



**VERİ MADENCİLİĞİ YAKLAŞIMI İLE PROJE BAŞVURU YAPMA
EĞİLİMİNİN TAHMİNİ**

Ayten MERMER

**YÜKSEK LİSANS TEZİ
ENDÜSTRİ MÜHENDİSLİĞİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

EKİM 2019

Ayten MERMER tarafından hazırlanan “VERİ MADENCİLİĞİ YAKLAŞIMI İLE PROJE BAŞVURU YAPMA EĞİLİMİNİN TAHMİNİ” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Endüstri Mühendisliği Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Dr. Öğr. Üyesi Murat ARIKAN

Endüstri Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

.....

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Başkan: Prof. Dr. Serpil EROL

Endüstri Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

.....

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Üye: Doç. Dr. Babek ERDEBİLLİ

Endüstri Mühendisliği Ana Bilim Dalı, Ankara Yıldırım Beyazıt Üniversitesi

.....

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

Tez Savunma Tarihi: 01/10/2019

Jüri tarafından kabul edilen bu çalışmanın Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum

.....

Prof. Dr. Sena YAŞYERLİ
Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

.....
Ayten MERMER
01/10/2019

VERİ MADENCİLİĞİ YAKLAŞIMI İLE PROJE BAŞVURU YAPMA EĞİLİMİNİN TAHMİNİ

(Yüksek Lisans Tezi)

Ayten MERMER

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Ekim 2019

ÖZET

Fazla miktardaki verinin; faydalı ve anlamlı bir takım kural ve örüntü çıkarılması amacıyla analiz edilmesine veri madenciliği denir. Veri madenciliği sürecinde, verinin yapısına ve uygulamanın amacına göre pek çok yöntem kullanılmaktadır. Karar ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri, veri madenciliğinde sınıflandırma ve tahmin için sıkça kullanılan yaklaşımlardandır. Bu çalışmada, proje fonlayan bir kuruluşa proje sunma potansiyeli olan araştırmacıların, veri tabanında kayıtlı yaş, cinsiyet, uzmanlık alan, eğitim bilgileri (lisans, yüksek lisans, doktora) ve çalıştığı kurum bilgileri ele alınarak bu kişilerin proje başvurusu yapma eğilimleri veri madenciliğinin sınıflandırma yöntemlerinden olan Karar ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri ile incelenmiştir. Analiz sonuçlarına göre, üç farklı algoritma da benzer ve yakın sonuçlar vermesine rağmen, Destek Vektör Makineleri ile yapılan sınıflandırmanın doğruluk oranı en yüksek bulunmuştur. Verilerin analizi SPSS programı veri madenciliği modülü IBM-Modeler ile gerçekleştirilmiştir.

Bilim Kodu : 90618

Anahtar Kelimeler : Veri Madenciliği, Karar Ağaçları, Yapay Sinir Ağları, Destek Vektör Makineleri

Sayfa Adedi : 69

Danışman : Dr. Öğr. Üyesi Murat ARIKAN

PREDICTION OF PROJECT PROPOSAL TENDENCY WITH DATA MINING

APPROACH

(M. Sc. Thesis)

Ayten MERMER

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

October 2019

ABSTRACT

Data Mining is the analysis of large amounts of data in order to derive meaningful and useful patterns and rules from the data. In Data Mining, many techniques are used according to the data structure and purpose of application. Decision trees, Artificial Neural Networks and Support Vector Machines are frequently used approaches for classification and forecasting in data mining. In this study, the information of age, gender, scientific expertise, education information (undergraduate, graduate, doctorate) and the institution information of researchers with the potential to present a project to the project funding institution are discussed and tendencies of these researchers to apply to various types of projects has been analyzed using the Decision Trees, Artificial Neural Networks and Support Vector Machines, which are the classification methods of data mining. According to the analysis results, although three different algorithms give similar and close results, the accuracy rate of the classification with Support Vector Machines is the best. Data analysis is performed with the SPSS IBM-Modeler data mining module.

Science Code : 90618

Key Words : Data Mining, Decision Trees, Artificial Neural Networks, Support Vector Machines

Page Number : 69

Supervisor : Assist. Prof. Dr. Murat ARIKAN

TEŞEKKÜR

Bu tez çalışmasının hazırlanması süresince ilgi ve desteğini esirgemeyerek bana yol gösteren tez danışmanım Dr. Öğr. Üyesi Murat ARIKAN'a, çalışmanın geliştirilmesi hususunda engin bilgi birikimi ve değerli tavsiyelerinden faydalandığım Prof. Dr. Serpil EROL ve Doç. Dr. Babek ERDEBİLLİ'ye, çalışmalarım boyunca bana sabırla destek olan tüm arkadaşlarıma, eğitim hayatım boyunca güven ve desteklerini bir an esirgemeyen sevgili aileme en içten duygularıyla teşekkür ederim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	x
SİMGELER VE KISALTMALAR.....	xi
1. GİRİŞ.....	1
2. LİTERATÜR TARAMASI	3
3. VERİ MADENCİLİĞİ	9
3.1. Veri Madenciliğinin Tanımı ve Amacı.....	9
3.2. Veri Madenciliğinin Uygulama Alanları.....	10
3.3. Veri Madenciliği Süreci.....	12
3.4. Veri Madenciliği Sürecini Etkileyen Faktörler.....	14
3.5. Veri Madenciliği Sürecinde Karşılaşılan Sorunlar	15
3.6. Veri Madenciliği Modelleri	16
3.6.1. Kümeleme (Clustering)	16
3.6.2. Birliktelik kuralları (Association rules) ve sepet analizi	17
3.7. Sınıflandırma (Classification).....	19
3.7.1. Karar ağaçları	20
3.7.2. Yapay sinir ağları.....	25
3.7.3. Destek vektör makineleri.....	31

4. VERİ MADENCİLİĞİ YÖNTEMLERİ İLE PROJE BAŞVURU YAPMA EĞİLİMİNİN İNCELENMESİ.....	35
4.1. Problemin Belirlenmesi ve Çalışmanın Amacı	35
4.2. Verinin Tanımı	36
4.3. Veri Ön İşleme.....	37
4.4. Yöntemin Belirlenmesi ve Modelin Oluşturulması.....	39
4.5. Tahmin Modellerini Etkileyen Değişkenlerin Belirlenmesi.....	43
5. BULGULAR.....	47
5.1. Karar Ağaçları Yöntemi ile Analiz.....	47
5.2. Yapay Sinir Ağları Yöntemi ile Analiz	50
5.3. Destek Vektör Makineleri Yöntemi ile Analiz	51
5.4. Tahmin Modellerinin Doğruluk Oranlarının Karşılaştırılması.....	52
6. SONUÇ VE ÖNERİLER.....	55
KAYNAKLAR	57
EKLER.....	63
EK-1. Uygulamada kullanılan ham veriye ilişkin örnek	64
EK-2. Girişimci ve yenilikçi üniversite endeksi 2017 yılı sıralaması.....	65
EK-3. Yapay sinir ağı modeli tahmin sonuçlarına ilişkin örnek	66
EK-4. Destek vektör makineleri ile oluşturulan modelin parametre değerleri.....	67
ÖZGEÇMİŞ	69

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. Çalışma konusuna uygun literatür özeti.....	7
Çizelge 3.1. Toplama fonksiyonu örnekleri.....	28
Çizelge 4.1. Ham veriden bir kesit	37
Çizelge 3.2. Aktivasyon fonksiyonu örnekleri.....	29
Çizelge 5.1. DVM model özeti	51
Çizelge 5.2. Modellerin performansına ilişkin özet çizelge.....	52

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 3.1. Veri madenciliği ile etkileşimli disiplinler	10
Şekil 3.2. Veri madenciliğinde izlenen süreç.....	12
Şekil 3.3. Karar ağacı yapısı	21
Şekil 3.4. Yapay sinir ağı yapısı	27
Şekil 3.5. DVM ayırıcı düzlem.....	31
Şekil 3.6. Doğrusal ayrılabilen ve ayrılamayan veri kümeleri.....	32
Şekil 4.1. Veri ön işleme algoritması	39
Şekil 4.2. Veri setine uygun sınıflandırma yöntemleri	40
Şekil 4.3. Modelde kullanılan bağımsız değişkenler	41
Şekil 4.4. Verinin bölünmesi (partition)	41
Şekil 4.5. Modellerin girdi ve hedef değişkenleri.....	42
Şekil 4.6. Oluşturulan modellere ilişkin ekran görüntüsü.....	43
Şekil 4.7. Karar ağacı modeline etkisi olan değişkenler	44
Şekil 4.8. YSA modeline etkisi olan değişkenler	45
Şekil 4.9. DVM modeline etkisi olan değişkenler	46
Şekil 5.1. Karar ağacı sonuç özeti.....	48
Şekil 5.2. Karar ağaçları modeline ilişkin doğruluk oranı	49
Şekil 5.3. YSA modeli özeti	50
Şekil 5.4. YSA ağ mimarisi	50
Şekil 5.5. YSA modeline ilişkin doğruluk oranı	51
Şekil 5.6. DVM modeline ilişkin doğruluk oranı	52

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklamalar

%

Yüzde

Kısaltmalar

Açıklamalar

CHAID

Chi-Squared Automatic Interaction Detector

C&RT

Classification and Regression Trees

DVM

Destek Vektör Makineleri (Support Vector Machines)

ID3

Iterative Dichotomiser 3

SLIQ

Supervised Learning in Ques

TÜBİTAK

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu

YSA

Yapay Sinir Ağları (Artificial Neural Networks)

1. GİRİŞ

Veri madenciliği, fazla miktardaki verinin; faydalı ve anlamlı bir takım kural ve örüntü çıkarmak amacıyla analiz edilmesi işlemidir. Çoğu zaman veri tabanlarında kayıtlı olan ham veri, karmaşıklığı ve büyüklüğü nedeniyle karar vericilere bir destek sağlayamamakta, bu veriyi çözümleyecek tekniklere ihtiyaç duyulmaktadır. Verilerin analiz edilmesinde istatistik, pek çok problemi çözmede önemli bir araçtır. Ancak, bazı karmaşık ve büyük verilerde istatistiğin kullanımı sınırlıdır. Bu gibi durumlarda, tanımlayıcı ve tahmin edici pek çok analiz yöntemlerini kapsayan veri madenciliğine gereksinim ortaya çıkmaktadır. Bu sebeple, ham veri yığınlarını, faydalı bilgi haline dönüştürerek anlam kazandıran veri madenciliği teknikleri son yıllarda büyük önem kazanmıştır.

Hizmet sektöründe faaliyet gösteren kuruluşların geniş veri kayıtları olmasına karşın, karar verme sürecinde bu veriden faydalı bilgi elde edilmesi noktasında yetersiz kalmaktadır. Bu tür kuruluşlarda istatistiksel çalışmalar yoğun olarak kullanılmakla birlikte, veri yığınlarında keşfedilmeyi bekleyen sayısız örüntü ve kuralları keşfetmede istatistik yetersiz kalmaktadır. Bu noktada çözüm olarak veri madenciliği önemli bir boşluğu doldurmaktadır. Verilerin nitelikli bir şekilde analiz edilmesi ve içerisinde gizlediği bilgilerin ortaya çıkarılması veri madenciliği ile mümkün olabilmektedir. Literatürde, kapsamlı veri yığınlarını tasnif ederek, önemli özellikleri tespit eden veya geleceğe ilişkin bir sonucu tahmin eden, veri madenciliğine ilişkin pek çok uygulama yer almaktadır. Bu örneklerin araştırılması ve özetlenmesi, ileride benzer özellikte problemlere çözüm önerisi oluşturma noktasında büyük önem arz etmektedir. Bu bilgiler ışığında, proje desteği veren bir kuruluşa başvuru yapan ve yapma potansiyeli olan kişilerin kapsamlı verilerinden, sınıflandırma yöntemi ile kuruluşun gerekli birtakım çalışmalarında kullanabileceği sonuçlar elde edilmiştir (Mermer, 2016).

Veri madenciliğinde, aynı amaca yönelik farklı sonuçlar üreten çok sayıda algoritma bulunmaktadır. Veri setinin özelliklerine göre en uygun algoritmayı bulmak ve yüksek performans gösteren modeli kurmak, sürecin en önemli adımlarındandır. Aynı veri setine hitap eden birden fazla algoritma bulunduğu durumlarda, en başarılı sonucu elde etmek için her bir algoritma ile model kurularak kıyaslama yapılması gerekmektedir. Bu tez çalışmasında, benzer çalışmalar dikkate alınarak; aynı amaç için kullanılan farklı

algoritmalar ile modeller oluşturulmuş ve model performanslarının kıyaslaması yapılmıştır. Bu sayede kullanılan veri setine uygun en iyi sınıflandırmayı yapabilecek algoritmayı tespit etmek amaçlanmıştır.

Veri madenciliği çalışmalarında, uygun veri setinin temin edilmesi çalışmaların önemli bir bölümünü oluşturmaktadır. Çalışmalarda uygulanan yöntemler ve süreç benzer olsa da, üzerinde çalışılan verinin niteliği ve niceliği; analiz sonuçlarının gerçek bir problemin çözümünde faydalı ve kullanışlı olması, veri madenciliği çalışmalarının önemini artırmaktadır. Kullanılan veri setinin, sağlıklı analiz yapmayı kolaylaştıracak büyüklükte ve çeşitlilikte olması, sonuçların ilgili kuruluşun politika düzenleyici ve karar verici birimleri tarafından kullanılabilir olması bu çalışmanın önemini ve gerekliliğini ortaya koymaktadır.

Tez çalışması kapsamında ikinci bölümde, kapsamlı bir literatür araştırması yapılarak veri madenciliğinin uygulama alanları, kullanılan yöntemler ve benzer problemlere çözüm oluşturmak amacıyla yapılan çalışmalar incelenmiş ve özet şekilde sunulmuştur. Üçüncü bölümde ise yapılan literatür taraması doğrultusunda veri madenciliğine ilişkin bilgiler sunularak izlenecek yöntem hakkında genel bilgi verilmiştir. Dördüncü bölümde, mevcut bir problemin çözümüne ilişkin farklı yöntemlerle kurulan modellerden bahsedilmiştir. Beşinci bölümde, uygulama bölümünde kurulan modellerin analiz sonuçlarına ve kıyaslamalara yer verilmiştir. Son bölüm olan sonuç ve öneriler kısmında ise tez çalışması yorumlanmış ve belirlenen problemin çözümüne sağladığı katkı ele alınmıştır.

2. LİTERATÜR TARAMASI

Veri madenciliği tekniklerinin 1990'lı yıllarda ortaya çıkmış olmasına rağmen, yapılan literatür taramasında 2000'li yıllardan itibaren kapsamlı çalışmalar yapıldığı tespit edilmiştir.

Özellikle uygulama yapılacak veri kümesine benzer verilerle yapılan çalışmalara yönelik literatür incelemesi sonucunda konuya ilişkin çalışmalar aşağıda özetlenmiştir:

Liao, Chu ve Hsiao (2012) ile Mukherjee ve diğerleri (2015) çalışmalarında veri madenciliğinin yaygınlaşmasından itibaren yapılan tüm uygulamaları inceleyerek, ayrıntılı bir literatür özeti oluşturmuşlardır. Araştırma sonuçlarını; uygulama alanlarına ve kullanılan tekniklere göre sınıflandırmışlardır.

Ngai (2009) ise benzer çalışmayı müşteri ilişkileri yönetimi konusuna ilişkin gerçekleştirmiştir. Müşteri ilişkileri yönetiminde literatürde yer alan bazı çalışmalarda sınıflandırma aracı olarak kullanılan yöntemler irdelenmiş ve tasnif edilmiştir. Bu özet niteliğindeki araştırmalar, konuya ilişkin yapılacak benzer çalışmalara ışık tutacak nitelikte karşılaştırmalar ve örnekler içermektedir.

Savaş, Topaloğlu ve Yılmaz (2012) yaptıkları çalışmada veri madenciliğinin uygulama alanlarına ayrıntılı bir biçimde değinmişlerdir. Türkiye'de veri madenciliği üzerine yapılan çalışmalar ve gerçekleştirilen uygulamalar incelenmiş ve özetlenmiştir. Bu çalışma, veri madenciliği sürecinin uygulama alanlarının yaygınlaşması noktasında önem arz etmektedir.

Kuzey (2012) doktora tezi kapsamında yaptığı çalışmada, bilgi çalışanlarının kurum performansına olan etkilerini ölçmek amacıyla karar ağaçları ve destek vektör makineleri yöntemleri üzerinde çalışmıştır. Çalışmada, anket yoluyla elde edilen geniş kapsamlı verinin sınıflandırma işleminde C&R karar ağacı ve destek vektör makineleri yöntemleri doğruluk oranı oldukça yüksek sonuçlar vermiştir. Karar ağacı yönteminin görsel sonuçlarının daha sade ve kolay yorumlanabilir olduğu, bununla birlikte destek vektör makineleri yönteminin de istatistik metodları ve makine öğrenmesini birleştiren etkili bir sınıflandırma yöntemi olduğu tespit edilmiştir.

Ding ve arkadaşları (2008) yaptıkları çalışmada finansal durum tahmininde destek vektör makinesi tabanlı bir algoritma geliştirmişlerdir. Çalışma neticesinde, önerilen modelin geleneksel istatistiksel yöntemlerden ve geri beslemeli Yapay Sinir Ağlarından daha başarılı olduğu sonucuna varılmıştır. Oluşturulan modelin daha kapsamlı veri kümelerine de uygulanabilirliği tartışılmıştır.

Raudys (2000), çalışmasında sınıflandırma yöntemi olarak yaygın bir şekilde kullanılan Destek Vektör Makinelerinin, gerçek hayatta yer alan problemlerin çözümünde kullanılabilirliğini tartışmıştır. Doğrusal veri setleri üzerinde etkili ve pratik bir sınıflayıcı olan Destek Vektör Makinelerinin; doğrusal olmayan verilere de birtakım veri dönüştürme algoritmaları sonrasında uygulanabileceği sonucuna varılmıştır.

Huang ve arkadaşları (2004), literatürde pek çok araştırmaya konu olan kredi derecelendirme analizi üzerine yapmış oldukları çalışmalarında, 21 bağımsız değişken içeren geniş bir veri kümesini sınıflandırmak amacıyla geri beslemeli yapay sinir ağları ile destek vektör makineleri yöntemini kıyaslamıştır. Çalışma sonucuna göre her iki yöntemin de doğruluk seviyesinin %80'in üzerinde olmasına rağmen, destek vektör makineleri yöntemi ile bir miktar daha fazla iyileşmenin sağlandığı belirtilmiştir. Ayrıca, yöntemler tarafından belirlenen değişken öneminin sıralamasındaki farklılığın nedenlerini araştırmak üzere yapılacak daha detaylı çalışmaların gerekliliği vurgulanmıştır.

Vadim (2018) çalışmasında problemlerin çözümünde veri madenciliğinden nasıl faydalanılacağına ilişkin önerilerde bulunmuş ve büyük veri kümelerindeki örüntüleri keşfetmek, tahminlerde bulunabilmek amacıyla veri madenciliği yöntemlerini sınıflandırmıştır. Benzer şekilde Wang ve diğerleri (2018) çoklu veri kaynaklarında veri madenciliğinin nasıl uygulanması gerektiği ve veriye uygun veri madenciliği yönteminin seçilmesine ilişkin analizler yapmıştır.

Huang ve arkadaşları (2004) döviz kurlarının tahmini üzerine yaptıkları çalışmalarında, Yapay Sinir Ağları yöntemini diğer tahmin yöntemleri ile karşılaştırmışlardır. Aynı yöntem kullanılan benzer çalışmaları inceleyerek doğruluk oranını etkileyen faktörleri analiz etmişlerdir. Çalışma sonucunda Yapay Sinir Ağlarının performansını etkileyen faktörlerin; bağımlı değişkenlerin seçimi, yapay sinir ağı mimarisi ve veri ön işleme süreci olduğu belirlenmiştir. Ayrıca çalışmada, Yapay Sinir Ağları ve Genetik Algoritmalar ile hibrit bir

yöntem geliştirilmiştir. Geliştirilen bu yeni yöntemin, yalnız Yapay Sinir Ağları veya Genetik Algoritma ile yapılan tahminlerden daha yüksek başarılı sonuçlar verdiği tespit edilmiştir.

Liao ve Wen (2007) çalışmalarında yapay sinir ağları üzerine 1995 yılından 2005 yılına kadar yapılmış olan çalışmaları özetleyen bir literatür taraması sunmuşlardır. Sınıflandırma ve kümeleme yöntemi olarak yaygın şekilde kullanılan Yapay Sinir Ağları metodolojisini ve uygulama alanlarını çalışmalar üzerinden özetlemişlerdir.

Tolle ve arkadaşları (2000) sağlık alanında gerçekleştirdikleri uygulamada, daha önce farklı yazılım araçları ile yapılan onkolojik hastalıkların erken teşhisini yapay sinir ağı modeli kullanarak gerçekleştirmişlerdir. Kan seviyelerine ilişkin sınıflandırma yapılan çalışmada, ileri beslemeli ve geri beslemeli yapay sinir ağı mimarisinden faydalanılmıştır. Kullanıcı dostu, az maliyetli ve etkin bir analiz yöntemi oluşturmak amaçlanmıştır. Yapay sinir ağı yönteminin bu amaca uygun, maliyeti az ve daha çok kullanıcıya hitap eden bir tahmin aracı olduğu çalışmada ispatlanmıştır.

Chang ve Chen (2009) sağlık alanında yaptıkları çalışmalarında, cilt kanserinin oluşmasına etki eden faktörleri belirlemek üzere karar ağaçları ve yapay sinir ağlarını sentezleyen bir algoritma geliştirmişlerdir. Çalışma sonucunda, yapay sinir ağlarının da kullanıldığı algoritmanın, yalnızca karar ağaçları ile sınıflandırmaya göre daha iyi sınıflandırma sağladığı tespit edilmiştir. Çalışma sonucunda, diğer sınıflandırma yöntemleri ile kombine edilerek oluşturulacak yeni algoritmaların sınıflandırma başarısının daha da iyi olabileceği tahmin edilmiştir.

Albayrak ve Yılmaz (2009) yaptıkları çalışmada İMKB verilerine karar ağacı sınıflandırma yöntemini uygulamış ve hizmet sektörü üzerine faaliyet gösteren firmaları birbirinden ayıran etkili değişkenleri belirlemiştir.

Du (2010) çalışmasında, sınıflandırma problemlerinde makine öğrenmesine dayalı sinir ağı algoritmalarının kullanımının da klasik istatistiksel sınıflandırma yöntemleri gibi yaygınlaştığını belirtmiştir. Makalesinde, makine öğrenmesine dayalı sınıflandırma yöntemlerine ilişkin matematiksel formüller tanımlanmıştır. Algoritmaların temelini

oluşturan matematiksel modellerin anlaşılması sonucunda, amaç fonksiyonuna en uygun modelin kolaylıkla kurulabileceği belirtilmiştir.

Li ve Wu (2010) çalışmalarında, finansal başarısızlık tahmininde sınıflandırma yöntemlerinden karar ağaçları, k-en yakın komşu algoritması ve destek vektör makineleri ve regresyon analizi yöntemlerini kullanarak karşılaştırma yapmışlardır. Analiz sonuçlarına göre, sınıflandırmanın performansını ölçen farklı değerlendirme kriterlerine göre yöntemlerin başarısının değiştiği tespit edilmiştir. Destek vektör makineleri ve regresyon analizi yöntemlerinin seçilen iki değerlendirme kriterine göre en iyi sınıflandırmayı yaptığı sonucuna ulaşılmıştır.

Çalış, Kayapınar ve Çetinyokuş (2014) çalışmalarında bilgisayar ve internet üzerine bir anket düzenlemiş ve karar ağaçları yardımıyla farklı demografik özellikteki kişiler için çıkarımlar yapılmıştır. Ayrıca çalışmada; karar ağacı algoritmaları karşılaştırılmış ve veri için doğruluk oranı en yüksek algoritma kullanılmıştır.

Cruz (2018) ve Yalçın (2019) gerçekleştirdikleri çalışmalarda sağlık sektöründe veri madenciliği tekniklerini uygulamış, sınıflandırma yöntemlerini ve karar ağacı yapısını kullanarak hastalıkların teşhisine yönelik tahmin modeli geliştirmişlerdir.

Li ve arkadaşları (2019) kentsel boru şebekelerinin güvenilirlik değerlendirilmesi üzerine yaptıkları çalışmalarında, destek vektör makinaları ve yapay sinir ağlarının sınıflandırma özelliklerinden faydalanmışlardır. Çalışmada, yapay sinir ağları ile model kurma sürecinin destek vektör makinelerine göre çok daha karışık ve uzun olmasına rağmen, yapay sinir ağlarının farklı veri türlerine daha fazla uyum sağlayabildiği tespit edilmiştir. İki farklı algorithmada da benzer ve yakın sonuçlar elde edilmekle birlikte, destek vektör makineleri ile kurulan modelin performansının daha iyi olduğu belirtilmiştir.

Veri madenciliği uygulamalarına ilişkin literatürde yer alan bazı diğer çalışmalar uygulama alanlarına göre sınıflandırılarak çizelge 2.1’de özetlenmiştir:

Çizelge 2.1. Çalışma konusuna uygun literatür özeti

Çalışma	Uygulama Alanı	Kullanılan Yöntem
Rygiels vd. (2012) Kim vd. (2006) Baesens vd. (2002) Emel ve Taşkın (2005) Wu vd. (2005) Zhang vd. (2005) Ahn vd. (2006) Lee ve Park (2005) Hosseini vd. (2010) Chen vd. (2005)	Pazarlama	Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Bayesyen Sınıflama) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Genetik Algoritma) Sınıflandırma (Karar Ağaçları) Kümeleme (K-Means Aloritması) Birliktelik Kuralları
Chen ve Hsu (2006) Zhong vd. (2005) Chang vd. (2011) Yıldırım vd. (2008) Breault vd. (2003)	Tıp	Sınıflandırma (Yapay Sinir Ağları) Kümeleme (K-Means Algoritması) Sınıflandırma (Regresyon Analizi) Sınıflandırma (Karar Ağaçları) Birliktelik Kuralları
Hsieh (2004) Mues vd. (2004) Kim (2006) Lee vd.(2006) Albayrak ve Yılmaz (2009) Katılmış (2018)	Bankacılık ve Finans	Birliktelik Kuralları Sınıflandırma (Karar Ağaçları) Sınıflandırma (Yapay Sinir Ağları) Sınıflandırma (Regresyon Analizi) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Regresyon Analizi)
Kantardzic (2002) Carvalho ve Freitas (2004) Zhang vd. (2004) Rokack ve Maimon (2008) Kuzey (2012) Çalış vd. (2014)	Mühendislik	Sınıflandırma (Destek Vektör Makineleri) Sınıflandırma (Genetik Algoritma) Birliktelik Kuralları (Apriori Algoritması) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Destek Vektör Makineleri) Sınıflandırma (Karar Ağaçları)
Özçınar (2006) Horner vd. (2010) Çizel (2018) Sheybani (2019) Alban ve Mauricio (2018)	Eğitim	Sınıflandırma (Regresyon Analizi) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları) Sınıflandırma (Karar Ağaçları)

3. VERİ MADENCİLİĞİ

Bir konuya ilişkin, tartışma, araştırma, bilgi edinme ve akıl yürütme sonucunda oluşan işlenmemiş, henüz yorum yapmaya imkan sağlayacak düzeyde sistemleştirilmemiş ham bilgi veri olarak adlandırılır. Veriler, gözlem ve kaydedilen işlemler yoluyla elde edilir. Elde edilen veriler üzerinde ön işleme ve analiz gibi işlemler uygulanarak, bu verilerin yorumlanabilecek ve raporlanacak hale gelmesi sağlanır. Veri madenciliği, genel anlamıyla, büyük ölçekli verilerin arasından faydalı bilgiye ulaşma, bilgiye ulaşmak amacıyla çalışmalar yapma sürecidir. Bu bölümde, veri madenciliğine ilişkin tanımlar ve veri madenciliği sürecine ilişkin bilgiler sunulmuştur.

3.1. Veri Madenciliğinin Tanımı ve Amacı

Büyük veri yığınları içerisinde gelecekle ilgili tahminde bulunabilmeye olanak sağlayan bağıntıların, bilgisayar programları ve algoritmalar aracılığıyla aranmasına veri madenciliği denir. Veri madenciliği tanımına benzer olarak, veri tabanlarında bilgi madenciliği (knowledge mining from databases), veri ve örüntü analizi (data/pattern analysis), bilgi çıkarımı (knowledge extraction), veri arkeolojisi gibi bazı kavramlar da literatüre geçmiştir. Veri madenciliği genel anlamıyla, eldeki verilerden örtülü, daha önce net olarak belirlenmemiş, önceden bilinmeyen fakat potansiyel olarak faydalı ve kullanışlı bilginin ortaya çıkarılmasıdır. Başka bir anlam olarak veri madenciliği, ham verilerin içerisinde yer alan desenlerin, kuralların, değişimlerin, ilişkilerin, düzensizliklerin ve istatistiksel anlam içeren yapıların veri madencisi ve bilgisayar yoluyla keşfedilmesidir. Veri madenciliği temel olarak, veri kümeleri arasında olan örüntülerin, uygun algoritmalar ile aranmasıyla ilgilenir. Veri kümeleri arasında yer alan ilişki, kurallar ya da özellikler bilgisayar programları aracılığıyla tespit edilir.

Veri Madenciliğinde amaç, daha önceden fark edilmemiş veri desenlerini tespit edebilmektir. Daha önce toplanan bilgilerin bir takım istatistik yöntemler kullanılarak incelenip ilgili kurumlarda ve yönetim destek sistemlerinde kullanılmak üzere değerlendirilmesidir. Veri madenciliğini uygulayan kişinin geleneksel yöntemlerde olduğunun aksine başlangıçta belirli bir amacı ya da varmak istediği bir sonuç yoktur. Yapılacak tüm analizlerden sonra elde edilen bilgilerin bir istatistikçi gözü ile incelenip daha

önceden düşünülmemiş kavramların ortaya çıkarılması, başarılı bir veri madenciliği süreci olarak kabul edilmektedir (Albayrak ve Yılmaz, 2009).

Veri madenciliği, istatistiksel yöntemler serisi olarak tanımlanabilse de, veri madenciliği, geleneksel istatistikten birkaç yönde farklılık gösterir. Veri madenciliğinde hedef, kolaylıkla mantıksal kurallara ya da görsel sonuç ve sunumlara dönüştürülebilecek nitel modellerin çıkarılmasıdır. Bu özelliği sebebiyle, veri madenciliği insan merkezlidir ve çoğunlukla insan – bilgisayar ara yüzü birleştirilir.

Veri madenciliği alanı, istatistik, makine bilgisi, veri tabanları ve yüksek performanslı işlem gibi temelleri de içermektedir. Benzer şekilde veri madenciliği; disiplinler arası doğasından dolayı veri tabanları, makine öğrenmesi, bilgi toplama, görselleştirme, optimizasyon gibi pek çok disiplinle de etkileşim halindedir. Veri madenciliğinin temellerini oluşturan disiplinler Şekil 3.1’de özetlenmiştir (Savaş vd, 2012).



Şekil 3.1. Veri madenciliği ile etkileşimli disiplinler

3.2. Veri Madenciliğinin Uygulama Alanları

Veri madenciliğinin günümüzde yaygın bir kullanım alanı bulunmaktadır. Özellikle, rekabetin oldukça güçlendiği piyasalarda, firmaların konumlarını sağlamlaştırmak adına birtakım değerlerinin yönetilmesinde veri madenciliği büyük rol oynamaktadır. Veri madenciliğinin en çok kullanıldığı alanların başında müşteri bilgileri gelmektedir. Bu kapsamda; başta pazarlama, bankacılık, sigortacılık gibi alanlar olmak üzere, iş dünyasında

hemen her alanda veri madenciliği yöntemleri ile çalışmalar yapılabilir. Bu çalışmalar kullanım yerlerine göre aşağıdaki gibi sınıflandırılabilir:

Pazarlama

- Pazar sepet analizi
- Mevcut müşterilerin sürekliliğinin sağlanması, yeni müşterilerin kazanılması
- Müşterilerin satın alma alışkanlıklarının belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların ortaya çıkarılması
- Müşteri ilişkileri yönetimi
- Satış tahmini

Bankacılık

- Müşterilerin kredi taleplerinin değerlendirilmesi
- Kredi kartı harcamalarına göre müşteri belirlenmesi
- Kredi kartı dolandırıcılıklarının tespiti
- Farklı finansal göstergeler arasındaki gizli korelasyonların tespit edilmesi

Tıbbi Araştırmalarda

- DNA içerisindeki genlerin sıralarının belirlenmesi
- Protein analizlerinin yapılması
- Hastalık haritalarının hazırlanması
- Hastalık teşhisleri
- Sağlık politikalarının belirlenmesi

Elektronik Ticaret

- E-CRM uygulamalarının yönetimi
- WEB sayfalarına sık yapılan ziyaretlerinin çözümlenmesi
- Kullanıcı davranışlarına göre web sitesinin güncellenmesi

Sigortacılık

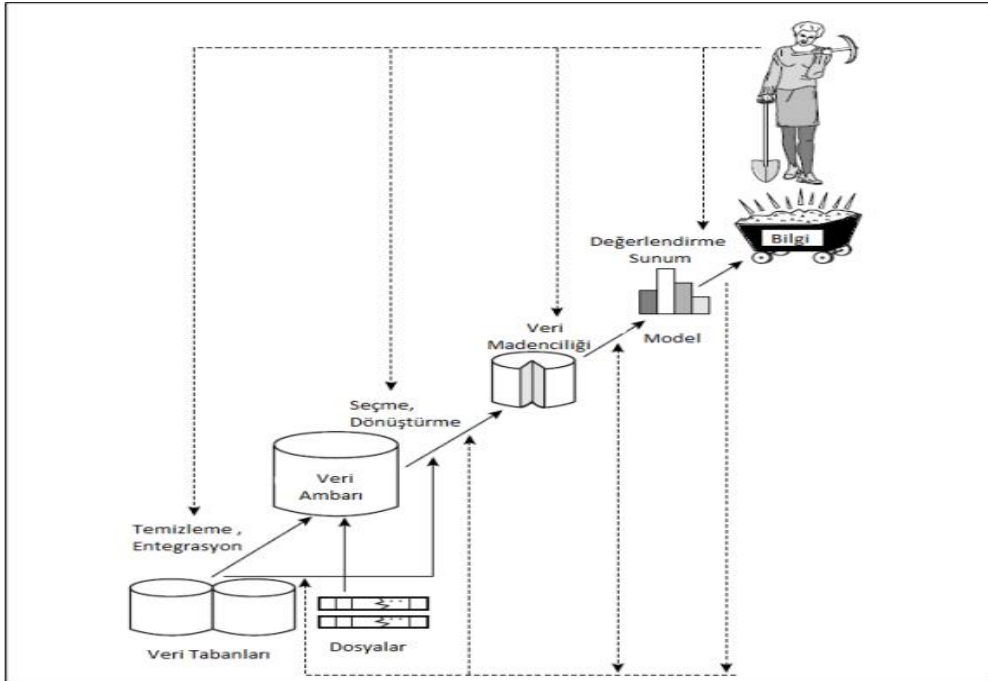
- Yeni poliçe talep edecek müşterilerin tahmin edilmesi
- Sigorta dolandırıcılıklarının tespit edilmesi
- Riskli müşteri gruplarının belirlenmesi

Telekomünikasyon

- İletişim ağlarında sorunlu bölgelerin tespiti
- Kaçak hat kullanımlarının belirlenmesi
- Kullanıcı davranışlarının belirlenmesi
- Müşteri davranışlarına göre yeni hizmetlerin sunulması

3.3. Veri Madenciliği Süreci

Veri madenciliği, birbiri ardına gelen adımlardan oluşan bir süreçtir. Verinin, amaca yönelik algoritmaları uygulamak üzere hazırlanması ve keşfedilen örüntülerin yorumlanmasına kadar bütün işlemler bu sürecin bir parçası olarak tanımlanabilir. Veri madenciliği sürecini özetleyen bir görsel, Şekil 3.2’de sunulmuştur (Han ve Kamber, 2006):



Şekil 3.2. Veri madenciliğinde izlenen süreç

Veri madenciliği çalışmalarında uygulanan adımlar sırasıyla aşağıdaki şekilde özetlenebilir:

1. Problemin (çalışmanın amacının) belirlenmesi,
2. Verilerin model için hazırlanması,
3. Uygun modelin kurulması ve değerlendirilmesi,
4. Modelin uygulanması,
5. Modelin takip edilmesi

Problemin belirlenmesi: Veri madenciliği üzerine yapılan çalışmaların başarıya ulaşmasının şartlarından biri, çalışmanın hangi amaçla, nerde kullanılacağını belirlenmesi ve çalışma sonucunda elde edilen bilgilerin ve raporların amaca uygunluğunun ölçülmesidir.

Verilerin hazırlanması: Model kurma aşamasından önce, veriler birtakım işlemlerden geçirilerek, modelin en doğru şekilde çalışması için hazırlanır. Bu işlemlerden bazıları, örneklem belirleme, dönüştürme, birleştirme, temizleme ve değer biçmedir. Modelin çalışmasına ilişkin problemler oluştuğunda, veri madencisi bu aşamaya dönerek yeniden değerlendirme yapmak durumunda kalmaktadır. Bu sebeple; verilerin hazırlanması adımı, veri madenciliği sürecinin en çok vakit alan adımı olarak düşünülebilir

Modelin kurulması ve değerlendirilmesi: İlk adımda belirlenmiş olan çalışmanın amacına en uygun modelin bulunması, daha önceki çalışmaların incelenmesi ve pek çok modelin kurularak denenmesi ile mümkün olmaktadır. Bu sebeple, verinin hazırlanması ve model kurulması adımları en iyiyi bulana kadar sürekli yenilenen aşamalardır.

Modelin uygulanması: Model kurulduktan sonra ortaya çıkan sonuçlara göre, geçerliliği kabul edilir ve/veya amaca uygun olarak başka modellerin kurulmasında alt model olarak kullanılabilir.

Modelin takip edilmesi: Sistemlerin yapılarında ve dolayısıyla üretilen verilerde zamanla ortaya çıkan değişiklikler, kurulan modellerin sürekli kontrol edilmesini ve gerekli görüldüğünde revize edilerek güncellenmesini zorunlu kılmaktadır.

3.4. Veri Madenciliği Sürecini Etkileyen Faktörler

Veri madenciliği temel olarak etkileyen faktörler aşağıdaki gibi sıralanabilir (Akpınar, 2000):

Veri: Veri madenciliğine konu olan ve gereklilik oluşturan en önemli faktördür. Veri madenciliğine konu olan veri yığınlarının boyutu ve miktarı arttıkça, kullanılan veri madenciliği yönteminin ve algoritmasının özellikleri değişmektedir. Uygulama yapılacak veri kümesinin niteliğine göre, veri ön işleme sürecinde izlenecek adımlar belirlenir.

Bilgisayar teknolojisi: Günümüzde, gelişmiş teknoloji ve bilgisayar ağı sayesinde dağınık ve büyük veri yığınları dahil her tür veri üzerinde algoritmalar uygulamak mümkün hale gelmiştir. Veri madenciliği uygulamaları bilgisayar teknolojisiyle doğru orantılı olarak gelişmektedir.

Donanım: Elektronik aletleri oluşturan parçalar olarak nitelendirilen donanım, veri madenciliği sürecini yüksek ölçüde etkileyen bir kavramdır. Günümüzde, yükselen işlemci hızı ve artan bilgisayar belleği sayesinde farklı amaçlara uygun veri madenciliği çalışmaları kolaylık kazanmıştır.

Bilimsel hesaplamalar: Karmaşık problemleri çözmek amacıyla geliştirilen algoritmalar modelleme ve simülasyon yazılımlarının geliştirildiği günümüzde, simülasyon da teorik bilgi ve deney kadar bilim insanları tarafından tercih edilen bir yöntemdir. Veri madenciliği, bu bilimsel hesaplamaları içeren yöntemleri birleştiren bir alan olması yönüyle dikkat çekmektedir.

Ticari eğilimler: Piyasada faaliyet gösteren işletmelerin, firmaların ve şirketlerin, artan rekabet hızına uyum sağlamak amacıyla en fazla kazancı, en az insan gücü ve maliyetle oluşturmak üzere veri madenciliği vb. faydalı süreçleri izlemesi zorunlu hale gelmiştir.

3.5. Veri Madenciliği Sürecinde Karşılaşılan Sorunlar

Veri madenciliği; veri tabanlarında yer alan ham veriyi girdi olarak kullanır. Bu sebeple de veri tabanlarında yer alan verinin güncel olmaması, eksikler içermesi ya da net olmaması gibi durumlar veri madenciliği sürecinde sorunlar oluşturur. Veri madenciliği sürecinde karşılaşılan bazı sorunlar aşağıdaki gibi özetlenebilir (Savaş vd, 2012).

- 1) Veri tabanı boyutu: Günümüzde, veri tabanı boyutları çok yüksek seviyelere ulaşmıştır. Temel algoritmalar ise küçük örneklemelere uygulanabilecek şekilde geliştirilmiştir. Bu algoritmaların çok daha fazla büyüklükte örneklemelerde kullanılabilmesi için daha fazla dikkat gerekmektedir. Örneklem büyüklüğünün artması, modelin başarısı açısından avantaj oluştursa da dikkatsizlik nedeniyle oluşabilecek hatalar büyük örneklemelerin dezavantajıdır.
- 2) Sınırlı Bilgi: Çoğu veri tabanında, yalnızca veriyi depolama özelliği bulunmaktadır. Bu sebeple, veri madenciliği sürecini kolaylaştıran, öğrenme görevine ilişkin herhangi bir özellik veri tabanlarında yer almamaktadır.
- 3) Eksik veri: Büyük veri kümelerinde, muhakkak eksik veriler bulunmaktadır. Bu eksiklikler uygulanacak olan istatistiksel analizlerde önemli sorunlar oluşturur. Bunun nedeni, istatistiksel analizlerin yapılması için kullanılan paket programların birçoğunun veri eksikliği olmayan durumlar için geliştirilmiş olmasıdır. Eksik veri bulunduran veri tabanlarında bu programların analiz sonuçlarının geçerliliği düşmektedir.
- 4) Aykırı veri: Veri tabanlarında yer alan veriler üzerinde oluşan sistem dışı hatalara gürültü denir. Veriler üzerinde oluşan gürültü, veri madenciliği sürecinde güvenilir sonuçlara ulaşmayı engellemektedir. Aynı şekilde bu gürültüler, geleceğe ilişkin yapılan tahminlerin de doğruluğunu etkiler. Veri madenciliği sürecinde, gürültülü verilerden kurtulmanın yolu, tutarsız ve küçük veriler yerine, anlamlı ve net veriler kullanılmasıdır.

3.6. Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller, ana hatlarıyla tanımlayıcı ve tahmin edici modeller olarak iki sınıfa ayrılır. Tanımlayıcı modeller; veride bulunan örüntü ve ilişkileri tespit etmek için kullanılır. Tahmin edici modeller ise geleceğe yönelik tahminlerde bulunmaya yardımcı olur. Modeller bazı uygulamalarda amaç olarak birbirlerinin yerine de tercih edilebilirler.

Tanımlayıcı modeller, veride mevcut bulunan örüntülerin ortaya çıkarılmasını amaçlar. Kümeleme ve Birlikte Kuralları veri madenciliğinde en sık kullanılan tanımlayıcı modellerdir.

Tahmin edici modeller ise, sonuçları bilinen bir miktar veriden faydalanılarak bir model geliştirilmesi ve sonuçları bilinmeyen bir veri kümesinin sonuçlarının tahmin edilmesini hedefler. Sınıflandırma ve regresyon analizi, tahmin edici modellerin temel iki türüdür.

3.6.1. Kümeleme (Clustering)

Kümeleme yönteminde temel amaç, çok boyutlu veriler içindeki doğal grupları (kümeleri) bulmaktır. Birbirine benzeyen özelliklerdeki elemanlar aynı kümede; benzemeyen özelliklere sahip elemanlar ise farklı kümelerde yer alır.

Kümeleme yönteminin tahmin edici modellerden olan sınıflandırmadan temel farkı, başlangıçta verilerin hangi özelliğe göre, hangi sayıda kümeleneceğinin bilinmemesidir. Sonuçların anlaşılabilmesi için küme hakkında bilgili olan birisinin kümeleri yorumlaması gerekmektedir. Küme oluştururken değişken seçimine dikkat edilmelidir. Bazı değişkenler üzerinde dönüşüm yapılarak yeni bir değişken oluşturulduğunda, dönüşümde kullanılan değişkenler kümenin dışında tutulmalıdır, çünkü kullanıcı bu değişkenleri zaten gruplamada kullanmıştır (Özkan, 2008).

Veri madenciliği çalışmalarında kullanılan pek çok kümeleme algoritması bulunmaktadır. Algoritmalar, yapılacak uygulamanın amacına, kullanılan veri türüne ya da kümelerin oluşma biçimine göre birbirinden ayrılırlar. Kümeleme algoritmalarını hiyerarşik ve hiyerarşik olmayan algoritmalar olmak üzere temel iki gruba ayırmak mümkündür:

Hiyerarşik Kümeleme Yöntemleri;

- Toplaşım Kümeleme Algoritmaları,
- Bölünür Kümeleme Algoritmaları

Hiyerarşik Olmayan Kümeleme Yöntemleri;

- Yer Değiştirmeli Algoritmalar,
- Olasılıksal Algoritmalar,
- Yoğunluğa Dayalı Bağlantılı Kümeleme,
- Yoğunluk Fonksiyonlu Kümeleme
- K-Medoid Yöntemler,
- K-Means Yöntemler, Yoğunluğa Dayalı Algoritmalar

3.6.2. Birliktelik kuralları (Association rules) ve sepet analizi

Veri içerisinde beraber görülen olayları, ürünleri, hareketleri ve davranışları keşfetmeye yarayan yöntemlere birliktelik kuralları adı verilir. Birliktelik kuralları; veri kümesi içerisinde varsa birbiriyle ilişkili olan değişkenlerin tespit edilmesi ve bu değişkenlerin aralarındaki örüntünün büyüklüğünün belirlenmesini hedefler. Değişkenler arasındaki ilişki, birliktelik kuralları analizi ile tanımlanır. Bu özelliği nedeniyle birliktelik analizi, bir veri kümesindeki kayıtlar arasındaki bağlantıları keşfeden denetimsiz veri madenciliği tekniğidir.

Birliktelik analizi, genellikle satış ve pazarlama alanında, süpermarket müşterilerinin satın alma davranışlarını ortaya koymak amacıyla kullanılır. Bu sebeple “pazar sepeti analizi” olarak da adlandırılır. Pazar sepet analizinde amaç ürünlerin birlikte alınmasını içeren kuralları tespit etmektir. Bu sayede sektörde yer alan şirketler, satış ve kârlarını artırmak üzere çalışmalar yapmaktadır. Örneğin, bir süpermarkette X ürününü alanlar yüksek ihtimalle Y ürününü de tercih ediyorsa; bir müşteri X ürününü alıp Y ürününü almadığında, o müşterinin Y ürününü alma potansiyeli oldukça yüksektir (Gülce, 2010). Firmalar bu analizlerin sonuçlarını, reklam ve ürün tanıtımı stratejilerinin belirlenmesinde, raf düzenlemelerinde, promosyon süreçlerinde kullanırlar.

Birliktelik kurallarının tespit edilmesinde çok sayıda teknik kullanılmaktadır. Algoritmalara göre, birliktelik kurallarının belirlenmesinde eşik bir değer bulunması gerekmektedir.

Algoritmanın en önemli adımı, faydalı bilgiyi gürültüden ayırt etmek üzere bu eşik değerin bulunmasıdır. Dikkat çeken birliktelik kurallarını, önemsiz olanlardan ayırt etmek için bazı ölçütler belirlenmiştir. Bu ölçütlerden destek (support) kriteri, veride ögeler arasındaki bağıntının ne sıklıkta bulunduğunu, güven (confidence) değeri ise belirlenen bir ögenin hangi olasılıkla diğer öge ile birlikte bulunacağını ifade etmektedir.

Birliktelik kuralı analizinde kullanılan çok sayıda algoritma bulunmaktadır. Bu algoritmalar arasında en yaygın kullanılan algoritma Apriori (önsel) algoritmasıdır.

Apriori algoritması:

1994 yılında, Agrawal ve Srikant tarafından geliştirilen algoritma, veri madenciliği sürecinde birliktelik kuralı analizi yapan algoritmalarından en fazla bilinen ve uygulanan algoritmadır. Apriori algoritmasının adı, bilgileri her seferinde bir önceki adımdan alması ve önsel bilgilerden faydalanması sebebiyle “önceki (prior)” kelimesinden türemiştir.

Apriori algoritmasının adımları aşağıdaki şekilde özetlenir;

1. Minimum destek (min.support) ve minimum güven değerleri (min.confidence) belirlenmesi
2. Öge kümeleri içerisinde bulunan her bir ögenin destek değerinin hesaplanması
3. Bulunan destek değerlerinden, minimum destek değerinden daha düşük destek değerine sahip olan ögelerin devre dışı bırakılması
4. Yukarıdaki işlemler sonucunda bulunan tekli birliktelikler dikkate alınarak ikili birlikteliklerin oluşturulması
5. Bulunan destek değerlerinden, minimum destek değerinden daha düşük destek değerine sahip olan öge kümelerinin devre dışı bırakılması
6. Yukarıdaki işlemler sonucunda bulunan ikili birliktelikler dikkate alınarak üçlü birlikteliklerin oluşturulması
7. Bulunan destek değerlerinden, minimum destek değerinden daha düşük destek değerine sahip olan üçlü öge kümelerinin devre dışı bırakılması
8. İzlenen adımlar sonrasında, en yüksek değere sahip ikili ve/veya üçlü birliktelik kurallarının çıkarılması

3.7. Sınıflandırma (Classification)

Sınıflandırma, dağınık bir şekilde olan verilerin belirli özelliklerine göre bir araya getirilmesi işlemidir. Sınıflandırma algoritması, bir araya getirme işleminin hangi özelliklere göre olacağını belirleyen algoritmadır. Sınıflandırma algoritması sonucunda oluşan bir sınıfta yer alan bir öğenin, sınıfta yer alan diğer öğelerle belirli ortak özelliklerinin bulunması gerekmektedir.

Sınıflandırma kullanım alanının fazla olması nedeniyle, en çok kullanılan veri madenciliği yöntemlerinden biridir. Resim, örüntü tanıma, hastalık tanıları, dolandırıcılık tespiti, kalite kontrol çalışmaları ve pazarlama konuları, sınıflandırma tekniklerinin en sık kullanıldığı alanlardır. Sınıflandırma yöntemi tahmin edici bir modeldir. Sınıflandırma ile öngörü yapılabilir. Sınıflandırma işleminde, eğitim veri seti ile en uygun model oluşturulur ve bir miktar test verisi modele uygulanarak doğruluğu hesaplanır. Havanın bir sonraki gün nasıl olacağını tahmin edilmesi ya da bir kutuda ne kadar mavi top olduğunun tahmin edilmesi aslında bir sınıflandırma işlemidir (Dunham, 2003: 8).

Sınıflandırma yöntemiyle ilgili bir model şu şekilde örneklendirilebilir: Satışlarını artırmak için kampanya düzenlemek isteyen bir firma, kampanyasına katılma ihtimali olan potansiyel müşterileri belirlemek için daha önceden satış yapmış olduğu müşterilerinin verilerini (satın alma sonuçlarını) kullanarak, hangi özelliklere sahip adayların kampanyaya katılabileceğini tahmin edebilir. Bu şekilde; kampanya kapsamında yalnızca satın alma potansiyeli yüksek müşterilere ulaşarak zaman ve kaynaklardan tasarruf edecektir. Benzer şekilde, bankacılık sektöründe yapılan fon transferleri sıklığına göre müşteri profilini ortaya çıkarmak ve bu profillerin ortak özelliklerini belirlemek de veri madenciliğinin sınıflandırma yöntemi altında yapılmaktadır.

Sınıflama kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Regresyon ile amaç, girdiler ile çıktıyı ilişkilendirecek modeli oluşturup, en iyi tahmine ulaşmaktır.

Sınıflandırma ve regresyon modellerinde kullanılan başlıca yöntemler şunlardır:

- Karar Ağaçları (Decision Tree)

- Yapay Sinir Ağları (Artificial Neural Networks)
- Destek Vektör Makineleri (Support Vector Machines)
- Genetik Algoritmalar (Genetic Algorithms)
- Regresyon Analizi (Regression)
- Bayesyen Sınıflandırma (Bayesian Classification)

Söz konusu yöntemler arasından Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri yöntemleri; çalışmanın amacı olan proje başvurusu yapan araştırmacıların profillerini belirlemek ve proje başvuru yapma eğilimine etkisi olan faktörleri tespit etmek amacıyla en uygun yöntem olması, kategorik veriye uygulanabilmesi ve sınıflandırma sonuçlarını görsel olarak özetleyerek tahmin olanağı sağlaması sebepleriyle uygulamada kullanılacak yöntemler olarak öne çıkmaktadır.

3.7.1. Karar ağaçları

En yaygın kullanılan sınıflandırma algoritmalarından biri karar ağaçlarıdır. Diğer veri madenciliği yöntemlerine göre kurulması oldukça kolay olan ve sonuçları görsel bir şekilde özetleyen karar ağaçları bu yapılarıyla sınıflandırma yöntemleri arasında öne çıkmaktadır.

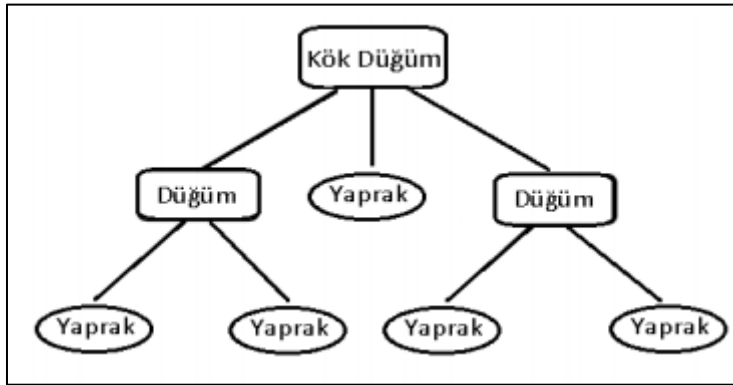
Karar ağaçları yöntemi, hem tanımlayıcı hem de tahmin edici özelliğe sahiptir. Ayrıca karar ağaçları ile oluşturulan modellerin başarısı oldukça yüksektir. Karar ağaçları tekniğinde, sınıflandırma yapmak üzere mevcut verilerden ağaç yapısı oluşturulur ve veri kümesinde yer alan kayıtlara bu ağaç yapısı uygulanır, elde edilen sonuçlara göre diğer kayıtların sınıflandırma işlemi gerçekleşir. Oluşturulan karar ağaçları sayesinde, yeni bir kayıt oluştuğunda bu kaydın hangi sınıfta yer alacağı öngörülür.

Sınıfları daha önceden belirlenen örnek veri kümesinden, tümevarım tekniği ile öğrenme gerçekleştiren ağaç biçimli bir karar yapısı çeşidi olan karar ağaçları, basit karar verme adımlarını uygulayarak, çok büyük miktarlardaki veri kayıtlarını, daha küçük veri kümelerine böler. Algoritmanın her adımında, en son oluşan veri gruplarında yer alan öğeler birbirleriyle çok daha benzer hale gelirler (Sun ve Li, 2008).

Karar ağaçları hem geleceğe dönük tahmin yapılmasına olanak sağlayan, hem de pek çok sisteme uygulanabilir olan bir sınıflandırma yöntemi olması nedeniyle en yaygın kullanılan sınıflandırma tekniğidir. Ayrıca karar ağaçları;

- Düşük maliyetli olması,
- Anlaşılmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolay olması,
- Güvenilirliklerinin iyi olması gibi nedenlerden ötürü de en yaygın kullanılan sınıflandırma tekniklerinden biridir (Çalış vd, 2014).

Karar ağaçları ile sınıflandırma süreci kök düğüm ile başlar. Niteliklerin değerine göre düğümler ilk olarak alt dallara ayrılır ve ayrılma süreci yaprak oluşana kadar sürer. Burada düğümler; test işlemine tabi tutulan nitelikleri gösterir ancak bilinmelidir ki tüm değişkenlerin ya da niteliklerin karar ağacında yer alması beklenmez. Dallar; teste tabi tutulan düğümlerin sonuçlarını gösterir. Elde edilen dallar ile tahmin edilecek sınıfa giden yol belirlenir. Dalın sonucunda sınıflama tamamlanmıyorsa tekrar düğüm oluşturulur. Yapraklar ise veri setindeki bir karar sınıfını temsil etmektedir. Karar ağacının yapısı Şekil 3.3’de gösterilmiştir (Diler, 2016):



Şekil 3.3. Karar ağacı yapısı

Karar ağacı yapısında yer alan ilk eleman, “kök düğüm” olarak da adlandırılan karar düğümüdür. Karar ağacı oluşturulurken kullanılan algoritmaya göre, her düğüm iki veya daha fazla dala sahip olabilir. İki dala sahip olan karar ağaçları ikili ağaç, ikiden daha fazla sayıda dala sahip olan karar ağaçları ise çoklu ağaç olarak nitelendirilir. Her dal başka bir karar düğümüyle, ya da ağacın sonuyla yani yaprak düğümle sonlanır. Karar ağacı yapılarında, elde edilen sınıfların maksimum saflıkta olması, yani oluşan sınıflar arasında

mesafenin en fazla olması amaçlanır. Kategorik verileri sınıflandırmak amacıyla oluşturulan karar ağaçları sınıflandırma ağacı, sürekli sayısal verileri tahmin etmek amacıyla kullanılan karar ağaçları ise regresyon ağacı olarak bilinir (Özçınar, 2006).

Karar ağacı temelli sınıflandırma uygulamalarının en yaygın kullanıldığı alanlar aşağıdaki gibidir (Diepen, 2006):

- Belirli bir sınıfın üyesi olmaya aday elemanların belirlenmesi,
- Parametrik modellerin kurulmasında kullanılması amacıyla fazla sayıdaki değişkenden en önemlilerinin seçilmesi,
- Çeşitli vakalarda düşük, orta, ve yüksek risk grupları gibi çeşitli kategorilerin oluşturulması,
- Gelecekteki olaylara ilişkin tahminler için kurallar oluşturulması,
- Yalnızca belirli alt gruplara özgü örüntülerin tanımlanması,
- Sürekli değişkenlerin kesikli değişkenlere dönüştürülmesi ve kategorilerin birleştirilmesi

Karar ağacı oluşturmak üzere, çalışma yapılacak çok sayıda algoritma bulunmaktadır. Kullanılan algoritmaya göre, oluşan karar ağaçlarının yapısı ve özellikleri değişiklik göstermektedir. Üzerinde çalışılan verinin niteliğine ve büyüklüğüne bağlı olarak, hedef ve bağımlı değişkenlerin özelliklerine göre kullanılacak algoritma seçilebilir. Ayrıca elde edilmek istenen sonuçlara göre de kullanılacak algoritma belirlenir. Karar ağaçları için geliştirilen algoritmalar aşağıda özetlenmiştir;

AID (1970) (Automatic Interaction Detector): Karar ağaçları için ilk algoritma ve yazılım Morgan ve Sonquist tarafından 1970'lerin başlarında geliştirilmiştir (Emel ve Taşkın, 2005).

C&RT (1984) (Classification and Regression Trees): Algoritma, Gini'ye dayalı ikili bölme işlemi üzerine kurulmuştur. Karar ağacı yapısındaki her bir düğümde iki adet dal bulunur. Budama işlemi ağacın karmaşıklık ölçüsüne dayanmaktadır. Algoritma sürekli hedef değişkeni ile çalışmaktadır.

ID3 (1986) (Iterative Dichotomiser 3): Geliştirilen ilk algoritmalarından olan bu algoritma, bilişim teorisi ve makine öğrenmesi temeline dayanır. Sonuç değerinde etkisi bulunan, ayırt

edici değişkenleri “entropi” kavramına göre belirler. Bu kavram, bir veri kümesinde mevcut olan düzensizliğin miktarını ölçmek amacıyla hesaplanır.

C4.5 ve C5.0 (1993): Sıklıkla kullanılan ID3 karar ağacı algoritmasının geliştirilmiş versiyonu olan C4.5 ve C5.0 algoritmaları daha çok büyük veri setleri için uygundur. Ayrıca bu algoritmalar ID3 algoritmasından farklı olarak sürekli değişken içeren veri setleri için de uygundur. C5.0 algoritmasının, daha yüksek doğruluk değeri için boosting algoritmasını kullanması nedeniyle, bu tür karar ağaçları “boosting ağaçları” olarak da adlandırılır. Her iki algoritmanın sınıflandırma yöntemi aynı olsa da C5.0 algoritması C4.5’in geliştirilmiş versiyonu olduğundan daha hızlı ve doğru sonuçlar vermektedir. (Çalış vd., 2014)

SLIQ (1996) (Supervised learning in Quest): Bu algoritma, sürekli ve kategorik verilerin sınıflanmasında kullanılabilir. Sürekli değişkenler üzerinde işlem yaparken maliyeti azaltmak için ağaç oluşturulurken önceden-sıralama tekniği kullanır. Önce genişlik ilkesiyle çalışan algoritma, birçok yaprağı oluşturur. Karar ağacını, dallara ayırmada gini endeksi kullanılır. Bu algoritmanın avantajları hızlı sonuç vermesi ve sonuçların doğruluğunun yüksek olmasıdır.

SPRINT (1996) (Scalable PaRallelizable INduction of Decision Trees): Bu algoritma, büyük veri kümeleri için uygundur. Tüm değişkenler için ayrı bir değişken listesi hazırlar. Her tabloda kullanılacak olan sınıf ve sıra numarası bulunur. Değişken sayısı kadar tablo oluşturulur ve sınıflandırma ağacı uygulanır. SPRINT algoritması da dallandırma kriteri olarak Gini indeksini kullanır (Silahtaroglu, 2016).

QUEST (1997) (Quick, Unbiased, Efficient Statistical Tree): Bu algoritma yalnızca ikili ağaçlar üretmektedir. İkili ağaç oluşturulmasının nedeni, budama ve doğrudan durma kuralı gibi yöntemlerin ikili ağaçlarda kullanılmasıdır. QUEST algoritması ile ağaç oluşturulurken, değişken seçimi ve bölünme eş zamanlı olarak yapılır. QUEST algoritmasının avantajları, ağacın dallanması esnasındaki önyargılı seçimin daha genel duruma getirilmesi ve hesaplama maliyetinin azaltılmasıdır (Çalış vd., 2014).

CHAID (1980) (Chi-Squared Automatic Interaction Detector): Bu algoritma, karar ağacında bölme ve budama işlemini ki-kare testi ile gerçekleştirir. Dalların sayısı en az iki olup, sınıf

kategorisine göre deęişiklik gösterir. Sürekli ve kategorik tüm deęişkenlere uygulanabilmesi algoritmanın avantajları arasındadır.

CHAID algoritması, karar ağacı yapısı oluşturmak için yaygın olarak kullanılan, istatistik temelli bir uygulamadır. Algoritma, karar ağacında bölme ve budama işlemini ki-kare testine karşılık gelen p deęerine göre belirler. CHAID analizi hedef sonuca etkili olan faktörleri belirleyerek, öngörücü bir model oluşturur.

Karar ağacı yapısında ikiden daha fazla dal oluşturabilmesi, kategorik ve sürekli tüm deęişkenlerle çalışabilir olması, deęişkenlere ağırlık deęerleri atanabilmesi, eksik deęerleri ayrı bir kategoride analiz etmesi algoritmanın en önemli avantajları olarak sıralanabilir. Ayrıca CHAID algoritmasında deęişkenlerin parametrik olma ve verilerin normal dağılması gibi zorunluluklar da bulunmamaktadır.

İstatistiksel olarak, deęişken deęerleri kategorik olduğunda ve bağımsız deęişkenler ile kategorik olarak ölçülebilen sonuçlar arasında ilişki gerektiren durumlarda oldukça yararlı sonuçlar üreten CHAID yöntemi, ki-kare parametrik olmayan istatistik yöntemini kullanmaktadır. CHAID, kesimleri ki-kare testi tarafından gösterilen bir bağımlılık ilişkisinin bir sonucu olarak yapılandırılan deęişkenlerin geri kalanıyla ilişkili bir kriter deęişkeni oluşturan tahmini bir analiz oluşturmaktadır. (Diaz vd., 2016)

İlk uygulamaları hastalık tanısı ve tıp araştırmaları üzerine olan CHAID algoritması daha sonra, satış ve pazarlama alanında tüketici sınıflarının belirlenmesi uygulamalarında kullanılmıştır.

CHAID algoritmasının adımları aşağıda yer almaktadır (Diler, 2016);

1. Veri setindeki her bir bağımsız deęişkenin kategorileri ile bağımlı deęişkenin kategorileri arasında çapraz tablo oluşturulur. Bunların arasında en az öneme sahip kategori çiftleri bulunur (en büyük p deęerine sahip olan). Burada bağımlı deęişkenin türüne göre p deęeri hesaplanacaktır;

Bağımlı deęişken sürekli ise F testi ile, bağımlı deęişken kategorik ise Ki-kare testi ile hesaplanır.

2. En büyük p değerine sahip bağımsız kategori çifti için, daha önce belirlenen alfa düzeyi ile p değeri kıyaslanır;
p değerinin alfadan büyük olması durumunda, bu iki kategori birleştirilir ve bir önceki adıma geri dönülür, p değeri alfadan küçük ise sıradaki adıma geçilir.
3. Bağımlı ve bağımsız kategori kümeleri için uygun bonferroni düzeltmesi kullanılarak, düzeltilmiş p değeri hesaplanır.
4. En önemli bağımsız değişken, en küçük düzeltilmiş p değerine sahip olan arasından seçilir ve bu değer önceden tanımlanmış alfa düzeyi ile kıyaslanır;
p değerinin alfa değerinden küçük ya da eşit olması durumunda ise düğüm belirlenmiş olan bağımsız kategori kümesi temel alınarak bölünür, p değeri alfa değerinden büyük ise bölünme olmaz, bu düğüm uç düğüm olur.
5. Bu işlem durma kuralları gerçekleşinceye kadar devam eder.

Exhaustive CHAID (1991): Bu algoritma CHAID algoritmasının modifiye edilmiş halidir. CHAID algoritmasına benzetmekle birlikte bu versiyonda hesaplama işlemleri fazla zaman almaktadır (Köktürk, 2012).

Karar ağaçları ile sınıflandırma algoritmaları arasından; tez çalışması kapsamında kullanılan veri setine uygunluğu ve sık kullanılan algoritmalarından biri olması nedeniyle CHAID algoritması seçilmiştir. Uygulamada kullanılan yöntemlere ilişkin bilgi, ayrıca dördüncü bölümde de sunulmuştur.

3.7.2. Yapay sinir ağları

Günümüzde, teknolojik gelişmeler konusunda oldukça ilerleme kaydedilmiş ve yapay zeka (artificial intelligence) kavramı ile insana ait düşünebilme ve öğrenme özelliklerinin bilgisayarlar tarafından taklit edilebilmesi çalışmaları önem kazanmıştır. Bilgisayarlara düşünme yetisinin kazandırılması olarak nitelendirilen yapay zeka çalışmaları, 1950'li yıllardan beri pek çok alanda kullanılarak etkisini sürdürmektedir.

Yapay zeka çalışmaları kapsamında önem kazanan alanlardan biri de Yapay Sinir Ağları'dır. Yapay Sinir Ağları (YSA) Teknolojisi, öğrenebilen bilgisayar sistemlerinin temelini

oluşturur. YSA, insan yapısındaki biyolojik sinir ağlarından yola çıkan ve insan beyninin çalışma mantığını temel alarak geliştirilen bir bilgi işleme mekanizmasıdır. YSA yönteminde, daha önce öğrenilmiş bilgilerden faydalanarak belirsiz ve karmaşık verilerden bilgi elde edilir. Katmanlardan oluşan YSA, birbirine bağlı nöronlardan meydana gelir. Bu nöronların matematiksel biçimde modellenmesi ile birlikte öğrenme gerçekleşmektedir. YSA yönteminde ilk olarak parametrik bir model öngörülmez ancak her bir uygulamanın ardından öğrenme gerçekleşmekte ve tecrübelerden yola çıkılarak algoritma oluşturulmaktadır (Argüden ve Erşahin, 2008). YSA, katmanlar içinde yer alan ve matematiksel bir modelle birbirine bağlı bulunan nöronların birleşmesinden oluşan ağ yapısıdır.

YSA'nın insan beyni ile benzerliği iki temel üzerine kurulur:

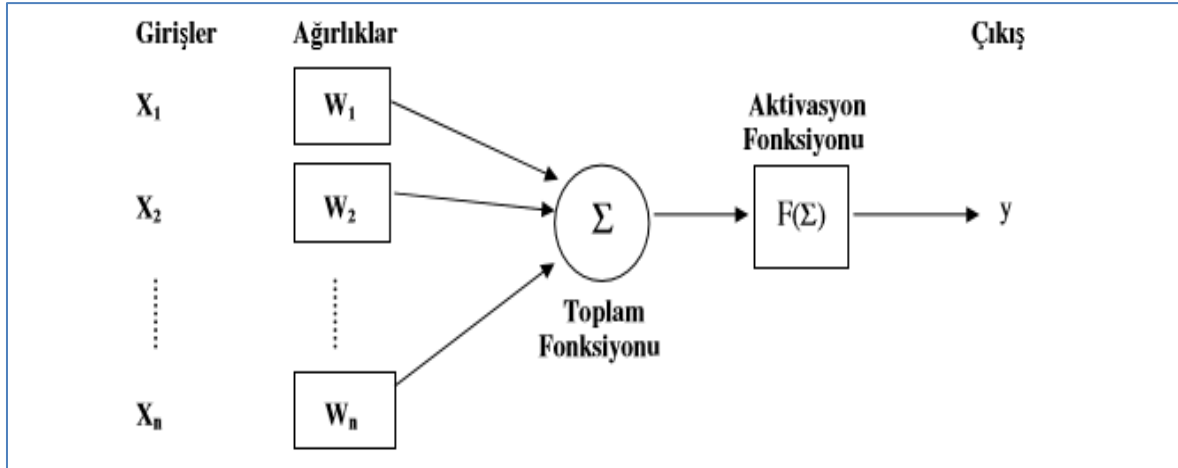
- a) İşlemci tarafından öğrenilen bilgi, belirli bir öğrenme işlemi sayesinde edinilir.
- b) İşlemciyi oluşturan yapay nöronlar arasındaki bağlantı ağırlıkları ile edinilen bilgi kaydedilir.

1949 yılında bir psikolog tarafından yazılan kitapta, yapay sinir hücrelerden oluşan ağın ağırlık değerlerini geliştiren bir öğrenme kuralı açıklanmıştır. “Hebb Öğrenme Kuralı” olarak adlandırılan bu kural, mühendislik alanında uygulanabilir bir değer taşıyorsa da pek çok yapay öğrenme kuralının temelini oluşturmuştur. YSA'nın tarihçesi, insana ait nörobiyoloji bilimine ilgi duyulması ve bu bilimin bilgisayar sistemleri üzerine uygulanabilmesi temeline dayanmaktadır. YSA'lar üzerine bugüne kadar yapılmış olan çalışmalar, birbirinin devamı niteliği taşımakta ve önceki çalışmaların eksik yönlerini tamamlamaktadır. Bu doğrultuda, daha önce yapılan çalışmalar da günümüzde yapılan çalışmalara temel oluşturmaktadır. (Öztemel, 2006).

YSA çoğunlukla, hava durumu tahminleri, ses ve görüntü tanıma, firmaların müşteri potansiyellerinin belirlenmesi, bankaların kredi kararlarının belirlenmesi, sağlık sektöründe hastalık teşhislerinin yapılması gibi pek çok önemli alana ilişkin uygulamalarda kullanılmaktadır.

YSA'da sistem, ağırlıklı düğümlerle birbirlerine bağlanmış sinirlerden oluşur. Sinirleri birbirine bağlayan nöronlar (düğümler), sinapslara bağlanan beyin nöronlarına benzemektedir. Sinir ağı modellerinden en yaygın olanı, algılama düğümlerinden oluşan

giriş düğümleri, bir veya birden fazla gizli katmandan oluşan hesaplama düğümleri ve son olarak çıktı düğümünü içeren çok katmanlı “perceptron” ağıdır. Giriş nöronları örneklemedeki değişken değerlerini ifade ederken, çıkış nöronları ise, diğer tüm örneklerin sınıflarını birbirinden ayıran ayırmıcılardır. Yapay sinir ağı yapısı genel olarak Şekil 3.4’de özetlenmiştir (Köktürk, 2012):



Şekil 3.4. Yapay sinir ağı yapısı

Şekil 3.4’e göre bir yapay sinir ağının temel yapısını oluşturan hücreler aşağıda sıralanmıştır:

Girdi (x): YSA’da ilk olarak dışarıdan girilen bilgilerdir. Şekil 3.4’de belirtildiği üzere; girdi değerleri $x(i)$ ($i=0,1,2,\dots,n$) matematiksel sembolleri ile gösterilir. Bu sonuç daha sonra bir çıktıya dönüştürülür. Burada, girdi vektörü $\{x_1, x_2, \dots, x_n\}$ olarak ve uygun ağırlıklar vektörü ise $\{w_{j1}, w_{j2}, \dots, w_{jn}\}$ olarak gösterilmektedir.

Ağırlık (w): YSA’da hücreler arasında bulunan bağlantıları ifade eden sayısal değerdir. Hücrelere gelen girdinin etkisini ve toplam fonksiyonu üzerindeki önemini belirtir. Şekil 3.4’de ağırlık değerleri ise $w(j)$ ($j=0,1,2,\dots,n$) matematiksel sembolleri ile gösterilmiştir. Sisteme tanımlanan girdi değerleri kendisine ait bağlantı ağırlığıyla çarpılarak en basit yapısıyla toplama fonksiyonuna iletilmektedir.

Toplama fonksiyonu: YSA’da bir hücreye net girdinin hesaplanmasından oluşan fonksiyondur. En yaygın kullanılan toplama fonksiyonu şekli, girdi değerlerinin ağırlıklarıyla çarpılarak toplanmasından oluşan fonksiyondur.

$$NET = w_{ji} x_i + \theta_j$$

Şeklinde ifade edilen bir toplama fonksiyonunda:

NET : YSA nöronunun net girdisini,

w_{ji} : j ve i nöronları arasında bulunan bağlantının ağırlığını,

x_i : i nöronunun çıktısını,

θ_j : eşik değeri belirtmektedir.

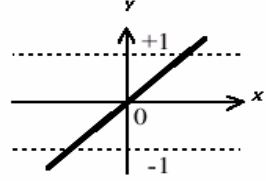
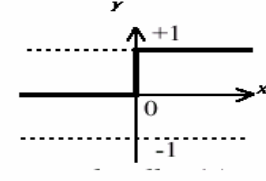
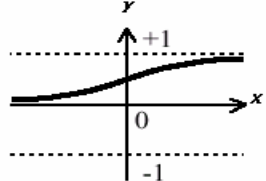
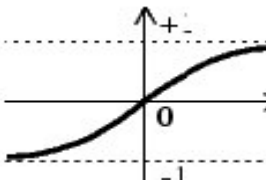
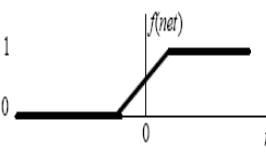
Literatürde yer alan bazı toplama fonksiyonları Çizelge 3.1’de özetlenmiştir (Öztemel, 2006):

Çizelge 3.1. Toplama fonksiyonu örnekleri

Toplam	$NET = \sum_i^n w_{ij} x_i + \theta_j$
Çarpım	$NET = \prod_i w_{ij} x_i$
Minimum	$NET = \text{Min} (w_{ij} x_i)$
Maksimum	$NET = \text{Max} (w_{ij} x_i)$
Çoğunluk	$NET = \sum_i^n \text{Sgn} (w_{ij} x_i)$

Aktivasyon fonksiyonu: YSA’da bir hücreye net girdinin işlenmesiyle hücre tarafından üretilen çıktının belirlemesini sağlayan fonksiyondur. Sigmoid fonksiyonu, YSA’da yaygın olarak kullanılan bir aktivasyon fonksiyonu türüdür. Bir yapay sinir ağı için en uygun aktivasyon fonksiyonu denemeler sonucunda oluşturulmaktadır. Literatürde kullanılan bazı aktivasyon fonksiyonları Çizelge 3.2’de özetlenmiştir (Kamruzzman vd., 2006):

Çizelge 3.2. Aktivasyon fonksiyonu örnekleri

Aktivasyon Fonksiyonu	Matematiksel Gösterimi	Matematiksel Gösterimi
Lineer Fonksiyon	$F(Net) = Net$	
Step Fonksiyonu	$F(Net) = \begin{cases} 1 & \text{eğer } Net \geq 0 \\ 0 & \text{eğer } Net < 0 \end{cases}$	
Sigmoid Fonksiyonu	$F(Net) = \frac{1}{1 + e^{-Net}}$	
Hiperbolik Tanjant Fonksiyonu	$F(Net) = \frac{e^{Net} - e^{-Net}}{e^{Net} + e^{-Net}}$	
Eşik Değer Fonksiyonu	$F(Net) = \begin{cases} 0 & \text{eğer } Net \leq 0 \\ Net & \text{eğer } 0 < Net < 1 \\ 1 & \text{eğer } Net \geq 1 \end{cases}$	

Çıktı (y): Aktivasyon fonksiyonu tarafından üretilen değer çıktı olarak adlandırılır. Çıktı değeri YSA'nın çıktısı olarak dış dünyaya sunulabileceği gibi, başka bir hücrenin girdisi olarak da kullanılabilir. Çıktı değeri, istatistikte yer alan bağımlı değişkeni ifade etmektedir. YSA'da oluşturulan bilginin dışarı aktarımını sağlayan katman çıktı katmanıdır.

Yapay sinir ağı, temel çalışma prensibi olarak birbirine benzemesine rağmen algoritma ve tasarım olarak ayırt edici özelliklere sahiptirler. Yapılarında bulunan nöronların yapılarına,

öğrenme biçimlerine ve bağlantılarında yer alan katman sayılarına göre yapay sinir ağları sınıflandırılır.

Bağlantı yapılarına göre Yapay Sinir Ağları;

- İleri Beslemeli Ağlar
- Geri Beslemeli Ağlar

Öğrenme şekillerine göre Yapay Sinir Ağları;

- Danışmanlı öğrenme
- Danışmansız öğrenme
- Destekleyici Öğrenme

Katman sayısına göre Yapay Sinir Ağları;

- Tek Katmanlı
- Çok Katmanlı Yapay Sinir Ağları olarak sınıflandırılmaktadır.

Farklı YSA ağ yapıları, karakteristik ve performans olarak birbirinden farklı özellikler göstermektedir. YSA'nın yapısı ağın algoritmasını yani modelleme sürecini etkilemesi yönünden önem arz etmektedir. Bir problemin çözümünde ağ tasarımı yapılırken çalışmaya en uygun olan ağ yapısı seçilir. Probleme uygun en iyi ağ mimarisinin seçilmesinde, önerilen bazı yöntemler bulunmasına rağmen, bu yöntemler gerçek tahminler için en iyi çözümü garantilememektedir. Özetle, ağ mimarisinin oluşturulmasında kesin ve net bir şekilde uygulanan yöntemler yoktur. Bu sebeple, yapay sinir ağı tasarımı, bilimsel bir yöntemden daha çok sanat olarak nitelendirilebilir.

Esnek yapıları sayesinde, YSA'lar yeni bilgilerin ortaya çıkması ve mevcut bilgilerde değişiklik olması durumlarında, bağlantı ağırlıkları yeniden belirlenerek ağı revize edebilirler. Bu özellik, YSA'yı geleneksel istatistik yöntemlerinden farklı kılan en önemli özelliklerdendir (Zhang vd., 1998).

YSA, gerçek hayata ilişkin doğrusal modellemesi zor ve karmaşık olan problemleri de kolaylıkla modelleyerek çözüme ulaştırabilmesi özelliğiyle ön plana çıkmaktadır. Ayrıca, gerçek problemlere ilişkin veri setlerinin yapılarında bulunan eksik veri değerleriyle de çalışabilme olanağı sağlaması YSA yönteminin avantajlarından.

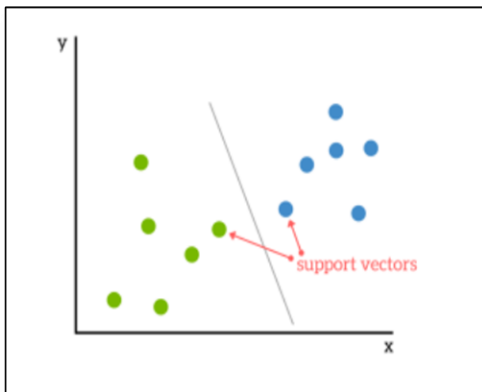
YSA yönteminin avantajlarına rağmen, bazı bilinen dezavantajları da bulunmaktadır. Bunlar; en iyi çözümü garanti etmemesi, öğrenme süresinin uzun olması ve gizli katman ve parametre sayısının net olarak bilinmemesi nedeniyle çalışma biçimini açıklamanın kolay olmamasıdır. YSA’da, diğer veri madenciliği yöntemlerinde olduğu gibi bilgiler bir veri tabanında veya modelin içinde depolanmadan, sinir ağının üzerinde saklanmaktadır. Bu özellik nedeniyle, YSA ile oluşturulan modellerin yorumlanması oldukça zordur.

Bahsi geçen olumsuz özelliklerine rağmen, pek çok problemin çözümünü YSA ile oluşturmak mümkün olmaktadır. Yöntemin dezavantajlarını minimuma indirmek için hem probleme hem de ağ yapısına ilişkin yeterli düzeyde bilgiye sahip olmak gerekmektedir. YSA, zor bir modelleme süreci sonrasında başarılı sonuçlar verebilen bir veri madenciliği yöntemidir (Öztemel, 2006).

3.7.3. Destek vektör makineleri

Destek Vektör Makineleri (DVM), sınıflandırma yöntemleri arasında oldukça yüksek performans gösteren ve oluşturulma fikri basit olan bir öğrenme algoritmasıdır (Hearst vd., 1998).

DVM’nin çalışma prensibi verileri n -boyutlu bir hiperdüzlem ile Şekil 3.5’de belirtilen şekilde en iyi iki sınıfa ayırmaktır. DVM yöntemi temel olarak yapay sinir ağları ile ilişkilidir. Sigmoid bir kernel fonksiyonu kullanan DVM, iki katmanlı, ileri beslemeli bir yapay sinir ağı özelliğiyle çalışır. (Haykin, 1999).

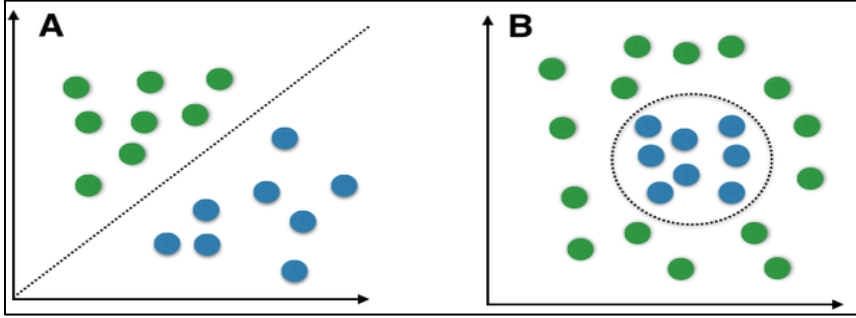


Şekil 3.5. DVM ayırıcı düzlem

DVM yönteminde, verilerin yer aldığı iki sınıfı optimal biçimde ayırabilen bir hiperdüzlem bulunur, bulunan hiperdüzlem sınıflama yaparken en iyi genelleştirmeyi tespit eder. Hiperdüzlemin genelleştirme yeteneği, öncelikle eğitim verisinden en iyi şekilde öğrenilir. Aynı şekilde, test verisi üzerinde de en iyi sınıflandırmayı yapacağını ve doğru tahminlerde bulunacağını ispat eder. (Kuzey, 2012).

DVM, verileri sınıflandırırken yalnızca istatistiksel kriterler dışında marj temelli geometrik bir ölçüt kullanmaktadır. Bu sayede, istatistiksel yöntemlere gereksinim duymadan marj maksimizasyonu ile sınıflandırma işlemini gerçekleştirmektedir (Melgani ve Bruzzone, 2004).

DVM yönteminde sınıflandırma yapılırken, bazı durumlarda veriler lineer bir doğruyla doğrudan sınıflara ayrılabilir. Bu doğru ile homojen sınıflar oluşturmak mümkün olmaktadır. Ancak, bazı durumlarda ise verileri homojen bir şekilde ayıran lineer bir düzlem bulunmamaktadır. Bu doğrultuda destek vektör makineleri; verilerin doğrusal ayrılma durumuna göre doğrusal ve doğrusal olmayan olmak üzere iki sınıfa ayrılmaktadır. Verilerin doğrusal ayrılma durumuna ilişkin görsel Şekil 3.6’da sunulmuştur.



Şekil 3.6. Doğrusal ayrılabilen ve ayrılamayan veri kümeleri

Doğrusal (Lineer) destek vektör makineleri

Her x_i örneğinin p adet niteliği olan girdi, $y_i \in \{1, -1\}$ ise örneklerin sahip olduğu sınıfı gösteren çıktı ve $x \in \mathbb{R}^d$ yüksek boyutlu girdi vektörünü temsil ettiği varsayılırsa; (x_i, y_i) ikililerinden oluşan n hacimli bir eğitim kümesi N verildiğinde, farklı sınıflara ilişkin verileri en iyi şekilde ayırt edecek aşağıdaki eşitlik sağlandığında,

$$w \cdot x + b \quad (1)$$

Eşitlikte yer alan w hiper düzlemin normali ve aynı zamanda ağırlık vektörü, b ise sabit değer olarak tanımlanmıştır. Gösterilen şekilde lineer hiper düzleminin bulunmasına imkan sağlayan denetimli makine öğrenme algoritmaları doğrusal destek vektör makineleri olarak adlandırılır. Aralarındaki uzaklık en fazla olan aşırı düzlemlerin üzerinde bulunan noktalar kümesi destek vektörleri olarak adlandırılır. Bu algoritmaların amacı, ayırt edilen iki sınıfa ilişkin destek vektörleri arasındaki uzaklığı maksimize eden bir ayırma hiper düzleminin bulunmasıdır.

Lineer destek vektör makineleri; verilerin doğrusal ayrılma ve belirli bir hata ile doğrusal ayrılma durumlarına göre iki sınıfa ayrılmaktadır. Doğrusal ayrılma durumunda; veri kümeleri direkt doğrusal bir hiper düzlem ile ayrılabilir. Bazı durumlarda ise doğrusal bir düzlem ile verilerin birbirinden ayrılması mümkün olmayabilir. Bu gibi problemlerde ise yanlış sınıflandırma ihtimalini azaltmak amacıyla bazı dönüşümler yapılarak problem kareli optimizasyon problemine dönüştürülür.

Doğrusal olmayan destek vektör makineleri

Gerçek hayatta karşılaşılan problemlerin birçoğunda verileri ayıran bir hiper düzlem bulmak mümkün olmamaktadır. Veri kümelerinin doğrusal bir fonksiyonla tam veya belirli bir hata ile ayrılamaması durumunda ise doğrusal olmayan sınıflandırıcılar kullanılır. Bu gibi durumlarda modele bazı değişkenler eklenerek çözüm sağlanır. Sınıflama işlemini gerçekleştiren fonksiyon lineer bir düzlem yerine en iyi ayırmayı yapan bir eğridir. Söz konusu eğriyi tespit etmek oldukça zor olmakla birlikte, doğrusal olmayan haritalama yaklaşımı yoluyla veri seti iki boyutlu düzlemden, üç boyutlu özellik uzayına transfer edilerek doğrusal ayırım sağlanır.

Veri setinin doğrusal olarak sınıflandırılmasının mümkün olmadığı durumlarda, her bir verinin, üç boyutlu özellik uzayına eşlenmesi ve bu uzayda doğrusal destek vektör makineleri yöntemi ile sınıflandırılması yöntemine çekirdek düzenlemesi adı verilir. Bileşenlerin, bir kural dizisine göre dönüştürülmesi işlemini esas alan çekirdek fonksiyonları (kernel functions), eğitim verisi üzerinde formülasyonun sağlanmasının ardından diğer verilere de uygulanabilecek hızlı bir çözüm sistemi oluşturur.

Destek Vektör Makineleri yönteminde en sık kullanılan çekirdek fonksiyonları aşağıda sıralanmıştır:

- Doğrusal Fonksiyon,
- Polinomiyal Fonksiyon,
- Radyal Tabanlı Fonksiyon,
- Sigmoid Fonksiyon.

Literatür araştırması sonucunda elde edilen bu bilgiler ışığında, tez çalışmasının bir sonraki bölümünde; akademik araştırma geliştirme çalışmalarını ve yenilikleri desteklemek amacıyla proje fonlayan bir kuruluşa araştırmacıların proje önerisi sunma eğilimine ilişkin analizler yapılmıştır. Analiz ve tahmin işlemlerinin gerçekleştirilmesinde, yukarıda bahsi geçen ve veri madenciliği çalışmalarında sıkça faydalanılan sınıflandırma yöntemlerinden Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri kullanılmıştır.

4. VERİ MADENCİLİĞİ YÖNTEMLERİ İLE PROJE BAŞVURU YAPMA EĞİLİMİNİN İNCELENMESİ

4.1. Problemin Belirlenmesi ve Çalışmanın Amacı

Araştırma projelerini değerlendiren ve destekleyen bir kurumda; farklı özelliklere sahip araştırmacılara hitap eden, başvuru dönemleri, destek miktarları, süreleri, başvuru kriterleri birbirinden farklı olan çok sayıda destek programı türü bulunmaktadır. Söz konusu proje destek programlarına üniversiteler, kamu kurumları ve özel kuruluşlarda görev yapan araştırmacılar bireysel olarak proje önerisinde bulunabilmektedir. Çoğu program türünde, başvuru sahibi kişinin, doktora ve dengi lisansüstü eğitim derecesine sahip olması başvuru şartlarından biridir. Söz konusu destek programlarına proje başvuruları kurumun çevrimiçi başvuru sistemi üzerinden yapılmaktadır. Bu sisteme giriş yapabilmek için bilim insanlarının öncelikle bilgi sistemine kayıtlı olmaları gereklidir. Söz konusu sistemde en az bir kez proje başvurusu yapan ya da yapmayan çok sayıda bilim insanının kaydı bulunmaktadır. Bilgi sistemine kayıt olabilmek için kişiler, cinsiyet, iş deneyimi, eğitim geçmişi, faaliyet alanı vb. özelliklerine ilişkin detaylı pek çok veriyi sisteme işlemektedir.

Proje destek programlarına yapılan başvuru sayısını artırmak ve programlara henüz hiç başvuru yapmayan bilim insanlarını söz konusu proje başvurularına özendirerek bilim ve teknoloji alanında rekabet gücünün artırılması, kurumun öncelikli hedefleri arasında yer almaktadır. Bu hedefler kapsamında kurumun proje yazma eğitimleri, tanıtım sunumları gibi pek çok özendirici faaliyeti bulunmaktadır. Ayrıca her bir program türüne özel güncelleme, iyileştirme ya da bilgilendirmeler başvuru yapan ya da yapması planlanan araştırmacılara e-posta ya da web duyurusu ile iletilmektedir.

Başvuru sayısını ve kalitesini artırmak üzere yapılan duyuru, anket, eğitim, tanıtım vb. faaliyetlerde sistemde kayıtlı araştırmacıların tümünü kullanmak yerine; ilgili destek programı türünün hedef kitlesi olan daha küçük bir sınıfı tercih etmek çalışmaların daha hızlı, verimli ve sonuç odaklı olmasını sağlayacaktır. Bu kapsamda, başvuru yapan ve başvuru potansiyeli bulunan bilim insanlarının ortak özelliklerinin incelenmesi ve sınıflandırılması, başvuru kararını etkileyen özelliklerin analiz edilmesi yapılacak çalışmalara ışık tutması bakımından büyük önem arz etmektedir.

Söz konusu kuruluşa proje başvurusu yapma potansiyeli olan araştırmacıların profilinin belirlenmesinin, ilgili kuruluşun çağrı bazında alacağı proje türlerinin öncelikli alanlarının belirlenmesinde önem arz edeceği düşünülmektedir.

Uygulamada; veri madenciliği yöntemleri aracılığıyla, fonlayıcı bir kuruluşa son on yıl içerisinde proje başvurusu yapan araştırmacıların başvuru yapma eğiliminde etkili olan özelliklerini belirlemek, araştırmacıları bu özelliklerine göre sınıflandırmak ve bundan sonraki dönemlerde yapacakları başvuru durumunu tahmin etmek amaçlanmıştır. Veriye en uygun sınıflandırma yöntemini kullanmak amacıyla; Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri ile sınıflandırma işlemi gerçekleştirilerek doğruluk oranları karşılaştırılmıştır.

Modelleme gereci olarak SPSS IBM Modeler programı kullanılmıştır. Bu program, veri madenciliği çözümleri ile hem istatistik kökenli algoritmaları hem de yapay zekâ kökenli algoritmaları görsel bir programlama ara yüzü altında sunmaktadır.

4.2. Verinin Tanımı

Uygulamada, fonlayıcı kuruluşun başvuru sisteminde mevcut olan geniş kapsamlı veri setinden faydalanılmıştır. Kullanılan veri seti; 2007-2017 yılları arasında kurumun hizmetlerinden faydalanan ya da proje başvurusu yapabilmek için gerekli şartları sağlayan tüm kişilerin bilgi sistemine kayıtlı 68.484 satır (kişi sayısı) ve 40 sütundan (kayıtlı özellik sayısı) oluşan kapsamlı bir veri kümesidir. Her bir araştırmacıya ait 40 adet değişken (özellik) içeren bu ham listede kişilerin aşağıdaki özelliklerinin detaylı verisi yer almaktadır:

Demografik bilgileri (Doğum yılı, cinsiyet, uyruk)

Eğitim Bilgileri (kişilerin lisans/yüksek lisans/doktora mezuniyet durumları ve tarihleri, diplomalarını aldıkları kurumlar, bu kurumların fakülte ve bölümleri vb.)

Çalışma Bilgileri (görev yaptıkları kurum tipi, adı, çalışma alanı, unvan vb.)

Proje destek programlarına başvurularına ilişkin bilgileri (son on yılda başvuru yaptıkları proje türleri, başvuru yapılan proje sayısı)

Bu bilgileri içeren geniş hacimli excel listesinin görüntüsünden bir kesit Çizelge 4.1’de sunulmuştur:

Çizelge 4.1. Ham veriden bir kesit

Id No	Doğum Yılı	Cinsiyet	Bölüm	Diploma Turu	Mezuniyet Yılı	Görev Unvan	Kurum Tipi	Şehir Ad
44661	1969	KADIN	Fizik	Yüksek Lisans	1994	Araştırma Görevlisi	Kamu	Kocaeli
44664	1968	ERKEK	Tıp	Lisans	1989	Doç. Dr.	Üniversite	İstanbul
44678	1973	ERKEK	Kimya Mühendisliği	Lisans	1994	Doç. Dr.	Kamu	Kocaeli
44826	1936	KADIN	Eczacılık	Doktora	1959	Prof. Dr.	Üniversite	İstanbul
44844	1958	KADIN	İnşaat Mühendisliği	Lisans	1979	Prof. Dr.	Üniversite	İzmir
44848	1971	KADIN	Bahçe Bitkileri	Doktora	1995	Yrd. Doç. Dr.	Üniversite	Osmaniye
44856	1960	KADIN	Mimarlık	Lisans	1984	Doç. Dr.	Üniversite	İstanbul
44857	1947	ERKEK	Kimya	Lisans	1971	Prof. Dr.	Üniversite	Ankara
44869	1977	ERKEK	İşletme	Yüksek Lisans	1999	Yrd. Doç. Dr.	Üniversite	Ankara
44877	1977	ERKEK	Bilgisayar Mühendisliği	Doktora	2001	Öğretim Görevlisi	Ticari (Özel)	İstanbul

Veri ön işleme süreci öncesi, ham veri setinin daha çok kayıt içeren bir bölümü EK-1’de ayrıca sunulmuştur.

4.3. Veri Ön İşleme

Veri madenciliği sürecinin en önemli adımlarından olan verinin ön işlenmesi aşamasında, sonuçların net ve anlaşılır olması amacıyla veride düzenlemeler yapılarak modelde kullanılacak olan değişkenler aşağıdaki şekilde sıralanmıştır:

“ID NO” sütunu, sisteme kayıtlı her bir bireyin ayırt edici numarası olarak tanımlanan bir değişkendir. Ham veride bu değişken, kişilerin lisans/yüksek lisans/doktora eğitimlerine ilişkin bilgilerine göre tekrar etmektedir. Program aracılığıyla bu satırlar sütunlara dönüştürülmüş ve “ID NO” değişkeninin satır bazında tekrar etmemesi sağlanmıştır.

Kişilerin sisteme manuel olarak giriş yaptıkları Anabilim Dalı, Fakülte, Bölüm gibi veriler net bir sınıflandırma yapabilmek amacıyla 5 ana alana (Sağlık Bilimleri, Sosyal ve Beşeri

Bilimler, Tarımsal Bilimler, Teknik Bilimler, Temel Bilimler kategorize edilmiş ve “Faaliyet Alanı” değişkeni oluşturulmuştur.

Karar ağaçlarının yapısının net ve kolay anlaşılır olması amacıyla kişilerin çalıştığı kurumlar ilk olarak; “Kamu”, “Özel Sektör” ve “Üniversite” olarak gruplandırılmıştır. Ardından üniversiteler, daha net bir sınıflandırma oluşturulması amacıyla, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından yayımlanan; üniversitelerin girişimcilik ve yenilikçilik performanslarına göre sıralandığı, “Girişimci ve Yenilikçi Üniversite Endeksi” 2017 yılı sıralamasına göre 5 gruba ayrılarak yeni bir değişken oluşturulmuştur. Türkiye’de olup bu listede yer almayan ya da yurtdışında yer alan üniversiteler ise “Diğer” olarak nitelendirilmiştir. 2017 yılında Türkiye Bilimsel ve Teknolojik Araştırma Kurumu tarafından hazırlanan, farklı göstergelere göre ülkemizin en girişimci ve yenilikçi 50 üniversitesinin sıralanmasından oluşan liste EK-2’de sunulmuştur. Kayıtlı kişilerin çalıştığı kurumun yer aldığı şehir verisi, genel kapsamlı sonuç elde edebilmek amacıyla şehrin yer aldığı coğrafi bölge bilgisine dönüştürülmüştür.

Lisans, yüksek lisans ve doktora eğitimleri verisine ilişkin, diplomanın alındığı ülke değişkeninin alt kategorilerinin fazla olması nedeniyle en fazla kayıt bulunan ülkeler (Türkiye, İngiltere, ABD, Almanya, Fransa, Kanada) haricindeki ülkeler “Diğer” seçeneği olarak belirlenmiş ve yeni bir değişken olarak tanımlanmıştır.

Yapılacak sınıflandırmanın amacının proje başvurusu yapma nihai sonucunun analizi olması nedeniyle, herhangi bir proje türüne en az bir kere başvuruda bulunan kişilerin durumu başvuru var (BVAR), henüz hiç başvuru yapmamış olan kişilerin durumu ise başvuru yok (BYOK) olarak nitelendirilmiştir. IBM Modeler programı üzerinden, sınıflandırmada kullanılacak modellere ilişkin veri hazırlık sürecinin ön işleme algoritması Şekil 4.2’de sunulmuştur:

Model	Build Time	No. Fields Used
 CHAID 1	1	9
 Neural Net 1	1	10
 LSVM 1	< 1	10

Şekil 4.2. Veri setine uygun sınıflandırma yöntemleri

Veri setine uygun algoritmalarından üç tanesi Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri olarak bulunmuştur. CHAID algoritması; değişkenler arasındaki ilişkinin ötesinde, değişkenlerin içerisindeki veri grupları arasındaki ilişkilerin de en alt detayına kadar araştırılmasına olanak sağlayan Karar Ağacı algoritmasıdır. Karar ağacı yapısını, istatistikte sıkça kullanılan ki-kare test yöntemi ile oluşturan algoritma, bu yönüyle istatistiğe dayalı bir algoritmadır. Bu yöntem, karar ağaçları algoritmaları içinde sürekli ve kategorik tüm değişken tipleri ile çalışabilmesi nedeni ile yaygın olarak kullanılmaktadır (Boysan ve Kayri, 2007). Çalışma kapsamında kullanılan veri setinde de yaş, tarih, başvuru sayısı gibi sürekli değişkenlerin yanı sıra, cinsiyet ve faaliyet alanı gibi kategorik değişkenler de bulunmaktadır. CHAID yöntemi, sürekli ve kategorik tüm değişken tiplerinin bulunduğu veri setlerinde, normal dağılımın şartlarını sağlama zorluğuna ve sıralı değişkenlerin gerekliliğine tabi olmayan bir teknik olması sebebiyle bu çalışma için en ideal Karar Ağacı algoritması olarak belirlenmiştir (Diepen ve Franses, 2006). Ayrıca bölünme sayısının ikiden fazla olduğu bu yöntemle oluşturulan karar ağaçlarını yorumlamak görsel açıdan oldukça kolaydır.

Model oluşturma sürecinde, öncelikle üzerinde ön işleme yapılan veri seti programa tanımlanmış; kişilere ilişkin veride yer alan revize edilmiş özellikler (bağımsız değişken), modelin girdileri olup, bu değişkenler şekil 4.4’de sunulmuştur:

Field	Measurement	Values	Missing	Check	Role
Blm Recono	Continuous	[44477.0,361759.0]	None		Input
Doğum Yılı	Continuous	[1901.0,1995.0]	None		Input
Cinsiyet	Nominal	BLINMIYOR,ERKEK,KADIN	None		Input
Yaş	Continuous	[21.0,115.0]	None		Input
Uyruk	Nominal	"ABD,Afganistan,Almanya,Ankara,Amanoluk,Ava...	None		Input
Lisans_Yük Anabilim Dalı	Flag	"	None		Input
Yüksek Lisans_Yük Anabilim Dalı	Nominal	"1. Beyin Cerrah...	None		Input
Doktora_Yük Anabilim Dalı	Nominal	"1. CERRAHI KLİNİĞİ","1. Ortopedi ve ...	None		Input
Lisans_Mezun Olduğu Kurum	Continuous	[50.0,25281.0]	None		Input
Lisans_Eğitimi Alan Kurum	Nominal	"1. Sınıf","2. Sınıf","3. Sınıf","4. Sınıf","5. Sınıf" Diğer	None		Input
Yüksek Lisans_Mezun Olduğu Kurum	Continuous	[50.0,25281.0]	None		Input
Yüksek Lisans_Eğitimi Alan Kurum	Nominal	"1. Sınıf","2. Sınıf","3. Sınıf","4. Sınıf","5. Sınıf" Diğer	None		Input
Doktora_Mezun Olduğu Kurum	Nominal	"1065","1738","2041","2220","2559","2744","3132"...	None		Input
Doktora_Eğitimi Alan Kurum	Nominal	"1. Sınıf","2. Sınıf","3. Sınıf","4. Sınıf","5. Sınıf" Diğer	None		Input
Lisans_Eğitim Ülke	Nominal	ABD,Almanya,Diğer,Fransa,İngiltere,Kanada,Türkiye	None		Input
Yüksek Lisans_Eğitim Ülke	Nominal	ABD,Almanya,Diğer,Fransa,İngiltere,Kanada,Türkiye	None		Input
Doktora_Eğitim Ülke	Nominal	ABD,Almanya,Diğer,Fransa,İngiltere,Kanada,Türkiye	None		Input
Lisans_Mezuniyet Yılı	Continuous	[1922.0,6034.0]	None		Input
Yüksek Lisans_Mezuniyet Yılı	Continuous	[1951.0,6046.0]	None		Input
Doktora_Mezuniyet Yılı	Continuous	[1990.0,14020.0]	None		Input
Lisans_Fakülte	Nominal	"[Fakülte denen birşey yokur]","2. Tıbbi Profila...	None		Input
Yüksek Lisans_Fakülte	Nominal	"[School of Agriculture]","DEPARTMENT OF WA...	None		Input
Doktora_Fakülte	Nominal	"[School of Agriculture]","...","1. Beyin Cerra...	None		Input
Lisans_Bölüm	Nominal	"[Değişim ile 1 sene eğitim]","...","..."	None		Input
Yüksek Lisans_Bölüm	Flag	"ENDÜSTRİ MÜHENDİSLİĞİ"	None		Input
Doktora_Bölüm	Nominal	"Biyoloji","DİŞ HEKİMLİĞİ","ENDÜSTRİ MÜHENDİS...	None		Input
Görev Ünvanı	Nominal	"Araştırma Görevlisi","Araştırma Görevlisi (Dr.)"	None		Input
Çalıştığı Kurum Adı	Nominal	"1065","1738","2041","2220","2559","2744","3132"...	None		Input
Çalıştığı Bölüm / Birim	Nominal	"...","Game Development Certification Program"...	None		Input
Çalıştığı Kurum	Nominal	"1. Sınıf","2. Sınıf","3. Sınıf","4. Sınıf","5. Sınıf" Diğer Ka...	None		Input
Şehir Adı	Nominal	"Adana,Adıyaman,Afyon,Aksaray,Alanya,Almatı,Am...	None		Input
Çalıştığı Şehir Bölge	Nominal	"Ardıcılık B","D. Anadolu B","Diğer","Ege B","G. Doğu...	None		Input
İsyan Kadrosu	Nominal	"Araştırma Görevlisi","Araştırma Görevlisi (Dr.)","A...	None		Input
Kurum Tipi	Nominal	Kamu,Özel,Üniversite	None		Input

Şekil 4.3. Modelde kullanılan bağımsız değişkenler

Programda, modelin veriyi öğrenmesi ve sınıflandırması için verinin bir bölümü eğitim (Training), kalan bölümü ise test (Testing) olmak üzere ikiye bölünmektedir. Bu sayede model öğretisi test edilebilmektedir. Bu uygulama neticesinde modelin veriyi ne denli iyi anladığı ve örüntüleri keşfettiği yorumlanabilir. Bu çalışmada, ilgili verinin %80'i eğitim, %20'i ise test verisine ayrılmıştır. Bu ayrım, Şekil 4.5'de sunulmuştur:

Partition

Generate Preview

Settings Annotations

Partition field: Partition

Partitions: ☒ Train and test ☐ Train, test and validation

Training partition size: 80 Label: Training Value = "1_Training"

Testing partition size: 20 Label: Testing Value = "2_Testing"

Validation partition size: 10 Label: Validation Value = "3_Validation"

Total size: 100%

Values: ☐ Use system-defined values ("1", "2" and "3") ☒ Append labels to system-defined values ☐ Use labels as values

☒ Repeatabable partition assignment

Seed: 1234567 Generate

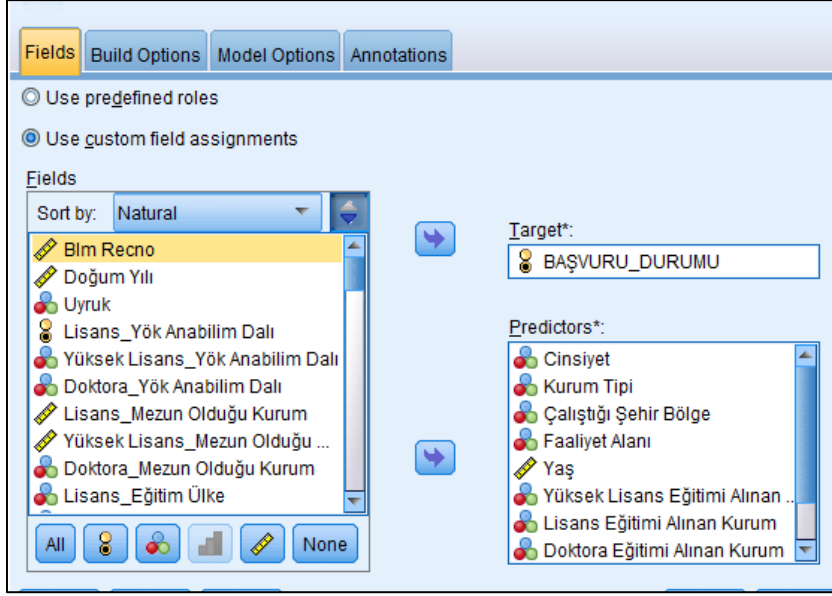
☐ Use unique field to assign partitions: []

OK Cancel Apply Reset

Şekil 4.4. Verinin bölünmesi (partition)

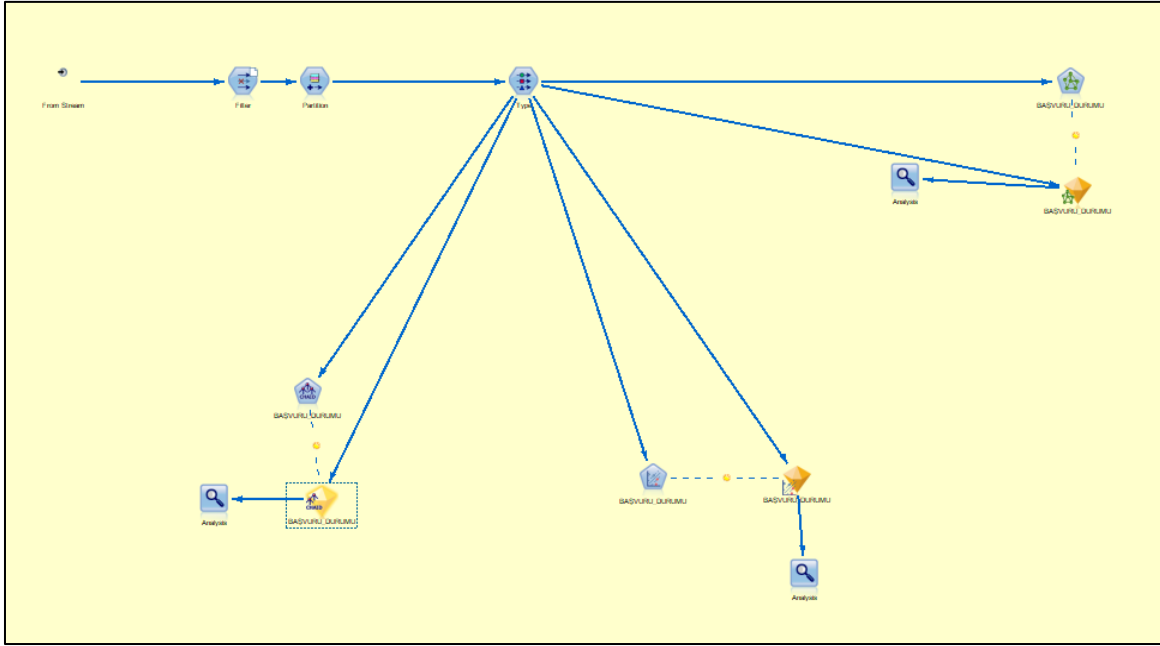
Verinin sınıflandırma yapılmak üzere bölünmesinin ardından, seçilen sınıflandırma yöntemlerinin uygulanabileceği bir model oluşturulmuştur. Model tasarlanması sürecinde,

veri kaynağına uygulanan her bir algoritmanın varsayılan parametreleri ile çalışılmıştır. Bunun sebebi; algoritmaların başarı düzeyini maksimize eden parametreler ile çalışmanın karşılaştırma sonuçlarına edeceği etkiyi ve subjektif bir karşılaştırma yapılması durumunu en aza indirmektedir. Her bir modelin varsayılan parametreleri ile şekil 4.6’de belirtilen girdi ve hedef değişkenler tanımlanmıştır.



Şekil 4.5. Modellerin girdi ve hedef değişkenleri

Model kurulmasına ilişkin yukarıda bahsedilen hazırlık sürecinin ardından her bir sınıflandırma yöntemi için ayrı ayrı model kurulmuş ve modelde yer alan bağımlı ve bağımsız değişkenler tanımlanmıştır. Oluşturulan modellere ilişkin veri akışı ve ekran görüntüsü şekil 4.7’de yer almaktadır:



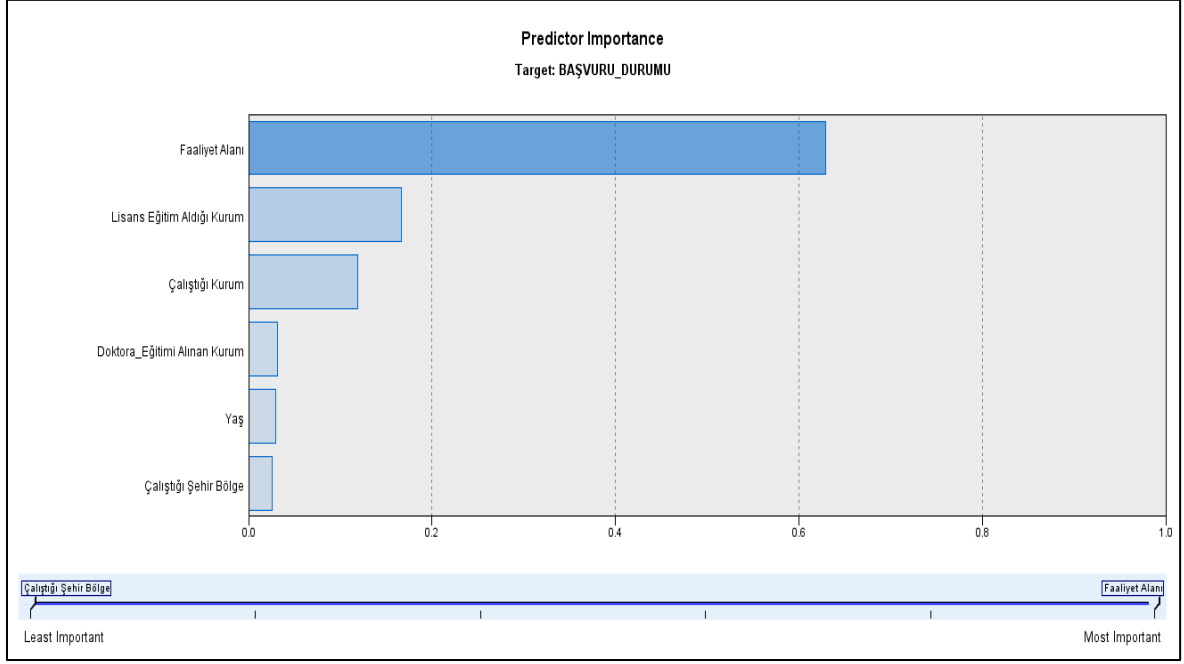
Şekil 4.6. Oluşturulan modellere ilişkin ekran görüntüsü

Modellerin kurulmasının ardından, her bir algoritma için program çalıştırılmış ve veri akışı gerçekleşmiştir. Oluşturulan model çıktılarına analiz işlemcisi eklenerek çıktılarına ilişkin performans ölçülmüştür.

4.5. Tahmin Modellerini Etkileyen Değişkenlerin Belirlenmesi

Uygulamaya ilişkin model kurulmasının ardından program aracılığıyla veri akışı gerçekleşmiş ve sınıflandırma ve tahmin modellerini etkileyen değişkenler her bir yöntem için belirlenmiştir.

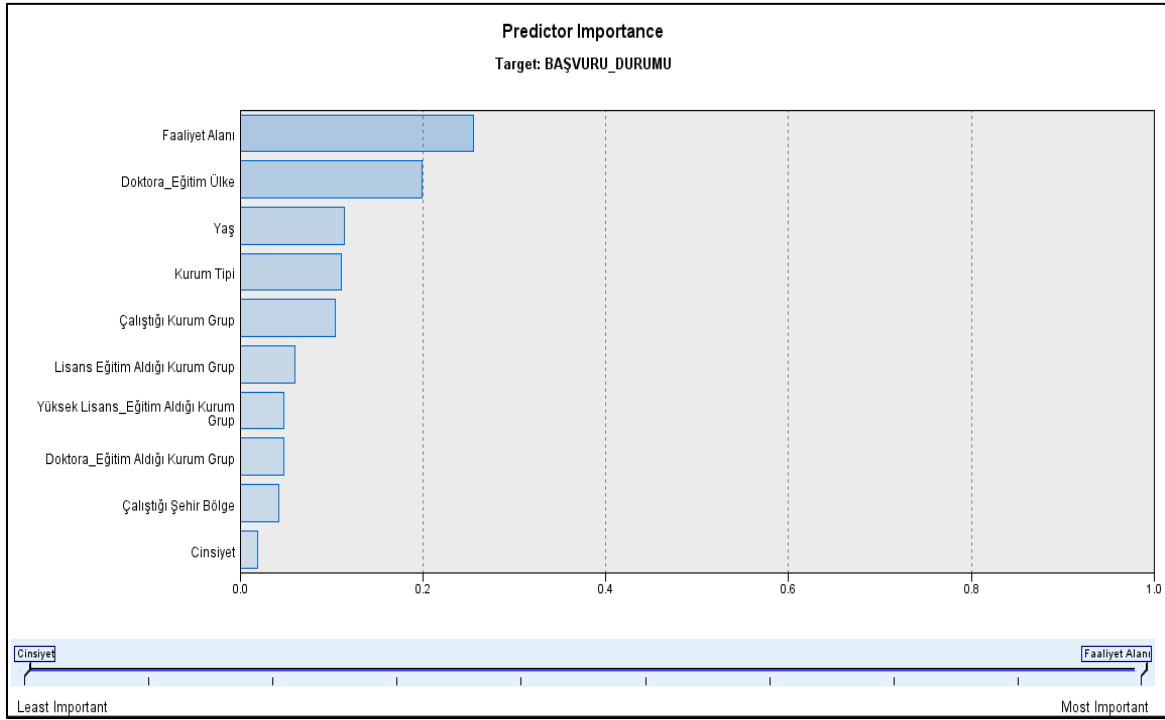
İlk olarak Karar Ağaçları ile oluşturulan modele bağımlı değişken olarak kaydedilen değişkenlerin başvuru yapma nihai sonucu etkileyenleri belirlenmiştir. Karar Ağaçları ile oluşturulan modele ilişkin algoritmaya göre, kişilerin başvuru yapma eğilimini en fazla etkileyen ve karar ağaçlarının oluşmasını sağlayan değişkenlerin histogramı Şekil 4.8’de sunulmuştur:



Şekil 4.7. Karar ağacı modeline etkisi olan değişkenler

Karar Ağaçları yöntemi ile oluşturulan, programlara başvuru yapma eğiliminde etkisi bulunan değişkenlere ilişkin histogram incelendiğinde; sınıflandırma ve tahmin işleminde en etkili olan değişkenin diğerlerine göre farkla kişilerin görev yaptığı “Faaliyet Alanı” değişkeninin olduğu görülmektedir. Buna göre, araştırmacıların görev yaptığı faaliyet alanı proje başvurusu yapma eğilimlerinde oldukça etkili bir değişken olarak bulunmuştur. Programlara başvuru yapma eğiliminde etkisi bulunan ikinci değişken, araştırmacıların lisans eğitimini almış olduğu kurum değişkenidir. Bununla birlikte kişilerin çalıştığı kurum, doktora eğitimi aldığı kurum, yaş ve çalıştıkları kurumun içinde yer aldığı şehir değişkenlerinin de sırasıyla sınıflandırmada az da olsa etkili olan özellikler olduğu dikkat çekmektedir.

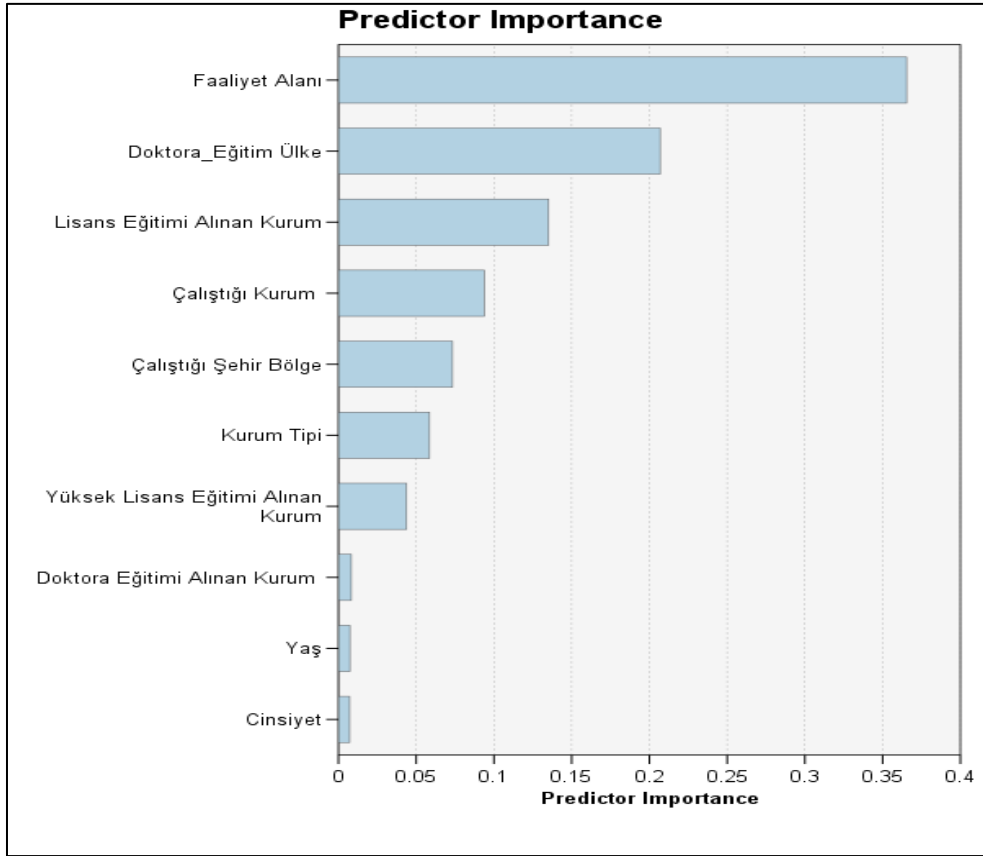
Uygulamada kullanılan ikinci sınıflandırma yöntemi olan Yapay Sinir Ağları ile oluşturulan modelin tahmin sonucuna katkıda bulunan bağımsız değişkenlerin önemlilik sırasına göre listelendiği grafik ise Şekil 4.9’da sunulmuştur:



Şekil 4.8. YSA modeline etkisi olan değişkenler

Histogram incelendiğinde; sınıflandırma ve tahmin işleminde en etkili olan değişken bu yöntemde de kişilerin görev yaptığı “Faaliyet Alanı” olarak belirlenirken, Karar Ağaçları ile oluşturulan modelden farklı olarak kişilerin doktora eğitimini almış olduğu ülkelerin, proje başvurusu yapma eğilimlerinde etkili bir değişken olduğu bu modelde tespit edilmiştir. Ayrıca bir önceki modelde etkili bulunmayan “Yüksek Lisans Eğitiminin Alındığı Kurum”, “Cinsiyet” gibi bazı değişkenlerin de bu modelde etkisi olduğu saptanmıştır. Yapay Sinir Ağları modelinin değişken önem sıralaması ve sayısı incelendiğinde, Karar Ağaçları modelinden çok daha fazla sayıda değişkenin modeli etkilediği belirlenmiştir. Tahmin modellerini etkileyen değişkenlerin sayısı ve önemlilik derecelerinin farklı olmasının sebebi, yöntemlerin model kurarken izlediği algoritmaların farklı olması olarak düşünülebilir.

Destek Vektör Makineleri ile oluşturulan modelin tahmin sonucuna katkıda bulunan bağımsız değişkenlerin önemlilik sırasına göre listelendiği grafik Şekil 4.10’da sunulmuştur:



Şekil 4.9. DVM modeline etkisi olan değişkenler

Destek Vektör Makineleri yöntemi ile oluşturulan tahmin modelinde de diğer modellerde olduğu gibi en etkili değişken “Faaliyet Alanı” olarak belirlenmiştir. Söz konusu değişkenin önem oranı bu modelde Karar Ağaçları ve Yapay Sinir Ağları ile oluşturulan modellere göre daha yüksektir. Sınıflandırma sonucuna en az etkisi olan değişkenler ise “Yaş” ve “Cinsiyet” olarak tespit edilmiştir. Modelde toplam 10 adet bağımsız değişken önemli bulunmuştur. Değişken sayısı, Yapay Sinir Ağları ile oluşturulan modele benzerlik göstermekle birlikte, Karar Ağaçları ile oluşturulan modelde kullanılan değişken sayısından oldukça fazladır.

Değişken önemlerinin belirlendiği modellerin çalıştırılmasıyla elde edilen bulgulara ve modellerin karşılaştırılmasına çalışmanın bir sonraki bölümünde yer verilmiştir.

5. BULGULAR

Bu bölümde, proje başvurusu yapma eğilimini tahmin etmek amacıyla Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri yöntemleri ile kurulan modellerin çıktıları sunulmuş ve yorumlanmıştır. Sistemde kayıtlı araştırmacıların proje başvuru yapma eğilimleri incelenmiş ve sınıflandırma sonuçlarında dikkat çeken hususlar yorumlanmıştır. Modelde ve karar ağacı yapısında yer alan ve sınıflandırma yapılan her bir kayıt, proje başvurusu yapan/yapmaya aday olan bir bilim insanını temsil etmektedir.

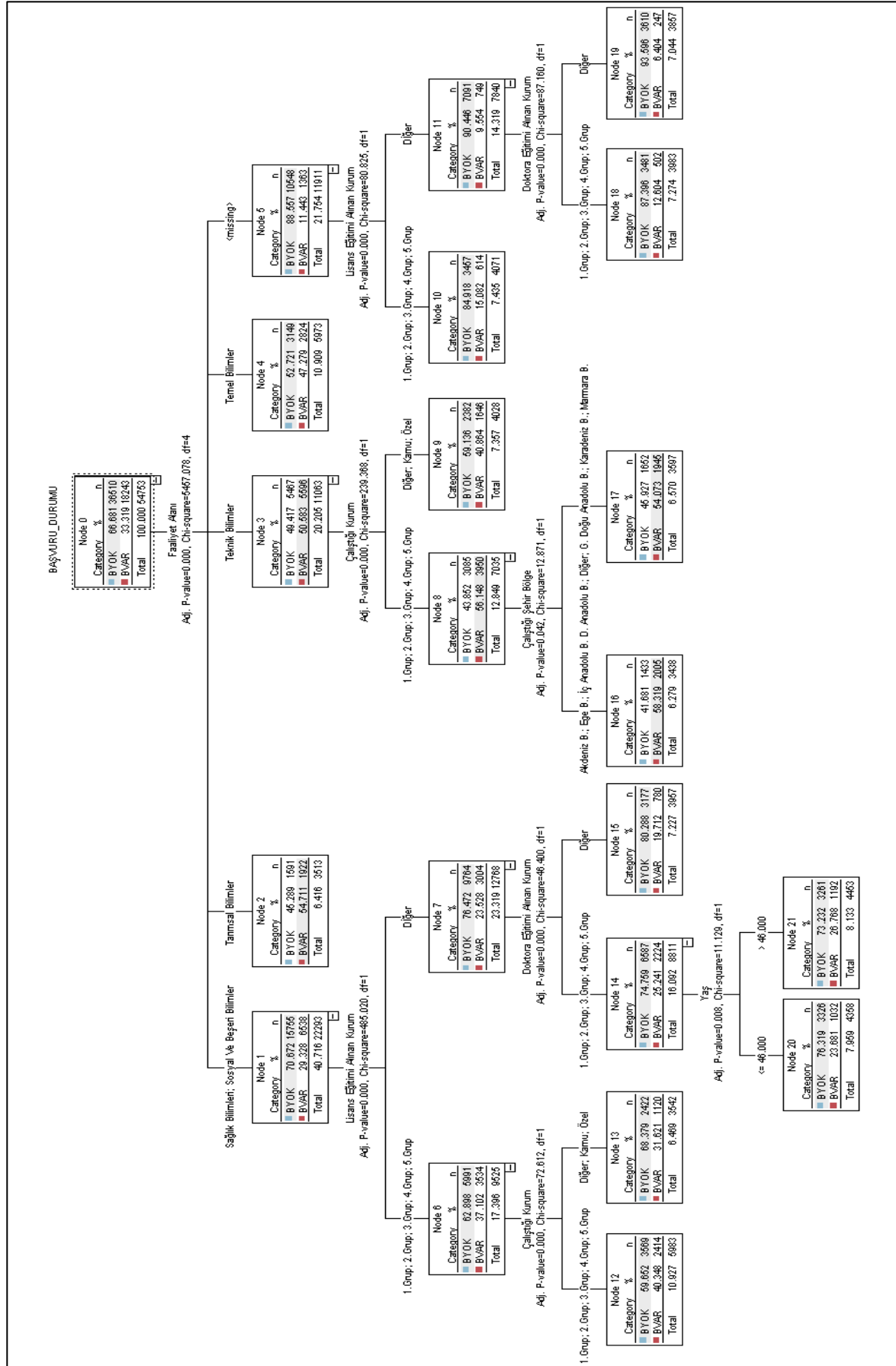
5.1. Karar Ağaçları Yöntemi ile Analiz

IBM Modeller Programı aracılığıyla veri setine en uygun karar ağacı algoritması olarak CHAID algoritması belirlenmiştir. Daha önce yöntemin belirlenmesi bölümünde bahsedilen avantajlara sahip olan algoritma ile proje başvuru yapma eğilimine ilişkin karar ağacı oluşturulmuştur.

Modele ilişkin karar ağacı yapısı şekil 5.1’de sunulmuştur. Başvuru durumuna ilişkin oluşturulan karar ağacı yapısı incelendiğinde; önemli hususlar aşağıdaki şekilde yorumlanmıştır:

Veri tabanında kayıtlı araştırmacılardan, programın test için seçtiği 54753 araştırmacıdan 18243’ü proje türlerinden herhangi birine başvuru yapmıştır. Bu sayı seçilen örneklemin %33’ünü oluşturmaktadır. Bununla birlikte, sisteme kayıtlı araştırmacıların yaklaşık %66’sı henüz proje başvurusunda bulunmamış kişilerden oluşmaktadır.

Proje başvurusu yapma nihai kararına en fazla etkisi olan bağımlı değişkenin, araştırmacıların görev yaptığı “Faaliyet Alanı” olduğu tespit edilmiştir. Buna göre; sisteme kayıt olan araştırmacılardan başvuru yapma eğilimi en yüksek oranda olan araştırmacıların tarımsal bilimler ve teknik bilimler alanında (Node 2 ve Node 3) çalışan araştırmacılar olduğu belirlenmiştir. Bu başvuru oranlarını temel bilimler ve sağlık bilimleri ile sosyal bilimler alanları takip etmektedir. (Node 4 ve Node 1). Karar ağacı yapısında, sınıflandırmanın ilk seviyesinde dikkat çeken bir diğer husus sisteme eksik bilgi ile kayıt olan araştırmacı sayısının oldukça fazla olmasıdır (Node 5).



Şekil 5.1. Karar ağacı sonuç özeti

Karar ağacı yapısının oluşmasında önemli olan ikinci değişkenler araştırmacıların lisans eğitimini aldıkları kurum ve çalıştıkları kurum değişkenleri olmuştur. Buna göre; Sağlık Bilimleri ile Sosyal ve Beşeri Bilimler faaliyet alanında görev yapan araştırmacılardan, girişimcilik ve yenilikçilik performanslarına göre sıralamada yer alan üniversitelerde lisans eğitimi alan araştırmacıların başvuru yapma eğilimleri diğerlerine göre daha yüksek orana sahiptir (Node 6 ve Node 10). Teknik Bilimler faaliyet alanında çalışan araştırmacılardan, üniversitelerde görev yapanların özel sektör ve kamuda görev yapanlara göre daha yüksek oranda başvuru yaptığı belirlenmiştir (Node 8).

Teknik Bilimler faaliyet alanında çalışan ve üniversitelerde görev yapan araştırmacıların sınıflandırılması işleminde etkili olan değişken türü kişilerin çalıştığı şehrin içinde yer aldığı coğrafi bölge değişkeni olmuştur. Bu araştırmacılar; Akdeniz, Ege ve İç Anadolu bölgelerinde görev yapan araştırmacıların başvuru oranı az farkla diğerlerine göre daha yüksek bulunmuştur (Node 16).

Karar ağacı yapısının son seviye dallanma yapısında en az etkili olan değişken yaş değişkenidir. Sağlık Bilimleri ve Sosyal Bilimler alanlarında çalışan araştırmacılardan, 46 yaşın üzerinde olanların başvuru oranı diğerlerine yakın olmakla birlikte % 3 daha fazladır (Node 21).

Karar Ağaçları yöntemi ile oluşturulan modelin başarısını ölçmek amacıyla, modele analiz işlemcisi eklenmiş ve oluşturulan ağacın doğruluk oranının (accuracy) %69,54 olduğu gözlenmiştir. Modelin doğruluk oranı şekil 5.2’de sunulmuştur:

Results for output field BAŞVURU_DURUMU

Comparing \$R-BAŞVURU_DURUMU with BAŞVURU_DURUMU

'Partition'	1_Training		2_Testing	
Correct	38,340	70.02%	9,528	69.54%
Wrong	16,413	29.98%	4,173	30.46%
Total	54,753		13,701	

Şekil 5.2. Karar ağaçları modeline ilişkin doğruluk oranı

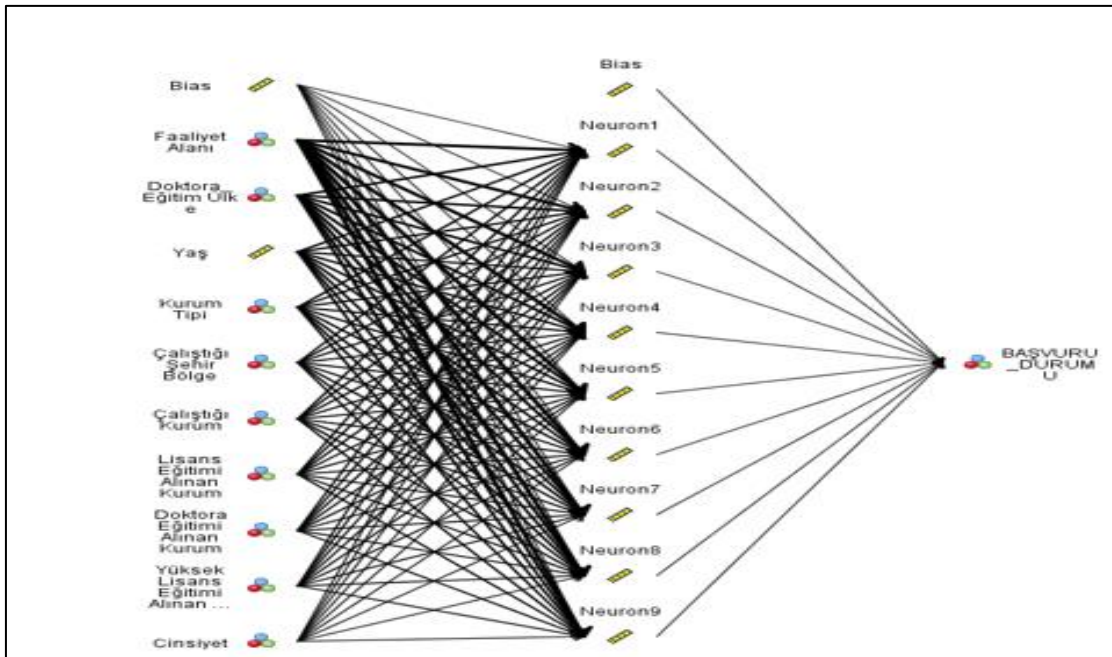
5.2. Yapay Sinir Ağları Yöntemi ile Analiz

Yapay Sinir Ağları yöntemine ilişkin modelin çalıştırılmasıyla tahmin sonuçları elde edilmiştir. Tahmin sonuçlarına ilişkin verilerin örneği EK-3’de sunulmuştur. En yüksek sınıflandırma performansını sağlaması amacıyla çok katmanlı yapay sinir ağı mimarisi kullanılan modele ilişkin özet şekil 5.3’de yer almaktadır.

Model Summary	
Target	BAŞVURU_DURUMU
Model	Multilayer Perceptron
Stopping Rule Used	Minimum accuracy exceeded
Hidden Layer 1 Neurons	9

Şekil 5.3. YSA modeli özeti

YSA ile oluşturulan analiz sonucuna göre, girdi katmanında 10 değişken, bir adet gizli katmanda 9 düğüm bulunmaktadır. Çıktı katmanında ise hedef değişken olan birim yer almaktadır. Modelin durdurma kriteri olarak elde edilebilecek en iyi doğruluk oranının hesaplanması seçilmiştir. YSA modelinin mimarisi Şekil 5.4’de gösterilmektedir.



Şekil 5.4. YSA ağ mimarisi

Yapay Sinir Ağları yöntemi ile oluşturulan modelin başarısını ölçmek amacıyla, modele analiz işlemcisi eklenmiş ve oluşturulan ağacın doğruluk oranının (accuracy) %66,6 olduğu gözlenmiştir. Modelin doğruluk oranı şekil 5.5’de sunulmuştur:

Results for output field BAŞVURU_DURUMU				
Comparing \$N-BAŞVURU_DURUMU with BAŞVURU_DURUMU				
'Partition'	1_Training		2_Testing	
Correct	36,419	66.52%	9,125	66.6%
Wrong	18,334	33.48%	4,576	33.4%
Total	54,753		13,701	

Şekil 5.5. YSA modeline ilişkin doğruluk oranı

5.3. Destek Vektör Makineleri Yöntemi ile Analiz

Destek Vektör Makineleri ile oluşturulan modele de bağımlı ve bağımsız değişkenler tanımlanmış, model parametreleri verinin yapısına uygun olarak seçilmiştir. Veriye en uygun matematiksel modelin lineer destek vektör makineleri ile oluşturulduğu bilgisi, program aracılığıyla elde edilmiştir. Modele ilişkin ceza parametresi 0,100 olarak hesaplanmıştır. DVM ile oluşturulan modele ilişkin özet bilgi çizelge 5.1’de sunulmuştur.

Çizelge 5.1. DVM model özeti

Target Field	BAŞVURU_DURUMU
Model Building Method	Linear SVM
Number of Predictors Input	10
Number of Predictors in Final Model	10
Regularization Type	L2
Penalty Parameter (Lambda)	0.100

Modelde yer alan değişkenlere ilişkin parametreler EK-4’de ayrıca sunulmuştur.

Destek Vektör Makineleri ile oluşturulan modelin başarısını ölçmek amacıyla, modele analiz işlemcisi eklenmiş ve oluşturulan ağacın doğruluk oranının (accuracy) %70,24 olduğu gözlenmiştir. Modelin doğruluk oranı şekil 5.6’da sunulmuştur:

Results for output field BAŞVURU_DURUMU				
Comparing \$L-BAŞVURU_DURUMU with BAŞVURU_DURUMU				
'Partition'	1_Training		2_Testing	
Correct	38,693	70.67%	9,623	70.24%
Wrong	16,060	29.33%	4,078	29.76%
Total	54,753		13,701	

Şekil 5.6. DVM modeline ilişkin doğruluk oranı

5.4. Tahmin Modellerinin Doğruluk Oranlarının Karşılaştırılması

Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri ile yapılan analizler sonucunda elde edilen doğruluk oranları çizelge 5.2’de karşılaştırılmıştır.

Çizelge 5.2. Modellerin performansına ilişkin özet çizelge

YÖNTEM	Eğitim Verisi				Test Verisi			
	Toplam	Doğru	Yanlış	Doğruluk Oranı %	Toplam	Doğru	Yanlış	Doğruluk Oranı %
KARAR AĞAÇLARI	54753	38340	16413	70,02	13701	9528	4173	69,54
YSA	54753	36419	18334	66,52	13701	9125	4576	66,60
DVM	54753	38693	16060	70,67	13701	9623	4078	70,24

Çizelge 5.2. incelendiğinde, tüm yöntemlerin kendi içinde tutarlı ve doğruluk oranları birbirine yakın sonuçlar verdiği gözlenmiştir. Ancak sıralama yapıldığında DVM yönteminin en yüksek doğruluk oranına sahip model olduğu belirlenmiştir (%70,24). Analiz sonuçlarına göre ikinci sırada Karar Ağaçları yönteminin doğruluk oranı (%69,54) gelmektedir. Sıralamada son olarak YSA yönteminin uygulama yapılan veri üzerinde en az doğruluk oranına (%66,6) sahip olduğu analiz sonucunda ortaya çıkmıştır.

Her üç yöntemin de yakın doğruluk oranları ile sınıflandırma yaptığı dikkate alındığında, yöntemlerin kendine özgü avantaj ve dezavantajlarının tercih sebebi olarak öne çıkacağı söylenebilir. Destek Vektör Makineleri yöntemi, diğer yöntemlere kıyasla daha hızlı sonuç vermiştir. CHAID algoritması diğer algoritmalara göre daha az değişken ile karar ağacı

oluşturmuştur. Karar Ağaçları yöntemi görsel olarak açıklayıcı ve kullanıcı dostu sonuçlar ortaya çıkarırken, Yapay Sinir Ağları ve Destek Vektör Makineleri yöntemleri ise genelleme yapabilmeleri, herhangi bir varsayıma gerek duymadan öğrenme işlemlerini gerçekleştirmeleri ve doğrusal olmayan veri kümelerinde de işlem yapabilmeleri sebepleriyle tercih edilebilir.

6. SONUÇ VE ÖNERİLER

Büyük veri yığınları içerisinde gizli örüntüleri keşfeden, ham veriyi anlaşılabilir ve yorumlanabilen faydalı bilgi haline getiren süreç veri madenciliği süreci olarak tanımlanabilir. Veri madenciliğinin amacı; bir anlam ifade etmeyen, karar verme üzerinde etkisi bulunmayan ham veriyi işleyerek faydalı bilgi haline getirmek ve bu bilgilerden yola çıkarak gelecekteki faaliyetlere yönelik tahminlerde bulunmaya olanak sağlamaktır.

Bu tez çalışmasında; proje fonlayan bir kuruluştaki yenilikçi ve faydalı bir uygulama yapmak amacıyla, proje başvurusu yapmak için gerekli şartlara sahip olan 68484 araştırmacının sisteme kayıtlı yaş, cinsiyet, çalıştığı alan, eğitim bilgileri (lisans, yüksek lisans, doktora) ve çalışma bilgileri (kurum bilgileri) ile on yıl içerisinde destek programı proje türlerine başvuru yapma sonucu üzerindeki etkileri; veri madenciliğinin temel araçlarından sınıflandırma yöntemi kullanılarak analiz edilmiştir. Veri setine IBM Modeler programı aracılığıyla, Karar Ağaçları, Yapay Sinir Ağları ve Destek Vektör Makineleri sınıflandırma yöntemleri uygulanmıştır. Uygulama sonucu oluşan model çıktıları analiz edilmiş ve yorumlanmıştır.

Analiz sonuçlarına göre, üç farklı algorithmada da benzer ve yakın sonuçlar elde edilmekle birlikte, Destek Vektör Makineleri ile yapılan sınıflandırma doğruluk oranı (%70,24) en yüksek sonucu vermiştir. Bu yöntemi %69,54 doğruluk oranı ile Karar Ağaçları ve %66,60 doğruluk oranı ile Yapay Sinir Ağları yöntemi takip etmiştir.

Algoritma sonuçlarına göre birbirine yakın doğruluk oranı hesaplanmıştır. Bu açıdan bakıldığında, herhangi bir algoritmanın kesin üstünlüğünden söz etmek mümkün olmamaktadır. Hesaplama süresi, modele katılan değişken sayısı vb. performans ölçütlerine göre algoritmaların başarısı sıralamasının değişmesi mümkündür. Bununla birlikte, yöntemlerin tercih sebebi olabilecek bazı avantajları ve dezavantajları bulunmaktadır. Destek Vektör Makineleri istatistik metotları ve makine öğrenmesini birleştiren etkili bir sınıflandırma yöntemi ve hızlı model kurma özellikleri ile ön plana çıkarken, Yapay Sinir Ağları ise genelleştirme yapabilme yeteneği, yeni veriler eklendiğinde kolaylıkla uyarlanabilir ve esnek bir yöntem olmasıyla tercih edilebilir. Karar ağaçları yönteminin ise

görsel sonuçları daha sade ve kolay yorumlanabilir olan, kullanıcı dostu bir sınıflandırma yöntemi olduğu gözlemlenmiştir.

Çalışma kapsamında kullanılan analiz ve tahmin yöntemleri sayesinde fonlayıcı kuruluşun proje destek programları bazında hitap ettiği kitlenin profili ve ortak özellikleri tespit edilebilecektir. Ayrıca çalışmada, araştırmacıların proje başvuru yapma eğilimlerini etkileyen faktörler ve önem sıraları belirlenmiştir. Başvuru sayısını ve kalitesini artırmak üzere yapılan duyuru, anket, eğitim, tanıtım vb. faaliyetlerde bu yöntemleri kullanarak yapılacak benzer analizler, çalışmaların daha hızlı, verimli ve sonuç odaklı olmasını sağlayacaktır. Bu kapsamda, analiz sonuçlarının bundan sonraki süreçte programlara ilişkin sunum, tanıtım, özendirme ve bilgilendirme çalışmaları başta olmak üzere pek çok faaliyete ışık tutacağı öngörülmektedir.

Gelecekte; daha kapsamlı ve geniş veri setleri ile yapılacak çalışmalar ile daha iyi doğruluk oranına sahip modeller oluşturulabilir. Veri tabanının daha sağlıklı ve tutarlı veri toplayacak şekilde iyileştirilmesi ile de sınıflandırma performansları yükseltilebilir. Algoritmaların birleştirilmesinden türeyecek hibrit algoritmalar ile daha başarılı sınıflandırma seviyelerine ulaşılabilir. Benzer çalışmalar ile hizmet üreten kuruluşların farklı süreçlerinde kullanabileceği karar destek sistemleri geliştirilebilir.

KAYNAKLAR

- Ahn, H., Kim, K. J., Han I. (2006). Hybrid Genetic Algorithms and Case Based Reasoning Systems For Customer Classification, *Expert Systems with Applications*, 23, 127-144.
- Akpınar H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 29 (1), 1-22.
- Alban, M., Mauricio D. (2018). Decision Trees For The Early Identification of University Students At Risk Of Desertion. *International Journal of Engineering&Technology*, 7(4), 51-54.
- Albayrak, A.S., Yılmaz, Ş.K. (2009). Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 14(1), 31-52.
- Argüden, Y. ve Erşahin, B. (2008). *Veri Madenciliği: Veriden Bilgiye, Masraftan Değere*. İstanbul: ArGe Danışmanlık Yayınları, 10.
- Baesens, B., Viane, S., Poel, D.V.D., Vanthienen, J., Dedene, G. (2002). Bayesian Neural Network Learning For Repeat Purchase Modeling in Direct Marketing. *European Journal of Operational Research*, 138, 191-211.
- Boysan, M., Kayri, M. (2007). Using CHAID Analysis in Researches And An Application Pertaining To Coping Strategies, *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, 40 (2), 133-149.
- Breault, J.L., Goodall, C.R., Fos, P.J. (2003). Data Mining a Diabetic Data Warehouse. *Artificial Intelligence in Medicine*, 27, 227.
- Carvalho, D.R., Freitas A.A. (2004). A Hybrid Decision Tree/Genetic Algorithm Method for Data Mining. *Information Sciences*, 163, 13-35.
- Chang C. L., Chen C.H. (2009). Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, 36,4035-4041.
- Chang, C., Wang C., Jiang B. C. (2011). Using Data Mining Techniques for Multi Diseases Prediction Modeling of Hypertension and Hyperlipidemia by Common Risk Factors. *Expert Systems with Applications*, 38, 5507-5513.
- Chen, Y., Tang K., Shen R. and Hu Y. (2005). Market Basket Analysis in a Multiple Store Environment. *Decision Support Systems*, 40 (2), 339-354.
- Chen, T. C., Hsu, T. C. (2006). A Gas Based Approach for Mining Breast Cancer Pattern. *Expert Systems with Applications*, 30, 674-681.
- Cruz, A. (2018). Predicting the Relapse Category in Patients with Tuberculosis: A Chi-Square Automatic Interaction Detector (CHAID) Decision Tree Analysis. *Open Journal of Social Sciences*, 6, 29-36.

- Çalış, A., Kayapınar, S., Çetinyokuş, T. (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları ile Bilgisayar ve İnternet Güvenliği Üzerine Bir Uygulama. *Endüstri Mühendisliği Dergisi*, 25, 3-4, 2-19.
- Çizel, R. (2018). Gender And Emotional Intelligence as Predictors Of Tourism Faculty Students' Career Adaptability. *Advances in Hospitality and Tourism Research (AHTR)*, 6(2), 188-204.
- Diaz-Perez, F.M., Bethencourt-Cejas M. (2016). Chaid Algorithm As An Appropriate Analytical Method For Tourism Market Segmentation. *Journal of Destination Marketing & Management*, 5-3, 275–282.
- Diepen, V., Franses M. (2006). Evaluating ChiSquared Automatic Interaction Detection. *Information Systems*, 31, 8.
- Diler, S. (2016). *Veri Madenciliği Süreçleri ve Karar Ağaçları Algoritmaları ile Bir Uygulama*, Yüksek Lisans Tezi, Van Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü, Van.
- Ding, Y., Song, X., Zeng, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Systems with Applications*, 34(4), 3081–3089.
- Du, K.-L., (2010). Clustering: A neural network approach, *Neural Networks*, 23, 89-107.
- Dunham, M. H. (2003). *Data Mining Introductory and Advanced Topics*. New Jersey: Pearson Education, 8.
- Emel, G., Taşkın, Ç. (2005). Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması. *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi*, 6, (2), 221-239.
- Gülce, A. (2010). *Veri Madenciliğinde Apriori Algoritması ve Apriori Algoritmasının Farklı Veri Kümelerinde Uygulanması*. Yüksek Lisans Tezi, Trakya Üniversitesi Fen Bilimleri Enstitüsü, Edirne.
- Han, J., Kamber, M. (2006). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers, 7.
- Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall Inc.
- Horner, S., Fireman, G., Wang E. (2010). The Relation of Student Behavior, Peer Status, Race, and Gender to Decisions About School Discipline Using CHAID Decision Trees and Regression Modeling. *Journal of School Psychology*, 48, 135–161.
- Hsieh, N. C. (2004). An Integrated Data Mining and Behavioral Scoring Model for Analyzing Bank Customers. *Expert Systems with Applications*, 27, 623-633.

- Hosseini, S., Maleki, A., Gholamian, M. (2010). Cluster Analysis Using Data Mining Approach to Develop CRM Methodology to Assess the Customer Loyalty. *Expert Systems with Applications*, 37, 5259–5264.
- Huang, Z., Chen, H., Chen, W., Hsu J. (2004) Credit Rating Analysis with Support Vector Machine and Neural Networks: A Market Comparative Study. *Decision Support Systems*, 37, 543-558.
- Huang, W., Lai, K.K., Nakamori, Y., Wang, S. (2004). Forecasting Foreign Exchange Rates With Artificial Neural Network: A Review, *International Journal of Information Technology & Decision Making*, 3(1), 145-165.
- Kantardzic, M., Djulbegovic B., Hamdan H. (2002). A Data Mining Approach to Improving Polycythemia Vera Diagnosis. *Computers & Industrial Engineering*, 43(4), 765-773.
- Kamruzzman, J., Rezaul K., Beggand and Ruhul A. Sarker. (2006). *Artificial Neural Networks in Finance and Manufacturing*. United States of America: Idea Group Publishing, 7-11.
- Katılmış, S. (2018). *Veri Madenciliği Yöntemleri ile Banka Dijital Kanallarına Eğilimli Müşterilerin Belirlenmesi*, Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.
- Kim, S.Y., Jung T.S., Suh E.H., Hwang H.S. (2006). Customer Segmentation and Strategy Development Based on Customer Lifetime Value: A case Study. *Expert Systems with Applications*, 31, 101-107.
- Kim, K. J. (2006). Artificial Neural Networks with Evolutionary Instance Selection for Financial Forecasting. *Expert Systems with Applications*, 30, 519-526.
- Köktürk, F. (2012). *K-En Yakın Komşuluk, Yapay Sinir Ağları ve Karar Ağaçları Yöntemlerinin Sınıflandırma Başarılarının Karşılaştırılması*, Doktora Tezi, Zonguldak Bülent Ecevit Üniversitesi Sağlık Bilimleri Enstitüsü, Zonguldak.
- Kuzey, C. (2012). *Veri Madenciliğinde Destek Vektör Makinaları ve Karar Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama*, Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
- Lee, J. H., Park, S. C. (2005). Intelligent Profitable Customers Segmentation System Based on Business Intelligence Tools, *Expert Systems with Applications*, 29, 145-152.
- Lee, T. S., Chiu, C. C., Chou, Y. C., Lu, C. J. (2006). Mining the Customer Credit Using Classification and Regression Tree And Multivariate Adaptive Regression Splines. *Computational Statistics & Data Analysis*, 50, 1113-1130.
- Li, H., Sun J., Wu J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and to classification mining methods, *Expert Systems with Applications*, 37, 5895–5904.

- Li, F., Wang W., Xu J., Yi J., Wang Q. (2019). Comparative study on vulnerability assessment for urban buried gaspipeline network based on SVM and ANN methods, *Process Safety and Environmental Protection*, 122, 23-32.
- Liao S.H., Wen C.H. (2007). Artificial neural networks classification and clustering of methodologies and applications – literature analysis from 1995 to 2005, *Expert Systems with Applications*, 32, 1–11.
- Liao, S., Chu, P., Hsiao P. (2012). Data Mining Techniques and Applications: A decade review from 2000 to 2011. *Expert Systems with Applications*, 39, 11303-11311.
- Hearst, MA., Dumais, ST., Osman, E., Platt, J., Schölkopf, B. (1998). Support Vector Machines. *Intelligent Systems and their Applications, IEEE*, 13(4), 18-28.
- Melgani F., Bruzzone L. (2004). Classification of Hyperspectral Remote Sensing Images with Support Vector Machines, *IEEE Transactions on Geoscience and Remote Sensing*, 42(8), 1778-1782.
- Mermer, A., (2016). *Veri Madenciliği Yöntemleriyle Araştırma Destek Programları Başkanlığı Destek Programlarına Başvuru Yapan Hedef Kitlenin Analizi*, Uzmanlık Tezi, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK), Ankara.
- Mues, C., Baesens, B., Files., C.M. Vanthienen, J. (2004). Decision Diagrams in Machine Learning: An Empirical Study On Real-Life Credit-Risk Data. *Expert Systems with Applications*, 27, 257-264.
- Mukherjee, S., Shaw, R., Haldar N., Changdar S. (2015). A Survey of Data Mining Applications and Techniques. *Computer Science and Information Technologie*, 6(5), 4663-4666.
- Ngai, E.W.T. (2009). Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. *Expert Systems with Applications*, 36, 2592-2602.
- Özçınar, H. (2006). *Kpss Sonuçlarının Veri Madenciliği Yöntemleri ile Tahmin Edilmesi*, Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.
- Özkan, Y. (2008). *Veri Madenciliği Yöntemleri*. İstanbul: Papatya Yayınevi, 191.
- Öztemel, E. (2006). *Yapay Sinir Ağları*. İstanbul: Papatya Yayıncılık, 91-100.
- Raudys, S. (2000). How good are support vector machines?, *Neural Networks*, 13, 17-19.
- Rokack, L., Maimon O. (2008). Data Mining with Decision Trees Theory and Applications. *Series in Machine Perception and Artificial Intelligence*, 69, 437-438.
- Rygielski, C., Wang, J., Yen, D. (2002). Data Mining Techniques for Customer Relationship Management. *Technology in Society*, 24, 483–502.
- Savaş, S., Topaloğlu N., Yılmaz, M. (2012). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 21, 1-23.

- Sheybani, F. (2019). Predicting the Individuals' Job Satisfaction And Determining The Factors Affecting It Using the CHAID Decision Tree Data Mining Algorithm. *European Journal of Engineering Research and Science*, 4(3), 6-9.
- Silahtaroglu, G. (2016). *Veri Madenciliği Kavram ve Algoritmaları*, (3. Baskı), İstanbul: Papatya Yayıncılık ve Eğitim, 59.
- Sun, J., Li, H. (2008). Data Mining Method for Listed Companies, Financial Distress Prediction. *Knowledge-Based Systems*, 21, 1.
- Tolle, K. M., Chen H., Chow H. (2000). Estimating Drug/Plasma Con-Centration Levels By Applying Neural Networks To Pharmacokinetic Data Sets, *Decision Support Systems*, 30, 139–152.
- Vadim, K. (2018). Overview of Different Approaches to solving problems of Data Mining. *Procedia Computer Science*, 123, 234–239.
- Wang, R., Ji, W., Liub, M., Wangc, X., Weng, J., Deng, S., Gao, S., Yuang, C. (2018). Review on Mining Data From Multiple Data Sources. *Pattern Recognition Letters*, 109, 120–128.
- Wu, C.H., Kao, S.C., Su, Y.Y. (2005). Targeting Customers via Discovery Knowledge for the Insurance Industry, *Expert Systems with Applications*, 29, 291-299.
- Yalçın, L. (2019). *Sağlık Sektöründe Veri Madenciliği*, Yüksek Lisans Tezi, Milli Savunma Üniversitesi Havacılık ve Uzay Teknolojileri Enstitüsü, İstanbul.
- Yıldırım, P., Uludağ, M., Görür, A. (2008). Hastane Bilgi Sistemlerinde Veri Madenciliği. *Akademik Bilişim Dergisi*, 429-434.
- Zhang G., Patuwo B. E., Hu M. Y. (1998). Forecasting With Artificial Neural Networks: The State Of The Art. *International Journal of Forecasting*, 14(1), 35-62.
- Zhang, S., Lu, J., Zhang, C. (2004). A Fuzzy Logic Based Method to Acquire User Threshold of Minimum-Support for Mining Association Rules. *Information Sciences*, 164, 1-16.
- Zhang, B., Valentine, I., Kemp P.D. (2005). A Decision Tree Approach Modelling Functional Group Abundance in a Pasture Ecosystem. *Agriculture Ecosystems&Environment*, 110, 279-288.
- Zhong, W., Altun, G., Harrison, R., Tai, P. C., Pan, Y. (2005). Improved K-means Clustering Algorithm for Exploring Local Protein Sequence Motifs. *IEEE Transactions On Nanobioscience*, 4, 255.

EKLER

EK-1. Uygulamada kullanılan ham veriye ilişkin örnek

ID	Doğum Yılı	Cinsiyet	Diploma Türü	Eğitim Ülke	Fakülte	Bölüm	Anabilim Dalı	Mezuniyet Yılı	Unvan	Kurum Tipi	Şehir	Başvuru Durumu
1	1963	ERKEK	Doktora	Türkiye	FEN FAKÜLTESİ	Fizik	NÜKLEER FİZİK	1989	Prof. Dr.	Üniversite	Kırıkkale	BYOK
1	1963	ERKEK	Yüksek Lisans	Türkiye	FEN FAKÜLTESİ	Fizik	NÜKLEER FİZİK	1997	Prof. Dr.	Üniversite	Kırıkkale	BYOK
2	1973	KADIN	Yüksek Lisans	Türkiye	FEN FAKÜLTESİ	BİYOMÜHENDİSLİK	MOLEKÜLER BİYOLOJİ VE GENETİK	1996	Prof. Dr.	Üniversite	İstanbul	BVAR
2	1973	KADIN	Tıpta Uzmanlık	Türkiye	FEN FAKÜLTESİ	BİYOMÜHENDİSLİK	MOLEKÜLER BİYOLOJİ VE GENETİK	2001	Prof. Dr.	Üniversite	İstanbul	BVAR
3	1954	ERKEK	Yüksek Lisans	Türkiye	KİMYA FAKÜLTESİ	KİMYA	KİMYA MÜHENDİSLİĞİ	1978	Prof. Dr.	Üniversite	Sakarya	BYOK
3	1954	ERKEK	Doktora	Türkiye	MÜHENDİSLİK FAKÜLTESİ	KİMYA	ANORGANİK KİMYA	1983	Prof. Dr.	Üniversite	Sakarya	BYOK
4	1968	ERKEK	Yüksek Lisans	Türkiye	SOSYAL BİLİMLER ENSTİTÜSÜ	İLAHİYAT	TEFSİR	1992	Yrd. Doç.	Üniversite	İstanbul	BVAR
4	1968	ERKEK	Doktora	Türkiye	SOSYAL BİLİMLER ENSTİTÜSÜ	İLAHİYAT	TEFSİR	2003	Yrd. Doç.	Üniversite	İstanbul	BVAR
4	1968	ERKEK	Doktora	Türkiye	FEN-EDEBİYAT FAKÜLTESİ	KİMYA	KİMYA	1994	Prof. Dr.	Üniversite	İstanbul	BVAR
5	1967	ERKEK	Yüksek Lisans	Türkiye	SOSYAL BİLİMLER ENSTİTÜSÜ	MALİYE	MALİYE	1992	Prof. Dr.	Üniversite	Ankara	BYOK
5	1967	ERKEK	Doktora	Türkiye	SOSYAL BİLİMLER ENSTİTÜSÜ	MALİYE	MALİYE	1996	Prof. Dr.	Üniversite	Ankara	BYOK
6	1963	KADIN	Yüksek Lisans	Türkiye	MÜHENDİSLİK FAKÜLTESİ	KİMYA MÜHENDİSLİĞİ	KİMYA MÜHENDİSLİĞİ	1988	Prof. Dr.	Üniversite	Kocaeli	BVAR
6	1963	KADIN	Doktora	Türkiye	MÜHENDİSLİK FAKÜLTESİ	KİMYA MÜHENDİSLİĞİ	KİMYA MÜHENDİSLİĞİ	1996	Prof. Dr.	Üniversite	Kocaeli	BVAR
7	1964	ERKEK	Yüksek Lisans	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	MAKİNE MÜHENDİSLİĞİ	TARIM MAKİNALARI	1991	Prof. Dr.	Üniversite	Konya	BVAR
7	1964	ERKEK	Doktora	Türkiye	GÜNEŞ ENERJİSİ ENSTİTÜSÜ	MAKİNE MÜHENDİSLİĞİ	ENERJİ	1996	Prof. Dr.	Üniversite	Konya	BVAR
8	1973	ERKEK	Tıpta Uzmanlık	Türkiye	TIP FAKÜLTESİ (İNGİLİZCE)	CERRAHİ TIP BİLİMİ	TIP EĞİTİMİ	1996	Prof. Dr.	Üniversite	Ankara	BYOK
8	1973	ERKEK	Yüksek Lisans	Türkiye	TIP FAKÜLTESİ	CERRAHİ TIP BİLİMİ	KULAK BURUN BOĞAZ HASTALIKLARI	2001	Prof. Dr.	Üniversite	Ankara	BYOK
9	1961	ERKEK	Yüksek Lisans	Türkiye	DEVLET KONSERVATUARLAR	Devlet Konservatuvarlığı	KOMPOZİSYON VE ORKESTRA ŞEFİ	1993	Prof.	Üniversite	Malatya	BYOK
9	1961	ERKEK	Doktora	Türkiye	DEVLET KONSERVATUARLAR	Devlet Konservatuvarlığı	KOMPOZİSYON VE ORKESTRA ŞEFİ	1998	Prof.	Üniversite	Malatya	BYOK
10	1968	ERKEK	Doktora	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	METALURJİ VE MALZEME MÜHENDİSLİĞİ	METALURJİ MÜHENDİSLİĞİ	1995	Doç. Dr.	Üniversite	Sakarya	BYOK
10	1968	ERKEK	Doktora	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	METALURJİ VE MALZEME MÜHENDİSLİĞİ	METALURJİ MÜHENDİSLİĞİ	2000	Doç. Dr.	Üniversite	Sakarya	BYOK
11	1972	ERKEK	Doktora	Türkiye	SOSYAL BİLİMLER ENSTİTÜSÜ	İŞLETME	İŞLETME	1997	Prof. Dr.	Üniversite	Adana	BVAR
11	1972	ERKEK	Yüksek Lisans	Türkiye	SOSYAL BİLİMLER ENSTİTÜSÜ	İŞLETME	YÖNETİM VE ORGANİZASYON	2002	Prof. Dr.	Üniversite	Adana	BVAR
12	1969	KADIN	Doktora	Türkiye	MÜHENDİSLİK VE FEN BİLİMLERİ ENSTİTÜSÜ	Integration of Computer and Mechanical Engineering	FİZİK	2001	Dr.	Kamu	Kocaeli	BVAR
12	1969	KADIN	Yüksek Lisans	Türkiye	Fen Bilimleri Enstitüsü	Integration of Computer and Mechanical Engineering	FİZİK	2012	Dr.	Kamu	Kocaeli	BVAR
13	1966	ERKEK	Doktora	Türkiye	MÜHENDİSLİK FAKÜLTESİ	Malzeme Bilimi ve Metalurji	METALURJİ VE MALZEME MÜHENDİSLİĞİ	1990	Prof. Dr.	Üniversite	Eskişehir	BVAR
13	1966	ERKEK	Doktora	ABD	MÜHENDİSLİK FAKÜLTESİ	Malzeme Bilimi ve Metalurji	MALZEME BİLİMİ VE MÜHENDİSLİĞİ	1994	Prof. Dr.	Üniversite	Eskişehir	BVAR
14	1964	ERKEK	Tıpta Uzmanlık	Türkiye	TIP FAKÜLTESİ	TIP BİLİMLERİ	TIP EĞİTİMİ	1987	Prof. Dr.	Üniversite	Ankara	BVAR
14	1964	ERKEK	Tıpta Uzmanlık	Türkiye	TIP FAKÜLTESİ	TIP BİLİMLERİ	ÇOCUK SAĞLIĞI VE HASTALIKLARI	1992	Prof. Dr.	Üniversite	Ankara	BVAR
14	1964	ERKEK	Yüksek Lisans	Türkiye	TIP FAKÜLTESİ	TIP BİLİMLERİ	GENETİK	1998	Prof. Dr.	Üniversite	Ankara	BVAR
15	1978	ERKEK	Yüksek Lisans	Türkiye	Mühendislik Fakültesi	İNŞAAT MÜHENDİSLİĞİ	YAPI	2007	Dr.	Üniversite	Erzurum	BYOK
15	1978	ERKEK	Doktora	Türkiye	Mühendislik Fakültesi	İNŞAAT MÜHENDİSLİĞİ	YAPI MALZEMESİ	2013	Dr.	Üniversite	Erzurum	BYOK
16	1972	ERKEK	Doktora	Türkiye	ZİRAAT FAKÜLTESİ	Bilgisayar Programlama	TARLA BİTKİLERİ	1997	Doç. Dr.	Üniversite	Samsun	BYOK
16	1972	ERKEK	Yüksek Lisans	Türkiye	ZİRAAT FAKÜLTESİ	Bilgisayar Programlama	TARLA BİTKİLERİ	2003	Doç. Dr.	Üniversite	Samsun	BYOK
17	1966	ERKEK	Yüksek Lisans	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	Fizik	FİZİK	1996	Doç. Dr.	Üniversite	Zonguldak	BVAR
17	1966	ERKEK	Doktora	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	Fizik	FİZİK	2003	Doç. Dr.	Üniversite	Zonguldak	BVAR
18	1972	ERKEK	Doktora	Türkiye	Mühendislik Fakültesi	BİLGİSAYAR MÜHENDİSLİĞİ	ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ	1996	Yrd. Doç.	Üniversite	Elazığ	BYOK
18	1972	ERKEK	Yüksek Lisans	Türkiye	Mühendislik Fakültesi	BİLGİSAYAR MÜHENDİSLİĞİ	ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ	2003	Yrd. Doç.	Üniversite	Elazığ	BYOK
19	1973	ERKEK	Doktora	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	UTD School of Engineering	KİMYA EĞİTİMİ	1997	Doç. Dr.	Kamu	Kocaeli	BYOK
19	1973	ERKEK	Yüksek Lisans	Türkiye	FEN BİLİMLERİ ENSTİTÜSÜ	UTD School of Engineering	ORGANİK KİMYA	2002	Doç. Dr.	Kamu	Kocaeli	BYOK
20	1963	KADIN	Doktora	Türkiye	KİMYA MÜHENDİSLİĞİ	Mühendislik Fakültesi	KİMYA MÜHENDİSLİĞİ	1988	Prof. Dr.	Üniversite	Kocaeli	BVAR
20	1963	KADIN	Yüksek Lisans	Türkiye	KİMYA MÜHENDİSLİĞİ	Mühendislik Fakültesi	KİMYA MÜHENDİSLİĞİ	1996	Prof. Dr.	Üniversite	Kocaeli	BVAR
21	1956	KADIN	Yüksek Lisans	Türkiye	PEYZAJ MİMARLIĞI	ZİRAAT FAKÜLTESİ	Peyzaj Mimarlığı	1986	Prof. Dr.	Üniversite	Tekirdağ	BYOK
22	1975	ERKEK	Yüksek Lisans	Türkiye	VETERİNER FAKÜLTESİ	VETERİNER FAKÜLTESİ	Veteriner Fakültesi	1992	Prof. Dr.	Üniversite	Samsun	BYOK

EK-2. Girişimci ve yenilikçi üniversite endeksi 2017 yılı sıralaması

Sıra	Üniversite	Toplam	Bilimsel ve Teknolojik Araştırma Yetkinliği	Fikri Mülkiyet Havuzu	İşbirliği ve Etkileşim	Girişimcilik ve Yenilikçilik Kültürü	Ekonomik Katkı ve Ticarileşme
1	SABANCI Ü.	90,97	19,6	15	25	12,7	18,8
2	ORTA DOĞU TEKNİK Ü.	87,31	19,8	10,9	25	12,8	18,8
3	GEBZE TEKNİK Ü.	85,18	18,5	9,8	20,6	11,4	25
4	İSTANBUL TEKNİK Ü.	82,38	15,9	9,9	24,2	13,6	18,8
5	BOĞAZİÇİ Ü.	82,16	18	10,7	24,9	10,7	17,9
6	İHSAN DOĞRAMACI BİLKENT Ü.	81,56	19,9	9,8	24,8	10	17
7	KOÇ Ü.	78,76	16,8	9	24,9	10,2	17,9
8	İZMİR YÜKSEK TEKNOLOJİ ENS.	77,5	20	9,5	23,1	7,3	17,6
9	ÖZYEGİN Ü.	74,36	16,2	6,9	22	11,9	17,3
10	YILDIZ TEKNİK Ü.	68,74	12,6	11,8	20,2	10,2	14,1
11	TOBB EKONOMİ VE TEKNOLOJİ Ü.	67,98	16,9	13,1	19,7	6,5	11,8
12	ABDULLAH GÜL Ü.	56,94	10,7	8,1	20	6,8	11,4
13	ATILIM Ü.	55,52	12,5	5,4	15,2	5,4	17,1
14	EGE Ü.	53,95	10,9	4,7	18,2	15	5,2
15	GAZİANTEP Ü.	52,71	8,5	9,4	13	12,3	9,5
16	HACETTEPE Ü.	52,5	12,7	6,2	16,7	12	4,9
17	SELÇUK Ü.	52,35	9,6	9,6	14,6	12,9	5,6
18	ERCIYES Ü.	51,42	10,5	4	16,1	11,3	9,6
19	ANADOLU Ü.	50,94	6,7	6,2	17,8	11,7	8,6
20	ULUDAĞ Ü.	47,41	7,6	7,4	14,3	12,9	5,3
21	GAZİ Ü.	46,25	10,9	5	12	12,4	6
22	İZMİR EKONOMİ Ü.	45,15	8	8,9	15,5	7,9	4,9
23	YEDİTEPE Ü.	43,93	9,1	10,3	14,6	7,1	2,8
24	KOCAELİ Ü.	43,28	6,5	3	12,1	8	13,7
25	SAKARYA Ü.	42,98	7	7,3	10,2	10,2	8,2
26	İSTANBUL ŞEHİR Ü.	42,31	11,1	5,1	11,8	7,9	6,4
27	PAMUKKALE Ü.	42,12	7,1	4,8	11	11,9	7,2
28	ANKARA Ü.	42,12	11,1	1,8	14	10,2	5
29	YAŞAR Ü.	41,92	8,3	7,3	16	10,3	0
30	DOKUZ EYLÜL Ü.	41,35	8,2	2,1	12,2	13,1	5,7
31	ÇUKUROVA Ü.	40,75	9,4	3	14,8	8,8	4,9
32	İSTANBUL Ü.	40,6	8,4	6,1	9,8	13,1	3,2
33	MERSİN Ü.	40,21	6,8	2,5	13	11,9	6
34	ÇANKAYA Ü.	39,14	8,1	7,2	9,7	5,1	9,1
35	DÜZCE Ü.	37,72	5,6	6	10	7,8	8,3
36	AKDENİZ Ü.	36,97	9	3,3	13,2	9,2	2,3
37	KARADENİZ TEKNİK Ü.	36,46	8,1	3,4	11,8	9,7	3,4
38	İSTANBUL BİLGİ Ü.	35,94	4,8	2,4	15,5	6,7	6,5
39	DUMLUPINAR Ü.	35,34	7,5	3,7	8,9	3,6	11,6
40	BURSA TEKNİK Ü.	35,18	11	7,9	15,5	0,8	0
41	HASAN KALYONCU Ü.	35,1	2,9	8,7	1,6	12,3	9,6
42	MEF Ü.	34,92	3,1	9,3	9,8	0,2	12,5
43	BAHÇEŞEHİR Ü.	34,6	7,3	3,5	14,3	5,4	4,1
44	IŞIK Ü.	34,05	3,7	4,2	7,7	9,7	8,8
45	SÜLEYMAN DEMİREL Ü.	33,59	9,5	4,8	8,9	4,3	6,1
46	GAZİOSMANPAŞA Ü.	33,17	7,5	2,3	11,4	7,4	4,5
47	TÜRK HAVA KURUMU Ü.	32,54	5	0,5	11,8	3,6	11,6
48	NAMIK KEMAL Ü.	32,27	6,5	5,1	7,3	2,6	10,9
49	FIRAT Ü.	31,44	6,4	3	7,1	10,5	4,5
50	MARMARA Ü.	30,44	7	2	12,1	6,7	2,7

EK-3. Yapay sinir ağı modeli tahmin sonuçlarına ilişkin örnek

ID	BAŞVURU DURUMU	VERİ TÜRÜ	MODELİN TAHMİNİ	PARAMETRE
1	BYOK	Eğitim	BYOK	0.952
2	BYOK	Eğitim	BYOK	0.807
3	BYOK	Eğitim	BYOK	0.954
4	BYOK	Test	BYOK	0.945
5	BYOK	Eğitim	BYOK	0.942
6	BYOK	Eğitim	BYOK	0.859
7	BVAR	Eğitim	BYOK	0.809
8	BYOK	Eğitim	BYOK	0.840
9	BYOK	Eğitim	BYOK	0.778
10	BYOK	Eğitim	BYOK	0.933
11	BYOK	Eğitim	BYOK	0.957
12	BYOK	Eğitim	BYOK	0.905
13	BYOK	Test	BYOK	0.777
14	BYOK	Eğitim	BYOK	0.919
15	BVAR	Test	BYOK	0.748
16	BYOK	Test	BYOK	0.829
17	BYOK	Eğitim	BYOK	0.932
18	BYOK	Eğitim	BYOK	0.866
19	BVAR	Eğitim	BYOK	0.877
20	BYOK	Eğitim	BYOK	0.935
21	BYOK	Eğitim	BYOK	0.833
22	BYOK	Test	BYOK	0.833
23	BVAR	Eğitim	BYOK	0.859
24	BYOK	Eğitim	BYOK	0.802
25	BYOK	Eğitim	BYOK	0.804
26	BVAR	Eğitim	BYOK	0.809
27	BYOK	Test	BYOK	0.883
28	BVAR	Test	BYOK	0.801
29	BYOK	Eğitim	BYOK	0.858
30	BYOK	Eğitim	BYOK	0.802
31	BVAR	Eğitim	BYOK	0.776
32	BYOK	Eğitim	BYOK	0.801
33	BYOK	Eğitim	BYOK	0.824
34	BYOK	Eğitim	BYOK	0.929
35	BYOK	Test	BYOK	0.885
36	BYOK	Test	BYOK	0.957
37	BYOK	Eğitim	BYOK	0.805
38	BVAR	Eğitim	BYOK	0.849
39	BYOK	Test	BYOK	0.851
40	BYOK	Eğitim	BYOK	0.784
41	BVAR	Eğitim	BYOK	0.803
42	BYOK	Eğitim	BYOK	0.808
43	BVAR	Test	BYOK	0.801

EK-4. Destek vektör makineleri ile oluşturulan modelin parametre değerleri

Intercept	0.297
Yaş	0.002
[Cinsiyet=ERKEK]	-0.046
[Cinsiyet=KADIN]	-0.024
[Lisans Eğitimi Alınan Kurum=1.Grup]	-0.006
[Lisans Eğitimi Alınan Kurum=2.Grup]	0.006
[Lisans Eğitimi Alınan Kurum=3.Grup]	0.056
[Lisans Eğitimi Alınan Kurum=4.Grup]	0.047
[Lisans Eğitimi Alınan Kurum=5.Grup]	0.023
[Lisans Eğitimi Alınan Kurum=Diğer]	0.170
[Yüksek Lisans Eğitimi Alınan Kurum=1.Grup]	0.064
[Yüksek Lisans Eğitimi Alınan Kurum=2.Grup]	0.062
[Yüksek Lisans Eğitimi Alınan Kurum=3.Grup]	0.026
[Yüksek Lisans Eğitimi Alınan Kurum=4.Grup]	0.011
[Yüksek Lisans Eğitimi Alınan Kurum=5.Grup]	0.024
[Yüksek Lisans Eğitimi Alınan Kurum=Diğer]	0.109
[Doktora Eğitimi Alınan Kurum =1.Grup]	0.030
[Doktora Eğitimi Alınan Kurum =2.Grup]	0.035
[Doktora Eğitimi Alınan Kurum =3.Grup]	0.056
[Doktora Eğitimi Alınan Kurum =4.Grup]	0.043
[Doktora Eğitimi Alınan Kurum =5.Grup]	0.070
[Doktora Eğitimi Alınan Kurum =Diğer]	0.062
[Doktora_Eğitim Ülke=ABD]	-0.160
[Doktora_Eğitim Ülke=Almanya]	-0.036
[Doktora_Eğitim Ülke=Diğer]	0.446
[Doktora_Eğitim Ülke=Fransa]	0.041
[Doktora_Eğitim Ülke=İngiltere]	-0.083
[Doktora_Eğitim Ülke=Kanada]	-0.081
[Doktora_Eğitim Ülke=Türkiye]	0.170
[Çalıştığı Kurum =1.Grup]	-0.137
[Çalıştığı Kurum =2.Grup]	-0.029

EK-4. (devam) Destek vektör makineleri ile oluşturulan modelin parametre değerleri

[Çalıştığı Kurum =3.Grup]	-0.004
[Çalıştığı Kurum =4.Grup]	0.003
[Çalıştığı Kurum =5.Grup]	0.046
[Çalıştığı Kurum =Diğer]	0.039
[Çalıştığı Kurum =Kamu]	0.100
[Çalıştığı Kurum =Özel]	0.279
[Çalıştığı Şehir Bölge=Akdeniz B.]	-0.093
[Çalıştığı Şehir Bölge=D. Anadolu B.]	-0.047
[Çalıştığı Şehir Bölge=Diğer]	0.444
[Çalıştığı Şehir Bölge=Ege B.]	-0.035
[Çalıştığı Şehir Bölge=G. Doğu Anadolu B.]	0.062
[Çalıştığı Şehir Bölge=Karadeniz B.]	-0.021
[Çalıştığı Şehir Bölge=Marmara B.]	-0.016
[Çalıştığı Şehir Bölge=İç Anadolu B.]	0.003
[Kurum Tipi=Kamu]	0.100
[Kurum Tipi=Özel]	0.279
[Kurum Tipi=Üniversite]	-0.082
[Faaliyet Alanı=Girilmemiş]	0.000
[Faaliyet Alanı=Sağlık Bilimleri]	-0.350
[Faaliyet Alanı=Sosyal Ve Beşeri Bilimler]	-0.215
[Faaliyet Alanı=Tarımsal Bilimler]	-0.757
[Faaliyet Alanı=Teknik Bilimler]	-0.594
[Faaliyet Alanı=Temel Bilimler]	-0.588

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : MERMER, Ayten
 Uyruğu : T.C.
 Doğum tarihi ve yeri : 30.01.1990,
 Ankara Medeni hali : Bekar
 Telefon : 0 (312) 298 11 51
 e-mail : ayten.mermer@tubitak.gov.tr



Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi / Endüstri Mühendisliği	Devam Ediyor
Lisans	Gazi Üniversitesi / Endüstri Mühendisliği	2012
Lise	Ankara Atatürk Anadolu Lisesi	2008

İş Deneyimi

Yıl	Yer	Görev
2012-Halen	TÜBİTAK	Uzman

Yabancı Dil

İngilizce

Yayınlar

Mermer, A. ve Arıkan, M. (2019, 19-21 Nisan). *Fonlayıcı Bir Kuruluşta Veri Madenciliği Yöntemleri Üzerine Bir Uygulama*. Paper presented at 3rd International Symposium on Innovative Approaches in Scientific Studies, Ankara.

Hobiler

Kitap Okuma



GAZİ GELECEKTİR..