T.C.

ÇANAKKALE ONSEKİZ MART ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ YÖNTEMLERİNİN HAYVANCILIKTA KULLANIMI

Orçun KÜÇÜKOĞLU

Zootekni Anabilim Dalı

Tezin Sunulduğu Tarih **08/06/2010**

Tez Danışmanı:

Doç. Dr. Mehmet MENDEŞ

ÇANAKKALE

YÜKSEK LİSANS TEZİ SINAV SONUÇ FORMU

Orçun KÜÇÜKOĞLU tarafından **DOÇ. DR. Mehmet MENDEŞ** yönetiminde hazırlanan **"VERİ MADENCİLİĞİ YÖNTEMLERİNİN HAYVANCILIKTA KULLANIMI"** başlıklı tez tarafımızdan okunmuş, kapsamı ve niteliği açısından bir Yüksek Lisans tezi olarak kabul edilmiştir.

Doç. Dr. Mehmet MENDEŞ

Danışman

Doç. Dr. Akın PALA

Jüri Üyesi

Yrd. Doç. Dr. Özcan ÖZEN

Jüri Üyesi

Sıra No:

Tez Savunma Tarihi: 08/06/2010

Prof Dr. İsmail TARHAN

Müdür

Fen Bilimleri Enstitüsü

İNTİHAL (AŞIRMA) BEYAN SAYFASI

Bu tezde görsel, işitsel ve yazılı biçimde sunulan tüm bilgi ve sonuçların akademik ve etik kurallara uyularak tarafımdan elde edildiğini, tez içinde yer alan ancak bu çalışmaya özgü olmayan tüm sonuç ve bilgileri tezde kaynak göstererek belirttiğimi beyan ederim.

Orçun KÜÇÜKOĞLU

TEŞEKKÜRLER

Bu tezin gerçekleştirilmesinde bana yol gösteren, eleştiri ve önerileri ile daha doğruya ulaşmama yardımcı olan saygı değer danışman hocam Doç. Dr. Mehmet MENDEŞ'e ve benden hiçbir zaman maddi ve manevi desteklerini esirgemeyen; aileme (Mehmet KÜÇÜKOĞLU, Ferize KÜÇÜKOĞLU, Onur KÜÇÜKOĞLU'na) sonsuz teşekkürlerimi sunarım…

Orçun KÜÇÜKOĞLU

KISALTMALAR VE SİMGELER LİSTESİ

$\mathbf{Y_i}$	Hedef Değişken Değeri
Y(k)	k Düğümünün Ortalaması
$\mathbf{X_i}$	Girdiler
$\mathbf{W_{ij}}$	Ağırlıklar
Σ	Toplam Fonksiyonu
F(Σ)	Aktivasyon Fonksiyonu
$\mathbf{Y_i}$	Model Tahmin Değeri
$\mathbf{Y_t}$	Gerçek Değer
N	Gözlem Sayısı
P	Önemlilik Değeri
χ**2**	Ki-Kare Değeri
$\bar{X}$	Ortalama
F (V_j)	Çıktı Değeri
$\mathbf{N_i}$	Başlangıç Veri Setindeki i Sınıfında Bulunan Deney Ünitelerinin Sayısı
$\mathbf{N_i(t)}$	t Düğümündeki i. Sınıfında Bulunan Deney Ünitelerinin Sayısı
R(k)	Nodun Varyansını ya da k Düğümünün Risk Değerini
VM	Veri Madenciliği
CA	Canlı Ağırlık
SA	Sınıflandırma Ağaçları
RA	Regresyon Ağaçları
SRA	Sınıflandırma Ve Regresyon Ağaçları
YSA	Yapay Sinir Ağları

HKO Hata Kareler Ortalaması

KHKO Hata Kareler Ortalamasının Karekökü

OMHY Ortalama Mutlak % Hata

OMH Ortalama Mutlak Hata

KIKT Küme İçi Kareler Toplamına

ÖZET

VERİ MADENCİLİĞİ YÖNTEMLERİNİN HAYVANCILIKTA KULLANIMI

Orçun KÜÇÜKOĞLU

Çanakkale Onsekiz Mart Üniversitesi

Fen Bilimleri Enstitüsü

Zootekni Anabilim Dalı Yüksek Lisans Tezi

Danışman: Doç. Dr. Mehmet MENDEŞ

08/06/2010, 45

Bu tez çalışmasında; Sınıflandırma ve Regresyon Ağaçları, Yapay Sinir Ağları ve k-Ortalamalar Kümeleme Yöntemi gibi farklı veri madenciliği yöntemlerinin hayvancılıkla ilgili çalışmalarda kullanımları üzerinde durulmuştur. Söz konusu yöntemlerden hem değişkenler arasındaki ilişkilerin araştırılmasında, hem tahmin yapma hem de sınıflama amacıyla yararlanılabilmektedir. Çalışmada önce bu yöntemlerin teorik temelleri hakkında detaylı bilgiler verilmiştir. Daha sonra bu yöntemlerin uygulaması yapılmıştır. Sınıflandırma Ağaçları, Türk Saanen ırkı keçilerinden elde edilen verilere uygulanmıştır. Diğer yöntemler ise Ross 308 hattından etlik piliçlerden elde edilen verilere uygulanmıştır. Değişkenler arasındaki ilişkilerin araştırılması ve tahmin yapma amacıyla Regresyon Ağaçları yönteminden, tespit edilen özelliklerden yararlanılarak hayvanların sınıflandırılması ve söz konusu sınıflandırma da etkili olan önemli değişkenlerin belirlenmesinde ise Sınıflandırma ağaçları, Yapay Sinir Ağları ve k-Ortalamalar Kümeleme Yönteminden yararlanılmıştır. Sonuç olarak dikkate alınan yöntemlerden hayvancılıkla ilgili çalışmalarda da etkin bir şekilde yararlanılabileceği görülmüştür.

Anahtar sözcükler: Veri Madenciliği, Sınıflandırma ve Regresyon Ağaçları, Yapay Sinir Ağları, k-Ortalamalar Kümeleme, Tahmin

ABSTRACT

THE USAGE OF DATA MINING METHODS IN ANIMAL SCIENCE

Orçun KÜÇÜKOĞLU

Çanakkale Onsekiz Mart Üniversity

Graduate School of Science and Engineering

Chair for Department of Animal Science Thesis of Master of Science

Advisor: Assoc. Prof. Dr. Mehmet MENDEŞ

08/06/2010, 45

In this thesis, the usage of data mining methods, namely Classification and Regression Tree, Artificial Neural Network and k-Means Clustering in animal science were studied. These methods can be used both for investigation of relations between variables, prediction and clustering. First, detailed information about the theoretical basis of the methods was provided. Second, the methods were applied. The classification tree method was applied to the data which were obtained from Turkish Saanen Goats, while the other methods were applied to data obtained from Ross 308 line broiler chickens. Regression Tree method was used to investigate the relations between the variables and for prediction of the slaughter weight (CA6) of broiler chickens, and the other methods were used for classification and to determine the variables which had significant effects on the response variables. The methods studied in this thesis may be utilized effectively in animal science research.

Keywords: Data Mining, Classification and Regression Trees, Artificial Neural Networks, k-Means Clustering, Estimate

İÇERİK **Sayfa**

BÖLÜM 1

GİRİŞ

Pek çok alanda olduğu gibi hayvancılıkla ilgili çalışmalarda da genellikle deney ünitelerinin birçok özelliğine ilişkin veri toplanmaktadır. Veriler ölçüm, tartım ve analiz etmek suretiyle elde edilebileceği gibi sayılarak, sıralanarak, var-yok şeklinde ya da kategorik olarak da elde edilebilir. Uygulamada yapılan çalışmalarda genel olarak veriler karmaşık bir şekildedir. Dolayısıyla pek çok durumda araştırıcılar büyük ve karmaşık veri setleri üzerinde çalışmak durumunda kalmaktadırlar. Bu durumda büyük emek verilerek ve masraf yapılarak elde edilen verilerin uygun istatistiksel yöntemlerle değerlendirilmesi hem elde edilecek bilginin kalitesi hem de varılacak sonuçların güvenilirliği bakımından oldukça önemlidir. (Swift, 2001; Mendeş ve Akkartal, 2009).

Uygulamada deney ünitelerinin tespit edilen özellikleri arasındaki ilişkilerin araştırılması, bu ilişkilerin uygun bir şekilde modellenmesi, üzerinde durulan özelliğe (sürekli ya da kesikli) ilişkin tahminlerde bulunulması, yapılacak tahminlerde etkili olabilecek özellik ya da özelliklerin belirlenmesi ve tespit edilen özellikler bakımından deney ünitelerinin sınıflandırılması amacıyla yapılan çalışmalardan elde edilen verilerin istatistiksel analizlerinde genellikle varyans analizi modelleri, regresyon analizi modelleri, ayırma (discriminant) analizi ve kümeleme (cluster) analizi gibi klasik istatistik tekniklerinden yararlanılmaktadır (Coşkun ve ark., 2004; Çamdeviren ve ark., 2005; Mendeş ve Akkartal, 2009). Ancak, bu tekniklerden beklenilen yararların sağlanabilmesi, çalışılan veri setlerinde normal/çok değişkenli normal dağılım, varyansların homojen olması / varyans-kovaryans matrislerinin homojen olması, bağımsız değişkenler arasında yüksek ilişkilerin bulunmaması (çoklu bağlantı problemi) gibi bir takım ön şartların yerine gelmesine bağlıdır (Draper ve Smith, 1998; Chatterjee ve Hadi, 2006). Diğer taraftan bu yöntemlerin gerektirdiği varsayımların sağlanması durumunda bile bu yöntemlerle her zaman merak edilen soruların cevapları verilememekte yani istenilen bilgilere ulaşılması pek mümkün olamamaktadır.

Bu durum özellikle karmaşık veri setleri ile çalışılması yani dikkate alınan özelliklerin bir kısmının kesikli bir kısmının sürekli olduğu, değişkenler arasında yüksek ilişkilerin bulunduğu, veri setinde kayıp gözlem ve uç değerlerin bulunduğu durumlarda çok daha da belirginleşmektedir (Breiman ve ark., 1984).Aynı zamanda bahsedilen klasik istatistiksel yöntemlerden yararlanılarak yüksek seviyeli interaksiyonlara ilişkin bilgi elde edilmesi de her zaman mümkün olamamaktadır.

İşte bu gibi durumlarda klasik yöntemlerin yerine Sınıflandırma ve Regresyon Ağaçları (classification and regression trees), Yapay Sinir Ağları (artificial neural network) ve k-ortalamalar kümelemesi (k-means clustering) gibi bazı veri madenciliği (data mining) yöntemlerinden etkin bir şekilde yararlanılabilmektedir (Breiman ve ark., 1984; Ribic ve Miller, 1998; Çamdeviren ve ark., 2005). Büyük ve karmaşık veri setlerinden bilgi üretmek ya da verileri bilgiye dönüştürmek veri madenciliği olarak tanımlanır (Swift, 2001). Dolayısıyla veri madenciliğinin amacı; büyük ve karmaşık veri gruplarında gizli kalmış bilgilerin ortaya çıkartılması ve bu bilgilerden araştırıcıların farklı şekillerde yararlanabilmesine olanaklarını araştırmasıdır (Özdemir ve ark., 2007; Özdemir ve ark., 2009).

Veri madenciliği yöntemlerinin (özellikle sınıflandırma ve regresyon ağaçları yöntemleri) daha ziyade bankacılık, pazarlama, sigortacılık, borsa, mühendislik, endüstri ve tıp gibi bazı alanlarda yaygın olarak kullanıldığı görülmektedir (Özekeş, 2003; Temel ve ark., 2005; Mendeş ve ark., 2008; Mendeş ve Akkartal, 2009). Halbuki, bu yöntemlerden tarımsal alanlarda da özellikle de hayvancılıkla ilgili çalışmalarda da çok etkin bir şekilde yararlanılabilir. Çünkü daha öncede belirtildiği gibi bu yöntemlerin aynı amaçla kullanılabilen klasik istatistiksel yöntemlere göre birçok avantajları vardır. Bu avantajlar özellikle büyük ve karmaşık veri setleri ile çalışıldığı ve tespit edilen değişkenler arasında yüksek ilişkilerin bulunduğu durumlarda çok daha belirginleşir. Aynı zamanda bu yöntemler sonucunda elde edilen bulgular grafiksel olarak sunulduğu için özellikle istatistikçi olmayan araştırıcılara sonuçların yorumlanma aşamasında büyük kolaylıklar sağlar.

Bu tez çalışmasında hem değişkenler arasındaki ilişkilerin araştırılmasında, hem sınıflandırma, hem de tahmin yapma amacıyla kullanılabilen sınıflandırma ve regresyon ağaçları (classification and regression tree: SRA), yapay sinir ağları (artificial neural network: YSA) ve k-ortalamalar kümelemesi (k-means clustering: KO) gibi dört farklı veri madenciliği yöntemlerinin hayvancılıkla ilişkili çalışmalarda uygulaması yapılacaktır.

BÖLÜM 2

ÖNCEKİ ÇALIŞMALAR

Veri madenciliği (VM) alanında pek çok çalışma bulunmaktadır. Her ne kadar söz konusu çalışmaların büyük bir kısmı tıp, mühendislik, endüstri ve sigortacılık alanında yapılmış çalışmalardan oluşmakta ise de, özellikle son yıllarda tarımla ilişkili çalışmalarda da bu yöntemlerden etkin bir şekilde yararlanılmaya başlandığı görülmektedir. Ancak hayvancılıkla ilgili çalışmalarda daha ziyade sınıflandırma ve regresyon ağaçları yöntemi ile yapay sinir ağlarından yararlanıldığı dikkati çekmiştir.

Yapay sinir ağlarından yararlanarak etlik piliçlerde asitesi tahmin etmek için yapılan çalışmada kandaki oksijen düzeyi, vücut ağırlığı, EKG, hematokrit, S dalgası ve kalp atım hızını yapay sinir ağı modeli için girdi değişkenleri olarak kullanılmıştır. Sonuçta yapay sinir ağları ile oluşturulmuş modellerden yararlanılarak asitesli hayvanların büyük bir doğrulukla (% 97) belirlenebileceği bildirilmiştir (Roush ve ark., 1997). Pulmoner hiper tansiyon sendromunun tahmin edilmesinde (Kirby ve ark., 1997), YSA'ları modeli ile mastitise neden olan mikroorganizmaları sınıflandırılmasında (Heald ve ark., 2000), etlik piliçlerde verimin tahmin edilmesinde (Salle ve ark., 2003), etlik piliçlerde kullanılan yem hammaddelerinin besin değerlerinin tahmin edilmesinde (Ahmadi ve ark., 2008), çiftlik koşulları ile süt proteini arasındaki ilişkinin tahmin edilmesinde (Fernandez ve ark., 2005), çevre kontrollü kümeslerde yetiştirilen erkek etlik piliçlerin büyümelerinin tanımlanmasında (Roush ve ark., 2006) yapay sinir ağlarından etkin bir şekilde yararlanmışlardır.

Çiftlik koşulları ile sütün içeriği arasındaki interaksiyonların belirlenmesinde (Barber ve ark., 2005), ergin yaş çocuklarda back depresyon skorlarına etki eden risk faktörlerinin ortaya konmasında (Çamdeviren ve ark., 2005), Brown-Swiss ırkı ineklerinde gerçek süt veriminin tahmin edilmesinde etkili olabilecek faktörlerin belirlenmesinde (Mendeş ve ark., 2008), etlik piliçlerde kesim ağırlığının tahmin edilmesinde etkili olan vücut ölçülerinin belirlenmesinde (Mendeş ve Akkartal, 2009) regresyon ağaçları yönteminden etkin bir şekilde yararlanmışlardır.

Kümeleme yöntemlerine ilişkin yapılan çalışmalar incelendiğinde söz konu yöntemlerin, koyunlarda döl verimi özellikleri, büyüme hızı, karkas kalitesi ile kuzuların et kalitesi arasındaki ilişkilerin belirlemesinde (Krogmeier ve ark., 1990), keçilerde laktasyon uzunluğunun sınıflandırılmasında (Bouloc ve Boichard, 1991) ve Türkiye'de yetiştirilen

çeşitli sığır ırkları arasındaki genetik ilişkilerin belirlemesinde (Önbeyaz ve ark., 1999) etkin bir şekilde yararlanılabildiği bildirilmiştir.

2.1. SINIFLANDIRMA VE REGRESYON AĞAÇLARI

Sınıflandırma ve Regresyon Ağaçları Yöntemi (SRA) özellikle büyük ve karmaşık veri setleri ile çalışıldığı ve değişkenler arasında doğrusal olmayan ilişkilerin söz konusu olduğu durumlarda araştırıcılara büyük avantajlar sağlamaktadır. Çünkü bu yöntem hem bağımsız değişkenler arasındaki yüksek seviyeli interaksiyonları da dikkate alarak bağımlı ve bağımsız değişkenler arasındaki ilişkilerin araştırılmasına hem de bağımlı değişkenin tahmin edilmesinde etkili olan bağımsız değişkenlerin belirlenmesine ve deney ünitelerinin sınıflandırılmasına imkan vermektedir (Çamdeviren ve ark., 2005). Söz konusu analizlerde Regresyon Ağaçları Yönteminden mi yoksa Sınıflandırma Ağaçları yönteminden mi yararlanılabileceği, bağımlı değişkenin yapısına bağlı olarak değişmektedir. Eğer bağımlı değişken sürekli (ölçüm veya tartım yoluyla) bir değişken ise Regresyon Ağaçları (Regression Tree; RA) yönteminden yararlanılır. Diğer taraftan bağımlı değişkenin kategorik yapıda olması halinde ise Sınıflandırma Ağaçları (Classification Tree; SA) yönteminden yararlanılabilir (Breiman ve ark., 2003).

Bu yönüyle Sınıflandırma ve Regresyon ağaçları (SRA) yöntemi, hem çoklu regresyon analizini, hem bağımlı değişkenin kategorik olduğu durumlarda kullanılan lojistik regresyon yöntemini hem de kümeleme ve diskriminant analizi yöntemlerini kapsamaktadır. Buna karşın, SRA yöntemi söz konusu tekniklerin (Varyans analizi modelleri, Regresyon modelleri ve Ayırma (discriminant) analizi gibi) varsayımlarını gerektirmemekte ve çalışılan veri setindeki değerlere müdahale etmeksizin kendi çevresinde bağımlı değişkenler ile bağımsız değişkenler arasında ilişkiler araştırabilmektedir. SRA yönteminin en önemli özelliklerinden birisi de çalışılan veri setindeki heterojenliğin dikkate alınarak, homojen alt gruplar oluşturarak bu alt gruplarda bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi açıklayabilmesi ve söz konusu ilişkiyi ağaç yapısı şeklinde görselleştirebilmesidir. SRA, kullandığı güçlü algoritma sayesinde sadece bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin yapısını araştırmanın yanında, dikkate alınan bağımsız değişkenlerin birbirleri ile olan etkileşimlerini de (interaksiyonlar) ortaya koyabilmektedir (Kayri ve Boysan, 2008; Mendeş ve Akkartal, 2009).

2.1.1. SINIFLANDIRMA AĞAÇLARI (SA)

Sınıflama Ağaçları (Classification Trees, SA) yöntemi, kategorik yapıdaki bağımlı değişkenin alacağı bir takım değerleri tahmin etmek üzere geliştirilen parametrik olmayan istatistiksel bir yöntemdir (Fu, 2003; Breiman ve ark., 2003). Sınıflandırma ağaçları, parametrik olmayan bir yöntem olduğu için bağımsız değişkenlerin dağılımlarına ilişkin herhangi bir varsayım gerektirmez. Dolayısıyla modele dahil edilecek bağımsız değişkenler; sürekli, kesikli, kategorik ya da sıralı olabilir (Yohannes ve Hoddinott, 2003). Bu yöntem, uygulamada yaygın olarak kullanılan çoklu regresyon, lojistik regresyon, diskriminant ve kümeleme analizi gibi birçok istatistik tekniğine göre pek çok avantajları bulunduğundan özellikle son yıllarda söz konusu yöntemlerin güçlü bir alternatifi olarak kullanılmaya başlanmıştır.

Uygulamada bu yöntemden daha ziyade sınıflandırma ve bağımlı değişkeni etkileyen önemli bağımsız değişkenlerin belirlenmesi amacıyla yararlanılmaktadır (Temel ve ark., 2005). SA yöntemi, elde edilen analiz sonuçlarını aynı zamanda görsel olarak ta sunabildiği için çok fazla bir istatistik bilgiye gerek duyulmadan ağaç şeklindeki model sonuçları kolayca yorumlanabilir. Bu yöntem bağımlı değişkeni etkileyebilecek bütün bağımsız değişkenleri ve bütün kombinasyonlarını modele katar ve en doğru sınıflandırmayı yapar (Yohannes ve Hoddinott, 2003; Breiman ve ark., 2003). Bu yöntem değişkenlerin kombinasyonlarına da baktığı için interaksiyonları da değerlendirmiş olur.

SA, aynı zamanda çalışılan veri setlerinde kayıp gözlemler ve uç ya da aykırı değerlerin bulunması durumlarında da rahatlıkla kullanılabilmektedir (Yohannes ve Hoddinott, 2003; Breiman ve ark., 2003). SA yönteminin yukarıda bahsedilen avantajlarının yanında bazı dezavantajları da bulunmaktadır. Bu yöntemin en önemli dezavantajı, elde edilen sonuçların bir olasılık modeline dayanmıyor olmasıdır. Yani veri setine uygun bir sınıflandırma ağacından alınmış tahmini sınıflandırmaya katkı sağlayabilecek olasılık düzeyi ya da güven aralığı yoktur. Bu sonuçların doğruluğuna duyulabilecek güven tamamen geçmiş verilere bağlıdır (Yohannes ve Hoddinott, 2003; Breiman ve ark., 2003). Bu yöntem, bütün verilerin içerisinde bulunduğu ve ana düğüm ya da kök düğüm (root node) denilen düğümden başlanarak devam eden ve her düğümde o düğüme uygulanabilen basit evet / hayır cevaplarına göre oluşan dalları içeren bir ağaç yapısından oluşur. Her düğümde uygulanan bu sorulara ayıraç, bu işlem ise ayırma olarak adlandırılır (Death ve Fabricius, 2000).

Eğer elimizde birden fazla bağımsız değişken bulunuyorsa değişen tek şey ayıraçların tüm değişken ve değişken kombinasyonlarını tek tek ele almasıdır. Bu durumda elde bulun tüm değişkenler ve bu değişkenlerin kombinasyonlarının yani interaksiyonlarının tanımlı bulunduğu aralıktaki tüm olası değerler birer ayıraç olarak kabul edilir ve mümkün olan bütün ayırmalar belirlenir. Bu ayırmalar sonucunda oluşan ağaçlardan homojen olmayan düğümler; çocuk (child) düğümü, homojen olan düğümlere ise terminal düğüm olarak adlandırılır. Bu hesaplamalar sonucu terminal düğümler kontrol grubu olarak adlandırılır ve bu düğümler yorumlanır. Herhangi bir düğümün heterojenlik değeri safsızlık (imprutiy) ölçüsü olarak adlandırılır. Bu değer safsızlık fonksiyonu kullanılarak hesaplanır ve 0 değerini alıyorsa düğümün homojen olduğu anlaşılır. Sınıflandırma ağaçlarında kullanılabilen değişik safsızlık ölçüleri (Gini, Twoning, Ki-Kare, G-Kare vb.) vardır. Bunlardan Gini ve Twoning ölçüleri uygulamada en yaygın olarak kullanılan ölçülerdir (Örekici, 2004; Breiman ve ark., 2003). Kullanılan bu safsızlık ölçüleri her hangi bir t düğümü için en iyi ayırmanın seçimini önemli bir şekilde etkilemektedir. Bu nedenle safsızlık ölçüleri en iyi ayırma kriterleri olarak da bilinir.

Sınıflandırma ağacı yönteminde bir sınıflandırma ağacı oluşturulurken ön olasılıklar (priori) kullanılır. Bu ön olasılıklar deney ünitelerinin hangi sınıfa atanacağının belirlenmesinde oldukça önemli rol oynarlar (Statsoft, 2003). j sınıfı için ön olasılık değeri (π_j) ile gösterilir. Bu değerler veri setinden ya da araştırıcı tarafından hesaplanır. Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile ise doğruluk oranı hesaplanır.

Doğruluk oranı = 1 – Hata oranı

Verilerin sınıflandırılabilmeleri için oluşturulan modellerin hata oranlarına karar vermek için risk matrisinden yararlanılır. Ayırma sonucunda herhangi bir düğüme atanacak olan en uygun sınıf ise aşağıdaki gibi tahmin edilir.

C(j/i): i sınıfı j sınıfı gibi sınıflamanın maliyeti (risk maliyeti katsayıları)

π_i: i. sınıfın önceki olasılığı

N_i: başlangıç veri setindeki i sınıfında bulunan deney ünitelerinin sayısı

$N_i(t)$: t düğümündeki i. sınıfında bulunan deney ünitelerinin sayısı

$$\frac{C(j/i)\,\pi_i\,N_i^{(t)}}{C(j/i)\ldots\pi_j\,N_j^{(t)}} > \frac{N_i}{N_j} \qquad \textbf{(2.1)}$$

j'nin bütün değerleri (j=1,2,…,k ve j≠i) için sağlanıyorsa t düğümüne en uygun olarak i sınıfı atanır (Levis, 2000).

Bu düğümün hesaplanmasından sonra düğümün yapısına göre bazı durumlarda birden fazla sınıf yukarıdaki eşitsizliği sağlayıp en uygun sınıf olarak kabul edilir ya da hiçbir sınıf bu eşitsizliği saplayamaz. Böyle durumlarda en uygun sınıfın belirlenmesinde çoğulluk ve minimum risk olmak üzere iki yönteme başvurulur (Breiman ve ark., 2003; Örekici, 2004). Bunlardan çoğulluk yöntemi; hatalı sınıflama maliyetini göz önüne almaksızın (eşit kabul eder) düğüm içerisinde en büyük orana sahip olan sınıfı en uygun sınıf olarak atar. Minimum risk yöntemi ise düğüm içersinde deney ünitelerinin sınıflara dağılımını göz önüne almaksızın (eşit kabul eder) düğüm içersinde hatalı sınıflama maliyetini minimum yapan sınıfı en uygun sınıf olarak kabul etmektedir (Statsoft, 2003). Sınıflama ağacı modelinde tekrarlanarak ikili bölünmelerle homojen alt sınıflar elde edilir ve ağaç bu şekilde büyümeye devam eder (Levis, 2000; Chipman ve ark., 2000; Breiman ve ark., 2003; Bevilachqua ve ark., 2003) . Dolayısıyla aynı veri setine ilişkin birçok ağaç yapısı elde etmek mümkündür. Bu durumda önemli olan nokta; optimum ağacın oluşturulmasıdır (Mendeş ve Akkartal, 2009). Optimum ağaç yapısına ise bazı hususların dikkate alınmasıyla ulaşılabilir. Öncelikle oluşturulacak ağacın büyümesinin ya da dallanmasının hangi koşullarda durdurulacağına karar vermek gerekir.

Bu durum aşağıdaki şekilde açıklanabilir:

1) Her çocuk düğümündeki gözlem sayısı 10 ya da bunun altında ise
2) Her düğümde grup içi homojenlik söz konusu ise
3) Ağacın düzey sayısında analizi yürüten kişi tarafından bir sınıflama yapıldıysa,
4) Yeni oluşacak düğümlerle fazla bir değişiklik oluşturmuyorsa
5) Ağacın artık büyümesinin uygun olamayacağı ve dolayısıyla da büyümenin durdurulması gerektiğine karar verilebilir (Levis, 2000; Breiman ve ark., 2003).

Optimum ağaç yapısının elde edilebilmesi için ya da oluşturulan ağacın optimum ağaç olup olmadığının belirlenmesinde dikkat edilmesi gereken diğer bir husus ise çalışılan veri setinin deney ya da eğitim (training) ve test olmak üzere iki parçaya bölünmesi gerekir. Aynı zamanda çalışılan veri setinde geçerlilik testinin de yapılması gerekir. Bunun için uygulamada daha ziyade 10-fold validasyon ya da verilerin %70 ve %30 olarak ayrılarak geçerlilik testinin

yapıldığı durumlarla oldukça sık karşılaşılır. Bu şekilde bir geçerlilik testinin yapılması sonucunda oluşturulan ağaç yapısından yararlanılarak yapılacak yorumların geçerli olup olmadığı ya da genelleştirilip genelleştirilemeyeceğine ilişkin bir sonuca ulaşılır. Kısaca oluşturulan ağacın uygun değer ağaç olup olmadığı hakkında araştırıcıya bir fikir verir.

2.2. REGRESYON AĞAÇLARI (RA)

Sayısal olarak ifade edilebilen bir bağımlı değişkenle bağımsız değişkenler arasındaki ilişki yapısını araştırmak amacıyla kullanılan bir yöntemdir. Bu yöntemde de sınıflandırma ağaçları yönteminde olduğu gibi bağımsız değişkenler sürekli, kesikli, kategorik ya da sıralı olabilir. (Örekici, 2004; Fu, 2003; Breiman ve ark., 2003; Mendeş ve Akkartal, 2009). Regresyon Ağaçları (RA) yöntemi, çoklu regresyon, lojistik regresyon, diskriminant ve kümeleme analizi gibi geleneksel istatistiksel yöntemlere göre birçok avantaja sahiptir. Çünkü RA, parametrik olmayan bir yöntem olduğu için çalışılan veri setinde diğer yöntemlerin gerektirdiği varsayımların yerine getirilmesini gerektirmez. Bu yöntem özellikle büyük ve karmaşık veri setleri ile çalışılması durumunda diğer yöntemlere göre daha avantajlıdır.

Çünkü büyük ve karmaşık veri setlerinde, bağımlı değişkeni etkileyen değişkenler ve bu değişkenlerin modeldeki önemlilik durumlarını basit bir ağaç yapısı ile görsel olarak sunabilmektedir. RA yönteminin bağımsız değişkenler arasındaki yüksek ilişkiden (çoklu bağlantı problemi) etkilenmemesi, kayıp gözlemlerin bulunması durumunda da tahminler yapmaya imkan sağlaması, değişkenler arasındaki yüksek ilişkileri dikkate alması ve başlangıçta her ne kadar sadece büyük veri setleri için geliştirilmiş olsa da artık küçük veri setlerinde de etkin bir şekilde yararlanılabilmesi, bu yöntemin klasik istatistik yöntemlerine göre diğer avantajları olarak gösterilebilir (Bremian ve ark., 1984; Talmon, 1986; Beckman ve ark., 1995; Chaudhuri ve ark., 1995; Steinber ve Colla, 1995; Honeycutt ve Gibson, 2003; Çamdeviren ve ark., 2005).

RA yöntemi bu avantajlarından dolayı hem tahmin yapma hem de sınıflandırma amacıyla birçok bilim dalında (ilaç, endüstri ve mühendislik) yaygın bir şekilde kullanılmaktadır (Çamdeviren ve ark., 2005; Mendeş ve Akkartal, 2009). Bu alanların dışında RTA yaygın olarak hayvanlarla ilgili birçok çalışmalarda da kullanılmaktadır. Regresyon ağacı, bütün bağımsız değişkenleri kullanılarak verileri alt gruplara ayırarak oluşturulan bir ağaçtır. İlk başta tek bir düğüm (kök ya da terminal düğüm) olarak başlar ve tüm gözlemleri de içeren ve birbirini takip eden alt düğümlere (çocuk düğümleri) ayrılarak dallanır. Bu

dallanma gözlemlerin homojen olmasına kadar devam etmektedir (Bremian ve ark., 1984; Bevilacqua ve ark., 2003; Çamdeviren ve ark., 2005).

Regresyon Ağaçlarını Oluşturma Adımları:

1) Bütün gözlemleri içeren ve ana düğüm (root node ya da main node) denilen tek bir düğüm ile analize başlanır.
2) m_c, herhangi bir "c" dalına ilişkin tahmin değerini, Vc; c dalının ayırma varyansını ve n_c'de bu daldaki gözlem sayısını göstermek üzere ayırma işlemi $S = \sum\sum(y_i - m_c)^2$ şeklinde ifade edilen S değeri minimum oluncaya kadar devam ettirilir. Bu ifade $S = \sum n_c V_c$ şeklinde de yazılabilir. m_c ise; $m_c = \sum cY_i / n_c$ dir.
3) Eğer düğüm içerisindeki bütün bireyler dikkate alınan tüm bağımsız değişkenler için aynı değerleri alıyorsa ağacın bölünme işlemi sonlandırılır. Aksi halde S minimize edilebilecek en son noktaya kadar indirilir. Eğer S' deki en büyük azalma belirlenen δ gibi bir eşik değerinden küçük ise bölünme durur.
4) Her bir yeni düğüm için 1. adıma geri dönülür.

RA yönteminden yararlanılarak tahmin ya da sınıflandırma işlemleri yapılırken her işlem adımında sadece bir bağımsız değişken dikkate alınır. Eğer birçok bağımsız değişken var ise hangisinin seçileceği genel olarak tesadüfen belirlenir (Çamdeviren ve ark., 2005; Mendeş ve Akkartal, 2009).

Uygun Regresyon ağacını oluşturabilmek için başvurulan etkin yollardan birisi, cross-validation yönteminin kullanılmasıdır. Ağaçlar için cross validation kullanılırken birçok ipucu vardır. Bunların en önemlisi alternatif büyüme ve budamadır. Veri grubu 2 parçaya ayrılır. Bir parça deney (training), diğer parça ise test amacıyla kullanılır. Ağaca önce büyüme daha sonra budama işlemi uygulanır. Budama uygulanmasının amacı ise önceki aşamalarda modele dahil edilmiş değişkenlerin ileriki aşamalarda tekrar modele dahil olmasını engellemektir. Budanmış ağaç diğer yarısına gelene kadar tekrar budanır. Bu budama işlemi ağacın boyu artık değişmeyene kadar devam eder.

Her bir k düğümü için R(k) nodun varyansını ya da k düğümünün risk değerini gösterir.

R(k) hesaplanırken; $R(k) = \frac{1}{N(k)} \sum_{i \equiv k} [Y_i - \overline{Y}(k)]^2$ **(2.2)**

N(k): düğümün k içerisindeki gözlem sayısını gösterir.

Y_i: hedef değişken değerini gösterir.

Y(k): k düğümünün ortalamasıdır.

Risk değeri bağımlı değişkenin birimine bağlıdır.

Hata varyansı ya da açıklanamayan varyasyon S_e^2 şöyle hesaplanır;

$$S_e^2 = \frac{\text{Risk}}{S_y^2} \quad \textbf{(2.3)}$$

Bu ifadeden yararlanılarak bağımlı değişkendeki varyasyonun oluşturulan ağaç tarafından açıklanabilen kısmı ise;

$$S_x^2 = 1 - S_e^2 \quad \textbf{(2.4)}$$

Düğümlerin homojen olup olmadıklarının bir ölçüsü olarak en küçük kareler sapmaları (LSD) kullanılmaktadır (Bremian ve ark., 1984; Karalic, 1992; Dietterich, 1998; Torgo, 1998; Bevilacqua ve ark., 2003). LSD kriteri, k düğümünün S kadar bölünmesini belirtir.

$$\Phi(S,k) = R(k) - P_L R(k_L) - P_R R(k_R) \quad \textbf{(2.5)}$$

Burada;

P_L= K_L içinde verilen ana k düğümündeki gözlem sayısını

K_L= Sol çocuk düğümünü

P_R= K_R içinde verilen ana k düğümdeki gözlem sayısını

K_R= Sağ çocuk düğümünü göstermektedir.

S (bölünme), $\Phi(S,k)$ değerini minimize etmek için seçilmiştir. Bu değer k düğümü içersindeki bütün gözlemlerin ölçümü ile ağacın ilerleme durumunu belirtir (Talmon, 1986; Cappelli ve ark., 2002). Ağaç oluşum süreci bölünmeyi yapan her bir çocuk düğümündeki bütün gözlemlerin ya da her bir çocuk düğümündeki her bir gözlemin bağımsız değişken ile aynı değeri alana kadar devam eder. Aynı değeri aldıktan sonra ağaç oluşumu tamamlanmış

olur. Ağaç oluşumu tamamlandıktan sonra her bir veri grubunun nasıl parçalandığı (dağıldığı) yazılır yani ağaç modeli yorumlanır (Breiman ve ark.,1984; Camdeviren ve ark., 2005).

2.3. YAPAY SİNİR AĞLARI (YSA)

Son yıllarda bilgisayar teknolojisindeki baş döndürücü gelişmeler beraberinde karmaşık ilişkilerin araştırılmasında kullanılabilen veri madenciliği yöntemlerinden etkin bir şekilde yararlanma olanaklarını da arttırmıştır. Bu yöntemlerden birisi de yapay sinir ağlarıdır (YSA). YSA pek çok alanda yaygın bir şekilde kullanılmaktadır. Son yıllarda ziraat alanında da yaygın bir şekilde kullanılmaya başlanmıştır. Ancak, hayvancılık alanında da yaygın kullanım alanı bulunmasına rağmen bu alanda söz konusu yöntemden yeteri kadar yararlanılamadığı dikkati çekmektedir. Yapay sinir ağları (YSA), insan beyninin en önemli özelliklerinden biri olan öğrenme ya da eğitilme yolu ile yeni bilgiler üretebilme ve gizli kalmış ilişkileri ortaya çıkarabilme gibi yetenekleri, herhangi bir destek almadan otomatik olarak gerçekleştirebilmek amacı ile geliştirilen bilgisayar sistemleridir (Öztemel, 2003). Dolayısıyla YSA; insan beyninden esinlenerek, öğrenme sürecinin matematiksel olarak modellenmesi uğraşı sonucu ortaya çıkmış bir yöntemdir. Bu nedenle, YSA üzerine yapılan çalışmalar ilk olarak beyni oluşturan biyolojik üniteler olan nöronların modellenmesi ve bilgisayar sistemlerinde uygulanması ile başlamış, daha sonraları bilgisayar sistemlerinin gelişimine de paralel olarak birçok alanda kullanılır hale gelmiştir. Diğer veri madenciliği yöntemlerinde olduğu gibi yapay sinir ağlarının bazı avantaj ve dezavantajları vardır. Yapay sinir ağları özellikle büyük ve karmaşık veri setleri ile çalışılması ve dikkate alınan değişkenler arasındaki ilişkilerin doğrusal olmadığı durumlarda hem tahmin yapma hem de sınıflandırma amacıyla etkin bir şekilde kullanılabilmesi, bu yöntemin en önemli avantajlarıdır. Bu yöntemin her hangi bir matematiksel modele ihtiyaç duymaması ve çalışılan veri setinde kayıp gözlemlerin ve uç değerlerin bulunması durumunda da çözüm üretebilmesi diğer önemli avantajları olarak karşımıza çıkmaktadır (Öztemel, 2003). Bu özelliklerinden dolayı YSA büyük ve karmaşık veri setlerinde değişkenler arasındaki ilişkilerin araştırılmasında kullanılabilmesinin yanında, üzerinde durulan özelliğe ilişkin yapılacak tahminlerde önemli düzeylerde etkili olan değişkenlerin belirlenmesinde, genelleme yapılmasında ve sınıflandırma amacıyla da kullanılabilmektedir. YSA aynı zamanda örneklerden elde edilen bilgiler ile kendi deneyim kazanarak daha sonra karşılaşabileceği benzer bir durumda benzer karar verebilmektedirler.

Ancak, verilerin analiz aşamasındaki karmaşık durum, uygun modelin ya da ağ yapısının belirlenmesi için belirli bir kuralın olmaması ve dolayısıyla da söz konusu ağ

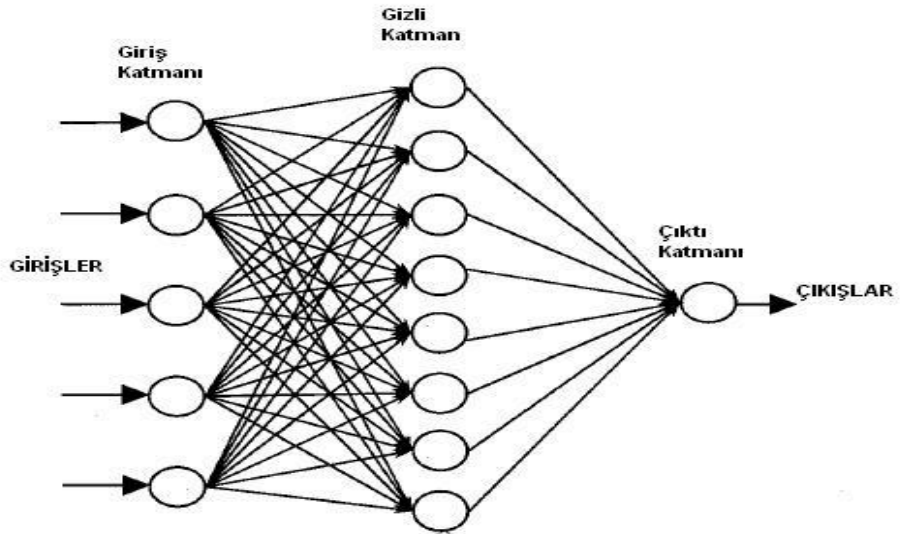
yapısının tamamen araştırıcının bilgi ve tecrübesine bağlı olması bu yöntemin dezavantajları olarak karşımıza çıkmaktadır. Bunun sonucunda da merak edilen konu hakkında ulaşılan sonucun uygun (optimum) sonuç olup olmadığı garanti edilememektedir (Yurtoğlu, 2005; Mamdouh, 2007).

2.3.1. Yapay Sinir Ağı Modeli

Bir YSA modeli; birbirleriyle bağlantılı olan sinirlerin bulunduğu katmanlardan oluşmaktadır. Bu katmanlar genel olarak; girdi katmanı, gizli katman ve çıktı katmanı olmak üzere üç bölümden oluşmaktadır. Girdi değişkenleri bu katmanlardan ilk olan girdi katmanından YSA modeline girer ve dış ortamla herhangi bir bağlantısı olmayan sinirlerden oluşan bir katmandır. Daha sonra bu değişkenler girdi katmanından geçerek ve çıktı katmanına iletmek ile görevli olan gizli katmana geçer. Buradan da son katman olan çıktı katmanına iletilirler. Yapay sinir ağının bir örnek modeli Şekil 1'de görülmektedir.

Bir YSA'nın kendisine verilen görevi yerine getirebilmesi için öncelikle ilgili olayın örnekleri ile eğitilerek (öğrenme) genelleme yapabilecek yeteneğe kavuşturulması gerekir. Bu sayede birbirleri ile aynı ya da benzer olaylara karşılık gelen çıktı setleri belirlenebilir. Dolayısıyla dikkate alınan değişkenler arasındaki ilişkilerin araştırılmasında ya da sınıflandırma amacıyla YSA yönteminden etkin bir şekilde yararlanılabilmesi için öncelikle veri setinin deney ya da eğitim (training) ve test olmak üzere iki kısma ayrılması gerekir (Yıldız, 1999; Güneri, 2001; Elmas, 2003). Böylece deney veri grubu üzerinden YSA'nın eğitimi yapılırken ya da genelleme yapabilme yeteneğine kavuşturulurken, test veri grubu ile de söz konusu öğrenme ya da eğitme işinin ne yeterli düzeyde olup olmadığının test edilmesi sağlanır. YSA'nın eğitiminde örnek genişliği oldukça önemlidir. Çünkü büyük hacimli örneklerle çalışılması, yapılacak tahminlerdeki güvenilirliği arttırmaktadır. Öğrenme sürecinde eğer çıkış değerleri de ağa veriliyorsa bu durumda denetimli (kontrollü) öğrenme söz konusu olur. Kontrollü öğrenmede çıkış değerleri ile ağın tahmin ettiği çıkış değerleri karşılaştırılarak hata miktarı bulunur ve bu hata kabul edilebilir seviyeye gelene kadar eğitime devam edilir. Diğer bir öğrenme tipi de kontrolsüz ya da denetimsiz öğrenmedir.

Burada kontrollü öğrenme şeklinden farklı olarak çıkış değerleri verilmez. Ağın çıkış değerlerini tahmin etmesi beklenir. Yapay Sinir Ağı Yapısı genel olarak Şekil 1'deki gibi gösterilebilir.

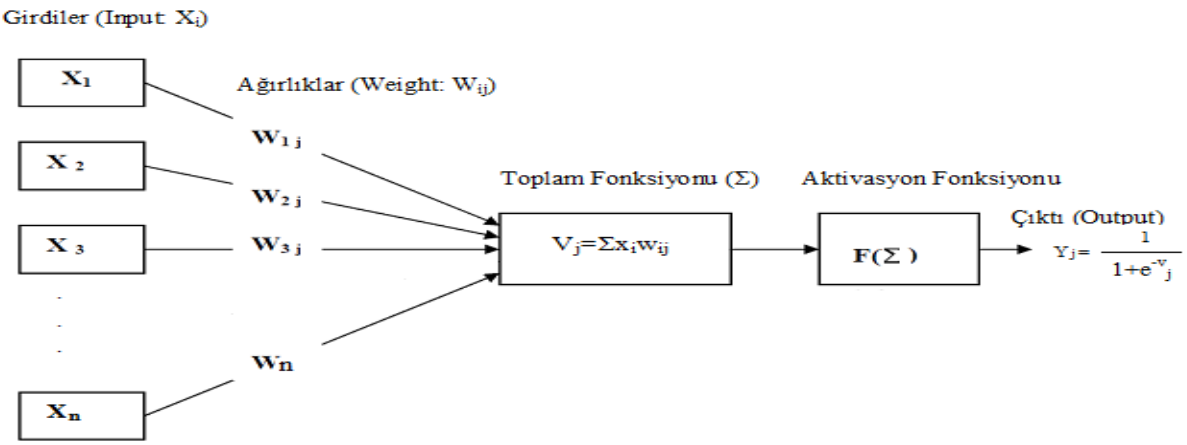


Şekil:1. Yapay Sinir Ağı Yapısı (Kurup ve Dudani 2002, Tolon ve Tosunoğlu 2008).

YSA yönteminde biyolojik sinir ağlarındaki sinir hücrelerinin yapısı taklit edilmektedir. Dolayısıyla yapay sinir ağlarında da sinir hücreleri söz konusudur. Ancak bu sinir hücreleri yapay sinir hücreleridir.

YSA daki bir sinir hücresi:

a) Girdiler (Input: X_i)
b) Ağırlıklar (Weight: W_{ij})
c) Toplam Fonksiyonu (Σ)
d) Aktivasyon Fonksiyonu ($F(\Sigma)$) ve
e) Çıktı (Output) olmak üzere 5 kısımdan oluşur (Tsoukalas ve Uhrig, 1997; Kurup ve Dudani 2002, Tolon ve Tosunoğlu 2008).



Şekil 2. Bir yapay sinir hücresinin yapısı.

Bu çalışmada bir gizli katman içeren çok katmanlı algılayıcılı ileri beslemeli (multilayer perceptron: MLP) YSA kullanılmıştır. Yapay sinir hücreleri girdi, gizli ve çıktı katmanı olmak üzere üç katman halinde tasarlanmıştır. Ağın eğitilmesinde tanjant hiperbolik (tanh) fonksiyonu kullanılmıştır. Yani kurulan YSA modelinde aktivasyon fonksiyonu olarak tanjant hiperbolik (tanh) aktivasyon fonksiyonu kullanılmıştır. Bu fonksiyon kullanılarak hesaplanan sinir hücresinin çıktı değeri F (V_j) ise aşağıdaki gibi hesaplanır (Şen, 2004).

$$F(V_j) = \frac{(e^{V_j} - e^{-V_j})}{(e^{V_j} + e^{-V_j})} \quad \textbf{(2.6)}$$

$$F(V_j) = \alpha V_j \quad \textbf{(2.7)}$$

$$F(V_j) = \frac{1}{(1 + e^{-\alpha V_j})} \quad \textbf{(2.8)}$$

$$F(V_j) = \exp\left(\frac{-V_j^2}{\sigma^2}\right) \quad \textbf{(2.9)}$$

Burada V_j yapay sinir hücresine karşılık gelen girdi değerini göstermektedir. Modelin performans kriterleri olarak uygulamada daha ziyade aşağıdaki ölçüler kullanılır. Bunlar;

Hata Kareler Ortalaması (HKO)

$$HKO = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_t)^2 \quad \textbf{(2.10)}$$

Hata Kareler Ortalamasının Karekökü (KHKO)

$$KHKO = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_t)^2} \quad \textbf{(2.11)}$$

Ortalama mutlak % hata (OMHY)

$$OMHY = \frac{1}{N}\left(\sum_{i=1}^{N}\left|\frac{Y_i - Y_t}{Y_i}\right|\right)x\,100 \quad \textbf{(2.12)}$$

Burada;

Y_i: model tahmin değerini ya da hesaplanan değeri

Y_t = gerçek değer

N= ise gözlem sayısını göstermektedir.

Ortalama mutlak hata (OMH)

$$OMH = \frac{1}{N}\left(\sum_{i=1}^{N}|Y_i - Y_t|\right) \text{ x } 100 \qquad \textbf{(2.13)}$$

2.4. k - ORTALAMALAR KÜMELEME YÖNTEMİ (k-means clustring)

Kümeleme analizi; n tane bireyi tespit edilen p tane ($p \geq 2$) özelliğinden yararlanarak kümelere ayrılmasında kullanılan çok değişkenli bir yöntemdir. Söz konusu bireyler kümelere ayrılırken aralarındaki benzerlik ya da farklılıklardan yararlanılır. Kısaca kümeleme analizi bireylerin p tane özelliğine göre hesaplanan bazı ölçüler kullanılarak (benzerlik ya da farklılık ölçüleri) homojen gruplara ayırmak amacıyla kullanılır. Oluşturulan kümeler kendi içinde homojen, kendi aralarında ise heterojendir. Kümeleme yöntemleri esas olarak aşamalı (hiyerarşik) ve aşamalı olmayan (hiyerarşik olmayan) yöntemler olmak üzere iki kısma ayrılır. Hiyerarşik kümeleme yöntemleri kümeleri birbiri ardına birleştirme sürecidir. Bir küme diğer herhangi bir küme ile bir kere birleştikten sonra daha sonra kesinlikle ayrılmaz (Fırat, 1995). Bu teknikte sonuçlar ağaç diyagramı olarak gösterilir ve bu diyagramlar dendogram olarak bilinir (Lorr, 1983). Hiyerarşik kümeleme yöntemlerinin kullanılabilmesi için başlangıçtaki küme sayısının bilinmesi gerekmez. Yani deney ünitelerinin ya da özelliklerin kaç kümeye ayrılacağının başta bilinmesi gerekmez. Bu yöntemlerden yararlanılarak deney üniteleri (gözlemler) kümelere ayrılırken ilk önce her bir deney ünitesi bir küme olarak kabul edilir. Daha sonra belirli benzerlik farklılık ölçülerine bağlı olarak hangi deney ünitelerinin hangi kümelere dahil edileceği belirlenerek kümeleme işlemi sürdürülür. Bu tez çalışmasında hiyerarşik olmayan kümeleme yöntemlerinden en çok kullanılanlardan birisi olan ve bir veri madenciliği yöntemi olarak da bilinen k-ortalamalar kümeleme yöntemi dikkate alınmıştır.

Hartigan ve Wong (1979) tarafından geliştirilen k-ortalamalar kümeleme yöntemi (k-means clustering) en eski kümeleme yöntemlerinden birisidir. Özellikle büyük veri setlerinde çalışılması durumunda n tane bireyin ya da deney ünitesinin tespit edilen p tane özelliğinden

yararlanılarak daha az sayıda k tane kümeye ya da gruba ayrılması amacıyla kullanılmaktadır. Bu k tane kümeye bölünme olayında kümeler içi benzerliğin en fazla, kümeler arası benzerliğin ise en düşük olmasını sağlamak k ortalama kümelemenin diğer bir amacıdır. Kümeler arası benzerlik ise; kümenin ağırlık merkezi olarak kabul edilen bir veri ile kümedeki diğer veriler arasındaki uzaklıkların ortalama değeri ile ölçülebilmektedir (Han ve Kamber, 2001; Berkhin, 2002). Söz konusu deney üniteleri kümelere ayrılırken kümeler içi kareler toplamının dolayısıyla da yanlış kümeleme (sınıflandırma) oranının minimize edilmesine çalışılır. Bu kümeleme yöntemi tespit edilen özelliklerin sürekli olması ve veri setinde uç değerlerin bulunmamasını gerektirmektedir. Bununla birlikte çalışılan veri setinde değişkenlerin bir kısmının kesikli olduğu durumlarda da kullanılabilmektedir. Ancak kesikli değişkenlerin bulunduğu durumlarda kullanılması pek önerilmemektedir.

Popüler bir kümeleme yöntemi olan k-ortalamalar kümeleme yönteminden yararlanılarak deney ünitelerinin kümelere ayrılabilmesi için söz konusu deney ünitelerinin kaç kümeye ayrılacağının bilinmesi gerekir. Yani küme sayısının baştan bilinmesi gerekir. Küme sayısı bilindiğinde k-ortalamalar kümeleme algoritması her kümenin merkezini (centroid) bularak söz konusu deney ünitelerini bu merkez noktasına yakınlıklarına göre hangi kümeye dahil edileceğini belirler. Her bir yeni gözlemin bir kümeye dahil edilmesi ile yeni bir merkez belirlenir ve bu işlem veri setindeki bütün gözlemler uygun kümelere ayrılıncaya kadar tekrarlanır.

k-ortalamalar kümeleme yöntemin algoritması aşağıdaki gibi özetlenebilir:

1) k-tane farklı küme belirlenir. Daha sonra çalışılan örnekteki bireyler tamamen rastgele ve k tane küme arasında dağıtılır.
2) Her bir kümenin merkezi (centroid) belirlenir.
3) Her küme içersinde her bir deney ünitesi, o kümenin merkezi arasındaki uzaklık ölçüleri hesaplanır.
4) Daha sonra deney üniteleri, hangi kümenin merkezine daha yakınsa o kümeye atanır.
5) Bu işlem tekrarlamalar (iterasyonlar) arasında fark kalmayıncaya kadar devam eder.

Daha öncede belirtildiği üzere k-ortalamalar kümelemede ilk aşama küme sayısının (k) belirlenmesidir. Bu sayının ikiden düşük olmaması ve maksimum küme sayısının da çalışılan toplam gözlem sayısı kadar olması beklenir (MacQueen, 1967; Mercer, 2003). Mesela p tane özelliği tespit edilen n tane gözlem değeri olsun ve biz bu n tane gözlemi p tane özelliğinden

yararlanarak k tane kümeye ayırmaya çalışalım. Bu durumda k. küme n_k tane gözlem içerecektir. k-ortalamalar kümeleme yönteminde farklı algoritmalardan kullanılarak kümeleme işlemi gerçekleştirilebilmektedir. Söz konusu algoritmalardan hangisinin daha uygun olduğunu belirlemek amacıyla küme içi kareler toplamına (KIKT) bağlı olarak hesaplanan uyum iyiliği ölçüsü olarak:

KIKT kullanılır. Bu ölçü ise;

$$KIKT_k = \left(\frac{np}{np-m}\right)\sum_{k=1}^{k}\sum_{i=1}^{p}\sum_{j=1}^{n_k}(1-\delta_{ijk})\,(z_{ij}-c_{ik})^2 \qquad \textbf{(2.14)}$$

şeklinde hesaplanır.

Buradaki c_{ik} ; k. kümenin i. değişken bakımından merkezini ya da ortalamasını,

n: gözlem sayısını

p: değişken sayısını

δ_{ijk} : k. gruptaki i.bireyin j. özelliğine ilişkin kayıp gözlem değerini ve

m: her kümenin merkezini göstermektedir.

z_{ij} ise i. değişken bakımından j. bireyin standardize edilmiş değeri olup

$$z_{ij} = \frac{X_i - \overline{X}_i}{S_i} \qquad \textbf{(2.15)}$$

şeklinde hesaplanır.

Dolayısıyla bu durumda her bir küme tarafından açıklanabilen varyasyon ise

$$V_k = \frac{KIKT_k}{KIKT_1}100 \qquad \textbf{(2.16)}$$

şeklinde hesaplanır.

Açıklanan varyasyon aynı zamanda optimum küme sayısının belirlenmesine de yardımcı olur. Bundan yararlanılarak uygun küme sayısı belirlenirken, açıklanan varyasyondaki düşüşün çok belirginleştiği değere karşılık gelen küme sayısı uygun küme

sayısı olarak alınır (Hintze, 2001). Bu yaklaşımın dışında verilerin kaç kümeye ayrılacağı ya da optimum küme sayısının kaç olduğuna ilişkin değişik araştırıcılar tarafından farklı yaklaşımlar geliştirilmiştir. Calinsky ve Harabasz (1974) $C = [iz(B)/k-1]/[iz(W)(n=k)]$ eşitliğini maksimum yapan k değerinin uygun küme sayısı olarak alınabileceğini bildirmiştir. Burada B ve W sırasıyla gruplar arası ve grup içi kareler toplamı matrisleridir (Atamer, 1992).

Günümüzde en yaygın kullanılan yaklaşımlardan birisi de $k = \sqrt{n/2}$ biçiminde belirtilmektedir (Tatlıdil, 1996; Tatlıdil, 2002). Uygun küme sayısının tam olarak belirlenememesi bu k-ortalamalar kümeleme yönteminin en dezavantajı olarak karşımıza çıkmaktadır.

BÖLÜM 3

MATERYAL VE YÖNTEM

3.1. MATERYAL

Sınıflandırma ve Regresyon Ağaçları, Yapay Sinir Ağları ve k-Ortalamalar Kümeleme gibi bazı veri madenciliği yöntemlerinin hayvancılıkla ilgili çalışmalarda kullanımlarını göstermek amacıyla yürütülen bu tez çalışmasının materyalini, Çanakkale Onsekiz Mart Üniversitesi Ziraat Fakültesi Zootekni Bölümünde 2000-2009 yılları arasında Türk Saanen keçileri ve Ross 308 hattından etlik piliçler üzerinde yürütülmüş çalışmalardan elde edilen veriler oluşturmuştur.

Çalışmada, 844 tane Türk Saanen keçisine ilişkin canlı ağırlık (g), vücut kondüsyon skoru (puanı), doğum yıl, doğum ayı, aşım yılı, aşım ayı, dönme sayısı ve dönme aralığı olmak üzere 8 özellik dikkate alınmıştır. Bu özelliklerden canlı ağırlık, vücut kondüsyon skoru ve dönme aralığı sürekli değişkenlerdir. Dönme sayısı sıralanmış (ordinal) nitelikte bir değişkendir. Doğum yılı, doğum ayı, aşım yılı ve aşım ayı ise kategorik değişkenler olarak dikkate alınmıştır. Bu özelliklerden vücut kondüsyon skorları el yordamı ile belirlenmiştir (1 ile 5 puan arasında). 224 tane etlik piliçten ise İncik Genişliği (mm), İncik Uzunluğu (mm), Göğüs Kemiği Uzunluğu (mm), Göğüs Genişliği (mm), Göğüs Çevresi (cm), Vücut Uzunluğu (cm), 2. Hafta Canlı Ağırlık (g), 6. Hafta Canlı Ağırlık (g), Cinsiyet (erkek ve dişi) ve Dönem (Yaz, Kış ve İlkbahar) gibi 11 farklı özelliğe ilişkin veriler dikkati alınmıştır. Bu özelliklerden canlı ağırlıklar hassas terazi alınmıştır. Göğüs çevresi ve vücut uzunluğu mezro ile ölçülürken diğer özellikler dijital kumpasla ölçülmüştür.

3.2. YÖNTEM

Bu çalışmada, Sınıflandırma Ağaçları Yöntemi Türk Saanen keçilerinden elde edilen verilere uygulanırken, Regresyon Ağaçları, Yapay Sinir Ağları ve k-Ortalamalar Kümeleme Yöntemi ise etlik piliçlerden elde edilen verilere uygulanmıştır. Sınıflandırma Ağaçları Yöntemi, Türk Saanen Keçilerinde dönme sayısına etkili olabilecek faktörlerin ya da değişkenlerin belirlenmesi amacıyla kullanılmıştır. Bu amaçla dönme sayısı, bağımlı değişken olarak, bu keçilerin canlı ağırlıkları, vücut kondüsyon puanları, doğum yılı, doğum ayı, aşım yılı, aşım ayı ve dönme aralığı bağımsız değişkenler olarak ele alınarak dikkate alınmıştır.

Regresyon Ağaçları Yöntemi, etlik piliçlere ilişkin 6. hafta canlı ağırlığın tahmin edilmesi ve söz konusu tahminlerin yapılmasında önemli düzeyde etkili olan bağımsız değişkenlerin belirlenmesi amacıyla kullanılmıştır. Bu amaçla 6.hafta canlı ağırlık bağımlı değişken olarak, İncik Genişliği, İncik Uzunluğu, Göğüs Kemiği Uzunluğu, Göğüs Genişliği, Göğüs Çevresi, Vücut Uzunluğu, 2. Hafta Canlı Ağırlık ve Cinsiyet ise bağımsız değişkenler olarak dikkate alınmıştır. Yapay Sinir Ağları Yöntemi, söz konusu piliçlerin cinsiyetlerine göre doğru sınıflandırılıp sınıflandırılmadığının belirlenmesinde yani sınıflandırma yapmak amacıyla kullanılmıştır. Bu amaçla Cinsiyet bağımlı değişken olarak, diğer değişkenler ise bağımsız değişkenler olarak dikkate alınmıştır. k-Ortalamalar Kümeleme Yöntemi ise bu piliçlerin tespit edilen özelliklerinden yararlanılarak bunların kümelere ayrılması ve söz konusu kümeleme işleminde önemli düzeyde etkili değişkenlerin belirlenmesi amacıyla kullanılmıştır.

BÖLÜM 4

BULGULAR ve TARTIŞMA

Regresyon ve Sınıflandırma Ağaçları, Yapay Sinir Ağları ve k-Ortalamalar Kümeleme gibi farklı veri madenciliği yöntemlerinin hayvancılıkla ilgili çalışmalara uygulanmasından elde edilen bulgular Şekil 3-10 ve Çizelge 1-13'da topluca verilmiştir.

4.1. Sınıflandırma Ağacı Yöntemine İlişkin Sonuçlar

Türk Saanen keçilerinden elde edilen veriler kullanılarak dönme sayısına etkili olabilecek faktörlerin belirlenmesi amacıyla yapılan sınıflandırma ağaçları sonuçları Şekil 3, Şekil 4, Çizelge 1-4'te verilmiştir. Söz konusu analizler yapılırken önce veriler training (deney ya da eğitim) ve test olmak üzere iki gruba ayrılmıştır. Training veri grubu üzerinden analizler yapılmıştır. Test veri grubu üzerinden de uygulanan eğitimin başarılı olup olmadığı test edilmiştir. Training veri grubuna bağlı olarak elde edilen sonuçlar Şekil 3'de, test veri grubu üzerinden yapılan analizler ise Şekil 4'de verilmiştir.

Sınıflandırma ağacı incelendiğinde (Şekil 3), keçilerin dönme sayılarına göre ≤ 22 gün, 22-41 gün, > 41 gün olmak üzere üç alt gruba ayrıldığı görülür. Dolayısıyla dönme sayısının tahmin edilmesinde 1.derecede etkili olan değişkenin ya da en fazla etkileyen değişkenin dönme aralığı olduğu söylenebilir.

Dönme aralığı ≤ 22 gün olan keçilerinin büyük bir kısmının (% 90,7) bir kez dönme göstereceği beklenirken, dönme aralığı 22-41 gün olan keçilerin genel olarak 1 (% 52,4) ya da 2 (% 47,6) defa dönme gösterebileceği beklenmektedir. Diğer taraftan dönme aralıkları 41 günden daha fazla olan keçilerin dönme aralıklarının yanında bir de aşım yıllarına göre alt gruplara ayrıldıkları görülmektedir. Dolayısıyla, dönme aralıkları 41 günden daha fazla olan keçilerin dönme sayılarına ilişkin güvenilir tahminlerde bulunabilmek için bunların dönme aralıklarının yanında bir de aşım yıllarının da dikkate alınması gerekmektedir.

Dönme aralıkları 41 günden fazla olan keçilerden 2002, 2003 ve 2005 yıllarında aşım yapılanların büyük bir kısmının 4 ya da 5 kez (% 33,3) dönmesi beklenirken, 2004, 2006 ve 2007 yıllarında aşım yapılan keçilerin ise büyük çoğunluğunun iki kez dönmesi (% 48,1) beklenmektedir.

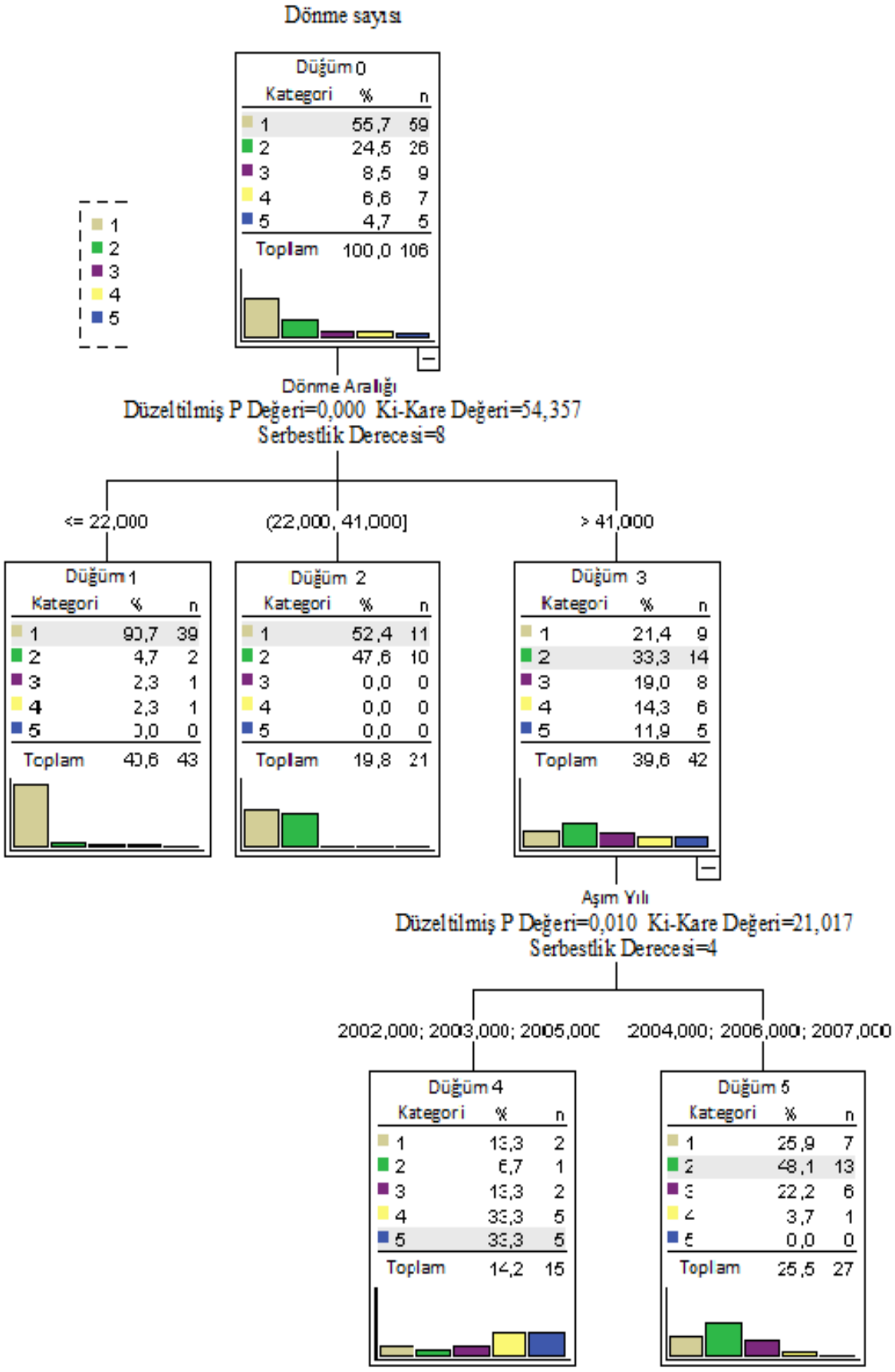
Şekil 3'deki ağaç yapısından yararlanılarak yapılacak olan tahminlerdeki doğruluk derecesi ise % 64,2 olarak belirlenmiştir (Çizelge 3). Test veri grubu üzerinden oluşturulan

ağaç yapısı (Şekil 4) incelendiğinde; training veri grubundaki gibi keçilerin dönme sayılarına göre ≤ 22 gün, 22-41 gün, > 41gün olmak üzere üç alt gruba ayrıldığı görülmektedir.

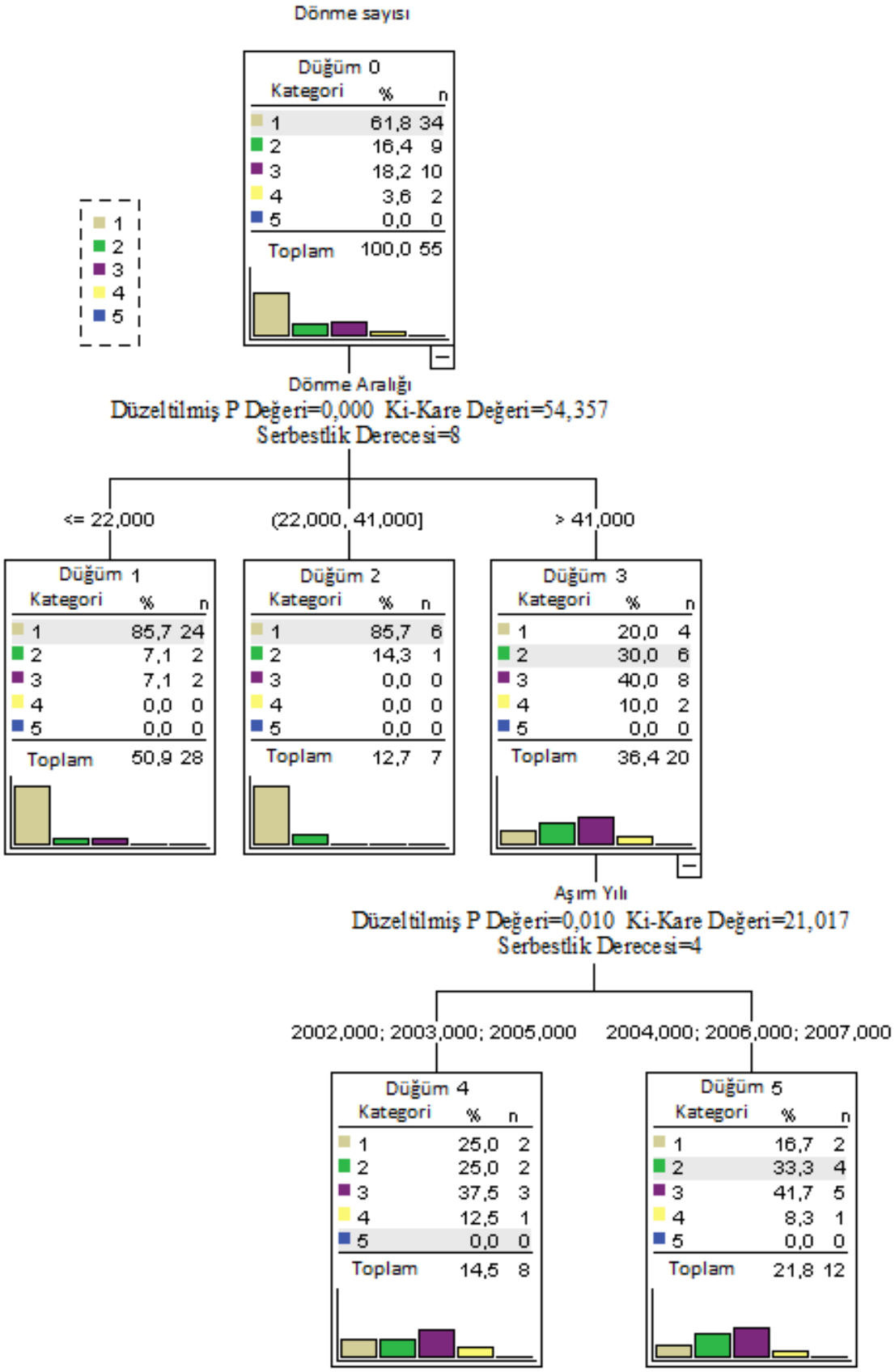
Dönme aralığı ≤ 22 gün olan keçilerinin büyük bir kısmının (% 85,7) bir kez dönme göstermesi beklenirken, dönme aralığı 22-41 gün olan keçilerin büyük bir çoğunluğunun 1 (% 85,7) defa dönme gösterebileceği beklenmektedir. Diğer taraftan dönme aralıkları 41 günden daha fazla olan keçilerin dönme aralıklarının yanında bir de aşım yıllarına göre alt gruplara ayrıldıkları görülmektedir. Dolayısıyla, dönme aralıkları 41 günden daha fazla olan keçilerin dönme sayılarına ilişkin güvenilir tahminlerde bulunabilmek için bunların dönme aralıklarının yanında bir de aşım yıllarının da dikkate alınması gerekmektedir. Dönme aralıkları 41 günden fazla olan keçilerden 2002, 2003 ve 2005 yıllarında aşım yapılanların büyük bir çoğunluğunun 3 kez (% 37,5) dönmesi beklenirken, 2004, 2006 ve 2007 yıllarında aşım yapılan keçilerin ise büyük çoğunluğunun 2 (% 33,3) ya da 3 (% 41,7) kez dönmesi beklenmektedir.

Şekil 4'deki ağaç yapısından yararlanılarak yapılacak olan tahminlerdeki doğruluk derecesi % 61,8 olarak belirlenmiştir (Çizelge 3).

Eğitim (Training) ve test veri grupları üzerinden oluşturulan ağaçlar (Şekil 3 ve Şekil 4) birlikte değerlendirildiğinde genel olarak bir uyumun olduğu görülür. Dolayısıyla elde edilen sonuçlar dikkate alınan deneme koşulları için güvenilir ve genelleştirilebilir sonuçlar olarak kabul edilebilir.



Şekil 3. Eğitim (Training) Veri Grubu Üzerinden Oluşturulan Sınıflandırma Ağacı.



Şekil 4. Test Veri Grubu Üzerinden Oluşturulan Sınıflandırma Ağacı.

Çizelge 1. Ağaç modelindeki düğümlerdeki (node) gözlem sayıları ve yüzdeleri

Örnek	Düğüm (Node)	1,00		2,00		3,00		4,00		5,00		Toplam	
		N	%	N	%	N	%	N	%	N	%	N	%
Eğitim	0	59	55,7	26	24,5	9	8,5	7	6,6	5	4,7	106	100,0
	1	39	90,7	2	4,7	1	2,3	1	2,3	0	0,0	43	40,6
	2	11	52,4	10	47,6	0	0,0	0	0,0	0	0,0	21	19,8
	3	9	21,4	14	33,3	8	19,0	6	14,3	5	11,9	42	39,6
	4	2	13,3	1	6,7	2	13,3	5	33,3	5	33,3	15	14,2
	5	7	25,9	13	48,1	6	22,2	1	3,7	0	0,0	27	25,5
Test	0	34	61,8	9	16,4	10	18,2	2	3,6	0	0,0	55	100,0
	1	24	85,7	2	7,1	2	7,1	0	0,0	0	0,0	28	50,9
	2	6	85,7	1	14,3	0	0,0	0	0,0	0	0,0	7	12,7
	3	4	20,0	6	30,0	8	40,0	2	10,0	0	0,0	20	36,4
	4	2	25,0	2	25,0	3	37,5	1	12,5	0	0,0	8	14,5
	5	2	16,7	4	33,3	5	41,7	1	8,3	0	0,0	12	21,8

Çizelge 2. Ağaç modelindeki düğümlerin önemlilik düzeyi

Örnek	Düğüm	Tahmin Edilen Kategori	Ana Düğüm	Önemli Olan Bağımsız Değişkenler		
				Değişken	p	$\chi 2$
Eğitim	0	1				
	1	1	0	Dönme Aralığı	0,000	54,357
	2	1	0	Dönme Aralığı	0,000	54,357
	3	2	0	Dönme Aralığı	0,000	54,357
	4	5	3	Aşım Yılı	0,010	21,017
	5	2	3	Aşım Yılı	0,010	21,017
Test	0	1				
	1	1	0	Dönme Aralığı	0,000	54,357
	2	1	0	Dönme Aralığı	0,000	54,357
	3	2	0	Dönme Aralığı	0,000	54,357
	4	5	3	Aşım Yılı	0,010	21,017
	5	2	3	Aşım Yılı	0,010	21,017

Çizelge 3. Ağaç modelindeki düğümlerin doğru sınıflandırma oranları

Örnek	Gözlenen	Tahmin Edilen	
		5	Doğruluk (%)
Eğitim	1	2	84,7
	2	1	50,0
	3	2	0,0
	4	5	0,0
	5	5	100,0
	Genel Doğru Sınıflandırma	14,2	64,2
Test	1	2	88,2
	2	2	44,4
	3	3	0,0
	4	1	0,0
	5	0	
	Genel Doğru Sınıflandırma	14,5	61,8

Çizelge 4. Ağaç modelindeki düğümlerin risk değerleri ve standart hatası

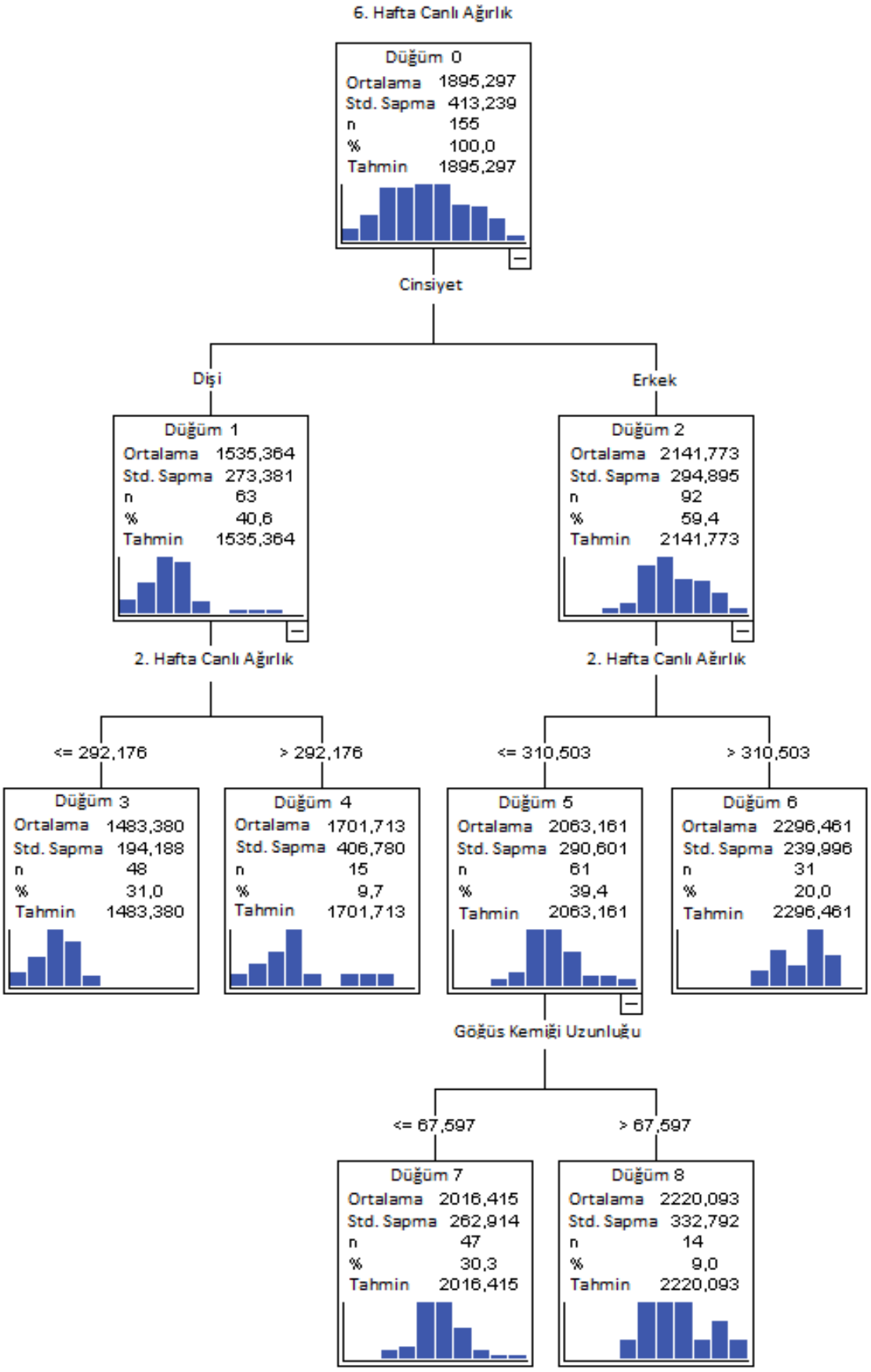
Örnek	Tahmin	Standart Hata
Eğitim	0,358	0,047
Test	0,382	0,066

4.2. Regresyon Ağacı Yöntemine İlişkin Sonuçlar

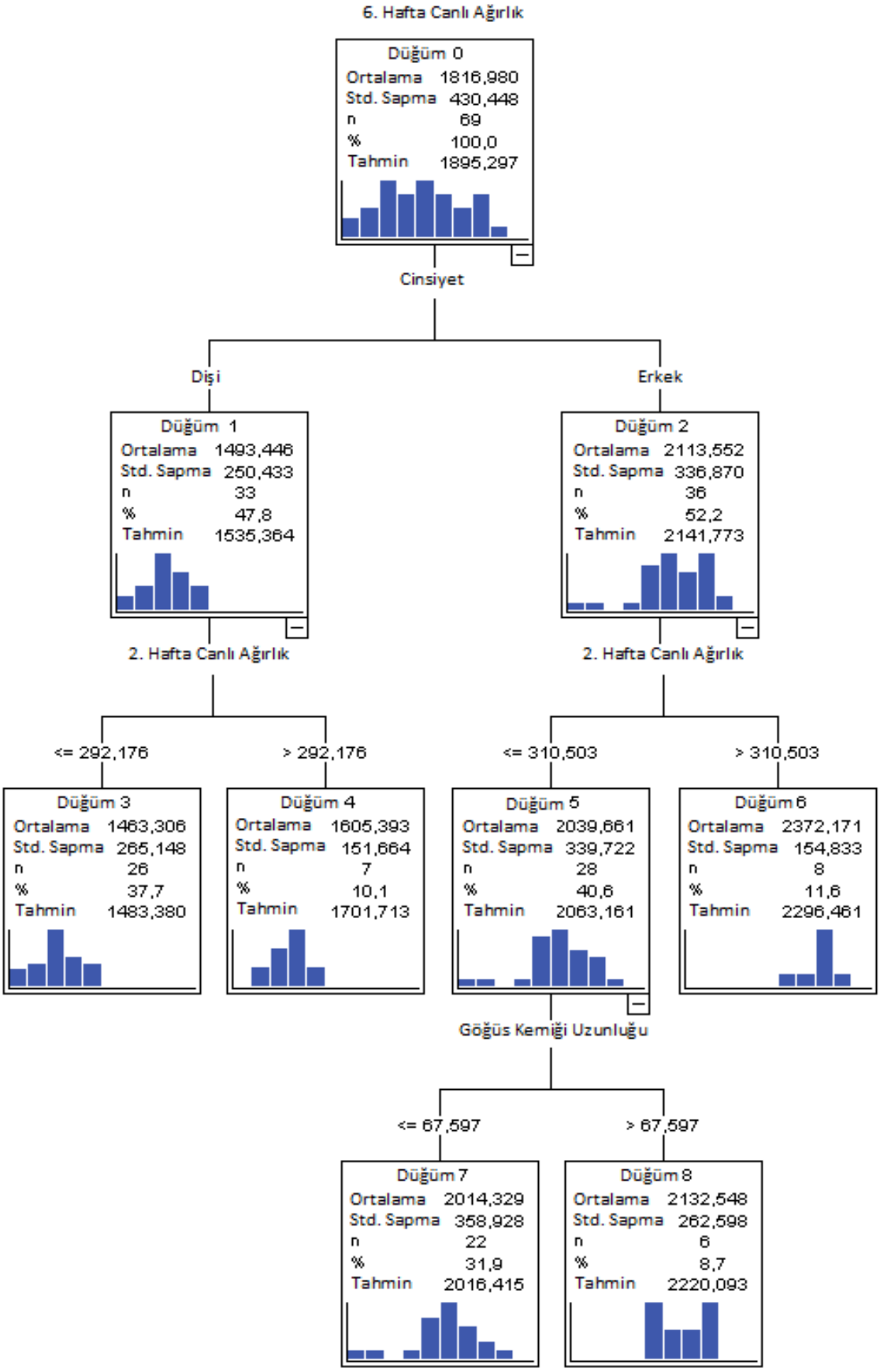
Kesim ağırlığına (6. hafta canlı ağırlığı) etkili olabilecek faktörlerin belirlenmesi amacıyla yapılan regresyon ağaçları sonuçları Şekil 5-8 ve Çizelge 5-6'te verilmiştir. Söz konusu analizler yapılırken önce veriler deney ya da eğitim (training) ve test olmak üzere iki gruba ayrılmıştır. Eğitim veri grubuna bağlı olarak elde edilen sonuçlar Şekil 3 de, test veri grubu üzerinden yapılan analizler ise Şekil 4 de verilmiştir.

Eğitim veri grubu üzerinden oluşturulan Regresyon ağacı incelendiğinde (Şekil 5), etlik piliçlerin kesim ağırlıklarının ilk önce cinsiyetlerine göre alt gruplara ayrıldığı görülmektedir. Dolayısıyla 6. hafta canlı ağırlığın tahmin edilmesinde 1. derecede etkili olan değişkenin cinsiyet olduğu söylenebilir. Dikkat edileceği üzere gerek erkekler gerekse de dişiler aynı zamanda 2. hafta canlı ağırlıklarına göre yeniden alt gruplara ayrılmışlardır. Bu durum söz konusu hayvanların sadece cinsiyetlerinden yararlanılarak kesim ağırlıklarının (6. hafta canlı

ağırlığı) tahmin edilmesinin yeterli olamayacağının bir göstergesidir. Dolayısıyla kesim ağırlıklarına ilişkin yapılacak tahminlerde 2. derecede etkili olan değişken, hayvanların 2. hafta canlı ağırlıklarıdır. Dişi hayvanlardan 2. hafta canlı ağırlıkları 292,176 g'dan daha fazla olan hayvanların kesim ağırlıklarının (1701,713 g), 2. hafta canlı ağırlıkları 292,176 g'dan daha düşük olan hayvanlardan (1483,380 g) daha yüksek olması beklenmektedir. Düğüm 3 ve Düğüm 4 olarak adlandırılan bu iki alt grup artık olabildiğince homojen dişi hayvanlardan oluşan alt gruplar oldukları için bunlar terminal düğüm olarak adlandırılır. Diğer yandan erkek hayvanların kesim ağırlıklarına ilişkin güvenilir tahminlerde bulunabilmek için 2. hafta canlı ağırlıklarının yanında birde bunların göğüs kemiği uzunluklarının dikkate alınması gerekir. Bu durum özellikle 2. hafta canlı ağırlıkları 310,503 g'dan daha düşük olan hayvanlar için geçerlidir. Çünkü 2. hafta canlı ağırlıkları 310,503 g'dan daha fazla olan hayvanlarda artık herhangi bir yeni bölünme ya da alt gruba ayrılma söz konusu değilken, 2. hafta canlı ağırlıkları 310,503 g'dan daha düşük olan hayvanların aynı zamanda göğüs kemiği uzunluklarına göre yeniden iki alt gruba ayrılmışlardır. 2. hafta canlı ağırlıkları 310,503 g'dan daha fazla olan erkek piliçlerin kesim ağırlıkları 2296,461 g olması beklenmektedir. 2. hafta canlı ağırlıkları 310,503 g'dan daha düşük olan hayvanlardan göğüs kemiği uzunlukları 67,597 mm den daha fazla olanların kesim ağırlıklarının (2220,093 g), göğüs kemiği uzunlukları 67,597 mm den daha kısa olanlardan (2016,415 g) daha fazla olması beklenmektedir. Dolayısıyla etlik piliçlerin kesim ağırlıklarına ilişkin yapılacak tahminlerde başta bunların cinsiyetleri olmak üzere 2. hafta canlı ağırlıklarının ve göğüs kemiği uzunluklarının dikkate alınması gerektiği sonucuna varılabilir. Bu durum Şekil 5'dan da kolayca görülebilmektedir. Kesim ağırlığının tahmin edilmesi ve dolayısıyla da yapılacak tahminlerde önemli etkiye sahip değişkenlerin belirlenmesi amacıyla oluşturulan ağacın uygun ya da yeterli olup olmadığının belirlenmesi amacıyla kriter olarak kullanılan R^2=0,64 olarak bulunmuştur. Dolayısıyla bu üç değişkenden yararlanılarak kesim ağırlığına ilişkin varyasyonun % 64'lük bir kısmı açıklanabilmektedir. Bu değer ise uygulama için genel olarak kabul edilebilecek bir değerdir. Bu değer aynı zamanda oluşturulan ağacın yeterli olabileceğinin de bir göstergesidir. Eğitim ve test veri grupları üzerinden oluşturulan ağaçlar (Şekil 5 ve Şekil 6) birlikte değerlendirildiğinde genel olarak bir uyumun olduğu görülür. Hem eğitim hem de test veri grupları üzerinden oluşturulan ağaçların uyumluluğunun bir ölçüsü olarak kullanılan Şekil 7'deki iki ayrı grafiğin birbirine oldukça benzerlik göstermesi de modelin doğruluğunun bir göstergesidir. Dolayısıyla oluşturulan ağaç uygun ağaç olarak kabul edilebilir nitelikte olup, bu ağaca bağlı olarak elde edilen sonuçlar dikkate alınan deneme koşulları için güvenilir ve genelleştirilebilir sonuçlar olarak kabul edilebilir.



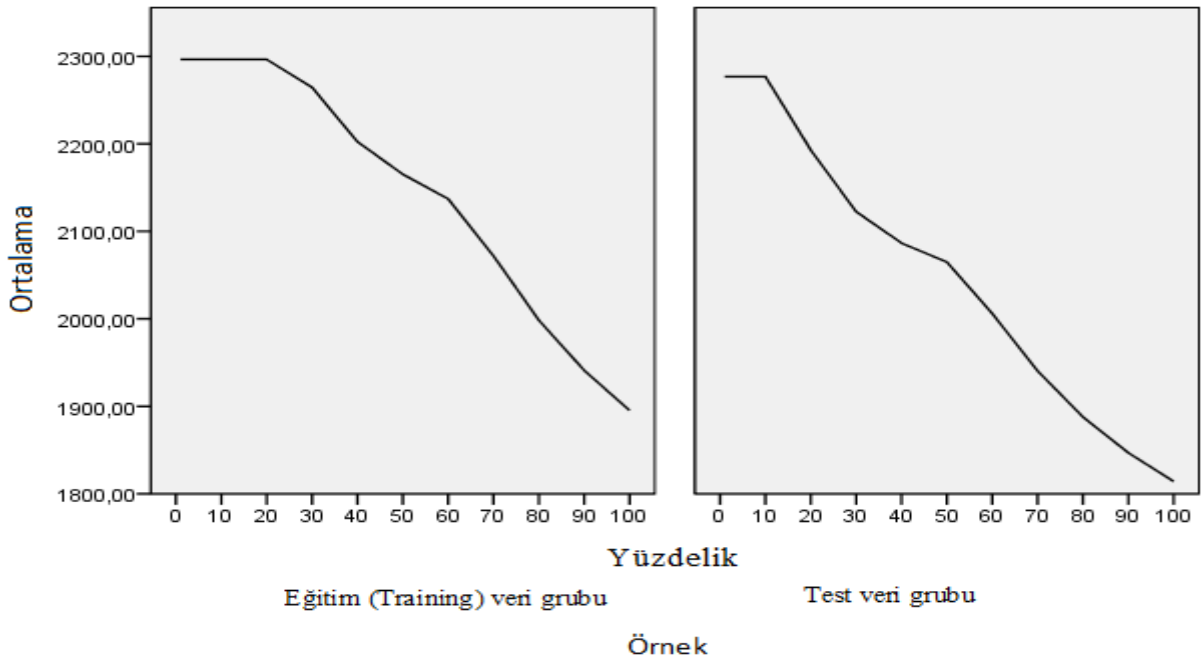
Şekil 5. Eğitim (Training) Veri Grubu Üzerinden Oluşturulan Regresyon Ağacı.



Şekil 6. Test Veri Grubu Üzerinden Oluşturulan Regresyon Ağacı.

Çizelge 5. Ağaç modelindeki düğümlerin gözlem sayısı, yüzdelik oranları ve ortalamaları

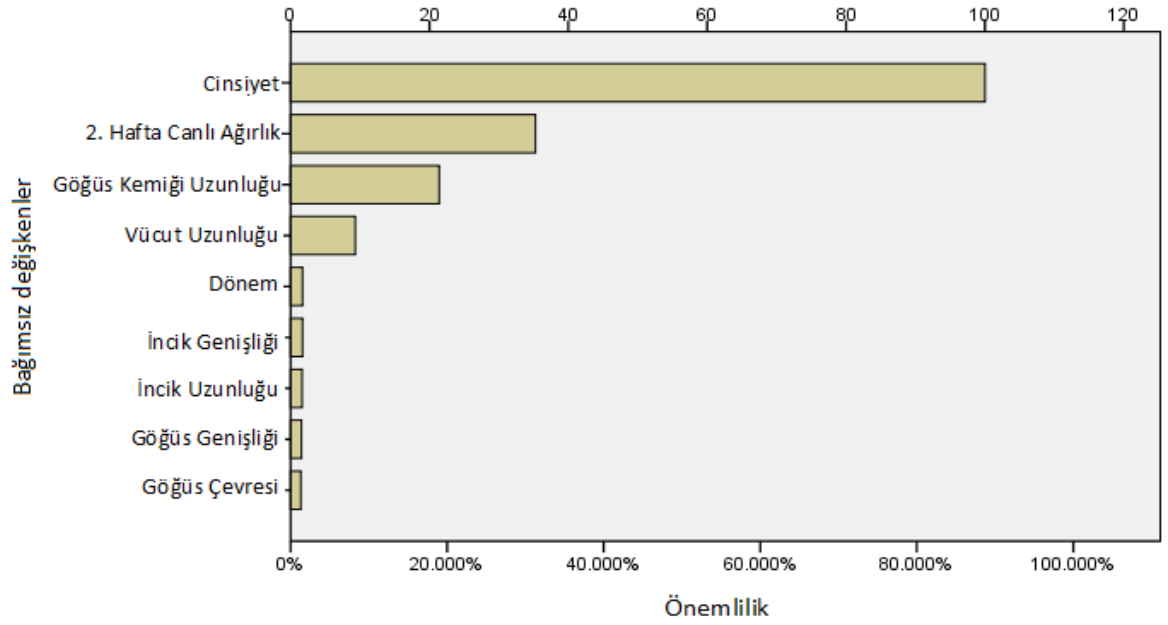
Örnek	Düğüm	N	%	$\overline{X}$
Eğitim	6	31	20,0	2296,461
	8	14	9,0	2220,093
	7	47	30,3	2016,415
	4	15	9,7	1701,713
	3	48	31,0	1483,380
Test	6	8	11,6	2372,171
	8	6	8,7	2132,548
	7	22	31,9	2014,329
	4	7	10,1	1605,393
	3	26	37,7	1463,306



Şekil 7. Eğitim (Training) ve Test Veri grupları Üzerinden Oluşturulan Regresyon Ağaçlarının Uyumları.

Çizelge 6. Ağaç modelindeki düğümlerin risk değerleri ve standart hatası

Örnek	Tahmin	Standart Hata
Eğitim	67330,889	9425,698
Test	76535,767	19175,942



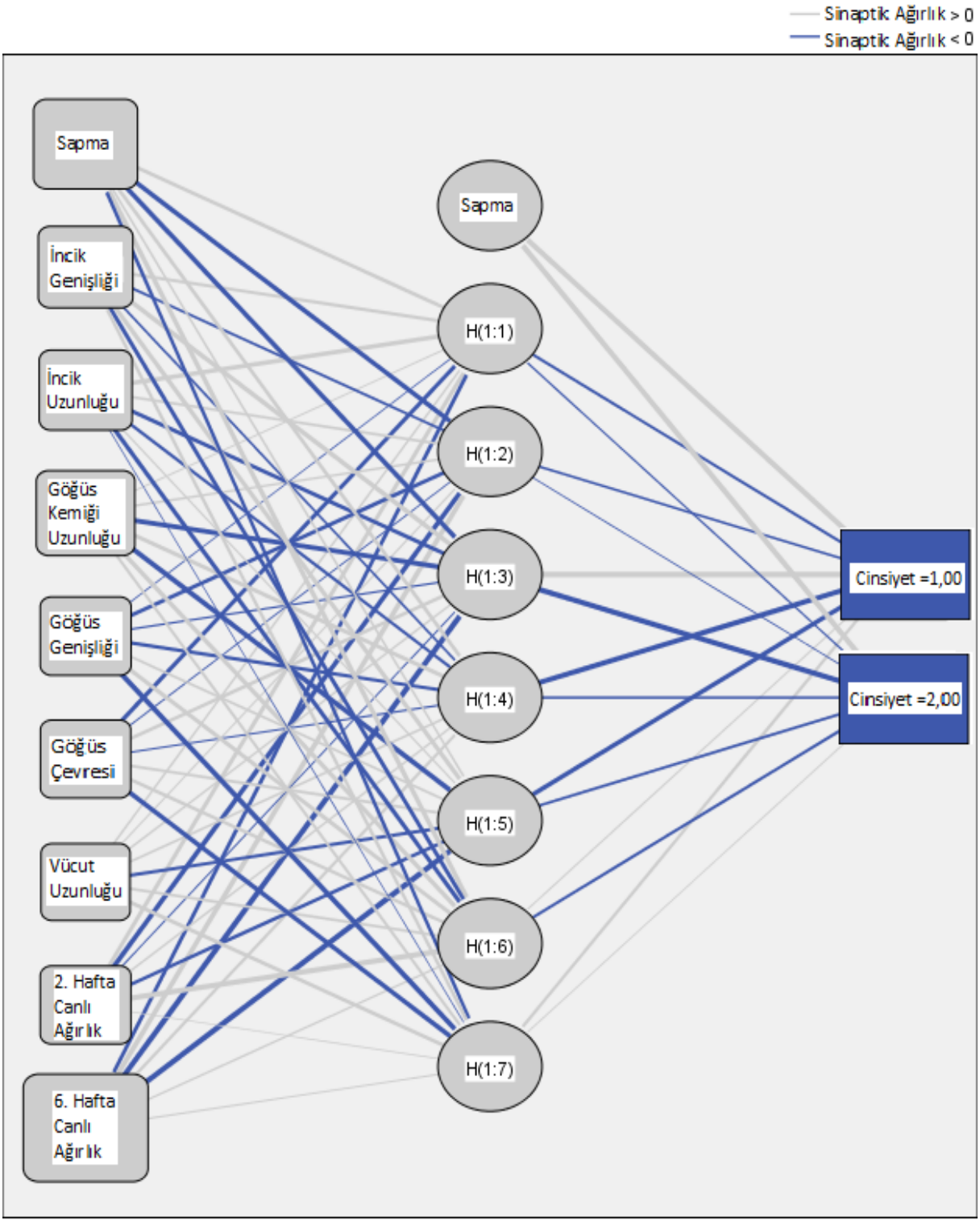
Şekil 8. Regresyon Ağacına Giren Değişkenlerin Önemlilik Düzeylerine Göre Sıralanmaları.

4.3. Yapay Sinir Ağları yöntemine ilişkin sonuçlar

Bu tez çalışmasında YSA yöntemi; 2. ve 6. hafta canlı ağırlık, incik genişliği, incik uzunluğu, göğüs kemiği uzunluğu, göğüs genişliği, göğüs çevresi, vücut uzunluğu ve cinsiyet olmak üzere 9 farklı özelliği tespit edilmiş etlik piliçlerin bu özelliklerden yararlanılarak cinsiyetlerinin doğru sınıflandırılıp sınıflandırılmadığının belirlenmesinde kullanılmıştır. Yani YSA, sınıflandırma ağaçları yönteminde olduğu gibi sınıflandırma yapmak amacıyla kullanılmıştır.

Bu amaçla cinsiyet değişkeni bağımlı değişken, diğer değişkenler ise girdi (input) ya da bağımsız değişkenler olarak modele dahil edilmiştir. 224 etlik piliçten 163 tanesine ilişkin veriler eğitim (training), 44 tanesi test ve 17 tanesi de geçerlilik testi (holdout ya da validatiton) amacıyla kullanılmıştır.

Bu çalışma için oluşturulan ağ yapısı Şekil 9'da verilmiştir. Şekil 9'dan da görüleceği üzere girdi değişken sayısının artması, oluşturulacak ağında daha karmaşık bir hale gelmesine neden olabilmektedir. Bunun sonucunda da hem yorumlama aşamasında sıkıntı çekilebilmekte hem de oluşturulan ağ yapısının uygun ağ olup olmadığı konusunda çelişkilere düşülebilmektedir.



Şekil 9. Oluşan ağ yapısı.

Öğrenme ya da eğitilme işleminde başarı kriteri olarak ise YSA'nın öğrenme ve tahmin yapabilme kalitesinin bir göstergesi olan hata kareler ortalaması (HKO) ve hatalı sınıflandırma %'si kullanılmıştır.

Eğitim seti için HKO=10.010 ve hatalı sınıflandırma oranı (%) 6,7 olarak bulunmuştur. Test veri grubunda ise HKO=3,496 ve hatalı sınıflandırma yüzdesi ise 6,8 olarak bulunmuştur. Eğitim ve test veri gruplarına ilişkin elde edilen bu bulguların birbirlerine yakın olmaları, YSA'nın yeterince eğitildiğin bir göstergesidir. Oluşturulan YSA'nın geçerli bir ağ olup olmadığının belirlenmesi amacıyla 17 veriden oluşan Holdout veri grubunda hatalı sınıflandırma yüzdesinin 5,9 olarak bulunmuş olması, söz konusu ağın yeterli ya da geçerli bir ağ olduğunu göstermektedir (Çizelge 7).

Çizelge 7. Eğitim, test ve gerçeklik testin hata kareler ortalaması ve hatalı sınıflandırma oranı

	HKO	Hatalı sınıflandırma oranı (%)
Eğitim	10,010	6,7
Test	3,496	6,8
Geçerlilik Testi		5,9

Söz konusu YSA'dan yararlanılarak piliçlerin cinsiyetlerine göre sınıflandırılmasına ilişkin sonuçlar Çizelge 8 de verilmiştir. Çizelge 8 incelendiğinde Eğitim (training), Test ve Geçerlilik (holdout) veri gruplarına ilişkin doğru sınıflandırma olasılıklarının birbirlerine oldukça yakın olduğu görülür. Söz konusu sınıflandırma olasılıklarının birbirlerine yakın olması, oluşturulan YSA'nın 224 etlik pilicin 8 özelliğinden yararlanılarak erkek ve dişi olarak sınıflandırma da oldukça etkili olduğu ve dolayısıyla güvenilirliği yüksek sonuçlar verdiğini göstermektedir. Kısaca söz konusu ağ yapısı uygun bir ağ yapısıdır.

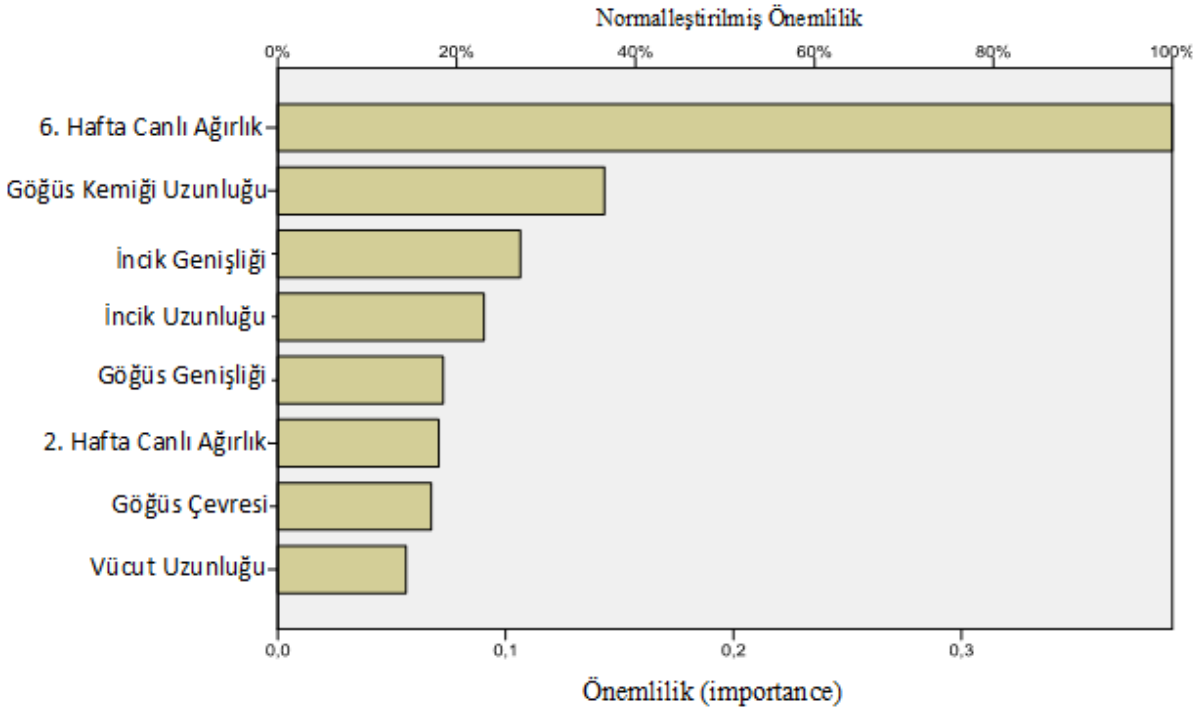
Çizelge 8. YSA'da ki yapılan testlere göre cinsiyetin doğru sınıflandırma oranları

Örnek	Gözlem	Tahmin		
		Dişi	Erkek	Doğru sınıflandırma oranı (%)
Eğitim	Dişi	64	3	95,5
	Erkek	8	88	91,7
	Doğru sınıflandırma oranı (%)	44,2	55,8	93,3
Test	Dişi	20	2	90,9
	Erkek	1	21	95,5
	Doğru sınıflandırma oranı (%)	47,7	52,3	93,2
Geçerlilik Testi	Dişi	7	0	100,0
	Erkek	1	9	90,0
	Doğru sınıflandırma oranı (%)	47,1	52,9	94,1

Oluşturulan YSA'nın söz konusu etlik piliçleri erkek ve dişi olarak ayırmasında önemli düzeyde etkili olan değişkenlerin önem sıralaması ise Çizelge 9 ve Şekil 10 da verilmiştir. Şekil 10 incelendiğinde piliçlerin cinsiyetlerine göre sınıflandırmasında 1.derecede etkili olan değişkenin 6.hafta canlı ağırlıkları olduğu görülür. 6. Hafta Canlı Ağırlığı, Göğüs Kemiği uzunluğu, İncik Genişliği ve İncik Uzunluğu olduğu görülür. Göğüs Genişliği, 2. Hafta Canlı Ağırlığı, Göğüs Çevresi ve Vücut Uzunluğu ise piliçlerin cinsiyetlere ayrılmasında diğer değişkenler kadar etkili olmadıkları görülmektedir.

Çizelge 9. Bağımsız değişkenlerin önemlilik değeri ve yüzdesi

Bağımsız Değişkenler	Önemlilik değeri (importance value)	%
İncik Genişliği	0,106	27,1
İncik Uzunluğu	0,091	23,1
Göğüs Kemiği Uzunluğu	0,143	36,5
Göğüs Genişliği	0,073	18,5
Göğüs Çevresi	0,067	17,2
Vücut Uzunluğu	0,056	14,3
2. Hafta Canlı Ağırlık	0,071	18,0
6. Hafta Canlı Ağırlık	0,393	100,0



Şekil 10. Yapay Sinir Ağına Giren Değişkenlerin Önemlilik Düzeylerine Göre Sıralanmaları.

Şekil 10'da görüldüğü gibi etlik piliçlerin cinsiyetlerine göre sınıflandırmasında 1.derecede etkili olan değişkenin 6. hafta canlı ağırlıkları yani kesim ağırlığı olduğu görülür. Bunu Göğüs Kemiği Uzunluğu, İncik Genişliği, İncik Uzunluğu izlemektedir. Göğüs Genişliği, 2. Hafta Canlı Ağırlık, Göğüs Çevresi ve Vücut Uzunluğu ise piliçlerin cinsiyetlere ayrılmasında diğer değişkenler kadar etkili olmadıkları görülmektedir. Değişkenlerin bu önemlilik düzeylerine göre sıralanmasında normalleştirilmiş önemlilik kullanılmıştır. Normalleştirilmiş önemlilik değişkenlerin (X-min)/(max-min) formülü ile yeniden ölçeklendirilmesidir. Normalleştirilmiş veriler 0 ile 1 arasında değişmektedir.

Çizelge 10. Değişkenlerin parametre tahminleri

Giriş Verileri	Gizli Tabaka							
	H (1:1)	H (1:2)	H (1:3)	H (1:4)	H (1:5)	H (1:6)	H (1:7)	Sapma (Bias)
Sapma (Bias)	0,265	-0,557	-0,623	0,267	0,139	0,558	-0,260	
İncik Genişliği	0,191	-0,157	0,449	-0,143	0,420	-0,368	0,215	
İncik Uzunluğu	0,396	0,155	-0,296	-0,179	0,092	-0,398	-0,017	
Göğüs Kemiği Uzunluğu	0,055	0,132	-0,616	0,302	-0,633	0,416	0,175	
Göğüs Genişliği	-0,024	-0,285	-0,105	-0,224	0,145	0,360	-0,689	
Göğüs Çevresi	-0,391	-0,045	0,226	-0,092	0,166	0,213	-0,525	
Vücut Uzunluğu	0,114	0,090	0,111	0,143	-0,210	0,161	0,283	
2. Hafta Canlı Ağırlık	0,387	-0,632	-0,073	0,112	-0,257	0,801	0,023	
Çıkış Verileri								
Dişi	-0,168	-0,144	0,861	-0,718	-0,426	0,089	0,193	0,843
Erkek	-0,102	-0,025	-0,825	-0,149	-0,164	-0,177	0,027	1,000

4.4. k-ortalamalar kümeleme yöntemine ilişkin sonuçlar

224 etlik pilicin 2. ve 6. hafta canlı ağırlık, incik genişliği, incik uzunluğu, göğüs kemiği uzunluğu, göğüs genişliği, göğüs çevresi ve vücut uzunluğu olmak üzere 8 özelliği dikkate alınarak bu piliçlerin k-ortalamalar kümeleme yöntemi ile kümelere ayrılması sonucunda elde edilen bulgular Çizelge 11, 12 ve 13 de verilmiştir. Dikkate alınan özelliklerden hangisi ya da hangilerinin kümeleme işleminde önemli etkide bulunduklarının belirlenmesi amacıyla her bir küme bir faktör seviyesi (grup) olarak dikkate alınmış ve tek yönlü varyans analizi tekniği uygulanmıştır (Çizelge 14).

Söz konusu piliçlerin kaç kümeye ayrılacağı yani başlangıçtaki küme sayısının belirlenmesinde her bir küme için küme içi kareler toplamı yüzdesi ya da kısaca varyasyon yüzdesinden yararlanılmıştır (burada sadece 4 küme için varyasyon yüzdeleri rapor edilmiştir). Bu varyasyon yüzdesinde keskin bir düşüşün olduğu küme sayısına karşılık gelen sayı, uygun küme sayısı olarak kabul edilir. Söz konusu varyasyon yüzdeleri incelendiğinde 3. Kümeden 4. Küme giderken kesin bir düşüşün olduğu dikkati çekmektedir. Dolayısıyla bu 224 pilicin söz konusu 8 özelliğinden yararlanılarak 3 kümeye ayrılmasının uygun olduğu sonucuna varılır (Çizelge 11). Oluşturulan 3 kümeye ilişkin tanıtıcı istatistikler ise Çizelge 12'de verilmiştir.

Çizelge 11. Uygun Küme Sayısının Belirlenmesine İlişkin Sonuçlar

Küme Sayısı	Varyasyon Yüzdesi
2	81,03
3	**73,28**
4	56,80

Her bir özelliğin piliçlerin kümelere ayrılmasında önemli etkiye sahip olup olmadıklarının belirlenmesi diğer bir ifade ile piliçlerin kümelere ayrılmasında önemli olan değişkenlerin belirlenmesi amacıyla kümelerin birer grup olarak kabul edilip tek yönlü varyans analizinin uygulanmasıyla elde edilen sonuçlar Çizelge 12'da görülebilir.

Çizelge 12. Tek Yönlü Varyans Analizi Sonuçları

Değişkenler	Gruplar Arası Kareler Ortalaması	Gruplar İçi Kareler Ortalaması	F- değeri	Önemlilik (P)
İncik Genişliği	3,94	2,14	1,84	0,162
İncik Uzunluğu	3,56	3,25	1,09	0,337
Göğüs Kemiği Uzunluğu	**250,73**	**40,20**	**6,24**	**0,002**
Göğüs Genişliği	9,37	9,23	1,01	0,364
Göğüs Çevresi	**14,61**	**2,45**	**5,97**	**0,003**
Vücut Uzunluğu	**764,67**	**123,91**	**6,17**	**0,002**
2. Hafta Canlı Ağırlık	**47667,53**	**3244,34**	**14,69**	**0,000**
6. Hafta Canlı Ağırlık	**1,636E7**	**29238,49**	**559,69**	**0,000**

Not: Koyu renkli yazılan özellikler bu piliçlerin kümelere ayrılmasında istatistiksel olarak önemli etkide bulunan özellikleri göstermektedir.

Çizelge 12 incelendiğinde İncik Genişliği ($p=0.162$), İncik Uzunluğu ($p=0.337$) ve Göğüs Genişliği ($p=0.364$) hariç, diğer 5 özelliğin piliçlerin kümelere ayrılmasında önemli düzeyde etkili oldukları görülür ($P<0.01$).

Çizelge 13. k- ortalama kümeleme yönteminde kümelere ilişkin tanıtıcı istatistikler

Değişkenler	Küme Ortalamaları		
	1. Küme	2. Küme	3. Küme
İncik Genişliği	10,10	10,40	10,56
İncik Uzunluğu	30,48	30,66	30,92
Göğüs Kemiği Uzunluğu	**60,28**	**61,21**	**63,85**
Göğüs Genişliği	35,85	36,57	36,27
Göğüs Çevresi	**14,47**	**15,07**	**15,36**
Vücut Uzunluğu	**97,28**	**99,20**	**103,61**
2. Hafta Canlı Ağırlık	**237,54**	**254,61**	**288,06**
6. Hafta Canlı Ağırlık	**1381,31**	**1830,29**	**2340,69**
n	66	82	76

Çizelge 13 incelendiğinde 1. kümenin 66, 2. kümenin 82 ve 3. kümenin ise 76 piliçten oluştuğu görülür. 1. küme Göğüs Kemiği Uzunluğu, Göğüs Çevresi, Vücut Uzunluğu, 2. Hafta Canlı Ağırlık ve 6. Hafta Canlı Ağırlık özellikleri düşük olan piliçlerin oluşturduğu küme ya da grup iken, 3. küme söz konusu özellikleri en yüksek olan 76 pilicin oluşturduğu kümedir. Dolayısıyla 224 piliç söz konusu özelliklerine göre düşük (1. küme), orta (2. küme) ve yüksek (3. küme) olmak üzere 3 kümeye ayrılmıştır.

Dikkat edileceği üzere İncik Genişliği, İncik Uzunluğu ve Göğüs Genişliği özellikleri bakımından kümeler arasında belirgin bir farklılık bulunmamaktadır. Dolayısıyla bu üç özellik, söz konusu 224 pilicin kümelere ayrılmasında önemli düzeyde etkide bulunmamışlarıdır.

TARTIŞMA

Uygulamada yapılan araştırmaların önemli bir kısmı n bireyden tespit edilen p tane özellik (değişken) arasındaki ilişkilerin araştırılması veya söz konusu özelliklerden yararlanılarak deney ünitelerinin ya da bireylerin sınıflandırılmasına yöneliktir. Bu amaçla genel olarak çoklu regresyon analizi, lojistik regresyon, kümeleme (cluster) analizi ve ayırma (discriminant) analizi gibi klasik istatistik tekniklerinden yararlanılmaktadır. Ancak uygulamada her zaman bu istatistik tekniklerinden yararlanılması mümkün olamamaktadır. Çünkü bu tekniklerden gerek ilişkilerin araştırılmasında gerekse de sınıflandırma yapmak amacıyla yararlanılabilmesi için çalışılan veri gruplarında bir takım ön şartların (normal/çok değişkenli normal dağılım, varyansların homojen olması / varyans-kovaryans matrislerinin homojen olması, bağımsız değişkenler arasında yüksek ilişkilerin bulunmaması (çoklu bağlantı problemi)) yerine getirilmesi gerekmektedir. Ancak uygulamada söz konusu ön şartların yerine gelmediği durumlarla oldukça sık karşılaşılmaktadır. Diğer taraftan adı geçen tekniklerin gerektirdiği ön şartların yerine gelmesi durumunda bile, değişkenler arasındaki ilişkilerin araştırılması ya da sınıflandırma yapılması amacıyla bu yöntemlerden yararlanılması, her zaman araştırıcıların merak ettikleri konular hakkında yeterli bilgi elde etmesine imkan verememekte ve dolayısıyla da araştırıcılar pek çok sorunun cevaplarına ulaşılamamaktadır. Bu durum özellikle büyük ve karmaşık veri setleriyle çalışıldığı durumlarda çok daha belirginleşmektedir.

Özellikle hayvancılık alanında yapılan çalışmalarda genel olarak büyük ve karmaşık veri setleriyle çalışıldığı göz önüne alındığında, ilişki araştırma ya da sınıflandırma amaçlarıyla yürütülen çalışma sonuçlarından elde edilen verilerin istatistiksel analiz aşamasında bahsedilen klasik yöntemlerden etkin bir şekilde yararlanılma imkanının çok fazla olmadığı görülür. İşte bu gibi durumlar için geliştirilen ve özellikle klasik istatistiksel tekniklere göre üstünlükleri giderek daha iyi anlaşılmaya başlanan Sınıflandırma ve Regresyon Ağaçları, Yapay Sinir Ağları ve k-ortalamalar Kümeleme Yöntemi gibi bazı veri madenciliği tekniklerden etkin bir şekilde yararlanılabilir. Bu noktadan hareketle yürütülmüş olan bu tez çalışmasında, Türk Saanen keçileri ve Ross 308 hattından etlik piliçlerde gerek ilişki araştırma ve gerekse de sınıflandırma amacıyla bu yöntemlerden nasıl yararlanılabileceği etraflıca açıklanmıştır. Dikkate alınan dört yöntemden Regresyon Ağaçları yöntemi ilişki araştırma ve tahminlerde bulunma amacıyla kullanılmıştır. Diğer üç yöntem ise sınıflandırma amacıyla kullanılmıştır.

BÖLÜM 5
SONUÇLAR VE ÖNERİLER

Yapılan analizler sonucunda özellikle büyük ve karmaşık veri gruplarıyla çalışılması durumlarında bu yöntemlerden yararlanılarak merak edilen birçok bilgiye ulaşılabildiği görülmüştür. Bu yöntemlerde yüksek seviyeli interaksiyonların (mesela üçlü, dörtlü, ….p_i x p_j seviyeli) etkileri daha etkin bir şekilde ortaya konulabilmiştir. Söz konusu yöntemler aynı zamanda analiz sonuçlarını grafiksel olarak ta sunmaktadırlar. Bu da okuyuculara (özellikle de istatistikçi olmayanlara) elde edilen bulguların daha kolay anlaşılmasına ve yorumlanmasına imkan sağlamaktadır. Bütün bu hususlar dikkate alındığında diğer pek çok alanda olduğu gibi söz konusu veri madenciliği yöntemlerinden hayvancılıkla ilgili çalışmalarda da etkin bir şekilde yararlanılabileceği sonucuna varılabilir.

KAYNAKLAR

Atamer B., 1992. Kümeleme Analizi (Cluster Analysis) ve Kümeleme Analizinin İlaç Sektörüne Uygulanması, *Yayınlanmamış Yüksek lisans Tezi, İstanbul.*

Barber D.G., Gobius N.R., Hannah I.J.C. ve Poppi D.P., 2005. The use of regression tree analysis to identify interactions between on-farm factors affecting milk protein content. *Animal Production in Australia 25, 214.*

Beckman R., Salzman G. ve Stewart C., 1995. Classification and regression trees for bone marrow immunophenotyping. *Cytometry 20: 210-217.*

Berkhin P., 2002. Survey of Clustering Data Mining Techniques, *San Jose, California, USA, Accrue Software Inc.*

Bevilacqua M., Braglia M. ve Montanari R., 2003. The classification and regression tree approach to pump failure rate analysis. *Reliab Eng. Systems Safe. 79: 59-67.*

Breiman L., Friedman J. ve Olshen R.A., 1984. Classification and regression trees. *Wadsworth: Belmont, CA Press.*

Breiman L., Friedman J.H., Olshen R.A. ve Stone, C.J., 1984. Classification and Regression Trees. *Chapman and Hall, Wadsworth Inc., New York, NY, USA.*

Breiman L., Friedman J. H., Olshen R. A., ve Stone CJ. 2003. Classification and Regression Trees. *Boca Raton. Florida: Chapman & Hall.*

Bouloc N. ve Boichard D., 1991. Clasification Of Lactation Curves in Goat. *Journal Of Dairy Science, 74: Supplement 1, 230.*

Calinski R. B. ve Harabasz J., 1974. A dendrite method for cluster analysis. *Communications in Statistics, 3(1): 1-27.*

Çamdeviren H., Mendeş M., Ozkan M., Toros F., Şaşmaz 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.*

Cappelli C., Mola F. ve Siciliano R., 2002. A statistical approach to growing a reliable honest tree. *Computational Statistics & Data Analysis. 38: 285-299.*

Chatterjee S. ve Hadi A. S., (2006). Regression Analysis by Example, (4th Edition). *New York: John Wiley and Sons.*

Chaudhuri P., Lo W.D., Loh W.Y. ve Yang C.C., 1995. Generalized regression trees. *Statistics. Sinica 5: 641-666.*

Chipman HA., George EI. ve Mcculloch RE., 2000. Hierarchical priors for bayesian CART Shrinkage. *Statistics and Computing, 10: 17-24.*

Coşkun S., Coşkun A., Kartal M. ve Bircan H., 2004. Lojistik Regresyon Analizinin İncelenmesi ve Diş Hekimliğinde Bir Uygulaması. *Cumhuriyet Üniversitesi Diş Hekimliği Fakültesi Dergisi Cilt:7 Sayı:1.*

De'ath G ve Fabricius K. 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology, 81: 3178-92.*

Dietterich T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computing. 10: 1895-1923.*

Draper N.R. ve Smith H., (1998). Applied Regression Analysis. 3rd Edition. *John Wiley and Sons, Inc. Canada.*

Elmas Ç., 2003. Yapay Sinir Ağları (Kuram, Mimari, Eğitim, Uygulama). *Seçkin Yayınları: Ankara.*

Fernandez C., Soria E., Sanchez-Seiquer P., Gomez-Chova L., Magdalena R., Martin J.D., Navarro M. J. ve Serrano A. J., 2005. Weekly milk prediction on dairy goats using neural networks. *Expert System with Applications, 14 (2), 305-318.*

Fu C., 2003. Combining loglinear model with classification and regression tree (CART): An application to birth data. *Computational Statistics & Data Analysis, 2003; (Article In Pres).*

Han J. ve Kamber M., 2001. Data Mining Concepts and Techniques, *Morgan Kauffmann Publishers Inc.*

Hartigan J. A. ve M. A. Wong 1979. "Algorithm AS 136: A k-means clustering algorithm". *In: Applied Statistics, 28. 1, s.100-108.*

Heald C.W., Kim T., Sischo W.M., Cooper J.B. ve Wolfgang D.R., 2000. A computerized mastitis decision aid using farm-based records: an artificial neural network approach, *Journal Of Dairy Science. 83. s.711–722.*

Hintze J.L., 2001. NCSS 2001 Statistical System for Windows. *Number Cruncher Statistical Systems. Kaysville, Utah.*

Honeycutt E. ve Gibson G., 2003. Use of regression methods to identify motifs that modulate germline transcription in *Drosophila melanogaster. Genet. Res. 83: 177-188.*

Güneri N. ve Apaydın A., 2004. "Öğrenci Başarılarının Sınıflandırılmasında Lojistik Regresyon Analizi ve Sinir Ağları Yaklaşımı", *Ticaret ve Turizm Eğitim Fakültesi Dergisi, Yıl: 2004, Sayı: 1, s.170 – 188.*

Karalic A., 1992. Linear regression in regression tree leaves. *International School for Synthesis of 624 İtalian Journal of Animal Science. 8, 615-624.*

Kayri M. ve Boysan M., 2008. Bilişsel Yatkınlık İle Depresyon Düzeyleri İlişkisinin Sınıflandırma Ve Regresyon Ağacı Analizi İle İncelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi 34: s.168-177.*

Kirby Y. K., McNew R. W., Kirby J. D. ve Wideman R. F. 1997. Evaluation of logistic versus linear regression models for predicting pulmonary hypertension syndrome (ascites) using cold exposure or pulmonary artery clamp models in broilers. *Poultry* Science, *76: 392–399.*

Krogmeier D., Dzapo V. ve Wassmuth R., 1990. Zusammenhange Zwischen Ausgewahlten Stofwechselparametern Und Leistungsmerkmalen Bei Mastlammern. *Journal of Animal Breeding Genetics, 107: 291-300.*

Kurup P. U. ve Dudani N. K., 2002. "Neural Networks for Profiling Stress History of Clays from PCPT Data.". *Journal of Geotechnical & Geoenvironmental Engineering. July 2002, Vol. 128, Issue 7, s. 569, 11p.*

Lewis R., 2000. An introduction to classification and regression tree (CART) analysis, *Academic Emergency Medicine, Erişim: http://www.saem.org/download/lewis1.pdf.*

Lorr M., 1983. Cluster Analysis for Social Sciences. *San Francisco: Jossey-Bass.*

MacQueen J. B., 1967. Some Methods for Classification and Analysis of Multivariate Observations, *Production Symp. Math. Statistics and Probability (5th), 281–297.*

Mendeş, M., ve Akkartal, E., 2009. Regression tree model in predicting slaughter weight of broiler. *İtalian Journal of Animal Science, 8, 615-624.*

Mendeş M., Yıldırır M., Küçükkebapçı M., Akkartal E., 2008. Regression tree methodology for determining factors affecting actual lactation milk yield in brown-swiss cattle. *Lucrări Ştiinţifice, Seria D, Vol. LI, Zootehnie, 254-258. The 37th International Session of Scientific Communications of the Faculty of Animal Science, Bucharest, Romania.*

Mercer D. P., (2003), "Clustering Large Datasets". Erişim: http://www.stats.ox.ac.uk/~mercer /documents/transfer.pdf. Erişim tarihi: 13.05.2005.

Momdouh R., 2007. Data Preparation For Data Mınıng Usıng SAS, *Mordan Kaufman Publishers is an İmprint of Elseviver. San Francisco, CA.*

Norusis M.J., 1993. SPSS For Windows Release 6.0 Advanced Statistics. *SPSS Inc., Chicago.*

Örekici G. 2004. Sınıflama ve Regresyon Ağaçları, Yüksek Lisans Tezi, *Mersin Üniversitesi Sağlık Bilimleri Enstitüsü, s.85.*

Özbeyaz C, Yıldız M. ve Çamdeviren H. 1999. Türkiye'de Yetiştirilen Çeşitli sığır Irkları Arasındaki Genetik ilişkiler. *Lalahan Hayvancılık Araştırma Enstitüsü Dergisi, 39 (1): 17-32.*

Özekes S., 2003. Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi, No.3, s.65-82.*

Özdemir G., T., Dolgun M.Ö., Şatır U., Deliloğlu S. ve Korkmaz H.E., 2007. 2005 Yılı Öğrenci Seçme Sınavı (ÖSS) Verileri Kullanılarak Öğrenci Profilinin Belirlenmesi, *5. İstatistik Kongresi, Antalya.*

Özdemir G., T., Dolgun, M.Ö., Oğuz D., 2009. Veri Madenciliğinde Yapısal Olmayan Verinin Analizi: Metin ve Web Madenciliği, *İstatistikçiler Dergisi 2 (2009) 48-58.*

Öztemel E. 2003. Artificial Neural Networks. 1st edition. *Papatya Yayıncılık, İstanbul. s.36-40.*

Ribic C. A. ve Miller T. W., 1998. Evaluation of alternative model selection criteria in the analysis of unimodal response curves using CART. *Journal of Applied Statistics; 25(5), 685-698.*

Roush W.B., Kirby Y.K., Cravener T.L. ve Wideman R.F., 1996. Artificial neural network prediction of ascites in broilers. *Poultry Science, 75 (12): 1479-1487.*

Roush W.B., Cravener T.L. Kirby Y.K. ve Wideman R.F. 1997. Probabilistic neural network prediction of ascites in broilers based on minimally invasive physiological factors. *Poultry Science, 76 (11):1513-1516.*

Roush W.B., Dozier W.A ve Branton S.L. 2006. Comparison of gompertz and neural network models of broiler growth. *Poultry Science, 85 (4): 794-797.*

Salle C. T. P., Guahyba A.S., Wald V.B., Silva A.B., Salle F.O. ve Nascemento V.P., 2003. Use of artificial neural networks to estimate production parameters of broilers breeders in the production phase. *British Poultry Science, 44 (2): 211–217.*

Statsoft 2003. Classification and Regression Trees. *Erişim: http://www.statsoft.com/textbook/ scart.html. Erişim Tarihi: 18.03.2003.*

Steinber G.D. ve Colla P., 1995. CART: Tree-structured Non-parametric Data Analysis. *Salford Systems Publishers, San Diego, CA, USA.*

Swift R., 2001. Accelerating Customer Relationship; *Prentice Hall PTR.*

Şen Z. 2004. Yapay Sinir Ağları İlkeleri, *Su Vakfı Yayınları, İstanbul.*

Talmon J. L., 1986. A multiclass nonparametric partitioning algorithm. *Pattern Recognition Letters. 4: 31-38.*

Tatlıdil H., 1996. Uygulamalı Çok Değişkenli İstatistiksel Analiz, *Cem Web Ofset Ltd. Şti. Ankara.*

Tatlıdil H., 2002. Uygulamalı Çok Değişkenli İstatistiksel Analiz, *Akademi Matbaası, s.424. Ankara.*

Temel G., Çamdeviren H. ve Akkuş Z., 2005. Sınıflama Ağaçları Yardımıyla Restless Legs Syndrome (RLS) Hastalarına Tanı Koyma, *İnönü Üniversitesi Tıp Fakültesi Dergisi 12 (2) 111-117.*

Tolon M. ve Tosunoğlu N. G., 2008. Tüketici Tatmini Verilerinin Analizi: Yapay Sinir Ağları ve Regresyon Analizi Karşılaştırması, *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 10 / 2. 247-259.*

Torgo L., 1998. A comparative study of reliable error estimators for pruning regression trees. In: H. Coelho (edition) Proc. of the Iberoamerican Conference on Artificial Intelligence. *Springer-Verlag, Porto, Portugal, Technical Report, s.98-100.*

Tsoukalas L.H. ve Uhrig E.R., 1997. Fuzzy and neural approaches in engineering. *John Wiley & Sons, Inc.*

Ümit F. S., 1995**.** "Kümeleme Analizi: İstihdamın Sektörel Yapısı Açısından Avrupa Ülkelerinin Karşılaştırılması", İ.Ü. *Sosyal Bilimler Dergisi, Cilt: III, Sayı:2, Temmuz, s.50-59.*

Yıldız B., 1999. Finansal Başarısızlığın Öngörülmesinde Yapay Sinir Ağı Kullanımı ve Ampirik Bir Çalışma, Yayımlanmamış Doktora Tezi, T.C. *Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü: Kütahya.*

Yohannes Y. ve Hoddinott J., 2003. Classification and Regression Trees: An introduction. *Erişim: http://www.ifpri.org/themes/mp18/techquid/tg03.pdf. Erişim Tarihi: 10.06.2003.*

Yurtoğlu E., 2005. "Yapay sinir ağları metodolojisi ile öngörü modellemesi: bazı makroekonomik değişkenler için Türkiye örneği", *Uzmanlık Tezi, Devlet Planlama Teşkilatı, Ekonomik Modeller ve Stratejik Araştırmalar Genel Müdürlüğü, Ankara, 3-43.*

EKLER LİSTESİ

Sayfa No

ÇİZELGELER LİSTESİ **Sayfa No**

ŞEKİLLER LİSTESİ **Sayfa No**

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı: Orçun KÜÇÜKOĞLU

Doğum Yeri: KARS

Doğum Tarihi: 27.11.1986

EĞİTİM DURUMU

Lisans Öğrenimi: Çanakkale Onsekiz Mart Üniversitesi Ziraat Mühendisliği

Yüksek Lisans Öğrenimi: Çanakkale Onsekiz Mart Üniversitesi Fen Bilimleri Enstitüsü Zootekni Anabilim dalı

Bildiği Yabancı Diller: İngilizce

BİLİMSEL FAALİYETLER

a) Yayınlar -SCI –Diğer: ____

b) Bildiriler -Uluslararası –Ulusal: ____

c) Katıldığı Projeler: ____

İŞ DENEYİMİ

Çalıştığı Kurumlar ve Yıl: ____

İLETİŞİM

E-posta Adresi: orcun_kucukoglu_19@hotmail.com