T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KARAR AĞAÇLARI VE YAPAY SİNİR AĞLARI KULLANARAK KASKO SİGORTALARINDA RİSK DEĞERLENDİRME

MERVE ŞAHİN

YÜKSEK LİSANS TEZİ
İSTATİSTİK BÖLÜMÜ ANABİLİM DALI
İSTATİSTİK PROGRAMI

DANIŞMAN
PROF. DR. GÜLHAYAT GÖLBAŞI ŞİMŞEK

İSTANBUL, 2018

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

KARAR AĞAÇLARI VE YAPAY SİNİR AĞLARI KULLANARAK KASKO SİGORTALARINDA RİSK DEĞERLENDİRME

Merve ŞAHİN tarafından hazırlanan tez çalışması 29.03.2018 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Bölümü Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof. Dr. Gülhayat Gölbaşı ŞİMŞEK

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

Yıldız Teknik Üniversitesi _____

Doç. Dr. Fatma NOYAN TEKELİ

İstanbul Üniversitesi _____

Prof. Dr. Dilek ALTAŞ

Yıldız Teknik Üniversitesi _____

ÖNSÖZ

Çalışmamın araştırılmasında, yürütülmesinde ve tamamlamasında engin bilgi ve tecrübelerini esirgemeyen değerli hocam Prof. Dr. Gülhayat Gölbaşı Şimşek'e , tüm eğitim hayatım boyunca maddi ve manevi şekilde her zaman yanımda olan sevgili aileme ve arkadaşlarıma bana desteklerinden ötürü sonsuz teşekkürlerimi sunarım.

Mart, 2018

Merve ŞAHİN

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ vii

KISALTMA LİSTESİ viii

ŞEKİL LİSTESİ ix

ÇİZELGE LİSTESİ x

ÖZET xi

ABSTRACT xiii

BÖLÜM 1

GİRİŞ 1

1.1 Literatür Özeti 1
1.2 Tezin Amacı 2
1.3 Hipotez 2

BÖLÜM 2

RİSK ANALİZİ VE SİGORTACILIK 3

2.1 Risk Nedir ? 3
2.2 Risk Analizi 4
2.2.1 Nitel Risk Analizi 4
2.2.2 Nicel Risk Analizi 4
2.3 Sigortacılıkta Risk Analizi 5

BÖLÜM 3

VERİ MADENCİLİĞİ 8

3.1 Veri Tabanlarında Bilgi Keşfi ve Süreci 8
3.1.1 Veri Hazırlama ve Problemi Tanıma 9
3.1.2 Veri Ön İşleme 9
3.1.3 Veri Dönüştürme 9

3.1.4 Veri Madenciliği ... 9
3.1.5 Yorumlama ve Değerlendirme ... 9
3.2 Veri Madenciliği Kavramı ... 10
3.3 Veri Madenciliği Metodolojisi ... 13
3.3.1 İş Hedefi Belirleme ... 13
3.3.2 Veriyi Tanıma ... 14
3.3.3 Veri Hazırlığı ... 14
3.3.4 Modelleme ... 14
3.3.5 Değerlendirme ... 14
3.3.6 Dağıtım ... 15

BÖLÜM 4

VERİ MADENCİLİĞİ TEKNİKLERİ ... 16

4.1 Karar Ağaçları ... 16
4.1.1 Karar Ağaçlarının Oluşumu ... 17
4.1.1.1 Bölünmüş Arama ... 18
4.1.1.2 Ayrıştırma Kriteri ... 18
4.1.1.3 Durdurma ve Budama Kuralı ... 19
4.1.2 Başlıca Karar Ağacı Algoritmaları ... 20
4.1.2.1 CART ... 20
4.1.2.2 CHAID ... 20
4.1.2.3 ID3 ... 21
4.1.2.4 C4.5 ... 21
4.1.3 Karar Ağaçları Neden Kullanışlıdır? ... 21
4.2 Yapay Sinir Ağları ... 22
4.2.1 Yapay Sinir Ağları Bileşenleri ... 23
4.2.2 YSA Varsayımları ... 24
4.2.3 Tek Katmanlı Sinir Ağları ... 24
4.2.4 Çok Katmanlı Sinir Ağları ... 26
4.3 Sınıflandırma Kalitesinin Ölçümü ... 28
4.3.1 Veride Ayrştırma ... 28
4.3.2 Sınıflandırma Performans Ölçütleri ... 29

BÖLÜM 5

UYGULAMA ... 31

5.1 Anakütlenin ve Değişkenlerin Tanımı ... 32
5.1.1 Betimsel İstatistikler ve Veri Dönüşümü ... 35
5.2 Karar Ağaçları Analizi ... 38
5.3 Yapay Sinir Ağları Analizi ... 46
5.4 Karar Ağacı ve Yapay Sinir Ağı Modelinin Karşılaştırılması ... 53

BÖLÜM 6

SONUÇ VE ÖNERİLER ... 55

KAYNAKLAR ... 57

ÖZGEÇMİŞ 61

YAYINLAR 62

SİMGE LİSTESİ

w	Yapay sinir ağı modeli için ağırlıklar
x	Yapay sinir ağı modeli için girdiler
y	Yapay sinir ağı modeli için çıktı

KISALTMA LİSTESİ

CART	Classification and Regression Trees
CHAID	Chi-Squared Automatic Interaction Detector
FN	False Negative
FP	False Positive
IBM	International Business Machines
KDD	Knowledge Discovery From Databases
MIT	Massachusetts Institute of Technology
MLP	Multilayer Perceptron
TN	True Negative
TP	True Positive
YSA	Yapay Sinir Ağları

ŞEKİL LİSTESİ

Sayfa

Şekil 2. 1 Bilgi Keşfi Süreci 9
Şekil 3. 2 Veri Madenciliği 12
Şekil 3. 3 Veri Madenciliği Aşamaları 13
Şekil 4. 4 Karar Ağacı Yapısı 17
Şekil 4. 5 Girdi Seviyesi Göre Olası Ayraç Sayısı 18
Şekil 4. 6 Doğru Ağaç Boyu Seçmek İçin Kullanılan Yaklaşımlar 19
Şekil 4. 7 Yapay Sinir Ağı Yapısı 23
Şekil 4. 8 Tek Katmanlı Sinir Ağı Yapısı 25
Şekil 4. 9 Çok Katmanlı Sinir Ağı Yapısı 26
Şekil 4. 10 Sigmoid Fonksiyonu 27
Şekil 5. 11 SAS Enterprise Miner Programı Arayüzü 32
Şekil 5. 12 Nominal ve İkili Ölçekli Değişkenler İçin Ki-Kare Değerleri 36
Şekil 5. 13 SAS Enterprise Miner Karar Ağaçları Adımları 38
Şekil 5. 14 Karar Ağacı Düğümü Ayrıştırma Özellikleri 39
Şekil 5. 15 Alt Katman Bölme Düğüm Özellikleri 39
Şekil 5. 16 Karar Ağacı Diyagramı 42
Şekil 5. 17 Yapay Sinir Ağı Modeli Kriterleri Ekranı 1 47
Şekil 5. 18 Yapay Sinir Ağı Modeli Kriterleri Ekranı 2 49
Şekil 5. 19 Yapay Sinir Ağı Kümülatif Kaldıraç Grafiği 51
Şekil 6. 20 SAS Enterprise Miner Model Karşılaştırma Adımları 53

ÇİZELGE LİSTESİ

Sayfa

Çizelge 2. 1 Sigorta Branşları 6
Çizelge 4. 2 Sınıflama Matrisi 29
Çizelge 4. 3 Sınıflandıma Performans Ölçütleri 30
Çizelge 5. 4 Değişkenlerin Tanımlanması 33
Çizelge 5. 5 Hedef Değişkenin Ordinal Ölçek Kriterleri 34
Çizelge 5. 6 BOLNUM Değişkeninin Kodlama Şekli 34
Çizelge 5. 7 MEDEDUR Değişkeninin Kodlama Şekli 34
Çizelge 5. 8 CINSIYET Değişkeninin Kodlama Şekli 35
Çizelge 5. 9 Analizde Kullanılan Değişkenlerin Ölçek Tipi Özeti 35
Çizelge 5. 10 Hedef Değişken Seviyelerinin Anakütleye Dağılımı 35
Çizelge 5. 11 Dönüşümü Öncesi Sürekli Değişkenlerin Betimsel İstatistikleri 36
Çizelge 5. 12 Dönüşümü Sonrası Sürekli Değişkenlerin Betimsel İstatistikleri 37
Çizelge 5. 13 Değişkenlerin Önemlilik Dereceleri 40
Çizelge 5. 14 Karar Ağacı Modeli Sınıflandırma Matrisi 41
Çizelge 5. 15 Karar Ağacı Uyum İstatistikleri 46
Çizelge 5. 16 Yapay Sinir Ağı Modeli Sınıflandırma Matrisi 49
Çizelge 5. 17 Yapay Sinir Ağı Skor Sıralaması 50
Çizelge 5. 18 Yapay Sinir Ağıı Uyum İstatistikleri 52
Çizelge 6. 19 Model Karşılaştırma Uyum İstatistikleri 54

ÖZET

KARAR AĞAÇLARI VE YAPAY SİNİR AĞLARI KULLANARAK KASKO SİGORTALARINDA RİSK DEĞERLENDİRME

Merve ŞAHİN

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

Bilişim sistemlerinin gün geçtikçe gelişmesiyle birlikte artık büyük verilerin elde edilmesi ve sistemlerde saklanması kolaylaşmıştır. Sistemlerde saklanan her bir veri, çözüm üretilmek istenen probleme göre uygun analiz yöntemleriyle anlamlandırılabilir. Bu büyük verilerden geleceğe yönelik anlamlı sonuçlar çıkarmaya kısaca Veri Madenciliği (Data Mining) denmektedir. Veri madenciliği günümüzde bankacılık, sigortacılık, telekomünikasyon, borsa, tıp vs. pek çok bilim dalının vazgeçilmez bir parçası haline gelmiştir. Özellikle büyük şirketler, raporlamalarında sık sık veri madenciliğinden faydalanmaktadır.

Veri madenciliği tekniklerinin temelini Bellek Tabanlı Yöntemler (Memory Based Reasoning), Sepet Analizi Tekniği (Market basket analysis), Yapay Sinir Ağları (Neural Networks), Karar Ağaçları (Decision Tree) ve Kümeleme Analizi (Cluster Analysis) oluşturmaktadır. Bu çalışmada sigortacılık sektörünün önemli alanlarından biri olan kasko branşında, özel bir şirketten alınan poliçe bilgileri üzerinden yeni gelen bir müşterinin hasar frekansını tahmin edebilmek için Yapay Sinir Ağları ve Karar Ağaçları yöntemlerinden faydalanılmıştır. Bu yöntemlerden Yapay Sinir Ağları; insan beyninin çalışma mekanizmasını taklit ederek beynin öğrenme, hatırlama genelleme yapma yoluyla yeni bilgiler üretebilme gibi temel işlevleri gerçekleştirmek üzere geliştirilmiştir. Karar ağaçları ise denetimli öğrenimin kullanıldığı veri madenciliği tekniklerindendir. Bu anlamda tahmin edilmesi gereken bir hedef değişken vardır. Bu çalışmada hasar frekansı

sigorta riski olarak kabul edilmiş olup bu hedef değişkeni etkilediği düşünülen bağımsız değişkenler ile SAS yazılımının Enteprise Miner modülü kullanılarak Yapay Sinir Ağları ve Karar Ağaçları yöntemleriyle modelleme çalışması yapılmıştır. Yapay Sinir Ağları ve Karar Ağaçları yöntemleriyle elde edilen modellerin sigorta riskini tahmin performansları karşılatırılmış olup, her ikisi de kabul edilebilir seviyede olmakla birlikte Karar Ağaçları yönteminin tahmin başarısı daha yüksek bulunmuştur.

Anahtar Kelimeler: Veri Madenciliği, Yapay Sinir Ağları, Karar Ağaçları, Sigortacılık

ABSTRACT

RISK ASSESSMENT IN CAR INSURANCE USING DECISION TREES AND ARTIFICIAL NEURAL NETWORKS

Merve ŞAHİN

Department of Statistics

MSc. Thesis

Adviser: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

As the information systems are growing day by day, it becomes easier to obtain bigger data and store it in systems. Each data stored in the systems can be interpreted by analysis methods which are appropriate to the problems desired to be solved. It is called Data Mining producing meaningful results for the future from the huge amount of data. Nowadays data mining has become an indispensable part of many sciences including banking, insurance, telecommunication, stock exchange, medicine etc. Especially large companies often use data mining in their reports.

Data mining techniques are based on Memory Based Reasoning, Market basket analysis, Neural Networks, Decision Tree and Cluster Analysis. In this study, Artificial Neural Networks and Decision Trees methods were used to estimate the damage frequency of a new customer based on policy information from a private company in the auto insurance branch, which is one of the important areas of the insurance industry. Artificial Neural Networks has been developed to realize brain's functions such as learning, producing new data by remembering and generalizing through imitating human brain's working mechanism. Decision trees are data mining techniques where supervised learning is used. In this sense there is a target variable to be estimated. In this study, damage frequency was accepted as insurance risk, and using independent variables affecting this target variable, insurance risk was modeled by Artificial Neural Networks and Decision Trees methods of SAS software using Enterprise Miner module. Performances of the of the models obtained by Artificial Neural Networks and Decision

Trees methods were at acceptable levels, and the predictive success of the Decision Trees method was found to be higher than Artificial Neural Networks.

Key Words: Data Mining, Neural Networks, Decision Tree, Insurance

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Gartner Group'a göre, veri madenciliği, şüpheli olmayan ilişkileri bulmak ve veriyi hem veri sahibinin anlaşılabilir hale getirmesi hem de yararlı olan yeni yollarla özetlemesi için (genellikle büyük) gözlemsel veri kümelerinin analizidir [1]. Veri madenciliği geçmişine bakılırsa, "Veritabanlarında Bilgi Keşfi" (KDD) terimi 1989 yılında ilk kez Gregory Piatetsky-Shapiro tarafından üretilmiş olup aynı zamanda KDD adındaki ilk atölye kurulmuştur. 1990'lı yıllarda ise "veri madenciliği" terimi ortaya çıkmıştır [2]. Perakende şirketleri ve finansal topluluklar, verileri analiz etmek, müşteri tabanını artırmak, eğilimleri tanımak ve müşteri talebini tahmin etmek gibi amaçlar için veri madenciliğini kullanmaya başlamıştır. Bu yöntem gittikçe yaygınlaşarak bankacılık, sigortacılık, reklamcılık vb. pek çok alanda kullanılmaya başlanmıştır. Veri madenciliğinin uygulama alanlarında, aynı zamanda tezin konusunu kapsayan sigortacılık sektöründe de bir takım çalışmalar olmuştur [2].

“Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması” isimli çalışmada, sigortacılık sektöründe hasar ihbarlarının asılsız olup olmadığına karar vermek amacıyla veri madenciliği tekniklerinden biri olan karar ağaçları, bu çalışmanın yöntemi olarak kullanılmıştır. Oluşan model sonucu İstanbul ve çevresinde tüzel kişiliklere ait araç ihbarlarının olumsuz sonuçlandırma olasılığı yüksek çıktığı gözlemlenmiştir [3].

“Applications Of Data Mining Techniques In Life Insurance” isimli çalışma, sigorta şirketlerinin modern veri madenciliği metodolojilerini kullanarak ve böylece maliyetleri düşürerek, kârlarını artırarak, yeni müşteriler kazanarak, mevcut müşterilerden

yararlanarak ve yeni ürünler geliştirerek nasıl yararlanabileceği konusunu içermektedir. Uygulama kullanılan örneklerde K-means, apriori ve K en yakın komşular algoritmaları kullanılmıştır [4].

Katamuri Samara'nın "Predictive Modeling With SAS Enterprise Miner" adlı kitabında yer alan, sigortacılık sektörüne ait karar ağaçları, lojistik regresyon ve yapay sinir ağaçları yöntemleriyle uygulanan risk analizleri de çalışmamıza yön vermede etkili olmuştur [5].

1.2 Tezin Amacı

Bu çalışmanın amacı sigortacılık sektörü kasko branşında, müşterilerin poliçe bilgileri üzerinden risk değerlendirmesi yaparak, yeni gelen bir müşterinin risk düzeyinin tahminini sağlamaktır.

1.3 Hipotez

Sigortacılık sektörü kasko branşında, yeni gelen bir müşterinin risk gruplarına atanmasında veri madenciliği yöntemlerinden karar ağaçları ve yapay sinir ağları kullanılabilir.

BÖLÜM 2

RİSK ANALİZİ VE SİGORTACILIK

2.1 Risk Nedir ?

Sıradan bir dilde risk kavramı genellikle olumsuz olaylar ya da etkilerin tehlikesiyle ilişkilendirilir. Ekonomik bakış açısından risk, geleceğin öngörülemezliğinden, bir bütçelendirilmiş değerden veya beklenen hedef değerden sapmanın olasılığını sunar [6]. Bir finansman perspektifinden risk, yatırımın getirisinin beklenenden farklı olma ihtimalini tanımlar. Risk yönetimi ise, yatırım ve iş karar verme sürecini belirlemek, analiz etmek, gelecekte oluşabilecek riskleri öngörüp ona göre bir yol haritası çizerek faydalı bir sonuca ulaşabilmek için kullanılan bir süreçtir.[7].

Hillson göre, risk tanımına yönelik iki seçenek mevcuttur. İlk olarak, pozitif etkilere sahip riskin fırsat olarak bilinen ve olumsuz etkilere sahip riskin tehdit olarak bilindiği bir terim olarak adlandırılmaktadır. İkinci olarak ise, riski yalnızca olumsuz etkileri veya tehdidi temsil etmek için kapsayan ve olumlu etkileri olan bir belirsizlik olma şansını ifade eden kapsamlı bir terimdir. Birinci seçenek, birçok uygulayıcı ve risk yönetimi araştırmacısı tarafından yaygın olarak kabul görmektedir.

Heldman'a göre, çoğu kişi diğer tarafı genelde gözden kaçırmaktadır. Bu nedenle, riski olumsuz sonuçlar açısından düşünme eğilimindedir. Riskler, projelere tehdit oluşturan potansiyel olaylardır ancak aynı zamanda riske atılabilecek potansiyel fırsatlar olarak da değerlendirilmelidir [8].

Risk Analizi

Şirketlerin üstlendiği faaliyetin uygulama alanlarını sağlıkla yürütebilmesi için etkili bir yönetim sistemi gereklidir. Bu yönetim sistemi projenin yaşam döngüsünün tüm aşamalarına ve ilgili tüm faaliyetlere uygulanmalıdır. Etkili yönetim sisteminin temel unsurlarından birisi, risk azaltma tedbirlerini uygulamaya koyma ihtiyacı üzerine karar vermeye yardımcı olmak için tehlikeyi belirlendikten sonra, tanımlanmış tehlikeden kaynaklanan risklerin olasılığını ve sonuçlarını belirlemek için risk analizi yapmaktır [9]. Sonuçlar ve olasılıklar daha sonra risk seviyesini belirlemek için birleştirilir. Bu analizler karar alıcılara, personel, çevre ve tesislerle ilgili risklerin özelliklerini sunar [10].

Risk analizi nitel veya nicel olarak yapılabilir.

2.1.1 Nitel Risk Analizi

Nitel bir analiz, ana risk kaynaklarının veya faktörlerin tanımlanmasını sağlar. Bu tanımlama yaygın olarak kontrol listeleri, varsayım analizleri veya beyin fırtınası yardımıyla yapılabilir [11]. Niteliksel teknikler temel alınarak risk olasılığı sonuçları yüksek, orta ve düşük terimlerde tanımlanır. Ardından risk düzeyi, olasılıkların ve sonuçların kombinasyonu ile belirlenir [10]. Nitel risk analizi olarak kullanılan tekniklerin bir kısmı aşağıdaki gibidir.

• Kontrol listeleri;

• Varsayımlar analizi;

• Bir riskin anlaşılma derecesini ölçebilmek adına, risklerle ilgili verilerin derecesini değerlendirmek ve verilerin güvenilirliğini incelemek için veri hassasiyeti sıralaması;

• Niteliksel terimlerle (çok yüksek, yüksek, orta ve benzeri) parametreleri tanımlamak için olasılık ve etki açıklaması;

• Riskler ve nedenleri arasındaki ilişkileri göstermek için neden-sonuç diyagramları [9].

2.1.2 Nicel Risk Analizi

Nicel bir risk analizi deterministik veya stokastik olarak yapılabilir. Deterministik yaklaşımda olasılık ve sonuçların tahmininde tek noktalı bir değer kullanılır. Monte Carlo

simülasyonu gibi stokastik yaklaşımda belirsiz girdiler, olasılık dağılımları olarak bilinen olası değerlerin aralığı ile temsil edilmektedir [10].

Niceliksel teknikler, düşük, orta ve yüksek gibi göreli terimler yerine sonuç ve olasılıkların gerçekçi değerini tahmin etmeye izin verir. Genellikle daha karmaşık teknikler içerir ve bilgisayar yazılımı gerektirir. Ancak, proje faaliyetleri hakkında yeterli ve ayrıntılı bilgi eksikliği, nicelik tekniklerinin her zaman mümkün olmamasına neden olmaktadır [10].

Kısaca risklerin tanımlanması ve öznel olarak tahmin edilmesine odaklanan analiz nitel bir analiz, risklerin objektif bir şekilde değerlendirilmesine odaklanan analiz ise nicel bir analizdir [11]. Nicel risk analizi olarak kullanılan bir takım teknikler aşağıdaki gibidir.

• Duyarlılık analizi

• Olasılıksal toplamlar

• Monte Carlo ve Latin Hypercube Simülasyonu

• Olasılıksal etki diyagramlarıdır [9].

2.2 Sigortacılıkta Risk Analizi

Sigorta, bir şahsın veya tüzel kişinin öngörülmüş bir rizikonun gerçekleştiği durumda ortaya çıkabilecek kayıplara karşı finansal koruma veya geri ödeme aldığı bir sözleşmedir [4]. Hızla gelişen sanayi, teknoloji, kentleşme ve daha birçok olgu insanların risklere karşı kendilerini koruma isteğini artırmış ve bu da sigortacılık sektörünün gün geçtikçe gelişip büyümesine büyük katkı sağlamıştır. Dünyanın en büyük endüstrilerinden biri olan sigortacılık sektöründe artan felaket kayıplarının etkileri ve nedenlerini ölçmek, endüstrinin gelişmesi ve çözüm önerilerinin artırılabilmesi için çok önemlidir. Bu etki ve nedenleri ölçmek, analizleri yapmak için veri madenciliği oldukça faydalı bir yöntemdir. Veri madenciliği, daha önce bilinmeyen kalıpları ortaya çıkarmak için büyük miktarda veriyi seçme, keşfetme ve modelleme süreci olarak tanımlanabilir [4]. Bu süreçler sonucunda sigorta şirketleri geleceklerini yönetme açısından pek çok fayda sağlar. Örneğin, veri madenciliği tekniklerini uygulayarak şirketler, müşterilerin satın alma kalıpları ve davranışları hakkındaki verileri tam olarak kullanabilir; aynı zamanda haksız kazançların azaltılmasına, sigortacılığın geliştirilmesine ve risk yönetiminin

geliştirilmesine yardımcı olmak için işletmeleriyle ilgili daha fazla bilgi sahibi olabilirler [4].

Riski yönetmek her sektör için çok önemlidir. Tezin genel konusu olan sigortacılık için risk yönetimini inceleyecek olursak; sigorta şirketleri, iki temel fonksiyonu yerine getiren finansal kuruluşlardır. Riskleri fiyatlandırarak sigortalarlar ve öngörülen riskler için ödenen primler ile yatırım yaparlar. Sigortacılar prim havuzlarını emlak ve finansal araçlara yatırırlar ve böylece ekonominin büyümesini ve gelişimini desteklemek için bir likidite kaynağı oluştururlar. Bu nedenle, sigorta sektörünün mali sağlığı ekonomi için son derecede önemlidir. Bu bağlamda, sigortacının ürünlerinin, yatırımlarının ve diğer gayretlerinin risklerini yönettiği uygulamalar büyük önem taşımaktadır [13].

Sigortacılık genel olarak hayat ve hayat dışı sigortalar olarak ikiye ayrılmaktadır. Her ülke kendi ihtiyacına göre farklı sigorta branşları oluşturulabilir. Türkiyede yaygın olarak satışı yapılan sigorta branşları aşağıdaki Çizelge 2.1'de gösterilmektedir.

Çizelge 2. 1 Sigorta Branşları

HAYAT SİGORTALARI	**HAYAT DIŞI SİGORTALAR**
Sağlık	Araç (Kasko, Trafik)
Ferdi Kaza	Yangın
Ölüm Hali	Kaza
Grup Hayat	Nakliye
Maluliyet	Tarımsal
Özel Durum	Makine

Tezin uygulama konusu olan kasko branşı, insanların sahip olduğu motorlu veya motorsuz araçlarını güvence altına alabildikleri bir sigorta branşıdır. Sigorta yaptıran şahıs "sigortalı" olarak adlandırılır. Sigortalı, sigorta şirketine belli bir bedel ödeyerek kasko hizmetini satın alır. Bu ödediği bedele ise "prim" adı verilmektedir. Satın alınan kasko hizmeti poliçenin şartlarına göre; aracın kazaya uğraması, çalınma vs. durumlarında sigortalının hasar bedelinin bir kısmını veya tamamını sigorta şirketinden tahsis etmesiyle gerçekleşir. Poliçenin imzalanması aşamasındaki prim ödemesi ise sigorta yapacak kişiden kişiye değişiklik göstermektedir. Aynı model araca sahip 2 kişinin

sigorta şirketine ödediklerine prim muhtemelen farklılık gösterecektir. Burada risk kavramı devreye girmektedir. Sigorta şirketleri tarafından her sigortalının kendine özgü bir risk seviyesi belirlenmektedir. Bu analizde, sigortalının yaşı, daha önce kaza yapıp yapmadığı, yaptıysa bunun adedi, yaşadığı bölge vs. pek çok değişken ele alınabilinir. Yapılan risk tespiti sonucu sigortalıya bir prim ödemesi çıkartılır. Bu durumda risk analizinin sigorta şirketleri için çok önemli olduğu aşikârdır.

BÖLÜM 3

VERİ MADENCİLİĞİ

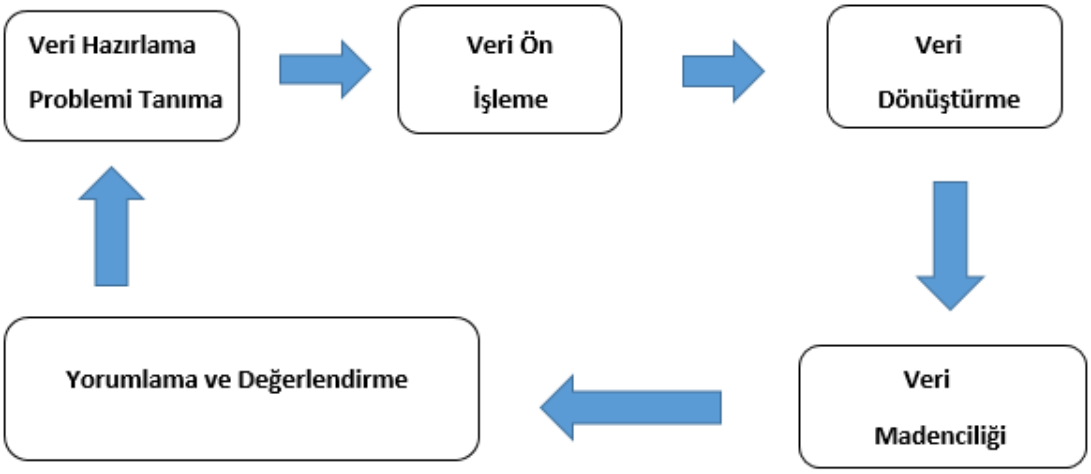
Dünyadaki teknolojik ilerlemeler gün geçtikçe gelişmekte ve farklı bir boyut kazanmaktadır. Saklanması ve kullanılması gereken verinin hacmi gittikçe büyümektedir. Bu büyük verilerden anlamlı sonuçlar çıkarmak için çeşitli teoriler geliştirilmektedir. Bu geliştirilen teori ve yöntemler veri tabanlarındaki bilgi keşfi sürecinin karşılığıdır. Verinin nasıl saklanması ve algoritmaların nasıl uygulanacağı, modellerin nasıl geliştirileceği ve yorumlanması gerektiği konusunda bir yol izler [14].

3.1 Veri Tabanlarında Bilgi Keşfi ve Süreci

Veri saf halde değersizdir. Önemli olan verinin içinden anlamlı bilgiler çıkarabilmektir. Verinin toplanması ve saklanması da bu anlamlı bilgileri çıkarabilmek adına önemli bir aşamadır. Veri toplama; veri tabanı tasarımı ile en uygun gösterimi kullanarak veritabanındaki girdilerin tanımı ve veri kalitesini içermektedir [15].

Veri tabanlarında bilgi keşfi sürecine genel olarak baktığımızda; bilgi toplama, algoritmalar geliştirme, süreçler ve veri toplama araçlarından potansiyel bilgiyi alma ,mekanizmalarının incelenmesi ve oluşturulması etrafında döner. Başka bir deyişle bilgi keşfi; veritabanlarını tanıyabilme, yeni, geçerli , potansiyel olarak yararlı ve sonuç olarak anlaşılabilir kalıpları / modelleri tanımlama işlemidir [16]. Bilgi keşfi çok disiplinli bir aktivitedir, interaktif ve tekrar eden bir süreçtir [17].

Bu sürecin yapısının aşamaları, aşağıda gösterilmiştir [18].



Şekil 2. 1 Bilgi Keşfi Süreci

3.1.1 Veri hazırlama ve Problemi tanıma

Bilgi keşfi sürecinin ilk aşamasıdır ve amacı uygulama alanını anlamak ve sorunu formüle etmektir. Bu adım, açık bir şekilde faydalı bilgiyi elde etmek, uygulama hedefine ve verilerin niteliğine göre uygun veri madenciliği yöntemlerini seçmek için bir ön şarttır.

3.1.2 Veri Ön İşleme

Veri temizleme ve entegrasyon işlemleri bu aşamadadır. Bu adım genelde analiz aşaması boyunca en uzun süren adımdır.

3.1.3 Veri Dönüştürme

Seçilen veriler doğru veri madenciliği için dönüştürülür.

3.1.4 Veri Madenciliği

Potansiyel olarak, yararlı kalıpların akıllı yöntemlerle çıkarıldığı aşamadır.

3.1.5 Yorumlama ve Değerlendirme

Bilgiyi temsil eden bu aşamadaki kalıplar yorumlanarak keşfedilen bilgi kullanıcıya aktarılır [18].

3.2 Veri Madenciliği Kavramı

"Veri madenciliği" terimi, istatistikçiler tarafından, 1960'lardan beri disiplinsiz verilerin araştırılmasını tanımlamak için aşağılayıcı bir terim olarak kullanılmıştır. "Veri tarama" ve "balık tutma" veri madenciliği için geçmişte yapılan benzetmeler olarak karşımıza çıkmaktadır. 1990'lı yıllarda, makine öğrenimi alanındaki araştırmacılar, işletmelerin daha iyi kararlar almalarını sağlayan kalıpları keşfetmek için algoritmalarını bu büyük veritabanlarına uygulamaya başlamışlar ve gelecek araştırmalar için hipotezler geliştirme konusunda aşama katetmişlerdir [19].

Bilgi teknolojilerine odaklı toplumlarda veri, her zaman araştırmaların en önemli kaynağı olmuştur [20]. Güncel bilgilerin hızlı bir şekilde gelişmesi, özellikle geniş alanda yüksek bir hızda yaygınlaşması nedeniyle, bilgi güncelleme ve bilgi üretme hızı günden güne artmaktadır.

Bunun yanında bilişim sistemlerinin de gün geçtikçe gelişmesiyle birlikte artık büyük verilerin elde edilmesi ve sistemlerde muhafaza edilmesi kolaylaşmıştır. Sistemlerde saklanan her bir veri, çözüm üretilmek istenen probleme göre uygun analiz yöntemleriyle anlamlandırılabilmektedir.Bu büyük verilerden geleceğe yönelik anlamlı sonuçlar çıkarmaya kısaca Veri Madenciliği (Data Mining) denilmektedir [21]. Bazı bilim insanları veri madenciliği hakkında çeşitli tanımlar ortaya koymuştur.

Gartner Group'a göre, "Veri madenciliği keşfetme sürecidir. Anlamlı miktarda yeni korelasyon, desen ve trendleri büyük miktarda elemek suretiyle depolarda depolanan verilerin, model tanıma teknolojilerinin yanı sıra istatistiksel ve matematiksel teknikler kullanılarak analiz etme sürecidir".

D.J. Hand, Heikki Mannila, Padhraic Smyth'e göre "Veri madenciliği, şüphe çekmeyen ilişkileri bulmak ve veriyi hem veri sahibi için anlaşılabilir hem de faydalı olan yeni yollarla özetlemek için (genellikle büyük) gözlemsel veri kümelerinin analizidir ".

Evanjelos Simoudis in Cabena ve çalışma arkadaşlarına göre, "Veri madenciliği, makine dilinde öğrenme, kalıp tanıma, istatistik, veri tabanları ve görselleştirme yöntemlerini, geniş veri tabanlarından bilgi çıkarma yöntemiyle ele alan, bir araya getiren disiplinlerarası bir alandır" [22].

Verinin büyüklüğü de araştırmaya konu olan problemi çözmede çok önemli bir role sahiptir. Sağlıklı bir sonuca varma adına verinin büyük olması önemli bir konudur. Büyük veri tabanlarının oluşması bir sürece tabiidir [21]. Dijital veri toplama ve depolama teknolojisindeki ilerleme, büyük veritabanlarının büyümesine katkı sağlamıştır. Bu, insanoğlunun her alanında, sıradan alanlardan (Süpermarket işlem verileri, kredi kartı kullanım kayıtları, telefon görüşmeleri detayları, ve hükümet istatistiklerinden) daha sıradışı (stronomik cisimlerin görüntüleri, moleküler veri tabanları ve tıbbi kayıtlar) alanlara uzanacak şekilde veri depolamasına olanak vermiştir. Bu veriler, çeşitli analizler yapılarak, veritabanının sahibine değerli olabilecek bilgileri çıkartmanın imkânını sağlamıştır [23]. 1980'lerin sonu ve 1990'ların başında, müşteri bilgilerinin büyük veri tabanlarına depolanabilme kolaylığının gelişmesi nedeniyle, özellikle de kredi kartı bankalarına sahip olan şirketler, müşterileri hakkında daha fazla öğrenme potansiyelini keşfetmek istemişlerdir [20].

Değişik müşteriler, pazarlar, çeşitli hizmetler ve ticaret ortamlarının çeşitliliği ve karmaşıklığı ile doğru ve zamanında karar verme için doğru bilgilere erişme ihtiyacı ile ilgili olarak firmaların işlevsel ve etkili bilgiyi bulmaları ve sınıflandırmaları zorunludur. Veri madenciliği bu firma ve kurumların ihtiyaçlarına cevap vermektedir. Veriler ne kadar büyük olursa ve ilişkileri de o kadar karmaşık hale gelir, verilere dahil olan bilgilere erişim de o kadar zor olur. Bu nedenle veri madenciliğinin bilgi keşfi yaklaşımı olarak rolü veriyi daha belirgin hale gelmektedir.

Günümüzde veri madenciliği bir karar destek sisteminde vazgeçilmez bir araçtır ve pazar segmentasyonu, müşteri hizmetleri, dolandırıcılık tespiti, kredi ve davranış puanlaması ve kıyaslama konularında kilit rol oynamaktadır. Bölgesel ve küresel rekabetçi pazarlar, üretimde, hizmet sunumunda, memnuniyetin artmasında ve müşteri cazibesinin artmasında kurumların (müşteri açısından) daha iyi bir yer edinmesini ve firmanın gelişmesini gerektirir [18].

Basitçe söylemek gerekirse, veri madenciliği büyük miktarda veriden bilgi çıkarma veya "araştırma" işlemi demektir. Veri madenciliği, teori ve veritabanının teknolojisi, yapay zeka, makine öğrenimi, istatistik vb. bütünleştiren disiplinler arası bir alandır. Veri

madenciliği, bir dereceye kadar bilgi teknolojisinin gelişmesinin kaçınılmaz bir ürünüdür ve verilerin birikiminin ileri bir aşamasındadır [24].

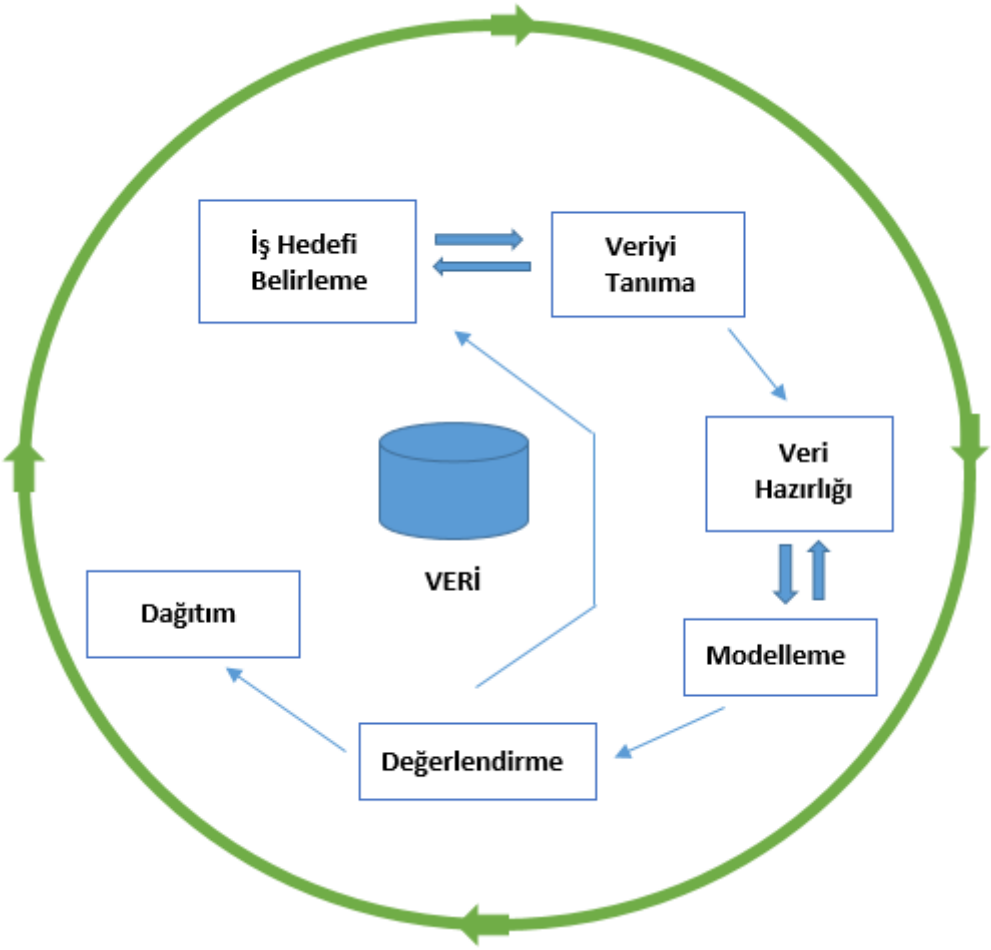
Şekil 3. 2 Veri Madenciliği

Veri madenciliğine birkaç örnek verecek olursak; Forbes dergisi, veri madenciliği ve tahminsel analitik yöntemlerin kullanılmasının, konjestif kalp yetmezliği geliştirme riski en yüksek olan hastaların belirlenmesine yardımcı olduğunu bildirmişir. IBM, 350.000 hastayla ilgili 3 yıllık verileri toplarken, kan basıncı, kilo ve reçete edilen ilaçlar gibi şeyleri de içeren 200'den fazla faktöre ilişkin ölçümlerde tahmini analitik kullanarak IBM, konjestif kalp yetmezliği riski taşıyan 8500 hastayı 1 yıl içinde tespit etmiştir [25].

MIT Teknoloji İncelemesi, Obama kampanyasının, Başkan Obama'nın Mitt Romney'den 2012 başkanlık seçimini kazanmasına yardımcı olan veri madenciliğini etkili bir şekilde kullandığını bildirmiştir. İlk önce muhtemel Obama seçmenlerini bir veri madenciliği modeli kullanarak tespit edilmiştir. Daha sonra yoklama sonuçlarını ilçe tarafından tahmin etmek için ayrı bir veri madenciliği modeli kullanılmıştır. Hamilton County, Ohio'nun önemli bir ilçesinde, model Obama'nın oyların% 56,4'ünü alacağını öngörmüştür. Gerçek oyların Obama payı % 56.6 olmuştur. Böylece tahminin sadece % 0.02 oranında kesilmiştir. Bu tür hassas öngörü gücü, kampanya personelinin kıt kaynaklarını daha verimli bir şekilde tahsis etmesini sağlamıştır [25].

3.3 Veri Madenciliği Metodolojisi

Bir veri madenciliği projesi, modellemeden daha fazlasını gerektirir. Modelleme aşaması, bir veri madenciliği projesinin uygulama sürecinde yalnızca bir safhadır. Kritik öneme sahip adımlar modelin oluşturulmasından önce olan adımlardır ve projenin başarısı üzerinde önemli bir etkiye sahiptir. Aşağıda veri madenciliği aşamaları Şekil 3.3'te gösterilmiştir [26].



Şekil 3. 3 Veri Madenciliği Aşamaları

3.3.1 Veri Madenciliği Metodolojisi Aşamaları

3.3.1.1 İş Hedefi Belirleme

Veri madenciliği projesi, iş hedefinin anlaşılması ve mevcut durumun değerlendirilmesi ile başlamalıdır. Kaynaklar ve sınırlamalar dahil, projenin parametreleri göz önüne

alınmalıdır. İş hedefi bir veri madenciliği hedefine dönüştürülmelidir. Başarı kriterleri tanımlanmalı ve bir proje planı geliştirilmelidir [26].

3.3.1.2 Veriyi Tanıma

Bu aşama, tanımlanan hedefe yönelik doğru adresleme için veri gereksinimlerini ve gerekli verilerin kullanılabilirliğini araştırmayı içermektedir. Bu aşama ayrıca, verileri anlamak, kullanılabilirlik ve kalitede olası sorunları tanımlamak için özet istatistikler ve görselleştirme araçları ile ilk veri toplama ve araştırmayı içerir [26].

3.3.1.3 Veri Hazırlığı

Veri hazırlama genellikle proje zamanının yaklaşık % 90'ını tüketir. Veri hazırlık aşamasının sonucu nihai veri kümesidir. Kullanılabilir veri kaynakları belirlendikten sonra seçilmeleri, temizlenmesi, yapılandırılması ve istenilen biçimde biçimlendirilmesi gerekir. Daha derinlemesine veri araştırması görevi, iş anlayışına dayalı kalıpları fark etmek için bu aşamada gerçekleştirilir [27].

3.3.1.4 Modelleme

İlk olarak, hazırlanan veri kümesi için kullanılacak modelleme teknikleri seçilir. Sonra, modelin kalitesini ve geçerliliğini doğrulamak için test senaryosu oluşturulur. Ardından, hazırlanan veri kümesinde modelleme aracını çalıştırarak bir veya daha fazla model oluşturulur. Son olarak, oluşturulan modellerin iş girişimleri ile uyumlu olduğundan emin olmak için paydaşları içeren modellerin titizlikle değerlendirilir [27].

3.3.1.5 Değerlendirme

Değerlendirme aşamasında, model sonuçları birinci aşamadaki hedefler bağlamında değerlendirilir. Bu aşamada, model sonuçlarında veya yeni faktörlerden keşfedilen yeni kalıplara bağlı olarak yeni iş gereksinimleri yaratılabilir. İş anlayışını kazanmak, veri madenciliğinde tekrar eden bir süreçtir. Dağıtım aşamasına geçmek için bu adımda kararı verilir [27].

3.3.1.6 Dağıtım

Projenin bulguları ve sonuçları bir raporda özetlenmekle birlikte, projenin son aşamasıdır. En iyi model bile, sonuçlarının dağıtılmadığı ve kuruluşların günlük pazarlama operasyonlarına entegre edilmemesi durumunda bir iş başarısızlığı haline gelecektir. Müşterilerin puanlanması ve sonuçların güncellenmesi için bir prosedür tasarlanmalı ve geliştirilmelidir. Dağıtım prosedürü ayrıca, model sonuçlarının işletme genelinde dağılımını ve organizasyonların veritabanlarında ve operasyonel CRM sisteminde yer almasını sağlayacaktır. Son olarak, bir bakım planı tasarlanmalı ve tüm süreç gözden geçirilmelidir [26].

BÖLÜM 4

VERİ MADENCİLİĞİ TEKNİKLERİ

Yaygın olarak kullanılan beş veri madenciliği tekniği aşağıda belirtilmiştir [17].

1-Sepet analizi tekniği (Market Basket Analysis)

2-Bellek Tabanlı Yöntemler

3-Kümeleme Analizi(Clustering)

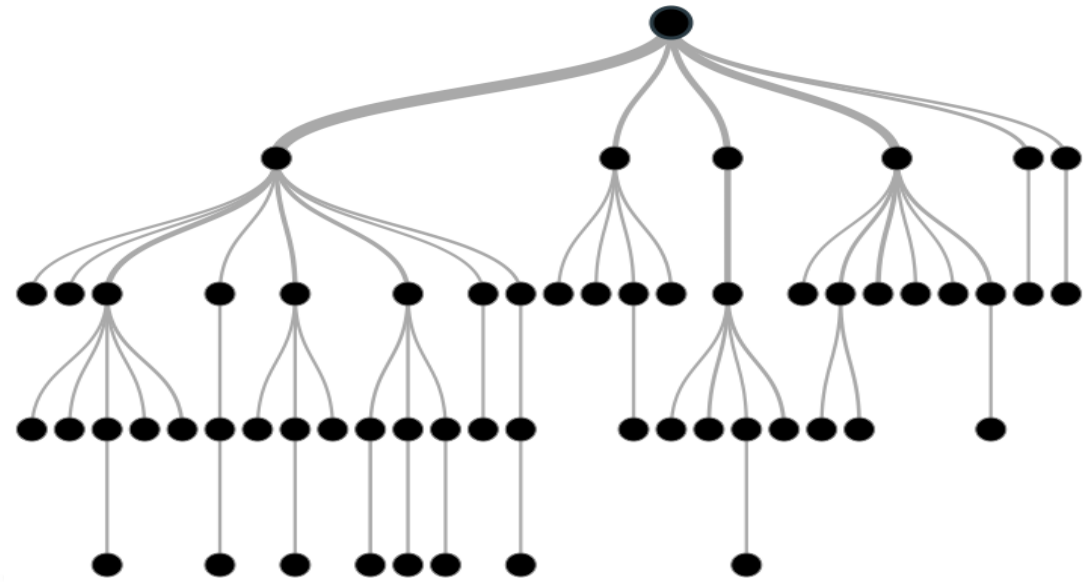
4-Karar Ağaçları (Decision Tree)

5-Yapay Sinir Ağları (Neural Network)

Bu tez çalışmasında karar ağaçları ve yapay sinir ağları teknikleri yöntemleri kullanılmıştır. Bu bölümde anılan iki veri madenciliği tekniği tanıtılmaktadır.

4.1 Karar Ağaçları

Veri madenciliğinde kara r ağacı, sınıflayıcıları ve regresyonları göstermek için kullanılan bir modeldir. Bazı düğümler ve dallardan oluşur. Sınıflandırma karar ağacında, yapraklar sınıfları belirtir. Karar, diğer düğümlerdeki (yaprak olmayan düğümlerdeki) bir veya daha fazla özel niteliklere göre yapılır. Karar ağacı, basit ve anlaşılır olduğu için veri madenciliğinde popüler bir tekniktir. Diğer bir deyişle, karar ağacı, çıktıyı yorumlamak için tüm içeriği bir uzmandan bağımsız olarak tek tek açıklar. Aslında, bu grafik bir yöntemdir ve bu yorumdan ötürü diğer sınıflandırma yöntemlerinden daha basittir. Bununla birlikte, çok sayıda düğüm karar ağacının grafik gösterimini zorlaştırabilir [18].



Şekil 4. 4 Karar Ağacı Yapısı

Bir karar ağacı modeli birkaç parçadan oluşur:

• Bir veri kümesinin her kaydını bir yaprak düğümüne atamak için düğüm tanımları veya kuralları,

• Her yaprak düğümünün posterior olasılıkları,

• Her yaprak düğümüne bir hedef seviyesinin atanması.

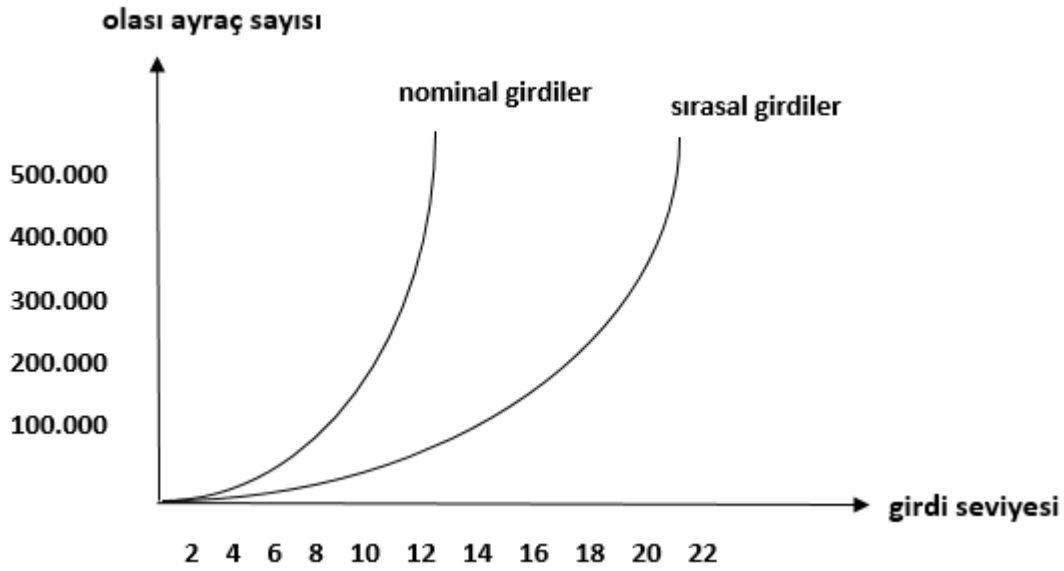
Düğüm tanımları, eğitim veri seti kullanılarak geliştirilmiştir ve girdi aralıkları cinsinden belirtilir. Posterior olasılıklar, eğitim veri setini kullanarak her düğüm için hesaplanır. Hedef seviyesinin her düğüme atanması, eğitim aşamasında eğitim veri setleri kullanılarak yapılır [28].

4.1.1 Karar Ağaçlarının Oluşumu

Ağaçların oluşumu sırasında bazı kriterlere dikkat edilerek ilerleme sağlanır. Takip edilmesi gereken adımlar aşağıda sıralanmıştır.

4.1.1.1 Bölünmüş Arama

- Hangi bölünmeler düşünülmelidir?



Şekil 4. 5 Girdi Seviyesi Göre Olası Ayraç Sayısı

Yukarıdaki Şekil 4.5'te de ifade edildiği gibi girdi sayısına bağlı olarak olası ayraç sayısı da artış göstermektedir. Göz önüne alınması gereken muhtemel bölünmelerin sayısı çoktur ancak en basit hallerdir. Ayrık arama algoritması, mümkün olan tüm bölümleri ayrıntılı olarak inceler. Dikkate alınması gereken olası bölünmeleri sınırlandırmak için çeşitli kısıtlamalar getirilir. En yaygın kısıtlama yalnızca ikili bölmelere bakmaktır. Diğer kısıtlamalar, kesintisiz girdileri, adımlı arama algoritmalarını ve örneklemeyi içerir [17].

4.1.1.2 Ayrıştırma Kriteri

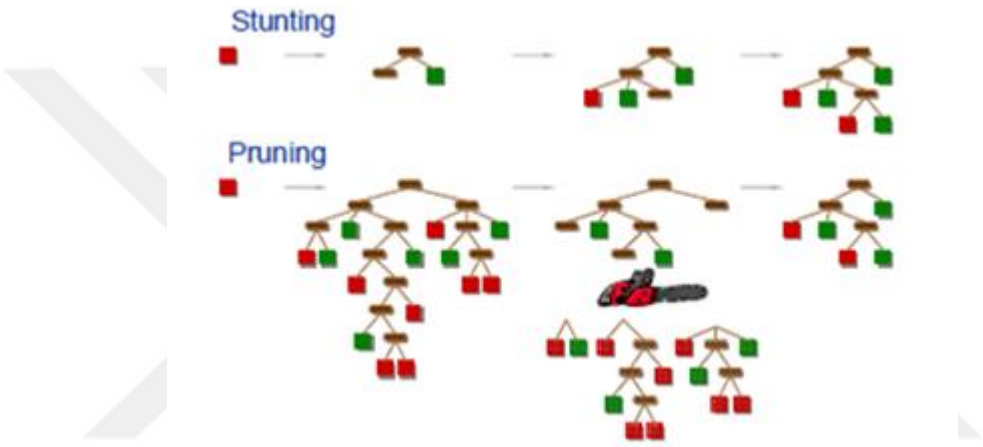
- En iyisi hangi bölümdür?

Bazı durumlarda, bölünme değeri açıktır. Beklenen hedef üst düğümdeki gibi alt düğümlerde aynı ise, iyileşme yapılmamıştır ve bölünme değersizdir. Bunun aksine, bölme saf düğümlerle sonuçlanırsa, bölünme tartışmasız en iyi sonuçtur. Sınıflandırma ağaçları için en yaygın kullanılan üç ayırma kriteri olan Pearson kare-kare testi, Gini indeksi ve entropi temel alınarak oluşturulmuştur. Her üçü de, alt düğümlerdeki sınıf dağılımlarındaki farkı ölçer. Üç yöntem genellikle benzer sonuçlar verir [17].

4.1.1.3 Durdurma ve Budama Kuralı

- Ayrılma ne zaman durmalıdır ve bazı budama işlemleri yapılmalı mıdır ?

Karar ağaçları için model karmaşıklığı yaprakların sayısı ile ölçülür. Bir ağaç, bütün yapraklar saf olana veya yalnızca bir kutu bulunana kadar bölünebilir. Bu ağaç, eğitim verilerine mükemmel bir uyum sağlayacak, ancak yeni verilere muhtemelen kötü öngörülerde bulunacaktır. Öbür ucunda, ağaç yalnızca bir yaprağa (kök düğüm) sahip olabilir. Doğru boydaki ağacı belirlemek için Şekil 4.6'da gösterilen şekilde iki yaklaşım vardır [28]:



Şekil 4. 6 Doğru Ağaç Boyu Seçmek İçin Kullanılan Yaklaşımlar

- Ağacın büyümesini önlemek için ön budama kurallarını kullanmak (**Stunting**).

Evrensel olarak kabul edilmiş önceden hazırlanmış bir kural, düğümün saf olması durumunda büyümeyi durdurmaktır. Diğer iki popüler kural, bir düğümdeki vaka sayısının belirtilenin altına düşmesi durumunda durdurulmasıdır. Sınırlama veya bölme belirli bir düzeyde istatistiksel olarak anlamlı olmadığında durur [28].

- Büyük bir ağaç yetiştirmek ve dalları budamak (**Pruning**).

Bu yaklaşımda ağacın performansını düşüren dallar budanır. En iyi (alt) ağacı belirlemek için bir değerlendirme kriteri gereklidir. Değerlendirme kriterleri, bekleme örnekleri üzerindeki performansa (doğrulama verileri veya çapraz doğrulama) dayalıdır. Maliyet veya kâr konuları değerlendirmeye dahil edilebilir. Ön budama işlemi, hesaplamanın

daha az talep edeceği halde, zayıf bölmelerin altında gerçekleşen gelecekteki bölünmelerin kaybolması riskini taşır. Fakat ön budama yaklaşımına göre daha çok tercih edilir [28].

4.1.2 Başlıca Karar Ağacı Algoritmaları

İstatistiksel, makine öğrenimi ve kalıp tanıma literatüründe yüzlerce karar ağacı algoritması önerilmiştir. En popüler olanları CART, CHAID, ID3 ve C4.5 'dir.

4.1.2.1 CART

CART (Classification and Regression Trees) Algoritması, Breiman ve arkadaşları tarafından, Sınıflandırma ve Regresyon Ağaçları temel alınarak hazırlanmıştır. Bir CART ağacı, bir düğümü iki öğrenme örneğini içeren kök düğümden başlayarak art arda iki düğüme ayırarak oluşturulmuş bir ikili karar ağacıdır. CART algoritmasının birçok varyasyonu vardır . Standart CART yaklaşımı ikili bölmelerle sınırlanmıştır. Olası tüm ikili bölmeler dikkate alınır. Veriler çok büyükse, düğüm içi örnekleme kullanılabilir. Standart bölme kriteri, sınıflandırma ağaçları için Gini indeksine ve regresyon ağaçları için varyans azaltmaya dayanmaktadır. Çok sınıflı problemler için diğer kriterler (iki yönlü kriter) ve regresyon ağaçları (en az mutlak sapma) kullanılmaktadır. Maksimum ağaç, v-katlama çapraz doğrulamasını kullanarak büyütülür ve budama yapılır. Yeterli veri varsa doğrulama verileri kullanılabilir [28].

4.1.2.2 CHAID

CHAID (Classification and Regression Trees) analizi, kategorik bir yanıt değişkeni ve diğer kategorik belirleyici değişkenler arasındaki ilişkileri keşfetmek için kullanılan bir algoritmadır. Çok sayıda kategorik değişkenli veri kümelerindeki kalıpları ararken yararlıdır ve ilişkileri kolayca görselleştirilebildiğinden verilerin özetlenmesinin kolay bir yoludur [29]. 1970'lerin sonunda bir istatistikçi Kass tarafından önerilen CHAID algoritması, karar ağacı geliştirme için denetlenmiş öğrenmenin istatistiksel olarak en popüler yöntemlerinden biridir. Esas olarak, çok değişkenli bağımlılık yöntemlerinden biri olan CHAID algoritması, kategorik bağımlı değişkene ve kategorik ve / veya metrik olabilen çoklu bağımsız değişkenler arasındaki ilişkinin saptanması için kullanılır.(Bu

durumda, kodlamaları ve kategorik değişkenlere dönüşümü daha önce yapılmalıdır) Kısaltma CHAID, Pearson'un ki-kare istatistiğine ve karşılık gelen p-değerine dayanan otomatik ve yinelemeli ağaç geliştirme prosedürünü belirtir. Algoritmanın adı ile gösterildiği gibi, bağımlı değişken kategorilerine göre heterojen popülasyonu homojen gruplara yinelemeli olarak bölmek için temel ölçüt Ki-kare testi istatistiğidir [30].

4.1.2.3 ID3

ID3 algoritması Quinlan tarafından ileri sürülmüştür. Sınıflandırmayı başlatmak için bilgi entropisi ve bilgi kazanma dereceleri göz önüne alınır. Bilgi teorisine dayanan klasik bir karar verme algoritmasıdır [32].

ID3 algoritması, eğitim örneklerinden elde edilen bilgilere (bilgi kazancı) dayalı bir ağaç oluşturur ve daha sonra test verilerini sınıflandırmak için kullanır. ID3 algoritması genellikle eksik değerler olmaksızın sınıflandırma için nominal öznitelikler kullanır [33].

4.1.2.4 C4.5

C4.5, karar ağacı formunda sınıflandırma kuralları oluşturmaya yönelik standart bir algoritmadır. ID3'ün bir uzantısı olarak, C4.5'te bölme niteliklerini seçmenin varsayılan ölçütleri, bilgi kazanım oranıdır. Bilgi kazanımını ID3'teki gibi kullanma yerine, bilgi kazanma oranı birçok değerli nitelikleri seçmenin sapmasından kaçınır [31]. Ayraçları sınıflandırırken; değişken eğer sürekli ise ikili ayraç yöntemini, kategorik ise çok yollu ayraç yönetimi kullanır. Ayraçların en iyisini belirlemede Entropi'yi kullanırken, durdurma kuralında Geriye Doğru Budama Kuralı esas alınır [18].

4.1.3 Karar Ağaçları Neden Kullanışlıdır?

Karar ağaçları, çoklu değişken (veya çoklu etki) analizlerinin bir şeklidir. Çok değişkenli analizlerin tüm biçimleri bir sonucu (veya hedefi) öngörmek, açıklamak, tanımlamak veya sınıflandırmamıza olanak tanır. Çok değişkenli analiz, birden fazla girdi değişkeninin, faktörün veya boyutun birleşik etkilerinin bir sonucudur.

Karar ağaçlarının bu çok değişken analiz kabiliyeti basit ve tek etkili ilişkilerin ötesinde çoklu etkiler bağlamında da yeni şeyleri keşfetmeye ve tanımlamaya olanak tanır. Çoklu

değişken analizi, mevcut problemi çözmede özellikle önemlidir, çünkü başarıyı belirleyen hemen hemen tüm kritik sonuçlar birden fazla faktöre dayanır.

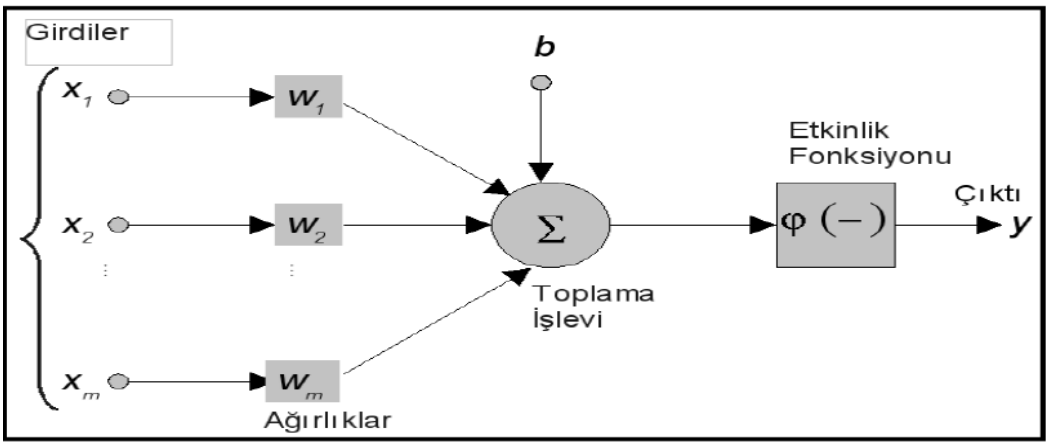
Karar ağaçlarının çok tercih edilme sebebi, kullanım kolaylığına, çeşitli veri ve ölçü seviyelerine sahip sağlamlıkta ve yorumlanabilirliğin kolaylığına dayanmaktadır. Karar ağaçları aşamalı olarak geliştirilir ve sunulur. Karar ağaçları ham verileri iş, mühendislik ve bilimsel konular hakkında artan bir bilgi ve farkındalık haline getirir ve bu bilgiyi basit ancak güçlü bir şekilde insan tarafından okunabilir kurallara yerleştirmemize olanak tanır [34].

4.2 Yapay Sinir Ağları

Biyolojik sinir ağı insan beyninin vazgeçilmez bir parçasıdır. Çok miktarda bilgiyi aynı anda işleyebilecek kadar karmaşık bir sistemdir. Biyolojik ağlar, farklı görsel girdileri çok hızlı tanıyabilir ve işleyebilir [35]. Yapay sinir ağları (YSA) da beyin gibi biyolojik sinir sistemlerinin bilgi işleme biçiminden esinlenen bir bilgi işleme paradigmasıdır. Genellikle belirli bir uygulama için yapılandırılmış yapar sinir ağı, belirli bir problemi çözmek için bir öğrenme süreci boyunca bir araya getirilmiş, nöronlar olarak adlandırılan çok sayıda birbirine paralel işlemci elemanlarından oluşur [36]. Düğüm olarak da adlandırılan nöronlar, sinir ağlarındaki temel işlem birimidir. Birçok nöron, tam bir ağ oluşturmak için birbirine bağlıdır.

Bir yapay sinir ağında, dış dünyadan girdi alan bir sensör katmanı vardır. Bu sensör nöronları, aktivasyonlarını ağırlıklı bağlantılar üzerinden diğer nöronlara iletirler. Sinir ağındaki son kat, çıkış katmanıdır ve çıktı katmanı sinir ağı içeren sistem tarafından kullanılır. Girdi ve çıktı katmanları arasında "gizli katmanlar" adı verilen nöronlar bulunur [37].

Aşağıdaki şekilde klasik bir yapar sinir ağı modeli gösterilmektedir [3].



Şekil 4. 7 Yapay Sinir Ağı Yapısı

4.2.1 Yapay Sinir Ağları Bileşenleri

• **Giriş(input values):** Yapay sinir ağlarına dış ortamlar veya diğer bir hücreden gelen verilerdir.

• **Ağırlıklar(weights):** Hücreye gelen bilgilerin etkisini, ağırlığını gösterir.

• **Toplama Fonksiyonu(sum function):** Hücreye gelen bilgilerle, bu hücrelerin ağırlıklarını çarpımını toplar ve o hücrenin net giriş bilgisinin hesaplanmasını sağlar.

• **Etkinlik Fonksiyonu(activation function):** Hücreye gelen net bilgiyi analiz ederek, hücrenin bu giriş bilgisine göre karşılık üreteceği çıkış bilgisinin belirlenmesini sağlar.

• **Çıkış(output):** Aktivasyon fonksiyonlarının oluşturduğu çıkış bilgileridir. Bu bilgi,dış dünyaya, başka bir hücreye ya da kendisine giriş bilgisi olarak iletilebilir [38].

Yapısında esnek oldukları ve yüksek derecede doğrusal olmayanlık içerdiği için, YSA'lar verideki çok karmaşık ilişkileri tanımlamak için eğitilebilir. Bu nedenle, sinir ağı modelleri, ekonomide ve finansta çeşitli model tanıma, optimizasyon ve tahmin problemlerinin çözümü için yararlı olduğu kanıtlanmıştır Bir sinir ağı, varsayımsal olarak çok basit hesaplama birimleri veya nöronlardan oluşur. Yapay bir nöron, biyolojik nöronun eklektik bir simülasyonudur ve kendi dendritleri, sinapsları, hücre gövdesi ve akson terminallerinden oluşur. Yakındaki hücrelerden veya çevreden uyarı alır ve modifiye bir aksiyon potansiyeli veya sinir sinyali üretir.

Stergious ve Siganos 'e göre YSA yaklaşımı, karmaşık veya yanlış verilerden anlam çıkarmak için benzersiz bir kabiliyete sahiptir ve insanlar veya diğer bilgisayar teknikleri tarafından fark edilemeyecek kadar karmaşık olan kalıpları veya eğilimleri tespit etmekte yararlıdır. Ağ, yüzlerce tekli ünite, suni nöron ya da işleme elemanından bir YSA'nın oluşturulduğunu programlamaktan ziyade, deneyimler yoluyla kural prosesini eğiterek, bağlantıları nasıl değiştirebileceğinizi, performansı iyileştirebileceğinizi, kalıpları tanımlamayı ve genellemeler geliştirmeyi öğretebilir [39].

YSA'nın davranışı, üniteler için belirtilen ağırlıklara ve girdi-çıktısına (aktarma fonksiyonu) bağlıdır . Bu işlevler üç kategoriden birine, yani, doğrusal, eşikli ve sigmoid işlevlere sahiptir. Bir nöronun transfer fonksiyonu, birçok özelliğe sahip olacak şekilde seçilir.Bu da nöronu içeren ağı geliştirir veya basitleştirir (Duch and Jankwoski, 2001) [39].

4.2.2 YSA Varsayımları

Biyolojik sinir ağları çok karmaşıktır. Buna karşılık, ağın matematiksel modeli çok daha basitleştirilmiştir ve birkaç varsayım üzerine kurulmuştur:

1-Tüm nöronlar senkronize edilir. Bu, bir sinirden diğerine geçen sinyalin tüm bağlantılar için aynı zaman aldığı anlamına gelir. Sinyal işleme de senkronize edilir ve tüm nöronlar için aynıdır.

2- Her nöron, giriş sinyalinin kuvvetine bağlı olarak nöronun çıkış sinyalini belirleyen bir transfer fonksiyonuna sahiptir. Bu işlev zamandan bağımsızdır.

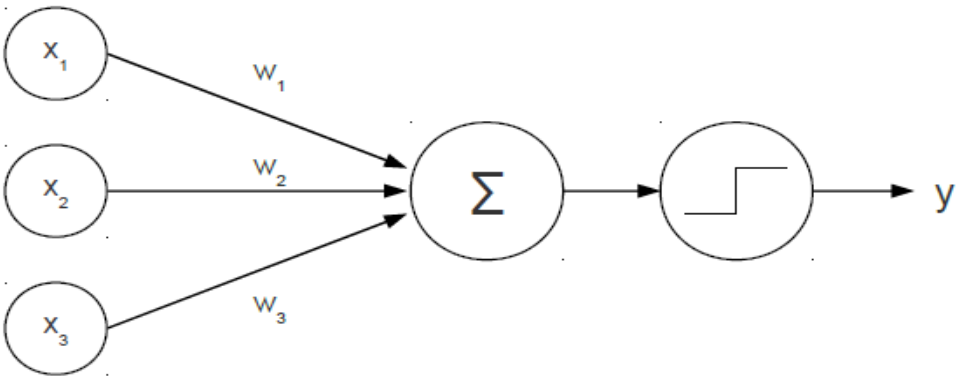
3- Sinyal sinapsı geçtiğinde doğrusal olarak değişir, yani sinyal değeri bazı sayılarla çarpılır. Bu sayıya sinaptik ağırlık denir. Sinaptik ağırlığın en önemli özelliği zamanla değişmesidir. Bu özellik, beynin farklı anlarda aynı girdide farklı tepki vermesini mümkün kılar [38].

4.2.3 Tek Katmanlı Sinir Ağları

Tek katmanlı sinir ağını en basit şekilde tanımlarsak, herhangi bir sayıda giriş ve tek bir çıkışa sahip tek bir nörondan oluşur. Ağırlık çarpanları, nörona yapılan girdilerin herbirine uygulanır ve sonuçlar toplanır. Daha sonra, eşik fonksiyonu, toplama dayalı

olarak nöronun çıktısını belirlemek için kullanılır. Bu çıktı noktanın sınıflandırmasını belirler. Matematiksel olarak, çıkış değeri y, aşağıdaki denklem kullanılarak belirlenir [40].

$$y = F_{eşik}(x1.w1 + x2.w2 + x3.w3) \quad (4.1)$$



Şekil 4. 8 Tek Katmanlı Sinir Ağı Yapısı

Ağırlıkların değerleri sinir ağının anahtarıdır. Nöronun girdi değerlerinin farklı kombinasyonlarına nasıl tepki vereceğini belirlerler. Doğru ağırlığın seçilmesi, eğitim boyunca denetlenmiş bir süreç aracılığıyla tekrar tekrar yapılır [40].

Tek bir nöron sınıflandırıcısının eğitimi, üretilen çıktı ile belirli bir girdi değeri kümesi için beklenen çıktı arasındaki karşılaştırmayı içerir. Her giriş değeri, iki değer arasındaki fark ve seçilen bir öğrenme oranı ile çarpılır. Sonuç daha sonra, girişin ağırlığına eklenir [40]. Bu açıklamanın formulasyonu aşağıdaki denklemde gösterilmektedir.

$$w_{yeni} = w_{eski} + oran_{\text{öğrenilmiş}} \left(çıktı_{beklenen} - çıktı_{gerçekleşen}\right).girdi \quad (4.2)$$

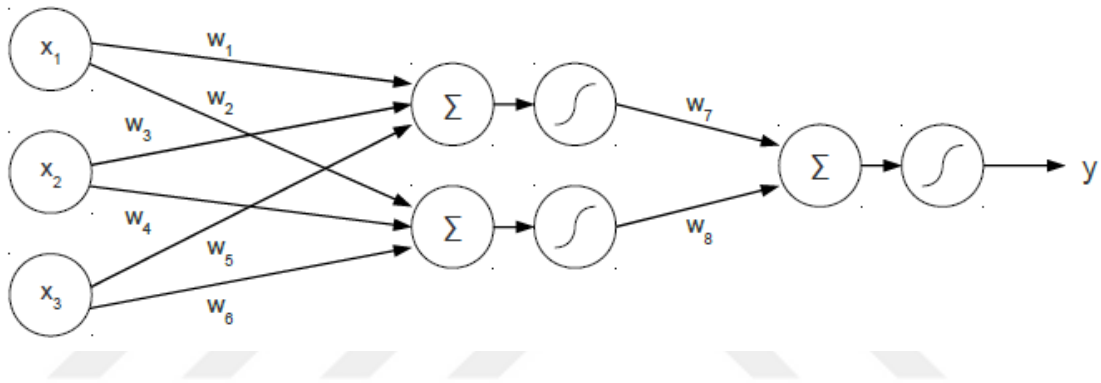
Nöronu her sınıf için birden çok numune ile eğittikten sonra, ağırlıkları gelecekteki sınıflandırma için, mümkün olan en düşük hata oranına sahip en uygun bir konfigürasyona yaklaşmalıdır. Bu, istenen çıktı tepkisini üretmek için giriş değerlerinin doğrusal toplamı için en iyi katsayıları bulmaya benzer [40].

Sınıflamanın doğrusal doğası nedeniyle sınıflar, uzayda bir hiper düzlem tarafından fiziksel olarak ayrılabilirlerse, yalnızca bir nöronla ayrılabilirler. Görüntü sınıflandırma

verileri, bu ayrışma seviyesinin yetersiz kalması için yeterince karmaşıktır. Bu uygulamalar için doğrusal olmayan bir sınır gereklidir ve daha karmaşık bir sinir ağı uygulanmalıdır [40]. Bu durum çok katmanlı sinir ağlarının konusudur.

4.2.4 Çok Katmanlı Sinir Ağları

Nöronlar kendi başlarına tüm sınıflandırma problemleri için yeterli olmamakla birlikte, doğrusal olmayan sınıflandırma durumlarını ele alabilen çok tabakalı sinir ağları için yapı taşları olarak da görev yapabilirler. Nöronların birkaç katmanını birbirine zincirleyerek, çok tabakalı bir algılayıcı olarak adlandırılan bir yapı oluşturulur. Bu daha karmaşık olan yapılar doğrusal olmayan sınıflandırmalar yapabilir [40].



Şekil 4. 9 Çok Katmanlı Sinir Ağı Yapısı

Bir çok tabakalı sınıflandırıcı sinir ağları, üç katmana sahiptir. Bir giriş katmanı, bir gizli katman ve bir çıktı katmanı şeklindedir. Gizli katman, her biri giriş değerlerinin ağırlıklı toplamlarını kabul eden ve bir eşik fonksiyonuna dayalı çıktı üreten nöronlara sahiptir. Çıktı katmanı, gizli tabaka çıktı değerlerinin ağırlıklı toplamlarını kabul eder ve bir eşik fonksiyonuna dayanan son sınıflandırıcı çıktılarını belirler. Çok katmanlı algılayıcıların (MLP (Multilayer Perceptron)) eşik fonksiyonu, eğitim amaçları için basamak fonksiyonunun türevlenebilir bir yakınlaştırması olmalıdır [40].

1980'lerin ortalarında Rumelhart, Hinton ve Williams tarafından verilen geri yayılım eğitim algoritmasının yayımlanmasıyla birlikte, bazen çok katmanlı algılayıcı (MLP) şebekeleri olarak adlandırılan çok katmanlı ileri besleme ağları, sinir ağı araştırmalarının temel dayanak noktası haline gelmiştir [41].

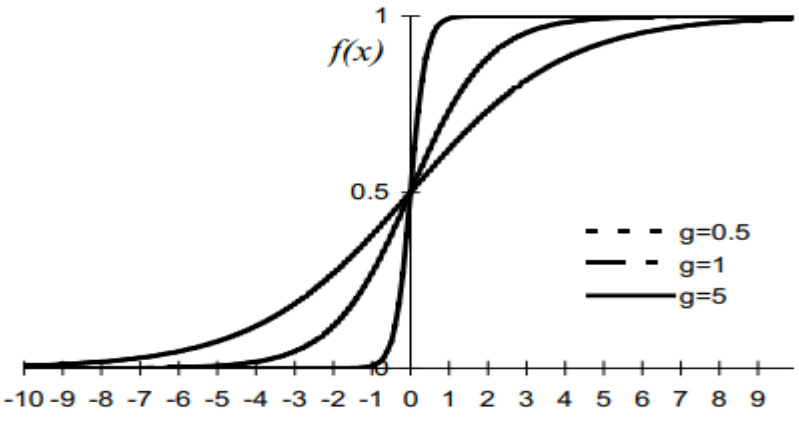
Bir sinir ağı, deneysel bilgiyi depolamak ve kullanım için hazır hale getirmek için doğal bir eğilimi olan, paralel dağıtılan bir işlemcidir. Bilgi, öğrenme yoluyla ağ tarafından edinilir. Nöronlar arasındaki bağlantıların gücü, bilgiyi depolamak için kullanılan sinaptik ağırlıklarla gerçekleştirilir. Öğrenme süreci, bilgiyi yakalamak için ağırlıkları bir öğrenme algoritması ile uyarlamanın bir prosedürüdür. Daha matematiksel olarak öğrenme sürecinin amacı, belirli bir ilişkiyi ağın giriş ve çıkış (çıktıları) arasında haritalamaktır [42].

Çok katmanlı ağların özellikleri de, ünitelerle kullanılan doğrusal olmayan doğruluklardan kaynaklanmaktadır. Ağdaki her nöron, ağdaki diğer nöronlardan girdi alır veya dış dünyadan girdi alır. Nöronların çıktıları diğer nöronlara veya dış dünyaya bağlıdır. Her girdi bir ağırlık ile nöronlara bağlıdır. Nöron, nöron için fiili çıktı üretmek için doğrusal olmayan bir aktarım fonksiyonundan geçen girdilerin ağırlıklı toplamını (aktivasyon olarak adlandırılır) hesaplar. En popüler doğrusal olmayan aktarım fonksiyonu sigmoidal tiptir [41].

Tipik bir sigmoid fonksiyonu şu şekildedir;

$$f(x) = \frac{1}{1+e^{-gx}} \tag{4.3}$$

g büyüdükçe, sigmoid fonksiyonu bir sinyal fonksiyonu haline gelir.



Şekil 4. 10 Sigmoid Fonksiyonu

4.3 Sınıflandırma Kalitesinin Ölçümü

Verileri sınıflandırırken kullanılan algoritmalar arasında hangi algoritmanın iyi olduğu konusunda seçim yapmak, analizin doğru sonuçlanması için çok önemlidir. Bu sınıflandırmayı yapmak için iki adımlı bir yol kullanılır [43].

4.3.1 Veride Ayrıştırma

Bu aşamada veri setinin bir kısmı bir model geliştirmek için iki veri kümesine bölünür. Birincisi, modeli eğitmek içindir. İkincisi, modelde budama yapmak içindir. Üçüncü bir veri seti, isteğe bağlı olarak modelin bağımsız bir değerlendirmesi için gereklidir. Bu üç veri seti sırasıyla Eğitim, Geçerlilik ve Test olarak SAS Enterprise Miner'da belirtilmektedir. Eğitim veri seti, kayıtları düğümlere atamak için kurallar geliştirmek (düğüm tanımları), her düğüm için posterior olasılıkların (hedefin her seviyesindeki vaka veya kayıtların oranı) hesaplanması ve her düğümün bir hedef seviyeye atanması (karar) görevlerini içerir. Doğrulama veri seti, bu modelleri değerlendirmek ve daha sonra en iyisini seçmek için kullanılır. Test veri seti, özellikle bir karar ağacı modelinin performansını diğer modellerin performansıyla karşılaştırmak istediğinde, nihai modelin bağımsız bir değerlendirmesi için kullanılır. Model verisinin yüzde kaçının bu üç amacın her birine tahsis edileceğine isteğe bağlı karar verebilir [5]. Veri setinin bölümü sırasında birkaç farklı yaklaşım söz konusudur.

-Yüzdesel Bölme : Bu yaklaşımda verinin belli bir yüzdesi test verisi için ayrılır. Bu oran genelde 1/3 olmakla birlike çalışmaya ve verinin özelliğine göre değişiklik gösterebilir.

-n-Bölme: Söz konusu teknikte veri n parçaya ayrılır. Parçaya ayrılan verinin ilk parçasından başlanarak n. parçaya gelene kadar tüm parçalar sırası geldikçe test verisi görevini üstlenir.

-1 Eksiltme: Veri ayrıştırmada kullanılan sonuncu yaklaşım olan 1 Eksiltme tekniği n-Bölme tekniği ile benzerlik göstermektedir. n sayısı gözlem sayısı ile eşit olarak alındığında n-bölme tekniği 1 Eksiltmeye dönüşmüş olur. Her gözlem test için kullanılırken diğer n-1 gözlem kullanılan algoritmada eğitim amaçlı rol alır [43].

4.3.2 Sınıflandırma Performans Ölçütleri

Algoritmalar sonucu oluşturulan sınıflandırmanın performansını değelendirirken birtakım ölçütlere bakılmaktadır. Sınıflandırma tekniği sonucu ortaya çıkan sonucu ve gerçek durumu gösteren sınıflama matrisi aşağıda yer almaktadır.

Çizelge 4. 2 Sınıflama Matrisi

		Gerçek Durum	
		Doğru (+)	Yanlış(-)
Sıfırlandırma Yöntemi Sonucu Tahmin	Doğru (+)	TP	FP
	Yanlış(-)	FN	TN

TP (True Positive):Doğru Pozitiflerin Sayısı

TN (True Negative):Doğru Negatiflerin Sayısı

FP (False Positive):Yanlış Pozitiflerin Sayısı

FN (False Negative):Yanlış Negatiflerin Sayısı

N = TP+ TN+ FP+ FN

Yukarıdaki tablo için; (yanlış negatif) false negative tahmini yanlış şekilde gerçekleşen olumsuz örnek miktarını, (doğru negaif) true negative tahmini doğru şekilde gerçekleşen olumsuz örnek miktarını, (yanlış pozitif) false positive tahmini yanlış şekilde gerçekleşen olumlu örnek miktarını ve (doğru pozitif) true positive tahmini doğru şekilde gerçekleşen olumlu örnek miktarını gösterir.

Bu verilerin kullanımıyla sınıflandırmada yaygın olarak kullanılan performans ölçütleri Çizelge 4.3'te gösterilmektedir.

Çizelge 4. 3 Sınıflandıma Performans Ölçütleri

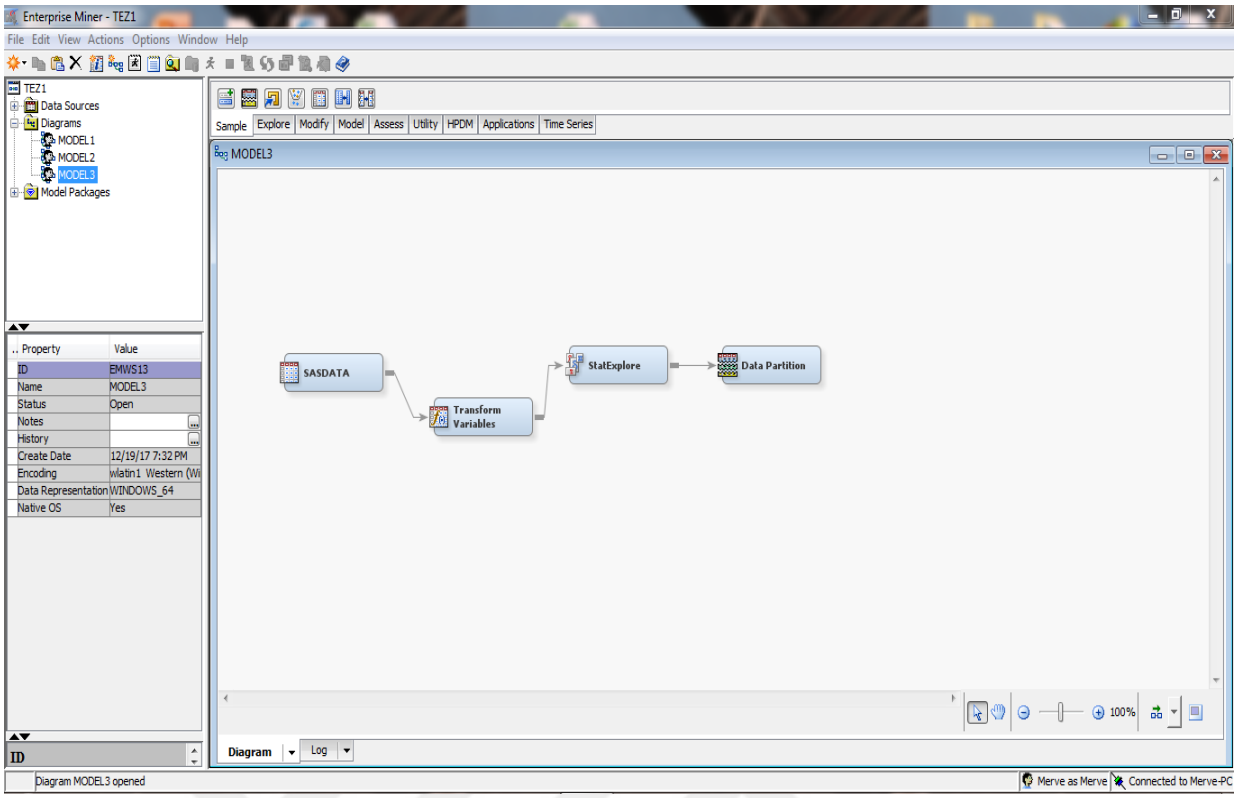
Doğruluk (Accuracy)	$$\boldsymbol{ACC = \frac{TP + TN}{N}}$$
F1 Ölçütü	$$\boldsymbol{F1 = \frac{2TP}{(2TP + FN + FP)}}$$
Duyarlılık	$$\boldsymbol{TPR = \frac{TP}{TP + FN}}$$
Belirlilik	$$\boldsymbol{TNR = \frac{TN}{TN + FP}}$$
Pozitif Kestirim Değeri	$$\boldsymbol{PPV = \frac{TP}{TP + FP}}$$
Matthews Kolerasyon Katsayısı	$$\boldsymbol{MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}}$$

BÖLÜM 5

UYGULAMA

Bu çalışmada sigortacılık sektörü kasko branşında, müşterilerin bir takım poliçe bilgileri üzerinden risk değerlendirmesi yapılmıştır. Uygulamadaki örnek birim sayısı 12.715' dir. Dikkate alınan değişkenler; sigortalının geçmiş ortalama hasar maliyeti, sigortalının yaşadığı bölge, aracın bölgedeki yoğunluğu, aracın ağırlığı, aracın kaç kilometrede olduğu, sigortalının cinsiyeti, kişinin medeni durumu, sigortalının yaşı, sigortalının deneyim yılı ve aracın bedelidir. Ulaşılmak istenen hedef değişken ise hasar frekansıdır. Uygulamanın amacı, bahsi geçen değişkenler ile sigorta şirketine yeni kayıt yaptıran bir kişinin hasar frekansını öngörmeye çalışmaktır. Modelleme teknikleri olarak karar ağaçları ve yapay sinir ağları kullanılmıştır. Bahsi geçen analizler SAS Enterprise Miner programı ile gerçekleştirilmiş olup uygulama aşamasında SAS Enterprise Guide programından da destek alınmıştır.

Aşağıdaki Şekil 5.11'de uygulamanın yapıldığı SAS Enterprise Miner programının arayüzü gösterilmiştir.



Şekil 5. 11 SAS Enterprise Miner Programı Arayüzü

5.1 Anakütlenin ve Değişkenlerin Tanımı

Uygulamaya ilk olarak 13.577 veri kümesi ile başlanmıştır. Özellikle yapar sinir ağları analizi için kayıp veriler modeli olumsuz yönde etkileyeceği için, kayıp verilerin bulunduğu satırlar tespit edilerek analiz dışı bırakılmıştır. Karar ağaçları kayıp veriler ile de çalışabilir özelliğe sahiptir [17].

Kayıp verlerin bulunduğu satırların analiz dışı bırakılmasından sonra, ilk aşamada aralık ölçek olan hedef değişkeni ordinal ölçek haline getirmek için bazı uç değerler analiz dışı bırakılmıştır ve sonuç olarak 12.715 adet veri ile analize başlanmıştır. Analizde kullanılan değişkenlerin isimleri, ölçekleri, modeldeki rolleri aşağıdaki Çizelge 5.4'te belirtilmiştir.

Çizelge 5. 4 Değişkenlerin Tanımlanması

Değişken ismi	Değişkenin Modeldeki Rolü	Değişken Tipi	Değişkenin Tanımı
ARACKM	girdi	aralık ölçek	kullanılan aracın kilometresi
ARACWG	girdi	aralık ölçek	kullanılan aracın ağırlığı
MEDEDUR	girdi	ikili ölçek	sigortalı medeni hal
CINSIYET	girdi	ikili ölçek	sigortalı cinsiyet
YAS	girdi	aralık ölçek	sigortalı yaş
ARBEDEL	girdi	aralık ölçek	aracın bedeli
ARACYOG	girdi	aralık ölçek	araç sayısı/ikamet edilen ilin nüfusu
GORTH	girdi	aralık ölçek	tazminat miktarı/hasar sayısı
BOLNUM	girdi	nominal ölçek	sigortalı yaşadığı bölge
DENYILI	girdi	aralık ölçek	sigortalının deneyim yılı
HASFREQ	-	aralık ölçek	ihbar hasar adet/kazanılan poliçe adet
HASORD	hedef	ordinal ölçek	HASREQ değişkeninin ordinal hali

Veri kümesinin oluşturulması analizin ilk aşamasını oluşturmuştur. Veri kümesi SASDATA şeklinde kodlanmıştır. Yukarıda belirtilen hedef değişkene etki eden bütün bağımsız değişkenler analizde aktif rol oynamıştır. Çizelge 5.4'te gösterilen değişkenleri açıklamak gerekirse; ARACKM değişkeni sigortası yapılan aracın sigorta yapılırken kaç kilometrede olduğunu, ARACWG değişkeni sigortası yapılan aracın ağırlığını, MEDEDUR değişkeni aracını sigorta ettiren kişinin medeni durumunu, CINSIYET değişkeni aracını sigorta ettiren kişinin cinsiyetini , YAS değişkeni aracını sigorta ettiren kişinin yaşını, ARBEDEL değişkeni sigortası yapılan aracın maddi değerini, ARACYOG değişkeni sigorta yaptıran kişinin ikamet ettiği şehirdeki araç sayısının şehirdeki nüfusa oranını, GORTH değişkeni sigorta yapan kişinin daha önceden yaşadığı kazalara bağlı olarak ödediği tazminat miktarının yaptığı kaza sayısına oranını, BOLNUM değişkeni aracını sigorta ettiren kişinin Türkiye'nin hangi bölgesinde yaşadığını, DENYILI değişkeni aracını sigorta ettiren kişinin araç kullanım deneyim yılını, HASRFEQ olarak isimlendirilen hasar frekansı değişkeni ise, bir branşta üretilen (bu çalışma için kasko branşında) poliçelerin hasarlanma oranını ifade etmektedir. Yani sigorta yaptıran kişinin daha önceden ihbar ettiği hasar sayısının,

kazanılan poliçe adedine bölümüdür. Bu değişken analizlerin hedef değişkenidir. Analizlerde sürekli yerine kesikli bir hedef değişken tercih edilmiştir. HASFREQ değişkeninin kategorik hali HASORD olarak adlandırılmıştır ve analizde bu şekilde kullanılmıştır. HASFREQ değişkeni analiz dışı bırakılmıştır. Aşağıdaki Çizelge 5.5'te ifade edildiği gibi kategorize edilmiştir.

Çizelge 5. 5 Hedef Değişkenin Ordinal Ölçek Kriterleri

HASORD = 0 eğer HASFREQ = 0 ise,
HASORD = 1 eğer 0 < HASFREQ < 0.40 ise,
HASORD = 2 eğer 0.40<HASFREQ <1.50

Veri kümesindeki değişkenlerden olan BOLNUM değişkeni aşağıdaki gibi kodlanmıştır.

Çizelge 5. 6 BOLNUM Değişkeninin Kodlama Şekli

BÖLGELER	
1	Akdeniz Bölgesi
2	Doğu Anadolu Bölgesi
3	Ege Bölgesi
4	Güneydoğu Bölgesi
5	İç Anadolu Bölgesi
6	Karadeniz Bölgesi
7	Marmara Bölgesi

Veri kümesindeki değişkenlerden olan CINSIYET ve MEDEDUR değişkenleri aşağıdaki gibi kodlanmıştır.

Çizelge 5. 7 MEDEDUR Değişkeninin Kodlama Şekli

KOD	MEDENİ DURUM
1	Evli
2	Bekar

Çizelge 5. 8 CINSIYET Değişkeninin Kodlama Şekli

KOD	CİNSİYET
1	Erkek
2	Kadın

5.1.1 Betimsel İstatistikler ve Veri Dönüşümü

Analizde kullanılan değişkenlerin ölçek tiplerinin özeti aşağıdaki tabloda özetlenmiştir. Görüldüğü üzere toplamda 10 adet bağımsız değişken bulunmaktadır. Bunlardan 2 tanesi ikili (binary) ölçek, 7 tanesi aralık (interval) ölçek, 1 tanesi nominal ölçek ve hedef değişken olan HASORD değişkeni de sıralı (ordinal) ölçek olarak analizde rol almıştır.

Çizelge 5. 9 Analizde Kullanılan Değişkenlerin Ölçek Tipi Özeti

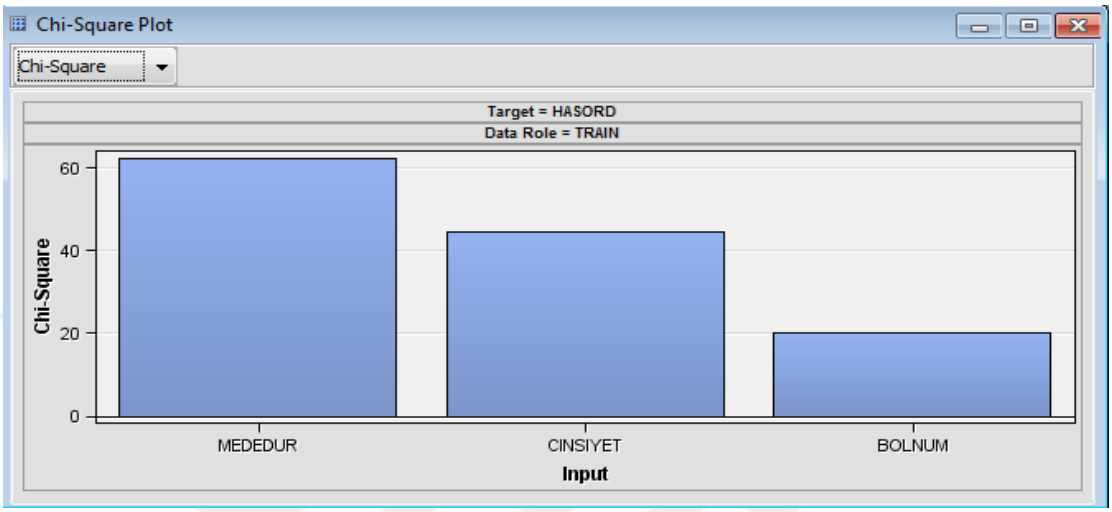
Değişkenin Rolü	Ölçme Düzeyi	Frekans
GİRDİ	İKİLİ	2
GİRDİ	ARALIK	7
GİRDİ	NOMINAL	1
HEDEF	ORDINAL	1

HASORD olan hedef değişkeninin “0”, ”1” ve “2” şeklinde kategorize edilmiş seviyelerinin, aşağıdaki tabloda belirtildiği üzere anakütleye dağılımı sırasıyla %57, %31 ve %11 şeklindedir.

Çizelge 5. 10 Hedef Değişken Seviyelerinin Anakütleye Dağılımı

Değişkenin Adı	Değişkenin Rolü	Seviye	Frekans	%
HASORD	HEDEF	0	7280	57.25
HASORD	HEDEF	1	3977	31.27
HASORD	HEDEF	2	1458	11.46

Aşağıdaki grafik, analizde kullanılan nominal ve ikili ölçekli değişkenlerin hedef değişken ile çapraz tablo (cross-table) ki-kare değerlerini göstermektedir. Medeni durumun en yüksek ki kare değerini aldığı, cinsiyet değişkeni ve sigortalının yaşadığı bölge değişkeninin medeni durum değişkenini takip ettiği gözlenmiştir.



Şekil 5. 12 Nominal ve İkili Ölçekli Değişkenler İçin Ki-Kare Değerleri

Analizlerde kullanılan değişkenlerin açıklamalarından sonra, sürekli değişkenlerin dağılımının ve aykırı değerlerin tespiti için Çizelge 5.11'de betimsel istatistikler açıklanmıştır.

Çizelge 5. 11 Dönüşümü Öncesi Sürekli Değişkenlerin Betimsel İstatistikleri

Değişken Adı	Ortalama	Standart Sapma	Kayıp Veri	Min.	Medyan	Max.	Çarpıklık	Basıklık
ARACKM	11230.65	8363.76	0	10	9522	301205	6.50	140.39
ARACWG	1208.59	193.00	0	740	1190	2717	1.65	5.96
ARACYOG	0.26	0.06	0	0.03	0.25	0.47	0.17	1.34
ARBEDEL	29556.02	17072.65	0	3811	26000	414067	4.91	52.35
DENEYIL	3.06	3.25	0	0	2	20	1.58	2.86
GORTH	500.61	1213.29	0	0	0	22579	6.14	60.49
YAS	46.08	12.601	0	18	45	92	0.38	-0.46

Tablodaki “missing” sütununda da görüldüğü üzere, veri kümesini oluşturma aşamasında kayıp veriler analiz dışında bırakıldığı için değişkenlerimizde kayıp veri gözlenmemiştir. Aralık ölçekteki değişkenlerin bazı istatistiksel değerleri incelediğinde; aracın sigortalanırken kaç kilometrede olduğu gösteren ARACKM değişkeninin, aracı bedelini gösteren ARBEDEL değişkeninin ve sigorta yapan kişinin daha önceden yaşadığı kazalara bağlı ödediği tazminat ortalamasını gösteren GORTH değişkenlerinin basıklık (kurtosis) değerlerinin yüksek çıktığı görülmektedir. Bu yüksek değerleri normalize ederek, değişken dağılımlarının normal dağılıma yaklaştırılması amacıyla dönüşüm (transformasyon) işlemi tercih edilmiştir. Bu dönüşüm işleminde program, değişkenleri normalize ederken en uygun dönüşümü (üstel, logaritmik, karekök vs.) yapmaktadır. Verilerin normallik dönüşümleri sonucu, yüksek değerlerin normalize edildiği betimsel istatistikler Çizelge 5.12’ de gösterilmektedir.

Çizelge 5. 12 Dönüşümü Sonrası Sürekli Değişkenlerin Betimsel İstatistikleri

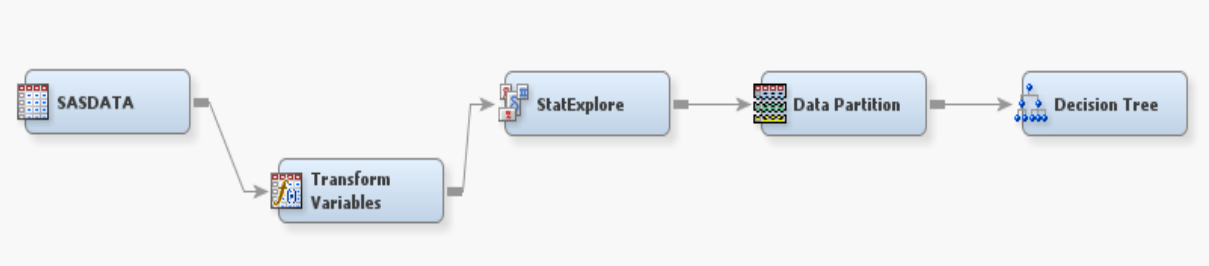
Değişken Adı	Ortalama	Standart Sapma	Kayıp Veri	Min.	Medyan	Max.	Çarpıklık	Basıklık
PWR_ARACKM	0.4232	0.0679	0	0	0.4215	1	0.1941	2.3209
PWR_ARACWG	0.6875	0.0688	0	0	0.6907	1	-0.0347	2.5922
ARACYOG	0.2656	0.0685	0	0.3366	0.2514	0.4769	0.1745	1.3489
LOG_ARBEDEL	0.0601	0.0362	0	0	0.0526	0.6931	3.7560	28.9699
SQRT_DENEYIL	0.3202	0.2248	0	0	0.3162	1	0.1425	-0.5937
LOG_GORTH	0.0207	0.0463	0	0	0	0.6931	4.8964	36.4400
SORT_YAS	0.5993	0.1428	0	0	0.6040	1	-0.1353	-0.4634

Verilerin dönüşümü işlemi tamamlandıktan sonra, analiz adımına geçmek için karar ağacı ve yapay sinir ağı analizlerinde eğitim, geçerlilik ve test kümeleri olarak kullanılacak veri kümelerinin oranları seçilmesi gerekmektedir. Bahsedilen oranların seçimi için tek tip kurallar bulunmamakla birlikte, sıklıkla kullanılan oranlar bulunmaktadır. Eğitim, geçerlilik ve test kümelerinin oranları sırasıyla %60, %30 ve %10 olarak belirlenmiştir. HASORD olan hedef değişkeninin “0”,”1” ve “2” şeklinde kategorize edilmiş hasarlama oranının anakütleye dağılımı sırasıyla %57, %31 ve %11 olup, oluşturulan eğitim,

geçerlilik ve test kümesinde bu dağılımın korunması amaçlanmıştır. Eğitim, geçerlilik ve test kümelerindeki veriler tabakalı örneklem yöntemine göre seçilmiştir.

5.2 Karar Ağaçları Analizi

Verilerin dönüşümü, betimsel istatistiklerin tablo şeklinde açıklanması ve anakütlenin eğitim, geçerlilik ve test bölümlerine ayrımı işleminden sonra karar ağaçları analizine geçilmiştir. Oluşturulacak karar ağacı her yaprak düğüm veya segment için beklenen hasar frekansını belirler. Karar ağacı analizinin yapılabilmesi için uygulanan adımlar aşağıdaki Şekil 5.13'te gösterilmiştir.



Şekil 5. 13 SAS Enterprise Miner Karar Ağaçları Adımları

SAS Enterprise Miner'da Karar Ağacı Düğümü Ayrıştırma Kriteri Özelliğini (Splitting Rule) seçerken, eğer hedef değişken ordinal ise, Ordinal Target Criterion özelliğindeki seçeneklerden ayırma yöntemlerinden Entropi veya Gini seçilebilir. Çalışmanın ilk bölümlerinde değinilen bu yöntemlerden bu çalışmada aşağıdaki tabloda gözüktüğü üzere Entropi yöntemi tercih edilmiştir. Diğer özellikler varsayılan şeklilde kullanılmıştır.

Splitting Rule	
Interval Target Crit	ProbF
Nominal Target Crit	ProbChisq
Ordinal Target Crite	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorica	5
Node	
Leaf Size	6
Number of Rules	5
Number of Surroga	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Şekil 5. 14 Karar Ağacı Düğümü Ayrıştırma Özellikleri

Alt Katman Bölme Düğüm Özelliğinde (Split Search Subtree Node), Method seçeneği için Değerlendirme (Assessment) ve Ölçme Ölçümü Özelliği (Assessment Measure) için Ortalama Kare Hata (Average Square Error) seçilmiştir. Bu seçenekler, farklı şekillerde denedikten sonra daha iyi başarı verdiği için tercih edilmiştir. Bu ekrandaki diğer seçenekler varsayılan şeklinde kullanılmıştır.

Node Split Search Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Şekil 5. 15 Alt Katman Bölme Düğüm Özellikleri

Gerekli özelliklerin seçiminden sonra karar ağacı düğümü çalıştırılmıştır. Çalışan düğüm sonucunda oluşan karar ağacındaki yaprak oluşumunu belirleyen değişkenler için elde edilen sonuçlar Çizelge 5.15'de gösterilmiştir. Analizde kullanılan bağımsız

değişkenlerden geçmiş ortalama tazminat miktarı (GORTH) değişkeni ağaç oluşumunda en büyük etkiye sahip olduğu ortaya çıkmıştır. Bu değişkeni, aracın kilometresi (ARACKM), sigortalının yaşı (YAS), sigortalının yaşadığı bölge (BOLNUM) ve sigortalının deneyim yılı (DENEYIL) değişkenleri takip etmiştir. GORTH, ARACKM, YAS, BOLNUM ve DENEYIL haricindeki analizde kullanılan diğer değişkenler karar ağacı oluşumunda çok az etkiye sahip olduğu için tabloda yer almamıştır.

Çizelge 5. 13 Değişkenlerin Önemlilik Dereceleri

Değişken	**Açıklama**	**Ayırma Kurallarının Sayısı**	**Önemlilik**	**Geçerlilik Veri Setine Ait Önemlilik**	**Geçerlilik Veri Setinin Önemliliğinin Eğitim Veri Setinin Önemliliğine Oranı**
LOG_GORTH	Transformed GORTH	3	1.000	1.000	1.000
PWR_ARACKM	Transformed ARACKM	1	0.0599	0.000	0.000
SQRT_YAS	Transformed YAS	2	0.0494	0.0247	0.4998
SQRT_DENEYIL	Transformed DENEYIL	1	0.0233	0.0000	0.0000
BOLNUM	BOLNUM	1	0.0265	0.0227	0.8562

Sonuçlanan analizin doğruluk, duyarlılık ve belirlilik oranlarını vererek modelin anlamlılığını ölçmeye yarayan sınıflama matrisi, bakılması gereken bir başka analiz çıktısıdır. Pek çok sınıflandırma performans ölçütleri olmakla birlikte doğruluk, duyarlılık ve belirlilik ölçütleri yaygın olarak kullanılan ölçütlerdendir. Bu çıktıda (yanlış negatif) false negative tahmini yanlış şekilde gerçekleşen olumsuz örnek miktarını, (doğru negaif) true negative tahmini doğru şekilde gerçekleşen olumsuz örnek miktarını, (yanlış pozitif) false positive tahmini yanlış şekilde gerçekleşen olumlu örnek miktarını ve (doğru pozitif) true positive tahmini doğru şekilde gerçekleşen olumlu örnek miktarını önceki bölümde belirtmiştir. Karar ağacı analizi sonucu elde edilen sınıflama matrisi aşağıdaki tabloda gösterilmiştir.

Çizelge 5. 14 Karar Ağacı Modeli Sınıflandırma Matrisi

Sınıflandırma Matrisi, Hedef = HASORD			
Yanlış Negatif	Doğru Negatif	Yanlış Pozitif	Doğru Pozitif
152	3369	705	389

Sınıflama matrisi yardımıyla hesaplanan doğruluk, duyarlılık ve belirlilik oranları;

$$Doğruluk = \frac{389 + 3369}{389 + 3369 + 152 + 705} = \%81$$

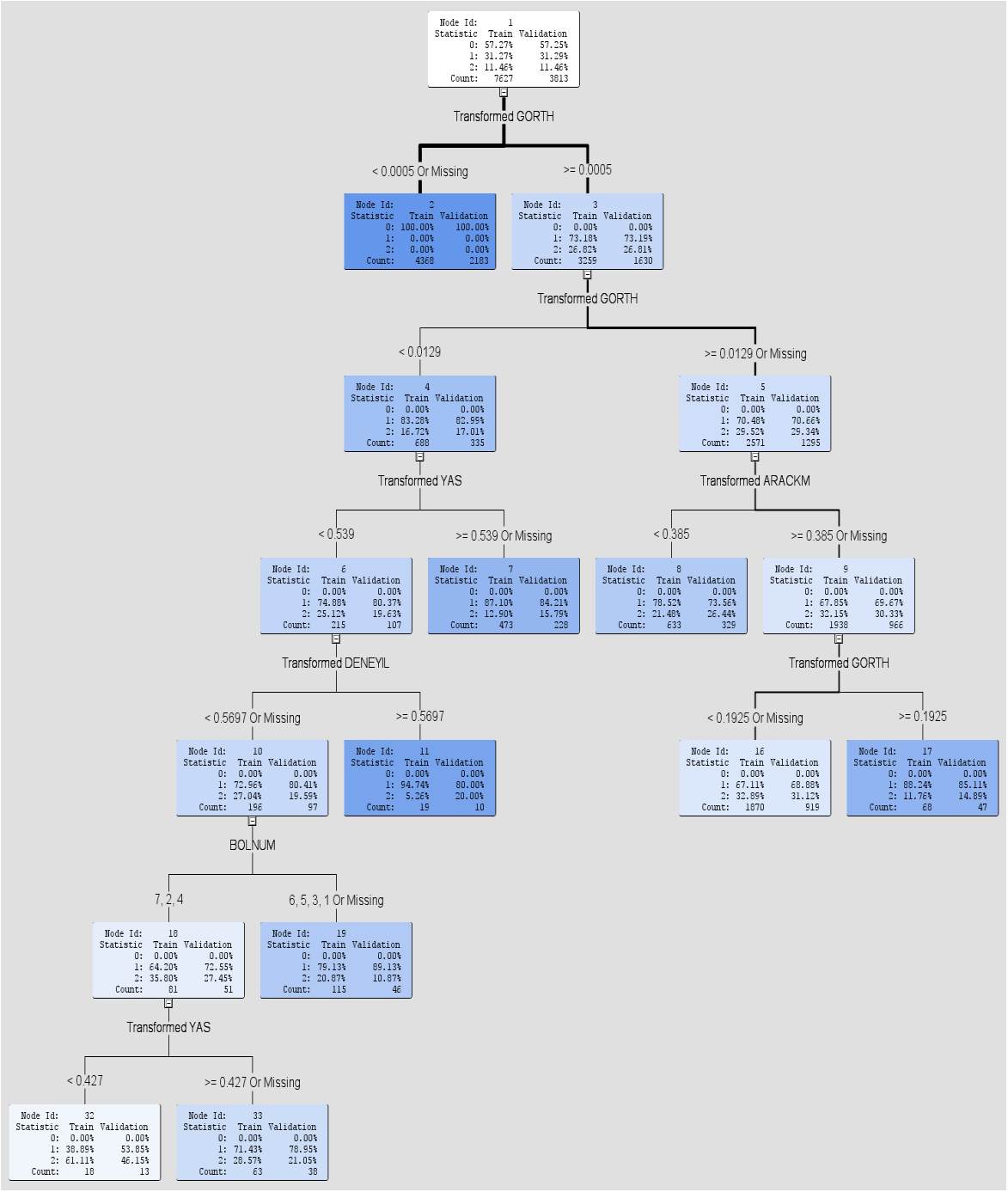
$$Duyarlılık = \frac{389}{389 + 152} = \%71$$

$$Belirlilik = \frac{3369}{3369 + 705} = \%82$$

şeklinde olup, karar ağacı algoritma sonucuna bakıldığında yapılanan tahminlemenin iyi bir seviye doğrulama oranıyla sonuçlandığı görülmüştür. Oluşan model sonucu duyarlılık ve belirlilik sonuçları da modelin başarısı açısından iyi bir seviyede gözlemlenmiştir.

Analiz sonucunda oluşan karar ağacı diyagramı Şekil 5.16'da görülmektedir.

Şekil 5. 16 Karar Ağacı Diyagramı



Entopi karar ağacı diyagramının kurallar dizisini gösteren şema yukarıda gösterilmektedir. Bu kuralları teker teker maddeler halinde yazacak olursak,

- Eğer GORTH<0.0005 veya kayıp veri ise

HASORD 0 için Öngörülen hasar frekansı : 0.100'dür.

- Eğer GORTH>=0.0005 veya kayıp veriyse ise

HASORD 1 için Öngörülen hasar frekansı : 0.73'dür.

HASORD 2 için Öngörülen hasar frekansı : 0.26'dır.

- Eğer GORTH>=0.0129 veya kayıp veriyse ve ARACKM<0.385 ise

HASORD 1 için Öngörülen hasar frekansı : 0.78'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.21'dir.

- Eğer GORTH<0.0129 veya kayıp veriyse ve ARACKM>=0.385 ise

HASORD 1 için Öngörülen hasar frekansı : 0.67'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.32'dir.

- Eğer GORTH<0.0129 ve ARACKM>=0.385 veya kayıp veriyse ve GORT<0.1925 ise

HASORD 1 için Öngörülen hasar frekansı : 0.67'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.32'dir.

- Eğer GORTH<0.0129 ve ARACKM>=0.385 veya kayıp veriyse ve GORT>=0.1925 ise

HASORD 1 için Öngörülen hasar frekansı : 0.80'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.11'dir.

- Eğer GORTH>=0.0129 ve YAS>=0.539 ise

HASORD 1 için Öngörülen hasar frekansı : 0.87'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.12'dir.

- Eğer ARACKM>=0.385 veya ve YAS<0.539 ise

HASORD 1 için Öngörülen hasar frekansı : 0.74'dür.

HASORD 2 için Öngörülen hasar frekansı : 0.25'dir.

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL>=0.5697 ise

HASORD 1 için Öngörülen hasar frekansı : 0.94'dür.

HASORD 2 için Öngörülen hasar frekansı : 0.05'dir.

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ise

HASORD 1 için Öngörülen hasar frekansı : 0.72'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.27'dir.

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ve BOLNUM=6,5,3,1 veya kayıp veri ise

HASORD 1 için Öngörülen hasar frekansı : 0.79'dur.

HASORD 2 için Öngörülen hasar frekansı : 0.20'dir.

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ve BOLNUM=7,2,4 ise

HASORD 1 için Öngörülen hasar frekansı : 0.64'dür.

HASORD 2 için Öngörülen hasar frekansı : 0.35'dir.

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ve BOLNUM=7,2,4 ve YAS>=0.427 ise

HASORD 1 için Öngörülen hasar frekansı : 0.71'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.28'dir.

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ve BOLNUM=7,2,4 ve YAS<0.427 ise

HASORD 1 için Öngörülen hasar frekansı : 0.38'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.61'dir.

Yukarıdaki kurallar dizisi incelendiğinde ilk düğümde müşterilerin GORTH değişkenine göre hasar gerçekleştirmeyen (HASORD=0) ve hasar gerçekleştiren (HASORD=1 ve 2) olarak ikiye ayrıldıkları görülmektedir.

Daha sonraki düğümlerdeki değişkenler de incelendiğinde, aracını sigorta yapmak için gelen potansiyel bir müşterinin GORTH, ARACKM, YAS, BOLNUM ve DENEYIL değişkenleri kullanılarak atanacağı risk grubuna göre, prim ödemesi değerlendirilebilir. Örneğin, yukarıda verilen 14 kural içerisinde HASORD =2 olduğu durumdaki en yüksek öngörülen hasar frekansı;

- Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ve BOLNUM=7,2,4 ve YAS<0.427 ise

HASORD 1 için Öngörülen hasar frekansı : 0.38'dir.

HASORD 2 için Öngörülen hasar frekansı : 0.61'dir.

adımında karşımıza çıkmıştır. Bu adıma göre yeni sigorta yaptıracak bir kişi için, Eğer GORTH<0.0129 ve YAS<0.539 ve DENEYIL<0.5697 ve YAS<0.427 ve BOLNUM= Marmara Bölgesi, Doğu Anadolu Bölgesi ve Güneydoğu Anadolu Bölgesi ise hasar yapma olasılığı en yüksektir.

Oluşturulan karar ağacı modelinin sonucunda, Eğitim, Geçerlilik ve Test veri setlerinde SAS Enterprise Miner Programı tarafından hesaplanabilen uyum istatistikleri Çizelge 5.15'te sunulmaktadır.

Çizelge 5. 15 Karar Ağacı Uyum İstatistikleri

Uyum İstatistikleri				
Hedef = HASORD				
Uyum İstatistikleri		Eğitim	Geçerlilik	Test
NOBS	Sum of Frequencies	7627.00	3813.00	1275.00
MISC	**Misclassification Rate**	**0.11**	**0.11**	**0.12**
MAX	Max. Absolute Error	0.95	0.95	0.95
SSE	Sum of Square Error	1233.57	630.76	211.44
ASE	Average Squared Error	0.05	0.06	0.06
RASE	Root Average Squared Error	0.23	0.23	0.24
DIV	Divisor for ASE	22881.00	11439.00	3825.00
DFT	Total Degrees of Freedom	15254.00	.	.
APROF	Average Profit for HASORD	1.89	1.89	1.88
PROF	Total Profit for HASORD	14384.00	7188.00	2401.00

Çıkan değerlere baktığımızda, modelin başarı oranını gösteren 'Misclassification Rate' sonucu geçerlilik veri setinde 0.11, test veri setinde ise 0.12 seviyesindedir. Bu sonuç, verilerin yanlış sınıflandırma yüzdesini göstermektedir. Verilerin doğru sınıflandırma oranı, yani modelin başarısı geçerlilik veri setine (100-11=89) %89, test veri setinde ise (100-12=88) %88 olarak karşımıza çıkmaktadır.

5.3 Yapay Sinir Ağları Analizi

Karar ağaçları analizinin ardından ikinci uygulanan model yapay sinir ağları modeli olmuştur. Bir modelde, dizideki ilk katman giriş katmanı ve son katman çıktı veya hedef katmanıdır. Giriş katmanı ve çıktı katmanı arasında bir takım gizli katmanlar olabilir. Gizli katmandaki birimler gizli birimler olarak adlandırılır. Gizli birimler ara hesaplamalar yapar ve sonuçları bir sonraki katmana geçirir. Gizli birimler tarafından gerçekleştirilen hesaplamaların amacı, önceki katmandan aldıkları girdileri birleştirmek ve birleştirilen değerler üzerinde bir matematiksel dönüşüm gerçekleştirmektir.

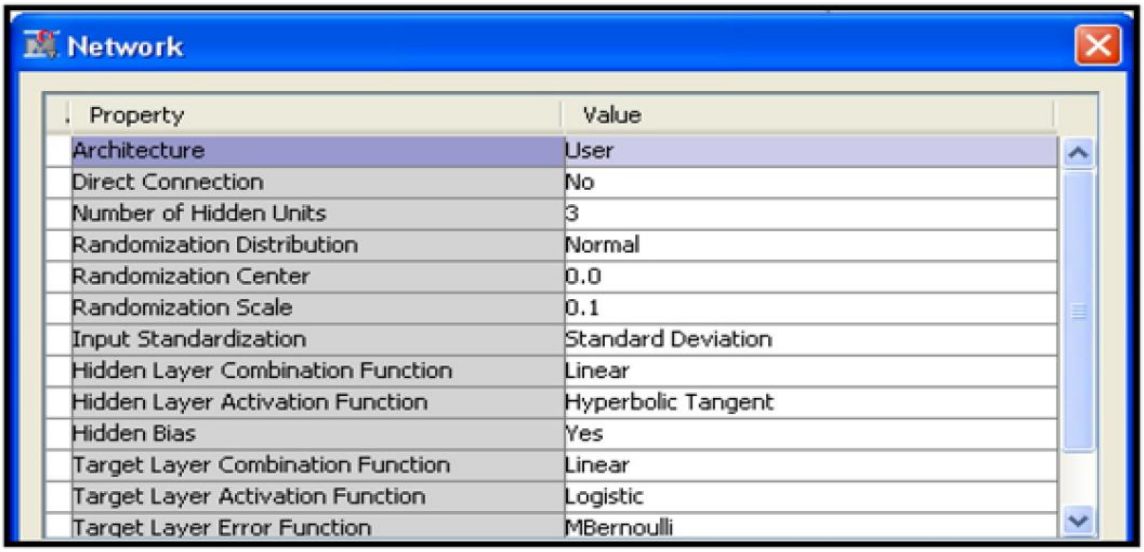
İki gizli katmana ve bir çıktı katmanına sahip bir sinir ağı, üç katmanlı bir ağ olarak adlandırılır. Giriş katmanındaki birimler, girişlerin standartlaştırılması dışında herhangi bir hesaplama yapmadığı için, ağın karakterizasyonunda girdi katmanı sayılmaz.

SAS Enterprise Miner'de, gizli birimler tarafından girdileri birleştirmek için kullanılan formüller Gizli Katman Kombinasyon Fonksiyonları (Hidden Layer Combination Function) bölümünde yer alır. Birleştirilmiş değerleri dönüştürmek için gizli birimler tarafından kullanılan formüller ise Gizli Katman Etkinleştirme Fonksiyonları (Hidden Layer Activation Function) bölümündedir. Birleşim ve aktivasyon işlemleri tarafından üretilen değerlere çıkışlar denir. Belirtildiği gibi, bir katmanın çıktıları bir sonraki katmanın girişleri haline gelir. Hedef katmandaki birimler ayrıca kombinasyon ve etkinleştirme işlemlerini gerçekleştirir.

Hedef katmandaki birimler tarafından girdileri birleştirmek için kullanılan formüllere Hedef Katman Kombinasyon Fonksiyonları (Target Layer Combination Function), birleştirilen değerleri dönüştürmek için kullanılan formüllere Hedef Katman Etkinleştirme Fonksiyonları (Target Layer Activation Function) denir.

Hedef katmanın ürettiği çıktıların yorumlanması, kullanılan Hedef Katman Etkinleştirme Fonksiyonuna bağlıdır. Gizli katmandaki ve hedef katmandaki kombinasyon ve aktivasyon işlevleri bir sinir ağı mimarisinin kilit unsurlarıdır.

SAS Enterprise Miner, bu işlevler için çok sayıda seçenek sunmaktadır. Dolayısıyla, Gizli Katman Kombinasyonu Fonksiyonu, Gizli Katman Etkinleştirme Fonksiyonu, Hedef Katman Kombinasyon Fonksiyonu ve Hedef Katman Etkinleştirme Fonksiyonunun kombinasyon seçenekleri fazladır ve her biri farklı bir sinir ağı modeli üretir [5].



Şekil 5. 17 Yapay Sinir Ağı Modeli Kriterleri Ekranı 1

Hedef katmanın ürettiği çıktıların yorumlanması kullanılan Hedef Katman Etkinleştirme Fonksiyonuna (Target Layer Activation Function) bağlıdır. Bu bağlamda çalışmadaki hedef değişken ordinal olduğu için Hedef Katman Etkinleştirme İşlevi değeri olarak Lojistik seçeneği seçilmiştir. Bu seçim, hedef katmanın çıktılarının yanıt verme ve yanıt vermeme olasılıklarını sağlar.

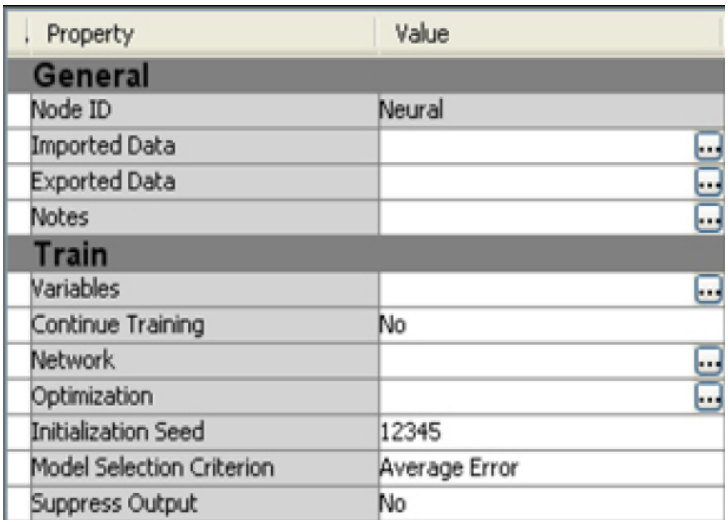
Gizli her bir birimde, girdilerin ağırlıklı toplamını hesaplamak için doğrusal kombinasyon fonksiyonu kullanılmıştır. Birinci gizli katmandaki ilk birim için i. kayıttaki girdilerin ağırlıklı toplamı aşağıdaki gibi hesaplanır:

$$\eta i11 = w011 + w111xi1 + w211xi2 + ... + wp11xip \quad (5.1)$$

Burada w111, w211, ... ,wp11, p girişlerinin her biri için bir ağırlık olarak daha sonra açıklanacak olan ve yinelemeli algoritma tarafından tahmin edilecek ağırlıklardır. w011'e bias (sapma) denir. ηi11 ise, veri kümesindeki i. kişi veya kayıt için ağırlıklandırılmış toplamıdır.

Her gizli ünite için ara çıktıyı hesaplamak adına bir hiperbolik teğet aktivasyon fonksiyonu kullanılmıştır. Analizde kullanılan hiperbolik teğet aktivasyon fonksiyonu dönüşümünün etkisi, ∞'den + ∞'a kadar değişebilen ηi11 değerlerini -1 ile +1 arasındaki daha dar aralıklarla eşlemektir [5].

Çıktı katmanında doğrusal kombinasyon fonksiyonu olarak lojistik aktivasyon fonksiyonu, hata fonksiyonu olarak Mbernoulli seçeneği seçilmiştir. Hedef Katman Kombinasyon İşlevi (Target Layer Combination Function) özelliği doğrusal (linear) olacak şekilde seçildiğinden, sinir ağı düğümü, gizli katman çıktılarını gelen her gözlem için doğrusal öngörücüler şeklinde hesaplar. HASORD hedef değişkeni 0, 1 ve 2 değerlerini alarak, sinir ağı düğümü iki doğrusal öngörücü hesaplar.



Şekil 5. 18 Yapay Sinir Ağı Modeli Kriterleri Ekranı 2

Çalışmada Model Seçimi Kriteri (Model Selection Criterion) özelliği Ortalama Hata (Average Error) olarak ayarlanmıştır. Eğitim süreci boyunca, Sinir Ağı düğümü, her iterasyonda olmak üzere bir dizi aday model oluşturur. Model Seçim Kriterini Ortalama Hata'ya ayarlandığında, sinir ağı düğümünde Geçerlilik veri seti kullanılarak hesaplanan en küçük hataya sahip olan model seçilir.

Yapay sinir ağı için gerekli analiz kriterleri seçildikten sonra analiz çalıştırılarak çıktıları alınmıştır. Doğruluk, duyarlılık ve belirlilik sonuçları aşağıdaki Çizelge 5.16'da gösterilmiştir.

Çizelge 5. 16 Yapay Sinir Ağı Modeli Sınıflandırma Matrisi

Sınıflandırma Matrisi, Hedef = HASORD			
Yanlış Negatif	Doğru Negatif	Yanlış Pozitif	Doğru Pozitif
265	6507	246	497

$$Doğruluk = \frac{497 + 6507}{497 + 6507 + 265 + 1327} = \%81$$

$$Duyarlılık = \frac{497}{497 + 265} = \%65$$

$$Belirlilik = \frac{6507}{6507 + 1327} = \%83$$

Doğruluk ve Belirlilk seviyeleri iyi bir sonuç gösterirken, Duyarlılık seviyesinin daha düşük çıktığı gözlemlenmiştir.

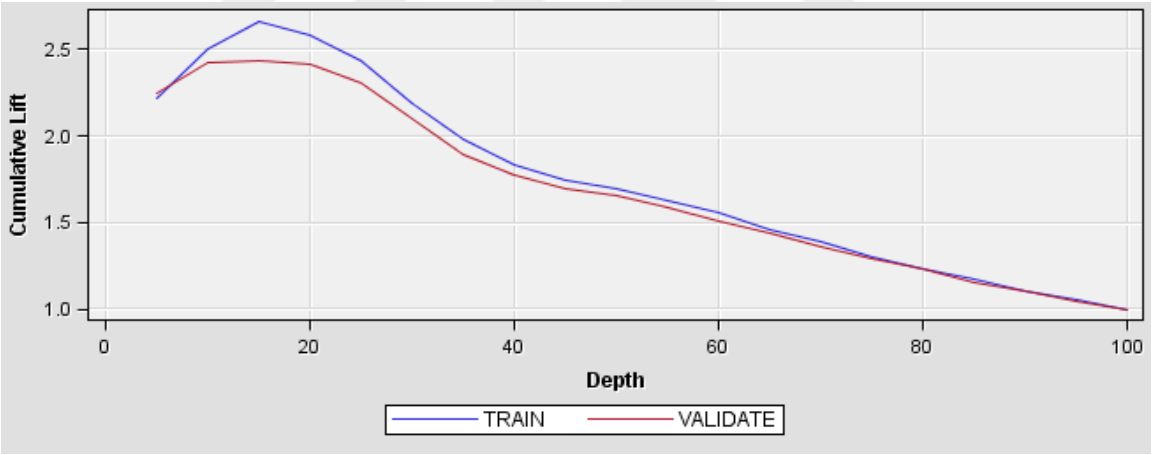
Çizelge 5. 17 Yapay Sinir Ağı Skor Sıralaması

Target Label	Data Role	Event	Cumulative % Response	% Captured Response	Cumulative % Captured Response	Depth	Lift	Cumulative Lift
	TRAIN	2	25.39267	11.0984	11.0984	5	2.215903	2.215903
	TRAIN	2	28.57143	13.84439	24.94279	10	2.771422	2.493298
	TRAIN	2	30.48035	14.98856	39.93135	15	2.992611	2.659881
	TRAIN	2	29.55439	11.67048	51.60183	20	2.33624	2.579077
	TRAIN	2	27.84478	9.153318	60.75515	25	1.832345	2.429887
	TRAIN	2	24.98908	4.691076	65.44622	30	0.936619	2.180683
	TRAIN	2	22.69663	3.89016	69.33638	35	0.778747	1.980631
	TRAIN	2	20.94395	3.775744	73.11213	40	0.755842	1.827683
	TRAIN	2	19.95339	5.263158	78.37529	45	1.05084	1.741242
	TRAIN	2	19.37598	6.17849	84.55378	50	1.236833	1.690854
	TRAIN	2	18.56973	4.576659	89.13043	55	0.916173	1.620495
	TRAIN	2	17.78458	4.004577	93.13501	60	0.799553	1.551979
	TRAIN	2	16.70028	1.601831	94.73684	65	0.32066	1.457358
	TRAIN	2	15.86439	2.173913	96.91076	70	0.435182	1.384413
	TRAIN	2	14.92746	0.800915	97.71167	75	0.159911	1.302651
	TRAIN	2	14.09374	0.686499	98.39817	80	0.137426	1.229896
	TRAIN	2	13.38886	0.915332	99.3135	85	0.183235	1.168385
	TRAIN	2	12.65841	0.114416	99.42792	90	0.022844	1.104642
	TRAIN	2	12.04803	0.457666	99.88558	95	0.091617	1.051376
	TRAIN	2	11.45929	0.114416	100	100	0.022904	1
	VALIDATE	2	25.65445	11.21281	11.21281	5	2.238454	2.238454
	VALIDATE	2	27.74869	13.04348	24.25629	10	2.603915	2.421184
	VALIDATE	2	27.7972	12.12815	36.38444	15	2.433927	2.425417
	VALIDATE	2	27.654	11.89931	48.28375	20	2.375502	2.412922
	VALIDATE	2	26.41509	9.382151	57.6659	25	1.872992	2.304823
	VALIDATE	2	24.03846	5.263158	62.92906	30	1.056233	2.097452
	VALIDATE	2	21.64794	3.203661	66.13272	35	0.639558	1.888869
	VALIDATE	2	20.24902	4.576659	70.70938	40	0.913654	1.766808
	VALIDATE	2	19.34732	5.263158	75.97254	45	1.056233	1.688131

Yukarıdaki Çizelge 5.17, yapay sinir ağı modeli skor sıralamasının bir bölümünü göstermektedir. Hedef ordinal olduğunda, SAS Enterprise Miner, hedef değiskenin en üst seviyesinde olma ihtimaline dayanarak kaldıraç tablolarını oluşturur. Bu çalışmada hedef değişkenin en üst seviyesi 2. seviyedir. Hedef sıralıyken SAS Enterprise Miner, test veri setindeki her kayıt için, modelin öngörülen olasılığını hesaplar. Daha sonra veri setini tahmin edilenin azalan sırasına göre sıralar ve veri kümesini 20'ye böler. Her bölüm için kaldırma oranı, hasord = 2 olan vakaların yüzdelikteki oranının, tüm veri kümesindeki

kayıp = 2 olan vakaların oranına eşittir. Yukarıdaki Çizelge 5.17 bu çalışma için kaldırma oranlarının bir kısmını göstermektedir. Söz konusu çizelgeye göre eğitim örneğinde, modelden elde edilen tahminler azalan sırada düzenlendikten sonra %10'luk yüzdelik için bakacak olursak tahmin edilenlerin hasar frekansı 2 olanların oranı bütün veri setinde hasar frekansı 2 olanların oranından 2.49 kat fazladır ve kümülatif yakalama oranı %24.94'tür. Bu oran geçerlilik örneğinde ise %10'luk yüzdelik için tahmin edilenlerin hasar frekansı 2 olanların oranı bütün veri setinde hasar frekansı 2 olanların oranından 2.42 kat fazladır ve kümülatif yakalama oranı %24.25'tir.

Eğitim örneğinde, %30'luk yüzdelik için bakacak olursak tahmin edilenlerin hasar frekansı 2 olanların oranı bütün veri setinde hasar frekansı 2 olanların oranından 2.18 kat fazladır ve kümülatif yakalama oranı %65.44'tür. Bu oran geçerlilik örneğinde ise %30'luk yüzdelik için tahmin edilenlerin hasar frekansı 2 olanların oranı bütün veri setinde hasar frekansı 2 olanların oranından 2.09 kat fazladır ve kümülatif yakalama oranı %62.92 olarak hesaplanmıştır.



Şekil 5. 19 Yapay Sinir Ağı Kümülatif Kaldıraç Grafiği

Yukarıdaki Şekil 5.19, Çizelge 5.17'nin grafik şeklinde genel özeti olmakla birlikte eğitim(training) ve geçerlilik(validation) veri setleri için kümülatif lift (kaldıraç) oranını göstermektedir. Grafiğe göre eğitim örneğinde, modelden elde edilen tahminler azalan sırada düzenlendikten sonra %20'lik yüzdelik değerinde tahmin edilenlerin hasar frekansı 2 olanların oranı bütün veri setinde hasar frekansı 2 olanların oranından 2.58 kat fazladır. Geçerlilik örneğinde ise modelden elde edilen tahminler azalan sırada

düzenlendikten sonra %20'lik yüzdelik değerinde tahmin edilenlerin hasar frekansı 2 olanların oranı bütün veri setinde hasar frekansı 2 olanların oranından 2.4 kat fazladır.

Çizelge 5. 18 Yapay Sinir Ağı Uyum İstatistikleri

Uyum İstatistikleri				
Hedef = HASORD				
Uyum İstatistikleri		Eğitim	Geçerlilik	Test
DFT	Total Degrees of Freedom	15254.00	.	.
DFE	Degrees of Freedom for Error	15168.00	.	.
DFM	Model Degress of Freedom	86.00	.	.
_NW_	Number of Estimated Weights	86.00	.	.
AIC	Akaike's Information Criterion	9486.57	.	.
SBC	Schwarz's Bayesian Criterion	10142.97	.	.
ASE	Average Square Error	0.11	0.11	0.12
MAX	Max. Absolute Error	0.97	0.94	0.96
DIV	Divisor for ASE	22881.00	11439.00	3825.00
NOBS	Sum of Frequencies	7627.00	3813.00	1275.00
RASE	Root Average Squared Error	0.34	0.34	0.34
SSE	Sum of Squared Errors	2614.33	1313.87	447.64
SUMW	Sum of Case Weights Times Freq	22881.00	11439.00	3825.00
FPE	Final Prediction Error	0.12	.	.
MSE	Mean Squared Error	0.11	0.11	0.12
RFPE	Root Final Prediction Error	0.34	.	.
RMSE	Root Mean Squared Error	0.34	0.34	0.34
AVERR	Average Error Function	0.41	0.41	0.41
ERR	Error Function	9314.57	4682.82	1579.32
MISC	**Misclassification Rate**	**0.25**	**0.25**	**0.27**
PROF	Total Profit for HASORD	13422.00	6692.00	2228.00
APROF	Average Profit for HASORD	1.76	1.76	1.75

Yapay sinir ağları analizinin en büyük problemi, yorumlama açısından zor olmasıdır. Bahsedilen tüm sonuçları kapsayan ve modelin genel başarısı hakkında yorum yapılmasını sağlayan yanlış sınıflandırma oranı (Misclassification Rate) aşağıdaki tabloda gösterilmektedir.

Oluşturulan yapay sinir ağı modelinin sonucunda; Eğitim, Geçerlilik ve Test veri setlerinde SAS Enterprise Miner Programı tarafından hesaplanabilen uyum istatistikleri yukarıdaki Çizelge 5.18'de gösterilmiştir. Çıkan değerlere baktığımızda, geçerlilik veri

setinde modelin başarı oranını gösteren 'Misclassification Rate' sonucu Geçerlilik ve seti için 0.25 seviyesindedir. Bu sonuç, verilerin yanlış sınıflandırma yüzdesini göstermektedir. Verilerin doğru sınıflandırma oranı yani modelin başarısı geçerlilik veri seti için (100-25=75) %75, test veri seti için (100-27=73= %73 olarak çıkmıştır.

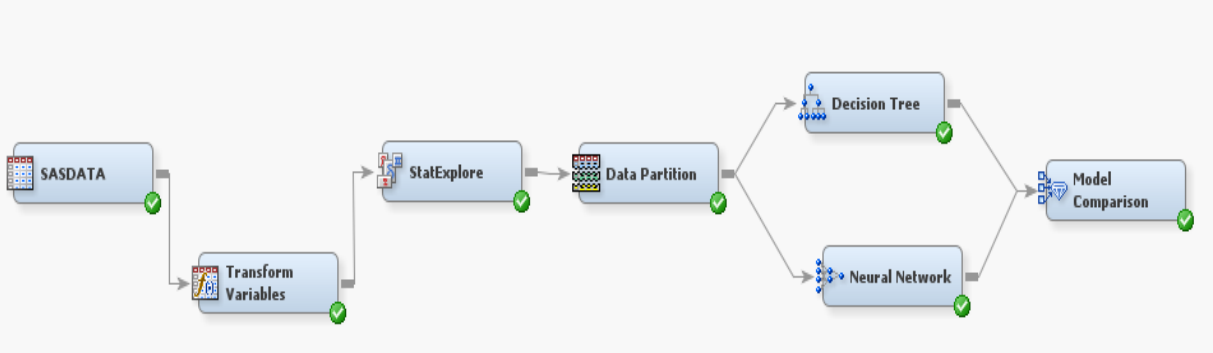
Model performansını değerlendirmek için, sigorta şirketi müşterilerini, hedef değişkenin gözlenen en üst seviyesi yerine HASORD'ın beklenen değerine göre sıralamalıdır. Hedefin beklenen değerine dayanan bir lift tablosunun nasıl oluşturulacağını göstermek için, aşağıdaki formülü kullanarak veri kümesindeki her kayıt için beklenen değeri oluşturulur.

- E (hasord|Xi)=Pr(hasord=0|Xi)*0+Pr(hasord=1|Xi)*1+Pr(hasord=2|Xi)*2

Daha sonra beklenen hasar frekansına göre müşterilere özel prim hesaplamalar vs. için bir yol izlenebilinir.

5.4 Karar Ağacı ve Yapay Sinir Ağı Modellerinin Karşılaştırılması

Karar ağaçları ve yapay sinir ağları modellerinin uygulama aşamasının ardından, bahsi geçen bu iki analiz aşağıdaki Şekil 6.20'de görülen adımlar gerçekleştirilerek karşılaştırılmıştır.



Şekil 6. 20 SAS Enterprise Miner Model Karşılaştırma Adımları

Bu iki analiz yönteminden başarı oranı yüksek çıkan yöntem Çizelge 6.19'da gözüktüğü üzere karar ağacı yöntemi olmuştur.

Çizelge 6. 19 Model Karşılaştırma Uyum İstatistikleri

Uyum İstatistikleri Model Karşılaştırma , Hedef : HASORD(_VAPROF_)					
Seçilen Model	Model Adları	Eğitim: Average Squared Error	**Eğitim: Misclassification Rate**	Geçerlilik: Average Squared Error	**Geçerlilik: Misclassification Rate**
Y	**Karar Ağaçları**	0.05391	**0.11407**	0.05512	**0.11487**
	Yapay Sinir Ağları	0.11426	**0.25200**	0.11486	**0.25413**

Karar ağacının yanlış sınıflandırma oranı eğitim ve geçerlilik veri setlerinde %11 olarak çıkmıştır. Bu durumda doğru sınıflandırma oranı yani modelin başarısı (100-11=89) %89 olarak karşımıza çıkmaktadır. Yapay sinir ağı modeli için yanlış sınıflandırma oranı eğitim ve geçerlilik veri setlerinde %25 olarak çıkmıştır. Verilerin doğru sınıflandırma oranı, yani modelin başarısı (100-25=75) %75 olarak çıkmıştır. İki model karşılaştırıldığında karar ağacı modelinin başarısı %89 ile daha yüksektir.

BÖLÜM 6

SONUÇ VE ÖNERİLER

Verinin önemi gün geçtikçe artmaktadır. Zamanla artan ve günümüzde veritabanlarında kolaylıkla depolanabilen verilerden anlamlı sonuçlar çıkarmak veri madenciliğinin uygulama alanlarını oluşturmaktadır. Artık her sektörde sıkça karşımıza çıkan veri madenciliği, işletmeler için ciddi bir öneme sahiptir. Bu yöntem ile ortaya çıkan sonuçlar işletmelerin aldıkları kararlara yön vermektedir. Şirketlerin üstlendiği faaliyetin uygulama alanlarını sağlıkla yürütebilmesi için, etkili yönetim sistemi gereklidir. Etkili yönetim sisteminin temel unsurlarından birisi, risk azaltma tedbirlerini uygulamaya koyma ihtiyacı üzerine karar vermeye yardımcı olmak için tehlikeyi belirlendikten sonra, tanımlanmış tehlikeden kaynaklanan risklerin olasılığını ve sonuçlarını belirlemek için risk analizi yapmaktır. Risk analizinde kullanılabilecek yöntemlerden biri de veri madenciliğidir. Türkiye'de sigortacılığın gelişerek veri kaynaklarının çoğalması ile bu sektörde de veri madenciliği uygulamalarına ilgi artmıştır. Sigorta yaptırmak için başvuruda bulunan yeni bir müşteriye karşı nasıl hareket edilmesi konusunda veri madenciliği, sigorta şirketleri için de yol gösterici olabilmektedir.

Sigorta, farklı uygulama alanlarında yapılabilir. Hayat ve hayat dışı olarak iki kategoriye ayrılır. Hayat sigortaları sigortalıların yaşam kalitelerini artırmak için yaptırılırken, hayat dışı sigortalar maddi hasarları karşılamaya yöneliktir. Hayat dışı sigorta türlerinden olan kasko sigortaları müşterilerin motorlu/motorsuz araçlarını güvenceye almak için yaptırdıkları bir sigorta türüdür.

Çalışmamızda kasko sigortası üzerinde, özel bir sigorta şirketinden elde edilen veriler ile veri madenciliği kullanılarak müşteriler için risk analizi yapılmıştır. Bağımsız değişken olarak göz önüne alınan değişkenler; geçmiş ortalama hasar maliyeti, sigortalının yaşadığı bölge, aracın yoğunluğu, aracın ağırlığı, aracın kaç kilometrede olduğu, sigortalının cinsiyeti, kişinin medeni durumu, sigortalının yaşı, sigortalının deneyim yılı ve aracın bedelidir. Ulaşılmak istenen hedef değişken ise hasar frekansıdır. Uygulamanın amacı, sigorta şirketine yeni kayıt yaptıran bir müşterinin poliçe bilgileri ve yaşadığı yerin demografik değişkenleri arasından seçilen bağımsız değişkenler kullanılarak hasar frekansını öngörmeye çalışmaktır.

Modelleme teknikleri olarak karar ağaçları ve yapay sinir ağları kullanılmıştır. İki model karşılaştırıldığında, yapay sinir ağları yönteminin müşterileri risk gruplarına atamadaki başarısı geçerlilik verisinde %75 ve test verisinde %73 iken karar ağacı modelinin başarısı geçerlilik verisinde %89 ve test verisinde %88 ile daha yüksek çıkmıştır. Genel olarak değerlendirildiğinde her iki yöntemin de risk değerlendirmede kullanışlı olduğu söylenebilmekle beraber karar ağaçları yöntemi ile daha başarılı öngörüler yapılabileceği sonucuna ulaşılmıştır. Uygulamada kullanılan geçmiş ortalama hasar maliyeti, aracın kaç kilometrede olduğu, sigortalının yaşı, sigortalının yaşadığı bölge ve sigortalının deneyim yılı değişkenleri karar ağacı oluşumunda en etkili değişkenler olarak bulunmuştur. Dolayısıyla bahsedilen bu değişkenler yardımı ile yeni gelen müşterinin öngörülen hasar frekansı hesaplanıp yüksek çıkması durumunda tahmini riske göre ücretlendirilecek primlerin belirlenmesi sağlanabilir. Yüksek ve düşük riske sahip sigortalılar için farklı politikalar belirlenebilir. Müşteriler öngörülen hasar frekanslarına göre sınıflandırılıp farklı risk guruplarına atanabilir. Türkiyede sigortacılık sistemi geliştikçe ele alınabilecek farklı değişkenler de ortaya çıkabilir. Yeni eklenecek değişkenlerle ve farklı nitelikteki veri setleri ile uygulama geliştirilip sigortacılık sektörüne ışık tutabilir.

KAYNAKLAR

[1] Mannila,H., (1996). “Data mining: machine learning, statistics and databases”, Conference: Scientific and Statistical Database Systems, 18-20 June 1996, Stockholm.

[2] Hackerbits, Data Mining History, https://hackerbits.com/data/history-of-data-mining/, 10 Kasım 2017.

[3] Muslu, D., (2009). Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, İTÜ Fen Bilimleri Enstitüsü, İstanbul.

[4] Devale,A.B. ve Kulkarni, Dr. V.R.,(2012). “Application of data mining techniques in life insurance” , International Journal of Data Mining & Knowledge Management Process (IJDKP), 2: 31-40.

[5] Sarma, K.S., (2013). Predictive Modeling with SAS® Enterprise Miner™ Practical Solutions for Business Applications, Second Edition, PhD Publisher: SAS Institute, North Carolina.

[6] Kriele, M. ve Wolf,J., (2014). Value-Oriented Risk Management of Insurance Companies, 1. Edition ,Publisher: Springer-Verlag London, New Jersey.

[7] Marton, P.D., (2015). Insurance and Financial Products to Mitigate Political and Credit Risk, Master Thesis, The University of Texas at Austin in Partial Fulfillment of the Requirements for the Degree of Master of Business Administration and Master of Global Policy Studies, Texas.

[8] Hui,W.C.C. ve Davalos, A.D.C., (2009). How Is Risk Assessment Performed In International Technology Projects, Master Thesis, Master in Strategic Project Management Umeå School of Business, Umeå.

[9] Gou,W., (2004). Development Of A Framework For Prelımınary Risk Analysis In Transportation Projects, Master Thesis, Worcester Polytechnıc Institute, Worcester.

[10] Bitaraf,S., (2011). Risk Assessment and Decision Support, Master Thesis, Chalmers University Of Technology Master of Science Thesis in the Master’s Programme , Sweden.

[11] Czirner,T., (2010). Risk management in new technology deployment projects,Master Thesis, Aalborg University M.Sc. In International Business Economics, Aalborg.

[12] Lookman Sithic,H. Ve Balasubramanian,T., (2013). "Survey of Insurance Fraud Detection Using Data Mining Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2:62-65.

[13] Shi, B., (2012). New Aspects of Product Risk Measurement and Management in the U.S. Life and Health Insurance Industries, Doctoral Dissertation, The University of Texas at Austin in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy, Austin.

[14] Öğüt,S.,Veri Madenciliği Kavramı ve Gelişim Süreci, http://www.sertacogut.com/blog/wpcontent/uploads/2009/03/sertac_ogut_veri_madenciligi_kavrami_ve_gelisim_sureci.pdf, 20 Ocak 2018.

[15] Qin,J. ve Norton M.J., (1999). "Knowledge Discovery in Bibliographic Databases", Library & Information Science Research, 22(4):433-435.

[16] Sirmakessis,S., (2005). Knowledge Mining: Proceedings of the Nemıs 2004 Final Conference, Springer-Verlag Berlin Heidelberg, 434, New York.

[17] Biçen,P., (2002). Veri Madenciliği: Sınıflandırma ve Tahmin Yöntemlerini Kullanarak Bir Uygulama,Yüksek Lisans Tezi, YTÜ Fen Bilimleri Enstitüsü, İstanbul.

[18] Farid,D., Sadeghi,H., Hajigol,E. ve Parirooy,N., (2016). "Classification of Bank Customers by Data Mining: a Case Study of Mellat Bank branches in Shiraz", International Journal of Management Accounting and Economics, 3:534-543.

[19] De Veaux, R., W. Hoerl, R. ve D. Snee,R., 2016. " Big Data and the Missing Links", Article in Statistical Analysis and Data Mining: The ASA Data Science Journal , 9(6): 383-460.

[20] Kaur,H. ve Wasan, S.K., (2006). "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science , 2(2): 194-200.

[21] Yan,H. ve Yan, L., (2016). "Analysis on a New Data Mining Algorithm of the Statistics Work", Journal TELKOMNIKA (Telecommunication Computing Electronics and Control), 14:42-46.

[22] Larosa, D.T., (2005). Discovering Knowledge In Data, 1st Edition, Published by John Wiley & Sons, New Jersey.

[23] Hand, D., Mannila H. ve Smyth, P., (2001), Principles of Data Mining, 1st Edition, A Bradford Book The MIT Press Cambridge, London.

[24] Chonweng,W. ve Scholten,D., (2013). "O2O E-Commerce Data Mining in Big Data Era", Journal-TELKOMNIKA, 14:396-402.

[25] Larose,D.T. ve Larose, C.D., (2015). Data Mining and Predictive Analytics, Second Edition,Wiley Series on Methods and Applications in Data Mining, New Jersey.

[26] Tsiptsis,K. ve Chorianopoulos,A., (2010). Data Mining Techniques in CRM Inside Customer Segmentation, 1st Edition, Publisher: John Wiley & Sons , Chichester.

[27] Zentut, Data Mining Processes, http://www.zentut.com/data-mining/data-mining-processes/, 2 Kasım 2017.

[28] Walsh,S., (2005). Applying Data Mining Techniques Using SAS® Enterprise Miner- Course Notes, SAS Institute Inc., North Carolina.

[29] Select Statistical Services, CHAID (Chi-square Automatic Interaction Detector), https://select-statistics.co.uk/blog/chaid-chi-square-automatic-interaction-detector/, 25 Temmuz 2017.

[30] Milanović,M. ve Stamenković,M., (2016). "Chaid decision tree: Methodological Frame and Application",Economic Themes, 54(4):571-572.

[31] Dai,W. ve Ji,W., (2014). "A MapReduce Implementation of C4.5 Decision Tree Algorithm",International Journal of Database Theory and Application, 7:49-60.

[32] He,Y.,Han,J. ve Zeng,S., Classification Algorithm based on Improved ID3 in Bank Loan Application, https://rd.springer.com/chapter/10.1007%2F978-1-4471-2386-6_148, 15 Ocak 2018.

[33] Hssina,B.,Merbouha,A.,Ezzikouri,H. ve Erritali,M., (2014) "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications (IJACSA), 13-19.

[34] Ville,B. Ve Neville,P., (2013). Decision Trees for Analytics Using SAS Enterprise Miner, 13, Publisher: SAS Institute, North Carolina.

[35] Aharkava,L., (2010). Artificial neural networks and self-organization for knowledge extraction, Master Thesis, Charles University in Prague Faculty of Mathematics and Physics , Prague.

[36] Ibiwoye,A.,Ajibola,O.E. ve Sogunro, A.B., (2012). "Artificial Neural Network Model for Predicting Insurance Insolvency", International Journal of Management and Business Research , 2 (1), 59- 68.

[37] Stanley,K.O., (2004). Efficient Evolution of Neural Networks through Complexification, Doctoral Dissertation ,The University of Texas at Austin in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy , Texas.

[38] İnanc, A., Yapay Sinir Ağları, https://www.slideshare.net/aybukeinanc/yapay, Temmuz 2017.

[39] Farhat,D., (2016). "The Performance Of Artıfıcıal Neural Networks And Tıer Structured Informatıon Transmıssıon In Identıfyıng Aggregate Consumptıon Patterns In New Zealand", Journal for Studies in Economics and Econometrics, 40:71-86.

[40] Radhamohan,R.S., (2010). Automatic Semiconductor Wafer Map Defect Signature Detection Using a Neural Network Classifier, Master Thesis,The

University Of Texas At Austin in Partial Fulfillment of the Requirements for the Degree of Master Of Scıence In Engıneerıng, Texas.

[41] Jun Chin, T., Multi-Layer Feed Forward Neural Networks, https://cs.adelaide.edu.au/~dsuter/Harbin_course/MultiPercept.pdf, 12 Kasım 2017.

[42] Hlavacek,B.M., (2014). Multilayer feedforward neural networks based on multi-valued neurons, Master Thesis, Masaryk University Faculty Of Informatics, Brno.

[43] Doç. Dr. Altunkaynak, B., (2017). Veri Madenciliği Yöntemleri ve R Uygulamaları, 1. Baskı, Seçkin Yayınları, Ankara .

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Merve ŞAHİN

Doğum Tarihi ve Yeri : 18.09.1992, Trabzon

Yabancı Dili : İngilizce

E-posta : mrvsahin34@gmail.com

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Lisans	İstatistik	Yıldız Teknik Üniversitesi	2015
Lise	Fen Bilimleri	Atatürk Anadolu Lisesi	2010

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2017- …	BDDK	İstatistikçi
2016-17	HDI Sigorta A.Ş.	Raporlama Uzman Yrd.
2014-15	Finnet Elektronik Yayıncılık Data İletişim San. Tic. Ltd. Şti	Destek Uzman Yrd.

YAYINLARI

Bildiri

1. Şahin, M. ve Gölbaşı Şimşek, G., (2017). “Veri Madenciliği Kullanılarak Sigortacılık Sektöründe Risk Değerlendirmesi”, III. International Balkan and Near Eastern Social Sciences Congress Series (IBANESS), 4-5 Mart 2017, Edirne.