SAĞLIK HİZMETLERİNDE ANONİMLİK:
DAĞITIK YAPILAR İÇİN İDEAL BİR VERİ PAYLAŞIM MODELİ

ANONYMITY IN HEALTHCARE SYSTEMS:
AN IDEAL DATA SHARING MODEL FOR DISTRIBUTED STRUCTURES

PELİN CANBAY

PROF. DR. HAYRİ SEVER

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

Bilgisayar Mühendisliği Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2014

SAĞLIK HİZMETLERİNDE ANONİMLİK:
DAĞITIK YAPILAR İÇİN İDEAL BİR VERİ PAYLAŞIM MODELİ

ANONYMITY IN HEALTHCARE SYSTEMS:
AN IDEAL DATA SHARING MODEL FOR DISTRIBUTED STRUCTURES

PELİN CANBAY

PROF. DR. HAYRİ SEVER

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

Bilgisayar Mühendisliği Anabilim Dalı İçin Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2014

PELİN CANBAY'ın hazırladığı "**Sağlık Hizmetlerinde Anonimlik: Dağıtık Yapılar İçin İdeal Bir Veri Paylaşım Modeli"** adlı bu çalışma aşağıdaki jüri tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**'nda **YÜKSEK LİSANS** tezi olarak kabul edilmiştir.

Yrd. Doç. Dr. Murat AYDOS
Başkan

Prof. Dr. Hayri SEVER
Danışman

Yrd. Doç. Dr. Erkut ERDEM
Üye

Yrd. Doç. Dr. Nazlı İKİZLER CİNBİŞ
Üye

Yrd. Doç. Dr. Burkay GENÇ
Üye

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS** tezi olarak onaylanmıştır.

Prof. Dr. Fatma SEVİN DÜZ
Fen Bilimleri Enstitüsü Müdürü

Sevgili Eşim ve Aileme

ETİK

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

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

beyan ederim.

29/12/2014

Pelin CANBAY

ÖZET

SAĞLIK HİZMETLERİNDE ANONİMLİK:

DAĞITIK YAPILAR İÇİN İDEAL BİR VERİ PAYLAŞIM

MODELİ

Pelin CANBAY

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Hayri SEVER

Aralık 2014, 86 Sayfa

Sağlık kuruluşları tarafından elde edilen (kayıt altına alınan) veriler, birçok alanda ileriye yönelik çözümler üretmek için olağanüstü fırsatlar sunmaktadır. Sağlık hizmetleri alanında faydalı sonuçlar üretebilmek için doğru (gerçek - tutarlı) verilerin paylaşılması gerekmektedir. Sağlık sistemlerinde tutulan kişisel sağlık kayıtlarının bireyler ile ilgili hassas bilgiler içermesinden dolayı, bu kayıtların sadece adı, soyadı ve kimlik numarası gibi bilgilerinin çıkarılarak başka hiçbir düzenleme yapılmadan doğrudan paylaşılması, bireysel mahremiyetin ihlal edilmesine sebep olmaktadır. Literatüre bakıldığında, mahremiyet ihlaline sebep olmadan, eldeki verilerden alınabilecek faydayı olabildiğince maksimum seviyede

tutmayı hedefleyen birçok mahremiyet korumalı veri paylaşım yaklaşımı geliştirilmiştir. Özellikle son yıllarda Mahremiyet Korumalı Veri Madenciliği (Privacy-Preserving Data Mining (PPDM)) ve Mahremiyet Korumalı Veri Yayıncılığı (Privacy-Preserving Data Publishing (PPDP)) yaklaşımları, kişisel veya kurumsal mahremiyeti korumak adına kapsamlı olarak çalışılmıştır.

Bu tez çalışmasında; Mahremiyet Korumalı Veri Madenciliği ve Mahremiyet Korumalı Veri Yayıncılığı yaklaşımları özetlenmiş, sağlık kayıtları çerçevesinde değerlendirilmiş ve sağlık hizmetlerinde, hem mahremiyeti koruyan hem de veri paylaşımına olanak sağlayan bir veri dağıtım modeli önerilmiştir.

Burada yapılan çalışmada, dağıtık sağlık kuruluşlarından toplanan verilere, alıcı kurumların ihtiyaçları doğrultusunda bölümleme yapan ve gerekli anonimleştirme ölçütlerini uygulayan, daha sonra bu anonim bilgileri alıcı kurumların hizmetine sunan ideal bir sistem modeli önerilmektedir. Önerilen model; dağıtık veri kümelerinin paylaşımına olanak sağlayan merkezi bir veri dağıtım sisteminin modelidir. Bu modelin gerçekleştiriminde, yatay ve dikey bölümleme teknikleri kullanılarak veriler ayrıştırılmış, daha sonra ayrıştırılan bu veriler k-anonimlik ve ℓ-çeşitlilik ölçütlerine tabi tutularak değerlendirilmiştir. Gerçekleştirme işlemleri olası iki farklı modele daha uygulanmış ve sonuçlar karşılaştırılmıştır. Yapılan işlemler sonucunda, önerilen modelin olası modellere göre hem veri kaybı hem de veri gizliliği açısından en ideal sonucu verdiği gözlemlenmiştir. Önerilen modelin amacı; mahremiyet (gizlilik) koruma ve veri faydası arasındaki dengeyi en ideal seviyede tutmayı sağlamaktır.

Anahtar Kelimeler: Mahremiyet koruma, gizlilik, veri tabanı bölümleme, k-anonimlik, veri dağıtımı

ABSTRACT

ANONYMITY IN HEALTHCARE SYSTEMS:

AN IDEAL DATA SHARING MODEL FOR DISTRIBUTED STRUCTURES

Pelin CANBAY

Master of Science, Department of Computer Engineering

Supervisor: Prof. Dr. Hayri SEVER

December 2014, 86 Pages

Data obtained or recorded by healthcare institutions, present extraordinary opportunities to produce forward solutions in many fields. Sharing accurate (real-consistent) data is necessary to produce useful results in healthcare field. Because of personal health records that are held by health systems include sensitive attributes about individuals, sharing these records after removing information like name, surname and identity number without any editing causes privacy disclosure. In literature, a lot of privacy-preserving data sharing approaches are developed that aims keeping the benefit that can be taken from existing data in maximum level. Especially in recent years, Privacy-Preserving

Data Mining and Privacy-Preserving Data Publishing approaches were studied comprehensively to protect personal or institutional privacy.

In this thesis, Privacy-Preserving Data Mining and Privacy-Preserving Data Publishing approaches were summarized, evaluated within the framework of health records and a data distribution model that faciliates both protecting privacy and data sharing was proposed.

In this work, an ideal system model was proposed that makes partitioning according to needs of recipient institutions and applies necessary anonymization criteria to the collected data from distributed health institutions, then presents this anonymous information to recipient institutions. The proposed model is a central data distribution system model that faciliates sharing of distributed data sets. In implementation of this model, horizontal and vertical partitioning techniques were used to decompose the data, then the decomposed data were evaluated by appliying k-anonymity and ℓ-diversity. The implementation processes were applied to two different models and results were compared. At the end, it was observed that the proposed model gave the ideal result in terms of both data loss and data privacy in comparision with likely models. The aim of the proposed model is keeping the balance between protecting privacy and data benefit in an ideal level.

Key Words: Privacy preserving, privacy, database partitioning, k-anonymity, data distribution

TEŞEKKÜR

Tez çalışmamın her aşamasında değerli katkı ve görüşleriyle yol gösteren, beni her zaman çalışmaya teşvik eden ve güven veren danışmanım Sayın Prof. Dr. Hayri SEVER'e, önemli yorum ve değerlendirmeleri ile katkıda bulunan jüri üyelerim Sayın Yrd. Doç. Dr. Murat AYDOS'a, Sayın Yrd. Doç. Dr. Nazlı İKİZLER CİNBİŞ'e, Sayın Yrd. Doç. Dr. Erkut ERDEM'e, Sayın Yrd. Doç. Dr. Burkay GENÇ'e, çalışmamın her aşamasında manevi olarak yanımda olan eşim Yavuz CANBAY'a ve aileme en içten teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

TABLOLAR

Sayfa

ŞEKİLLER

Sayfa

1. GİRİŞ

1.1. Tanım

Günümüzde bilgi teknolojileri, her alanda veri üretilmesine, toplanmasına ve depolanmasına olanak sağlamaktadır. Depolanan bu veriler, ileriye yönelik faydalı yenilikler geliştirebilmek adına büyük bir önem teşkil etmektedir. Bu amaç doğrultusunda, özellikle veri madenciliği uygulamaları ile eldeki verilerden anlamlı sonuçlar çıkararak birçok alanda faydalı bilgiler elde edilmiştir. Dolayısıyla depolanan verilerin, araştırma toplulukları ile paylaşılabilir olması büyük önem taşımaktadır. İşbirliği ve veri paylaşımı, biyomedikal araştırmaların temeli haline gelmiştir. Amacı; "Tüm araştırma topluluklarına verileri ulaşılabilir hale getirmek" olan Uluslararası Kanser Genom Birliği (ICGC) [1], bu alanda yapılan uluslararası projelere bir örnektir. Özellikle sağlık sektöründe, dağıtık haldeki veri ambarlarında tutulan verilerin paylaşılmasına olan ihtiyaç gün geçtikçe artmaktadır. Fakat bireyler veya kurumlar ile ilgili bilgilerin bulunduğu bu veri ambarları, ifşa edilmesi istenmeyen birçok hassas bilgiyi de içermektedir. Özellikle sağlık alanında depolanan verilerin analiz edilebilmesi sonucu atılabilecek birçok adım, bireylerin mahremiyetini korumak adına kısıtlanmıştır.

Sağlık hizmetlerinde faydalı sonuçlar elde edebilmek için gerçek (doğru-tutarlı) verilerin paylaşılabilir olması gerekmektedir. Kişisel sağlık verileri sistemlerde tutulduğu haliyle bireylerle ilgili hassas bilgiler içerdiğinden dolayı, sağlık kayıtlarının olduğu gibi paylaşılması bireylerin mahremiyetinin ihlal edilmesine sebep olacaktır. Sağlık alanında bireysel mahremiyeti korumak adına; 1996 yılında, HIPAA (Health Insurance Portability and Accountability Act) [2] adı verilen mahremiyet yasaları, Amerika Birleşik Devletleri tarafından ulusal mevzuat olarak yürürlüğe konulmuştur. Türkiye'de ise 2005 yılında Sağlık Bakanlığı tarafından, "Veri Güvenliği Hakkında Genelge 2005/153" [3] başlığı altında, sağlık alanında, bireysel mahremiyetin korunması için dikkat edilmesi gereken unsurlar genelge olarak yayımlanmıştır.

Hükümetler tarafından bireysel mahremiyetin koruma altına alınmasıyla kurumlar, ellerindeki verileri yayımlamadan önce üzerinde bazı düzenlemeler yapmak zorunda kalmıştır. Bireylerin mahremiyetini korumak için, kişisel bilgiler içeren bir veri kümesinin paylaşılacağı veya yayımlanacağı zaman, açık-tanımlayıcılar olarak

adlandırılan ve bireyi doğrudan tanımlayan genellikle isim, adres ve telefon numarası gibi verilerin, veri kümesinden çıkartılarak veya bu verilerin şifrelenerek paylaşılması yolu tercih edilmiştir. Ancak, bu durum başlarda mahremiyeti sağlayan bir çözüm olarak görünse de, birçok veri kümesinde yetersiz kalmıştır. Kişisel bilgilerin tutulduğu veri kümelerinde, bireysel mahremiyeti korumak için veri sahiplerinin anonimleştirilmesi; yani, verilerin sahiplerinin birey olarak ayırt edilemez olması gerekmektedir. Bireysel anonimliği sağlamak için, yarı-tanımlayıcılar olarak adlandırılan ve bir bireyi tanımlamak için tek başına yeterli olmamasına rağmen başka verilerle bir araya getirilerek bir bireyi tanımlayabilme imkânı sağlayabilecek; yaş, cinsiyet ve ırk gibi ayırt edici öz niteliklerin de düzenlenmesi gerekliliği ortaya çıkmıştır.

Mahremiyet koruma yaklaşımları temelde; verilerin gizliliğini korumayı garantilerken, aynı zamanda verilerden alınacak faydayı da maksimum düzeyde tutmayı amaçlamaktadır. Çünkü veri faydası ve veri gizliliği arasında zıt bir denge söz konusudur. Örneğin; bir veri kümesini hiçbir gizleme işlemine tabi tutmadan olduğu gibi paylaşmak o veri kümesinden alınacak faydayı maksimum yapacaktır, fakat veri kümesi saldırılara ve her türlü mahremiyet ifşasına açık hale gelecektir. Aynı durumun tersi de söz konusudur. Yani bir veri kümesini tamamen gizler ve hiçbir bilgi paylaşılmaz ise o veri kümesinden herhangi bir tespit veya ifşa etme işlemi yapılamayacak, dolayısı ile güvenlik ve mahremiyet maksimum düzeyde korunurken veri faydası minimum düzeye inecektir. Kısacası; mahremiyet korumalı algoritmalarda veri faydası arttıkça mahremiyet koruması azalmakta, mahremiyet koruması arttıkça veri faydası da azalmaktadır. Literatürde yapılan tüm çalışmalarda, veri faydası ve veri gizliliği arasındaki dengeyi ideal seviyede tutmak üzerine odaklanılmıştır. Mahremiyet korumalı çalışmalarda veri gizliliğinden kasıt; veri sahiplerinin mahremiyetini korumak için, bireyleri anonim bir birey, kurumları ise anonim bir kurum haline getirmektir.

Günümüze kadar, bireysel veya kurumsal anonimliği sağlamak adına birçok çalışma yürütülmüştür. Özellikle son yıllarda, bireysel mahremiyetin gizliliğini ve güvenliğini sağlayabilmek için Mahremiyet Korumalı Veri Madenciliği (Privacy-Preserving Data Mining (PPDM)) ve Mahremiyet Korumalı Veri Yayını (Privacy-Preserving Data Publishing (PPDP)) yaklaşımları kapsamlı olarak çalışılmıştır. Ek olarak problemi çözmek adına, istatistiksel çıkarım toplulukları,

kriptografik iletişim toplulukları ve veri tabanı kullanım toplulukları gibi pek çok araştırma topluluğu tarafından da bu konuda araştırmalar ve çalışmalar yapılmıştır. Mahremiyet Korumalı Veri Madenciliği (PPDM) alanında yapılan çalışmalar, kişisel veya kurumsal gizliliğin korunması adına pek çok algoritma ve modele sahiptir [4]. Bu yaklaşımın amacı, geleneksel veri madenciliği teknikleriyle büyük miktardaki verilerden ilgili çıkarımın yapılması ile beraber aynı zamanda hassas verinin korunmasıdır. Mahremiyet Korumalı Veri Yayını (PPDP) ise veri madenciliği sonuçlarından ziyade, özel veriler içeren tabloların veya veri tabanlarının yayımlanabilmesi üzerine odaklanmıştır [5]. PPDP, kayıt sahiplerinin kimliklerini gizleyerek veriyi anonimleştirir ve güvenle yayımlanabilmesini sağlar. İstatistiksel çıkarımlar üzerine çalışan topluluklar, istatistiksel tablolar için mahremiyet korumalı yayınlama üzerine çalışırken, bu alandaki pek çok güncel çalışmada İstatistiksel Açığa Çıkarma Kontrolü (Statistical Disclosure Control SDC) metotları üzerine odaklanmıştırlar [6]. Veri tabanı toplulukları genellikle, veri tabanı bölümleme ve sorgu denetleme gibi farklı metotları kullanarak hassas verinin güvenliğinin sağlanması için diğer topluluklarla ortak çalışmalar yürütmüşlerdir [7]. Kriptoloji topluluğu hassas verinin ifşa edilmeden farklı kurumların verilerini güvenli bir şekilde paylaşmaları için genel fonksiyonlar üzerine yoğunlaşmıştır. Bu alanda son zamanlarda yapılan çalışmalarda Güvenli Çok-parçalı Hesaplama (Secure Multiparty Computation (SMC)) üzerinde durulmuştur [8]. Bazı durumlarda çalışmaların paralel hatları oldukça benzer olmasına rağmen topluluklar geniş çaplı bir perspektif edinmek için etkili bir şekilde birleşememişlerdir.

Bu çalışmada, mahremiyet koruma alanında pek çok yaklaşım özetlenmiş ve sağlık kayıtları çerçevesinden değerlendirilmiştir. Veri kümesinin gizliliğinin korunması amacıyla literatürde birçok teknik geliştirilmiştir. Bu tekniklerden randomizasyon gibi pek çok teknik, veri madenciliği sonuçlarını görüntülemek veya güvenli veri yayımı yapabilmek adına, kayıt seviyesindeki verilere, düzenli gürültü ekleme yaparak, verileri bozmaktadır [4, 9]. Bu durumda sağlık kayıtları için, kullanılan veri kullanışsız olmaktadır. Örneğin; kanser hastalarının verilerinden oluşan bir veri kümesinin içerisinde 15 yaşında erkek hasta bulunmazken; eklenen gürültüler ile belirli sayıda, 15 yaşında erkek kanser hastası içeren bir veri kümesi elde edilebilir. Bu tip mahremiyet korumalı teknikler birçok market-tabanlı

yaklaşımlarda verimli bir şekilde kullanılabilirken, sağlık alanında böyle bir yaklaşım tercih edilmemektedir. Yapılan bu tez çalışmasında; değerlerin güvenilirliğini bozmadan gizliliği koruyan veri teknikleri ile ilgili çalışmalar değerlendirilmiştir. Bu tip teknikler veriye gürültü eklemek yerine verileri genelleştirmekte ve bu sayede verilerin spesifikasyonunu azaltarak veri kümesinin anonimleştirilmesini sağlamaktadır [10]. Bu tez çalışmasında modellerin gerçekleştiriminde kullanılan k-anonimlik [10-13] ve ℓ-çeşitlilik [14] ölçütlerinin sağlanması için genelleştirilme ve baskılama tekniği kullanılmıştır. Literatürdeki birçok çalışma temel olarak, veri sahipleri teşhis edilebilir (identifiable) olan veri kümelerinin kimliksizleştirilmesi (de-identification) veya anonimleştirilmesi için teknikler geliştirilmesi üzerine yapılmıştır. Bu çalışmada, sağlık hizmetlerinde, dağıtık haldeki verilerin merkezi ve güvenilir bir yapı üzerinden, gerekli mahremiyet ölçütlerini kullanarak, farklı kurum veya kuruluşlarla paylaşılabilmesini sağlayan bir model tanımı üzerine odaklanılmıştır.

1.2. Katkılar

Bu çalışma öncelikle, mahremiyet korumalı yaklaşımlar konusunda yapılan birçok çalışmanın tanıtılması, incelenmesi ve özetlenmesi bakımından bir araştırma çalışması niteliği taşımaktadır. Bunun yanında, mahremiyet korumalı yaklaşımların doğru bir sistem üzerinde kullanılması ile sağlık hizmetlerinde veri paylaşımı hususunda meydana gelen problemlerin çözülmesi hedeflenmiştir. Bu amaç doğrultusunda; sağlık hizmetlerinde dağıtık halde bulunan verilerden merkezi bir yapı aracılığıyla maksimum düzeyde fayda elde edebilecek, aynı zamanda veri sahiplerinin mahremiyetini koruyacak paylaşımlar yapılabilmesi gerekliliği ortaya çıkmaktadır. Yapılacak paylaşımlarda veri sahiplerinin mahremiyetini korurken aynı zamanda veri faydasını en üst düzeyde tutabilecek, basit ama güçlü bir sistem modeli önerilmektedir.

Önerilen sistem modeli güvenilir merkezi bir sistemi temel almakta ve istenen mahremiyet ölçütleri doğrultusunda veri faydasını arttırmaktadır. Ayrıca, özellikle çok boyutlu veri kümelerinde gerçekleştirilen anonimleştirme işlemlerinin zorluğu, önerilen modelde kullanılan veri bölümleme yöntemleri ile aza indirilmiştir.

Veri madenciliği sonuçları ve güvenli veri transferinden ziyade temel olarak, genel sistem modelinin nasıl olması gerektiği üzerinde durulmuştur. Sağlık verilerinin

dağıtım aşamasında, kişisel sağlık verilerinin sadece alıcı kurum tarafından ihtiyaç duyulan (istenen ve merkezi sistem yöneticileri tarafından onaylanan) kısmı, veri tabanı bölümleme metotları kullanılarak ayrıştırılmaktadır.

Dağıtık verilerin güvenilir bir merkez tarafından mahremiyet korumalı ve veri faydası yüksek bir biçimde paylaşılabilmesine olanak sağlamak için önerilen model, bu amaç doğrultusunda önerilen diğer modellerle ve iki olası modelle daha karşılaştırılmıştır. Karşılaştırmalar sonucunda ne tür yapılar için nasıl modeller kullanılması gerektiği açıklanmış ve modeller arası avantajlar ve dezavantajlar değerlendirilmiştir.

1.3. Organizasyon

Çalışmanın ilk bölümünde; sağlık hizmetlerinde veri paylaşımının gerekliliği ve bu paylaşımlardaki mahremiyetin korunmasına olan ihtiyaç tanımlanmıştır. Yapılan çalışmalar ve bu tez çalışmasının farkına kısaca değinilmiş ve çalışmanın katkıları kısaca belirtilmiştir. İkinci bölümde; yapılan çalışmanın anlaşılabilir olması için bu konuda literatüre yerleşen tanım, ölçüt ve ayrımların açıklaması yapılmıştır. Üçüncü bölümde; konu ile alakalı yapılan çalışmalar tanıtılmış ve bu tez çalışmasında hangi aşamalarda ne şekilde kullanıldığı belirtilmiştir. Dördüncü bölümde bu çalışmanın amacı, ne tür problemlere çözüm olarak sunulduğu ve literatürdeki benzer bir çalışmadan farkları ve ortak yanları belirtilmiştir. Beşinci bölümde çalışmada kullanılan materyal ve metotlar anlatılmıştır. Altıncı bölümde çözüme yönelik önerilen modelin gerçekleştirimi yapılmış ve elde edilen sonuçlar açıklanmıştır. Yedinci bölümde yapılan çalışmanın avantaj ve dezavantajları değerlendirilmiş ve önerilen modelin ne tür yapılarda kullanılabilirliği tartışılmıştır. Son bölümde ise çalışmada elde edilen sonuçlar ve bu sonuçlar doğrultusunda ileriye dönük yapılabilecek çalışmalar değerlendirilmiştir.

2. GENEL BİLGİLER

2.1. Genel Tanımlar

Bir amaca yönelik olarak toplanan veri kümeleri içerisinde, kişiler ve kurumlar ile ilgili birçok öznitelik bulunabilir. Bu özniteliklerden yola çıkarak veri sahiplerinin teşhis edilebilmesinin önüne geçmek için, veri sahiplerinin kimliksizleştirilmesi veya anonimleştirilmesi gerekmektedir. Eldeki veriler üzerinde düzenlemeler yaparak veri sahiplerini anonimleştirmek için öncelikle veri kümesindeki özniteliklerin sınıflandırılması gerekmektedir [11]. Anonimleştirme yaklaşımlarında temel olarak veri yayıncısının elinde bir tablo bulunduğu ve bu tabloda 4 tip verinin bulunduğu varsayılmaktadır. Bunlar;

- Açık tanımlayıcılar (Explicit Identifier): Adı, soyadı, telefon numarası ve sosyal güvenlik numarası gibi kayıt sahibini doğrudan tanımlayan öznitelikler kümesi.
- Yarı tanımlayıcılar (Quasi Identifier): Yaş, adres, cinsiyet gibi, kayıt sahibini tanımlamak için tek başına yeterli olmamasına rağmen, veri kümesinde bulunma durumuna bağlı olarak veya başka verilerle bir araya getirilerek bir bireyi tanımlayabilecek potansiyele sahip olan öznitelikler kümesidir. Yarı tanımlayıcı öznitelikler kısaca QID olarak belirtilirken bu özniteliklerin değer kümeleri qid olarak gösterilmektedir.
- Hassas öznitelikler (Sensitive attributes): Hastalık bilgisi, gelir bilgisi, engel durumu gibi, veri sahiplerine özel hassas bilgiler kümesi. S ile gösterilir ve bir kümede birden çok olabilmektedir.
- Hassas olmayan öznitelikler (Non-Sensitive attributes): Diğer üç gruba girmeyen özniteliklerdir. Bu özniteliklerin ifşa edilmesi herhangi bir tehlike teşkil etmemektedir. Genellikle bu tip bireylerle ilgili ayırt edici bilgiler içermeyen veriler tablolarda yer almamaktadır.

Tablo 2.1'de bir veri yayıncısının elinde bulundurduğu varsayılan ve yayımlamadan önce anonimleştirilmesi gereken bir tablo gösterilmektedir. Tabloda üç tip veri gösterilmektedir çünkü bu aşamada hasta ile ilgili hiçbir anlam ifade etmeyen bir öznitelik hala bulunamamıştır.

Tablo 2.1. Mahrem veriler içeren tablo örneği

Açık Tanımlayıcılar			Yarı Tanımlayıcılar						Hassas Veriler
Kimlik No	**Ad Soyad**	**Tel**	**Cinsiyet**	**Yaş**	**Doğum Yeri**	**Posta Kodu**	**Medeni Hali**	**Mesleği**	**Hastalık**
			Bay	20	Bursa	16001	Bekâr	Terzi	HIV
			Bay	24	Mardin	47001	Bekâr	Öğretmen	Hepatit
			Bay	28	Muş	49001	Evli	Kasap	HIV
			Bay	32	Ankara	06001	Evli	Mühendis	Ülser
			Bay	36	İstanbul	32001	Evli	Doktor	Astım
			Bayan	20	Bursa	16001	Bekâr	Kuaför	HIV
			Bayan	24	Mardin	47001	Bekâr	Doktor	Ülser
			Bayan	28	Muş	49001	Evli	Ev Hanımı	Hepatit
			Bayan	32	Ankara	06001	Evli	Öğretmen	HIV
			Bayan	36	İstanbul	32001	Evli	Doktor	Astım

Şirketler içi yaygın kanı, kişisel verilerin bulunduğu veri kümelerinden açık tanımlayıcı bilgilerin çıkartılmasıyla veri sahiplerinin kimliksizleşeceği yani tanımlanamayacağı (de-identify), bir başka ifadeyle anonimleşeceği yönündedir. Çünkü veri kümesi anonim olarak görünmektedir. Fakat kalan veriler, yayımlanan veri kümesindeki tek olma özellikleriyle veya başka bir veri kümesindeki verilerle eşleştirilmesine dayanarak veri sahiplerinin teşhis edilmesinde (re-identification) kullanılabilmektedir [15]. Birkaç özniteliğin eşleşmesi bir bireyi doğrudan veya neredeyse doğrudan tanımlayabilir. [13]'de yapılan çalışmada, sadece (5 haneli posta kodu, cinsiyet ve doğum tarihi) bilgileri ile Amerika Birleşik Devletleri'nde toplumun %87'sinin (248 milyon insandan 216 milyonunun) doğrudan tanımlanabileceği belirtilmiştir. Bu doğrultuda yapılan birçok çalışmada; bir veri

kümesinden tüm açık tanımlayıcılar çıkarılsa bile, kalan veriler kullanılarak veri sahiplerinin tanımlanabileceği gösterilmiştir. Bu durum literatürde Bağlantı Saldırısı (Linking Attack) olarak adlandırılır [5]. Bu tip saldırılardan korunmak için veri yayıncısının, herhangi bir veri dönüşüm tekniği kullanarak verileri anonim bir tabloda sunması gerekmektedir. Veri kümelerine yapılabilecek saldırılar ve bu saldırılardan korunmak için geliştirilen modeller Bölüm 2.3'de detaylandırılmaktadır.

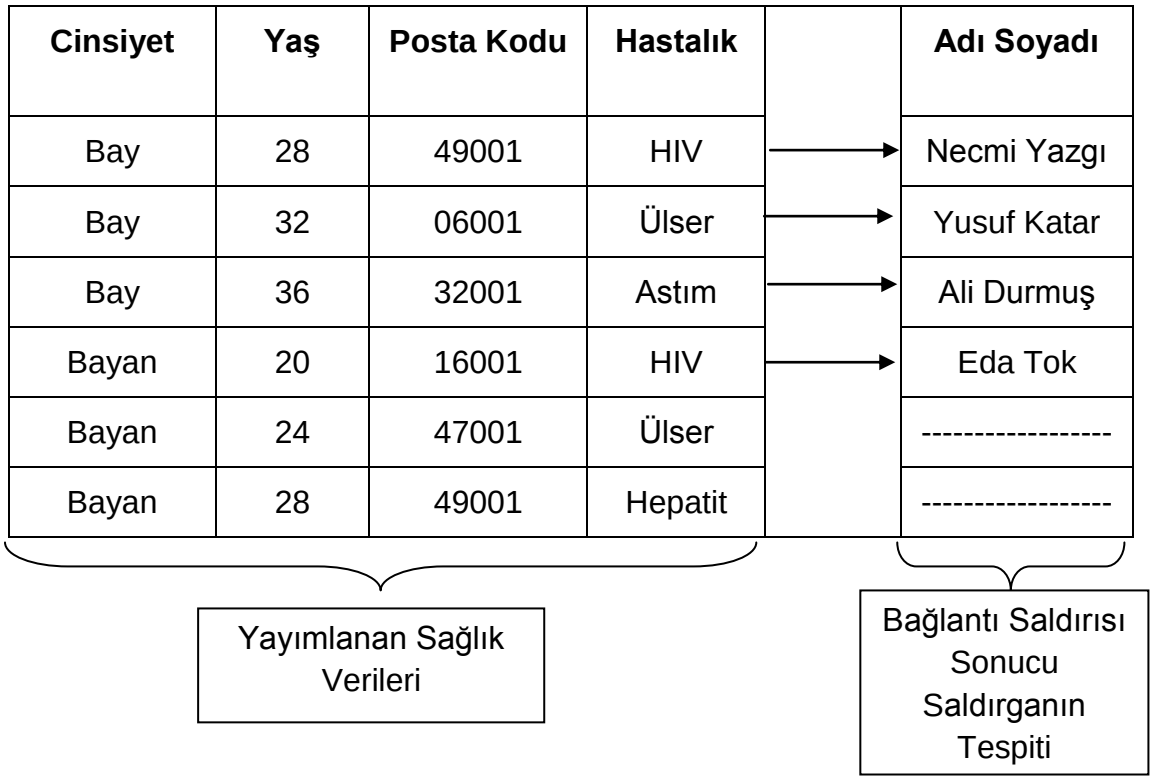
2.2. Mahremiyet Koruma - Anonimleştirme

İçerisinde sahiplerinin ifşa edilmesini istemedikleri hassas veriler bulunan veri kümelerinin güvenle paylaşılabilmesi için, veri sahiplerinin gizliliğinin (mahremiyetinin) korunması gerekmektedir. Bu tip veri kümelerinde, mahremiyet korumanın temelde nasıl olması gerektiği 1977 yılında, Dalenius [16] tarafından net bir tanım ile belirtilmiştir. Tanıma göre: "Yayımlanan bir veri kümesi, arka planda başka kaynaklardan bilgiler elde etmiş olsa bile bir saldırgana, o veri kümesine erişimi yokmuş gibi veri sahipleri ile ilgili herhangi bir ekstra bilgi edinmesine izin vermemelidir". Aşağıda açık tanımlayıcı bilgilerin çıkarılmasına rağmen yapılabilen basit bir bağlantı saldırısı örneği gösterilmiştir. Saldırganın, Tablo 2.2'deki kayıt sahiplerinin Şekil 2.1'de yayımlanan tabloda varlığını bilmesi ile dört kayıt sahibinin hastalık bilgilerine ulaşmış olur.

Tablo 2.2. Yayımlanan bir anket örneği

Ad Soyad	Cinsiyet	Yaş	Posta Kodu	Mesleği
Erhan Doğru	Bay	20	16001	Terzi
Hakan Söner	Bay	24	47001	Öğretmen
Necimi Yazgı	Bay	28	49001	Kasap
Yusuf Katar	Bay	32	06001	Mühendis
Ali Durmuş	Bay	36	32001	Doktor
Eda Tok	Bayan	20	16001	Kuaför
Yeşim Güler	Bayan	36	147001	Doktor
Sevgi Çelik	Bayan	48	25001	Ev Hanımı

Tablo 2.2 içerisinde mahrem bilgi içermediğinden dolayı yayımlanmasında sakınca görülmeyen basit bir anket sonucunda bile ulaşılabilen bilgileri içermektedir. Tablo 2.2'ye erişimi olan bir saldırgan, Şekil 2.1'de yayımlanan kimliksizleştirildiği düşünülen verileri eşleştirerek mahremiyet çıkarımı yapabilir. Bu durum Bağlantı saldırısına basit bir örnektir. Şekilde de görüleceği üzere sadece yaş, cinsiyet ve posta kodu bilgileri kullanılarak birçok kayıt tekilleştirilebiliyor ve kime ait olduğunun tespiti kolayca gerçekleştirilebiliyor.



Şekil 2.1. Bağlantı Saldırısı ile Mahremiyetin İfşa Edilmesi

Örneğin, yayımlanan bir veri kümesinde X şahsı ile ilgili özellikler ve hastalık verisi bulunsun. Bir saldırgan X şahsı ile ilgili her özelliği biliyor olsa bile yayımlanan veri kümesinden X şahsının hastalık verisine doğrudan ulaşamamalı. Bu mahremiyet güvenliğini sağlamak için veri kümesinin anonimleştirilmesi, yani kayıt sahiplerinin bir şekilde birbirinden ayırt edilemez olması gerekir. Bu doğrultuda farklı anonimleştirme ölçütleri geliştirilmiştir. Örneğin k-anonimlik ölçütü, Bağlantı saldırılarından korunmak için geliştirilmiş bir anonimlik ölçütüdür. Bu ölçüt, yayımlanacak bir veri kümesinde en az k tane kayıttan oluşan denklik sınıflarının kendi içinde yarı tanımlayıcı öznitelikler bakımından birbirinden ayırt edilemiyor olmasını gerektirir. Bu yapıyı sağlamak için veri kümesine bazı veri dönüşüm teknikleri uygulanır. Tablo 2.1'deki tablonun 3-anonimlik (k=3) ölçütüne uygun hali Tablo 2.3'te gösterilmektedir. Bölüm 2.4'te, anonimleştirme ölçütlerinde kullanılan veri dönüşüm teknikleri açıklanmaktadır.

Tablo 2.3. K-anonimlik (k=3) ölçütüne göre anonimleştirilmiş tablo örneği

Cinsiyet	Yaş	Doğum Yeri	Posta Kodu	Medeni Hali	Mesleği	Hastalık
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Hepatit
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Ülser
Bay	28-36	*	*	Evli	*	HIV
Bay	28-36	*	*	Evli	*	Ülser
Bay	28-36	*	*	Evli	*	Astım
Bayan	28-36	*	*	Evli	*	Hepatit
Bayan	28-36	*	*	Evli	*	HIV
Bayan	28-36	*	*	Evli	*	Astım

En az k tane kayıtın bulunduğu birbirinden ayırt edilemeyen bu gruplara Denklik Sınıfları (Equivalence Classes) denmektedir ve anonimleştirmede kullanılan önemli bir yapıdır.

2.3. Saldırı Yöntemleri ve Saldırılara Karşı Geliştirilen Güvenlik Ölçütleri

Saldırı prensiplerine göre geliştirilen güvenlik modelleri iki kategoriye ayrılmıştır [5]. İlk kategori, bağlama saldırıları (Linking Attack) olarak bilinir ve saldırganın erişim durumuna göre değişiklik gösterir. Bu saldırılarda saldırganın, kayıt sahibi ile ilgili bazı bilgilere sahip olduğu ve bu bilgileri kullanarak yayınlanan tablolarda kayıt ile bağlantı kurarak hassas bilgilerine ulaşabildiği farz edilir. Saldırganın, yayınlanan veri tablosundaki bir kayıta bağlantı elde etmesi kayıt bağlama (record linkage), hassas bir özniteliğe bağlantı elde etmesi öznitelik bağlama (attribute linkage), yayınlanan tablonun kendisine bir bağlantı elde etmesi tablo bağlama (table linkage) olarak adlandırılır.

İkinci kategori ise, bilgilendirici olmayan (uninformative principle) temele dayanır. Yayınlanan tablo saldırgana, önceki bilgilerine ek olarak bir miktar bilgi sunar. Eğer

saldırganın önceki ve sonraki bilgileri arasında farklılık meydana gelmişse bu durum olasılık saldırısına (probabilistic attack) açık hale gelecektir. Aşağıda bazı önemli saldırılar ve bu saldırılara çözüm olarak geliştirilen bazı güvenlik ölçütleri verilmiştir.

2.3.1. Bağlantı Saldırıları ve Önerilen Güvenlik Ölçütleri

2.3.1.1. Kayıt Bağlama Saldırısı

Bazı yarı tanımlayıcı özniteliklerin değerleri, yayımlanan tablodaki bir grup kaydı tanımlayabilir. Eğer saldırganın elindeki öznitelik değerleri ile bu tablodaki değerler eşleşirse kayıt sahibinin mahremiyeti tehlike altında olur. Saldırgan, bir grup kayıt içinde sadece birkaç kayıt arasından elindeki verilerle eşleşen bir kayıt sahibini gruptan ayırt edebilirse bu bir bağlantı saldırısı olur. Örneğin; bir hastanenin, bir grup kaydı bir araştırma merkezine yayımlamak istediğini farz edelim. Aynı zamanda araştırma merkezinin başka bir veri kümesine erişiminin de olduğunu ve her iki veri kümesinde de aynı veri sahiplerinin verilerinin olduğu varsayılmıştır. Bu iki tablodaki en yaygın özniteliklerin (yaş, cinsiyet ve meslek) eşleştirilmesi bir bireyin hastalığı tespit edilebilirliği Tablo 2.2 ve Şekil 2.1'de gösterilmiştir.

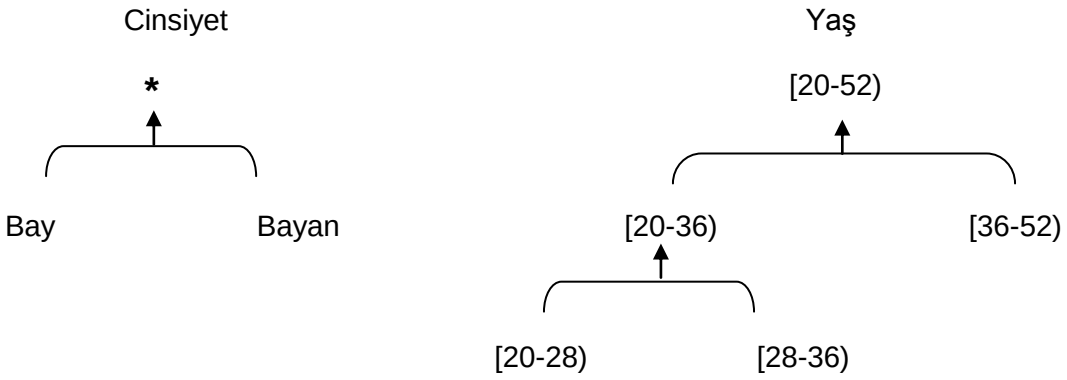
2.3.1.2. Kayıt Bağlama Saldırısından Korunma Ölçütleri

Kayıt bağlama saldırıları, yarı tanımlayıcı özniteliklerin tekilleştirilebilmesi ile yapılabilmektedir. Bu saldırılardan korunmak için [10-13] çalışmalarında k-anonimlik ölçütü önerilmiştir. Bu ölçüt, bir kayıtın tek başına tanımlanamaması için en az k-1 tane, aynı yarı-tanımlayıcı değerlere sahip olan kayıtların bulunduğu bir grupta olmasını gerektirir. Başka bir deyişle, en az k tane kayıtın bulunduğu bir grupta, tüm kayıtların yarı tanımlayıcı öznitelik değerleri birbiriyle aynı olmalıdır. Sonuç olarak saldırgan en az k tane kaydı birbirinden ayırt edemeyecektir. Bu tür gruplar literatürde denklik sınıfları olarak adlandırılmaktadır. Şekil 2.2'de, Tablo 2.1'de verilen tablonun, 3-anonimlik ölçütü olan Tablo 2.4 üzerinden denklik sınıflar vurgulanmıştır.

Tablo 2.4. Anonim (3-anonim) tablodaki denklik sınıfları

Cinsiyet	Yaş	Doğum Yeri	Posta Kodu	Medeni Hali	Mesleği	Hastalık
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Hepatit
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Ülser
Bay	28-36	*	*	Evli	*	HIV
Bay	28-36	*	*	Evli	*	Ülser
Bay	28-36	*	*	Evli	*	Astım
Bayan	28-36	*	*	Evli	*	Hepatit
Bayan	28-36	*	*	Evli	*	HIV
Bayan	28-36	*	*	Evli	*	Astım

Yukarıda Tablo 2.4'te yapılan anonimleştirme işlemi için her özniteliği göre bir sınıflandırma ağacı belirlenmiş olmalıdır. Örneğin yaş ve cinsiyet öznitelikleri için kullanılan sınıflandırma ağacı veya genelleştirme hiyerarşisi Şekil 2.2'de gösterilmiştir. Genelleştirme hiyerarşileri ile ilgili detaylı bilgi ve mahremiyet ölçütlerinde kullanılabilen bazı veri dönüşüm teknikleri Bölüm 2.4'te özetlenmiştir.



Şekil 2.2. Cinsiyet ve Yaş Özniteliklerinin Sınıflandırma Ağacı

Kayıt bağlama saldırısından korunmak için farklı veya benzer yapıda birçok ölçüt geliştirilmiştir fakat en yaygın kullanılanı k-anonimliktir.

2.3.1.3. Öznitelik Bağlama Saldırısı

Öznitelik bağlama saldırılarında saldırgan, hedef kayıt sahibini tam olarak tanımlayamamış olabilir fakat yayınlanan veri kümesinden kayıt sahibinin hassas verisini çıkarabilir. Kayıt bağlama saldırıları için geliştirilen ölçütler genellikle sadece yarı-tanımlayıcı öznitelikler üzerinden çıkarım yapılmasının önüne geçmiştir. Fakat kayıt bağlama saldırısına karşı güvenlik ölçütü kullanılan bir tabloda, denklik sınıflarındaki hassas özniteliklerin dağılımı kayıt sahibinin hassas verisinin bulunmasına sebep olabilir. Örneğin 4-anonim ölçütünde bir tabloda, denklik sınıflarından birindeki tüm kayıtlar aynı hassas öznitelik değerine sahip ise, o denklik sınıfındaki özelliklere uyan kayıt sahiplerinin hassas verisi bilinmiş olur. Tablo 2.5'te 3-anonim bir tablodaki öznitelik bağlama saldırısı örneklenmiştir. Tabloda verilen bilgilere göre, tabloda kaydı bulunan, 28-36 yaşları arasındaki tüm evli baylar hepatit hastalığına sahiptir.

Tablo 2.5. Anonim (3-anonim) bir tabloda öznitelik bağlama saldırısı örneği

Cinsiyet	Yaş	Doğum Yeri	Posta Kodu	Medeni Hali	Mesleği	Hastalık
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Hepatit
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Ülser
Bay	28-36	*	*	Evli	*	Hepatit
Bay	28-36	*	*	Evli	*	Hepatit
Bay	28-36	*	*	Evli	*	Hepatit
Bayan	28-36	*	*	Evli	*	Hepatit
Bayan	28-36	*	*	Evli	*	HIV
Bayan	28-36	*	*	Evli	*	Astım

2.3.1.4. Öznitelik Bağlama Saldırısından Korunma Ölçüleri

Kayıt bağlama saldırılarına karşı geliştirilen ve sadece yarı tanımlayıcı özniteliklerin düzenlenmesini gerektiren yöntemlerin saldırılardan korunmak için yetersiz kalması üzerine öznitelik bağlama saldırısını da önleyecek ölçütler önerilmiştir. Bu ölçütlerden en yaygın kullanılanı ℓ-çeşitlilik (ℓ-diversity) yaklaşımıdır [14]. K-anonimlik yaklaşımının tamamlayıcısı olarak da görülebilen bu yaklaşım, anonim bir tablodaki denklik sınıflarının her birinde en az ℓ tane iyi-temsil edilmiş (well-represented) hassas öznitelik bulunması gerekliliğini savunur. 3-anonim ve 3-çeşitlilik ölçütüne uygun bir tablo Tablo 2.6'da örneklenmiştir.

Tablo 2.6. Anonim (3-anonim) bir tabloda 3-çeşitlilik (ℓ=3) örneği

Cinsiyet	Yaş	Doğum Yeri	Posta kodu	Medeni Hali	Mesleği	Hastalık
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Hepatit
*	20-28	*	*	Bekâr	*	HIV
*	20-28	*	*	Bekâr	*	Ülser
Bay	28-36	*	*	Evli	*	HIV
Bay	28-36	*	*	Evli	*	Ülser
Bay	28-36	*	*	Evli	*	Astım
Bayan	28-36	*	*	Evli	*	Hepatit
Bayan	28-36	*	*	Evli	*	HIV
Bayan	28-36	*	*	Evli	*	Astım

Bir diğer önemli ölçüt ise t-yakınlıktır (t-closeness) [17]. t-yakınlık ölçütü, ℓ-çeşitlilik ölçütünün daha gelişmiş halidir ve hassas bilgilerin tüm anonim tablodaki yakınlığının denklik sınıflarda da sağlanması gerektiğini savunur. Örneğin a hassas verisi tüm tabloda %95 oranında bulunurken b hassas verisi %5 oranında bulunmakta olsun. 2-çeşitliliğin sağlandığı bir denklik sınıfında b verisinin %50 oranında bulunması, o gruptakilerin b özniteliğine sahip olma ihtimalini %5'ten

%50'ye çıkaracaktır. Karmaşıklık (skewness) saldırısı olarak bilinen bu duruma karşı geliştirilen t-yakınlık ölçütü, hassas bilgilerin tüm tablodaki dağılımının denklik sınıflarda da korunması gerektiğini savunur.

Öznitelik bağlama saldırısından korunmak için benzer yapıda birçok ölçüt geliştirilmiştir fakat en yaygın kullanılanlar ℓ-çeşitlilik ve t-yakınlıktır.

2.3.1.5. Tablo Bağlama Saldırısı

Bağlantı saldırılarından kayıt bağlama ve öznitelik bağlama saldırılarında, saldırganın kayıt sahibinin yarı tanımlayıcı öznitelik değerlerini bildiği farz edilir. Fakat bazı durumlarda bir kayıtın yayınlanan bir tabloda varlığı veya yokluğu o kayıtın hassas bilgisine ulaşılmasına sebep olabilir. Örneğin bir hastanenin belirli bir hastalık ile ilgili bir tablo paylaştığı durumu ele alındığında, bir kayıt sahibinin o tabloda olduğunun tespit edilmesi ile hassas verisi korunmamış olur. Bu durum tablo bağlama saldırısı olarak adlandırılır ve bir veri kümesinden birden çok farklı tablo yayımlanacağı zaman kullanılır.

2.3.1.6. Tablo Bağlama Saldırısından Korunma Ölçüleri

Tablo bağlama saldırıları, aynı tablodan yayımlanmış iki veya daha fazla farklı tablonun eşleştirilebilmesiyle yapılan saldırılardır. Tablo bağlama saldırılarından korunmak için önerilen yöntemlerden biri δ–varlık (δ-presence) ölçütüdür [18]. δ–varlık ölçütü, herhangi bir kayıtın varlığının tanımlanabilme olasılığının belirlenen bir aralıkta $\delta = (\delta_{min}, \delta_{max})$ sınırlanması gerektiğini savunan bir ölçüttür. Tablo bağlama saldırısından korunmak için, yapılacak yayınların bu aralık dikkate alınarak ayarlanması gerekmektedir.

2.3.1.7. Olasılık Saldırıları ve Önerilen Güvenlik Ölçütleri

Olasılık saldırılarında saldırganın kurabileceği bağlantılardan ziyade, yayımlanan tablo ile birlikte saldırganın olasılığa dayalı bilgileri üzerinde bir gelişme olup olmadığına odaklanılmıştır. Yayınlaman bir tablo ile birlikte, eğer saldırganın hedef kayıt ile ilgili ihtimale dayalı bilgilerinde bir değişiklik oluyor ise bu durum yayımlanan tablonun olasılık saldırısına açık olduğu anlamına gelir. Burada amaç, anonim tablonun herhangi bir bireysel kayıt ile ilgili bilgilendirici olmayan bir yapıda sunulmasını sağlamaktır.

Olasılık saldırısından korunmak için, ε-Diferansiyel Mahremiyet (ε-Differantial Privacy) [19] ve Dağıtık Mahremiyet (Distributional Privacy) [20] gibi ölçütler geliştirilmiştir. ε-Diferansiyel Mahremiyette, kayıt sahibinin verisinin yayınlanan tabloda olup olmaması durumundaki risklerin karşılaştırılması sonucu aradaki farkın ε'dan düşük olması gerektiğini savunur. Dağıtık Mahremiyette, ε-diferansiyel mahremiyetten daha sıkı bir mahremiyet kavramı vardır. Dağıtık mahremiyet ölçütüne göre, dağıtık birimlerden çekilen verilerden oluşan tabloda sadece altyapıdaki dağılım hakkında bilgi çıkarılmasına izin verilmeli, başka hiçbir ek çıkarım yapılamamalıdır. Saldırı modelleri ve sunulan bazı çözümler aşağıdaki tabloda özetlenmiştir.

Tablo 2.7. Saldırı modelleri ve sunulan gizlilik ölçütleri

Gizlilik Ölçütü	**Saldırı Modeli**			
	Kayıt Bağlama	Öznitelik Bağlama	Tablo Bağlama	Olasılık Saldırısı
k- Anonimlik	✓			
MultiR k-Anonimlik	✓			
ℓ-Çeşitlilik	✓	✓		
Güven Sınırlama		✓		
(α,k)-Anonimlik	✓	✓		
(X,Y)-Gizlilik	✓	✓		
(k,e)-Anonimlik		✓		
(ϵ,m)-Anonimlik		✓		
Kişiselleştirilmiş Gizlilik		✓		
t-Kapalılık		✓		✓
δ-Varlık			✓	
(c,t)-İzolasyon	✓			✓
ϵ-Diferansiyel Gizlilik			✓	✓
(d,y)-Gizlilik			✓	✓

2.4. Güvenlik Modellerinde Kullanılan Veri Dönüşüm Teknikleri

Farklı kurum ve kuruluşlar tarafından kayıt altına alınan verilerden özellikle kişisel kayıtlardan faydalı çıkarımlar yapabilmek için, öncelikle bu verilerin veri sahiplerinin mahremiyetini koruyacak şekilde kullanılabilir hale getirilmesi yani verilerin değiştirilmesi, düzenlenmesi (modifiye edilmesi) gerekmektedir. Veri kümelerindeki özniteliklere farklı veri dönüşüm teknikleri uygulanarak veri sahiplerinin anonimizasyonu sağlanabilir.

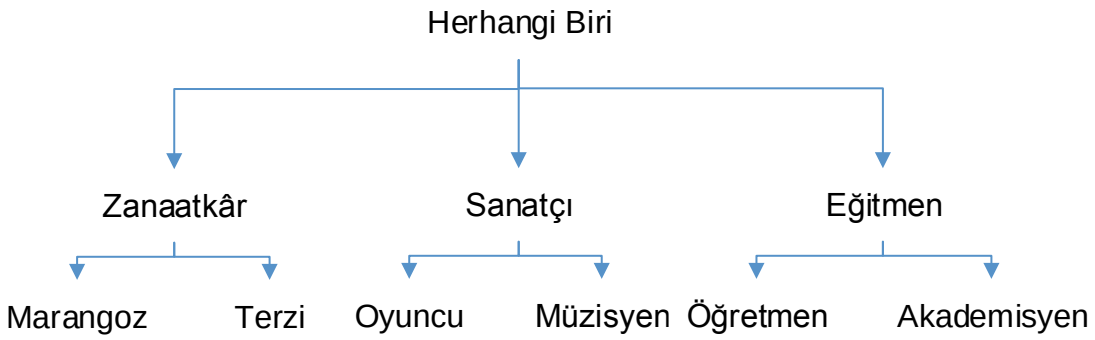
Bir veri kümesini anonimleştirmek ile o veri kümesindeki verileri kullanarak kayıt sahibinin tanımlanmasının veya kayıt sahibi ile ilgili başka bir bilgiye bağlantı kurulmasının önüne geçmiş olunur. Bu bağlamda ele alınması gereken iki önemli unsur bulunmaktadır. Bunlar veri faydası ve veri gizliliğidir. Çünkü veri faydası ve veri gizliliği arasında zıt bir denge söz konusudur. Veri ne kadar çok modifiye edilir ise o kadar çok mahremiyet ve güvenlik sağlayacak fakat aynı oranda veriden alınacak fayda da azalacaktır. Bu durumun tersi de söz konusudur. Yani veri ne kadar az modifiye edilir ise o kadar az mahremiyet garantilenecek fakat veriden alınacak fayda daha fazla olacaktır. Yapılan tüm anonimleştirme işlemlerinde amaç bu iki kavram arasındaki ideal dengeyi sağlamaktır.

Orijinal bir veri kümesinden (T) anonim bir veri kümesi (T') elde edebilmek için literatürde kullanılan veri dönüşüm yöntemleri ve sağlık alanında kullanılabilirliği aşağıda açıklanmıştır.

2.4.1. Genelleştirme ve Baskılama

Genelleştirme işleminde veri kümesinin anonimliğinin sağlanması için, veri kümesinde bulunan belirgin bir değer, verilen genelleştirme hiyerarşisi kullanılarak kendisinden daha genel bir değerle değiştirilir. Örneğin, Şekil 2.3.'de verilen Meslek özniteliğine bağlı genelleştirme hiyerarşisine göre ata düğüm olan Sanatçı, çocuk düğüm olan Müzisyen'den daha geneldir. Veri kümesinde, değeri müzisyen olan Meslek özniteliğinin kök düğüme genelleştirilmesine baskılama denir. Baskılama durumunda verinin değeri veri kümesinde (*) ile gösterilir, genelleştirmenin son seviyesidir. Şekil 2.3'te Meslek özniteliğinin 2 seviyeli genelleştirmesi örneklenmiştir.

Genelleştirme işlemleri farklı düzenlerde olabilmektedir. Literatürdeki yaygın genelleştirme düzenleri; Alt-ağaç Genelleştirme Düzeni [21], alt-link genelleştirme düzeni [22], hücresel genelleştirme düzeni [23] ve tüm yarı-tanımlayıcılara aynı genelleştirme düzeninin kullanıldığı yapı olan tüm-alan (full domain) genelleştirme düzenidir [22].



Şekil 2.3. Meslek Özniteliğinin Genelleştirme Hiyerarşisi

Genelleştirme ve baskılama tekniği, birçok market tabanlı veri kümelerinin anonimleştirilesi için uygun değildir. Çünkü uygulandığı veri kümesinin değerlerini bozmadan belirginliği azalttığından, özellikle istatistiksel çıkarım yapılacak paylaşımlarda, diğer veri dönüşüm tekniklerine göre daha fazla veri kaybına sebep olur. Fakat sağlık hizmetleri gibi gerçekliği bozulmamış, tutarlı verilerle çalışılması gereken ve mahremiyet koruma sınırının net olarak belirlenebildiği alanlarda bu veri dönüşüm tekniğinin kullanılması uygundur.

2.4.2. Anatomizasyon ve Permutasyon

Genelleştirme ve baskılama tekniğinin aksine, verilerin modifikasyonundan ziyade yarı-tanımlayıcılar ile hassas öznitelikler arasındaki ilişkiyi koparmaya yönelik yapılan çalışmadır. Yarı tanımlayıcılar ve hassas öznitelikler farklı tablolarda verilir, iki tabloda ortak olarak bir GrupID bulunur. Yarı-tanımlayıcı öznitelikleri aynı değerde olan kayıtlar bir grupta toplanır ve gruplar içindeki farklı hassas öznitelik sayısı arttıkça bağlantı saldırısı riski azalır. Permutasyon işlemi de [24] benzer yaklaşım ile nümerik hassas öznitelikler üzerinde çalışılmıştır. Tablo 2.8'de verilen bir veri kümesinin anatomizasyon uygulanmış örneği Tablo 2.9 ve Tablo 2.10'da gösterilmiştir.

Tablo 2.8. Anatomizasyon uygulanacak orijinal tablo

Yaş	Cinsiyet	Hastalık
30	Bay	Hepatit
30	Bay	Hepatit
30	Bay	HIV
32	Bay	Hepatit
32	Bay	HIV
32	Bay	HIV
36	Bayan	Grip
38	Bayan	Grip
38	Bayan	Kanser
38	Bayan	Kanser

Tablo 2.8'de orijinal hali verilen veri kümesine anatomizasyon uygulanması için öncelikle yarı tanımlayıcıların ve hassas verilerin birbirinden ayrılması daha sonra yarı tanımlayıcıların gruplanarak almış oldukları GrupID'ler ile hassas verilerin gruplanması gerekmektedir. Gruplar oluşturulurken Bölüm 2.4.1'de kullanılan hiyerarşiye benzer bir yapı kullanılır. Özniteliklerin belirli aralıklarla gruplanarak birbirine benzetilmesi sağlanır. Yarı tanımlayıcıların ayrılmış ve gruplanmış hali Tablo 2.9'da gösterilmektedir.

Tablo 2.9. Yarı tanımlayıcılar için Anatomizasyon örneği

Yaş	Cinsiyet	GrupID
30	Bay	1
30	Bay	1
30	Bay	1
32	Bay	1
32	Bay	1
32	Bay	1
36	Bayan	2
38	Bayan	2
38	Bayan	2
38	Bayan	2

Tablo 2.9'da belirtilen gruplara göre hassas öznitelik tablosu da anatomizasyon işleminde GrupID değerlerine göre birbirinden ayrılır ve orijinal tablodaki gibi hastalığın öznitelikleri değil hangi grupta kaç tane olduğu bilgisi verilir.

Tablo 2.10. Anatomizasyon uygulanmış hassas öznitelik tablosu

GrupID	Hastalık	Sayı
1	Hepatit	3
1	HIV	3
2	Grip	2
2	Kanser	2

Anatomizasyon tekniğinin en önemli avantajı yarı-tanımlayıcıların ve hassas bilgilerin modifiye edilmemeleridir. Bu teknikte veri faydası yüksektir fakat mahremiyet koruma düşüktür. Bu teknik, veri faydası bakımından market-tabanlı veri paylaşımları için kullanılabilecek tekniklerden biridir, fakat sağlık hizmetlerinde hassas öznitelikler ile yarı tanımlayıcılar arasındaki ilişkilerin paylaşımlarda büyük önem arz ettiği göz önüne alındığında, kullanımı pek uygun olmamaktadır.

2.4.3. Perturbasyon (Bozma- Karıştırma)

Perturbasyon tekniğinde, veri kümesine düzenli gürültüler eklenerek, bazı değerlerin yerleri değiştirilerek veya orijinal veri kümesinin bazı istatistiksel özelliklerinin temel alınmasıyla sentetik veri üretilerek veri kümesinin anonimleştirilmesi sağlanmıştır. Amaç mahremiyet korumalı istatistiksel veri toplama çalışmalarında veri faydasını yüksek tutmaktır.

Ortalama, standart sapma gibi istatistiksel sonuçların önemli olduğu paylaşımlarda, anonimleştirme işlemlerinde başlıca kullanılan teknik perturbasyondur. Verilere istatistiksel sonuçları değiştirmeyecek şekilde rastgele gürültüler eklenebilir, değerlerin yerleri değiştirilebilir veya orijinal verinin istatistiksel dağılımı göz önüne alınarak yapay veri üretilebilir.

Diğer anonimleştirme teknikleriyle karşılaştırıldığında bu teknik sonucunda yayımlanan veriler sentetik verilerdir. Başka bir ifadeyle verilerin gerçekliği bozulmuştur. Verilerin dağılımı, istatistiksel oranları gibi bilgilerin paylaşılması gereken alanlarda yüksek veri faydası ve yüksek koruma sağlamaktadır. Fakat sağlık hizmetlerinde, paylaşılan veriler üzerinde yapılan çalışmalar verilerin az fayda sağlasa bile gerçekliğinin bozulmamış olmasını gerektirir. Örneğin; kanser hastalarının verilerinden oluşan bir veri kümesinin içerisinde 15 yaşında erkek hasta bulunmazken; eklenen gürültüler ile belirli sayıda 15 yaşında erkek kanser hastası içeren bir veri kümesi elde edilebilir. Dolayısı ile bu tekniğin sağlık hizmetlerinde kullanılması uygun görülmemektedir. Bu teknikte, yayımlanan verilerden, rastgele gürültüleri filtreleyerek gerçek veri kümesini elde edebilen saldırılar da gerçekleştirilmiştir [25].

2.4.4. Sentetik Elektronik Tıbbi Kayıt Üretimi

Hastaların muayene kayıtlarını, klinik aktivitelerini, laboratuvar sipariş ve sonuçlarını, radyoloji sipariş ve sonuçlarını ve bireyler ile ilgili sosyo-demografik özellikleri içeren bilgilerin bulunduğu dijital yapılara Elektronik Sağlık Kayıtları (ESK) denir. Sağlık kurumları tarafından depolanabilen bu bilgiler, özellikle canlı-gözlemleme [26] (bio-surveilance) uygulamaları başta olmak üzere, toplum sağlığı izleme, farmakolojik çalışmalar ve sağlık tehditlerinin gözlemlenmesinde kullanılabilmektedir. Fakat hem yetersiz ve düzensiz kayıt tutulması hem de mahremiyetin korunması gerekliliğinden sağlık kuruluşları tarafından tutulan bu kayıtların erişimi sınırlandırılmıştır. İlgili çalışmaların test edilebilmesi, özellikle salgın hastalıkların davranışlarının tespit edilebilmesi ve belirli hastalıklar için hasta gelişiminin gözetlenmesi için bir takım çalışmalar sentetik tıbbi veri üretimi üzerine yoğunlaşmıştır. Bu amaç doğrultusunda belirlenen hastalık ile ilgili gerçek ESK'lar bir model olarak alınıp sentetik ESK'lar üreten bir dizi yapı tanımlanmıştır.

[27]'de hastalık takip sistemlerinin değerlendirilebilmesi için sentetik ESK'lar üreten ve dağıtan bir metot tanımlanmıştır. Bu metotta; temelde, geçmişe ait gerçek sağlık kayıtları, uzmanlar tarafından oluşturulmuş hastalık modeli, hastalığın yayılma modeli gibi birçok model girdi olarak alınıp sentetik veri üretimi sağlanmıştır. Bireysel mahremiyetin korunması adına sentetik veri üretimi başlangıcında RBNR [28] (The Realistic But Not Real) metodu olarak bilinen ve kayıt sahiplerinin yaş ve muayene günlerinin değerlerinin belirli sınırlar arasında değiştirilmesine dayanan veri bozma tekniği kullanılmıştır. Sonuç olarak bu sentetik kayıtlar hastalık başlangıç durumu, laboratuvar tahlil ve sonuçlarını da içeren gerçek tıbbi kayıtların bir taklidi olarak üretilmektedir.

Sentetik ESK'ların yapımı ve doğrulanması amacıyla yapılan çalışmada [29]; gerçek ESK'ları bir model olarak kullanarak sentetik ESK üretebilmek için bir dizi algoritma tanımlanmıştır. Çalışmada; algoritmaları ve protokolleri test etmek, geliştirmek veya doğrulamak için standartlaştırılmış EMR kümelerinin olmadığı dolayısıyla tutarlı ve tam sentetik ESK üretilmesi gerektiği savunulmuştur. Çalışmanın amacı; bir pilot gruptaki topluluk için sentetik temel ESK'lar üretmektir. Kullanılan gerçek ESK'lar, BioSense [30] programından bir dizi anonimleştirilmiş kayıt kümesi olarak temin edilmiştir. Temin edilen kayıtlar, muayene verilerinin

analizi (ilk şikâyet, muayene sonucu, çalışma ve son tespit), klinik aktiviteleri, laboratuvar ve radyoloji tetkik ve sonuçlarını içermektedir. Çalışma üç temel adıma dayanır;

1. Hastaların temel karakteristik özellikleri, hastalıklar ve yaralanmalar benzetilerek sentetik hastaların üretilmesi
2. Her hastanın alabileceği, gerçek verilerdeki modeli temel alan, tıbbi bakım modelinin tanımlanması
3. Bakım modelinin sentetik hastaya adaptasyonu

Aynı çalışma veriye dayalı yaklaşımlar için [31]'de tekrar ele alınmıştır. Veriye-dayalı sentetik ESK'lar oluşturmak; gerçek ESK'lardan, hastalar ile ilgili bakım örüntüsü çıkarmak, hastalık frekansını belirlemek, hastalık tablosu ve alt-hastalık tabloları belirlemekten ibarettir. Burada sentetik ESK'lar, gerçek ESK'larla aynı olmamalı onları taklit etmelidir. Fakat yapılan çalışmalar sonucunda üretilen sentetik kayıtlar da gerçek kayıtlarla aynı dağılımı içereceğinden (örneğin; 4-11 yaş aralığında, 07.03.2010 ile 07.08.2010 tarihleri arasında mide bulantısı şikayetiyle gelen hastaların çoğunun zehirlenmiş olması), hastaların sosyo-demografik özellikleri sentetik ESK üretmeden önce de anonimleştirilmelidir. Bu çalışmada anonimleştirilmiş ve temizlenmiş verilerin bile tek başına canlı-gözlem sistemlerinde kullanılamadığı görülmüş, dolayısıyla başta hastalık verileri olmak üzere gerçek verilerin tüm özellikleri, benzetim yaklaşımı kullanılarak (belirli değer aralıkları vererek) hasta modelleri oluşturulmuştur.

MIDAS (Models of Infectious Disease Agent Study) [32] projesinde; bulaşıcı hastalıkların yayılması temel alınarak sentetik veri üretilmesi için işaretler gösterilmektedir. Fakat bu aşamada diğer toplulukların altyapısı için veya başka bulaşıcı hastalık kurbanları için ESK'lar üretilememektedir.

Archimedes [33] projesinde; çoğu kronik koşullar göz önüne alınarak, hastalar için matematiksel modeller açıklanmıştır. Bu model psikolojik gidişatları, hastalığın etkilerini, testleri ve tehlikeleri içermekte ise de pratikte var olan ESK'lardaki çeşitliliği ve veri düzensizliğini içermemektedir.

MIMIC [34] projesinde; hastalığın başlangıcından gözlemlenen zaman serilerinin istatistiksel özellikleri üzerinden, tam sentetik zaman serileri olarak modellense de tam ESK'lar üretilememektedir. Normal bir bayandan alınan bir hormon salgısının

zaman serileri deneysel olarak gözlemlenmiş ve taklit edilerek o hormonun sentetik veri grupları [35]'de yapılan çalışmada oluşturulmuştur. Bazı araştırmacılar, algoritmaları değerlendirmek ve salgın tespit performansını EARS (Early Aberration Reporting System) ile karşılaştırmak için, yarı-sentetik veriler (rastlantısal veri akışlarına, salgın zaman çizelgesini ekleyerek oluşturulan veriler) kullanmışlardır [36].

2.5. Mahremiyet Korumada Kullanılan Bilgi Kaybı Metrikleri

Mahremiyet korumalı metotlar geliştirilirken, işlem sonunda elde kalacak olan bilgi de göz önünde bulundurulmalıdır. Verilen mahremiyet ölçütlerini sağlayan anonim bir T' tablosu üretmek ve mümkün olabildiğince az veri kaybına sebep olmak, genel bir anonimleştirme problemidir. Anonimleştirme işleminin bir amacı mahremiyet korumayı sağlamaksa diğer amacı da anonim verinin pratik olarak faydasını korumaktır. Anonimleştirilmiş veri kümesinden elde edilecek veri faydasını ölçmek için farklı metrikler bulunmaktadır. Bir bilgi metriği, orijinal tablodaki veri kalitesini göz önüne alarak anonim tablodaki veri kalitesini ölçer. Literatürde bulunan bazı bilgi kaybı metrikleri, bunların avantaj ve dezavantajları aşağıda özetlenmiştir.

2.5.1. En az Bozukluk (Minimal Distortion-MD)

Kısaca MD olarak adlandırılan bu metrik, orijinal veri ile anonim veri arasındaki benzerliği ölçer. Genelleştirilen veya baskılanan her değere bir ceza verilmiştir. Bu metrik [11, 13, 37]'de kullanılmıştır. Örneğin, 3 meslek değerinin 1 kategoride genelleştirilmesi 3 birim bozukluğa, 20 mesleğin 1 kategoride genelleştirilmesi veya baskılanması 20 birim bozukluğa sebep olur. Bu metrik her bir özniteliğe tek tek uygulanan bir metriktir.

2.5.2. Ortalama Denklik Sınıfı Boyutu (Average Equivalence Class Size)

Veri dönüşümleri sonucunda elde edilen denklik sınıflarının boyutunu temel alan bir bilgi kaybı metriğidir. Orijinal veri kümesindeki yarı tanımlayıcıların gerçek değerini hesaba katmadan bilgi kaybının hesaplandığı genel amaçlı bir metriktir [38].

2.5.3. Ayırt Edilebilirlik Metriği (Discernability Metric-DM)

Anonimleştirmedeki veri faydasını ölçmek için kullanılan genel amaçlı bir metriktir [39]. Ayırt Edilebilirlik Metriği (DM), yarı tanımlayıcılar üzerinden verilen belirli cezalarla ortalama denklik sınıflarının boyutunu azaltabilmek için çalışır. Çünkü aynı denklik sınıfı içerisinde ne kadar çok kayıt var ise o kadar az belirgin bilgi o kayıtlar için korunur. Orijinal veri kümesindeki yarı tanımlayıcıların gerçek değerini hesaba katmadan bilgi kaybının hesaplandığı genel amaçlı bir metriktir. Eğer bir kayıtın bulunduğu grup boyutu |T[qid]| ise, o kayıta verilen ceza da |T[qid]| kadardır. Böylece, bir denklik sınıfının cezası $|T[qid]|^2$ olacaktır. Dolayısı ile genelleştirilmiş bir T tablosunun toplam cezası aşağıdaki gibi hesaplanır.

$$DM\ (T) = \sum_{\mathrm{qid(i)}} |\mathrm{T[qid]}|^2$$

3. İLGİLİ ÇALIŞMALAR

Günümüze kadar, bireysel veya kurumsal anonimliği sağlamak adına birçok çalışma yürütülmüştür. Özellikle son yıllarda, bireysel mahremiyetin gizliliğini ve güvenliğini sağlayabilmek için Mahremiyet Korumalı Veri Madenciliği (Privacy-Preserving Data Mining (PPDM)) ve Mahremiyet Korumalı Veri Yayını (Privacy-Preserving Data Publishing (PPDP)) yaklaşımları kapsamlı olarak çalışılmıştır. Ek olarak problemi çözmek adına, istatistiksel çıkarım toplulukları, kriptografik iletişim toplulukları ve veri tabanı kullanım toplulukları gibi pek çok araştırma topluluğu tarafından da bu konuda araştırmalar ve çalışmalar yapılmıştır. Mahremiyet Korumalı Veri Madenciliği (PPDM) alanında yapılan çalışmalar, kişisel veya kurumsal gizliliğin korunması adına pek çok algoritma ve modele sahiptir [4]. Bu yaklaşımın amacı, geleneksel veri madenciliği teknikleriyle büyük miktardaki verilerden ilgili çıkarımın yapılması ile beraber aynı zamanda hassas verinin korunmasıdır. Mahremiyet Korumalı Veri Yayını (PPDP) ise veri madenciliği sonuçlarından ziyade, özel veriler içeren tabloların veya veri tabanlarının yayımlanabilmesi üzerine odaklanmıştır [5].

PPDP, kayıt sahiplerinin kimliklerini gizleyerek veriyi anonimleştirir ve güvenle yayımlanabilmesini sağlar. İstatistiksel çıkarımlar üzerine çalışan topluluklar, istatistiksel tablolar için mahremiyet korumalı yayınlama üzerine çalışırken, bu alandaki pek çok güncel çalışmada İstatistiksel Açığa-çıkarma Kontrolü (Statistical Disclosure Control SDC) metotları üzerine odaklanmıştırlar [6]. Veri tabanı toplulukları genellikle, veri tabanı bölümleme ve sorgu denetleme gibi farklı metotları kullanarak hassas verinin güvenliğinin sağlanması için diğer topluluklarla ortak çalışmalar yürütmüşlerdir [7]. Kriptoloji topluluğu hassas verinin ifşa edilmeden farklı kurumların verilerini güvenli bir şekilde paylaşmaları için genel fonksiyonlar üzerine yoğunlaşmışlardır. Bu alanda son zamanlarda yapılan çalışmalarda Güvenli Çok-parçalı Hesaplama (Secure Multiparty Computation (SMC)) üzerinde durulmuştur [8]. Bazı durumlarda çalışmaların paralel hatları oldukça benzer olmasına rağmen topluluklar geniş çaplı bir perspektif edinmek için etkili bir şekilde birleşememişlerdir.

Bu tez çalışmasında konuyla alakalı geliştirilen birçok yaklaşım temel alınmıştır. Aşağıda bu çalışmalar kısaca özetlenmiştir.

3.1. Veri Tabanında Tutulma Şekline Göre Veri Türleri

Mahremiyet korumalı yaklaşımlarda, veri tabanlarında tutulma şekillerine göre veriler ve onlara bağlı olarak mahremiyet koruma çalışmaları da farklılık göstermektedir. Literatürde ele alınan veri türleri ve o türlere bağlı yapılan çalışmalar aşağıda tanıtılmıştır.

3.1.1. Homojen Veriler

Homojen veya yapısallaştırılmış veriler, veri tabanlarında belirli özellikler ve o özelliklere bağlı değerler olarak tutulmaktadır. Her satır bir kaydı ve her sütun o kayıtın bir özelliğini temsil etmektedir. Her değerin bağlı olduğu bir öznitelik ve bir kayıt vardır. Birçok mahremiyet korumalı çalışma homojen tekil bir veri tabanı üzerinden geliştirilmişken, dağıtık veri tabanlarında bu homojen veri tabanları üzerinden yatay ve dikey bölümlenmiş olmak üzere iki farklı çalışma ele alınmıştır.

3.1.2. Yatay Bölümlenmiş Dağıtık Homojen Veriler

Özellikle aynı hizmeti veren fakat dağıtık halde bulunan (sağlık hizmetleri gibi) kurumların, farklı kayıt sahipleri ile ilgili aynı öznitelikleri kayıt altına aldıkları farz edilmiştir. Farklı veri tabanlarında farklı kayıtlar için aynı öznitelik değerlerini tutan veri tabanlarına Yatay Bölümlenmiş veri tabanları denir. Bu veri tabanları üzerinden mahremiyet korumalı ortak hesaplamalar yapabilecek [40] örneğin Naive Bayes Sınıflandırma [41], SVM Sınıflandırma [42], Birliktelik Analizi [43] ve Kümeleme [44-46] gibi birçok mahremiyet korumalı veri madenciliği çalışması yürütülmüştür. Ayrıca, İşbirlikçi Filtreleme [47] ve ortak istatistiksel analizler yapan [48, 49] çalışmalar da yürütülmüştür.

Bu tez çalışmasında temel aldığımız çalışmalardan olan ve yatay bölümlenmiş dağıtık homojen veri tabanları için mahremiyet korumalı veri yayını yapan çalışmalar da gerçekleştirilmiştir [50, 51].

3.1.2.1. Dikey Bölümlenmiş Dağıtık Homojen Veriler

Özellikle ortak bir yapıya bağlı fakat aynı veri sahiplerine dağıtık halde farklı hizmetler sunan kurumların bulundurdukları yapıdır. Yani, aynı kayıt sahipleri ile ilgili, dağıtık veri tabanlarında farklı özniteliklerin tutulduğu farz edilmiştir. Farklı

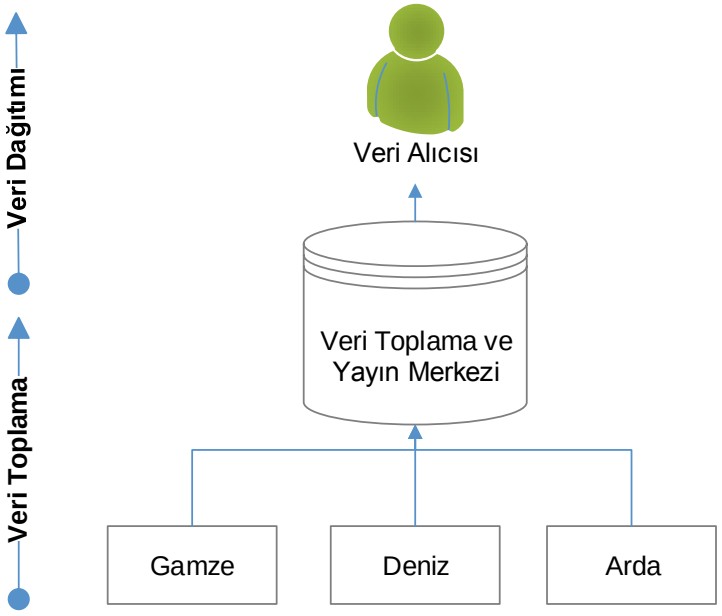
veri tabanlarında, aynı kayıt sahipleri için farklı öznitelikler tutan veri tabanlarına Dikey Bölümlenmiş veri tabanları denir. Bu veri tabanları üzerinden mahremiyet korumalı ortak hesaplamalar yapabilecek [52], örneğin Naive Bayes Sınıflandırma [53], SVM Sınıflandırma [54], Birliktelik Analizi [7] ve Kümeleme [55] gibi birçok mahremiyet korumalı veri madenciliği çalışması yürütülmüştür. Ayrıca, dikey bölümlenmiş homojen veriler üzerinden Karar Ağacı Sınıflandırması [56] ve kriptografik yaklaşım çalışmaları da yürütülmüştür.

3.1.3. Heterojen Veriler

Hastalar ile ilgili yazılmış tıbbi dokümanlar, raporlar vb. gibi heterojen veya yapılandırılmamış veriler de veri tabanlarında tutulabilir. Bu tür yapılandırılmamış verilerin anonimleştirilmesi [57] ve dağıtık halde bulunan bu tür veriler üzerinden mahremiyet korumalı işlemler yapılabilmesi üzerine çalışmalar yürütülmüştür [58]. Bu çalışmalar genel olarak heterojen veriyi, istatistiksel öğrenme yaklaşımı olan Şartlı Rastgele Alanlar (Conditional Random Field (CRF)) [59] tabanlı Varlık Adı Tanımlama (Named Entity Recognition (NER)) [60] yaklaşımını kullanarak homojen veriye dönüştürmek ve bu veriyi kimliksizleştirmek üzerine yoğunlaşmıştır.

3.2. Veri Toplama ve Veri Dağıtım Modelleri

Tekil bir veri tabanı elde etmek için geniş çaplı veri anonimleştirme veya mahremiyet korumalı veri yayını çalışmaları yapılmıştır. Veri toplama ve veri yayınının temel yapısı Şekil 3.1'de gösterilmiştir. Veri toplama aşamasında; bir veri toplayıcı, kayıt sahiplerinden veya kayıt tutan organizasyonlardan verileri toplar. Yayın aşamasında da veri toplayıcı, alıcı tarafların çeşitli analiz veya veri madenciliği işlemlerini yürütebilmeleri için, topladığı verileri yayınlar [5]. Şekil 3.1'de gösterilen yapı kayıt sahipleri ile ilgili verilerin yayını için yetersiz ve tehlikeli bir yapıdır. Çünkü bu tür veriler kayıt sahiplerinin hassas bilgilerini içerebilmektedir. Bu yapıdaki veriler kayıt sahiplerinin bilinmesini istemedikleri özelliklerini ifşa etmede ve onları tanımlamada kullanılabilir. Özellikle bireysel mahremiyetin korunması adına, veri yayınlama işlemlerinde, yayınlanacak verilere bir mahremiyet koruma katmanı eklenmelidir.

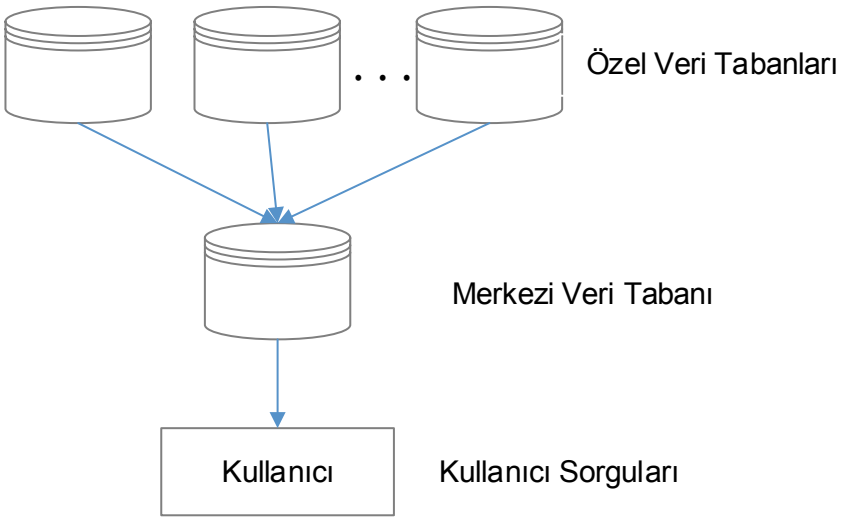


Şekil 3.1. Veri Toplama ve Veri Yayımlamanın Basit Hali

Yapılan çalışmalarda veri toplama ve yayınlama işlemleri için iki tip yapı kullanılır. Bu yapılar veri sağlayıcı kurumların güven durumlarına göre belirlenmiştir. Kullanılan iki yapı ve hangi durumlarda kullanılabileceği aşağıda belirtilmiştir.

3.2.1. Güvenilir Üçüncü Taraf Yapısı

Veri yayıncısının güvenilir olduğu ve kayıt sahiplerinin, mahremiyetin korunacağına güvenerek verilerini güvenle paylaştığı yapıdır. Bu yapıda veri yayıncısı kurumlardan veya kişilerden topladığı verilerin ya analiz sonuçlarını anonimleştirip yayınlar, ya da doğrudan verileri anonimleştirip yayınlar. Bu yapıda, merkezi sistem görevi yürüten üçüncü tarafın güvenilmez olması veya bir saldırı karşısında savunmasız kalması durumunda, tüm veri sahiplerinin mahremiyeti tehlikeye gireceğinden genellikle özel ve küçük hesaplamalar gerektiren işlemler için uygun değildir. Fakat büyük çaplı analizler yapan veya dağıtık birçok küçük yapının ortak verilerini kullanmak isteyen yapılar için, gerekli güvenlik önlemleri alındığı sürece ideal bir yapıdır. Şekil 3.2'de güvenilir bir merkezi sorgulama yapısı örneği gösterilmektedir.



Şekil 3.2. Güvenilir Merkezi Sistem Örneği

Yapılan bu tez çalışmasında, sağlık hizmetlerinde kullanılmak üzere Güvenilir Merkezi bir üçüncü taraf yapısı önerilmektedir.

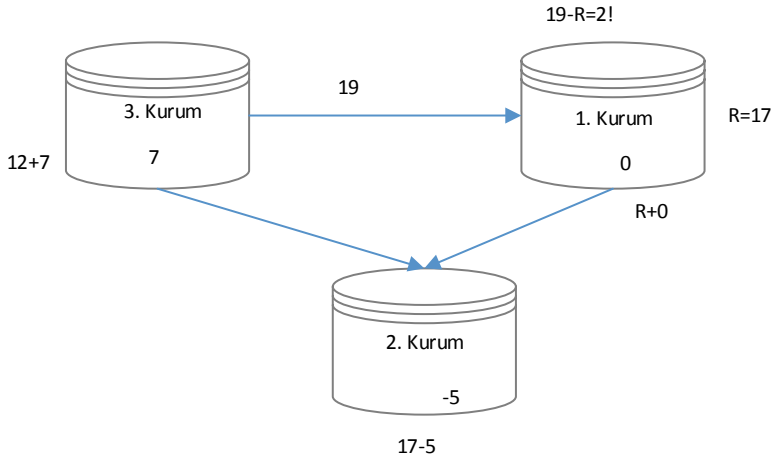
3.2.2. Güvenilmeyen Ortamda Dağıtık Verilerden Ortak Hesaplamalar

Veri yayıncısının güvenilir olmadığı yapıdır. Bu yapıda kayıt sahipleri veri yayıncısına veya yayıncının birlikte çalıştığı kişi veya kurumlara güvenmezler. Bu yapı, aslında ayrıca bir veri yayıncısının olmadığı yapıdır. Bazı kurumların verileri üzerinden ortak analizler yapmak istemesi fakat kendi verileri ile ilgili bilgilerin diğer kurumlar tarafından bilinmesini istememeleri temel alınarak geliştirilmiştir. Bu yapıda, veri sağlayıcı kurumlar arasında bazı iletişimler gerçekleştirilir. Bu iletişimlerde orijinal veride herhangi bir açığa çıkarma olmaması için iletişim rastgele seçilen bir anahtarla şifrelenebilir. Yapılan iletişimler sonucu her kurum kendi verisi dışında sadece yapılan analiz sonucunu öğreniyor ve diğer kurumların verileri ile ilgili ek bir bilgi alamıyor ise bu yapı güvenli yani mahremiyet korumalı olarak sağlanmış demektir.

Kurumlar arası güvenli iletişimlerle ortak analizler yapılabilmesine ortam sunan bu çalışmalar, özellikle veri madenciliği üzerinde geliştirilmiştir. İki kurum arasında mahremiyet korumalı olarak; karar ağacı (decision tree) sınıflandırması [61], iki veya daha fazla kurum arasında birliktelik kuralı (association rule) analizleri [7, 43] yapılmıştır. Ayrıca veri madenciliği çalışmaları ile ilgili, güvenli toplama (secure sum), güvenli küme bileşimi (secure set union), güvenli küme boyutları etkileşimi (secure size of set interaction) ve güvenli skaler çarpım

(secure scaler product) gibi birçok güvenli iletişim protokolü önerilmiştir [62]. Bu çalışmalarda bir yayın veya başka bir kurum ile veri paylaşımı söz konusu değildir.

Şekil 3.3'te toplama işleminin güvenli bir şekilde taraflar arası gerçekleştirildiği bir protokol gösterilmektedir. [62]'de önerilen bu protokol ile taraflar arası gerçekleşen şifreli işlemler sonucunda taraflar sadece kendi verileri ve sonuç bilgisine ulaşmaktadır, diğer tarafların verilerini görememektedirler.



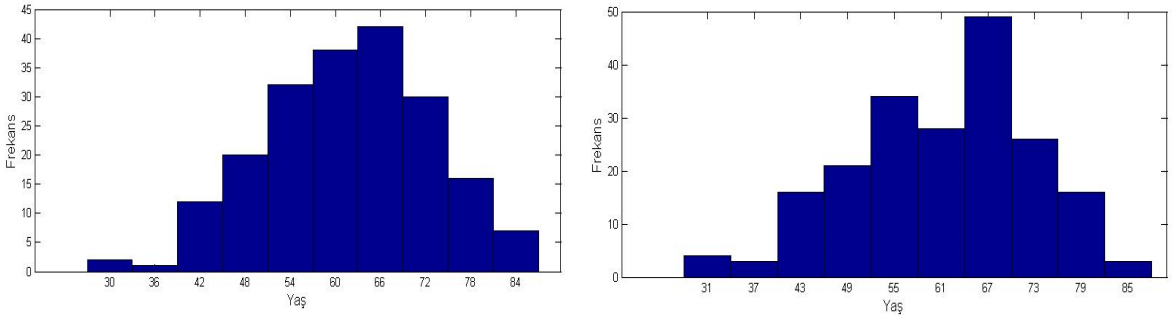
Şekil 3.3. Güvenli Toplam Hesaplaması

3.3. Mahremiyet Korumalı Veri Madenciliği

Mahremiyet korumalı veri madenciliği işlemleri için birçok algoritmik teknik değerlendirilmiştir. Çalışma [63]'de, konuyla alakalı geliştirilen algoritmalar ve modeller değerlendirilmiş, ilgi alanlarına göre bölümlenmiş ve farklı yaklaşımlarla karşılaştırılmıştır. Mahremiyet korumalı yaklaşımlarda veri madenciliği teknikleri, genellikle veriyi modifiye ederek veya veri dönüşümleri yaparak mahremiyet ihlalini engellemeye çalışmışlardır.

Mahremiyet korumalı veri madenciliği yaklaşımlarında, kayıt sahibinin gizliliği, çoğunlukla veri değerinin bozulmasıyla korunmaktadır [4]. Çalışmada orijinal veri dağılımına benzer bir dağılım üretebilmek için orijinal veriye gürültü eklenir. Gürültü olarak rastgele bir veri dağılımının bilgisi kullanılır. Gürültü eklenerek bozulmuş bu veriler ile kayıtların kimlikleri ifşa edilemeyecek ve bu veriler, veri madenciliği uygulamalarında güvenle kullanılabilecektir. Fakat [64]'de gösterildiği gibi; bir saldırgan kullanılan rastgele gürültüyü, orijinal gürültüden filtreleyerek gizliliği ihlal edebilir. Bu durumun yanında, bozulmuş veriler, orijinal veride

bulunmayan bilgiler içerdiğinden tam bir doğruluk sağlamaz dolayısıyla güvenilir değildir. Basit bir örnek olarak Şekil 3.4'te gerçek bir veri kümesindeki yaş değerinin dağılımı ve bu dağılımın ortalama ve standart sapma değerlerine bağlı kalarak üretilmiş bir dağılım görülmektedir. Bu örnekte ortalama ve standart sapma verileri kullanılarak bir olasılık yoğunluk fonksiyon değeri hesaplanmıştır. Bu kapsamda her bir öznitelik değeri, ortalama ve standart sapma değerlerine bağlı kalarak Normal Olasılık Yoğunluk Fonksiyonu (normal probability density function) kullanılarak rastgele olarak üretilmiştir. Şekil 3.4'te hem gerçek veri dağılımı hem de üretilen sentetik verilerin histogramları gösterilmektedir.



Şekil 3.4. Yaş Özniteliğinin Orijinal ve Sentetik Histogramı

Sağlık hizmetlerinde verilerin doğruluğu ve güvenilirliği hayati önem arz etmektedir. Bundan dolayı yapılan çalışmada orijinal veriye benzer bir dağılımı kullanmaktan ziyade orijinal, gerçek verilerin elde edilebilmesi üzerine odaklanılmıştır.

3.4. Mahremiyet Korumalı Veri Yayını

Mahremiyet Korumalı Veri Yayını (Privacy-Preserving Data Publishing - PPDP) alanında yapılan çalışmalar, veri madenciliği veya veri analizi sonuçlarından ziyade, genellikle verinin yayın aşamasında mahremiyet korumasının sağlanması üzerine odaklanılmıştır. Mahremiyet korumalı veri yayıncılığı uygulamalarının temel amaçlarından biri, kayıt sahibinin kimliğini gizleyerek veriyi anonimleştirmeyi başarabilmektir. Çalışmalar genellikle, hassas bilgi çıkarımı, arka plan saldırısı ve çeşitli bilgi metrikleri ile gizliliğin ölçülmesi üzerine odaklanılmıştır. Mahremiyet korumalı veri yayıncılığı uygulamalarında mahremiyet korumasını sağlayabilmek için, randomizasyon [4] ve genelleştirme gibi çeşitli veri dönüşüm teknikleri kullanılmaktadır. Günümüzde birçok mahremiyet korumalı veri yayını tekniği, veri

madenciliği veya veri analizlerinin sağlıklı bir şekilde yapılabilmesi için, yayınlanabilir anonim verilerin kullanımına göre değerlendirilmiştir. Yapılan çalışmada; farklı kuruluşlarla, kişisel sağlık kayıtlarının anonim bir şekilde paylaşılabilmesi adına, mahremiyet korumalı veri yayını çalışmaları önemli bir rol oynamaktadır.

3.4.1. Dağıtık Veriler için K-Anonimlik Algoritmaları

Farklı alıcılardan toplanan veriler üzerine anonimleştirme yapabilmek ve istenen veri faydası ve mahremiyet dengesini sağlayabilmek, anonimlik uygulamalarında hedeflenen amaçlardan biridir. Yatay bölümlenmiş verilerde k-anonimlik uygulanması [65]'de tartışılmıştır. Her veri alıcısının aynı zamanda kayıt sahibi olan bir müşteri olduğu farklı durumlar üzerine çalışılmıştır. Veri kayıtlarının hem hassas bilgiyi hem de yarı-tanımlayıcı bilgileri içerdiği farz edilerek; k tane aynı yarı-tanımlayıcı özniteliklere sahip kayıt bir araya gelene kadar hassas bilgiler şifreli halde kalmaktadır. İki taraf arasında dikey bölümleme ile k-anonim protokolü [66]'de tanımlanmıştır. Genellikle yapılan çalışmalarda iki tarafın, aynı yarı-tanımlayıcılar üzerinde, aynı genelleştirme aralığında anlaşması sonucu ortak çalışmalar yürütülmüştür. Veri dağıtılmadan önce, iki tarafın anlaşmasına göre genelleştirmenin nasıl yapılacağı [67]'de tartışılmış ve k-anonimlik sağlanmıştır.

3.5. Çok Boyutluluk ve Veri Bölümleme

Anonimleştirme yaklaşımlarının kullanımında genellikle veri kümeleri çok boyutlu yapıdadır. Veri kümelerinin boyutunun artması anonimleştirme işlemlerinde hesaplama maliyetinin de artması anlamına gelmektedir. Anonimleştirme yaklaşımları aşırı hesaplama maliyetinden dolayı çok boyutluluk karşısında zayıf kalmaktadır. Optimal bir k-anonimleştirme işleminin NP-Hard (Non-deterministic Polynomial-time Hard) olduğu [68]'da gösterilmiştir. Çok boyutlu veri kümelerinde anonimleştirmenin zorluğu ve bu veri kümelerinin ne tür saldırılardan etkilenebileceği [69]'de tartışılmıştır.

Bilgi paylaşımı, mahremiyet ve güvenlik uygulamalarında, veri bölümleme yaklaşımları, genellikle kriptografik metotlar kullanılarak uygulanmıştır [8]. Farklı gruplar arası bilgi paylaşımı kriptografik protokollerin kullanımıyla güvenlik sağlanmasını gerektirir [62]. Yapılan çalışmada bu gereklilik üzerinde

durulmamıştır. Çünkü, gösterilen modellere isteğe göre farklı güvenlik protokolleri entegre edilebilir. Yapılan çalışmada, verilerin, veri toplayıcılarından farklı veri alıcılarına nasıl doğru bir şekilde iletilebileceği üzerinde durulmuştur. Farklı veri alıcılarının farklı bilgi ihtiyacı olduğu göz önüne alındığında, veri bölümleme teknikleri kullanılarak, bu ihtiyacın veri faydası ve mahremiyet dengesini en iyi şekilde sağlayacak yapılar içerisinde karşılanması sağlanmıştır.

Literatürde; yatay bölümleme, dikey bölümleme ve hibrit bölümleme olmak üzere 3 tip bölümleme tekniği, veri paylaşımı yöntemlerinde kullanılmaktadır. Yatay bölümlemede [70], farklı veri sahiplerinin aynı öznitelikleri alınırken, dikey bölümlemede [71], aynı veri sahiplerinin farklı öznitelikleri ele alınır. Hibrit bölümlemede ise; yatay ve dikey bölümleme ya aynı anda ya da ardışık olarak kullanılır. Konu ile ilgili ayrıntılı bilgi Bölüm 3.1'de verilmiştir. Yapılan çalışmada; farklı grupların veri ihtiyacının belirlenmesi adına yatay bölümleme, çok boyutluluğun maliyetinin azaltılması adına ise dikey bölümleme metotlarının kullanımıyla, genel olarak veri faydasını maksimize edebilmek için hibrit bölümleme kullanılmıştır.

4. DAĞITIK VERİ KÜMELERİNDE ANONİMLEŞTİRME VE PAYLAŞIM MODELLERİ

Son yıllarda, yayımlanacak tek bir veri kümesi üzerinde gizliliğin korunması için potansiyel çözüm sunan birçok anonimleştirme ölçütü ve veri yayın stratejisi geliştirilmiştir. Yapılan çalışmaların büyük bir kısmı, genelleştirme, baskılama, gürültü ekleme veya permutasyon gibi yöntemlerle verileri k-anonimlik gibi mahremiyet ölçütlerine uygun hale getirecek algoritmalar üzerine yoğunlaşmıştır. Böylece istenen veri kümesine bu algoritmaların uygulanmasıyla anonimliğin sağlanması hedeflenmiştir [5]. Fakat başta sağlık sektörü olmak üzere birçok alanda veriler dağıtık kurumlar tarafından kayıt altına alınmakta ve bu dağıtık verilerden toplu olarak anlamlı çıkarımlar yapabilmek, araştırma toplulukları açısından çok zorlaşmaktadır.

Bu tez çalışmasında dağıtık kurumlar tarafından kayıt altına alınan verilerin araştırma toplulukları ile paylaşılamaması probleminin çözümü üzerine odaklanılmıştır. Bu sorunun çözümü için, verilerin kurumlardan toplanıp mahremiyet korumalı bir şekilde araştırma topluluklarına ulaşmasını sağlayan bir sistem modeli oluşturulması gerekmektedir. Bu yapıyı sağlayacak birçok sistem geliştirilebilir fakat hem veri faydası hem de veri gizliliği bakımından en ideal dengeyi sağlayacak sistemi oluşturmak zordur.

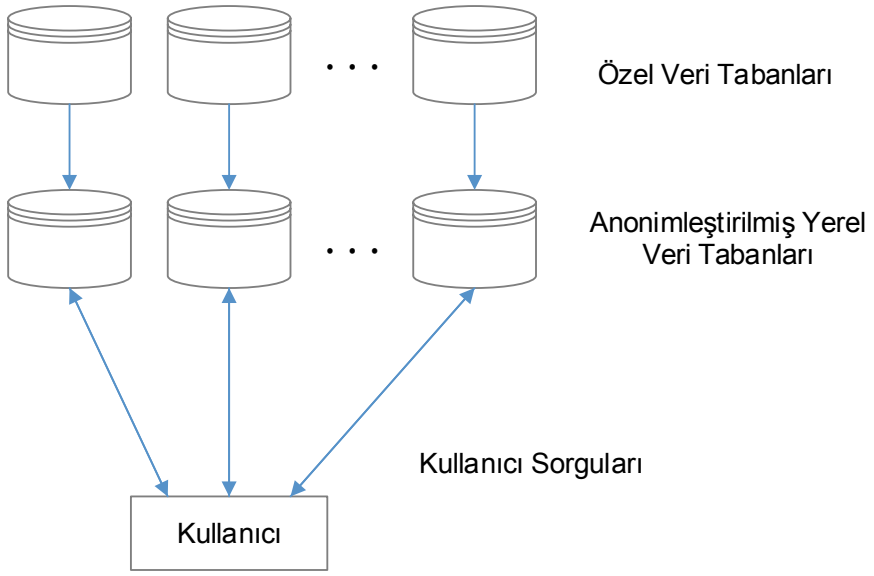
Bu tez çalışmasında önerilen model, literatürde aynı problemin ele alındığı, Emory Üniversitesi tarafından geliştirilen HIDE (Health Information DE-identification) [72] projesi kapsamında önerilen model ile karşılaştırılmıştır.

4.1. HIDE Projesinde Önerilen Model

Sağlık hizmetlerinde dağıtık halde bulunan yapılandırılmış veya yapılandırılmamış verilerden ortak paylaşımlar yapılabilmesini sağlamak adına Emory Üniversitesi tarafından HIDE (Health Information DE-identification) projesi geliştirilmektedir. Bu projede dağıtık verilerin paylaşılması üzerine bazı yaklaşımlar gösterilmektedir [50, 51, 58].

Çalışmada, merkezi veya merkezi olmayan veri paylaşım modeline göre olası iki yaklaşım gösterilmiş ve bir model önerilmiştir. Gösterilen yaklaşımlardan ilki, basit bir yaklaşım olarak her kurumun bağımsız bir şekilde kendi veri kümesini

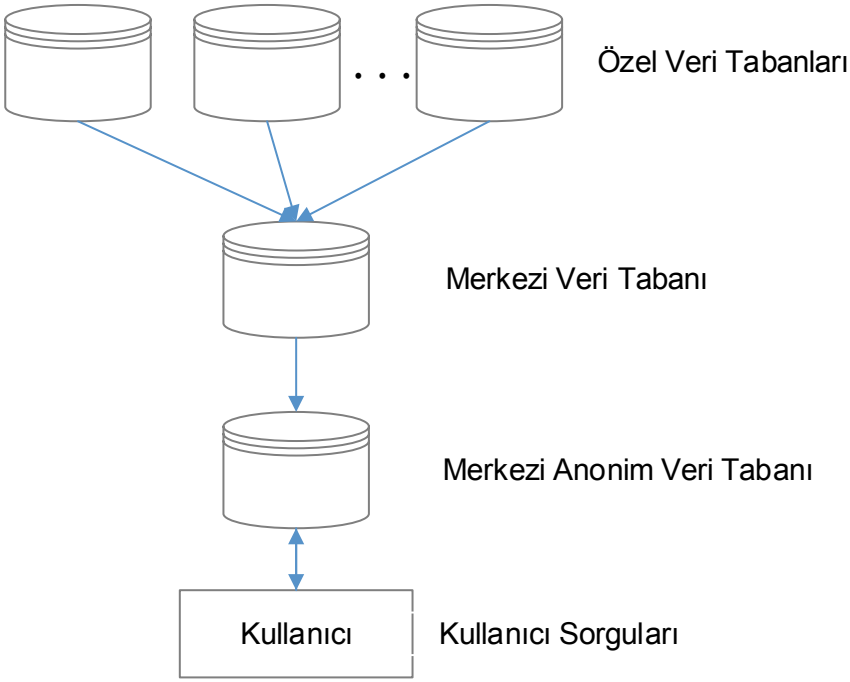
anonimleştirdiği ve elde edilen anonim veri kümesini sorgulamaya açık hale getirdiği bir sistem modelidir. Bahsedilen model Şekil 4.1'de gösterilmiştir. Bu yaklaşımın en temel sorunu, her kurum bağımsız olarak kendi verisini anonimleştirdiğinden dolayı çok fazla veri kaybının ortaya çıkacak olmasıdır. Ek olarak hangi anonimleştirilmiş verinin hangi kurumdan geldiği de açığa çıkacaktır.



Şekil 4.1. Bağımsız Anonimleştirme Modeli

Şekil 4.1'deki yaklaşım, merkezi olmayan veri dağıtım yapısına uygun bir yaklaşımdır. Hem yüksek veri kaybı hem de tüm dağıtık kurumlara doğrudan bağlantının sebep olacağı iş yükü ve güvenli bağlantı maliyeti, bu yaklaşımı kullanışsız kılmaktadır.

[50] ve [51]'da ilk yaklaşıma alternatif olarak; her kurum tarafından güvenilen bir üçüncü tarafın varlığını kabul ederek işlemleri bu merkezi ve güvenilir yapı üzerinden gerçekleştiren bir sistem modeli gösterilmiştir. Bu alternatif yapıda; veri sahipleri, üzerinde hiçbir değişiklik yapmadan veri kümelerini merkezi sisteme iletirler, güvenilir merkezi sistem homojen yapıda ve yatay bölümlenmiş bu veri kümelerini birleştirerek ortak (bütünleşik) bir veri kümesi elde eder. Daha sonra bu anonim veri kümesi, merkezi sistem üzerinden sorgulamaya açık hale getirilir.



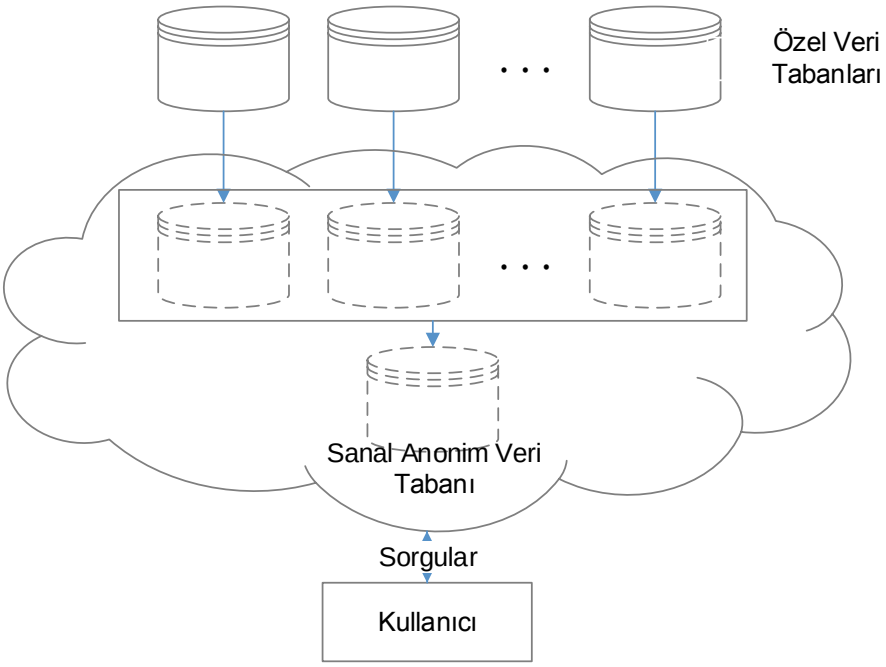
Şekil 4.2. Güvenilir Merkezi Anonimleştirme Modeli

Bu yaklaşımın en önemli sorunu çok fazla miktarda verinin tek seferde anonimleştirilmesidir. Çok boyutlu ve büyük veri kümelerinde anonimleştirme işlemlerinin NP-Hard olduğu [73]'de ispatlanmıştır. Dolayısı ile merkezi sistemin zorlu işlemleri kaldırabilecek yapıda olması gerekmektedir.

Şekil 4.2'deki modeldeki gibi güvenilir bir merkezi üçüncü taraf yapısı her zaman uygun olmamaktadır. Saldırganlar tarafından sunucunun ele geçirilmesi durumunda, tüm kurumlar ve veri sahipleri ile ilgili mahremiyet ortadan kalkacaktır. Dolayısı ile veri alıcıların merkezi yapıya doğrudan erişimleri sakıncalıdır.

Yukarıda gösterilen iki olası yaklaşıma ek olarak HIDE projesi kapsamında, dağıtık verilerin anonimleştirilip paylaşımı ile ilgili önerilen yapı Şekil 4.3'te gösterilmiştir [51, 58]. Bu yaklaşımda; sanal, bütünleşik ve anonim bir veri tabanı üretmek için veri sağlayıcılar dağıtılmış protokollere katılımda bulunurlar. Dikkat edilmesi gereken, anonimleştirilmiş verilerin hala bağımsız veri tabanlarında bulunduğu ve bu verilerin birleştirilmesi ve tekrar anonimleştirilmesi güvenli dağıtılmış protokoller (secure distributed protocol) üzerinden gerçekleştirildiğidir. Bu yaklaşım temelini güvenli çok-taraflı hesaplamalar (Multi-party Computation (MPC)) probleminden alır [74, 75]. Lokal anonimleştirilmiş veri kümeleri güvenli birleştirme protokolleri (Secure Union Protocols) [43, 76] kullanılarak birleştirilebilir, sonrasında

yayımlanabilir veya sorgulanabilen sanal bir veri tabanında hizmet verebilir. Daha sonraki durumda, her bağımsız veri tabanı gelen sorguyu kendi anonim veri tabanında çalıştırır ve anonimliği garantilemiş sonuçları toplaması için güvenli birleştirme protokollerine bağlanır.



Şekil 4.3. HIDE Projesinde Önerilen Bağımsız Anonimleştirme Sanal Merkezi Yayın Modeli

HIDE projesinde önerilen Şekil 4.3'deki yapı ile Şekil 4.1'de gösterilen yapı arasında, sanal bir ortamda veri kümelerinin birleştirilmesiyle sadece güvenlik problemi iyileştirilmiştir. Önerilen modelde gelen sorgular doğrudan kurumlara gitmediğinden hem saldırganın doğrudan kurumların veri tabanına erişmesi engellenmiş hem de toplu sonuçların tek bir merkezi sorgu ile görüntülenmesi iş yükünü veri alıcı tarafından azaltmıştır. Fakat Şekil 4.1'de gösterilen ilk yapıdaki büyük miktardaki veri kaybı aynı kalmaktadır. Dolayısı ile sunulan çözüm, sunduğu veri faydası açısından yeterli değildir.

4.2. Yapılan Çalışma

Yapılan bu tez çalışmasında, mahremiyet korumalı ve veri faydası yüksek bir veri paylaşım modeli örneklenmiştir. Çalışmada, dağıtık kurumlardan toplanan yatay bölümlenmiş verilerin, belirli mahremiyet koruma ölçütlerine uygun olarak ve en az

veri kaybı ile başka kurumlarla paylaşılabilmesini sağlamak hedeflenmiştir. Bu sistemin, merkezi bir yapının yönetiminde ve alıcı kurumların ihtiyaçları doğrultusunda tasarlanması gerekliliği göz önüne alınarak ideal bir sistem modeli önerilmiştir.

Bu çalışmada önerilen modelin, HIDE projesinde önerilen modele göre 2 farklı yanı vardır. İlki, güvenilir üçüncü taraf yapısına göre tasarlanmasıdır. Güvenilir bir merkezi yapı vardır ve işlemler ağırlıklı olarak bu yapı üzerinden gerçekleşir. Diğer fark ise HIDE projesinde gösterilen modeller sorgulama tabanlı veri yayını üzerine geliştirildiğinden veri kaybı yüksektir. Fakat yapılan çalışmada veri dağıtımı söz konusudur ve sadece merkezi sistem tarafından onaylanan kurumlara istekleri doğrultusunda dağıtım yapılmaktadır. Bu durum paylaşılan verinin faydasını arttırmaktadır.

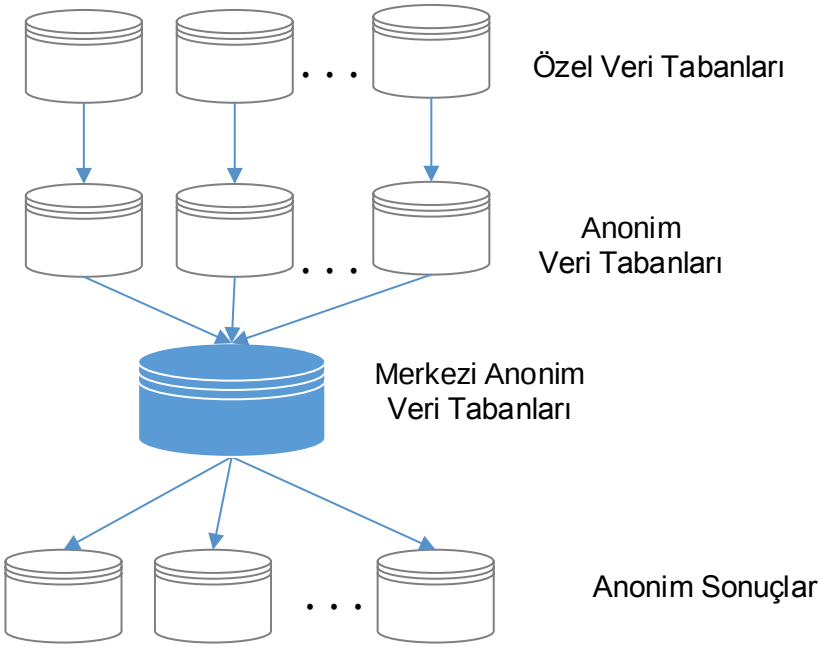
Bu çalışmada da HIDE projesinde olduğu gibi 2 olası model gösterilmiş ve bir model önerilmiştir. Önerilen bu modelin HIDE projesindeki modellerle benzerliği, farklılığı ve önerilen modele karşı avantaj ve dezavantajları aşağıda belirtilmiştir.

4.2.1. Güvenilmeyen Ortamda Veri Dağıtımı

Eğer veri sağlayıcısı, verilerinin orijinal haline kimsenin erişmesine izin vermek istemez ise (diğer veri sahipleri veya merkezi bir yapının erişimine izin vermeyecek şekilde paylaşmak ister ise) bu durum güvenilmeyen (untrusted) ortamlarda veri paylaşımına girer. Bu durumlar, anonimleştirme işlemlerinin veri sağlayıcısı tarafında yapılmasını gerekir. Her veri göndericisinin kendi veri kümesini bağımsız olarak anonimleştirdiği, güvenilmeyen üçüncü taraf (untrusted third party) yaklaşımına göre tasarlanmış bir yapı Şekil 4.1'de gösterilmektedir. Bu yapı; her veri göndericisinin, verilen anonimleştirme ölçütlerine göre (örneğin; k-anonim ve ℓ-çeşitlilik yöntemlerini k=10 ve ℓ=4 olacak şekilde) verilerini anonimleştirdiği yapıdır. Bu yapıda merkezi sistem güvenilmez olup veri sahipleri, verilerini doğrudan üçüncü bir tarafa göndermek istemezler. Dolayısıyla merkezi sistemde veriler toplanmadan önce dağıtık olarak anonimleştirilmiş olur.

Veri sahipleri, anonimleştirdikleri verilerini merkezi sisteme gönderirler ve merkezi sistem tüm bu verileri birleştirerek genel bir veri kümesi elde eder. Veri alıcı tarafların istekleri/sorguları doğrultusunda merkezi sistemde, doğrudan anonim veriler üzerinden, yatay ve dikey bölümleme teknikleri kullanılarak farklı tablolar

üretilir. Üretilen bu tablolara ihtiyacı olan alıcı tarafın erişebilmesi sağlanır. Bu senaryoda, veri gönderen taraflar için güvenli bir yapı sağlanmıştır. Ancak veri alıcı taraflar için bilgi kaybı çok yüksektir. Bu yapı, özellikle banka hesaplarında veri sahiplerinin gizliliğini maksimum seviyede tutacağından kullanılması uygundur, fakat sağlık hizmetlerinde kullanım için veri faydası az olacaktır.



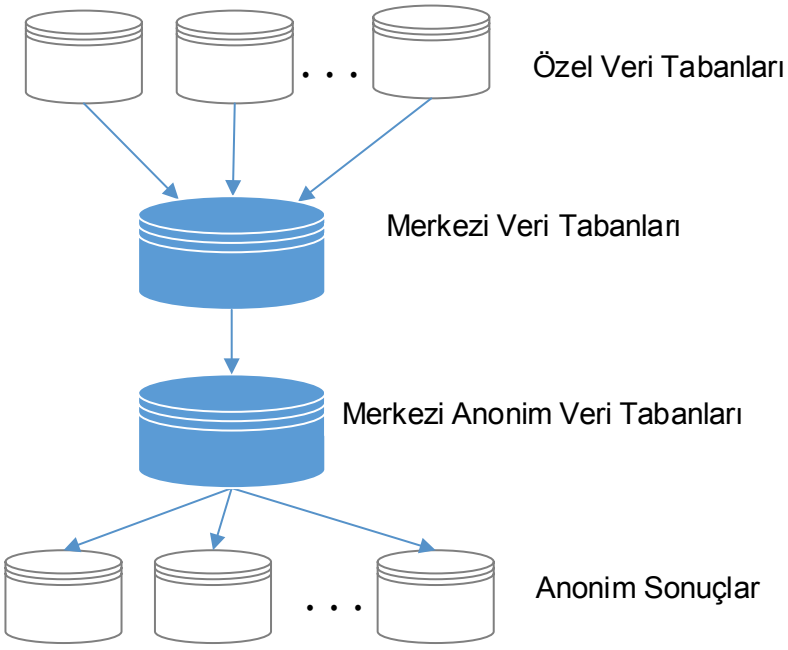
Şekil 4.4. Bağımsız Anonimleştirme ve Merkezi Dağıtım Modeli

Bu model HIDE projesinde önerilen ve Şekil 4.1'de gösterilen modellere çok yakın bir modeldir. Şekil 4.1 ile tek farkı sorgulama tabanlı olmamasıdır. Önerilen modelle tek farkı ise anonim verilerin sanal merkezi sistemde değil belirlenen bir kurumda bir araya gelmesidir. HIDE projesinde önerilen modeldeki farklı aşamalar da bu modelde kullanılabilir. Örneğin anonim veri tabanlarının bulunduğu merkezi yapıda, verilerin alındığı kurumların da anonimliğini sağlamak için ikinci bir anonimleştirme yapılabilir.

4.2.2. Güvenli Merkezi Sistemden Ortak Veri Dağıtımı

Veri sahiplerinin, bir Güvenli İletişim Protokolü (Secure Transmission Protocol) kullanarak verilerini merkezi sisteme ilettikleri yapı Şekil 4.5'de gösterilmiştir. Bu yapı; merkezi sistemin güvenilir olduğu, güvenilir üçüncü parti (trusted third party) yaklaşımına göre tasarlanmış bir yapıdır. Bu senaryoda; veri sahiplerinden toplanan orijinal verilerin, merkezi sistemde bir araya getirilip, tüm verilere aynı

anonimleştirme ölçütlerinin uygulanarak anonimleştirildiği yapı gösterilmektedir. Tüm verilerin oluşturduğu veri kümesinin anonimleştirilmesi ile veri kümesi, oluşabilecek birçok saldırıdan korunmuş olmasına rağmen bu yapı çok fazla işlem yükü gerektirecektir. Veri alıcı tarafların istekleri/sorguları doğrultusunda merkezi sistemde, doğrudan anonim veriler üzerinden, yatay ve dikey bölümleme teknikleri kullanılarak farklı tablolar üretilir. Üretilen bu tablolara ihtiyacı olan tarafın erişebilmesi sağlanır. Bu senaryodaki yapının, ilk senaryodaki yapıdan daha fazla veri faydası sağladığı fakat işlem maliyeti olarak daha maliyetli olduğu Bölüm 6.'da yapılan deneylerle gösterilmiştir.



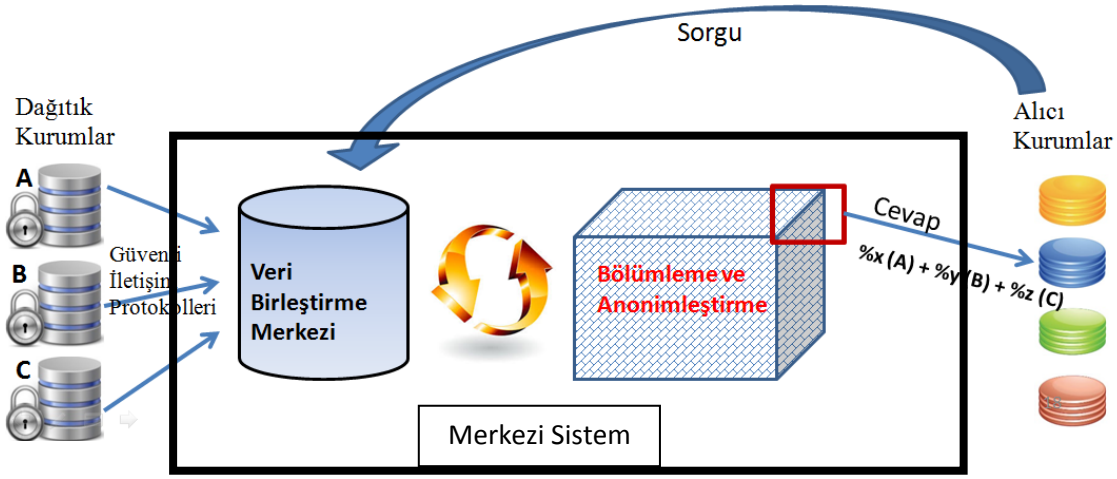
Şekil 4.5. Merkezi Anonimleştirme ve Merkezi Dağıtım Modeli

Bu model HIDE projesinde gösterilen olası modellerden Şekil 4.2'de gösterilen güvenilir merkezi anonimleştirme modeline yakın bir modeldir. İki model arasındaki tek fark birinin sorgulama tabanlı diğerinin ise dağıtım tabanlı olmasıdır.

4.2.3. Güvenli Merkezi Sistemden İdeal Dağıtım Modeli

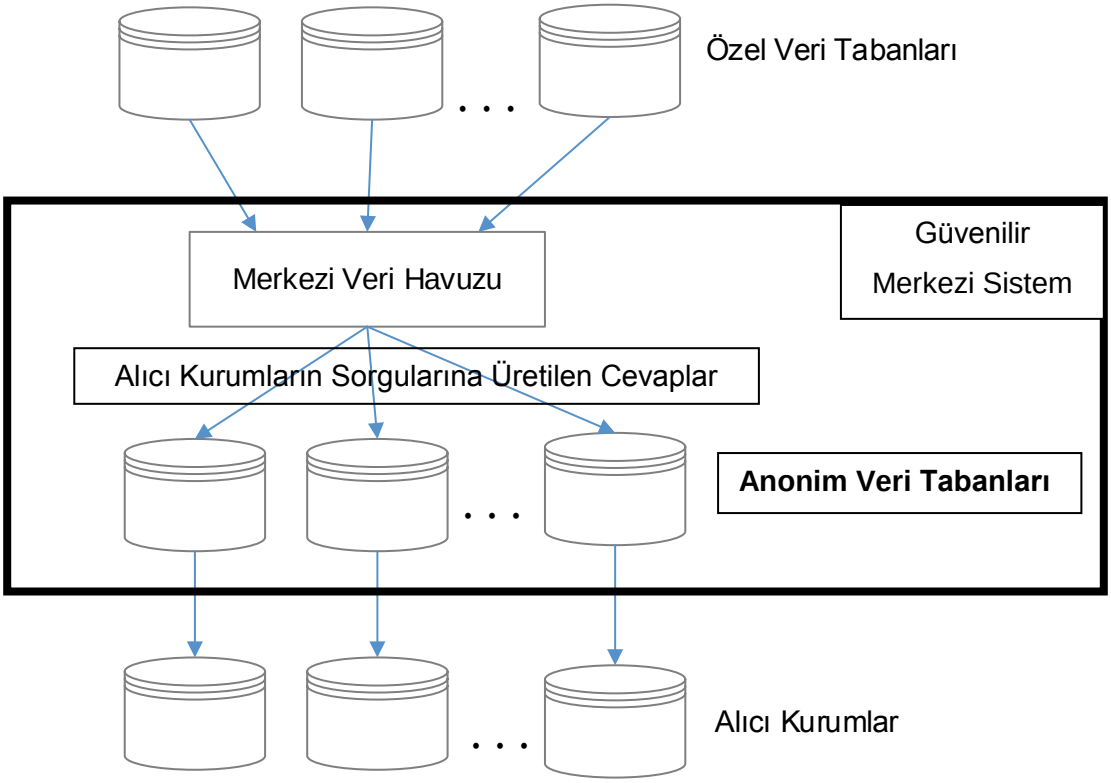
Gösterilen ilk iki yapıyla kıyaslandığında, maksimum veri faydası sağlayacak olan yapı Şekil 4.6'da gösterilmiştir. Bu senaryoda; veri sahipleri bir Güvenli İletişim Protokolü (Secure Transmission Protocol) kullanarak verilerini merkezi sisteme iletirler. Bu yapı; merkezi sistemin güvenilir olduğu, güvenilir üçüncü parti (trusted third party) yaklaşımına göre tasarlanmış bir yapıdır.

Merkezi veri tabanında, veri sahiplerinden toplanan orijinal veriler birleştirilir. Tüm veriler bir araya getirildikten sonra veri alıcı tarafların istekleri/sorguları doğrultusunda yatay ve dikey bölümleme teknikleri kullanılarak tablolar üretilir. Orijinal verilerden oluşan bu tablolar istenen ölçütlere göre anonimleştirilir. Anonimleştirilmiş tablolar, alıcı tarafların erişimine sunulur. Bu yapı kullanılarak veri faydasının arttırıldığı fakat tablo bağlama gibi bazı saldırılara karşı korumanın azaldığı ve bu tip saldırılar için anonimleştirme aşamasında δ-varlık gibi ek ölçütlerin belirlenmesi gerektiği, dağıtım senaryosundan anlaşılabilmektedir.



Şekil 4.6. Merkezi Dağıtım ve Dağıtık Anonimleştirme Modeli

Gösterilen mimarileri anlatan senaryolarda veri sahiplerinden alınan veri setlerinin homojen olduğu (tüm veri sahipleri aynı veri taslağına fakat farklı veri sahiplerinin bilgilerine sahip olduğu). Tüm veri sahiplerinin, bireysel sağlık kayıtlarına sahip sağlık hizmetleri sistemleri olduğu fakat veri alıcı tarafların sadece sağlık hizmetleri sistemleri olmadığı aynı zamanda farklı istatistiksel çalışmalar yürüten sistemler de olabileceği farz edilmiştir.



Şekil 4.7. Merkezi Dağıtım ve Dağıtık Anonimleştirme Taslağı

Daha basit ve anlaşılır bir yapıda göstermek gerekirse, Şekil 4.7'deki yapı önerilen modelin kaba taslak tasarlanmış halidir. Taslaktan da anlaşılacağı üzere her alıcı kurumun sadece ihtiyaç duyduğu bölümler anonimleştirilerek alıcı kurumların hizmetine sunulmaktadır.

5. MATERYAL VE METOT

Önerilen modelin gerçekleştirimi üç aşamadan oluşmaktadır. En önemli aşama merkezi sistemde toplanan verilerin alıcı kurumların talepleri doğrultusunda bölümlenmesi ve elde edilen her tablonun istenen anonimlik ölçütlerini sağlaması işlemidir. İlk aşama dağıtık kurumların veri tabanlarında bulunan verilerin merkezi sistemde toplanması ve birleştirilmesi işlemidir. Son aşama ise gerekli ölçütlere göre anonimleştirilmiş tabloların alıcı kurumlara dağıtılmasıdır. Aşağıda bu üç aşamayı gerçekleştirebilmek için kullanılan materyal ve metotlar detaylandırılmıştır.

5.1. Anonimleştirmede Kullanılan Ölçütler

5.1.1. K-Anonimlik

Mahremiyet korumalı sistemlerde, en yaygın kullanılan yöntem verilerin anonimleştirilmesidir. Anonimleştirme işlemlerinde amaç, kullanılan veri kümesinden kimlik tespiti yapılabilmesini engellemektir. Yani; bir saldırganın veri setinin özel bir parçasına bağlantı kurarak herhangi bir bireyle ilgili hassas bir bilgiyi açığa çıkarmasını engellemektir. Bir bireyin tanımlanabilme tehlikesini aza indirmek için önerilen en yaygın anonimleştirme yöntemlerinden biri k-anonimlik'dir [12]. K-anonimlik yönteminde, genelleştirme (generalization) ve baskılama (suppression) tekniklerinin, yarı-tanımlayıcı özniteliklere uygulanması ile bireysel kayıtların tekilleştirilmesi engellenerek gizliliği korunmaya çalışılmıştır [10]. Genelleştirme tekniği ile belirginliği azaltmak için, özniteliğin değeri belli aralıklara genişletilir. Baskılama tekniğinde ise, özniteliğin değeri tamamen kaldırılır. Bu durum genelleştirmenin son seviyesidir ve verinin değeri hakkında bilgi vermez. Bölüm 2'de orijinal veri ve anonimleştirilmiş veri örnekleri gösterilmiştir.

K-anonimlik yaklaşımı; veri kümesinde bulunan her kayıt için en az k-1 tane kayıttan, yarı tanımlayıcılar (QID) bakımından, ayırt edilemez olmasını gerektirir. En az k tane, yarı-tanımlayıcı öznitelikler bakımından birbirinden ayırt edilemeyen kayıtların oluşturduğu gruplara denklik sınıfı (equivalence class) ve bu kayıtların her birine öğe (tuple) denir. K-anonimlik sağlayan bir T' tablosu denklik sınıflarından oluşmalıdır.

Yapılan çalışmada kullanılan veri kümesinde k için; 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 değerleri verilerek veri kümesinin farklı k değerlerinde anonimleştirmede kaybettikleri veriler ile oluşan denklik sınıflarının boyutları karşılaştırılmıştır. Örneğin k=50 değeri için her denklik sınıfında en az 50 tane birbirinden ayırt edilemeyen kayıt bulunması istenmektedir.

5.1.2. L-Çeşitlilik

ℓ-çeşitlilik [14] yöntemi, öznitelik açığa çıkarma (attribute disclosure) olarak bilinen saldırılara karşı korunmak için önerilmiştir. Bu saldırılar ile saldırgan, veri kümesinin herhangi bir özel parçasına bağlantı kurması gerekmeden, bir bireyle ilgili ek bilgi çıkarımı yapabilmektedir. ℓ-çeşitlilik yöntemi, k-anonimlik yönteminden sonra, k-anonimlik yönteminin tamamlayıcısı olarak geliştirilmiştir. Çünkü k-anonimlik yöntemi anonimleştirmede yarı-tanımlayıcı öznitelikleri kullandığından hassas öznitelikler tarafından yapılabilecek çıkarımları engelleyememektedir. ℓ-çeşitlilik yöntemi, her denklik sınıfının en az ℓ tane iyi-temsil edilmiş (well-represented) hassas öznitelik içermesi gerektiğini, yani her denklik sınıfında en az ℓ tane ayrık hassas öznitelik değeri bulunması gerektiğini savunur.

Yapılan çalışmada kullanılan veri kümesinde ℓ için 2 değeri kullanılmıştır. Kullanılan hassas verinin değerlerine göre bu sayı arttırılıp azaltılabilir. Çalışmada bir örnek teşkil etmesi bakımından 2 değeri verilmiştir. Daha yüksek ℓ değerlerinin verilmesi önerilen model ile diğer modeller arasında aynı etkiyi yapacağından verilen değerin boyutu bu çalışma için fazla ayırt edici olmamaktadır.

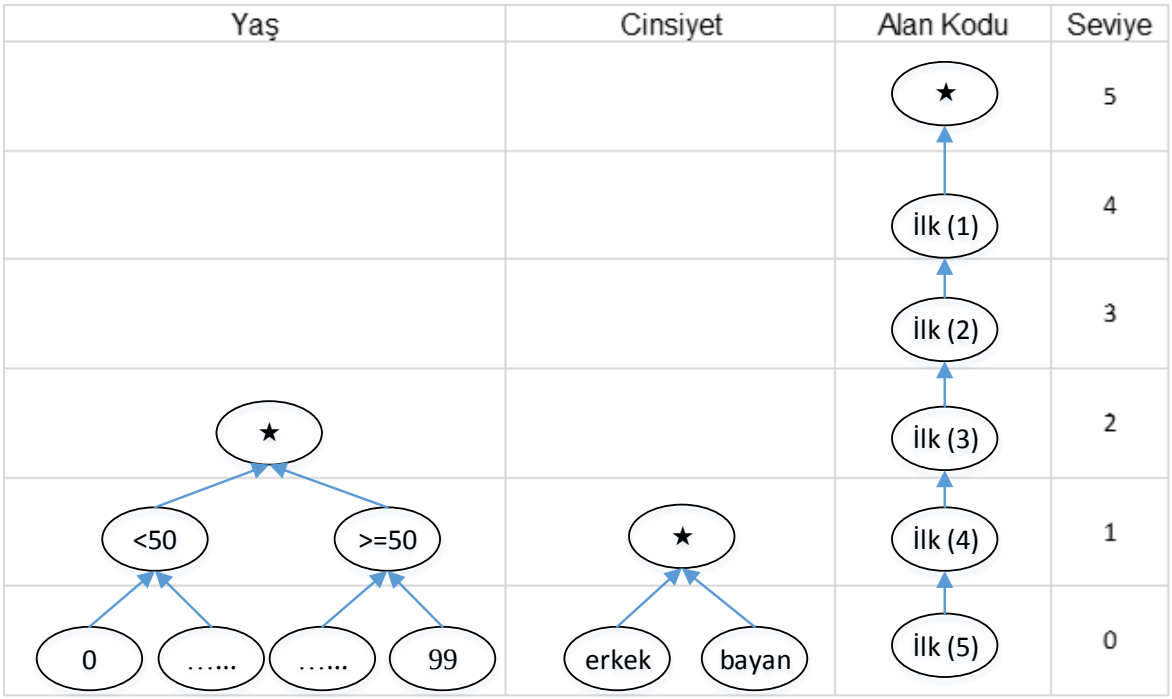
5.1.3. Genelleştirme Ölçütü

Yapılan çalışmada k-anonimliği ve ℓ-çeşitliliği sağlamak için veri dönüşüm tekniklerinden genelleştirme ve baskılama yöntemi kullanılmaktadır. Biyomedikal alanda, çok boyutlu global kodlama yapısını kullanan global-optimal anonimleştirme algoritmalarının kullanımı için tüm alan genelleştirmesi önerilmektedir. Global-optimal anonimleştirme tanımlanan optimal çözümlerden arama alanının oluşturulmasıdır, örneğin belirlenen metriğe göre en az bilgi kaybını sağlayan veri dönüşüm sonuçları bir arama alanı sağlamaktadır.

Çok boyutlu global kodlama yapısı aynı genelleştirme kuralının benzer aynı kayıt çiftlerine uygulandığı yapıdır. Verilerin genelleştirilmesi için, genelleştirme

ölçütlerinden tüm-alan genelleştirmesi (full-domain generalization) [22] kullanılmıştır. Tüm-alan genelleştirmesi, veri kümesinde bulunan özniteliklerin her birinin genelleştirme hiyerarşisine göre daha genel bir yapıya seviyeli olarak dönüştürülebildiği yapıdır. Bu yapıda örneğin, bir denklik sınıfında, bir kayıttaki bir öznitelik değerinin genelleştirme aralığı diğer değerlerinki ile aynıdır. Yani 23 yaşı için bir üst seviye genelleştirme aralığı, bir kayıtta 20-25 aralığındayken başka bir kayıtta 22-27 aralığında olamaz. Bir özniteliğin aynı seviyelerdeki genelleştirmesi tüm tabloda sabittir.

Aşağıdaki şekilde yaş ve cinsiyet öznitelikleri için örnek bir genelleştirme hiyerarşisi örneklenmiştir.

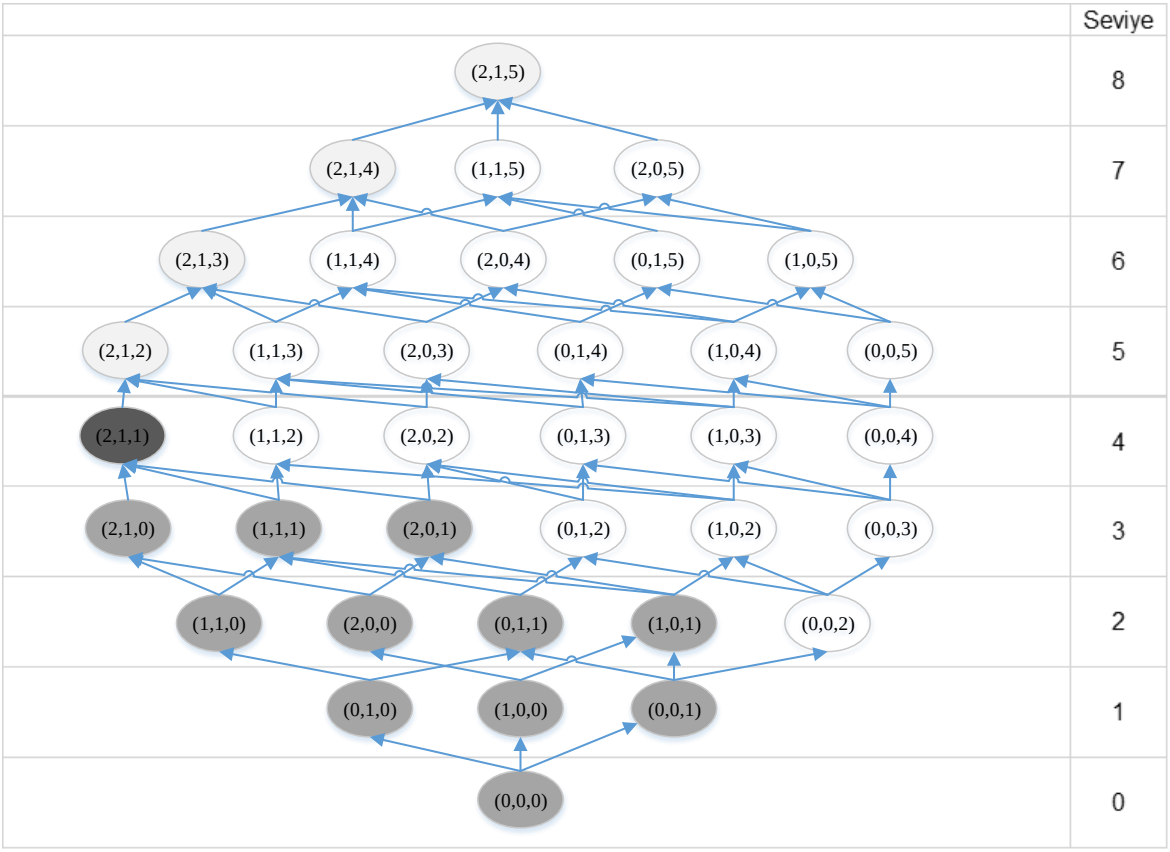


Şekil 5.1. Yaş, Cinsiyet ve Posta Kodu Öznitelikleri için Genelleştirme Örneği

Şekil 5.1'deki yukarı oklar özel durumdan daha genel bir duruma geçişin işaretidir. Şekilde görüldüğü gibi 0 seviyesi genelleştirmenin en alt yani verinin en özel halidir. Yaş için 2, cinsiyet için 1 değeri ise genelleştirmenin en üst halidir yani veri baskılanmıştır

Bu teknik Genelleştirme Örüntüsü (Generalization Lattice) adı verilen veri yapısı ile kademeli bir genelleştirme hiyerarşisi oluşturabilme olanağı sağlar. Genelleştirme için önerilen bu teknik, kullanılan Flash algoritmasında [77] (Lattice and Predictive

Tagging) ile yürütülmüştür. Şekil 5.2'deki genelleştirmeye göre oluşturulmuş örnek bir örüntüye ve tahmine dayalı bir etiketleme örüntüsü aşağıdaki şekilde görülmektedir.



Şekil 5.2. Yaş, Cinsiyet ve Posta Kodu için Genelleştirme Örüntüsü

Yukarıdaki şekilde (0,0,0) ile belirtilen düğüm, özniteliklerin en spesifik halidir. Oklarla yukarıya doğru çıkıldıkça genelleştirme yapılır ve en üstte (2,1,5) ile belirtilen düğüm en genel düğümdür ve bu düğümde tüm öznitelik değerleri baskılanmıştır. Şekilde koyu gri ile belirtilen yerler gerekli mahremiyet ölçütlerini karşılayamayan düğümlerdir. Siyah olan düğüm ise gerekli mahremiyet ölçütlerini sağlayan ve en az veri kaybını sağlayacak düğümdür. Açık gri olan düğümler ise beyaz düğümlere göre daha az veri kaybı sağlayacak düğümlerdir.

5.2. Veri Toplama ve Birleştirmede Kullanılan Yöntemler

Sağlık hizmetlerinde kurumlar dağıtık haldedir. Dağıtık halde depolanan veriler üzerinden ortak hesaplamalar yapabilmek için birçok mahremiyet korumalı veri madenciliği veya istatistiksel yaklaşım önerilmiştir [63]. Fakat doğrudan veri paylaşımı yapılan çalışma sayısı sınırlıdır. [78]' da yapılan çalışmada, web

servisleri kullanılarak bağlı bulunduğu birçok kaynakta bulunan veri tabanlarından çekilen verilerin tek bir uygulamada birleştirilmesi ile istenen analizlerin yapılması sağlanmıştır. Bu yapı, “data mashup” olarak Google gibi devasa sunucuları olan yapılarda kullanılmaktadır. “Data mahsup” yapısını kullanabilecek bir merkezi sistem üzerinde çalışan bir uygulamaya dağıtık kurumların bağlanıp verilerini doğrudan o yapı üzerinden işletmeleri en uygun yöntemdir. Bu çalışmada ek bir uygulama veya servis kullanmadan dağıtık kurumların merkezi sisteme kendi veri tabanlarına erişim izni verdiği varsayılmıştır. Çalışmada, yatay bölümlenmiş veri kümeleri üzerinden model gerçekleştirimi yapılacağı için basit veri tabanı işlemlerinden birleştirme işlemi kullanılarak veriler bir araya getirilebilir. Önerilen modelin esnekliğinden dolayı iletişimde farklı güvenli iletişim protokolleri veya şifreleme teknikleri kullanılabilir. Merkezi bir üçüncü taraf yapısı maliyeti arttırmaktadır fakat daha düzenli bir sistem sunmaktadır.

5.3. Veri Bölümleme Ve Dağıtım

Birçok durumda, dağıtık mahremiyet korumalı veri madenciliği yaklaşımları, bölünmüş/parçalı veri kümelerinden toplu ortak sonuçlar elde etmeye çalışır. Bölünme farklı veri kümelerinde, yatay veya dikey gerçekleşmiş olabilir. Yatay bölünmede; aynı öznitelikler, farklı kayıt sahiplerinden alınıp veri kümelerinde depolanırken, dikey bölünmede; aynı kayıt sahiplerinin farklı öznitelikleri veri kümelerinde depolanmıştır. Dağıtık yapılar için geliştirilen yöntemlerin çoğunda temel amaç; farklı katılımcılar arasından toplanan tüm veri kümeleri üzerinden yararlı istatistiksel hesaplamalar yapılabilmesine olanak sağlanmasıdır.

Dağıtık veri madenciliği problemi genellikle kriptografik alanlarda çalışılmış ve Güvenli Çok-Partili Hesaplama (Secure Multi-party Computation-SMC) [8] gibi ortak fonksiyon hesaplamaları üzerinde durulmuştur. Yapılan çalışmada; veri bölümleme metotları, veri alıcıların isteklerinin sorgulamalarla merkezi veri tabanından çıkarılmasında kullanılmıştır.

Ülke çapında fayda sağlamak için, özellikle sağlık hizmetlerinde, farklı organizasyonlardan veri toplamak ve toplanan verileri, potansiyel tehlikeleri önceden görebilmek için değerlendirebilmek hayati bir öneme sahiptir. Dolayısıyla gösterilen modeller sadece veri göndericileri değil farklı bilimsel çalışma yapan kurumları da kapsamalıdır. Toplanan verilerin, farklı kurumlara ihtiyaçları

doğrultusunda, maksimim veri faydası ile dağıtılabilmesi için bölümlenmesi gerekmektedir. Yapılan çalışmada ortak hesaplama yöntemleri detaylandırılmayacaktır. Çünkü bu durum alıcı tarafların farklı gereksinimleri olduğu durumlarda değişmektedir. Yapılan çalışmada, veri toplama aşamasında olduğu gibi veri dağıtma aşamasında da veri tabanı işlemleri kullanılmıştır. Alıcı kurumlar için bölümlenen tablo veya veri kümelerine alıcı kurumların verilen erişim izinleri ile ulaşabildiği varsayılmıştır. Merkezi yapılar için kullanılabilecek uygulama web servisleridir. Böylece istenen kurumlarla istendiği şekilde veri alış-verişinde bulunulabilir.

5.4. Veri Kümesi ve Kurulum

2004 yılında kurulan, Kaliforniya Üniversitesi Irvine Makine Öğrenmesi Veri Ambarı (Machine Learning Repository) tarafından Amerika Birleşik Devletleri gelir sayım (census income) veri tabanından bir kısım, ana veri kümesi olarak bu tez çalışmasında kullanılmıştır [79]. Yetişkin veri kümesi (Adult database) olarak adlandırılan bu veri kümesinin orijinal içeriğinde 48842 kayıt örneği, 14 öznitelik bulunmaktadır. Fakat kullanılan veri kümesi ile amaç, önerilen modelin örneklemini gerçekleştirmek olduğundan veri kümesi üzerinde bazı düzenlemeler yaparak içeriğinde bulunan kayıp değerlerin ve bağlı bulunduğu kayıtların etkisi çıkarılmıştır. İşlem yükünü hafifletmek ve işleyişi daha net bir biçimde görüntüleyebilmek adına benzer özniteliklerden bir kısmı çıkarılmıştır. Düzenlemeler sonucunda 29967 kayıt ve 7 öznitelik kullanıma uygun hale gelmiştir. Kullanılan öznitelikler ve o özniteliklere bağlı alan boyutu Tablo 5.1‘de gösterilmiştir.

Tablo 5.1. Kullanılan veri kümesinin özellikleri

Öznitelik	**Alan Boyutu**
Yaş	74 [17-91]
Cinsiyet	2 [bay-bayan]
Eğitim Seviyesi	16
Çalışma Sınıfı	7
Mesleği	14
Haftalık Çalışma Saati	100 [0-100]
Gelir Durumu	2 [ort. üstü- ort. altı]

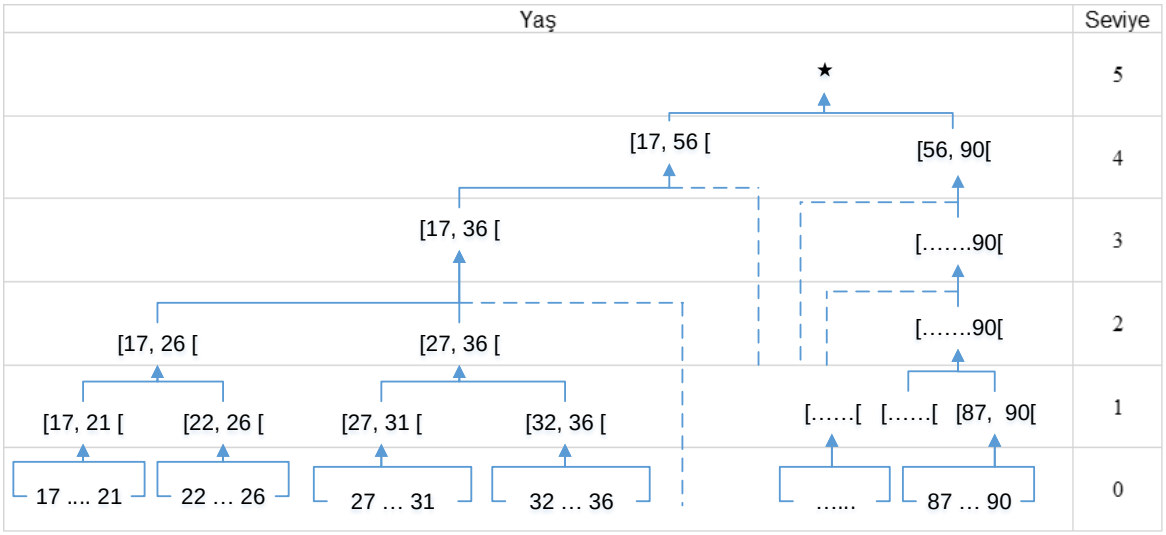
Olası modeller ve önerilen modelin gerçekleştirimi için Java tabanlı, açık kaynak kodlu veri anonimleştirme sisteminin bir parçası olan ARX 2.2 veri anonimleştirme aracı (ARX Data Anonymization Tool) [77] kullanılmıştır.

Yetişkin veri kümesi, ilk olarak yatay bölümleme teknikleri kullanılarak bölümlenmiş ve her bölüm dağıtık yapıdan bir kurumun veri kümesi kabul edilerek işleme alınmıştır. Aynı özniteliklerin değerlerinin farklı kayıt sahiplerine göre oluşturulmuş bu veri kümelerinde daha sonra yapılan birleştirme işlemleri veri tabanı birleştirme işlemleridir.

Merkezi sistemden yapılan veri paylaşımında ise alıcı kurumların sorguları doğrultusunda elde edilen tablolar, alıcı kurumların erişimine açık hale getirilmiştir. Bu yapının güvenli bir şekilde oluşturulması için gerekli olan izinler ve kullanılması gereken protokoller bu çalışmada detaylandırılmamıştır. İletişimde gerekli tüm güvenlik önlemlerinin sağlandığı varsayılmıştır.

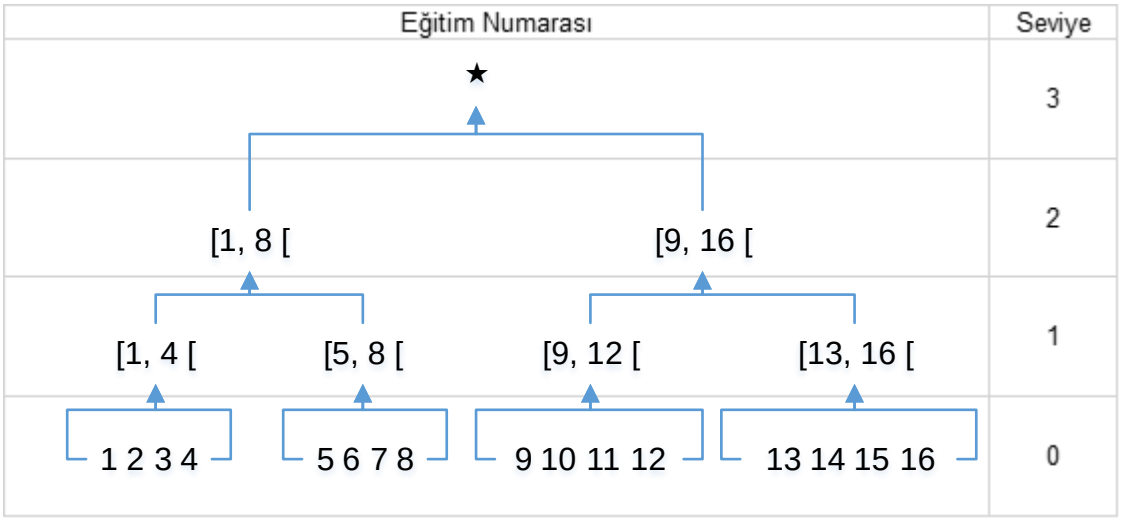
6. DENEY VE BULGULAR

Bu tez çalışmasında, yaygın olarak kullanılan k-anonimlik ve ℓ-çeşitlilik anonimleştirme ölçütleri kullanılarak, dağıtık kurumların güvenli bir şekilde veri paylaşımı yapabilecekleri bir sistem modeli önerilmektedir. Önerilen bu modelin gerçekleştirimi için öncelikle kullanılan veri kümesi bölümlenip 3 farklı dağıtık kurum veri kümesi elde edilmiştir. Her kurumda ortalama 9990 kayıt bulunmaktadır. Sonuçlarda dengesizlik olmaması için örneklenen her modelde tüm veri kümelerine aynı genelleştirme hiyerarşisi uygulanmıştır. Genelleştirme hiyerarşisi kullanıcı tarafından kolay anlaşılabilmesi açısından basit bir sıralama düzenine göre oluşturulmuştur. Yaş, cinsiyet, eğitim seviyesi ve haftalık çalışma saati öznitelikleri için tüm veri kümelerindeki özniteliklere uygulanan genelleştirme hiyerarşileri aşağıda gösterilmektedir.



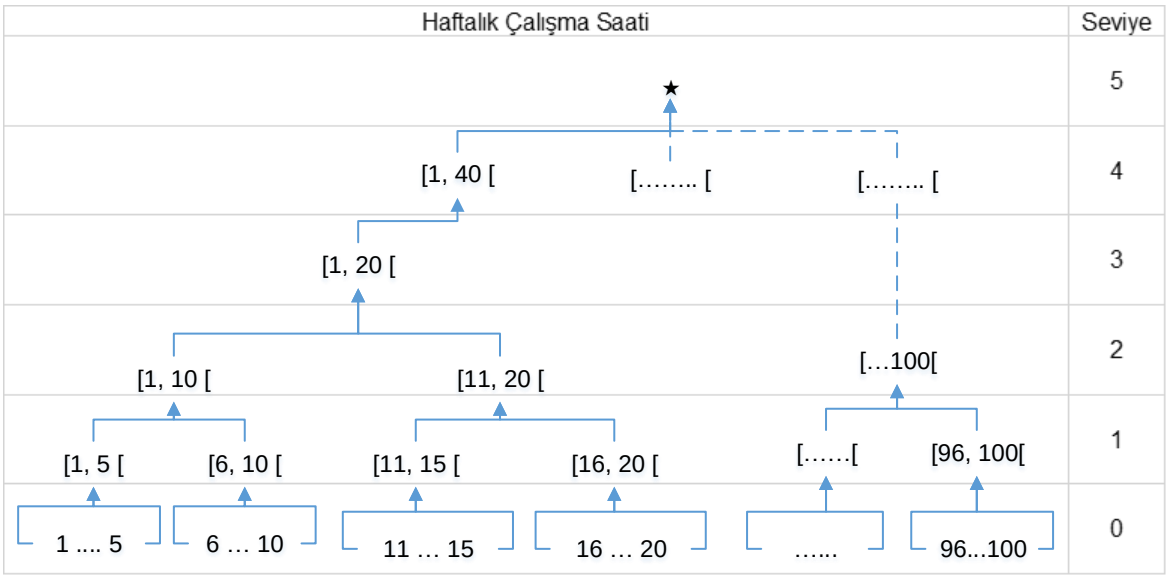
Şekil 6.1. Yaş Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

Kullanılan veri kümesi yetişkin bireyler ile ilgili bilgiler içerdiğinden yaş özniteliğinin değer aralığı 17 ile 90 arasındadır. Bu özniteliğe 6 seviyeli genelleştirme uygulanmıştır. İlk seviyede verilerin kendi değerleri bulunurken 2. seviyede ardışık 5 değerin olduğu bir gruplama söz konusudur. 3. seviyede ardışık 10 değer gruplanırken 4. seviyede grup hacmi 20'ye yükseltilmiştir. 5. seviyede grup hacmi 40 değere çıkartılmış ve son seviyede baskılama uygulanmıştır.



Şekil 6.2. Eğitim Seviyesi Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

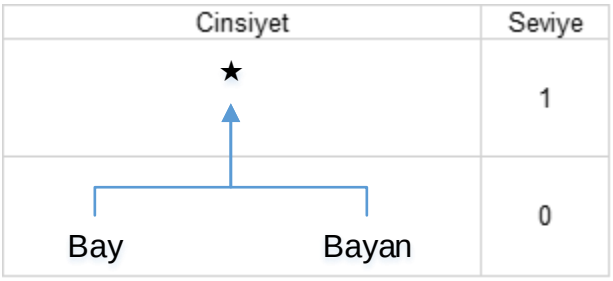
Kullanılan veri kümesinde eğitim seviyesi 16 kademeli olarak belirtilmiştir. Ülkemizdeki 4+4+4 yapısından esinlenilerek ikinci seviyedeki genelleştirmede gruplar ardışık 4 kademeden oluşmaktadır. Üçüncü seviyede grup hacimleri 8 değere çıkarılmış ve bir üst seviye baskılama seviyesi olarak belirlenmiştir.



Şekil 6.3. Haftalık Çalışma Saati Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

Kullanılan veri kümesinde kayıt sahiplerinin haftalık çalışma süreleri 0 ile 100 aralığındadır. Bu özniteliğe 6 seviyeli genelleştirme uygulanmıştır. İlk seviyede verilerin kendi değerleri bulunurken ikinci seviyede ardışık 5 değerin olduğu bir

gruplama söz konusudur. Üçüncü seviyede ardışık 10 değer gruplanırken dördüncü seviyede grup hacmi 20'ye yükseltilmiştir. Beşinci seviyede grup hacmi 40 değere çıkartılmış ve son seviyede baskılama uygulanmıştır.



Şekil 6.4. Cinsiyet Özniteliğine Uygulanan Genelleştirme Hiyerarşisi

Cinsiyet özniteliği sadece iki farklı değerden oluştuğu için sadece 1 seviye genelleştirilebilmiştir o da baskılanmıştır.

Benzer şekilde çalışma sınıfı ve meslek özniteliklerine de gereken genelleştirme hiyerarşisi uygulanmıştır. Anonimleştirme işlemlerinde kullanılacak genelleştirme seviyeleri, belirlenen mahremiyet ölçütü ve girilen hiyerarşi dikkate alınarak oluşturulmaktadır.

Veri kümeleri, dağıtık bölümler ve genelleştirme hiyerarşilerinin belirlenmesiyle birlikte yapılan düzenlemeler sonucunda olası iki model ile önerilen model (3. model), 2 ile 100 arasında verilen k değerleri ile anonimleştirilmiştir. Her anonimleştirme işleminde girilen k değerine ek olarak ℓ-çeşitlilik ölçütü de hassas özniteliğin maksimum alan boyutu 2 olmasından dolayı ℓ=2 olarak kullanılmıştır. Anonimleştirme işlemlerinde veri kümesinin maksimum ne kadar dışlanabileceğini belirtmek gerekir. Kullanılan veri kümesinde bazen tek bir kayıtın anonimleştirilmesi fazla veri kaybına neden olacağından o kayıt dışlanabilir böylece veri kaybı miktarı daha az olabilir. Bu durum dışlama miktarı (outlyers ratio) olarak belirtilmektedir. Bu çalışmada, modeller arası eşitliği sağlamak için tüm modellerin gerçekleştiriminde maksimum dışlama oranı 0,05 olarak ayarlanmıştır.

Anonimleştirme işlemleri sonucunda alıcı kurumların sahip olacağı tablodaki en ideal (optimum) veri dönüşümleri aşağıdaki tabloda gösterilmiştir.

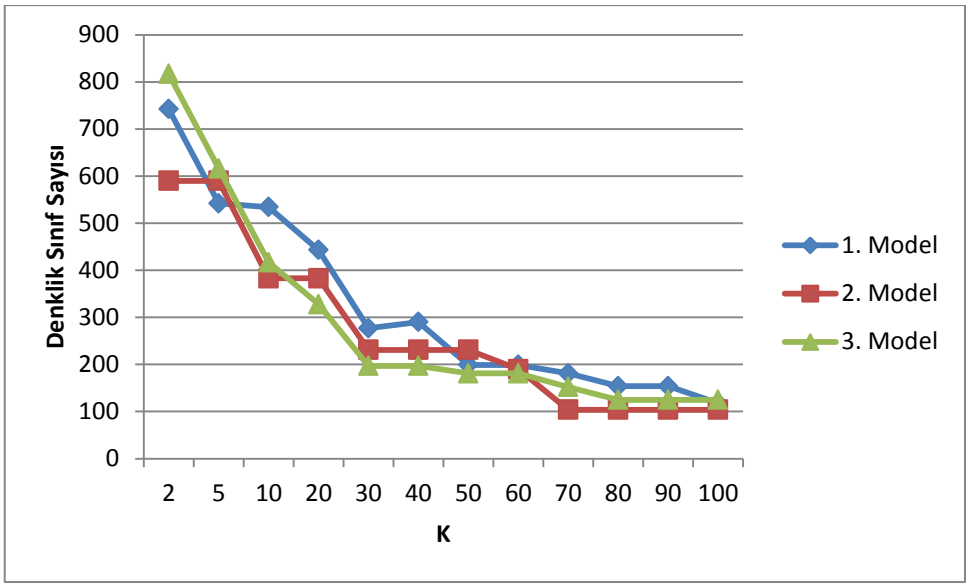
Sadece veri kümesindeki çalışma sınıfı, eğitim seviyesi, mesleği, cinsiyeti ve gelir durumu öznitelikleri ile ilgilenen bir kuruma, anonimleştirme işlemleri sonucu 3 modelin de çıkardığı en ideal anonim tablolardaki dönüşümler Tablo 6.1 ‘de verilmiştir.

Tablo 6.1. Karşılaştırılan modellerin farklı k değerlerindeki veri dönüşümleri

K Değerleri	**1. Model**	**2. Model**	**3. Model**
K=2	3,3,0,0,0	2,3,0,0,0	0,0,0,1,0
K=5	3,3,0,0,0	2,3,0,0,0	0,0,0,1,0
K=10	2,3,0,1,0	3,3,0,0,0	3,0,0,0,0
K=20	3,3,0,1,0	3,3,0,0,0	2,0,0,1,0
K=30	3,3,0,1,0	3,3,0,1,0	3,0,0,1,0
K=40	3,3,0,1,0	3,3,0,1,0	3,0,0,1,0
K=50	3,3,3,0,0	3,3,0,1,0	2,1,0,0,0
K=60	3,3,3,0,0	3,3,0,0,0	1,2,0,0,0
K=70	3,3,1,0,0	3,3,0,1,0	1,2,0,0,0
K=80	3,3,0,1,0	3,3,0,1,0	3,1,0,0,0
K=90	3,3,0,1,0	3,3,0,1,0	3,1,0,0,0
K=100	3,3,0,1,0	3,3,0,1,0	2,2,0,0,0

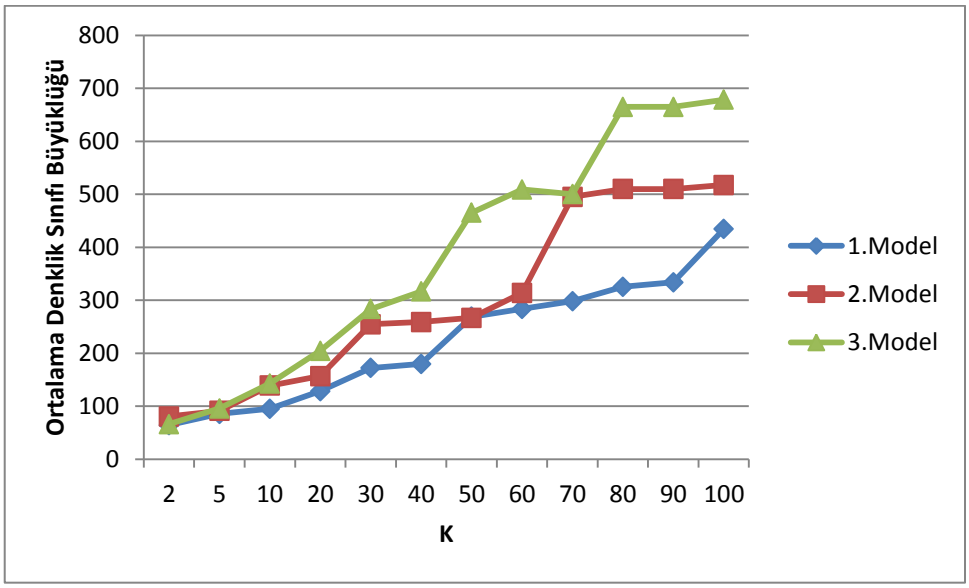
Genelleştirmeler sırasıyla; çalışma sınıfı, eğitim seviyesi, mesleği, cinsiyeti ve gelir durumu özniteliklerdir.

Sağlıklı karşılaştırma yapabilmek adına tüm modellerde aynı anonimleştirme ölçütleri aynı oranlarda kullanılmıştır. Anonimleştirme işlemleri sonucunda dikkat edilmesi gereken önemli noktalardan biri denklik sınıfları sayısıdır. Gerçekleştirilen modellerin k-anonimlik ölçütüne göre oluşturduğu denklik sınıflar Şekil 6.5’te gösterilmektedir.



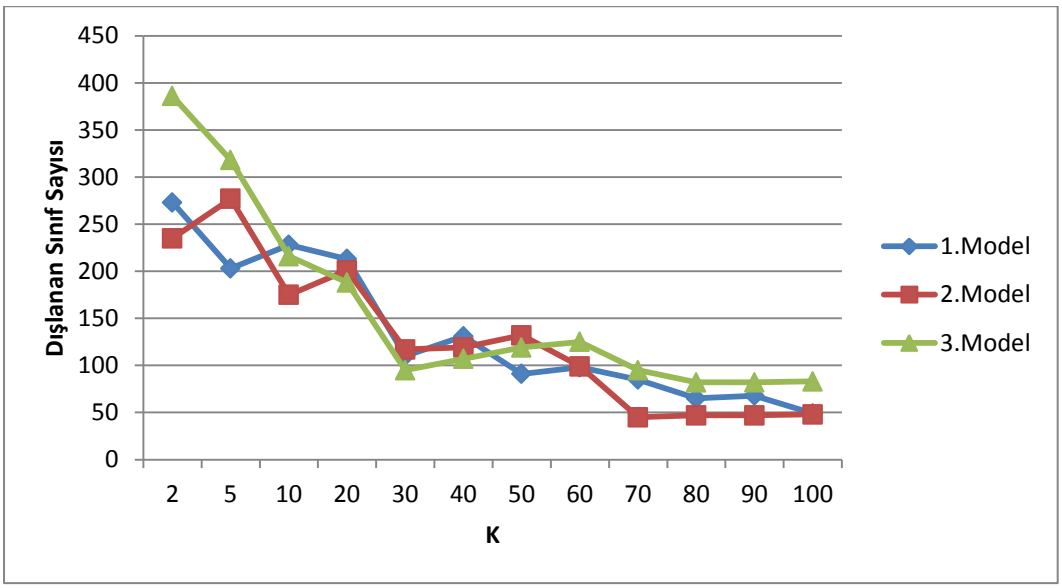
Şekil 6.5. Modellerin Ürettikleri Denklik Sınıfı Sayıları

Anonim tablolar denklik sınıflardan oluşur fakat denklik sınıflarının boyutu da yani her sınıfta kaç kayıt bulunduğu da veri kaybı tespiti için göz önünde bulundurulan önemli noktalardan biridir. Gerçekleştirilen modellerin k-anonimlik ölçütüne göre oluşturduğu denklik sınıfların boyutları Şekil 6.6'da gösterilmektedir.



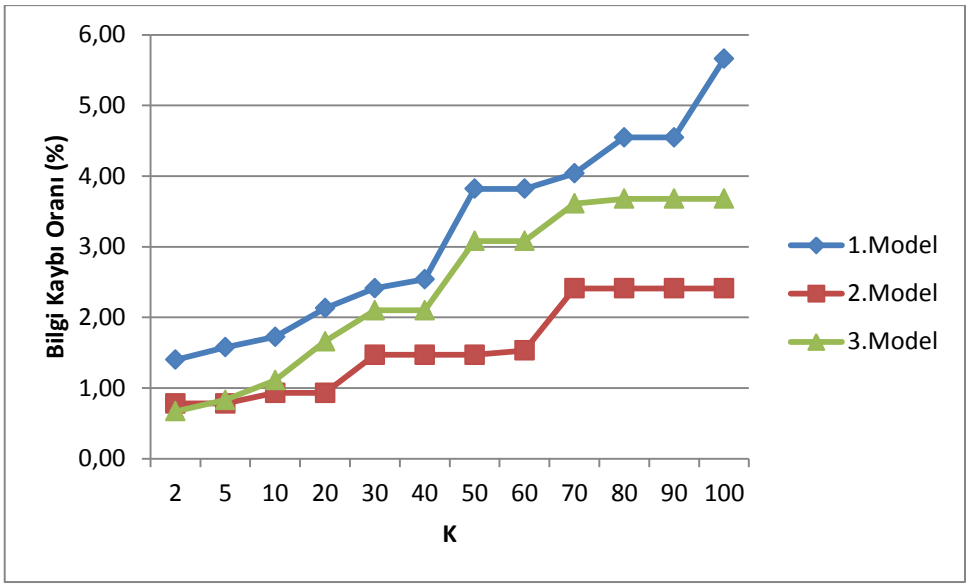
Şekil 6.6. Modellerin Ürettikleri Denklik Sınıflarının Büyüklüğü

0,05 olarak her modele verilen dışlama miktarına göre modellerde dışlanan sınıf miktarları Şekil 6.7'de gösterilmektedir. Verilen miktara göre modeller tutarlı bir dışlama oranı vermişlerdir. Dolayısı ile oranlar arası anlamlı bir ayrım söz konusu değildir.



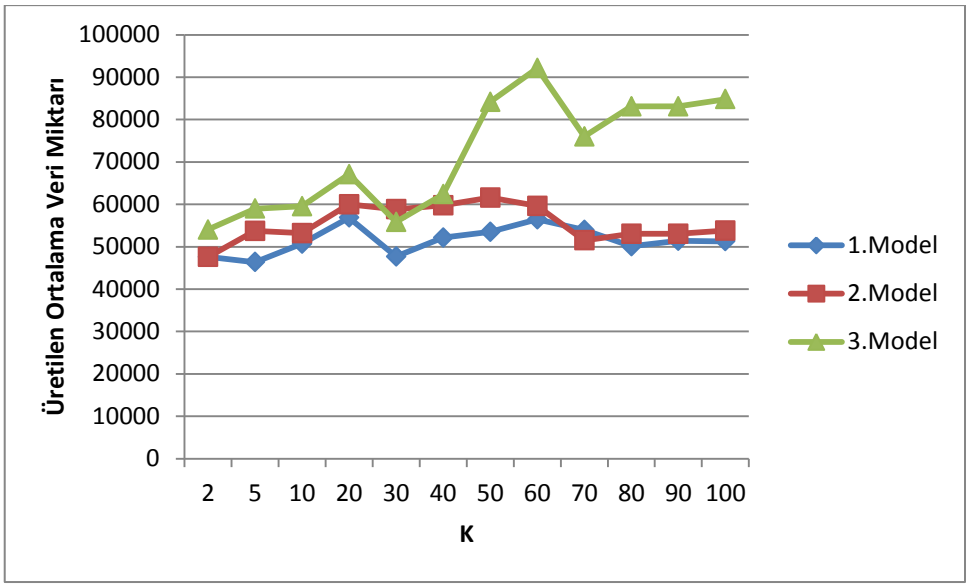
Şekil 6.7. Modellerin Dışladıkları Sınıf Sayıları

Anonimleştirme işlemleri sonucu elde edilen veri kümesindeki veri kaybının ölçümü için "Monotonik Uniform Discernability Metric" kullanılmıştır. Bu metrik hem veri kaybı değerlerini yüzdelik dilim üzerinden hesaplamakta hem de denklik sınıflarının boyutlarını az, sayılarını fazla tutmayı amaçlayarak veri kaybını ölçmektedir. Ayırt Edilebilirlik Metriğine göre, modellerin anonimleştirme işlemleri sonucunda elde edilmiş veri kaybı değerleri Şekil 6.8'deki grafikte gösterilmiştir.



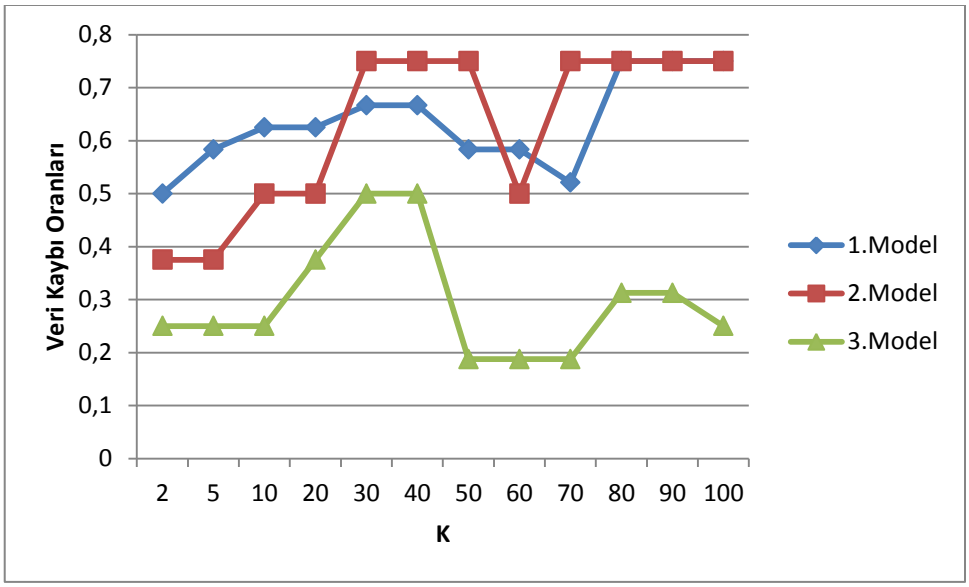
Şekil 6.8. Ayırt Edilebilirlik Metriği ile Modellerin Bilgi Kaybı

Şekil 6.8'de verilen değerler beklendiği gibi sonuç vermemektedir, çünkü birinci ve ikinci modellerde anonimleştirilen veri miktarı üçüncü modelde anonimleştirilen veri miktarının yaklaşık 3/2 katıdır. Girilen sorgular doğrultusunda ilk iki modelde altı öznitelik anonimleştirilmiş iken son modelde dört öznitelik anonimleştirilmiştir. Örnek veri kümesine iki özniteliğin eklenmesi tüm kümenin veri faydasını arttırırken mevcut dört özniteliğin veri kaybını arttırmıştır. Çünkü anonimleştirme işlemi, eklenen iki özniteliğin de göz önüne alınmasıyla yapılmaktadır. Bu durumu Şekil 6.10'daki grafikte daha net olarak gösterilmektedir. Ayrıca bu durum daha açık olarak aşağıdaki grafikten anlaşılabilmektedir. Şekil 6.9'da her modelde alıcılara sunulan ortalama veri miktarı (denklik sınıfı X ortalama sınıf boyutu) gösterilmektedir.



Şekil 6.9. Modellerin Sundukları Ortalama Veri Miktarı

Önerilen modelde özniteliklerin kurumların ihtiyaçları doğrultusunda belirlenmesinden dolayı her özniteliğin eşit şekilde ölçüldüğü veri kaybı metriklerinin kullanımı uygun değildir. Sadece Tablo 6.1'deki genelleştirme oranlarından yola çıkarak En az Bozukluk (Minimal Distortion) metriğinin, sadece istenen özniteliklerdeki bozulmanın hesaplanması sonucu modellerden elde edilen bilgi kaybı aşağıdaki gibidir.



Şekil 6.10. En Az Bozukluk Metriği ile Modellerin Bilgi Kaybı

7. TARTIŞMA

Dağıtık sistemlerden toplanan veriler ile mahremiyet korumalı veri dağıtımı yapabilen bir sistem modelinin önerildiği bu çalışmada, özellikle mahremiyet korumalı veri yayıncılığı yaklaşımlarından yararlanılmıştır. Mahremiyet korumalı veri yayıncılığı ve mahremiyet korumalı veri madenciliği yaklaşımlarının detaylı olarak araştırıldığı bu çalışmada birçok yaklaşımdan esinlenilerek çalışmalar yürütülmüştür.

Yapılan çalışma öncelikle literatürde bulunan mahremiyet koruma çalışmalarından farklı olarak veri madenciliği sonuçları veya veri yayıncılığı üzerinde durulmamış, alıcı kurumlar için anonim verilerdeki veri faydasını arttıracak bir sistem üzerine odaklanılmıştır. Literatürde var olan veri tabanı bölümleme yöntemleri, anonimleştirme ölçütleri ve veri dönüşüm teknikleri kullanılarak gerçekleştirilen model güvenilir üçüncü taraf yapısının kullanımıyla anonim verilerdeki veri faydasının artmasına yol açmıştır.

Benzer amaç doğrultusunda, HIDE projesi tarafından önerilen sistem modeliyle kıyaslandığında önerdiğimiz model verilerin güvenliği bakımından daha maliyetli olmasına rağmen veri kaybını azaltarak alıcı kurumların çalışmalarına daha fazla katkı sağlayacaktır.

Bu tez çalışmasındaki model temel olarak resmi sağlık otoriteleri tarafından kullanılabilecek bir model niteliği taşımaktadır. Ülkemiz sağlık otoritelerinden örneğin Sağlık Bakanlığı, merkezi ve güvenilir üçüncü taraf yapısı olarak kabul edilebilir. Bu durumda bu otorite tarafından toplanan verilerin, bilimsel çalışmalar yapılması, araştırma ve geliştirme imkânları sunulması adına farklı araştırma toplulukları ile paylaşılması için güvenilir bir model kullanılması gerekmektedir. Önerilen modelin güvenilirliği ve esnekliği sayesinde kayıt sahipleri açısından gereken mahremiyet ölçütleri uygulanabilir ve daha verimli bilimsel çalışmaların yürütülmesi sağlanabilir.

Son olarak önerilen model üç aşamalı bir yapı gerektirmektedir. Bu aşamaların her biri ayrı bir çalışma konusu olup farklı problem çözümlerinde kullanılmaktadır. Örneğin veri toplama aşamasında literatürde birçok protokol geliştirilmiş veya farklı sistemler kullanılmıştır fakat yapılan çalışmaların çoğunda verilerin homojen olarak dağıtıldığı varsayılmıştır. Gerçekleştirilen bu çalışma da temelde bazı varsayımlar

üzerine yürütülmüştür fakat bu varsayımların bir kısmı mevcut durumda pek mümkün olmamaktadır. Örneğin tüm dağınık kurumların yatay bölümlenmiş homojen veri tabanları kullandığı varsayılmıştır fakat mevcut durumda böyle bir yapı söz konusu değildir. Merkezi bir yönetim için olması gereken ideal yapı gösterilmiştir. Çalışmanın amacı veri sahiplerinin mahremiyetini istenen ölçütlerde korurken aynı zamanda veri faydasını arttırmak olduğundan, diğer aşamalarda ele alınan varsayımların çözümü üzerine bir çalışma yapılmamıştır. Diğer aşamaların farklı uygulamalar kullanılarak gerçekleştirilmesi sonucu etkilememektedir. Çünkü veri faydasının arttırılması ikinci aşamada gerçekleştirilmektedir.

8. SONUÇ VE ÖNERİLER

Bu tez çalışmasında birçok mahremiyet korumalı yaklaşım özetlenmiş ve sağlık hizmetleri çerçevesinde değerlendirilmiştir. Mahremiyet korumalı yaklaşımların doğru bir sistem üzerinde kullanılması ile sağlık hizmetlerinde veri paylaşımı hususunda meydana gelebilecek problemlerin çözülmesi hedeflenmiştir. Bu amaçla, sağlık hizmetlerinde dağıtık halde bulunan verilerden, merkezi bir yapı aracılığıyla veri sahiplerinin mahremiyetini koruyacak ve aynı zamanda maksimum düzeyde fayda sağlayacak paylaşımlar yapabilecek ideal bir sistem modeli önerilmektedir.

Yapılan çalışmada, diğer çalışmalardan farklı olarak veri madenciliği veya güvenli ortak hesaplamalar üzerine teknik geliştirmekten ziyade bu tekniklerin kullanıldığı modelin nasıl olması gerektiği üzerine odaklanılmıştır. Özellikle sağlık verilerinin dağıtık veri tabanlarından, mahremiyet korumalı ve veri kaybı en aza indirilmiş şekilde araştırma topluluklarına ulaştırılmasını sağlayacak bir model üzerinde durulmuştur. Çalışmada önerilen modele ek olarak iki olası modelin daha gerçekleştirimi yapılmış ve bu modellerin ne tür yapılar için kullanımının uygun olduğu, avantaj ve dezavantajları ile belirtilmiştir.

Önerilen sistem modeli güvenilir merkezi bir yapının varlığını temel almakta ve istenen mahremiyet ölçütleri doğrultusunda veri faydasını da arttırmaktadır. Ayrıca, özellikle çok boyutlu veri kümelerinde gerçekleştirilen anonimleştirme işlemlerinin zorluğu, önerilen modelde kullanılan veri bölümleme yöntemleri ile aza indirilmiştir.

Yapılan çalışmalar sonucunda veri kaybı ölçümü için kullanılan mevcut metriklerin gerekli ihtiyaçları karşılamada yetersiz kaldığı gözlemlenmiştir. Mevcut metrikler tüm özniteliklerin değer kaybını eş bir biçimde hesaplamaktadır. Fakat yapılan çalışmada görülmüştür ki her öznitelik aynı oranda tanımlayıcı olmadığından, aynı oranda değerlendirmeye alınmamaları gerekmektedir. İleriki çalışmalarda özniteliklerin tanımlayıcılık oranları göz önüne alınarak bilgi kaybını belirleyen ve veri dönüşümlerini bu doğrultuda gerçekleştiren metrikler üzerine çalışılması planlanmaktadır. Bahsedilen metriğin oluşturulması ile birlikte önerilen sistemin daha sağlıklı sonuçlar vermesi beklenmektedir.

Önerilen modelin, çevirim içi sorgulama yapılabilecek bir model haline dönüştürülmesi de ileriki çalışmalarda planlanmaktadır. Planlanan bu model alıcı kurumlar için akan veriler üzerinden aktif olarak hizmet sağlayacaktır. Bu sayede periyodik olarak değil güncel olarak alıcı kurumlara veri akışı sağlanabilecektir.

KAYNAKLAR

[1] *B. Consortium Goals*. İnternet Sayfası: http://icgc.org/icgc/goals-structure-policies-guidelines/b-consortium-goals, Erişim Tarihi: 04.01.2014

[2] "Health insurance portability and accountability act of 1996 (HIPAA)", Public Law Gazette, 1–349, **1996**.

[3] "Veri Güvenliği Hakkında Genelge 2005/153", **2005**.

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

[5] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments", *ACM Computing Surveys (CSUR),* 42, **2010**.

[6] J. Domingo-Ferrer, "A survey of inference control methods for privacy-preserving data mining", *Privacy-preserving data mining*, Springer, 53-80, **2008**.

[7] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 639-644, **2002**.

[8] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining", *ACM SIGKDD Explorations Newsletter,* 4,12-19, **2002**.

[9] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 247-255, **2001**.

[10] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information", *PODS*, 188, **1998**.

[11] P. Samarati, "Protecting Respondents' Identities in Microdata Release" *IEEE Transactions on Knowledge and Data Engineering,* 13, 1010-1027, **2001**.

[12] P. Samarati and L. Sweeney, "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression", Computer Science Laboratory, SRI International, **1998**.

[13] L. Sweeney, "k-anonymity: a model for protecting privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 10, 557-570, **2002**.

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

[15] L. Sweeney, "Weaving technology and policy together to maintain confidentiality", *J Law Med Ethics,* 25, 98-110, 82, Summer-Fall **1997**.

[16] T. Dalenius, "Towards a methodology for statistical disclosure control", *Statistik Tidskrift,* 15, 2-1, **1977**.

[17] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity", *ICDE*, 106-115, **2007**.

[18] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases", *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 665-676, **2007**.

[19] C. Dwork, "Differential privacy", *Automata, languages and programming*, Springer, 1-12, **2006**.

[20] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy", *Journal of the ACM (JACM),* 60,12, **2013**.

[21] B. C. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation", *Proceedings of the 21st International Conference on Data Engineering,* 205-216, **2005**.

[22] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity", *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 49-60, **2005**.

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

[24] X. Xiao and Y. Tao, "Personalized privacy preservation", *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, 229-240, **2006**.

[25] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques", *Third IEEE International Conference on Data Mining*, 99-106, **2003**.

[26] Z. R. Mnatsakanyan, D. J. Mollura, J. R. Ticehurst, M. R. Hashemian, and L. M. Hung, "Electronic medical record (EMR) utilization for public health surveillance", *AMIA Annual Symposium Proceedings*, 480, **2008**.

[27] J. S. Lombardo and L. J. Moniz, "TA Method for Generation and Distribution", *Johns Hopkins Apl Technical Digest,* 27, 356, **2008**.

[28] T. Burr, R. Klamann, S. Michalak, and R. Picard, "Generation of Synthetic Biosense Data", Los Alamos National Laboratory Report LAUR-05-7841, **2005**.

[29] L. Moniz, A. L. Buczak, L. Hung, S. Babin, M. Dorko, and J. Lombardo, "Construction and validation of synthetic electronic medical records", *Online journal of public health informatics,* 1, **2009**.

[30] *Subsyndromes Presentation,* İnternet Sayfası: http://www.cdc.gov/Biosense/files/PHIN2007, Erişim Tarihi: 01.01.2014

[31] A. L. Buczak, S. Babin, and L. Moniz, "Data-driven approach for creating synthetic electronic medical records," *BMC medical informatics and decision making,* 10,59, **2010**.

[32] *Models of Infectious Disease Agent Study (MIDAS)*. İnternet Sayfası: http://www.epimodels.org/midas/Rabout.do, Erişim Tarihi: 03.02.2014

[33] *ARCHIMEDES Project*, İnternet Sitesi: http://www.archimedesmodel.com, Erişim Tarihi: 24.01.2014

[34] *Project Mimic,* İnternet Sayfası: http://www.projectmimic.com/, Erişim Tarihi: 02.02.2014

[35] M. L. Johnson, L. Pipes, P. P. Veldhuis, L. S. Farhy, D. G. Boyd, and W. S. Evans, "AutoDecon, a deconvolution algorithm for identification and characterization of luteinizing hormone secretory bursts: Description and validation using synthetic data", *Analytical biochemistry,* 381, 8-17, **2008**.

[36] R. E. Watkins, S. Eagleson, B. Veenendaal, G. Wright, and A. J. Plant, "Disease surveillance using a hidden Markov model", *BMC medical informatics and decision making,* 9, 39, **2009**.

[37] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the Datafly System", *Proceedings of the AMIA Annual Fall Symposium*, 51, **1997**.

[38] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity", *Proceedings of the 22nd International Conference on Data Engineering*, 25-25, **2006**.

[39] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization", *Proceedings of 21st International Conference on Data Engineering*, 217-228, **2005**.

[40] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining", **2002**.

[41] M. Kantarcıoglu, J. Vaidya, and C. Clifton, "Privacy preserving naive bayes classifier for horizontally partitioned data", *IEEE ICDM workshop on privacy preserving data mining*, 3-9, **2003**.

[42] H. Yu, X. Jiang, and J. Vaidya, "Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data", *Proceedings of the 2006 ACM symposium on Applied computing*, 603-610, **2006**.

[43] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data", *IEEE Transactions on Knowledge and Data Engineering,* 16, 1026-1037, **2004**.

[44] A. Inan, S. V. Kaya, Y. Saygın, E. Savaş, A. A. Hintoğlu, and A. Levi, "Privacy preserving clustering on horizontally partitioned data," *Data & Knowledge Engineering,* 63, 646-666, **2007**.

[45] G. Jagannathan and R. N. Wright, "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data", *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 593-599, **2005**.

[46] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright, "A New Privacy-Preserving Distributed k-Clustering Algorithm", *SDM*, 494-498, **2006**.

[47] H. Polat and W. Du, "Privacy-preserving top-N recommendation on horizontally partitioned data", *Proceedings of The International Conference on Web Intelligence*, 725-731, **2005**.

[48] W. Du, Y. S. Han, and S. Chen, "Privacy-Preserving Multivariate Statistical Analysis: Linear Regression and Classification", *SDM*, 222-233, **2004**.

[49] W. Du and M. J. Atallah, "Privacy-preserving cooperative statistical analysis", *Proceedings of 17th Annual Computer Security Applications Conference,* 102-110, **2001**.

[50] L. X. P. Jurzcyk, "Privacy-preserving data publishing for horizontally partitioned databases", The CM International Conference on Information and Knowledge Management, **2008**.

[51] P. Jurczyk and L. Xiong, "Distributed anonymization: Achieving privacy for both data subjects and data providers," *Data and Applications Security XXIII*, Springer, 191-207, **2009**.

[52] C. Dwork and K. Nissim, "Privacy-Preserving Datamining on Vertically Partitioned Databases", The 24th Annual International Cryptology Conference, **2004**.

[53] J. Vaidya and C. Clifton, "Privacy Preserving Naïve Bayes Classifier for Vertically Partitioned Data", *SDM*, 522-526, **2004**.

[54] H. Yu, J. Vaidya, and X. Jiang, "Privacy-preserving svm classification on vertically partitioned data", *Advances in Knowledge Discovery and Data Mining*, Springer, 647-656, **2006**.

[55] J. Vaidya and C. Clifton, "Privacy-preserving k-means clustering over vertically partitioned data", *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 206-215, **2003**.

[56] J. Vaidya and C. Clifton, "Privacy-preserving decision trees over vertically partitioned data", *Data and Applications Security XIX*, Springer, 139-152, **2005**.

[57] J. Gardner, L. Xiong, K. Li, and J. J. Lu, "HIDE: heterogeneous information DE-identification", *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 1116-1119, **2009**.

[58] P. Jurczyk and L. Xiong, "Towards privacy-preserving integration of distributed heterogeneous data", *Proceedings of the 2nd PhD workshop on Information and knowledge management*, 65-72, **2008**.

[59] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", **2001**.

[60] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification", *Lingvisticae Investigationes,* 30, 3-26, **2007**.

[61] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining", The 20th Annual International Cryptology Conference on Advances in Cryptology, **2000**.

[62] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining", *ACM SIGKDD Explorations Newsletter,* 4, 28-34, **2002**.

[63] C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms", Springer, **2008**.

[64] Z. Huang, W. Du, and B. Chen, "Deriving private information from randomized data" *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, 37-48, **2005**.

[65] S. Zhong, Z. Yang, and R. N. Wright, "Privacy-enhancing k-anonymization of customer data", *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 139-147, **2005**.

[66] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity", *Data and Applications Security XIX*, Springer, 166-177, **2005**.

[67] K. Wang, B. C. Fung, and G. Dong, "Integrating private databases for data analysis", *Intelligence and Security Informatics*, Springer, 171-182, **2005**.

[68] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity", *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 223-228, **2004**.

[69] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality", *Proceedings of the 31st international conference on Very large data bases*, 901-909, **2005**.

[70] Y. Lindell and B. Pinkas, "Privacy preserving data mining", *Advances in Cryptology,* 36-54, **2000**.

[71] C. Dwork and K. Nissim, "Privacy-preserving datamining on vertically partitioned databases", *Advances in Cryptology,* 528-544, **2004**.

[72] J. G. Li Xiong. *HIDE: Health Information DE-identification*. İnternet Sayfası: http://www.mathcs.emory.edu/hide/index.html , Erişim Tarihi: 01.01.2014

[73] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality", The 31st international conference on Very large data bases, **2005**.

[74] S. Goldwasser, "Multi party computations: past and present," in *Proceedings of the sixteenth annual ACM symposium on Principles of distributed computing*, 1-6, **1997**.

[75] O. Goldreich, "Secure multi-party computation", **1998**.

[76] S. Böttcher and S. Obermeier, "Secure set union and bag union computation for guaranteeing anonymity of distrustful participants", *Journal of Software,* 3, 9-17, **2008**.

[77] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. A. Kuhn, "Flash: efficient, stable and optimal k-Anonymity", *2012 International Conference on Privacy, Security, Risk and Trust, and 2012 International Confernece on Social Computing*, 708-717.

[78] N. Mohammed, B. Fung, K. Wang, and P. C. Hung, "Privacy-preserving data mashup", *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 228-239, **2009**.

[79] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "{UCI} Repository of machine learning databases", **1998**.

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı	:	Pelin CANBAY
Doğum Yeri	:	Diyarbakır
Medeni Hali	:	Evli
E-posta	:	paktas.bm@gmail.com
Adresi	:	Huzur Mah. 1632.Sok 24/12 Ankara/Çankaya

Eğitim

Lise	:	Melik Ahmet Lisesi (Diyarbakır)
Lisans	:	Harran Üniversitesi (Şanlıurfa)
Yüksek Lisans	:	Hacettepe Üniversitesi (Ankara)

Yabancı Dil ve Düzeyi

İngilizce – Upper

İş Deneyimi

Kahramanmaraş Sütçü İmam Üniversitesi	Araştırma Görevlisi / 2012-2012
Hacettepe Üniversitesi	Araştırma Görevlisi / 2012 - Devam

Deneyim Alanları

Bilgisayar Bilimleri

Tezden Üretilmiş Projeler ve Bütçesi

Tezden Üretilmiş Yayınlar

Pelin Aktaş, Hayri Sever, Murat Aydos, “Role-based privacy-preserving health records distribution”, 2nd International Conference on e-Health and Telemedicine, 89-96, **2014**.

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar