



**T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**



DOKTORA TEZİ

**ETKİN GİZLİLİK KORUMASI İÇİN RASGELE SAYI ÜRETİCİNE
DAYALI VERİ ANONİMİZASYON MODELİNİN
GELİŞTİRİLMESİ**

Merve KANMAZ

**DANIŞMAN
Prof. Dr. Ahmet SERTBAŞ**

**II. DANIŞMAN
Doç. Dr. Muhammed Ali AYDIN**

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

İSTANBUL-2021

Bu çalışma, 13.09.2021 tarihinde aşğıdaki jüri tarafından Bilgisayar Mühendisliğı Anabilim Dalı, Bilgisayar Mühendisliğı Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Ahmet SERTBAŞ(Danışman)
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Prof. Dr. Abdül Halim ZAIM
İstanbul Ticaret Üniversitesi
Mühendislik Fakültesi

Dr. Öğr. Üyesi Özgür Can TURNA
İstanbul Üniversitesi-Cerrahpaşa
Mühendislik Fakültesi

Doç. Dr. Şerif BAHTİYAR
İstanbul Teknik Üniversitesi
Bilgisayar ve Bilişim Fakültesi

Dr. Öğr. Üyesi Oğuzhan ÖZTAŞ
İstanbul Üniversitesi- Cerrahpaşa
Mühendislik Fakültesi



20.04.2016 tarihli Resmi Gazete’de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi-Cerrahpaşa’nın abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü’nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ÖNSÖZ

Bu çalışma, İstanbul Üniversitesi-Cerrahpaşa Lisansüstü Enstitüsü Bilgisayar Mühendisliği Ana Bilim Dalı’nda yapılan “Etkin Gizlilik Koruması İçin Rasgele Sayı Üreticine Dayalı Veri Anonimizasyon Modelinin Geliştirilmesi” adlı doktora tez çalışmasını içermektedir. Tez çalışması kapsamında yayınlanan makalelerin bilgileri aşağıdaki gibidir:

Kanmaz, M., Aydın M. A., Sertbaş A., 2021. A New Geometric Data Perturbation Method for Data Anonymization Based on Random Number Generators. Journal of Web Engineering , Vol. 20 6, 1–24, doi: 10.13052/jwe1540-9589.20614 (Science Citation Index Expanded)

Bu tez çalışması süresince gösterdiği her türlü destek ve yardımlarından dolayı çok değerli danışman hocalarım Prof. Dr. Ahmet SERTBAŞ ve Doç. Dr. Muhammed Ali AYDIN’ a en içten dileklerimle teşekkür ederim.

Eğitim hayatım boyunca maddi ve manevi desteklerini hiçbir zaman esirgemeyen ve her zaman yanımda olan sevgili aileme en içten dileklerimle teşekkür ederim.

Doktora eğitimim süresince maddi ve manevi olarak her zaman yanımda olan, desteğini hiçbir zaman esirgemeyen eşim Burak KANMAZ’ a sonsuz teşekkürlerimi sunarım.

Tezimi henüz kendisi henüz iki aylıkken doktora sürecine başladığım kızım Zeynep Naz’a armağan ediyorum

Eylül 2021

Merve KANMAZ

İÇİNDEKİLER

Sayfa No

ÖNSÖZ	iv
İÇİNDEKİLER.....	v
ŞEKİL LİSTESİ	vii
TABLO LİSTESİ.....	ix
SİMGE VE KISALTMA LİSTESİ	x
ÖZET	xiv
SUMMARY	xvi
1 GİRİŞ.....	1
2 GENEL KISIMLAR.....	5
2.1 VERİ MAHREMİYETİ.....	5
2.1.1 Sezgisel Yaklaşımlar	11
2.1.1.1 K-Anonimlik	11
2.1.1.2 L-Çeşitlilik.....	11
2.1.1.3 T-Yakınlık	12
2.1.2 Yeniden Yapılandırma Tabanlı Yöntemler	12
2.1.3 Kriptolojik Yöntemler	12
2.2 GİZLİLİK KORUMAYA YÖNELİK ÇALIŞMALAR	13
2.3 ANONİM VERİNİN KULLANILABİLİRLİĞİ.....	17
2.4 DAĞITIK VERİ ORTAMLARI.....	17
3 MALZEME VE YÖNTEM.....	25
3.1 GEOMETRİK PERTÜRBASYON	25
3.1.1 Çarpımsal Dönüşüm	25
3.1.2 Öteleme Dönüşüm	27
3.1.3 Gürültü Ekleme	28
3.1.4 Çoklu Sütun Gizliliği.....	29
3.2 RASGELE SAYI ÜRETEÇLERİ VE LİNEER KONGRÜANS MODEL	30
3.3 VERİ SETLERİ VE BÖLÜMLEME	31
3.4 PERFORMANS METRİKLERİ	33
3.4.1 Sınıflandırma Doğruluğu.....	33
3.4.2 F-Ölçütü.....	34

3.4.3	Atak Dayanıklılığı	34
3.4.4	Çalışma Zamanı.....	35
3.4.5	Friedman Rank Testi	35
3.5	SINIFLANDIRMA YÖNTEMLERİ.....	35
3.5.1	Naive Bayes Sınıflandırıcılar	35
3.5.2	J48 Sınıflandırıcı	36
3.5.3	Karar Tablosu Sınıflandırıcı (Decision Table)	37
3.5.4	OneR Sınıflandırıcı.....	37
3.6	ATAK TÜRLERİ	37
3.6.1	Saf Tahmin Yöntemi (Naive Interference / Estimation)	37
3.6.2	ICA(Bağımsız Bileşen Analizi) Saldırısı	38
3.6.3	I/O (Girdi/Çıktı) Saldırısı	38
3.7	ÖNERİLEN YÖNTEM: RSUGP	39
3.8	ÖNERİLEN RSUGP MODELİNİN SPARK ÜZERİNDE UYGULANMASI	43
4	BULGULAR.....	51
4.1	SINIFLANDIRMA DOĞRULUĞU SONUÇLARI.....	51
4.2	F-ÖLÇÜTÜ SONUÇLARI.....	58
4.3	ATAK DAYANIKLILIĞI SONUÇLARI.....	61
4.4	FRIEDMAN RANK TESTİ SONUÇLARI	62
4.5	ÇALIŞMA ZAMANI	63
5	TARTIŞMA VE SONUÇ	65
	KAYNAKLAR.....	69
	ÖZGEÇMİŞ	79

ŞEKİL LİSTESİ

	Sayfa No
Şekil 1.1: Yetkisiz erişim sembolize edilmesi.....	2
Şekil 2.1: Niteliklerin örnek gösterimi	6
Şekil 2.2: Gizlilik korumalı veri madenciliği yöntemleri.....	10
Şekil 2.3: Verinin 3V gösterimi.....	18
Şekil 2.4: Hadoop ekosistem gösterimi	19
Şekil 2.5: Ana düğüm ve işçi düğümlerin gösterimi	20
Şekil 2.6: Spark bileşenleri gösterimi.....	21
Şekil 2.7: MapReduce hesaplama süreci	23
Şekil 2.8: Zeppelin not defteri örnek ekran gösterimi	24
Şekil 3.1: Rotasyon matrisi gösterimi.....	26
Şekil 3.2: Çarpımsal dönüşüm örneği.....	26
Şekil 3.3: Öteleme dönüşüm örneği	28
Şekil 3.4: 5 kat çapraz geçerleme örnek gösterimi	33
Şekil 3.5: Konfüzyon matrisi.....	34
Şekil 3.6: Rasgele sayı üreticilerine bağlı geometrik veri pertürbasyon modeli	40
Şekil 3.7: Lineer Kongrüans Üreteçler kullanarak gürültü üretme	41
Şekil 3.8: Lineer Kongrüans Akış Diyagramı	43
Şekil 3.9: Ambari ara yüzü bağlantı ekranı	44
Şekil 3.10: Ambari üzerinde Hadoop bileşenleri ve sistemin genel görüntüsü.....	45
Şekil 3.11: Ambari üzerinde kaynak yöneticisi (Resource Manager) görüntüsü	46
Şekil 3.12: HDFS üzerine dosya yükleme görüntüsü.....	46
Şekil 3.13: HIVE ile veri seti yükleme görüntüsü.....	47
Şekil 3.14: HIVE sorgusuyla örnek veri çekme	47

Şekil 3.15: Zeppelin başlangıç ekranı.....	48
Şekil 3.16: Zeppelin yorumlayıcı seçimi	49
Şekil 3.17: Zeppelin kütüphaneleri koda eklenmesi.....	49
Şekil 3.18: Zeppelin üzerinde pyspark ile rotasyon matrisi oluşturma	50
Şekil 4.1: Önerilen modelin FMR değerlerinin kıyaslanması	63
Şekil 4.2: Çalışma süreleri.....	64



TABLO LİSTESİ

Sayfa No

Tablo 2.1 : Genelleştirme yapılmadan önce hasta verileri	7
Tablo 2.2: Genelleştirme yapıldıktan sonra hasta verileri	8
Tablo 2.3: Gizleme yapılmadan önce hasta verisi	8
Tablo 2.4: Gizleme yapıldıktan sonra hasta verileri	8
Tablo 2.5: Gürültü eklenmemiş orijinal veri seti	9
Tablo 2.6: Gelir sütununa gürültü eklenmiş veri seti	9
Tablo 2.7: PPDm yöntemlerinin kıyaslanması	14
Tablo 3.1: Kullanılan veri setleri ve özellikleri	32
Tablo 4.1: Önerilen algoritmanın veri setinin tamamı eğitim ve test verisi olarak kullanıldığı sınıflandırma doğruluğu sonuçları	52
Tablo 4.2: Önerilen algoritmanın 2-kat çapraz geçerleme sınıflandırma doğruluğu sonuçları	53
Tablo 4.3: Önerilen algoritmanın 5-kat çapraz geçerleme sınıflandırma doğruluğu sonuçları	54
Tablo 4.4: Önerilen algoritmanın 10-kat çapraz geçerleme sınıflandırma doğruluğu sonuçları	55
Tablo 4.5: Önerilen modelin sınıflandırma sonuçlarının tek tablo üzerinde gösterimi	57
Tablo 4.6: Önerilen algoritmanın 2-kat çapraz geçerleme için F-ölçütü sonuçları	58
Tablo 4.7: Önerilen algoritmanın 5-kat çapraz geçerleme için F-ölçütü sonuçları	59
Tablo 4.8: Önerilen algoritmanın 10-kat çapraz geçerleme için F-ölçütü sonuçları	60
Tablo 4.9: Önerilen model atak dayanıklılığı sonuçları	61
Tablo 4.10: Tüm sonuçların FMR değerleri ile kıyaslanması	62

SİMGE VE KISALTMA LİSTESİ

Simgeler	Açıklama
a	: Lineer kongrüans üreteçlerde kat sayısı
c	: Lineer kongrüans üreteçlerde artırım sayısı
D	: Orijinal veri seti
E[D]	: farkların standart sapması
G(X)	: Geometrik pertürbasyon sonrası X veri seti
I	: Birim matris
ID	: Tanımlayıcı nitelikler
k	: Çapraz geçirmede kat sayısı/k-anonimlikte k sayısı
l	: l-çeşitlilikte l sayısı
m	: Lineer kongrüans üreteçlerde mod değeri
min(Y_i)	: Y. sütundaki minimum değer
max(Y_i)	: Y. sütundaki maksimum değer
N	: Matris boyutu/ sınıflandırmada negatif grupların sayısı
N_i	: Önerilen algoritmada i. gürültü matrisi sütunu
p_i	: Sütunun gizlilik seviyesi
P	: Sınıflandırmada pozitif grupların sayısı
P (A)	: Bayes teoreminde A olayının gerçekleşme olasılığı
P (B)	: Bayes teoreminde B olayının gerçekleşme olasılığı
P (A B)	: Bayes teoreminde B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı
P (B A)	: Bayes teoreminde A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı
QID	: Yarı tanımlayıcı nitelikler
R	: Rotasyon matrisi
SA	: Önerilen yöntemdeki hassas nitelikler
SA_i	: i. hassas nitelik
t	: t- yakınlıkta t değeri
T	: Öteleme matrisi gösterimi
Var[D]	: Farkların varyansı
VoD	: Farkların varyansı

w_i	: Sütunun ağırlık seviyesi
X	: Pertürbasyon uygulanacak veri seti
X_i	: i. sütun değeri
Y_i	: i. sütun değeri
Y_{si}	: i. sütun pertürbasyon sonrası değeri
$ D $	: Orijinal D matrisinin boyutu
$ SA $	: Hassas nitelik sayısı
$\emptyset_1$	: minimum gizlilik garantisi
$\emptyset_2$	: ortalama gizlilik garantisi
Δ :	: Gürültü matrisi



Kısaltmalar	Açıklama
AES	: Advanced Encryption Standard (Gelişmiş Şifreleme Standardı)
BUG	: Buttom Up Generalization (Aşağıdan Yukarıya Genelleştirme)
DC	: Data Condensation (Veri Yoğunlaştırma)
DT	: Decision Table (Karar Tablosu)
FN	: False Negatives (Yanlış Negatifler)
FMR	: Friedman Rank
FP	: False Positives (Yanlış Pozitifler)
HDFS	: Hadoop Distributed File System (Hadoop Dağıtık Dosya Sistemi)
HiveQL	: Hive Query Language (Hive Sorgu Dili)
HRNG	: Hardware Random Number Generator (Donanımsal Rasgele Sayı Üreteçleri)
ICA	: Independent Component Analysis (Bağımsız Bileşen Analizi)
ID	: Identifier (Tanımlayıcı)
ID3	: Iterative Dichotomiser 3(Yinelemeli Dikotomizör 3)
I/O	: Input Output(Girdi / Çıktı)
IoT	: Internet of Things (Nesnelerin İnterneti)
IP	: Internet Protocol (İnternet Protokolü)
k-NN	: k-Nearest Neighbours (k-En Yakın Komşu)
KVKK	: Kişisel Verileri Koruma Kanunu
MDR	: Mean Difference Recommending (Ortalama Fark Önerisi)
ML	: Machine Learning (Makine Öğrenmesi)
NB	: Naive Bayes
NI	: Naïve Interference (Naif Yansıtma)
OneR	: One Rule (Tek Kural)
PRNG	: Pseudo Random Number Generator (Söзде Rasgele Sayı Üreteçleri)
RAM	: Random Access Memory(Rasgele Erişimli Bellek)
RDD	: Resilient Distributed Datasets (Esnek Dağıtılmış Veri Kümeleri)
RP	: Rotation Perturbation (Rotasyonel Pertürbasyon)
RSU	: Rasgele Sayı Üreteçleri
RSUGP	: Rasgele Sayı Üreteçlerine bağlı Geometrik Pertürbasyon
QID	: Quasi Identifier (Yarı Tanımlayıcı)
SA	: Sensitive Attributes (Hassas Nitelikler)
SQL	: Structured Query Language (Yapılandırılmış Sorgu Dili)

TDS	: Top Down Specialization (Yukarıdan Aşağıya Uzmanlaşma)
TN	: True Negatives (Doğru Negatifler)
TP	: True Positives (Doğru Pozitifler)
VM	: Virtual Machine (Sanal Makine)
YARN	: Yet Another Resource Negotiator (Başka Kaynak Müzakerecisi)
3V	: Volume, Velocity and Variety (Hacim, Hız ve Çeşitlilik)
7V	: Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value (Hacim, Hız, Çeşitlilik, Değişkenlik, Doğruluk, Görselleştirme ve Değer)



ÖZET**DOKTORA TEZİ****ETKİN GİZLİLİK KORUMASI İÇİN RASGELE SAYI ÜRETİCİNE DAYALI
VERİ ANONİMİZASYON MODELİNİN GELİŞTİRİLMESİ****Merve KANMAZ****İstanbul Üniversitesi-Cerrahpaşa****Lisansüstü Eğitim Enstitüsü****Bilgisayar Mühendisliği Anabilim Dalı****Danışman : Prof. Dr. Ahmet SERTBAŞ****II. Danışman : Doç. Dr. Muhammed Ali AYDIN**

Teknolojinin hızlı gelişimi ve hayatımızın her alanına girmesi sonucunda verinin hacmi ve değeri oldukça önemli bir çalışma alanı olmuştur. Verinin bu denli kıymetlenmesi kişilerin bilgileri açısından bazı sonuçlar doğurmuştur. Veri kimliksizleştirme, kişisel verilerin güvenliği açısından bu konulardan en önemlisidir. Bu alanda birçok çalışma yapılmıştır ve yapılmaya devam edilmektedir. Bu çalışmada ise hassas niteliklerin kimliksizleştirilmesi için RSUGP isimli bir yöntem önerilmiştir. Geometrik veri pertürbasyonunda klasik olarak kullanılan gauss gürültüsü ya da rasgele gürültü yöntemlerinin yerine rasgele sayı üreteçlerine bağlı yeni bir gürültü modeli önerilmiştir. Önerilen RSUGP yöntemi Hadoop dosya sistemi HDFS üzerinde tutulan veri setlerinin Apache Spark ile işlenmesi sonrasında sınıflandırma doğruluğu, F- ölçütü, atak dayanıklılığı açısından 6 farklı veri tabanı ve 4 farklı sınıflandırma yöntemi ile test edilmiş, bulgular sonuçlar kısmında paylaşılmıştır. Sonuçlar önerilen yöntemin sınıflandırma doğruluğu, atak dayanıklılığı ve çalışma zamanı açısından kıyaslanan diğer yöntemlere göre daha başarılı olduğunu ortaya koymuştur.

Eylül 2021, 97 sayfa.

Anahtar kelimeler: Gizlilik Korumalı Veri Madenciliği, Veri Anonimleştirme, Veri Pertürbasyonu, Geometrik Pertürbasyon, Rasgele Sayı Üreteçleri



SUMMARY**Ph.D. THESIS****DEVELOPMENT OF DATA ANONYMIZATION MODEL BASED ON
RANDOM NUMBER GENERATOR FOR EFFICIENT PRIVACY
PRESERVING****Merve KANMAZ****Istanbul University-Cerrahpasa****Institute of Graduate Studies****Department of Computer Engineering****Supervisor : Prof. Dr. Ahmet SERTBAŞ****Co-Supervisor : Assoc. Prof. Dr. Muhammed Ali AYDIN**

With the technology's rapid development and its involvement in all areas of our lives, the volume and value of data have become a significant field of study. Valuation of the data to this extent has produced some consequences in terms of people's knowledge. Data anonymization is the most important of these issues in terms of the security of personal data. Much work has been done in this area and continues to being done. In this study, we proposed a method called RSUGP for the anonymization of sensitive attributes. A new noise model based on random number generators has been proposed instead of the Gaussian noise or random noise methods, which are being used conventionally in geometric data perturbation. The proposed RSUGP method has been tested with 6 different databases and 4 different classification methods in terms of classification accuracy, F- score, attack resistance, execution time and Friedman rank test. After the data sets kept on the Hadoop file system HDFS are processed with Apache Spark, and the findings are shared in the results section. Experiments show that the proposed method was more successful than the other two classification accuracy, attack resistance, and runtime.

September 2021, 97 pages.

Keywords: Privacy-Preserving Data Mining, Data Anonymization, Data Perturbation, Geometric Perturbation, Random Number Generators



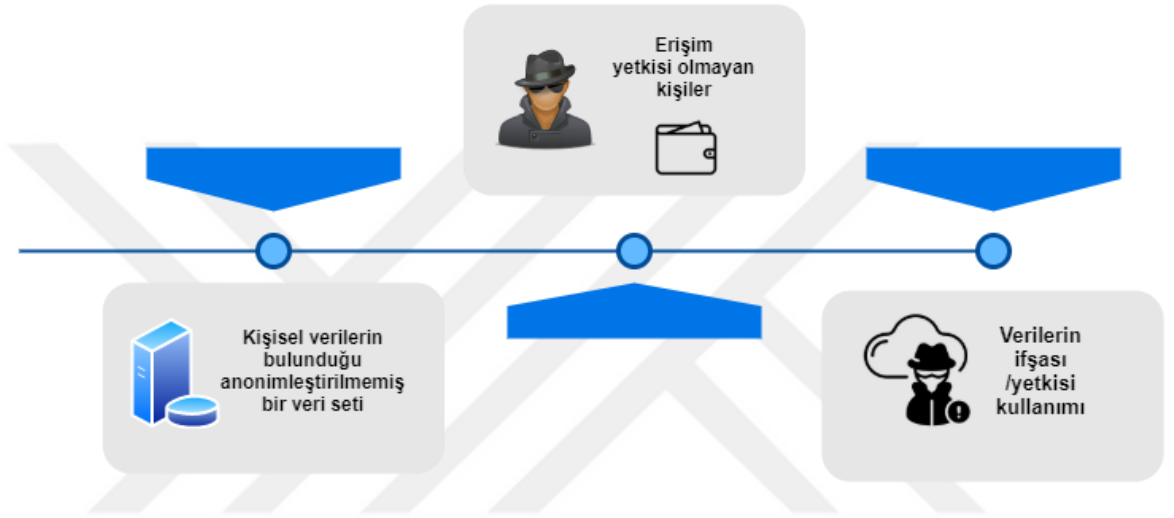
1 GİRİŞ

Günümüz teknoloji çağında veri önemli bir bilgi kaynağı olarak görülmektedir. Teknolojinin gelişmesi, sosyal medya uygulamalarının artması, nesnelerin interneti gibi kavramlarında hayatımıza girmesiyle birlikte üretilen ve işlenmesi gereken verilen boyutları da oldukça artmıştır. Daha öncelerde içerik sağlayıcısı konumunda olan internet bireylerin içeriklerini oluşturup paylaşabildiği, kendilerine ait bir sanal bir dünya oluşturabildiği bir platforma dönüşmüştür. Mobil teknolojilerinde artmasıyla birlikte kişilerin masa başı olmayan ortamlarda da bu platforma istedikleri her saniye erişim sağlayabilmesi, kullanılan ve üretilen veride ciddi bir büyümeye yol açmıştır[1]. Büyüyen boyut ile verinin değeri de büyümüştür[2]. Artık veri çağında yaşadığımız ve çağımızın petrolünün veri olduğu bile söylenmektedir. Hal böyle olunca çok farklı büyük veri kaynakları vardır. Örneğin bir uçak motoru saniyede 5000 farklı bileşen için veri üretmektedir. Sosyal medyada yapılan paylaşımların istatistiklerini tutan bir siteden alınan bilgilere göre sadece Twitter’da günlük 500 milyon tweet atılmaktadır, Youtube’da ise dakikada 90 bin video izleme yapılmaktadır[3]. Yine video, resim uygulamaları dolayısı ile üretilen içerikler, e -ticaret verileri, internet reklamları, tıbbi cihazlardan okunan veriler, hasta takip verileri ve en önemlisi nesnelerin interneti kavramı ile cihazların üretmiş olduğu veriler büyük veri kavramındaki büyük sıfatını tanımlamak için oldukça yeterli örneklerdir.

Bu derece büyük verinin toplanması ve depolanması ile ilgili teknolojik gelişmelerin de hız kazanmasıyla birlikte veriden anlamlı bilgiler çıkarma bilimi olan veri madenciliği çalışmaları da oldukça önem kazanmıştır. Birçok alanda veri sayısallaştırılarak hammadde haline getirilmekte ve yeni verilerin üretilmesine olanak sağlamaktadır. Verilerin analitik yöntemler ile işlenmesi sonrasında veriler ekonomik bir değer haline getirilerek kazanç sağlamaktadır[4]. Kazanılan bu değerle birlikte organizasyonlar rekabet avantajı sağlamaktadırlar. Bu kazanımlar sadece organizasyonlara değil sağlık bilimleri, kamu, uzay bilimleri, güvenlik gibi alanlarda yapılan çalışmalara da katma değer katmaktadır.

Böylelikle hem günlük yaşamımızın hem de iş hayatının vazgeçilmez bir parçası olan verinin, büyük veri kavramının da 3Vsi olarak adlandırılan hacim(volume), hız(velocity) ve çeşitlilik(variety) değerleri hızlıca büyümektedir. Bu değerler büyürken depolama, veri işleme ve güvenlik problemlerini de beraberinde getirmektedir. Farklı kaynaklardan farklı

konfigürasyonlarla gelen verileri ilişkisel veri tabanı yöntemleri işlemek için yetersiz kalmakta ve yapısal olmayan bu verileri işleyecek ve depolayabilecek dağıtık ortamlara ihtiyaç duyulmaktadır[5]. Bununla birlikte kişiler ve kuruluşlar hakkında önem arz eden hassas bilgilerin, madencilik çalışmalarında ya da bilgiye erişimin olduğu herhangi bir alanda gizliliğinin sağlanması da ortaya çıkan problemlerden biridir. Burada kullanılan gizlilik terimi, verinin sadece saldırganlardan gelecek olan tehditlerden korunması olarak düşünülmemeli, yetkisiz erişimler ve yetkisiz kullanımların önüne geçilmesi hedefi olarak düşünülmelidir[6,7].



Şekil 1.1: Yetkisiz erişim sembolize edilmesi

Geleneksel (ilişkisel) veri tabanları için birçok başarılı kimliksizleştirme ve veri güvenliği çözümü bulunmaktadır[8,9,10,11,12]. Ama yeni bir teknoloji olan Hadoop, Spark, Cloud gibi dağıtık ortamlar üzerinde veri mahremiyetini sağlarken verilere erişim konusunda performans problemi yaratmayacak yenilikçi çözümlere ihtiyaç duymaktadır. Birçok kuruluş sahip olduğu büyük miktardaki veriyi yakın zamana kadar ilişkisel veri tabanları üzerinde tutmaktaydı. Kendi zamanlarının vazgeçilmezi olan bu makineler ile çalışan sistemler verinin büyüme hızına karşı, depolama ve verinin işleniş açısından performans üst limitine ulaşmışlardır. Bu limit karşısında donanımların artırılması hem aşırı maliyetli hem de ihtiyacı karşılamakta yetersiz kalmaktadır. Tüm bu nedenlerden ötürü büyük miktarda veri için dağıtık veri depolama ve veri işleme konusunda kabul görmüş ihtiyaca göre genişleyebilecek maliyeti nispeten düşük Hadoop gibi platformlara geçiş ihtiyacı oluşmuştur. İlişkisel veri tabanı için kullanılan geleneksel kimliksizleştirme yöntemleri büyük veri işleyen platformlar için kullanılamamaktadır. Büyük verilerin işlendiği bu tür sistemler, hesaplamalarını yaparken verileri sistemdeki alt makinalara dağıtır ve hesaplamaların tamamlanmasından sonra sonucu

birleřtirip geriye döndürür. Bu süreçte hem alt makinalarda hem de en üstte toplanan bilgiler için mahremiyetin sağlanması gerekmektedir. Bu nedenle ilişkisel veri tabanı için kullanılan uygulamalarda kullanılan geleneksel kimliksizleştirme yöntemleri büyük veri işleyen bu tarz sistemler için uygun olmamaktadır. Bu problemin giderilmesi için ya yeni bir yaklaşım ortaya konulmalı ya da hali hazırda kullanılan kimliksizleştirme çözümleri bu sistemler için uyarlanmalıdır.

Verinin kimliksizleştirilmesindeki en önemli olay, kimliksizleştirilen verinin kullanılabilirliğinin devamı sağlamaktadır. Aksi takdirde yapılan işlemin eldeki verilerin rasgele değerlerle değiştirilmesinden bir farkı kalmayacaktır. Bu şekilde verilerin değerlerini yitirmeden değiştirilerek yayınlanması işlemine gizlilik korumalı veri yayınlama denilmektedir. Orijinal veri seti üzerindeki veriler 4 farklı nitelik olarak değerlendirilmektedir. Bunlar tanımlayıcı, yarı tanımlayıcı, hassas ve hassas olmayan nitelikler olarak sınıflandırılmaktadır. Kimliksizleştirme çalışmaları hassas niteliklerin saldırganlar ya da erişim yetkisi olmayan kişilerin eline geçmesini engellemek için yapılmaktadır.

Gizlilik koruması için kullanılan anonimleştirme yöntemleri beş tanedir. Bunlar genelleştirme, gizleme, anatomizasyon, permütasyon ve pertürbasyon yöntemleridir. Genelleştirme değerlerin daha genel ifadeler ile değiřtirmesidir. Gizleme işlemi ise veri içerisindeki belirli değerleri kaldırmak ya da *,# gibi karakterlerle saklama işlemidir. Anatomizasyon işleminde ise verilerdeki hassas nitelikler ile yarı tanımlayıcılar farklı tablolara ayrılarak aradaki ilişki yok edilir. Permütasyon işleminde veriler gruplara ayrılır ve her grup içindeki hassas nitelikler kendi içinde karışırılır. Son olarak pertürbasyonda ise verilerin orijinal değerleri değiştirilerek yerine yapay değerler konulur. Böylece asıl değer gizlenmiş olur. Sonuç olarak tüm bu yöntemler sonrasında veride bir miktar kayıp olmakta ve kullanılabilirliği azalmaktadır.

Literatürde yapılmış olan çalışmalar incelendiğinde kimliksizleştirme çalışmalarının yaygın olarak devam ettiğı görülmüştür. Özellikle daha geniş kapsamı olması nedeniyle pertürbasyon yöntemi aktif olarak kullanılmaktadır. Buradan yola çıkarak yapılan arařtırmalar sonucunda bu tez çalışmasında da pertürbasyon yöntemi kullanılmıştır. Geometrik veri pertürbasyonu ve rasgele sayı üreteçleri kullanılarak yeni bir kimliksizleştirme modeli ortaya konulmuştur. Geometrik veri pertürbasyonu ile verilerin birden fazla aşama ile kimliksizleştirilmesi sağlanmıştır. Ayrıca bilgisayar bilimleri alanında sıklıkla kullanılmakta olan rasgele sayı üreteçleri ile pertürbasyon yöntemi kuvvetlendirilmiş, böylece ortaya kapsamlı bir model

ıkmıřtır. alıřmanın literatre katkısı tam olarak budur. Rasgele sayı retelerinin zellikle kriptografik alıřmalarda kullanılıyor olması teze fikrini vermiřtir.

Tez alıřmasının geri kalanı řu řekilde dzenlenmiřtir: Genel Kısımlar blmnde; tez konusu kapsamında literatrde bugne kadar yapılmıř olan alıřmalardan bahsedilmiřtir. Malzeme ve Yntem blmnde; tez alıřması kapsamında nerilen modelin kullandığı geometrik veri pertrbasyonu ve rasgele sayı reteleri detaylıca anlatılmıřtır. Ayrıca kullanılan veri setleri, sınıflandırma yntemleri, nerilen sistemin kullanılacağı platform anlatılmıřtır. Bulgular blmnde; deęerlendirme lt olarak sınıflandırma doęruluęu, F- lt , atak dayanıklılığı, alıřma zamanı ve Friedman rank testi sonuları dięer yaklařımlarla karřılařtırmalı olarak sunulmuřtur. Son olarak tartıřma ve sonu blmnde; nerilen modelin literatre katkısından, performans sonularından ve mevcut algoritmalarından hangi aılardan daha etkili olduęundan bahsedilmiřtir.

2 GENEL KISIMLAR

Bu bölümde veri anonimleştirme yöntemleri detaylıca anlatılarak literatürde var olan çalışmalardan bahsedilecektir.

2.1 VERİ MAHREMİYETİ

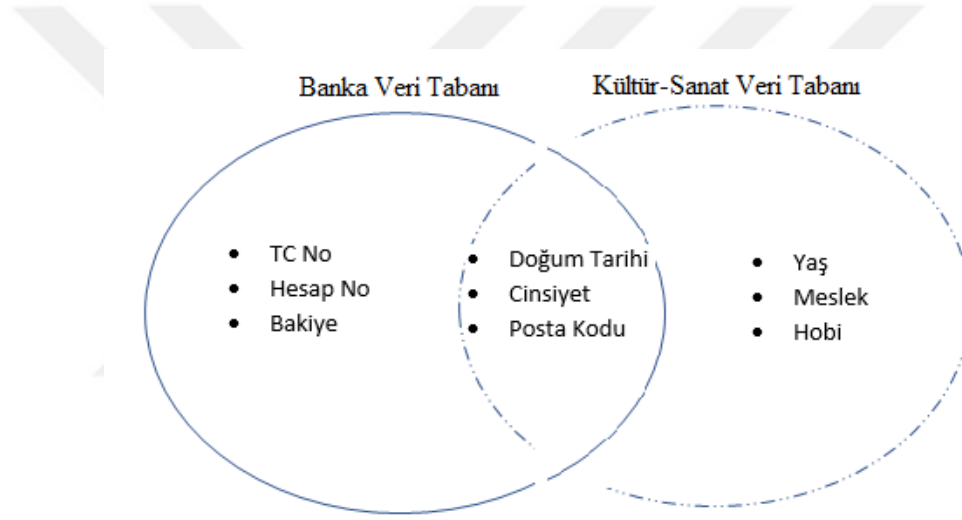
Veri madenciliği var olan veri tabanının analiz edilerek veriden anlamlı veriler çıkarma işlemine denilmektedir. Son yıllarda kümeleme, sınıflandırma gibi veri madenciliği algoritmaları oldukça güçlü analizler elde etmişlerdir. Yapılan bu analizlerle birlikte özellikle sağlık, finansal ve güvenlik alanlarında verilerin gizliliği konusu oldukça önemli hale gelmiştir. Dünya çapındaki yenilikçi organizasyonlar, daha yüksek değerli müşterileri bulmak ve onlara hitap etmek, satışlarını artırmak, ürün tekliflerini yeniden yapılandırmak için ve hata veya sahtekârlıktan kaynaklanan kayıpları en aza indirmek için zaten veri madenciliğini kullanmaktadırlar.

Hal bu olunca kişisel bilgiler hem kişiler hem de kurumlar açısından oldukça önem arz etmektedir. Kurumlar verilerden çıkarılacak anlamlı bilgilerle ilgilenirken, şahıslar kendilerine ait olan bu bilgilerin gizliliğiyle ilgilenmektedir. Böylelikle gizlilik korumalı veri madenciliği kavramı ortaya çıkmış, üzerinde çalışılmaya devam edilen önemli çalışma alanlarından biri haline gelmiştir.

Bu noktada olay bireysellikten çıkıp genel bir güvenlik durumu oluşturduğu için özel hayat ve kişisel verinin gizlilik koruması adına hem ulusal hem de uluslararası düzenlemeler yapılmıştır. Bu anlamda ilk geniş kapsamlı uluslararası sözleşme Avrupa Konseyi tarafından 28 Ocak 1981 tarihinde kabul edilen 108 sayılı “Kişisel Verilerin Otomatik İşleme Tabi Tutulması Karşısında Şahısların Korunmasına Dair Sözleşme” olmuştur[13]. Ayrıca ülkemizde anayasada bulunan “herkes, özel hayatına ve aile hayatına saygı gösterilmesini isteme hakkına sahiptir. Özel hayatın ve aile hayatının gizliliğine dokunulamaz” hükmü gereğince hazırlanan 6698 sayılı “Kişisel Verilerin Korunması Kanunu[14]” da kişisel veri, özel nitelikli kişisel veri ve kişisel verinin işlenmesi kavramlarını tanımlamakta ve kişisel verinin işlenme şartlarını açıklamaktadır[15].

Veri kimliksizleştirme çalışmaları iki grup tehditten koruma amaçlı yapılmaktadır. Bunlar kullanıcı kimlik ifşası ve hassas niteliklerin ifşası olarak sınıflandırılmaktadır. Kimlik ifşasında veri seti içerisindeki bazı niteliklerin farklı veri setleri içerisindeki verilerle eşleştirilmesi sonucunda veri sahiplerinin ortaya çıkarılması işlemidir. Hassas niteliğin ifşasında ise, kişilerin başkaları tarafından bilinmesi istenmeyen bilgilerinin tespit edilmesidir. Bunun önlenmesi için verilere kimliksizleştirme ya da anonimleştirme denilen yöntemler uygulanmaktadır.

Kimliksizleştirme veriyi, ona sahip olan kişiden bağımsız bir hale getirme işlemidir[16]. Bu işlem sonrasında başka verilerle eşleştirilse dahi, hiçbir şekilde belli bir kimlik ile eşleştirilemez ya da dolaylı yoldan kimlik belirlenemez durumdadır. Kimliksizleştirme işleminin yapılması için her kayıt aşağıda belirtilen niteliklere ayrılır (Şekil 2.1).



Şekil 2.1: Niteliklerin örnek gösterimi

- **Tanımlayıcı(Identifier-ID):** Kişiyi başka bir özelliğe ihtiyaç duymadan tek başına tanımlayabilen özelliklerdir. TC kimlik numarası, hesap numarası gibi.
- **Yarı Tanımlayıcı (Quasi-identifier-QID):** Kişiyi tek başına tanımlayamayan ancak diğer veri setleriyle ilişkilendirilerek tanımlayabilecek olan özellik setleridir. Cinsiyet ve posta kodu gibi.
- **Hassas Nitelik(Sensitive attribute-SA):** Kişinin diğer bireylerden gizli kalması gereken özellikleridir. Maaş bilgisi, not bilgisi, hastalık bilgisi gibi.

- **Hassas Olmayan Nitelik(Non-Sensitive attribute-SA):** Bilgisi erişimi olmayan kişilerin eline geçse dahi kişiyi ifşa etmeyecek olan özelliklerdir. Hobi, yorumları, sigorta şirketi gibi.

Gizlilik korumalı veri madenciliğinde veriler yayınlanmadan önce anonimleştirme tekniklerinden birisi kullanılarak hassas niteliklerin gizliliği sağlanır. Bu yöntemler aşağıdaki gibidir:

Genelleştirme(Generalization): Yarı tanımlayıcı olan niteliklerin , niteliği daha az betimleyecek olan değerler ile değiştirilmesi işlemidir. Yani ilgili veri, kişiye özel bir değer değil de daha genel bir değere çevrilir[17-20]. Örneğin veri Tablo 2.1 üzerinde gösterildiği gibi olsun. Bu durumda mesleği bilinen bir kişinin test sonucunun pozitif veya negatif olması direk anlaşılır durumdadır. Tablo 2.2 de gösterildiği gibi meslek sütununda genelleştirme yapılmıştır. Böylece meslek dalı bilinse bile detayı tablo üzerinde yazmadığı için hastalık durumu direk bilinemeyecektir. Genelleştirme toplam rakamlar üzerinden değerlendirme yapılan durumlarda daha sık tercih edilir. Örneğin 12345678901 TC kimlik numarasına sahip olan bir birey internet üzerinden yapmış olduğu alışverişinde deterjan alırken aynı zaman yumuşatıcıda almış olsun. Yapılan raporlama da deterjan alanların %x' i aynı zamanda yumuşatıcı da almaktadır şeklinde şahıslardan bağımsız bir değerlendirme yapılabilir. Genelleştirme türleri; tam alan genelleme (full domain generalization), hücre genelleme (cell generalization), çok boyutlu genelleme (multidimensional generalization), kardeş genelleme (sibling generalization) ve alt ağaç genelleme (subtree generalization) olarak bilinmektedir.

Tablo 2.1 : Genelleştirme yapılmadan önce hasta verileri

Yaş	Cinsiyet	Meslek	Test Sonucu
22	K	Bilgisayar Mühendisi	Pozitif
35	E	Ressam	Negatif
46	E	Müzisyen	Pozitif
28	K	Endüstri Mühendisi	Pozitif
17	K	Matematik Öğretmeni	Negatif
44	E	Tiyatrocu	Negatif
35	K	Kimya Öğretmeni	Pozitif

Tablo 2.2: Genelleştirme yapıldıktan sonra hasta verileri

Yaş	Cinsiyet	Meslek	Test Sonucu
22	K	Mühendis	Pozitif
35	E	Sanatçı	Negatif
46	E	Sanatçı	Pozitif
28	K	Mühendis	Pozitif
17	K	Öğretmen	Negatif
44	E	Sanatçı	Negatif
35	K	Öğretmen	Pozitif

Gizleme (Suppression): Bu yöntemde bilinen bazı değerler, özel karakterler ile yer değiştirilerek gerçek değeri gizlenmiş olur. Gizleme işlemindeki amaç tahmin edilme olasılığını azaltmaktır. Gizleme yöntemi, hücre üzerinde, kayıt üzerinde ve değer üzerinde ayrı ayrı yapılabilir.

Tablo 2.3: Gizleme yapılmadan önce hasta verisi

Yaş	Cinsiyet	Posta Kodu	Test Sonucu
22	K	34030	Pozitif
35	E	34040	Negatif
46	E	34040	Pozitif
28	K	34050	Pozitif
17	K	34055	Negatif
44	E	34030	Negatif
35	K	34045	Pozitif

Tablo 2.3 veriler üzerinde herhangi bir anonimleştirme işlemi yapılmamış halidir. Tablonun bu durumunda kişinin posta kodu bilgisine sahip olan birisi test sonucunu elde edebilir. Tablo 2.4 ise Posta Kodu sütunundaki değerler üzerinde gizleme yapıldıktan sonraki durumu göstermektedir. Posta kodunun son iki hanesi * karakteri ile gizlenerek , semt bilgisinin görünmesinin önüne geçilmiştir.

Tablo 2.4: Gizleme yapıldıktan sonra hasta verileri

Yaş	Cinsiyet	Posta Kodu	Test Sonucu
22	K	340**	Pozitif
35	E	340**	Negatif
46	E	340**	Pozitif
28	K	340**	Pozitif
17	K	340**	Negatif
44	E	340**	Negatif
35	K	340**	Pozitif

Anatomizasyon (Anatomization): Hassas verileri içeren tablo yarı tanımlayıcılar ve hassas nitelikler olmak üzere iki ayrı tabloya parçalanır[21]. Yani veriler genelleştirilmez ya da gizlenmez, kişi bilgisini ifşa edemeyecek şekilde iki ayrı tabloda tutulur. İki tabloda aradaki ilişkiyi koparmamak adına ortak bir nitelik barındırır. Bu niteliğe grup numarası adı verilir.

Permütasyon (Permutation): Bu yöntemde veri gruplara ayrılır ve her grup içerisindeki hassas veriler karıştırılarak kimiksizleştirme işlemi yapılır[22].

Pertürbasyon (Perturbation): Pertürbasyon yönteminde veri seti üzerindeki değerler, gerçekte değeri olmayan sentetik değerler ile değiştirilir . Bu değişim ile verilerin dağılımı ve istatistiksel hesaplamalarında herhangi bir bozulma söz konusu olmaz. Veri değiş tokuşu (kayıt çiftlerinin kendi aralarında yer değiştirmesi), gürültü ekleme (orijinal veriye üzerindeki alana eşit miktarda veri ekleme) , sentetik değer oluşturma (anlam ifade etmeyen veriler belirleme) gibi seçenekleri vardır. Tablo 2.5 de orijinali verilmiş veri setinin, Tablo 2.6 üzerinde Gelir sütunu üzerine gürültü ekleyerek pertürbasyon yapılmış hali mevcuttur.

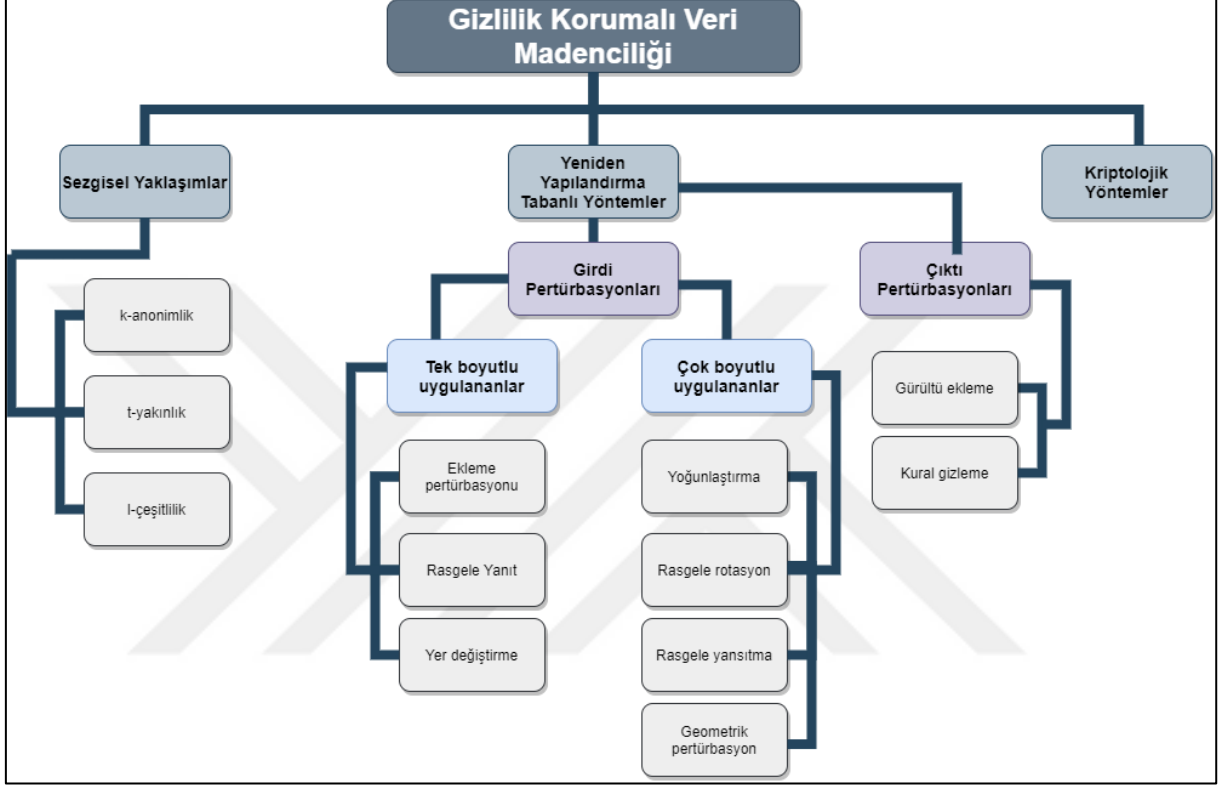
Tablo 2.5: Gürültü eklenmemiş orijinal veri seti

Yaş	Cinsiyet	İl	Gelir
22	K	Adana	15000
35	E	İzmir	7500
46	E	İstanbul	34000
28	K	İstanbul	22000
17	K	Trabzon	5000
44	E	İzmir	25000
35	K	Gaziantep	12000

Tablo 2.6: Gelir sütununa gürültü eklenmiş veri seti

Yaş	Cinsiyet	İl	Gelir
22	K	Adana	25000
35	E	İzmir	17500
46	E	İstanbul	44000
28	K	İstanbul	32000
17	K	Trabzon	15000
44	E	İzmir	35000
35	K	Gaziantep	22000

Veri mahremiyetini korumaya yönelik kullanılan yöntemler sezgisel yaklaşımlar, yeniden yapılandırma tabanlı yöntemler ve kriptolojik yöntemler olarak sınıflandırılmaktadır[23]. Şekil 2.2 üzerinde bu yöntemler gösterilmiştir.



Şekil 2.2: Gizlilik korumalı veri madenciliği yöntemleri

Anonimleştirme tabanlı çözüm süreçlerinde dört farklı rol bulunmaktadır. Bunlar veri sahipleri, veriyi paylaşan kurum ya da kişiler, paylaşılan bu veriye erişim yetkisi bulunanlar ve erişim yetkisi olmayıp saldırı düzenleyenler. Bu rollere bakarak kişilerin hassas nitelikleri yayınlanmadan önce saldırganların ya da yetkisiz kişilerin erişim kısıtlaması için anonimleştirilerek yayınlanırlar. Bu aşamada kurumlar ya da kişiler verileri yayınlamadan önce bazı süreçler uygulayabilirler.

- Tek sürüm yayınlama modelinde veri setinin daha sonra üzerine yeni veri eklenmeyecek ve anonimleştirme işleminde herhangi bir güncellemeye gidilmeyecek şekilde yayınlanmasıdır.

- Veriler anonimleştirme sonrasında az da olsa bilgi kaybına uğrarlar. Veri yayıncısı bu durumun önüne geçmek için, yarı tanımlayıcılarla farklı gruplar oluşturarak anonimleştirdiği veriyi ayrı parçalar halinde sunar. Bu modele paralel yayınlama denilmektedir.
- Sıralı yayınlama modelinde ise, verilerin devamlı olarak arttığı düşünülerek, anonimleştirme sonrasında gelen yeni veriler ile tekrar anonimleştirme yapılır. Burada sistemi devamlı besleme söz konusudur ve birden fazla anonim versiyon oluşturulur.

2.1.1 Sezgisel Yaklaşımlar

Sezgisel yöntemler daha çok gizlenen bilginin sınırlarının belirlenmesi için kullanılan istatistiksel yöntemlerdir. Bu yöntemlerden en bilinenleri k -anonimlik, l -çeşitlilik ve t -yakınlıktır.

2.1.1.1 K -Anonimlik

Anonimleştirilmiş veri kümelerinde yarı tanımlayıcı değerlerin doğru kombinasyonlarla bir araya getirilmesi halinde bilgilerin kolayca saptanabildiği tespit edilmiştir[24]. Bu sebeple anonim hale getirilmiş olan verinin bazı istatistiksel yöntemler ile güvenilirliğinin saptanması gerektiği fikri ortaya çıkmıştır.

Samarati ve Sweeney tarafından önerilen bu modelde[25,18] saldırganın, bir şekilde yarı tanımlayıcı değerleri elde etmesi halinde, yayınlanan verideki her kaydın verilerdeki en az $k-1$ tane kayıttan ayırt edilememesi sağlanmaktadır. Yani belirli alanlarda birden fazla kişinin bulunması sağlanarak, bazı durumlarda tekil özellik gösteren kişilerin bilgilerinin ifşa edilmesinin önüne geçmeyi sağlamaktadır[26]. Daha büyük k değeri, daha yüksek korumayı ifade etmektedir ancak değer büyüdükçe verideki bozulma da artmaktadır[27-30].

2.1.1.2 L -Çeşitlilik

$K=4$ anonimliğin eksikliği üzerine yapılan çalışmalarda hassas nitelik çeşitliliğinin az olduğu durumlarda, k -anonimliğin gizliliği yeterinde korumadığı ve verilerden çıkarım yapılabilceği tespit edilmiştir[24]. Bunun üzerine yarı tanımlayıcı ve hassas niteliklerin arasındaki ilişkiye dayanan l -çeşitlilik ilkesi öne sürülmüştür[31]. Yönteme göre seçilen bir yarı tanımlayıcı grup içerisinde en az l çeşit hassas nitelik bulunuyorsa, l -çeşitlilik sağlanmıştır denilir.

2.1.1.3 T-Yakınlık

İ-çeşitlilik kişisel verilerde çeşitlilik sağlayarak ifşaların önüne geçmesine rağmen verilerin içeriği ve hassasiyet durumu ile ilgilenmemesinden ötürü yeteri kadar koruma sağlamadığından yola çıkılarak t-yakınlık ilkesi öne sürülmüştür[32]. Bu yöntemde değerlerin birbirlerine yakınlık dereceleri hesaplanır ve değerlere göre veriler alt sınıflara ayrılarak anonim hale getirilir.

Birçok çalışmada t değeri sabit eşik değer olarak kabul edilmektedir ve önerilen modeller bu değere göre şekillenmektedir[33-36]. Buna karşın bazı çalışmalar da t değerini sabit bir değer olarak değil de kullanılan veri setine göre dinamik olarak hesaplamaktadır[37].

2.1.2 Yeniden Yapılandırma Tabanlı Yöntemler

Yeniden yapılandırma tabanlı yöntemlere pertürbasyon teknikleri de denilmektedir. Pertürbasyon veri ya da veri setlerinin kendileri ile aynı dağılıma sahip yapay veriler ile değiştirilmesi olarak kullanılmaktadır. Verinin değiştirilmesi/bozulmasındaki en önemli ve en büyük zorluk gizliliği korurken verinin kalitesinin ve dengesinin bozulmamasını sağlamaktır. Gizlilik korumalı veri madenciliğinde birçok yaklaşım bulunmakla birlikte veri pertürbasyonu basit ve etkin bir yöntem olmasından dolayı sıklıkla kullanılan yöntemlerdendir[22]. Veri pertürbasyonu girdi pertürbasyonu ve çıktı pertürbasyonu olarak iki sınıfta incelenmektedir. Girdi pertürbasyonu eklemeli gürültü ya da çarpımsal gürültü yöntemlerinin birini uygularken, çıktı pertürbasyonunda gürültü ekleme ve kural gizleme[38-40] yaklaşımları birlikte uygulanır[41]. Girdi pertürbasyonu; eklemeli pertürbasyon[42,43], rasgele yanıt[44] ve yer değiştirme[45-48] olarak tek boyutlu uygulanırken, yoğunlaştırma[49], rasgele rotasyon[50-54], rasgele yansıtma[55] ve geometrik pertürbasyon[56] olarak çok boyutlu uygulanabilmektedir.

2.1.3 Kriptolojik Yöntemler

Kriptolojik yöntemler klasik şifreleme algoritmalarının anonimleştirme modellerine uygulandığı yöntemlerdir. Sıklıkla kullanılmasının nedeni kriptolojik yöntemler ile hem iyi bir gizlilik modeli tanımlanabilmesi hem de bu alanda genişçe araç setlerinin bulunmasıdır[57,58]. Burada unutulmamalıdır ki kriptolojik yöntemler çıktı pertürbasyonu sağlamaz yalnızca hesaplama sürecindeki gizlilik sızıntılarını önler. Ayrıca zaman karmaşıklığı açısından diğer

yöntemlere göre daha büyük bir karmaşıklığa sahiplerdir[23]. Farklı gizlilik protokolleri kullanılarak yapılan çalışmalara literatürde rastlanmaktadır[58-62].

2.2 GİZLİLİK KORUMAYA YÖNELİK ÇALIŞMALAR

Gizlilik korumalı veri madenciliği, verinin madencilik yapılarak anlam çıkarılmasından önce gizliliğini sağlayacak çalışmaların yapılmasıdır. Kişisel bilgileri içeren veri setlerinin yayınlanmadan önce bazı uygulamalara tabi tutularak bu bilgileri ifşa etmeyecek konuma getirilmektedirler. Bunlar anonimizasyon, pertürbasyon, rasgele yanıt, yoğunlaştırma ve kriptolojik çalışmalar olarak sınıflandırılabilir.

- **Anonimizasyon:** Veriyi tüm doğrudan ya da dolaylı tanımlayıcılarından ayırma işlemidir. Bununla birlikte veri, farklı veri setleri ile eşleştirilse dahi, kişi bilgisi ifşa edilemeyecektir. Genelleştirme, gizleme ve permutasyon gibi yöntemleri barındırmaktadır.
- **Rasgele Yanıt:** Verilerin toplanma aşamasında kişisel bilgilerden arındırılması sürecini kapsayan basit bir yöntemdir. Kolayca uygulanabilir ancak bilgi kaybı yüksektir. Birden fazla özelliğe aynı anda uygulanamaz.
- **Pertürbasyon:** Verinin gerçek değerinin bilinmesi istenmediği durumlarda uygulanan yöntemlerdir. Verinin orijinal değerini, bilgi kaybı olmadan değiştirmeyi amaçlar. Genellikle matematiksel bazı yöntemlerle veriyi bozar ve hassas nitelikler için kullanılır.
- **Yoğunlaştırma:** Verileri önceden tanımlanmış boyutlar kullanılarak, verileri daha fazla gruba dağıtarak yoğunlaştırma yapılır. Verilerin orijinali ile değil kopyası ile çalışılır.
- **Kriptolojik:** Hassas niteliklerin şifreleme yöntemleri ile değerlerinin gizlenmesini temel alan yaklaşımdır.

Anlatılan yöntemlerin veri madenciliği yöntemleri açısından kıyaslanması Tablo 2.7 de verilmiştir.

Tablo 2.7: PPDM yöntemlerinin kıyaslanması

PPDM Teknikleri	Metotlar	Veri Madenciliği Yöntemleri		
		Sınıflandırma	Kümeleme	BirlikteKuralı
Anonimleştirme	Genelleştirme, Gizleme, Permutasyon	✓	✓	✓
Pertürbasyon	Gürültü Ekleme, Yer Değiştirme, Mikro-Birleştirme	✓	✓	✓
Rasgele Yanıt	Gürültü Ekleme, Yeniden Örnekleme	✓		
Yoğunlaştırma	Birleştirme	✓		
Kriptolojik	Şifreleme, De- şifreleme			

Bir önceki bölümde anlatılan yöntemler literatürde etkin olarak kullanılarak anonimleştirme çalışmaları yapılmaya devam edilmektedir. X Zhang ve diğ. 2014 yılında yayınladıkları makalede[7] bulut üzerinde MapReduce fonksiyonu kullanarak iki aşamalı TDS yöntemi geliştirmişlerdir. Kavitha S. diğ. tarafından yapılmış olan çalışmada[62] ise yine bulut üzerinde tutulan veriler üzerinde iki aşamalı TDS(Top Down Specialization) yöntemi uygulanmıştır. İki aşamada Map(haritala) ve Reduce(azalt) fonksiyonları önerilen modele göre tasarlanarak performans ve doğruluk oranının artışı desteklenmiştir. Önerilen yöntemin uygulaması için bir yazılım tasarlanmış ve iki aşamalı TDS ile merkezileştirilmiş (centralized) TDS yöntemi kıyaslanarak önerilen modelin önemi vurgulanmıştır. Bir diğer yakın çalışmada ise Xuyun Zhanga ve diğ. bulut üzerindeki verileri kullanarak TDS ve BUG yöntemlerini kullanarak hibrit bir model önermişlerdir[63]. Uğur Sopaoğlu tarafından yayınlanan [64] çalışmada büyük veriler için Apache Spark ortamı üzerinde TDS ve k-anonimlik kullanılarak kimliksizleştirme yapacak yeni bir model önerilmiştir. Adult veri seti kullanılarak önerilen model uygulanmış ve düğüm sayıları artırılarak önerilen modelin karşılaştırılması yapılmıştır.

Sezgisel yöntemlerde literatürde sıkça kullanılan yöntemlerdendir. S Madan ve diğ. Tarafından yayınlanan çalışmada bulut üzerinde tutulan verilerin anonimleştirilmesi için k-anonimlik ve DragonFly algoritması kullanılarak hibrid bir yapı oluşturulmuştur[65]. Tochukwu ve diğ. ise

yayınladıkları bir çalışmada mekânsal bir indeks oluşturma'nın k- anonimlik yöntemi ile çok benzeştiğinden yola çıkılarak, mekânsal indekslemenin anonimleştirme için kullanılabileceğini düşünmüş ve artırılmış anonimleştirme yöntemi önermişlerdir[66]. Önerilen model neticesinde R-ağacı indeksine dayalı anonimleştirme yönteminin diğer yöntemlere göre daha iyi sonuç verdiği sunulmuştur.

2018 yılında yayınlanan bir çalışmada veri anonimleştirme için çok boyutlu k-anonimlik kullanılmıştır. Makalede MDR (mean difference recommending tree) kullanılarak Mondrian yöntemi ile kıyaslanmıştır[67]. F Amiri ve diğ. tarafından yayınlanan çalışmada ise bilgi yayınlamada bazı atakların önüne geçilmesi için k-anonimlik ve β -benzerlik yöntemleri kullanılarak gizliliği kuvvetlendirilmiş hibrit bir yapı önerilmiştir [68].

Can Eyüboğlu ve diğ. tarafından yayınlanan çalışmada [69] literatürde farklı alanlarda sıkça kullanılan kaos teorisine bağlı yeni bir pertürbasyon yöntemi önerilmiştir. K-anonimlik kullanılan bu yöntemde gizlilik açısından oldukça başarılı sonuçlar elde edilmiştir. Bu alanda farklı bir çalışma da Zakariae El Ouazzani ve Hanan El Bakkali tarafından yapılmıştır. Yayınladıkları çalışmada büyük veri üzerinde k-anonimlik de k değerinin baştan belirlenmeden uygulanabildiği dinamik bir yapı sağlanmıştır[22].

2017 yılında Yavuz Canbay, Şeref Sağıroğlu tarafından yayınlanan çalışmada[70] ise anonimleştirme işlemi için Hadoop'a göre hız ve performans açısından daha iyi gördükleri Spark kullanılmıştır. Önerilen modelde HDFS(Hadoop Distributed File System) üzerinde tutulan veriler, Spark üzerinde oluşturulan RDDlere (Resilient Data Distribution) aktarılması ve daha sonra bilinen bir kimliksizleştirme yöntemi uygulanarak tekrar HDFS üzerinde tutulması modellenmiştir. Nandini Prasaad ve çalışma arkadaşları tarafından yayınlanan çalışmada [71], Hadoop üzerinde tutulan büyük veriler üzerinde kimliksizleştirme yapılması için yarı tanımlayıcı(QID) seçimi için yeni bir yöntem önerilmiştir.

2014 yılında yayınlanan çalışmada ise hassas değerlerin çoklu seçimine ve kategorik olarak yakın olmasının daha iyi sonuçlar vereceğini ortaya koyan bir MapReduce modeli önerilmiştir[72]. Chamikara ve diğ. ise P2RoCAI isimli önerdikleri yöntem için durağan veri yerine IoT üzerinden sağladıkları akan veri kullanmışlardır. Makalelerinde veri kimliksizleştirme için pertürbasyon yöntemini kullanmışlardır[41]. Önerilen yöntem veri yoğunlaştırma (data condensation) ve rotasyonel pertürbasyon (rotation perturbation) olarak iki

aşamalıdır. DC aşaması hem akan hem durağan veri üzerinde uygulanabilirken, RP yalnızca durağan veri üzerinde uygulanabilmektedir.

Yapılan çalışmalar yalnızca sezgisel ve yapılandırma tabanlı değildir. Yapay sinir ağları gibi farklı çalışma alanları da anonimleştirme için kullanılabilmektedir. Andrew J. Ve arkadaşlarının yayınladıkları çalışmada bulut üzerinde tutulan veriler için kimliksizleştirme yapılırken Mondrian anonimleştirme tekniği uygulanmıştır [73]. Makalenin öne çıkan diğer yanı ise özellik belirlenmesi için DNN(Deep Neural Network) kullanılmış olmasıdır. 2017 yılında Jesu ve diğ. tarafından önerilen yöntemde her bir hassas değerden tekdüze dağılımına sahip anonimleştirilmiş kümeler elde etmek için bir k en yakın komşu(knn) kümeleme yöntemi kullanılmıştır. Önerilen model gizlilik koruması için kümelemeyi temel alan ve benzerlik saldırısına ve olasılıksal çıkarım saldırısına dayanıklı bir anonimleştirme yöntemidir[74]. 2019 yılında Naixuan Guo ve diğ. Tarafından yayınlanan çalışmada mikro verilerde bir çeşit denklik sınıfının var olduğu ispat edilerek o sınıfı doğal eşdeğer sınıf (natural equivalence class) olarak tanımlanmıştır. Yayınladıkları çalışmada buldukları denklik sınıfı yardımı ile daha hızlı ve başarılı bir anonimleştirme yöntemi önermişlerdir[75].

Geometrik veri pertürbasyonu üzerine yapılmış olan çalışmada [76] yazarlar farklı veri madenciliği yöntemleri ile uyumlu çalışan 3 aşamalı bir model önermişlerdir. Ayrıca çalışma içerisinde çoklu sütun gizliliğiyle ilgili detaylı analizler yayınlamışlardır. Farklı bir çalışmada[77] kişisel sağlık kayıtlarının bulunduğu veri seti geometrik pertürbasyonda Gauss gürültü modeli kullanılarak anonimleştirme modeli geliştirilmiş ve çalıştırılma süreleri açısından AES şifreleme yöntemi ile karşılaştırılmıştır. Yine benzer bir çalışmada gauss gürültü modeli ve sağlık hizmeti veri seti kullanılarak geometrik pertürbasyon yapılmıştır[78]. [79] deki yazarlar ise geometrik veri pertürbasyonu ve ona uygun bir sınıflandırma modeli geliştirmişlerdir. Sınıflandırma ve kümeleme yöntemleri gizlilik korumalı veri madenciliği için önemli konulardır. [80] deki çalışmada k-ortalımalı kümeleme yöntemi kullanılarak kümelenen veriler, küme merkezleri üzerinde farklı açılar ile rotasyon işlemi kullanılarak pertürbasyon yapılmıştır.

Başka bir çalışmada [81] anonimleştirme için 4 boyutlu rotasyon modeli önerilmiştir. Bu modelde veriler ikiye ayrılarak önce xy sonra zw eksenleri üzerinde döndürülerek değerleri değiştirilmiştir. Kablosuz sensor ağlarda iletilmek üzere sensor düğüm üzerinde toplanan verilerin gizliliği de önemli konulardan biridir. Yapılan bir çalışmada[82] bu gizliliği

sağlayabilmek için geometrik veri pertürbasyonu kullanılmıştır. Geometrik veri pertürbasyonu yöntem aşamalar barındırmaktadır. Bu aşamaların sırasını konu edinen bir çalışmada[83] ise rotasyon, gürültü ekleme, ölçekleme gibi adımların sırasını değiştirmenin sonuca etkisi paylaşılmıştır. Geniş çaplı bir başka çalışmada ise geometrik veri pertürbasyonu detaylıca analiz edilmiştir[84]. Yine aynı çalışma içerisinde yazarların verileri ayırma yöntemi ile geometrik yaklaşımı birleştirerek önerdikleri modelden elde ettikleri başarılı sonuçlar paylaşılmıştır. Aynı yazarların farklı bir çalışmasında[85] ise Laplace gürültü yöntemi kullanılarak etkin ve güvenilir bir pertürbasyon yöntemi önerilmiştir.

2.3 ANONİM VERİNİN KULLANILABİLİRLİĞİ

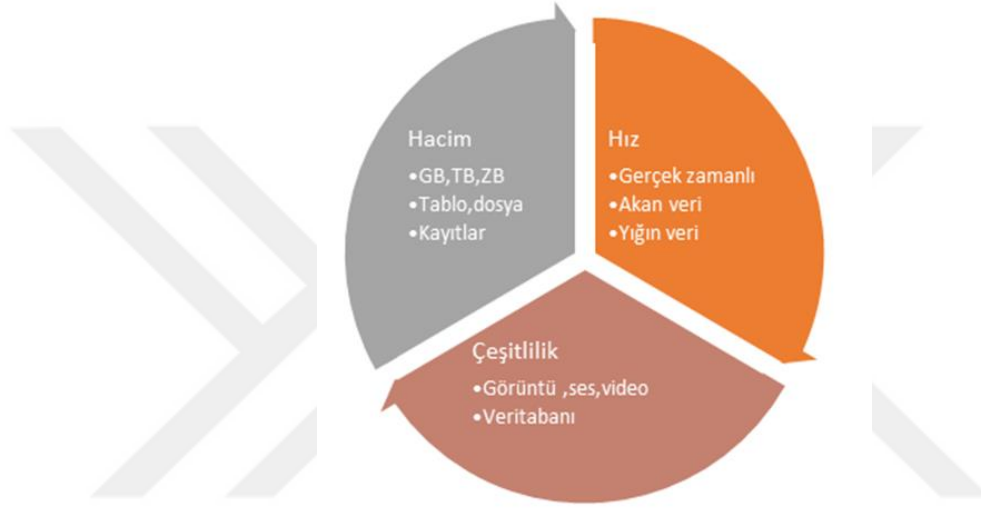
Gizlilik korumalı veri madenciliği, veri üzerinde gizlilik ihmali yapılmadan anlamlı veriler çıkarılarak işlevselliğin korunmasını amaçlar. Gizlilik koruması yüksek seviyede olan bir anonimleştirme algoritması ilk bakışta tercih edilebilir görünmekle birlikte verinin kullanılabilirliğini negatif yönde etkileyebilir. Yani çok iyi anonimleştirilmiş bir veri seti, orijinali ile aynı anlamı taşıyamayacağı için veri seti anonimdir ancak analiz için kullanılamaz durumda olabilir. Bu nedenle verinin kullanılabilirliği ve gizliliği arasındaki dengenin iyi kurulmuş olması gerekmektedir[86,87]. Literatürde yapılan çalışmalar incelendiğinde önerilen anonimleştirme çalışmalarının başarısı bazı metrikler ile ölçülmektedir. Bu metrikler aşağıdaki gibi özetlenmiştir:

- Minimum bozulma (minimal distortion) [88]
- Sınıflandırma metriği (classification metric) [89]
- Bilgi dengeleme metriği (information trade-off metric) [90]
- Bilgi kaybı (information loss) [91]
- Normalize ortalama eşdeğerlik sınıfı boyutu metriği (normalized average equivalence class size metric) [92]
- Atak dayanıklılığı (attack resistance) [93]

2.4 DAĞITIK VERİ ORTAMLARI

Büyük verinin adından anlaşılacağı üzere en önemli ve en belirgin özelliği klasik depolama yöntemleri ile saklanamayacak hacme sahip olmasıdır. Ancak büyük veriyi büyük kılan şeyin

yalnızca hacmi olmadığı araştırmacılar tarafından öne sürülmüştür. Böylece büyük verinin 3V'si olarak adlandırılan hacim(volume), hız(velocity) ve çeşitlilik(variety) kavramlar ile büyük veri tanımlanır olmuştur. Daha sonra bazı araştırmacılar 3V'nin tanım için yeterli olmadığını, büyük veriyi üç özelliği haricinde değer(value) ve gerçeklik(veracity) ile de tanımlamak gerektiğini öne sürmüşlerdir. Hatta görselleştirilebilme (visualization) ve değişkenlik(variability) ile 7V olarak tanımlamalarda literatürde yaygın olarak görülmektedir[94]. Farklı yaygın görüşler olmasına rağmen hepsi temel olarak verinin 3V özelliğini almaktadır.



Şekil 2.3: Verinin 3V gösterimi

Verinin 3V'si olarak adlandırılan bu özelliklerden ilki hacim verinin boyutunu ifade etmektedir. Çeşitlilik ise farklı veri kaynaklarından gelen verinin farklı tipte olmasını belirtmektedir. Örneğin; ses, video, metin... vs. gibi. Hız ise verinin boyutundaki artışı ifade etmektedir. Farklı kaynaklardan yüksek hacimde ve farklı tipte üretilen bu verileri işlemek için var olan teknoloji ve yöntemlerin yetersiz kalması, yeni oluşum ihtiyacı doğurmuştur. Bu sebepten dolayı klasik hesaplama ve saklama yöntemleri yerine dağıtık dosya saklama ve hesaplama sistemlerinin kullanımı yaygınlaşmıştır.

2005 yılında Doug Cutting ve Michael J. Cafarella tarafından Nutch arama motoruna büyük verileri işlemek için kullanılan bir hesaplama yöntemi olan map-reduce eklemek için yapılan çalışmalar sonucunda ortaya Hadoop ürünü çıkmıştır ve 2006 yılında Apache projeyi satın alarak geliştirmeye devam etmiştir. Apache Hadoop büyük hacimli veri setlerinin sunucu

kümeleri üzerinde, belirli programlama modelleriyle birlikte dağıtık bir şekilde işlenmesine olanak sağlayan, açık kaynak kodlu, ölçeklenebilir, hata dayanıklılığı olan bir yazılım kütüphanesidir(framework). Büyük hacimli veriyi makul sürede ve maliyette işlemek ve depolamak için kullanılır[95]. Yahoo, Facebook, Amazon, Netflix gibi büyük şirketler Hadoop ekosistemini (Şekil 2.4) kullanmaktadır. Hadoop sisteminin avantajları:

- Açık kaynak kodludur, yüksek maliyetler gerektirmez.
- Basit sunucular üzerinde çalıştığı için donanım arızalarının olabileceği düşünülerek hata korumalı bir yapıda tasarlanmıştır. Herhangi bir düğüm arızalandığı anda hemen tespit edilerek yerine bir başka düğüm devreye girmektedir.
- Ölçeklenebilir bir sistemdir , ihtiyaç halinde yeni düğümler eklenebilir.
- Aynı anda birçok kullanıcıya erişim imkânı sunar. Yazılımcılar için de yine aynı anda birden çok yazılımcının kod yazabilmesine imkân verir.

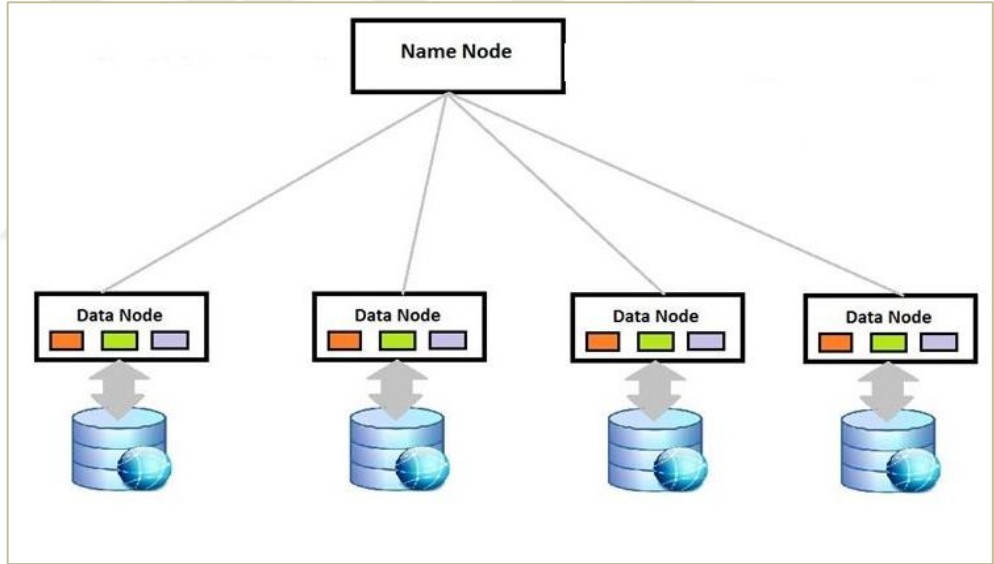


Şekil 2.4: Hadoop ekosistem gösterimi

HDFS(Hadoop Distributed File System): Hadoop' un en temel bileşenlerinden biri olan HDFS çok büyük hacimli verileri sıradan sunuculardan oluşan kümeler üzerinde depolayan ve makul cevaplama süresi içinde işleyebilen dağıtık bir dosyalama sistemidir. HDFS sunucuları üzerine kopyalanan veriler aynı anda ve hızlı bir şekilde erişilebilir olurlar. HDFS de büyük veriler küçük dosya blokları olarak saklanır.

HDFS'in çalışma mantığı düzenleme ve silme işlemlerini gerçekleştirmek değil, bir defa veriyi yazdıktan sonra defalarca hızlı bir şekilde okuma yapmaktır. Özel bir marka sunucu ya da donanım gerektirmez. Anında cevap dönülmesi beklenen uygulamalar için uygun değildir.

HDFS saklama ve işleme sürecini ana düğüm(namenode) ve işçi düğümler (datanode) ile yönetilir (Şekil 2.5). Ana düğümler bloklar halinde tutulan verilerin, dağıtılması, dağıtıldığı yerlerin bilinmesi, gerektiğinde silinmesi gibi durumlardan sorumlu olan makinedir. İşçi düğümler ise kendine verilen veri bloklarını saklamaktan sorumlu olan sıradan sunuculardır. Herhangi bir işçi düğüm üzerinde bir hata olması durumunda, problem hemen ana düğüm tarafından ele alınarak, hatalı ya da bozuk olan işçi düğüm içerisindeki veriler diğer işçi düğümler üzerine kopyalanır . Böylelikle herhangi bir veri kaybı söz konusu olmaz.

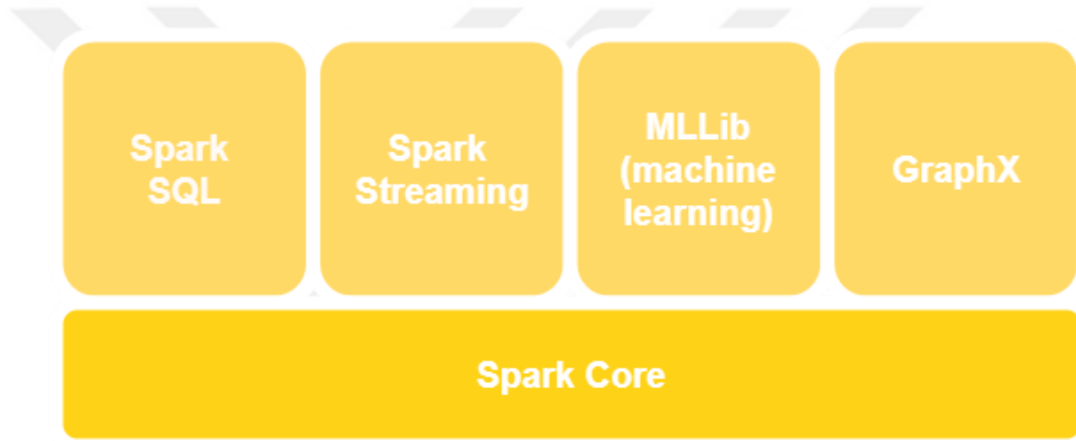


Şekil 2.5: Ana düğüm ve işçi düğümlerin gösterimi

SPARK: Hadoop ile uyumlu çalışan ancak Hadoop zorunluluğu olmayan, mapreduce alternatifi açık kaynak kodu veri işleme Apache projesidir. Veri okuma, yazma, programlama, SQL, makine öğrenmesi, akış hizmeti(streaming), graf hesaplama[88] gibi işlemleri yapabilen, içerisinde paralel veri işleme yapabilmek için gerekli kütüphaneleri barındıran büyük çaplı bir hesaplama motorudur. Hadoop MapReduce diski kullanırken, Spark RAM kullanarak verileri işlemektedir. Bu özelliğinden dolayı Hadoop MapReduce' a göre 100 kat daha hızlı

çalışabilmektedir. Bir diğer önemli özelliği ise farklı programlama dilleri desteğidir. Spark içerisinde Java, Scala, Python ve R dili ile kodlama yapılabilmektedir.

Spark çekirdeği altında farklı bileşenler ile işlem yapılması mümkündür (Şekil 2.6). Bunlar Spark SQL, Spark MLLib, Spark Streaming ve Spark GraphXdir. Spark SQL ile büyük veriler yapısal hale getirildikten sonra klasik SQL sorguları çalıştırılabilmektedir. Spark MLLib ise durağan veriler üzerinde makine öğrenmesi uygulamalarını çalıştırabilecek alt yapıyı sağlamaktadır. Akan verileri işlemek için Spark Streaming kullanılırken, GraphX ile graf analizi yapılmaktadır.



Şekil 2.6: Spark bileşenleri gösterimi

Spark, cluster mod ve local mod olarak iki türlü çalıştırılabilmektedir. Cluster mod çalışma modeline göre 3 farklı tipte çalışabilmektedir. Eğer herhangi bir cluster yöneticisi kullanılmayacaksa Spark Standalone Cluster , yönetici kullanılacaksa Apache Mesos ya da YARN kullanılabilmektedir.

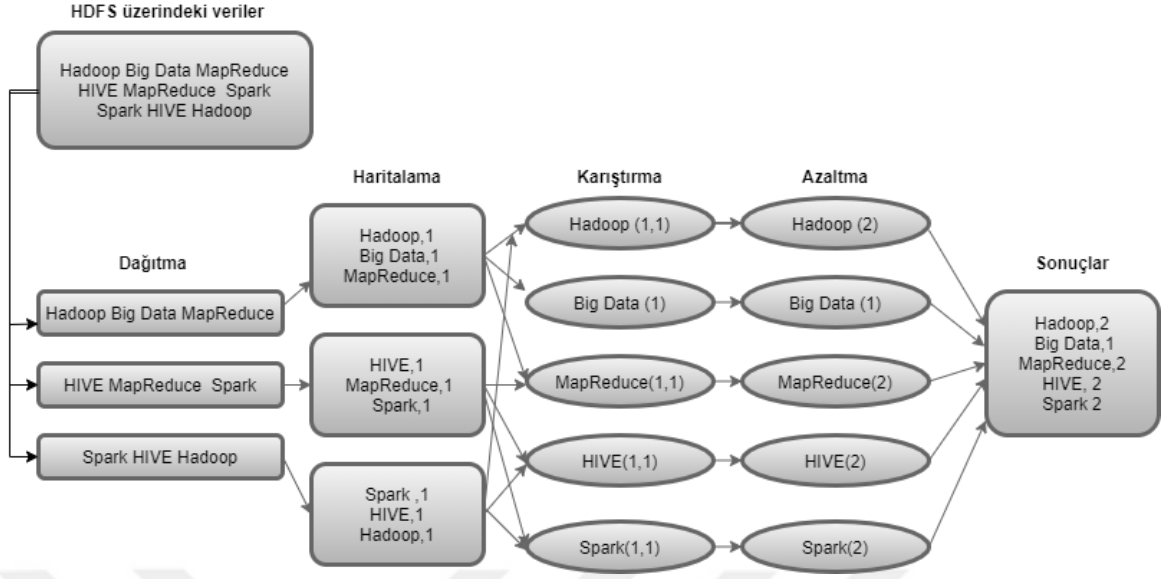
Spark genel olarak RDD'lerden (Resilient Distributed Datasets - Esnek Dağıtılmış Veri Kümeleri) oluşmaktadır. İşlenecek olan nesne yığınları parçalanarak kümeler(cluster) üzerine dağıtılır. Bu yapısı ile paralel veri işleme olanağı sağlayarak hızlı ve esnek bir yapı sağlamaktadır.

Bu çalışmanın motivasyonu Apache Spark'ın veriyi hızlı işleyebilme özelliğinde faydalanarak geometrik veri pertürbasyonu tabanlı yeni bir anonimleştirme modeli oluşturmaktır.

YARN: YARN(Yet Another Resource Negotiator) Hadoop kaynaklarını yöneten bileşendir. YARN kaynak yönetim sürecini her bir sunucu kümesi için bir kaynak yönetici(Resource Manager), bulunan her bir düğüm için düğüm yönetici(Node Manager) ve her uygulama için ise Application Master bileşenleri ile gerçekleştirir. Kaynak yönetici sunucu kümesi üzerinde iş ve uygulama önceliklerine göre kaynak ayırma işlemini gerçekleştirir. Düğüm yöneticisi, her bir işçi düğüm üzerinde kaynak takibi yaparak belli zaman aralıklarında kaynak yöneticisini bilgilendirir. Application Master ise işçi düğüm üzerinde her bir uygulama için bulunur. Uygulamanın çalışabilmesi için yeteri kadar kaynak bulunması durumunu kontrol eder ve düğüm yöneticisi ile kaynak yöneticisi haberleşmesini koordine eder.

Kaynak yönetimini ayrı bir süreç olarak ele alması dolayısı ile YARN, Hadoop HDFS üzerinde tutulan verilerin MapReduce haricinde farklı şekilde işlenebilmesine olanak sağlamıştır.

MAPREDUCE: Hadoop ekosistemi üzerinde bulunan verileri kısa sürede işlemeye yarayan bir programlama modeli ve bileşendir. Google'ın verileri indeksleme ihtiyacına karşılık ortaya çıkmıştır. Map(parçala) ve reduce (hesapla) olmak üzere iki aşamadan oluşur. Büyük sunucu kümeleri üzerinde dağıtık ve paralel çalışabilme özelliği sağlar. MapReduce çalışma mantığı aşağıdaki şu şekilde özetlenebilir(Şekil 2.7): HDFS üzerinde tutulan veriler, parçalara ayrılarak sunucular üzerine gönderilir (map) . Her bir sunucu üzerindeki değerler ayrı ayrı hesaplanır. Ayrı hesaplanan bu değerler daha sonra özet bir veri olarak toplanır(reduce). Bu şekilde büyük olan verinin paralel olarak işlenmesi sağlanarak hızlı bir sonuç elde edilmiş olur.

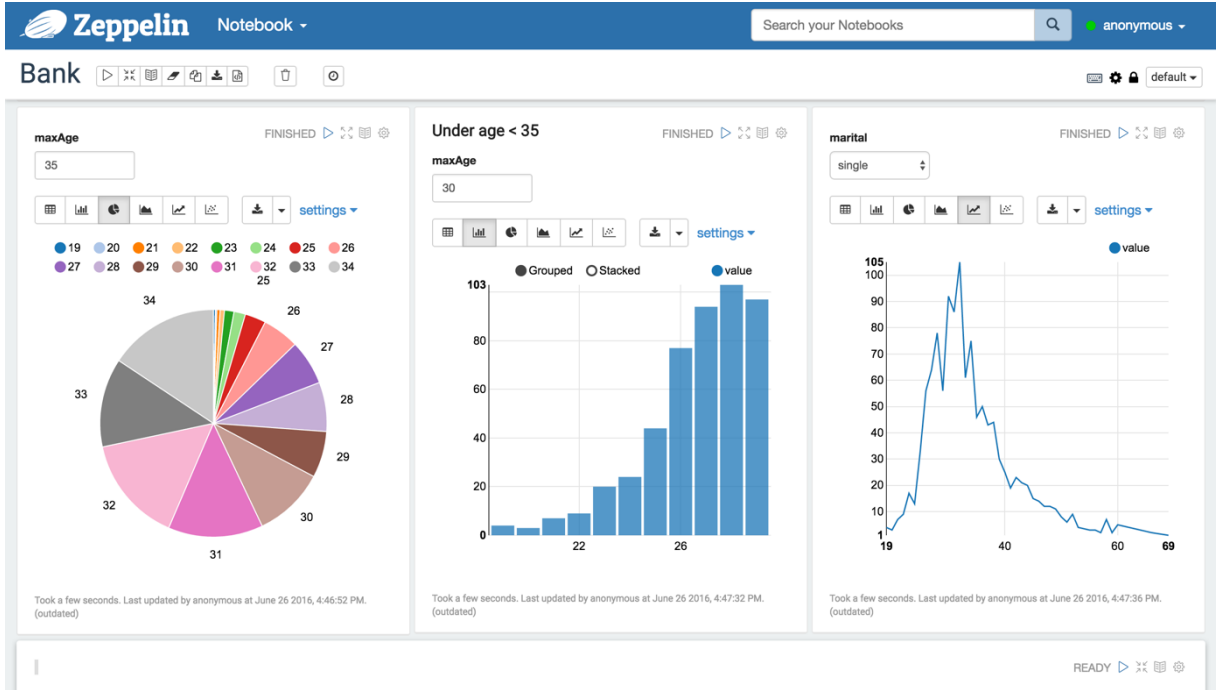


Şekil 2.7: MapReduce hesaplama süreci

HIVE: Büyük veri ortamına geçiş sonrasında klasik programlama yetenekleri haricinde farklı bir yapı ile karşılaşılmıştır. MapReduce ile işlenen HDFS üzerindeki verilerden basit bir sorgu yapabilmek için bile satılınca koda ihtiyaç vardır. Bu nedenle kişilerin var olan klasik SQL sorgulama yeteneklerini hızlı ve etkin bir şekilde kullanmasını sağlayacak bir yapı ihtiyacı oluşmuştur. Hadoop gibi dağıtık dosya sistemleri üzerinde tutulan büyük hacimli veri setlerini okuma, yazma yönetme işlemlerini kısa sürede gerçekleştirebilmek için oluşturulmuş, klasik SQL sorgularına benzeyen sorgularla bu işlemleri yapabilmeyi sağlayan bir Hadoop projesi olarak HIVE geliştirilmiştir. SQL seviyesinde sorgular için kullanımı idealdir ancak Büyük hacimli verilerle uğraştığı için kayıt seviyesinde güncelle(update) ve sil(delete) işlemlerini barındırmayan HiveQL sorgu dilini kullanmaktadır.

MAHOUT: Açık kaynak kodlu makine öğrenmesi kütüphanesidir. Özellikle reklam şirketlerinin ihtiyaç duyduğu büyük verilen analiz edilerek anlamlı verilerin çıkarılması işlemlerinde kullanılan önemli bir projedir.

ZEPELİN: SQL, Scala, Python, R vb programlama dili veya programlama süreçleri için kullanılabilen web tabanlı bir not defteri uygulamasıdır. Apache Spark, Apache Flink, Python, R, JDBC, Markdown ve Shell gibi birçok yorumlayıcıyı desteklemektedir.



Şekil 2.8: Zeppelin not defteri örnek ekran gösterimi

3 MALZEME VE YÖNTEM

Bu bölümde tez çalışması kapsamında önerilen rasgele sayı tabanlı gizlilik koruma modelinin detayları ile çalışmanın performans analizleri yapılırken kullanılan veri setleri, veri seti bölümlendirme yöntemleri, sınıflandırma yöntemleri ve yöntemin dayanıklılığının tespitinde kullanılan atak yöntemlerinden bahsedilecektir.

3.1 GEOMETRİK PERTÜRBASYON

Yeniden yapılandırma tabanlı anonimleştirme yöntemlerinden biri olan Geometrik Veri Pertürbasyonu, verilerin belli miktarlarda döndürme, öteleme ve gürültü eklenmesi ile elden edilen yeni değerlerin orijinal veriler ile yer değiştirilerek kimliksizleştirme yapan yöntemlerden biridir. Formül üzerinde RX-Rotasyon, T-Dönüşüm, Δ - Gürültü Ekleme adımları olarak işlenmektedir.

$$G(X) = RX + T + \Delta \quad (3.1)$$

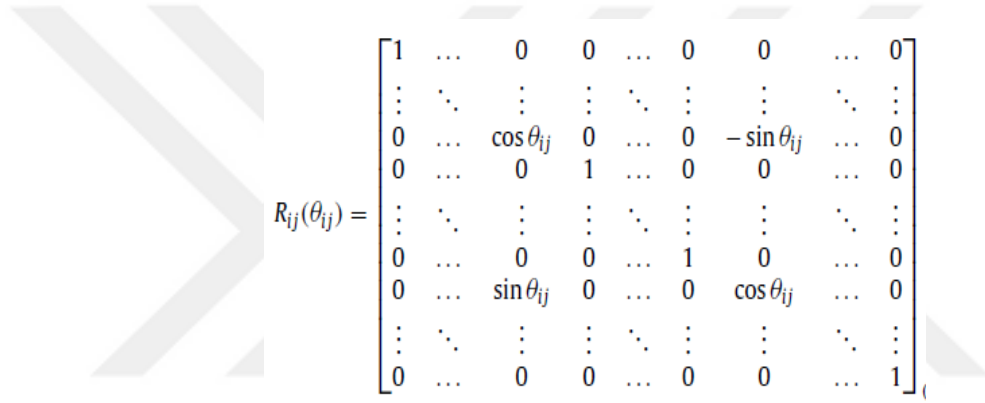
3.1.1 Çarpımsal Dönüşüm

Veri rotasyonu D veri seti üzerinde bulunan hassas verilerin çarpımsal dönüşüm yöntemiyle pertürbasyonun sağlanmasıdır. Rotasyonel pertürbasyon $G(X) = RX$ olarak tanımlanır. Nümerik hassas değerlerden oluşan veri seti dönüşüm için bir matris ile çarpılır. Bu matris rasgele rotasyon matrisi[50] ya da rasgele projeksiyon[55] matrisi olabilir. Burada kullanılan rotasyon matrisi özel bir matristir ve bazı kurallara göre oluşturulması gerekmektedir[96].

Rasgele projeksiyon matrisi ise $R = \sqrt{\frac{d}{k}} R_0$ olmak üzere $R_{k \times d}$ olarak tanımlanmaktadır. Burada R_0 matrisi rasgele tanımlanmış satır vektörleri ortogonal olan bir matristir[97]. Rotasyon matrisi verilerin dönüşümü sonrasında mesafeyi tam olarak korurken, projeksiyon matrisi yaklaşık olarak koruma yapar. Bu nedenle rasgele rotasyon matrisi daha fazla tercih edilmektedir. Rotasyon matrisi N boyutlu koordinat sistemi üzerinde nokta veya noktaların belirlenen bir eksen üzerinde belirlenen bir açı miktarı kadar döndürülmesi için kullanılan matristir (Formül 1). Burada kullanılan rotasyon matrisinin bazı kurallar dahilinde oluşturulmaktadır (Algoritma 1). i satırları j ise sütunları belirtmek üzere:

- R matrisi üzerindeki tüm satır ve sütunlar ortogonal olmalı
- Herhangi bir j sütunu için $\sum_{i=1}^d r_{ij}^2 = 1$
- Herhangi iki sütun j ve k için $\sum_{i=1}^d r_{ij}r_{ik} = 0$
- $R^T R = R R^T = I$ (birim matris)
- Rotasyon matrisinin herhangi bir sütununun veya satırın sırasının yerinin değiştirilmesi sonucu da bir rotasyon matrisidir.

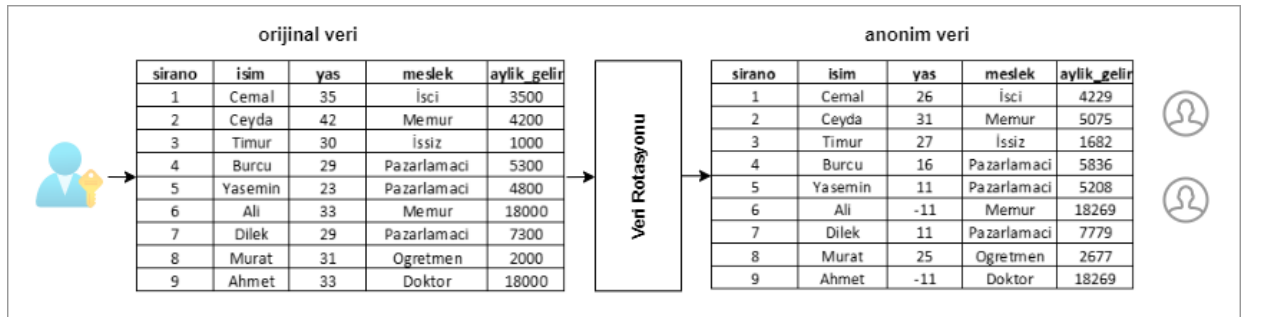
Buna göre oluşturulan N boyutlu bir rotasyon matrisi Şekil 3.1 de gösterilmiştir. Bu kurallara ek olarak Haar Dağılımı kullanılarak da rotasyon matrisi oluşturulabilir[98].



$$R_{ij}(\theta_{ij}) = \begin{bmatrix} 1 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \cos \theta_{ij} & 0 & \dots & 0 & -\sin \theta_{ij} & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & \sin \theta_{ij} & 0 & \dots & 0 & \cos \theta_{ij} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 1 \end{bmatrix}$$

Şekil 3.1: Rotasyon matrisi gösterimi

Şekil 3.2 üzerinde örnek veri setinin yas ve aylık_gelir hassas niteliklerinin saat yönünün $\alpha=13,7$ derece olacak şekilde döndürülme sonucu izlenmektedir.



Şekil 3.2: Çarpımsal dönüşüm örneği

Algoritma 1: Rotasyon Matrisi Üretme

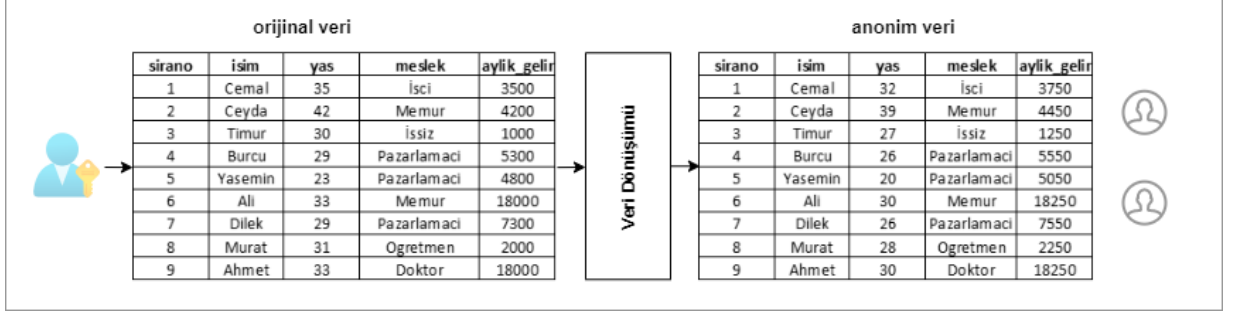
Girdi: n <-- veri setinden seçilen nitelik sayısı, α <-- döndürme açısı

Çıktı: R <-- çok boyutlu rotasyon matrisi

- 1: $V = \{1, 2, 3, \dots, n\}$
- 2: $C = \{(i, j) \mid i, j \in V \text{ ve } i \neq j\}$
- 3: $I_n = n$ boyutlu birim matris
- 4: $N = \binom{n}{2}$ kombinasyon
- 5: $I_3 = I_n$
- 6: **for** $k=1$ to N **do**
- 7: $A = \{(i_k, i_k), (j_k, i_k), (i_k, j_k), (j_k, j_k)\} \setminus \{(i_k, j_k)\}$ (C içinde k . Küme olmak üzere)
- 8: $I_2 = I_n$
- 9: $I_2(i_{k1}, j_{k1}) = \cos(\alpha)$ (i_{k1}, j_{k1}) A içinde 1. küme
- 10: $I_2(i_{k2}, j_{k2}) = \sin(\alpha)$ (i_{k2}, j_{k2}) A içinde 2. küme
- 11: $I_2(i_{k3}, j_{k3}) = -\sin(\alpha)$ (i_{k3}, j_{k3}) A içinde 3. küme
- 12: $I_2(i_{k4}, j_{k4}) = \cos(\alpha)$ (i_{k4}, j_{k4}) A içinde 4. küme
- 13: $I_3 = I_3 \times I_2$
- 14: **end for**
- 15: $M = I_3$

3.1.2 Öteleme Dönüşüm

Veri dönüşümü D veri seti üzerinde bulunan hassas verilerin bir miktar ötelenmesiyle pertürbasyonun sağlanmasıdır. Nümerik hassas değerlerden oluşan veri seti dönüşüm için bir matris ile toplanır. Bu matris veri setindeki her aynı nitelik için aynı pozitif ya da negatif değerlerden oluşan bir matristir. Şekil 3.4 üzerinde örnek veri setinin yas ve aylık_gelir hassas niteliklerine eklenen (-3,250) değerleri sonucu izlenmektedir.



Şekil 3.3: Öteleme dönüşüm örneği

3.1.3 Gürültü Ekleme

Klasik bir gürültü ekleme yöntemi veri seti üzerindeki değerler üzerine rasgele üretilmiş olan değerlerin eklenmesi şeklinde açıklanır. Orijinal veri seti üzerindeki herhangi bir nümerik alanı temsil eden $\mathbf{X}$ sütunu ($x_1, x_2, x_3, \dots, x_n$) değerlerinden oluşmaktadır. Gürültü matrisi $\mathbf{R}$ ile orijinal matrisin toplanması sonucunda seçilen bu sütuna gürültü eklenerek pertürbasyon işlemi yapılmış $\mathbf{Y}$ sütunu $\mathbf{Y}=\mathbf{X}+\mathbf{R}$ elde edilmiş olur. Sonuç olarak elde edilen sütun ($x_1 + r_1, x_2 + r_2, \dots, x_n + r_n$) olarak ifade edilir.

İncelenen çalışmalar sonucunda geometrik veri pertürbasyonunda bu zamana kadar benzer rotasyon ve dönüşüm yöntemleri kullanılmıştır. Gürültü modeli olarak bazı çalışmalarda rasgele gürültü eklenmiştir. Ancak çalışmaların çoğunda veriye yani verinin standart sapması ve ortalamasını kullanması sebebiyle daha etkin sonuçlar alınan Gauss gürültüsü kullanılmıştır. Gauss her seferinde veriye bağlı olarak rasgele gürültü değeri üretmektedir.

Buradan yola çıkarak rasgelelik kavramı üzerine yoğunlaşmıştır. Rasgelelik istatistik, oyun teorisi, benzetim, nümerik analiz, eğlence gibi birçok alan için ihtiyaç duyulan bir kavramdır. Özellikle kriptolojik uygulamalar çoğunlukla rasgele üretilen sayılarla işlem yapmaktadır. Oturum anahtarları, imza anahtarları ve parametreleri, kimlik doğrulama protokolleri, geçici anahtarlar, sıfır bilgi ispatı, blok şifreler için başlangıç vektörleri ve yan kanal saldırılarına karşı koruma önlemleri olan körleştirme ve maskeleye işlemleri rasgele sayılara gereksinim duyan kriptolojik uygulamalardan sadece birkaçıdır[99,100].

3.1.4 Çoklu Sütun Gizliliği

Diğer anonimleştirme yöntemlerinin aksine rotasyonel pertürbasyon yönteminde kimliksizleştirilmesi istenen tüm sütunlar birlikte ele alınır, ayrı ayrı işlem yapılmaz . Bu nedenle gizliliğin ölçülmesi de birlikte ele alınmalıdır. Veri setimiz üzerindeki nümerik verileri barındıran bir matris olarak ele alalım. Y ise bu matris üzerindeki herhangi bir sütunu, Y' ise bu sütunun pertürbasyon sonrası durumunu ifade etsin. $D = Y' - Y$ olarak ifade ettiğimiz sütunun ilk ve son hali arasındaki fark, $E[D]$ farkın standart sapması, $Var[D]$ ya da VoD farkların varyansı olacaktır. Çoklu sütunlar üzerinde çalıştığımız için standart sapma ve varyans değerleri bizim için anlamlı veriler olmayacaktır.

Veri seti üzerinde bulunan kendi tanımlı olduğu alan içerisinde değer almaktadır. Örneğin maaş verisi 5000-10000 arasında değişirken, yaş aralığı 18-65 arasında değişmektedir. Geometrik veri pertürbasyonunda yukarıda da anlatıldığı üzere anonimleştirme yapılmak istenen sütunlar birlikte ele alınır. Yaş ve maaş gibi çok farklı değer aralıklarına sahip olan bu iki alanı rotasyon ile döndürme işlemini yapmak anlamsız sonuçlar alınmasına sebep olmaktadır. Bu nedenle verilerin değişimlerini anlamlı bir şekilde ifade edebilmemiz için bu alanlar üzerinde normalizasyon uygulanması gerekmektedir. Normalizasyonun veri setindeki sayısal değişkenlerin değerlerini, değerler aralığındaki farklılıkları bozmadan ortak bir ölçeğe dönüştürmek için yapılır. Normalizasyon sonrasında tüm veriler 0-1 aralığında olacaktır. Bunun için aşağıdaki formül (3.2) her sütun için ayrı ayrı uygulanır.

$$Y_{si} = \frac{Y_i - \min(Y_i)}{\max(Y_i) - \min(Y_i)} \quad (3.2)$$

Formülün her bir sütuna uygulanması sonucunda her Y sütunu $[0,1]$ arası değere dönüştürülmüş olur. Normalizasyon uygulanan sütun Y_{si} olarak düşünülürse Y'_{si} bu sütunun pertürbasyon sonrası olarak adlandırılır. Buna göre verilerin normalizasyon öncesi ve sonrası arasındaki fark olan $D' = Y'_{si} - Y_{si}$ farkının standart sapması gizlilik seviyesi olarak kabul edilmektedir. Her sütunun veri seti içerisindeki taşıdığı önem farklı olduğu için, burada bir ağırlık parametresi olarak w_i ve yine her sütunun gizlilik seviyesi p_i olarak belirlenmiştir. Buna göre minimum

gizlilik garantisi $\emptyset_1$ ve ortalama gizlilik garantisi ise $\emptyset_2$ olarak ařağıdaki řekilde ifade edilmiřtir(3.3).

$$\emptyset_1 = \min \left\{ \frac{p_i}{w_i} \right\} \quad \emptyset_2 = \frac{1}{m} \sum_{i=1}^m \frac{p_i}{w_i} \quad (3.3)$$

Burada sřtunların ağırlık derecelerinin aynı olduğı kabul edilirse třm w_i deęerleri 1 olacağı iin yukarı formřller ařağıdaki gibi gřncellenecektir.

$$\emptyset_1 = \min \{p_i\} \quad \emptyset_2 = \frac{1}{m} \sum_{i=1}^m p_i \quad (3.4)$$

3.2 RASGELE SAYI ŐRETELERİ VE LİNEER KONGRÜANS MODEL

Rasgelelik sayısal analiz, oyun teorisi gibi bilgisayar teknolojilerinde sıklıkla kullanılan kavramlardan biridir. Őzellikle kriptolojik uygulamalarda anahtar řretimi ve dağıtımında, kimlik doęrulama mekanizmalarında , řifre řretme iřlemlerinde rasgele sayılara ihtiya duyulmaktadır. Rasgele sayı řreteleri, ıktı olarak rasgele sayı veren sayısal veya fiziksel kaynaklardan řretilmiř olan sayılardır. İstatiksel olarak birbirinden bağımsızlardır ve řretilen sayıların aralarında bir korelasyon yoktur. Bu rasgele sayıların řretilbilmesi iin RSU(Rasgele Sayı Őreteleri) denilen bir kavram řretilmiřtir. Rasgele sayı řreteleri kullandıkları mekanizmaya gřre ikiye ayrılır.

- **Donanım Rassal Sayı Őreteci** (Gerek Rassal Sayı Őreteci, GRSŐ) : Isıl gřrřltř, fotoelektrik etkisi ,termal gřrřltř gibi bilgisayarın fiziksel ۆzelliklerinden faydalananarak, bilgisayar programı kullanmadan rasgele sayı řretme yřntemidir. Bu sistemler kararsız dinamik sistemlerdir ve davranıřları kaos teorisi ile aıklanmaktadır. Bu yřntemle řretilen sayılar tamamen rasgele olduğı yani bir kurala gřre řretilmediğı iin kararsızlardır ve kestirilemezlerdir.

Bu sistemler yazılım ve donanım tabanlı olarak tasarlanabilmektedir. Donanım tabanlı sistemler giriř parametresi olarak belirlenen donanım biriminden gelen sinyali baz alırlar. Yazılım tabanlı sistemler ise Mouse hareketleri ya da klavye basma sřresi gibi parametreleri kullanarak rasgele sayı řretmektedir.

- **Sözde Rassal (rastgele) Sayı Üreteci** (Pseudorandom Number Generator, PRNG) : Algoritmik bir mantık ile bir başlangıç koşulundan başlayarak rasgele sayı üretme yöntemidir. Üretilen sayılar arasında bir ilişki kurulamayacak şekildedir ancak bu şekilde üretilen sayılar belirli bir kurala göre oluşturulduğu için tamamen rasgele olmamakta ve bu yüzden sözde rassal sayılar olarak bilinmektedir. Ancak kolaylık, hız ve masrafsız olması özellikleri de düşünülünce çokça kullanılan yöntemlerdendir.

En yaygın olarak kullanılan modeli Lineer Kongrüans Üreteçlerdir. Lineer Kongrüans Üreteçleri saat aritmetiği de denilen sisteme benzer çalışmaktadır.

$m > 0$ bir doğal sayı, $0 \leq a < m$, $0 \leq c < m$ ve $0 \leq X_0 < m$ olmak üzere $X_0 \in \{1, 2, \dots, m-1\}$ olacak şekilde başlangıç değeri seçilir ve belirlenen tekrar sayısı kadar Formül (3.5) uygulandıktan sonra rasgele sayı değeri ortaya çıkmış olur.

$$X_{i+1} = aX_i + c \pmod{m} \quad (3.5)$$

Rasgele sayı üreteçleri, bir başlangıç değerini giriş parametresi olarak baz aldığı ve tamamen rasgele sayılar üretmediği için kullanım alanı kısıtlıdır. Ancak kullanım açısından hem düşük maliyetli hem de hızlı çalışmaktadırlar. Ancak aynı girdi değeri için aynı sonucu üretebilme özelliğinden dolayı yaygın olarak kullanılmaktadır.

3.3 VERİ SETLERİ VE BÖLÜMLEME

Önerilen modelin performans sonuçları 6 farklı veri seti üzerinde uygulanmıştır. Kullanılan bu veri setleri veri kimliksizleştirme çalışmalarında yaygın olarak kullanılan farklı boyutlardaki veri setlerinden belirlenmiştir. Veri setleri üzerinde anonimleştirme yapılırken, nümerik ve hassas olan değerler belirlenmiştir. Kullanılan bu veri setlerindeki veriler ve hangi nitelikler üzerinde kimliksizleştirme yapıldığı bilgisi Tablo 3.1 üzerinde paylaşılmıştır.

Önerilen model nümerik hassas verileri kullanarak çalışmaktadır. Bu neden içerisinde birçok alan olan veri setlerinden hassas niteliklerin seçilmesi gerekmektedir. Tablo 3.1 incelendiğinde son sütunda kullanılan özellikleri gösteren bölümde her veri seti üzerinden anonimleştirmek üzere seçilen alan isimleri belirtilmiştir. Kullanılan veri tabanlarının dört tanesi California Üniversitesi-Irvine, Makine Öğrenmesi Deposunda (Machine Learning Repository), iki tanesi ise makine öğrenmesi yöntemleri için kullanılan veri setlerinin bulunduğu OpenML(Open

Machine Learning) üzerinde online olarak erişime açıktır. Tablonun alt kısmındaki bilgilendirme satırlarında ise bu veri tabanlarına erişim sağlanan linkler bulunmaktadır.

Tablo 3.1: Kullanılan veri setleri ve özellikleri

Veri Setleri	Kayıt Sayısı	Nitelik Sayısı	Sınıf Sayısı	Kullanılan Özellikler
Iris ¹	150	4	3	<i>SepalLength</i> , <i>SepalWidth</i> , <i>PetalLength</i> , <i>PetalWidth</i>
Heart StatLog ²	270	14	2	<i>resting_blood_pressure</i> , <i>serum_cholesterol</i>
Wine Quality-White ³	4898	12	7	<i>volatile acidity</i> , <i>citric acid</i> , <i>residual sugar</i> , <i>chlorides</i>
Fried ⁴	40768	11	2	<i>x1</i> , <i>x2</i> , <i>x3</i> , <i>x4</i> , <i>x5</i> , <i>x6</i> , <i>x7</i> , <i>x8</i> , <i>x9</i> , <i>x10</i>
Bank Management ⁵	45212	17	2	<i>age</i> , <i>balance</i>
Electricity ⁶	45313	9	2	<i>nswprice</i> , <i>nswdemand</i>

¹ <https://archive.ics.uci.edu/ml/datasets/iris>

² [https://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](https://archive.ics.uci.edu/ml/datasets/statlog+(heart))

³ <https://archive.ics.uci.edu/ml/datasets/wine+quality>

⁴ <https://www.openml.org/d/901>

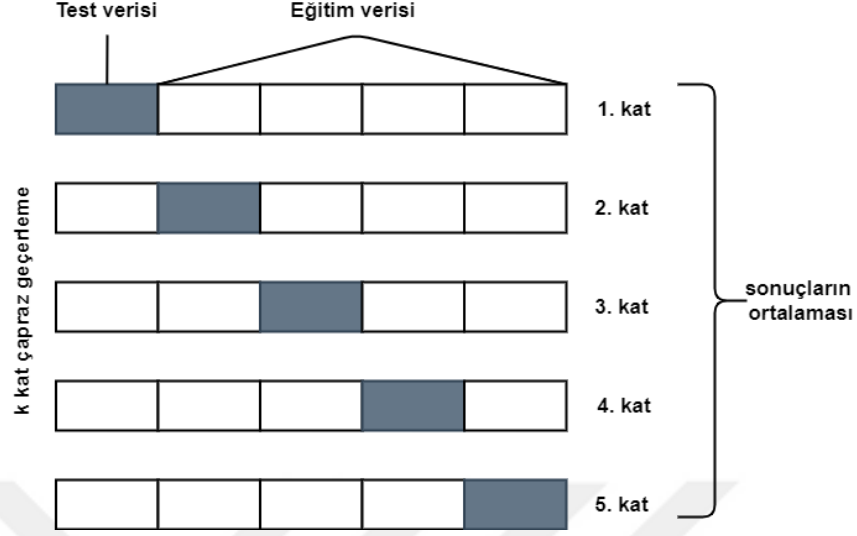
⁵ <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

⁶ <https://www.openml.org/d/151>

Yukarıda belirtilen veri setleri ile modelin sınıflandırma doğruluğu performansı değerlendirilirken iki farklı şekilde değerlendirme yapılmıştır. Birincisinde modelin eğitim ve testi için veri setinin tamamı kullanırken, ikinci olarak k- kat çapraz geçerleme(k-fold cross validation) kullanılmıştır.

k- kat çapraz geçerleme yönteminde veri seti eşit boyutlu k adet rasgele alt kümelere bölünür. Sınıflandırma modeli eğitildikten sonra k defa test edilir. Her test ediliş basamağında k-1 kat eğitim için, geriye kalan 1 kat ise test için kullanılır. Sonuç ise tekrarlanan bu sonuçların ortalaması alınarak hesaplanır(3.6)

$$\text{Doğruluk Oranı} = \frac{1}{k} \sum_{i=1}^k A_i \quad (3.6)$$



Şekil 3.4: 5 kat çapraz geçerleme örnek gösterimi

Şekil 3.4 üzerinde 5 kat çapraz geçerleme modeli gösterilmiştir. Çalışmalarda genel olarak 10 kat çapraz geçerleme kullanılmaktadır. Bu çalışma içerisinde ise 2, 5 ve 10 kat çapraz doğrulama kullanılarak önerilen modelin doğruluk oranı sınanmıştır.

3.4 PERFORMANS METRİKLERİ

Bu bölümde önerilen modelin sınanması için kullanılan sınıflandırma doğruluğu, atak dayanıklılığı, F- ölçütü, çalışma zamanı ve Friedman rank testi metrikleri açıklanacaktır.

3.4.1 Sınıflandırma Doğruluğu

Sınıflandırma doğruluğu, sınıflandırma sonrasında doğru sınıflara yerleştirilmiş olan veri kümelerinin oranını ifade etmektedir (Şekil 3.5). TP pozitif etiketlenen pozitif gruplar, TN negatif olarak etiketlenen negatif gruplar ve T pozitifler ve N negatifler olmak üzere toplam sayıyı göstermek üzere Aşağıdaki gibi bir formülle gösterilir (3.7):

$$\text{Sınıflandırma Doğruluğu} = (TP + TN)/(P + N) \quad (3.7)$$

Gerçek sınıf	Tahmin edilen sınıf		
	Pozitif	Negatif	Toplam
Pozitif	TP	FN	P
Negatif	FP	TN	N
Toplam	P'	N'	$P+N$

Şekil 3.5: Konfüzyon matrisi

3.4.2 F-Ölçütü

F1 Score ismi ile de bilinen F-ölçütü; sınıflandırma sonuçlarını değerlendirmek için kesinlik (*precision*) ve duyarlılık (*recall*) ölçütlerini kullanarak elde edilen değerlendirme kriterlerinden birisidir. TP, TN, FP, FN, T ve P değerleri sınıflandırma doğruluğu bölümünde verilen değerler ile aynı olmak üzere duyarlılık ve kesinlik ölçütleri aşağıdaki gibi gösterilir (Formül 3.8 ve 3.9):

$$Kesinlik = \frac{TP}{TP + FP} \quad (3.8)$$

$$Duyarlılık = \frac{TP}{TP + FN} \quad (3.9)$$

Kesinlik ve duyarlılık ölçütlerinden elde edilen değerler kullanılarak F-ölçütü değeri (3.10) ile hesaplanır:

$$F - ölçütü = \frac{2 \times Kesinlik \times Duyarlılık}{Duyarlılık + Kesinlik} \quad (3.10)$$

3.4.3 Atak Dayanıklılığı

Literatür incelendiğinde matris çarpımına bağlı kimliksizleştirme yöntemlerine karşı kullanılan veriden orijinal veriyi çıkarmaya amaçlı çeşitli saldırı yöntemleri bulunmaktadır[101]. Bunlardan en sık kullanılanları Bağımsız Bileşen Analizi-ICA(Independent Component Analysis), Girdi/Çıktı saldırısı(I/O attacks) ve Naif Girişim-NI(Naive Inteference)'dır. Buradaki ölçülen değer orijinal veriler ile değeri değişmiş veriler arasındaki farkın standart sapması, yani verilerin değişim miktarı olarak düşünülmelidir.

3.4.4 Çalışma Zamanı

Çalışma zamanı (execution time, run time), bir programın başlatılması ile bitmesi arasında geçen süreye denilmektedir. Yürütme süresi olarak da bilinir. Programın başarılı kabul edilebilmesi için kabul edilebilir makul bir süre içerisinde işini bitirmesi beklenmektedir.

3.4.5 Friedman Rank Testi

Sınıflandırma doğruluğu, F-ölçütü ve atak dayanıklılığı sonuçları parametrik olmayan istatistiksel sına yöntemlerinden biri olan Friedman's testi[102] kullanılarak değerlendirilmiştir. Bu teste göre farklı koşullarda birden fazla kez tekrar eden testlerin sonuçlarının birbirine göre değerlendirmesi yapılır ve bir değer üretilir. Üretilen bu değer yüksek olması, yöntemin diğer yöntemlere göre belirgin bir fark ortaya koyduğunu ifade eder.

$$F = \left[\frac{12}{Nk(k+1)} + \sum_{i=1}^k R_i^2 \right] - 3N(k+1) \quad (3.11)$$

3.5 SINIFLANDIRMA YÖNTEMLERİ

Bu çalışma kapsamında önerilen modelin sınıflandırma doğruluk oranının ölçülmesi için literatürde bu tarz anonimleştirme çalışmalarında sıklıkla kullanılan Naive Bayes, J48, Karar Tabloları(Decision Trees) ve OneR sınıflandırıcılar kullanılmıştır.

3.5.1 Naive Bayes Sınıflandırıcılar

İstatistiksel sınıflandırma yöntemlerinden biridir ve Bayes teoremine dayanarak işlem yapar. Herhangi bir sınıflandırma yapmadan önce olasılıksal olarak hangisine daha uygun olduğunu kıyaslayarak işlem yapmaktadır.

1812 yılında Thomas Bayes tarafından bulunan koşullu olasılık hesaplama teoremi olan Bayes; rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi göstermektedir. Buna göre: $P(A|B)$ = B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı, $P(A)$ = A olayının gerçekleşme olasılığı, $P(B|A)$ = A olayı gerçekleştiğinde B olayının gerçekleşme olasılığı ve $P(B)$ = B olayının gerçekleşme olasılığı olmak üzere denklem olarak şu şekilde ifade edilmektedir:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)} \quad (3.12)$$

Naive Bayes; Gauss (Gaussian) Naive Bayes, Çokterimli (Multinomial) Naive Bayes ve Bernoulli Naive Bayes olarak üç farklı türde çalışmaktadır. Önemli bir varsayım yaparak sınıflandırma işlemini yapar; buna göre hiçbir durumun olasılığı sıfır olmamalıdır. Sistem bir defa eğitildikten sonra etkin çalışsa da ilk eğitim aşaması zaman ve kaynak kullanımı açısından maliyetlidir. Basitçe Naive Bayes algoritması şu şekilde çalışmaktadır:

- Öncelikle sınıflandırılacak özellik için veri setinde frekans tablosu oluşturularak, her bir değerin tekrarlanma durumu incelenir.
- Bu frekans tablosundan yola çıkarak her bir sınıfın olasılığı hesaplanarak olasılık tablosu oluşturulur.
- Son olarak elde edilen bu değerlere göre (3.12) Bayes formülü uygulanarak yüksek olasılıklı olan sınıf seçilir.

Naive Bayes sınıflandırıcılar, gerçek zamanlı tahminleme, çoklu sınıf tahminleme, metin sınıflandırma, istenmeyen (spam) filtreleme gibi birçok alanda kullanılmaktadır[103].

3.5.2 J48 Sınıflandırıcı

ID3 olarak da bilinen J48 sınıflandırıcılar sık kullanılan karar ağacı algoritmalarından biridir. C4.5 karar ağacı algoritmasının Weka'ya uygulanmasıyla oluşturulmuştur. , J. Ross Quinlan tarafından geliştirilen bu algoritma böl ve yönet mantığını uygulamaktadır. Shannon'ın Bilgi Kuramı'ndan[104] faydalanarak karar ağacını entropi değerini baz alarak optimize etmeyi amaçlar. Basitçe çalışma adımları:

- İlk olarak hedef sınıf için entropi değeri hesaplanır.
- Her sınıf için ayrı ayrı bilgi değeri(information) ve sınıfın bilgi kazanımı(information gain) hesaplanır.
- En yüksek bilgi kazanımına sahip olan sınıf üzerinden ağaç yapısı oluşturularak dallanma gerçekleştirilir.
- İlk sınıf bu şekilde tespit edildikten sonra işlem yinelemeli(recursive) olarak tekrar edilerek diğer sınıf bilgileri elde edilir.

3.5.3 Karar Tablosu Sınıflandırıcı (Decision Table)

Karar tabloları sayısal tahminleme için kullanılan özelleştirilmiş karar ağacı yöntemlerinden biridir ve Wekada kurallar(rules) kısmında listelenir. Hesaplama açısından daha basit bir mantığı bir mantığa sahiptir. Orijinal veri seti ile aynı sayıda özniteliğe sahip olan bir tablo ile veri setini özetler. Buna göre karar tablosu çoğunluk sınıflandırıcı oluşturur. En iyi ilk aramayı(best-first search) kullanarak özellik alt kümelerini değerlendirir ve değerlendirme için çapraz doğrulama kullanır. Bir sınıf belirlenirken tablonun çoğunluğu yerine, bir karar tablosu girişi tarafından kapsanmayan her bir örnek için sınıfı belirlemek için en yakın komşu(k nearest neighbour) yöntemini kullanır.

3.5.4 OneR Sınıflandırıcı

Holte tarafından 1993 yılında önerilen kural tabanlı sınıflandırma yöntemlerinden biridir. ZeroR algoritmasının gelişmiş versiyonu olarak bilinir. Tek bir özelliği baz alarak sınıflandırma yaptığı için yüzeysel daha basit bir sınıflandırıcı olarak kabul edilir. Kuralı oluşturmak için konfüzyon matrisi oluşturur, yani var olan özelliklerin sınıfı nasıl etkilediğini belirlemek için matrisler oluşturur. Daha sonra hangi özelliğin daha yüksek etkileme kapasite olduğunu tespit ettikten sonra, o özelliğe göre sınıflandırmayı gerçekleştirir.

3.6 ATA K TÜRLE Rİ

Etkin bir gizlilik koruma algoritması aynı zamanda olası saldırıları da göze önüne alarak değerlendirilmelidir. Yapılan çalışmalar göstermektedir ki sadece pertürbasyon öncesi ve sonrası durum arasındaki farkı gözetmek yeterli değildir[93]. Çalışmalar incelendiği bu alanda yapılan birçok saldırı çeşidi olduğu bilinmektedir[101]. Saf tahmin yöntemi saldırganın orijinal veri seti hakkında herhangi bir bilgiye sahip olmadığını varsayarken, yeniden yapılandırma tabanlı saldırganın sütun dağılımı hakkında az da olsa bilgi sahibi olduğu durumları düşünür ve son olarak mesafe çıkarım saldırıları ise orijinal veriler üzerinde bazı noktaların bilinmesi durumunda kullanılmaktadır. Tez kapsamında önerilen model 3 farklı saldırı yöntemi ile test edilmiştir.

3.6.1 Saf Tahmin Yöntemi (Naive Interference / Estimation)

Çıkarım saldırısı da denilen bu yöntemde orijinal veriler hakkında herhangi bir bilgi sahibi olmadan, değiştirilmiş veriler üzerinde analizler yapılarak bilgi elde etme yöntemidir. En basit

şekliyle tek tek deneme yoluyla bilgi edinme olarak düşünülebilir. Her bir çıkarım saldırısı farklıdır çünkü veri setinin bir özneliliğini detaya inerek bulmaya çalışır. Geometrik veri pertürbasyonunda çıkarım saldırısına karşı dayanıklı olma durumu normalize edilmiş orijinal veri seti ile pertürbasyon sonrası oluşan verilerin arasındaki fark olarak değerlendirilmektedir. Bu fark ne kadar büyükse veriler o kadar değiştirilmiş ve kırılması zor hale getirilmiştir kabul edilir.

3.6.2 ICA(Bağımsız Bileşen Analizi) Saldırısı

Daha önceki bölümlerde anlatıldığı gibi geometrik veri pertürbasyonunda aynı birden fazla sütunu anonimleştirmek mümkündür. Verilerin koordinat düzlemi üzerinde belirli bir miktarda açı kadar rotasyonu ve ötelenmesi sonrasında, aynı anda birden fazla nitelik birbirine bağlı olarak anonimleştirilmiş olur. Tek bir özneliliğin değişim miktarını aradaki fark olarak hesaplansa da çoklu sütun değişiminin ifade edilişi daha önceki bölümlerde anlatılmıştır.

Bu nedenle burada kullanılabilecek bir diğer saldırı yeniden yapılandırma türü saldırı olan ICA(Bağımsız Bileşen Analizi)'dir. ICA modeli bozulan veri seti üzerinden, orijinal satır vektörlerini tahmin etmeye çalışır. Bu yüzden ICA saldırılarına dayanıklı bir modelin çoklu sütun gizliliğinde yüksek bir farka sahip olması beklenir, pertürbasyon sonrası ve öncesi değerlerin farklarının varyansı olarak kabul edilir.

3.6.3 I/O (Girdi/Çıktı) Saldırısı

Bu saldırı yönteminde ise saldırganın orijinal veri setinden bazı verileri bilmesi ve pertürbasyon sonrası yayınlanan veriler ile eşleştirmesi sonrasında elinde bulunan eşleştirilmiş çiftleri kullanarak diğer verilerden çıkarım yapmasına dayanmaktadır. Saldırgan elindeki eşleştirilmiş veri çiftleri ile bir bilgi matrisi oluşturarak bilgi çıkarımı yapmaya çalışır. Dayanıklılığı sağlayacak olan parametre Öklid mesafesi olarak kabul edilir. Bilinen bu girdi/çıktı saldırı tipi saldırganın elinde az bir miktarda veri olsa bile etkili sonuçlar verebilmektedir[105].

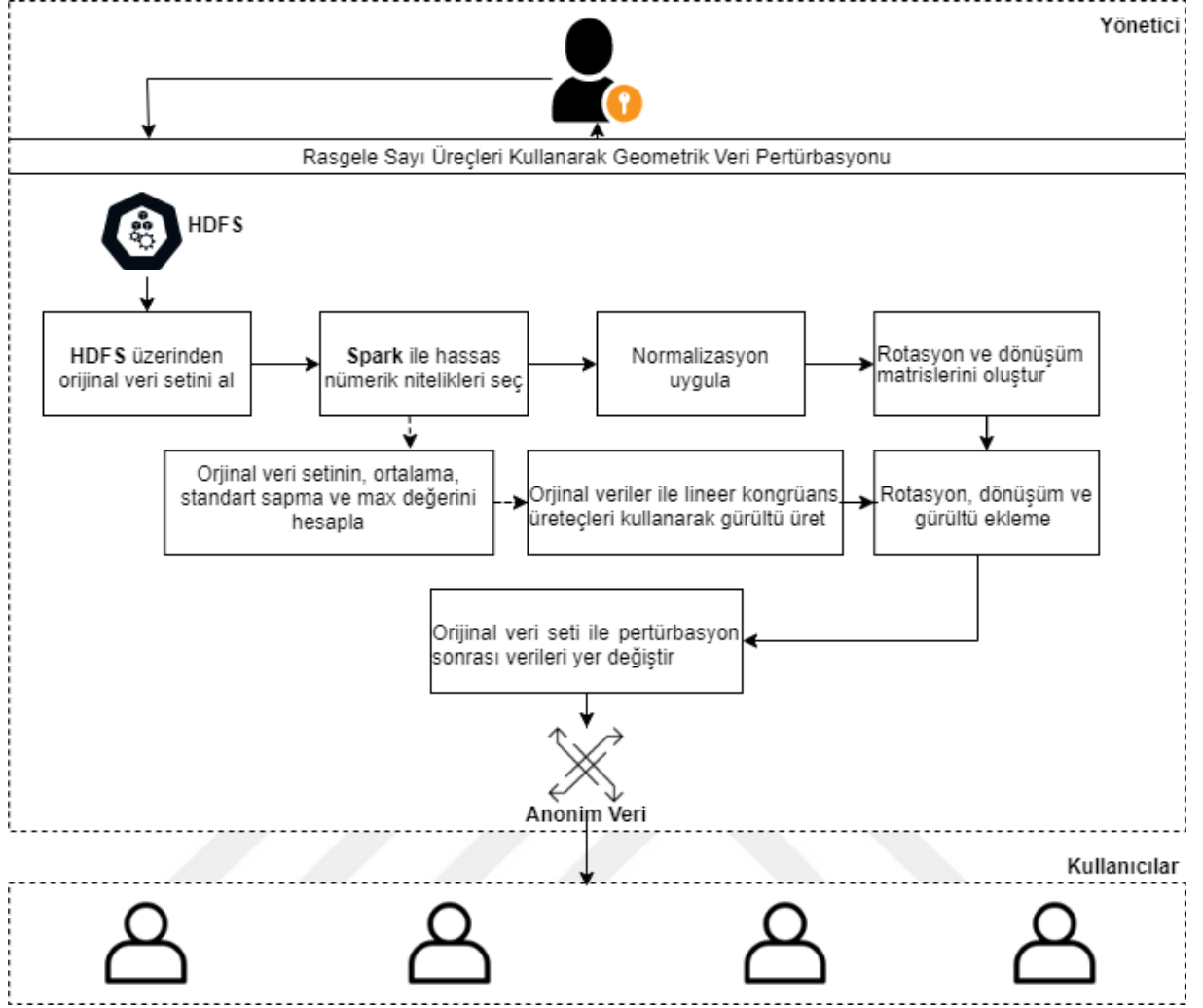
Çalışma içerisinde saldırganın eldeki verinin %10unu bildiği varsayılmıştır. Ayrıca tekrar sayısı 10, sigma parametresi ise 0,3 olarak belirlenmiştir.

3.7 ÖNERİLEN YÖNTEM: RSUGP

Şimdiye kadar daha önceki bölümlerde anlatılan yöntemlerle ilgili birçok çalışma yapılmıştır ve güncel bir konu olması sebebiyle etkin anonimleştirme modelleri üzerine günümüzde çalışılmaya devam edilmektedir. Literatürdeki çalışmaların da detaylıca incelenmesiyle pertürbasyon yöntemlerinin başarısı izlenmiştir. Geometrik veri pertürbasyonu matris çarpımına dayalı etkin yöntemlerden biridir. Ancak çalışmalar incelendiğinde kullanılan gürültü modellerinin ya 0-1 arası rasgele oluşturulmuş bir gürültü, ya da Gauss gürültü olarak seçildiği izlenmiştir. Buradan yola çıkılarak literatürde kullanılmamış yeni bir gürültü modeli içeren geometrik veri pertürbasyonu ihtiyacı fark edilmiştir.

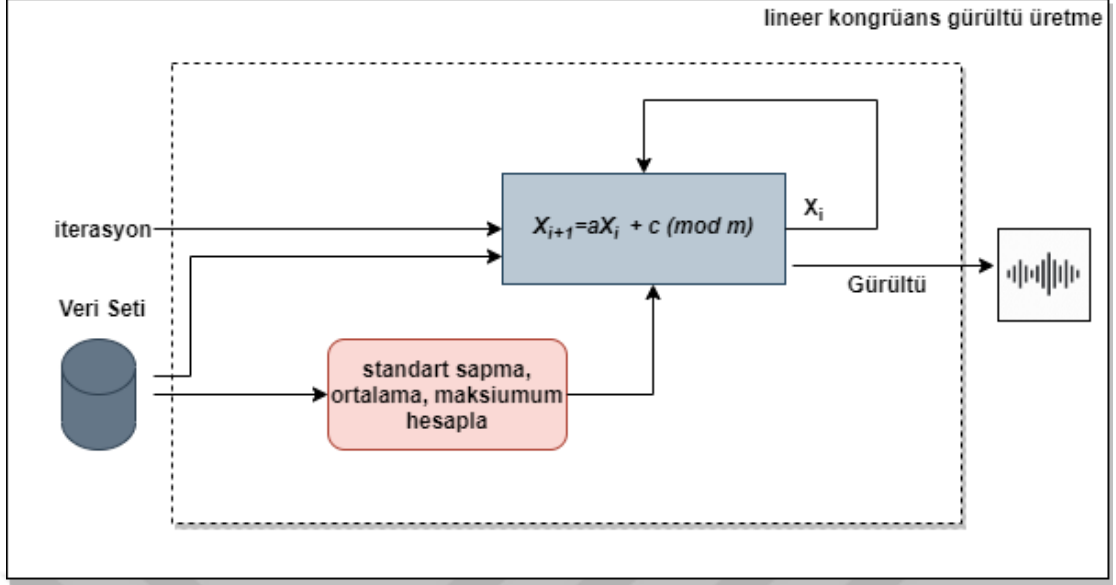
Bu tez çalışması kapsamında etkin bir gizlilik koruması için rasgele sayı üreteçleri ve geometrik veri pertürbasyonu tabanlı yeni bir yöntem önerilmiştir: RSUGP(Rasgele Sayı Üreteçleri tabanlı Geometrik Pertürbasyon). RSUGP, oluşturduğu gürültü modeli ile geri döndürülemez bir girdi pertürbasyon mekanizmasıdır. Gürültü oluşturmak için rastgele sayı üreteçlerini kullanmak, anonimleştirme için yeni bir yöntemdir ve bu çalışmanın ana katkılarından biridir. Ana katkımız, veri setinden bağımsız ve tüm sayısal özniteliklere uygulanabilen, gizliliği koruyan bir veri yayınlama algoritması geliştirmektir. RSUGP' nin daha iyi sınıflandırma doğruluğu, daha iyi saldırı direnci ve karşılaştırılan yöntemlerden daha hızlı ile daha fazla gizlilik garantisi sağladığı deneysel olarak kanıtlanmış ve sonuçlar bulgular kısmında paylaşılmıştır.

Bu tez çalışması kapsamında rasgele sayı üreteçlerine bağlı yeni bir geometrik veri pertürbasyonu modeli önerilmiştir. Önerilen algoritmanın modeli Şekil 3.6 üzerinde gösterilmiştir.



Şekil 3.6: Rasgele sayı üreteçlerine bağlı geometrik veri pertürbasyon modeli

Rasgele sayı üretmede kullanılan Lineer Kongrüans üreteçlerinde a , c ve m parametrelerinin alacağı farklı değerlere karşılık modelin vereceği sonuç değişmektedir. İncelenen çalışmalarda Gauss gürültüsünü başarılı yapan verinin kendisini kullanarak gürültü oluşturması olduğu fikrinden yola çıkılarak, Lineer Kongrüans Üreteçleri düzenlenmiştir. Buna göre verinin kendisi başlangıç değeri olarak, verinin en yüksek değeri m değeri olarak, standart sapması a değeri olarak ve son olarak da verinin ortalaması ise c değeri olarak değerlendirilmiştir. Böylelikle verinin kendisine uygun bir gürültü modeli oluşturulmuştur, tahmin edilebilirlik minimuma indirilmiştir. Lineer Kongrüans üreteçlere bağlı rasgele gürültü üretme modeli Şekil 3.7 da gösterilmiştir.



Şekil 3.7: Lineer Kongrüans Üreteçler kullanarak gürültü üretme

Tez kapsamında sunulan etkin gizlilik koruması modeli için uygulanan Lineer Kongrüans Gürültü üretme algoritması Algoritma 2 de, akış diyagramı ise Şekil 3.8 de sunulmuştur.

Algoritma 2: RSU Gürültü Üretme

Girdi: Orijinal veri seti D , hassas nümerik değerler $SA(SA_1, SA_2, SA_3 \dots SA_N)$

Çıktı: RSU gürültü değerleri N

Başlangıç Değerleri: $iterasyon$

- 1: $d = |D|$
 - 2: $s = |SA|$
 - 3: **for** $i = 1$ to s **do**
 - 4: $attStd_i =$ hassas niteliğin standart sapması
 - 5: $attMean_i =$ hassas niteliğin ortalaması
 - 6: $attMax_i =$ hassas niteliğin maksimum değeri
 - 7: **for** $j = 1$ to d **do**
 - 8: $N_i = D_i$
 - 9: **for** $k = 1$ to $iterasyon$ **do**
 - 10: $N_{ij} = \text{mod}(attStd_i * N_{ij} + attMean_i, attMax_i)$
 - 11: **end for**
 - 12: $D_i = N_i$
 - 13: **end for**
 - 14: **end for**
 - 15: üretilen gürültü değerleri çıkış değeri olarak dışarı aktarılır
 - 16: $N = D$
-

Önerilen model şu şekilde çalışmaktadır:

Adım 1: HDFS üzerinde tutulan orijinal veri seti D üzerinden, belirlenen nümerik hassas nitelikler $SA(SA_1, SA_2, SA_3..SA_N)$ pyspark sorgusu ile çekilir.

Adım 2: Seçilen bu niteliklere normalizasyon uygulanır.

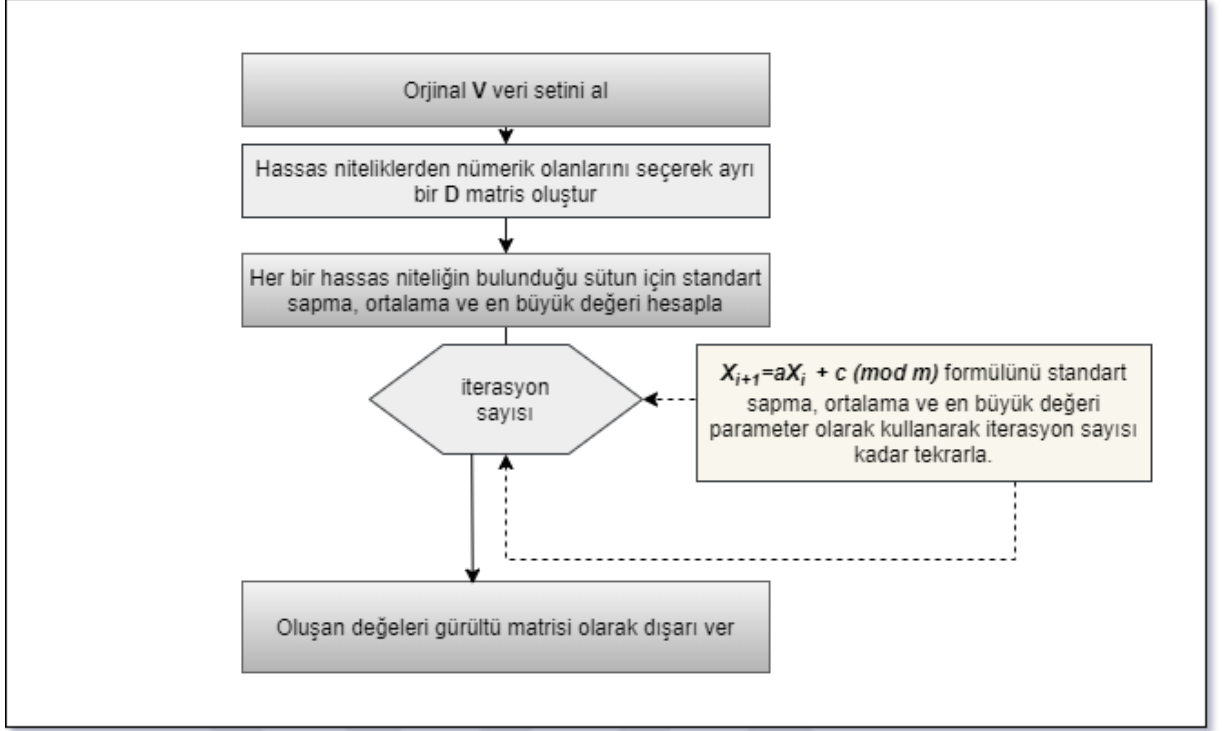
Adım 3: Rotasyon ve öteleme matrisleri oluşturulur.

Adım 4: Seçilen bu nitelikler önce rotasyon sonra da öteleme işlemine tabi tutulur. Ve de-normalizasyon işlemi yapılarak veriler kendi değer aralıklarına geri döndürülür.

Adım 5: Normalizasyon öncesi orijinal verilerden seçilen nümerik hassas niteliklerin her birinin ayrı ayrı standart sapma, ortalama ve maksimum değerleri hesaplanarak rasgele sayı üreteçlerine bağlı gürültü üretme fonksiyonuna verilir.

Adım 6: Lineer Kongrüans Gürültü ekleme modeli çalıştırılarak her bir nitelik için uygun gürültü oluşturulur. Bu gürültüler daha sonra rotasyon ve öteleme sonrası de-normalizasyon yapılmış olan değerlerin üzerine eklenir.

Adım 7: Orijinal veri seti üzerinden seçilen nümerik hassas niteliklerin değerleri, önerilen gizlilik korumalı anonimleştirme modeli ile üretilen yeni değerler ile yer değiştirilerek yayınlanır.



Şekil 3.8: Lineer Kongrüans Akış Diyagramı

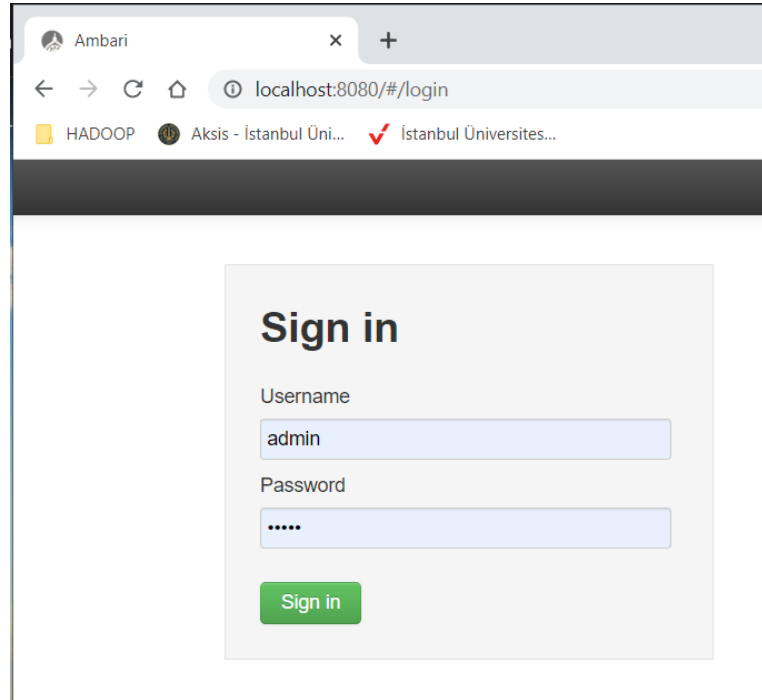
3.8 ÖNERİLEN RSUGP MODELİNİN SPARK ÜZERİNDE UYGULANMASI

Bu tez çalışmasının amacı etkin bir gizlilik koruma modeli oluşturmaktır. Ayrıca önerilen bu modelin dağıtık veri ortamları üzerinde uygulanması da yine bu kapsamda ele alınmıştır. Daha önceki bölümlerde anlatıldığı üzere büyük verilerin daha hızlı ve etkin bir biçimde işlenmesi ve depolanabilmesi için geleneksel yöntemler yerine artık dağıtık veri ortamları tercih edilmektedir.

Apache Hadoop da büyük miktardaki verilerin hesaplama problemlerini çözmek için kullanılan ve birçok bilgisayardan oluşan açık kaynak kodlu bir yazılım araçları topluluğudur. İçerisinde farklı görevleri yerine getirmek üzere tasarlanmış birçok alt programı barındırmaktadır. Apache Spark ise büyük ölçekli verileri hızlıca işleme kapasitesine sahip açık kaynak kodlu bir veri işleme motorudur. Bu çalışma içerisinde Hadoop HDFS(Hadoop Distributed File System) üzerinde tutulan veri setleri, Spark ile işlenerek anonimleştirme gerçekleştirilmiştir. Uygulanan model mimarisi Şekil 3.5 üzerinde gösterilmiştir.

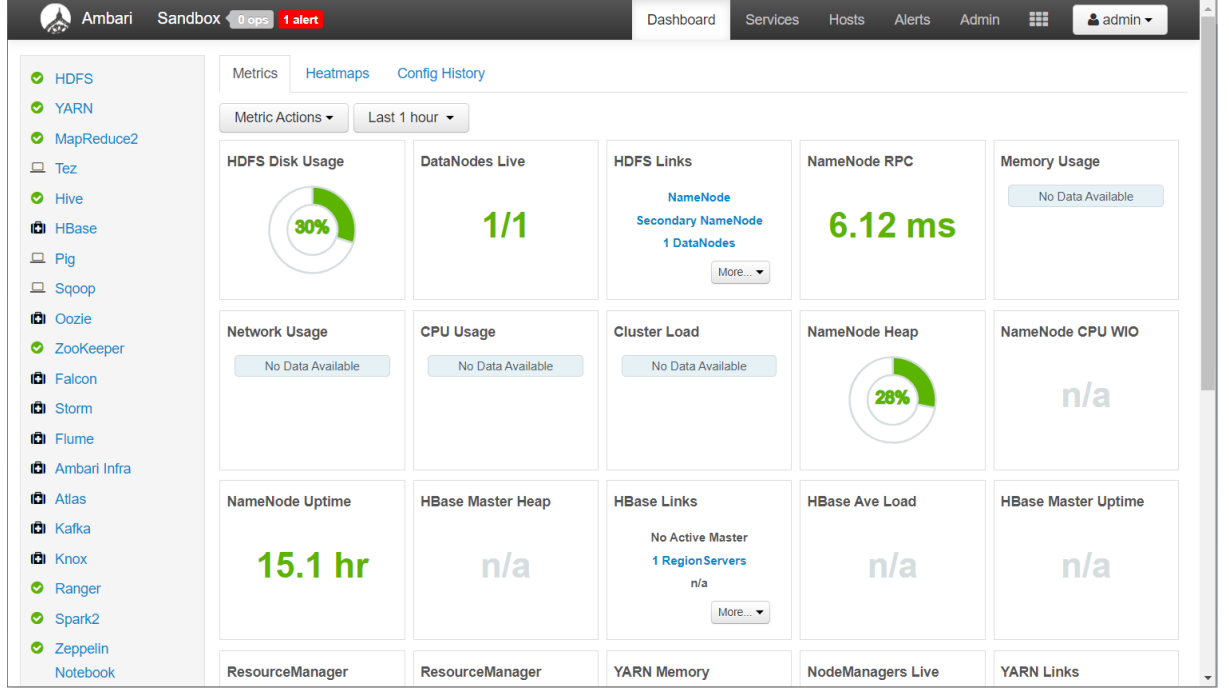
Tez kapsamında kullanılan ortamlar ve kurumları aşağıda anlatılmıştır:

1. Sanal makine olarak Oracle VM Virtual Box 6.1 kurulur.
2. Windows bilgisayardan uzak Linux sunucusuna SSH(Secure Shell Host) bağlantısı sağlamak ve komut satırını kullanabilmek için Putty kurulumu gerçekleştirilir.
3. Sandbox HDP VirtualBox 2.6.5 kurulumu için “.ova” uzantılı dosya indirilerek, Virtual Box ile içe aktar denilerek kurulumu başlatılır. Ekranda bazı IP ve port numaraları belirene kadar kurulumun tamamlanması beklenir.
4. Putty açılarak sanal makinenin IP adresi 127.0.0.1 ve port numarası 2222 yazılarak bağlantı sağlanır.
5. Kurulum ve bağlantılar tamamlandıktan sonra localhost:1080 ile Ambari ara yüzüne erişim sağlanır (Şekil 3.9).
6. Diğer erişim IP leri ise:
 - Sistemde bulunan 1 adet işçi düğüm(data node): sandbox-dp.hortonworks.com:50010
 - Sistemde bulunan 1 adet ana düğüm(name node): sandbox-dp.hortonworks.com:8020
 - Kaynak yöneticisi(Resource Manager) için : <http://localhost:8088/cluster>
 - Düğün yöneticisi (Node Manager) için: <http://localhost:8042/node>



Şekil 3.9: Ambari ara yüzü bağlantı ekranı

Gerekli kurumlar sıkıntısızca gerçekleştirildikten sonra Şekil 3.10 da gösterildiği gibi Hadoop bileşenleri başarıyla başlatılmış olur. Açılan ekranda sistemin genel durumu ile ilgili bilgilerde izlenebilmektedir. Sol tarafta servislerin yanında bir uyarı işareti olmaması demek servislerin sıkıntısızca çalıştığı anlamına gelmektedir.



Şekil 3.10: Ambari üzerinde Hadoop bileşenleri ve sistemin genel görüntüsü

Kaynak yöneticisi sistemde o anda hangi uygulamalar çalışıyor, hangi işlemler YARN dan kaynak alabildi, alamadıysa neden alamadı diye ayarlamaların ve görüntülemelerin yapıldığı Hadoop bileşenidir. Şekil 3.11 da kaynak yöneticisinin anlık durumunu gösteren ekran bulunmaktadır.



All Applications

Cluster

- About
- Nodes
- Node Labels
- Applications
- NEW
- NEW_SAVING
- SUBMITTED
- ACCEPTED
- RUNNING
- FINISHED
- FAILED
- KILLED
- Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes
1	0	1	0	2	2.44 GB	2.93 GB	0 B	2	8	0	1

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[MEMORY]	<memory:250, vCores:1>

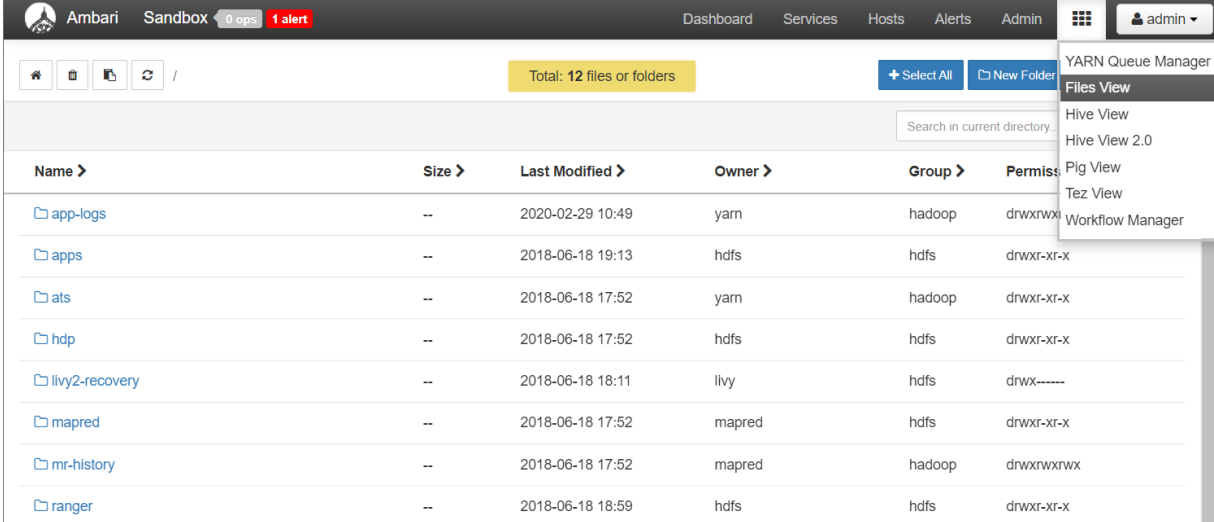
Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCoers	Allocated Memory MB
application_1629400346685_0001	zeppelin	Zeppelin	SPARK	default	0	Thu Aug 19 22:16:41 +0300 2021	N/A	RUNNING	UNDEFINED	2	2	2500

Showing 1 to 1 of 1 entries

Şekil 3.11: Ambari üzerinde kaynak yöneticisi (Resource Manager) görüntüsü

HDFS verilerin depolandığı bileşendir. İçerisine iki türlü veri yüklenebilir. Bir dosya yükleme ve görüntüleme kısmı olan Files View bölümünde yeni dosya yükle seçeneği ile yüklenebilir (Şekil 3.12).



Ambari Sandbox 0 ops 1 alert

Dashboard Services Hosts Alerts Admin

admin

Total: 12 files or folders

+ Select All New Folder

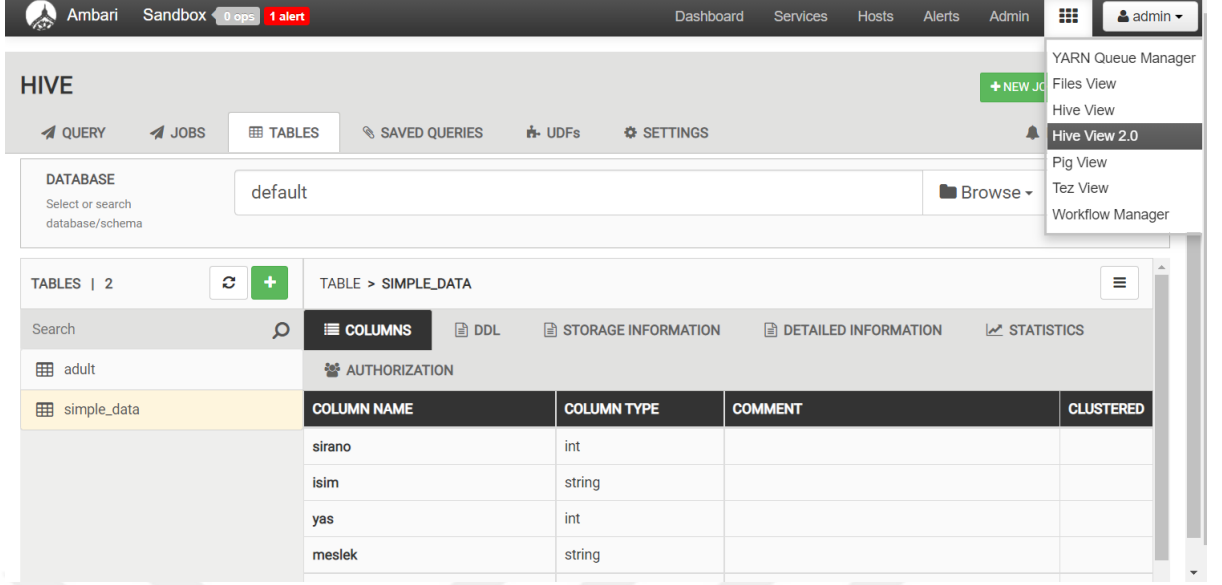
Search in current directory...

Name	Size	Last Modified	Owner	Group	Permissions
app-logs	--	2020-02-29 10:49	yarn	hadoop	drwxrwx--
apps	--	2018-06-18 19:13	hdfs	hdfs	drwxr-xr-x
ats	--	2018-06-18 17:52	yarn	hadoop	drwxr-xr-x
hdp	--	2018-06-18 17:52	hdfs	hdfs	drwxr-xr-x
livy2-recovery	--	2018-06-18 18:11	livy	hdfs	drwx-----
mapred	--	2018-06-18 17:52	mapred	hdfs	drwxr-xr-x
mr-history	--	2018-06-18 17:52	mapred	hadoop	drwxrwxrwx
ranger	--	2018-06-18 18:59	hdfs	hdfs	drwxr-xr-x

YARN Queue Manager
Files View
Hive View
Hive View 2.0
Pig View
Tez View
Workflow Manager

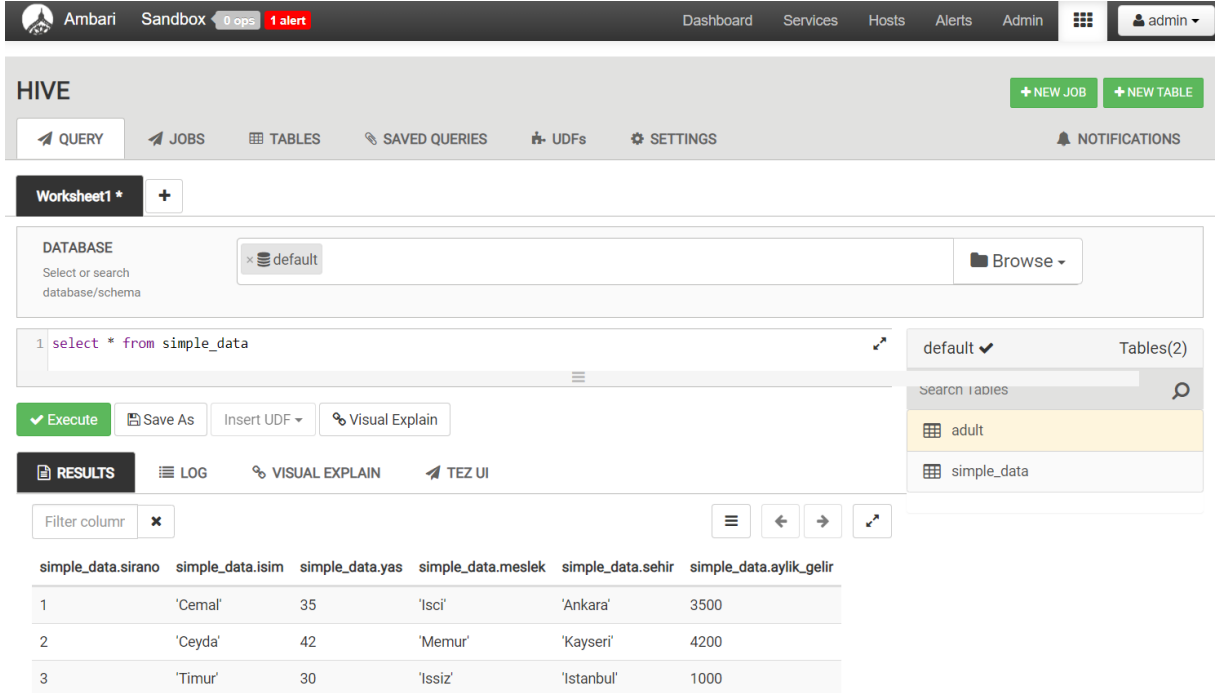
Şekil 3.12: HDFS üzerine dosya yükleme görüntüsü

İkinci bir seçenekte veri setlerinin veri tabanı özelliklerini koruyarak, sütun bilgilerinin ve veri tiplerinin ayarlanarak yüklenebildiği seçenektir (Şekil 3.13)



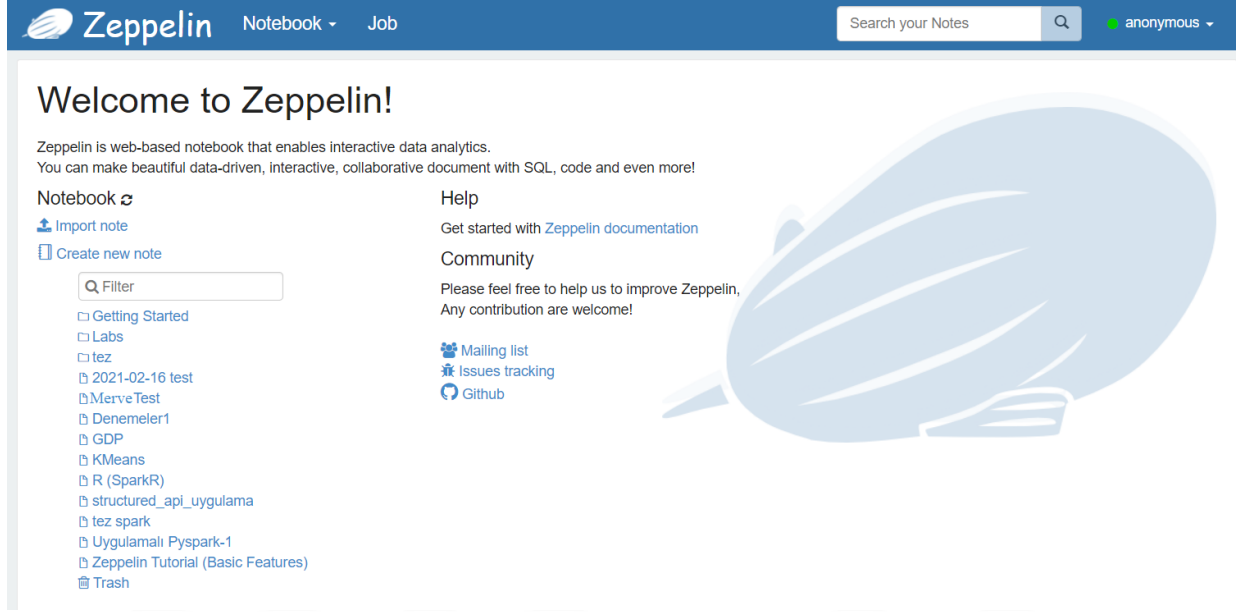
Şekil 3.13: HIVE ile veri seti yükleme görüntüsü

Daha önceki bölümlerde anlatıldığı üzere Hadoop farklı görevleri gerçekleştirmek üzere birçok bileşene sahiptir. HIVE bunlardan birisidir. Dağıtık veri ortamında kişilerin alışkın olduğu SQL sorgu tipini kullanarak veri ekleme silme ve düzenleme işlemlerini yapabilmesi için HIVE kullanılmaktadır. Bu nedenle veri yükleme yapılan aynı ekrandan sorgu yapabilmek mümkündür (Şekil 3.14).



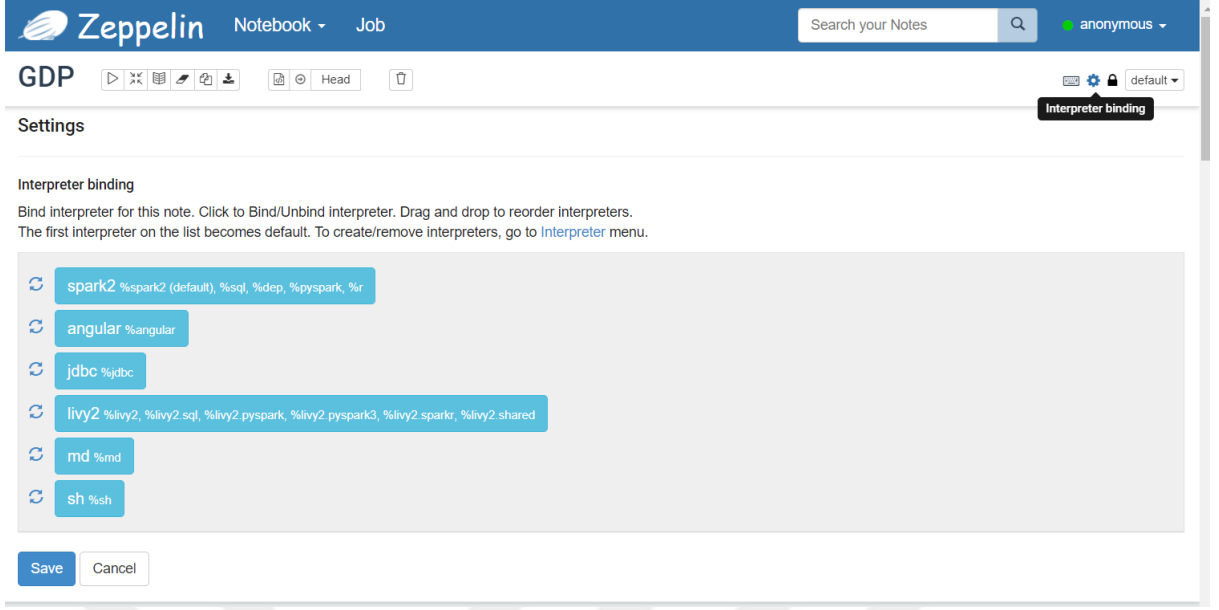
Şekil 3.14: HIVE sorgusuyla örnek veri çekme

Veri seti HDFS üzerine yüklendikten sonra anonimleştirme işlemini gerçekleştirmek için Spark kullanılmaktadır. Kod yazma kolaylığı sağlaması açısından Zeppelin Notebook kullanmak kolaylık sağlamaktadır. Bu yüzden Şekil 3.10 da sol alt köşede bulunan menü üzerinden Zeppelin seçilir ve yazım ekranına ulaşılır (Şekil 3.15).

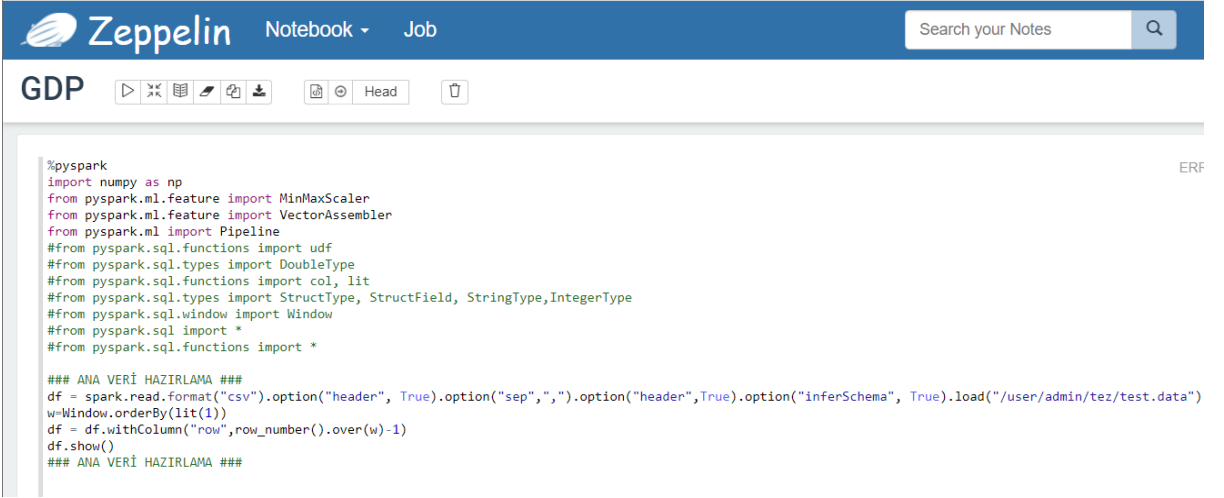


Şekil 3.15: Zeppelin başlangıç ekranı

Kodlamaya başlamadan önce yorumlayıcı(interpreter) seçimi yapılmalıdır. Seçenekler spark2, angular, jdbc, livy2, md, sh olarak liste olarak bulunmaktadır. Burada pyspark kodlama yaptığımız için spark2 seçilerek devam edilir(Şekil 3.16).



Şekil 3.16: Zeppelin yorumlayıcı seçimi



Şekil 3.17: Zeppelin kütüphaneleri koda eklenmesi

Gerekli kütüphaneler eklenerek kodlamaya başlanır (Şekil 3.17). Gerekli bütün düzenlemeler ve kütüphane eklemeler bitirildikten sonra Şekil 3.6’te anlatılan önerilen RSUGP modelinin pyspark ile kodlaması tamamlanır. Elde edilen anonim veriler ile orijinal verilerin yerleri değiştirilerek verinin pertürbasyon işlemi tamamlanmış olur. Şekil 3.18 da örnek kod bölümü gösterimi bulunmaktadır.



The image shows a Zeppelin Notebook interface. At the top, there's a blue header with the Zeppelin logo, 'Notebook' and 'Job' tabs, a search bar, and a user profile 'anonymous'. Below the header, the notebook title 'GDP' is visible. The main area contains a PySpark script for data normalization and rotation. The script calculates min, max, and average salary, then normalizes the 'age' and 'salary' columns using MinMaxScaler. It then rotates the scaled data by 61 degrees using a rotation matrix R. Finally, it denormalizes the rotated data back to the original scale.

```

salary_max = df.agg({"salary": "max"}).collect()[0][0]
salary_min = df.agg({"salary": "min"}).collect()[0][0]
salary_avg = df.agg({"salary": "avg"}).collect()[0][0]
### MIN-MAX DEĞERLER ###

### NORMALİZASYON ###
print('### NORMALİZASYON ###')
for i in ["age", "salary"]:
    assembler = VectorAssembler(inputCols=[i], outputCol=i+"_Vect")
    scaler = MinMaxScaler(inputCol=i+"_Vect", outputCol=i+"_Scaled")
    pipeline = Pipeline(stages=[assembler, scaler])
    df = pipeline.fit(df).transform(df).withColumn(i+"_Scaled", unlist(i+"_Scaled")).drop(i+"_Vect")
    df.show()
### NORMALİZASYON ###

### ROTASYON MATRİSİ ###
print('### ROTASYON MATRİSİ ###')
theta = np.radians(61)
c, s = np.cos(theta), np.sin(theta)
R = np.array(((c, -s), (s, c)))
print(R)
data = np.array(df.select('age_Scaled', 'salary_Scaled').collect())
rotated = data.dot(R)

df_rotated = sc.parallelize(rotated).map(lambda x: x.tolist()).toDF(["age_Scaled", "salary_Scaled"])
df_rotated.show()

### ROTASYON MATRİSİ ###

### DENORMALİZASYON ###
print('### DENORMALİZASYON ###')
df_denormalized = df_rotated.withColumn("age_Denormalized", col("age_Scaled") * (age_max - age_min) + age_min).withColumn(
    "salary_Denormalized", col("salary_Scaled") * (salary_max - salary_min) + salary_min)
df_denormalized.show()

```

Şekil 3.18: Zeppelin üzerinde pyspark ile rotasyon matrisi oluşturma

4 BULGULAR

Tez kapsamında önerilen algoritma 16GB RAM, Intel Core i5-1035G1 işlemci ve Windows 10 işletim sistemine sahip kişisel bilgisayar üzerinde sanal makineye kurulu, Hadoop ve Spark kullanılarak gerçekleştirilmiştir. Pyspark kodlama kolaylığı sağlaması açısından Ambari arayüzü üzerinden Zeppelin Notebook editör kullanılmıştır. 6 farklı veri seti ile performans analizleri yapılan modelin, sınıflandırma doğruluğu ve F- ölçütü sonuçları Weka 3.8 kullanılarak elde edilmiştir. Atak dayanıklılığı için ise Matlab R2018a üzerinde FastICA[106] paketi kullanılarak sonuçlar elde edilmiştir. Elde edilen sonuçların çoklu değerlendirilebilmesi için Friedman rank testi IBM SPSS Statistics 26 kullanılmıştır.

4.1 SINIFLANDIRMA DOĞRULUĞU SONUÇLARI

Önerilen algoritmanın sınıflandırma doğruluğunun ölçülebilmesi için Naive Bayes, J48, Decision Table ve One R olmak üzere 4 farklı sınıflandırma algoritması kullanılmıştır. 2 farklı eğitim modeli kullanılarak sonuçlar elde edilmiştir. Önce veri setlerinin tamamı hem eğitim hem de test için kullanılmıştır. İkinci olarak k-kat çapraz geçirme ile performans sonuçları elde edilmiştir. Değerlendirme yapılması açısından kullanılan veri setlerinin büyüklükleri ve seçilen nümerik hassas değer sayısı farklı farklı belirlenmiştir.

Modelin sınıflandırma doğruluğu; orijinal sınıflandırma doğruluğu, rasgele gürültü üretilerek oluşturulmuş kimliksizleştirme sınıflandırma doğruluğu, Gauss gürültü modeli kullanılarak oluşturulmuş sınıflandırma doğruluğu olmak üzere 3 farklı sonuç ile kıyaslanmıştır.

İlk olarak veri setlerinin tamamı hem eğitim hem de test verisi olarak kullanılarak elde edilen değerlendirme sonuçları Tablo 4.1 üzerinde paylaşılmıştır. Tablo içerisindeki değerler incelendiğinde orijinal sınıflandırma doğruluğuna en yakın sonucun önerilen RSUGP modeli ile üretilen RSU gürültü olduğu izlenmektedir. Değerlendirme tablosunun en alt satırında karşılaştırılan modellerin FMR(Friedmans Rank Test) sonuçları bulunmaktadır. FMR değerleri incelendiğinde yine önerilen RSUGP modelinin diğerlerine kıyasla en yüksek sonucu verdiği izlenmektedir.

Tablo 4.1: Önerilen algoritmanın veri setinin tamamı eğitim ve test verisi olarak kullanıldığı sınıflandırma doğruluğu sonuçları

Veri Setleri	Algoritmalar	Naive Bayes	J48	Decision Table	One R
Iris	Orijinal	96,00%	98,00%	96,00%	96,00%
	Rasgele Gürültü	91,94%	88,59%	87,24%	87,24%
	Gauss Gürültü	90,95%	90,42%	84,30%	86,25%
	RSU Gürültü	83,22%	98,65%	97,31%	93,28%
Heart	Orijinal	85,18%	91,48%	85,18%	76,29%
	Rasgele Gürültü	84,07%	91,11%	82,96%	76,29%
	Gauss Gürültü	84,81%	92,59%	85,15%	76,29%
	RSU Gürültü	84,81%	92,59%	85,18%	76,29%
Wine Quality-White	Orijinal	72,02%	44,87%	37,23%	60,77%
	Rasgele Gürültü	97,97%	44,87%	36,91%	98,89%
	Gauss Gürültü	97,91%	44,87%	36,91%	98,67%
	RSU Gürültü	97,59%	44,87%	36,56%	98,10%
Fried	Orijinal	86,53%	96,60%	84,56%	70,85%
	Rasgele Gürültü	70,92%	74,20%	67,89%	68,53%
	Gauss Gürültü	74,06%	78,11%	75,23%	74,02%
	RSU Gürültü	81,50%	93,17%	79,77%	71,16%
Bank Management	Orijinal	88,05%	94,12%	90,89%	89,46%
	Rasgele Gürültü	87,76%	93,58%	90,89%	89,46%
	Gauss Gürültü	88,07%	93,78%	90,89%	89,46%
	RSU Gürültü	88,12%	93,82%	90,85%	89,46%
Electricity	Orijinal	73,05%	96,37%	82,77%	78,20%
	Rasgele Gürültü	60,56%	90,75%	75,25%	68,31%
	Gauss Gürültü	61,38%	87,92%	73,41%	69,84%
	RSU Gürültü	61,02%	94,37%	75,25%	72,64%
FMR Değerleri	Rasgele Gürültü: 1,69	Gauss Gürültü: 2,00	RSU Gürültü: 2,31		

k- kat çapraz geçerleme yöntemi makine öğrenmesi çalışmalarında k defa geçerlemenin tekrar edildiği ve modelin her seferinde veri setinin farklı gruplarıyla eğitilmesi ve test edilmesini sağlayan istatistiksel bir modelidir. Eğitim ve test kümesinde veri setinin her seferinde farklı

gruplarını kullanmasından dolayı oldukça başarılı sonuçlar vermektedir. Bu nedenle önerilen modelin performans analizleri için sırasıyla 2-kat (Tablo 4.2), 5-kat (Tablo 4.3) ve 10-kat (Tablo 4.4.) çapraz geçerleme kullanılarak sonuçlar bir de bu şekilde değerlendirilmiştir.

Tablo 4.2: Önerilen algoritmanın 2-kat çapraz geçerleme sınıflandırma doğruluğu sonuçları

Veri Setleri		2-Kat çapraz geçerleme			
		NB	J48	DT	OneR
Iris	Orijinal	96.6	93.3	94.0	94.0
	Rasgele Gürültü	92.0	84.4	84.4	87.3
	Gauss Gürültü	88.6	87.5	83.3	83.3
	RSUGP	95.5	90.2	87.3	88.4
Heart (statlog)	Orijinal	82.2	80.0	74.0	67.7
	Rasgele Gürültü	81.8	80.3	73.8	67.7
	Gauss Gürültü	76.5	71.2	70.8	67.7
	RSUGP	82.9	79.6	74.0	67.7
Wine Quality	Orijinal	53.6	44.8	48.7	39.5
	Rasgele Gürültü	50.6	44.8	48.7	35.6
	Gauss Gürültü	50.4	44.8	48.7	34.5
	RSUGP	55.8	44.8	48.7	39.4
Fried	Orijinal	86.4	88.3	83.0	68.8
	Rasgele Gürültü	71.9	69.9	69.2	60.0
	Gauss Gürültü	70.6	68.7	69.2	60.2
	RSUGP	79.7	76.4	73.7	62.4
Bank Management	Orijinal	87.8	90.1	90.0	88.4
	Rasgele Gürültü	87.4	90.0	90.0	88.0
	Gauss Gürültü	85.5	88.9	89.1	88.0
	RSUGP	87.8	90.1	90.0	88.4
Electricity	Orijinal	72.9	87.88	78.9	75.4
	Rasgele Gürültü	70.6	81.2	77.5	70.1
	Gauss Gürültü	60.6	79.3	72.1	64.8
	RSUGP	72.7	87.7	78.8	72.3
FMR Değerleri	Rasgele Gürültü		1,94		
	Gauss Gürültü		1,25		
	RSUGP		2,81		

Tablo 4.3: Önerilen algoritmanın 5-kat çapraz geçerleme sınıflandırma doğruluğu sonuçları

Veri Setleri		5-Kat çapraz geçerleme			
		NB	J48	DT	OneR
Iris	Orijinal	96.0	96.0	92.6	92.6
	Rasgele Gürültü	92.0	83.0	86.6	84.6
	Gauss Gürültü	88.7	87.5	83.5	83.5
	RSUGP	95.6	90.5	87.3	88.5
Heart (statlog)	Orijinal	85.1	77.7	77.0	70.7
	Rasgele Gürültü	84.0	75.7	75.0	70.7
	Gauss Gürültü	76.8	71.2	71.8	70.7
	RSUGP	84.8	78.2	77.7	70.7
Wine Quality	Orijinal	56.7	44.8	47.1	44.0
	Rasgele Gürültü	51.4	44.8	47.0	36.7
	Gauss Gürültü	51.3	44.8	47.0	34.6
	RSUGP	56.7	44.8	47.1	43.5
Fried	Orijinal	86.5	88.9	83.3	69.1
	Rasgele Gürültü	71.9	69.9	69.9	61.2
	Gauss Gürültü	70.7	68.7	69.3	60.5
	RSUGP	79.7	76.9	74.1	62.5
Bank Management	Orijinal	87.9	90.4	90.0	88.5
	Rasgele Gürültü	87.6	90.1	90.0	88.1
	Gauss Gürültü	85.5	89.0	89.1	88.2
	RSUGP	87.9	90.3	90.0	88.5
Electricity	Orijinal	72.9	90.5	79.7	75.8
	Rasgele Gürültü	70.7	82.5	76.1	71.8
	Gauss Gürültü	60.6	81.6	72.7	64.8
	RSUGP	72.6	89.3	79.4	72.8
FMR Değerleri	Rasgele Gürültü			1,92	
	Gauss Gürültü			1,19	
	RSUGP			2,90	

Tablo 4.2, Tablo 4.3 ve Tablo 4.4 incelendiğinde çapraz geçerlemedeki kat sayısını ifade eden k değerinin artması sınıflandırma doğruluk oranlarının artmasına sebep olmaktadır. Yine tablolar incelendiğinde orijinal sınıflandırma doğruluk oranına en yakın sonuçlar neredeyse tüm satırlarda izlendiği üzere önerilen RSUGP modeline aittir.

Önerilen yöntem zaman zaman orijinali ile birebir aynı doğruluk oranına sahipken, bazı yerlerde sapmalar olmaktadır. Tablolardaki bu değerlerin gözle değerlendirilmesi ve modellerin birbirine göre üstünlüklerinin saptanması mümkün görünmemektedir. Bu nedenle her tablonun en alt satırında FMR değerleri bulunmaktadır. FMR sonuçları modellerin birbirine göre durumlarını incelemektedir. Bu değerler kıyaslandığında önerilen RSUGP modelinin en yüksek FMR değerine sahip olduğu, yani diğer yöntemlere göre daha iyi sonuç verdiğini anlamaktayız.

Tablo 4.4: Önerilen algoritmanın 10-kat çapraz geçerleme sınıflandırma doğruluğu sonuçları

Veri Setleri		10-Kat çapraz geçerleme			
		NB	J48	DT	OneR
Iris	Orijinal	96.0	96.0	92.6	92.0
	Rasgele Gürültü	91.3	86.0	88.0	80.0
	Gauss Gürültü	88.8	87.7	83.5	83.5
	RSUGP	95.6	90.8	87.3	88.5
Heart (statlog)	Orijinal	83.7	76.6	82.5	72.5
	Rasgele Gürültü	84.0	75.7	82.0	72.4
	Gauss Gürültü	76.9	72.8	71.8	72.4
	RSUGP	84.4	76.5	82.5	72.5
Wine Quality	Orijinal	56.7	44.8	44.5	45.0
	Rasgele Gürültü	51.1	44.8	44.3	36.8
	Gauss Gürültü	51.1	44.8	44.2	34.7
	RSUGP	56.7	44.8	44.5	44.5
Fried	Orijinal	86.5	89.4	83.4	69.2
	Rasgele Gürültü	71.9	70.0	69.7	61.5
	Gauss Gürültü	71.2	68.8	69.5	60.7
	RSUGP	79.7	77.4	74.0	62.7
Bank Management	Orijinal	88.0	90.3	90.0	88.5
	Rasgele Gürültü	87.6	90.0	90.0	88.1
	Gauss Gürültü	85.5	89.5	89.1	88.2
	RSUGP	88.0	90.2	90.0	88.5
Electricity	Orijinal	72.9	91.1	79.9	76.0
	Rasgele Gürültü	71.3	88.2	76.4	71.5
	Gauss Gürültü	60.6	82.7	73.0	64.7
	RSUGP	72.8	90.2	79.5	72.8
FMR Değerleri	Rasgele Gürültü			1,90	
	Gauss Gürültü			1,21	
	RSUGP			2,90	

Tablodan da görüldüğü gibi önerilen algoritma, mevcut algoritmalara kıyasla Naive Bayes ve J48 sınıflandırma algoritmalarının tüm durumlarında daha iyi veya eşit performans göstermektedir. Karşılaştırılan yöntemlerin Karar Tablosunda sınıflandırıcı sonuçları çoğunlukla aynıdır. OneR sınıflandırıcıda, RSUGP en yakın yöntem olmasına rağmen, RSUGP ile orijinal sonuç arasındaki fark çok daha fazladır. J48 ve Naive Bayes sınıflandırıcı, tüm algoritmalar için sınıflandırma doğruluğu açısından iki sınıflandırıcıdan daha iyidir.



Tablo 4.5: Önerilen modelin sınıflandırma sonuçlarının tek tablo üzerinde gösterimi

Veri Setleri		2-kat Çapraz Geçerleme				5-kat Çapraz Geçerleme				10-kat Çapraz Geçerleme			
		NB	J48	DT	OneR	NB	J48	DT	OneR	NB	J48	DT	OneR
Iris	Orijinal	96.6	93.3	94.0	94.0	96.0	96.0	92.6	92.6	96.0	96.0	92.6	92.0
	Rasgele Gürültü	92.0	84.4	84.4	87.3	92.0	83.0	86.6	84.6	91.3	86.0	88.0	80.0
	Gauss Gürültü	88.6	87.5	83.3	83.3	88.7	87.5	83.5	83.5	88.8	87.7	83.5	83.5
	RSUGP	95.5	90.2	87.3	88.4	95.6	90.5	87.3	88.5	95.6	90.8	87.3	88.5
Heart (statlog)	Orijinal	82.2	80.0	74.0	67.7	85.1	77.7	77.0	70.7	83.7	76.6	82.5	72.5
	Rasgele Gürültü	81.8	80.3	73.8	67.7	84.0	75.7	75.0	70.7	84.0	75.7	82.0	72.4
	Gauss Gürültü	76.5	71.2	70.8	67.7	76.8	71.2	71.8	70.7	76.9	72.8	71.8	72.4
	RSUGP	82.9	79.6	74.0	67.7	84.8	78.2	77.7	70.7	84.4	76.5	82.5	72.5
Wine Quality	Orijinal	53.6	44.8	48.7	39.5	56.7	44.8	47.1	44.0	56.7	44.8	44.5	45.0
	Rasgele Gürültü	50.6	44.8	48.7	35.6	51.4	44.8	47.0	36.7	51.1	44.8	44.3	36.8
	Gauss Gürültü	50.4	44.8	48.7	34.5	51.3	44.8	47.0	34.6	51.1	44.8	44.2	34.7
	RSUGP	55.8	44.8	48.7	39.4	56.7	44.8	47.1	43.5	56.7	44.8	44.5	44.5
Fried	Orijinal	86.4	88.3	83.0	68.8	86.5	88.9	83.3	69.1	86.5	89.4	83.4	69.2
	Rasgele Gürültü	71.9	69.9	69.2	60.0	71.9	69.9	69.9	61.2	71.9	70.0	69.7	61.5
	Gauss Gürültü	70.6	68.7	69.2	60.2	70.7	68.7	69.3	60.5	71.2	68.8	69.5	60.7
	RSUGP	79.7	76.4	73.7	62.4	79.7	76.9	74.1	62.5	79.7	77.4	74.0	62.7
Bank Management	Orijinal	87.8	90.1	90.0	88.4	87.9	90.4	90.0	88.5	88.0	90.3	90.0	88.5
	Rasgele Gürültü	87.4	90.0	90.0	88.0	87.6	90.1	90.0	88.1	87.6	90.0	90.0	88.1
	Gauss Gürültü	85.5	88.9	89.1	88.0	85.5	89.0	89.1	88.2	85.5	89.5	89.1	88.2
	RSUGP	87.8	90.1	90.0	88.4	87.9	90.3	90.0	88.5	88.0	90.2	90.0	88.5
Electricity	Orijinal	72.9	87.88	78.9	75.4	72.9	90.5	79.7	75.8	72.9	91.1	79.9	76.0
	Rasgele Gürültü	70.6	81.2	77.5	70.1	70.7	82.5	76.1	71.8	71.3	88.2	76.4	71.5
	Gauss Gürültü	60.6	79.3	72.1	64.8	60.6	81.6	72.7	64.8	60.6	82.7	73.0	64.7
	RSUGP	72.7	87.7	78.8	72.3	72.6	89.3	79.4	72.8	72.8	90.2	79.5	72.8
FMR Değerleri	Rasgele Gürültü	1,94				1,92				1,90			
	Gauss Gürültü	1,25				1,19				1,21			
	RSUGP	2,81				2,90				2,90			

4.2 F-ÖLÇÜTÜ SONUÇLARI

Önerilen yöntem sınıflandırma doğruluğu haricinde bir de F-ölçütü ile değerlendirilmiştir. Sınıflandırma doğruluğunda olduğu gibi dört farklı sınıflandırıcı kullanılmış, sonuçlar 6 farklı veri seti üzerinde k çapraz geçerleme ile sınanmıştır. 2-kat, 5-kat ve 10-kat çapraz geçerleme kullanılarak değerlendirilen sonuçlar minimum ve maksimum değerler değil de ortalama F-ölçütü sonuçları kullanılarak elde edilmiştir. 2-kat çapraz geçerleme sonuçları Tablo 4.6, 5-kat çapraz geçerleme sonuçları Tablo 4.7 ve 10-kat çapraz geçerleme sonuçları Tablo 4.8 ile verilmiştir.

Tablo 4.6: Önerilen algoritmanın 2-kat çapraz geçerleme için F-ölçütü sonuçları

Veri Setleri		2-kat çapraz geçerleme			
		NB	J48	DT	OneR
Iris	Orijinal	0,967	0,934	0,94	0,94
	Rasgele Gürültü	0,864	0,92	0,92	0,91
	Gauss Gürültü	0,85	0,92	0,88	0,9
	RSUGP	0,94	0,953	0,927	0,926
Heart (statlog)	Orijinal	0,822	0,8	0,741	0,678
	Rasgele Gürültü	0,805	0,765	0,707	0,65
	Gauss Gürültü	0,785	0,774	0,703	0,653
	RSUGP	0,826	0,789	0,793	0,686
Wine Quality	Orijinal	0,536	0,448	0,487	0,395
	Rasgele Gürültü	0,506	0,448	0,487	0,356
	Gauss Gürültü	0,504	0,448	0,487	0,345
	RSUGP	0,558	0,448	0,487	0,394
Fried	Orijinal	0,864	0,883	0,83	0,688
	Rasgele Gürültü	0,719	0,699	0,692	0,6
	Gauss Gürültü	0,706	0,687	0,692	0,602
	RSUGP	0,82	0,883	0,81	0,624
Bank Management	Orijinal	0,88	0,897	0,888	0,857
	Rasgele Gürültü	0,882	0,893	0,882	0,856
	Gauss Gürültü	0,881	0,893	0,885	0,855
	RSUGP	0,88	0,894	0,888	0,857
Electricity	Orijinal	0,729	0,8788	0,789	0,754
	Rasgele Gürültü	0,706	0,812	0,775	0,701
	Gauss Gürültü	0,606	0,793	0,721	0,648
	RSUGP	0,727	0,877	0,788	0,723
FMR Değerleri	Rasgele Gürültü			1,83	
	Gauss Gürültü			1,34	
	RSUGP			2,61	

Tablo 4.7: Önerilen algoritmanın 5-kat çapraz geçerleme için F-ölçütü sonuçları

Veri Setleri		5-kat çapraz geçerleme			
		NB	J48	DT	OneR
Iris	Original	0,96	0,96	0,927	0,927
	Random Noise	0,882	0,93	0,92	0,89
	Gauss Noise	0,88	0,92	0,91	0,87
	RSUGP	0,95	0,953	0,928	0,9
Heart (statlog)	Original	0,852	0,777	0,771	0,707
	Random Noise	0,815	0,765	0,707	0,685
	Gauss Noise	0,812	0,765	0,754	0,703
	RSUGP	0,837	0,778	0,771	0,707
Wine Quality	Original	0,567	0,448	0,471	0,44
	Random Noise	0,514	0,448	0,47	0,367
	Gauss Noise	0,513	0,448	0,47	0,346
	RSUGP	0,567	0,448	0,471	0,435
Fried	Original	0,865	0,889	0,833	0,691
	Random Noise	0,719	0,699	0,699	0,612
	Gauss Noise	0,707	0,687	0,693	0,605
	RSUGP	0,84	0,769	0,82	0,625
Bank Management	Original	0,882	0,9	0,889	0,858
	Random Noise	0,88	0,899	0,885	0,856
	Gauss Noise	0,88	0,898	0,89	0,856
	RSUGP	0,881	0,9	0,89	0,858
Electricity	Original	0,729	0,905	0,797	0,758
	Random Noise	0,707	0,825	0,761	0,718
	Gauss Noise	0,606	0,816	0,727	0,648
	RSUGP	0,726	0,893	0,794	0,728
FMR Değerleri	Random Noise		1,92		
	Gauss Noise		1,45		
	RSUGP		2,85		

Tablo 4.5, Tablo 4.6 ve Tablo 4.7 incelendiğinde önerilen modelin orijinal veri setine en yakın F-ölçütü değerlerine sahip olduğu görülmektedir. Verilerin birbirine çok yakın değerler olması dolayısı ile objektif değerlendirme yapılabilmesi için elde edilen sonuçlar Friedman rank testi ile değerlendirilmiştir. Her bir kat değerlendirme sonuçları ayrı ayrı ölçülmüş ve FMR değerleri her tablonun en alt satırında sunulmuştur. Bu sunulan değerlere göre önerilen RSUGP modeli

en yüksek rank değerine sahiptir. Bu değer yüksek olması sonucun kıyaslanan diğer modellere göre ne kadar fark gösterdiğini ortaya koymaktadır.

Yine tablolar incelendiğinde Naive Bayes ve J48 sınıflandırma sonucunda elde edilen F-ölçütü değerlerinin diğerlerine göre daha iyi sonuç verdiği, Karar tablosunun çok büyük bir fark göstermediği izlenmektedir.

Tablo 4.8: Önerilen algoritmanın 10-kat çapraz geçerleme için F-ölçütü sonuçları

Veri Setleri		10-kat çapraz geçerleme			
		NB	J48	DT	OneR
Iris	Original	0,96	0,96	0,927	0,92
	Random Noise	0,885	0,94	0,92	0,89
	Gauss Noise	0,89	0,94	0,91	0,89
	RSUGP	0,92	0,967	0,953	0,9
Heart (statlog)	Original	0,837	0,767	0,826	0,726
	Random Noise	0,825	0,765	0,707	0,71
	Gauss Noise	0,828	0,745	0,734	0,703
	RSUGP	0,833	0,766	0,826	0,726
Wine Quality	Original	0,567	0,448	0,445	0,45
	Random Noise	0,511	0,448	0,443	0,368
	Gauss Noise	0,511	0,448	0,442	0,347
	RSUGP	0,567	0,448	0,445	0,445
Fried	Original	0,865	0,894	0,834	0,692
	Random Noise	0,719	0,7	0,697	0,615
	Gauss Noise	0,712	0,688	0,695	0,607
	RSUGP	0,86	0,889	0,83	0,627
Bank Management	Original	0,882	0,898	0,889	0,857
	Random Noise	0,88	0,899	0,887	0,856
	Gauss Noise	0,882	0,899	0,888	0,856
	RSUGP	0,882	0,898	0,898	0,857
Electricity	Original	0,729	0,911	0,799	0,76
	Random Noise	0,713	0,882	0,764	0,715
	Gauss Noise	0,606	0,827	0,73	0,647
	RSUGP	0,728	0,902	0,795	0,728
FMR Değerleri	Random Noise	2,20			
	Gauss Noise	1,67			
	RSUGP	2,90			

Veri Setleri	Algoritmalar	ICA _{ort}	ICA _{min}	IO _{ort}	IO _{min}	NI _{ort}	NI _{min}
Iris	Rasgele Gürültü	0,83	0,71	0,34	0,08	1,97	1,95
	Gauss Gürültü	0,78	0,77	0,58	0,5	1,92	1,84
	RSU Gürültü	0,85	0,78	0,48	0,19	1,98	1,96
Heart	Rasgele Gürültü	0,82	0,73	0,06	0,02	1,89	1,83
	Gauss Gürültü	0,75	0,73	0,67	0,55	1,93	1,92
	RSU Gürültü	0,77	0,73	0,65	0,56	1,98	1,96
Wine Quality-White	Rasgele Gürültü	0,72	0,67	0,02	0,46	1,8	1,77
	Gauss Gürültü	0,72	0,72	0,5	0,2	1,89	1,85
	RSU Gürültü	0,75	0,69	0,5	0,03	1,89	1,79
Fried	Rasgele Gürültü	0,71	0,7	0,57	0,39	1,63	1,49
	Gauss Gürültü	0,71	0,66	0,53	0,32	1,64	1,5
	RSU Gürültü	0,74	0,69	0,45	0,4	1,61	1,47
Bank Management	Rasgele Gürültü	0,99	0,99	0,01	0,01	1,95	1,95
	Gauss Gürültü	0,82	0,7	0,53	0,35	1,98	1,97
	RSU Gürültü	0,85	0,71	0,45	0,33	1,99	1,99
Electricity	Rasgele Gürültü	0,8	0,51	0,57	0,51	1,93	1,9
	Gauss Gürültü	0,84	0,77	1,51	0,4	1,91	1,89
	RSU Gürültü	1,08	0,85	1,46	1,4	1,99	1,95
FMR Değerleri	Rasgele Gürültü: 1,64		Gauss Gürültü: 1,97		RSUGürültü: 2,39		

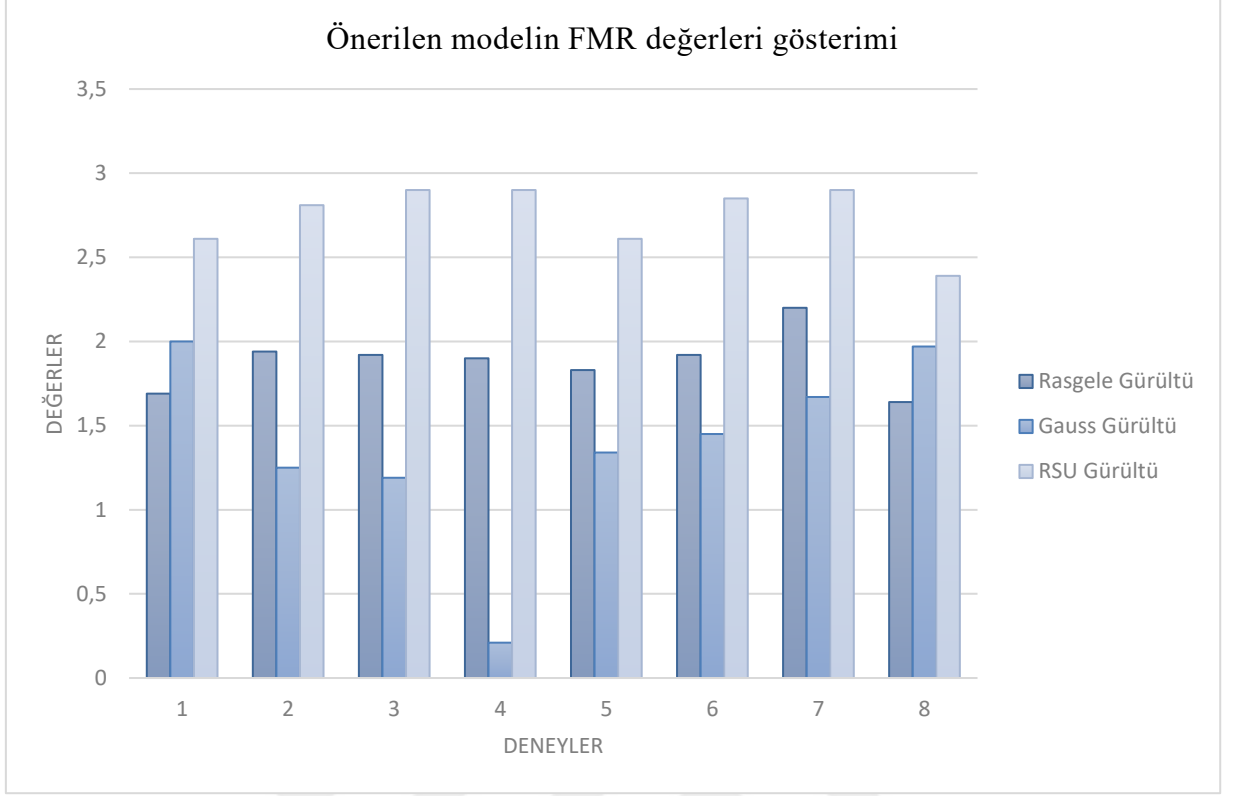
4.4 FRIEDMAN RANK TESTİ SONUÇLARI

Daha önceki bölümlerde anlatıldığı gibi parametrik olmayan değerlerin kıyaslanması için kullanılan istatistiksel sına yöntemlerinden biri de Friedman Rank testidir. Önerilen modelin performans ölçülmesi için sınıflandırma doğruluğu, F- ölçütü ve atak dayanıklılığı kullanılmıştır. Bu parametreler 6 farklı veri seti üzerinde 4 farklı sınıflandırma ve 3 farklı saldırı yöntemi ile test edilmiştir. Elde edilen bu sonuçların birbirlerine göre ne ifade ettiği ne kadar fark ortaya koyduğu Friedman rank testi ile ölçülmüştür. Elde edilen sonuçlar Tablo 4.10 üzerinde gösterilmiştir.

Tablo 4.10: Tüm sonuçların FMR değerleri ile kıyaslanması

FMR uygulanan durumlar	Rasgele Gürültü	Gauss Gürültü	RSU Gürültü
Aynı test ve eğitim seti durumu	1,69	2	2,61
k=2 kat çapraz geçерleme	1,94	1,25	2,81
k=5 kat çapraz geçерleme	1,92	1,19	2,9
k=10 kat çapraz geçерleme	1,9	0,21	2,9
k=2 kat çapraz geçерleme	1,83	1,34	2,61
k=5 kat çapraz geçерleme	1,92	1,45	2,85
k=10 kat çapraz geçерleme	2,2	1,67	2,9
Atak sonuçları	1,64	1,97	2,39

Tablo 4.10 ‘daki değerin yüksek olması, yöntemin diğer yöntemlere göre belirgin bir farklılık gösterdiğini gösterir. Deneyin test istatistiklerinin χ^2 değeri 24, serbestlik derecesi 2 ve p değeri 34.889 idi. Önerilen RSUGP modelinin tüm kıyaslamalarda en yüksek rank değerine sahip olduğu ve kıyaslanan diğer yöntemlere göre belirgin bir fark ortaya koyduğu izlenmektedir. Bu değerin grafik olarak gösterimi Şekil 4.1 de gösterilmiştir.

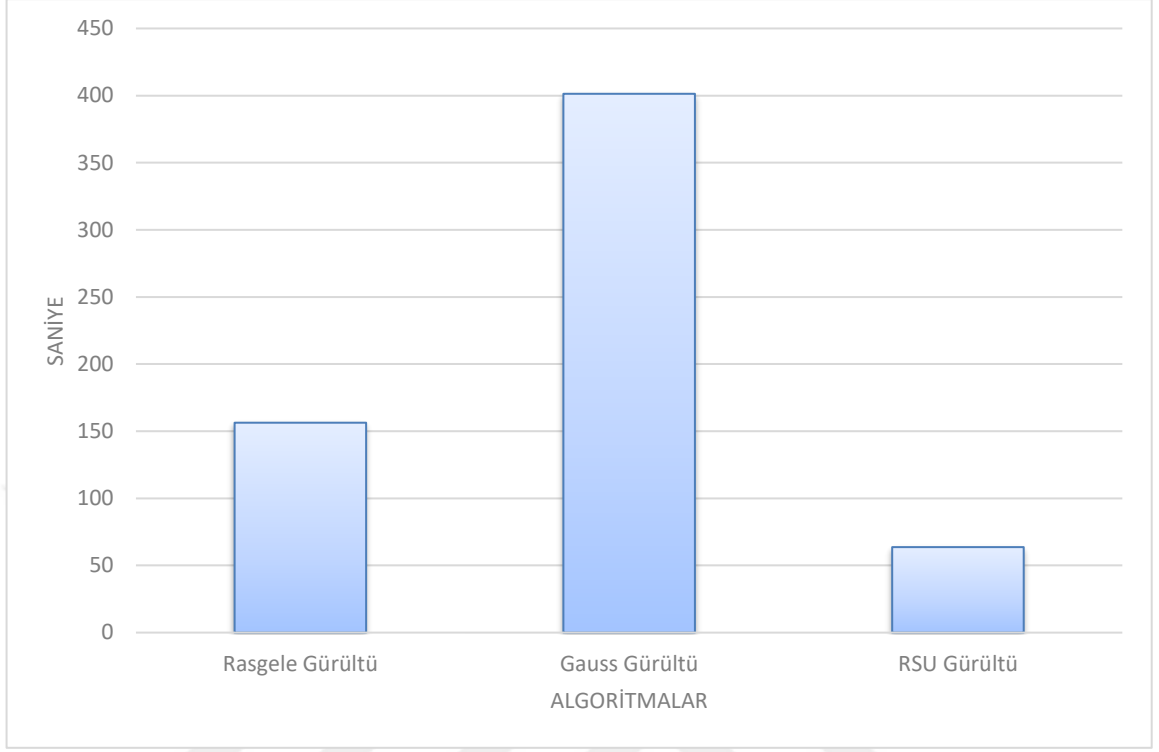


Şekil 4.1: Önerilen modelin FMR değerlerinin kıyaslanması

4.5 ÇALIŞMA ZAMANI

Bu çalışmada 6 farklı boyuttaki veri seti üzerinde 3 farklı yöntem çalıştırılmıştır. Veri setleri üzerinde kimliksizleştirilmek üzere belirlenen hassas nitelik sayısı ile veri setinin büyüklüğü de artınca yöntemlerin çalışma süreleri de ona bağlı olarak artmaktadır.

Gauss gürültü, olasılık yoğunluk fonksiyonu ile tanımlanan rasgele değişkenlerin istatistiksel davranışı olarak tanımlanmaktadır. Grafik incelendiğinde verinin tamamına göre gürültü üreten Gauss modelinin diğerlerine göre daha uzun sürede çalıştığı görülmektedir. Rasgele sayı üreteçlerine bağlı olarak gürültü üreten önerilen model ise gürültü üretmek için verinin tamamını değil de sadece üzerinde bulunduğu hassas niteliğin tamamını belirlenen tekrar sayısı kadar kullanarak gürültüyü üretmektedir. Böylelikle hem verinin kendisinden gürültü üretmekte hem de verinin tamamıyla ilgilenmediği için çok daha kısa sürede çalışmaktadır. Çalıştırılan bu yöntemlerin ortalama çalışma sürelerinin birbirlerine göre durumu Şekil 4.2 teki grafik üzerinde gösterilmiştir.



Şekil 4.2: Çalışma süreleri

5 TARTIŞMA VE SONUÇ

Gelişen teknoloji, sosyal ağlar, nesnelerin interneti kavramlarının hızlı veri oluşturulması ve oluşturulan bu verinin anlamlı kullanımını sağlamaktadır. Özellikle her alanda oluşan pazar rekabeti, bu verilerin işlenmesini ve verilerden kişilere ya da kullanılan alanı bağlı anlamlı bilgiler çıkarımını bir artı özellik olarak değil bir gereklilik haline getirmiştir. Buda verileri günümüz kıymetli hazinesi haline getirmektedir. Kişiler ve kişilerin verileri bu nedenle kıymetlenmesi birçok sektörün gözünü bu veriler üzerinden çıkarılacak anlamlı verilere çekmektedir. Fakat anlamlandırılan ve paylaşılan bu veri her ne olursa olsun şahıslara veya nesnelere aittir ve mahremiyetinin sağlanması gerekmektedir.

Bu sağlanacak olan gizlilik sadece saldırganlardan gelen şifre, kimlik avcılığı olarak düşünülmemelidir, paylaşılmasına ve anlamlandırılmasına izin verilmeyen verilerin gizliliğinin her durumda korunması demektir. Basit düşünmek gerekirse, en ufak bir internet alışverişinde dahi kimlik bilgisi gibi şahıslara özel bilgilerin, ya da bir hastane kaydındaki rahatsızlık bilgilerinin orada çalışan ve konu ile ilgisi olmayan herhangi bir personelin eline geçmemesi ve buna bir erişim kısıtlaması konulması gerekmektedir. Ancak günümüzde web üzerinde yapmış olduğumuz web aramalarından bile ihtiyacımız olan ürün/gereksinim ile ilgili bilgilerin analiz edilip önümüze reklam olarak gelmesi bu gizliliğin ihlal edildiğini gözler önüne sermektedir.

Bu konu sadece kişiler değil devletlerinde sorumluluğu altında olması dolayısı ile ülkeler tarafından çıkarılan kişisel verileri korumaya yönelik kanunlarla bu durumun önüne geçmek hedeflenmektedir. Kanun kapsamında yapılan çalışmalar artık sadece devletlerin çalışma alanı değil, bu korumayı sağlayacak olan teknoloji şirketlerinin de güncel çalışma konuları arasındadır. Bir yandan geliştirilen yöntemlerle birlikte, kullanılan verinin hacim ve hızı gibi değişen özelliklerine karşıda uyumlu geliştirmeler gerekmektedir.

Verilerin kişisel değerlerden çeşitli yöntemlerle arındırılması işlemine veri kimliksizleştirme ya da anonimleştirme denilmektedir ve bu alanda birçok çalışma yapılmaktadır. Burada dikkat edilmesi gereken en önemli nokta veriyi bu bilgilerden arındırıp farklılaştırdığımızda hala eskisiyle aynı değere sahip olması gerektiğidir. Gizlilik korumalı veri madenciliği yöntemlerindeki en önemli parametre gizliliği korunmuş olan verinin, bilgi kaybını minimum seviyede tutmaktır.

Bu tez çalışmasında da gizlilik koruma yöntemlerinden biri olan veri pertürbasyonu için rasgele sayı üreteçlerine bağlı yeni bir geometrik veri pertürbasyon yöntemi sunulmuştur. Rasgele sayı üreteçleri oyun teorisi ve kriptoloji gibi bilgisayar teknolojilerinde hali hazırda başarıyla kullanılan bir üreteçtir. Tanımlandığı aralıklarda rasgele değerler üretmek üzere tasarlanmışlardır. Geometrik veri pertürbasyonu ise nümerik değerlerin matris çarpımı ya da toplamı işlemleriyle değerlerinin değiştirilmesini sağlamaktadır. Birden fazla adımın bulunduğu geometrik veri pertürbasyonu modeli kullanılan adım sayısına göre özelleştirilebilir. Literatürde yapılan çalışmalar incelendiğinde geometrik veri pertürbasyonun en özelleştiren kısmın verilere eklenen gürültü adımı olduğu ancak birçok çalışma aynı gürültü modeli ile çalıştığı görülmüştür. Buradan yola çıkılarak yeni bir gürültü modelinin geometrik veri pertürbasyonuna etkisi fikri oluşmuştur.

Bu çalışma kapsamın önerilen rasgele sayı üreteçlerine bağlı geometrik veri pertürbasyonu RSUGP modeli, verinin anonimleştirme sonrasında kullanılabilirliğini engellemeyecek, hızlı, etkin ve atak dayanıklılığı olan yeni bir modeldir. Önerilen algoritmanın başarılı olmasının nedeni, verinin kendisini ve kendisinden çıkardığı bazı parametreleri kullanarak gürültü üretmesidir. Ayrıca önerilen yöntem veri setinden ve özneliliklerden bağımsız olarak çalışabilmesi ile de farklı verilere kolayca uygulanabilmektedir. Veri setindeki hassas değerleri saklamaya yönelik bir yöntem olduğu için direk sonuç vermektedir. Çalışmanın bunlar haricindeki önemli bir katkısı ise dağıtık veri ortamlarından Hadoop dosya sistemi HDFS üzerinde tutulan verilerin Apache Spark ile işlenmesi ve hızlı sonuçlar elde edilmesidir. Uygun donanımlar sağlandığında önerilen model daha büyük veri setleri için kolayca uygulanabilecektir.

Önerilen bu yöntem farklı boyutlardaki altı farklı veri setler üzerinde farklı adette nümerik hassas değeri seçilerek değerlendirmeye alınmıştır. Model sınıflandırma doğruluğu, F-ölçütü, atak dayanıklılığı, çalışma süresi ve Friedman rank sonuçlarına göre literatürdeki çalışmalar ile kıyaslanmış ve başarılı sonuçlar elde edilmiştir. Elde edilen sonuçlar bulgular kısmında detaylıca paylaşılmıştır. Önerilen yöntemin deneysel sonuçları, modelin etkin gizlilik koruması için rahatlıkla kullanılabilecek bir yöntem olduğunu göstermektedir.

Sınıflandırma sonuçlarına bakıldığında önerilen modelin orijinal sınıflandırma doğruluğuna en yakın doğruluk oranına sahip olduğu görülmektedir. En yakın olmakla birlikte önerilen RSUGP yönteminin kıyaslanan diğer yöntemlere göre ne kadar iyileştirme yaptığının ölçülmesi için

yapılan Friedman rank testi sonuçlarına göre, model 1-1,5 kat daha iyi sonuç vermektedir. İçeriği tamamen nümerik hassas değerlerden oluşan veri setleri için önerilen modelin sınıflandırma sonucu doğruluğu orijinale en yakın olmasına rağmen, yakınlık miktarı farklı alanlara sahip olan veri setlerine göre bir miktar fazladır. Sınıflandırıcılar açısından değerlendirmek gerekirse, J48 ve Naive Bayes sınıflandırıcılarda elde edilen sonuçlar önerilen modelin neredeyse tüm değerlendirmelerinde en yüksek doğruluk oranına sahiptir. Karar Tablosu sınıflandırıcılar ise kıyaslanan 3 modelde yaklaşık olarak aynı sonucu ifade ettiği için, değerlendirme açısından belirgin bir fark ifade etmemektedir. OneR sınıflandırıcı modelinde ise önerilen model orijinale en yakın doğruluk oranına sahip olmasına rağmen orijinale uzaklığı diğer sınıflandırıcılara göre bir miktar fazladır. Buradan yola çıkarak söylenebilir ki önerilen model için en başarılı sınıflandırıcılar J48 ve Naive Bayes sınıflandırıcılardır.

Atak dayanıklılığı sonuçları incelendiğinde hem minimum dayanıklılık hem de ortalama dayanıklılık açısından önerilen model kıyaslanan diğer yöntemlere göre çok daha dayanıklıdır. Sonuçlar tablosuna bakıldığı zaman kıyaslanan tüm yöntemler, Naif Girişim(Naive Inference) saldırısına karşı yüksek dayanıklılığa sahiptir. Yine tüm modeller açısından değerlendirmek gerekirse Girdi/Çıktı Saldırılarına(I/O Attack) karşı geometrik veri pertürbasyonun yeterince dayanıklı olmadığı görülmektedir. Ancak önerilen model ile saldırı direnci bir miktar güçlendirilmiştir.

Çalışma zamanları açısından kıyaslanan yöntemler incelendiğinde önerilen modelin Gauss gürültü ve rasgele gürültü kullanan modellere göre çok daha kısa sürede sonuç aldığı görülmektedir. Özellikle Gauss gürültü kullanan yöntemlerde gürültü üretme süresinin oldukça uzun olduğu tespit edilmiştir. Önerilen RSUGP yönteminde hem diğer değerlendirme sonuçlarında oldukça başarılı değerler edilen bu yöntemin, çalışma zamanı açısından da başarısı ortaya konulmuştur. Gauss gürültü modelinin verinin kendisini kullanarak gürültü üretmesi, rasgele gürültü üretmeninde basit ve hızlı hesaplanabilirlik avantajlarını kullanması dolayısı ile, önerilen modelden bu sonuçlar elde edilmiştir.

Ayrıca alınan bu hızlı sonuçlar Apache Spark'ın verileri hızlı işleme yeteneği ile elde edilmiştir. Tez kapsamında Apache Spark üzerinde tek ana düğüm ve tek işçi düğüm ile verileri işlemesi sağlamış herhangi bir paralel çalışma modeli kullanılmamıştır. Gerekli donanımlar sağlanarak kurulacak paralel veri işleme modeli ile çok daha büyük boyutlu verilerin çok daha kısa süreler içinde işlenebileceği düşünülmektedir. Bu çalışmanın ileriki adımlarında gerekli donanımsal

gereklilikler sağlanarak Apache Spark'ın paralel veri işleme performansı ile çok daha büyük verileri hızlı ve etkin bir şekilde anonimleştirmesi planlanmaktadır.

Çalışma geometrik veri pertürbasyonunda daha önce kullanılmamış bir gürültü modelinin başarısını ortaya koymakla birlikte bu alanda yapılabilecek farklı çalışmalara da yön vermektedir. Bu noktada kaotik haritalar ya da donanımsal ya da yazılımsal elde edilebilecek farklı rasgelelik mekanizmalarının kullanılması gelecek çalışmalar için hedeflenmektedir.



KAYNAKLAR

- [1]. Derinözlü, C., 2017, Büyük Veri ve Mahremiyet Big Data and Privacy, *TUBITAK 1.Ulusal Bulut Bilişim ve Büyük Veri Sempozyumu B3S17 Sempozyumu Bildiri Kitabı*
- [2]. Chen, Chiang and Storey, 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), p.1165.
- [3]. Internet Live Stats , 2021. [online] Available at: <https://www.internetlivestats.com/twitter-statistics/> [Ziyaret Tarihi: 28.05.2019].
- [4]. AKTAN, E., 2018. Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu. *Bilgi Yönetimi*, 1(1), pp.1-22.
- [5]. Zafar, R., Yafi, E., Zuhairi, M. F. ve Dao, H. , 2016, Big Data: The NoSQL and RDBMS Review. *2016 International Conference on Information and Communication Technology (ICICTM)*, Kuala Lumpur, Malaysia, 120-126
- [6]. Canbay, Y., Vural, Y. and Sağiroğlu, Ş., 2019, Mahremiyet Korumalı Büyük Veri Yayınlama İçin Kavramsal Model Önerileri. *Journal of Polytechnic*.
- [7]. Xuyun Zhang, Yang, L., Chang Liu and Jinjun Chen, 2014. A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud. *IEEE Transactions on Parallel and Distributed Systems*, 25(2), pp.363-373.
- [8]. Anonim, 2021. *Pseudonymization / Anonymization.*, Prime Factors. Available at: <https://www.primefactors.com/solutions/pseudonymization-anonymization/> [Ziyaret Tarihi: 01.08.2021].
- [9]. Anonim, 2021. *Libelle DataMasking | GDPR-compliant data.* [online] En.libelle.com. Available at: <https://en.libelle.com/products/datamasking> [Ziyaret Tarihi: 01.08.2021].
- [10]. Anonim, 2021. *ARX - Data Anonymization Tool | A comprehensive software for privacy-preserving microdata publishing.* [online] Arx.deidentifier.org. Available at: <https://arx.deidentifier.org/> [Ziyaret Tarihi: 01.08.2021].
- [11]. Manolis Terrovitis, D., 2021. *Amnesia Anonymization Tool - Data anonymization made easy.* [online] Available at: <https://amnesia.openaire.eu/> [Ziyaret Tarihi: 01.08.2021].
- [12]. Anonim, 2021. *Data anonymization Tool for Sensitive Data Protection | BizDataX.* [online] BizDataX. Available at: <https://bizdatax.com/data-anonymization/> [Ziyaret Tarihi: 01.08.2021].

- [13]. Kişisel Verileri Koruma Kurumu, (2018), *Kişisel Verilerin Korunması Kanununa İlişkin Uygulama Rehberi*. ,Ankara: KVKK Yayınları. Sayfa:11
- [14]. T.C. Resmî Gazete. Kişisel verilerin otomatik işleme tabi tutulması karşısında bireylerin korunması sözleşmesi., 30.1.2016. , Sayı: 2016/8576, *Başbakanlık Basımevi*, Ankara.
- [15]. Aktan, E. ,2018,. Büyük veri: Uygulama alanları, analitiği ve güvenlik boyutu. *Bilgi Yönetimi*, 1(1), 1-22.
- [16]. Ranjan, A., and Ranjan, P., 2016, April, Two-phase entropy based approach to big data anonymization. In *2016 International Conference on Computing, Communication and Automation (ICCCA)* ,76-81.
- [17]. Sweeney, L., 1997, Guaranteeing anonymity when sharing medical data, the Datafly System, *AMIA Annual Fall Symposium*, American Medical Informatics Association, 51-55.
- [18]. Samarati, P. and Sweeney, L., 1998, *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression*, Technical Report, SRI International, 101-132.
- [19]. Sweeney, L., 2002a, Achieving *k*-anonymity privacy protection using generalization and suppression, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10 (5), 571-588.
- [20]. Xiao, X., and Tao, Y., 2006, Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* ,2016 July, 229-240
- [21]. Yao, L., Chen, Z., Wang, X., Liu, D., and Wu, G., 2019, Sensitive label privacy preservation with anatomization for data publishing. *IEEE Transactions on Dependable and Secure Computing*.
- [22]. El Ouazzani, Z., and El Bakkali, H. ,2017, A new technique ensuring privacy in big data: Variable t-closeness for sensitive numerical attributes. In *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)* , October 2017, 1-6.
- [23]. Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., Theodoridis, Y. , 2004, State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1), 50-57.

- [24]. Kişisel Verilerin Silinmesi, Yok Edilmesi veya Anonim Hale Getirilmesi Rehberi, *Kişisel Verileri Koruma Kurumu- KVKK*, Kasım 2017, Ankara - ISBN : 978-975-19-6807-4
- [25]. Samarati, P. and Sweeney, L., 1998, Generalizing data to provide anonymity when disclosing information. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART (PODS)*. ACM, 1998a, New York, 188
- [26]. Fung, B. C., Wang, K., Chen, R., and Yu, P. S., 2010, Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4), 1-53.
- [27]. Chen, T., Lee, W., Chen, J., Kao, Y. and Hou, P., 2013. Reversible privacy preserving data mining: a combination of difference expansion and privacy preserving. *The Journal of Supercomputing*, 66(2), pp.907-917.
- [28]. Kabir ,S.M.A, Youssef, A.M., Elhakeem A.K., 2007, On data distortion for privacy preserving data mining. *In: Proceedings of the 20th Canadian conference on electrical and computer engineering*, 308–311, doi:10.1109/CCECE.2007.83
- [29]. Xiao-Dan, W. U., Dian-Min, Y. U. E., Feng-Li, L. I. U., Yun-feng, W., and Chao-Hsien, C., 2006, Privacy preserving data mining algorithms by data distortion. *In 2006 International Conference on Management Science and Engineering*, October 2016, 223-228
- [30]. Zhu, D., Li, X. and Wu, S., 2009. Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. *Decision Support Systems*, 48(1), 133-140.
- [31]. Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkatasubramanian, M., 2007, *l*-diversity: Privacy Beyond *k*-Anonymity, *ACM Transactions on Knowledge Discovery from Data*, 1 (1), 3.
- [32]. Li, N., Li, T. and Venkatasubramanian, S., 2007, *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity, *IEEE 23rd International Conference on Data Engineering (ICDE 2007)*, 15-20 April 2007, Istanbul, Turkey, 106-115.
- [33]. Roy, D. and Kumar Jena, S., 2013. Determining *t* in *t*-closeness using Multiple Sensitive Attributes. *International Journal of Computer Applications*, 70(19), 47-51.
- [34]. Soria-Comas, J., Domingo-Ferrer, J., Sanchez, D. and Martinez, S., 2015. *t*-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11), 3098-3110.

- [35]. Domingo-Ferrer., J., On the Connection between t-Closeness and Differential Privacy for Data Releases, 2013, *10th Int. Conf. Security Cryptography (ICETE 2013)*, 478–481.
- [36]. Domingo-Ferrer, J. and Soria-Comas, J., 2015. From t-closeness to differential privacy and vice versa in data anonymization, *Knowledge-Based Systems*, 74, pp.151-158.
- [37]. Aldeen, Y. A. A. S., Salleh, M., and Razzaque, M. A., 2015, A comprehensive review on privacy preserving data mining, *SpringerPlus*, 4(1), 1-36.
- [38]. Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., & Verykios, V., 1999, Disclosure limitation of sensitive rules. In *Proceedings 1999 Workshop on Knowledge and Data Engineering Exchange (KDEX'99)*(Cat. No. PR00453), 1999, November, 45-52
- [39]. Saygin, Y., Verykios, V. and Clifton, C., 2001. Using unknowns to prevent discovery of association rules. *ACM SIGMOD Record*, 30(4), 45-54.
- [40]. Verykios, V., 2013. Association rule hiding methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 28-36.
- [41]. Chamikara, M., Bertok, P., Liu, D., Camtepe, S. and Khalil, I., 2018. Efficient data perturbation for privacy preserving and accurate data stream mining. *Pervasive and Mobile Computing*, 48, 1-19.
- [42]. Muralidhar, K., Parsa, R. and Sarathy, R., 1999. A General Additive Data Perturbation Method for Database Security. *Management Science*, 45(10), pp.1399-1415, doi: 10.1287/mnsc.45.10.1399
- [43]. Agrawal, R. and Srikant, R., 2000, Privacy-preserving data mining, *ACM SIGMOD Record*, 29 (2), 439-450.
- [44]. Tang, J., Korolova, A., Bai, X., Wang, X., and Wang, X. Privacy Loss in Apple's Implementation of Differential Privacy on macOS 10.12. arXiv:1709.02753 (2017). <https://arxiv.org/abs/1709.02753>.
- [45]. Dalenius, T. and Reiss, S., 1982. Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6(1), 73-85, doi: 10.1016/0378-3758(82)90058-1.
- [46]. Fienberg S., McIntyre J.: Data Swapping: Variations on a Theme by Dalenius and Reiss. *Technical Report, National Institute of Statistical Sciences*, 2003.
- [47]. Hasan, A., Jiang, Q., Luo, J., Li, C. and Chen, L., 2016. An effective value swapping method for privacy preserving data publishing. *Security and Communication Networks*, 9(16), 3219-3228, doi: 10.1002/sec.1527.

- [48]. Estivill-Castro, V., and Brankovic, L. 1999, Data swapping: Balancing privacy against precision in mining for logic rules. In *International Conference on Data Warehousing and Knowledge Discovery*, Springer, Berlin, Heidelberg. 1999, August, 389-398.
- [49]. Aggarwal, C. C. and Yu, P. S., 2004, A condensation approach to privacy preserving data mining, In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, 183-199.
- [50]. Chen K., and Liu L., 2005, A random rotation perturbation approach to privacy preserving data classification, In: *Proceedings of international conference on data mining (ICDM)* Houston, 2005.
- [51]. Chen, K., and Liu, L. ,2005, Privacy preserving data classification with rotation perturbation. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, 4-, doi: 10.1109/ICDM.2005.121
- [52]. Z. Lin, J. Wang, L. Liu and J. Zhang, Generalized random rotation perturbation for vertically partitioned data sets, *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 2009,159-162, doi: 10.1109/CIDM.2009.4938644.
- [53]. Li, F., Zhang, R., Xu, Y., Liu, J., and Li, J.,2016, Privacy preservation based on rotation perturbation in weighted social networks. In *2016 16th International Symposium on Communications and Information Technologies (ISCIT)*, 2016 September ,206-211.
- [54]. Kadampur, M. A., and Somayajulu, D. V.,2010, Privacy preserving technique for Euclidean distance based mining algorithms using a wavelet related transform. In *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, Berlin, Heidelberg, 2010 September, 202-209.
- [55]. Liu, K., Kargupta H. and Ryan J., 2016, Random projection-based multiplicative data perturbation for privacy preserving distributed data mining, In *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92-106, January 2006, doi: 10.1109/TKDE.2006.14.
- [56]. Chen, K. and Liu, L., 2010, Geometric data perturbation for privacy preserving outsourced data mining., *Knowledge and Information Systems*, 29(3), 657-695. Doi: 10.1007/s10115-010-0362-4
- [57]. Patel, N., and Patel, S., 2016, Geometric Data Perturbation Techniques in Privacy Preserving On Data Stream Mining.
- [58]. Lindell and Pinkas, 2002, Privacy Preserving Data Mining. *Journal of Cryptology*, 15(3), 177-206, doi: 10.1007/s00145-001-0019-2

- [59]. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., and Zhu, M. Y., 2002, Tools for privacy preserving data mining, *SIGKDD Explorations* 4, 2, 28-34.
- [60]. Kantarcioglu, M., and Clifton, C., 2004, Privately computing a distributed k-nn classifier, *In European conference on principles of data mining and knowledge discovery*, 2004 September, Springer, Berlin, Heidelberg ,279-290.
- [61]. Vaidya, J., & Clifton, C.,2004, Privacy preserving naive bayes classifier for vertically partitioned data, *In Proceedings of the 2004 SIAM international conference on data mining* Society for Industrial and Applied Mathematics , 2004 April, 522-526.
- [62]. Kavitha S., Yamini S. and Raja Vadhana P., An evaluation on big data generalization using k-Anonymity algorithm on cloud, *2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO)*, 2015, 1-5, doi: 10.1109/ISCO.2015.7282237.
- [63]. Zhang, X., Liu, C., Nepal, S., Yang, C., Dou, W. and Chen, J., 2014, A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud, *Journal of Computer and System Sciences*, 80(5), 1008-1020, doi: 10.1016/j.jcss.2014.02.007.
- [64]. U. Sopaoglu and O. Abul, A top-down k-anonymization implementation for apache spark, *2017 IEEE International Conference on Big Data (Big Data)*, 2017, 4513-4521, doi: 10.1109/BigData.2017.8258492.
- [65]. Madan, S., and Goswami, P, 2018, A privacy preserving scheme for big data publishing in the cloud using k-anonymization and hybridized optimization algorithm, *2018 international conference on circuits and systems in digital enterprise technology (ICCSDET)* , 2018 December, 1-7.
- [66]. Iwuchukwu, T., DeWitt, D. J., Doan, A., and Naughton, J. F, 2007, K-anonymization as spatial indexing: Toward scalable and incremental anonymization, *2007 IEEE 23rd International Conference on Data Engineering*, 2007 April,1414-1416.
- [67]. K. Liu, C. Kuo, W. Liao and P. Wang, Optimized Data de-Identification Using Multidimensional k-Anonymity, 2018, *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*,1610-1614, doi: 10.1109/TrustCom/BigDataSE.2018.00235.
- [68]. Amiri, F., Yazdani, N., Shakery, A. and Chinaei, A., 2016, Hierarchical anonymization algorithms against background knowledge attack in data releasing, *Knowledge-Based Systems*, 101, 71-89, doi: 10.1016/j.knosys.2016.03.004.

- [69]. Eyupoglu, C., Aydin, M., Zaim, A. and Sertbas, A., 2018, An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques, *Entropy*, 20(5), 373, doi: 10.3390/e20050373.
- [70]. Y. Canbay and S. Sağiroğlu, Big data anonymization with spark, *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, 833-838, doi: 10.1109/UBMK.2017.8093543.
- [71]. Nandini Prasaad K.S. and Pratheek T.R., Providing anonymity using top down specialization on Big Data using hadoop framework, *2015 Annual IEEE India Conference (INDICON)*, 2015,1-6, doi: 10.1109/INDICON.2015.7443205.
- [72]. X. Zhang et al., Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud, *IEEE Transactions on Computers*, 1 Aug. 2015, vol. 64, no. 8, 2293-2307, doi: 10.1109/TC.2014.2360516.
- [73]. Andrew, J., Karthikeyan, J., & Jebastin, J. , 2019, Privacy preserving big data publication on cloud using mondrian anonymization techniques and deep neural networks, *2019 5th international conference on advanced computing & communication systems (ICACCS)*, 2019 March, 722-727).
- [74]. Nayahi, J. and Kavitha, V., 2017, Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop, *Future Generation Computer Systems*, 74, 393-408, doi:10.1016/j.future.2016.10.022
- [75]. N. Guo, M. Yang, Q. Gong, Z. Chen and J. Luo, Data Anonymization Based on Natural Equivalent Class, 2019, *IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 22-27, doi: 10.1109/CSCWD.2019.8791905.
- [76]. K. Chen and L. Liu, Privacy-Preserving Multiparty Collaborative Mining with Geometric Data Perturbation, *IEEE Transactions on Parallel and Distributed Systems*, December 2009, vol. 20, no. 12, 1764-1776, , doi: 10.1109/TPDS.2009.26.
- [77]. Balasubramaniam, S. and Kavitha, V., 2015, Geometric Data Perturbation-Based Personal Health Record Transactions in Cloud Computing, *The Scientific World Journal*, 2015, 1-9, doi: 10.1155/2015/927867
- [78]. Reddy, V. and Rao, B., 2018, A Combined Clustering and Geometric Data Perturbation Approach for Enriching Privacy Preservation of Healthcare Data in Hybrid Clouds, *International Journal of Intelligent Engineering and Systems*, 11(1), 201-210, doi:10.22266/ijies2018.0228.21.

- [79]. Darshna R., Avani J., Geometric Data Perturbation Using Clustering Algorithm, 2015, *International Journal Of Advances In Cloud Computing And Computer Science (IJACCCS)*, volume-1,issue-1, 2454-4078.
- [80]. S. S. S. Dhiraj, A. M. A. Khan, W. Khan and A. Challagalla, 2009, Privacy preservation in k-means clustering by cluster rotation, *TENCON 2009 - 2009 IEEE Region 10 Conference*, , 1-7, doi: 10.1109/TENCON.2009.5396140.
- [81]. T. Javid and M. K. Gupta, 2019, Privacy Preserving Classification using 4-Dimensional Rotation Transformation, *8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 279-284, doi: 10.1109/SMART46866.2019.9117391.
- [82]. K. Sreekumar and E. Baburaj, 2012, Privacy preservation using Geometric Data perturbation and fragmentation approach in Wireless Sensor Networks,*IET Chennai 3rd International on Sustainable Energy and Intelligent Systems (SEISCON 2012)*, 1-4, doi: 10.1049/cp.2012.2184.
- [83]. Oliveira, S. R., and Zaiane, O. R., 2010, Privacy preserving clustering by data transformation, *Journal of Information and Data Management*, 1(1), 37-37.
- [84]. Chamikara, M., Bertok, P., Liu, D., Camtepe, S. and Khalil, I., 2020, Efficient privacy preservation of big data for accurate data mining, *Information Sciences*, 527, 420-443, doi:10.1016/j.ins.2019.05.053
- [85]. Chamikara, M., Bertok, P., Liu, D., Camtepe, S. and Khalil, I., 2019, An efficient and scalable privacy preserving algorithm for big data and data streams, *Computers & Security*, 87, 101570, doi: 10.1016/j.cose.2019.101570
- [86]. Sopaoğlu, U., 2020, *Büyük veri ve akan verinin mahremiyet korumalı anonimleştirilmesi.*, Thesis (PhD), TOBB Ekonomi ve Teknoloji Üniversitesi.
- [87]. Eyüpoğlu, C. , 2018, *Büyük veride etkin gizlilik koruması için yazılım tasarımı*, Thesis (PhD), İstanbul Üniversitesi
- [88]. Sweeney, L., 2002b, k-Anonymity: A model for Protecting Privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10 (5), 557-570.
- [89]. Iyengar, V. S., 2002, Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, July 2002, 279-288.

- [90]. B. C. M. Fung, K. Wang and P. S. Yu, 2005, Top-down specialization for information and privacy preservation, *21st International Conference on Data Engineering (ICDE'05)*, 205-216, doi: 10.1109/ICDE.2005.143.
- [91]. E. Bertino, B. C. Ooi, Y. Yang and R. H. Deng, 2005, Privacy and ownership preserving of outsourced medical data, *21st International Conference on Data Engineering (ICDE'05)*, 521-532, doi: 10.1109/ICDE.2005.111.
- [92]. LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., 2006, Mondrian multidimensional kanonymity, *22nd International Conference on Data Engineering (ICDE'06)*, 3-7 April 2006, Atlanta, GA, USA, 25.
- [93]. Chen, K., Sun, G., & Liu, L. , 2007, Towards attack-resilient geometric data Perturbation, *In proceedings of the 2007 SIAM international conference on Data mining*, 2007 April, 78-89, Society for Industrial and Applied Mathematics.
- [94]. Berman, J.J., 2013, *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*, Morgan Kaufmann, Waltham, MA, USA, ISBN: 978-0-12-404576-7
- [95]. Hadoop, 2018. Apache Hadoop. [online] Hadoop.apache.org. Available at: <http://hadoop.apache.org/> [Ziyaret Tarihi: 1 Ağustos. 2021].
- [96]. Sadun, L., 2008, *Applied linear algebra.*, Oxford: American mathematical Society.
- [97]. Vempala, S., 2004, *The random projection method*. Providence, R.I: American Mathematical Society.
- [98]. Jiang, T., 2006, How many entries of a typical orthogonal matrix can be approximated by independent normals?. *The Annals of Probability*, 34(4), 1497-1529.
- [99]. ARSLAN TUNCER, S. and GENÇ, Y., 2019, İnsan Hareketleri Tabanlı Gerçek Rasgele Sayı Üretimi. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 8(1), 261-269, doi: 10.17798/bitlisfen.445548
- [100]. F. Özkaynak, 2015, Kriptolojik Rasgele Sayı Üreteçleri , *Tbv - Türkiye Bilişim Vakfı Bilgisayar Bilimleri Ve Mühendisliği Dergisi* , Cilt : 8 - Sayı : 2, 37-44.
- [101]. Okkalioglu, B., Okkalioglu, M., Koc, M. and Polat, H., 2015, A survey: deriving private information from perturbed data. *Artificial Intelligence Review*, 44(4), 547-569, doi: 10.1007/s10462-015-9439-5
- [102]. Anonim, 2021. *Friedman Test in SPSS Statistics - How to run the procedure, understand the output using a relevant example | Laerd Statistics.* [online]

Statistics.laerd.com. Available at: <https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php> [Ziyaret Tarihi: 23 Mayıs 2021].

- [103]. Anonim, 2017. *Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/> [Ziyaret Tarihi 8 Ağustos. 2021].
- [104]. Anonim, 2017. *Information theory - Wikipedia*. [online] En.wikipedia.org. Available at: https://en.wikipedia.org/wiki/Information_theory [Ziyaret Tarihi: 7 Ağustos. 2021].
- [105]. Giannella, C., Liu, K. and Kargupta, H., 2013, Breaching Euclidean distance-preserving data perturbation using few known inputs, *Data & Knowledge Engineering*, 83, 93-110, doi: 10.1016/j.datak.2012.10.004
- [106]. Gert, H., 2005, The FastICA package for MATLAB. <http://www.cis.hut.fi/projects/ica/fastica/>.

Kanmaz, M., and Aydin, M. A. , 2017, Indoor localization for wireless sensor network and DV-hop, *PressAcademia Procedia*, 5(1), 88-95.

Yiner Z. , Kanmaz M. , Aydın M. A., 2017, Optik Ağlar Anahtarlama Yöntemleri ve Enerji Tüketimi, *8th International Advanced Technologies Symposium(IATS'17)*, Elazığ, Türkiye, 19 -22 Ekim 2017.

Kanmaz M., 2017, Arttırılmış Gerçeklik ve Uygulama Alanları, *Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı*, Antalya, Türkiye, 5 -08 Ekim 2017

Kanmaz M.,Turna Ö.C. ,Aydın M.A, 2011, Pasif Optik Ağlar ve WDM-PON Gelişimi , *Akademik Bilişim* ,Malatya, Türkiye, 2 -04 Şubat 2011, 295-300

