**TÜRK HAVA KURUMU ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# MÜLKİYET BİLGİLERİNİN PAYLAŞILMASINDA KİŞİSEL VERİLERİN MAHREMİYETİNİN KORUNMASI

**YÜKSEK LİSANS TEZİ**

**BARIŞ ANKAY**

**Elektrik ve Bilgisayar Anabilim Dalı**

**Elektrik ve Bilgisayar Mühendisliği Programı**

**EYLÜL 2019**

**TÜRK HAVA KURUMU ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

# MÜLKİYET BİLGİLERİNİN PAYLAŞILMASINDA KİŞİSEL VERİLERİN MAHREMİYETİNİN KORUNMASI

**YÜKSEK LİSANS TEZİ**

**BARIŞ ANKAY**

**1303617005**

**Elektrik ve Bilgisayar Anabilim Dalı**

**Elektrik ve Bilgisayar Mühendisliği Programı**

**Tez danışmanı: Yrd. Doç. Dr. Meltem YILDIRIM İMAMOĞLU**

Türk Hava Kurumu Üniversitesi Fen Bilimleri Enstitüsü'nün 1303617005 numaralı Yüksek Lisans öğrencisi, Barış ANKAY ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı Mülkiyet Bilgilerinin Paylaşılmasında Kişisel Verilerin Mahremiyetinin Korunması, aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

**Tez Danışmanı :** Yrd. Doç. Dr. Meltem Yıldırım İmamoğlu
Türk Hava Kurumu Üniversitesi

**Jüri Üyeleri :** Prof. Dr. Ahmet COŞAR
Türk Hava Kurumu Üniversitesi

Prof. Dr. Halit OĞUZTÜZÜN
Orta Doğu Teknik Üniversitesi

Yrd. Doç. Dr. Meltem Yıldırım İmamoğlu
Türk Hava Kurumu Üniversitesi

**Tez Savunma Tarihi : 16.09.2019**

**TÜRK HAVA KURUMU ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜ'NE**

Yüksek Lisans Tezi olarak sunduğum, Mülkiyet Bilgilerinin Paylaşılmasında Kişisel Verilerin Mahremiyetinin Korunması adlı çalışmamın, tarafımdan akademik etik ve kurallara aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım kaynakların kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

**16.09.2019**
**Barış ANKAY**

## TEŞEKKÜR

Yüksek Lisans tez çalışma sürecinde beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan desteğini ve yardımını hiçbir zaman esirgemeyen tez danışmanım değerli Yrd. Doç. Dr. Meltem Yıldırım İmamoğlu'na teşekkürlerimi sunarım.

Hayatım boyunca her zaman yanımda olan, maddi ve manevi desteklerini hiçbir zaman esirgemeyen aileme teşekkürlerimi, sevgi ve saygılarımı sunarım.

Eylül, 2019 Barış ANKAY

# İÇİNDEKİLER

# TABLO LİSTESİ

# ŞEKİL LİSTESİ

## KISALTMALAR

**CM** : Classification metric
**DF** : De-identified File
**EC** : EquivalenceClass
**EMD** : Earth Mover Distance)
**EMD** : EarthMoverDistance
**Pid** : Person identifier
**PPDM** : Privacy-preserving data mining
**PPDP** : Privacy Preserving Data Publishing
**Pt** : Person Table
**QID** : Quasi-IDentifier
**S** : Sensitive

# ÖZET

## MÜLKİYET BİLGİLERİNİN PAYLAŞILMASINDA KİŞİSEL VERİLERİN MAHREMİYETİNİN KORUNMASI

ANKAY, Barış
Yüksek Lisans, Elektrik ve Bilgisayar Anabilim Dalı
Tez Danışmanı: Yrd. Doç. Dr. Meltem YILDIRIM İMAMOĞLU
Eylül-2019, 100 sayfa

Kamu kurumları daha hızlı, güvenli ve kaliteli hizmet verebilmek için birbirleri ile günümüzde veri paylaşmaktadırlar. Bu tez çalışmasında stratejik olarak da önem arz eden mülkiyet verilerinin paylaşılmasında anonimliğin sağlanmasına yönelik uygulanabilecek yöntemler gösterilmiştir. Uygulanan bu yöntemlerle mahremiyetin korunmasına ilişkin risk analizi yapılmıştır. Bu çalışma ile ilgili kişi ve kurumlarda farkındalık oluşmasına katkıda bulunmak amaçlanmıştır.

Mülkiyet verilerinin kurum ve kuruluşlar arasındaki paylaşımı birçok alanda yeni çözümler ve fırsatlar sunmaktadır. Mülkiyet sisteminde tutulan veriler aynı zamanda kişisel bilgileri de içermektedir. Bu verilerin doğrudan düzenlemeye (anonimliği sağlayıcı uygulamalar) tabi tutulmadan paylaşılması kişisel mahremiyetin ifşasına sebep olmaktadır. Kişisel verilerin korunmasına yönelik adı, soyadı ve T.C. kimlik numarası gibi alanların bu verilerden çıkarılarak sunulması bireysel mahremiyetin ihlal edilmesini engelleyememektedir. Yapılan araştırmada mahremiyetin korunmasına yönelik literatürde birçok mahremiyet korumalı veri paylaşım yaklaşımı geliştirildiği tespit edilmiştir. Bu yaklaşımlar veri faydasını maksimum düzeyde tutmak için kullanılmaktadır. Özellikle son dönemde Mahremiyet Korumalı Veri Yayıncılığı (Privacy-Preserving Data Publishing (PPDP)) ve Mahremiyet Korumalı Veri Madenciliği (Privacy-Preserving Data Mining (PPDM)) yaklaşımları veri mahremiyetinin korunması adına sıkça ve kapsamlı olarak çalışılmıştır. Bu çalışmada mülkiyet verilerinin belirlenen alanlarında k-anonimlik, $\ell$-

çeşitliliği ve t-yakınlığı uygulanmıştır. Bu kapsamda anonimleştirilen veriler üzerinde Savcı Riski, Gazeteci Riski ve Pazarlamacı Riski hesaplanmıştır.

Bu tez çalışmasının amacı; uygulanan anonimlik modelleri ile yapılan risk ölçümleri, veri kaybını minimum tutarak veri faydasının maksimum olduğu en ideal sonucu verdiği anonimleştirme seviyelerinin gözlemlenerek. Yapılan risk analizleri sonucunda hesaplanan risk değerlerinin sıfıra yaklaştığı gözlemlenmiştir.

Bu bağlamda uygulanan anonimlik modellerinin kişisel mahremiyetin korunmasına yönelik başarılı olduğunu söyleyebiliriz.

**Anahtar Kelimeler:** Mülkiyet verisi, mahremiyet koruma, gizlilik, k-anonimlik, risk ölçümleri

## ABSTRACT

### PRIVACY PRESERVING PERSONAL INFORMATION IN THE SHARİNG OF LAND REGISTER DATA

ANKAY, Barış

Master, Department of Elektrical and Computer

Thessis Supervisor: Yrd. Doç. Dr. Meltem YILDIRIM İMAMOĞLU

Jane-2019, 100 page

The, Public institutions share data with each other in order to provide faster, more secure and quality services. In this thesis, the methods that can be applied to provide anonymity in sharing property data which is also strategically important are shown. With these methods, risk analysis regarding privacy protection was performed. The aim of this study is to contribute to raising awareness in related people and institutions.

The sharing of property data between institutions and organizations offers new solutions and opportunities in many areas. Data held in the property system also includes personal information. Sharing this data without being directly regulated (anonymity practices) causes disclosure of personal privacy. Name, surname and T.C. submission of fields such as identification number from this data cannot prevent the violation of individual privacy. In the research, it has been determined that many privacy protected data sharing approaches have been developed in the literature regarding privacy protection. These approaches are used to maximize data benefit. Especially in the recent period, Privacy-Preserving Data Publishing (PPDP) and Privacy-Preserving Data Mining (PPDM) approaches have been frequently and comprehensively studied for the protection of data privacy. In this study, k-anonymity, l-diversity and t-affinity were applied in the identified areas of property data. In this context, Prosecutor Risk, Journalist Risk and Marketing Risk were calculated on the anonymized data.

The aim of this thesis is; risk measurements made with the anonymity models applied, keeping the data loss to a minimum, by observing the anonymization levels which yield the most ideal result with maximum data benefit. As a result of the risk analyzes, it was observed that the calculated risk values approached zero.

We can say that the anonymity models applied in this context are successful for the protection of personal privacy.

**Keywords:** Property data, privacy protection, confidentiality, k-anonymity, risk measurements

# BİRİNCİ BÖLÜM

## GİRİŞ

Kamu ve özel sektörün günlük işleyişleri esnasında kamu kaynakları kullanılarak ürettiği sayısal bilgilerin toplanması ve bu bilgilerden çeşitli çıkarımlarda bulunulması önemli fırsatlar sunmaktadır. Bilgilerin üretimi, saklanması ve diğer bilgi sistemlerine girdi olarak kullanılması kamu ve özel sektörünün daha hızlı, etkin ve kaliteli hizmet verebilmesi için temel bileşen halini almıştır. Sağlık, nüfus, finans, eğitim ve mülkiyet konularında hizmet veren elektronik uygulamaların kullanımı hızla yaygınlaşmaktadır. Bu uygulamalar aracılığıyla toplanan veriler içerisinde kişisel verilerde yer almaktadır. Bu veriler içerisinde kişilere ait sağlık, maaş, adres, nüfus, tapu, sabıka kayıtları gibi hassas ve özel veriler bulunmaktadır. Uygulamalar aracılığıyla toplanan verilerden çıkarımlar ve analizler yapılarak fayda sağlanmasına araştırmacılar veya kurumlar tarafından ihtiyaç duyulmaktadır. Veriler paylaşılmadan önce, veri sahiplerinin mahremiyetini koruyan tedbirlerin alınması kişisel mahremiyet açısından önemlidir. Mahremiyet koruyucu yöntemlerin uygulanmasında paylaşılacak verilerden elde edilecek değerin dikkate alınması gerekmektedir.

Kamu kurumları ve özel sektöre ait işletmelerin kendi iş süreçlerine bağlı olarak kullandıkları bilgileri elektronik ortamda tutmaya başlamalarıyla birlikte bu bilgiler analiz edilebilir ve araştırmalar yapılabilir hale gelmektedir. Bunlardan elde edilen çıkarımlar kamu ve özel sektör tarafından ileriye dönük politikaların hazırlanması esnasında rehber olarak kullanılmaktadır. Örneklemek gerekirse; vergi kayıplarının önüne geçilmesi, yatırım yapılacak sektörlerin belirlenmesi, yatırım için kurulacak fabrikalara arazilerin bulunması ve kişilerin harcama eğilimlerinin belirlenmesi gibi konularda ilgili veriler kullanılmaktadır. Fakat bu verilerin kullanımı sırasında kişisel veri mahremiyetinin ön plana çıktığı görülmektedir.

Mahremiyetin korunması konusunda yeterli düzeyde önlem alınmaması mahremiyet odaklı birçok saldırıya davet çıkarmaktadır. Paylaşılan verilerde veri

sahiplerinin kimliklerinin ortaya çıkartılması ve hassas bilgilerinin ifşa edilmesine yönelik birçok saldırı düzenlenmekte ve bu bilgiler kötü niyetli kişi/kurum eline geçmektedir.

Mahremiyete yönelik büyük saldırılardan birisi 2016 yılında Yahoo yaptığı açıklamada tarihteki en büyük veri sızıntısını yaşadıklarını açıklamıştır. 2013 yılında 500 milyon olarak tahmin ettikleri sızıntının aynı yılın sonunda 1 milyara ulaştığını, günümüzde ise sızıntıdan 3 milyar e-posta kullanıcısının etkilendiğini açıklamıştır. Yahoo önlem olarak tüm kullanıcılarının e-posta şifrelerini değiştirmelerini istemiştir [3].

2016 yılında 7 milyondan fazla müşterisi olan Tesco Bank yaptığı açıklamada 40 bin müşterisinin hesap bilgilerinin sızdırıldığını ve bunlardan 20 bin müşterinin hesaplarından usulsüz para aktarıldığını tespit ettiklerini açıklamıştır. Önlem olarak Tesco Bank yönetim kurulu yaptığı açıklama ile bir süreliğine ilgili hesapları işleme kapatma kararı almış ve müşterilerin mali kayıplarını karşılayacaklarını bildirmiştir [4].

2017 yılında 23 ülkede faaliyet gösteren uluslararası bir finans kuruluşu olan Equifax'a karşı gerçekleştirilmiştir. Equifax'ın web sayfalarındaki zafiyeti kullanan saldırgan bir yazılım aracılığı ile Amerika Birleşik Devletleri'ndeki (ABD) 143 milyon tüketicinin sosyal güvenlik numarası, doğum tarihi, adres bilgileri vb. sızdırılmıştır. Ayrıca İngiltere'deki 400 bin kullanıcının da bu saldırıdan etkilendiği firma tarafından bildirilmiştir [2].

Dünyada olduğu gibi ülkemizde de çok sayıda mahremiyet ihlal ve saldırıları yaşanmaktadır. Kişilerin kimliklerini doğrudan tanımlamada kullanılan kimlik numaralarının ifşa edilmesi ile ilgili ülkemizde birçok olay yaşanmıştır. Bunlardan en önemlisi 2016 yılında ülkemizde 49.611.709 kişinin TC kimlik numarası, ad, soyad, anne adı, baba adı, cinsiyet, doğum yeri, doğum tarihi, Nüfus kayıt yeri ve açık adres bilgilerini içeren kişisel bilgilerinin sızdırılması olmuştur. Sızdırılan bu verilerin boyutu 6,6 GB olarak açıklanmıştır [5].

2008 yılında gerçekleşen bir başka olay ise Resmî Gazetede 8,5 milyon kişinin TC Kimlik ve sigorta numarasını içeren KEY ödemeleri listesinin yayınlanmasıdır [6].

25 Aralık 2010'da yapılan İş Yeri Hekimliği ve İş Güvenliği Uzmanlığı Sınavına katılan 2 bin 119 adayın kimlik numarası, baba adı ve doğum tarihi gibi kişisel bilgilerinin ilgili kamu sitesinin web sayfasında yayınlanması ülkemizdeki mahremiyet ihlallerine verilebilecek bir diğer örnektir [8].

Yukarıda yer alan örnek olaylardan da anlaşılacağı üzere mahremiyete yönelik saldırı ve ihlaller artarak devam etmektedir. Mahremiyet ihlallerinin ve saldırılarının

en aza indirgenmesi amacıyla yeni mahremiyet koruyucu modellerin geliştirilmesi ve mahremiyet bilincinin arttırılmasına ihtiyaç duyulmaktadır. Avrupa Birliği vatandaşlarını etkileyen herhangi bir kişisel veri ihlalini 72 saat içinde raporlamayı gerektiren mevzuatı 25 Mayıs 2018 yılında yürürlüğe koymuştur [65]. Bu düzenlemeye uymayan şirketlere para cezası uygulanacaktır.

Yukarıdaki ihlaller ve ihlallere karşı alınan tedbirlerin önemi her geçen gün artmakta ve daha fazla ülke/kurumlar tarafından ön plana çıkarılmaktadır. Bir ülkenin stratejik öneme sahip ve korunması gereken iki önemli verisi vardır. Bunlar sırası ile Nüfus ve Mülkiyet verisidir. Kamu ve özel sektör hemen hemen tüm çalışma alanlarında öncelikle nüfus ve mülkiyet bilgisine ihtiyaç duymaktadır. Örneğin Tarım Bakanlığı çiftçilere ait toprakları belirlemek için mülkiyet ve nüfus verisine, Gelir idaresi vergi kaçaklarının tespiti için yine mülkiyet ve nüfus bilgisine ihtiyaç duymaktadır.

Mülkiyet verilerinin paylaşımında yaşanabilecek mahremiyet sorunları ile bu sorunların çözümü ve mahremiyetin korunmasına yönelik yöntemler bu tez çalışmasının araştırma konusudur. Ayrıca yapılan araştırmalar ve değerlendirmelere göre mülkiyet verilerinin paylaşılmasında kişisel verilerin mahremiyetinin korunmasını sağlayan bir çözüm önerisi getirilecektir.

Tez genel olarak beş kısımdan oluşmaktadır. Birinci kısımda Mülkiyet bilgilerinin paylaşılmasında kişisel verilerin mahremiyetinin korunması konusunda problemin tanımı yapılmaktadır. İkinci kısımda kişisel mahremiyetin korunması ile ilgili literatür taraması yapılmış, bu çalışmaların genel olarak amacı açıklanmıştır. Üçüncü kısımda çalışmamızda kullandığımız materyaller ve kullanılan yöntemler anlatılmakta, deneysel çalışmalarda verilerimizin kullanımı ile ilgili bilgiler verilmektedir. Dördüncü kısımda çalışmamızda yaptığımız deneysel çalışmalar ve sonuçları anlatılmakta, deneysel çalışmalardan çıkarımlar yapılmaktadır. Beşinci kısımda çalışmamız ile ulaşılan sonuçlar, karşılaşılan zorluklar ve öneriler anlatılmaktadır.

# İKİNCİ BÖLÜM

# LİTERATÜR ÖZETİ

Kamu ve özel sektör verilerinin birbirleri ile karşılıklı etkileşimde ve paylaşımda olduğu düşünüldüğünde kişisel veri mahremiyeti ön plana çıkmaktadır. Bu mahremiyetin korunabilmesi ise anonimlik kavramı ile bağlantılı olmaktadır. Yakın zamanda yapılan araştırmalar iki yaklaşımı ön planda tutmaktadır. Bunlar, Mahremiyet Korumalı Veri Madenciliği (Privacy-Preserving Data Mining (PPDM)) ve Mahremiyet Korumalı Veri Yayını'dır (Privacy-Preserving Data Publishing (PPDP)) [39].

Kişisel veri mahremiyeti ile ilgili araştırmalar incelendiğinde genel olarak iki başlık altında toplanmakta oldukları görülmektedir. Bunlardan ilki veri yayınlama ve ikincisi veri modeli üzerine yapılan araştırmalardır [39]. Bu çalışmada kişisel veri mahremiyeti konusunda yapılan araştırmaların oldukça fazla olması sebebiyle bu iki başlık altında yapılan araştırmalara yer verilecektir.

## 2.1 Genel Tanımlar

Belirli bir amaca yönelik olarak toplanan verilerden daha fazla fayda sağlayabilmek için verinin veya veriden elde edilen bazı istatistiklerin paylaşılması gerekmektedir. Paylaşılan bu veri bünyesinde farklı öznitelikleri barındırmaktadır. Bu özniteliklerden yararlanılarak verinin kimliksizleştirilmesi veya anonimleştirilmesi gerekmektedir. Bu veriler üzerinde düzenlemeler yapılarak anonimlik sağlanmaktadır. Bu anonimleştirmenin yapılabilmesi için verinin özniteliklerinin 4 ayrı tip de sınıflandırılması gerekmektedir [9]. Bu sınıflandırmalar;

1. Açık (Doğrudan) tanımlayıcılar (Explicit Identifier): Adı, soyadı, telefon numarası, sosyal güvenlik numarası ve ehliyet gibi tekil numaralar kayıt sahibini doğrudan tanımlayan öznitelikler kümesi [9].

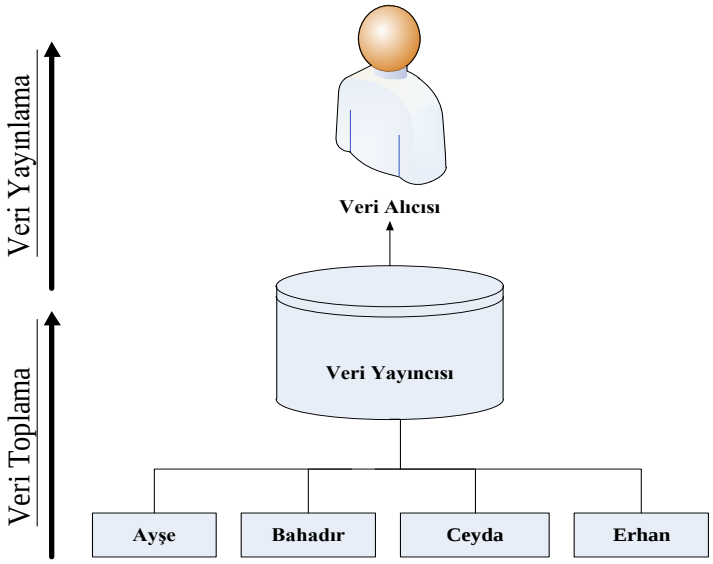
2. Yarı tanımlayıcılar (Quasi Identifier): Yaş, adres, posta kodu, cinsiyet gibi, kayıt sahibini tanımlamak için tek başına yeterli olmamasına rağmen, veri kümesinde bulunma durumuna bağlı olarak veya başka verilerle bir araya getirilerek bir kişiyi tanımlayabilecek potansiyele sahip olan öznitelikler kümesidir. Yarı tanımlayıcı öznitelikler kısaca QID olarak belirtilirken bu özniteliklerin değer kümeleri qid olarak gösterilmektedir [9].
3. Hassas öznitelikler (Sensitive attributes): Hastalık bilgisi, gelir bilgisi, engel durumu gibi, veri sahiplerine özel hassas bilgiler kümesi. S ile gösterilmekte ve bir kümede birden çok olabilmektedir [9].
4. Hassas olmayan öznitelikler (Non-Sensitive attributes): Diğer üç gruba girmeyen özniteliklerdir. Bu özniteliklerin ifşa edilmesi herhangi bir tehlike teşkil etmemektedir. Bireylerle ilgili ayırt edici bilgiler içermeyen verilerden oluşmaktadır [9].

## 2.2 Mahremiyet Modelleri

Bu çalışmalar ile genel olarak verinin toplanması ve yayınlama süreçlerindeki mahremiyet ihlalleri incelenmektedir. Bu ihlalleri tipik bir senaryo ile ifade etmek gerekirse; Veri toplama ve yayınlama için tipik bir senaryo Şekil 2.1'de açıklanmaktadır [39]. Veri toplama aşamasında, veri yayıncısı kayıt sahiplerinin verisini toplamaktadır. Veri yayınlama aşamasında, toplanan verileri bir veri madencisine veya veri alıcısına yayınlamaktadır. Sürecin bu kısmında, veri madenciliği önem taşımaktadır. Bu aşamada yapılan iş sadece model oluşturma ile sınırlı değildir. Örneğin, bir hastane, hastalardan veri toplamakta ve hasta kayıtlarını farklı bir tıbbi merkez için yayınlamaktadır. Bu örnekte, hastane veri yayıncısı, hastalar kayıt sahipleri, tıbbi merkez ise veri alıcısı olmaktadır. Tıbbi merkezde yapılan veri madenciliği, diyabetli erkek sayısının basit bir sayımından karmaşık bir küme analizi yapılmasına kadar her şeyi kapsayabilmektedir.

İki veri yayıncısı modeli bulunmaktadır. Güvenilmeyen modelde, veri yayıncısı güvenilir değildir ve kayıt sahiplerinden hassas bilgileri belirleme girişiminde bulunmaktadır. Bu durumda çözüm için çeşitli şifreleme teknikleri; İsimsiz iletişimler ve istatistiksel yöntemler kullanılabilmektedir. Veri yayıncısından veri sahiplerinin kimliğini göstermeden anonim olarak kayıtları toplaması önerilmektedir. Güvenilir

modelde, veri yayıncısı güvenilirdir ve kayıt sahiplerinin kişisel bilgilerini kendi veri alıcısına sunmaya hazırdır.




**Şekil 2.1:** Veri Toplama ve Veri Yayınlama

Uygulamada, her veri yayınlama senaryosunun, veri yayıncısının, veri alıcılarının ve veri yayımlama amacının kendi varsayımları ve gereksinimleri bulunmaktadır. İzleyen kısımda pratik veri yayınındaki birkaç varsayım ve özellik verilmektedir:

Uzman olmayan veri yayıncısı: Veri yayıncısının, veri alıcısı adına veri madenciliği yapma bilgisine sahip olması gerekmemektedir. Herhangi bir veri madenciliği faaliyeti, veri alıcısı tarafından verinin alınması ile başlar. Bu nedenle veri yayıncısı sonraki aşamada bilgi sahibi değildir. Örneğin, Kaliforniya'daki hastaneler Web ‘de hasta kayıtlarını yayınlamaktadır. Hastaneler, alıcıların kim olduğunu ve alıcıların veriyi nasıl kullanacaklarını bilmemektedir. Hastane veri madenciliğinin sonucunu yayınlaması gerektiği için veya genel tıbbi araştırmayı desteklediği için hasta kayıtlarını yayınlamaktadır. Bu nedenle, veri yayıncısının böyle bir senaryoda yayınlanmak üzere verileri anonimleştirmekten fazlasını yapmasını beklemek makul olmamaktadır.

Diğer senaryolarda, veri yayıncısı veri madenciliği sonuçlarıyla ilgilenmektedir. Ancak analiz yapmak için kurum içi uzmanlığından yoksundur ve bu nedenle veri madenciliği faaliyetlerini veri alıcısına bırakmaktadır. Bu durumda veri madenciliği, alıcı tarafından gerçekleştirilmektedir. Veri yayıncısı veriyi korumak için belirli desen türlerini koruyan özelleştirilmiş bir veri kümesi kullanabilmektedir.

PPDP'de veri alıcısının aynı zamanda bir saldırgan olabileceği varsayımı da bulunmaktadır [39]. Örneğin, veri alıcısı, güvenilir bir ilaç araştırması şirketi olabilir; bununla birlikte, şirketteki tüm personelin de güvenilir olduğunu garanti edilememektedir. Böylesi durumlar, PPDP'ye ilişkin sorunların çözümü amacıyla şifreleme yaklaşımdan daha farklı çözümler üretilmesini gerekli kılmıştır. Yalnızca yetkili ve güvenilir alıcılara açık metinlere erişmek için özel anahtar verilmektedir. PPDP 'deki en büyük zorluk, gizlilik ve anonim bilginin kullanılabilirliğinin eşzamanlı olarak korunabilmesidir. Bu durum veri madenciliğinin sonucu değil de bireylerin bazı bilgilerinin yayınlanması (yani mikro veriler) olgusunu vurgulamaktadır. Bu gereksinim sınıflayıcılar, ilişkilendirme kuralları veya bireylerin grupları hakkındaki istatistikler gibi veri madenciliği sonuçlarını yayınlamaktan daha katı kuralları beraberinde getirmektedir [39].

Örneğin Netflix, sonucu etkilemeyecek bazı bilgileri değiştirerek kullanıcıların film derecelendirme bilgilerini yayınlamaya başlamıştır. Böylece alıcılar yayınlanan bilgiler üzerinde daha esnek araştırma yapabilmektedirler. Netflix'in veri madenciliği sonuçlarını değil de verinin kendisini yayınlaması veri keşfi konusunda daha fazla esnekliğe sahip olunmasını sağlamaktadır. Belli bir deseni içeren işlemleri görselleştirmek, farklı modelleme yöntemleri ve parametreleri denemek gibi serbestilerin ortaya çıkabilmesi verilerden elde edilen sonuçların değil bazı veri parçalarının yayınlanabilmesi ile mümkün olmaktadır. Böylece uzman olmayan bir veri alıcısı veriyi işleyebilir [66]. Örneğin, Netflix ilgilenen tarafların verileri nasıl analiz edeceğini önceden bilmiyor. Bazı temel "bilgi parçaları" yayınlanan verilerde saklanmaktadır ve bu bilgi parçaları veri setine ait spesifikasyonları değiştirmemektedir.

Kayıt seviyesindeki doğruluk: Bazı veri yayımlama senaryolarında, yayınlanan her kaydın gerçek hayatta mevcut bir kişiye karşılık gelmesi önemlidir. Hasta kayıtları örneğini düşünün. İlaç araştırmacısı (veri alıcısı), test edilen ilacın daha önce bilinmeyen yan etkilerini keşfetmek için gerçek hasta kayıtlarını incelemek zorunda kalabilir. Yayınlanmış bir kayıt, gerçek hayatta mevcut bir hastaya karşılık gelmezse, veri madenciliği sonuçlarından beklenen fayda sağlanamamaktadır. Rastgele seçilmiş sentetik veriler bu gereksinimi karşılamamaktadır. Şifrelenmiş bir kayıt gerçek hayatta bir hastaya karşılık gelmesine rağmen, şifreleme, temsil edilen hasta üzerinde hareket etmeyi ve bir sonuca varmayı kısıtlar.

### 2.2.1 K-Anonimliği (K-Anonymity)

Mahremiyetin korunması söz konusu olduğundan en yaygın kullanılan yöntem verilerin anonimleştirilmesidir. Verilerin anonimleştirilmesi, saldırganın veri kümesinde yer alan bireye ait hassas veriye ulaşılmasının engellenmesidir.

Kişisel verilerin anonim hale getirilmiş olması için verilerin, veri sorumlusu veya alıcı grupları tarafından geri döndürülmesi veya başka verilerle eşleştirilmesi gibi kayıt ortamı ve ilgili faaliyet alanı açısından uygun tekniklerin kullanılması yoluyla kimliği belirli veya belirlenebilir bir gerçek kişiyle ilişkilendirilemez hale getirilmesi gerekir [18].

K-anonimlik yönteminde, sırası ile genelleştirme (generalization) ve baskılama (suppression) tekniklerinin, yarı tanımlayıcı ve hassas özniteliklere uygulanması ile bireysel kayıtların tekilleştirilmesi engellenerek (Kayıtlar genelleştirilerek) gizliliği korumaya çalışılmaktadır [40]. Genelleştirme tekniğinde ise belirginliği azaltmak için, özniteliğin değeri belirli aralıklarla genişletilir. Baskılama tekniğinde ise özniteliğin değeri tamamen kaldırılır. Bu durum genelleştirmenin son seviyesidir ve verinin değeri hakkında bilgi vermemektedir.

K-anonimliğinde veri tablosunda bulunan her $qid$ için aynı $qid$ ile en az $k-1$ tane daha kayıt olması gerekmektedir. Böyle bir $qid$ altında gruplandırılan hassas değerlerin bir $q*-blokta$ bulunduğu söylenmektedir. Böyle bir anonimleştirme tekniğinin sonucu Tablo 2.1'de görülmektedir. K-anonimliliğinin etkisi, bir saldırganın bir kişiyi maksimum $1/k$ olasılıkla rekora bağlaması olmaktadır [52].

Diğer benzer $qid$'leri elde etmek için önerilen yöntem genelleştirmedir [56]. Öznitelikler için üç değer türü vardır. Güncel alan (W), sayısal (R) ve kategorik (T). Bunların haricinde de türler bulunmakta ve ses, görüntü vb. gibi alanlarda kullanılmaktadır. Sayısal değerler için, genellemeler sayıları o sayıyı içeren aralıklara dönüştürerek gerçekleştirilir. Örneğin, 36 yaş (30,40) veya 3* olarak genelleştirilebilmektedir. Kategorik değerler için bir taksonomi ağacı kullanılabilir. Böyle bir ağaç Şekil.2.2'de görülmektedir. Şekil incelendiğinde, dansçının, herhangi birine veya sanatçıya genelleştirilebileceği gözlemlenebilir. Örneğimizde, anonimliği elde etmek için yalnızca iş ve yaş genelleştirilmesi gerçekleştirilmiştir. Bu durum, bir saldırganın en fazla $1/3$ olasılığa sahip bir rekora bağlanabileceği anlamına gelmektedir.

**Tablo 2.1:** Hasta Tablosu

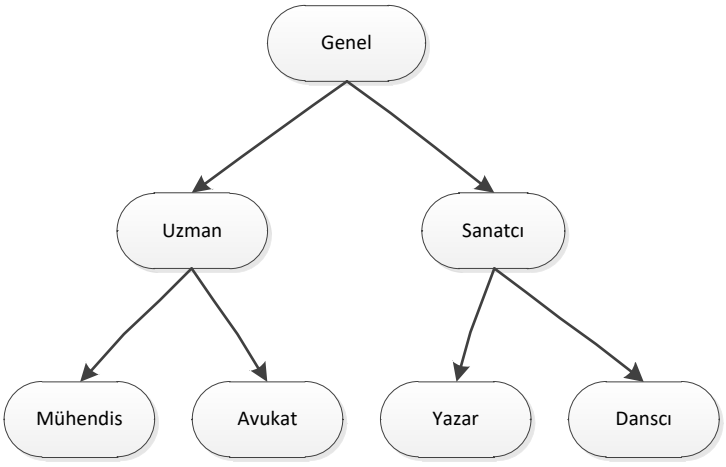
| İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|
| Mühendis | Erkek | 35 | Hepatit |
| Mühendis | Erkek | 38 | Hepatit |
| Avukat | Erkek | 38 | AİDS |
| Yazar | Kadın | 35 | Grip |
| Yazar | Kadın | 35 | AİDS |
| Sanatçı | Kadın | 35 | AİDS |
| Sanatçı | Kadın | 36 | AİDS |

**Tablo 2.2:** Dış Bağlantı Tablosu

| Name | İş | Cinsiyet | Yaş |
|---|---|---|---|
| Ayşe | Dansçı | Kadın | 36 |
| … | … | … | … |

**Tablo 2.3:** Hastalar İçin Anonim Tablo

| İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | AİDS |
| Sanatçı | Kadın | [35-40) | Grip |
| Sanatçı | Kadın | [35-40) | AİDS |
| Sanatçı | Kadın | [35-40) | AİDS |
| Sanatçı | Kadın | [35-40) | AİDS |




**Şekil 2.2:** İş İçin Taksonomi Ağacı

### 2.2.2 (X-Y)- Anonimliği (X-Y Anonymity)

Önceki örneğimizde, tablo bireysel olarak en fazla bir rekor içermekteyken, bu örnekte bir kişinin birden fazla kaydı olabileceği bir tablo verilmektedir. Her bireyin tabloda üç kayıta sahip olması halinde, her $q*$ bloğunda $k/3$ ayrı kişi olmaktadır. 3-anonimlik durumunda, bu bir $q*$-bloğunun yalnızca bir ayrı bireyi içerdiği anlamına gelmektedir.

Bu sorunun üstesinden gelmek için Wang ve Fung tarafından $(X-Y)$ kimlik doğrulaması önerilmektedir. $X$ ve $Y$, tablodaki tüm öznitelikler üzerinde iki ayrık öznitelik kümesi olsun. Her sette bulunan özellikler veri yayıncısı tarafından belirlenmektedir. Model, $X$ üzerindeki her bir değerin $Y$ üzerinde en az $k$ farklı değerle bağlantılı olması (birlikte gerçekleşmesi) gerektiğini ifade etmektedir. Önceki örneği her bir birey için bir kimlik ile genişleten Tablo 2.4'e görülmektedir. $X = \{iş, cinsiyet, yaş\}$ ($QID$ ile aynı) ve $Y = \{pid\}$ olarak düzenlenmiştir. $(X,Y)$-anonimliki sağlanırsa, $X$, $k$ farklı pid'lerle ilişkilendirilir. Bu, her $q*$ -bloğunun daha sonra, yukarıda belirtilen sorunu aşarak, $k$ farklı bireyler içerdiği anlamına gelmektedir [17].

**Tablo 2.4:** Hastalık İçin Anonim Tablo

| id | İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|---|
| 1 | Uzman | Erkek | [35-40) | Hepatit |
| 2 | Uzman | Erkek | [35-40) | Hepatit |
| 1 | Uzman | Erkek | [35-40) | AİDS |
| … | … | … | … | … |

**Tablo 2.5:** Hastalığın Ayırt Edici Olduğu Anonim Tablo

| Pid | İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|---|
| 1 | Uzman | Erkek | [35-40) | Hepatit |
| 2 | Uzman | Erkek | [35-40) | Hepatit |
| 1 | Uzman | Erkek | [35-40) | AİDS |
| … | … | … | … | … |

### 2.2.3 Çoklu İlişkisel Anonimlilik (MultiR- Anonymity)

K-anonimlik üzerine yapılan çalışmaların çoğunun sadece bir tabloyu anonimleştirdiğini düşündüğü için, MultiRelational-anonimlik kavramı Nergiz ve

diğerleri tarafından ortaya atılmıştır. Gerçek hayatta, yayınlanmış bir veri kümesi genellikle tablolar arasındaki ilişkileri içermektedir. Bu nedenle, her bir tabloyu ayrı ayrı anonimleştirmek, birleştirildiğinde özel bilgileri ortaya çıkarabilmektedir. Bunu modelleyebilmek için yazarlar, kümenin, pid (personel tanımlayıcısı) ve bazı hassas değerler içeren PT' ye (personel tablosu) sahip olduğunu düşünmektedirler. Dahası, veri setleri $n$ adet tablo içermektedir. $T1 \ldots Tn$, yabancı bir anahtardır ve $QID$'deki bazı özellikler ile hassas nitelikleri içermektedir. $T$ tüm tabloların birleşimi olarak tanımlanmaktadır:

$$T = PT \bowtie T_1 \bowtie \cdots \bowtie T_n \qquad \text{(Denklem } \mathbf{2.1}\text{)}$$

T'deki her bir kayıt sahibi için MultiR anonimliğini elde edebilmek için, aynı QID'yi paylaşan en az k - 1 sayıda kayıt sahibi olmalıdır. İlk tablolar için Tablo 2.6'e ve katılma sonrası durum için Tablo 2.7'de gösterilmektedir. Birliğe katıldıktan sonra, ilk örnek elde edilmektedir ve anonimleştirildiğinde MultiR anonimliğini elde edilmektedir [18].

**Tablo 2.6:** Çoklu İlişkisel Veri Seti Örnekleri

| Pid | İş | Hastalık |
|---|---|---|
| 1 | Mühendis | Hepatit |
| 2 | Mühendis | Hepatit |
| 3 | Avukat | AİDS |
| 4 | Yazar | Grip |
| 5 | Yazar | AİDS |
| 6 | Dansçı | AİDS |
| 7 | Dansçı | AİDS |

| Pid | Cinsiyet | Yaş |
|---|---|---|
| 1 | Erkek | 35 |
| 2 | Erkek | 38 |
| 3 | Erkek | 38 |
| 4 | Kadın | 35 |
| 5 | Kadın | 35 |
| 6 | Kadın | 35 |
| 7 | Kadın | 36 |

**Tablo 2.7:** Pid Üzerindeki Birleştirilmiş Tablolar Üzerinde Anonimleştirme

| Pid | İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|---|
| 1 | Uzman | Erkek | [35-40) | Hepatit |
| 2 | Uzman | Erkek | [35-40) | Hepatit |
| 3 | Uzman | Erkek | [35-40) | AİDS |
| 4 | Sanatcı | Kadın | [35-40) | Grip |
| 5 | Sanatcı | Kadın | [35-40) | AİDS |
| 6 | Sanatcı | Kadın | [35-40) | AİDS |
| 7 | Sanatcı | Kadın | [35-40) | AİDS |

### 2.2.4 Bayes Teoremi (Bayes Optimal Privacy)

Machanavajjhala ve arkadaşları "Bayes Optimal Gizlilik" olarak adlandırılan ideal bir gizlilik kavramını sunmaktadır. Bu düşünce hem veri yayımcısının hem de saldırganın hassas ve hassas olmayan özniteliklerin ortak dağılımı hakkında tam bilgiye sahip olduğunu varsaymaktadır. Arka plan bilgisini nitelikler üzerinde bir olasılık dağılımı olarak modellemek için bu varsayım kullanılmaktadır. Ayrıca bir saldırganı, daha önceki (verilerin yayınlanmasından önce) ve sonraki düşünce modeline dayanarak (veriler yayınlandıktan sonra) Bayes formüllerini nicelleştirmektedirler. Önceki ve sonraki düşünce modeline dayanarak, Machanavajjhala ve arkadaşları resmi olarak üç gizlilik ilkesini tanımlamaktadır [56].

**Tanım 1 (Olumlu Bilgilendirme)** Orta olasılık, yüksek ihtimalle hassas öznitelik değerini, yani saldırganın bireyin duyarlı değeri olan (s) posterior değerinin $1-\alpha, \alpha > 0$ 'dan büyük olduğunu doğru olarak tanımlamaktadır.

**Tanım 2 (Olumsuz Bilgilendirme)** Bir düşman, olası hassas niteliklerin bazı olasılıklarını yüksek ihtimalle ortadan kaldırabilir, yani saldırgan kişinin duyarlı değeri $(s)$ için $\varepsilon, \varepsilon > 0$'dan küçük olduğu posterior verisi.

**Tanım 3 (Bilgilendirici ilke)** Yayınlanan veriler, arka plan bilgisinin ötesinde bazı ek bilgileri düşmana sağlamaktadır. Bu, ön ve arka plan bilgisi arasındaki farkın küçük olması gerektiği anlamına gelmektedir.

Pek çok eksiklik nedeniyle Bayes Optimal Gizlilik modeli pratikte uygulanamamaktadır.

**Yetersiz Bilgi:** Yayıncı, genel popülasyonda duyarlı ve hassas olmayan özelliklerin tam ortak dağılımını bilmemektedir.

**Bilinmeyen diğer bilgiler:** Bir düşmanın bu ortak dağılım bilgisine sahip olma olasılığı da düşüktür. Yine de veri yayıncısı bir saldırganın ne kadar arka plan bilgisine sahip olabileceğini bilmemektedir.

**Örnek Düzeyli Bilgi:** Bu gizlilik modeli, olasılıksal olarak modellenemeyen arka plan bilgisine karşı koruma sağlayamamaktadır. Örneğin, Ayşe Bahadır 'ın grip olmadığını söyledi.

**Çoklu düşmanlar:** Her biri farklı düzeyde arka plan bilgisi olan farklı saldırganların bulunması olası olabilmektedir. Bu, veri yayıncısının tüm bunları hesaba katması gerektiği anlamına gelmektedir, zira farklı arka plan bilgisi düzeyleri farklı çıkarımlara neden olabilmektedir.

### 2.2.5 $\ell$ -Çeşitlilik ($\ell$ - Diversity)

Optimal modelin sınırlamalarını aşmak için Machanavajjhala ve arkadaşları "çeşitlilik" kavramını önermektedir.

($\ell$-çeşitlilik) Tablodaki her $q^*$ bloğu, hassas nitelikteki ($S$) için en az $\ell$ "iyi temsil edilen" değer içeriyorsa, bir tablonun çok çeşitli olduğu söylenir [56].

Bu gizlilik kavramı "iyi temsil edilen" kavramın nasıl tanımlandığına bağlı olarak oluşturulabilmektedir. Kavram ile ilgili üç tanım bulunmaktadır: Belirgin $\ell$ -çeşitlilik, özyinelemeli $l$-çeşitlilik ve Çoklu özellikli $\ell$ -çeşitliliğidir. Bunlar aşağıda sırasıyla sunulmaktadır.

### 2.2.6 Belirgin Çeşitlilik (Distinct $\ell$ -Diversity)

$\ell$-çeşitliliğinin en basit tanımı Her $q^*$ bloğunun, hassas öznitelik $S$ için en azından $\ell$ farklı değer içermesi gerektiği anlamına gelmektedir. Ayrıca, $k = \ell$ durumunda k-anonimliği de ortaya çıkmaktadır.

Farklı $\ell$-çeşitlilik, mahremiyet konusunda çok güçlü bir kavram değildir, çünkü farklı değerlere sahip olması değerler hakkında hiçbir şey söylememektedir. Bunun üstesinden gelmek için Machanavajjhala ve arkadaşları Entropiye dayalı "iyi tanımlanmış" bir tanım vermek için tablodaki her $q^*$ bloğunda formüldeki değişkenlerin bulunması durumunda, bir tablonun $l$-çeşitli olduğunu belirtmektedirler [20]:

$$-\sum_{s\in S} P(qid,s)\log\left(P(qid,s)\right) \geq \log(l) \quad \text{(Denklem 2.2)}$$

Burada, $P(qid,s)$, $q^*$ bloğundaki hassas değer ($s$) olan kesirlerin frekanslarıdır. Değerlerin frekansları daha düzgün hale geldikçe entropinin artması sebebiyle "iyi temsil edilen" kavramına yaklaşmaktadır.

### 2.2.7 Özyinelemeli $(\mathbf{C},\ell)$-Çeşitlilik (Recursive $(\mathbf{C},\ell)$-Diversity)

"İyi temsil edilenler" için üçüncü tanım daha sezgisel olmaktadır. Buradaki ana fikir, sık değerlerin çok sık olmaması ve göreli olarak sık olmayan değerlerin de çok nadir olmaması gerektiğidir.

**Tablo 2.8:** Özyinelemeli Hasta Tablosu

| İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | HIV |
| Uzman | Erkek | [35-40) | HIV |
| Sanatcı | Kadın | [35-40) | HIV |
| Sanatcı | Kadın | [35-40) | HIV |
| Sanatcı | Kadın | [35-40) | HIV |
| Sanatcı | Kadın | [35-40) | Hepatit |
| Sanatcı | Kadın | [35-40) | Hepatit |
| Sanatcı | Kadın | [35-40) | Grip |

Duyarlı öznitelik $S$ 'nin değerleri frekansa göre sıralandığında. $f_i$, $q*$ -blokta en sık görülen değeri ifade etmektedir [39]. Aşağıdaki formül her $q*$ -blok için özyinelemeli $(c, l)$-diverse'i ifade etmektedir:

$$f_1 < c \sum_{i=l}^{n} f_i \qquad \text{(Denklem } \mathbf{2.3}\text{)}$$

$(c)$ Kullanıcı tanımlı bir değerdir. $c$ değerinin 1'den büyük veya eşit olması durumunda 2-çeşitliliğin sağlandığı anlamına gelmektedir. Daha yakından baktığımızda, bu tanımlamayı kullanarak, her $q*$-bloğunun, $S$ için en azından $\ell$ farklı hassas değerini içerdiğini görebiliriz. Bir diğer gözlem, $c$ için daha yüksek değerlerin gizlilik riskinin daha yüksek olmasına yol açmasıdır. $c < n, n = |q* - blok|$ için, en azından bir parça arka plan bilgisi, q *-b bloğu içindeki tuple'lerin $\frac{c}{n}$ 'sini ortadan kaldırabilmektedir [39].

### 2.2.8 Çoklu Özellik Yakınlığı (Multi-Attribute $\ell$-Diversity)

Nitelikler arasındaki korelasyon nedeniyle birden fazla hassas nitelik göz önüne alındığında, yeni sorunlar ortaya çıkmaktadır. Örneğin, bir saldırgan bazı hassas nitelikleri ortadan kaldırmak için arka plan bilgisi kullanıyorsa, s ile ilişkili diğer hassas değerleri de ortadan kaldırabilmektedir. Bunun üstesinden gelmek için Multi-

Attribute $\ell$ -diversity önerilmektedir çünkü $QID$ niteliklerini $Q_1 \ldots Q_n$ ve hassas nitelikler $S_1 \ldots S_m$. olarak varsayar.

Çoklu Özelleştirme-Çeşitliliği: Bir tablo çok-özniteliğe sahipse çok değişkenli olarak nitelenmektedir, çünkü tüm $i = 1..m$, için, $S_i$, tek hassas nitelik olarak ele alındığında tablo $\ell$-çeşitlidir ve $Q_1 \ldots Q_n S_I \ldots S_{i-1}\, S_{i+I} \ldots\ S_m$, $QID$ özniteliklerdir [39].

### 2.2.9 T-Yakınlığı (T-CLOSENESS)

Bir önceki bölümde, $\ell$-çeşitliliğin nitelik bağlantı saldırılarına karşı nasıl korunduğu tanımlanmıştır. Yine de, $\ell$-çeşitliliğinin nitelik açıklamasını engelleyemediği iki tür saldırı bulunmaktadır [20].

**Çarpıklık Saldırısı** (Skewness attack) Örneğin, bir hassas özniteliğe sahip tablo seçilsin: bir STD testinin sonucu. Geri kalan %99'u olumsuz olsun. Verilerin çarpıklığına bağlı olarak, değerlerin yarısının pozitif ve diğer yarısının negatif olduğu bir veri setine sahip olunabilir. Böyle bir durumda, bu $q*$ bloğuyla bağlantılı birinin pozitif olma olasılığının $\%50$ olmaktadır, ancak genel nüfusta %1 olacağı düşünülmektedir. Başka bir örnekte, bir $q*$-bloğunun $\%98'i$ pozitif ve $\%2'si$ negatif değerlere sahip olabilir. Bu bloğa bağlı olan biri neredeyse kesinlikle STD'ye sahip olarak kategorize edilebilir. Her iki örnekte de 2-çeşitlilik farklıdır.

**Benzerlik saldırısı** (Similarity attack) Bir $q*$ bloğundaki değerlerin farklı olması durumunda semantik olarak başka bir benzeri problemi ortaya çıkmaktadır. Örneğin, $q*$ bloğundaki tüm hastaların bir çeşit akciğer hastalığı olsun. Farklı değerlere sahip olmak, özellik açıklamaya karşı koruma sağlamamaktadır.

Bu soruna bir çözüm, t-yakınlık kavramıdır. Yazarlar, $q*$-blok yerine "eşdeğerlik sınıfı" (EC) kavramını kullanmaktadır.

$(t - yakınlık)$ Tablodaki her eşdeğerlik sınıfı ($q*$-block) için, gruptaki hassas değerlerin dağılımıdır ($t$ bütün nüfus).

Ancak dağılımlar arasındaki mesafeyi doğru bir şekilde nasıl ölçeceğiz? Genelde varyasyonel mesafe veya Kullback-Leibler ayrımı bunu ölçebilmektedir. Ancak, bizim durumumuzda, bir ekstra kısıtlamaya söz konusudur. Aynı zamanda semantik mesafede ölçülmelidir. Yazarlar bu amaçla sezgisel olarak, dağıtım kütlesini ikisi arasında hareket ettirerek, bir dağılımı diğerine dönüştürmek için gereken asgari iş olarak ölçülen bir kavram olan. Earth Mover Distance (EMD) kullanmaya karar

verdiler. Biri sayısal diğeri kategorik değerler için olan iki formül bulunmaktadır. $P$ ve $Q$'nun, her dağılımda sırasıyla, eleman $i$'nin $p_i$ ve $q_i$ olasılıklarıyla iki dağılım olsun. Ayrıca,$r_i = p_i - q_i$ olsun. Buna göre $P$ ve $Q$ arasındaki uzaklık:

$$D[P,Q] = \frac{1}{m-1}\sum_{i=1}^{i-m}\left|\sum_{j=1}^{j=i} r_j\right| \qquad \text{(Denklem } \mathbf{2.4}\text{)}$$

Sayısal bir örnek yardımcı olabilir. $P = \{3k, 4k, 5k\}$ ve $Q = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$ iki maaş dağılımı olsun. Böylece $p_i = \frac{1}{3}$ ve $q_i = \frac{1}{9}$ olur. Aşağıdaki değerler $r_1 \ldots 9 = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}, -\frac{1}{9}\}$ Sezgisel olarak, $r_1$ miktarını ilk elemandan ikinci sıraya, $r_1 + r_2$ 'yi ikinci elemandan üçüncü elemana vb. hareket ettirerek. 3.1 denklemi uygulanırsa, $P$ ve $Q$ arasındaki mesafeyi elde ederiz

$$D[P,Q] = \frac{1}{9-1}\left(\frac{2}{9} + \frac{4}{9} + \frac{6}{9} + \frac{5}{9} + \frac{4}{9} + \frac{3}{9} + \frac{2}{9} + \frac{1}{9}\right) = 0.375$$

Sezgisel olarak, $P$ ve $Q$ arasındaki mesafe büyüktür. Bir kişi bu EC'ye bağlı kalırsa, saldırgan, EC'deki tüm değerlerin küçük olması nedeniyle, kişinin küçük bir maaşa (tam dağıtım ile karşılaştırıldığında) sahip olduğuna karar verebilir.

**Tablo 2.9:** 3-Çeşitlilik / Yakınlık-(0.375) Tablosu

| Posta Kodu | Yaş | Maaş |
|---|---|---|
| 476** | 2* | 3k |
| 476** | 2* | 4k |
| 476** | 2* | 5k |
| 479** | ≥ 40 | 6k |
| 479** | ≥ 40 | 11k |
| 479** | ≥ 40 | 8k |
| 476** | 2* | 4k |
| 476** | 2* | 7k |
| 476** | 2* | 10k |

**Tablo 2.10:** Yakınlık-(0.167) Tablosu

| Posta Kodu | Yaş | Maaş |
|---|---|---|
| 4767* | 2* | 3k |
| 4767* | 2* | 5k |
| 4767* | 2* | 9k |
| 4790* | $\geq 40$ | 6k |
| 4790* | $\geq 40$ | 11k |
| 4790* | $\geq 40$ | 8k |
| 4760* | 2* | 4k |
| 4760* | 2* | 7k |
| 4760* | 2* | 10k |

$D\,[P,Q]\,0.375$ sezgiselliği kanıtlar. $t$ için iyi değerler genellikle $0{,}1$ den küçüktür. Örneğin, t-yakınlığından önce Tablo 2.9'de, t-yakınlığı uygulandıktan sonra, Tablo 2.10'de gösterilmiştir.

Yakınlık, k-anonimlik ve $\ell$-çeşitlilik sorunlarını çözmektedir, ancak kendi sınırlamaları bulunmaktadır. Bu sınırlamalardan birincisi, farklı hassas değerler için farklı koruma seviyeleri belirleme esnekliğinden yoksun olmasıdır. İkincisi, EMD mesafesi sayısal değerler için uygun olmamaktadır. Üçüncüsü, verilerin bozulması çok fazladır çünkü her EC'nin benzer bir dağılımı olmalıdır.

### 2.2.10 (N, T) Yakınlığı ((N, T)-Closeness)

(n, t)-yakınlık, t-yakınlığın sınırlamalarını telafi etmeye çalışmaktadır. Yakınlığın bir uzantısı olarak verileri daha az deforme eden bir algoritmaya sahiptir. Tüm eşdeğerlik sınıflarının (EC) dağılımını tam tablo dağılımına olan uzaklığa zorlamak yerine, EC'nin veri kümesinin yeterince büyük bir popülasyonunu $t$ 'de olmasını şart koşmaktadır [20].

**((n, t) -closeness)** $E_1$ denkleminin, $E_1$'in doğal bir üst seti olması, $n$ 'den fazla elemana sahip olması ve $E_1$'in doğal bir üst kümesi olması durumunda bir EC $E_2$ varlığında, (n, t) $E_1$ ve $E_2$'nin hassas nitelik dağılımları arasındaki fark en fazla $t$ kadar olur. Tablodaki tüm eşdeğerlik sınıfları varsa, bir tablo (n, t)-yakınlığı özelliğine sahiptir [20].

Eşdeğerlik sınıfının doğal üst seti, belirlenen değer sınırlarını genişleterek onu içeren bir denklik sınıfıdır. Örneğin $E_1$'i tablo 2.12'deki ilk EC olarak düşünelim. $E_1$

böylece ($posta\ kodu = '476\ **', yaş = [20{,}29]$) olarak tanımlanır. Bu durumda "doğal" ($posta\ kodu = '476\ **', yaş = [20{,}29]$) verisini içerdiğinden bir üst set olacaktır.

Örneğin, Tablo 2.11 ve 2.12'e görülmektedir. Tablo 2.12, 0.1 yakınlığı dikkate almamakta ancak (1000, 0.1) açıklığı dikkate almaktadır. Üç eşdeğerlik sınıfı vardır. İkincisi, kendisinin doğal bir üst kümesidir (1000 unsur içerir) (1000, 0,1)–yakınlığa ulaşmaktadır. İlk ve üçüncü elementlerin 1000 elementi yoktur ve dağılım değişmemektedir.

**Tablo 2.11:** Hasta Tablosu

| Posta Kodu | Yaş | Hastalık | Sayı |
|---|---|---|---|
| 47673 | 29 | Kanser | 100 |
| 47674 | 21 | Grip | 100 |
| 47605 | 25 | Kanser | 200 |
| 47602 | 23 | Grip | 200 |
| 47605 | 43 | Kanser | 100 |
| 47904 | 48 | Grip | 900 |
| 47906 | 47 | Kanser | 100 |
| 47907 | 41 | Grip | 900 |
| 47603 | 34 | Kanser | 100 |
| 47605 | 30 | Grip | 100 |
| 47604 | 36 | Kanser | 100 |
| 47607 | 32 | Grip | 100 |

**Tablo 2.12:** Yakınlık-(1000,0.1)

| Posta Kodu | Yaş | Hastalık | Sayı |
|---|---|---|---|
| 476** | 2* | Kanser | 300 |
| 476** | 2* | Grip | 300 |
| 479** | 4* | Kanser | 200 |
| 479** | 4* | Grip | 1800 |
| 476** | 3* | Kanser | 200 |
| 476** | 3* | Grip | 200 |

**Belirsizliklerin kimliği** $D\ [P, P] = 0$; inanç değişmezse hiç bilgi kazanılmaz.

**Negatif Olmama** Bilgiler yine pozitif olabilir; $D\ [P, Q] \geq 0$

**Olasılık Ölçekleme** Birinin inanç ihtimalini sabit bir γ ile artırmak, ilk inançta daha büyük inanç için daha büyük olmalıdır; $(\alpha \rightarrow \alpha + \gamma) > (\beta \rightarrow \beta + \gamma);\ \alpha < \beta$

**Sıfır Olasılık Tanımlanabilirliği** Mesafe metriği dağılımlardaki sıfır olasılık değerlerini işleyebilmelidir (örneğin KL ayrımı yok).

**Semantik farkındalık** Metrik, eğer varsa, değerler arasındaki semantik mesafeyi yansıtmalıdır.

Önceden önerilen EMD metriğinin olasılık ölçeklemesi yoktur. Bu nedenle, yukarıda belirtilen tüm hususlara dikkate alan yeni bir metrik önerildi. İki adımdan oluşan metriğin ilk adımı çekirdek yumuşatma. Nadaraya-Watson ağırlıklı ortalamayı her iki dağıtımda da uyguluyorlar:

$$\hat{p} = \frac{\sum_{j=1}^{m} K(d_{ij})}{\sum_{j=1}^{m} K(d_{ij})} \qquad \text{(Denklem } \mathbf{2.5}\text{)}$$

burada $K(.)$ çekirdek fonksiyonu ve $d_{ij}$, hassas $s_i$ ve $s_j$ arasındaki mesafe olmaktadır. Bunun en büyük avantajı, anlamlar arasındaki mesafeyi belirleyen değerler arasında bir uzaklık matrisi gerektirmesidir. Mesafeyi hesaplamak için iki yol vardır [20].

**Tablo 2.13:** Koruyucu Düğümlü Hasta Tablosu

| Yaş | Cinsiyet | Posta Kodu | Hastalık | GN |
|---|---|---|---|---|
| 5 | m | 12000 | mide ülseri | Mide hastalığı |
| 9 | m | 14000 | hazımsızlık | hazımsızlık |
| 6 | m | 18000 | zatürre | Solun yolu enfeksiyonu |
| 8 | m | 19000 | bronşit | bronşit |
| 12 | m | 22000 | zatürre | Solun yolu enfeksiyonu |
| 19 | m | 24000 | zatürre | Solun yolu enfeksiyonu |
| 21 | k | 58000 | grip | Ø |
| 26 | k | 36000 | gastrit | gastrit |
| 28 | k | 37000 | zatürre | Solun yolu enfeksiyonu |
| 56 | k | 33000 | grip | grip |

İki hassas değer $s_1$ ve $s_2$ verildiğinde sol formül sayısal ve kategorik durumda doğru olan formüldür:

$$d_{ij} = \frac{|s_i - s_j|}{R}$$ (Denklem **2.6**)

$$d_{ij} = \frac{h(s_i - s_j)}{H}$$ (Denklem **2.7**)

Burada, $R$ öznitelik etki alanı aralığı, $H$ taksonomi ağacının yüksekliği ve $h\ (-)$, uzaklık ağacındaki en düşük ortak atanın yüksekliğidir.

Çekirdek yumuşatma işleminden $\hat{P}$ ve $\hat{Q}$ elde edilir. Şimdi, $D\ [\hat{P}, \hat{Q}]$ mesafesini hesaplamak yerine, Jensen-Shanon ayrımı kullanarak mesafenin tahmini $D\left[\hat{P}, \hat{Q}\right]$ hesaplanmaktadır:

$$JS[P, Q] = \frac{1}{2}(KL[P, avg(P, Q)] + KL[Q, avg(P, Q)])$$ (Denklem **2.8**)

Burada $avg(P, Q)$ ortalama dağılımı $(P + Q)/2$'dir.

Ayrıca, (n,t)-yakınlığını başarmak için bir algoritma önermektedirler. Bu, Mondrian algoritmasına dayanmaktadır. Ağaç bölümleri vardır: bölme için bir boyut seçme, bölme değeri seçme ve bölümlemenin gizlilik koşullarını ihlal edip etmediğini kontrol etme. Onların modifikasyonu, üçüncü bölüm için bir kontrol algoritmasından oluşur.

## 2.3 Mahremiyet Saldırıları

### 2.3.1 Bağlantı Saldırıları (Countering Linkage Attacks)

Bağlantı saldırıları, adından da anlaşılacağı üzere, bir kişiyi belirli bir tabloda bir rekor ya da bir değere bağlamak ya da tablonun kendi içinde yokluğun varlığını belirlemek için çalışmaktadır. Bu tür saldırılara sırasıyla kayıt bağlantısı, öznitelik bağlantısı ve tablo bağlantısı denmekte ve aşağıda daha ayrıntılı olarak ele alınmaktadır.

### 2.3.2 Bağlantı kaydı (Record linkage)

Yukarıda belirtildiği gibi kayıt bağlantı saldırıları, bir kişiyi yayınlanmış bir veri kümesindeki bir rekora bağlamaya çalışmaktadır. Tablo 2.14'de gösterilen hasta tablosunu ele alalım. Burada Hastalık hassas nitelik olarak kabul edilmektedir. Bireylerin kimliği hakkında çok fazla şey söylenemez. Bununla birlikte, bir saldırgan Ayşe'nin hastaneye gittiğini biliyorsa ve Ayşe'nin 36 yaşındaki kadın dansçı olduğunu aynı zamanda yaş, cinsiyet ve iş niteliklerini biliyorsa saldırgan hangi kaydın Ayşe'ye ait olduğunu bulabilmektedir. Bağlantı, Ayşe için qid = {yaş = 36, cinsiyet = kadın, iş = dansçı} değerlerini alan QID temel alınarak yapılır.

**Tablo 2.14**: Bağlantı Kaydını Gösteren Hasta Tablosu

| İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|
| Mühendis | Erkek | 35 | Hepatit |
| Mühendis | Erkek | 38 | Hepatit |
| Avukat | Erkek | 38 | AİDS |
| Yazar | Kadın | 35 | Grip |
| Yazar | Kadın | 35 | AİDS |
| Sanatçı | Kadın | 35 | AİDS |
| Sanatçı | Kadın | 36 | AİDS |

| Name | İş | Cinsiyet | Yaş |
|---|---|---|---|
| Ayşe | Dansçı | Kadın | 36 |
| … | … | … | … |

### 2.3.3 Özellik Bağlantısı (Attribute Linkage)

Bir öznitelik bağlantı saldırısı söz konusu olduğunda, hedef hangi hassas değerin mağdura ait olduğunu belirlemektir. k-anonimlik ve varyasyonları sadece kayıt bağlantı saldırılarına karşı koruma sağlamaktadır. İki saldırı öznitelik bağlantı kategorisine girer bunlar, homojen saldırı ve arka plan bilgisi saldırısıdır.

### 2.3.4 Homojenlik saldırısı

Homojenlik saldırısı durumunda, problem, bir $q*$-blokundaki tüm hassas değerlerin aynı değere sahip olması gerçeğidir. Örneğin, 3-isimsiz hasta tablosu olan Tablo 2.15'i ele alalım. Saldırganın Ayşe'yi (35 yaşındaki kadın, yazar) tanıdığını ve

bu tabloyu yayınlayan hastaneye gittiğini varsayalım. Verilerdeki çarpıklığa bağlı olarak, saldırgan, yazar 35 yaşında ve aynı hastalığa yakalanmış tüm kadınları görebilmektedir: Saldırgan, Ayşe'nin de HIV olduğuna karar verir. Bu saldırı resmi olarak olumlu açıklama olarak bilinmektedir.

**Tablo 2.15:** Anonim-3 Hasta Tablosu

| İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | HIV |
| Sanatçı | Kadın | [35-40) | HIV |
| Sanatçı | Kadın | [35-40) | HIV |
| Sanatçı | Kadın | [35-40) | HIV |

### 2.3.5 Artalan bilgi saldırısı (Background knowledge attack)

Bu saldırıda, saldırgan bir mağdurun hassas nitelikleri ve olası değerlerini ortadan kaldırmak için arka plan bilgisini kullanmaktadır. 3-isimsiz veriyi içeren Tablo 2.16'deki verilerde çarpıklık yoktur, bu nedenle homojen bir saldırı gerçekleşememektedir. Saldırganın Bahadır'a bakmak istediğini varsayılsın (38, erkek, avukat). Tabloya bakarak, Bahadır'ın sadece grip veya hepatit olduğuna karar verilebilirdi. Şimdi saldırgan Bahadır'ı bildiğini ve ayrıca, görünür belirtilerin olmamasından dolayı gribe yakalanmadığını da biliyor olsun. Bu bilgiyi kullanarak, saldırgan, Bahadır'ın sahip olabileceği tek hastalığın hepatit olduğuna inanmaktadır. Bu saldırıyı içine alan gizlilik ilkesine olumsuz açıklama denmektedir.

**Tablo 2.16:** Anonim-3 Hasta Tablosu

| İş | Cinsiyet | Yaş | Hastalık |
|---|---|---|---|
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | Hepatit |
| Uzman | Erkek | [35-40) | Grip |
| Sanatçı | Kadın | [35-40) | Grip |
| Sanatçı | Kadın | [35-40) | HIV |
| Sanatçı | Kadın | [35-40) | HIV |

## 2.4 Veri Modelleri

Şimdiye kadar ele alınan tüm çalışmalar, ilişkisel ve istatistiksel verileri anonimleştirmeye odaklanmaktadır. Son yıllardaki araştırmalar yayıncılık işlem verilerinin, hareketli nesne verisinin ve metinsel verilerin acil tehdit ve hassas bilgi sızıntılarına neden olabileceğini göstermektedir [56]. Aşağıda, gizlilik tehditlerine karşı bazı koruma çözümleri ile birlikte bu ilişkisel olmayan veri türleri tartışılmaktadır.

### 2.4.1 Yüksek Boyutlu İşlem Verileri

Yüksek boyutlu verileri yayınlamak ticari ve kamusal faaliyetlere ait günlük operasyonların bir parçasıdır. Yüksek boyutlu verilerin klasik bir örneği, işlem veritabanlarıdır. Her işlem bir kayıt sahibine karşılık gelir ve büyük bir evrenden seçilmiş bir grup öğeden oluşur. İşlemlerin örnekleri, web sorguları, tıklama akışları, e-postalar, pazar sepetleri ve tıbbi notlardır. Bu tür veriler çoğunlukla zengin bilgiler içerir ve veri madenciliği için mükemmel bir kaynaktır. Ayrıntılı işlem verileri, muhtemelen hassas bilgiler içeren bir kayıt sahibinin hayatının elektronik bir görüntüsünü sağlamaktadır.

Yakın tarihli bir olay, işlem verilerini yayınlamaktan kaynaklanan gizlilik tehditlerini göstermektedir [66]: AOL, araştırma amacıyla kamuya sorgu günlükleri veritabanını yayınlamıştır. Bununla birlikte, sorgu terimlerini inceleyerek, 4417749 numaralı AOL kullanıcısı, Lilburn'de yaşayan 62 yaşındaki bir dul olan Bayan Thelma Arnold'a geri dönmüştür. Bir sorgu bir adres veya ad içermese bile, bir kayıt sahibi (bu örnekte AOL kullanıcısı), kayıt sahibine yeterince benzeyen sorgu terimleri kombinasyonlarından yeniden tanımlanabilmektedir. Bu güvenlik açığı sadece AOL kullanıcılarının özel bilgilerinin ifşa edilmesine değil, aynı zamanda veri yayıncılarının araştırma amaçlı anonim işlem verileri sunma konusundaki isteklerine de kısıtlama getirmektedir. Kumar ve arkadaşları ayrıca, bazı "referans" sorgu günlüklerinde de göstergenin eş zamanlı oluşumlarına dayalı olarak karma işlevinin ters çevrilmesiyle bazı belirteç-karma anonim sorgu günlüklerinin ifşa olabileceğini göstermişlerdir. Açıkçası, işlem verisi için uygun bir anonimleştirme yöntemine ihtiyaç bulunmaktadır.

İşlem verileri genellikle yüksek boyutludur. Örneğin, Amazon.com birkaç milyon katalog öğesine sahiptir. Her boyut, kayıt veya özellik bağlantısı için kullanılan olası bir QID özniteliği olabilir; Bu nedenle, k-anonimliği gibi geleneksel anonimlik modellerinin kullanılması, tüm boyutların tek bir QID'ye dâhil edilmesini gerektirmektedir. Yüksek boyutluluk sorunu nedeniyle, k küçük olsa bile, k-anonimliğini sağlamak için birçok veri bastırılmalı veya en üst değerlere genelleştirilmelidir. Açıkçası, bu gibi isimsiz veriler, veri analizi için yararsız hale gelmektedir [66].

Yüksek boyutlu verilerin anonimleştirilmesine ilişkin bazı yeni çalışmalar bulunmaktadır. Ghinita ve arkadaşları genel fikri, yakın çevredeki işlemleri bir araya toplayıp sonra da her bir grubu çeşitlendirilmiş hassas değerler grubuyla ilişkilendirmek olan bir permütasyon yöntemi önermektedir. Gerçek hayatta gizlilik saldırısında, saldırganın hedef bilgi birikimini toplamak için gereken çaba sebebiyle, hedef mağdurun tümüyle tanımlayıcı niteliklerini bilebileceği ihtimal dâhilinde değildir. Dolayısıyla, saldırganın arka plan bilgisini gizlilik modeli ile sınırlamak mantıklıdır. Terrovitis ve arkadaşları, işlemleri genelleştirerek k-anonimleştirmek için bir algoritma önermektedirler. Xu ve arkadaşları geleneksel k-anonimlik modelini, saldırganın hedef mağdurun en fazla m adet hareket öğesini bildiğini varsayarak genişletmişlerdir. Özellikle, Xu ve arkadaşlarının gizlilik modeli, (1) yayınlanan tablodaki $m$'den büyük olmayan her öğe setinin en az $k$ kayıt tarafından paylaşılmasını ve (2) hassas değerin $s$'den çıkarılma güven aralığının, maksimum güven aralığı eşik değerinin altındadır ($h$) bulgularını içermektedir [27]. Sonuçlar, veri kullanımını büyük ölçüde artırabildiğini göstermektedir. Başka bir çalışmada, Xu ve arkadaşları sık kullanılan öğeleri veri yardımcısı olarak korumayı önermektedirler [27]. Öğelerin üssel olarak artmasından kaynaklanan ölçeklenebilirlik darboğazı ile uğraşmak için Xu ve arkadaşları, gizlilik gereksinimini ve sık öğe setlerini ihlal eden öğeleri temsil etmek için sınırlar adı verilen maksimum ve minimum öğe setlerini kullanmaktadırlar. Her iki çalışmada da; Xu ve arkadaşları işlem verisi için taksonomi ağaçlarını düz ve fanout olma eğiliminde olduğundan, genelleme yerine madde bastırmada kullanmaktadırlar. Bu durumda, genelleme kullanmak öğe bastırma işleminden daha fazla bilgi kaybına sebep olmaktadır. Aggarwal ve Yu, eskiz tabanlı bir uygulama için bir anonimlik modeli oluşturmuşlar ve orijinal verinin eskiz temelli gizlilik koruyan temsillerini oluşturmak için kullanmışlardır. Eskiz temelli yaklaşım, daha özgün özellik değerlerinin ağırlıklı bir toplamını üretmek için her biri farklı rastgele ağırlık

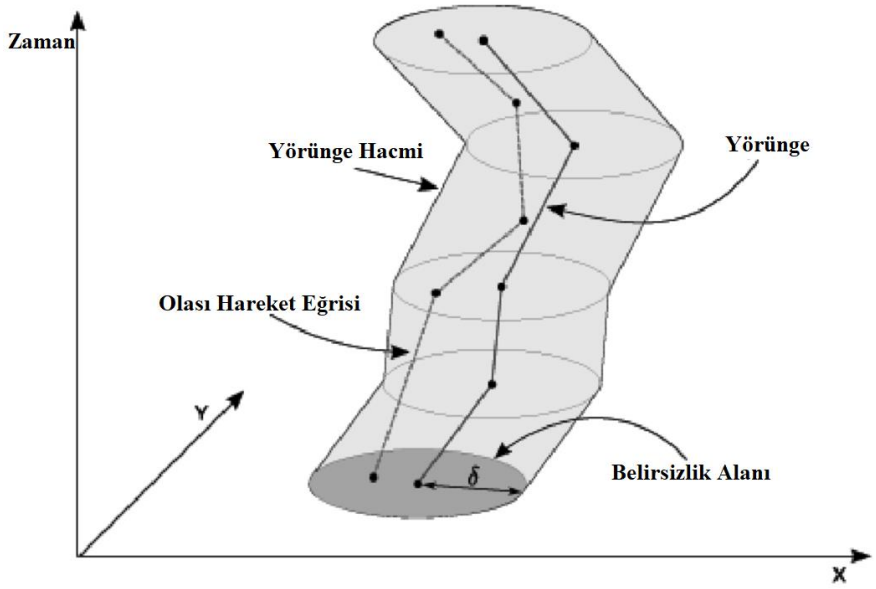
kümesini kullanan çok daha küçük sayıda özellikli yeni bir gösterim üreterek verilerin boyutunu düşürmektedir [30]. Bu teknik, veri seyrek olduğu sürece yüksek boyutlu veri kümeleri için oldukça etkilidir. Eskiz tabanlı yöntem, birçok toplu mesafe önleminin etkili bir şekilde yeniden yapılandırılmasına izin verirken gizlilik koruması sağlamaktadır. Bu nedenle, kümeleme ve sınıflandırma gibi çeşitli veri madenciliği algoritmaları için kullanılabilmektedir.

### 2.4.2 Nesne Verilerini Taşıma

Konum Tabanlı Hizmetler (LBS), cep abonelerine belirli fiziksel konumlarına dayanarak sağlanan bilgi servisleridir. Son yıllarda, abonelerden gelen talep artışı nedeniyle çeşitli konuma dayalı hizmetler geliştirilmiştir. Telekomünikasyon teknolojisinin ilerlemesiyle yaşam kalitesinin yükselmesine rağmen, yapılan araştırmalar, potansiyel LBS kullanıcılarının %24'ünün konumlarını diğer kişisel verileriyle birlikte açıklanmasından ciddi olarak endişe duyduğunu ortaya koymuştur. Nesne verilerini hareket ettirmek, benzersiz özelliklerinden dolayı geleneksel veri tabanı, veri madenciliği ve gizlilik koruyan teknolojilere yeni zorluklar getirmektedir çünkü zamana bağlı, konuma bağlı ve yüksek hacimli akış verisi üretilmektedir. Aşağıdaki örnek, hareket eden nesne verilerini yayınlamaktan kaynaklanan gizlilik tehditlerini göstermektedir.

**Tablo 2.17:** Hastaya Özgü Yol Tablosu T

| **Pid** | **Yol** | **Hastalık** |
|---|---|---|
| 1 | *(a1→d2→b3→e4→f6→c7)* | HIV |
| 2 | *(b3*→e4→f6→e8) | Grip |
| 3 | *(b3*→c7→e8) | Grip |
| 4 | *(d2*→f6→c7→e8) | Alerji |
| 5 | *(d2*→c5→f6→c7) | HIV |
| 6 | *(c5*→f6→e9) | Alerji |
| 7 | *(d2*→c5→c7→e9) | Ateş |
| 8 | *(f6*→c7→e9) | Ateş |




**Şekil 2.3:** Zaman ve Mekânsal Yörünge Hacmi [33]

Bir hastane, tablo 2.17'e göre hastaya özgü yol tablosunu veri analizi için üçüncü bir tarafa bırakmak istiyor olsun. Hasta isimleri ve Pid gibi açık tanımlayıcılar kaldırılmıştır. Her kayıt, hastanedeki bir hastanın hastalıklar gibi hastaya özgü (hassas) bazı bilgileri içermektedir. Hastanın ziyaret ettiği konum, zaman damgasını gösteren çiftler dizisini ($loc_i t_i$) içerir. Örneğin, Pid 3'ün yolu (b3 → c7 → e8), yani hastanın sırasıyla 3, 7 ve 8 zaman damgalarında b, c ve e konumlarını ziyaret ettiği anlamına gelir.

Bir saldırgan eşleme için süreci QID olarak kullanarak kayıt ve / veya öznitelik bağlantılarını gerçekleştirmek için kullanmak istemektedir. (1) Bağlantı kaydedilir: saldırgan, hedef kurban Ayşe'nin $e$ ve $c$'yi zaman damgaları 4 ve 7'de sırasıyla ziyaret ettiğini bildiğini varsayalım. Ayşe'nin kaydı, hassas değeri (bu durumda HIV) ile birlikte benzersiz bir şekilde tanımlanabilir çünkü Pid 1 e4 ve c7 içeren tek kayıttır. (2) Öznitelik bağlantısı: saldırgan, başka bir hedef kurbanın Bahadır'ın d2 ve f6'yı (Pid 1,4,5) eşleştirdiğini bildiğini varsayarsa, saldırgan Bahadır'ın HIV'e sahip olduğunu $2/3 = \%67$ ile tahmin edebilir.

Hareketli nesneleri anonimleştirmeye yönelik birkaç yeni eser var. Abul ve arkadaşları geleneksel $k$-anonimlik modelini hareketli nesneleri anonimleştirmek için genişletmişlerdir. Şekil 2.2'de, en az $k$ adet hareket eden cisimin yolunun yarıçapı δ'ında görünmesini sağlamaktadır. Hastane ortamında konuma dayalı verilerin gizlilik tehditlerini incelendi ve geleneksel anonimleştirme operasyonlarına ek olarak, orijinal yollara gürültü eklenerek uzay çevirimi araştırılmıştır; böylece daha fazla cisimin aynı

anda ve mekânsal yörüngede yer alması mümkün olabilmiştir. Konumlar hassas bilgilerdir ve saldırganın bilmediği hedef mağdurun ziyaret ettiği bazı hassas yerler bu sayede bulunabilmektedir. Fung ve arkadaşları, yüksek boyutlu RFID hareket eden nesne verilerini anonimleştirmek için ilk çalışmayı sunmuşlardır. Önerilen gizlilik modeli, LKC gizliliği olarak betimlenmektedir. $L$'den büyük olmayan uzunluğa sahip her RFID hareketli yolunun en az $K-1$ diğer hareketli yollarla paylaşılmasını ve önceden belirlenmiş herhangi bir duyarlı değeri tahmin için güvenin $C$'den büyük olmamasını sağlamaktadır.

Papadimitriou ve arkadaşları, zaman serisi verilerin yayınlanmasındaki gizlilik konusunu ve zaman serisi sıkılaştırılabilirliği ile kısmi bilgi gizleme arasındaki dengeyi ve bunların bireylerin değerlerini değiştirecek şekilde nasıl bir belirsizlik yaratması gerektiğine ilişkin temel etkileri incelemiştir. Çalışma, pertürbasyonun orijinal verilere "benzer" olmasını sağlayarak, verilerin yapısını daha iyi korumakta ve aynı zamanda ihlalleri zorlaştırmaktadır. Bununla birlikte, veriler daha sıkıştırılabilir olursa, gerçek değerler sızdırıldıkça, belirsizliğin bir kısmı giderilebilmektedir.

### 2.4.3 Metin Verileri

Önceki çalışmaların çoğu, yapısal ya da yarı yapılandırılmış verilerin anonimleştirilmesine odaklanmıştır. Metin belgeleri gibi yapısal olmayan veriler ne olmalıdır? Saygin ve arkadaşları, metin belge depolarında örtülü ve açık gizlilik tehditlerini açıklamaktadır. Metin belgelerinin ayrıştırılması, hassas bilgileri bir kişide veya bir belgedeki hassas bilgilere bağlanabilecek bağlantı bilgilerini kaldırmayı içerir. Bu çalışma henüz olgunlaşma aşamasındadır.

Kokkinakis ve Thurin, bazı önceden tanımlanmış hassas varlık türlerini karşılayan klinik metinden tüm cümleleri tanımlayarak ve kasten kaldırarak hastane taburcu mektuplarını otomatik olarak anonimleştiren bir sistem geliştirmişlerdir [30]. Tanıma aşamasına, altta yatan genel adlandırılmış varlık tanıma sistemini kullanarak bunu yaptılar.

Chakaravarthy ve diğerleri [2008], önceden tanımlanmış hassas cisim tipleri içeren evreleri basitçe kaldırmak yerine, en az karışık bir dokümanı ayrıştırarak ERASE sistemini sunmuşlardır. Bir veritabanını kendi bağlamıyla ilişkilendirmek için dışsal bilgi gerekmektedir. ERASE, korunan varlıkların, bağlamlarının belirli koşullarını kaldırarak açıklamasını engellemektedir; böylece kalan belge metninden

korumalı varlık çıkarılabilmektedir. k-güvenliği, k-anonimliği ile aynı şekilde tasnif edilir. Her korunan varlık ile kesişim noktası en az $k$ varlık içeriyorsa, bir dizi k-güvenli olmaktadır. Ardından önerilen problem, k-güvenliğini sağlayan bir belgenin maksimum kararlılıktaki alt kümesini bulmaktadır. Chakaravarthy ve arkadaşları k-emniyetini sağlamak için hem küresel bir optimal algoritma hem de etkin bir açgözlü algoritma önermiş ve değerlendirmiştir [30].

## 2.5 Tanımlayıcılar

### 2.5.1 Veri Gizliliğini Koruyarak Yayınlama (Privacy Preserving Data Publishing PPDP)

Veri anonimleştirmesi ile uğraşan birçok araştırma bulunmaktadır. En çok Açık Veri fikri (verilerin nihai kullanımını bilmeden yayımlamak) araştırma gruplarının ilgisini çekmektedir [39]. Çalışmalar, arka plan saldırıları, hassas özniteliklerin çıkarılması, genellemeyi ve veri yararını ölçen çeşitli kavramları dikkate alan çalışmalar ile genişletilmektedir.

### 2.5.2 Gizlilik Koruması

Farklı anonimleştirme teknikleri tartışılmadan önce yayınlanan verilerin gizliliğini korumanın ne anlama geldiği anlaşılmalıdır. Gizliliğin korunmasına ilişkin ilk tanım 1977 yılında Dalenius tarafından yapılmıştır [27]. Buna göre; Gizlilik koruması, yayınlanan verilere erişim, saldırganın, herhangi bir saldırganın diğer kaynaklardan edindiği bilgi birikiminin varlığıyla bile, veritabanına erişim iznine kıyasla herhangi bir hedef kurban hakkında ekstra bir şey öğrenmesine izin vermemesi olarak tanımlanmaktadır.

Bu tanım iyi olabilir ancak çok katıdır. Çünkü Dwork, arka plan bilgisinin varlığı nedeniyle böyle bir mutlak gizliliğin korunmasının imkânsız olduğunu kanıtlamıştır [30]. Bu kısıtlamayı ortadan kaldırmak için, PPDP literatürü daha rahat ve daha pratik bir tanımlama benimsemiştir. Revize edilen tanıma göre ise gizlilik koruması saldırganın, yalnızca belirli bir arka plan bilgisi içeriğine sahip olduğu varsayımı altında, veritabanına erişemeden sadece yayınlanan verilere erişerek herhangi bir hedef kurban hakkında ekstra bir şey öğrenmesine izin verilmemesidir.

### 2.5.3 Tablolar ve nitelikler (Tables and attributes)

PPDP hakkındaki literatürlerin çoğu, veri yayıncısının birkaç özellik türünü içeren tek bir tablo bıraktığını varsayarak başlamaktadır. Bu varsayım genellikle daha sonra ilgili çalışmada bırakılır. Literatür, dört farklı özellik türü arasında ayrım yapmaktadır (tablo sütunları) [39].

Tanımlayıcılar (Identifiers) Belirli bir tabloda bulunan hassas bilgilerin bazı parçalarını tamamen veya belirsiz bir şekilde tanımlayan nitelikler veya bunların bir kümesidir. Bir veri kümesi yayınlandığında, bu nitelikler daima kaldırılır. Örneğin SSN, pasaport numarası ve isim gibi.

### 2.5.4 Yarı tanımlayıcılar (Quasi-identifiers (QID)

Belirli bir anonim tabloda bulunan kişileri benzersiz şekilde tanımlamak veya harici bilgilerle bağlantı oluşturmak için kullanılan bir dizi nitelikleri temsil etmektedirler. Bunlar, ilk bakışta zararsız görünebilecek nitelikleri (posta kodu, cinsiyet, yaş v.b.) içermektedirler. Ancak Sweeney'e göre, ABD nüfusunun %87'si büyük olasılıkla yalnızca QID'e (posta kodu, cinsiyet, yaş) göre tahmin edilebilmektedir [39]. Hangi niteliklerin yarı tanımlayıcılar olarak ele alınacağını belirlemek veri yayımcısına düşmektedir. Daha sonra gösterileceği gibi, bu kategoriye çok fazla özellik eklenmesi, verinin faydasını önemli ölçüde etkilemekte ve bireylerin mahremiyetine karşı bir risk oluşturmaktadır. PPDP literatürü, bir saldırganın bir bireyin QID değerlerini tam olarak bildiğini kabul eder. Genellikle, yalnızca bağlantı saldırılarıyla uğraşan gizlilik modelleri bu kavramı kullanmaktadırlar.

Hassas özellikler (Sensitive attributes (S) Bu öznitelikler mağdurun duyarlı olduğu düşünülen değerleri içermektedir. Bunlara örnek olarak maaş ve hastalık verilebilir.

Hassas olmayan öznitelikler ((Non-sensitive attributes (NS)) Bu öznitelikler, daha önce bahsedilen kategorilerin herhangi birine girmeyen sütunlardan oluşmaktadır.

Özniteliklerin sınıflandırıldığı sıralama, yukarıdaki ile aynıdır: tanımlayıcılar, yarı tanımlayıcılar ve daha sonra hassas niteliklerdir. Geri kalanlar hassas olmayan nitelikler olarak değerlendirilmektedir. Böylece, çözüm bir saldırganın bir kişiyi hassas bilgilere bağlama becerisini engellemeye dayanmalıdır.

Gizlilik modelleri (Privacy models) Mahremiyeti düşündüğünüzde ikiliği gözlemleyebilirsiniz. Bir yandan, bir şahsın mahremiyetine, diğer yandan da şahsınızın mahremiyetine sahipsiniz. Literatürde tartışılan pek çok gizlilik modeli bulunmaktadır. Modelleri gruplamanın bir yolu, önlemeye çalıştıkları saldırı türüne dayanmaktadır. Fung ve arkadaşları iki kategoriyi tanımlamışlardır. Bunlar: bağlantı saldırılarına karşı çıkan gizlilik modelleri ve olasılıklı saldırılardır.

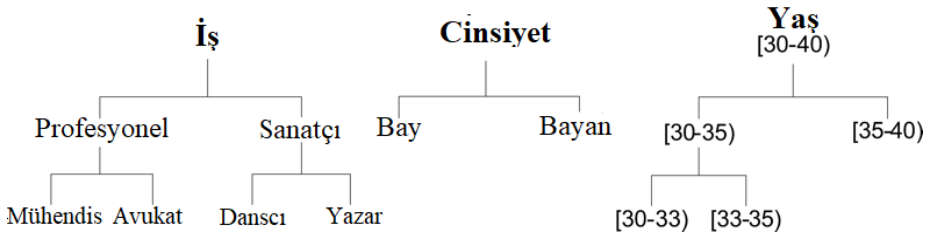
## 2.6 Ölçümün Faydaları (Measures of Utility)

### 2.6.1 Genel Amaçlı Bilgi Metriği

Veri yayıncısı çoğu durumda, yayınlanan verilerin alıcı tarafından nasıl analiz edileceğini bilmemektedir. Bu anlamda veri madenciliği görevinin bilinmekte olduğunu varsayan PPDM'den (privacy-preserving data mining) çok farklıdır. PPDP'de, örneğin, veriler Web'de yayınlanabilmekte ve bir alıcı, veriyi kendi amacına göre analiz edebilmektedir. Alıcı için iyi bir bilgi metriği, başka bir alıcı için iyi olmayabilir. Bu senaryolarda makul bir bilgi metriği, asgari bozulma ilkesine dayanan orijinal verilerle anonim veriler arasındaki "benzerliği" ölçmektedir [56]; Minimum bozulma metriğinde veya MD 'de, genelleştirilen veya bastırılan bir değerin her birine bir ceza uygulanır. Örneğin, 10 Mühendis örneğini profesyonelleştirmek, 10 birim bozulmaya ve bu örneklerin herhangi bir mesleğe daha da genelleştirilmesine neden olur; bu metrik, arama metriği olarak kullanılmaktadır ve tek bir öznitelik için önlemdir.

ILoss, bilgi kaybını yakalamak için spesifik bir değeri genel bir değere genelleme için önerilen bir veri metriğidir [56]:

$ILoss(V_g) = \frac{|v_g|-1}{|D_A|}$ 1, Burada $|v_g|$, $v_g$ özniteliğindeki alan adı değerlerinin sayısıdır. Bu veri metriği, tüm orijinal veri değerlerinin taksonomideki yapraklarda olmasını gerektirir. ILgoss $(v_g) = 0$ vg tablonun orijinal bir veri değeri ise, ILoss $(v_g)$ kelimeleri ile, $(v_g)$ile genelleştirilmiş alan değerlerinin fraksiyonunu ölçer.




**Şekil 2.4:** İş, Cinsiyet, Yaş İçin Taksonomi Ağaçları

Örneğin, Şekil 2.4'de bir örnek Dansçının sanatçıya genelleştirilmesi

$ILoss(Sanatcı) = \frac{2-1}{4}$ = 0.25. Genelleştirilmiş bir kaydın kaybı $r$ tarafından verilir.

$$ILoss(r) = \sum_{v_g \in r} \left(w_i \times ILoss(v_g)\right) \quad \text{(Denklem } \mathbf{2.9}\text{)}$$

$w_i$, $v_g$'nin Ai niteliğinin ceza ağırlığını belirten pozitif bir sabittir. Genelleştirilmiş bir tabloda $T$'nin genel kaybı ise şöyledir:

$$ILoss(T) = \sum_{r \in T} \quad ILoss(r) \quad \text{(Denklem } \mathbf{2.10}\text{)}$$

Hem MD hem de ILoss, bir kaydı diğer kayıtlardan bağımsız olarak genellemek için bir ceza öngörmektedir. Örneğin, 99 Mühendisi ve Avukat'ı genelleştirmek, 50 Mühendisi ve 50 avukat örneğini genelleştirme gibi aynı cezaya sahip olacaktır. Her iki durumda da, 100 örnek ayırt edilemez hale getirilir. Aradaki fark, genellemeden önce ilk örnekte 99 örneğin ayırt edilemez olduğu, ancak ikinci durumda yalnızca 50 örneğin ayırt edilemez olmasıdır. Bu nedenle, ikinci durum, daha orijinal olarak ayırt edilebilir kayıtların ayırt edilemez hale gelmesine neden olmaktadır. Tanınırlık ölçütü olan DM, QID 'e göre diğer kayıtlardan ayırt edilemez olduğu için her kaydın bir cezasını yükleyerek bu kayıp kavramını ele almaktadır. Bir kayıt boyut $s$ grubuna aitse, kayıt cezası $s$ olur. Bu veri metriği, kayıtları QID'e göre ayırt edilemez hale getirmeye çalışan k-anonimliğine tamamen aykırı çalışmaktadır.

Tam etki alanı genelleme şemasındaki minimum derecede anonim bir tablo aramaya rehberlik etmek için, ayırt edici nitelik veya DA olarak adlandırılan basit bir arama metriği kullanmaktadır. Sezgisel veri, genelleme verilerinde en fazla sayıda ayrımsal değeri olan özniteliktir. Bu tür basit sezgisel bilgilerin yalnızca arama yönlendirme amacına hizmet ettiğini, ancak adsız bir tablonun kullanımını nicelleştirmediğini unutmamak gerekmektedir.

### 2.6.2 Özel Amaçlı Bilgi Metriği

Verilerin amacı yayın sırasında biliniyorsa, bilgilerin korunması için anonimleştirme sırasında amaç hesaba katılabilmektedir. Örneğin, veriler tablodaki bir hedef niteliğin sınıflandırılmasını modellemek için yayınlanıyorsa, ayrımları hedef özniteliğindeki sınıf etiketlerini ayırt etmek için gerekli olan değerleri genelleştirmemeniz önemlidir. Sıkça sorulan bir soru, verilerin amacı biliniyorsa, neden veri yerine veri madenciliği sonuçlarını çıkarma ve yayınlama (sınıflandırıcı gibi) yoluna gidilmediğidir? Cevap, veri madenciliği sonuçlarının yayınlanmasının, uzman olmayan veri yayıncısı için pratik olmayan veya veri alıcısı için arzu edilmeyen algoritmik düzeyde bir taahhüt olmasıdır. Uygulamada, veriden belirli bir amaç için çıkarımda bulunmanın birçok yolu bulunmakta ve genellikle veri alınana ve farklı yollar denenene kadar hangisinin en iyisi olduğu bilinmemektedir. Gerçek hayattan bir örnek, Netflix verilerinin (New York Times, 2 Ekim 2006) yayımlanmasıdır. Netflix, katılımcılara belirli bir şekilde onları sınırlandırmak yerine istedikleri analizleri gerçekleştirmede en büyük esnekliği sağlamak istemektedir [66].

Somut olarak ele almak gerekirse, bu çalışmanın amacı yayınlanan verilerdeki eğitim vakaları ile nüfus vakalarını önceden belirlenmiş bazı sınıflandırma sorunlarını ele almaktır. Eğitim vakaları, sınıflandırma modelini artıracak yararlı sınıflandırma bilgilerini hem de sınıflandırma modelini bozabilmektedir. Özellikle yararlı sınıflandırma bilgisi, hedef sınıfları ayırt edebilen ve yalnızca eğitim vakaları için değil, gelecekteki vakalarda da olan bilgileri içermektedir. Aksine, yararsız gürültü yalnızca eğitim vakaları için geçerli olmaktadır. Açıkçası, yalnızca sınıflandırmaya yardımcı olan kullanışlı sınıflandırma bilgileri muhafaza edilmelidir. Örneğin, bir hastanın doğum yılını ele alalım, yaşlı insanlarda hastalık daha sık ortaya çıkıyorsa, akciğer kanserinin sınıflandırılmasına yönelik bilginin bir parçası olabilir, ancak tam doğum tarihi muhtemelen gürültü olacaktır. Bu durumda, doğum tarihini doğum yılına genellemek, gürültüyü ortadan kaldırdığı için sınıflandırmaya yardımcı olur. Bu örnek, tüm genel amaçlı metrikler ve optimum k-anonimleştirmesi tarafından benimsenen verilerde bozulmayı asgariye indirmenin doğru soruna hitap etmediğini göstermektedir.

Sınıflandırma hedefine hitap etmek için, bozulma gelecekteki vakalarda sınıflandırma hatası ile ölçülmelidir. Çoğu senaryoda gelecekteki veriler mevcut olmadığından, en gelişmiş yöntemler, eğitim verileri üzerindeki doğruluğu ölçmeye çalışmaktadır. Bu yöntemler yararlı sınıflandırma bilgisinin farklı öznitelik

kombinasyonları tarafından ele alındığını ileri sürmektedirler. Genelleme ve bastırma bu yararlı "sınıflandırma yapılarının" bir kısmını yok edebilir, ancak yardımcı olması için diğer faydalı yapılar ortaya çıkabilir. Bazı durumlarda, genelleme ve bastırma gürültü kaldırıldığından sınıflandırma doğruluğunu bile geliştirebilir.

Kayıt sınıfının çoğunluk sınıfı olmadığı bir gruba bastırılmış veya genelleştirilmiş her kayıt için bir ceza vermektir. Sezgi, bir grupta çoğunluk olmayan bir sınıfa sahip olan bir kayıt çoğunluk sınıfı olarak sınıflandırılacaktır, bu bir hatadır; çünkü kayıtın orijinali sınıfa katılmamaktadır.

CM, bir veri metriğidir ve bu nedenle, eğitim verisinde değişiklik yapılmasını cezalandırır. Bu, sınıflandırma hedefine tam olarak yönelik değildir; bu da işe yaramayan gürültüyü faydalı sınıflandırma bilgilerine genelleştirerek olur. Sınıflandırma için, daha anlamlı bir yaklaşım, bazı yöntemlere göre "iyi" bir isimsizleştirme araştırması yapmaktır. Başka bir deyişle, bir veri metriğini optimize etmek yerine, bu yaklaşım, aramadaki her adımda anonimleştirme işlemlerini sıralamak için bir arama metriği kullanmaktadır. Bir anonimleştirme işlemi, yararlı sınıflandırma bilgisini koruduğu takdirde yüksek olarak sıralanmaktadır. Arama metriği farklı anonimleştirme algoritmaları ile uyarlanabilir. Örneğin, aç gözlü bir algoritma veya tepeye tırmanma optimizasyon algoritması, belirli bir arama metriği için anonimleştirme operasyonlarının minimal bir sırasını tanımlamak için kullanılabilir.

Ne bir veri metriği, ne de bir arama metriği gelecekteki vakalar için iyi bir sınıflandırmayı garanti edememektedir. Anonim verilerin bir sınıflandırıcı oluşturarak anonimleştirme etkisini deneysel olarak değerlendirmek ve test vakalarında nasıl performans gösterdiğini görmek gerekmektedir. Bu konuda Az sayıda eser bulunmaktadır (Fung ve diğerleri 2005, 2007; Iyengar 2002; LeFevre ve diğerleri 2006b; Wang ve diğerleri 2004], Bayardo ve Agrawal [2005] gibi pek çok kişi sınıflandırma sorununun çözümüne yönelik bir girişimde bulunmak için böyle deneyleri gerçekleştirmişlerdir [39].

### 2.6.3 Takas Amaçlı Bilgi Metriği

Özel amaçlı bilgi ölçümleri, belirli bir veri madenciliği görevi için verilerin yararlılığını korumayı amaçlamaktadır. Maksimum bilgiyi elde eden anonimleştirme işlemi, diğer kimlik doğrulama işleminin gerçekleştirilemeyeceği çok fazla gizliliğin kaybedebilmesi anlamına gelmektedir. Takas metriği fikri anonimleştirme işleminde

hem gizlilik hem de bilgi gereksinimlerini göz önüne almakta ve iki gereklilik arasında en uygun dengeyi belirlemektedir.

Bilgi / gizlilik dengesi ilkesine dayalı bir arama metriği ortaya koyuldu. Anonim bir tablonun, çocuk değerlerine genel bir değeri tekrar tekrar uzmanlaştırarak arandığını varsayalım. Her uzmanlaşma işlemi, genel değeri içeren her grubu, her çocuk değeri için bir grup halinde bölmektedir. Her uzmanlık işlemi, $IG(s)$ olarak belirtilen bazı bilgiler alır ve bazı gizlilikleri, $PL(s)$ kaybeder. Bu arama metriği, en üst düzeye çıkaran uzmanlık alanını tercih etmektedir [39].

$$IGPL(s) = \frac{IG(s)}{PL(s)+1} \quad \text{(Denklem } \mathbf{2.11}\text{)}$$

IG(s) ve PL(s) seçimi bilgi metriğine ve gizlilik modeline bağlıdır. Örneğin, sınıflandırma analizinde, IG(s), genel bir grubun çeşitli uzman gruplara uyarlanmasından sonra sınıf entropisinin azalması olarak tanımlanan bilgi kazanımı olabilir. Alternatif olarak, $IG(s)$ yürütüldükten sonra MD ile ölçülen bozulmanın azalması olabilir. k-anonimleştirme için gizlilik kaybını $PL(s)$ 'nin özniteliğini içeren tüm $QID_j$'de anonimliğin ortalama azalması ile ölçmüştür [39].

$$PL(s) = avg\{A(QID_j) - A_s(QID_j)\} \quad \text{(Denklem } \mathbf{2.12}\text{)}$$

Burada $A(QID_j)$ ve $A_s(QID_j)$'in isimlendirilmesinden önce ve sonra $QID_j$ anonim olduğunu belirtilmektedir. Bir varyasyon, PL(s) sıfıra ayarlayarak bilgi kazanımı en üst düzeye çıkarmaktadır. Maksimum bilgiyi yakalamak için elde edilen uzmanlığın, diğer uzmanlıkların gerçekleştirilemeyeceği gizliliği kaybetmesidir.

Bilgi / gizlilik dengesi ilkesinin bir genelleme $g$ seçmek için de kullanılabileceğini unutulmamalıdır; bu durumda,

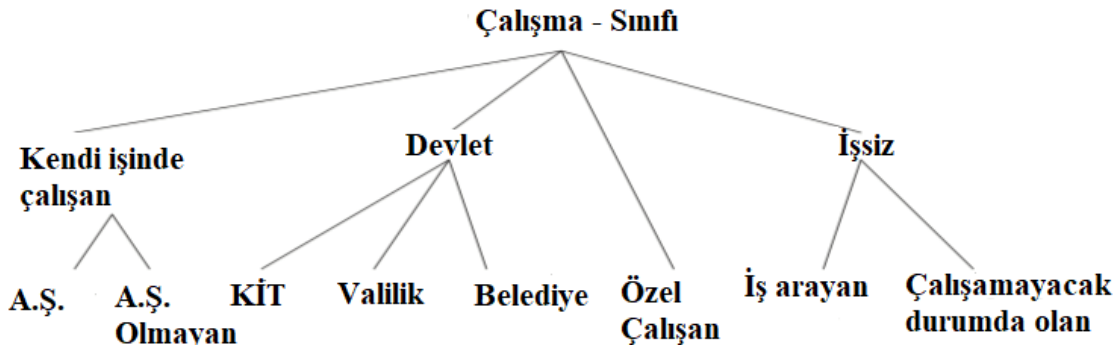
$$IGPG(g) = \frac{IG(g)}{PL(g)+1} \quad \text{(Denklem } \mathbf{2.13}\text{)}$$

burada $IL(g)$ bilgi kaybını, $PG(g)$, $g$'yi oluşturarak gizlilik kazanımını gösterir [39].

### 2.6.4 Tanınabilirlik Metriği (Discernibility Metric)

En genel durum, veriler çoklu kullanımlar için dağıtıldığı zaman ortaya çıkar. Ayrıca, kullanımın dağıtım zamanında bilinmediği durum da bu duruma dâhil edilmelidir. Bu durumda, daha önceki çalışmalarda olduğu gibi bazı bilgi kaybı kavramlarını yakalayan bir metrik tanımlanmalıdır. Metriğimiz, önceki çalışmalarımızda izin verilenden daha esnek genellemeleri ele almalıdır.

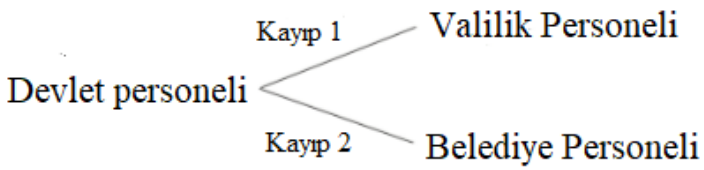
Potansiyel olarak tanımlayıcı tüm sütunlarda yer alan bilgilerin bu durumda da aynı derecede önemli olduğu varsayılmaktadır. Genelleştirmeler ve bastırmalara (genel kayıp metriği LM ile gösterilir) bağlı toplam bilgi kaybı, bu sütunların her biri için normalleştirilmiş bir bilgi kaybı toplamak suretiyle hesaplanmaktadır. Her bir sütun türüne ait kayıp hesaplaması bir sonraki olarak tanımlanmalıdır. Bir sütun için bir bilginin kaybı, sütundaki her giriş için ortalama kayıp olarak hesaplanmalıdır.




**Şekil 2.5:** Çalışma Türü Hakkında Bilgi İçeren Taksonomi Ağacı

Bir taksonomi ağacını ($T$) genelleştirilmiş kategorik bilgileri içeren potansiyel olarak tanımlayıcı bir sütun olarak düşünelim (Şekil 2.5). Dönüşüm yapılan tablodaki genelleştirilmiş değerin, $T$ taksonomisi ağacında $P$ düğümüne (ör. Devlet) karşılık geldiğini ve bu sütundaki bir girişi düşünün (ör. Belediye, Devlet). Bir yaklaşım, bir yaprak düğümü değeri ayrıştırılamadığında zararı nicelikselleştirmektedir. Genellemeye bağlı olarak başka bir değerden. Örneğin, 'Belediye' değerini 'Devlet' değerine yaygınlaştırmak, Devlet çalışanın ve Belediye çalışanın değerlerinin birbirinden ayrılamayacağını ifade etmektedir. İlişkili kayıp, Şekil 2.5'de gösterildiği gibi modellenebilir. Denemelerde, iki genel kategorik değer arasındaki belirsizlik için aynı genelleme kaybını varsayarak modeli basitleştirilebilir. $T$'deki yaprak düğümlerinin toplam sayısı, $M$ ile gösterilsin. Alt ağaçtaki düğüm $P$ köklü yaprak düğümlerinin sayısı $M_p$ olsun. Bu basitleştirilmiş modeli kullanarak ve genelleştirilmiş

düğümün taksonomi ağacının kökü olduğu durumda en kötü durumu kullanarak normalleştirme, bu girişin kaybı olarak ($M_p$ - 1) / (M - 1)'e yol açar. Daha önceki örnekte, Belediyenin Devlete genelleştirildiğinde normalleştirilmiş kayıp 2/7 'dir. Bastırılmış bir girişin kaybı, genelleştirilmiş değer ağacın köküne karşılık geldiğinde ortaya çıkan kayıp ile aynıdır [39].




**Şekil 2.6:** Kategorik Özelliklerin Genelleştirilmesine Bağlı Kayıp.

Örneğin, bir üretici, belirli bir ürün kategorisiyle ilgilenen müşterileri modellemekle ilgilenebilir. Bir perakendeci tarafından toplanan müşteri profili ve işlem verileri, bu tür modeller oluşturmak için kullanılabilir. Kilit bir soru, kimlik ifşasına ilişkin kısıtlamaları yerine getirirken bu modelleri doğru bir şekilde oluşturup oluşturamayacağımızdır. Modelin doğruluğu, genellemeler ve bastırmalar nedeniyle bilgi kaybına bağlıdır. Genellemelerden kaynaklanan toplama nedeniyle hedef değişkenlerdeki (tahmini modelleme için) saflığın kaybını ölçen metrikleri tanımlanabilir ve böylece spesifik prediktif modelleme yöntemine uyarlama yapılmadığı için, muhafazakâr bir yaklaşım olarak düşünülebilir.

### 2.6.5 Bilgi Kaybı Metriği (Classification Metric)

Potansiyel olarak tanımlayan sütunlarda içeriğin genelleştirilmesi veya bastırılması, bu sütunları kullanarak sınıfların ayrımını zayıflatmaktadır. k-anonimlik kısıtı tablodaki çoklu satırları (en azından k), potansiyel olarak tanımlayan sütunlar için genelleştirilmiş değerlerin aynı kombinasyonunu zorlamaktadır. Benzersiz bir genel değer kombinasyonuna sahip tüm satırların aynı gruba ait olduğu söylenebilir. Her satır r için, $G(r)$ ait olduğu grubu belirtmektedir. Farklı sınıf etiketlerine sahip bir $G$ grubundaki satırların, potansiyel olarak tanımlayıcı sütunlar kullanılarak ayrımı yapılamamaktadır. Bu nedenle, doğru sınıflandırma için, $G$'deki tüm satırların aynı sınıf etiketine sahip olması tercih edilmektedir. Şimdilik, tanımlayıcı olmayan sütunlardaki bilgilerin $G$'deki satırları farklı etiketlerle ayırt edebileceğini göz ardı

edilecektir. Dolayısıyla, bu kullanım metriği, farklı etiketli satırlar içeren saf olmayan gruplar cezalandırılacaktır.

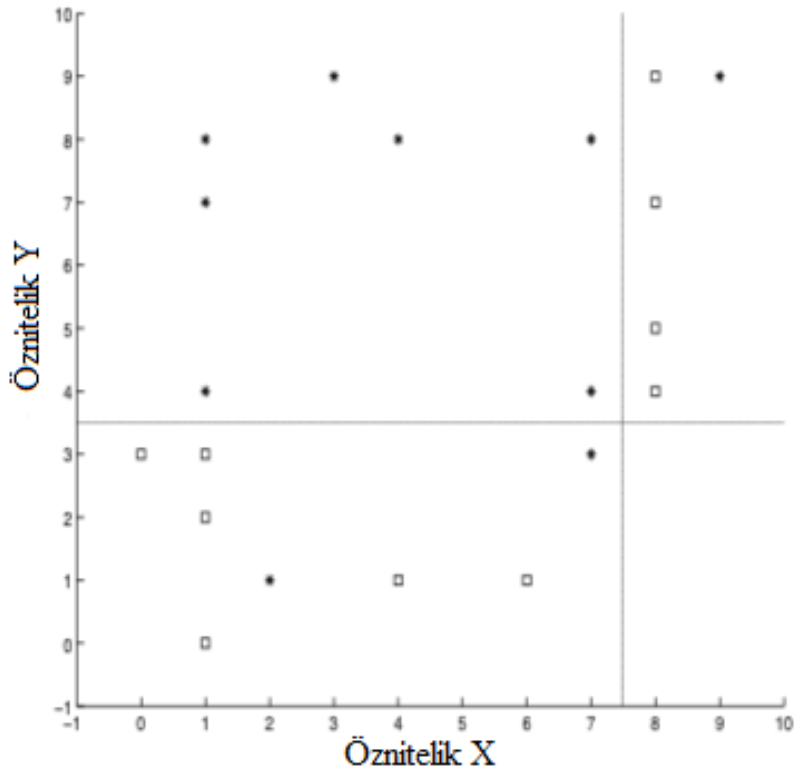
Sınıflama metriği CM, Denklem 1'de, toplam satır sayısı N ile normalize edilen tablodaki her satır için ayrı cezaların toplamı olarak tanımlanmaktadır.

$$CM = \frac{\sum_{all\ rows}(row\ r)}{N} \qquad \text{(Denklem } \mathbf{2.14}\text{)}$$

Bir satır $r$ bastırılırsa veya sınıf etiketi sınıfı $(r)$, $G$ grubunun çoğunluk sınıfı etiketi çoğunluğu $(G)$ değilse cezalandırılır [56].

$$penalty(row\ r) = \begin{cases} 1 \quad if\ r\ suppressed \\ 1 \quad if\ class(r \neq) majorty(G(r)) \\ 0 \quad otherwise \end{cases} \qquad \text{(Denklem } \mathbf{2.15}\text{)}$$

Şekil 2.7'de sayısal değerlerle (0 ila 10 arasında) potansiyel olarak tanımlayıcı iki özniteliğe $(X, Y)$ sahip bir örnek kullanılarak gösterilmiştir. Sabit çizgiler { [0,7.5], (7.5,10]}, öznitelik $X$ için genelleme ve { [0,3.5], (3.5,10}}, öznitelik $Y$ için genellemenin olduğu bir çözümdür. Bu çözümü ele alalım Belirtilen k-isimsizlik kısıtı $k = 5$. olduğunda hiçbir satır bastırılmamalıdır $X = 9$, $Y = 9$ olan satır, sınıf etiketi $*$ kendi grubundaki çoğunluk etiketi 2 olmadığından 1 cezaya katkıda bulunur ($X\varepsilon\ (7.5, 10]$ ve $Y\ \varepsilon\ (3.5, 10)$ ile tanımlanan grup) CM metriği, bu durumda 0,15 değerindedir, çünkü 20 noktadan 3'ü gruplarında çoğunluk sınıfına sahip değildir.




**Şekil 2.7:** Sayısal Özniteliklere Sahip X ve Y Grafiği

Sınıflandırma metriği, Sınıf C1'e sahip bir veri noktasını sınıf C2'ye sahip olarak yanlış sınıflandırmanın maliyetini gösteren bir maliyet matrisini içerecek şekilde genişletilebilir. Bu ceza fonksiyonunu orijinal sınıftan bir grubun çoğunluğuna sınıflandırmak için bir satır $r$ 'yi yanlış sınıflandırmanın maliyetini yansıtacak şekilde değiştirerek yapılmaktadır. Bastırılmış satırlar bu durumda ayrı bir grup olarak da ele alınabilir.

Sonra, tablonun sütunlarından biri olan $V$ bağımlı değişkeni için bir regresyon modeli oluşturmak için dönüştürülmüş tabloyu kullanılmış olsun. Bu modele girdi olarak izin verilen sütunlar belirtilir ve potansiyel olarak tanımlayıcı sütunları içerebilir.

Potansiyel olarak tanımlayıcı sütunlar için benzersiz değer kombinasyonlarını temel alan bir grubun önceki kavramını kullanarak, bağımlı değişkenin modellemesi, gruplar içindeki değerinin değişkenliği (saflığı) tarafından etkilenmektedir. Her grubun saflığı herhangi bir dağılım ölçeği ile nicelleştirilebilir. Özellikle, varyans gibi klasik bir ölçüt veya medyanın mutlak sapması gibi sağlam bir önlem, iyi bir çözüm olabilir. Regresyon metriği, tüm satırlar tek bir grupta olduğunda dispersiyonla normalleştirilir.

Tahmini modelleme için birden fazla hedef değişken varsa, normalleştirilmiş metrikleri, her hedef için kullanıcı tarafından belirlenen bir ağırlık temel alınarak, ağırlıklı toplam kullanılarak birleştirilebilir.

### 2.6.6 Ayırt Edilebilirlik Metriği (Discernibility Metric)

Daha önce bahsedilen eserlerden her biri, istenilen anonimleştirmeleri modellemek için kendi benzersiz metriklerini sunamaktadır. Kullanılan algoritmanın büyük kısmı metrik agnostiktir ve basit bir maliyet sınırlayıcı fonksiyon tanımlayarak belirli bir metriği kullanmanın örneklendiği şekilde önerilmektedir. Sunuyu somut tutmak için maliyet sınırlamasının iki örnek metrikle nasıl gerçekleştirildiği gösterilmelidir. Ele alınan iki yaklaşımın önemli özelliği neredeyse tüm değişkenleri olan diğer ölçütlere kolayca adapte edilebilir olmalarıdır.

Maliyet metrikleri, tipik olarak uygulanan bastırma veya genellemeler sonucunda oluşan bilgi kaybını ölçmektedir. Her anonimleştirme, dönüşüm veya bastırma ile ilgili bilgi kaybını yansıtan her bir grupta bir "ceza" meydana getirmek olarak düşünülebilir. Doğal olarak, maliyet metriklerinin çoğu, gruplanmış bir tuple'yi, en azından tuple'nin genellemesine kadar cezalandırmaktadır.

**Tablo 2.18:** Basit 3 Nitelikli / 9 Değerli Tablo İçin Değer Alanları Örneği [46]

| **Yaş** | | | **Cinsiyet** | | **Medeni Hal** | | | |
|---|---|---|---|---|---|---|---|---|
| [10-29] | [30-39] | [40-49] | E | K | Evli | Dul | Boşanmış | Hiç Evlenmemiş |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |

Toplam sipariş verildiğinde, değerler sipariş boyunca konumlarına göre tanımlanabilir. Her bir öznitelik alanından en az olan değer (* ile işaretlenmiştir) herhangi bir geçerli genellemede görülmelidir ve dolayısıyla örtük olarak ele alınabilir.

Kullandığımız ilk metrik, belirli bir k ayarıyla izin verildiği ölçüde, tuple'ler arasında ayırdedilebilirliği korumak için arzuyu basit bir şekilde yakalamaya çalışan metriktir. Bu belirlenebilirlik metriği, dönüşümlü veri kümesindeki kaç tuple'nin kendisinden ayırt edilemeyeceğine dayanarak her bir kümeye bir ceza atar. Bastırılmamış bir grup, boyut $j$'nin indüklenmiş eşdeğerlik sınıfına girerse, o gruba $j$'nin bir cezası verilir. Eğer bir grup bastırılırsa, girilen veri kümesinin büyüklüğü olan $D$'ye bir ceza atanır. Bu ceza, bastırılmış bir kümenin, veri kümesindeki başka herhangi

bir kümeden ayırt edilememesi gerçeğini yansıtmaktadır. E kümeleri, G anonimleştirmesiyle indüklenen D'de tuple'lerin denklik sınıflarını ifade eder (Bir bastırma modeli(Modeling Tuple Suppression)) [39].

Kullanılan bir diğer ilginç maliyet metriği ilk olarak Iyengar tarafından önerilmiştir. Bu metrik, tuplelere kategorik bir sınıf etiketi atandığında, uygulanan denklik sınıfları sınıf etiketine göre tek biçimli olan türlerden oluşan anonimleştirmeler üretmek için uygulanabilir. Bu sınıflandırma metriği, indüklenen eşdeğerlik sınıfı içinde çoğunluk sınıfına aitse, bastırılmamış bir gruba ceza atamaz. Diğer tüm tuple'ler 1 değerine cezalandırılır. Aşağıdaki gibi ifade edilir [46].

$$C_{DM}(g,k) = \sum_{\forall E s.t. |E| \geq k} |E|^2 + \sum_{\forall E s.t. |E| < k} |D||E| \qquad \text{(Denklem } \mathbf{2.17}\text{)}$$

Yukarıdaki deyim içerisindeki azınlık fonksiyonu, bir dizi sınıf etiketli tuple kabul etmektedir ve o gruba göre herhangi bir azınlık sınıfına ait olan tuple alt kümesini döndürmektedir. Daha önce olduğu gibi, bu ifadeden elde edilen ilk miktar bastırılmamış olan tuple'leri cezalandırır ve ikincisi ise tuple'leri bastırır. Iyengar, bu bilinçli metriğin, sınıf gözetlenemeyen metriklerden daha iyi sınıflandırma modelleri üreten anonimleştirilmiş veri kümeleri ürettiğini göstermiştir.

Bir kümeyi bastırmak oldukça ciddi bir işlem olduğundan, izin verilen bastırma sayısına sert bir sınır getirmek istenebilir. Bu, her iki metrik için ifadede, bastırılmış tuple'lerin sayısı (k'den daha küçük olan eşdeğer sınıf boyutlarının toplamı) bu sınırı aştığında, sonsuz bir maliyet getiren bir koşul ekleyerek modellenebilir..

## 2.7 Metrik

Tanımlama algoritmalarının pratikte uygulanması, veriyi yeniden tanımlama olasılığının ölçülebilmesini gerektirir. Böyle bir ölçüm, gözetim görevlisine, yeniden tanımlama olasılığının yüksek olup olmadığını bildirmektedir. Olasılık yüksekse, kimlik tespit yöntemlerinin uygulanması gerekmektedir. Bu, yeniden tanımlama olasılığını ölçmek için belirli ölçütlerin yanı sıra değerlerini yorumlama yönergelerine de ihtiyaç olduğu anlamına gelmektedir. Bu bölümde, kimlik belirlemek için yeniden tanımlama olasılığını ölçen ve yorumlayan bir dizi metrik ve karar kuralı sunulmaktadır.

### 2.7.1 Basit ve Türemiş Metrikler

Bir veri kümesi için yeniden tanımlama riskini değerlendirdiğimizde, o veri kümesindeki her kaydın başarılı bir şekilde yeniden tanımlanmasına ilişkin bir olasılık tayin edilmektedir. Kimlik ifşası için yeniden tanımlama olasılığı, o kayda doğru bir kimlik atanma olasılığı anlamına gelmektedir. $\theta_i = 1, \dots, n$ ve $n$, veri kümesindeki toplam kayıt sayısı olduğu yerde $\theta_i$ tarafından doğru olarak yeniden tanımlanabilen bir kayıt olasılığı belirtilmek zorundadır. Buna dayanarak, bir takım türetilmiş metrikler geliştirilebilir.

Açıklanan veri kümesindeki eşdeğerlik kümeleri $J$ olsun ve $|J|$ veri kümesindeki eşdeğerlik sınıflarının sayısı olsun. Uygulamada, aynı denklik sınıfındaki tüm kayıtların aynı olasılık değeri, $\theta_i$ olacaktır. Dolayısıyla, eşdeğerlik sınıfı için $j$ $j \in J$ olduğu olasılık $\theta_j$'ye yaklaşır [46].

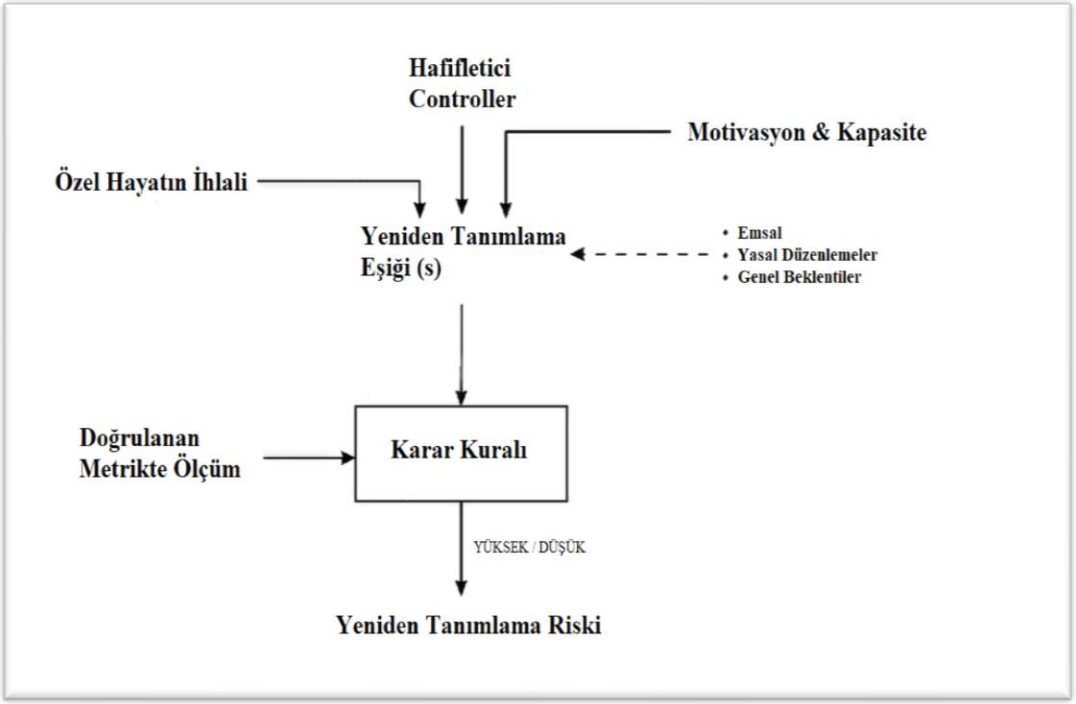
Türetilen metriklerin tümü, bu basit bireysel eşdeğerlik sınıfına ait veri setine uygulanmaktadır. Tutarlılık için, elde edilen metriklerin tümü 0 ile 1 arasında bir değere sahip olacak şekilde ölçeklendirilir.

Olasılığın çok yüksek olup olmadığına bakmak için türemiş metriklerin ikili bir değere dönüştürülmesi gerekmektedir. Ölçümün sonunda elde edilen karar ikili bir sonuç olarak ortaya çıkmaktadır. Metrik değeri yorumlamak için verilen karar kuralları bunlardır.

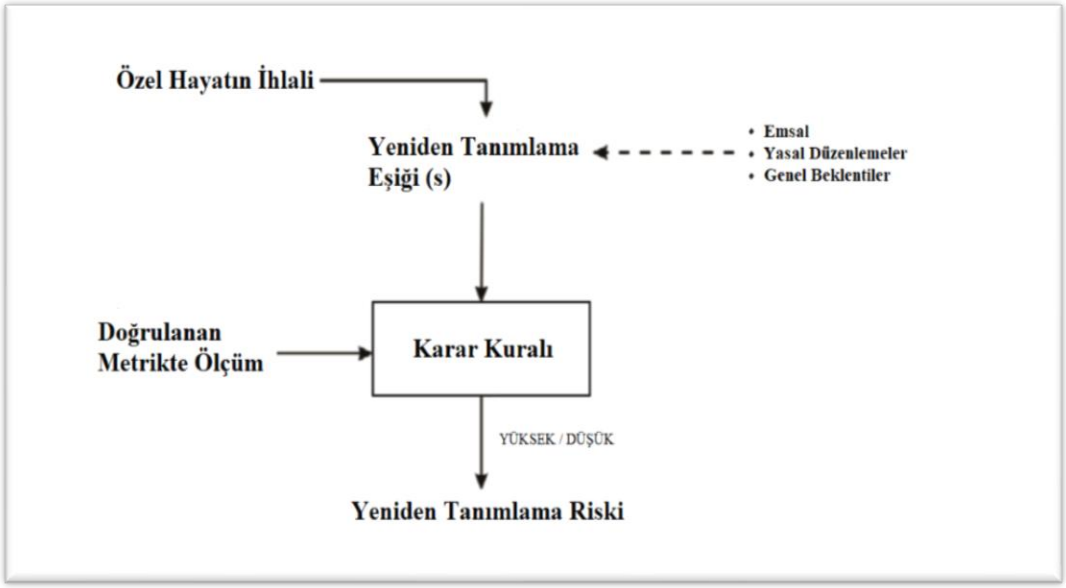
Karar kuralı, türetilen metrikteki ölçümleri yeniden tanımlama riskine dönüştürmektedir. Dolayısıyla, karar kuralının riskin yüksek olduğunu belirlemesi durumunda, düşük tanıma yöntemlerini düşük bir değere getirmek için gerekli olmaktadır. Karar kuralı, riskin düşük olduğunu tespit ederse, kimliği belirlemeye gerek bulunmamaktadır. Bu eşik kullanımı ile sağlanmaktadır. Eşikler, türetilen metriğin kabul edilebilir olup olmadığını belirlemektedir. Eşiğe karar verme, risk yönetimindeki öznel bir bileşendir.

DF üzerinde erişim kısıtlamaları varsa, veri saklama sorumlusu üç risk unsurunu dikkate almalıdır, bunlar: kontrolleri azaltmak, mahremiyete girmek ve amaçları ve kapasiteyi belirlemek şeklinde sıralanabilir. Bu, Şekil 2.8'da gösterilmiştir. Ayrıca, daha önce de belirtildiği gibi, (1) aynı DF için eşiklerin emsalleri, (2) düzenleyici gereklilikler ve sinyaller (3) makul eşikler hakkında halkın beklentileri kabul edilebilir eşikler açısından eşik üzerinde karar verirken dikkate alınması gereken faktörler olarak sıralanabilir.

Erişim kısıtlamaları yoksa, Şekil 2.9'de gösterildiği gibi yalnızca mahremiyet konusunun ihlali düşünülmelidir.




**Şekil 2.8:** Türemiş Metriklerin ve Karar Kurallarının Yeniden Tanımlama Riskini Nasıl Belirlediğini Gösteren Şema: Erişim Kısıtlamalı DF




**Şekil 2.9:** Türemiş Metriklerin Ve Karar Kurallarının Yeniden Tanımlama Riskini Nasıl Belirlediğini Gösteren Şema: Erişim Kısıtlamaları Olmaksızın DF

İlk türetilen metrik, bir tekrar tespit olasılığı, eşiğin üzerinde olan kayıtların oranını değerlendirmektedir. Eşik $\tau$ ile gösterilir [39]:

$$R_a = \frac{1}{n}\sum_{j\in J} f_j \times I(\theta_j > \tau$$ (Denklem **2.18**)

burada $I(-)$ gösterge fonksiyonudur (parametreler doğruysa 1, aksi halde 0 döndürür) ve $f_j$veritabanında eşdeğerlik sınıfı $j$'nin büyüklüğüdür. İlerleyen paragraflarda $\tau$ değerinin nasıl ayarlanacağı tartışılacaktır. $R_a$'nın değeri çok yüksekse, veri kümesinin kabul edilebilir olmayan yeniden tanımlama riski bulunduğu kabul edilmektedir.

Yeniden tanımlama olasılığı yüksek olan kayıtların maksimum oranı eşiğini $\alpha$ ile göstereceğiz. Dolayısıyla $R_a$ 'nın karar kuralı aşağıdaki gibidir:

$$D_a = \begin{cases} HIGH, R_a > \alpha \\ LOV, R_a \leq \alpha \end{cases}$$ (Denklem **2.19**)

$R_a$ değeri eşikten yüksekse yeniden tanımlama riski yüksek kabul edilmektedir. Dikkat edileceği üzere, $D_a$ karar kuralında, $\tau$ ve $\alpha$ işleyişi için birden fazla eşik gerekmektedir.

Türemiş metriklerden bir diğeri de en kötü durum senaryosunu ele almakta ve en yüksek yeniden tanımlama olasılığına sahip eşdeğerlik sınıfının tüm veri setini temsil ettiği varsayılmaktadır.

$$R_b = max_{j\in J}(\theta_j)$$ (Denklem **2.20**)

Bu oldukça sıkı bir standarttır, çünkü yeniden tanımlama olasılığı düşük olan kayıtlar bile, yeniden tanımlama olasılığı yüksek kayıtlar kadar cezalandırılır. Örneğin, $\theta_i$ için yüksek bir değere sahip bir milyon kayıt veri setinde tek bir kayıt olsa bile, $R_b$ değeri bu değeri alacaktır. Bu metriğin uygun olacağı durumlar, (1) veri saklama sorumlusu oldukça muhafazakâr davranmayı ertelemek istediğinde ve (2) düşmanın zeki olduğunu varsayılırsa ve dikkatini en çok tekrar eden kayıtlara odakladığı durumlar olacaktır. Tanımlama olasılığının yüksek olduğu durum ikinci durumun yapılması en mantıklı olduğu durumlardır ve bu nedenle $R_b$'nin kullanılması uygun olur. Bununla birlikte, bir düşmanın en yüksek riskli kayıtları hedef alabileceğini varsaymak mantıklı değilse, o zaman $R_b$'nin uygun olup olmadığı tartışmaya açık bulunmaktadır. $R_b$ için karar kuralı

$$D_b = \begin{cases} HIGH, R_b > \tau \\ LOV, R_b \leq \tau \end{cases}$$ (Denklem **2.21**)

Bu karar kuralının eşiğinin yukarıdaki Eşitlikte görüldüğü gibi olmasına dikkat edilmelidir; çünkü bu eşik ile aynı anlama gelir: bir kaydın yeniden belirlenebileceği en yüksek olasılıktır.

Bir diğer türetilmiş metrik, veri kümesindeki tüm kayıtların ortalama olasılığını almaktadır. Aslında beklenen değerdir. Bu beklenen değer aşağıdaki gibi ifade edilir:

$$R_c = \frac{1}{n}\sum_{j \in J} f_j \, \theta_j$$ (Denklem **2.22**)

Bu türden türetilmiş metriğin karar kuralı,

$$D_c = \begin{cases} HIGH, R_c > \lambda \\ LOV, R_c \leq \lambda \end{cases}$$ (Denklem **2.23**)

Burada, λ, yeniden tespit edilebilen kayıtların maksimum oranını temsil etmektedir.

$R_c$ metriği, bir düşman tarafından yeniden tanımlanacak kayıtların oranını vermektedir (ortalama olarak). Yüzeysel olarak, yeniden tanımlama olasılığı yüksek olan kayıtların oranını ölçen $R_a$ metriğine benzemektedir. Bununla birlikte, $R_a$ 'da açıklanan kayıtlar mutlaka yeniden tanımlanacak kayıtlar olmayabilir.

Örneğin, bir veri kümesinin 100 kayıt olduğunu söylersek, $\tau = 0.2$, $\alpha = 0$, $\lambda = 0.2$ ve sadece iki kayıtta tekrar tanımlama θj = 0.3 olasılığı bulunmaktadır, geri kalan kısmı $\theta_j$ = 0.1 olur. $R_a$ metriğini kullandığımızda, düşmanın tek bir kaydı tekrar belirlemeye çalışacağını ve yeniden tanımlamak için $\theta_j > \tau$ ile bu iki kaydın birinin seçebileceği varsayılmaktadır. Böyle bir durumda, risk kabul edilmemelidir, çünkü seçildiyse eğer, iki kayıt tekrar belirlenme olasılığına sahip olmaktadır. Öte yandan, $R_c$ metriği kullanıldığında saldırgan, bazı kayıt defteriyle eşleştirerek veri kümesindeki tüm kayıtları yeniden belirlemeye çalışacak ve muhtemelen on doğru eşleşmeye ulaşabilecektir. Böylesi bir durumda ulaşılan eşleşmeler $\lambda$ 'dan daha küçük bir oran ifade ettiği için bu, kabul edilebilir bir risk olarak algılanabilecektir.

Yeniden tanımlanan risk ölçütleri tartışılırken, ne tür bir metrik kullanımından bahsedildiği ve kullanılan karar kuralından emin olmak önemli olmaktadır. Kullanılan

eşikleri açıklığa kavuşturmamız gerekmektedir; Sonuçlar doğru yorumlanmalıdır. Bu sonuçları yorumlamada, $\tau = 0.2$ ve $\tau = 0.05$ arasında büyük bir fark olacaktır.

Türetilmiş metriklerin bir özeti, Tablo 2.20'deki karar kurallarına göre Tablo 2.19'da verilmiştir ve eşiklerin bir özeti ve bunların yorumları ise Tablo 2.21'de verilmektedir.

**Tablo 2.19:** Türemiş Yeniden Tanımlama Metriklerinin Özeti [39]

| Türemiş Risk Metriği | Yorumlama |
|---|---|
| $R_a = \frac{1}{n}\sum_{j \in J} f_j \times I(\theta_j > \tau$ | Yeniden tanımlama olasılığı bir eşiğin üzerinde olan kayıtların oranı |
| $R_b = max_{j \in J}(\theta_j)$ | Tüm kayıtlar arasında veri kümesindeki yeniden tanımlamanın maksimum olasılığı |
| $\mathrm{R_c} = \frac{1}{\mathrm{n}}\sum_{\mathrm{j \in J}} \mathrm{f_j}\,\theta_\mathrm{j}$ | Ortalama olarak doğru yeniden tanımlanabilen kayıtların oranı |

**Tablo 2.20:** Karar Kurallarının Özeti [39]

| Türemiş Risk Metriği | Yorumlama |
|---|---|
| $D_a = \begin{cases} HIGH, R_a > \alpha \\ LOV, R_a \leq \alpha \end{cases}$ | Yeniden tanımlama olasılığı yüksek kayıtların, Kabul edilebilir kayıtlara oranı. |
| $D_b = \begin{cases} HIGH, R_b > \tau \\ LOV, R_b \leq \tau \end{cases}$ | Tüm kayıtlar arasında en olası yeniden tanımlamana bilen kayıtların, kabul edilebilir kayıtlara oranı. |
| $D_c = \begin{cases} HIGH, R_c > \lambda \\ LOV, R_c \leq \lambda \end{cases}$ | Yeniden tespit edilebilen ortalama kayıtların kabil edilebilir kayıtlara oranı. |

**Tablo 2.21**: Eşiklerin Özeti [39]

| Eşik | Yorumlama |
|---|---|
| $\tau$ | Tek bir kaydı en yüksek doğru bir şekilde yeniden tanımlama olasılığı. |
| $\alpha$ | Veri saklama alanının, kabul edebileceği tekrar tanımlama olasılığının yüksek olduğu kayıtların oranı |
| $\lambda$ | Veri saklamacısı için kabul edilebilir doğru şekilde yeniden tanımlanabilen kayıtların ortalama oranı |

### 2.7.2 Risk Metriği

İki basit yeniden tanımlama metriği veya daha doğrusu, $\theta_j$ 'nin iki örneği bulunmaktadır. Onları farklı kılan başlıca kriter, saldırganın belli bir şahsın DF'de olup olmadığını bilmesidir. Bu birey, yeniden tanımlanan kişidir ve hedef olarak

belirlenecektir. Hedef, saldırganın zaten hakkında arka plan bilgisine sahip olduğu belirli bir kişi olabilir. Örneğin, bu düşmanın komşusu, iş arkadaşı, eski eş veya ünlü bir kişi olabilir. Bu kişi düşmanın tanıdığı gibi görülebilmelidir. Ya da hedef, seçmen kayıt defteri gibi bir nüfus listesinden seçilmiş bir birey olabilir. Örneğin, saldırgan bir veri saklayıcısını ambargo altına almak ya da ortaya çıkarmak isteyen bir gazeteci ise, gazeteci herhangi bir kaydın yeniden tanımlanması amacına ulaşacağından rastgele bir hedef seçebilmektedir.

Hedefin $DF$ 'de olup olmadığını saldırgan biliyorsa buna savcı riski denmektedir. Hedefin $DF$ 'de olup olmadığını bilmiyorsa, buna gazeteci riski denmektedir [39].

Hedefin, $DF$ 'de olup olmadığını saldırgan nasıl bilebilir? Üç koşuldan herhangi biri doğruysa, veri saklama sorumlusunun savcı riski ile ilgilenmesi gerekmektedir; $DF$ bütün nüfusu temsil etmektedir (örn. Bir nüfus kayıt defteri) veya bu nüfustan geniş bir örneklem grubu. Eğer bütün nüfus açıklanırsa, saldırgan hedefin $DF$ 'de olduğunu kesin olarak bilecektir. Tersi de doğru olur. Saldırgan bir nüfus kayıt defterinde bir tanıdık veya rasgele bir kimliği yeniden belirlemeye çalışsın. Ayrıca, büyük bir örneklem grubu, hedefin $DF$ 'de olmasının muhtemel olduğu anlamına gelmektedir. Nüfus kayıtlarının örnekleri, eyalet ya da devlet hastalık kayıtları ya da doğumlar ya da hastane boşalmaları gibi belirli olaylarla ilgili bilgi toplayan kayıtlardır.

$DF$ bir nüfus kayıt defteri değildir, bir popülasyondan alınan bir örnektir ve $DF$ 'de kimin olduğu kolayca tespit edilebilmektedir. Örneğin, $DF$, örnek bir uyuşturucudan alınan veri kümesiyse, gençlerin anketini kullanmaktadır; bu durumda bir ebeveyn, büyük bir olasılıkla, çocuğunun katılmasına izin vermesi nedeniyle onun katıldığını bilmektedir.

$DF$ bir örnektir ve $DF$ 'deki bireyler örneklemin bir parçası olduklarını kendileri belirmektedir. Örneğin, bir klinike ait deneme veri setinin kamuya açıklandığı varsayılsın. Klinik araştırmalarında yer alanlar genellikle ailelerine, arkadaşlarına ve hatta tanıdıklarına bir denemeye katıldığını bildirir. Bir tanıdık, DF 'deki bu kendini açığa çıkaran katılımcıların verilerinden birini yeniden belirleme girişiminde bulunabilir. Bireyler, bir çalışma veya hastalık kayıt defterinin bir parçası olduklarını ortaya koyabilecek bilgileri bloglar veya sosyal paylaşım sitesi sayfaları aracılığıyla kendileri verebilirler. Bununla birlikte, her zaman bireylerin kayıtlarının bir veri kümesinde olduğunu bilmeleri mümkün olmamaktadır. Örneğin, rıza alınmadığında veya bireylerin gelecek araştırmalarda kullanılacak olan verilere veya doku örneğinin

bir izine tabi olduğu hallerde mevcut verileri kullanan çalışmalar için bireyler, kayıtlarının belirli bir DF'de olduğunu bilmemektedir.

Bir veri seti yukarıdaki ölçütleri karşılamıyorsa, veri saklama sorumlusu gazeteci veya savcı riski hakkında endişe etmemelidir (yani, biri ya da ikisi, ikisi birden değil). İki risk türü arasındaki ayrım oldukça önemlidir, çünkü risk ölçme veya tahmin edilme biçimi açısından farklılık göstermektedirler.

Veri sahipleri belirli bir veri türüne ilişkin olarak savcı veya gazeteci riski öngördüklerinde bu tür ölçümleri uygulayabilmektedirler.

Burada nüfus terimi mutlaka bir coğrafi bölgenin veya belli özellikleri olan grupları kapsamamaktadır. Nüfustan kastedilen şey veri kümesidir. Örneğin kanseri olan herkes kayıt defterinde olacağından bu popülasyona dahil olacaktır. Belirli bir hastalığı coğrafi bir sınır dahilinde olan veya belirli bir demografik (örn., Etnik köken, evde konuşulan dil veya yaş grubu) olan tüm hastaların bir veri kümesi bir nüfus olarak kabul edilmektedir.

### 2.7.3 Savcı ve Gazeteci Riski Örnekleri

Aşağıdakiler, her bir risk türünün ne zaman uygulandığını göstermek için gerçek örneklerdir [56]:

BORN olarak bilinen Ontario doğum kaydı, yeni anne ve bebeklerin araştırma ve halk sağlığı araştırmaları da dâhil olmak üzere belirli amaçlar için sağladığı veri setlerini içermektedir. Normal olarak, tüm nüfus verileri seti veya belirli ölçütleri (örneğin, doğum yeri veya doğum periyodu) karşılayan herkes açıklanmaktadır. Kayıt defteri bütün popülasyonu (diğer bir deyişle eyaletteki tüm doğumları) içerdiğinden, bir saldırgan, herhangi bir annenin kayıt defterinden açıklanan veri kümesinde olacağını kesin olarak bilmektedir. Bu durumda savcılık riski geçerli olur. Bununla birlikte, kayıt defterinden rastgele bir örnek açıklanırsa, o zaman gazeteci riski söz konusu olmaktadır. Bunun nedeni, saldırganın, belirli bir anne / bebeğin bu numuneye dâhil edilip edilmediğini bilmemesidir.

Bir hastanenin acil servisine gelen hastalar hakkındaki veriler açıklanırsa, savcı riski söz konusudur. Bunun nedeni, saldırganın belirli bir kişinin acil servise gittiği konusunda bilgisi olursa, düşman da hastanın verisinin bu veri setinde olduğunu bilmesidir. Bununla birlikte, yalnızca influenza (Bir tür Grip, Viral enfeksiyon) benzeri bir hastalık şikayetiyle gelen hastalar hakkındaki veriler açıklanıyorsa, düşman

hedef bireyin influenzaya sahip olduğunu bilmediği takdirde gazeteci riski söz konusu olurdu. Böyle bir durumda, düşman hedef kitlenin veri kümesinde olup olmadığını anlayamamaktadır.

Bir devlet sağlık sigortacısı, sağlık bilgilerini bir DF olarak kamuya açık hale getirmeye karar vererek hastaların %10'unun rassal bir örneği hakkındaki verileri DF'ye dâhil etmiştir. Bu durumda gazeteci riski geçerlidir, çünkü bir saldırgan bu rastgele örneklemin içinde kim olduğunu bilmemektedir ve hastalar bu numunede olup olmadıklarını kendileri bile bilmemektedirler.

### 2.7.4 Savcı Riskinin Ölçülmesi

Saldırganın, belirli bir hedefi, Ayşe'yi tekrar belirlemeye teşebbüs ettiğini varsayalım. Ayrıca, Ayşe'nin $DF$'de olduğunu da biliyor olsun bu nedenle savcı riski söz konusu olmaktadır.

Şekil 2.10'da gösterilen örnekte, hasta bilgilerini ve bazı reçeteli bilgileri (DIN-ilaç tanımlama numarası) içeren orijinal bir veri seti bulunmaktadır. Doğrudan tanımlayan değişken olan hasta adı bastırılmış. Tanımlama, doğum yılını genelleştirerek uygulanmıştır. Artık bir $DF$ var. Saldırgan Ayşe hakkında bazı bilgileri, yani doğum yılı ve cinsiyeti bilmektedir. Saldırgan ayrıca, Ayşe'nin $DF$'de olduğunu da bilmektedir [56].

| ORJİNAL VERİTABANI | | | |
|---|---|---|---|
| **TANIMLAYICI** | **YARI TANIMLAYICI** | | |
| **İsim - Soyisim** | **Cinsiyet** | **Doğum Tarihi** | **Gürültü (DIN)** |
| Ahmet Ertürk | Bay | 1979 | 2046059 |
| Ali Can | Bay | 1982 | 719839 |
| Mustafa Taş | Bay | 1979 | 2241497 |
| Fatma Gül | Bayan | 1995 | 2046059 |
| Ayşe Sert | Bayan | 1987 | 392537 |
| Can Deniz | Bay | 1995 | 363766 |
| Muharrem Tatlı | Bay | 1998 | 544981 |
| Birgül Sev | Bayan | 1984 | 293512 |
| Cem Saygıner | Bay | 1979 | 544981 |
| Zehra Dağlı | Bayan | 1995 | 596612 |
| Ümit Atlıhan | Bay | 1987 | 725765 |

**Tanımlama**

| YARI TANIMLAYICI | | |
|---|---|---|
| **Cinsiyet** | **Doğum Tarihi** | **Gürültü (DIN)** |
| Bay | 1979 | 2046059 |
| Bay | 1982 | 719839 |
| Bay | 1979 | 2241497 |
| Bayan | 1995 | 2046059 |
| Bayan | 1987 | 392537 |
| Bay | 1995 | 363766 |
| Bay | 1998 | 544981 |
| Bayan | 1984 | 293512 |
| Bay | 1979 | 544981 |
| Bayan | 1995 | 596612 |
| Bay | 1987 | 725765 |

**Eşleştirme**

Ayşe Sert

Cinsiyet: Bayan

Doğum Tarihi: 1987

İFŞA DOSYASI

**Şekil 2.10:** Savcının, Arka Plan Bilgisine Sahip Olduğu Belli Bir Hedef Kişiye, Ayşe'ye Ait Bir Kaydın Yeniden Belirlenmeye Çalışıldığı Savcı Riskinin Gösterilmesi [56]

Saldırgan Ayşe'nin DF 'de olduğunu bildiği için, doğum ve cinsiyet açısından eşleşme yapabilir çünkü eşleşen iki kayıt bulunmaktadır. Dolayısıyla eşdeğer sınıf büyüklüğü 2'ye eşittir. Saldırgan bu iki kaydın hangisinin Ayşe ile ilgili olduğunu bilmediğinden rastgele birini seçecektir. Bu nedenle, yeniden tanımlama olasılığı 0.5'dir.

Daha genel olarak, doğru yeniden tanımlama olasılığı şu şekilde verilmektedir [56]:

$$p\theta_j = {}^{1}/_{f_j} \qquad \text{(Denklem } \mathbf{2.24}\text{)}$$

$f_j$, $DF$ 'de eşleşen eşdeğerlik sınıfının boyutudur. Burada $p$ ifadesi savcı riski olduğu belirtilmek için kullanılmaktadır.

Uygulamada, veri saklama birimi, saldırganın Ayşe'yi hedef alacağını önceden bilememektedir. Saldırgan, $DF$ 'teki herhangi bir kişiyi hedefleyebilir. Dolayısıyla, veri saklama birimi, tüm denklik sınıfları için $p\theta_j$değerini hesaplamalıdır.

Düşman, $DF$ 'de olduğu bilinen insanlar hakkında bilgi içeren nüfus kayıtlarına sahipse, kayıt defterindeki bireyleri küçük eşdeğerlik sınıflarında hedef olarak seçebilmektedir. Bu durumda, düşman en küçük eşdeğerlik sınıflarını hedefleyebilir ve $pR_b$ metriği korunmak için mantıklı olacaktır. Bununla birlikte, $DF$ 'de olduğu bilinen kişilerin kayıt defteri bilinmiyorsa, veri saklama sorumlusunun konfor seviyesine bağlı olarak maksimum risk veya ortalama risk (metrikleri) uygun olmaktadır.

### 2.7.5 Gazeteci Riskinin Ölçülmesi

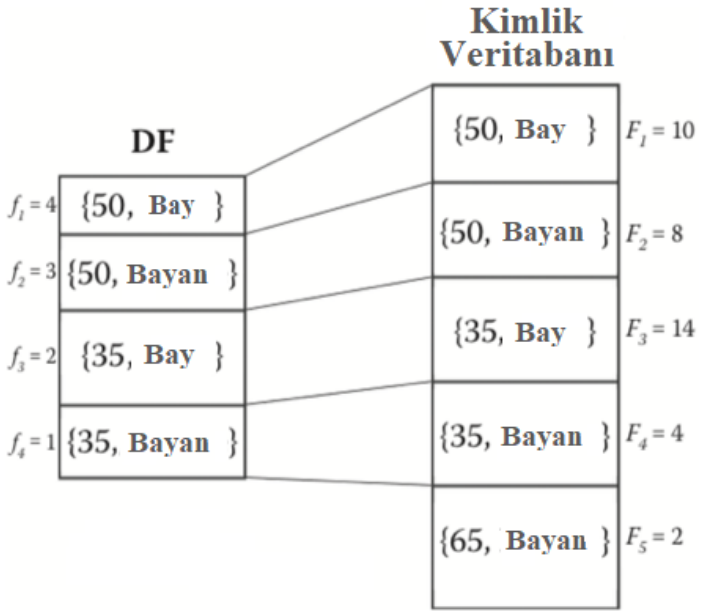
Gazeteci riski veri seti için bir takım örneklem olması durumunda söz konusudur. Açıklamalı veri seti bir nüfus kayıt defteriyse, saldırgan büyük olasılıkla hedefin nüfusta olduğunu biliyor olacaktır. Bunun tersinin mutlaka doğru olmadığını unutmayın: Bir veri seti örnek ise bu, gazeteci riski olduğu anlamına gelmez.

Gazeteci riskinde iki tür yeniden tanımlama saldırısında söz konusudur: (1) saldırgan, belirli bir bireyi hedef almaktadır ve (2) her bireyi hedef almaktadır. Birincisinde saldırgan, belirli bir birey (örneğin, bir komşu veya ünlü bir kişi) hakkında arka plan bilgisine sahipken, ikincisinde, hangi bireyin hedeflendiği önem taşımamaktadır.

Gazeteci riski için, düşmanın açıklanan veri setini başka bir veritabanıyla eşleştirdiği varsayılmaktadır. Şimdilik, açıklanan veri kümesinin bir üst kümesi olduğu bilinen bir tanımlama veritabanının olduğunu varsayılsın. Amerika Birleşik Devletleri'ndeki seçmen kayıt listesi genelde popülasyonun tamamını temsil etmektedir ancak uygun vatandaşların yaklaşık üçte ikisi oy kullanmak için kayıt

yaptırabilmektedir. Ancak, varsayım seçmen listesinin tüm nüfusu kapsadığına dair olarak kabul edilmektedir.

Bu senaryoda, $K$ tanımlama veritabanında denklik sınıfları kümesi olsun, $|K|$ ise tanımlama veritabanındaki eşdeğerlik sınıflarının sayısı olsun. $J \subseteq K$ olur $J = \{x|\forall x: x \in K \wedge f_x > 0\}$. Tanımlama veritabanındaki bir eşdeğerlik sınıfındaki $j$ kayıtlarının, $j \in K$ için $F_j > 0$ olduğu ve tanımlama veritabanındaki kayıtların toplam sayısı, $F_j$ ile gösterilsin [56].




**Şekil 2.11:** Bir Tanımlama Veritabanından Çizilen Örnek Bir ***DF*** Örneği

$$N = \sum_{j \in K} F_j \qquad \text{(Denklem } \mathbf{2.25}\text{)}$$

Açıklanan veritabanındaki eşdeğerlik sınıfı boyutu $f_j$ ile verilmekte ve bu örnekteki toplam kayıt sayısı aşağıdaki gibi olmaktadır:

$$n = \sum_{j \in J} f_j \qquad \text{(Denklem } \mathbf{2.26}\text{)}$$

Açıklanan veritabanının, tanımlama veritabanından basit bir rastgele örnek olduğunu varsayılmaktadır.

Şekil 2.11'deki örneğe göre saldırgan 50 yaşındaki bir erkek olan komşusunu, yani belirli bir bireyi yeniden tanımlamak istesin. Saldırgan, komşunun tanımlama veri

tabanında olduğunu bilmektedir. O zaman, seçilen 50 yaşındaki hedefin $DF$ 'de olması olasılığı ve hedefin $DF$ 'de olduğu göz önüne alındığında, arka plan bilgisi ile doğru eşleşme olasılığı göz önüne alınmalıdır. Hedef kişinin eşdeğerlik sınıfı $j$ 'de olduğu varsayılırsa, olasılık aşağıdaki gibi olur.

$$\frac{f_j}{F_j} \cdot \frac{1}{f_j} = \frac{1}{F_j} \qquad \text{(Denklem } \mathbf{2.27}\text{)}$$

Daha genel olarak, doğru eşleşme olasılığı şudur:

$$j\theta_j = {}^{1}/_{F_j} \qquad \text{(Denklem } \mathbf{2.28}\text{)}$$

Herhangi bir eşdeğerlik sınıfı için fj> 0 olur. Bunun, gazeteci riski olduğunu belirtmek için J alt simgesi kullanılsın. Rakip en küçük eşdeğerlik sınıflarına odaklanırsa, uygun risk metriği olur.

$$j\mathrm{R}_b = max_{j \in J}({}^{1}/_{F_j} \qquad \text{(Denklem } \mathbf{2.29}\text{)}$$

Saldırgan $DF$'den bir kayıt seçip, bir tanımlama veritabanıyla eşleştirmeye ve ayrıca tanımlama veritabanından bir kayıt seçerek eşleştirmeye çalışıyor olabilir.

Rakibin en küçük eşdeğerlik sınıflarına odaklanmayacağı biliniyorsa, ortalama riski ölçülebilmektedir. $DF$'nin kanser hastaları hakkındaki verileri içerdiği varsayılsın. Bir tanıdığa ait kaydı tekrar belirlemeye çalışan bir düşman, en küçük eşdeğerlik sınıflarını hedeflemeyecektir, ancak tanıdıklarına ait eşdeğerlik sınıflarında eşleşme yakalamaya çalışacaktır. Böyle bir durumda ortalama risk önemli olmaktadır.

Bir tanımlama veritabanına karşı eşleştirilen $DF$'den rastgele seçilmiş bir kayıt için, yeniden tanımlamanın ortalama olasılığı şöyledir:

$$j\mathrm{R}_{c1} = \frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j} \quad \text{(Denklem } \mathbf{2.30}\text{)}$$

Öte yandan, tanımlama veritabanından, $DF$ 'ye eşleştirilen rastgele seçilmiş bir kayıt için, yeniden tanımlamanın ortalama ihtimali şöyledir:

$$j\mathrm{R}_{c2} = \frac{|J|}{\sum_{j\in J} F_j} \quad \text{(Denklem } \mathbf{2.31}\text{)}$$

Bir saldırganın hangi saldırı yöntemini kullanacağını önceden bilmek mümkün olmadığından, genel ortalama risk, şöyle ifade edilebilir:

$$j\mathrm{R}_c = max\left(\frac{|J|}{\sum_{j\in J} F_j}, \frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}\right) \quad \text{(Denklem } \mathbf{2.32}\text{)}$$

Uygulamada, veri saklama sorumlusu gazeteci riskini hesaplamak için tanımlama veritabanına erişememektedir. Dolayısıyla, $J\theta_j$ değeri açıklanan veritabanındaki bilgileri kullanarak hesaplanmalıdır. Bunun için çeşitli yöntemler geliştirilmiştir.

Veri saklamacısının bir tanımlama veritabanına sahip olmamasının iyi nedenleri bulunmaktadır. Genellikle, bir nüfus veritabanı elde etmek pahalı olmaktadır. Veri sahibi birden çok popülasyonu koruması gerekecektir ve böylelikle bir nüfus veri tabanı elde etmek için efor ve masrafın çarpımını hesaplamak oldukça yüksek bir maliyete sebep olacaktır. Örneğin, Kanada'da yeniden tanımlama saldırıları için kullanılabilen yarı-kamusal kayıt defterlerini kullanan tek bir mesleğe özgü veritabanının maliyeti 150.000 ila 188.000 $ arasındadır. Ticari veritabanları nispeten daha pahalı olmaktadır. Amerika Birleşik Devletleri'nde, Alabama'daki seçmen kayıt listesinin maliyeti 280.000 Louisiana için 5.000 New Hampshire için 8.000 Wisconsin için 12.000 ve Batı Virginia için 17.000 dolardan fazla olarak gerçekleşmiştir [56]. Bir saldırgan nüfus kayıtlarına erişmek için yasadışı hareketler de yapabilmektedir. Örneğin, gizlilik yasası ve Kanada Seçim Yasası seçmen listelerinin seçim faaliyetlerini yürütmek ve desteklemek için kullanılmasını kısıtlamaktadır. Bir terörist

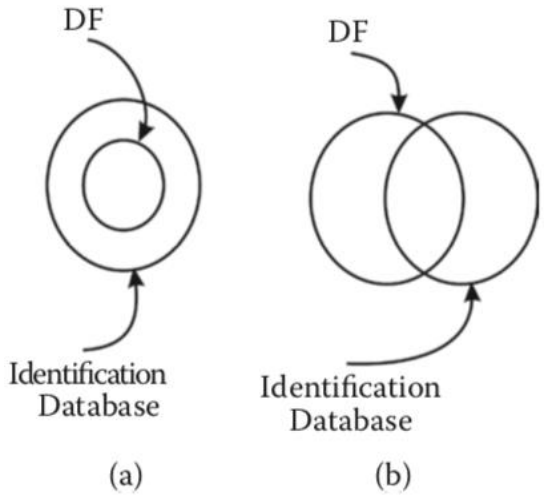
grubu desteklediği iddia edilen bir yardım kuruluşunun fon artırımı için Kanada seçmen listelerini elde ettiği gerçek hayattan bir örnek bulunmaktadır. Yasal bir veri sahibi, sorumlusu bu tür fiillere girmez ve bu nedenle uygun bir tanımlama veritabanına sahip olamayabilir. Birleşik Devletlerde bazı eyaletlerde seçmen listeleri yalnızca seçimlerle ilgili amaçlar için kullanılabilmektedir. Bir saldırgan bu listelere girebilmek için çeşitli yollar deneyebilir. Bu koşullar altında, veri saklama görevlisinin $p\theta_j$ 'yi tahmin etme kabiliyeti önem kazanmaktadır.

Veri saklama birimi, bireysel $p\theta_j$ değerlerini hesaplamak için türetilen bir risk türünü seçmelidir. Tablo 2.22'de $\tau = 0.33$ olduğu varsayılarak hesaplanan gazeteci riskin ölçümü gösterilmektedir.

**Tablo 2.22:** DF 'Nin Senaryosunun Kimlik Veri Tabanının Uygun Bir Alt Kümesi Olması İçin Türetilmiş Üç Farklı Gazeteci Risk Ölçümünün Hesaplanması [56]

| Türetilmiş Risk Ölçümü | Denklem | Risk Değerleri |
|---|---|---|
| $JR_a$ | $\frac{1}{n}\sum_{j\in J} f_j \times / \left(\frac{1}{F_j} > \tau\right)$ | 0 |
| $JR_b$ | $\max_{j\in J}\left(\frac{1}{F_j}\right) = \frac{1}{\min_{j\in J}(F_j)}$ | 0.25 |
| $JR_c$ | $max\left(\frac{\lvert J\rvert}{\sum_{j\in J} F_j}, \frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}\right)$ | 0.12 |

Şimdiye kadar $DF$'nin tanımlama veritabanından bir örnek olduğu varsayılmıştır. Bununla birlikte, başka durumlar için de birçok örnek bulunmaktadır. $DF$ ile tanımlama veritabanı arasında bireylerin tam örtüşmediği durumlarda mevcuttur. Şekil 2.12 'de (b) olarak gösterilmiştir. $DF$ 'nin tanımlama veritabanının bir alt kümesi olduğu durum, şekil (a) da gösterilmektedir.




**Şekil 2.12:** Gazeteci Riski İçin İki Senaryo [56]

Senaryo (b) uyarınca, $DF$'teki bazı bireyler de tanımlama veritabanında bulunmaktadır. Sayımın tam nüfus veri tabanının olduğunu varsayılmaktadır. Daha sonra bireyler sayım sonucunda $DF$ oluşturmak için eşit olasılıkla örneklenmektedir.

### 2.7.6 Pazarlamacı Riskinin Ölçülmesi

Saldırgan, açıklanan veri kümesinden hangi kayıtların doğru olarak yeniden tanımlandığını bilmediğini dikkate almamaktadır. Bunun yerine dikkate aldığı, açıklanan veri kümesindeki doğru olarak yeniden tanımlanan kayıtların oranının yeterince yüksek olmasıdır.

Pazarlamacı risk metriği doğru olarak yeniden tanımlanan kayıtların (beklenen) oranında aranmalıdır. Bu terim, saldırganın pazarlama amacıyla iki veritabanını açıklanan veritabanındaki kişilere eşleştirdiği temel senaryoyu temsil etmek için kullanılır.

Pazarlamacı riskinin hesaplanması gereken iki durumu var. Birincisi, açıklanan veri tabanı, kimlik veritabanıyla aynı kişilere sahip olmasıdır. Örneğin, bir reçete veritabanı tüm seçmenleri içeriyorsa ve seçmen listesi ile eşleşiyorsa. İkinci ve daha olası durum, ifşa edilen verileri tanımlama veritabanından bir alt set / numune olduğunda gerçekleşir. Örneğin, açıklanan reçete kayıtları veritabanı, seçmen listesinin kapsadığı nüfusun bir alt kümesidir [55].

Veri saklamakla yükümlü birimin bir veritabanını anonimleştirmeye ihtiyaç olup olmadığına karar vermek için, pazarlamacının riskini hesaplamak istediğini veya açıklama kontrol faaliyetlerine ihtiyaç duyulduğunu varsayıyoruz. Veri saklamakla

yükümlü birimin kimlik veritabanına sahip olması pek olası değildir. Örneğin, reçete kayıtlarını satan bir eczane zinciri, pazarlamacı riskinin yüksek olup olmadığını belirlemek için bir saldırganın, nüfus kimlik veritabanını elde etmek için faaliyet gösterdiği tüm ülkelerdeki tüm seçmen listesinin satın alınamayacağından, risk ölçütü yalnızca açıklanan veritabanından hesaplanmalıdır.

Yukarıdaki ilk durumda, açıklanan veriler tanımlama veritabanıyla aynı olup bu nedenle tanımlama veritabanına erişim sorun değildir. Bununla birlikte, veritabanını saklamakla yükümlü birimde iki veritabanının bulunmamasının birçok nedeni vardır. Örneğin bir nüfus veritabanı tutmanın maliyetinin yüksek olması gibi.

Önemli bir soru, bir veri saklama sorumlusu, kayıtların doğru olarak yeniden tanımlanmasının beklenen oranının çok yüksek olduğuna ne zaman karar verir? Örneğin Kanser kayıtlarına ilişkin eski verileri, araştırma amaçlı sunmak için kabul edilebilir yüksek riskli kayıtların %5 ve %20'lik dilimlerde olması beklenmektedir. Bunlar, pazarlamacı risk değerlerinin kabul edilebilirlik eşiklerini belirlemek için bir temel olarak kullanılabilir.

Kimlik bilgilerini araştırma amacı ile sunmak için yukarıda bahsi geçen iki risk önlemi tanımlanmıştır [55]. Birincisi, savcının riski;

$U = D.$ olarak hesaplanır:

$$R_p = {}^{1}\Big/_{\min\limits_{j}\,(f_j)} \qquad \text{(Denklem } \mathbf{2.33}\text{)}$$

İkincisi, $U \subset D$ zamanında uygulanabilen ve

$$R_j = {}^{1}\Big/_{\min\limits_{j}\,(F_j)} \qquad \text{(Denklem } \mathbf{2.34}\text{)}$$

olarak hesaplanan gazeteci riski. Her iki durumda da risk ölçümü, tek bir kaydı yeniden belirlerken daha kötü ihtimali yakalar; oysa pazarlamacının riski için, doğru olarak yeniden tanımlanacak olan kayıtların beklenen sayısını (oranını) değerlendiririz. Bir diğer önemli fark ise bu pazarlamacı riski, $U$'daki hangi kayıtların yeniden

tanımlanması muhtemel olduğunu belirlemeye yardımcı olmaz. Bununla birlikte, gazeteci ve savcı risk önlemleri ile en yüksek riskli kayıtları tespit etmek ve bunları sunmak için verileri kontrol eylemine odaklanılmalıdır.

Şu anda, pazarlamacı riskini kontrol etmek için özel olarak tasarlanmış herhangi bir algoritma bulunmamaktadır. Bununla birlikte, pazarlamacı riskini kontrol etmek için mevcut k-anonimlik algoritmaları kullanılmaktadır.

Bir saldırganın, pazarlamacı riskinin bir miktar eşiğin altına, yani $\tau$ değerinin altında olmasını istediğini varsayalım [55].

$$\text{Sonra } \frac{1}{n}\sum_j \frac{f_j}{F_j} \leq \left[\frac{1}{\min\limits_j (F_j)} . \frac{\sum_j f_j}{n}\right] = \frac{1}{\min\limits_j (F_j)} \qquad \text{(Denklem } \mathbf{2.35}\text{)}$$

Bu nedenle, $R_j \leq \tau$ 'i sağlayarak pazarlamacı riskinin bu eşiğin altında olmasını sağlayabiliriz. Bu eşitsizliği garanti etmek için herhangi bir k-anonimlik algoritması kullanılabilir. K-anonimlik algoritmalarının dezavantajı, gerekenden daha fazla özdeşleşmeye neden olabilmesidir. Pazarlamacı risk değeri uygulamada $R_j$'den biraz daha küçük olabilir. Örneğin, 3 eşdeğerlik sınıfı $F_j \in \{5,20,23\}$ ve benzersizlerin bulunduğu örnekle bir popülasyon veri seti düşünün. Bu durumda, pazarlamacı risk değeri, $R_j$ değerinin yarısı olacaktır. Bu nedenle, mevcut k-anonimlik algoritmalarını kullanmak şimdilik uygun bir yaklaşım olsa da, araştırma topluluğunun pazarlamacı riskini doğrudan yöneten algoritmalar geliştirmesi gerektiği düşünülmektedir.

Bir saldırganın bir kimlik veritabanına sahip olması durumunda, tek bir kişinin kimliğini yeniden tanımlamak veya mümkün olduğunca çok kişiyi yeniden tanımlamak için kullanabilir. Birinci durumda ya savcı ya da gazeteci risk metrikleri ölçümleri ve ikinci durumda pazarlamacı risk metriği ölçümü kullanılmaktadır. Bu nedenle, bir risk önleminin seçimi, saldırganın düşüncesine bağlı olacaktır. Karar vermesi zor olsa da pazarlamacı riskinin uygulanabilir olduğu ve değerlendirilmesi ve yönetilmesi gereken birincil riski temsil ettiği açık senaryolar mevcuttur.

Örnek bir senaryo, açıklanan veritabanındaki tüm bireylere bir ürünü pazarlama konusunda motive olan bir saldırgan olduğunu düşünelim. Bu durumda, saldırgan, bireyleri yeniden tanımlamak için seçmen listesi gibi bir kimlik veritabanı kullanabilir. Saldırganın hangi kayıtların yeniden tanımlandığını bilmesine gerek yoktur. Çünkü bir bireyi pazarlama kampanyasına dahil etmenin artan maliyeti düşüktür. Doğru

tanımlamaların beklenen sayısı yeterince yüksek olduğu sürece, bu saldırgana yeterli bir dönüş sağlayacaktır. Bir pazarlama potansiyelinin var olduğunu bilen veri saklama birimi, pazarlamacı riskini tahmin eder ve böyle bir ifşa için önlem alır.

İkinci bir senaryo, veri saklama sorumlusu veriyi birden çok tarafa ifşa ettiğinde ortaya çıkar. Örneğin kayıt defteri etnik kökenli ve sosyo-ekonomik göstergelere sahip bir A veri setini bir araştırmacıya, zihinsel sağlık bilgileri olan bir veri setini başka bir araştırmacıya açıklayabilir. Her iki veri seti de hastalar üzerinde aynı çekirdek demografik bilgileri paylaşıyor. Kayıt defteri, verilerin hassasiyeti ve grup zarar potansiyeli nedeniyle etnik köken, sosyoekonomik ve zihinsel sağlık verilerini aynı araştırmacıya bırakmaz. Bunu farklı araştırmacılarla paylaşır. Bununla birlikte, iki araştırmacı A ve B'yi bir araya getirebilir ve kayıt defterini taleplerine karşı bağlayabilir. Verileri açıklamadan önce, veri yöneticileri, araştırmacıların verileri birbirine bağlamış olması halinde, ortak demografik veriler üzerinde doğru bir şekilde eşleştirilebilen beklenen kayıt sayısını değerlendirmek için pazarlamacı riskini değerlendirebilir. Çekirdek demografik bilgilerin ayrıntısına kadar anonimleştirerek pazarlamacı riskini minimuma indirir.

Acil durumda getirilen tüm hastaların bir listesinin hastanede bulunduğu üçüncü bir senaryo düşünün. $D$ ' Bu veriler daha sonra sendromik durum ve genel durumsal farkındalığı sağlamak için $\acute{D}$ olarak bir belediye halk sağlığı birimine gönderilmektedir. Veri setinde benzersiz tanımlayıcılar bulunmamaktadır. Ancak kamu sağlık biriminde bir ihlal meydana geldi ve kayıtların %10'u $U$, bir saldırgan tarafından ele geçirildi. Halk sağlığı birimi, bu hastaların verilerine ulaşıldığını hastalara bildirmekle yükümlüdür. $\acute{D}$ Tanımlanmadığından, halk sağlığı birimi, hastanenin yardımıyla hastaları belirleyerek onlara bildirimde bulunmalıdır. Bildirimde bulunulan hastalar ne kadar fazla olursa, halk sağlığı biriminin maliyeti de artar ve muhtemelen telafi masraflarını da arttırır. Yapılması en kolay, fakat maliyeti en fazla olan çözüm $D$ 'teki tüm hastaları bilgilendirmek için hastaneyle birlikte çalışmaktır. Bununla birlikte, halk sağlığı ünitesi $\acute{\lambda}$ 'ı tahmin etmek ve hastaneden orijinal $\acute{D}$ verileri ile uyuşmayan alt kümenin eşleştirilip eşleştirilmesinin yeterli derecede yüksek bir başarı oranına sahip olup olmayacağını belirlemek için $U$ kullanabilir. $\acute{\lambda}$ yüksekse, halk sağlığı ünitesi $U$'i $D$'e bağlamayı ve sadece yasal bildirim şartına uyacak en uygun maliyetli seçeneği olacak şekilde yeniden tanımlanan

hastalara bildirilmesini isteyecektir. Eğer $\acute{\lambda}$ düşük ise, $\acute{D}$ 'deki tüm hastalar, ihlal alt grubuna dahil olsun veya olmasın, bunların %90'ı ihlalden etkilenmese de bildirilir.

### 2.7.7 Türemiş Metriklerin ve Karar Kuralı Uygulanması (Applying the Derived Metrics and Decision Rule)

Tanıma dışı bağlamında, $\alpha$'yı sıklıkla sıfıra eşit olarak alınır; Dolayısıyla $R_a$ tipi türetilmiş ölçümler, bize ne kadar bastırılması gereken kayıt olduğunu bildirmektedir. Örneğin, $R_a = 0.053$ değeri ise, $\tau$ değerinden daha yüksek bir risk içeren kayıtların bulunmamasını sağlamak için kayıtların sadece %5'inden fazlasının bastırılması gerekmektedir [55].

$R_a$ tipi metrikler değerleri tamamlayıcı niteliktedir. Örneğin, $\tau$ eşiğimiz 0.33 'ten yüksek olan $R_b = 0.4$ 'e sahipsek, $R_b = 0.053$ değeri, kayıtların sadece %5 'inin tüm uçtaki bu yüksek riske neden olmaktadır.

$R_c$ türevi metriklerin de özel bir yorumu bulunmaktadır: Saldırgan, $DF$ 'deki tüm kayıtları bir tanımlama veritabanıyla eşleştirmeye çalışırsa, $DF$ 'deki kayıtların ortalama oranı doğru olarak yeniden tanımlanacaktır. Bu yorumla birlikte, bu türetilmiş metrikler, bize kaç kayıtın yeniden tanımlandığının bir ölçüsü olduğunu ve bu nedenle bunu pazarlamacı riski olarak etiketlediğimizi ortaya koymaktadır. Tanımlama veritabanının $DF$ ile tam olarak aynı kayıtlar ve veri konuları olup olmadığına bağlı olarak iki pazarlamacının risk ölçütü $mR_1$ ve $mR_2$ olarak gerçekleşmektedir.

Eğer riskin yüksek olduğu belirlenirse, bu riski azaltmak için tanımlama yöntemleri uygulanabilir.

### 2.7.8 Metrikler Arasındaki İlişki (Relationship Among Metrics)

Göz önüne alınması gereken diğer bir nokta, risk ölçümleri arasındaki ilişkidir. Aynı veri setleri ve eşikleri için, sayısal olarak bir dizi eşitsizlik bulunmaktadır. Tez eşitsizlikleri, hangi metriklerin kullanılacağına karar vermek için pratikte faydalı olabilmektedir.

Çünkü tüm $j$ için, $f_j \leq F_j, j \in J$ için $pR_b \geq jR_b$ olmaktadır ve eğer bir sabit $j$ için $I(1/F_j > \tau)$ doğruysa, o zaman $I(1/f_j > \tau)$ da doğru olmalıdır; bu da $pR_a \geq jR_a$ anlamına gelmektedir [56].

**Tablo 2.23:** Yeniden Tanımlama Risk Ölçümlerinin Özeti [56]

| Risk Türü | Denklem | Notlar ve Durum |
|---|---|---|
| **Savcı Riski** | $pR_a = \frac{1}{n}\sum_{j\in J} f_j \times / \left(\frac{1}{f_j} > \tau\right)$<br>$pR_b = \frac{1}{\min_{j\in J}(f_j)}$<br>$pR_c = \frac{\lvert J\rvert}{n}$ | Burada $f_j$ DF 'deki denklik sınıfının büyüklüğüdür. Eğer DF, tüm popülasyonla aynıysa, o zaman $f_j = F_j$ ve $F_j$ popülasyondaki denklik sınıfı büyüklüğüdür. |
| **Gazeteci Riski** | $JR_a = \frac{1}{n}\sum_{j\in J} f_j \times / \left(\frac{1}{F_j} > \tau\right)$<br>$JR_b = \frac{1}{\min_{j\in J}(F_j)}$<br>$JR_c = max\left(\frac{\lvert J\rvert}{\sum_{j\in J} F_j}, \frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}\right)$ | Bu ölçütler, DF'nin tanımlama veri tabanının uygun bir alt kümesi olduğu durumlar için uygundur. Nüfus kayıtlarının tamamı hazır olmadıkça, değerin veri sorumlusu tarafından tahmin edilmesi gerekecek, bu durumda sadece sayılabilir. |
| **Pazarlamacı Riski** | $mR_1 = \frac{\lvert J\rvert}{N}$ | Bu metrik, tüm veritabanlarını iki veritabanında eşleştirmek için uygundur; sonra n=N (yani, rakip, DF 'deki ile aynı kişilerle ilgili kayıtları içeren bir tanımlama veritabanına ulaşır) |
| | $mR_2 = \frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}$ | Bu metrik, n<N ve açıklanan veri setinin tanımlama veri tabanınına ($J \subseteq$ K) uygun bir alt kümesi olduğu durumlar için uygundur. Tüm nüfus kaydı hazır olmadıkça $\frac{1}{F_j}$ 'nin değeri veri saklama uzmanı tarafından tahmin edilmelidir, bu durumda $F_j$ sadece sayılabilir. Bu, DF 'deki kayıtların (ortalama) doğru eşleşmesini sağlar |

Bu ilişkilerin temel öneme sahip olmasının nedeni, birden fazla risk türünün yönetiminin yapılması gerektiğine karar verildiğinde, bu ilişkilere ilişkin hangi ölçütlerin gerçekte hesaplanacağına ve hangi risklerin yönetileceğine karar verilmesinde yardımcı olabilecekleridir.

# ÜÇÜNCÜ BÖLÜM

# MATERYAL VE YÖNTEM

Bu çalışmada amaç; mülkiyet verileri kullanılarak kişilerin kimlik bilgilerinin ve mahrem bilgilerinin ifşa edilmemesi için mevcut yöntemlerle ilgili çalışmaların yapılmasıdır. Mülkiyet verilerinin özniteliklerinin belirlenmesi ve kişisel verilerin tahmin edilmesinin önüne geçilmesidir. Mülkiyet verilerindeki öznitelikler kullanılarak mahremiyetin ifşası konusunda 3 test yapılacaktır.

Test-1; Belirlenen özniteliklere anonymity gizlilik modeli seçilerek 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 değerlerinin uygulanması ve bu değerler karşısında Savcı, Gazeteci, Pazarlamacı risklerinin gözlemlenmesidir.

Test-2; Hassas olarak belirlenen özniteliklere diversity gizlilik modeli uygulanarak Savcı, Gazeteci, Pazarlamacı risklerinin gözlemlenmesidir.

Test-3; Hassas olarak belirlenen özniteliklere closeness gizlilik modeli uygulanarak Savcı, Gazeteci, Pazarlamacı risklerinin gözlemlenmesidir.

## 3.1 Veri Seti Yapısı ve Özellikleri

Çalışmada kullanılan veri seti 2 ana başlık altında toplanmaktadır. Birinci kısım parsel verisi (Mülkiyet verisi), ikinci kısım ise kişi bilgileridir. Bu veri seti 9.873 kişinin parsel verisi ve nüfus verisini içermektedir. Sentetik olarak oluşturulan Veri seti üzerinde öncelikle yapılan işlem ilçe ve mahalle isimleri aynı olan data üzerinde değişiklik yapılmasıdır. Oluşturulan test verisinde ilçe isimleri ve mahalle isimleri aynı olan birçok kayda rastlanılmıştır. Örneğin Kale Mahallesi birçok ilimizde mevcuttur. Bu nedenle tekilliği sağlamak için ilçe isimlerinin aynı olması durumunda İl ismi ilçe isminin başına getirilerek düzeltme yapılmıştır (Ağrı Merkez, Kars Merkez vb.). Mahalle verisi içinse aynı olan isimlerin önüne ilçe isimleri getirilerek benzersizlik sağlanmıştır (Çaycuma ilçesi Cumhuriyet mahallesi, Çankaya ilçesi Cumhuriyet mahallesi).

Mülkiyet verisi borçlar kanunu, medeni kanun vb. kanunların gereği olarak oldukça karmaşık bir yapıya sahiptir. Bu nedenle, tapu mevzuatında geçen bazı durumlarla ilgili kısa bilgilerin verilmesi gerekmektedir.

**Akitli işlemler;** Tapu sicil müdürlüğünde resmi senet düzenlenerek yapılan işlemlerdir. Satış, bağış, taksim, ölünceye kadar bakma akdi, ipotek ve irtifak hakları akitli işlemlere örnek gösterilebilir. Tapu Kanunu'nun 26. Maddesinde mülkiyet ve mülkiyet dışındaki ayni hakların kurulması ve devri için tapu sicil müdürlüğünde resmi senet düzenleneceği belirtilmiştir. Şahsi haklar için resmi senet düzenlenmez.

**Akitsiz işlemler;** Tapu Kanunu'nun 26. Maddesinin dışında kalan, ayni hakların kurulması ve devri ile ilgili olmayan kişisel hakların şerhi, terkin, düzeltme, değişiklik, ayırma, birleştirme, mirasın intikali gibi hak sahibinin veya malikin tek taraflı iradesi ile tapu sicil müdürlüğünde yapılan işlemlerdir. Akitsiz işlemler için resmi senet yerine Tescil İstem Belgesi düzenlenir.

**Paylı Mülkiyet;** Müşterek mülkiyet, birden çok kişinin bir şeye (taşınmaza) payları belirli bir şekilde sahip olmalarıdır.

**Elbirliği mülkiyet;** Birden çok kişinin bir taşınmaz mal üzerinde pay oranları açıkça gösterilmeden malik olmaları demektir.

**Aleniyet;** Açıklık. Tapu sicil bilgilerinin sadece ilgililerine açık olması ilkesi anlamındadır.

**Devletin Sorumluluğu;** Medeni Kanun'un 1007. maddesi uyarınca, tapu sicilinin tutulmasından kaynaklanan zararlardan kusursuz dahi olsa Devletin sorumlu olması demektir. Sicili tutan tapu memuru ise, ancak kusurlu olması halinde ve kusuru oranında sorumludur.

**Kat İrtifakı;** Bir arsa üzerinde yapılmakta veya ilerde yapılacak olan bir veya birden çok yapının bağımsız bölümleri üzerinde, yapı tamamlandıktan sonra geçilecek kat mülkiyetine esas olmak üzere, o arsanın maliki veya paydaşları tarafından arsa payına bağlı olarak kurulan irtifak hakkına kat irtifakı denmektedir.

**Kat Mülkiyeti;** Tamamlanmış bir yapının bağımsız bölümleri üzerinde, o gayrimenkulün maliki veya ortak malikleri tarafından kat mülkiyeti kanunu hükümlerine göre kurulan bir mülkiyet türüdür. Kat mülkiyeti dikey kat mülkiyeti ve yatay (yaygın) kat mülkiyeti olarak ikiye ayrılmaktadır.

**Yevmiye;** Tapu siciliyle ilgili olarak tapu sicil müdürlüğünde yapılan her işlem ile red (onaylanmayan) edilen istemlerin tarih, zaman ve sıra numarasına göre kaydedildiği deftere yevmiye defteri; bu defterden alınarak tapu kütüğünde yapılan

tescil veya terkinin yanında belirtilen tarih ve numaraya da, yevmiye tarihi ve numarası denmektedir.

Sentetik olarak oluşturulan bu veriler üzerinde öncelikle (processing) düzeltme yapılarak eğitim ve test işlemleri için uygun hale getirildikten sonra çalışmalarımızda kullanılmaktadır.

## 3.2 Veri Setinin Yapısı

Veri seti kişilerin; mülkiyete ait taşınmaz bilgisi ve mülkiyetin sahibine ait nüfus verilerini içermektedir. Bunlar tablo 3.1 içinde verilmiştir.

**Tablo 3.1:** Mülkiyet Verisine Ait Öznitelik Alanları

| **Mülkiyete verisi** |
|---|
| Taşınmaz No |
| İl |
| İlçe |
| Mahalle / Köy |
| Ada |
| Parsel |
| Alan |
| Nitelik |
| Pafta |
| **Kişi verisi** |
| TC No |
| Adı |
| Soyadı |
| Cinsiyet |
| Yaş |
| Baba Adı |
| Anne Adı |
| Doğum Yeri |
| Nüfusa Kayıtlı İl |
| Nüfusa Kayıtlı İlçe |
| Cilt No |
| Aile No |
| Sıra No |

**Taşınmaz Numarası;** Tüm zeminlerde benzersizliğin sağlanması için verilen numara (Parsel ve üzerindeki bina ayrı ayrı zemin olarak adlandırılır.)

**İl;** Parselin hangi il içerinde yer aldığını belirtmektedir.

**İlçe;** Parselin hangi ilçe içerisinde yer aldığını belirtmektedir.

**Mahalle / Köy;** Parselin hangi mahalle veya köy sınırları içerisinde yer aldığını belirtmektedir.

**Ada Numarası:** Tüm çalışma alanı içindeki adaların benzersizliğin sağlanması için verilen numaradır. (Ada Numarası ilçe içinde benzersizdir. Fakat ilçe içerisinde mahalle aktarımlarından dolayı bu benzersizlik kısmen de olsa bozulmuştur.)

**Parsel Numarası;** Tüm ada içindeki parsellerde benzersizliğin sağlanması için verilen numaradır.

**Alan:** Mevcut parsellere ait yüzölçümü bilgisini içermektedir.

**Nitelik:** Parselin kullanım amacına yönelik bilgi içeren alandır.

**Pafta:** Parsele ait pafta numarasını içermektedir.

**TC;** Kişiyi benzersiz kılmak için verilen numaradır.

**Ad;** Kişinin Adı

**Soyad;** Kişinin Soyadı.

**Cinsiyet;** Kişinin Cinsiyeti.

**Yaş:** Kişinin yaş bilgisinin tutulduğu alandır.

**Baba Adı;** Kişinin baba adı.

**Anne Adı;** Kişinin anne adı

**Doğum Yeri;** Kişinin doğum yeri

**Nüfusa Kayıtlı İl;** Kişinin nüfusa kayıtlı olduğu il

**Nüfusa Kayıtlı İlçe;** Kişinin nüfusa kayıtlı olduğu ilçe

**Cilt No;** Kişinin cilt numarası

**Aile No;** Kişinin aile numarası

**Sıra No;** Kişinin sıra numarası

Veri setinde kullanılacak olan öznitelikler için belirlenen yari tanımlayıcılar ve hassas nitelikler aşağıda gösterilmiştir.

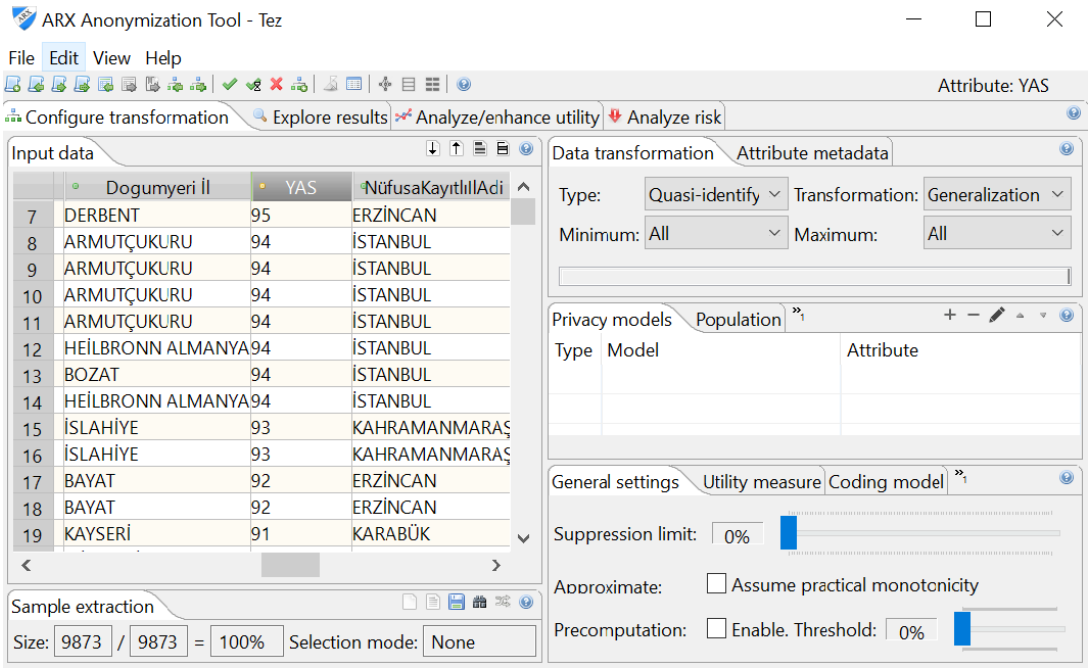
**Tablo 3.2:** Test İçin Kullanılan Öznitelikler

| | |
|---|---|
| Yarı tanımlayıcılar (Quasi-identifying) | Cinsiyet, Yaş, Nüfusa kayıtlı il |
| Hassas değerler (Sensitive) | Taşınmaz No, Mahalle/Köy, Alan |

### 3.2.1 Yaş Özniteliğinin Yarı Tanımlayıcı Olarak Özellikleri

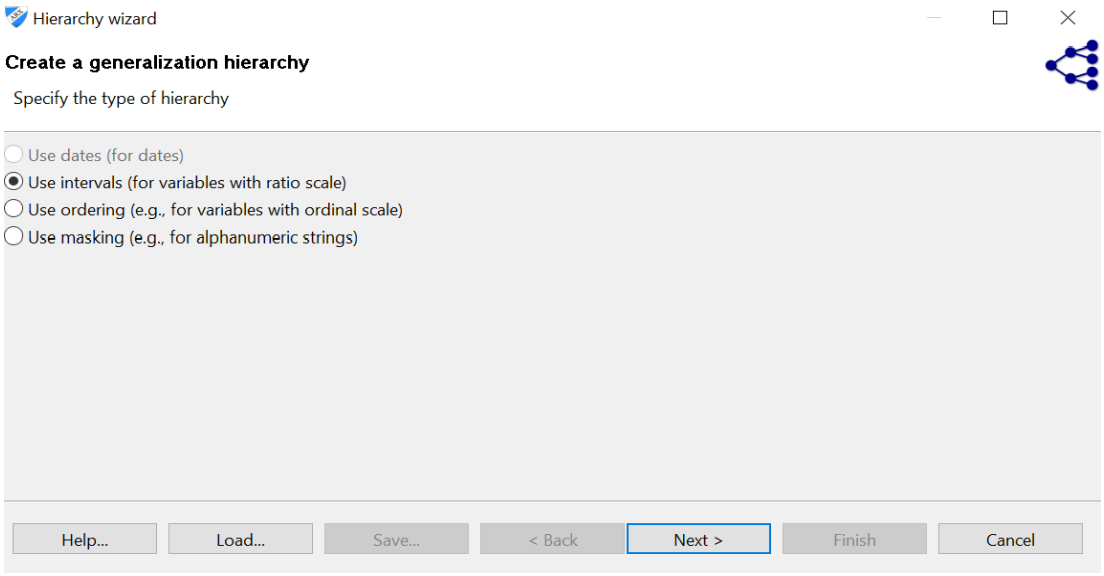
Doğum tarihi alanı mülkiyet verisi içinde bulunan ve kişi bilgilerini kapsayan kısım içinde yer almaktadır. Veri seti içerindeki bu alan hesaplama yöntemi ile yaş özniteliği oluşturulmuştur.

Veri seti içerisine yaş özniteliği yarı tanımlayıcı olarak seçilmiştir. Bu nedenle ARX programında yaş özniteliği için hiyerarşi kurulmuştur. İlk olarak ARX' e yüklenen veri setinden yaş öz niteliği seçilerek "Data Transfarmation" sekmesinin altında "Type" alanından "Quasi-identifying" seçilmiştir.



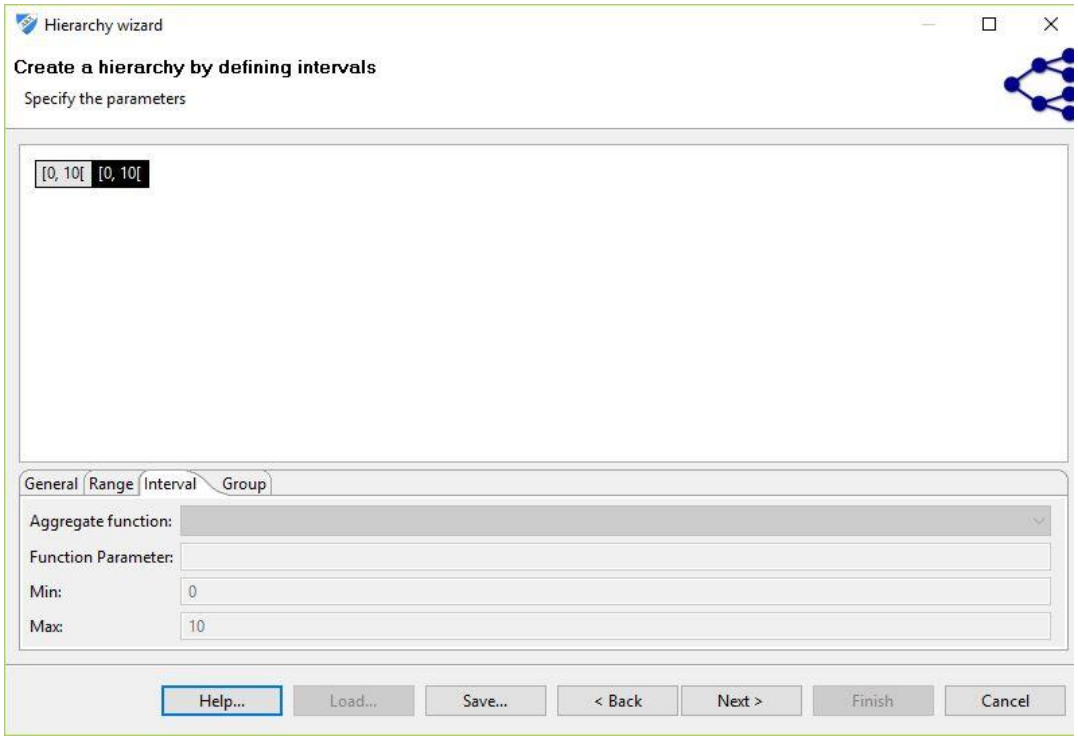

**Şekil 3.1:** Mahalle Özniteliğinin Yarı Tanımlayıcı (Quasi-İdentifying) Olarak Seçilmesi

Bu alanın seçilmesi ile "Edit" menüsü altında bulunan "create hierarchy" çalıştırılmıştır. Ekrana gelen ekrandan "use intervals (for variables with ratio scale) seçilerek "Next" tıklanmıştır.



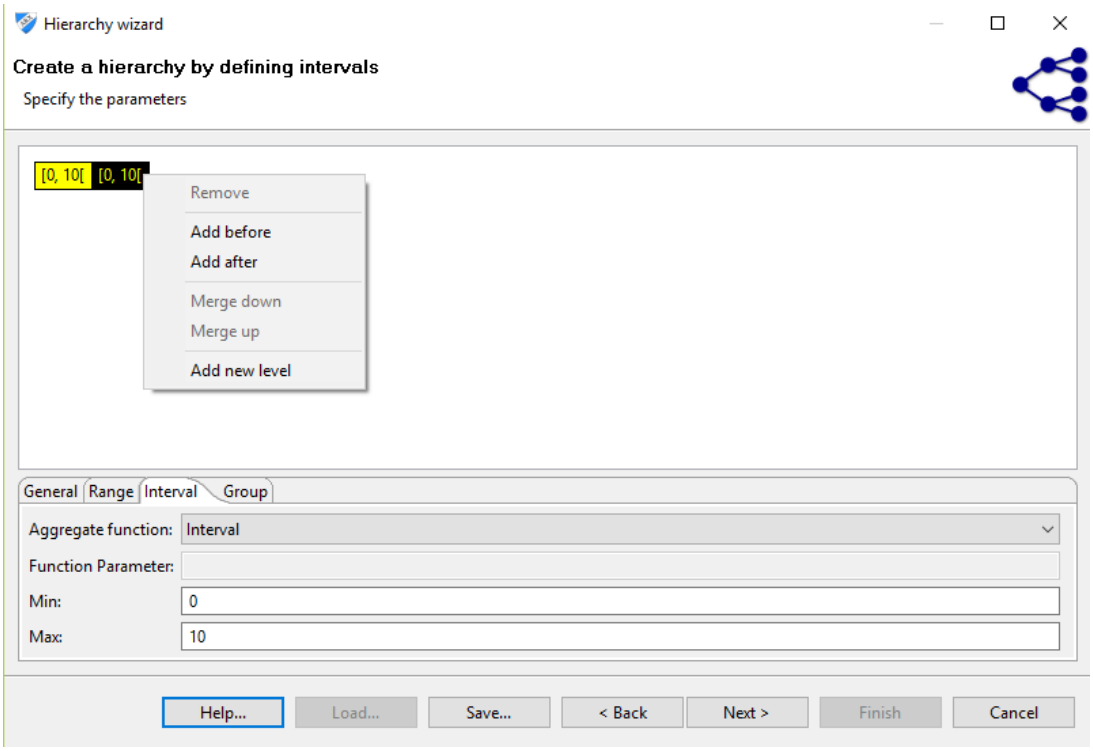

**Şekil 3.2:** Yaş Özniteliği İçin Aralık Kullanımı ile Hiyerarşinin Seçimi

Açılan ekran dan "Intervaval" sekmesinin altında "Aggregate function" alandan "Intervaval" seçilmiştir. "Min" değerinin karşısına "0" değeri, "Max" değerinin karşısına ise "10" değeri girilmiştir.



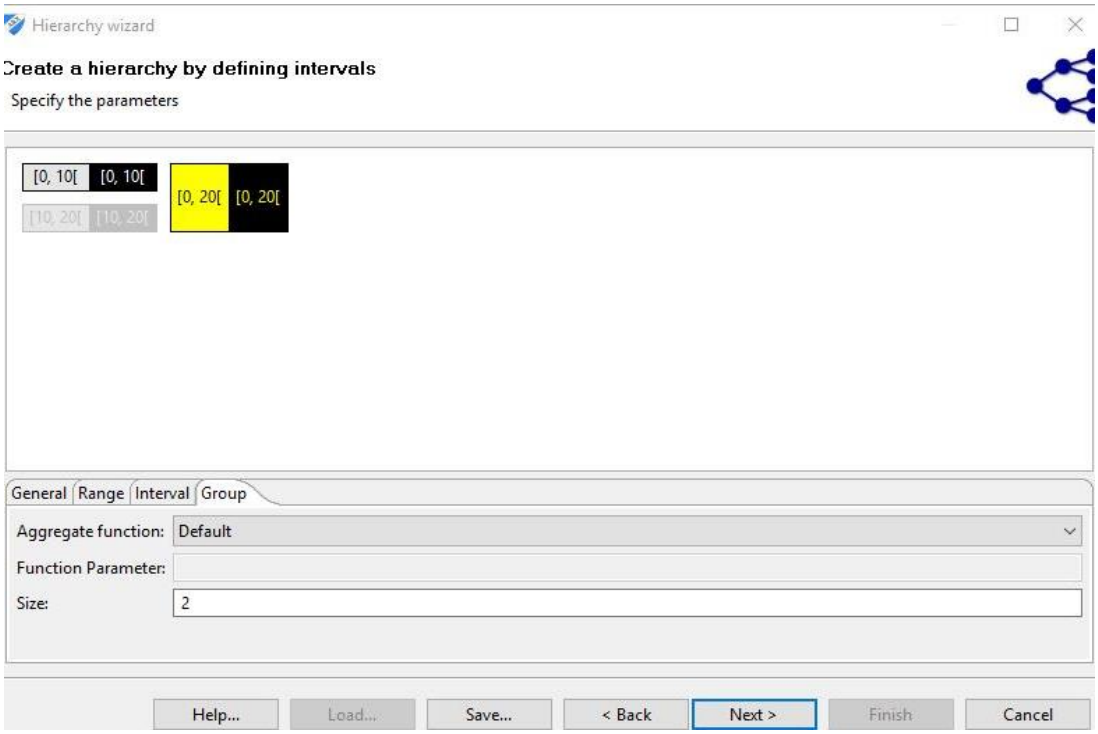

**Şekil 3.3:** Yaş Özniteliği İçin Birinci Seviye Gösterim Aralığının Beş Olarak Belirlenerek Anonimleştirilmesi

Ardından girilen değerlerin göründüğü kısma Mouse ile sağ tıklanarak "Add new level" seçilmiştir.



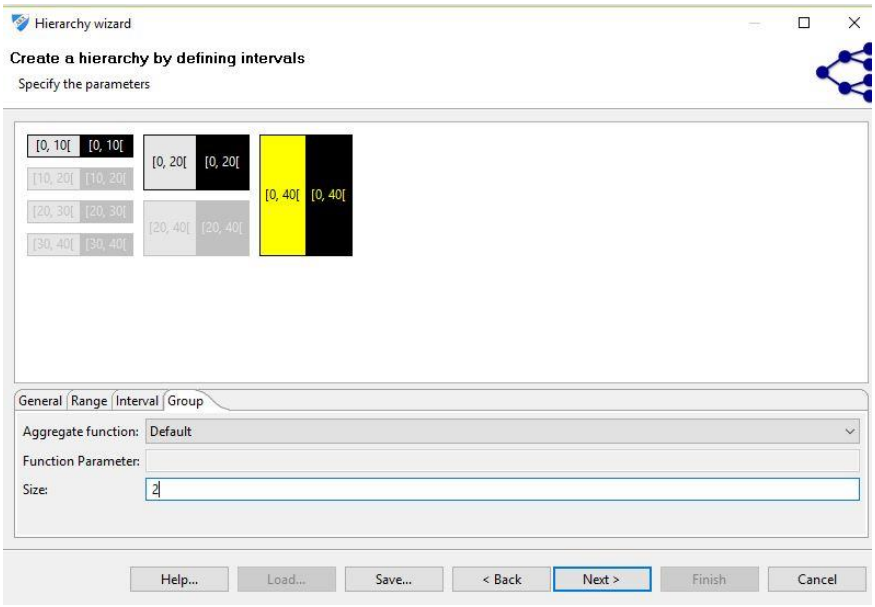

**Şekil 3.4:** Yaş Özniteliği İçin İkinci Seviyenin Eklenmesi

Açılan ekrandan "Group" sekmesinin altındaki "size" alanına "2" değeri girilerek yaşın bu katmanda aralık değerleri 10 ile sınırlandırılmıştır.



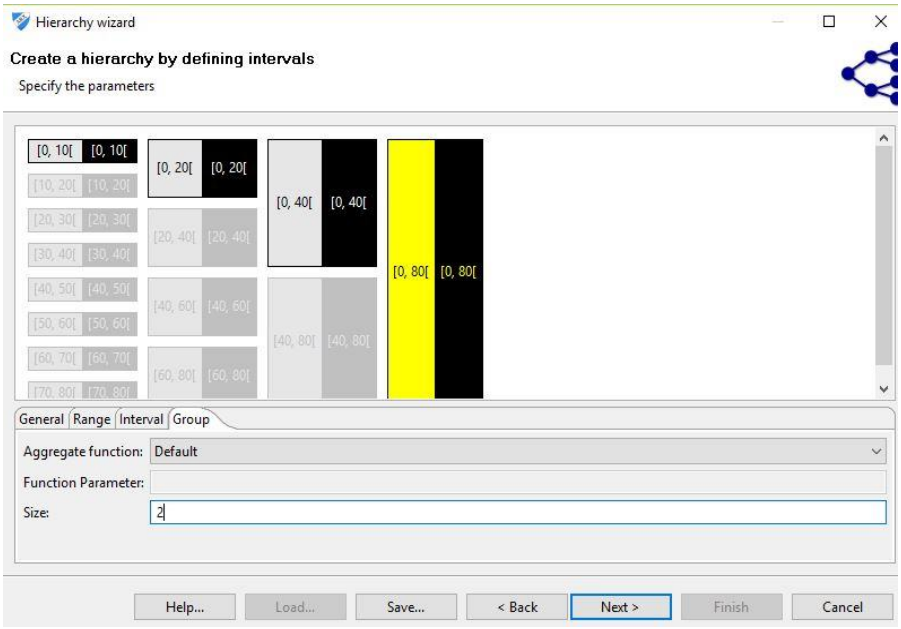

**Şekil 3.5:** Yaş Özniteliğinin İkinci Seviye İçin Aralık Değerinin 2 Seçilmesi

Ardından girilen değerlerin göründüğü kısma Mouse ile sağ tıklanarak "Add new level" seçilmiştir. Açılan ekrandan "Group" sekmesinin altındaki "size" alanına "2" değeri girilerek yaşın bu katmanda aralık değerleri 20 ile sınırlandırılmıştır.

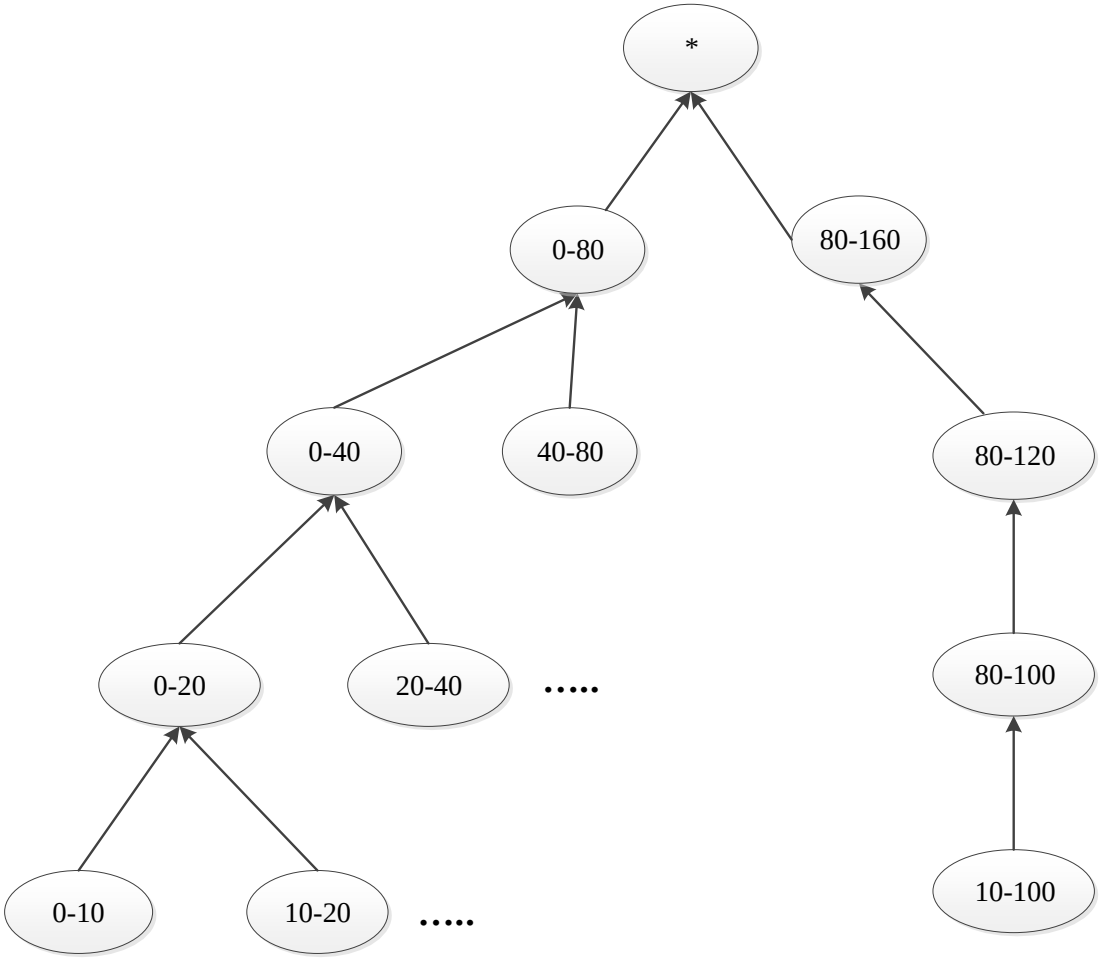



**Şekil 3.6:** Yaş Özniteliğinin Üçüncü Seviye İçin Aralık Değerinin 2 Seçilmesi

Üçüncü Seviyeyi tamamlamak için Mouse sağ tıklanarak "add after" sekmesi seçilmiştir. "Size" değeri 2 verilmiştir. Bu işlem maximum olarak girdiğimiz değere gelinceye kadar devam ettirilmiştir. "Next" diyerek bu kısım tamamlanmış olmaktadır.




**Şekil 3.7:** Yaş Özniteliğinin Üçüncü Seviye İçin Maksimum Değere Eşitlenmesi




**Şekil 3.8:** Yaş Özniteliğinin Taksonomi Ağacında Anonimleştirilmesi

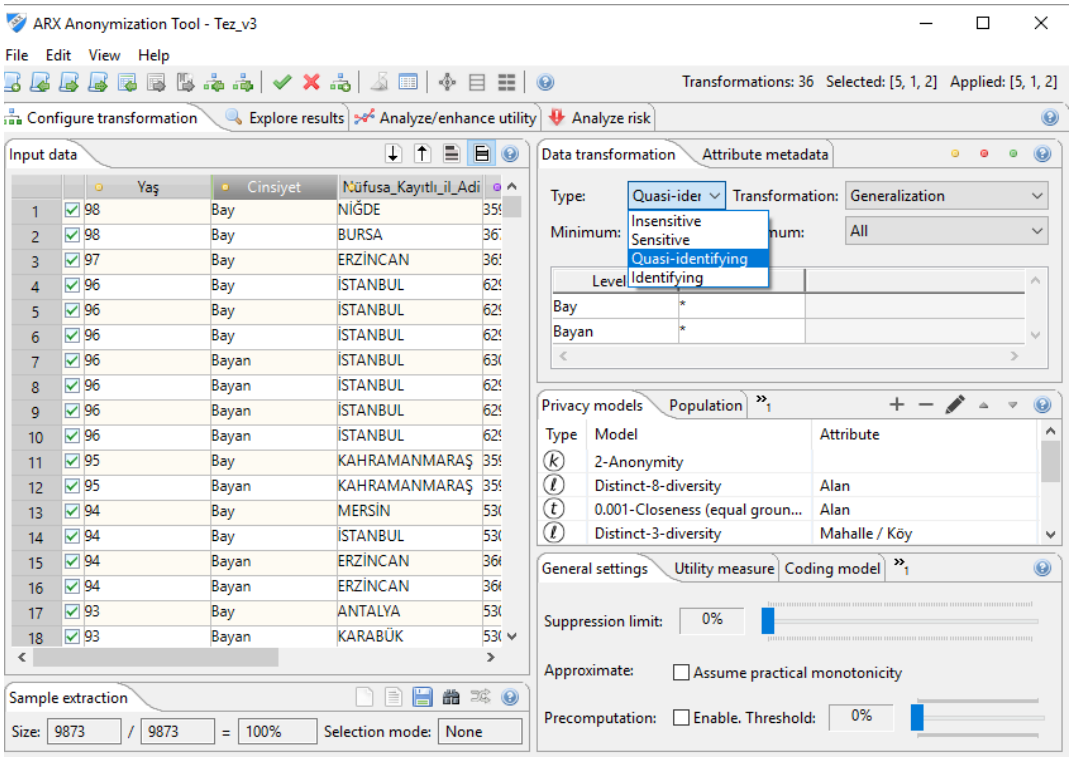
Şekil 3.8'de görüldüğü üzere test verisinin içinde 98 yaşında kişilerin olduğu görülmektedir. Bu nedenle ilk seviyede (0-10….90-100) olarak değerler anonimleştirilmiştir. İkinci seviyede (0-20….80-100) olarak anonimleştirilmiştir Üçüncü seviyede (0-40….80-120) olarak anonimleştirilmiştir. En son seviye olarak da (*) bastırılmıştır. Bu sayede 5 seviyeden oluşan bir anonimleştirme hiyerarşisi tanımlanmıştır.

### 3.2.2 Cinsiyet Özniteliğinin Yarı Tanımlayıcı Olarak Özellikleri

Cinsiyet alanı mülkiyet verisinde bulunan ve kişi bilgilerini kapsayan kısım içinde yer almaktadır.

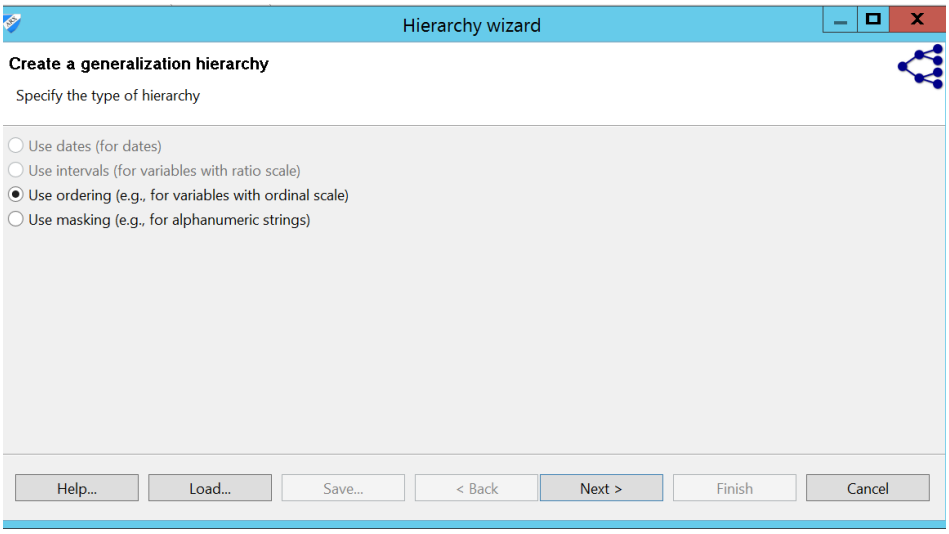
Veri seti içerisinde bulunan cinsiyet özniteliği yarı tanımlayıcı olarak seçilmiştir. Bu nedenle ARX programında cinsiyet özniteliği için hiyerarşi kurulmuştur. İlk olarak

ARX e yüklenen veri setinden cinsiyet öz niteliği seçilerek "Data Transformation" sekmesinin altında "Type" alanından "Quasi-identifying" seçilmiştir.



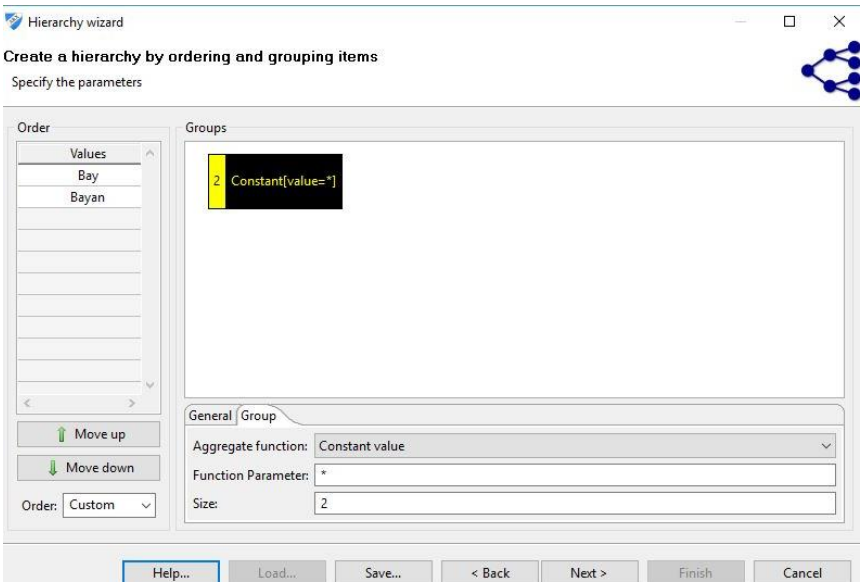

**Şekil 3.9:** Cinsiyet Özniteliğinin Yarı Tanımlayıcı (Quasi-İdentifying) Olarak Seçilmesi

Bu alanın seçilmesi ile "Edit" menüsü altında bulunan "create hierarchy" çalıştırılmıştır. Gelen ekrandan "use masking (e.g., for variables with ordinal scale) seçilerek "Next" tıklanmıştır.



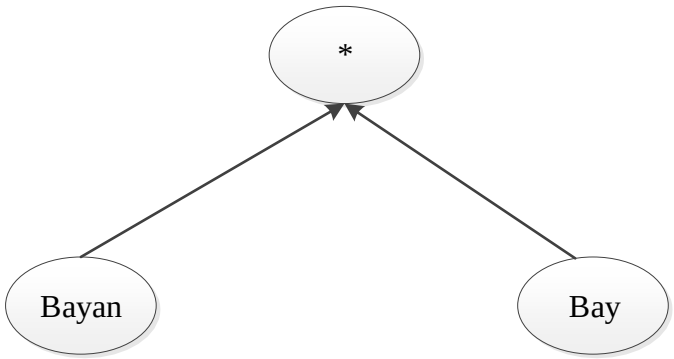

**Şekil 3.10:** Cinsiyet Özniteliği İçin Sıralı Hiyerarşinin Seçimi

Açılan ekrandan “Groups” altında gelen “1 Set” alanı seçilerek “Groups” altında bulunan “Aggregate function” alanından “Constant value” seçilmiştir. Ardından “Function Parameter” alanından “*” karakteri girilmiştir. Ardından “Size” alanına 2 girilmiştir.




**Şekil 3.11:** Cinsiyet Özniteliğinin (*) Karakterine Anonimleştirilmesi

Şekilde görüldüğü üzere test verimizin içinde bulunan cinsiyet alanı (*) yıldız şeklinde gösterime (Bastırılmış) anonimleştirilmiştir.

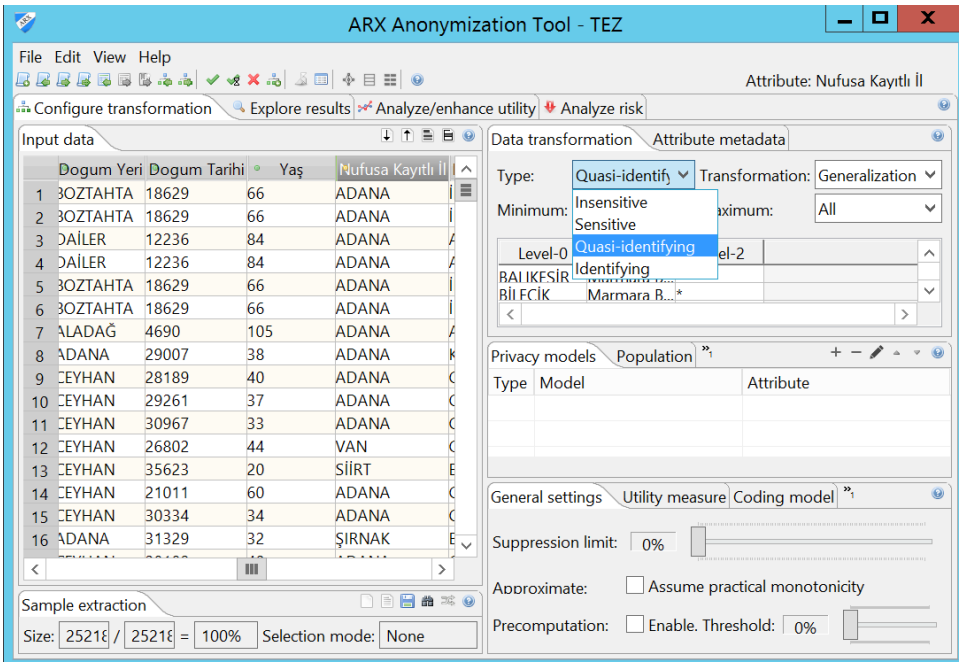



**Şekil 3.12**: Cinsiyet Özniteliğinin Anonimleştirilmesinin Taksonomi Ağacında Gösterimi

### 3.2.3 Nüfusa Kayıtlı İl Özniteliğinin Yarı Tanımlayıcı Olarak Özellikleri

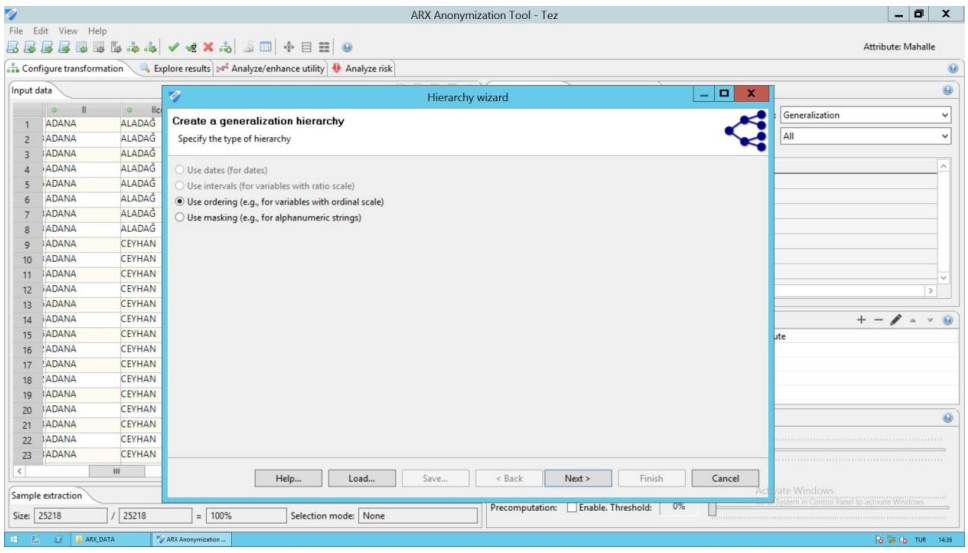
Nüfusa kayıtlı il özniteliği mülkiyet verisinde bulunan kişinin nüfusa kayıtlı il bilgisini içermektedir.

Veri seti içerisinde bulunan Nüfusa kayıtlı il özniteliği yarı tanımlayıcı olarak seçilmiştir. Bu nedenle ARX programında Nüfusa kayıtlı il için hiyerarşi kurulmuştur. İlk olarak ARX e yüklenen veri setinden Nüfusa kayıtlı il özniteliği seçilerek "Data Transfarmation" sekmesinin altında "Type" alanından "Quasi-identifying" seçilmiştir.



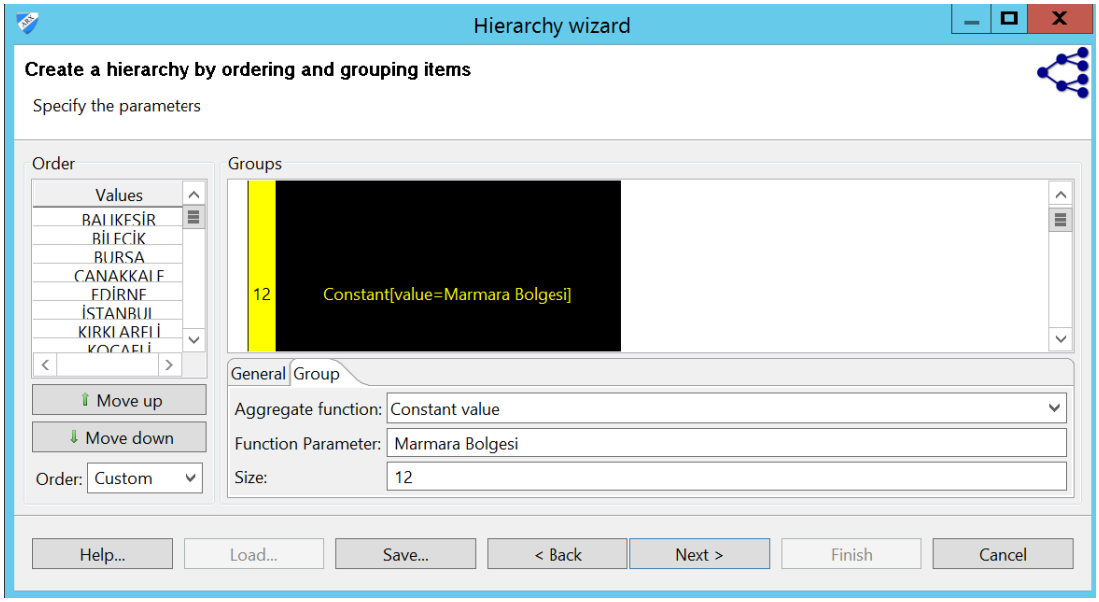

**Şekil 3.13:** Nüfusa Kayıtlı İl Özniteliğinin Yarı Tanımlayıcı (Quasi-İdentifying) Olarak Seçilmesi

Bu alanın seçilmesi ile "Edit" menüsü altında bulunan "create hierarchy" çalıştırılmıştır. Ekrana gelen ekrandan "use ordering (e.g., for variables with ordinal scale) seçilerek "Next" tıklanmıştır.



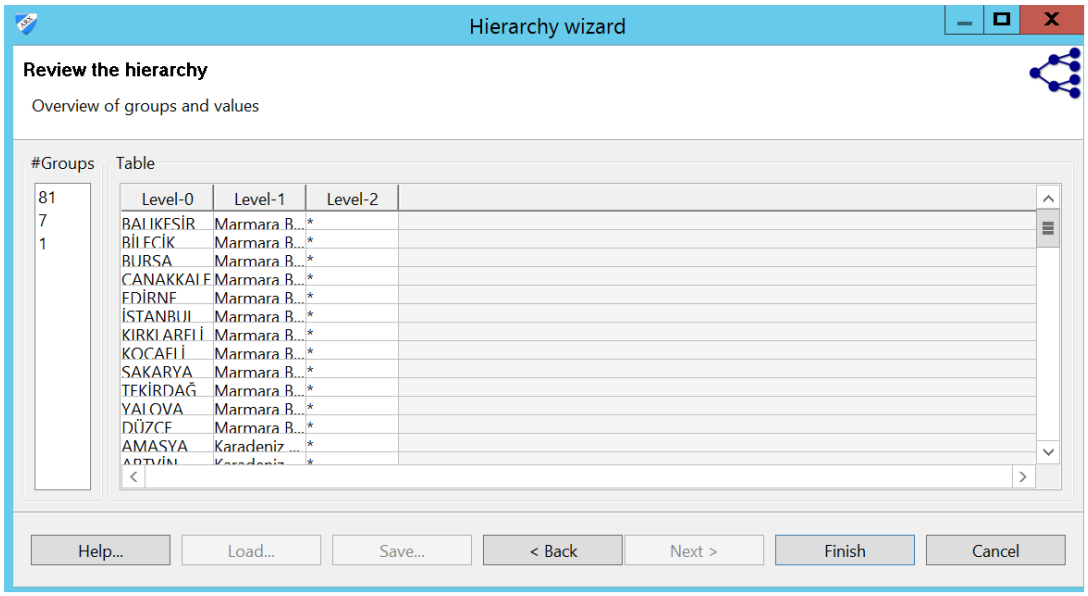

**Şekil 3.14:** Nüfusa Kayıtlı İl Özniteliği İçin Sıralı Hiyerarşinin Seçimi

Açılan ekrandan “Groups” altında gelen “1 Set” alanı seçilerek “Groups” altında bulunan “Aggregate function” alanından “Constant value” seçilmiştir.



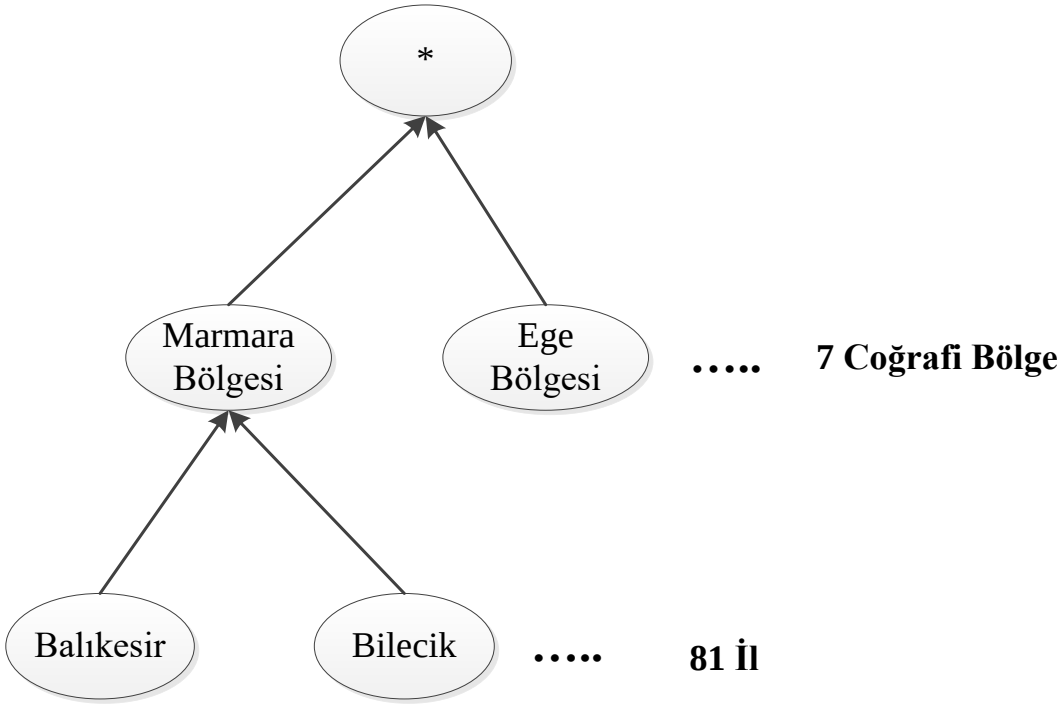

**Şekil 3.15:** Nüfusa Kayıtlı İl Özniteliğinin Anonimleştirilmesi

Ardından “Function Parameter” alanına Coğrafi bölgelerin isimleri girilmiştir ve bu bölgeye ait il sayısı “Size” alanına girilmiştir. Sistemin sağlıklı ve çalışabilir olması için soldaki il isimlerinin sırası girilecek coğrafi bölgelere göre sıralı olmalıdır. Veri setinde bu sıralama mevcut değildi. Bu nedenle veriler sıralanmıştır. Hiyerarşi kurulurken illeri bu sıraya uygun girmek gerekmektedir.




**Şekil 3.16:** Nüfusa Kayıtlı İl Özniteliğinin 2 Seviyede Anonimleştirilmesi

Şekilde görüldüğü üzere test verisi içinde bulunan Alan özniteliği ilk olarak coğrafi bölgelere anonimleştirilmiş ve ardından da (*) yıldız şeklinde bastırılmıştır.

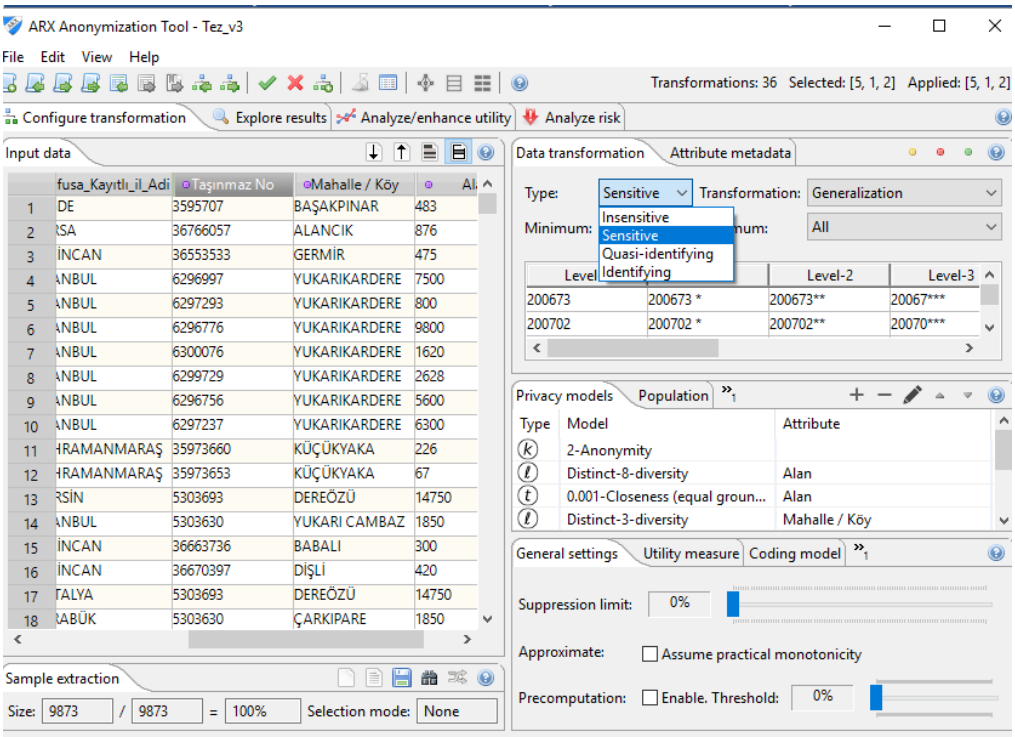



**Şekil 3.17:** Nüfusa Kayıtlı İl Özniteliğinin Anonimleştirilmesinin Taksonomi Ağacında Gösterimi

### 3.2.4 Taşınmaz No Özniteliğinin Hassas Değer Olarak Özellikleri

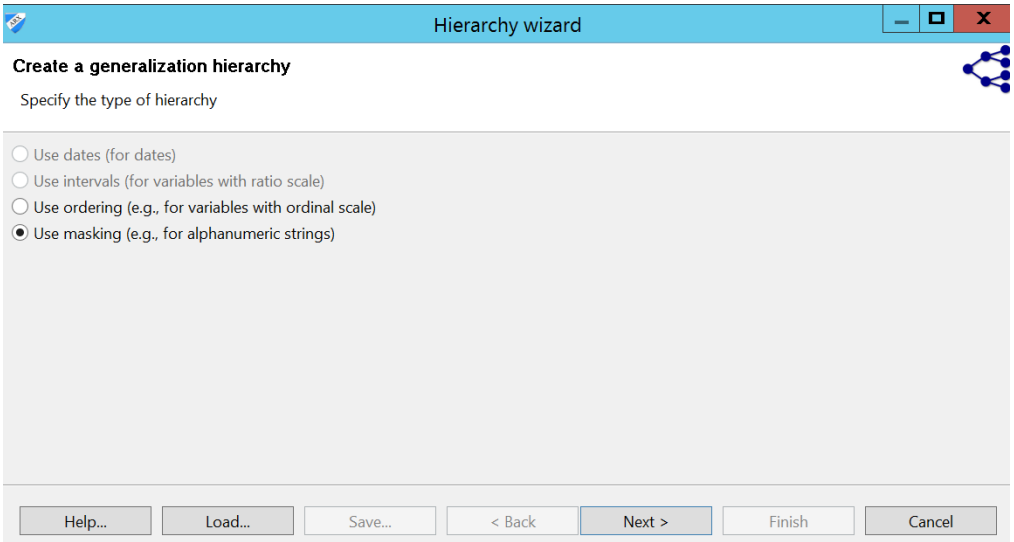
Taşınmaz No mülkiyet verisinde bulunan öznitelik bilgisidir. Bu öznitelik bilgisi 8 haneli benzersiz numaralardan oluşmaktadır. Veri seti üzerinde yapılan çalışmada tüm Taşınmaz no değerleri her seviyede sağdan başlayarak "*" yıldız şeklinde gösterime anonimleştirilmiştir. 8. Son seviyede (*) şeklinde bastırılmıştır.

Veri seti içerisinde bulunan Taşınmaz No özniteliği hassas olarak seçilmiştir. Bu nedenle ARX programında Taşınmaz No için hiyerarşi kurulmuştur. İlk olarak ARX' e yüklenen veri setinden Taşınmaz No özniteliği seçilerek "Data Transformation" sekmesinin altında "Type" alanından "Sensitive" seçilmiştir.



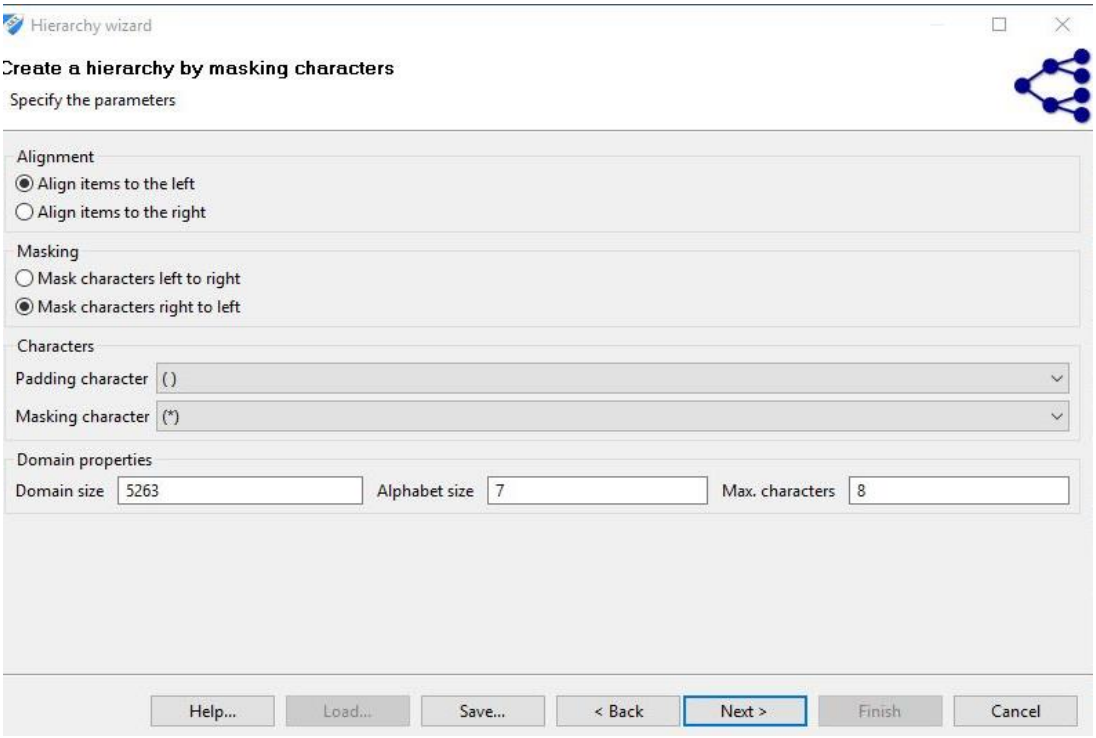

**Şekil 3.18:** Taşınmaz No Özniteliğinin Hassas (Sensitive) Olarak Seçilmesi

Bu alanın seçilmesi ile "Edit" menüsü altında bulunan "create hierarchy" çalıştırılmıştır. Ekrana gelen ekrandan "use masking (e.g., for alpnumeric string) seçilerek "Next" tıklanmıştır.



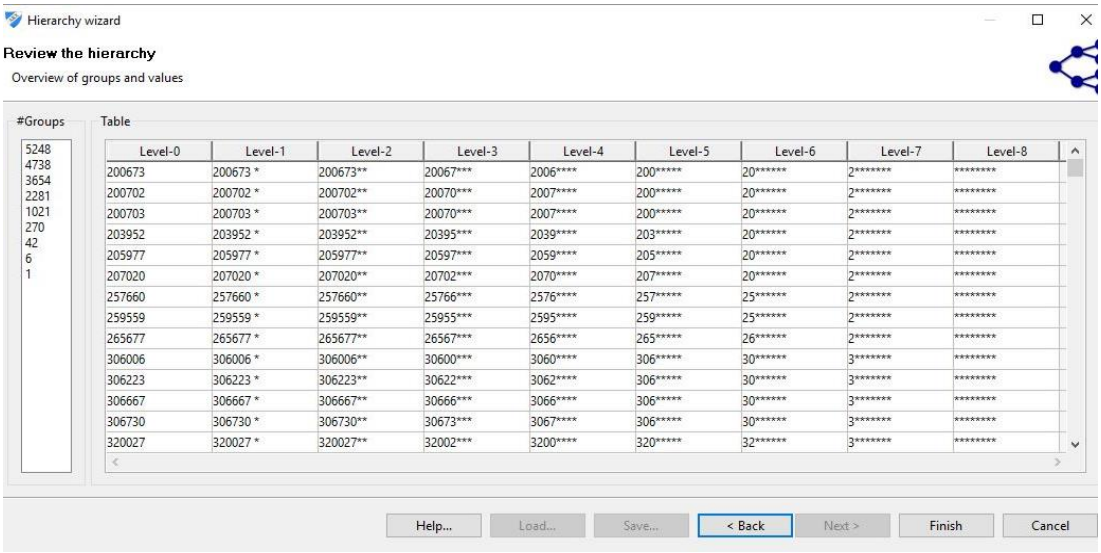

**Şekil 3.19:** Taşınmaz No Özniteliği İçin Maskeleme Hiyerarşinin Seçimi

Açılan ekran üzerinden "Masking" sekmesinin altında "Mask characters right to left" radio butonu işaretlenmiştir. "Characters" sekmesinin altında "Masking character" olarak "(*)" seçilmiştir.




**Şekil 3.20:** Taşınmaz No Özniteliğinin (*) Karakteri İle Anonimleştirilmesi

"Next" butonuna tıklandığında açılan ekran üzerinde Altı basamaktan oluşan Taşınmaz No değeri üzerinden sekiz seviyede sağdan sola doğru maskeleme yapıldığı görülmektedir.




**Şekil 3.21:** Taşınmaz No Özniteliğinin 8 Seviyede Anonimleştirilmesi

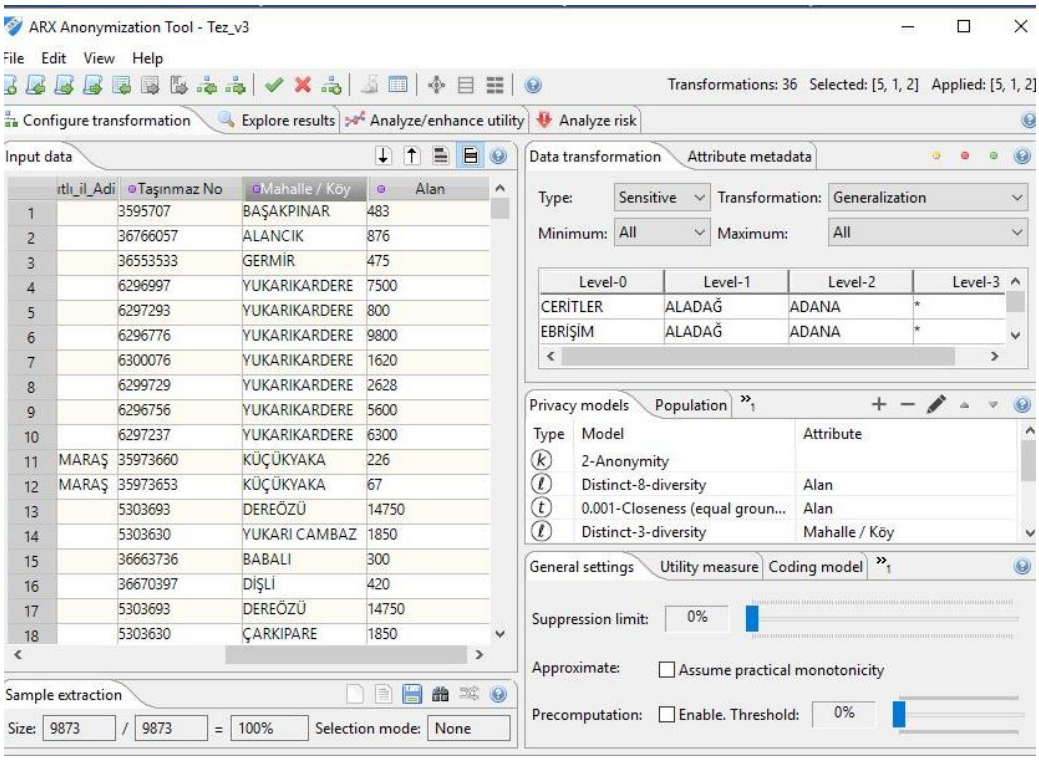
Şekilde görüldüğü üzere test verisi içinde bulunan Taşınmaz No alanı sekiz seviyede (*) yıldız şeklinde gösterime anonimleştirilmiştir.


**Şekil 3.22:** Taşınmaz No Özniteliğinin Anonimleştirilmesinin Taksonomi Ağacında Gösterimi

### 3.2.5 Mahalle / Köy Özniteliğinin Hassas Değer Olarak Özellikleri

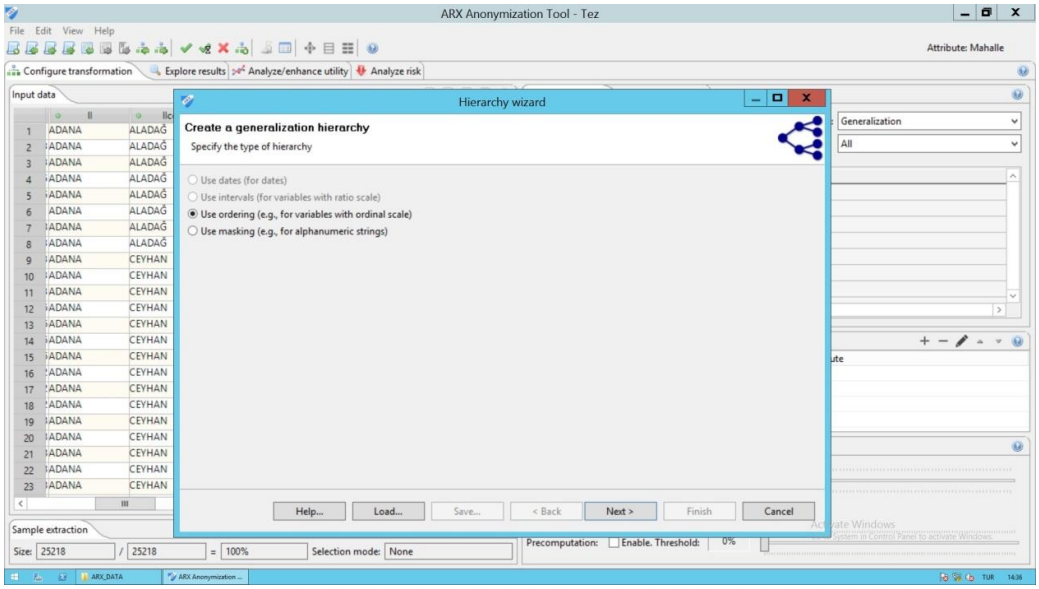
Mülkiyet verisi, içerisinde yer alan mahalle/köy özniteliği hassas olarak alınmıştır. Test veri setinde benzersiz 943 adet mahalle verisi bulunmaktadır. Verilerin karışık olmasından dolayı yaklaşık 9.873 satırdan oluşan veri içerisinde bu mahalleler sırası ile incelenerek ilçe ve il ile ilişkilendirilerek benzersizlik elde edilmiştir.

Veri seti içerisinde bulunan mahalle/köy özniteliği hassas olarak seçilmiştir. Bu nedenle ARX programında mahalle/köy den ilçeye oradan da il özniteliğine doğru hiyerarşi kurulmuştur. İlk olarak ARX e yüklenen veri setinden mahalle öz niteliği seçilerek “Data Transfarmation” sekmesinin altında “Type” alanından “Sensitive” seçilmiştir.



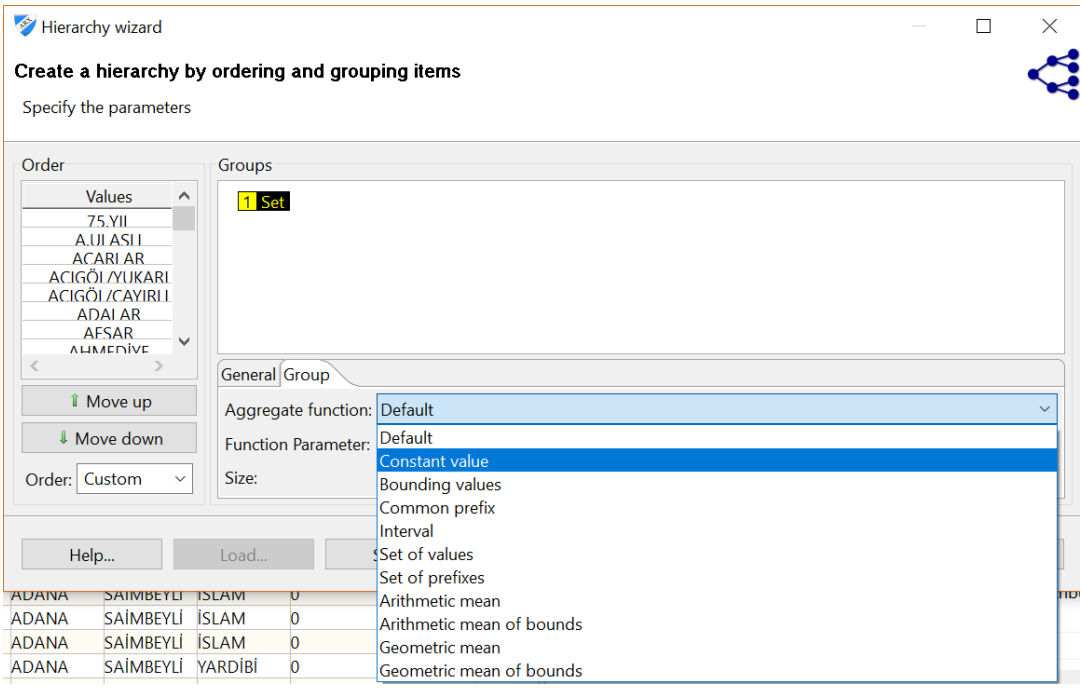

**Şekil 3.23:** Mahalle/Köy Özniteliğinin Hassas (Sensitive) Olarak Seçilmesi

Bu alanın seçilmesi ile "Edit" menüsü altında bulunan "create hierarchy" çalıştırılmıştır. Ekrana gelen menüden "use ordering (e.g., for variables with ordinal scale) seçilerek "Next" tıklanmıştır.



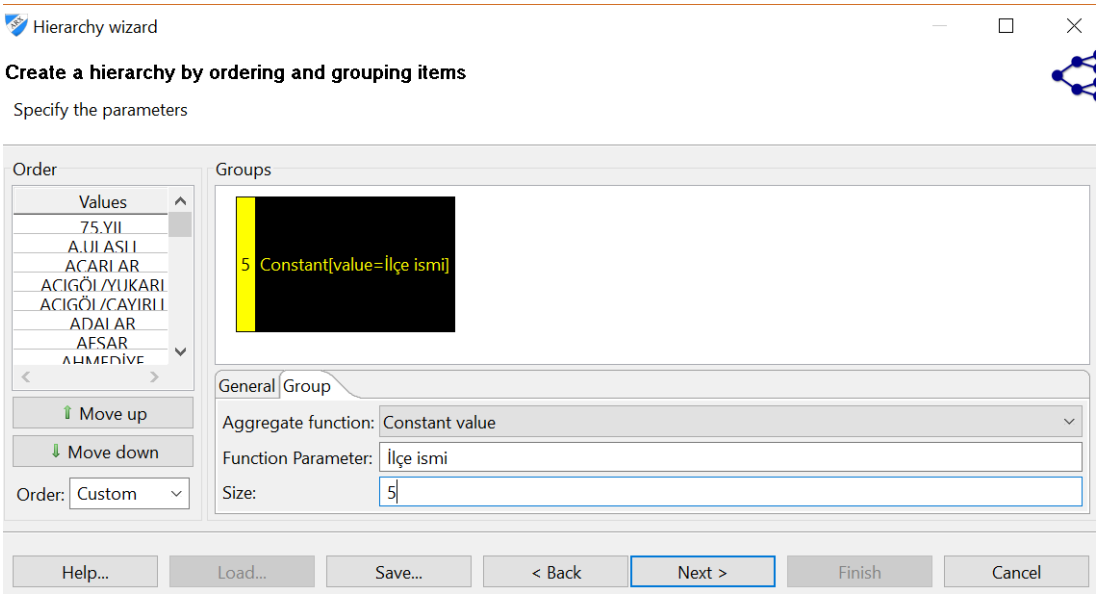

**Şekil 3.24:** Mahalle/Köy Özniteliği İçin Sıralı Hiyerarşinin Seçimi

Açılan ekrandan "Groups" altında gelen "1 Set" alanı seçilerek "Groups" altında bulunan "Aggregate function" alandan "Constant value" seçilmiştir.



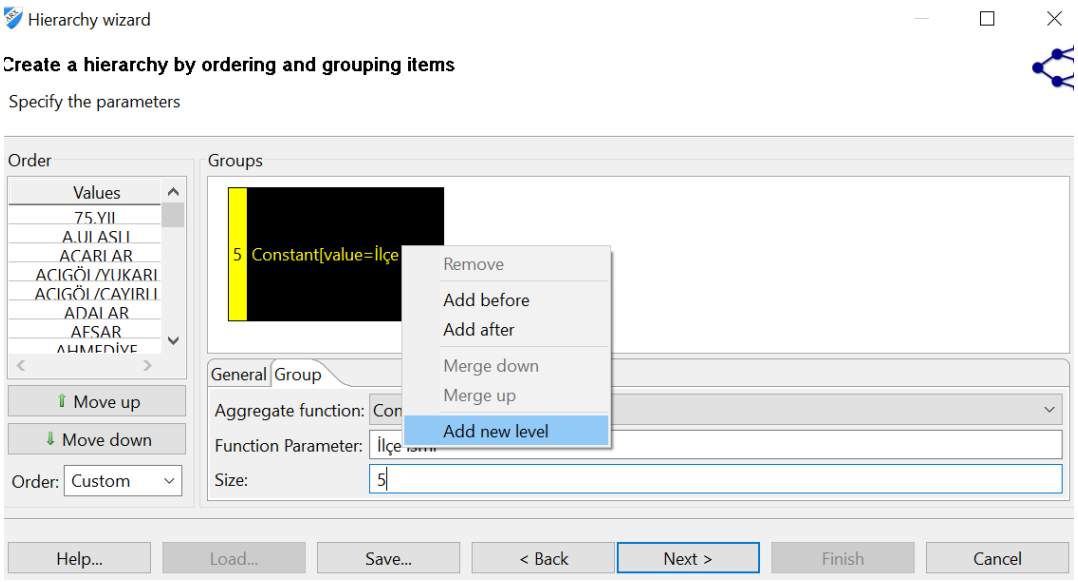

**Şekil 3.25:** Mahalle/Köy Özniteliği İçin Bir Üst Seviye Olan İlçeye Anonimleştirilmesi

Ardından "Function Parameter" alanına ilçenin ismi ve bu ilçeye ait mahalle sayısı da "Size" alanına girilmiştir. Sistemin sağlıklı ve çalışabilir olması için soldaki mahalle isimlerinin sırası gireceğiniz ilçeye ve ile göre sıralanmalıdır. Veri setinde bu sıralama mevcut bulunmamaktaydı. Bu nedenle veriler sıralanmıştır. Hiyerarşi kurulurken mahalle, ilçeler ve illeri bu sıraya uygun girmek gerekmektedir.



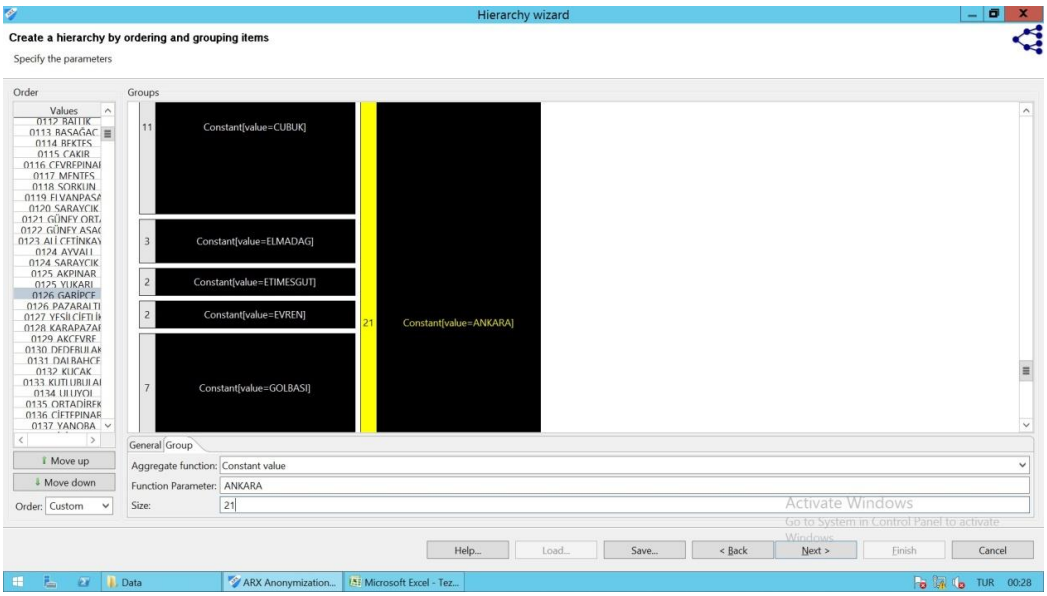

**Şekil 3.26:** Mahalle/Köy Özniteliğinin Birinci Seviye İçin Anonimleştirilmiş Görünümü

Bu şekilde girilen ilçeler bittikten sonra “Groups” altında ilçelerin en üst kısmına çıkılmıştır. Buradan ilk girilen ilçe ismin üstüne mause ile sağ tıklayarak açılan menüden “add new level” sekmesi işaretlenmiştir.

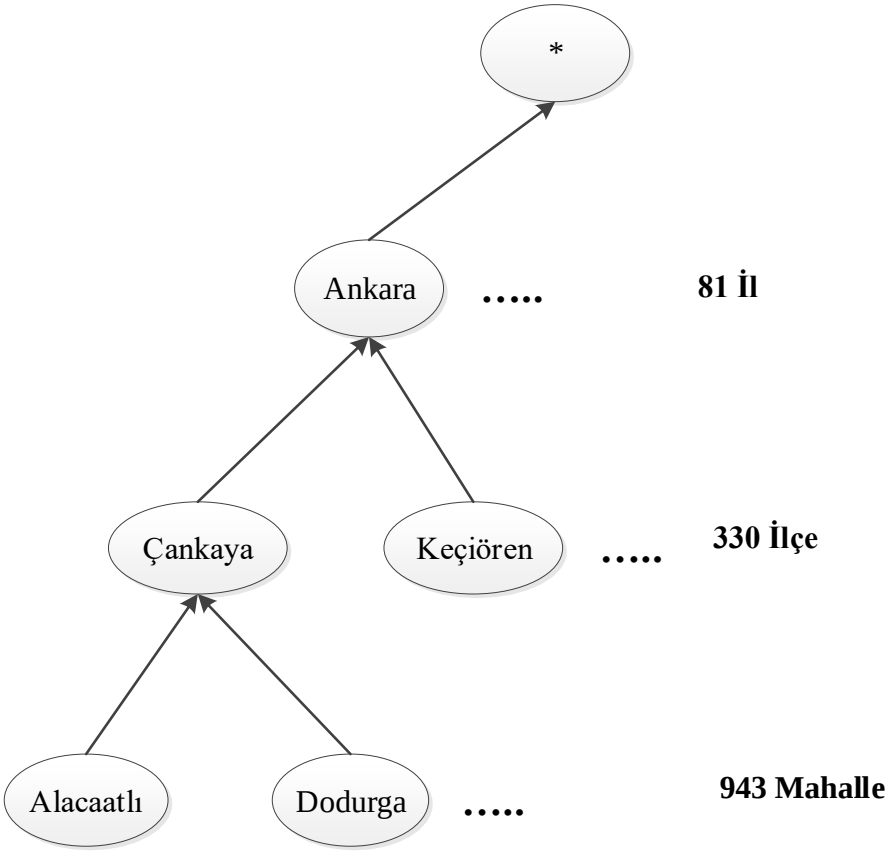



**Şekil 3.27:** İlçe Seviyesinin Bir Üst İl Seviyesine Anonimleştirilmesi

Böylece mahalle/köy den ilçe verisine anonimleştirilen veriler artık il seviyesine taşınmıştır. Açılan ekrandan “Groups” altında gelen “ilçe ismi” alanı seçilerek “Groups” altında bulunan “Aggregate function” alanından “Constant value” seçilmiştir. Ardından “Function Parameter” alanından il ismi ve bu il’e ait ilçe sayısı da “Size” alanına girilmiştir. Bütün işlemler bitirildikten sonra “Next” diyerek işlem bitirilmiştir.




**Şekil 3.28:** Mahalle/Köy Özniteliği İçin İl İsimlerinin Tamamlanmış Görünümü



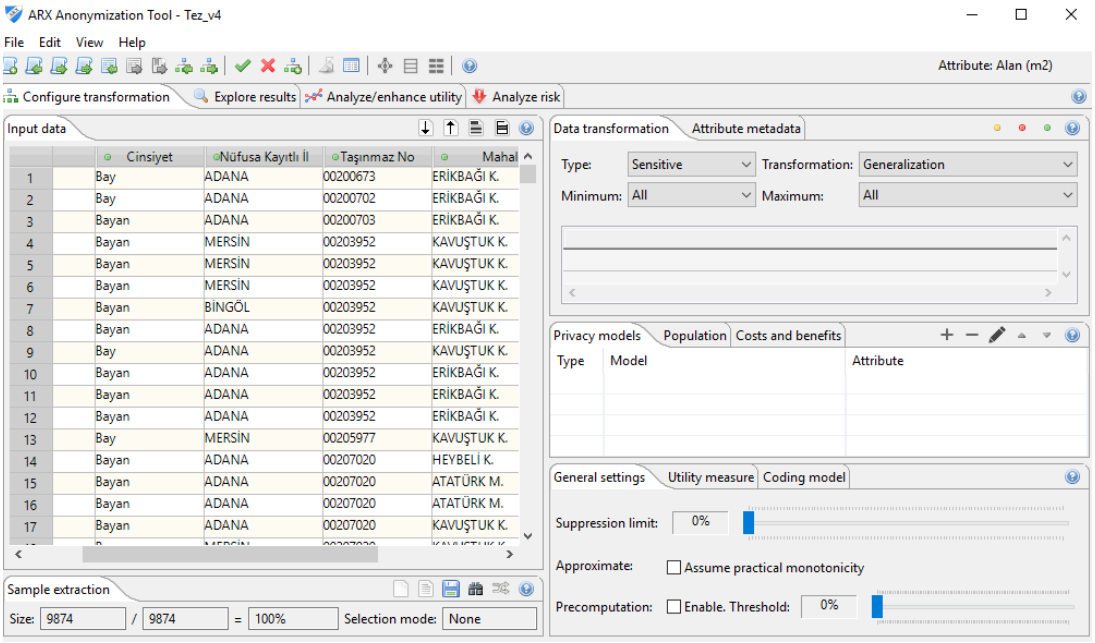

**Şekil 3.29:** Mahalle Özniteliğinin Anonimleştirilmesinin Taksonomi Ağacında Gösterimi

Şekil 3.29'de görüldüğü üzere test verisinin içinde 943 mahalle sırası ile bir üst veri tipi olan 330 ilçe verisine, oradan 81 il verisine ve son olarak da yıldız şeklinde gösterime anonimleştirilmiştir. Mahalle verisi 4. Seviyeye kadar anonimleştirilmiştir. En son seviyede * ile gösterim kullanılarak baskılama kullanılmıştır.

### 3.2.6 Alan özniteliğinin hassas değer olarak özellikleri

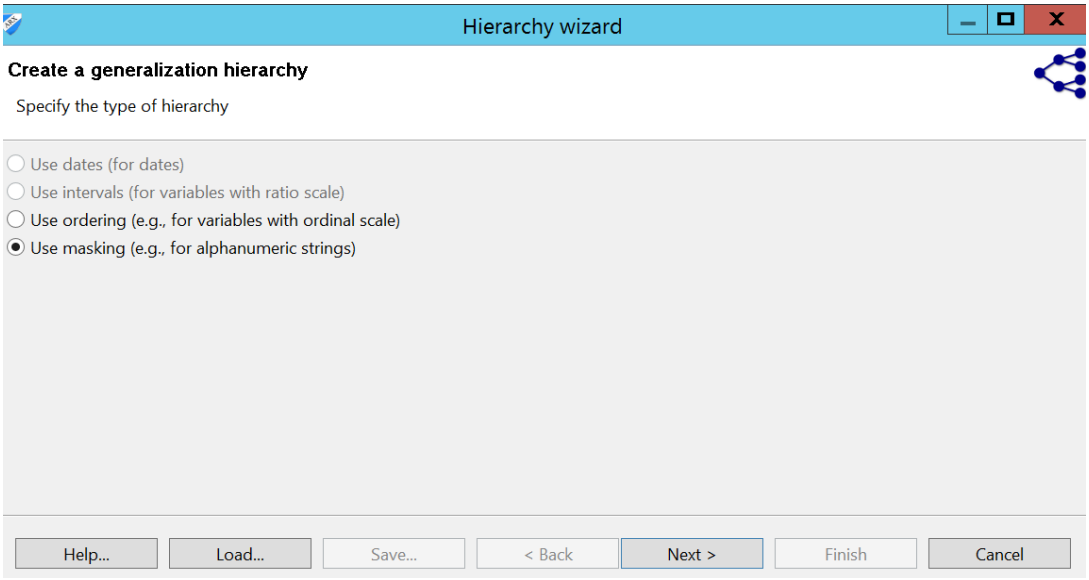
Alan özniteliği mülkiyet verisinde bulunan ve Parsellerin alan bilgisini metrekare cinsinden içermektedir.

Veri seti içerisinde bulunan Alan özniteliği hassas olarak seçilmiştir. Bu nedenle ARX programında Alan için hiyerarşi kurulmuştur. İlk olarak ARX' e yüklenen veri setinden Alan özniteliği seçilerek "Data Transfarmation" sekmesinin altında "Type" alanından "Sensitive" seçilmiştir.



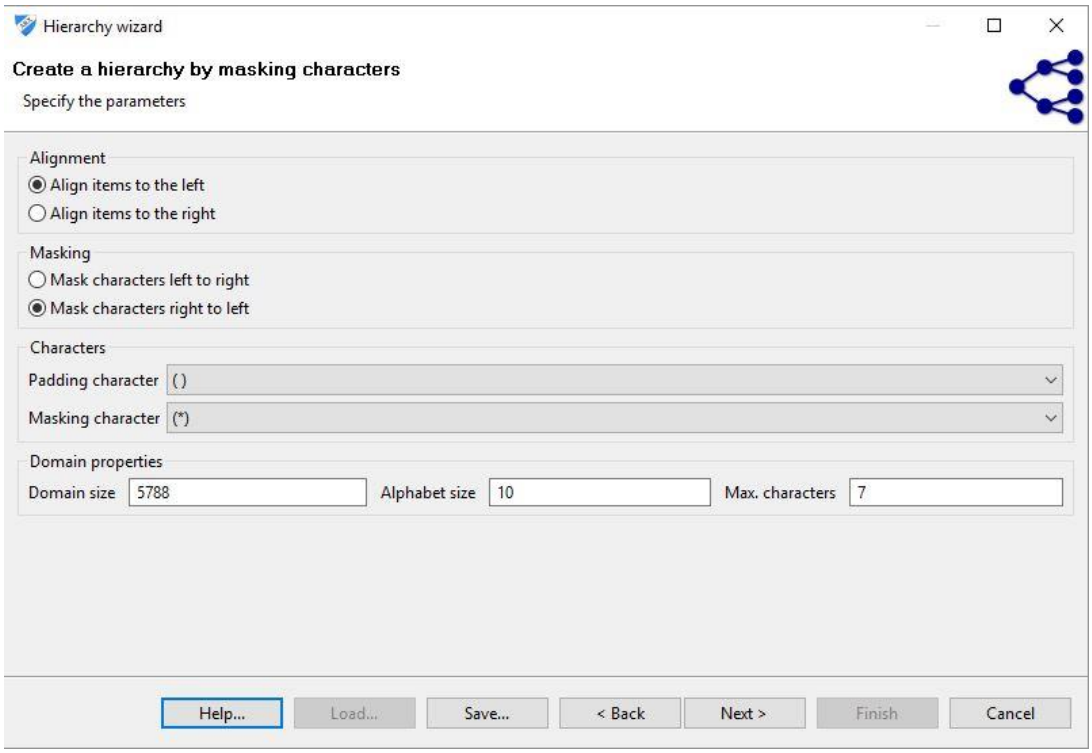

**Şekil 3.30:** Alan Özniteliğinin Hassas (Sensitive) Olarak Seçilmesi

Bu alanın seçilmesi ile "Edit" menüsü altında bulunan "create hierarchy" çalıştırılmıştır. Ekrana gelen ekrandan "use masking (e.g., for variables with ordinal scale) seçilerek "Next" tıklanmıştır.



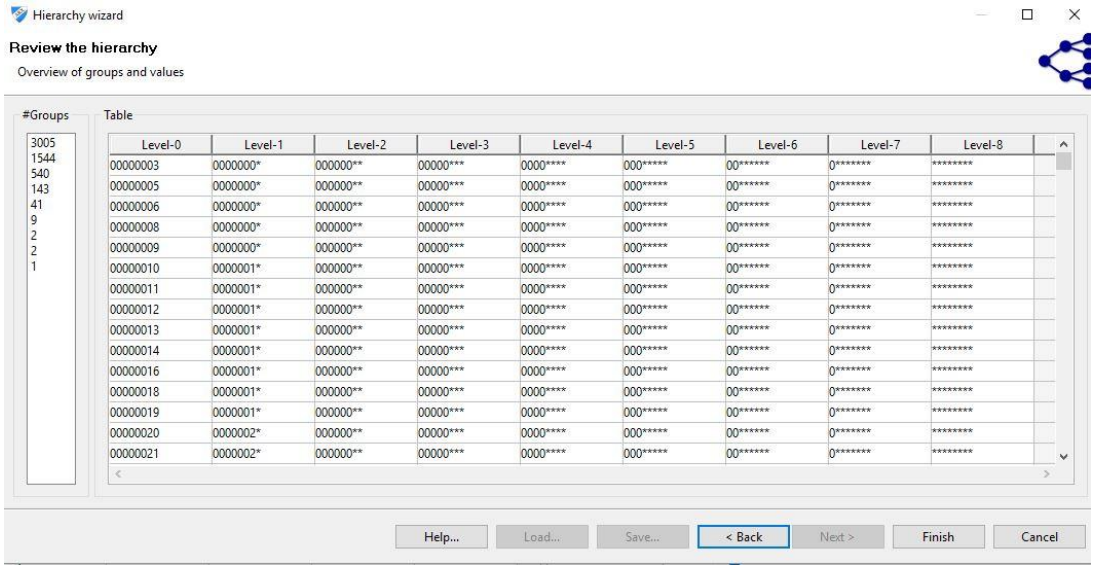

**Şekil 3.31:** Alan Özniteliği İçin Maskeleme Hiyerarşinin Seçimi

Açılan ekran üzerinden "Masking" sekmesinin altında "Mask charecters right to left" radio butonu işaretlenmiştir. "Charecters" sekmesinin altında "Masking charecter" olarak "(*)" seçilmiştir.



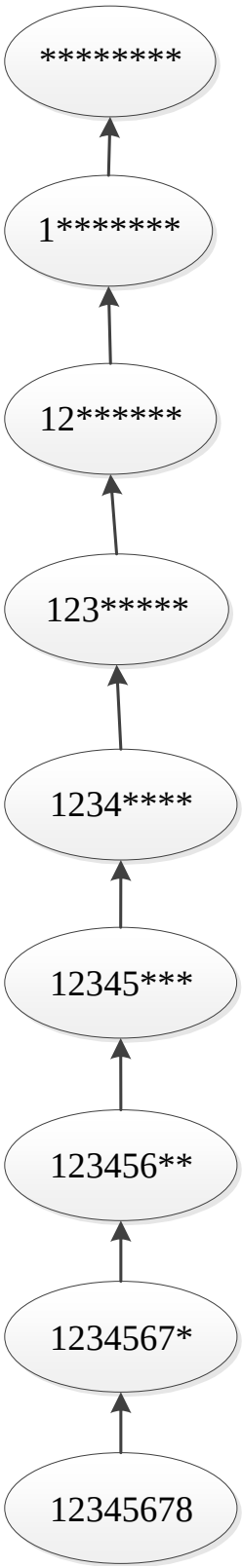

**Şekil 3.32:** İşlem Durum Özniteliğinin (*) Karakteri İle Anonimleştirilmesi

"Next" butonuna tıkladığımızda açılan ekran üzerinde yedi basamaktan oluşan Alan değeri üzerinden yedi seviyede sağdan sola doğru maskeleme yapıldığı görülmektedir.




**Şekil 3.33:** Alan Özniteliğinin 8 Seviyede Anonimleştirilmesi

Şekilde görüldüğü üzere test verisi içinde bulunan Alan özniteliği yedi seviyede (*) yıldız şeklinde gösterime anonimleştirilmiştir.




**Şekil 3.34:** Alan Özniteliğinin Anonimleştirilmesinin Taksonomi Ağacında Gösterimi

# DÖRDÜNCÜ BÖLÜM

# DENEYSEL ÇALIŞMALAR

Tez çalışmasının bu aşamasında yaygın olarak kullanılan gizlilik modellerinden k-anonimlik, $\ell$-çeşitlilik ve t-yakınlık ölçütleri kullanılarak savcı riski, gazeteci riski ve pazarlamacı riski arx yazılımı kullanılarak gösterilecektir. Çalışmanın bu aşamasında 9.873 kayıtlı ve Ülkemizin tamamını kapsayacak şekilde oluşturulan kayıtlar kullanılacaktır. Genelleştirme hiyerarşisi kolay anlaşılabilmesi için basit ve sıralı olarak aşağıda verilmiştir. Yaş, Cinsiyet, Nüfusa kayıtlı il, Taşınmaz no, Mahalle / Köy ve Tapu alanı özniteliklerinin veri kümesinin tamamına uygulanan genelleştirme hiyerarşileri aşağıda gösterilmektedir.

**Tablo 4.1:** Yaş Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

| Yaş | Seviye |
|---|---|
| * | 5 |
| [0, 80] [81, 160] | 4 |
| [0, 40] [41, 80] [81, 120] | 3 |
| [0, 20] [21, 40] [80, 100] | 2 |
| [0, 10] [11, 20] [21, 30] [31, 40] [90, 100] | 1 |
| 6…10……20…… 98 | 0 |

Kullanılan veri kümesi her yaştan bireyi kapsayabildiğinden yaş özniteliği minimum 6 ve maksimum 98 değerlerini kapsamaktadır. Başlangıçta özniteliğin kendi değerleri yer alırken, 1.Seviyede ardışık 10 değer, 2. Seviyede ardışık 20 değer, 3.

Seviyede ardışık 40 değer, 4. Seviyede ise grup hacmi ardışık 80 değere çıkarılmış 5. ve son seviye olarak (*) şeklinde gösterim ile baskılanmıştır.

**Tablo 4.2:** Cinsiyet Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

| Cinsiyet | Seviye |
|---|---|
| * | 1 |
| Bay Bayan | 0 |

Kullanılan veri kümesinde cinsiyet özniteliği Bay ve Bayan olmak üzere iki farklı değer aldığından başlangıç seviyesinde özniteliğin kendi değerleri yer alırken son seviye olarak (*) şeklinde gösterim ile baskılanmıştır.

**Tablo 4.3:** Doğum Yeri İl Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

| Nüfusa Kayıtlı İl | Seviye |
|---|---|
| * | 2 |
| [İç Anadolu] [Ege] [Doğu Anadolu] …… [Karadeniz] | 1 |
| Ankara… Konya……...Zonguldak | 0 |

Kullanılan veri kümesinde Nüfusa kayıtlı il özniteliği içinde iller yer aldığından 1. Seviye olarak coğrafi bölgelere genelleştirilmiş ve son seviye olarak (*) şeklinde gösterim ile baskılanmıştır.

**Tablo 4.4:** Taşınmaz No Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

| Taşınmaz No | Seviye |
|---|---|
| ******** | 8 |
| 1******* | 7 |
| 12****** | 6 |
| 123***** | 5 |
| 1234**** | 4 |
| 12345*** | 3 |
| 123456** | 2 |
| 1234567* | 1 |
| 12345678 | 0 |

Kullanılan veri kümesinde Taşımaz no özniteliği 8 basamaklı sayılardan meydana gelmektedir. Başlangıç seviyesinde özniteliğin kendi değerleri yer alırken 1. Seviye olarak sayını son hanesini (*) şeklinde gösterime genelleştirilmiş ve 7. Seviyeye kadar sürekli sağdan sola doğru (*) şekilleri artırılarak genelleştirilmeye devam edilmiştir. 8. ve son seviyede ise sayıların tamamı (*) gösterilerek baskılanmıştır.

**Tablo 4.5:** Mahalle / Köy Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

| Mahalle / Köy | Seviye |
|---|---|
| * | 3 |
| İl | 2 |
| İlçe | 1 |
| Mahalle | 0 |

Kullanılan veri kümesinde Mahalle / Köy özniteliği ülkemizde bulunan Mahalle ve köy isimlerinden meydana gelmektedir. Başlangıç seviyesinde özniteliğin kendi değerleri yer alırken 1. Seviye olarak Mahalle/Köy 'ün bağlı bulunduğu ilçeye genelleştirilmiş, 2. Seviye olarak ilçelerin bağlı olduğu illere genelleştirilmiş ve son seviye olarak Mahalle/Köy özniteliğinin tüm değerleri (*) gösterilerek baskılanmıştır.

**Tablo 4.6:** Tapu Alanı Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

| Tapu Alanı | Seviye |
|---|---|
| ******** | 8 |
| 1******* | 7 |
| 12****** | 6 |
| 123***** | 5 |
| 11234**** | 4 |
| 12345*** | 3 |
| 123456** | 2 |
| 1234567* | 1 |
| 12345678 | 0 |

Kullanılan veri kümesinde Tapu alanı özniteliği 6 basamaklı sayılardan meydana gelmektedir. Başlangıç seviyesinde özniteliğin kendi değerleri yer alırken 1. Seviye olarak sayının son hanesini (*) şeklinde gösterime genelleştirilmiş ve 5. Seviyeye kadar sürekli sağdan sola doğru (*) şekilleri artırılarak genelleştirilmeye devam edilmiştir. 6. ve son seviyede ise sayıların tamamı (*) gösterilerek baskılanmıştır.

Anonimleştirme işlemi öncesinde Belirlenen özniteliklere (Yaş, Cinsiyet, Doğum Yeri İl, Taşınmaz Numarası, Mahalle/Köy, Tapu alanı) uygulanan Tip seçimleri aşağıdadır.

**Tablo 4.7:** Belirlenen Özniteliklere Uygulanan Tip Seçimleri

| | |
|---|---|
| Yaş | Quasi-identifying |
| Cinsiyet | Quasi-identifying |
| Doğum Yeri İl | Quasi-identifying |
| Taşınmaz Numarası | Sensitive |
| Mahalle / Köy | Sensitive |
| Tapu Alanı | Sensitive |

Veri kümesinde yapılan anonimleştirme işlemi sırasında k-anonimliği için 2, 3, 4, 5, 10 ve 100 değerleri uygulanmıştır. $\ell$-çeşitliliği için hassas olarak seçilen Taşınmaz numarası ve alan öznitelikleri için maximum boyut (Genelleştirme Seviyesi) $\ell$=8, Mahalle/Köy özniteliği için işe $\ell$=3 uygulanmıştır. T-yakınlığı için ise standart olarak gelen t=0,01 değeri kullanılmıştır.

Yapılan bu çalışmada veri kümesinde bulunan öznitelikler yarı tanımlayıcı olarak seçilmiş ve uygulanan k-anonimliği modeli ile savcı, gazeteci ve market riski için değerleri gözlemlenecektir. Sağlıklı bir karşılaştırma işlemi olabilmesi için Uygulanan k-anonimliği değerlerine aynı işlemler aynı sıra ile yapılarak sonuçlar alındı.

Birinci test; Veri kümesinde yapılan bu genelleştirme işlemi ile savcı riski, gazeteci riski ve pazarlamacı riski için sırası ile 2, 3, 4, 5, 10 ve 100 değerleri uygulanmış ve aşağıdaki değerler alınmıştır.

**Tablo 4.8:** Belirlenen K-Anonimliği Değerleri İçin Gözlemlenen Savcı Riski, Gazeteci Riski ve Pazarlamacı Riski Değerleri

| **Belirlenen Risk Türleri** | | | |
|---|---|---|---|
| **k-anonimliği** | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 2 | 50,00 | 50,00 | 0,88 |
| 3 | 0,44 | 0,44 | 0,14 |
| 4 | 0,44 | 0,44 | 0,14 |
| 5 | 0,44 | 0,44 | 0,14 |
| 10 | 0,44 | 0,44 | 0,14 |
| 100 | 0,44 | 0,44 | 0,14 |

İkinci test; Veri kümesi içinde hassas öznitelik olarak seçilen alanların teker teker ele alınarak) $\ell$ çeşitlilik ve t yakınlık için belirlenen değerlerin uygulanarak savcı riski, gazeteci riski ve pazarlamacı riskinin için aşağıdaki değerler alınmıştır.

Taşınmaz no özniteliğinin hassas olarak seçilmesi ve $\ell$=1, $\ell$=2, $\ell$=3, $\ell$=4, $\ell$=5, $\ell$=6, $\ell$=7, $\ell$=8 çeşitliliği ve t=1 t=0,1 t=0,01 t=0,001 yakınlığı için alınan sonuç değerleri;

**Tablo 4.9:** Belirlenen $\ell$ -Çeşitlilik Değerleri İçin Gözlemlenen Savcı Riski, Gazeteci Riski Ve Pazarlamacı Riski Değerleri

| **Taşınmaz No** | | | |
|---|---|---|---|
| $\ell$ çeşitlilik | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 2 | 0,4355 | 0,4355 | 0,4355 |
| 3 | 0,1418 | 0,1418 | 0,1418 |
| 4 | 0,1418 | 0,1418 | 0,1418 |
| 5 | 0,1418 | 0,1418 | 0,1418 |
| 6 | 0,1418 | 0,1418 | 0,1418 |
| 7 | 0,1418 | 0,1418 | 0,1418 |
| 8 | 0,1418 | 0,1418 | 0,1418 |

**Tablo 4.10:** Belirlenen T-Yakınlığı Değerleri İçin Gözlemlenen Savcı Riski, Gazeteci Riski ve Pazarlamacı Riski Değerleri

| **Taşınmaz No** | | | |
|---|---|---|---|
| t-yakınlığı | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 1 | 0,1621 | 0,1621 | 0,1621 |
| 0,1 | 0,0101 | 0,0101 | 0,0101 |
| 0,01 | 0,0101 | 0,0101 | 0,0101 |
| 0,001 | 0,0101 | 0,0101 | 0,0101 |

Mahalle/Köy özniteliğinin hassas olarak seçilmesi ve belirlen ve $\ell$=1, $\ell$=2, $\ell$=3 çeşitliliği ve t=1 t=0,1 t=0,01 t=0,001yakınlığı değerler için alınan sonuçlar;

**Tablo 4.11:** Belirlenen $\ell$-Çeşitlilik İçin Gözlemlenen Savcı Riski, Gazeteci Riski Ve Pazarlamacı Riski Değerleri

| **Mahalle/Köy** | | | |
|---|---|---|---|
| $\ell$ çeşitlilik | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 2 | 0,4355 | 0,4355 | 0,4355 |
| 3 | 0,1418 | 0,1418 | 0,1418 |

**Tablo 4.12:** Belirlenen T-Yakınlığı İçin Gözlemlenen Savcı Riski, Gazeteci Riski Ve Pazarlamacı Riski Değerleri

| **Mahalle/Köy** | | | |
|---|---|---|---|
| t-yakınlığı | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 1 | 0,8812 | 0,8812 | 0,8812 |
| 0,1 | 0,0101 | 0,0101 | 0,0101 |
| 0,01 | 0,0101 | 0,0101 | 0,0101 |
| 0,001 | 0,0101 | 0,0101 | 0,0101 |

Alan (m2) özniteliğinin hassas olarak seçilmesi ve $\ell$=1, $\ell$=2, $\ell$=3, $\ell$=4, $\ell$=5, $\ell$=6, $\ell$=7, $\ell$=8 çeşitliliği ve t=1 t=0,1 t=0,01 t=0,001 yakınlığı için alınan sonuç değerleri;

**Tablo 4.13:** Belirlenen $\ell$-Çeşitlilik İçin Gözlemlenen Savcı Riski, Gazeteci Riski Ve Pazarlamacı Riski Değerleri

| **Alan (m2)** | | | |
|---|---|---|---|
| $\ell$ çeşitlilik | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 2 | 0,4355 | 0,4355 | 0,4355 |
| 3 | 0,1418 | 0,1418 | 0,1418 |
| 4 | 0,1418 | 0,1418 | 0,1418 |
| 5 | 0,1418 | 0,1418 | 0,1418 |
| 6 | 0,1418 | 0,1418 | 0,1418 |
| 7 | 0,1418 | 0,1418 | 0,1418 |
| 8 | 0,1418 | 0,1418 | 0,1418 |

**Tablo 4.14:** Belirlenen T-Yakınlığı İçin Gözlemlenen Savcı Riski, Gazeteci Riski Ve Pazarlamacı Riski Değerleri

| **Alan (m2)** | | | |
|---|---|---|---|
| t-yakınlığı | **Savcı Risk** | **Gazeteci Risk** | **Pazarlamacı Risk** |
| 1 | 0,2937 | 0,2937 | 0,2937 |
| 0,1 | 0,0101 | 0,0101 | 0,0101 |
| 0,01 | 0,0101 | 0,0101 | 0,0101 |
| 0,001 | 0,0101 | 0,0101 | 0,0101 |

# BEŞİNCİ BÖLÜM

# TARTIŞMA

Bu tez çalışması sırasında mülkiyet verileri çerçevesinde birden fazla mahremiyet korumalı yaklaşımlar özetlenerek değerlendirilmiştir. Mülkiyet verilerinin paylaşılması sırasında; mahremiyet korumalı yaklaşımların, meydana gelebilecek problemlerin çözülmesinde rol alabileceği düşünülmüştür. Mülkiyet verilerinin paylaşılmasında seçilen öznitelikler ile mahremiyetin korunması noktasında maksimum fayda sağlayarak ideal bir çözüm önerilmektedir.

Önerilen bu çözümle hesaplanan riskler dikkate alınarak faydanın en üst seviyede kalması düşünülmüştür. Seçilen öznitelikler en fazla kullanılanlardır. Yapılan çalışmada risk analizi sırasında seçilen tüm özniteliklerin risk analizlerinin hesaplanması noktasında eşit ağırlığa sahip olduğu görülmüştür. Fakat bazı özniteliklerin mahremiyeti ortadan kaldırma noktasında diğerlerinden farklı olarak hesaplanması gerektiği görülmüştür. Bu nedenle bazı özniteliklerin hesaplanmasında ağırlık puanlaması yapılarak risk analizi yapılması yönünde bir çalışmanın ileriki aşamalarsa yapılması düşünülmektedir.

# ALTINCI BÖLÜM

# SONUÇ

Çeşitli kamu kurum ve kuruluşları yapacakları çalışmalar adına, mülkiyet verilerinin anonimleştirilerek mahremiyet korumalı hale getirilmesine örnek olması için bu çalışma yapılmıştır. Mahremiyet korumalı veri yayıncılığı ve mahremiyet korumalı veri madenciliği üzerine yapılan detaylı araştırmalar üzerine ilgili kamu kurumlarına örnek teşkil etmesi için bu çalışma gerçekleştirilmiştir.

Yapılan bu çalışma öncelikle literatürde bulunan sağlık verileri üzerine yapılan araştırmalar incelenerek mahremiyet korumalı mülkiyet verilerinin yayınlanması ve mahremiyet korumalı veri madenciliği yapılması ele alınmıştır. Literatürde yapılan çalışmalar dan esinlenerek yapılan bu çalışmada anonimleştirme ve fayda analizleri doğrultusunda belirlenen üç farklı riskin analizi yapılmıştır. Literatürde en fazla kullanılan savcı riski, gazeteci riski ve pazarlamacı riskleri değerlendirilerek sonuçlar yayınlanmıştır.

Bu çalışma neticesinde mülkiyet verilerinde bulunan farklı öznitelikler incelenerek bunlar üzerinde anonimleştirme teknikleri uygulanmıştır. Çalışma sırasında farklı özniteliklere uygulanan farklı anonimleştirme teknikleri gerçek hayatta, kurum ve kuruluşların uygulama sırasında yaralanabilecekleri bir örnek model olması için tasarlanmıştır.

Örnek model baz alınarak yapılan risk analizleri sonuçlarına göre belirlediğimiz özniteliklerde anonimlik modeli uygulanmadan önce savcı, gazeteci ve pazarlamacı risk oranı için 98,43 dür. Aynı risk çeşitleri için k=2 olduğunda riskin 50 ‘ye düştüğü ve k=3 olduğunda 0,44 olduğu ve daha sorra k'nın değeri değişse de risk miktarının değişmediği görülmüştür.

Hassas özniteliklerden Taşınmaz no için $\ell$-çeşitlilik modeli uygulandığında savcı, gazeteci ve pazarlamacı riski $\ell$=2 değeri için 0,4355 olurken $\ell$=3 değeri için 0,1418 olduğu ve sonrasında $\ell$ değerindeki değişimlere karşı duyarsız olduğu

gözlemlenmiştir. Aynı risk türleri için t=1 değeri için 0,1621 olduğu ve t=0,1 değeri için ise 0,0101 olduğu ve daha sonraki t değerleri için sonucun değişmediği ve duyarsız olduğu görülmüştür.

Mahalle/Köy özniteliği için $\ell$-çeşitlilik modeli uygulandığında savcı, gazeteci ve pazarlamacı riski $\ell$=2 değeri için 0,4355 olurken $\ell$=3 değeri için 0,1418 olduğu ve sonrasında $\ell$ değerindeki değişimlere karşı duyarsız olduğu gözlemlenmiştir. Aynı risk türleri için t=1 değeri için 0,8812 olduğu ve t=0,1 değeri için ise 0,0101 olduğu ve daha sonraki t değerleri için sonucun değişmediği ve duyarsız olduğu görülmüştür.

Alan özniteliği için $\ell$-çeşitlilik modeli uygulandığında savcı, gazeteci ve pazarlamacı riski $\ell$=2 değeri için 0,4355 olurken $\ell$=3 değeri için 0,1418 olduğu ve sonrasında $\ell$ değerindeki değişimlere karşı duyarsız olduğu gözlemlenmiştir. Aynı risk türleri için t=1 değeri için 0,2937 olduğu ve t=0,1 değeri için ise 0,0101 olduğu ve daha sonraki t değerleri için sonucun değişmediği ve duyarsız olduğu görülmüştür.

Ayrıca (3.2.4. ile 3.2.6. gibi) alan ve taşınmaz no özniteliklerine uygulanan anonimleştirme metotlarının benzer olması neticesinde risk analizi sonuçlarının aynı kaldığı görülmektedir. Alan ve taşınmaz no özniteliğine uygulanan savcı, gazeteci ve pazarlamacı risk analizlerinde $\ell$-çeşitlilik modelinin 2. seviye ile birlikte 0,1418 değeri aldığı ve sonrasında değerlerin değişmediği gözlenmektedir.

Yapılan analizler ışığında; alan ve taşınmaz numarası gibi sayılardan oluşan ve aynı anonimleştirme tekniklerinin uygulanması neticesinde, risk analizlerinde aynı oranda sonuca etki etmelerine rağmen uygulamada taşınmaz numarasının farklı olarak değerlendirilmesi gerektiği düşünülmektedir. Web de çalışan bazı uygulamalara taşınmaz numarası girildiğinde diğer özniteliklere doğrudan ulaşma imkânı vardır. Bu nedenle tek başına öznitelikleri değerlendirmek yerine uygulamalardaki kullanım alanlarını dikkate alarak anonimleştirme yapılması ve bunların risk analizlerininde farklı değerlendirme ölçütleri kullanılarak yapılmasının uygun olacağı kanaatine varılmıştır.

Ülkemizde mülkiyet verisi yayınlama noktasında söz sahibi olan kurumlar tarafında da kullanılabilecek bu anonimleştirme metotları ve belirlenen risklere göre yapılan analizler bilimsel çalışmaların yürütülmesine de katkı sağlayacaklardır.

# KAYNAKÇA


[1] A. Korolova, “Protecting Privacy When Mining and Sharing User Data,” Standford Univ., no. August, 2010.

[2] https://www.techworld.com/security/uks-most-infamous-data-breaches-3604586/ (01.01.2018).

[3] https://www.trthaber.com/haber/bilim-teknoloji/yahoo-mail-skandalini-dogruladi-301852.html (01.01.2018).

[4] https://www.theguardian.com/business/2016/nov/08/tesco-bank-cyber-thieves-25m (01.01.2018).

[5] https://www.Databreaches.Net/Turkish-Citizenship-Database-Leak/ (01.01.2018).

[6] http://www.ufukotesi.com/yazigoster.asp?yazi_no=20080818 (01.01.2018).

[7] http://Resources.Infosecinstitute.Com/2018-Cyber-Security-Predictions/ (01.01.2018).

[8] http://www.milliyet.com.tr/siyaset/bakanlik-sinava-girecekleri-desifre-etti-adaylar-isyan-etti-1324192 (01.01.2018).

[9] https://www.kvkk.gov.tr/yayinlar/KİŞİSEL VERİLERİN SİLİNMESİ, YOK EDİLMESİ VEYA ANONİM HALE GETİRİLMESİ REHBERİ.pdf. (01.01.2018).

[10] T.-P. Hong, K.-T. Yang, C.-W. Lin, and S.-L. Wang, “Evolutionary privacy-preserving data mining,” in World Automation Congress (WAC), 2010.

[11] R. N. Wright, Z. Yang, and S. Zhong, “Distributed Data Mining Protocols for Privacy: A Review of Some Recent Results,” in Secure Mobıle Ad-Hoc Networks And Sensors, vol. 4074, 67–79, 2006.

[12] D. L. Chaum, “Untraceable electronic mail, return addresses, and digital pseudonyms,” Commun. ACM, vol. 24, no. 2, 84–90, 1981.

[13] S. L. Warner, “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias,” J. Am. Stat. Assoc., vol. 60, no. 309, 63–69, 1965.

[14] California Office of Statewide Health Planning and Development, “California Inpatient Data Reporting Manual, Medical Information Reporting For California, Seventh Edition,” in Office of Statewide Health Planning and Development., no. September, 23, 2014,

[15] K. El Emam, “Data Anonymization Practices in Clinical Research,” Heal. (San Fr., 1–15, 2006.

[16] L. Sweeney, “k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,” Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 05, 557–570, 2002.

[17] B. C. M. Fung, K. Wang, and P. S. Yu, “Anonymizing classification data for privacy preservation,” IEEE Trans. Knowl. Data Eng., vol. 19, no. 5, 711–725, 2007.

[18] M. E. Nergiz, C. Clifton, and A. E. Nergiz, “Multirelational k-anonymity,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 8, 1104–1117, 2009.

[19] M. Machanavajjhala, A., Kıfer, D., Gehrke, J., And Venkıtasubramanıam, “l-diversity: Privacy be- yond k-anonymity.,” ACM Trans. Knowl. Discov. Data 1, 2007.

[20] L. Ninghui, L. Tiancheng, and S. Venkatasubramanian, “t-Closeness: Privacy beyond k-anonymity and ℓ-diversity,” in Proceedings - International Conference on Data Engineering, 106–115, 2007.

[21] J. Li, Y. Tao, and X. Xiao, “Preservation of proximity privacy in publishing numerical sensitive data,” in International conference on management of data, no. 220, 473, 2008,

[22] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional K-anonymity,” in Proceedings - International Conference on Data Engineering, vol., 25, 2006.

[23] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “-Diversity: Privacy Beyond,” Discovery, vol. 1, no. 1, 146, 2007.

[24] R. Kumar, J. Novak, B. Pang, and A. Tomkins, “On anonymizing query logs via token-based hashing,” Proc. 16th Int. Conf. World Wide Web - WWW ’07, 629, 2007.

[25] C. C. Aggarwal, “On K-anonymity and the Curse of Dimensionality,” 31st Int. Conf. Very Large Data Bases, 901–909, 2005.

[26] G. Ghinita, Y. Tao, and P. Kalnis, “On the anonymization of sparse high-dimensional data,” in Proceedings - International Conference on Data Engineering, 715–724, 2008,

[27] M. Terrovitis and N. Mamoulis, “Privacy preservation in the publication of trajectories,” in Proceedings - IEEE International Conference on Mobile Data Management, 65–72, 2008.

[28] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu, “Anonymizing transaction databases for publication,” in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08, 767, 2008.

[29] Y. Xu, B. C. M. Fung, K. Wang, A. W. C. Fu, and J. Pei, “Publishing sensitive transactions for itemset utility,” in Proceedings - IEEE International Conference on Data Mining, ICDM, 1109–1114, 2008,

[30] C. C. Aggarwal and P. S. Yu, “On Privacy-Preservation of Text and Sparse Binary Data with Sketches,” in In Proceedings of the SIAM International Conference on Data Mining (SDM)., 57–67, 2013

[31] N. Alon, Y. Matias, and M. Szegedy, “The space complexity of approximating the frequency moments,” in Journal of Computer and System Sciences, vol. 58, no. 1, 137–147, 1999.

[32] E. Beinat, “Privacy and Location-based Services: Stating the Policies Clearly,” in GeoInformatics, vol. 4, 14–17, 2001.

[33] O. Abul, F. Bonchi, and M. Nanni, “Never walk alone: Uncertainty for anonymity in moving objects databases,” in Proceedings - International Conference on Data Engineering, 376–385, 2008.

[34] B. Malin and E. Airoldi, “The effects of location access behavior on re-identification risk in a distributed environment,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4258 LNCS, 413–429, 2006.

[35] S. Papadimitriou, F. Li, G. Kollios, and P. Yu, “Time series compressibility and privacy”, 459–470, 2007.

[36] B. Thuraisingham and E. Ferrari, “Web and information security,” IRM Press, 933–933, 2005.

[37] D. Kokkinakis and A. Thurin, “Anonymisation of Swedish Clinical Data,” in Artificial Intelligence in Medicine, 237–241, 2007.

[38] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania, “Efficient techniques for document sanitization,” Proc. 17th ACM Conf. Inf. Knowl. Manag., 843, 2008.

[39] R. Liu and H. Wang, “Privacy-preserving data publishing,” Proc. - Int. Conf. Data Eng., vol. 42, no. 4, 305–308, 2010.

[40] L. Sweeney, “Achıevıng K-Anonymıty Prıvacy Protectıon Usıng Generalızatıon And Suppressıon,” Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 05, 571–588, 2002.

[41] P. Samarati, “Protecting respondents’ identities in microdata release,” IEEE Trans. Knowl. Data Eng., vol. 13, no. 6, 1010–1027, 2001.

[42] L. Sweeney, “Datafly: a system for providing anonymity in medical data,” Database Secur. XI, 356–381, 2015.

[43] R. Chi-Wing, J. Li, A. W.-C. Fu, and K. Wang, “(A, K)-Anonymity,” Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD ’06, 754, 2006.

[44] K. Wang and B. C. M. Fung, “Anonymizing sequential releases,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’06, 414, 2006.

[45] X. Xiao and Y. Tao, “Personalized privacy preservation,” in Proceedings of the 2006 ACM SIGMOD international conference on Management of data - SIGMOD ’06, 229, 2006,

[46] R. J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” in Proceedings - International Conference on Data Engineering, 217–228, 2005,

[47] S. A. VINTERBO, “Privacy: A machine learning view.,” IEEE Trans. Knowl. Data Engin., 939–948, 2004.

[48] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, “Utility-based anonymization using local recoding,” in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’06, 785, 2006.

[49] M. E. Nergiz and C. Clifton, “Thoughts on k-anonymization,” Data Knowl. Eng., vol. 63, no. 3, 622–645, 2007.

[50] B. C. M. Fung, K. Wang, and P. S. Yu, “Top-down specialization for information and privacy preservation,” in Proceedings - International Conference on Data Engineering, 205–216, 2005.

[51] V. S. Iyengar, “Transforming data to satisfy privacy constraints,” in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’02, 279, 2004.

[52] K. El Emam and F. K. Dankar, “Protecting Privacy Using k-Anonymity,” J. Am. Med. Informatics Assoc., vol. 15, no. 5, 627–637, 2008.

[53] K. Benitez and B. Malin, “Evaluating re-identification risks with respect to the HIPAA privacy rule,” J. Am. Med. Informatics Assoc., vol. 17, no. 2, 169–177, 2010.

[54] U. S. C. Bureau, “Voting and Registration in the Election of November 2010 - Detailed Tables,” Voting and Registration, 2010.

[55] F. K. Dankar and K. El Emam, “A method for evaluating marketer re-identification risk,” in Proceedings of the 1st International Workshop on Data Semantics - DataSem ’10, 1, 2010,

[56] K. El Emam, S. Jabbouri, S. Sams, Y. Drouet, and M. Power, “Evaluating common de-identification heuristics for personal health information,” J. Med. Internet Res., vol. 8, no. 4, 2006.

[57] B. S, “Bell S. Alleged LTTE front had voter lists. National Post, July 22, 2006,” National Post, 2006.

[58] http://www.webcitation.org/5Xe4UWJKP. (01.01.2008).

[59] K. El Emam, “Risk-based de-identification of health data,” IEEE Secur. Priv., vol. 8, no. 3, 64–67, 2010.

[60] P. Samarati and L. Sweeney, “Generalizing data to provide anonymity when disclosing information (abstract),” in Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems - PODS ’98, 188, 1998.

[61] C. Pelin, “Sağlık Hizmetlerinde Anonimlik: Dağıtık Yapılar İçin İdeal Bir Veri Paylaşım Modeli,” 2014.

[62] M. Barbaro, “A Face Is Exposed for AOL Searcher No. 4417749,” New York Times, no. 4417749, 1–3, 2009.

[63] J. Domingo-Ferrer and J. M. Mateo-Sanz, “Practical data-oriented microaggregation for statistical disclosure control,” IEEE Trans. Knowl. Data Eng., vol. 14, no. 1, 189–201, 2002.

[64] H. Gökçe and O. Abul, “Sensitive knowledge hiding application,” in 2010 National Conference on Electrical, Electronics and Computer Engineering, ELECO 2010, 558–562, 2010.

[65] https://tr.euronews.com/2019/01/25/ab-ye-genel-veri-koruma-tuzugu-ihlal-edildigi-gerekcesiyle-8-ayda-95-bin-sikayet-ulasti, (01.02.2019)

[66] http://www.privacyanalytics.ca/,(01.01.2019)

# ÖZGEÇMİŞ

## KİŞİSEL BİLGİLER

**Adı Soyadı** : Barış ANKAY

**Uyruğu** : T.C.

**Doğum Yeri ve Tarihi** : Ankara – 17.07.1974

**Medeni Hali** : Evli

**Adres** : Eryaman 171. Sokak İlke Sitesi H5-2E/A blok No:5 Etimesgut/ANKARA

**E-Posta Adresi** : bankay@gmail.com

**İletişim (Telefon)** : 0 532 378 33 76

## EĞİTİM

**Lise** : Ömer Seyfettin Lisesi (Ankara) – 1993

**Ön Lisans** : Mersin Üniversitesi/Harita Kadastro (Mersin) – 1997

**Lisans** : Anadolu Üniversitesi/İktisat (Eskişehir) – 2004

**Lisans** : Ahmet Yesevi Üniversitesi/Bilgisayar Mühendisliği (Ankara) – 2008

**Yüksek Lisans** : Türk Hava Kurumu Üniversitesi/Elektrik ve Bilgisayar Mühendisliği (Ankara) – Devam Ediyor

## İŞ DENEYİMİ

**TKGM, Rize Kadastro**

Kadastro Teknikeri 1997-1999

**TKGM, Ankara Genel Müdürlük**

Tapu otomasyon 1999 – 2009

Veritabanı Yöneticisi 2009 – 2015

Şube Müdürü (vekâleten) 2015 – 2017