HAYVANCILIKTA KULLANILAN FARKLI VERİ MADENCİLİĞİ ALGORİTMALARININ KARŞILAŞTIRILMASI

Musa YILMAZ
Yüksek Lisans Tezi

ZOOTEKNİ ANABİLİM DALI

1.Danışman: Doç. Dr. Ecevit EYDURAN
2.Danışman: Prof. Dr. Ömer AKBULUT

2017
Her hakkı saklıdır

IĞDIR ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

HAYVANCILIKTA KULLANILAN FARKLI VERİ MADENCİLİĞİ ALGORİTMALARININ KARŞILAŞTIRILMASI

Musa YILMAZ

ZOOTEKNİ ANABİLİM DALI

IĞDIR
2017

Her hakkı saklıdır

Doç. Dr. Ecevit EYDURAN ve Prof. Dr. Ömer AKBULUT danışmanlığında Musa YILMAZ tarafından hazırlanan bu çalışma 15/08/2017 tarihinde aşağıdaki jüri üyeleri tarafından Zootekni Anabilim Dalı'nda Yüksek lisans tezi olarak kabul edilmiştir.

Başkan: Prof.Dr. Ömer AKBULUT İmza

Üye: Doç.Dr. Ecevit EYDURAN İmza

Üye: Yrd.Doç.Dr. Mehmet Kazım KARA İmza

Üye: Yrd.Doç.Dr. Şenol ÇELİK İmza

Üye: Yrd.Doç.Dr. İsa YILMAZ İmza

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ….... / …... /2017 tarih ve 2017/ ….. sayılı kararı ile onaylanmıştır.

(İmza)

……

Doç. Dr. Süleyman TEMEL

Enstitü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada orijinal olmayan her türlü kaynağa eksiksiz atıf yapıldığını bildiririm.

Musa YILMAZ

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge, şekil ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZET

HAYVANCILIKTA KULLANILAN FARKLI VERİ MADENCİLİĞİ ALGORİTMALARININ KARŞILAŞTIRILMASI

YILMAZ, Musa

Iğdır Üniversitesi

Yüksek Lisans Tezi, Zootekni Anabilim Dalı

1. Danışmanı: Doç. Dr. Ecevit EYDURAN

2. Danışmanı: Prof. Dr. Ömer AKBULUT

Eylül 2017,39 Sayfa

Bu çalışmanın amacı, bazı veri madenciliği algoritmalarının koyunlardaki doğum tipi bakımından tahmin performansını karşılaştırmaktır. Bu amaç için, CART, CHAID, Exhaustive CHAID, Naive Bayes ve MLP algoritmaları uygulanmıştır. Tahmin performansı içerisinde, en iyi algoritmayı bulmak için, doğruluk oranı (%), duyarlılık, özgüllük ve ROC eğrisi altındaki alan hesaplanmıştır. Üzerinde çalışılan veriler, MLP algoritmaları için eğitim seti (% 80) ve test seti (% 20) olmak üzere iki ana kısma ayrılmıştır. Araştırma kapsamında, cinsiyet (dişi ve erkek), çiftlikler (Mastung, Quetta, Noshkı), ana yaşı, doğum ağırlığı, kuzulama mevsimi ve yıl bağımsız değişkenler olarak kullanılmıştır. Araştırmada kullanılan tüm algoritmaların üstün performans gösterdiği görülmüştür.

Anahtar Kelimeler: CART, CHAID, Exhaustive CHAID, ROC eğrisi, Veri Madenciliği

ABSTRACT

COMPARISON OF DIFFERENT DATA MINING ALGORITHMS USED IN ANIMAL SCIENCE

YILMAZ, Musa

Iğdır University

Msc. Thesis/Department of Animal Science

1st Supervisor: Assoc. Prof. Dr. Ecevit EYDURAN

2nd Supervisor: Prof. Dr. Ömer AKBULUT

September 2017,39 Pages

The goal of this study was to compare predictive performance of some data mining algorithms in birth type of sheep. For this goal, CART, CHAID, Exhaustive CHAID, Naive Bayes and MLP algorithms were applied. Accuracy rate (%), sensitivity, specifity, kappa statistic and area under ROC criteria were calculated for finding the best algorithm in the predictive performance. The studied data were split into two main parts, training set (80%) and testing set (20%) for only MLP algorithm predictors The data included under the study were gender (male and female), farms (Mastung,Quetta, Noshkı) dam age and weight at lambing, lambing season and year, respectively. As a result, it was determined that all the algorithms used in the investigation showed superior performance.

Key words: CART, CHAID, Data Mining, Exhaustive CHAID, ROC curve

ÖNSÖZ ve TEŞEKKÜR

Veri madenciliği farklı sınıflama yöntemleri ile kıyaslandığında birçok avantaja sahip olduğu açık bir şekilde gözlemlenebilmektedir. Nominal, ordinal ve sürekli değişkenler, veri madenciliği algoritmaları kapsamında etkili bir şekilde kullanılabilir. Hayvancılık alanında doğum tipini etkileyen faktörlerin belirlenmesi, üzerinde çalışılan ırka ait standartların ortaya konulması pratik açıdan önemlidir. Bu bakımdan bu çalışmada kullanılan Mengali koyunlarını doğum tipi bakımından incelemek amacıyla algoritmaların performansları değerlendirilmiştir. Üzerinde durulan bağımlı değişken kategorik ise sınıflama, sürekli değişken ise regresyon analizi kapsamında düşünülebilir. Özellikle incelenen değişkenlere ilişkin dağılımsal varsayımların ihlal olması durumundan veri madenciliği algoritmalarının kullanılması önerilebilir. Bu amaçla, CART, CHAID, Exhaustive CHAID, MLP ve Naive Bayes veri madenciliğine örnek olarak verilebilir. Hayvancılık alanında iki seviyeli (cinsiyet ve doğum tipi) bağımlı değişkenler için bu algoritmalar kullanılabilir.

Bu tezin yazım ve hazırlama aşamasında bana yol gösteren, tecrübe ve önerileri ile beni destekleyen, maddi ve manevi daha iyiye ulaşmam konusunda destek olan saygıdeğer danışman hocalarım Doç.Dr.Ecevit EYDURAN ve Prof.Dr. Ömer AKBULUT'a, Zootekni bölümü öğretim üyelerinden Yrd.Doç.Dr. M. Kazım KARA hocama ve verilerini kullanmamıza izin veren Prof. Dr. Mohammad Masood TARIQ hocama teşekkürlerimi sunarım.

Musa YILMAZ

Eylül, 2017

İÇİNDEKİLER

Sayfa No

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler

% ……..	Yüzde
cm …….	Santimetre
g ……...	Gram
Kg …….	Kilogram
mm ……...	Miligram

Kısaltmalar

CART ……...…	Sınıflandırma ve Regresyon Ağacı
CHAID …….....	Otomatik Ki-Kare Etkileşim Belirleme Analizi
DA…….	Doğum Ağırlığı
Exhaustive CHAID	Geniş Otomatik Ki-Kare Etkileşim Belirleme Analizi
MLP……..	Çok Katmanlı Algılayıcı Algoritması
ROC eğrisi……	İşlem Karakteristik Eğrisi
SA …….....	Sınıflandırma Ağacı
VM …….…....	Veri Madenciliği
YSA …….....	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

Sayfa No

ÇİZELGELER DİZİNİ

Sayfa No

1. GİRİŞ

Veri madenciliği (VM) günümüzde oldukça hız kazanan, aynı zamanda teknolojik açıdan ilerleyen bir bilgi teknolojisi olmakla birlikte büyük veri setlerinin içerisinde saklı kalmış olan istenilen önemli bilginin açığa çıkarılması için uygulanan bir yöntem olarak tanımlanmaktadır (Küçükönder ve ark., 2014). VM, diğer bir ifade ile; geniş çapta ve karmaşık durumda olan veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin ortaya çıkarılması işlemi olarak ifade edilmektedir (Albayrak ve Yılmaz, 2009).

VM'de genel amaç, toplanmış olan bilgilerin bir takım istatistik yöntemlerle incelenip ilgili kurum ve yönetim destek sistemlerinde uygulanmak üzere değerlendirilmesidir. VM, çok büyük miktarda bilginin depolandığı veri tabanlarından, amaç doğrultusunda, gelecek ile ilgili tahminler yapmamızı sağlayacak, anlamlı olan veriye ulaşma ve veriyi kullanma biçimidir (Savaş ve ark., 2012). Dijital verilerin, son yıllarda çok fazla artması ve bu verilerin büyük veri tabanlarında kaydedilmesi ile birlikte, zaman periyodları içerisinde bu verilerden en verimli biçimde faydalanma ihtiyacı doğmuştur. Bu yüzden Veri Tabanlarında Bilgi Keşfi-VTBK (Knowledge Discovery in Databases) adı altında sürekli ve yeni arayışlar ortaya çıkmıştır. Söz konusu yöntem veri tabanı bakış açısı, makine öğrenimi ve istatistiksel perspektif gibi üç farklı unsurun kombinasyonu ile uygulanmaktadır (Karabatak ve İnce, 2004).

Özellikle büyük örnek genişliğine sahip veri setlerindeki çeşitli ilişkilerin incelenmesi bakımından VM algoritmaları kullanılması iyi bir seçenektir (Grzesiak and Zaborski, 2012). VM; önceden bilinmeyen, geçerli ve uygulanabilir olup aynı zamanda istenilen değerli bilginin, verilerden elde edilmesi sürecidir. Bilgi değer oluşturur. VM deki muhtemel faydalar çok büyük olabiliyor. VM, veriler içerisinde gizli olan değerli bilgiyi açığa çıkartarak firmalarda önemli rekabet avantajı sağlayabilmektedir. Dünyanın önde gelen firmaları, veri madenciliği tekniklerini; müşteri ilişkileri yönetiminden, kredi derecelendirmeye; risk analizinden, satış tahminlerine kadar pek çok alanda başarıyla kullanmaktadır (Alkan ve Falay, 2007). Yapay zeka ve yapay sinir ağları ile birlikte hemen hemen bütün VM teknikleri, yüksek öğretimdeki bilim

adamlarının buluşuydu, fakat VM ilk defa yüksek öğretimde uygulanmamakla beraber, yüksek öğretim, VM için hala önem arz etmektedir (Luan, 2002).

VM, verilerin farklı bir bakış açısından analiz edilmesi ve kullanışlı bilgi halinde özetlenme sürecidir. Teknik olarak VM, büyük ve birbiriyle ilişkili veri tabanları içinde düzinelerce alan arasında korelasyonlar ve düzenler bulma sürecidir. Bir başka ifade ile veri madenciliği büyük ve karmaşık verilerde beklenmeyen patikaların, değerli yapıların ve ilginç ilişkilerin keşfedilmesi bilimidir (Tüzüntürk, 2010). VM'de diskriminant analiz yöntemi oldukça önem arz eden bir uygulama metodudur. Diskriminant Analizi (Discriminant Analysis), temelleri 1930'larda Fisher tarafından ortaya çıkarılan biyoloji, davranış bilimleri ve finans alanlarında sıklıkla kullanılan çok değişkenli istatistik yöntemlerden biridir. Diskriminant analizinin, Sınıflandırma (Classification) ve Ayırma (Discrimination) olmak üzere temel iki amacı bulunmaktadır: Analizin bu işlevlerinden dolayı eğer diskriminant analizi bir ayırma fonksiyonu belirlemeye yönelik olarak uygulandıysa Tanımlayıcı Diskriminant Analizi (Descriptive Discriminant Analysis) ve eğer sınıflama amacıyla uygulanmış ise Ayırıcı Diskriminant Analizi (Predictive Discriminant Analysis) olarak adlandırılır. (Yakut ve Elmas, 2013).

Sınıflandırma (Classification) ve Regresyon (Regression) Modelleri Tahmin etmede faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleridir. Sınıflamada tahmin edilen bağımlı değişken kategoriktir. Regresyonda ise bağımlı değişken süreklidir. Sınıflama ve regresyon modellerinde karar ağaçları, yapay sinir ağları, genetik algoritmalar, K-en yakın komşu ve Naive-Bayes gibi teknikler kullanılmaktadır. Karar Ağaçları (Decision Trees) Veri madenciliğinde karar ağaçları, kurulmasının ucuz olması, yorumlanmalarının kolay olması, veritabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir. Yapay Sinir Ağları (YSA), temelde tamamen insan beyni örneklenerek geliştirilmiş bir teknolojidir. Bilindiği gibi; öğrenme, hatırlama, düşünme gibi tüm insan davranışlarının temelinde sinir hücreleri bulunmaktadır. Naive-Bayes algoritmasında her kriterin sonuca olan etkilerinin olasılık olarak hesaplanması temeline dayanmaktadır. Kümeleme Modelleri (Clustering) ise bu modelde amaç üyelerinin birbirlerine çok

benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. K-En Yakın Komşu (K-Nearest Neighbor), VM'de sınıflama amacıyla kullanılan bir diğer teknik ise örnekseme yoluyla öğrenmeye dayanan k-en yakın komşu algoritmasıdır. Bu teknikte tüm örneklemler bir örüntü uzayında saklanır. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dahil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemi bulur. VM fonksiyonel açısından bakılacak olursa, VM aktiviteleri 3 sınıf altında toplanmıştır, bunlar; Keşif (Discovery), Tahmini modelleme (Predictive modeling) ve Adli analiz (Forensic analysis) olmak üzere (Ayık ve ark., 2007).

Hayvancılık alanında ise önceki yıllarda yapılan çalışmalarda, sınıflandırma (classification) amacıyla kullanılan CART, CHAID, EXHAUSTIVE CHAID, Naive Bayes, Yapay Sinir Ağları (YSA) gibi VM algoritmalarının önemi vurgulanmıştır. Elde edilen sonuçların hayvancılık açısından önemli olduğu belirtilmiş ve bu açıdan incelenen özelliklerden yola çıkarak ırkların karakterizasyonu, cinsiyetin karakterizasyonu, döl tutma oranının karakterizasyonu, mastitisin karakterizasyonu gibi konularda en uygun veri madenciliği algoritmasının kullanımı hayvancılıkta doğru stratejilerin geliştirilmesi bakımından önemli olacaktır. Ayrıca sınıflandırma amacıyla en uygun algoritmanın kullanılması, üzerinde çalışılan ırklara ait ırk standartlarının doğru tanımlanması açısından yetiştiricilere katkı sağlayacaktır. Kullanımı ve yorumlaması kolay olan VM algoritmaları ile hayvancılıkta istenilen ya da istenmeyen özelliği ön plana çıkaran bağımsız değişkenler ve bu değişkenlere ait kombinasyonların tanımlanması ileride yapılacak çalışmalar için temel oluşturacaktır. Bu çalışmada, koyunlarda ikizlik oranı ile ilgili olarak bazı veri madenciliği algoritmalarının sınıflama performanslarının karşılaştırılması yapılmıştır. Bu amaç için, CART, CHAID, Exhaustive CHAID, Naive Bayes algoritmaları kullanılmıştır. En iyi algoritmayı belirlemek amacıyla, doğru sınıflama oranı, duyarlılık, özgüllük ve ROC eğrisi altında kalan alan gibi ölçütler hesaplanmıştır.

2. KAYNAK ÖZETLERİ

Nisbet et al. (2009), SPSS, STATISTICA ve SAS paket programlarını kullanarak CART, CHAID, Boosted Classifiers ve Regression, MARS, Regresyon ve Sınıflama için Random Forest, Makine Öğrenme (Bayesyan, Destek Vektörleri, En yakın komşuluk, Bağımsız Bileşenler gibi ileri veri madenciliği algoritmalarını detaylı olarak incelemişlerdir.

2.1.Sağlık Sektöründe Veri Madenciliği

Ozgulbaş ve Koyuncugil. (2007), 2004 yılına ilişkin Türkiye'de kamu hastanelerinin mali karlarının belirlenmesi VM metodu kapsamında Ki-kare (Chi-square) ve CHAID karar ağacı algoritması uygulamışlardır. Bu çalışmada 2004 yılındaki veriler kullanılmıştır. Böylece, Türkiye'de Sağlık Bakanlığı tarafından işletilen 645 kamu hastanesini kapsayan bir çalışma yürütülmüştür. Çalışma sonucunda %9.15 (59 hastane) mali performansı açısından kötü olduğu, %90.85 finansal performansı açısından iyi olduğunu belirlenmiştir. Kamu hastaneleri, devam eden ulusal reformun bir parçası olarak mali problemlerine acil çözüme ihtiyaç duymakta ve bireysel olarak işletilen döner sermaye fonları ve genel bütçeden tahsis edilen fonların bir karışımı olarak finanse edilmiştir. Mali performansları açısından kamu hastaneleri finansal açıdan, CHAID karar ağacı ile incelenmiştir.

Sut ve Şimşek. (2011), kazalarda kafa yaralanması sonucu oluşan ölüm oranını tahmin etmek amacıyla sınıflama performansı bakımından altı farklı karar ağacı algoritmasını (CART, CHAID, Exhaustive-CHAID, QUEST, and Boosted Tree Classifiers and Regression (BTCR)) birbiriyle karşılaştırılmıştır. Değerlendirilen algoritmaların performansları, hassasiyet oranı (sensitivity), özgüllük oranı (specificity), pozitif/negatif tahmin oranı (positive/negative predictive) ve isabet oranı (accuracy rate) gibi bakımından ölçütleri kullanarak (0.801 ile 0.954) karşılaştırmıştır. Ayrıca, tüm algoritmalara ait ROC eğrisi altında kalan alanlar tahmin edilmiştir. Kafa yaralanması sonucu oluşan ölüm oranının sınıflandırılmasında, ROC eğrisi altında kalan en büyük alana sahip algoritmanın BTCR olduğu (0.954) ve bu algoritmaya ilişkin isabet oranının %93 olduğu belirlenmiştir. Bu çalışmada, kafa yaralanmaları sonucu oluşan ölüm oranına ilişkin doğru tahminler yapılması bakımından, BTCR algoritmasının kullanılmasının yararlı olabileceği sonucuna varılmıştır.

Camdeviren et al. (2007), yapmış oldukları bu çalışmada sınıflandırma ağaç modeli (CART) ile lojistik regresyon modeli karşılaştırarak sosyal-demografik risk faktörleri belirlemek için farklı postpartum periyodlarda 1447 kadının hangi depresyon durumundan etkilenmiş olduğunu belirlemeyi amaçlamıştır

2.2.Sanayi ve Hizmet Sektöründe Veri Madenciliği

Albayrak ve Yılmaz. (2009), İMKB endeksinde sanayi ve hizmet sektöründe faaliyet gösteren 173 işletmenin 2004-2006 yıllarına ait yıllık finansal göstergeleri CHAID algoritması ile analiz etmişlerdir. Bu algoritma için çapraz geçerlilik kuralı seçilmiş ve farklı ağaç yapılarını izlemek için algoritmanın durdurma kurallarına (stopping rules) uygun değişik dügüm sayıları denenmiştir. Karar ağaçlarında hedef değişken seçilen sektör bulgularına göre araştırma kapsamında İMKB100 endeksinde yer alan 173 işletmeden %73.4'ünün (127 işletme) sanayi sektöründe, %26.6'sının (46 işletme) hizmet sektöründe olduğu anlaşılmaktadır. CHAID karar ağacı ile işletmelerin sektör profilleri oluşturulmuş ve işletmeler 7 farklı profilde sınıflandırılmıştır. En önemli bağımsız değişkenler sırasıyla işletme sermayesinin net satışlara oranı, stok devir hızı ve ekonomik rantabilite oranı değişkenleridir.

Yakut ve Elmas. (2013), yapmış oldukları çalışmada, işletmelerin finansal başarısızlığının veri madenciliği ve diskriminant analizi modelleri ile değerlendirmiştir. Daha sonra kontrol grubu ve veri seti kullanarak İMKB'de işlem gören 140 sanayi işletmesinin 2005-2008 yılları arasındaki finansal başarısızlıkları veri madenciliği ve diskriminant analizi modelleri ile tahmin ederek hangi yöntemin daha iyi sonuç verdiğini tespit etmişlerdir.

Coşkun ve Baykal. (2011), bu çalışmada WEKA programı ile farklı sınıflandırma yöntemlerine ait algoritmalar kullanılarak modeller oluşturmuştur. Çalışma sonuçlarına bakıldığında J48 algoritmasının model testine ilişkin %86.36 doğruluk derecesiyle en iyi sonucu ortaya çıkardığı söylenebilir. Doğruluk ölçütü çok basit ve önemli bir kriterdir. Bu ölçüte göre J48 algoritmasını sırasıyla KStar, Lojistik Regresyon ve NaiveBayes algoritmaları takip etmektedir.

2.3.Hayvancılık Alanında Veri Madenciliği

Grzesiak and Zaborski. (2012), yaptıkları çalışmada CART (Classification and Regression Tree), CHAID (Chi-Square Automatic Interaction and Detection), MARS (Multivariate Adaptive and Regression Splines), ANN (Artificial Neural Network), ve diğer makine öğrenme metotları (Naive Bayes Classifier (NBC), Destek Vektör Makinaları (Support Vector Machines =SVM), ve k-en yakın komşuluklar (k-nearest neighbors k-NN) gibi veri madenciliği metotlarının teorik alt yapıları ve bu metotların konusunda detaylı bilgiler vermiştir.

Bayram et al. (2015), Gümüşhane ilinde bulunan özel bir süt sığırı işletmesinde 2004 ile 2006 yılları arasında buzağılayan 613 Siyah Alaca ineğin 947 yavrulama kaydı kullanmış, genetiksel olmayan bazı faktörlerin bu ineklerde güç ve ölü doğurma etkilerini CHAID algoritması ile belirlemiştir.

Caraviello et al. (2006), çalışmasında büyük süt işletmelerinde yetiştirilen laktasyon dönemindeki Holstein ırkı ineklerin laktasyon dönemindeki üreme özeliklerini etkileyen faktörleri belirlemeyi amaçlamıştır. Bu amaç için, 103 çiftçi üzerinde bakım ve idare, iş gücü, beslenme, barınak imkânları, üreme, genetik seleksiyon, iklim ve süt verimi ile ilgili bir anket çalışması gerçekleştirmiştir. İlkine doğurma oranı için 31076 laktasyon kaydı, 14804 inek ve 317 bağımsız değişken kullanılırken laktasyonun 150. gündeki genetik durumu için 17587 laktasyon kaydı, 9516 inek ve 341 bağımsız değişken kullanılmıştır. İlkine doğurma oranı için alternatif bir karar ağacı algoritmasının kayıtları %75,6 oranında doğru sınıfladığı belirlenmiştir.

Grzesiak *et al*. (2011), süt sığırlarında döl tutmayı etkileyen faktörlerin etkisini belirlemek amacıyla Naive Bayes (NBC) ve sınıflandırma ve regresyon ağacı (CART) metotlarını kullanmışlardır.Çalışmada, laktasyon sayısı, suni tohumlama mevsimi, ineğin tohumlama yaşı, ineklerde HF genlerinin oranı, gebelik oranı, gebelik süresi, süt protein ve yağ verimi ve bir önceki buzağılamadaki cinsiyeti bağımsız değişken olarak kullanılmıştır. Bağımlı değişken, iyi (bir ya da iki suni tohumlamadan sonra bir ineğin gebe kalması) ve kötü (ikiden fazla suni tohumlamadan sonra bir ineğin gebe kalması) döl tutma olmak üzere iki seviyeli (binary) olarak değerlendirilmiştir. Tahminlerin doğruluğu (accuracy classification) %83 bulunmuştur. NBC'ye nazaran CART algoritması zayıf döl tutma sınıfının hassasiyet oranı (sensitivity) tanımlanmasında daha

etkili bulunmuştur (P<0.01). Özgüllük oranı (Specificity) her iki algoritma için benzer bulunmuş olup, CART algoritmasında döl tutma düzeyini belirleyen değişkenler arasında, buzağılamadan döl tutmaya kadar geçen süre, buzağılama aralığı, ortalama vücut kondisyonu ile suni tohumlamadaki kondisyon arasındaki fark değişkenleri çok önemli bulunmuştur. Sonuç olarak bu çalışmada, özellikle CART algoritması kullanımının, ineklerin uygun yapay tohumlama zamanının tespitinde yararlı olabileceği sonucuna varılmıştır.

Karabağ et al. (2010), kınalı keklikler üzerinde yürüttükleri bir çalışmada, sınıflandırma ağacı yöntemini kullanarak çıkış gücüne etki eden yumurta özeliklerini belirlemişlerdir. Çalışmada yumurta dış özeliklerinden yumurta ağırlığı, yumurta hacmi, yumurta uzunluğu ve yumurta genişliğinin çıkış üzerinde önemli etkiye sahip olduğu %75.6 isabetle tahmin etmişlerdir. Çıkış gücü 18.1 g'dan hafif yumurtalar için %56.0 olarak belirlenirken, 18.1 g'dan ağır yumurtalarda ise %80 olarak tespit edilmiştir. Yumurta hacmi ve yumurta genişliği sırasıyla 27.2 ve 3.14 cm'den büyük olduğunda çıkış gücü yaklaşık %82.1 bulunmuştur.

Piwczynski (2009), Polonyanın Pomarze ve Kujavay bölgesinde bulunan on sürüden yaşları 2 ile 8 arasında değişen 6586 baş Polonya Merinosu üreme performans indeksi bakımından (reproductive performance index) değerlendirmiştir. Çalışmada çiftleştirilen koyunlardan elde edilen kuzu sayısına ilişkin meydana gelen varyasyondan sorumlu değişkenleri tanımlanmak amacıyla CART sınıflama algoritması kullanmıştır. En önemli bağımsız değişkenlerin ana yaşı, sürü, doğum tipi ve koyunların 16. ay canlı ağırlığı olduğu tespit edilmiştir. Ayrıca, çiftleştirilen her koyun başına elde edilen kuzu sayısı üzerinde modele alınan faktörlerinin (sürü, ana yaşı, doğum tipi, vucut ağırlığı) etkili olduğu bildirilmiştir.

Piwczynski and Sitkowska. (2012a), 455 baş Polonya Holstein ırkı inekte somatik hücre sayısını etkileyen laktasyonla ilgili faktörlerin (laktasyon sırası, sürü büyüklüğü, verim düzeyi, buzağılama yılı, buzağılama mevsimi, test günü mevsimi, laktasyon aşaması ve test sağımında elde edilen süt miktarı) etkilerini belirlemek amacıyla sınıflama ağacı ve lojistik regresyon metotlarını karşılaştırmışlardır. Sınıflama ağacının oluşmasında Entropy fonksiyonu ve Gini katsayısı gibi ayırma ölçütleri temel alınmıştır. Modellerin uyum kalitesi, kare hata (squared error), yanlış sınıflama oranı

(misclassification rate), cumulative lift, kolmogorov-smirnov statistic ve ROC eğrisi altında kalan alan gibi ölçütlere göre değerlendirilmiştir. Somatik hücre sayısı bakımından en iyi modelleme, Entropy fonksiyonunu temel alan sınıflama ağacı tekniği ile elde edilmiştir. Elde edilen sonuçlar değerlendirildiğinde, yüksek süt verimli ineklerin bulunduğu sürülerde daha özel bakım koşullarının sağlanmasının önemli olduğu anlaşılmıştır.

Piwczynski *et al*. (2012b), tarafından 20044 baş Polonya Merinos kuzularının ölüm oranından sorumlu olan faktörleri tespit etmek amacıyla logistik regresyon ve sınıflama ağacı metotlarını kullanmışlardır. Bu kuzuların doğumdan 100. güne kadar olan yaşama gücü binomiyal olarak (canlı ve ölü) ifade edilmiştir. Araştırıcılar, Gini indeks ve Entropy fonksiyonu gibi ayırma ölçütlerine göre iki farklı model geliştirmişlerdir. Karar ağacı modelleri ve çoklu regresyonlar; karesel hata, yanlış sınıflandırma oranı, kolmogorov-smirnov ve ROC eğrisi altında kalan alan gibi ölçütlere değerlendirilmiştir. Entropy fonksiyonu ve Gini indeksini esas alan sınıflama ağaçlarının, yaşama gücüyle ilgili özelliklere (doğumdan 100. güne kadar olan yaşama gücü) ilişkin varyasyonu açıklamada daha isabetli sonuçlar verdiğini tespit etmişlerdir. Ayrıca, en iyi sınıflama modeli Gini indeksi ile elde edilmiş, geliştirilen bu modelde farklılaşmayı sağlayan en önemli bağımsız değişkenlerin sürü, kuzuların doğum yılı ve tipi olduğu belirlenmiştir.

Piwczynski *et al* (2013), 1257 baş Holstein ineğin olduğu bir populasyonda, ölü doğum (stillbirths) ve buzağılama kolaylığını etkileyen faktörlerin etkilerini belirlemek amacıyla bazı sınıflama algoritmaları kullanmışlardır.Bu araştırmada CART ve QUEST algoritmaları kullanılarak elde edilen sınıflama ağaçları, üç ayırma ölçütü (Pearson chi-squared, Entropy function ve gini index) ve beş uyum iyiliği ölçütüne (squared error, misclassificaton rate, cumulative lift, kolmogorov-smirnov statistic ve ROC eğrisi altında kalan alana) göre değerlendirilmiştir. Araştırıcılar Ki-kare istatistiği ve Entropy fonksiyonunu esas alan sınıflama ağaçları, buzağılama kolaylığına ilişkin varyasyonu daha doğru tanımlamışlardır. Ancak, ayırma kriterlerine bakılmaksızın, ölü doğum özelliğine ilişkin elde edilen sınıflama ağaç modelleri birbirine benzer bulunmuştur. Araştırmada, buzağılama kolaylığı üzerinde etkili olan değişkenlerin önem sıralamasına bakıldığında, en etkili değişkenin buzağıların canlı ağırlığı olduğu ve

bunu laktasyon sırası, yetiştirme sistemi, gebelik süresi uzunluğu ve buzağı cinsiyeti değişkenlerinin izlediği bildirilmiştir. Ölü doğum üzerinde sadece buzağı doğum ağırlığı değişkeninin etkili olduğu belirlenmiştir. Bu çalışma sonunda, grafiksel bir model olan sınıflama algoritmalarının buzağılamadan sorumlu olan faktörlerin belirlenmesinde yarar sağlayabileceği bildirilmiştir.

Küçükönder et al. (2014), yapmış oldukları çalışmada, Japon bıldırcını yumurtalarının döllülük üzerine etkisi olan mevsim seleksiyon ve yerleşim sıklığı faktörlerinin etkilerini belirlemişlerdir. Araştırmada kullanılan sınıflandırma algoritmaları sırasıyla YSA, RBF Network, Naive Bayes, KStar, ve Ridor algoritmalarıdır. Bu algoritmalara göre oluşturulan modellerin karşılaştırmasında Kappa istatistiği, Ortalama Mutlak Hata (OMH), Ortalama Hata Karekök (OHK), Göreli Mutlak Hata (GMH) ve Göreli Hata Karekök (GHK) performans ölçütleri kullanılmıştır. Analizler sonucunda, yaptıkları karşılaştırmada, performans kriter değerleri sırasıyla OMH: 0.002, OHK: 0.05, GMH: %1.07, GHK: %14.50 ve Kappa: 0.98 olan Ridor algoritmasının en az hata ile sınıflama yaptığı görülmüştür.

Topal et al. (2010), çalışmalarında, CHAID algoritması kullanarak, esmer alabalıklarda cinsiyet tanımlaması yapmışlardır. Alabalıklarda cinsiyet belirlemeyi, üretken periyod esnasındaki morfolojik özellikleri ölçerek gerçekleştirmişler. Yapılan bu çalışmada, vücut uzunluğu ve ağırlığı, eni, çatal uzunluğu, yağlı yüzgeç uzunluğu, baş uzunluğu ve toplam uzunluk üretken olmayan periyod esnasında ölçmüşlerdir. Cinsiyet belirleme metodu'nun CHAID olduğu saptanmıştır

Çetin ve Mikail. (2016), yaptıkları çalışmada, k-ortalama yaklaşımı, k-en yakın komşu yaklaşımı, çok değişkenli uyarlanır regresyon eğrileri (Multivariate Adaptive Splines, MARS), Bayes sınıflandırıcıları (Naive Bayesian Classifier, NBC), yapay sinir ağları (Artificial Neural Networks, ANN) destek vektör makineleri (Support Vector Machines, SVM), karar ağaçları gibi VM yöntemleri hakkında önemli bilgiler vermişlerdir.

Eyduran et al. (2013), çalışmalarında, tarım alanının bir kolu olan hayvan bilimi ile ilgili veriler için sınıflandırma ağacı metodu uygulaması kullanarak örnek bir çalışma oluşturmaktadırlar. Sınıflandırma ağacı diyagramında, cinsiyet, bağımlı

değişken olarak göz önünde bulundurularak 138 Mengali Kuzusu verileri değerlendirilmiştir.

Üçkardeş et al (2014), Sınıflandırma ağacı algoritması (CHAID) algoritması kullanmışlardır. Japon bıldırcınlarında, mevsim ve kafeste stoklama yoğunluğu ve verimlilik üzerine genotipin etkisini açığa çıkarmayı hedeflemiştir.

3. MATERYAL ve METOT

3.1. Materyal

Bu çalışmanın materyalini Pakistanın farklı tarım işletmelerinde yetiştirilen (Araştırma istasyonları: Mastung, Quetta ve Noshki) Mengali koyun ırkına ait veriler oluşturmuştur. Bu amaçla elde edilen verilerden kesikli varyasyon gösteren doğum tipi (tekiz ve ikiz) ve çoklu doğum durumları veri madenciliği algoritmaları kullanılarak istatistiki değerlendirmeler yapılmıştır. Bağımsız değişken olarak, ana yaşı, ana ağırlığı doğum mevsimi, doğum ağırlığı ve sütten kesim ağırlığı kullanılmıştır.

3.2. Metot

Bu tez çalışmasında, sınıflandırma algoritmaları olarak CART, CHAID, Exhaustive CHAID, C5.0, Naive Bayes, ve Yapay Sinir Ağları algoritmalarından MLP kullanılmıştır. Söz konusu algoritmaların karşılaştırılmasında aşağıdaki model değerlendirme kriterleri baz alınmıştır.

3.2.1. Veri madenciliği algoritmaları

3.2.1.a. CART algoritması

İlk olarak Breiman *et al.* (1984) tarafından geliştirilen CART veri madenciliği algoritması, veri setindeki (learning sample) tüm bireyleri kapsayan kök düğümden başlayarak, yinelemeli bir şekilde devam ederek, homojen düğümlere ulaşana kadar her evrede bir düğümden yanlızca iki yavru düğüm oluşturan ikili karar ağacı algoritmasıdır. CART non-parametrik bir yöntem olmanın yanı sıra, bir veri setinin dağılımı ile ilgili herhangi bir varsayım gerektirmez ve önemli olan bağımsız değişkenleri otomatik bir şekilde ağaç yapısına dahil eder (Nisbet *et al*., 2009).

CART hem kategorik hem de sürekli değişkenleri kullanma yolu ile sınıflama ve regresyon problemlerinin çözümünde karar ağaçlarını kullanarak non-parametrik bir istatistiksel metot sergilemektedir. Bağımlı değişken kategorik olduğunda,elde edilen ağaçlar sınıflama ağaçları (Classification Tree) olarak adlandırılırlar. Bağımlı değişken sürekli olduğu durumda ise oluşan ağaçlar regresyon ağaçları (Regression Tree) olarak adlandırılır.

CART algoritmasının her bölünme durumu yanlızca bir bağımsız değişkenin değerine bağlıdır. Ayrıca, CART algoritması ile oluşturulacak bir karar ağacında

yorumlama basitliğinin sağlanması noktasından tahminleme performansını azaltan gereksiz olmayan dalların ya da çocuk düğümlerin ağaç yapısından budamak için SPSS programında budama (pruning) işlemi aktif hale getirilmelidir. Ancak budama işlemini, CHAID ve Geniş (Exhaustive) CHAID veri madenciliği algoritmaları için SPSS programı otomatik bir şekilde yapmaktadır. Diğer iki algoritmadan farklı olarak CART algoritması ile şekillendirilen bir karar ağacında aynı değişkenin bir dalda bir yada daha fazla bölünme (surrogate splits) olması durumu da vardır. Diğer taraftan, CART algoritmasının diğer algoritmalara kıyasla daha çok dallanma potansiyeli olan bir algoritma olduğu söylemek mümkündür. Böyle oluşumları önlemek yada minimize etmek için, üzerinde çalışılan veri setini öğrenme (Learning, Training) ve test (testing, validation) setlerine ayrılması, ebeveyn-yavru düğümde olası birey sayısının ve ağaç derinliği (tree depth) sayısının titizlikle ayarlanması gerektiği tavsiye edilir. Yaygın olarak öğrenme ve test setleri için 80:20 oranının kullanımı önerilir. Çok iyi bir ağaç yapısı için bu iki set için hesaplanan model kalite ölçütlerinin birbirine yakın olması çok önemlidir. Başka bir seçenek olarak iki set oluşturmak yerine çapraz geçerlik (validation) 10 alınıp, veri seti için ortalama bir hata değeri hesaplamak mümkündür.

Bir bireyin bağımlı değişkene ait ya da bütün bağımsız değişkenlere ait değerler kayıpsa, CART algoritmasında bu birey analiz kapsamına girmez.

$\boldsymbol{Y}$ ……Bağımlı değişken sıralı, isimsel ya da sürekli olabilir.

$\boldsymbol{X_m, m = 1, \dots, M}$…...Bütün bağımsız değişkenlerin şekillendiği set

ñ = $\{\boldsymbol{x_n, y_n}\}_{\boldsymbol{n=1}}^{\boldsymbol{N}}$……..Öğrenme örneğini kapsayan veri setinin tümü

ñ($\boldsymbol{t}$)……..t düğüme atanan öğrenme örnekleri

$\boldsymbol{w_n}$……n. bireye ait ağırlık ile ilgili durum ağırlığı,

$\boldsymbol{f_n}$……n. bireye ait frekans ağırlığı,

Bağımlı değişken SKA gibi sürekli olduğu zaman Gini, Twoing ve Sıralı Twoing ayırma kriterleri yerine ayırma kriteri (splitting criterion)

$\Delta(\delta(t)) = i(t) - P_L.i(t_L) - P_R.i(t_R)$ $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$……..…..…(3.1)

En Küçük Kareler Sapma (LSD) safsızlık ölçümü ile beraber kullanılır.

$$i(t) = \frac{\sum_{n \in \hbar(t)} w_n f_n (y_n - \bar{y}(t))^2}{\sum_{n \in \hbar(t)} w_n f_n} \quad \text{.......(3.2)}$$

$$p_L = N_w(t_L)/N_w(t)\,, p_R = N_w(t_R)/N_w(t)\,, N_w(t) = \sum_{n \in \hbar(t)} w_n f_n \quad \text{.......(3.3)}$$

Burada;

P_L ....t. düğümden sol yavru düğüme atanan bireylerin (kuzuların) oranı,

P_R ....t. düğümden sağ çocuk düğüme atanan bireylerin (kuzuların) oranı,

$i(t)$ ....t düğüme ilişkin safsızlık ölçüsü,

$i(t_L)$ ......sol yavru düğüme ilişkin safsızlık ölçüsü ve

$i(t_R)$ ......sağ yavru düğüme ilişkin safsızlık ölçüsünü temsil etmektedir (Oruçoğlu, 2011).

$$\bar{\boldsymbol{y}}(\boldsymbol{t}) = \frac{\sum_{n \in \hbar(t)} w_n f_n y_n}{N_w(t)} \quad \text{.......(3.4)}$$

Variable importance (Değişkenlik önemi); CART algoritmasında elde edilen son ağaç (T) ile ilgili olarak X bağımsız değişkeninin önem derecesi, X değişkeninin katkı (improvement) değerlerinin tüm düğümlerdeki ağırlıklı toplamı olarak tanımlanır.

$$\boldsymbol{M}(\boldsymbol{X}) = \sum_{t \equiv T}^{\Delta(\tilde{S} \quad X = T)} \quad \text{.......(3.5)}$$

X bağımsız değişkenine ait önem derecesi VI (X), en yüksek önem dereceli bağımsız değişkene oranla X değişkeninin sınıflama ağacına sağladığı katkıyı gösteren normalleştirilmiş miktar olarak ifade edilir. Önem derecesi 0 ile 100 arasında değişir. En yüksek önem derecesine sahip değişken 100 değerini alır.

$$VI(X) = \frac{M(X)}{\max_{x} M(X) \times 100} \quad \text{.......(3.6)}$$

Kategoriksel Bağımlı Değişken (Categorical Dependent Variable); Y kategorik ise, orada üç bölme kriterine ulaşılır: Bunlar Gini, Twoing ve Sıralı Twoing kriterleridir. Düğümdeki t, olasılıklara izin verir $p(j,t), p(t)$ ve $P(j|t)$ tarafından tahmin edilir.

$$p(j,t) = \frac{\pi(j)N_{w,j}(t)}{N_{w,j}} \quad \text{.....(3.7)}$$

$$p(t) = \sum_j p(j,t) \quad \text{.....(3.8)}$$

$$P(j|t) = \frac{p(j,t)}{p(t)} = \frac{p(j,t)}{\sum_j^{p(j,t)}} \quad \text{.....(3.9)}$$

$$N_{w,j=}\sum_{n\equiv n}^{wnfnI(y_n=j)} \quad \text{.....(3.10)}$$

$$N_{w,j}\left(\boldsymbol{t}\right) = \sum_{n\equiv n(t)}^{wnfnI(y_n=j)} \quad \text{.....(3.11)}$$

I (*a* = *b*) *a* = *b*, 0, aksi takdirde 1 değerini alır.

3.2.1.b. CHAID ve geniş (exhaustive) CHAID

CART ile kıyaslandığında, bir düğümün çok sayıda ayrılarak bölünmesine (multiple splitting) izin veren CHAID algoritması Kass (1980) tarafından geliştirilmiştir. Bu bakımından CHAID'e benzeyen Geniş (Exhaustive) CHAID algoritması ise Biggs *et al.*, (1991) tarafından geliştirilmiştir.

CART ikili ağaç yapısı oluştururken CHAID çoklu ağaç yapısı oluşturmaktadır. CHAID algoritmasında kategorik değişkenlere ait veri kümesi, bağımlı değişkeni en iyi yorumlayacak şekilde ayrıntılı olarak homojen alt gruplara bölünmektedir. CHAID sürekli ve kategorik değişkenler üzerinde uygulama yapabilmesi, ağaç yapısında her düğümü ikiden fazla alt gruba ayırabilmesi gibi nedenler dolayısıyla günümüzde çok tercih edilen bir algoritmadır.

SPSS programında, ağaç derinliği her iki CHAID algoritması için 3 olmasına karşın, CART için bu derinlik 5 olarak ayarlanmıştır. Ancak, ağaç derinliği değişikliği yapılabilir.

İsimsel Bağımlı Değişken (Nominal Dependent Variable) gözlenen hücre frekansları ve beklenen hücre frekansları Pearson ki-kare istatistiğini veya olasılık oranı istatistiğini hesaplamak için kullanılır. P-değeri bu iki istatistiğin her birine bağlı olarak hesaplanır. Pearson Ki-kare istatistiği ve olasılık oran istatistiği, sırasıyla;

$X^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$ $G^2 = 2\sum_j^J \sum_{i=1}^I n_{ij} \ln(n_{ij}) / \hat{m}_{ij}$ ……...(3.12)

Burada $n_{ij} = \sum_{n \in D} f_n I\,(x_n = i \hat{} y_n = j)$……...……..…..(3.13)

gözlenen hücre frekansı ve $\hat{m}_{ij}$ hücre için olasılık olarak beklenen hücre frekansı $(x_n = i, y_n = j)$…….….....(3.14)
bağımsızlık modelini takip eder. Burada p-değerine karşılık gelen Pearson'un Ki-kare testi için $p = \Pr({x_d}^2 > x^2)$……..…(3.15)
tarafından verilir yada olasılık oran testi için $p = \Pr({x_d}^2 > G^2)$……(3.16)
verilir. Burada ${x_d}^2$ $d = (I-1)(J-1)$özgürlük dereceleri ile ki-kare dağılımını takip eder. Ağırlık durumu olmaksızın beklenen hücre frekansının tahmin edilmesi ise şu şekilde yapılır;

$\hat{m}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$.....(3.17)

Burada $\boldsymbol{n_i = \sum_{j=1}^{J_t} n_{ij},\ n_{.j} = \sum_{i=1}^{I_t} n_{ij},\ n = \sum_{j=1}^{J_t} \sum_{i=1}^{I_t} n_{ij}}$ ……...…(3.18)
durum ağırlıkları belirlenmiş ise, bağımsızlık sıfır hipotezi altında beklenen hücre frekansı oluşturulur. $m_{ij} = \bar{w}_{ij} - 1 a_1 \beta_j$……..….(3.19)
burada a_i ve β_j tahmin etmek için parametrelerdir.

$\bar{w}_{ij} = \frac{w_{ij}}{n_{ij}},\ w_{ij} = \sum_{n \in D} w_n f_n I\,(x = i \wedge y_n = j\,)$……..……...(3.20)

Parametre tahminleri $\hat{\alpha}_i, \hat{\beta}_i$ ve bu yüzden $\hat{m}_{ij}$, aşağıda tekrarlanan yöntemden kaynaklanmıştır.

1. $k = o, a_i^{(0)} = \beta_i^{(0)} = 1, m_{ij}^{(0)} = \bar{w}_{ij}$……..……...(3.21)

2. $a_i^{(k+1)} = \frac{n_{i.}}{\sum_j \bar{w}_{ij} - 1 \beta_j^{(k)}} = a_i^{(k)} \frac{n_{i.}}{\sum_j m_{ij}^{(k)}}$…….……..…(3.22)

3. $\beta_i^{(k+1)} = \frac{n_{.j}}{\sum_i \bar{w}_{ij} - 1 a_i^{(k+1)}}$……..….……...(3.23)

4. $m_{ij}^{(k+1)} = \bar{w}_{ij} - 1 a_i^{(k+1)} \beta_i^{(k+1)}$ ……..…….……...(3.24)

5. eğer $\max_{ij} \left| m_{ij}^{(k+1)} - m_{ij}^{(k)} < \varepsilon, \right|$……..……..(3.25)

durdurma ve çıkış $a_i^{(k+1)}, \beta_i^{(k+1)}$ ve $m_{ij}^{(k+1)}$…….……(3.26)

nihai tahmin olarak kullanılır. Aksi durumda, $k = k + 1$, 2. Adıma gidilir. Sıralı Bağımlı Değişken (Ordinal Dependent Variable); *Y* bağımlı değişken, kategorik sıralı olursa, *X* in ve *Y* nin bağımsızlık hipotezi önemsiz olduğunda, *X* in kategorisi olan satır ile ve *Y* nin sınıfı olan sütun ile satır etkileri modele karşı test edilir. Beklenen hücre frekanslarının iki seti, $\hat{m}_{ij}$(bağımsızlık hipotezi altında) ve $\hat{\hat{m}}_{ij}$ (bir satır etki modeli takip eden hipotez altında) tahmin edilir. Olabilirlik oranı istatistiği ve p-değeri aşağıdaki gibidir.

$H^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{\hat{m}}_{ij} \ln(\hat{\hat{m}}_{ij} / \hat{m}_{ij})$ p=Pr$(x_{I-1}^2 > H^2)$…….……(3.27)

Satır etki modeli altında beklenen hücre frekanslarının tahmin (estimation of expected cell frequencies under row effects model) edilmesi ise; Satır etki modelinde, *Y*'nin sınıflarına ait sayısal verilere ihtiyaç vardır.

$m_{ij} = \bar{w}_{ij} - 1 a_1 \beta_j \gamma_i$ Burada;

$\bar{s} = \sum_{j=1}^{J} w_{.j} s_j I \sum_{j=1}^{J} w_{.j}$ $w_{.j} = \sum_i w_{ij}, a_i, \beta_j$ ve γ_i …….……(3.28)

bilinmeyen parametreleri tahmin etmek için kullanılır.

1.$k = 0, a_i^{(0)} = \beta_j^{(0)} = \gamma_i^{(0)} = 1, m_{ij}^{(0)} = \bar{w}_{ij}$……..……...(3.29)

2.$a_i^{(k+1)} = \frac{n_{.j}}{\sum_j \bar{w}_{ij} - 1 \beta_j^{(k)} (\gamma_i^{(k)})^{(s_j - \bar{s})}} = a_i^{(k)} \frac{n_{i.}}{\sum_j m_{ij}^{(k)}}$……..…(3.30)

3.$\beta_j^{(k+1)} = \frac{n_{.j}}{\sum_i \bar{w}_{ij} - 1 a_i^{(k+1)} (\gamma_i^{(k)})^{(s_j - \bar{s})}}$……..……..…..(3.31)

4.$m_{ij}^{*} = \bar{w}_{ij} - 1 a_i^{(k+1)} \beta_j^{(k+1)} (\gamma_i^{(k)})^{(s_j - \bar{s})},$ $G_i = 1 + \frac{\sum_j (s_j - \bar{s})(n_{ij} - m_{ij}^{*})}{\sum_j (s_j - \bar{s})^2 m_{ij}^{*}}$……(3.32)

5. $\gamma_i^{(k+1)} = \begin{cases} \gamma_i^{(k)} G_i & G_i > 0 \\ \gamma_i^{(k)} \end{cases}$ ……...……...(3.33)

6. $m_{ij}^{(k+1)} = \bar{}_{w_{ij}} - 1a_i^{(k+1)}\beta_j^{(k+1)}(\gamma_i^{(k+1)})(s_j - \bar{}_s)$……....(3.34)

7. $max_{i,j}^{\left|m_{ij}^{(k+1)} - m_{ij}^{(k)}\right| < \varepsilon,}$ ……..…..(3.35)

$a_i^{(k+1)}, \beta_j^{(k+1)}, \gamma_i^{(k+1)}\ ve\ m_{ij}^{(k+1)}$ ……..……..(3.36)

son tahmin olarak kullanılır, aksi halde, $k = k + 1$, 2. adıma gidilir.

3.2.1.c. Çok katmanlı algılayıcı algoritması (MLP Multilayer perceptron algorithms)

Çok katmanlı algılayıcı algoritması, bir ileri besleme sistemi olup. iki gizli katmana kadar ağı denetler. Bir yada daha fazla hedef değişkenin tahmin hatasını en aza indirgeyen bir yöntemdir. Tahminler ve hedefler kategoriksel ve ölçek değişkenlerin bir karışımı olabilir.

$X^{(m)} = (x_1^{(m)}, \ldots, x_P^{(m)})$……..…...(3.37)

Giriş, şekil *m*, *m* = 1, …,*M*.

$Y^{(m)} = (y_1^{(m)}, \ldots, y_R^{(m)}$……..…(3.38)

Hedef, şekil *m*.

I .....Katman sayısını giriş katmana indirgeme.

J_i ……Katmandaki birim sayısı *i*. Birim indirgeme $J_0 = P$, $J_i = R$.

Γ^c ……..Kategorik çıkışın seti.

Γ ……...Ölçek çıktıların seti.

Γ_h ……..$Y^{(m)}$ nin alt vektör seti, 1-*c* nin kodu *h*th kategorik değişken.

$a_{i:j}^m$ ……*i* Katmanın *j* birimi, şekil *m*, = 0, …, J_i;*i* = 0, …, *I*. $W_{i:j,k}$, *i*-1 katmanından giden ağırlık ile *i* katmanı için *j* birimi ve *k* birimi $a_{i-1:j}^m$ bağlantılı ağırlıklar değil ve herhangi bir *j*, için $w_{i:j,0}$ yok.

$\sum_{j=0}^{Ji-1} wi{:}\, j, k a_{i-1:j}^m$, *i*=1, …,*I*, γ_i (*c*) *i* ……..…(3.39)

katmanı için etkinleştirme işlevi. W bütün ağırlıkları içeren ağırlık vektörü ($w_{1:0,\ 1,}\ w_{1:0,\ 2,}\ w_{ı:Jı\text{-}1,\ Jı}$). Mimari (Architecture); çok katmanlı algılayıcı için genel mimari; giriş katmanı; $J = P$ birimi, $a_{0:1}, \ldots, a_{0:J0}$; ile $a_{0:j} = x_j$. Gizli katmanı; J_i birimleri, $a_{i:1}, \ldots, a_{i:Ji}$; ile $a_{i:k} = \gamma_i (c_{i:k})$ ve $c_{i:k} = \sum_{i=0}^{Ji-1} w_{i:j,k} a_{i-\ 1:j}$……..(3.40)

burada $a_{i\text{-}1:0} = 1$. Çıkış katmanı: $J_ı = R$ birimleri;

$a_{ı:1}, \ldots, a_{ı:Jı}$; ile $a_{ı:k} = \gamma ı (c_{ı:k})$……..…(3.41)

ve $c_{ı:k} = \sum_{j=0}^{J1} w_{ı:j,k}\ a_{i\text{-}1:j}$ burada $a_{i\text{-}1:} = 1$…….……(3.42)

3.2.1.d. Nive Bayes algoritması

Naive Bayes sınıflandırma algoritması, adını Matematikçi Thomas Bayes'den alan bir sınıflandırma ve kategorilendirme algoritmasıdır. Naive Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış bir dizi hesaplama ile, sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar.

Naıve Bayes modeli, basitliliğinden ve istikrarlı hesaplamalar yapabildiğinden dolayı sınıflandırma ve olasılık seçiminde başarısından dolayı tercih edilen eski bir yöntemdir.

J_0 ……Olasılıkların toplam sayısı.

X ……Kategorik olasılık vektörü X' = ($X_1, \ldots, X_J$), buradaki J tahmin edilen olasılıkların sayısı.

M_j ……...X_j olasılığı için sınıflandırma sayısı.

Y ……....Kategorik hedef değişkeni.

K ……....Y nin kategori sayısı.

N ……Model ve durumların toplam sayısı.

N_k ……...Y=k ile olan durumların sayısı.

N^j_{mk} …….Y=k ve X_j=m ile olan durumların sayısı.

π_k …….Y=k nın olasılığı.

P^j_{mk} …….Y=k da verilen X_j nin olasılığı.

Naive Bayes modeli, hedef sınıf için verilen her bir olasılığın koşula bağlı olarak bağımsızlık modeline dayalı bir modeldir. Bayesian ilkesi, sonraki en geniş olasılık durum atamasıdır. Bayes teorisi, verilen X, Y nin sonraki atamasıdır.

$$P\ (\ Y = k|\mathrm{X} = \mathrm{x}) = \frac{P(\mathrm{X}=\mathrm{x}|Y=k)P(Y=k)}{\sum_{i=1}^{K} P(\mathrm{X}=\mathrm{x}|Y=i)P(Y=i)} \quad \text{.....(3.43)}$$

Modelde J olasılığı olarak görülen $X_{1, \ldots,}X_J$ dir. Naıve Bayes modeli, verilen hedef için koşula bağlı olarak bağımsız değişkenin $X_1, \ldots, X_J$ olduğunu varsayar.

$$P\ (\ \mathrm{X} = \mathrm{x}|Y = k\ ) = \prod_{j=1}^{J} P\ (\ X_j = x_j|Y = k\ ) \quad \text{.....(3.44)}$$

Bu olasılıklar, aşağıdaki denklemlerle hesaplanarak deneme verisinden tahmin edilir:

$$\pi_k = P\ (\ Y = k\ ) = \frac{Nk+\lambda}{N+K\lambda} \quad \text{.....(3.45)}$$

$$P^{j}_{mk} = P\ (X_j = m|Y = k) = \frac{N^{j}_{mk+f}}{\sum_{\iota=1}^{Mj} N^{j}_{\iota k+Mjf}} \quad \text{.....(3.46)}$$

Buradaki N_k, eksik olmayan bir şekilde Y sayısına dayalı hesaplanmıştır. Nj_{mk} Xj ve Y nin eksik olmayan tüm çiftlerine dayalı hesaplanır. λ ve f sıfır yada çok küçük hücre hesaplamalarında ortaya çıkan problemlerin üstesinden gelmek için kullanılır. Bu tahminler

$$\lambda = f = \frac{1}{N} \quad \text{.....(3.47)}$$

Veri geçişi için, ilgili tüm sayıları toplamaya ihtiyaç vardır. Burada J = 0 özel bir durumdur;

$$P(Y = k|\mathrm{X} = \mathrm{x}) = P(Y = k) \quad \text{.....(3.48)}$$

3.2.1.e. Geçerlilik ölçüleri

Kalitatif özelliklerin testlerle değerlendirilmelerinde denekler çoğunlukla; yeterli-yetersiz, başarılı-başarısız, geçer-geçmez, hasta-sağlam, döl tutan-tutmayan, tekiz-ikiz ve erkek-dişi vb. şeklinde değerlendirmek mümkündür. Bu tip bir testin geçerliğinin istatistiksel olarak belirlenmesi işlemi, testin "*karar*"geçerliği süreci ile alakalıdır. Bu durumda geliştirilen test, denekleri; başarılı-başarısız, yeterli-yetersiz, vb. olarak doğru sınıflama özelliğine sahip olmasıdır. Tarafımızdan hayvancılıkta farklı stratejiler geliştirmek amacıyla hayvancılığa ait veriler için en uygun veri madenciliği algoritmasının belirlenmesinde, kalitatif testlerin geçerliğini bulmak için, Alpar (2012) tarafından tanımlanan yöntemler kullanılacaktır.

Çizelge 3.1. Geçerlik ölçülerine ilişkin geçerlik katsayısı

Geliştirilen test sonucu	Gerçek Durum		Toplam
	Başarılı	Başarısız	
Başarılı	A(GB)	B(YP)	A+B
Başarısız	C(YN)	D(GBZ)	C+D
Toplam	A+C	B+D	A+B+C+D

A: Gerçekte başarılı olup, kullanılan algoritmaya göre de başarılı olarak belirlenenlerin sayısını verir ve Gerçek Başarlılar-GB olarak adlandırılır.

D: Gerçekten başarısız olup, kullanılan algoritmada da başarısız olanların sayısını verir ve gerçek başarısızlar-GBZ olarak adlandırılır.

C: Gerçekten başarılı olup, kullanılan algoritmanın yanlışlıkla başarısız dediği gözlem sayısıdır ve yanlış negatif-YN olarak adlandırılır.

B: Gerçekten başarısız olup kullanılan algoritmanın yanlışlıkla başarılı dediği gözlemlerin sayısıdır ve yanlış pozitif YP olarak adlandırılır.

Doğru sınıflama oranı (Accuracy rate); geçerlik katsayısı ya da doğru sınıflama oranı (*c*), her iki durumda da uyumlu olan gözlerdeki (A ve D) sıklıkların toplam gözlem sayısına (A+B+C+D) bölünmesi ile bulunur. Doğru sınıflama oranı yüksek olması, algoritmalar ile yapılacak sınıflamanın güvenilir olduğu anlamına gelmektedir.

$$C=\frac{A+D}{A+B+C+D} \quad \text{.....(3.49)}$$

Doğru sınıflama yüzdesi olan *c* değeri 0,50 ve 0,50'nin altında ise, kullanılan algoritma ile yapılan sınıflandırmanın şans eseri olduğu yorumu yapılır. *C* değerinin 1,0'a yakın olması istenir. Ancak herhangi bir algoritmanın tam olarak doğru bir sınıflama yaptığını söylemek için duyarlılık, seçicilik ve ROC eğrisi altında kalan alan ölçütlerinin de baz alınması gerekmektedir.

3.2.1.f. Duyarlılık (sensitivity) ve özgüllük (spesifity) ölçütleri

Kullanılacak algoritmaların performansı (başarısı) ile ilgili olarak amaca uygun bazı oranlarda tanımlanabilir. Bu oranlar aşağıda özetlenmiştir.

A/(A+C): Gerçekten başarılı olanların % kaçının kullanılan algoritma sonucunda başarılı bulunduğunu gösterir ve " gerçek başarılılar içinde kullanılan algoritmanın başarılıları ayırt edebilme yeteneği " olarak tanımlanır.

D/(B+D): Gerçekten başarısız olanların % kaçının kullanılan algoritma sonucunda başarısız olduğunu gösterir ve "gerçek başarısızlar içinde, kullanılan algoritmanın başarısızları ayırt edebilme yeteneği" olarak tanımlanır.

C/(A+C): Yanlış negatif oranı olup, gerçek sınıflama sonucunda başarılı olanlar arasından kullanılan algoritma sonucunda yanlışlıkla başarısız olarak sınıflandırma yüzdesidir.

B/(B+D): Yanlış pozitif oranı olup, gerçek sınıflandırma sonucunda başarısız olanlar arasından kullanılan algoritma sonucunda yanlışlıkla başarılı olarak nitelendirilenlerin yüzdesidir.

Geçerliğe ilişkin iki önemli ölçüt A/(A+C) ve D/(B+D) oranları olup, ideal bir testte bu oranların 1 ya da 1'e yakın olması arzu edilir. Bu açıklamaya göre, A/(A+C)' duyarlılık; D/(B+D)'ye ise seçicilik denir ve **Duyarlık**=A/(A+C), **Seçicilik**=D/(B+D) dir. Bu bağlamda;

A/(A+B) : Kullanılan algoritma sonucunda başarılı olanların gerçekten başarılı olma olasılığını verir ve **pozitif kestirim değeri** adını alır.

D/(C+D) Kullanılan algoritma sonucunda başarısız olanların gerçekten başarısız olma olasılığını verir ve **negatif kestirim değeri** adını alır.

3.2.1.g. ROC eğrisi yaklaşımı

ROC eğrisi, tanı testinin kendi doğruluğunu tanımlaması ve testler arasında güvenilir bir karşılaştırma yapmaya olanak sağlaması açısından sıklıkla kullanılmaktadır (Alpar, 2010).

ROC eğrisinin grafiksel yaklaşımı ölçümlerin duyarlılığı ve özgüllüğü arasındaki ilişkilerin daha rahat kavranmasını sağlar. Birbirinden bağımsız iki grup (başarılı-başarısız ya da hasta-sağlam), sonucu sayısal veri türünden elde edilen bir test yardımıyla bilmek istendiğinde genelde **ROC** eğrisinden yararlanılır. Hayvancılıkla ilgili verilerin değerlendirilmesinde de **ROC** eğrisinden faydalanılmaktadır.

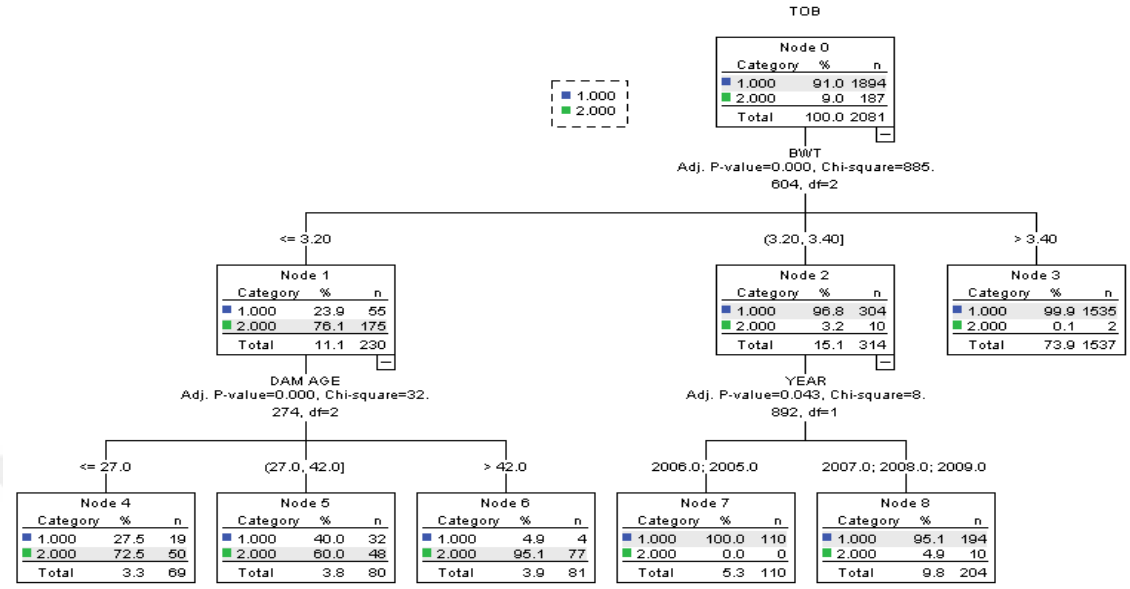
ROC eğrisi çizebilmek için bireylerin gerçek durumunun standart olarak adlandırılan bir test ile kesin olarak belirlenmesi gerekir. Sonuçları sürekli sayısal veri türünde olduğunda, pozitif ya da negatif olarak ayırmak için belirlenen kesim noktasına bağlı olarak başarılı-başarısız olma oranı değişir. Bu nedenle, **ROC** eğrisini çizmek için elde edilen tüm değerler kesim noktası olarak alınıp, her değer için doğru pozitif (geçerlilik) ve doğru negatif (seçicilik) değerleri hesaplanır. **ROC** eğrisinin dikey ekseninde doğru pozitif (duyarlılık) ve yatay ekseninde de yanlış pozitif (geçerlilik) değerleri yer alır. Sayısal verideki her kesim noktası için elde edilen duyarlılık ve 1-özgüllük değerleri kullanılarak **ROC** eğrisi çizilir.

4. BULGULAR ve TARTIŞMA

CART, CHAID, Exhaustıve CHAID, Naive Bayes, MLP ve C5.0 algoritmaları ile bu algoritmaların karşılaştırılmalı performansları ve yapılan istatistiki değerlendirmeler Çizelge 4.1 - 4.11 de gösterilmiştir.

4.1. CHAID Algoritmasına İlişkin Bulgular

CHAID algoritması ile oluşturulan ağaç diyagramı Şekil 4.1 de gösterilmiştir. Bu diyagrama göre; Düğüm 1de, Doğum ağırlığı 3.20 kg ve daha hafif olan kuzuların ikiz olma ihtimali % 76.1olarak tahmin edilmiştir. Düğüm 2 de, 3.20 kg < doğum ağırlığı ≤ 3.40 kg olan kuzuların tekiz olma ihtimali % 96.8 olarak tahmin edilmiştir. Düğüm 3, doğum ağırlığı 3.40 kg dan ağır olan kuzuların tekiz olma ihtimali % 99.9 olarak bulunmuştur. Düğüm 4, Doğum ağırlığı, ≤ 3.20 kg ve ana yaşı ≤ 27 ay olan kuzuların ikiz olma olasılığı % 72.5 olduğu tespit edilmiştir. Düğüm 5, doğum ağırlığı ≤ 3.20 kg ve ana yaşı 27 < ay ≤ 42 ay olan anaların ikiz doğurma olasılığı % 60 olarak bulunmuştur. Düğüm 6, doğum ağırlığı ≤ 3.20 kg olan 42 ay dan daha yaşlı anaların ikiz kuzu doğurma olasılığı % 95.1 olarak tespit edilmiştir. Düğüm 7, doğum ağırlığı 3.20 < ana yaşı ≤ 3.40 kg olan ve 2005 ve 2006 yıllarında doğan kuzuların tamamı % 100'ü tekiz olarak bulunmuştur. Düğüm 8 de, 2007, 2008 ve 2009 yıllarında 3.2 < Doğum ağırlığı < 3.4 kg arasında kuzu doğuran koyunların tekiz kuzu doğurma ihtimali % 95.1olduğu belirlenmiştir.



Şekil 4.1. CHAID algoritmasına ait ağaç diyagramı

Çizelge 4.1. CHAID algoritmasına ilişkin model özeti

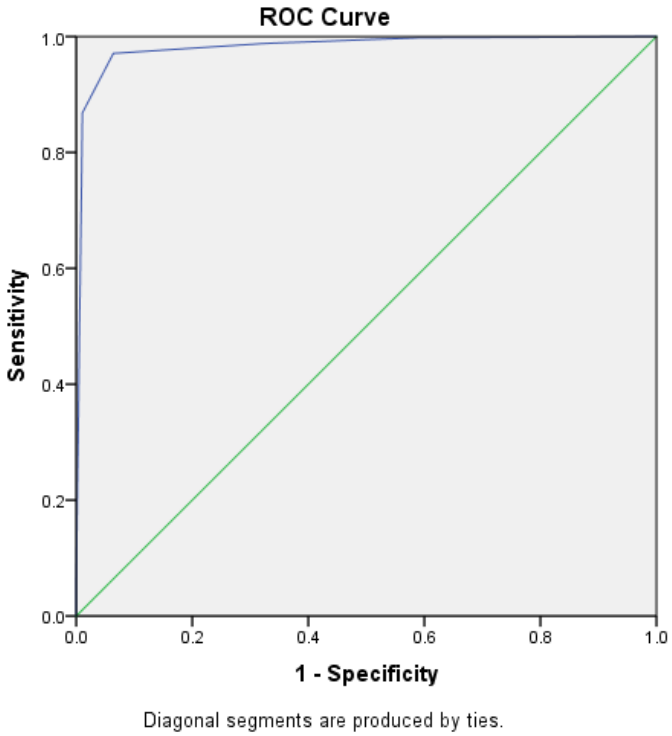
Özellikler	İlerleme metodu	CHAID
	Bağımlı değişken	Doğum tipi
	Bağımsız değişkenler	Yıl, Yer, Ana yaşı, Ana ağırlığı, Doğum mevsimi, Cinsiyet, Doğum ağırlığı, Sütten kesim ağırlığı
	Geçerlilik	Çapraz geçerlilik
	Maksimum ağaç derinliği	3
	Ebeveyn düğümdeki minimum durum	100
	Çocuk düğümdeki minimum durum	50
Sonuçlar	Dahil edilen bağımsız değişkenler	Doğum ağırlığı, Ana yaşı, Yıl
	Düğüm sayısı	9
	Terminal düğüm sayısı	6
	Derinlik	2

Çizelge 4.2. CHAID algoritmasına ait tahmin ve standart hata

Metot	Tahmin	Std. Hata
Resubstitution	0.032	0.004
Çapraz geçerlilik	0.032	0.004

Çizelge 4.3. CHAID algoritmasına ait sınıflandırma

	Tahmin edilen		
Gözlemlenen	Tekiz	İkiz	Yüzdelik oran (%)
Tekiz	1839	55	97.1
İkiz	12	175	93.6
Genel yüzde	88.9	11.1	96.8

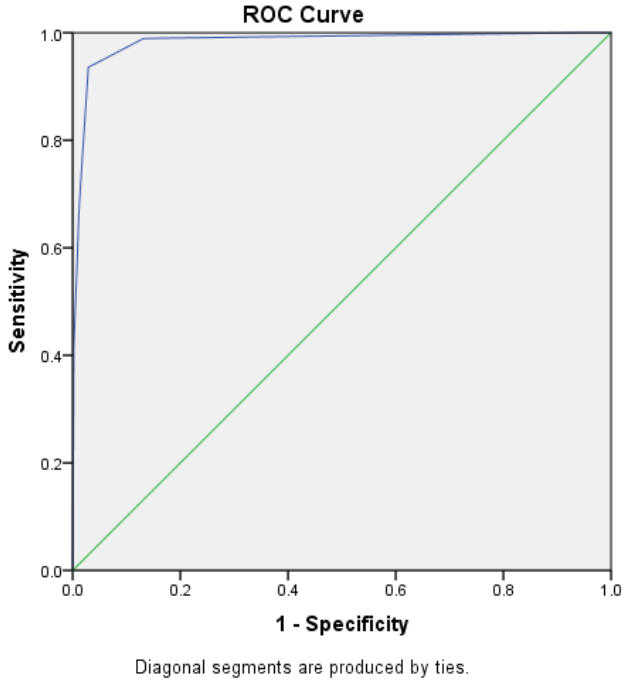


Şekil 4.2 CHAID algoritmasına ait ROC eğrisi (tekiz)

Çizelge 4.4. CHAID algoritmasına ait ROC eğrisi sonuçları (tekiz)

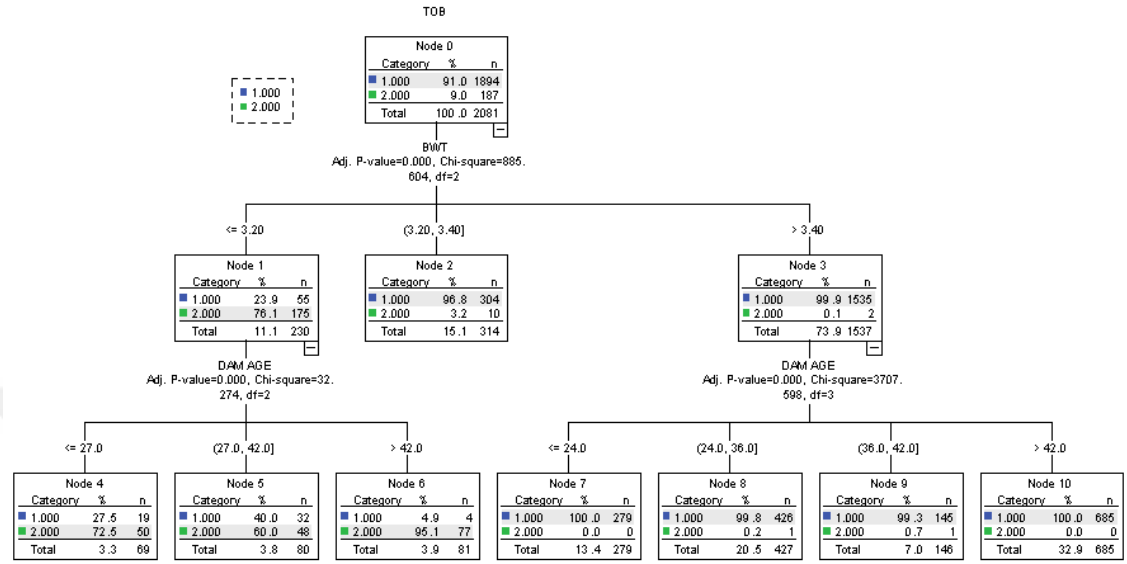
Eşit yada büyük ise pozitif	Duyarlılık	1-Özgüllük	Özgüllük	A+B
0,9506	1,000	1,000	0,000	1,000
0,1624	0,998	0,588	0,410	1,408
0,3377	0,988	0,321	0,667	1,655
0,6755	**0,971**	**0,064**	**0,907**	**1,878**
0,9748	0,869	0,011	0,858	1,727
0,9993	0,058	0,000	0,058	0,116
20.000	0,000	0,000	0,000	0,000

Çizelge 4.4 te gösterildiği gibi, CHAID algoritmasının ROC eğrisi sonuçlarına göre, tekiz olma olasılığı 0.6755 ya da daha büyük olan kuzuların tekiz, bu değerden düşük olasılığa sahip kuzuların ise ikiz olacağı söylenebilir.



Şekil 4.3 CHAID algoritmasına ait ROC eğrisi (ikiz)

4.2. Geniş (Exhaustive) CHAID Algoritmasına İlişkin Bulgular



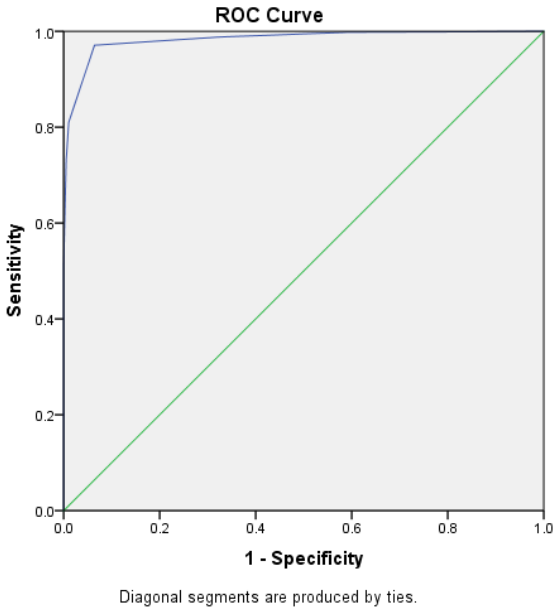
Şekil 4.4 Geniş (exhaustive) CHAID algoritmasına ait ağaç diyagramı

Çizelge 4.5. Geniş (exhaustive) CHAID algoritmasına ait model özeti

Özellikler	İlerleme metodu	EXHAUSTIVE CHAID
	Bağımlı değişken	Doğum tipi
	Bağımsız değişkenler	Yıl, Yer, Ana yaşı, Ana ağırlığı, Doğum mevsimi, Cinsiyet, Doğum ağırlığı, Sütten kesim ağırlığı
	Geçerlilik	Çapraz Geçerlilik
	Maksimum ağaç derinliği	3
	Ebeveyn düğümdeki minimum durum	100
	Çocuk düğümdeki minimum durum	50
Sonuçlar	Dahil edilen bağımsız değişkenler	Doğum ağırlığı, Ana yaşı
	Düğüm sayısı	11
	Terminal düğüm sayısı	8
	Derinlik	2

Çizelge 4.6. Geniş (exhaustive) CHAID algoritmasına ait tahmin ve standart hata

Metot	Tahmin	Std. Hata
Resubstitution	0.032	0.004
Çapraz geçerlilik	0.032	0.004



Şekil 4.5 Geniş (exhaustive) CHAID algoritmasına ait ROC eğris (tekiz)

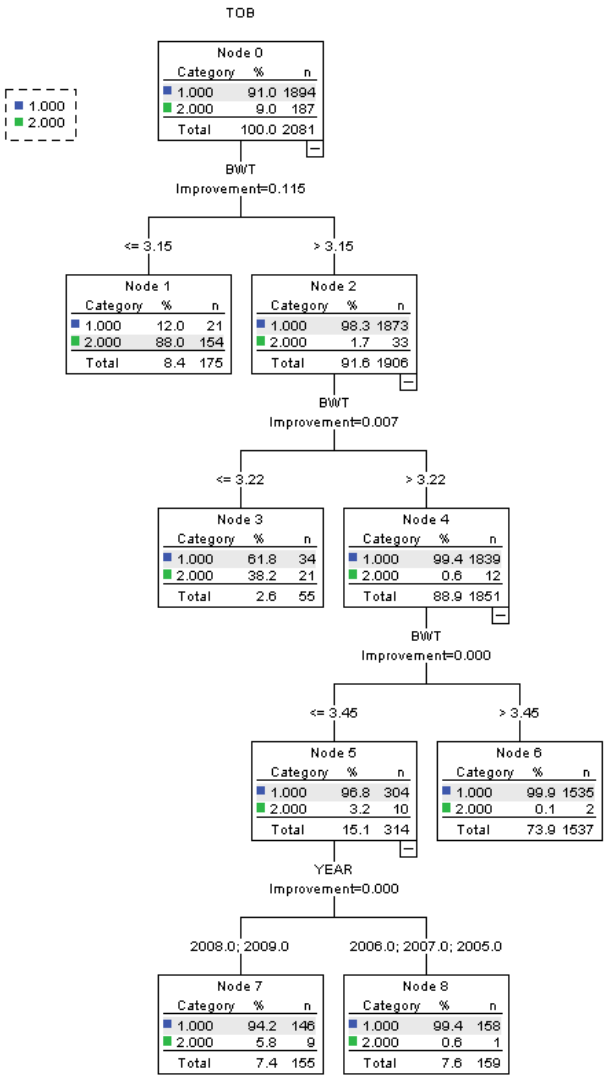
Çizelge 4.7. Geniş (exhaustive) CHAID algoritmasına ait ROC eğrisi sonuçları (tekiz)

Eşit yada büyük ise pozitif	**Duyarlılık**	**1-Özgüllük**	**Özgüllük**	**A+B**
0.9506	1	1	0.000	1.000
0.1624	0.998	0.588	0.410	1.408
0.3377	0.988	0.321	0.667	1.655
0.6841	**0.971**	**0.064**	**0.907**	**1.878**
0.9807	0.81	0.011	0.799	1.609
0.9954	0.734	0.005	0.729	1.463
0.9988	0.509	0	0.509	1.018
20.000	0	0	0.000	0.000

Çizelge 4.7 incelendiğinde, Geniş (exhaustive) CHAID algoritmasına ait ROC eğrisi sonuçlarına göre, tekiz olma olasılığı 0.6841 ya da daha büyük olan kuzuların tekiz, bu değerden düşük olasılığa sahip kuzuların ise ikiz olacağı söylenebilir.

4.3. CART Algoritmasına İlişkin Bulgular

CART algoritmasına ilişkin ağaç diyagramı Şekil 4.6 da gösterilmiştir.



Şekil 4.6 CART algoritmasına ait ağaç diyagramı

Çizelge 4.8. CART algoritmasına ait model özeti

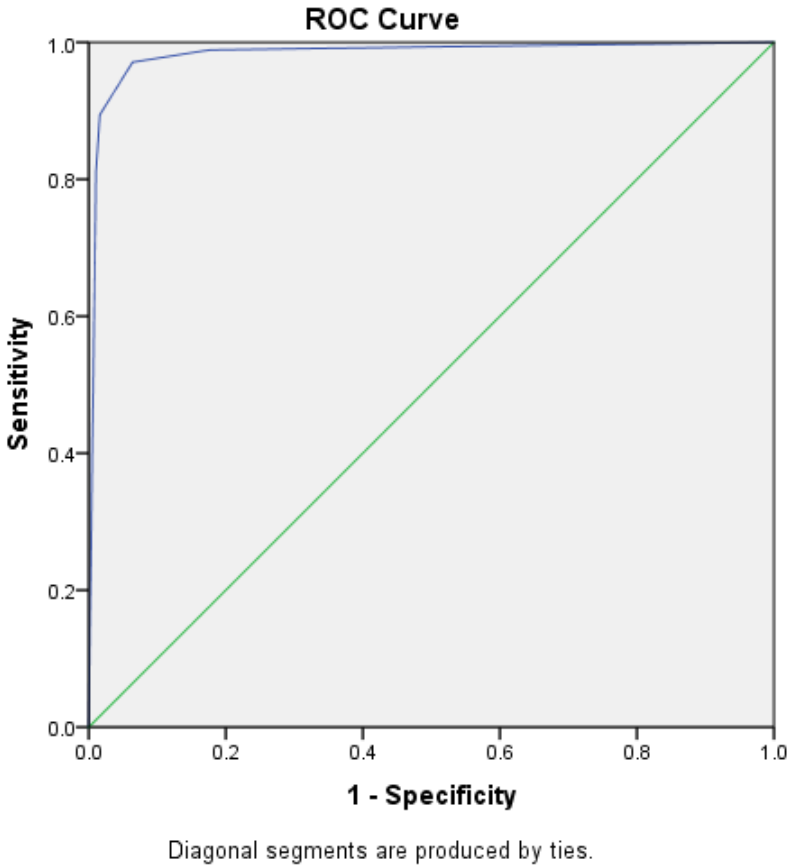
Özellikler	İlerleme metodu	CRT
	Bağımlı değişken	Doğum tipi
	Bağımsız değişkenler	Yıl, Yer, Ana yaşı, Ana ağırlığı, Doğum mevsimi, Cinsiyet, Doğum ağırlığı, Sütten kesim ağırlığı
	Geçerlilik	Çapraz geçerlilik
	Maksimum ağaç derinliği	5
	Ebeveyn düğümdeki minimum durum	100
	Çocuk düğümdeki minimum durum	50
Sonuçlar	Dahil edilen bağımsız değişkenler	Doğum ağırlığı, Sütten kesim ağırlığı, Ana yaşı, Ana ağırlığı, Yıl, Yer, Doğum mevsimi, Cinsiyet
	Düğüm sayısı	9
	Terminal düğüm sayısı	5
	Derinlik	4

Çizelge 4.9. CART algoritmasına ait tahmin ve standart hata

Metot	Tahmin	Std. Hata
Resubstitution	0.026	0.003
Çapraz geçerlilik	0.026	0.003

Çizelge 4.10. CART algoritmasına ait sınıflandırma

	Tahmin edilen		
Gözlemlenen	Tekiz	İkiz	Yüzdelik oran (%)
Tekiz	1873	21	98.9
İkiz	33	154	82.4
Genel yüzde	91.6	8.4	97.4



Şekil 4.7 CART algoritmasına ait ROC eğrisi (tekiz)

Çizelge 4.11. CART algoritmasına ait ROC eğrisi sonuçları (tekiz)

Eşit yada büyük ise pozitif	**Duyarlılık**	**1-Özgüllük**	**Özgüllük**	**A+B**
0.0000	1.000	1.000	0.000	1.000
0.3691	0.989	0.176	0.813	1.802
0.7801	**0.971**	**0.064**	**0.907**	**1.878**
0.9678	0.894	0.016	0.878	1.772
0.9962	0.810	0.011	0.799	1.609
10.000	0.000	0.000	0.000	0.000

Çizelge 4.11 sonuçlarına baktığımızda, ROC eğrisi sonuçlarına göre, tekiz olma olasılığı 0.7801 ya da daha büyük olan kuzuların tekiz, bu değerden düşük olasılığa sahip kuzuların ise ikiz olacağı söylenebilir.

4.4. Naive Bayes Algoritmasına İlişkin Bulgular

Naıve Bayes algoritmasına göre tekizlik ve ikizlik durumuna ilişkin elde edilen duyarlılık değeri % 98 olarak bulunmuştur. özgüllük ise % 87 olarak bulunmuştur.

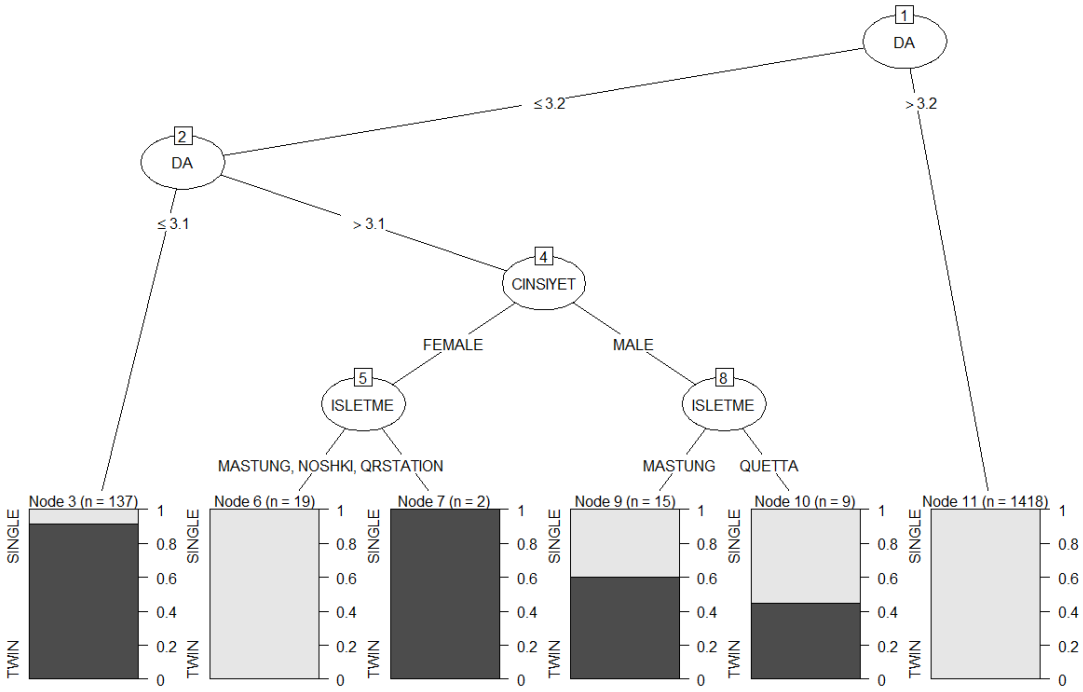
4.5. MLP Algoritmasına İlişkin bulgular

Yapay sinir ağları Training set tekiz oranı % 98.07, tekiz sensitivity % 98.7, ikiz spesifity % 91.43. Testing doğruluk oranı % 97.3, sensitivity % 98.63, spesifity % 82.05 olarak sonuçlanmıştır.

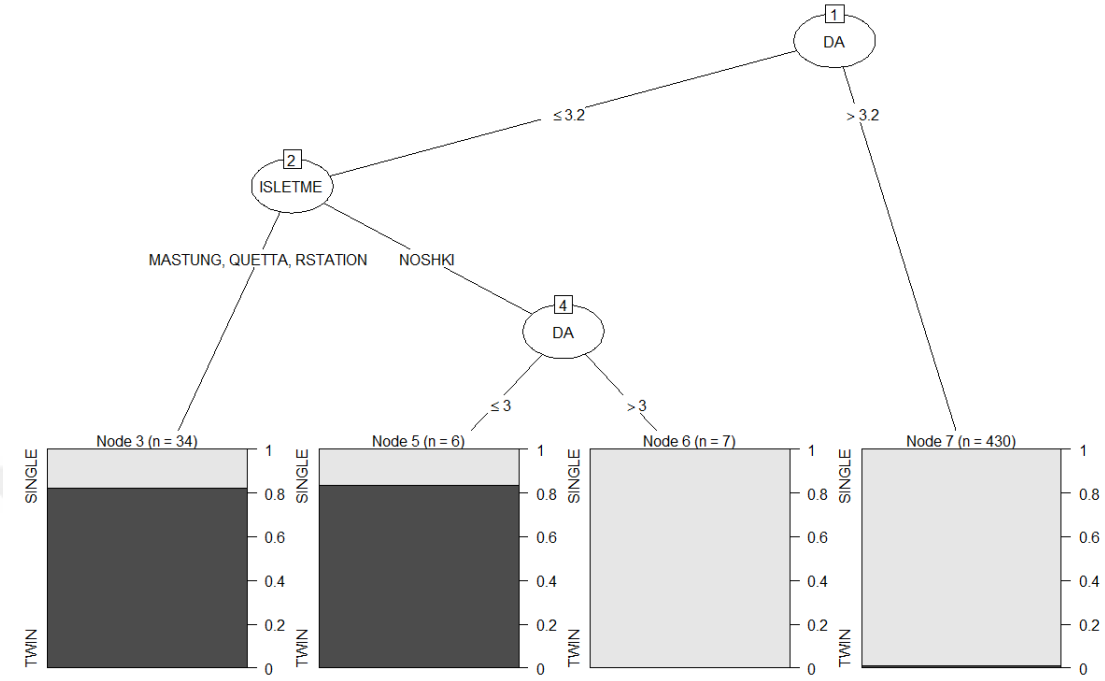
4.6. C5.0 Algoritmasına İlişkin Bulgular

C5.0 algoritmasına ait sınıflama ağaçları Şekil 4.8. ve 4.9. da gösterilmiştir.

C5.0 algoritması için sensitivity % 98.41, spesifity %87.0, genel doğru sınıflama oranının %97.5 olarak bulunmuştur.



Şekil 4.8. C5.0 algoritmasına ait ağaç diyagramı



Şekil 4.9. C5.0 Algoritmasına ait ağaç diyagramı

Sut ve Şimşek (2011) kazalarda kafa yaralanması sonucu oluşan ölüm oranını tahmin etmek amacıyla sınıflama performansı bakımından farklı karar ağacı algoritmalarını (CART, CHAID, Exhaustive-CHAID) birbiriyle karşılaştırmış. İncelenen algoritmaların performansları, duyarlılık oranı (sensitivity), özgüllük oranı (specificity), pozitif/negatif tahmin oranı (positive/negative predictive) ve isabet oranı (accuracy rate) ölçütleri kullanarak (0.801 ile 0.954) bulmuşlar. Ayrıca, tüm algoritmalara ait ROC eğrisi altında kalan alanları hesaplamışlar. (P<0.001). ROC eğrisinin altında kalan en küçük alana sahip algoritmanın CART (0.801) olduğu ve bu algoritma için isabet oranının % 91.1 olduğunu saptamıştır. Bizim çalışmamızın ROC eğrisi sonucu ise genel yüzdelik % 91.6 olarak gözlemlenmiştir. Tekizlik oranı ise % 98. 9 bulunmuştur. Bizim bulduğumuz doğru sınıflama oranı sut ve şimşekten yüksek bulunmuştur.

5. SONUÇ ve ÖNERİLER

Bu çalışmada tarım işletmesinde yetiştirilen hayvanlardan elde edilen ve kesikli varyasyon gösteren doğum tipi (tekiz ve ikiz) bakımından , CART, CHAID, Exhaustive CHAID, Naive Bayes, yapay sinir ağları, C5.0 algoritması gibi sınıflandırma algoritmalarının performansları karşılaştırılmıştır. Tekizlik referans alındığında elde edilen analiz sonuçlarına göre;

1. CART sınıflama algoritması için sensitivity % 98.9, spesifity % 82.4, genel doğru sınıflama oranının %97.4 olduğu,
2. CHAID ve Exhaustive CHAID sınıflama algoritmaları için sensitivity % 97.1, spesifity % 93.6, genel doğru sınıflama oranının %96.8 olduğu,
3. Multilayer perceptron YSA sınıflama algoritması için sensitivity training set için % 98.7 ve testing set %98.63, spesifity training set için % 91.43 ve testing set için %91.43, genel doğru sınıflama oranının ise training set için %98.07 ve testing set için 97.3 olduğu,
4. Naive Bayes sınıflama algoritması için sensitivity %98, spesifity %87 ve genel sınıflama oranının %97 olduğu,
5. C5.0 algoritması için duyarlılık % 98.41, özgüllük %87.0, genel doğru sınıflama oranının %97.5 olduğu belirlenmiştir

Çalışmadan elde edilen bu sonuçlara göre kullanılan bütün algoritmalarının sınıflama performanslarının oldukça iyi olduğu söylenebilir. Buna karşın, büyük veri setleri için CART, CHAID, Exhaustive CHAID ve C5.0 gibi sınıflama ağacı üreten algoritmaların kullanılması önerilebilir. CART, CHAID, Exhaustive CHAID algoritmalarına ilişkin en iyi sınıflama performansı elde etmek için ebeveyn ve çocuk düğümünün oranları dikkatli bir şekilde ayarlanmalıdır. İlk etapta başlangıç olarak ebeveyn düğüm sayısının N/6, çocuk düğümünün ise N/12 olarak alınması önerilebilir. CART algoritması aşırı dallanan bir algoritma olduğu için elde edilen sınıflama ağacının daha kolay ve etkili yorumlanması bakımından ebeveyn ve çocuk düğüm oranının çok dikkatli ayarlanması gerekebilir.

KAYNAKLAR

Adamczyk, K., Zaborski, D., Grzesiak, W., Makulska, J., Jagusiak, W., 2016. Recognition of culling reasons in polish dairy cow using data mining methods. ***Computers and Electronics in Agriculture. Elsevier***. 127, 26–37.

Akaike, H., 1974. A new look at the statistical model identification. ***Statistical Model Identification IEEE Transactions on Automatic Control*** AC-19, 716–723.

Alpar, R., 2000. Sağlık, spor ve eğitim bilimlerinde örneklerle Uygulamalı istatistik ve Geçerlilik-Güvenirlik. ***Detay yayıncılık***, Ankara.

Alkan, A., Falay, E. 2007. Kamu Uygulamalarında Çözüm Veri Madenciliğinde. ***Strateji Bülteni Dergisi.*** Eylül-Ekim. Sayı 5. 6-8.

Ayık, Y. Z., Özdemir, A., Yavuz, U., 2007. Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle ilişkisinin Veri Madenciliği Tekniği İle Analizi**.** ***Sosyal Bilimler Enstitüsü Dergisi.*** 10(2), s444-446.

Albayrak, A. S., Yılmaz, K., 2009. Veri Madenciliği: Karar Ağacı Algoritmaları Ve İMKB Verileri Üzerine Bir Uygulama***. Veri Madenciligi: Karar Ağacı Algoritmaları ve MKB Verileri***, 14(1) s31-52.

Bayram, B., Topal, M., Aksakal, V., 2015. Investigate the Effects of Non-genetic Factors on Calving Difficulty and Stillbirth Rate in Holstein Friesian Cattle Using the CHAID Analysis. ***Kafkas Universitesi Veteriner Fakultesi Dergisi.***21(5), s645-652

Biggs, D., B. de Ville, ve E. Suen. 1991. A method Of Choosing multiway Partitions For Classification And Decision Trees. ***Journal of Applied Statistics***,18, 49-62.

Caraviello, D.Z., Weigel, K.A., Craven, M., Gianola, D., Cook, N.B., Nordlund, K.V., Fricke,P.M., Wiltbank, M.C., 2006. Analysis of Reproductive Performance of Lactating Cows on larce Dairy Farms Using Machine Learning Algorithms***. J. Dairy Sci.*** 89.4703-4722

Coşkun, C., Baykal, A., 2011. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. ***Akademik Bilişim'11-XIII. Akademik Bilişim Konferansı Bildirileri 2-4 şubat İnönü Üniversitesi***, s51-54

Çamdeviren H., Mendeş, M., Ozkan M., Toros F., gaçmaz T. ve Oner S., 2005. ***Determination of depression risk factors in children and adolescents by regression tree methodology***. Acta Med. Okayama 59(1):19-26.

Çetin, F. A., Mikail, N., 2016. Hayvancılıkta Veri Madenciliği Uygulamaları. ***Türkiye Tarımsal Araştırmalar Dergisi.*** 3(1) 79-88.

Çalış, A., Kayapınar, S., Çetinyokuş, T., 2014. Veri Madenciliğinde Karar Ağacı Algoritmaları İle Bilgisayar Ve İnternet Güvenliği Üzerine Bir Uygulama. ***Endüstri Mühendisliði Dergisi***. 25(3-4) s 2-19.

Eyduran, E., Tatlıyer, A., Tarıq, M. M., Waheed, A., 2013., Antalya Application Of Classification And Regression Tree Methods In Agriculture. ***Ulusal Tarım Kongresi***

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. ***From datamining to discovery knowledge in databases. AI Magazine*** 3(17):37-54.

Glymour, C., Madigan, D., Pregibon, D., Smyth, P., 1997. Statistical Themes and Lessons for Data Mining. ***Data Mining and Knowledge Discovery1***, s11-28.

Goodman, L. A., 1979. Simple Models For The Analysis Of Association İn Cross-Classifications Having Ordered Categories. ***Journal of the American Statistical Association,*** 74, 537-552.

Grzesiak, W., Zaborski, D., 2012. Examples of the Use of Data Mining Methods in Animal Breeding. ***Data Mining Applications in Engineering and Medicine http://dx.doi.org/10.5772/50893***

Grzesiak, W., Zaborski, D., Sablik, P., Pilarczyk, R., 2011. Detection of Difficult Conceptions İn Dairy Cows Using Selected Data Mining Methods. ***İnstitute of Genetics and Animal Breeding, jastrzebiec, Poland.***

Animal Science Papers and reports vol. 29(4), 293-302.

Grzesiak, W., Zaborski, D., Sablik, P., Zukiewicz, A., Dybus A., Szatkowska, I., 2010. Detection of cows with insemination problems using selected classification models. ***Computers and Electronics in Agriculture*** 74.265–273.

Hanley, J.A., McNeil, B.J., 1982. ***The meaning and use of the area under a receiver operating characteristic*** (ROC) curve. ***Radiology*** 143, 29–36.

Hush, D., Horne, B., 1993. Progress in supervised neural networks. ***IEEE Signal Processing Magazine*** 10, 8–39.

Jing L. 2002. Data mining and its applications in higher education. ***New Directions For Institutional Research*** 113, 17-36.

Karabağ, K., Alkan. S., Mendeş, M.(2010). Kınalı Keklik (Alectoris chukar) Yumurtalarında Çıkış Gücüne Etki Eden Faktörlerin Sınıflandırma Ağacı Yöntemi ile Belirlenmesi. ***Kafkas Univ. Vet. Fak. Dergisi*** *16*(5): 723-727.

Larose, T. L., 2006. Discovering Knowledge in Data An ***Introduction to Data Mining***. ***PWN, Warszawa***, Polish edition. ***Wıley İnterscience***, s11.

Liddle, A.R., 2008. Information criteria for astrophysical model selection. ***Monthly Notices of the Royal Astronomical Society:*** Letters 377, L74–L78.

Lisboa, P.J.G., Ifeachor, E.C., Szczepaniak, P.S., 2000. Classifying spinal measurements using a radial basis function network. In: ***Artificial Neural Networks in Biomedicine***. Springer-Verlag, London, pp. 93–104.

Marciniak, A., Korbicz, J., Ku's, J., 2000. Data preprocessing. In: Duch, W., Korbicz, J.,Rutkowski, L., Tadeusiewicz, R. (Eds.), ***Biocybernetics and Biomedical Engineering, Neural Networks. AOW Exit, Warszawa***, p. 6, 62.

Menard, S., 1995. Applied Logistic Regression Analysis. Sage Publications Series: ***Quantitative Applications in the Social Sciences***, p. 106.

Mendeş, M., Akkartal, E., 2009. Regression Tree Analysis For Predicting Slaughter Weight in Broilers. ***Italian J. Anim.*** Sci, 8: 615-624.

Oğuzlar A. 2003. Veri ön izleme. ***Erciyes Üniversitesi iktisadi ve idari Bilimler Fakültesi Dergisi*** 21.67-76.

Oruçoğlu, O. 2011. ***Holstein Irkı İneklerin 305 Günlük Süt Verimini Etkileyen Çevre Faktörlerinin Regresyon Ağacı ile Belirlenmesi*** (Yüksek lisans tezi) Süleyman Demirel Üniversitesi**,** Fen Bilimleri Enst. Isparta.

Ozgulbaş, N., ve Koyuncugil, A. S., 2007. Financial Profiling Of Public Hospitals***: An Application By Data Mining. Int J Health Plann Mgm***t 2009; 24: 69–83.

Küçükönder, H., Üçkardeş, F., Nariç, D., 2014. Hayvancılık Alanında Bir Veri Madenciliği Uygulaması: ***Japon Bıldırcını Yumurtalarında Döllülüğe Etki Eden Bazı Faktörlerin Belirlenmesi***. ***Kafkas Üniv Vet Fak Derg*** 20 (6): 900-908.

Piwczynski, D., 2009 Using Classification Trees in Statistical Analysis of Discrete Sheep Reproduction Traits. ***Bydgoszcz*** 85-84, ***Mazowiecka*** 28 Poland.

Piwczynski, D., Sitkowska, B., 2012a. Statistical Modelling of Somatic Cell Cvounts Using the Claassification Tree Technique. ***Archiv Tierzuccht*** 55 4, Pg 332-345

Piwczynski, D., Nogalski, Z., Sitkowska, B., 2013. ***Statistical Modeling of Calving Ease and Stillbirths in Dairy Cattle Using the Classification Tree Technique***. ***Livestock Science***. 154 Pg 19-24

Piwczynski, D., Sitkowska, B., Wisniewska, E., 2012b. Aplication of Classification Trees and Logistic Regression to Determine Factors Responsible for Lamb Mortality. ***Small ruminant research*** 103 pg 225-231

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by backpropagating errors. ***Nature*** 323, 533–536.

Savaş, S., Topaloğlu, N., Yılmaz, M., (2012). Veri Madenciliği Ve Türkiyedeki Uygulama Örnekleri. ***İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*** Yıl:11 Sayı: 21 Bahar 2012 s. 1-23

Song, T.T., 1997. Analysis of correlated ROC areas in diagnostic testing. ***Biometrics*** 53, 370–382.

Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. ***Communications in Statistics-Theory and Methods*** A7, 13–26.

Süt, N. ve Şimşek, O. 2011. Comparisson Of Regression Tree Data Mining Methods For Prediction of Mortality in Head İnjury. ***Nowember, December*** 2011, Pages 15534-15539.

Takma, Ç., Atıl, H., Aksakal, V., 2012. Çoklu Doğrusal Regresyon ve Yapay Sinir Ağı Modellerinin Laktasyon Süt Verimlerine Uyum Yeteneklerinin Karşılaştırılması. ***Kafkas Üniversitesi Veteriner Fakültesi Dergisi***, 18(6), 942.

Tüzüntürk, S. (2010). Veri Madenciliği Ve İstatistik. ***Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi Cilt XXIX***, Sayı 1, 2010, s. 65-90.

Topal,M., Yağanoğlu, A. M., Sönmez, A. Y., 2010. Using Discriminant and CHAID Analysis Methods to Identify Sex in Brown Trout (Salmo trutta fario) by Morphometric Features. ***The Israeli Journal of Aquaculture Bamidgeh*** 62(4), 251-259

Uckardes, F., Narinc, D., Kuçukonder, H., Rathert, T. C., 2014. ***Application of Classification Tree Method to Determine Factors Affecting Fertility in Japanese quail Eggs***. 4(8): 1017-1023

Vapnik, V.N., Chervonenkis, A., 1971. On the uniform convergence of relative frequencies of events to their probabilities. ***Theory of Probability and its Applications*** 16, 264–280.

Yakut, E., Elmas, B., (2013). İşletmelerin Finansal Başarısızlığının Veri Madenciliği ve Diskriminant Analizi Modelleri İle Tahmin Edilmesi. ***Afyon Kocatepe Üniversitesi, İİBF Dergisi*** (C.XV, S I, 2013)

ÖZGEÇMİŞ

1982 yılında Ordu İli Ünye İlçesinde doğdu. İlköğretimini Ünye Karadere okulunda, Ortaöğretimi ise Açık öğretim ortaokulu ve Lisesinde tamamladı. 2009 yılında Iğdır Üniversitesi, Meslek Yüksekokulu Makine Bölümünü kazandı ve 2011 yılında 3.'lükle bitirdi. Yine 2009 Yılında Anadolu Üniversitesi Açık öğretim Fakültesi Sosyoloji Bölümüne kayıt yaptırdı ve 2013 yılında mezun oldu. 2014 yılında Iğdır Üniversitesi Fen Bilimleri Enstitüsü Zootekni (Biyometri ve Genetik) Anabilim Dalında Lisansüstü öğrenimine başladı. Evli olup, bir çocuk babasıdır.