T.C.

GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ

MÜHENDİSLİK VE FEN BİLİMLERİ

ENSTİTÜSÜ

ELEKTRONİK POSTALARIN ADLİ ANALİZİNDE YAZAR ANALİZİ TEKNİKLERİNİN KULLANILMASI

EKİN EKİNCİ

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

GEBZE

2013

T.C.

GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ

MÜHENDİSLİK VE FEN BİLİMLERİ

ENSTİTÜSÜ

ELEKTRONİK POSTALARIN ADLİ ANALİZİNDE YAZAR ANALİZİ TEKNİKLERİNİN KULLANILMASI

EKİN EKİNCİ

YÜKSEK LİSANS TEZİ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

TEZ DANIŞMANI

DOÇ. DR. YUSUF SİNAN AKGÜL

GEBZE

2013

YÜKSEK LİSANS TEZİ JÜRİ ONAY SAYFASI

G.Y.T.E. Mühendislik ve Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 07.12.2012 tarih ve 2012/44 sayılı kararıyla oluşturulan jüri tarafından 20.12.2012 tarihinde tez savunma sınavı yapılan Ekin Ekinci'nin tez çalışması Bilgisayar Mühendisliği Anabilim Dalında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE (TEZ DANIŞMANI)	:	Doç. Dr. Yusuf Sinan AKGÜL
ÜYE	:	Prof. Dr. Yaşar BECERİKLİ
ÜYE	:	Yrd. Doç. Dr. Hidayet TAKÇI

ONAY

G.Y.T.E. Mühendislik ve Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./…./…. tarih ve ……/…… sayılı kararı.

İMZA/MÜHÜR

ÖZET

TEZİN BAŞLIĞI: ELEKTRONİK POSTLARIN ADLİ ANALİZİNDE YAZAR ANALİZİ TEKNİKLERİNİN KULLANILMASI

YAZAR ADI: EKİN EKİNCİ

Bilişim teknolojisinde yaşanan hızlı gelişmeler sonucunda elektronik postalar günümüzün en önemli iletişim araçlarından biri haline gelmiştir. Pek çok alanda kullanıcısına kolaylık sağlayan elektronik postalar suçlular için de ilgi çekici bir ortam olmaktadır. Gerçek sahibi belli olmayan kötü niyetli elektronik postalar bilişim suçları içerisinde yerini alırken, bu tür elektronik postaların gerçek sahibini belirlemek için yazar belirlemenin gerekliliği ortaya çıkmıştır.

Çalışmada, elektronik postaların yazarlarını adli bilimler açısından belirlemek amacıyla; mesaj gövdesine dayalı olarak elektronik postaların güvenliğini arttırabilen ayrıca elektronik postaların görünen sahibi yerine asıl sahibini bulabilen 5 yazarın 250 mesajının kullanıldığı uygulama gerçekleştirilmiştir ve elektronik postalar ile aynı karakteristiği gösteren haber grubu mesajları veri seti olarak kullanılmıştır. Veri setinden çıkartılan 49 tane metinsel özellik; sınıflandırma yöntemleri algoritmalarından J48 karar ağacı, Naive Bayes, çok katmanlı yapay sinir ağları, SMO (Sıralı Minimal Optimizasyon), Bagging ve AdaBoostM1 ile işlenmiştir. Sonuçlar F-ölçümüne göre değerlendirilmiştir. Tekli sınıflandırıcılardan çok katmalı yapay sinir ağları %82.4 en başarılı yöntem olarak gözlemlenirken gruplamalı sınıflandırıcılardan Bagging (SMO) %86.1 ile AdaBoostM1 (J48) de %84.4 ile en iyi genellemeyi sağlamıştır. Tüm sonuçlar değerlendirildiğinde eldeki veri seti için Bagging'in AdaBoostM1'e göre daha iyi bir genelleme yaptığı gözlemlenmiştir. Yöntemlerin başarılarındaki farkın temel nedeninin seçilen veri setine, veri setine uygulanan önişleme adımlarına, çıkarılan metinsel özelliklere ve algoritma parametrelerine bağlı olduğu anlaşılmıştır. Çalışmadan çıkarılan sonuç ise, yazar belirlemenin gerçek suçluyu belirlemede oldukça başarılı bir yöntem olduğu ve adli bilimler üzerine çalışan bilim adamlarına büyük fayda sağlayacağıdır.

SUMMARY

THESIS TITLE: USING AUTHORSHIP ANALYSIS TECHNIQUES IN FORENSIC ANALYSIS OF ELECTRONIC MAILS

AUTHOR: EKİN EKİNCİ

As a result of rapid advances in information technology, electronic mail has become one of today's most important communication tools. Electronic mail which provides conveniences to its users in many cases, is also an attractive environment for criminals. Malicious electronic mail whose actual owner is unknown is taking place in computer crimes and authorship attribution has become necessary for determining the actual owner of these electronic mails.

In this study, an application which can improve security of electronic mails based on message body and can find actual owner of electronic mails instead of known owner for 250 messages from 5 writers was achieved with the aim of identifying authors of electronic mails in terms of forensic science and newsgroup messages which have same characteristic with electronic mails were used as dataset. 49 textual measures which were extracted from dataset were processed with J48 decision tree, NaiveBayes, MLP, SMO (Sequential Minimal Optimization), Bagging and AdaBoostM1 classifying methods. The results were evaluated according to the F-measure. While single classifier MLP has been observed to be the most successful with 82.6%, in ensemble classifiers, Bagging (SMO) and AdaBoostM1 (J48) have provided the best generalization with 86.1% and 84.4% respectively. Considering all results, Bagging provided better generalization than AdaBoostM1. The main reason of difference between the success rates of methods depends on selected dataset, preprocessing step which is applied to dataset, extracted textual features and parameters of algorithms. Authorship identfying is very successful method for identifying real criminal and useful for scientists working on forensic science .

TEŞEKKÜR

Çalışmam konusunda benden desteğini esirgemeyen, yorum ve kaynak edinme yönünden yardımlarını gördüğüm ve iki yıl boyunca danışmanlığımı üstlenen ve çalışmamda büyük emekleri olan Cumhuriyet Üniversitesi öğretim üyelerinden saygıdeğer hocam Yrd. Doç. Dr. Hidayet TAKÇI'ya, bu konuda bana çalışma fırsatı sunan ve yeni yöntemler ile çalışmamı sağlayan saygıdeğer danışman hocam Doç. Dr. Yusuf Sinan AKGÜL'e, tez yazım aşamasında önerilerinden ve eleştirilerinden yararlandığım Kocaeli Üniversitesi değerli öğretim üyelerinden Yrd. Doç. Dr. Önder EKİNCİ'ye teşekkürlerimi sunarım.

Beni bugüne kadar sürekli destekleyen, her zaman yanımda olan ve bana güvenen başta merhum babam Murat EKİNCİ olmak üzere annem Nergiz EKİNCİ'ye ve kardeşim Başak EKİNCİ'ye de teşekkürlerimi sunarım.

İÇİNDEKİLER DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

SMO:	Sıralı Minimal Optimizasyon (Sequential Minimal Optimization)
MLP:	Multilayer Perceptron
Bagging:	Bootstrap Aggregating
MS:	Milattan Sonra
IP:	İnternet Protokolü
DVM:	Destek Vektör Makineleri
YSA:	Yapay Sinir Ağları
ID3:	Iterative Dichotomiser
DNA:	Deoksiribonükleik Asit
ARPANET:	Advanced Research Projects Agency Network
POS:	Part-of-Speech
PCA:	Principal Component Analysis
DP:	Doğru Pozitif
YN:	Yanlış Negatif
YP:	Yanlış Pozitif
DN:	Doğru Negatif

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Üzerine tarihçilerin, hukukçuların, felsefecilerin ve bilim adamlarının yıllar boyunca araştırma yaptıkları, yasalarla sınırlarını ve yaptırımlarını belirledikleri suç, insanlık tarihinin başlangıcından beri bilinen bir kavramdır. Her ne kadar yasalar caydırıcı etki yapsa da insanoğlunun doğası gereği suç, toplum hayatında her zaman var olmuştur ve gelişen teknolojinin de etkisiyle ortaya çıkan yeni türleri ile birlikte varlığını sürdürmeye devam etmektedir.

Sosyal bir sorun olan ve toplumu meydana getiren bireyleri ekonomik, psikolojik ve yaşam standartları açısından olumsuz yönde etkileyen suç, Stanciu tarafından "Sosyal toplumun çoğunluğu tarafından tehlikeli sayılan ihmal ya da icra niteliğindeki hareketlerdir." şeklinde tanımlanmıştır [Dönmezer, 1984]. Jhering tarafından ise "Toplum halinde yaşama şartlarına yönelmiş her türlü saldırılardır." şeklinde ifade edilmiştir [Dönmezer, 1984].

Genel olarak trafik, cinsel, hırsızlık, dolandırıcılık, kaçakçılık, narkotik, şiddet, yangın ve bilişim şeklinde sınıflandırılan suçun aydınlatılmasında adli analiz olarak adlandırılan bir süreç izlenmektedir [Chen et al, 2004]. Suç ile ilgili delillerin toplanması, muhafaza edilmesi ve incelenmek üzere ilgili birimlere gönderilmesi şeklinde adımları olan adli analiz süreci adli bilimlerin çalışma alanına girmektedir.

Adli tıp, balistik, kriminoloji ve adli bilişim başta olmak üzere 18 alt dalı bulunan adli bilimlerin geçmişi MS 700'lü yıllara dayanmaktadır. MS 700'lü yıllarda Çinlilerin dokümanların ve kilden yapılan heykellerin gerçek sahiplerini belirlemek için parmak izini kullanması ile ilk uygulaması ortaya çıkan adli bilimler [Inman and Rudin, 2000], "Tıp, fen ve sosyal bilimler alanlarındaki bilgilerin adaletin hizmetine sunulması ile ilgilenen bir dal." olarak tanımlanmaktadır [Hancı, 2005].

1980'ler bilişim teknolojisinde yaşanan hızlı gelişmelere tanıklık etmiş ve günlük hayatın bir parçası haline gelen bilgisayar ve uygulamaları özellikle internet kullanımı ile kişi ve kurumlar için vazgeçilmez bir hal almıştır. Bilişim teknolojisinin yaşadığı bu hızlı gelişim süreci suça eğilimli bazı kişilerin ilgisini çekmiş ve gerçek

dünya suçlarının yanında bilişim suçu olarak adlandırılan ve son yıllarda adı sıkça duyulan sanal dünya suçları ortaya çıkmıştır.

Bilişim suçu, "Suçu işleyenlerin kazanç sağladığı, kurbanların ise zarara uğradığı bilgisayar ve teknolojinin kötüye kullanılması ile oluşan maksatlı bir davranış." olarak tanımlanmaktadır [Parker, 1989]. Bilişim suçlarının çözümlenmesi amacıyla ortaya çıkan adli bilişim ise temel olarak kanıtın elde edilmesi, tanımlanması ve analiz edilmesi adımlarından oluşmaktadır [Hu, 2009].

Bilişim Suçunun işlenmesinde önemli bir araç olan internet, ilk olarak 1969 yılında ortaya çıkmıştır. Dünya çapında yayın kapasitesine sahip olan internet, mesafeleri göz ardı ederek insanların bilgisayar aracılığı ile etkileşimini ve bilgi paylaşımında bulunmalarını sağlamaktadır [Leiner et al, 1997]. İnternetin süregelen gelişimi ve sahip olduğu güçlü özellikleri sayesinde insanların bilgisayarlar aracılığı ile yazılı haberleşmelerini sağlayan elektronik postalar önemli bir iletişim aracı haline gelmiştir [Gribbin, 2011]. Ekonomik ve pratik olmaları açısından eğitimden endüstriye, günlük yaşamdan kamuya kadar pek çok alanda kullanılan elektronik postalar saldırganların da en çok ilgi duyduğu ortamlardan biri olmuştur.

Elektronik postalar günümüzde önemli bilginin yetkisiz iletimi, istenmeyen posta gönderimi, tehdit mesajı gönderme, yasaklı propaganda yapma, dolandırıcılık, virüs gönderme, korsan yazılım ve eser dağıtımı, cinsel içerikli mesaj gönderimi ve bir başkasının hesabına girip onun adına hoş olmayan mesajlar atarak veya mesajın gönderim sırasında içeriğini değiştirerek yasal kullanıcının zor durumda bırakılması gibi kötü amaçlarla kullanılmaktadır [Ma et al, 2008]. Bu tür durumları önlemede virüs programı, güvenlik duvarı ve şifre koruma kullanımı yetersiz kaldığı gibi sahte kimlik kullanan gerçek suçluları belirlemede geriye dönük IP taraması da her zaman başarılı olamamaktadır. Bahsedilen mevcut yöntemlerin suçu önlemede ve suçluyu belirlemede yetersiz oluşları bilişim suçları ile mücadelede yeni bir arayışa neden olmuş ve yazar analizinin adli bilişim içerisindeki gerekliliği ortaya çıkmıştır.

Adli bilişim içerisinde bir yöntem olarak kullanılan yazar analizi aynı zamanda metin madenciliği görevleri arasında yer almaktadır. Eldeki dokümanlardan belli bir amaç çerçevesinde önceden bilinmeyen ancak potansiyel olarak faydalı bilginin

çıkarılması şeklinde tanımlanan [Visa, 2001] metin madenciliği; doğal dil işleme ve veri madenciliği disiplinlerinin sentezinden oluşmaktadır. İlk ciddi uygulaması 1950'lerde gerçekleştirilen doğal dil işleme, doğal dildeki metinlerin yapısal ve anlamsal çözümlemesini gerçekleştirmektedir [Nadkarni et al, 2011]. 1980'lerin sonunda ortaya çıkan ve sınıflandırma, kümeleme ve birliktelik kuralları keşfi görevlerinden oluşan veri madenciliği ise eldeki veriden bilginin keşfi olarak tanımlanmaktadır [Han et al, 2012]. Doğal dil işleme ve veri madenciliği yazar analizi içerisindeki temel adımları oluşturmaktadır.

Yazar analizi, yazara ait metinlerden çıkarılan ve yazarı belirlemeye yarayan metinsel özelliklerin (kelime sayısı, cümle sayısı, fonksiyonel kelimeler ve yazım hataları gibi) kullanımı ile eldeki metnin yazarına ait niteleyici bir sonuca varmayı amaçlamaktadır [Abbasi, 2005]. Yazar analizi; yazar belirleme, yazarı karakterize etme ve benzerlik tespiti olmak üzere üçe ayrılmaktadır [Zheng et al, 2006]. Yazar belirleme, ele alınan metnin mevcut yazarlara ait olup olmadığının tespitini yaparken; yazarı karakterize etme yazarla ilgili cinsiyet, yaş, eğitim durumu gibi bilgileri eldeki metinlerden çıkarmayı amaçlamakta; benzerlik tespiti ise verilen iki metin arasındaki benzerliği bulmaktadır [Stamatatos, 2009]. Bu üç alt dal arasında en çok uygulama alanına sahip olan yazar belirleme, çıkarılan metinsel özellikleri veri madenciliğindeki çeşitli sınıflandırma yöntemleri ile işleyip bir sonuca varmayı amaçlamaktadır.

Yazar belirleme üzerine bilinen ilk çalışma, 1887 yılında Mendenhall tarafından yapılmış olup çalışmada metinsel özellik olarak kelimeyi meydana getiren karakter sayıları kullanılmıştır [Mendenhall, 1887]. Bu çalışmayı 1932 yılında Zipf'in, 1939 ve 1944 yıllarında da Yule'un yaptığı istatistiksel çalışmalar izlemiştir [Stamatatos, 2009]. Yazar belirleme konusunda milat kabul edilen çalışma ise Mosteller ve Wallace tarafından 1963 yılında yapılan çalışma olmuştur. Bu çalışmada metinsel özellik olarak fonksiyonel kelimeler, çıkarılan bu fonksiyonel kelimeleri işlemek için ise sınıflandırma yöntemlerinden Bayes Teoremi kullanılmıştır [Mosteller and Wallace, 1963]. Fonksiyonel kelimelerin yazar belirleme çalışmalarında popüler bir metinsel özellik olarak kullanılması bu çalışmadan sonra başlamıştır.

Elektronik ortamdaki metinlerin (e-posta, haber grubu mesajları, forum ve blog mesajları gibi) yaygınlaşması ile ise, bu ortamdaki metinler için de yazar belirleme çalışmalarının gerekliliği doğmuştur. 2000 yılında Oliver De Vel seçtiği 5 yazara ait elektronik postalar için sınıflandırma yöntemlerinden DVM'yi kullanarak 38 tane metinsel özelliği işlemiştir [De Vel, 2000]. Arapça ve İngilizce forum mesajları ile yapılan çalışmada DVM ve Karar Ağaçları Yöntemi kullanılarak %97'ye varan başarı sağlanmıştır [Abbasi and Chen, 2005]. 2012 yılında yapılan çalışmada ise elektronik postaların yapısından ve yazar belirleme için kullanılan yöntemlerden genel olarak bahsedilmiştir [Bogawar and Bhoyar, 2012].

Bu çalışmanın amacı ise mesaj gövdesine dayalı olarak elektronik postaların güvenliğini arttırabilen ayrıca gönderilen elektronik postaların görünen sahibi yerine asıl sahibini bulabilen örnek bir uygulama gerçekleştirmektir. Çalışmanın temelini oluşturan yazar belirleme, kim tarafından yazıldığı belli olan dokümanlara dayalı olarak yeni gelen ve yazarı belli olmayan dokümanların yazarını belirlemeyi amaçlamaktadır.

Çalışma temel olarak; veri kümesinin elde edilmesi, veri kümesi üzerinde ön işleme yapıp metinsel özelliklerin çıkartılması ve çıkartılan metinsel özelliklerin veri madenciliğindeki çeşitli sınıflandırma yöntemleri ile işlenmesi adımlarından oluşmaktadır. Elektronik postalar mahremiyet ve güvenlik açısından elde edilmesi oldukça güç olan bir veri kümesidir ve aynı zamanda sadece gönderildiği kişiye özeldir. Bu nedenle çalışmada elektronik postalar yerine elektronik postalar ile aynı karakteristiği gösteren ve herkes tarafından görülebilen haber grubu mesajlarının kullanılması uygun görülmüş ve www.newskolik.net isimli haber grubu sitesinden 5 yazar için 250 tane mesaj (yazar başına 50 mesaj) seçilmiştir. Seçilen mesajlar ön işlemeye tabi tutulmuş ve ön işlemeye tabi tutulan bu mesajlardan 49 tane metinsel özellik çıkartılmıştır. Çıkartılan bu metinsel özellikler; eğitim seti üzerinden tek bir model oluşturup bu modele göre değerlendirme yapan Karar Ağaçları, Bayes Sınıflandırıcılar, YSA ve DVM ve içerisine aldığı temel sınıflandırıcılar ile beraber (YSA, DVM gibi) eğitim seti üzerinden birden fazla model oluşturup bu modellerin birleşimine göre değerlendirme yapan Gruplamalı Sınıflandırıcılar ile işlenmiştir.

Metinsel özelliklerin işlenmesi adımında; ID3 ve C4.5 karar ağacı algoritmalarından geliştirilen J48 karar ağacı algoritması, büyük veri setleri üzerinden hızlı bir şekilde karar ağacı oluşturabilmesi nedeniyle bu çalışmada tercih edilmiştir. Metin sınıflandırmada oldukça hızlı olan Naive Bayes algoritması, büyük veri setleri üzerinde yüksek performans ile çalışması ile bilinmektedir ve metinsel özellikleri işlemek amacıyla seçilmiştir. Doğrusal olmayan problemlerin çözümünde en sık kullanılan ve bir YSA algoritması olan Çok Katmanlı YSA algoritması ve standart DVM algoritmalarına göre basit, daha hızlı ve gerçeklenmesi kolay olan SMO algoritması [Platt, 1999] da yine metinsel özellikleri işlemek üzere bu çalışmada kullanılmıştır. Çalışmada Gruplamalı Sınıflandırıcılardan Bagging (Bootstrap Aggregating) ve AdaBoostM1 algoritmalarından faydalanılmıştır. Kolay gerçeklenmesi ve güçlü performansı ile ön plana çıkan ve yeniden örneklendirme mantığı ile çalışan Bagging, güçlü genelleme özelliği nedeni ile tercih edilmiştir [Wang et al, 2011]. İki sınıfa sahip veri setlerini sınıflandırmada kullanılan ve yeniden ağırlıklandırma prensibi ile çalışan AdaBoost algoritmasının çok sınıflı veri setlerine uygulanması için geliştirilen AdaBoostM1 algoritması da zayıf sınıflandırıcılardan güçlü bir sınıflandırıcı oluşturması ve iyi bir genelleme özelliğine sahip olması nedeniyle bu çalışmada kullanılmıştır [Doğan and Akay, 2010].

Algoritmaların eldeki veri kümesine 10 kat çapraz doğrulama kullanılarak uygulanması ile her yöntem için bir model oluşturulmuştur. İstatistiksel bir yöntem olan 10 kat çapraz doğrulama sonucunda oluşan modellerin performanslarını belirlemede veri madenciliğinde sıkça kullanılan ölçümlerden F-ölçümü göz önünde bulundurularak bir sonuca varılmıştır. Uygulama sonucunda; eldeki veri kümesine, veri kümesine uygulanan önişleme adımlarına, çıkarılan metinsel özelliklere ve algoritma parametrelerine bağlı olarak tekli sınıflandırıcılar içerisinde en başarılı yöntemin %82.4 F-ölçümü ile YSA olduğu gözlemlenmiştir. Gruplamalı Sınıflandırıcılarda Bagging SMO %86.1 F-ölçümü ile en iyi genellemeyi sağlarken, AdaBoostM1 J48 %84.8 ile en iyi genellemeyi sağlamıştır. Tüm sonuçlar göz önünde bulundurulduğunda eldeki veri kümesi için Bagging'in AdaBoostM1'e göre daha iyi bir genelleme özelliğine sahip olduğu gözlemlenmiştir.

2. ADLİ BİLİMLER

“Tıp, fen ve sosyal bilimler alanlarındaki bilgilerin adaletin hizmetine sunulması ile ilgilenen bir dal.” olarak tanımlanan adli bilimler günümüzdeki önemli bilim dallarından biridir [Hancı, 2005]. MS 700’lü yıllarda Çinlilerin parmak izini dokümanların ve kilden yapılan heykellerin gerçek sahiplerini belirlemek için kullanması adli bilimlere ait bilinen ilk uygulama olmakla birlikte milattan önce yaşayan insanların parmak izlerinin yaptıkları resimler ve taş oymaları üzerinde bulunması adli bilimlerin başlangıcı olarak kabul edilmektedir [Inman and Rudin, 2000].

Adalet ile ilgili sorulara bilimsel olarak cevap veren ve çok geniş bir çalışma alanına sahip olan adli bilimler tıp, böcek bilimi, ses bilimi gibi çeşitli bilgileri adli vakaların çözümü için kullanan başlıca 18 alt dala sahiptir [Fienberg, 2007]. Bu alt dallar adli tıp, kriminoloji, adli toksikoloji, adli entomoloji, adli patoloji, adli psikiyatri, adli seroloji, adli antropoloji, adli otomotiv, adli bilişim, balistik, adli palinoloji, adli diş hekimliği, adli foniatri, adli sanat, adli hemşirelik, adli meteoroloji ve adli veterinerlik şeklinde sınıflandırılmaktadır.

Adli bilimler tümüyle suç ve suçun aydınlatılması ile ilgilenen bir bilim dalıdır. Üzerine tarihçilerin, hukukçuların, felsefecilerin ve bilim adamlarının yıllar boyunca araştırma yaptıkları suç, Stanciu tarafından "Sosyal toplumun çoğunluğu tarafından tehlikeli sayılan ihmal ya da icra niteliğindeki hareketlerdir." şeklinde tanımlanmıştır [Dönmezer, 1984]. Günümüzde ise suç, "Yasalara aykırı davranış." olarak tanımlanmaktadır [TDK, 2005].

Trafik, cinsel, hırsızlık, dolandırıcılık, kaçakçılık, narkotik, şiddet, yangın ve bilişim şeklinde sınıflandırılan suçun aydınlatılmasında bilimsel bir sürecin takibi gerekli olmaktadır (Çizelge 1.1.) [Chen et al, 2004]. Adli analiz olarak adlandırılan ve adli bilimlerin çalışma alanına giren bu süreç delilin belirlenmesi, toplanıp muhafaza edilmesi ve incelenmek üzere ilgili birimlere gönderilmesi şeklinde işlemektedir

Çizelge 1.1.: Suç Türleri [Demirbaş, 2001].

Suç Türü	Açıklaması
Trafik	Trafik kurallarına uymama, aşırı hız ve uykulu, alkollü ve ilaç alarak araç kullanımı ile can ve mal güvenliğini tehlikeye atma veya öldürme, araç ile bir kişiye vurup kaçma
Cinsel	Teşhircilik, çocuk, genç ve kadınların cinsel suistimali
Hırsızlık	Bir başkasının mal varlığının, devlete ait gizli bilgilerin çalınması
Dolandırıcılık	Sahte çek verme, kalpazanlık, evrakta sahtecilik, tefecilik, yolsuzluk, uluslararası para dolandırıcılığı
Kaçakçılık	Doku ve organ kaçakçılığı, insan kaçakçılığı, gümrük kaçakçılığı, silah ve tarihi eserlerin uluslararası kaçakçılığı
Narkotik	Uyuşturucu madde kullanma ve satma, uluslararası uyuşturucu madde kaçakçılığı
Şiddet	Cinayet, aile içi şiddet, kurumlarda şiddet, holiganlık, terörizm
Yangın	Kasıtlı yangın çıkarma
Bilişim	Sanal dolandırıcılık, istenmeyen posta, virüs, tehdit mesajı gönderimi, sanat eserlerinin internet üzerinden yasadışı dağıtımı

Deliller, işlenen suç ile ilgili olay yerinde bırakılan ve çağımız bilimsel yöntemlerinin kullanımı ile suçu aydınlatmada önleyici hizmet olarak görev yapan polise en büyük yardımcı olarak karşımıza çıkmaktadır. Olayın başlangıcından bitişine kadar geçtiği tüm ortamları ifade eden olay yeri, değişik biçimlerde ortaya çıkar ve delillerde bu ortamlara göre şekillenir. Buradaki temel mantık işlenen suç ile ortam arasında kesin bir etkileşimin varoluşudur. Bu etkileşim sonucunda ortamda oluşan deliller adli bilimler içersindeki alt dalların yardımı ile suçluyu belirlemede kullanılmaktadır.

Suçu aydınlatma sürecinin birinci aşamasını oluşturan delillerin belirlenmesi polisi sonuca yaklaştıran en önemli adımdır. Deliller; suç ile ilişkili olması, mahkemeye sunulabilir olması, inandırıcı olması ve gelecekte meydana gelebilecek olaylara ışık tutması özelliklerini taşımalıdır. Bu özelliklere sahip deliller belirlendikten sonra ikinci aşamaya geçilir. İkinci aşamada; elde edilen deliller kaybolmadan, bozulmadan ve uygun koşullar altında toplanıp muhafaza edilir. Üçüncü ve son aşama ise sonuç aşamasıdır. Toplanıp muhafaza edilen deliller laboratuvar ortamında incelenir ve olay aydınlatılır. Üç aşamadan oluşan adli analiz sürecinin tamamında bilimsel bir yol izlenir.

Yukarıda da bahsedildiği üzere adli bilimler, suçun işlendiği andan suç aydınlatılıncaya kadar geçen sürede önleyici hizmetlere her aşamada yardımcı olan bir bilim dalıdır. Adli bilimler ile adli analiz sürecinde doğru kararlar verilmesi sağlanırken gerçek suçlunun tespiti de hızlı ve doğru bir şekilde gerçekleşmektedir. En önemlisi ise milattan önceki dönemlerden bugüne kadar sürekli gelişerek ve güçlenerek varlığını sürdüren adli bilimler hukuksal düzenin temel yapı taşını oluşturmaktadır.

3. BİLİŞİM SUÇLARI VE ADLİ BİLİŞİM

1980'ler bilişim teknolojisinde yaşanan hızlı gelişmelere tanıklık etmiş ve günlük hayatın bir parçası haline gelen bilgisayar ve uygulamaları özellikle internet kullanımı ile kişi ve kurumlar için vazgeçilmez bir hal almıştır. Bilişim teknolojisinin yaşadığı bu hızlı gelişim süreci suça eğilimli bazı kişilerin ilgisini çekmiş ve gerçek dünya suçlarının yanında bilişim suçu olarak adlandırılan ve son yıllarda adı sıkça duyulan sanal dünya suçları ortaya çıkmıştır.

"Suçu işleyenlerin kazanç sağladığı, kurbanların ise zarara uğradığı bilgisayar ve teknolojinin kötüye kullanılması ile oluşan maksatlı bir davranış." olarak tanımlanan [Parker, 1989] bilişim suçu, Perry tarafından bilgisayar ve teknolojinin dolandırıcılık, sahtecilik, saldırı, kopyalama ve değiştirme gibi kötü niyetlerle kullanılması ile kişi ve kurumların zor durumda bırakılması şeklinde tanımlanmaktadır [Perry, 1986].

Bilişim suçlarının işlenmesinde etkin rol üstlenen internet, her ne kadar 1962 yılında Licklider'in "Galaktik Ağ" fikri ile ortaya atılmış olsa da 1969 yılında ARPANET projesi kapsamında 4 üniversitenin birbirine bağlanması ile adını duyurmuştur [Leiner et. al., 2009]. Dünya çapında yayın kapasitesine sahip olan internet; bilgiye kolay ve hızlı erişimden iletişime, alışverişten bankacılık işlemlerine kadar insan hayatını kolaylaştıran pek çok özelliğe sahiptir. Bu güçlü özellikleri ile internet bir yandan günlük hayatın önemli ve vazgeçilmez bir parçası haline gelirken diğer yandan suçlular için de cazip bir ortam olmaktadır.

Bilişim suçları genel olarak bilişim sistemine hukuka aykırı haksız müdahale ile kazanç sağlama, çocuk ve kadın istismarı, internet dolandırıcılığı ve sahtecilik, kimlik hırsızlığı, uyuşturucu madde ticareti, kumar oynatma, yazılım ve sanat eseri korsanlığı, forumlar veya oluşturulan web sayfaları üzerinden terörizm ve ırkçılık gibi yasadışı yayın yapılması, elektronik postların tehdit unsuru olması şeklinde sınıflandırılmaktadır.

Bilişim sistemine hukuka aykırı haksız müdahale ile kazanç sağlama; bilişim sistemine girerek sistemdeki veri ve programların bozulması, değiştirilmesi, verilerin ele geçirilmesi, erişimin kısıtlanması şeklinde tanımlanmaktadır [Taşkın, 2008]. Sistemdeki veri ve programlara verilen zararlar sistemin sahibini zor durumda bırakmakla birlikte psikolojik ve maddi zararlar verebilir. Bununla beraber kişiye ait ve sistemde bulunan yazışma, görüntü ve ses verilerini ele geçirip haberleşme özgürlüğünün ihlali ve özel hayata müdahale suçları işlenmiş olur. Elde edilen veriler şantaj amaçlı kullanılarak ölüme sebebiyet veren sonuçlar doğurabilir. Bilişim sistemlerine ve web sitelerine izinsiz girerek belli bir amaca yönelik mesaj iletimi yapılması amaçlanan internet korsanlığı da bu suç türü içerisinde yer almaktadır.

Çocuk ve kadın istismarı da son yıllarda sıkça duyulan bilişim suçları arasında yer almaktadır. Çocuk ve kadınlara ait kötü niyetli resim ve video dosyalarının bilgisayarlar üzerinde kaydedilmesi ve internet üzerinden dağıtılması aynı zamanda kadın ticaretinin de internet üzerinden yapılması ciddi bir suç teşkil etmektedir.

Phishing olarak bilinen internet dolandırıcılığı elektronik postalar üzerinden gerçekleştirilmektedir. Özellikle bankalardan geliyor gibi hazırlanan elektronik postalar, içerilerinde bulundurdukları linkler ile bankanın resmi web sitesinin birebir kopyası olan sahte bir siteye yönlendirmeyi amaçlamaktadır. Bu link üzerinden sahte siteye girip kullanıcı adlarını ve parolalarını veren kullanıcılar internet dolandırıcılığının mağduru olmaktadır [Ranganayakulu, 2011]. Sahtecilik ise sosyal siteler ve sohbet odaları üzerinden bir başkasının adına hesap açıp bu hesap üzerinden kötü niyetli işlemler gerçekleştirmek şeklinde tanımlanmaktadır.

Kimlik hırsızlığı, bir başkasına ait bilgilerin yasa dışı yollardan elde edilip dolandırıcılık amacı ile kullanılması şeklinde tanımlanabilir. İnternet bankacılığı ile ilgili kullanıcı adı ve parola, elektronik posta, sohbet odaları veya sosyal paylaşım sitelerine ait üyelik bilgilerinin elde edilmesi amaçlanmaktadır. Edinilen kişisel bilgiler ile banka hesaplarının boşaltılması, kredi kartlarının kullanılması, elde edilen kullanıcı adı ve parola bilgileri ile birlikte yasal kullanıcının zor durumda bırakılması, yasal kullanıcı üzerinden dolandırıcılık yapma suçları son günlerde medyada da sıkça yer almaktadır.

Günümüzde uyuşturucu madde ticareti gelişen teknoloji ile birlikte internet üzerinden yapılmaya başlanmıştır. Sanal dünya üzerinden yazışmalar yaparak ticaretin ne şekilde ve nasıl yapılacağı gibi ayrıntılar belirlenmekte ve para akışı da internet üzerinden gerçekleşmektedir. İnternet üzerinden kumar oynama, oynatma ve kumar oynanmasına yer sağlama da bilişim suçları arasında yer almaktadır. Bu konuda en bilinen örnek ise spor müsabakaları için internet siteleri üzerinden oynanan bahislerdir.

Yazılımların, müziklerin, basılı eserlerin, filmlerin, televizyon programlarının ve oyunların kişisel bilgisayarlara kopyalanması, paylaşılması, satılması şeklinde tanımlanan yazılım ve sanat eseri korsanlığı yasal sahipleri maddi ve manevi yönden zor durumda bırakmaktadır [Jamil and Zaki, 2011]. Yasalar ise bu tür suçları önlemede çoğu zaman yetersiz kalmaktadır. Yasadışı örgütler de bilişim teknolojilerini kendi amaçları için aktif bir şekilde kullanmaktadır. Forumlar ve oluşturdukları web siteleri üzerinden terörizm içerikli ve ırkçı yayınlar yaparak toplumdaki bireyler üzerinde psikolojik baskı kurmaktadırlar.

İnsanların bilgisayarlar üzerinden internet aracılığı ile yazılı haberleşmesini sağlayan elektronik postalar da suç işlemek için uygun bir ortam oluşturmaktadır. Bir başkasının hesabına girip onun adına hoş olmayan elektronik posta gönderimi, elektronik postalar aracılığı ile zararlı kod gönderimi, istenmeyen posta gönderimi gibi durumlarla sıklıkla karşılaşılmaktadır. Bilişim suçlarının aydınlatılmasında adli bilimlerin bir alt dalı olan ve bilişim sistemlerinin adli açıdan değerlendirilmesini ve analizini yapan adli bilişimden yararlanılmaktadır [Hankins et. al., 2009]. Adli bilişim suç ile ilgili olabilecek tüm delillerin elde edilmesi, uygun koşullarda muhafaza edilmesi, incelenmek üzere ilgili birimlere gönderilmesi ve elde edilen sonuçların raporlanıp ilgili kurumlara sunulması adımlarından oluşmaktadır.

Adli bilişim içerisindeki bir suçun aydınlatılmasında bilişim sistemleri üzerinden suçu aydınlatmak üzere elde edilen deliller dijital delil olarak adlandırılmaktadır. Dijital deliller; işlenen bilişim suçunu aydınlatabilmeli, suç ile arasında bir neden sonuç ilişkisi bulunmalı, bulunduğu ortamdan herhangi bir kayba uğramadan elde edilmeli, delillerin analizinde bilimsel bir yol izlenmeli, deliller

üzerinden analiz sonunda elde edilen sonuçlar inandırıcı olmalı ve deliller ileriki zamanlarda gerçekleşecek aynı olaylarda aynı şekillerde adli bilişim uzmanlarının karşısına çıkmalıdır [Henkoğlu, 2011].

Bilişim suçlarının aydınlatılmasında kullanılan dijital deliller; bilişim sistemleri üzerinden elde edilen tarih ve zaman bilgileri, silinmiş dosyalar, internet geçmişi kayıtları, sohbet kayıtları, videolar, sesler, fotoğraflar, içerisinde çeşitli kişilere ait kullanıcı adı ve parolaların bulunduğu metin dosyaları, gönderilen elektronik postalar, adres defterleri, korsan olarak indirilen sanat eserleri, yazılımlar, yazılımlara ait sahte lisanslar şeklinde sıralanabilir. Bahsedilen bu deliller; bilgisayarlardan, veri depolama birimlerinden, cep telefonlarından, dijital fotoğraf makinelerinden, dijital kameralardan elde edilmektedir.

Dijital delillerin elde edilmesinde adli bilişim uzmanlarının görev alması delillerin güvenliği açısından oldukça önemlidir. Olay yerine ilk varıldığında uçucu delil olarak adlandırılan ve bilişim sistemlerinin geçici kayıt bölgelerinde tutulan verilerin elde edilmesi gerekmektedir [Henkoğlu, 2011]. Elektrik kesildiğinde kaybolacak olan bu veriler bazı durumlarda işlenen suç ile ilgili önemli ipuçları oluşturabilir. Uçucu deliller elde edildikten sonra sıra bilişim sistemleri üzerinde kayıtlı olan diğer delillerin elde edilmesi aşamasına gelir ve bu amaçla dijital delillerin bulunduğu bilişim sisteminin imajının alınması gerekir. İmaj alımı ile bilişim sisteminin tüm birimleri birebir kopyalanmaktadır. Alınan imaj ile diskin tüm birimleri okunabilir, silinmiş dosyalara erişim sağlanabilir, şifrelenmiş dokümanlar belirlenebilir, çeşitli mail servislerinin oluşturdukları dosyalar tanınabilir, tüm dosya uzantıları tanınabilir. Burada dikkat edilmesi gereken ise alınan imajda herhangi bir değişikliğin yapılmamasıdır. İmaj üzerine hash algoritması ile imaja özel bir anahtar üretilir ve imajda herhangi bir değişiklik yapılmadığı garanti edilir.

Bilişim suçlarının aydınlatılmasında dijital delillerin yanı sıra fiziksel delillerden de faydalanılır. Suçun işlendiği bilişim sistemi, parmak izinin bulunabileceği çeşitli donanımlar, etrafta suç ile ilgili olduğu düşünülen kağıt parçaları da incelenmek üzere ilgili birimlere gönderilir. Elde edilen delillerin muhafaza edilmesi de ayrıca dikkat edilmesi gereken bir durumdur. Manyetik

ortamda bulunan delillerin ısıdan, kimyasallardan, çizilme, kırılma gibi ortam ve durumlardan kesinlikle korunması gerekir. Deliller elde edildikten sonra inceleme ve analiz için adli bilişim laboratuvarlarına gönderilir.

Analiz aşaması ciddi bir teknik bilgi gerektirmektedir. Elde edilen delillerin ne şekilde ele alınacağı, olayla ilişkili olup olmadığı ve olayı aydınlatıcı özelliğe sahip olup olmadığı tespit edilmelidir. Bilişim sistemindeki dosyaların oluşturulma ve değiştirilme zamanları, dosya sahiplikleri ve silinen dosyalar ile ilgili bilgi alma, kriptolu verileri çözme, zararlı kodların bulunduğu dosyaları tespit etme, uzantıları değiştirilmiş dosyaları belirleme analiz adımlarını oluşturmaktadır. Bununla birlikte işletim sisteminin tuttuğu kayıtlardan da analiz aşamasında yararlanılmaktadır. Gönderilen elektronik postalar üzerinden adli analizi mesaj başlıklarının incelenmesi ile yapılmaktadır. Yalnız bu yöntem uzmanları gerçek suçluya götürmede her zaman başarılı olamamaktadır. Bu nedenle elektronik postaların adli analizinde yeni yöntemlerin kullanılmasının gerekliliği ortaya çıkmıştır. Deliller üzerinden gerekli tüm analizler yapıldıktan sonra adli birimlere sunulması için uygun bir raporlama yapılmalıdır. Raporda, deliller, olaylar ile ilişkileri ve analiz sonuçları verilmelidir.

3.1. Elektronik Postaların Adli Analizi

İnsanların bilgisayarlar aracılığı ile yazılı haberleşmelerini sağlayan elektronik postalar popüler bir iletişim aracıdır [Gribbin, 2011]. Hızlı ve ekonomik olmaları nedeniyle insanlar arasında bilgi ve belge paylaşımı gibi pek çok alanda kullanılan elektronik postalar suça eğilimli kişilerin de ilgisini çeken bir ortam olarak karşımıza çıkmaktadır.

Elektronik postalar adli bilişim açısından ele alındığında günümüzde; önemli bilginin yetkisiz iletimi, zararlı kod (virüs, truva atları, solucanlar), phishing, tehdit mesajı, cinsel içerikli mesaj ve istenmeyen posta (spam) gönderimi, yasaklı propaganda yapma, yazılım ve sanat eserlerinin dağıtımı, dolandırıcılık gibi kötü amaçlarla kullanılmaktadır. Sahte hesap açılarak, kimlik hırsızlığı ile elde edilen bilgileri kullanıp başkasının hesabına girerek veya mesajın gönderimi sırasına içeriğini değiştirerek işlenen suçlar yasal kullanıcıları zor durumda bırakmaktadır.

Bu tür durumları önlemede virüs programı, güvenlik duvarı, şifre koruma kullanımı yetersiz kalmaktadır ve suçlarla mücadelede etkili yöntemlerin kullanılmasının gerekliliği doğmaktadır.

Adli bilişim elektronik postalar ile işlenen suçların aydınlatılmasında gönderilen elektronik postaya ait başlık bilgilerini kullanarak gerçek suçluyu bulmayı amaçlamaktadır. Bilişim suçları içerisinde ayrı bir önemi olan elektronik postalar gövde ve başlık kısmı olmak üzere iki kısımdan oluşmaktadır. Mesajın gövdesi gönderilen mesajın metin kısmına karşılık gelmektedir. Elektronik postaların adli analizinde kullanılan başlık kısmı da çeşitli alt alanlardan oluşmaktadır ve bunlar aşağıda maddeler halinde belirtilmiştir.

- Subject: Gönderilen mesajın içeriğini özetleyen ve genellikle bir iki kelimeden oluşan konu alanına karşılık gelmektedir.
- From: Mesajın hangi mail adresinden geldiğini gösteren alandır.
- Date: Mesajın gönderilme tarihi ve zamanını göstermektedir.
- To: Mesajın gönderildiği mail adresini göstermektedir.
- Return-path: Elektronik posta cevaplandığında mesajın hangi adrese iletileceğini göstermektedir.
- Received: Elektronik postaların başlık bilgisi incelendiğind ebirden fazla Received alanı görülmektedir. Gönderilen mesajın iletilirken geçtiği tüm elektronik posta hizmet sunucularının kaydını içermektedir.
- Content-type: İçerik türü ve karakter tanımlamalarını içerir.
- Message-id: Elektronik postaya verilen numaradır ve sadece gönderildiği elektronik postaya özel olarak üretilir.

Mesaj başlığında bulunan Received alanı adli bilişimcileri gerçek suçluya götürmede kullanılmaktadır. Received alanı mesajın çıktığı yerden vardığı yere kadar izlediği yolu gösterir ve buradan mesajın gönderildiği bilgisayarın IP adresi tespit edilerek gerçek suçluya ulaşılması amaçlanır ancak durum her zaman bu kadar basit olmayabilir. Anonim bir ortam olan internete güvenli ya da güvensiz pek çok alanda girilmektedir ve bunların başında da internet kafeler gelmektedir. Sahte hesap ile internet kafeden gönderilen mesajların gerçek sahibini belirlemek neredeyse imkansızdır. Bunun yanında yurtdışı elektronik posta servilerini kullanarak

gönderilen mesajlar da adli bilişimcilerin işlerini zora sokmakta [Özmutlu ve Özmutlu, 2003], mesajın gönderildiği bilgisayar ağı ile internet arasında bulunan güvenlik duvarı da gerçek suçluyu belirlemede sorun oluşturmaktadır. Var olan sorunun çözümlenememesi ve mevcut yöntemlerin bu sorunu çözmede yetersiz oluşu elektronik postaların gerçek sahiplerini tespit etmek için adli bilişim içinde yazar analizinin bir alt dalı olan yazar belirlemeye ihtiyaç olduğunu açıkça göstermektedir.

4. VERİ MADENCİLİĞİ

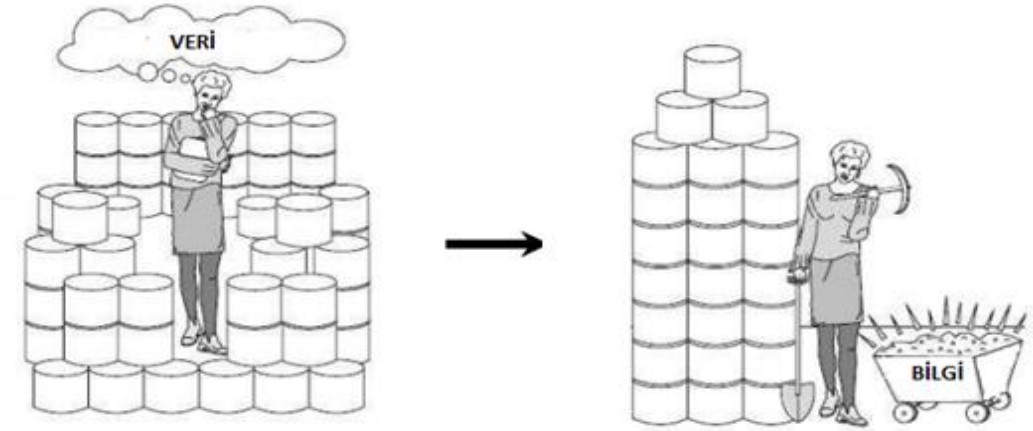
Bilgisayarların daha ucuz ve güçlü hale gelmeleri ile birlikte artan kullanımları kişi ve kurumların her alanda bu teknolojiden yararlanmaları için bir zorunluluk oluşturmaktadır. Gelişen teknolojinin de katkılarıyla günümüzün en önemli araçlarından biri haline gelen bilgisayarlar güçlü ve her geçen gün artan depolama alanları ile daha da dikkat çekmektedir. Depolama boyutlarındaki bu artışın en önemli nedeni ise günlük bazda terabaytlar seviyesinde sağlık, ticaret, market, bankacılık, eğitim, sosyal hayat, bilimsel ve daha pek çok alandaki verinin özellikle internet üzerinden de erişilmesi ile veritabanlarına kaydedilmesinin gereksinimidir.

Veritabanlarında tutulan veri ham veridir ve birden fazla alandan oluşan bu veri içerisinde gözle görülemeyecek önemli bilgiler ve ilişkiler bulunmaktadır. Bilgi çağını yaşadığımız bu dönemde eldeki veriden bilginin ve ilişkilerin çıkartılması için ise güçlü yöntemlere ihtiyaç vardır. Geleneksel istatistiksel yöntemler bu ihtiyaçları karşılayabilecek durumda değildir çünkü istatistik özet veriler ile ilgilenir, veriler genellikle indirgenmiştir ve veri boyutu büyük olmamalıdır [Hand, 1998].

1980'lerde bilgi teknolojilerinde yaşanan gelişmelerin de doğal sonucu olarak ortaya çıkan veri madenciliği; dolaylı anlatılan, önceden bilinmeyen, potansiyel olarak faydalı bilginin eldeki veriden keşfi olarak tanımlanmaktadır (Şekil 4.1.). Veri madenciliği ile eldeki veriler arasındaki gözle görülmeyen ilişkiler ortaya koyulurken, elde edilen bilgiler ışığında da geleceğe yönelik tahminler yapılabilmektedir. Artan veri boyutu ile başa çıkan, birbirinden farklı alanlardan oluşan kayıtlar üzerinden bilgiyi ve ilişkileri çıkaran, hızlı olan, dağıtık veri ile çalışabilen veri madenciliği; istatistik, yapay zeka ve makine öğrenmesi ve veritabanı disiplinlerinin sentezinden oluşmaktadır.

Örnek verilerden hareketle popülasyon hakkında yorumlama, genelleme ve tahmin istatistiksel yöntemlerle yapılmaktadır. İstatistik temel bir disiplindir ve veri madenciliği başta olmak üzere pek çok bilim istatistikten yararlanmaktadır. Veri madenciliği ve istatistiğin kesiştiği nokta ise her iki disiplinin de veri içerisindeki bilgi ile ilgilenmeleridir [Hand, 1999]. İstatistiğin tersine sezgisel yaklaşımlar ile

verilere yaklaşan yapay zeka ile yapay zekanın bir alt kolu olan ve istatistiksel yöntemleri de içinde barındıran makine öğrenmesi eldeki veriden öğrendiklerini başka veriler üzerinden çıkarım yapmak için kullanmaktadır [Kumar and Bhardwaj, 2011]. Veritabanları ise veri madenciliğinde kullanılacak verileri depolamada kullanılan bir teknoloji olarak karşımıza çıkmaktadır.



Şekil 4.1.: Veri Madenciliği [Han et al, 2012].

Veri madenciliği son yıllarda bankacılık, elektronik ticaret, sağlık, adli olaylar, yazar analizi, pazarlama gibi alanlarda yaygın bir şekilde kullanılmaktadır. Bankacılık alanında kredi kartı dolandırıcılıklarının değerlendirilmesi, müşterilere uygun fırsatların sunulması; elektronik ticarette müşterinin geçmiş verilerine dayalı olarak yeni öneriler yapılması; sağlıkta hastalıkların tespit edilmesi; adli olaylarda geçmiş veriler üzerinden tahmin yaparak gelecekte meydana gelebilecek olayların önlenmesi, gerçek suçlunun belirlenmesi; yazar analizinde eldeki metinlerin yazarlarını belirleme; pazarlamada ise müşterilere yönelik kampanyaların belirlenmesi, müşteri memnuniyeti gibi alanlarda kullanılmaktadır. Veri madenciliği, tüm bu alanlar için elde edilmiş verilerden bilgiyi çıkararak önemli başarılar sağlamaktadır.

Veri madenciliğindeki veri, veritabanlarındaki alanlara kaydedilmiş sayısal ve mantıksal ifadelerin tümüne karşılık gelmekte iken; verinin işlenmesi ile elde edilen anlamlı veri ise bilgi olarak tanımlanmaktadır. Veriler öznitelik adı verilen alanlar

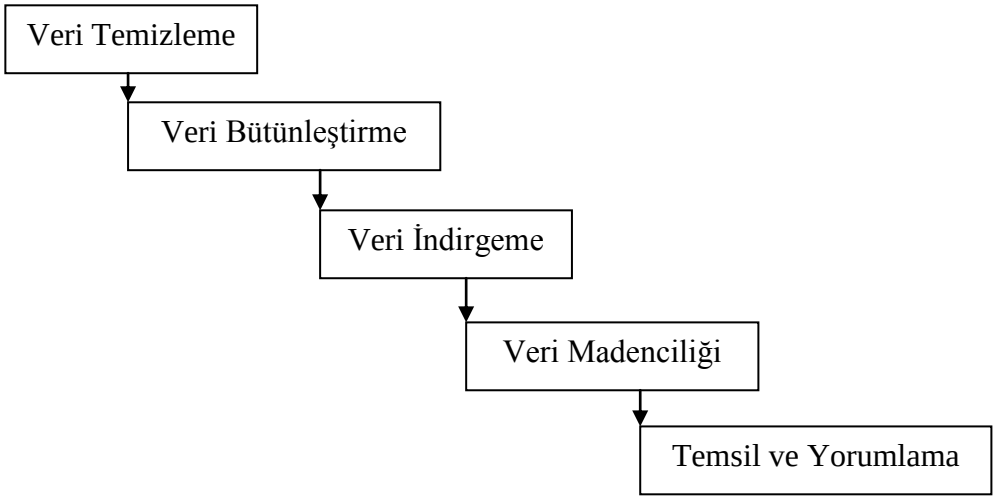
altında toplanmaktadır. Örneğin; yaş bir öznitelik alanı oluştururken yaşın değeri veri olarak tanımlanmaktadır. Veri madenciliğinde kategorik, binary, ordinal ve sayısal olmak üzere 4 çeşit öznitelik türü vardır. Kategorik özniteliklerde veri göz rengi, eğitim durumu gibi alanların alacağı sözel değerlere karşılık gelmekteyken; binary öznitelikte ise veri 0 ya da 1 değerini almaktadır. Ordinal öznitelikte veri, belli bir ölçüte göre büyükten küçüğe veya küçükten büyüğe sıralama değeri almaktadır. Örneğin; veriler sıcak, ılık ve soğuk şeklinde ifade edilmektedir. Sayısal öznitelikler de adından da anlaşılacağı üzere verilerin sayısal değerler aldığı alanlara karşılık gelmektedir. Aslında öznitelikleri sürekli ve ayrık olmak üzere iki sınıfta ifade etmek de mümkündür. Sürekli öznitelikler gerçek sayılar ve tamsayılara karşılık gelmekte iken ayrık öznitelikler de kategorik öznitelikleri, binary öznitelikleri, ordinal öznitelikleri karşılamaktadır. Ayrık öznitelikler öğrenci numarası ya da çalışanın sicil numarası gibi tamsayı değerlerini de karşılamaktadır.

Eldeki veriden çalışılacak alan ile ilgili yararlı bilginin çıkartılması için ise veri üzerinde Şekil 4.2.'deki adımların izlenmesi gerekmektedir. Burada karşımıza ilk olarak veri temizleme adımı çıkmaktadır. Üzerinde çalışılacak verilerde eksiklik, tutarsızlık veya hata var ise böyle verilere gürültü adı verilmektedir ve bu tür veriler çeşitli yöntemler ile temizlenmelidir. Eldeki veri temizlendikten sonra veri bütünleştirme ile farklı veritabanlarındaki verilerin tek bir veritabanı altında birleştirilmesi, farklı türdeki verilerin aynı türe dönüştürülmesi amaçlanmaktadır [Tüzüntürk, 2010]. Veri indirgeme ile sonucu değiştirmeyeceği düşünülen veriler ya da öznitelikler eldeki veri kümesinden çıkarılmaktadır. Veri kümesindeki çok büyük ve çok küçük değerler veri madenciliği için bir sorun teşkil edeceğinden bu durumun çözülmesi için veri dönüştürmeden yararlanılır. Veri madenciliğinde kullanılan en yaygın veri dönüştürme yöntemi Min-Max Normalleştirmesidir. Min-Max Normalleştirmesi ile amaç eldeki verileri 0-1 aralığına çekmektir.

$$x^* = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (4.1.)$$

Öznitelik alanındaki veriyi (X) 0-1 aralığına indirgemek için 4.1'deki eşitlikten yararlanılmaktadır. X_{min} öznitelik alanındaki en küçük değer, X_{max} ise en büyük değerdir. X* ise verinin 0-1 aralığındaki karşılığını temsil etmektedir. Burada veriler

üzerinde gerekli işlemler ve dönüşümler yapıldıktan sonra sıra veri madenciliği görevlerinin eldeki veri kümesine uygulanması adımına gelir. Eldeki veri kümesine görevler uygulandıktan sonra sonuçlar grafikler ile temsil edilir ve sonuçlar üzerinde yorumlama yapılır.



Şekil 4.2.: Veri Madenciliği Süreci [Yalçın, 2008].

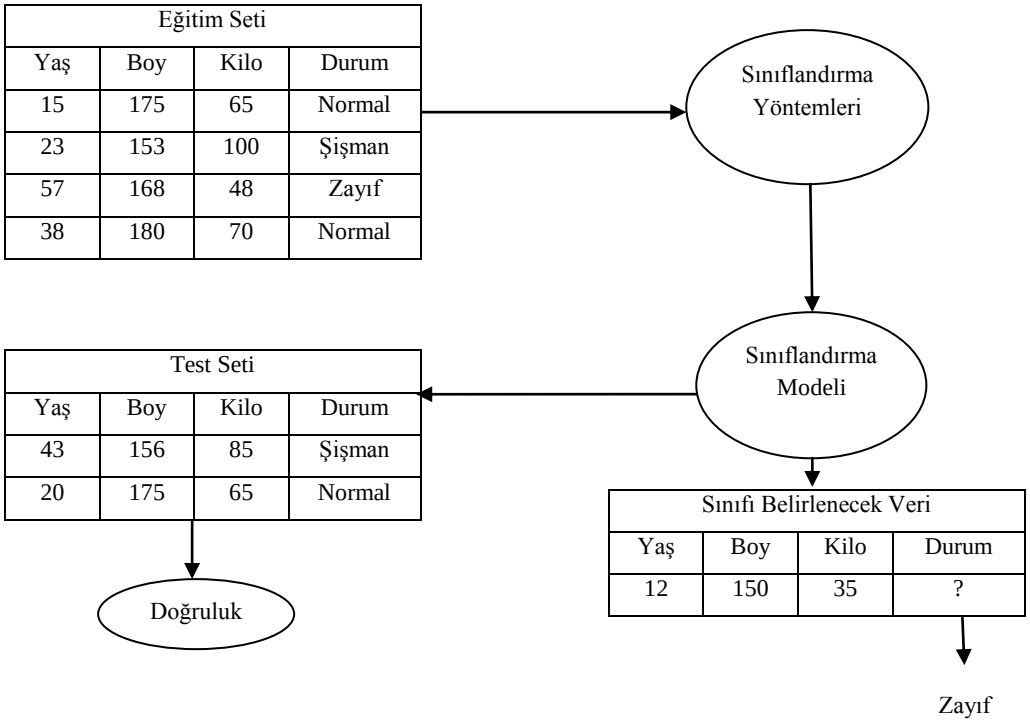
Veri madenciliği içerisinde eldeki veriden bilginin çıkartılmasında en büyük rol ise veri madenciliği görevlerine düşmektedir. Veri madenciliği; sınıflandırma, kümeleme, birliktelik kuralları keşfi olmak üzere üç temel göreve ayrılmaktadır. Bu görevler eldeki veriye ve veriden çıkarılması istenen bilgiye göre çeşitli uygulamalarda kullanılmaktadır. Bu çalışma ise sınıflandırma tabanlı bir çalışmadır.

4.1. Sınıflandırma

Günümüzde; kredi kartı verilirken kişinin riskli olup olmadığının tespitinde, doküman sınıflandırmada, yazar belirlemede, tümör hücrelerinin iyi huylu olup olmadığının tespitinde ve daha pek çok alanda kullanılan sınıflandırma, eldeki veri içerisinden gizli kalmış örüntülerin çıkartılması olarak tanımlanmaktadır. Çıkartılan bu gizli örüntüler bilgiye karşılık gelmektedir ve bu bilginin kullanımı ile geleceğe yönelik tahmin yapılabilmektedir.

Üzerinde sınıflandırma yapılacak veri, girdilerden (öznitelikler) ve bu girdiler ile ilişkili olan çıktı (sınıf etiketleri) alanlarından oluşmaktadır. Sınıflandırmada temel kural verilen girdiler ile çıktı arasındaki ilişkinin çözülmesine dayanmaktadır. Öğretmenli öğrenme olarak adlandırılan bu yöntem sınıflandırmanın da temel mantığını oluşturmaktadır.

Eldeki veri üzerinde sınıflandırma yapılırken Şekil 4.3.'teki belli adımların izlenmesi gerekir. İlk olarak veri eğitim seti ve test seti olmak üzere iki kısma ayrılır. Eğitim seti üzerinden sınıflandırma yöntemleri ile bir model oluşturulur. Bu model girdiler ile çıktılar arasındaki ilişkileri barındırmaktadır. Daha sonra bu model test setine uygulanır ve oluşan modelin test seti üzerinden doğruluğu tespit edilir. Sınıflandırmadaki amaç elde edilen modelin başarısına göre yeni gelen ve sınıfı bilinmeyen verileri doğru sınıflara atamaktır.



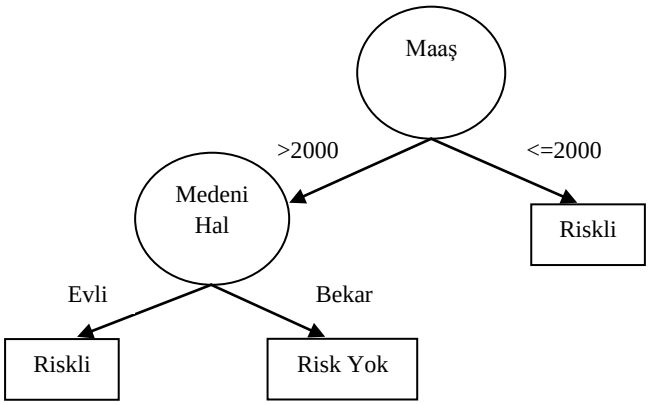
Şekil 4.3.: Sınıflandırma.

Sınıflandırma içerisinde yer alan ve eğitim seti üzerinden model oluşturulmasını sağlayan yöntemlerden; karar ağaçları, bayes sınıflandırıcılar, yapay sinir ağları, destek vektör makineleri ve gruplamalı sınıflandırıcılar bu çalışmada

kullanılmıştır. Bunların yanı sıra bayes ağları, sık örüntüler, kural tabanlı yaklaşım ve k en yakın komşu olmak üzere çeşitli sınıflandırma yöntemleri de mevcuttur ve pek çok uygulamada kullanılmaktadır.

4.1.1. Karar Ağaçları

Anlaşılır olması, kolay ve düşük maliyetli kurulumu, veritabanlarına kolayca uyum sağlaması, gürültüye karşı duyarlı olması ve güvenirlikleri ile karar ağaçları günümüzde en sık tercih edilen veri madenciliği yöntemlerinden bir tanesidir [Barros et al, 2012]. Öğretmenli öğrenme mantığı ile çalışan karar ağaçları adında da anlaşılacağı üzere ağaç yapılı bir veri yapısıdır. Yönlü bir ağaç olan karar ağaçları; karar düğümleri, dal ve yapraklardan oluşmaktadır ve en üstteki karar düğümü kök düğüm olarak adlandırılmaktadır. Ağaçtaki karar düğümleri öznitelikleri, dallar öznitelik değerlerini, yaprak düğümler ise sınıf etiketlerini tutmaktadır. Karar ağaçlarında her düğüme sadece bir dal gelebilir ancak bir düğüm üzerinden birden fazla dallanma olabilir [Rokach and Maimon, 2008]. Bir kayıt geldiğinde kök düğümdeki öznitelikten başlanır özniteliğin tuttuğu veriye göre dallardan biri seçilir ve sınıf etiketinin tutulduğu yaprağa inene kadar bu işlem devam eder. Kök düğümlerinden yaprağa inildiğinde ise artık karar ağaçları üzerinden yeni gelen kaydın sınıf etiketi belirlenmiş olur. Şekil 4.4'te bir kişiye kredi verilmesindeki riskin tespiti için örnek bir karar ağacı gösterilmiştir. Şekilden de anlaşılacağı üzere daire ile gösterilenler öznitelikleri tutmakta iken dikdörtgen ile gösterilenler de sınıf etiketlerini tutmaktadır.



Şekil 4.4.: Karar Ağacı.

Karar ağaçlarındaki en önemli konu ise dallanmanın hangi düğümden başlayacağıdır yani pek çok öznitelik bulunduran veri kümesinde hangi özniteliğin kök düğüm olacağıdır. Burada açgözlü bir yaklaşım vardır ve kayıtlar belli kıstasları optimize eden bir özniteliğe dayalı olarak dallanırlar. Sınıf etiketleri belli olan kayıtların tümü üzerinde bu yaklaşım gerçekleştirilir. Bir öznitelik kök düğüm olarak belirlendikten sonra ise geriye kalan öznitelikler için de aynı işlem özyinelemeli olarak ve bütün kayıtlar aynı sınıfa ait oluncaya kadar devam eder. Sonuç olarak "eğer ise" yapısında kurallar oluşur. Şekil 4.4.'te elde edilen kurallardan bir tanesi; "Eğer Maaş>2000 ise ve Eğer Medeni Hal Evli ise Risklidir." şeklindedir. Dallanma kıstasının belirlenmesinde en sık kullanılan yaklaşım ise entropiye dayalı algoritmalardır. Bu algoritmalar ID3, C4.5 ve J48 olmak üzere üçe ayrılmaktadır ve çalışmada J48 karar ağacı algoritmasından yararlanılmıştır. Karar ağaçları günümüzde kredi riskinin tespitinde, yazar belirleme, metin sınıflandırma gibi alanlarda sıklıkla kullanılmaktadır.

4.1.1.1. J48 Algoritması

ID3 ve C4.5 algoritmalarından geliştirilen J48 karar ağacı algoritması günümüzde sıklıkla kullanılmaktadır. Ouinlan tarafından geliştirilen ve basit hesaplamalar ile büyük veri setleri üzerinden hızlı bir şekilde karar ağacı oluşturabilen ID3 ve C4.5 çalıştıkları öznitelik türlerine göre farklılık göstermektedir. ID3 yalnızca kategorik öznitelikler ile çalışırken, ID3'ün gelişmiş versiyonu olan C4.5 sayısal öznitelikler ile çalışmaktadır. ID3 temelli olan C4.5'te sayısal öznitelikler kategorik özniteliğe dönüştürüldükten sonra ID3 algoritması eldeki veri kümesine uygulanmaktadır. Sayısal öznitelikler kategorik özniteliğe dönüştürülürken o özniteliğe ait veri değerlerinin ortalaması alınır ortalamadan büyük olanlar büyük, küçük ve eşit olanlar ise küçük veya eşit olarak değer alır. J48 algoritması ise hem kategorik hem de sayısal öznitelikler ile çalışması ile kolaylık sağlamakta ve güçlü olmaktadır. Bu üç algoritmanın en önemli ortak özelliği ise hangi özniteliğin kök düğüm olacağının belirlenmesinde entropiden faydalanmalarıdır.

Entropi, bir sistemdeki belirsizliğin ölçüsüne verilen addır. Olasılık tabanlı bir yöntem olan entropiyi bir örnek ile açıklamak gerekirse; G bir oyun olsun ve bu

oyunda m tane sayı birbirinden bağımsız olarak {n1, n2, n3, …, n_m} şeklinde üretilsin [Özkan, 2008]. Üretilen n_i sayılarının olasılığı P={p_1, p_2, p_3, …, p_m} şeklindedir. Bu G oyununun entropisi ise eşitlik 4.2.'ye göre hesaplanmaktadır.

$$H(G) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (4.2.)$$

Her ne kadar bu üç algoritma entropi tabanlı olsa da hangi düğümün kök düğüm olacağının ve dallanmanın ne şekilde gerçekleşeceğinin belirlenmesinde bilgi kazancı kullanılmaktadır. Bilgi kazancı; sınıf etiketleri üzerinden elde edilen entropi değerinden, ilgili öznitelikten elde edilen entropi değerinin çıkartılması ile elde edilmektedir. Ağacın oluşumu sırasında en yüksek bilgi kazancı sağlayan öznitelik kök düğüm olmaktadır ve diğer düğümlerin belirlenmesinde aynı işlem özyinelemeli şekilde devam etmektedir. Aşağıda J48 algoritmasının Çizelge 4.1.'deki veri kümesine uygulanması ile bir karar ağacı elde edilmektedir.

Çizelge 4.1.: Veri Kümesi.

Cinsiyet	**Gelir**	**Yaş>25**	**Risk Durumu**
Bay	Yüksek	Evet	Risk Yok
Bayan	Yüksek	Hayır	Riskli
Bay	Düşük	Hayır	Riskli
Bay	Yüksek	Hayır	Risk Yok
Bayan	Düşük	Evet	Riskli
Bay	Düşük	Evet	Risk Yok
Bay	Düşük	Hayır	Riskli
Bay	Yüksek	Hayır	Risk Yok
Bay	Düşük	Evet	Risk Yok
Bay	Düşük	Evet	Risk Yok

Birinci aşamada sınıf etiketi olan Risk Durumu için entropi hesabı yapılır ama öncelikle Risk Durumu'nun iki farklı değerinin olasılığı $p_{\text{Risk Yok}}=\frac{2}{10}$, $p_{\text{Riskli}}=\frac{8}{10}$ şeklinde hesaplanır.

$$H(Risk\ Durumu) = -\left(\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right) = 0.97$$

İkinci aşamada Cinsiyet, Gelir ve Medeni Hal öznitelikleri için aldıkları değerlerin olasılıkları Risk Durumu'na göre kendi içlerinde hesaplanıp entropi hesabı yapılır. Son olarak da bilgi kazancı bulunur.

$$H(Cinsiyet_{Bay}) = -\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) = 0.81$$

$$H(Cinsiyet_{Bayan}) = -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

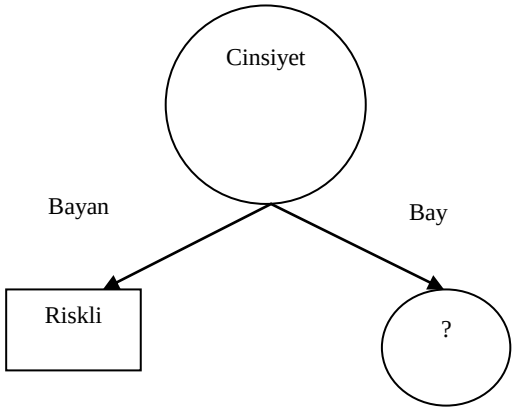
$$H(Cinsiyet, Risk\ Durumu\ ) = \left(\frac{8}{10}0.81\right) + \left(\frac{2}{10}0\right) = 0.65$$

$$Kazanç(Cinsiyet, Risk\ Durumu) = 0.97 - 0.65 = 0.32$$

$$Kazanç(Gelir, Risk\ Durumu) = 0.97 - 0.88 = 0.09$$

$$Kazanç(Yaş, Risk\ Durumu) = 0.97 - 0.85 = 0.12$$

Sonuçlara baktığımızda en yüksek kazancın Cinsiyet özniteliği ile sağlandığı görüldüğünden kök düğüm cinsiyet olarak belirlenmiştir (Şekil 4.5.).



Şekil 4.5.: Kök Düğümü Belli Olan Karar Ağacı.

Cinsiyet özniteliği Çizelge 4.1.'de de görüldüğü üzere Bay ve Bayan olmak üzere iki değer almaktadır Bayan değeri için sınıf etiketi her koşulda Riskli olduğu için hemen yaprağa inilmekte ve sınıf etiketi belirlenmektedir. Sıra ikinci düğümü belirlemeye geldiğinde ise veri kümesinde Çizelge 4.2.'de görüldüğü gibi sadece 8 kayıt bulunmaktadır. Burada elde kalan veri kümesinden sadece Risk Durumu, Gelir ve Yaş öznitelikleri için entropi ve bilgi kazancı hesabı yapılarak bir sonraki düğümün belirlenmesi amaçlanmaktadır.

Çizelge 4.2.: Veri Kümesi.

Cinsiyet	**Gelir**	**Yaş>25**	**Risk Durumu**
Bay	Yüksek	Evet	Risk Yok
Bay	Düşük	Hayır	Riskli
Bay	Yüksek	Hayır	Risk Yok
Bay	Düşük	Evet	Risk Yok
Bay	Düşük	Hayır	Riskli
Bay	Yüksek	Hayır	Risk Yok
Bay	Düşük	Evet	Risk Yok
Bay	Düşük	Evet	Risk Yok

$$H(Risk\ Durumu) = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{6}{8}\log_2\frac{6}{8}\right) = 0.81$$

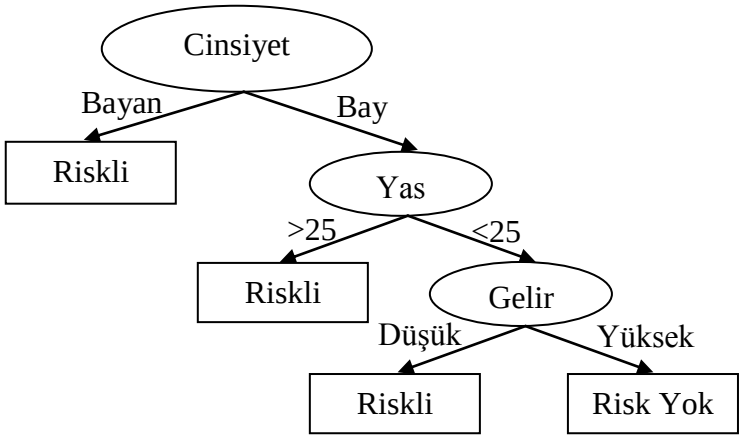
$$H(Gelir, Risk\ Durumu\ ) = \left(\frac{3}{8}0\right) + \left(\frac{5}{8}0.97\right) = 0.61$$

$$H(Yaş, Risk\ Durumu\ ) = \left(\frac{4}{8}0\right) + \left(\frac{4}{8}1\right) = 0.5$$

$$Kazanç(Gelir, Risk\ Durumu) = 0.81 - 0.61 = 0.2$$

$$Kazanç(Yaş, Risk\ Durumu) = 0.81 - 0.5 = 0.31$$

Bu aşamada da en yüksek kazanç Yaş özniteliğinden sağlandığı için ağaca yaş özniteliği düğüm olarak eklenmiştir. Yaş özniteliği de ağaca eklendikten sonra elde sadece Gelir özniteliği kalmıştır. Gelir özniteliği de aldığı iki değere göre ağaca yerleştirilmiş ve ağaç Şekil 4.6.'daki halini almıştır.



Şekil 4.6.: Karar Ağacının Son Hali.

4.1.2. Bayes Sınıflandırıcılar

Sınıflandırma problemini çözmede olasılık tabanlı bir çerçeve sunan bayes sınıflandırıcılar istatistiksel bir yöntem olarak karşımıza çıkmaktadır. Öğretmenli öğrenme yöntemini kullanarak sınıflandırma yapan bayes sınıflandırıcılar Bayes Teoremi tabanlı çalışmaktadır.

Şartlı olasılık üzerine kurulu olan Bayes Teoremi adını 18. yüzyılda yaşayan İngiliz istatistikçi Thomas Bayes'ten almaktadır. Bir olayın gerçekleşmesi için bazı koşulların olması gerekiyorsa buna şartlı olasılık adı verilmektedir [Özkan, 2008]. Şartlı olasılığı bir örnek üzerinden açıklamak gerekirse E_1 ve E_2 kesişen iki olay olsun. E_1 olayının gerçekleşmesi E_2 olayının gerçekleşmesine bağlı ise bu durum Eşitlik 4.3. ile ifade edilir. Aynı şekilde E_2 olayının gerçekleşmesi E_1 olayının gerçekleşmesine bağlı ise bu durumdaki şartlı olasılık da Eşitlik 4.4.'te gösterilmiştir.

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} \tag{4.3.}$$

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} \tag{4.4.}$$

Bayes Teoremi daha önceden de belirtildiği gibi şartlı olasılık üzerine kuruludur ve olaylar ile hipotezler arasındaki olasılığı hesaplamaktadır. R, veri

kümesinde bulunan bir kayıt (olay), H ise R kaydının sınıfını C olduğunu iddia eden hipotez olsun. Bayes Teoremi'nin amacı R kaydının C sınıfına ait olma olasılığını H hipotezi üzerinden hesaplamaktır, kısacası burada hesaplanmak istenen P(H|R)'tir. P(H|R), H hipotezinin R olayı üzerinde koşullandırılması ile oluşan sonrasal olasılık olarak tanımlanmaktadır. Bayes Teoremini bir örnek ile açıklamak gerekirse; bir öğretmen fizik dersinden kalan öğrencilerin %50'sinin matematik dersinden de kaldığını bilmektedir. Bu durum P(R|H)'a karşılık gelmektedir. P(R|H); R olayının H hipotezi üzerinden koşullandırılması ile oluşan sonrasal olasılık olarak tanımlanmaktadır. Herhangi bir öğrencinin fizik dersinden kalma olasılığı tüm öğrenciler içerisinde %5 olarak bilinmektedir. Bu durum P(H) ile ifade edilir ve H hipotezinin öncesel olasılığı olarak tanımlanmaktadır. Herhangi bir öğrencinin tüm öğrenciler içerisinde matematik dersinden kalması ise %40 olarak bilinmektedir. Bu durum da P(R) ile ifade edilir, burada P(R) R olayının öncesel olasılığı olarak açıklanmaktadır. Burada hesaplanmak istenen P(H|R) ise bir kişi matematikten kalmış ise o kişinin fizikten kalma olasılığıdır. P(H|R)'in hesaplanması yani Bayes Teoremi Eşitlik 4.5.'te verilmiştir.

$$P(H|R) = \frac{P(R|H)P(H)}{P(R)} \tag{4.5.}$$

Bayes sınıflandırıcılar, günümüzde sıklıkla metin sınıflandırma, yazar belirleme, veterinerlik, tıp gibi alanlarda kullanılmaktadır. Bayes sınıflandırıcılar içerisinde, sınıflandırma yapmak amacıyla Naive Bayes algoritması kullanılmaktadır.

4.1.2.1. Naive Bayes Algoritması

Geniş veri tabanlarına uygulandığında hızlı bir şekilde ve yüksek doğruluk ile çalışan Naive Bayes algoritması performans açısından tercih edilen bir sınıflandırma algoritmasıdır. Naive Bayes algoritmasındaki temel mantık; veri kümesini oluşturan her öznitelik birbirinden bağımsız olmasıdır ve buna sınıf koşulu bağımsızlığı adı verilmektedir [Mahdi et al, 2011]. Naive Bayes algoritmasının çalışma adımları aşağıda verilmiştir [Han et al, 2012].

- R, n tane özniteliğe ve m tane sınıf etiketine sahip eğitim seti, T'de R ile aynı özniteliklere sahip ve sınıf etiketi belli olmayan bir kayıt olsun. Buradaki amaç Naive Bayes algoritmasını kullanarak T'nin sınıf etiketini belirlemektir.
- Sınıf etiketleri C={C_1, $C_{2,}$ …, C_m} şeklinde olsun. Ancak P(C_i|T) sonrasal olasılığını maksimum yaparsa T C_i sınıf etiketine sahiptir denilebilmektedir. P(C_i|T)'yi hesaplamak için Eşitlik 4.6. kullanılmaktadır.

$$P(\mathrm{C_i}|T) = \frac{P(T|C_i)P(C_i)}{P(T)} \tag{4.6.}$$

- P(T) değeri tüm sınıflar için aynı olacağından eldeki formülden bu değer çıkartılabilir. Ayrıca Naive Bayes sınıf koşulu bağımsızlığı özelliğini taşıdığı için formül daha basit bir forma dönüştürülür ve Eşitlik 4.7. elde edilir.

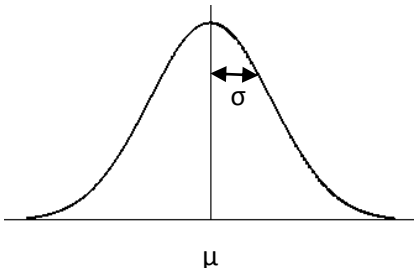
$$P(T|C_i) = \prod_{k=1}^{n} P(t_k|C_i) \tag{4.7.}$$

- Eğer öznitelikler kategorik verilerden oluşuyorsa hesaplamada artı bir işlem yapmaya gerek kalmaz yalnız elde sürekli öznitelikler varsa burada ek işlem yükü ortaya çıkar ve $P(t_k|C_i)$'yi hesaplamak için eşitlik 4.8.'deki Gauss Dağılımı'ndan (Normal Dağılım) yararlanılır.

$$P(t_k|C_i) = \mathrm{g}(\mathrm{t_k}, \mu_{\mathrm{C_i}}, \sigma_{\mathrm{C_i}}) = \frac{1}{\sqrt{2\pi}\sigma_{\mathrm{C_i}}} \mathrm{e}^{-\frac{\left(\mathrm{t_k}-\mu_{\mathrm{C_i}}\right)^2}{2\,\sigma^2_{\mathrm{C_i}}}} \tag{4.8.}$$

- Sürekli bir olasılık dağılımı olan ve bilimsel uygulamalarda sıklıkla kullanılan Gauss dağılımımda ortalama (μ) ve standart sapma (σ) olmak üzere iki parametre vardır.
- Eşitlik 4.9., sürekli bir x değişkeni için Gauss Dağılımı'ndaki olasılık yoğunluk fonksiyonuna karşılık gelmektedir.

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x}\, \mathrm{e}^{-\frac{(\mathrm{x}-\mu_{\mathrm{x}})^2}{2\,\sigma_x^2}} \tag{4.9.}$$



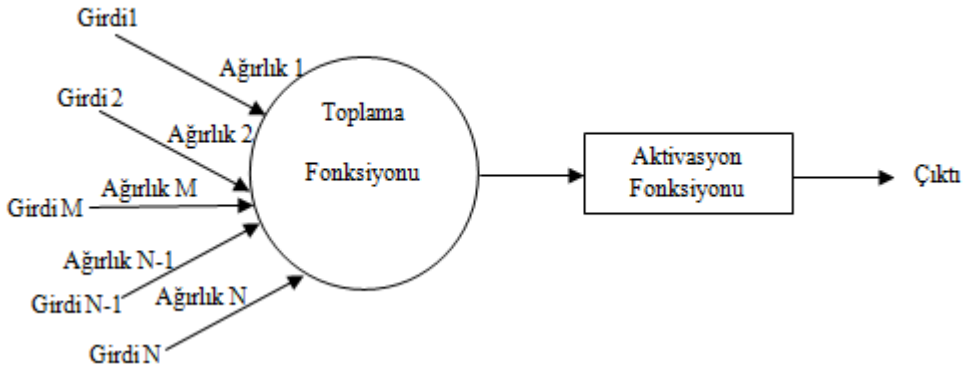
Şekil 4.7.: Gauss Olasılık Dağılımı.

- Eğrinin tepe noktası ortalamaya (μ) karşılık gelmektedir. Şekil 4.7.'de de görüldüğü üzere ortalamaya göre simetrik olan bu eğri bir çana benzediği için çan eğrisi adını almaktadır. Gauss dağılımında ortalama, mod ve medyan birbirine eşittir. Eğrinin genişliği ise standart sapmaya göre belirlenmektedir.
- Naive Bayes algoritmasında ise μ_{C_i} sürekli öznitelik değerinin ortalamasını, σ_{C_i} ise standart sapmasını göstermektedir. Sürekli bir değişken olan t_k da $-\infty \leq t_k \leq \infty$ arasında değişmektedir.
- Tüm hesaplamalar yapıldıktan sonra $P(T|C_i)\ P(C_i)$ hangi sınıf etiketi için maksimum çıkarsa T'nin o sınıfa ait olduğu anlaşılır.

4.1.3. Yapay Sinir Ağları

İnsan beyninden esinlenerek geliştirilen ve insan beyninin öğrenme yolu ile yeni bilgiler üretebilen, ilişkilendirme ve genelleme yapabilen yapay sinir ağları; birbirlerine ağırlıklı bağlantılar aracılığı ile bağlanan ve her biri kendi dağıtık belleğine sahip işlem elemanları ile paralel şekilde bilgi işleyen bilgisayar sistemlerine denir.

Kendilerine verilen örnekler üzerinden öğrenen ve daha sonra yeni gelen bir örneğe karşı nasıl tepki verebileceğini bilen yapay sinir ağları yapay sinir hücrelerinin birbirlerine bağlanmalarından oluşmaktadır. Temel bir yapay sinir hücresi Şekil 4.8.'de gösterildiği gibi 5 temel elemandan oluşmaktadır.



Şekil 4.8.: Yapay Sinir Hücresi [Öztemel, 2006].

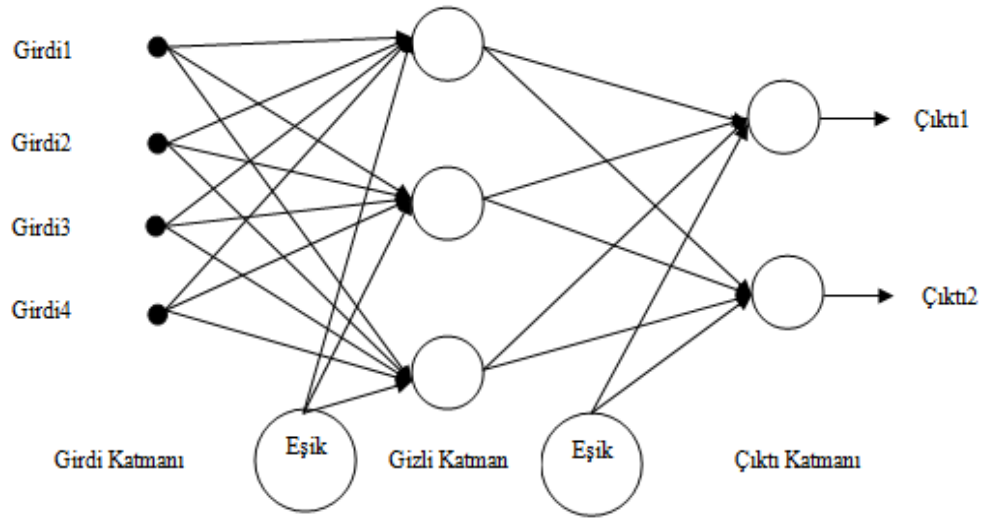
Yapay sinir hücresindeki Girdi değerleri; çevreden, bir başka sinir hücresinin çıkışından veya kendi çıkışından elde ettiği bilgiyi hücreye getirmektedir. Bu girdi değerlerinin geldiği hücredeki önemini ve hücre üzerindeki etkisini ise Ağırlık değerleri göstermektedir. Toplama fonksiyonu, yapay sinir hücresine gelen girdiyi Eşitlik 4.10.'a göre hesaplanmaktadır. Toplama fonksiyonu sonunda elde edilen değer aktivasyon fonksiyonu ile işlenir ve en son olarak da sinir hücresine gelen bilgilere karşı en son olarak çıktı üretilir. Çeşitli aktivasyon fonksiyonları olmasına rağmen en çok kullanılan aktivasyon fonksiyonu Eşitlik 4.11.'de gösterilen "Sigmoid" fonksiyonudur. Yapay sinir hücreleri ile ilgili en önemli özelliklerden bir tanesi her hücreye birden fazla girdi gelebilir ama bu girdilere karşılık sadece bir çıktı üretilir.

$$Toplama\ Fonksiyonu(NET) = \sum_{i=1}^{n} Girdi_i Ağırlık_i \qquad (4.10.)$$

$$Çıktı = \frac{1}{1+e^{-NET}} \qquad (4.11.)$$

Bir araya gelen yapay sinir hücreleri yapay sinir ağlarını oluşturur. Yapay sinir ağları Şekil 4.9.'da gösterildiği gibi temel olarak 3 katmandan oluşmaktadır. Girdi katmanı; birden fazla yapay sinir hücresinden oluşan girdi katmanında dış dünyadan gelen bilgiler herhangi bir işleme tabi tutulmaksızın gizli katmana iletilir. Girdi katmanındaki hücrelere sadece bir girdi gelir. Gizli katman da girdi katmanı gibi birden fazla yapay sinir hücresinden oluşur. Girdi katmanındaki hücrelerin her

birinden gelen bilgi gizli katmandaki tüm hücrelere girdi olarak gelir. Kısacası gizli katmandaki hücrelere girdi katmandaki hücre sayısı kadar girdi gelir. Girdi katmanından gelen bilgiler gizli katmanda işlenip çıktı katmanına gönderilir. Bir yapay sinir ağında birden fazla gizli katman bulunabilir. Çıktı katmanında da gizli katmandan gelen bilgiler işlenir ve çıktı katmanındaki her bir hücreden çıktılar elde edilir.



Şekil 4.9.: Çok Katmanlı Yapay Sinir Ağı Yapısı.

Yapay sinir ağlarında öğrenme hücreler arasındaki bağlantıların ağırlıklarının belirlenmesine karşılık gelmektedir. Başlangıçta rastgele belirlenen ağırlık değerleri eğitim setindeki örnekler ağa verildikçe değişime uğramakta ve ağ eğitildiğinde bağlantılar son halini almış olmaktadır [Öztemel, 2006]. Ağ öğrendiğinde ise artık olay üzerinde genelleme yeteneğine sahip olmaktadır.

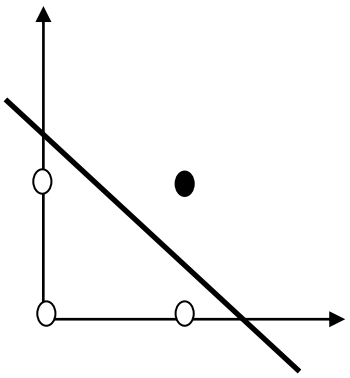
Sınıflandırma, örüntü tanıma, tahmin gibi alanlarda kullanılan yapay sinir ağlarının popüler olma nedenleri; eksik bilgi ile çalışabiliyor olmaları, hataya toleranslı olmaları, yeni olayları hızlı bir şekilde öğrenmeleri, görülmemiş örnekler hakkında bilgi üretebilmeleri gibi özelliklere sahip olmalarıdır. Bunların yanı sıra kara kutu olarak nitelendirilen yapay sinir ağlarının nasıl öğrenme gerçekleştirdiği tam olarak anlaşılamamaktadır. Yapay sinir ağları içinde doğrusal olmayan problemlerin çözümünde ise Çok Katmanlı Yapay Sinir Ağı algoritması kullanılmaktadır.

4.1.3.1. Çok Katmanlı YSA Algoritması

1958 yılında Rosenblat tarafından geliştirilen ve tek katmanlı bir algılayıcı olan Percetronlar yapay sinir ağları tarihinde milat olarak kabul edilmekle birlikte sadece doğrusal problemleri çözebilmekteydi [Öztemel, 2006]. Doğrusal problemden kasıt problem uzayının bir doğru ya da düzlemle sınıflara ayrılabilmesi demektir. Bunun en basit örneği de AND mantıksal fonksiyonu (Çizelge 4.3.) ile Şekil 4.10.'da gösterilmektedir.

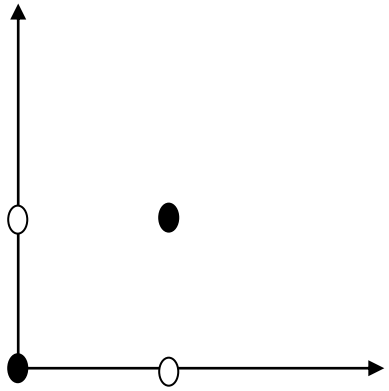
Çizelge 4.3.: AND Mantıksal Fonksiyonu.

Girdi1	Girdi2	Çıktı
0	0	0
0	1	0
1	0	0
1	1	1



Şekil 4.10.: AND Fonksiyonu.

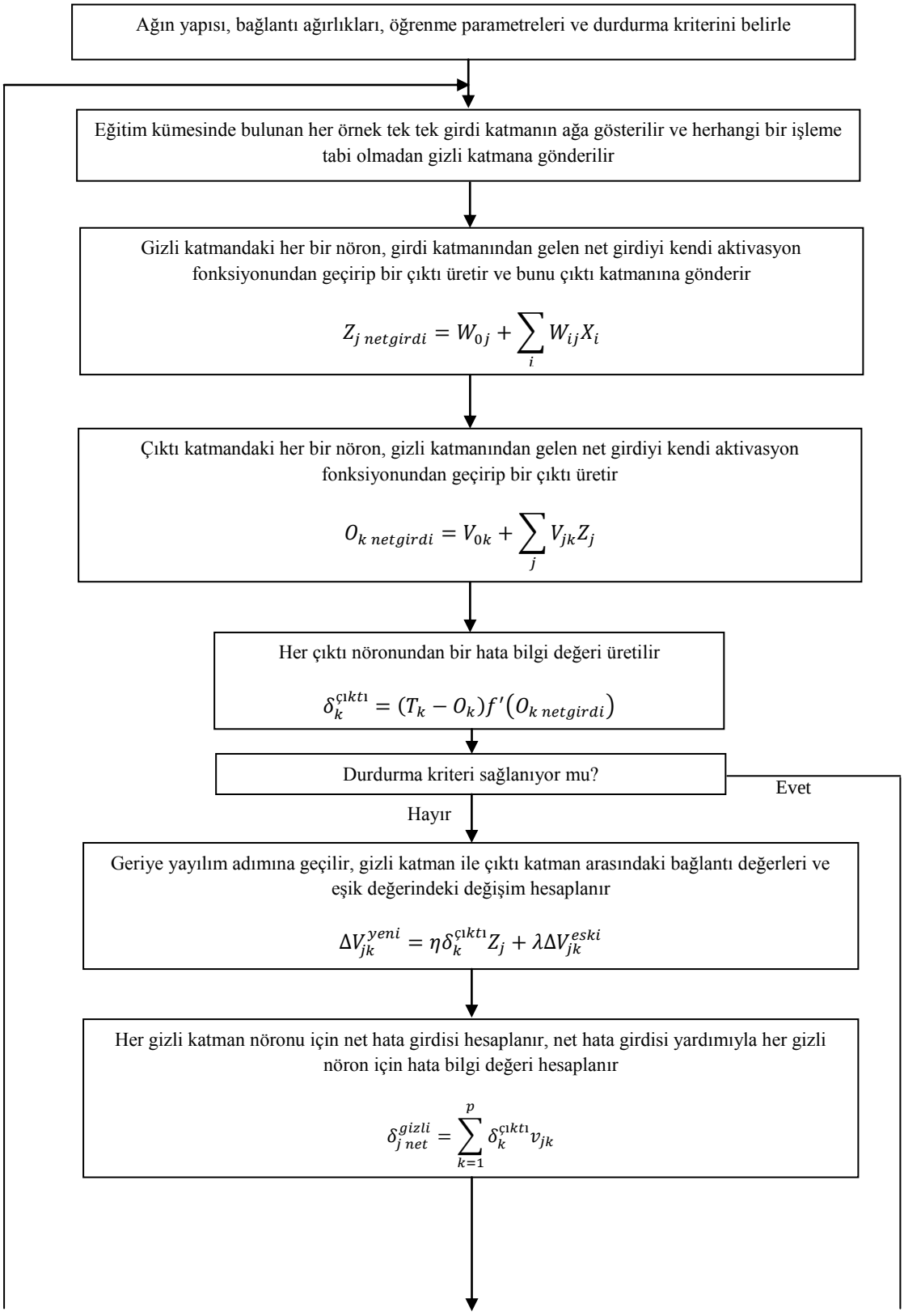
XOR problemine bakıldığında ise tek katmanlı yapay sinir ağlarının bu konuda başarısız olduğu (Şekil 4.11.), ve yeni bir yapay sinir ağı modeline ihtiyaç olduğu anlaşılmıştır. Bu durum üzerine temeli Perceptronlar olan çok katmanlı yapay sinir ağı modeli ortaya çıkmıştır.



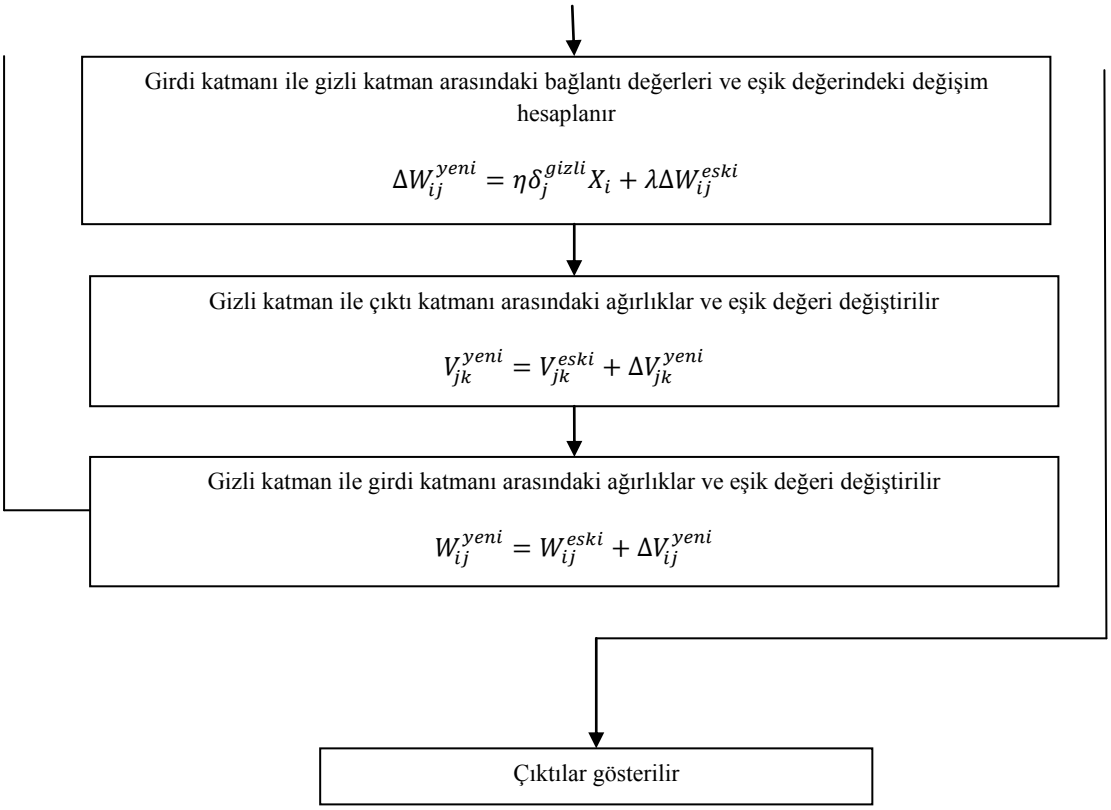
Şekil 4.11.: XOR Fonksiyonu.

Çok katmanlı yapay sinir ağları, bir girdi katmanı, bir ya da daha fazla gizli katman ve bir çıktı katmanından oluşmaktadır. Kendisine verilen girdiler ve bu girdilere karşılık verilen çıktılardan öğrenme gerçekleştiren bu yapay sinir ağı modelinde genelleştirilmiş delta öğrenme kuralı kullanılmaktadır. Genelleştirilmiş delta öğrenme kuralı ileri doğru besleme ve geriye yayılım olmak üzere iki aşamada öğrenmeyi sağlamaktadır. İleri doğru beslemede ağ kendisine verilen girdilere göre çıktıları üretirken, geriye doğru yayılımda da bağlantı ağırlıkları güncellenmektedir. Bağlantı ağırlıklarının güncellenmesindeki amaç hata değerinin en aza düşürülmesidir.

Çok katmanlı yapay sinir ağlarının çalışma mantığı genel olarak özetlenecek olursa ilk aşamada sınıflandırılması gereken örnekler toplanıyor. Bir kısmı eğitim geri kalan kısmı da test seti olmak üzere ikiye ayrılıyor. Örnekler elde edildikten sonra ise ağın kaç tane gizli katmanı olacağı ve girdi, gizli ve çıktı katmanlarında kaç tane sinir hücresi olacağı belirleniyor. Bu adımdan sonra ağın bağlantı ağırlıkları, ileri beslemede kullanılacak toplama ve aktivasyon fonksiyonları, eşik değerler, ağırlıkların değişim miktarını belirleyen öğrenme katsayısı ve bir önceki iterasyondaki değişimin bir kısmının yeni değişim miktarına eklenmesi şeklinde tanımlanan momentum katsayısı bilgilerinin belirlenmesi gerekiyor. Tüm bu adımlardan sonra eğitim setindeki örnekler ki bu örnekler sayısal olmak zorunda ağa gösteriliyor. Örnekler ağa gösterilerek ileri doğru hesaplama yapılıyor. Gelen çıktı ile beklenen çıktı karşılaştırılıp hata belirleniyor ve eğer hata istenilen seviyede değilse geriye doğru yayılım aşamasına geçiliyor (Şekil 4.12.) [Hamzaçebi, 2011].



Şekil 4.12.: Çok Katmanlı YSA Algoritması Akış Diyagramı [Hamzaçebi, 2011].



Şekil 4.12.Devam: Çok Katmanlı YSAAlgoritması Akış Diyagramı [Hamzaçebi, 2011].

Yukarıda akış diyagramı verilen algoritmada X ile girdileri, T de hedef çıktıyı temsil etmektedir. $Z_{j\ netgirdi}$, gizli katman nöronlarına girdi katmanından gelen net girdiyi göstermekteyken; Z_j, $Z_{j\ netgirdi}$ değerinin aktivasyon fonksiyonundan geçirilerek çıktı katmanına gönderilen değere karşılık gelmektedir. $O_{k\ netgirdi}$ gizli katmandan çıktı katmanına gelen net girdiye karşılık gelmekteyken; O_k, $O_{k\ netgirdi}$ değerinin aktivasyon fonksiyonundan geçirilerek çıktıyı oluşturmaktadır. W gizli nöronlar için eşik değerini, V ise çıktı nöronları için eşik değerini temsil etmektedir. λ, momentum sayısını; η, öğrenme katsayısını; δ, ağırlık düzeltme sayısını göstermektedir. Parametreler ile ilgili açıklamalar Çizelge 4.4.'te verilmiştir. Çok katmanlı yapay sinir ağlarında 2 tane durdurma kriteri vardır ya belli iterasyonlardan sonra ya da belli bir hata düzeyine inildikten sonra öğrenme durdurulur.

Çizelge 4.4.: Çok Katmanlı YSA Parametreleri.

Parametre	Açıklaması	Parametre	Açıklaması
X_i	i. Girdi	$f'(O_{k\ netgirdi})$	Aktivasyon Fonksiyonunun Türevi
W_{0j}	Gizli Katmandaki Nöronlar için Eşik Değeri	$\delta_k^{çıktı}$	k. Çıktı Nöronu için Hata Bilgi Değeri
W_{ij}	Girdi Katmanındaki i. Nöron ile Gizli Katmandaki j. Nöron Arasındaki Ağırlık Değeri (i=1,…,n)	η	Öğrenme Katsayısı
$Z_{j\ netgirdi}$	Gizli Katmandaki j. Nörona Gelen Net Girdi	λ	Momentum Katsayısı
Z_j	Gizli Katmandaki j. Nörona Gelen Net Girdinin Aktivasyon Fonksiyonundan Geçmiş Hali	ΔV_{jk}^{yeni}	Çıktı Katmanı ile Gizli Katman Arasındaki Eşik Değeri ve Bağlantı Ağırlıklarının Değişim Miktarı
V_{0k}	Çıktı Katmandaki Nöronlar için Eşik Değeri	$\delta_{j\ net}^{gizli}$	j. Gizli Nöron için Hata Bilgi Değeri
V_{jk}	Gizli Katmanındaki j. Nöron ile Çıktı Katmandaki k. Nöron Arasındaki Ağırlık Değeri	ΔW_{jk}^{yeni}	Gizli Katman ile Girdi Katmanı Arasındaki Eşik Değeri ve Bağlantı Ağırlıklarının Değişim Miktarı
$O_{k\ netgirdi}$	Çıktı Katmandaki k. Nörona Gelen Net Girdi	V_{jk}^{yeni}	Çıktı Katmanı ile Gizli Katman Arasındaki Eşik Değeri ve Bağlantı Ağırlıklarının Yeni Değeri
O_k	Çıktı Katmandaki k. Nörona Gelen Net Girdinin Aktivasyon Fonksiyonundan Geçmiş Hali	W_{jk}^{yeni}	Gizli Katman ile Girdi Katmanı Arasındaki Eşik Değeri ve Bağlantı Ağırlıklarının Yeni Değeri
T_k	Çıktı Katmanındaki k. Nörondan Beklenen Çıtı Değeri		

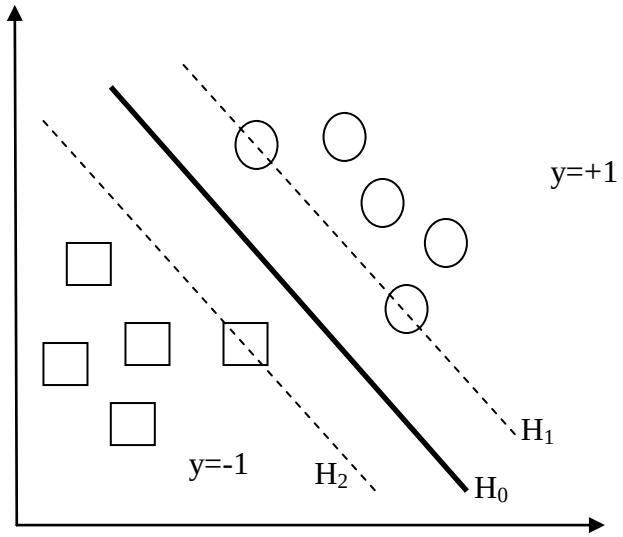
4.1.4. Destek Vektör Makineleri

Doğrusal ve doğrusal olmayan problemlerin sınıflandırılmasında kullanılan destek vektör makineleri istatistik tabanlı bir yöntemdir. Vapnik tarafından geliştirilen bu yöntem yapısal risk küçültme esasına dayanmaktadır [Vapnik, 1995]. Yapısal risk küçültme, iki sınıfı birbirinden ayırırken eğitim seti üzerindeki hataları

minimuma indirmeyi amaçlamaktadır [De Vel, 2000]. Destek vektör makineleri ile amaçlanan eğitim setindeki iki sınıfı birbirinden en iyi ayıran üstündüzlemi ve eğitim seti üzerindeki destek vektörlerini belirlemektir [Teng et al, 2004].

Destek vektör makineleri ilk olarak iki sınıflı doğrusal olarak ayrılabilen örneklerin sınıflandırılmasında kullanılırken; günümüzdeki gerçek problemler daha fazla sınıftan oluşmakta ve doğrusal olarak da ayrılamamaktadır. Bu nedenle zamanla doğrusal olmayan problemlerin sınıflara ayrılması için destek vektör makineleri kullanılmaya başlanmıştır.

Doğrusal olarak ayrılabilen sınıflarda T veri kümesinin (x_1, y_1), (x_2, y_2), …, (x_n, y_n) şeklinde ifade edildiği varsayılsın. i=1, …, n oluncaya kadar $y_i \in \{+1,-1\}$ şeklinde x_i değerlerinin sınıf etiketini tutmaktadır [Özkan, 2008]. Doğrusal olarak iki sınıfa ayrılabilen x_i değerleri ve bunların iki sınıfa doğrusal olarak ayrıldığını gösteren üstündüzlem Şekil 4.13.'te gösterilmiştir.



Şekil 4.13.: Doğrusal Ayrılabilen Veriler.

T veri kümesini ayırmada pek çok hiper düzlem kullanılabilir ancak burada aranan iki sınıfı birbirinden ayıran en büyük boşluklu hiper düzlemleri bulmaktır. H_1 ve H_2 bu iki hiper düzleme karşılık gelmektedir. H_0 ise bu iki hiper düzlem arasında yer alan üstündüzlem olmaktadır. H_1 ve H_2 üzerindeki noktalar da destek vektörleri

olarak adlandırılmaktadır. Hiper düzlemler ve üstündüzlem için yazılan denklemler 4.12., 4.13. ve 4.14.'te gösterilmiştir.

$$w \cdot x_i + b \geq +1 \quad her\ y_i = +1 \tag{4.12.}$$

$$w \cdot x_i + b \leq -1 \quad her\ y_i = -1 \tag{4.13.}$$

$$w \cdot x + b = 0 \tag{4.14.}$$

Üstündüzlem sınırının maksimum olması için ||w|| değerinin minimum hale gelmesi gerekir, minimum hale getirmek için 4.15.'ten yararlanılır ve 4.16. sonuç olarak elde edilir.

$$\frac{1}{2}||w||^2 \tag{4.15.}$$

$$y_i(w \cdot x_i + b) \geq 1 \tag{4.16.}$$

Bu ifade Lagrange Çarpanları kullanılarak çözümlenirse 4.17. elde edilir. Bu ifadeden yola çıkarak elde edilen karar fonksiyonu da Eşitlik 4.18.'de verilmiştir.

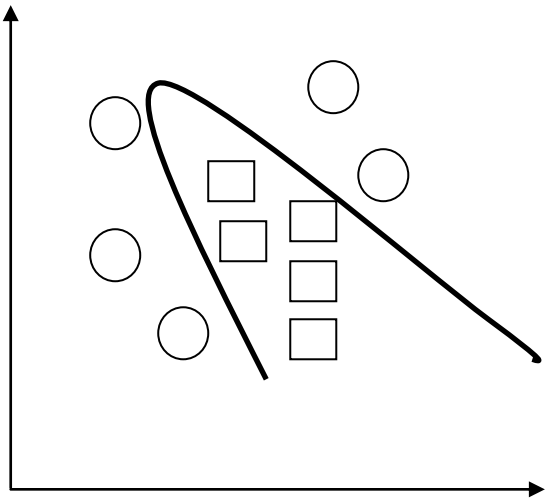
$$(L, w, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{k} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{k} \alpha_i \tag{4.17.}$$

$$f(x) = sign\left(\sum_{i=1}^{k} \alpha_i y_i (x x_i) + b\right) \tag{4.18.}$$

Eldeki veri kümesi Şekil 4.14'teki gibi doğrusal olarak ayrılamayabilir. Bu durumda en az hata yapan düzlemi bulmak gerekir. Bu durumda doğrusal olmayan problemlerin çözümünde mevcut denklemlere yeni değişkenlerin eklenmesi gerekmektedir. $\xi^t \geq 0$ koşulunu sağlayan ve hataları ifade eden gevşek değişkenler eldeki denkleme eklenerek doğrusal olmayan problemlerin çözümü sağlanır (4.19., 4.20.).

$$w \cdot x_i + b \geq +1 - \xi^t \quad her\ y_i = +1 \tag{4.19.}$$

$$w \cdot x_i + b \geq -1 - \xi^t \quad her\ y_i = -1 \tag{4.20.}$$



Şekil 4.14.: Doğrusal Ayrılamayan Veriler.

Doğrusal olmayan problemlerin sınıflandırılmasındaki çözüm veriyi daha büyük boyutlu bir uzaya taşımaktır. Veri daha büyük boyutlu uzaya taşındıktan sonra çekirdek fonksiyonları kullanarak karar fonksiyonu elde edilir. Kullanılan bu çekirdek fonksiyonları doğrusal (4.21.), polinom (4.22.), radyal tabanlı (4.23.) olmak üzere üçe ayrılmaktadır [Özkan, 2008].

$$K(x_i, x_j) = x_i^T x_j \tag{4.21.}$$

$$K(x_i, x_j, c, d) = (c + x_i^T x_j)^d \tag{4.22.}$$

$$K(x_i, x_j, \sigma) = e^{\frac{||x_i - x_j||}{2\sigma^2}} \tag{4.23.}$$

Destek vektör makineleri yüksek boyutlu veri ile çalışırken oldukça başarılı sonuçlar ürettiği için metin sınıflandırma ve yazar belirleme uygulamalarında sıklıkla tercih edilmektedir. Destek vektör makineleri içerisinde SMO algoritması sıklıkla kullanılmaktadır.

4.1.4.1. SMO Algoritması

Standart destek vektör makinesi algoritmasının gelişmiş bir versiyonu olan SMO, gerçekleştirilmesi oldukça basit ve oldukça hızlı sınıflandırma yapması nedeniyle gerçek zamanlı problemlerin çözümünde sıklıkla kullanılmaktadır. [Keerthi et al, 2001].

Platt tarafından geliştirilen SMO, optimizasyon problemi türü olan quadratik programlamada etkin bir çözüm sunmaktadır. Quadratik problemlerin çözümünü bu problemleri alt quadratik problemlere bölerek sağlayan SMO büyük boyuttaki veri setleri üzerinde de başarılı bir sınıflandırma sağlamaktadır [Platt, 1999]. SMO algoritmasında amaç seçilen iki Lagrange çarpanlarını her defasında yeni çarpanlar ile değiştirerek en uygun çözümü sağlamaktır.

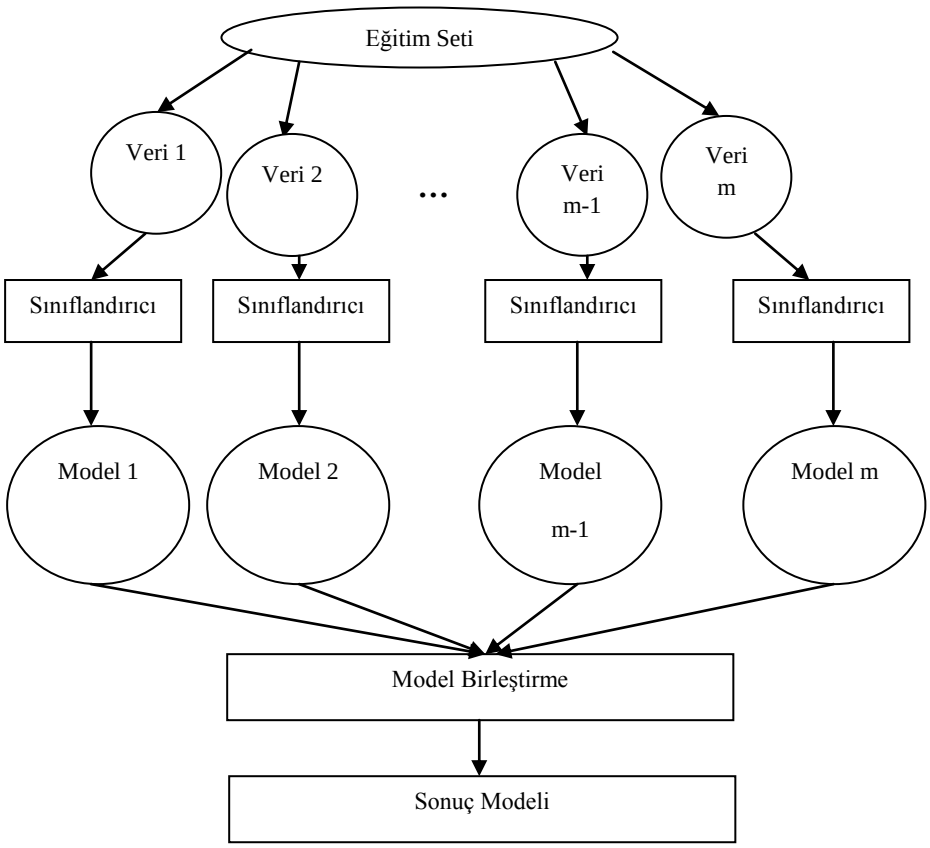
4.1.5. Gruplamalı Sınıflandırıcılar

Bir filmi izlemeden önce o filmi izlemiş olanların fikirlerini dikkate almak, bir ürünü almadan önce o ürünü kullanmış olanların yorumlarını göz önünde bulundurmak, "Kim Milyoner Olmak İster?" adlı bilgi yarışmasında seyirci jokerini kullanıp çoğunluğun fikrini dikkate alarak karar vermek ve benzeri şekilde sıralanabilecek pek çok örnek 1979 yılında ortaya çıkan Gruplamalı Sınıflandırıcıların altında yatan temel mantığı en iyi şekilde açıklamaktadır.

Tek bir sınıflandırıcının tüm problem uzayını temsil edemeyeceği fikri ile ortaya çıkan [Dasarathy and Sheela, 1979] Gruplamalı Sınıflandırıcılarda, sonuç modeli oluşturulurken, oluşturulan bu modelin tek bir sınıflandırıcıya -ki bu zayıf bir sınıflandırıcı olarak kabul ediliyor- bağımlı olması yerine birbirinden bağımsız iki veya daha fazla sınıflandırıcıdan -güçlü bir sınıflandırıcı- çoğunluk oylaması ile belirlenmesi amaçlanmaktadır. Burada bahsi geçen ve Gruplamalı Sınıflandırıcılar için önemli bir kavram olan bağımsızlık ile anlatılmak istenen veri kümesi üzerinden üretilen hataların her sınıflandırıcı için çeşitlilik göstermesidir. Böylece herhangi bir sınıflandırıcının yaptığı hatadan tüm sistem etkilenmez, sistem bu hatayı telafi

edebilir. Böylece risk azalır, iyi bir genelleme, daha iyi bir sınıflandırma ve daha iyi bir başarı sağlanmış olur [Polikar, 2006].

Bahsedilen bu özellikleri ile her geçen gün daha da cazip hale gelen ve üzerine pek çok uygulama geliştirilen Gruplamalı Sınıflandırıcılar zaman içerisinde araştırmacıların ilgi alanına daha çok girmiş ve yeni gelişmeleri de peşinden getirmiştir. Bunların başında da eldeki veri kümesinin farklı versiyonlarına aynı sınıflandırıcının uygulanması gelmektedir (Şekil 4.15.). Gruplamalı Sınıflandırıcıların bu versiyonlarında veri kümesi üzerinden tek bir hipotez (model) öğrenmek yerine hipotezlerin bir kümesi öğrenilmektedir [Wang et al, 2011]. Çalışmada kullanılan Bagging ve AdaBoostM1 algoritmaları da bu mantık ile sonuç modellerini üretmektedir.



Şekil 4.15.: Gruplamalı Sınıflandırıcılar.

4.1.5.1. Bagging Algoritması

Eğitim setinin yeniden örneklenmesi mantığı ile çalışan Bagging (Bootstrap Aggregating), 1996 yılında Breinman tarafından geliştirilen gerçeklenmesi kolay, performansı güçlü, varyans ve bias düşümü sağlayan bir sınıflandırma algoritması olarak uygulamalarda yaygın bir şekilde kullanılmaktadır [Liang et al, 2011].

Temel çalışma mantığı eldeki eğitim setinin yeniden örneklenmesi olan Bagging, bunu eğitim seti içerisinden seçilen örneklerin eğitim setiyle aynı veya farklı boyutlardaki değişik eğitim setlerini (bootstrap) oluşturması ile sağlamakta, oluşturulan bu eğitim setlerinde aynı örneklerin birden çok tekrarı ise kaçınılmaz bir hal almaktadır. Bu temel mantık çerçevesinde farklı eğitim setleri ile çalışarak tek bir eğitim setine dayalı sonuç üretilmesi engellenmekte bu da varyans düşümünü sağlamaktadır.

Eğitim setinden yeni eğitim setlerinin oluşturulması adımından sonra ise sıra oluşturulan eğitim setlerine seçilen bir temel sınıflandırıcının - ki bu sınıflandırıcı zayıf bir sınıflandırıcıdır - (YSA, DVM gibi) paralel şekilde uygulanması adımına gelmektedir. Burada dikkat edilmesi gereken nokta ise seçilen temel sınıflandırıcının sabit olmamasıdır. Sabit olmamak ile kastedilen, eğitim setinde yapılan ufak bir değişikliğin sınıflandırıcının ürettiği modelde büyük değişikliklere yol açmasıdır [Breiman, 1996]. Oluşturulan bu farklı modellerin çoğunluk oylaması ile oluşan sonuç model ise iyi bir genelleme ve yüksek performans ile bias düşümünü sağlamaktadır (Şekil 4.16.).

Girişler:

- n tane sınıf etiketi bulunan eğitim seti, E
- Kullanılacak temel sınıflandırıcı, T
- İterasyon Sayısı, I
- Bootstraplar E'nin yüzde kaçı, Y

For i=1 to I

a) E'nin yüzde Y lik kısmına karşılık gelen bootstrap alınsın, E_i
b) Temel sınıflandırıcı olan T E_i üzerinden model oluştursun, m_i
c) Oluşturulan model m_i Gruplamaya eklensin, G

Test:

a) Sınıf etiketi belirlenmek istenen örnek alınsın, o
b) G içerisindeki tüm m ler o örneği üzerinde denensin
c) En çok oy alan sınıfa göre o'nun sınıf etiketini belirle (çoğunluk oylaması)

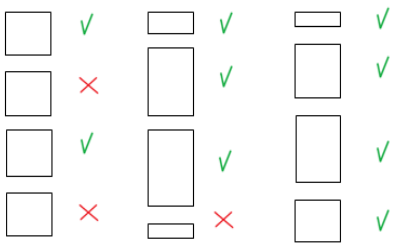
Şekil 4.16.: Bagging Algoritması [Polikar, 2006].

4.1.5.1. AdaBoostM1 Algoritması

İlk olarak 1997 yılında tanıtılan ve sadece iki sınıf etiketine sahip eğitim setleri üzerinden model kuran ve ağırlık güncellemesi esasına dayanan AdaBoost algoritması, ikiden fazla sınıf etiketine sahip gerçek hayat problemlerini sınıflandırmada yetersiz kalmış ve bir arayış içerisine girilmiştir. Bu arayış sonunda AdaBoostun gelişmiş bir versiyonu olan AdaBoostM1 algoritması ortaya çıkmıştır.

AdaBoostM1 algoritması, eğitim seti içerisindeki örneklerin ağırlıklarının belirli iterasyon sayısı kadar güncellenmesi esasına dayalı güçlü bir Gruplamalı Sınıflandırıcı olarak parmak izi sınıflandırmadan [Liu, 2010], mermer yüzeylerdeki görüntülerin sınıflandırılmasına [Doğan and Akay, 2010] kadar pek çok alanda kullanılmaktadır. Burada ağırlık ile kastedilmek istenen ise eğitim seti içerisindeki örneklerin ne kadar önemli olduğudur. Başlangıçta her örnek için tek bir ağırlık değeri belirlenir ve temel sınıflandırıcı model bu ağırlık üzerinden oluşturur. Oluşturulan model sonucunda bazı örnekler doğru sınıflandırılırken bazıları da yanlış sınıflandırılabilir. AdaBoostM1 algoritması için yanlış sınıflandırılan ve öğrenilmesi güç olan bu örnekler önem teşkil etmektedir. Şekil 4.17.'de gösterildiği gibi bir örnek doğru bir şekilde sınıflandırıldıysa o örneğin ağırlığı düşürülür, yanlış sınıflandırıldıysa ise de ağırlığında artış meydana gelir. Bir sonraki iterasyonda temel

sınıflandırıcı artık yeni ağırlık değerleri ile birlikte yeni bir model oluşturur ve bu işlem belli bir iterasyon sayısına kadar seri bir şekilde devam eder (Şekil 4.18.). Sonuç model eldeki modellerin ağırlıklı çoğunluk oylaması ile belirlenir.



Şekil 4.17.: AdaBoostM1'de Ağırlık Güncellemesi.

Girişler:

- n tane sınıf etiketi bulunan eğitim seti, E=[(x_j, y_j)], j=1, …, n
- Kullanılacak temel sınıflandırıcı, T
- İterasyon Sayısı, I

Başlat: $w_j = \frac{1}{n}$, j=1, …, n (w_j: j. örneğin ağırlığı)

For i=1 to I

a) Temel sınıflandırıcı olan T, güncel ağırlıkları kullanarak E_i üzerinden model oluştursun, m_i

b) Oluşturulan model m_i'nin hatası, ε_i, belirlensin

$$m_i: \varepsilon_i = \sum_{j=1}^{n} w_j^i, \; y_j \neq m_i(x_j)$$

c) m_i nin ağırlığı belirlensin

$$m_i: \alpha_i = \frac{1}{2} \ln\left(\frac{1-\varepsilon i}{\varepsilon i}\right)$$

d) E içerisinde doğru sınıflandırılan örneklerin ağırlıkları güncellensin

$$w_j = w_j e^{-\alpha i}$$

e) E içerisinde yanlış sınıflandırılan örneklerin ağırlıkları güncellensin

$$w_j = w_j e^{\alpha i}$$

f) Ağırlıklar, w_j, normalize edilsin

Test:

d) Sınıf etiketi belirlenmek istenen örnek alınsın, o

e) G içerisindeki tüm m ler o örneği üzerinde denensin

f) En çok oy alan sınıfa göre o'nun sınıf etiketini belirle (çoğunluk oylaması)

Şekil 4.18.: AdaBoostM1 Algoritması [Polikar, 2006].

5. YAZAR BELİRLEME

19. yüzyıldan itibaren üzerine çalışmalar yapılan yazar belirleme; yazarı bilinmeyen bir dokümanın yazarını, yazarı bilinen dokümanlar üzerinden belirlemeyi amaçlamaktadır. Yazar belirleme hiyerarşik yapı içerisinde üçüncü sırada yer alırken bir üst seviyesinde yazar analizi, yazar analizinin üzerinde de metin madenciliği yer almaktadır.

Her geçen gün artan veri miktarı bu verileri yönetmeyi ve içerisinden önemli olan ancak keşfedilmemiş bilgiyi çıkarmayı gerekli hale getirmiş ve metin madenciliği kavramı ortaya çıkmıştır. Doğal dilde yazılmış olan metinlerden belirli bir amaç dahilinde yararlı bilginin çıkarılması ile ilgilenen metin madenciliği doğal dil işleme ve veri madenciliği disiplinlerinin sentezinden oluşmaktadır [Witten, 2005].

Metin madenciliği, eldeki metinler üzerinden belli adımların gerçekleşmesini gerektirmektedir ve doğal dil işleme ile veri madenciliği de bu adımlar içerisinde yer almaktadır. Metin madenciliği yapılabilmesi için ilk olarak metin koleksiyonlarının oluşturulması gerekmektedir. Günümüzde internetin de varlığı ile artık metin koleksiyonlarının oluşturulması ve depolanması da oldukça kolay bir şekilde gerçekleşmektedir. Metinler elde edildikten sonra üzerinde gerekli önişleme adımlarının yapılması gerekmektedir. Önişleme ile amaçlanan eldeki metni yine eldeki metin içinde gizli kalmış bilgiyi çıkarabilecek hale dönüştürmektir. Metinler üzerinde yapılacak önişleme çalışılacak amaca göre farklılıklar göstermekle birlikte temel önişleme adımları; noktalama işaretleri, sayı ve özel karakterlerin eldeki metinlerden çıkartılması, büyük harflerin küçük harflere dönüştürülmesi, metni meydana getiren ve herhangi bir anlamı olmayan fonksiyonel kelimelerin eldeki metinlerden ayıklanması, yazım hatalarının düzeltilmesi, kelimelerin kök ve gövdelerinin belirlenmesi şeklinde sıralanmaktadır. Bu önişleme aşamasında doğal dil işleme devreye girmektedir. Doğal dil işleme; doğal dil üzerine inceleme, çözümleme, yorumlama, bilgi çıkarma, üretme yapan bilgisayar sitemi şeklinde tanımlanmaktadır [Oğuzlar, 2011]. Önişleme adımından sonra çalışılacak alana ve istenilen bilgiye yönelik olarak metinsel özelliklerin eldeki metinlerden çıkartılması

aşamasına geçilmektedir. Eldeki metinlerden çıkartılacak metinsel özellikler karakter tabanlı, kelime tabanlı, söz dizilimsel, anlamsal ve uygulamaya özel özellikler olmak üzere beş başlık altında toplanabilmektedir (Çizelge5.1.) [Stamatatos, 2009].

Çizelge 5.1.: Metinsel Özellikler.

Metinsel Özellikler	Türleri
Karakter Tabanlı Özellikler	Harf Sayısı Noktalama İşareti Sayısı Rakam Sayısı Büyük/Küçük Harf Sayısı Özel Karakterlerin Sayısı Karakter n-gramlar
Kelime Tabanlı Özellikler	Kelime Sayısı Ortalama Kelime Uzunluğu Kelime Zenginliği Cümle Sayısı Ortalama Cümle Uzunluğu Kelime n-gramlar Bir/İki Kere Geçen Kelimeler İkilemeler Kısaltmalar
Söz Dizilimsel Özellikler	Cümle ve Söz Öbeklerinin Kullanımı Kelime Türleri Devrik Cümle
Anlamsal Özellikler	Kelimelerin Anlamsal Bağlantıları Fonksiyonel Kelimeler
Uygulamaya Özel Özellikler	Yazım Hataları Format Hataları Noktalama Hataları Mimik Kullanımı Selamlama Kullanımı İmza Kullanımı

Metin dediğimizde aklımıza ilk gelen karakterler, karakterlerin birleşmesi ile oluşan kelimeler ve kelimelerin de birleşmesi ile oluşan cümleler gelmektedir. Karakter tabanlı metinsel özellikler; metni meydana getiren harf, rakam, noktalama ve özel karakterler ve bunlar ile ilgili çıkarılabilecek her türlü özelliklerdir. Karakter tabanlı özellikler dilden bağımsız şekilde kolayca eldeki metinden çıkartılabilmektedir. Bir dildeki karakterlerin kümesi bellidir ve bu karakterlerin sayısı ile ilgili özellikler rahatlıkla çıkartılabilir. Karakter tabanlı metinsel özelliklerde gözümüze çarpan karakter n-gramlardır. Buradaki n bir sayıya karşılık gelmektedir mesela 3-gramlar gibi. Elimizde Şekil 5.1.'deki gibi bir cümle olsun buradan 3-gramları çıkararak aşağıdaki sonucu elde ederiz.

Örnek Cümle: Hava sıcak. Cümleden elde edilen 3-gramlar: Hav, ava, va_, a_s, _sı, sıc, ıca, cak şeklindedir.

Şekil 5.1.: 3-gramlar.

Kelime tabanlı metinsel özellikler; kelimeler ve cümleler ile ilgili sayısal özellikleri göstermektedir. Bir metindeki toplam kelime ve cümle sayısı bunların başında gelmektedir. Ortalama kelime uzunluğu, harf ve rakamların toplam sayısının toplam kelime sayısına oranı iken kelime sayısının cümle sayısına oranı da ortalama cümle uzunluğu olarak hesaplanmaktadır. Kelime zenginliği, bir metindeki farklı kelimelerin sayısının tüm kelimelerin sayısına oranı olarak tanımlanmaktadır. Metindeki farklı kelimelerin sayısının belirlenmesinde doğal dil işlemeden yararlanılarak kelime kök ve gövdesine indirgenmektedir. Kelime n-gramlar aynı karakter n-gramlar gibi çalışmaktadır. n sayısına göre kelimeler sırayla ele alınmaktadır. Hem karakter hem de kelime n-gramlarda eldeki n-gramın metindeki toplam sayısı yani frekansı özellik olarak kullanılmaktadır. Yine doğal dil işlemenin kullanılması ile bir kere ve iki kere geçen kelimelerin sayısı hesaplanmaktadır. İkilemeler ve kısaltmalar da sıklıkla metinsel özellik olarak kullanılmaktadır.

Söz dizilimsel metinsel özellikler içerisinde; cümle ve söz öbeklerinin kullanımı, kelime türleri ve devrik cümle kullanımı bulunmaktadır. Cümle ve söz öbekleri incelenirken Sözcük Türü (POS/Part-of-Speech) etiketlerinden yararlanılmaktadır. Bir cümleyi ya da söz öbeğini meydana getiren kelimeleri cümledeki durumlarına göre isim, sıfat, fiil vb. şekillerde etiketleyen Sözcük Türü etiketleri kelimelerin morfolojik analizini gerçekleştirmektedir [Oğuzlar, 2011]. Aynı şekilde kelime türlerini belirlerken de doğal dil işleme ve Sözcük Türü etiketlerinden yararlanılmaktadır. Devrik cümleleri oluşturan kelimelerin normal sıralanması yerine farklı şekilde sıralanması ile oluşan bir durumdur. Söz dizilimsel metinsel özellikler yazarlar ile ilgili önemli ipuçları vermektedir.

Anlamsal metinsel özellikler adından da anlaşılacağı üzere kelimeler arasındaki anlamsal bağlantıyı belirleyen özelliklerdir. Metindeki kelimelerin eş anlamlıları, alt ve üst sınıfları gibi bilgiler anlamsal özelliklerin kullanımı ile elde edilmektedir. Fonksiyonel kelimeler de bir metni meydana getirirken olmazsa olmazlar arasında

yer almaktadır. Çalışılacak alana göre fonksiyonel kelimelerin metinsel özellik olarak kullanılıp kullanılmayacağına karar verilir.

Uygulamaya özel metinsel özellikler ile kast edilen ise elektronik ortamda bulunan metinlerin edebi metinlerden farklı olarak bulundurduğu özelliklerdir. Elektronik ortamdaki metinlerde (elektronik posta, forum mesajları, haber grubu mesajları vs.) en sık hatalar ile karşılaşılmaktadır. Bunlar yazım, format ve noktalama hatalarıdır. Yazım hataları metni meydana getiren kelimelerin yazım yanlışlığıdır, format hatası büyük ve küçük harflerin yanlış kullanımı, noktalama hatası ise noktalama işaretlerinin yanlış ve eksik kullanımından kaynaklanan hatalardır. Bu tür ortamdaki metinlerde mimik kullanımına da sıklıkla rastlanmaktadır. Mimik, yazarın duygularını ifade etmede kullandığı noktalama işaretlerinin kombinasyonundan oluşmaktadır (☺, ☹) [Aydın vd., 2004]. Elektronik ortamdaki metinlerde selamlama ve imza da sıklıkla kullanılmaktadır. Selamlama, yazarın alıcıya söylediği "merhaba, selam, iyi akşamlar" gibi hitaplara karşılık gelmekteyken imza ise yazarın, metni tamamlarken kullandığı ismi, lakabı ya da iletişim bilgileri şeklinde tanımlanmaktadır.

Metin madenciliğindeki tüm bu aşamalar gerçekleştirildikten sonra çıkarılan metinsel özelliklerden bilginin çıkarılmasına aşamasına geçilmekte ve bu aşamada veri madenciliği kullanılmaktadır. Veri madenciliği ile bilgi çıkartıldıktan sonra sonuçlar değerlendirilerek başarı belirlenmektedir.

Metin madenciliği; metin sınıflandırma, metin kümeleme, metin özetleme, otomatik çeviri ve yazar analizi olmak üzere beş görevden oluşmaktadır. Metin sınıflandırma; eldeki metni, metinsel özellikleri çıkartıp önceden tanımlı sınıflardan birine dahil etmek şeklinde tanımlanmaktadır [Yıldız vd., 2007]. Metin kümeleme, öğretmensiz öğrenme yöntemlerini kullanarak benzer metinsel özellikleri gösteren metinleri aynı grup altında toplamayı amaçlamaktadır. Metin özetleme, eldeki metinden ana fikrin çıkartılmasını amaçlamaktadır [Chong and Chen, 2009]. Otomatik çeviri, diller arasında bilgisayar aracılığı ile çeviri yapılmasıdır, benzer aileye mensup olan diller arasında çeviri daha kolay gerçekleşmektedir [Orhun ve ark, 2011]. Yazar analizi ise yazara ait metinlerden çıkarılan ve yazarı belirlemeye

yarayan metinsel özelliklerin (kelime sayısı, cümle sayısı, fonksiyonel kelimeler ve yazım hataları gibi) kullanımı ile eldeki metnin yazarına ait niteleyici bir sonuca varmayı amaçlamaktadır [Abbasi, 2005]. Yazar analizi; yazarı karakterize etme, benzerlik tespiti ve yazar belirleme şeklinde üçe ayrılmaktadır.

Yazarı karakterize; etme eldeki metinden çıkarılan metinsel özelliklerin kullanımı ile yazarın yaş, cinsiyet, eğitim durumu gibi bilgileri hakkında bir çıkarım yapmayı amaçlamaktadır. 2002 yılında yapılan bir çalışmada eldeki elektronik postalarda sorunları çözmede bayanların daha duygusal bir ifade kullanmaları erkeklerin de bağımsızlıkları ve baskınlıkları ile ön plana çıktıkları gözlemlenmiştir [Corney et al, 2002]. Yine 2002 yılında türe ve cinsiyete göre bir sınıflandırma yapılmış ve cinsiyete göre ayrımda %80 başarı elde edilmiştir [Koppel et al, 2002]. Benzerlik tespiti; iki metin arasındaki benzerliğin tespitinde metinsel özelliklerin kullanılması ile gerçekleştirilmektedir. Özellikle intihal tespitinde bu yöntem sıklıkla kullanılmaktadır. Shakespeare'nin yazdığı eserler ile ilgili her zaman bir kuşku oluşmuş ve Shakespeare'nin intihal yaptığı ile ilgili görüşler ortaya atılmıştır. 1991 yılında yapılan bir çalışmada Shakespeare'in Edward de Vere'den çalıntı yapmadığı ikisi arasındaki farklılıkların ortaya koyulması ile ispat edilmiştir ve bu amaçla kelime ve cümle uzunlukları gibi metinsel özellikler kullanılmıştır [Elliot and Valenza, 1991]. İntihal tespiti için yapılan bir çalışmada ise paragraflar arası kelime zenginliği, ortalama kelime frekansı metinsel özellik olarak kullanılırken, veri madenciliği yöntemlerinden discriminant analizi ve destek vektör makineleri kullanılmıştır [Eissen at al, 2007].

Üzerine en çok araştırma yapılan ve bu çalışmanın da konusu olan yazar belirlemedeki yeni gelen bir metnin yazarının mevcut yazarlardan biri olup olmadığını tespit etmede metinsel özelliklerin ve veri madenciliğindeki sınıflandırma yöntemlerinin kullanılması olarak tanımlanmaktadır. Yazar belirlemedeki temel fikir; edinilmiş davranışların sürekli tekrarlanması şeklinde tanımlanan alışkanlıkların, yazarların ürettikleri metinlerde de kendini göstermesidir. Edinilen yazma alışkanlıkları, yazarların metinleri oluştururken aynı metinsel özellikleri aynı biçimde sürekli kullanması şeklinde ortaya çıkmaktadır. Sürekli tekrarlanan bu özellikler eldeki metnin yazarını belirlemede önemli bir ipucu oluştururken, geçmişten

günümüze yazar belirleme üzerine yapılan çalışmalar da göstermiştir ki her yazar "metinsel bir parmak izine" sahiptir [Baayen et al, 2002].

Yazar belirlemenin tarihçesi 19. Yüzyılda Mendenhall'ın yaptığı çalışamaya dayanmaktadır. Mendenhall edebi metinler üzerine yaptığı çalışmada kelimelerdeki karakter sayısını esas alan bir yaklaşım izlemiştir. Bir örnekle açıklamak gerekirse bir metindeki bir harfli, iki harfli kelime frekanslarına göre karakteristik eğriler oluşturmuş ve buna göre bir sınıflandırma yapmıştır [Mendenhall, 1887]. Bu yaklaşımın günümüzde pek bir geçerliliği bulunmamaktadır. Mendenhall'ın yaptığı bu çalışmadan sonra yazar belirleme üzerine yapılan çalışmalar 1932 yılına kadar bir duraklama evresine girmiştir. 1932 yılında Zipf yazar belirleme üzerine istatistiksel bir çalışma gerçekleştirmiştir [Stamatatos, 2009]. Bu çalışmayı 1939 ve 1944 yılında Yule'un yaptığı istatistiksel çalışmalar izlemiştir. 1939 yılında yapılan çalışmada bir cümleyi oluşturan toplam kelime sayısına göre belirlenen cümle uzunluğu metinsel özellik olarak kullanılmış ve istatistiksel yöntemlerin kullanımı ile benzerlik tespiti yapılmıştır [Yule, 1939]. Bu çalışmanın amacı cümle uzunluğunun bir yazarı ne kadar tasvir gücüne sahip olduğunu belirlemektir. Yule'un 1944 yılında yaptığı çalışmada Poisson dağılımı eldeki metinlerdeki isimlerin frekansına uygulanarak yazar belirleme gerçekleştirilmiştir [Yule, 1944].

Yazar belirlemede kilometre taşı olarak kabul edilen çalışma ise 1963 yılında Mosteller ve Wallace tarafından gerçekleştirilmiştir. Federalist Paper'larda yazarı belli olmayan 77 tane metnin yazarını belirlemek için yapılan bu çalışmada metinsel özellikler olarak fonksiyonel kelimeler, veri madenciliği yöntemi olarak da Bayes Teoremi kullanılmıştır ve 12 tanesinin yazarının Madison olduğu anlaşılmıştır [Mosteller and Wallace, 1963]. Bu çalışmadan sonra fonksiyonel kelimeler pek çok çalışmada vazgeçilmez metinsel özellik oldu [Koppel et al, 2009]. 1985 yılında Holmes tarafından yapılan çalışmada pek çok metinsel özellik eldeki veri kümesi için değerlendirilmiş ve bir kere, iki kere ya da daha fazla geçen kelime frekanslarının konudan bağımsız olan metinlerde avantajlı olduğu sonucuna varılmıştır [Holmes, 1985]. 1998 yılında yapılan çalışmada ise Morton'un geliştirdiği CUSUM tekniği ele alınmıştır. CUSUM tekniğinde cümle uzunluğu dikkate alınarak grafik çizilmiş ve bu

grafiğe göre değerlendirme yapılmıştır [Holmes, 1998]. Adli bilimlerde sıkça kullanılmış olan bu yöntem pek de başarılı bulunmamıştır.

Günlük bir gazeteden alınan Yunanca metinler üzerine yapılan bir çalışmada 10 yazar, her yazar için 20 metin ve 22 metinsel özellik seçilmiştir. Seçilen 22 tane metinsel özellik PCA ile işlenmiş ve 100 kelimeden az olan metinlerde yazar belirlemenin başarısız olduğu sonucuna varılmıştır [Stamatatos et al, 1999]. Almanca için yapılan çalışmada 8 yazar üç konuda üçer makale yazmış ve çıkarılan metinsel özellikler PCA ve Discriminant Analysis ile işlenmiştir [Baayen, 2002]. Günlük bir gazeteden 40 yazar için alınan metinlerden 39 tane metinsel özellik seçilip çıkartılmış ve çıkartılan özellikler ki-kare testi ile işlenmiştir. En başarılı sonuçlar (%90) 5 yazar için kelime, noktalama işareti ve 2-gram metinsel özelliklerinin birleşimi ile sağlanmıştır [Grieve, 2007]. 2011 yılında yapılan çalışmada, eski çalışmaların aksine yeni gelen metnin mevcut yazarlardan birine ait olma zorunluluğunun bulunmayışıdır. Bu çalışmada çıkarılan metinsel özellikler üzerine kosinüs benzerliği uygulanarak yazar belirleme gerçekleşmiştir [Koppel et al, 2011]. Türkçe için 2003 yılında yapılan çalışmada 18 yazar için on beşer tane metin veri kümesi olarak seçilmiş ve veri kümesinden çıkartılan 22 tane metinsel özellik çıkartılmıştır. Üç farklı sınıflandırıcı ile farklı sayılardaki metinsel özellikler işlenmiş ve ortalama %84 başarı elde edilmiştir [Diri and Amasyalı, 2003]. 2006 yılında yapılan çalışmada yine 18 yazar için toplamda 630 metin veri kümesi olarak günlük bir gazeteden seçilmiştir. Her yazar için on tane metinsel özellik Naive Bayes, destek vektör makineleri, C4.5 ve bir karar ağacı olan Random Forest ile işlenmiş ve destek vektör makineleri ile en başarılı sonuçlar elde edilmiştir [Diri ve vd., 2006]. 2007 yılında Taş ve Görür yaptıkları çalışmada 20 yazar için 35 tane metinsel özellik çıkartmıştır. Daha sonra seçilen bu özelliklerden 22 tanesi 15 tane veri madenciliği yöntemi ile işlenerek Naive Bayes Multinomial ile %80 başarı elde edilmiştir [Taş and Görür, 2007].

Elektronik ortamdaki metinlerin (elektronik posta, haber grubu mesajları, forum mesajları gibi) artması ile bu ortamdaki metinler için de yazar belirlemenin gereksinimi ortaya çıkmıştır. Elektronik ortamdaki metinler üzerine yapılan ilk çalışma 2000 yılında De Vel tarafından gerçekleştirilmiştir. Bu çalışmada 5 yazar

için seçilen elektronik postalar üzerinden 38 tane metinsel özellik çıkartılmış ve çıkartılan bu özellikler destek vektör makineleri ile işlenmiştir, uygulama sonucunda %62 başarı sağlanmıştır [De Vel, 2000]. 2001 yılında yapılan bir çalışmada 4 yazara ait 253 elektronik posta yazar belirleme için kullanılmıştır. Eldeki metinlerden çıkartılan 211 tane metinsel özellik destek vektör makineleri kullanılarak işlenmiş ve 2-gramlardan en başarılı sonuçlar üretilmiştir [Corney et al, 2001]. Yine aynı yıl yapılan başka bir çalışmada İngiliz olan 4 erkek haber grubu yazarının 4 haber grubu için yazdığı 1259 metin veri kümesi olarak kullanılmıştır. Çalışmada toplam 191 tane metinsel özellik destek vektör makineleri ile işlenmiştir [De Vel et al, 2001]. 2003 yılında yapılan bir çalışmada 11 yazara ait ortalama 200 kelimeden oluşan 480 elektronik posta veri kümesi olarak tercih edilmiştir. Metinsel özellikler C4.5 ve destek vektör makineleri kullanılarak işlenmiştir. Hatalar tek başına metinsel özellik olarak ele alındığında C4.5'in daha başarılı olduğu gözlemlenmiştir [Koppel and Scheler, 2003].

İngilizce ve Arapça forum mesajları üzerine yapılan çok dilli yazar belirlemede metinsel özelliklerin işlenmesi için C4.5 ve destek vektör makineleri kullanılmıştır. Her dil için 5 yazara ait yirmişer mesaj seçilmiş ve destek vektör makinelerinin daha başarılı olduğu gözlemlenmiştir [Abbasi and Chen, 2005]. Yine çok dilli bir çalışma 2006 yılında yapılmıştır. Bu çalışmada İngilizce ve Çince haber grubu mesajları veri kümesi olarak kullanılmış ve yazarların alışkanlıklarının söz dizilimsel özellikleri ayırt edici hale getirdiği tespit edilmiştir [Zheng et al, 2006]. Abbasi ve Chen'in 2008'de yaptıkları çalışmada 4 adet çevrimiçi veri kümesi ve 4 adet yöntem hem yazar belirleme hem de benzerlik tespiti için kullanılmıştır. Yazar sayısı 25, 50 ve 100 olarak belirlenerek yapılan deneylerde 25 yazar için %96 başarı sağlanmıştır [Abbasi and Chen, 2008]. Sık örüntüler temelli olan çalışmada her yazar için farklı örüntüler tanımlanmış ayrı yazarlar için aynı olan örüntüler elenmiştir. 6 yazar için yirmişer elektronik posta kendilerinin geliştirdiği AutoMiner isimli algoritma ile işlenmiştir [Iqbal et al, 2008].

AutoMiner algoritmalarının geliştiricileri 2011 yılında geliştirdikleri AutoMiner2 algoritması ile 158 yazarın kırkar tane elektronik postasından oluşan veri kümesi üzerinde %70 ile %90 arasında başarı elde etmişlerdir. AutoMiner'dan

farkı bir şüpheliye ait elektronik postalar alt gruplara bölünür ve bu alt gruplardan sık örüntüler çıkartılır [Iqbal et al, 2011]. 2012 yılında yapılan çalışmada elektronik postaların yazarlarının belirlenmesi ile ilgili genel bilgiler verilmiş ve metin madenciliğinden farkları anlatılmıştır [Bogawar and Bhoyar, 2012]. Türkçe için 2012 yılında yapılan çalışmada elektronik postalar yerine elektronik postalar ile aynı karakteristiği gösteren haber grubu mesajları yazarları belirlenmek amacı ile kullanılmıştır. Çalışmada 5 yazara ait 250 metinden çıkartılan 43 metinsel özellik J48, çok katmanlı YSA ve destek vektör makineleri ile işlenmiş ve %83 F-ölçümü ile karar ağaçları en başarılı yöntem olarak tespit edilmiştir [Ekinci ve Takçı, 2012].

6. UYGULAMA

6.1. Veri Kümesinin Elde Edilmesi

Sadece gönderildiği kişilerin görebildiği elektronik postalar, mahremiyet ve güvenlik açısından elde edilmesi oldukça güç olan bir veri kümesidir. Bu durum herkes tarafından görülebilen ve elektronik postalarla aynı karakteristiği gösteren bir veri kümesine ihtiyacı doğurmuştur. Bu ihtiyaç sonucunda haber grubu mesajları elektronik postalar yerine çalışmalarda veri kümesi olarak kullanılmaya başlanmıştır [De Vel, 2001].

Haber grupları, ortak ilgi alanına sahip kullanıcıların belirli gruplar dahilinde siyasetten sağlığa, hukuktan teknolojiye çeşitli konularda mesaj yazdıkları ortamlar şeklinde tanımlanabilir. Elektronik postaların aksine herkes tarafından görülebilen haber grubu mesajları, elektronik postalar gibi gönderici, alıcı, tarih ve zaman ve konu gibi yapısal alan ile gönderilen mesajın gövdesine karşılık gelen yapısal olmayan alandan oluşması nedeniyle bu çalışmada veri kümesi olarak tercih edilmiştir.

Çalışmada kullanılacak haber grubu mesajları www.newskolik.net adlı haber grubu sitesinden elde edilmiştir. Çeşitli haber gruplarından çeşitli konularda 5 yazar için Çizelge 6.1.'de gösterildiği gibi toplamda 250 tane metin seçilmiştir. Veri kümesi MS Access veritabanında depolanmıştır. Veri kümesi oluşturulduktan sonra eldeki veri kümesi üzerinde önişleme yapılması aşamasına geçilmiştir.

Çizelge 6.1.: Haber Grubu Mesajlarına Ait Veri Kümesi.

Yazar	**Mesaj Sayısı**	**Mesaj Boyutu (Kelime Sayısı)**	
		Minimum	**Maksimum**
Yazar1	50	23	86
Yazar2	50	18	98
Yazar3	50	26	86
Yazar4	50	21	76
Yazar5	50	19	114

6.2. Önişleme Adımı

Klasik adli bilişim yöntemlerinde gönderilen elektronik postaların, haber grubu mesajlarının gerçek sahiplerini belirlemek için yapısal alan bilgilerinden yararlanılmaktadır ancak bu bilgiler önleyici hizmetleri gerçek suçluya götürmede her zaman başarılı olamamaktadır. Yazar belirleme ile gerçek suçluya ulaşmak için elektronik ortamdaki metinlerin gövdeleri ile yani yapısal olmayan alanı ile ilgilenmek gerekmektedir. Bu nedenle önişlemeye, eldeki veri kümesi üzerinden ilk olarak yapısal alan bilgilerinin çıkartılması ile başlanmıştır. Yapısal alan bilgileri eldeki mesajlardan çıkartıldıktan sonra sıra yapısal olmayan alan üzerinde yapılacak önişleme adımına gelmiştir.

Yapısal olmayan alan mesajın gövdesidir ve bu alan selamlamalardan, karakterlerden, kelime ve cümlelerden, kısaltmalardan, alıntılardan, cevaplama metinlerinden, linklerden ve imzalardan oluşmaktadır. Yapısal olmayan bu alan üzerinde yapılacak ilk önişleme ise yazara ait olmayan alıntıların ve cevaplama metinlerinin çıkartılmasıdır. Alıntılar, cevaplama metinleri ve imza alanındaki telefon numaraları ve elektronik posta adresleri çıkardıktan sonra linkler de eldeki mesajlardan çıkartılmıştır. Önişleme yapılmasındaki en önemli neden amaca uygun olmayan bilgilerden mesajı temizleyerek gerekli bilgiye daha kısa sürede ulaşmaktır. Önişleme ile sadece gereksiz bilgi eldeki metinlerden çıkartılmaz, önişleme ile yazar belirleme için önemli olan bilgiye de erişim sağlanır. Önişleme adımı metinsel özelliklerin çıkartılması aşamasında da devam etmektedir.

6.3. Metinsel Özelliklerin Çıkartılması

Bu çalışma için eldeki veri kümesinden karakter tabanlı, kelime tabanlı, söz dizilimsel, anlamsal ve uygulamaya özel özellikler başlıkları altında toplamda Çizelge 6.2.'de gösterildiği gibi 49 metinsel özellik çıkartılmıştır. Elektronik ortamdaki metinlerin edebi metinlere göre kısa olmaları, pek çok yazım ve noktalama hatası içermeleri, kısaltma içermeleri yazar belirleme için bir sorun teşkil ediyor gibi görünse de bu tür bilgilerin metinsel özellik olarak kullanılması ile yazar belirlemede başarılı sonuçlar elde edilmektedir.

Çizelge 6.2.: Metinsel Özellikler.

Özellik No	Özellik	Özellik No	Özellik
1	Karakter Sayısı	14	Ortalama Cümle Uzunluğu
2	Harf Sayısı	15	Selamlama
3	Rakam Sayısı	16	İmza (Yazarın İsmi/Lakabı)
4	Noktalama Sayısı	17	Devrik Cümle
5	Büyük Harf Sayısı	18	İkilemeler
6	Küçük Harf Sayısı	19	Kelime Zenginliği
7	Mimik	20	Bir Kere Geçen Kelime
8-10	Hatalar	21	İki Kere Geçen Kelime
11	Kelime Sayısı	22-41	Fonksiyonel Kelimeler
12	Cümle Sayısı	42	Kısaltmalar
13	Ortalama Kelime Uzunluğu	43-49	Kelime Türleri

Yapılan çalışmada metinsel özellik olarak öncelikle karakterler ile ilgili özellikler çıkartılmıştır. Çıkarılması en kolay olan ve herhangi bir önişleme adımı gerektirmeyen bu özelliklerden harf, rakam, noktalama ve özel karakterlerin oluşturduğu toplam karakter sayısı, harf sayısı, rakam sayısı, noktalama sayısı, büyük ve küçük harf sayıları eldeki veri kümesinden çıkartılmıştır. Elektronik ortamdaki metinlerde en sık karşılaştığımız mimik kullanımı da özellik olarak yine bu çalışmada tercih edilmiş ve metinler mimik içerip içermediklerine göre 0 ya da 1 olarak değer almıştır. Bu adımdan sonra eldeki veri kümesi üzerinde hatalar ile ilgili bir önişleme yapılmıştır. Hatalar yazar belirlemede zorluk çıkarmak ile birlikte veri kümesinden çıkartılan önemli özelliklerdir. Veri kümesinden; yazım hatası, format hatası ve noktalama hatası çıkartılmıştır. Hatalar eldeki veri kümesinde olup olmamalarına göre 0 ya da 1 değerini almıştır. Bu hatalar çıkartıldıktan sonra kelimeler ve cümleler ile ilgili doğru sonuçlar alabilmek adına yazım hataları ve noktalama hataları manuel olarak düzeltilmiştir. Bir metinde geçen toplam kelime sayısı ve cümle sayısı çıkartıldıktan sonra ortalama kelime uzunluğu ve ortalama cümle uzunluğu hesaplanmıştır. Uygulamaya özel olan metinsel özelliklerden selamlama ve imza yazarlar ile ilgili önemli bilgiler içermekle birlikte bu tür bilgilerin kullanılmasının bir alışkanlık olduğu da gözlemlenmiştir. Devrik cümlenin kullanımı da yine yazar ile ilgili bir bilgi olduğu için eldeki veri kümesinden metinsel özellik olarak çıkartılmış ve metinler devrik cümle içerip içermediklerine göre 0 ya da 1 değerine göre etiketlenmişlerdir. İkilemeler de devrik cümleler gibi metinlerde bulunup bulunmamalarına göre 0 ya 1 değerini almışlardır. Bu adımlardan sonra çıkartılacak metinsel özellikler için yine bir önişleme adımı gerekmiş ve Türkçe doğal dil kütüphanesi olan Zemberek'ten yararlanılmıştır.

Zemberek sadece Türkiye Türkçesi için değil diğer Türk dilleri için de çözümler üreten bir doğal dil işleme kütüphanesidir. Zemberek; yazım denetiminden, kelime kök ve gövdesini bulma, kelimeyi eklerine ayırma, biçimsel analiz yapma gibi pek çok işlevi barındırmaktadır [Akın ve Akın, 2007]. Bu çalışmada Zemberek kelimelerin kök ve gövdelerini belirlemek amacı ile kullanılmıştır. Kelimelerin kök ve gövdeleri belirlenmeden önce veri kümesinde yine bir önişleme yapılarak tüm rakamlar, noktalama işaretleri, özel işaretler ve mimikler eldeki metinlerden çıkartılmıştır ve elde sadece kelimeler kalmıştır. İlk olarak eldeki veri kümesinden kelime zenginliği, bir ve iki kere geçen kelimeler çıkartılmıştır. Fonksiyonel kelimeler çıkartılmadan önce Türkçe için fonksiyonel kelimeler belirlenmiştir. 100 civarında belirlenen bu fonksiyonel kelimeler üzerinde bir eleme yapılmış ve en az 75 metinde geçen fonksiyonel kelimeler eldeki veri kümesine dahil edilmiştir ve elimizde sadece 20 tane fonksiyonel kelime kalmıştır. Bunun en önemli nedeni ise haber grubu mesajlarının elektronik postalar gibi kısa olmalardır. Çıkartılan fonksiyonel kelimeler eldeki veri kümesi için frekanslarına göre etiketlenmiştir. Elektronik ortamdaki metinlerde en sık karşılaştığımız kısaltmaların çıkartılması için de Zemberek'ten faydalanılmıştır. Eldeki veri kümesinden; sıfat, fiil, zamir, zarf, edat, bağlaç, ünlem olmak üzere 7 adet kelime türü frekansları ile birlikte çıkartılmıştır. Son adımda da karakter sayıları, kelime ve cümle sayıları üzerinde min-max normalleştirmesi yapılarak değerleri 0-1 aralığına indirgenmiştir. Metinsel özelliklerin çıkartılması amacıyla geliştirilen kod Java programlama dili ile NetBeans 7.1 ortamında geliştirilmiştir. Zemberek doğal dil işleme kütüphanesi de yazılan kod içerisine entegre edilmiştir.

6.4. Veri Madenciliğinin Uygulanması ve Model Değerlendirme

Çıkartılan 49 tane metinsel özelliğin işlenmesi için eldeki veri kümesine sırasıyla J48 karar ağacı, Naive Bayes, çok katmanlı yapay sinir ağları, SMO, Bagging ve AdaBoostM1 algoritmaları Java dili platformuna WEKA kütüphanesinin dahil edilmesi ile uygulanmıştır. Bir veri madenciliği aracı olan WEKA sınıflandırma, kümeleme ve birliktelik kurallarını kendisine verilen çeşitli formattaki verilere uygulaması ile bilinen güçlü bir yazılımdır.

Sınıflandırma konusunda oldukça başarılı olan algoritmalardan gürültüye karşı duyarlı olan J48 karar ağacı, hızlı sonuçlar üreten Naive Bayes, doğrusal olmayan verileri ayırmadaki performansı ile bilinen çok katmanlı yapay sinir ağları (öğrenme katsayısı 0.2, momentum katsayısı 0.6 alınmıştır), yüksek boyutlu verileri sınıflandırmada gösterdiği başarı ile SMO, iyi genelleştirme ve güçlü performansları ile Bagging ve AdaBoostM1 (her iki algoritma da temel sınıflandırıcı olarak J48, NaiveBayes, Çok Katmanlı YSA ve SMO'yu kullanmıştır) eldeki veri kümesine uygulanıp sonuçların elde edilmesinde 10 kat çapraz doğrulamadan yararlanılmıştır (Bkz. Bölüm 4). 10 kat çapraz doğrulama, eldeki veri kümesini 10 eşit parçaya bölerek her defasına bu 10 parçadan birini test seti geri kalanları eğitim seti olarak kullanıp sınıflandırma sonucu üretmektedir. Bu işlem 10 defa gerçekleştirilmekte ve 10 kat çapraz doğrulamanın kullanılması ile sınıflandırmadan elde edilen sonuçlar daha güvenilir bir hale gelmektedir.

Eldeki veri kümesine sınıflandırma algoritmalarının 10 kat çapraz doğrulama kullanılarak uygulanması ile her bir yöntem için bir sınıflandırma modeli elde edilmiştir. Elde edilen bu modellerin performanslarının belirlenmesinde Çizelge 6.3.'teki karıştırma matrisinden yararlanılmaktadır.

Çizelge 6.3.: Karıştırma Matrisi.

	TAHMİN EDİLEN SINIF		
GERÇEK SINIF		Sınıf=1	Sınıf=0
	Sınıf=1	a (Doğru Pozitif)	b (Yanlış Negatif)
	Sınıf=0	c (Yanlış Pozitif)	d (Doğru Negatif)

Karıştırma matrisindeki a (DP), gerçek sınıfı 1 olan ve sınıflandırma sonucunda da sınıf 1'e atanan kayıtların sayısını göstermektedir. b (YN), gerçek sınıfı 1 olan ancak sınıflandırma sonucunda sınıf 0'a atanan yanlış kayıtları yani yanlış sınıflandırılan kayıtların sayısını göstermektedir. c (YP), gerçek sınıfı 0 olan ancak sınıflandırma sonucunda sınıf 1'e atanan kayıtları tutmaktadır. c'de b gibi yanlış sınıflandırılan kayıtların sayısına karşılık gelmektedir. d (DN) gerçek sınıfı sınıflandırma sonucunda da tahmin edilen sınıfı 0 olan kayıtların sayısını tutmaktadır. Karıştırma matrisini oluşturan bu sayısal veriler, veri madenciliğinde

kullanılan doğruluk, kesinlik, duyarlılık ve F-ölçümü ile ilgili hesaplamalarda kullanılmaktadır.

Doğruluk, doğru sınıflandırılan kayıtların sayısının yanlış sınıflandırılan kayıtların sayısına oranı olarak tanımlanmaktadır (Eşitlik 6.1.). Doğruluk, diğer bir adıyla sınıflandırıcının doğru tahmin oranıdır. Kesinlik (p); gerçek sınıfı ve tahmin edilen sınıfı 1 olan kayıtların, tahmin edilen sınıfı 1 olan kayıtlara oranı şeklinde tanımlanmaktadır (Eşitlik 6.2.). Duyarlılık (r), gerçek sınıfı ve tahmin edilen sınıfı 1 olan kayıtların gerçek sınıfı 1 olan kayıtlara oranıdır (Eşitlik 6.3.). Yalnız kesinlik ve duyarlılık oranlarını başarı ölçümü olarak kullanılması geçerli değildir. Bu durumu açıklamak gerekirse; kesinlik değerinin 1 olması demek sınıfı 1 olarak belirlenen kayıtların gerçekten de sınıf 1'e ait olduğudur ancak kesinlik değeri bize sınıfı 1 olup 0 olarak etiketlenen kayıt sayısını göstermemektedir. Duyarlılık değerinin 1 olması ise gerçek sınıfı 1 olan tüm kayıtların tahmin edilen sınıfları da 1 ancak sınıfı 0 olup tahmin edilen sınıfı 1 olan kayıtların sayısı bu ölçüm için de bilinmemektedir [Han et al, 2012]. Bu nedenle daha güvenli bir ölçüme ihtiyaç duyulmuştur ve F-ölçümü ortaya çıkmıştır. F-ölçümü kesinlik ve duyarlılık ölçümlerinin harmonik ortalaması alınarak bulunmaktadır (Eşitlik 6.4.).

$$Doğruluk = \frac{DP+DN}{DP+YN+YP+DN} \tag{6.1}$$

$$Kesinlik = \frac{DP}{DP+YP} \tag{6.2.}$$

$$Duyarlılık = \frac{DP}{DP+YN} \tag{6.3.}$$

$$F - \text{ölçümü} = \frac{2\times kesinlik\times duyarlılık}{kesinlik+duyarlılık} \tag{6.4.}$$

5 yazarın 250 metninden oluşan veri kümesine uygulanan algoritmalar sonucunda elde edilen performans ölçümleri Çizelge 6.4.'te, 6.5.'te ve 6.6.'da, yanlış sınıflandırılan örnek sayıları Çizelge 6.7.'de gösterilmiştir. Algoritmaların model oluşturma zamanı açısından değerlendirilmesi de Çizelge 6.8.'de gösterilmiştir.

Algoritmaların zaman açısından değerlendirildiği makine 64 bit işletim sistemine, 6 GB RAM'e ve 2.5 GHz CPU hızına sahiptir.

Çizelge 6.4.: Tekli Sınıflandırıcıların Ortalama Sınıflandırma Başarıları.

Başarı Ölçümleri	J48	Naive Bayes	Çok Katmanlı YSA	SMO
Doğruluk	0.784	0.788	0.824	0.812
Kesinlik	0.786	0.79	0.825	0.817
Duyarlılık	0.784	0.788	0.824	0.812
F-Ölçümü	0.784	0.788	0.824	0.814

Çizelge 6.5.: Bagging için Ortalama Sınıflandırma Başarıları.

Başarı Ölçümleri	Bagging J48	Bagging Naive Bayes	Bagging Çok Katmanlı YSA	Bagging SMO
Doğruluk	0.812	0.772	0.840	0.860
Kesinlik	0.815	0.773	0.841	0.863
Duyarlılık	0.812	0.772	0.840	0.860
F-Ölçümü	0.813	0.772	0.840	0.861

Çizelge 6.6.: AdaBoostM1 için Ortalama Sınıflandırma Başarıları.

Başarı Ölçümleri	AdaBoostM1 J48	AdaBoostM1 Naive Bayes	AdaBoostM1 Çok Katmanlı YSA	AdaBoostM1 SMO
Doğruluk	0.848	0.776	0.824	0.812
Kesinlik	0.851	0.776	0.825	0.817
Duyarlılık	0.848	0.776	0.824	0.812
F-Ölçümü	0.848	0.775	0.824	0.812

Çizelge 6.7.: Algoritmaların Yanlış Sınıflandırdığı Örnek Sayıları.

	J48	Naive Bayes	Çok Katmanlı YSA	SMO
Tekli	54	53	44	47
Bagging	47	57	40	35
AdaBoostM1	38	56	44	47

Çizelge 6.8.: Model Oluşturulma Zamanları.

Algoritmalar	Model Oluşturma Zamanları (Saniye Cinsinden)	Model Oluşturma Zamanları (Bağıl Zaman)
J48	0.16	1.172×10^{-3}
Naive Bayes	0.05	0.366×10^{-3}
Çok Katmanlı YSA	6.29	46×10^{-3}
SMO	0.27	1.977×10^{-3}
Bagging J48	0.74	5.418×10^{-3}
Bagging Naive Bayes	0.07	0.513×10^{-3}
Bagging Çok Katmanlı YSA	136.57	1000×10^{-3}
Bagging SMO	2.75	20×10^{-3}
AdaBoostM1 J48	1.69	12×10^{-3}
AdaBoostM1 Naive Bayes	1.09	7.981×10^{-3}
AdaBoostM1 Çok Katmanlı YSA	68.94	505×10^{-3}
AdaBoostM1 SMO	1.5	11×10^{-3}

7. SONUÇLAR VE ÖNERİLER

Son dönemlerde suça sıklıkla alet edilen yöntemlerden biri olan elektronik postaların adli bilimler açısından analizinde mevcut yöntemler açısından yetersizlikler bulunmaktadır. Daha çok elektronik postaların başlık bilgilerine göre suç ve suçluyla mücadele eden bu mevcut yöntemler maalesef ihtiyaçları karşılayamamakta bununla birlikte gerçek suçluyu belirlemek ise büyük bir önem taşımaktadır. Bu nedenle; elektronik postaların gerçek yazarlarını adli bilimler açısından belirlemek amacıyla; mesaj gövdesine dayalı olarak elektronik postaların analizini yerine getiren ve böylece elektronik postaların görünen sahibi yerine asıl sahibini bulabilen ayrıca elektronik postaların güvenliğini arttırabilen örnek bir uygulama bu çalışma ile gerçekleştirilmiştir.

Geçmişten günümüze elektronik postalar üzerine yapılan yazar belirleme çalışmalarında farklı veri ve özellik setleri ile farklı sayıda yazarlar kullanılmıştır. Elektronik postalar üzerine bilinen ilk uygulamada De Vel, 5 yazar için 274 elektronik postadan oluşan bir veri kümesi kullanmış ve bu veri kümesinden 38 metinsel özellik çıkartmıştır [De Vel, 2000]. Corney ve arkadaşları 2001 yılında yaptıkları çalışmada 4 yazara ait 253 elektronik postadan 211 metinsel özellik çıkartmıştır [Corney et al, 2001]. Yapılan bu çalışmanın deneysel kısmında ise haber grubu mesajları veri kümesi olarak kullanılmıştır. Veri kümesinde 5 yazara ait 250 haber grubu mesajı bulunmakta olup deneysel çalışmanın özellik setinde bu metinlerden elde edilen 49 adet metinsel özellik yer almıştır.

Mahremiyet ve güvenlik açısından elektronik postaların elde edilmesinin güç olması elektronik postalar ile aynı karakteristiği gösteren haber grubu mesajlarına yöneltmiş, haber grubu mesajlarının elde edildiği kaynakta mesajların büyük çoğunluğunun yazar belirleme için yeterli özellikleri bulundurmaması nedeni ile yazar sayısı, veri kümesi boyutu ve çıkarılan metinsel özelliklerde kısıtlamalar yapılmıştır. Mevcut kaynaktaki mesajlar oldukça kısa ve genellikle tek kelime, link ve ekleri içermesi kaynaktaki verilerin büyük çoğunluğunun yazar belirleme için yetersiz olduğunu göstermiştir.

Veri kümesinden çıkartılan metinsel özellikler J48, Naive Bayes, Çok Katmanlı YSA ve SMO algoritmaları yardımıyla sınıflandırma başarısı açısından test edilmiştir. Bu algoritmaların metinsel özellikleri işlemede seçilmesinin esas nedeni yazar belirleme üzerine yapılan çalışmalarda en sık tercih edilen ve başarılı yöntemler olmalarıdır. Karar ağaçları hem gürültüye karşı duyarlı ve güvenilir olması hem de elektronik ortamdaki metinlerin yazarlarını belirlemede tercih edilmesi nedeniyle kullanılmıştır [Koppel and Scheler, 2003]. Karar ağacı algoritması olarak seçilen algoritma hem kategorik hem de sayısal veriler ile çalışabilmesi nedeniyle J48 olmuştur. Bayes sınıflandırıcılar yazar belirlemede kullanılan ilk yöntemdir ve özellikle metin sınıflandırmada başarılı sonuçlar ürettiği gözlemlenmiştir. Başarılı bir sınıflandırma yapan yapay sinir ağları ve doğrusal olmayan veriler üzerinde çalışan çok katmanlı yapay sinir ağları bugüne kadar genellikle edebi metinlerin yazarlarını belirlemek için kullanılmıştır. Destek vektör makineleri ise elektronik ortamdaki metinlerin yazarlarını belirlemede ilk ve en sık kullanılan yöntemdir [De Vel, 2000]. Gürültüye karşı duyarlı olması ve yüksek boyutlu verilerde başarılı sınıflandırma yapması ile bilinen destek vektör makineleri içerisindeki SMO algoritmasının tercih nedeni ise WEKA'nın sağladığı bir algoritma olmasıdır. Gruplamalı Sınıflandırıcılarda ise Bagging ve AdaBoostM1 algoritmaları kolay gerçekleştirim, güçlü performans ve iyi bir sınıflandırıcı olmaları nedeni ile tercih edilmişlerdir.

J48, NaiveBayes, Çok Katmanlı YSA ve SMO kendi içlerinde değerlendirildiklerinde %82.4 ile Çok Katmanlı YSA'nın en başarılı yöntem olduğu gözlemlenmiştir. Gruplamalı Sınıflandırıcılardan Bagging temel sınıflandırıcı olarak SMO'yu kullandığında en iyi genellemeyi sağlarken AdaBoostM1 J48 ile en iyi genellemeyi sağlamıştır. Gruplamalı Sınıflandırıcılar kendi aralarında değerlendirildiğinde ise Baggingin daha iyi bir genelleme sağladığı gözlemlenmiştir. NaiveBayes'in temel sınıflandırıcı olarak kullanıldığı Gruplamalı Sınıflandırıcılarda ise genelleme başarısı gözlemlenmemiştir.

Ardıl'ın 2009 yılındaki çalışmasında, metin sınıflandırmada oldukça hızlı olduğunu gözlemlediği Naive Bayes [Ardıl, 2009] bizim çalışmamızda da 0.05 saniye ile en hızlı çalışan algoritma olmuştur. Ayrıca NaiveBayes'i temel

sınıflandırıcı olarak alan Gruplamalı Sınıflandırıcılar da diğerlerine göre daha hızlı çalışmıştır.

Çalışmadan elde edilen en önemli sonuç ise, yazar belirlemenin gerçek suçluyu belirlemede oldukça başarılı bir yöntem olduğu ve adli bilimler üzerine çalışan bilim adamlarına büyük fayda sağlayacağıdır.

Yazar belirleme önceleri sadece edebi metinlerin yazarlarını tespit etmede kullanılırken son zamanlarda artan yetenekler dolayısıyla adli bilimlerde de tercih edilir bir yöntem haline gelmiştir. Çalışma etkinliğini arttırabilmek için bu çalışmaya ek bazı çalışmaların yapılması elzemdir.

1) Yazar belirleme suç analizi amacıyla kullanılacağı zaman bir sınıflandırma probleminden çok bir doğrulama problemi olarak ele alınmalıdır. Devam eden çalışmalarda doğrulamaya dayalı elektronik postaların yazarları bulunabilir.
2) Aynı zamanda yazar belirleme kapalı sınıf bir sınıflandırma problemi olmakta bundan dolayı da o sınıfta yer alan bütün elemanların tanımlanması gerekmektedir. Mevcut haliyle 5 kişi gibi az sayıda elemanı destekleyen örnek veri kümesi gerçek uygulamalarda olması gereken sayıdan düşüktür, daha etkili bir deneysel çalışma için yazar adedi artırılmalıdır.
3) Hangi metin uzunluklarında sistemin ne kadar doğru sonuç verdiği bulunabilir.
4) Daha fazla sayıda sınıflandırıcı kullanımı daha uygun olacaktır ve gelecek çalışmalarda yazar adediyle birlikte sınıflandırıcı adedi artırılmalıdır.
5) Yazar belirleme problemleri kimlik doğrulama sistemleri ile birlikte kullanılmaya müsait bir konudur. Bu konuda geliştirmeler yapılabilir.

KAYNAKLAR

[1] S. Dönmezer, Kriminoloji. İstanbul: Filiz Kitabevi, 1984.

[2] H. Chen, W. Chung, J.J. Xu, G. Wang and Y. Qin, "Crime Data Mining: A General Framework and Some Examples," *IEEE Computer Society*, vol. 37, no. 4, pp. 50-56, 2004.

[3] K. Inman and N. Rudin, Principles and Practice of Forensic Science: The Profession of Forensic Science. CRC Press, 2000.

[4] İ. H. Hancı, "Adli Bilişim Bilimi ve Diğer Bilimlerle Olan İlişkisi," İzmir: Adli Bilişim Çalıştayı Sunumu, 2005.

[5] D. B. Parker, Computer Crime Criminal Justice Resource Manual. Washington D.C, 1989.

[6] D. Hu, " Exploratory Study on Computer Forensic Technology ," *IEEE Computer Society*, vol. 2, pp. 608-611, 2009.

[7] B. M. Leiner, V. G. Cef, D. D. Clark, R. E. Kahn, L. Kleinrock, D. C. Lynch, J. Postel, L. G. Roberts and S. S. Wolff, "The Past and Future History of the Internet", *Communication of the ACM*, vol. 40, no. 2, pp. 102-108, 1997.

[8] A. Gribbin, "A Brief History of the Internet", New Statesman, vol.140, no. 5066, pp.30, 2011.

[9] J. Ma, Y. Li, G. Teng, F. Wang and Y. Zhao, "Sequential Pattern Mining for Chinese E-mail Authorship Identification", *IEEE Computer Society*, 2008.

[10] A. Visa, "Technology of Text Mining", *P. Perner (Ed.): MLDM 2001*, LNAI 2123, pp. 1–11, 2001.

[11] P. M. Nadkarni, L. Ohno- Machado and W. W. Chapman, "Natural Language Processing: An Introduction ", *Journal Of The American Medical Informatics Association: JAMIA [J Am Med Inform Assoc]*, vol. 18, no. 5, pp. 544-51, 2011.

[12] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques. USA: Morgan Kaufmann, 2012.

[13] A. Abbasi and H. Chen, "Applying Authorship Analysis To Extremist-Group Web Forum Messages", *IEEE Computer Society*, vol. 20, no. 5, pp.67-75, 2005.

[14] R. Zheng, J. Li, H. Chen and Z. Huang, “A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques”, *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378-393, 2006.

[15] E. Stamatatos, "A Survey of Modern Authotship Attribution Methods", *Journal of the American Society for Information Science and Technology,* vol. 60, no. 4, pp. 538-556, 2009.

[16] T. C. Mendenhall, “Characteristic Curves of Composition”, *American Association for the Advancement of Science*, vol. 9, no. 214, pp. 237-246, 1887.

[17] F. Mosteller and D. L. Wallace, “Inference in Authorship Problem”, *Journal of the American Statistical Association*, vol 58, no. 302, pp. 275-309, 1963.

[18] O. De Vel, “Mining E-mail Authorship”, *ACM International Conference on Knowledge Discovery and Data Mining (KDD’2000)*, vol. 30, no. 4, pp. 50-56, 2000.

[19] P. S. Bogawar and K. K. Bhoyar, "Email Mining: A Review", *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 429-434, 2012.

[20] J. C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", *Advances in Kernel Methods*, vol. 12, pp. 185-208, 1999.

[21] G. Wang, J. Hao, J. Ma and H. Jiang, "A Comparative Assessment of Ensemble Learning for Credit Scoring", *Expert Systems with Applications*, vol. 38, pp. 223-230, 2011.

[22] H. Doğan and O. Akay, "Using AdaBoost classifiers in a hierarchical framework for classifiying surface images of marble slabs", *Expert Systems with Applications*, vol. 37, pp. 8814-8821, 2010.

[23] S. E. Fienberg, "Editorial: Statistics and Forensic Science", *The Annals of Applied Statistics,* vol. 1, no. 2, pp. 285-286, 2007.

[24] Türk Dil Kurumu, "Türkçe Sözlük". Ankara: Türk Dil Kurumu Yayınları, 2005.

[25] T. Demirbaş, Kriminoloji. Ankara: Seçkin Yayıncılık, 2001.

[26] R. L. Perry, Computer Crime. New York: Franklin Watts, 1986.

[27] Ş. C. Taşkın, Bilişim Suçları. Bursa: Beta, 2008.

[28] D. Ranganayakulu, L. Kavisankar and C. Chellappan, “Enhanced E-Mail Authentication Against Spoofing Attacks”, *European Journal of Scientific Research*, vol. 54, no. 1, pp 165-175, 2011.

[29] D. Jamil and H. Zaki, “Software Piracy Does Not Hurt Anyone?”, *International Journal of Engineering Science and Technology*, vol. 3, no. 4, pp. 3467-3471, 2011.

[30] R. Hankins, T. Uehara and J. Liu, "A Comparative Study of Forensic Science and Computer Forensics", *Third IEEE International Conference on Secure Software Integration and Reliability Improvement*, pp. 230-239, 2009.

[31] T. Henkoğlu, Adli Bilişim Dijital Delillerin Elde Edilmesi ve Analizi. İstanbul: Pusula Yayıncılık, 2011.

[32] H. C. Özmutlu ve S. Özmutlu," Bilgisayar Ağları Aracılığı ile Gerçekleştirilebilecek Suçlar ve Yaşanan Sorunlar", 1. Polis Bilişim Sempozyumu, sf. 84-87, 2003.

[33] D. J. Hand, "Data Mining: Statistics and More?", The American Statistical Association, vol. 52, no. 2, pp. 112-118, 1998.

[34] S. Tüzüntürk, "Veri Madenciliği ve İstatistik", *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, Cilt XXIX, sayı 1, s. 65-90, 2010.

[35] Y. Özkan, Veri Madenciliği Yöntemleri. İstanbul: Papatya Yayıncılık, 2008.

[36] D. J. Hand, "Statistics and Data Mining: Intersecting Disciplines", *SIGKDD Explorations*, vol. 1, no. 1, pp. 16-19, 1999.

[37] D. Kumar and D. Bhardwaj, "Rise of Data Mining: Current and Future Application Areas", *IJCSI International Journal of Computer Science Issues*, vol. 8, no. 1, pp. 256-260, 2011.

[38] R. C. Barros, M. P. Masgalupp, A. C. P. L. F. de Carvalho and A. A. Freitas, “A Survey of Evolutionary Algorithms for Decision-Tree Induction”, *IEEE Transactions on Systems, Man, and Cybernetıcs—Part C: Applications and Reviews*, vol. 42, no. 3, pp. 291-311, 2012.

[39] L. Rokach and O. Maimon, Data Mining With Decision Trees. Singapore: World Scientific Publishing Co. Pte. Ltd, 2008.

[40] A. Mahdi, A. Razali and A. AlWakil, "Comparison of Fuzzy Diagnosis with K-Nearest Neighbor and Naïve Bayes Classifiers in Disease Diagnosis",

Broad Research in Artificial Intelligence and Neuroscience, vol. 2, no. 2, pp. 58-66, 2011.

[41] E. Öztemel, Yapay Sinir Ağları. İstanbul: Papatya Yayıncılık, 2006.

[42] C. Hamzaçebi, Yapay Sinir Ağları. Bursa: Ekin Basım Yayın Dağıtım, 2011.

[43] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer-Verlac, 1995.

[44] G. Teng, M. Lai, J. Ma and Y. Li, "E-mail Authorship Mining Based on SVM for Computer Forensic", *Proceedings of Third International Conference on Machine Learning and Cybernetics*, vol. 2, pp.1204-1207, 2004.

[45] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya and K. R. K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design", *Neural Computation*, vol. 13, pp. 637-649, 2001.

[46] B. V. Dasarathy and B. V. Sheela, "A Composite Classifier System Design: Consepts and Methodology", *Proceedings of the IEEE*, vol. 67, no. 5, pp. 708-713, 1979.

[47] R. Polikar, "Ensemble Based Systems in Decision Making", *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45, 2006.

[48] G. Liang, X. Zhu and C. Zhang, "An Emprical Study of Bagging Predictors for Different Learning Algorithms", *Proceedings of Twenty-Fifth AAAI Conference on Artifical Intelligence*, pp. 1802-1803, 2011.

[49] L. Breiman, "Bagging Predictors", *Machine Learning*, vol. 24, pp. 123-140, 1996.

[50] M. Liu, "Fingerprint classification based on AdaBoost learning from singularity features", *Pattern Recognition*, vol. 43, pp. 1062-1070, 2010.

[51] I. H. Witten, "Text Mining", *Practical handbook of internet computing*, pp. 1-23, 2005.

[52] A. Oğuzlar, Temel Metin Madenciliği. Bursa: Dora Basım Yayın Dağıtım, 2011.

[53] C. H. Aydın, Y. Hoşcan ve A. E. Özkul, Temel Bilgi Teknolojileri. Eskişehir: Anadolu Üniveristesi, 2004.

[54] H. K. Yıldız, M. Gençtav, N. Usta, B. Diri ve M. F. Amasyalı, "Metin Sınıflandırmada Yeni Özellik Çıkarımı", *15. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, sf. 1-4, 2007.

[55] L. H. Chong and Y. Y. Chen, Text Summarization for Oil and Gas News Article, "P*roceedings of World Academy of Science, Engineering and Technology*", vol. 41, pp. 291-294, 2009.

[56] M. Orhun, E. Adalı ve A. C. Tantuğ, "Uygurcadan Türkçeye Bilgisayarlı Çeviri", *İTÜ Dergisi,* vol. 10, no. 3, sf. 3-14, 2011.

[57] M. Corney, O. De Vel, A. Anderson and G. Mohay, "Gender-Preferential Text Mining of E-Mail Discourse", *Computer Security Applications Conference,* pp. 282-289, 2002.

[58] M. Koppel, S. Argamon and A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender", Literary and Linguistic Computing, vol. 17, no. 4, pp. 401-412, 2002.

[59] W. Elliot and R. Valenza, "Was the Earl of Oxford the True Shakespeare? A Computer Aided Analysis", *Notes and Queries*, vol. 38, no.4, pp.501-506, 1991.

[60] S. M. Zu Eissen, B. Stein and M. Kulig, "Plagiarism Detection without Reference Collections", *Advances in Data Analysis*, pp. 359-366, 2007.

[61] H. Baayen, H. V. Halteren, A. Neijt and F. Tweedie, "An Experiment in Authorship Attribution", *JADT 2002*, pp. 69-75, 2002.

[62] U. Yule, "On-Sentence Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship", *Biometrika*, vol. 30, no. 3 /4, pp. 363-390, 1939.

[63] U. Yule, "The Statistical Study of Literary Vocabulary", *Journal of the Royal Statistical Society*, vol. 107, no. 2, pp. 129-131, 1944.

[64] M. Koppel, J. Scheler and S. Argamon, "Computational Methods in Authorship Attribution", *Journal of the American Society for Information Science & Technology*, vol. 60, no. 1, pp. 9-26, 2009.

[65] D. I. Holmes, "The Analysis of Literary Style-A Review", *Journal of the Royal Statistical Society*, vol.148, no. 4, pp. 328-341, 1985.

[66] D. I. Holmes, "The Evolution of Stylometry in Humanities Scholarship", *Literary and Linguistic Computing*, vol. 13, no. 3, pp. 111-117, 1998.

[67] E. Stamatatos, N. Fakotakis and G. Kokkinakis, "Automatic Authorship Attribution", *Proceedings of EACL'99*, pp.158-164, 1999.

[68] J. Grieve, "Quantiative Authorship Attribution An Evaluation of Techniques", *Literary and Linguistic Computing*, vol. 22, no. 3, pp. 251-270, 2007.

[69] M. Koppel, J. Scheler and S. Argamon, "Authorship Attribution in the Wild", Lang Resources & Evaluation, vol. 45, pp. 83-94, 2011.

[70] B. Diri and M. F. Amasyalı "Automatic Author Detection for Turkish Texts", *Yıldız Teknik Üniversitesi*, pp. 1-8, 2003.

[71] B. Diri, M. F. Amasyalı ve F. Türkoğlu, "Farklı Özellik Vektörleri ile Türkçe Dokümanların Yazarlarının Belirlenmesi", *Yıldız Teknik Üniversitesi*, 2006.

[72] T. Taş and A. K. Görür, "Author Identification for Turkish Texts", *Journal of Art and Sciences*, no. 7, pp. 151-161, 2007.

[73] M. Corney, A. Anderson, G. Mohay and O. De Vel, "Identifying the authors of Suspect Email", *Communications of the ACM*, pp. 1-19, 2001.

[74] O. De Vel, A. Anderson, M. Corney and G. Mohay, "Multi-Topic E-mail Authorship Attribution Forensics", *ACM Sigmoid Records*, vol. 30, no. 4, pp.55-62, 2001.

[75] M. Koppel and J. Scheler, "Exploiting Stylistic Idiosyncrasies for Authorship Attribution", *Citeseer*, vo. 3, no. 2000, pp. 69-72, 2003.

[76] A. Abbasi and H. Chen, "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace", ACM Transactions on Information Systems, vol. 26, no. 2, pp. 1-29, 2008.

[77] F. Iqbal, R. Hadjidj, B. C. M. Fung and M. Debbabi, "A Novel Approach of Mining Write-Prints for Authorship Attribution in E-Mail Forensics", Digital *Investigation*, vol. 5, no. 1, pp. 42-51, 2008.

[78] F. Iqbal, H. Binsalleeh, B. C. M. Fung and M. Debbabi, "A Unified Data Mining Solution for Authorship Analysis in Anonymous Textual Communications", *Information Sciences*, pp.1-15, 2011.

[79] E. Ekinci, H. Takçı, "Elektronik Postaların Adli Analizinde Yazar Analizi Tekniklerinin Kullanılması", *20. Sinyal İşleme ve İletişim Uygulamaları Kurultayı*, sf. 1-4, 2012.

[80] M. D. Akın ve A. A. Akın, "Türkçe Dilleri için Açık Kaynaklı Doğal Dil İşleme Kütüphanesi: Zemberek", *Elektrik Mühendisliği*, no. 431, sf. 38-44, 2007.

[81] E. Ardıl, " Esnek Hesaplama Yaklaşımı ile Yazılım Hata Kestirimi", Trakya Üniversitesi Yüksek Lisans Tezi, 2009.

ÖZGEÇMİŞ

Ekin Ekinci, 19.12.1987 de Trabzon’ da doğdu. İlk öğretimini Cumhuriyet İ.Ö.O’ da tamamladı. Orta öğretimini Trabzon Lisesi'nde (Süper Lise) 2005 yılında tamamladıktan sonra aynı yıl Çanakkale 18 Mart Üniversitesi Bilgisayar Mühendisliği bölümünü kazandı ve 2009 yılında bu bölümden mezun oldu. 2009 yılında Gebze Yüksek Teknoloji Enstitüsü'nde yüksek lisans yapmaya başladı. 2010 yılında Kocaeli Üniversitesi Bilgisayar Mühendisliği Bölümü’nde Araştırma Görevlisi olarak göreve başladı ve hala Kocaeli Üniversitesi’nde Araştırma Görevlisi olarak görev yapmaktadır.