T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

MOTOKARAVAN SİGORTACILIĞI TAHMİN MODELLEMESİ VE UYGULANAN YÖNTEMLERİN KARŞILAŞTIRILMASI

Yüksek Lisans Tezi

Tezi Hazırlayan : **Yasin KAYA**

İstanbul , 2017

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

MOTOKARAVAN SİGORTACILIĞI TAHMİN MODELLEMESİ VE UYGULANAN YÖNTEMLERİN KARŞILAŞTIRILMASI

Yüksek Lisans Tezi

Tezi Hazırlayan :

Yasin KAYA

Öğrenci No:

140820040

Tez Danışmanı :

Yrd. Doç. Dr. Atınç YILMAZ

İstanbul , 2017

YEMİN METNİ

Yüksek lisans tezi olarak sunduğum "Motokaravan Sigortacılığı Tahmin Modellemesi ve Uygulanan Yöntemlerin Karşılaştırılması" başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere uygun şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmamın içinde kullanıldıkları her yerde bunlara atıf yapıldığını belirtir ve onurumla doğrularım.11.05.2017

Yasin KAYA

T.C.
BEYKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ SAVUNMA SINAVI SONUÇ TUTANAĞI

Beykent Üniversitesi
Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Aşağıda tez adı belirtilen yüksek lisans öğrencisi 16082004 no'lu YASİN UNAYA'in 11/05/2017 tarihinde yapılan tez savunma sınavı[1] sonucunda 45 dakika süreyle sunduğu ve savunduğu tezi hakkında[2] oybirliğiyle KABUL kararı verilmiştir.

Bilgilerinize saygılarımızla arz ederiz.

Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ
Programı : BİLGİSAYAR MÜHENDİSLİĞİ
Tez Başlığı[3] : Metakarayen Sigortacılığı Tahmin Modellemesi ve Uygulanan Yöntemlerin Karşılaştırılması.

Tez Sınav Jürisi	Öğretim Üyesi	İmza
Danışman :	Yrd. Doç. Dr. Atınç Yılmaz	
Üye :	Doç. Dr. Gökhan Silahtaroğlu	
Üye :	Yrd. Doç. Dr. Ediz SAYLAM	

[1] Jüri üyeleri söz konusu tezin kendilerine teslim edildiği tarihten itibaren en geç bir ay içinde toplanarak öğrenciyi tez savunma sınavına alır. Belirlenen günde yapılamayan jüri toplantısı, katılanların hazırladığı bir tutanakla enstitü yönetimine bildirilir. Bu durumda jüri en geç onbeş gün içinde toplanarak adayı tez savunma sınavına alır. Tez savunma sınav süresi en az 45 dakikadır. Yüksek lisans tez savunma sınavı, tez çalışmasının sunulması ve bunu izleyen soru-yanıt bölümlerinden oluşur ve dinleyiciye açıktır. (Beykent Lisansüstü eğitim ve Öğretim Yönetmeliği-Madde30-3)

[2] Tez sınavının tamamlanmasından sonra jüri, tez hakkında "kabul", "düzeltme" veya "red" kararı verir. Jüri başkanı, jüri üyelerince imzalanmış sınav tutanağını, tez sınavını izleyen üç gün içinde ilgili enstitü yönetimine teslim eder. Tezi başarısız bulunan öğrencinin Enstitü ile ilişiği kesilir. Tezi hakkında düzeltme kararı verilen öğrenci en geç üç ay içinde gerekli düzeltmeleri yaparak ve yönetmelikte belirtilen usullere uygun olarak tezini aynı jüri önünde yeniden savunur. Bu savunma sınavında da tezi kabul edilmeyen öğrencinin enstitü ile ilişiği kesilir.(Beykent Lisansüstü eğitim ve Öğretim Yönetmeliği-Madde30-4)

[3] İleride doğabilecek aksaklıkların engellenmesi için tezin başlığının yazılması gerekmektedir.

MOTOKARAVAN SİGORTACILIĞI TAHMİN MODELLEMESİ VE UYGULANAN YÖNTEMLERİN KARŞILAŞTIRILMASI

Tezi Hazırlayan : **Yasin KAYA**

ÖZET

Sigortacılık, gelişmiş ve gelişmekte olan ülkelerin finansal yapılarına olumlu şekilde etki eden sektörlerdendir. Türkiye'de sigortacılık sektörünün son yıllarda çift haneli oranlarda büyüdüğü ve Gayri Safi Yurtiçi Hasıla'ya olan katkısının artmış olduğu görünmektedir. Bu durumun daha iyi seviyelere çıkarak devam etmesini sağlamak için teknolojinin ve makinelerin kabiliyetlerinden faydalanmak gerekmektedir. Veriler arasındaki gözle görünmeyen ilişkilerin tespitinde kullanılan veri madenciliği de sigortacılığın gelişmesi için yararlanılması gereken alanlardan birisidir. Veri madenciliği aktüeryal hesaplamalarda ve her türlü sigorta suistimallerinin tespitinde kullanılır. Günümüzde birçok veri madenciliği metot ve algoritması mevcuttur. Veri kümelerinin karakteristiğine göre hangisinin kullanılacağına karar verilebilir. Bu çalışmada, Avrupa'da popüler olan ve gerekli çalışmalar yapıldığında ülkemizde de gelişme potansiyeli gösterebilecek olan motokaravan sigortacılığı üzerine model geliştirilmiştir. Hangi özellikteki kişilerin bu sigortayı yaptırabileceğine yönelik tahminlemeler yapılmıştır. Modelleme çalışmasında karar ağaçları tekniğinin kullanıldığı J48 algoritması, lojistik regresyonun kullanıldığı algoritma ve yapay sinir ağlarının kullanıldığı MultilayerPerceptron algoritması koşulmuştur. MultilayerPerceptron algoritması en iyi doğruluk oranına sahip olmuştur. J48 ise MultilayerPerceptron algoritmasına yakın bir doğruluk oranına sahiptir. Lojistik regresyonda ise ciddi bir başarı elde edilememiştir. Elde edilen sonuçların motokaravan sigortası yaptırma potansiyeline sahip olan kişilere ulaşmaya yönelik çalışmaları kolaylaştıracağı düşünülmektedir.

Anahtar kelimeler: veri madenciliği, sigortacılık sektörü, motokaravan sigortası, karar ağaçları, lojistik regresyon, yapay sinir ağları.

PREDICTIVE MODELLING IN MOTOR CARAVAN INSURANCE AND COMPARISON OF METHODS APPLIED

Presented By **: Yasin KAYA**

ABSTRACT

Insurance is a sector that affects the financial structures of developed and developing countries positively. It appears that the insurance sector in Turkey has grown in double digits in recent years and the contribution to the Gross Domestic Product has increased. It is necessary to take advantage of the abilities of technology and machines to ensure that this situation continues at better levels. Data mining, which is used to identify invisible relationships between data, is one of the areas that need to be exploited for the development of insurance. Data mining is used in actuarial calculations and in the detection of all types of insurance fraud. Many data mining methods and algorithms are available today. Which one to use depends on the nature of the data sets. In this study, a model has been developed on motocaravan insurance, which is popular in Europe and can demonstrate the development potential in our country when necessary studies are carried out. Estimates have been made on which property persons can make this insurance. In the modeling study, the J48 algorithm that uses the decision tree technique, the algorithm in which the logistic regression is used, and the MultilayerPerceptron algorithm which uses artificial neural networks are run. The MultilayerPerceptron algorithm has the best accuracy. J48 is in second place. J48 has an accuracy ratio close to the MultilayerPerceptron algorithm. No serious success has been achieved in the logistic regression. It is thought that the results obtained will facilitate the studies to reach those who have the potential to have a motocaravan insurance.

Key Words: data mining, insurance sector, motocaravan insurance, decision trees, logistic regression, artificial neural networks.

TEŞEKKÜR

Yüksek lisans ders ve tez sürecimde yaptığı olumlu katkı ve yorumlarıyla beni yönlendiren hocalarım Yrd. Doç. Dr. Atınç YILMAZ'a ve Yrd. Doç. Dr. Ediz ŞAYKOL'a teşekkür ederim. Ayrıca, hayatımın her aşamasında olduğu gibi bilgisayar başında geçen tez sürecimde de desteğini esirgemeyen sevgili eşim Ayşe KAYA'ya teşekkür ederim.

İÇİNDEKİLER

Sayfa No.

TABLOLAR LİSTESİ

Sayfa No.

ŞEKİLLER LİSTESİ

Sayfa No.

KISALTMALAR

KDD : Knowledge Discovery on Databases

DIKW : Data Information Knowledge Visdom

VTBK : Veri Tabanlarında Bilgi Keşfi

ETL : Extract Transformation Load

OLTP : Online Transaction Processing

SQL : Structured Query Language

CRM : Customer Relationship Management

HRM : Human Resource Management

SVM : Support Vector Machines

KNN : K-Nearest Neighbor

1.GİRİŞ

İnsanlık, içinde bulunduğumuz milenyum çağı ile beraber çok büyük değişimleri ve gelişimleri çok hızlı yapacak seviyelere ulaşmıştır. Modern çağın başlangıcından itibaren devam eden kol gücünün azalması ve var olan işlerin makinelere yaptırılması süreci 1970'lerden itibaren daha farklı bir formata bürünmüştür. Bilgisayarlar ortaya çıkmış ve manuel olarak çok uzun zamanlarda yapılacak olan işler salise veya saniyeler boyutuna indirgenmiştir. Bilgisayar veya diğer bir ifadeyle makine ile ilgili gelişmeler eksponansiyel bir şekilde artmıştır. İnternetin de doğması ve dünyaya yayılmasıyla dev sunucular (server) kurulmuş ve daha önce kağıtlara yazılan verilerin disklerde depolanması sağlanmıştır. İteratif olarak ilerleyen gelişmeler yeni sektörlerin ortaya çıkmasını sağlamıştır. Depolanan verilerin birikmesi geçmişe dönük gelişen olaylardan geleceğe yönelik çıkarımlar yapılması fikrini doğurmuştur. Veri tabanında Bilgi Keşfi (VTBK), 1990'lı yıllardan itibaren büyük bir hızla ilerlemiştir. Günümüzde devasa boyutlara ulaşan veri tabanlarında tutulan veriler sayesinde, veri madenciliği çok önemli bir alan haline gelmiştir. Geleceğe dönük yatırım yapan her kurumsal şirket, özel veya kamu kurumu olması farketmeksizin, veri madenciliği konusunda yatırım yapmak zorundadır. Dijital dönüşümü ilke haline getiren inovatif şirketlerin ayakta kalabileceği bir çağa geçiş yapmış bulunmaktayız. Eldeki verilerden profesyonel anlamda yararlanmak ve veri madenciliğinin gücünü kullanmak da bu dönüşüme katkı sağlayacak unsurlardandır.

Veri madenciliği, veri bilimi, iş zekası gibi alanların hepsinin üst başılığı veri analitiğidir. Var olan verileri tutmak için başlı başına büyük bir maliyet gerekmesinin yanında, anlık olarak sunuculara gelen veri akışı ve kayıt altına almalar (log), tüm işlerin sistematik bir şekilde işlemesini gerekli kılmıştır. Veri madenciliği alanında önemli algoritmalar, yaklaşımlar, araçlar geliştirilmiştir. Veri madenciliğinin kullanıldığı yerlerden birisi de sigortacılık sektörüdür. Bu çalışmada da sigortacılık sektörüne yönelik modelleme çalışması yapılmaktadır. Ayrıca, veri madenciliğinde en çok kullanılan algoritmalardan olan karar ağaçları, lojistik regresyon ve yapay sinir ağları tekniklerinin kullanıldığı algoritmalar karşılaştırılmıştır. Üçünün de birbirine karşı avantajlı ve dezavantajlı olduğu durumlar vardır. Bu çalışmada bu konular üzerine yoğunlaşılmıştır.

Çalışmanın temel amacı ülkemizde yaygın olmasa da Avrupa'nın bazı ülkelerinde çok popüler olan motokaravan sigortasını hangi özellikteki kişilerin yaptırdığının tespit edilmesidir. Ayrıca kullanılan karar ağaçları, lojistik regresyon ve yapay sinir ağları yöntemlerinin sonuçlarını karşılaştırarak çıkarımlar yapmaktır. Önümüzdeki onlarca yılın popüler konularından biri olacağı şimdiden belli olan veri analitiği dünyasına ufak da olsa bir katkı sunmaya çalışılmıştır.

Çalışma yedi ana bölümden oluşmaktadır.

İlk olarak 3. bölümde Türkiye ve dünyada sigortacılık sektörünün durumu mercek altına alınmıştır. Sigortacılık sektöründe veri madenciğinin nasıl kullanıldığından bahsedilmiş ve motokaravan sigortacılığı hakkında bilgi verilmiştir.

4. bölümde veri madenciliği kavramı üzerinde durulmuştur. Verinin tanımı yapılmış, enformasyon, bilgi ve bilgelik kavramları açıklanmıştır. Veri madenciliğinin tarihçesine inilmiş ve KDD süreci uçtan uca ele alınmıştır. Daha sonra veri madenciliğinin tanımı yapılmış, veri madenciliğini oluşturan alanlardan ve çeşitli sektörlerde veri madenciliğinin nasıl ve hangi amaçlarla kullanıldığından bahsedilmiştir.

5. bölümde veri madenciliği süreçleri ele alınmıştır. Öncelikle model, teknik, yöntem ve algortima kavramları açıklanmıştır. Veri madenciliğinde kullanılan semma metodolojisine değinilmiş ve veri ön işlemesi hakkında bilgi verilmiştir. Ön işleme sürecinde veri tiplerinin neler olduğundan, istatistiki değerlerin nasıl yorumlanacağından, veri temizleme işleminin nasıl yapılacağından, boş ve gürültülü değer problemlerinin nasıl giderileceğinden, veri birleştirme, dönüştürme ve indirgeme işlemlerinin nasıl yapılacağından, korelasyon analizlerinin nasıl yapılacağından bahsedilmiştir. Daha sonra ise kurulan modelin değerlendirme kriterlerinin neler olduğundan ve modellerin karşılaştırmasının nasıl yapılması gerektiğinden bahsedilmiştir.

6. bölümde ise veri madenciği metotlarından bahsedilmiştir. Tahminleyici (predictive), tanımlayıcı (descriptive) ve buyrukçu (prescriptive) modelleme çeşitlerinden ve bunları oluşturan yöntem ve algoritmalar hakkında bilgi verilmiştir.

7. bölümde kullanılan teknolojiler ve araçlar hakkında bilgi verilmiştir. Modelleme çalışmasında aşama aşama olarak hangi araçların hangi özelliğinden faydalanıldığı şekillerle beraber aktarılmıştır.

8. bölümde veri kümesi üzerinde uygulamaya geçilmiştir. Öncelikle veri kümesi ve kullanılan teknolojiler tanıtılmıştır. Daha sonra veri üzerinde ön işleme çalışması yapılmış ve istatistiki değerler yorumlanmıştır. Korelasyon analizleri yapılmış ve modele girecek olan değişkenler seçilmiştir. Modelleme çalışması karar ağacı tekniğini kullanan J48, lojistik regresyon ve yapay sinir ağları tekniğini kullanan MultilayerPerceptron algoritmaları koşularak yapılmıştır. Yapılan çalışmanın sonucunda modeller yorumlanmış ve karşılaştırılarak en iyi olan model seçilmiştir.

9. bölümde ise çalışmada elde edilen sonuçlar yazılmıştır. Sigotacılık sektörüne katkı sağlayacağı düşünülen doneler sıralanmıştır.

2. BENZER ÇALIŞMALAR

Sigoracılık sektörüne ait veri madenciliği çalışmaları ve karar ağaçları, lojistik regresyon ve yapay sinir ağlarının karşılaştırılması konularıyla ilgili daha önceden yayımlanmış olan bazı tez ve makaleler aşağıda sıralanmıştır.

"Sigortacılık Sektöründe Müşteri İlişkileri Yönetimi Yaklaşımıyla Veri Madenciliği Teknikleri ve Bir Uygulama" [1] isimli çalışmada, sigortacılık sektöründe müşteri ilişkileri yönetimi incelenmiş ve birliktelik kuralları, sınıflandırma ve kümeleme gibi veri madenciliği yöntemleri kullanılmıştır. Müşterilerin sınıflandırılması ve davranış olasılıklarının tahmini gibi konularda çalışılmıştır. Ayrıca ürün ile müşteri ve şirket ile müşteri ilişkileri ortaya koyularak poliçe artışlarına yardımcı olacak yollar araştırılmıştır.

"Aktüeryal Modellemede Bulanık Destek Vektör Makineleri" [2] isimli çalışmada, aktüerya hesaplarında kullanılacak olan bir yöntem geliştirmek istenmiştir. Belirsizlik durumundaki sigorta prim hesaplaması, risk ölçümü gibi konuları içeren aktüerya için destek vektör makineleri (DVM) ve bulanık regresyon çözümlemesi kullanılmıştır. Sigorta problemlerinde güvenilir modellerin geliştirilmesi, sigorta şirketinin finansal istikrarı için çok önemlidir. Bu bağlamda, gizli yapıları tanımlayabilen güçlü bulanık modellere ulaşılması hedeflenmiştir.

"Müşteri İlişkileri Yönetimi İçin Veri Madenciliği Kullanılması ve Sigortacılık Sektörü Üzerine Bir Uygulama" [3] isimli çalışmada, Türkiye'de bulunan bir sigorta şirketinden elde edilen veriler üzerinde Apriori, K-Means ve Kohonen Ağları algoritmaları koşulmuştur. Daha sonra da müşteri ilişkileri yönetimine ilişkin bilgiler elde edilmiştir.

"Hava Kirliliği Tahmininde Çoklu Regresyon Analizi ve Yapay Sinir Ağları Yönteminin Karşılaştırılması" [4] isimli çalışmada, hava kirliliğinin yani havadaki SO2 konsantrasyonunun tahmin edilmesi için çoklu doğrusal regresyon, bulanık sinir ağları ve yapay sinir ağları teknikleri ile modeller kurulmuş ve ortaya çıkan sonuçlar karşılaştırılmıştır. Yaklaşımlara ait sonuçlar üretebilmek için kullanılan veriler; uygulama alanı olarak seçilen Sivas ili şehir merkezine ait 1990-2004 yıllarındaki hava kirliliği ve

meteorolojik verileridir. Üretilen sonuçlar içerisinde yapay sinir ağları modeli çok başarılı sonuç elde etmiştir. Bulanık sinir ağları tarafından üretilen model ikinci iyi sonucu üretmiştir.

"Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması" [5] isimli çalışmada, kredi talebinde bulunan kişilerin geri ödemelerinde problem yaşanıp yaşanmayacağı tahminlenmiştir. Böylece bankalar kredi taleplerindeki riskleri öngörebileceklerdir. Tahminlemede lojistik regresyon ve yapay sinir ağları kullanılmıştır. Çalışmada kullanılan veriler Türkiye'de bulunan bir bankadan sağlanmıştır. Bankanın müşterilerinden 1639 kişi rastgele bir şekilde seçilmiştir. Bağımlı değişken, müşterilerin kredisinin geri ödemesini düzenli gerçekleştiren ve düzenli gerçekleştirmeyen olarak iki sınıfa ayrılmıştır. Yapılan tahminleme sonuçlarına bakıldığında lojistik regresyon modelinde elde edilen doğruluk oranı % 65,1 ve yapay sinir ağları modelindeki ise % 70,3 olarak ölçülmüştür.

"Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression" [6] isimli çalışmada, öğrencilerin akademik performansı hakkında tahminleme yapılmıştır. 206 öğrencinin bilgileriyle modelleme çalışması yapılmış ve modelde öğrencilerin demografik verileri ve GPA'ları değişken olarak kullanılmıştır. SAS Enterprise Miner aracı kullanılarak yapılan modellemede yapay sinir ağları , karar ağaçları ve lineer regresyonun üçü ile de %80 in üzerinde doğru tahminlenmiş sonuçlar elde edilmiştir. En iyi sonucu yapay sinir ağları vermiştir.

"Yapay Sinir Ağları ve Lojistik Regresyon Yöntemleri İle Meme Kanseri Koltuk Altı Lenf Nodu Durumunun Belirlenmesi" [7] isimli çalışmada, meme kanseri hastalarının klinik ve patolojik verileri incelenerek yapay sinir ağları ile koltuk altı lenf nodlarının durumunun belirlenmesi amaçlanmıştır. Ankara Onkoloji Eğitim ve Araştırma Hastanesi ve Ankara Numune Eğitim ve Araştırma Hastanesi'nden alınan 270 meme kanseri hastasının patolojik ve klinik verileri üzerinde çalışılmıştır. Sınıflandırma yapmak için lojistik regresyon analizi ve ileri beslemeli geri yayılımlı çok katmanlı yapay sinir ağları kullanılmıştır. Oluşturulan modeller test verileri ile test edilmiştir. Lojistik regresyon ve seçilen yapay sinir ağları modelleri kıyaslandığında yapay sinir ağları değerleri daha

başarılı olmuştur. YSA'nın korelasyon katsayısı 0,872, duyarlılık değeri %88,8, belirlilik değeri %97,2, doğruluk değeri %94,4 olarak elde edilmiştir.

"Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması" [8] isimli çalışmada, sigortacılık sektörü için kurulmuş olan hasar ihbar veri tabanında bulunan verilerden suistimal (fraud) riski tahminleme çalışması yapılmıştır. Veri madenciliğinde yer alan sınıflandırma yönteminlerinden olan karar ağacının kullanıdığı bir algoritma uygulanmıştır. Bu doğrultuda, hedef değiken ve onu etkileyecek nitelikler belirlenmiştir. Modelleme çalışması yapılmış ve sonuçlar değerlendirilerek risk maddesi oluşturup oluşturmadığına dair yorumlar geliştirilmiştir.

"Yükseköğretimde Öğrenci Başarılarının Sınıflandırılmasında Yapay Sinir Ağları ve Lojistik Regresyon Yöntemlerinin Kullanılması" [9] isimli çalışmada, üniversite öğrencileri arasında yapılan ankette akademik başarıları etkileyen durumlar ele alınmıştır. Çalışmanın amacı anketteki verilere göre öğrenen bir sistem kurarak farklı programlara yeni kaydolan öğrencilerin gelecekte nasıl bir başarı elde edeceklerine dair tahmin yapılmasını sağlamaktır. Ayrıca kullanılan sınıflandırma teknik ve algoritmaların performansları karşılaştırılmıştır. Ankete Ankara Üniversitesi'nde toplam 419 üçüncü sınıf öğrencisi katılmıştır. Kullanılan lojistik regresyon modeli aracılığıyla başarılı öğrencilerin %51,1'ini doğru sınıflandırılmıştır. Başarısız olan öğrencilerin sınıflandılmasındaki doğruluk oranı ise %77 olarak ölçülmüştür. Çok katmanlı algılayıcı model olarak oluşturulan yapay sinir ağları aracılığıyla ise başarılı öğrencilerin %51,1'i doğru sınıflandırılırken, başarısız olan öğrencilerin %83,95'ini doğru sınıflandırılmıştır.

"Lojistik Regresyon Analizi (LRA), Yapay Sinir Ağları (YSA) Ve Sınıflandırma Ve Regresyon Ağaçları (C&Rt) Yöntemlerinin Karşılaştırılması Ve Tıp Alanında Bir Uygulama" [10] isimli çalışmada, Tokat Gaziosmanpaşa Üniversitesi Tıp Fakültesi hastanesi veri tabanından çekilen Üroloji Polikliniği hastalarına ait veri kümesi üzerinde karşılaştırması yapılmıştır. Bu veri kümesi oluşturulurken kesin olarak prostat kanseri olduğu tanısı konmuş olan hastalar yaklaşık beş yıllık bir veri yığını içinden süzülmüştür. Sonuç olarak kullanılabilecek 118 adet kesin prostat kanseri tanısı olan hasta bulunmuştur. Bu hastaların, prostat kanseri teşhisinde kullanılan yaş, genetik yatkınlık, rektal tuşe kontrolü ve PSA değerleri tespit edilmiştir. Ayrıca yukarıdaki parametrelere sahip 118

adet prostat kanseri tanısı olmayan hasta tespit edilerek toplamda 236 adet hastaya ait bir veri kümesi oluşturulmuştur. Çalışmada hedef değişken olarak prostat kanseri tanısı alınmış, bağımsız değişkenler ise yaş, genetik yatkınlık, rektal tuşe kontrolü ve PSA değeri olarak tespit edilmiştir. Yöntemlerin sınıflandırma başarıları açısından veri kümesi üzerinde en iyi sınıflandırmayı yapay sinir ağları (YSA) gerçekleştirmiştir. %87,29 doğruluk ve 0,929 AUC ile yapay sinir ağları algoritmasına girilen kayıtlar diğer yöntemlere göre daha doğru şekilde sınıflandırılmıştır. İkinci sırayı ise %83,90 doğruluk ve 0,924 AUC ile lojistik regresyon analizi yöntemi, üçüncü sırayı ise %81,78 doğruluk ve 0,828 AUC ile C&RT karar ağacı algoritması almıştır. Diğer taraftan yöntemlerin Kappa istatistiği yönünden karşılaştırılmasında YSA 0,746 ile ilk sırada, LRA 0,678 ile ikinci ve C&RT 0,636 ile üçüncü sıradadır. Sonuç olarak doğruluk oranlarına göre sıralandığında en iyi modelin yapay sinir ağları olduğu görülmüştür. Yapay sinir ağlarından sonra lojistik regresyon analizinin başarısı gelmektedir. Başarı sırasında son sırada ise C&RT olduğu görülmüştür.

"Comparison Of The Decision Tree, Artificial Neural Network and Linear Regression Methods Based On The Number and Types Of İndependent Variables and Sample Size" [11] isimli çalışmada, örneklem büyüklüğü ,bağımsız değişkenlerin sayısı , tipleri ve sınıflarının değişmesinin veri madenciliği ve istatistik tekniklerinin performansını nasıl etkiledikleri karşılaştırmalı olarak incelenmiştir. Örneklem büyüklükleri olarak 100, 500, 1000 ve 10000 elemanlı kümeler kullanılmıştır. Değişken sayısı ve örneklem büyüklüğünün hesaba katılmadığı durumda, sürekli bağımsız değişkenler için lineer regresyonu kullanmanın karar ağaçları ve yapay sinir ağlarını kullanmaya göre daha iyi sonuçlar ortaya koyduğu gözlemlenmiştir. Sürekli ve kategorik bağımsız değişkenler için, kategorik değişken sayısının bir olduğu durumda lineer regresyon; kategorik değişken sayısının iki veya daha fazla olduğu durumlarda ise yapay sinir ağları en iyi sonuçları vermiştir. Kategorik değişken sayısının artması, yapay sinir ağlarının performansını da arttırmıştır.

"Finansal Veri Madenciliği" [12] isimli çalışmada, çeşitli veri madenciliği teknikleri kullanılarak, bir bankaya ait veri tabanından elde edilen kredi kartı ve kredilerin statüsü verileri üzerinde müşteri profili ve müşteri segmentasyonu uygulamaları

gerçekleştirilmiştir. Banka verisi, kredi kartı tipi verisi ve kredi statüleri verisi olmak üzere ikiye ayrılarak incelenmiştir. Karar ağaçlarından J48 algoritması, yapay sinir ağlarından ÇKA ve regresyon analizinden LR algoritması veri kümesi üzerinde kullanılan sınıflandırma teknikleri ve algoritmalarıdır. Kredi kartı tipi verisine uygulanan sınıflandırma algoritmaları, başarı oranları olarak birbirine son derece yakın sonuçlar vermişlerdir. Bu algoritmalardan J48 ve ÇKA algoritması, verinin sınıflandırılmasında aynı başarı oranını yakalamıştır. Fakat modelin anlamlılığının bir ölçütü olan Kappa statistik değerinin J48 algoritmasında daha iyi sonuç vermesinden dolayı modellemede J48 algoritmasının kullanılması uygun görülmüştür. Kredi statüsü verisine uygulanan sınıflandırma algoritmaları içerisinde en iyi başarı oranı, az bir farkla LR algoritması tarafından gerçekleştirilmiştir. Kredi kartı tipi veri kümesi üzerinde uygulanan Simple K-Means algoritması sonucunda müşterilerin baskın özelliklerine göre 3 farklı müşteri grubu elde edilmiştir. Bu müşteri grupları incelendiğinde; 1. gruba ait müşterilerin hesaplarından yaptıkları otomatik ödeme işlemlerinin diğer gruptaki müşterilerden daha fazla olduğu; 2.grupta yer alan müşterilerin en genç müşteri grubunu oluşturduğu; 3.grupta yer alan müşterilerin ise yaptıkları otomatik ödemelerin ve hesapta kalan para miktarlarının en az olduğu müşteri grubunu oluşturduğu görülmüştür.

"Yapay Sinir Ağları ve Çoklu Regresyon Analizinin Karşılaştırılması" [13] isimli çalışmada, doğrusal regresyon (linear regression) analizi ile yapay sinir ağlarının tahminleme başarıları karşılaştırılmıştır. Çalışmada kullanılan veri kümesi için yapılan tahminlemelerde yapay sinir ağları, en küçük kareler yöntemi ve Robust regresyon tekniklerinden M-kestiricilerinin başarıları karşılaştırılmıştır. Sonuç olarak yapay sinir ağlarının diğerlerinden daha iyi sonuçlar ürettiği görülmüştür.

"K - En Yakın Komşuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması" [14] isimli çalışmada, Bülent Ecevit Üniversitesi'ne ait olan hastaneden elde edilen veriler kullanılmıştır. Kadın Hastalıkları ve Doğum Polikliniği'nde gerçekleşen erken ve zamanında doğumlara ait gebe verileri üzerinde sınıflandırma tekniklerinden faydalanılmıştır. Kullanılan karar ağacı, k-en yakın komşuluk (KNN) ve yapay sinir ağlarına (YSA) ait başarılar karşılaştırılmıştır. Çalışmanın sonucunda YSA için %90,8 oranında, KNN için %78,3 oranında ve karar

ağacı tekniğinin kullanıldığı algoritma için ise %82,5 oranında doğruluk oranı elde edilmiştir.

"Churn Analysis and Prediction With Decision Tree and Artificial Neural Network" [15] isimli çalışmada, yapay sinir ağları ve karar ağaçları kullanılarak müşteri kayıp analizi gerçekleştirilmiştir. Analiz sonucunda yapay sinir ağlarının daha iyi sonuçlar verdiği görülmüştür. Yapay sinir ağlarının doğruluk oranı %97, karar ağaçlarınınki ise %81 olarak ölçülmüştür.

"Yapay Sinir Ağları, Lojistik Regresyon ve Karar Ağaçları Uygulamaları İle Kariyer Başarısı Tahmini Akademisyenler Üzerine Bir Araştırma" [16] isimli çalışmada, Holland'a ait olan tipoloji kuramından yararlanılmıştır. Çalışmanın amacı, kariyer tercihleri ile kariyer başarısı arasındaki ilişkiyi tespit etmektir. Akademisyenlerin mevcut mesleklerinin kişilik tipi ile mesleki ilgi ölçeğine göre ortaya çıkan mesleki ilgi profili kıyaslanmıştır. Çalışma sonucunda, Türkiye'deki akademisyenlerin büyük bir çoğunluğunun Holland'ın tipoloji kuramı çerçevesinde oluşturulan mesleki ilgi profiliyle uyumlu olmayan bir ilgi alanı seçtiği görülmüştür. Çalışma kapsamında bir takım istatistiksel analizler yapılarak değişkenler arasındaki ilişkiler ortaya konulmuştur. Bununla birlikte akademisyenlerin kariyer başarılarına ait verilerin tahmin etme çalışmaları yapılmıştır. Tahminleme yapmak için karar ağaçları, lojistik regresyon ve yapay sinir ağları olmak üzere üç farklı metot kullanılmıştır. 28 farklı veri kümesi kullanılarak bu yöntemlerin farklı veri kümelerine verdikleri tepkiler karşılaştırılmıştır. En iyi başarı sonuçları karar ağaçları yöntemi ile elde edilmiştir.

"Kredi Kartı Değerlendirme Tekniklerinin Karşılaştırılması" [17] isimli çalışmada ise kredi değerlendirmesi için kullanılan istatistiksel tekniklerin etkinliği karşılaştırmıştır. Lojistik regresyon analizi, diskriminant analizi, regresyon ağacı ve yapay sinir ağları tekniklerinden faydanılan çalışmada başarıyı ölçmek için birinci tip hata, ikinci tip hata ve doğruluk oranı değerlerine bakılmıştır. Birinci tip hata ve doğruluk oranına göre bakıldığında regresyon ağacı, ikinci tip hataya göre bakıldığında ise yapay sinir ağları en iyi sonucu vermiştir.

Bu çalışmada yukarıda hakkında bilgi verilen çalışmaların bazılarıyla sigortacılık sektöründe çalışılmasından dolayı, bazılarıyla da model kurma aşamasında yapay sinir

ağları, karar ağaçları ve lojistik regresyon kullanılmasından dolayı benzerlikler vardır. Çalışmanın sigortacılık üzerine yapılan çalışmalardan farkı ise motokaravan sigortacılığı üzerine yoğunlaşılmış olmasıdır. Hangi özelliklere sahip olan müşterilerin motokaravan sigortası yaptıracağına ilişkin bir tahminleme yapılmıştır. Kullanılan modellerin aynı olduğu çalışmalarda ise verisi kullanılan sektör ve alanlar farklılık göstermektedir. Çalışma sonucunda en iyi doğruluk oranı, yapay sinir ağları tekniğini kullanan algoritmanın uygulanmasıyla elde edilmiştir.

3. SİGORTACILIK SEKTÖRÜ

Sigorta (insurance); kanun ve mukavele ile tespiti yapılan çerçevesi belirlenmiş bir riskten benzer derecede tehdit alan belli bir sayının üstündeki grupların, tesadüf eseri ortaya çıkan, para birimiyle ölçülebilen hasar olayından kaynaklanacak olan zararın beraberce üstesinden gelinmesi amacıyla bir araya getirilmesiyle meydana gelen organizasyona denir [43]. Sigortacılık sayesinde rizikolarla kollektif bir birliktelik yoluyla baş edilebilir. Aynı rizikonun paylaşılması söz konusu olduğu için sigorta bir risk transfer mekanizmasıdır.

Sigortacılığın tarihi 4000 yıl öncesine dayanmaktadır. O yıllarda çok önemli bir ticaret merkezi durumun olan Babil'de, kervan sahibi tüccarların kendilerinden borç aldığı sermaye sahipleri, kervan soygunu gibi bir durum yaşandığında tüccarların kendilerine ödeyecekleri borçları siliyorlardı. Bunun yanında borcu alırken, taşımış oldukları riske karşılık ana para üzerinden de belli bir miktar para alıyorlardı. Kral Hammurabi bu uygulamayı yasalaştırdı. Hammurabi Kanunlarında, saldırıya maruz kalan kervanların uğradığı zararların tüm kervanlarca paylaşılması ilkesi vardır ve bu paylaşma kara taşımacılığındaki ilk örnektir. M.Ö. 600'lü yıllarda ise Hindu'lar arasında sigorta özelliği taşımakta olan kredi anlaşmaları yapılmaya başlanmıştı. Prim esasına dayalı sigortacılığın ilk örneklerine M.S. 1250'li yıllarda İtalya'nın Venedik, Floransa ve Cenova şehirlerinde rastlanmıştır. Tarihteki ilk sigorta poliçesi 23 Ekim 1347 tarihinde, Cenova Limanı'ndan Mallorca'ya gidecek olan "Santa Clara" adlı geminin yükü için düzenlendi. Ayrıca ilk sigorta şirketi 1424'te, Cenova'da kuruldu. Sigorta için yazılmış ilk kanuni mevzuat ise 1435'te yayımlanmış olan Barselona Fermanı'dır. 1700'lü yıllar ve sonrasında özellikle istatistik biliminin de gelişmesiyle sigortacılık faaliyetleri hız kazanmıştır. 20. yüzyılın başlarından itibaren sigorta şirketleri tam kurumsal bir şekilde faaliyetlerini hızlandırmış ve finansal açıdan ülkelerin çok önemli kurumları haline gelmişlerdir [44].

Sigorta kavramının varlığından söz edebilmek için şu unsurların bulunması gerekmektedir:

- Riskin (riziko) var olması
- Tehlikenin aynı ya da benzer olması

- Ortaya çıkan kaybın karşılanması
- Prim ödenmiş olması

Trafik, kasko, yeşilkart, hayat ve sağlık sigortaları başta olmak üzere çok fazla çeşitte sigorta branşı vardır. Yeni risk tanımlarının oluşması ve farkındalığın artması gibi sebeplerle gün geçtikçe branş sayısı artmaktadır.

Sigorta şirketleri, belli bir dönem için (genellikle 1 yıl) sigorta hizmeti almak isteyenlerle belirli bir prim karşılığında anlaşır. Ödenen prime göre teminat üst limiti belirlenir ve o dönem içerisinde hasarın veya istenmeyen durumun gerçekleşmesi durumunda masraflar sigorta şirketi tarafından karşılanır. Dönem bitmeden sigortalı, sigortacıyla olan anlaşmasını iptal edebilir. İptal durumunda sigorta hizmetinin alınmadığı zaman aralığı ölçüsünde sigorta ettirene prim iadesi gerçekleştirilir. Sigorta ettiren, sigortacıya prim ödemesi yapan özel veya tüzel kişidir. Sigortalı ise sigorta hizmetinden yararlanan kişidir. Sigorta ettirenle sigortalı aynı kişi olmayabilir. Örneğin, hayat ve sağlık sigortaları şirketler tarafından toplu yapıldığı için sigorta ettirenle sigortalı genellikle farklıdır. Sigorta şirketleri, acente veya aracılarla anlaşarak sigorta poliçesi düzenleyebilmektedir. Acenteler şirketlerin temsilcisi olarak müşterilerle sigorta anlaşması yapabilirler.

Sigorta şirketlerinin yaptığı sigortaların da sigortalanması gerekebilir. Bu işlemin yapılmasına reasürans (re-insurance) denir. Özellikle teminat üst limiti çok yüksek olan sigorta poliçeleri için olası bir riziko durumunda hasarın karşılanabilmesi için sigortacının reasürans yaptırması çok önemlidir.

3.1. Türkiye'de Sigortacılık

Türkiye'deki sigortacılık faaliyetleri 1870'li yıllarda başlamıştır. İngiliz sigorta şirketleri 1872'de Türkiye'de temsilcilik açmışlardır. 1893 yılında açılan Osmanlı Umum Sigorta Şirketi ilk yerli sigorta şirketi olmuştur. Sigorta şirketleri 1939 yılında Ticaret Bakanlığı'na, 1987 yılında ise Hazine ve Dış Ticaret Müsteşarlığı'na bağlanmıştır.

Branş bazında bakıldığında, 1999 depremlerinin ardından "Doğal Afet Sigortaları Kurumu" (DASK) kurulmuştur. Bireysel emeklilik sistemi (BES) 27 Ekim 2003 tarihinde aktif olmuştur. 14 Haziran 2005 tarihinde çıkan kanun kapsamında tarım sigortalarının verilerinin tutulması için Sigorta Havuzu (TARSİM) kurulmuştur. 16 Aralık 2003 tarihinde ise Trafik Sigortası Bilgi Merkezi (TRAMER) kurulmuştur. Trafik poliçesi üreten bütün sigorta şirketlerinin 2003 yılı başından itibaren tüm poliçe, hasar ve ödeme kayıtları TRAMER sistemine transfer edilmiştir. Günümüzde de hala devam eden bir uygulama olarak yeni üretilen poliçe ve hasar kayıtlarının transferi günlük olarak sisteme aktarılmaktadır. Diğer sigorta branşları için de benzer uygulama devreye alınmıştır. Sağlık Sigortası Bilgi Merkezi (SAGMER), Hayat Sigortası Bilgi Merkezi (HAYMER) ve Sigorta Hasar Takip Merkezi (HATMER) kurulmuş ve tüm bu kurumlar 9 Ağustos 2008 tarihinde faaliyetlerine başlayan Sigorta Bilgi Merkezi (SBM) altında toplanmışlardır [44].

T.C. Başbakanlık Hazine Müsteşarlığı verilerine göre, 2015 yılı sonu itibariyle ülkemizde toplam 59 sigorta ve 1 reasürans şirketi faaliyette bulunmaktadır. Bu şirketlerin 4'ü hayat, 19'i hayat/emeklilik, 36'sı hayat dışı şirkettir. Türkiye'de sigortacılık sektörü 2015 yılında toplam 31,1 milyar TL prim üretimi gerçekleştirmiştir. Toplam üretimin 27,3 milyar TL'lik kısmı (%88) hayat dışı sigortalarda, 3,8 milyar TL'lik kısmı (%12) ise hayat sigortalarında gerçekleştirilmiştir. Bir önceki yıla göre prim üretimi %19,5 oranında artmış ve 1998 yılı sabit fiyatlarıyla %11,2 oranında büyümüştür. Hayat grubu prim üretimi %6,7 oranında, hayat dışı prim üretimi ise %11,8 oranında reel artış sağlamıştır. Tanzim edilen poliçeler karşılığında yıl içinde sigortalılara toplam 86,1 trilyon TL teminat sağlanmıştır ve bu ülkemizin Gayri Safi Yurtiçi Hasıla (GSYH)'sının 44 katına eşittir [45].

3.2. Dünyada Sigortacılık

Sigortacılık, önemli bir kurumsal tasarruf organı olarak görülmektedir ve gelişmiş ülkelerde daha etkindir. Sektörün gelişim düzeyini ölçmek için kullanılan en önemli uluslararası göstergeler arasında kişi başına düşen prim miktarı, ülkelerin toplam primleri ve yıllık toplam primin Gayri Safi Milli Hasıla'ya oranı gösterilmektedir.

Swiss Reinsurance Company (Swiss Re) adlı şirketin Sigma Raporu'nda yayınladığı verilere göre, 2015 yılında dünyada sigortacılık sektöründe reel prim üretimi %3,8 artmıştır. Tablo 1, Tablo 2 ve Tablo 3'te gösterileren veriler ışığında hayat sigortalarında %4, hayat dışı sigortalarda ise %3,6 oranında artış olduğu söylenebilir. Gelişmiş ülkelerin üretimindeki büyüme %2,5 (hayat %2,5, hayat dışı %2,6) olurken, gelişmekte olan pazarlardaki büyüme %9,8 (hayat %12, hayat dışı %7,8) olarak gerçekleşmiştir. Toplam prim ise 4,554 trilyon Amerikan doları (USD) olmuş ve bir önceki yıla göre %3,5 artmıştır. Kıta bazında prim üretiminde 1,589 trilyon Amerikan doları ile Amerika ilk sırayı alırken, ikinci sırada 1,468 trilyon Amerikan doları ile Avrupa ve üçüncü sırada 1,350 trilyon Amerikan doları ile Asya vardır.

Tablo.1. Kıt'a ve Birlikler Bazında 2014 ve 2015 Yılları Toplam Sigorta Prim Üretimi

	Prim Miktarı (milyon ABD Doları Cinsinden)		**Düzeltilmiş Enflasyon Oranındaki Yüzde Değişim**		**Dünya Pazarındaki Yüzde Oranı**	**GSYH İçindeki Prim Oranı Yüzdesi**	**Kişi Başına Düşen Prim Miktarı**
	2015	**2014**	**2015**	**2014**	**2015**	**2015**	**2015**
Amerika	1 589 385	1 576 073	3,6	0,7	34,9	6,42	1 610
Avrupa	1 468 878	1 695 091	1,2	3,4	32,26	6,89	1 634,4
Asya	1 350 974	1 313 874	8,2	6,1	29,67	5,34	311,7
Afrika	64 123	70 116	2,4	3,9	1,41	2,9	54,7
Okyanusya	80 426	99 557	-4,5	15,1	1,77	5,58	2 065
Dünya	4 553 785	4 754 710	3,8	3,5	100	6,23	621,2
Gelişmiş Pazarlar	3 704 063	3 926 402	2,5	2,6	81,34	8,12	3 439,6
Gelişmekte Olan Pazarlar	849 723	828 308	9,8	7,6	18,66	2,92	135
OECD Ülkeleri	3 602 190	3 837 557	2,4	2,4	79,1	7,6	2 717,7
G7 Ülkeleri	2 809 967	2 931 527	2,7	1,5	61,71	8,05	3 637,1
AB Ülkeleri	1 352 516	1 559 618	1,4	3,6	29,7	7,57	2 411,9
NAFTA Ülkeleri	1 456 464	1 425 266	3,5	0,1	31,98	7,01	3 006,7
ASEAN Ülkeleri	87 921	88 354	8,1	6,8	1,93	3,35	128,2

Amerika'nın birinci olmasını sağlayan hayat dışı sigortada yüksek prim üretmesidir. Asya kıtası hayat sigortasında ilk sıradadır. Japonya ve Güney Kore gibi gelişmiş ülkeler buna katkı sağlamış.

Türkiye'nin prim üretiminin dünyadaki toplam üretim içindeki oranı %0.25'tir. Avrupa'daki üretim içindeki oranı ise %0.76'dır.

Tablo.2. Kıt'a ve Birlikler Bazında 2014 ve 2015 Yılları Hayat Sigortaları Prim Üretimi

	Prim Miktarı (milyon ABD Doları Cinsinden)		Düzeltilmiş Enflasyon Oranındaki Yüzde Değişim		Dünya Pazarındaki Yüzde Oranı	GSYH İçindeki Prim Oranı Yüzdesi	Kişi Başına Düşen Prim Miktarı
	2015	**2014**	**2015**	**2014**	**2015**	**2015**	**2015**
Amerika	668 037	660 808	4,2	-0,5	26,36	2,7	676,7
Avrupa	872 115	1 002 559	1,2	5,7	34,42	4,16	987,2
Asya	904 569	886 462	7,8	5,1	35,7	3,59	209,8
Afrika	43 704	47 605	2,8	5,1	1,72	1,97	37,3
Okyanusya	45 393	58 159	-7,8	27,5	1,79	3,15	1 165,5
Dünya	2 533 818	2 655 593	4	4,3	100	3,47	345,7
Gelişmiş Pazarlar	2 089 765	2 232 193	2,5	3,8	82,47	4,61	1 953,7
Gelişmekte Olan Pazarlar	444 052	423 399	11,7	6,8	17,53	1,52	70,6
OECD Ülkeleri	1 979 925	2 130 920	2,2	3,5	78,14	4,19	1 500
G7 Ülkeleri	1 531 861	1 614 697	2,7	2,1	60,46	4,46	2 014,5
AB Ülkeleri	820 286	944 837	1,3	5,9	32,37	4,68	1 492,1
NAFTA Ülkeleri	613 299	598 688	3,8	-0,9	24,2	2,95	1 266,1
ASEAN Ülkeleri	57 172	57 269	8,4	8,7	2,26	2,36	90,3

Tablo.3. Kıt'a ve Birlikler Bazında 2014 ve 2015 Yılları Hayat Dışı Sigortaların Prim Üretimi

	Prim Miktarı (milyon ABD Doları Cinsinden)		Düzeltilmiş Enflasyon Oranındaki Yüzde Değişim		Dünya Pazarındaki Yüzde Oranı	GSYH İçindeki Prim Oranı Yüzdesi	Kişi Başına Düşen Prim Miktarı
	2015	**2014**	**2015**	**2014**	**2015**	**2015**	**2015**
Amerika	921 347	915 266	3,1	1,6	45,61	3,72	933,3
Avrupa	596 763	692 533	1,1	0,3	29,54	2,73	647,2
Asya	446 405	427 411	9,2	8,3	22,1	1,74	102
Afrika	20 419	22 511	1,3	1,2	1,01	0,92	17,4
Okyanusya	35 033	41 398	0,1	1,1	1,73	2,43	899,5
Dünya	2 019 967	2 099 118	3,6	2,4	100	2,77	275,6
Gelişmiş Pazarlar	1 614 298	1 694 209	2,6	1,1	79,92	3,51	1 485,9
Gelişmekte Olan Pazarlar	405 670	404 909	7,8	8,6	20,08	1,29	64,4
OECD Ülkeleri	1 622 265	1 706 637	2,6	1	80,31	3,4	1 217,8
G7 Ülkeleri	1 278 105	1 316 831	2,6	0,8	63,27	3,59	1622,6
AB Ülkeleri	532 230	614 780	1,5	0,4	26,35	2,89	919,8
NAFTA Ülkeleri	843 165	826 578	3,3	0,9	41,74	4,06	1 740,6
ASEAN Ülkeleri	30 749	31 085	7,6	3,5	1,52	0,99	37,9

3.3. Karavan Sigortacılığı

Karavan (caravan) veya diğer adıyla mobil ev (mobile home), gezmeyi ve belli aralıklarla kamp yapmayı seven insanların başvurduğu araçlardan biridir. Motokaravanlar için trafik sigortası yaptırmak zorunludur. Ülkemizdeki karavan ve motokaravan sayısının, resmi olarak bilinmemekle beraber birkaç bin adet civarında olduğu tahmin edilmektedir. Avrupa'da ise karavan kültürü daha gelişmiş bir seviyededir. Avrupa Karavan Federasyonu (European Caravan Federation) verilerine göre Avrupa'da 2015 yılı sonu itibariyle yaklaşık 4 milyon karavan ve 1.5 milyondan fazla motorlu karavan bulunmaktadır.

Tablo.4. Avrupa'daki Motokaravan Sayısı

Ülke	2015
Avusturya	23 800
Belçika	45 000
Danimarka	16 000
Finlandiya	52 600
Fransa	436 100
Almanya	460 000
Büyük Britanya	205 000
İtalya	206 500
Hollanda	85 000
Norveç	5 700
Portekiz	10 000
İspanya	29 500
İsveç	79 300
İsviçre	34 000
Slovenya	4 850
Diğer	10 100
Toplam	1 703 450

Tablo 4'te bulunan verilere bakıldığında, en fazla motokaravanın Almanya ve Fransa'da bulunduğu görünmektedir. Avrupa'da motokaravanlar için yapılan sigortalarda kazadan kaynaklı hasar, çalıntı, yangın, temel geçici tamiratlar, çekici masrafları, ön cam kırılması gibi durumlar için teminat verilmektedir.

3.4. Sigortacılıkta Veri Madenciliği

Diğer sektörlerde olduğu gibi sigortacılık sektöründe de veri madenciliği kullanılmaktadır. Yeni poliçe talebinde bulunacak müşterilerin tahmin edilmesi, suistimallerin (fraud) tespit edilmesi, riskli müşteri guruplarının belirlenmesi, branşa göre prim tesbiti gibi konularda kullanılabilir. Veri madenciliği uygulamalarından faydalanmak ekonomik bakımdan da olumlu sonuçlar doğurmaktadır. Prim miktarlarının belirlenmesi sigorta yaptırma isteğini etkilemektedir. Bu yüzden optimize edilmiş bir fiyat belirlenmelidir. Suistimalleri engellemek, hem sigorta şirketlerini fazla maliyetten kurtarır

hem de vatandaşın daha az prim ödemesine yardımcı olur. Riskli kişilerin tespit edilmesi sayesinde suistimallerin ve dolayısıyla hukuki süreçlerin önüne geçilmiş olur.

4. VERİ MADENCİLİĞİ KAVRAMI

Veri madenciliğinin tanımına geçmeden önce veri ve bilgi kavramlarından, KDD sürecinden ve kısaca veri ambarından bahsetmek yararlı olacaktır.

4.1. Veri Kavramı

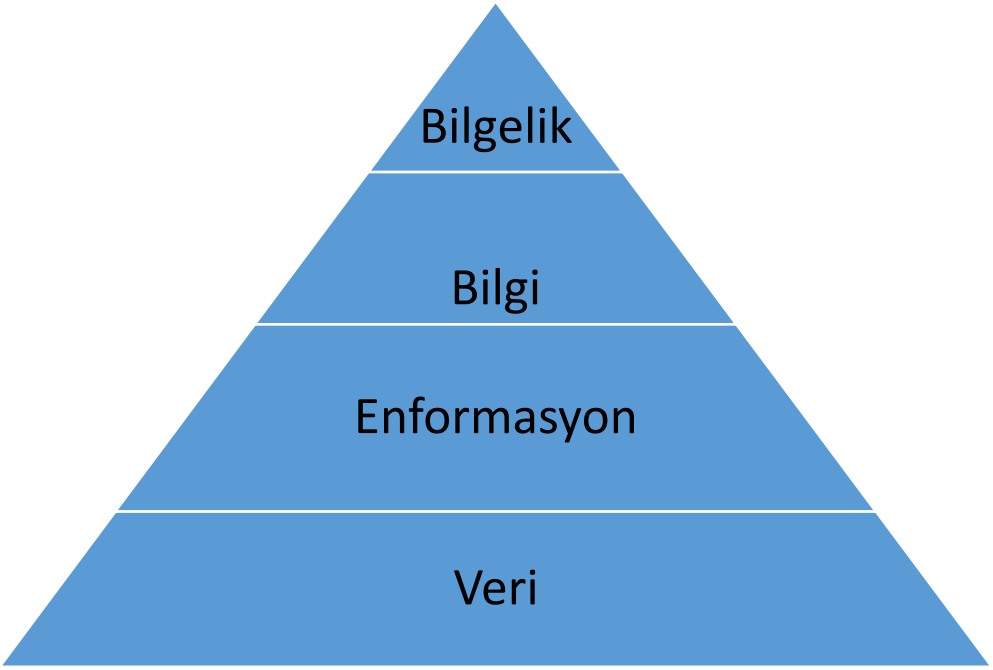
Veri kavramının evrilme süreci ele alırken Şekil 1'de gösterildiği gibi dört aşamadan geçmektedir.

a. Veri

b. Enformasyon

c. Bilgi

d. Bilgelik



Şekil.1. DIKW (Bilgi Piramidi, Knowledge Pyramid)

4.1.1. Veri

Veri kelimesini kelime anlamı itibariyle Latince'deki "datum" kelimesi karşılamaktadır. Veri, bilişim terimi olarak bakıldığında var olan ancak işlenmemiş yani ham halde bulunan kayıt olarak adlandırılır. Bu kayıtlar ilişkisi ortaya konmamış ve düzenlenmemiştir. Dolayısıyla anlam kazanmamıştır. Ancak bu durumun istisnaları vardır. İşlenerek üst katmana geçmiş olan (enformasyon haline gelmiş) bir veri, daha sonra farklı bir amaç için tekrar veri kullanılabilmektedir [18].

Kavramsal olarak bakıldığında veri, kaydedilmiş olan her türlü durum, fikir ve olaydır. Bu bakış açısıyla değerlendirildiğinde çevremizde bulunan her nesneye bir veri olarak bakılabilir. Semboller, sinyaller, sayılar, kelimeler veya işaretler birer veridir. Veriler, metrik değerlerle ölçülebilir. Örneğin, izlediğimiz videoların zaman uzunlukları, herhangi iki şehir arasındaki km cinsinden uzaklık, bir şehrin bir yıl boyunca gün gün ortaya çıkan °C cinsinden hava sıcaklıkları da veridirler. Verinin içerisinde yorum yoktur. Mekanik olmayan şeyler de veri kapsamına girebilir. Hisler, bilinçaltı, görüş, fikir vs de veriye dönüştürülebilir. İstatistik bilimcilerin sık kullandığı anket uygulamalarında bunlara rastlanmaktadır. Veri ham olduğu için otomatik süreçler yardımıyla işlenebilir ve yorumlanabilir.

4.1.2. Enformasyon

İkinci safhada ise enformasyon vardır. Verilerin düzenlenmiş, ilişkilendirilmiş, anlam kazanmış, işlenmiş haline enformasyon denir. Enformasyon, içerisinde bilgi barındırma potansiyeli yüksek olan bir seviyedir. Enformasyon, dış dünya ile olan ilişkilerde, belirsizlik (uncertainty) düzeyinde azalmaya sebep olan her türlü uyarana verilen addır. Kısacası format kazanmış, yapılandırılmış ve anlam katılmış veriler bütünü şeklinde düşünülebilir [18].

Enformasyon ile verileri organize bir hale getirmiş oluruz. Kelimelerin cümleleri, sayılar ve sembollerin denklemleri oluşturması gibidir. Soru sorma ve cevap verme yapılabilir. Enformasyon seviyesinde kim, ne, nerede gibi sorular sorulabilir. Ancak nasıl sorusu sorulamaz.

4.1.3. Bilgi

Bilgi, bu süreçteki üçüncü aşamadır. Sözlükteki anlamı "insan aklının erebileceği olgu, gerçek ve ilkeler bütünü, malumat" tır. Enformasyonun bilgiye dönüşmesi, onu algılamak, özümsemek ve sonuç çıkarmakla gerçekleşir. Bilgi felsefi olarak ise insanların maddi ve toplumsal anlıksal etkinliğinin ürünüdür [18].

Bilgi, zamanla farklı kaynaklardan sentezlenen enformasyonun teoriler, aksiyomlar veya yapılara oturtulmasıdır. Tecrübe, değerler ve öngörüden etkilenir. Bilgi kişiseldir (subjektif). Nesnel hale gelirse enformasyon seviyesine indirgenmiş olur.

4.1.4. Bilgelik

Bilgelik, yukarıdaki üç kavramın üstünde bulunur. Bilgi topluluklarının bir araya getirilip bir senteze dönüştürülmesiyle ortaya çıkar. Tecrübe, yetenek, öngörü gibi bireysel özellikler bilgelik elemanları arasındadır. Neyin bilindiğinin ve en iyinin ne olduğunun dikkate alınarak en uygun davranışın sergilenmesidir [18].

Bilgelik, farkında olmak, paradigmaya sahip olmak, etraflıca hakimiyet ve "neden" sorusuna cevap vermektir. Neden sorusuna cevapla birlikte olayların oluşuna ve varlıklara saygı ve kabul ediş beraber gelir ve beraberinde bütünün anlaşılması vardır.

Bilgelik, doğru ve yanlış, etik ve etik dışı kavramlarını içerir, kendi değer yargılarını ve var oluşunu tanımlar.

Yukarıdaki dört seviyeye bilgisayarda yapılan veri çalışmaları bakış açısıyla baktığımızda, veriye tablolarda hiçbir değişime uğramadan ilk defa doldurulan alanlar, enformasyona çeşitli manipülasyonlara uğratılarak elde edilen malumat, bilgiye ise yapay zeka ve veri madenciliği gibi süreçlerden geçirilerek elde edilen bir olgu gözüyle bakılabilir. Bilgisayarlar için bilgelik seviyesi şu an için söz konusu değildir. İnsan gibi düşünebilen ve kendine has bir etik dünyası olacak makinelerin varlılığı öteden beri tartışılsa da en azından yakın gelecekte gerçekleşmesi zor görünmektedir.

4.2. Veri Madenciliği Tarihçesi

Veri madenciliği, bilişim dünyasındaki gelişim sürecinde olağan bir sonuç olarak ortaya çıkmış ve bugünkü noktalara ulaşmıştır. İnsanlık tarihinin ilk yıllarından günümüze kadar veri her zaman yorumlanmış, biriktirilmiş ve veriden anlamlı bilgiler elde edilmek istenmiştir. Bilgisayarın icadıyla birlikte verinin tutulması ve değerlendirilme şeklinde format değişikleri olmuştur.

1950'li yıllarda icat edilen ilk bilgisayarlar aracılığıyla sayma ve diğer basit işlemler yapılmaya başlanmıştır.

İlk olarak 1960'lı yılların ortasından sonra, bilgisayarların veri analizine dair problemleri çözmek için kullanılmaya başlamasıyla veri madenciliği uğraşı ortaya çıkmıştır. O dönemde, bilgisayarların yardımı ile, yeteri kadar uzun yapılan bir taramanın ardından istenen verilere ulaşılabileceği gerçeği kabullenilmiştir. Yapılan bu işleme veri taraması, veri yakalaması gibi adlar verilmiştir [19].

1970'li yıllarda ilişkisel veri tabanı kavramı ortaya çıkmıştır. E.F. Codd tarafından ilişkisel veri tabanları makalesi yayımlanmıştır. Aynı yıllarda basit makine öğrenimi

çalışmaları gerçekleştirilmiştir. P. Chen'in çalıştığı varlık bağıntı modelleri de o yıllarda elde edilen kazanımlardan olmuştur.

1980'li yıllarda SQL (Structured Query Language) standart bir dil olarak kabul edilmiştir. Veri tabanı yönetim sistemleri (VTYS) yaygınlaşmıştır. Özel sektörde, kamuda ve bilimsel çalışmalarda kullanılması yaygınlaşmıştır. O yıllarda kurumsal şirketler, ürünlerini, müşterilerini ve rakiplerini mercek altına alabileceği ve CRM projeleri için temel oluşturabilecekleri veri tabanlarını oluşturmuşlardır.

1989'da Gregory Piatetsky-Shapiro tarafından düzenlenen bir organizasyonla ilk KDD (Knowledge Discovery in Databases) çalıştayı yapılmıştır. (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991 yılında KDD ile ilgili temel kavramların anlatıldığı "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesinin yayınlanması ile süreç daha da hızlanmıştır. 1992'de veri madenciliğinin uygulanması için yazılım çalışması gerçekleştirilmiştir. Böylece veri analizi, sadece geleneksel istatistiksel yöntemlerle manuel işlemler aracılığıyla yapılmak yerine, algoritmik bilgisayar modülleri tarafından yapılmaya başlanmıştır. 1995 yılında ise KDD çalıştayı, ACM SIGKDD Conference on Knowledge Discovery and Data Mining adı verilerek düzenli hale getirilmiştir. Bundan sonraki süreçte bilim insanları veri madenciliğine temelinde istatistik, makine öğrenmesi (machine learning), veri tabanı (database) ve otomasyon gibi disiplinlerin ve kavramların yattığı çeşitli yaklaşımlar getirmeye başlamışlardır [20].

İstatistik, verilerin yorumlanması, değerlendirilmesi ve analizleri konusunda yardımcı olan bir yöntemler topluluğudur ve bilgisayarın aktif kullanımından önce veri ile ilgili çalışmalar onun vasıtasıyla yapılmıştır. Makinelerin veri analizinde kullanılmaya başlamasıyla birlikte istatistiksel çalışmalar da hız kazanmıştır. Bilgisayarlarla yapılabilen işlemler sayesinde daha önce yapılması çok zor olan istatistiki araştırmalar da kolaylaşmıştır. Geliştirilen bilgisayar araçlarının (tool) arka planında istatistiksel hesaplamaların algoritmaları yatmaktadır. Verilerin, büyük yığınlar içerisinden sistemli bir şekilde çıkarılması ve yapılan analizler sonucunda bilgi elde edilmesi sürecinde veri madenciliği ve istatistik alanlarının iyi bir çalışma birlikteliği içine girdikleri görülmüştür.

Veri madenciliği, istatistiğin yanında veri tabanı ve makine öğrenimi (machine learning) alanlarıyla paralel bir şekilde ilerlemiştir. Yapay zeka alanının altında yer alan makine öğrenimi, diğer adıyla yapay öğrenme, bilgisayarların var olan veride bulunan deseni öğrenerek yeni veriler üzerinde uygulamasıdır. 1980'li yıllara kadar makineler, insanın öğrenmesine benzeyen bir yapıda oluşturulmaya çalışılmıştır. Ancak 1980'lerden sonra bu yaklaşım tarzı değişmiş ve makineler belirli alanlarda algoritmalar üretmek için inşa edilmiştir. Bütün bu gelişmeler istatistik ve yapay öğrenme kavramlarını, veri madenciliği ile entegre hale getirmiştir [19].

21. yüzyılın başından itibaren küreselleşmenin etkisiyle rekabetin artması ve buna bağlı olarak şirketlerin veri madenciliğini bir adım öne çıkmak için yeni yollar aramasına katkı sağlayacak önemli unsurlardan biri olarak görmesinin etkisiyle bu alandaki çalışmalar daha bir hız ve önem kazanmıştır. Bilgisayar sayısının ciddi bir şekilde artması, internetin yayılması, kurumsal ve hatta şahsi anlamda tüm verilerin kağıtlardan veri tabanlarına aktarılması, mobil dünyanın gelişmesi, sosyal medyanın ortaya çıkması sonucu anlık olarak akan verilerin devasa boyutlara ulaşması da veri madenciliği aracılığla bilgi üretimini hızlandıran sebepler arasındadır. Sadece 2012 yılı içerisinde internet ortamında ve sosyal medyada ortaya çıkan yazılar, yazının hayat kazanmasından sonra geçen yaklaşık 6000 yıllık süre zarfında yazılmış olan tüm içerikten fazladır [21]. 2012 yılından sonraki yıllarda artarak devam eden bir sanal ortam yazılığı içeriği vardır. Tahminlere göre 2020 yılında, 2012'de üretilen içeriğin 50 katından daha fazlası üretilecektir [22]. Bu kadar büyük miktarlarda üretilen verinin boşa gitmesi düşünülemez.

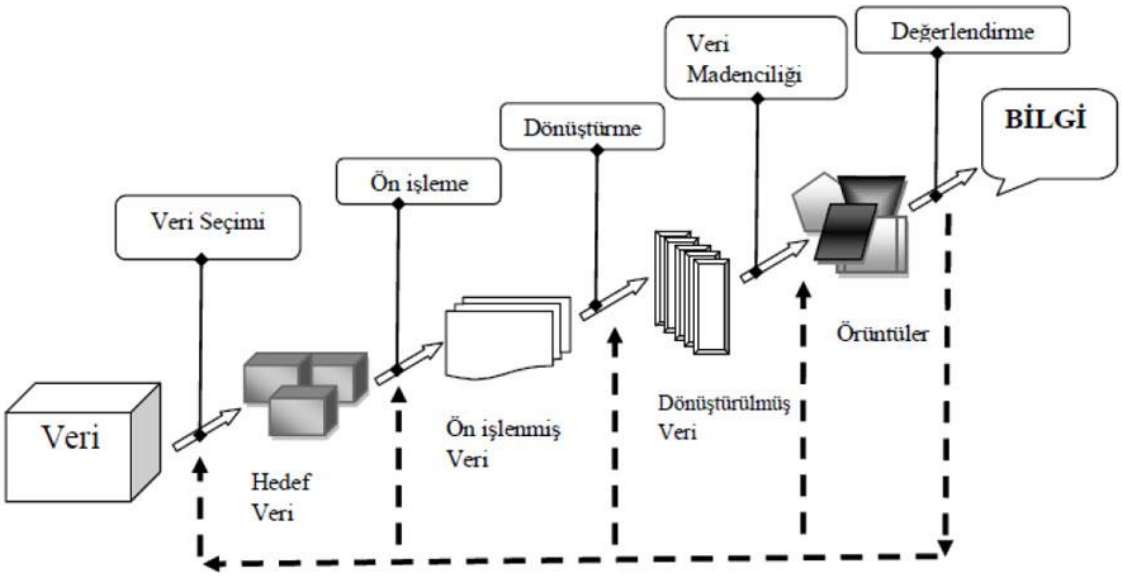
Bu açıdan bakıldığında, geleneksel veri madenciliği de format değiştirmektedir. Akan çok büyük miktardaki verinin anlık olarak değerlendirilip anlık modeller oluşturacağı yapılar oluşturulmaya başlanmıştır. Büyük Veri (Big Data) çalışmaları sayesinde yüksek hacim ve çeşitlilikteki verinin kısa sürede işlenmesi ve sosyal mecraların analiz edilmesi mümkündür.

Görüldüğü üzere veri madenciliğinin dinamik bir yapısı vardır ve gelişmeler eksponansiyel bir hızla ilerlemektedir. Sürekli yeni algoritmalar ve metotlar ortaya çıkmakta ve buna uygun alt yapıyı sağlayan teknoloji de çok hızlı ilerlemektedir. Önümüzdeki yıllarda veri madenciliğinde aktif rol oynamak isteyen kişilerin ve

kurumların, bu alt yapıyı sağlayan teknolojileri üreten tarafta olmasa bile en azından çağa uygun şekilde takip edebilmesi, inovatif bir gelecek için çok hayati bir öneme sahiptir.

4.3. KDD Süreci

Veri Tabanlarında Bilgi Keşfi (VTBK), veri madenciliğini de içine alan bir süreçtir. Verinin ham halinden başlayarak bilgi (knowledge) elde edilmesine kadar geçeceği aşamaları gösterir. VTBK metodolojisi ile, verinin nasıl depolanacağına, veriye erişileceğine, algoritmaların çok büyük seviyedeki veri kümelerine nasıl ölçeklenebileceğine, kullanışlı modellerin nasıl üretilebileceğine, sonuçların nasıl yorumlanabileceğine odaklanılır. 1989 yılında yapılan çalıştayla ortaya çıkmış ve verinin değerlendirilmesi üzerine çok önemli sonuçlar vermiştir. Çalıştay günümüzde de hala devam etmektedir.



Şekil.2. VTBK (KDD) Süreci

Şekil 2'de görüldüğü gibi VTBK işleyişine bakıldığında veri (data) ile başladığı görülmektedir. Bilgiye giden yolda veri başlangıç noktasıdır. Bu yoğun ve kompleks yapının içinden hedef veri seçilir. Hangi amaçla yola çıkıldığı baştan belli olmak zorundadır. Daha sonrasında seçilen veriler bir ön işlemeye tabi tutulur. Çünkü ham verinin içinde formattan boş değere kadar istenmeyen veya modelin kurulmasında olumsuz etkisi olacak olan değerler olabilir. Ön işlemeden sonra veri üzerinde gerekli olan transformasyonlar gerçekleştirilir. İstediğimiz enformasyonları, uygun bağlantılarla (SQL'deki join kodunu kullanarak) yeni tablolar oluşturarak elde ederiz. Buraya kadar olan kısımda modele girebilecek varlıkları elde etmiş oluruz. Artık veri madenciliği algoritmalarının ve yöntemlerinin zamanı gelmiştir. Uygulamalar ve yorumlamalar sonucunda bilgi elde edilmiş ve hangi amaçla kullanılacaksa ona hazır hale gelmiştir. Son aşamada başarısızlık ve başa dönme gibi durumlar yaşanabilir. Böyle bir durumla karşılaşmamak için her adımı dikkatli bir şekilde yürütmek çok önemlidir.

Veri ambarları bu süreçte önemli rol oynamaktadır. Veri ambarları, veri madenciliği sürecinin gerçekleştirilmesi için gerekli olan veriyi sağlayan özel bir veri tabanıdır. İçerisinde birçok farklı kaynaktan gelen veri bulunmaktadır. Sadece güncel değil, tarihsel olarak da veriyi barındırır. Farklı kaynaklardan gelen verileri istenen şekilde kompoze etmeyi sağlar. Bu birleşmeyi veriyi çekme (E-Excract), dönüştürme (T-Transformation) ve yükleme (L-Load) işlemleriyle (ETL) gerçekleştirir. Ayrıca OLTP yapısından ayrı bir şekilde ele alındığı için hizmet sunulan sistemde performans problemi yaşanmamasını sağlar [23].

4.4. Veri Madenciliği Tanımı

Veri madenciliği, büyük miktardaki veri koleksiyonu arasından geleceğe dair tahminleme yapılmasını sağlayan ilişki ve kuralların bilgisayar algoritmaları kullanılarak aranmasıdır. Yakın geleceğin, çok marjinal şeyler olmadığı sürece yakın geçmişten çok farklı olmayacağını düşünürsek geçmişteki verilerden çıkarılmış olan kuralların gelecek için fikir vereceği açıktır [24]. Madencilik (mining) kelimesi, meslek olarak kullanılan

manayla benzerlik gösterir. Maden işinde yapılan şey yerin altından değerli ve kullanışlı ürünleri yeryüzünde kullanılmak üzere çıkarmak eylemidir. Veri madenciliğinde veriyi toprak olarak düşünebiliriz. Çok büyük kütle ve hacme sahiptir. Bu kütlenin içinde yer alan değerli bilgileri yüzeye çıkarmak, o bilgi sayesinde yeni keşifler yapmak ve yeni politikalar belirlemek işi de aslında tam manasıyla madenciliktir. Doğru ve kullanışlı bilgiyi elde etmek için doğru yeri kazmak, doğru derinliğe inmek ve inmeyi sağlayacak değişkenleri doğru seçmek çok önemlidir.

Veri madenciliğinde önceden belli olmayan desenleri (pattern) yakalamak esastır. Öngörülemeyen ilişkiler keşfetmek başarıdır. Zaten var olan etkileri bulmak, veri madenciliğinin konusu değildir. Örneğin, arama motorunda bir kelime yazıp onunla alakalı bilgilere ulaşmak, bir iş yerindeki çalışanların yaş gruplarına göre maaş dağılımının histogramını çıkarmak, veri tabanında bir SQL sorgusu yazarak veri çekmek veri madenciliği yapmak değildir. Ancak benzer hobilere sahip insan gruplarının özelliklerini bulma, yüksek gerilim hattının yanında ikamet etmenin kanser riskini arttırma eğilimi olup olmadığı, kredi vermek isteyen bir bankanın hangi müşteriye ne kadar kredibilite sağlayacağının belirlenmesi, bir GSM şirketinin hangi müşteriye hangi kampayayı sunarsa en optimal sonuca ulaşağının belirlenmesi, bir süpremarketin yeni bir ürün için potansiyel müşteri kitlesi belirlemesi gibi konular veri madenciliğinin konusudur.

Veri madenciliğinin türevi olarak görülen ve kaynakları itibariyle özelleşen yeni alanlar bulunmaktadır. Bunlar metin madenciliği, web madenciliği, sosyal ağ madenciliği ve resim madenciliğidir. Metin madenciliğiyle benzer özelliklere sahip metinler gruplanabilir. Bu yapılırken dilin karakteristik özellikleri kullanılır. Sosyal ağlarda yapılan madencilikle çok fazla çıktı elde edilebilir. Örneğin twitterda yapılan bir çalışmayla insanların ruh haletiyle televizyonda izlenen programların içeriğinin benzerlik gösterip göstermediği tahmin edilebilir. Grup ve arkadaşlık analizleri yapılabilir. Resim madenciliği ile resimden cinsiyet ve yaşı söyleyebilecek modeller geliştirilebilir. Veri madenciliğinin çıktıları iş zekası uygulamalarında, raporlamalarda ve görselleştirme uygulamalarında kullanılabilir.

4.5. Veri Madenciliğini Oluşturan Alanlar

Şekil 3'te gösterildiği gibi veri madenciliği sürecinde farklı alanlardan ve disiplinlerden yararlanmak gerekmektedir.

Şekil.3. Veri Madenciliğinin Diğer Disiplinlerle Olan İlişkisi

Veri madenciliği tamamen KDD sürecinin içerisinde ve veri tabanı, istatistik, örüntü tanıma(pattern recognition), sinir ağlarının ve makine öğrenmesinin kesişiminde yer alır. Makine öğrenmesi, yapay zekanın altında yer alan bir disiplindir. Yani her makine öğrenmesi çalışması aynı zamanda yapay zeka (artificial intelligence) çalışmasıdır. Ancak tersi doğru değildir. Veri tabanı, içerisinde barındırdığı veriler sayesinde KDD içerisindeki aşamalara ve veri madenciliğine katkı sağlar. Ancak tamamen KDD içerisinde yer almaz. Çünkü bilgi keşfi yapılabilmesinin yanında müşteriye anlık erişim sağlanması, bakım gibi sadece veri tabanını ilgilendiren konular da vardır. İstatistik, veri madenciliğinde kullanılan önemli disiplinlerdendir. İstatistiksel çalışmalarda teori bazlı ilerlenir ve bir varsayımın doğruluğu araştırılır. İstatistiksel yöntemlerin kullanılması için yapısal (structural) veri tabanı gerekmektedir. Bilgisayarın gelişmesiyle beraber makine öğrenmesi ve yapay sinir ağları için geliştirilen algoritmalar sayesinde veri

madenciliğinde çok önemli mesafeler katedilmiştir. Makine öğrenmesinde sezgisel yaklaşım ön plana çıkmaktadır ve öğrenme işleminin başarısı arttırılmaya çalışılır. Veri madenciliği teori bazlı ve sezgisel yaklaşımları birleştirmiştir. Yeni teknolojilerin gelişmesiyle çok büyük miktarda veri kütlesi pratik ve doğru bir şekilde modellenebilecektir.

4.6. Veri Madenciliğinin Kullanıldığı Sektörler

Veri madenciliği, bir işletmenin veri ambarlarında yer alan verilerdeki gizli eğilimleri, bu eğilimlerin nedenlerini, birbirleriyle olan ilişkilerini, verilerde görülen desenlerin tarzlarını görünür kılan yöntemler topluluğu, teknikler olarak görülebilir [25]. Veri madenciliği , veriden kullanışlı örüntüler ve öngörüler elde etmek suretiyle karar verme sürecini geliştirmek için kullanılır. Veriyi farklı perspektiflerden inceleyen, kullanılabilir desenler ve keşfedilen ilişkileri özetleyen analitik bir sürecin adıdır. Birçok sektörde, kendine özgü dünyasının problemlerini çözmek için kullanılır.

Finansal alanda veri madenciliği, müşteri davranışları ve kredi kartı harcamaları, borsa ve diğer finansal araçların analizi, kara para aklama algılama, hedeflenmiş pazarlama (müşteriye özel kampanya, sms, internette neyi araştırdıysa ona göre bir kampanya vs), XRM (CRM + HRM), müşteri sadakati (Customer Churn analysis), kredi taleplerine verilecek yanıtın belirlenmesi, kredi kartlarındaki suistimallerin (fraud) tespiti ve harcama alışkanlıklarına göre müşteri gruplarının belirlenmesi konularında çözümler sunar.

Perakende sektöründe veri madenciliği, çok boyutlu raporlar (müşteri, ürün, zaman, şube vs), kampanya oluşturma ve kampanyanın başarı analizi, müşteri ilişkileri yönetimi (CRM), ürün tavsiyeleri, satış tahmini, müşteri değerlendirme, mevcut müşterilerin kaybedilmemesi ve yeni müşterilerin kazanılmasına dair stratejilerin belirlenmesi, müşterilerin satın almadaki kriterlerinin belirlenmesi, demografik özelliklerin belirlenmesi, pazar sepeti analizi ve raf analizlerinde kullanılır.

Telekom alanında veri madenciliği hileli aramaların yakalanması, müşteri profillenmesi, kaçak hat kullanımlarının belirlenmesi, CRM, müşteri sadakati (Customer Churn analysis), iletişimi sağlayan ağlarda problemli bölgelerin tespiti, kullanıcı alışkanlıklarının belirlenmesi ve buna göre yeni ürünlerin sunulması, GSM kullanım görselleştirmesinde kullanılır.

Son yılların trend konularından biri olan biyoenformatik alanında veri madenciliği, hastalıkların coğrafi olarak haritalarının hazırlanması, tanı koyma, kanser riski, sağlık politikası ortaya koyma, protein veya gen dizilimlerinin analizi, görselleştirmesi, indekslenmesi, kategorilenmesi veya aranmasında kullanılır.

Kolluk kuvvetlerinin saldırganları yakalamak için yapacağı çalışmalarda veri madenciliği, akan verinin analizi, davranış analizi, monitor ve alarm mekanizmaları, görselleştirme ve sorgu araçları konularında kullanılabilir.

Tüm dünyada çok önemli noktalara gelen elektronik ticarette veri madenciliği, web sayfalarına yapılan ziyaretlerin analizi ve sayfanın buna göre yenilenmesi, e-CRM uygulamaları, satış artışı sağlamak için hangi kampanyaların hangi müşteri grubuna yapılması gerektiğinin belirlenmesi konularında kullanılabilir.

Sigortacılık sektöründe veri madenciliği, poliçesini yeniletmek istemesi muhtemel müşterilerin tahmin edilmesi, riskli olan müşteri gruplarının belirlenmesi, sigortacılık suistimallerinin tespit edilmesi, branşa göre prim tespiti gibi konularda kullanılabilir.

Bu sektörlerin dışında eğitim, turizm ve otelcilik, taşımacılık ve ulaşım ve yerel hizmetler alanlarında da veri madenciliği uygulamaları kullanılabilir.

5. VERİ MADENCİLİĞİ SÜREÇLERİ

Veri madenciliği sadece değişkenleri modele koyup çıktı üretmesi beklenen bir süreç olarak düşünülmemelidir. Veriden bilgiye giden yolda birçok işlem gerekmektedir. Bu bölümde, modellemenin öncesinde ve sonrasında yapılması gerekenler anlatılacaktır.

5.1. Model ve Algoritma Kavramları

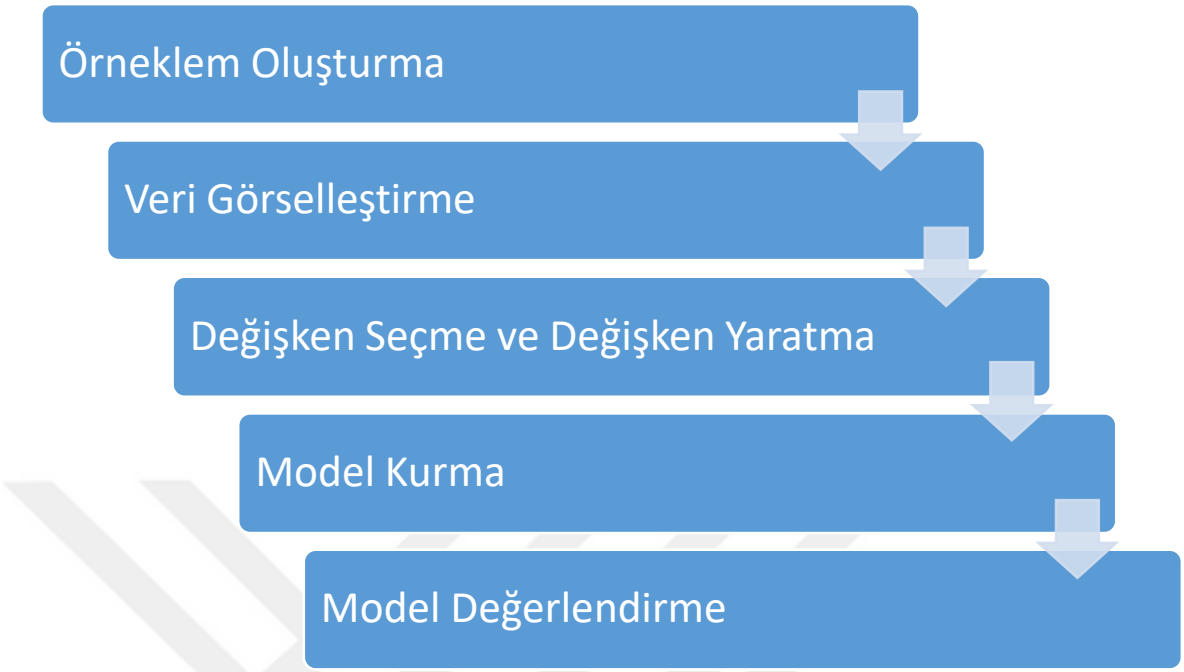
Öncelikle kavramlardan başlamak faydalı olacaktır. Model, teknik ve algoritma kavramları arasındaki farklar ele alınacaktır.

Model, olayı veya sistemi tanımlamak için kullanılır. Kavramsallaştırma gücünü arttırır. Veri madenciliğinde sınıflandırma, kümeleme ve bağlantı analizi birer modeldir. Modeller yaklaşım tarzını belirler. Örneğin sınıflandırma modeli uygulanacaksa gözetimli öğrenme (supervised learning), kümeleme modeli uygulanacaksa gözetimsiz öğrenme (unsupervised learning) kullanılır. Modelin uygulanması için teknikler yani yöntemler kullanılır. Örneğin, karar ağaçları, yapay sinir ağları, bayes sınıflandırması birer tekniktir. Tekniğin makinelerde hayata geçebilmesi için algoritmalar yazılır ve kullanılır. Örneğin J48 algortiması ile karar ağacı tekniğinin sonuçları elde edilir. Bir tekniğin birden fazla algoritması olabilir.

5.2. Semma

Veri madenciliğinde sonuca gitmek için belli prensipler kullanılır. Bunlardan kabul görmüş iki tane metodoloji crisp_dm (Cross-Industry Standard Process for Data Mining) ve semmadır. Bu çalışmada semma metodolojisi kullanıldığı için onun hakkında bilgi verilecektir. Semma (Sample, Explore, Modify, Model, Assess), 1966 yılında

kurulan ve analitik yaklaşımlar konusunda ürünler geliştiren SAS (The Statistical Analysis System) şirketinin veri modelleme prensibidir. Şekil 4'te gösterildiği gibi öncelikle Sample aşamasında örneklem seçilir. Bunun temel amacı, çok büyük veri kümeleri üzerinde çalışırken makine performansından kaynaklı yavaşlıklara takılmamaktır. Anakütleyi en iyi temsil ettiği düşünülen örneklem rastgele (random) şekilde seçilir. Ayrıca gözetimli öğrenme yapılıyorsa veri üç parçaya ayrılır. Bunlar öğrenme verisi (train data, %40), doğrulama verisi (validation data, %30) ve test verisidir (test data, %30). Bazen test verisi Sample içinde seçilmez ve model oluşturulduktan sonra test etmek için kullanılır. Bu durumda öğrenme ve doğrulama verilerinin ikisi de eşit sayıda olacak şekilde bölüm gerçekleştirilir. Explore aşamasında görselleştirme yapılarak bağımsız değişkenlerin (variable, attribute, feature) genel görüntüsü ortaya çıkarılmaktadır. Sürekli bir değişkense normal dağılıp dağılmadığına, kategorik bir değişkense frekanslara bakılmalıdır. Anomali tespiti de diyebileceğimiz bu işlemler sonucunda veriler daha anlamlı hale gelecektir. Ayrıca gözetimsiz öğrenme yolu takip ediliyorsa kümeleme (clustering) teknikleri de burada kullanılmaktadır. Modify aşamasında hedef değişkeni en iyi açıklayan değişkenlerin seçilmesi ve diğerlerinin modele girmeden elenmesi sağlanır. Değişkenin türüne göre korelasyon sonuçları incelenir. Ayrıca veri dönüşümleri de gerçekleştirilebilir. Model aşamasında temel veri madenciliği teknikleri uygulanır. Assess aşamasında sonuçlar değerlendirilir ve modellerin doğruluk karşılaştırılması yapılır [26].



Şekil.4. SEMMA Metodolojisinin Şeması

5.3. Problemin Tanımlanması

Veri madenciliğinin en önemli noktalarından birisi amacın (aim) ne olduğudur. Amacın çok net bir şekilde çizilmesi gerekir. Aksi taktirde yapılan çalışma rotası belli olmayan bir yelkenliden farksız olur. Hedef değişkenin ne olacağı, gözetimli mi, yarı gözetimli mi yoksa gözetimsiz öğrenme mi uygulanacağı baştan belirlenmeli ve ona göre yol izlenmelidir. Veri ambarından oluşturulacak olan veri marketleri (data mart) de ona göre seçilmelidir. Bu çalışmada, amaç hangi özelliklere sahip kişilerin karavan (mobile home) sigortası yaptırdığının belirlenmesi ve yapılacak kampayalarda hangi müşteri kitlesine hitap edilmesi gerektiğini ortaya çıkaracak bilgiyi elde etmektir.

5.4. Veri Ön İşleme (Data Preprocessing)

Veri ön işleme, modelleme çalışmalarının en fazla zaman harcanan ve doğru bir şekilde yapıldığı taktirde güzel sonuçların elde edilebileceği ve KDD aşamalarının başarıyla bitirilmesini sağlayacak bir süreçtir. Semma sürecini de hızlandırır. Çünkü orada

yapılacak bazı işlemler önceden bitmiş olur. Veri kalitesini arttırmak amacıyla uygulanır. Özellikle son yıllarda veri sayısının artmasıyla birlikte ön işleme sürecinde kullanılan tekniklere olan ihtiyaç daha da artmıştır.

5.4.1. Veri Tipleri

Veri tipleri, verinin karakteristiğinin bilinmesi ve ona uygun işlem yapılması için ipucu verir. Bağımsız değişkenlerin türü bilinmeli ve ona göre ön işlemeden geçirilmelidir. Bazen tür değişikliğine gidilebilir. Örneğin doğum yılı tarih formatındayken, yaşa çevrilecek sürekli bir değişkene dönüştürürebilir. Başlıca veri tipleri, nominal, sıralı (ordinal), ikili (binary), sayısal (numeric) verilerdir.

Nominal tipli değişkenler nesne isimleri veya sembolleri olabilir. Örneğin saç rengi, evlilik durumu, Türkiyenin coğrafi bölgeleri nominal değişkenlerdir. Nominal tipli değişkenlere kategorik değişkenler de denir. Sıralı değişkenlerde sıra veya derecelendirme söz konusudur. İçecek boyutu (small, medium, large), iş yeri ünvanları (uzman, kıdemli uzman, takım yöneticisi, müdür, direktör), müşteri memnuniyeti (0,1,2,3,4) sıralı değişkenlerdir. İkili değişkenlerde iki kategori bulunmaktadır. Sigara içme durumu ve sigorta poliçesi olma durumu ikili değişkenlerdir. Sayısal değişkenlere sürekli değişkenler de denir. Tam sayı veya reel sayıdırlar. Bir aralık veya oranla ölçeklendirilebilirler. Sıcaklık, sigorta poliçe sayısı, internet bankacığı sistenine giriş sayısı sayısal değişkene verilebilecek örneklerdendir [27].

5.4.2. İstatistiki Değerleri Yorumlama

Mevcut veri kümesiyle alakalı fikir verebilecek önemli istatistiki ölçüler vardır. Bu ölçülerin değerlendirilmesinde verilerin tipleri önemlidir. Çünkü kategorik değişkenler için anlamlı olan ölçümler ve sonuçlar, sürekli değişkenler için birşey ifade etmeyebilir.

Tam tersi de geçerlidir. İkili değişkenler, karakteristik açıdan kategorik değişkenlerle benzerlik gösterirler. Kategorik değişkenlerde frekans dağılımı önem arz etmektedir. Sayısal bir ölçüyle ifade edilmedikleri için hangi durumun kaç tane örnekte varlık gösterdiği önemlidir. Aykırılık (outlier) analizi elde edilen frekans ve yüzdelere bakılarak yapılabilir. Sürekli değişkenler için ise ortalama (mean), minimum değer, maksimum değer, standard sapma, standard hata, boş değerler (null), varyans, mod, genişlik (range), çeyreklik değerler (quartile), ortanca (medyan) gibi ölçümlere bakılır. Mod, değişken için en fazla bulunan değerdir. Genişlik, maksimum değerle minimum değer arasındaki farktır. Çeyreklik değerler, %25 (Lower Quartile), %50 , %75 (Upper Quartile) değerleridir. Değişkenin aldığı değerlerin minimum olanından maksimum olanına kadar sıralanıp yüz parçaya bölündüğünde dilimdeki 25. , 50. Ve 75. sayıya denk gelen sayılardır. Medyan ise ortadaki sayıdır. Değişkende tek adet değer varsa ortadaki değer bellidir. Eğer çift adet değer varsa ortada iki değer vardır ve medyan bu iki değerin aritmetik ortalamasıdır. Değişkenleri yorumlamada bu ölçümler önemlidir. Değişkenin nasıl bir dağılım gösterdiği, anlamlı olup olmadığı, hedef değişkeni ifade etme gücünün olup olamayacağı, sürekli bir değişken olarak kalması mı yoksa bayrak (flag) ataması yapıp gölge (dummy) değişken olarak mı modele sokulması gerektiği gibi konulara bu ölçümlere bakarak karar verilir. Eğer bayrak ataması yapılacaksa hangi değerden sonrası için 1 ve öncesi için 0 değerinin verileceği, minimum ve maksimum değerler, %1, %5, %10, %25, %50, %75, %90 ve %99'daki değerlere bakılarak belirlenir.

5.4.3. Veri Temizleme (Data Cleaning)

Modellenecek olan verinin anlamlı sonuçlar verebilmesi için kesin olmak, tutarlı olmak, inandırıcı olmak ve yorumlanabilir olmak gibi belli özelliklere sahip olması gerekmektedir. Veri kalitesinin arttırılması için veri temizliği yapmak gerekmektedir. Veri temizliği ile var olan kirli veriyi silmenin yanında boş ve gürültülü değer problemlerine bu problemlerin karakteristiğine uygun çözümler getirmek de amaçlanmıştır.

5.4.3.1. Boş (Null) Değerler

Modellemede elde edilecek sonucun doğruluğunun yüksek olması ve uygulanabilir olması için modele girecek olan verinin de doğru olması gerekir. Ancak verinin doğruluğu çeşitli sebeplerle %100 sağlanamayabilmektedir. Bunun en sık karşılaşılan örneklerinden birisi bazı değerlerin boş olmasıdır. Tabloda boş değer olmasının başlıca sebepleri arasında donanımsal bozukluklar, veri girişinin eksik yapılması, zorunlu olmayan alanların önemsenmemesi, uyuşmazlık nedeniyle silinen veriler olması, yaşını, kilosunu söylemek istemeyen müşterilerin olması vardır. Problemi çözmek için çeşitli yöntemler geliştirilmiştir. Bunlardan bazıları aşağıdaki gibidir [27].

a. Boş değerler ihmal edilebilir.
b. Boş değerler elle doldurulabilir.
c. İlgili değişkenin dolu olan değerlerinin ortalaması boş olan değerlere yazılabilir. Böylece değişkenin istatistiksel değerleri korunmuş olur.
d. Önemli bir kategorik değişken bazında ortalama almak suretiyle doldurma yapılabilir. Bu değişkenim seçimi iş bilgisiyle olmalıdır.

Ön işleme sürecinde boş bırakılan değişken değerleri veri madenciliği araçlarında (tool) genellikle görmezden gelinir.

5.4.3.2. Gürültülü (Noisy) Değerler

Bir değişkenin genel görüntüsünün dışına çıkmış değerlerdir. 1000 tane değere sahip sayısal bir değişkenin 995 değerinin 1 ile 50 arasında olup diğer 5 değerinin 500 ile 600 arasında olması, gürültülü değere örnektir. Ölçümle elde edilen değişkenlerde, ölçümlerdeki hatalardan kaynaklanabilir. Ayrıca veri toplama araçlarındaki hatalar, veri giriş promlemleri ve veri iletim problemleri de gürültülü veri oluşmasına sebep olabilir. Verideki böyle anormal değerleri yumuşatmak (smoothing) gerekmektedir. Bunun için izlenmesi gereken yöntemler vardır [27]:

a. Kutulama (Binning)
b. Lineer Regresyon
c. Aykırılık Analizi (Outlier Analysis)

5.4.4. Veri birleştirme (Data Integration)

Modelin kurulması esnasında kullanılacak olan tablodaki değişkenler çok farklı kaynaklardan (veritabalarından veya veri ambarlarından) bir araya gelmiş olabilir. Farklı kaynaklardan toplanan verinin bir araya getirilmesi işlemine veri birleştirme veya bütünleştirme denir. Veriler, farklı kaynaklarda farklı formatlarda tanımlanmış olabilir. Örneğin, kişilerle ilgili toplanan verilerde cinsiyet değişkeni birisinde "Erkek" ve "Kadın", başka birisinde 1 ve 0, diğer bir kaynakta ise "E" ve "K" şeklinde olabilir. Sayısal bir değişken bir veri tabanında "integer" tipinde tutulurken başka bir veri tabanında "float" tipinde tutulabilir. Tüm bu farklılıklar çeşitli problemlere sebep olabileceğinden dolayı birleştirme aşamasında aynı formatta bir araya getirilmelidir.

Veri birleştime, modele girecek olan veriyi zenginleştireceği için önemlidir. Çünkü modele seçeceğimiz değişkenleri belirlerken alternatifin fazla olmasını sağlar. Özellikle amacımız tahminleyici bir modelleme yapmak ise hedef (target) değişkeni en iyi açıklayan değişkenleri modelde kullanmak gerekmektedir. Bunun için hedef değişkeni etkileyeceği düşünülen mümkün olduğunca fazla değişken üretip içlerinden en iyilerini seçmek gerekir. Veri ambarlarında dönüştürülmüş olan değişkenlerden birkaç tanesinin belirli bir kombinasyonla bir araya gelmesiyle yeni değişkenler üretilebilir. Böyle değişkenlere türetilmiş değişkenler denir. Örneğin, uzunluk ve genişlik değerlerinin verildiği bir tabloda alan bilgisi türetilebilir. Uzunluk ve genişlik değişkenlerin hedef değişkeni açıklama gücü zayıf olmasına rağmen bir araya geldiklerinde güçlü bir değişken ortaya çıkarabilirler.

5.4.5. Veri Dönüştürme (Data Transformation)

Veri dönüştürme, modele girecek olan değişkenin formatına ve yapısına olumlu manada manipülasyon yapmaktır. Örneğin, doğum tarihini yaşa dönüştürmek, günlük satış miktarını haftalık satış miktarına dönüştürmek model içinde daha kullanışlı olabilir. Kategorik değişkenler genelleştirilip daha az kategori oluşacak şekilde düzenlenebilir.

Değişkenlerin varyanslarının ve ortalamalarının birbirlerinden yüksek miktarda farklı olduğu durumlarda ortalaması ve varyansı büyük olan değişkenin diğerlerine göre daha fazla olur. Ayrıca değişkenlerin sahip olduğu çok küçük ve çok büyük değerler de çözümlemelerin sağlıklı biçimde yapılmasını engeller. Bundan dolayı bu değişkenler üzerinde normalleştirme veya standartlaştırma yapılması gerekmektedir. Veri dönüştürme işlemi olan normalleştirme yöntemlerinden en çok kullanılan üç tanesi aşağıdaki gibidir [27].

a. Min-maks normalleştirmesi

Veri üzerinde doğrusal bir dönüşüm yapmayı sağlar. En büyük ve en küçük sayısal değerler belirlenir ve diğer değerler buna uygun biçimde 0 (minimum değer) ile 1 (maksimum değer) arasındaki sayılara dönüştürülür.

Dönüştürme bağıntısı :

$$X^*=(X - X_{min})/(X_{max} - X_{min}) \quad (1)$$

Denklem 1'de bulunan simgeler aşağıdaki anlamlarda kullanılmıştır:

- X^*: dönüştürülmüş değerleri
- X : gözlem değeri
- X_{min} : en küçük gözlem değeri
- X_{max} :en büyük gözlem değeri

b. Z-skor normalleştirme

Verilerin ortalaması ve standart hatası düşünülerek yeni değerlere dönüştürülmesi esasına dayanmaktadır. Sıkça kullanılan bir dönüşüm şeklidir.

Dönüştürme bağıntısı :

$$X^* = (X - \mu)/\sigma x \quad (2)$$

Denklem 2'de bulunan simgeler aşağıdaki anlamlarda kullanılmıştır:

- X* : dönüştürülmüş değerleri
- X : gözlem değerlerini,
- μ : verilerin aritmetik ortalamasını,
- σx : gözlem değerlerinin sapmasını

c. Ondalık ölçekleme

Sürekli değişkenlerin değerlerinin ondalık kısmı üzerinden normalleştirme sağlanır. Hareket edecek olan ondalık büyüklük, değişkenin maksimum mutlak değerine göre değişir.

Dönüştürme bağıntısı :

$$X^* = X /(10)j \quad (3)$$

Denklem 3'te bulunan simgeler aşağıdaki anlamlarda kullanılmıştır:

- X*: dönüştürülmüş değerleri
- X : gözlem değeri
- j=max(|X*|) < 1 yapan en küçük tam sayı

5.4.6. Veri İndirgeme (Data Reduction)

Veri kümesi çok büyük olduğunda algoritmaların çalışması çok zaman alır ve bazılarının uygulanması makinelerin sınırlı hızından dolayı oldukça zorlaşır. Bu yüzden aynı veri kümesinin daha küçük bir örneğinde işlem yapmak daha efektif olacaktır. Ancak bu veri kümesinin asıl veri kümesiyle benzerlik göstermesi gerekmektedir. Bu amaçla geliştirilen çeşitli teknikler aşağıda listelenmiştir.

a. Boyut indirgeme (Dimension Reduction)

Değişkenlerden gereksiz olanların veya açıklama gücü zayıf olanların kaldırılmasıdır. Telefon numarası, e-mail gibi veriler çoğu zaman modelde anlam ifade etmez. Bu yüzden bu tür değişkenler ön işleme sürecinde elenmelidir.

b. Kesikli hale getirme (Discretization)

Bazı algoritmalarda sadece kategorik değişkenlerle işlem yapıldığından dolayı sürekli değişkenlerin değerlerinin kesikli hale dönüştürülmesi gerekmektedir. Bu yüzden verinin karakteristiğine göre kesikli bir hale getirilirler. Örneğin, normalde sürekli olan yaş değişkeninin 1-15, 16-25, 26-35, 36-45, 46+ olacak biçimde aralıklara bölünmesi.

5.4.7. Korelasyon Analizleri

Değişkenler, madenciliği yapılacak olan bir veri kümesinin özelliklerini keşfetmemizi sağlar. Değişken ismi istatistikte kullanılır ve farklı disiplinlere göre farklı isimleri vardır. Veri ambarı (datawarehouse) alanında boyut (dimension), makine öğrenmesinde (machine learning) özellik (feature) denmektedir. Veri madenciliği ve veri tabanı profesyonelleri ise nitelik (attribute) demektedir.

Değişkenlerin türleri vardır ve kendi aralarındaki ilişkiler buna göre şekillenir. Değişkenler arası ilişkinin ölçülerinden birisi korelasyondur. Korelasyon, iki tane rassal değişkenin arasındaki ilişkinin gücünün ve yönünün belirlenmesine yardımcı olan bir

ölçümdür[28]. Korelasyon katsayısı -1 ve +1 sayıları arasındaki değerleri alır. Korelasyon katsayısının 0’ dan büyük olduğu durumda pozitif yönlü bir ilişki, 0’dan küçük olduğu durumda ise negatif yönlü bir ilişki vardır. Korelasyon ile birlikte değişme gibi bir algı oluşmaktadır. Ancak bu her zaman doğru değildir. İki tane bağımsız değişken arasında yüksek korelasyon çıkması onların birbirini etkilediğini göstermez. Meşhur korelasyon örneklerinden biri bu durumu açıklar. İskandinavya'da 19. yüzyılın sonunda ve 20. yüzyılın başında yapılan leylek ve çocuk doğumları incelendiğinde çok yakın bir pozitif yönlü korelasyon bulunmuştur [28]. Buradan leyleklerin yeni doğan çocukları getirdikleri anlamı çıkmaz. Benzerlik olmasında ekonomik gelişmeler ve şehirleşme sebepleri ön plana çıkmıştır. Yani azalmanın nedeni sebep-sonuç ilişkisinden kaynaklı ortaya çıkmamıştır.

Model için seçilecek değişkenleri belirlemede korelasyon analizi oldukça önemlidir. Korelasyon katsayısını yorumlamada önemli noktalardan birisi değişkenlerin bağımlı ya da bağımsız olmasıdır. Bağımlı bir değişkenle (hedef değişken) bağımsız değişken arasındaki korelasyonun yüksek çıkması, bağımsız değişkenin hedef değişkeni açıklama gücünün yüksek olabileceğini göstermektedir. Burada nedensellik bakış açısıyla da bakmak gereklidir. Buna iş bakış açısı da denmektedir. Çünkü korelasyon, nedensellik değildir. Hedef değişkenle yüksek korelasyonlu bir ilişki çıkmış olsa bile reel dünyada aslında öyle bir sebep sonuç ilişkisi olmayabilir. Bu durumun tersi de geçerlidir. Düşük korelasyon olması o değişkenle hedef değişken arasında sebep sonuç ilişkisinin olmadığı anlamına gelmez. Genel olarak baktığımızda korelasyon katsayıları fikir verir ama nihai sonucu söylemez. Sonuç olarak, ne tamamen korelasyona dayalı değişken seçimi ne de tamamen iş bakış açısıyla bakıp değişken seçmek doğru olmaz. Bunun yerine ikisine de bakıp öyle karar vermek veri madenciliği felsefesine daha uygundur.

Korelasyon analizi yapmanın çeşitli yöntemleri vardır. Sürekli değişkenler için formülü Denklem 4’te yer alan Pearson Korelasyon Katsayısı hesaplaması en çok kullanılan yöntemler arasındadır. Sıralı değişkenler için Spearman Korelasyon Katsayısı ve Goodman and Kruskal's gamma değerleri yorumlanır. Kategorik değişkenler için ise Cramer’s V ve Somers’ D değerlerine bakılır. Cramer’s V değeri, iki tane bağımsız kategorik değişken arasındaki ilişkiyi belirlemek için kullanılır. Somers’ D değeri ise,

kategorik bağımsız bir değişkenle (tahminleyici değişken) kategorik bağımlı bir değişken arasındaki ilişkiyi incelemek için kullanılır.

Pearson korelasyon katsayısı şu şekilde bulunur :

$$r = (\sum x_i \, y_i) / (\sum x_i^2 \sum y_i^2)^{1/2} \quad (4)$$

Burada, r Pearson korelasyon katsayısını, x ve y ise aralarındaki korelasyon katsayısı hesaplanacak olan değişkenleri belirtmektedir.

Spearman korelasyon katsayısının hesaplanmasında d (distance) kullanılır. Öncelikle korelasyon katsayıları belirlenecek olan iki değişkenin değerleri küçükten büyüğe sıralanır ve 1'den başlayarak artan numaralar verilir. Daha sonra iki değişken (x ve y) için karşılıklı elemanların sıra sayısı (x_i, y_i) arasındaki fark (x_i - y_i) bulunur. Bu değer d bilinmeyenini verir. Hangi değişkenin sırasının başa yazılacağının bir önemi yoktur. Çünkü korelasyon katsayısı hesaplamasında d'nin karesi alınacağı için iki durumda da aynı sayı elde edilecektir. Korelasyon denklemi şu şekildedir:

$$\rho = 1 - (6 \sum d_i^2 / n(n^2 - 1)) \quad (5)$$

Burada, n gözlem sayısını, ρ Spearman korelasyon katsayısını ve d ise yukarıda anlatılmış olan uzaklık değerini vermektedir.

Modelde kullanılacak olan değişkenleri belirlemek için yapılan çalışmada korelasyon analizini bağımsız değişken ile hedef değişken arasındaki ilişki ve bağımsız değişken ile bağımsız değişken arasındaki ilişki şeklinde ele almak gerekir. Öncelikle değişkenlerin hedef değişkenle olan korelasyon katsayılarına bakılır. Hedef değişkenle yüksek korele olan ve iş bakış açısıyla da modele girmesi önemli bulunan değişkenler seçilir. Daha sonra bu değişkenlerin kendi aralarındaki korelasyon katsayılarına bakılır. Eğer yüksek çıkarsa iki değişkenden birini devre dışı bırakmak gerekecektir. Çünkü değişkenlerin ayrı ayrı güçleri iyi olsa da ikisinin de modele girmesi durumunda birbirlerinin güçlerini etkisiz hale getirmeleri söz konusudur. Böyle bir durumda genellikle hedef değişkenle olan korelasyon katsayısı daha yüksek olan değişken seçilir. Ancak başka değişkenlerle olan ilişkiler ve tüm kombinasyonu hesaba katmak gerekeceği için hepsini birlikte değerlendirmek gerekmektedir. Korelasyon katsayısı, %70' den fazlaysa yüksek kabul edilir. Ancak amaçlanan işe göre bu oran değişebilir. Özellikle

dengesiz (imbalanced) bir veri kümesiyle çalışılıyorsa yüksek bir korelasyon bulmak güç olabilir. Çünkü hedef değişkenin ikili olduğunu varsayarsak çok az bir kısmı 1 olduğu için, korelayonun yüksek çıkma ihtimali daha düşüktür.

5.5. Modelin Kurulması ve Değerlendirilmesi

Ön işleme süreci bittikten sonra veri kümesi, üzerinde model koşulacak bir yapıya ulaşır. Hangi yaklaşımların ve tekniklerin kullanılacağı baştan belirlenmelidir. Veri toplama stratejisi de ona göre olmalıdır. Bundan sonraki süreçte modelin kurulması ve değerlendirmesi devreye girmektedir.

5.5.1. Modelleme ve Modellerin Karşılaştırılması

Modelleme, tek başına birşey ifade etmeyen veriden kıymetli bir kaynak hüviyetindeki bilgiye ulaşmadaki işlemler bütünüdür. Modellemenin sonucunda ortaya çıkan ürüne de model denir. Veri kümesi model kurmak için hazır hale geldiğinde veri madenciliği teknik ve algoritmaları kullanılır. Kullanılacak tekniklerin birden fazla algoritması mevcut olabilir. Örneğin, karar ağaçlarını C 4.5 veya C 5 algoritmasıyla oluşturabiliriz. Hedeflediğimiz sonuçları elde etmek için birden fazla model kurabiliriz. Kurduktan sonra en iyisini seçeriz. En iyi modeli seçmek için çeşitli yöntemler vardır. Bunlar arasında Karışıklık Matrisi (Confusion Matrix) çok kullanılmaktadır. Doğru tahmin edilen veri sayının toplam veri sayısına oranı doğruluk oranını vermektedir. Model seçiminde doğruluk oranı önemlidir.

Tablo.5. Karışıklık Matrisi

	Öngörülen sınıf (Predicted Class)		
Gerçek Sınıf (Actual Class)		C_1 (Positive)	C_2 (Negative)
	C_1 (Positive)	True positive TP	False negative FN
	C_2 (Negative)	False positive FP	True negative TN

Karışıklık matrisinde yer alan alanlardan "True positive" (TP) alanı, doğru tahminlenen olumlu örneklerin sayısını, "False negative" (FN) alanı yanlış tahminlenen olumsuz örneklerin sayısını, "False positive" (FP) alanı yanlış tahminlenen olumlu örneklerin sayısını, "True negative" (TN) alanı ise doğru tahmin edilen olumsuz örneklerin sayısını vermektedir.

Modelin doğruluğu doğru tahminlenen örnek sayısının toplam örnek sayısına oranıdır.

$$\text{Doğruluk} = (TP + TN) / (TP + TN + FP + FN) \quad (6)$$

Modelin duyarlılığı (sensitivity) tüm doğru örneklerin arasından ne kadar doğrunun tahmin edildiğinin bir ölçüsüdür.

$$\text{Duyarlılık} = TP / (TP + FN) \quad (7)$$

Modelin belirliliği (specificity) tüm yanlış örneklerin arasından ne kadar yanlışın tahmin edildiğinin bir ölçüsüdür.

$$\text{Belirlilik} = TN / (TN + FP) \quad (8)$$

Karışıklık matrisinde ölçülen kriterlerin dışında çalışma hızı, eksik veya kirli verilerden ne kadar etkilendiği, ölçeklendirilebilir olması yani veri sayısı arttıkça sonucun değişmemesi ve yorumlanabilir olması da modelin başarı kriterleri arasındadır. En iyi modeli seçmek, yapılmak istenen işe göre değişir. Bazı modellerde hız önemlidir. Ancak bazen hız hiç önemli olmayabilir. Örneğin, kişilerin belirlenen özelliklerine göre kanser hastası olup olmayacağıyla alakalı yapılan bir tahminlemede hızdan ziyade doğruluk çok önemlidir.

5.5.2. Değerlendirme

Modeller kurulduktan ve model seçildikten sonra değerlendirilmesi ve hedefler ile uyumlu olup olmadığının kontrol edilmesini gerekir. Öncelikle modelin hedefleri ne ölçüde karşıladığı değerlendirilir. Model gerçek veriler ile test edilir. Kaliteli bir şekilde çalışan modelin belli bir zamandan sonra güncellemesinin yapılması ve tekrar şekillendirilmesi gerekebilir.

6. TEMEL VERİ MADENCİLİĞİ METOTLARI

Veri kümesi model kurmak için hazır hale geldiğinde çeşitli metotlar uygulanır. Başlangıçta belirlenen amaç doğrultusunda teknikler uygulanır. İzlenecek yola göre üç farklı modelleme çeşidi vardır.

6.1. Tahminleyici (Predictive) Modelleme

Geçmişe ait verinin desenlerini çözerek geleceğe yönelik çıkarımlar yapmayı sağlar. Tahminleyici modelleme, tahminlenecek olan değişkenin tipine göre ikiye ayrılır. Hedef değişken kategorik ise sınıflandırma, sürekli ise regresyon modellemeleri kullanılır.

6.1.1. Sınıflandırma (Classification)

İki veya daha fazla sınıf içeren kategorik değişkenlerin modellenmesinde kullanılır. Gözetimli öğrenme metodu kullanılarak geçmişteki veri kullanılarak gelecekte olacaklar için tahminleme yapılmaktadır. Sınıfların net bir şekilde belirlenmesi gerekmektedir. Sınıflandırma ile tahminleyici değişkenlerin sahip olduğu değerlere göre tahminlenen değişkenin hangi sınıfına ait olduğunun öğrenilmesi ve daha sonraki verilerde bunun kullanılması amaçlanmaktadır. Örneğin, bir malda bulunan özelliklerle müşteri özellikleri eşleştirilebilir ve bir müşteri için uygun ürünler veya bir ürün için uygun müşteri profilleri belirlenebilir. Çoğunlukla genç erkeklerin okuduğu bir dergiye reklam vermek isteyen bir otomotiv şirketi geçmiş müşteri hareketlerinin analizi ile bu grup için hangi model araçların reklamının uygun olacağını sınıflandırma yardımıyla tespit edebilir. Birçok sınıflandırma tekniği ve algoritması vardır. Bunlardan bazıları aşağıda belirtilmiştir.

6.1.1.1. Bayes Sınıflandırması

Bayes sınıflandırması, istatistiksel bir sınıfladırmadır. Mevcut sınıflandırılmış verileri kullanıp yeni bir verinin var olan sınıflardan herhangi birine girme olasılığı hesaplanır [30]. İngiliz matematikçi Thomas Bayes tarafından geliştirilen Bayes teoremini temel almaktadır. Bayes teoremi aşağıdaki gibi ifade edilebilir:

$$P(C_i|X) = (P(X|C_i) P(C_i)) / P(X) \quad (9)$$

Burada,

$X = (X_1, X_2, \ldots, X_n)$, veri kümesindeki her bir örnektir. n sayısı kümedeki bağımsız değişken sayısını gösterir.

C_i, kaç farklı sınıf olduğunu, bir başka deyişle hedef değişkende kaç farklı kategori olduğunu gösterir.

Eğer X örneğini oluşturan boyutlar (değişkenler, özellikler) birbirinden bağımsız ise aşağıdaki denklem kullanılabilir.

$$P(C_i|X) = P(x_1|C_i) P(x_1|C_i) P(x_2|C_i) ..... P(x_n|C_i) \quad (10)$$

Bayes algoritması çalışırken öncelikle öğrenme kümesinde yer alan hedef değişkenin sınıflarının bulunma sıklığı ($P(C_i)$) hesaplanır. Daha sonra $P(X|C_i)$ ve $P(X)$ değerleri hesaplanır ve X örneğinin C_i sınıfı içinde yer alma olasılığı hesaplanır. Tüm sınıflar için bu işlem uygulanır ve en yüksek olasılıklı sınıf seçilir. Örneğin cinsiyet, kilo, boy ve yaş değişkenlerine bakılarak giysi bedeni tahminlemesi yapılırken veya kredi kartı harcamalarına bakılarak müşteri sınıfı belirlenirken Bayes sınıflandırması kullanılabilir.

6.1.1.2. Destek Vektör Makineleri (Support Vector Machine)

Destek vektör makineleri, sınıflandırma için kullanılan makine öğrenimi algoritmalarından biridir. Sınıfların ayrılmasını sağlayan optimum bir çizginin bulunmasını sağlar. Çizginin sınıflara en uzak noktalardan çizilmesi gerekmektedir. İlk

olarak iki sınıflı doğrusal verilerin modellenmesinde kullanılmıştır. Daha sonra ikiden fazla sınıflı ve doğrusal olmayan verilerin modellemesi de SVM aracılığıyla yapılmıştır. SVM, sınıfların birbirinden ayrılmasında en uygun karar fonksiyonun tahmin edilmesi (hiper-düzlemin tanımlanması) prensibine göre çalışmaktadır [31]. Destek Vektör Makineleri, sınıflandırma işlemini kareli optimizasyon işlemine dönüştürerek yapmaktadır. Dolayısıyla öğrenme aşamasında yapılan işlemlerin sayısında azalma olmakta ve böylece diğer algoritmalara göre daha hızlı çözümler sağlanmaktadır. Bundan dolayı, hacmi büyük veri kümelerinde büyük avantaj oluşturmaktadır. [32]. SVM, dış bükey optimizasyonuna dayalı bir algoritmadır. Dağılımdan bağımsız bir öğrenme kabiliyeti vardır. Sınıflandırma ve örüntü tanıma problemlerinin çözümünde kullanılmak üzere Vapnik tarafından geliştirilmiştir [31]. SVM'in temelleri Vapnik-Chervonenkis (VC) teorisine dayanmaktadır.

SVM kullanımında kernel (çekirdek) fonksiyon seçimi ve parametre optimizasyonu önemli bir role sahiptir. Kernel fonksiyonu sınıfların hangi şekillerle birbirinden ayrılacağının göstergesidir. Temel olarak dört farklı kernel fonksiyonu vardır: linear (doğrusal), polynomial, radial basis function (RBF) ve sigmoid.

Veri modellemesinde en sık kullanılan SVM algoritmalarının başında Sequential Minimal Optimisation (SMO) gelir. Örneğin WEKA aracında SMO kullanılmaktadır. SVM^{light} da popüler algoritmalar arasındadır.

6.1.1.3. K-En Yakın Komşu (KNN)

KNN, mesafeye dayalı sınıflandırma algoritmalarından biridir. Yeni kaydın sınıfı belirlenirken diğer kayıtlarla olan uzaklığı belirlenir. Belirlenen mesafelerden en yakın k tanesi seçilir. Algoritmanın ismi burdan gelmektedir. Seçilen kayıtlar en çok hangi sınıfta yer alıyorsa, yeni kaydın da o sınıfta yer aldığı kabul edilir.

K değeri algoritma çalışmadan önce belirlenir. Değerin yüksek seçilmesi birbirine benzemeyen noktaların bir araya toplanmasına sebep olur. Genellikle kullanılan k değerleri 3,5 ve 7'dir [30].

Mesafe ölçümünün nasıl yapılacağı da önemli noktalardandır. Sıkça kullanılan mesafe ölçümleri Euclid (Öklid), Manhattan ve Minkowski uzaklık ölçüleridir.

a. Euclid mesafesi

En çok kullanılan ölçüm türüdür. $P = (p_1, p_2, \ldots, p_n)$ ve $Q = (q_1, q_2, \ldots, q_n)$ noktaları arasındaki mesafe (d) aşağıdaki gibidir:

$$d = (\sum (p_i - q_i)^2)^{1/2} \quad (11)$$

b. Manhattan mesafesi

Manhattan mesafesi, kareli ölçüm yapar. İki nokta arasındaki mesafeyi ölçerken tek boyutta hareket eder; iki veya daha fazla boyutta ilerlemez. P ve Q noktaları arasındaki uzaklık (d):

$$d = \sum |p_i - q_i| \quad (12)$$

ile hesaplanır.

c. Minkowski mesafesi

Öklid ve Manhattan mesafelerinin genelleştirilmiş halidir. $X = (x_1, x_2, \ldots, x_n)$ ve $Y = (y_1, y_2, \ldots, y_n)$ noktaları arasındaki uzaklık (d) aşağıdaki gibidir:

$$d = (\sum |x_i - y_i|^p)^{1/p} \quad (13)$$

KNN algoritmasının çalışma metodolojisi şu şekilde sıralanabilir [30]:

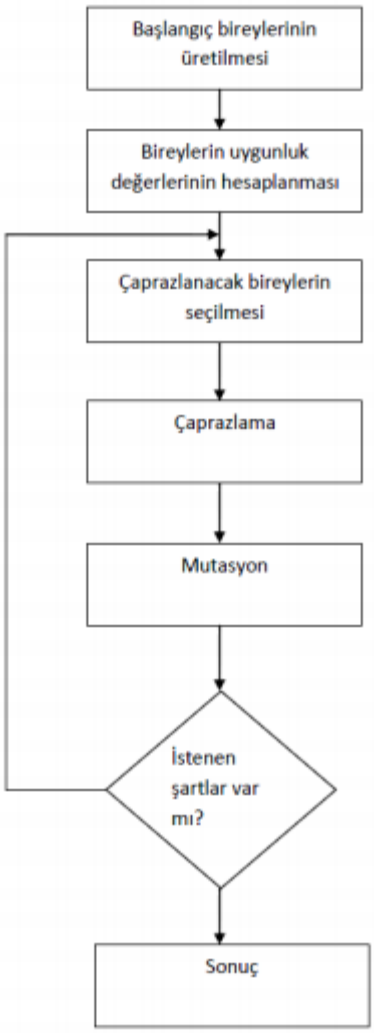
1. Mesafe ölçüm türü belirlenir.
2. K sayısı belirlenir.
3. Yeni kayda en yakın k adet kayıt belirlenir.
4. Bu k adet kaydın en çok içinde bulunduğu sınıf belirlenir.
5. Yeni kaydın bu sınıfa ait olduğu kabul edilir.

6.1.1.4. Genetik Algoritmalar

Genetik algoritmalar, ilk olarak Michigan Üniversitesi'nde makine öğrenmesi üzerine çalışan John Holland tarafından 1975 yılında ortaya atılmış ve Holland'ın öğrencisi olan David E. Goldberg tarafından geliştirilmiştir. Goldberg, gaz hattıyla alakalı bir problemin çözümünde genetik algoritmaları kullanmıştır [34]. Doğal seleksiyon ve evrim teorisi ilkeleri kullanılarak en iyinin hayatta kalması üzerine inşa edilmiş bir optimizasyon yöntemidir.

Genetik Algoritmalarda gen, kromozom ve popülasyon kavramları bulunmaktadır. Parametreler genleri, genlerin koleksiyonu da kromozomu temsil etmektedir. Her fert kromozomlarla işaretlenen popülasyonlardan oluşmaktadır. Popülasyon uygunluğu, belli kurallara göre minimize veya maksimize edilir. Yeni gelecek olan nesil, rastsal olarak oluşan enformasyon değişimi yoluyla meydana gelen dizilerin arasında hayatta kalmayı başaranların birleştirilmesiyle elde edilir [35]. Yapı gereği kötü bireyler (uygun olmayan çözümler) elenmektedir. Genetik Algoritmalardaki akış diyagramı Şekil 5'te gösterilmiştir.

Mevcut topluluktaki bireylerden çaprazlama ve mutasyon işlemleri uygulanacak olanlar seçilir ve yeni topluluk meydana gelir. Darwin'in teorisine göre iyi bireylerin yaşamını sürdürmesi ve yeni bireylerin bunlardan oluşması söz konusudur. Bundan dolayı seçilme yöntemlerinde daha iyi olan bireylerin seçilmesinin ihtimali daha fazladır. Rulet seçilimi, Turnuva seçilimi ve Sıralı seçilimi en bilinen seçilim yöntemleridir. Rulet seçiliminde, bir bireyin seçilme olasılığı, o bireyin uygunluk değerinin popülasyondaki toplam değere oranıdır. Sıralı seçilimde ise, en kötü uygunluğa sahip olan kromozoma 1 değeri verilir ve sonraki kromozomlara 2'den başlayarak artan değerler verilir. Son olarak turnuva seçiliminde topluluğun içinden rastsal bir şekilde k tane birey seçilir ve daha sonra bunlar arasından uygunluğu en yüksek olanı seçilir.



Şekil.5. Genetik Algoritmalar Akış Diyagramı

Genetik Algoritmaları kullanmanın birçok avantajı bulunmaktadır:

- Kavramlar kolay bir şekilde tasarlanmaktadır.
- Karmaşık problemler kolay bir şekilde çözümlenebilmektedir.
- Çok sayıda değişkenle çalışılabilmektedir.
- Karmaşık bir şekilde bulunan amaç fonksiyonları kolay bir şekilde optimize edilebilmektedir.
- Paralel makineler kullanılarak çalıştırılabilmektedir.
- Kısa sürede çalışabilmektedir.

Diğer taraftan bazı dezavantajlar da barındırmaktadır. Örneğin, optimizasyon aşamasında tüm farklı durumlar gözden geçirilmediği için en iyi çözüm elde edilmeyebilir.

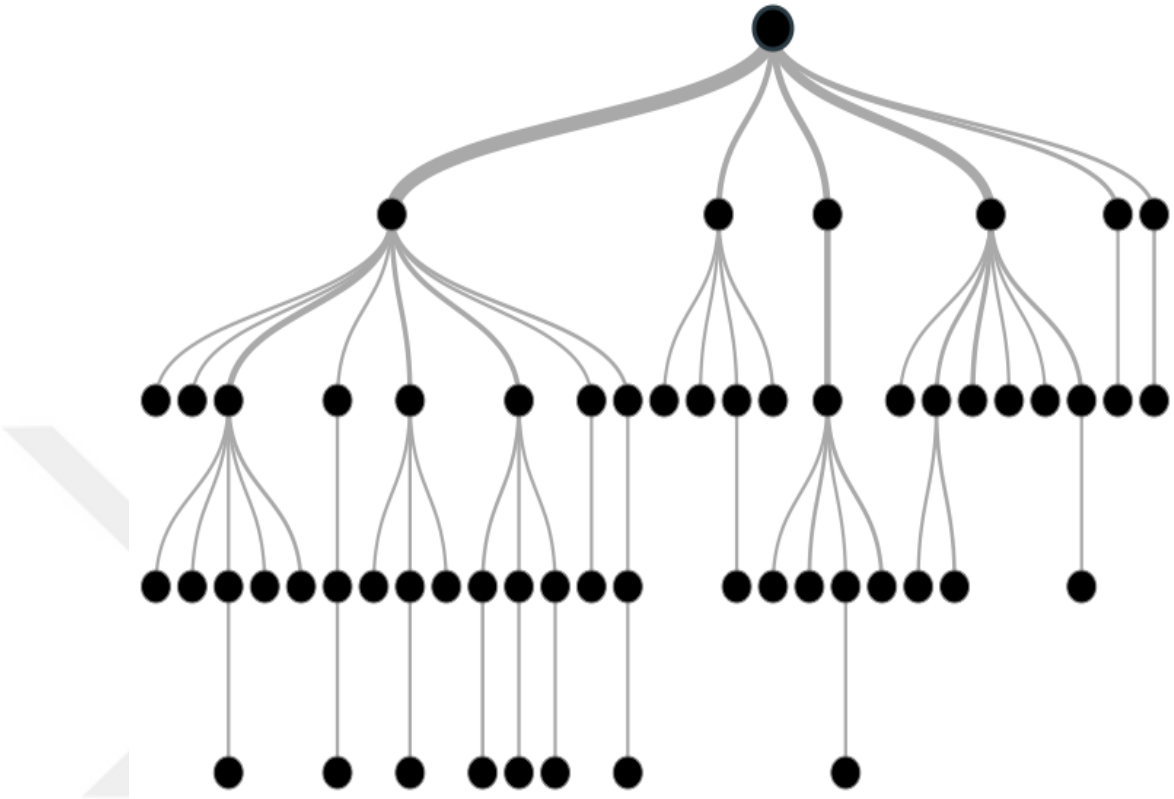
Genetik algoritmalar bilgi sistemleri, ekonomik ve sosyal sistem modelleri, finans, pazarlama, tesis yerleşimi, hücresel üretim, sistem güvenilirliği, araç rotalama gibi birçok alanda kullanılmaktadır.

6.1.1.5. Karar Ağaçları

Karar ağaçları, sınıfları belli olan örnek veri kümesinden Şekil 6'da görüldüğü gibi ağaç şeklinde bir karar yapısı kurularak gelecekteki verinin sınıfının tahminlenmesini sağlar. Tüme varım mantığı kullanılmaktadır. Ağaç tersten büyümektedir. En tepede kök bulunmaktadır. Kökten sonraki kısımda karar düğümleri, dallar ve uç yapraklar bulunmaktadır. Karar düğümlerinde, düğüme gelen verilerin test edilmesi, soruların sorulması ve düğümden sonraki yönelimler belirlenir. Bu yönelimler dallarla temsil edilir. Karar düğümlerinde girdi ve çıktı vardır. Yapraklar ise hedef değişkenin içindeki kategorileri barındırır. Yapraklar son noktalardır ve burdan sonra dal bulunmamaktadır. Karar ağacında kök ve karar düğümleri kendi iç yapılarına göre iki veya daha fazla dala ayrılmaktadır.

Karar ağacı algoritmaları temel olarak aşağıdaki sırayla çalışmaktadır:

- Öğrenme kümesinde hedef değişkeni belirlemede en iyi olduğu düşünülen nitelik köke yerleştirilir.
- Kökte bulunan nitelik kategorik veya sayısal bir değişken olmasının farkları gözetilerek uygun dallara ayrılır.
- Kökten ayrılan her bir dal karar düğümlerine bağlanır. Bu düğümler kökteki niteliğin seçilmesine benzer bir şekilde seçilir.
- Karar düğümleri yapraklar oluşuncaya kadar oluşturulur. Örnekleri bölecek bir nitelik kalmamışsa veya örneklerin hepsi aynı sınıfa aitse işlem sonlandırılır.



Şekil.6. Karar Ağacının Yapısı

Kök ve karar düğümlerine yerleştirilecek olan niteliklerin belirlenmesinde bazı yöntemler kullanılmaktadır. Bunlardan en çok kullanılan ikisi Entropi (Entropy) değeri ve Gini indeksidir. Entropi değeri, enformasyon kazancının (information gain) ölçümünde kullanılır. Rastsallığın, belirsizliğin ve beklenmeyen bir durumun ortaya çıkmasının ölçütüdür. Eldeki bütün veriler tek bir sınıfa ait ise herhangi bir şaşırma olmaz ve entropi sıfır olur. Örneğin, bir veri kümesindeki herkesin aynı basketbol takımının taraftarı olması. Entropi, 0 ile 1 arasında değer alır.

Entropi matematiksel olarak şu şekilde tanımlanabilir:

$$H(D) = \sum p(x_i).\log(1/\ p(x_i)) \qquad (14)$$

Burada, < p_1, p_2, …, p_n> olasılıkları ifade eder ve toplamları 1'e eşittir. D kümesi herhangi bir değişkenin elemanlarını içermektedir. Hedef değişken için ve kazanımı hesaplanacak olan değişken için bu hesaplama yapılır. Kazanımı ifade eden formül aşağıdaki gibidir:

$$\text{Kazanım} = H(D) - \sum p(D_i).H(D_i) \quad (15)$$

Kazanımı en yüksek olan nitelik kök olarak seçilir. Bu işlem, düğüm seçiminde de benzer şekilde kullanılır.

Gini indeksi ise aşağıdaki gibi hesaplanır.

$$\text{Gini}(K) = 1 - \sum p_j^2 \quad (16)$$

Burada, K gini indeksi hesaplanacak olan değişkenin elemanlarının kümesidir. p_j ise, j sınıfının K kümesi içindeki sıklığıdır. K kümesi K_1 ve K_2 alt kümelerine bölünmüş ise aşağıdaki formül geçerlidir:

$$\text{Gini}_{\text{bölünmüş}}(K) = \sum (n_i/n).\text{gini}(K_i) \quad (17)$$

Hedef değişken dışındaki tüm değişkenler için gini indeksi hesaplanır ve en küçük gini değerine sahip olan değişken köke atanır. Benzer işlemler, düğümlerde de yapılır.

Karar ağaçları kural üretmenin yanında parametre değerleri de sunar. Bunlardan en önemlilerinden birisi kaldıraçtır (Lift). Kaldıraç, herhangi bir düğümde bulunan kayıtların, tüm ağaçla kıyaslandığında ne kadarlık bir kısmının hedef sınıfa ait olduğunun göstergesidir. Örneğin, %155 kaldıraç değerine sahip bir düğümde bulunan kayıtların 1,55 kat daha fazla oranla belirlenmiş olan sınıfa aittir [30].

Ağaçtaki aşırı öğrenmenin (overfitting) probleminin önüne geçmek için budama (prunning) işlemi uygulanır. Aşırı öğrenme eldeki verilere uygulandığında iyi sonuçlar verir. Ancak yeni veriler üzerinde daha az hassastır. Budama istatistiksel olarak düşük öneme sahip olan dalların ağaçtan arındırılmasıdır. Ağaçta çok fazla düğüm ve dal oluşursa, yapraklara ulaşan veri sayısı azalır ve ağacın hassasiyeti azalır. Budama

sayesinde gereksiz ayrıntılardan arındırılmış bir ağaç elde edilir ve tahminlerin doğruluk oranı artar. Ağacın kurulumu esnasında veya kurulduktan sonra da yapılabilir.

Karar ağacı tekniğini kullanan çok sayıda algoritma geliştirilmiştir. ID3, C4.5, C5, CART (C&RT), SLIQ, CHAID ve SPRINT algoritmaları bunlardan en popüler olanlarıdır. ID3 algoritması Sydney Üniversitesi'nde J. Ross tarafından geliştirilmiştir. En ayırıcı özelliği bulurken entropi kavramından faydalanır. C4.5 ve C5, ID3 algoritmasının gelişmiş versiyonlarıdır [30].

Tablo 6'da kullanımı yaygın olan karar ağacı algoritmaları ve temel özellikleri verilmiştir.

Tablo.6. Karar Ağacı Algoritmaları ve Özellikleri [36]

KARAR AĞACI ALGORİTMASI	ÖZELLİKLER
C&RT	Budama işlemi, karar ağacının karmaşıklığına göre yapılmaktadır. Regresyonu ve sınıflandırmayı destekleyen bir yapısı vardır. Gini katsayısı kullanılır.
C4.5 ve C5	Düğümlerden çıkan dal sayısı çok olabilir. Tahmin edicinin kategori sayısı dalların sayısına eşittir. Dalların ve düğümlerin belirlenmesi için entropi ve bilgi kazancı kullanılır. Yapraktaki hata oranı, budamanın şeklini belirlemektedir.
CHAID (Chi-Squared Automatic Interaction Detector)	Bölme işlemi, ki-kare (χ^2) testi kullanılarak gerçekleştirilir. Dal adeti iki ile başlar ve tahmin edici değişkenin kategori sayısı kaçsa o kadar değer alabilir.
SLIQ (Supervised Learning in Quest)	Hızlı bir şekilde ölçeklendirilebilir bir algoritmadır. Budamanın da hızlı yapıldığı söylenebilir.
SPRINT (Scalable Parallelizable Induction of Decision Tree)	Veri sayısı çok fazla olan kümeler üzerinde kullanışlıdır.

Karar ağaçlarının kurulması, anlaşılması, yorumlanması ve ağaçtan kural çıkartılması kolaydır. Kategorik değişkenlerin modellemesini basit bir şekilde yapabilir. Sürekli ve kesikli değerlerin ikisi için de modelleme yapabilir. Ancak sürekli değişkenleri tahminlemede çok başarılı değildir.

6.1.1.6. Yapay Sinir Ağları

Yapay sinir ağları (artifical neural networks), insandaki sinir sistemini oluşturan ağlardan esinlenerek geliştirilmiş olan bir bilgi işleme sistemidir. İki sinir sistemi arasındaki kavram benzerlikleri Tablo 7'de verilmiştir. Yapay olarak sinir hücrelerinin farklı pozisyonlarda bir araya gelmesinden oluşur. Yapısında Şekil 7'de gösterilen katmanlar bulunmaktadır.

Yapay sinir ağlarıyla alakalı çalışmalar 1940 lı yıllara kadar dayanmaktadır. 1942'de, McCulloch ve Pitts tarafından hücre modeli ortaya konmuştur. 1949 yılında hücre bağlantılarını ayarlamada kullanmak için ilk öğrenme kuralı Hebb tarafından geliştirilmiştir. 1958'de Rosenblatt tarafından algılama ve öğrenme kuralı geliştirilmiş ve günümüzde kullanılan kuralların temelleri atılmıştır. 1960 ile 1962 yılları arasında Widrow ve Hoff LMS kuralını geliştirdi. 1969'da Minsky ve Papert algılayıcının karmaşık lojik fonksiyonlarda kullanılamayacağı sonucuna vardılar. 1970'ler yapay sinir ağları çalışmalarının aktif olmadığı yıllardı. Daha sonraki yıllarda XOR probleminin de çözülmesiyle bu alandaki çalışmalar tekrar popüler hale gelmiş ve çok sayıda yeni algoritma ve model ortaya konmuştur. 1984 yılında Kohonen kendi kendini düzenleyen haritayı tanımladı ve Kohonen ağlarını geliştirdi. 1986'da Rumelhart tarafından geri yayılım tekrar ortaya çıkartıldı. 1988 yılına gelindiğinde ise Chua ve Yang, yapay hücresel sinir ağlarını geliştirdiler [37].

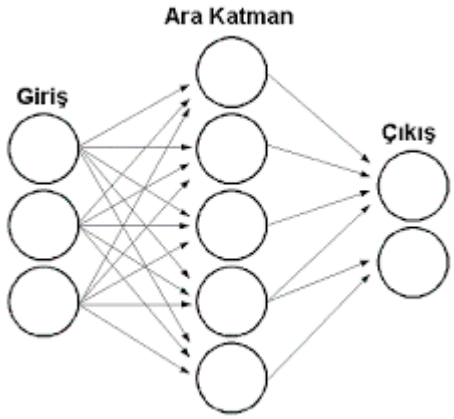
İnsandaki sinir hücresi (nöron), gövde (cell body), çekirdek (nucleus), akson (axon), birçok sinir ucu (dendrite) ve aksonla diğer nöronun sinir ucu arasındaki uzantılardan (synapse) oluşur. Biyolojik sinir hücrelerinin yapay sinir hücreleriyle kavram benzerlikleri aşağıdaki gibidir:

Tablo.7. Biyolojik Sinir Hücrelerinin Yapay Sinir Hücreleriyle Kavram Benzerlikleri

Yapay Sinir Ağı	Sinir Sistemi
İşlem Elemanı	Nöron
Toplama Fonksiyonu	Dentrit
Aktivasyon Fonksiyonu	Hücre Gövdesi
Eleman Çıkışı	Akson
Ağırlıklar	Sinaps

Biyolojik sinir hücreleri arasındaki iletişim sinapslar vasıtasıyla kurulur. Sinir hücresinde işlenen bilgiler aksonlar vasıtasıyla diğer hücrelere gönderilirler. Buna benzer yapıda yapay sinir hücrelerinde de dışarıdan alınan veriler toplama fonksiyonu ile toplanır. Toplanan veriler daha sonra aktivasyon fonksiyonundan geçirilerek çıktıya dönüştürülür ve ağın sahip olduğu bağlantılar üzerinden diğer hücrelere gönderilirler. Çeşitli toplama ve aktivasyon fonksiyonları bulunmaktadır. Yapay sinir hücrelerinin birbirlerine bağlanmasını sağlayan bağlantıların değerlerine ağırlık değerleri denir. Ağı oluşturan birbirlerine paralel şekilde yer alan girdi katmanı, ara katmanlar ve çıktı katmanı bulunmaktadır. Veriler bu ağa girdi katmanından giriş yaparlar. Ara katmanlarda işlendikten sonra çıktı katmanına iletilirler. Veri işleme kavramı, ağa gelen verilerin ağın ağırlık değerleri kullanılarak çıktıya dönüştürülmesine verilen addır. Ağın her bir girdiye doğru çıktılar üretebilmesinin yolu ağırlıkların değerlerinin doğru olmasıyla alakalıdır. Doğru ağırlık değerlerinin bulunması için ağın eğitilmesi gerekmektedir. Ağırlık değerleri rastgele atanır. Eğitim sürecinde değiştirilir. En doğru değerler bulunmaya çalışılır. Bu süreç ağın sahip olduğu eğitim kümesindeki kayıtların tamamı için doğru sonuçlar verinceye kadar tekrarlanır. Doğru çıktılar alındıktan sonra test kümesindeki örnekler ağa gösterildiğinde doğru cevaplar alınırsa ağın eğitilmiş olduğu kabul edilir. Ağırlıkların anlamı tam olarak bilinmediği için yapay sinir ağlarına "kara kutu" adı verilmiştir. Yapay sinir ağlarına gelen girdilere uygulanan işlemlerde ağırlıkların önemi büyüktür. Hatta ağın sahip olduğu zekanın bu ağırlıklarda gizli olduğu söylenebilir. Yapay sinir ağlarına dair çalışmalar başladığı günden beri bir çok model geliştirilmiştir. Bir yapay sinir ağı modelini toplama fonksiyonunun, ağın topolojisinin, aktivasyon fonksiyonunun, öğrenme

stratejisinin ve öğrenme kuralının karakterize ettiği söylenebilir. Sık kullanılan yapay sinir ağları, tek ve çok katmanlı algılayıcılar, LVQ, ART ağları, SOM, Elman ağı gibi ağlardır [38].



Şekil.7. Yapay Sinir Ağlarının Katmanları

Yapay sinir ağlarında (YSA) kullanılan terimler istatistik alanındakilerden isim olarak genellikle farklıdır. Ancak işlevsel olarak benzerlik göstemektedirler. Tablo.8'de terminoloji farklılıkları belirtilmiştir.

Tablo.8. YSA ve İstatistik Terminolojileri [50]

YSA Terminolojisi	İstatistik Terminolojisi
Yapay Sinir Ağı	Model
Ağırlık	Parametre
Girdi	Bağımsız değişken
Çıktı	Tahmin değeri
Hedef	Bağımlı değişken
Hata	Artık
Hata çizgisi	Güven aralığı

YSA, doğrusal değildir ve çözmeye çalıştığı problemdeki değişikliklere göre ağırlık ayarlaması yapar. YSA'nın hataları tolere etme kabiliyeti yüksektir. Hücrelerin bağlanma şekillerine göre geri beslemeli ve ileri beslemeli olmak üzere iki farklı ağ yapısı vardır. İleri beslemeli ağlarda işlemler ileriye doğru devam etmektedir. Yani veri akışı

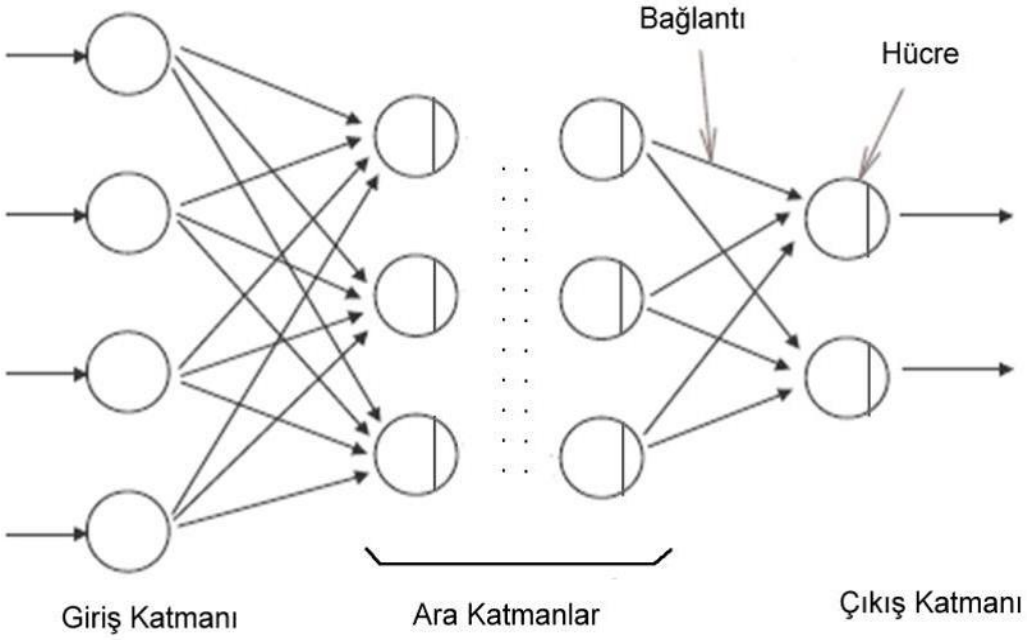
ileri yönlüdür. Geri beslemeli ağlarda ise bir hücre veya daha fazlasının çıkışı kendisine ya da gerideki bir hücreye verilir. Katman içindeki hücreler arasında olabileceği gibi farklı katmanlardan hücreler arasında da olabilir.

Bir yapay sinir ağında en kritik noktalardan birisi gizli katmanında kaç tane nöron olacağına karar vermektir. Sonucu direk olarak vermese de etkilemektedir. Nöron sayısı gerekli olan sayıdan daha az olursa alçak uyum (underfitting) meydana gelir ve bu durum nöronlar arasında taşınan verinin karmaşıklaşmasına neden olur. Diğer taraftan nöron sayısı gerekli olan sayıdan fazla olursa aşırı uyum (overfitting) meydana gelir. Bunlardan dolayı gizli katmanda bulunan her nöronu eğitecek kadar eğitim kümesi bulunmama problemi ortaya çıkabilir. Ayrıca yeterli miktarda eğitim verisi bulunsa bile geç öğrenmeyle karşı karşıya kalınabilir. Dolayısıyla, gizli katmanda bulunan sinir hücrelerinin eğitimi normalden çok daha uzun sürebilir.

Yapay sinir ağları, ortamlara hızlı bir şekilde uyum sağlayabilen, eksik veriyle çalışma kabiliyetine sahip olan, belirsizlik durumlarında karar verme özelliğine sahip olan, meydana gelen hatalara müsamahalı bir şekilde davranan bir tekniktir. Bazı olumsuzluklara da sahiptir. Ağ yapısı ve parametrelerin belirli bir standarda sahip olmaması, problemlerin sayısal (numeric) veriler tarafından gösterilebilmesi, eğitim sürecinin bitiş noktasının bilinmemesi bunlardan bazılarıdır. Bunlara rağmen yapay sinir ağlarına olan ilgi sürekli artmaktadır. Özellikle sınıflandırmada, örüntü tanımada, sinyal filtreleme işleminde, optik karakter taşımada, verilerin sıkıştırılması ve optimizasyonu için yapılan çalışmalarda, veri madenciliğinde, parmak izi tanımada, en iyi rotayı belirlemede, malzeme analizinde ve tıbbi analizlerde sıklıkla kullanılan teknikler arasındadır [38].

Günümüzde en çok kullanılan yapay sinir ağları modeli çok katmanlı algılayıcı (ÇKA) ağlarıdır. XOR probleminin çözümü için yapılan çalışmalar neticesinde ortaya çıkmıştır. Birden fazla ara katman kullanılır ve böylece karmaşık problemlerin çözümü kolaylaştırılır. Veri akışı girdi katmanında başlar ve ara katmanlar üzerinden çıktı katmanına geçiş yapar. Şekil 8'deki gibi bir yapıda olan çok katmanlı algılayıcılarda olayların öğrenilmesini örnek seçimi, girdilerin ve çıktıların ağa sunumu ve sayısal

gösterimi, ağırlıkların atanan başlangıç değerleri, öğrenme oranı ve momentum oranının kaç olacağı, örneklerin ağa sunumu, ağırlıkların değiştirilme anları, girdilerin ve çıktıların ölçeklendirilmesi, algoritmayı durdurmanın kriterlerinin belirlenmesi, ağların budanması ve büyütülmesi gibi faktörler etkilemektedirler [38].



Şekil.8. Çok Katmanlı Algılayıcı (ÇKA) Ağlarının Yapısı

6.1.2. Regresyon Analizi

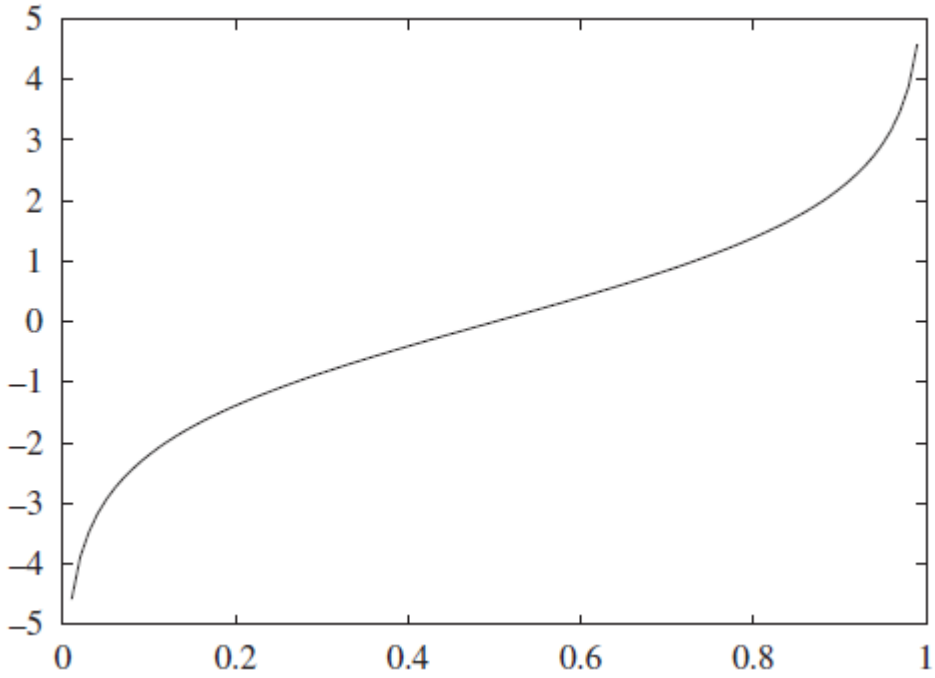
Tahminleyici modellemede, sınıfladırmanın yanında kullanılan diğer bir yöntem regresyondur. Regresyon, hedef değişkenin kategorik olduğu durumlarda da kullanılmasına rağmen daha çok sürekli olduğu durumlarda tahminleme yapmak için kullanılmaktadır. Örneğin skorlama çalışmasında regresyon kullanılabilir. Regresyon, istatistiksel bir metodolojidir. Bağımsız değişkenlerle bağımlı değişken arasında bir denklem kurulur. Regresyonun şekli oluşturulacak denkleme göre değişmektedir. Denklemin doğrusal olması durumunda lineer (doğrusal) regresyon, polinom şeklinde

olması durumunda polinomsal (polynomial) regresyon, logaritmik şeklinde olması durumunda ise lojistik (logistic) regresyon ismi verilmektedir.

6.1.2.1. Lojistik Regresyon (Logistic Regression)

Lojistik regresyon, hem kategorik değişkenlerin hem de sürekli değişkenlerin modellenmesinde kullanılmaktadır. Bu yüzden regresyon analizinin yanında sınıflandırma teknikleri arasında da yer almaktadır.

Lojistik regresyon ismini, bağımlı değişkene uygulanan lojit dönüştürmesinden (logit transformation) almaktadır. Şekil 9'da bu dönüştürme gösterilmiştir.



Şekil.9. Lojit Dönüşümü (Logit Transformation)

Lojistik regresyon (LR) tekniğini uygulayabilmek için, bağımsız değişkenlerle bağımlı değişken arasında doğrusal, üstel veya polinom ilişki olması gerekir. LR ilişki türü ne olursa olsun lojit bir ilişki olduğunu varsayar ve ilişkinin şeklini doğrusallaştıran logaritmik dönüştürmeler yapar.

Lineer regresyonda sapmaların karesinin en az olması için çaba sarfedilirken, lojistik regresyonda ise herhangi bir olayın meydana gelme olasılığını maksimum yapmak amaçlanmaktadır. Lojistik regresyonda bağımsız değişkenlerde normal dağılım gerekli değildir.

LR'da odds ve odds oranı (odds ratio) kavramları vardır. Herhangi bir olayın olma ihtimalinin (p) olmama ihtimaline (1-p) oranı o olayın odds'unu verir.

$$\text{Odds} = p / (1-p) \tag{18}$$

Örneğin bir zar atıldığında 5'in gelme ihtimali 1/6, gelmeme ihtimali 5/6' dır. Dolayısıyla 5 gelme olayının odds'u 1/5 = 0.2' dir. Lojistik regresyonun anahtar kavramı olan lojit, odds değerinin doğal logaritmasıdır. İki tane olayın odds değerlerinin oranına ise odds oranı adı verilmektedir [36].

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k \tag{19}$$

$$\text{logit}(p) = \ln (p / (1 - p)) \tag{20}$$

LR, bağımsız değişkenler arasındaki yüksek korelasyona oldukça duyarlıdır. Eğer böyle bir durum söz konusu ise bağımsız değişkenler arasında çoklu bağlanım (multicollinearity) var demektir. Bu problemin önüne geçebilmek için yüksek korele değişkenlerin uygun yöntemlerle çıkarılarak sadece düşük korele bağımsız değişkenlerin modele dahil edilmesi gerekmektedir. Aksi taktirde modellemedeki başarıyı çok iyi seviyelere getirebilecek olan değişkenlerin tahminleme gücü azalmış olur.

LR tekniği, standart ve aşamalı (stepwise) şeklinde iki farklı temel yöntemle uygulanabilmektedir. Standart yöntemde tüm ortak nitelikler (covariates) modelde bulunur ve bloklar için parametre kestirimi yapılır. Aşamalı yöntemler ise ileriye doğru (forward) ve geriye doğru (backward) olmak üzere iki farklı şekilde uygulanır [40].

İleriye doğru yönteminde analize sabit terim dâhil edilerek başlanır. Sonra puan istatistiklerine (score statistics) göre, modele değişken eklenir. En önemli istatistiğine sahip olan değişkenden anlamlı en son istatistiğe sahip olan değişkene kadar modele değişken alımı devam eder.İşlemin kesme noktası 0.05'tir (α). Her bir adımda dışarıda bırakılacak olan değişken olup olmadığı araştırılır. Bunu gerçekleştirmenin üç yolu

bulunmaktadır. Bunlardan ilki olabilirlik oran istatistiğidir (likelihood ratio statistics). Mevcut model ile modelde bulunma öneminin testi yapılan bağımsız değişkenin dışarıda bırakıldığı durumda kurulan model karşılaştırılır. Bu değişkenin dışarıda bırakılması mevcut modele göre anlamlı bir şekilde daha kötü sonuç veriyorsa bu değişken modelin uyumunu iyileştirdiği için modelde tutulması gerekir. İkinci yol ise durum indekstir (Condition Index). Durum indeksi olabilirlik oran istatistiğine göre daha az duyarlıdır ve bu yüzden kullanımı tavsiye edilmez. Son yol ise Wald istatistiğidir. Wald istatistiği, tüm bağımsız (independent) değişkenler için lojistik regresyon katsayısının anlamlılığını test eder. Diğer bir ifadeyle, herhangi bir lojit katsayısı için null hipotezini test eder. Wald istatistiği, standardize olmamış olan bir lojistik regresyon katsayısının, kendi standart hatasına oranının karesidir. Ayrıca, en çok olabilirlik kestirimlerinin (maximum likelihood estimations) normal dağıldığını varsayar. Bu üç ölçüt arasından en iyisi olarak olabilirlik oran istatistiği görülmektedir[40].

Geriye doğru yönteminde ise analize tüm bağımsız değişkenler dâhil edilerek başlanır. Daha sonra modele katkı sağlayıp sağlamadıklarına bakılarak elenirler veya modelde tutulurlar. En az katkı sağlayan ilk önce uzaklaştırılır prensibine dayanmaktadır. Geriye doğru yöntemler, ileriye doğru yöntemlere göre daha çok tercih edilmektedir. Bunun en temel sebebi baskılama etkisidir. Baskılama etkisi, bağımsız bir niteliğin etkisinin sabit olması durumunda, başka bir niteliğin anlamlı ve önemli bir etkiye sahip olması durumudur. İleriye doğru yönteminin, geriye doğru yöntemine kıyasla anlamlı bağımsız değişkenleri eleme olasılığı daha yüksektir.

4.2. Tanımlayıcı (Descriptive) Modelleme

Tanımlayıcı modellemede veriyi keşfetmek, veri üzerinde tespit yapmak, problemi algılamak, profilleme yapmak amaçlanmaktadır. Gelecekteki bir olayı tahmin etmek bu modelleme çeşidinin alanına girmez. Kümeleme, anomali tespiti (anomaly detection), birliktelik kuralları ve ilişki analizi ve temel bileşen analizi (principal component analysis) tanımlayıcı modelleme içerisinde yer alır.

6.2.1. Kümeleme (Clustering)

Kümeleme tekniği, grup veya küme sayısı belirtilmemiş olan veri kümesindeki desenleri ve benzerlikleri keşfederek verileri ayrık kümelerde toplamayı amaçlamaktadır. Birbirine benzeyen verilerin yaklaşması ve benzemeyenlerin uzaklaşması sağlanarak kümeleme yapılmaktadır. Makine öğrenmesi tekniklerinden gözetimsiz öğrenme (unsupervised learning) yöntemi kullanılmaktadır. Hiyerarşik ve hiyerarşik olmayan olmak üzere iki farklı kümeleme tekniği bulunmaktadır. Hiyerarşik kümeleme tekniklerinde, kümeler peşpeşe birleştirilir ve bir grup, diğeri ile birleştikten sonra, daha sonraki adımlarda ayrılamaz. Toplama ve ayırma yöntemleri kullanılır. Hiyerarşik olmayan kümeleme tekniklerinde ise verilerin belirli sayıda kümeye atanması söz konusudur. Bir veri iterasyondan sonra yeni bir kümenin elemanı olabilir. bu tekniğin en çok kullanılan algoritmaları k-means (k-ortalama) ve en çok olabilirliktir [41].

K-means kümeleme algoritmasında izlenen yol şöyle sıralanabilir:

- k (küme sayısı) belirlenir.
- Verilerin arasından rastgele k tanesi seçilir.
- Kullanılan mesafe ölçüsüne göre (genellikle Öklid mesafesi kullanılır) diğer veriler bu k tane veriden hangisine en yakınsa onun grubuna atanır. Bu şekilde boşta veri kalmayacak şekilde k tane küme ortaya çıkar.
- Oluşan k tane kümenin ortalaması veya merkez elemanı hesaplanır. Yeni değerler baz alınarak aynı işlem tekrar uygulanır ve yine k tane küme elde edilir.
- Bu iteratif işlem merkez değeri ve dolayısıyla kümelerde değişiklik olmayıncaya kadar devam eder.

K-means algoritmasındaki temel amaç birbirine benzer özellikler taşıyan verileri bir kümeye toplamaktır. Karakteristik bakımdan uzak olan verileri ise farklı bir kümeye atamaktır. K sayısının anlamlı olması önemlidir. Kümeleme yapılacak olan verilerin özelliklerine göre değişmektedir.

6.2.2. Birliktelik Kuralı (Association Rule)

Veri tabanlarında toplanan veri arttıkça, verilerin arasındaki ilişkilerin ortaya çıkarılması da değerli hale gelmiştir. Birliktelik kuralı (association rule) bu ihtiyacı karşılamak üzere ortaya çıkmıştır. Geçmiş verilerin arasındaki birliktelik davranışlarının tespiti sonucunda geleceğe yönelik çalışmalar yapılmasını destekler. Birliktelik kuralının matematiksel modeli 1993'te Agrawal, Imielinski ve Swami tarafından oluşturulmuştur [42].

Birliktelik kuralının tanımı söyledir:

$$A_1, A_2, ..., A_k \Rightarrow B_1, B_2, ..., B_n \quad (21)$$

Ai ve Bj, yapılan iş veya nesneleri ifade eder. "A_1,A_2,...,A_k" nesnelerinin ortaya çıktığı durumda, "B_1,B_2, ...,B_n" nesnelerinin genellikle aynı aksiyon içinde yer aldıkları anlamına gelmektedir. Birliktelik kuralı, alabileceği en küçük değeri belirlenmiş olan destek ve güvenirlik değerlerini sağlamalıdır [42].

Birliktelik kuralına ilişkin ilk algoritma AIS algoritmasıdır. En çok kullanılan ise Apriori algoritmasıdır. Minimum destek ve güven aralığı kavramları kullanılır. Pazar sepet analizi yapmak için sıklıkla kullanılır. Örneğin, Apriori algoritmasının çalışması sonucunda şu cümle kurulabilir:

%33 destek ve %60 güven seviyesinde, elma alan kişiler muz ve simit de alır [30].

4.3. Buyrukçu (Prescriptive) Modelleme

Buyrukçu veya öngörülü modelleme, karar vermeden önce muhtemel sonuçları bildirmek için gelecekteki kararların etkisini ölçmeye çalışır. Model geliştirmek için makine öğrenimindeki tavsiyeci (recommender) algoritmalar kullanılabilir. Öngörülü modelleme, iş kuralları, makine öğrenimi ve hesaplama modellemesi prosedürleri gibi tekniklerin ve araçların bir kombinasyonunu kullanır. Reklamcılık ve pazarlamada

sıklıkla kullanılır. Pazarlama optimizasyonu yapmak, müşterinin özelliklerine tavsiyede bulunmak, ürün müşteri eşleştirmesi yapmak, doğru ürünü doğru zamanda tedarik etmek gibi önemli işlerde bu modellemeden yararlanılır. Örneğin, bir internet sitesine girildiğinde, daha önceki girişlerde aranan ürünlere benzer olanların görülmesi, online kitap satış sitesinde kitap satın alan kişiye satın aldığı kitabı satın alanların başka hangi kitapları satın aldıklarının gösterilmesi, öngörülü modellemenin çalışma alanına girmektedir.

7. KULLANILAN TEKNOLOJİLER

Veri ön işlemesinden model kurma, karşılaştırma ve seçmeye kadar tüm süreçler Weka aracı (tool) ve SAS kuruluşuna ait olan araçlar kullanılarak gerçekleştirilmiştir.

Weka (Waikato Environment for Knowledge Analysis) aracı, Yeni Zelanda'da bulunan Waikato Üniversitesi'nde yazılmıştır. 1993 yılında C diliyle yazılmaya başlanmıştır. 1997 yılında ise Java diliyle tekrar geliştirilmiş ve sonraki versiyonları da Java ile yazılmıştır. Açık kaynak kodludur. Açıldığında Şekil 10'daki gibi bir arayüzle karşılaşılır. Makine öğrenmesi yazılımıdır ve birçok istatistik yönteminin de uygulanabileceği bir platformdur. Dolayısıyla veri madenciliği çalışmaları için çok kullanışlıdır [46].

Şekil.10. Weka Aracının Açılış Arayüzü

Weka'da ARFF (Attribute-Relation File Format) dosyaları kullanılır. ARFF, değişkenlerin istenen formatta tanımlanmasını sağlayan ASCII metin dosyasıdır. Weka üzerinde çalışılacak bir veri kümesinin ARFF dosya formatında Şekil 11'de gösterildiği gibi üç temel kavram vardır. Bunlardan ilki başlık kısmının yazıldığı @RELATION, ikincisi değişkenlerin girildiği @ATTRIBUTE ve üçüncüsü verilerin girildiği @DATA kısmıdır. Veriler girilirken aralarına virgül koyulur. Değişkenleri yazarken @ATTRIBUTE yazdıktan sonra bir boşluk bırakarak değişken türü ve tekrar boşluk bırakarak değişken ismi yazılır. Değişken ve başlık isimlerinde büyük küçük harf

duyarlılığı vardır ve boşluk bırakmadan yazılmak zorundadır. Zorunluluklar yerine getirilmezse Weka'ya yükleme yapılamaz. Yorum yazmak istendiğinde satırın başına "%" (yüzde) işareti koyulur [47].

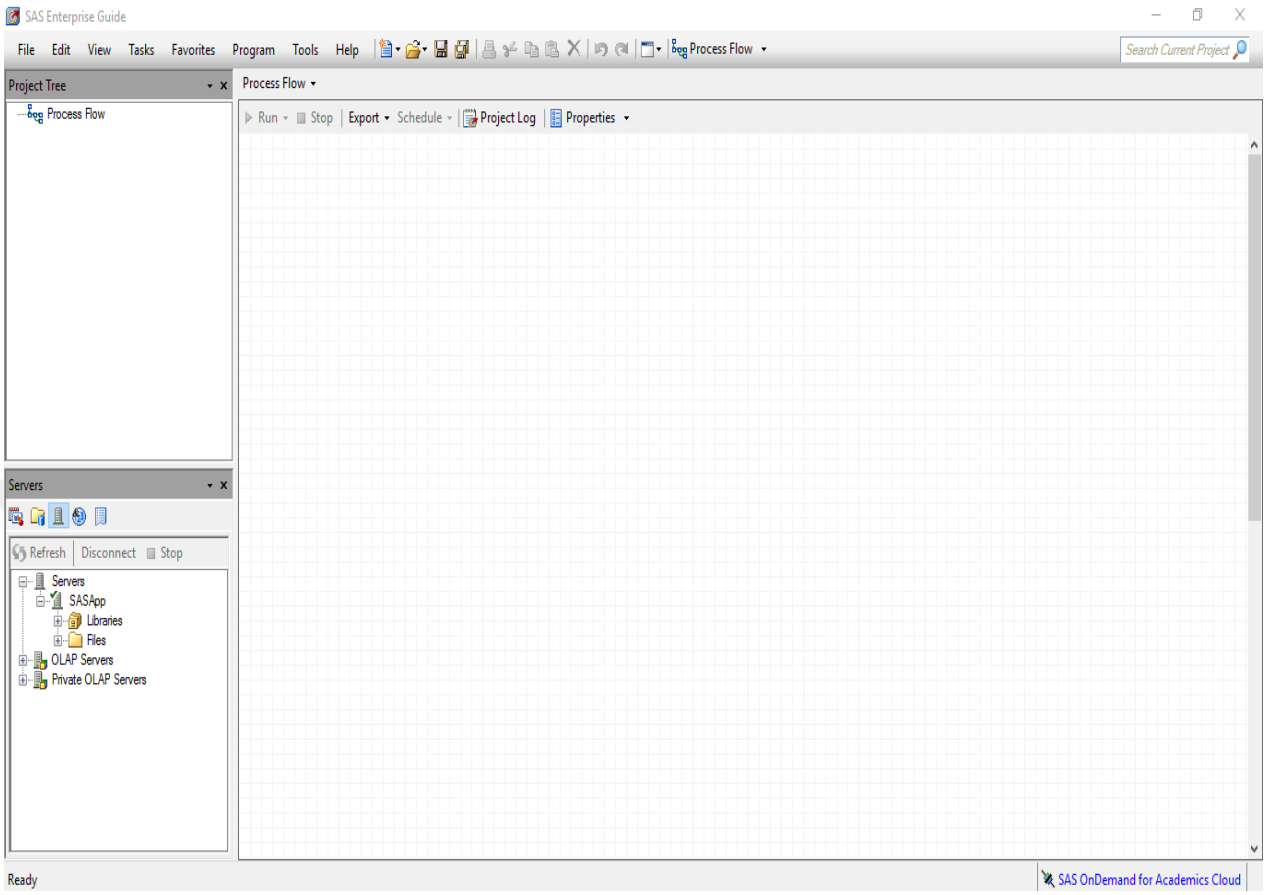
```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
```

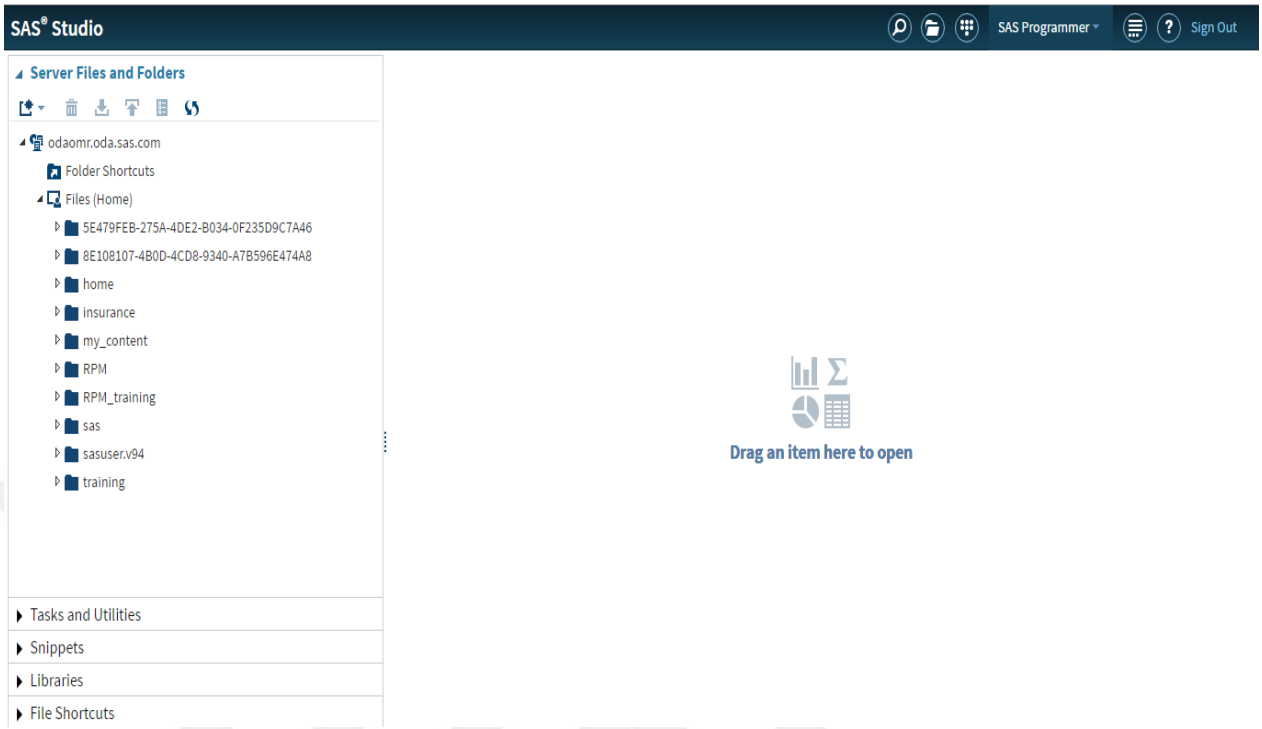
Şekil.11. ARFF Dosya Yapısı

Diğer kullanılan araçlar SAS (Statistical Analysis System) kuruluşuna aittir. ABD'de bulunan North Carolina Devlet Üniversitesi'nde yapılan çalışmalar sonucunda 1976 yılında kurulmuştur. Veri Analitiği alanının dünyadaki ilk kuruluşlarından biridir. 139 ülkede faaliyet gösteren SAS'ın birçok alanla ilgili aracı mevcuttur [48]. Bu çalışmada SAS'ın akademik çalışmalar için kullanım hakkı tanıdığı SAS OnDemand for Academics araçları üzerinde işlem yapılmıştır. Şekil 12'de arayüzü gösterilen SAS Enterprise Guide aracının masaüstü kurulumu yapılmış ve web tabanlı çalışan SAS Studio aracı kullanılmıştır. SAS Studio aracının arayüzü Şekil 13'te gösterilmiştir.



Şekil.12. SAS Enterprise Guide Arayüzü

Enterprise Guide aracı, birçok istatistik sonucunun görülebileceği menülere sahiptir. Servers kısmında verilerin tutulduğu kütüphane ve tablo isimleri vardır. Project Tree kısmında sunucuda yer alan veya SAS kodlarıyla oluşturulan tablolar, tablolara uygulanan işlemler, yazılacak kodların tutulduğu program dosyaları bulunur. Process flow kısmında ise tablolardan elde edilen istatistikler ve yeni tabloların akışı görülebilir.



Şekil.13. SAS Studio Arayüzü

8. UYGULAMA

Uygulamada kullanılmış olan sigortacılık verisi, ABD'de bulunan University of California, Irvine (UCI) tarafından biriktirilen, erişime açık ve akademik amaçlı kullanılmak zorunda olan veri kümeleri arasından elde edilmiştir. Veriyi bu üniversite veri tabanına aktaran grup, Sentient Machine Research (SMR) isimli Amsterdam merkezli bir gruptur. Veriler Hollanda merkezli bir sigorta şirketine aittir. Kullanılan değişkenler Hollanda sigorta yönetmeliklerine uygun seçilmiştir. Bu veri kümesinde müşterilere ait veriler bulunmakta ve karavan sigortası yaptırıp yapırmadıkları belirtilmektedir. Veri kümesi kullanılarak hangi özellikteki müşterilerin karavan sigortası yaptırabileceğinin belirlenmesi amaçlanmıştır.

Veri kümesi, CoIL 2000 Challenge ödüllü yarışmasında kullanılmak üzere toplanmıştır. CoIL (Computational Intelligence and Learning Cluster), Avrupa Birliği tarafından fonlanan 4 farklı oluşumun ortak bir iş birliğidir. Bu oluşumlarda çalışılan alanlar yapay sinir ağları, bulanık mantık, evrimsel hesaplama ve makine öğrenimidir [51]. Yarışma için yapılan çalışmalarda tahminleme alanında birinci olan çalışmada Naive Bayes algoritmasıyla çalışılmıştır. Çalışmada en iyi tahmin ediciler olarak müşterinin bir araba sahibi olması, satın alma gücü, özel üçüncü taraf sigortası olması, bot poliçesine sahip olması, sosyal güvenlik sigorta poliçesi ve yangın poliçesi için ödenen prim değişkenleri belirlenmiştir [52]. Tanımlayıcı modelleme alanında birinci olan çalışmada ise evrimsel yerel arama algoritması (evolutionary local search algorithms), ki-kare (chi-square) testi ve birliktelik kuralları kullanılmıştır [53]. Yapay sinir ağlarının, genetik algoritmaların, bulanık mantığın, kümelemenin (clustering), regresyon ağacının ve destek vektör makinesinin (DVM) kullanıldığı çalışmalar yapılmıştır. Bu çalışmada ise, tahminleyici modelleme kullanılmıştır. Diğer çalışmalardan farklı olarak karar ağaçları tekniğini kullanan C 4.5 (J48) algoritması, lojistik regresyon kullanılmıştır. Ayrıca yapay sinir ağları tekniğinin kullanıldığı çok katmanlı algılayıcı yapısındaki MultilayerPerceptron algoritması kullanılmıştır. Bu yönleriyle özgün bir çalışmadır.

8.1. Veri Kümesi ve Özellikleri

Veri kümesi, öğrenme (train) ve validation (doğrulama) için bir küme ve test için bir küme olmak üzere iki parça halinde alınmıştır. Veri kümelerinin öğrenme ve test kümelerine ayrılması motokaravan sigortasını yaptıran kişilerin oranı değişmeyecek şekilde yaklaşık %60'a %40 şeklinde yapılmıştır. Öğrenme veri kümesi 5822 ve test veri kümesi de 4000 kayıttan oluşmaktadır. İki kümede de bu sigortayı yaptıranların oranı %5,95'tir. Kümeler, 85 bağımsız ve 1 bağımlı değişken olmak üzere toplam 86 değişkenden oluşmaktadır. İlk 43 değişken müşterilerin sosyodemografik özellikleriyle ilgilidir. Posta kodu esas alınmıştır. Aynı posta kodunda yaşayan kişilerin aynı sosyodemografik özelliklere sahip olduğu varsayılmıştır. Bu özellikler için sıralı değişken kullanılmıştır ve 0 ile 9 arasındaki sayılar kullanılmıştır ve bu değerlerin her birisi belli bir yüzdeyi veya yüzde aralığını temsil eder.

0 : %0

1 : %1 - %10

2 : %11 - %23

3 : %24 - %36

4 : %37 - %49

5 : %50 - %62

6 : %63 - %75

7 : %76 - %88

8 : %89 - %99

9 : %100

Örneğin, bir değişkenin değeri 3 ise müşterinin yaşadığı yerdeki (aynı posta kodu) kişilerin %24 ile %36 arası kadarı bu özelliği taşıyor demektir. Diğer 42 değişken ürün sahipliği ile ilgilidir. Hedef değişken ise müşterinin karavan (mobile home) sigortası yaptırıp yaptırmamasıdır. Bazı değişken bilgileri aşağıda verilmiştir:

1. Customer_type : Müşteri tipini bildirir. 10 farklı kategori mevcuttur. Yaşam tarzına ait özelliklerdir.
2. Number of houses : Müşterinin sahip olduğu ev sayısını belirtir.
3. Married : Müşterinin yaşadığı yerdeki evli oranını belirtir.
4. High level education : Müşterinin yaşadığı yerdeki yüksek seviye eğitime sahip kişi oranını belirtir.
5. High status : Müşterinin yaşadığı yerdeki yüksek statü sahibi kişilerin oranını belirtir.
6. Entrepreneur : Müşterinin yaşadığı yerdeki girişimci oranını belirtir.
7. Farmer : Müşterinin yaşadığı yerdeki çiftçi oranını belirtir.
8. Skilled labourers : Müşterinin yaşadığı yerdeki vasıflı işçi oranını belirtir.
9. Rented house : Müşterinin yaşadığı yerdeki kirada yaşayan insanların oranını belirtir.
10. Home owners : Müşterinin yaşadığı yerdeki ev sahibi olan insanların oranını belirtir.
11. One Car : Müşterinin yaşadığı yerdeki bir arabası olan insanların oranını belirtir.
12. Private health insurance : Müşterinin yaşadığı yerdeki özel sağlı sigortası yaptıran insanların oranını belirtir.
13. Contribution fire policies : Müşterinin yangın poliçesi yaptırmak için ödediği prim miktarıdır.
14. Number of car policies : Müşterinin araba poliçelerinin sayısını belirtir.
15. Number of life insurances : Müşterinin hayat poliçelerinin sayısını belirtir.
16. Mobile home policies : Müşterinin karavan sigortası poliçelerinin sayısını belirtir.

Öğrenme veri kümesinde bulunan 5822 müşterinin sadece 348'i karavan sigortası yaptırmıştır. Yani hedef değişkenin yaklaşık %6'sı 1 değerini taşımaktadır. Dolayısıyla %94'ü 0 değerini taşımaktadır. Bundan dolayı öğrenme kümesi dengesiz (imbalanced) bir veri kümesidir. Model sonucunu değerlendirirken sadece doğruluk kriterine bakılsaydı tüm değerlere 0 atanır ve zaman maliyetinden kurtulmuş olunurdu. Çünkü bu durumda model tüm değerlere 0 atayacak ve yaklaşık %94 doğrulukla çalışacaktı. Ancak buradaki asıl amaç 1'lerin doğru tahmin edilmesi olduğu için başka yöntemler geliştirmek gerekmektedir. Dengesiz veri kümesi sık karşılaşılan problemlerden birisidir. Örneğin,

kanser olma riskinin modellendiği bir veri kümesinde kanser olma sayısının olmama sayısına göre çok düşük olmasından dolayı benzer bir problem söz konusudur. Bu problemin çözümü için belli yöntemler vardır. Bunlar şu şekilde sıralanabilir:

- Daha fazla veri toplamak
- Performans metriklerini değiştirmek
- 0 ve 1 gelen durumların sayısını eşitlemek
- 1 gelen durumların sayısını arttırmak (Synthetic data)
- Algoritma değiştirmek
- Aykırılık analizi, anomali tespiti gibi farklı varyasyonalarla yeni bakış açıları kazanmak

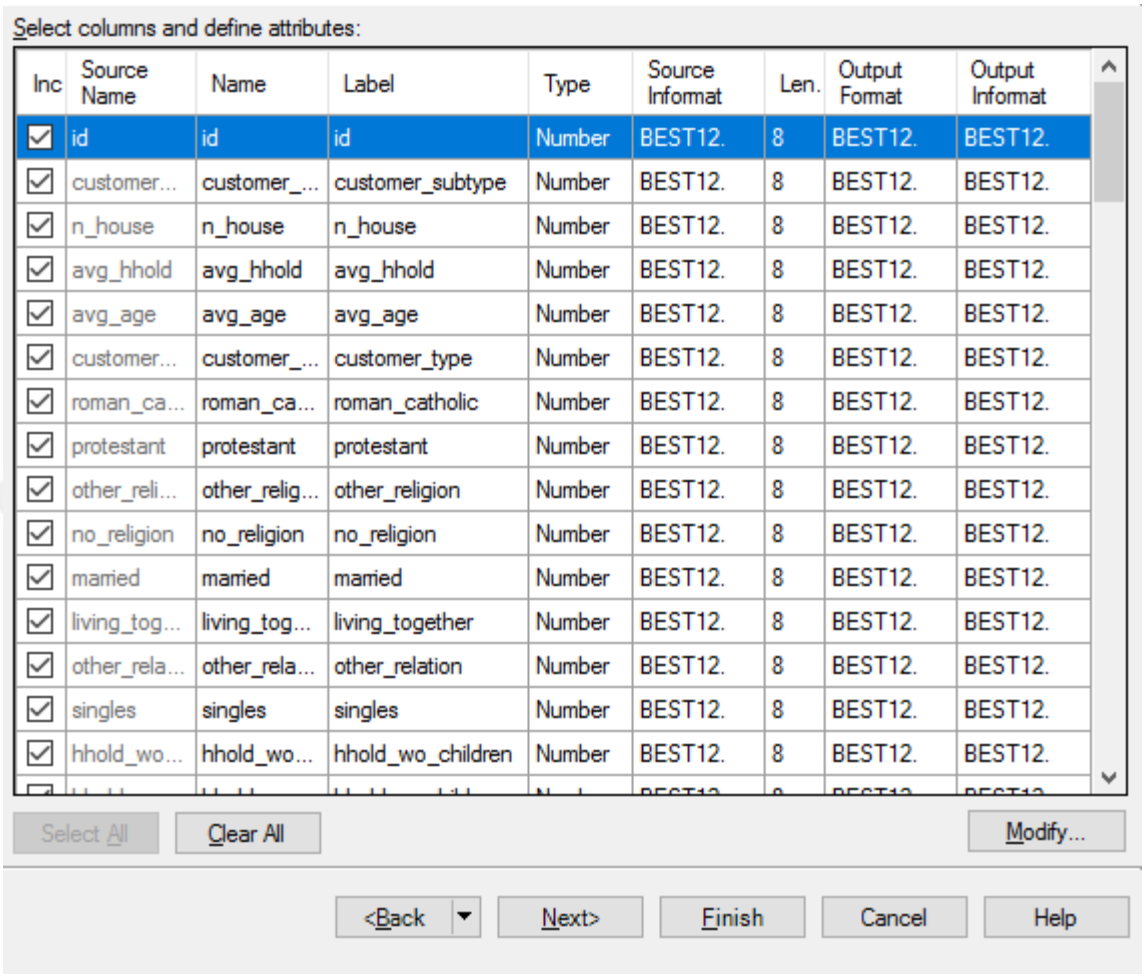
Veri ön işleme ve sonrasında bu yöntemlerden birkaç tanesine başvurularak problem aşılmaya çalışılmıştır.

8.2. Veri Ön İşleme Süreçlerinin Uygulanması

Veri ön işlemesi aşamasında SAS Enterprise Guide ve SAS Studio araçları kullanılmıştır. 5. bölümde anlatılan süreçler burada işletilmiş ve Semma metodolojisinin bir aşaması olan ön işleme kriterleri uygulanmıştır.

8.2.1. Verilerin SAS Enterprise Guide Platformuna Aktarılması

Veri kümesi, notepad formatından değişken isimleri de eklenerek Microsoft Office Excel ortamına aktarılmıştır ve dosyaya “training_sas” adı verilmiştir. Daha sonra SAS Enterprise Guide (EG) aracında bulunan ve Şekil 14’te gösterilen dışarıdan veri kümesi alma (import data) özelliği kullanılarak bu ortama aktarılmıştır.



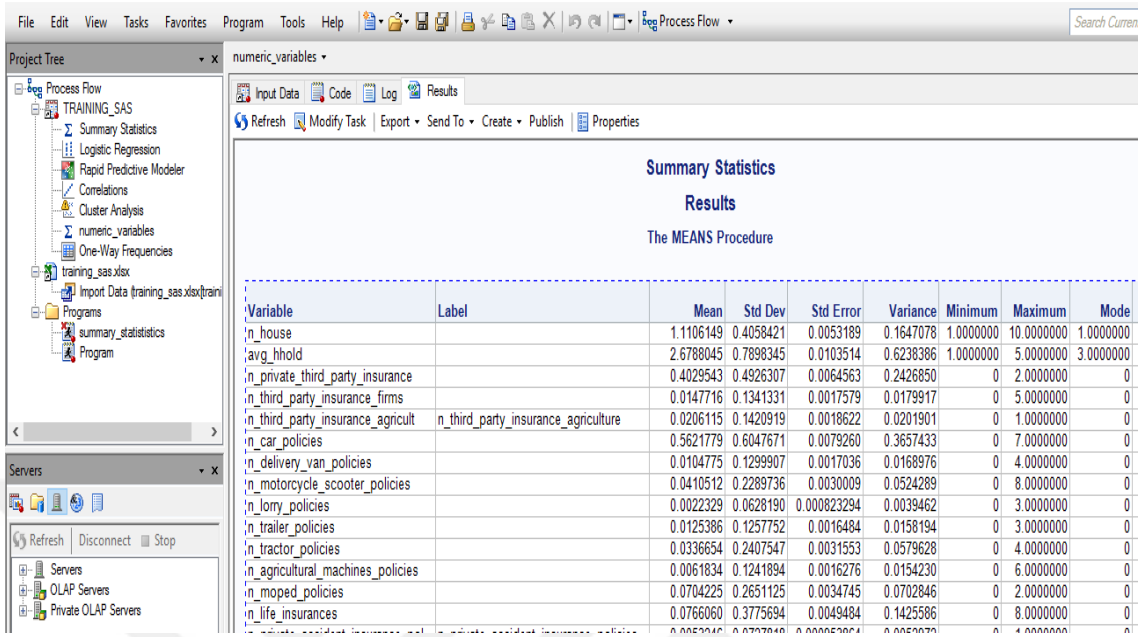
Şekil.14. Veri Kümesinin Excel'den EG'a Aktarılması için Kullanılan Arayüz

Aktarım sırasında verinin formatını EG kendisi otomatik ayarlamaktadır. Eğer formatta değişiklik yapmak istenirse "Modify..." butonuna tıklanarak yapılabilir. Herhangi bir değişiklik yapılmazsa EG excelden aldığı tabloyu "WORK" kütüphanesine atar. Burası, geçici oluşturulan tabloların bulunduğu yerdir ve program kapatıldıktan sonra otomatik olarak silinir. Bunun yerine tablonun kalıcı bir kütüphaneye alınması maliyet açısından daha faydalıdır. Böylece, EG platformuna alınan tablo, SAS'ın sağladığı Cloud veri tabanı sistemi sayesinde SAS Studio aracına giriş yapıldığında da kullanılabilir.

8.2.2. İstatistiki Göstergelerle Veri Dönüştürme İşlemleri

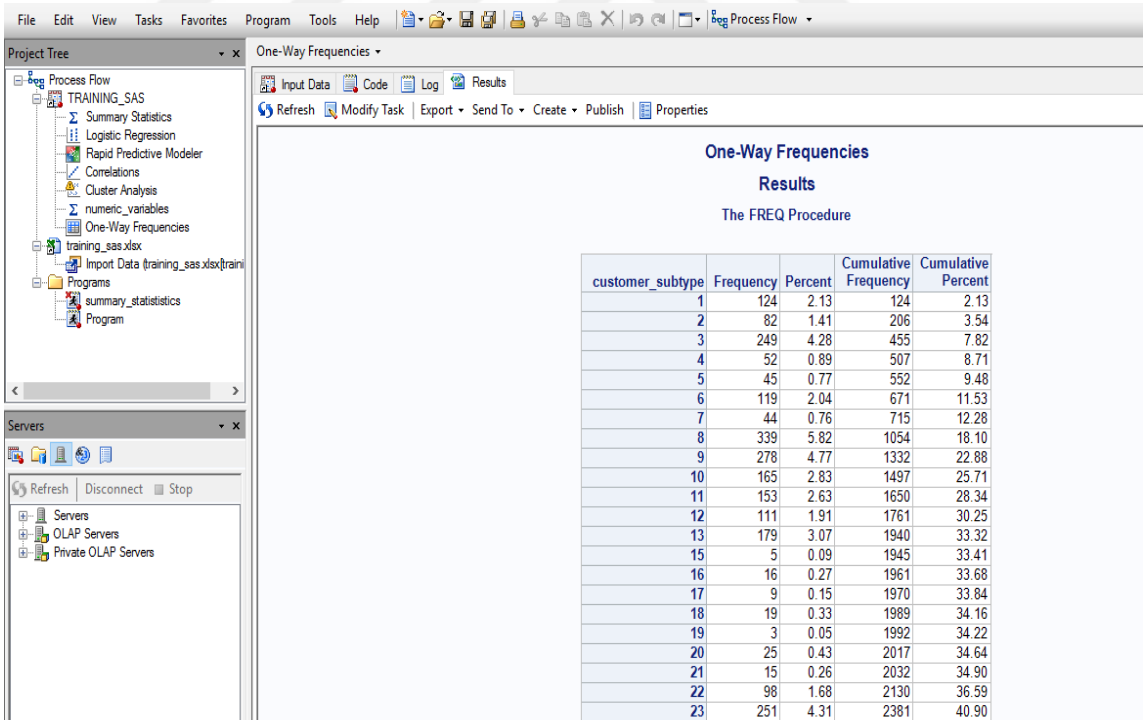
Tablonun SAS ortamına alınmasından sonra veri ön işleme çalışmalarına başlanmıştır. Tabloda boş (null) değer bulunmamaktadır. Hedef değişken (mobile_home_insurance) ikili (binary) bir değişkendir ve {0,1} değerlerini almaktadır. Diğer değişkenlerin 23 tanesi sayısal (numeric), 2 tanesi nominal, 60 tanesi sıralı (ordinal) değişkendir.

Bir sonraki basamak olarak 5. bölümde anlatılan veri ön işleme süreci kapsamındaki diğer aşama olan istatistiki değerlerin tespiti aşamasına geçilmiştir. Değişkenlerin özelliklerini yorumlayabilmek için bazı istatistiki değerlerin bilinmesi gerekmektedir. EG'ta sürekli değişkenler için "Summary Statistics" ve kategorik değişkenler için "One Way Frequency" uygulamaları kullanılarak bu değerlere ulaşılabilir. Şekil 15'te genel görünümü verilen Summary Statistics uygulamasında sürekli değişkenler için ortalama (mean), standart sapma, standart hata, varyans, minimum ve maksimum değerler, mod, genişlik (range), toplam değer (sum), boş olan değer sayısı, boş olmayan değer sayısı, yüzde 1 değeri, yüzde 5 değeri, yüzde 10 değeri, yüzde 25 değeri (lower quartile), medyan, yüzde 75 değeri (lower quartile), yüzde 90 değeri, yüzde 95 değeri, yüzde 99 değeri gibi göstergeler mevcuttur. Şekil 16'daki görünüme sahip olan One Way Frequency uygulamasında ise kategorik değişkenler için frekans değeri, frekans oranı, kümülatif (cumulative) frekans değeri ve kümülatif frekans oranı gibi özellikler vardır.



Variable	Label	Mean	Std Dev	Std Error	Variance	Minimum	Maximum	Mode
n_house		1.1106149	0.4058421	0.0053189	0.1647078	1.0000000	10.0000000	1.0000000
avg_hhold		2.6788045	0.7898345	0.0103514	0.6238386	1.0000000	5.0000000	3.0000000
n_private_third_party_insurance		0.4029543	0.4926307	0.0064563	0.2426850	0	2.0000000	0
n_third_party_insurance_firms		0.0147716	0.1341331	0.0017579	0.0179917	0	5.0000000	0
n_third_party_insurance_agricult	n_third_party_insurance_agriculture	0.0206115	0.1420919	0.0018622	0.0201901	0	1.0000000	0
n_car_policies		0.5621779	0.6047671	0.0079260	0.3657433	0	7.0000000	0
n_delivery_van_policies		0.0104775	0.1299907	0.0017036	0.0168976	0	4.0000000	0
n_motorcycle_scooter_policies		0.0410512	0.2289736	0.0030009	0.0524289	0	8.0000000	0
n_lorry_policies		0.0022329	0.0628190	0.00082329	0.0039462	0	3.0000000	0
n_trailer_policies		0.0125386	0.1257752	0.0016484	0.0158194	0	3.0000000	0
n_tractor_policies		0.0336654	0.2407547	0.0031553	0.0579628	0	4.0000000	0
n_agricultural_machines_policies		0.0061834	0.1241894	0.0016276	0.0154230	0	6.0000000	0
n_moped_policies		0.0704225	0.2651125	0.0034745	0.0702846	0	2.0000000	0
n_life_insurances		0.0766060	0.3775694	0.0049484	0.1425586	0	8.0000000	0

Şekil.15. Summary Statistics Menüsünün Genel Görünümü



customer_subtype	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	124	2.13	124	2.13
2	82	1.41	206	3.54
3	249	4.28	455	7.82
4	52	0.89	507	8.71
5	45	0.77	552	9.48
6	119	2.04	671	11.53
7	44	0.76	715	12.28
8	339	5.82	1054	18.10
9	278	4.77	1332	22.88
10	165	2.83	1497	25.71
11	153	2.63	1650	28.34
12	111	1.91	1761	30.25
13	179	3.07	1940	33.32
15	5	0.09	1945	33.41
16	16	0.27	1961	33.68
17	9	0.15	1970	33.84
18	19	0.33	1989	34.16
19	3	0.05	1992	34.22
20	25	0.43	2017	34.64
21	15	0.26	2032	34.90
22	98	1.68	2130	36.59
23	251	4.31	2381	40.90

Şekil.16. One Way Frequency Menüsünün Genel Görünümü

Sürekli değişkenler için olan göstergelere bakıldığında bazı değişkenlerin çok az sayıda 0'dan farklı değeri vardır. Yüzde 99 değeri dahi 0'dır. Bu değişkenleri birleştirme yoluna gitmek veya korelasyon analizine tabi tutulmadan önce elemek gerekmektedir.

Sonuçlara bakıldığında “n_delivery_van_policies”, “n_private_accident_insurance_pol”, “n_surfboard_policies”, “n_family_accidents_insurance_pol”, “n_agricultural_machines_policies”, “n_lorry_policies”, “n_boat_policies”, “n_disability_insurance_policies”, “n_third_party_insurance_firms” ve “n_social_security_insurance_poli” değişkenleri bu sebepten dolayı modelleme çalışmasından çıkarılmıştır. Benzerlikler ve iş bilgisi ile düşünüldüğünde “n_third_party_insurance_agricult”, “n_trailer_policies” ve “n_tractor_policies” değişkenleri birleştirilmiş ve herhangi birinin değeri 0’dan farklıysa 1, değilse 0 alacak şekilde bir gölge (dummy) değişken oluşturulmuştur. Bu dönüştürme, sas kodunda bulunan “case when” yöntemiyle gerçekleştirilmiştir. Aynı işlem “n_moped_policies”, “n_motorcycle_scooter_policies” ve “n_bicycle_policies” değişkenleri için de uygulanmıştır.

Modellenecek olan veri kümesinin en büyük zorluklarından birisi, müşterilerin demografik özellikleri yerine yaşadığı bölgenin sosyodemografik özellikleri verilmiş olmasıdır. Bu durum birden fazla değişkeninin tek bir değişken altında birleştirilmesini engellemiştir. Örneğin, gelir durumuyla alakalı olan “Income_lt30”, “Income_btw_30and45” ve “Income_btw_45and75” değişkenleri bir tane değişkenin altında toplanabilirdi. Ancak verinin karakteristiğinden dolayı yapılamamaktadır. Bu değişkenlerin yüksek korele çıkması olasıdır. Korelasyon analizinde daha iyi yorum yapılabilecektir.

Kategorik değişkenlerle ilgili frekans analizine bakıldığında ve EG aracının yardımıyla yapılan analizin sonucunda şöyle bulgular elde edilmiştir:

- “Customer_subtype” değişkeninin aldığı değerlere bakıldığında en fazla motokaravan sahibi olan grubun orta sınıf aileler (middle class families) oldukları görülmektedir. 249 müşterinin 51 tanesi motokaravan sigortasına sahiptir. SAS koduyla bu enformasyon elde edilmiştir.
- “Customer_type” değişkenine bakıldığında ise orta büyüklükte ve kalabalık ailelerin motokaravan sigortası yaptırdığı görülmektedir. Yüksek kalite hayat yaşayanlar ve çiftçilik yapanlar arasında ise bu sigortaya sahip olan kişi sayısı çok düşüktür.

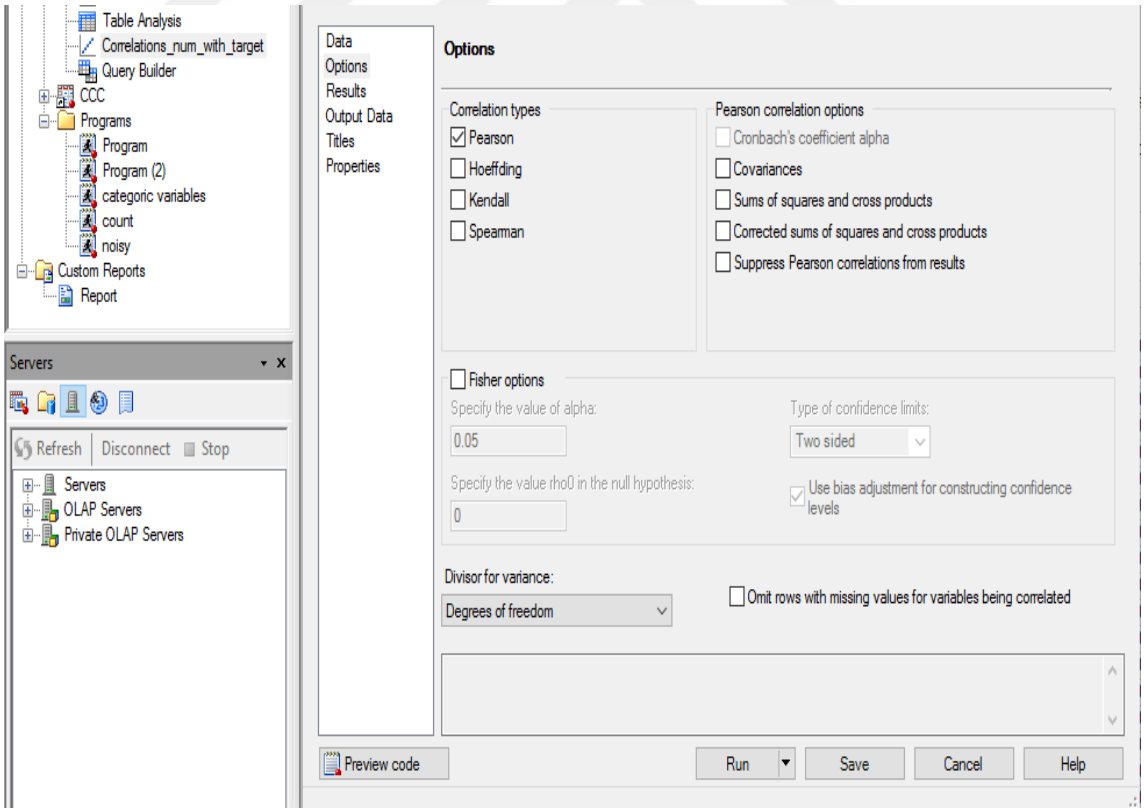
- Kategorik değişkenler arasında gürültülü (noisy) değerlere sahip olanlar bulunmaktadır. Bu değerlerin sistemde bir değişim yapmasının önüne geçmek için daha yumuşak (smooth) hale getirilmesi gerekmektedir. Örneğin, "Roman_catholic" değişkeni "0,1,2,3" değerlerinin dışında çok az değer almıştır. Bu yüzden 3'ten büyük olan değerler 3'e sabitlenmiştir. Buna benzer değişikler "no_religion", "living_together", "high_education", "farmer", "unskilled_labourers", "social_class_B2", "cars_two" ve "car_policies" değişkenlerinde de yapılmıştır.
- Bazı sıralı değişkenlerin değerlerinin neredeyse tamamı 0'dır (%0). Bu değişkenlerin modele performans yükünden başka katkısı yoktur. Dolayısıyla korelasyon analizi aşamasından önce sistemden çıkarılması gerekmektedir. Örneğin, "private_accident_insurance_policies" değişkeninde 5791 kayıt, "lorry_policies" değişkeninde 5813 kayıt, "agricultural_machines_policies" değişkeninde 5801 kayıt, "disability_insurance_policies" değişkeninde 5799 kayıt, "surfboard_policies" değişkeninde ise 5819 kayıt 0 değerine sahiptir. Bu değişkenlerin silinmesi için de SAS koduna başvurulmuştur.

8.2.3. Korelasyon Analizleri ve Değişkenlerin Seçimi

Veri ön işlemenin son basamağı, modelleme çalışmasının önemli aşamalarından birisi olan modele girecek değişkenlerin belirlenmesi işlemidir. Eğer modele katkısı olmayan değişkenler kullanılırsa birçok olumsuz sonuç ortaya çıkabilir. Böyle değişkenler hedef değişkeni anlamlı bir şekilde açıklayamadığı gibi, açıklayan değişkenlerin de gücünü zayıflatabilmektedir. Bundan dolayı, bu değişkenlerin belirlenip model kurma aşamasında elenmesi gerekmektedir. Korelasyon analizi sayesinde modele girdiğinde en iyi sonucun elde edileceği değişken koleksiyonu seçilebilmektedir.

Korelasyon analizinin nasıl yapılacağı değişkenlerin tipine göre değişmektedir. Sürekli değişkenler için Pearson korelasyon katsayısı kullanılmıştır. İstatistiki açıdan incelendiğinde modele katkı sağlamayacak olan değişkenler çıkarıldıktan sonra geriye

kalan sürekli değişkenlerin hedef değişkenle olan korelasyon ilişkisine bakılmış ve en iyileri seçilmiştir. En iyinin seçilmesinin birden fazla yolu olabilir. Korelasyon katsayısı en yüksek olan belli sayıda değişken seçilebilir. Bu sayı modeli kuran kişi tarafından belirlenebilir. İkinci olarak, belli bir yüzdenin üzerindeki katsayılar alınabilir. Ayrıca, korelasyon katsayısı düşük olsa da hedef değişkeni belirleme noktasında katkı sağlayacağı düşünülen değişkenler de modele alınır. Bunun kararı iş bilgisi (business know how) ile verilir. Dolayısıyla, değişkenler hem korelasyon katsayısı değerlerine bakılarak hem de iş bilgisi bakış açısıyla yorumlanarak seçilir. SAS EG üzerinde korelasyon analizi yapmak için "PROC CORR" prosedürü kullanılabilir. Burada "PROC" prosedürün ve "CORR" korelasyonun kısaltılmış halidir. SAS'ın kendine özgü bir sözdizimi (syntax) ile yazılır ve korelasyon hesaplanır. Ayrıca Şekil 17'de gösterildiği gibi aracın sağlamış olduğu "Correlations" menüsünden analizi yapılacak olan değişkenler işaretlenerek de katsayılar hesaplanabilir.



Şekil.17. SAS Enterprise Guide Üzerinde Korelasyon Analizi

“PROC CORR” prosedürü ile hesaplanan iki veya daha fazla değişken arasındaki korelasyon katsayısı, varsayılan (default) olarak Pearson korelasyon katsayısıdır.

Korelasyon analizi sonucunda hedef değişkeni açıklama gücü yüksek olduğu düşünülen sürekli değişkenler modele girmek üzere seçilmiştir.

Kategorik değişkenlerden hangilerinin modele gireceğinin belirlenmesi için Somers’D ve Cramer’s V değerlerinin yorumlanması önemlidir. Bağımsız değişkenlerle hedef değişken arasındaki ilişki için Somers’D, bağımsız değişkenler arasındaki ilişki için ise Cramer’s V değerine bakmak gerekmektedir. Somers’D değeri belli bir değerin üzerinde olan değişkenlerin kendi aralarındaki Cramer’s V değerlerine bakılır. Kendi aralarında korele oldukları görülen değişkenlerden Somers’D değeri yüksek olanı alınır ve diğer değişken elenir. İş bakış açısıyla değerlendirilip diğer değişkenin de alındığı durumlar olabilir.

Somers’D değerinin bulunması için “proc logistic” prosedürü ve Cramer’s V değerinin bulunması için ise “proc freq” prosedürü kullanılmıştır.

Somers’D değeri en yüksek olan 20 değişken seçilmiş ve aralarındaki korelasyonu değerlendirmek için ki-kare ve Cramer’s V değerleri yorumlanmıştır. Sonuç olarak modele girecek olan kategorik değişkenler de belirlenmiştir.

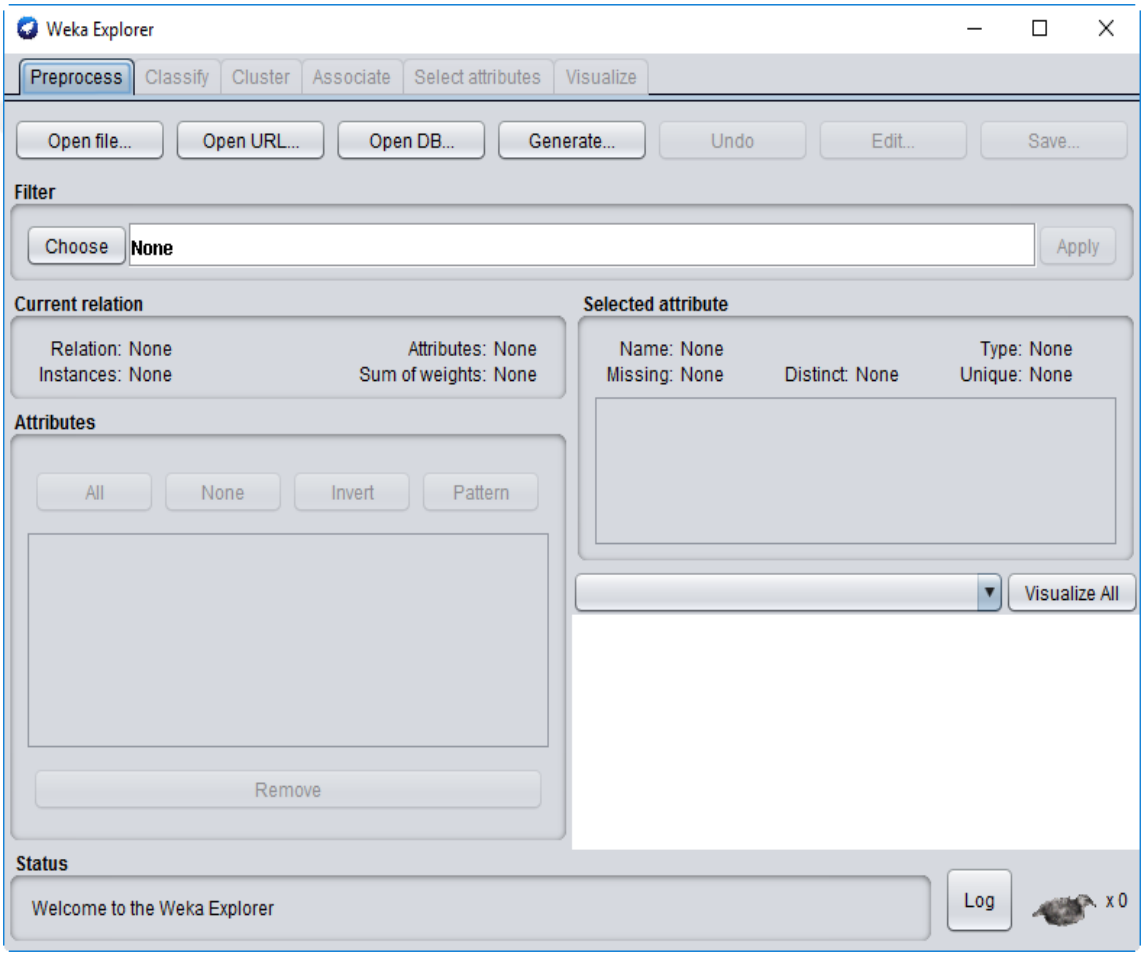
Sonuç olarak modele 12 adet bağımsız değişken ve hedef değişken olan “mobile_home_policies” girmiştir.

8.3. Modelleme

Değişken seçimi yapıldıktan sonra model kurma aşamasına gelinmiştir. Model kurma aşamasında üç teknikten faydalanılmıştır. Bunlar karar ağaçları, lojistik regresyon ve yapay sinir ağlarıdır. Modelleme çalışmasında Weka aracından faydalanılmıştır.

8.3.1. Karar Ağaçları

Uygulamada, karar ağaçları tekniğinin kullanıldığı J48 algoritması kullanılmıştır. C4.5 karar ağacı, Weka'da J48 algoritması olarak adlandırılmıştır. Weka'nın ilk girişteki arayüzünün "Applications" kısmında yer alan "Explorer" butonuna tıklandığında Şekil 18'de gösterilen arayüz açılmaktadır.



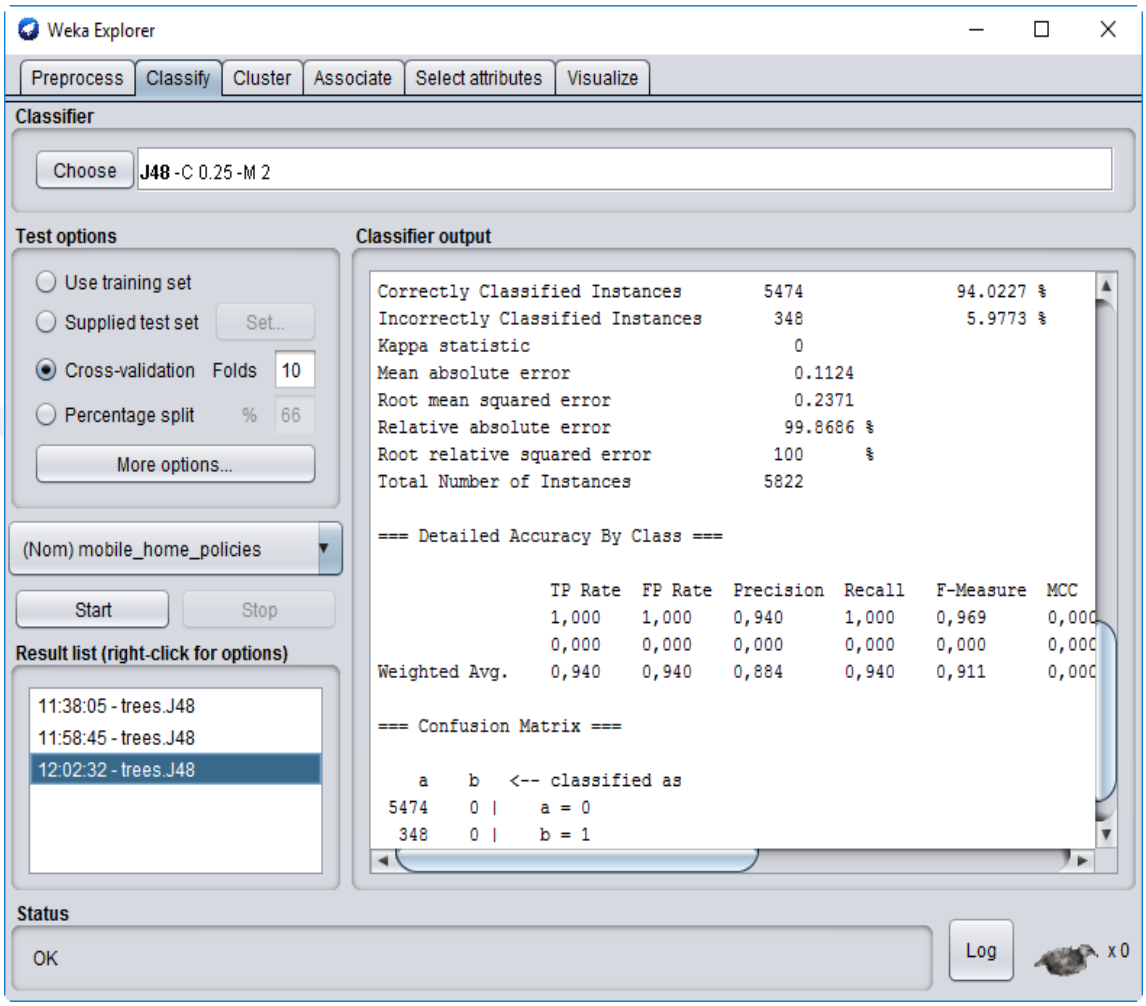
Şekil.18. Weka Explorer Arayüzü

Explorer arayüzünde yer alan "Open file" butonuna tıklanarak ARFF formatındaki veri kümesi Weka'ya aktarılır. Veri kümesi üzerinde yapılacak istatistiki değer görüntüleme ve veri ön işleme süreci "Preproccess" sekmesinde yer alan "Filter" alanında yapılabilmektedir.

Seçilen değişkenler üzerinde model kurulacağı için diğer değişkenler ve o değişkenlere ait değerlerin silinmesi gerekmektedir. Bu işlem elle (manuel olarak) veya Weka kullanılarak yapılabilir. Aktarılan veri kümesinde bulunan değişkenler Preproccess sekmesinde yer alan "Attributes" kısmında görünmektedir. Burada çıkarılmak istenen değişkenler "Remove" butonuna tıklanarak kaldırılabilir. Değişken kaldırıldığında onun değerleri de otomatik olarak silinmektedir. Bu işlemler uygulanmış ve aynı işlemleri diğer algoritmalarda tekrarlamamak için kümenin yeni hali "Save..." butonu ile kaydedilmiştir.

Model kurma aşaması "Classify" sekmesi altında bulunan sınıflandırma ve regresyon algoritmaları aracılığıyla yapılacaktır. Veri kümesi üzerinde herhangi bir değişiklik yapmadan modeli kurmak istediğimizde model yüksek bir doğruluk oranı ile kurulmaktadır. Ancak hedef değişkenle ilgili tüm değerler 0 (sıfır) olarak tahmin edilmiştir. Bunun sebebi, dengesiz bir veri kümesiyle çalışılmasıdır. Kümede, toplam 5822 kayıt bulunmaktadır. Şekil 19'da görüldüğü üzere kayıtların 348 tanesinde hedef değişkenin değeri 1 (bir), diğer 5474 kayıtta ise 0 (sıfır)'dır.

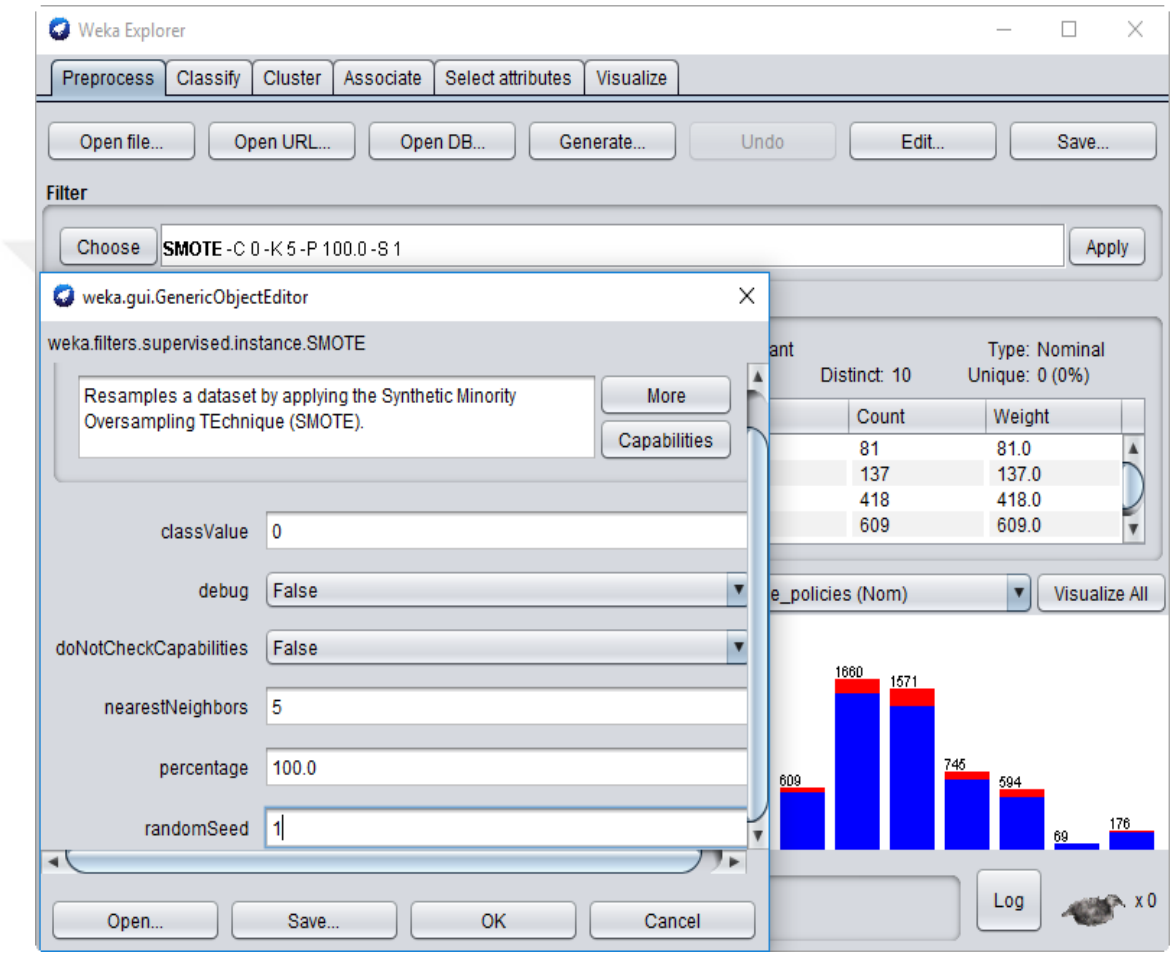
Tüm değerler 0 (sıfır) tahmin edildiği için Karışıklık Matrisi'nde 1 (bir) sütunundaki toplam sayı değerleri 0 (sıfır)'dır. Ayrıca model %94,0227 gibi yüksek bir doğruluk oranına sahiptir. Lojistik regresyon ve yapay sinir ağları modellerinde de durum değişmemektedir. Motokaravan sigortası yaygın olmayan bir sigorta türü olduğu için yeni gelecek olan test veri kümelerinde de iyi tahminlemeler ve %90' dan daha fazla oranda doğruluklar elde edilmiştir. Ancak, modellemedeki asıl amaç 1 (bir) olan değerleri doğru tahmin etmek olduğu için buna yönelik teknikler kullanılmıştır. Dengesiz veri kümesinin çözümleri olan iki yöntem birlikte kullanılmıştır. Bunlar "SMOTE" ve "SpreadSubsample" yöntemleridir ve "Preproccess" sekmesinde yer alan "Filter" bölümündeki "supervised" kısmında bulunmaktadırlar.



Şekil.19. Dengesiz Veri Kümesinde J48 Algoritmasının Sonucu

SMOTE (Synthetic Minority Oversampling Technique) yöntemiyle istenen verilere benzer özellikte veriler türetilir.. SMOTE ile kaydı az olan hedef değişken değerinin daha fazla kayda sahip olması sağlanır. Kayıtların sayısı arttırılırken benzerlik derecelerinin ölçüsü belirlenirken KNN algoritması kullanılır. Veriler, baz alınan uzaklık ölçüsüne göre en yakın komşuların özelliğini taşıyacak şekilde çoğaltılır. Bu işlem sayesinde makinenin öğrenmesi daha iyi olur ve daha iyi sonuçlar elde edilebilir. Ancak makinenin aşırı öğrenmesi, tekrara düşmesi veya ezberlemesi gibi bir tehlike de söz konusu olabilir. Bu yüzden Şekil 20'de bulunan SMOTE uygulama arayüzündeki yüzde artış (percentage) seçeneğinin ve en yakın komşuluk derecesinin (nearestNeigbors) kaç olması gerektiğine dikkatli karar vermek gerekir. Percentage seçeneği 100 yapılıp

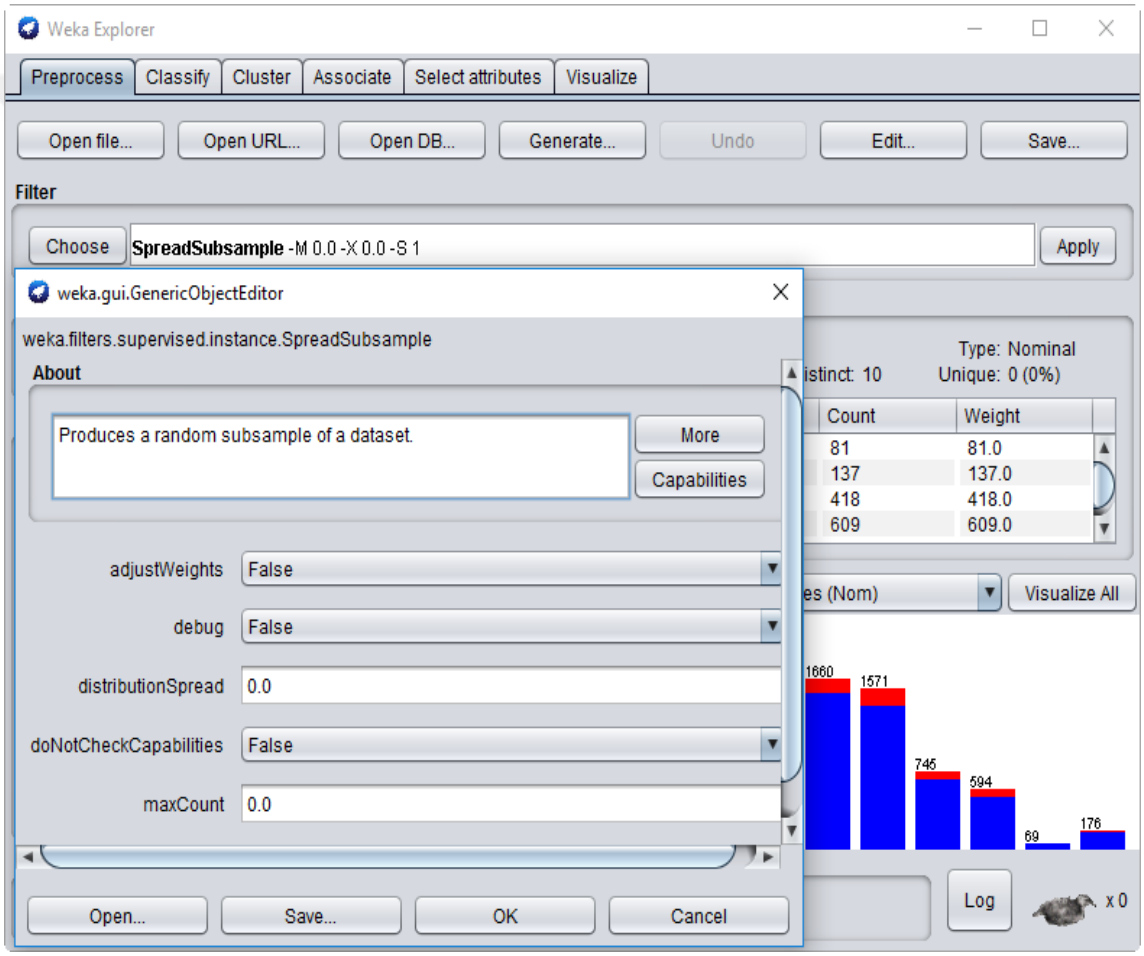
uygulanırsa az sayıda kayda sahip olan hedef değişken değerinin sayısı %100 artar ve iki katına çıkmış olur. nearestNeigbors derecesinin varsayılan değeri 5'tir. Eğer percentage ve nearestNeigbors değerleri dikkatli seçilmezse doğrulama (validate) verisinde iyi sonuçlar verse bile test verisindeki tahminleme gücü zayıf olur.



Şekil.20. SMOTE Yönteminin Uygulama Arayüzü

SpreadSubsample yönteminde ise SMOTE yönteminin tersi bir işlem uygulanmaktadır. SMOTE sayesinde veri sayısı arttırılmış olur. Ancak SpreadSubsample ile veri sayısında azalma yaşanır. Verinin karakteristiğine göre farklı seçenekler kullanmak mümkün olsa da hedef değişkenin değerlerinin kayıt sayısının eşitlenmesi yaygındır. Şekil 21'de bulunan "distributionSpread" alanı kayıtların hangi orana sahip olacaklarını gösterir. Örneğin, bu alanın değeri 1.0 olduğunda kayıtlar eşit olacak demektir. Sayı değiştiğinde Böylece kayıt sayısı çok yüksek olan değerlerin, kayıt sayısı az olanlara göre daha fazla ağırlık göstermesinin önüne geçilmiş olur. Bunun yanında

makinenin az olan değerin özelliklerini de öğrenmesi sağlanmış olur. Kullanılan veri kümesi için düşünüldüğünde 1 (bir) olan kayıtların özelliklerinin daha iyi öğrenir. Sayıların eşitlenmesi sürecinde 0 (sıfır) değerine sahip olan kayıtlardan hangilerinin eleneceği de önemlidir. Elemeyi sağlayan algoritma, kalan kayıtların 0 (sıfır) değerinin özelliklerini en iyi şekilde temsil etmesini sağlayacak şekilde çalışmaktadır. Buna rağmen öğrenme kayıplarının yaşanması muhtemeldir. Ancak hiçbir ön işleme tabi tutmadan yapılacak modelleme çalışmasına göre çok daha iyi sonuçlar elde edileceği de açıktır.



Şekil.21. SpreadSubsample Yönteminin Uygulama Arayüzü

Kullanılan veri kümesinin karakteristiği incelendiğinde SMOTE ve SpreadSubsample yöntemlerinin beraber uygulanmasına karar verilmiştir. Öncelikle SMOTE uygulanmış ve percentage değeri 200 olarak belirlenmiştir. Daha sonra da SpreadSubsample uygulanmış ve distributionSpread değeri 1.0 olarak belirlenmiştir. Öğrenme (train) veri kümesi üzerinde Cross – Validation uygulanmıştır. "Folds" alanına

10 yazılmıştır. Cross – Validation bu şekilde uygulandığında rastgele şekilde belirlenen verilerle öğrenme kümesi %90 ve doğrulama kümesi %10 olacak şekilde bölünme sağlanır. Daha sonra tüm verilerin doğrulama kümesinde tekrarsız olarak yer alması sağlanacak şekilde 9 farklı dağılım daha oluşturulur. Son olarak işlem yapılan 10 durumdaki değerlerin ortalaması alınarak nihai bir sonuca ulaşılır. Bu durumda J48 algoritmasının doğruluk oranı %94,0247 gibi iyi bir değere sahip olmuştur. Sonrasında ise aynı algoritma test kümesine uygulanmış ve aşağıdaki tabloda yer alan değerlere ulaşılmıştır.

Tablo.9. J48 Algoritmasının Test Kümesi Üzerinde Çalışması Sonucu Oluşan Karışıklık Matrisi

	Öngörülen sınıf (Predicted Class)		
Gerçek Sınıf (Actual Class)		0	1
	0	3078	684
	1	68	170

Karışıklık matrisinde ortaya çıkan değerlere bakıldığında 0 ve 1 değerlerinin çoğunluk olarak doğru sınıflandırıldığı görülmektedir. Denklem 6'da belirtilen modelin doğruluğu aşağıdaki gibi bulunmuştur.

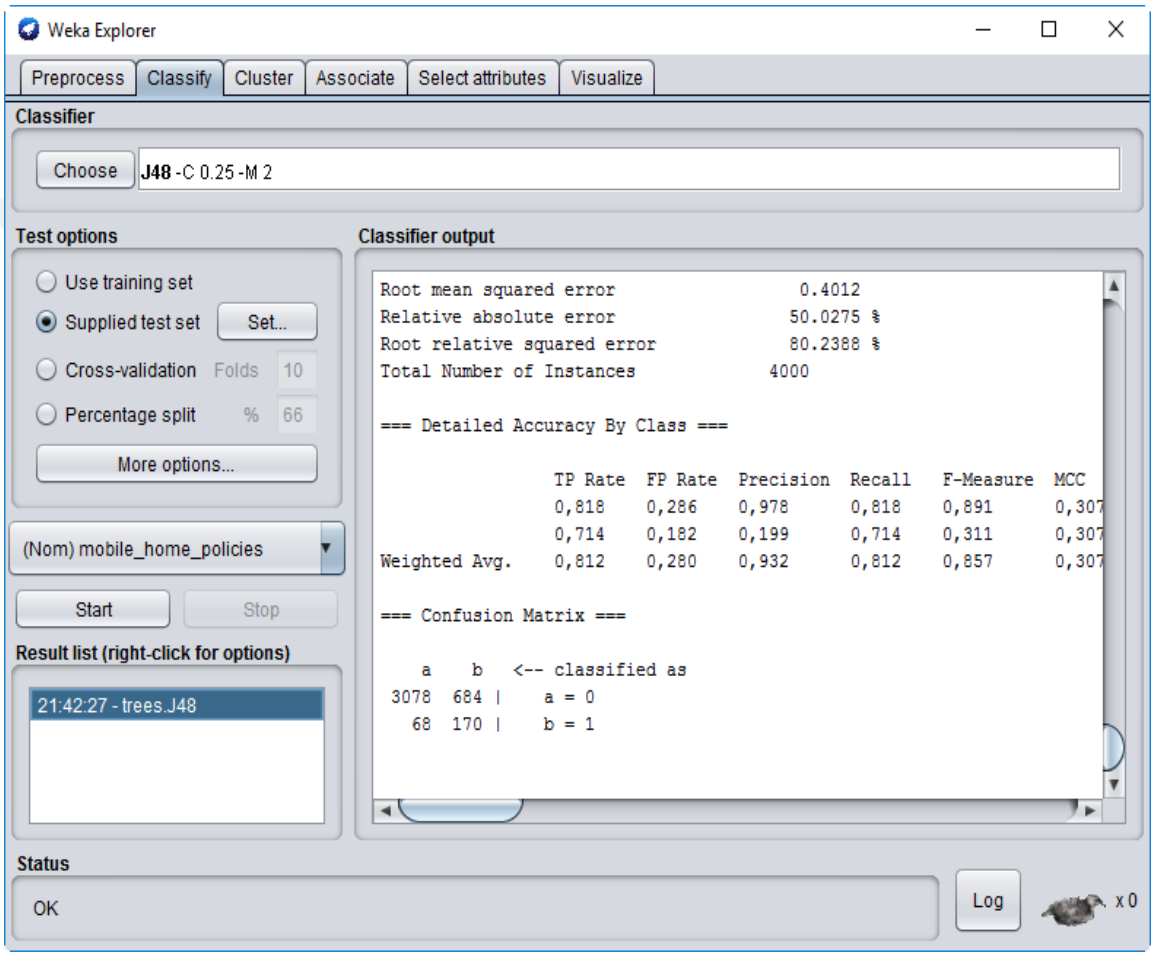
Doğruluk = (170 + 3078) / 4000 = %81,2

Test veri kümesini yüksek bir doğruluk oranıyla tahminleyen bir algoritma elde edimiştir. Modelleme çalışmasındaki asıl amaç motokaravan sigortası yaptıran kişileri bulmak olduğu için duyarlılık değerine bakılmalıdır. Şekil 22'de görüldüğü gibi test kümesinde yer alan 238 motokaravan sigortası yaptıran kişinin 170'i doğru tahmin edilmiştir. Motokaravan sigortası yaptırmayan kişilerin de 3078 tanesi doğru tahmin

edilmiştir. Denklem 7'de belirtilen modelin duyarlılığı ve Denklem 8'de belirtilen modelin belirliliği aşağıda hesaplanmıştır.

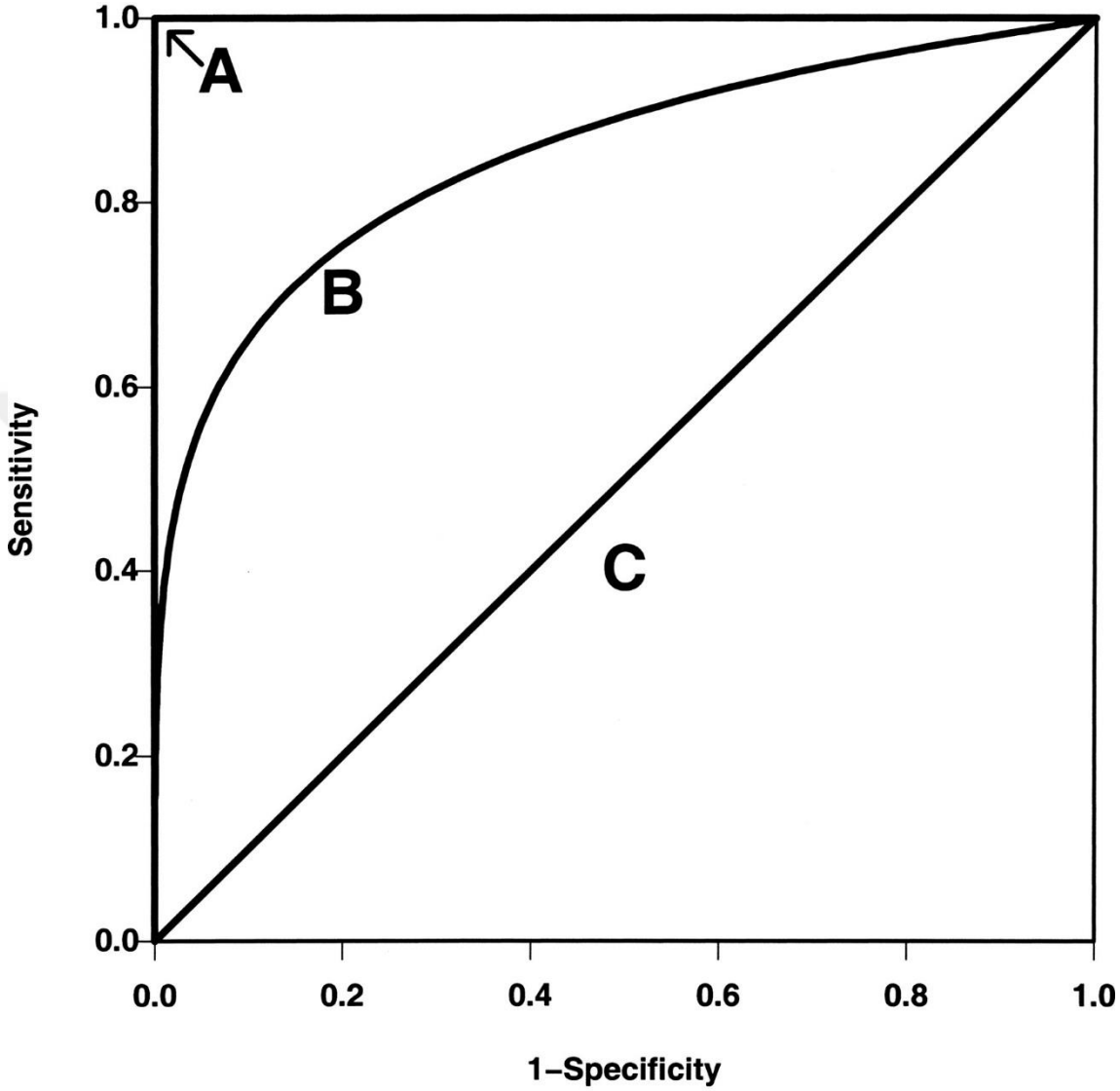
Duyarlılık = 170 / 238 = %71,42

Belirlilik = 3078 / 3762 = %81,81



Şekil.22. Weka'da J48 Algoritmasının Test Veri Kümesi Üzerinde Çalışmasının Sonucu

Model kurma işleminin başarısını test etmenin yollarından birisi de ROC (Receiver Operating Characteristic) eğrisinin yorumlanmasıdır. TP oranıyla FP oranının karşılaştırılması yapılır. Diğer bir ifadeyle Şekil 23'te görüldüğü gibi duyarlılık ile (1 - belirlilik) karşılaştırılması yapılır. X ekseninde (1 - belirlilik) ve y ekseninde duyarlılık yer almaktadır. ROC eğrisi olası bütün durumlar, yanlış sınıflandırmadan kaynaklı maliyetler ve sınıf dağıtımları hakkında bilgi vermektedir. ROC eğrisi veri madenciliğinin yanında tıp, veteriner hekimliği, radyoloji ve psikoloji gibi alanlarda da kullanılmaktadır.



Şekil.23. ROC Eğrisi

Kusursuz bir kesinliğe sahip olan ideal ROC eğrisi (0,0) noktasından (0,1) noktasına (yani y ekseni boyunca) ve yatay olarak ise (0,1) noktasından grafiğin sağ üstünde bulunan (1,1) noktasına doğru çizilir. Rastgele ve üzerinde uğraşılmadan yapılan bir tahminlemenin grafiğinin (0,0) noktasından (1,1) noktasına doğru köşegen halinde olması muhtemeldir. Uygun olan bir test ise bu iki uç eğrinin arasında ve grafiğin üst kısmına yakın bir ROC eğrisi gösterir. Test eğrisinin x eksenine ve sağa yakın olduğu durumda ise testin doğru sonuçtan ziyade yanlış sonuç verdiği anlaşılır. ROC eğrisinin altında kalan alana ROC alanı denilmektedir. Şeklin tamamı bir kenarı 1 birim olan kare

olduğu için toplam alan 1 birim karedir. Dolayısıyla, ROC alanı ne kadar 1 birim kareye yakınsa o kadar iyi bir sonuç olduğu anlamına gelmektedir. Karar ağacı algoritmasının test kümesine uygulanması sonucunda elde edilen ROC alanı 0,817 birim karedir. Buradan modelin iyi bir ROC eğrisine sahip olduğu söylenebilir. Model doğru pozitiflere, yanlış pozitiflerden daha çok ulaşmıştır. En mükemmel olan ROC eğrisine yakın bir eğri oluşmuştur. Modelin hassasiyeti ile kesinliği arasında iyi bir denge kurulmuştur.

J48 algoritmasının inşasında entropi ve enformasyon kazanımı kullanılmaktadır. Bu bağlamda kök kısmına gelen değişken "car_policies" olmuştur. "car_policies" değişkeni sıralı bir değişkendir ve araba sigortası için ödenen primi ifade eder. Bunun dışında hedef değişkeni belirlemek için modelde kullanılan başka değişkenler de vardır. Bunlar aşağıda verilmiştir.

- Married : Kişinin bulunduğu yerde hangi oranda evli kişi olduğu gösterir.
- High_education : Kişinin bulunduğu yerde hangi oranda yüksek eğitim seviyesine sahip kişi olduğu gösterir.
- High_status : Kişinin bulunduğu yerde hangi oranda yüksek statüye sahip kişi olduğu gösterir.
- Home_owners : Kişinin bulunduğu yerde hangi oranda oturduğu evin sahibi olan kişi olduğu gösterir.
- 1_car : Kişinin bulunduğu yerde hangi oranda bir arabaya sahip kişi olduğu gösterir.
- Private_health_insurance : Kişinin bulunduğu yerde hangi oranda özel sağlık sigortasına sahip kişi olduğu gösterir.
- Average_income : Kişinin bulunduğu yerde hangi oranda ortalama gelire sahip kişi olduğu gösterir.
- Purchasing_power_class : Kişinin bulunduğu yerde hangi oranda satın alma gücü yüksek olan kişi olduğu gösterir.
- Private_3rd_party_insurance : Kişinin bulunduğu yerde hangi oranda üçüncü parti sigorta yaptıran kişi olduğu gösterir.
- Fire_policies : Kişinin bulunduğu yerde hangi oranda yangın sigortası yaptıran kişi olduğu gösterir.

Bütün bu değişkenler hedef değişkeni belirlemede önemli rol oynamışlardır. Karar ağacı algoritmalarının önemli avantajlarından birisi de bu şekilde kolay yorumlanabilir olmasıdır.

8.3.2. Lojistik regresyon

Lojistik regresyon da Şekil 24'te görüldüğü gibi J48 algoritması gibi "Explorer" ekranındaki "Classify" sekmesinde yer almaktadır. Veri kümesi Weka'ya alındıktan sonra lojistik regresyon algoritması veri üzerinde Cross – Validation yöntemi kullanılarak uygulanmış ve %80,5779 oranında başarılı tahminleme yapılmıştır.

Karar ağacında olduğu gibi lojistik regresyonda da SMOTE uygulanmış ve percentage değeri 200 olarak belirlenmiştir. Daha sonra da SpreadSubsample uygulanmış ve distributionSpread değeri 1.0 olarak belirlenmiştir. Elde edilen karışıklık matrisi Tablo 10'da verilmiştir.

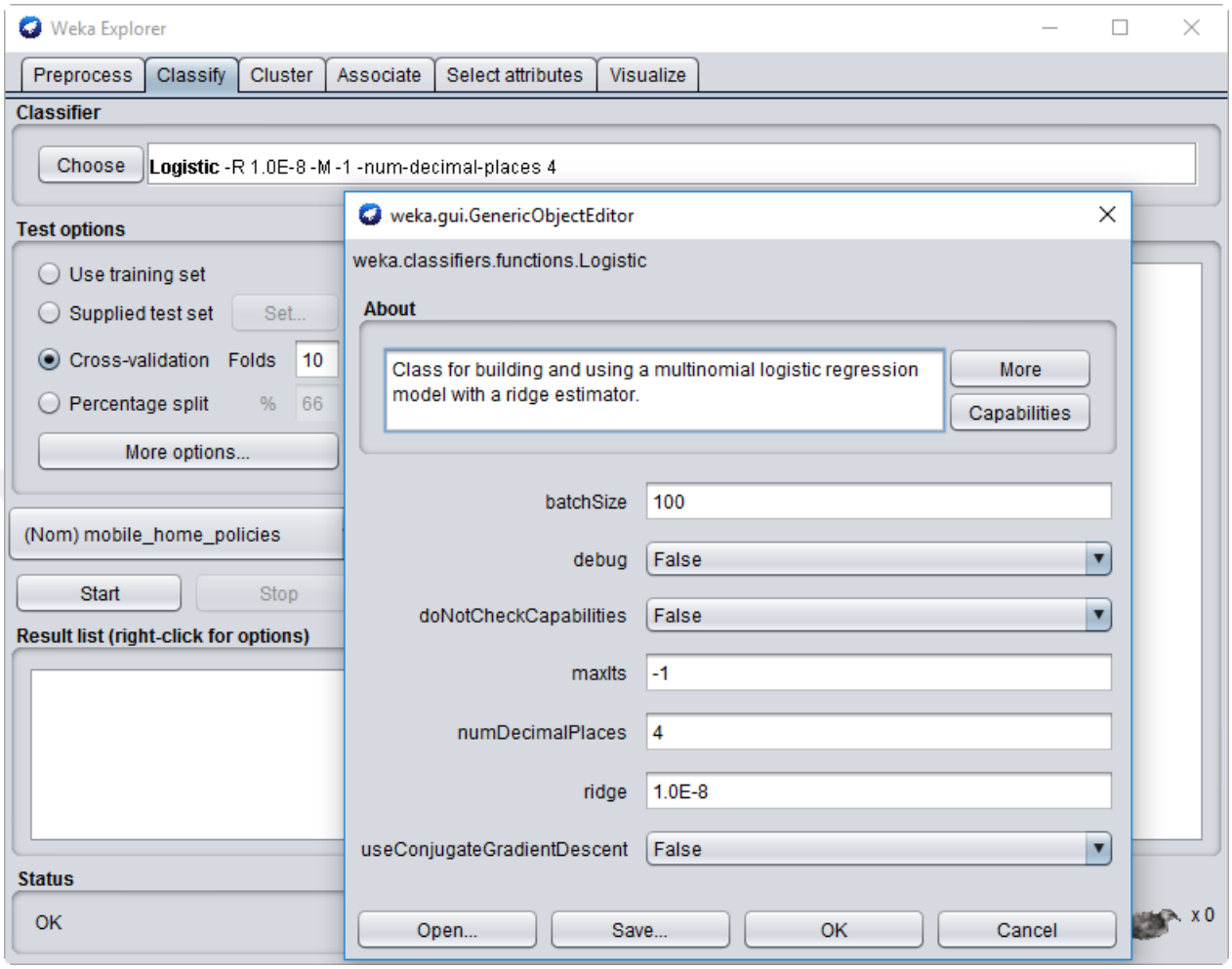
Modelin doğruluğu aşağıdaki gibi bulunmuştur.

Doğruluk = (157 + 2544) / 4000 = %67,525

Modelin duyarlılığı ve belirliliği aşağıda hesaplanmıştır.

Duyarlılık = 157 / 238 = %65,96

Belirlilik = 2544 / 3762 = %67,62



Şekil.24. Weka'da Lojistik Regresyon Uygulaması Arayüzü

Tablo.10. Lojistik Regresyon Algoritmasının Test Kümesi Üzerinde Çalışması Sonucu Oluşan Karışıklık Matrisi

	Öngörülen sınıf (Predicted Class)		
Gerçek Sınıf (Actual Class)		0	1
	0	2544	1218
	1	81	157

Modelin doğruluğu orta seviyenin biraz üzerindedir. Motokaravan sigortası yaptıranları tahminlemedeki gücü düşük değildir. Ancak daha iyi seviyelerde bir sonuç elde etme beklentisi mevcut olduğu için istenen ölçüde başarılı olduğu söylenemez.

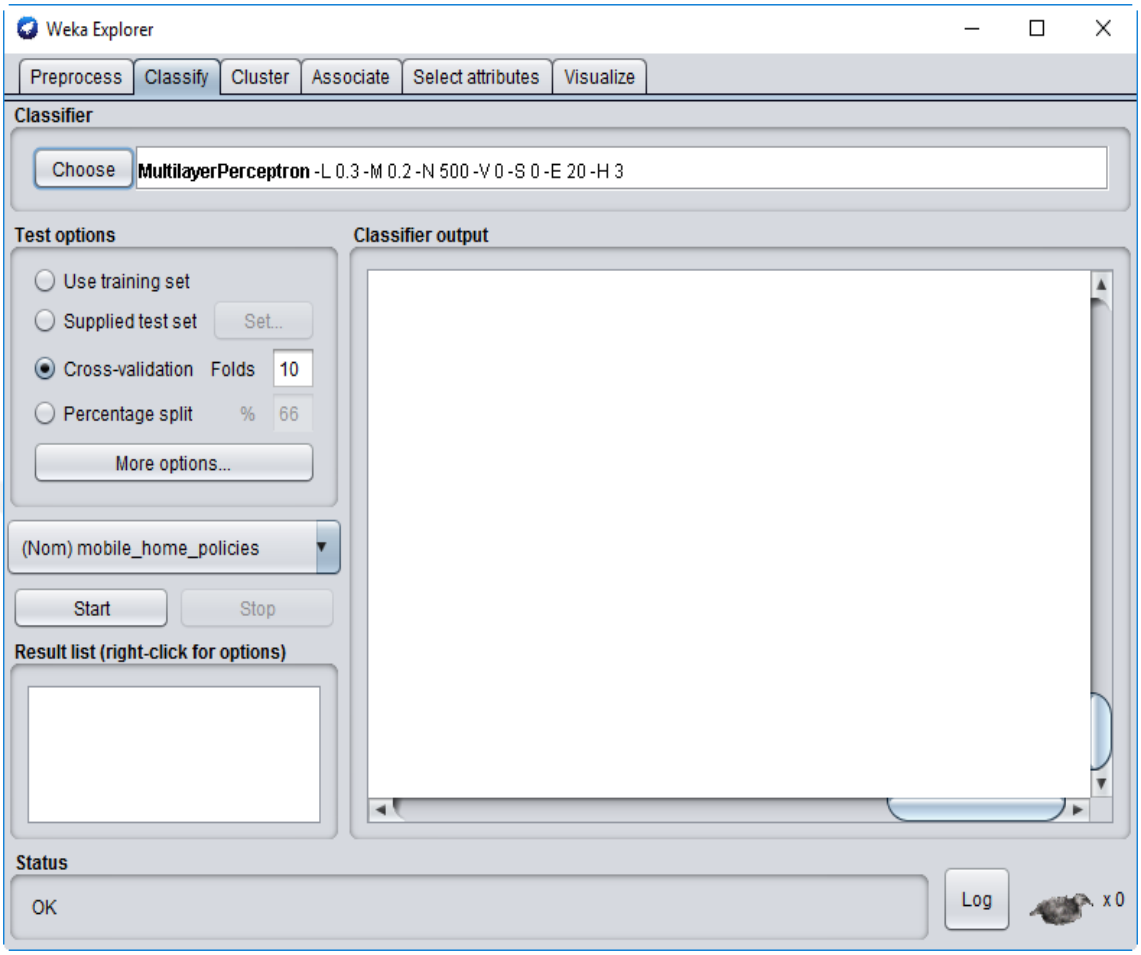
ROC alanı 0,719 olarak hesaplanmıştır. Bu değer ROC eğrisinin iyi kaliteye sahip olduğunu gösterir. Lojistik regresyon modeli de doğru pozitiflere, yanlış pozitiflerden daha çok ulaşmıştır. Ancak en mükemmel olan ROC eğrisine çok yakın bir eğri oluşmamıştır. Modelin hassasiyeti ile kesinliği arasında normal bir denge kurulmuştur.

Değişken bazında bakıldığında ise "Odds Ratio" değerlerinin değerlendirilmesi gerekmektedir. En yüksek "Odds" değerine sahip olan değişken "High_status" değişkenidir. Yüksek statüyü belirten bu değişken lojistik regresyon algoritmasının en başarılı açıklayıcı değişkenidir. Daha sonra "Private_health_insurance" ve "Home_owners" değişkenleri gelmektedir.

Lojistik regresyon tekniği kullanılarak kurulmuş olan bu modelde iyi bir sonuç elde edilememiştir. Bundan dolayı diğer iki model arasında bir tercih yapılacaktır.

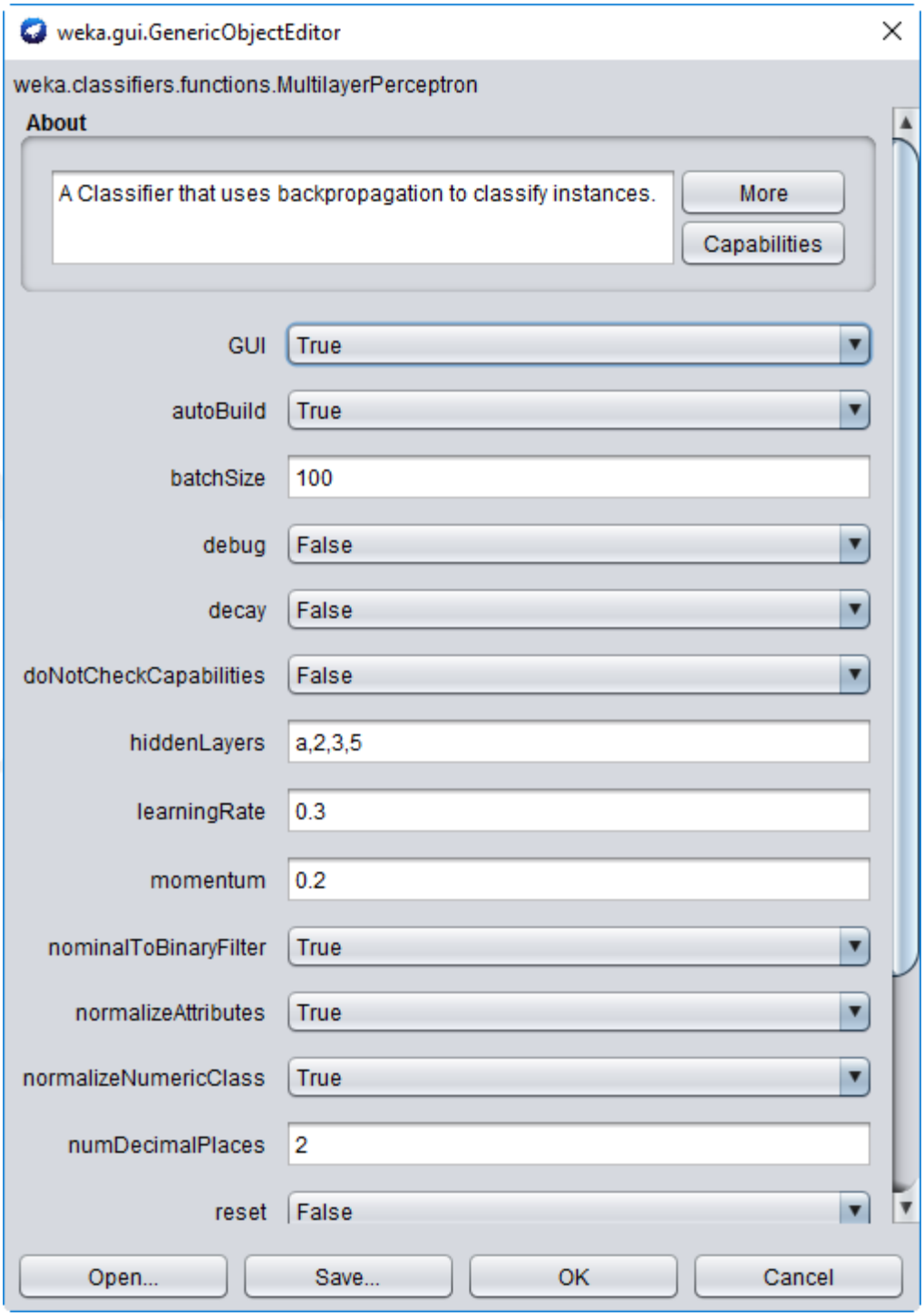
8.3.3. Yapay Sinir Ağları

Yapay sinir ağları, Lojistik regresyon ve J48 algoritması gibi "Explorer" ekranındaki "Classify" sekmesinde yer almaktadır. Weka'daki arayüzü Şekil 25'te gösterilmiştir.



Şekil.25. Weka'da Çoklu Katman Kullanılan Yapay Sinir Ağları Algoritmasının Arayüzü

Kullanılan arayüzde algoritmanın özellikleri değiştirmek istendiğinde "MultilayerPerceptron" satırına tıklanır ve açılan yeni ekranda istenen değişiklikler yapılır. Bu yeni ekran Şekil 26'da gösterilmiştir. Modelin yapay sinir ağları şeklini görebilmek için "GUI" (The Graphical User Interface) seçeneği "True" yapılmalıdır. Modelin otomatik olarak inşa edilmesi için "autoBuild" seçeneği "True" yapılmalıdır. "hiddenLayers" seçeneğinde kaç tane gizli katman olacağına karar verilir. "learningRate" seçeneğinde geri yayılım için öğrenme oranının ne olacağı belirlenir. 0 ile 1 arasında değerler alır. Varsayılan değeri 0,3 olarak kararlaştırılmıştır. "momentum" seçeneğinde geri yayılım için momentum oranına karar verilir. Bu seçenek de 0 ile 1 arasında değerler alır. Varsayılan değeri 0,2 olarak belirlenmiştir.



Şekil.26. Weka'da Yapay Sinir Ağları Algoritmasının Özellikler Arayüzü

Öğrenme veri kümesi Weka'ya alındıktan sonra MultilayerPerceptron algoritması veri üzerinde Cross – Validation yöntemi kullanılarak uygulanmış ve % 94,913 oranında başarılı tahminleme yapılmıştır.

Karar ağacı ve lojistik regresyonda olduğu gibi yapay sinir ağlarında da percentage değeri 200 olarak belirlenen SMOTE uygulanmıştır. Sonrasında ise diğer algoritmalarda olduğu gibi SpreadSubsample uygulanmış ve distributionSpread değeri 1.0 olarak belirlenmiştir. İşlem sonucunda elde edilen karışıklık matrisi aşağıda verilmiştir.

Tablo.11. Multilayerperceptron Algoritmasının Test Kümesi Üzerinde Çalışması Sonucu Oluşan Karışıklık Matrisi

	Öngörülen sınıf (Predicted Class)		
Gerçek Sınıf (Actual Class)		0	1
	0	3174	588
	1	42	196

Modelin doğruluğu aşağıdaki gibi bulunmuştur.

Doğruluk = (196 + 3174) / 4000 = %84,25

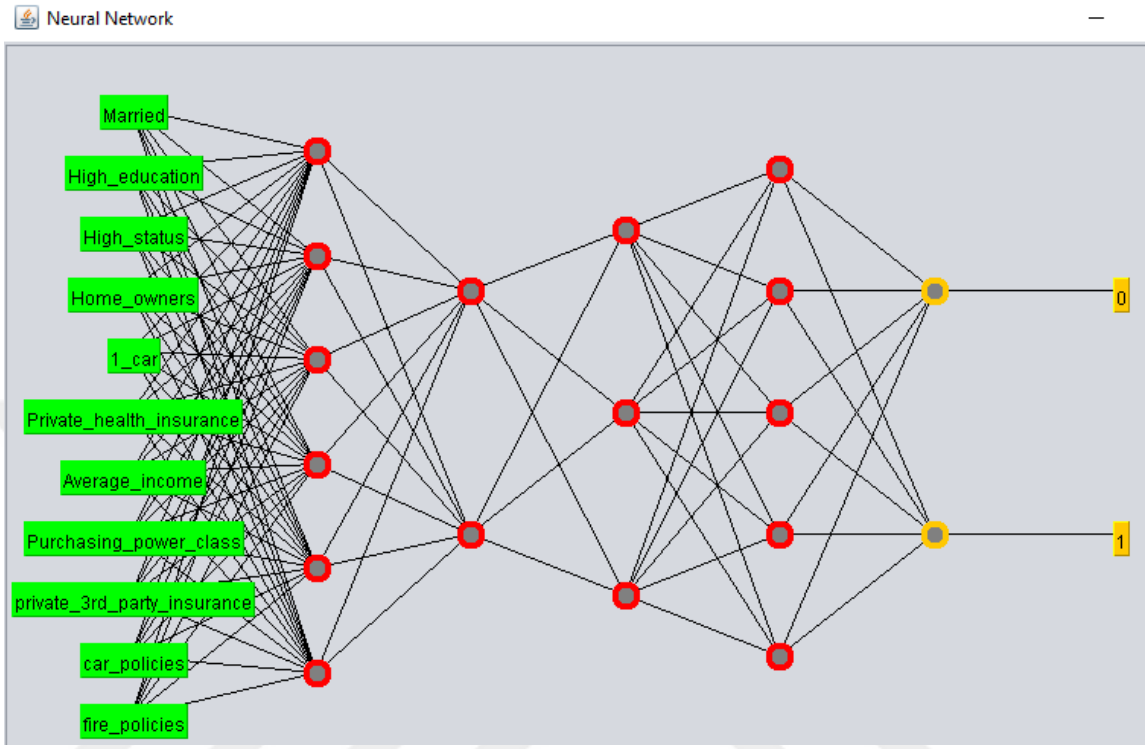
Modelin duyarlılığı ve belirliliği aşağıda hesaplanmıştır.

Duyarlılık = 196 / 238 = %82,35

Belirlilik = 3174 / 3762 = %84,37

Modelin doğruluğu iyi bir seviyededir. Motokaravan sigortası yaptıranları tahminlemedeki gücü yüksektir. ROC alanı 0,881 olarak hesaplanmıştır. Bu değer ROC eğrisinin çok iyi kaliteye sahip olduğunu gösterir. Modelin doğru pozitiflere, yanlış pozitiflerden daha çok ulaşmış olduğu söylenebilir. En mükemmel olan ROC eğrisine yakın bir eğri oluşmuştur. Modelin hassasiyeti ile kesinliği arasında çok iyi bir denge kurulmuştur.

Modelin sinir ağları şeklinde oluşan görüntüsü Şekil 27'de gösterilmiştir.



Şekil.27. Weka'da Oluşan Yapay Sinir Ağlarının Görüntüsü

Nitelik ve sınıflar uzayda doğrusal bir şekilde ayrılmadığı için (not linearly separable) tek katmanlı bir yapay sinir ağı tasarlamak iyi bir sonuç vermemektedir. Bu yüzden gizli katmanlar kullanılmıştır. Yapılan denemeler ve modelin öğrenmesi çalışmasında yukarıdaki sonuçları veren yapay sinir ağı yapısında "hiddenLayers" olarak (a,2,3,5) kullanılmıştır. Burada a sayısı (Nitelik sayısı + Sınıf sayısı)/2 formülü kullanılarak elde edilir. Şekilde de görüldüğü gibi a için (11 + 2)/2 = 6,5 olduğundan dolayı 6 değeri bulunmuştur. 3 gizli katman kullanılmış ve bu katmanlarda 2,3 ve 5 olmak üzere toplam 10 nöron kullanılmış ve en iyi sonuca ulaşılmıştır.

Değişken bazında bakıldığında ise modelin yorumlanması kolay değildir. Yapay sinir ağlarının genel problemlerinden birisi yorumlanamamasıdır. Ancak sonuç odaklı bakıldığında iyi bir sonuç elde edildiği söylenebilir. Modelin çalışma hızı, diğer algoritmalarla karşılaştırıldığında oldukça yavaştır.

8.4. Modellerin Karşılaştırılması ve Bulgular

Modelleme çalışmalarının üçü de tamamlandığında belli kriterler ışığında bir tanesi seçilir ve devreye alınır. Bundan sonraki aşamada, sigorta şirketleri modelin çıktıları ışığında kampanya belirleyebilir. Hangi değişkenlerin hedef değişkeni etkilediği ortaya çıkarıldığı için o değişkenlere sahip olan veya daha çok sahip olan kişilere odaklanılmalıdır. Sosyal medya reklamı yapılacaksa o kişilerden oluşan gruplara reklam verilmelidir. Cep telefonu mesajları o kişilere atılmalıdır. Böylece hedef kitleye en az maliyetle ulaşılabilir.

Yapılan çalışmalar sonucunda modellere ait sonuçlar ve karşılaştırmalar Tablo 12'de verilmiştir.

Tablo.12. Modellerin Karşılaştırılması

	Karar Ağacı (J48)	Lojistik Regresyon	Yapay Sinir Ağları (MP)
Öğrenme Kümesi Doğruluk Oranı	%94,0247	%80,5779	%94,913
Öğrenme Kümesi Duyarlılık Oranı	%95,9	%84,6	%94,7
Öğrenme Kümesi Belirlilik Oranı	%92,3	%76,8	%95,2
Test Kümesi Doğruluk Oranı	%81,2	%67,525	%84,25
Test Kümesi Duyarlılık Oranı	%71,42	%65,96	%82,35
Test Kümesi Belirlilik Oranı	%81,81	%67,62	%84,37
ROC Alanı	0,817	0,719	0,881
Çalışma Süresi	0,53 saniye	3,5 saniye	1 saat 17 dk
Yorumlama	Kolay	Kolay	Zor
Maliyet	Düşük	Düşük	Yüksek

Öğrenme veri kümesi, öğrenme (train) ve doğrulama (validate) kümesi olarak ikiye ayrılmıştır. Öğrenme kümesi sonuçlarına bakıldığında küçük bir farkla en iyi doğruluk oranına yapay sinir ağları ile ulaşılmıştır. Ancak motokaravan sigortası yaptıranları tahminlemede (duyarlılık) karar ağaçları en iyi algoritma olmuştur. Lojistik regresyonda ise sigorta yaptıranlar iyi tahminlense de yaptırmayanlar iyi tahminlenemediği için doğruluk oranı olması daha düşük çıkmıştır. Ayrıca diğer algoritmalara göre daha düşük sonuç vermiştir. Motokaravan sigortası yaptırmayanları tahminlemede en iyi algoritma yapay sinir ağları olmuştur.

Modelleme çalışmasının performansını ölçmedeki temel aşama öğrenme kümesinde elde edilen algoritmanın test edilmesidir. Bağış yapan şirketten alınan 4000 müşteriye ait verilerden oluşan test kümesi üzerinde öğrenilen algoritma koşulmuştur. Sonuçlara bakıldığında en iyi doğruluk oranına %84,25 ile yapay sinir ağları sahip olmuştur. İkinci sırada, %81,2 doğruluk oranı ile karar ağaçları yer alır. Lojistik regresyon ise %67,525 doğruluk oranıyla en kötü başarıyı göstermiştir. Modelin duyarlılığına bakıldığında ise motokaravan sigortası yaptıranları tahminlemede en iyi başarıyı yine yapay sinir ağları göstermiştir. Motokaravan sigortası yaptıran 238 müşterinin 196 tanesini doğru tahmin etmiştir. Karar ağaçları tekniğini kullanan J48 algoritmasında ise %71,42 duyarlılık oranı elde edilmiştir. Sigortayı yaptıran 238 müşterinin 170 tanesi doğru tahmin edilmiştir. Lojistik regresyonun duyarlılığı %65,96 oranındadır. 238 müşterinin 157 tanesi doğru tahmin edilmiştir. Motokaravan sigortası yaptırmayanların tahminlemesi için bir ölçü olan belirlilik oranında ise %84,37 ile yapay sinir ağları birinci, %81,81 ile J48 ikinci ve %67,62 ile lojistik regresyon son sırada yer almıştır. Sonuç olarak istenen durumu en iyi tahminleyen algoritmanın yapay sinir ağları tekniğini kullanan MultilayerPerceptron algoritması olduğu belirlenmiştir. İkinci sırada J48 algoritması vardır. İyi bir tahminleme yaptığı söylenemeyecek olan lojistik regresyon algoritması ise son sıradadır.

Duyarlılık ve (1 - Belirlilik) değerlerini y ve x koordinatlarına koyarak modeli değerlendirmede bir ölçü olarak ortaya çıkan ROC alanı değerlerine bakıldığında yine en iyi algoritmanın 0,881 ile MultilayerPerceptron olduğu görülmektedir. Dolayısıyla, MultilayerPerceptron algoritmasının ROC eğrisinin (0,1) ile (1,1) noktalarını birleştiren

doğru parçasına yakın bir yerden geçtiği söylenebilir. İkinci sırada yer alan J48 algoritmasının ise ROC alanı 0,817 olarak ölçülmüştür.

Çalışma sürelerine bakıldığında ise J48 ve lojistik regresyon algoritmalarının çok hızlı olduğu görülmektedir. MultilayerPerceptron algoritmasının ise diğer ikisine kıyasla oldukça uzun sürdüğü görülmektedir. Bu algoritmanın çalışmasının 1 saat 17 dakika sürdüğü görülmektedir. Sürenin uzunluğunda özellikle yapay sinir ağlarındaki katman sayısı etkilidir. Bellekte uzun süre yer kaplaması ve sonuçların diğer algoritmalara göre daha geç ortaya çıkmasından dolayı MultilayerPerceptron algoritmasının maliyeti yüksektir. Çalışma süresini etkileyen faktörlerden birisi de değişkenlerin tipleridir. Algoritmasların sayısal değişkenleri yorumlaması daha hızlı olmaktadır. Kategorik değişkenlerle çalışmak daha uzun sürmektedir. Ayrıca sürenin uzunluğu, elde edilmek istenen bilginin karakteristiği göz önünde bulundurulduğunda çok kritik bir değişken değildir. Anlık yorumlanması gereken modellerin kurulduğu çalışmalarında önemlidir.

Diğer bir açıdan bakıldığında ise karar ağacı algoritmasını yorumlamak kolaydır. Hangi değişkenin kökte olduğu, dalların hangi durumlarda kaça ayrıldığı gibi konular kolay bir şekilde açıklanabilmektedir. Ancak yapay sinir ağlarının açıklanması ve yorumlanması kolay değildir. Gizli katmanın da devreye girmesi de bunda etkili olmaktadır. J48 algoritmasına bakıldığında kökte araba sigortasına verilen primi belirten değişken vardır. Kökten ayrılan dalların ucunda ise evlilik ve ortalama gelir değişkenleri bulunmaktadır. Dolayısıyla, kurulan karar ağacı algoritması motokaravan sigortası yaptıranlara ulaşmak isteyen kurumlara fikir verecek yapıdadır.

Sonuç olarak en başarılı model yapay sinir ağları tekniğinin kullanıldığı MultilayerPerceptron algoritmasıyla kurulan modeldir. Ancak yorumlanması zordur. Bu yüzden doğruluğu nispeten daha düşük olsa da karar ağaçları algoritması da kullanılabilir. Bu iki algoritmanın başarısına bakıldığında veri ön işleme aşamasında doğru yaklaşımlar geliştirilerek modelde kullanıldığında en doğru sonuçları veren değişkenlerin seçildiği görülmüştür.

9. SONUÇ

Dijitalleşmenin ön plana çıktığı günümüzde bilgisayar ve bilgisayar benzeri makinelerin yükü ve görevinde ciddi artışlar yaşanmaktadır. Mobil cihazların da yaygınlaşmasıyla internetin günlük hayattaki yeri çok artmıştır. Dünya nüfusunun yarısının internetle entegre bir şekilde yaşadığı tahmin edilmektedir. Bunların doğal sonucu olarak veri tabanlarına hızlı veri akışı olmaktadır. Sadece sosyal medya mecralarındaki anlık veri üretimi düşünüldüğünde bu akışın artarak devam edeceği düşünülebilir. Bu kadar büyük bir kütle ve hacme sahip bir yığının boşa akması düşünülemez. Veri madenciliği, bu noktada devreye girmektedir. Birçok işte ve alanda daha önceki yaşananlar daha sonra yaşanacak olanlarla ilişkilidir. Örneğin, ülkelerin veya şirketlerin bir yılda oluşturdukları katma değer miktarı birbirine yakın yıllarda genellikle yakın oranlarda artar veya azalır. Bir öğrencinin ülke çapında yapılan deneme sınavlarında iki milyon öğrenci arasındaki sıralaması 100 000 (yüz bin) ise asıl sınavda ilk 100 (yüz) öğrenci arasına girmesi beklenemez. Dolayısıyla, verilerden birşeyler öğrenmek ve bilgi elde etmek günümüzün önemli meselelerinden bir tanesi haline gelmiştir. Makinelerin yardımıyla var olan veri kümelerinden anlamlı sonuçlar çıkarmak, kurumu ve bireyleri ciddi bir maliyetten kurtarır ve daha spesifik çözümler sunar. Kurulan modeller çok ekstrem şeyler olmadıkça belli bir süre kullanılabilir. Zaman ilerledikçe farklı parametrelerin ortaya çıkması, işin doğasının değişmesi ve teknolojinin getirdiği etkilerden dolayı modeli tekrar kurmak gerekir. Bu yüzden veri madenciliği bir süreçtir.

Model kurmak için veriyi bir ön işlemeden geçirmek gerekmektedir. Bu yapılmadığı taktirde iyi bir modelin ortaya konması çok zordur. Veri kümesinde boş veya gürültülü değerler olabilir. Bu problemlerin çözülmesi gerekmektedir. Sonucu olumlu manada etkileyeceği düşünülüyorsa başka bir veri kümesinden veri ithal edilebilir. Buna veri birleştirmesi denmektedir ve uygun formatta yapılması çok önemlidir. Değişkenlerin veri tipleri aynı olmalıdır. İstatistiki sonuçlar modele girecek değişkenleri belirlemede kullanılır. Bu sonuçlar sayesinde bazı değişkenlerin elenmesine, bazılarının dönüştürülmesine karar verilebilir. Bunların yanında türetilmiş değişken kullanmak önemlidir. Bazı değişkenler tek başına birşey ifade etmezken bir araya geldiklerinde

önemli bir değişken haline gelebilirler. Ayrıca veri ambarları sayesinde yeni değişkenler üretilerek modellenecek veri kümesinin içine koyulabilir. Gözetimli (supervised) bir öğrenmede ayrıca değişkenlerin hedef değişkeni açıklama güçlerini belirlemede korelasyon analizleri kullanılır. Ayrıca bağımsız değişkenlerin birbirleri arasındaki korelasyonu da incelemek gerekmektedir. Yüksek korelasyon katsayısına sahip bağımsız değişkenler, ortada çoklu bağlanım (multicollinearity) problemi olduğundan dolayı, modelde birbirlerinin etkilerini düşürebilir ve kötü bir model kurulmasına sebep olabilirler [49].

Veri madenciliğinde geliştirilmiş birçok model, yaklaşım, teknik ve algoritma bulunmaktadır. Yaklaşımlardan başlıcaları tahminleyici, tanımlayıcı ve buyrukçu modelleme yaklaşımlarıdır. Tahminleyici modellemede sınıflandırma ve regresyon yer almaktadır. Tanımlayıcı modellemede kümeleme ve birliktelik kuralları kullanılır. Buyrukçu modellemede ise tavsiyeci sistemler (recommender systems) kullanılır.

Veri madenciliği birçok sektörde ve alanda kullanılmaktadır. Bunlardan bir tanesi de sigortacılık sektörüdür. Ülkemizde sigortacılık sektöründe son yıllarda önemli gelişmeler yaşanmış, ekonomimizdeki yeri ve GSYH içindeki oranı artmıştır. Bu artışın temel sebebi hayat dışı sigortalarda yaşanan olumlu gelişmelerdir.

Bu çalışmada, motokaravan sigortacılığı konusunda bir modelleme yapılmıştır. Karavan ile gezi ve turizm, ülkemizde yaygınlaşmış olmasa da Almanya ve Büyük Britanya başta olmak üzere Avrupa'da oldukça popülerdir. Hollanda merkezli bilimsel veri bağışçısı bir şirketten alınan veri kümeleriyle yapılan çalışmada karar ağaçları, lojistik regresyon ve yapay sinir ağları tekniklerini kullanan algoritmalar koşulmuştur. Veriler 2000 yılına ait olsa da motokaravan sigortacılığıyla alakalı ciddi bir değişim olmadığı için elde edilen sonuçlar günümüzde de geçerliliğini korumaktadır. Modelin ön işleme süreci ve çeşitli istatistiki değerlerin bulunması için ABD merkezli SAS şirketinin masaüstü kurulumu olan SAS Enterprise Guide ve web tabanlı SAS Studio araçları kullanılmıştır. En iyi bağımsız değişkenler seçildikten sonra model kurma aşamasına geçilmiştir. Bundan sonraki çalışmalar Yeni Zelanda'da bulunan Waikato Üniversitesi'nin çalışmalarıyla ortaya çıkan Java ile yazılmış olan Weka aracıyla yapılmıştır. J48, lojistik regresyon ve MultilayerPerceptron algoritmalarının çalışması sonucunda sonuçlar elde edilmiştir. En

iyi doğruluk oranına sahip olan model MultilayerPerceptron algoritmasının çalışması sonucu elde edilmiştir. Algoritmanın test kümesi üzerinde çalışması sonucunda %84,25 doğruluk oranına ulaşılmıştır. J48 algoritması %81,2 doğruluk oranı ile ikinci en iyi sonuçları vermiştir. Lojistik regresyonda ise bu iki algoritmanın sonuçlarına göre daha kötü oranlar ortaya çıkmıştır.

Modellerin çalışma sürelerine bakıldığında J48 ve lojistik regresyon algoritmalarının çok kısa sürdüğü, MultilayerPerceptron algoritmasının ise çok uzun sürdüğü görülmektedir. Dolayısıyla, yapay sinir ağlarının kullanıdığı algoritmaların çalışma süresi, karar ağaçları ve lojistik regresyonu kullanan algoritmaların çalışma sürelerine göre oldukça uzundur. Bu yüzden daha maliyetli olduğu söylenebilir.

Model sonuçlarını değerlendirmede önemli noktalardan birisi de yorumlanabilirliktir. Karar ağaçlarının kolay bir şekilde yorumlandığı ve yapay sinir ağlarının ise yorumlanmasının kolay olmadığı görülmektedir. J48 algoritmasına bakıldığında motokaravan sigortası yaptırmaya etki eden değişkenlerin araba sigortasına ödenen prim miktarı, müşterinin evlilik durumu, ortalama gelirin seviyesi, eğitim seviyesi, toplumsal statü, yangın poliçesine ödenen prim miktarı, müşterinin satın alma gücü, müşterinin bir tane arabaya sahip olup olmadığı, müşterinin ev sahibi olup olmadığı ve müşterinin özel sağlık sigortasına sahip olup olmadığı gibi değişkenler olduğu tespit edilmiştir. Sigorta şirketleri motokaravan sigortası yaptırmak isteyen kişilere yönelik kampanya yapacakları zaman müşterilerin bu bilgilerine göre hareket edebilirler. Reklam panolarına verilecek olan reklamlara buna göre içerik hazırlanabilir. İnternette ve sosyal medya ortamında verilecek olan reklamlar bu özelliklere sahip kişilere ve bu kişilerin yoğunlukta yaşadığı bölgelere ve internet giriş saatlerine göre verilebilir. Ayrıca cep telefonu mesaj içerikleri ve kime gönderileceği gibi konular da bu değişkenlere sahip olan kişilere gönderilebilir. MultilayerPerceptron algoritmasında bu şekilde yorumlama yapılamamaktadır. Ancak doğruluk oranı yüksek olduğu için modelin seçtiği kişiler üzerine yoğunlaşılabilir. Veri madenciliğinde önemli olan noktalardan birisi de gözle görülemeyecek ilişkileri ortaya koymaktır. Yapay sinir ağları, yorumlaması güç olsa da ilişkileri doğru bir şekilde ortaya koyduğu için kullanımı her geçen gün artan bir tekniktir.

Sonuç olarak motokaravan sigortası alması muhtemel olan potansiyel müşteriler iyi bir doğruluk ve duyarlılık oranıyla belirlenmiştir. Bu yönüyle çalışma hedefine ulaşmıştır. Sonuçların ve yorumların sigortacılık sektörüne katkı sağlayacağı düşünülmektedir. Motokaravan sigortasının Avrupa'da olduğu gibi ülkemizde de gelişmesinin önünde herhangi bir engel yoktur. Doğal güzellikleri ve keşfedilecek yerleriyle ülkemiz çok önemli bir doğa kampüsüdür. Bunun değerlendirilmesi için yetkili kurumların ve sigorta şirketlerinin bu konuda yatırım yapması beklenmektedir.

Çalışmanın bundan sonraki aşamasında Türkiye'deki motokaravan sigortacılığı için bir çalışma yapılabilir. Yetkili kurumların tutacağı veriler sayesinde yapılacak böyle bir çalışma, ülkemizin sigortacılık sektörüne katkıda bulunacaktır.

KAYNAKLAR

[1] Kasap, E. (2007). "Sigortacılık Sektöründe Müşteri İlişkileri Yönetimi Yaklaşımıyla Veri Madenciliği Teknikleri Ve Bir Uygulama". Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi.

[2] Başer, F. (2013). "Aktüeryal Modellemede Bulanık Destek Vektör Makineleri". Fen Yayınlanmamış Doktora Tezi, Ankara Üniversitesi.

[3] Erol, B. (2013). "Müşteri İlişkileri Yönetimi İçin Veri Madenciliği Kullanılması ve Sigortacılık Sektörü Üzerine Bir Uygulama". Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi.

[4] Yüksek, A. G. (2007). "Hava kirliliği tahmininde çoklu regresyon analizi ve yapay sinir ağları yönteminin karşılaştırılması". Yayınlanmamış Doktora Tezi, Cumhuriyet Üniversitesi.

[5] Budak, H. ve Erpolat, S. (2012). "Kredi Riski Tahmininde Yapay Sinir Ağları ve Lojistik Regresyon Analizi Karşılaştırılması". AJIT-e: Online Academic Journal of Information Technology Fall/Güz 2012 – Vol./Cilt: 3 – Num./ Sayı: 9.

[6] Ibrahım, Z. ve Ruslı, D. (2007). "Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression". 21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur.

[7] Karakış, R. (2009). "Yapay Sinir Ağları Ve Lojistik Regresyon Yöntemleri İle Meme Kanseri Koltuk Altı Lenf Nodu Durumunun Belirlenmesi". Yayınlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi.

[8] Muslu, D. (2009). "Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması". Yayınlanmamış Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi.

[9] Çırak, G. ve Çokluk, Ö. (2013). "Yükseköğretimde Öğrenci Başarılarının Sınıflandırılmasında Yapay Sinir Ağları ve Lojistik Regresyon Yöntemlerinin

Kullanılması” , Mediterranean Journal of Humanities mjh.akdeniz.edu.tr III/2, 2013, s.71-79.

[10] Kuyucu, Y. E. (2012). “Lojistik Regresyon Analizi (Lra), Yapay Sinir Ağları (Ysa) Ve Sınıflandırma Ve Regresyon Ağaçları (C&Rt) Yöntemlerinin Karşılaştırılması Ve Tıp Alanında Bir Uygulama”. Yayınlanmamış Yüksek Lisans Tezi, Gaziosmanpaşa Üniversitesi.

[11] Kim, Y. S. (2006). “Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size” , CI Division, SK telecom, 11, Euljiro 2-ga, Jung-gu, Seoul, 100-999, Republic of Korea.

[12] Aras, Ü. (2008). “Finansal Veri Madenciliği”. Yayınlanmamış Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi.

[13] Akbilgiç, O. ve Keskintürk, T. (2008). “Yapay Sinir Ağları ve Çoklu Regresyon Analizinin Karşılaştırılması”. İstanbul Üniversitesi İşletme Fakültesi , Sayısal Yöntemler Anabilim Dalı Yönetim Yıl: 19 Sayı: 60, Haziran 2008.

[14] Köktürk, F., Büyükuysal, M. Ç. ve Sümbüloğlu, V. (2012). “K - En Yakın Komsuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması” , Uluslararası Katılımlı XIV. Ulusal Biyoistatistik Kongresi, 4-7 Eylül 2012 Kadir Has Kongre Merkezi ,Kayseri.

[15] Çimenli, S. (2015). “Churn Analysis And Prediction With Decision Tree And Artificial Neural Network”. Yayınlanmamış Yüksek Lisans Tezi, Kadir Has Üniversitesi.

[16] Er, O. (2016). “Yapay sinir ağları, lojistik regresyon ve karar ağaçları uygulamaları ile kariyer başarısı tahmini Akademisyenler üzerine bir araştırma”. Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü İşletme Anabilim Dalı.

[17] Çinko, M. (2006). “Kredi Kartı Değerlendirme Tekniklerinin Karşılaştırılması” , İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi Yıl:5 Sayı:9 Bahar 2006/1 s.143-153.

[18] Akpınar, H., Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği, İ.Ü. İşletme Fakültesi Dergisi, C:29, S:1 / Nisan 2000.

[19] Ergün, K., Veri Madenciliğine Giriş (Introduction To Data Mining) (b.t), 23.03.2017, http://kergun.baun.edu.tr/veri_madenciligi_hafta1.pdf

[20] Dolgun, Ö., Veri Madenciliği ve Tarihçesi, http://ozgurdolgun.com/?p=19, 13.08.2015.

[21] Şeker, Ş. E. (2014). Sosyal Ağlarda Akan Veri Madenciliği. YBS Ansiklopedisi, 1(3). S. 21.

[22] Sas Big Data, https://www.youtube.com/watch?v=-Gj93L2Qa6c, 15.02.2017.

[23] Lane, P., Oracle9i Data Warehousing Guide Release 2 (9.2),

https://docs.oracle.com/cd/B10500_01/server.920/a96520/concept.html , 2002.

[24] Alpaydın, E. (2000). Zeki Veri Madenciliği : Ham Veriden Altın Bilgiye Ulaşma Yöntemleri, Bilişim 2000 Eğitim Semineri.

[25] Özmen, S. (2003). Ag Ekonomisinde Yeni Ticaret Yolu: E-Ticaret, İstanbul Bilgi Üniversitesi Yayınları.

[26] Data Mining Using SAS Enterprise Miner : A Case Study Approach, Second Edition. https://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf, s.4-16.

[27] Han,J., Kamber, M. ve Pei, J., Data Mining Concepts and Techniques (3rd Edition), Elsevier, Waltham,2012. s.40-43

[28] Korelasyon, https://tr.wikipedia.org/wiki/Korelasyon, 10.01.2017.

[29] Argüden, M. ve Erşahin, B., Veri Madenciliği Veriden Bilgiye, Masraftan Değere. s.24-25.

[30] Silahtaroğlu, G., "Kavram ve Algoritmalarıyla Temel Veri Madenciliği", İstanbul, Papatya Yayıncılık, 2008.

[31] Vapnik, V.N., The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[32] Osowski, S., Siwekand, K., and Markiewicz, T. (2004). MLP and SVM Networks – a Comparative Study. Proceedings of the 6th Nordic Signal Processing Symposium – NORSIG.

[33] Khan, M., K-Nearest Neighbor Classification on Spatial Data Streams Using P-Trees, 6. Pasifik Asya Knowledge Discovery and Data Mining Konferansı PAKKDD'02, Taiwan, 2002, s.517-518.

[34] Goldberg, D.E.,1989, Genetic Algorithms in Search, Optimization, and Machine Learning, New York: Addison Wesley.

[35] Angeline, P.J., 1995, "Evolution revolution: An introduction to the special track on genetic and evolutionary programming," IEEE Expert Intelligent Systems and their Applications 10, June pp.6-10.

[36] Emel, G. G., Taşkın, Ç. 2005. "Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması," Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, cilt 6, s. 221-239.

[37] Yapay Sinir Ağlarının Tarihçesi, erişim tarihi: 14.10.2004. http://backpropagation.netfirms.com/tarihce.htm

[38] ÖZTEMEL, E., Yapay Sinir Ağları, Papatya Yayıncılık, 3. Basım, İstanbul, 2012, s.45-59.

[39] Mertler, C. A., & Vannatta, R. A. (2005). Advanced and multivariate statistical methods: Practical application and interpretation (3rd ed.). Glendale, CA: Pyrczak Publishing.

[40] Field, A. (2005). Discovering statistics using SPSS (2nd ed.). London: Sage.

[41] Johnson, R. A. And D. W. Wichern (1988). Applied Multivariate Statistical Analysis: (2nd Ed.) Prentice Hall, Englewood Cliffs, New Jersey

[42] Agrawal R. ve Srikant R., (1995), "Mining Sequential Patterns", 11th International Conference on Data Engineering, Taipei, Taiwan, 3-14.

[43] Prof. Dr. Kamuran Pekiner , "Sigorta İşletmeciliğinin Prensipleri (Hesap Bünyesi Alt Başlık)",Üçüncü baskı ,Istanbul ,Formül Matbaası, 1981, s.17.

[44] Sigortanın Tarihi, http://www.tsb.org.tr/sigortanin-tarihi.aspx?pageID=438, 03.03.2017.

[45] T.C Başbakanlık Hazine Müsteşarlığı Sigortacılık Ve Bireysel Emeklilik Sektörleri 2015 Yılı Faaliyet Raporu Özet Bilgiler.

[46] Weka, http://www.cs.waikato.ac.nz/ml/weka/, 10.12.2016.

[47] Arff (book version), https://weka.wikispaces.com/ARFF+%28book+version%29, 13.12.2016.

[48] Sas Hakkında, https://www.sas.com/tr_tr/company-information.html#stats, 29.12.2016.

[49] Larose, D. T., Data Mining Methods and Models, John Wiley & Sons, Inc., New Jersey,2006, s. 117.

[50] Yapay Sinir Ağları Nedir, http://kod5.org/yapay-sinir-aglari-ysa-nedir/, 10.11.2016.

[51] Preface, http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/Preface.html, 13.05.2016.

[52] Elkan, C., 2000, CoIL Challenge 2000 Entry, University of California, San Diego, Department of Computer Science and Engineering 0114.

[53] Kim, Y. ve Street W. N., 2000, CoIL Challenge 2000: Choosing and Explaining Likely Caravan Insurance Customers, University of Iowa, Management Sciences Department.

ÖZGEÇMİŞ

1988 yılında Manisa'da doğdum. İlköğretim eğitimimi Manisa'da, lise eğitimimi Balıkesir'de tamamladım. 2005 yılında başladığım Boğaziçi Üniversitesi Fen Edebiyat Fakültesi Matematik bölümününden 2011 yılında mezun oldum. 2011 yılında İstanbul Teknik Üniversitesi Sosyal Bilimler Fakültesi İktisat bölümde yüksek lisansa başladım. Dersleri bitirdim ve tez aşamasındayım. Diğer taraftan İstanbul'da özel bir şirkette iş analisti olarak işe başladım. 2015 yılından beri farklı bir şirkette uzman veri analisti olarak çalışıyorum.

Aday: Yasin KAYA