Journal of the Faculty of Engineering and Architecture of Gazi University 33:4 (2018) 1355-1368

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

ρ-Gain: A Utility Based Data Publishing Model

Yılmaz Vural, Murat Aydos
Department of Computer Engineering, Hacettepe University, Ankara, 06800, Turkey

Highlights:

- Privacy preserving data publishing
- Data anonymization and privacy models
- Data utility and outlier equivalence classes

Keywords:

- ρ-Gain model
- Data anonymization
- Data publishing
- Data privacy
- Data utility

Article Info:

Received: 14.03.2017
Accepted: 23.05.2017

DOI:

10.17341/gazimmfd.416433

Acknowledgement:

Correspondence:

Author:Yılmaz Vural
e-mail:
yilmazvural@gmail.com
phone: +90 312 297 7193

Abstract

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.

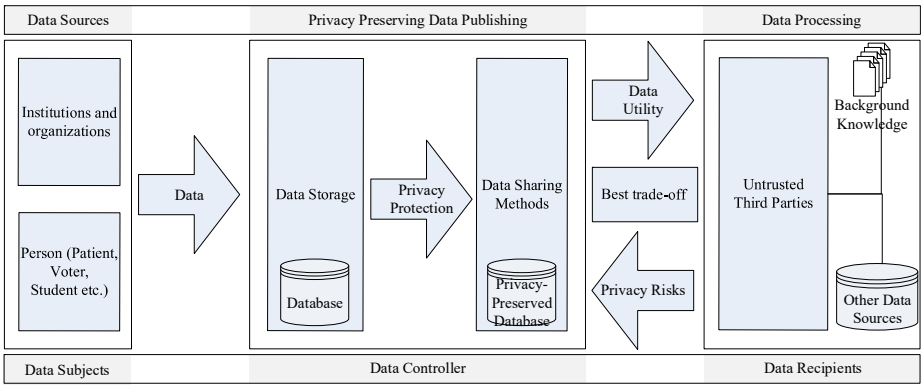


Figure A. Data publishing process

Purpose:
The aim of this study is to define and construct a new anonymization model for data publishing process given the above Figure A with high data utility and low privacy risk for data publishing purpose.

Theory and Methods:
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 and 1- Diversity privacy models were used together with ρ-iterations to reduce the privacy risks. The Average Equivalence Class metric was used to measure data utility. We have used two real-world datasets, most of which have been used in previous studies on data privacy. The datasets were prepared for the anonymization process. In this context, we removed the missing values and selected one SA and seven QID attributes of each dataset. To evaluate our model, we used ARX, which is a comprehensive open-source software for anonymizing data We performed two kinds of experiments to evaluate our solution. First, we compared our model with related work in terms of data utility. In these experiments, we used suppression limits of 100% in order to find the actual size of the OEC. In the second set of experiments, we showed that our model did not cause any negative impact on the privacy risk estimates.

Results:
Our proposed model was tested for data utility and privacy risks with two real-world datasets. Our experimental evaluation results for data utility and privacy risks were compared against previous studies. According to utility results, when our model was applied to the datasets, it was observed that the data utility is increased. Subsequently, privacy (re-identification) risks were not negatively affected by this process. In summary, the experimental results show that our model increased data utility without compromising privacy.
Conclusion:

The use of the OEC has a positive effect on the data utility, while causing no negative impact on privacy risks. One of the main contributions of our work is to improve data utility using the OEC.

Journal of the Faculty of Engineering and Architecture of Gazi University 33:4 (2018) 1355-1368

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

ρ-Kazanım: fayda temelli veri yayınlama modeli

Yılmaz Vural*, Murat Aydos
Hacettepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, Beytepe, Ankara, 06800,Türkiye

Ö N E Ç I K A N L A R

- Mahremiyet koruyucu veri yayınlama
- Anonimleştirme ve mahremiyet modelleri
- Aykırı kayıtlar ve veri faydası

Makale Bilgileri
Geliş: 14.03.2017
Kabul: 23.05.2017

DOI:
10.17341/gazimmfd.416433

Anahtar Kelimeler:
ρ-Kazanım modeli,
veri yayınlama,
veri anonimleştirme,
veri mahremiyeti,
veri faydası

ÖZET

Veri mahremiyeti, veri sahiplerinin mahremiyet riskleri ile veri paylaşımının taraflara sağlayacağı fayda arasındaki en iyi dengeyi bulmaya çalışan zor bir problemdir. Anonimleştirme teknikleri veri mahremiyeti probleminin çözümünde kullanılan etkin bir yöntemdir. Anonimleştirmeyle oluşan eşdeğer sınıflar veri faydasına göre fayda sağlayan ve aykırı olmak üzere iki sınıfa ayrılır. Faydalı eşdeğer sınıflar mahremiyet-fayda dengesini koruyan kayıtları içerir. Bu çalışmada, veri faydası olmayan kayıtları içeren aykırı eşdeğer sınıf üzerinde çalışılmıştır. Aykırı eşdeğer sınıf içerisinde yer alan kayıtların kazanımının veri faydasını artırmaya yönelik etkisini ortaya koyan, fayda temelli ρ-Kazanım modeli önerilmiştir. Önerilen model içerisinde mahremiyet risklerinin en aza indirgenmesinde k-Anonimlik ve l-Çeşitlilik mahremiyet modelleri birlikte kullanılmıştır. Eşdeğer sınıflar üzerinde işlemler yapıldığından veri faydasının ölçümünde eşdeğer sınıflar ortalaması metriği kullanılmıştır. Çalışma sonucunda elde edilen bulgulara göre, ρ-Kazanım modeli, veri faydasında iyileşmeyi sağlarken, mahremiyet risk tahminlerinde anlamlı bir olumsuzluğa yol açmamıştır. Veri mahremiyeti risklerini arttırmadan veri faydasını iyileştiren, fayda temelli ρ-Kazanım modelinin veri mahremiyeti probleminin çözümünde etkin bir rol oynayacağı gözlemlenmiştir.

ρ-Gain: a utility based data publishing model

H I G H L I G H T S

- Privacy preserving data publishing
- Data anonymization and privacy models
- Data utility and outlier equivalence Classes

Article Info
Received: 14.03.2017
Accepted: 23.05.2017

DOI:
10.17341/gazimmfd.416433

Keywords:
ρ-Gain model,
data anonymization,
data publishing,
data privacy,
data utility

ABSTRACT

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 and 1- Diversity 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

*Sorumlu Yazar/Corresponding Author: yvural@hacettepe.edu.tr / Tel: +90 0312 297 7193

1. GİRİŞ (INTRODUCTION)

Elektronik toplum olma yönünde hızla ilerlerken, sağlık, nüfus, finans, eğitim, yerel yönetimler, mülkiyet ve adli konularda hizmet veren elektronik uygulamaların kullanımı hızla yaygınlaşmaktadır. Uygulamalar aracılığıyla toplanan verilerin büyüklüğü ve çeşitliliği her geçen gün artmaktadır [1]. Toplanan veriler içerisinde demografik veriler, sağlık verileri, adli bilgiler, ticari bilgiler, tweetler, e-postalar, fotoğraflar, videolar ve konum bilgileri gibi kişisel ve hassas bilgilerde yer almaktadır [2, 3]. Yasal zorunluluklar, yeni ürünlerin geliştirilmesi, mevcut hizmet kalitesinin arttırılması, bilimsel araştırmaların yapılması ve kamuoyunun bilgilendirilmesi amacıyla toplanan bu veriler muhataplarıyla paylaşılır. Toplanan verilerin paylaşılması araştırmacılar ve kurumlara önemli fırsatlar sunarken mahremiyetle ilgili riskleri beraberinde getirir [4, 5]. Veri mahremiyeti, veri sahiplerinin mahremiyeti ile veri paylaşımının taraflara sağlayacağı fayda arasındaki en iyi dengeyi bulmaya çalışan zor bir problemdir. Yeterli mahremiyet önlemleri alınmadan yapılan veri paylaşımları sonucunda dünyada ve Türkiye'de mahremiyet ihlalleri yaşanmıştır [6-8].

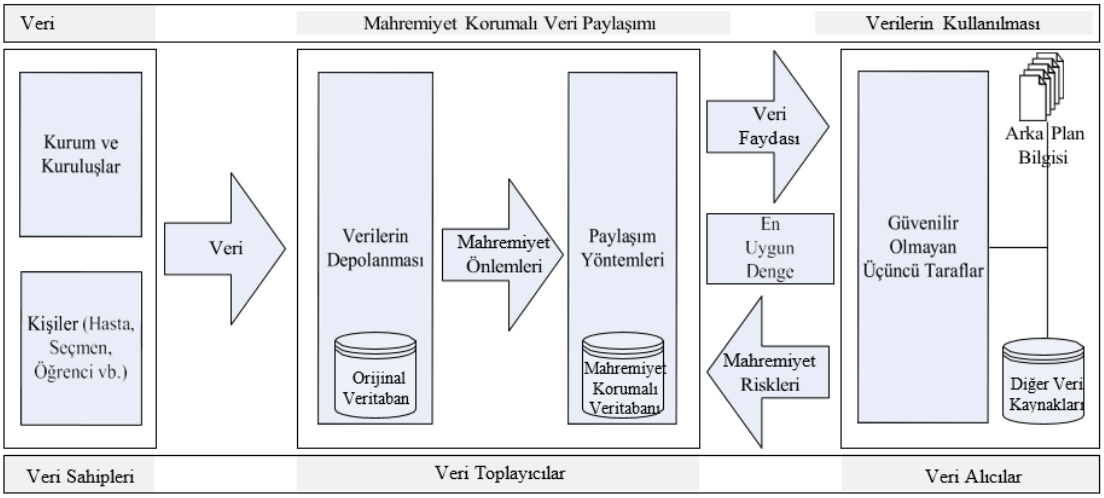
Veri mahremiyeti probleminin tarafları ile sorumluluklarının anlaşılabilmesi için veri toplama ve paylaşma sürecinin bilinmesi önemlidir. Verilerin toplanmasından paylaşımına kadar gerçekleşen süreci ifade etmek için bu çalışma kapsamında hazırlanan yeni bir çizim Şekil 1'de sunulmuştur. Bu süreçte sırasıyla veri sahipleri, veri toplayıcı ve veri alıcılar olmak üzere üç önemli taraf bulunmaktadır [9].

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 bildirimler veya diğer nedenlerle güvenilir olduklarını varsaydıkları veri toplayıcısına verilerini iletir. Veri toplayıcılar, veri sahiplerinin mahremiyetini koruyarak verilerin güvenli olarak, ilgililerine veya halka açık olarak paylaşılmasını sağlayan kişi, kurum ve kuruluşlardır. Güvenilir olduğu varsayılan veri toplayıcılar bu süreçte çeşitli yöntemlerle toplamış oldukları verileri depolamakta, kullanmakta ve paylaşmaktadır. Mahremiyet riskleri ile veri faydası arasındaki dengeyi sağlamakla yükümlü olan veri toplayıcılar veri toplanması ve paylaşılması sürecinde mahremiyetin korunmasından birinci derecede sorumludur.

Veri alıcılar, paylaşılan veriler üzerinde analizler veya işlemler yaparak verilerden ihtiyaçları doğrultusunda fayda sağlamaya çalışan güvenilir olmadığı varsayılan üçüncü taraflardır. Veri alıcısı kılığına girmiş saldırganlar, meraklı veya görevini kötüye kullanan gerçek ve tüzel kişilerin varlığı veri alıcılarını güvensiz kılar. Kötü niyetli veri alıcıların başka kaynaklardan elde ettiği veya daha önceden sahip olduğu verileri, mahremiyet korumalı verilerle eşleştirmesi sonucunda kimlik, öznitelik veya üyelik ifşalarının yapıldığı mahremiyet ihlalleri meydana gelir [9].

Veri toplayıcılar, topladıkları verileri çoğunlukla öznitelik ve kayıtlardan oluşan mikro veri tablosu biçiminde paylaşırlar. Mikro veri tablosundaki öznitelikler, her bir kayıtta yer alan muhatapları hakkında verdikleri bilgilere göre, kimlik tanımlayıcı öznitelik (Identifier-ID), birleşik tanımlayıcı öznitelik (Quasi Identifier-QID), hassas öznitelik (Sensitive Attribute-SA) ve hassas olmayan öznitelik (Non Sensitive Attribute-NSA) olmak üzere dört grupta sınıflandırılır [10]. Orijinal mikro veri tablosu T (ID, QID, SA, NSA), anonimleştirilmiş mikro veri tablosu T* (QID*, SA) biçiminde gösterilir.

Anonimleştirilmiş veri kümelerinde birleşik tanımlayıcı öznitelik (QID*) değeri aynı olan kayıtları içeren gruplar eşdeğer sınıflar (Equivalence Class -EC) olarak adlandırılır. Eşdeğer sınıflar EC (QID*, SA) biçiminde gösterilir. Eşdeğer sınıflar içerisinde yer alan kayıtların veri alıcıya sunduğu faydaya göre faydalı (Utility Equivalence Class-



Şekil 1. Verilerin toplanması ve paylaşılması (Data collecting and sharing)

UEC) ve aykırı (Outlier Equivalence Class -OEC) [11] sınıflar olmak üzere ikiye ayrılır. UEC fayda-mahremiyet dengesini sağlayan kayıtları içerir. Aykırı eşdeğer sınıflar tamamen baskılanmanın etkisiyle en yüksek seviyedeki mahremiyet korunmasına sahip fayda sağlamayan kayıtları içerir.

1.1. Anonimleştirme ve Mahremiyet Modelleri
(Anonymization and Privacy Models)

Anonimleştirme, verinin tipi ve biçimi korunarak veri faydası açısından kabul edilebilir düzeyde yapılan veri detayını azaltma (reduction) temelli mahremiyet koruyucu dönüştürme işlemidir. Genelleştirme ve baskılama yaygın olarak kullanılan anonimleştirme tekknikleridir [12]. Genelleştirme, özniteliklere ait değerlerin biçimsel ve anlamsal bütünlüğü korunarak daha az detay içerecek şekilde yeniden ifade edilmesidir. Örnek genelleştirme hiyerarşileri Şekil 2'de verilmiştir.

Şekil 2a'da sayısal Yaş özniteliği 5 seviyede, Şekil 2b'de kategorik Cinsiyet özniteliği 2 seviyede, Şekil 2c'de kategorik Ülke özniteliği 4 seviyede genelleştirilmiştir. Alt seviyeden üst seviyeye doğru genelleştirildikçe ($G_i \rightarrow G_{i+1}$) veri faydası azalmakta, mahremiyet koruması artmaktadır. Veri faydası ve mahremiyet dengesine göre farklı hiyerarşi seviyelerinden seçilecek öznitelikleri gösteren düğümler çözüm için kullanılır. Düğümlerin tamamı, ardıl ve öncül ilişkilerini gösteren Hasse çizgeleriyle [13] ifade edilir. Bir diğer anonimleştirme tekniği olan baskılama [12], seçilen özel bir karakterle (* vb.) öznitelik değerlerinin değiştirilmesini sağlar. Genelleştirme ve baskılamanın birlikte kullanılması anonimleştirmede başarılı sonuçlar verir.

Veri mahremiyeti gereksinimlerinin sağlanması amacıyla anonimleştirme tekniklerini kullanan mahremiyet modellerine ihtiyaç duyulur. k-Anonimlik (k- Anonymity) [15], l-Çeşitlilik (l-Diversity) [16] ve t-yakınlık (t-Closeness) [17] modelleri yaygın olarak kullanılır. k-Anonimlik [15], anonimleştirilen verilerin en az k sayıda kayıt içeren eşdeğer sınıflardan oluşmasını garanti eder. k-Anonimlik modeli kimliklerin ifşa edilmesine karşı koruma sağlarken özniteliklerin ifşa edilmesine karşı koruma sağlamaz. l-Çeşitlilik [16], eşdeğer sınıflar içerisindeki SA değerlerinin en az l çeşitlilikte olmasını garanti eder. l-Çeşitlilik modelinin anlamsal yakınlıkları ayırt edememesi üzerine t-yakınlık modeli önerilmiştir [17].

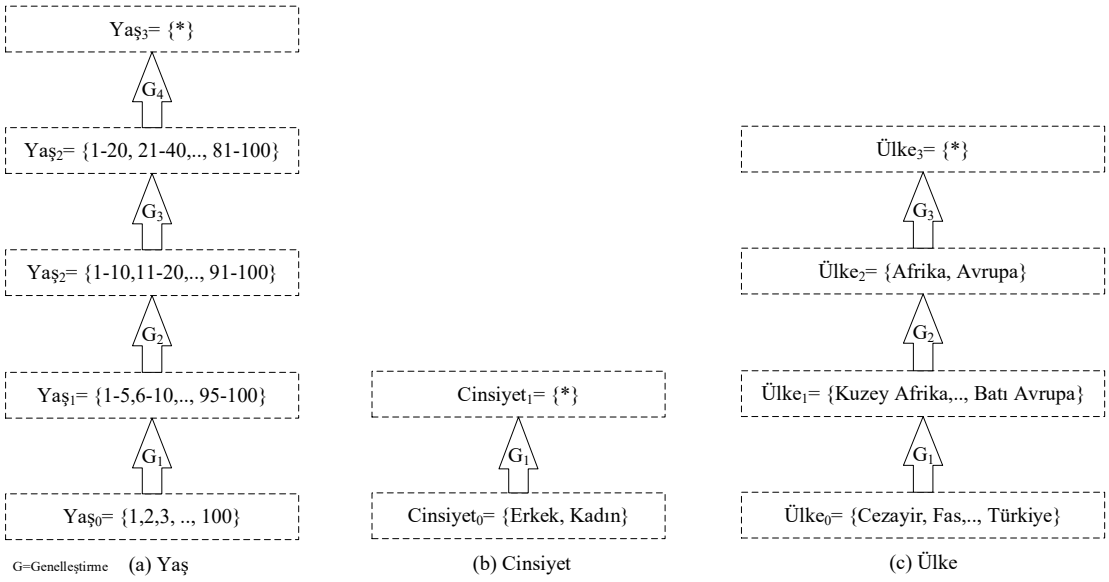
1.2. Veri Faydası ve Mahremiyet Riskleri
(Data Utility and Privacy Risks)

Veri faydası, anonimleştirme sırasında veride meydana gelen kayıplar ile ölçülür [18,19]. Veri kayıpları eşdeğer sınıf veya öznitelik temelli metrikler kullanılarak hesaplanır. Veri kayıpları en aza indirildiğinde veri faydası en yüksek olacaktır. Veri kaybının ölçümünde yaygın olarak kullanılan metrikler ile öne çıkan özellikleri takip eden paragraflarda özetlenmiştir [20].

Ortalama Eşdeğer Sınıf Sayısı (Avarage Equivalence Class Size - AECS): Anonimleştirmeyle oluşan eşdeğer sınıfların sayısının büyüklüğüne bağlı olarak veri kaybını ölçen bu metrik, LeFevre ve arkadaşları tarafından önerilmiş olup $|T|$ anonim verideki kayıt sayısını, H anonimleştirilmiş veri kümesindeki eşdeğer sınıf sayısını ve k en küçük eşdeğer sınıftaki kayıt sayısını göstermek üzere Eş.1'deki gibi hesaplanır [21].

$$AECS_{Loss} = (|T|/H)/k \quad (1)$$

Ayırt Edilebilirlik (Discernibility Method): Ayırt edilebilirlik metriği (DM, eşdeğer sınıflar içerisindeki her bir kaydın diğer kayıtlardan farklılığına göre ceza verilmesi



Şekil 2. Örnek genelleştirme hiyerarşileri (Generalization hirearchies examples)

esasına göre veri kaybını ölçer. DM yönteminde very kaybı, eşdeğer sınıfların sayısı $|E|$, olmak üzere Eş.2'deki gibi hesaplanır [22].

$$DM(T)=\Sigma(|E|) \quad (2)$$

Minimal Bozulma (Minimal Distortion): Minimal bozulma yöntemine göre genelleştirilen özniteliklere ceza puanı uygulanır. Uygulanan ceza puanlarının toplamı her bir öznitelik için veri kaybının hesaplanmasında kullanılır [23-25].

Bilgi Kaybı (Information Loss): Kategorik öznitelikler için veri kaybının ölçülmesi amacıyla Lyengar tarafından önerilmiştir [26]. Bu metriğe göre veri kaybı $|V_g|$, ilgili düğüme ait çocukların sayısı, $|DA|$ ise V_g'nin A niteliğindeki değerlerinin sayısı olmak üzere Eş. 3'deki gibi hesaplanır.

$$ILoss(v_g) = |V_g|-1/|DA| \quad (3)$$

Veri mahremiyetinin sağlanmasındaki bir diğer önemli bileşen kimlik veya öznitelik bilgilerinin ifşa risklerinin hesaplanmasıdır. Bu risklerin hesaplanmasında savcı (prosecutor), gazeteci (journalist) ve pazarlamacı (marketer) olmak üzere 3 temel yaklaşım kullanılır [27].

Savcı yaklaşımında, saldırgan yayınlanan veri içerisinde kurbanın olduğunu bilir. Bu varsayım altında n yayınlanan kayıtların toplam sayısı olmak üzere, j eşdeğer sınıfı, f_j kurbanla eşleşen eşdeğer sınıfların sayısını gösterdiğinde savcı yaklaşımlı risk Eş. 4'deki gibi hesaplanır [27].

$$_p\theta_j = 1/f_j \quad (4)$$

Gazeteci yaklaşımında, saldırgan yayınlanan veri içerisinde kurbanın olup olmadığını bilemez. Gazeteci riskinin uygulanabilmesi için yayınlanan verilerin bütününü temsil etmeyen örnek bir veri kümesine ihtiyaç duyulur. Riskin hesaplanmasında örnek küme ile tanımlama veritabanı eşleştirilir. Tanımlama veritabanında denklik sınıfı j, kayıt sayısı F_j olmak üzere gazeteci yaklaşımlı riski Eş. 5'deki gibi hesaplanır [27].

$$_j\theta_j = 1/F_j \quad (5)$$

Pazarlamacı yaklaşımında, saldırgan anonim veri içerisinde herhangi bir kayıt yerine mümkün olan en yüksek sayıda kaydın kimliğinin yeniden tanımlanmasına çalışır. Savcı ve gazeteci yaklaşımları bireye yönelik risk ölçümü yaparken, pazarlamacı yaklaşımı daha çok toplulukla ilgilenir. Örnek veritabanında eşdeğer sınıf j, eşdeğer kayıt sayısı f_j, tanımlama veritabanında eşdeğer sınıf J, eşdeğer kayıt sayısı F_j olmak üzere pazarlamacı yaklaşımlı risk Eş.6'daki gibi hesaplanır [28].

$$_m\theta_j = f_j/F_j \quad (6)$$

Bu çalışmada, verilerin toplanmasından paylaşılmasına kadar olan süreç içerisinde mahremiyet problemi açıklanmış, anonimleştirme yöntemleri ve yaygın olarak kullanılan mahremiyet modelleri incelenmiştir. Veri faydasının hesaplanması ile mahremiyet risklerinin tahmininde kullanılan metrikler gözden geçirilmiştir. Eşdeğer sınıflardan bahsedilerek UEC ve OEC kavramları açıklanmıştır. ρ-Kazanım olarak adlandırılan veri faydasını arttırıcı fayda temelli yeni bir model önerilmiştir. Önerilen model veri faydası ve mahremiyet koruması açısından farklı modellerle karşılaştırılmış ve başarımı değerlendirilmiştir.

Bu amaçla aşağıdaki sorulara cevap aranmıştır.

- Eşdeğer sınıf ayrımının veri faydası ve mahremiyet riskleri üzerindeki etkisi nedir?
- Aykırı eşdeğer sınıfın geri kazanımının veri faydası ve mahremiyet riskleri üzerindeki etkisi nedir?

2. YÖNTEM (METHOD)

Çalışma kapsamında mahremiyet riskleri ve veri faydası dengesini koruyarak veri faydasını arttırıcı yeni bir model önerilmiştir. Önerilen model içerisinde k-Anonimlik ve l-Çeşitlilik modelleri ifşa risklerinin azaltılmasında kullanılmıştır. Veri faydasının ölçümünde eşdeğer sınıflar ortalaması (AECS) metriği kullanılmıştır.

$$T^*=\rho\text{-}Kazanım\ (T) \quad (7)$$

Eş.7 veri faydasını arttırmak için paylaşılacak veri setine uygulanır. Eş.7'de T başlangıç veri seti, T* anonimleştirilmiş veri setini ve ρ iterasyon sayısını gösterir. Tablo 1'de önerilen modelin algoritması verilmiştir.

ρ-Kazanım, yüksek veri faydasına sahip mahremiyet korumalı kayıtları veri alıcıları için yayınlar. Başlangıçta mikro veri tablosunda yer alan öznitelikler (A_1, A_2, A_3,.., A_n) muhatapları hakkında verdikleri bilgilere göre (ID, QID, SA, NSA) sınıflarına ayrılır (Adım 2). ID ve NSA sınıfında yer alan öznitelikler yayınlanacak veri kümesinden çıkartılır (Adım 3). QID ve SA özniteliklerinden oluşan veri kümesinin başlangıçta bir kopyası alınır (Adım 4). Bu tablo anonimleştirilmiş aykırı kayıtların orijinal haline getirilebilmesi için kullanılır. QID içerisinde yer alan özniteliklerin genelleştirme hiyerarşileri oluşturulur (Adım 5-6). Mahremiyet modelleri için gerekli olan parametreler (k, l, ρ) belirlenerek anonimleştirme öncesi hazırlıklar tamamlanır (Adım 7-10).

QID özniteliklerinin anonimleştirmesinde k-Anonimlik ve l-Çeşitlilik algoritmaları birlikte kullanılır (Adım 12). Anonimleştirme sonucunda oluşan veri kümesinde ($S^*_{dataset}$) en az k sayıda kayıt içeren eşdeğer sınıflardan (Equivalence Class- EC) oluşur. Eşdeğer sınıflar içerisinde anonimleştirilmiş birleşik tanımlayıcı kümesi (QID*) ile l-Çeşitlilik gereksinimini sağlayan SA özniteliği yer alır. Tüm eşdeğer sınıflar (All Equivalence Class- TEC) faydalı (UEC) ve aykırı (OEC) kayıtları içermesine göre iki sınıfa ayrılır (Adım 13). UEC içerisindeki QID* ve SA özniteliklerine sahip kayıtlar anonim paylaşım tablosuna ($T^*_{dataset}$) taşınır (Adım15). Veri faydasını arttırmada kullanılacak OEC

Tablo 1. ρ-Kazanım modeli algoritması (ρ-Gain model algorithm)

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 ρ for ρ-Kazanım
10	ρ' = 0
11	do {
12	$S^{*}_{dataset}$ {QID*} ←Anonimleştir ($S_{dataset}$ {QID}, k-Anonimlik, l-Çeşitlilik)
13	$TEC_{dataset}$ {UEC, OEC}←Eşdeğer_Sınıfları_Oluştur ($S^{*}_{dataset}$ {QID*})
14	for all $TEC_{dataset}$ {UEC, OEC} do
15	$T^{*}_{dataset}$ {QID*, SA}← Faydalı_Kayıtlar ($TEC_{dataset}$ {UEC})
16	$TEC'_{dataset}$ {QID*, SA}← Aykırı_Kayıtlar ($TEC_{dataset}$ {OEC})
17	Tablo_Hazırla ($S_{dataset}$ {QID}, SA)
18	$S_{dataset}$ {QID, SA}← Aykırı_Kayıtları_Eşleştir ($TEC'_{dataset}$ {QID*,SA}; $S'_{dataset}$ {QID, SA})
19	ρ'= ρ' + 1
20	} while (ρ' = ρ)
21	Veri_Yayınla ($T^{*}_{dataset}${QID*}, SA})

içerisindeki kayıtlar aykırı kayıtlar tablosuna ($TEC'_{dataset}$) taşınır (Adım16). Aykırı kayıtları içeren $TEC'_{dataset}$ tablosu orijinal kayıtları içeren $S'_{dataset}$ tablosu yardımıyla anonimleştirme öncesine döndürülerek, yeniden anonimleştirme için hazırlanan $S_{dataset}$ tablosuna taşınır (Adım17-18). Veri faydasını arttırmak amacıyla ρ iterasyon defa bu işlem uygulanır (Adım 11-20). Veri faydası arttırılan $T^{*}_{dataset}$ tablosu veri alıcılar için yayınlanır (Adım 21).

2.1. Deneyler (Experimental setup)

Modelin uygulanması için yapılacak deneylerde, literatürde daha önceki çalışmalarda yaygın olarak kullanılan [29-31] k=5 ve l=2 değerleri seçilmiştir. Deneyler, 64-bit Oracle JVM çalıştıran, dört çekirdekli 2,6 GHz Intel Core i7 CPU'ya sahip ARX 3.5.1 çalışan masaüstü bilgisayarda yapılmıştır.

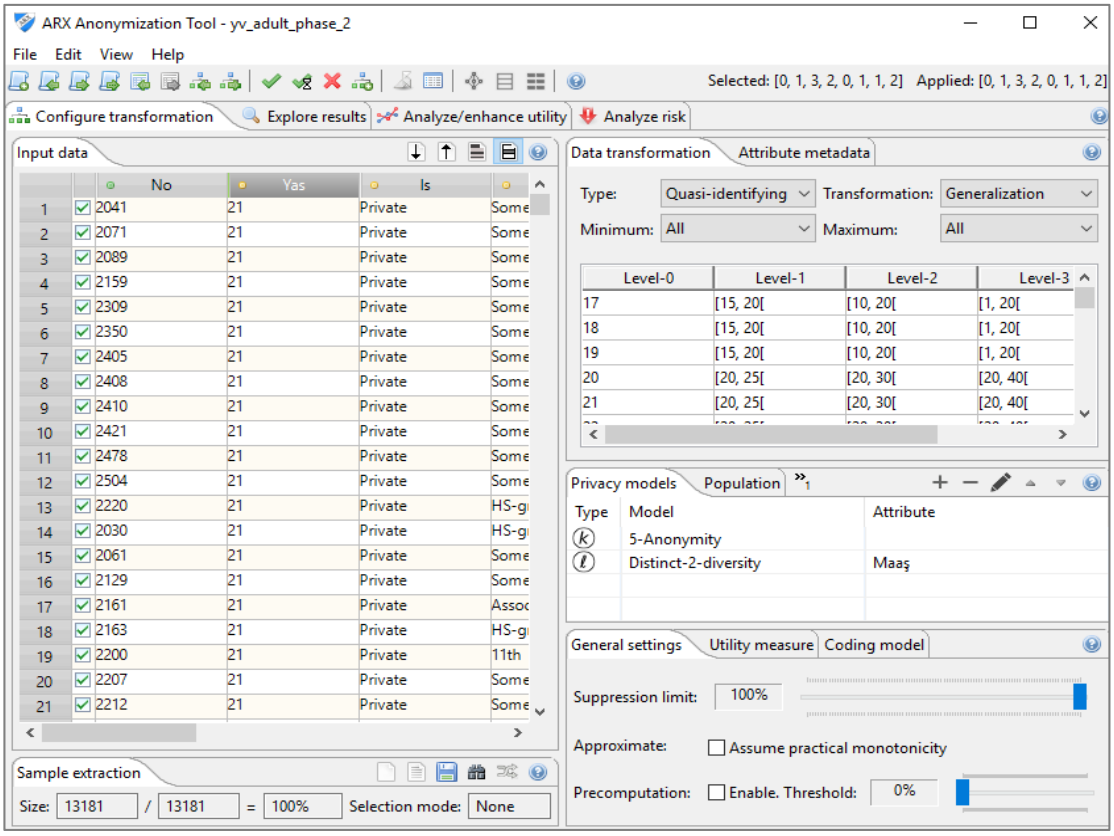
2.1.1. ARX mahremiyet aracı (ARX privacy tool)

Deneyler için literatürde birçok çalışmada kullanılan açık kaynaklı ARX aracı seçilmiştir [29-31]. ARX, mahremiyet ihlallerine neden olabilecek saldırıları hafifletmek veya önlemek amacıyla anonimleştirme yöntemlerini içeren veri faydası ile risklerin mahremiyet modellerine göre analiz edilmesini sağlayan açık kaynak kodlu Java programlama dilinde geliştirilmiş bir yazılımdır. ARX öznitelikleri sınıflandırarak, genelleştirme, baskılama veya kümeleme teknikleriyle anonimleştirilmesini sağlar. QID genelleştirme örüntüsünü oluşturarak, örüntü içinde arama alanını azaltmak için çoklu budama yapan Flash algoritmasını kullanır [29].

2.1.2. Veri seti (Data set)

Deneylerde, Barry Becker tarafından 1994 Nüfus Sayımı veri tabanından çıkartılan Adult [32] veritabanı kullanılmıştır. Veriler gözden geçirilerek düzenlenmiştir. Şekil 3'de gösterildiği gibi ARX aracılığıyla üzerinde işlem yapılmak üzere hazırlanmıştır. T_{adult} veri setinde yer alan öznitelikler sınıflandırılarak genelleştirme seviyeleri ile birlikte Tablo 2'de gösterilmiştir.

Veri kümelerindeki öznitelikler arası ilişkinin dikkate alınması sınıflandırma başarısını etkileyen önemli bir etkendir [33]. Hassas öznitelik tanımlaması üzerinde veri faydası nedeniyle herhangi bir işlem yapılmamıştır. Veri kümesi içerisindeki farklı değerlere bakılarak SA seçimi yapılmıştır. T_{adult} veri setinde yer alan birleşik tanımlayıcı sınıfındaki özniteliklerden Yaş özniteliğine ait genelleştirme hiyerarşisi Şekil 4'de, İş özniteliğine ait genelleştirme hiyerarşisi Şekil 5'de ve Eğitim özniteliğine ait genelleştirme hiyerarşisi ise Şekil 6'da verilmiştir. Şekil 4'de sayısal Yas



Şekil 3. T_{adult} Veri setinin hazırlanması (T_{adult} Data set preparation)

Tablo 2. Adult veri seti (Adult data set)

Veri Seti Adı	QID (Genelleştirme Seviyesi)	SA
T_{ADULT}	Cinsiyet (2), Yas (5), Irk (2), Medeni Durum (3), Egitim (4), Ulke (3), Is (3), Pozisyon (3), Eğitim (4)	Maaş

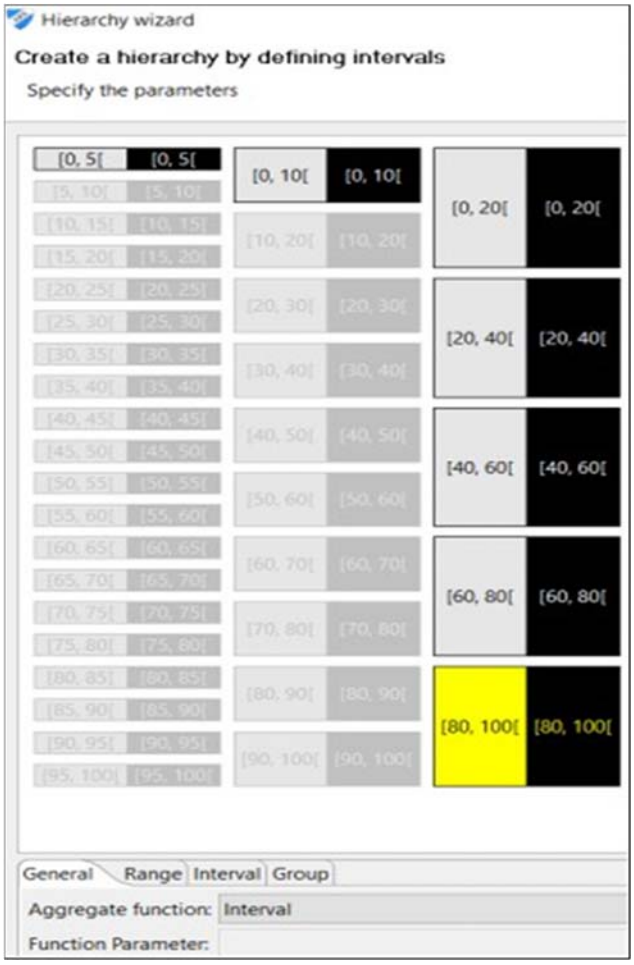
özniteliği aralıklara ayrılarak, Şekil 5 ve Şekil 6'da gösterilen kategorik İş ve Egitim özniteliği daha az detaylara sahip olacak şekilde sınıflandırılarak genelleştirme hiyerarşileri oluşturulmuştur. T_{adult} veri seti hazırlandıktan sonra ρ-Kazanım öncesi AECS metriğine göre veri faydası sonuçları Şekil 7'de verilmiştir. Veri kaybı azaldıkça veri faydası artacaktır. Şekil 7'de verilen ρ-Kazanım öncesi veri faydası ölçümü 25,47466 olarak bulunmuştur. Bu ölçüm sonucunda toplam 1184 adet eşdeğer sınıf (UEC ve OEC) oluşmuştur. Veri mahremiyeti probleminin çözümünde bir diğer parametre risk tahminleridir. Savcı, gazeteci, pazarlamacı saldırgan modelleri için ρ-Kazanım öncesi risk tahminleri Şekil 8'de verilmiştir.

Şekil 8'de verilen ρ-Kazanım öncesi risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Ayrıca sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 6,96 olduğu görülmektedir. Bu çalışmada saldırgan kurbanın yayınlanan veri seti içerisinde olduğu bilgisine sahip olduğu varsayıldığından ρ-Kazanım öncesi savcı risk grafiği Şekil 9'da verilmiştir Grafik incelendiğinde kayıtların %4,24'nün yüksek risk aralığında (16,7-20) olduğu ve başlangıç riskinin %20 olduğu görülmektedir.

3. SONUÇLAR VE TARTIŞMALAR
(RESULTS AND DISCUSSIONS)

2-Kazanım (T_{Adult}) için elde edilen sonuçlara ait bulgular bu bölümde verilmiştir. Öznitelik genelleştirmesi sonrasında T veri seti literatürde mahremiyeti sağlamaya yönelik en sık kullanılan 5-Anonimlik ve 2-Çeşitlilik modelleri ile anonimleştirilmiş ve sonuçlar aşağıda verilmiştir.

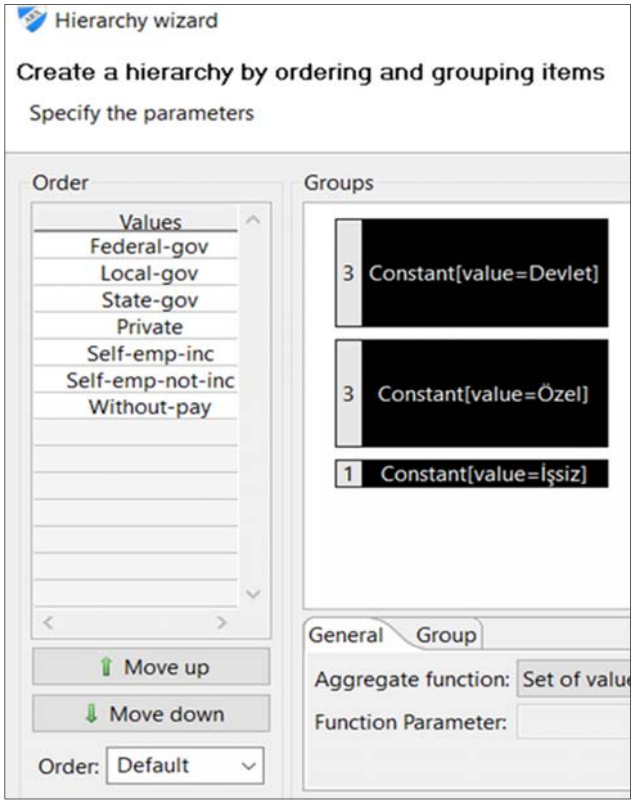
Şekil 10'da çözüm düğümüne göre AECS bilgi metriği ile elde edilen veri faydası sonuçları OEC dikkate alınarak gösterilmiştir. Şekil 10 incelendiğinde 1184 adet eşdeğer sınıf içerisinde, en az 5, en fazla 13181 ve ortalama %25,47466 kaydın yer aldığı görülmüştür. ρ-Kazanım öncesi veri faydasına katkı sağlayan 1183 adet UEC bulunmuştur. UEC içerisindeki 16981 kayıt yayınlanacak T* kümesine aktarılmıştır. Veri faydası olmayan ve tamamen bastırılan toplam verinin %43,7'sini oluşturan OEC içerisinde 13181 kayıt olduğu görülmüştür.



Şekil 4. Yaş özniteliği genelleştirme hiyerarşisi
(Generalization hierarchy of age attribute)

Şekil 10'da verilen fayda sonuçlarına göre ρ'=0 durumundaki veri faydası %25, 47466 olarak bulunmuştur. ρ'=0 durumunda aykırı eşdeğer sınıf kayıtları Şekil 11'de gösterilmiştir. Şekil 11 incelendiğinde resmin sağ tarafında tamamen bastırılan OEC kayıtları sol tarafında ise orijinal halleri görülmektedir. Çözüm düğümüne göre savcı, gazeteci, pazarlamacı saldırgan modelleri için başlangıç risk tahminleri Şekil 8'de daha önce verilmiştir. Şekil 8'de verilen risk tahminlerine göre en yüksek risk olan %20 ρ'=0 durumundaki risk olarak kabul edilmiştir. ρ'=0 durumunda tamamen bastırılan veri faydası olmayan 13181 adet aykırı kayıt ρ-Kazanım algoritmasına göre orijinal haline getirilmiştir. ρ-Kazanım modeline göre ρ=1 durumu için 13181 adet kayıt yeniden anonimleştirilmiş ve Şekil 12'de gösterilmiştir

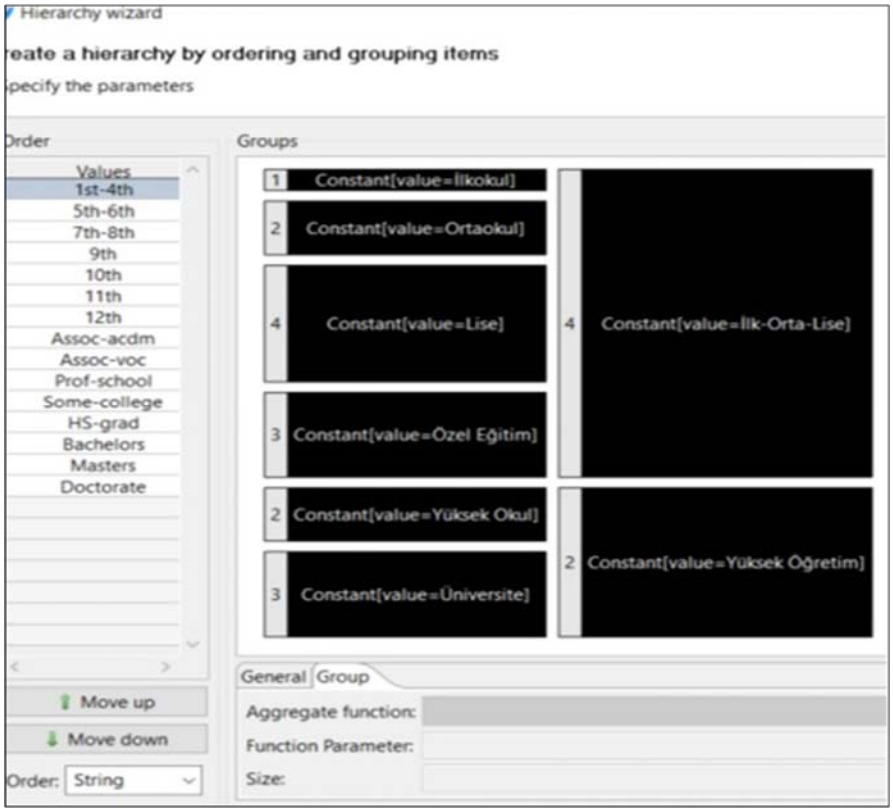
Şekil 12'ye göre ρ=1 durumunda elde edilen bulgularda 480 adet EC olduğu görülmüştür. En büyük EC'nin 7796 kayıt içeren OEC olduğu tespit edilmiştir. 5385 adet kayıt içeren UEC'deki kayıtlar ise yayınlanacak olan T* içerisine taşınmıştır. ρ=1 durumunda OEC kayıtlarının %40,85'nin faydaya dönüştüğü tespit edilmiştir. ρ=1 durumu için risk tahminleri Şekil 13'de ve savcı risk grafiği Şekil 14'de verilmiştir.



Şekil 5. İş özniteliği genelleştirme hiyerarşisi
(Generalization hierarchy of work attribute)

Şekil 13'de verilen ρ=1 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Ayrıca sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 8.89 olduğu görülmektedir. ρ=1 için savcı risk grafiği Şekil 14'de verilmiştir. Grafik incelendiğinde kayıtların %4,27'sinin yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=1 durumunda en yüksek riskin değişmediği görülmüştür. ρ=2 durumunda veri faydası Şekil 15'de verilmiştir. Şekil 15'e göre ρ=2 durumunda elde edilen bulgularda 119 adet EC olduğu görülmüştür. En büyük EC'nin 6060 kayıt içeren OEC olduğu tespit edilmiştir. 1736 adet kayıt içeren UEC'deki kayıtlar ise yayınlanacak olan T* içerisine taşınmıştır. ρ=2 durumunda OEC kayıtlarının %22,26'sının faydaya dönüştüğü tespit edilmiştir. ρ=2 durumu için risk tahminleri Şekil 16'da ve savcı grafiği Şekil 17'de verilmiştir.

Şekil 16'da verilen ρ=2 durumundaki risk tahminleri ve bu tahminlerden etkilenecek kayıt sayıları yüzdelik olarak verilmiştir. Ayrıca sırasıyla savcı, gazeteci ve pazarlamacı risklerinin %20, %20 ve % 6.79 olduğu görülmektedir. ρ=2 için savcı risk grafiği Şekil 17'de verilmiştir. Grafik incelendiğinde kayıtların %4,14'nün yüksek risk aralığında (16,7-20) olduğu görülmektedir. Bulgularda ρ=2 durumunda en yüksek riskin değişmediği görülmüştür. Yapılan deneysel değerlendirmelere ek olarak Tablo-3'de ρ-kazanım modelinin literatürde yaygın olarak kullanılan diğer modellerle veri faydası ve mahremiyet koruması açısından karşılaştırılması gösterilmiştir.



Şekil 6. Eğitim özniteliği genelleştirme hiyerarşisi (Generalization hierarchy of education attribute)

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

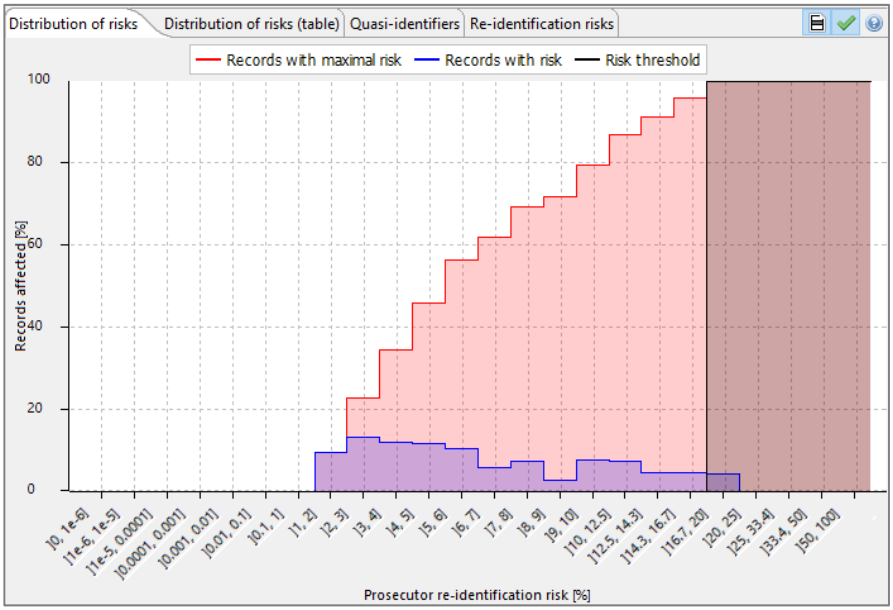
Measure	Including outliers
Average class size	25.47466 (0.08446%)
Maximal class size	13181 (43.70068%)
Minimal class size	5 (0.01658%)
Number of classes	1184
Number of records	30162
Suppressed records	13181 (43.70068%)

Şekil 7. Veri faydası sonuçları (Data utility results)

Re-identification risks | Population uniques | Population

Measure	Value [%]
Lowest prosecutor risk	1.25%
Records affected by lowest risk	0.47111%
Average prosecutor risk	6.96661%
Highest prosecutor risk	20%
Records affected by highest risk	4.24003%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	6.96661%

Şekil 8. ρ-Kazanım öncesi riskler (Risks before ρ-Gain)

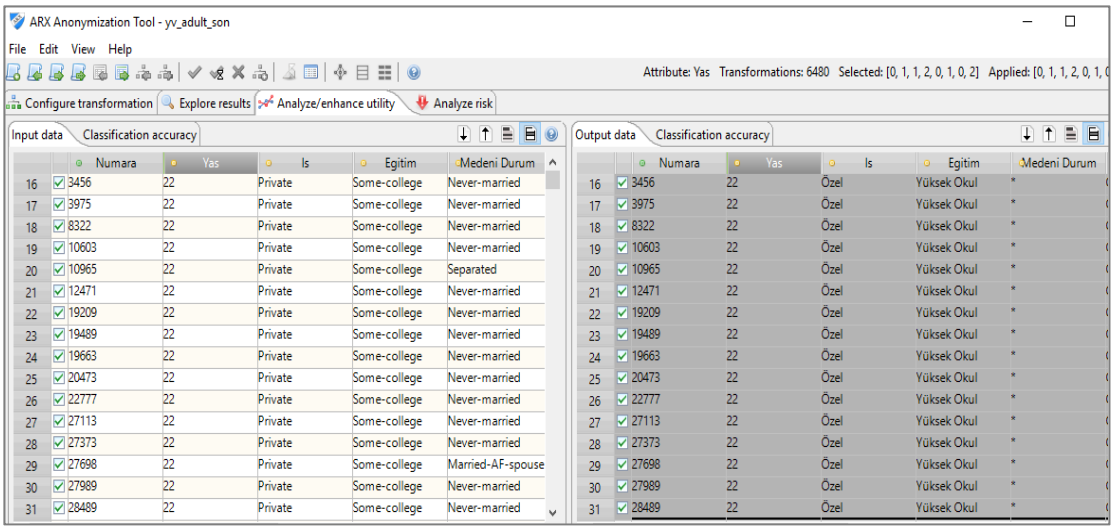


Şekil 9. ρ-Kazanım öncesi savcı riski (Prosecutor risk before ρ-Gain)

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

Measure	Including outliers	Excluding outliers
Average class size	25.47466 (0.08446%)	14.35418 (0.08453%)
Maximal class size	13181 (43.70068%)	80 (0.47111%)
Minimal class size	5 (0.01658%)	5 (0.02944%)
Number of classes	1184	1183
Number of records	30162	16981 (56.29932%)
Suppressed records	13181 (43.70068%)	0

Şekil 10. ρ'=0 durumunda OEC (OEC for ρ'=0)



Şekil 11. ρ'=0 durumunda OEC kayıtları (OEC records for ρ'=0)

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

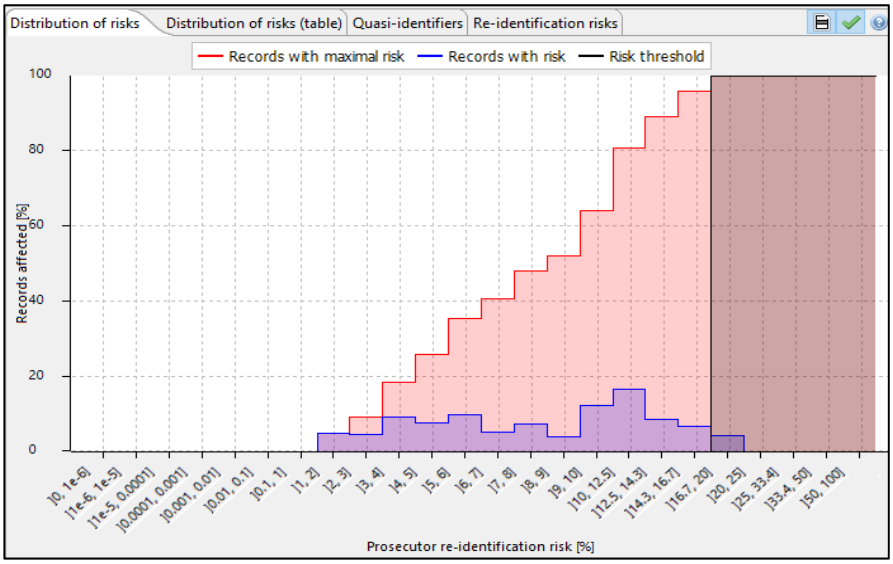
Measure	Including outliers	Excluding outliers
Average class size	27.46042 (0.20833%)	11.24217 (0.20877%)
Maximal class size	7796 (59.14574%)	80 (1.48561%)
Minimal class size	5 (0.03793%)	5 (0.09285%)
Number of classes	480	479
Number of records	13181	5385 (40.85426%)
Suppressed records	7796 (59.14574%)	0

Şekil 12. ρ=1 durumunda veri faydası sonuçları (Data utility results for ρ=1)

Re-identification risks | Population uniques | Population

Measure	Value [%]
Lowest prosecutor risk	1.25%
Records affected by lowest risk	1.48561%
Average prosecutor risk	8.89508%
Highest prosecutor risk	20%
Records affected by highest risk	4.27112%
Estimated prosecutor risk	20%
Estimated journalist risk	20%
Estimated marketer risk	8.89508%

Şekil 13. ρ=1 için risk tahminleri (Estimated risks for ρ=1)



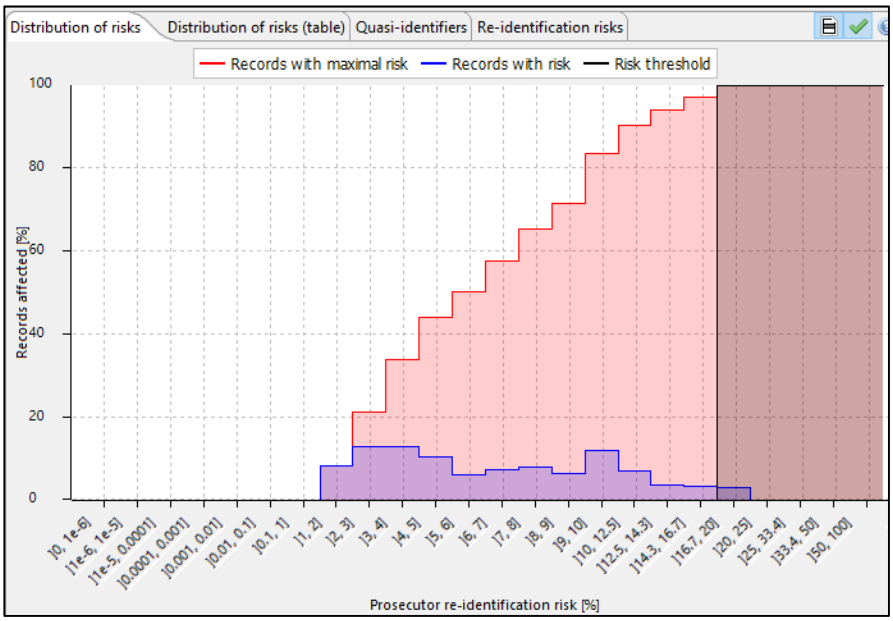
Şekil 14. ρ=1 için savcı risk grafiği (Prosecutor risk for ρ=1)

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

Measure	Including outliers	Excluding outliers
Average class size	65.51261 (0.84034%)	14.71186 (0.84746%)
Maximal class size	6060 (77.73217%)	72 (4.14747%)
Minimal class size	5 (0.06414%)	5 (0.28802%)
Number of classes	119	118
Number of records	7796	1736 (22.26783%)
Suppressed records	6060 (77.73217%)	0

Şekil 15. ρ=2 durumunda veri faydası sonuçları (Data utility results for ρ=2)

Şekil 16. ρ=2 için risk tahminleri (Estimated risks for ρ=2)



Şekil 17. ρ=2 için savcı risk grafiği (Prosecutor risk for ρ=2)

Tablo 3. Mahremiyet modellerinin karşılaştırılması (Comparison of privacy models)

Model	Veri Kaybı (AECS)	Koruma	EC Kullanımı
k-Anonimlik (k=5)	16,68	Kimlik İfşası	UEC
k-Anonimlik ve l-Çeşitlilik (k=5, l=2)	25,47	Kimlik ve Nitelik İfşası	UEC
k-Anonimlik ve t-Yakınlık (k=5, t=0.001)	33,33	Kimlik ve Nitelik İfşası	UEC
ρ-Kazanım (k=5, l=2, ρ=2)	13,54	Kimlik ve Nitelik İfşası	UEC ve OEC

Tablo-3'de k-Anonimlik yöntemi tek başına uygulandığında veride yaşanan kaybın % 16,68 olduğu tespit edilmiştir. Ancak bu yöntem yayınlanan verileri kimlik ifşalarına karşı korunurken öznitelik ifşalarına karşı korunmasız durumdadır. k-Anonimlik ve l-Çeşitlilik modelleri birlikte uygulandığında veride yaşanan kaybın % 25,47 olduğu tespit edilmiştir. k-Anonimlik ve t-Yakınlık modelleri birlikte uygulandığında veride yaşanan kaybın % 33,33 olduğu tespit edilmiştir. Son olarak ρ-kazanım modeli uygulandığında veride yaşanan kaybın % 13,54 olduğu tespit edilmiştir. Son üç model kimlik ve nitelik ifşalarına karşı koruma sağlamaktadır. Ancak ρ-Kazanım modeli diğer modellerden farklı olarak OEC üzerinde işlem yapabilmektedir. Bulgulara göre OEC üzerinde yapılan işlemlerin veri faydasına olan katkısı görülmüş ve yüksek seviyedeki mahremiyet risk değerinin değişmediği tespit edilmiştir.

4. SONUÇLAR (CONCLUSIONS)

Mahremiyet korumasından ödün vermeden veri faydasının arttırılması veri paylaşımlarının başarısı için önemli bir gelişmedir. Bu çalışmada mahremiyet riskleri ve veri faydası arasındaki denge korunarak veri faydasının arttırılması problemi üzerinde durulmuştur. Veri kalitesini bozan ve

eşdeğer sınıflar içerisindeki gruplandırmaları olumsuz etkileyen aykırı kayıtların veri faydasına olan etkisi çalışılmıştır. Bu çalışmada eşdeğer sınıflar kendi arasında veri faydası olan (UEC) ve veri faydası olmayan (OEC) şeklinde ikiye ayrılmıştır. Veri faydası olmayan aykırı kayıtları içeren OEC üzerinde veri faydasını arttırıcı işlemler yapılmıştır. Ayrıca OEC üzerinde yapılan bu işlemlerin mahremiyet risklerine olan etkisi araştırılmıştır. Bu doğrultuda, OEC üzerinde işlem yapabilen yeni bir model olan ρ-Kazanım modeli tanımlanmış ve test edilmiştir. ρ-Kazanım modelinin başarımının değerlendirilebilmesi amacıyla literatürde yaygın olarak kullanılan k-Anonimlik, l- Çeşitlilik ve t-Yakınlık modelleriyle veri faydası ve mahremiyet koruması açısından karşılaştırılmıştır. Elde edilen sonuçlara göre, başlangıçta toplam verinin yarıya yakın bir oranda aykırı kayıt içerdiği ve veri faydası sağlamadığı görülmüştür. Başlangıçta mahremiyet risk tahmini ise %20 olarak bulunmuştur. Önerilen modelin uygulanmasıyla ilk iterasyonda elde edilen bulgularda veri faydasında iyileşme olduğu gözlemlenmiş ve mahremiyet risklerinde değişiklik olmamıştır. İkinci iterasyonda ilk iterasyonda olduğu gibi veri faydasındaki iyileşme devam etmiş ve en yüksek risk tahmini yine değişmemiştir.

Sonuç olarak, eşdeğer sınıf ayrımının veri faydasını arttırırken mahremiyetten ödün vermediği, aykırı eşdeğer sınıfın geri kazanımının veri faydasına olumlu etkisi olduğu, aykırı eşdeğer sınıfın geri kazanımının mahremiyet riskleri üzerinde olumsuz bir etki yapmadığı görülmüştür. Buna göre, ρ-Kazanım modelinin OEC üzerinde işlem yapabildiği, k-Anonimlik ve l-Çeşitlilik modellerini birlikte kullandığından kimlik ve öznitelik ifşalarına karşı koruma sağladığı ve mahremiyetten ödün vermeden veri faydasını arttırdığı 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.

KAYNAKLAR (REFERENCES)

1. Samarati P., Protecting respondent's privacy in micro data release, IEEE Transaction on Knowledge and Data Engineering, 13 (6), 1010-1027, 2001.
2. Korolova A., Protecting privacy while mining and sharing user data, Doktora Tezi, Stanford Üniversitesi, Bilgisayar Mühendisliği Bölümü, Amerika, 2012.
3. Abul O., Gokce H., Knowledge Hiding from Tree and Graph Databases, Data & Knowledge Engineering, 72, 148-171, 2012.
4. Verykios S.V., Bertino E., Fovino N.I., Provenza P.L., Saygin Y., Theodoridis Y., State-of-the-art in Privacy Preserving Data Mining, ACM SIGMOD Record, 33 (1), 50-57, 2004.
5. Çelik C., Bilge, H.Ş., Feature Selection with Weighted Conditional Mutual Information, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (4), 585-596, 2015.
6. Mahmood S., New Privacy Threats for Facebook and Twitter Users, Seventh International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Victoria, Kanada, 164-169, 2012.
7. Barbaro M., Zeller M., A Face Is Exposed for AOL Searcher No. 4417749, http://www.nytimes.com/2006/08/09/technology/09aol.html. Yayın tarihi Ağustos 9, 2016. Erişim tarihi Mart 14, 2017.
8. Wagas, Someone Hacked and Leaked Entire Turkish Citizenship Database Online, https://www.hackread.com/turkish-citizenship-database-hacked-leaked. Yayınlanma tarihi Nisan, 2016. Erişim tarihi Mart 14, 2017.
9. Fung B. C. M., Wang K., Chen R., Yu P. S., Privacy-preserving data publishing: A survey of recent developments, ACM Computing Surveys (CSUR), 42 (4), 523-553, 2010.
10. Lin W., Yang D., Wang J., Privacy preserving data anonymization of spontaneous ADE reporting system dataset, BMC Medical Informatics and Decision Making, 10 (1), 21-35, 2016.
11. Kohlmayer F., Prasser F., Kuhn KA.,The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss, Journal of Biomedical Journal of Biomedical Informatics, 58, 37-48, 2015.
12. Xu X., Ma T., Tang M., TianA W., Survey of privacy preserving data publishing using generalization and suppression, International Journal on Applied Mathematics & Information Sciences, 8 (3), 1103-1116, 2014.
13. Brüggemann R., Patil PG., Partial Order and Hasse Diagrams, Ranking and Prioritization for Multi-Indicator Systems, Springer, New York, 13-23, 2011.
14. Ferrer-Domingo J., Mateo-Sanz J.M., A Comparative Study of Microaggregation Methods, Qüestiió Journal, 22 (3), 511–526, 1998.
15. Sweeney L., k-Anonymity: A model for protecting privacy, International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 10 (5), 557–570, 2002.
16. Machanavajjhala A., Kifer D., Gehrke J., Venkitasubramaniam M., L-Diversity: Privacy beyond k- anonymity, International Conference on Data Enginnering (ICDE), Atlanta, ABD, 24, 2006.
17. Li N., Li T., Venkatasubramanian S., t-Closeness: Privacy beyond k-anonymity and l-diversity, International Conference on Data Engineering (ICDE), İstanbul, Türkiye, 106–115, 2007.
18. Lengdong W., Hua H., Osmar RZ., Utility Enhancement for Privacy Preserving Health Data Publishing, 9th International Conference on Advanced Data Mining and Applications, Springer Berlin Heidelberg, Berlin, 311-322, 2013.
19. 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, Paris, Fransa, 517-526,2009.

20. Hua M., Pei J., A Survey of Utility-based Privacy-Preserving Data Transformation Methods, Privacy-Preserving Data Mining: Models and Algorithms, Editör: Aggarwal C., Yu PS., Springer, Boston, 207-237,2008.

21. LeFevre K., DeWitt DJ., Ramakrishnan R., Mondrian Multidimensional k-Anonymity, International Conference on Data Enginnering (ICDE), Atlanta, ABD, 25-36, 2006.

22. Bayardo R. J., Agrawal R., Data privacy through optimal k-anonymization, International Conference on Data Enginnering (ICDE), Tokyo, Japonya, 217–228, 2005.

23. Sweeney L., Datafly: A system for providing anonymity in medical data, Eleventh International Conference on Database Securty XI: Status and Prospects, Londra, İngiltere, 356–381, 1998.

24. Sweeney L., Achieving k-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 571–588, 2002.

25. Wang K., Fung BCM., Anonymizing sequential releases, 12th ACM SIGKDD, Philadelphia, ABD, 414–423, 2006.

26. Lyengar VS., Transforming data to satisfy privacy constraints, 8th ACM SIGKDD, Edmonton, Kanada, 279–288, 2002.

27. El Emam K., Guide to the De-Identification of Personal Health Information, CRC Press, Florida, ABD, 2013.

28. Dankar F.K., El Emam K., A method for evaluating marketer re-identification risk", EDBT/ICDT Workshops, Lausanne, İsviçre, 1-10, 2010.

29. Prasser F., Bild R., Eicher J., Spengler H., Kohlmayer F., Kuhn KA., Lightning: Utility-Driven Anonymization of High-Dimensional Data, Transactions on Data Privacy, 9 (2), 161–185, 2016.

30. Prasser F., Kohlmayer F., Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool, Medical Data Privacy Handbook, Editörler: Divanis AG., Loukides G., Springer, İsviçre, 111-145, 2015

31. Kohlmayer F., Prasser F., Kuhn KA., The Cost of Quality: Implementing Generalization and Suppression for Anonymizing Biomedical Data with Minimal Information Loss, Journal of Biomedical Informatics, 58, 37-48, 2015.

32. The UCI Machine Learning Repository, Adult Data Set, https://archive.ics.uci.edu/ml/datasets/Adult. Yayınlanma tarihi 1994. Erişim tarihi Mart 14, 2017.

33. Akben S. B., Alkan A., Density-Based Feature Extraction To Improve The Classification Performance In The Datasets Having Low Correlation Between Attributes, Journal of the Faculty of Engineering and Architecture of Gazi University, 30 (4), 597-603, 2015.