ρ-KAZANIM: MAHREMİYET KORUMALI FAYDA TEMELLİ VERİ YAYINLAMA MODELİ

ρ-GAIN: PRIVACY PRESERVING UTILITY-BASED DATA PUBLISHING MODEL

YILMAZ VURAL

YRD. DOÇ. DR. MURAT AYDOS
Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin
Bilgisayar Mühendisliği Anabilim Dalı için Öngördüğü
DOKTORA TEZİ olarak hazırlanmıştır.

2017

YILMAZ VURAL'ın hazırladığı "ρ-KAZANIM: MAHREMİYET KORUMALI FAYDA TEMELLİ VERİ YAYINLAMA MODELİ" adlı bu çalışma aşağıdaki jüri tarafından BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI'nda DOKTORA TEZİ olarak kabul edilmiştir.

Prof. Dr. Mehmet Önder EFE

Başkan ……

Yrd. Doç. Dr. Murat AYDOS ……

Danışman

Doç. Dr. Ahmet Burak CAN ……

Üye

Doç. Dr. Mehmet TEKEREK ……

Üye

Doç. Dr. Suat ÖZDEMİR ……

Üye

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından DOKTORA TEZİ olarak onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU

Fen Bilimleri Enstitüsü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

☐ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**
(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, tezinin arama motorlarının önbelleklerinde kalmaya devam edebilecektir.)

☒ **Tezimin/Raporumun** 12.07.2022 **tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**
(Bu sürenin sonunda uzatma için başvuruda bulunmadığım taktirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı ve ya tamamının fotokopisi alınabilir)

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını istemiyorum, ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐ **Serbest Seçenek/Yazarın Seçimi**

12 / 07 / 2017

(İmza)

Öğrencinin Adı Soyadı

Yılmaz YURAL

Sevgili Aileme;

Bilhassa şimdiden çok özlediğim, yaşadıkça özleyeceğim

Kıymetli Babacığım

AHMET VURAL

anısına...

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

11/07/2017

YILMAZ VURAL

ÖZET

ρ-KAZANIM: MAHREMİYET KORUMALI FAYDA TEMELLİ VERİ YAYINLAMA MODELİ

Yılmaz VURAL

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Yrd. Doç. Dr. Murat AYDOS

Haziran 2017, 134 sayfa

Veri mahremiyeti, veri sahiplerinin mahremiyet riskleri ile veri paylaşımının taraflara sağlayacağı fayda arasındaki en iyi dengeyi bulmaya çalışan zor bir problemdir. Mahremiyet korumalı yaklaşımlar veri mahremiyeti probleminin çözümünde yaygın olarak kullanılır. Mahremiyet korumalı yaklaşımların uygulanmasında anonimleştirme tekniklerinden yararlanılır. Anonimleştirme veri detaylarını azaltarak mahremiyeti koruyan fayda temelli dönüştürme tekniğidir. Anonimleştirilen veriler benzerliklerine göre eşlenik sınıf adı verilen gruplar içerisinde toplanır. Eşlenik sınıflar veri faydasına göre, fayda sağlayan (Utility Equivalence Class-UEC) ve aykırı (Outlier Equivalence Class-OEC) olmak üzere iki sınıfa ayrılır. Faydalı eşlenik sınıflar mahremiyet gereksinimlerini sağlayarak veri alıcılarına fayda sunan kayıtları içerir. Aykırı eşlenik sınıf, mahremiyet gereksinimlerini sağlayamadığı için tamamen baskılanan veri faydası olmayan kayıtları içerir. Bu çalışmada, eşlenik sınıf ayrımının veri faydası ve mahremiyet riskleri üzerindeki etkisi incelenmiş, aykırı eşlenik sınıf içerisinde yer alan kayıtların veri faydası açısından geri kazanımı konusu araştırılmıştır. Bu kapsamda mahremiyetten ödün vermeden eşlenik sınıf ayrımı yaparak veri faydasını arttıran fayda temelli ρ-Kazanım modeli önerilmiştir. Önerilen model k-Anonimlik, ℓ-Çeşitlilik ve t-Yakınlık modellerinin makul kombinasyonlarına uygulanarak test edilmiştir. Test sonuçlarının veri faydası açısından değerlendirilmesinde eşlenik sınıfları dikkate alan metrikler kullanılmıştır. Elde edilen bulgulara göre, ρ-Kazanım modeli, veri faydasında iyileşmeyi sağlarken, mahremiyet risk tahminlerinde olumsuz bir değişime yol açmamıştır. Veri mahremiyeti risklerini arttırmadan veri faydasını iyileştiren, fayda temelli ρ-Kazanım modelinin veri mahremiyeti probleminin çözümünde etkin bir rol oynayacağı gözlemlenmiştir.

Anahtar Kelimeler: ρ-Kazanım modeli, anonimleştirme, aykırı eşlenik sınıf, faydalı eşlenik sınıf, fayda temelli veri yayınlama, veri mahremiyeti, veri faydası.

ABSTRACT

ρ-GAIN: UTILITY-BASED PRIVACY PRESERVING DATA PUBLISHING MODEL

Yılmaz VURAL

PhD, Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Murat AYDOS

June 2017, 134 pages

Data privacy is a difficult problem that tries to find the best balance between the privacy risks of data owners and the utility of data sharing to the third parties. Anonymization is the most commonly applied technique to overcome data privacy problems. The equivalence classes, the natural outcome of anonymization process, are classified according to the data utility in two main categories: Utility and Outlier Equivalence Classes (UEC, OEC). The utility equivalence class contains records that have been suppressed by anonymization techniques for privacy concerns. Meanwhile, the outlier equivalence class contains records that have been fully suppressed by anonymization techniques resulting in no data utility. In this study, ρ-Gain model, which focus on the effect of outlier equivalence class for increasing data utility, was proposed. In the proposed model, k-Anonymity, ℓ-Diversity and t-Closeness privacy models were used together with ρ-iterations to reduce the privacy risks. The Average Equivalence Class metric was used to measure data utility. According to the findings obtained from the study, the ρ-Gain model improved the data utility, but did not cause a significant negative impact on privacy risk estimates. With the use of the proposed ρ-Gain model as an anonymization technique, we have shown that the data utility has improved while keeping the data privacy risk with no significant change.

Keywords: ρ-GAIN model, anonymization, data privacy, data utility, utility-based data publishing, outlier equivalence class, utility equivalence class.

TEŞEKKÜR

Tez çalışmamın her aşamasında değerli katkılarıyla yol gösteren ve beni her zaman çalışmaya teşvik eden tez danışman hocam, Yrd. Doç. Dr. Murat AYDOS'a, önemli yorum ve yönlendirmeleriyle katkıda bulunan değerli hocalarım Prof. Dr. Hayri SEVER ve Prof. Dr. Şeref SAĞIROĞLU'na, jüri üyelerim Prof. Dr. Mehmet Önder EFE'ye, Doç. Dr. Ahmet Burak CAN'a, Doç. Dr. Suat ÖZDEMİR'e ve Doç. Dr. Mehmet TEKEREK'e teşekkür ederim.

Tez çalışmam süresince hep yanımda olan ve beni destekleyen, kıymetli eşim Şebnem'e, biricik kızım Beril'e kardeşlerim Erdal ve Sibel'e, hayatlarını ailesine adamış kıymetli anneciğim Cennet VURAL'a ve tez savunmasının hemen öncesinde kaybettiğim, sevgili babacığım Ahmet VURAL'a minnettarım.

Çalışmalarımda bana yardımcı olan beni her zaman teşvik eden ismini sayamadığım tüm dostlarıma destekleri için ayrıca teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÇİZELGELER LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

SİMGELER VE KISALTMALAR

Simgeler

ρ	Ro
δ	Delta

Kısaltmalar

ABD	Amerika Birleşik Devletleri
AEM	Ayırt Edilebilirlik Metriği
ARX	Veri Mahremiyet Aracı
API	Uygulama Programlama Arayüzü
ARX	Açık kaynak veri anonimleştirme aracı
CAT	Cornell Anonymization Toolkit
CM	Sınıflandırma Yöntemi
CPM	Kesinlik Cezası
CPU	Merkezi İşlem Birimi
DGH	Alan Genelleştirme Hiyerarşisi
DVD	Çok Amaçlı Sayısal Disk
EC	Eşlenik Sınıflar
EM	Entropi Metrik
EMD	Toprak Taşıyıcı Mesafe
ESO	Eşlenik Sınıflar Ortalaması
GID	Grup Numarası Özniteliği
HM	Yükseklik Metriği
ID	Tanımlayıcı Öznitelik
IMDB	İnternet Film Veritabanı
JVM	Java Sanal Makinesi
KDD-CUP	Veri madenciliği yarışması
KL	Kullback-Leibler Uzaklığı
KLM	KL-Sapma Metriği
KM	Kesinlik Metriği
LM	Kayıp Metriği
MD	En Az Bozulma

NETFLİX	İnternet üzerinden hizmet veren medya sağlayıcısı
NP	Polinomsal zamanda çözülemeyen problem
NSA	Hassas Olmayan Öznitelik
OEC	Aykırı Eşlenik Sınıf
PK	Posta Kodu
PM	Duyarlılık Metriği
PPDM	Mahremiyet Korumalı Veri Madenciliği
PPDP	Mahremiyet Korumalı Veri Yayınlama
PQT	Yarı Tanımlayıcı Öznitelik Tablosu
PST	Hassas Öznitelik Tablosu
PT	Genel Tablo
QID	Yarı Tanımlayıcı Öznitelik Kümesi
QIT	Yarı Tanımlayıcı Öznitelik Tablosu
SA	Hassas Öznitelik
SDC	İstatistiksel İfşa Kontrolü
SECRETA	C++ dilinde görsel bir anonimleştirme aracıdır
SQL	Structured Query Language
ST	Hassas Öznitelik Tablosu
TA	Aykırı eşlenik sınıf içindeki toplam aykırı kayıt sayısı
TEC	Eşlenik Sınıflar Toplamı
TF	Faydalı eşlenik sınıf sayıların toplamı
TIAMAT	A Tool for Interactive Analysis of Microdata Anonymization Techniques
TM	Trade-Off Metrik
UEC	Faydalı Eşlenik Sınıf
UTD	Java dilinde veri kümelerinin anonimleştirilmesi için kullanılan yazılımıdır.
VGH	Değer Genelleştirme Hiyerarşileri
XML	Genişletilebilir İşaretleme Dili
YM	Yükseklik Metriği

1 GENEL BİLGİLER

Bu bölümde, veri mahremiyeti ve tezin yapısı hakkında genel bilgilere yer verilmiştir. Veri mahremiyeti probleminden bahsedilerek, tezin amaç ve hedefleri ile kapsam ve yapısı hakkında genel bilgiler verilmiştir.

1.1 Mevcut Problem

Veri mahremiyeti, veri sahiplerinin mahremiyetinin korunması ile veri paylaşımının taraflara sağlayacağı fayda arasındaki en iyi dengeyi bulmaya çalışan zor bir problemdir [1]. Elektronik bilgi toplumu olma yönünde hızla ilerlerken, sağlık, nüfus, finans, eğitim, yerel yönetimler, mülkiyet ve adli konularda hizmet veren elektronik uygulamaların kullanımı hızla yaygınlaşmaktadır [2]. Bu uygulamalar aracılığıyla toplanan veriler her geçen gün artmakta, işlenmekte, paylaşılarak hızla yayılmaktadır [3].

Paylaşılan veriler içerisinde demografik veriler, sağlık verileri, adli bilgiler, ticari bilgiler, tweetler, e-mailler, fotoğraflar, videolar ve konum bilgileri gibi kişisel ve hassas bilgilerde yer almaktadır [4]. Yasal zorunluluklar, yeni ürünlerin geliştirilmesi, mevcut hizmet kalitesinin arttırılması, bilimsel araştırmaların yapılması ve kamuoyunun bilgilendirilmesi amacıyla toplanan veriler fayda amaçlı paylaşılır. Verilerin paylaşılması araştırmacılar ve kurumlara önemli fırsatlar sunarken beraberinde mahremiyetle ilgili problemleri getirir [5].

Mahremiyet problemlerine bağlı yaşanan saldırılar bilgi sistemleri yerine doğrudan veri sahiplerini etkiler [6-8]. Kimlik bilgileri, sağlık bilgileri, mahkeme bilgileri, borç veya alacak bilgileri gibi mahrem bilgilerin kişilerin izni olmadan bilgisi dışında kullanılmasıyla mahremiyet ihlalleri yaşanır. Yaşanan bu ihlaller sonucunda veri sahipleri toplum önünde itibarsızlaştırmadan adli olaylara kadar gidebilecek birçok istenmeyen olaya maruz kalabilir. Veri paylaşımında yeterli düzeyde veya hiçbir mahremiyet önleminin alınmamasına bağlı olarak yaşanan bu olayların önüne geçmek amacıyla mahremiyet korumalı yaklaşımların kullanılması gerekir [9].

Mahremiyet korumalı yaklaşımlar mahremiyet korumasının yanında veri faydasını da dikkate alır. Mahremiyeti korunan veriler fayda sağlamak amacıyla, veri alıcıları ile paylaşılır. Fayda amaçlı paylaşılan verilerde mahremiyet riskleri ile veri faydası gereksinimleri mahremiyet korumalı yaklaşımlar tarafından sağlanır. Mahremiyet korumalı yaklaşımlar veri faydası ile mahremiyet riskleri arasındaki dengeyi gözeterek verilerin uygun yöntemlerle ilgililerine paylaşılmasını sağlarlar.

Veri faydasını iyileştirmek için gereğinden fazla alınacak önlemler mahremiyeti olumsuz yönde etkilerken gereğinden fazla mahremiyet koruyucu önlemlerde veri faydasını olumsuz etkiler. Veri faydası ağırlıklı çalışmalarda dikkat edilmesi gereken husus mahremiyet risklerinin olumsuz bir değişim göstermemesidir. Bu çalışmada mahremiyet korumalı veri faydası ağırlıklı bir model geliştirilmiş ve test edilerek etkinliği gösterilmiştir.

1.2 Tezin Amaç ve Hedefleri

Mahremiyet korumalı veri paylaşımında mahremiyetten ödün vermeden veri faydasının artırılarak en yüksek düzeyde tutulabilmesi bu tez çalışmasının temel motivasyonudur. Bu doğrultuda tez çalışmasının amaç ve hedefleri aşağıda verilmiştir.

- Mahremiyet korumalı veri yayınlanmasında verinin toplanmasından paylaşılmasına kadar geçen mahremiyet koruma sürecinde yer alan tarafların ve sorumluluklarının ortaya konulması.
- Veri faydası ile mahremiyet riskleri arasındaki ilişkinin ortaya konulması.
- Mahremiyet koruma yöntemlerinin ve modellerinin incelenmesi.
- Anonimleştirme temelli mekanizmaların incelenmesi ve uygulanması.
- Veri kaybı metriklerinin incelenmesi.
- Mahremiyet riskleri ve metriklerinin incelenmesi.
- Eşlenik sınıfların veri faydası üzerindeki etkilerinin incelenmesi.
- Aykırı eşlenik sınıf kullanımının mahremiyet risklerine olan etkisinin araştırılması.
- Veri faydası ve mahremiyet dengesini gözeterek fayda arttırıcı veri yayınlama modelinin önerilmesi
- Önerilen modelin gerçekleyerek test edilmesi
- Önerilen modelin mevcut çözümlerle fayda ve risk açısından karşılaştırılarak etkinliğinin gösterilmedi.

1.3 Tezin Kapsamı

Tez kapsamında mahremiyet korumalı yaklaşımlar incelenmiştir. Mahremiyet korumalı yaklaşımların gereksinimlerini yerine getirmede kullanılan yaygın mahremiyet modelleri araştırılmıştır. Seçilen mahremiyet modellerinin uygulanmasında veri modeli olarak mikro veriler seçilmiştir. Mahremiyet tehditleri incelenerek risk metrikleri gözden geçirilmiş ve anonimleştirilmiş verilerdeki tehditler fayda bakış açısıyla incelenmiştir. Mahremiyet

korumasıyla meydana gelen veri kaybının ölçülmesi amacıyla veri kaybı metrikleri incelenmiştir. Bu tez çalışmasında istatistiksel ve kriptografik yöntemler veri faydasını olumsuz etkilediği için kapsam dışında bırakılmış olup fayda temelli anonimleştirme yöntemleri kullanılmıştır.

1.4 Tezin Yapısı

Çalışmanın yapısı aşağıda maddeler halinde verilmiştir.

- Giriş bölümünde; veri mahremiyeti konusu ana hatlarıyla özetlenmiş, çalışmanın amacı, hedefleri, kapsamı ve tezin yapısı hakkında bilgiler verilmiştir.
- İkinci bölümde; veri mahremiyeti konusunda literatüre yerleşen kavramlar, modeller, riskler, metrikler ve tehditlerden bahsedilmiştir.
- Üçüncü bölümde; çalışmanın materyal ve yöntemi tanıtılmıştır.
- Dördüncü bölümde; mahremiyet korumalı fayda temelli veri yayınlama modelinin gerçekleştirimi yapılarak test edilmiştir. Testlere göre elde edilen sonuçlar mevcut modellerle karşılaştırılmış ve önerilen modelin veri faydası etkinliği gösterilmiştir.
- Beşinci bölümde, çalışmanın sonuçları ve gelecek çalışmalar hakkında bilgi verilerek öneriler sunulmuştur.

2 VERİ MAHREMİYETİ

Mahremiyet kavramı Warren ve Brandeis [10] tarafından Mahremiyet Hakkı başlığıyla yayımlanan makalelerinde ilk defa ele alınmıştır. Bu çalışmada, mahremiyet hem yalnız bırakılma hakkı hem de her bireyin dokunulmaz bir kişiliğe sahip olma hakkı olarak tanımlanmıştır. Mahremiyete ihlâlleri veya endişeleri, kişisel yaşantıyı ve bireyselliği zayıflatarak toplumu birbirinden ayırt edilemeyen anonim bireyler haline getirir [11]. Mahremiyet ihlalleri kişilerin mahrem bilgilerinin ifşa edilmesiyle meydana gelir. Günümüzde, mahremiyet ihlalleri çoğunlukla korunmasız veri paylaşımları sonucunda meydana gelir. Paylaşılan verilerin mahremiyetinin korunması veri sahiplerinin mahremiyetini doğrudan etkileyen önemli bir konudur.

Veri mahremiyetinin korunamadığı durumlarda veri sahibinin mahremiyetini ihlal eden ihlaller yaşanır. Çevrimiçi yayıncılık ve DVD satış sitesi Netflix'in kullanıcıların geçmiş oylamalarına dayanan film öneri sistemini geliştirmek için 2006'da başlattığı ödüllü yarışma bu kapsamda verilecek iyi örneklerden biridir. Netflix 500,000 kadar abonesinin film derecelendirmeleriyle ilgili yaklaşık 100 milyon kaydı içeren veri kümesini bu yarışma için yayınlamıştır. Aboneleri tanımlayan kişisel bilgiler (ad, soyad, IP adresi vb.) yayınlanan kayıtlardan çıkarılmıştır. Aboneleri birbirinden ayırt etmek amacıyla sayısal numaralar kayıtlara eklenerek yarışma için yayınlanmıştır. Ancak, 2007'de Austin Üniversitesi'nden iki araştırmacı, yayınlanan veri kümelerini İnternet Film Veritabanı (IMDB) üzerindeki film derecelendirmeleriyle eşleştirerek abonelerin kimliklerinin yeniden tanımlanabileceğini göstermiştir [12]. Kişileri doğrudan tanımlayan alanların mahremiyeti koruyabileceği yanılgısını gösteren başka örneklerde yaşanmıştır [13-16].

Mahremiyet ihlalleri çoğunlukla kimlikleri dolaylı yönden tanımlayan bilgilerin bir araya gelmesi veya yayınlanan verilerin farklı verilerle eşleştirilmesiyle meydana gelmektedir. 1990 yılında ABD'de sayım uygulamasıyla toplanan cinsiyet, posta kutusu ve doğum tarihi gibi doğrudan tanımlayıcı olmayan bilgilerin kullanılarak ABD nüfusunun %87'sinin kimliklerinin tespit edilebileceği Sweeney tarafından raporlanmıştır [17, 18].

Veri mahremiyeti gereksinimlerinin sağlanabilmesi amacıyla mahremiyet modelleri geliştirilmiştir. Mahremiyet gereksinimlerini sağlayacak modeller anonimleştirme ve kriptografi temelli algoritmalara ihtiyaç duyar [19]. Kriptografi temelli mahremiyet koruyucu algoritmalar, şifrelenmiş verinin depolanmasını, paylaşımını ve analizini mümkün kılar [20]. Kriptografik algoritmaların kullanımında anlamı (semantiği) gizlenmiş veriler

üzerinde işlem yapmanın kısıtları ve maliyetleri vardır [21]. Bu kısıtlar ile maliyet etkinlik dikkate alındığında özellikle halka açık verilerin yayınlanmasında kriptografik algoritmaların kullanılması tercih edilen bir yöntem değildir. Bu tez çalışmasında verilerin yayınlanmasında kriptografik çözümlere göre daha fazla probleme uygulanabilen fayda temelli anonimleştirme yöntemleri üzerinde çalışılmıştır.

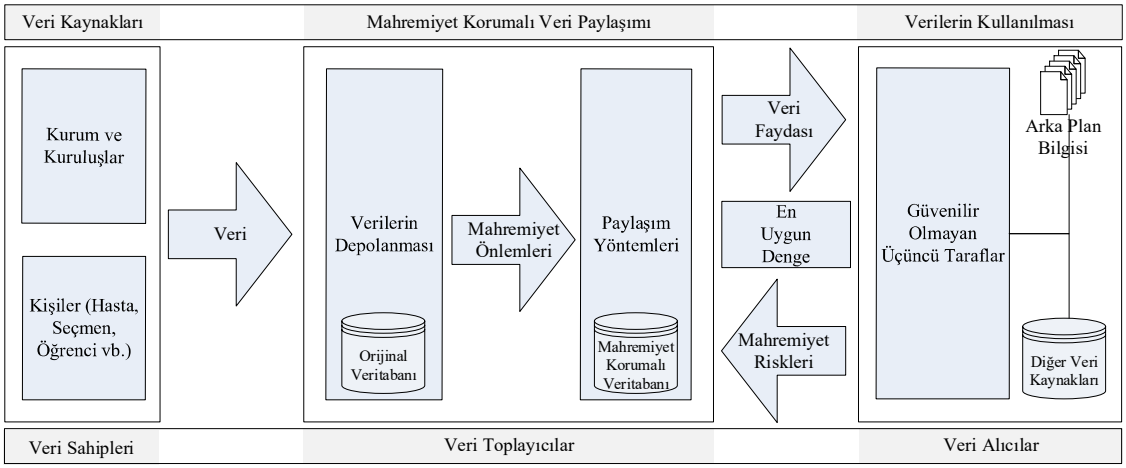
Mahremiyet korumalı yaklaşımlarda anonimleştirme teknikleri yaygın olarak kullanılır. Anonimleştirme, verinin tipi ve biçimi korunarak kimlik bilgilerinden arındırılma ve detayların azaltılması amacıyla yapılan mahremiyet koruyucu işlemlerdir [22]. Mahremiyet koruma modellerinin omurgasını oluşturan anonimleştirme mekanizmaları, genelleştirme ve bastırma [23], anatomi [24], permütasyon [25], pertürbasyon [26] olmak üzere veri üzerinde yapılan işlemlere göre sınıflandırılır. Genelleştirme ve bastırma, öznitelikler üzerinde yapılacak dönüşümlerle, anatomi ve permütasyon, öznitelikler arasındaki ilişkileri düzenleyerek pertürbasyon ise, orijinal verilere anlamlı gürültü katarak mahremiyetin korunmasını sağlar.

Mahremiyet koruma önlemleri sonucunda bilgilerde kayıplar meydana gelir. Mahremiyet korumalı yaklaşımlar bilgi kayıplarının ve mahremiyet risklerinin ölçülmesi amacıyla metriklere ihtiyaç duyar [27]. Koruyucu önlemlere bağlı olarak verilerde meydana gelen kayıpların bilinmesi veriden sağlanacak faydanın değerlendirilmesi açısından önemlidir. Ayrıca, anonimleştirme öncesi ve sonrası mahremiyet riskleri ile veri faydasının ölçülmesi mahremiyet modellerinin başarısını belirler [28].

Takip eden alt bölümlerde, mahremiyet koruma süreci, veri modelleri, mahremiyet probleminin tanımı, veri mahremiyeti konusunda alanyazına yerleşen kavramlar, modeller, riskler, metrikler ve tehditlerden bahsedilmiştir.

2.1 Mahremiyet Koruma Süreci

Veri mahremiyeti probleminin tarafları ile sorumluluklarının anlaşılabilmesi için mahremiyet koruma sürecinin bilinmesi önemlidir. Mahremiyet koruma sürecinde veri sahipleri, veri toplayıcı (sağlayıcı) ve veri alıcılar olmak üzere üç önemli taraf vardır. Verilerin toplanması ve paylaşılmasında mahremiyet koruma sürecini gösteren ve bu çalışma kapsamında çizilen yeni bir çizim Şekil 2-1'de verilmiştir. Şekilde mahremiyet koruma sürecindeki taraflar ile bu taraflar arasındaki ilişki ve sorumluluklar gösterilmiştir.



Şekil 2-1 Mahremiyet koruma süreci

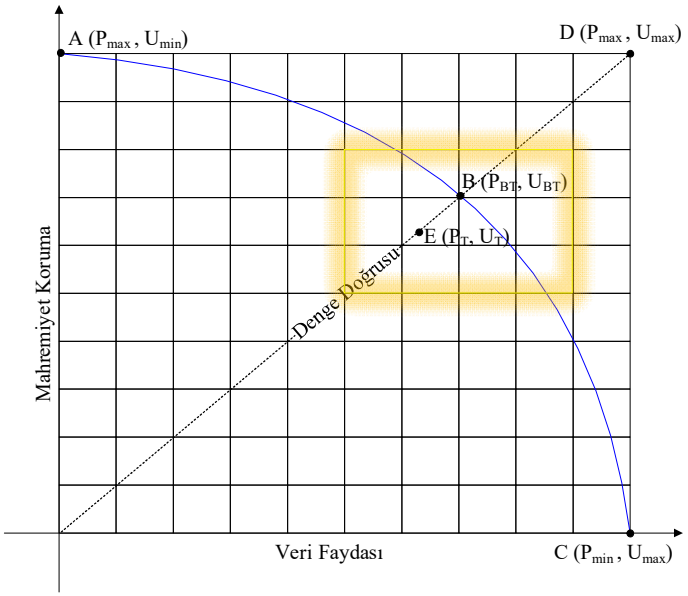
Veri sahipleri, paylaşılan veriler içerisinde kimlik ve hassas bilgileri yer alan mahremiyetleri korunması gereken kişi, kurum ve kuruluşlardır. Veri sahipleri hizmet aldıkları uygulamalar, doldurdukları formlar, katıldıkları anketler, yasal olarak yapmak zorunda oldukları bildirimler veya diğer nedenlerle güvenilir olduklarını varsaydıkları veri toplayıcısına verilerini iletir. Veri sahipleri ile veri toplayıcılar arasında yasal veya teknik önlemlerle güven ilişkisi sağlanmak zorundadır.

Veri toplayıcılar, veri sahiplerinin mahremiyetini koruyarak verilerin güvenli olarak ilgilileriyle veya halka açık olarak paylaşılmasını sağlayan kişi, kurum ve kuruluşlardır. Güvenilir olduğu varsayılan veri toplayıcılar farklı yöntemlerle toplamış oldukları verileri depolamakta, kullanmakta ve anonimleştirerek paylaşmaktadır. Veri yayıncı veya veri sağlayıcı olarak da adlandırılan veri toplayıcılar mahremiyet riskleri ile veri faydası arasındaki dengenin sağlanmasından yükümlüdür. Dengeyi sağlayabilmek için veri toplayıcılar mahremiyet korumalı yaklaşımları kullanarak veri alıcılarının fayda taleplerini mahremiyet korumalı paylaşım yöntemlerini kullanarak yerine getirir.

Veri alıcılar, paylaşılan veriler üzerinde analizler veya işlemler yaparak verilerden niyetleri veya ihtiyaçları doğrultusunda fayda sağlamaya çalışan güvenilir olmadığı varsayılan üçüncü taraflardır. Veri alıcısı kılığına girmiş saldırganlar, meraklı veya görevini kötüye kullanan gerçek ve tüzel kişilerin varlığından dolayı veri alıcılarının güvenilir olmadıkları varsayılır. Kötü niyetli veri alıcılar veya saldırganlar farklı kaynaklardan elde ettiği veya daha önceden sahip olduğu verileri yayınlanan verilerle eşleştirerek mahremiyet ihlallerinin yaşanmasına neden olur.

2.2 Mahremiyet Problemi

Veri mahremiyeti problemi, İstatistiksel İfşa Kontrolü (Statistical Disclosure Control-SDC), Mahremiyet Korumalı Veri Madenciliği (Privacy Preserving Data Mining-PPDM) ve Mahremiyet Korumalı Veri Yayınlama (Privacy Preserving Data Publishing-PPDP) başlıkları altında birbirleriyle yakından ilişkili üç grupta farklı topluluklar tarafından çalışılır. SDC, veri mahremiyeti probleminin çözümünde istatistiksel yöntemleri kullanır [29]. PPDM, veri sahiplerinin kimliklerinin veya hassas bilgilerinin ifşa edilmesini engelleyerek veriler üzerinde birden fazla araştırmacının birlikte çalışmasını mümkün kılan veri madencilik yöntemidir [30-32]. Mahremiyet korumalı veri paylaşımında, kolay ve ekonomik bir çözüm olduğu için yaygın bir yöntem olarak kullanılan PPDP, veri madenciliği ve diğer analizlerin veri alıcıları tarafından yapılabilmesini mümkün kılar [33-36]. Mahremiyet ve fayda arasındaki ilişki Şekil 2-2'de gösterilmiştir.



Şekil 2-2 Mahremiyet–Fayda eğrisi

Şekil 2-2'deki A (P_{max}, U_{min}) ve C (P_{min}, U_{max}) noktaları fayda ve koruma eğrisi üzerinde en yüksek ve en düşük değerlerin yer aldığı başlangıç noktalarıdır. D (P_{max}, U_{max}) ve E (P_T, U_T), noktaları denge(trade-off) doğrusu üzerinde yer alır. B (P_{BT}, U_{BT}) noktası ise eğri ve denge(trade-off) doğrusu kesişiminde yer alır. Bu noktaların incelenmesi koruma ve fayda arasındaki ilişkinin anlaşılması açısından önemlidir. A noktasında yüksek seviyede koruma olduğundan (şifreleme, vb.) verilerden fayda sağlanamaz.

C noktasında mahremiyet önlemleri alınmadığından veri faydası en üst düzeyde ancak veriler saldırılara karşı korunmasız durumdadır (anonimleştirme öncesi durum). E ve B noktalarını içerisine alan sarı renkle gösterilen ve veri toplayıcı tarafından belirlenen çözüm kümesine denk gelen tüm noktalar aday çözüm noktaları olarak seçilir. Çözüm kümesi üzerinde yer alan E noktasında mahremiyet ile veri faydası arasında denge sağlanmış, B noktasında ise en iyi denge yakalanmıştır. En iyi dengenin sağlandığı B noktası en uygun dengenin sağlandığı nokta olarak adlandırılır. D noktası ise mahremiyet ile veri faydasının en yüksek değerlerini alan, gerçekte hiç bir zaman ulaşılamayacak ideal çözüm veya referans noktası olarak adlandırılır.

2.3 Mikro Veri Modeli

Mahremiyet korumalı veri paylaşımında mikro veri modellerini kullanan mikro veri dosyaları yaygın olarak kullanılır. Bir mikro veri dosyası, muhatapları (bireyler) hakkında bir dizi kayıt içeren tablodan oluşur [37]. Her muhatap için verilen bilgiler öznitelikler veya değişkenler ile ifade edilir. Mikro veri tablosu, her bir satırın muhatabına atıf yapıldığı m satır ve her bir satırı oluşturan alanların gösterildiği n kolonla boyutunda bir matristir. Satır verinin muhatabıyla ilgili bir kaydı (R), sütun ise o verinin muhatabına ait bir özniteliği (A) gösterir. Örnek mikro veri tablosu Çizelge 2-1'de verilmiştir.

Çizelge 2-1 Mikro veri modeli

T	A_1	A_2	A_3	…	A_n
R_1	Veri $_{1,1}$	Veri $_{1,2}$	Veri $_{1,3}$	…	Veri$_{1,n}$
R_2	Veri $_{2,1}$	Veri $_{2,2}$	Veri $_{2,3}$	…	Veri$_{2,n}$
R_3	Veri $_{3,1}$	Veri $_{2,2}$	Veri $_{2,3}$	…	Veri$_{3,n}$
...	…	…	…	…	…
R_m	Veri $_{m,1}$	Veri $_{2,2}$	Veri $_{2,3}$	…	Veri $_{m,n}$

Mikro veri tablolarında, sayısal, kategorik, zaman serileri, sunucu günlükleri, sorgu günlükleri gibi farklı veri tiplerine sahip veriler bulunmaktadır.

Çizelge 2-1'de m tane muhatabın (kayıtlar) n tane özniteliğini içeren mikro veri tablosu örneği verilmiştir. Mikro veri tablosundaki (T), öznitelikler (A), her bir kayıtta (R) yer alan

muhatapları hakkında verdikleri bilgilere göre, kimlik tanımlayıcı (ID), yarı tanımlayıcı (QID) ve hassas (SA) olmak üzere 3 grupta sınıflandırılır ve T (ID, QID, SA) biçiminde ifade edilir [33, 36].

Açık tanımlayıcılar, ilgili kaydın kimliğini açık bir şekilde doğrudan tanımlamak için kullanılabilen özniteliklerdir. Pasaport numarası, TC kimlik numarası, telefon numarası tipik tanımlayıcı öznitelik örnekleridir. Yarı tanımlayıcılar, tek başına kullanıldıklarında kimlik tanımlayamayan ancak bir araya geldiklerinde kimliklerin tanımlanabilmesini sağlayan özniteliklerdir. Posta kodu, doğum tarihi ve cinsiyet iyi bilinen yarı tanımlayıcı öznitelik örnekleridir. Hassas öznitelikler, ilgili kaydın sahibi hakkında mahrem bilgiler içeren veri faydası yüksek özniteliklerdir. Maaş, din ve hastalık bilgisi bilinen hassas öznitelik örneklerindendir. Açık tanımlayıcılar dışındaki mikro veri tablosunda yer alan her bir özniteliğin yarı tanımlayıcı olabilmesi mahremiyet problemini zorlaştırır.

Mahremiyet koruma yaklaşımları öznitelik sınıfları üzerinde farklı işlemler yapar. ID öznitelikler, ilgili kaydın kimliğini gizlemek amacıyla yayınlanan verilerden kaldırılır veya şifrelenir. İlgili kaydın muhataplarının kimliklerinin dışarıdan sağlanan bilgilerle yeniden tanımlanmasını önlemek için QID öznitelikler üzerinde mahremiyet koruyucu işlemler yapılır. Genellikle yayınlanan verilerin faydasını korumak için SA öznitelikler üzerinde mahremiyet koruyucu işlemler yapılmaz. Yarı tanımlayıcılar, dış kaynaklardan gelen bilgilerle eşleştirildiklerinde kişileri tanımlayabilen özniteliklerdir. Yayınlanacak mikro veri tablosu $T(A_1, A_2, ..., A_n)$ olsun, $d \leq n$ olmak üzere $QID_T = \{A_1, A_2, \ldots, A_d\} \subseteq \{A_1, A_2, \ldots, A_n\}$, veri sahiplerinin kimliğini ortaya çıkarmak için dış bilgi ile eşleştirilen en az sayıdaki öznitelik grubudur [38].

2.4 Mahremiyet Koruma Modelleri

Geleneksel yöntemler mahremiyet koruması amacıyla paylaşılacak verideki ID özniteliklerini şifreleyerek veya kaldırarak verileri kimliksiz olarak yayınlar [39]. Kimliksiz verilerin yarı tanımlayıcı öznitelikleriyle, dışarıdan sağlanan bilgilerin eşleştirilebilmesi geleneksel korumayı atlatarak mahremiyet ihlallerinin oluşmasına sebep olur [40, 41]. Eşleştirme temelli yapılan saldırıları önleyebilmek amacıyla, yarı tanımlayıcı öznitelikler üzerinde mahremiyet koruyucu işlemler yapılır. Mahremiyet gereksinimlerinin sağlanabilmesi amacıyla mahremiyet modelleri kullanılır.

Eşleştirme saldırılarına örnek olarak, Çizelge 2-2'de hastaların sağlık bilgileri verilmiştir. Çizelge 2-3Şekil 2-3'de ise kamuya açık haldeki Çayyolu semtindeki öğretmenlerin listesini içeren kimliksizleştirilmiş tablo verilmiştir.

Çizelge 2-2 Kimliksizleştirilmiş sağlık verileri

ID		**QID**			**SA**
*	*	02.09.1970	E	06152	Hepatit
*	*	20.09.1970	K	06143	Kardiyomiyopati
*	*	12.09.1970	K	06148	Egzema
*	*	05.09.1970	E	06155	Zatürre
*	*	01.08.1960	K	06154	Felç
*	*	02.08.1960	K	06153	Felç
*	*	10.08.1960	E	06140	Felç
*	*	20.08.1960	E	06141	Felç
*	*	07.08.1970	K	06141	Yüksek Kolesterol
*	*	05.08.1970	K	06142	Eritema
*	*	09.07.1958	E	06232	Diyabet
*	*	25.08.1970	E	06153	Yüksek Kolesterol
*	*	02.09.1960	E	06147	Hepatit
*	*	05.09.1960	E	06145	Grip
*	*	30.09.1960	K	06159	Kardiyomiyopati

Çizelge 2-3 Öğretmen listesi

İsim	**Adres**	**Şehir**	**PK**	**Doğum**	**Cinsiyet**	**Ders**	**Okul**
Aras Kutlu	Park Cad.	ANKARA	06232	09.07.1958	Erkek	Fizik	Lise
Beril Vural	İncek Bulvarı	ANKARA	08230	06.06.1970	Kadın	Biyoloji	Lise
----	----	----	----	----	----	----	----

Verilen örnekte, doğum gününü, cinsiyeti ve posta kodunu simgeleyen QID özniteliği, yayınlanan sağlık verilerinin öğretmen listesiyle bağlanması için kullanıldığında kişilerin yeniden tanımlanmasına ve hastalıklarının açığa çıkmasına sebep olur. Bu örnekte, kimliksizleştirilmiş medikal veri içerisinde yer alan 09.07.1958'de doğmuş ve 06232 bölgesinde yaşayan erkek diyabet hastasının, Park Cad. Çayyolu adresinde oturan lise fizik öğretmeni Aras Kutlu olduğu açığa çıkmış olur.

2.4.1 k-Anonimlik

Veri bağlama (eşleştirme) saldırıları sonucunda kimliklerin yeniden tanımlanmasının engellenebilmesi amacıyla Sweeney ve arkadaşları [42, 43] tarafından önerilmiştir. k-Anonimlik modeli yayınlanan veri kümesi içinde yer alan bir kaydın en az k-1 tane kayıttan yarı tanımlayıcı öznitelikleri üzerinden ayırt edilemeyeceğini garanti eder. k-Anonimlik modeliyle QID özniteliğinin kimlik tanımlama özelliği zayıflatılır. Bir tablonun k-Anonimlik gereksinimlerini sağlayabilmesi için yayınlanan tablodaki tüm kayıtların k-Anonim olması gerekir. k-Anonimlik modeliyle yarı tanımlayıcı öznitelikler üzerinden yapılacak eşleştirme saldırılarına bağlı kimliklerin yeniden tanımlanma ihtimali 1/k'ya düşer. İlk bakışta basit bir problem olarak görünmesine karşılık optimum k-anonimliği sağlamanın NP-Zor bir problem olduğu ispatlanmış [44] ve yaklaşık çözümler üretilmeye çalışılmıştır. Yayınlanacak tablo T, tablo içerisindeki kayıtlar $\{R_1, R_2, \ldots, R_m\}$, yarı tanımlayıcı öznitelikleri $\{QID_1, QID_2, .., QID_d\}$ olsun. R [QID] R kaydı için QID özniteliğini gösterdiğinde, eğer T tablosu k-anonim ise, $R\,[QID] = R_1\,[QID_1] = R_2\,[QID_1] = \ldots = R_{k-1}\,[QID_1]$ eşitliğini sağlayan en az $R_1, R_2, \ldots, R_{k-1}$ kayıt vardır.

Yayınlanmış bir tabloda QID değeri birebirinden ayırt edilemeyen kayıtları içeren gruplar eşlenik sınıflar (Equivalence Class-EC) olarak adlandırılır [45]. T tablosunun yayınlanmış hali ise T* ile gösterilir. Verileri k-anonim hale getirmek için, ilk adım mikro veri tablosundaki yarı tanımlayıcıların belirlenmesidir. Yarı tanımlayıcı özniteliklerinin doğru olarak seçilmesinde yarı tanımlayıcıların harici kaynaklardan elde edilecek bilgi olup olmamasına bakılır [46]. Ancak harici kaynaklardaki bilgilerin doğru tahmin edilmesi zordur. Bunun yerine hassas öznitelikler dışında kalan diğer öznitelikler yarı tanımlayıcı olarak seçilir. Çizelge 2-4'de 3-Anonim şartını sağlayan tablo örnek olarak verilmiştir.

Çizelge 2-4 3-Anonim tablo

T*	QID			SA
	PK	D.Tarihi	Cinsiyet	Hastalık
EC_1	120**	1967	E	Hepatit
	120**	1967	E	Kardiyomiyopati
	120**	1967	E	Egzema
EC_2	120**	1970	K	Zatürre
	120**	1970	K	Felç
	120**	1970	K	Felç
EC_3	118**	1964	E	Felç
	118**	1964	E	Felç
	118**	1964	E	Yüksek Kolesterol
	118**	1964	E	Eritema

Eşlenik sınıflar içerisinde yer alan her bir kaydın, QID (PK, DTarihi, Cinsiyet) değeri en az diğer 2 kayıtla eşittir. QID özniteliklerinin eşleştirme saldırılarına açık olması nedeniyle, T tablosu QID öznitelikleri korumalı (sanitization) olarak yayınlanmıştır.

k-Anonimlik modelinde mahremiyet koruması için tek bir tablo üzerinde işlem yapılacağı varsayılmıştır. Bu varsayım altında birden fazla tabloya k-Anonimlik modeli uygulandığında yüksek bilgi kaybı yaşanarak koruma işlemi başarısızlıkla sonuçlanabilir. Bu durumu sınırlamak için Nergiz ve arkadaşları MultiR k-anonimlik kavramını önermişlerdir [47]. Bu model, ilişkisel bir veritabanı içerisinde pid tanımlayıcı özniteliğiyle kişisel bilgileri içeren PT tablosu ve bu tabloyla ilişkili, T_1, T_2, ... T_n tablolarının olduğunu, ilgili tüm tabloların birleşmesiyle PT $\bowtie$ $T_1$$\bowtie$ T_2, …, $\bowtie$ T_n oluşan birleştirilmiş tablo üzerinde k-anonimlik yönteminin çoklu tablolara uygulanmasını sağlar.

K-anonimlik yaklaşımında yayınlanan veri kümesindeki her bir veri sahibi için yalnızca bir kayıt olduğu varsayılmıştır. Wang ve Fung, veri kümesinde aynı veri sahibinin birden fazla kayda sahip olduğu durumlar için (X, Y)-Anonimlik kavramını önermişlerdir [48]. Bu model, X üzerindeki her değerin Y üzerinde en azından k farklı değerle bağlantılı olmasını sağlar. Yeniden kimliklendirme saldırılarına karşı farklı modeller de önerilmiştir [49, 50]

2.4.2 ℓ-Çeşitlilik

k-Anonimlik, kimlik ifşasına karşı koruma sağlarken, yarı tanımlayıcı öznitelik kümesinde yer almayan hassas özniteliklerin ifşasına karşı koruma sağlayamaz. Machaanavajjhala ve

ark. [51] k-anonimlik modelinin bu sorununu vurgulayarak hassas öznitelikleri koruyan ℓ-Çeşitlilik yöntemini önermişlerdir. Yayınlanacak tablo T*, eşlenik sınıflar {EC_1, EC_2, .., EC_j} ve tablodaki hassas öznitelik {SA}, $0 < i \leq j$ olmak üzere, EC_i [SA] ise i. EC içerisinde yer alan hassas değerlerin sayısını göstersin. Eğer T* tablosu ℓ-Çeşitli ise, T* tablosunda EC_i [SA] ≥ ℓ eşitliğini sağlayan {EC_1, EC_2, . . . , EC_j } eşlenik sınıflar vardır. Yayınlanan tablodaki her bir eşlenik sınıf içerisinde en az ℓ sayıda hassas öznitelik varsa yayınlanan tablo ℓ-çeşitliği sağlar. ℓ-Çeşitlilik modelinin örnekleri aşağıda maddeler halinde verilmiştir.

- *Farklı ℓ-Çeşitlilik*, her bir eşlenik sınıf içerisinde birbirinden farklı en az ℓ sayıda hassas değer vardır. Farklı ℓ-çeşitlilik, hassas özniteliklerin olasılıksal çıkarım oranını 1/ ℓ 'ye eşit veya daha düşük bir değere getirir.
- *Entropi ℓ-Çeşitlilik*, eşlenik sınıflardaki her bir hassas öznitelik entropisinin belirlenen bir alt sınırın üzerinde kalmasını sağlar. Eşlenik sınıf EC'nin entropisi Eşitlik 2-1'de gösterilmiştir.

$$Entropy\ (EC) = -\sum_{s \in Dom(s)} f(EC, s)\ x \log \mathrm{f}\,(\mathrm{EC}, \mathrm{s}) \qquad 2\text{-}1$$

  Eşitlik 2-1'de f (EC, s) EC eşlenik sınıfındaki kayıtlarda yer alan hassas öznitelik fonksiyonunu göstermek üzere, entropi ℓ-çeşitliliğe sahip bir tabloda, Entropi (EC) ≥ log ℓ eşitliği sağlanmalıdır. Bir tablonun entropi ℓ-çeşitlilik gereksinimini sağlaması için, her bir eşlenik sınıf entropisinin değeri en az log ℓ olması gerekir. Değerlerin çok sık görülmesi durmunda, tablonun entropisi düşer.
- *Özyinelemeli (c,ℓ)-Çeşitlilik*, eşlenik sınıflardaki hassas değerlerin sıklıklarının ayarlanmasını sağlar. EC içerisinde yer alan hassas değerler S_1, …, S_m olsun, r_1=n (EC_i, S_1), ..., r_m= n (EC_i, S_m) EC_i içerisindeki hassas değerlerin sayısını azalan sırada gösterdiğinde c sabit olmak üzere if $r_1 < c(r_2 + \ldots + r_m)$ şartını sağlayan EC_i özyinelemeli (c, ℓ)-Çeşitlidir. Bir tablonun özyinelemeli (c, ℓ)-Çeşitliliğe sahip olabilmesi için tabloda yer alan tüm EC'lerin özyinelemeli (c,ℓ)-çeşitliliğe sahip olması gerekir.

Öznitelik ifşalarına karşı farklı modeller de kullanılmıştır [52-55]

2.4.3 t-Yakınlık

ℓ-Çeşitlilik, güçlü bir mahremiyet modeli olmasına rağmen, Li ve ark. [56] çarpık veri dağılımına sahip mikro veri tablolarıyla karşılaşıldığında mahremiyet koruması için ℓ-

çeşitlilik ilkesinin yetersiz olduğunu göstererek t-Yakınlık modelini önermişlerdir. ℓ-Çeşitlilik, hassas değerler arasındaki anlamsal yakınlıklara ve eşlenik sınıf içindeki hassas değerlerin dağılımının genel dağılımdan önemli ölçüde farklı olmasına bağlı olarak yapılacak olan çarpıklık saldırılarına karşı mahremiyet korumasında yetersiz kalır.

t-Yakınlık, eşlenik sınıf içerisindeki SA dağılımı ile tablo genelindeki SA dağılımın farkının t eşik değerini geçemeyeceğini garanti eder. Bir tablonun t-Yakın olabilmesi için tablo içerisindeki tüm eşlenik sınıfların t-Yakınlık gereksinimini sağlaması gerekmektedir. t-Yakın bir tabloda hassas değerlerin dağılımını bilen bir saldırgan eşlenik sınıflar hakkında sınırlı bir bilgiye sahip olur. t-Yakınlık modeline göre, saldırganın tablo genelindeki dağılımdan edindiği bilgi ile her bir eşlenik sınıftan edindiği bilgi arasındaki fark t eşik değerini geçtiğinde öznitelik ifşaları meydana gelir.

Tablo genelindeki hassas değerlerin dağılımı Q eşlenik sınıf içerisindeki hassas değerlerin dağılımı P olduğunda iki dağılım arasındaki mesafenin D(P,Q) hesaplanmasında değişimsel mesafe (variational distance), Kullback-Leibler(KL), Toprak Taşıyıcı Mesafe (Earth Mover's Distance-EMD) gibi farklı algoritmalar kullanılır [57, 58]. t-Yakınlık modelinde P ve Q dağılımları arasındaki uzaklığın hesaplanmasında EMD algoritması yaygın olarak kullanılır. EMD, dağılımlar arasındaki mesafeleri öznitelik tipine göre (sayısal, kategorik vb.) hesaplar. Eşlenik sınıf içerisindeki dağılımlar P = (p_1, p_2, ...p_m), ve tablo genelindeki dağılımlar Q = (q_1, q_2, ...q_m) olduğunda, sayısal öznitelik (yaş, maaş vb.) dağılımları arasındaki mesafenin hesaplanması Eşitlik 2-2'de verilmiştir.

$$D[P,Q] = \frac{1}{m-1}\sum_{i=1}^{m} \left| \sum_{j=1}^{j=i} (p_j - q_j) \right| \qquad 2\text{-}2$$

EMD ile sayısal olmayan kategorik hassas değerlerin (hastalık vb.) dağılımları arasındaki mesafenin hesaplanmasında, Eşit Mesafe (Equal Distance) veya Hiyerarşik Mesafe (Hierarchical Distance) olmak üzere iki farklı yaklaşım kullanılır [59]. Eşit mesafe yaklaşımında iki hassas değer arasındaki mesafenin 1 olduğu kabul edilir. Eşit mesafenin hesaplanması Eşitlik 2-3'de verilmiştir.

$$D[P,Q] = \frac{1}{2}\sum_{i=1}^{m}(p_j - q_j) \qquad 2\text{-}3$$

Kategorik bir özniteliğin iki değeri arasındaki uzaklık, bu iki değerin genelleştirme ağacına göre göre aynı değere genelleştirildiği minimum düzeydir. Genelleştirme hiyerarşisi ile P ve Q dağılımları göz önüne alındığında, i öğesine karşılık gelen yaprak düğümün mesafesi p_i-q_i, yaprak olmayan düğümlerin ekstralarının toplamı olarak tanımlanır. Ekstra fonksiyonu özyinelemeli olarak Eşitlik 2-4'de verilmiştir.

$$extra(N) = \begin{cases} p_i - q_i, & N\ bir\ yapraksa \\ \sum_{C \in Child(N)} extra(C), & diğer\ durumlarda \end{cases} \qquad 2\text{-}4$$

Child (N) N'nin altındaki çocuk düğümlerin kümesi, pos_ekstra fonksiyonu, aynı seviyedeki ekstra değerler toplamının pozitif olduğu düğümler, neg_ekstra fonksiyonu, aynı seviyedeki extra değerler toplamının negatif olduğu düğümler olmak üzere pozitif ve negatif ekstra fonksiyonları Eşitlik 2-5 ve Eşitlik 2-6'da verilmiştir.

$$pos_{extra(N)} = \sum_{C \in Child(N) \wedge extra(C) > 0} |extra(C)| \qquad 2\text{-}5$$

$$neg\_extra(N) = \sum_{C \in Chil\ (N) \wedge extra(C) < 0} |extra(C)| \qquad 2\text{-}6$$

N'nin alt düğümleri arasındaki hareketlerin maliyetinin hesaplanması (N) Eşitlik 2-7'de verilmiştir.

$$Cost(N) = \frac{height(N)}{H} min(pos\_extra(N), neg\_extra(N)) \quad 2\text{-}7$$

EM, Hiyerarşik Mesafe hesaplaması Eşitlik 2-8'de verilmiştir.

$$D[P,Q] = \sum_N Cost(N) \quad 2\text{-}8$$

2.4.4 δ-Mevcudiyet

Açık kaynaklar, sosyal ağlar, yazılı ve görsel basın, sohbet ve gerçek dünyadaki ilişkilerden elde edilebilen arka plan bilgileri mahremiyet saldırılarının ve ihlallerinin yaşanmasında önemli rol oynar. Arka plan bilgisine sahip saldırganın yayınlanan verilerde kurbanın olup olmadığını bilmesi önemli bir mahremiyet zafiyeti oluşturur. Üyelik bilgisine ve arka plan bilgisine sahip olan saldırgan veri bağlama yöntemleriyle yapacağı saldırılar sonucunda yeniden kimliklendirme yapabilir.

ℓ-Çeşitlilik ve k-Anonimlik modelleri kimlik ve öznitelik ifşalarına karşı koruma sağlarken üyelik ifşalarına karşı koruma sağlayamaz. Üyelik bilgisinin keşfini zorlaştırarak mahremiyet riskini azaltmak amacıyla Nergiz ve ark. [60] δ-Mevcudiyet modelini önermiştir. Temel yaklaşım, yayınlanan veri kümesinin saldırganın arka plan bilgisini temsil eden genel veri kümesinin alt kümesi olarak modellenmesidir.

δ-Mevcudiyet, saldırganın veri bağlama saldırısında kullanabileceği P tablosuna eriştiğini varsayar. Yayınlanan veri kümesiyle P tablosu arasında bağlantı kurmanın bir mahremiyet riski olduğu göz önüne alındığında, T tablosuna bu riski azaltacak uygun bir sterilizasyon işleminin yapılması gerekir. P harici genel tablo, T özel tablosunun yayınlanmış hali T* olmak üzere T'nin T* dönüşümü için δ-mevcudiyet formülü Eşitlik 2-9'da verilmiştir.

$$\begin{cases} \delta = (\delta_{min}, \delta_{max}) ; \\ if \quad \delta_{min} \leq \mathcal{P}(t \in T \mid P, T^*) \leq \delta_{max} \ \ \forall\, t \in P \end{cases} \quad 2\text{-}9$$

Eşitlik 2-9'da (δ_{min}, δ_{max})-Mevcudiyet modeli, genel veri kümesinde (P) yer alan bir kaydın, yayınlanacak veri kümesinde (T*) yer alması ihtimalinin δ_{min} ve δ_{max} arasında olmasını garanti eder.

2.4.5 Diferansiyel Mahremiyet

Diferansiyel mahremiyet veri kümesindeki kayıtların tamamı yerine tekil bir kaydın mahremiyetini koruyan sözdizimsel model olup, veritabanına yapılacak olan ekleme veya çıkarmaların sorgulama sonuçlarını değiştirmeyeceğini garanti eder [50, 61]. Bu model, veritabanı üzerinde çalıştırılan sorgulara cevap verilmesiyle meydana gelebilecek çıkarımlara karşı mahremiyeti koruyarak tek bir kayıt düzeyinde farklılıkları olan komşu veritabanları üzerinde çıkarım yapmaya çalışan saldırgana ek bilgi verilmemesini sağlar. D veritabanına sadece bir kayıt ekleme veya çıkarmayla D' veritabanı (komşu veritabanları) oluşsun, $A:D^n \rightarrow \hat{y}$ rastgele seçilmiş D, D' üzerinde çalışan bir fonksiyon, $D \wedge D' \in D^n$, çıktı kümesi $Y \subset \hat{y}$ ve $\varepsilon > 0$ olduğunda olduğunda Diferansiyel mahremiyetin tanımı Eşitlik 2-10'da verilmiştir.

$$PR[A(D) \in Y] \leq e^{\varepsilon} PR[A(D') \in Y] \qquad 2\text{-}10$$

Eşitlik 2-10'da ε kullanıcı tarafından belirlenen ayarlanabilir bir değer olup D ve D' komşu veritabanlara gönderilen iki sorgunun sonuçları arasındaki farkın ihmal edilebilecek kadar küçük olmasını yani sorgu sonuçlarının birbirinden ayırt edilememesini sağlar. Diferansiyel mahremiyet modeli, ortalama maaş gibi bir sorgunun komşu veritabanı üzerinde çalıştırılmasıyla elde edilen sonuçların $e^{(\varepsilon)}$'dan farklı olamayacağını garanti eder. Sorgu sonuçları arasındaki farkın ihmal edilebilecek kadar küçük olması, saldırganın sorgu sonucunun hangi komşu veritabanından geldiğini tahmin edememesini sağlar. Böylece iki veri tabanı arasındaki tekil bir kaydın varlığı ortaya çıkmamış olur.

Diferansiyel mahremiyet modelinde gürültü eklemede kullanılan en yaygın yöntem Laplace dağılımıdır [62]. Laplace dağılımında gürültü ölçeği sorgu sonucunun hassasiyetine göre ayarlanır. $D \wedge D' \in D^n$ olmak üzere $f:D \rightarrow R$ fonksiyonun hassasiyeti Eşitlik 2-11'de verilmiştir.

$$\Delta f = max_{D,D'}|f(D) - f(D')| \quad 2\text{-}11$$

Hassasiyeti (Δf) aralarında bir kayıt fark bulunan iki veritabanının sorgusundan dönen en büyük fark olarak tanımladığımızda gürültü Laplace (Δf/ε) dağılımıyla hesaplanır. Hassasiyet ile ölçeklenmiş olan Laplace dağılımı, uygun gürültü miktarını üretecek ve bu değer sorgu sonucuna eklendiğinde sonuç üzerinde mahremiyet-fayda dengesini gözeten mahremiyet koruması sağlanmış olacaktır.

2.5 Veri Anonimleştirme

Mahremiyet korumalı yaklaşımlarla verilerin yayınlanması sürecinde mahremiyet gereksinimlerinin karşılanması ve ihlallerin oluşmaması amacıyla mahremiyet modelleri kullanılır. Mahremiyet modellerinin uygulanmasında anonimleştirme tekniklerinden faydalanılır. Anonimleştirme kimlik ve hassas özniteliklerin ifşasının önlenmesi amacıyla mahremiyet modelleri tarafından yarı tanımlayıcı öznitelikler üzerinde yapılan dönüşüm işlemleridir. Anonimleştirmeyle verinin tipi ve biçimi korunarak paylaşılmış veri kümelerinde yer alan veri sahiplerinin kimlik bilgileri ve hassas verilerinin ifşa edilmesi zorlaştırılır. Veri anonimleştirme ilk defa 1981 yılında Chaum tarafından önerilmiş ve ilk uygulama Jakobsson tarafından yapılmıştır [63, 64].

Anonimleştirmenin kabul edilebilir düzeyde veri kaybıyla yapılması veri faydası açısından önemlidir. Veri kayıplarındaki artış veri kalitesini düşürerek paylaşılan veriden sağlanan faydanın azalmasına hatta tamamen yok olmasına yol açar. Anonimleştirme genellikle veri toplayıcılar veya mahremiyet problemine göre veri sahipleri tarafından tablo içerisindeki niteliklerin veya kayıtların eşlenik sınıflara ayrılmasıyla yapılır. Aynı eşlenik sınıf içindeki kayıtlar yarı tanımlayıcı öznitelikler üzerinden ayırt edilemez. Yaygın olarak kullanılan anonimleştirme yöntemleri takip eden alt bölümlerde özetlenmiştir.

2.5.1 Genelleştirme

Genelleştirme bilgisayar bilimleri, istatistik, biyoloji, biyoinformatik gibi birçok alanda veri analizi için kullanılan yaygın bir yöntemdir [65]. Verinin anlamsal bütünlüğüne sadık kalarak daha az detay içerecek şekilde ifade edilmesini sağlayan genelleştirme yöntemi anonimleştirmede yaygın olarak kullanılan detay azaltma tekniklerindendir [43].

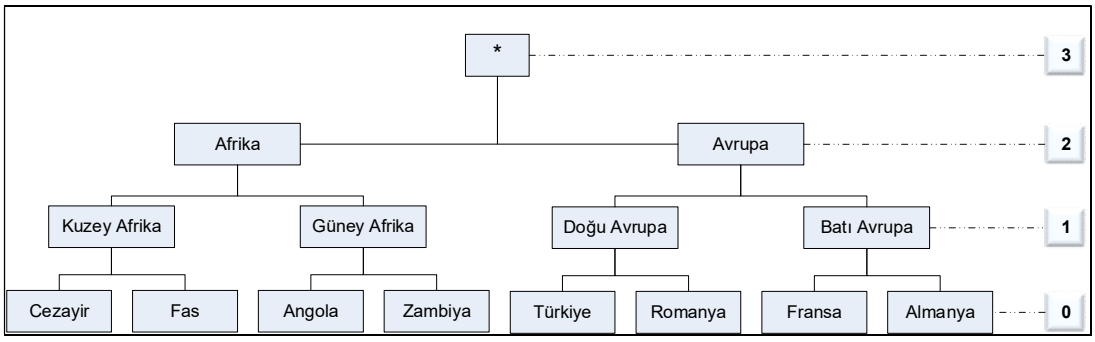
Genelleştirme hiyerarşileri mahremiyet korumalı yaklaşımlarda Sweeney [42] ve Samarati [66] tarafından güçlü ve yararlı bir teknik olarak önerilmiş ve kullanılmıştır.

Genelleştirmeler mikro veri tablolarında yer alan öznitelik değerlerinin tiplerine göre yapılır. Örneğin, sayısal değerler belirli aralıklar içerisinde, sınıflandırılması mümkün olan kategorik değerler ise daha üst seviyede veya çok sayıda farklı içeren posta kodu gibi değerler maskeleme yöntemleri ile kapsayıcı bir üst küme ile genelleştirilir [67]. Verileri içeren tablo T, dom (A_i, T) ise T tablosundaki A özniteliğin alanını, T_i (A_1,.., A_n), T_j (A_1,.., A_n) öznitelikleri aynı olan iki tabloyu, t [A], T tablosundaki A özniteliğini içeren kaydı, $T_j \gtrsim T_i$ T_i tablosunun genelleştirilmiş halinin T_j tablosu olduğunu göstermek üzere genelleştirmenin tanımı Eşitlik 2-12 de verilmiştir.

$$T_j \gtrsim T_i, iff \begin{cases} |T_j| = |T_i| & (1) \\ \forall A_z \in \{A_1, .., A_n\} : dom\,(A_z, T_i) \leq_D dom(A_z, T_j) & (2) \\ t_i[A_z] \leq_V t_j[A_z] \;;\; A_z \in \{A_1, .., A_n\} & (3) \end{cases} \qquad 2\text{-}12$$

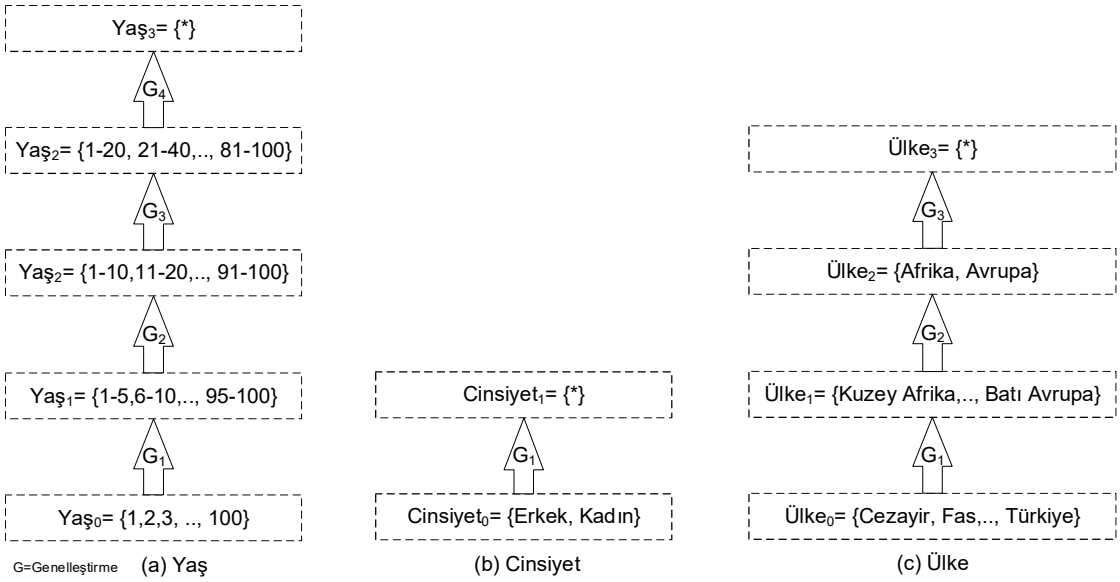
Eşitlik 2-12'de, T_j ve T_i tabloları aynı sayıda kayıt (tuple) içerir (1). T_i veya genelleştirilmiş T_i* tablosundaki her bir niteliğin etki alanı Tj* tablosundaki her bir niteliğin etki alanına eşit (2), T_i veya genelleştirilmiş T_i * tablosundaki her bir t_i kayıt içerisindeki özniteliklerin değeri T_j içindeki her bir kaydın t_j öznitelik değerine eşittir (3).

Verilerin genelleştirilmesi gösteriminde, alan genelleştirme ve değer genelleştirme olmak üzere iki farklı hiyerarşik gösterim kullanılır [68]. Alan Genelleştirme Hiyerarşisi (Domain Generalization Hierarchy-DGH) özniteliğin karakteristiğine göre alan uzmanları tarafından oluşturulur. Değer Genelleştirme Hiyerarşileri (Value Generalization Hierarchy -VGH) ise yapraklardan köklere doğru gidildikçe bir değerin daha az detayda ifade edilmesini sağlayan ağaç gösterimini kullanır [69]. Ülke özniteliğine ait örnek VGH gösterimi Şekil 2-3'de verilmiştir.



Şekil 2-3 Ülke özniteliği değer genelleştirme hiyerarşisi

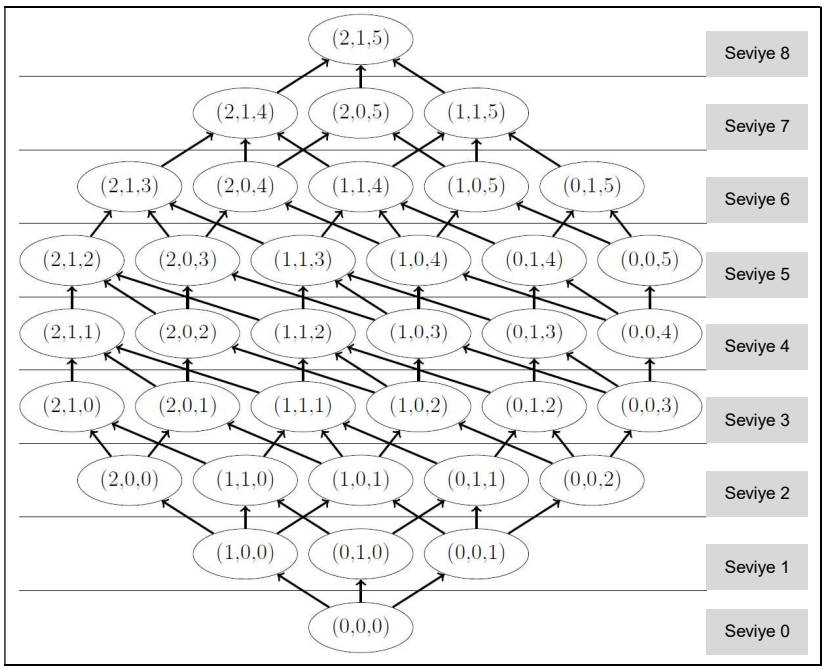
Verilerin genelleştirilmesi sırasında Şekil 2-3'de örneği verilen yüksekliği h=3 olan hiyerarşi ağaçlarının yapraklarında orijinal öznitelik değerleri yer alırken köke doğru daha genel ifadeler yer alır. İfadeler genelleştikçe veriden sağlanan fayda azalmakta verinin mahremiyeti artmaktadır. Diğer bir gösterim örneği olarak, Yaş, Cinsiyet ve Ülke öznitelikleri için DGH gösterimi Şekil 2-4'de verilmiştir.



Şekil 2-4 Yaş, cinsiyet, ülke öznitelikleri alan genelleştirme hiyerarşisi

Şekil 2-4 (a)'da sayısal veri tipine sahip Yaş özniteliği 5 detay seviyede, Şekil 2-4 (b)'de kategorik veri tipine sahip olan Cinsiyet özniteliği 2 detay seviyede, Şekil 2-4 (c)'de kategorik veri tipine sahip olan Ülke özniteliği 4 yüksekliğinde gruplandırılarak özniteliklerin genelleştirme hiyerarşileri oluşturulmuştur. Her bir öznitelik hiyerarşinin alt seviyesinden üst seviyesine genelleştirildikçe ($G_i \rightarrow G_{i+1}$) veri detayları budandığından fayda

azalmakta, mahremiyet artmaktadır. Veri faydası ve mahremiyet dengesine göre farklı hiyerarşi seviyelerinden seçilecek birleşik tanımlayıcı özniteliklerini gösteren düğümler çözüm için kullanılır. Düğümlerin tamamı, ardıl ve öncül ilişkileriyle birlikte Hasse çizgeleriyle gösterilir [70, 71]. Yaş, cinsiyet ve posta kodundan oluşan QID özniteliklerine ait genelleştirme çözüm örüntüsünü gösteren örnek bir Hasse çizgesi Şekil 2-5'de verilmiştir [72].



Şekil 2-5 QID genelleştirme örüntüsü örneği

Şekil 2-5'de 3 öznitelikten oluşan olası tüm QID genelleştirmeleri içeren arama veya çözüm uzayı yönlü Hasse çizgesiyle sıralı olarak gösterilmiştir. Her bir düğüm öncül düğümünde yer alan QID öznitelik kümesindeki herhangi bir özniteliğin genelleştirme seviyesindeki değişimiyle oluşur. Örüntü içerisindeki her bir aday düğüm QID öznitelik kümesi için uygulanacak olası genelleştirme kuralını gösterir. Örneğin {1,0,4} aday düğümü çözüm olarak QID öznitelik kümesine uygulandığında, yaş özniteliği 1. seviyeden, cinsiyet özniteliği değişmeden, posta kodu özniteliği ise 4. seviyeden genelleştirilerek dönüşümü sağlanır. En yüksek veri faydası {0, 0, 0} (infumum), en yüksek mahremiyet koruması ise {2, 1, 5} (supremum) düğümleridir.

Anonimleştirme çalışmalarında çözüm düğümüne göre verileri yeniden kodlanmasında (Recoding) yerel kodlama (local recoding) ve küresel kodlama (global recoding) olmak

üzere iki farklı yöntemle yapılır [73]. Küresel kodlamada, detay bir değerin genelleştirilmesinde tüm kayıtlar için aynı genelleştirme hiyerarşisi kullanılır. Yerel kodlama ise aynı öznitelik değerinin genelleştirilmesinde farklı kayıtlarda farklı değerlere izin verir [74]. Küresel ve yerel kodlama örnekleri Çizelge 2-5'de gösterilmiştir.

Çizelge 2-5 Yerel ve küresel kodlama örnekleri

T	**QID**		
ID	PK	Yaş	C
1	53712	24	E
2	53711	25	K
3	53711	30	E
4	53711	30	K
5	53712	32	E
6	53712	32	K

a) Orijinal tablo

T*	**QID**		
ID	PK	Yaş	C
1	53712	[24-32]	*
5	53712	[24-32]	*
6	53712	[24-32]	*
2	53711	[24-32]	*
3	53711	[24-32]	*
4	53711	[24-32]	*

b) Küresel kodlama

T*	**QID**		
ID	PK	Yaş	C
1	5371*	[24-30]	*
2	5371*	[24-30]	*
3	5371*	[24-30]	*
4	5371*	[30-32]	*
5	5371*	[30-32]	*
6	5371*	[30-32]	*

c) Yerel kodlama

Çizelge 2-5 (a) da 3 ve 4 numaralı kayıtlar küresel kodlamada aynı eşlenik sınıfta yer alırken, yerel kodlamanın PK özniteliğinin farklı şekilde genelleştirilmesine izin vermesinden dolayı farklı eşlenik sınıflar içerisinde yer almıştır. Yerel kodlama küresel kodlamaya göre daha az bilgi kaybına sebep olmasına rağmen, en uygun çözümün bulunması için kullanılan çözüm uzayının çok büyümesi nedeniyle optimum çözümü bulmanın maliyeti yüksektir. Küresel kodlamada en uygun çözümün bulunması için LeFevre ve ark. Incognito algoritmasını önermişlerdir [75].

Yeniden kodlamanın başka bir sınıflandırması ise hiyerarşi ağacındaki genelleştirme uygulamasına göre yapılır. Aynı genelleştirme kuralı, ağacın tamamında uygulanıyorsa tam genelleştirme (Full-Domain), alt bir ağaçta uygulanıyorsa, kısmı genelleştirme (Subtree) olarak adlandırılır [76]. Kısmı olarak yeniden kodlama farklı seviyelerde farklı genelleştirme kuralının uygulanmasına izin verir. Kısmı kodlama çözüm uzayında optimuma en yakın çözümün bulunması için Iyengar [77] genetik algoritmayla sezgisel aramayı, LeFevre kd-tree [78] yaklaşımlarını önermişlerdir.

Yeniden kodlamanın seçimine veri faydası-mahremiyet dengesi açısından yapılan değerlendirmelere göre veri toplayıcılar karar verir. Yeniden kodlamalar çoğunlukla yarı tanımlayıcı öznitelikleri üzerinde yapılmasına rağmen veri mahremiyeti açısından yapılan

değerlendirmeler sonucunda hassas öznitelikler üzerinde de yapılabilir [79]. Örneğin hastalık özniteliğine ait gastrit ve ülser değerleri veri faydası açısından yaşanacak kayıp tolere edilebilir düzeyde ise mide hastalığına genelleştirilebilir. Çok nadir görülen hastalıkların diğer hastalıklar sınıfına genelleştirilmesi ise mahremiyetin korunması açısından hassas özniteliklerin genelleştirilmesine verilebilecek bir diğer örnektir.

Mahremiyet modelleri tarafından kullanılan genelleştirme uzayında en iyi genelleştirme çözümünün bulunmasının polinomsal zamanda çözülemeyen zor bir problem olduğu Meyerson [80]ve Bayardo [81] tarafından gösterilmiştir. Zor problemlerde polinomsal zamanda en iyi çözümü bulmak mümkün olmadığından genelleştirme problemlerinin çözümünde optimuma en yakın çözüm için literatürde çalışmalar yapılmıştır [82].

2.5.2 Baskılama

Baskılama tek bir değer (cell suppression), kayıt (tupple suppression) veya öznitelik (attribute) seviyesinde yapılan karakter maskeleme işlemidir [83]. Genelleştirme yöntemleriyle birlikte tamamlayıcı olarak kullanıldığında etkin sonuçlar verir. Genelleştirme ve baskılama tekniğiyle yapılan anonimleştirme işlemleri sonucunda veri mahremiyeti gereksinimini sağlayamayan aykırı (outlier) kayıtlar meydana gelir. Örneğin 5-Anonimlik modelinin genelleştirme ve baskılama yöntemiyle uygulandığı bir anonimleştirme işleminde $k \leq 4$ olan tüm durumlar için aykırı kayıtlar meydana gelecektir. Tamamen bastırılan kayıtlar aykırı eşlenik sınıf (outlier equivalence class-OEC) adı verilen tek bir eşlenik sınıfta toplanır. Tek bir eşlenik sınıfta toplanan tüm aykırı kayıtlar $k \geq 5$ mahremiyet gereksinimini karşılayarak mahremiyet probleminin çözümüne katkıda bulunur [72]. Genelleştirme ve baskılama örneği Çizelge 2-6'da verilmiştir.

Çizelge 2-6 Baskılama örneği

QID			SA
64	K	06100	Ülser
64	K	06100	Ülser
64	E	06100	Ülser
64	E	06100	Ülser
63	E	34100	Lösemi
63	E	34100	Lösemi
64	K	34060	Lösemi
64	K	34060	Lösemi
61	E	34100	Zatürre
61	E	34100	Zatürre
61	K	35100	Zatürre
62	E	35060	Kalp

a) Orijinal Tablo

QID			SA
[60-64]	K	061*	Ülser
[60-64]	K	061*	Ülser
[60-64]	E	061*	Ülser
[60-64]	E	061*	Ülser
[60-64]	E	341*	Lösemi
[60-64]	E	341*	Lösemi
[60-64]	K	340*	Lösemi
[60-64]	K	340*	Lösemi
[60-64]	E	341*	Zatürre
[60-64]	E	341*	Zatürre
[60-64]	K	351*	Zatürre
[60-64]	E	350*	Kalp

b) 2-Anonim Tablo

QID			SA
[60-64]	K	061*	Ülser
[60-64]	K	061*	Ülser
[60-64]	E	061*	Ülser
[60-64]	E	061*	Ülser
[60-64]	E	341*	Lösemi
[60-64]	E	341*	Lösemi
[60-64]	K	340*	Lösemi
[60-64]	K	340*	Lösemi
[60-64]	E	341*	Zatürre
[60-64]	E	341*	Zatürre
[60-64]	*	35**	Zatürre
[60-64]	*	35**	Kalp

c) 2-Anonim Tablo

Çizelge 2-6 (a)'da orijinal hali verilen tablo k-Anonimlik modeline göre k=2 seçilerek genelleştirme ve baskılama yöntemleriyle anonimleştirilmiştir. 2-Anonimlik şartını sağlayamayan sınıflandırılamamış iki aykırı kayıt görülmüş olup bu kayıtların yayınlanabilmesi amacıyla baskılama düzeyi arttırılarak tüm tablonun 2-Anonim hale gelmesi sağlanmıştır. Genelleştirmeyle birlikte baskılamanın formal tanımı Eşitlik 2-13'de verilmiştir.

$$T_j \gtrsim T_i, iff \begin{cases} |T_j| \leq |T_i| & (1) \\ \forall A_z \in \{A_1, .., A_n\} : \ dom\,(A_z, T_i) \leq_D dom(A_z, T_j) & (2) \\ t_i[A_z] \leq_V t_j[A_z] \ ; \ A_z \in \{A_1, .., A_n\} & (3) \end{cases} \qquad 2\text{-}13$$

Eşitlik 2-12 ve Eşitlik 2-13 incelendiğinde T_i'de görülen kayıtların T_j'de karşılıklarının olmayabileceği (1) nolu eşitlik satırında gösterilmiştir. Bu durum, mahremiyet gereksinimlerini karşılamak üzere T_j'de baskılanmış olan kayıtların tablodan kaldırıldığını gösterir. Baskılamanın gereğinden fazla yapılması veri faydasını olumsuz etkilediği için ihtiyacın ötesinde baskılamaya izin verilmemelidir. T_i orijinal, T_j k-Anonimlik gereksinimlerini karşılayan T_i'nin genelleştirilmiş hali olmak üzere baskılamanın ihtiyaç duyulan seviyede kalması için gerekli Eşitlik 2-14'de verilmiştir.

$$En\ az\ baskılama\ iff \begin{cases} \forall T_z : T_z \gtrsim T_i, D(V_{i,z}) = D(V_{i,j}) & (1) \\ T_z\ tablosu\ k-anonim\ \rightarrow |T_z| \leq |T_j| & (2) \end{cases} \quad \text{2-14}$$

Genelleştirme ve baskılama, mahremiyet gereksinimlerinin sağlanmasına yönelik iki farklı anonimleştirme yaklaşımıdır. Tek başına genelleştirmenin yetersiz olduğu durumlarda baskılama devreye girerek anonimleştirmenin iyileştirilmesine yardımcı olur. Tek başına kısmi baskılama değer alanı yüksek olan özniteliklerde (PostaKodu vb.) kullanılır. Tamamen baskılama ise veri faydası açısından istenmeyen bir durumdur. Sonuç olarak her iki yaklaşım birlikte uygulandığında anonimleştirme açısından en iyi sonuçlar alınır.

2.5.3 Anatomi

Genelleştirme ve baskılama mahremiyet korumasında etkili çözümler sunmasına rağmen, anonimleştirme sırasında öznitelikler üzerinde yapılan değişikliklere bağlı meydana gelen kayıplar veri analizinin doğruluğunu etkiler. Anatomi yöntemi öznitelikler üzerinde değişiklik yapmak yerine öznitelikler arasındaki ilişkinin zayıflatılması prensibine göre çalışır. Xiao ve Tao genelleştirme temelli algoritmalara kıyasla verimliliği arttırarak bilgi kaybını en aza indirgeyen ve ℓ-çeşitlilik gereksinimlerini sağlayan anatomi yöntemini önermişlerdir [24].

Bu yöntemde, öznitelik ilişkisinin zayıflatılması amacıyla yayınlanacak tablo QIT ve ST olmak üzere iki tabloya ayrılır. Yarı tanımlayıcı öznitelikler QIT tablosunda, hassas öznitelik ise ST tablosunda tutulur. QIT ve ST tablolarının bağlantısının kurulabilmesi amacıyla GroupID ortak özniteliği her iki tabloya eklenir. Aynı gruptaki tüm kayıtlar, her iki tabloda da aynı GroupID değerine sahip olacağından her bir gruptaki hassas değerlerle yarı tanımlayıcı değerler tam olarak eşleşir. Eğer bir grup ℓ farklı hassas değere sahip ve her bir hassas değer grupta tam olarak bir defa yer alıyorsa, bir kaydı GroupID üzerinden hassas bir değere bağlama olasılığı 1 / ℓ olur. Bu durumda öznitelik bağlama riskleri ℓ kadar arttırılarak düşürülebilir. Anatomi yöntemiyle anonimleştirmenin formal tanımı Eşitlik 2-15'de verilmiştir.

$$T(A) \rightarrow \{QIT;\ ST\}\ iff \begin{cases} QIT = (A_1, A_2, .., A_d; i);\ T(QID) = \{A_1, .., A_d\} \\ ST = (GroupID; A_s; Count) \end{cases} \quad 2\text{-}15$$

Anatomi örneği Çizelge 2-7'de verilmiştir. Çizelgede anatomi yöntemiyle karşılaştırma yapılabilmesi amacıyla orijinal tablonun genelleştirilmiş gösterimi de verilmiştir.

Çizelge 2-7 Anatomi örneği

QID		SA
40	E	Hepatit
40	E	Hepatit
40	E	Lösemi
42	E	Hepatit
42	E	Lösemi
42	E	Lösemi
46	K	Grip
48	K	Grip
48	K	Ülser
48	K	Ülser

a) Orijinal tablo

QID		SA
[40-45]	E	Hepatit
[40-45]	E	Hepatit
[40-45]	E	Lösemi
[40-45]	E	Hepatit
[40-45]	E	Lösemi
[40-45]	E	Lösemi
[46-51]	K	Grip
[46-51]	K	Grip
[46-51]	K	Ülser
[46-51]	K	Ülser

b) Genelleştirilmiş tablo

QID		GID
40	E	1
40	E	1
40	E	1
42	E	1
42	E	1
42	E	1
46	K	2
48	K	2
48	K	2
48	K	2

c) QIT tablosu

GID	SA	Σ
1	Hepatit	2
1	Lösemi	2
2	Grip	2
2	Ülser	2

d) ST tablosu

Çizelge 2-7 (a)'da SA = {Hastalık} ve QID = {Yaş, Cinsiyet} özniteliklerine sahip örnek tablo anatomi yöntemi ile ℓ =2 değeriyle yayınlanmak isteniyor. İlk olarak Çizelge 2-7 (b)'de gösterilen genelleştirme işlemi sonucunda her bir farklı en az ℓ tane SA değerinin sadece aynı grupta yer aldığı E_1= {[40-45], E, {SA}} ve E_2= {[46-51], K, {SA}} grupları oluşturulur. Bu gruplara numara verilerek Çizelge 2-7 (c)'deki QIT tablosunun oluşturulması amacıyla orijinal tablodaki tüm SA değerleri grup numarasını gösteren GID özniteliğiyle değiştirilir. Çizelge 2-7 (d)'deki ST tablosunun oluşturulması amacıyla her bir grup içinde yer alan SA değerlerinin toplamı bulunarak ST tablosuna eklenen Toplam özniteliği (Σ) içerisine yazılır. QIT ve ST tablolarının mahremiyet gereksinimlerini sağladıkları kontrol edilerek her iki tablo anatomi yöntemiyle veri alıcıları için yayınlanır.

Anatomi yönteminin en büyük avantajı hem QIT hem de ST'deki verilerin değiştirilmeden yayınlanmasıdır. Anatomik tablolar genelleştirilmiş tablolara göre QID ve SA değerlerini içeren çoklu (aggregate) sorgular için daha doğru sonuçlar verir. Örneğin, Çizelge 2-7 (c) ve

(d) tablolarına göre araştırmacının ülser hastalığı olan 48 yaşındaki hasta sayısını araştırdığını varsayalım. Orijinal Çizelge 2-7 (a) incelendiğinde 2 tane ülser hastasının olduğu görülür. Çizelge 2-7 (c) QIT tablosunda GID = 2 olan 4 kayıttan 2'sinin ülser hastalığı olması 2/4 olarak bulunur. 48 yaşında olan hastaların sayısının da 3 olması nedeniyle anatomik tablolardan beklenen sayı 3 × 2/4 = 1,5 olarak bulunur. Bu sayı tablonun genelleştirilmiş halinin gösterildiği Çizelge 2-7 (b)'de verilen tablodaki sayıdan (2/4) daha doğru bir sayıdır.

Anatomi tekniğinde iki tablonun yayınlanması veri alıcılar için yeni araçların ve algoritmaların tasarlanmasını gerektirmiştir. Bu durum veriden fayda sağlayacak veri alıcılar için maliyet gerektiren bir durumdur. Özellikle sınıflandırma kümeleme gibi temel veri madenciliği araçları için yeni yaklaşımların iki tablolu veri yayınlamaya uygun olarak yeniden çalışılması gerekmektedir. Verilerde herhangi bir değişiklik yapılmamasına bağlı olarak veri faydasının arttırılması bu tekniğin güçlü tarafını oluştururken, özellikle arka plan bilgilerinin desteğiyle yapılan üyelik ifşa saldırılarına karşı yeterli koruma sağlayamaması ve yeni yatırımlar gerektirmesi ise zayıf yönüdür.

2.5.4 Permütasyon

Anatomi yönteminin arka plan bilgisiyle yapılan eşleştirme saldırılarına karşı etkisiz olması nedeniyle Zhang ve ark. anatominin gelişmiş bir versiyonu olan permütasyon yöntemini önermiştir [25, 53]. Mahremiyet korumasında anatomiye göre daha etkili bir yöntem olup öznitelikler arasındaki ilişkinin permütasyon yöntemiyle zayıflatılmasını sağlar.

Anatomi yönteminde olduğu gibi mahremiyetin korunması amacıyla verilerin yayınlanmasında iki tablo kullanılır. QID öznitelikleri PQT tablosunda ve SA öznitelikleri PST tablosunda tutulur. Hem PQT hem de PST'ye ortak öznitelik eklenerek iki tablo arasındaki bağlantı sağlanır. Anatomi yönteminde tablolar doğrudan yayınlanırken permütasyon yönteminde aynı gruptaki tüm kayıtlar kendi aralarında permütasyon yöntemiyle karıştırılarak yayınlanır. Permütasyon yöntemiyle anonimleştirmenin formal tanımı Eşitlik 2-16'da verilmiştir.

$$T(P) \rightarrow \{PQT; PST\}\ iff \begin{cases} PQT = (\propto_{i1}(t).A_1, \propto_{i2}(t).A_2, .., \propto_{id}(t).A_d; i) \\ t_i \in [A_d];\ A_d \in \{A_1, .., A_d\};\ \ \propto_{ij} : 1 \leq j \leq d \\ QID = \{A_1, .., A_d\} \\ PST = (i; \propto_{i1}(t).A_s) \end{cases} \quad 2\text{-}16$$

Permütasyon yöntemiyle verilerin yayınlanması örneği Çizelge 2-8'de verilmiştir.

Çizelge 2-8 Permütasyon örneği

QID		SA
55	E	Anfizem
40	E	Kanser
60	K	Nezle
45	K	Gastrit
80	K	Dispepsi
35	E	Nezle
40	K	Pnömoni
30	K	Gastrit
10	E	Bronşit

a) Orijinal tablo

QID		SA
[40-80]	*	Anfizem
[40-80]	*	Kanser
[40-80]	*	Nezle
[40-80]	*	Gastrit
[40-80]	*	Dispepsi
[10-50]	*	Nezle
[10-50]	*	Pnömoni
[10-50]	*	Gastrit
[10-50]	*	Bronşit

b) Genelleştirilmiş tablo

QID		GID
40	K	1
80	E	1
60	K	1
55	K	1
45	E	1
40	E	2
10	K	2
35	E	2
30	K	2

c) PQT tablosu

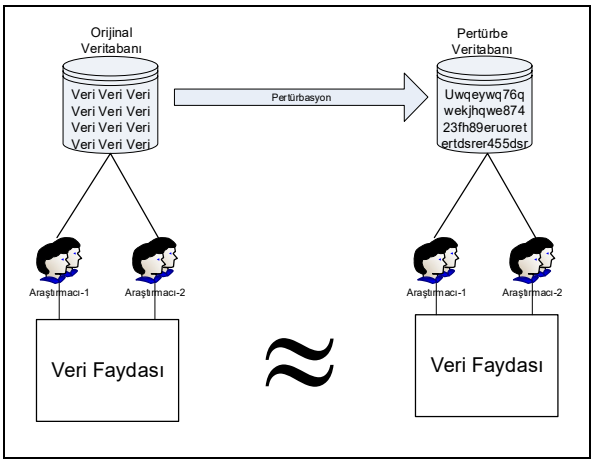
GID	SA
1	Anfizem
1	Kanser
1	Nezle
1	Gastrit
1	Dispepsi
2	Nezle
2	Pnömoni
2	Gastrit
2	Bronşit

c) PST tablosu

Permütasyon, anatomiye göre arka plan saldırılarına karşı daha dirençli bir yöntem olup, özellikle sayısal hassas nitelikler ve çoklu sorgulamalarda anatomi yöntemine yakın veri faydası sunar.

2.5.5 Pertürbasyon

Veri pertürbasyonu, mahremiyet koruması sağlanacak veriler üzerinde istatistiksel açıdan tolere edilebilecek düzeyde değişiklikler yapılmasını sağlayan anonimleştirme tekniğidir [84-86]. Yayınlanacak verilerin türüne ve ihtiyaç duyulan mahremiyet gereksinimlerine bağlı olarak uygulanabilecek tablo veya kayıt düzeyinde uygulanacak pertürbasyon yöntemleri vardır. Veri pertürbasyonun örnek gösterimi Şekil 2-6'da verilmiştir.



Şekil 2-6 Pertürbasyon yöntemi

Yayınlanacak veri kümesindeki her bir kayıt kayıt seri bir şekilde pertürbe işleminden geçirilir. T tablosundaki t kaydı, p olasılığı ile t [A_s] hassas öznitelik değerini korur ve A_s'yi (1-p) olasılığı ile hassas öznitelik alanından rasgele bir değerle değiştirir. Pertürbasyon işlemi sonucunda mahremiyet saldırılarına karşı dayanıklı verilerin oluşturulması amaçlanırken veride meydana gelen kayıpların en az seviyede olması hedeflenir [87]. Şekil 2-6' da gösterildiği gibi orijinal veritabanı ve pertürbe veritabanı üzerinde yapılan istatistiki analizler ile madencilik sonuçlarının benzerlik oranı veri faydası açısından önemlidir. Benzer sonuçlar veride meydana gelen kaybın tolere edilebilir düzeyde ve mahremiyet koruması için p olasılık değerinin çok küçük olduğunu gösterir.

Orijinal verilere anlamlı gürültülerin eklenmesi yaygın olarak kullanılan pertürbasyon yöntemlerindendir. Verilere gürültü eklenmeden önce hesaplanan istatistiki değerlerin pertürbe veriler ile yeniden elde edilebileceği literatürde birçok çalışma ile gösterilmiştir [88, 89]. Bu çalışmalarda verilere gürültü ekleyerek veri mahremiyetinin korunduğu, orijinal dağılıma yakın değerlerin yeniden hesaplanabilmesiyle tolere edilebilecek düzeyde veri kaybıyla veri faydasının korunabildiği gösterilmiştir.

Gürültü ekleyerek veri bozmanın mahremiyeti korumada etkin olmadığı, yeniden dağılımın oluşturulması ile ilgili saldırılara karşı zayıflıkları olduğu literatürdeki farklı çalışmalarda gösterilmiştir [86, 90, 91]. Gürültü ekleme yönteminin mahremiyet korumadaki zafiyetlerini gidermek için eklemeden farklı olarak çarpma yöntemi ile ilgili çalışmalarda yapılmıştır. Ancak çarpma yönteminin de benzer saldırılara karşı korunmasız olduğu görülmüştür [92-94].

Veri bozmada bir diğer yaklaşım mikro agregasyon yöntemidir [95-97]. Bu yöntemde orijinal veriler k veya 2k arasında değer içeren kümelere ayrılır. Kümeler oluşturulurken,

mevcut veri kayıtlarından rassal bir kayıt ilgili kümenin merkezi olarak seçilir. Veri kümesinden merkeze en yakın olan k-1 kayıt bulunarak ilk grup oluşturulur. İlgili küme içerisindeki tüm değerler kümenin merkezi olarak belirlenen değerle değiştirilir. Bir sonraki grup oluşturulmadan önce oluşturulan grup veri kümesinden çıkartılır. k tane grup oluşuncaya kadar bu işlem devam ettirilir. Gruplar için, ortalama ve kovaryans gibi istatistiksel değerler hesaplanır. Bu istatistikî bilgiler, orijinal veri kümesiyle benzer istatistiksel özelliklere sahip anonim verileri oluşturmak için kullanılır. Oluşturulan veri kümelerinin kullanılmasıyla mahremiyeti korunan veriler elde edilir. Elde edilen verilerden fayda sağlamak amacıyla madencilik işlemleri kullanılarak veri alıcısının ihtiyacına yönelik analizler yapılır.

2.6 Mahremiyet Tehditleri

Arka plan bilgileri ile veri bağlama yöntemleri mahremiyete yönelik tehditlerin başında gelir [98]. Yayınlanan veriler ile halka açık veya önceden edinilmiş arka plan bilgilerin bağlanmasıyla yapılan eşleştirmeler sonucunda ifşalar meydana gelir. Arka plan bilgisine sahip saldırgan sahip olduğu bilgiler ile yayınlanan veriler arasında kayıt, hassas öznitelik veya tablo düzeyinde bağlantı kurarak saldırı düzenleyebilir [19, 99, 100]. Bu saldırılar sonucunda kimlik [101-103] öznitelik [104, 105] ve üyelik ifşaları [60] yaşanır. İfşaların mikro veri modeliyle gösterimi Çizelge 2-9'da gösterilmiştir.

Çizelge 2-9 Örnek hastane kayıtları

T*	PK	Yaş	Cinsiyet	Hastalık	
Üyelik İfşası (3)	06***	[70-75]	Erkek	Diyabet	Kimlik İfşası (1)
	34***	[45-50]	Kadın	Migren	Öznitelik İfşası (2)
	34***	[45-50]	Kadın	Migren	
	34***	[45-50]	Kadın	Migren	

Çizelge 2-9'da gösterilen ifşalara sebebiyet veren mahremiyet tehditleri takip eden alt başlıklarda açıklanmıştır.

2.6.1 Arka Plan Bilgileri

Saldırganlar tarafından farklı yollardan elde edilen arka plan bilgileri mahremiyet saldırılarının ve ihlallerinin yaşanmasında önemli rol oynar. Arka plan bilgileri farklı kuruluşlar tarafından yayınlanan verilerden, sosyal ağlardan, gazete ve dergilerden, gerçek dünyadaki sosyal ilişkilerden ve diğer yollardan elde edilebilen hassas olmayan bilgilerdir. Hassas olmayan bu bilgiler veri bağlama yöntemleriyle yayınlanan verilerle eşleştirildiğinde mahremiyet ihlalleri yaşanır. Chen ve ark. [106, 107] mahremiyetin korunmasında arka plan bilgilerini ölçebilen çerçeve bir yapı önermişlerdir. Önerilen çerçeve yapı veri yayıncılarına arka plan bilgileri hakkında fikir vererek tedbir almalarına yardımcı olur. Martin ve ark. [108] veri yayıncıları için arka plan bilgileriyle yapılabilecek saldırılarda en kötü senaryonun dikkate alınması gerektiğini savunan bir yaklaşım ile arka plan bilgisinin ifade edilebilmesini sağlayan bir dil önermişlerdir. Ayrıca en kötü senaryoda hassas bilgilerin ifşa ölçümü için polinomsal zamanda çalışan bir algoritma geliştirmişlerdir.

2.6.2 Kimlik İfşası

Arka plan bilgisine sahip bir saldırganın kamuya açık kimlik bilgileri içeren veritabanlarıyla bu veritabanların alt kümesi olan yayınlanmış kimliksiz verilerin kayıt düzeyinde yarı tanımlayıcılar üzerinden eşleştirilmesi sonucunda kimlik ifşaları meydana gelir [109]. Kimliksizleştirilmiş verileri hedef alan bu saldırı yönteminde, saldırgan kimliksiz yayınlanan veri içerisindeki kurbana ait hassas bilgileri öğrenerek kurbanın kimliğini hassas bilgileriyle birlikte ifşa eder. Kimlik ifşasına örnek olarak Çizelge 2-9 incelendiğinde 74 yaşındaki kurbanın yayınlanan tabloda yer aldığını bilen saldırgan Diyabet hastasının kim olduğunu öğrendiğinde ilgili kaydın kimliğini hassas öznitelikleriyle yeniden tanımlamış olur.

2.6.3 Öznitelik İfşası

Saldırgan sahip olduğu arka plan bilgileri ile yayınlanan tablodaki özniteliklerin homojen dağılımına bağlı olarak kurbanın hassas bilgilerini veri bağlama yapmadan öğrenebilir [51]. Öznitelik ifşasına örnek olarak, Çizelge 2-9'da verilen tabloda saldırgan 43 yaşındaki kadın kurbanın bu tabloda olduğunu bilmektedir. Saldırgan tablodaki hangi kaydın kurbana ait olduğunu öğrenemez ancak migren rahatsızlığı olduğu bilgisinin çıkarımını kolayca yaparak kimliğini tanımlayamadığı kurbanının hassas özniteliğini ifşa eder.

2.6.4 Üyelik İfşası

Saldırgan kurbanın yayınlanan bir veri kümesinde olup olmadığını öğrendiğinde herhangi bir bilgiyi ifşa edemez ancak yayınlanan tabloya göre üst seviye çıkarımlar yapabilir [60]. Çizelge 2-9'da verilen örnek tabloda kurbanın yer aldığını bilen bir saldırgan kurbanın bu tabloyu yayınlayan hastanenin bir hastası olduğunu çıkarımını yaparak üyelik ifşasını gerçekleştirir. Bundan sonraki süreçte saldırgan kurbanın kimlik ve hassas özniteliklerinin ifşası için üyelik ifşasından elde etmiş olduğu bilgiyi geliştirerek arka plan bilgilerini arttırmaya ve bunları kullanacağı kamuya açık diğer veritabanlarını bulmaya çalışır.

2.6.5 Tehditlere Göre Modeller

İfşa temelli saldırıları önleyebilme kabiliyetlerine göre mahremiyet koruma modelleri, Çizelge 2-10'da verilmiştir [98].

Çizelge 2-10 Mahremiyet modelleri

Mahremiyet Modelleri	**Önlediği İfşa Tipleri**		
	Kimlik	Öznitelik	Üyelik
k-Anonimlik	✓		
k^m-Anonimlik	✓		
Çoklu R K-anonimlik	✓		
k-Harita	✓		
ℓ-Çeşitlilik	✓	✓	
(α,k)-Anonimlik	✓	✓	
(X, Y)-Mahremiyet	✓	✓	
(k, e)-Anonimlik		✓	
(ε,m)-Anonimlik		✓	
t-Yakınlık		✓	
δ-Mevcudiyet			✓
(c, t)-İzolasyonu	✓		
ε –Diferansiyel Mahremiyet			✓
(d,γ)-Mahremiyet			✓
Dağıtık Mahremiyet			✓

İfşa saldırılarını önlemek amacıyla geliştirilen tüm modellerin amacı yayınlanmış verilerle diğer kaynaklardan elde edilen arka plan bilgilerinin saldırganlar tarafından eşleştirilmesini engellemeye çalışır.

2.7 Bilgi Metrikleri

Veri yayıncıları yalnızca verinin korunarak yayınlanmasından değil yayınlanan verinin alıcılarına sağladığı fayda ve kaliteden de sorumludur. Mahremiyet koruyucu yaklaşımlar anonimleştirme yöntemleriyle verileri kimliksizleştirir. Anonimleştirme işleminden sonra veride meydana gelen kayıpların miktarı veriden sağlanan faydayı ölçmek açısından önemlidir [110]. Veri kaybının ölçümü veri alıcıları ile veri yayıncıları tarafından farklı açılardan önemlidir. Veri alıcıları açısından yapacakların analizin doğruluğunu ve başarısını etkilerken, veri yayıncıları için, mahremiyet probleminin çözümündeki en uygun dengenin bulunabilmesi açısından önemlidir [111]. Veri faydasını ölçmek için anonimleştirilmiş tablo ile orijinal tablo arasındaki benzerlikten faydalanarak veri kaybının ölçülmesini sağlayan metriklere ihtiyaç duyulur [98, 112]. Kayıp ölçen metrik değerinin yüksek olması bilgi kaybının fazla olduğunu gösterir. Bilgi kaybının ölçülmesi için literatürde kullanılan genel ve özel metrikler takip eden alt bölümlerde özetlenmiştir.

2.7.1 Bozulma Metriği

Genelleştirilmiş veya baskılanmış her bir öznitelik değerine bağımsız olarak ceza puanı verilmesiyle bilgi kaybını hesaplayan genel bir metriktir [43, 66]. Orijinal ve anonimleştirilmiş veri arasındaki benzerliğe göre ölçüm yapan bu metrikte, uygulanan ceza puanlarının toplamı ilgili öznitelik için anonimleştirmeye bağlı bilgi kaybının ölçülmesini sağlar [113].

2.7.2 Yükseklik Metriği

Yükseklik metriği (YM), genelleştirme hiyerarşisindeki yüksekliğe göre bilgi kaybının hesabını yapan genel bir metriktir [83]. Genelleştirme yönteminin tek başına kullanıldığı durumlarda Samarati tarafından önerilmiştir [66]. Meyerson ve Williams ise genelleştirmeye ek olarak baskılanan öznitelik değerlerine ceza puanı verilmesini önererek bilgi kaybının ölçülmesini önermişlerdir [80]. Samarati tarafından önerilen metrikte, $A_d \in \{A_1,..,A_d\}$;

QID= { A_1,..,A_d } ve Y(A_i) A_i özniteliğinin yüksekliğini göstermek üzere YM hesaplaması, Eşitlik 2-17'de verilmiştir.

$$M_{YM}\ (A_1, .., A_d) = \sum_{i=1}^{d} Y(A_i) \qquad 2\text{-}17$$

YM en yüksek 1 (orijinal tablo) ile en düşük 0 (tamamen baskılanmış) değerlerini alabilir. YM metriğine göre genelleştirmesi yüksek olan öznitelikler kısa olan özniteliklere göre daha doğru sonuçlar verir. YM metriğinin hassasiyet problemi olarak da tanımlanan bu durum aynı tablo içerisinde farklı yüksekliklere sahip özniteliklerin aynı seviyede genelleştirilmelerine rağmen farklı oranda bilgi kaybı oluşur. Farklı yükseklikteki özniteliklerin bilgi kayıplarının aynı doğrulukta ölçülememesi bu metriğin zayıf yönüdür. Şekil 2-4'de genelleştirme hiyerarşisi verilen Cinsiyet ve Ülke öznitelikleri hassasiyet problemine örnek olarak verilebilir. Hiyerarşi yüksekliği 2 olan Cinsiyet özniteliği ile hiyerarşi yüksekliği 4 olan Ülke özniteliği için birinci kademe genelleştirmelerde Cinsiyet özniteliği veri kayıpları sırasıyla %100 ve %25 olur.

2.7.3 Duyarlılık Metriği

YM metriğindeki hassasiyet problemini dikkate alan duyarlılık metriği (AEM), genelleştirme hiyerarşilerin farklı yüksekliklerini dikkate alarak bilgi kaybını hesaplar [43]. QID öznitelik sayısı d, kayıt sayısı m, A_{xy} (x,y) tablo hücresinin genelleştirme yüksekliği, m_{hi} maksimum genelleştirme yüksekliği olmak üzere, en yüksek hiyerarşi seviyesine göre diğer yüksekliklerin normalize edilmesini sağlayan AEM hesaplaması, Eşitlik 2-18'de verilmiştir.

$$M_{DM}\ (A_1, .., A_d) = 1 - \frac{\sum_{i=1}^{d}\ \sum_{j=1}^{m} \frac{A_{ij}}{mh_i}}{m.d} \qquad 2\text{-}18$$

AEM en yüksek 1 (orijinal tablo) en düşük 0 (tamamen baskılanmış) değerlerini alır. Baskılama olmadan sadece genelleştirme kullanılması durumunda AEM değerinin hesaplanması Eşitlik 2-19'da verilmiştir.

$$M_{DM*}\,(A_1,..,A_d) = 1 - \frac{\sum_{i=1}^{d} \frac{A_i}{mh_i}}{d} \qquad 2\text{-}19$$

Küresel kodlamanın kullanıldığı genelleştirme yöntemleri ile bilgi kaybını hesaplayan farklı çalışmalarda kullanılmıştır [114].

2.7.4 Kayıp Metriği

Genelleştirmeden kaynaklanan bilgi kaybını ölçmek için bir diğer genel metrik olan kayıp metriği (KM) Iyengar ve arkadaşları tarafından önerilmiştir [77]. A_i kategorik öznitelik, L_x ilgili düğüme ait yaprakların sayısı, L_{topi} A_i özniteliğinin genelleştirme hiyerarşisindeki toplam yaprak sayısını, L_x x-köklü alt ağacın yaprak sayısı, L_{topi} kolon i için genelleme hiyerarşisinde toplam yaprak sayısını göstermek üzere LM hesaplaması Eşitlik 2-20'de gösterilmiştir.

$$M_{KMC}\,(A_1,..,A_d) = 1 - \frac{\sum_{i=1}^{d} \sum_{j=1}^{m} \frac{L_{R_{ij}} - 1}{L_{Top_i} - 1}}{m.d} \qquad 2\text{-}20$$

KM en yüksek 1 (orijinal tablo) en düşük 0 (tamamen baskılanmış) değerlerini alır. U'_{ij} genelleştirilmiş U_{ij} orijinal tablodaki (i,j) hücresinin en yüksek aralığını, L'_{ij} genelleştirilmiş L_{ij} orijinal tablodaki (i,j) hücresinin en düşük aralığını göstermek üzere sayısal öznitelik değerleri için KM hesaplaması ise Eşitlik 2-21'de verilmiştir.

$$M_{KMN}\,(A_1,..,A_d) = 1 - \frac{\sum_{i=1}^{d} \sum_{j=1}^{m} \frac{U'_{ij} - L'_{ij}}{U_i - L_i}}{m.d} \qquad 2\text{-}21$$

2.7.5 Eşlenik Sınıflar Ortalaması

Öznitelik bazında uygulanan KM metriğinin eşlenik sınıflar düzeyinde uygulanması amacıyla LeFevre ve ark. [78] eşlenik sınıf büyüklüğünün ortalamasına göre (ESO) bilgi kaybını ölçen genel bir metrik önermişlerdir. Metriğin hesaplanmasında eşlenik sınıfların normalize edilmesine bağlı olarak ortalama grup büyüklüğü azalır. Buna bağlı olarak veri kalitesinde dolaylı bir iyileşme görülür. Anonimleştirilmiş tablodaki kayıtların sayısı R, eşlenik sınıf EC, p en küçük eşlenik sınıf boyutu olmak üzere ESO hesaplaması Eşitlik 2-22'de verilmiştir.

$$M_{ESO}\ (A_1, .., A_d, p) = \frac{R}{\sum_{i=1}^{d} EC(i)\,.\,p} \qquad 2\text{-}22$$

Metriği kullanan mahremiyet modeline göre p değeri değişir. Örneğin kullanılan model k-anonimlik ise p değeri k'ya eşit olarak seçilir. ESO sterilizasyon işlemlerinden sonra ortalama eşlenik sınıf sayısını bularak grup büyüklüklerini normalize eder. Eşlenik sınıfların boyutunda herhangi bir kısıtlama yoksa normalizasyon işlemi gözardı edilebilir [115].

2.7.6 Ayırt Edilebilirlik Metriği

Bayardo ve ark. [44] eşlenik sınıfların içinde yer alan toplam kayıt sayısına verilen ceza puanlamasıyla bilgi kaybını ölçen ayırt edilebilirlik metriğini (AEM) önermişlerdir. Eşlenik sınıf içindeki kayıt sayısı |EC|, toplam kayıt sayısı R, ve baskılanmış kayıtlardan oluşan eşlenik sınıf O olmak üzere AEM metriğiyle bilgi kaybının hesaplanması Eşitlik 2-23'de verilmiştir.

$$M_{AEM}\ (A_1, .., A_d) = \sum_{\forall EC | EC \notin O}^{d} (|EC|^2) + \sum_{\forall EC | EC \in O}^{d} (R.\,|EC|) \qquad 2\text{-}23$$

Eşitlik 2-23'de baskılanmayan kayıtları içeren eşlenik sınıflara kendi büyüklükleri kadar ceza verilirken, baskılanan kayıtları içeren eşlenik sınıflara toplam kayıt sayısı kadar ceza verilir. Veriler tamamen baskılandığında |EC| =R, d=R.R olur. Bu durumda metriğin en büyük değeri R^2 olarak bulunur. Veriler tamamen orijinal halindeyken metriğin en küçük değeri ise benzer işlemlerle hesaplandığında R olarak bulunur.

AEM, için en ideal durum mahremiyet gereksinimlerinin tam olarak sağlandığı durumdur. Örneğin k-Anonimlik modelinde tüm grup büyüklüklerinin k'ya eşit olması en ideal durum olup bu durumda k-Anonimlik için en ideal AEM değeri $R.k^2$ olacaktır. Bu metriğe göre eşlenik sınıf büyüklükleri ile veri kaybı ilişkili olduğundan eşlenik sınıflar içerisinde ne kadar çok fazla kayıt varsa veri kaybı o kadar fazla olacaktır.

2.7.7 Entropi Metriği

De Waal ve Willenborg, Kooiman ve ark. [116], bilgi kayıplarını ölçmede Shannon entropisinin kullanımını önermişlerdir. Entropi metriği (EM) Küresel kodlama yöntemiyle sadece genelleştirme işlemleri ile aonimleştirilen verilerin bilgi kaybının ölçümünde kullanılır. EM'nin matemiksel ifadeleri Eşitlik 2-24 ve 2-25'de verilmiştir.

$$M_{EM}\ (A_0,..,A_{d-1}) = \sum_{i=0}^{d-1}\sum_{j=0}^{m-1}\big(X(i,j,A_i)\big).\log_2 X\ (i,j,A_i) \qquad 2\text{-}24$$

$$\big(X(i,j,k)\big) = \frac{\sum_{l=0}^{m-1} I(i,l,0) = R(i,j,0)}{\sum_{l=0}^{m-1} I(i,l,k) = R(i,j,k)} \qquad 2\text{-}25$$

Eşitlikler'de verilen matematiksel gösterimde, QID öznitelik sayısı d, toplam kayıt sayısı m, $0 \geq i \geq n\text{-}1$ ve $0 \geq j \geq m\text{-}1$ olmak üzere R (i, j, A) ise i. satır, j. kolonda yer alan özniteliğin k. seviyeden genelleştirilmesini, I(x) gösterge fonksiyonu ise x doğru olduğunda 1, diğer durumlarda 0 değerini alır.

2.7.8 Sınıflandırma Yöntemi

Genelleştirilen veya baskılanan kayıtlara ceza puanı uygulayarak bilgi kaybının ölçülmesini sağlayan sınıflandırma yöntemi metriği (SYM) Iyengar tarafından önerilen sınıflandırmaya özel önerilen bir metriktir [77]. Genelleştirilmiş T tablosu için sınıflandırma metriğine göre ceza puanı kaydın baskılanmış veya genelleşmiş olması durumunda pen (r)=1 diğer durumlarda pen (r)=0 olarak hesaplanır. CM, değeri ne kadar düşük olursa bilgi kaybı o kadar az olacaktır. Kayıt bazında hesaplanan ceza puanlarının toplamının ortalaması alınarak yayınlanan tablonun cezası Eşitlik 2-26'ya göre hesaplanır.

$$M_{SYM} = \sum_{i=1}^{|R|} \left(\frac{pen(R_i)}{|R|} \right) \quad 2\text{-}26$$

2.7.9 Denge Metriği

Fung ve ark. veri faydası ve mahremiyet arasındaki dengeyi gözeten özel amaçlı arama metriğini bilgi kaybını ölçmek amacıyla önermişlerdir [33]. Denge metriği (DEM) anonimleştirme sürecinde mahremiyet ve veri faydasını dikkate alarak iki gereksinim arasındaki dengenin kurulmasına odaklanır. Anonimleştirmenin her aşamasında s özniteliği için bilgi kazancı IG(s) ve mahremiyet kaybı PL(s) hesaplanır ve s özniteliği için en uygun durum elde edilmeye çalışılır. Denge metriğine göre bilgi kaybının hesaplanması Eşitlik 2-27 'de verilmiştir.

$$M_{DEM} = \frac{IG(s)}{PL(s) + 1} \quad 2\text{-}27$$

IG ve PL seçimi mahremiyet modeline göre değişiklik gösterir. Örneğin, k-anonimlik için, Fung ve ark. PL (s)'yi, QID_j'de anonimliğin ortalama azalması ile Eşitlik 2-28'de gösterildiği şekilde ölçmüştür.

$$PL(s) = avg\{A(QID_j) - A_s(QID_j)\} \quad 2\text{-}28$$

Burada A (QID_j) ve A_s (QID_j) QID_j'in sırasıyla anonimleştirmeden önce ve sonraki halini gösterir. IL(g) bilgi kaybını, PG(g) mahremiyet kazancını göstermek üzere Veri faydası açısından en uygun genelleştirmenin (g) seçilmesi Eşitlik 2-29'da verilmiştir.

$$M_{IGPL} = \frac{IL(g)}{PG(g) + 1} \quad 2\text{-}29$$

2.7.10 KL-Sapma Metriği (KL-divergence-KLM)

KL uzaklığı, iki olasılık dağılımı arasındaki uzaklığın ölçülmesinde kullanılır [117]. Yayınlanacak verilerdeki öznitelik değerlerinin genel dağılımını dikkate alarak

anonimleştirilmiş verilerin bilgi kaybını ölçen KL-Uzaklık metriği (KLM) Kullback-Leibler tarafından önerilmiştir [117]. KL-Sapma metriğinin uygulanabilmesi amacıyla orijinal verilerdeki bir kaydın olasılık dağılımı p_1, anonimleştirme sonucunda elde edilen tablonun olasılık dağılımı p_2'ye dönüştürülür. Anonimleştirilmiş verileri bir olasılık dağılımına dönüştürmenin yolları Chen ve ark. [118]tarafından gösterilmiştir. KL-sapması p_1 ve p_2 için Eşitlik 2-30'da verilmiştir.

$$M_{KLM}\ (p_1, p_2) = \sum_r p_1(r) \log \frac{p_1(r)}{p_2(r)} \qquad 2\text{-}30$$

Bilgi kaybının ölçülmesinde anonimleştirilmiş değer ile orijinal tablo arasındaki uzaklığın ölçümünde kullanılır.

2.7.11 Kesinlik Ceza Metriği

Kesinlik cezası metriği Terrovitis ve ark. [73] tarafından eşlenik sınıfların tanımındaki doğruluk kaybını değerlendiren genel amaçlı bir metrik olarak önerilmiştir. Anonimleştirilmiş tablo T*, orijinal tablo T, $QID_1..QID_d$ yarı tanımlayıcı öznitelikler, r ϵ T, d=|QID|, $1 \leq i \leq d$ (x_i,y_i) yarı tanımlayıcı öznitelik üzerindeki bir aralık olmak üzere, QID üzerindeki r kaydı için normalize edilmiş kesinlik ceza puanının hesaplanması Eşitlik 2-31'de verilmiştir.

$$M_{KCP_{QIDi}}(r) = \frac{|\ y_i - x_i|}{|\ QID_i|} \qquad 2\text{-}31$$

$w_{i,\ öznitelik}$ ağırlığı olmak üzere, r kaydı üzerindeki NCP;

$$NCP(r) = \sum_{i=1}^{d} w_i\ .M_{KCP_{QIDi}}(r) \qquad 2\text{-}32$$

sonuç olarak T tablosu için M_{CPM} hesaplaması Eşitlik 2-33'de verilmiştir.

$$M_{CPM}(T) = \sum_{r \in T^*} NCP(r) \qquad 2\text{-}33$$

2.8 Kimlik İfşa Metrikleri

Bu bölümde mahremiyet tehditlerinin başında gelen kimlik ifşasına yönelik risklerin ölçülmesini sağlayan risk metriklerine yer verilmiştir. Veri faydası ile mahremiyet risklerinin dengelenmesinde veri kaybının ölçülmesinin yanında mahremiyet risklerinin de ölçülmesi gerekir. Bu ölçümler veri yayıncılarını mahremiyet riskleri ile veri faydası konusunda bilgilendirerek gerekli önlemleri almasını sağlar. Mahremiyet risklerinin ölçülmesine yönelik metriklerin yanı sıra bu metrik değerlerinin yorumlanmasında kullanılacak karar kurallarına da ihtiyaç vardır [102, 119, 120]. Takip eden alt bölümlerde, kimlik ifşaları için yeniden tanımlama olasılığının ölçülmesi ve yorumlanması için türetilmiş metrikler ile karar kuralları, bu metriklerin uygulandığı basit metrikler ve teklik metriği sunulmuştur.

2.8.1 Türetilmiş Risk Metrikleri

Arka plan bilgileri ile kamuya açık veri tabanlarını dikkate alarak kimliksiz olarak yayınlanmış veri kümesindeki bir kaydın kimliğinin ifşa edilmesine ait risklerin hesaplanmasında türetilmiş risk metrikleri kullanılır [119]. Risk metriklerinin gösteriminde θ fonksiyonu kullanılır. $1 \leq i \leq n$ olmak üzere, i kaydının ifşa edilme ihtimali θ_i, j eşlenik sınıfının ifşa edilme ihtimali ise θ_j fonksiyonları ile gösterilir. Tek bir eşlenik sınıfa uygulanan θ_j risk metriğinin verilerin tamamına uygulaması için 0-1 arasında değer alacak şekilde normalize edilmesi gerekir. Normalizasyon sonucunda olasılığın yüksek olup olmadığına karar vermek için metrik değerini yorumlayan karar verme kurallarına ihtiyaç duyulur. Karar verme kuralıyla, metrik ölçümleri yeniden kimliklendirme riskine dönüştürülür ve hesaplanan riskler belirlenen eşik değerlere göre yorumlanır. Eşik değerler veri yayıncılarının daha önceki deneyimleri veya veri alıcılarının beklentilerine göre belirlenir. I (x) gösterge fonksiyonu olup (x doğruysa 1, değilse 0 döndürür), f_j, eşlenik sınıf j'nin büyüklüğü olmak üzere, eşik değer τ' ya göre olasılığı yüksek olan kayıtların kimlik ifşa riskinin hesaplanması Eşitlik 2-34'de verilmiştir.

$$R_a = \frac{1}{n} \sum_{j \in J} f_j \, . \, I(\theta_j > \tau) \qquad 2\text{-}34$$

Kimlik ifşa riski yüksek olan kayıtların eşik değeri α ile gösterilirse R_a için karar kuralı eşitlik 2-35'de verilmiştir.

$$D_a = \begin{cases} YÜKSEK & R_a > \alpha \\ DÜŞÜK & R_a \leq \alpha \end{cases} \quad 2\text{-}35$$

En kötü durum senaryosunda ise kayıtlar içerisinde en yüksek kimlik ifşa olasılığına sahip kayıtların tüm kayıtların ifşa olasılığı olarak alındığı türetilmiş metrik hesaplaması Eşitlik 2-36'da verilmiştir.

$$R_b = max_{j \in J}(\theta_j) \quad 2\text{-}36$$

Doğrudan yüksek riskli kayıtların hedef alınabileceği durumlar için geçerli olan R_b metriği için karar kuralı Eşitlik 2-37 'de verilmiştir.

$$D_b = \begin{cases} YÜKSEK, R_b > \tau \\ DÜŞÜK & R_b \leq \tau \end{cases} \quad 2\text{-}37$$

Veri kümesindeki tüm kayıtların ortalama olasılığını alarak ifşa riskleri için beklenen değeri veren bir diğer türetilmiş metrik olan R_c hesaplaması Eşitlik 2-38'de verilmiştir.

$$R_c = \frac{1}{n} \sum_{j \in J} f_j \,.\, \theta_j \quad 2\text{-}38$$

R_c metriğinin karar kuralı Eşitlik 2-39'da verilmiştir.

$$D_c = \begin{cases} YÜKSEK, R_c > \lambda \\ DÜŞÜK & R_c \leq \lambda \end{cases} \quad 2\text{-}39$$

R_c metriği, ortalama bir risk hesaplaması yaparak yeniden kimliklendirilecek kayıtların oranını verir. İlk bakışta, R_a metriğine benzer gibi görünsede R_a metriğinde saldırganın sadece yüksek riskli kayıtları yeniden tanımlayacağı varsayılırken, R_c metriği ise tüm saldırganın tüm kayıtları yeniden tanımlamaya çalışacağını varsaymaktadır.

Türetilmiş risk metriklerinin birlikte gösterimi ve yorumları Çizelge 2-11'de verilmiştir.

Çizelge 2-11 Türetilmiş risk metrikleri yorumları

Türetilmiş Risk Metrikleri	Yorumlar
$R_a = \frac{1}{n}\sum_{j\in J} f_j \,.\, I(\theta_j > \tau$	Belirlenen eşiğin üzerinde olan kayıtlar için uygulanır.
$R_b = max_{j\in J}(\theta_j)$	En yüksek risk tüm kayıtlar için seçilir.
$R_c = \frac{1}{n}\sum_{j\in J} f_j \,.\, \theta_j$	Tüm kayıtla içerisinde doğru olarak yeniden kimliklendirilen kayıtların ortalaması için seçilir.

Türetilmiş risk metriklerinin karar kurallarının birlikte gösterimi ve yorumları Çizelge 2-12'de verilmiştir.

Çizelge 2-12 Karar kuralları yorumları

Karar Kuralları	Açıklama
$D_a = \begin{cases} YÜKSEK & R_a > \alpha \\ DÜŞÜK & R_a \leq \alpha \end{cases}$	Yeniden kimliklendirme olasılığı yüksek kayıtların kabul edilebilir/kabul edilemez oranı
$D_b = \begin{cases} YÜKSEK, R_b > \tau \\ DÜŞÜK & R_b \leq \tau \end{cases}$	Tüm kayıtlar için yeniden kimliklendirmenin kabul edilebilir/kabul edilemez ihtimali
$D_c = \begin{cases} YÜKSEK, R_c > \lambda \\ DÜŞÜK & R_c \leq \lambda \end{cases}$	Yeniden kimliklendirilebilir kayıtların ortalama kabul edilebilir/kabul edilemez oranı

Türetilmiş risk metriklerinin karar kurallarında kullandığı eşik değerlerin birlikte gösterimi ve yorumları Çizelge 2-13'de verilmiştir.

Çizelge 2-13 Eşik değer yorumları

Eşik Değeri	Yorumlama
τ	Tek kaydın doğru şekilde en yüksek yeniden kimliklendirme olasılığı
α	Yeniden kimliklendirme olasılığının yüksek olduğu kayıtların oranı
λ	Yeniden kimliklendirilebilen kayıtların ortalama oranı

2.8.2 Basit Risk Metrikleri

Saldırgan kimlik ifşası için seçmiş olduğu kurbanı hakkında farklı seviyelerde bilgi sahibi olabilir. Kurban tanıdık veya hakkında hiçbir şey bilinmeyen rastgele seçilen birisi olabilir. Basit risk metrikleri saldırganın kurban hakkında edindiği bilgilerle ilgili varsayımlar altında türetilmiş risk metriklerini kullanarak kimlik ifşa risklerini hesaplar. Saldırganın kurbanı hakkındaki bilgi seviyesine göre kimlik ifşa risklerini hesaplayan basit risk metrikleri savcı, gazeteci ve pazarlamacı olmak üzere üç kategoride incelenir [119].

Saldırgan kurbanın yayınlanan veri içerisinde yer alıp almadığı yani üyelik bilgisine sahipse kimlik ifşa riskinin hesaplanmasında savcı yaklaşımı uygulanır. Yayınlanan veri toplumun genelini veya büyük bir bölümünü temsil ediyorsa, kurban yayınlanan veri içerisinde yer alıyor veya kendisi beyan ediyorsa, saldırgan sahip olduğu arka plan bilgileriyle kurbanın yayınlanan veri içerisinde yer aldığını biliyorsa kimlik ifşa riskinin hesaplanmasında savcı yaklaşımı kullanılır.

f_j eşlenik sınıf boyutu, ${}_p\theta_j$ ifadesi savcı riskini göstermek üzere ifşa riskinin hesaplanması, Eşitlik 2-40'da verilmiştir [121].

$$ {}_p\theta_j = \frac{1}{f_j} max_{j \in J}(\theta_j) \qquad 2\text{-}40 $$

Yayınlanan veri kümesi savcı yaklaşımı gereksinimlerini karşılamıyorsa veri yayıncı savcı riski yerine gazeteci yaklaşımı ile kimlik ifşa riskini hesaplar. Gazeteci riski için, saldırganın yayınlanan veri kümesinin alt kümesi olduğu varsayılan tanımlama veritabanına erişimi olduğunu ve kimliksiz verilerle eşleştirdiğini varsayıyoruz.

Gazeteci yaklaşımında tanımlama veritabanındaki eşlenik sınıfların büyüklüğü |K|, J ⊆ K olmak üzere J = {x | ∀x: x ∈ K ∧ fx> 0}. Eşlenik sınıf içindeki kayıtların sayısı j ve j ∈ K f_j > 0 olmak üzere tanımlama veritabanındaki kayıtların toplam sayısı F_j ile gösterildiğinde gazeteci riski ${}_{jo}\theta_j$'nın hesaplanması Eşitlik 2-41'de gösterilmiştir [119].

$$ {}_{jo}\theta_j = \frac{1}{F_j} \qquad 2\text{-}41 $$

Saldırgan yayınlanmış veri kümesi içerisinde herhangi bir kayıt yerine mümkün olan en yüksek sayıda kaydın yeniden kimliklendirilmesiyle ilgileniyor ve kurbanları rastgele bir yöntemle seçtiği durumlarda ifşa riskinin hesaplanmasında pazarlamacı yaklaşımı kullanılır. Savcı ve gazeteci yaklaşımları bireysel risk hesaplaması yaparken, pazarlamacı yaklaşımı daha çok toplulukla ilgilenir.

Pazarlamacı riskinin hesaplanmasında, yayınlanacak veritabanında eşlenik sınıf j, eşlenik sınıf büyüklüğü f_j, tanımlama veritabanında eşlenik sınıf J, eşlenik sınıf büyüklüğü F_J olmak üzere pazarlamacı riski ${}_m\theta_j$ hesaplanması Eşitlik 2-42'de verilmiştir [122].

$$ {}_m\theta_j = \frac{f_j}{F_j} \qquad 2\text{-}42 $$

İfşa risklerinin hesaplanmasında veri yayıncı basit metrik yaklaşımını seçtikten sonra türetilmiş metrik türünü seçerek ifşa risklerini hesaplar. İfşa risklerinin hesaplanmasında kullanılan metrikler ve yorumları Çizelge 2-14'de verilmiştir.

Çizelge 2-14 Risk metrikleri hesaplamaları

Basit Risk	Türetilmiş Risk Metrikleri	Yorum
Savcı	$_pR_a = \frac{1}{n}\sum_{j\in J} f_j \,.\, I(\frac{1}{f_j} > \tau)$	f_j, yayınlanan verideki eşlenik sınıf büyüklüğüdür. Yayınlanan veri tüm popülasyonla aynı ise, ($f_j = F_j$) olduğu durumlar için uygundur.
	$_pR_b = \frac{1}{min_{j\in J} f_j}$	
	$_pR_c = \frac{\lvert J\rvert}{n}$	
Gazeteci	$_{jo}R_a = \frac{1}{n}\sum_{j\in J} f_j \,.\, I(\frac{1}{F_j} > \tau)$	Bu metrikler, yayınlanan verinin tanımlama veritabanının bir alt kümesi olduğu durumlar için uygundur.
	$_{jo}R_b = \frac{1}{min_{j\in J} F_j}$	
	$_{jo}R_c = \left\{ \frac{\lvert J\rvert}{\sum_{j\in J} F_j} ; \frac{1}{n}\sum_{j\in J} \frac{f_j}{F_j} \right\}$	
Pazarlamacı	$_mR_1 = \frac{\lvert J\rvert}{N}$	Bu metrik, iki veritabanındaki tüm kayıtların (n = N) eşleştirilebildiği durumlar için uygundur;
	$_mR_2 = \frac{1}{n}\sum_{j\in J} \frac{f_j}{F_j}$	Bu metrik, n<N olduğu ve açıklanan veri kümesinin, tanımlama veritabanının ($J \subseteq K$) uygun bir alt kümesi olduğu durum için uygundur.

2.8.3 Biriciklik Metriği

Eşlenik sınıflar yerine kayıtların bulunduğu durumlarda biriciklik metriği basit metriklerle birlikte kullanılarak kimlik ifşa riskinin ölçülmesinde kullanılır [123]. Savcı yaklaşımıyla biriciklik metriğinin hesaplanmasında, yayınlanan veri kümesindeki benzersiz kayıtların oranı eşitlik 2-43'de gösterildiği şekilde hesaplanır [124].

$$U_1 = \frac{1}{n}\sum_{j\in J} I(f_j = 1) \qquad 2\text{-}43$$

U_1 metriğinin karar kuralı ise Eşitlik 2-44'de verilmiştir.

$$D_1 = \begin{cases} YÜKSEK, U_1 > \chi \\ DÜŞÜK \quad U_1 \leq \chi \end{cases} \qquad 2\text{-}44$$

Gazeteci yaklaşımının uygulandığı durumlarda farklı biriciklik ölçümleri önerilmiştir [101, 125]. Gazeteci riski altında, yayınlanan veri kümesinde tek olan bir kayıt, tanımlama veritabanında tek olmayabilir. Bu nedenle, önerilen bu metriklerden birisi tek bir kaydın tanımlama veritabanında da tek olma ihtimalini dikkate alarak Eşitlik 2-45'de gösterildiği şekilde hesaplanır.

$$U_2 = \frac{\sum_{j \in J} I(f_j = 1, F_j = 1)}{\sum_{j \in J} I(f_j = 1)} \qquad 2\text{-}45$$

U_2 metriğinin karar kuralı ise Eşitlik 2-46'da verilmiştir.

$$D_2 = \begin{cases} YÜKSEK, U_2 > \omega \\ DÜŞÜK \quad U_2 \leq \omega \end{cases} \qquad 2\text{-}46$$

2.9 Bölüm Sonucu

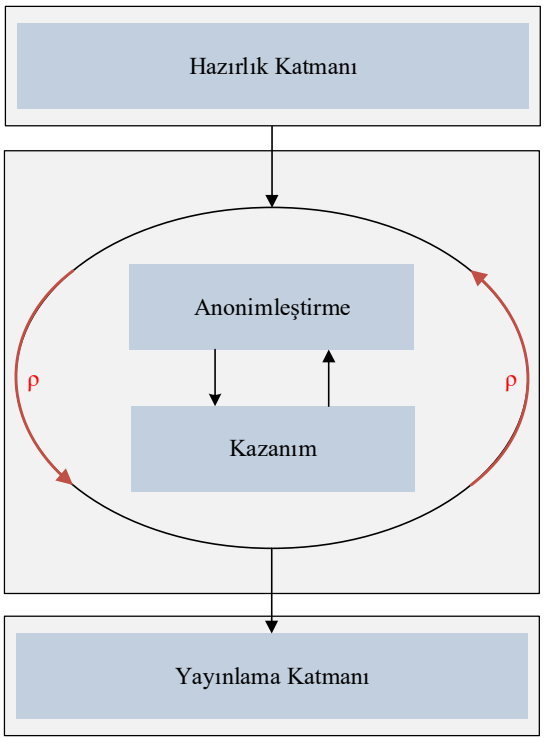
Bu bölümde mahremiyet koruma süreci açıklanmış, bu süreçte yer alan taraflar ile sorumluluklarından bahsedilerek veri mahremiyeti problemi hakkında bilgi verilmiştir. Veri mahremiyeti probleminin çözümünde veriden sağlanan fayda ile mahremiyet koruma arasındaki ilişki açıklanmıştır. Mahremiyet korumalı veri paylaşımında veri yayınlamanın ekonomik ve etkin bir çözüm olduğu vurgulanarak, veri paylaşımında yaygın olarak kullanılan mikro veri modelinin yapısı anlatılmıştır. Mikro veri modeli içerisinde yer alan özniteliklere ait sınıflandırma hakkında bilgi verilerek, mahremiyet korumalı yaklaşımlarda kullanılan modeller incelenmiştir. Mahremiyet modellerinin uygulanmasında kullanılan anonimleştirme yöntemlerinden bahsedilerek mahremiyet tehditleri hakkında bilgi verilmiştir. Veri faydası ve mahremiyet koruma dengesinin sağlanmasında ihtiyaç duyulan ölçümlerin yapılmasını sağlayan bilgi ve risk metrikleri açıklanmıştır.

Bu bölümde yapılan inceleme ve araştırmalar sonucunda tez çalışmasında önerilen modelin gerçekleştiriminin nasıl ve hangi yöntemlerle yapılacağı konusu netleştirilmiş takip eden bölümde materyal ve yöntemler başlığı adı altında anlatılmıştır.

3 MATERYAL ve YÖNTEM

Tez çalışması kapsamında mahremiyetten ödün vermeden veri faydasını arttırıcı ρ-Kazanım adı verilen yeni bir model önerilmiştir. Bu bölümde model gerçekleştiriminin nasıl ve hangi yöntemlerle yapılacağı konusu katmanlara ayrılarak anlatılmıştır. Önerilen model hazırlık, kazanımsal anonimleştirme ve yayınlama olmak üzere üç katmandan oluşmaktadır.

Önerilen modelin blok diyagramı Şekil 3-1'de gösterilmiştir.



Şekil 3-1 ρ-Kazanım blok diyagramı

Şekil 3-1'de verilen blok diyagramına göre hazırlık katmanında yayınlanacak tablo bir sonraki aşama için hazırlanır. Bu katmanda, anonimleştirme aracı, öznitelik sınıflandırılması, koruma modeli, anonimleştirme teknikleri, tehditlere göre mahremiyet model gereksinimleri, bilgi metrikleri, risk metrikleri fayda dikkate alınarak seçilir. Bir sonraki aşama olan kazanımsal anonimleştirme katmanında, hazırlanan tablo veri faydasının arttırılması amacıyla kazanımsal anonimleştirme sürecine girer. Yayınlama aşamasında ise önerilen model ile veri faydası arttırılmış anonim kayıtlar veri alıcıları için yayınlanır. Takip eden alt bölümlerde, önerilen modeli gerçekleştirebilmek için kullanılan materyal ve yöntemler katmanlara ayrılarak anlatılmıştır.

3.1 Hazırlık Katmanı

Önerilen fayda temelli mahremiyet koruma modelinin temelinde, mahremiyetten ödün vermeden veri faydasının mümkün olduğunca arttırılmasına yönelik işlemler yapılır. Veri faydasının arttırılmasının temelinde en az seviyede bilgi kaybıyla anonimleştirmenin yapılması yer alır. Ancak anonimleştirme seviyesindeki azalma mahremiyet korunma seviyesini olumsuz etkileyeceği için farklı yöntemlere ihtiyaç vardır. Bu çalışmada anonimleştirmenin seviyesinin azaltılması yerine verimliliğin arttırılması üzerinde çalışılmıştır. Bu katmanda kazanımsal anonimleştirme öncesinde ihtiyaç duyulan çalışmalar takip eden alt bölümlerde açıklanmıştır.

3.1.1 Mahremiyet Araçları

Veri yayıncılar verilerin anonimleştirilmesinde hızlı ve kolay bir yöntemle anonimleştirme yapan mahremiyet araçlarından faydalanır. ρ-Kazanım modelinin gerçekleştirilebilmesi amacıyla öncelikle mahremiyet araçları üzerinde araştırmalar yapılmıştır. Mahremiyet araçları, mahremiyet modellerini ve anonimleştirme tekniklerini kullanarak veri yayınlama konusunda veri yayıncılara veya veri sahiplerine yardımcı olmak üzere geliştirilen açık kaynak veya ticari yazılımlardır. Tez kapsamında seçilecek mahremiyet aracıyla, mahremiyet modellerinin pratikte çalışma yöntemlerinin anlaşılabilmesi, mevcut modellerin karşılaştırılarak zayıf veya güçlü yönlerinin ortaya konulması, tez kapsamında önerilen modelin gerçekleştiriminin yapılması ve deneysel çalışmalarla sonuçlarının yorumlanarak etkinliğinin gösterilmesi konularında faydalanılmıştır. Veri anonimleştirmede yaygın olarak kullanılan araçlar takip eden alt bölümlerde özetlenmiştir.

3.1.1.1 CAT

CAT (Cornell Anonymization Toolkit) mahremiyet aracı, Cornell Üniversitesi tarafından C++ programlama dilinde geliştirilmiş Windows işletim sisteminde çalışan ticari olmayan bir yazılımdır. CAT, ℓ–Çeşitlilik ve t-Yakınlık modelleri, Incognito [75] arama algoritması, küresel kodlama ile manuel kayıt bastırma tekniklerini destekler. Veri kümelerinin anonimleştirilebilmesi için metin dosyalarına ve genelleştirme hiyerarşilerinin manuel olarak yazılıma yüklenmesi gerekir [126]. Çok fazla manuel işlem olması ve mahremiyet model desteğinin yeterli olmamasından dolayı tercih edilmemiştir.

3.1.1.2 TIAMAT

TIAMAT (A Tool for Interactive Analysis of Microdata Anonymization Techniques), Java programlama dilinde geliştirilmiş açık kaynak olmayan anonimleştirme yazılımıdır. TIAMAT, k-Anonimlik modelini, Mondrian [78] arama algoritmasını ve küresel yeniden kodlamayı destekler. Genelleştirme hiyerarşileri için basit bir grafik editörüne sahip olan TIAMAT risk değerlendirme yeteneğine sahip değildir. SQL dilini destekleyen veri tabanlarıyla otomatik olarak haberleşerek veri işlemlerini yapabilmekte ve istatistiki dağılımları veri yayıncıya rapor olarak sunabilmektedir [127]. Açık kaynak olmaması ve risk değerlendirme yapamaması gibi zayıflıklarından dolayı tez çalışmasında kullanılmamıştır.

3.1.1.3 ARX

İlişkisel veri kümeleri üzerinde anonimleştirme işlemleri yapan görsel ve kapsamlı bir araç olan ARX k-Anonimlik, ℓ-Çeşitlilik, t-Yakınlık ve δ-Mevcudiyet modellerini destekleyen açık kaynak veri anonimleştirme aracıdır. Veri faydası ve mahremiyet açısından risk analizleri ve kullanıcı geri bildirimlerini dikkate alarak etkin bir anonimleştirme yapılmasını sağlar. ARX veri yayıncının ihtiyaçları doğrultusunda genelleştirme yapılmasını sağlayan kapsamlı bir genelleştirme editörüne sahiptir. Veri faydasının ölçülmesi amacıyla birçok bilgi kaybı ve risk metriğini destekleyerek ölçüm sonuçlarını grafiksel olarak kullanıcıya sunar. ARX'in diğer önemli özelliği ise API desteğiyle entegrasyon konusunda veri paylaşımcılara sunduğu kolaylık ve esnekliktir [128]. Desteklediği modeller, risk değerlendirme yeteneğinin olması, etkileşimli kullanıcı arayüzü, hiyerarşi editörü özelliklerinden dolayı çalışma kapsamında mahremiyet aracı olarak ARX yazılımı tercih edilmiştir.

3.1.1.4 SECRETA

SECRETA (A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms), ilişkisel (R), işlemsel (T) ve R-T (ilişkisel-işlemsel) veri kümeleri üzerinde anonimleştirme işlemlerini karşılaştırmalı olarak yapan açık kaynak olmayan C ++ dilinde geliştirilmiş görsel bir anonimleştirme aracıdır. SECRETA ilişkisel veriler için yukarıdan aşağıya, aşağıdan yukarıya, Incognito arama algoritmaları, küresel ve yerel yeniden kodlamayı destekler. Risk değerlendirme kabiliyeti olmayan SECRETA, kullanıcının çözüm alanını görselleştirmesine ve göz atmasına izin vermez [129]. Risk

değerlendirme yeteneğinin olmaması ve açık kaynak koda sahip olmamasından dolayı tez çalışmalarında tercih edilmemiştir.

3.1.1.5 UTD

Dallas Üniversitesi (UTD) tarafından geliştirilen, platform bağımsız Java dilinde geliştirilmiş açık kaynak veri anonimleştirme yazılımıdır. UTD yazılımı k-Anonimlik, ℓ-Çeşitlilik ve t-Yakınlık modelleri ile Datafly [130] ve Incognito [75] arama algoritmaları ile Mondrian [78] ve anatomi [24] yeniden kodlama tekniklerini destekler. Büyük veri kümeleriyle ölçeklenebilirlik sorunları olan UTD, grafiksel bir arabirime sahip değildir. XML dosyası aracılığıyla yapılandırılabilen UTD, risk analizleri veya riske dayalı anonimleştirme yöntemlerini uygulayamamaktadır [131]. Grafiksel ara birimi olmaması, risk değerlendirme yeteneğinin olmaması gibi zayıflıklarından dolayı tez çalışması kapsamında tercih edilmemiştir.

3.1.1.6 Mahremiyet Araç Seçimi

Önceki bölümlerde özetlenen mahremiyet araçları Çizelge 3-1'de farklı parametreler üzerinden birbirleriyle karşılaştırılmıştır.

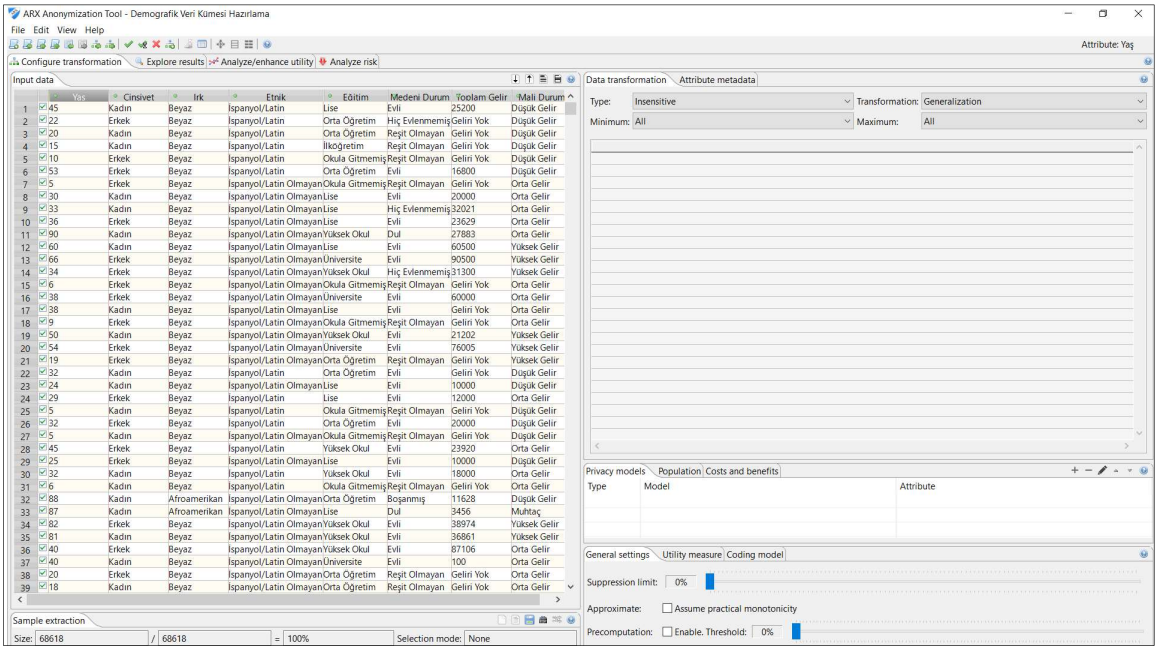
Çizelge 3-1 Mahremiyet araçlarının karşılaştırılması

	UTD	**CAT**	**TIAMAT**	**SECRETA**	**ARX**
Model	k, ℓ, t	ℓ, t	k	-	k, ℓ, t, δ
Genelleştirme	Var	Var	Hayır	Var	Var
Baskılama	Kısmi	Yok	Yok	Yok	Var
Risk Ölçme	Yok	Sınırlı	Yok	Yok	Var
Bilgi Ölçme	Yok	Yok	Yok	Yok	Var
Veri Formatı	CSV	Manuel	CSV	CSV	CSV, Excel, DBMS
Hiyerarşi	Yok	Yok	Kısmi	Var	Var
Çözüm Uzayı	Yok	Yok	Yok	Yok	Var
GUI	Yok	Var	Yok	Var	Var
Dil	Java	C++	Java	C++	Java

Çizelge 3-1'de ticari olmayan mahremiyet araçları karşılaştırılmıştır. Desteklediği modeller, arama uzayı oluşturması, risk ve fayda ölçümü, genelleştirme hiyererşi editörü, küresel, yerel kodlama, mikroagregasyon yöntemlerini uygulayabilmesi açısından avantajları olan Çizelge 3-1'de öne çıkan özellikleri verilen ARX mahremiyet aracı tercih edilmiştir. ARX 3.5.1 sürümü 64-bit Oracle JVM çalıştıran, dört çekirdekli 2,6 GHz Intel Core i7 CPU özelliklerine sahip masaüstü bilgisayarda çalıştırılmıştır.

3.1.2 Veri Kümeleri

Tez çalışması kapsamında literatürde veri anonimleştirme konusundaki birçok çalışmada kullanılmış gerçek dünya verilerini içeren 1994 ABD nüfus sayımı örneği ADULT, 1998 KDD CUP ve DEMOGRAFİK veri kümeleri kullanılmıştır [132, 133]. Özellikle ADULT veri kümesi, birçok farklı çalışmada kullanıldığı için mahremiyet modellerinin değerlendirilmesinde kullanılan de-facto bir standart haline gelmiştir. Çalışma kapsamına alınan veri kümelerinde kayıp değerler ve benzer öznitelikler çıkarılarak sadeleştirmeler yapılmıştır. Yapılan sadeleştirmeler sonucunda veri kümeleri deneysel çalışmalar için kullanıma hazır hale getirilmiştir. Çalışma kapsamında kullanılan DEMOGRAFİK veri kümesinin ARX ortamına aktarılmış hali Şekil 3-2'de gösterilmiştir.



Şekil 3-2 DEMOGRAFİK veri kümesi düzenlemesi

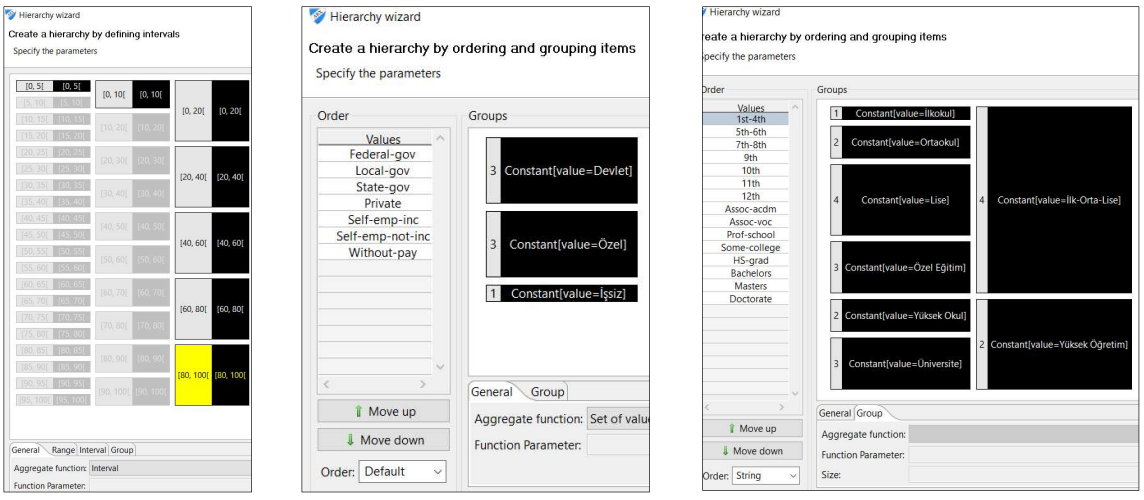
Veri kümeleri ARX ortamında kullanıma hazır hale getirildikten sonra öznitelikler sınıflandırılarak genelleştirme hiyerarşileri oluşturulmuştur. Yarı-tanımlayıcı öznitelikleri

seçerken, kimlik ifşasına yardımcı olacak veri bağlama riskleri mümkün olan öznitelikler seçilmiştir. Ayrıca, her bir veri kümesi için hassas öznitelikler seçilmiştir. Veri kümelerinde yer alan öznitelikler sınıflandırılarak genelleştirme seviyeleri ile birlikte Çizelge 3-2'de gösterilmiştir

Çizelge 3-2 Öznitelik sınıflandırma

Veri Kümesi Adı	QID (Genelleştirme Yüksekliği)	SA
T_{ADULT}	cinsiyet (2), yaş (5), ırk (2), medeni_durum(3), eğitim (4), ülke (3), pozisyon (3), maaş (2)	pozisyon
$T_{DEMOGRAFİK}$	yaş (5), cinsiyet (2), ırk (2), etnik (2), eğitim (3), medeni_durum (2), toplam gelir (3)	etnik
T_{CUP}	posta_kodu (6), yaş(5), cinsiyet (2)	gelir

Çizelge 3-2'de verilen veri kümelerinde yer alan özniteliklere ait hiyerarşi yükseklikleri, 2-6 arasında değişmektedir. Veri kümeleri öznitelik sınıflandırılması yapıldıktan sonra öznitelik tiplerine göre genelleştirme hiyerarşileri hazırlanır. Örnek olarak ADULT veri kümesinde yer alan farklı tiplerdeki { Yaş, İş, Eğitim } birleşik tanımlayıcı özniteliklerine ait genelleştirme hiyerarşi örnekleri Şekil 3-3'de verilmiştir.

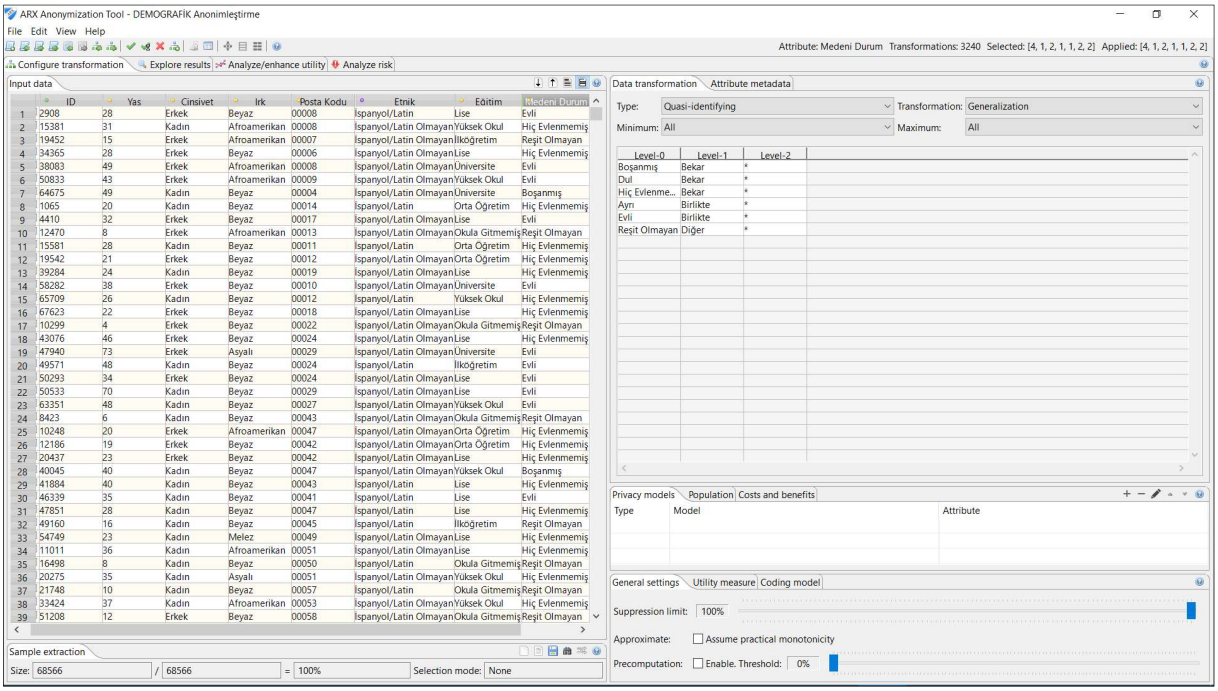


(a) Yaş genelleştirme (b) İş genelleştirme (c) Eğitim genelleştirme

Şekil 3-3 QID genelleştirme hiyerarşi örnekleri

Veri kümeleri için genelleştirme hiyerarşileri hazırlandıktan sonra veri kümeleri anonimleştirme için hazır hale gelir. Veri faydasının iyileştirilmesi açısından dönüşüm

işlemlerinde genelleştirme hiyerarşilerinin hangi seviyelerde kullanılacağı konusunda kısıtlamalar konularak çözüm uzayı küçültülebilir. DEMOGRAFİK veri kümesi Şekil 3-4'de gösterildiği üzere dönüşümler üzerinde herhangi bir kısıtlamaya gidilmeden anonimleştirilmeye hazır hale getirilmiştir.



Şekil 3-4 Düzenlenmiş DEMOGRAFİK veri kümesi

3.1.3 Arama Uzayı

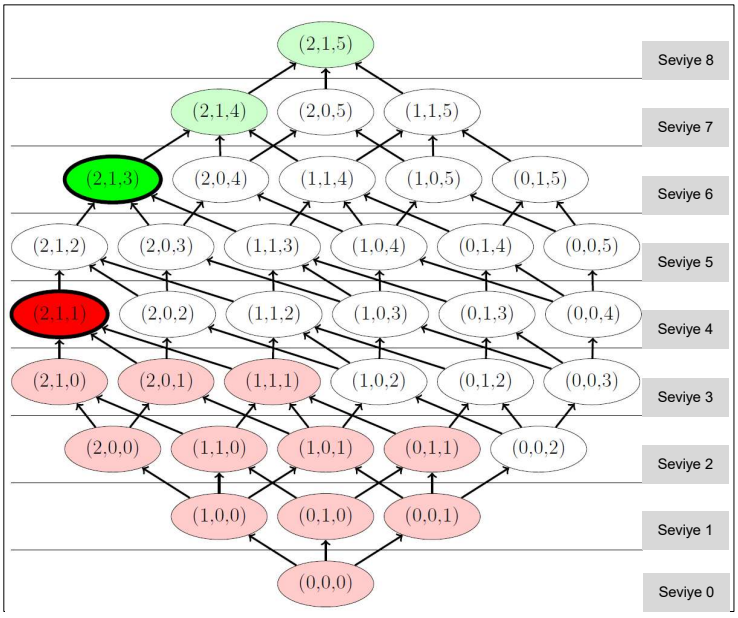
Tez kapsamında yapılan çalışmalarda mahremiyet gereksinimlerinin sağlanması için anonimleştirme tekniklerinden veri faydasını olumlu etkileyen genelleştirme ve baskılama yöntemi kullanılmıştır. Dönüşüm işlemlerini yaparak arama uzayının oluşturulmasında küresel kodlama yapısını kullanan tüm alan genelleştirilmesi seçilmiştir. Arama uzayının büyüklüğü özniteliklerin genelleştirme hiyerarşisine bağlıdır. Yarı tanımlayıcı özniteliklerin genelleştirme hiyerarşilerinin yüksekliği çarpımı arama uzayındaki toplam düğüm sayısını verir.

Genelleştirme ve baskılama tekniğiyle özniteliklerin küresel olarak yeniden kodlanmasıyla oluşturulan arama uzayında mahremiyet gereksinimini sağlayan en uygun çözüm düğümünün bulunması gerekir. Arama uzayında veri faydası ile mahremiyet riskleri arasındaki en uygun dengeyi sağlayan çözüm düğümünün bulunması NP-Zor problemdir [44]. Çözüm düğümünün bulunmasında arama algoritmaları kullanılır [75, 78, 134, 135].

Optimum çözümün bulunmasında, monotonluk özelliğiyle arama uzayını küçülterek arama performansını arttıran Flash [135] arama algoritması kullanılmıştır.

Monotonluk, çözüm uzayında optimum düğümün bulunmasında performans açısından önemli bir özelliktir [136]. Monotonluğun temelinde eşlenik sınıfların birleştirilmesi vardır. İki anonim eşlenik sınıfın birleşmesiyle oluşan anonim eşlenik sınıf birleşen eşlenik sınıfların özelliğini taşır. Örneğin 4-Anonim iki eşlenik sınıf birleştirildiğinde oluşan eşlenik sınıf yine en az 4-Anonim olacaktır. Benzer olarak ℓ-Çeşitliliği ele aldığımızda ℓ-Çeşitli iki eşlenik sınıfın birleşmesiyle oluşan yeni eşlenik sınıf içerisinde en azından ℓ çeşit hassas değer olacaktır.

Arama uzayında yer alan mahremiyet gereksinimlerini sağlayamayan düğümlerin öncülü de anonim olamayacağından öncüller içerisinde arama yapılmasına gerek olmayacaktır. Benzer olarak mahremiyet gereksinimlerini sağlayan bir düğümün yer aldığı soncul düğümlerde çözüme aday düğümler olarak arama uzayı içerisinde kalacaktır. Monotonluk özelliğinin örneklenmesi amacıyla genelleştirme yükseklikleri sırasıyla (3, 2, 6) olan yarı tanımlayıcı özniteliklere ait genelleştirme örüntüsü Şekil 3-5'de verilmiştir [72].



Şekil 3-5 Arama uzayı

Şekil 3-5'de verilen örnekte yarı tanımlayıcı özniteliklere ait hiyerarşi yüksekliklerinin çarpımıyla (3x2x6) arama uzayının büyüklüğü 36 düğüm olarak hesaplanır. Arama uzayını

36 düğümden daha az düğüme indirgeyerek arama performansını arttırmak amacıyla monotonluk özelliği kullanılmıştır.

Şekil 3-5'de verilen örüntü içerisinde Seviye-6'da bulunan koyu yeşil (2, 1, 3) düğümü mahremiyet gereksinimlerini sağlayan anonim bir düğümdür. Bu düğümün takipçi düğümleri olan ve açık yeşil ile gösterilen seviye-7'deki (2, 1, 4), seviye-8'deki (2, 1, 5) düğümleri monotonluk özelliğinden dolayı optimum çözüm düğümlerinin arandığı alan içinde yer alacaktır. Benzer şekilde seviye-4'de bulunan koyu kırmızı (2, 1, 1) düğümü mahremiyet gereksinimlerini sağlayamadığından arama alanı dışına çıkartılmıştır. Arama alanı dışında kalan bu düğümün öncül düğümleri olan diğer tüm düğümlerde monotonluk özelliğinden faydalanılarak arama alanı dışında bırakılmıştır. Yeşil renkli düğümler arama uzayının yeni sınırları olup kırmızı renge boyanan düğümler ise arama alanı dışında bırakılmıştır.

3.1.4 Modeller ve Parametreler

Literatürde yapılan çalışmalar veri faydası açısından incelendiğinde yapılan çalışmaların büyük bir çoğunluğunda, genelleştirme ve baskılama teknikleriyle anonimleştirme yapıldığı tespit edilmiştir. Mahremiyet gereksinimlerinin sağlanmasında ise en yaygın k-Anonimlik ile ℓ-Çeşitlilik ve t-Yakınlık modellerinin uygun birleşimlerinin birlikte kullanıldığı görülmüştür [128, 137, 138]. Tez çalışması kapsamında önerilen modelle yaygın olarak kullanılan k-Anonimlik, ℓ-Çeşitlilik ve t-Yakınlık modelleri karşılaştırılmıştır. Anonimleştirmede ise genelleştirme ve baskılama tekniklerinin kullanılmıştır. k-Anonimlik, ℓ-Çeşitlilik, t-Yakınlık k, (k, ℓ) , (k, t), ve (k, ℓ, t) modellerinin uygun kombinasyonları oluşturularak deneyler yapılmıştır. Parametre olarak veri faydası ve mahremiyet dengesini sağlamak için literatürde daha önce kullanılan k = 5, ℓ =2, t=0.2 değerleri seçilmiştir [51, 56, 72].

3.1.5 Hazırlık Katmanı Algoritması

Çalışma kapsamında önerilen ρ-Kazanım modelinin hazırlık katmanına ait algoritması Çizelge 3-3'de verilmiştir.

Çizelge 3-3 ρ-Kazanım hazırlık katmanı algoritması

Algoritma ρ-Kazanım (Hazırlık Katmanı)	
1	$T^*_{dataset}$ =**ρ-Kazanım** ($T_{dataset}$)
	Girdi: Mikro veri tablosu $T_{dataset}$ {A_1, A_2, A_3 ,.., A_n}
	Çıktı: Anonim mikro veri tablosu $T^*_{dataset}$ (QID*, SA)
2	$S_{dataset}$ {ID, {QID}, SA, NSA}←Öznitelik_Sınıflandır ($T_{dataset}$ {A_1, A_2, A_3 ,.., A_n})
3	$S_{dataset}$ {{QID}, SA}←Öznitelik_Çıkart ($S_{dataset}$ {ID, {QID}, SA, NSA}; {ID, NSA})
4	$S'_{dataset}$ {{QID}, SA}←$S_{dataset}$ {{QID}, SA}
5	for all $S_{dataset}$ {QID{ A_1, A_2, A_3 ,.., A_j}} do
6	Genelleştirme_Hiyerarşisi_Oluştur ($S_{dataset}$ { QID{ A_1, A_2, A_3 ,.., A_j}})
7	input k for k-Anonimlik
8	input ℓ for ℓ-Çeşitlilik
9	input t for t-Yakınlık
10	input ρ for ρ-Kazanım
11	ρ' = 0

Çizelge 3-3'de verilen tüm işlemler adımlara göre aşağıda açıklanmıştır.

Adım-1: ρ-Kazanım algoritması girdi olarak orijinal veri kümesini ($T_{dataset}$) alırken, çıktı olarak veri faydası yüksek mahremiyet korumalı kayıtları içeren anonim veri kümesini ($T^*_{dataset}$) veri alıcıları için yayınlar.

Adım 2: Başlangıçta mikro veri tablosunda yer alan sınıflandırılmamış öznitelikler (A_1, A_2, A_3 ,.., A_n) muhatapları hakkında verdikleri bilgilere göre (ID, QID, SA, NSA) sınıflarına ayrılır.

Adım 3: ID ve NSA sınıfında yer alan öznitelikler yayınlanacak veri kümesinden çıkartılır.

*Adım 4:*QID ve SA özniteliklerinden oluşan veri kümesinin başlangıçta bir kopyası alınır. Kopya tablo anonimleştirilmiş aykırı k ayıtların orijinal haline getirilebilmesi için kullanılır.

Adım 5-6: QID içerisinde yer alan özniteliklerin genelleştirme hiyerarşileri oluşturulur

Adım 7-10: Mahremiyet modelleri için gerekli olan parametreler (k, ℓ, t, ρ) belirlenerek anonimleştirme öncesi hazırlıklar tamamlanır. Parametreler seçilen anonimleştirme

algoritmalarına göre değişiklik gösterir. k-Anonimlik, ℓ-Çeşitlilik, t-Yakınlık modellerinin (k), (k, ℓ), (k, t), ve (k, ℓ, t) kombinasyonları oluşturulacağı için bu modellere ait parametreler girilmiştir.

Adım 11: İterasyon döngü kontrolü olarak seçilen ρ' değeri sıfırlanır.

Hazırlık katmanında işlemler tamamlanarak kazanımsal anonimleştirme katmanına geçilir.

3.2 Kazanımsal Anonimleştirme Katmanı

Hazırlık katmanında anonimleştirme için hazırlıkları tamamlanan veri kümesi üzerinde veri faydasını arttırıcı işlemler kazanımsal anonimleştirme katmanında yapılır. Bu katmanda, veri faydasının arttırılması için aykırı (outlier) kayıtlar üzerinde çalışılmıştır. Bu çalışmanın bir diğer katkısı olan faydaya göre eşlenik sınıfların kategorize edilmesi yine bu katmanda yer alır. Bu çalışma kapsamında, geleneksel anonimleştirme sonucunda elde edilen eşlenik sınıflar veri faydasına göre aykırı eşlenik sınıf (Outlier Equivalence Class-OEC) ve faydalı eşlenik sınıflar (Utility Equivalence Class-UEC) olarak sınıflandırılmıştır. Veri faydasını arttırıcı işlemler OEC üzerinde yapılmıştır. Kazanımsal anonimleştirme katmanında yapılan işlemler takip eden alt bölümlerde verilmiştir.

3.2.1 Fayda Ölçümü

ρ-Kazanım modeli eşlenik sınıflar düzeyinde işlem yaptığından veri faydası ölçümü eşlenik sınıflar düzeyinde yapılmıştır. Eşlenik sınıflara göre ölçüm yapan metriklerden eşlenik sınıflar ortalaması (ESO) ve ayırt edilebilirlik metriği (AEM) ρ-Kazanım modelinin veri faydası ölçümünde kullanılmıştır.

Veri faydası ölçümünün anlaşılabilmesi amacıyla DEMOGRAFİK veri kümesine ait kayıtların ESO ve AEM metrikleriyle ölçüm örneği Şekil 3-6'da verilmiştir.

Average class size	8.48589 (0.01238%)
Maximal class size	6304 (9.19406%)
Minimal class size	5 (0.00729%)
Number of classes	8080
Number of records	68566

a) ESO ölçüm örneği

Average class size	22.85533 (0.03333%)
Maximal class size	43 (0.06271%)
Minimal class size	5 (0.00729%)
Number of classes	3000
Number of records	68566

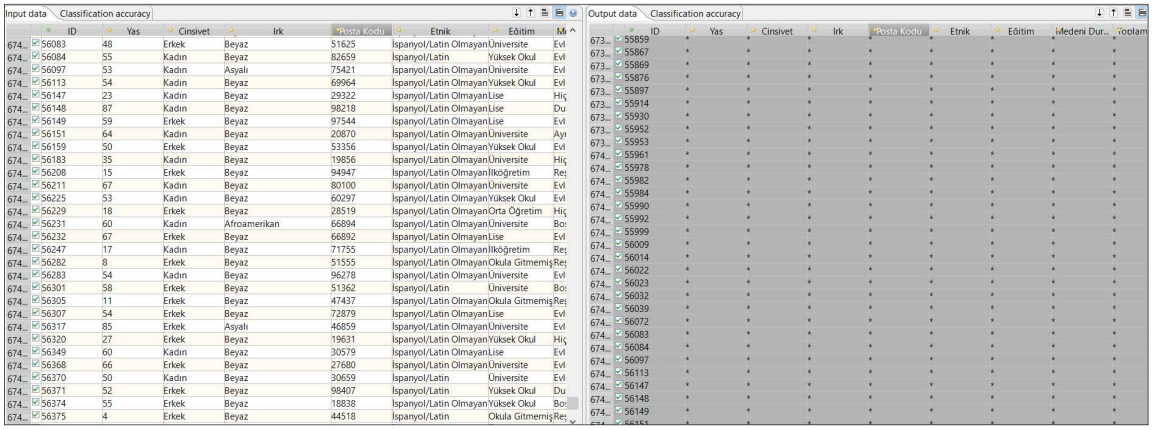
b) AEM ölçüm örneği

Şekil 3-6 Örnek veri faydası ölçümleri

Şekil 3-6'da verilen ölçüm örneklerinde veri faydasının ölçülmesinde ortalama sınıf sayıları dikkate alınacaktır. Ortalama sınıf sayısının küçük olması ölçülen veri kaybının az olduğunu gösterir. Veri kaybının az olması veri faydasının yüksek olması anlamına geldiğinden veri faydası ölçümlerinde ortalama eşlenik sınıf sayısının düşürülmesi veri faydasının arttırıldığı anlamına gelecektir. Bilgi metrikleri aracılığıyla ölçülen değerler veri faydası açısından değerlendirildiğinde en küçük eşlenik sınıf ortalamasına sahip olunması veri faydasının arttırılması olarak yorumlanır.

3.2.2 Aykırı Kayıtlar

Aykırı kayıtlar geleneksel anonimleştirme yöntemleri sonucunda herhangi bir eşlenik sınıfta yer bulamayan ve tamamen bastırılmış veri faydası olmayan dışlanmış kayıtlardır. DEMOGRAFİK veri setine ait aykırı kayıtların bir örneği Şekil 3-7'de gösterilmiştir.



Şekil 3-7 Aykırı kayıt örneği

Şekil 3-7 incelendiğinde resmin sağ tarafında tamamen bastırılan aykırı kayıtları sol tarafında ise aykırı kayıtların orijinal halleri görülmektedir. Anonimleştirme sonucunda tamamen bastırılan kayıtların dışarıda bırakılması eşlenik sınıf ortalamasını düşüreceğinden veri faydasını olumlu etkiler. ρ-Kazanım modelinin veri faydasını arttırıcı iki özelliği vardır. Bunlardan birincisi, aykırı kayıtların yeniden kullanılmasıyla aykırı kayıt sayısını azalması ve eşlenik sınıf sayısının artarak veri faydasındaki iyileşmedir. Azaltılan aykırı kayıtların yayınlanan veri kümesinin dışında tutulmasıyla oluşan eşlenik sınıf ortalamasındaki düşüş ise ρ-Kazanım modelinin veri fayda arttırıcı bir diğer özelliğidir.

Aykırı kayıtların dışarıda tutulmasının DEMOGRAFİK veri kümesinden elde edilen 44108 kayıt üzerinden verilen örnek bir gösterimle veri faydasına olan etkisi Şekil 3-8'de gösterilmiştir.

Measure	Including outliers	Excluding outliers
Average class size	95.06034 (0.21552%)	81.82937 (0.21598%)
Maximal class size	6221 (14.10402%)	1102 (2.90865%)
Minimal class size	5 (0.01134%)	5 (0.0132%)
Number of classes	464	463
Number of records	44108	37887 (85.89598%)
Suppressed records	6221 (14.10402%)	0

Şekil 3-8 Aykırı kayıtların veri faydasına etkisi

Şekil 3-8'de veri faydası açısından aykırı kayıtları içeren ölçümlerle içermeyen ölçümler karşılaştırılmıştır. Aykırı kayıtları içeren durumda eşlenik sınıf ortalaması 95,06 olarak ölçülmüştür. Aykırı kayıtları içermeyen durumda ise eşlenik sınıf ortalaması 81,82 olarak ölçülmüştür. Aykırı kayıtların dışarıda tutulmasıyla azalan eşlenik sınıf sayısı ortalaması veri faydasındaki iyileşmenin göstergesidir.

Şekil 3-8'de verilen ölçümde tüm kayıtların %14'üne denk gelen aykırı eşlenik sınıf olduğu görülmüştür. ρ-Kazanım modeli aykırı eşlenik sınıf üzerinde yapacağı işlemlerle mahremiyet gereksinimlerini sağlayan faydalı eşlenik sınıflar üretecektir. Üretilen faydalı eşlenik sınıflar 95,06 olan genel eşlenik sınıf ortalamasını düşürerek veri faydasında iyileşme sağlayacaktır. Ayrıca eşlenik sınıf üretimi sonrasında geriye kalan atık aykırı kayıtların dışarıda bırakılması sonucunda 81,82 olan eşlenik sınıf ortalamasını daha aşağıya çekecektir.

ρ-Kazanım modelinin aykırı eşlenik sınıf üzerinde yaptığı işlemler Eşitlik 3-1'de verilmiştir.

$$U_{\rho}\left((|UEC| \leq |UE\ '|) \wedge (|OEC| \geq |OE\ '|)\right) \qquad 3\text{-}1$$

Bu eşitlikte;

|OEC| :Başlangıçtaki aykırı eşlenik sınıf içindeki kayıt sayısı

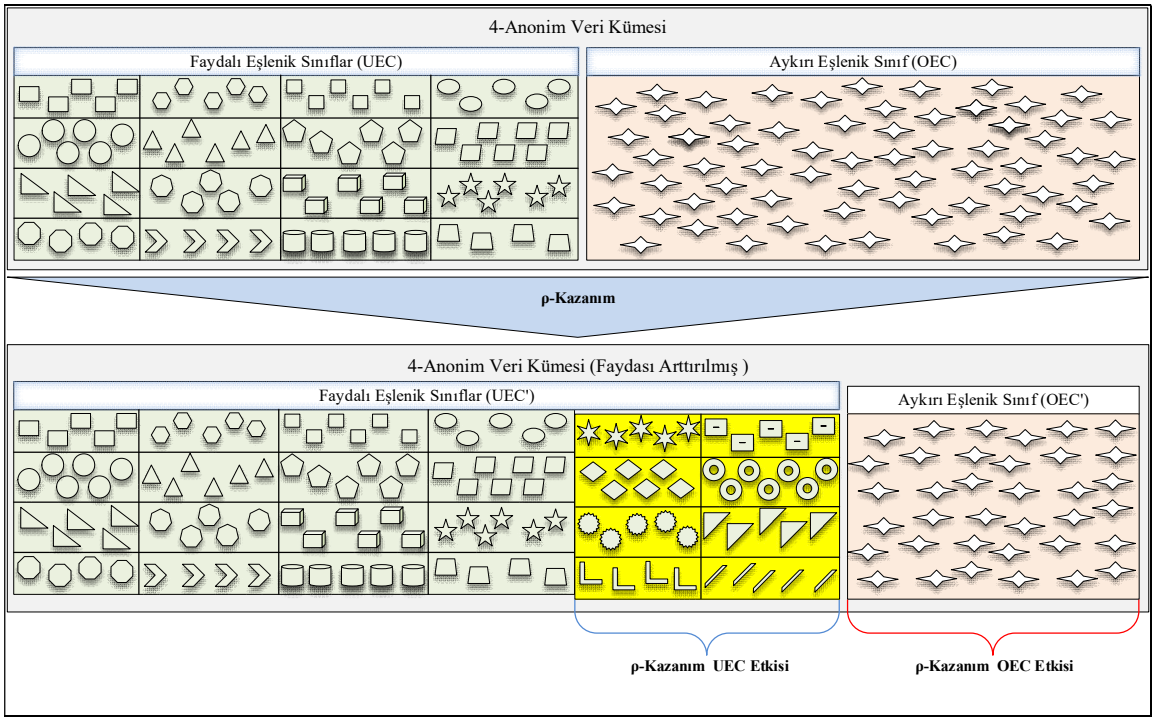
|UEC| :Başlangıçtaki faydalı eşlenik sınıf sayısı

|OEC'| : ρ-Kazanım modeli sonrası aykırı eşlenik sınıf kayıt sayısı (atık aykırı kayıt)

|UEC'| : ρ-Kazanım modeli sonrası faydalı eşlenik sınıf sayısı

U_ρ : ρ-Kazanım modelinin bilgi kaybı açısından ölçümünde kullanılacak gösterge fonksiyonu (içerisindeki değer doğruysa 1, diğer durumlarda 0 değeri alır)

Anonim bir veri kümesinde eşlenik sınıflar üzerinde ρ-Kazanım modelinin etkisini gösteren yeni bir çizim Şekil 3-9'da gösterilmiştir.



Şekil 3-9 ρ-Kazanım modelinin OEC ve UEC etkisi

Şekil 3-9'da geleneksel anonimleştirme sonucunda temsili 16 tane faydalı eşlenik sınıf oluştuğu görülmektedir. Ayrıca veri faydası olmayan aykırı kayıtları içeren aykırı eşlenik sınıf ayrı bir küme olarak gösterilmiştir. ρ-Kazanım Modeli sonrasında aykırı eşlenik sınıf üzerinde yapılan işlemlerle faydalı eşlenik sınıfların (UEC') sayısı artarken aykırı eşlenik sınıfın (OEC') boyutu azalmaktadır.

3.2.3 Risk Ölçümü

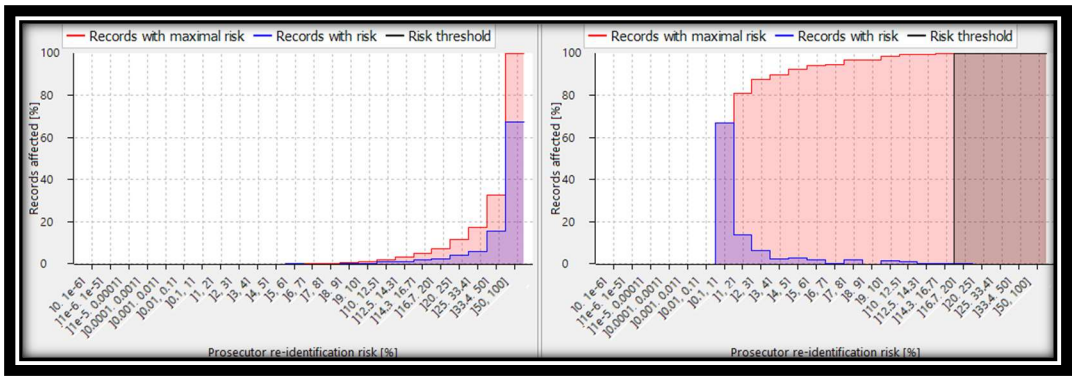
Veri faydasındaki iyileşmenin mahremiyet risklerini nasıl etkilediği ρ-Kazanım modelinin başarısının ölçülmesinde önemli bir parametredir. Mahremiyet korumalı yaklaşımların başarımının değerlendirilmesinde veri faydası ile mahremiyet risklerinin birlikte değerlendirilmesi gerekir. ρ-Kazanım risk ölçümünde saldırganın yayınlanan veri içerisinde kimliğini yeniden tanımlamak istediği kişinin olduğu bilgisine sahip olduğu varsayımı yapılarak en yüksek riske sahip olan basit risk metriklerinden savcı yaklaşımı kullanılacaktır. ADULT veri kümesi için savcı, gazeteci ve pazarlamacı yaklaşımlarına ait örnek değerler Şekil 3-10'da verilmiştir.

Measure	Value [%]
Lowest prosecutor risk	5.55556%
Records affected by lowest risk	0.31629%
Average prosecutor risk	78.96679%
Highest prosecutor risk	100%
Records affected by highest risk	67.10596%
Estimated prosecutor risk	100%
Estimated journalist risk	100%
Estimated marketer risk	78.96679%

Measure	Value [%]
Lowest prosecutor risk	0.23585%
Records affected by lowest risk	8.55874%
Average prosecutor risk	1.57449%
Highest prosecutor risk	20%
Records affected by highest risk	0.20186%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	1.57449%

Şekil 3-10 Risk ölçümleri

Şekil 3-10'da verilen örnek ölçüm incelendiğinde anonimleştirme öncesi (sol taraf) ortalama savcı riski %78,96 olarak ölçülürken, kayıtların tamamı savcı ve gazeteci yaklaşımı dikkate alındığında yüksek riskten etkilenmektedir. Anonimleştirme sonrasında (sağ) ise ortalama savcı riski %1,57 olarak ölçülürken en yüksek riskten etkilenen kayıtların %20'ye düştüğü görülür. Şekil 3-10'daki değerlere göre çizilen grafiksel gösterim Şekil 3-11'de verilmiştir.



Şekil 3-11 Risk grafiği örneği

Şekil 3-11'de benzer olarak sol grafik anonimleştirme öncesi, sağ grafik ise anonimleştirme sonrasını göstermektedir. Deneysel değerlendirmelerde ρ-Kazanım modelinin uygulandığı her bir iterasyonda ölçülen risk değerlerinde anlamlı bir değişiklik olmaması modelin başarı göstergesi olacaktır. ρ-Kazanım modeli gösterge fonksiyonu Eşitlik 3-2'de verilmiştir.

$$P_\rho({}_pR_b \leq {}_pR'_b) \qquad 3\text{-}2$$

Bu eşitlikte;

${}_pR_b$: Geleneksel anonimleştirme sonucunda savcı yaklaşımıyla hesaplanan risk değerini,

${}_pR'_b$: ρ-Kazanım modelinin uygulanması sonucunda elde edilen risk değerini;

P_ρ : ρ-Kazanım modeli gösterge fonksiyonu (içerisindeki değer doğruysa 1, diğer durumlarda 0 değeri alır)

3.2.4 ρ-Kazanım Karar Kuralı

ρ-Kazanım modelinin veri faydası ile mahremiyet risk dengesini korumadaki başarısının yorumlanmasında karar kuralına ihtiyaç vardır. Eşitlik 3-3'de verilmiştir.

$$D_\rho = \begin{cases} BAŞARILI & I\ (U_\rho \wedge P_\rho) = 1 \\ BAŞARISIZ & I\ (U_\rho \wedge P_\rho) = 0 \end{cases} \qquad 3\text{-}3$$

Bu eşitlikte;

$D_{\rho:}$: ρ-Kazanım karar kuralı,

I : Karar verme gösterge fonksiyonu (içerisindeki değer doğruysa 1, diğer durumlarda 0 değeri alır)

ρ-Kazanım modelinin uygulanması sonucunda mahremiyetten ödün vermeden veri faydası arttırıldığında karar kuralı (D_ρ) BAŞARILI değerini, diğer durumlarda ise BAŞARISIZ değerini döndürerek veri yayıncısını bilgilendirecektir.

3.2.5 Kazanımsal Anonimleştirme Katmanı Algoritması

ρ-Kazanım modelinin kazanımsal anonimleştirme katmanına ait kısmi algoritması Çizelge 3-4'de verilmiştir.

Çizelge 3-4 ρ-Kazanım kazanımsal anonimleştirme katmanı algoritması

Algoritma ρ-Kazanım (Kazanımsal Anonimleştirme Katmanı)	
11	……..……..
12	do {
13	$S^*_{dataset}$ {QID*, SA} ←Anonimleştir ($S_{dataset}$ {QID, SA} ; (k-Anonimlik, ℓ-Çeşitlilik))
14	$TEC_{dataset}$ {UEC, OEC}←Eşlenik_Sınıfları_Ayrıştır ($S^*_{dataset}$ {QID*, SA})
15	for all $TEC_{dataset}$ {UEC, OEC} do
16	$T^*_{dataset}$ {QID*, SA}← Faydalı_Eşlenik_Sınıflar ($TEC_{dataset}$ {UEC})
17	$TEC'_{dataset}$ {QID*, SA}← Aykırı_Kayıtlar ($TEC_{dataset}$ {OEC})
18	Tablo_Hazırla ($S_{dataset}$ {QID}, SA)
19	$S_{dataset}$ {QID, SA}← Aykırı_Kayıtları_Eşleştir ($TEC'_{dataset}$ {QID*,SA}; $S'_{dataset}$ {QID, SA})
20	ρ'= ρ ' + 1
21	} while (ρ' = ρ)
22	……..

Çizelge 3-4'de verilen tüm işlemler adımlara göre aşağıda açıklanmıştır.

Adım-12: ρ-iterasyon başlangıcı

Adım-13: QID öznitelik kümesinin anonimleştirmesinde seçilen yardımcı modellerin (k-Anonimlik, ℓ-Çeşitlilik, t-Yakınlık) uygun kombinasyonları kullanılır. Bu örnekte (k,ℓ) kombinasyonu kullanılmıştır. Deneysel çalışmalar ve bulgular bölümünde (k), (k,ℓ), (k,t) (k,ℓ,t) kombinasyonları çalışılmıştır.

Adım-14: (k,ℓ)-Anonimleştirme sonucunda $S^*_{dataset}$ anonim veri kümesi oluşur. Anonim veri kümesi içerisinde en az k sayıda faydalı kayıt içeren faydalı eşlenik sınıflar (UEC) ile tüm aykırı kayıtları içeren aykırı eşlenik sınıf (OEC) bulunur. Anonim veri kümesi içerisinde yer

alan tüm eşlenik sınıflar (All Equivalence Class- TEC) veri faydasına göre faydalı eşlenik sınıflar (UEC) ve aykırı eşlenik sınıf (OEC) olarak sınıflandırılır.

Adım 15-17: Faydalı eşlenik sınıflar paylaşım tablosuna ($T^*_{dataset}$), OEC içerisindeki aykırı kayıtlar kazanım tablosuna ($TEC'_{dataset}$) taşınır. Bu işlem anonim veri kümesi içerisindeki tüm eşlenik sınıflar için tekrarlanır.

Adım 18: $S_{dataset}$ tablosu kazanım kayıtlarının taşınabilmesi için hazırlanır.

Adım 19: Aykırı kayıtları içeren $TEC'_{dataset}$ kazanım tablosu orijinal kayıtları içeren $S'_{dataset}$ tablosu yardımıyla anonimleştirme öncesine döndürülerek, kazanım için hazırlanan $S_{dataset}$ tablosuna taşınır.

Adım 20: ρ' iterasyon kontrol değişkeni arttırılır.

Adım 12-21: Kazanımsal anonimleştirme katmanındaki kazanım döngüsü için kullanılan iterasyon katsayısı kadar bu işlemler tekrarlanır.

Kazanımsal anonimleştirme katmanında yapılan işlemler tamamlanarak yayınlama katmanına geçilir.

3.3 Yayınlama Katmanı

ρ-Kazanım modelinin son aşaması olan yayınlama katmanında veri faydası ve mahremiyet konusunda mahremiyet gereksinim kontrolleri yapılır. Kontroller sonrasında atık aykırı kayıtlar yayınlanacak kayıtlardan çıkarılır. Faydalı kayıtları içeren UEC eşlenik sınıfları veri alıcıları için yayınlanır. ρ-Kazanım modelinin tüm aşamalarını gösteren ρ-Kazanım algoritması Çizelge 3-5'de verilmiştir.

Çizelge 3-5 ρ-Kazanım algoritması

Algoritma ρ-Kazanım	
1	$T^*_{dataset}$ =**ρ-Kazanım** ($T_{dataset}$)
	Girdi: Mikro veri tablosu $T_{dataset}$ {A_1, A_2, A_3 ,.., A_n}
	Çıktı: Anonim mikro veri tablosu $T^*_{dataset}$ (QID*, SA)
2	$S_{dataset}$ {ID, {QID}, SA, NSA}←Öznitelik_Sınıflandır ($T_{dataset}$ {A_1, A_2, A_3 ,.., A_n})
3	$S_{dataset}$ {{QID}, SA}←Öznitelik_Çıkart ($S_{dataset}$ {ID, {QID}, SA, NSA}; {ID, NSA})
4	$S'_{dataset}$ {{QID}, SA}←$S_{dataset}$ {{QID}, SA}
5	for all $S_{dataset}$ {QID{ A_1, A_2, A_3 ,.., A_j}} do
6	Genelleştirme_Hiyerarşisi_Oluştur ($S_{dataset}$ { QID{ A_1, A_2, A_3 ,.., A_j}})
7	input k for k-Anonimlik
8	input ℓ for ℓ-Çeşitlilik
9	input t for t-Yakınlık
10	input ρ for ρ-Kazanım
11	ρ' = 0
12	do {
13	$S^*_{dataset}$ {QID*, SA} ←Anonimleştir ($S_{dataset}$ {QID, SA} ; (k-Anonimlik, ℓ-Çeşitlilik))
14	$TEC_{dataset}$ {UEC, OEC}←Eşlenik_Sınıfları_Ayrıştır ($S^*_{dataset}$ {QID*, SA})
15	for all $TEC_{dataset}$ {UEC, OEC} do
16	$T^*_{dataset}$ {QID*, SA}← Faydalı_Eşlenik_Sınıflar ($TEC_{dataset}$ {UEC})
17	$TEC'_{dataset}$ {QID*, SA}← Aykırı_Kayıtlar ($TEC_{dataset}$ {OEC})
18	Tablo_Hazırla ($S_{dataset}$ {QID}, SA)
19	$S_{dataset}$ {QID, SA}← Aykırı_Kayıtları_Eşleştir ($TEC'_{dataset}$ {QID*,SA}; $S'_{dataset}$ {QID, SA})
20	ρ'= ρ ' + 1
21	} while (ρ' = ρ)
22	Mahremiyet_Kontrol ($T^*_{dataset}$)
23	Yayınla ($T^*_{dataset,}$ UEC)

3.4 Bölüm Sonucu

Bu bölümde tez çalışması kapsamında ρ-Kazanım modelinin uygulanabilmesi için gerekli olan materyal ile yöntemler örnekler verilerek anlatılmıştır. Modelin algoritması katmanlara bölünerek ayrıntılı olarak açıklanmış ve her katmanın görevi detaylandırılmıştır. Modelin uygulanması sonucunda mahremiyet riskleri ile veri faydasının nasıl ölçüleceği ve elde edilen sonuçların nasıl değerlendirileceği konusunda veri yayıncıya yardımcı olacak karar kuralı hakkında açıklamalar yapılmıştır. Takip eden bölümde tez çalışması kapsamında önerilen modelin deneysel çalışmaları, sonuçları ve karşılaştırmaları verilmiştir.

4 DENEYSEL ÇALIŞMALAR VE BULGULAR

Bu bölümde tez çalışmasında önerilen modelin gerçekleştirilmesi sonucunda elde edilen deneysel değerlendirmelere ve bulgulara yer verilmiştir. Materyal ve yöntem bölümünde tanımlandığı şekilde deneysel çalışmalar yapılarak sonuçları değerlendirilmiştir.

4.1 Öznitelik Sınıflandırma ve Genelleştirme

Önerilen modelin gerçekleştiriminde kullanılacak veri kümeleri düzenlenerek anonimleştirmeye hazır hale getirilmiştir. Düzenleme sonrasında veri kümelerinde yer alan öznitelikler sınıflandırılmıştır. QID öznitelikleri için genelleştirme yükseklikleri belirlenmiştir. ARX editörü aracılığıyla öznitelik tiplerine göre (sayısal, kategorik vb.) QID genelleştirme hiyerarşileri oluşturulmuştur. Takip eden alt bölümlerde her bir veri kümesi üzerinde yapılan sınıflandırma ve hiyerarşi oluşturma işlemleri anlatılmıştır.

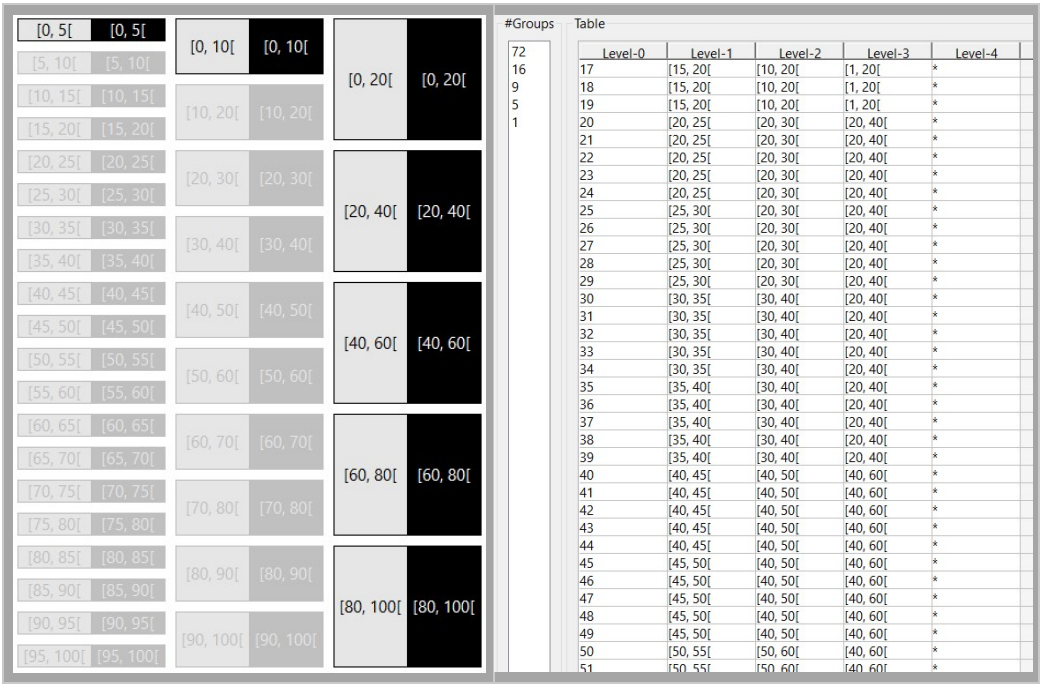
4.1.1 ADULT

ADULT veri kümesinin sınıflandırılmış öznitelikleri genelleştirme yükseklikleriyle birlikte Çizelge 4-1'de verilmiştir.

Çizelge 4-1 ADULT veri kümesi öznitelikleri

Veri Kümesi Adı	QID (Genelleştirme Yüksekliği)	SA
T_{ADULT}	yaş (5), cinsiyet (2), ırk (2), medeni_durum(3), eğitim (4), ülke (3), pozisyon (3), maaş (2)	meslek

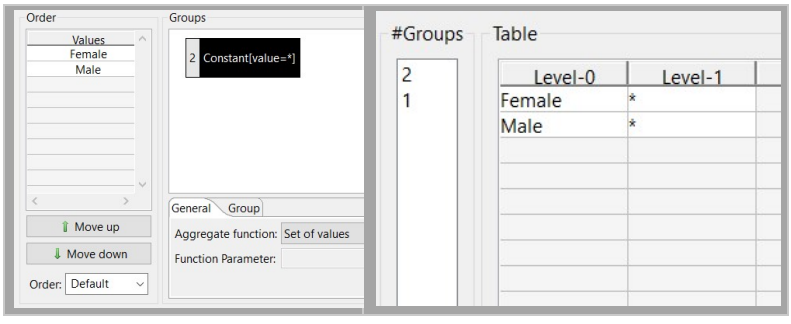
ADULT veri kümesi içerisinde yer alan sayısal öznitelikler belirlenen aralıklarda, kategorik öznitelikler farklı detay seviyelerinde genelleştirilmiştir. ADULT veri kümesindeki sayısal yaş özniteliği farklı aralıklara ayrılarak 4 yüksekliğinde genelleştirme hiyerarşisi oluşturulmuştur. Oluşturulan yaş özniteliği genelleştirme hiyerarşisi Şekil 4-1'de gösterilmiştir.



Şekil 4-1 ADULT Yaş özniteliği genelleştirme hiyerarşisi

Şekil 4-1'de verilen yaş özniteliğinin değer aralığı 17 ile 90 arasında yer almaktadır. Baskılama dahil beş seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 72 grup içinde yer alırken ikinci seviyede 16 grup içerisinde ardışık 5 değer yer almaktadır. Üçüncü seviyede ardışık 10 değer 9 grupta toplanırken, 4. seviyede ardışık 20 değer 5 grupta toplanmıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

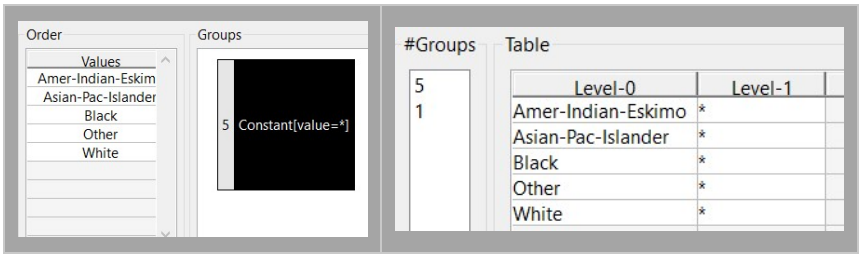
ADULT veri kümesinde yer alan kategorik cinsiyet özniteliğine ait genelleştirme hiyerarşisi Şekil 4-2'de verilmiştir



Şekil 4-2 ADULT Cinsiyet özniteliği genelleştirme hiyerarşisi

Şekil 4-2'de verilen ADULT veri kümesi kategorik cinsiyet özniteliğinin değerleri {Female, Male} olup baskılama dahil iki seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer iki grup içinde yer alırken ikinci seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

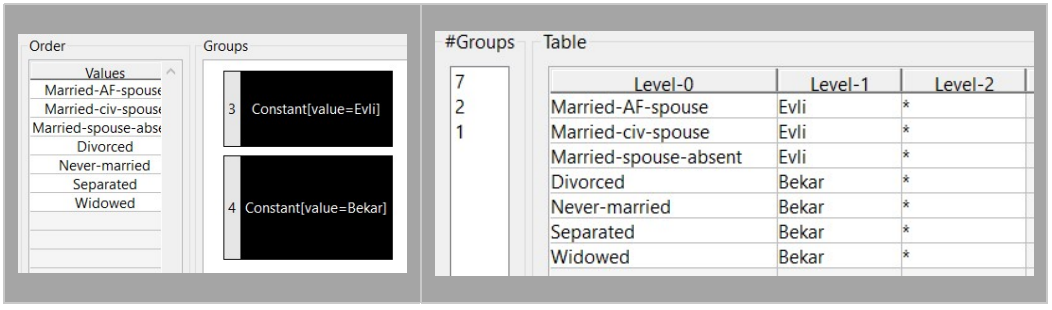
ADULT veri kümesinin bir diğer kategorik özniteliği ırk' a ait genelleştirme hiyerarşisi Şekil 4-3'de verilmiştir.



Şekil 4-3 ADULT Irk özniteliği genelleştirme hiyerarşisi

Şekil 4-3'de verilen ADULT veri kümesi kategorik ırk özniteliğinin değerleri {Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White} olup baskılama dahil iki seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer beş grup içinde yer alırken ikinci seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

ADULT veri kümesinin bir diğer kategorik özniteliği medeni durum'a ait genelleştirme hiyerarşisi Şekil 4-4'de verilmiştir

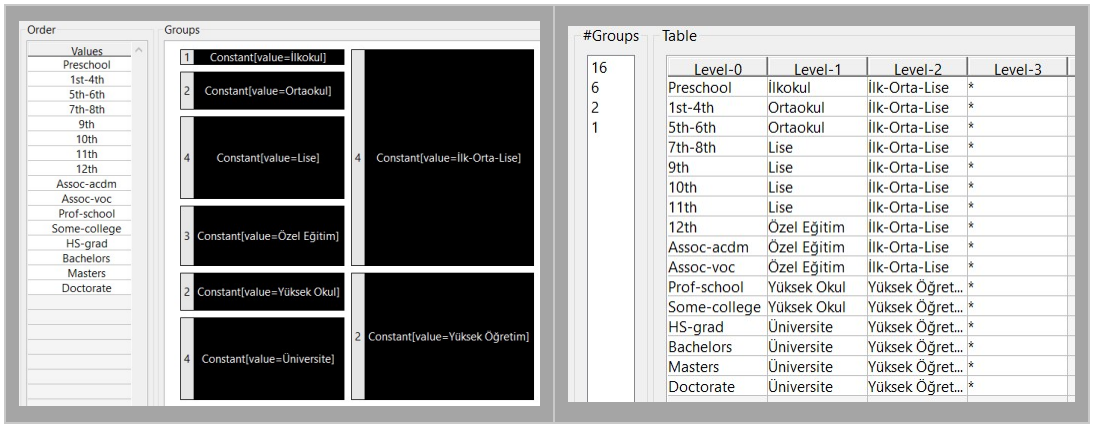


Şekil 4-4 ADULT Medeni durum özniteliği genelleştirme hiyerarşisi

Şekil 4-4'de verilen ADULT veri kümesi kategorik medeni_durum özniteliğinin değerleri {Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Divorced, Never-

married, Separated, Widowed } olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer yedi grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {Evli, Bekâr} olarak iki grupta yer almıştır. Üçüncü seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

ADULT veri kümesinin bir diğer kategorik özniteliği eğitim'e ait genelleştirme hiyerarşisi Şekil 4-5'de verilmiştir.

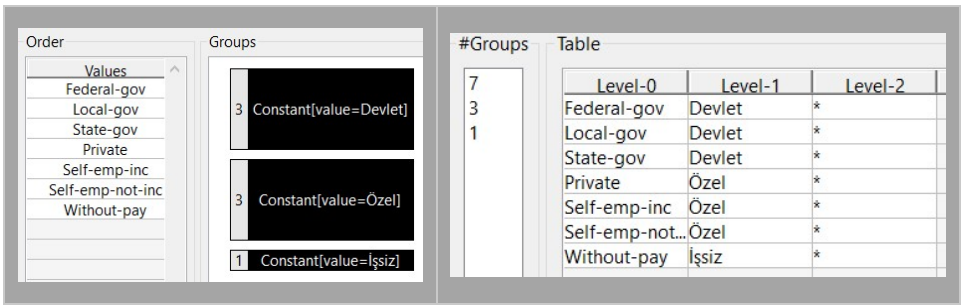


Level-0	Level-1	Level-2	Level-3
Preschool	İlkokul	İlk-Orta-Lise	*
1st-4th	Ortaokul	İlk-Orta-Lise	*
5th-6th	Ortaokul	İlk-Orta-Lise	*
7th-8th	Lise	İlk-Orta-Lise	*
9th	Lise	İlk-Orta-Lise	*
10th	Lise	İlk-Orta-Lise	*
11th	Lise	İlk-Orta-Lise	*
12th	Özel Eğitim	İlk-Orta-Lise	*
Assoc-acdm	Özel Eğitim	İlk-Orta-Lise	*
Assoc-voc	Özel Eğitim	İlk-Orta-Lise	*
Prof-school	Yüksek Okul	Yüksek Öğret...	*
Some-college	Yüksek Okul	Yüksek Öğret...	*
HS-grad	Üniversite	Yüksek Öğret...	*
Bachelors	Üniversite	Yüksek Öğret...	*
Masters	Üniversite	Yüksek Öğret...	*
Doctorate	Üniversite	Yüksek Öğret...	*

Şekil 4-5 ADULT Eğitim özniteliği genelleştirme hiyerarşisi

Şekil 4-5'de verilen ADULT veri kümesi kategorik eğitim özniteliğinin değerleri {Preschool,1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Assoc-acdm, Assoc-voc, Prof-school, Some-College, Hs-grad, Bachelors, Masters, Doctorate} olup baskılama dahil dört seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 16 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {İlkokul, Ortaokul, Lise, Özel Eğitim, Yüksekokul, Üniversite} altı grupta yer alırken üçüncü seviyede ise öznitelik değerleri{İlk-Orta_Lise, Yüksek Öğretim}iki grupta yer almıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

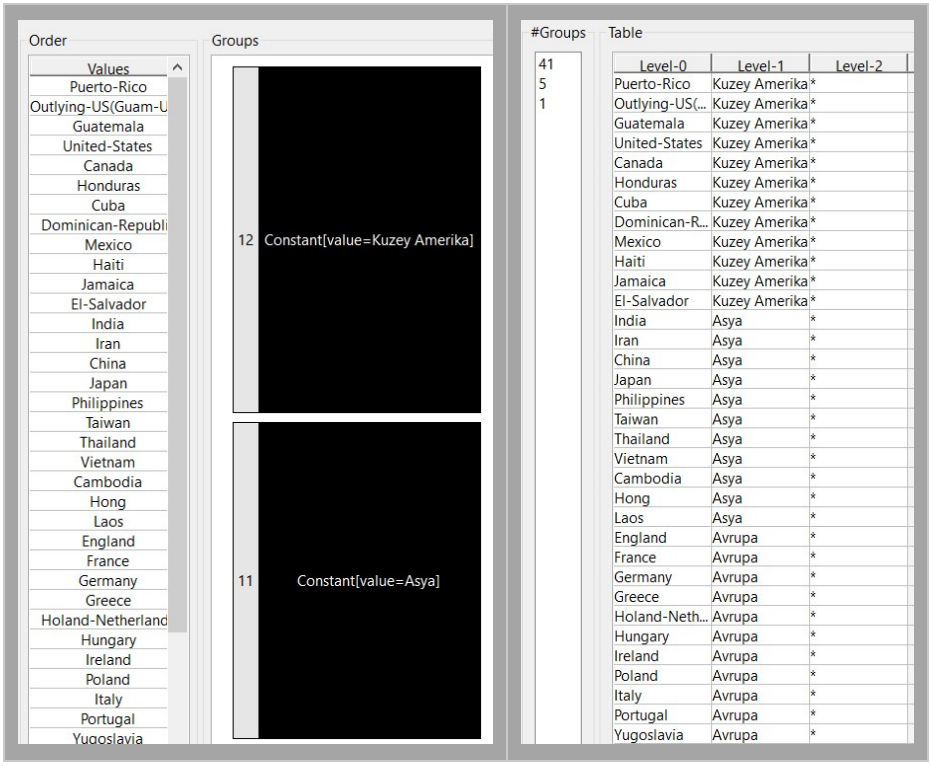
ADULT veri kümesinin bir diğer kategorik özniteliği olan iş özniteliğine ait genelleştirme hiyerarşisi Şekil 4-6'da verilmiştir.



Şekil 4-6 ADULT İş özniteliği genelleştirme hiyerarşisi

Şekil 4-6'da verilen ADULT veri kümesi kategorik iş özniteliğinin değerleri {Federal-gov, local-gov, state-gov, private, Self-emp-inc-, Self-emp-not, Without-pay } olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 7 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {Devlet, Özel, İşsiz} üç grupta yer alırken üçüncü seviyede ise öznitelik tamamen baskılanarak tek bir grupta (*) toplanmıştır.

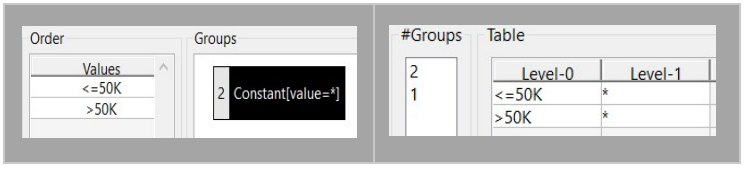
ADULT veri kümesinin bir diğer kategorik özniteliği olan ülke özniteliğine ait genelleştirme hiyerarşisi Şekil 4-7'de verilmiştir



Şekil 4-7 ADULT Ülke özniteliği genelleştirme hiyerarşisi

Şekil 4-7’da verilen ADULT veri kümesi kategorik ülke özniteliğinin değerleri {Peurto-Rico, Outlying US, Guatemala, United States ,.., Portuqal, Yugoslavia, Scotland, Columbia,Trinadad&Tobocco, Peru, Nicaragua, Ecuador, South Africa } olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 41 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {Kuzey Amerika, Asya, Avrupa, Güney Amerika, Afrika } beş grupta yer almıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) yer almıştır.

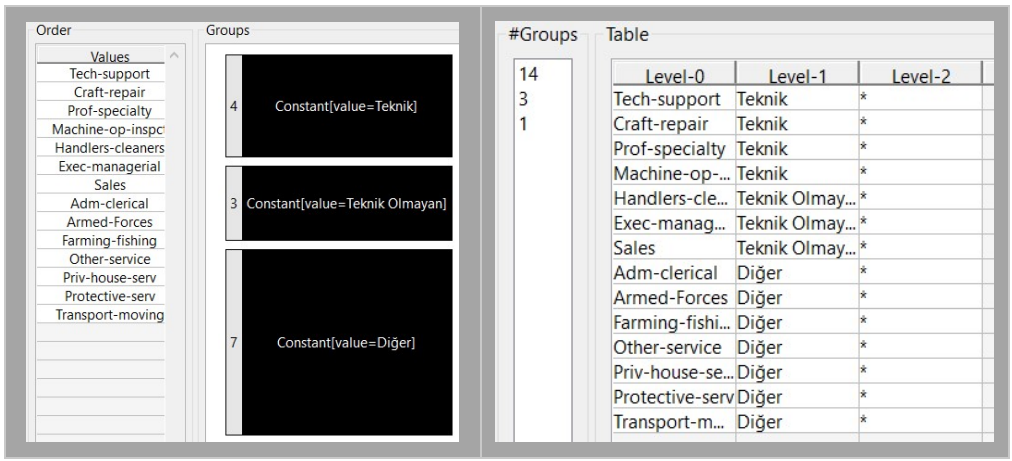
ADULT veri kümesinin bir diğer kategorik özniteliği olan maaş özniteliğine ait genelleştirme hiyerarşisi Şekil 4-8Şekil 4-9’de verilmiştir



Şekil 4-8 ADULT Maaş özniteliği genelleştirme hiyerarşisi

Şekil 4-8Şekil 4-9’de ADULT veri kümesi kategorik maaş özniteliğinin değerleri {≤ 50K, > 50K } olup baskılama dahil iki seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 2 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri baskılanarak tek bir (*) grupta yer almıştır.

ADULT veri kümesinin bir diğer kategorik özniteliği olan pozisyon özniteliğine ait genelleştirme hiyerarşisi Şekil 4-9’da verilmiştir.



Şekil 4-9 ADULT Pozisyon özniteliği genelleştirme hiyerarşisi

Şekil 4-9’da verilen ADULT veri kümesi kategorik pozisyon özniteliğinin değerleri {Tech-support, Craft-repair, Prof-speciality, Machine-op-inspct, Handlers-cleaners, Exec-managerial, Sales, Adm-clerical, Armed- forces, Farming-fishing, Other-srvice, Priv-house-serv, Protective-serv, Transport-moving } olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 14 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {Teknik, Teknik Olmayan, Diğer } üç grupta yer almıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) yer almıştır.

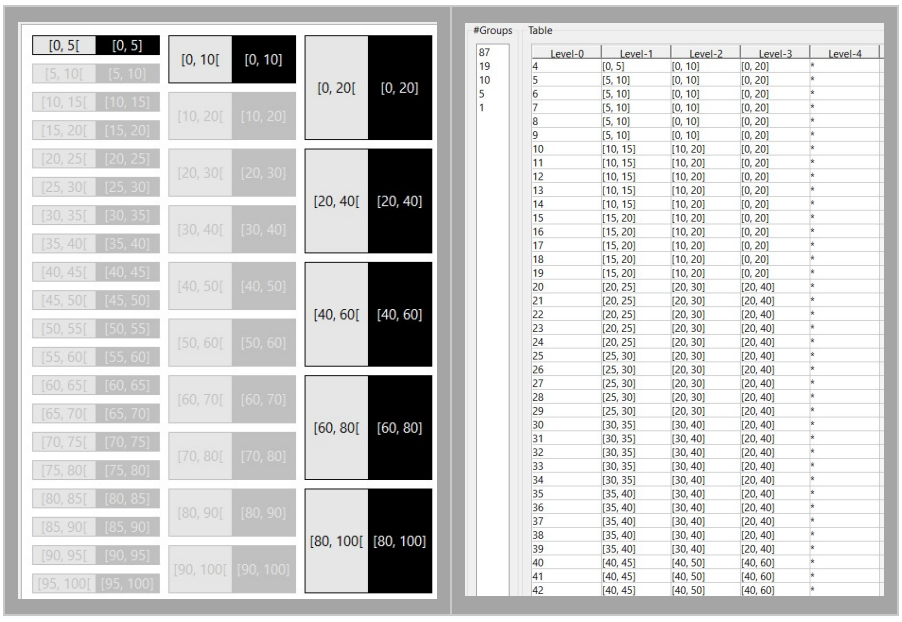
4.1.2 DEMOGRAFİK

DEMOGRAFİK veri kümesinin sınıflandırılmış öznitelikleri genelleştirme yükseklikleriyle birlikte Çizelge 4-2’de verilmiştir.

Çizelge 4-2 DEMOGRAFİK veri kümesi öznitelikleri

Veri Kümesi Adı	QID (Genelleştirme Yüksekliği)	SA
$T_{DEMOGRAFİK}$	yaş (5), cinsiyet (2), ırk (3), posta kodu (6), etnik (2), eğitim (3), medeni_durum (2), toplam gelir (3)	etnik

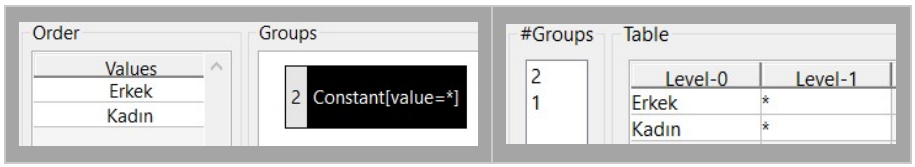
DEMOGRAFİK veri kümesi sayısal yaş özniteliğine ait genelleştirme hiyerarşisi Şekil 4-10’de verilmiştir.



Şekil 4-10 DEMOGRAFİK Yaş özniteliği genelleştirme hiyerarşisi

Şekil 4-10'de DEMOGRAFİK veri kümesi için verilen yaş özniteliğinin değer aralığı 1 ile 87 arasında olup baskılama dahil beş seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 87 grup içinde yer alırken ikinci seviyede 19 grup içerisinde ardışık 5 değer yer almaktadır. Üçüncü seviyede ardışık 10 değer 9 grupta toplanırken, 4. seviyede ardışık 20 değer 5 grupta toplanmıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

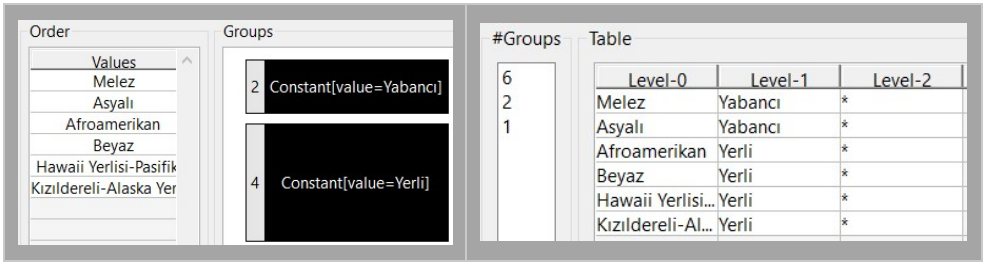
DEMOGRAFİK veri kümesinin kategorik tipindeki cinsiyet özniteliğine ait genelleştirme hiyerarşisi Şekil 4-11'de verilmiştir



Şekil 4-11 DEMOGRAFİK Cinsiyet özniteliği genelleştirme hiyerarşisi

Şekil 4-11'de verilen DEMOGRAFİK veri kümesi kategorik cinsiyet özniteliğinin değerleri {Erkek, Kadın} olup baskılama dahil iki seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer iki grup içinde yer alırken son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

DEMOGRAFİK veri kümesinin bir diğer kategorik özniteliği olan ırk özniteliğine ait genelleştirme hiyerarşisi Şekil 4-12'de verilmiştir

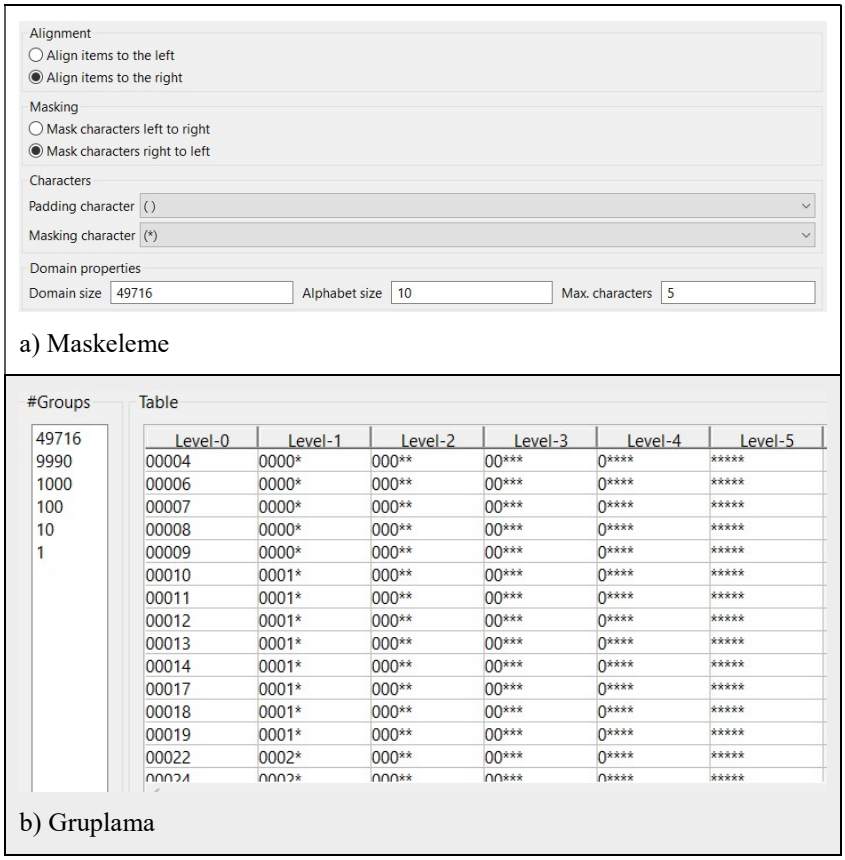


Şekil 4-12 DEMOGRAFİK Irk özniteliği genelleştirme hiyerarşisi

Şekil 4-12'de verilen DEMOGRAFİK veri kümesi kategorik ırk özniteliğinin değerleri {Melez, Asyalı, Afroamerikan, Beyaz, Hawaii Yerlisi-Pasifik, Kızıldereli-Alaska Yerlisi} olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer altı

grup içinde yer alırken, ikinci seviyede her bir değer iki grup içinde {Yabancı, Yerli} yer almıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

DEMOGRAFİK veri kümesi için alfanümerik tipindeki posta kodu özniteliğine ait genelleştirme hiyerarşisi Şekil 4-13'de verilmiştir



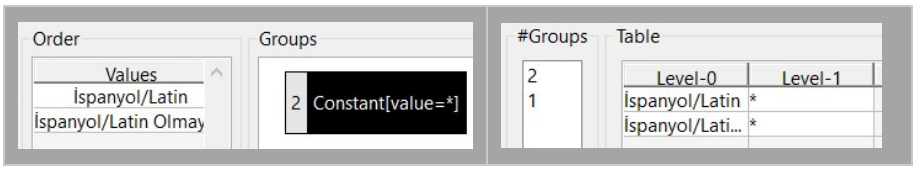
a) Maskeleme

b) Gruplama

Şekil 4-13 DEMOGRAFİK Posta kodu özniteliği genelleştirme hiyerarşisi

Şekil 4-13 (a)'da verilen DEMOGRAFİK veri kümesi alfanümerik posta kodu özniteliğinin değerleri 49716 farklı değer olup Şekil 4-13 (b)'de gösterildiği gibi baskılama dahil altı seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 49716 grup içinde yer alırken ikinci seviyede 9990, üçüncü seviyede 1000, dördüncü seviyede 100, beşinci seviyede 10, altıncı seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

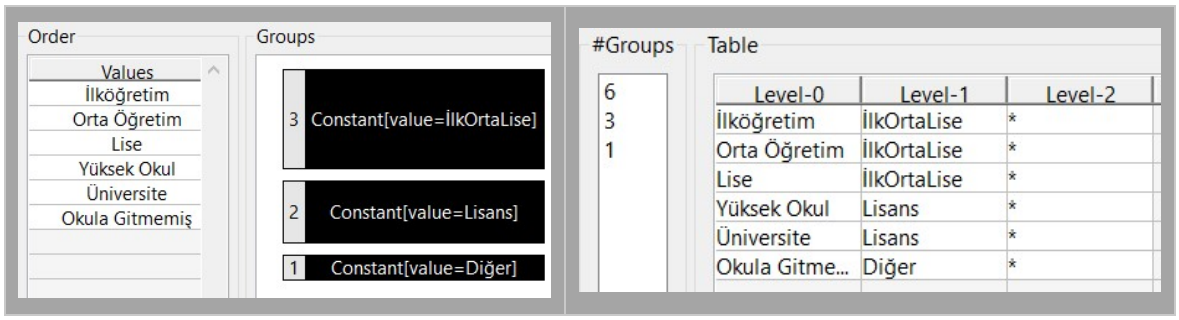
DEMOGRAFİK veri kümesinin bir diğer kategorik özniteliği olan etnik özniteliğine ait genelleştirme hiyerarşisi Şekil 4-14'de verilmiştir



Şekil 4-14 DEMOGRAFİK Etnik özniteliği genelleştirme hiyerarşisi

Şekil 4-14'de verilen DEMOGRAFİK veri kümesi kategorik etnik özniteliğinin değerleri {İspanyol/Latin, İspanyol/Latin Olmayan } olup baskılama dahil iki seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer iki grup içinde yer alırken ikinci seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

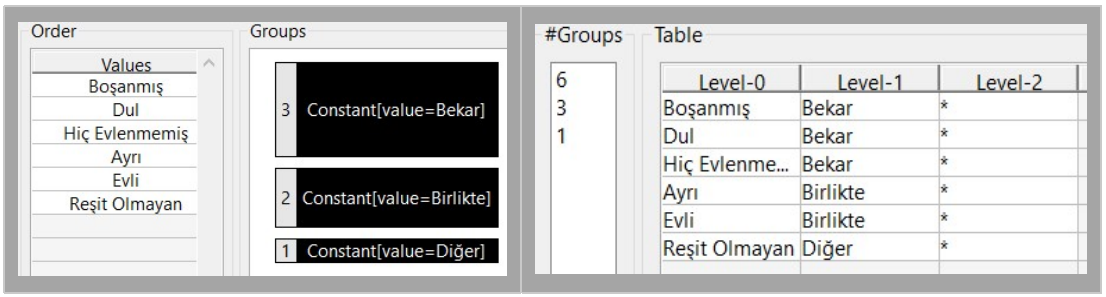
DEMOGRAFİK veri kümesinin bir diğer kategorik özniteliği olan eğitim özniteliğine ait genelleştirme hiyerarşisi Şekil 4-15'da verilmiştir



Şekil 4-15 DEMOGRAFİK Eğitim özniteliği genelleştirme hiyerarşisi

Şekil 4-15'da verilen DEMOGRAFİK veri kümesi kategorik eğitim özniteliğinin değerleri {İlköğretim, Ortaöğretim, Lise, Yüksekokul, Üniversite, Okula Gitmemiş } olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 6 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {İlkortalise, Lisans, Diğer} üç grupta yer almıştır. Son seviyede ise öznitelik tamamen baskılanarak tek bir grupta (*) toplanmıştır.

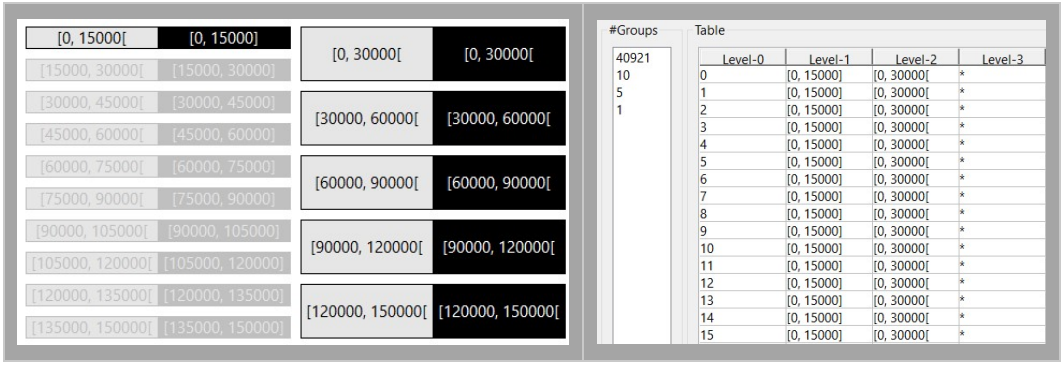
DEMOGRAFİK veri kümesinin bir diğer kategorik özniteliği olan medeni durum özniteliğine ait genelleştirme hiyerarşisi Şekil 4-16'de verilmiştir.



Şekil 4-16 DEMOGRAFİK Medeni durum özniteliği genelleştirme hiyerarşisi

Şekil 4-16'de verilen DEMOGRAFİK veri kümesi medeni durum özniteliğinin değerleri {Boşanmış, Dul, Hiç Evlenmemiş, Ayrı, Evli, Reşit olmayan} olup baskılama dahil üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 6 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {Bekâr, Birlikte, Diğer} üç grupta yer almıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) yer almıştır.

DEMOGRAFİK veri kümesi için bir diğer sayısal öznitelik olan toplam gelir özniteliğine ait genelleştirme hiyerarşisi Şekil 4-17'de verilmiştir



Şekil 4-17 DEMOGRAFİK Toplam gelir özniteliği genelleştirme hiyerarşisi

Şekil 4-17'de DEMOGRAFİK veri kümesi için verilen sayısal toplam gelir özniteliğinin değer aralığı 0 ile 150.000 arasında olup baskılama dahil beş seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 40921 grup içinde, ikinci seviyede değerler 10 grup içerisinde üçüncü seviyede 5 grup içerisinde toplanmıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

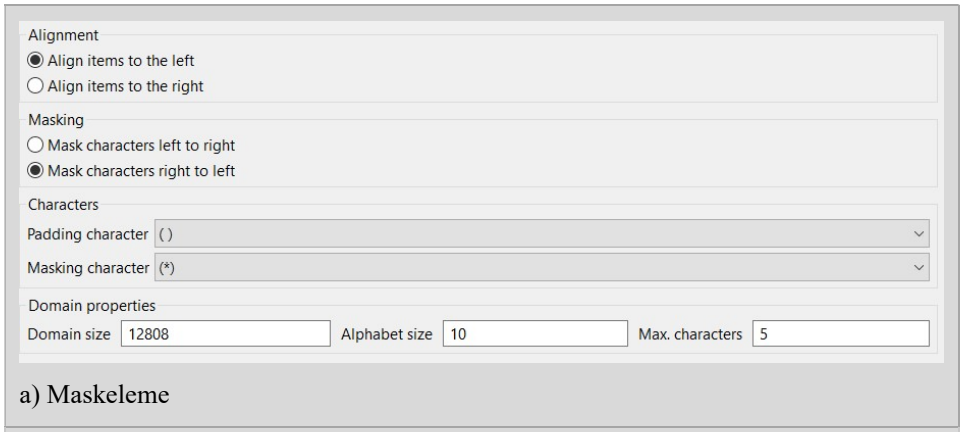
4.1.3 CUP

CUP veri kümesinin sınıflandırılmış öznitelikleri genelleştirme yükseklikleriyle birlikte Çizelge 4-3'de verilmiştir.

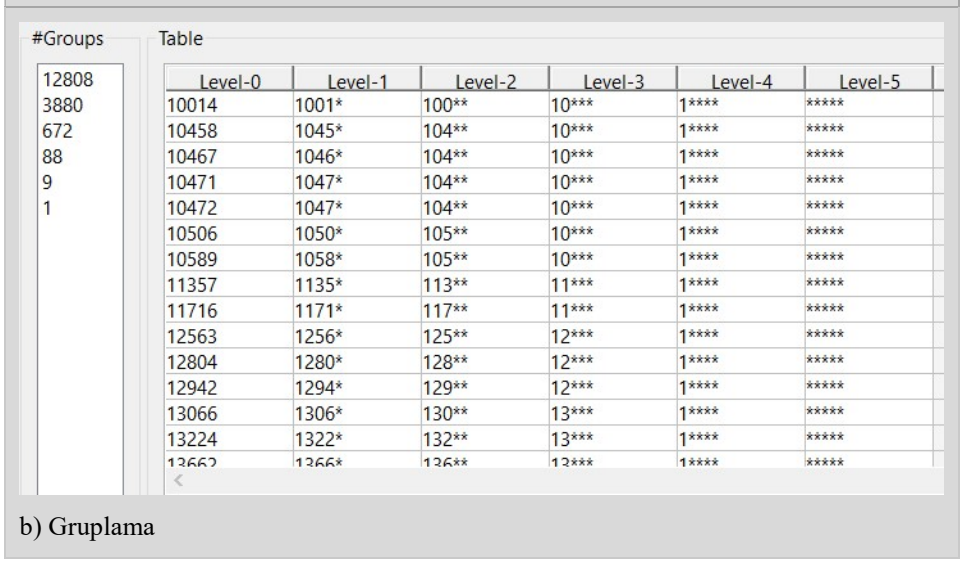
Çizelge 4-3 CUP veri kümesi öznitelikleri

Veri Kümesi Adı	QID (Genelleştirme Yüksekliği)	SA
T_{CUP}	posta kodu (6), yaş(5), cinsiyet (2)	gelir

CUP veri kümesinin alfanümerik tipindeki posta kodu özniteliğine ait genelleştirme hiyerarşisi Şekil 4-18'de verilmiştir



a) Maskeleme



b) Gruplama

Şekil 4-18 CUP Posta kodu özniteliği genelleştirme hiyerarşisi

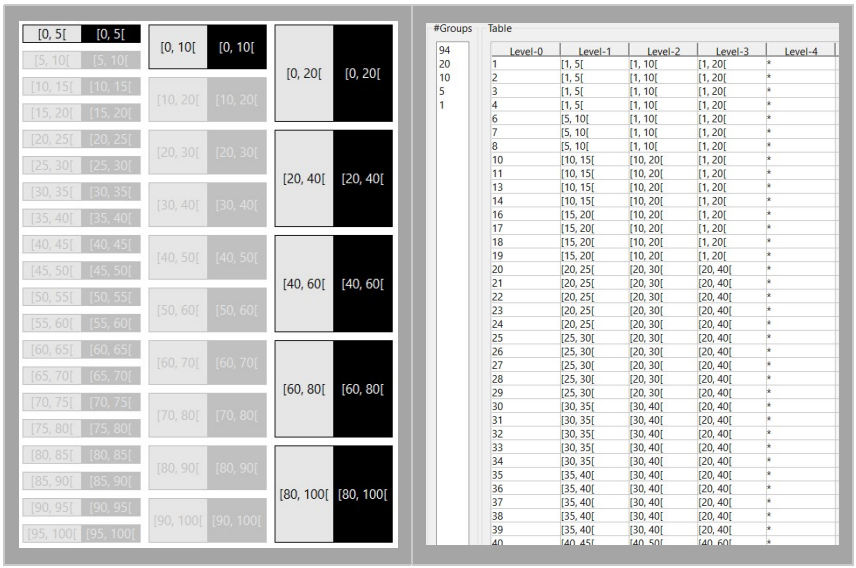
Şekil 4-18 (a)'da gösterilen 12808 farklı değer alan alfanümerik posta kodu özniteliği için baskılama dahil altı seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 12808 grup içinde yer alırken ikinci seviyede 3880, üçüncü seviyede 672, dördüncü seviyede 88, beşinci seviyede 9, altıncı seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

CUP veri kümesinin kategorik cinsiyet özniteliğine ait genelleştirme hiyerarşisi Şekil 4-19'da verilmiştir

Şekil 4-19 CUP Cinsiyet özniteliği genelleştirme hiyerarşisi

Şekil 4-19'da verilen CUP veri kümesi kategorik cinsiyet özniteliğinin değerleri {F, M} olup baskılama dahil iki seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer iki grup içinde yer alırken ikinci seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

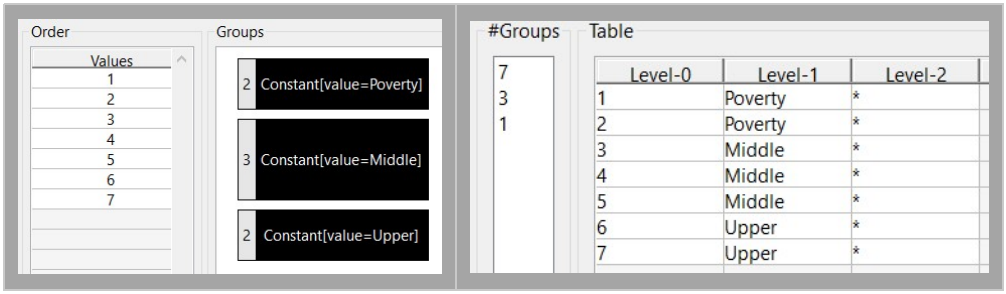
DEMOGRAFİK veri kümesindeki sayısal yaş özniteliğine ait genelleştirme hiyerarşisi Şekil 4-20'de verilmiştir.



Şekil 4-20 CUP Yaş özniteliği genelleştirme hiyerarşisi

Şekil 4-20'de verilen sayısal yaş özniteliğinin değer aralığı 1 ile 94 arasında olup baskılama dahil beş seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 94 grup içinde yer alırken ikinci seviyede 20 grup içerisinde ardışık 5 değer yer almaktadır. Üçüncü seviyede ardışık 10 değer 10 grupta toplanırken, 4. seviyede ardışık 20 değer 5 grupta toplanmıştır. Son seviyede ise öznitelik değerleri tamamen baskılanarak tek bir grupta (*) toplanmıştır.

CUP veri kümesinin bir diğer kategorik özniteliği olan gelir özniteliğine ait genelleştirme hiyerarşisi Şekil 4-21'de verilmiştir.



Şekil 4-21 CUP Gelir özniteliği genelleştirme hiyerarşisi

Şekil 4-21'de verilen CUP veri kümesi kategorik eğitim özniteliğinin değerleri {1, 2, 3, 4, 5, 6, 7 } olup üç seviyeli genelleştirme uygulanmıştır. İlk seviyede her bir değer 7 grup içinde yer alırken ikinci seviyede ise öznitelik değerleri {Poverty, Middle, Upper }üç grup içinde yer almıştır. Son seviyede ise öznitelikler tamamen baskılanarak tek bir grupta (*) toplanmıştır.

4.2 Deneysel Değerlendirmeler

Bu bölümde veri kümeleri üzerinde ρ-Kazanım modelinin uygulanmasını sağlayan deneysel çalışmalar yapılmıştır. Deneysel çalışmalar ρ-Kazanım modelinin (k), (k,ℓ), (k,t) (k,ℓ,t) kombinasyonları üzerinde yapılmıştır. Deneysel çalışmalarda veri faydası ölçümünde eşlenik sınıflar üzerinden ölçüm yapan ayırt edilebilirlik (AEM) ile eşlenik sınıflar ortalaması (ESO) metrikleri kullanılmıştır. ρ-Kazanım modelinin başarımının ölçülmesinde ρ-Kazanım karar kuralı uygulanmıştır. Eşlenik sınıfa dayalı ölçümlerde ESO ve AEM metriklerinin ölçüm hassasiyetlerini karşılaştırmak amacıyla veri kümeleri üzerinde deneyler yapılarak hangi metriğin daha başarılı olduğu araştırılmıştır. Ölçümlerde;

TF (UEC): Faydalı eşlenik sınıf sayılarının toplamı,

TA (OEC) : Aykırı eşlenik sınıf içindeki toplam aykırı kayıt sayısını,

göstermektedir.

Bu amaçla k-Anonimlik modelinin ADULT, DEMOGRAFİK ve CUP veri kümeleri üzerinde 7 farklı k değeri (2, 3, 4, 5, 10, 20, 50, 100) kullanılarak ESO metriği ile ölçülen değerler Çizelge 4-4'de gösterilmiştir.

Çizelge 4-4 ESO ölçüm değerleri

Parametreler	ADULT	CUP	DEMOGRAFİK
k=2	TF(UEC)=5014 TA(OEC)=8009 ESO=6,01	TF(UEC)=13944 TA(OEC)=16107 ESO=4,26	TF(UEC)=19431 TA(OEC)=15011 ESO=3,52
k=3	TF(UEC)=3075 TA(OEC)=8401 ESO=9,80	TF(UEC)=8134 TA(OEC)=17152 ESO=7,31	TF(UEC)=13274 TA(OEC)=9815 ESO=5,16
k=4	TF(UEC)=2297 TA(OEC)=8661 ESO=13,12	TF(UEC)= 5893 TA(OEC)=21507 ESO=10,08	TF(UEC)=9081 TA(OEC)=15195 ESO=7,54
k=5	TF(UEC)=1817 TA(OEC)=7549 ESO=16,59	TF(UEC)=4448 TA(OEC)=7264 ESO=13,36	TF(UEC)=8079 TA(OEC)=6304 ESO=8,48
k=10	TF(UEC)=951 TA(OEC)=8167 ESO=31,68	TF(UEC)=2201 TA(OEC)=8348 ESO=27,00	TF(UEC)=3884 TA(OEC)=14496 ESO=17,64
k=20	TF(UEC)=536 TA(OEC)=9662 ESO=56,16	TF(UEC)=1116 TA(OEC)=19109 ESO=53,23	TF(UEC)=1990 TA(OEC)=178 ESO=34,43
k=50	TF(UEC)=250 TA(OEC)=6668 ESO=120,16	TF(UEC)=507 TA(OEC)=12617 ESO=117,06	TF(UEC)=990 TA(OEC)=468 ESO=69,18
k=100	TF(UEC)=146 TA(OEC)=8374 ESO=205,18	TF(UEC)=262 TA(OEC)=16170 ESO=226,11	TF(UEC)=396 TA(OEC)=15917 ESO=172,71

Eşlenik sınıflara göre ölçüm yapan AEM metriğinin ADULT, DEMOGRAFİK ve CUP veri kümeleri üzerinde 7 farklı k değeri (2, 3, 4, 5, 10, 20, 50, 100) kullanılarak ölçülen değerleri Çizelge 4-5'de gösterilmiştir.

Çizelge 4-5 AEM ölçüm değerleri

Parametreler	ADULT	CUP	DEMOGRAFİK
k=2	TF(UEC)=665 TA(OEC)=62 AEM=45,28	TF(UEC)=1059 TA(OEC)=67 AEM=56,10	TF(UEC)=2389 TA(OEC)=79 AEM=28,66
k=3	TF(UEC)=371 TA(OEC)=55 AEM=81,08	TF(UEC)=1037 TA(OEC)=111 AEM=57,29	TF(UEC)=2300 TA(OEC)=257 AEM=29,79
k=4	TF(UEC)=361 TA(OEC)=85 AEM=83,32	TF(UEC)= 964 TA(OEC)=110 AEM=61,62	TF(UEC)=2233 TA(OEC)=458 AEM=30,69
k=5	TF(UEC)=356 TA(OEC)=105 AEM=84,48	TF(UEC)=952 TA(OEC)=158 AEM=62,40	TF(UEC)=2187 TA(OEC)=646 AEM=31,51
k=10	TF(UEC)=180 TA(OEC)=121 AEM=166,64	TF(UEC)=446 TA(OEC)=85 AEM=133,03	TF(UEC)=1788 TA(OEC)=3533 AEM=38,32
k=20	TF(UEC)=164 TA(OEC)=335 AEM=182,8	TF(UEC)=223 TA(OEC)=85 AEM=265,48	TF(UEC)=1018 TA(OEC)=14090 AEM=67,28
k=50	TF(UEC)=55 TA(OEC)=330 AEM=538,60	TF(UEC)=112 TA(OEC)=29 AEM=526,26	TF(UEC)=410 TA(OEC)=33283 AEM=166,82
k=100	TF(UEC)=51 TA(OEC)=618 AEM=580,03	TF(UEC)=112 TA(OEC)=29 AEM=526,26	TF(UEC)=163 TA(OEC)=50065 AEM=418,08

Çizelge 4-4 ve Çizelge 4-5'deki ölçülen değerler eşlenik sınıf sayıları açısından karşılaştırıldığında ESO metriğinin daha hassas ölçümler yaptığı görülmüştür. Ayrıca ESO metriği ile veri faydasının ölçülmesinde en uygun sonuçların k=5 değerinde aldığı

görülmüştür. Takip eden alt bölümlerde ρ-Kazanım modelinin (k), (k,ℓ), (k,t) (k,ℓ,t) kombinasyonları üzerinde ADULT veri kümesi kullanılarak deneysel çalışmalar yapılmıştır.

4.2.1 ρ-Kazanım (k) Deneyi

ρ-Kazanım (k) deneyiyle, ρ-Kazanım modelinin k-Anonimlik modeline uygulanmasıyla veri faydası ile mahremiyet risklerinin nasıl değiştiği gösterilmiştir. Model parametreleri olarak tez çalışması kapsamında yapılan araştırmalarda en iyi sonuçları veren k=5 ve ρ=2 değerleri seçilmiştir. Veri faydasının ölçümünde eşlenik sınıf ortalaması metriği, mahremiyet risklerinin değerlendirilmesinde savcı yaklaşımı kullanılmıştır.

Deney çalışmaları kapsamında ADULT veri kümesi üzerinde *2-Kazanım (5)* ile 5-Anonimlik sonuçları veri faydası ve mahremiyet riskleri açısından karşılaştırılmıştır. 5-Anonimlik durumundaki (ρ'=0) veri faydası sonuçları Şekil 4-22'de verilmiştir.

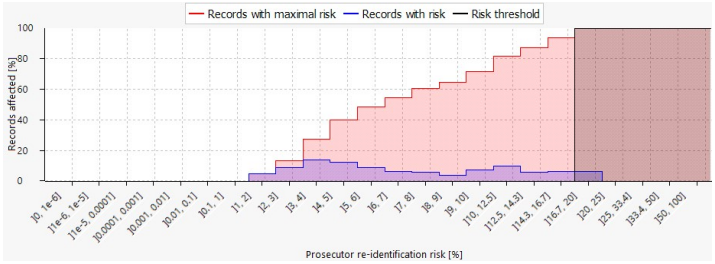
Measure	Including outliers	Excluding outliers
Average class size	16.59076 (0.05501%)	12.44524 (0.05504%)
Maximal class size	7549 (25.02818%)	65 (0.28745%)
Minimal class size	5 (0.01658%)	5 (0.02211%)
Number of classes	1818	1817
Number of records	30162	22613 (74.97182%)
Suppressed records	7549 (25.02818%)	0

Şekil 4-22 (k=5, ρ'=0) durumunda veri faydası sonuçları

Şekil 4-22 incelendiğinde, en az 5, en fazla 7549 kayıt içeren 1818 adet eşlenik sınıf olduğu görülmüştür. Veri faydası açısından yapılan aykırı kayıtları dahil ederek yapılan ölçümde ESO değeri 16,59 olarak bulunmuştur. ρ'=0 durumunda veri faydasına katkı sağlayan 1817 adet UEC bulunmuştur. 22613 kaydı içeren 1817 adet UEC yayınlanacak T* kümesine aktarılmıştır. Veri faydası olmayan ve tamamen bastırılan veri kümesindeki kayıtların %25,02'sini oluşturan OEC içerisinde 7549 kayıt olduğu görülmüştür. Veri faydası sonuçlarına göre ρ'=0 durumundaki aykırı kayıtlı veri faydası 83,41 (100-16,59) olarak bulunmuştur.

ρ'=0 durumundaki risk tahminleri ve savcı risk grafiği Şekil 4-23'de verilmiştir.

Measure	Value [%]
Lowest prosecutor risk	1.53846%
Records affected by lowest risk	0.28745%
Average prosecutor risk	8.0352%
Highest prosecutor risk	20%
Records affected by highest risk	6.43435%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	8.0352%



a) Tüm riskler b) Savcı riski

Şekil 4-23 (k=5, ρ'=0) durumunda risk sonuçları

Şekil 4-23 (a)' de verilen ρ'=0 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 8,03 olduğu görülmektedir. Bu çalışmada saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınacaktır.

ρ'=0 durumunda savcı risk grafiği Şekil 4-23 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %6,43'nün yüksek risk aralığında (16,7-20) olduğu ve başlangıç riskinin ise %20 olduğu görülür. Savcı yaklaşımına göre hesaplanan %20 değeri ρ'=0 durumundaki risk olarak kabul edilmiştir. ρ'=0 durumunda tamamen bastırılan veri faydası olmayan 7549 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ'=0 durumundan ρ=1 durumuna geçilmiştir.

ρ=1 durumu için 7549 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ilk iterasyonunda ρ=1 durumunda elde edilen veri faydası sonuçları Şekil 4-24'de gösterilmiştir.

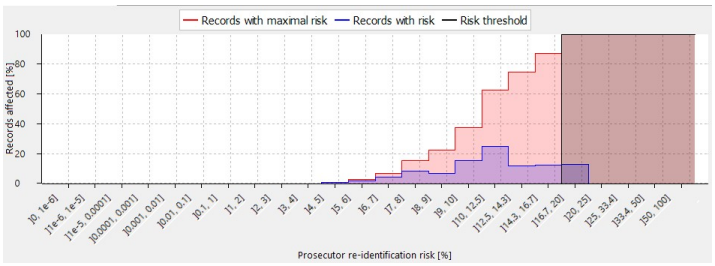
Measure	Including outliers	Excluding outliers
Average class size	9.99868 (0.13245%)	8.05172 (0.13263%)
Maximal class size	1478 (19.57875%)	21 (0.34591%)
Minimal class size	5 (0.06623%)	5 (0.08236%)
Number of classes	755	754
Number of records	7549	6071 (80.42125%)
Suppressed records	1478 (19.57875%)	0

Şekil 4-24 (k=5, ρ=1) durumunda veri faydası sonuçları

Şekil 4-24 incelendiğinde ρ=1 durumunda elde edilen bulgularda en az 5, en fazla 1478 kayıt içeren 755 adet eşlenik sınıf olduğu görülmüştür. ρ=1 durumunda veri faydasına katkı sağlayan 754 adet UEC bulunmuştur. 6071 kaydı içeren 754 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=1 durumunda OEC kayıtlarının %80,42'sinin faydaya dönüştüğü tespit edilmiştir.

ρ=1 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-25'de verilmiştir.

Measure	Value [%]
Lowest prosecutor risk	4.7619%
Records affected by lowest risk	0.34591%
Average prosecutor risk	12.4197%
Highest prosecutor risk	20%
Records affected by highest risk	12.93032%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	12.4197%



a) Tüm riskler b) Savcı riski

Şekil 4-25 (k=5, ρ=1) durumunda risk sonuçları

Şekil 4-25 (a)' da verilen ρ=1 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 12.41 olduğu görülmektedir. Saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınacaktır. ρ=1 durumunda savcı risk grafiği Şekil 4-25 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %12,93'nün yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=1 durumunda en yüksek riskin değişmediği görülmüştür. ρ=1 durumunda tamamen bastırılan veri faydası olmayan 1478 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ=1 durumundan ρ=2 durumuna geçilmiştir.

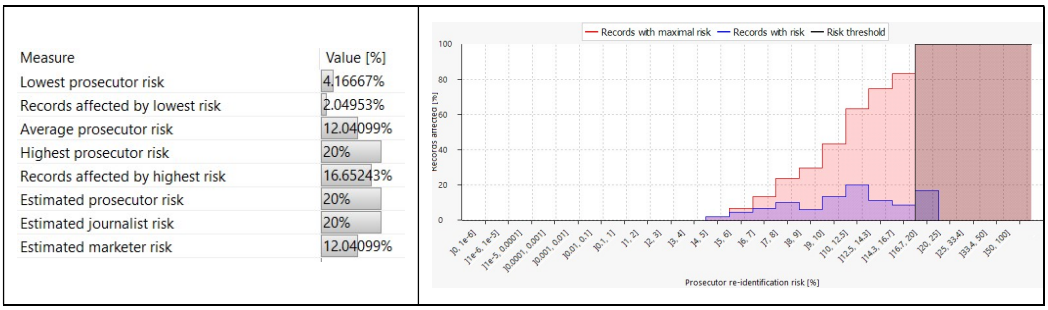
ρ=2 durumu için 1478 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ikinci iterasyonunda ρ=2 durumunda elde edilen veri faydası sonuçları Şekil 4-26'da gösterilmiştir.

Measure	Including outliers	Excluding outliers
Average class size	10.40845 (0.70423%)	8.30496 (0.70922%)
Maximal class size	307 (20.77131%)	24 (2.04953%)
Minimal class size	5 (0.33829%)	5 (0.42699%)
Number of classes	142	141
Number of records	1478	1171 (79.22869%)
Suppressed records	307 (20.77131%)	0

Şekil 4-26 (k=5, ρ=2) durumunda veri faydası sonuçları

Şekil 4-26 incelendiğinde ρ=2 durumunda elde edilen bulgularda en az 5, en fazla 307 kayıt içeren 142 adet eşlenik sınıf olduğu görülmüştür. ρ=2 durumunda veri faydasına katkı sağlayan 141 adet UEC bulunmuştur. 1171 kaydı içeren 141 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=2 durumunda OEC kayıtlarının %79,22'sinin faydaya dönüştüğü tespit edilmiştir.

ρ=2 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-27'de verilmiştir.



a) Tüm riskler b) Savcı riski

Şekil 4-27 (k=5, ρ=2) durumunda risk sonuçları

Şekil 4-27 (a)' da verilen ρ=2 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 12,04 olduğu görülmektedir. Saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınmıştır. ρ=2 durumunda savcı risk grafiği Şekil 4-27 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %16,65'nin yüksek

risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=2 durumunda en yüksek riskin değişmediği görülmüştür.

ρ-Kazanım (k) deneyi sonucunda veri faydası, mahremiyet koruması ve UEC / OEC karşılaştırmaları Çizelge 4-6'da gösterilmiştir.

Çizelge 4-6 ρ-Kazanım (k) değerlendirmesi

Model	ESO	UEC	OEC
k-Anonimlik (k=5)	16,59	%74,97	%25,03
ρ-Kazanım (k=5, ρ=2)	11,11	%98,98	%1,02

Çizelge 4-6'da k-Anonimlik modeli ile ρ-Kazanım (k) ölçüm sonuçları karşılaştırılmıştır. k-Anonimlik modelinde aykırı kayıtların dahil olduğu durumda veri kaybı 16,59 olarak ölçülmüştür. Tüm kayıtların içerisinde aykırı kayıt sayıların oranı %25,03 olarak ölçülmüştür. ρ-Kazanım (k) modelinin aykırı kayıtların dahil olduğu durumda veri kaybı 11,11 olarak ölçülmüştür. ρ-Kazanım modelinin uygulanmasıyla veride yaşanan kayıp %33,03 oranında azalmıştır.

Bir diğer karşılaştırma UEC ve OEC oranları üzerinden yapılmıştır. k-Anonimlik modelinde UEC oranı %74,97 OEC oranı %25,03 ρ-Kazanım (k) modelinde UEC oranı %98,98, OEC oranı %1,02 olarak ölçülmüştür. ρ-Kazanım (k) modelinin uygulanmasıyla aykırı kayıt sayıların oranı %24,01 oranında azalarak veri faydasına dönüştüğü tespit edilmiştir.

ρ-Kazanım modelinin k-Anonimlik modelinin oluşturduğu OEC üzerinde yaptığı fayda arttırıcı işlemler sayesinde veri kaybını azalttığı dolayısıyla veri faydasını arttırdığı bu deneysel çalışmayla gösterilmiştir. Veri faydasındaki iyileşmeye rağmen tüm iterasyonlarda en yüksek mahremiyet risk ölçümünün %20 olarak kaldığı gözlemlenmiş k-Anonimlik mahremiyet risklerine kıyasla risk değerinin değişmediği tespit edilmiştir.

4.2.2 ρ-Kazanım (k,ℓ) Deneyi

ρ-Kazanım (k,ℓ) deneyiyle, ρ-Kazanım modelinin (k,ℓ)-Anonimlik modeline uygulanmasıyla veri faydası ile mahremiyet risklerinin nasıl değiştiği gösterilmiştir. Model parametreleri olarak tez çalışması kapsamında yapılan araştırmalarda en iyi sonuçları veren

k=5, ℓ=2 ve ρ=2 değerleri seçilmiştir. Veri faydasının ölçümünde eşlenik sınıf ortalaması metriği, mahremiyet risklerinin değerlendirilmesinde savcı yaklaşımı kullanılmıştır.

Deney çalışmaları kapsamında ADULT veri kümesi üzerinde *2-Kazanım (5,2)* ile (5,2)-Anonimlik sonuçları veri faydası ve mahremiyet riskleri açısından karşılaştırılmıştır. (5,2)-Anonimlik durumundaki (ρ'=0) veri faydası sonuçları Şekil 4-28'de verilmiştir.

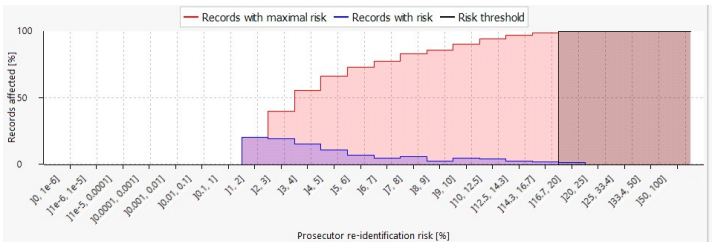
Measure	Including outliers	Excluding outliers
Average class size	29.19845 (0.09681%)	19.93702 (0.0969%)
Maximal class size	9587 (31.78503%)	99 (0.48117%)
Minimal class size	5 (0.01658%)	5 (0.0243%)
Number of classes	1033	1032
Number of records	30162	20575 (68.21497%)
Suppressed records	9587 (31.78503%)	0

Şekil 4-28 (k=5, ℓ=2, ρ'=0) durumunda veri faydası sonuçları

Şekil 4-28 incelendiğinde, en az 5, en fazla 9587 kayıt içeren 1033 adet eşlenik sınıf olduğu görülmüştür. Veri faydası açısından aykırı kayıtları dahil ederek yapılan ölçümde ESO değeri 29,19 olarak bulunmuştur. ρ'=0 durumunda veri faydasına katkı sağlayan 1032 adet UEC bulunmuştur. 20575 kaydı içeren 1032 adet UEC yayınlanacak T* kümesine aktarılmıştır. Veri faydası olmayan ve tamamen bastırılan veri kümesindeki kayıtların %31,78'ni oluşturan OEC içerisinde 9587 kayıt olduğu görülmüştür. Veri faydası sonuçlarına göre ρ'=0 durumundaki aykırı kayıtlı veri faydası 70,81 (100-29,19) olarak bulunmuştur.

ρ'=0 durumundaki risk tahminleri ve savcı risk grafiği Şekil 4-29'da verilmiştir.

Measure	Value [%]
Lowest prosecutor risk	1.0101%
Records affected by lowest risk	0.48117%
Average prosecutor risk	5.0158%
Highest prosecutor risk	20%
Records affected by highest risk	1.45808%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	5.0158%



a) Tüm riskler b) Savcı riski

Şekil 4-29 (k=5, ℓ=2, ρ'=0) durumunda risk sonuçları

Şekil 4-29 (a)' da verilen ρ'=0 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 5,01 olduğu görülmektedir. Bu çalışmada saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınacaktır.

ρ'=0 durumunda savcı risk grafiği Şekil 4-29 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %1,45'nin yüksek risk aralığında (16,7-20) olduğu ve başlangıç riskinin ise %20 olduğu görülmüştür. Savcı yaklaşımına göre hesaplanan %20 değeri ρ'=0 durumundaki risk olarak kabul edilmiştir. ρ'=0 durumunda tamamen bastırılan veri faydası olmayan 9587 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ'=0 durumundan ρ=1 durumuna geçilmiştir.

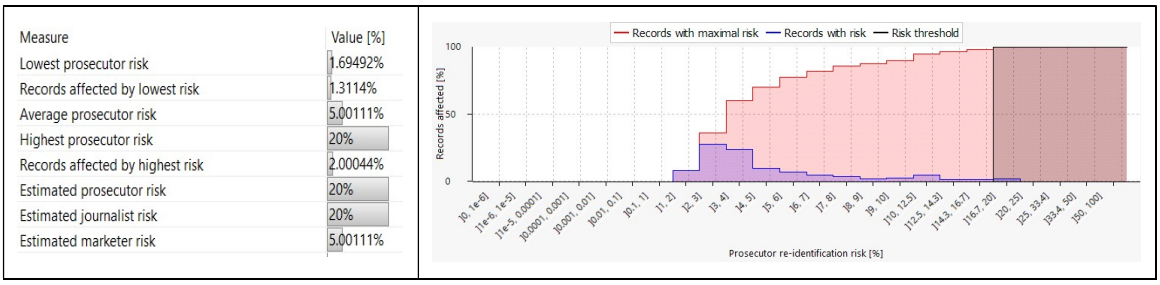
ρ=1 durumu için 9587 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ilk iterasyonunda ρ=1 durumunda elde edilen veri faydası sonuçları Şekil 4-24Şekil 4-30'da gösterilmiştir.

Measure	Including outliers	Excluding outliers
Average class size	42.42035 (0.44248%)	19.99556 (0.44444%)
Maximal class size	5088 (53.07187%)	59 (1.3114%)
Minimal class size	5 (0.05215%)	5 (0.11114%)
Number of classes	226	225
Number of records	9587	4499 (46.92813%)
Suppressed records	5088 (53.07187%)	0

Şekil 4-30 (k=5, ℓ=2, ρ=1) durumunda veri faydası sonuçları

Şekil 4-30 Şekil 4-24incelendiğinde ρ=1 durumunda elde edilen bulgularda en az 5, en fazla 5088 kayıt içeren 226 adet eşlenik sınıf olduğu görülmüştür. ρ=1 durumunda veri faydasına katkı sağlayan 225 adet UEC bulunmuştur. 4499 kaydı içeren 225 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=1 durumunda OEC kayıtlarının %46,92'sinin faydaya dönüştüğü tespit edilmiştir.

ρ=1 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-31'de verilmiştir.



a) Tüm riskler b) Savcı riski

Şekil 4-31 (k=5, ℓ=2, ρ=1) durumunda risk sonuçları

Şekil 4-31 (a)' da verilen ρ=1 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 5,00 olduğu görülmektedir. Saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınacaktır. ρ=1 durumunda savcı risk grafiği Şekil 4-31 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %2,00'nın yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=1 durumunda en yüksek riskin değişmediği görülmüştür. ρ=1 durumunda tamamen bastırılan veri faydası olmayan 5088 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ=1 durumundan ρ=2 durumuna geçilmiştir.

ρ=2 durumu için 5088 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ikinci iterasyonunda ρ=2 durumunda elde edilen veri faydası sonuçları Şekil 4-32'de gösterilmiştir.

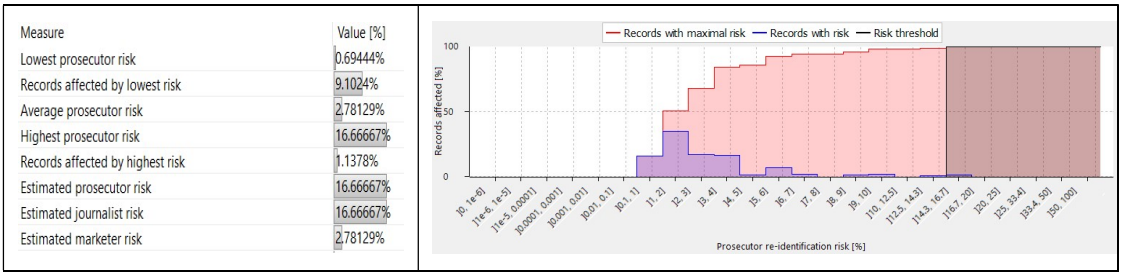
Measure	Including outliers	Excluding outliers
Average class size	113.06667 (2.22222%)	35.95455 (2.27273%)
Maximal class size	3506 (68.90723%)	144 (9.1024%)
Minimal class size	6 (0.11792%)	6 (0.37927%)
Number of classes	45	44
Number of records	5088	1582 (31.09277%)
Suppressed records	3506 (68.90723%)	0

Şekil 4-32 (k=5, ℓ=2, ρ=2) durumunda veri faydası sonuçları

Şekil 4-32 incelendiğinde ρ=2 durumunda elde edilen bulgularda en az 5, en fazla 3506 kayıt içeren 45 adet eşlenik sınıf olduğu görülmüştür. ρ=2 durumunda veri faydasına katkı

sağlayan 44 adet UEC bulunmuştur. 1582 kaydı içeren 44 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=2 durumunda OEC kayıtlarının %31,09'nun faydaya dönüştüğü tespit edilmiştir.

ρ=2 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-33'de verilmiştir.



a) Tüm riskler b) Savcı riski

Şekil 4-33 (k=5, ℓ=2, ρ=2) durumunda risk sonuçları

Şekil 4-33 (a)' da verilen ρ=2 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %16,66, %16,66 ve % 2,78 olduğu görülmektedir. Saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınmıştır. ρ=2 durumunda savcı risk grafiği Şekil 4-33 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %1,13'nün yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=2 durumunda en yüksek riskin azaldığı görülmüştür.

ρ-Kazanım (k,ℓ) deneyi sonucunda veri faydası, UEC ve OEC karşılaştırmaları Çizelge 4-7'de gösterilmiştir.

Çizelge 4-7 ρ-Kazanım (k, ℓ) değerlendirmesi

Model	ESO	UEC	OEC
(k=5, ℓ =2)-Anonimlik	29,19	%68,25	%31,75
ρ-Kazanım (k=5, ℓ =2, ρ=2)	23,18	%88,37	%11,63

Çizelge 4-7'de (k,ℓ)-Anonimlik modeli ile ρ-Kazanım (k,ℓ) ölçüm sonuçları karşılaştırılmıştır. (k,ℓ)-Anonimlik modelinde aykırı kayıtların dahil olduğu durumda veri kaybı 29,19 olarak ölçülmüştür. Tüm kayıtların içerisinde aykırı kayıt sayılarının oranı %31,78 olarak ölçülmüştür. ρ-Kazanım (k,ℓ) modelinin aykırı kayıtların dahil olduğu durumda veri kaybı 23,18 olarak ölçülmüştür. ρ-Kazanım modelinin uygulanmasıyla veride yaşanan kayıp %20,76 oranında azalmıştır.

Bir diğer karşılaştırma UEC ve OEC oranlarıyla yapılmıştır. (k,ℓ)-Anonimlik modelinde UEC oranı %68,25, OEC oranı %31,75 ρ-Kazanım (k,ℓ) modelinde UEC oranı %88,37, OEC oranı %11,63 olarak ölçülmüştür. ρ-Kazanım (k,ℓ) modelinin uygulanmasıyla aykırı kayıt sayıların oranı %20,12 oranında azalarak veri faydasına dönüştüğü tespit edilmiştir.

ρ-Kazanım modelinin OEC üzerinde yaptığı fayda arttırıcı işlemler sayesinde veri kaybının azaldığı dolayısıyla veri faydasının arttığı bu deneysel çalışmayla gösterilmiştir. Veri faydasındaki iyileşmeye rağmen tüm iterasyonlarda en yüksek mahremiyet risk ölçümünün %20 olarak kaldığı hatta son iterasyonda %16,66'ya düştüğü tespit edilmiştir.

4.2.3 ρ-Kazanım (k,t) Deneyi

ρ-Kazanım (k,t) deneyiyle, ρ-Kazanım modelinin (k,t)-Anonimlik modeline uygulanmasıyla veri faydası ile mahremiyet risklerinin nasıl değiştiği gösterilmiştir. Model parametreleri olarak tez çalışması kapsamında yapılan araştırmalarda en iyi sonuçları veren k=5, t=0,2 ve ρ=2 değerleri seçilmiştir. Veri faydasının ölçümünde eşlenik sınıf ortalaması metriği, mahremiyet risklerinin değerlendirilmesinde savcı yaklaşımı kullanılmıştır.

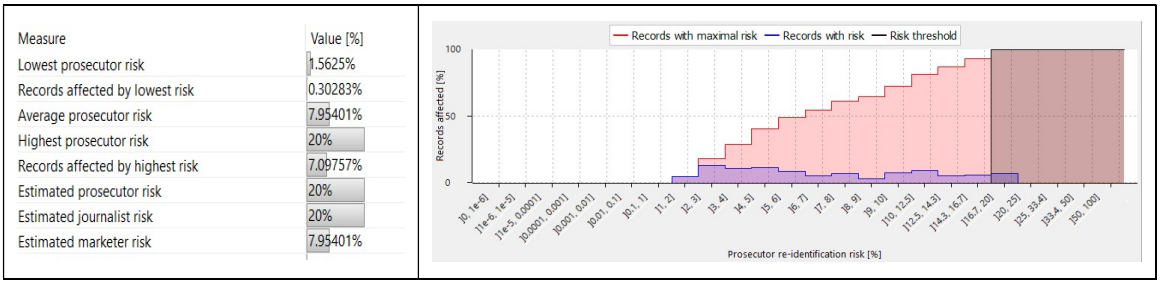
ADULT veri kümesi üzerinde ρ'=0 durumunda veri faydası sonuçları Şekil 4-34'de verilmiştir.

Measure	Including outliers	Excluding outliers
Average class size	17.93222 (0.05945%)	12.57228 (0.05949%)
Maximal class size	9028 (29.9317%)	64 (0.30283%)
Minimal class size	5 (0.01658%)	5 (0.02366%)
Number of classes	1682	1681
Number of records	30162	21134 (70.0683%)
Suppressed records	9028 (29.9317%)	0

Şekil 4-34 (k=5, t=0,2, ρ'=0) durumunda veri faydası sonuçları

Şekil 4-34 incelendiğinde, en az 5, en fazla 9028 kayıt içeren 1682 adet eşlenik sınıf olduğu görülmüştür. Veri faydası açısından aykırı kayıtları dahil ederek yapılan ölçümde ESO değeri 17,93 olarak bulunmuştur. ρ'=0 durumunda veri faydasına katkı sağlayan 1681 adet UEC bulunmuştur. 21134 kaydı içeren 1681 adet UEC yayınlanacak T* kümesine aktarılmıştır. Veri faydası olmayan ve tamamen bastırılan veri kümesindeki kayıtların %29,93'nü oluşturan OEC içerisinde 9028 kayıt olduğu görülmüştür. Veri faydası sonuçlarına göre ρ'=0 durumundaki aykırı kayıtlı veri faydası 82,07 (100-17,93) olarak bulunmuştur.

ρ'=0 durumundaki risk tahminleri ve savcı risk grafiği Şekil 4-35'de verilmiştir.



a) Tüm riskler b) Savcı riski

Şekil 4-35 (k=5, t=0,2, ρ'=0) durumunda risk sonuçları

Şekil 4-35 (a)' da verilen ρ'=0 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 7,95 olduğu görülmektedir. Bu çalışmada saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınmıştır.

ρ'=0 durumunda savcı risk grafiği Şekil 4-35Şekil 4-29 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %7,09'nun yüksek risk aralığında (16,7-20) olduğu ve başlangıç riskinin ise %20 olduğu görülmüştür. Savcı yaklaşımına göre hesaplanan %20 değeri ρ'=0 durumundaki risk olarak kabul edilmiştir. ρ'=0 durumunda tamamen bastırılan veri faydası olmayan 9028 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ'=0 durumundan ρ=1 durumuna geçilmiştir.

ρ=1 durumu için 9028 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ilk iterasyonunda ρ=1 durumunda elde edilen veri faydası sonuçları Şekil 4-24Şekil 4-36'da gösterilmiştir.

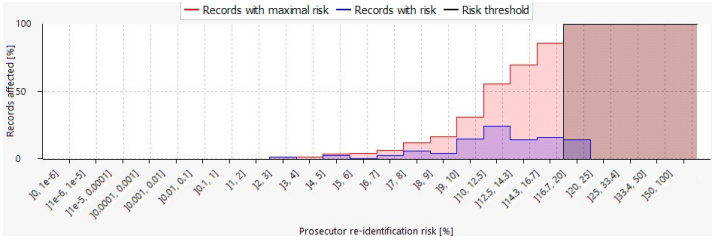
Measure	Including outliers	Excluding outliers
Average class size	14.8243 (0.1642%)	7.72533 (0.16447%)
Maximal class size	4331 (47.97297%)	38 (0.80903%)
Minimal class size	5 (0.05538%)	5 (0.10645%)
Number of classes	609	608
Number of records	9028	4697 (52.02703%)
Suppressed records	4331 (47.97297%)	0

Şekil 4-36 (k=5, t=0,2, ρ=1) durumunda veri faydası sonuçları

Şekil 4-36 Şekil 4-24incelendiğinde ρ=1 durumunda elde edilen bulgularda en az 5, en fazla 4331 kayıt içeren 609 adet eşlenik sınıf olduğu görülmüştür. ρ=1 durumunda veri faydasına katkı sağlayan 608 adet UEC bulunmuştur. 4697 kaydı içeren 608 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=1 durumunda OEC kayıtlarının %52,02'sinin faydaya dönüştüğü tespit edilmiştir.

ρ=1 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-37'de verilmiştir.

Measure	Value [%]
Lowest prosecutor risk	2.63158%
Records affected by lowest risk	0.80903%
Average prosecutor risk	12.94443%
Highest prosecutor risk	20%
Records affected by highest risk	14.47733%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	12.94443%



a) Tüm riskler b) Savcı riski

Şekil 4-37 (k=5; t=0,2; ρ=1) durumunda risk sonuçları

Şekil 4-37 (a)' da verilen ρ=1 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 12,94 olduğu görülmüştür. Saldırgan kurbanın yayınlanan veri seti

içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınacaktır. ρ=1 durumunda savcı risk grafiği Şekil 4-37 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %14,47'sinin yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=1 durumunda en yüksek riskin değişmediği görülmüştür. ρ=1 durumunda tamamen bastırılan veri faydası olmayan 4331 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ=1 durumundan ρ=2 durumuna geçilmiştir.

ρ=2 durumu için 4331 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ikinci iterasyonunda ρ=2 durumunda elde edilen veri faydası sonuçları Şekil 4-38'de gösterilmiştir.

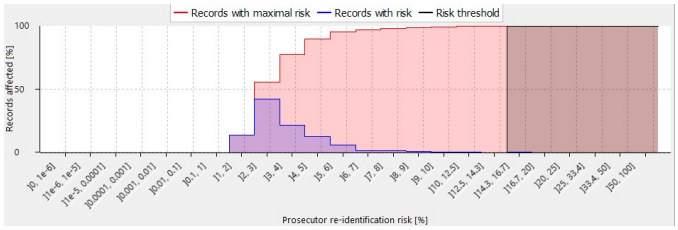
Measure	Including outliers	Excluding outliers
Average class size	61 (1.40845%)	30.3 (1.42857%)
Maximal class size	2210 (51.02748%)	64 (3.01744%)
Minimal class size	6 (0.13854%)	6 (0.28289%)
Number of classes	71	70
Number of records	4331	2121 (48.97252%)
Suppressed records	2210 (51.02748%)	0

Şekil 4-38 (k=5; t=0,2; ρ=2) durumunda veri faydası sonuçları

Şekil 4-38Şekil 4-32 incelendiğinde ρ=2 durumunda elde edilen bulgularda en az 5, en fazla 2210 kayıt içeren 70 adet eşlenik sınıf olduğu görülmüştür. ρ=2 durumunda veri faydasına katkı sağlayan 70 adet UEC bulunmuştur. 2121 kaydı içeren 70 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=2 durumunda OEC kayıtlarının %48,97'sinin faydaya dönüştüğü tespit edilmiştir.

ρ=2 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-39'da verilmiştir.

Measure	Value [%]
Lowest prosecutor risk	1.5625%
Records affected by lowest risk	3.01744%
Average prosecutor risk	3.30033%
Highest prosecutor risk	16.66667%
Records affected by highest risk	0.28289%
Estimated prosecutor risk	16.66667%
Estimated journalist risk	16.66667%
Estimated marketer risk	3.30033%



a) Tüm riskler b) Savcı riski

Şekil 4-39 (k=5; t=0,2; ρ=1) durumunda risk sonuçları

Şekil 4-39 (a)' da verilen ρ=2 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %16,66, %16,66 ve % 3,30 olduğu görülmektedir. Saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınmıştır. ρ=2 durumunda savcı risk grafiği Şekil 4-39 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %0,28'nin yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=2 durumunda en yüksek riskin azaldığı görülmüştür.

ρ-Kazanım (k,t) deneyi sonucunda veri faydası, UEC ve OEC karşılaştırmaları Çizelge 4-8'de gösterilmiştir.

Çizelge 4-8 ρ-Kazanım (k, t) değerlendirmesi

Model	ESO	UEC	OEC
(k=5; t=0,2)-Anonimlik	17,93	%70,06	%29,94
ρ-Kazanım (k=5; t=0,2;ρ=2)	12,78	%92,67	%7,33

Çizelge 4-8'Çizelge 4-7de (k,t)-Anonimlik modeli ile ρ-Kazanım (k,t) ölçüm sonuçları karşılaştırılmıştır. (k,t)-Anonimlik modelinde aykırı kayıtların dahil olduğu durumda veri kaybı 17,93 olarak ölçülmüştür. Tüm kayıtların içerisinde aykırı kayıt sayılarının oranı %29,94 olarak ölçülmüştür. ρ-Kazanım (k,t) modelinin aykırı kayıtların dahil olduğu durumda veri kaybı 12,57 olarak ölçülmüştür. ρ-Kazanım modelinin uygulanmasıyla veride yaşanan kayıp %28,72 oranında azalmıştır.

Bir diğer karşılaştırma UEC ve OEC oranlarıyla yapılmıştır. (k,t)-Anonimlik modelinde UEC oranı %70,06 OEC oranı %29,94 ρ-Kazanım (k,t) modelinde UEC oranı %92,67, OEC oranı %7,33 olarak ölçülmüştür. ρ-Kazanım (k,t) modelinin uygulanmasıyla aykırı kayıt sayıların oranı %22,61 oranında azalarak veri faydasına dönüştüğü tespit edilmiştir.

ρ-Kazanım modelinin OEC üzerinde yaptığı fayda arttırıcı işlemler sayesinde veri kaybının azaldığı dolayısıyla veri faydasının arttığı bu deneysel çalışmayla gösterilmiştir. Veri faydasındaki iyileşmeye rağmen tüm iterasyonlarda en yüksek mahremiyet risk ölçümünün %20 olarak kaldığı hatta son iterasyonda %16,66'ya düştüğü tespit edilmiştir.

4.2.4 ρ-Kazanım (k, ℓ ,t) Deneyi

ρ-Kazanım (k,ℓ,t) deneyiyle, ρ-Kazanım modelinin (k,ℓ,t)-Anonimlik modeline uygulanmasıyla veri faydası ile mahremiyet risklerinin nasıl değiştiği gösterilmiştir. Model parametreleri olarak tez çalışması kapsamında yapılan araştırmalarda en iyi sonuçları veren k=5, ℓ=2, t=0,2 ve ρ=2 değerleri seçilmiştir. Veri faydasının ölçümünde eşlenik sınıf ortalaması metriği, mahremiyet risklerinin değerlendirilmesinde savcı yaklaşımı kullanılmıştır.

ADULT veri kümesi üzerinde ρ'=0 durumunda veri faydası sonuçları Şekil 4-40'da verilmiştir.

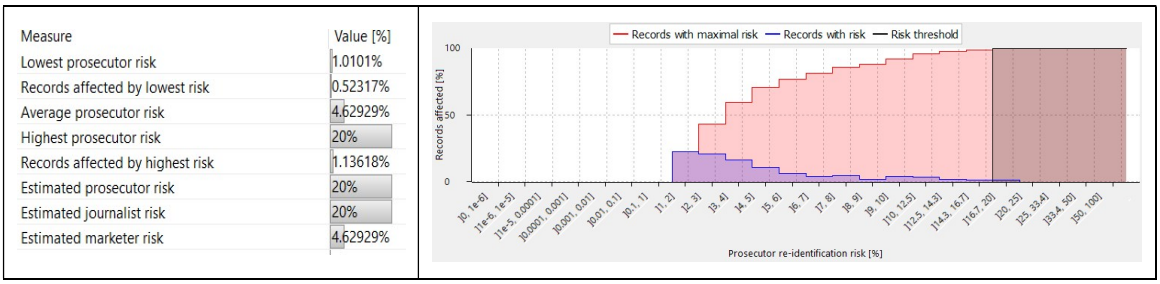
Measure	Including outliers	Excluding outliers
Average class size	34.39225 (0.11403%)	21.6016 (0.11416%)
Maximal class size	11239 (37.26212%)	99 (0.52317%)
Minimal class size	5 (0.01658%)	5 (0.02642%)
Number of classes	877	876
Number of records	30162	18923 (62.73788%)
Suppressed records	11239 (37.26212%)	0

Şekil 4-40 (k=5, ℓ=2, t=0,2, ρ'=0) durumunda veri faydası sonuçları

Şekil 4-40 incelendiğinde, en az 5, en fazla 11239 kayıt içeren 877 adet eşlenik sınıf olduğu görülmüştür. Veri faydası açısından aykırı kayıtları dahil ederek yapılan ölçümde ESO değeri 34,39 olarak bulunmuştur. ρ'=0 durumunda veri faydasına katkı sağlayan 876 adet UEC bulunmuştur. 18923 kaydı içeren 876 adet UEC yayınlanacak T* kümesine

aktarılmıştır. Veri faydası olmayan ve tamamen bastırılan veri kümesindeki kayıtların %37,26'sını oluşturan OEC içerisinde 11239 kayıt olduğu görülmüştür. Veri faydası sonuçlarına göre ρ'=0 durumundaki aykırı kayıtlı veri faydası 65,61 (100-34,39) olarak bulunmuştur.

ρ'=0 durumundaki risk tahminleri ve savcı risk grafiği Şekil 4-41'de verilmiştir.



Measure	Value [%]
Lowest prosecutor risk	1.0101%
Records affected by lowest risk	0.52317%
Average prosecutor risk	4.62929%
Highest prosecutor risk	20%
Records affected by highest risk	1.13618%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	4.62929%

a) Tüm riskler b) Savcı riski

Şekil 4-41 (k=5, ℓ=2, t=0,2, ρ'=0) durumunda risk sonuçları

Şekil 4-41 (a)' da verilen ρ'=0 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 4,62 olduğu görülmektedir. Bu çalışmada saldırganın kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınmıştır.

ρ'=0 durumunda savcı risk grafiği Şekil 4-41 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %1,13'nün yüksek risk aralığında (16,7-20) olduğu ve başlangıç riskinin ise %20 olduğu görülmüştür. Savcı yaklaşımına göre hesaplanan %20 değeri ρ'=0 durumundaki risk olarak kabul edilmiştir. ρ'=0 durumunda tamamen bastırılan veri faydası olmayan 11239 adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ'=0 durumundan ρ=1 durumuna geçilmiştir.

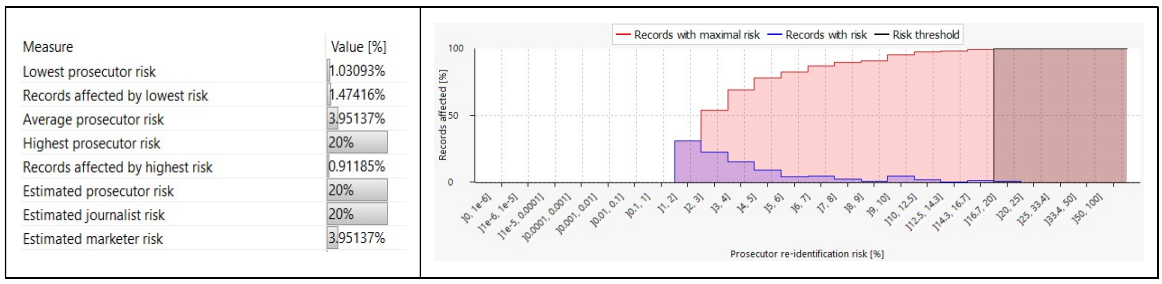
ρ=1 durumu için 11239 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ilk iterasyonunda ρ=1 durumunda elde edilen veri faydası sonuçları Şekil 4-24Şekil 4-42'de gösterilmiştir.

Measure	Including outliers	Excluding outliers
Average class size	43.0613 (0.38314%)	25.30769 (0.38462%)
Maximal class size	4659 (41.45387%)	97 (1.47416%)
Minimal class size	5 (0.04449%)	5 (0.07599%)
Number of classes	261	260
Number of records	11239	6580 (58.54613%)
Suppressed records	4659 (41.45387%)	0

Şekil 4-42 (k=5, ℓ=2, t=0,2, ρ=1) durumunda veri faydası sonuçları

Şekil 4-42 Şekil 4-24incelendiğinde ρ=1 durumunda elde edilen bulgularda en az 5, en fazla 4659 kayıt içeren 261 adet eşlenik sınıf olduğu görülmüştür. ρ=1 durumunda veri faydasına katkı sağlayan 260 adet UEC bulunmuştur. 6580 kaydı içeren 260 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=1 durumunda OEC kayıtlarının %58,54'nün faydaya dönüştüğü tespit edilmiştir.

ρ=1 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-43'de verilmiştir.



a) Tüm riskler b) Savcı riski

Şekil 4-43 (k=5, ℓ=2, t=0,2, ρ=1) durumunda risk sonuçları

Şekil 4-43 (a)' da verilen ρ=1 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 3,95 olduğu görülmüştür. Saldırganın kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınacaktır. ρ=1 durumunda savcı risk grafiği Şekil 4-43 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %0,91'inin yüksek risk aralığında (16,7-20) olduğu görülmüştür. Bulgularda ρ=1 durumunda en yüksek riskin değişmediği görülmüştür. ρ=1 durumunda tamamen bastırılan veri faydası olmayan 4659

adet aykırı kayıt ρ-Kazanım algoritmasına uygun olarak orijinal haline getirilmiş ρ=1 durumundan ρ=2 durumuna geçilmiştir.

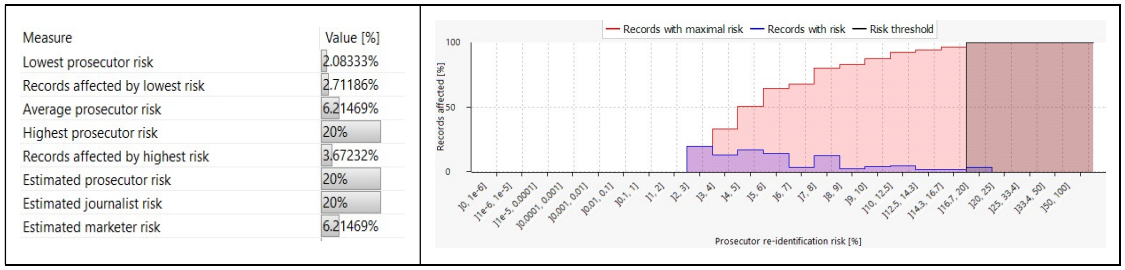
ρ=2 durumu için 4659 adet kayıt ρ-Kazanım algoritmasına uygun olarak kazanımsal anonimleşme işlemine alınmıştır. Kazanımsal anonimleşmenin ikinci iterasyonunda ρ=2 durumunda elde edilen veri faydası sonuçları Şekil 4-44'de verilmiştir.

Measure	Including outliers	Excluding outliers
Average class size	41.97297 (0.9009%)	16.09091 (0.90909%)
Maximal class size	2889 (62.00901%)	48 (2.71186%)
Minimal class size	5 (0.10732%)	5 (0.28249%)
Number of classes	111	110
Number of records	4659	1770 (37.99099%)
Suppressed records	2889 (62.00901%)	0

Şekil 4-44 (k=5, ℓ=2, t=0,2, ρ=2) durumunda veri faydası sonuçları

Şekil 4-44Şekil 4-38Şekil 4-32 incelendiğinde ρ=2 durumunda elde edilen bulgularda en az 5, en fazla 2889 kayıt içeren 111 adet eşlenik sınıf olduğu görülmüştür. ρ=2 durumunda veri faydasına katkı sağlayan 110 adet UEC bulunmuştur. 1770 kaydı içeren 110 adet UEC yayınlanacak T* kümesine aktarılmıştır. ρ=2 durumunda OEC kayıtlarının %37,99'unun faydaya dönüştüğü tespit edilmiştir.

ρ=2 durumu için risk tahminleri ve savcı risk grafiği Şekil 4-45'de verilmiştir.



a) Tüm riskler b) Savcı riski

Şekil 4-45 (k=5, ℓ=2, t=0,2, ρ=2) durumunda risk sonuçları

Şekil 4-45 (a)' da verilen ρ=2 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 6,21 olduğu görülmüştür. Saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından bu duruma en uygun olan savcı yaklaşımına göre yapılan risk hesaplaması dikkate alınmıştır. ρ=2 durumunda savcı risk grafiği Şekil 4-45 (b)'de verilmiştir. Grafik incelendiğinde kayıtların %3,67'sinin yüksek risk aralığında (16,7-20) olduğu görülmüştür. Bulgularda ρ=2 durumunda en yüksek riskin değişmediği görülmüştür.

ρ-Kazanım (k, ℓ t) deneyi sonucunda veri faydası, UEC ve OEC karşılaştırmaları Çizelge 4-9'da gösterilmiştir.

Çizelge 4-9 ρ-Kazanım (k, ℓ,t) değerlendirmesi

Model	ESO	UEC	OEC
k=5, ℓ=2, t=0,2	34,39	%62,74	%32,76
ρ-Kazanım (k=5, ℓ=2, t=0,2 ρ=2)	24,20	%90,59	%9,41

Çizelge 4-9'da (k,ℓ,t)-Anonimlik modeli ile ρ-Kazanım (k,ℓ,t) ölçüm sonuçları karşılaştırılmıştır. (k,ℓ,t)-Anonimlik modelinde aykırı kayıtların dahil olduğu durumda veri kaybı 34,39 olarak ölçülmüştür. Tüm kayıtların içerisinde aykırı kayıt sayılarının oranı %37,26 olarak ölçülmüştür. ρ-Kazanım (k,ℓ,t) modelinin aykırı kayıtların dahil olduğu durumda veri kaybı 23,18 olarak ölçülmüştür. ρ-Kazanım modelinin uygulanmasıyla veride yaşanan kayıp %29,63 oranında azalmıştır.

Bir diğer karşılaştırma UEC ve OEC oranlarıyla yapılmıştır. (k,ℓ,t)-Anonimlik modelinde UEC oranı %62,74 OEC oranı %32,76 ρ-Kazanım (k,ℓ,t) modelinde UEC oranı %90,59 OEC oranı %9,41 olarak ölçülmüştür. ρ-Kazanım (k,ℓ,t) modelinin uygulanmasıyla aykırı kayıt sayıların oranı %23,35 oranında azalarak veri faydasına dönüştüğü tespit edilmiştir.

ρ-Kazanım modelinin OEC üzerinde yaptığı fayda arttırıcı işlemler sayesinde veri kaybının azaldığı dolayısıyla veri faydasının arttığı bu deneysel çalışmayla gösterilmiştir. Veri faydasındaki iyileşmeye rağmen tüm iterasyonlarda en yüksek mahremiyet risk ölçümünün %20 olarak değişmediği tespit edilmiştir.

4.3 Bölüm Sonucu

Bu bölümde tez çalışması kapsamında önerilen model, farklı mahremiyet modellerine uygulanmıştır. Modelin uygulanmasıyla ortaya çıkan veri faydası ve mahremiyet riskleri model uygulanmadan önceki değerlerle karşılaştırılmıştır. ρ-Kazanım (k), ρ-Kazanım (k,ℓ), ρ-Kazanım (k,t), ρ-Kazanım (k,ℓ,t) deneysel çalışmalarında elde edilen bulgular ρ-Kazanım modelinin veri faydasını iyileştirdiğini ortaya koymuştur. Veri faydasındaki iyileşmenin mahremiyet risklerini olumsuz etkilemediği görülmüştür. Deneysel çalışmalarda elde edilen iyileştirme oranları ve iyileştirmeyi sağlayan OEC-UEC dönüşüm oranları Çizelge 4-10'da verilmiştir.

Çizelge 4-10 ρ-Kazanım genel değerlendirme

Model	Veri Faydası Artışı	OEC-UEC Dönüşümü
ρ-Kazanım (k=5, ρ=2)	%33,03	%24,01
ρ-Kazanım (k=5, ℓ =2, ρ=2)	%20,76	%20,12
ρ-Kazanım (k=5; t=0,2;ρ=2)	%28,72	%22,61
ρ-Kazanım (k=5, ℓ=2, t=0,2 ρ=2)	%29,63	%23,35

Çizelge 4-10'da verilen sonuçlar incelendiğinde OEC kayıtlarının %20 ile %24 arasında UEC'ye dönüşerek kazanım sağladığı, elde edilen bu kazanımın %20 ile %33 arasında veri faydasında iyileşme sağladığı görülmüştür. ρ-Kazanım modeli deneysel çalışmaları sonucunda veri faydası ile mahremiyet risk dengesini korumadaki başarısının yorumlanmasında Eşitlik 3-3'de verilen karar kuralı uygulanmıştır. UEC sayısının artması, OEC sayısının azalması ve risklerde değişim olmaması dikkate alındığında karar kuralı BAŞARILI değerini döndürmüştür.

5 SONUÇLAR

Veri mahremiyeti, veri sahiplerinin mahremiyet riskleri ile veri paylaşımının taraflara sağlayacağı fayda arasındaki en iyi dengeyi bulmaya çalışan zor bir problemdir. Veri mahremiyetinin korunmasıyla ilgili gerekli tedbirlerin alınmaması veya yetersiz olmasına bağlı olarak hassas verilerin kötüye kullanılmasıyla ortaya çıkan mahremiyet problemleri her geçen gün çeşitlenerek artmakta ve kayıplara yol açmaktadır. Mahremiyet ihlallerinin oluşmaması için mahremiyet koruma sürecinde tarafların sorumluluğunu bilerek veri paylaşımında bulunması gerekmektedir.

Mahremiyet koruma sürecinde veri sahipleri, veri toplayıcı (yayıncı) ve veri alıcılar olmak üzere üç önemli sorumlu taraf vardır. Veri sahipleri, paylaşılan veriler içerisinde kimlik ve hassas bilgileri yer alan mahremiyet farkındalığı olması gereken kişi, kurum ve kuruluşlardır. Veri toplayıcılar, veri sahiplerinin mahremiyetini koruyarak verilerin güvenli olarak ilgililerine veya halka açık olarak paylaşılmasını sağlayan kişi, kurum ve kuruluşlardır. Veri alıcılar, paylaşılan veriler üzerinde analizler veya işlemler yaparak verilerden niyetleri veya ihtiyaçları doğrultusunda fayda sağlamaya çalışan güvenilir olmadığı varsayılan üçüncü taraflardır. Mahremiyet koruma sürecinde kullanılan yaklaşımlar veri mahremiyeti probleminin çözümünde veri yayıncılar tarafından yaygın olarak kullanılır.

Mahremiyet korumalı yaklaşımlar mahremiyet gereksinimlerinin yerine getirilmesini sağlar. Mahremiyet gereksinimlerinin yerine getirilmesinde farklı mahremiyet modelleri tek başına veya birlikte kullanılır. Bu çalışma kapsamında incelenen k-Anonimlik, ℓ-Çeşitlilik, t-Yakınlık literatürde yaygın olarak kullanılan mahremiyet modellerindendir. Mahremiyet modellerinin uygulanabilmesi amacıyla birçok koruyucu yöntemden faydalanılır. Bu yöntemlerin başında anonimleştirme teknikleri gelmektedir.

Anonimleştirme veri detaylarını budamak amacıyla öznitelikler üzerinde yapılan fayda temelli dönüşüm işlemleridir. Anonimleştirmenin kabul edilebilir düzeyde yapılması mahremiyet koruması açısından önemlidir. Anonimleştirmenin gereğinden az yapılması mahremiyet problemlerine, gereğinden fazla yapılması veri kayıplarının çoğalarak veri kalitesinin bozulmasına yol açar. Anonimleştirmenin en uygun dengede yapılması mahremiyet koruma sürecindeki tüm tarafların talep ettiği önemli bir mahremiyet gereksinimidir.

Mahremiyet korumalı yaklaşımları kullanan veri yayıncılar, yalnızca verinin korunarak yayınlanmasından değil yayınlanan verinin alıcılarına sağladığı fayda ve kaliteden de sorumludur. Koruma seviyelerinin belirlenebilmesi, alınan önlemlerin ne düzeyde olduğunun anlaşılabilmesi amacıyla mahremiyet riskleri ile veri faydasının birlikte değerlendirilmesi gerekir. Veri faydasının ölçülmesinde veri kayıplarının ölçümünde kullanılan metriklerden yararlanılır. Veri kaybının ölçülmesinde literatürde birçok genel ve özel metrik kullanılmış ve bu çalışma kapsamında özetlenmiştir. Bilgi kaybını ölçen metrikler öznitelikler veya eşlenik sınıflar düzeyinde ölçümler yapar. Tez çalışması kapsamında yapılan çalışmaların eşlenik düzeyde işlemler yapmasına bağlı olarak veri faydasının ölçümünde eşlenik sınıf temelli çalışan eşlenik sınıf ortalaması metriği ile kullanılmıştır.

Mahremiyet korumalı yaklaşımların başarımının değerlendirilmesinde veri faydasının yanında mahremiyet risklerinin de değerlendirilmesi gerekir. Veri faydası ile mahremiyet risklerinin dengelenmesi veri mahremiyeti probleminin çözümündeki temel problemdir. Mahremiyet riskleri konusunda yapılan ölçümler veri yayıncılarını mahremiyet riskleri konusunda bilgilendirerek yeterli düzeyde önlemler almasını sağlar. Tez çalışması kapsamında kimlik ifşaları için yeniden tanımlama olasılığının ölçülmesi ve yorumlanması için türetilmiş metrikler ile karar kuralları, bu metriklerin uygulandığı basit metrikler ve biriciklik metriği incelenmiştir. En yüksek risk oranına sahip olan savcı yaklaşımı adı verilen basit bir risk metriği kullanılmıştır.

Anonimleştirilen veriler benzerliklerine göre eşlenik sınıf adı verilen gruplar içerisinde toplanır. Tez çalışması kapsamında eşlenik sınıflar veri faydasına göre, fayda sağlayan (Utility Equivalence Class-UEC) ve aykırı (Outlier Equivalence Class-OEC) olmak üzere iki sınıfa ayrılmıştır. Faydalı eşlenik sınıflar mahremiyet gereksinimlerini sağlayan veri alıcılarına fayda sunan kayıtları içerirken, aykırı eşlenik sınıf, mahremiyet gereksinimlerini sağlayamadığı için tamamen bastırılan veri faydası sunamayan kayıtları içerir.

Tez çalışmasında sunulan eşlenik sınıf ayrımının veri faydası ve mahremiyet riskleri üzerindeki etkisi incelenerek aykırı eşlenik sınıf içerisinde yer alan kayıtların veri faydası açısından kazanımı konusu araştırılmıştır. Veri kalitesini bozarak faydasını azaltan, eşlenik sınıflar içerisindeki gruplandırmaları olumsuz etkileyen ve mahremiyet gereksinimlerini sağlayamayan kayıtların tamamen bastırılarak dışlanmasının veri faydasına olan etkisi de ayrıca çalışılmıştır. Bu kapsamda mahremiyetten ödün vermeden veri faydasını arttıran ρ-Kazanım olarak adlandırdığımız mahremiyet korumalı fayda temelli bir model tanımlanmış

ve test edilmiştir. ρ-Kazanım modeliyle eşlenik sınıf ayrımı yapılmış, veri faydası olmayan aykırı kayıtları içeren OEC üzerinde veri faydasını arttırıcı işlemler yapmıştır. Ayrıca OEC üzerinde yapılan bu işlemlerin mahremiyet risklerine olan etkisi araştırılmıştır.

ρ-Kazanım modeli literatürde yaygın olarak kullanılan k-Anonimlik, ℓ-Çeşitlilik ve t-Yakınlık veri yayınlama modellerine farklı kombinasyonlarla uygulanarak başarımı değerlendirilmiştir. Farklı veri kümeleri üzerinde yapılan deneysel çalışmalarda elde edilen sonuçlar veri faydası ve mahremiyet riskleri açısından karşılaştırılmıştır. Elde edilen sonuçlara göre, OEC kayıtlarının %20 ile %24 arasında UEC'ye dönüşerek kazanım sağladığı, elde edilen bu kazanımın %20 ile %33 arasında veri faydasına dönüştüğü görülmüştür. ρ-Kazanım modeli deneysel çalışmaları sonucunda veri faydası ile mahremiyet risk dengesini korumadaki başarısının yorumlanmasında karar kuralı kullanılmıştır. UEC sayısının artmasına bağlı eşlenik sınıf sayılarındaki yükselme, OEC sayısının azalmasına bağlı olarak veri faydasındaki iyileşme dikkate alındığında karar kuralına göre ρ-Kazanım BAŞARILI değerini döndürmüştür.

Sonuç olarak, veri faydasına göre eşlenik sınıf ayrımı yaparak iteratif yapıda fayda optimizasyonu yapabilen ρ-Kazanım modelinin uygulanmasıyla tüm iterasyonlarda elde edilen bulgularda veri faydasında iyileşme olduğu görülürken mahremiyet risklerinde olumsuz bir değişiklik olmamıştır. Bu sonuçlar doğrultusunda, aykırı eşlenik sınıfların dışlanarak yeniden kullanımının veri faydası üzerine olumlu etkisinin olduğu ve dışlanan aykırı kayıtların geri kazanımının mahremiyet risklerini olumsuz etkilemediği görülmüştür.

Gelecek çalışmalarda, UEC içerisinde yer alan kayıtların büyüklüğünün veri faydası ile mahremiyet arasındaki denge ile OEC'yi nasıl etkileyeceği araştırılmalıdır. OEC geri kazanımında en uygun iterasyon sayısının (ρ) bulunması ise bir diğer önemli araştırma konusudur. ρ-Kazanım içerisinde farklı sınıflandırma algoritmaları kullanılmasının UEC ve OEC üzerindeki etkileri de araştırılmalıdır. Son olarak ρ-Kazanım modeline özgü geliştirilecek eşlenik sınıf temelli bilgi metrikleri üzerinde çalışılmalıdır.

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı: Yılmaz VURAL

Doğum Yeri: Kahramanmaraş

Medeni Hali: Evli

E-posta: yvural@hacettepe.edu.tr

Adresi: Türksat A.Ş Konya Yolu 40. Km Gölbaşı ANKARA

Eğitim

Lise: Sütçü İmam Lisesi

Lisans: Trakya Üniversitesi, Bilgisayar Mühendisliği

Y. Lisans: Gazi Üniversitesi, Bilgisayar Mühendisliği

Doktora: Hacettepe Üniversitesi, Bilgisayar Mühendisliği

Yabancı Dil ve Düzeyi

İngilizce, İyi seviyede

İş Deneyimi

Textiplik A.Ş Yazılım Mühendisi	1996-1998
Kahramanmaraş Sütçü İmam Üniversitesi	1998-2000
STM A.Ş Kıdemli Yazılım Mühendisi	2000-2010
Türksat A.Ş Proje Yöneticisi	2010-Devam

Deneyim Alanları

Bilgi güvenliği, veri mahremiyeti, proje yönetimi

Tezden Üretilmiş Projeler ve Bütçesi

-

Tezden Üretilmiş Yayınlar

Vural, Y., Aydos M., ρ-Kazanım: Fayda Temelli Veri Yayınlama Modeli, *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, (Kabul Mayıs-2017).

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

Vural, Y., Aydos M., A New Approach to Utility-based Privacy Preserving in Data Publishing, *17th IEEE International Conference on Computer and Information Technology (IEEE CIT-2017)* Helsinki, Finland, 21-23 August, 2017 (Kabul Mayıs-2017).

KAYNAKLAR

[1] Xu, L., Jiang, C., Wang, J., Yuan, J., Ren, Y., Information security in big data: Privacy and data mining, *IEEE Access,* 2, 1149-1176, **2014**.

[2] Samarati, P., Protecting respondents identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering,* 13, 6, 1010-1027, **2001**.

[3] Korolova, A., *Protecting privacy when mining and sharing user data*, Doktora Tezi, Bilgisayar Bilimleri, Stanford Üniversitesi, **2012**.

[4] De Capitani Di Vimercati, S., Foresti, S., Livraga, G., Samarati, P., Data privacy: Definitions and techniques, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 20, 06, 793-817, **2012**.

[5] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., Theodoridis, Y., State-of-the-art in privacy preserving data mining, *ACM Sigmod Record,* 33, 1, 50-57, **2004**.

[6] Xu, Y., Fung, B. C. M., Wang, K., Fu, A. W. C., Pei, J., Publishing sensitive transactions for itemset utility, *2008 Eighth IEEE International Conference on Data Mining*, 1109-1114, 15-19 Dec., Pisa, Italy, **2008.**

[7] Mahmood, S., New privacy threats for facebook and twitter users, *2012 Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, 164-169, 12-14 Nov, Victoria, BC, Canada, **2012.**

[8] Chen, B.-C., Kifer, D., Lefevre, K., Machanavajjhala, A., Privacy-preserving data publishing, *Foundations and Trends® in Databases,* 2, 1–2, 1-167, **2009**.

[9] Canbay, P. Sever, H., The effect of clustering on data privacy, *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 277-282, December 9-11, Miami, Florida, USA, **2015.**

[10] Warren, S. D. Brandeis, L. D., The right to privacy, *Harvard law review,* 193-220, **1890**.

[11] Yüksel, M., Mahremiyet hakkina ve bireysel özgürlüklere felsefi yaklaşımlar, *Ankara Üniversitesi SBF Dergisi,* 64, 1, 275-298, **2009**.

[12] Narayanan, A. Shmatikov, V., Robust de-anonymization of large sparse datasets, *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111-125, 18-21 May 2008, Oakland, California, USA, **2008.**

[13] Erlandsson, F., Boldt, M., Johnson, H., Privacy threats related to user profiling in online social networks, *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 838-842, 3-5 Sep., Amsterdam, Netherlands, **2012.**

[14] Yang, L., Xue, H., Li, F., Privacy-preserving data sharing in smart grid systems, *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 878-883, 3-6 Nov., Venice, Italy, **2014.**

[15] Khan, J., Abbas, H., Al-Muhtadi, J., Survey on mobile user's data privacy threats and defense mechanisms, *Procedia Computer Science,* 56, 376-383, **2015**.

[16] Linn, J., Technology and web user data privacy - a survey of risks and countermeasures, *IEEE Security & Privacy,* 3, 1, 52-58, **2005**.

[17] Sweeney, L., K-anonymity: A model for protecting privacy, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.,* 10, 5, 557-570, **2002**.

[18] Sweeney, L., Simple demographics often identify people uniquely, *Health (San Francisco),* 671, 1-34, **2000**.

[19] Aggarwal, C. C. Yu, P. S., *A general survey of privacy-preserving data mining models and algorithms*, in *Privacy-preserving data mining: Models and algorithms*, (eds: Aggarwal, C. C. Yu, P. S.), Boston, MA: Springer US, 11-52, **2008**.

[20] Pinkas, B., Cryptographic techniques for privacy-preserving data mining, *SIGKDD Explor. Newsl.,* 4, 2, 12-19, **2002**.

[21] Brearty, S. M., Farrelly, W., Curran, K., Preserving data privacy with searchable symmetric encryption, *2016 27th Irish Signals and Systems Conference (ISSC)*, 1-7, 21-22 June, London, United Kingdom, **2016.**

[22] Fung, B. C. M., Wang, K., Yu, P. S., Anonymizing classification data for privacy preservation, *IEEE Transactions on Knowledge and Data Engineering,* 19, 5, 711-725, **2007**.

[23] Thakkar, A., Bhatti, A. A., Vasa, J., *Correlation based anonymization using generalization and supprssion for disclosure problems*, in *Advances in intelligent informatics*, (eds: El-Alfy, E.-S. M., Thampi, S. M., Takagi, H., Piramuthu, S., Hanne, T.), Cham: Springer International Publishing, 581-592, **2015**.

[24] Xiao, X. Tao, Y., Anatomy: Simple and effective privacy preservation, *Proceedings of the 32nd international conference on Very large data bases*, September 12-15, Seoul, Korea, 139-150, **2006.**

[25] He, X., Xiao, Y., Li, Y., Wang, Q., Wang, W., Shi, B., *Permutation anonymization: Improving anatomy for privacy preservation in data publication*, in *New frontiers in applied data mining: Pakdd 2011 international workshops, shenzhen, china, may 24-27, 2011, revised selected papers*, (eds: Cao, L., Huang, J. Z., Bailey, J., Koh, Y. S., Luo, J.), Berlin, Heidelberg: Springer, 111-123, **2012**.

[26] Banu, K. S., Santhi, V., Tripathy, B. K., Non-cryptographic security to data: Distortion based anonymization techniques, *2014 International Conference on Advances in Engineering and Technology (ICAET)*, 1-5, Juner 13-15, Hammamet, Tunisia, **2014.**

[27] Brickell, J. Shmatikov, V., The cost of privacy: Destruction of data-mining utility in anonymized data publishing, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 70-78, Las Vegas, Nevada, USA, **2008.**

[28] Paintsil, E. Fritsch, L., A taxonomy of privacy and security risks contributing factors, *IFIP PrimeLife International Summer School on Privacy and Identity Management for Life*, 52-63, **2010.**

[29] Fayyoumi, E. Oommen, B. J., A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases, *Softw. Pract. Exper.,* 40, 12, 1161-1188, **2010**.

[30] Saranya, K., Premalatha, K., Rajasekar, S. S., A survey on privacy preserving data mining, *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 1740-1744, February 26-27 Coimbatore, India, **2015.**

[31] Abuwardih, L. A., Shatnawi, W., Aleroud, A., Privacy preserving data mining on published data in healthcare: A survey, *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, July 13-14, Amman, Jordan, **2016.**

[32] Wang, J., Luo, Y., Zhao, Y., Le, J., A survey on privacy preserving data mining, *2009 First International Workshop on Database Technology and Applications*, 111-114, April 25-26, Wuhan, Hubei, China, **2009.**

[33] Fung, B. C. M., Wang, K., Chen, R., Yu, P. S., Privacy-preserving data publishing, *ACM Computing Surveys,* 42, 4, 1-53, **2010**.

[34] Fung, B. C. M., Wang, K., Fu, A. W. C., Pei, J., *Anonymity for continuous data publishing*, in *Proceedings 11th international conference on extending database technology: Advances in database technology*, 264-275, March 25-29, Nantes, France, **2008**.

[35] Li, T., Li, N., Zhang, J., Molloy, I., Slicing: A new approach for privacy preserving data publishing, *IEEE Transactions on Knowledge and Data Engineering,* 24, 3, 561-574, **2012**.

[36] Sattar, A. S., Li, J., Ding, X., Liu, J., Vincent, M., A general framework for privacy preserving data publishing, *Knowledge-Based Systems,* 54, 276-287, **2013**.

[37] Bethlehem, J. G., Keller, W. J., Pannekoek, J., Disclosure control of microdata, *Journal of the American Statistical Association,* 85, 409, 38-45, **1990**.

[38] Lodha, S. Thomas, D., *Probabilistic anonymity*, in *Privacy, security, and trust in kdd*, Springer, 56-79, **2008**.

[39] Xu, Y., Wang, K., Fu, A. W.-C., Yu, P. S., Anonymizing transaction databases for publication, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 767-775, August 24-27, Las Vegas, NV, USA, **2008.**

[40] Vatsalan, D., Christen, P., O'keefe, C. M., Verykios, V. S., An evaluation framework for privacy-preserving record linkage, *Journal of Privacy and Confidentiality,* 6, 1, 3, **2014**.

[41] Golle, P., Revisiting the uniqueness of simple demographics in the us population, *Proceedings of the 5th ACM workshop on Privacy in electronic society*, 77-80, October 30-November 03, Alexandria, VA, USA, **2006.**

[42] Sweeney, L., K-anonymity: A model for protecting privacy, *Int J Uncertainty Fuzziness Knowledge Based Syst,* 10, **2002**.

[43] Sweeney, L., Achieving k-anonymity privacy protection using generalization and suppression, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 10, 05, 571-588, **2002**.

[44] Kenig, B. Tassa, T., A practical approximation algorithm for optimal k-anonymity, *Data Mining and Knowledge Discovery,* 25, 1, 134-168, **2012**.

[45] Zhou, B., Han, Y., Pei, J., Jiang, B., Tao, Y., Jia, Y., Continuous privacy preserving publishing of data streams, *Proceedings of the 12th International Conference on*

Extending Database Technology: Advances in Database Technology, 648-659, March 23 - 26, Saint-Petersburg, Russian Federation, **2009.**

[46] Torra, V. Navarro-Arribas, G., *Data privacy: A survey of results*, in *Advanced research in data privacy*, (eds: Navarro-Arribas, G. Torra, V.), Cham: Springer International Publishing, 27-37, **2015**.

[47] Nergiz, M. E., Clifton, C., Nergiz, A. E., Multirelational k-anonymity, *IEEE Transactions on Knowledge and Data Engineering,* 21, 8, 1104-1117, **2009**.

[48] Wang, K. Fung, B., Anonymizing sequential releases, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 414-423, August 20-23, Philadelphia, PA, USA, **2006.**

[49] Wong, R. C.-W., Li, J., Fu, A. W.-C., Wang, K., (α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 754-759, August 20-23, Philadelphia, PA, USA, **2006.**

[50] Chawla, S., Dwork, C., Mcsherry, F., Smith, A., Wee, H., Toward privacy in public databases, *Theory of Cryptography Conference*, 363-385, February 10-12, Cambridge, MA, USA, **2005.**

[51] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M., L-diversity: Privacy beyond k-anonymity, *22nd International Conference on Data Engineering (ICDE'06)*, 24-24, April 3-7, Atlanta, GA, USA, **2006.**

[52] Wang, K., Fung, B. C., Philip, S. Y., Handicapping attacker's confidence: An alternative to k-anonymization, *Knowledge and Information Systems,* 11, 3, 345-368, **2007**.

[53] Zhang, Q., Koudas, N., Srivastava, D., Yu, T., Aggregate query answering on anonymized tables, *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, 116-125, April 11-15, İstanbul, Turkey, **2007.**

[54] Li, J., Tao, Y., Xiao, X., Preservation of proximity privacy in publishing numerical sensitive data, *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 473-486, June 09-12,Vancouver, Canada, **2008.**

[55] Xiao, X. Tao, Y., Personalized privacy preservation, *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 229-240, June 27-29, Chicago, IL, USA, **2006.**

[56] Li, N., Li, T., Venkatasubramanian, S., T-closeness: Privacy beyond k-anonymity and l-diversity, *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, 106-115, April 11-15, İstanbul, Turkey, **2007.**

[57] Li, N., Li, T., Venkatasubramanian, S., Closeness: A new privacy measure for data publishing, *IEEE Transactions on Knowledge and Data Engineering,* 22, 7, 943-956, **2010**.

[58] Rubner, Y., Tomasi, C., Guibas, L. J., The earth mover's distance as a metric for image retrieval, *International journal of computer vision,* 40, 2, 99-121, **2000**.

[59] Ling, H. Okada, K., An efficient earth mover's distance algorithm for robust histogram comparison, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 29, 5, 840-853, **2007**.

[60] Nergiz, M. E., Atzori, M., Clifton, C., Hiding the presence of individuals from shared databases, *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 665-676, June 11-14, Beijing, China, **2007.**

[61] Dwork, C., Differential privacy: A survey of results, *International Conference on Theory and Applications of Models of Computation*, 1-19, April 25-29, Xi'an, China, **2008.**

[62] Xiao, X., Wang, G., Gehrke, J., Differential privacy via wavelet transforms, *IEEE Transactions on Knowledge and Data Engineering,* 23, 8, 1200-1214, **2011**.

[63] Chaum, D. L., Untraceable electronic mail, return addresses, and digital pseudonyms, *Communications of the ACM,* 24, 2, 84-90, **1981**.

[64] Jakobsson, M., Juels, A., Rivest, R. L., Making mix nets robust for electronic voting by randomized partial checking, *USENIX security symposium*, 339-353, August 05 - 09, USA, **2002.**

[65] Ciriani, V., Vimercati, S. D. C., Foresti, S., Samarati, P., K-anonymous data mining: A survey, *Privacy-preserving data mining,* 105-136, **2008**.

[66] Samarati, P., Protecting respondents' identities in microdata release, *IEEE Trans. on Knowl. and Data Eng.,* 13, 6, 1010-1027, **2001**.

[67] Agresti, A. Kateri, M., *Categorical data analysis*, Springer, **2011**.

[68] Samarati, P. Sweeney, L., Generalizing data to provide anonymity when disclosing information, *PODS*, 88, June 01-04, Seattle, Washington, USA, **1998.**

[69] Campan, A., Truta, T. M., Cooper, N., P-sensitive k-anonymity with generalization constraints, *Trans. Data Privacy,* 3, 2, 65-89, **2010**.

[70] Atkinson, M. D., Sack, J.-R., Santoro, N., Strothotte, T., Min-max heaps and generalized priority queues, *Communications of the ACM,* 29, 10, 996-1000, **1986**.

[71] Prasser, F. Kohlmayer, F., *Putting statistical disclosure control into practice: The arx data anonymization tool*, in *Medical data privacy handbook*, Springer, 111-148, **2015**.

[72] Kohlmayer, F., Prasser, F., Kuhn, K. A., The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss, *Journal of biomedical informatics,* 58, 37-48, **2015**.

[73] Terrovitis, M., Mamoulis, N., Kalnis, P., Local and global recoding methods for anonymizing set-valued data, *The VLDB Journal—The International Journal on Very Large Data Bases,* 20, 1, 83-106, **2011**.

[74] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A. W.-C., Utility-based anonymization using local recoding, *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 785-790, August 20 - 23, Philadelphia, PA, USA, **2006.**

[75] Lefevre, K., Dewitt, D. J., Ramakrishnan, R., Incognito: Efficient full-domain k-anonymity, *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 49-60, June 13 - 17, Baltimore, MD, USA, **2005.**

[76] Xu, Y., Ma, T., Tang, M., Tian, W., A survey of privacy preserving data publishing using generalization and suppression, *Applied Mathematics & Information Sciences,* 8, 3, 1103-1116, **2014**.

[77] Iyengar, V. S., Transforming data to satisfy privacy constraints, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, July 23-25, Edmonton, Alberta, Canada, 279-288, **2002.**

[78] Lefevre, K., Dewitt, D. J., Ramakrishnan, R., Mondrian multidimensional k-anonymity, *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, 25-25, April 3-7, Atlanta, GA, USA, **2006.**

[79] Tian, H. Zhang, W., Extending ℓ-diversity to generalize sensitive data, *Data & Knowledge Engineering,* 70, 1, 101-126, **2011**.

[80] Meyerson, A. Williams, R., On the complexity of optimal k-anonymity, *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 223-228, June 13-18, Paris, France, **2004.**

[81] Bayardo, R. J. Agrawal, R., Data privacy through optimal k-anonymization, *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, 217-228, April 05-08, Washington, DC, USA, **2005.**

[82] Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A., Achieving anonymity via clustering, *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 153-162, June 27 - 29, Chicago, IL, USA, **2006.**

[83] Samarati, P. Sweeney, L., Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression, *SRI International,* **1998**.

[84] Rastogi, V., Suciu, D., Hong, S., The boundary between privacy and utility in data publishing, *Proceedings of the 33rd international conference on Very large databases*, 531-542, September 23-28, Vienna, Austria, **2007.**

[85] Muralidhar, K. Sarathy, R., A theoretical basis for perturbation methods, *Statistics and Computing,* 13, 4, 329-335, **2003**.

[86] Liu, K., Giannella, C., Kargupta, H., *A survey of attack techniques on privacy-preserving data perturbation methods*, in *Privacy-preserving data mining: Models and algorithms*, (eds: Aggarwal, C. C. Yu, P. S.), Boston, MA: Springer US, 359-381, **2008**.

[87] Oganian, A. Karr, A. F., *Combinations of sdc methods for microdata protection*, in *Privacy in statistical databases: Cenex-sdc project international conference, psd 2006, rome, italy, december 13-15, 2006. Proceedings*, (eds: Domingo-Ferrer, J. Franconi, L.), Berlin, Heidelberg: Springer Berlin Heidelberg, 102-113, **2006**.

[88] Agrawal, R. Srikant, R., Privacy-preserving data mining, *ACM Sigmod Record*, 29, 2, 439-450, **2000.**

[89] Evfimievski, A., Randomization in privacy preserving data mining, *ACM Sigkdd Explorations Newsletter,* 4, 2, 43-48, **2002**.

[90] Kargupta, H., Datta, S., Wang, Q., Sivakumar, K., On the privacy preserving properties of random data perturbation techniques, *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 99-106, November 19-22, Washington, DC, USA, **2003.**

[91] Huang, Z., Du, W., Chen, B., Deriving private information from randomized data, *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 37-48, June 14-16, Baltimore, Maryland, **2005.**

[92] Kim, J. Winkler, W., Multiplicative noise for masking continuous data, *Statistics,* 01, **2003**.

[93] Liu, K., Kargupta, H., Ryan, J., Random projection-based multiplicative data perturbation for privacy preserving distributed data mining, *IEEE Transactions on Knowledge and Data Engineering,* 18, 1, 92-106, **2006**.

[94] Liu, K., Giannella, C., Kargupta, H., An attacker's view of distance preserving maps for privacy preserving data mining, *European Conference on Principles of Data Mining and Knowledge Discovery*, 297-308, **2006.**

[95] Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz, J. M., Sebé, F., Efficient multivariate data-oriented microaggregation, *The VLDB Journal—The International Journal on Very Large Databases,* 15, 4, 355-369, **2006**.

[96] Domingo-Ferrer, J. Mateo-Sanz, J. M., Practical data-oriented microaggregation for statistical disclosure control, *IEEE Transactions on Knowledge and Data Engineering,* 14, 1, 189-201, **2002**.

[97] Domingo-Ferrer, J. Torra, V., Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery,* 11, 2, 195-212, **2005**.

[98] R. H. Khokhar, B. C. M. Fung, F. Iqbal, D. Alhadidi, J. Bentahar. Privacy-preserving data mashup model for trading person-specific information. Electronic Commerce Research and Applications (ECRA), 17:19-37, May-June **2016**.

[99] Wong, R. C.-W., Fu, A. W.-C., Wang, K., Pei, J., Minimality attack in privacy preserving data publishing, *Proceedings of the 33rd international conference on Very large databases*, 543-554, September 23-28, Vienna, Austria, **2007.**

[100] Duncan, G. Lambert, D., The risk of disclosure for microdata, *Journal of Business & Economic Statistics,* 7, 2, 207-217, **1989**.

[101] Skinner, C. Holmes, D., Estimating the re-identification risk per record in microdata, *Journal of Official Statistics,* 14, 4, 361, **1998**.

[102] Dankar, F. K., El Emam, K., Neisa, A., Roffey, T., Estimating the re-identification risk of clinical data sets, *BMC Medical Informatics and Decision Making,* 12, 1, 66, **2012**.

[103] Winkler, W. E., Masking and re-identification methods for public-use microdata: Overview and research problems, *International Workshop on Privacy in Statistical Databases*, 231-246, **2004.**

[104] Sun, X., Sun, L., Wang, H., Extended k-anonymity models against sensitive attribute disclosure, *Computer Communications,* 34, 4, 526-535, **2011**.

[105] Domingo-Ferrer, J. Torra, V., A critique of k-anonymity and some of its enhancements, *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*, 990-993, **2008.**

[106] Chen, B.-C., Lefevre, K., Ramakrishnan, R., Privacy skyline: Privacy with multidimensional adversarial knowledge, *Proceedings of the 33rd international*

conference on Very large databases, 770-781, September 23-28, Vienna, Austria, **2007.**

[107] Peng, Y., Kou, G., Shi, Y., Chen, Z., A descriptive framework for the field of data mining and knowledge discovery, *International Journal of Information Technology & Decision Making,* 7, 04, 639-682, **2008**.

[108] Martin, D. J., Kifer, D., Machanavajjhala, A., Gehrke, J., Halpern, J. Y., Worst-case background knowledge for privacy-preserving data publishing, *2007 IEEE 23rd International Conference on Data Engineering*, 126-135, April 11-15, İstanbul, Turkey, **2007.**

[109] Sweeney, L., Computational disclosure control, *A Primer on Data Privacy Protection,* **2001**.

[110] Li, T. Li, N., On the tradeoff between privacy and utility in data publishing, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, June 28 - July 01, Paris, France, 517-526, **2009.**

[111] Domingo-Ferrer, J. Torra, V., Disclosure control methods and information loss for microdata, *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies,* 91-110, **2001**.

[112] Ghinita, G., Karras, P., Kalnis, P., Mamoulis, N., Fast data anonymization with low information loss, *Proceedings of the 33rd international conference on Very large databases*, 758-769, September 23 - 28, University of Vienna, Austria, **2007.**

[113] Fung, B. C., Wang, K., Fu, A. W.-C., Philip, S. Y., *Introduction to privacy-preserving data publishing: Concepts and techniques*. CRC Press, **2010**.

[114] Gionis, A. Tassa, T., K-anonymization with minimal loss of information, *IEEE Transactions on Knowledge and Data Engineering,* 21, 2, 206-219, **2009**.

[115] Sui, P. Li, X., A privacy-preserving approach for multimodal transaction data integrated analysis, *Neurocomputing,* 253, 56-64, **8/30/ 2017**.

[116] De Waal, A. Willenborg, L., Information loss through global recoding and local suppression, *Netherlands Official Statistics,* 14, 17-20, **1999**.

[117] Kullback, S. Leibler, R. A., On information and sufficiency, *The annals of mathematical statistics,* 22, 1, 79-86, **1951**.

[118] Chen, B.-C., Kifer, D., Lefevre, K., Machanavajjhala, A., Privacy-preserving data publishing, *Found. Trends databases,* 2, 2, 1-167, **2009**.

[119] El Emam, K., Risk-based de-identification of health data, *IEEE Security & Privacy,* 8, 3, 64-67, **2010**.

[120] El Emam, K., *Guide to the de-identification of personal health information*. CRC Press, **2013**.

[121] El Emam, K. Dankar, F., Re-identification risk in de-identified databases containing personal information, *Google*, **2012**.

[122] Dankar, F. K. Emam, K. E., A method for evaluating marketer re-identification risk, *Proceedings of the 2010 EDBT/ICDT Workshops*, 1-10, March 22, Lausanne, Switzerland, **2010.**

[123] Sweeney, L., Uniqueness of simple demographics in the us population, *Technical report*, Carnegie Mellon University, **2000**.

[124] El Emam, K., Brown, A., Abdelmalik, P., Evaluating predictors of geographic area population size cut-offs to manage re-identification risk, *Journal of the American Medical Informatics Association,* 16, 2, 256-266, **2009**.

[125] Skinner, C. J. Elliot, M., A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society: series B (statistical methodology),* 64, 4, 855-867, **2002**.

[126] Xiao, X., Wang, G., Gehrke, J., Interactive anonymization of sensitive data, *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, 1051-1054, June 29-July 02, Rhode Island, USA, **2009.**

[127] Dai, C., Ghinita, G., Bertino, E., Byun, J.-W., Li, N., Tiamat: A tool for interactive analysis of microdata anonymization techniques, *Proceedings of the VLDB Endowment,* 2, 2, 1618-1621, **2009**.

[128] Prasser, F., Kohlmayer, F., Lautenschläger, R., Kuhn, K. A., Arx-a comprehensive tool for anonymizing biomedical data, *AMIA Annual Symposium Proceedings*, 984-993, **2014.**

[129] Poulis, G., Gkoulalas-Divanis, A., Loukides, G., Skiadopoulos, S., Tryfonopoulos, C., Secreta: A system for evaluating and comparing relational and transaction anonymization algorithms, **2014**.

[130] Sweeney, L., *Datafly: A system for providing anonymity in medical data*, in *Database security xi*, Springer, 356-381, August 10-13, Lake Tahoe, California, USA **1998**.

[131] Anonim, *Utd data security and privacy tool,* http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php, (Haziran, **2017**).

[132] Lichman, M., *{uci} machine learning repository,* http://archive.ics.uci.edu/ml, (Haziran, **2017**).

[133] Gong, Q., *Partition and anatomize anonymization,* https://github.com/qiyuangong/PAA/blob/master/data/demographics.csv, (Haziran, **2017**).

[134] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J.-P., Walker, M., Chowdhury, S., Vaillancourt, R., A globally optimal k-anonymity method for the de-identification of health data, *Journal of the American Medical Informatics Association,* 16, 5, 670-682, **2009**.

[135] Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., Kuhn, K. A., Flash: Efficient, stable and optimal k-anonymity, *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 708-717, September 3-5, Amsterdam, Netherlands, **2012.**

[136] Kohlmayer, F. M., *Datenschutz und biomedizinische forschung: Konzepte und lösungen für anonymität*, Doktora Tezi, Enformatik Fakültesi, Sağlık Bilişimi, Münih Teknik Üniversitesi, Münih, **2015**.

[137] Aggarwal, C. C., On k-anonymity and the curse of dimensionality, *Proceedings of the 31st international conference on Very large databases*, 901-909, October 04-06, 2005, Italy, **2005.**

[138] Prasser, F., Bild, R., Eicher, J., Spengler, H., Kohlmayer, F., Kuhn, K. A., Lightning: Utility-driven anonymization of high-dimensional data, *Transactions on Data Privacy,* 9, 2, 161-185, **2016**.

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 11/07/2017

Tez Başlığı / Konusu: ρ-KAZANIM : MAHREMİYET KORUMALI FAYDA TEMELLİ VERİ YAYINLAMA MODELİ

Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 106 sayfalık kısmına ilişkin, 11/07/2017 tarihinde tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 2'dir.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar hariç/~~dâhil~~
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza
11.07.2017

Adı Soyadı:	YILMAZ VURAL
Öğrenci No:	N12149520
Anabilim Dalı:	BİLGİSAYAR MÜHENDİSLİĞİ
Programı:	BİLGİSAYAR MÜHENDİSLİĞİ
Statüsü:	☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Yrd. Doç. Dr. Murat AYDOS

(Unvan, Ad Soyad, İmza)