İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21(42), Güz 2022, 422-441.

İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

İstanbul Commerce University Journal of Science

http://dergipark.org.tr/ticaretfbd

Derleme Makalesi / Review Article

BÜYÜK VERİDE ANONİMLEŞTİRME TEKNİKLERİ VE SALDIRI TÜRLERİ: UYGULAMA ÖRNEKLERİ

ANONYMIZATION TECHNIQUES AND ATTACK TYPES IN BIG DATA: APPLICATION EXAMPLES

Hamza Talha GÜMÜŞ[1] **Can EYÜPOĞLU[2]**

https://doi.org/10.55071/ticaretfbd.1086750

Sorumlu Yazar / Corresponding Author
ceyupoglu@hho.msu.edu.tr

Geliş Tarihi / Received
12.03.2022

Kabul Tarihi / Accepted
06.07.2022

Öz

Veri kavramı ortaya çıktığından beri kişiye özgü olan bilgilerimiz işlenmeye başlanmıştır. Veri kavramı sonraki yıllarda kişisel veri konusunu oluşturarak veri güvenliği ve mahremiyeti kavramlarının ortaya çıkmasına sebep olmuştur. Kişiye ait ve özgü olarak tanımlanan kişisel veri kavramı mahremiyetin önemini bir kez daha vurgulamıştır. Gizlilik ya da gizli olma durumu şeklinde tanımlanan mahremiyet kavramı, her geçen gün veriler geliştikçe ve arttıkça daha fazla oranda önem kazanmaktadır. Son yıllarda araştırmacılar tarafından farklı anonimleştirme teknikleri geliştirilmiş ve bu teknikler sayesinde veri koruması artırılmıştır. Veri mahremiyeti önem kazanırken veri hırsızlığı kavramı da ortaya çıkmış ve belirli saldırı türleri geliştirilmiştir. Bu saldırı türlerine yönelik geliştirilen anonimleştirme teknikleri, veri kaybına yol açsa da kişisel veriye ulaşma ihtimalini büyük ölçüde azaltmaktadır. Bu çalışmada büyük veride anonimleştirme teknikleri ve saldırı türleri incelenmiş ve mahremiyet koruması konusu üzerinde durulmuştur.

Anahtar Kelimeler: Anonimleştirme teknikleri, büyük veri gizliliği, büyük veride saldırı türleri, mahremiyet koruması.

Abstract

Since the concept of data emerged, our personal information has been started to be processed. The concept of data has created the subject of personal data in the following years, leading to the emergence of the concepts of data security and privacy. The concept of personal data, which is defined as personal and private, has once again emphasized the importance of privacy. The concept of privacy, which is defined as confidentiality or the state of being confidential, gains more and more importance as data develop and increase day by day. In recent years, different anonymization techniques have been developed by researchers and data protection has been increased thanks to these techniques. While data privacy has gained importance, the concept of data theft has also emerged and certain types of attacks have been developed. Anonymization techniques developed for these types of attacks greatly reduce the possibility of accessing personal data, even if it leads to data loss. In this study, anonymization techniques and attack types in big data are examined and privacy protection is emphasized.

Keywords: Anonymization techniques, attack types in big data, big data privacy, privacy protection.

[1]Milli Savunma Üniversitesi, Hezârfen Havacılık ve Uzay Teknolojileri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, İstanbul, Türkiye. hamzatalhagumus@gmail.com, Orcid.org/0000-0001-7360-8138.

[2]Milli Savunma Üniversitesi, Hava Harp Okulu, Bilgisayar Mühendisliği Bölümü, Yeşilyurt, İstanbul, Türkiye. ceyupoglu@hho.msu.edu.tr, caneyupoglu@gmail.com, Orcid.org/0000-0002-6133-8617.

1. GİRİŞ

Günümüzde teknolojinin gelişmesi ile veri setleri büyük önem kazanmıştır. Veriler kullanılırken hali hazırda ülkemizde ve ülke birlik kuruluşları tarafından kişisel bilgilerin korunması için belirli çalışmalar yapılmış ve tasarılar sunulmuştur. Genel anlamda veri setleri içerisinde bulunan kişisel bilgilerin paylaşılması sınırlanmış ve paylaşım sağlanması için gerekli şartlar belirlenmiştir. En bilinen kanunlar dünyada Avrupa Birliği Genel Veri Koruma Tüzüğü (General Data Protection Regulation-GPDR) ve ülkemizde ise 6698 sayılı Kişisel Verilerin Korunması Kanunudur (KVKK). Bu kanunlar gereğinde veriler paylaşılırken belirli yöntemlerin kullanılması bazı verilerin gizlenmesi şart koşulmuştur. Bir insana ait olan veriler kişisel olup izni dışında paylaşması yasaklanmış fakat ticari durum göz önüne alındığında verilerin içerisinde bulunan bazı bilgiler anonimleştirilerek paylaşılmasına izin verilmiştir. Aynı durum söz konusu iken veri paylaşımları sırasında ya da veri kümesinin bulunduğu kişi veya kurum içerisinde de verinin korunması şart koşulmuştur. Kanunlaşma süreci henüz yeni olsa da veriler uzun yıllardır korunmaya çalışılmaktadır. İlk olarak 1977 yılında Dalenius mahremiyetin korunmasını "Yayımlanan bir veri kümesi, arka planda başka kaynaklardan bilgiler elde etmiş olsa bile bir saldırgana, o veri kümesine erişimi yokmuş gibi veri sahipleri ile ilgili herhangi bir ekstra bilgi edinmesine izin vermemelidir" şeklinde açıklamıştır. Veriler her ne kadar anonimleştirilmiş olsa da belirli bilgiler ve var olan eski verilerin karşılaştırılması sonucunda açığa çıkabilmektedir. Bu durumu engellemek için belirli anonimleştirme yöntemleri geliştirilmiş ve mahremiyet sağlanmaya çalışılmıştır. Anonimleştirme teknikleri artarken veriye düzenlenebilecek saldırı yöntemleri de aynı düzeyde artmış ve saldırı yöntemlerine karşı da teknikler güçlendirilmiştir.

Çalışmanın diğer bölümleri şu şekilde düzenlenmiştir: Bölüm 2'de nitelik türleri, anonimleştirme teknikleri, anonim hale getirme yönteminin seçilmesi, bilgi kaybını ölçme ve saldırı türleri ele alınmaktadır. Bölüm 3'te literatürdeki çalışmalar özetlenerek karşılaştırılmaktadır. Bölüm 4'te ise çalışmanın genel sonuçlarından bahsedilmektedir.

2. BÜYÜK VERİDE GİZLİLİK KORUMASI

Veri koruması belirli kanunlar çerçevesinde gerçekleşmektedir. Fakat anonimleştirme teknikleri ile verilerde kişisel veriler belirli aralıklar ile anonimleştirilerek verinin kişiye ait olduğunu anlama ihtimali düşürülmektedir. Veri koruma iki şekilde gerçekleştirilmektedir. İlk aşama gizlilik korumalı veri koruma şeklindedir. Burada veriler anonimleştirilerek tespit edilmesi zor hale getirilmektedir. Bir diğer aşama olan veriden bilgi çıkarmada ise ilk aşamada anonimleştirilen verilerden bilgilerin çıkarılması şeklindedir. Çıkan bilgi kullanımına göre anlamlı veri sınıfındadır (Eyüpoğlu ve ark., 2017).

2.1. Nitelik Türleri

Bir küme halinde bulunan veriler dört farklı sınıfta değerlendirilmektedir. Bunlar; doğrudan ya da açık tanımlayıcı, dolaylı ya da yarı tanımlayıcı, hassas nitelikler ve hassas olmayan nitelikler olarak sınıflandırılmaktadır (Afyonluoğlu, 2019; Eyüpoğlu, 2018).

Açık tanımlayıcı (explicit identifier–ID): Bir kişinin kimliğini belirten niteliklerdir. Bir başka deyişle kişiye ait olan ve kişiden başkasında bulunmayan, belirli bir kişiyi o bilgiyle tanımlayabileceğimiz veriler bu sınıf içerisinde yer almaktadır. Örnek olarak, TC kimlik numarası, pasaport numarası, ehliyet numarası, cep telefon numarası, sosyal güvenlik numarası, isim, soy isim gibi nitelikler açık tanımlayıcı nitelik olarak değerlendirilmektedir (Afyonluoğlu, 2019; Eyüpoğlu, 2018).

Yarı tanımlayıcı (quasi-identifier–QID): Kişiyi tam olarak tanımlamayan yanında gerekli olan bilgiler ile ya da başka veri setlerinden elde edilebilecek veriler ile kişinin tanımlanmasını sağlayan veriler bu sınıf içerisinde yer almaktadır. Yarı tanımlayıcı veriler sadece kendileri kullanılarak bir anlam ifade etmemektedir ve bir kişiyi tanımlamamaktadır. Tanımlama yapılabilmesi için kişi ile ilgili önceden bilgi sahibi olunması ya da farklı veri kümelerinde bulunan eşleşmeler ile kişisel bilgiye ulaşarak kişi tanımlanabilmektedir. Örnek olarak, yaş, adres, posta kodu, cinsiyet, doğum tarihi, doğum yeri, medeni hal ve meslek verilebilmektedir (Afyonluoğlu, 2019; Eyüpoğlu, 2018).

Hassas nitelikler (sensitive attributes–SA): Kişiye özel olan, kişiye bağlı paylaşılan veya paylaşılmayan, kişinin hassas bilgileri bu sınıf içerisinde yer almaktadır. Gelir bilgisi ve sağlık bilgisi en bilinen hassas niteliklerdir (Afyonluoğlu, 2019; Eyüpoğlu, 2018).

Hassas olmayan nitelikler (non-sensitive attributes–NSA): Belirlenmesi ya da ele geçmesi durumunda kişiye ilişkin bilgi edinilemeyen verilere hassas olmayan nitelikler denilmektedir. Açık tanımlayıcı, yarı tanımlayıcı ve hassas nitelik olmayan tüm nitelikler bu nitelik sınıfı içerisinde yer almaktadır (Afyonluoğlu, 2019; Eyüpoğlu, 2018).

Veriler çok farklı alanlarda kullanılsa da literatürde en önemli olan verinin, sağlık verisi olduğu yapılan araştırmalardan görülmektedir. Yayımlanan makale ve tez çalışmalarında, veri anonimliği içinde sağlık alanı diğer alanlara göre daha fazla incelenmiştir. Bu incelemenin sebebi hassas veri dışında açık tanımlayıcı ve yarı tanımlayıcı verilerin de sağlık alanında bulunmasıdır. Sosyoekonomik veriler olarak bilinen veriler kişiyi tanımlar niteliktedir. Bu verilere örnek olarak isim, adres, doğum tarihi, aile, ırk, cinsiyet, evlilik durumu, meslek, gelir kaynağı, etnik köken, eğitim ve iş durumu verilebilir. Bir diğer durum sosyal medya ağlarında da mevcuttur. Günümüzde teknolojinin gelişimi ile gelen dijitalleşme sürecinde ortaya çıkan mevcut ağlarda da kişi açık tanımlayıcı, tanımlayıcı ve hassas nitelikli verilerini kullanmaktadır (Afyonluoğlu, 2019).

Sağlık alanı ve iş sosyal ağ platformu üzerine işlenen verilerde bir karşılaştırma yapılarak açık tanımlayıcı, yarı tanımlayıcı ve hassas veriler Tablo 1'de gösterilmektedir (Ünal, 2017).

Tablo 1. Tanımlayıcı Nitelikler ile Tıbbı Veri ve Sosyal Ağ Karşılaştırılması

	Tıbbı Veriler	**İş Sosyal Ağ Platformu**
Tanımlayıcı	İsim, soy isim, TC kimlik numarası	İsim, soy isim, cep telefon numarası, e-posta adresi
Yarı Tanımlayıcı	Yaş, adres, posta kodu, cinsiyet, doğum tarihi, doğum yeri, medeni hal, meslek bilgisi	Yaş, adres, cinsiyet, doğum tarihi, posta kodu, meslek bilgisi
Hassas Nitelik	Hastanın kimliği, öyküsü, şikâyetleri, ameliyat raporları, patoloji raporları, acil servis raporları, progres notları, yardımcı hizmet raporları, sosyal hizmet değerlendirmeleri, bakımını kim yapmış, taburculuk özeti, değerlendirme raporları; Hastaya sunulan hizmet ne, maliyeti ne kadar, ne zaman sunulmuş, nerede sunulmuş, hizmetin gerekçesi ne, çıktıları veya etkisi ne olmuş	Sağlık bilgisi, gelir durumu

Tanımlayıcı, yarı tanımlayıcı ve hassas niteliklerden oluşan örnek bir veri seti Tablo 2'de gösterilmektedir.

Tablo 2. Tanımlayıcı, Yarı Tanımlayıcı ve Hassas Nitelik Örnek Tablo

Tanımlayıcı			**Yarı Tanımlayıcı**					**Hassas Nitelik**	
TC Kimlik No	**İsim**	**Soy İsim**	**Cinsiyet**	**Yaş**	**Posta Kodu**	**Meslek**	**Medeni Hali**	**Sağlık Bilgisi**	**Gelir Bilgisi**
12345678902	Ali	Yılmaz	Erkek	25	06310	Bilgisayar Mühendisi	Bekâr	Hepatit	4250
23456789012	Mehmet	Gündüz	Erkek	29	06370	Ressam	Bekâr	Akciğer Kanseri	3500
34567890124	Veli	Kaya	Erkek	36	06050	Elektrik Ustası	Evli	Lenf Kanseri	5750
45678901234	Ayşe	Demir	Kadın	23	06120	Elektronik Mühendisi	Bekâr	Hepatit	4250
56789012346	Sevim	Şahin	Kadın	27	06165	Oyuncu	Evli	Grip	14650
67890123456	Ülkü	Çelik	Kadın	32	06210	Elektrik Teknisyeni	Bekâr	Beyin Tümörü	5250
78901234568	Mustafa	Aslan	Erkek	30	06260	Makine Teknisyeni	Evli	Hepatit	5500
89012345678	Ahmet	Çetin	Erkek	43	06340	İnşaat Ustası	Evli	Astım	5875
90123456780	Zeynep	Kara	Kadın	45	06450	Tarih Öğretmeni	Bekâr	Ülser	5300
12345678990	Fatma	Doğan	Kadın	48	06378	Edebiyat Öğretmeni	Evli	Gastrit	5600

2.2. Anonimleştirme Teknikleri

Veri setlerini anonimleştirmek için farklı yöntemler uygulanmaktadır. Bu yöntemler genel olarak beş başlık halinde tanımlanmıştır. Bu başlıklar; genelleme, gizleme, anatomizasyon, permütasyon, pertürbasyon şeklindedir. Fakat ülkemizde bulunan kişisel verileri koruma kurumu üç üst başlık şeklinde anonimleştirme yöntemlerini tanımlamıştır. Bunlar; değer düzensizliği sağlamayan anonim hale getirme yöntemleri, değer düzensizliği sağlayan anonim hale getirme yöntemleri ve anonim hale getirmeyi kuvvetlendirici istatistik yöntemler şeklindedir. Bu başlıkları altında dağılım yapılırken, değer düzensizliği sağlamayan anonim hale getirme yöntemleri; değişkenleri çıkartma, kayıtları çıkartma, alt ve üst sınır kodlama, bölgesel gizleme, örnekleme, değer düzensizliği sağlayan anonim hale getirme yöntemleri; mikro-birleştirme, veri değiş-tokuşu, gürültü ekleme, tekrar örnekleme ve anonim hale getirmeyi kuvvetlendirici istatistik yöntemler; *k*-anonimlik, *l*-çeşitlilik, *t*-yakınlık, diferansiyel gizlilik şeklindedir. Farklı kaynaklarda genelleştirme, baskılama, kovalara ayırma ve hibrit yöntemler de bulunmaktadır. Anonimleştirme teknikleri başlığı altında yirmi bir farklı teknik incelenmiştir (Eyüpoğlu, 2018; KVKK, 2017).

2.2.1. Literatürdeki anonimleştirme teknikleri

Bu bölümde genelleme, genelleştirme, gizleme, baskılama, anatomizasyon, permütasyon, pertürbasyon, kovalara ayırma ve hibrit yöntemler ele alınmaktadır (Eyüpoğlu, 2018; KVKK, 2017; Vural, 2018).

2.2.1.1. Genelleme

Genelleme, veri setinde QID değerlerinin yani yarı tanımlayıcı niteliklerin bütünlüğünün korunarak genel anlam ile ifade edilmesidir. Yarı tanımlayıcı nitelikler üst anlamlarda kullanılarak genelleme yapılması şeklinde gerçekleştirilen anonimleştirme tekniğidir (KVKK, 2017). Örnek olarak yarı tanımlayıcı bir nitelik olan meslek bilgisinde bilgisayar mühendisi ve elektronik mühendisi bulunsun. Genelleştirme yapılırken temel meslek genellenerek mühendis olarak yazılmaktadır. Bu en basit genelleştirme örneğidir. Tablo 2'deki meslek verilerinin

genelleştirilmiş hali Tablo 3'te gösterilmektedir. Genelleştirme, farklı kaynaklarda genelleme anlamında kullanılan anonimleştirme tekniğidir (Vural, 2018).

Tablo 3. Genelleme Örneği

İsim	**Soy İsim**	**Meslek**	**Genelleştirilmiş Meslek Bilgisi**
Ali	Yılmaz	Bilgisayar Mühendisi	Mühendis
Mehmet	Gündüz	Ressam	Sanatçı
Veli	Kaya	Elektrik Ustası	Usta
Ayşe	Demir	Elektronik Mühendisi	Mühendis
Sevim	Şahin	Oyuncu	Sanatçı
Ülkü	Çelik	Elektrik Teknisyeni	Teknisyen
Mustafa	Aslan	Makine Teknisyeni	Teknisyen
Ahmet	Çetin	İnşaat Ustası	Usta
Zeynep	Kara	Tarih Öğretmeni	Öğretmen
Fatma	Doğan	Edebiyat Öğretmeni	Öğretmen

2.2.1.2. Gizleme

Gizleme işleminde, verilerden alınan tanımlayıcı ve yarı tanımlayıcı nitelikler bir karakter yardımı ile gizlenmektedir. Örnek olarak bir kişinin isim ve soy isim bilgileri bir karakter ile değiştirilir ve kişinin bilgisindeki tanımlayıcı nitelik kaldırılmış olur (KVKK, 2017). Tablo 4'te gizleme örnekleri verilmiştir. Tablonun E kısmında kayıt gizleme, D kısmında hücre gizleme ve A, B, C kısımlarında posta kodu için değer gizleme uygulanmıştır. Baskılama, farklı kaynaklarda gizleme anlamında kullanılan anonimleştirme tekniğidir (Vural, 2018).

Tablo 4. Gizleme Örneği

	Tanımlayıcı		**Yarı Tanımlayıcı**		
	İsim	**Soy İsim**	**Yaş**	**Posta Kodu**	**Meslek**
A	Ali	Yılmaz	25	06***	Bilgisayar Mühendisi
B	Mehmet	Gündüz	29	06***	Ressam
C	Veli	Kaya	36	06***	Elektrik Ustası
D	****	*****	23	06120	Elektronik Mühendisi
E	*****	*****	*	*	Oyuncu
F	Ülkü	Çelik	32	06210	Elektrik Teknisyeni

2.2.1.3. Anatomizasyon

Anatomizasyon, tanımlayıcı nitelikler dışında, yarı tanımlayıcı ve hassas niteliklerin arasındaki anlamlı bağlantının koparılması ile gerçekleştirilmektedir. Yöntem iki farklı tablo olarak yayımlanma ile yapılmaktadır. Yarı tanımlayıcı ve hassas nitelikler farklı tablolarda yayımlanır. Aralarındaki bağlantı verilere verilen grup numarası ile sağlanmaktadır. Gerekli durumlarda veriler ayrı ayrı paylaşılmaktadır (Eyüpoğlu, 2018; KVKK, 2017). Kovalara ayırma, farklı kaynaklarda anatomizasyon anlamında kullanılan anonimleştirme tekniğidir (Vural, 2018). Tablo 5'te yarı tanımlayıcı nitelik ve Tablo 6'da hassas nitelik tablosu Tablo 2'den ayrılarak yapılmıştır. Verilen sıra numaraları ile tablolar karşılaştırıldığında yarı tanımlayıcı ve hassas niteliklerin kime ait olduğu belirlenmese bile nitelikler anlamlandırılmaktadır.

Tablo 5. Yarı Tanımlayıcı Nitelik

Yarı Tanımlayıcı			
Sıra No	**Cinsiyet**	**Yaş**	**Posta Kodu**
1	Erkek	25	06310
2	Erkek	29	06370
3	Erkek	36	06050
4	Kadın	23	06120
5	Kadın	27	06165

Tablo 6. Hassas Nitelik

Hassas Nitelik		
Sıra No	**Sağlık Bilgisi**	**Gelir Bilgisi**
1	Hepatit	4250
2	Akciğer Kanseri	3500
3	Lenf Kanseri	5750
4	Hepatit	4250
5	Grip	14650

2.2.1.4. Permütasyon

Permütasyon yönteminde, anatomizasyonda olduğu gibi yarı tanımlayıcı ve hassas nitelikler kullanılmaktadır. Genelleme ve gizleme yöntemi birlikte kullanılmaktadır. Veri içerisinde gruplara ayırma gerçekleştirilir ve değerler karıştırılır. Anlamlı bilgi yeterliliğini kaybeder, fakat veriler grup içerisinde korunduğundan işlevi kaybolmamaktadır (Eyüpoğlu, 2018). Hibrit yöntem, farklı kaynaklarda permütasyon anlamında kullanılan anonimleştirme tekniğidir (Vural, 2018). Tablo 7'de Tablo 2 üzerinde permütasyon yöntemi kullanılarak anonimleştirilen veri seti görülmektedir.

Tablo 7. Permütasyon Örneği

Cinsiyet	**Yaş**	**Sağlık Bilgisi**	**Gelir Bilgisi**
Erkek	<30	Enfeksiyon	<5000
Erkek	<30	Kanser	<5000
Erkek	>30	Kanser	>5000
Kadın	<30	Enfeksiyon	<5000
Kadın	<30	Enfeksiyon	>5000
Kadın	>30	Kanser	>5000

2.2.1.5. Pertürbasyon

Pertürbasyon, büyük veri içerisinde bulunan verilerin anlamsızlaştırılması ile gerçekleştirilmektedir. Veriler farklı veri değerleri ile değiştirilerek saldırılara karşı koruma sağlanması düşünülmektedir. Fakat istatiksel olarak farklılık minimum düzeydedir (Eyüpoğlu, 2018). Pertürbasyon örnekleri Bölüm 2.2.3'te ele alınmaktadır.

2.2.2. Değer düzensizliği sağlamayan anonim hale getirme yöntemleri

Bu yöntemlerde, verilerde ekleme/çıkarma işlemleri yapılmamaktadır. Veri kümesi içerisinde bulunan değer sütun ya da değer satırlarında değişiklik yapılarak verilere anonimleştirme

uygulanmaktadır. Sadece ilgili alan korunur ve verinin genel anlamda bütününde bozulma sağlayarak anonimleştirme gerçekleştirilir. İlgili alandaki veriler işlemler sonrasında anlamlı veri olarak bütünlüğünü korumaktadır. Örnek olarak değişkenleri çıkartma, kayıtları çıkartma, alt ve üst sınır kodlama, bölgesel gizleme ve örnekleme yöntemleri mevcuttur (KVKK, 2017).

2.2.2.1. Değişkenleri çıkartma

Değişkenleri çıkartma, veri içerisinde bulunan değişkenlerin çıkarılmasıyla elde edilmektedir. Bir ya da daha fazla değişken tamamen çıkarılarak oluşturulan anonimleştirme tekniğidir. Kamu niteliğinde hassas verilerde, istatiksel yöntem dışından bir alanda kullanılacak verilerde ve uygun tekniğin olmadığı verilerde kullanılmaktadır (KVKK, 2017). Tablo 8'de değişken çıkartma örneği gösterilmektedir.

Tablo 8. Değişken Çıkartma Örneği

TC Kimlik No	**İsim**	**Soy İsim**
12345678902	Ali	Yılmaz
23456789012	Mehmet	Gündüz
34567890124	Veli	Kaya
45678901234	Ayşe	Demir
56789012346	Sevim	Şahin

2.2.2.2. Kayıtları çıkartma

Kayıtları çıkartma, veri kümesi içerisinde bulunan ve tekil olan kayıtların ortadan kaldırılması ile gerçekleştirilir. Böylece saldırı esnasında ifşanın kolaylaşması engellenir ve anonimlik güçlendirilir. Veri kümesi diğer kayıtlarla karşılaştırılsa bile kayıt bulunamadığı için tahmin yok seviyesine erişilebilmektedir (KVKK, 2017). Örnek olarak bir veri kümesinde meslek bilgisi önce genelleştirme tekniği uygulanarak düzenlensin. Sonraki aşamada Tablo 9'da tekillik ifade eden sanatçı bilgisi görülmektedir. Meslek bilgisi yerine sadece sanatçının olduğu satırın çıkarılması kayıt çıkarmaya örnek olarak verilebilir.

Tablo 9. Kayıtları Çıkartma Örneği

İsim	**Soy İsim**	**Genelleştirilmiş Meslek Bilgisi**
Ali	Yılmaz	Mühendis
Mehmet	Gündüz	Sanatçı
Ayşe	Demir	Mühendis
Ülkü	Çelik	Teknisyen
Mustafa	Aslan	Teknisyen

2.2.2.3. Alt ve üst sınır kodlama

Alt ve üst sınır kodlama, veri kümesi içerisinde bir değişken tanımlanmasıyla, bu değişken grubu içerisinde bulunan değerleri birleştirerek elde etme yöntemidir. Genel olarak değişken düşük veya yüksek olarak tanımlanır ve yeni yapılan tanımlama ile değişkenlerin değerleri değiştirilerek anonimleştirme sağlanır (KVKK, 2017). Örnek olarak Tablo 10'da bir veriye ait yaş ve gelir bilgisi görülmektedir. Net olarak yaş için 32 değeri, gelir bilgisi için 5000 değeri referans alındığında, 32 yaş altı ve eşiti küçük, 32 yaş üstü ise büyük, aynı şekilde gelir bilgisinde 5000 altı ve eşiti düşük, 5000 üstü yüksek olarak tanımlansın. Bu tanımlama

sonucunda Tablo 10'daki yaş ve gelir bilgisi niteliklerinin alt ve üst sınır kodlama uygulanmış değerleri Tablo 11'deki olmaktadır.

Tablo 10. Yaş ve Gelir Bilgisi

İsim	Soy İsim	Yaş	Gelir Bilgisi
Ali	Yılmaz	25	4250
Mehmet	Gündüz	29	3500
Veli	Kaya	36	5750
Ayşe	Demir	23	4250
Sevim	Şahin	27	14650
Ülkü	Çelik	32	5250
Mustafa	Aslan	30	5500
Ahmet	Çetin	43	5875
Zeynep	Kara	45	5300
Fatma	Doğan	48	5600

Tablo 11. Alt ve Üst Kodlama Örneği

İsim	Soy İsim	Yaş	Gelir Bilgisi
Ali	Yılmaz	Küçük	Düşük
Mehmet	Gündüz	Küçük	Düşük
Veli	Kaya	Büyük	Yüksek
Ayşe	Demir	Küçük	Düşük
Sevim	Şahin	Küçük	Yüksek
Ülkü	Çelik	Küçük	Yüksek
Mustafa	Aslan	Küçük	Yüksek
Ahmet	Çetin	Büyük	Yüksek
Zeynep	Kara	Büyük	Yüksek
Fatma	Doğan	Büyük	Yüksek

2.2.2.4. Bölgesel gizleme

Bölgesel gizleme, tahmin edilme olasılığını düşürmek ve anonimliği artırmak için kullanılan bir tekniktir. Bir durum kişi ile ilgili bilgi verirken aynı zamanda aile ile ilgilide de bilgi verdiğinde çevre tarafından ifşa edilme ve tahmin edilme olasılığı arttığı durumlarda kullanılmaktadır (KVKK, 2017). Örnek olarak Tablo 12'de kistik fibroz hastalığı şüphesi olan bireyler ve test sonuçları görülmektedir. Sadece genetik olarak aktarılan bu hastalık, kişi ile ilgili bilgi verdiği için aile hakkında da bilgi vermektedir. Ayrıca 1. ve 4. kayıtlar çocuklar ile ilgili olduğundan istisnai durum yaratmaktadır ve çıkarım yapılma ihtimalini artırmaktadır. Tablo 13'te bölgesel gizleme yöntemi kullanılmış ve çocuklara ait kayıtlarda gizleme gerçekleştirilmiştir.

Tablo 12. Kistik Fibroz Test Sonuçları

Cinsiyet	Yaş	Meslek	Test Sonucu
Erkek	7	Öğrenci	Pozitif
Erkek	29	Ressam	Pozitif
Erkek	21	Öğrenci	Negatif
Kadın	6	Öğrenci	Pozitif
Kadın	27	Oyuncu	Negatif

Tablo 13. Bölgesel Gizleme Örneği

Cinsiyet	Yaş	Meslek	Test Sonucu
Erkek	-	Öğrenci	Pozitif
Erkek	29	Ressam	Pozitif
Erkek	21	Öğrenci	Negatif
Kadın	-	Öğrenci	Pozitif
Kadın	27	Oyuncu	Negatif

2.2.2.5. Örnekleme

Örnekleme yönteminde, bir veri kümesine ait kümeden bir küme oluşturarak işleme başlanılmaktadır. İlk veri kümesi içerisinde bulunan herkes bu alt kümede yer alamayacağı için tahmin etme olasılığı düşürülmüş olur (KVKK, 2017). Örnek olarak Z kuşağı üzerinde araştırma yapmak isteyen bir firma açık bir veri kümesine erişerek bilgilere ulaşabilmektedir. Doğum yerleri üzerinden araştırma yapmayı planlayan firma üzerinden yapılacak işlemde yaşanılan yer bilgisi de bulunmaktadır. Eğer örnekleme ile anonimleştirme gerçekleşmişse bu bilgilerden birisi yani doğum yeri ve yaşanan yer bilgilerinden birisi yok edilecek ve tahmin olasılığı düşürülecektir.

2.2.2.6. Global kodlama

Global kodlama, Bölüm 2.2.2.3'te anlatılan alt ve üst sınır kodlama yönteminin farklı bir türevidir. Alt ve üst sınır kodlamada rakamsal değerler söz konusu iken rakam içermeyen durumlarda global kodlama kullanılmaktadır. Tahmin olasılığını düşürmek için kullanılmaktadır. Veri seti içerisinde bulunan değerler genellemede olduğu gibi farklılaşır ya da ortak anlamda buluşarak verideki değerler değişmektedir (KVKK, 2017). Tablo 14'te görülen hekim bilgileri Tablo 15'te global kodlama ile düzenlenmiş ve tahmin olasılığını düşürülmüştür.

Tablo 14. Hekim Bilgileri

Cinsiyet	Meslek	Medeni Hali
Erkek	Diş Hekimi	Bekâr
Erkek	Psikiyatrist	Bekâr
Erkek	Hematolog	Evli
Kadın	Cerrah	Bekâr
Kadın	Kardiyolog	Evli
Kadın	Nörolog	Bekâr

Tablo 15. Global Kodlama Örneği

Cinsiyet	Meslek	Medeni Hali
Erkek	Hekim	Bekâr
Erkek	Hekim	Bekâr
Erkek	Hekim	Evli
Kadın	Hekim	Bekâr
Kadın	Hekim	Evli
Kadın	Hekim	Bekâr

2.2.3. Değer düzensizliği sağlayan anonim hale getirme yöntemleri

Bu yöntemler, genel olarak bozulma üzerine gerçekleştirilen anonimleştirme teknikleridir. Veri kümesi üzerindeki istatistiksel sonuçlar değişmeden kayıt değerlerinde bozulma yapılması amaçlanmaktadır. Mikro birleştirme, veri değiş tokuşu, gürültü ekleme ve tekrar örnekleme gibi farklı yöntemleri mevcuttur (KVKK, 2017).

2.2.3.1. Mikro birleştirme

Mikro birleştirme yönteminde, elde olan veri kümesi içerisindeki veriler, öncelikle anlam oluşturacak şekilde sıralanmaktadır. Sonra öncelik olarak alt kümeler oluşturulur. Oluşturulan alt kümelerde seçilen değişken değerlerin ortalaması alınarak eldeki veri ile değiştirilmektedir. Bu sayede istatiksel durumlarda çıkacak olan sonuç da değişmeyecektir (KVKK, 2017). Tablo 2 üzerinde gelir bilgisine göre sıralama yapılarak elde edilen veri kümesi Tablo 16'daki gibidir. Tablo 17'de ise bu veri kümesinin mikro birleştirme tekniği uygulanmış hali gösterilmektedir. Burada ilk olarak gelir bilgisi için yakın değerlerin olduğu kayıtlar üçerli olarak gruplandırılmış ve ardından bu grupların ortalama değeri ilgili kayıtlardaki değerler ile değiştirilmiştir.

Tablo 16. Düzenlenmiş Veri Kümesi

Cinsiyet	**Yaş**	**Posta Kodu**	**Meslek**	**Gelir Bilgisi**
Erkek	25	06310	Bilgisayar Mühendisi	4250
Kadın	23	06120	Elektronik Mühendisi	4250
Kadın	45	06450	Tarih Öğretmeni	5300
Kadın	48	06378	Edebiyat Öğretmeni	5600
Erkek	36	06050	Elektrik Ustası	5750
Erkek	43	06340	İnşaat Ustası	5875

Tablo 17. Mikro Birleştirme Örneği

Cinsiyet	**Yaş**	**Posta Kodu**	**Meslek**	**Gelir Bilgisi**
Erkek	25	06310	Bilgisayar Mühendisi	4600
Kadın	23	06120	Elektronik Mühendisi	4600
Kadın	45	06450	Tarih Öğretmeni	4600
Kadın	48	06378	Edebiyat Öğretmeni	5742
Erkek	36	06050	Elektrik Ustası	5742
Erkek	43	06340	İnşaat Ustası	5742

2.2.3.2. Veri değiş tokuşu

Veri değiş tokuşu, veri kümesi içerisinde benzer bilgiler içeren kayıtlardaki bir değişkenin, diğer bir kayıttaki değişken ile değiştirilmesidir (KVKK, 2017). Bir meslek grubundaki bireyin gelir bilgisinin, aynı meslek grubu içerisindeki diğer bir bireyin meslek bilgisi ile değiştirilmesi veri değiş tokuş işlemine örnek olarak gösterilebilir. Tablo 18'de meslek ve gelir bilgisinin olduğu veri kümesi görülmektedir. Veri değiş tokuş işlemi sonrasında oluşan veri kümesi ise Tablo 19'da gösterilmektedir. Tabloda görüldüğü üzere mesleğin usta olduğu 2. ve 4. kayıtların gelir bilgileri değiş tokuş yapılmıştır. Ayrıca mesleğin öğretmen olduğu 5. ve 6. kayıtlarda da gelir bilgileri birbirleri ile değiştirilmiştir.

Tablo 18. Meslek ve Gelir Bilgisi

Cinsiyet	Yaş	Meslek	Gelir Bilgisi
Erkek	25	Mühendis	4250
Erkek	36	Usta	5750
Kadın	23	Mühendis	4250
Erkek	43	Usta	5875
Kadın	45	Öğretmen	5300
Kadın	48	Öğretmen	5600

Tablo 19. Veri Değiş Tokuşu Örneği

Cinsiyet	Yaş	Meslek	Gelir Bilgisi
Erkek	25	Mühendis	4250
Erkek	36	Usta	5875
Kadın	23	Mühendis	4250
Erkek	43	Usta	5750
Kadın	45	Öğretmen	5600
Kadın	48	Öğretmen	5300

2.2.3.3. Gürültü ekleme

Gürültü ekleme, veri kümesi içerişinde belirli bozulmalar yapılarak sağlanmaktadır. Bu bozulma ekleme ve çıkarma işlemi sonucunda oluşmaktadır. Yöntem, sayısal yani rakamsal değerler üzerinde gerçekleşmektedir ve tüm değerlere eşit şekilde uygulanmaktadır (KVKK, 2017). Örnek olarak, Tablo 2'de bulunan gelir bilgisi sayısal niteliğini kullanalım. Tablo 2'nin düzenlenmiş hali Tablo 20'de görülmektedir. Tablo 21'de ise gürültü ekleme yöntemi kullanılarak verilerde bozulma yapılmıştır. Bu bozulma gelir düzeyine +1000 işlemi uygulanarak gerçekleştirilmiştir.

Tablo 20. Gelir Bilgisi Tablosu

Cinsiyet	Yaş	Meslek	Gelir Bilgisi
Erkek	25	Mühendis	4250
Erkek	36	Usta	5750
Kadın	23	Mühendis	4250
Erkek	43	Usta	5875
Kadın	45	Öğretmen	5300
Kadın	48	Öğretmen	5600

Tablo 21. Gürültü Ekleme Örneği

Cinsiyet	Yaş	Meslek	Gelir Bilgisi
Erkek	25	Mühendis	5250
Erkek	36	Usta	6750
Kadın	23	Mühendis	5250
Erkek	43	Usta	6875
Kadın	45	Öğretmen	6300
Kadın	48	Öğretmen	6600

2.2.4. Anonim hale getirmeyi kuvvetlendirici istatistik yöntemler

Anonim hale getirmeyi kuvvetlendirici istatistik yöntemler, kişisel verilerde güvenlik sağlama amacıyla geliştirilmiştir. Bu yöntemler kullanılarak veri kümesi içerisinde bulunan kişilerin kimlik tespitinin yapılma olasılığı düşürülmeye çalışılmaktadır. Bu sayede anonimlik güçlendirilmektedir. Bir diğer amaç ise küme içerisinden elde edilecek faydanın da yüksek tutulmasıdır. *k*-anonimlik, *l*-çeşitlilik, *t*-yakınlık ve diferansiyel gizlilik gibi teknikler, anonim hale getirmeyi kuvvetlendirici istatistik yöntemlerden bazılarıdır (Eyüpoğlu, 2018; KVKK, 2017; Vural, 2018).

2.2.4.1. *k*-anonimlik

Veri anonimleştirmede birçok farklı model kullanılmaktadır. Bu modellerin başında *k*–anonimlik modeli gelmektedir. *k*–anonimlik, Sweeney (2002) tarafından önerilmiş ve geliştirilmiştir. Oluşturulan model ile belirli bir veri kümesinde bulunan veriler anonimleştirilmektedir. Bu modelde veri kümesinde bulunan belirli ölçütlere göre kişisel verilerin ya da kişiye özgü verilerin tanımlanmasının ve kimlik tespitinin engellenmesi amaçlanmaktadır. Bu sebeple kişisel veriler işlenirken veri kümesi içerisinde anonimleştirme işlemi ile kimlik tespiti bulma ihtimali azalacaktır. Aynı verilere sahip olan kişilerin tanımlayıcı nitelikleri çıkarılarak işlenir. Sonraki işlem ise hassas nitelikli verilerin tespitini zorlaştırmaktır. Bu işlem *k*–1 kuralı ile gerçekleştirilmektedir. *k* adet kayıt *k*–1 seviyesinde yenilenerek yeni bir tablo oluşturulmaktadır (Eyüpoğlu, 2018; KVKK, 2017; Sweeney, 2002; Vural, 2018). Tablo 22'de cinsiyet, yaş ve posta kodunun yarı tanımlayıcı, sağlık bilgisinin ise hassas nitelik olduğu örnek bir veri kümesi mevcuttur. Bu veri kümesi üzerinde *k*–anonimlik (*k*=2) tekniği uygulanmıştır. Yarı tanımlayıcı nitelikler üzerinde gizleme ve genelleme yöntemleri kullanılarak 4 adet eşdeğerlik grubu elde edilmiştir. Sonuç olarak 2-anonim grupların olduğu veri kümesi Tablo 23'te görülmektedir.

Tablo 22. Örnek Veri Kümesi

Sıra No	Cinsiyet	Yaş	Posta Kodu	Sağlık Bilgisi
1	Erkek	35	15325	Gastrit
2	Kadın	38	15340	Ülser
3	Kadın	42	15160	Lenf Kanseri
4	Kadın	49	15620	Lenf Kanseri
5	Erkek	52	15755	Diyabet
6	Erkek	58	15480	Gastrit
7	Erkek	65	15830	Karaciğer Kanseri
8	Erkek	66	15260	Karaciğer Kanseri

Tablo 23. *k*–anonimlik Örneği (*k*=2)

Sıra No	Cinsiyet	Yaş	Posta Kodu	Sağlık Bilgisi
1	*	<40	153**	Gastrit
2	*	<40	153**	Ülser
3	Kadın	[40-50]	15***	Lenf Kanseri
4	Kadın	[40-50]	15***	Lenf Kanseri
5	Erkek	[50-60]	15***	Diyabet
6	Erkek	[50-60]	15***	Gastrit
7	Erkek	>60	15***	Karaciğer Kanseri
8	Erkek	>60	15***	Karaciğer Kanseri

2.2.4.2. *l*-çeşitlilik

l–çeşitlilik, Machanavajjhala ve ark. (2007) tarafından *k*–anonimlik modelinin zayıflıklarının üstesinden gelmek için önerilen bir modeldir. Aynı nitelik değerlerinin olduğu gruplardaki hassas niteliklere odaklanarak çeşitlilik oluşturma amaçlanmaktadır (Eyüpoğlu, 2018; KVKK, 2017; Sweeney, 2002; Vural, 2018). Tablo 22'deki örnek veri kümesi üzerinde 2-anonimlik uygulanması ile oluşturulan Tablo 23'te görüldüğü üzere 2. ve 4. gruplardaki hassas nitelik değerleri aynıdır. Yani burada örneğin cinsiyet için erkek ve yaş için >60 değerlerine sahip olan bir kişiyi tanıyan saldırgan, bu kişinin karaciğer kanseri olduğunu kolay bir şekilde öğrenebilir. Sadece yarı tanımlayıcı niteliklerin değerlerine odaklanmanın yeterli olmadığı bu gibi durumlarda *l*–çeşitlilik yöntemi devreye girmektedir. Tablo 22'deki örnek veri kümesinde 4-anonimlik ve 3-çeşitlilik uygulanarak elde edilen veri kümesi Tablo 24'te gösterilmektedir. Tablodan görüldüğü üzere her bir eşdeğerlik grubu içerisinde 4 kayıt ve 3 farklı sağlık bilgisi değeri vardır. Böylece aynı bilgilere sahip olan saldırganın, kişinin karaciğer kanseri olduğunu öğrenme ihtimali azalmıştır. Sonuç olarak daha iyi bir anonimleştirme sağlanmıştır.

Tablo 24. *l*–çeşitlilik Örneği (*k*=4, *l*=3)

Sıra No	Cinsiyet	Yaş	Posta Kodu	Sağlık Bilgisi
1	*	[30-50]	15***	Gastrit
2	*	[30-50]	15***	Ülser
3	*	[30-50]	15***	Lenf Kanseri
4	*	[30-50]	15***	Lenf Kanseri
5	Erkek	[50-70]	15***	Diyabet
6	Erkek	[50-70]	15***	Gastrit
7	Erkek	[50-70]	15***	Karaciğer Kanseri
8	Erkek	[50-70]	15***	Karaciğer Kanseri

2.2.4.3. *t*-yakınlık

t–yakınlık, *l*–çeşitlilik yönteminin kişisel veri üzerinde uygulanırken yetersiz kalması durumunda uygulanmaktadır. Veri kümeleri öncelikle alt sınıflara ayrılmaktadır. Bu ayrılma veri kümesinin içerisinde birbirlerine yakın verilerin gruplanması ile sağlanmaktadır. Gruplama şekillerinde *t*–yakınlık ve *l*–çeşitlilik yöntemlerinin yanı sıra kendi içerisinde de mahremiyet sağlamak için gruplama yapılmaktadır (Eyüpoğlu, 2018; KVKK, 2017; Sweeney, 2002; Vural, 2018).

2.2.4.4. Diferansiyel gizlilik

Diferansiyel gizlilik, Dwork (2008) tarafından ortaya atılan bir yöntemdir. *k*–anonimlik, *l*–çeşitlilik ve *t*–yakınlık yöntemlerini kullanarak anonimleştirme sağlamaktadır. Veri kümeleri veri tabanı üzerinde bulunmakta ve koruma, veri tabanına yapılacak olan saldırılara karşı olarak yapılmaktadır. Bir veri tabanı içerisinde bulunan değerlere gürültü ekleme yöntemi ile anonimleştirme sağlanması amaçlanmaktadır. Değerlere gürültü ekleme diferansiyel gizlilikte en fazla kullanılan yöntemdir (Vural, 2018).

2.3. Anonim Hale Getirme Yönteminin Seçilmesi

Veriler, veri sorumluları tarafından işlenmektedir. Bir veri sorumlusu seçeceği anonimleştirme yöntemi Bölüm 2.2 içerisinde işlenen bilgiler dahilinde kendi elindedir. Veri sorumlusu, veriyi işleyen kurum ve veriyi kullanan kişiler hukuk karşısında sorumlu olmaktadır. Veri kontrolörü, veriyi işlerken verinin niteliği, çeşitliliği, fiziki ortamda bulunan yapısı, işleme sıklığı,

dağıtıklık/merkezlilik oranı, büyüklüğü, sağlanan fayda, aktarılan taraf güvenliği, işleme aracı, bozulma durumu, zararı, zarar ile ortaya çıkacak olan etkiyi, yetki kontrolünü ve saldıra karşı dayanıklılığını dikkate almalıdır (KVKK, 2017).

2.4. Bilgi Kaybını Ölçme

Veriler, işlenirken hem kanun tarafından oluşan sebepler hem de kendi verilerini koruma dahilinde işlenmektedir. Bu sebeple çeşitli anonimleştirme yöntemleri uygulandığı Bölüm 2.2'de anlatılmıştır. Veriler işenirken bu yöntemler sebebi ile ya da gönderimde gerçekleşen aksilikler dahilinde bozulmaya uğramaktadır. Yine aynı şekilde saldırı karşısında da veriler bozulmakta ve kayba uğramaktadır. Bu durumlar veri kümesi bilgi kaybı olarak tanımlanabilmektedir. Veri kümesinin kullanılabilirliği gizlilik kadar önemli olmaktadır. Farklı yöntemler ile bilgi kaybı ölçülebilmektedir. Bu yöntemlere örnek olarak; Kullback–Leibler uzaklığı, minimal bozulma, belirlenebilirlik metriği, ağırlık kesinlik cezası, bilgi teorik metrikleri ve normalize ortalama eşdeğerlik sınıf boyutu metriği gösterilebilir (KVKK, 2017; Vural, 2018).

Kullback–Leibler uzaklığı, göreli entropi ya da göreli belirsizlik olarak bilinmektedir. Uzaklık, iki farklı olasılık arasında ölçülmektedir. Ölçümün farkı ise bilgi temelli olmasıdır. Mesafe her zaman pozitif değer olmalıdır. Mesafenin sıfır olması iki veri arasında bir fark olmadığı anlamına gelmektedir. Minimal bozulma, ceza puanı sistemine göre ölçülmektedir. Genelleştirme işlemi yapılarak toplam genelleştirilme yapılan değer miktarı kadar ceza işlenir. İşlenen ceza puanları toplanarak veride anonimleştirme hesabı yapılmaktadır. Bu sayede bilgi kaybı ölçülmektedir. Bu yöntem literatürde veri fayda metriği ismi ile de anılmaktadır. Ayırt edilebilme metriği (discernibility method-DM) de minimal bozulma da olduğu gibi ceza puanı sistemi üzerinden hesaplanmaktadır. Bu yöntemde bir veri a büyüklüğünde bir gruba dahil ise bu veri a^2 ceza puanına sahiptir. Büyük veride ise bu işlem $DM(T)=\Sigma(a_i)^2$ olacak şekilde hesaplanmaktadır (Bayardo & Agrawal, 2005; Vural, 2018).

Sınıflandırma yöntemi, Iyengar (2002) tarafından önerilmiştir. Veriler hem anlamlı bilgi hem de gürültü ekleme yöntemi ile bozulmuş bilgi içermektedir. Fakat bilgi elde edilirken veriyi gürültüden kurtarmak gereklidir. Trade–off yöntemi, Fung ve ark. (2005) tarafından önerilmiştir. Mahremiyet ve fayda ikileminde denge sağlamaya odaklıdır. Aradaki denge kurulduğunda veri kaybı bulunabilmektedir. I-Loss, verilerin kategorize edilmesi ile gerçekleştirilmektedir. Bir değer genelleme mantığı ile genelleştirilir ve formül yardımı ile bulunur. $I\text{-}Loss(v_g)=|V_g|-1/|DA|$ formülü ile hesaplanan I-Loss yönteminde, V_g düğüme ait sayı, DA ise V_g'nin A tabanında değer sayısını ifade etmektedir (Vural, 2018).

2.5. Saldırı Türleri

Büyük veri kullanımında mahremiyet konusu çok büyük öneme sahiptir. Mahremiyet modellerinin belirli açıklıkları bulunmaktadır. Bu açıklıklar üzerinden yapılan saldırılar ayrıca yeni modellerin gelişmesi için de öncülük etmektedir. Kimlik ifşası/bağlantı saldırısı, homojenlik saldırısı, benzerlik saldırısı, geçmiş bilgisi saldırısı, olasılıksal çıkarım saldırısı, arka plan bilgi saldırıları, çarpıklık saldırıları, anlamsal benzerlik saldırıları, minimalite saldırılar ve de-finetti saldırılar bu saldırılara örnek olarak gösterilebilir (Eyüpoğlu, 2018; Koca & Aydın, 2017; Vural, 2018).

2.5.1. Kimlik ifşası/bağlantı saldırısı

Bu saldırı şeklinde saldırganlar; hassas verileri, yarı tanımlayıcı nitelikleri üzerinden inceleyerek ortaya çıkarmaya çalışmaktadır. Bir başka deyişle mevcut veri setinde yarı tanımlayıcı değerler üzerinden bir kişinin hassas nitelik değerinin bulunmaya çalışılması kimlik ifşası saldırısı olarak

tanımlanmaktadır. Bu saldırıyı önlemek için *k*–anonimlik yöntemi kullanılmaktadır (Eyüpoğlu, 2018).

Örnek olarak Tablo 22'deki 58 yaşındaki erkeği tanıyan bir saldırgan kişinin hassas verisi olan sağlık bilgisine ulaşabilmektedir ve gastrit hastası olduğu öğrenebilmektedir. Ancak Tablo 22'deki veri kümesinin *k*–anonimlik uygulanmış hali olan Tablo 23'te aynı saldırgan kişinin gastrit hastası olduğu sonucuna varamamaktadır. Kişinin diyabet hastalığına sahip olma ihtimali gastrit hastası olma ihtimali ile aynıdır. Sonuç olarak Tablo 23'teki veri kümesi kimlik ifşası/bağlantı saldırısına karşı dayanıklıdır.

2.5.2. Homojenlik saldırısı

Bir veri seti üzerinde *k*–anonimlik yöntemi uygulanarak veri kümesi saldırılara karşı dayanıklı hale getirilir. Ancak anonim grupların olduğu kayıtlarındaki hassas niteliklerin değerlerinin aynı olduğu durumda *k*–anonimlik etkisini kaybetmektedir. Bu tür saldılar, homojenlik saldırısı olarak adlandırılmaktadır. Homojenlik saldırısı ihtimalinin olduğu durumlarda *l*–çeşitlilik yöntemi kullanılmaktadır. Homojenlik saldırısının engellenmesi için benzer hassas nitelikleri aynı grup içerisine almamaya dikkat edilmelidir. Kayıt çoğaltma yöntemi bu saldırıya karşı kullanılabilmektedir. Kayıt çoğaltma ile homojen hassas nitelikler heterojen hale gelmekte ve saldırı sonucu tanımlanma riski azalmaktadır (Eyüpoğlu, 2018; Vural, 2018).

Tablo 23'teki 2. ve 4. eşdeğerlik gruplarındaki hassas nitelik değerleri aynıdır ve bu durum homojenlik saldırısına yol açmaktadır. *l*–çeşitlilik yönteminin (*k*=4, *l*=3) uygulandığı Tablo 24'te görüldüğü üzere bu saldırının üstesinden gelinmiştir.

2.5.3. Benzerlik saldırısı

Paylaşılan anonim veri kümelerinde hassas değerler farklı olsa da benzerlik gösterebilmektedir. *l*–çeşitlilik yöntemi uygulansa bile kişilere ait hassas bilgiler ifşa olabilmektedir. Bu saldırılara benzerlik saldırıları denilmektedir (Eyüpoğlu, 2018).

Örnek olarak Tablo 23'teki 1. eşdeğerlik grubundaki hassas nitelikler aynı olmasa da benzerdir. Tablo üzerinden saldırı yapan kötü niyetli bir kişi, tanıdığı kişinin sağlık bilgisinin yani hastalığının mide ile ilgili olduğunu öğrenebilir.

2.5.4. Geçmiş bilgisi saldırısı

Veri seti ile ilgili genel bilgiye sahip ya da veri seti içerisinde olan kişisel bilgiler dahilinde bilgi sahibi bir saldırganın, veri kümesi içerisinde bu bilgilerini kullanarak anlamlandırma yapmasına geçmiş bilgisi saldırısı denilmektedir (Eyüpoğlu, 2018; Vural, 2018).

2.5.5. Olasılıksal çıkarım saldırısı

Bir veri seti içerisinde aynı gruplandırma içerisinde olan bir hassas nitelik değeri diğer hassas nitelik değerlerine göre daha fazla bulunuyorsa saldırgan, olasılıksal olarak daha fazla olan değeri düşünmektedir. Bu duruma olasılıksal çıkarım saldırısı denilmektedir. Ayrıca bu durum saldırganın çoğunluk olan hassas nitelik için belirli yarı tanımlayıcı değerlerde daha fazla bulunduğunu öğrenmesine sebep olmaktadır (Eyüpoğlu, 2018; Vural, 2018).

2.5.6. Arka plan bilgi saldırısı

Teorik olarak Bölüm 2.5.4'te anlatılan geçmiş bilgisi saldırısına benzemektedir. Saldırgan öncelik olarak farklı kurum ve kuruluşlar tarafından yayımlanan veri setleri üzerinden elde ettiği bilgileri kullanmaktadır. Bu veri setleri sosyal medya, dergi ve gazete olabilmektedir. Veri setlerinin bağlama yöntemi ile iki farklı kaydın bilinen bilgi dahilinde değerlendirilmesi, veri mahremiyet sorununa yol açmaktadır. Saldırılar esnasında kullanılan bilgileri engellemek pek mümkün değildir. Bu sebeple genel olarak kişisel veri kuralları geçerlidir ve güvenlik önlemleri dikkate alınmalıdır (Chen ve ark., 2007; Vural, 2018).

2.5.7. Çarpıklık saldırıları

Bir veri kümesi içerisinde hassas niteliklerin istatiksel olarak değerlendirilmesi ve bu durum dikkate alınarak dağılım yapılması mahremiyet için oldukça önemlidir. Bir anlamda Bölüm 2.5.5'te anlatılan olasılıksal çıkarım saldırısına benzemektedir. Fazla sayıda bulunan hassas nitelik değerleri genel istatistiki dağılımda çarpıklık oluşturmaktadır. Değerler çarpık olduğu için saldırı karşısında zayıf olacak ve çarpıklık giderilmediği sürece koruma sağlanılamayacaktır (Vural, 2018; Xu ve ark., 2010).

2.5.8. Anlamsal benzerlik saldırıları

Yayımlanmış anonim bir veri seti, gruplar dahilinde hassas nitelik değerlerinin birbirinden farklı olması mahremiyet için yeterli olmamaktadır. Anlamsal benzerlikler yani sezgisel benzerliklerin yardımı olması sebebiyle homojenlik saldırısı ile benzemektedir. Ayrıca anlamsal benzerlik saldırısı, Bölüm 2.5.3'te anlatılan benzerlik saldırısı ile de benzemektedir. Engellenmesi için öncelikle veri seti içerisinde yer alan benzerlik durumlarının hesaplanması gerekmektedir. Sonra bu değerlerin farklı gruplara ayrılması ve aynı şekilde farklı gruplarda yer alması sağlanmaktadır. Bir başka deyişle *t*–yakınlık yöntemi kullanılmalıdır (Vural, 2018; Wang ve ark., 2014).

2.5.9. Minimalite saldırıları

Saldırıların, anonimleştirme algoritmaları veya sistemler hakkında bilgi sahibi olunduğunda gerçekleşebileceği söylenmektedir (Wong ve ark., 2007). Bu saldırı türü minimal durumlar düşünülerek kurulmuştur. Bir başka deyişle anonimleştirme en alt düzeyde kalmalı ve gerekli durum dışında anonimleştirme yapılmamalıdır (Vural, 2018; Wong ve ark., 2007).

2.5.10. De-finetti saldırıları

De-finetti saldırıları, Kifer (2009) tarafından ortaya atılan ve de-finetti teoremi temelli bir fikirdir. Bu teorem ile değiştirilebilirlik kavramı birlikte kullanılarak mahremiyetin ifşası araştırılmıştır. Arka plan bilgisine ihtiyaç duymadığı için diğer saldırı türlerinden ayrılmaktadır. Makine öğrenmesi temellidir. İlgili kayıt üzerinden hassas olmayan nitelikler dahilinde öğrenme yapılarak saldırı gerçekleştirilmektedir (Kifer, 2009; Vural, 2018).

3. LİTERATÜRDEKİ ÇALIŞMALARIN KARŞILAŞTIRILMASI

Bu bölümde son yıllarda yapılan veya temel teşkil eden çalışmalar incelenmektedir. Büyük veride mahremiyet ya da gizlilik koruması, anonimleştirme yöntemleri ve saldırı türleri konusunda yapılan literatürdeki çalışmalar Tablo 25'te özetlenmekte ve karşılaştırılmaktadır. Tabloda çalışmanın ismi, yazarları, yayımlandığı yıl ve konusu hakkında bilgilere yer

verilmektedir. Araştırma alanının önemli ve güncel olduğu incelenen çalışmaların sonuçlarından görülebilmektedir.

Tablo 25. Literatürdeki Çalışmalar

Çalışma	Yazar	Yıl	Çalışmanın Konusu
Büyük veride kişi mahremiyetinin korunması	Eyüpoğlu ve ark.	2017	Büyük veride güvenlik ve mahremiyet alanında var olan çalışmalar incelenmiştir.
Büyük veride siber güvenlik açıkları ve güvenlik yöntemleri üzerine bir araştırma	Koca ve Aydın	2017	Büyük veri güvenliğinin sağlanmasına yönelik çalışmalar araştırılmıştır.
Veri mahremiyeti: saldırılar, korunma ve yeni bir çözüm önerisi	Vural	2018	Büyük veride güvenlik ve mahremiyet alanında var olan çalışmalar incelenmiş ve çoğaltma tekniği öne sürülmüştür.
k-anonymity: a model for protecting privacy	Sweeney	2002	*k*–anonimlik modeli ileri sürülmüştür.
t-closeness: privacy beyond *k*-anonymity and *l*-diversity	Li ve ark.	2007	*t*–yakınlık modeli önerilmiştir.
Sağlık hizmetlerinde anonimlik: dağıtık yapılar için ideal bir veri paylaşım modeli	Canbay	2014	Sağlık alanında veri anonimleştirmede yeni model önerilmiştir.
Büyük veri uygulamalarında kişisel veri mahremiyeti	Akıncı	2019	Büyük veride güvenlik ve mahremiyet için sorunların tespiti ve çözüm önerileri geliştirilmesi anlatılmaktadır.
Kişisel verilerin anonimleştirilmesinin iyileştirilmesine yönelik bir model geliştirilmesi ve e-devlet alanında uygulanması	Afyonluoğlu	2019	Veriden sağlanan faydanın en aza indirildiği ve anonimleştirmenin en yüksek seviyede yapıldığı bir algoritma modeli öne sürülmüştür.
Büyük veri ve açık veri analitiği: yöntemler ve uygulamalar	Sağıroğlu	2017	Büyük veri analitiğinde güvenlik ve mahremiyetin önemi ve gelişimi bu çerçeve içerisinde yapılan çalışmalar ve uygulamalar anlatılmıştır.
Tıbbi belgeleme	Ünal	2017	Tıp alanı içerisinde belgeleme şekilleri anlatılmaktadır.
Kişisel verilerin silinmesi, yok edilmesi veya anonim hale getirilmesi rehberi	KVKK	2017	6698 sayılı kanuna göre verilerin işlenme koşulları ve anonimleştirme teknikleri anlatılmaktadır.
Büyük veride etkin gizlilik koruması için yazılım tasarımı	Eyüpoğlu	2018	Büyük veride güvenlik ve mahremiyet alanında etkin gizlilik algoritması önerilmiştir.
Mahremiyet korumalı büyük veri yayınlama için kavramsal model önerileri	Canbay ve ark.	2020	Mahremiyet koruma kullanan veri yayımlama modelleri araştırılmış ve karşılaştırılmıştır. Ayrıca yeni bir model önerilmiştir.

Büyük veri analitiği, güvenliği ve mahremiyeti	Sağıroğlu ve ark.	2016	Büyük veri analitiği, güvenliği ve mahremiyeti incelenmiştir.
An efficient big data anonymization algorithm based on chaos and perturbation techniques	Eyupoglu ve ark.	2018	Kaos ve pertürbasyon tekniklerine dayanan yeni bir büyük veri anonimleştirme algoritması önerilmiştir.
Derin ögrenme ile büyük veri kümelerinden saldırı türlerinin sınıflandırılması	Ahmetoğlu ve Daş	2019	Saldırılara karşı geliştirilen model sunulmuştur.
Privacy and security problems in healthcare 4.0	Kara ve Eyüpoğlu	2020	Sağlık 4.0'daki mahremiyet ve güvenlik sorunları ele alınmıştır. Ayrıca bu sorunların çözümüne ilişkin geliştirilen tekniklerden bahsedilmiştir.
Anonymization methods for privacy-preserving data publishing	Kara ve Eyüpoğlu	2021	Gizlilik korumalı veri yayınlama için geliştirilen anonimleştirme yöntemleri ve modelleri incelenmiştir.

4. SONUÇ

Bu çalışmada büyük veri kavramı içerisinde güvenlik ve mahremiyetin öneminden bahsedilmiştir. Büyük veri çerçevesinde verilerin gizliliğinin korunması için gerekli anonimleştirme yöntemleri incenmiş, kişisel bilgilerin ifşasına yönelik geliştirilen saldırı türleri ve bu saldırı türlerine karşı geliştirilen anonimleştirme tekniklerine değinilmiştir. Anonimleştirme teknikleri, örnek veri kümeleri üzerinde uygulanarak konunun daha iyi anlaşılması amaçlanmıştır.

Günümüzde veri kavramının ön plana çıkması ve hukuksal düzenlemelerden doğan sorumluluklar verilerin işlenmesi ve saklanmasına etki etmiş, güvenlik kavramının önemini göstermiştir. Bu yükümlülükler karşısında veri bilimciler anonimleştirme tekniklerini geliştirmiş, ayrıca verilere karşı yapılan saldırı durumlarında veri kaybı ve bozulması yaşanmaması için farklı teknikler geliştirilmeye devam etmektedir. Dijital dünyanın her geçen gün gelişmesi ve verilerin dijital ortamlara aktarılması; veri güvenliğinin ve mahremiyetinin önemini göstermeye devam etmektedir.

Yazarların Katkısı
Yazarların makaleye katkıları eşit orandadır.

Teşekkür
Makaleye değerli yorumları ile katkı sağlayan hakemlere teşekkür ederiz.

Çıkar Çatışması Beyanı
Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı
Yapılan çalışmada araştırma ve yayın etiğine uyulmuştur.

KAYNAKÇA

Afyonluoğlu, M. (2019). *Kişisel verilerin anonimleştirilmesinin iyileştirilmesine yönelik bir model geliştirilmesi ve E-devlet alanında uygulanması* [Doktora tezi]. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Ahmetoğlu, H., & Daş, R. (2019, September). Derin öğrenme ile büyük veri kümelerinden saldırı türlerinin sınıflandırılması. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)* (1-9). IEEE.

Akıncı, A. N. (2019). *Büyük veri uygulamalarında kişisel veri mahremiyeti* [Uzmanlık tezi]. T.C. Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı, Sektörler ve Kamu Yatırımları Genel Müdürlüğü, Ankara.

Bayardo, R. J., & Agrawal, R. (2005, April). Data privacy through optimal *k*-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*, 217-228, IEEE.

Canbay, P. (2014). *Sağlık hizmetlerinde anonimlik: Dağıtık yapılar için ideal bir veri paylaşım modeli* [Yüksek lisans tezi]. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Canbay, Y., Vural, Y., & Sağıroğlu, Ş. (2020). Mahremiyet korumalı büyük veri yayınlama için kavramsal model önerileri. *Politeknik Dergisi*, 23(3), 785-798.

Chen, B. C., LeFevre, K., & Ramakrishnan, R. (2007). Privacy skyline: Privacy with multidimensional adversarial knowledge. *In 2007 International Conference on Very Large Data Bases (VLDB)*, 770-781.

Dwork, C. (2008, April). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation* (pp. 1-19). *Springer*, Berlin.

Eyüpoğlu, C. (2018). *Büyük veride etkin gizlilik koruması için yazılım tasarımı* [Doktora tezi]. İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Eyüpoğlu, C., Aydın, M. A., Sertbaş, A., Zaim, A. H., & Öneş, O. (2017). Büyük veride kişi mahremiyetinin korunması. *Bilişim Teknolojileri Dergisi*, 10(2), 177-184.

Eyupoglu, C., Aydin, M. A., Zaim, A. H., & Sertbas, A. (2018). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, 20(5), 373.

Fung, B. C., Wang, K., & Yu, P. S. (2005, April). Top-down specialization for information and privacy preservation. In *21st International Conference on Data Engineering (ICDE'05)*, 205-216. IEEE.

Iyengar, V. S. (2002, July). Transforming data to satisfy privacy constraints. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 279-288.

Kara, B. C., & Eyüpoğlu, C. (2020, October). Privacy and security problems in healthcare 4.0. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1-12. IEEE.

Kara, B. C., & Eyüpoğlu, C. (2021, October). Anonymization methods for privacy-preserving data publishing. In *3rd International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2021)*. Berlin.

Kifer, D. (2009, June). Attacks on privacy and deFinetti's theorem. In *2009 ACM SIGMOD International Conference on Management of Data,* 127-138.

Koca, M., & Aydın, M. A. (2017, October). A survey of cyber security vulnerabilities and security methods on big data. In *8th International Advanced Technologies Symposium* 1-7.

KVKK. (2017). Kişisel verilerin silinmesi, yok edilmesi veya anonim hale getirilmesi rehberi. *Kişisel Verileri Koruma Kurumu*. Ankara.

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, 106-115. IEEE.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). *l*-diversity: Privacy beyond *k*-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3-es.

Sağıroğlu, Ş. (2017). Büyük veri ve açık veri analitiği: yöntemler ve uygulamalar. *Grafiker Yayınları.* Ankara.

Sağıroğlu, Ş., Sinanç Terzi, D., Terzi, R., Canbay, Y., Gündüz, S., Arslan, B., Ayaydın, A., & Gökalp, A. B. (2016). Büyük veri analitiği, güvenliği ve mahremiyeti. *Gazi Üniversitesi.* Ankara.

Sweeney, L. (2002). *k*-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557-570.

Ünal, N. (2017). Tıbbi belgeleme. *Anadolu Üniversitesi Yayınları*, Eskişehir.

Vural, Y. (2018). Veri mahremiyeti: Saldırılar, korunma ve yeni bir çözüm önerisi. *Uluslararası Bilgi Güvenliği Mühendisliği Dergisi, 4*(2), 21-34.

Wang, H., Han, J., Wang, J., & Wang, L. (2014). (l, e)-diversity--a privacy preserving model to resist semantic similarity attack. *Journal of Computers, 9*(1), 59-65.

Wong, R. C. W., Fu, A. W. C., Wang, K., & Pei, J. (2007, September). Minimality attack in privacy preserving data publishing. In *33rd International Conference on Very Large Data Bases* 543-554.

Xu, Y., Wang, K., Fu, A. W. C., & Wong, R. C. W. (2010, April). Publishing skewed sensitive microdata. In *2010 SIAM International Conference on Data Mining*, 84-93.