

T.C.
GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ
MÜHENDİSLİK VE FEN BİLİMLERİ
ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE MAHREMİYETİN
SAĞLANMASI

Uğur KAVZA
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ
ANA BİLİM DALI

GEBZE

2010

T.C.
GEBZE YÜKSEK TEKNOLOJİ ENSTİTÜSÜ
MÜHENDİSLİK VE FEN BİLİMLERİ
ENSTİTÜSÜ

VERİ MADENCİLİĞİNDE MAHREMİYETİN
SAĞLANMASI

Uğur KAVZA
YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

TEZ DANIŞMANI
Dr. Hidayet TAKÇI

GEBZE
2010

ÖZET

TEZİN BAŞLIĞI: VERİ MADENCİLİĞİNDE MAHREMİYETİN SAĞLANMASI

YAZAR ADI: UĞUR KAVZA

Son yıllarda internet kullanımının yaygınlaşması ile daha fazla kişisel ve kurumsal bilgi gün yüzüne çıkmıştır. Genellikle mahrem öğeler içeren bu bilgilerin büyük bir kısmı; finans kuruluşları, GSM şirketleri, elektronik ticaret sistemleri, devlet kurumları ve buna benzer yerler tarafından kontrolsüz olarak kayıt altına alınıp saklanabilmektedir.

Saklanan veriler üzerinde yapılacak titiz çalışma ve analizler sonucu mahrem olabilecek kişisel veya kurumsal bilgiler ortaya çıkabilmektedir. Özellikle veriler üzerinde veri madenciliği gibi analiz araçları ile gizli bilgiler açığa çıkarılabilmektedir.

Gizli kalması gereken bilgilerin ortaya çıkması insanlar için endişe verici bir durumdur. Bu endişenin ortadan kalkması için veri madenciliğinin mahremiyeti ortadan kaldırmayacağının garantisi verilmelidir. Bunun için de, mahrem verilerin ortaya çıkmasını engelleyecek bir takım teknikler geliştirilmelidir.

Bu tez çalışmasında mahremiyeti korumak için ortaya konan teknikler incelenmiş ve önemli bir veri madenciliği tekniği olan sınıflandırma için öznitelik seçimi yardımıyla mahremiyet korunmaya çalışılmıştır. En az sayıda mahrem nitelik ile en yüksek sınıflandırma başarısı bulunmaya çalışılmıştır.

SUMMARY

THESIS TITLE: PRESERVING PRIVACY ON DATA MINING

AUTHOR: UĞUR KAVZA

In last years more personal and corporate information come to light with the proliferation of using internet. Generally a great part of this information including privacy items are saved chaining unrestrainedly by financial establishments, GSM companies, electronic commerce systems, public institutions and places like this.

With the precise tries and analyses doing on saved data, personal and corporate information that can be secret can emerge. Especially on data, privacy information can be emerged with analyses tools like data mining.

Emerging of confidential information is a threatening situation for people. For removing of this worry must be given guarantee about not removing privacy of data mining. So, a set of techniques must be improved obstructing of emerging of the secret data.

In this thesis study, presented techniques to protect privacy are examined and privacy is tried to being protected with the help of attribute selection for classification that is an important data mining technique. Maximum classification success is tried to find with the least number of secret quality.

TEŞEKKÜR

Tez çalışmasında bana sürekli destek olan ve yardımlarını esirgemeyen sayın hocam Dr. Hidayet TAKÇI' ya,

Tez çalışmam sırasında bana yardımlarını esirgemeyen ve destek olan arkadaşlarıma,

Bugünlere gelmemde büyük emeği olan, hayatım boyunca maddi ve manevi destekleri ile her zaman yanımda olan aileme en içten dileklerimle teşekkür ederim.

İÇİNDEKİLER DİZİNİ

	<u>Sayfa</u>
ÖZET	iv
SUMMARY	v
TEŞEKKÜR	vi
İÇİNDEKİLER DİZİNİ	vii
SİMGELER VE KISALTMALAR DİZİNİ	ix
ŞEKİLLER DİZİNİ	x
ÇİZELGELER DİZİNİ	xi
1. GİRİŞ	1
2. VERİ MADENCİLİĞİNDE MAHREMİYET	4
2.1 Veri Madenciliği Nedir?	4
2.1.1 Veri Madenciliği Aşamaları	5
2.1.2 Veri Madenciliği İşlevleri	7
2.1.3 Veri Madenciliğinin Uygulama Alanları	11
2.2 Mahremiyet	13
2.3 Veri Madenciliğinde Mahremiyet Kavramı	16
2.4 Veri Madenciliğinde Mahremiyet Ölçütleri	17
2.4.1 Bilgiyi Sınırlandırma	18
2.4.2 Yeterli Bilgi Edinme	18
2.4.3 Açığa Çıkmasını Önleme	19
2.4.4 Gerçeği Saklama	19
2.5 Yapılan Çalışmalar	20
2.5.1 Veri Koleksiyonu ile Mahremiyeti Sağlama	20
2.5.2 Sonuç Kontrolü ile Mahremiyeti Sağlama	21
2.5.3 Bilgi Paylaşımı ile Mahremiyeti Sağlama	21
3. ÖNERİLEN YÖNTEM	23
3.1 Yöntem Bileşenleri	24
3.1.1 İşlenecek Veri	24
3.1.2 Öznitelik Seçme Adımı	24

3.1.3 Seçilmiş Veri Setleri	25
3.1.4 Sınıflandırma Adımı	25
3.2 Yöntem Modeli	26
3.3 Önerilen Yöntemin Değerlendirilmesi	27
4. YAPILAN ÇALIŞMA	29
4.1 Veri	31
4.2 Öznitelik Seçimi	32
4.3 Mahrem Olmayan Veri Seti	33
4.4 Sınıflandırma Algoritmaları	35
4.4.1 ID3 Algoritması	35
4.4.2 C-RT Algoritması	37
4.4.3 Naive Bayes Algoritması	37
4.4.4 MLP Algoritması	38
4.5 Veri Madenciliği Yazılımı (Tanagra)	39
4.6 Yapılan Çalışmanın Modeli	40
4.7 Deney	42
4.8 Sınıflandırma Kazancı	43
5. SONUÇ	46
KAYNAKLAR	48
ÖZGEÇMİŞ	51

SİMGELER VE KISALTMALAR DİZİNİ

GÇTİ : Güvenli Çok Taraflı İletişim.

MLP : Multi Layer Perceptron.

UCI : Kaliforniya Üniversitesi, Irvine kampusu.

VTBK : Veri Tabanlarında Bilgi Keşfi.

ŞEKİLLER DİZİNİ

<u>Sekil</u>	<u>Sayfa</u>
2.1. Veri Madenciliğinin Diğer Disiplinlerle İlişkisi	4
2.2. Veri Madenciliği Aşamaları	6
2.3. Veri Madenciliği Klasik İşlem Süreci	6
2.4. Veri Madenciliği İşlevleri	8
3.1. Model Blok Diyagramı	26
4.1. Tanagra Grafiksel Ara Yüzü	40
4.2. Yapılan Çalışma Modeli	41
4.3. Sınıflandırma Sonuçları	43

ÇİZELGELER DİZİNİ

<u>Çizelge</u>	<u>Sayfa</u>
Çizelge 4.1. Yetişkin Verisi.	31
Çizelge 4.2. Veri seti örnekleri.	34
Çizelge 4.3. Sınıflandırma Kazancı.	44

1. GİRİŞ

Bilişim teknolojilerindeki hızlı gelişme ve ilerlemeler yaşantımıza birçok yenilik ve kolaylık getirmiştir. Özellikle son yıllarda artan bilgisayar ve internet kullanımı sonucu; kişiler, kurumlar ve kuruluşlar işlerinin büyük kısmını artık elektronik ortamlarda gerçekleştirmeye başlamışlardır. Bunun sonucunda; e-ticaret, e-kurum, e-devlet, e-imza, e-posta gibi kavramlar hızla klasik çalışma biçimlerinin yerini almıştır. Bu değişimi günlük yaşantımızda neredeyse her alanda görebiliriz. Örneğin bankacılık işlemlerini bankaya gitmeden evimizdeki, iş yerimizdeki kişisel bilgisayarlarımızla veya cep telefonlarımızla yapabiliriz. Elektronik ortamlarda; fatura ödeme, pasaport başvurusunda bulunma, seyahat rezervasyonları yapma, çalıntı cep telefonlarını sorgulama artık sıradan olaylar haline gelmiştir.

Teknolojideki bu baş döndürücü gelişmelere paralel olarak kişiler ve kurumlara ait özel veriler bilgisayar ortamında saklanmaya başlamıştır. Elektronik ortamlarda tutulan bu veriler; bilgi hırsızlığı, elektronik saldırılar, ilgili kuruluşların kendi çalışanlarınca oluşturulabilecek potansiyel iç saldırılar gibi çok geniş bir yelpazeye sahip kaynaklardan gelen tehdit ve tehlikelerle karşı karşıyadır. Özellikle kişisel ve kurumsal bilgilerin izinsiz olarak elde edilmesi veya değiştirilmesi konusundaki tehditler artarak sürmektedir. Kişiler ve kurumlar bu tehlikeler karşısında giderek daha riskli bir duruma gelmektedirler. Ayrıca hizmetlerin internet ortamında sunulması, bilgisayar ağları arasındaki geçişler ve bilgilerin halka açık sistemlerde paylaşılması gibi uygulamaların artması sonucu özel verilere erişimin denetlenmesi güçleşmektedir.

Her yeni teknolojiye olduğu gibi, iletişim ve bilişim teknolojileri de üzerinde çalışılması gereken yeni sorunları beraberinde getirmiştir. Bu sorunlardan belki de en önemlisi ise kişisel ve kurumsal mahremiyetin korunmasıdır.

İnternet, sosyal paylaşım ağları, cep telefonları, GPS cihazları, RFID etiketlerinin kullanıldığı teknolojiler, bireylerin zaman ve mekân bağlamındaki davranışları hakkında şimdiye kadar görülmemiş boyutlarda verinin toplanmasına olanak sağlamıştır. Toplanan bu veriler bir yandan; şehir ve trafik planlama,

reklamcılık, lojistik vb. konularda çok önemli, yararlı ve ticari değeri olan enformasyonun ortaya çıkartılmasına olanak sağlarken, diğer yandan bireylerin mahremiyetini tehlikeye atmaktadır. Bu gelişmeler ile birlikte mahrem veriler gün yüzüne çıkmaya başlamıştır. Veri çokluğu özellikle de mahremiyet manasında ciddi sıkıntılar meydana getirebilmektedir [Cavoukian,1998].

Veri madenciliği teknikleriyle ortaya çıkarılan gizli bilgiler kötü niyetli kişiler tarafından kullanılabilir. Bu yüzden mahrem bilgiler içeren veriler üzerinde daha dikkatli çalışılmalıdır. Aksi takdirde gizliliğin ortadan kalkması endişesi ile insanlar yanlış veya eksik bilgi vermek gibi davranışlar içerisinde bulunabilirler. Bu durum veri analizi ve bilgi çıkarım işlemleri için uygun olmayan bir ortamın doğmasına sebep olabilir. Çünkü mahremiyet korkusu ile eksik verilen bilgiler, doğruluğu düşük veya yanlış bilgilerin ortaya çıkmasına sebep olacaktır.

Örneğin e-ticaret müşterilerinin tutumunu anlamak için yapılan bir anketin sonuçlarına göre, gizliliğin garanti edilmediği sistemlerde katılımcıların sadece %27'lik bir kısmının doğru bilgi verdiği saptanmıştır. Kalan %73'lük kısım eksik veya yanlış bilgi vermiştir. Bu tür bilgiler ile doğru veri analizlerinin yapılması imkânsız hale gelmektedir [Cranor et al.,1999].

Gerek doğru veri analizi, gerekse mahremiyetin korunması için veri madenciliğinde mahremiyetin sağlanması önemli bir konudur. Bu kapsamda çeşitli teknikler geliştirilmiş ve halen geliştirilmeye devam etmektedir.

Kişilik hakları göz önüne alındığında; toplanan verilerin kimler tarafından toplanacağı, nasıl ve ne süreyle saklanacağı, bu verilere kimlerin hangi koşullar altında erişeceği, bu verilerden yararlı bilgi çıkartılırken hangi teknik ve yöntemlerin kullanılacağı, çıkan yararlı bilginin ticari haklarının kime ait olacağı ve hangi uluslararası kanunlar çerçevesinde değerlendirileceği gibi konuların acilen tartışılması, toplumun bu konuda bilinçlendirilmesi ve gerekli düzenlemelerin hızla hayata geçirilmesi için ivedilikle çalışmalara başlamak gerekmektedir.

Bu kapsamda veri madenciliğinde mahremiyetin sağlanması için bazı teknikler ortaya konulmuştur. Veriler üzerinde çeşitli işlemler yapılması ile mahremiyetin sağlanmasına çalışan *Veri Koleksiyonu ile Mahremiyeti Sağlama* tekniği [Agrawal and Aggarwal, 2001] [Yang et al., 2005] verilerin karartılması, değiştirilmesi, şifrelenmesi, genelleştirilmesi gibi yaklaşımlar önermektedir. Veriler üzerinde değişiklik yapmak yerine, ortaya çıkan sonuçların korunmasına yönelik olarak *Sonuç Kontrolü ile Mahremiyeti Sağlama* tekniği [Wang et al., 2004] [Srikant et al., 2005] kullanılarak sorgu denetlemeleri ve yaklaşık sonuçlar üretme yöntemi ile mahrem verilerin korunması sağlanmaktadır. Farklı sistemlerde mevcut olan verilerin birleştirilmesi sonucu ortaya çıkabilecek mahrem verilerin engellenmesi için ise *Bilgi Paylaşımı ile Mahremiyeti Sağlama* tekniği [Asonov et al., 2004] [Evfimievski et al., 2003] [Goldreich, 2004] ortaya konmuştur. Bu yaklaşım mahrem verilerin dağıtık olarak tutulmasını ve bu verilerin birleştirilmesi esnasında çeşitli teknikler kullanılarak güvenliğinin sağlanmasını esas almaktadır. Bu çalışmalar ile veri madenciliğinde mahremiyetin sağlanması ve verilerin açığa çıkması engellenerek gizli kalması sağlanmaktadır.

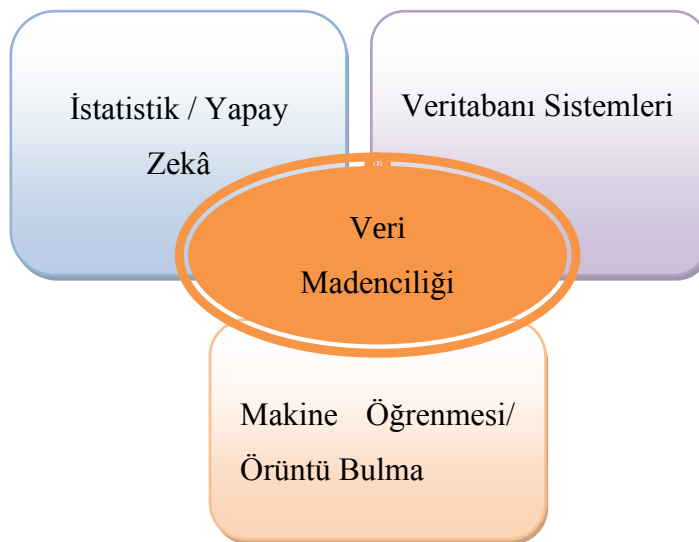
Bu tez çalışmasında veri madenciliğinde mahremiyetin sağlanması için yeni bir yaklaşım ortaya konmaktadır. Veri madenciliği tekniklerinden sınıflandırma tekniği esas alınarak mahremiyeti koruyucu veri madenciliğinin nitelik seçimi ile ne şekilde yerine getirebileceği deneysel olarak ortaya konulacaktır. Tezin organizasyonu şu şekildedir; bu bölümde giriş mahiyetinde bilgiler sunulduktan sonra ikinci bölümde mahremiyet ve veri madenciliği ile ilgili kavramlar açıklanmakta ve veri madenciliğinde mahremiyet ile ilgili yapılan çalışmalar ve tekniklerden bahsedilmektedir. Üçüncü bölümde veri madenciliğinde mahremiyeti sağlamak için ortaya koyduğumuz sistem ve bu sistemin bileşenleri ayrıntılı olarak anlatılmaktadır. Dördüncü bölümde, önerilen sistemin uygulaması olan bir çalışma ve bu çalışmanın adımları gösterilmektedir. Beşinci bölümde ise veri madenciliğinde mahremiyetin sağlanması konusunda ortaya konan çalışmanın sonuçlarından bahsedilmektedir.

2. VERİ MADENCİLİĞİNDE MAHREMİYET

2.1 Veri Madenciliği Nedir?

Gelişen teknoloji, yeryüzünde bugüne kadar hiç görülmemiş boyutlarda bilginin üretilmesini ve bunların veri tabanlarında saklanmasını mümkün hale getirmiştir. Bu gelişim, olağanüstü hacimlere ulaşan veri tabanlarındaki verilerden nasıl faydalanılabileceği problemini ortaya çıkarmıştır. Geleneksel sorgulama ve raporlama araçlarının bu probleme çözüm olamaması, Veri Tabanlarından Bilgi Keşfi (VTBK) veya daha özel bir kavram olarak veri madenciliği kavramını ortaya çıkarmıştır. Veri madenciliği ile ilgili olarak bugüne kadar birçok tanım yapılmış olup bunlardan belli başlı kabul görenler şöyledir. *Veri madenciliği*; veriler içinde yer alan desenlerin, ilişkilerin, değişimlerin, düzensizliklerin, kuralların ve istatistiksel olarak önemli olan yapıların yarı otomatik olarak keşfedilmesidir [Frawley et al., 1991].

Bir başka tanıma göre *Veri madenciliği*; büyük veri tabanlarından, çok net olmayan, üstü kapalı, önceden bilinmeyen ancak potansiyel olarak kullanışlı olabilecek bilginin çıkarılmasıdır [Fayyad et al., 1994]. Yani gizli kalan bilgilerin ortaya konmasında önemli rol oynar ve onların açığa çıkmasına yardımcı olur.



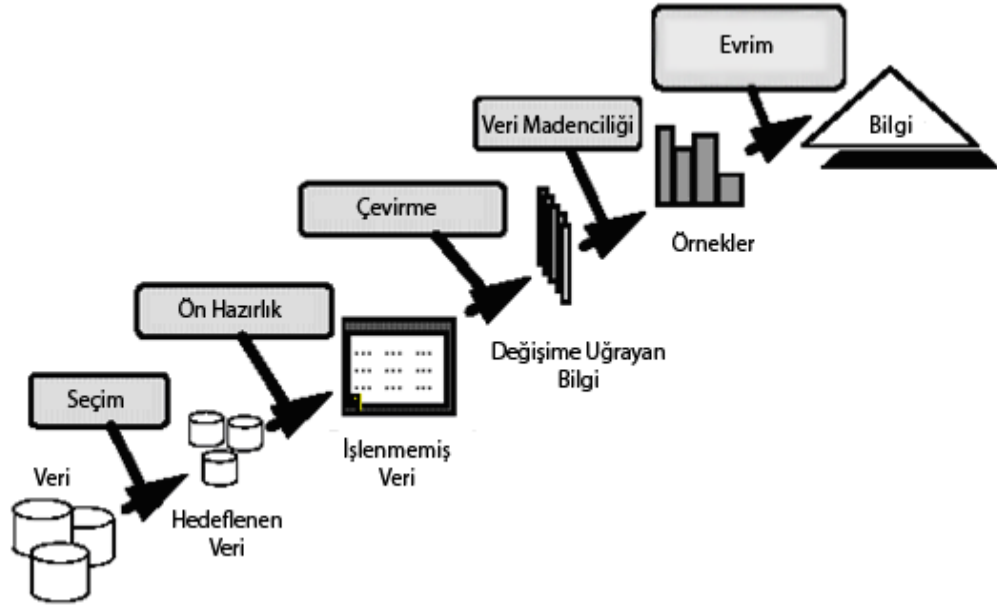
Şekil 2.1. Veri Madenciliğinin Diğer Disiplinlerle İlişkisi

Veri Madenciliği; büyük veri tabanlarında yer alan örüntülerin, birlikteliklerin, anormalliklerin ve çeşitli yapıların yarı otomatik bir sistem ile keşfidir [Berry and Linoff, 2000].

Veri Madenciliği, ortaya çıkış süreci ve kullandığı tekniklerden dolayı disiplinler arası bir özelliğe sahiptir (Bkz. Şekil 2.1.). Veri madenciliğinin temel görevleri olan sınıflandırma, kümeleme, birliktelik kuralı keşfi, sıralı örüntü keşfi ve anormallik tespiti aynı zamanda makine öğrenmesi, yapay zeka ve istatistik disiplinlerinin çalışma alanına girmektedir. Veri madenciliği, diğer disiplinlere ait bu birikimlerin veri tabanı sistemleri ile kullanılmasını sağlamaktadır.

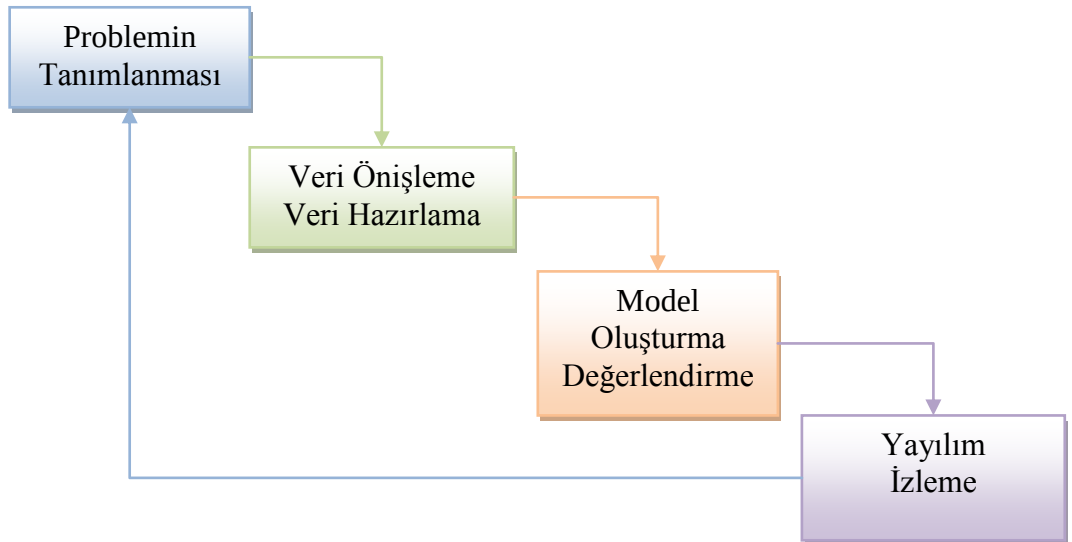
2.1.1 Veri Madenciliği Aşamaları

Veri madenciliği tekniklerinin uygulanabilmesi için öncelikle veriler toplanır. Bu veriler ham veridir ve genellikle veri ambarında tutulur. İlk bakışta bu veriler arasından bir ilişki çıkarabilmek mümkün değildir. Verinin tamamı ile çalışmak performans açısından ciddi problemler oluşturacağı için bu veriler arasından seçim yapılarak hedeflenen veri elde edilir. Veri ambarının küçük parçaları üzerinde çalışılır. Bu veriler ön işleme sokularak işlenmiş veri elde edilir. Bu işlem sırasında veriler arasından ayıklama yapılarak gürültülü ve ilişkisiz veriler temizlenir. İşlenmiş veriler dönüştürme işlemine sokularak artık kolaylıkla işlemler uygulanabilecek hale gelirler. Sonra Veri madenciliği algoritmaları uygulanarak bilgi hakkında örnekler, kümeler, örüntüler ve kurallar çıkarılır. Burada elde edilen sonuçlar önceden keşfedilmeyen veya gizli olan bir bilgiye ulaşmamızı sağlayabilir. Elde edilen örüntüler, kümeler ve sınıflardan yola çıkılarak kararlar verilir ve bilgi kullanılır. Veri madenciliği aşamaları şekil 2.2’de gösterilmektedir [Han and Kamber, 2001].



Şekil 2.2. Veri Madenciliği Aşamaları.

Veri madenciliği çalışmalarının başarılı olabilmesi için sistematik bir işlem sürecinin takip edilmesi gerekmektedir. Genel olarak veri madenciliği işlem süreci; problemin tanımlanması, ön işleme teknikleri ile verilerin hazırlanması, modelin ortaya konulması, değerlendirilmesi ve izlenmesi aşamalarından oluşmaktadır [Tan et al., 2005]: (Bkz. Şekil 2.3.).

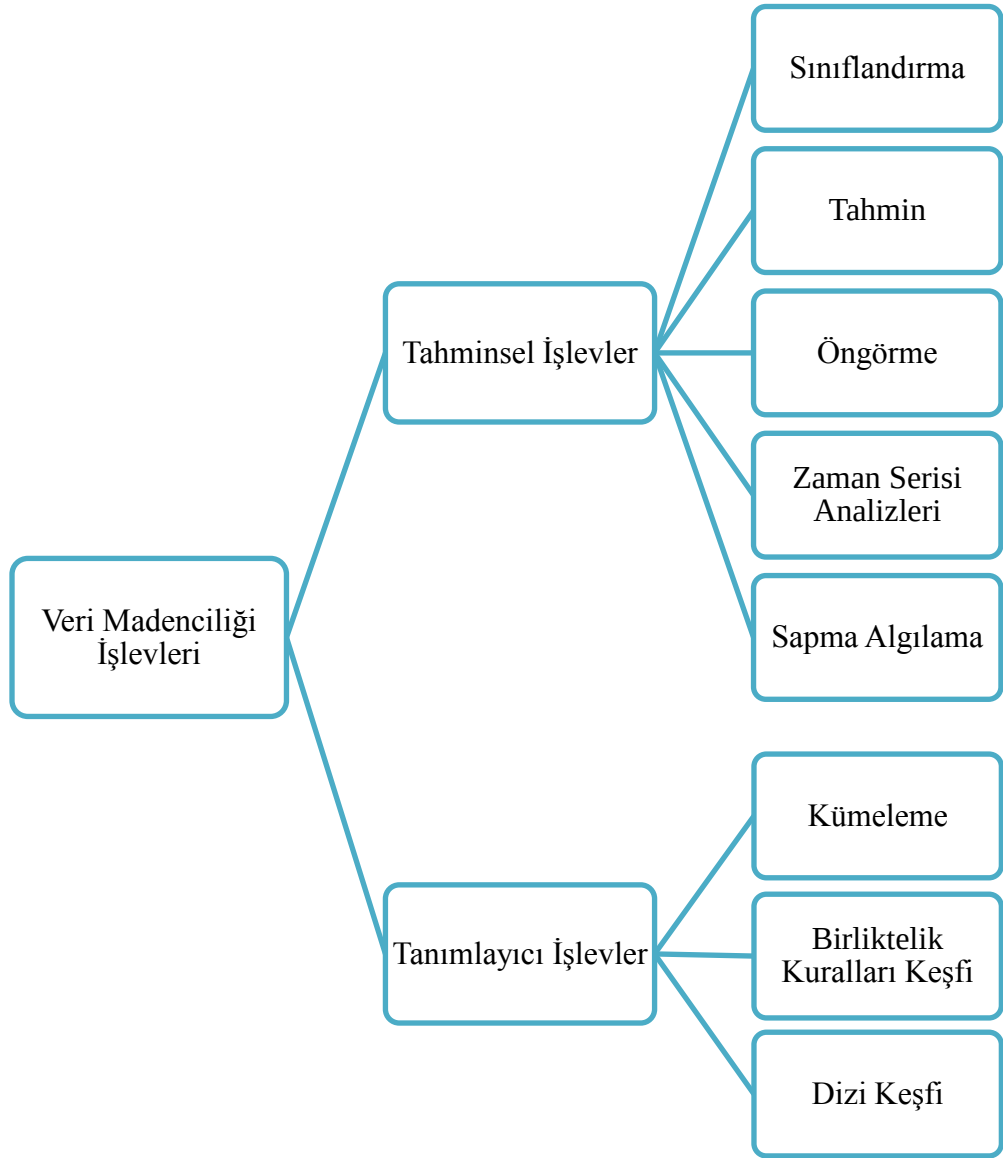


Şekil 2.3. Veri Madenciliği Klasik İşlem Süreci.

2.1.2 Veri Madenciliği İşlevleri

Veri madenciliğinde kullanılan modeller işlevlerine göre, tahminsel (predictive) ve tanımlayıcı (descriptive) olmak üzere iki ana başlık altında incelenmektedir. Tahminsel modellerde denetimli veriler yardımıyla modeller meydana getirilir. Bunlar üzerinden anahtar parametreler ve değişkenler yardımı ile ileride sonuçları bilinmeyen veri kümelerinin sonuçlarının ortaya çıkarılması amaçlanır. Tanımlayıcı modeller karar vermeye yardımcı olma ve eldeki verinin tanımlanması odaklıdır. Eldeki veriler üzerinden belirli örüntü veya izlerin ortaya çıkarılmasını hedeflemektedir. Bu iki ana başlık altında yer alan veri madenciliği modelleri, gördükleri işlevlere göre aşağıdaki gibi sıralanabilir [Tan et al., 2005]: (Bkz. Şekil 2.4.)

- Sınıflandırma
- Tahmin
- Öngörme
- Zaman Serisi Analizleri
- Kümeleme
- Birliktelik Kuralları Keşfi
- Dizi Keşfi
- Sapma Tespiti



Şekil 2.4. Veri Madenciliği İşlevleri.

Sınıflandırma:

İnsanoğlu dünya üzerindeki birçok konu ve olayı daha iyi anlamak ve anlatabilmek için sınıflandırma yapar, kategorilere ayırır ve derecelendirir. Örneğin, canlıları türlerine, doğadaki maddeleri ortak özelliklerine göre ve olayları önem derecelerine göre kategorilere ayırır. Sınıflandırma, veri madenciliği işlevleri arasında en yaygın olanıdır. Sınıflandırma işleminde öğrenme denetimlidir ve hedef değişken değerleri kategorik veya ikili (0/1) değerlerden oluşur.

Sınıflandırma işleminde amaç, yeni karşılaşılan bir girdinin özelliklerinin incelenip, bu girdinin daha önce tanımlanmış olan sınıflardan hangisine ait olduğuna karar vermektir. Gerçek hayatta karşımıza çıkan bazı veri madenciliği sınıflandırma örnekleri arasında; kredi risk puanlama, kampanyalara göre müşteri kitlesinin tespiti, çevrimiçi/canlı/görev kritik sistemler üzerinde yapılan işlemlerde yer alan şüpheli ve yasal olmayan işlemlerin tespiti ve benzerleri sıralanabilir.

Tahmin:

Bu modellerin amacı, sınıflandırma işlevi gibi, bilinmeyen bir hedef değişken için bir değer belirlemektir. Sınıflandırma ile benzer gibi görünmesine rağmen, sınıflandırma işlevinde hedef değişkenlerin kategorik bir değere sahip olması, tahmin işlevinde ise söz konusu değerlerin süreklilik gösteren bir değere sahip olması ile birbirlerinden ayrılırlar. Verilen girdiler sonucunda süreklilik arz eden mesafe, borç miktarı, tahmini süre gibi değerler üretilir.

Öngörme:

Öngörü işlevi sınıflandırma ve tahmin işlevine çok benzer bir yapıdadır. Ancak öngöründe yapılan sınıflandırma, gelecek için tahmin edilen belirli bir davranışa ya da belirli bir değere göre yapılır. Bu işlevi gören modellerde hedef değişken değerleri kategorik veya sayısal olabilir.

Öngörü işleminde yapılan sınıflandırmanın doğru olup olmadığını test etmenin tek yolu “bekle ve gör” prensibidir [Han and Kamber, 2001]. Öngörüye örnek olarak enflasyon tahmini, deprem tahmini, şirketlerin müşteri istek oranlarının tahmini ve otellerin yazlık rezervasyon tahminleri verilebilir.

Zaman Serisi Analizleri:

Zaman serisi, zaman içinde gözlemlenen geçmiş ölçümlerin bir dizisi olarak ifade edilir. Zaman serisi analizi, bir değişkenin zamana bağlı olarak değişen değerlerini inceleyerek gelecekte alacağı değerleri tahmin etmek amaçlanmaktadır.

Zaman serisi analizi ile eğer veri setinde mevcut ise eğilimleri, mevsimsellikleri ve devresellikleri tespit eder.

Eldeki veri kümesinden yola çıkılarak bulunmaya çalışılan hedef değişkenin değeri (sinyal) belli periyotlara (yıl, ay, gün, saat, dakika vb.) göre analiz edilir. Analizde öğrenmede kullanılan veri setinin periyodu ile sonuç değerlerinin verildiği zaman periyotları aynıdır. Örneğin bir satış tahmini için yapılan zaman serisi analizinde haftalık veriler kullanılıyorsa, sonuç değerleri de haftalık satış tahminleri verecektir. Zaman serisi analizi örneğin, makine ve cihazların arıza, bakım ve ömür tahminlerinde, otomatik para çekme makinelerinin nakit yönetiminde kullanılır.

Kümeleme:

Kümeleme, veri kümesindeki verileri belirli kriterlere göre gruplamayı amaçlar. Grup içi benzerliğin maksimum, gruplar arası benzerliğin minimum olacak şekilde gruplandırmada yapılır. Oluşan her gruba “küme” denir. Kümeleme sınıflandırma işlemine benzerlik göstermesine rağmen önceden bilinen sınıfların olmayışı ile birbirlerinden ayrılırlar. Diğer bir ifade ile oluşturulan kümelerin, önceden belirlenmiş bir etiket ile etiketlenmemesi aralarındaki farkı oluşturur. Kümeleme modelleri denetimsiz öğrenme yaklaşımına uygundur.

Birliktelik Kuralları Keşfi:

Veri kümesi içeride sık tekrar eden ve birlikte yer alan öznitelik değerlerinin ve bunların aralarındaki ilişkilerin keşfedilmesidir. Başka bir ifade ile bir işlem kaydında (transaction) bir elemanın meydana gelme olasılığını diğer elemanların meydana gelme olasılıklarından tahmin etmek için kurallar bulunmasıdır.

Birliktelik kuralı keşfi uygulamaları pazarlama ve bankacılık gibi alanlarda karşımıza çıkmaktadır. Pazarlama sektöründe pazar sepeti analizinde yaygın olarak kullanılır. Örneğin bir süpermarket zinciri elinde bulunan işlem kayıtlarından müşteri eğilimlerini tespit etmek ve bunlar için özel kampanyalar düzenlemek istediğinde birliktelik kuralları yöntemini kullanabilecektir.

Dizi Keşfi :

Dizi keşfi veri kümesindeki birbirini takip eden zaman örüntülerinin tespitini amaçlar. Birliktelik kuralı keşfine benzeyen dizi/ardışık keşfinde bir arada yer alan değerlerin aynı anda olması yerine belirli bir zaman aralığı ve sıra ile olmasını inceler.

Sapma Tespiti:

Sapma tespiti bütün popülasyonun genel karakteristiğine uymayan örneklerin bulunmasıdır. Popülasyonun ortak davranış örüntüsünden veya özetinden bir normal davranış profili oluşturulur. Saha sonra bu normal profilden bariz bir şekilde farklılık gösterenler tespit edilmeye çalışılır. Grafik tabanlı, istatistik tabanlı, uzaklık tabanlı ve model tabanlı türleri vardır.

2.1.3 Veri Madenciliğinin Uygulama Alanları

Veri madenciliğinin uygulama alanları çok geniştir, çünkü her alanda veriler birikmektedir. Bu verilerin işlenmesi ve kullanılabilecek bilgiler elde edilmesi için veri madenciliği gereklidir. Genel olarak veri madenciliğinin uygulandığı alanlar:

Bilim: Astronomi, Biyoinformatik, İlaç Keşfi

Web: Arama Motorları, Web Analizi, Metin Madenciliği (haber, e-mail gibi)

Devlet: Terörle Mücadele, Kanun Yaptırımı, Vergi Kaçakçılarının Profilinin Çıkarılması

İş Hayatı: Reklam, CRM (Müşteri İlişkileri Yönetimi), Müşteri Modelleme, Yatırım Değerlendirme ve Karşılaştırma, Sağlık, Üretim, Spor/Eğlence, Telekom, Hedef Pazarlama, Bankacılık

Bunun gibi daha birçok alanda veri madenciliğinden faydalanılabilir. Veri madenciliği uygulamaları sonucunda bu sektörlerin kazanabileceği faydalar:

- Sektörler, müşterilerinin örüntülerini çıkararak müşterilerine daha iyi hizmete verebilirler. Hem müşteri memnuniyeti sağlamış olurlar hem de daha fazla satış yapabilirler.

- En iyi müşteriler belirlenebilir ve müşteri bölümlenmesi yapılarak bu müşterilere yönelik yeni satış stratejileri oluşturulabilir.

- Kuruluşlar tarafından düzenlenecek çeşitli kampanyalarda mevcut müşteri kitlesinin seçimi ve bu müşterilerin davranış özelliklerine yönelik kampanya şartlarının oluşturulması sağlanabilir.

- Kuruluşlar veri analizi ile müşterilerini kişiselleştirilmiş ürün ve hizmetler hakkında bilgilendirebilirler.

- Günümüzde var olan rekabet ortamında kuruluşların hızlı ve kendi için en doğru kararı almalarını sağlayabilirler.

- Özellikle finans sektöründe mevcut müşterileri bölümlere ayırıp, kredi risk davranış modelleri oluşturarak, yeni başvuruda bulunan müşterilere karşı riskin minimuma indirilmesi sağlanabilir.

- Mevcut müşterilerin ödeme performansları incelenip kötü ödeme performansı gösteren müşterilerin ortak özellikleri belirlenerek, benzer özelliklere sahip tüm müşteriler için yeni risk yönetim politikaları oluşturulabilir.

- Spor alanında istatistiksel veriler elde ederek sporcuların performanslarını artırıcı çalışma programları hazırlanabilir.

- Suçluların davranış profilleri çıkarılarak benzer suçları işleyen kişiler tespit edilebilir. Aynı şekilde saldırı hazırlığındaki teröristlerin örüntüleri keşfedilerek terör saldırılarının önceden engellenmesi sağlanabilir.

- Sağlık alanında ise doğru zamanda karar almak hastanın hayatı için çok önemli olduğu aşikârdır. Hastanın yaş, cinsiyet, kilo gibi verileri, laboratuvar verileri, hastanede uygulanan tedavi yöntemlerini başarı oranı, kullanılan cihazların yeterliliği gibi veriler incelenerek hasta için en iyi tedavi yöntemleri bulunabilir.

- Hastanelere yapılan tedavi taleplerinin bölgelere, zamana ve ihtiyaca göre değerlendirmesi salgın hastalık riskinin ilk aşamada tespiti, kontrolü ve kaynak planlama açısından faydalı olur.

- Birçok sektörde müşteri kaybı önemli sorunlardandır. Kuruluşlar hangi müşterilerini kaybedebileceklerini önceden belirleyebildikleri takdirde bu müşterilerini elde tutma amaçlı stratejiler geliştirebilir, düşük maliyetli ve etkili kampanyalar düzenleyebilirler.

- Arama motorları Web'deki link yapılarından ve sayfa içeriklerinden yararlanarak aranan dokümanı bulabiliyor.

Bu kadar geniş bir alana hitap etmesi ister istemez veri madenciliğini olmazsa olmazlar arasında olmasını sağlıyor. Veri madenciliğinin gizli bilgilerinde ortaya çıkmasını sağlaması bazen dezavantaj olmaktadır. Özellikle insanın içerisinde bulunduğu verilerde mahremiyet içeren bilgilerin saklı kalması gerekmektedir. Yalnız veri madenciliği kullanılarak ister istemez bu bilgilerin açığa çıkmasına sebep olunabilir. Bu durumun gerçekleşmesi istenmez. Bu yüzden veri madenciliği çalışmalarında mahremiyetin sağlanması için bazı çalışmaların yapılması gerekmektedir.

2.2 Mahremiyet

Mahremiyet kavramı ilk kez 1890 yılında Amerikalı yargıç Brandeis tarafından "bireyin yalnız bırakılma hakkı" olarak tanımlanmıştır. O zamandan günümüze kadar mahremiyet kavramı pek çok kez gündeme gelmiştir; ancak sınırları net olarak çizilememiştir. Özgürlüklerin geniş olması ve sınırlandırılmaması bunun en önemli nedenidir [Manning, 1997].

Bireylerin kişisel mahremiyetlerini koruma ve mahremiyetlerine saygı gösterilmesini talep etme hakkı vardır. Bu husus Avrupa İnsan Hakları Sözleşmesi'nin *Özel hayatın ve aile hayatının korunması* başlıklı 8. Maddesinde şöyle ifade edilmiştir [inhak-bb.adalet.gov.tr, 2006]:

1. Herkes özel ve aile hayatına, konutuna ve haberleşmesine saygı gösterilmesi hakkına sahiptir.

2. *Bu hakkın kullanılmasına bir kamu otoritesinin müdahalesi, ancak ulusal güvenlik, kamu emniyeti, ülkenin ekonomik refahı, dirlik ve düzenin korunması, suç işlenmesinin önlenmesi, sağlığın veya ahlakın veya başkalarının hak ve özgürlüklerinin korunması için, demokratik bir toplumda, zorunlu olan ölçüde ve yasayla öngörülmüş olmak koşuluyla söz konusu olabilir.*

Özel hayat, bireylerin tüm dış etkilerden ve baskılardan uzak, kendi başına özgürce faaliyetlerini sürdürdüğü ve dokunulmazlığı bulunan yaşamsal bir alandır. Toplumsal bir varlık olan insanın, birçok faaliyette toplumla birlikte onun bir parçası olarak hareket etmesine rağmen, her bireye kendi kişiliğini tüm saflığı ile yansıtabileceği özel bir alanı anayasal özgürlük olarak tanımak temel insan haklarının gereğidir.

Burada kişisel veri kavramı ortaya çıkmaktadır. Peki, bu mahremiyeti içeren bu kişisel veri nedir? Kişisel verilerin neler olduğunun genel tanımı “Kişisel Verilerin Korunması Kanun Taslağı”nda şu şekilde yapılmıştır: *Kişisel veri, belirli veya kimliği belirlenebilir bir kişiye ilişkin bütün bilgilerdir.* 1995 yılında yayınlanmış olan Directive 95/46/EC numaralı Avrupa Birliği direktifi kişisel verinin tanımını ise [Duji, 1996]: *Kim olduğu belli veya belirlenebilen bir gerçek kişiye ait tüm bilgilerdir. Belirlenebilen bir kişi doğrudan veya dolaylı olarak bir kimlik numarası referansından veya kişiye ait fiziksel, psikolojik, ekonomik, kültürel veya sosyal bilgilerden yola çıkarak tespit edilebilen bir kişidir.*

Kişisel Veriler: Genel yaşam ve sektörsel mahremiyet ihtiyaçlarından yola çıkılarak aşağıdaki veriler kişisel veri olarak değerlendirilebilecek verilere örnektir:

Yaşam şekline ilişkin kişisel veriler: Ayrımcılığa uğramamak ve haysiyetin korunmasıyla ilişkili olarak dinsel inanç, cinsel tercih, etnik köken, suç geçmişi, politik eğilimler ve kişisel özel aktivitelere ilişkin bilgiler hassas bilgilerdir.

Finansal kişisel veriler: Suçlular tarafından suiistimale ve kimlik hırsızlığına hedef olmamak için kişinin mali varlığı, sahip olduğu hisse ve hesaplar, borçları yaptığı alışverişler, kredi kartlarına ilişkin veriler kritiktir. Ayrıca bu bilgiler kişinin nerede ve kimlerle bulunduğu, sağlık bilgilerine ilişkin bilgiler de ortaya

çıkartılabileceğinden ve varlık bilgisinin toplumsal açıdan da özel sayılmasından dolayı hassastır.

İnternet kullanımına ilişkin kişisel veriler: e-posta'lar, internet siteleri ile paylaşılan kişisel veriler mahrem olarak değerlendirilebilir. Ayrıca internet kullanımından kaynaklanan saldırı yüzeyi sosyal paylaşım siteleri, dosya saklama hizmeti, internet erişimine ilişkin iz kayıtlarının hizmet sağlayıcı ve sunucu sahipleri tarafından tutulabiliyor olması nedenleriyle artmaktadır.

Sağlıkla ilgili kişisel veriler: Sağlık verileri kişilerin iş güvenliğini, toplum içindeki statüsünü ve sigorta kapsamını etkileyebileceğinden hassas verilerdir. Ayrıca sağlık verileri kişilerin sosyal yaşantısı ve psikolojik durumları hakkında bilgi edinilmesine neden olabilir. Biyometrik veriler de kişisel veriler arasındadır.

Politik kişisel veriler: Toplum barışı ve seçme özgürlüğü açılarından politik veriler mahrem verilerdir.

Mahremiyet her ne kadar bireylere ait özel bilgiler diye tanımlansa da kurum ve kuruluşları da ilgilendiren bir kavramdır. Bu bakış açısıyla olaya bakıldığında mahremiyet kişisel ve kurumsal olmak üzere ikiye ayrılır.

Kurumsal mahremiyet, o kuruma ait gizli kalması gereken verileri içerir. Söz konusu kurumun faaliyet alanı ve faaliyet gösterdiği ortama göre kurumsal verilerin kritiklik derecesi değişebilmekle birlikte aşağıdaki verilerin kurumlar açısından genel olarak kritik olduklarını söyleyebiliriz:

Stratejik Planlar: Rekabet açısından kritik olabilecek stratejik yatırım, ürün, ya da faaliyet değişimi planları hassas olabilir.

Mali Bilgiler: Özellikle halka açık şirketlerde pay sahipleri ve kamunun eşit bilgi alma hakkı dolayısıyla mali bilgilerin kontrollü biçimde oluşturulması ve yayınlanması gerekmektedir. Özel kurumlarda bu bilgiler rekabet açısından kritik olabilir.

Personel Bilgisi: Hem kişisel verilerin korunmasına ilişkin düzenlemeler hem de rekabet veya kurumu hedef alabilecek sosyal mühendislik saldırılarına karşı personel bilgileri ve iletişim bilgileri hassastır. Ayrıca maaş bilgileri de kurum içi huzurun ve kişisel mahremiyetin sağlanması için kritiktir.

Müşteri Bilgileri: Özellikle denetim, sağlık, telekomünikasyon, finans hizmeti veren kurumlar için müşteri bilgileri hem kişisel bilgilerin korunmasına ilişkin kanunlar hem de sektörsel düzenlemeler nedeniyle son derece hassastır. Bu bilgilerin korunmasında gösterilecek yetersizlik ayrıca kuruma ticari olarak da olumsuz etki yapacaktır.

Kritik Sistem, Prosedür ve Erişim Bilgileri: Kurumu hedef alan taraflara yardımı olabilecek kritik altyapı ve prosedür bilgileri hassastır.

2.3 Veri Madenciliğinde Mahremiyet Kavramı

Veri madenciliği ve mahremiyet birbirine zıt gibi görünen, birinin devamı diğerine bağlı olan iki önemli konudur. Veri madenciliğinin doğası gereği gizli kalmış bilgileri veri yığınlarından çıkarma yeteneği vardır, Mahremiyet ise gizli kalması gereken bilgilerin korunmasını zorunlu kılar. Burada bir zıtlık var olsa da bu kavramların birlikte ele alınması ile bir uyumun yakalanması söz konusudur. Veri madenciliği ile mahremiyet arasındaki sınırı doğru şekilde çizerek bu iki konu birleştirilebilir. Daha doğrusu veri madenciliği yaparken mahremiyet korunarak hem gizli bilgilerin bir kısmı bulunabilir hem de bazı gizli bilgiler korunabilir.

Eğer veri analizleri sırasında mahremiyet korunmazsa istenmeyen durumlarla karşılaşmak mümkündür. Gizli kalması gereken bazı verilerin paylaşılması sonucu meydana gelen önemli olayların birkaçı şöyledir:

1993 yılında Amerika’da Maryland eyaletinde yaşayan insanlara daha iyi bir sağlık hizmeti vermek için tüm sağlık kayıtları elektronik ortama aktarılmıştır. Bir banker bu verilere erişerek kayıtları inceleyip kendi müşterilerinin hastalıklarını tespit etmiş ve ölümcül hastalığı olan müşterilerinden borçlarını ödemesini istemiştir [Etzioni, 1999].

1989 yılında Amerikalı kadın oyuncu Rebecca Lucile Schaeffer, bir hayranının motorlu taşıtlar dairesinden ehliyet bilgilerine ulaşması sonucu ev adresini elde edilmesiyle evinin önünde vurularak öldürülmüştür. Bu olay sonrası böyle bir durumla karşılaşmamak için 1994'te Driver's Privacy Protection Act kanunu çıkartılmıştır [epic.org/privacy/drivers,1994][Karras, 1999].

1988 yılında Robert Bork ismindeki yargıç yüksek mahkemeye aday gösterilmiştir. Ancak bir gazeteci yargıcın film satın aldığı video şirketinden ne tür filmler kiraladığını öğrenerek haber yapmış ve bunu yargıcın aleyhine kullanarak bazı güçlüklerle karşılaşmasına sebep olmuştur. Bu olaydan sonra da Video Privacy Protection Act kanunu çıkartılmıştır [epic.org/privacy/vppa, 1988].

Bunlar gizli kalması gereken bilgilerin açığa çıkartılması sonucu geçmişte gerçekleşmiş olaylardan sadece birkaçıdır. Günümüzde de böyle olaylar ne yazık ki devam etmektedir. Özellikle facebook gibi sosyal ağlarda insanların özel bilgilerini ve sosyal arkadaşlık çevrelerini deşifre etmesi istenmedik durumlarla karşılaşmalarına sebep olmaktadır.

Veri madenciliği çalışmalarıyla mahrem verilerin istenmeden ortaya çıkarılabilmesi mümkün hale geldiğinden veri madenciliği uygulamalarında mahrem bilgilerin saklanması ve gizlenmesi esas alınmalıdır. Aksi takdirde bu bilgilere ulaşan kötü niyetli kişiler, diğer kişi ve kurumlara doğrudan veya dolaylı olarak ciddi zararlar verebilirler. Bu kişiler elde ettikleri mahrem bilgileri kullanarak X kişi veya X kurum adına birçok suç işleyebilir. Sonuç olarak veri madenciliği ve mahremiyet birlikte kullanılmalıdır. Yoksa gizli kalması gereken bilgilerin açığa çıkması endişesinin önüne geçilmesi mümkün olmayacaktır.

2.4 Veri Madenciliğinde Mahremiyet Ölçütleri

Mahremiyetin daha net bir şekilde ortaya konması için bazı ölçütlerden bahsedebiliriz. Bu ölçütlere uyarak mahrem verilerin açığa çıkmasını önleyebiliriz. Mahrem veriler üzerinde çeşitli değişiklikler yapılması veya verilerin sistemlerde kısıtlı bir şekilde paylaşılması gibi yöntemler ile mahrem verilerin açığa çıkması

engellenmiş olabilir [Verykios et al., 2004]. Bu kapsamda *Bilgiyi Sınırlandırma*, *Yeterli Bilgi Bilme*, *Açığa Çıkmasını Önleme* ve *Gerçeği Saklama* şeklinde ifade edilen ölçütler olarak ortaya konulabilir [Clifton et al., 2004].

2.4.1 Bilgiyi Sınırlandırma

Mahremiyetin daha etkin bir şekilde ortaya konması için verinin bazı sınırlara çekilmesi gerekebilir. Bu sınırlar sayesinde veriler bazı değerlere yakınsanarak mahrem bilgilerin ortaya çıkması engellenmiş olur. Veriye gürültü eklenerek ya da veri genelleştirilerek sınırlı veri haline dönüştürülür [Agrawal and Aggarwal, 2001].

Bu yaklaşım kullanılarak veri madenciliğinde mahremiyeti sağlayan çeşitli yöntemler geliştirilmiştir. Amaç burada veriyi belirli kurallar ile kısıtlayarak içerisinde barındırdığı gizli kalması gereken bilgilerin ortaya çıkmasını engellemektir.

2.4.2 Yeterli Bilgi Edinme

Yeterli bilgi edinme ölçütü ile gereksiz verilerin sistemden uzak tutulması sağlanır. Bu sayede gereksiz bazı bilgilerin etkisi ile ortaya çıkacak mahrem veriler engellenmiş olur. Bu nedenle özellikle veri madenciliği uygulamalarında yeterli bilgiye sahip olma ölçütü göz önüne alınarak uygulamalar gerçekleştirilmelidir. Böylece mahrem verilerin ortaya çıkmasını engellemeye yardımcı olabilecek adımlar atılmış olur [Clifton et al., 2004].

Yeterli bilgi edinme ölçütü veriye erişim kontrolünün olmasını da sağlar. Veriye erişmek için erişim gerekçesi ve giriş yetkisi talep edilir. Bu sayede veriye erişim kontrol altına alınarak daha güvenli bir ortam sağlanır.

Yeterli bilginin daha fazlasına sahip olmak güvenlik açıklıkları meydana getireceğinden gereğinden fazla olan verinin ayrıştırılarak veya sisteme dâhil edilmeyerek kullanılmalıdır.

2.4.3 Açığa Çıkmasını Önleme

Gizli kalması gereken verilerin açığa çıkmasını önlemek gereklidir. Bu veri mahrem bilgi, özel bir kural veya gizli kalması gereken bir bilgi olabilir. Özellikle veriler üzerinde etkili olan veri madenciliği teknikleri kullanılarak bu bilgilerin ortaya çıkartılması mümkündür. Veri madenciliği sonucundan açığa çıkabilecek mahrem verilerin saklı kalması için sonuçlar üzerinde bazı işlemler yapılarak mahremiyet sağlanabilir. Mahremiyeti sağlama ölçütlerinden biri olan verilerin açığa çıkmasını önlemek için sınıflandırma metodu kullanmak etkili yöntemlerden biridir [Clifton et al., 2004].

Veri madenciliği sonuçları üzerinde yapılan sorguların kontrolden geçilerek sorguların istenmeyen sonuçları veya bilgileri döndürmesi engellenebilir. Böylece uygulama sonuçlarıyla istenmeden ortaya konan mahrem bilgilerin açığa çıkması önlenmiş olur.

2.4.4 Gerçeği Saklama

Mahremiyeti sağlayabilecek ölçütlerden biride verilerin saklanmış bir şekilde ifade edilmesidir. Veriler çeşitli teknikler kullanılarak farklı formatlara dönüştürülebilir. Kriptografi teknikleri kullanılarak veriler şifreli bir şekilde saklanabilir. Veri madenciliği sunucusunda bu şifrelenmiş veriler çözülerek işleme alınır ve verilerin güvenli bir ortamda saklanması sağlanır.

Farklı bir yöntem olarak veriler farklı değerlere dönüştürülerek gerçek değerleri saklanabilir. Örneğin bir yaş değeri belli bir aralığa çekilerek gerçek değerinin başkaları tarafından bilinmesini engellemek mümkündür [Clifton et al., 2004] [Verykios et al., 2004]. Yöntem ne olursa olsun amaç verilerin gerçek değerleri dışında ifade edilerek mahremiyetin sağlanmasıdır.

2.5 Yapılan Çalışmalar

Mahremiyetin korunması için, veri madenciliğinde gizlilik içeren veriler üzerinde titiz ve dikkatli çalışılması gerekmektedir. Bu gereksinim veri madenciliği uygulamalarında mahremiyetin sağlanmasını gerçekleştirecek bazı tekniklerin ortaya çıkmasını sağlamıştır.

Bu teknikleri üç temel kategoriye ayırarak inceleyebiliriz.

2.5.1 Veri Koleksiyonu ile Mahremiyeti Sağlama

Temel hedef veriler üzerinde işlemler yaparak mahremiyetin korunmasıdır. Bunlar çeşitli tekniklerle gerçekleştirilmektedir [Zhang et al., 2007] [Verykios et al., 2004] [Aggarwal et al., 2008].

Veri karartma tekniği, veriye gürültü ekleyerek mahrem bilgilerin açığa çıkmasını engellemeye çalışan bir tekniktir. Bu sayede mahrem veriler değiştirilerek yâda karıştırılarak veri madenciliği uygulamasında kullanılır. [Verykios et al., 2004] [Aggarwal et al., 2008].

Veri; rastgele gürültü (örneğin yaşı 23'den 30'a değiştirilmesi, bölgenin İstanbul'dan Ankara yapılması gibi.) eklenerek ya da belli kurallara dayalı genelleştirmelerle (örneğin 23 yaş değeri 21-25 değerine genellenir, İstanbul verisi Türkiye olarak genellenir.) karartılır.

Veri değiştirme yöntemi, farklı veri sağlayıcıların olduğu bir sistemde her veri sağlayıcının kendi verisini diğer veri sağlayıcı ile değiştirerek veriyi karıştırması şeklinde yerine getirilir. Bu şekilde veri sahibinin gizlenmesi hedef alınır. Yalnız bu teknik güvenli ortamlarda kullanılmalıdır. Çünkü veri değişimleri sonucu mahrem veriler sistemde kolaylıkla açığa çıkartılabilmektedir [Zhang et al., 2007].

Şifreleme metodu, verilerin toplanması sırasında oluşturulan anahtar değerler ile şifrelenir ve veri madenciliği sunucusunda bu şifrelenmiş veriler çözülerek işleme alınır ve verilerin güvenli bir ortamda saklanması sağlanır [Verykios et al., 2004].

K-anonymity ve L-diversity modelleri, dolaylı yollardan mahrem bilgilerin ortaya çıkmasını engellemek için kullanılan modellerdir. Bazı kayıtların bir araya gelmesi sonucu mahrem bilgilerin açığa çıkması gerçekleşebilir. K-anonymity metoduyla veriler, gizleme ve genelleştirme teknikleri ile ifade edilir. L-diversity metodu ise *k*-anonymity modelinin zayıf veya eksik taraflarını örtmek için kullanılır [Aggarwal et al., 2008].

2.5.2 Sonuç Kontrolü ile Mahremiyeti Sağlama

Veri madenciliği uygulamalarının sonuçları üzerine çalışmalar yapılarak mahrem bilgiler ortaya çıkartılabilmektedir [Zhang et al., 2007]. Böylece uygulama sonuçlarıyla istenmeden ortaya konan mahrem bilgilerin açığa çıkması önlenir.

Sorgu denetleme yöntemi, veri madenciliği sonuçları üzerinde yapılan sorguların kontrolden geçirilerek sorguların çalışması denetlenir. Eğer sorgu mahrem verilerin açığa çıkmasını sağlamıyorsa sorgu çalıştırılır. Bu işlem veri madenciliğinin kısıtlanmasına sebep olsa da mahremiyeti sağlamada etkin rol oynar.

Veritabanı yaklaşım tekniği, sistemdeki verilerin karıştırılmış olarak tutulmasını ve sorguların yaklaşık sonuçları içermesini önerir. Bu sayede mahrem veriler korunur. Buradaki verilerin karıştırılması olayı verilerin toplanması aşamasında olabileceği gibi sorguların çalıştırılması zamanında da olabilir [Zhang et al., 2007].

2.5.3 Bilgi Paylaşımı ile Mahremiyeti Sağlama

Farklı yerlerde mevcut olan verilerin birleştirilmesi, ortaya özel verilerin çıkmasına neden olabilir. Örneğin hastane kayıtları ile cep telefon satışlarının tutulduğu ayrı sistemlerin verilerinin paylaşılması sonucu hastane kayıtlarındaki

hastaların hangi model telefon kullandıkları ve aynı telefonu kullanan insanların hangi ortak hastalıklara sahip olduğunun bilgisine erişilebilir [Vaidya et al., 2004].

Güvenli Çok Taraflı İletişim (GÇTİ) (Secure Multiparts Communication) metodu, verilerin farklı veri tabanlarında saklanarak sonradan birleştirilmesini esas almaktadır. Böylece güvenlik ve doğruluk ideal seviyede yakalanmaktadır [Clifton et al., 2004] [Vaidya et al., 2004]. Burada kriptografi kullanılarak daha güvenli bir ortam sağlanabilir [Clifton et al., 2004].

Özellikle internet ortamında saklı kalması gereken veriler birden fazla kaynak arasında iletişim yoluyla paylaşılıp, dağıtık olarak tutulabilmektedir. Bu yapıda kullanılan şifreleme protokolü bir sistem ile diğer sistemin haberleşmesinde kriptografi kullanılarak iletişim gerçekleştirme tekniğidir [Aggarwal et al., 2008]. Matematiksel bir yaklaşım ile verilerin şifrelenerek paylaşılması ve sonra bu verilerin tekrar eski hallerine getirilmesini içerir. Anahtar değeri iki sistem arasında paylaşılıp şifrelenen veri çözülerek veri madenciliğinin uygulanması sağlanılır [Verykios et al., 2004].

İletişim protokolünü kullanan sistemlerin verilere erişiminin engellenmesi amaçlanmaktadır. Bu yaklaşım yarı güvenilir kısıtlama ve kötü niyetli kısıtlama olmak üzere iki tekniği içermektedir.

Yarı güvenilir kısıtlama yöntemi, mahrem verilere ulaşmak isteyen kişilerin tasarlanan iletişim protokolünü kullanacağı varsayılmaktadır. Bu protokolde mahrem veriler çeşitli tekniklerle saklanmaktadır veya iletişim için farklı bir yol kullanılarak mahrem veriler saklı tutulmaya çalışılmaktadır. Genel olarak RSA şifreleme algoritması kullanılır [Zhang et al., 2007]. Bu sayede veri madenciliği sunucusu işleyeceği veriyi çözerek işler ve şifreleme anahtarını bilmeyen sistemlerden mahrem veriler saklanmış olur.

Kötü niyetli kısıtlama yöntemi, mahrem verileri elde etmek isteyen kişilerin veri madenciliği sonuçları ile de ilgilenmesi söz konusudur. Burada veri madenciliği sonuçlarının da kısıtlı erişimi sağlanarak kötü niyetli kişilerin ulaşımı engellenmeye çalışılır [Zhang et al., 2007].

3. ÖNERİLEN YÖNTEM

Günümüzde teknolojinin ve internetin kullanımının yaygınlaşması sonucu elektronik ortamlarda kişisel ve kurumsal veriler gün geçtikçe artmaktadır. Bu verilerin büyük bir kısmı bizim kontrolümüz dâhilin de olmayan finans kuruluşları, gsm şirketleri, e-ticaret siteleri, devlet kurumları gibi yerler tarafından kayıt altına alınarak saklanmaktadır.

Veriler tek başlarına anlamsız olabilirler; fakat üzerlerinde analizler ve çalışmalar yapıldığında gizli kalması gereken bilgilerin ortaya çıkması sağlanabilmektedir. Özellikle veri madenciliği kullanılarak büyük ölçekli veri alanları içerisinde anlamsız gibi görünen veriler üzerinde çalışmalar yapılarak anlamlı ve keşfedilmemiş bilgiler ortaya konulabilmektedir. İster istemez bu bilgilerin içerisinde mahrem verilerin ortaya çıkması kaçınılmaz bir durumdur.

Bu açığı kapatmak için veri madenciliği uygulamalarında mahrem bilgilerin ortaya çıkmasını önleyecek metotlar veya teknikler kullanılmalıdır. Bu sayede veri madenciliği çalışmaları ile mahremiyetin korunması birlikte gerçekleştirilecek ve gizliliğin ortaya çıkması konusunda var olan endişeler azaltılacaktır.

Bu tez çalışmasında veri madenciliğinde mahremiyetin sağlanması için yeni bir yöntem ortaya konmaktadır. Bu yöntem veri madenciliğinde mahremiyeti sağlayan çalışmalara benzer, yeni ve etkin bir yöntemdir.

Önerdiğimiz yeni yöntemde; mevcut veri seti içerisinde mahrem ve mahrem olmayan nitelikler seçilerek en az sayıda mahrem veriyle en fazla bilgi nasıl elde edilir sorusuna cevap aranmaktadır. Böylece, veri madenciliği ile mahremiyet arasında bir arakesit bulunmaya çalışılmaktadır. Bu arakesit noktası, en az mahrem nitelik, en yüksek veri madenciliği başarısıdır. Veri madenciliği tekniği olarak sınıflandırma algoritmaları kullanılmaktadır. En iyi sınıflandırma sonucunu veren nitelikler bu çalışmada bulunarak bunlar arasındaki mahrem, mahrem olmayan ayrımı ortaya konulmasına çalışılmaktadır.

3.1 Yöntem Bileşenleri

Önerilen sistem *İşlenecek Veri*, *Öznitelik Seçme Adımı*, *Seçilmiş Veri Setleri* ve *Sınıflandırma Adımı* bileşenlerini içermektedir.

3.1.1 İşlenecek Veri

Veri madenciliği veriler üzerinde çalışır. Verilerin olmadığı bir yapıda veri madenciliği tekniklerinin uygulanması söz konusu değildir. Dolayısı ile işlenecek veri bu çalışmanın en temel yapı taşıdır. İçerisinde mahrem verilerin bulunduğu veriler üzerinde yapılan çalışmalar ise veri madenciliğinde mahremiyetin açığa çıkması problemini doğurmaktadır.

Veriyi sunan nitelikler mahrem ve mahrem olmayan nitelikler diye ikiye ayrılmaktadır. Bu sistemde önemli olan nitelikler gizli kalması gereken mahrem niteliklerdir. Bu niteliklerin dikkate alınması gereklidir. Dikkatli bir şekilde mahrem verilerden ayrıştırılmış veriler sayesinde mahremiyetin ortaya çıkması engellenmiş olacaktır.

3.1.2 Öznitelik Seçme Adımı

Sınıflandırıcıya verilecek öznitelik sayısı, tüm bileşimleri denemeye elverişli değilse, sınıflandırıcının işini kolaylaştıracak bir ön işleme gerek duyar. Sınıf bilgisinin elde edileceği özniteliklerden bazıları bu bilgiyi içermiyor olabilir. Bu özniteliklerin kullanılması bir anlam ifade etmeyeceğinden sınıflandırma işleminden önce çıkartılarak işleme alınması önlenir [Huang et al., 2007]. Bu sayede sınıflandırma gücü olmayan öznitelikler ayıklanmış olur. Aksi takdirde öznitelik uzayının boyutu artar ve sınıflandırıcı için boyut laneti de sınıflandırıcının başarısını azaltıcı etki yapar. Bu nedenlerden ötürü öznitelik seçme kullanılır.

Her özneliğin elde edilmesi toplam sistem için fazladan bir işlem yükü demektir ki bu da işlem zamanını arttırır. Gerçek zamanlı uygulamalarda bu sorun da

kullanılan öznitelik sayısının sınırlandırılmasını gerektirir. Bu yüzden öznitelik seçmeye ihtiyaç duyulur.

Özellikle mahrem verilerin sınıflandırılmasında sınıflandırma başarısını çok fazla artırmayan mahrem niteliklerin belirlenerek çıkartılması hem mahrem verinin veri madenciliğine girerek oluşabilecek gizli verinin açığa çıkma kaygısını önlememize yarayacak hem de sınıf bilgisi barındırmayan verilerin sınıflandırma işleminden uzak tutulması sağlanmış olacaktır.

3.1.3 Seçilmiş Veri Setleri

Öznitelik seçimi ile seçilmiş niteliklerden oluşan dizilerdir. Bu veri seti, gizli kalması gereken mahrem verilerin eksiltilmesi ile elde edilir. Böylece mahrem niteliklerin azaltıldığı veri setleri elde edilmiş olur.

Buradaki veriler kullanılarak sınıflandırma işlemi gerçekleştirilir. Seçilmiş veri setleri sayesinde mahremiyetin sağlanması hedeflenmektedir. Çünkü veri setlerimiz mümkün olduğu kadar mahrem verilerden ayrıştırılmış bir şekilde ifade edilmektedir. Bu veri setlerinin veri madenciliği uygulamasında kullanılması ile veri madenciliğinde mahremiyet sağlanmış olacaktır.

3.1.4 Sınıflandırma Adımı

Veri içerisinde öznitelik seçimi ile elde edilen seçilmiş nitelik setlerine veri madenciliği sınıflandırma işleminin uygulandığı bileşendir. Nesneleri bir veya bir kaç önceden tanımlanmış kategoriye atama anlamına gelen sınıflandırma, veri sınıflarının ortaya konmasında veya gelecek veri eğilim tahmin modellerinin oluşturulmasında kullanılan bir veri analiz yöntemidir.

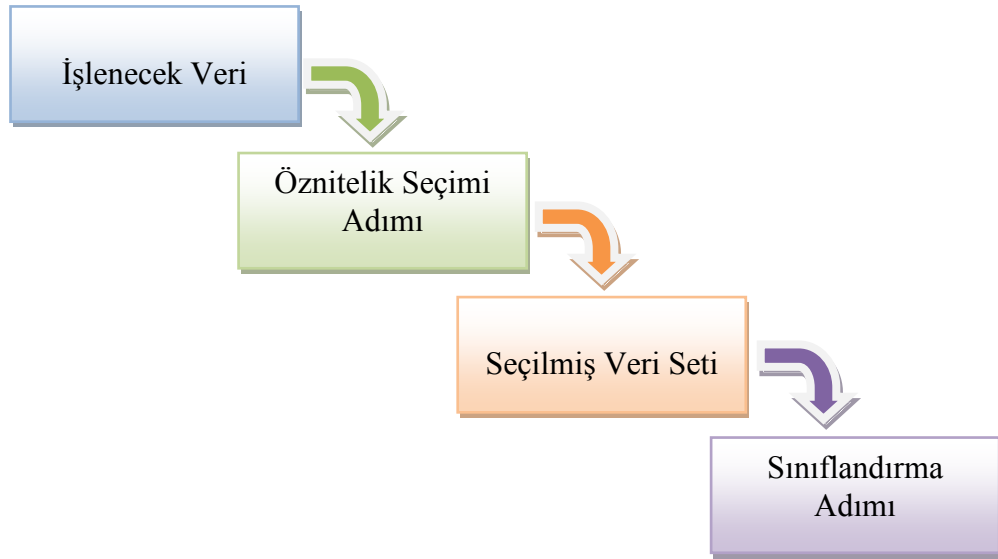
Mahremiyet ile ilgili çalışmalarda sınıflandırma tercih edilen tekniklerden biridir. Bunun nedeni çoğu veri madenciliği probleminin sınıflandırma ile ifade edilmesinden kaynaklanmaktadır. Veri madenciliğinde sınıflandırma işlemi için çeşitli algoritmalar kullanılmaktadır. Bu algoritmalar yardımı ile seçilmiş nitelik setleri sınıflandırılmaya çalışılmaktadır.

İçerisinde mahrem nitelikler barındırmayan veri setleri ile yapılan sınıflandırma adımı ile elde edilen sınıflandırma başarısı analiz edilerek sınıflandırmanın ve mahremiyetin birlikte gerçekleştirilip gerçekleştirilmediği kontrol edilmektedir.

3.2 Yöntem Modeli

Veri madenciliğinde mahremiyeti sağlamak için ortaya koyduğumuz model yöntem bileşenlerinin adım adım izlenmesi ile oluşmaktadır.

Veri, Öznitelik Seçimi Adımı, Seçilmiş Veri Seti ve Sınıflandırma Adımı bileşenlerinin oluşturduğu model diyagramı aşağıda ifade edilmektedir. (Bkz. Şekil 3.1.)



Şekil 3.1. Model Blok Diyagramı

Mahremiyeti koruyucu veri madenciliği önerimiz sınıflandırma modeli yardımıyla gösterilecektir. Bu modelde; tahminci değişkenlerimizi mahrem ve mahrem olmayan nitelikler diye iki kısma ayırdıktan sonra sınıflandırma işlemine girecek olan mahrem verilerin eksiltilmesi için öznitelik seçimi adımı gerçekleştirilir. Bu adım ile mahrem olmayan niteliklerle ne oranda başarılı sonuçlar elde edebileceğimiz görülecektir.

Çalışmada; farklı nitelik değerleri ile farklı sınıflandırma modelleri uygulanmıştır. Öznitelik seçimi gerçekleştirilmeden yapılan sınıflandırma ile bizim yaklaşımımız ile elde edilen sınıflandırma sonuçları karşılaştırılarak yapılan işlemin doğruluğu test edilmiştir.

3.3 Önerilen Yöntemin Değerlendirilmesi

Verileri içerisinde var olan mahrem bilgilerin kişilere ve kurumlara zarar vermemesi için veriler üzerinde bu özel bilgiler dikkate alınarak çalışmalar ortaya konulmalıdır. Bu kapsamda veri madenciliğinde mahremiyetin sağlanabilmesi için çeşitli çalışmalar ve yöntemler ortaya konmaktadır. Bu tez çalışmasında ortaya koyduğumuz yöntem bu çalışmalar arasında yer alabilecek bir yapıdadır.

Üçüncü bölümde mahremiyetin sağlanması için yapılan çalışmalar *Veri Koleksiyonu ile Mahremiyeti Sağlama*, *Sonuç Kontrolü ile Mahremiyeti Sağlama* ve *Bilgi Paylaşımı ile Mahremiyeti Sağlama* olmak üzere 3 kategoride ele alınmıştır.

Veri Koleksiyonu ile Mahremiyeti Sağlama tekniği altında yapılan çalışmalar veriler üzerinde çeşitli işlemler gerçekleştirerek mahremiyeti sağlamaya çalışmaktadır. Veriler üzerinde gerçekleştirilen değiştirme, genelleştirme, gürültü ekleme gibi işlemler mahrem verilerin tamamen ortadan kaldırmamaktadır. Yapılan işlemler çeşitli yöntemler ile geri dönüştürülerek gizli kalması gereken bilgilere ulaşmak mümkündür. Veriler üzerinde gerçekleştirilen şifreleme tekniği dış ortamlara kapalı sistemlerde başarılı sonuçlar vermektedir. Fakat güvenli olmayan ve dışarıya açık olan sistemlerde şifreleme anahtar değerinin elde edilmesi tehlikesi dolayısı ile mahremiyetin açığa çıkmasını doğuracak endişelere neden olmaktadır [Kargupta et al., 2003].

Sonuç Kontrolü ile Mahremiyeti Sağlama tekniği altında yapılan çalışmalar veriler üzerinde veri madenciliği uygulamaları yapıldıktan sonra elde edilen sonuçların mahrem bilgiler içermemesini sağlayacak teknikleri içermektedir. Veri madenciliği yapıldıktan sonra elde edilen sonuç verileri içerisinden çeşitli sorgular yardımıyla mahrem veriler açığa çıkartılabilir. Bu sorgulama işlemlerinin güvenli

olup olmadığını kontrol edebilecek bir mekanizma yardımıyla gizli kalması gereken bilgilerin açığa çıkması önlenabilir. Sorgu denetleme yöntemi olarak bilinen bu yöntem özellikle istatistiksel verilerin mevcut olduğu sistemlerde başarılı sonuçlar vermektedir. Yalnız dinamik olarak sorguların güvenilir olup olmadığının kontrolü çalışma anında zaman alan bir işlemdir. Bu yüzden bazı sorgular uzun süren işlemler sonrası cevap alabilirler. Bazı iyi niyetli sorguların sistem tarafından güvenilir sorgu olmadığı şeklinde işleme alınması bu sorgulamaların cevaplanmaması sorununu doğurmaktadır [Zhang and Zhao, 2007]. Bu sorun veri madenciliği sonuçlarını etkileyecek bir durumdur. Diğer bir yöntem olan veritabanı yaklaşım tekniği verilerin karışık olarak tutulmasını ve sorguların yaklaşık sonuçlar içermesini sağlayarak mahremiyeti sağlamaya çalışmaktadır. Veriler ne kadar karıştırılırsa karıştırılsın içerisindeki mahrem bilgileri içerdiği için mahremiyetin sağlanacağını garanti edememektedir [Srikant et al., 2005]

Bilgi Paylaşımı ile Mahremiyeti Sağlama tekniği ise verilerin farklı veritabanlarında saklanarak sonradan birleştirilmesini esas almaktadır. Bu teknik içerisinde yer alan yarı güvenilir sistem mahrem verilere ulaşmak isteyen kişilerin tasarlanan iletişim protokolünü kullanacağı varsayılmaktadır. Bu protokol üzerinden veriye erişimi ya da engellemeyi gerçekleştirmektedir. Bu yöntem çok pratik kullanım sahasına sahip değildir. Ayrıca şifreleme tekniklerinin de kullanılması bazı performans sorunlarını ortaya çıkarmaktadır [Zhang and Zhao, 2008].

Öznitelik seçimi ile mahrem verilerden ayrıştırılan verilerin kullanılmasını içeren yöntemimiz *Veri Koleksiyonu ile Mahremiyeti Sağlama* tekniğinin altında yer alabilecek bir çalışmadır. Veri Karartma tekniğine yakın olan yöntemimiz veri içerisinde var olabilecek mahrem verileri en az seviyede içerir. Bu yaklaşım veri karartma yöntemindeki mahrem verilerin tekrar elde edilmesi veya mahrem bilgilerin ortaya çıkartılabilmesini önleyecek bir yöntemdir. Ayrıca özellikle mahrem verilerden ayrıştırılarak elde edilen veri setinin daha az nitelik içermesi sınıflandırma işlem yükünün düşürülmesini sağlamaktadır. Dolayısı ile performans sorunu ile karşılaşmamızı engelleyecek bir durum sağlanmış olur.

4. YAPILAN ÇALIŞMA

Veri madenciliğinde yapılan çalışmalar ile mahrem verilerin gün yüzüne çıkabilmesi mümkündür. Bu bizim istemediğimiz bir durumdur. Çünkü mahrem bilgilere ulaşan kötü niyetli kişiler bu bilgilerin sahibi olan kişilere, kurumlara veya kuruluşlara zarar verebilirler. Bu istenmeyen sonuçlarla karşılaşmamak için veri madenciliği uygulamalarında mahrem verilerin saklanması ve gizlenmesi esas alınmalıdır. Yoksa gizli kalması gereken bilgilerin ortaya çıkması ve bunun doğurabileceği endişenin giderilmesi mümkün olmayacaktır.

Veri madenciliğinde mahremiyeti sağlamak adına ortaya koyduğumuz yöntemde gizli kalması gereken bilgilerin ortaya çıkmasını sağlayabilecek mahrem veriler öznitelik seçimi ile ayrıştırılır.

Özellikle mahrem verilerin sınıflandırma başarısına etkisinin az olduğu niteliklerin seçilerek ayrıştırılması hem mahrem verinin ortaya çıkmasını engelleyebilecek hem de sınıflandırma başarısını değiştirmeyecektir. Böylece olabildiğince az mahrem veri barındıran veri seti ile veri madenciliği teknikleri gerçekleştirilerek hedeflenen işler yapılabilir.

Bu yaklaşım çerçevesinde yapılmış olan bu çalışmada veri madenciliğinde mahremiyetin sağlanması amaçlanmaktadır. Yukarı da önerdiğimiz metodun temel alındığı çalışmada sistem bileşenlerinin oluşturulduğu şu adımlar izlenmiştir;

1. Üzerinde çalışılacak veri elde edilir.
2. Veri içerisinde var olan mahrem öznitelikler seçilir.
3. Mahrem olan öznitelikler veri içerisinde çıkarılır.
4. Mahrem olmayan ya da olabildiğince az mahrem veri barındıran veriler veri madenciliği sınıflandırma tekniğinde kullanılır.

Yapılan çalışmada öznitelik seçimi yapılmayan sınıflandırma sonuçları ile öznitelik seçimi yapılarak mahrem verilerin ayrıştırıldığı veri setleri ile yapılan

sınıflandırma sonuçları karşılaştırılır. Bu işlem yapılan sınıflandırma işleminin doğruluğunu kontrol etmek için gereklidir.

Bu çalışmada Kaliforniya Üniversitesi, Irvine kampusu Makine Öğrenmesi Ambarı (UCI Machine Learning Repository) topluluğu tarafından yayınlanan yetişkinlere ait yetişkin veri kümesi kullanılmaktadır [cml.ics.uci.edu, 1994]. Yetişkin veri kümesi içerisinde kategorik ve sayısal olmak üzere meslek, eğitim, gelir, ülke, cinsiyet, ırk, yaş gibi alanları içeren elli binin üzerinde kayıt mevcuttur. Bu kayıtlar sınıflandırma problemlerinde kullanılmak üzere üretilmiştir.

Yetişkin veri kümesi içerisinde mahrem sayılabilecek ülke, cinsiyet, ırk, yaş gibi verilerin ayrıştırılması öznelik seçimini sağlar. Buradaki mahrem verilerden sınıflandırma işlemine etkisinin az olduğu nitelikler belirlenerek kapsam dışında bırakılması için bu niteliklerin oluşturdukları kombinasyonlar veri setleri olarak hazırlanır.

Veri madenciliği araçlarından biri olan Tanagra [eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html, 2005] yazılımı kullanılarak bu çalışma gerçekleştirilmektedir. Tanagra açık kaynak olan veri madenciliği uygulamalarından biridir. Fransa Lyon Üniversitesi tarafından geliştirilmiştir. Veri madenciliğinin daha kolay ve etkili kullanılmasını amaçlamaktadır. Bu anlamda kolay kullanım imkânı sağlar. Tanagra yazılımına giriş parametresi olarak yetişkin veri kümesi verilerek sınıflandırma işlemi yapılır.

Sınıflandırma işlemi için C-RT, ID 3, Navie Bayes ve Multi Layer Perceptron (MLP) algoritmaları kullanılmıştır. C-RT algoritması sınıflandırma ve regresyon ağaçları yapısını, ID3 algoritması karar ağaçları yapısını, Navie Bayes olasılık tabanlı sınıflandırma algoritmasını, Multi Layer Perceptron (MLP) algoritması ise yapay sinir ağlarını temel alan algoritmalarıdır.

Giriş değeri olarak sisteme girilen veri setlerinin sınıflandırma algoritmalarına göre sınıflandırma başarıları analiz edilerek sınıflandırmanın ve mahremiyetin birlikte gerçekleştirilebileceği gözlenmeye çalışılmaktadır.

Yapılan çalışmayı daha net ortaya koymak için sistem bileşenlerini ve içeriklerini ayrı ayrı ele almak gerekir.

4.1 Veri

Veri olarak 1994 yılına ait Census veritabanında bulunan yetişkinlere ait bilgilerin mevcut olduğu bir veri tablosu kullanılmaktadır. Bu veri tablosu Kaliforniya Üniversitesi, Irvine kampüsü Makine Öğrenmesi Ambarı (UCI Machine Learning Repository) topluluğu tarafından yayınlanmaktadır. Yetişkin verisi sınıflandırma problemlerinde kullanılmak üzere üretilmiştir. Sınıflandırma bilgisi niteliği yetişkin veri kümesinin verilerinin hangi sınıfa ait olduğunu belirtmektedir.

Yetişkin veri kümesi içerisinde kategorik ve sayısal olmak üzere iki tür veri mevcuttur. (Bkz. Çizelge 4.1.)

Çizelge 4.1. Yetişkin Verisi.

Alan Adı	Veri Tipi	Veri
Yaş	Sürekli	10, 14, 27, 34, 50
Çalışma Alanı	Kategorik	Özel Sektör, Serbest Meslek, Kamu
Son Ağırlık Değeri	Sürekli	77516, 83311, 338409
Eğitim Seviyesi	Kategorik	Lise, Lisans, Yüksek Lisans, Doktora
Okul Sınıfı	Sürekli	13, 9, 8, 11
Medeni Durumu	Kategorik	Bekâr, Evli, Dul
Meslek Bilgisi	Kategorik	Satıcı, Öğretmen, Mühendis, Doktor
Akrabalık Derecesi	Kategorik	Eş, Öz çocuk, Aileden değil, Anne
İrk	Kategorik	Uzak Doğulu, Siyah, Eskimo
Cinsiyet	Kategorik	Erkek, Bayan
Gelir Bilgisi	Sürekli	1200, 2450, 300, 678

Gider Bilgisi	Sürekli	0, 100, 230, 900
Haftalık Çalışma Saati	Sürekli	40, 13, 36
Ülke Bilgisi	Kategorik	Türkiye, İngiltere, Almanya
Sınıf Bilgisi	Kategorik	Sınıf 1, Sınıf 2

4.2 Öznitelik Seçimi

Sınıflandırma işleminde kullanılacak verilerin ön bir işleme tabi tutulması ile daha iyi bir sonuç elde edilebilir. Bu ön işlem verilerin içerisindeki faydasız bilgilerin ayıklanmasını içerdiği kadar mahremiyetin sağlanması için önemli olabilecek yahut veri madenciliği işlemine tabi tutulmaması gereken niteliklerinde çıkartılmasını sağlar. Bu sayede mahrem niteliklerin ortaya çıkmaması sağlanabilir.

Ortaya konan model için seçilen yetişkin veri kümesi içerisinde bulunan mahrem sayılacak veriler öznitelik seçimi ile seçilerek sınıflandırma işleminde kullanılması engellenmektedir. Yetişkin veri kümesi içerisinde bulunan nitelikler arasından mahrem sayılan nitelikler seçilerek veri madenciliğine verilecek olan giriş değerlerinden çıkartılmaktadır.

Bütün özniteliklerin giriş verisi olarak verilmesi işlem yükünün artmasına sebep olabilir. Hem bu yükün dengelenmesi için hem de mahrem verilerin açığa çıkmasını engellemek için bir ön işlemin gerçekleştirilmesi gereklidir. Bu ön işlemi gerçekleştirmek için yetişkin veri kümesinde bulunan mahrem sayılabilecek nitelikler seçilerek giriş veri setinden çıkartılmaktadır.

Mahrem verilerin eksiltildiği veri setleri ile sınıflandırma işlemi gerçekleştirilerek mahrem verilerin sınıflandırmaya olan etkisi incelenmektedir. Özellikle sınıflandırma başarısını çok fazla etkilemeyen mahrem niteliklerin belirtilerek çıkartılması mahrem verilerin veri madenciliğine girmesi ile olabilecek mahrem verilerin açığa çıkmasını engeller.

4.3 Mahrem Olmayan Veri Seti

Veri madenciliği işlemine girmemesi gereken ve sınıflandırmayı en az şekilde etkileyebilecek mahrem verilerin ayrıştırılması ile elde edilen verileri içermektedir. Böylece mahrem niteliklerden arındırılmış bir veri seti elde edilmiş olur. Bu veri seti kullanılarak sınıflandırma işlemi gerçekleştirilir.

Öznitelik seçimi ile elde edilen veri seti ile mahremiyetin sağlanması hedeflenmektedir. Veri seti mümkün olduğu kadar mahrem verilerden ayrıştırılmış bir şekilde elde edilmeye çalışılır.

Sınıflandırmaya etkisinin en az olduğu ve sınıflandırmaya girmemesi gereken mahrem niteliklerin belirlenmesi için yetişkin veri kümesinden öznitelik seçimi ile elde edilebilecek veri setleri oluşturularak veri madenciliğine giriş değeri olarak teker teker verilir ve sonuçlar değerlendirilir.

Yetişkin veri kümesi içerisinde çıkarılan mahrem niteliklerin kombinasyon ile oluşturacakları veri setleri elde edilir. Buradaki veriler kullanılarak sınıflandırma işlemi gerçekleştirilir. Seçilmiş veri setleri sayesinde mahremiyetin sağlanması hedeflenmektedir. Çünkü veri setlerimiz mümkün olduğu kadar mahrem verilerden ayrıştırılmış bir şekilde ifade edilmektedir. Bu veri setlerinin veri madenciliği uygulamasında kullanılması ile veri madenciliğinde mahremiyet sağlanmış olacaktır.

Yetişkin veri kümesi içerisinde mahrem veri olarak kabul ettiğimiz ülke, ırk, cinsiyet ve yaş olmak üzere dört alan seçilmiştir. Bu mahrem verilerin oluşturacağı veri seti dizileri ile sınıflandırma işlemi gerçekleştirilerek sınıflandırmaya olan etkisi hesaplanmaktadır.

Bu dört mahrem niteliklerin oluşturabileceği veri seti dizilerinin belirlenmesi için kombinasyon kullanılır. Kombinasyon bir nesne grubu içerisinde, sıra gözetmeksizin yapılan seçimler olarak tanımlanır.

Kombinasyon denklemi:

$$C(n, r) = C \binom{n}{r} = \frac{n!}{r! \cdot (n - r)!} \quad (4.1.)$$

Dört adet mahrem niteliğin oluşturacağı veri seti dizilerinin sayısının hesaplanması için aşağıdaki kombinasyonlar hesaplanıp toplanır.

$$\text{Veri Seti Sayısı} = C \binom{4}{0} + C \binom{4}{1} + C \binom{4}{2} + C \binom{4}{3} \quad (4.2.)$$

Bu kombinasyonların hesaplanması sonucu 16 tane farklı veri setleri oluşturulmaktadır. Bu veri setlerinin içerdiği mahrem veriye göre gösterimi aşağıdaki çizelgede ifade edilmektedir. (Bkz. Çizelge 4.2.)

Çizelge 4.2. Veri seti örnekleri.

Veri seti	Ülke	İrk	Cinsiyet	Yaş
X ₁	Yok	Yok	Yok	Yok
X ₂	Yok	Yok	Yok	Var
X ₃	Yok	Yok	Var	Yok
X ₄	Yok	Yok	Var	Var
X ₅	Yok	Var	Yok	Yok
X ₆	Yok	Var	Yok	Var
X ₇	Yok	Var	Var	Yok
X ₈	Yok	Var	Var	Var
X ₉	Var	Yok	Yok	Yok
X ₁₀	Var	Yok	Yok	Var
X ₁₁	Var	Yok	Var	Yok
X ₁₂	Var	Yok	Var	Var
X ₁₃	Var	Var	Yok	Yok
X ₁₄	Var	Var	Yok	Var
X ₁₅	Var	Var	Var	Yok
X ₁₆	Var	Var	Var	Var

Çizelge 4.2.' de görüldüğü gibi ülke, ırk, cinsiyet, yaş alanlarının birleşimleri ile oluşturulan mahrem veri setleri diğer alanlarla birleştirilerek seçilmiş veri setleri elde edilmektedir. Bu veri setleri sınıflandırma algoritmalarında kullanılır.

4.4 Sınıflandırma Algoritmaları

Verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflandırma olarak adlandırılır; veri madenciliğinde önemli bir konudur. Sınıflandırmaya dayalı modeller daha önceden bilinen değerlere göre tahminde bulunmaya yönelik bir modeldir. Sınıflandırma işleminde elimizde ki sınıf veya istatistiksel değimiyle bağımlı değişken hem sınıfsal hem de sürekli değer taşıyabilir; bu anlamda regresyon ve çok terimli regresyona yaklaşılmaktadır. Ancak, veri madenciliği çerçevesinde bu istatistiksel yöntemlerin dışında sınıflandırma işleminde *Bayesyen Sınıflandırma Algoritması*, *Karar Ağaçlarına Dayalı Algoritmalar*, *Yapay Sinir Ağları Temelli Algoritmalar* ve *K-en Yakın Komşu Algoritması* gibi birçok teknik ve algoritma geliştirilmiştir.

4.4.1 ID3 Algoritması

ID3 algoritması karar ağaçları yardımıyla sınıflandırma işlemlerini yerine getirmek için geliştirilmiştir. Karar ağaçları, en çok kullanılan sınıflandırmaya dayalı modeldir. Karar ağaçlarının sıkça kullanılmasında ki en büyük etken kullanımı ve sonuçlarının anlaşılmasının en kolay yöntem olmasından kaynaklanmaktadır. Temel olarak karar ağaçlarından sonuç elde etmenin 2 aşaması vardır. Birincisi ağacın kurulmasıdır. Karar ağacının kurulmasında, daha hızlı ve güvenli bir sonuç alınması için verileri en büyük parçaya bölecek sorular kök düğümden itibaren sorulmaya başlanmalıdır. Karar ağacı yapısı oluşturulduktan sonra, ikinci aşamaya geçilip eldeki veri ağacın uygun dalına yerleştirilir.

Ağaç oluşturmada yapılan işlemlerden bir tanesi de budama işlemidir. Budama ağaçta oluşmuş sonucu etkilemeyen ve sınıflanmaya herhangi bir katkısı olmayan dalların ağaçtan alınmasıdır. Bir bakıma gereksiz ayrıntıların sonuçtan

çıkartılması işlemidir. Ağaçtan birçok düğüm ve dal oluşursa, ağacın alt dallara ve yapraklarına ulaşan veri sayısı da azalacaktır; bu da ağacın hassasiyetini azaltacaktır.

ID3 algoritması entropi tabanlı bir algoritmadır. Karar ağacında hangi niteliğe göre dallanmanın yapılacağını belirlemek üzere bu entropi kavramına başvurulur. Entropi, bir sistemdeki belirsizliğin ölçüsüne denir. Entropi eldeki verilerin birbirinden farklılığına dayalı bir formül olarak düşünülürse aynı değere sahip 2 ifadenin entropileri 0 çıkmaktadır.

Entropi formüsel olarak ifade etmek için S bir kaynak olsun. Bu kaynağın $\{m_1, m_2, \dots, m_n\}$ olmak üzere n adet mesaj üretebildiğini varsayalım. Tüm mesajlar birbirinden bağımsız olarak üretilmektedir ve m_i mesajlarının üretilme olasılıkları $P=\{P_1, P_2, \dots, P_n\}$ olasılık dağılımına sahip mesajları üreten S kaynağının entropisi $H(S)$ şu şekildedir [Shannon, 1948]:

$$H(S) = - \sum_{i=1}^n P_i \log_2(P_i) \quad (4.3.)$$

ID3 algoritması veritabanındaki 2 değişkenin dallanmasına göre farklı değerleri almaktadır. Bu değerlerin farkına kazanım demek ve hangi değişkenin kazanımı fazla ise bu değişkene göre de ağacı dallandırmaktadır. Kazanım formülü aşağıdaki gibi hesaplanır.

Sınıf niteliğini ifade eden T, sınıf niteliği olmayan bir X niteliğinin değerine bağlı olarak $T_1, T_2, \dots, T_n$ alt kümelerine ayrılırsa, T'nin bir elemanının sınıfını belirlemek için gerekli bilgi, T_i nin bir elemanının sınıfının belirlenmesinde gerekli olan bilginin ağırlıklı ortalaması olarak kabul edilir. O halde bu tanıma bağlı olarak T nin bir elemanının sınıfını belirlemek için gerekli bilgi şu şekilde hesaplanır [Quinlan,1993]:

$$H(X, T) = \sum_{i=1}^n \frac{|T_i|}{|T|} H(T_i) \quad (4.4.)$$

T veritabanını X testine göre bölmekle elde edilen bilgileri ölçmek için kazanç ölçütü adı verilen bir ifadeye başvurulur. Bu ölçüt şu şekilde hesaplanır:

$$\text{Kazanç}(X, T) = H(T) - H(X, T) \quad (4.5.)$$

Burada ayırma işlemi yapılırken $\text{Kazanç}(X, T)$ değerinin en yüksek olması hedeflenir. En yüksek bilgi kazancını sağlayan X testi seçilir ve sınıflandırma işlemi yapılır.

4.4.2 C-RT Algoritması

C-RT algoritması sınıflandırma ve regresyon ağaçlarını kullanarak sınıflandırma işlemi gerçekleştirir. Sınıflandırma ve regresyon ağaçları veri madenciliğinin sınıflandırma ile ilgili konuları arasında yer alır. C-RT karar ağacı, her bir karar düğümünden itibaren ağacın iki dala ayrılması ilkesine dayanır [Bremian et al., 1998]. Yani bu tür karar ağaçlarında ikili dallanmalar söz konusudur. O halde bir düğümde seçme işlemi yapıldığında, düğümlerden sadece iki dal ayrılabilir.

C-RT algoritmasında, bir düğümde belirli bir kriter uygulanarak bölünme işlemi gerçekleştirilir. Bunun için önce tüm niteliklerin var olduğu değerler göz önüne alınır ve tüm eşleşmelerden sonra iki bölümde elde edilir. Bu bölünmeler üzerinde seçme işlemi uygulanır.

C-RT tekniği ID3 algoritmasında olduğu gibi en iyi dallara ayırma kriterini seçmek için entropiden yararlanır. C-RT dallara ayırma kriterini hesaplarırken kayıp verileri önemsemez.

4.4.3 Naive Bayes Algoritması

Naive bayes Algoritması istatistiksel tekniklerin kullanılarak sınıflandırılmasını sağlar. Temel olarak *Bayes teoremine* dayanmaktadır. Bayes teoremi olasılık hesabında önemli bir yere sahiptir ve bu teoreme dayanarak sınıflandırma işlemi yapmak mümkündür.

Sınıflandırma işlemi başında X kümesi sınıf üyeliği bilinmeyen veri kümesi olarak kabul edilir. H ise bu X veri örneğinin C sınıfına ait olduğunu iddia eden hipotez olarak ortaya konur. H 'nin C sınıfına ait olduğu varsayımı ile $P(H \setminus X)$ olasılığı hesaplanır. Burada $P(H \setminus X)$, H hipotezinin X üzerinde koşullandırılmasıyla ilgili “sonrasal olasılık” olarak bilinir.

Örneğin bir torbada bazı cisimlerin bulunduğunu varsayalım. Elimizdeki bilgiler X 'i tanımlar. Cisimlerin yuvarlak ve kırmızı olduğunu bildiğimizi varsayalım. Hipotezimiz ise “bu bir elmadır” olsun. Bu durum $P(H)$ bir “önsel olasılık” olarak karşımıza çıkacaktır. Yani başlangıçta bu olasılığın ne olduğunu biliyoruz. Ancak $P(X \setminus H)$ olasılığı ise H koşulu üzerinden kurulduğundan bir “sonrasal olasılık” olarak değerlendirilir. Yani, bir X 'in kırmızı ve yuvarlak olması olasılığı biliniyorsa “bu bir elmadır” sonucunu doğurur. Bu durumda Bayes bağlantısı şu şekli alır:

$$P(H \setminus X) = \frac{P(X \setminus H)P(H)}{P(X)} \quad (4.6.)$$

Naive bayes algoritması elde ki mevcut sınıflandırmış verilerden yola çıkarak yeni gelen bir verinin hangi sınıfa uyduğunu bulmak için bu formülü esas alır.

4.4.4 MLP Algoritması

Multi Layer Perceptron (MLP) algoritması yapay sinir ağlarını temel alarak sınıflandırma işlemini gerçekleştirir. Yapay sinir ağları, insan beyninin problem çözme şeklini ve matematiksel işleyişini makinelere kazandırılması için geliştirilmişlerdir.

Yapay sinir ağlarında öğrenme, ilişkilendirme, genelleme ve sınıflandırma gibi işlemler, beynin paralel ve yüksek yoğunluklu ara bağlantılı yapısını modellenmesi sonucunda gerçekleştirilmektedir. Öğrenme yolu ile kazanılan bilgi, snaptik ağırlıklar olarak bilinen sinirler arası bağlantılarda saklanır. Problemlerin en uygun çözümünü bulmak maksadıyla bahse konu ağırlıkların ve serbest parametrelerin belirlenmesi işlemi sonucunda öğrenme gerçekleşmektedir.

Sınıflandırma işlemini gerçekleştirmek için veri madenciliğinde var olan Multi Layer Perceptron (MLP) algoritması ile birlikte yapay sinir ağı modeli kullanılır. Sınıflandırılan veriler giriş değeri olarak çok katmanlı bir yapay sinir ağına verilir ve sistemin sınıflandırma işlemi için gerekli öğrenimleri kazanması beklenir. Bu kazanım nöronlar arasındaki değerlerin hesaplanması sonucu elde edilir. Sisteme verilen sınıf bilgisi bilinmeyen giriş verileri önceden oluşturulan öğrenme bilgisi sayesinde sınıflandırılır.

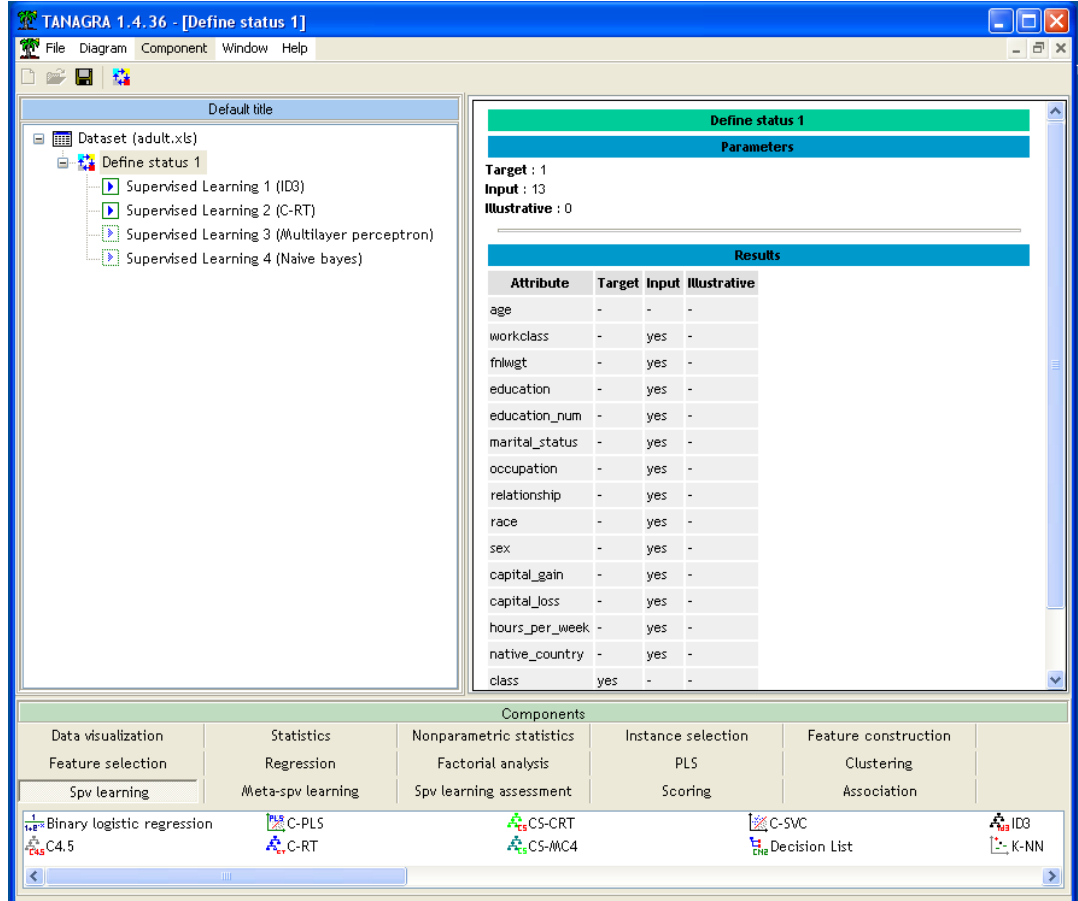
4.5 Veri Madenciliği Yazılımı (Tanagra)

Teknolojinin gelişimiyle bilgisayar ortamında ve veritabanlarında tutulan veri miktarının artması, yeni veri toplama yolları, otomatik veri toplama aletleri, veritabanı sistemleri, bilgisayar kullanımının artması, büyük veri kaynakları (İş dünyası: Web, e-ticaret, alışveriş, hisse senetleri gibi.), bilim dünyası (Uzaktan algılama ve izleme, simülasyonlar gibi.), toplum (haberler, dijital kameralar, Youtube, Facebook gibi.) neden veri madenciliği sorusuna cevap vermektedir.

Veri Madenciliği uygulamalarını gerçekleştirmek için yazılımlara ihtiyaç duyulur. Bu kapsamda, SPSS Clementine, Excel, SPSS, SAS, Angoss, KXEN, SQL Server, MATLAB gibi ticari ve RapidMiner(YALE), WEKA, R, C4.5, Orange, KNIME, Tanagra gibi açık kaynak olmak üzere birçok yazılım geliştirilmiştir.

Bu çalışmada Tanagra yazılımı kullanılmaktadır. Veri madenciliği yazılımlarından biri olan Tanagra, Fransa Lyon Üniversitesi tarafından açık kaynaklı bir yazılım olarak geliştirilmektedir. Akademik ve araştırma projelerinde kullanılması hedeflenen Tanagra, açık kaynaklı olduğu için kullanıcılar tarafından da geliştirilmektedir. Böylece ihtiyaç duyulan algoritmalar ve gereksimler Tanagra'ya eklenebilmektedir. Kullanım kolaylığı ve ücretsiz olması Tanagra'nın kullanıcılara sunduğu diğer avantajlardır.

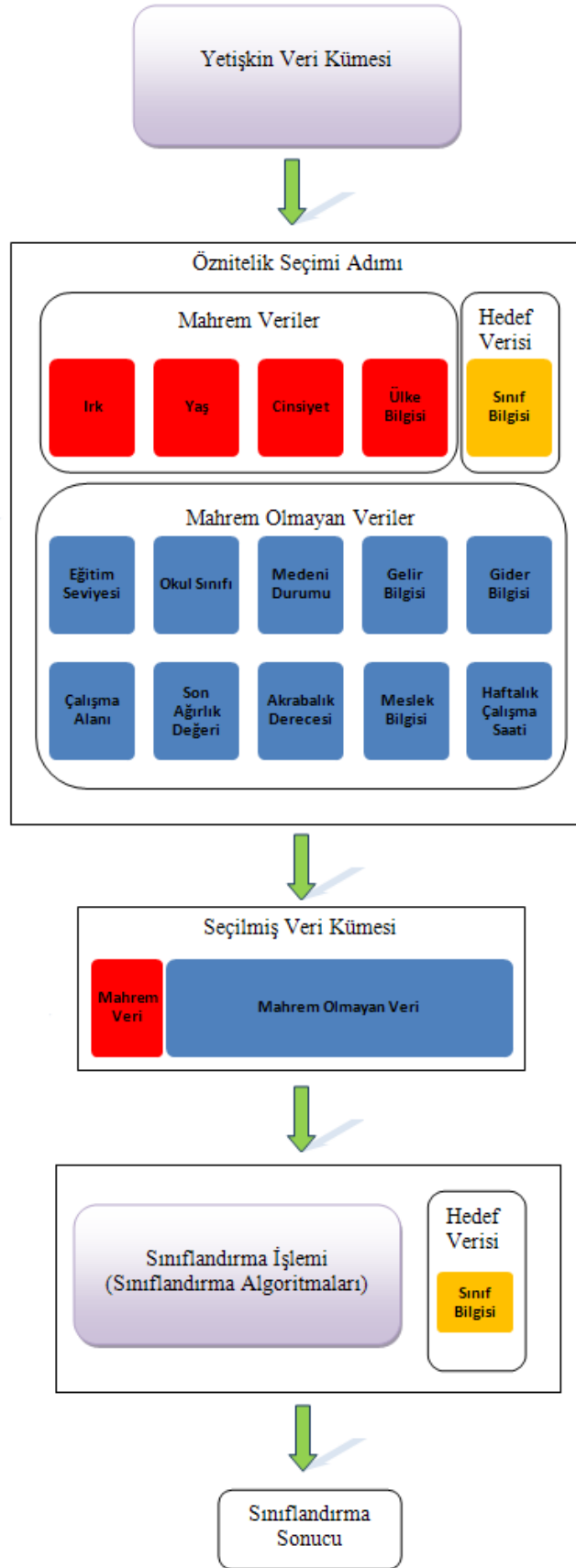
Yetişkin veri kümesinden alınan veri setinin girdi değeri olarak verilmesi ve bu veri setine veri madenciliği sınıflandırma algoritmalarının seçilmesini gösteren ara yüz şekil 4.1.'de gösterilmektedir.



Şekil 4.1. Tanagra Grafiksel Ara Yüzü.

4.6 Yapılan Çalışmanın Modeli

Yetişkin veri kümesi ele alınarak içerisinde bulunan öznitelikler mahrem ve mahrem olmayan nitelikler olarak ayrıştırılmıştır. Yetişkin veri kümesi içerisinde var olan sınıf bilgisi verisi sınıflandırma işlemi için kullanılacak bir ölçüttür. Mahrem verilerin kombinasyonlarının mahrem olmayan veriler ile oluşturduğu veri setleri sınıflandırma işlemine verilmiştir. Hedef veri olarak sisteme verilen sınıflandırma bilgisi yardımıyla sınıflandırma sonuçları elde edilmiştir. Elde edilen sınıflandırma başarısı ile en az mahrem veriyi barındırarak en az sınıflandırma başarısı kaybını sağlayan veri setleri seçilerek veri seti içerisinde kullanılmayacak mahrem nitelikleri belirlemeye çalışılmıştır. Veri madenciliğinde mahremiyetin sağlanması için yapılan çalışmada ortaya konan bileşenlerin bir araya gelmesi sonucu oluşan model şekil 4,2’de ifade edilmektedir.



Şekil 4.2. Yapılan Çalışma Modeli.

4.7 Deney

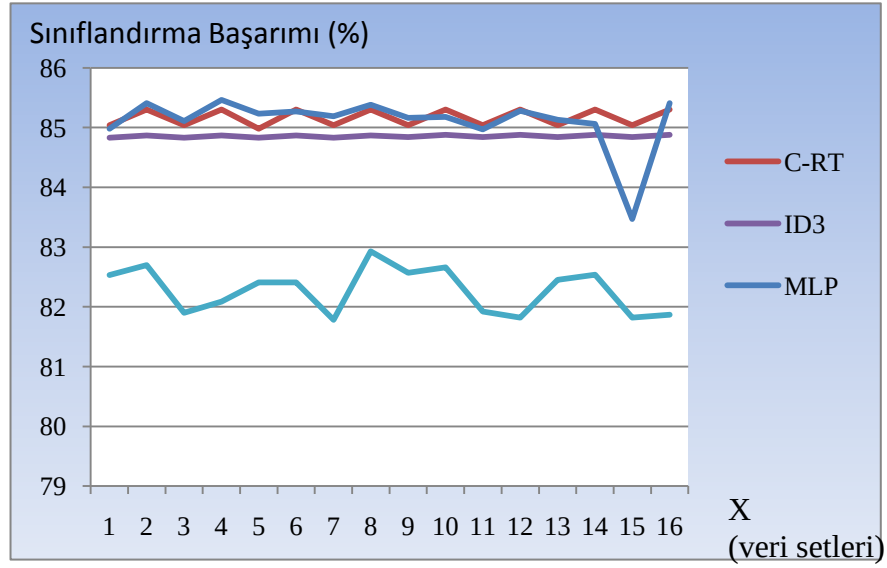
Veri madenciliğinde mahremiyetin sağlanması için ortaya konan modelin gerçekleştirilmesi gerekmektedir. Gerçekleştirilecek bu model ile mahrem verilerin açığa çıkmasının önlenmesinin gözlenmesi amaçlanmaktadır.

Yetişkin veri kümesi içerisinde var olan mahrem verilerin seçilerek çıkartılması ile elde edilecek veri seti dizileri deney için giriş verileri olarak belirlenmektedir. Ülke, ırk, cinsiyet ve yaş olmak üzere belirlenen dört mahrem veri setlerinde olabilecek tüm kombinasyonlarla eklenip çıkartılmaktadır. Özellikle veri içerisindeki mahrem niteliklerin eksiltiyle elde edilen veri seti dizileri mahrem bilgilerin açığa çıkmasını engelleyecektir. Bu diziler sınıflandırma işlemi için giriş değeri olarak sisteme verilmektedir.

Sınıflandırma işlemi için sınıflandırma ve regresyon ağaçları yapısını kullanan C-RT algoritması, karar ağaçları yapısındaki ID3 algoritması, olasılık tabanlı Navie Bayes algoritması, yapay sinir ağlarını temel alan Multi Layer Perceptron (MLP) algoritması kullanılmaktadır.

Bu algoritmalar ile birlikte sınıflandırma işlemi gerçekleştirilmektedir. Veri madenciliği yazılımı olan Tanagra yazılımı kullanılarak veri setleri sınıflandırma algoritmaları ile birlikte çalıştırılmaktadır. Bu işlem sonucunda elde edilen sonuçlar analiz edilerek sınıflandırmanın ve mahremiyetin birlikte gerçekleştirilip gerçekleştirilmediği kontrol edilmektedir.

Yetişkin veri kümesi içerisinde bulunan mahrem niteliklerin oluşturduğu kombinasyonlar veri madenciliği yazılımına giriş değerleri olarak verilmektedir. Bu veri setlerinin her bir algoritma için sınıflandırma sonuçları elde edilerek sınıflandırma başarısının sonuçları şekil 4.2.'de gösterilmektedir.



Şekil 4.3. Sınıflandırma Sonuçları.

Gerçekleştirilen deneyler sonucunda C-RT, ID3 ve MLP sınıflandırma algoritmaları ile elde edilen sınıflandırma başarısı %85 civarı iken Navie Bayes sınıflandırma algoritması ile elde edilen sınıflandırma başarısı %82 civarındadır. Bu sınıflandırma başarısı sınıflandırma algoritmalarına giriş değeri olarak verilen veri seti örneklerine göre %1 oranında değişim göstermektedir. Yani mahrem veriler içeren veri setleri ve mahrem veriler içermeyen veri setleri arasındaki sınıflandırma başarısı hemen hemen aynı değere sahip olmaktadır.

4.8 Sınıflandırma Kazancı

Veri içerisinde bulunan mahrem ve mahrem olmayan niteliklerin birbirinden ayrıştırılarak sınıflandırmanın maksimum düzeyde başarılı olmasını beklemek en ideal çözüme ulaşmamız demektir.

Veri madenciliğinde hem mahrem verilerin saklı kalmasını sağlamak hem de yapılacak işin doğru ve etkin bir şekilde gerçekleştiğini görmek için yaptığımız çalışmanın sonuçlarını analiz etmemiz gerekmektedir. Burada dikkat etmemiz gereken mahrem bilginin ortaya çıkmasını engellemek olduğu kadar sınıflandırma başarısının da minimum kayıp ile elde edilmesini sağlamaktır.

Sınıflandırma başarısını hesaplamak için sınıflandırma kazancını hesaplamak gereklidir. Sınıflandırma kazancı sınıflandırma algoritmasına giren nitelik sayısı ile elde edilen sınıflandırma sonucu arasında bir bağıntı ile ifade edilebilir. Bu bağıntıyı kurduktan sonra sınıflandırma kazancını daha net bir şekilde ortaya koyup analiz edebiliriz.

Sınıflandırma işlemine giren veri seti içerisindeki nitelik sayısı n olsun. Sınıflandırma sonucu ise k olsun. Sınıflandırma kazancı sınıflandırma sonucunu sınıflandırmaya giren veri setindeki nitelik sayısına bölerek elde edebiliriz. Bu işlemin formülü aşağıda gösterilmektedir:

$$\text{Sınıflandırma Kazancı} = \frac{k}{n} \quad (4.7.)$$

Bu formüle göre yapılan çalışma ile elde edilen sonuçların sınıflandırma kazancı hesabı aşağıda gösterilmektedir. (Bkz. Çizelge 4.3.)

Çizelge 4.3. Sınıflandırma Kazancı.

Veri Seti	C-RT	ID3	MLP	Navie Bayes
X ₁	8.50	8.48	8.50	8.25
X ₂	7.75	7.72	7.76	7.52
X ₃	7.73	7.71	7.74	7.45
X ₄	7.11	7.07	7.12	6.84
X ₅	7.73	7.71	7.75	7.49
X ₆	7.11	7.07	7.11	6.87
X ₇	7.09	7.07	7.10	6.82
X ₈	6.56	6.53	6.57	6.38
X ₉	7.73	7.71	7.74	7.51
X ₁₀	7.11	7.07	7.10	6.89

X ₁₁	7.09	7.07	7.08	6.83
X ₁₂	6.56	6.53	6.56	6.29
X ₁₃	7.09	7.07	7.09	6.87
X ₁₄	6.56	6.53	6.54	6.35
X ₁₅	6.54	6.53	6.42	6.29

Öznitelik seçimi ile oluşturulan veri setleri ile yapılan sınıflandırma işleminde sınıflandırma kazancının yüksek olması sınıflandırma işleminin başarılı olduğunu göstermektedir.

Sınıflandırma kazancı hesaplamaları dikkatli bir şekilde incelendiğinde veri setinin az olduğu ve sınıflandırma değerinin yüksek olduğu değerlerde kazancında yüksek olduğu görülmektedir. (Bkz. Çizelge 4.3.)

5. SONUÇ

İletişim ve bilişim teknolojisindeki hızlı gelişmeler ve ilerlemeler sonucu sosyal yaşantımıza birçok yenilik getirmektedir. Özellikle internetin yaygın bir şekilde kullanılması, teknolojik ve sosyal etkileşimin artması mahrem veriler dâhil olmak üzere birçok verinin toplanıp saklanmasına imkân yaratmaktadır. Bu veriler üzerinde çeşitli çalışmalar yapılarak faydalı sonuçlar elde edilebileceği gibi kişileri ya da kurumların aleyhlerine olabilecek bilgilerinde elde edilmesi mümkündür. Bu bilgiler kişilik haklarına aykırı ve insanın mahremiyetini gözetmiyorsa büyük bir tehlike doğurmaktadır. Bu tehlikenin ilerleyen zaman içerisinde tehdit haline dönüşebilmesi ve kişilere ya da kurumlara zarar vermesi mümkündür.

Veri madenciliği teknikleri kullanılarak veriler içerisinde gizli kalmış mahrem bilgilerin açığa çıkması mümkündür. Açığa çıkan kişisel ve özel bilgilerimizin kötü niyetli insanlar tarafından kullanılmaması için mahrem verilerin daha dikkatli bir şekilde ele alınması gereklidir. Aksi takdirde gizliliğin ortadan kalkması endişesi ile insanlar yanlış veya eksik bilgi vermek gibi davranışlar içerisinde bulunabilirler. Bu durum veri analizi ve bilgi çıkarım işlemleri için uygun olmayan bir ortamın doğmasına sebep olabilir. Çünkü mahremiyet korkusu ile eksik verilen bilgiler doğruluğu düşük veya yanlış bilgilerin ortaya konmasına sebep olacaktır.

Veri madenciliği kullanımı ile ortaya çıkabilecek mahrem bilgilerin oluşturabileceği tehditlerin ve zararların ortadan kalkması için veri madenciliğinde mahremiyetin göz önünde bulundurulması gereklidir. Bu tez çalışmasında mahremiyete biraz daha fazla önem ve öncelik vererek saklı kalması gereken bilgilerin gün ışığına çıkmasına engel olunması amaçlanmıştır.

Ortaya koyduğumuz sistem çerçevesinde eldeki veri kümesi içerisindeki mahrem nitelikler öznitelik seçimi mantığıyla ayrıştırılarak mahremiyeti sağlayacak veri seti haline getirilmeye çalışılmıştır. Mahrem verilerden olabildiğince uzak olan mahrem veri setleri veri madenciliği sınıflandırma işlemine girdi parametresi olarak verilerek sınıflandırılmaya çalışılmıştır.

Veri madenciliği sınıflandırma algoritmaları (C-RT algoritması, ID3 algoritması, Naive Bayes algoritması, Multi Layer Perceptron Algoritması) ile gerçekleştirilen sınıflandırma işlemi sonuçları analiz edildiğinde mahremiyetin sınıflandırma işleminin birlikte gerçekleştirilebildiği gözlenmektedir.

Mahrem verilerin ayrıştırılarak elde edildiği veri setlerinin sınıflandırma kazançları mahrem verileri de içeren veri setlerinin sınıflandırma kazançlarından yüksek olduğu da görülmektedir.

Yapılan çalışma çerçevesinde hem sınıflandırma başarısının hem de sınıflandırma kazancının yüksek olması mahremiyetin sağlanabileceğini göstermektedir [Kavza ve Takçı, 2009].

Sonuç olarak mahrem verilerden arındırılarak oluşturulan veri setleri ile gerçekleştirilecek veri madenciliği uygulamaları mahremiyeti sağlayabilmektedir. Veri koleksiyonları ile mahremiyeti sağlama yöntemlerinde var olan veri karartma tekniğine yakın olan bu yaklaşım ile mahrem verilerin kontrol altına alınarak sistemden uzaklaştırılması sonucu çalışmamızın hedeflediği amaca ulaşmış oluyoruz.

KAYNAKLAR

- [1] Ann Cavoukian,” Data Mining: Staking a Claim on Your Privacy”, Information and Privacy Commissioner's Report, Ontario, Canada, 1998.
- [2] L. F. Cranor, J. Reagle, and M. S. Ackerman. “Beyond concern: Understanding net users’ attitudes about online privacy”, Technical report, AT&T Labs-Research, 1999.
- [3] D. Agrawal, C. C. Aggarwal, “On the design and quantification of privacy preserving data mining algorithms”, ACM SIGMOD, pp. 247-255, 2001.
- [4] Zhiqiang Yang, Sheng Zhong1, Rebecca N. Wright, “Privacy-Preserving Classification of Customer Data without Loss of Accuracy”, International Conference on Data Mining (SDM), 2005.
- [5] L. Wang, S. Jajodia, D. Wijesekera, "Securing OLAP Data Cubes Against Privacy Breaches", Proc. 25th IEEE Symposium on Security and Privacy, pp. 161-175, 2004.
- [6] R. Srikant, R. Agrawal, D. Thomas, "Privacy Preserving OLAP", Proc. 25th ACM SIGMOD International Conference on Management of Data, pp. 251 - 262 , 2005.
- [7] D. Asonov, R. Agrawal, R. Srikant, "Enabling Sovereign Information Sharing Using Web Services", Proc. 23rd ACM SIGMOD International Conference on Management of Data, pp. 873-877, 2004.
- [8] A. Evfimievski, R. Agrawal, R. Srikant, "Information Sharing Across Private Databases", Proc. 22nd ACM SIGMOD International Conference on Management of Data, pp. 86-97, 2003.
- [9] O. Goldreich, “The Foundations of Cryptography”, vol. 2. Cambridge, United Kingdom: Cambridge University Press, 2004.
- [10] Frawley William J., Piatetsky-Shapiro Gregory, Matheus Christopher J., “Knowledge Discovery in Databases: An Overview.”, Knowledge Discovery in Databases, 1-30, 1991.

- [11] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., “Advances in data mining and Knowledge Discovery”, AAAI Pres, USA ,1994.
- [12] Berry, M.J.A. and Linoff, G.S., “Mastering Data Mining”, John Wiley & Sons, New York, 2000.
- [13] Han, J.; Kamber, M., “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers Inc., 2001.
- [14] Tan P. N., Kumar V., Steinbach M., “Introduction to Data Mining”, Addison-Wesley, 769 pp, 2005.
- [15] Rita C. Manning ,”Liberal and Communitarian Defenses of Workplace Privacy,” Journal of Business Ethics,16: 817-823,1997.
- [16] <http://www.inhak-bb.adalet.gov.tr>, “Avrupa İnsan Hakları Sözleşmesi Metni -Özel hayatın ve aile hayatının korunması”, 2006.
- [17] Slobodan Duji, “The Directive of the European Union on the Protection of Individual regarding the processing of Personal Data, published in, Public service”, , No. 1, 51-74, 1996.
- [18] Amitai Etzioni,”The Limits of Privacy”, New York: Basic Books, page 258,1999.
- [19] <http://epic.org/privacy/drivers/>, “Driver's Privacy Protection Act”,1994.
- [20] Angela R. Karras , “The Constitutionality of the Driver's Privacy Protection Act: A Fork in the Information Access Road”, 52 Fed. Comm. L.J. 125 1999.
- [21] <http://epic.org/privacy/vppa/>, “Video Privacy Protection Act”,1988.
- [22] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridis, “State-of-the-art in Privacy Preserving Data Mining”, SIGMOD Record 33(1): 50-57, 2004.
- [23] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, “Defining Privacy for Data Mining, Next Generation Data Mining”, AAAI/MIT Press, 2004.
- [24] Nan Zhang, Wei Zhao, “Privacy-Preserving Data Mining Systems”, IEEE Computer, Vol. 40, No. 4, Pages 52-58, 2007.

- [25] Charu C. Aggarwal, Philip S.Yu, "Privacy Preserving Data Mining: Models and Algorithms", Springer Publishing Company, 2008.
- [26] Jaideep Vaidya, Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When", IEEE Educational Activities Department, 2004.
- [27] Wenliang Du, Zhijun Zhan, "Building Decision Tree Classifier on Private Data", IEEE Int'l Conf. Privacy, Security, and Data Mining, pp. 1-8, 2002.
- [28] Wenliang Du, Zhijun Zhan, "Using Randomized Response Techniques for Privacy-Preserving Data Mining", ACM, 2003.
- [29] Yue Huang, Paul McCullagh, Norman Black, Roy Harper, "Feature Selection and Classification Model Construction on Type 2 Diabetic Patient's Data", Artificial Intelligence in Medicine 41(3): 251-262, 2007.
- [30] Nan Zhang, Wei Zhao, "Privacy-Preserving Data Mining Systems", IEEE Computer Society Press, 2007.
- [31] Nan Zhang, Wei Zhao, "Privacy Protection Against Malicious Adversaries in Distributed Information Sharing Systems", IEEE Transactions on Knowledge and Data Engineering, 2008.
- [32] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the Privacy Preserving Properties of Random Data Perturbation Techniques," Proc. Third IEEE International Conference on Data Mining, pp. 99-106, 2003.
- [33] <http://cml.ics.uci.edu>, "Yetişkin veri kümesi", 1994.
- [34] <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>, "Tanagra", 2005.
- [35] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.
- [36] J. Ross Quinlan "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers, Inc., 1993.
- [37] F. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, "Classification and Regression Trees", Chapman & Hall, 1998.
- [38] U. Kavza, H. Takçı, "Veri Madenciliğinde Mahremiyetin Sağlanması", 3. Uluslararası Bilgisayar ve Öğretim Teknolojileri Sempozyumu, 2009.

ÖZGEÇMİŞ

1983 doğumlu olan Uğur Kavza, Erciyes Üniversitesi Bilgisayar Mühendisliği 2006 lisans mezunu olup 2007 yılı güz döneminde Gebze Yüksek Teknoloji Enstitüsü Bilgisayar Mühendisliği Bölümü'nde yüksek lisans eğitimine başladı. 2006 yılından beri bilgisayar mühendisi olarak çalışmaktadır.