T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

PERTÜBASYON YÖNTEMİ İLE HASSAS VERİ GÜVENLİĞİNE YÖNELİK ÇOK DEĞİŞKENLİ VERİLER İÇİN TAHMİN ANALİZİ

YÜKSEK LİSANS TEZİ

İlker İLTER

Endüstri Mühendisliği Anabilim Dalı

Mühendislik Yönetimi Bilim Dalı

HAZİRAN 2023

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

PERTÜBASYON YÖNTEMİ İLE HASSAS VERİ GÜVENLİĞİNE YÖNELİK ÇOK DEĞİŞKENLİ VERİLER İÇİN TAHMİN ANALİZİ

YÜKSEK LİSANS TEZİ

İlker İLTER

Endüstri Mühendisliği Anabilim Dalı

Mühendislik Yönetimi Bilim Dalı

Tez Danışmanı: Doç. Dr. Safiye TURGAY

HAZİRAN 2023

İlker İLTER tarafından hazırlanan "Pertübasyon Yöntemi İle Hassas Veri Güvenliğine Yönelik Çok Değişkenli Veriler İçin Tahmin Analizi ." adlı tez çalışması 12.06.2023 tarihinde aşağıdaki jüri tarafından oy birliği ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı Mühendislik Yönetimi Bilim Dalı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı :	**Doç. Dr. Merve CENGİZ TOKLU** Sakarya Üniversitesi	.....
Jüri Üyesi :	**Doç. Dr. Safiye TURGAY** Sakarya Üniversitesi	.....
Jüri Üyesi :	**Dr. Öğr. Fatih VARÇIN** Sakarya Uygulamalı Bilimler Üniversitesi	.....

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "PERTÜBASYON YÖNTEMİ İLE HASSAS VERİ GÜVENLİĞİNE YÖNELİK ÇOK DEĞİŞKENLİ VERİLER İÇİN TAHMİN ANALİZİ" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete' de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

12/06/2023

İlker İLTER

TEŞEKKÜR

Yüksek lisan eğitimim süresince, değerli bilgi ve deneyimlerini esirgemeyen, her daim teşvik eden ve titizlikte beni yönlendiren değerli danışman hocam Doç. Dr. Safiye Turgay'a minnet ve şükranlarımı sunarım.

İlker İLTER

İÇİNDEKİLER

Sayfa

KISALTMALAR

CA : Sınıflandırma doğruluğu

EM : Eksponansiyel mekanizma

FC : Keskinlik ve duyarlılığın harmonik ortalaması

FMR : Friedman sıralama testi değeri

FP : Yanlış pozitif tahmin sayısı

FN : Yanlış negatif tahmin sayısı

GBM : Gradyanı artırılmış makineleri yöntemi

GM : Gauss mekanizması

ID : Tanımlayıcı nitelik

KNN : K en yakın komşu yöntemi

LM : Laplace mekanizması

LightGBM : Hafif gradyanı artırılmış makineleri yöntemi

LR : Lojistik regresyon yöntemi

NN : Sinir ağları yöntemi

NSA : Hassas olmayan nitelik

P : Hassasiyet, pozitif olarak tahmin edilenlerin gerçekteki pozitif olma oranı

QID : Yarı tanımlayıcı nitelik

NSA : Hassas olmayan nitelik

R : Duyarlılık, gerçek pozitif doğru şekilde tespit edilme sayısı

RDP : Rotasyon temelli veri bozma yöntemi

SA : Hassas nitelik

SPSS : İstatistiksel analiz programı

STD : Standart sapma

SVM : Destek vektör makineleri

TN : Doğru negatif tahmin sayısı

TP : Doğru pozitif tahmin sayısı

XGBoost : Gradyanı artırılan karar ağaçları yöntemi

TABLO LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

PERTÜBASYON YÖNTEMİ İLE HASSAS VERİ GÜVENLİĞİNE YÖNELİK ÇOK DEĞİŞKENLİ VERİLER İÇİN TAHMİN ANALİZİ

ÖZET

Veri madenciliği, büyük verilerin, nesnelerin interneti (loT) yaygın kullanımı ile imalattan, sağlığa, adaletten, bankacılığa kadar tüm sektörlerde veri gizliliği kavramını gündeme getirmektedir. Veri madenciliği, büyük verilerden ilginç ve önceden bilinmeyen bilgileri keşfetme sürecidir. Dolayısıyla verilerin işlenmesi sürecinde veri gizliğinin sağlanması gerekliliği de ortaya çıkmaktadır. Veri gizliliğinin korunması ile veri madenciliğinin gerçekleştirilmesi daha sağlıklı ve etkin bir ortamda veri analizi sağlayacaktır. Bu çalışmada statik veya dinamik ortamlardaki veri; pertürbasyon yöntemleri ile gizliliği sağlanacak, öncesi ve sonrası verilerin çok boyutlu yapısına dikkate alarak analizleri yapılacaktır.

Teknolojinin gelişimi ile büyük veri kullanımı da artan bir hızla yaygınlaşmaktadır. Verilerin depolanması, analiz edilmesi ve gizliliğinin sağlanması konuları, geliştirilmesi gereken algoritma yöntemlerini de beraberinde getirmiştir. Gizlilik koruması veri bozulması, kayıt gizliliğinin korunması, veri tabanında yer alan değerlerin anlamını ve değişkenler arasındaki ilişkiyi bozmadan saklama tekniğidir. Yarı tanımlı ve hassas sayısal verilerin gizliliğinin korunmasına esas alan bu çalışmada pertürbasyon yöntemlerine yer verilmiştir.

Pertürbasyon yöntemleri, verilerin gizliliğini korurken veri analizine olanak sağlamak için kontrollü gürültü veya rastgelelik eklemek için kullanılan matematiksel tekniklerdir. Rastgele yanıtlama, farklılaştırılmış gizlilik, güvenli çoklu taraf hesaplama, gürültü eklemesi, örnekleme ve birleştirme gibi çeşitli yöntemler, hassas bilgilerin ifşa edilmesini veya istismarını engellemek için kullanılır. Bu yöntemler, makine öğrenimi, istatistik ve kriptografi alanlarında veri gizliliğini sağlamak için başarılı bir şekilde uygulanmaktadır. Bununla birlikte, uygulama dikkatli bir şekilde tasarlanmalıdır, böylece veri doğruluğunu tehlikeye atmaz veya analize önyargı getirmez. Genel olarak, pertürbasyon yöntemleri çeşitli alanlarda veri gizliliğini koruma konusunda umut verici bir yaklaşım sunar.

Çalışmada veri gizliliğinin bununla birlikte veri güvenliliğinin sağlanması için çok boyutlu rotasyona dayalı pertürbasyon ve rastgele üretilmiş gürültü ekleme mekanizmaları 4 veri seti, 6 sınıflandırma yöntemi yardımı ile analiz edilmiştir.

Veri setleri sınıflandırma yöntemleri öncesi aykırı değerleri, boş değerleri farklı veri madenciliği yöntemleri yardımıyla değiştirilmiştir. Veri setlerindeki sayısal değerler normalizasyon yöntemleri ile, kategorik değerler tek çizgi ve sahte değişken kodlama yöntemleri ile sınıflandırma öncesi ön işlemeye tabi tutulmuştur.

Dört veri seti, test ve eğitim verileri olmak üzere çapraz doğrulama yöntemi ile kümelere ayrılmıştır. Orijinal veri setleri lojistik regresyon (LR), k en yakın komşu

(KNN), yapay sinir ağları (NN), destek vektör makinaları (SVM), gradyanı artıran karar ağaçları (XGBoost), hafif gradyanı artırılmış makinaları yöntemleri ile doğruluk değerleri hesaplanmıştır.

Veri setlerindeki sayısal iki yarı tanımlayıcı değişken önce 110 derece döndürülerek sonra ortalamaları ve standart sapmaları nezdinde gauss gürültüsü eklenerek doğruluk değerleri f1 skorları hesaplanmıştır.

Hesaplanan Friedman sıralama değerleri yardımıyla da sınıflandırma çıktıları kıyaslanmış en iyi sınıflandırma yöntemi dört veri seti için bulunmuş ve performans metrikleri yorumlanmıştır.

Çalışmada geometrik pertürbasyon yöntemleri ile nümerik verilerin gizliliği korunmaya çalışılmış ve sınıflandırma yöntemleri ile de bu değişimin doğurduğu bilgi kaybının miktarı analiz edilmiştir.

Çalışmanın ana katkısı, mahremiyeti koruyan güvenlik tahmini analizi için bir çerçeve önerisidir. Çerçeve, veri toplama, pertürbasyon teknikleri, analiz metodolojileri ve mahremiyetin korunmasının değerlendirilmesi dahil olmak üzere çok değişkenli hassas verilere pertürbasyon yöntemlerinin uygulanmasıyla ilgili adımları özetlemektedir.

Çalışmanın bir diğer katkısı, güvenlik tahmini için verilerin faydasını korurken mahremiyetindeki etkinliklerini değerlendirerek farklı pertürbasyon yöntemlerinin karşılaştırmalı bir analizini sunmaktır.

PREDICTION ANALYSIS FOR MULTIVARIATE DATA WITH RESPECT TO SENSITIVE DATA SECURITY USING THE PERTURBATION METHOD

SUMMARY

With the widespread use of big data, the Internet of Things (LoT), data mining brings the concept of data privacy to the agenda in all sectors from manufacturing, health, justice to banking. Data mining is the process of discovering interesting and previously unknown information from big data. Therefore, the necessity of ensuring data confidentiality in the process of processing data also arises. Protecting data privacy and performing data mining will provide data analysis in a healthier and more effective environment. In this study, data in static or dynamic environments; Confidentiality will be ensured with perturbation methods, and before and after data will be analyzed by considering the multidimensional structure.

With the development of technology, the use of big data is becoming widespread at an increasing pace. The storage, analysis and confidentiality of data has brought about algorithm methods that need to be developed. Privacy protection is the technique of data corruption, protection of record confidentiality, storage without disturbing the meaning of the values in the database and the relationship between variables. Perturbation methods are included in this study, which is based on the protection of the confidentiality of semi-defined and sensitive numerical data.

Sensitive data analysis is crucial to ensuring that appropriate security measures are in place to protect the confidentiality and integrity of information. One approach to handling sensitive data is data perturbation methods, which involves adding random noise or making small changes to data to maintain privacy while allowing analysis.

Perturbation methods are mathematical techniques used to add controlled noise or randomness to allow data analysis while maintaining the confidentiality of the data. Various methods such as random response, differentiated privacy, secure multi-party computing, noise addition, sampling, and aggregation are used to prevent the disclosure or abuse of sensitive information. These methods are being successfully applied in the fields of machine learning, statistics, and cryptography to ensure data privacy. However, the implementation should be carefully designed so that it does not compromise data accuracy or introduce bias into the analysis. Overall, perturbation methods offer a promising approach to protecting data privacy in a variety of areas.

In the context of security prediction analysis with multivariate data, we can apply data perturbation methods to protect sensitive information while maintaining the usefulness of the data for analysis. Examples of common techniques used in the perturbation process include:

Random Response: This technique involves adding random noise to data by offering controlled randomness during data collection or analysis. It ensures that individual

data points are not easily attributed to specific individuals, increasing confidentiality while maintaining the overall statistical characteristics of the dataset.

Differential Privacy: Differential privacy provides a mathematical framework for adding interference to data in a way that guarantees confidentiality. By adding carefully calibrated noise, meaningful analysis of the aggregated data is allowed while maintaining the confidentiality of individual data points.

Data Masking: In this approach, sensitive data is modified or transformed into less sensitive values. For example, instead of using the exact ages of individuals, you can use age ranges or categories to obscure the exact values.

Synthetic Data Generation: Synthetic data generation involves creating artificial data sets that mimic the statistical properties of the original data. By generating synthetic data that is not directly linked to sensitive information, privacy can be maintained while enabling analysis.

It is important to note that the choice of perturbation method depends on the specific requirements of your analysis and the sensitivity of the data. In addition, it is important to comply with legal and ethical rules, such as obtaining appropriate consent when working with sensitive information and adhering to data protection regulations. It proposes a privacy-preserving approach for security prediction analysis that uses perturbation methods on multivariate sensitive data. The goal is to protect the confidentiality and integrity of data while maintaining the usefulness of the information for security prediction purposes.

Before the classification methods of the data sets, the outliers and the null values were replaced with the help of different data mining methods. Numerical values in the data sets were pre-processed by normalization methods and categorical values were pre-processed before classification by single line and pseudo variable coding methods.

The four data sets are divided into sets by cross-validation, including test and training data. The accuracy values of the original data sets were calculated by methods of logistic regression (LR), k nearest neighbor (KNN), neural networks (NN), support vector machines (SVM), gradient increasing decision trees (XGBoost), slightly gradient increased machines.

The accuracy values F1 scores were calculated by first rotating the two semi-descriptive variables in the data sets by 110 degrees and then adding Gaussian noise in terms of their means and standard deviations.

With the help of the calculated Friedman ranking values, the classification outputs were compared, the best classification method was found for four data sets and the performance metrics were interpreted.

In the study, the confidentiality of numerical data was tried to be protected with geometric perturbation methods and the amount of information loss caused by this change was analyzed with classification methods.

The main contribution of this study is the proposal for a framework for security prediction analysis that protects privacy. The framework outlines the steps involved in applying perturbation methods to multivariate sensitive data, including data collection, perturbation techniques, analysis methodologies, and assessment of privacy protection.

It also provides a comparative analysis of different perturbation methods, assessing their effectiveness in protecting privacy while maintaining the usefulness of the data for security prediction. It discusses trade-offs between privacy and data use and provides insights into choosing the most appropriate perturbation method based on the specific requirements of the analysis. Privacy concerns surrounding sensitive data have led to the development of perturbation methods to protect individual privacy while allowing for meaningful analysis. In this study, it aims to contribute to the field of security prediction analysis by recommending and evaluating perturbation techniques specifically tailored for multivariate sensitive data. The contributions of this research lie in the following areas:

New Perturbation Techniques: Innovative perturbation methods designed to address the unique challenges of multivariate safety prediction analysis are discussed. By adding controlled randomness to these techniques, data masking approaches are used, offering different privacy guarantees, and synthetic data generation is used to ensure the confidentiality of sensitive data.

Privacy-Benefit Swap Assessment: Includes a comprehensive evaluation of the privacy-benefit trade-offs associated with various perturbation methods. By quantitatively assessing the impact of each technique on both privacy protection and data use, researchers and practitioners can make informed decisions about the most appropriate method for their specific security prediction analysis needs.

Practical Implementation Guidelines: Acknowledging the practical considerations for applying perturbation methods in real-world security forecasting systems, this document provides guidelines and best practices for the appropriate application of perturbation techniques. It handles data collection, perturbation algorithms, analysis methodologies, and the evaluation of privacy protection, ensuring that practitioners comply with legal and ethical standards when making accurate security predictions.

Comparative Analysis: A comparative analysis of existing perturbation methods applied to multivariate sensitive data is conducted in the context of security prediction analysis. By comparing the performance, strengths, and limitations of these methods, facilitating the selection of appropriate techniques for specific applications, researchers gain insight into their effectiveness and suitability for different scenarios.

Future Research Guidelines: By exploring privacy-preserving techniques and security prediction analysis, artificial intelligence techniques, and machine learning algorithms that protect data privacy, addressing the challenges of evolving privacy regulations and investigating the impact of perturbation techniques on different security prediction models.

The contributions of the study contribute to the advancement of security prediction analysis that protects privacy using perturbation methods on multivariate sensitive data. In data security situations involving sensitive information, researchers and practitioners can confidently analyze data while protecting individual privacy, thereby encouraging secure and responsible data-driven decision-making.

1. GİRİŞ

Dijitalleşme ile dünya, insanların ve makinaların çok büyük miktarlarda kişisel verinin üretilmesine olanak sağlamaktadır. E ticaretten, sağlık sektörüne, adaletten, sosyal medyaya, akıllı şehirlerden, otonom araçlara, savunma sanayiden birçok teknolojik alana verilerin mahremiyetinin korunması önemli bir hal almaktadır.

Dijitalleşme, daha fazla insanın teknolojiyle uğraşmasına, sosyal medya ortamlarında vakit geçirmesine, işlerini bu ortamlarda gerçekleştirmesine ve böylelikle bu ortamlarda daha çok ayak izi bırakmasına yani onlara ait tanımlayıcı kişisel verinin depolanmasına imkân vermektedir.

Kişisel veri, bir kişinin tanımlanabilir olması durumundaki her türlü bilgiye denir. İsim, kimlik numarası, ikametgâhı, elektronik posta adresi, cep telefonu numarası gibi bilgiler kişisel veriye örnek olarak verilebilir. Ayrıca, bir kişinin sosyal medya hesapları, fotoğrafları, coğrafi konumu, banka hesap bilgileri, sağlık durumu, cinsiyeti, ırkı, dinî inancı ve diğer demografik bilgileri de kişisel veri olarak kabul edilir.

Gelişen teknolojiyle birlikte üretilen ve depolanan verinin boyutu akıl almaz şekilde artmakta, bu veri içerinde çok sayıda kişisel veri de bulunmaktadır. Verinin ortaya çıkmasından veri depolarında muhafaza edilene dek güvenlik önemli bir unsurken, verinin analizi ile verinin mahremiyeti öne çıkarmaktadır.

Veri mahremiyeti insanların temel haklarından biridir ve bu hak korunmalıdır. Kişisel verilerin ifşa edilmesi, bir bireyin özel hayatını ve güvenliğini tehdit edebilir ve bu nedenle veri koruma yasaları ve politikaları gereklidir. İnsanların güvenliğini sağlamak için, bireylerin güvenlik bilincinin artırılması ve bilgi güvenliği yöntemlerine uyması önemlidir.

Ayrıca şirketlerin ve kurumların da veri koruma politikalarını uygulaması ve siber saldırıları gözeterek onlara karşı etkin bir şekilde güvenlik önlemlerini alması gerekmektedir. Veri mahremiyeti, dijital dünyada güvenliği sağlamak ve toplumun huzurunu korumak için büyük bir önem taşımaktadır. Bu verinin paylaşılması,

kullanılması veriyi üreten insanın mahremetiyle ilgili bir husus olduğu için birçok ülkede yasal düzenlemelere tabidir.

Kişisel Verilerin Korunması Kanunu, Türkiye'de 2016 yılında kabul edilen bir yasadır. Bu kanunun amacı, kişisel verilerin işlenmesiyle ilgili olarak kişisel verilerin gizliliğini, güvenliğini ve işlenmesine ilişkin diğer hakları korumaktır. Bu kanun kapsamında, kişisel verilerin işlenmesi sadece belirli koşullar altında mümkündür. Bu koşullar arasında, veri sahibinin açık rızası, yasal bir yükümlülük veya sözleşmenin yerine getirilmesi gibi nedenler yer almaktadır.

Veri mahremiyeti herkesin koruması altında olması gereken bir haktır. Kişisel verilerin ifşa edilmesi, bireyin özel yaşamına ve hatta güvenliğine zarar verebilir. Kişisel bilgilerin yanlış ellerde kullanımı, kimlik hırsızlığı, dolandırıcılık, itibar kaybı, ayrımcılık gibi istenmeyen birçok soruna neden olabilir. Veri koruma yasaları ve önlemleri, kişisel verilerin yetkisiz erişim, kullanım, ifşa ve yok edilmesine karşı etkili bir şekilde korunması için hayati öneme sahiptir.

Ayrıca, şirketlerin ve kurumların da veri koruma politikalarını sıkı bir şekilde uygulaması ve siber saldırılara karşı güvenlik önlemlerini alması gerekmektedir. Veri mahremiyetinin korunması, dijital dünyada güvenliği sağlamak ve toplumun huzurunu muhafaza etmek açısından büyük önem arz eder.

Çalışma beş bölümden meydana gelmekte olup ilk bölümde literatür taramasında veri ve veri mahremiyetiyle ilgili uygulamalara değinilmiştir.

Çalışmanın ikinci bölümünde genel manada veri madenciliğinde gizliliğin korunması metotlarıyla ilgili tanımlamalara değinilmiştir.

Çalışmanın üçüncü bölümünde uygulamada kullanılan gereç ve yöntemler anlatılmıştır.

Çalışmanın dördüncü bölümünde ise daha önce teorik bilgileri aktarılmış olan metodun uygulamasına yer verilmiştir. Burada altı sınıflandırma algoritması kullanılarak dört veri seti için iki boyutlu rotasyon pertürbasyonu ve rastgele gauss gürültüsü ekleme metotları bir araya getirilerek veri mahremiyeti sağlanmaya çalışılmıştır. Yöntemlerin uygulama süreci adım adım açıklanarak, yönteme ait kavram ve hesaplamalar üzerinde durulmuştur.

Çalışmanın beşinci ve son bölümünde ise iki boyutlu rotasyon pertürbasyonu ve gauss mekanizması metotlarının, veri gizliliğini ne derece sağladığına yönelik performansları ölçülmüş ve yorumlanmıştır.

2. LİTERATÜR ARAŞTIRMASI

Son birkaç yıl içinde, gizlilik koruma veri madenciliği çalışmalarında farklı araştırma çevrelerince yeni birçok yaklaşım önerildi. Başlangıçta, rastgele eklemeler ve çarpma gibi temel yöntemler kullanıldı. Nitekim bu yöntemler hemen hemen tüm saldırılara açıktı. Daha sonra, veri kullanımı ve gizlilik arasında denge sağlayan verimli teknikler de önerildi. Bazı temel yaklaşımlar (Aggarwal ve Philip, 2008) veri bozulması, veri değiş-tokuşu, k-anonimleştirme, şifreleme tabanlı yöntemler, kural gizleme yöntemleri ve güvenli dağıtılmış madencilik teknikleri olarak sayılabilir.

Olasılık dağılımı ve veri değeri bozulma yaklaşımı olmak üzere iki temel veri bozulma yaklaşımı vardır. Olasılık dağılımı yaklaşımında (Liew ve diğerleri, 1985), veriler aynı dağılımdan başka bir örnek ile değiştirilir. Veri değeri bozulmasında ise, veri öğeleri ekleyici gürültü, çarpıcı gürültü veya diğer rastgeleleştirme metotları ile bozulur.

Gürültü Ekleyici Bozulma, veri setini gürültü ekleyerek bozar. Genellikle Gauss dağılımı, gürültü değerini oluşturmak için kullanılır. Gürültünün korelasyonu, orijinal veriye benzerse, gizliliğin daha iyi korunması sağlanır. Prensipal Bileşen Analizi ve Bayes Tahmini Teknikleri, rastgeleleştirme tekniklerinin yeniden yapılandırma hassasiyetini tahmin etmek için yaygın olarak incelenmiştir (Huang ve diğerleri, 2005). Bozulma yöntemlerinin diğer yöntemleri, çarpıcı bozulma (Chen ve Liu, 2008), döndürme bozulması (Huang ve diğerleri, 2005; Chen ve Liu, 2011) ve çok boyutlu bozulma (Chen ve Liu, 2005) şeklindedir. Başka bir yaklaşımda (Oliveira ve Zaane, 2004), veriye öncelikle logaritmik dönüşüm uygulanır, daha sonra önceden tanımlanmış birçok değişkenli Gauss gürültüsü eklenir. Gürültü eklenmiş verinin anti-logu alınır.

Veri değiştirme (Fienberg ve McIntyre, 2004) yönteminde, hassas öznitelik değerleri kayıtlar arasında değiştirilerek, hassas veriler hakkında belirsizlik yaratılır. K-Anonimlik modeli (Sweeney, 2002; Gionis ve Tassa, 2009) veri genelleştirme ve baskılama yöntemlerini kullanır ve veriler yalnızca her bir kişinin bilgilerinin en az (k-1) diğer kişilerden ayırt edilemeyeceği durumlarda yayınlanır. KD-ağacı tabanlı

pertürbasyon yönteminde (Li ve Sarkar, 2006) veriler küçük alt kümeler halinde ayrılır ve alt kümelerdeki hassas veriler, alt küme ortalaması kullanılarak pertürbe edilir.

İstatistiksel özellikleri koruyarak gizlilik düzeyini artırmak için çarpımsal rastgele projeksiyon matrislerine dayalı bir gizlilik koruma dağıtık veri madenciliği tekniği (Liu vd., 2006) önerilir. Pinkas tarafından önerilen şifreleme teknikleri (2002) de gizlilik koruma amaçlı veri madenciliği için önerilir. Chen ve Liu, birden fazla geometrik pertürbasyonu farklı taraflar tarafından tercih edilerek kavramsal anahtarlar kullanarak güvenli bir şekilde birleştiren çok paydaşlı iş birliği gizlilik koruma madenciliği yöntemi önerir (Chen ve Liu, 2009). Kural gizleme yaklaşımında (Verykios vd., 2004) veri tabanı, hassas kuralları gizlemek için dönüştürülür. Yeni veri madenciliği algoritmaları, özellikle PPDM için, rastgele karar ağacı (Vaidya vd., 2014), değiştirilmiş Bayesian ağı (Yang ve Wright, 2006) ve SVM sınıflandırıcısı (Lin ve Chen, 2011) gibi önerilir.

Veri takası hakkındaki ilk araştırma Resis (1980) tarafından yapılmıştır. Bu makalede, yazarlar öncelikle bir kayıt içindeki hassas bilgilerin değiştirilmesine odaklanmıştır. Takas yöntemi hassas değişkenlerin varyant dağılımını korur. Aşağıdaki farklı yazar perspektifleri takas tekniğiyle ilgilidir:

Verykios, K. Bertino, I.N.Fovino, L.P.Provenza, Y.Saygin ve Theodoridis (2004), merkezi ve dağıtık veriler için veri madenciliği yöntemlerinin gizlilik korumasına dayanan yeni bir teknik sunmuştur.

W. Du ve Zhan (2002), orijinal verileri korumak için güvenli skaler, güvenli toplama ve güvenli birleştirme kullanarak karar ağaçları oluşturma yöntemlerini açıklamıştır. Bu yöntemin ana dezavantajı, sistem performansını düşüren birden çok veri tabanı taraması yapmasıdır.

R. Aggarwal ve R.Srikanth (2000), kullanıcının özel verilerini korumak için rastgele veri bozulma tekniği kullanarak yeni bir pertürbasyon yöntemi tanımladılar ve karar ağaçları oluşturdular. Bu yöntem, daha önceki yöntemlere kıyasla daha düşük bir doğruluk elde etti.

HillolKargupta, SouptikDatta, Qi Wang ve K.Siva Kumar (2003), kullanıcı verilerinin rastgele gürültü üreten farklı rastgele yöntemlerle gizlilik koruyucu veri madenciliği sergiledi. Rastgele matrisler oluşturmak için farklı rastgele tabanlı yöntemler tanımladılar.

TanveerJahan, G.Narasimha, C.V.GururRao (2012), gizlilik koruyucu veri madenciliği için bozulmuş verilerde kümeleme konusunu detaylandırdılar.

Lindell ve B.Pinkas (2002), daha küçük veri tabanları için kullanılan farklı şifreleme tabanlı gizlilik koruyucu veri madenciliği tekniklerine odaklandı. Bu makale, şifreleme yöntemlerini kullanarak yüksek güvenlik elde etti.

Shweta T, ShanshankKhanna, T.Sugandha ve Ankitha (2014), hassas verilerin gizliliğini korumak için veri madenciliği sırasında etkili bir hibrit karar ağacı teknikleri tasarlamışlardır. Şifreleme işlemi için özel karakterler ve ASCII kodları kullanılmıştır. Bu teknik, tıbbi veri kümelerini korumak için çok faydalıdır.

A.Srivastava ve G.Srivastav (2015), E-Sağlık kayıtlarında bireysel özel verileri korumak için K-Anonim tekniklerini kullandılar ve tıbbi veri kümelerinde en iyi doğruluk sonuçlarını elde ettiler.

Yifeng XU ve Jie Liu (2010), rasgele yanıt yöntemi ve geometrik veri bozulma yöntemini sergilemiştir. Bu yöntem yalnızca sayısal veri kümelerini koruyabilir ve mevcut tekniklere göre daha iyi gizlilik koruması sağlar.

Jie Liu ve Yifeng XU (2010), rasgele yanıt teknikleriyle geometrik veri pertürbasyonu kullanarak bir yöntem geliştirdi. Bu makale sadece sayısal veri kümesine odaklanmaktadır ve kategorik veri kümeleri için uygulanabilir değildir. Mevcut veri pertürbasyon tekniklerine kıyasla iyi bir doğruluk seviyesi sağlamaktadır.

Chhinkaniwala H ve Garg S (2011), gizlilik koruma amaçlı veri madenciliği ile ilgili farklı teknikleri ve zorlukları tanımlamışlardır. Bu makalede, yazarlar farklı gizlilik koruma tekniklerinin sınıflandırılması üzerine odaklanmışlar ve mevcut gizlilik koruma veri madenciliği yöntemlerinin çeşitli sınırlamalarını tartışmışlardır.

M.Reza ve SomayyehSeifi (2011), veri pertürbasyon yöntemleri kullanarak gizlilik koruma veri madenciliğini değerlendirmek için yeni bir sınıflandırma tekniği tasarlamışlardır.

H. Chhinkaniwala ve S. Garg (2013), demet değerine dayalı etkili bir Çarpımsal Pertürbasyon tekniği sunmuşlardır. Doğruluğu artırmak için K-Means Kümeleme algoritmasını kullandılar ve geri çağırma, hassasiyet ve CMM hesapladılar.

Mr.Kiran Patel (2013), Gizlilik koruma veri madenciliği için veri akışlarının sınıflandırılması için yeni bir yaklaşım tanımladı. Veri ön işleme ve veri akışı veri

madenciliği gibi iki temel veri madenciliği adımını içeriyordu. Minimum bilgi kaybı ile Hoeffding tekniğini tanıttılar.

G. Manikandan ve ark. (2013), iyi bir doğruluk elde edilen veriler için Normalizasyon kullanarak veri dönüştürme tekniği önerdiler ve veri madenciliği algoritmalarının performansını artırdılar.

Tarique Ahmad ve diğerleri (2014), bir veri kümesinde hassas özniteliklerin gizliliğini korumak için min-maks normalizasyonuna dayalı bir yaklaşım tanımlamışlardır. Veri madenciliği başlamadan önce orijinal veri kümesi değerleri min-maks normalizasyonu kullanılarak değiştirilir. Deneyler, önerilen k-means algoritmasının hem doğruluğun hem de gizliliğin korunduğunu kanıtladı.

Patel Brijal ve diğerleri (2015), bir veri kümesinde hassas öznitelikleri korumak için kümeleme algoritması kavramını açıklamışlardır. Önerilen yöntem, veri kümesindeki hassas ve kritik bilgileri korumada başarılı olmuş ve minimum bilgi kaybı ile iyi veri madenciliği sonuçları elde etmiştir.

Anjana Patel ve diğerleri (2016), k-means kümeleme algoritması kullanarak değiştirilmiş ve rastgeleleştirilmiş veriler üzerinde geometrik veri pertürbasyonu kavramını göstermişlerdir. Önerilen yöntem doğruluğu kontrol etmek için kullanılmış ve iyi bir doğruluk seviyesi elde edilmiştir. Deneysel sonuçlar, önerilen yöntemin orijinal veri değerlerini bilgi kaybı olmadan yeniden oluşturabileceğini kanıtlamıştır.

3. VERİ GİZLİLİĞİNİN KORUNMASI

Büyük veride gizlilik koruması, özellikle hassas verilerin (kişisel, tıbbi, finansal vb.) toplandığı büyük veri kümelerinde, veri madenciliği işlemleri yaparken gizliliği muhafaza etmek maksadıyla geliştirilen yöntem ve bir dizi teknikten ibarettir. Bu teknikler, verilerin anonimleştirilmesi, gizlenmesi veya özetlenmesi gibi yöntemleri içerebilir ve veri madenciliği işlemlerinin gerçekleştirilmesi sırasında minimum veri miktarını korurken, hassas verilerin ifşa edilmesini önler. Büyük veride gizlilik koruması, gizliliğin korunması ve veri analizinin güvenliği arasındaki dengeyi sağlamak için önemlidir.

3.1. Gizliliği Koruyan Veri Madenciliği Teknikleri (Privacy Preserving Data Mining)

Kişisel verilerin gizliliğini koruyarak veri madenciliği yapılmasına olanak tanıyan bir yöntemdir. Gizliliği koruyan veri madenciliği yöntemleri, veri sahiplerinin gizlilik kaygılarını dikkate alarak, özel bilgilerin korunmasını sağlar ve bu verilerin analizi ve kullanımı için özel algoritmalar kullanır.

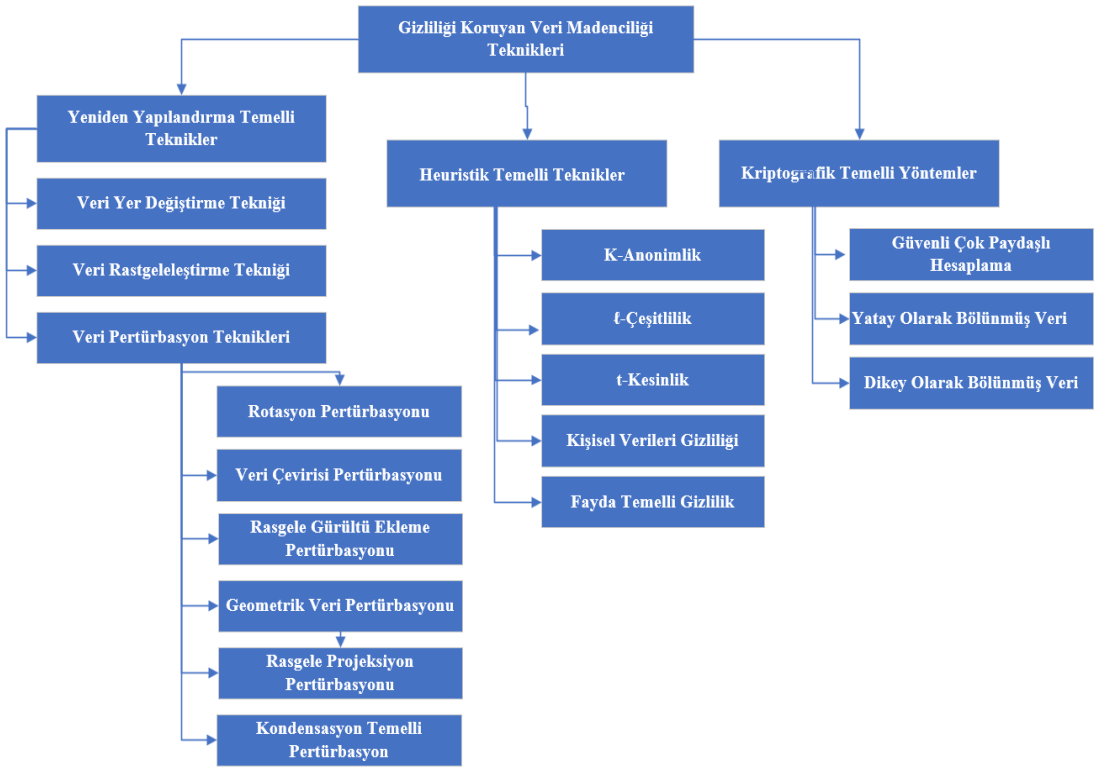
Gizliliği koruyan veri madenciliği yöntemleri, kişisel verilerin gizliliğini muhafaza etmek maksadıyla birçok metoda yer verir. Bu metotlar, verilerin gizliliğini koruyarak, verilerin analiz edilmesi ve kullanılmasını sağlar. Gizliliği koruyan veri madenciliği teknikleri, özellikle sağlık, elektronik ticaret ve haberleşme ve finans gibi sektörlerde, kişisel ve hassas verilerin korunmasına yöneliktir.

Gizliliği koruyan veri madenciliği yöntemlerinin amacı, veri sahiplerinin kişisel verilerinin korunmasını sağlamak ve bu verilerin analizi ve kullanımı için özel yöntemler kullanmaktır. Bu yöntemler, veri sahiplerinin gizlilik haklarını korurken, aynı zamanda verilerin analizi ve kullanımına olanak tanır.

Gizliliği koruyan veri madenciliği teknikleri, veri madenciliği çalışmalarında kullanılan bazı yöntemleri içerir. Bu yöntemlere örnek, veri anonimleştirme, veri maskeleme, veri karıştırma ve gizli veri madenciliği verilebilir. Böylece yöntemler,

kişisel verilerin gizliliğini korurken, veri sahiplerinin gizlilik haklarını da korumuş olur.

Gizliliği koruyan veri madenciliği yöntemleri, kişisel verilerin gizliliğinin korunması ve veri madenciliği çalışmalarının yapılabilmesi arasında bir denge sağlar. Bu yöntemler, veri sahiplerinin gizlilik haklarını korurken, aynı zamanda verilerin analizi ve kullanımına olanak tanır. Gizliliği koruyan veri madenciliği teknikleri Şekil 3.1'de gösterilmiştir.



Şekil 3.1. Gizliliği koruyan veri madenciliği teknikleri.

Gizliliği koruyan veri madenciliği tekniklerinden bazıları şunlardır:

3.1.1. Yeniden yapılandırma temelli teknikleri

Son zamanlarda önerilen birçok teknik, madenciliği gerçekleştirmek için verileri pertürbe etme ve toplu düzeyde dağılımları yeniden oluşturma yoluyla gizlilik koruma sorununu ele almaktadır.

3.1.1.1. Veri pertürbasyon teknikleri

Pertürbasyon, istatistiksel açıklama kontrolünde kullanılan bir yöntemdir. Basitlik, verimlilik ve istatistiksel bilgiyi koruma özelliği gibi içsel özellikleri vardır. Pertürbasyonda, orijinal değerler, istatistiksel bilgi pertürbe edilmiş verilerden

hesaplanan istatistiksel bilgiye, orijinal verilerden büyük ölçüde farklı olmayacak şekilde bazı sentetik veri değerleriyle değiştirilir.

Pertürbe edilmiş veri kayıtları gerçek dünya kayıt sahipleriyle aynı olmadığından, saldırgan düşünceli bağlantıları yapamaz veya mevcut verilerden hassas bilgi kurtaramaz.

Pertürbasyon yaklaşımında, her boyutu bağımsız olarak ele almak üzere her dağılım tabanlı veri madenciliği algoritması varsayımsal olarak çalışır.

Sınıflandırma gibi veri madenciliği yöntemleri ile ilgili bilgi, ara-ölçü öznitelikleri arasındaki ilişkilere gizlenmiştir. Bu, pertürbasyon yaklaşımının farklı öznitelikleri bağımsız olarak ele alması nedeniyle olur. Bu nedenle, dağılım tabanlı veri madenciliği algoritmaları, çok boyutlu kayıtlarda mevcut gizli bilginin kaybı açısından içsel bir dezavantaja sahiptir. Pertürbasyon yaklaşımının dezavantajlarını yöneten gizlilik koruma veri madenciliğinin başka bir dalı ise kriptografik tekniklerdir.

Verilerde küçük rastgele değişiklikler yaparak orijinal verilerin gizliliğini korur. Bu teknik, hassas verilerin tamamen gizlenmesi yerine verilerin gizliliğini korumak için kullanılır.

Örneğin, bir veri kümesindeki her yaştaki kişinin yaşı birkaç yıl artırılabilir veya azaltılabilir. Bu değişiklikler, verilerin analiz edilebilirliğini korurken, kişisel bilgilerin ifşa edilmesini önler. Ancak, yanlış pertürbasyon teknikleri kullanılırsa, verilerin analiz edilebilirliği de etkilenebilir, bu nedenle doğru yöntemlerin seçimi ve uygulanması çok önemlidir.

Verilerde küçük rastgele değişiklikler yaparak orijinal verilerin gizliliğini korur. Bu teknik, hassas verilerin tamamen gizlenmesi yerine verilerin gizliliğini korumak için kullanılır. Örneğin, bir veri kümesindeki her yaştaki kişinin yaşı birkaç yıl artırılabilir veya azaltılabilir. Bu değişiklikler, verilerin analiz edilebilirliğini korurken, kişisel bilgilerin ifşa edilmesini önler. Ancak, yanlış pertürbasyon teknikleri kullanılırsa, verilerin analiz edilebilirliği de etkilenebilir, bu nedenle doğru yöntemlerin seçimi ve uygulanması çok önemlidir.

Pertürbasyon, eklenmiş gürültü veya veri değişimi veya sentetik veri oluşturma kullanılarak yapılabilir.

Pertürbasyon yöntemlerinden en çok tercih edilenlerini çalışmada ele aldık. Bunlar;

1. Rotasyon pertürbasyonu: Bir veri kümesindeki özelliklerin döndürülerek değiştirildiği bir gizliliği koruyan veri madenciliği yöntemidir. Bu yöntemde, veri kümesindeki özellikler birbirleriyle ilişkili olduğu için, özelliklerin dönüştürülmesi, orijinal veri kümesindeki yapısal ilişkileri koruyarak verilerin gizliliğini koruyabilir.

Örneğin, bir veri kümesinde "yaş" ve "gelir" özellikleri birbirleriyle ilişkilidir. Rotasyon pertürbasyonu kullanılarak, bu özellikler bir miktar döndürülerek değiştirilebilir. Böylece, verilerin gizliliği korunurken, orijinal veri kümesindeki yapısal ilişkiler korunabilir. Ancak, bu yöntem değiştirilen verilerin anlaşılabilirliğini zayıflatabilir.

Rotasyon pertürbasyonu, sınıflandırma ve kümeleme gibi işlemlerde gizliliğin korunması için kullanılır. Bu yöntem, temel bileşen analizine dayanarak kullanılır ve değerlerin değiştirilmesi için dik metrikler kullanılır. Ayrıca, G(x) = RX formülü kullanılarak geometrik veri dönüştürmesi yapılır, burada R bir dönüşüm matrisi ve X orijinal veri setidir. Rotasyon pertürbasyonu ayrıca mesafe koruma özelliği de taşır. Çalışmada çok boyutlu veri setimizin gizlemek istediğimiz iki özniteliğini saat yönünde rastgele bir açıda döndürerek veri gizliliği sağlanmasına çalışılmıştır. Üçüncü bölümde metot daha ayrıntılı olarak işlenecektir.

2. Veri çevirisi pertürbasyonu: Bir veri kümesindeki hassas verileri muhafaza etmek maksadıyla kullanılan bir tekniktir. Bir öznitelik değerine sabit bir değer eklenerek veriler değiştirilir. Veri çevirisi yöntemi, eklenen değerin özelliklerin dağılımını bozmadan verilerin gizliliğini korumak için sıklıkla kullanılır.

3. Rasgele gürültü ekleme pertürbasyonu: Literatürde diferansiyel mahremiyet diye anılır.

Gauss, Laplace, Eksponansiyel mekanizma olmak üzere yaygın üç kullanımı vardır. Çalışmamızda rotasyon pertürbasyonu birlikte gauss mekanizması kullanılmıştır Üçüncü bölümde metot daha ayrıntılı olarak işlenecektir.

4. Geometrik veri pertürbasyonu: Metot üç adımdan oluşur. İlk adımda, orijinal veri kümesi gibi rastgele değerler kullanılarak rastgele bir veri kümesi oluşturulur. Bu rastgele veri kümesi saat yönünde döndürülür ve daha sonra orijinal veri kümesi ile çarpılır. Üçüncü adımda devriği alınmış veri kümesine gürültü ile bozulmuş veri kümesi eklenir.

5. Rasgele projeksiyon pertürbasyonu: Veri noktalarını orijinal çok boyutlu

uzaydan rastgele seçilen başka bir uzaya yansıtma tekniği olarak tanımlanır (Liu, Kargupta ve Ryan, 2006). Projeksiyon bozulmasının mantığı, Johnson-Lindenstrauss Lemma'sı (Johnson ve Lindenstrauss, 1984) tarafından desteklenen yaklaşık mesafe korumasına dayanır.

Herhangi bir veri kümesinin öklid uzayında, herhangi iki noktanın çiftli mesafesinin küçük hata ile korunacağı başka bir uzaya gömülebileceğini gösterir. Sonuç olarak, model kalitesi yaklaşık olarak korunabilir.

6. Kondensasyon temelli pertürbasyon: Çok boyutlu bir pertürbasyon tekniği olarak tipik bir yöntemdir ve çoklu sütunlar için kovaryans matrisini korumayı amaçlar. Bu nedenle, karar sınırının şekli gibi bazı geometrik özellikler iyi korunur. Rastgeleleştirme yaklaşımından farklı olarak, tüm "pertürbe edilmiş veri kümesini oluşturmak için bir bütün olarak birden fazla sütunu pertürbe eder. Pertürbe edilmiş veri kümesi kovaryans matrisini koruduğundan, birçok mevcut veri madenciliği algoritması doğrudan pertürbe edilmiş veri kümesine uygulanabilir ve algoritmaların değiştirilmesi veya yeni geliştirilmesi gerekmez (Aggarwal ve Yu, 2004).

Kondensasyon yaklaşımı şöyle özetlenebilir: İlk olarak, orijinal veri kümesi k-kayıt gruplarına bölünür. Her grup, iki adımda oluşur. Mevcut kayıtlardan bir kaydın rastgele seçilmesi ve merkez olarak belirlenmesi, ardından merkezin (k-1) en yakın komşusunun diğer (k-1) üyesinin bulunması seçilen k kayıt, bir sonraki grup oluşturulmadan önce orijinal veri kümesinden çıkarılır. Her grup küçük bir yerelliğe sahip olduğundan, dağılımı ve kovaryansı korumak için yaklaşık olarak bir k kaydı yeniden oluşturmak mümkündür. Kayıt yeniden oluşturma algoritması, her grup için özvektörleri ve özdeğerleri korumaya çalışır

Ancak, kondensasyon yaklaşımının veri gizliliğini korumada zayıf olduğunu söylenebilir. Her grup içindeki yerelliğin boyutu ne kadar küçük olursa, yeniden oluşturulmuş k kayıtlarının korunan kovaryans kalitesi o kadar iyi olur.

3.1.1.2. Veri yer değiştirme tekniği

Yer değiştirme tekniği özellikle hassas verilerin değiştirilmesi gerektiği durumlarda kullanılır ve verilerin mahremiyetini korumayı amaçlar. Bu teknik, verilerin içeriğini değiştirmeden verilerin sırasını rastgele değiştirerek verilerin gizliliğini korur. Bu uygulama esnasında, herhangi bir veri öğesinin kimliği korunur, ancak konumları değişir.

Örneğin, bir araştırmacı bir anket çalışması yapmak istediğinde, katılımcılardan gelen verileri analiz etmek için bu yöntemi kullanabilir. Bu yöntem sayesinde, verilerin içeriği korunurken, herhangi bir katılımcının özel bilgileri ortaya çıkmaz.

Veri yer değiştirme tekniği, özellikle kişisel verilerin toplandığı ve paylaşıldığı alanlarda, örneğin tıp, sosyal bilimler, pazarlama gibi alanlarda yaygın olarak kullanılmaktadır**.**

3.1.1.3. Veri rastgeleleştirme tekniği

Verilerin gizliliğini korumak maksadıyla istatistiksel yöntemlerden yararlanır. Bu teknikte, veriler, herhangi bir müşterinin doğru bilgi içeren veri mi yoksa yanıltıcı bilgi içeren veri mi sunduğunu merkezi yerin belirli bir eşiği aşmayacak şekilde karıştırılır. Her bir kullanıcının aldığı bilgi karıştırılır ve örnek yani kullanıcı sayısı arttıkça, bu kullanıcıların toplam bilgileri iyi bir doğrulukla tahmin edilebilir. Bu, karar ağacı sınıflandırması için çok değerlidir. Veri toplama işlemi rastgeleleştirme yöntemiyle iki adımda gerçekleştirilir. İlk adımda, veri sağlayıcılar verilerini rastgeleleştirerek veri alıcısına aktarır. İkinci adımda, veri alıcısı yeniden oluşturma algoritması kullanarak verilerin orijinal dağılımını yeniden oluşturur.

Rastgeleleştirme yöntemi oldukça basittir ve diğer kayıtların dağılımını bilme gerekliliği yoktur. Bu nedenle, rastgeleleştirme yöntemi veri toplama zamanında uygulanabilir. Anonimleştirme işlemini yapmak için güvenilir bir sunucuda tüm orijinal kayıtların bulundurulması gerekliliği yoktur. Ancak, rastgeleleştirme yanıtına dayalı bir gizliliği koruyan veri madenciliği tekniğinin zayıflığı, tüm kayıtların yerel yoğunluklarına bakılmaksızın eşit işlem görmesi gerçeğidir. Bu, aykırı kayıtların daha sıkı bir şekilde karşı koyma saldırısına maruz kalmasına neden olur. Bu sorunun bir nedeni, tüm kayıtlara gereksiz bir gürültü eklemektir. Ancak, veri madenciliği amacına uygun sonuçlar vermediği için yeniden oluşturulan dağılım veri kullanımını azaltır.

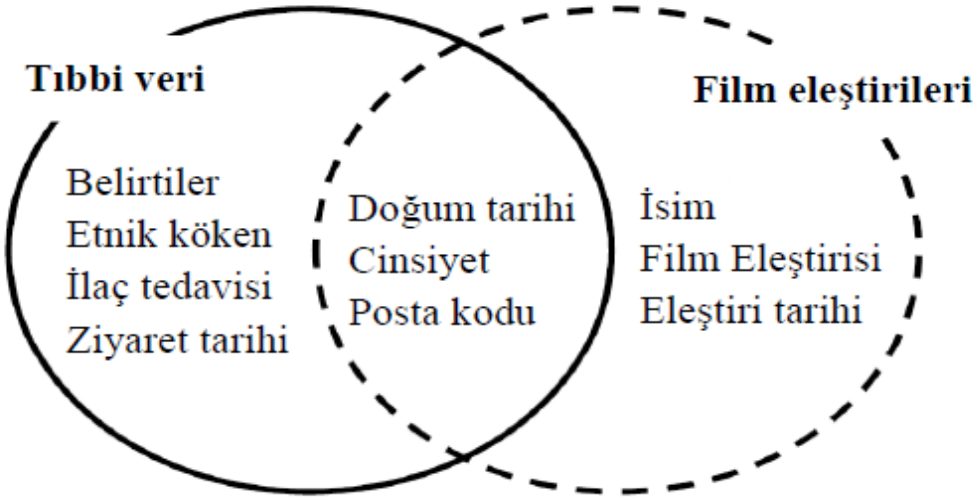
3.1.2. Heuristik temelli teknikler

Veri tabanı topluluğundaki araştırmacılar, kayıtları "grup tabanlı" bir şekilde işleyen ve özel gizlilik ölçütlerini koruyan özelleştirilmiş kayıtların yayınlanmasına olanak sağlayan yöntemler geliştirmişlerdir. Bir kaydın belirli yerel kayıtlar hakkında global bilgileri kullanarak dönüştürülmesi, belirli gizlilik metriklerinin korunmasını sağlar. Bu değiştirilmiş kayıtlar, saldırılar tarafından yeniden yapılandırılma korkusu olmadan yayınlanabilir.

Belirli bir kaydın, kayda bağlı bir bireyi benzersiz bir şekilde tanımlayan yarı tanımlayıcı özniteliklerini ve üçüncü taraflar tarafından ilişkilendirilmemesi gereken hassas öznitelikleri içerdiği varsayılır. Tanımlayıcı, yarı tanımlayıcı ve hassas öznitelikler Şekil 3.2.'de örnek olarak gösterilmiştir.

3.1.2.1. Tanımlayıcı nitelik (Identifier-ID)

Tanımlayıcı (Identifier-ID), bir nesneyi veya bir varlığı benzersiz bir şekilde tanımlamak için kullanılan bir kimlik numarasıdır. Bu kimlik numarası, bir veri tabanında veya bir sistemdeki diğer kaynaklarda ilgili varlık hakkında bilgi depolamak, aramak ve yönetmek için kullanılır. Örneğin, bir işletmenin çalışanlarını yönetmek için bir veri tabanı kullanıyorsanız, her çalışan için bir tanımlayıcı atanabilir ve bu tanımlayıcılar, çalışanların bilgilerinin doğru bir şekilde yönetilmesine ve raporlanmasına yardımcı olabilir. Tanımlayıcılar genellikle benzersiz ve sabit bir biçimde oluşturulur ve bu nedenle bir varlık hakkında her zaman aynı tanımlayıcı kullanılır.



Şekil 3.2. Hassas, tanımlayıcı ve yarı tanımlayıcı öznitelikler.

3.1.2.2. Yarı tanımlayıcı nitelik (quasi identifier)

Yarı tanımlayıcı bir veri kümesindeki bireyleri doğrudan tanımlayamasa da diğer verilerle birleştirildiğinde bir bireyi tanımlamak için kullanılabilecek verilerdir. Yani, yarı tanımlayıcıların, bir bireyin kimliğini belirlemek için yeterli olmasa da diğer verilerle bir araya getirildiğinde bireyleri tanımlayabilecek özelliklerdir. Örneğin, bir ad, adres ve doğum tarihi kombinasyonu, bir kişinin kimliğini doğrudan tanımlamaz, ancak diğer verilerle birleştirildiğinde kişinin kimliği belirlenebilir. Bu nedenle, yarı

tanımlayıcılar gizliliği korumak için korunması gereklidir.

3.1.2.3. Hassas nitelik (sensitive attribute)

Hassas nitelik, bir veri kümesindeki bireylerin özellikleri hakkında duyarlı ve kişisel olarak tanımlayıcı olan özniteliklerdir. Bir sağlık veri kümesinde, kişilerin tıbbi durumları, tedavi geçmişleri veya genetik bilgileri hassas niteliklere örnek olabilir.

Bir kişinin hassas niteliklerinin ifşa edilmesi, gizlilik ihlali olarak kabul edilir ve kişisel mahremiyetlerinin tehlikesinin önlenmesi için, ivedi karşı tedbirler almak gerekir. Hassas niteliklerin gizliliğinin korunması, veri koruma ve gizlilik yöntemleri açısından önemlidir.

3.1.2.4. Hassas olmayan nitelikler (non-sensitive attribute)

Hassas olmayan nitelik, kişisel bilgiler veya hassas verilerle ilgili olmayan veri özellikleridir. Örneğin, bir kişinin adı veya adresi hassas bilgi olabilirken, yaş veya cinsiyet gibi diğer özellikler hassas olmayan niteliklerdir. Hassas olmayan nitelikler genellikle veri analizinde kullanılan diğer veri özellikleridir ve veri madenciliği modellerinin oluşturulmasında önemli bir rol oynarlar. Ancak, hassas olmayan nitelikler de hassas verilerin ifşa edilmesiyle ilişkilendirilebilir, bu nedenle gizlilik koruma teknikleri tüm veri özelliklerini korumak için tasarlanmalıdır.

Anonimleştirme, kayıt sahiplerinin kimlik bilgileri ve/veya hassas verilerinin gizlenmesi gereken bir yaklaşımı ifade eder. Hassas verilerin analiz için saklanması gerektiği varsayılır. Açık kimlik belirleyicileri kaldırmak açıktır, ancak neredeyse kimlik bilgileri, halka açık verilerle ilişkilendirildiğinde gizlilik ihlali riski vardır. Bu tür saldırılara ilişkin terimler "bağlama saldırıları" olarak adlandırılır. Örneğin, cinsiyet, ırk ve posta odu gibi öznitelikler, seçmen listesi gibi halka açık kayıtlarda mevcuttur.

Kayıt eşleştirmeyi önlemek, k-anonimlik, nitelik eşleştirmeyi ve kayıt bağlamayı engellemek maksadıyla l-çeşitlilik gibi yöntemler, olasılıksal saldırılara karşı ve nitelik eşleştirmesi için t-yakınlık (Fung ve diğerleri, 2010) kullanılmaktadır.

K-anonimlik, ℓ-çeşitlilik ve t-kesinlik gibi üç grup tabanlı yöntem varyasyonu vardır.

3.1.2.5. K-Anonimlik

Anonimleştirme teknikleri, verilerin kimliğini korurken belirli bir veri kümesinin yararlı bilgilerini korumak için kullanılan tekniklerdir. Anonimleştirme, bir veri

kümesindeki kişisel bilgilerin (örneğin ad, adres, doğum tarihi) kaldırılması veya değiştirilmesiyle başlar. Daha sonra, veri kümesindeki diğer niteliklerin (örneğin yaş, cinsiyet, posta kodu) değiştirilmesi, gizlenmesi veya gruplandırılması gibi teknikler kullanılarak veri kümesinin anonimleştirilmesi sağlanır.

3.1.2.6. ℓ-Çeşitlilik

Bir gizlilik koruma tekniği olarak kullanılan grup tabanlı bir yöntemdir. K-anonimlik yönteminde, bir grup içindeki kayıtların yarı tanımlayıcı öznitelikleri aynı olacak şekilde anonimleştirilir.

Ancak bu durum, grup içindeki hassas özniteliklerin korunmasını sağlamaz. ℓ-çeşitlilik yöntemi, bir grup içindeki kayıtların yarı tanımlayıcı özniteliklerine ek olarak hassas özniteliklerinin de belirli bir çeşitlilik seviyesine sahip olmasını sağlamaktadır. Böylece, saldırganlar bir kaydın hassas özniteliklerini tahmin etmeye çalıştığında, birden fazla farklı değer olma ihtimali yüksek olur. Bu, hassas özniteliklerin korunmasını sağlamaktadır.

3.1.2.7. t-Kesinlik

Bir gizlilik koruma tekniği olarak kullanılan grup tabanlı bir yöntemdir. k-anonimlik ve ℓ-çeşitlilik gibi diğer yöntemlerde olduğu gibi, bir grup içindeki kayıtların yarı tanımlayıcı öznitelikleri anonimleştirilir. Ancak bu yöntemlerde, hassas özniteliklerin dağılımı göz ardı edilir ve grup içindeki kayıtların hassas öznitelikleri aynı dağılımı paylaşır. t-kesinlik yöntemi ise, bir grup içindeki kayıtların hassas öznitelikleri için belirli bir kesinlik seviyesi sağlar. Yani, bir grup içindeki kayıtların hassas öznitelikleri, gerçek dağılımlarından belirli bir mesafe kadar uzaklıkta olabilirler. Bu yöntem sayesinde, hassas özniteliklerin korunması sağlanırken, diğer yöntemlerde olduğu gibi, yarı tanımlayıcı öznitelikleri de anonimleştirilir.

3.1.2.8. Kişisel verileri gizliliği

Bu teknik, farklı birçok gizlilik koruma yöntemini içerebilir, ancak kişisel verilerin korunması bağlamında farklı tekniklerin birleştirilmesi ile oluşturulur.

Kişisel veriler gizliliği teknikleri, kişisel verilerin korunması için farklı tekniklerden yararlanır. Bunlar arasında örneğin farklılaştırılmış gizlilik, homomorfik şifreleme, veri maskelenmesi gibi teknikler yer alabilir.

Farklılaştırılmış gizlilik, veriye rastgele gürültü ekleyerek kişilerin kimliklerinin

tanınmasını zorlaştırır. Homomorfik şifreleme, verilerin şifreli olarak depolanmasını ve hesaplamaların bu şifreli veriler üzerinde yapılmasını sağlar.

Veri maskelenmesi teknikleri, kişilerin kimliklerini belirlemek için kullanılabilecek özellikleri gizleyerek verilerin gizliliğini korur.

Kişisel veriler gizliliği teknikleri, verilerin saklanması, işlenmesi ve paylaşımı sırasında gizlilik risklerinin azaltılmasına yardımcı olur. Bu teknikler, özellikle sağlık gibi hassas verilerin saklanması veya paylaşımı söz konusu olduğunda, özellikle önemlidir.

3.1.2.9. Fayda temelli gizlilik koruma

Veri kullanışlılığı (yani analiz veya diğer amaçlar için verilerin yararlılığı) ile gizlilik koruma ihtiyacı arasında bir denge yaklaşımıdır. İki hedef arasında bir denge bulmayı amaçlar.

Verilerin kullanışlılığı artırılmak istendiğinde, kişisel bilgilerin gizliliğini korumak amacıyla çeşitli teknikler kullanılabilir. Bu teknikler arasında veri anonimleştirme, veri bölme, veri gizleme ve veri şifreleme gibi yöntemler yer alabilir. Amacı, verilerin yararlılığını en üst seviyede tutarak ve kişisel bilgilerin gizliliğini korumak arasında bir denge sağlamaktır.

3.1.3. Kriptografik temelli yöntemler

Gizliliğin bir modelini sunarak, kanıtlama ve nitelendirme metodolojileri de dahil olmak üzere gizliliğin tanımlanmış bir model sunması nedeniyle büyük popülerlik kazanmıştır. İkinci olarak, kriptografik algoritmaların geniş bir araç seti ile gizlilik koruyan veri madenciliği algoritmalarını uygulamak maksadıyladır. Ancak kriptografi, bir hesaplama işleminin çıktısını korumaz. Bunun yerine, hesaplama işlemi sırasında gizlilik sızıntılarını önler. Veri kaynakları ağ üzerinde dağılmış olabilir ve merkezi bir konumda toplamak hesaplama ve iletişim kaynaklarındaki kısıtlamalar nedeniyle mümkün olmayabilir.

Dağıtılmış uygulamalar, verileri dikey veya yatay veri modeli olarak adlandırılan iki farklı yapıda depolarlar.

Dağıtılmış veri madenciliği, verileri tek bir konuma toplamadan dağıtılmış bir ortamda veri madenciliği yapmak için algoritmalar sağlar. Ancak gizlilik nedenleriyle, organizasyonlar birbirleriyle verilerini paylaşmaktan çekinebilirler.

Güvenli çok paydaşlı hesaplama teknikleri, dağıtılmış veri madenciliğinde gizlilik korumaya yönelik bir yaklaşım olarak kullanılır. İlk güvenli çok paydaşlı hesaplama, verileri iki parti arasında bölümlendirmek için teoride güvenli devre değerlendirmesi için yapılmıştır.

Herhangi bir fonksiyonu hesaplamak için kullanılabilir ve hesaplanan çıktıdan başka hiçbir şeyi her iki tarafa da açığa vurmadan yapar. Ancak, veri madenciliği genellikle milyonlarca veya milyarlarca veri öğesini içerdiğinden, bu protokollerin iletişim maliyetleri bu amaçlar için pratik olmaktan çıkar.

Bu, verimli iletişim karmaşıklığına sahip probleme özgü protokollerin arayışına yol açmıştır. Lindell ve Pinkas, protokollerinde kriptografik teknikler kullandılar.

3.1.3.1. Güvenli çok paydaşlı hesaplama

Birden fazla tarafa ait özel verilerin işlenmesi veya hesaplamaların yapılması sırasında gizliliğin korunmasını maksadıyla başvurulan bir metottur. Bu metot, özel verilerin ortak bir veri merkezinde toplanmasını gerektirmeden, verilerin güvenli bir şekilde işlenmesine olanak tanır. Güvenli çok paydaşlı hesaplama protokolleri, farklı taraflar arasında veri bölünmüş olsa bile, özel verilerin ifşa edilmesini önleyecek şekilde tasarlanmıştır. Bu sayede, gizli verilerin korunması sağlanarak, verilerin işlenmesi için güvenli bir ortam oluşturulur.

3.1.3.2. Yatay olarak bölünmüş veri

Bir veri kümesini, her veri kaynağını ayrı ayrı gösteren farklı kayıtların her birine bölme işlemidir. Yatay bölünmüş bir veri kümesi örneği, bir şirketin çalışanlarına ait veri tabanları olabilir. Bu veri kümesi her bir kaydın ayrı bir çalışan hakkındaki bilgileri içerdiği birçok kayıttan oluşur. Bu veri tabanı yatay olarak bölünebilir ve farklı veri kaynaklarına dağıtılabilir. Örneğin, bir kaynak sadece çalışanların adı ve soyadı bilgilerini içerebilirken, diğer bir kaynak çalışanların maaş bilgilerini içerebilir. Veriler bu şekilde yatay olarak bölündüğünde, her veri kaynağına erişim sağlamak için birleştirme işlemi yapmak gerekebilir.

3.1.3.3. Dikey olarak bölünmüş veri

Bir veri kümesinin farklı sütunlarına veya özelliklerine göre ayrılmasıdır. Örneğin, bir kişinin adı, yaşadığı şehir, adresi, posta kodu, telefon numarası gibi özelliklerinin ayrı ayrı tutulduğu bir veri kümesi dikey olarak bölünmüş veriye örnek olarak verilebilir.

Dikey olarak bölünmüş veriler, veri tabanı tablolarında sütun tabanlı veya dikey tabanlı olarak depolanabilir.

Dikey olarak bölünmüş veriler, gizlilik koruması sağlamak için kullanılabilecek bir gizliliği koruyan veri madenciliği tekniğidir.

3.2. Gizlilik Korumalı Veri Yayınlama

Gizlilik Korumalı Veri Yayınlama, hassas verilerin açık bir şekilde yayınlanmasını mâni olmak maksadıyla başvurulan bir metottur. Metot, özellikle büyük veri kümeleriyle ilgilenen kuruluşlar için önemlidir.

Veri yayınlama işlemi, birçok kez verilerin toplumun yararına kullanılmasına yardımcı olmak için yapılırken, aynı zamanda kişisel bilgilerin açık bir şekilde paylaşılmasıyla gizlilik riski taşır.

Gizlilik korumalı veri yayınlama teknikleri, örneğin anonimleştirme, veri pertürbasyonu, veri gizleme gibi yöntemler kullanarak, hassas verilerin anonimleştirilmesi veya modifiye edilerek yayınlanmasını sağlar. Bu yöntemler, verilerin gizliliğini korumaya yardımcı olurken, verilerin hala kullanılabilirliğini ve doğruluğunu korur.

Veri gizliliği ve kullanılabilirliği zıt ilişkilidir. Verileri son kullanıcılara, analistlere yayınlamadan veya muhafaza etmeden evvel bir dizi değişikliğe uğratmak için birden fazla metot önerilmiştir. (Fung ve diğ., 2010; Wong ve Fu, 2010). PPDP, gizliliği korumak için sıklıkla anonimleştirme tekniklerine başvurur. Veriler, hassasiyeti yüksek ve özel olarak kabul edilen birden fazla kayıttan oluşur.

Verilerin paylaşımını ve dağıtımını daha güvenli kılmak maksadıyla, içerisinde birçok teknik ve yöntemi barındıran bir disiplin olarak tanımlanabilir. Bu teknikler, özellikle duyarlı verilerin anonimleştirilmesi ve gizliliğinin korunması amacıyla kullanılır. PPDP, istatistiksel analizde, veri madenciliğinde sıklıkla tercih edilir.

3.3. Gizlilik ve Kullanılabilirlik Dengesi

Yüksek seviyeli veri anonimleştirme, verinin gizliliğinin iyi korunduğunu gösteren bir parametredir. Fakat bu durum verinin kullanılabilirliğini etkiyebilir. Bu durumda veriden çıkarılan değeri azalır. Büyük verilerde gizlilik ve kullanılabilirlik arasındaki dengeyi sağlamaya çalışmak çok önemlidir. Bu denge her daim gözetilmelidir. Bilgi

kaydı verinin kullanılabilirliğindeki azalma ile ölçülür.

Bilgi kaybını saptamak üzere literatürde çeşitli metotlar vardır. Gizlilik korumalı veri yayınlama metotlarından gizlilik ve kullanılabilirlik arasındaki dengeyi yakalayabilmek için sıklıkla açgözlü yaklaşım tercih edilir.

Bu metotlarda bilgi kaybı ve kişisel bilgilerin korunması için belli ölçütler tanımlanarak çoklu tablolar oluşturulur. Anonimleştirme işlemleri boyunca optimum gizlilik ihtiyaçlarını sağlamaya çalışır. Açgözlü bir metottun çıktısı en az bilgi kaybı olan tablodadır (Mehmood ve diğ., 2016).

Gizliliği ölçmek halihazırda zor bir işlemdir. Aslında, bir veri sadece bir veri sahibinden toplanmış ise veri sahibi, üçüncü bir şahısla ne miktarda ve ne çeşit bilgi paylaşacağını kendisi karar verir.

Üçüncü şahıslara verilen bilgi sonraki aşamalarda bazı bilgi kayıplarına maruz kalma ihtimali her daim olabilmektedir. Kişisel verilerinin gizliliği konusunda ihtiyatlı bireyler, gizlilikleri hakkında daha az endişe duyanlardan daha fazla kayıp görebilirler (Mehmood ve diğ., 2016).

Bir veri madenciliği algoritmasının performansı aşağıdaki kriterlerle ölçülür:

1. Performans: Bir madencilik algoritmasının performansı, gizlilik kriterlerini elde etmek için gereken zaman açısından ölçülür.
2. Veri Faydalılığı: Verilerin PPDM algoritmalarının olmadığı durumlarda sağlanabilecek sonuçları sağlama işlevindeki kayıpları veya bilgi kaybını ölçer.
3. Belirsizlik düzeyi: Gizlenen hassas bilginin hala ne kadar belirsiz bir şekilde tahmin edilebileceği ölçüsüdür.
4. Direnç: Çeşitli veri madenciliği algoritmalarına ve modellerine karşı gösterilen toleransı ölçen bir kriterdir.

Bu kriterlerin tümü daha iyi bir şekilde özellik koruyan algoritmaların değerlendirilmesi için belirlenmiştir. Ancak gizlilik ve bilgi kaybı ölçümü en önemli iki kriterdir. Gizlilik ölçümü veya gizlilik metriği, bir öznitelik değerinin ne kadar yakından tahmin edilebileceğini gösteren bir ölçüttür. Eğer özgün veri kümesi daha yüksek bir güvenilirlikle tahmin edilebilirse, gizlilik düşüktür. Bilgi kaybındaki hassasiyetsizlik, veri madenciliği amacının başarısız olmasına neden olur. Bu yüzden, gizlilik ve bilgi kaybı arasında bir denge kurulması gerekmektedir.

4. GEREÇ VE YÖNTEMLER

Çalışmanın bu bölümünde yer verilen veri setlerinin özelliklerine, dönüştürme işlemlerine, modellenmesine, model performansını ölçmek için bölünme yöntemlerine, performans metriklerine, sınıflandırma yöntemlerine, gizlilik koruma algoritmasına ve gizliliği korunmuş veri setlerinin tanımlanmasındaki yöntemlere, kullanılan yazılım altyapılarına yer verilmiştir.

4.1. Veri Setleri

Önerilen veri gizliliği yöntemlerinin performans göstergelerini kıyaslayabilmek için öncelikle sınıflandırma yöntemlerinin kullanılmasına karar verilmiştir. Çalışmada sınıflandırma yöntemleri vasıtasıyla tahminleme yapılabilecek gerçek veri setleri tercih edilmiştir.

Bu veri setlerinin seçim kararında bir diğer tercih nedeni de içerisinde tanımlayıcı, yarı tanımlayıcı veriler içermesi ve gizlilik analizi yapmaya olanak tanımasıdır. Ayrıca veri setleri açık kaynaklı platformlar üzerinden herkesin ulaşabileceğe diğer çalışmalarda kıyaslamalara olanak verebilecek verilerden oluşmaktadır.

Bu çalışmada gizlilik analizleri kapsamında dört veri setine yer verilmiştir.

4.1.1. Titanic veri seti

Titanic veri seti, Titanic gemisinin yolculuğu sırasında hayatta kalan ve hayatını kaybeden yolcuların bazı özelliklerini içeren bir veri setidir.

Titanic veri seti data.world (data. world, Inc., Austin, Teksas merkezli bir özel şirkettir. Şirket, 2015 yılında kurulmuştur ve amacı, dünya genelindeki verileri daha erişilebilir hale getirmek ve insanların bu verileri işlemelerine ve anlamalarına yardımcı olmaktır. Veri seti 1310 kayıt içermektedir ve eksik değeri olmayan toplam kayıt sayısı 1043'tür. Veri seti bu çalışmada kullanılmadan önce eksik değer içeren kayıtlardan silinmiştir. Veri setindeki öznitelik sayısı 7 sayısal ve 7 kategorik olmak üzere 14'tür. Bu veri setinde yer alan tanımlayıcı nitelikler: name, pclass, sex, age, ticket, sibsp, fare, parch, cabin, embarked, boat, body, home.dest, survived değişkenleridir.

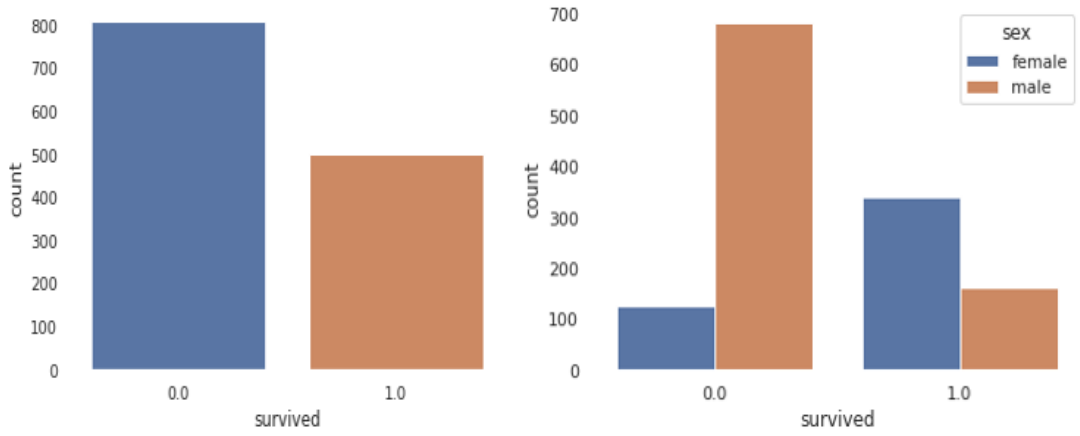
Titanic veri setinin detaylı öznitelikleri Tablo 3.1'de gösterilmektedir. Bunlardan tanımlayıcı özelliğe sahip nitelikler olduğundan dolayı name, ticket, cabin, home.dest veri setinden kaldırılmıştır.

Yine veri setinde yer alan ve içerinden çok fazla eksik veri içeren boat, body değişkenleri de veri setinden kaldırılmıştır.

Böylelikle gizliliği korunacak ve sınıflandırılacak veri setindeki nitelik sayısı 8'e indirilmiştir. Bu değişkenler sex, age, pclass, sibsp, fare, parch, embarked, survived çalışmada kullanılmıştır. Veri setinde döndürülen ve gauss gürültüsü eklenen özellikler "*age" ve "pclass"* değerleridir.

Tablo 4.1. Titanic veri seti.

Nitelik Adı	Nitelik Tanımı	Nitelik Türü	Tanım Kümesi
sex	Cinsiyet	Kategorik	male, female
age	Yaş	Sayısal	[0,1667-80]
pclass	Yolcu Sınıfı	Sayısal	[1-3]
sibsp	Eş/Kardeş Sayısı	Sayısal	[0-8]
fare	Ücret	Sayısal	[0-512329]
parch	Ebeveyn/Çocuk Sayısı	Sayısal	[0-9]
embarked	Gemiye Biniş Limanı	Kategorik	S,Q,C
survived	Hayatta Kalma Durumu	Sayısal	0-1
name	Yolcu Adı	Kategorik	AllenMiss.Elisabeth Watson
ticket	Bilet Numarası	Kategorik	113781-24160
cabin	Kabin Numarası	Kategorik	B5,C22,C23,C24 vs.
boat	Bindiği Bot Numarası	Kategorik	[1-1170]
body	Ölenlerin Bulunup/Bulunamaması	Sayısal	[1-328]
home.dest	Ev Adresi	Kategorik	Montreal, PQ / Chesterville, ON



Şekil 4.1. Titanic veri seti python programı ile veri analizi/görselleştirme.

4.1.2. Göğüs kanseri veri seti

Bu veri seti, meme kanseri hakkında bilgiler içeren Birleşmiş Milletler Kalkınma Programı Nepal ile iş birliği sonucunda ortaya çıkan içinde meme kanseri olan hastalara ait 569 adet kayda ve 32 değişkene sahip bir gerçek veri setidir.

Bu veriler, hastaların demografik özellikleri, klinik bulguları ve teşhis sonuçları gibi meme kanseri ile ilgili önemli bilgileri içermektedir. Bu veri seti meme kanseri belirtileri ile hücrelerin anormal büyümesini tahmin etmeye yönelik sınıflandırma yöntemlerinin kullanımına imkân verdiği için tercih edilmiştir.

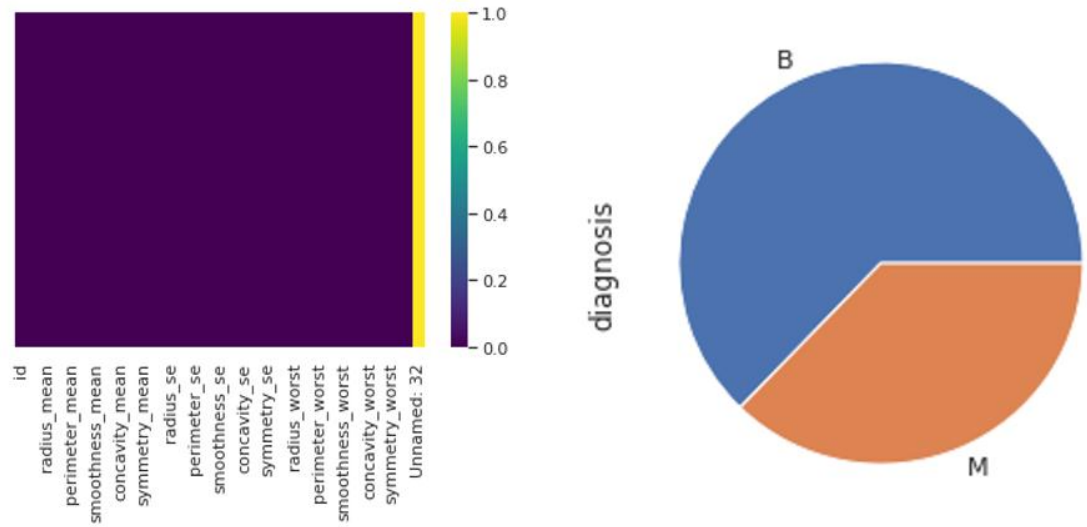
Bu veri setinde yer alan tanımlayıcı nitelik id hasta kayıt numarası değişkeni veri setinden çıkarılmıştır. Ayrıca Unnamed32 değişkenin içeriğinde veri olmadığı için bu veri setinde kullanılmamıştır. Göğüs kanseri veri setinin detaylı öznitelikleri Tablo 4.2'de gösterilmektedir. Veri setinde döndürülen ve gauss gürültüsü eklenen özellikler *texture_mean ve radius_mean* değerleri rastgele seçilmiştir.

Tablo 4.2. Göğüs kanseri veri seti.

Nitelik Adı	Nitelik Türü	Tanım Kümesi
id	Sayısal	[8670-911M]
diagnosis	Kategorik	[M-B]
radius_mean	Sayısal	[6.98-28.1]
texture_mean	Sayısal	[9.71-39.3]
perimeter_mean	Sayısal	[43.8-189]
area_mean	Sayısal	[144-2.5K]
smoothness_mean	Sayısal	[0.05-0.16]
compactness_mean	Sayısal	[0.02-0.35]
concavity_mean	Sayısal	[0-0.43]
concave_points_mean	Sayısal	[0-0.2]

Tablo 4.2. (Devamı) Göğüs kanseri veri seti.

Nitelik Adı	Nitelik Türü	Tanım Kümesi
symmetry_mean	Sayısal	[0.11-0.3]
fractal_dimension_mean	Sayısal	[0.05-0.1]
radius_se	Sayısal	[0.11-2.87]
texture_se	Sayısal	[0.36-4.88]
perimeter_se	Sayısal	[0.76-22]
area_se	Sayısal	[6.8-542]
smoothness_se	Sayısal	[0-0.03]
compactness_se	Sayısal	[0-0.14]
concavity_se	Sayısal	[0-0.4]
concave_points_se	Sayısal	[0-0.5]
symmetry_se	Sayısal	[0.1-0.08]
fractal_dimension_se	Sayısal	[0-0.03]
radius_worst	Sayısal	[7.93-36]
texture_worst	Sayısal	[12-49.5]
perimeter_worst	Sayısal	[50.4-251]
area_worst	Sayısal	[185-4.25K]
smoothness_worst	Sayısal	[0.07-0.22]
compactness_worst	Sayısal	[0.03-1.06]
concavity_worst	Sayısal	[0-1.25]
concave_points_worst	Sayısal	[0-0.29]
symmetry_worst	Sayısal	[0.16-0.66]
fractal_dimension_worst	Sayısal	[0.06-0.21]
Unnamed:32	Sayısal	N/A



Şekil 4.2. Göğüs kanseri veri seti python programı ile veri analizi/görselleştirme.

4.1.3. Kalp krizi veri seti

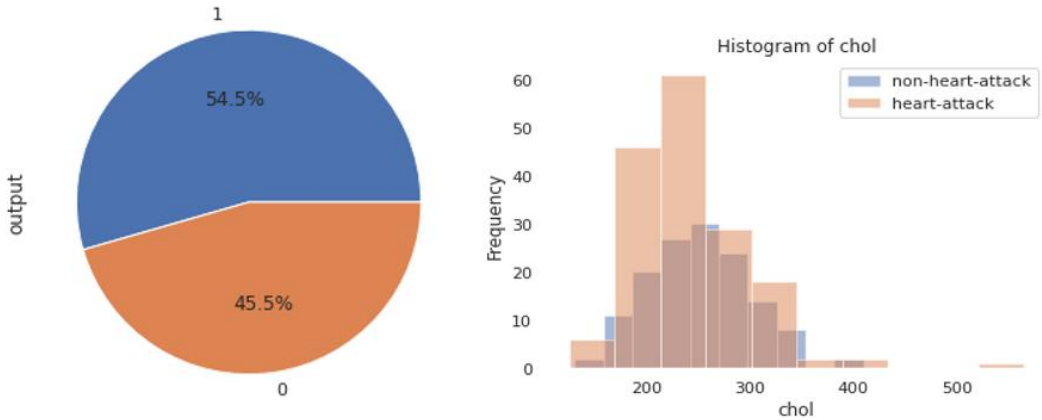
Tabloda 1988 yılına kadar dayanan ve dört veri tabanını içeren bir veri kümesi gösterilmektdir. Cleveland, Macaristan, İsviçre ve Long Beach V. Bunlar arasında 76 öznitelik bulunmaktadır, ancak çalışmada bu veri setinin 14 adet özelliğiyle birlikte

303 adet kayıttan oluşan bir alt kümeye yer verilmiştir. "output" değişkeni, hastada kalp hastalığının varlığına işaret eder. 0 = hastalık yok ve 1 = hastalık olan olmak üzere tam sayı değerine sahiptir.

Ayrıca Unnamed32 değişkenin içeriğinde veri olmadığı için bu veri setinde kullanılmamıştır. Kanser veri setinin detaylı öznitelikleri Tablo 4.4.' de gösterilmektedir. Veri setinde döndürülen ve gauss gürültüsü eklenen özellikler *age ve chol* değerleri seçilmiştir.

Tablo 4.3. Kalp krizi veri seti.

Nitelik Adı	Nitelik Türü	Nitelik Tanımı	Tanım Kümesi
age	Sayısal	Hasta Yaşı	[29-77]
sex	Sayısal	Cinsiyeti	1 male,0 female
cp	Sayısal	Göğüs Ağrısı Çeşidi	1,2,3
trtbps	Sayısal	Kan Basıncı	[94-200]
chol	Sayısal	Kolesterol Seviyesi	[126-564]
fbs	Sayısal	Açlık Kan Şekeri	0,1
restecg	Sayısal	EKG Sonucu	0,1,2
thalachh	Sayısal	Maks. Kalp Atım Hızı	[71-202]
exng	Sayısal	Angijo Durumu	0,1
oldpeak	Sayısal	Efor Testi Sonucu ST	[0-6.2]
slp	Sayısal	Efor Segmentinin Eğilimi	0,1,2
caa	Sayısal	Kapalı Damar Sayısı	0,1,2,3,4
thall	Sayısal	Kusur Sayısı	1,2,3
output	Sayısal	Kalp Krizi Riski	0,1



Şekil 4.3. Kalp krizi veri seti python programı ile veri analizi/görselleştirme.

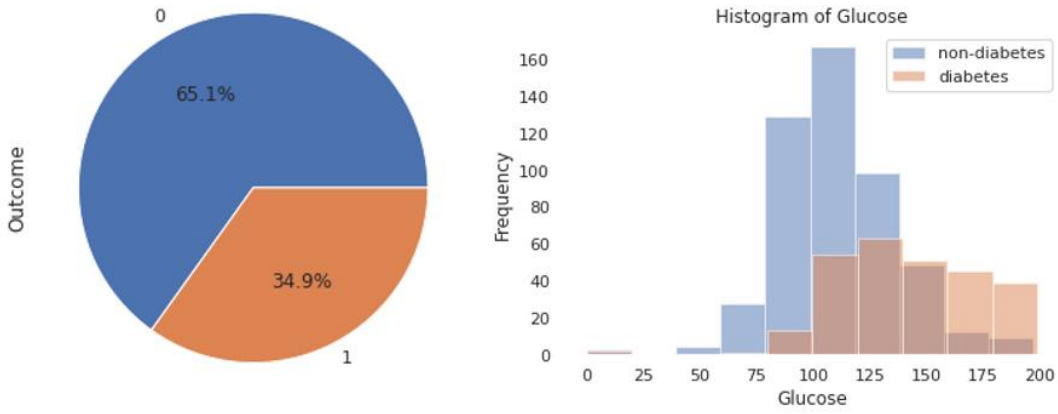
4.1.4. Diyabet veri seti

Diyabet veri seti hâlihazırda Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü'nden temin edilmiştir. Veri setinin amacı, veri kümesine dahil edilen belirli

teşhis ölçümlerine dayanarak bir hastanın diyabet olup olmadığını sınıflandırma yöntemleri ile tahmin etmektir. Bu örneklemelerin seçimine çeşitli kısıtlamalar getirilmiştir. Özellikle, veri setindeki bütün hastalar, 21 yaşında veya daha büyük ve Hint kökenli kadınlardır. 9 adet özelliğiyle birlikte 768 adet kayıttan oluşan veri seti "Outcome" değişkeni, diyabet hastalığının varlığına işaret eder. 0 = hastalık yok ve 1 = hastalık olan olmak üzere tam sayı değerine sahiptir. Veri setinde döndürülen ve gauss gürültüsü eklenen özellikler "*Age" ve "Glucose"* değerleridir. Diyabet veri seti öznitelikleri Tablo 4.5' de verilmiştir.

Tablo 4.4. Diyabet Veri Seti.

Nitelik Adı	Nitelik Türü	Nitelik Tanımı	Tanım Kümesi
Glucose	Sayısal	Kandaki Glikoz Miktarı	[0-199]
BloodPressure	Sayısal	Kan Basıncı	[0-122]
SkinThickness	Sayısal	Derinin Kalınlığı	[0-99]
Insulin	Sayısal	Kandaki İnsulin Miktarı	[0-846]
BMI	Sayısal	Vücut Kitle Endeksi	[0-67.1]
DiabetesPedigreeFunction	Sayısal	Diyabet Yüzdesi	[0.08-2.42]
Age	Sayısal	Yaş	[21-81]
Outcome	Sayısal	Diyabet Hastası	0,1



Şekil 4.4. Diyabet veri seti python programı ile veri analizi/görselleştirme.

4.2. Veri Ön İşleme Dönüştürme Yöntemleri

Veri ön işleme ve dönüştürme yöntemleri, veri analitiği ve makine öğrenimi projelerinde kullanılan veri hazırlama süreçleridir. Sayısal ve kategorik dönüştürme işlemleri olmak üzere ikiye ayrılabilir.

4.2.1. Sayısal değişkenlerde ölçeklendirme yöntemleri

Sayısal değişkenlerin ölçeklerinin farklı olması, verilerin karşılaştırılabilir olmasını zorlaştırabilir. Bu nedenle, verilerin bir arada değerlendirilmesi için ölçeklendirme yöntemleri kullanılabilir. Sayısal değişkenlerde kullanılan ölçeklendirme yöntemleri aşağıdaki gibi sıralanabilir**:**

4.2.1.1. Normalleştirme (min-maks ölçeklendirme)

Bu yöntem, verilerin belirli bir aralıkta yer almasını sağlar. Verilerin minimum ve maksimum değerleri belirlenir ve bu değerlere göre veriler belirli bir aralıkta ölçeklenir. Örneğin, verilerin 0-1 aralığına ölçeklenmesi gibidir. Çalışmada bazı veriler için min-maks normalleştirme yöntemi kullanılmıştır.

4.2.1.2. Standartlaştırma (z-skoru)

Bu yöntemde, verilerin ortalaması ve standart sapması hesaplanır. Verilerin ortalaması çıkarılarak, standart sapmaya bölünür ve bu sayede verilerin standart sapmaya göre ölçeklenmesi sağlanır**.**

4.2.1.3. Skor dönüşümü

Bu yöntemde, verilerin normal dağılımına uygun hale getirilmesi amaçlanır. Verilerin logaritması, karekökü veya tersi gibi dönüşümler kullanılarak, verilerin normal dağılıma daha yakın hale getirilmesi sağlanır.

4.2.1.4. Robust ölçeklendirme

Bu yöntem, verilerin dağılımında yer alan aykırı değerlerin etkisini azaltmak için kullanılır. Medyan ve çeyrekler arası aralık gibi aykırı değerlerden etkilenmeyen istatistikler kullanılarak, veriler ölçeklenir. Çalışmada bazı veriler için robust ölçeklendirme yöntemi kullanılmıştır.

4.2.2. Kategorik değişkenlerde dönüştürme yöntemleri

Kategorik veriler, genellikle nominal veya ordinal ölçekleme seviyelerine sahip veri tipleridir ve sayısal olmayan kategorilere veya sınıflara ait verileri temsil eder. Kategorik verilerin dönüştürülmesi, veri analitiği, makine öğrenimi, istatistiksel analiz ve veri işleme gibi çeşitli disiplinlerde geniş kullanımı olan bir yöntemdir. Kategorik verilerin dönüştürülmesi için kullanılan bazı yöntemler:

4.2.2.1. Etiketleme kodlama (label encoding)

Bu yöntemde, kategorik veriye sıralı ve benzersiz sayısal etiketler atanır. Her bir kategori, bir sayıya dönüştürülür. Bu yöntem, ordinal kategorik verilerin dönüştürülmesinde kullanılabilir, yani kategorilerin doğrusal bir sırası veya hiyerarşisi bulunuyorsa tercih edilir.

4.2.2.2. Tek-çizgi kodlama (one-hot encoding)

Bu yöntemde, her kategori için ayrı bir sütun oluşturulur ve o kategoriye ait veriler için 1 veya 0 gibi ikili değerler atanır. Bu yöntem, nominal kategorik verilerin dönüştürülmesinde kullanılabilir, yani kategorilerin arasında doğrusal bir sıra veya hiyerarşi yoksa tercih edilir.

4.2.2.3. İkili kodlama (binary encoding)

Bu yöntemde, kategorik veri, ikili (0 ve 1) değerler kullanarak dönüştürülür. Kategoriler, ikili bitlerle temsil edilir ve her bir bit, bir kategoriye ait bir özniteliği belirtir. Bu yöntem, veri boyutunu azaltmak için kullanılabilir, özellikle çok sayıda kategori içeren verilerde etkili olabilir.

4.2.2.4. Sayım kodlama (count encoding)

Bu yöntemde, her kategori, o kategorinin veri kümesindeki görünme sayısı ile kodlanır. Bu yöntem, nadir kategorilerin etkisini azaltmak için kullanılabilir ve nadir kategorilerin sık kategorilere oranla daha yüksek bir ağırlık taşımasını sağlayabilir.

4.2.2.5. Hedef kodlama (target encoding)

Bu yöntemde, kategorik değişken, hedef değişkenin sınıf etiketleri ile kodlanır. Bu yöntem, sınıf dengesizliğinin olduğu sınıflandırma problemlerinde veya hedef değişkenle ilişkili bilgiyi kodlamak istediğimiz durumlarda kullanılabilir.

4.2.2.6. Sahte değişken kodlama (dummy variable)

Kategorik verilerin dönüştürülmesi için kullanılan bir kodlama yöntemidir. Sahte değişken yöntemi, nominal kategorik verilerin dönüştürülmesi için kullanılır ve her bir kategori için ayrı bir sütun oluşturulur ve 1 veya 0 gibi ikili değerlerle doldurulur.

Sahte değişken yöntemi, her bir kategoriye ait bir ikili sütun oluşturarak kategorileri sayısal formda ifade eder. Eğer bir veri örneği bir kategoriye aitse ilgili sütun 1 olarak işaretlenir, diğer tüm sütunlar ise 0 olarak kalır. Bu yöntem, kategorik verilerin sayısal

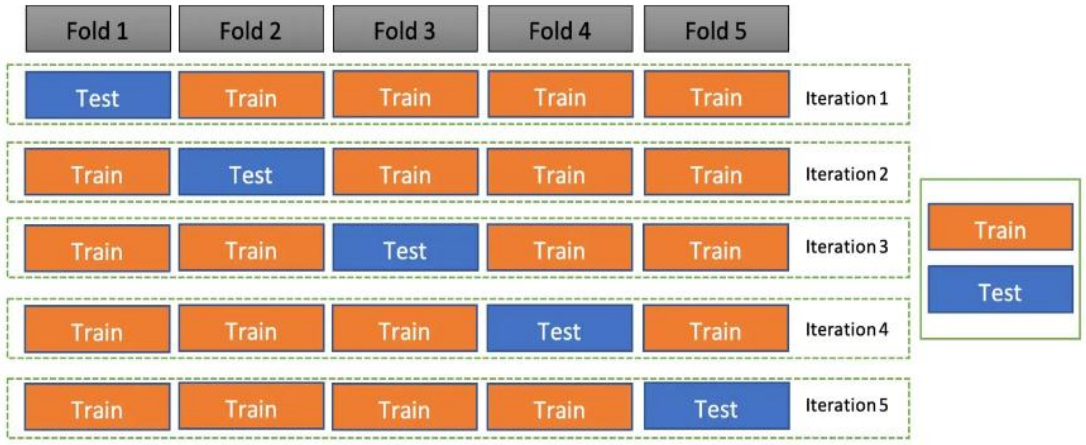
formda ifade edilmesini sağlar ve bu sayede makine öğrenimi algoritmalarının kategorik verileri anlamasına olanak tanır.

Sahte değişken kodlaması, özellikle nominal kategorik verilerin olduğu durumlarda tercih edilir, yani kategorilerin arasında doğrusal bir sıra veya hiyerarşi bulunmuyorsa kullanılır. Örneğin, cinsiyet (erkek, kadın), ülke (Türkiye, ABD, Çin), meslek (doktor, avukat, mühendis) gibi kategorik verilerin sahte değişken kodlaması yapılabilir.

Çalışmada cinsiyet (gender/sex) ve ülke (country) değişkeni için sahte değişken kodlaması yapılmıştır.

4.3. Veri Bölümleme Yöntemleri

K-kat çapraz doğrulama (k-fold cross-validation), bir makine öğrenimi modelinin performansını değerlendirmek için yaygın olarak kullanılan bir yöntemdir. Veri seti k sayıda ancak eşit kayıtta parçaya bölünür ve her biri sırayla test ve eğitim verisi olarak kullanılır. Yani, modelde, her seferinde farklı bir alt küme kullanılarak k defa eğitilir ve k defa test edilir. Bu yöntem, veri setinin her bir parçasının test verisi olarak kullanılması sayesinde modelin genelleştirilmesinin ölçülmesine yardımcı olur. Sonuçlar, k defa ölçümün ortalaması alınarak hesaplanır. k değeri, çapraz geçerleme sırasında kullanılacak alt kümelerin sayısını belirler ve genellikle 5 veya 10 olarak seçilir. k-kat çapraz geçerleme, aşırı örenme gibi sorunları azaltarak modelin performansını daha güvenilir bir şekilde değerlendirmeye yardımcı olur. Deneysel çalışmalar ideal *k* kat sayısının 5 olduğunu göstermektedir (Olson ve Delen, 2008). *k*'nın 5 olduğu *k*-kat çapraz geçerlemenin görselleştirilmesi Şekil 4.5' de (Olson ve Delen, 2008) verilmektedir.



Şekil 4.5. Çapraz doğrulama yöntemi (k=5).

Avantajları:

- Daha doğru performans ölçümü: K-kat çapraz doğrulama, modelin performansını daha doğru bir şekilde ölçmek için kullanılabilir. Bu yöntem sayesinde, modelin aşırı öğrenme ezberleme ihtimali azaltılabilir.
- Verimli kaynak kullanımı: K-kat çapraz doğrulama, sınırlı bir veri kümesi üzerinde çalışırken, veri kümesinin tamamını eğitim ve test için kullanmamızı sağlar. Bu, verimli bir veri kullanımı anlamına gelir.
- Parametre ayarlaması için kullanılabilir: K-kat çapraz doğrulama, parametrelerin en iyi şekilde ayarlanmasına yardımcı olabilir. Bu nedenle, modelin performansını daha da iyileştirebilir.

Dezavantajları:

- Yavaş çalışabilir: K-kat çapraz doğrulama, eğitim ve test setlerinin birden çok kez yeniden oluşturulması gerektiğinden, diğer doğrulama yöntemlerine göre daha yavaş çalışabilir.
- Hiperparametrelerin seçimi zor olabilir: K-kat çapraz doğrulama ile birçok hiperparametrenin (model parametreleri) ayarlanması gerekebilir. Bu, modelin performansını iyileştirmek için fazladan çaba gerektirir.
- Veri kümesinin boyutuna bağımlı: Veri kümesinin boyutu, K-kat çapraz doğrulama sonuçlarını etkileyebilir. Küçük veri setleri için, K-kat çapraz doğrulama sonuçları güvenilir olmayabilir.
- Yapısal problemlerle karşılaşılabilir: K-kat çapraz doğrulama, belirli bir veri kümesinin yapısına göre özelleştirilmiş bir şekilde uygulanabilir. Veri kümesinin yapısına göre farklı K-kat çapraz doğrulama uygulamaları yapılması gerekebilir.

4.4. Sınıflandırma Yöntemleri

Veri madenciliğinde sınıflandırma, veri kümesinde yer alan nesnelerin belirli özelliklerine göre sınıflara ayrılması ve yeni veri kayıtlarının hangi sınıfa ait olduğunun tahmin edilmesi işlemidir. Sınıflandırma yöntemleri, veri kümesinde yer alan nesnelerin sınıflandırılmasını gerçekleştiren algoritmalar veya modellerdir.

Sınıflandırmada, veri tabanındaki kayıtlar, niteliklerine göre farklı sınıflara atanır ve

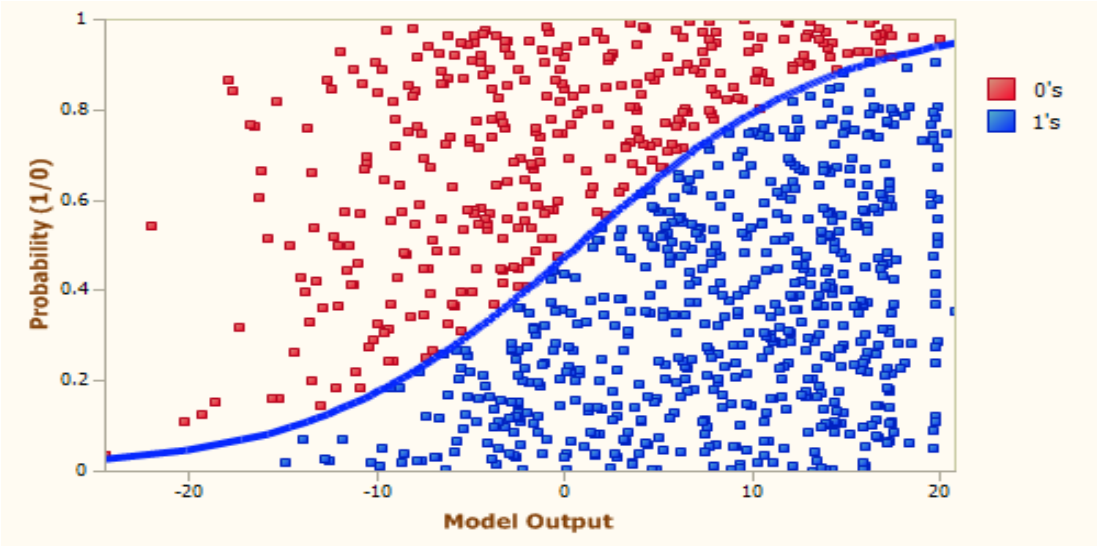
gelecekte veri kümesine dahil olacak yeni kayıtların hangi sınıfa ait olduğu belirlenmeye çalışılır. Sınıflandırma modeli, veri tabanındaki nesnelerin özelliklerini isimleriyle ilişkilendirir. Aynı veya farklı olan nesnelerin benzerliklerini ve farklılıklarını belirler ve daha sonra değerlendirilen özelliklere dayanarak nesnenin adını tahmin eder. Bir çocuğun cinsiyetini sınıflandırma süreci gibi. İlk başta çocuğun cinsiyet kavramı veya sınıflandırma bilgisi yoktur. Daha sonra anne, baba, teyze, amca, kendinden büyük ve küçük erkek ve kız çocuklarını görerek bir veri tabanı oluşturur. Çocuk, bu veri tabanına dayanarak kadınlar ve erkekler arasındaki temel farkları belirler ve sonrasında bilgilerine dayanarak tanımadığı bir çocuğun kız mı yoksa erkek mi olduğuna karar verir. Aslında bu tamamen bir sınıflandırma sürecidir.

Bu tez çalışmasında, önerilen algoritmanın sınıflandırma doğruluğu performansını, Lojistik Regresyon, K En Yakın Komşu, Yapay Sinir Ağları, Destek Vektör Makineleri, Gradyanı Artırılan Karar Ağaçları, Hafif Gradyanı Artırılmış Makineleri sınıflandırıcıları kullanılarak incelenmiştir.

4.4.1. Lojistik regresyon yöntemi (logistic regression)

Lojistik Regresyon, istatistiksel bir sınıflandırma yöntemidir ve veri setindeki bağımsız değişkenlerin bir veya daha fazla bağımlı değişken üzerindeki etkisini tahmin etmek için kullanılır. Genellikle ikili sınıflandırma problemleri için kullanılsa da çok sınıflı sınıflandırma problemlerinde de kullanılabilir.

Lojistik Regresyon, veri setindeki bağımsız değişkenlerin doğrusal bir kombinasyonunu kullanarak bağımlı değişkenin olasılığını hesaplar. Bu olasılık, [0,1] aralığında bir değer alır ve genellikle 0.5 eşik değeri kullanılarak bir sınıflandırma kararı alınır (Şekil 4.6.)



Şekil 4.6. Lojistik regresyon ikili sınıflandırma grafiği.

Lojistik Regresyon, çeşitli optimizasyon algoritmaları kullanılarak hesaplanabilir. En yaygın olarak kullanılan optimizasyon algoritması gradyan azaltımı algoritmasıdır. Bu algoritma, modelin parametrelerini veri setine uygun hale getirmek için iteratif olarak günceller.

Gradyan azaltımı, bir fonksiyonun minimum noktasını bulmak için kullanılan bir yöntemdir. Bu fonksiyonun değerini en aza indirecek noktayı bulmak için fonksiyonun türevini hesaplar ve türevin negatif yönünde ilerleyerek minimuma yaklaşmaya çalışır.

Gradyan azaltımı yöntemi, bir öğrenme algoritması olarak da kullanılır. Özellikle, makine öğrenmesindeki regresyon ve sınıflandırma problemlerinde kullanılır. Öğrenme sürecinde, bir hata fonksiyonunun değeri, model parametrelerinin (ağırlıkların) değiştirilmesiyle azaltılmaya çalışılır. Bu değişiklikler, gradyan azaltımı yöntemiyle hesaplanan bir gradyan vektörüne göre yapılır.

Lojistik Regresyon algoritmasını uygulamak için bir sınıf sağlar. Bu sınıf, önceden belirlenmiş parametrelerle (varsayılan değerler veya kullanıcının belirlediği) bir model oluşturmak için kullanılabilir. Daha sonra, bu model, veri setindeki bağımsız değişkenlerin bir kombinasyonunu kullanarak bağımlı değişkenin olasılığını hesaplamak için kullanılabilir.

Avantajları:

- Basit ve hızlı: Lojistik regresyon, modelleme ve tahmin yapma işlemlerinde hızlı ve basit bir algoritmadır.
- İyi performans: Lojistik regresyon, iyi performans gösterir ve doğru bir şekilde kullanılırsa, yüksek doğruluk oranları sağlar.
- Anlaşılabilir sonuçlar: Lojistik regresyon sonuçları, olasılık değerleri şeklinde elde edilir ve kolayca yorumlanabilir.
- Az veri gereksinimi: Lojistik regresyon, az miktarda veri ile de kullanılabilir ve daha az veriye ihtiyaç duyar.

Dezavantajları:

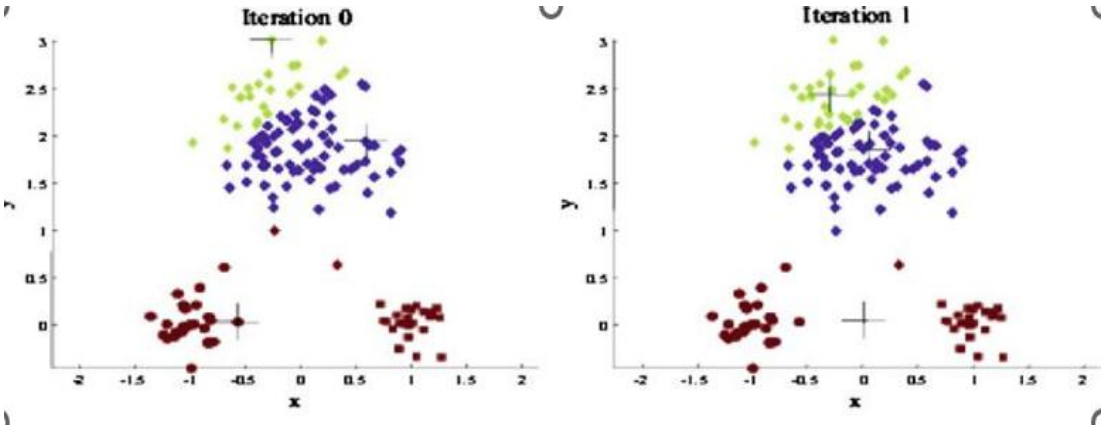
- Sadece iki sınıflandırma için kullanılabilir: Lojistik regresyon, sadece iki sınıf arasındaki sınıflandırma için kullanılabilir.
- Aşırı uyuma yatkınlığı: Lojistik regresyon, bazı durumlarda eğitim verilerine aşırı uyum sağlayabilir ve genelleştirme yapmakta zorlanabilir. Bu durum, düzenlileştirme yöntemleri ile önlenmelidir.
- Bağımlılık varsayımı: Lojistik regresyon, veriler arasındaki bağımlılık varsayımına dayanır. Bu varsayım, bazı durumlarda gerçeği yansıtmayabilir ve modelin doğruluğunu azaltabilir.
- Aykırı değerlere duyarlılık: Lojistik regresyon, aykırı verilere çok duyarlıdır. Bu nedenle, veri kümesindeki aykırı değerlerin tanımlanması ve yönetilmesi önemlidir.

4.4.2. K En yakın komşu yöntemi (kneighbors classifier)

K-En Yakın Komşuluk (KNN) yöntemi, 1967 yılında Edward Feigenbaum ve Julian Feldman tarafından ilk kez tanıtıldı. Bu yönteme göre, sınıflandırma işlemi veri tabanındaki her kaydın birbirine olan uzaklığına dayanarak gerçekleştirilir. Ancak, bazı durumlarda kaydın hangi ürünleri satın alabileceğini tahmin etmek zor olabilir. Bebek bezinin yanında bira alınması örneği verilebilir.

Dolayısıyla, KNN yöntemi bu tür karmaşık ilişkileri ortaya çıkarmada önemli bir rol oynar. Birliktelik kuralları, veriler arasındaki ilişkiyi destek ve güven kriterleri ile hesaplar. Destek kriteri, öğeler arasındaki ilişkinin ne kadar sık olduğunu ifade

ederken, güven kriteri, A öğesinin B öğesi ile birlikte olma olasılığını ifade eder. Örneğin, bir mağazada 500 müşterinin pantolon satın aldığı ve bunların 400'ünün pantolonun yanında gömlek de aldığı düşünüldüğünde, pantolon alındığında gömlek de almanın güven değeri %80 olarak hesaplanır. Bu da pantolon alan müşterilerin %80'inin aynı zamanda gömlek de aldığı sonucuna ulaştırır. Birliktelik kuralları, sık tekrarlanan öğelerin belirlenmesi ve bu sık tekrarlanan öğelerden güçlü birliktelik kurallarının oluşturulması adımlarını içerir. Bu kurallar, perakende sektörü başta olmak üzere, mağaza raflarının tasarımı, promosyon için hangi ürünlerin seçilmesi, özel ürünlere karar verme, katalog tasarımı gibi birçok alanda kullanılır.



Şekil 4.7. K En yakın komşu verilerin yakınlık durumlarına göre kümelenmesi.

K en yakın komşu (KNN) yöntemi, sınıflandırma ve regresyon problemlerinde kullanılan bir makine öğrenmesi algoritmasıdır. Bu yöntem, önceden tanımlanmış bir veri kümesindeki en yakın K sayıda veri örneğine dayalı olarak yeni bir veri örneğinin sınıflandırılmasını veya tahmin edilmesini sağlar.

Uygulama aşamaları şu şekildedir:

Veri kümesi hazırlanır ve sınıflandırma veya regresyon amacına uygun şekilde etiketlenir.

- Yeni bir veri örneği verilir.
- Bu yeni örnek ile veri kümesindeki diğer örnekler arasındaki uzaklıklar hesaplanır. Öklid mesafesi genellikle kullanılan bir uzaklık ölçüsüdür.
- En yakın K adet örnek belirlenir.
- Sınıflandırma problemleri için, bu K örneğin sınıfları incelenir ve en çok tekrar eden sınıf, yeni veri örneğinin sınıfı olarak tahmin edilir. Regresyon

problemleri için, K örneğin ortalaması alınır ve bu ortalama, yeni veri örneğinin tahmin değeri olarak kullanılır.

KNN yöntemi, basit ve anlaşılır bir algoritma olması nedeniyle sık kullanılan bir yöntemdir. Ancak, büyük veri kümelerinde yüksek hesaplama maliyeti nedeniyle uygulanması zor olabilir.

Avantajları:

- Basit ve kolay: KNN, basit bir algoritma olup uygulaması kolaydır.
- Eğitimli veriye ihtiyaç duymaz: KNN, eğitim verilerine ihtiyaç duymaz, bu nedenle çok az örnek veri ile de kullanılabilir.
- Çok sınıflı sınıflandırma: KNN, çok sınıflı sınıflandırmada da kullanılabilir ve bu durumda da yüksek doğruluk oranları sağlayabilir.
- Esnek: KNN, değişen veya geliştirilen veri setleri için esnek bir yapıya sahiptir.

Dezavantajları:

- Bellek kullanımı: KNN, büyük boyutlu veri setlerinde bellek kullanımı nedeniyle yavaş çalışabilir ve bellek kullanımı açısından dezavantajlıdır.
- Ağırlıklandırma problemleri: KNN, ağırlıklandırma problemleri ile karşılaşabilir. Bu nedenle, bazı örneklerin diğerlerinden daha fazla etkili olmasına neden olabilir ve yanlış sınıflandırma sonuçlarına neden olabilir.
- Boyut problemleri: KNN, boyutluluk problemleri ile karşılaşabilir. Veri setinin boyutu arttıkça, KNN'nin doğruluğu da azalabilir.
- Ölçeklendirme: KNN, ölçeklendirme yapılması gerektiğinde doğruluğunu arttırır. Bu nedenle, önceden işlenmiş bir veri seti kullanmadan önce ölçeklendirme yapmak gerekebilir.
- Belirsizlik: KNN, belirli bir örneğin birden fazla sınıfa ait olabileceği belirsizlik durumları ile başa çıkmakta zorlanabilir.

4.4.3. Sinir ağları yöntemi (neural network)

Sinir ağı, ilk kez 1943 yılında Warren McCulloch ve Walter Pitts tarafından önerilmiştir. Ancak, daha sonraki yıllarda, sinir ağlarının geniş uygulama alanlarına yönelik çalışmaları, özellikle 1980'lerde yoğunlaşmıştır.

Bir yapay sinir ağı beş aşamadan meydana gelmektedir. Bu aşamalar girdiler, çıktılar, ağırlık, toplam ve aktivasyon fonksiyonlarıdır. (Şen, 2004). Girdi, yapay sinir ağının öğrenilmesini istediğimiz veri seti tanımlarken, ağırlıklar, girdilerin çıktılara eşlemelerinin yapılacağı eğitim aracını anlatır. Böylece ağırlıklar en uygun hale getirilerek girdiler çıktılarla eşleştirilir. Ağırlıkların nasıl belirleneceği ve değiştireceği öğrenme algoritmaları vasıtasıyla yapılır.

Aynı zamanda, yapay sinir ağı, biyolojik sinir sistemlerinden matematiksel olarak ilham alarak, verilere dikkate değer bir şekilde uyum sağlayabilir ve öğrenebilirler.

Nöral ağlar, girdi, gizli ve çıktı katmanlarındaki yapay nöronlardan (veya düğümlerden) oluşur ve aralarındaki bağlantıları yöneten ağırlıklar bulunur. Toplam fonksiyonu, girdilerin ağırlıklı toplamını hesaplar ve aktivasyon fonksiyonu, sonucu çıktıya yönlendirmek için kullanılır. Eğer aktivasyon fonksiyonun sonucu belirlenmiş bir eşik değerinin altında ise çıktı üretilmez; ancak eşik değerini aşarsa çıktı oluşturulur. Kullanılan fonksiyonlar arasında doğrusal, adım, eşik, sigmoid, hiperbolik ve Gauss fonksiyonları bulunur. Ağın nihai çıktısı, transfer fonksiyonunun sonucunu aktivasyon fonksiyonundan geçirerek elde edilir ve sonuç, sonraki nöronlara girdi olarak kullanılabilir veya dış çevreye yönlendirilen çıktı olarak kullanılabilir.

Literatürde, sınıflandırma amaçlı yapay sinir ağı metodolojisinin birçok uygulaması bulunmaktadır. Örneğin, pazar segmentasyonu (Fish, Barnes ve Aiken, 1995), finansal başarı (Zhang vd., 1999), sağlık koşulları (Ottenbacher vd., 2001), gen tahmini (Chandra ve Babu, 2014) ve eğitim başarısı (Toprak, 2017) gibi alanlarda sınıflandırma çalışmaları yapılmaktadır.

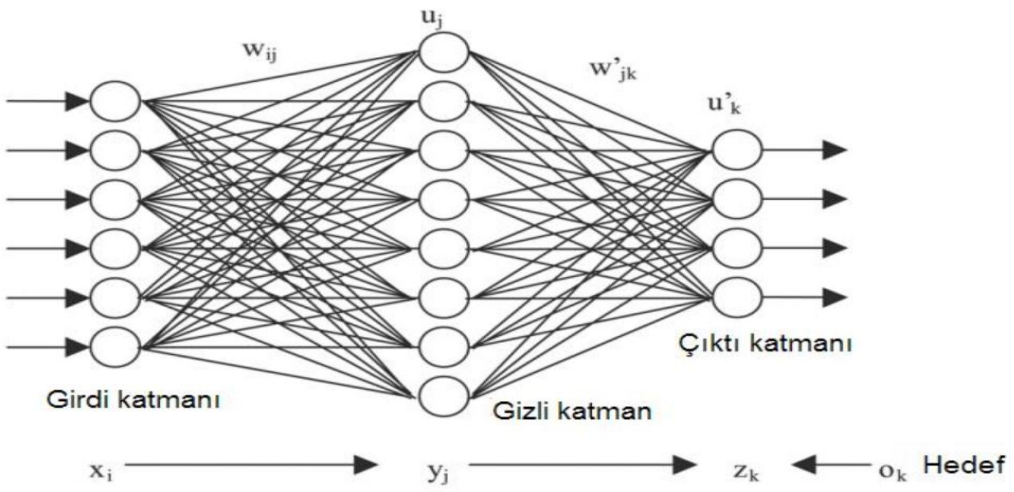
Yapay sinir ağı, karmaşık veri işleme problemlerini çözmek için güçlü bir araç olarak matematiksel olarak biyolojik sinir sistemlerinden esinlenmiştir. Bu modeller, biyolojik nöronların etkileşimlerini ve aktivasyonlarını taklit ederek verilere dikkate değer bir şekilde uyum sağlayabilir ve öğrenebilirler.

Sinir ağı, çeşitli yapay sinir hücrelerinin (nöronlar) birbirleriyle bağlantılarını ve aktivasyonlarını modelleyen bir matematiksel modeldir. Bu nöronlar, bir girdi katmanından başlayarak, ara katmanlar ve bir çıkış katmanı aracılığıyla birbirleriyle bağlantılıdır. Bu bağlantılar, ağırlıklar olarak adlandırılan parametrelerle belirlenir.

Sinir ağları, örüntü tanıma, sınıflandırma, tahmin, kontrol ve optimizasyon gibi birçok alanda kullanılabilir. İnsan vücudundaki sinir sistemine benzer şekilde, sinir ağları da denetimli veya denetimsiz öğrenme süreçleriyle eğitilebilirler.

Sinir ağı uygulama adımları şu şekildedir:

- Veri toplama ve önişleme: Sinir ağının eğitilmesi için gereken veriler toplanır ve ön işleme adımları (veri temizleme, ölçeklendirme vb.) yapılır.
- Sinir ağı tasarımı: Sinir ağı mimarisi, girdi ve çıktı sayısı, katman sayısı ve nöron sayısı gibi parametreler belirlenir.
- Eğitim verilerinin ayrılması: Veri seti, eğitim, doğrulama ve test setleri olarak ayrılır.
- Sinir ağı eğitimi: Veri setleri kullanılarak sinir ağı eğitilir. Bu süreçte, ağın parametreleri (ağırlıklar) optimize edilir.
- Sinir ağı doğrulama ve testi: Eğitilmiş sinir ağı, doğrulama seti ve test seti ile değerlendirilir. Ağın doğruluğu ve performansı değerlendirilir.



Şekil 4.8. Çok katmanlı yapay sinir ağı örneği.

Avantajları:

- Yüksek doğruluk: Sinir ağı, yüksek doğruluk oranlarına sahip bir algoritmadır. Bu, özellikle büyük ve karmaşık veri setleri için geçerlidir.
- Öğrenme kabiliyeti: Sinir ağı, yeni verileri öğrenme kabiliyetine sahiptir ve öğrenme sürecindeki yanıtlara dayanarak gelecekteki tahminler yapabilir.

- Ölçeklenebilirlik: Sinir ağı, büyük veri setleriyle çalışırken ölçeklenebilirlik açısından avantajlıdır. Bu nedenle, büyük ölçekli endüstriyel uygulamalarda sıklıkla kullanılır.
- Mimarilerde esneklik: Sinir ağı, farklı mimariler ve katmanlar kullanarak farklı modeller oluşturma kabiliyetine sahiptir.
- Veri ön işleme gereksinimleri: Sinir ağı, veri ön işleme gereksinimlerinde sınırlıdır. Bazı diğer algoritmaların aksine, özellikle kategorik veriler ve eksik veriler gibi bazı veri kusurlarıyla daha iyi baş edebilir.

Dezavantajları:

- Eğitim gereksinimleri: Sinir ağı, eğitim sürecinde önemli miktarda veri ve zaman gerektirir. Ayrıca, uygun bir model seçmek için birçok parametre ayarlanması gerekir.
- Hesaplama gücü: Sinir ağı, hesaplama gücü gereksinimleri açısından yüksektir ve bu nedenle yüksek miktarda donanım gereksinimleriyle çalışabilir.
- Modelin yorumlanması: Sinir ağı, modelin yorumlanması açısından sınırlamalara sahiptir. Bu nedenle, modelin kararları açıklayabilme kabiliyeti sınırlı olabilir.
- Aşırı uydurma: Sinir ağı, aşırı uydurma problemleriyle karşılaşabilir ve bu nedenle modelin genelleştirilebilirliğini azaltabilir.
- Değişken önem düzeyleri: Sinir ağı, değişken önem düzeylerini tahmin etmek için kullanılamaz. Bu, veri setindeki en önemli değişkenleri belirlemeye yardımcı olabilen diğer algoritmalarla karşılaştırıldığında bir dezavantaj olabilir.

4.4.4. Destek vektör makineleri yöntemi (Support Vektor Machine)

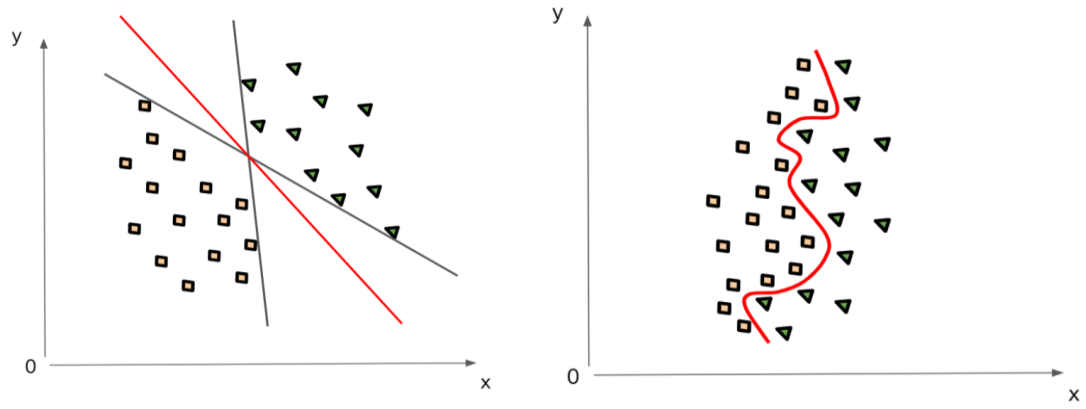
Bugünün makine öğrenimi uygulamalarında, destek vektör makineleri (SVM) tüm iyi bilinen algoritmalar arasında en sağlam ve doğru yöntemlerden birini sunması nedeniyle mutlaka denemeye değer olarak kabul edilmektedir.

Sağlam bir teorik temeli vardır, sadece bir düzine örnek gerektirir ve boyut sayısından bağımsızdır. Ayrıca, SVM için etkili eğitim yöntemleri de hızla geliştirilmektedir.

İki sınıflı bir öğrenme görevinde, SVM'nin amacı eğitim verilerindeki sınıf üyesi arasındaki ayırımı belirlemek için en iyi sınıflandırma fonksiyonunu bulmaktır. "En iyi" sınıflandırma fonksiyonu için ölçüt geometrik olarak gerçekleştirilebilir. Doğrusal olarak ayrılabilir bir veri kümesi için, doğrusal bir sınıflandırma fonksiyonu iki sınıfın ortasından geçen bir ayırma hiper düzlemini f(x) temsil eder. Bu fonksiyon belirlendiğinde, yeni veri örneği xn, sadece f(xn)'nin işaretini test ederek sınıflandırılabilir. xn pozitif sınıfa aitse, f(xn) > 0'dır.

Birçok doğrusal hiper düzlem olduğu için, SVM'nin ek olarak garanti ettiği şey, iki sınıf arasındaki marjın maksimize edilerek en iyi işlevin bulunmasıdır. Basitçe ifade etmek gerekirse, marj, hiper düzlem tarafından tanımlanan iki sınıf arasındaki boşluk veya ayırma olarak tanımlanır.

Geometrik olarak, marj, en yakın veri noktaları arasındaki en kısa mesafeye karşılık gelir. Bu geometrik tanım, marjı nasıl en üst düzeye çıkaracağımızı keşfetmemize olanak tanır, böylece sonsuz sayıda hiper düzlem olmasına rağmen, sadece birkaçı SVM çözümü olarak uygun olur (Şekil 4.9).



Şekil 4.9. Destek vektör makinesi verilerin ayrılması.

SVM'nin neden en genel geçerliliğe sahip olduğunu iddia ettiği neden, en iyi genelleme yeteneği sunmasıdır. Bu, sadece eğitim verilerindeki en iyi sınıflandırma performansını (örneğin doğruluğu) değil, aynı zamanda gelecekteki verilerin doğru sınıflandırılması için de alan bırakması anlamına gelir.

Avantajları:

Yüksek doğruluk: SVM, doğru parametre ayarı ile yüksek doğruluk oranlarına sahip olabilir.

- Veri boyutuna uygun: SVM, yüksek boyutlu veri setleri üzerinde de iyi performans gösterebilir. Bu nedenle, özellikle görüntü işleme veya doğal dil işleme gibi alanlarda sıklıkla kullanılır.
- Esnek çekirdek fonksiyonları: SVM, verilerin farklı şekillerde ayrılabilmesi için çeşitli çekirdek fonksiyonları kullanabilir. Bu nedenle, birçok veri kümesinde farklı çekirdek fonksiyonları kullanarak daha iyi sonuçlar elde edilebilir.
- Aşırı öğrenmeye dirençli: SVM, aşırı öğrenme problemleriyle karşılaşma olasılığını azaltır. Bu, özellikle düşük boyutlu veri setleri için avantajlıdır.

Dezavantajları:

- Ölçeklenebilirlik: SVM, büyük veri setleriyle çalışırken ölçeklenebilirlik açısından dezavantajlıdır. Bu nedenle, büyük ölçekli endüstriyel uygulamalarda bazen kullanılmaz.
- Eğitim zamanı: SVM, büyük veri setleri üzerinde eğitim yaparken zaman alabilir. Bu, özellikle doğru parametrelerin seçilmesi için daha uzun bir süre gerektiğinde geçerlidir.
- Parametre seçimi: SVM'nin doğru çalışabilmesi için doğru parametrelerin seçilmesi gerekir. Bu, süreçte doğru parametreleri bulmak için daha fazla çaba gerekebilir.
- Modelin yorum yapılabilirliği: SVM, modelin yorum yapılabilirliği açısından sınırlamalara sahiptir. Modelin kararları açıklayabilme kabiliyeti sınırlı olabilir.

4.4.5. Gradyanı artırılan karar ağaçları yöntemi (XGBoost)

Gradyanı Artırılan Karar Ağaçları, 2014 yılında Tianqi Chen tarafından geliştirilmiş bir makine öğrenimi algoritmasıdır. GBM'nin (Gradyan Artırımlı Makine) dezavantajlarını aşarak daha hızlı ve daha az bellek kullanarak daha iyi sonuçlar verir.

Gradyan artırımlı karar ağaçları ile birleştirilmiş bir algoritmadır ve gradyan yükseltme altyapısında geliştirilmiştir. Aşırı öğrenmeyi önleme, eksik veri ile çalışabilme, çapraz doğrulama ve ağaç budama gibi özelliklere sahip bir altyapısı bulunmaktadır. Sınıflandırma ve regresyon görevlerinde kullanılır. Bu yöntemin çalışma mantığı, veriyi tek tek değerlendirmek yerine onları küçük parçalara bölmek

üzerine kuruludur ve daha iyi tahmin sonuçları elde etmesi beklenir. Modelin performansı için parametre ayarlarına dikkat ederek en uygun değerleri seçmek önemlidir. Modelde ağaç derinliği, gamma (aşırı uyumu önlemek için), öğrenme oranı, alt örnek (eğitim için rastgele seçilen alt veri kümesi) gibi parametreler kullanılır.

Gradyan artırma yöntemi, regresyon ağaçlarını iteratif bir şekilde modele ekleyerek ve türev alınabilir bir kayıp fonksiyonunu optimize ederek regresyon ağaçları topluluğunu geliştiren bir yöntemdir (Natekin & Knoll, 2013).

Bu yöntemin en son uygulamalarından biri Gradyanı Artırılan Karar Ağaçları algoritmasıdır. Gradyanı Artırılan Karar Ağaçları Algoritması, Regularize Edilmiş Boosting ve Stokastik Boosting tekniklerini uygulayarak işlem süresini azaltır ve bellek kaynaklarının optimum kullanımını sağlar (Chen & Guestrin, 2016).

Ayrıca, düzenleme parametrelerini ayarlayabilme özelliği sayesinde aşırı uyumlanmayı önler ve modeldeki ağaçların karmaşıklığını kontrol ederek yüksek doğruluk oranlarına ulaşır. XGBoost algoritması, Gradyan Artırılmış Karar Ağaçlarının optimize edilmiş bir versiyonudur ve yüksek tahmin gücü, aşırı öğrenmeyi önleme ve eksik verileri etkili bir şekilde yönetme özellikleri ile bilinir.

XGBoost Algoritması Gradyanı Artırılan Karar Ağaçlarının çeşitli düzenlemeler sonucu optimize edilmiş halidir. Yüksek tahmin gücü elde edebilmesi, aşırı öğrenmeyi önlemesi, boş verileri yönetmesi özellikleri ile iyi bir performans gösterir.

Gradyanı Artırılan Karar Ağaçları yöntemini kullanarak, birden fazla karar ağacı oluşturarak ve bu ağaçların sonuçlarını birleştirerek çalışır. Bu yöntem, birçok makine öğrenimi yarışmasında ve gerçek dünya problemlerinde üstün performans göstermiştir.

Gradyanı Artırılan Karar Ağaçları uygulama adımları şu şekildedir:

- Veri toplama ve önişleme: Veriler toplanır ve ön işleme adımları (veri temizleme, ölçeklendirme vb.) yapılır.
- Veri kümesinin bölünmesi: Veri kümesi, eğitim ve test setleri olarak ayrılır.
- Gradyanı Artırılan Karar Ağaçları modelinin oluşturulması: Model oluşturma aşamasında, ağaçların sayısı, derinliği, düğüm ağırlığı ve öğrenme oranı gibi parametreler belirlenir.
- Model eğitimi: Eğitim seti kullanılarak model eğitilir.

- Model değerlendirme: Eğitilmiş model, test seti kullanılarak değerlendirilir.
- Model iyileştirme: Modelin hata oranını azaltmak için, parametreler ayarlanarak ve aşırı uyum önlemek için düzenlileştirme yöntemleri kullanarak model iyileştirilebilir.

Gradyanı Artırılan Karar Ağaçlarının avantajları şunlardır:

- Yüksek doğruluk: Birçok veri kümesinde yüksek doğruluk sağlar.
- Hız: Hızlı bir algoritmadır ve büyük veri kümelerinde bile hızlı sonuçlar verir.
- Performans: Performans ölçümleri gibi metrikleri iyileştirmek için özelleştirilebilir.
- Uygunluk: Çeşitli veri tipleriyle uyumludur.

Gradyanı Artırılan Karar Ağaçlarının dezavantajları şunlardır:

- Parametre ayarlaması: Gradyanı Artırılan Karar Ağaçlarının yüksek performans göstermesi için parametrelerin doğru ayarlaması gerekir.
- Aşırı öğrenme riski: Gradyanı Artırılan Karar Ağaçlarının, aşırı uyum riski taşıyan bir algoritmadır. Bu nedenle, düzenlileştirme yöntemleri kullanılmalıdır.
- Anlaşılması zor: Gradyanı Artırılan Karar Ağaçlarının iç işleyişi oldukça karmaşıktır ve anlaşılması zor olabilir.

4.4.6. Hafif gradyanı artırılmış makineleri yöntemi (LightGBM)

Hafif Gradyanı Artırılmış Makineleri Yöntemi (LightGBM), bir makine öğrenimi algoritmasıdır ve ağaç tabanlı bir yöntemdir. LightGBM, büyük veri kümesi üzerinde hızlı, yüksek performanslı ve yüksek doğruluklu tahminler yapma yeteneği ile bilinir. LightGBM, Microsoft tarafından geliştirilmiştir ve 2016 yılında ilk kez duyurulmuştur.

LightGBM, geleneksel Gradyan Artırımı yöntemine dayanır, ancak bazı önemli yenilikler içerir. En dikkat çekici özellikleri arasında düşük bellek kullanımı, yüksek hızlı eğitim süreçleri ve büyük veri kümesinde yüksek performans bulunmaktadır. LightGBM, büyük veri kümesinde hızlı tahminler yapma yeteneği ile tanınır ve çeşitli endüstriyel uygulamalarda kullanılır.

LightGBM'in avantajları şunlardır:

- Hızlı eğitim süreçleri: LightGBM, büyük veri kümesinde hızlı bir şekilde eğitim yapabilen bir yöntemdir. Düşük bellek kullanımı ve özel hesaplama teknikleri sayesinde, diğer geleneksel Gradyan Artırımı yöntemlerine göre daha hızlı eğitim süreçleri sunar.
- Yüksek performans: LightGBM, büyük veri kümesinde yüksek performans sağlar. Yüksek doğruluklu tahminler yapma yeteneği ile bilinir ve birçok endüstriyel uygulamada başarıyla kullanılır.
- Ölçeklenebilirlik: LightGBM, büyük veri kümesini kolayca işleyebilen bir ölçeklenebilirlik sunar. Büyük veri kümesinde yüksek performanslı tahminler yapabilme yeteneği, veri büyüklüğüne bağlı olarak ölçeklendirilebilir.
- Özelleştirilme: LightGBM, kullanıcıların modeli özelleştirmelerine izin veren birçok hiper parametre sunar. Bu, kullanıcıların modeli kendi veri kümesine ve gereksinimlerine uygun hale getirmelerine olanak tanır.

LightGBM'in dezavantajları şunlardır:

- Veri dengesizliği: LightGBM, veri dengesizliği ile karşılaşıldığında doğruluk oranını düşürebilir. Dengesiz veri kümesi içeren uygulamalarda dikkatli olunmalı ve dengeleme teknikleri kullanılmalıdır.
- Aşırı uyuma LightGBM, aşırı uyuma eğilimli olabilir, özellikle model aşırı karmaşıklaştığında elde edilen modelin test verilerine genelleme yapma yeteneğini azaltabilir. Uygun düzenleme teknikleri kullanılarak aşırı uyumanın önüne geçilmelidir.
- İnterpretasyon zorluğu: LightGBM, karmaşık bir model yapısına sahiptir ve modelin içsel yapısal yorumlanması zor olabilir. Modelin içindeki ağaç yapısı ve etkileşimlerini anlamak bazen karmaşık olabilir, bu nedenle modelin yorumlanması ve açıklanması zorlu olabilir.
- Donanım gereksinimleri: LightGBM, hızlı eğitim süreçleri ve yüksek performans için yüksek işlem gücü gerektirebilir. Bu, düşük donanım kaynaklarına sahip sistemlerde LightGBM'in performansının düşebileceği anlamına gelebilir.

4.5. Gizlilik Koruma Yöntemleri

Çalışmamda literatürde var olan iki yönteme yer verilmiştir. Diferansiyel temelli yaklaşımlardan Gauss ve Laplace Mekanizmaları ve pertürbasyon temelli rotasyona dayalı yöntemler aşağıda anlatılmıştır. Çalışmada literatürde var olan iki metot birleştirilmiş mahremiyet koruma seviyesi artırılmıştır.

4.5.1. Diferansiyel temelli mahremiyet yöntemleri

Diferansiyel temelli mahremiyet yöntemleri, veri analitiğinde, istatistiksel verilerin paylaşılmasını veya yayınlanmasını güvenli hale getirmek için kullanılan tekniklerdir. Bu yöntemler, hassas verilerin gizliliğini koruyarak, veri analizine olanak tanır.

Diferansiyel mahremiyet, veri analitiğinde kullanılan ve Dwork ve ark. tarafından önerilen bir mahremiyet koruma modelidir. Bu model, verilere giriş, model ve çıkış seviyelerinde gürültü ekleyerek mahremiyet sağlar.

K-anonimlik, l-çeşitlilik, t-yakınlık ve δ-mevcudiyet gibi geleneksel yaklaşımların en önemli zafiyeti, arka plan bilgisi saldırılarına karşı tam bir koruma sağlayamamalarıdır. Bununla birlikte, diferansiyel mahremiyet modeli bu sorunu ele alır ve yüksek bir mahremiyet sunar.

Diferansiyel mahremiyet, veri analitiğinde hassas bilgileri koruyabilmek için verilere gürültü ekleyerek mahremiyet sağlar. Hedef, veri analiz sürecinde gürültü ekleyerek orijinal verilerle aynı istatistiksel sonuçları elde etmek ve hassas bilgilerin ifşa edilmesini önlemektir.

Giriş seviyesinde, diferansiyel mahremiyet verilere gürültü ekleyerek veri mahremiyetini korur. Model seviyesinde, modelin yapısını veya parametrelerini koruyarak mahremiyeti sağlar. Çıkış seviyesinde ise, diferansiyel mahremiyet, modelin çıkışını koruyarak mahremiyeti sağlar.

Diferansiyel mahremiyet, veri analizinde güçlü bir mahremiyet koruma yöntemi olarak kabul edilir ve yaygın olarak kullanılır. Bununla birlikte, her yöntemin kendi avantajları ve dezavantajları vardır ve doğru yöntemi seçmek, veri mahremiyeti ve analiz doğruluğunu sağlamak için dikkatli bir değerlendirme gerektirir. Ayrıca, diferansiyel mahremiyetin uygulandığı senaryo ve veri tabanlarına bağlı olarak farklı parametre ayarları ve teknikler kullanılabilir.

Anonimleştirme, veri kümesi veya sorgu sonucunu orijinal veriye mümkün olduğu kadar yakın ancak anonim hale getiren çeşitli dönüşümler uygulayarak sağlanan bir mahremiyet koruma yöntemidir. Genelleştirme, baskılama, gürültü ekleme gibi yaklaşımlar, veri kümesi veya sorgu sonucunu anonimleştirme sürecinde kullanılan yapıları temsil eder. Bu yapılar sayesinde anonimleştirilen veri kümesi veya sorgu sonucu üzerinde geliştirilen bir model veya analizin doğruluğu, orijinal veri üzerinde geliştirilecek bir model veya analizin doğruluğundan daha düşük seviyede olacaktır.

Nitekim, bir verinin veya bir sorguya diferansiyel mahremiyetin uygulanması ile elde edilecek yöntemin başarısı orijinal veri üzerinde geliştirilecek yöntemin başarından daha düşük olması beklenir.

Bu süreçte, veri üzerinde analiz arasındaki farkı minimize etmek hedeflenir. f sorgusunun sonuçlarındaki maksimum fark hassasiyet, Δf, olarak adlandırılır. Diferansiyel mahremiyet, D veri kümesine uygulanan rasgele bir M mekanizması ile gerçekleştirilir. ε-Diferansiyel Mahremiyet, bir M mekanizmasının herhangi bir S çıktı kümesi için herhangi D1 ve D2 komşu veri kümeleri için aşağıdaki şartı sağlaması durumunda mümkün olur.

Diferansiyel Mahremiyet; bir M mekanizması, her çıktı kümesi S için ve herhangi iki komşu veri kümesi D1 ve D2 için aşağıdaki şartı sağlarsa ε-diferansiyel mahremiyete ulaşır (3.1).

$$\Pr[\mathrm{M}(\mathrm{D1}) \in \mathrm{S}] \leq \exp(e)\ \Pr[\mathrm{M}(\mathrm{D2}) \in \mathrm{S}] \tag{4.1}$$

(ε, δ)-Diferansiyel Mahremiyet; bir M mekanizması, her çıktı kümesi S için ve herhangi iki komşu veri kümesi D1 ve D2 için aşağıdaki şartı sağlarsa (ε, δ)-diferansiyel mahremiyeti sağlar. Bir analist, veri tabanına sorgu ("Toplam (SUM)", "Sayma (Count)", "Ortalama (Mean)") göndermek istediğinde 3.2'deki denklem ile hesaplanır.

$$\Pr[\mathrm{M}(\mathrm{D1}) \in \mathrm{S}] \leq \exp(e)\ \Pr[\mathrm{M}(\mathrm{D2}) \in \mathrm{S}] + \delta \tag{4.2}$$

Sorgu veri tabanında gerçekleştirilir ve elde edilen gerçek cevap diferansiyel mahremiyet mekanizmasına gönderilir. Mekanizma gerçek cevaba gürültü ekler ve gürültülü cevap analiste gönderilir. Bu sayede analist, gerçek veriye erişmek yerine yaklaşık cevaplar elde eder.

Şekil 4.10. Farklılaştılmış gizlilik süreci/veri mahremiyeti.

Farklılaşmış Gizlilik Süreci: U veri alanını temsil etsin ve |U| bu alanın boyutunu ifade etsin. Diyelim ki, bir tablo veri kümesi d-boyutlu uzayda r kaydı içermektedir. D1 ve D2, yalnızca bir kayıtta farklılaşan komşu veri kümeleridir ve komşu veri kümeleri olarak adlandırılır. Bu f, veri kümesinden çeşitli sorgu sonuçlarını döndüren bir sorgu işlevidir ve sorgu işlevlerinin grubu F tarafından temsil edilir (Şekil 4.10.).

Diferansiyel mahremiyetin amacı, D1 ve D2 veri kümeleri arasındaki f sorgularının sonuçlarıdır.

Bu tanımlamalara göre, ε-Diferansiyel Gizlilik temel diferansiyel gizlilik olarak tanımlanırken, (ε, δ)-Diferansiyel Gizlilik $\delta > 0$ için yaklaşık diferansiyel gizlilik olarak tanımlanır.

Bir sorgunun hassasiyeti, mekanizmanın uygulayacağı gürültü miktarını belirler. Bir veri kümesine f sorgusu uygulandığında, hassasiyet sonucuna eklenecek gürültü miktarını belirler.

Literatürde diferansiyel gizlilik kapsamında global ve lokal olmak üzere iki farklı hassasiyet türü kullanılmaktadır. Global hassasiyet, komşu veri kümelerine uygulanan sorgular arasındaki maksimum farkı verir ve eklenecek gürültü miktarını belirler.

Tüm veri kümelerine uygulanan sorgular arasındaki maksimum farkı dikkate aldığı için veri kümesine bağımlı olmayan bir yapıdır ve sorgu bağımlı bir yapıdır. Sorgu

hassasiyeti düşük olduğunda kullanılması tercih edilir. Örneğin, "Sayım" sorgusunun hassasiyeti 1 olarak kabul edilir.

Lokal hassasiyet, bazı sorgular için global hassasiyetin doğru sonuçların üretilmesini garanti edemediği durumlarda kullanılacak veri kümesini sınırlamayı gerektirir. Bu durumda, belirli bir ön tanımlı eşik değeri sağlayan komşu veri kümelerinin hassasiyetinin ölçülmesi gerekmektedir. Bu amaçla geliştirilen lokal hassasiyet, "Medyan" gibi sorgularda tercih edilir. "Sayım" ve "Aralık" sorguları için lokal hassasiyet, global hassasiyetle aynıdır."

Literatürde diferansiyel mahremiyeti sağlamak için 3 yöntem vardır. Çalışmamızda bunlardan Gauss Mekanizmasına yer verilmiştir.

4.5.1.1. Gauss mekanizması

Diferansiyel mahremiyet için Gauss mekanizması kullanılarak elde edilen (ε, δ)-diferansiyel mahremiyet tanımı aşağıda sunulmuştur.

Rasgeleleştirilmiş bir fonksiyon K, yalnızca bir kaydın farklı olduğu D1 ve D2 komşu veri kümeleri için (ε, δ)-diferansiyel mahremiyeti sağlar, böylece her $S \subseteq Range(K)$ için kullanılabilir (3.3).

$$Pr[K(D1) \in S] \leq \exp(e) \text{ x } Pr[K(D2) \in S] + \delta \tag{4.3}$$

Gauss mekanizması, hassas verilerin ortalamasını veya toplamını koruyarak, gürültü ekleyerek çalışır. Verilerin ortalamasına veya toplamına gürültü ekleyerek, tahminlerin doğruluğunu korurken, tam veriyi ifşa etmez. Gauss mekanizması, normal dağılıma sahip gürültü ekleyerek çalışır ve gürültü miktarı, mahremiyet düzeyini ayarlamak için kullanılabilir.

4.5.1.2. Laplace mekanizması

Laplace mekanizması, hassas verilere Laplace dağılımından gürültü ekleyerek çalışır. Laplace dağılımı, asimetrik bir olasılık dağılımıdır ve gürültü miktarı, verinin hassasiyetine bağlı olarak ayarlanabilir.

Laplace mekanizması, verilere gürültü ekleyerek, tahminlerin doğruluğunu korurken, veri mahremiyetini sağlar.

Laplace mekanizması kullanarak ε-farklısal gizlilik için bir tanım aşağıda sunulmuştur.

Bir rastgeleleştirme fonksiyonu K, komşu veri kümeleri D1 ve D2 için $S \subseteq Range(K)$ olacak şekilde aşağıdaki koşulu sağladığında ε-farklısal gizlilik sağlar (3.4).

$$\mathrm{E}[K(D1) \in S] \leq \exp(e) \text{ x } E[K(D2) \in S] \tag{4.4}$$

Bir sorgu fonksiyonu f için, bu fonksiyonun hassasiyeti komşu veri kümeleri D1 ve D2 için aşağıdaki gibi tanımlanır (3.5).

$$f\colon D \to Rd, \text{ f sorgusunun L1-hassasiyeti } \Delta f = \max\|f(D1) - f(D2)\|1 \tag{4.5}$$

Sonuç olarak, ε-farklısal gizlilik için ilgili mekanizma aşağıdaki gibi oluşturulur, burada ε gizlilik parametresini ve $Lap(0, \Delta f/\varepsilon)$ Laplace dağılımını temsil eder, 0 ortalama ve $\Delta f/\varepsilon$ ölçekleme (standart sapma: $\sqrt{2}\Delta f/\varepsilon$):

$$K(D) = f(D) + Lap(0, \Delta f/\varepsilon) \tag{4.6}$$

ε değerinin gizlilik koruması ile ters orantılı ve sorgu fonksiyonunun hassasiyeti ile doğru orantılı olduğu unutulmamalıdır.

4.5.1.3. Eksponansiyel mekanizması

Eksponansiyel mekanizması, verilere gürültü eklerken, verilerin hassasiyetini korumak için çalışır. Eksponansiyel mekanizması, gürültüyü, hassasiyet ve mahremiyet düzeyini kontrol etmek için kullanılan bir parametreye dayanarak ayarlar. Eksponansiyel mekanizması, verilere gürültü ekleyerek, mahremiyeti koruyarak, veri analizini mümkün kılar.

Sayısal olmayan sorgular için diferansiyel mahremiyeti sağlamak için kullanılan bir mekanizmadır. Bu mekanizma, veri kümesi D için elde edilen çıktı $\phi \in \Phi$'yi değerlendirmek için bir skor fonksiyonu olan q(D, ϕ) kullanmaktadır. Bu skor fonksiyonu farklı uygulamalara göre değişebilir.

$\varepsilon xq(D, \phi)$ ifadesi q skor fonksiyonunun hassasiyetini temsil etsin. Bir K mekanizması, aşağıdaki senaryo için ε-diferansiyel mahremiyeti sağlar (3.7).

$$K(D) = \{\exp(\varepsilon xq(D, \phi)/2) \text{ x } \Delta q \tag{4.7}$$

Eğer ϕ' için orijinal ve alternatif sonuçların oranı exp(ε)'den küçük veya buna eşitse ile doğru orantılı bir ihtimal için ϕ'ı geri gönder.

4.5.2. Pertürbasyon temelli mahremiyet yöntemleri

Çalışmada ele alınan konuları anlamak için öncelikle temel kavramlardan olan veri mahremiyetinde rotasyona dayalı iki boyutlu veri pertürbasyonuna yer veriyoruz. Veri bozma yaklaşımına dayalı yöntemler, olasılık dağılımı kategorisi ve sabit veri bozma kategorisi olmak üzere iki ana kategoriye ayrılır.

Olasılık dağılımı kategorisinde, güvenlik kontrol yöntemi orijinal veri tabanını aynı dağılımdan başka bir örnek veya dağılım kendisi ile değiştirir. Diğer yandan, literatürde ele alınan sabit veri bozma yöntemleri genellikle sayısal veri veya kategorik veri için özel olarak geliştirilmiştir. Bu yöntemler genellikle ikincil kullanım için dönüştürülmüş bir veri tabanının oluşturulmasını gerektirir ve tek bir nitelik için basit bir yöntemden çoklu nitelik yöntemlerine kadar gelişmiştir. Her durumda, bu yöntemler, ortalama değeri 0 olan bir gürültü terimini ekleyerek ve böylece ortalamanın tahmininde yanlılık yaratmadan veriyi bozmayı içerir. Bu çalışmada, sabit veri bozma yöntemlerine odaklanıyoruz.

En basit haliyle, sabit veri bozma yöntemleri güvenilir bir niteliği X'e bir gürültü terimi ε ekleyerek pertürbe edilmiş bir nitelik Y elde etmeyi içerir. Bu yöntem çoklu nitelikli veri tabanları için kullanıldığında, veritabanındaki her nitelik diğerlerinden bağımsız olarak pertürbe edilir. Genel olarak, bu yöntem Y = X + ε olarak tanımlanır, burada ε, bilinen bir varyansa sahip (örneğin, Uniform, Normal gibi) bir olasılık dağılımından çekilmiştir. Bu yöntemlere Eklenti Veri Pertürbasyonu (ADP) denir. ADP yöntemlerinin yanı sıra, Bölgesel Çarpanlı Veri Pertürbasyonu (MDP) bireylerin gizliliğini korurken toplu istatistikler sağlamak için kullanılabilir. Bu tür bir yöntemde, tek bir güvenilir nitelik X için pertürbe edilmiş nitelik Y şu şekildedir: Y = X ε, burada ε 'nin ortalaması 1.0 ve belirtilen bir varyansı vardır. ε 'nin ortalaması 1.0 olduğu için ortalamanın tahmininde yanlılık yaratmaz.

MDP yöntemi birden fazla güvenilir niteliğin bozulmasında kullanıldığında, her nitelik diğer niteliklerden bağımsız olarak pertürbe edilmelidir.

Gürültü Ekleme R = Orijinal veri setinin iki değişkenine ortalamaları kadar gürültü eklenir (3.9).

$$Y = X + \varepsilon \tag{4.8}$$

Orijinal veri seti(x, y) = {3, 2, 2, 0, 5, 9, 7, -2, 5 ,1},{4, 5, 4, 2, 7, 8, 7, 1, 6 ,3}

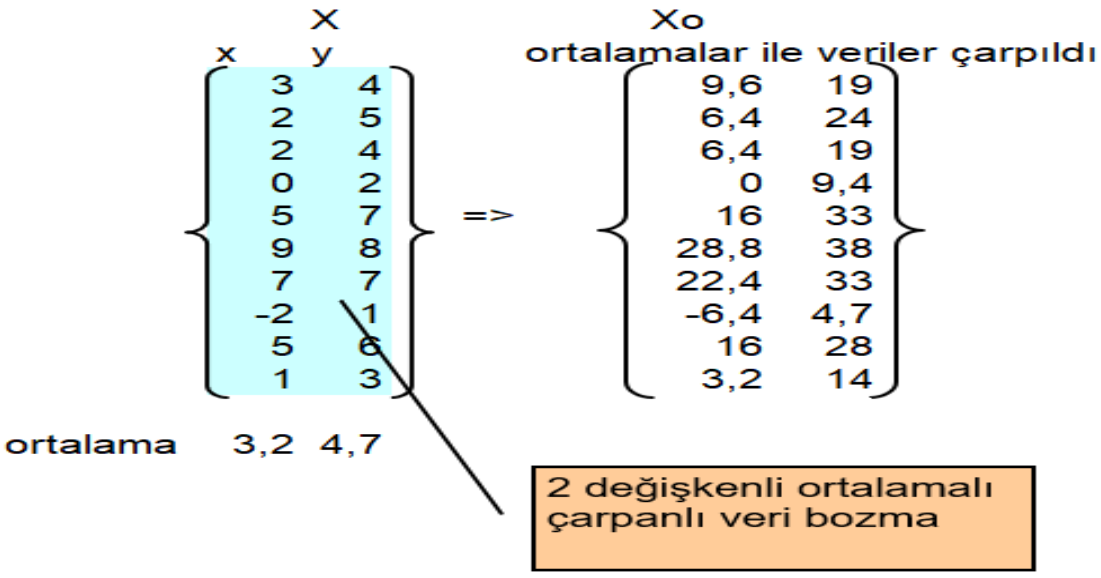
Gürültü eklenmiş veri seti(x, y) = {-0.2,-1.2,-1.2, -3.2, 1.8, 5.8, 3.8, -5.2, 1.8 ,-2.2},{-0.7, 0.3, -0.7, -2.7, 2.3, 3.3, 2.3, -3.7, 1.3 ,-1.7}

Şekil 4.11. Çarpanlı veri pertürbasyonu

Çarpanlı R = Orijinal veri setinin iki değişkenine ortalamaları kadar ε ile çarpıldı Y = X ε

Orijinal veri seti(x, y) = {3, 2, 2, 0, 5, 9, 7, -2, 5 ,1},{4, 5, 4, 2, 7, 8, 7, 1, 6 ,3}

Ortalamalar ile çarpılmış veri seti(x, y) = {9.6, 6.4, 6.4, 0, 16, 28.8, 22.4, -6.4, 16 ,-3.2},{19, 24, 19, 9.4, 3.3, 38, 33, 4.7, 28 ,14}



Şekil 4.12. Bölgesel çarpanlı veri pertürbasyonu.

4.5.3. Döndürme temelli veri bozma yöntemi (rotation Data perturbation)

Döndürme Temelli Veri Bozma Yöntemi, RDP olarak kısaltılan, önceki yöntemlerimizden farklı bir şekilde çalışır. Bu durumda, gürültü terimi bir açı θ'dir. Saat yönünde ölçülen döndürme açısı θ, gizli niteliklerin gözlemlerine uygulanan bir dönüşümdür. Önceki yöntemlerin aksine, RDP bazı gizli niteliklere birden fazla kez uygulanabilir. Örneğin, bir dönüşüm uygulandığında, iki koordinatın değerleri etkilenir. İki boyutlu ayrık uzayda, X ve Y koordinatları etkilenir.

Üç boyutlu ayrık uzayda veya daha yüksek boyutlarda, iki değişken etkilenir ve diğerleri herhangi bir değişiklik olmadan kalır. Gizliliği korumak için tüm gizli niteliklerin bozulması için bir veya daha fazla dönüşüm dönüşümünün uygulanması gereklidir. RDP algoritmasının taslağı aşağıdaki gibi verilmiştir:

Döndürme matrisi R = Orijinal veri setinin bir değişkeninin 180 derece döndürülmüş örneği

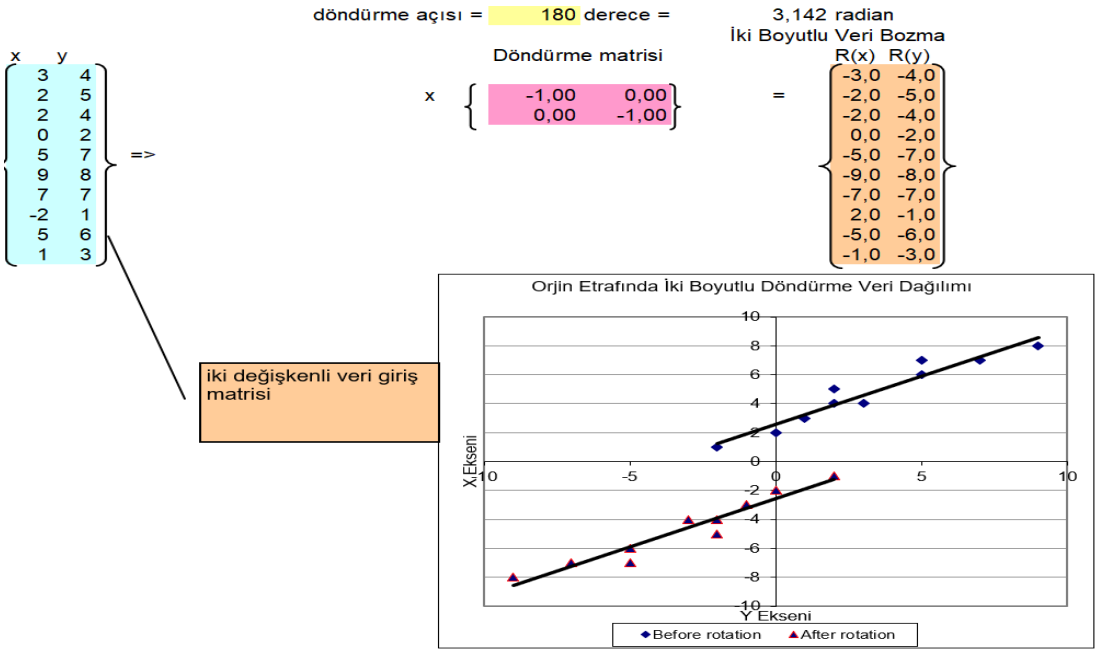
Orijinal veri seti(x) = 19, 26, 34, 34, 41, 54, 54, 54, 26.2, 20.4, 3, 46.5

Döndürme matris (R) = 46.5, 3, 20.4, 26.2,54 ,54, 54, 41, 34, 34, 26, 19

Döndürme matrisi R = Orijinal veri setinin iki değişkeninin 180 derece döndürülmüş örneği

Orijinal veri seti(x, y) = {3, 2, 2, 0, 5, 9, 7, -2, 5 ,1},{4, 5, 4, 2, 7, 8, 7, 1, 6 ,3}

Döndürülmüş veri seti(x, y) = {-3,-2,-2, 0, -5, -9, -7, 2, -5 ,-1},{-4, -5, -4,-2, -7, -8, -7, -1, -6 ,-3} (Şekil 4.13.)



Şekil 4.13. Döndürme temelli iki boyutlu veri pertürbasyon yöntemi.

4.6. Performans Metrikleri ve Yöntemleri

Performans metrikleri ve yöntemleri, veri madenciliğinde sınıflandırma algoritmalarının performansını değerlendirmek ve karşılaştırmak için kullanılır. Bu metrikler, modelin doğruluğunu, güvenilirliğini ve tahmin yeteneğini değerlendirmek için kullanılır ve farklı modellerin performansını karşılaştırmak için objektif bir temel sağlar.

Performans metrikleri ve yöntemleri aşağıdaki gibi işlevlere sahiptir**:**

4.6.1. Performans değerlendirmesi

Sınıflandırma modellerinin performansını değerlendirmek için kullanılır. Bir modelin doğruluk, hassasiyet, özgüllük ve F1-skoru gibi performans metrikleri ile ne kadar iyi çalıştığını değerlendirebiliriz.

4.6.2. Model karşılaştırması

Farklı sınıflandırma modellerinin performansını karşılaştırmak için kullanılır. Birden fazla modelin farklı performans metriklerini karşılaştırarak hangi modelin daha iyi çalıştığını belirleyebiliriz.

4.6.3. Model seçimi

Farklı sınıflandırma modelleri arasında en uygun olanını seçmek için kullanılır. Performans metrikleri ve yöntemleri, modellerin hangi özelliklerinin daha iyi olduğunu değerlendirmemize ve hangi modelin veri setine daha uygun olduğunu belirlememize yardımcı olur.

4.6.4. Model ayarlama

Model parametrelerini ve hiper parametrelerini ayarlamak için kullanılır. Performans metrikleri, modelin parametrelerini ve hiper parametrelerini ayarlayarak daha iyi bir performans elde etmek için kullanılabilir**.**

4.6.5. Uygulama geliştirme

Sınıflandırma modellerini geliştirmek ve optimize etmek için kullanılır. Performans metrikleri, modelin performansını izlememize ve modeli sürekli olarak geliştirmemize yardımcı olur.

Bazı performans metrik ve yöntemleri şunlardır:

4.6.6. Sınıflandırma doğruluğu (classification accuracy)

Bir sınıflandırma modelinin doğru sınıflandırma oranını ölçen bir performans metriğidir. Sınıflandırma doğruluğu, toplam doğru tahmin edilen örneklerin, toplam örnek sayısına oranını temsil eder. Yani, doğru sınıflandırılan örneklerin yüzde cinsinden bir değerdir.

Sınıflandırma doğruluğu, aşağıdaki formül ile hesaplanabilir:

Sınıflandırma Doğruluğu = (Doğru Tahmin Edilen Örnek Sayısı) / (Toplam Örnek Sayısı)

Örneğin, bir sınıflandırma modeli 1000 örnek üzerinde çalıştırıldığında, 800 örneği doğru sınıflandırdıysa, sınıflandırma doğruluğu şu şekilde hesaplanabilir:

Sınıflandırma Doğruluğu = 800 / 1000 = 0.8 veya %80

Sınıflandırma doğruluğu, bir modelin doğru sınıflandırma yeteneğini ölçer ve yüksek bir doğruluk oranı, modelin iyi performans gösterdiğini gösterir. Ancak, sınıflandırma doğruluğu tek başına yeterli bir performans metriği değildir ve diğer metriklerle birlikte değerlendirilmelidir. Özellikle dengesiz sınıf dağılımlarına sahip veri

setlerinde, sınıflandırma doğruluğu yanıltıcı olabilir ve diğer metrikler daha ayrıntılı bir performans değerlendirmesi sağlayabilir.

4.6.7. Hassasiyet (precision)

Sınıflandırma algoritmalarındaki bir performans metriği olup, pozitif olarak tahmin edilen örneklerin gerçekten pozitif olma oranını ölçer. Hassasiyet, bir modelin yanlış pozitif tahminlerinin sayısını minimize etme yeteneğini değerlendirir.

Hassasiyet, aşağıdaki formülle hesaplanır:

Hassasiyet, pozitif olarak tahmin edilen veri örneklerinin gerçek pozitif veri örneklerine oranını temsil eder.

$$\text{Hassasiyet} = (TP) / (TP + FP) \tag{4.9}$$

TP (True Positive): Gerçek pozitif tahmin sayısı, yani pozitif olarak tahmin edilen ve gerçekten pozitif olan örneklerin sayısı.

FP (False Positive): Yanlış pozitif tahmin sayısı, yani negatif olarak tahmin edilen ancak gerçekte pozitif olmayan örneklerin sayısı.

Hassasiyet, modelin pozitif tahminlerinin ne kadar doğru olduğunu ölçer. Yüksek hassasiyet değeri, modelin pozitif tahminlerinin doğru ve güvenilir olduğunu gösterir.

4.6.8. Duyarlılık (Recall veya Sensitivity)

Sınıflandırma algoritmalarındaki bir performans metriği olup, gerçek pozitif örneklerin doğru bir şekilde tespit edilme oranını ölçer. Duyarlılık, aynı zamanda "duyarlılık", "hassaslık" veya "gerçek pozitif oranı" olarak da adlandırılır.

Duyarlılık, aşağıdaki formülle hesaplanır:

$$\text{Duyarlılık} = (TP) / (TP + FN) \tag{4.10}$$

Burada:

TP (True Positive): Gerçek pozitif tahmin sayısı, yani pozitif olarak tahmin edilen ve gerçekten pozitif olan örneklerin sayısı.

FN (False Negative): Yanlış negatif tahmin sayısı, yani pozitif olarak tahmin edilmeyen ancak gerçekte pozitif olan örneklerin sayısı.

Duyarlılık, modelin pozitif sınıfı kaçırma durumlarına karşı ne kadar duyarlı olduğunu ölçer. Yüksek duyarlılık değeri, modelin gerçek pozitifleri tespit etme yeteneğinin yüksek olduğunu gösterir.

4.6.9. F1 skoru (F1 score)

Sınıflandırma algoritmalarındaki bir performans metriği olup, kesinlik ve duyarlılık metriklerinin harmonik ortalamasını temsil eder. F1 Skoru, hem sınıflandırma modelinin doğru pozitif tahminleri (kesinlik) hem de pozitif sınıfı kaçırmama durumlarına dikkate alarak modelin performansını değerlendirir. F1 Skoru, bir modelin hem kesinlik hem de duyarlılık açısından dengeli bir performans sergileyip sergilemediğini ölçer.

F1 Skoru, aşağıdaki formülle hesaplanır:

$$\text{F1 Skoru} = 2 * (\text{Kesinlik} * \text{Duyarlılık}) / (\text{Kesinlik} + \text{Duyarlılık}) \quad (4.11)$$

F1 Skoru, 0 ile 1 arasında değer alır, 1 en iyi performansı, 0 en kötü performansı temsil eder. F1 Skoru ne kadar yüksekse, modelin hem kesinlik hem de duyarlılık açısından daha iyi bir performans sergilediği kabul edilir.

4.6.10. Karışıklık matrisi (confusion matrix)

Sınıflandırma algoritmalarının performansını değerlendirmek için kullanılan bir metriktir. Karışıklık matrisi, gerçek sınıf etiketleri ve modelin tahmin ettiği sınıf etiketleri arasındaki ilişkiyi görsel olarak sunar. Karışıklık matrisi, sınıflandırma problemlerinde modelin doğruluk, hassasiyet, duyarlılık, özgünlük gibi performans metriklerini hesaplamak için kullanılır.

Karışıklık matrisi, genellikle 2x2 boyutunda bir matris olarak temsil edilir. 2 sınıflı sınıflandırma problemlerinde dört temel kavramı içerir:

True Positive (TP) (Gerçek Pozitif): Modelin doğru bir şekilde pozitif olarak tahmin ettiği veri örneklerinin sayısını temsil eder. Yani, gerçek sınıfı pozitif olan veri örnekleri doğru bir şekilde pozitif olarak tahmin edilmiştir.

Yanlış Pozitif (False Positive FP): Modelin yanlış bir şekilde pozitif olarak tahmin ettiği veri örneklerinin sayısını temsil eder. Yani, gerçek sınıfı negatif olan veri örnekleri yanlış bir şekilde pozitif olarak tahmin edilmiştir.

Gerçek Negatif (True Negative TN): Modelin doğru bir şekilde negatif olarak tahmin ettiği veri örneklerinin sayısını temsil eder. Yani, gerçek sınıfı negatif olan veri örnekleri doğru bir şekilde negatif olarak tahmin edilmiştir.

Yanlış Negatif (False Negative FN): Modelin yanlış bir şekilde negatif olarak tahmin ettiği veri örneklerinin sayısını temsil eder. Yani, gerçek sınıfı pozitif olan veri örnekleri yanlış bir şekilde negatif olarak tahmin edilmiştir.

Karışıklık matrisi, bu dört kavramı kullanarak sınıflandırma modelinin performansını değerlendirir. Karışıklık matrisi, bu kavramları görsel olarak sunarak modelin sınıflandırma sonuçlarını daha iyi anlamamıza yardımcı olur.

4.6.11. Friedman sıralama testi

Friedman sıralama testi, veri grupları arasındaki normallik sağlanmadığı durumlarda ortalama değerlerin karşılaştırmasını yapmaya imkân veren sık kullanılan amacı istatistiksel veri analizi metodudur.

Friedman sıralama testi, birbirine bağlı birden çok ölçüm kümesinin karşılaştırılmasında kullanılan bir non-parametrik istatistiksel testtir. Bu testte, farklı koşullar altında yapılan tekrarlı testler birbirleriyle karşılaştırılır ve sonuçlar arasında anlamlı bir fark olup olmadığına karar verilir. Test sonucunda elde edilen bir p değeri, örnekler arasındaki farkın istatistiksel olarak anlamlı olup olmadığını gösterir. Friedman sıralama testi, sadece iki örneklem arasındaki farklılığı değil, birden çok örneklem arasındaki farklılığı da değerlendirebilir. Uygulamada SPSS Programı yardımıyla sınıflandırma metotlarının çıktılarının (doğruluk değerlerinin) karşılaştırılması maksadıyla tercih edilmiştir.

4.7. Yazılım Gereçleri

Uygulamadaki yazılım gereçlerinden açık kaynak kodlu Anaconda, Jupyter Notebook, Python ve Python Numpy, Pandas, Seaborn, Matplotlib, Scikit-learn kütüphanelerinden, Minitab, IBM SPSS gibi istatistiksel yazılım gereçlerinden yararlanılmıştır.

4.7.1. Anaconda navigator

Anaconda, açık kaynaklı bir veri bilimi ve makine öğrenimi platformudur. Anaconda, Python programlama diline dayanmaktadır ve Python tabanlı veri analitiği ve makine öğrenimi projelerini desteklemek için birçok popüler kütüphaneyi içermektedir.

Anaconda, kullanıcı dostu bir kullanıcı arabirimi olan Anaconda Navigator ile gelir ve Python, Jupyter Notebook ve diğer popüler geliştirme araçlarını içerir. Anaconda'nın diğer avantajları, sanal ortamlar oluşturma, paket yönetimi, ortam yönetimi, raporlama ve paylaşma gibi veri bilimi ve makine öğrenimi projelerinde kullanışlı araçlar sunmasıdır. Çalışmada Anaconda Navigator içerisindeki Jupyter Notebook geliştirme aracı kullanılmıştır.

4.7.2. Jupyter notebook

Jupyter Notebook, açık kaynaklı bir projedir ve popüler bir veri bilimi, makine öğrenimi ve bilimsel hesaplama ortamıdır. Jupyter Notebook, kod hücreleri ve metin hücreleri olarak adlandırılan iki farklı tür hücre içerir. Kod hücrelerine Python kodu veya diğer programlama dillerinde kod yazabilir ve doğrudan çalıştırabilirsiniz. Metin hücreleri ise açıklamalar, belgelendirme veya Markdown formatında yazılmış metin içerebilir.

Jupyter Notebook, kodunuzu parçalara ayırarak adım adım çalışmanıza, sonuçları görsel olarak görmenize ve belgelendirme oluşturmanıza olanak tanır. Ayrıca, Jupyter Notebook'ta oluşturulan belgeleri HTML, PDF, Markdown, LaTex gibi farklı formatlarda kaydedebilir ve paylaşabilirsiniz.

Jupyter Notebook, veri bilimciler, araştırmacılar, geliştiriciler ve eğitimciler arasında yaygın olarak kullanılan bir araçtır ve veri analizi, makine öğrenimi, model geliştirme, veri görselleştirme ve prototip geliştirme gibi birçok uygulama alanında kullanılabilir. Çalışmada 4.9.2 versiyonlu Jupyter Notebook arayüzü kullanılmıştır.

4.7.3. Pyhton

Python okunabilirliği kolay, modüler ve yorumlanabilir bir betik dildir. Son yıllarda bilimsel hesaplama yöntemlerinin başında gelen yapay zekâ alanında yaygın şekilde kullanılmaktadır. Sonuçları kolaylıkla okuma, analiz etme ve görselleştirme imkânı sayesinde araştırmacılar tarafından tercih edilmektedir. Platform bağımsız olması en büyük avantajlarından bir tanesidir. Çalışmada 3.9.12 versiyonlu Python ürünü kullanılmıştır.

4.7.4. Pyhton kütüphaneleri

Anaconda, birçok veri bilimi, makine öğrenimi ve yapay zekâ alanında kullanılan popüler kütüphaneleri (örneğin, NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, Keras, PyTorch vb.) içinde barındırır ve bu kütüphanelerle ilgili bağımlılıkları yönetir. Bu, kullanıcıların ayrı ayrı bu kütüphaneleri kurmak ve yönetmek yerine tek bir platformda hepsini bir arada kullanmalarına olanak tanır. Çalışmada kullanılan Python kütüphaneleri aşağıdaki gibidir.

4.7.4.1. NumPy

Python programlama dili için açık kaynaklı bir bilimsel hesaplama kütüphanesidir. NumPy, büyük, çok boyutlu diziler ve matrisler üzerinde hızlı ve verimli hesaplamalar yapmak için geliştirilmiştir. NumPy, temel matematiksel işlemler, rastgele sayı üretimi, doğrusal cebir, dizi indeksleme ve dilimleme, veri dönüştürme ve daha birçok matematiksel ve bilimsel hesaplama işlevini içerir.

NumPy, Python'daki listelerden daha hızlı ve daha verimlidir, çünkü dizi ve matrislerde homojen veri tipleri kullanır ve bellek yönetimini optimize eder. NumPy, aynı zamanda veri analitiği, veri işleme, veri bilimi ve makine öğrenimi gibi birçok alanda da yaygın olarak kullanılan bir temel kütüphanedir. Çalışmada 1.21.5 versiyonlu NumPy kütüphanesi kullanılmıştır.

4.7.4.2. Pandas

Python programlama dili için açık kaynaklı bir veri analitiği ve veri işleme kütüphanesidir. Pandas, etkili veri analitiği ve veri manipülasyonu için geliştirilmiş yüksek performanslı veri yapıları ve veri analitiği araçları sağlar.

Pandas, temel olarak iki ana veri yapıları üzerinde çalışır:

- Veri Çerçevesi: İki boyutlu, etiketli ve sütun-tabanlı bir veri yapısıdır. Excel'deki gibi bir tabloyu veya SQL'deki gibi bir veri tabanını temsil edebilir. Veri çerçevesi, farklı veri tiplerini içeren sütunlar halinde organize edilmiş verileri kolayca işlemek ve analiz etmek için kullanılır.
- Seriler (Series): Bir boyutlu etiketli bir veri yapısıdır. Veri çerçevesinin herhangi bir sütununu temsil edebilir ve tek bir sütunun verilerini ve etiketlerini içerir.

Pandas, veri analitiği ve veri işleme süreçlerinde bir dizi işlevi destekler, örneğin veri temizleme, veri dönüşümü, veri filtreleme, veri gruplama, veri birleştirme, zaman serisi analizi, istatistiksel analiz ve daha birçok veri manipülasyonu işlemi için kullanılabilir. Çalışmada 1.4.2 versiyonlu Pandas kütüphanesi kullanılmıştır.

4.7.4.3. Seaborn

Python programlama dili için açık kaynaklı bir veri görselleştirme kütüphanesidir. Matplotlib kütüphanesine dayanarak geliştirilen Seaborn, veri görselleştirme süreçlerini daha kolay ve hızlı hale getirmek için daha yüksek seviyeli bir arayüz sunar.

Seaborn, çekici ve bilgilendirici veri görselleştirmeleri oluşturmak için kullanılır. Görselleştirmelerde renk paletleri, stil şablonları ve istatistiksel grafikler gibi gelişmiş özelliklere sahip olmasıyla bilinir. Seaborn, veri analitiği ve veri görselleştirme süreçlerini daha hızlı ve etkili hale getirerek veriye görsel bir şekilde daha derinlemesine bakmaya yardımcı olur. Çalışmada 0.11.2 versiyonlu Seaborn kütüphanesi kullanılmıştır.

4.7.4.4. Matplotlib

Python programlama dili için açık kaynaklı bir veri görselleştirme kütüphanesidir. Veriye grafikler, çizimler ve görsel sunumlar ekleyerek veri analitiği ve veri görselleştirme süreçlerini destekler.

Matplotlib, kullanıcıların çeşitli grafik türlerini oluşturmasına olanak tanır, örneğin çizgi grafikleri, bar grafikleri, histogramlar, dağılım grafikleri, pasta grafikleri, kutu grafikleri, 3D grafikler ve daha birçok farklı grafik türü. Aynı zamanda grafiklere etiketler, başlıklar, eksenler, renk paletleri ve stil özellikleri eklemek gibi ayrıntılı özelleştirme seçenekleri sunar.

Matplotlib, kullanıcı dostu bir API (Application Programming Interface) sağlar ve diğer Python kütüphaneleriyle uyumludur. Ayrıca geniş bir kullanıcı topluluğuna sahip olması, dokümantasyonun zengin olması ve sürekli olarak güncellenmesi gibi avantajları bulunur.

4.7.4.5. Scikit-learn

Yaygın kullanılan makine öğrenimi lineer regresyon, lojistik regresyon, karar ağaçları, destek vektör makineleri, rastgele ormanlar, KNN ve daha birçok algoritmayı içerir.

Scikit-learn, aynı zamanda model seçimi, model değerlendirme ve model optimizasyonu gibi önemli makine öğrenimi süreçlerini destekler. Çapraz doğrulama, hiper parametre optimizasyonu, model performans değerlendirmesi, model seçimi ve model ayarlama gibi işlemleri kolayca gerçekleştirmek için kullanıcı dostu bir arayüz sunar.

Scikit-learn, akademik araştırmalardan endüstriyel uygulamalara kadar birçok alanda yaygın olarak kullanılır.

Veri analitiği, görüntü işleme, doğal dil işleme, tıbbi veri analizi, finansal analiz, reklam tahminleri, müşteri segmentasyonu, hedefleme ve daha birçok alanda kullanıcıların makine öğrenimi modelleri oluşturmasına, değerlendirmesine ve kullanmasına yardımcı olur. Çalışmada 1.0.2 Scikit-learn versiyonlu Python kütüphanesi kullanılmıştır.

4.7.5. Minitab

Minitab, Amerika Birleşik Devletleri'nde bulunan ve yazılım geliştirme, istatistiksel danışmanlık ve eğitim hizmetleri sunan Minitab, LLC şirketi tarafından geliştirilmiştir. Şirket, 1972 yılında Pennsylvania State üniversitesinde öğrenci olan Barbara F. Ryan ve Thomas A. Ryan tarafından kurulmuştur.

Minitab, günümüzde dünya genelinde 100'den fazla ülkede yaygın olarak kullanılan bir yazılımdır ve birçok endüstri, akademik kurum ve araştırma merkezi tarafından tercih edilmektedir.

Minitab, istatistiksel veri analizi ve kalite kontrolü için kullanılan bir yazılımdır. Minitab, verilerin analizini kolaylaştıran araçlar, grafikler ve tablolar sunar. Minitab, çok çeşitli istatistiksel testlerin yanı sıra doğrusal ve çoklu regresyon, varyans analizi, faktör analizi, zaman serisi analizi ve daha birçok analiz yöntemini destekler. Minitab, araştırmacılar, öğrenciler ve endüstriyel kullanıcılar gibi çeşitli kullanıcılar tarafından kullanılmaktadır. Minitab, basit arayüzü ve kullanıcı dostu özellikleri ile tanınır ve istatistiksel analiz ve kalite kontrolü için yaygın olarak kullanılan bir yazılımdır. . Çalışmada 17.1.0 versiyonlu yazılımı kullanılmıştır.

4.7.6. Ibm Spss

SPSS (Statistical Package for the Social Sciences), sosyal bilimler ve istatistik alanlarında kullanılan bir istatistiksel analiz yazılımıdır. SPSS, verilerinizi analiz etmek, grafikler oluşturmak ve sonuçları raporlamak için bir dizi araç sağlar.

SPSS, sıklık tabloları, çapraz tablolar, ortalama, standart sapma, regresyon analizi, faktör analizi, kümeleme analizi, t testleri, ANOVA, MANOVA, ANCOVA, non-parametrik testler gibi birçok temel ve gelişmiş istatistiksel analiz yöntemlerini destekler.

SPSS, sosyal bilimler, psikoloji, işletme, sağlık, eğitim ve birçok diğer disiplinde araştırmacılar tarafından yaygın olarak kullanılmaktadır.

Friedman sıralama testi için 29.0.1.0 versiyonlu SPSS programından yararlanılmıştır.

4.7.7. Microsoft 365

Tablolar Microsoft'un Excel uygulaması, blok diyagramlar Visio uygulaması yardımıyla hazırlanmıştır.

4.7.7.1. Microsoft excel

Excel, Microsoft tarafından geliştirilen ve elektronik tablolama işlevi sunan bir ofis yazılımıdır. İşletmeler, finans kurumları, okullar ve birçok farklı sektörde kullanılmaktadır. Excel, verileri hızlı bir şekilde girme, düzenleme ve analiz etme yeteneği sunar.

Excel, kullanıcıların sayılar, metin, tarihler ve diğer verileri hücreler içinde saklamasına ve bu verileri formüller kullanarak hesaplamasına olanak tanır. Kullanıcılar, verileri tablo, grafik veya pivot tablo gibi farklı biçimlerde görselleştirebilirler. Excel ayrıca, verilerin sıralanması, filtrelenmesi ve aranması gibi çeşitli veri işleme işlevleri sunar.

4.7.7.2. Microsoft visio

Microsoft Visio, Microsoft tarafından geliştirilen diyagram çizme yazılımıdır. Visio, iş süreçleri, organizasyon şemaları, akış şemaları, ağ diyagramları, elektrik devreleri, bina planları ve diğer türden çizimleri oluşturmak için kullanılır.

4.8. Donanım Gereçleri

Çalışmada Huawei marka AMD Ryzen 5 3500U 2.10 GHz işlemcili, 8 Gigabyte sanal bellekli, 64 bit Windows 11 Home işletim sistemli bir mobil bilgisayar kullanılmıştır.

5. UYGULAMA

5.1. Uygulamanın Amacı

Teknolojinin gelişmesi birlikte büyük veri kullanımı da artan bir hızla yaygınlaşmaktadır. Verinin depolanması, analiz edilmesi ve gizliliğinin sağlanması konuları geliştirilmesi gereken algoritma yöntemlerini de beraberinde getirmiştir. Veri gizliliğinin sağlanması bununla birlikte veri güvenliğinin sağlanması tamamen blok zinciri yaklaşımı ile verilerin pertürbasyon yöntemi ile verilerin kısmen parçalara ayrılması ve işlenmesi durumudur.

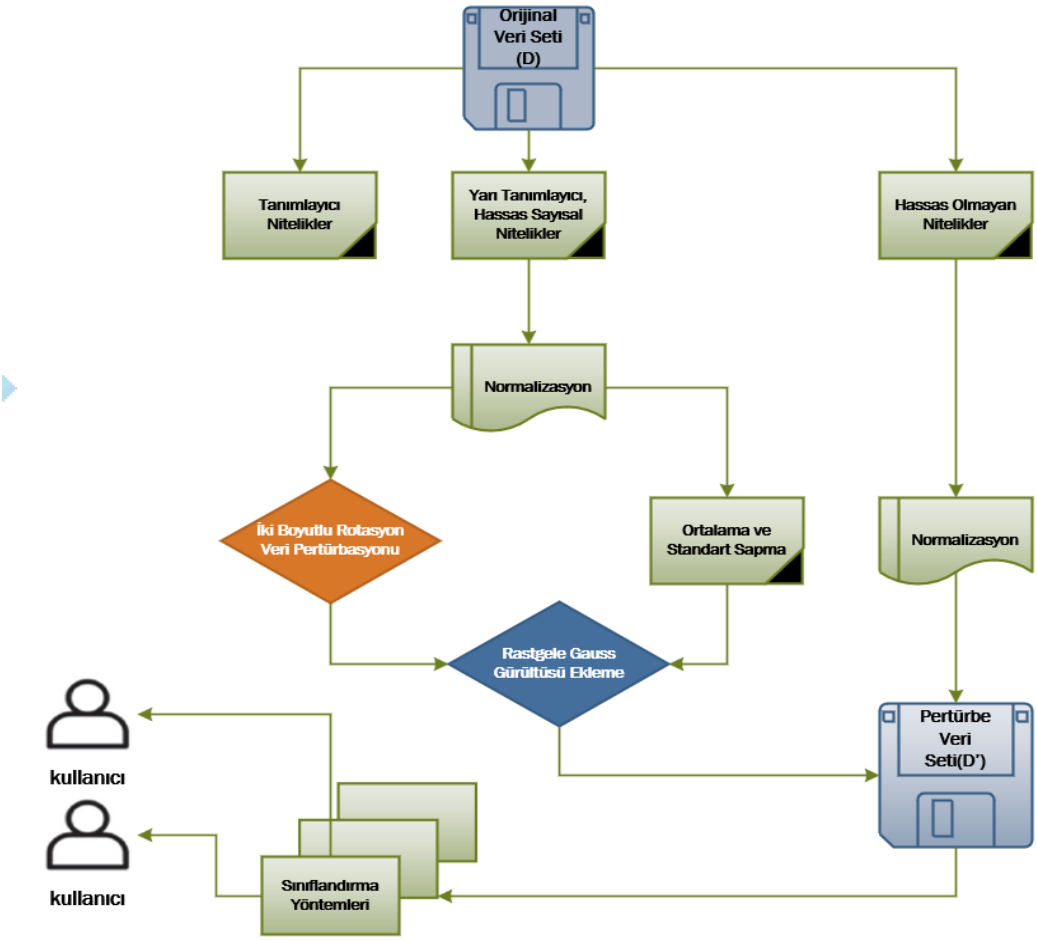
Bu proje kapsamında, normalleştirme, geometrik döndürme, doğrusal regresyon ve skaler veri çarpımı ile hassas veri madenciliğinde karşılaştırmalı sınıflandırma kullanılarak verilerin analiz işlemi gerçekleştirilmiştir.

Çalışmada literatürde yaygın olarak tercih edilmiş veri güvenliği yöntemlerinden ikisi birleştirilmiştir. İki boyutlu döndürmeye bağlı ve rastgele gauss mekanizmasının birlikte uygulandığı bir yöntem önerilmiştir. Önerilen yöntem ile veri gizliliğinin artırılması hedeflenmiştir. İki boyutlu yarı tanımlayıcı hassas özellikler öncelikle ikili gruplar halinde bir araya getirilmiş, rastgele bir açı mertebesinde saat yönünü istikametinde iki boyutlu döndürme işlemine tabi tutulmuştur. Sonrasında her bir özelliğinin standart sapma ve ortalamaları hesaplanarak bu veriler gauss mekanizmasına girdi olarak sunulmuştur. Böylelikle gauss mekanizması da rastgele daha önemlisi özelliklerin standart sapma ve ortalamalara bağlı bir gürültü oluşturmuştur. Evvela tanımlayıcı nitelikler veri setlerinden çıkarılmıştır. Bunlar isim, soyisim, kimlik numaraları, bilet numaralarıdır. Yarı tanımlayıcı özellikli hassas ikili veri çerçeveleri, 110 derece saat yönünde döndürülüp, gauss gürültüsü de eklenerek veri gizliliği artırılmıştır. Dört veri seti altı sınıflandırma algoritmasına tabi tutulmuş ve sınıflandırma doğrulukları karşılaştırılmıştır. Veri doğruluğu, veri güvenliği performans metrikleri açısından değerlendirmiştir.

Çalışmanda kullanılan yöntemin performansını anlamak için anonimleştirme öncesi ve sonrası sınıflandırma yöntemlerinin doğruluk değerlerindeki, F1 skorlarındaki,

istatistiksel özelliklerindeki değişim miktarları Friedman sıralama testleri yardımıyla ölçülmüştür.

Temel yapının blok diyagramı Şekil 5.1.' de verilmiştir.



Şekil 5.1. Uygulamanın işlem adımları/temel yapının blok diyagramı.

5.2. Uygulamanın Adımları

Bu bölümde önerilen pertürbasyon metodunun işleyişi adım adım ele alınmıştır.

• Adım-1 Veri madenciliği yöntemlerinin seçimi**:** Tahmine dayalı sonuçlarının kıyaslamaya elverişli olması ve öncesi ve sonrası durumların analizine, büyük veri kümeleriyle çalışmaya imkân vermesi, günlük hayattan farklı veri tiplerine uyumluluk göstermesi, çeşitli endüstrilerde yaygın olarak kullanılıyor olması nedeniyle sınıflandırma yöntemleri tercih edilmiştir.

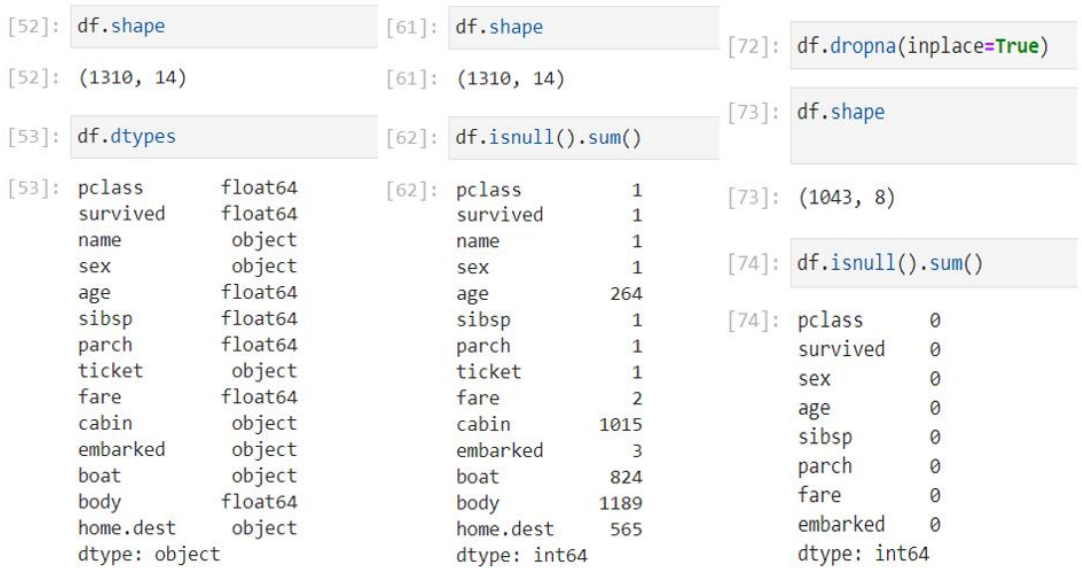
Çalışmada eski ve yeni karma sınıflandırma yöntemlerimden Lojistik Regresyona, K-En Yakın Komşuluk Yöntemine, Yapay Sinir Ağlarına, Destek Vektör Makinelerine,

Gradyanı Artırılan Karar Ağaçlarından XGboost ve LightGBM sınıflandırma yöntemlerine yer verilmiştir.

• Adım-2 Veri setlerinin seçimi: Çalışmada, seçime bağlı sınıflandırma algoritmalardan denetimli öğrenmeye imkân veren, içerinde yarı tanımlayıcı hassas veriler barındıran, çalışmayı okuyan herkesin ulaşabildiği, açık kaynak kodlu Titanik, göğüs kanseri, diyabet ve kalp krizi dört gerçek veri setine yer verilmiştir.

• Adım-3 Veri ön işlemesi: Veri seçimi aşamasından sonra, veri ön işleme adımına geçilmiştir. Bu aşamada, veriler temizlenir, düzenlenir ve hazırlanır. Veri önişleme aşaması, verilerin doğruluğunu ve tutarlılığını arttırmak için önemlidir.

Titanik Veri Seti Ön İşleme: Titanik veri seti içerinde kayıp verilerin olduğu 1310 kayıt ve 14 değişkenden oluşan bir veri setidir. Kayıp veriler tamamı silme yöntemi ile veri setinden çıkarılmıştır. Çalışmada kullanılan veri setinin kayıt sayısı 1043 adete ve özellik sayısı da 8 adete inmiştir. (Şekil 5.2, Tablo 5.1.)



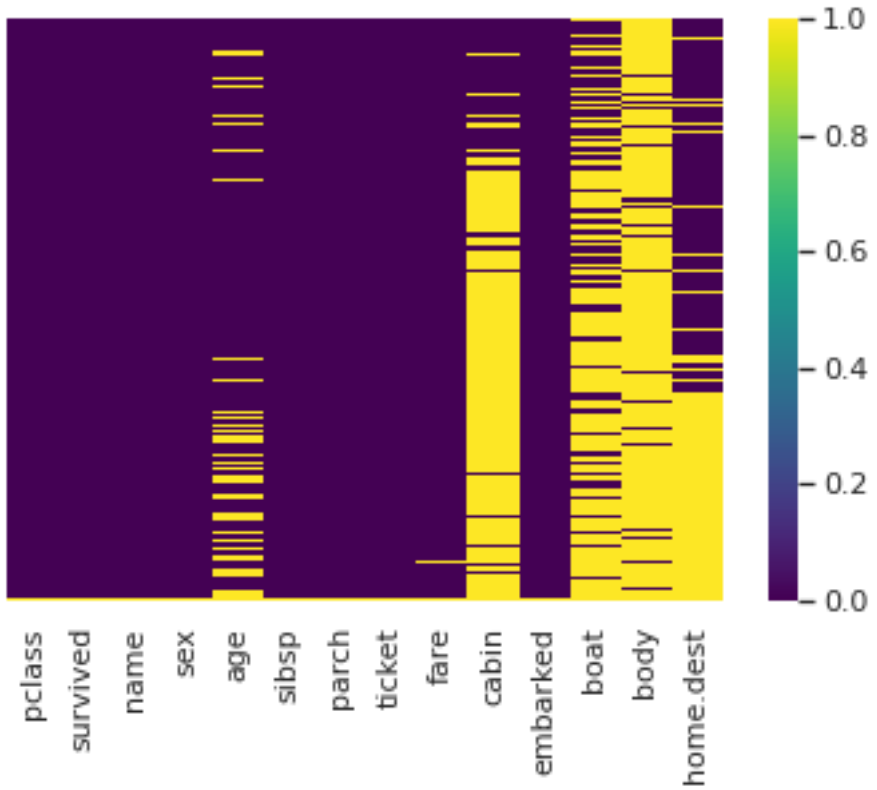
Şekil 5.2. Titanik veri seti ön işleme örnek python kod çıktısı.

Veri setinde yer alan cinsiyet, gemiye biniş limanı ve bilet sınıfı değişkenleri tek çizgi kodlama yapılarak ikili sayı sistemine göre düzenlenmiştir. Veri seti hedef ve tahmin olmak üzere nitelikler kapsamında yatayda ikiye ayrılmıştır.

Tablo 5.1. Titanik veri seti ön işleme örnek python kod çıktısı.

	survived	age	sibsp	parch	fare	isMale	S	C	Q	pc1	pc2	pc3
Adet	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043	1043
Ortalama	0,41	29,81	0,5	0,42	36,6	0,63	0,2	0,05	0,75	0,27	0,25	0,48
Standart	0,49	14,37	0,91	0,84	55,75	0,48	0,4	0,21	0,43	0,44	0,43	0,5
Min	0	0,17	0	0	0	0	0	0	0	0	0	0
25%	0	21	0	0	8,50	0	0	0	0	0	0	0
50%	0	28	0	0	15,75	1	0	0	1	0	0	0
75%	1	39	1	1	35,08	1	0	0	1	1	0,5	1
Maks	1	80	8	6	512,33	1	1	1	1	1	1	1

• Göğüs Kanseri Veri Seti Ön İşleme: Göğüs Kanseri veri seti içerinde sadece bir nitelik tamamının eksik verilerden oluştuğu 569 kayıt ve 33 değişkenden oluşan bir veri setidir. Kayıp veri Unnamed32 ve tanımlayıcı ve id veri setinden çıkarılmıştır (Şekil 5.3.). Çalışmada kullanılan veri setinin kayıt sayısı 569 adete ve özellik sayısı da 31 adete ve inmiştir. Veri seti hedef ve tahmin olmak üzere nitelikler kapsamında yatayda ikiye ayrılmıştır.



Şekil 5.3. Göğüs kanseri veri seti ön işleme eksik veriler grafiği.

Diyabet Veri Seti Ön İşleme: Göğüs kanseri içerinde eksik verilerin olduğu 768 kayıt ve 8 değişkenden oluşan bir veri setidir. Özniteliklerin ortalamaları, standart sapmaları, minimum, maksimum, ortanca, kartel değerleri Tablo 5.2.'de olduğu gibi hesaplanmıştır. Kayıp veri Glucose, BloodPressure, SkinThickness, Insulin, BMI değişkenlerinin ortancaları hesaplanarak eksik veriler tamamlanmıştır. Ayrıca veri setindeki numerik ve kategorik değişkenler sınıflandırılmış ve bazı değişkenler tek çizgi kodlama yapılarak ikili sayı sistemine göre düzenlenmiştir. Veri seti hedef ve tahmin olmak üzere nitelikler kapsamında yatayda ikiye ayrılmıştır.

Tablo 5.2. Diyabet seti ön işleme örnek python kod çıktısı.

	Adet	**Ortalama**	**Std S.**	**Min**	**25%**	**50%**	**75%**	**Maks**
Pregnancies	768	3,85	3,37	0	1,00	3,00	6,0	17,0
Glucose	768	120,89	31,97	0	99,00	117,00	140,3	199,0
BloodPressure	768	69,11	19,36	0	62,00	72,00	80,0	122,0
SkinThickness	768	20,54	15,95	0	0,00	23,00	32,0	99,0
Insulin	768	79,80	115,24	0	0,00	30,50	127,3	846,0
BMI	768	31,99	7,88	0	27,30	32,00	36,6	67,1
DiabetesPedigreeFun	768	0,47	0,33	0,078	0,24	0,37	0,6	2,4
Age	768	33,24	11,76	21	24,00	29,00	41,0	81,0
Outcome	768	0,35	0,48	0	0,00	0,00	1,0	1,0

Kalp Krizi Veri Seti Ön İşleme: İçerinde eksik verisi olamayan 303 adet kayıt ve 14 adet değişkenden oluşan tamamı nümerik bir settir (Şekil 5.4.). Glucose, BloodPressure, SkinThickness, Insulin, BMI değişkenlerinin ortanca hesaplanarak eksik veriler tamamlanmıştır. Ayrıca veri setindeki numerik ve kategorik değişkenler sınıflandırılmış ve bazı değişkenler tek çizgi kodlama yapılarak ikili sayı sistemine göre düzenlenmiştir. Veri seti hedef ve tahmin olmak üzere nitelikler kapsamında yatayda ikiye ayrılmıştır.

```
df
```

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Şekil 5.4. Kalp Krizi Seti Ön İşleme Örnek Python Kod Çıktısı.

• Adım-4 Pertürbe edilecek sayısal değişkenlerin belirlenmesi: Çalışmadaki veri setlerinde rotasyon ve gauss gürültüsü eklenecek öznitelikler tanımlayıcı sayısal değişkenler arasından seçilmiştir. Bunlar değişkenler aşağıdaki gibidir.

- Titanik: age, fare
- Göğüs Kanseri: radius_mean, texture_mean
- Diyabet: age, glucose
- Kalp Krizi: age, chol

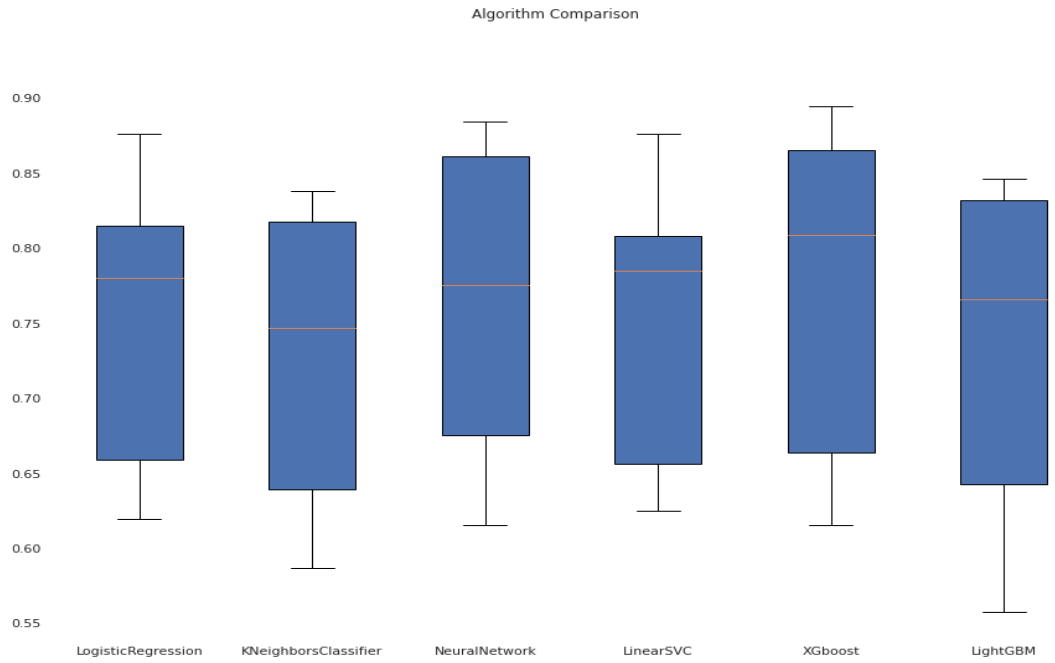
• Adım-5 Sayısal değişkenlerin ölçeklendirilmesi: Titanik, göğüs kanseri ve kalp krizi veri setindeki age, fare, radius_mean, texture_mean, age ve chol değişkenleri min-maks normalizasyonu 0,1 aralığında standartlaştırılmıştır.

Diyabet veri setinde yer alan tüm nümerik değişkenler ise Pregnancies, GlucosBloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age robust ölçeklendirme yöntemine göre standartlaştırılmıştır.

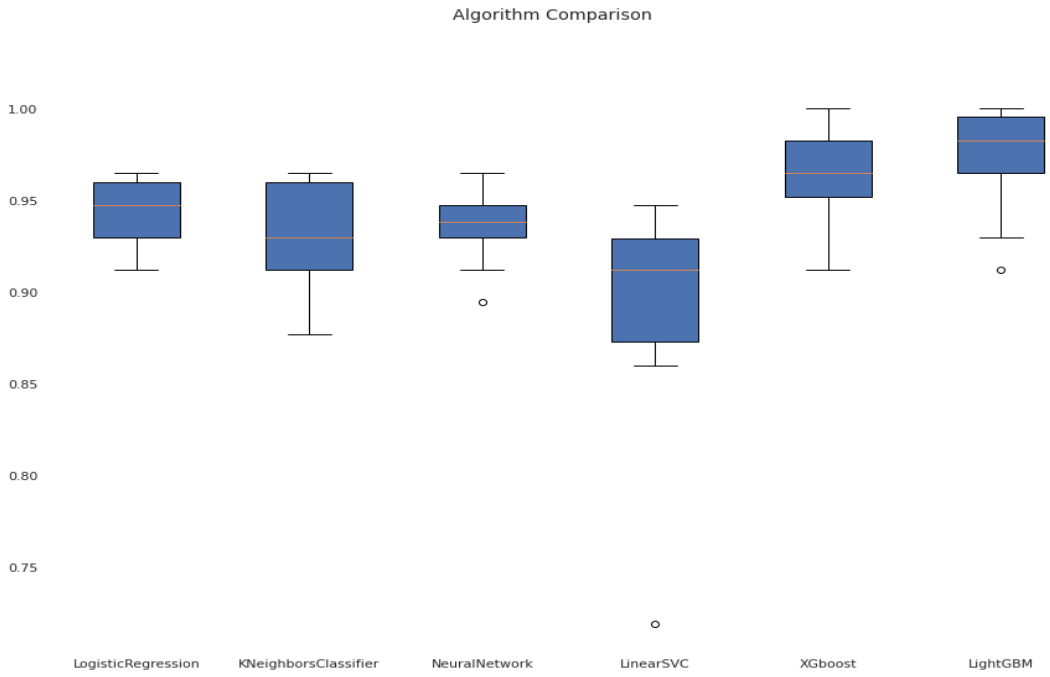
• Adım-6 Orijinal veri setinin kümelere ayrılması: Veri setleri sınıflandırma öncesi doğruluk performansını artırmak maksadıyla çapraz doğrulama denilen yönteme başvurulmuştur. Bu yöntem ile veri setini k kümeye ayırarak gerçekleştirilir ve her bir kümeye sırayla test seti olarak atanır. Diğer k-1 kümeler ise eğitim seti olarak

kullanılır. Çalışmada aşırı öğrenmeye ve veri dengesizliklerine yol almayacak şekilde k değeri 10 olarak belirlenmiştir.

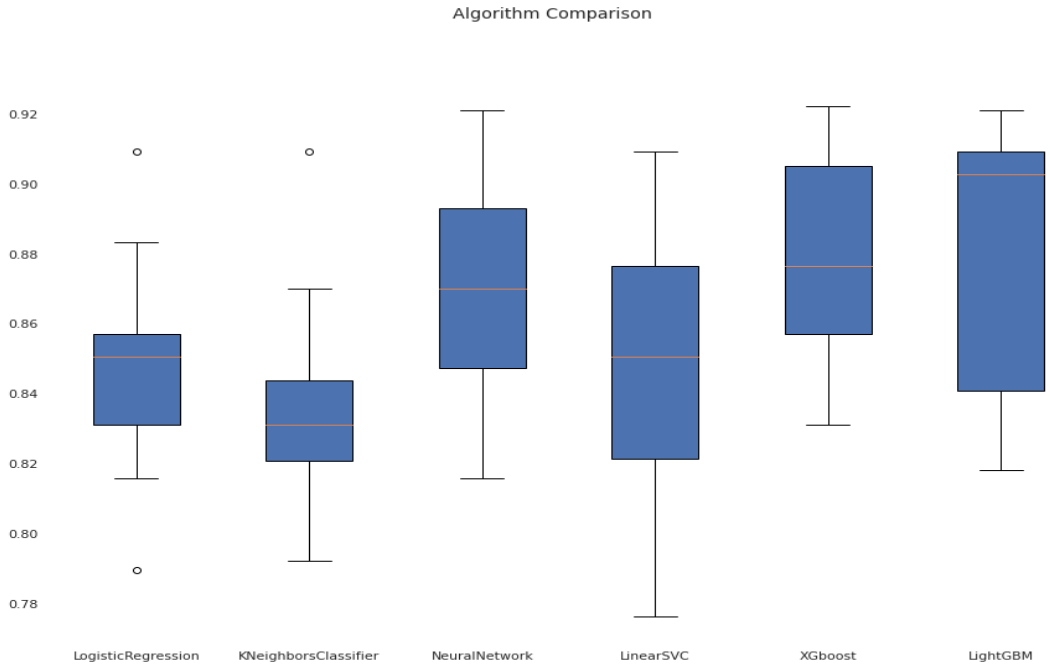
• Adım-7 Orijinal veri setinin sınıflandırılması: Test ve eğitim olarak kümelere ayrılan orijinal veri setleri 6 sınıflandırma yöntemi ile doğrulukları, kesinlik, duyarlılık, F1 Skorları ve bunlara ait standart sapmaları Python programı yardımı ile hesaplanmıştır. Dört veri seti için de doğruluk ortalama ve standart sapma değerlerinin karşılaştırıldığı kutu diyagramları Şekil 5.5, Şekil 5.6, Şekil 5.7, Şekil 5.8'de verilmiştir.



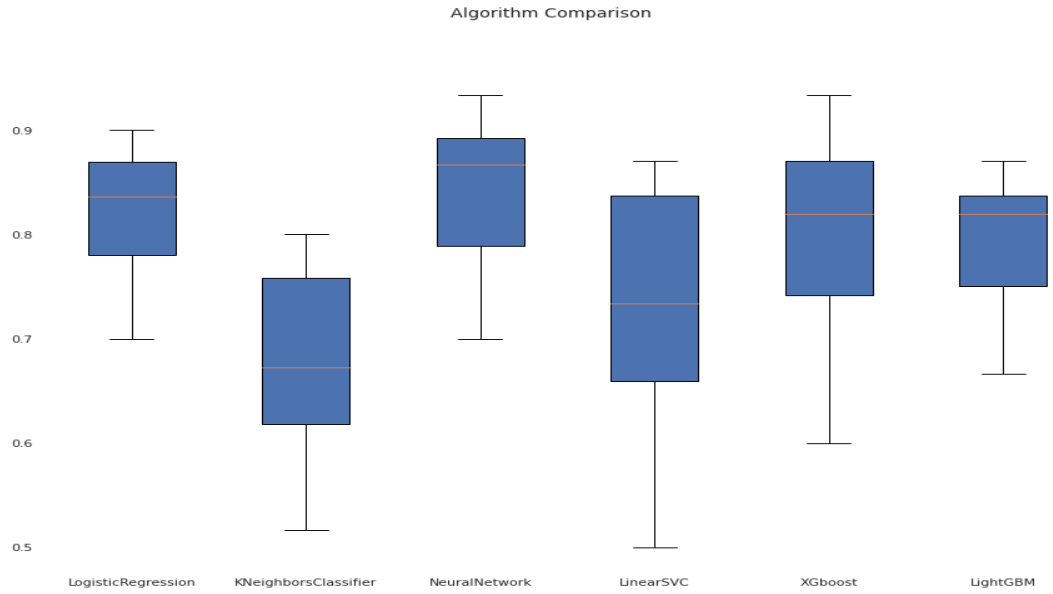
Şekil 5.5. Titanik veri seti sınıflandırma doğruluk sonuçları kutu grafiği.



Şekil 5.6. Göğüs kanseri veri seti sınıflandırma doğruluk sonuçları kutu grafiği.



Şekil 5.7. Diyabet veri seti sınıflandırma doğruluk sonuçları kutu grafiği.

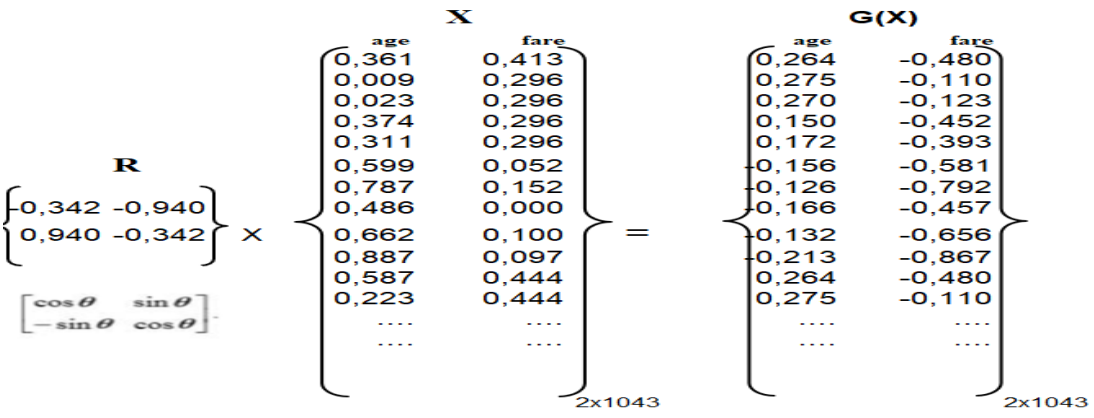


Şekil 5.8. Kalp krizi veri seti sınıflandırma doğruluk sonuçları kutu grafiği.

• Adım-8 Rotasyon pertürbasyonu: Orijinal veri setleri daha önce belirlenen ve standartlaştırılan ikili nümerik değişkenler sırasıyla rastgele seçilen (çalışmada saat yönünde 110 derece) koordinat sisteminde saat yönünde öklid mesafesi korunacak şekilde döndürülerek pertürbe edilmişlerdir. Döndürme işlemi her veri seti için aşağıdaki koordinat sistemlerinde öncesi ve sonrası olacak şekilde gösterilmiştir.

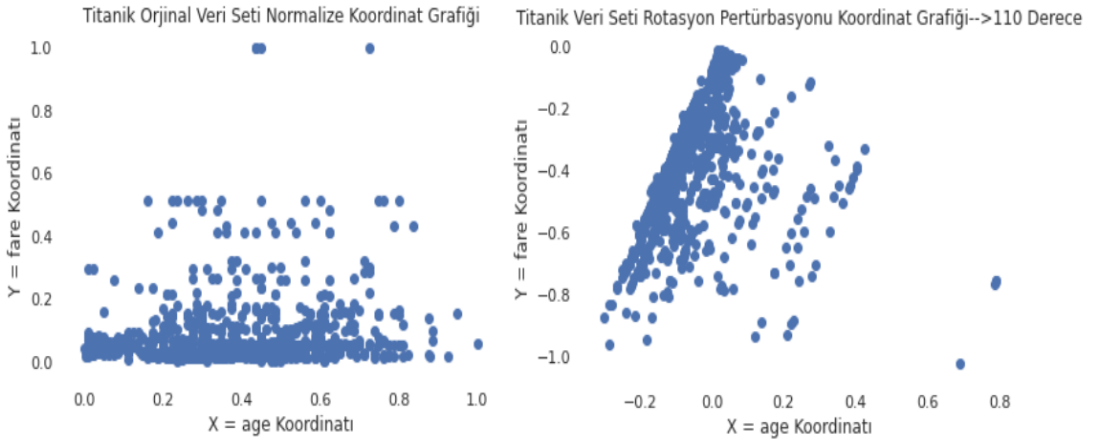
$$G(X) = RX \tag{5.1}$$

Burada 5.1'deki denklemde verilen 110 dereceye karşılık gelen R matris aşağıdaki gibidir. X ise 2xn veri gizliliği sağlanmak istenen yarı tanımlayıcı veri kümesidir (Şekil-5.9).

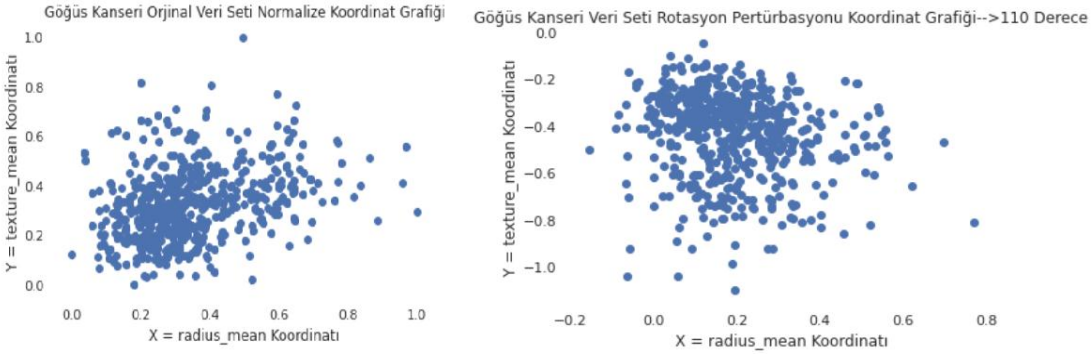


Şekil 5.9. Titanik veri setinin normalizasyonu sonrası rotasyon örneği.

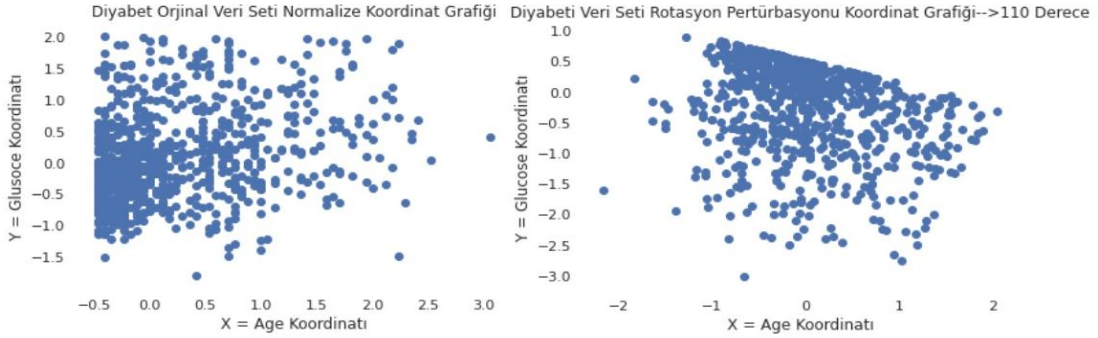
110 dereceye karşılık gelen veri setlerinin koordinat düzleminde döndürülmüş öncesi ve sonrası halleri ayrı ayrı Şekil 5.10, Şekil 5.11, Şekil 5.12, Şekil 5.13'te gösterilmiştir.



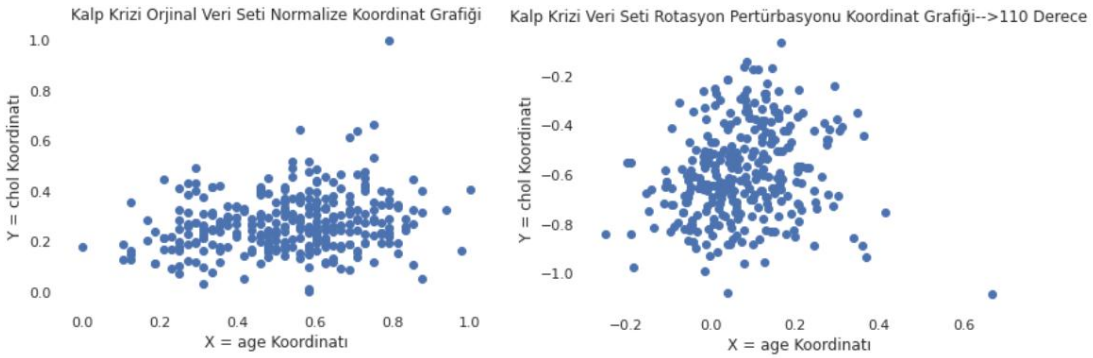
Şekil 5.10. Titanik veri setinin age ve fare değişkenlerinin koordinat düzleminde döndürülmesi.



Şekil 5.11. Göğüs kanseri veri setinin radius_mean ve texure_mean değişkenlerinin koordinat düzleminde döndürülmesi.



Şekil 5.12. Diyabet veri setinin age ve glucose değişkenlerinin koordinat düzleminde döndürülmesi.

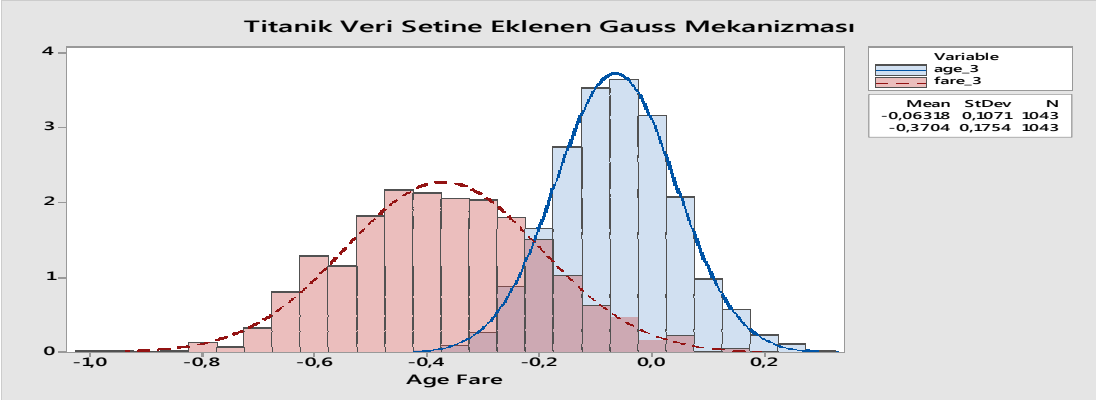


Şekil 5.13. Kalp krizi veri setinin age ve chol değişkenlerinin koordinat düzleminde döndürülmesi.

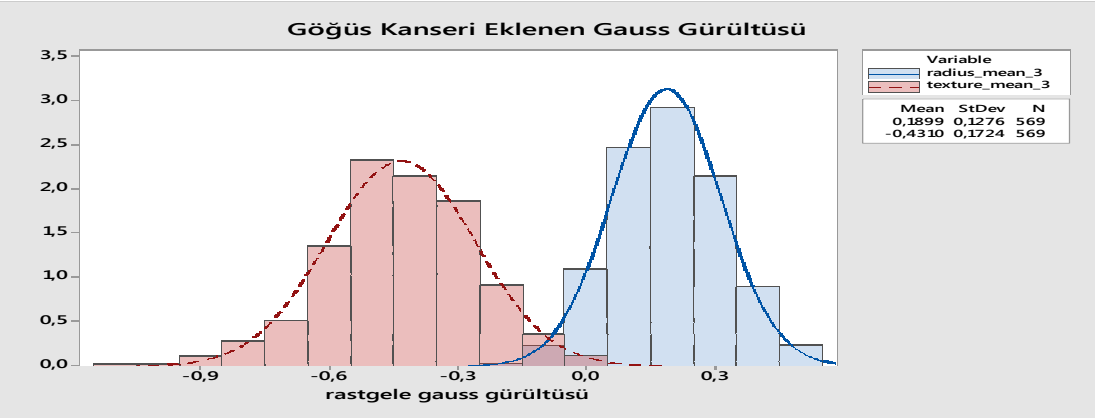
• Adım-9 Rastgele gürültü ekleme: Döndürülmüş veri kümesinin ayrı ayrı her bir niteliğinin ortalaması ve standart sapması hesaplanır. Gauss mekanizmasına girdi olarak sunulur. Böylelikle rotasyona tabi tutulan özelliklerin her birine rastgele bir gauss gürültüsü eklenmiş olur. Burada maksat veri gizliliğini bir seviye daha artırmaktır. Yarı tanımlayıcı özellikteki değişkenler saldırılara daha dirençli hale getirilmeye çalışılmaktadır. Tüm veri setleri için eklenen gauss gürültülerinin histogram diyagramları aşağıda verilmiştir.

$$\Delta = \text{Gauss Gürültüsü} = [(1/ (var* sqrt(2*PI)))*(pow(e,(-pow (x-mean), 2)/2*pow (var, 2)))] \quad (5.2)$$

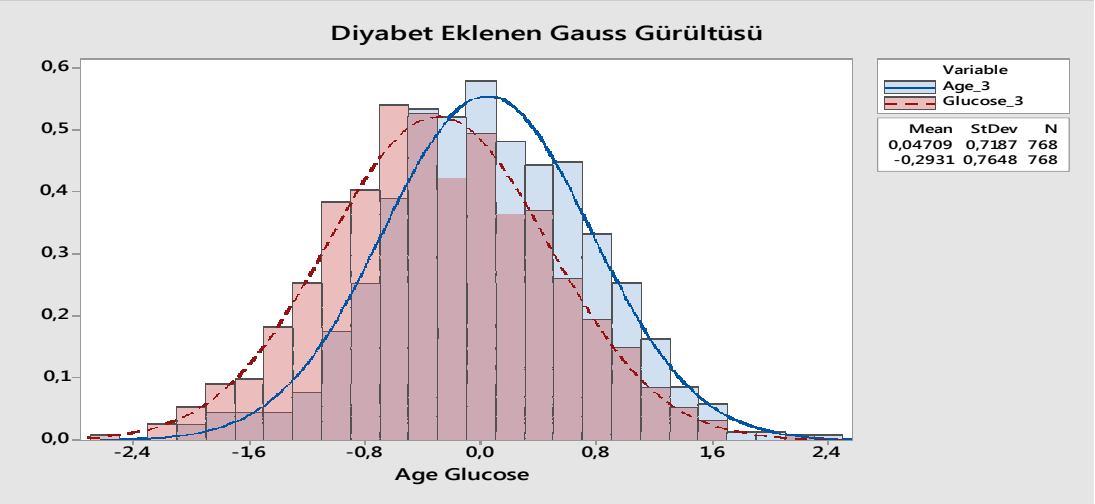
Burada G(X) = RX + Δ 110 derece karşılık gelen R matrisini, X ise 2_{xn} veri gizliliği sağlanmak istenen yarı tanımlayıcı veri kümesini, Δ eklenen rastgele gürültüyü temsil eder. Eklenen rastgele gauss gürültülerine ait histogram grafikleri ayrı ayrı Şekil 5.14, Şekil 5.15, Şekil 5-16, Şekil 5.17'de verilmiştir.



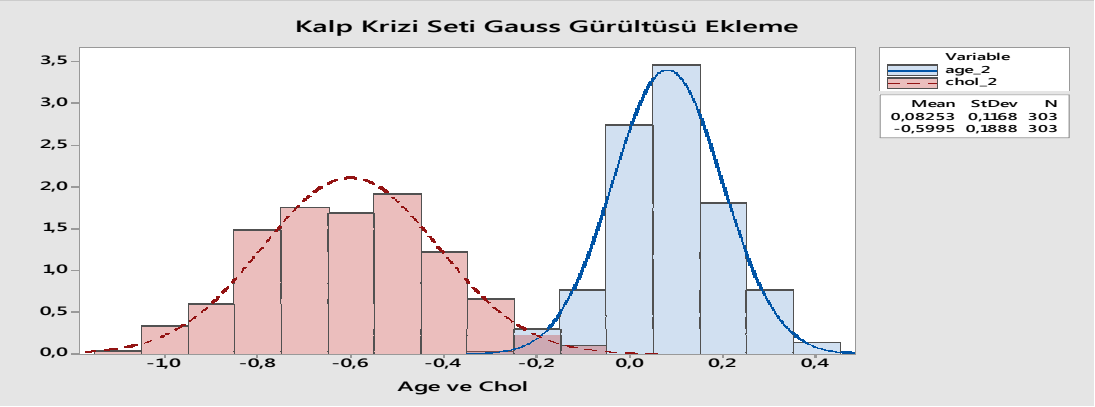
Şekil 5.14. Titanik veri setinin age ve fare değişkenlerine eklenen gauss gürültüleri.



Şekil 5.15. Göğüs kanseri veri setinin radius_mean ve texture_mean değişkenlerine eklenen gauss gürültüleri.



Şekil 5.16. Diyabet Veri Setinin Age ve Glucose Değişkenlerine Eklenen Gauss Gürültüleri.



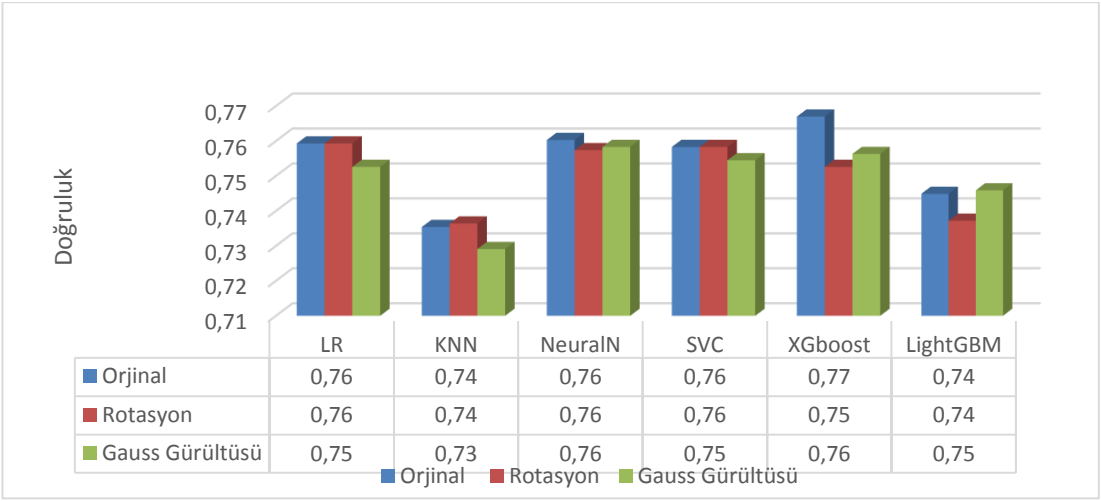
Şekil 5.17. Kalp krizi veri setinin age ve chol değişkenlerine eklenen gauss gürültüleri.

• Adım-10 Pertürbe veri setinin kümelere ayrılması: Rotasyon ve gürültü eklenmiş veri setleri aşırı öğrenmeye ve veri dengesizliklerine yol almayacak şekilde k değeri 10 seçilerek çapraz doğrulamaya tabi tutulmuştur.

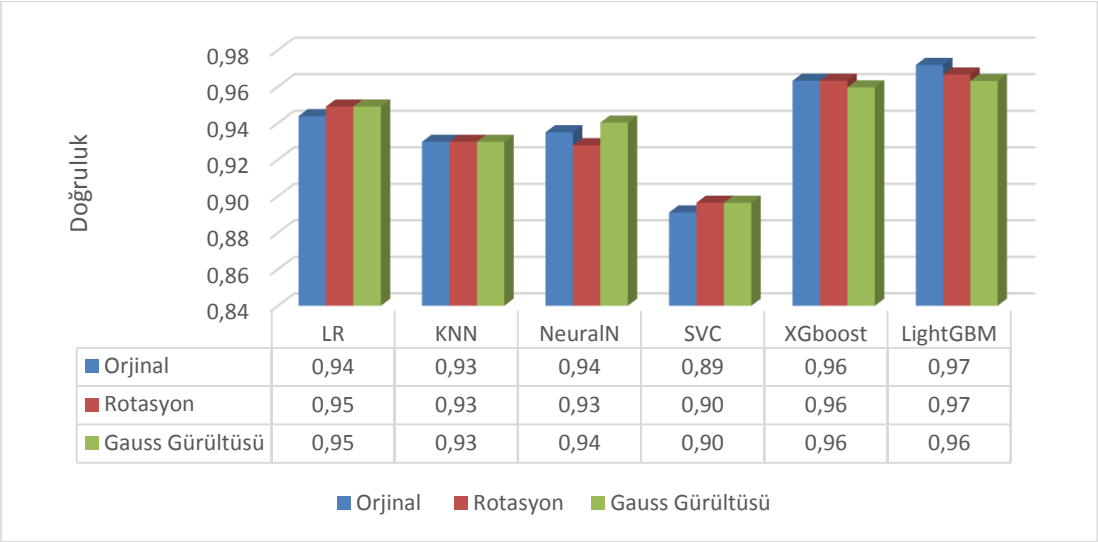
• Adım-11 Pertürbe edilmiş veri setlerinin sınıflandırılması: Test ve eğitim olarak kümelere ayrılan önce rotasyona sonrasında gürültü eklenmiş veri setleri yine aynı 6 sınıflandırma yöntemi ile doğrulukları, kesinlik, duyarlılık, f1 skorları ve bunlara ait standart sapmaları Python programı yardımı ile hesaplanmıştır. Hesaplanan sınıflandırma yöntemlerinin doğruluk sonuçları Tablo5.3.'de bir arada verilmiştir. Ayrı ayrı veri setlerinin pertürbe edilmiş ve edilmemiş hallerinin çubuk diyagramları da Şekil 5.18, Şekil 5.19, Şekil 5.20, Şekil 5.21'de gösterildiği gibidir.

Tablo 5.3. Pertürbe edilmiş veri setlerinin sınıflandırma doğruluk değerleri.

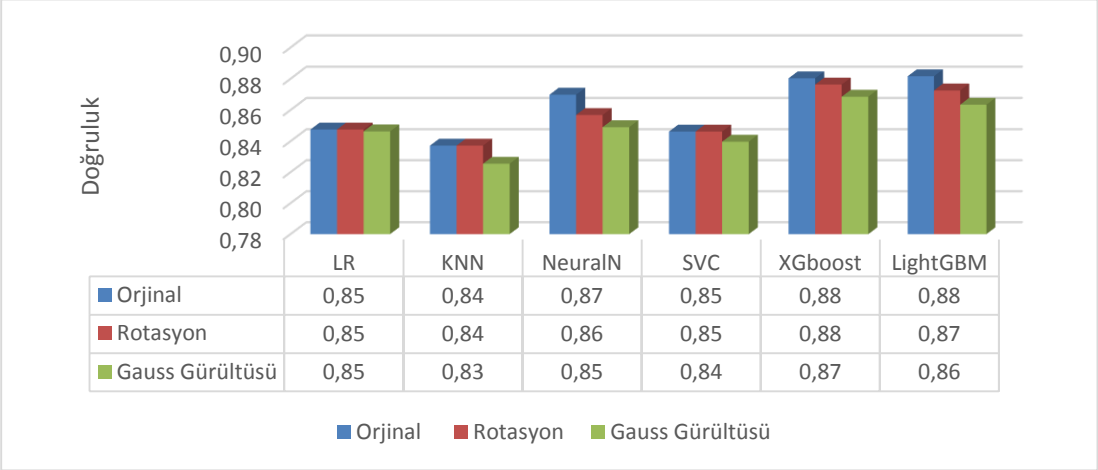
	LR	KNN	NN	SVM	XGboost	LightGBM
Orijinal	0,76	0,74	0,76	0,76	0,77	0,74
RDP	0,76	0,74	0,76	0,76	0,75	0,74
GM	0,75	0,73	0,76	0,75	0,76	0,75
Orijinal	0,85	0,84	0,87	0,85	0,88	0,88
RDP	0,85	0,84	0,86	0,85	0,88	0,87
GM	0,85	0,83	0,85	0,84	0,87	0,86
Orijinal	0,94	0,93	0,94	0,89	0,96	0,97
RDP	0,95	0,93	0,93	0,90	0,96	0,97
GM	0,95	0,93	0,94	0,90	0,96	0,96
Orijinal	0,82	0,67	0,84	0,73	0,80	0,81
RDP	0,82	0,67	0,84	0,73	0,80	0,82
GM	0,83	0,67	0,83	0,72	0,81	0,82



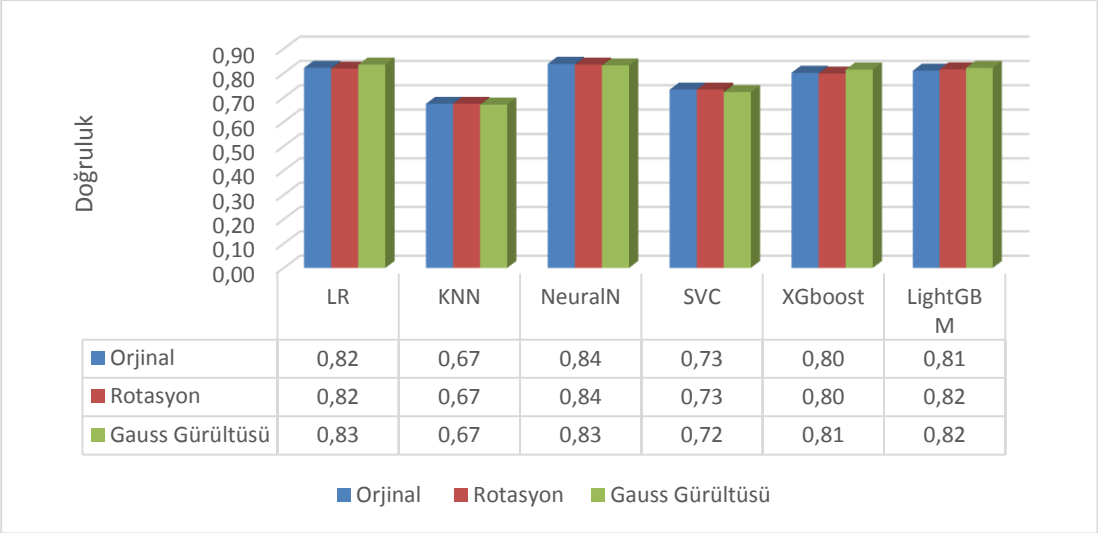
Şekil 5.18. Titanik veri setinin sınıflandırma yöntemlerinin doğruluk değerleri.



Şekil 5.19. Göğüs kanseri veri setinin sınıflandırma yöntemlerinin doğruluk değerleri.



Şekil 5.20. Diyabet veri setinin sınıflandırma yöntemlerinin doğruluk değerleri.



Şekil 5.21. Kalp krizi veri setinin sınıflandırma yöntemlerinin doğruluk değerleri.

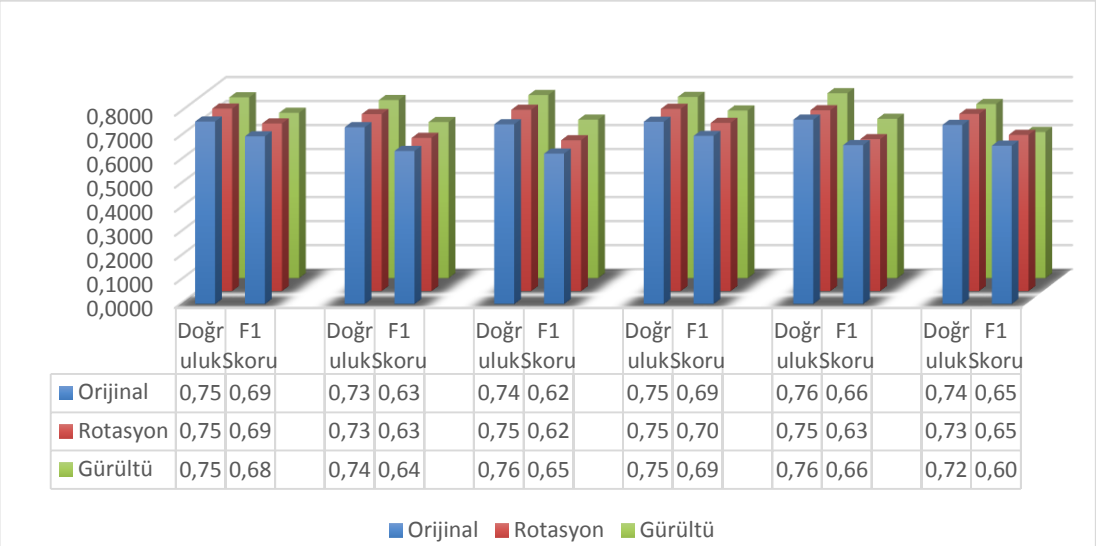
5.3. Uygulamanın Performans Göstergeleri

Rotasyona tabi tutulmuş ve rastgele gauss gürültü eklenmiş veri setlerinin doğruluk, kesinlik, duyarlılık ve f1 skorları hesaplanmış Tablo 5.4' de bir arada verilmiştir.

Tablo 5.4. Veri setlerinin sınıflandırma doğruluk, keskinlik, duyarlılık, F1 skor değerleri.

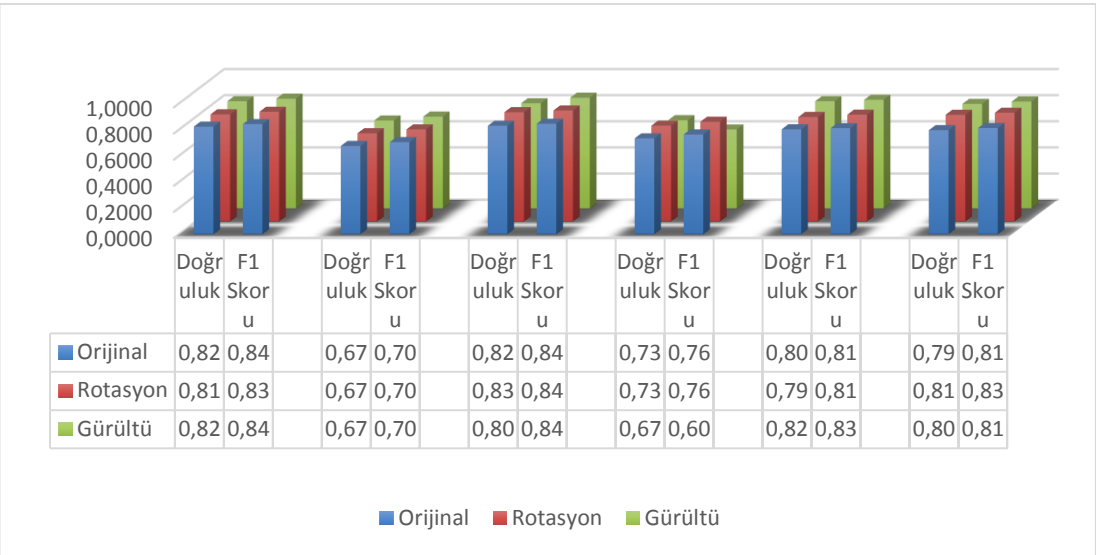
	Titanic			Diyabet			Göğüs Kanser			Kalp Krizi		
LR	**Orijinal**	**Rotasyon**	**GM**	**Orijinal**	**Rotasyon**	**GM**	**Orijinal**	**Rotasyon**	**GM**	**Orijinal**	**Rotasyon**	**GM**
Doğruluk	0,76	0,76	0,75	0,85	0,85	0,85	0,94	0,95	0,95	0,82	0,82	0,82
Kesinlik	0,72	0,72	0,71	0,79	0,79	0,79	0,94	0,95	0,95	0,82	0,81	0,82
Duyarlılık	0,70	0,70	0,69	0,77	0,77	0,76	0,91	0,92	0,92	0,88	0,88	0,88
F1 Skoru	0,70	0,70	0,69	0,78	0,78	0,77	0,92	0,93	0,93	0,84	0,84	0,84
KNN												
Doğruluk	0,74	0,74	0,74	0,84	0,84	0,83	0,93	0,93	0,93	0,67	0,67	0,67
Kesinlik	0,70	0,70	0,71	0,77	0,77	0,76	0,92	0,93	0,93	0,69	0,69	0,69
Duyarlılık	0,65	0,65	0,66	0,77	0,77	0,75	0,89	0,88	0,88	0,72	0,72	0,72
F1 Skoru	0,64	0,64	0,65	0,77	0,77	0,75	0,90	0,90	0,90	0,70	0,70	0,70
NN												
Doğruluk	0,75	0,75	0,76	0,86	0,86	0,85	0,93	0,94	0,94	0,83	0,84	0,81
Kesinlik	0,73	0,73	0,77	0,84	0,83	0,80	0,92	0,94	0,95	0,83	0,82	0,82
Duyarlılık	0,62	0,61	0,64	0,77	0,78	0,76	0,91	0,85	0,85	0,90	0,88	0,90
F1 Skoru	0,63	0,63	0,66	0,80	0,79	0,78	0,89	0,89	0,91	0,84	0,85	0,85
SVM												
Doğruluk	0,76	0,76	0,75	0,85	0,85	0,84	0,89	0,90	0,90	0,73	0,73	0,68
Kesinlik	0,71	0,71	0,71	0,78	0,78	0,78	0,92	0,97	0,97	0,72	0,72	0,76
Duyarlılık	0,70	0,70	0,70	0,77	0,77	0,76	0,82	0,75	0,75	0,87	0,87	0,60
F1 Skoru	0,70	0,70	0,70	0,78	0,78	0,77	0,85	0,84	0,84	0,76	0,76	0,61
Xgboost												
Doğruluk	0,77	0,75	0,77	0,88	0,88	0,86	0,96	0,96	0,96	0,80	0,80	0,82
Kesinlik	0,74	0,73	0,75	0,86	0,84	0,82	0,96	0,97	0,97	0,81	0,80	0,84
Duyarlılık	0,65	0,62	0,64	0,80	0,81	0,78	0,94	0,93	0,93	0,83	0,85	0,84
F1 Skoru	0,66	0,63	0,66	0,82	0,82	0,80	0,95	0,95	0,95	0,81	0,82	0,83
LightGBM												
Doğruluk	0,74	0,74	0,72	0,88	0,87	0,86	0,97	0,97	0,97	0,79	0,82	0,80
Kesinlik	0,68	0,68	0,66	0,85	0,82	0,82	0,97	0,96	0,96	0,80	0,82	0,81
Duyarlılık	0,68	0,65	0,62	0,81	0,81	0,78	0,96	0,95	0,95	0,85	0,85	0,84
F1 Skoru	0,66	0,65	0,61	0,83	0,81	0,80	0,96	0,96	0,96	0,81	0,83	0,82

Tüm veri setlerinin sırasıyla sınıflandırma algoritmalarının sonuçlarına ait doğruluk ve f1 skor değerlerinin karşılaştırılmasına imkân verecek şekilde çubuk diyagramları ile Şekil 5.22, Şekil 5.23, Şekil 5.24, Şekil 5.25' de gösterilmiştir.



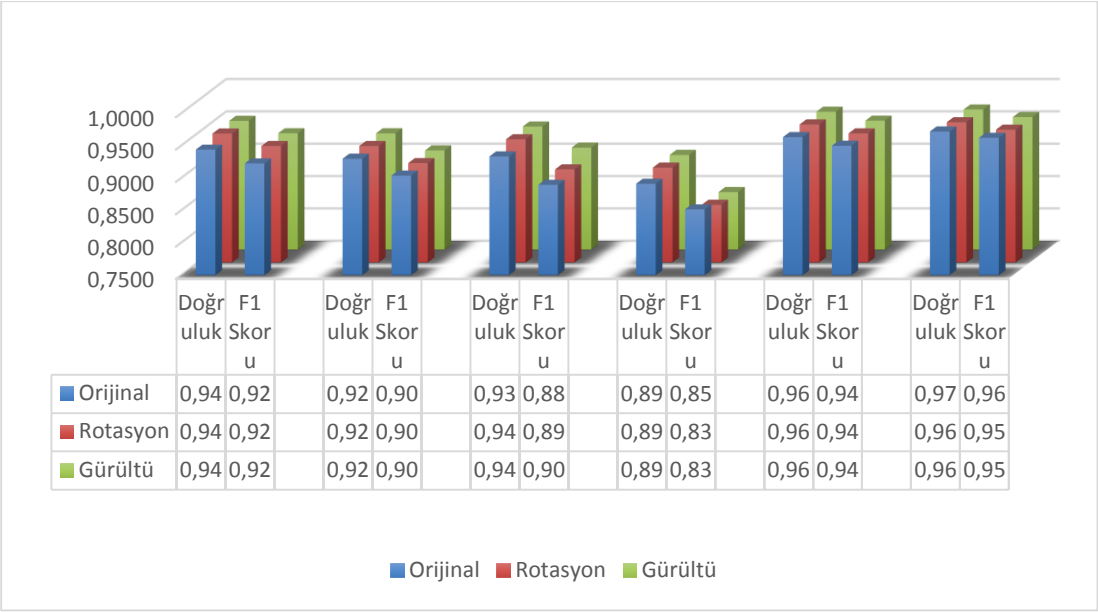
	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru
Orijinal	0,75	0,69	0,73	0,63	0,74	0,62	0,75	0,69	0,76	0,66	0,74	0,65
Rotasyon	0,75	0,69	0,73	0,63	0,75	0,62	0,75	0,70	0,75	0,63	0,73	0,65
Gürültü	0,75	0,68	0,74	0,64	0,76	0,65	0,75	0,69	0,76	0,66	0,72	0,60

Şekil 5.22. Titanik veri setinin sınıflandırma yöntemlerinin doğruluk ve F1 skor değerleri.



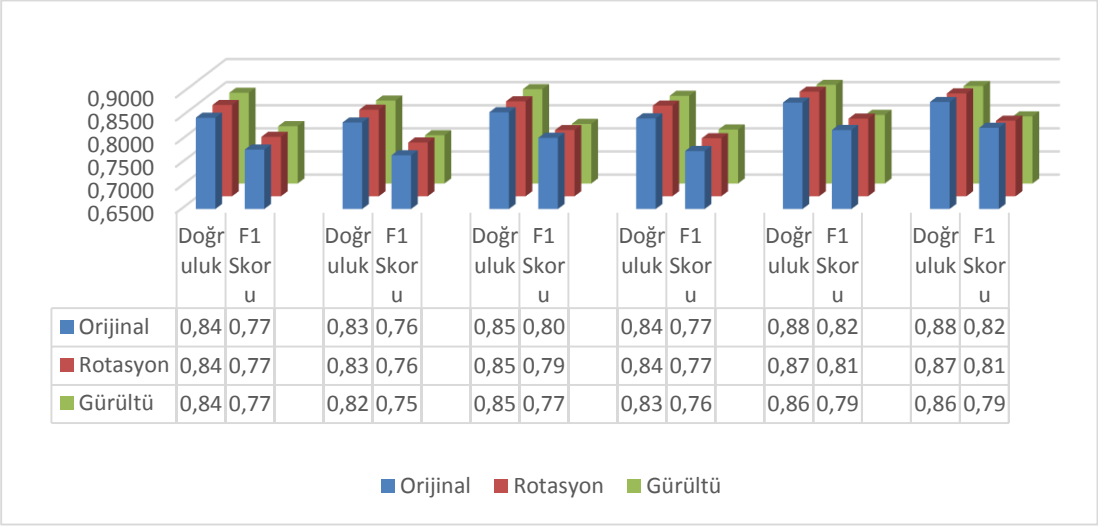
	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru
Orijinal	0,82	0,84	0,67	0,70	0,82	0,84	0,73	0,76	0,80	0,81	0,79	0,81
Rotasyon	0,81	0,83	0,67	0,70	0,83	0,84	0,73	0,76	0,79	0,81	0,81	0,83
Gürültü	0,82	0,84	0,67	0,70	0,80	0,84	0,67	0,60	0,82	0,83	0,80	0,81

Şekil 5.23. Göğüs kanseri veri setinin sınıflandırma yöntemlerinin doğruluk ve F1 skor değerleri.



	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru
Orijinal	0,94	0,92	0,92	0,90	0,93	0,88	0,89	0,85	0,96	0,94	0,97	0,96
Rotasyon	0,94	0,92	0,92	0,90	0,94	0,89	0,89	0,83	0,96	0,94	0,96	0,95
Gürültü	0,94	0,92	0,92	0,90	0,94	0,90	0,89	0,83	0,96	0,94	0,96	0,95

Şekil 5.24. Kalp krizi veri setinin sınıflandırma yöntemlerinin doğruluk ve F1 skor değerleri.



	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru	Doğruluk	F1 Skoru
Orijinal	0,84	0,77	0,83	0,76	0,85	0,80	0,84	0,77	0,88	0,82	0,88	0,82
Rotasyon	0,84	0,77	0,83	0,76	0,85	0,79	0,84	0,77	0,87	0,81	0,87	0,81
Gürültü	0,84	0,77	0,82	0,75	0,85	0,77	0,83	0,76	0,86	0,79	0,86	0,79

Şekil 5.25. Diyabet veri setinin sınıflandırma yöntemlerinin doğruluk ve F1 skor değerleri.

Birbirine bağlı birden çok ölçüm kümesinin karşılaştırılmasında kullanılan bir non-parametrik istatistiksel sıralama yöntemi olan Friedman sıralama testi ile ortalama rank değerleri hesaplanmış Tablo 5.5'te verilmiştir. Rank değerlerine göre doğruluk ve F1 skorları arasındaki farkın en az olduğu veri seti, sonuçlar nezdinde kalp krizi (FMR_D:53 ve FMR_F1:52,67) veri setidir.

Tablo 5.5. Veri setlerinin doğruluk ve F1 skorlarına göre ortalama rank değerleri.

Performans Metrikleri	Veri Seti	FMR
Doğruluk	Titanic	13,87
	Göğüs Kanseri	38
	Kalp Krizi	53
	Diyabet	17,13
F1 Skoru	Titanic	9
	Göğüs Kanseri	29,73
	Kalp Krizi	52,67
	Diyabet	30,6

Dört veri setinde en iyi sınıflandırmayı sonucunu ortaya koyan yöntemi belirlemek için Friedman sıralama testinden yararlanılmıştır. IBM SPPS programı yardımıyla sınıflandırma sonuçlarının ortalama rankları hesaplanmış olup sonuçlar Tablo 5.6'da ve Tablo 5.7'de verilmiştir.

Bu sonuçlara göre en yüksek rank değerine karşılık en iyi sınıflandırmayı yapan yöntemin Xgboost (FMR: 4,92) olduğu görülmektedir. En iyi sınıflandırmayı yapan diğer sınıflandırma yöntemleri sırasıyla Lojistik Regresyon, LightGBM, Yapay Sinir Ağları, Destek Vektör Makinaları, En Yakın Komşu'dur.

Sınıflandırma yöntemlerinin ortalama rank değerleri.

Sınıflandırma Yöntemleri	FMR
Lojistik Regresyon	4,25
K en Yakın Komşu	1,33
Yapay Sinir Ağı	4,08
Destek Vektör Makinaları	2,33
Xgboost	4,92
LightGBM	4,08

Tablo 5.6. Sınıflandırma yöntemlerinin ortalama rank değerleri ile kıyaslanması.

Veri Seti		LR	KNN	NN	SVM	Xgboost	LightGBM
Titanic	Orijinal	0,76	0,74	0,75	0,76	0,77	0,74
	Rotasyon	0,76	0,74	0,75	0,76	0,75	0,74
	Gauss	0,75	0,74	0,76	0,75	0,77	0,72
Diyabet	Orijinal	0,85	0,84	0,86	0,85	0,88	0,88
	Rotasyon	0,85	0,84	0,86	0,85	0,88	0,87
	Gauss	0,85	0,83	0,85	0,84	0,86	0,86
Göğüs Kanser	Orijinal	0,94	0,93	0,93	0,89	0,96	0,97
	Rotasyon	0,95	0,93	0,94	0,90	0,96	0,97
	Gauss	0,95	0,93	0,94	0,90	0,96	0,97
Kalp Krizi	Orijinal	0,82	0,67	0,83	0,73	0,80	0,79
	Rotasyon	0,82	0,67	0,84	0,73	0,80	0,82
	Gauss	0,82	0,67	0,81	0,68	0,82	0,80
	FMR	**4,25**	**1,33**	**4,08**	**2,33**	**4,92**	**4,08**

Sınıflandırma yöntemlerinin doğruluk değerlerinin minimum, maksimum, ortalama ve standart sapma, ortanca değerleri Tablo 5.8' de verilmiştir. Bu sonuçlara bakarak da en iyi sınıflandırma yönteminin Friedman Rank Mean değeri 4,92 olan Xgboost olduğu görülmektedir.

Tablo 5.7. Sınıflandırma Yöntemlerinin İstatistiksel Metriklerinin Tablosu.

	N	**Ortalama**	**Std**	**Min**	**Maks**	**Yüzdelik Oranlar**		
						1.Kartil	Ortanca	3.Kartil
LR	12	0,84%	0,07%	0,75%	0,95%	0,77%	0,83%	0,92%
KNN	12	0,79%	0,10%	0,67%	0,93%	0,69%	0,79%	0,91%
NN	12	0,84%	0,07%	0,75%	0,94%	0,77%	0,84%	0,91%
SVM	12	0,80%	0,08%	0,68%	0,90%	0,74%	0,80%	0,88%
Xgboost	12	0,85%	0,08%	0,75%	0,96%	0,78%	0,84%	0,94%
LightGBM	12	0,84%	0,09%	0,72%	0,97%	0,76%	0,84%	0,95%

Sınıflandırma doğruluğu, üç farklı sonuçla karşılaştırılmıştır. Orijinal verinin sınıflandırma doğruluğu, 110 derece döndürülen veri kümesinin ve sonrasında eklenen rastgele gauss gürültüsünün sınıflandırma yöntemleri ile ölçülen doğruluk değerleri Tablo 5.9' da verilmişti

Tablo 5.8. Orjinal ve pertürbe edilmiş verilere ait sınıflandırma yöntemlerinin doğruluk değerleri.

Veri Seti		LR		KNN		NN		SVM		XGboost		LightGBM	
		Doğruluk	**Std**	**Doğruluk**	**Std**	**Doğruluk**	**Std**	**Doğruluk**	**Std**	**Doğruluk**	**Std**	**Doğruluk**	**Std**
	Orijinal	0,76	0,09	0,74	0,10	0,76	0,11	0,76	0,08	0,77	0,09	0,74	0,09
Titanic	Rotasyon	0,76	0.09	0,74	0,10	0,76	0,11	0,76	0,08	0,75	0,10	0,74	0,08
	Gauss	0,75	0.09	0,73	0,10	0,76	0,09	0,75	0,07	0,76	0,09	0,75	0,09
	Orijinal	0,85	0.03	0,84	0.03	0,87	0,03	0,85	0,04	0,88	0,03	0,88	0,04
Diyabet	Rotasyon	0,85	0.03	0,84	0,03	0,86	0,03	0,85	0,04	0,88	0,03	0,87	0,04
	Gauss	0,85	0.03	0,83	0,04	0,85	0,03	0,84	0,04	0,87	0,02	0,86	0,03
	Orijinal	0,94	0,02	0,93	0,03	0,94	0,02	0,89	0,06	0,96	0,03	0,97	0,03
Kanser	Rotasyon	0,95	0,02	0,93	0,03	0,93	0,02	0,90	0,05	0,96	0,03	0,97	0,02
	Gauss	0,95	0.02	0,93	0,03	0,94	0,02	0,90	0,04	0,96	0,03	0,96	0,02
	Orijinal	0,82	0.07	0,67	0,09	0,84	0,07	0,73	0,12	0,80	0,01	0,81	0,07
Kalp	Rotasyon	0,82	0.06	0,67	0,09	0,84	0,07	0,73	0,07	0,80	0,07	0,82	0,04
	Gauss	0,83	0.06	0,67	0,09	0,83	0,08	0,72	0,07	0,81	0,07	0,82	0,06

Doğruluk değerlerinin minimum, maksimum, ortalama ve standart sapma, ortanca değerleri Tablo 5.10.' da verilmiştir. Orijinal veri kümelerine ait değerlerin rotasyon ve gürültüye maruz bırakılan veri kümelerindeki değerlerle kıyaslandığında nerede ise eşit olduğu görülmektedir.

Tablo 5.9. Orjinal ve pertürbe edilmiş veri setlerinin istatistiksel metriklerinin tablosu.

	N	**Ortalama**	**StandartS**	**Min**	**Maks**	**Yüzdelik Oranlar**		
						1.Kartil	Ortanca	3.Kartil
Orijinal	24	83%	8%	67%	97%	76%	83%	89%
Rotasyon	24	83%	8%	67%	97%	76%	84%	89%
Gauss	24	83%	9%	67%	97%	76%	83%	89%

6. SONUÇ VE ÖNERİLER

Bu çalışmada önerilen yaklaşım sayısal hassas verilerin gizliğini korumaya yönelik iki aşamalı bir pertürbasyon tekniğidir. Veri setlerinden seçilen hassas sayısal ikili veri öznitelikleri rastgele bir açıda koordinat düzleminde döndürülmüş, sonrasında bu özniteliklere ortalamaları ve standart sapmaları girdi olacak şekilde rastgele gürültü eklenmiştir. Öncesinde orijinal veri setleri ve sonrasında pertürbe edilmiş veri setleri bilgi kaybı, doğruluk kaybı kriterleri gözetilerek altı farklı sınıflandırma yöntemi ile karşılaştırılmıştır.

Seçilen altı sınıflandırma yöntemi arasında Xgboost, Lojistik Regresyon, LightGMB ve Yapay Sinir Ağları yöntemlerinin pertürbasyon sonrası bilgi kaybının en düşük seviye kaldığı Friedman sıralama testiyle ölçülerek görülmüştür. Kullanılan veri setleri içerisinde bilgi kaybının en az olduğu ve en iyi sınıflandırma doğruluğu derecesine sahip yöntem Gradyanı Artırılan Karar Ağaçları olduğu görülmüştür (Xgboost).

Gelecekteki çalışmalarda veri kümesinin hacmini genişletilebilir, daha farklı sınıflandırma yöntemleri tercih edilebilir. İki boyutlu döndürme işlemi veri seti içerisinde yer alan hassas tüm ikili özniteliklere uygulanabilir. Laplace, eksponansiyel, gauss gürültü mekanizmaları rotasyon pertürbasyonu sonrası eklenerek kendi aralarında veri gizliliği, bilgi kaybı atak direnci mertebesinde kıyaslanabilir.

KAYNAKLAR

Aggarwal, C. C. (2015). Privacy-preserving data mining. In Data Mining (pp.663–693). Springer. doi:https://doi.org/10.1007/978-3-319-14142-8.

Aggarwal, C. C., & Yu, P. S. (2004). A condensation approach to privacy preserving data mining. In EDBT (pp.183–199). Springer volume 4. doi:https://doi.org/10.1007/ 978-3-540-24741-8_12.

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In ACM Sigmod Record (pp. 439–450). ACM volume 29. doi:https://doi.org/10.1145/335191.335438.

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy pre-serving data mining. SpringerPlus, 4, 694. doi:https://doi.org/10.1186/s40064-015-1481.

Aloysius, J. A., Hoehle, H., Goodarzi, S., & Venkatesh, V. (2018). Big data initiatives in retail environments: Linking service process perceptions to shopping outcomes. Annals of operations research, 270 , (pp.25–51). doi:https://doi.org/10.1007/s10479-016-2276-3.

Aydındağ Bayrak, E. , Kırcı, P. , Ensari, T. , Seven, E. & Dağtekin, M. (2022). Göğüs Kanseri Verileri Üzerinde Makine Öğrenmesi Yöntemlerinin Uygulanması. Journal of Intelligent Systems: Theory and Applications, 5 (1) , 35-41. doi: 10.38016/jista.966517

Bettini, C., & Riboni, D. (2015). Privacy protection in pervasive systems: State of the art and technical challenges. Pervasive and Mobile Computing, 17 , (pp.159–174). doi:https://doi.org/10. 1016/j.pmcj.2014.09.010.

Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2016). A threat to friendship privacy in facebook. In International Conference on Availability, Reliability, and Security (pp. 96–105). Springer. doi:https://doi.org/10.1007/978-3-319-45507-5_7.

Canbay, Y. & Sağıroğlu, Ş. (2020). Derin Öğrenmede Diferansiyel Mahremiyet. Uluslararası Bilgi Güvenliği Mühendisliği Dergisi, 6 (1) , (pp.1-16). DOI: 10.18640/ubgmd.750310

Capraro, V., & Perc, M. (2018). Grand challenges in social physics: In pursuit of moral behavior. Frontiers in Physics, 6 , (pp.107). doi: https://doi.org/10.3389/fphy.2018.00107.

Chamikara, M. A. P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive and Mobile Computing, 48 , (pp.1–19). doi:https://doi.org/10.1016/j.pmcj.2018.05.003.

Chen, K., & Liu, L. (2005). A random rotation perturbation approach to privacy preserving data classification. The Ohio Center of Excellence in Knowledge-Enabled Computing, . doi:https://corescholar.libraries.wright.edu/knoesis/916/.

Chen, K., & Liu, L. (2011). Geometric data perturbation for privacy preserving outsourced data mining. Knowledge and Information Systems, 29, (pp.657–695). doi:https://doi.org/10.1007/s10115-010-0362-4.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. ACM Sigkdd Explorations Newsletter, 4 , (pp.28–34). doi:https: //doi.org/10.1145/772862.772867.

Cuzzocrea, A. (2015). Privacy-preserving big data management: The case of olap. Big Data: Algorithms, Analytics, and Applications, (pp.301–326). doi: https://books.google.com.au/books?isbn=1482240564.

Dwork, C., Roth, A. et al. (2014). The algorithmic foundations of diff erential privacy. Foundations and Trends in Theoretical Computer Science, 9 , (pp.211–407). doi:http://dx.doi.org/10.1561/ 0400000042.

E. & Özcan, E. (2022). Makine Öğrenme Teknikleri Kullanılarak Kükürt Giderme İşleminde Kullanılan Malzeme Miktarının Tahmini. Journal of Intelligent Systems: Theory and Applications, 5 (1) , 57-63. DOI: 10.38016/jista.993853

Erlingsson, ´U., Pihur, V., & Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC conference on computer and communications security (pp.1054–1067). ACM. doi:https://doi.org/10.1145/2660267. 2660348.

Eyüpoğlu, C., Aydın, M.A., Zaim, A.H. and Sertbaş, A., 2018, An Efficient Big Data Anonymization Algorithm Based on Chaos and Perturbation Techniques, Entropy, 20 (5), 373, 1-18 (Science Citation Index Expanded).

Eyüpoğlu, C., Aydın, M.A., Sertbaş, A., Zaim, A.H. and Öneş, O., 2017, Preserving Individual Privacy in Big Data, International Journal of Informatics Technologies, 10 (2), ss.177-184 (TÜBİTAK ULAKBİM TR Dizin).

Gai, K., Qiu, M., Zhao, H., & Xiong, J. (2016). Privacy-aware adaptive data encryption strategy of big data in cloud computing. In Cyber Security and Coud Computing (CSCloud), 2016 IEEE 3rd International Conference on (pp.273–278).IEEE. doi:http://doi.ieeecomputersociety. org/10.1109/CSCloud.2016.52.

H. Çavşi Zaim, E. Yolaçan ve E. Gülbandılar , "Banka Ödemelerinde Dolandırıcılığın Çizge Madenciliği ve Makine Öğrenimi Algoritmalarıyla Tespiti", Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, c. 12, sayı. 4, ss. 615-625, Eyl. 2021, doi:10.24012/dumf.1002110

Hasan, A., Jiang, Q., Luo, J., Li, C., & Chen, L. (2016). An eff ective value swapping method for privacy preserving data publishing. Security and Communication Networks, 9 , (pp.3219–3228). doi:https://doi.org/10.1002/sec.1527.

Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O., Mous-said, M., Johansson, A., Krause, J., Schutte, S. et al. (2015). Saving human lives: What complex-ity science and information systems can contribute. Journal of statistical physics, 158, (pp.735–781). doi:https://doi.org/10.1007/s10955-014-1024-9.

Howell, D. C. (2016). Fundamental statistics for the behavioral sciences. Cengage Learning. doi:https://books.google.com.au/books?isbn=1305652975.

Jalili, M., & Perc, M. (2017). Information cascades in complex networks. Journal of Complex Networks, 5 , (pp.665–693). doi:https://doi.org/10.1093/comnet/cnx019.

Jones, H. (2012). Computer Graphics through Key Mathematics. Springer London : Imprint: Springer. doi: https://books.google.com.au/books?id=f7gPBwAAQBAJ.

Kabir, W., Ahmad, M. O., & Swamy, M. (2015). A novel normalization technique for multimodal biometric systems. In Circuits and Systems (MWSCAS), 2015 IEEE 58th International Midwest Symposium on (pp.1–4). IEEE. doi:https://doi.org/10.1109/MWSCAS.2015.7282214.

Kairouz, P., Oh, S., & Viswanath, P. (2014). Extremal mechanisms for local diff erential privacy. In Advances in neural information processing systems (pp.2879–2887). doi: http://papers. nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.

Kargupta, Hillol & Datta, Souptik & Wang, Q. & Sivakumar, Krishnamoorthy. (2003). On the privacy preserving properties of random data perturbation techniques. Proceedings - IEEE International Conference on Data Mining, ICDM.(pp.99-106). 10.1109/ICDM.2003.1250908.

Kerschbaum, F., & H¨arterich, M. (2017). Searchable encryption to reduce encryption degradation in adjustably encrypted databases. In IFIP Annual Conference on Data and Applications Security and Privacy (pp.325–336). Springer. doi:https://doi.org/10.1007/978-3-319-61176-1_18.

Kieseberg, P., & Weippl, E. (2018). Security challenges in cyber-physical production systems. In International Conference on Software Quality (pp.3–16). Springer. doi:https://doi.org/10. 1007/978-3-319-71440-0_1.

Li, P., Li, J., Huang, Z., Gao, C.-Z., Chen, W.-B., & Chen, K. (2017). Privacy-preserving outsourced classification in cloud computing. Cluster Computing, (pp.1–10). doi:https://doi. org/10.1007/s10586-017-0849-9.

Liu, K., Kargupta, H., & Ryan, J. (2006). Random projection-based multiplicative data pertur-bation for privacy preserving distributed data mining. IEEE Transactions on knowledge and Data Engineering, 18 , (pp.92–106). doi:https://doi.org/10.1109/TKDE.2006.14.

Li G, Xue R (2018) A new privacy-preserving data mining method using non-negative matrix factorization and singular value decomposition. Wireless Personal Commun 102 (pp.1799-1808). doi:https://doi.org/10.1007/s11277-017-5237-5

Li, C., Palanisamy, B., Reversible spatio-temporal perturbation for protecting location privacy, Computer Communications, Volume 135, 2019, (pp.16-27). ISSN 0140-3664,https://doi.org/10.1016/j.comcom.2018.12.003.

Liu C, Chen S, Zhou S et al (2019) A novel privacy preserving method for data publication. Inf Sci 501(pp.421–435). doi: https://doi.org/10.1016/j.ins.2019.06.022

Lindell, Y. and Pinkas, B., Privacy preserving data mining. In Advances in Cryptology – CRYPTO ’00, volume 1880 of Lecture Notes in Computer Science, (pp.36–54). Springer-Verlag, 2000.

Manogaran, G., Thota, C., Lopez, D., Vijayakumar, V., Abbas, K. M., & Sundarsekar, R.(2017). Big data knowledge system in healthcare. In Internet of things and big data technolo-gies for next generation healthcare (pp.133-157). Springer. doi:https://doi.org/10.1007/ 978-3-319-49736-5_7.

Maruskin, J. (2012). Essential Linear Algebra. Solar Crest Publishing, LLC. doi: https: //books.google.com.au/books?id=aOF3-hx3u1kC.

Muralidhar, K., Parsa, R., & Sarathy, R. (1999). A general additive data perturbation method for database security. management science, 45, (pp.1399-1415). doi:https://doi.org/10.1287/mnsc. 45.10.1399.

Paeth, A. W. (2014). Graphics Gems V (Macintosh Version). Academic Press. doi: https: //books.google.com.au/books?isbn=1483296695.

Park, K.-j., & Ryou, H.-b. (2003). Anomaly detection scheme using data mining in mobile environment. Computational Science and Its Applications ICCSA, (pp.978-978). doi:https: //doi.org/10.1007/3-540-44843-8_3.

Qin, Z., Yang, Y., Yu, T., Khalil, I., Xiao, X., & Ren, K. (2016). Heavy hitter estimation over set-valued data with local privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp.192-203). ACM. doi:https://doi.org/10. 1145/2976749.2978409.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In Security and Privacy (SP), 2017 IEEE Symposium on (pp.3-18). IEEE. doi:https://doi.org/10.1109/SP.2017.41.

Soria-Comas, J., & Domingo-Ferrer, J. (2016). Big data privacy: challenges to privacy prin-ciples and models. Data Science and Engineering, 1 ,(pp.21-28). doi:https://doi.org/10.1007/s41019-015-0001.

Steel, E., & Fowler, G. (2010). Facebook in privacy breach. The Wall Street Journal, 18 . doi: https://www.wsj.com/articles/SB10001424052702304772804575558484075236968.

Torra, V. (2017). Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer. doi:https://doi.org/10.1007/978-3-319-57358-8.

Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. In Handbook of Big Data Technologies (pp.851-895). Springer. doi:https://doi.org/10.1007/978-3-319-49340-4_25.

Wei, Z., Wu, Y., Yang, Y., Yan, Z., Pei, Q., Xie, Y., & Weng, J. (2018). Autoprivacy: automatic privacy protection and tagging suggestion for mobile social photo. Computers & Security, . doi:https://doi.org/10.1016/j.cose.2017.12.002.

Wen, Y., Liu, J., Dou, W., Xu, X., Cao, B., & Chen, J. (2018). Scheduling workï¬‚ows with privacy protection constraints for big data applications on cloud. Future Generation Computer Systems,. doi: https://doi.org/10.1016/j.future.2018.03.028.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. doi: https://books.google.com.au/books?isbn= 0128043571.

Wong, R. C.-W., Fu, A. W.-C., Wang, K., & Pei, J. (2007). Minimality attack in privacy preserving data publishing. In Proceedings of the 33rd international conference on Very large data bases (pp.543-554). VLDB Endowment. doi: https://dl.acm.org/citation.cfm?id=1325914.

ÖZGEÇMİŞ

Ad-Soyad : İlker İLTER

ÖĞRENİM DURUMU:

- **Lisans** **:** 1998, Uludağ Üniversitesi, Mühendislik Fakültesi, Elektronik Mühendisliği
- **Yükseklisans** **:** 2021, Sakarya Üniversitesi, Kalite Yönetimi Anabilim Dalı, Kalite Yönetimi
- **Yükseklisans** **:** 2022, Sakarya Üniversitesi, İşletme Anabilim Dalı, İşletme
- **Yükseklisans** **:** 2023, Sakarya Üniversitesi, Endüstri Mühendisliği Anabilim Dalı, Mühendislik Yönetimi

MESLEKİ DENEYİM VE ÖDÜLLER:

- 2001-2006 yılları arasında bilişim sektöründe sistem mühendisi olarak çalıştı.
- 2006 yılından bu yana savunma sanayi sektöründe sistem üretim mühendisi olarak görev yapmaktadır.

TEZDEN TÜRETİLEN ESERLER:

- Ilter, I., Turgay, S., 2023. Privacy Enhancement with Perturbation Method for Multidimensional Grid, *Journal of Artificial Intelligence*, 6(4), 2371-8412. (pp.31-39). doi: http://dx.doi.org/10.23977/jaip.2023.060405.
- Turgay, S., Ilter, I., 2023. Perturbation Methods for Protecting Data Privacy: A Review of Techniques and Applications, *Automation and Machine Learning*, 4(2), 2516-5003. (pp.31-41). doi: http://dx.doi.org/10.23977/autml.2023.040205.