleri elde edilerek -20^{o} de saklanmıştır. Serum ve kolostral IgG konsantrasyonları ELISA kit kullanılarak ölçülmüştür. Aynı şekilde serum ve kolostrumda total protein, Gamma Glutamil Trasnfereraz (GGT) ve Albumin (ALB) analizleri ticari spektrofotometrik kitler kullanılarak ölçülmüştür. Orijinal veri setinde 347 kayıt (satır), 42 nitelik (sütun) bulunmaktadır. Orijinal veri setindeki niteliklere detaylı bilgi Çizelge 4. 1'de verilmiştir.

Çizelge 4. 1 Ham veri setindeki nitelikler

	Özellikler	Kısaltma	İçerik
1.	Kuzu Numarası	lambno	Numerik
2.	Immunoglobulin G	IgG	Numerik
3.	Gamma Glutamil Trasnfereraz	GGT	Numerik
4.	Laktoferrin	LT	Numerik
5.	Total Protein	TP	Numerik
6.	Albumin	ALB	Numerik
7.	Doğum Ağırlığı	BW	Numerik
8.	28. Gün Sonundaki Canlı Ağırlık	WG28	Numerik
9.	İlk 28 Gün Sonundaki Günlük Canlı Ağırlık Kazanımı	MDG28	Numerik
10.	84. Gün Sonundaki Canlı Ağırlık	WG84	Numerik
11.	İlk 84 Günlük Periyottaki Canlı Ağırlık Kazanımı	MDG84	Numerik
12.	28-84 Gün Aralığındaki Toplam Canlı Ağırlık Kazanımı	WGPN	Numerik

Çizelge 4. 1 Ham veri setindeki nitelikler (devam)

13.	28-84 Gün Aralığındaki Günlük Canlı Ağırlık Kazanımı	MDGPN	Numerik
14.	Anne Hastalık Durumu	AH	Nominal {0=Sağlıklı, 1=Hasta}
15.	Neonatal Hastalık Durumu	NGH	Nominal Nominal {0=Sağlıklı, 1=Hasta}
16.	Neonatal Periyot Hastalık Sonucu	NHS	Nominal Nominal {0=Sağlıklı, 1=Hastalandı Yaşıyor, 2=Öldü}
17.	Neonatal Dönemde Hastalandığı Gün	DAYI	Numerik
18.	Hastalandığı Hafta	DAYIG	Numerik
19.	1. Hafta Hastalananlar	1week	Nominal {0=Sağlıklı, 1=Hasta}
20.	Hastalığın Teşhisi	NT	Nominal {0=Sağlıklı, 1=İshal, 2=Pnömoni, 3=Septisemi, 4=Halsizlik, 5=Pmöoenteritis}
21.	28-84 Günlük Periyotta Hastalık Takibi	PNGH	Nominal {0=Sağlıklı, 1=Hasta, 2=Öldü}
22.	Post-neonatal Periyot Hastalık Sonucu	PNHS	Nominal {0=Sağlıklı, 1=Hastalandı Yaşıyor, 2=Öldü}
23.	28-84 Günlük Aralıkta Hangi Gün Hastalandığı	DAYII	Numerik
24.	Hastalık Nedeni	PNT	Nominal {0=Sağlıklı, 1=İshal, 2=Pnömoni, 3=Septisemi, 4=Halsizlik, 5=Pmöoenteritis, 6=Diğer }
25.	İlk 3 Ay Genel Hastalık Takibi	PWGH	Nominal {0=Sağlıklı, 1=Hasta, 2=Öldü}
26.	İlk 3 Ay Genel Hastalık Sonucu	PWHS	Nominal {0=Sağlıklı, 1=Hastalandı Yaşıyor, 2=Öldü}
27.	İlk 3 Ayda Ne Zaman Hastalandığı	DAYIII	Numerik
28.	Neonatal ve Post-neonatal Dönemde Hastalananların Gruplandırılması	DAYIIIG	Numerik
29.	İlk 3 Ay Genel Teşhis	PWT	Nominal {0=Sağlıklı, 1=İshal, 2=Pnömoni, 3=Septisemi, 4=Halsizlik, 5=Pmöoenteritis, 6=Diğer}
30.	Anne Yaşı	AGE	Numerik
31.	Anne Doğum Sayısı	PARITY	Numerik
32.	İkizlik	TWIN	Nominal {0=İkiz, 1=Tek}
33.	Cinsiyet	GENDER	Nominal {0=Dişi, 1=Erkek}
34.	İşletme	FARM	Nominal {0,1}
35.	Immunoglobulin G	IgGK	Numerik
36.	Gamma Glutamil Trasnfereraz	GGTK	Numerik
37.	Laktoferrin	LTK	Numerik
38.	Total Protein	PRTK	Numerik
39.	Albumin	ALBK	Numerik
40.	Immunoglobulin G	IgGA	Numerik
41.	Gamma Glutamil Trasnfereraz	GGTA	Numerik
42.	Total Protein	TPA	Numerik

Veri setindeki IgG, GGT, LT, TP ve ALB nitelikleri; kuzunun ilk süt/kolostrum seviyesini göstermektedir. Doğumdan 24 sonra ölçülmektedir. Kolostrum kuzular için önemli bir enerji kaynağı olup, yeterince alınmadığı taktirde kuzular hastalanır veya ölebilirler. Çünkü kuzularda plasental yapıdan dolayı anneden yavruya başta hastalıklara karşı koruyucu antikorlar olmak üzere yaşam için gerekli olan birçok maddenin geçişi olmaz. Kuzuların hastalıklarından korunması ve normal gelişmesi için gerekli olan tüm maddeler annelerin doğumdan sonra ürettikleri kolostrumda bulunmaktadır. Bu nedenle kolostrumun yeterli alınması oldukça önemelidir ve yetersizliği doğumdan sonra ilk 24 saatte ölçülen bu kan parametreleri ile belirlenebilmektedir. Özellikle neonatal dönemde gelişen hastalıklarda kolostrumda bulunan maddelerin yeterli alınmasıyla doğrudan ilişkilidir. Yani kolostrum seviyesi ne kadar düşük olursa kuzunun hastalanma/ölme riski o kadar yüksektir.

Veri setindeki IgGK, GGTK, LTK ve PRTK nitelikleri; annedeki kolostrumda ilgili parametrelerin seviyesini göstermektedir. Annedeki ilk süt yani kuzu doğduktan sonra annesini emmeden önce anneden alınan kan örnekleridir.

Veri setindeki IgGA, GGTA ve TPA nitelikleri; annede doğum yapmadan 10-15 gün önce anneden alınan kan örnekleri olup doğum öncesi kolostrum seviyesini göstermektedir.

BW; kuzunun doğum ağırlığını göstermekte olup doğumdan hemen sonra, kolostrumu almadan önce baskül yardımıyla ölçülmektedir.

Kuzunun kilo kazanımı takip etmek için kuzular neonatal ve post-neonatal dönemlerde baskül yardımıyla tartılmış ve bu sonuçlar şu şekilde kayıt altına alınmıştır. WG28; kuzunun 28. Gün sonundaki (neonatal dönemdeki) canlı ağırlığını, MDG28; neonatal dönemdeki günlük kilo kazanımını, WG84; kuzunun 84. gün sonundaki (post-neonatal dönemdeki) canlı ağırlığını, MDG28; post-neonatal dönemdeki günlük kilo kazanımını göstermektedir. Bu iki dönemin tamamını kapsayan (28 - 84 gün) kilo kazanımı WGPN, günlük kilo kazanımı ise MDGPN niteliklerinde tutulmaktadır.

AH; annenin doğumdan önceki sağlık durumunu (sağlıklı/hasta) göstermektedir.

NGH; neonatal dönemde kuzunun sağlık durumunu (sağlıklı/hasta) göstermektedir.

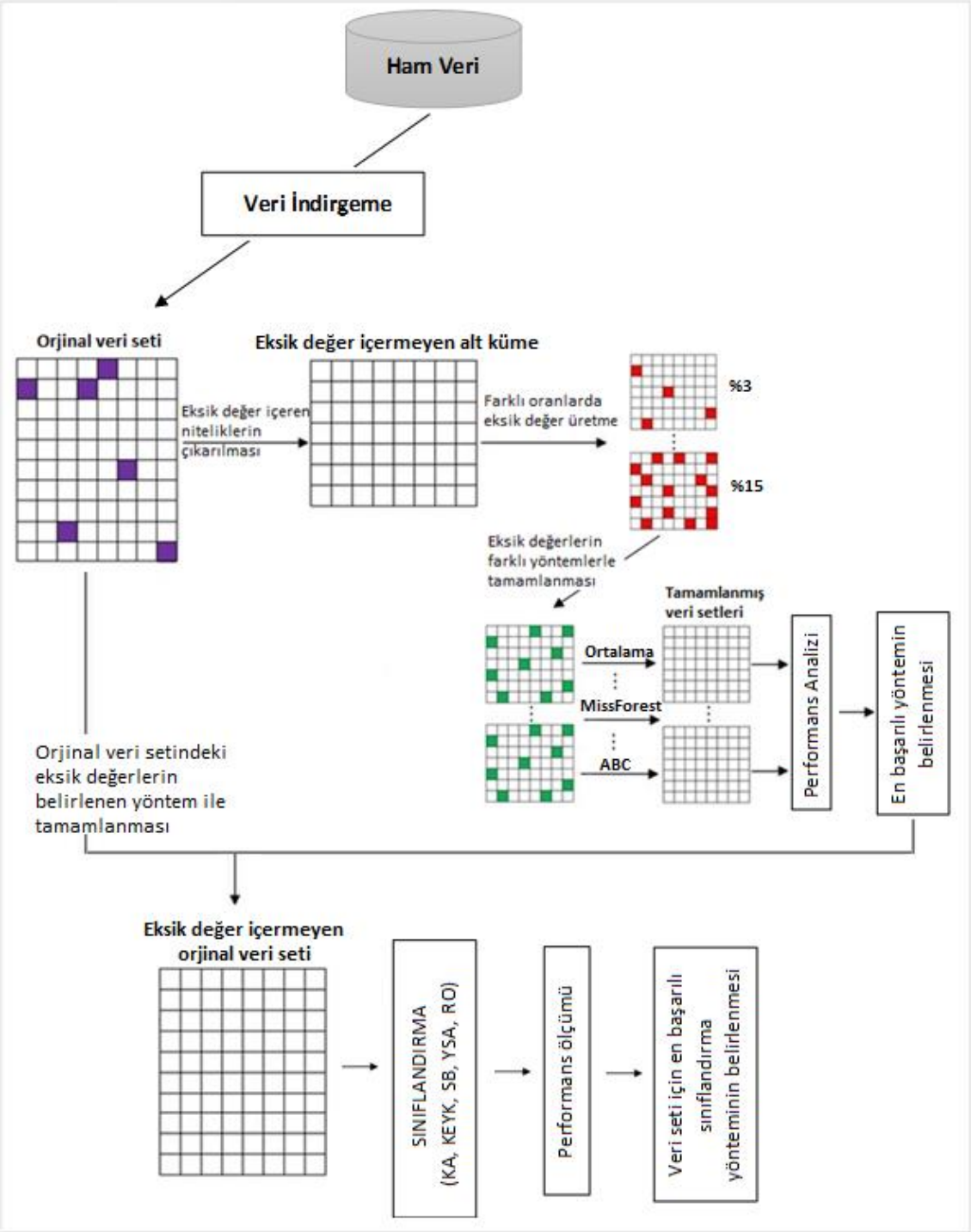
NHS; neonatal dönemin sonunda kuzunun sağlık durumunun sonucunu göstermektedir. Yani kuzunun neonatal dönem sonunda sağlıklı, hasta veya hastalandı fakat yaşamını sürdürmekte olduğu bilgisi kayıt altına alınmıştır.

NT; neonatal dönemde kuzu hastalığının teşhisi yani sağlıklıysa sağlıklı olduğunu, hasta ise hangi hastalık (ishal, pnömoni, septisemi, halsizlik, pmöoenteritis ve diğer) olduğu bilgisi kayıt altına alınmıştır.

Veri setinde ayrıca annenin yaşı (age), annenin doğum sayısı (parity), kuzunun tek veya ikiz doğduğu bilgisi (twin), cinsiyeti (gender) ve hangi işletme (farm) olduğu bilgilerde kayıt altına alınmıştır.

4.2 Yöntem

Tez kapsamında izlenen yöntem Şekil 4. 1'de verilmiştir.



Şekil 4. 1 Tez kapsamında izlenen yöntem

Bu tez çalışmasında ham veriye veri indirgeme adımı uygulandıktan sonra sınıflandırma için kullanılacak veri seti elde edilir. Bu veri seti eksik değerler içermektedir. Eksik değerlerin nasıl tamamlandığı önemli bir konu olup veri madenciliği sonuçlarını etkilemektedir. Bu nedenle veri setindeki eksik değerleri tamamlamak için literatürde sıklıkla kullanılan 5 yöntemden (ortalama, ortanca, knn, mice, missForest) ve geliştirilen

ABC yönteminden yararlanılarak veri seti için en başarılı yöntemin tespit edilmesi amaçlanmıştır. Bunun için orijinal veri setinde eksik değer içerin gözlemler veri setinden çıkarılmıştır. Elde edilen eksiksiz orijinal veri seti üzerinde yaklaşık %3, %5, %7, %10, %12 ve %15 oranlarında rastgele eksik değer üretilmiştir. Eksik değer içeren bu veri setleri ortalama, ortanca, kNN, mice, missForest ve ABC yöntemleriyle tamamlanmıştır. Bu yöntemlerin başarıları RMSE ölçütüne göre karşılaştırılarak en az hata yapan yöntem yani en başarılı eksik değer tamamlama yöntemi belirlenmiştir.

Sonra, orijinal veri setindeki eksik değerler belirlenen bu en başarılı yöntem ile tamamlanmıştır. Veri setindeki eksik değerler tamamlandıktan sonra kuzularda hastalık sınıflandırması yapılmıştır. Sınıflandırmada modelin başarısı seçilen eğitim ve test kümesine bağlı orak değişmektedir. Örneğin modelin eğitiminde kullanılan veri kümesi test örneklerini içeriyorsa çok yüksek oranda başarı elde edilecektir. Yine bu gibi yanlılıkları ortadan kaldırmak için sınıflandırma yöntemleri uygulanmadan önce veri seti %70 eğitim kümesi, %30 doğrulama kümesi olarak ayrılmıştır. Sınıflandırma modelleri bu eğitim kümesi eğitilmiştir. Her bir farklı eğitim kümesi için 10-kat çapraz geçerlilik (cross-validation) uygulandıktan sonra model başarısı daha önce %30 oranında ayrılan doğrulama veri seti ile ölçülmüştür. Model başarısını değerlendirmek için ise doğruluk, dengeli doğruluk, duyarlılık, seçicilik, kesinlik, f-ölçütü, kappa ve AUC kriterleri kullanılmıştır.

Daha sonra, özellik seçim tekniklerinden olan bilgi kazanımı yöntemi ile hastalıklarda hangi özelliklerin risk faktörü olduğu ve özellik sayısının azaltılmasında sınıflandırma performansının nasıl değiştiğini gözlemlenmiştir. Ayrıca kümülatif dağılım fonksiyonu) ve ortak bilgili yöntemlerini kullanarak ölümler için kan değerlerine ait eşik seviyesi belirlenmiştir.

Son olarak, verilerin kolay bir şekilde bilgisayar ortamında tutulabilmesi için bir veri tabanı oluşturulmuştur. Bu veri tabanına bilgi girişini kolaylaştırabilmek için masaüstü ve mobil arayüzler tasarlanmıştır. Ayrıca bu arayüzler üzerinden basit istatistiksel analizlerde yapılabilmektedir. Tez çalışmasında, hem masaüstü hem de mobil tarafı JAVA SE kullanılarak geliştirilmiştir. JAVA SE seçilmesindeki en büyük etken iki platformu aynı teknoloji ile geliştirip kod uyumu yakalamaktır. Veri tabanı kısmında Postgresql,

Webserver olarak Nodejs, View olarak masaüstü için Swing kütüphanesi, Mobil için Android studio'nun kendi bileşenleri ve analizler için de R dili kullanılmıştır.

BÖLÜM 5

UYGULAMA

5.1 Veri Ön İşleme Tekniklerinin Veri Setine Uygulanması

Orijinal veri setinde 347 kayıt (satır), 42 nitelik (sütun) bulunmaktadır (Çizelge 4.1). Orijinal veri setindeki bazı niteliklerin değerleri eksiktir. Bu niteliklerin eksiklik oran ve yüzdeleri Çizelge 5.1'de verilmiştir.

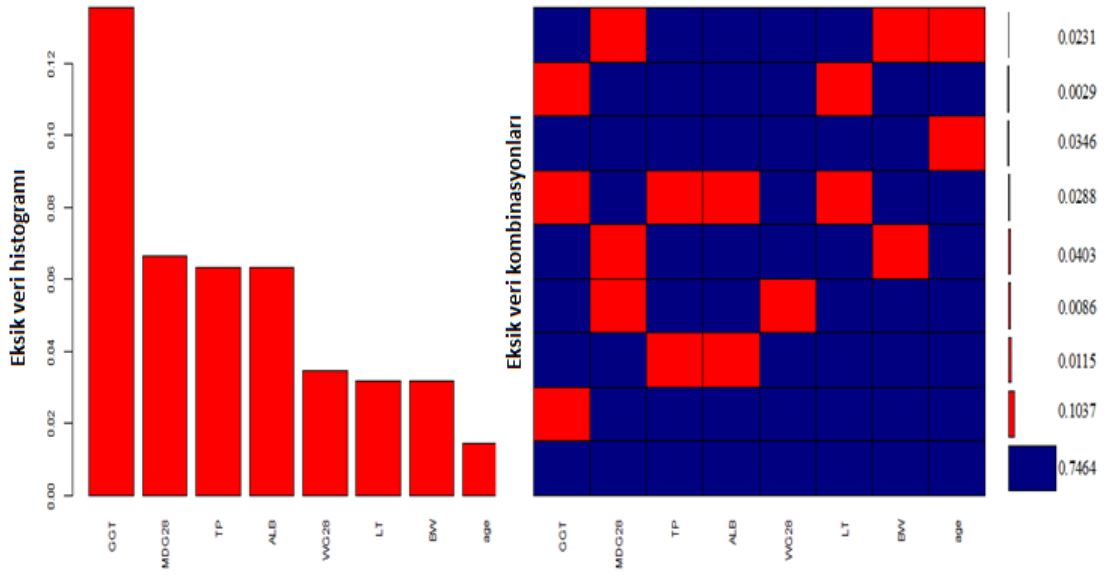
Çizelge 5. 1 Eksik değer içeren nitelikler, eksiklik oran ve yüzdeleri

Nitelik	Oran	Yüzde	Nitelik	Oran	Yüzde
GGT	47/347	13.5	GGTK	73/347	21
LT	11/347	0.3	PRTK	75/347	21
TP	22/347	0.6	ALBK	79/347	21.6
ALB	22/347	0.6	IgGK	178/347	51.3
BW	11/347	0.3	LTK	154/347	44.4
WG28	12/347	0.35	IgGA	306/347	88.2
MDG28	23/347	0.66	ALBA	306/347	88.2
WG84	31/347	0.89	TPA	306/347	88.2
MDG84	42/347	12			
WGPN	33/347	9.5			
MDGPN	33/347	9.5			
Age	5/347	1.44			
Parity	5/347	1.44			

Veri setinde kuzu doğduktan hemen sonra alınan kan örnekleri 347 kuzuya ait olup bu kuzuların emzirildikten sonra annelerinden doğumdan sonra ilk 4 saat içerisinde alınan kolostrum seviyelerini gösteren IgGK, GGTK, PRTK, LTK ve ALBK özelliklerindeki eksik veri oranı oldukça yüksektir. Aynı şekilde annelere doğumdan 10-15 gün önce alınan ve

kuzularda ölçülen parametrelerin içeren IgGA, ALBA ve TPA özelliklerinin eksiklik yüzdeleri yüksek olduğundan bu nitelikler veri setinden çıkarılmıştır.

Daha önce de belirtildiği gibi kuzu ölümleri büyük oranda neonatal dönemde gerçekleşmekte olup, doğumdan sonra kuzunun anneden kolostrumu yeterli alıp almaması ile ilgilidir. Çünkü kuzuların hastalıklarından korunması ve normal gelişmesi için gerekli olan tüm maddeler annelerin doğumdan sonra ürettikleri ilk süt/kolostrumda bulunmaktadır. Alınan kolostrumun yetersizliği doğumdan sonra ilk 24 saatte ölçülen IgG gibi çeşitli kan parametreleri ile belirlenebilmektedir. Ancak bu durum post-neonatal dönemde etkisini kaybederek hastalıkların gelişmesinde işletmenin fiziksel ve çevresel koşulları, aşılama gibi faktörlerin etkili olduğu bilinmektedir [90]. Bu nedenle veri setinden post-neonatal bilgiler çıkarılarak analizlerin neonatal kuzular üzerinde yürütülmesine karar verilmiştir. Tez çalışmasında analizlerin doğru bir şekilde yapılabilmesi için veri setinden neonatal kuzuların hastalık durumu ile ilişkisiz veya doğrudan hastalıklarla ilişkisi olan özellikler çıkartılmıştır. Sonuç olarak tez çalışmasında 347 (60 hasta, 287 sağlıklı) gözlem, 14 nitelik ve 1 sınıf etiket bilgisi kullanılmıştır. Bu niteliklerdeki eksik değerlerin oranları ve niteliklerin birlikte ne oranda eksik değer içerdikleri Şekil 5.1'de verilmiştir.



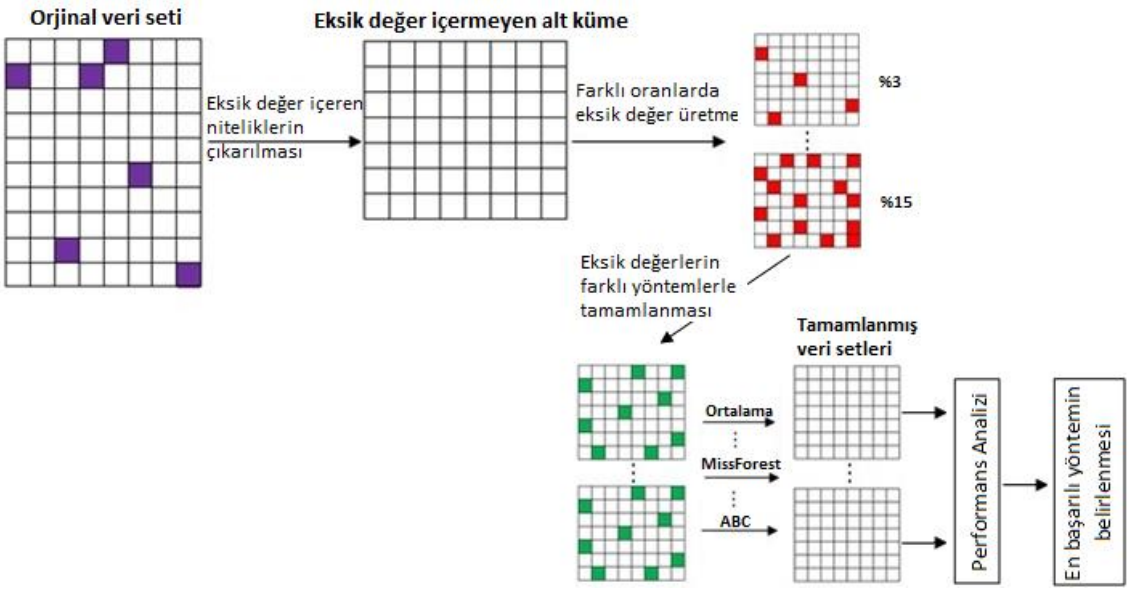
Şekil 5. 1 Özelliklerin eksik veri oranları ve kombinasyonları

Şekil 5.1'de sütunlar özelliklerin eksiklik oranlarını gösterirken satırlar özelliklerin birlikte eksik değer içerme oranını göstermektedir. Şekildeki sütunlar incelendiğinde IgG, anne

hastalık durumu (AH), ikizlik (twin), cinsiyet (gender), çiftlik (farm) ve sınıf etiketi yani neonatal kuzunun hastalık durumunu gösteren özelliklerde tüm hücreler yeşil olup eksik değer içermemektedir. En fazla eksik değere GGT özelliği sahip olup onu sırasıyla MDG28, TP, ALB, WG28, LT, BW, yaş ve doğum sayısı özellikleri takip etmektedir. Eksik değer içeren özellikler içerisinde kan seviyelerini gösteren 5 özellikten (IgG, GGT, TP LT ve ALB) 4'nün eksik değer içerdiği ve GGT kan seviyesi özelliğinin diğer kan seviyelerine göre neredeyse iki katından fazla eksik değer içerdiği görülmektedir.

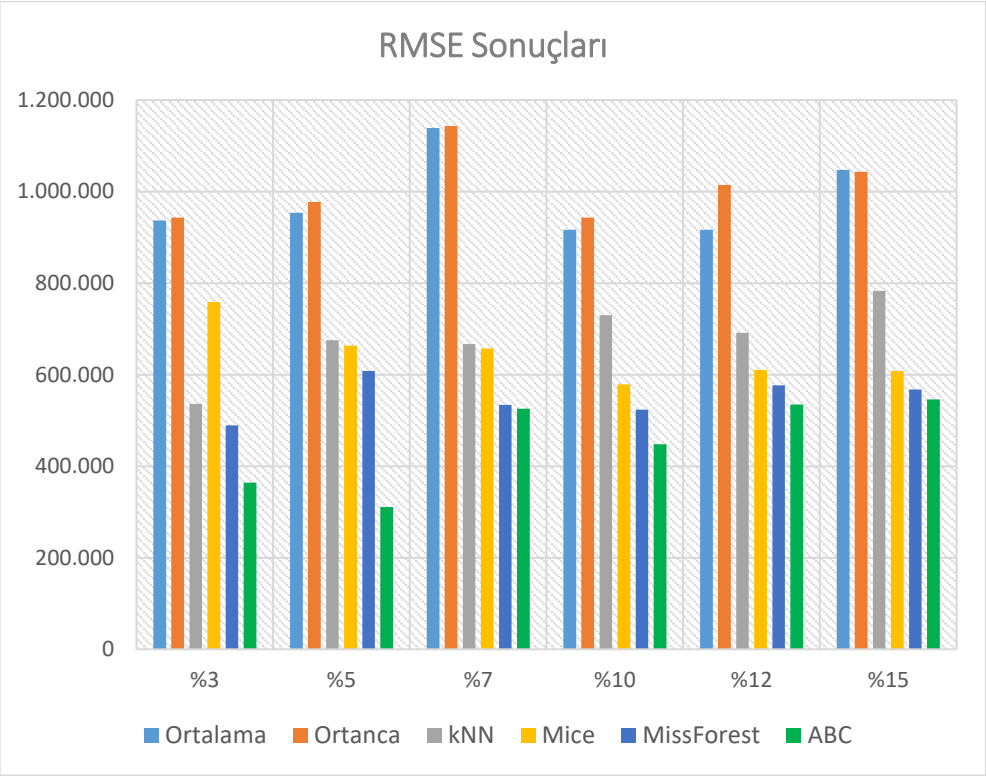
Şekil 5.1'deki satırlar incelendiğinde veri setindeki örneklerin yaklaşık %75'inin eksik değer içermediği, yaklaşık %14'nün sadece GGT özelliğinin eksik olduğu, yaklaşık %3'ünde dört kan seviyesinin (GGT, TP, LT ve ALB) birlikte eksik olduğu, yaklaşık %3'ünde sadece yaş özelliğinin eksik olduğu görülmektedir. Sonuç olarak GGT özelliğindeki eksik değer oranı diğer özelliklerdeki eksik değer oranından 2 ile 10 kat daha fazla olduğu gözlemlenmiştir.

Bölüm 3'de de değinildiği gibi eksik değerler analizlerin yapılmasında sorun teşkil ettiğinden, bunların tamamlanması gerekmektedir. Ancak literatürde eksik değerleri tamamlamak için çok fazla yöntem geliştirilmiştir. Hangi yöntemin eksik değerleri tamamlamada daha başarılı olduğunu tespit edebilmek için; eksik değerler ortalama, ortanca, kNN, mice, missForest ve geliştirilen ABC yöntemi ile tamamlanmıştır. Daha sonra bu yöntemlerin performansları RMSE sonuçlarına göre karşılaştırılarak veri seti için en başarılı yöntem belirlenmiştir. Eksik değerleri tamamlamada izlenilen süreç Şekil 5.2'de gösterilmiştir.



Şekil 5. 2 Veri setindeki eksik değerlerin tamamlanması süreci

Öncelikle orijinal veri setinde eksik değer içeren nitelikler veri setinden çıkartılmıştır. Elde edilen eksiksiz veri seti, orijinal veri setinin bir alt kümesidir. Eksik değer içermeyen bu veri seti üzerinde yaklaşık %3, %5, %7, %10, %12 ve %15 oranlarında eksik değer oluşturulmuştur. Farklı oranlarda eksik değer içeren bu 6 veri setindeki eksik değerler ortalama, ortanca, kNN, mice, missForest ve ABC ile tamamlanmıştır. Daha sonra bu yöntemlerin performanslarını değerlendirmek için RMSE sonuçları karşılaştırılarak veri seti için en başarılı eksik değer tamamlama yöntemi belirlenmiştir. Elde edilen RMSE sonuçları Şekil 5.3'de sunulmuştur.



Şekil 5. 3 Eksik değer tamamlama yöntemlerinin RMSE sonuçları

RMSE değeri hata miktarının büyüklüğünü ifade ettiği için bu değerlerin düşük olması yöntemin başarısını arttırmaktadır. 6 farklı eksik değer tamamlama yöntemi için hesaplanan ortalama RMSE değerleri incelendiğinde literatürde eksik veri tamamlamada sıklıkla kullanılan ortalama ve ortanca yöntemlerinin birbirine yakın ve istatistiksel yöntemlere göre çok daha yüksek hata değerine sahip olduğu görülmektedir. Geliştirilen ABC yöntemi ise en düşük RMSE değerine sahip olarak en başarılı yöntemdir.

Bu tez çalışmasında kullanılan veri seti için en başarılı eksik veri tamamlama yönteminin ABC olduğu belirlenmiş olup eksik değerler bu yöntem ile tamamlanmıştır. Veri setindeki eksik değerler tamamlandıktan sonra özellikler ve bu özelliklere ait sonuçlar Çizelge 5.2'de verilmiştir. Çalışmada bağımsız grup t-testi (Independent Samples t Test) kullanılmıştır.

Çizelge 5. 2 Veri setindeki özellikler ve bu özelliklere ait istatistiksel sonuçlar

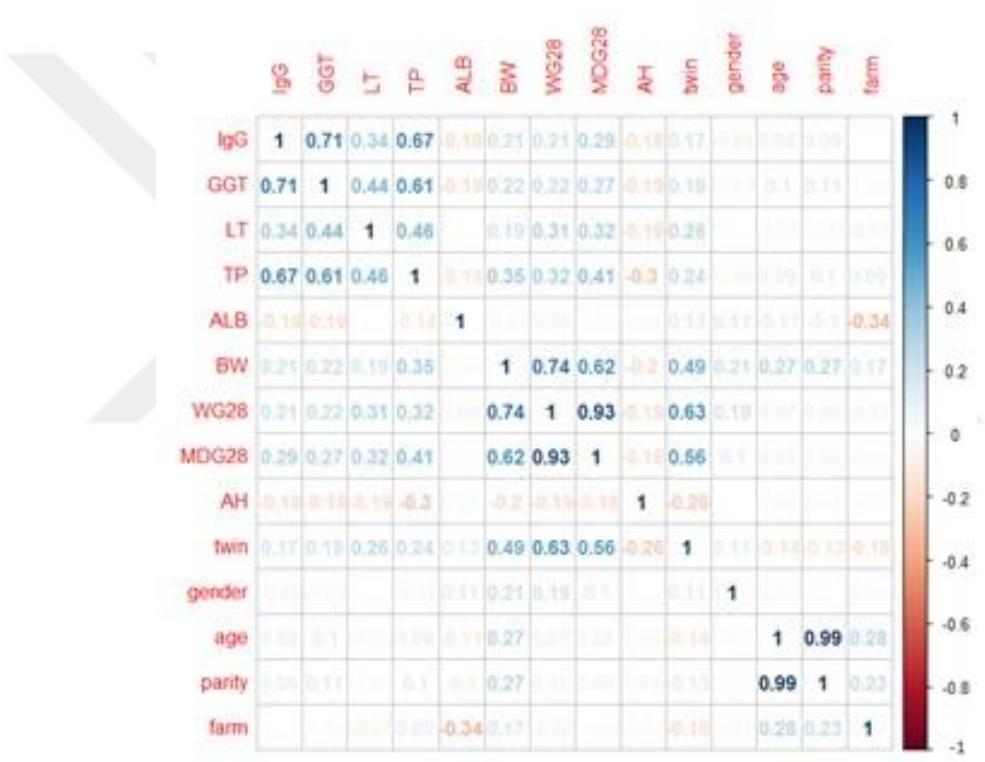
Nümerik nitelikler					
Özellikler	**Min.**	**Maks.**	**Ort.**	**Ortalama (Hasta, n=60)**	**Ortalama (Sağlıklı, n=287)**
IgG****	19	5302	2196	1526±1268	3114 ± 1121
GGT**	38	7517	2382	1780±1655	2855±1410
LT**	354	2194	1052	955±312	1064±321
TP****	21	117	73	62±18	78±11
ALB*	32	51	40	41±4	41±4
BW****	2260	5900	4028	3641±708	4143±593
WG28****	4364	14016	8804	7377±1975	9378±1913
MDG28****	14	340	169	130±53	185±55
Age	1	6	3	3±1	3±1
Parity	1	5	2	2±1	2±1
Kategorik nitelikler					
Özellikler	**Hasta Kuzu (n = 60)**			**Sağlıklı Kuzu (n = 287)**	
AH**	Hasta n=11	Sağlıklı n=49		Hasta n=6	Sağlıklı n=281
Twin	İkiz n=19	Tek n=41		İkiz n=73	Tek n=214
Gender	Erkek n=22	Dişi n=38		Erkek n=134	Dişi n=153
farm*	Farm1 n=12	Farm2 n=48		Farm1 n=18	Farm2 n=269
**** P<0.0001, *** P<0.001, ** P<0.01, * P<0.05, n=Hayvan sayısı					

Öncelikle, kuzularda hastalık risk faktörlerini en basit düzeyde inceleyecek olursak; hasta olan grup ile sağlıklı olan grubun IgG, GGT, LT, TP, ALB, BW, WG28, MDG8, AH ve farm değerleri ortalamaları istatistiksel olarak anlamlı farklılık göstermiştir.

Risk faktörleri ile hastalık varlığının ilişkisi incelendiğinde; hastalık varlığı ile IgG seviyesinin düşük olması arasında anlamlı ilişki saptanmıştır. Hasta olanların ortalama

IgG seviyesi 1526±1268 iken, sağlıklı olan kuzuların ortalama IgG seviyesi 3114 ± 1121 olarak saptanmıştır (p=1.768e-05). Hastalık varlığı ile GGT seviyesinin düşük olması arasında anlamlı ilişki saptanmıştır. Hasta olanların ortalama GGT seviyesi 1780±1655 iken, sağlıklı olan kuzuların ortalama GGT seviyesi 2855±1410 olarak saptanmıştır (P=0.002). Aynı şekilde LT (P=0.009), TP (P=2.753e-07), ALB (P=0.0397), BW (p=1.086e-05), WG28 (P=02.514e-08), MDG28 (P=1.014e-08), özellikleri de istatistiksel olarak anlamlı bulunmuştur. Diğer risk faktörlerinden anne yaşı, anne doğum sayısı, ikizlik ve cinsiyet ile hastalık varlığı arasında istatistiksel olarak anlamlı ilişki saptanmamıştır.

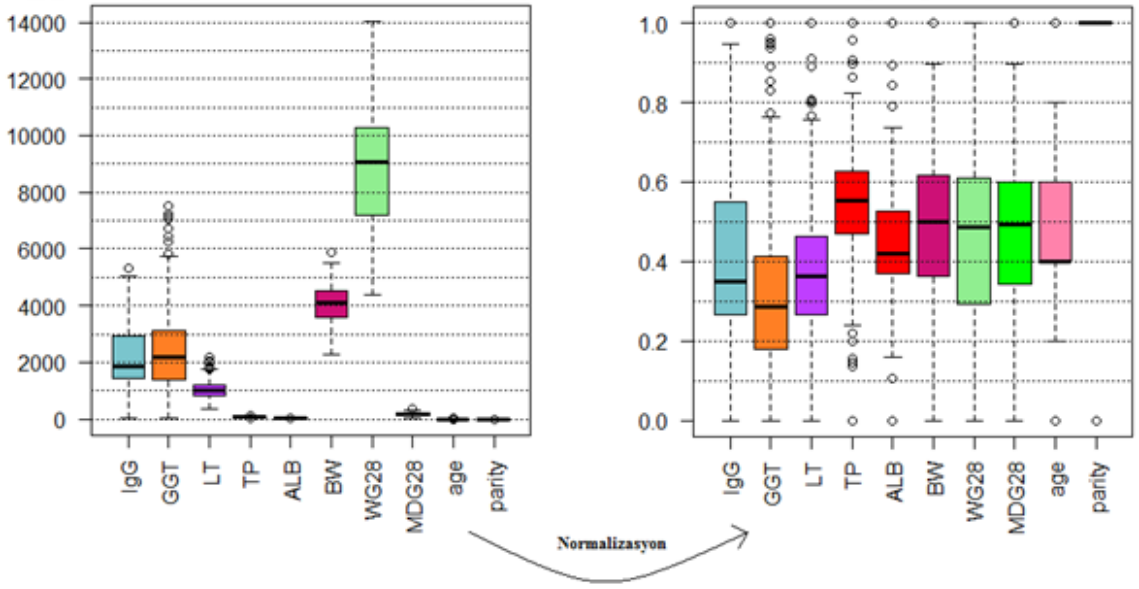
Veri setindeki özelliklerin ilişki matrisi Şekil 5.4'de verilmiştir.



Şekil 5. 4 Özellikler arasındaki ilişki matrisi

İlişki (korelasyon) matrisindeki koyu mavi renk değişkenler arasındaki güçlü pozitif ilişkiyi gösterirken koyu kırmızı renk ise güçlü negatif ilişki olduğunu göstermektedir. Değişkenler arasındaki ilişkinin gücü arttıkça renk koyulaşmaktadır. Korelasyon matrisi incelendiğinde IgG özelliğinin hem GGT hem de TP özelliği ile arasında yüksek korelasyon olduğu görülmektedir.

Çizelge 5.2’de numerik özelliklerin aldıkları değer aralıklarının birbirinden çok farklı olduğu görülmektedir. Örneğin IgG kan seviyesinin en küçük değeri 19 iken en büyük değeri 5302, 28. gün sonundaki vücut ağırlığının en küçük değeri 4364 iken en büyük değeri 14016 veya anne yaşının en küçük değeri 1 iken en büyük değerinin 6 olduğu görülmektedir. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğunda büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır. Özellikler uzaklık ölçüsü alan yöntemlerde normalize edilmiş veri kullanımı daha doğru sonuçlar üretilmesini sağlayacaktır. Şekil 5.5'de normalize edilmemiş veri seti ve minimum-maksimum yöntemi ile normalize edilmiş veri setine ait kutu grafiği verilmiştir.



Şekil 5. 5 Orijinal veri seti ve normalize edilmiş veri seti değer aralıkları

Kutu grafiği ile sürekli değişkenin hem büyüklüğünü, hem de dağılımını incelememiz mümkündür. Şekil 5.5’de solda veri setinin normalize edilmemiş orijinal değerlerine ait kutu grafiği, sağda ise veri setine minimum-maksimum normalizasyon uygulandıktan sonraki kutu grafiği gösterilmiştir. Ham verinin kutu grafiği incelendiğinde IgG, GGT ve WG28 parametrelerinin değer aralıklarının LT, TP, ALB, BW ve MDG28 özelliklerinin aldıkları değer aralığından çok daha geniş olduğu görülmektedir. Normalizasyon uygulandıktan sonra ise bu farkın azaldığı yani özelliklerin aldıkları değerlerin birbirlerine daha yaklaştığı görülmektedir. Ham verideki özelliklerin ortalamalarının birbirinden çok

uzak değerlere sahip oldukları, normalizasyon uygulanmış veri setindeki özelliklerin ortalamalarının birbirine çok yakın değerler olduğu görülmektedir.

Literatürde veri normalleştirme için farklı yöntemler kullanılmaktadır. Veri setine uygulanan normalizasyon yöntemlerinden hangisinin daha başarılı olduğunu tespit edebilmek için veriler k-ortalama yöntemi ile analiz edilmiştir. Kümeleme sonucunda elde edilen etiketler ile elimizdeki kuzuya ait etiketler (sağlıklı / hasta) karşılaştırılarak kümeleme başarısı saflık (purity) ve entropi (entropy) kriterlerine göre ölçülmüştür. Elde edilen sonuçlar Çizelge 5.3'de verilmiştir.

Çizelge 5. 3 Normalizasyon yöntemlerinin saflık ve entropi sonuçları

Normalizasyon Tekniği	Saflık (Purity)	Entropi (Entropy)
Orijinal Veri	0.6455	0.9051
Min-Max Normalizasyonu	0.5504	0.9725
Z-Score Normalizasyonu	0.7061	0.9108
Ondalık Ölçeklendirme	0.5706	0.9895
Sigmoid Normalizasyonu	**0.7349**	**0.8594**

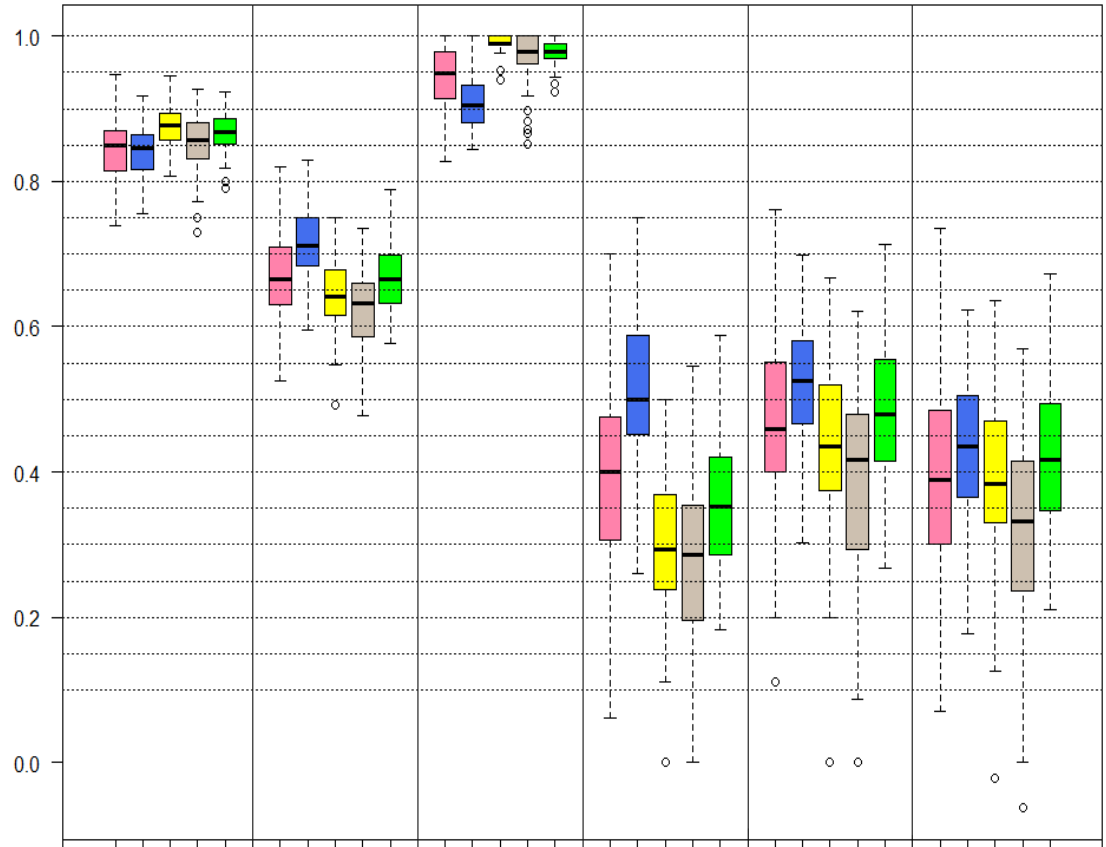
Kümelerin kalitesini değerlendirmede saflık ve entropi kriterleri en sık kullanılan ölçütlerdendir. Entropi düzensizliğin ölçütü olup değeri büyüdükçe kümelerin düzensiz olduğunu göstermektedir. Saflık ise entropi ile ters orantılıdır. Bir kümenin düzensizliği arttıkça saflığı azaldığından analiz sonuçlarında saflık değerinin yüksek entropi değerinin ise düşük olması istenilen bir durumdur. Sigmoid normalizasyon tekniği uygulandıktan sonra veri setinin %73 saflık ve %86 düzensizlik oranıyla diğer normalizasyon tekniklerinden daha başarılı olduğu gözlemlenmiştir. Sigmoid normalizasyon yöntemi verinin dağılımından etkilenmediğinden diğerlerinden daha iyi sonuç ürettiği düşünülmektedir.

5.2 Neonatal Kuzularda Hastalık Sınıflandırması

Sınıflandırıcı model için çalışacak algoritma, yeni doğan bir kuzunun neonatal dönemde biyokimyasal ve fiziksel toplamda 14 özelliğe bakarak kuzunun hasta ya da sağlıklı sınıflarından hangisine ait olduğunu belirleyecektir. Sınıflandırma modellerinde J48, Saf

bayes, k-en yakın komşu, yapay sinir ağları ve rastgele orman algoritmaları sıklıkla tercih edilmektedir. En iyi model başarımını test etmek için bahsedilen sınıflandırıcı model algoritmaları R yazılımı aracılığıyla test edilmiştir.

Neonatal kuzularda hastalık sınıflandırması yaparken veri seti %70 eğitim kümesi, %30 doğrulama kümesi olarak ayrılmıştır. Sınıflandırma modelleri bu eğitim kümesi eğitilmiştir. Her bir farklı eğitim kümesi için 10-kat çapraz geçerlilik uygulandıktan sonra model başarısı daha önce %30 oranında ayrılan doğrulama veri seti ile ölçülmüştür. Veri seti J48, K-EYK, SB, YSA ve RO sınıflandırıcıları ile değerlendirilmiştir. Model başarımını karşılaştırmada doğruluk, dengeli doğruluk, seçicilik, duyarlılık, F-ölçütü, kappa istatistiği ve AUC kriterlerinden yararlanılmıştır. Sınıflandırma yöntemleri 100 kere tekrarlanmış olup, elde edilen sonuçlar kutu grafiğiyle Şekil 5.6'da sunulmuştur. 100 tekrar sonucunda elde edilen ortalama sonuçlar ise kutu grafiğinin altında verilmiştir.



	Doğruluk	Dengeli Doğ.	Seçicilik	Duyarlılık	F-Ölçütü	Kappa
J48	0,8438	0,6698	0,9431	0,3966	0,4692	0,3864
SB	0,8427	0,7132	0,9100	0,5164	0,5226	0,4304
KEYK	0,8745	0,6395	0,9968	0,2821	0,4300	0,3834
YSA	0,8500	0,6210	0,9730	0,2689	0,3671	0,3067
RO	0,8681	0,6647	0,9744	0,3548	0,4707	0,4081

Şekil 5. 6 Sınıflandırma yöntemlerinin performans sonuçları

Birçok çalışmada sadece doğruluk ölçütü değeri göz önüne alınarak en yüksek değere sahip olan model en iyi performansa sahip model olarak nitelendirilir. Oysa sınıflar arası uygun dağılım gösteren dengeli veri setinde doğruluk değeri, dengesiz veri setine göre daha iyi performans göstermektedir. Eşit sınıf dağılımına sahip veri setinde doğruluk değeri ile dengeli doğruluk değeri aynı sonuca sahip olurken, veri setinin dengesiz olması durumunda ise dengeli doğruluk değeri doğruluk değerinden daha düşüktür. Nitekim neonatal kuzularda hastalık sınıflandırması sonuçları incelendiğinde (Şekil 5.6) en yüksek doğruluk değerine göre en iyi performansı KEYK ve RO modelleri gösterirken, dengeli doğruluk kriterine göre bu durum değişerek en iyi performansı SB modelinin gösterdiği görülmektedir.

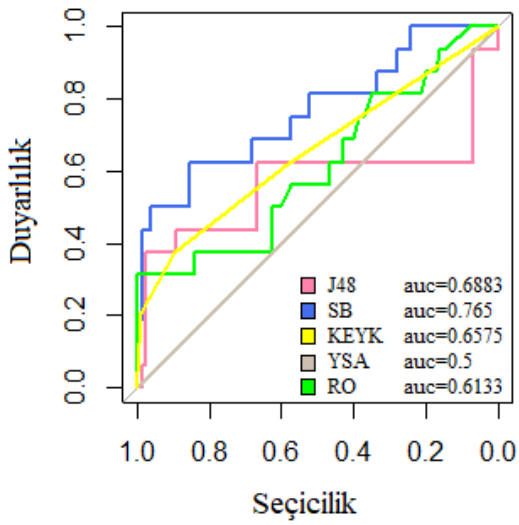
Seçicilik oranı bize gerçekte sağlıklı olan kuzuların ne kadarının sistem tarafından doğru tahmin edildiğini verir. Yaklaşık 0.997 oranı ile sağlıklı kuzuları en başarılı şekilde tespit eden yöntemin KEYK olduğu görülmektedir. Her ne kadar sağlıklı kuzuları doğru tahmin etmek önemli olsa da bizim için asıl önemli olan hastalanabilecek kuzuların tespit edilerek önlemler alınmasıdır. Bu nedenle de duyarlılık ölçüsü önemsenecek ve göz önünde bulundurulacak kıstastır.

Duyarlılık oranı gerçekte hasta olan kuzuların ne kadarının sistem tarafından hasta olarak tahmin edildiğini verir. Yaklaşık 0.5164 duyarlılık oranıyla en yüksek başarıyı SB yöntemi göstermiştir. Veri setindeki 347 kuzudan 60 tanesi hastadır. Kuzuların içinden hastayı tahmin etme ihtimali normalde yaklaşık %17 iken, veri madenciliği yöntemi ile bu oran yaklaşık %52'dir. Hasta olabilecek kuzuların önceden tahmin edilmesi, erken tedavi için fayda sağlayacaktır.

F-ölçütü, kesinlik (precision) ve duyarlılığın ağırlıklandırılmış ortalamasını göstermektedir. Başlı başına kesinlik ve duyarlılığa bakmak yerine F-ölçütüne bakmak daha doğru sonuç vermektedir. 0.5226 F-ölçütü oranıyla en yüksek başarıyı yine SB yöntemi göstermiştir.

Kappa değeri sınıflandırıcı modelin doğru sınıflandırma başarımı hakkında bilgi vererek başarının şans faktörüne bağlı olup olmadığı hakkında fikir vermektedir. SB yöntemi yaklaşık 0.435 kappa değerine sahiptir. Kappa değerinin 0.40'dan büyük olmasının makul olduğu araştırmacılar tarafından bildirilmiştir [86]. SB modelinin 0,435 kappa değeri nedeniyle tutarlı tahminler yaptığını göstermektedir.

Özellikle sağlık alanında yapılan modellerin değerlendirilmesinde sıklıkla kullanılan AUC kriteri sonuçları Şekil 5.7'de gösterilmiştir.



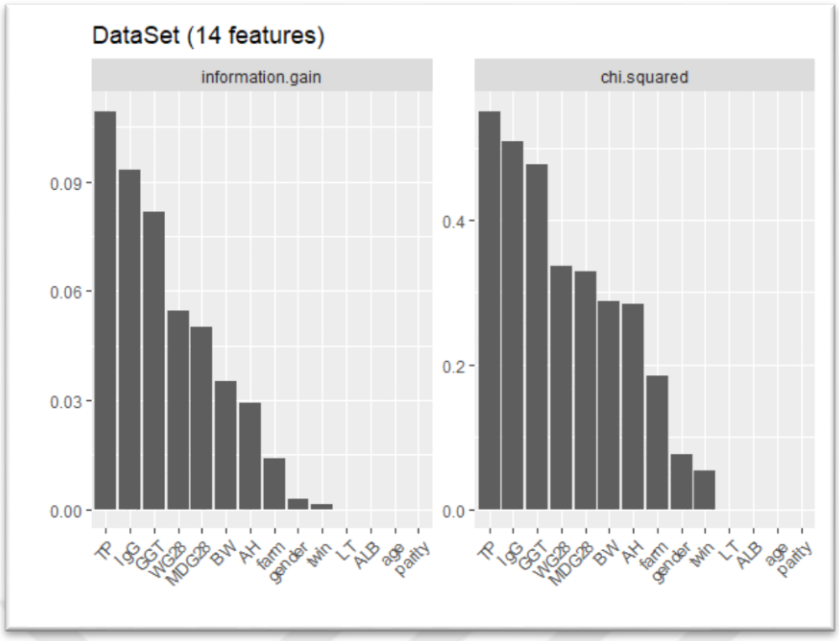
Şekil 5. 7 Sınıflandırma yöntemlerinin AUC sonuçları

Sınıflandırma yöntemlerinin AUC grafiği incelendiğinde, 0.765 oranı ile SB yönteminin diğer yöntemlerden daha başarılı olduğu görülmektedir. Bu da bize SB modelinin verileri sınıflandırmada ayrım gücünün yüksek olduğunu gösterir.

Veterinerlik alanında veri toplama oldukça güç, zaman alıcı ve maliyeti yüksek bir süreçtir. Bu nedenle daha az veri kullanılarak daha başarılı sınıflandırma yapmak zamandan ve emekten kazanç sağlayacaktır. Bu nedenle tez kapsamında bilgi kazancına (information gain) dayanan özellik seçme algoritması kullanılmıştır.

Bilgi kazancına dayanan özellik seçme algoritması veri kümesinde bulunan ilgisiz, gereksiz, fazla veya bilgi kazancı düşük olan özellikleri atmayı amaçlamaktadır. Bu işlem ile sınıfları en iyi şekilde ayırt eden özelliklerin alt kümesi tanımlanır [93], [94]. Böylece ayırt edici gücü yüksek özellikler seçilerek özellik veri kümesinin boyutu azaltılmış olur [95].

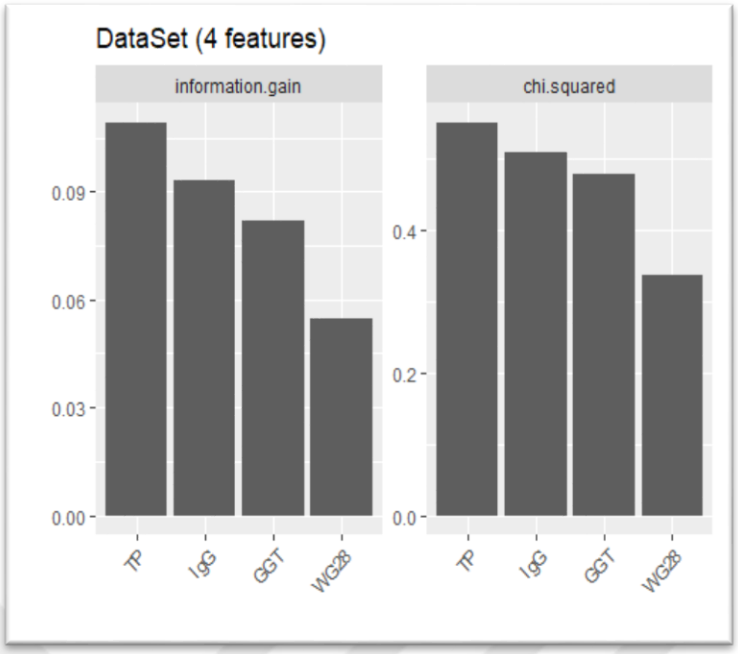
14 özellikten oluşan veri setine bilgi kazancına dayanan özellik seçim algoritması uygulandıktan sonra özelliklerin önem derecesini gösteren grafik Şekil 5.8'de gösterilmiştir. Bunun için R'da 'generateFilterValuesData' fonksiyonunun "information.gain" metodu kullanılmıştır.



Şekil 5. 8 Bilgi kazancı yöntemine göre özelliklerin önem derecesi

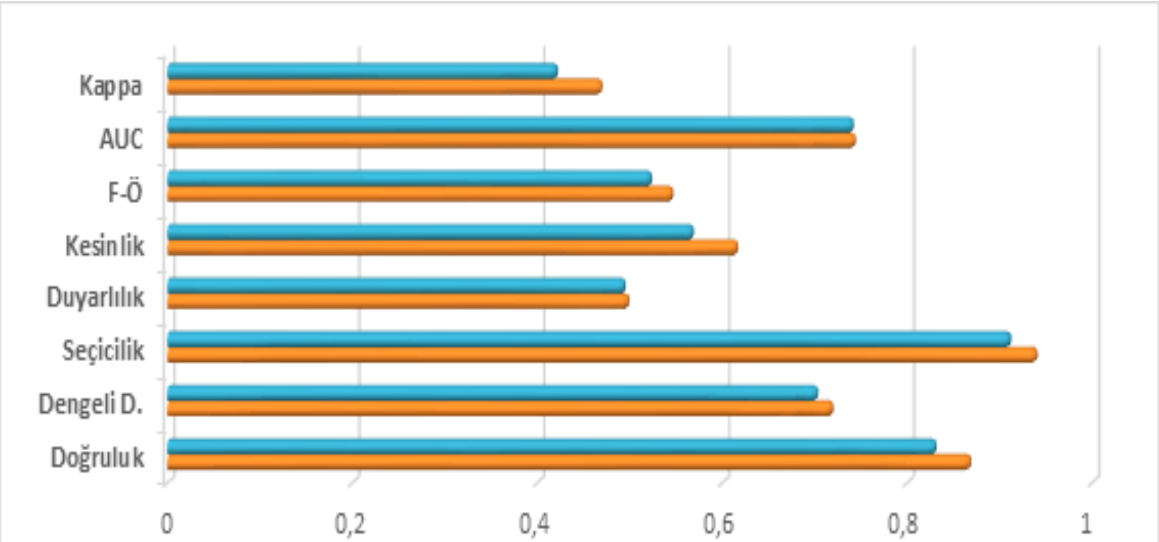
Şekil 5.8'de özelliklerin önem dereceleri sağa doğru azalmakta olup LT, ALB, age ve parity özelliklerinin hastalık durumu (sağlıklı/hasta) üzerinde önemli olmadığı görülmektedir. En önemli özelliğin TP olduğu ve onu aralarında yüksek korelasyon bulunan IgG ve GGT özelliklerinin takip ettiği görülmektedir. Nitekim yapılan çalışmalarda IgG seviyesinin hastalıklarla ilişkili olduğu ve neonatal sağlıklı kuzularda IgG ile büyüme performansı arasında direk bir ilişki olduğu bildirilmiştir [92].

Bilgi kazancı yöntemine göre özelliklerin önem oranları hesaplandıktan sonra, eşik seviyesi olarak önem oranı 0.5'den (%50'den) fazla olan özellikler alınarak diğer özellikler elimine edilmiştir. Önem oranı %50'den fazla olan özellikler Şekil 5.9'da verilmiştir.



Şekil 5. 9 Önem oranı %50'den fazla olan özellikler

Şekil 5.9'da görüldüğü gibi kuzunun sağlıklı veya hasta olması üzerinde en etkili olan özellikler TP, IgG, GGT ve WG28'dir. Özellik seçimi, sınıflandırma başarımı, sınıflandırma zamanı, öğrenme için gereken örneklerin sayısı ve sınıflandırma başarım maliyeti üzerinde önemli etkiye sahiptir [95]. Sınıflandırma başarımı üzerinde özellik seçiminin etkisini gözlemleyebilmek için, 14 özellikten oluşan veri seti ile özellik seçimi yaptıktan sonra elde edilen 4 özellikten oluşan veri seti hastalık durumuna göre SB yöntemi ile 10 kez sınıflandırılmıştır. Elde edilen ortalama sonuçlar Şekil 5.10'da sunulmuştur. 10 tekrar sonucunda elde edilen ortalama sonuçlar ise grafiğinin altında verilmiştir.



	Doğruluk	Dengeli D.	Seçicilik	Duyarlılık	Kesinlik	F-Ö	AUC	Kappa
SB	0,8299	0,7021	0,9106	0,4937	0,5674	0,5223	0,7404	0,4206
ös_SB	0,8673	0,7183	0,9391	0,4975	0,6153	0,5451	0,7426	0,4688

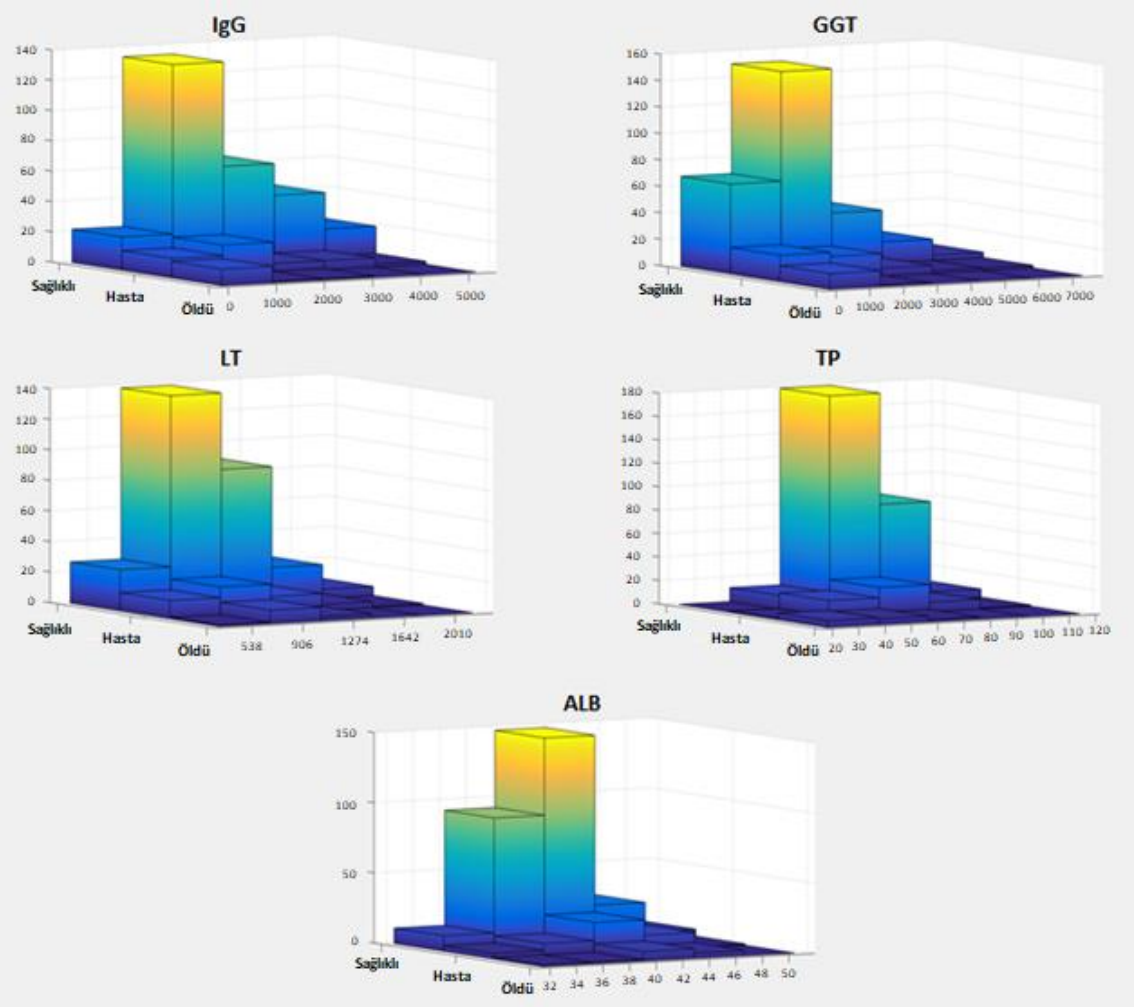
Şekil 5. 10 SB sınıflandırıcı yönteminin özellik seçimi yapılmadan önce ve yapıldıktan sonraki sınıflandırma başarısı

Özellik seçimi sonrası yapılan sınıflandırma sonucunda elde edilen bulgular incelendiğinde tüm model değerlendirme kritelerinde özellik seçimi sonrası (ös_SB) bir artış olduğu görülmektedir. Bu da bize sadece 4 özellik (TP, IgG, GGT ve WG28) kullanılarak yüksek performanslı sınıflandırma yapılabileceğini göstermektedir. Yani neonatal kuzularda hasta ve sağlıklıları ayırmada TP, IgG, GGT ve WG28 özelliklerinin önemli rol oynadığı söylenebilir.

5.3 Kuzu Ölümlerinde Eşik Kan Değeri Belirleme

Kümülatif dağılım fonksiyonu (Cumulative Distribution Function-Cdf) sürekli artan bir fonksiyon olup, X ekseni rastgele değişkenin alacağı değerleri, Y ekseni bu değerleri alma olasılıklarının eklenik toplamını gösterir. Tez çalışmada bu fonksiyonun kullanılmasındaki amaç ölen veya hastalanan kuzuların yüzde kaçının hangi kan değerin altında olduğunu gözlemleyebilmektir. Ayrıca kan değerlerinin hangi hastalık türünde nasıl seyrettiği fikrine sahip olabilmek için de kümülatif dağılım fonksiyonundan yararlanılmıştır.

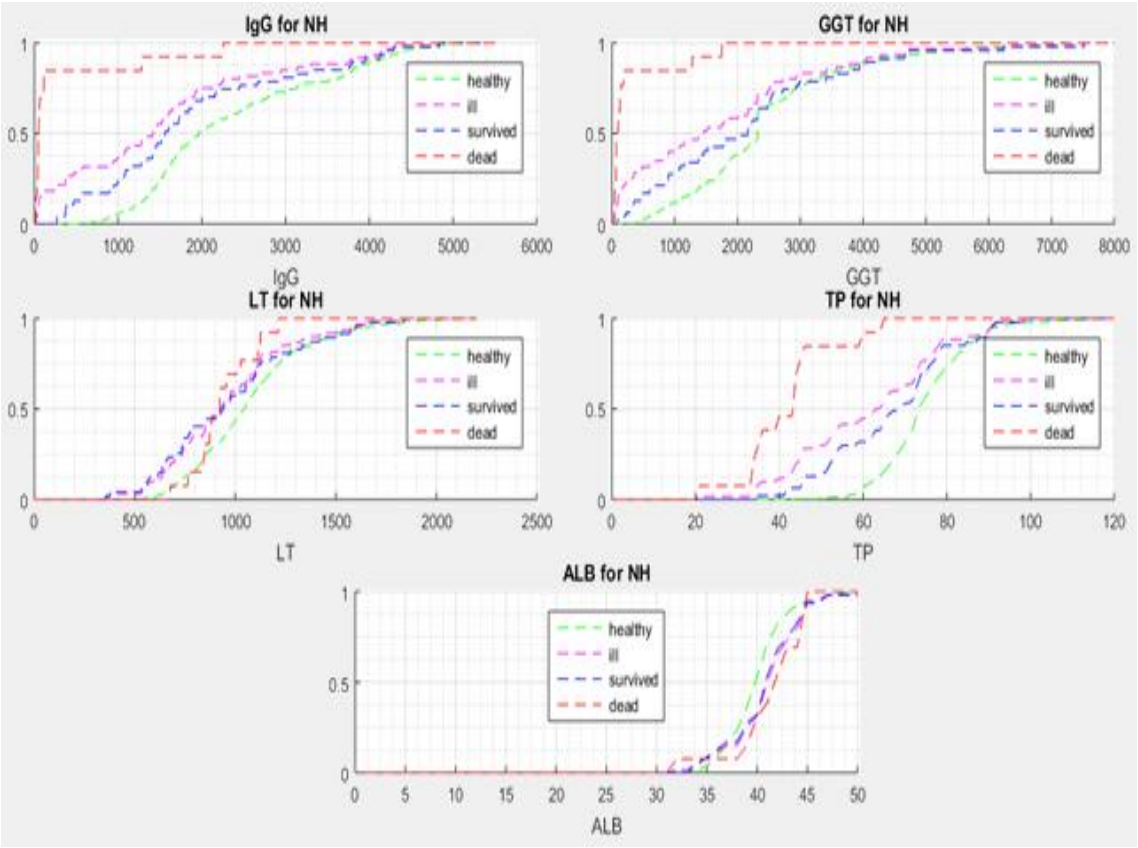
Şekil 5.11’de kan değerlerinin neonatal hastalık sonuçlarına göre (sağlıklı/hasta/öldü) dağılımları görülmektedir. Grafiklerde X ekseni kan değişkenlerinin aldıkları değerleri, Y ekseni ise kuzu sayısını göstermektedir.



Şekil 5. 11 Kan değerlerinin hastalık durumlarına göre dağılım grafiği

Şekil 5.12’de kan seviyeleri niteliklerinin neonatal hastalık durumlarına göre kümülatif dağılımları verilmiştir. Elde edilen kümülatif dağılım fonksiyonu incelendiğinde, neonotal dönemde ölen (kırmızı eğri) kuzuların IgG, GGT ve TP kan değerlerinin sağlıklı (yeşil eğri), hasta (pembe eğri) ve yaşamını sürdüren (mavi eğri) kuzulardan daha düşük değere sahip olduğu ve ölen kuzuların hasta ve sağlıklılardan net bir şekilde ayrıştığı görülmektedir. Hasta, yaşamını sürdüren ve sağlıklı kuzulara ait kan değerleri birbirlerine yakın olup neredeyse tüm durumlarda daha yüksek kan değerlerine sahip olduğu görülmektedir. Bu da ölen kuzularda IgG, GGT, TP ve LT değerinin düşük olduğunu göstermektedir. ALB kan değeri diğer kan değerlerinden daha farklı sonuç

göstermektedir. Genel olarak sağlıklı kuzuların ALB değeri ölü ve hasta kuzuların ALB değerinden daha küçük olduğu görülmektedir.



Şekil 5. 12 Neonatal dönemdeki hastalık sonuçlarına göre kan değerlerinin kümülatif dağılım fonksiyonu

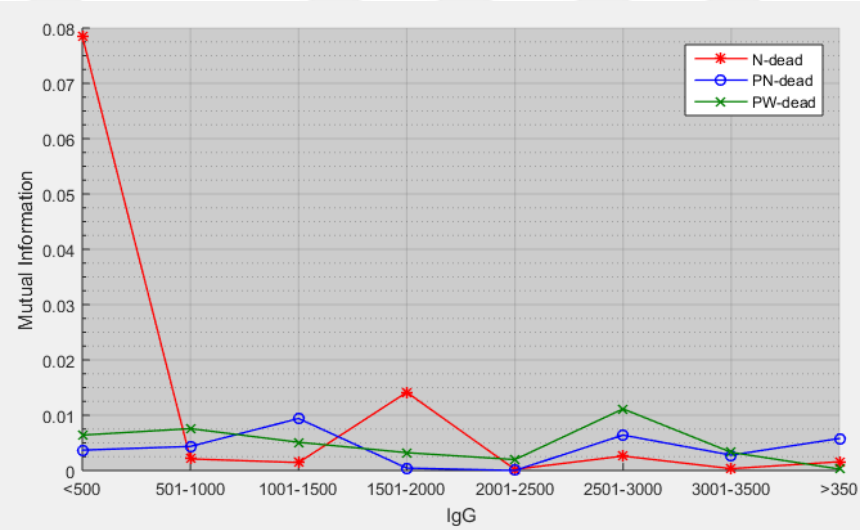
Kan değerleri ayrıklaştırırken Şekil 5.11 ve kümülatif dağılımlardan elde edilen grafik incelenmiş olup gözlemsel olarak yaklaşık değerler üzerinden ayrıklaştırılma yapılmıştır. IgG ve GGT kan değeri <500, 501-1000, 1001-1500, 1501-2000, 2001-2500, 2501-3000, 3001-3005, >3500 şeklinde ayrıklaştırılmıştır. LT kan değeri <400, 401-800, 801-1200, 1201-1600, 1601-2000, >2000 şeklinde ayrıklaştırılmıştır. TP kan değeri <30, 31-40, 41-50, 51-60, 61-70, 71-80, >80 şeklinde ayrıklaştırılmıştır. ALB kan değeri ise <35, 36-40, 41-45, >45 şeklinde ayrıklaştırılmıştır.

Kan değerleri ayrıklaştırıldıktan sonra eşik seviyesi belirlemek için kan değerleri ile ölümler arasındaki ortak bilgi grafikleri incelenmiştir. Ortak bilgi olasılık ve bilgi teorisinde, iki rassal değişkenin paylaştığı bilgiyi ölçer. Yani bu değişkenlerden herhangi birinin bilinmesinin bir diğeri hakkındaki belirsizliği ne kadar düşürdüğünü ölçer. Eğer iki değişken birbirleri hakkında çok az bilgi içeriyorsa, karşılıklı bilgileri 0'a yakındır. Diğer

durumda ise, eğer iki değişken (X ve Y) aynı bilgileri taşıyorlarsa, yani X değerlerini belirlerken Y'den bilgi içeriyorsa bu durumda karşılıklı bilgi 0'dan farklıdır ve değer ne kadar büyük ise iki değişken arasındaki ortak bilgi o kadar fazladır.

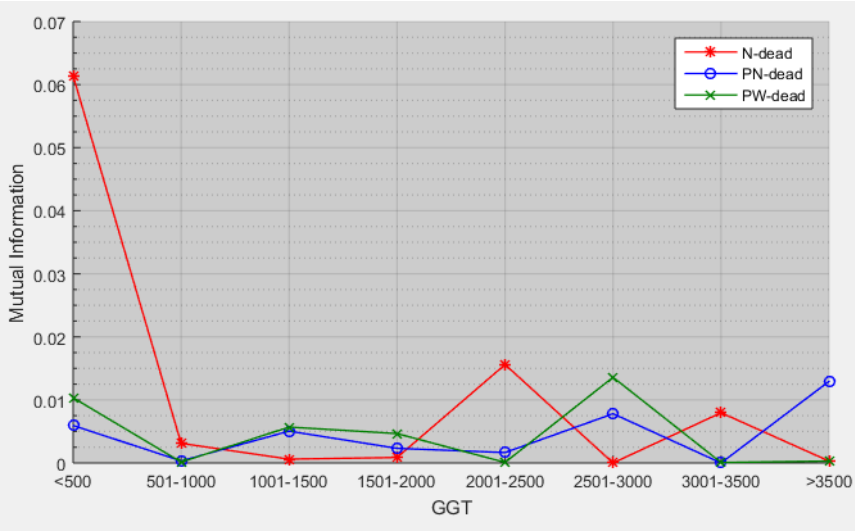
Ölümler ile kan değerleri arasındaki en yüksek ortak bilgi değeri yani grafik üzerinde zirve (peak) yaptığı noktanın aldığı kan değeri eşik değer olarak kabul edilebilir.

Şekil 5. 13'da IgG kan değer aralıklarının ölümler ile arasındaki ilişki grafiği verilmiştir. Grafik incelendiğinde IgG kan değeri 500'ün altında iken neonatal dönemde en fazla ölümlerin olduğu, post-neonatal dönemde hayvanın bağışıklık sistemi artarak ve çevresel faktörlerin ölümler üzerinde daha fazla etkili olmasından dolayı IgG kan değer aralığının 1001-1500'e yükseldiği görülmektedir.



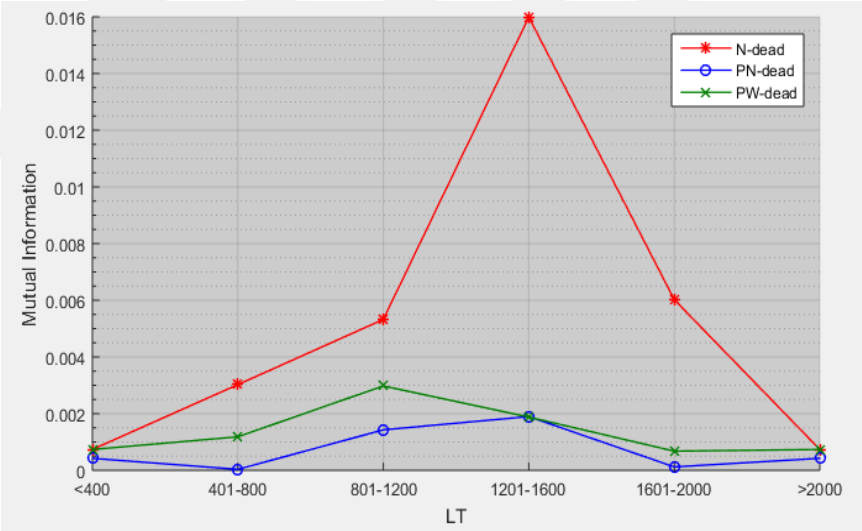
Şekil 5. 13 IgG kan seviyesinin ölümler ile ilişki grafiği

Şekil 5. 13'de GGT kan değer aralıklarının ölümler ile arasındaki ilişki grafiği verilmiştir. Grafik incelendiğinde GGT kan değeri 500'ün altında iken neonatal dönemde en fazla ölümlerin olduğu, görüldüğü GGT kan değeri 3500'den büyük olma durumunda ise post-neonatal dönemde ölümlerin en fazla olduğu görülmektedir.



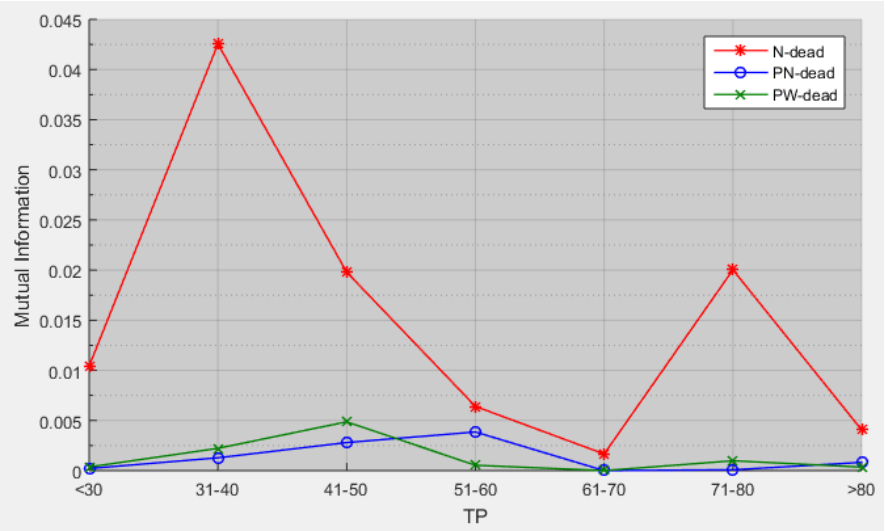
Şekil 5. 14 GGT kan seviyesinin ölümler ile ilişki grafiği

Şekil 5.15'de LT kan değer aralıklarının ölümler ile arasındaki ilişki grafiği verilmiştir. LT kan değeri 1201-1600 aralığında iken hem neonatal hem de post-neonatal dönemde en fazla ölümlerin olduğu görülmektedir.



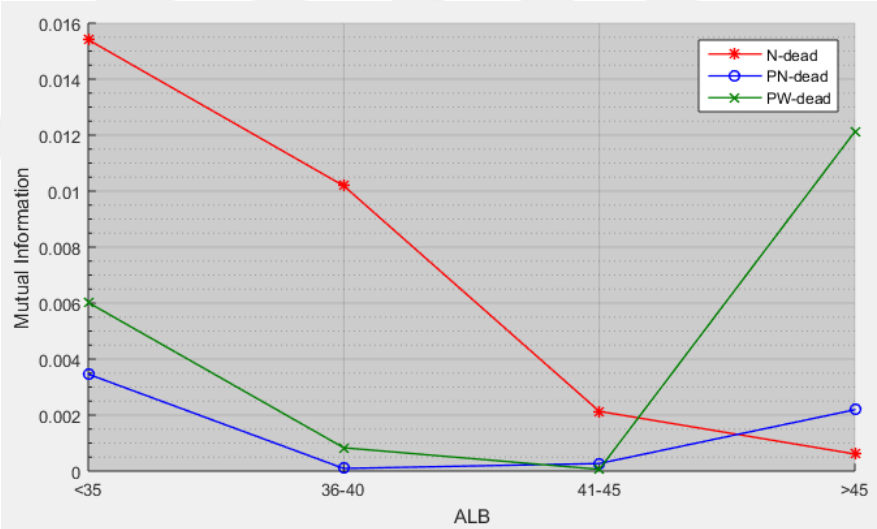
Şekil 5. 15 LT kan seviyesinin ölümler ile ilişki grafiği

Şekil 5.16'da TP kan değer aralıklarının ölümler ile arasındaki ilişki grafiği verilmiştir. Grafik incelendiğinde TP değeri 31-40 aralığında olduğu durumda neonatal dönemde en fazla ölümlerin olduğu, TP değeri 51-60 aralığında olduğu durumda ise post-neonatal dönemde en fazla ölümlerin olduğu görülmektedir.



Şekil 5. 16 TP kan seviyesinin ölümler ile ilişki grafiği

Şekil 5.17'de ALB kan değer aralıklarının ölümler ile arasındaki ilişki grafiği verilmiştir. Grafik incelendiğinde ALB değerinin <35 olduğu durumda en fazla neonatal ve post-neonatal ölümlerin olduğu görülmektedir.

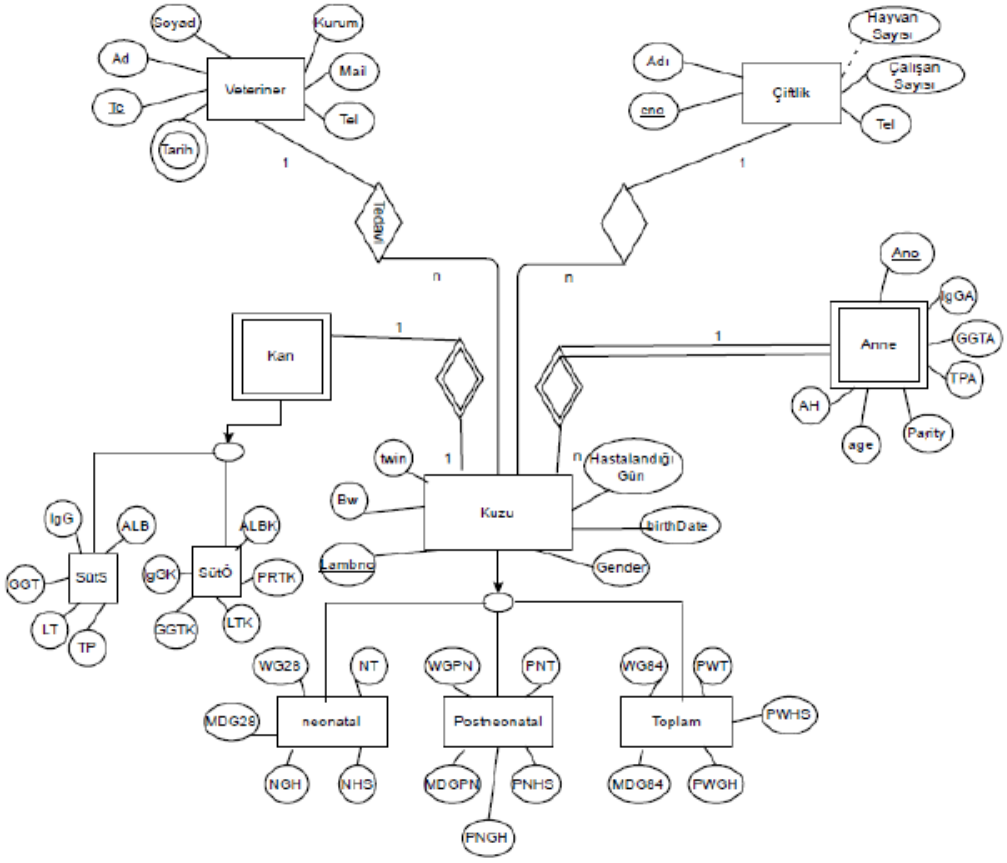


Şekil 5. 17 ALB kan seviyesinin ölümler ile ilişki grafiği

5.4 Masaüstü ve Mobil Uygulama

Hayvan sağlığında tahmin ve analizlerin gerçekleştirilebilmesi için öncelikle veriye erişimin kolay olması gerekmektedir. Günümüzde sağlık kurumlarında kullanılan hasta bilgi sistemleri sayesinde, hasta verileri kolay bir şekilde depolanıp çeşitli analizler yapılabilmektedir. Ancak veterinerlik alanında veriyi elde etmek güç ve elde edilen bu verilerin veri tabanlarında tutulması, paylaşılması henüz mümkün değildir. Buda

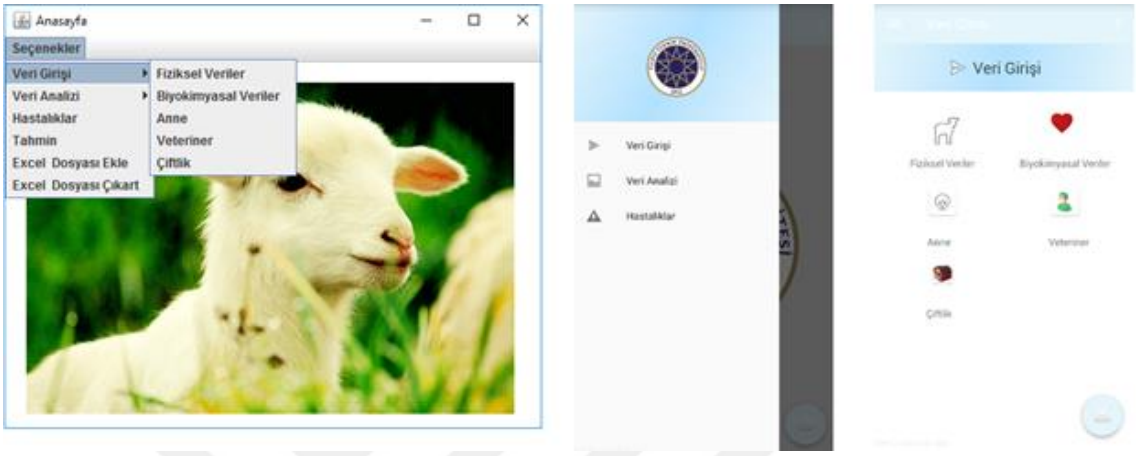
hayvanlar üzerinde yapılacak analizleri güçleştirmektedir. Literatürde veterinerlik alanı ile bilgisayar bilimleri arasındaki disiplinler arası çalışmanın oldukça az olduğu bilinmektedir [3]. Bunun en büyük sebebinin de verilerin depolanıp, paylaşıldığı bir ortamın olmamasından kaynaklandığı düşünülmektedir. Bu nedenle tez çalışmasında veteriner hekimin veriyi ofis, klinik, saha gibi ortamlarda kolay bir şekilde veri tabanında depolayabilmesi için masaüstü ve mobil arayüzler geliştirilmiştir. Hem masaüstü hem mobil arayüzler üzerinden, ortak kullanılan veritabanında erişerek verileri kolay bir şekilde depolayıp çeşitli istatistiksel analizler gerçekleştirebilecektir. Uygulama geliştirilmeden önce nesneler ve aralarındaki ilişkileri, bağlantıları tasvir eden Varlık Bağlantı (Entity Relationship, ER) modeli Şekil 5.18'de verilmiştir.



Şekil 5. 18 ER diyagramı

Şekil 5.19'de masaüstü ve mobil uygulamanın anasayfa ekranı sunulmuştur. Kullanıcı anasayfa ekranında bulunan 'seçenekler' sekmesine tıkladığında sistemde yapılabilecek işlemler görülmektedir. Veri girişi bölümünde kullanıcı; fiziksel, biyokimyasal, anne,

veteriner ve çiftliğe ait veriler ile ilgili işlemleri gerçekleştirebilir. Veri analizi bölümünde veritabanındaki veriler üzerinde bazı istatistiksel analizler yapabilir. Hastalıklar bölümünde kuzuya ait hastalıkları görüntüleyip yeni hastalık girebilir. Excel dosya ekle seçeneğinden sisteme kolayca veri seti yükleyebilir veya sistemden kendi bilgisayarına ya da mobil cihazına veritabanındaki veri setini indirebilir.



Şekil 5. 19 Anasayfa modülü - masaüstü ve mobil uygulama

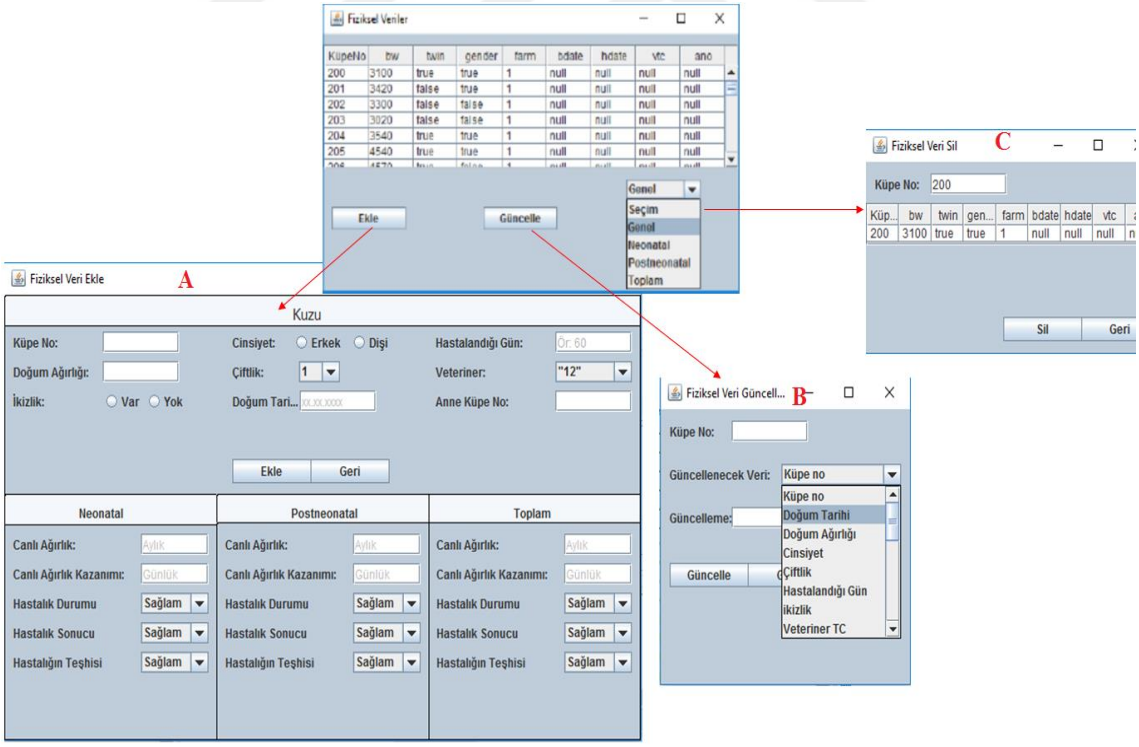
Veritabanındaki kuzulara ait fiziksel veriyi görüntüleme, veritabanına fiziksel veri ekleme, fiziksel veriyi veritabanından silme ve veritabanındaki fiziksel veriyi güncelleme işlemlerinin yapılabildiği masaüstü uygulaması Şekil 5.20, mobil uygulaması ise Şekil 5.21'da gösterilmiştir.

Bu masaüstü ve mobil arayüz yardımı ile kullanıcı veritabanındaki fiziksel veriyi görüntüleyebilir ve buradan veri ekleme, veri güncelleme ve veri silme işlemlerini gerçekleştirebilir. Kullanıcı bu arayüz üzerinden nonatal (0-28. gün), post-neonatal (28-84. gün) veya tamamı (0-84. gün) olacak şekilde istediği döneme ait kuzu bilgisini görüntüleyebilir.

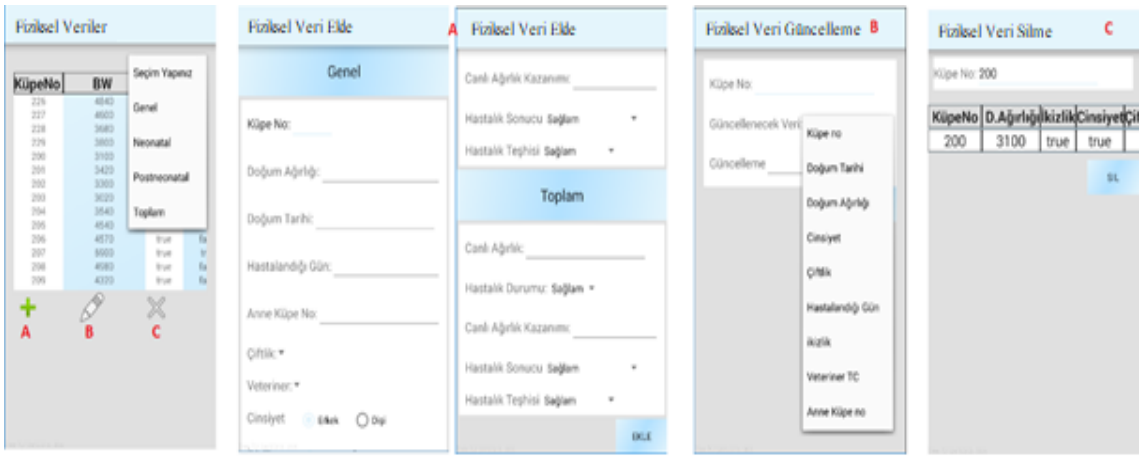
Kullanıcı masaüstü arayüzünde bulunan 'Ekle' ve mobil arayüzünde bulunan ' ' butonuna bastığında açılan Şekil 5.20(A) veya 5.21(A) arayüzleri ile veritabanına fiziksel veri ekleme işlemini gerçekleştirir. Bu arayüzler üzerinden veritabanına kuzunun küpe numarası, annesinin (koyun) küpe numarası, ağırlık bilgileri, muayene eden hekim, cinsiyet, hastalık sonucu (sağlıklı, hasta, ölü), hastalık teşhisi (sağlıklı, ishal, halsizlik, pnömoni, septisemi, pmöoenteritis, diğer) gibi bilgileri veritabanında depolayabilir.

Kullanıcı masaüstü arayüzünde bulunan ‘Güncelle’ ve mobil arayüzünde bulunan ‘ ’ butonuna bastığında açılan Şekil 5.20(B) veya 5.21(B) arayüzleri ile veritabanındaki fiziksel veriyi güncelleme işlemini gerçekleştirir. Kullanıcı kuzuya ait veriyi güncelleyebilmek için öncelikle kuzunun küpe numarasını girmelidir. Daha sonra listeden, güncellenecek özelliği seçerek yeni değerini girip güncelleme işlemini gerçekleştirir.

Kullanıcı masaüstü arayüzünde bulunan ‘Sil’ ve mobil arayüzünde bulunan ‘ ’ butonuna bastığında açılan Şekil 5.20(C) veya 5.21(C) arayüzleri ile veritabanında bulunan fiziksel veriyi silme işlemini gerçekleştirir. Veri silme işlemi kuzunun küpe numarası üzerinden gerçekleştirilmektedir. Kuzunun küpe numarası girildikten sonra yanlış verinin silinmesini önlemek amacıyla, o küpe numarasına ait bilgiler kullanıcıya gösterilmektedir. Eğer silinmesi gereken veri doğru ise sil butonuna basarak veriyi veritabanından siler, eğer başka bir kuzuya ait veri silinecek ise yeni küpe numarası girilmelidir.



Şekil 5. 20 Fiziksel veri (A) ekleme, (B) güncelleme, (C) silme modülleri- masaüstü uygulama



Şekil 5. 21 Fiziksel veri (A) ekleme, (B) güncelleme, (C) silme modülleri - mobil uygulama

Veritabanındaki kuzulara ait biyokimyasal veriyi görüntüleme, veritabanına biyokimyasal veri ekleme, biyokimyasal veriyi veritabanından silme ve veritabanındaki biyokimyasal veriyi güncelleme işlemlerinin yapılabildiği masaüstü uygulaması Şekil 5.22, mobil uygulaması ise Şekil 5.23'de gösterilmiştir.

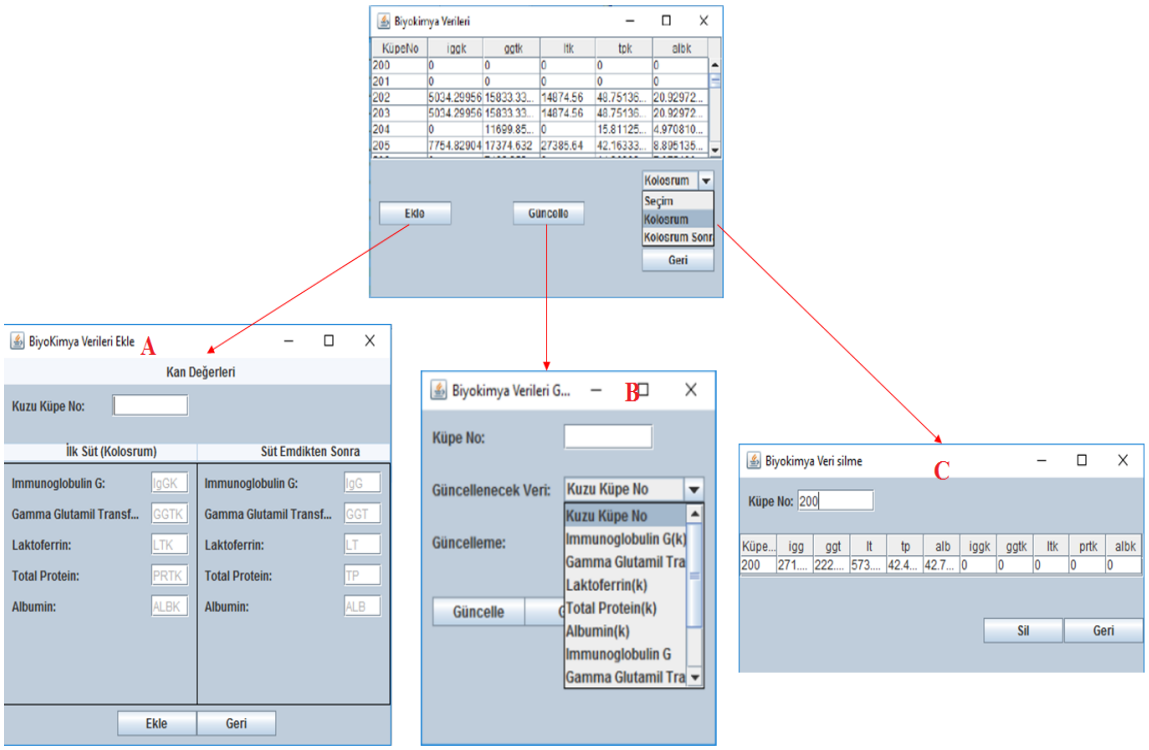
Kullanıcı kuzunun süt emmeden alınan kan seviyeleri üzerinden işlem yapmak isterse listeden kolostrumu, eğer süt emdikten sonra alınan kan seviyeleri üzerinden işlem yapmak isterse kolostrum sonrasını seçer.

Kullanıcı masaüstü arayüzünde bulunan 'Ekle' ve mobil arayüzünde bulunan ' + ' butonuna bastığında açılan Şekil 5.22(A) veya 5.23(A) arayüzleri ile veritabanına biyokimyasal veri ekleme işlemini gerçekleştirir. Bu arayüzler üzerinden veritabanına kuzunun küpe numarası girildikten sonra kuzuya ait IgG, GGT, LT, TP ve ALB kan seviyeleri veya IgGK, GGTK, LT, PRTK ve ALBK kan seviyeleri eklenir.

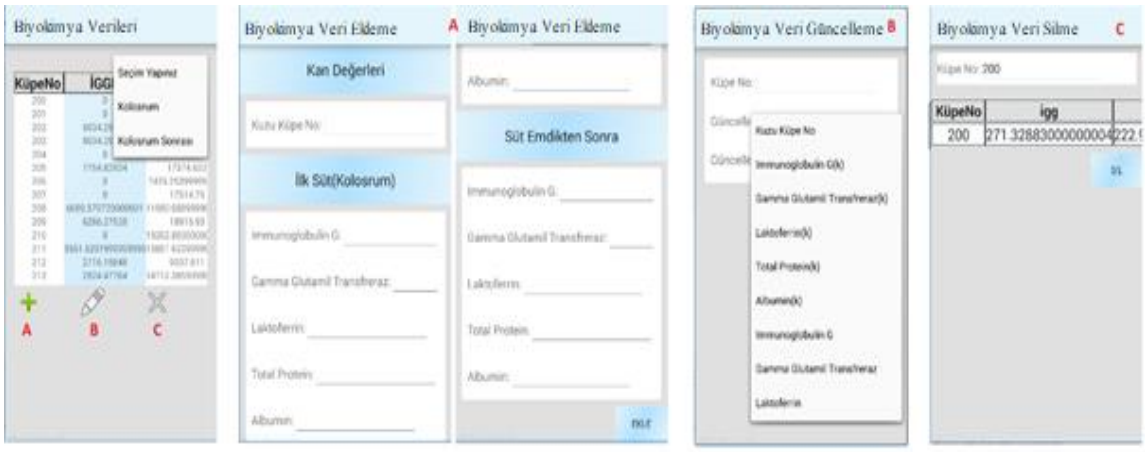
Kullanıcı masaüstü arayüzünde bulunan 'Güncelle' ve mobil arayüzünde bulunan ' ' butonuna bastığında açılan Şekil 5.22(B) veya 5.23(B) arayüzleri ile veritabanındaki biyokimyasal veriyi güncelleme işlemini gerçekleştirir. Kullanıcı kuzuya ait biyokimyasal veriyi güncelleyebilmek için öncelikle kuzunun küpe numarasını girmelidir. Daha sonra listeden, güncellenecek kan seviyesini seçerek yeni değerini girip güncelleme işlemini gerçekleştirir.

Kullanıcı masaüstü arayüzünde bulunan ‘Sil’ ve mobil arayüzünde bulunan ‘ ’ butonuna bastığında açılan Şekil 5.22(C) veya 5.23(C) arayüzleri ile veritabanında bulunan biyokimyasal veriyi silme işlemini gerçekleştirir.

Veri silme işlemi kuzunun küpe numarası üzerinden gerçekleştirilmektedir. Kuzunun küpe numarası girildikten sonra yanlış verinin silinmesini önlemek amacıyla, o küpe numarasına ait bilgiler kullanıcıya gösterilmektedir. Eğer silinmesi gereken veri doğru ise sil butonuna basarak veriyi veritabanından siler, eğer başka bir kuzuya ait veri silinecek ise yeni küpe numarası girilmelidir.



Şekil 5. 22 Biyokimyasal veri (A) ekleme, (B) güncelleme, (C) silme modülleri - masaüstü uygulama



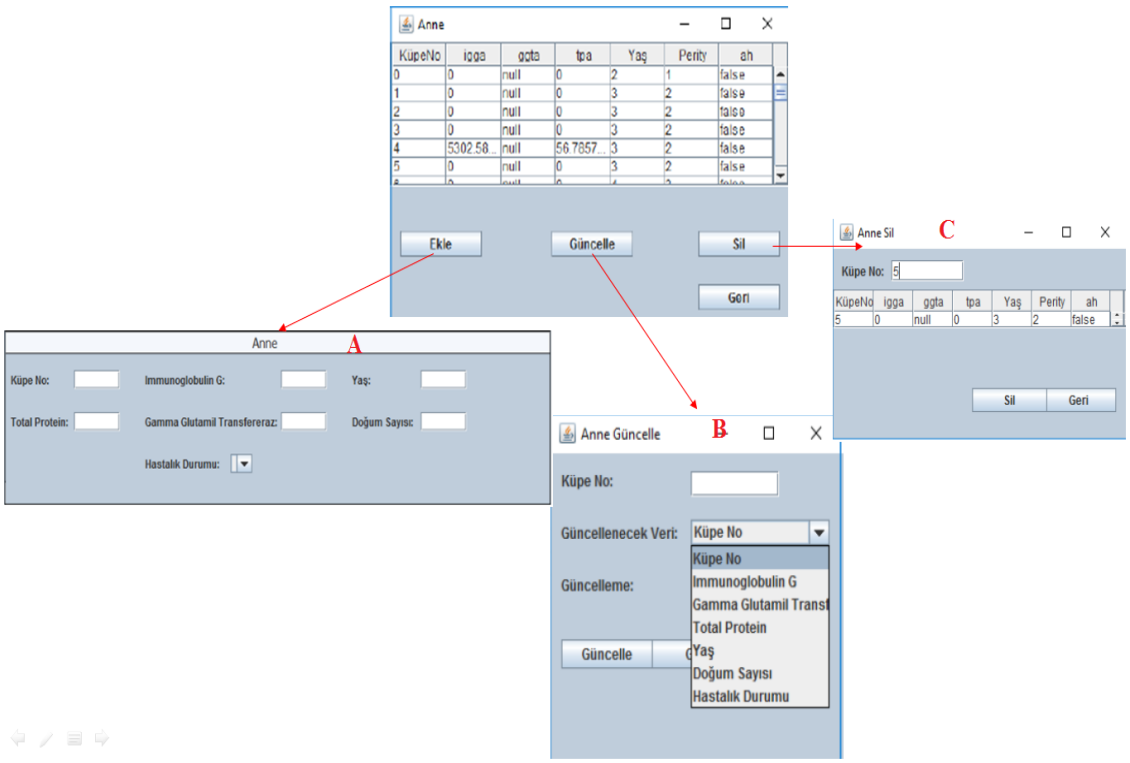
Şekil 5. 23 Biyokimyasal veri (A) ekleme, (B) güncelleme, (C) silme modülleri - mobil uygulama

Veritabanındaki anneye ait bilgileri görüntüleme, ekleme, güncelleme ve silme işlemlerinin yapılabildiği masaüstü uygulaması Şekil 5.24, mobil uygulaması ise Şekil 5.25'da gösterilmiştir.

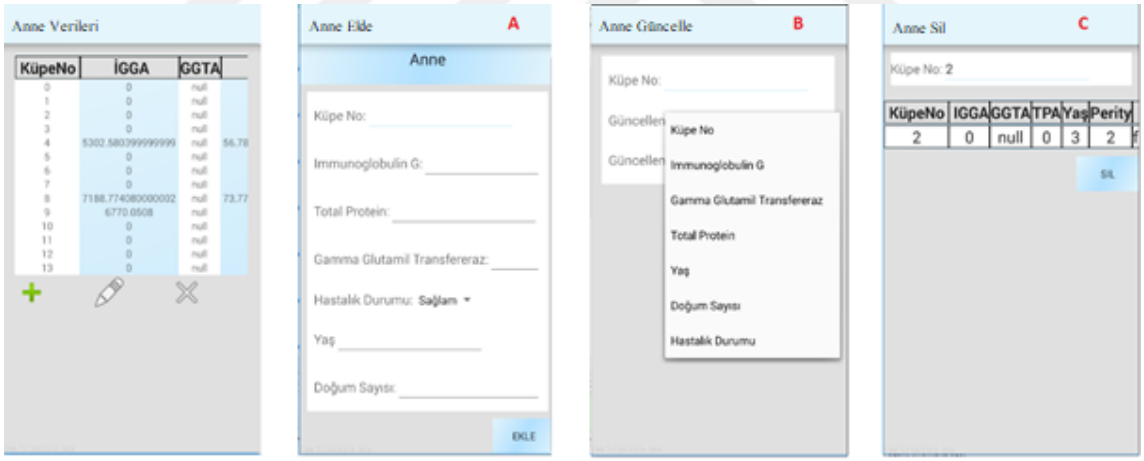
Kullanıcı masaüstü arayüzünde bulunan 'Ekle' ve mobil arayüzünde bulunan ' + ' butonuna bastığında açılan Şekil 5.24(A) veya 5.25(A) arayüzleri ile veritabanında anneye ait bilgileri depolayabilir. Burada anneye ait küpe numarası, doğumdan önce alınan kan seviyeleri (IgGA, GGTA, TPA), yaşı, doğum sayısı ve hastalık durumu (sağlıklı/hasta) bilgisi veritabanına eklenir.

Kullanıcı masaüstü arayüzünde bulunan 'Güncelle' ve mobil arayüzünde bulunan ' ' butonuna bastığında açılan Şekil 5.24(B) veya 5.25(B) arayüzleri ile veritabanında bulunan anneye ait bilgileri güncelleme işlemini gerçekleştirir. Kullanıcı anneye ait bilgileri güncelleyebilmek için öncelikle küpe numarasını girmelidir. Daha sonra listeden, güncellenecek bilgiyi seçerek yeni değerini girip güncelleme işlemini gerçekleştirir.

Kullanıcı masaüstü arayüzünde bulunan 'Sil' ve mobil arayüzünde bulunan ' ' butonuna bastığında açılan Şekil 5.24(C) veya 5.25(C) arayüzleri ile veritabanında bulunan biyokimyasal veriyi silme işlemini gerçekleştirir. Veri silme işlemi annenin küpe numarası üzerinden gerçekleştirilmektedir.



Şekil 5. 24 Anne verisi (A) ekleme, (B) güncelleme, (C) silme modülleri - masaüstü uygulama



Şekil 5. 25 Anne verisi (A) ekleme, (B) güncelleme, (C) silme modülleri - mobil uygulama

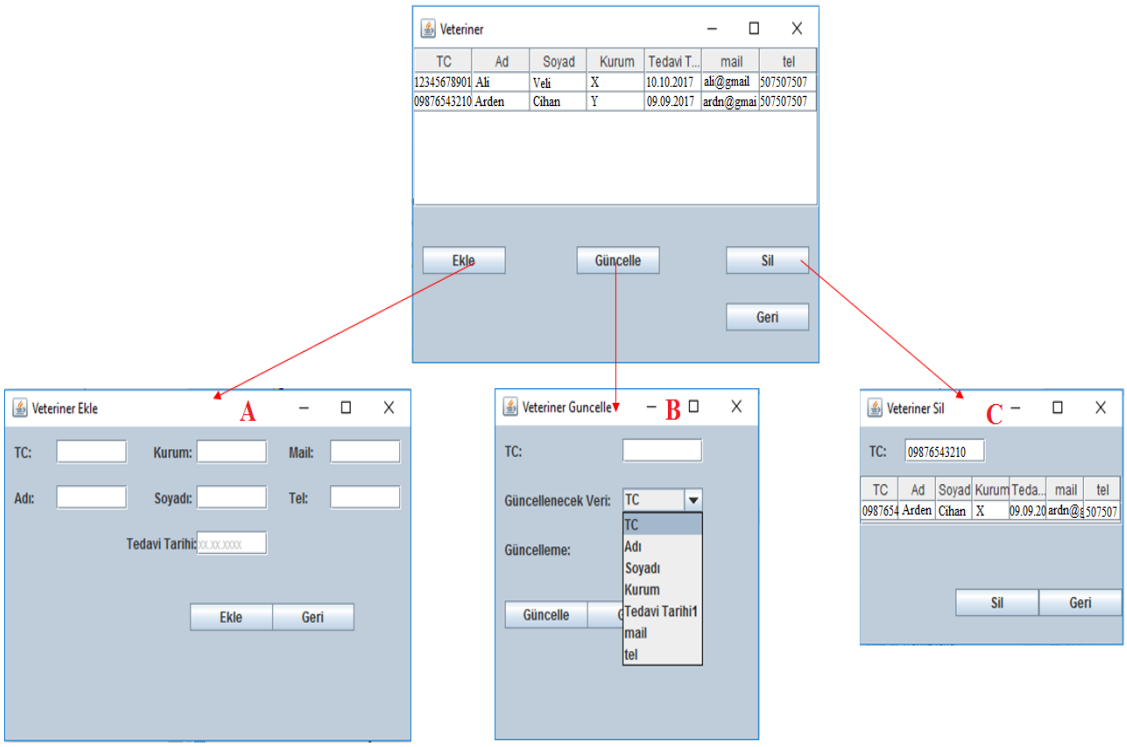
Veterinere ait bilgileri görüntüleme, ekleme, güncelleme ve silme işlemlerinin yapılabildiği masaüstü uygulaması Şekil 5.26, mobil uygulaması ise Şekil 5.27'de gösterilmiştir.

Kullanıcı masaüstü arayüzünde bulunan 'Ekle' veya mobil arayüzünde bulunan ' + ' butonuna bastığında açılan Şekil 5.26(A) veya 5.27(A) arayüzleri ile veritabanında

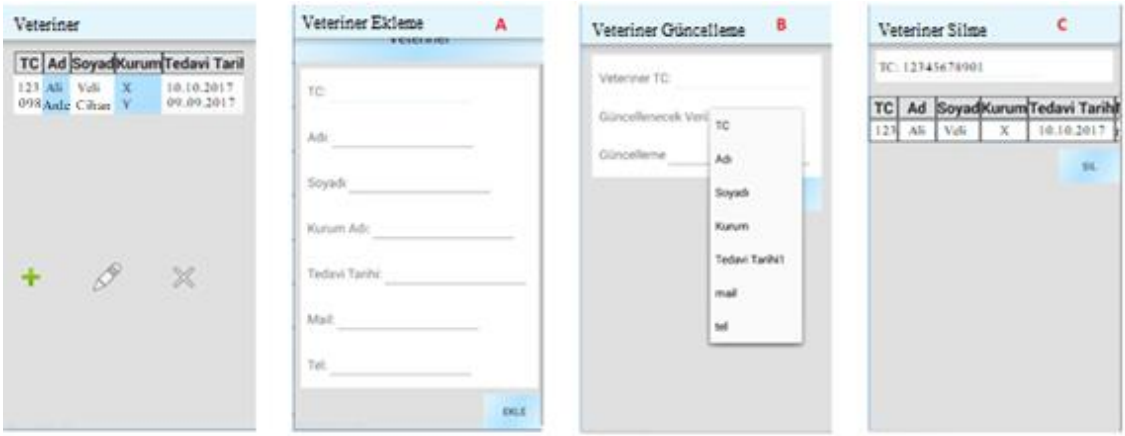
veterinere ait bilgileri ekler. Burada veteriner hekimin TC kimlik numarası, adı, soyadı, çalıştığı kurum, mail, telefon ve tedaviyi yaptığı tarih bilgisi veritabanına eklenir.

Masaüstü arayüzünde bulunan 'Güncelle' veya mobil arayüzünde bulunan ' ' butonuna basıldığında açılan Şekil 5.26(B) ve 5.27(B) arayüzleri ile veritabanında bulunan veteriner hekime ait bilgilerin güncelleme işlemi gerçekleştirilir. Sayfadaki listeden güncellenecek veri seçilip, yeni değeri girildikten sonra güncelleme işlemi gerçekleştirir.

Masaüstü arayüzünde bulunan 'Sil' veya mobil arayüzünde bulunan ' ' butonuna basıldığında açılan Şekil 5.26(C) ve 5.27(C) arayüzleri ile veritabanında bulunan veteriner hekim bilgisi silme işlemini gerçekleştirir. Veri silme işlemi için veterinerin TC kimlik girildikten sonra silme işlemi gerçekleştirilir.



Şekil 5. 26 Veteriner (A) ekleme, (B) güncelleme, (C) silme modülleri - masaüstü uygulama



Şekil 5. 27 Veteriner (A) ekleme, (B) güncelleme, (C) silme modülleri - mobil uygulama

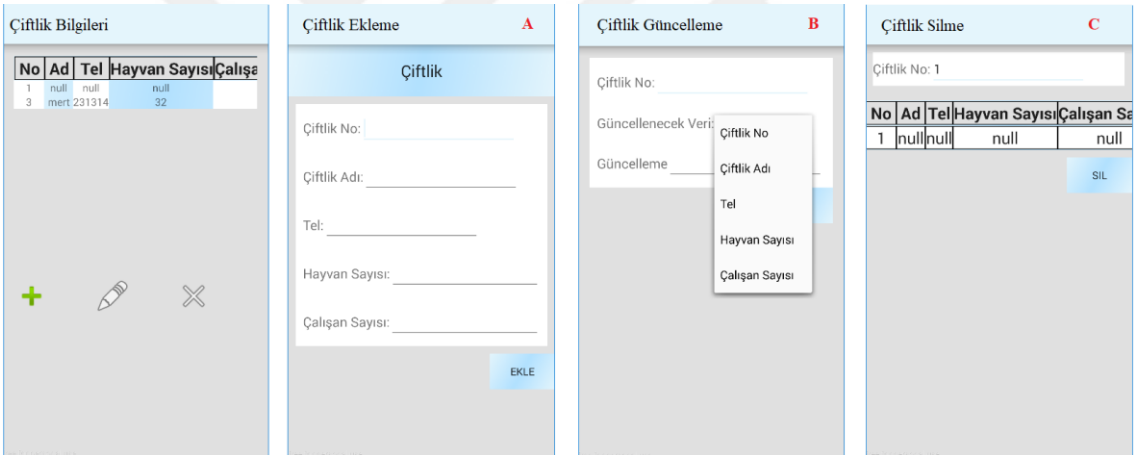
Çiftliğe ait bilgilerin görüntülendiği, ekleme, silme ve güncelleme işlemlerinin yapıldığı masaüstü uygulaması Şekil 5.28, mobil uygulaması Şekil 5.29'de gösterilmiştir.

Kullanıcı masaüstü arayüzünde bulunan 'Ekle' veya mobil arayüzünde bulunan ' + ' butonuna bastığında açılan Şekil 5.28(A) veya 5.29(A) arayüzleri ile veritabanında çiftliğe ait bilgileri ekler. Burada çiftliğe bir numara girildikten sonra çiftlik adı, çiftlikteki hayvan sayısı, çiftlikteki çalışan sayısı ve çitlik telefon numarası veritabanına eklenir.

Masaüstü arayüzünde bulunan 'Güncelle' veya mobil arayüzünde bulunan ' ' butonuna basıldığında açılan Şekil 5.28(B) ve 5.29(B) arayüzleri ile veritabanındaki çitlik bilgileri güncellenebilir. Sayfadaki listeden güncellenecek veri seçilip, yeni değeri girildikten sonra güncelleme işlemi gerçekleştirir.

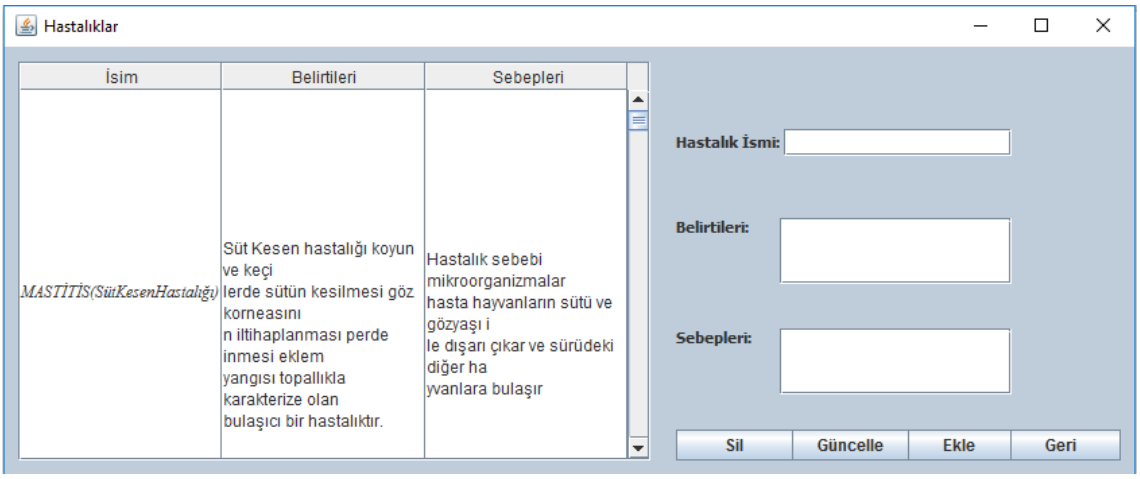
Masaüstü arayüzünde bulunan 'Sil' veya mobil arayüzünde bulunan ' ' butonuna basıldığında açılan Şekil 5.28(C) ve 5.29(C) arayüzleri ile veritabanında bulunan çiftlik bilgisi silinebilir. Veri silme işlemi için çiftliğe ait atanmış numara girildikten sonra silme işlemi gerçekleştirilir.

Şekil 5. 28 Çiftlik (A) ekleme, (B) güncelleme, (C) silme modülleri - masaüstü uygulama



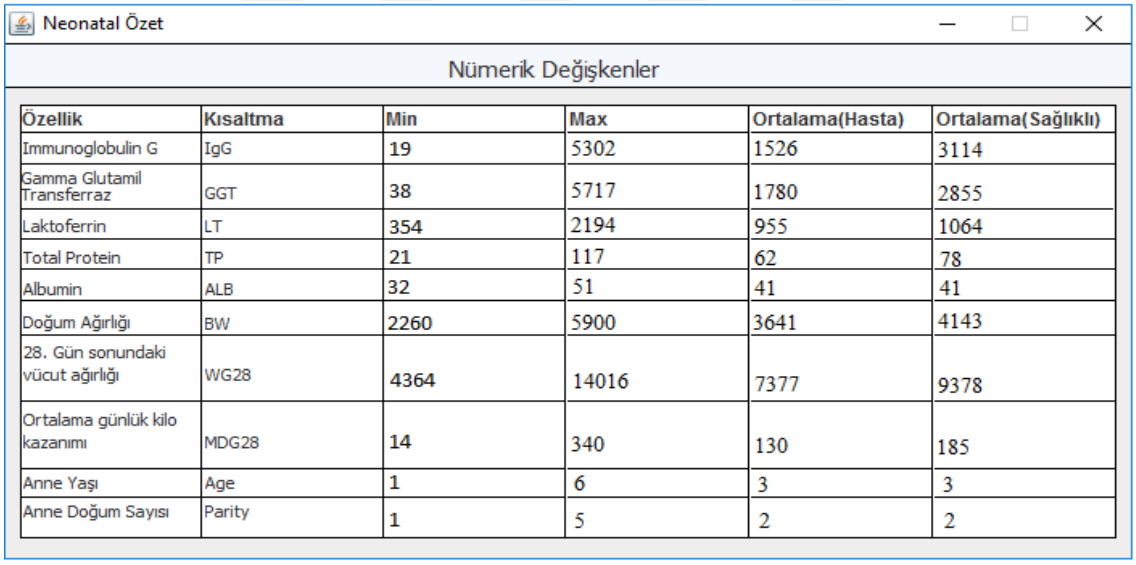
Şekil 5. 29 Çiftlik (A) ekleme, (B) güncelleme, (C) silme modülleri - mobil uygulama

Veritabanındaki hastalıkları görüntüleme, hastalık ekleme, güncelleme ve silme işlemlerinin gerçekleştirildiği arayüz Şekil 5.30'da verilmiştir. Bu arayüzde hastalık ismi girildiğinde eğer veritabanında öyle bir hastalık varsa sol tarafta hastalığın adı, belirtileri ve sebepleri bilgisi görüntülenir. Aynı zamanda bu hastalığın bilgileri güncellenip, silinebilir. Eğer girilen isimde bir hastalık yoksa hastalığın belirtileri ve sebepleri girildikten sonra bu hastalık veritabanına eklenebilir.



Şekil 5. 30 Kuzu hastalıkları ekleme, silme ve güncelleme modülü

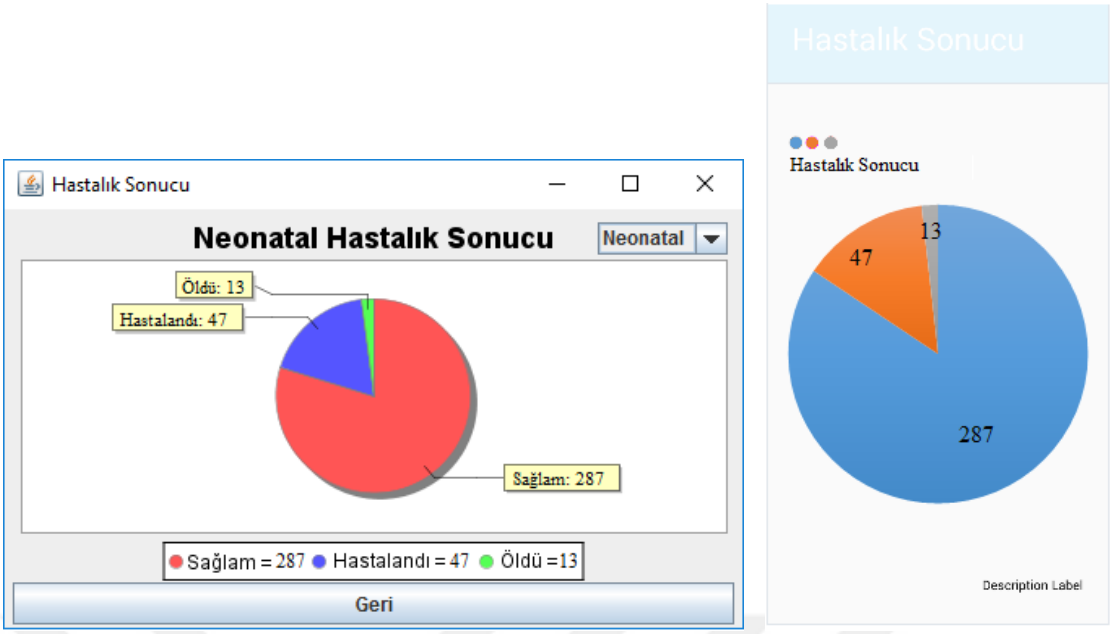
Anasayfa arayüzünde bulunan özet seçeneği tıklandığında, kuzuya ait nümerik nitelikler, niteliğin kısaltması, niteliğin minimum ve maksimum değeri, hasta kuzularda ilgili niteliğin ortalama değeri ve sağlıklı kuzularda niteliğin ortalama değeri görüntülenebilir (Şekil 5.31).



Özellik	Kısaltma	Min	Max	Ortalama(Hasta)	Ortalama(Sağlıklı)
Immunoglobulin G	IgG	19	5302	1526	3114
Gamma Glutamil Transferraz	GGT	38	5717	1780	2855
Laktoferrin	LT	354	2194	955	1064
Total Protein	TP	21	117	62	78
Albumin	ALB	32	51	41	41
Doğum Ağırlığı	BW	2260	5900	3641	4143
28. Gün sonundaki vücut ağırlığı	WG28	4364	14016	7377	9378
Ortalama günlük kilo kazanımı	MDG28	14	340	130	185
Anne Yaşı	Age	1	6	3	3
Anne Doğum Sayısı	Parity	1	5	2	2

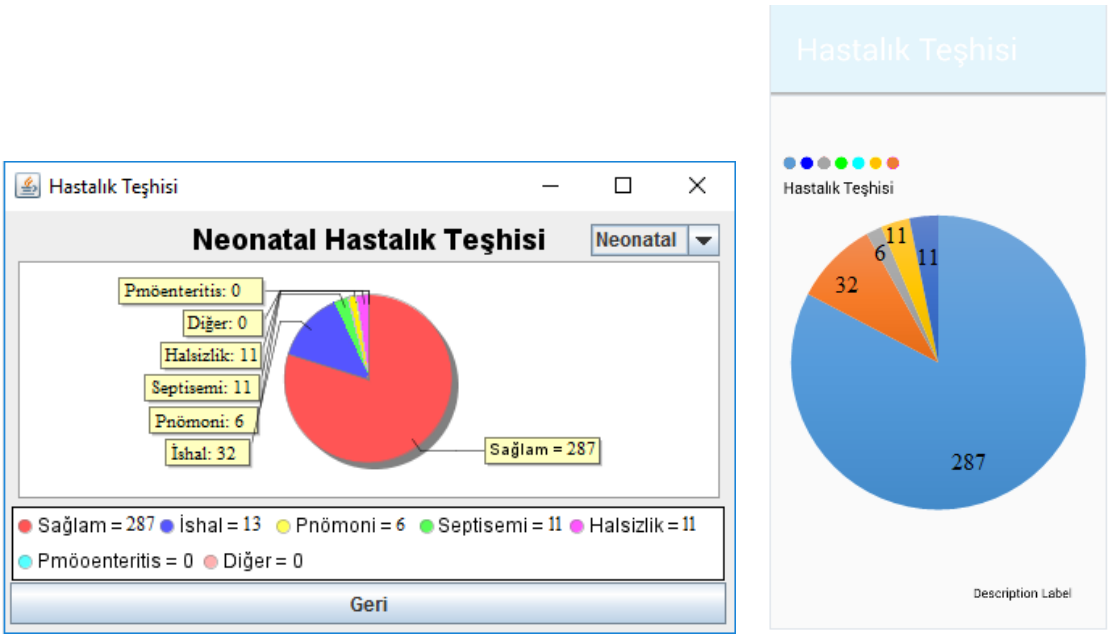
Şekil 5. 31 Neonatal kuzulara ait niteliklerin istatistiksel bilgileri

Kuzuların neonatal dönemdeki sağlık durumlarına ait bilgilerin analizleri masaüstü ve mobil uygulamasında Şekil 5.32'de pasta grafiği üzerinde gösterilmiştir. Grafik incelendiğinde neonatal dönemde 347 kuzudan 287'si sağlıklı, 47'si hasta iken 13'ü neonatal dönemde yaşamını yitirmiştir.



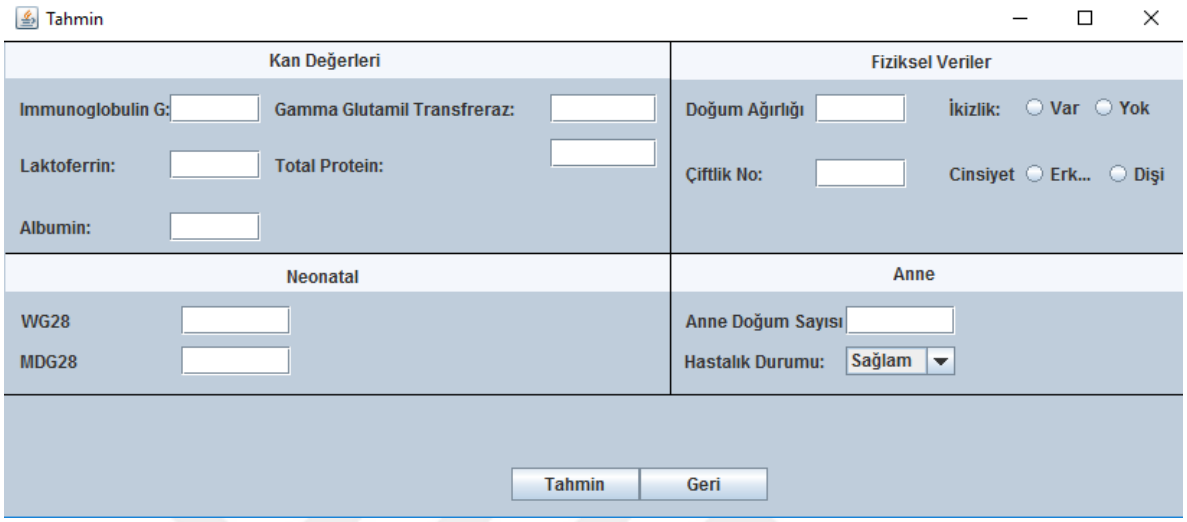
Şekil 5. 32 Masaüstü ve mobil uygulamada neonatal dönemde kuzuların hastalık durumu grafiği

Kuzuların neonatal dönemde yakalandıkları hastalıklara ait bilgilerin analizleri masaüstü ve mobil uygulamasında Şekil 5.33'de pasta grafiği üzerinde gösterilmiştir. Grafik incelendiğinde neonatal dönemde 287 kuzunun sağlıklı, 60 hasta kuzunun da 32'si ishal, 6'sı pnömoni, 11'i halsizlik ve 11'inin septisemi hastalığına yakalandığı görülmektedir.

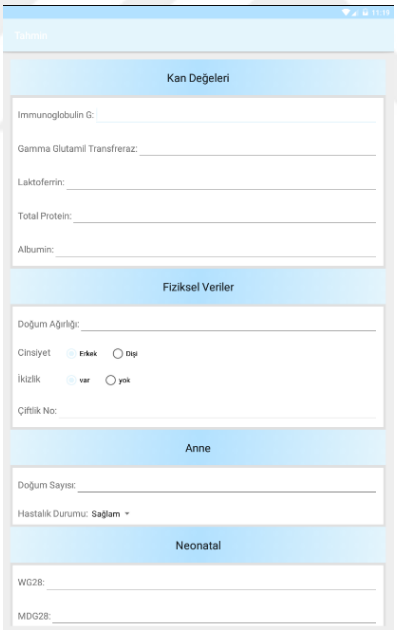


Şekil 5. 33 Masaüstü ve mobil uygulamada neonatal dönemde kuzuların hastalık teşhis grafiği

Kuzunun sağlık durumunu tahmin etmede kullanılan masaüstü uygulama Şekil 5.34'de, mobil uygulama ise Şekil 5.35'de verilmiştir. Uygulamalarda kuzuya ait kan seviyeleri fiziksel veriler, kilo kazanım bilgileri ve anneye ait bilgiler girildikten sonra tahmin butonuna basarak kuzunun hastalanma oranı kullanıcıya sunulmaktadır.



Şekil 5. 34 Kuzunun sağlık durumu tahmini - masaüstü uygulama



Şekil 5. 35 Kuzunun sağlık durumu tahmini - mobil uygulama

BÖLÜM 6

SONUÇ VE ÖNERİLER

Bu tez kapsamında, literatürdeki çalışmalardan farklı olarak veri madenciliği yöntemleri kullanılarak kuzularda teşhis, prognoz ve risk faktörleri tespit edilmeye çalışılmıştır. Veterinerlik alanındaki çalışmalarda, hayvanlardan elde edilen veri setleri genellikle istatistiksel yöntemlerle analiz edilmektedir. Hayvanlarda bilgisayar destekli tanı ile hastalıkların önceden tahmin edilmesi noktasında istatistiksel analizler yetersiz kalmakta olup, veri madenciliği yöntemlerine ihtiyaç duyulmaktadır. Hayvanlarda bilgisayar destekli tanı, erken teşhis ve tedavi konusunda veteriner hekimlere yardımcı olarak bilime ve Türkiye ekonomisine katkı sağlayacaktır.

Tez çalışmasında, 2009 yılında Kars ilinde bulunan iki koyun çiftliğinden toplanmış 301 Akkaraman melezi koyun ve bunlardan doğan 347 kuzuya ait toplam 42 özellikten oluşan veri seti kullanılmıştır.

Araştırma Sorularına Verilen Yanıtlar

1. Veri setindeki eksik değerler neden ve nasıl tamamlanmıştır?

Analizler için kullanılacak olan klasik ve modern istatistiksel yöntemlerin hemen hemen hepsi veri setinin eksiksiz olduğu varsayımı altında geliştirilmiştir. Bu nedenle veri setlerindeki eksik değerler sorun oluşturur ve eksik değerlerin tamamlanması gerekmektedir. Eksik değerlerin tamamlanmasında kullanılan yöntem önem arz etmekte olup sınıflandırma başarısını etkilemektedir. Literatürde veri setindeki eksik değerlerin tamamlanması için birçok yöntem kullanılmakta olup detaylı bilgi Bölüm 3.1.2 de sunulmuştu.

Evrimsel yöntemler sınıflandırma ve optimizasyonda yaygın olarak kullanılmaktadır. Ancak eksik değer tamamlamada evrimsel yöntemlerin kullanımı çok yenidir ve Yapay Arı Kolonisi (ABC) yöntemi kullanılarak eksik değerler tamamlanmamıştır.

Tez çalışmasında ver setindeki eksik değerler, tek değer atama yöntemlerinden; ortalama ve ortanca, istatistiksel yöntemlerden; kNN, mice ve missForest, evrimsel yöntemlerden ise geliştirilen ABC yöntemi ile tamamlanmıştır. Bu yöntemlerin başarıları Ortalama Karesel Hatanın Karekökü (RMSE) değerine göre karşılaştırılmıştır. Çalışma sonucunda en küçük RMSE değeriyle yani gerçek değere en yakın tahminler üreten yöntemin ABC olduğu gözlemlenmiştir.

2. Veri setindeki veri dağılımı normal midir? Veri normal dağılım göstermiyorsa hangi normalizasyon yöntemi kullanılarak veriler normalize edilebilir?

Kutu grafiği ile görselleştirilen veri setinin normal dağılım göstermediği Bölüm 5.1'de görülmüştür. Verileri normalize etmede birçok yöntem kullanılmaktadır. Bu tez çalışmasında kullandığımız veri seti için en başarılı normalizasyon tekniğini belirlemek için literatürde sıklıkla kullanılan minimum-maksimum, ondalık ölçeklendirme, z-değeri ve sigmoid normalizasyon yöntemleri karşılaştırılmıştır. Başarı karşılaştırması için bu yöntemler ile veriler normalize edildikten sonra k-ortalama kümeleme yöntemi ile analiz edilmiştir. Kümeleme başarısı saflık ve entropi kriterlerine göre ölçülmüş olup, 0.7349 saflık ve 0.8594 entropi değeriyle en başarılı normalizasyon tekniğinin sigmoid olduğu belirlenmiştir.

3. Hasta kuzuları sınıflandırmada en başarılı yöntem hangisidir?

Neonatal kuzularda hastalık sınıflandırması için; karar ağacı (J48), Saf bayes, k-en yakın komşu, yapay sinir ağı ve rastgele orman sınıflandırma algoritması kullanılmıştır. Sınıflandırıcıların performansları doğruluk, dengeli doğruluk, seçicilik, duyarlılık, f-ölçütü, kappa ve AUC ölçütlerine göre karşılaştırılmıştır. Model başarım ölçütleri incelendiğinde, 0.8427 doğruluk, 0.7132 dengeli doğruluk, 0.5164 duyarlılık, 0.5226 f-ölçütü, 0.4304 kappa ve 0.765 AUC değeri ile en başarılı yöntemin Saf Bayes olduğu gözlemlenmiştir. Kappa değerinin 0,40'dan büyük olması modelin tutarlı tahminler yaptığını göstermektedir. Özellikle sağlık alanında yapılan modellerin değerlendirilmesinde kullanılan AUC değerinin 0,765 olması geliştirilen modelin verileri sınıflandırmada ayrım

gücünün yüksek olduğunu belirtir. Veri setinde 287 sağlıklı ve 60 hasta kuzu olmasından dolayı dengeli bir sınıf dağılımı söz konusu değildir. Nitekim doğruluk değerinin, dengeli doğruluk değerinden daha yüksek olması bize dengesiz bir sınıf dağılımının olduğunu göstermektedir. 347 (60 hasta, 287 sağlıklı) kuzu içerisinden hasta kuzuları tahmin etme oranı yaklaşık %17 iken veri madenciliği yöntemi ile bu oran yaklaşık %52'dir.

Bilgisayar destekli tanı ile yeni doğan kuzulardan hasta olabilecekler tahmin edilerek erken teşhis ve tedavide veteriner hekime yardımcı olunabilecektir. Erken teşhis ve tedavi sayesinde hastalıklar ve ölümlerdeki azalma ile ülke ekonomisine katkı sağlanabilecektir.

4. Daha az özellik kullanılarak daha başarılı sınıflandırma yapmak mümkün müdür? Daha az özellik kullanmak neden önemlidir? Hasta ve sağlıklı kuzuları ayırmada önemli rol oynayan özellikler hangileridir?

Birçok alanda verilerin toplanması zordur ancak hayvancılık alanında verilerin toplanması hem çok daha zor hem de çok masraflıdır. Bu nedenle başarılı sınıflandırma ne kadar az özellik ile yapılırsa o kadar emek, zaman ve maliyet azalır.

Tez çalışmasında bilgi kazanımı (information gain) yöntemi ile özellik seçimi yapıldığında, özellik sayısı 14'ten 4'e düşürüldüğünde sınıflandırma doğruluğunun %4 arttığı gözlemlenmiştir. Yani özellik sayısı yaklaşık %70 azaltılarak daha başarılı sınıflandırma elde edilmiştir. Bu da bize sadece 4 özellik (TP, IgG, GGT ve WG28) kullanılarak yüksek performanslı sınıflandırma yapılabileceğini göstermektedir. Yani neonatal kuzularda hasta ve sağlıklıları ayırmada TP, IgG, GGT ve WG28 özelliklerinin önemli rol oynadığı söylenebilir.

5. Kuzu ölümleri ile kan seviyeleri arasındaki ilişkiye bakıldığında ölümlerin en fazla gerçekleştiği kan değer aralıkları nelerdir?

Kümülatif dağılım fonksiyon sonuçlarına göre ayrıklaştırılan kan değerleri, ölümler ile kan değerleri arasındaki en yüksek ortak bilgi (Mutual Information) temel alınarak eşik değer belirlenmiştir. Buna göre neonatal dönemde en fazla ölümlerin gerçekleştiği kan değer aralıkları şu şekildedir; IgG < 500, GGT < 500, LT 1201-1600 aralığında, TP 31-40 aralığında ve ALB < 35.

6. Veteriner hekime yardımcı masaüstü ve mobil uygulama neden gereklidir?

Veteriner hekime yardımcı ücretsiz masaüstü ve mobil uygulama ile karşılaşılmamıştır. Tez çalışmasında kuzulara ait verinin depolandığı, hastalığa yönelik analizlerin yapıldığı ve hasta kuzuların takibinin sağlanmasını amaçlayan mobil ve masaüstü bir yazılım geliştirilmiştir. Bu da, sahadan elde edilen veriye veri madenciliği tekniklerinin uygulanması ve etkin olarak kullanılabilir olması açısından ayrıca önem arz etmektedir.

Geliştirilen mobil ve masaüstü uygulaması ile veteriner hekim, hayvana ait tüm bilgileri veritabanına kolay bir şekilde girerek kayıt altına alabilmektedir. Mobil ve masaüstü uygulamasının ortak veritabanı ile iletişime geçmesi sayesinde veri girişi, veri depolanması kolaylaşırken veri doğruluğu ve tutarlılığı sağlanır. Ayrıca gereksiz veri tekrarları engellenmiş olur. Geliştirilen veritabanı sayesinde veri madenciliği ile ilgili yapılacak çalışmalarda veri ambarı yönünden sağlam bir temel oluşturacaktır. Çünkü sağlık alanında hasta ve hastalıkla ilgili bilgilerin doğru bir şekilde elde edilmesi, işlenmesi, analizi, değerlendirilmesi, sunulması ve arşivlenmesi önemlidir. Veri tabanları, veriyi veri madenciliğine hazırlayarak veri madenciliği sürecini başlatırlar. Aynı zamanda bilgisayar destekli tanı için bir altyapı oluşturulmuş olur. Mobil uygulama ile yazılım aktif olarak sahada kullanılabileceği için hem verinin depolanması kolaylaşacak hem de hekim tarafından analizler hızlı bir şekilde gerçekleştirilebilecektir. Masaüstü uygulama sayesinde ise hekimler kliniklerde veriyi arşivleyip, analizleri gerçekleştirebilecektir.

Tez çalışmasında kullanılan veri seti dengesiz bir dağılım göstermektedir. Gelecek çalışmalarda veri madenciliği yöntemleriyle hasta kuzu sayısı arttırılarak analizler yeniden yapılabilir. Ölümlerde etkili olan kan seviyeleri için farklı yöntemler kullanılarak eşik seviyesi belirlenebilir.

KAYNAKLAR

[1] Alpaydin, E., (2014). Introduction to machine learning: MIT press.

[2] Koh, H.C. ve Tan, G., (2011). "Data mining applications in healthcare", Journal of healthcare information management, 19: 65.

[3] Cihan, P. Gökçe, E. ve Kalipsiz, O., (2017). "A Review of Machine Learning Applications in Veterinary Field", Kafkas Universitesi Veteriner Fakultesi Dergisi, 23: 673-680.

[4] Milli Eğitim Bakanlığı, Hayvan Yetiştiriciliği Küçükbaş Hayvan Seçimi, http://www.megep.meb.gov.tr/mte_program_modul/moduller_pdf/K%C3%BC%C3%A7%C3%BCkba%C5%9F%20Hayvan%20Se%C3%A7imi.pdf, 30 Nisan 2016.

[5] Gökçe, E., (2007). Neonatal Kuzu Morbidite ve Mortalitelerinin Klinik Sebepleri ve Muhtemel Risk Faktörlerinin Belirlenmesi, Doktora Tezi, Kafkas Üniversitesi Sağlık Bilimleri Enstitüsü, Kars.

[6] Akıllı, A. Atıl, H. Takma, Ç. ve Ayyılmaz, T., (2016). "Fuzzy Logic-Based Decision Support System for Dairy Cattle", Kafkas Universitesi Veteriner Fakultesi Dergisi, 22: 13-19.

[7] Hempstalk, K. McParland, S. ve Berry, D., (2015). "Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows", Journal of dairy science, 98: 5262-5273.

[8] Küçükönder, H. Üçkardeş, F. ve Narinç, D., (2014). "Hayvancılık Alanında Bir Veri Madenciliği Uygulaması: Japon Bıldırcını Yumurtalarında Döllülüğe Etki Eden Bazı Faktörlerin Belirlenmesi", Kafkas Universitesi Veteriner Fakultesi Dergisi, 20.

[9] Lewis, F. Brülisauer, F. ve Gunn, G., (2011). "Structure discovery in Bayesian networks: An analytical tool for analysing complex animal health data", Preventive veterinary medicine, 100: 109-115.

[10] Karabag, K. Alkan, S. ve Mendes, M., (2010). "Classification tree method for determining factors that affecting hatchability in Chukar Partridge (Alectoris chukar) eggs", Kafkas Univ. Vet. Fak. Derg, 16: 723-727.

[11] Ortiz-Pelaez, Á. ve Pfeiffer, D.U., (2008). "Use of data mining techniques to investigate disease risk classification as a proxy for compromised biosecurity of cattle herds in Wales", BMC veterinary research, 4: 24.

[12] Gokce, E. Atakisi, O. Kirmizigul, A.H. Unver, A. ve Erdogan, H.M., (2014). "Passive immunity in lambs: serum lactoferrin concentrations as a predictor of IgG concentration and its relation to health status from birth to 12 weeks of life", Small ruminant research, 116: 219-228.

[13] Teke, E.Ç. Orhan, H. Küçüksille, E.U. Bilginturan, S. ve Teke, H., (2013). Veri Madenciliği Süreci ile Siyah Alaca Sığırlarda Canlı Ağırlık Tahmini, ed^eds. 8. Ulusal Zootekni Bilim Kongresi. Çanakkale, 365.

[14] Ghotoorlar, S.M. Ghamsari, S.M. Nowrouzian, I. ve Ghidary, S.S., (2012). "Lameness scoring system for dairy cows using force plates and artificial intelligence", The Veterinary record, 170: 126-126.

[15] Takma, Ç. Atıl, H. ve Aksakal, V., (2012). "Çoklu doğrusal regresyon ve yapay sinir ağı modellerinin laktasyon süt verimlerine uyum yeteneklerinin karşılaştırılması", Kafkas Üniversitesi, Veterinerlik Fakültesi Dergisi, 18: 941-944.

[16] Hermann-Bank, M.L. Skovgaard, K. Stockmarr, A. Strube, M.L. Larsen, N. Kongsted, H. Ingerslev, H.-C. Mølbak, L. ve Boye, M., (2015). "Characterization of the bacterial gut microbiota of piglets suffering from new neonatal porcine diarrhoea", BMC veterinary research, 11: 139.

[17] Dupuy, C. Morignat, E. Maugey, X. Vinard, J.-L. Hendrikx, P. Ducrot, C. Calavas, D. ve Gay, E., (2013). "Defining syndromes using cattle meat inspection data for syndromic surveillance purposes: a statistical approach with the 2005–2010 data from ten French slaughterhouses", BMC veterinary research, 9: 88.

[18] Kiliç, İ. ve Özbeyaz, C., (2010). "Bulanık kümeleme analizinin koyun yetiştiriciliğinde kullanımı ve bir uygulama", Kocatepe Veterinary Journal, 3.

[19] Warns-Petit, E. Morignat, E. Artois, M. ve Calavas, D., (2010). "Unsupervised clustering of wildlife necropsy data for syndromic surveillance", BMC veterinary research, 6: 56.

[20] Gürcan, S. ve Akçapınar, H., (2002). "Alman Et ve Karacabey Merinosu koyunlarının canlı ağırlık, vücut ölçüleri ve yapağı inceliği yönünden kümeleme analizi ile incelenmesi", Turkish Journal of Veterinary and Animal Sciences, 26: 1255-1261.

[21] Bozkurt, Y. Aydoğan, T. ve Tüzün, C.G., (2013). Açıkta besi (feedlot) sisteminde yetiştirilen esmer ve siyah alaca ırkı hayvanların sayısal görüntü işleme ve yapay sinir ağları yöntemi ile performans ve karkas özelliklerinin saptanması, ed^eds.: Tübitak Projesi, Proje no: 111O269.

[22] McEvoy, F.J. ve Amigo, J.M., (2013). "Using machine learning to classify image features from canine pelvic radiographs: evaluation of partial least squares discriminant analysis and artificial neural network models", Veterinary Radiology & Ultrasound, 54: 122-126.

[23] Slósarz, P. Stanisz, M. Boniecki, P. Przybylak, A. Lisiak, D. ve Ludwiczak, A., (2011). "Artificial neural network analysis of ultrasound image for the estimation of intramuscular fat content in lamb muscle", African Journal of Biotechnology, 10: 11792.

[24] Saidani, K. Lopez-Sandez, C. ve Pablo, D.-F., (2016). "Effect of climate on the epidemiology of bovine hypodermosis in Algeria", Kafkas Universitesi Veteriner Fakultesi Dergisi, 22: 147-154.

[25] Awaysheh, A. Wilcke, J. Elvinger, F. Rees, L. Fan, W. ve Zimmerman, K.L., (2016). "Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats", Journal of Veterinary Diagnostic Investigation, 28: 679-687.

[26] Boujenane, I. ve El Aimani, J., (2015). "Incidence and occurrence time of clinical mastitis in Holstein cows", Turkish Journal of Veterinary and Animal Sciences, 39: 42-49.

[27] Amrine, D.E. White, B.J. ve Larson, R.L., (2014). "Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease", Computers and electronics in agriculture, 105: 9-19.

[28] Piwczyński, D. Sitkowska, B. ve Wiśniewska, E., (2012). "Application of classification trees and logistic regression to determine factors responsible for lamb mortality", Small ruminant research, 103: 225-231.

[29] Sandholm, T. Brodley, C. Vidovic, A. ve Sandholm, M., (1996). "Comparison of regression methods, symbolic induction methods and neural networks in morbidity diagnosis and mortality prediction in equine gastrointestinal colic".

[30] Pala, T., (2013). Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi, Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, istanbul.

[31] Persidis, A. ve Persidis, A., (1991). "Medical Expert Systems: An Overview", Journal of Management in Medicine, 5: 27-34.

[32] TIGEM, Hayvancılık Sektör Raporu, http://tarim.kalkinma.gov.tr/wp-content/uploads/2014/10/2013-TIGEM-HAYVANCILIK-SEKTOR-RAPORU.pdf, 15 Mayıs 2017.

[33] Türkiye İstatistik Kurumu, (2017). "Hayvancılık İstatistikleri", Ankara.

[34] Gökçe, E. Kırmızıgül, A.H. Atakişi, O. ve Erdoğan, H.M., (2013). "Risk factors associated with passive immunity, health, birth weight and growth performance in lambs: III. The relationship among passive immunity, birth weight gender, birth type, parity, dam's health and lambing season", Kafkas Univ Vet Fak Derg, 19: 741-747.

[35] Altuğ, N. Özdemir, R. ve Cantekin, Z., (2013). "Ruminantlarda Koruyucu Hekimlik: I. Aşı Uygulamaları", Journal of Faculty of Veterinary Medicine, Erciyes Üniversitesi Veteriner Fakültesi Dergisi, 10.

[36] Eales, F. ve Small, J., (1980). "Summit metabolism in newborn lambs", Research in veterinary science, 29: 211-218.

[37] Pakkanen, R. ve Aalto, J., (1997). "Growth factors and antimicrobial factors of bovine colostrum", International Dairy Journal, 7: 285-297.

[38] Kuralkar, P. Kuralkar, S. ve PGIVAS, A., (2010). "Nutritional and Immunological Importance of Colostrum for the new born", Veterinary World, 3: 46-47.

[39] Selk, G., (1998). "Management factors that affect the development of passive immunity in the newborn calf", Beef Cattle Handbook-2240. Extension Beef Cattle Resource Committee: 1-7.

[40] Arthington, J., (2001). "Technologies for delivering passive immunity to newborn calves".

[41] Rice, D.N. ve Rogers, D.G., (1990). "Colostrum quality and absorption in baby calves".

[42] Güngör, Ö. ve Baştan, A., (2004). "Gebe ineklerde uygulanan aşıların kolostrum ve buzağıda IgG konsantrasyonu üzerine etkileri", Ankara Üniv Vet Fak Derg, 51: 7-1.

[43] Maden, M. Altunok, V. Birdane, F.M. Aslan, V. ve Nizamlıoğlu, M., (2001). Sağlıklı Kuzularda Kolostral Pasif Transfer Durumunun Belirlenmesinde Serum Gamma-Glutamil Transferaz Enzim Aktivitesinin Önemi, ed^eds. Konya: Tubitak Projesi, Proje no: VHAG-1588 (1199V1102).

[44] Fayyad, U. Piatetsky-Shapiro, G. ve Smyth, P., (1996). "From data mining to knowledge discovery in databases", AI magazine, 17: 37.

[45] Chapman, P. Clinton, J. Kerber, R. Khabaza, T. Reinartz, T. Shearer, C. ve Wirth, R., (2000). "CRISP-DM 1.0 Step-by-step data mining guide".

[46] Han, J. Pei, J. ve Kamber, M., (2011). Data mining: concepts and techniques: Elsevier.

[47] Tan, P.-N., (2006). Introduction to data mining: Pearson Education India.

[48] Osborne, J.W., (2012). Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data: Sage Publications.

[49] Allison, P.D., (2003). "Missing data techniques for structural equation modeling", Journal of abnormal psychology, 112: 545.

[50] Pigott, T.D., (2001). "A review of methods for missing data", Educational research and evaluation, 7: 353-383.

[51] D'Agostino, R.B., (2007). "Overview of missing data techniques", Topics in Biostatistics: 339-352.

[52] Buhi, E.R. Goodson, P. ve Neilands, T.B., (2008). "Out of sight, not out of mind: strategies for handling missing data", American journal of health behavior, 32: 83-92.

[53] Cheema, J., (2012). Handling missing data in educational research using SPSS: George Mason University.

[54] Baraldi, A.N. ve Enders, C.K., (2010). "An introduction to modern missing data analyses", Journal of school psychology, 48: 5-37.

[55] Acock, A.C., (2005). "Working with missing values", Journal of Marriage and family, 67: 1012-1028.

[56] Sinharay, S. Stern, H.S. ve Russell, D., (2001). "The use of multiple imputation for the analysis of missing data", Psychological methods, 6: 317.

[57] Rubin, D.B., (1976). "Inference and missing data", Biometrika, 63: 581-592.

[58] Allison, P.D., (2002). "Missing data: Quantitative applications in the social sciences", British Journal of Mathematical and Statistical Psychology, 55: 193-196.

[59] Little, R.J., (1988). "A test of missing completely at random for multivariate data with missing values", Journal of the American Statistical Association, 83: 1198-1202.

[60] Roth, P.L., (1994). "Missing data: A conceptual review for applied psychologists", Personnel psychology, 47: 537-560.

[61] Alpar, R., (2013). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Detay Yayıncılık, Ankara.

[62] Schafer, J.L., (1999). "Multiple imputation: a primer", Statistical methods in medical research, 8: 3-15.

[63] Tabachnick, B.G. Fidell, L.S. ve Osterlind, S.J., (2001). "Using multivariate statistics".

[64] Schafer, J.L. ve Graham, J.W., (2002). "Missing data: our view of the state of the art", Psychological methods, 7: 147.

[65] Soley-Bori, M., (2013). "Dealing with missing data: Key assumptions and methods for applied analysis", Boston University.

[66] Little, R.J. ve Rubin, D.B., (2014). Statistical analysis with missing data: John Wiley & Sons.

[67] Stekhoven, D.J. ve Bühlmann, P., (2011). "MissForest—non-parametric missing value imputation for mixed-type data", Bioinformatics, 28: 112-118.

[68] Buuren, S. ve Groothuis-Oudshoorn, K., (2011). "mice: Multivariate imputation by chained equations in R", Journal of statistical software, 45.

[69] Olson, D.L. ve Delen, D., (2008). Advanced data mining techniques: Springer Science & Business Media.

[70] Fayyad, U. ve Irani, K., (1993). "Multi-interval discretization of continuous-valued attributes for classification learning".

[71] Larose, D.T., (2014). Discovering knowledge in data: an introduction to data mining: John Wiley & Sons.

[72] Özkan, Y., (2008). Veri madenciliği yöntemleri: Papatya Yayıncılık Eğitim.

[73] Cihan, P. Kalipsiz, O. Ve Gökçe, E., (2017). "Hayvan Hastalığı Teşhisinde Normalizasyon Tekniklerinin Yapay Sinir Ağı ve Özellik Seçim Performansına Etkisi", Electronic Turkish Studies, 12.

[74] Bishop, C.M., (1995). Neural networks for pattern recognition: Oxford university press.

[75] Jayalakshmi, T. ve Santhakumaran, A., (2011). "Statistical Normalization and Back Propagationfor Classification", International Journal of Computer Theory and Engineering, 3: 89.

[76] Roiger, R.J., (2017). Data mining: a tutorial-based primer: CRC Press.

[77] Kuhn, M., (2015). "Caret: classification and regression training", Astrophysics Source Code Library.

[78] Wang, J., (2005). Encyclopedia of data warehousing and mining: IGI Global.

[79] Kröse, B. Krose, B. van der Smagt, P. ve Smagt, P., (1993). "An introduction to neural networks".

[80] Breiman, L., (2001). "Random forests", Machine learning, 45: 5-32.

[81] Witten, I.H. Frank, E. Hall, M.A. ve Pal, C.J., (2016). Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann.

[82] Dirican, A., (2001). "Tani Testi Performanslarının Değerlendirilmesi ve Kıyaslanması", Cerrahpaşa Tıp Dergisi, 32.

[83] Brodersen, K.H. Ong, C.S. Stephan, K.E. ve Buhmann, J.M., (2010). "The balanced accuracy and its posterior distribution": IEEE.

[84] Cohen, J., (1960). "A coefficient of agreement for nominal scales", Educational and psychological measurement, 20: 37-46.

[85] Fleiss, J.L. Levin, B. ve Paik, M.C., (2013). Statistical methods for rates and proportions: John Wiley & Sons.

[86] Landis, J.R. ve Koch, G.G., (1977). "The measurement of observer agreement for categorical data", Biometrics: 159-174.

[87] DeLong, E.R. DeLong, D.M. ve Clarke-Pearson, D.L., (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", Biometrics: 837-845.

[88] Krzanowski, W.J. ve Hand, D.J., (2009). ROC curves for continuous data: CRC Press.

[89] Robin, X. Turck, N. Hainard, A. Tiberti, N. Lisacek, F. Sanchez, J. ve Müller, M., (2014). "pROC: Display and analyze ROC curves. R package version 1.7. 3", R Found. Stat. Comput., Vienna.

[90] Gökçe, E. Atakişi, O. Kırmızıgül, A.H. ve Erdoğan, H.M., (2013). "Risk Factors Associated with Passive Immunity, Health, Birth Weight and Growth Performance in Lambs: I. Effect of Parity, Dam's Health, Birth Weight, Gender,

Type of Birth and Lambing Season on Morbidity and Mortality", Kafkas Universitesi Veteriner Fakultesi Dergisi, 19

[91] Gökçe, E. Atakişi, O. Kırmızıgül, A.H. ve Erdoğan, H.M., (2013). "Risk Factors Associated with Passive Immunity, Health, Birth Weight and Growth Performance in Lambs: II. Effects of Passive Immunity and Some Risk Factors on Growth Performance During the First 12 Weeks of Life", Kafkas Universitesi Veteriner Fakultesi Dergisi, 19: 619-627.

[92] Gökçe, E. Atakişi, O. Kırmızıgül, A.H. ve Erdoğan, H.M., (2013). "Risk Factors Associated with Passive Immunity, Health, Birth Weight and Growth Performance in Lambs: III- The Relationship among Passive Immunity, Birth Weight, Gender, Birth Type, Parity, Dam's Health, and Lambing Season", Kafkas Universitesi Veteriner Fakultesi Dergisi, 19.

[93] Liu, H. ve Motoda, H., (2007). Computational methods of feature selection: CRC Press.

[94] Ma, B.L.W.H.Y. ve Liu, B., (1998). "Integrating classification and association rule mining".

[95] Bow, S.T., (2002). Pattern recognition and image preprocessing: CRC press.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Pınar CİHAN

Doğum Tarihi ve Yeri : 15.03.1986, Diyarbakır

Yabancı Dili : İngilizce

E-posta : pkaya@nku.edu.tr

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Bilgisayar Müh.	Yıldız Teknik Üniversitesi	2013
Lisans	Bilgisayar Müh.	Bahçeşehir Üniversitesi	2010
Lise	Fen Bilimleri	Özel Amid Lisesi	2003

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2018	Namık Kemal Üniversitesi	Araştırma Görevlisi
2012	Yıldız Teknik Üniversitesi	Araştırma Görevlisi
2011	Namık Kemal Üniversitesi	Araştırma Görevlisi

YAYINLARI

Makale

1. CİHAN P., GÖKÇE E., KALIPSIZ O., A Review of Machine Learning Applications in Veterinary Field, Kafkas Univ Vet Fak Derg, vol. 23, pp. 673-680, 2017.

2. CİHAN P., KALIPSIZ O., GÖKÇE E., Hayvan Hastalığı Teşhisinde Normalizasyon Tekniklerinin Yapay Sinir Ağı ve Özellik Seçim Performansına Etkisi, Electronic Turkish Studies,, vol. 12, pp. 59-70, 2017.

Bildiri

1. CİHAN P., KALIPSIZ O., GÖKÇE E., Veri Madenciliği, Veri Ön İşleme Tekniklerinin Hayvancılık Verisine Uygulanması, Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı, 2016 .

2. CİHAN P., KALIPSIZ O., GÖKÇE E., Veterinerlik Alanında Bir Veritabanı Tasarımı, 2. Uluslararası Mühendislik Mimarlık ve Tasarım Kongresi, 2017.