T.C.

BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

OTOKODLAYICI TABANLI GİZLİLİĞİ KORUYAN ORTAK FİLTRELEME

YÜKSEK LİSANS TEZİ

ELİF TUĞÇE AÇIL

TEZ DANIŞMANI

DR. ÖĞR. ÜYESİ ALPER YARGIÇ

BİLECİK, 2023

10580846

T.C.

BİLECİK ŞEYH EDEBALİ ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

OTOKODLAYICI TABANLI GİZLİLİĞİ KORUYAN ORTAK FİLTRELEME

YÜKSEK LİSANS TEZİ

ELİF TUĞÇE AÇIL

TEZ DANIŞMANI

DR. ÖĞR. ÜYESİ ALPER YARGIÇ

BİLECİK, 2023

10580846

BEYAN

"**Otokodlayıcı Tabanlı Gizliliği Koruyan Ortak Filtreleme**" adlı yüksek lisans tezinin hazırlık ve yazımı sırasında bilimsel araştırma ve etik kurallarına uyduğumu, başkalarının eserlerinden yararlandığım bölümlerde bilimsel kurallara uygun olarak atıfta bulunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, tezin herhangi bir kısmının Bilecik Şeyh Edebali Üniversitesi veya başka bir üniversitede başka bir tez çalışması olarak sunulmadığını, aksinin tespit edileceği muhtemel durumlarda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

<table>
<tr><td colspan="4">Bu çalışmanın,
Bilimsel Araştırma Projeleri (BAP), TÜBİTAK veya benzeri kuruluşlarca desteklenmesi durumunda; projenin ve destekleyen kurumun adı proje numarası ile birlikte, ETİK KURUL onayı alınması durumunda ise ETİK KURUL tarih karar ve sayı bilgilerinin beyan edilmesi gerekmektedir.</td></tr>
<tr><td>DESTEK ALINMIŞTIR</td><td></td><td>DESTEK ALINMAMIŞTIR</td><td>X</td></tr>
<tr><td colspan="4">Destek alındı ise;</td></tr>
<tr><td colspan="4">Destekleyen kurum;</td></tr>
<tr><td colspan="2">Desteğin Türü</td><td colspan="2">Proje Numarası</td></tr>
<tr><td colspan="2">1- BAP (Bilimsel Araştırma Projesi)</td><td colspan="2" rowspan="3"></td></tr>
<tr><td colspan="2">2- TÜBİTAK</td></tr>
<tr><td colspan="2">Diğer;……</td></tr>
<tr><td colspan="4">ETİK KURUL onayı var ise;</td></tr>
<tr><td colspan="2">ETİK KURUL karar tarih/sayı:</td><td colspan="2">……/……..</td></tr>
</table>

Elif Tuğçe AÇIL

09.10.2023

ÖNSÖZ

Çalışmalarım boyunca yardımını ve desteğini esirgemeyen, değerli bilgilerini benimle paylaşarak bana yol gösteren danışman hocam Dr. Öğr. Üyesi Alper YARGIÇ'a sonsuz teşekkürlerimi sunarım.

Çalışmalarıma katkı sağlayan sayın hocam Dr. Öğr. Üyesi Zeynep BATMAZ'a vakit ayırdığı ve bilgi birikimini paylaştığı için teşekkürü borç bilirim.

Tez çalışmalarımın değerlendirmesindeki kıymetli tavsiyeleri ve bana bugüne dek sağladığı tüm katkıları için sayın hocam Doç. Dr. Emre DANDIL'a teşekkürlerimi sunarım.

Son olarak maddi ve manevi desteğini her zaman hissettiren değerli aileme tüm emekleri için sonsuz teşekkürlerimi sunarım.

Elif Tuğçe AÇIL

2023

ÖZET

OTOKODLAYICI TABANLI GİZLİLİĞİ KORUYAN ORTAK FİLTRELEME

Gizliliği koruyan ortak filtreleme sistemleri, kullanıcıların mahremiyetlerini ihlal etmeden onlara kişiselleştirilmiş öneriler sunan etkili yaklaşımlardır. Ancak rastgele karıştırma tekniklerine dayalı veri gizleme yaklaşımları gerçek kullanıcı verisi üzerinde bozulmaya neden olduğundan sistemin öneri üretme başarısını olumsuz yönde etkilemektedir. Özellikle yüksek gizlilik seviyelerine ulaşmak için yapılan veri gizleme işlemlerinde öneri doğruluğunda ortaya çıkan kayıplar oldukça fazladır.

Bu çalışmada, rastgele karıştırma teknikleri ile maskelenen gerçek kullanıcı oy değerlerinin öneri doğruluğu kayıplarını hafifletmek için otokodlayıcı tabanlı bir öneri üretme yaklaşımı kullanılmıştır. Book-Crossing referans alınarak üretilen veri seti üzerinde rastgele karıştırma teknikleri kullanılarak farklı gizlilik seviyelerinde üretilen maskelenmiş veri setlerinin öneri doğrulukları geleneksel hafıza tabanlı komşuluk algoritması ve otokodlayıcı tabanlı öneri üretme yaklaşımları ile analiz edilmiştir. Otokodlayıcı tabanlı öneri üretme sistemi çeşitli gizli katman sayıları (2, 3 ve 4) ve aktivasyon fonksiyonları (tanh, elu, selu ve lineer) ile test edilmiştir. Deneysel çalışmalar sonucunda, rastgele karıştırma tekniği kullanılarak maskelenmiş veri seti üzerinde otokodlayıcı tabanlı öneri üretme yaklaşımının geleneksel komşuluk tabanlı yaklaşımlarına göre tahmin doğruluğunu önemli ölçüde arttırdığı gösterilmiştir. Değişken gizlilik seviyelerine göre ortalama mutlak hata değerleri incelendiğinde, geleneksel komşuluk tabanlı ortak filtreleme yaklaşımda sırasıyla en düşük ve en yüksek gizlilik seviyelerindeki hata değerleri 1,395 ve 2,249 iken, otokodlayıcı tabanlı ortak filtreleme yaklaşımında bu değerler 1,208 ve 1,313 olarak elde edilmiştir. Sonuç olarak, otokodlayıcı tabanlı ortak filtreleme sistemi yükselen gizlilik seviyelerine bağlı olarak veri setinde ortaya çıkan bozulmaları daha iyi tolere edebilmekte, kullanıcıya yüksek mahremiyet seviyeleri sağlandığında da yüksek doğrulukta öneriler üretebilmektedir.

Anahtar kelimeler: Ortak Filtreleme, Otokodlayıcı, Gizliliği Korunan Ortak Filtreleme, Öneri Sistemleri

ABSTRACT

AUTOENCODER-BASED PRIVACY-PRESERVING COLLABORATIVE FILTERING

Privacy-preserving collaborative filtering systems are effective approaches that provide personalized recommendations to users without violating their privacy. However, data disguising approaches based on randomized perturbation techniques cause corruption in genuine user data and negatively affect the system's success in generating recommendations. Especially in data disguising processes to achieve high privacy levels, the losses in recommendation accuracy are quite high.

In this study, an autoencoder-based recommendation generation approach was used to alleviate the recommendation accuracy losses of genuine user ratings disguised by randomized perturbation techniques. The recommendation accuracies of disguised data sets produced at different privacy levels using randomized perturbation techniques on the data set produced with Book-Crossing as a reference were analyzed with traditional memory-based neighborhood algorithms and autoencoder-based recommendation generation approaches. The autoencoder-based recommendation generation system was tested with various hidden layer numbers (2, 3, and 4) and activation functions (tanh, elu, selu, and linear). As a result of experimental studies, it was shown that the autoencoder-based recommendation generation approach on the disguised data set using a randomized perturbation technique significantly increased the recommendation accuracy compared to traditional neighborhood-based approaches. When the mean absolute error values were examined via varying privacy levels, the error values at the lowest and highest privacy levels were 1.395 and 2.249, respectively, in the traditional neighborhood-based collaborative filtering approach, while these values were 1.208 and 1.313 in the autoencoder-based collaborative filtering approach. In conclusion, the autoencoder-based collaborative filtering system can better tolerate the distortions that occur in the data set due to increasing privacy levels and can produce high-accuracy recommendations when high privacy levels are provided to the user.

Keywords: Collaborative Filtering, Autoencoder, Privacy-Preserving Collaborative Filtering, Recommendation Systems.

İÇİNDEKİLER

Sayfa

KISALTMALAR VE SİMGELER LİSTESİ

$\boldsymbol{\mu}$ **:** Ortalama

$\boldsymbol{\sigma}$ **:** Sigma

a **:** Aktif Kullanıcı

BX : Book-Crossing Veri Seti

BXM : BX Veri Setinin Maskelenmiş Alt Kümesi

BXN : BX Veri Setinin Normalize Edilmiş Alt Kümesi

DÖ : Derin Öğrenme

elu : Üstel Doğrusal Birim (Exponential Linear Unit)

G : Gerçek Kullanıcı Derecelendirme Vektörü

GKOF : Gizliliği Koruyan Ortak Filtreleme

MAE : Ortalama Mutlak Hata (Mean Absolute Error)

N : Normal Dağılım

OF : Ortak Filtreleme

ÖS : Öneri Sistemleri

PK : Pearson Korelasyonu

q **:** Hedef Ürün

R : Rastgele Sayı Vektörü

RKT : Rastgele Karıştırma Tekniği

RMSE : Kök Ortalama Kare Hata (Root Mean Square Error)

selu : Ölçeklendirilmiş Üstel Doğrusal Birim (Scaled Exponential Linear Unit)

tanh : Hiperbolik Tanjant Aktivasyon Fonksiyonu

TABLOLAR LİSTESİ

Sayfa

ŞEKİLLER LİSTESİ

Sayfa

1. GİRİŞ

Teknolojinin gelişmesi ve internet tabanlı servislerin kullanımının yaygınlaşması ile insanlar geniş bir dijital içerik yelpazesine erişim sağlamış ve günlük rutinlerini çevrimiçi servisler üzerinden gerçekleştirmeye başlamışlardır. Çevrim içi alışveriş siteleri, müzik-video platformları ve sosyal medya mecraları gibi platformlar kullanıcılarına farklı kategorilerde ürünler/hizmetler ve içerikler sunmaktadır. Bu platformlardaki içeriklerin ve kullanıcı sayısının her geçen gün artmasına bağlı olarak, işlenmesi ve depolanması gereken veri miktarı da her geçen gün artmaktadır. Bu durum, aşırı bilgi yükleme problemi olarak adlandırılan soruna neden olmakta ve kullanıcıların karar verme sürecini zorlaştırmaktadır (Ricci, Rokach ve Shapira, 2015). Öneri Sistemleri (ÖS) bu problemin üstesinden gelmek için internet servisleri tarafından sıklıkla kullanılmaktadır. Bu sistemler kullanıcıların geçmiş tercihlerini, satın alma alışkanlıklarını ve ürün/hizmet beğeni davranışlarını analiz edip, ilgilenebilecekleri ürün/hizmet veya içerikleri tahmin ederek kullanıcılara öneri olarak sunarlar. ÖS'nin sağlamış olduğu kişiselleştirilmiş deneyim imkânı hem kullanıcıya hem de firmalara pek çok avantaj sağlamaktadır. Örneğin sistem tarafından sunulan öneriler sayesinde kullanıcı ilgi alanlarına uygun ürünlere çok daha az çabayla ulaşabilmekte, zamandan ve enerjiden tasarruf edebilmektedir. Kullanıcılara sunulan önerilerin pek çok farklı kategori ve çeşitte olması, kullanıcıya daha kolay şekilde yeni deneyimlere ulaşma ve yeni ürünler keşfetme konusunda imkân sağlamaktadır. ÖS bu gibi avantajları ile kullanıcının alışveriş memnuniyetini yükselterek, firmalara kullanıcı sadakati sağlamaktadır. Ayrıca sistemin kullanıcıya önerdiği yeni ürünler ile firmalar satış hacmini artırarak kazancını yükseltebilmekte, kullanıcının alışveriş alışkanlıklarına ve geçmiş tercihlerine göre kişiselleştirilmiş promosyonlar oluşturabilmektedir (Weinsberg, 2012).

ÖS, hem firmalara hem de kullanıcılara birçok katkı sağlamakla birlikte, hala gelişmeye ve iyileştirilmeye ihtiyaç duyulan yönleri bulunmaktadır. Özellikle kullanıcı veri gizliliği ve öneri doğruluğu bu konulardan başlıcalarıdır. Bu çalışmada, kullanıcıların geçmişte deneyimledikleri ürünler/hizmetler için firmalarla paylaştıkları kişisel verilerinin korunmasına ve gizlenen kullanıcı verileri üzerinden daha doğru önerilerin üretilmesine odaklanılmıştır. Verinin gizlenmesi için sıkça başvurulan Rastgele Karıştırma Tekniği (RKT) ile gerçek kullanıcı derecelendirme değerleri maskelenmiştir. Maskeleme işleminden dolayı ortaya çıkan öneri doğruluğundaki kayıpları hafifletmek için, literatürdeki geleneksel komşuluk tabanlı tahmin üretme yaklaşımlarına göre daha yüksek bir doğruluk oranı sunan otokodlayıcı tabanlı

bir Derin Öğrenme (DÖ) yaklaşımı kullanılmış ve otokodlayıcıların maskelenmiş veri seti üzerindeki tahmin doğrulukları incelenmiştir.

1.1. Ortak Filtreleme

Bir ÖS temel olarak; bir ürün/hizmet için tahmin üretme ve üretilen tahminlere göre kişiselleştirilmiş ürün/hizmet listeleri oluşturma fonksiyonlarını sağlar (Sarwar vd., 2001). Tahmin üretme teknikleri açısından ÖS'de yaygın olarak kullanılan tekniklerden biri Ortak Filtreleme (OF) tekniğidir (Adomavicius ve Tuzhilin, 2005).

OF sistemleri, kullanıcıların ilgilenebilecekleri ancak daha önce deneyimlemedikleri ürünler/hizmetler hakkında kullanıcılarına tahminler sunan etkili yazılım araçlarıdır. Kullanıcılarının daha önce deneyimledikleri çeşitli hizmet ve ürünlere verdikleri derecelendirme değerlerinden yola çıkarak bu tahminler yapılır. OF sistemlerinin öneri üretmek için kullandığı temel yaklaşım, geçmişte benzer beğeni profillerine sahip kullanıcıların gelecekte de benzer beğeni profillerine sahip olacağı varsayımıdır. Bu varsayımdan yola çıkarak, OF sistemleri kullanıcılarının internet servisleri tarafından sunulan ürünler ya da hizmetler hakkındaki derecelendirme verilerini toplayıp analiz ederek tahminler üretir. Örneğin Amazon, YouTube, Netflix, Spotify gibi platformlar OF yaklaşımını sistemlerinde etkin olarak kullanarak, kullanıcılarına kişiselleştirilmiş öneriler sunmaktadırlar.

OF sistemleri temel olarak hafıza tabanlı ve model tabanlı olmak üzere iki temel yaklaşım kullanmaktadır (Bergner vd., 2012; Al-Mani vd., 2022). Hafıza tabanlı yaklaşımda sistemden öneri isteyen aktif kullanıcıya öneri üretebilmek için *kullanıcı*×*ürün* matrisi olarak adlandırılan n adet kullanıcının m adet ürüne verdiği derecelendirme değerlerini saklayan $n \times m$ boyutlu bir veri setinin tamamı kullanılmaktadır. Hafıza-tabanlı yaklaşımda tahmin üretmek için komşuluk belirleme işlemi temel olarak ürün tabanlı ve kullanıcı tabanlı olarak gerçekleştirilir (Thorat vd, 2023). Sistem, ürünler veya kullanıcılar arasındaki benzerlikleri analiz ederek komşulukları belirler ve bu doğrultuda tahmin üretir (Afoudi vd, 2021). Komşulukların belirlenmesi için benzerlikler hesaplanırken uzaklık ve entropi tabanlı benzerlik metrikleri, pearson korelasyonu ve kosinüs benzerliği gibi metrikler kullanılabilmektedir (Fkih, 2022).

Model tabanlı yaklaşımlar, hafıza tabanlı yaklaşımlardan farklı olarak kullanıcı ve ürün matrisini tahmin modelini oluşturmak için kullanmaktadır (Ning vd., 2015). Model tabanlı yaklaşımlar daha karmaşık ilişkileri kavrayabilme yeteneği sayesinde geleneksel yöntemlere göre daha başarılı sonuçlar verebilmektedirler (Al-Mani vd., 2022). Ancak her iki yaklaşımında

birbirine üstünlük sağladığı yönleri bulunmaktadır. Örneğin, hafıza tabanlı yaklaşımlar ölçeklenebilirlik, soğuk-başlatma ve seyreklik gibi problemlerden mustarip olabilirken, model-tabanlı yaklaşımlar bu problemlerle başa çıkmakta daha başarılı olmaktadır. Bu nedenle her iki yöntemin de avantajlarının değerlendirilebileceği hibrit yaklaşımlar önerilmiştir (Bobadilla vd., 2013; Dwivedi ve Islam, 2023). Hibrit yaklaşımlar ile geleneksel yöntemlere kıyasla daha başarılı önerilerin üretilmesi sağlanmıştır (Dwivedi ve Islam, 2023).

1.2. OF Sistemlerinin Problemleri ve Literatürde Yapılan Çalışmalar

OF sistemleri günümüzde pek çok platformda tercih edilse de veri seyrekliği, soğuk başlatma, veri mahremiyeti ve öneri doğruluğu gibi birçok probleme sahiptir (Casino vd., 2013). Bu zorluklar arasında mahremiyetin korunması ve mahremiyet korunurken yüksek doğrulukta öneri üretilmesi kritik bir öneme sahiptir.

1.2.1. Mahremiyet riskleri ve rastgele karıştırma teknikleri

OF sistemleri, yüksek doğrulukta öneriler üretebilmek için kullanıcıların geçmişte ürünler/hizmetler için verdikleri derecelendirme değerlerini kullanır, analiz eder ve saklar. *OF* sistemlerinde, kullanıcıların deneyimlediği ürün/hizmetler için verdiği derecelendirme değerleri ve deneyimlemediği ürün/hizmetlerin kümesi gizli veriler olarak kabul edilmiştir (Polat ve Du, 2005). Kullanıcıların sistem ile paylaştığı verilerin işlenmesi mahremiyet açısından risk teşkil etmektedir. Sistemdeki herhangi bir güvenlik ihlali kullanıcıların kişisel tercihlerinin, bir kategoriye karşı ilgisinin, satın alma geçmişi ve alışveriş alışkanlıklarının ifşa olmasına sebebiyet verebilmektedir (Shayong vd., 2006). Bu tür veriler kullanıcıların özel hayatı hakkında üçüncü kişilerin bilgi sahibi olmasına yol açabileceği için kişisel kullanıcı verilerinin gizliliğinin ve güvenliğinin sağlanması gereklidir. Gerçek kullanıcı derecelendirme değerlerinin kullanılması kullanıcının demografik bilgileri, yaşam tarzı özellikleri, alışveriş alışkanlıkları, gelir durumu, yaşı ve cinsiyeti gibi kişisel tanımlayıcıların tahmin edilmesi için basamak olarak kullanılabilmektedir (Yargıç ve Bilge, 2017). Bir kullanıcının tercih ettiği müzik, izlediği film ya da ilgisini çeken ürünler, onun yaşını ve cinsiyetini tanımlamak için kullanılabilir (Chaabane vd. 2012; Shyong vd., 2006). Doğrudan yaş ve cinsiyet bilgileri özel sayılmasa da bu tür bilgilere dayanarak elde edilen detaylar özel hayata müdahale olarak değerlendirilebilir. Bu tür verilere dayanarak kullanıcının rızası olmaksızın reklam gönderilmesi veya kişiye özel fiyatlandırma yapılması olası riskler arasındadır (Weinsberg vd., 2012). Dahası, bir kişinin ürünlere verdiği derecelendirmeler ve yorumlar, gelir durumu ya da yaşam tarzı hakkında da bilgi verebilir. Shyong vd., bir bireyin kimliği sadece posta kodu, yaş

ve cinsiyet bilgileri kullanılarak %87 olasılıkla saptanabilir (Shyong vd., 2006). Bu bilgiler doğrultusunda, mahremiyetleri konusunda güvensiz hisseden kullanıcılar herhangi bir öneri sistemini kullanmayı tamamen reddetmeyi seçebilmekte veya sisteme kasıtlı olarak yanlış verileri sunabilmektedir (Berkovsky vd., 2007; Ackerman vd., 1999). Kullanıcıların paylaştığı yanlış veriler sistemin ürettiği önerilerin doğruluğunu olumsuz etkileyerek kullanıcı memnuniyetinin düşmesine yol açabilmektedir. Buna göre, OF sistemlerindeki mahremiyet riskleri öneri doğruluğu problemini de tetiklemektedir. Bu nedenle OF sistemlerini etkin olarak kullanan platformların kullanıcı mahremiyetini sağlaması ve bu esnada öneri doğruluğunu muhafaza edebilmesi kritik bir konudur.

Gizliliği Korunan Ortak Filtreleme (GKOF) sistemleri OF sistemlerinin mahremiyet konusundaki zayıflıkları üzerine yapılan çalışmalar sonucu ortaya çıkmış kritik öneme sahip bir alandır. GKOF sistemlerinde kullanıcı ve sistem mahremiyetini sağlamak için kriptografik yaklaşımlar, veri gizleme ve karıştırma temelli yöntemler, anonimleştirme ve diferansiyel gizlilik gibi yöntemler ortaya konmuştur (Bilge vd., 2013).

GKOF sistemleri içerisinde kullanıcı mahremiyetini korumak için kullanılan RKT kullanıcılara ait gerçek ürün/hizmet derecelendirme değerlerini sisteme göndermeden önce maskeleyerek kullanıcı gizliliğini sağlamayı hedeflemektedir. Bu işlem, belirli dağılımlarla elde edilmiş olan rastgele sayıların gerçek kullanıcı derecelendirme değerleri üzerine eklenmesiyle gerçekleştirilmektedir (Polat ve Du, 2003; Polat ve Du, 2005). Bu kapsamda Polatidis vd. (2017), OF sistemlerinde kullanıcı gizliliğini sağlamak için çok düzeyli bir gizlilik yöntemi önermiştir. Bu yöntemde, derecelendirme değerleri sunucuya gönderilmeden önce kullanıcı tarafından belirlenecek üç farklı mahremiyet seviyesine göre maskelenmektedir. Böylece kullanıcıların farklı seviyelerdeki mahremiyet ihtiyaçlarına göre gerçek kullanıcı derecelendirme değerleri farklı düzeylerde maskelenerek kullanıcıya sağlanan gizlilik ile tahmin doğruluğu arasında bir denge kurulması hedeflenmektedir. Yalçın ve Bilge (2023), popülerlik yanlılığı sorununa karşı RKT'ne dayalı veri gizleme prosedürünü analiz etmektedir. Yapılan çalışmada farklı mahremiyet seviyelerine sahip kullanıcı karakterleri oluşturularak farklı öneri üretme algoritmalarının öneri doğrulukları üzerine etkisi incelenmektedir. Bir diğer çalışmada Yargıç ve Bilge (2019), geleneksel tek-kriterli ortak filtreleme sistemlerine ek olarak çok-kriterli ortak filtreleme sistemlerinde kullanıcı mahremiyetini sağlamak için RKT'ni kullanmaktadır. Birden fazla ölçütün maskelendiği bu yaklaşımda, doğruluk kayıplarını telafi edebilmek için her bir alt kriterin doğasında bulunan bilgi miktarını referans alarak ölçütlerin mahremiyet seviyesini kontrol eden entropi tabanlı bir rastgelelik belirleme prosedürü

sunulmuştur. Bu çalışmalara ek olarak, RKT tabanlı koruma mekanizmalarına göre daha yüksek seviyede gizlilik elde etmek için RKT ve diferansiyel gizlilik metodunun birleştirildiği hibrit gizlilik koruma yaklaşımları da bulunmaktadır (Liu vd. 2017).

RKT, kullanıcıların mahremiyetini sağlamak için etkili bir yöntem olsa da sistemin maskelenmiş veri setleri üzerinde tahmin üretmesi doğruluk kayıplarına sebep olmaktadır. Bu nedenle; kullanıcı gizliliğinin RKT ile sağlandığı durumlarda mahremiyet seviyesi ve öneri doğruluğu arasında bir denge kurulması ve maskelenmiş veri setleri üzerinde tahmin doğruluğunu iyileştirmeye yönelik geleneksel tahmin üretme yaklaşımlarının yerine daha yenilikçi yaklaşımlarının kullanılması gerekmektedir.

1.2.2. DÖ tabanlı yaklaşımlar

OF sistemlerinin temel amacı büyük veri setleri üzerinden kullanıcılara yüksek doğrulukta ve kişiselleştirilmiş öneriler sunmaktır. Son yıllarda birçok alanda kullanılan ve başarılı sonuçlar veren DÖ teknikleri OF sistemlerinde de yaygın olarak kullanılmaktadır. DÖ teknikleri, veri setlerinden gizli ve karmaşık özellikleri çıkarabildikleri için OF sistemlerinde tahmin doğruluğunu artırmak amacıyla kullanılmıştır. DÖ yöntemleri, karmaşık mimarilere (birden çok gizli katman, döngü veya paylaşılan ağırlıklar ve havuzlama katmanlarına) sahip bir makine öğrenimi dalıdır ve temeli yapay sinir ağlarına dayanmaktadır (Kiran vd., 2020). DÖ yöntemleri; bilgisayarla görme, görüntü işleme gibi pek çok araştırma alanında büyük bir başarı elde etmiştir (Lecun vd., 2015). DÖ yöntemlerinin bilgisayar bilimlerindeki popülaritesi ve başarısı, ÖS alanında çalışan araştırmacıların da dikkatini çekmiştir ve çalışmalarında ölçeklenebilirlik, veri seyrekliği, soğuk başlatma gibi zorlukların hafifletilmesi için DÖ tekniklerinden faydalanmışlardır (Georgiev ve Nakov, 2013; Strub vd., 2015; Yang vd., 2017). Son zamanlarda yapılan birçok çalışma, DÖ yöntemlerinin ÖS üzerindeki etkinliğini doğrulamaktadır (Khan vd., 2021).

Otokodlayıcılardır ÖS'lerinde sıklıkla kullanılan bir DÖ yaklaşımıdır. Denetimsiz bir DÖ yöntemi olarak otokodlayıcılar, veri boyutunu azaltma, özellik çıkarma ve veri setini yeniden oluşturmadaki başarılı performansı nedeniyle yaygın olarak kullanılmaktadır (Zhang vd., 2020). Ayrıca, ÖS'lerinde kullanıcıların ve öğelerin özelliklerini daha iyi analiz edebileceği için önerilerin kalitesini arttırmaktadır (Zhang vd., 2020). Unger vd., kullanıcıların zaman içinde değişen tercihlerini yüksek boyutlu veri setlerinden anlayabilmek için otokodlayıcıları kullanmıştır (Unger vd. 2018). Önerdiği yaklaşımda otokodlayıcılar yüksek boyutlu veriyi

öğrenerek gizli bir temsile indirger, elde edilen temsiller sayesinde kullanıcının sevebileceği ürünlerin kategorisi tahmin edilmektedir.

Otokodlayıcıların ÖS için kullanımının ana örneklerinden biri olan AutoRec adlı çalışma, Sedhain vd., tarafından sunulmuştur (Sedhain vd., 2015). Kullanıcı ve ürün tabanlı olmak üzere 2 farklı yaklaşımla yürütülen çalışmada, veri setinin düşük boyutlu temsilleri öğrenilerek öneriler üretilmektedir. Önerilen yöntemin geleneksel ÖS'ne göre daha yüksek öneri doğruluğu sağladığı gözlemlenmiştir. AutoRec çalışmasından ilham alan Haghighi vd., en doğru tahminlerin, öğelerin kullanıcıya neden önerildiği hakkında fikir sahibi olunmayan kara kutular tarafından üretilme eğiliminde olduğuna dikkat çekmektedir (Haghighi vd., 2019). Çalışmalarında otokodlayıcıları komşuluk tabanlı tahmin üreten bir yaklaşımla birlikte kullanarak daha açıklanabilir bir öneri sistemi tasarladıklarını belirtmektedirler. Sonuçlarını AutoRec ile karşılaştırdıklarında, en az AutoRec kadar tahmin doğruluğu elde ederken, *en-iyi-N* tavsiyedeki açıklanabilirliği geliştirdikleri gözlemlenmiştir.

Bathla vd. (2020), otokodlayıcıları seyreklik ve soğuk başlatma durumlarında öneri doğruluğunu artırmak için çalışmak üzere uyarlamıştır. Önerilen yaklaşımda, otokodlayıcıları eğitirken kullanıcıların açık ve örtülü derecelendirme değerleri birlikte kullanılarak daha güvenilir bir veri seti elde edilmektedir. Sonuçlar literatürdeki benzer çalışmalar ile karşılaştırıldığında, daha yüksek öneri doğruluğu gözlenirken, otokodlayıcıların kullanımı ve güvenilir derecelendirme değeri verileri, tahmin hatasındaki azalmanın nedeni olarak gösterilmektedir (Bathla vd., 2020). Heidari vd. (2022), Bathla vd.'ye benzer bir amaçla ÖS'lerin başarısını düşüren en önemli faktörlerden olan seyreklik ve soğuk başlatma problemleriyle baş edebilecek bir yaklaşım önermiştir. Önerilen yaklaşım beş temel adımdan oluşmaktadır ve boyutsallık azaltma amacıyla otokodlayıcılar kullanılmıştır. Önerilen yaklaşımın deneysel çalışmaları MovieLens-100K, Movielens-1M ve Book-Crossing[1] (BX) veri setleri kullanılarak yapılmıştır. Kullanılan veri setlerinin arasında BX veri seti en seyrek matris olmasına rağmen önerilen yöntemle sağlanan iyileştirme oranının en yüksek gözlemlendiği veri seti de BX'dir. Kullanılan otokodlayıcı yapısında [1-4] aralığında değişen katman sayıları kullanılarak hata üzerindeki etkisi gözlemlenmiş ve en düşük hata değeri 4 katmanlı otokodlayıcı kullanıldığında elde edilmiştir. Al Sbou ve Rahim (2022), aynı problemler için otokodlayıcı ile özellik çıkarımı yaparak önerilerin doğruluğunu yükseltmeyi amaçlamıştır. Üç farklı otokodlayıcı yapısını karşılaştıran bu çalışmada MovieLens-100K veri seti kullanılmıştır. Otokodlayıcıların ÖS'nin öneri kalitesini yükseltmede ve seyreklik, soğuk

[1] https://grouplens.org/datasets/book-crossing/

başlatma gibi problemlerle baş etmede etkili bir teknik olduğu görülmüştür. Ibrahim vd., kullanıcı ve ürün özelliklerini yakalamak için DÖ kullanmıştır ve bunu OF sistemi ile birleştirerek hibrit bir öneri sistemi sunmuştur. Önerilen yaklaşımda kullanıcılara kişiselleştirilmiş öneriler üretmek için OF'den yararlanılırken, yüksek boyutlu verilerin işlenmesinde, özellik çıkarımında, öğrenme sürecindeki hata oranını azaltılmasında ve daha yüksek öneri doğruluğu elde edilmesinde DÖ'nin avantajlarından faydalanılmıştır (Ibrahim vd., 2023).

Batmaz (2019) tarafından, otokodlayıcıların çok-kriterli veri setlerinde kullanıldığı ve OF sistemlerinin temel problemlerinden olan öneri doğruluğu ve veri seyrekliği probleminin çözümüne yönelik bir çalışma yapılmıştır. Literatürde yapılan önceki çalışmalara ilave olarak veri setlerindeki gizli özelliklerden faydalanmayı sağlayacak AE-simMCCF adlı otokodlayıcı tabanlı yaklaşımı sunmuşlardır. Geleneksel OF sistemlerindeki benzerlik hesaplama adımında ham kullanıcı verileri yerine otokodlayıcı kullanılarak çıkarılan gizli özellikler kullanılmıştır. Bu sayede önerilen yöntem veri seyrekliğinin sebep olduğu düşük öneri doğruluğu problemini hafifletmiştir. Önerilen yaklaşımda kullanıcıların derecelendirme değerlerine ek olarak otokodlayıcılarla kullanıcıların yorum tabanlı profillerinden çıkarılan özellikler de kullanılmıştır. AE-simMCCF yaklaşımının deneysel çalışmaları farklı katman sayıları ve aktivasyon fonksiyonları kullanılarak Yahoo!Movies ve TripAdvisor veri setleri üzerinde yapılmıştır. Deneysel çalışmalar kodlayıcı katman sayısının artmasıyla birlikte gizli özelliklerin daha başarılı çıkarılabilmesinden dolayı öneri doğruluğunun yükseldiğini göstermiştir. Ayrıca diğer çok-kriterli OF algoritmalarıyla kıyaslandığında da daha yüksek öneri doğruluğu sağladığı belirtilmiştir. Literatürdeki otokodlayıcı tabanlı çalışmalar genellikle daha yüksek tahmin doğruluğu sağlamak veya soğuk başlatma ve seyreklik sorunlarıyla başa çıkmak için geliştirilmiştir. ÖS için kritik bir sorun olan mahremiyet konusunda yaptığı HETEDP çalışması ile Wei vd., diferansiyel gizlilik kullanarak veri mahremiyetini sağlarken öneri üretme aşamasında otokodlayıcı kullanılan bir yaklaşımı literatüre kazandırmıştır (Wei vd., 2022).

1.3. Amaç ve Katkılar

Öneri sistemlerinde sıklıkla kullanılan OF sistemleri, kullanıcıların kişisel değerlendirmelerini sistemle paylaşmasından dolayı mahremiyet riskleri oluşturmaktadır. Bu mahremiyet riskleri ile baş edebilmek için önerilen gizliliği koruma yaklaşımlarından biri olan RKT, kullanıcı mahremiyetini sağlamakta başarılı olmakta fakat öneri doğruluğunu olumsuz yönde etkilemektedir. Bu tez çalışmasının amacı; OF sistemlerinde kullanıcıların ham derecelendirme verilerinin kullanımından kaynaklanan mahremiyet risklerini RKT ile

azaltırken RKT'nin veride yarattığı bozulmadan kaynaklanan öneri doğruluğundaki düşüşü otokodlayıcı tabanlı bir yaklaşımla hafifletmektir.

Geleneksel komşuluk tabanlı OF sistemlerinin mahremiyet riskleri hem kullanıcı tarafında hem de firmalar tarafında pek çok dezavantaja neden olmaktadır ve bu risklerin giderilmesi pek çok açıdan önem arz etmektedir. Bu noktada geliştirilen GKOF sistemleri ise mahremiyet sorunlarına büyük ölçüde çözüm sağlamaktadır. Orijinal verilerin saklanması konusunda bu yaklaşımlar başarılı olsa da sistemin kullanıcıya yüksek doğrulukta öneriler üretebilmesi için kullanıcılara ait yeterli sayıda ve gerçek derecelendirme değerlerine ihtiyacı vardır. RKT gibi veri maskeleme teknikleri ise veriye gürültü ekleyerek veri setindeki gerçek kullanıcı derecelendirme değerlerini bozmakta ve dolayısıyla öneri doğruluğunda düşüşe sebebiyet vermektedir. RKT ile mahremiyeti sağlanan geleneksel *OF* sistemlerinde gizlilik ve doğruluk arasında denge kurulması kritik öneme sahip bir konudur (Yargıç ve Bilge, 2019). Bu noktada ilk aşamada; RKT ile sağlanan farklı mahremiyet seviyelerine göre geleneksel komşuluk tabanlı OF sistemlerinin öneri doğrulukları incelenmiştir. Kullanılan veri seti için sağlanabilecek en yüksek mahremiyet seviyesi ve öneri doğruluğu için dengeli bir mahremiyet seviyesi tespit edilmiştir. İkinci aşamada; RKT'nden kaynaklanan öneri doğruluğu kaybı için otokodlayıcı tabanlı tahmin üretme yaklaşımı kullanılmıştır. OF sistemlerinde DÖ teknikleri, ham veriden gizli ve karmaşık özellikleri çıkarabilme ve farklı kaynaklardan bilgileri bir araya getirebilme becerileri dolayısıyla tahmin doğruluğunu yükseltmek için sıklıkla tercih edilmiştir. Ham kullanıcı verileri ile yapılan çalışmalarda öneri doğruluğunu yükselttiği görülen otokodlayıcıların, RKT ile maskelenmiş veriler kullanıldığında da yüksek öneri doğruluğu sağladığı görülmüştür.

Bu doğrultuda maskelenmiş veriler kullanılarak otokodlayıcı tabanlı bir öneri sistemi oluşturulmuştur. Oluşturulan otokodlayıcı tabanlı sistemde farklı gizlilik seviyelerinde maskelenmiş veri setleri ile deneyler yapılarak öneri doğrulukları incelenmiştir.

Özetle, yapılan çalışmanın mevcut literatüre katkıları şu şekilde sıralanabilir:

- Tahmin üretmek için ham veri setlerinde sıklıkla kullanılan otokodlayıcılar, RKT ile maskelenen veriler ile kullanılmıştır.
- RKT ile maskelenmiş veri seti için, otokodlayıcıda kullanılan aktivasyon fonksiyonları ve gizli katman sayısının öneri doğruluğu üzerindeki etkileri incelenmiştir.
- RKT'nin veri setinde yarattığı bozulmadan kaynaklanan öneri doğruluğu kaybı azaltılmıştır.

- Otokodlayıcı tabanlı ortak filtreleme sisteminin yükselen gizlilik seviyelerine bağlı olarak veri setinde ortaya çıkan bozulmaları daha iyi tolere edebildiği ve kullanıcıya yüksek mahremiyet seviyeleri sağlandığında da yüksek doğrulukta öneriler üretebildiği gösterilmiştir.

2. MATERYAL VE YÖNTEM

Bu bölümde OF sistemlerinde geleneksel komşuluk tabanlı ve otokodlayıcı tabanlı yaklaşımlar ile tahmin üretme süreci ve veri mahremiyetini sağlamak için kullanılan RKT ile ilgili genel bilgiler verilmektedir.

2.1. OF Sistemlerinde Öneri Üretme Süreci

OF sistemlerinde öneri üretme süreci n adet kullanıcının m adet ürün için verdiği derecelendirme değerlerinden oluşan $n \times m$ boyutlu bir matris üzerinde gerçekleştirilir. Sistemden bir hedef ürün (q) için öneri talebinde bulunan aktif kullanıcı (a), sistemle geçmişte tercih ettiği ürünleri, derecelendirme değerlerini paylaşmış olmalıdır. Tüm bu koşullar sağlandığı takdirde geleneksel *OF* sistemlerinde komşuluğa dayalı olarak gerçekleştirilen öneri üretme süreci en genel haliyle iki adımda gerçekleştirilmektedir;

i. öneri talebinde bulunan *a*'ya benzer tercih eğilimleri gösteren kullanıcılar, yani komşular tespit edilir.

ii. komşu kullanıcıların geçmiş tercihlerine göre sistem *a* için kişiselleştirilmiş öneriler üretir.

Kullanıcı derecelendirme değerlerinin tutulduğu ve üzerinden öneri üretme işlemlerinin gerçekleştirildiği örnek bir kullanıcı×ürün matrisi Tablo 2.1'de görülmektedir.

Tablo 2.1. Örnek kullanıcı×ürün Matrisi

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	1	3	1	3	3	5	4	5
u_2	1	3	1	3	3	**?**	4	5
u_3	1	3	1	3	3	3	4	3
u_4	2	2	2	2	2	3	1	3
u_5	2	2	2	2	2	3	1	3

Tablo 2.1'de gösterilen $\{u_{1,\ldots,}u_5\}$ kullanıcıları, $\{i_{1,\ldots,}i_8\}$ ise ürünleri ve kesişim hücrelerindeki değerler kullanıcının ürün için verdiği derecelendirme değerini ifade etmektedir. Örneğin tabloda a olarak belirlenen kullanıcı u_2, q olarak belirlenen i_6 için öneri istemektedir. Geleneksel komşuluk tabanlı OF sistemlerinde öneri üretme süreci a'ya en fazla benzerlik gösteren N adet kullanıcı profilini belirleyerek başlamaktadır. Benzer profilleri belirlemek için literatürde çeşitli yöntemler kullanılmakla birlikte Pearson Korelasyonu (PK) iki veri arasındaki korelasyonun hangi yönde ve ne seviyede güçlü olduğunu ölçmek için sıklıkla kullanılan bir benzerlik metriğidir. PK ile iki kullanıcı arasındaki benzerliği hesaplamak için kullanılan eşitlik, Denklem 2.1'de verilmiştir.

$$PK(a,u) = \frac{\left(\Sigma_{i\in I}\left(r_{a,i}-\overline{r_a}\right)\times\left(r_{u,i}-\overline{r_u}\right)\right)}{\left(\sqrt{\Sigma_{i\in I}\left(r_{a,i}-\overline{r_a}\right)^2}\sqrt{\Sigma_{i\in I}\left(r_{u,i}-\overline{r_u}\right)^2}\right)} \quad (2.1)$$

Denklem 2.1'de, *a* ve *u* kullanıcılarının her ikisinin de derecelendirdiği ürünler *I* ile temsil edilmiştir. $r_{a,i}$ *ve* $r_{u,i}$ *a* ve *u* kullanıcılarının *i* ürününe verdiği derecelendirme değerini ifade etmektedir. $\overline{r_a}$ *ve* $\overline{r_u}$ *a* ve *u* kullanıcılarının geçmişte değerlendirdiği tüm ürünler için derecelendirme değerlerinin ortalamasıdır. Eşitliğin sonucunda elde edilen değer [-1,1] aralığındadır ve bu iki kullanıcı arasındaki ilişkiyi sayısal olarak ifade etmektedir. Elde edilen sonucun yorumlanması ise aşağıda verilen maddelere göre yapılmaktadır;

i. hesaplanan PK değeri -1'e yaklaşması bu iki veri arasında negatif korelasyon olduğu, yani benzerliğin ters orantılı değiştiği şeklinde yorumlanmaktadır.

ii. hesaplanan PK değeri 0'a yaklaştıkça bu iki veri arasında korelasyon olmadığı sonucuna ulaşılmaktadır.

iii. hesaplanan PK değeri 1'e yaklaştıkça iki veri arasında pozitif korelasyon olduğu şeklinde yorumlanmaktadır.

Özetle; PK değeri 1'e ne kadar yakın hesaplanırsa, bu iki kullanıcı arasındaki benzerlik o kadar yüksektir.

Birbirine en yüksek benzerlik gösteren *N* adet kullanıcı profillerinin belirlenmesi ve bu profillerin tercihlerine göre öneriler üretilmesi *en-iyi N* (*top-N)* öneri yaklaşımıdır. Veri setindeki diğer kullanıcılar ve *a* arasındaki benzerlik PK ile hesaplandıktan sonra ilk *N* adet kullanıcı profili *a*'nın komşusu olarak belirlenerek öneri üretme sürecinde belirlenen *N* kullanıcının *q* ürünü için verdiği derecelendirme değerleri *a*'ya öneri üretilirken referans alınacaktır. *a* kullanıcısının *q* ürünü için öneri üretme işlemi Denklem 2.2'de verilmiştir.

$$p_{a,q} = \frac{\Sigma_{u\in N}\left(r_{u,q} - \overline{r_u}\right)PK(a,u)}{\Sigma_{u\in N}\, PK(a,u)} \quad (2.2)$$

Denklem 2.2'de *N, a'*ya en yüksek benzerliği gösteren *N* adet kullanıcıyı, yani komşuluk kümesini temsil etmektedir. $r_{u,q}$, *u* kullanıcısının *q* ürününe verdiği derecelendirme değerini ve PK*(a,u)*, *a* ve *u* arasındaki benzerlik değerini ifade etmektedir.

2.2. RKT

RKT'nin temel prensibi, öneri sistemine kullanıcının ham derecelendirme değerlerini göndermek yerine maskelenmiş derecelendirme değerlerini göndererek gizlilik sağlamaktır.

Gizliliği RKT ile sağlanan sistemlerde kullanıcının ham verileri sunucuyla paylaşılmamakta fakat maskelenmiş veriler kullanılarak hala kullanıcıya kişiselleştirilmiş öneriler üretilebilmektedir (Polat ve Du 2005). RKT'nin çalışma mekanizması, kullanıcının gerçek derecelendirme değerleri (G) ve üzerine eklenen rastgele sayı vektörü (R) olarak adlandırıldığında, sunucuya G yerine G+R vektörünü göndermesidir.

RKT kullanılan sistemlerde mahremiyet farklı seviyelerde sağlanabilmektedir. Kullanıcının orijinal verilerini bozarak gizlilik sağlayan RKT, kullanıcının ihtiyacını karşılayacak mahremiyet seviyesini sağlamak üzere ve veri setinin değerlendirme ölçeğine bağlı olarak orijinal verileri farklı seviyelerde bozmaktadır. Bu işlem için derecelendirme değerleri üzerine eklenecek olan R vektörü gizlilik seviyesini belirleyen ve σ katsayısı olarak adlandırılan bir parametre ile kontrol edilmektedir. R, σ katsayısına bağlı olarak, normal ve uniform gibi dağılımlara göre üretilebilmektedir (Bilge vd., 2013). Uniform dağılıma göre üretilen R $\left[-\sqrt{3\sigma}, \sqrt{3\sigma}\right]$ aralığında üretilirken normal dağılıma göre üretilen R ortalaması (μ) sıfıra eşit olacak şekilde ve kullanıcının derecelendirme değerlerinin standart sapmasına bağlı olarak üretilmektedir.

Veri maskeleme işleminden önce ham kullanıcı verilerinde kullanıcı ürün değerlendirme alışkanlıklarından dolayı ortaya çıkabilecek sapmaları hafifletmek için veri normalizasyonu gerçekleşmektedir. Kullanıcılar deneyimledikleri ürünleri derecelendirirken belirli standartlara bağlı kalmazlar. Araştırmacılar, kullanıcılar bir ürün derecelendirirken her kullanıcının derecelendirme sürecini farklı algılayabileceğini düşünmektedir (Ning vd., 2015). Bazı kullanıcılar beğendikleri veya beğenmedikleri ürünler için en yüksek veya en düşük derecelendirme değeri verebilirken, bazı kullanıcılar daha detaylı bir değerlendirme yaparak değerlendirme ölçeğinin tümündeki derecelendirmeleri kullanabilir. Bu iki farklı değerlendirme karakterine sahip kullanıcılar verdikleri derecelendirme değerleri eşit olmasa dahi benzer profillerde olabilmektedirler. Normalizasyon işlemi, kullanıcıların derecelendirme değerlerini genel bir ölçeğe uyumlu hale dönüştürmeyi amaçlamaktadır (Althbiti vd., 2020). Tablo 2.2'de [1,10] değer aralığında derecelendirme değerlerine sahip temsili bir veri seti ve bu derecelendirme değerlerinin z-skor normalizasyonu ile normalize edilmiş değerleri bulunmaktadır. Tablodaki derecelendirme değerlerine bakıldığında, kullanıcıların farklı değer aralıkları içerisinde değerlendirme yapmış olsalar bile benzer profillerde olabildikleri görülmektedir. $u_1$ve u_2 ile u_3 ve u_4 aynı ürünleri aynı derecelendirme değerleriyle derecelendirmemelerine rağmen, z-skor normalizasyonu uygulandığında profillerinin tamamen uyumlu olduğu sonucuna varılmaktadır.

Tablo 2.2. Kullanıcıların Orijinal Derecelendirme Değerleri ve z-skor ile Normalize Edilmiş Derecelendirme Değerleri

		u_1	u_2	u_3	u_4
i_1	**Derecelendirme**	5,00	8,00	2,00	3,00
	z-skor	0,53	0,53	-1,00	-1,00
i_2	**Derecelendirme**	4,00	7,00	2,00	3,00
	z-skor	0,00	0,00	2,00	3,00
i_3	**Derecelendirme**	6,00	9,00	3,00	4,00
	z-skor	1,07	1,07	1,00	1,00
i_4	**Derecelendirme**	1,00	4,00	3,00	4,00
	z-skor	-1,60	-1,60	1,00	1,00

Veri seti içerisindeki derecelendirmelerin, kullanıcı profiline aykırı derecelendirme değerlerine sahip olması sistem tarafından üretilen önerilerin doğruluğunu olumsuz etkilemektedir (Bilge ve Yargıç, 2017). Bu olumsuz etkinin hafifletilmesi amacıyla, kullanıcının derecelendirme değerlerine maskeleme işleminden önce z-skor normalizasyonu uygulanır. Z-skor normalizasyonu Denklem 2.3'te verilmiştir (Herlocker vd., 1999).

$$z_{ui} = \frac{r_{ui} - \overline{r_u}}{\sigma_u} \tag{2.3}$$

Bu eşitlikte r_{ui}, bir kullanıcının i için derecelendirme değeri, $\overline{r_u}$ kullanıcının derecelendirme değerlerinin ortalamasını ve σ_u u kullanıcısına ait derecelendirme değerlerinin standart sapmasını temsil etmektedir.

Veri normalizasyon işlemi gerçekleştirildikten sonra kullanıcı verileri RKT ile maskelenmektedir. Bu işlem temel olarak Prosedür 2.1'de gösterilmektedir. Yapılan işlemler sırasıyla:

i. *Gizlilik parametresi σ 'nın belirlenmesi:* Kullanıcının gizlilik ihtiyacı ve kullanılan veri setinin değerlendirme ölçeğine göre mahremiyet seviyesini tanımlayacak σ katsayısı, sunucu tarafından belirlenen $(0, \sigma_{max}]$ aralığındaki bir değer olarak belirlenir.
ii. *Rastgele sayı vektörünün üretilmesi:* Belirlenen σ değerine ve N dağılıma göre R oluşturulur.
iii. *Maskelenmiş derecelendirme değerlerinin elde edilmesi:* Önceki adımlarda elde edilmiş olan z-skor uygulanmış gerçek derecelendirme değerleri vektörü ile R toplanarak maskelenmiş derecelendirme değerleri vektörü elde edilir.

Prosedür 2.1. RKT Uygulanma Adımları (Bilge ve Polat, 2013)

Giriş: kullanıcı × ürün vektörü (V), σ_{max}

Çıkış: Maskelenmiş kullanıcı×ürün vektörü (V')

Z-skor normalizasyonu → Z

1. $\overline{V} \leftarrow Ortalama\,(V)$
2. $\sigma_V \leftarrow StdSapma(V)$
3. $for\ all\ items\ in\ V\big(i \leftarrow 1\ to\ length(V)\big)\ do$
4. $Z_i \leftarrow \left(V_i - \overline{V_i}\right) \div \sigma_V$
5. $end\ for$

Gizlilik parametresi σ belirlenir

6. $\sigma \leftarrow (0, \sigma_{max}]$

Normal dağılım ve σ değerine göre R üretilir

7. $R \leftarrow RastgeleSayiUret(\sigma)$

Z-skor ile normalize edilmiş vektöre R vektörü ekleyerek derecelendirme değerlerini maskele

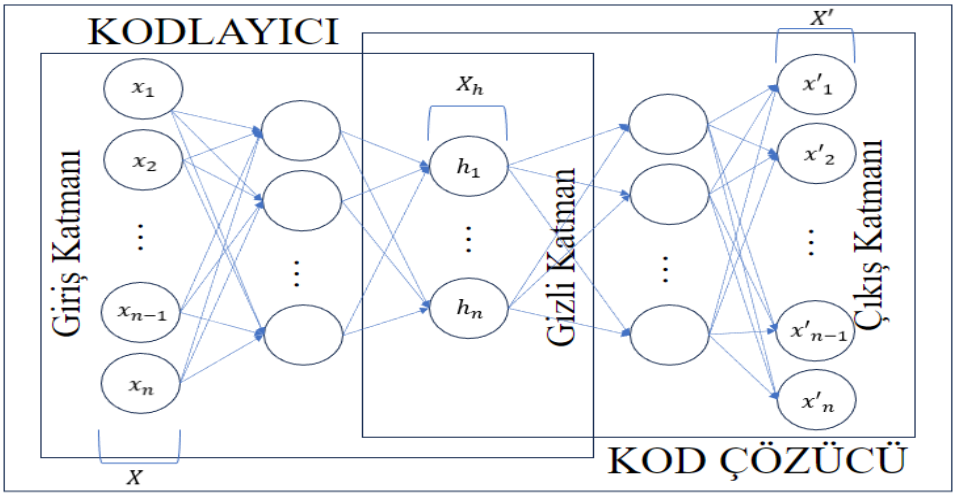
8. $V' = Z + R$
9. $Return\ V'$

2.3. Otokodlayıcılar

Otokodlayıcılar ÖS'de verilerin sahip oldukları gizli ilişkileri çıkarmak amacıyla kullanılmaktadırlar. Otokodlayıcılar gizli katmanlarında giriş verisini sıkıştırır ve yeniden oluşturur. Bu esnada özellik çıkarımı yaparak verilerdeki gizli temsilleri öğrenir ve daha sonrasında bunları kullanarak kullanıcılara veya öğelere ilişkin öneriler üretir.

Otokodlayıcılar kodlayıcı ve kod çözücü olmak üzere iki ana parçadan oluşurlar ve bu parçalar giriş katmanı, gizli katmanlar ve çıkış katmanı olmak üzere *l* adet katman içermektedir (Goodfellow ve Bengio, 2016). Giriş katmanı, otokodlayıcının besleneceği verilerin ağa verildiği katmandır. Giriş katmanından alınan veriler ağırlıklandırılarak gizli katmanlara doğru iletilir. Kodlayıcı kısmında giriş katmanından sonra nöron sayıları her katmanda bir öncekine göre daha az sayıdadır. Nöron sayıları azaldıkça katman çıkışındaki veri de daha küçük boyutlu olmaktadır. Böylece kodlayıcı kısmındaki en az nöron sayısına sahip gizli katmanın çıktısı, giriş verisinin düşük boyutlu temsili olarak kullanılmaktadır. Bu nedenle otokodlayıcıların kodlayıcı kısmı boyut indirgeme ve özellik çıkarımı için kullanılmaktadır. Kod çözücü katmanda ise,

kodlayıcı katmanda düşük boyutlu hale getirilen veri gittikçe nöron sayıları artan gizli katmanlar arasında iletilerek yeniden oluşturulmaya çalışılır. Kod çözücü katman mimari olarak kodlayıcı katmanın simetriğidir ve giriş katmanı ile çıkış katmanının nöron sayıları birbirine eşittir. Şekil 2.1'de birden çok gizli katman içeren bir otokodlayıcı modeli örneklenmiştir. Klasik yapay sinir ağlarından temel farkı giriş katmanındaki nöron sayısı ve çıkış katmanındaki nöron sayısının aynı olmasıdır.



Şekil 2.1. Otokodlayıcı Yapısı

Şekil 2.1'de dört katmanlı bir otokodlayıcının kodlayıcı ve kod çözücü kısımları görülmektedir. Kodlayıcı kısmında, giriş katmanına beslenen veri seti *X* olarak gösterilmiştir ve $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$vektörlerinden oluşmaktadır. Giriş katmanından otokodlayıcıya beslenen veriler ağırlıklandırılarak bir sonraki katmana iletilir. Kodlayıcıda sıradaki her gizli katmanda azalan nöron sayıları ile *X* kodlanır ve en az sayıda nöron içeren gizli katmana ulaşır. Bu katmanın nöronları şekilde X_h olarak gösterilmiştir ve çıktısı $\{h_1, \ldots, h_n\}$, *X* veri setinin boyutunun indirgenmiş halidir. Bu katmandan sonra X_h, kod çözücü kısmındaki katmanlara iletilerek çıkış katmanında *X* veri seti yeniden elde edilmeye çalışılmaktadır. X_h, nöron sayıları kodlayıcıya simetrik olacak şekilde artan katmanlar arasında iletilir ve en son giriş katmanıyla eşit nöron sayısındaki çıkış katmanına ulaşır. Çıkış katmanında yeniden elde edilen veriler $\{x'_1, \ldots, x'_n\}$olarak şekilde görülmektedir ve çıkış katmanında yeniden oluşturulmuş veri seti *X'* ile ifade edilmiştir.

Otokodlayıcı tüm bu süreçleri tamamlarken Denklem 2.4, 2.5 ve 2.6'yı kullanmaktadır (Batmaz, 2019).

$$f(x) = \varphi(W_1 x + b_1) \tag{2.4}$$

Denklem 2.4'te verilen parametreler x, φ ve b_1 sırasıyla giriş verilerini, aktivasyon fonksiyonunu ve bias vektörünü ifade etmektedir.

$$g\big(f(x)\big) = \delta(W_2 f(x) + b_2) \tag{2.5}$$

Denklem 2.5'te verilen δ, W, b_2 ve sırasıyla doğrusal olmayan bir fonksiyonu, kodlayıcı katman ile kod çözücü katman arasındaki ağırlıkları ve bias vektörünü temsil etmektedir.

$$\Sigma_{x \in X} \left|\left| x - g\big(f(x)\big) \right|\right|_2^2 \tag{2.6}$$

Verilen bir giriş verisi için, otokodlayıcı öğrenme işlemini Denklem 2.6'yı minimize ederek gerçekleştirir.

Derin sinir ağlarında olan hız ve aşırı öğrenme (overfitting) problemi otokodlayıcılar için de geçerlidir. Aşırı öğrenme problemi, ağın eğitim verilerine aşırı uyum sağlaması, her bir detayı ezberlemesi sonucu genelleme yeteneğinin azalması ve test verilerindeki farklı girişlere yüksek hata değerine sahip sonuçlar vermesi problemidir. Aşırı öğrenme probleminden muzdarip bir model, eğitim verilerinde yüksek doğrulukta sonuçlar vermekte fakat test verilerinde oldukça yüksek hata oranına sahip olmaktadır. Bu sorunu önleyebilmek için çeşitli yöntemler mevcuttur. Bu yöntemlerden bazıları model karmaşıklığının azaltılması, önceden eğitilmiş (pre-trained) modeller kullanma, L1 ve L2 düzenlileştirmeleridir. L2 düzenlileştirmesi aşırı uyum problemini azaltmanın etkili bir yoludur (Phaisangittisagul, 2016). Ağırlıkların büyük değerlere ulaşmasını engeller ve dağılımını yayar, böylece girişlerden çıkarılan özelliklere dengeli bir katkı sağlanır. L2 düzenlileştirmesinde, düzenlileştirmenin şiddetini kontrol etmek amacıyla λ hiperparametresi kullanılır ve seçilen λ değeri modelin performansında oldukça etkilidir.

2.4. Veri Seti ve Değerlendirme Ölçütleri

Bu bölümde yapılan çalışmada kullanılan veri seti, deneysel çalışmalarda öneri doğruluğu ve veri mahremiyetini ölçeklemek için kullanılan metrikler açıklanmaktadır. Çalışmada, kullanıcıların kitap değerlendirmelerinden oluşan BX veri seti kullanılmıştır. Öneri doğruluğu ölçeklemek için MAE ve RMSE kullanılmıştır. Ayrıca maskeleme prosedürü sonrasında veri setinin mahremiyet seviyesini ölçeklemek için Agrawal ve Aggarwal (2001) tarafından literatüre kazandırılan diferansiyel entropi tabanlı veri mahremiyeti ölçekleme metriği kullanılmıştır.

2.4.1. BX veri seti

BX veri seti, Cai-Nicolas Ziegler tarafından, kitap paylaşım grubu olan Book-Crossing topluluğunun Ağustos-Eylül 2004 tarihleri arasındaki 4 haftalık bir taraması sonucunda oluşturulmuştur. BX veri seti 271.379 kitap hakkında 1.149.780 derecelendirme değeri sağlayan 278.858 kullanıcı bilgisini demografik bilgileriyle (yaş, ülke/eyalet) birlikte içermektedir. BX veri seti users, books ve ratings olmak üzere 3 ayrı tablodan oluşmaktadır. Users tablosu, kullanıcı kişisel bilgilerini içermekte ve varsa konum ve yaş demografik verilerini sağlamaktadır. Books tablosu kitapların ISBN bilgilerini içermektedir. Ayrıca, Amazon Web Servislerinden alınan kitap başlığı, yazar, yayın yılı ve yayıncı bilgileri bu tabloda verilmektedir. Ratings tablosu kullanıcıların kitap derecelendirme değerlerini içermektedir. Kullanıcı tarafından geçmişte okunup derecelendirilen kitaplar 0-10 değer aralığı içinde ifade edilmiş olup, kullanıcı tarafından okunmuş fakat derecelendirilmemiş kitaplar 0 ile ifade edilmektedir.

Tablo 2.3. BX Veri Seti ve Oluşturulan Alt Kümeleri

	BX	**BXA**	**BXN**
Kullanıcı Sayısı	105.283	77.805	756
Kitap Sayısı	340.553	185.972	1173
Derecelendirme Değeri Sayısı	1.149.780	433.671	21.690
Seyreklik	%99,9	%99,9	%97,6

Bu çalışmada sadece ratings tablosundaki değerlendirilmiş kitapların derecelendirme değerleri kullanılmıştır, bu sebeple 0 olarak kaydedilmiş değerlerin bulunduğu satırlar tablodan çıkarılmıştır. 0 değerleri çıkarıldıktan sonra elde edilen alt küme BXA olarak adlandırılmıştır. BXA veri seti %99,9 seyreklikte bir veri setidir. BXA veri seti mevcut haliyle oldukça seyrek bir yapıya sahiptir, bu nedenle ön işlemlere tabii tutularak az 15 kez değerlendirme yapmış kullanıcılar ve farklı kullanıcılar tarafından en az 10 kez değerlendirilmiş kitaplardan oluşan bir alt küme belirlenmiştir. Elde edilen düzenlenmiş veri seti BXN olarak isimlendirilerek deneysel çalışmalarda kullanılmıştır. BXN veri setinde seyreklik %97,6'ya gerilemiştir. BX, BXA ve BXN veri kümelerinin detaylı bilgileri Tablo 2.3 'te özetlenmiştir. Deneysel çalışmalarda kullanılan BXN veri seti, 756 kullanıcının 1173 kitap için verdiği 21.690 adet derecelendirme değerini içermektedir.

2.4.2. Öneri doğruluğunu ölçeklemek için kullanılan metrikler

Deneysel sonuçları değerlendirmek için Ortalama Mutlak Hata (Mean Absolute Error - MAE) ve Hataların Ortalama Kare Kökü (Root Mean Squre Error - RMSE) ölçütleri kullanılmıştır. Denklem 2.7 ve 2.8 sırasıyla MAE ve RMSE değerlerinin hesaplanması için kullanılan eşitlikler verilmektedir.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - g_i| \tag{2.7}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(p_i - g_i)^2}{n}} \tag{2.8}$$

Denklem 2.7 ve Denklem 2.8'de görülen g_i ve p_i, sırasıyla gerçek kullanıcı derecelendirme değerini ve tahmin edilen derecelendirme değerini ifade etmektedir.

2.4.3. Mahremiyet ölçekleme metriği

Veri setine RKT uygulandıktan sonra elde edilen gizlilik seviyesini ölçmek için Agrawal ve Aggarwal (2001) tarafından önerilen diferansiyel entropi tabanlı mahremiyet ölçekleme metriği kullanılmıştır. Bu yöntemde gizlilik seviyesi, veri setinde bulunan değerlerin gözlemlenme yoğunluğuna bağlı olarak hesaplanmaktadır. Veri setinde rastgele seçilen bir değer *x* ve *x*'in gözlemlenme yoğunluğu *p* olsun. Denklem 2.9'da, *x*'in diferansiyel entropisi *h(x)*'i hesaplamak için kullanılan eşitlik verilmiştir.

$$h(X) = -\int_{\Omega_x} f_x(x)\log_2 f_x(x)dx \tag{2.9}$$

Denklem 2.9'da *x*, Ω_x kümesinde tanımlı olup hesaplanan diferansiyel entropi değeri belirsizliğini ifade etmektedir. Bir *x* değişkeninin belirsizliği ne kadar yüksek olursa tahmin edilmesinin o kadar zor olacağı ve sağlayacağı mahremiyet seviyesinin artacağı varsayılmaktadır. Önerilen yöntemde diferansiyel entropi kullanılarak gizlilik seviyesinin tanımlanması Denklem 2.10'a göre yapılmaktadır.

$$\prod(\mathrm{X}) = 2^{h(\mathrm{x})} \tag{2.10}$$

Bu yöntem ile veri setindeki herhangi bir değerin gizlilik seviyesi hesaplanmaktadır. Maskelenmiş derecelendirme değerlerinin gizliliğinin hesaplanması ise, maskelenmiş verinin sunucu tarafından bilindiği durumda ham kullanıcı verilerinin ne seviyede gizli kalabildiği ile

ilişkili olduğundan koşullu diferansiyel entropi kullanılarak yapılmaktadır. Maskeleme işleminde kullanılan R'nin ortalama koşullu mahremiyetini tanımlayan eşitlik $\Pi(V,P) = 2^{H(V|P)}$ olup, P maskelenmiş derecelendirme vektörünü, V orijinal derecelendirme vektörünü ve $2^{H(V|P)}$ maskelenmiş o derecelendirme verileri için orijinal derecelendirme verilerinin koşullu diferansiyel entropisini ifade etmektedir. Maskelenmiş verilerin açığa çıkması durumunda orijinal verilerin sahip olacağı mahremiyet seviyesi koşullu gizlilik kaybı üzerinden hesaplanmaktadır. Orijinal derecelendirme değerlerinin maskelenmiş derecelendirme değerlerine göre koşullu gizlilik kaybı Denklem 2.11'e göre hesaplanmaktadır (Agrawal ve Aggarwal, 2001).

$$\Pr(V|P) = 1 - 2^{H\left(V\middle|P - H(P)\right)} \tag{2.11}$$

Maskelenmiş verilerin açığa çıkması durumunda orijinal verilerin sahip olacağı mahremiyet seviyesinin hesaplanması için Denklem 2.12 kullanılmaktadır.

$$\Pi(V,P) = \Pi(V) \times (1 - \Pr(V|P)) \tag{2.12}$$

3. BXN VERİ SETİ KULLANILARAK OLUŞTURULAN RKT TABANLI GKOF SİSTEMİ

Bu bölümde BXN veri seti kullanılarak RKT tabanlı bir geleneksel komşuluk tabanlı GKOF sistemi oluşturulmuştur. Çelişen öneri doğruluğu ve mahremiyet hedefleri arasında bir denge bulmak için farklı seviyelerde mahremiyet sağlanarak geleneksel komşuluk tabanlı yaklaşımlar ile elde edilen öneri doğrulukları karşılaştırılmıştır.

3.1. RKT ile Mahremiyet Sağlanması

BXN veri seti Prosedür 2.1'deki işlem adımları takip edilerek maskelenmiştir. Prosedür 2.1 ile ilk olarak veri setindeki derecelendirme değerlerinde standardizasyon sağlamak için z-skor normalizasyonu uygulanmıştır. Bir sonraki adımda, sağlanmak istenen mahremiyet seviyesine göre verilerin ne kadar bozulacağını tanımlayan σ katsayısı belirlenmiştir. σ katsayısı belirlenirken referans alınan aralık ise kullanılan veri seti ve servis sağlayıcının empirik olarak belirleyeceği *sigmaMax* (σ_{max}) katsayısı ile ilişkilidir. BXN veri seti [1-10] aralığındaki derecelendirme değerlerinden oluşmaktadır, dolayısıyla σ değeri (1,10] aralığında değişmektedir. Belirlenen σ değerine göre N dağılım referans alınarak elde edilen rastgele sayı vektörü oluşturulmaktadır. Veri maskeleme işleminin son adımındaysa, z-skor ile normalize edilen derecelendirme değerleri vektörü ile rastgele sayı vektörü toplanarak maskelenmiş derecelendirme değerleri elde edilmektedir. Sonuç olarak elde edilen maskelenmiş derecelendirme değerlerini tutan veri seti BXM olarak adlandırılmıştır.

3.2. Maskelenmiş Derecelendirme Değerleri ile Geleneksel Komşuluk Tabanlı GKOF Sistemlerinde Öneri Üretme Süreci

Maskelenmiş kullanıcı verileri kullanılarak öneri üretmek için öncelikle, Denklem 2.1'de verilmiş olan PK eşitliğini yeniden düzenleyerek elde edilen Denklem 3.1 ile benzerlik değerleri hesaplanmaktadır (Polat ve Du,2005; Bilge ve Polat, 2013).

$$PK_{au} \approx PCC'_{au} = \sum_{i}^{m} z'_{ai} \times z'_{ui} \tag{3.1}$$

Denklem 3.1'de z'_{ai} ve z'_{ui} sırasıyla a ve u kullanıcılarının i ürünü için z-skor normalizasyonu uygulanmış ve maskelenmiş derecelendirme değerleri vektörlerini temsil etmektedir. Denklem 3.1 kullanılarak kullanıcılar arasındaki benzerlik değeri hesaplandıktan sonra, kullanıcıya q ürünü için maskelenmiş veri ile öneri üretme işlemi Denklem 3.2 kullanılarak gerçekleştirilir (Polat ve Du, 2005, Bilge ve Polat, 2012).

$$P_{aq} \approx P'_{aq} = \overline{r_a} + \sigma_a \frac{\Sigma_{u=1}^{k} PK'_{au} \times z'_{uq}}{\Sigma_{u=1}^{k} PK'_{au}} \quad (3.2)$$

Denklem 3.2'de $\bar{r}_a$ kullanıcının ortalama derecelendirme değerini, σ_a kullanıcının değerlendirmelerinin standart sapmasını, PK'_{au} a ve u kullanıcıları arasında hesaplanan benzerliği ve z'_{uq} u kullanıcısının q ürünü için değerlendirmesinin z-skor ile normalize edilmiş versiyonunu temsil etmektedir. Denklem 3.2 kullanılarak, oluşturulan GKOF sisteminde maskelenmiş veri seti kullanılarak öneri doğruluğundan kabul edilebilir seviyede kayıpla öneri üretilebilmektedir (Polat ve Du, 2005; Bilge ve Polat, 2013).

3.3. Deneysel Yaklaşımlar ve Elde Edilen Sonuçlar

Maskelenmiş veri seti üzerinde tahmin üretme aşamasında en-iyi N öneri üretme yaklaşımı kullanılmıştır. Bu yöntemde, a'ya en yüksek benzerliği gösteren N adet kullanıcının derecelendirme değerleri, a'ya öneri üretmek için referans alınmaktadır. Yapılan çalışmada N değeri deneysel olarak belirlenmiştir ve diğer tüm deneylerde sabit tutulmuştur. N değerinin belirlenmesi için, maskelenmemiş gerçek kullanıcı verisi üzerinde komşu sayılarıyla deneyler yapılmış ve en yüksek doğruluğu veren komşu sayısı sabit N değeri olarak belirlenmiştir. Gerçek kullanıcı derecelendirme değerleri üzerinde yapılan deneylerde değişken komşuluk değerleri ile elde edilen öneri doğrulukları Tablo 3.1'de verilmiştir.

Tablo 3.1. Gerçek Derecelendirme Değerleri ile Yapılan Deneylerde Değişken Komşuluk Sayısının Tahmin Üretme Doğruluğuna Etkisi

Komşu Sayısı (*N*)	**Öneri Doğruluğu (MAE)**
5	1,282
10	1,227
15	1,237
30	1,243
50	1,236

Tablo 3.1'de sonuçları sunulan deneysel çalışma kapsamında maskelenmemiş gerçek derecelendirme değerleri kullanılarak yapılan tek deney olup, oluşturulan GKOF sisteminde geleneksel komşuluk tabanlı yaklaşım ile en yüksek doğruluk seviyesine ulaşacak komşu sayısının tespit edilmesi amaçlanmıştır. Elde edilen sonuçlar göz önünde bulundurulduğunda, komşu sayısı 10 olarak seçildiğinde en yüksek doğrulukla tahmin üretme işlemiş gerçekleşmiştir. Bu nedenle N değeri tüm deneylerde 10 olarak sabit tutulmuştur

Tahmin oluşturma sürecinde, örneklem seçiminde ortaya çıkabilecek rastlantısallıkların öneri doğruluğu üzerindeki olumsuz etkisini hafifletmek için 10 kat çapraz doğrulama (10-fold cross-validation) metodolojisi kullanılmıştır. 10 kat çapraz doğrulama, kullanıcıların vermiş olduğu derecelendirme değerlerini 10 eşit parçaya bölen, ardından her bir parçayı sırasıyla test seti olarak kullanırken, geri kalan 9 parçayı eğitim seti olarak kullanan bir yöntemdir. Bu işlem, her bir parçanın test seti olarak kullanıldığı 10 farklı eğitim ve test kombinasyonu sağlayarak sistemi eğitme aşamasında örneklem seçiminde ortaya çıkabilecek rastlantısallıkların etkisini hafifletmektedir.

3.4. Sigma Değerinin Öneri Doğruluğuna Etkisi

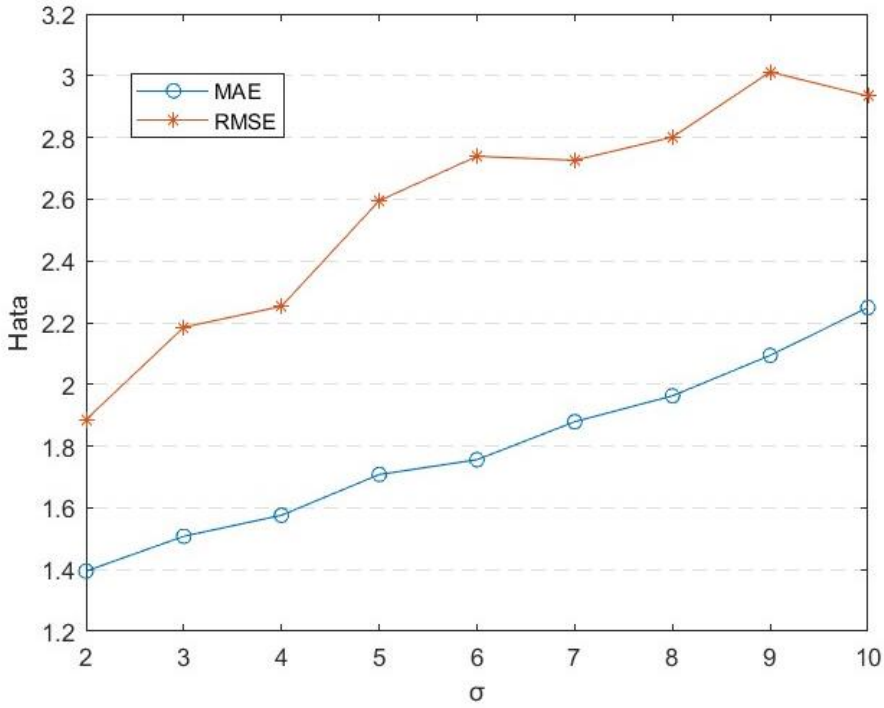
RKT'de mahremiyet seviyesini belirleyen sigma (σ) katsayısının öneri doğruluğu üzerindeki etkisinin test edilmesi için BXN veri setinin derecelendirme ölçeğine göre (1,10] aralığında değerler seçilmiştir. σ katsayısı ile yapılan her deneyde gerçek kullanıcı davranışlarını simüle edebilmek için belirlenen σ katsayısı [0,1] aralığı içerisinde üretilen rastgele bir sayı ile çarpılmaktadır. Bu sayede sistem tarafından seçilen en yüksek σ katsayısı üzerinden hesaplanan farklı bir σ değeri ile kullanıcı derecelendirmeleri maskelenebilmektedir. Hem yapılan bu işlem sonucunda ortaya çıkan rastlantısallık hem de RKT ile üretilen rastgele sayı vektöründeki rastlantısallık nedeniyle ortaya çıkabilecek sapmaların öneri doğruluğu üzerindeki etkisinin hafifletilmesi için deney koşulları sabit tutularak her bir senaryoda deneyler 5 kez tekrar edilmiştir. Tekrar edilen 5 deney sonucunda elde edilen hata değerlerinin aritmetik ortalaması alınarak nihai sonuç olarak Tablo 3.2'de sunulmuştur.

Tablo 3.2. Farklı σ Değerleri ile Maskelenen Veriler ile Üretilen Önerilerin Hata Değerleri

σ Değeri	MAE	RMSE
2	1,395	1,886
3	1,508	2,185
4	1,576	2,253
5	1,708	2,595
6	1,756	2,739
7	1,879	2,726
8	1,963	2,801
9	2,094	3,801
10	2,249	3,012

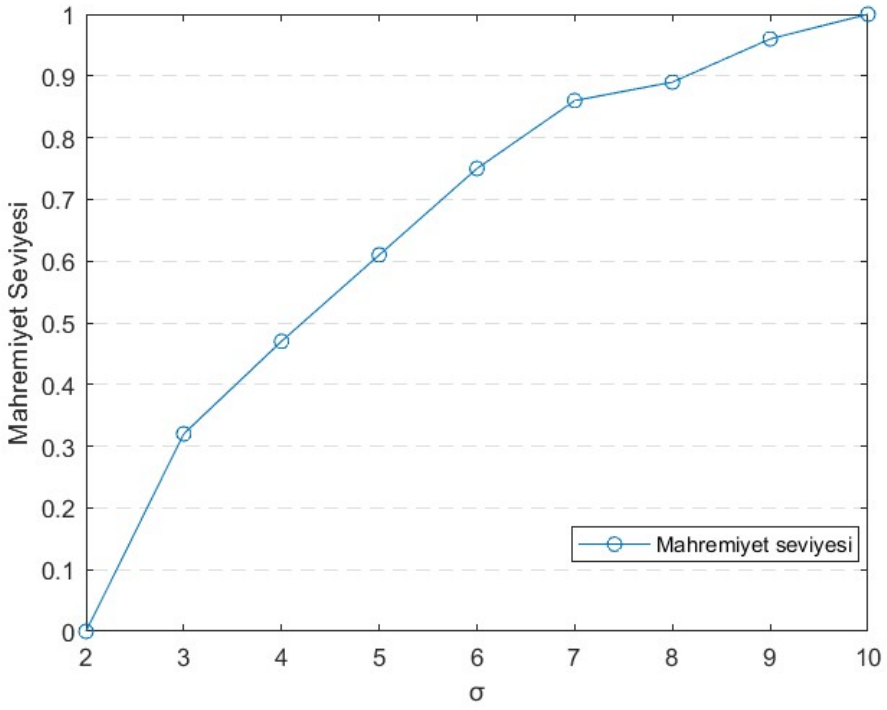
Tablo 3.2'de görüldüğü gibi, en düşük MAE değeri en küçük σ değeri olan 2 için elde edilirken ve en düşük tahmin doğruluğu σ =10 değeri için elde edilmiştir. σ değeri büyüdükçe, normal dağılıma göre üretilen rastgele sayıların değer aralığı da büyüdüğünden gerçek

derecelendirme değerleri üzerindeki bozulma da σ değeri ile ilişkili olarak artmaktadır. Bu durum kullanıcıların gerçek derecelendirme değerleri üzerinde daha fazla bozulmaya neden olduğundan sistemin ürettiği önerilerin doğruluğunda azalmaya ve daha yüksek hataya neden olur. Elde edilen hata değerleri incelendiğinde, $\sigma = 2$ ve $\sigma = 10$ için maskelenen veriler kullanılarak üretilen öneriler arasında MAE ve RMSE değerleri göz önünde bulundurulduğunda, hata oranlarının σ değeriyle birlikte artış gösterdiğinin gözlemlenebildiği grafik Şekil 3.1'de sunulmuştur.



Şekil 3.1. σ Değerine Göre Değişen MAE ve RMSE Değerleri

Farklı σ seviyelerinde sağlanan mahremiyeti ölçmek için Bölüm 2.4'te detayları verilen diferansiyel entropi tabanlı gizlilik ölçme metriği kullanılmıştır. Denklem 2.10 kullanılarak farklı σ seviyeleri için yapılan mahremiyet analizinin sonuçlarının karşılaştırılması için elde edilen değerlere 0-1 normalizasyonu uygulanmıştır. Bu sayede elde edilen değerler 0-1 aralığına getirilerek karşılaştırılabilirliği kolaylamıştır. Elde edilen mahremiyet seviyeleri Şekil 3.2'de sunulmuştur. σ değeri arttıkça üretilen rastgele sayı vektörünün değer aralığının genişlemesi verideki bozulmanın da paralel olarak artmasına neden olur. Verideki bozulma arttığında veri setinden çıkarılabilecek bilgi miktarı azalmaktadır ve kullanıcıların derecelendirme değerlerinin tahmin edilebilmesi zorlaşmaktadır. Böylece, kullanıcının mahremiyet seviyesi de yükselmektedir. σ değerinin yükselmesiyle birlikte artan mahremiyet seviyesi Şekil 3.2'de görülmektedir.



Şekil 3.2. σ Değerine Göre Mahremiyet Seviyesinde Oluşan Değişiklik

3.5. Sonuçlar

Bu bölümde RKT kullanılarak farklı σ değerlerinde elde edilen mahremiyet seviyeleri ve tahmin doğrulukları üzerindeki etkisi sunulmuştur. Gizliliği sağlamak amacıyla kullanılan yöntem RKT, BXN veri setine bağlı olarak (2,10] aralığında değişen seviyelerde σ değerlerinde uygulanmıştır. Değişen σ değerine göre öneri doğruluğundaki kayıp MAE ve RMSE kullanılarak hesaplanmıştır. σ değeri yükseldikçe artan hata değerleri gerçek derecelendirme değerlerinin orijinal halinden uzaklaştırılarak gizlilik sağlanmasının öneri doğruluğu ile doğrudan ilişkili olduğu ve öneri doğruluğunu olumsuz etkilediği sonucuna ulaşılmaktadır. Bununla birlikte artan σ seviyesiyle birlikte kullanıcıya sağlanan gizlilik seviyesinin de artış görülmektedir. σ değerinin artmasıyla birlikte sağlanan gizliliğin yükselmesi, derecelendirme değerlerinin dağılımının daha geniş bir aralığa dağılması ve herhangi bir değer için hesaplanabilecek diferansiyel entropinin yükselmesi ile ilişkilidir.

GKOF sistemlerinde kullanıcının veri mahremiyeti sağlanırken aynı zamanda da tahmin oluşturma doğruluğunda ortaya çıkabilecek kayıplarının hafifletilmesi hedeflenmektedir. Şekil 3.1'de gösterilen gizlilik seviyesindeki artışa göre MAE'de gözlemlenen artışın 7'den büyük değerlerde yavaşladığı görülmüştür. Şekil 3.2'de gösterilen mahremiyet grafiğine göre ise, σ değeri 7 seviyesine gelene kadar mahremiyet seviyesinin hızla yükseldiği ve 7'den daha yüksek

değerlerde mahremiyet seviyesindeki artışın yavaşladığı görülmektedir. Bu nedenle BXN için ideal σ parametresi gizlilik ve tahmin doğruluğu arasında dengenin sağlanması için 7 olarak belirlenmektedir ve gizlilik seviyesini olabildiğince yüksek tutarken makul bir tahmin doğruluğu kaybı ile öneri üretilmektedir. Deneysel çalışmalardan elde edilen sonuçlara göre, artan σ değeri ile birlikte veri gizliliği artarken öneri doğruluğu azalmaktadır. Bu nedenle geleneksel komşuluk tabanlı GKOF sistemleri yerine daha yüksek doğruluk seviyelerinde tahmin üretebilecek yeni yaklaşımlara ihtiyaç duyulmaktadır (Yargıç ve Açıl, 2022).

4. OTOKODLAYICI TABANLI GKOF SİSTEMİ

Bu bölümde, RKT ile maskelenen verilerin geleneksel komşuluk tabanlı OF sistemlerinde ortaya çıkardığı öneri doğruluğundaki azalma problemini hafifletmek için otokodlayıcılar kullanılmıştır. Otokodlayıcıların ÖS problemlerindeki başarısı gözlemlenmiş ve pek çok çalışma ile desteklenmiş olsa da RKT ile maskelenen veri setleri üzerinde otokodlayıcı tabanlı yaklaşımların öneri doğruluğu üzerine etkisi incelenmemiştir. Literatürdeki maskelenmemiş derecelendirme verileri kullanılarak yapılan çalışmalarda öneri doğruluğunun geleneksel komşuluk tabanlı tahmin üretme yaklaşımlarına göre daha başarılı sonuçlar verdiği görülmüştür. Otokodlayıcı, derecelendirme değerlerindeki gizli ve doğrusal olmayan özellikleri öğrenmektedir ve bu özelliklere dayalı olarak öneri üretmektedir. Bu tez kapsamında otokodlayıcı tabanlı öneri üretme yaklaşımında RKT ile maskelenmiş veriler kullanıldığında da öneri doğruluğunun yükseldiği görülmüştür. Kullanılan yöntemde Bölüm 3'te tanımlanan yaklaşım ile maskelenen gerçek kullanıcı verileri otokodlayıcı girişine verilmiştir ve üretilen önerilerin doğrulukları geleneksel komşuluk tabanlı GKOF sistemi tarafından üretilen önerilerin doğruluklarıyla MAE ve RMSE metrikleri kullanılarak karşılaştırılmıştır. Öneri doğruluklarını karşılaştırma işlemi kullanılan aktivasyon fonksiyonları ve katman sayılarına göre farklı senaryolar altında test edilmiştir ve en yüksek doğruluk seviyesine erişen model belirlenmiştir.

4.1. ÖS'nde Otokodlayıcılar

Geleneksel komşuluk tabanlı GKOF sistemlerinde geleneksel veri maskeleme yöntemlerinden biri olan RKT'nin kullanıcı mahremiyeti sağlarken öneri doğruluğu üzerinde olumsuz etkileri olduğu Bölüm 3'te sunulmuştur. RKT ile veri maskeleme işlemi kullanıcıların gerçek derecelendirme değerlerinde bozulmaya neden olduğu için sistemin öneri üretme başarısını olumsuz yönde etkilemektedir. Yapılan çalışmada öneri doğruluğundaki kayıpları azaltmak için otokodlayıcılar kullanılmaktadır.

Genellikle boyut küçültme ve gizli özelliklerin öğrenilmesi için kullanılan denetimsiz bir öğrenme modeli olan otokodlayıcılar, ÖS'nde sıklıkla kullanılan DÖ yaklaşımlarından biridir. Otokodlayıcı tabanlı ÖS'nin, geleneksel OF sistemlerine göre daha yüksek öneri doğruluğu sağlamaktadır (Unger, 2015; Unger vd., 2016; Zhang vd., 2020). ÖS için özelliklerin çıkarılması ve kullanıcı tarafından oluşturulan verilerin ve ürün verilerinin modellenmesi kullanıcılara sunulan önerilerin kalitesinde büyük iyileştirmeler sağlayabilmektedir

(Karatzoglou ve Hidasi, 2017). Özellikle otokodlayıcılar, özellik çıkarma amacıyla kullanıldığında öneri doğruluğunu yükselten yöntemlerdendir (Strub vd., 2016).

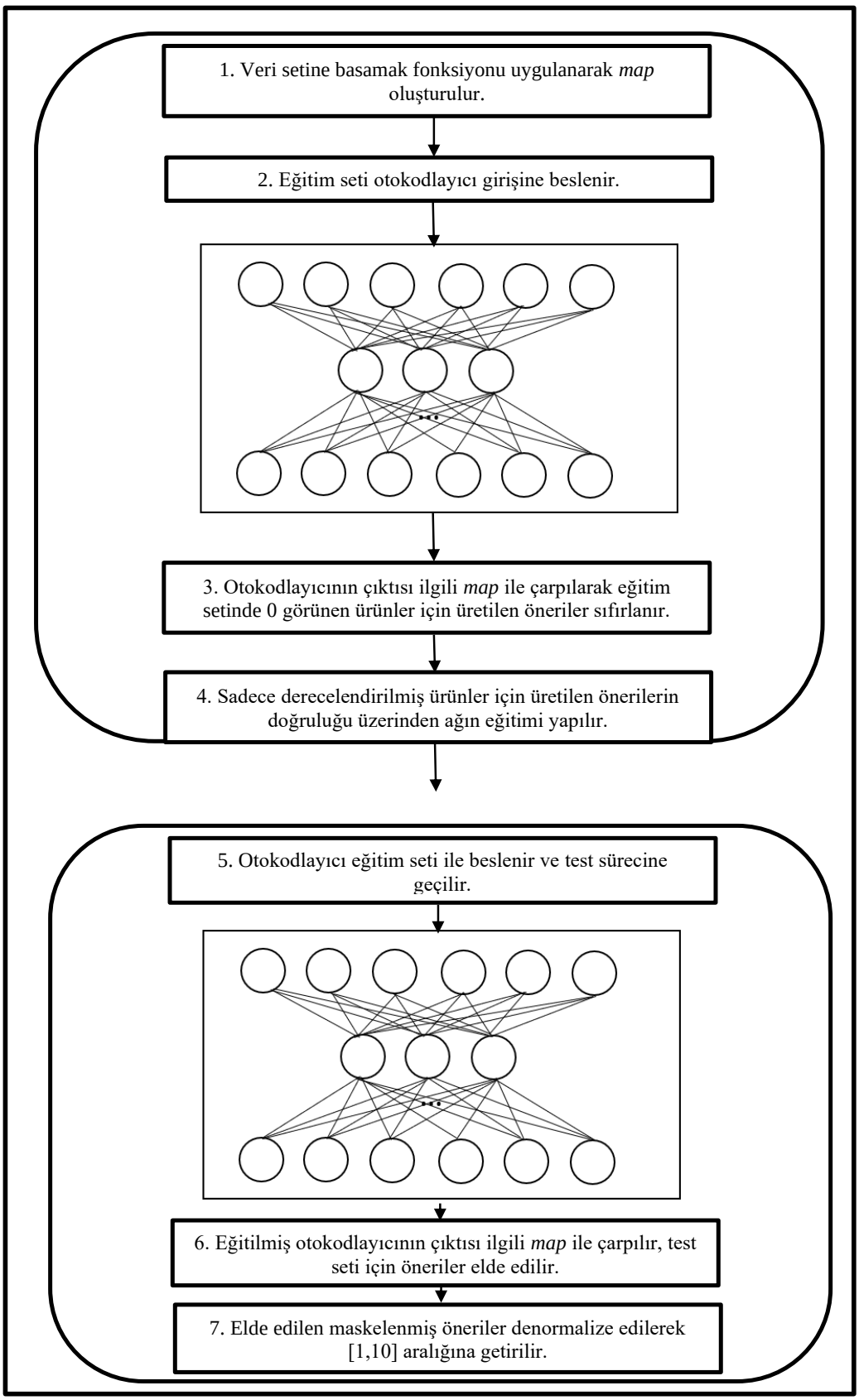
4.2. Otokodlayıcı ve RKT Tabanlı Gizliği Koruyan ÖS

RKT ile maskelenmiş olan BXM veri seti kullanılarak otokodlayıcı tabanlı yöntemle kullanıcılara öneriler üretilmiş ve öneri doğrulukları incelenmiştir. Kullanılan veri seti Tablo 4.1'de temsili olarak gösterilmiştir. $u_1 - u_{756}$ arasında kullanıcılar ve $i_1 - i_{1173}$arasında ürünler bulunduran BXM veri seti, kullanıcıların ürünler için verdiği derecelendirme değerlerinin maskelenmiş hallerini ($\{(r_{1,1}, r_{1,2}, \dots, r_{1173}),\ (r_{756,1}, r_{756,2}, \dots, r_{756,1173})\}$ olarak temsil edilmiştir) bulundurmaktadır. Kullanılan otokodlayıcı mimarisinde giriş katmanındaki düğüm sayısı, mevcut veri setindeki ürün sayısına eşittir ve 1173 adettir. Veri setinin otokodlayıcıya beslenmesi Tablo 4.1'de temsili olarak gösterilen veri setinin satırlarında görülen maskelenmiş derecelendirme vektörleri ile gerçekleştirilmektedir.

Tablo 4.1. BXM Veri Setinin Temsili Gösterimi

	$\boldsymbol{i_1}$	$\boldsymbol{i_2}$	**…**	$\boldsymbol{i_{1173}}$
$\boldsymbol{u_1}$	$r_{1,1}$	$r_{1,2}$	…	$r_{1,1173}$
…	…	…	…	…
$\boldsymbol{u_{756}}$	$r_{756,1}$	$r_{756,2}$	…	$r_{756,1173}$

Tablo 4.1'de görülen her bir satırdaki kullanıcı derecelendirme vektörleri çoğunluğu 0 değerlerinden oluşan seyrek vektörlerdir. Otokodlayıcı bu seyrek verileri girdi olarak almaktadır ve kodlayıcı kısmında gizli ve doğrusal olmayan özelliklerini öğrenerek daha küçük boyutlu bir temsilini oluşturmaktadır. Ardından kod çözücü katmana boyutu azaltılmış veriler aktarılarak giriş verisinin yeniden oluşturulması sağlanmaktadır. Bu aşamada yeniden oluşturulan veriler, sistemin giriş verilerinden öğrendiği özelliklere göre ürettiği öneriler olmaktadır. Ancak otokodlayıcının eğitimi sırasında değerlendirilmemiş ürünler için üretilen tahminlerin etkili olmaması gerekmektedir (Batmaz, 2019; Sedhain vd., 2015; Strub vd., 2015). Eksik verilerden kaynaklanan hataların otokodlayıcının eğitiminde görmezden gelinmesi için BXM veri setine basamak fonksiyonu uygulanarak elde edilen, dolu olan değerlerin "1" ve boş olan değerlerin "0" olarak bulunduğu *map*'ten faydalanılır (Batmaz, 2019). BXM veri seti kullanılarak otokodlayıcıların eğitilmesi ve eğitilmiş otokodlayıcılar kullanılarak öneri üretme işlem adımları Şekil 4.1.'de gösterilmiştir.



Şekil 4.1. Otokodlayıcı Tabanlı ÖS'nin Tahmin Üretme İşlem Adımları

Şekil 4.1'de verilen otokodlayıcı eğitimi ve otokodlayıcı ile öneri üretme adımları şu şekilde detaylandırılmaktadır:

BXM veri seti kullanıcıların değerlendirmediği ürünler için 0 değeri içerdiğinden oldukça seyrek vektörlerden oluşmaktadır. Otokodlayıcının çıkışından elde edilen veriler ise kullanıcıların hem derecelendirdiği hem de derecelendirmediği ürünler için üretilmiş öneriler içermektedir. Veri setinin orijinal halinde derecelendirilmemiş olan ürünlerin otokodlayıcının çıkış verilerinde de tespit edilebilmesi için BXM veri setine basamak fonksiyonu uygulanarak oluşturulan *map* kullanılmaktadır. Ağın eğitiminde, otokodlayıcının girişine beslenen eğitim verileri ile ağ, veri setinin gizli özelliklerini belirlemektedir ve buna bağlı olarak veriyi çıktı katmanında yeniden oluşturmaktadır. Ağın eğitiminde kullanılan geri yayılım algoritması ile çıktı katmanında elde edilen verilerden yüksek hata değerine neden olanlar için ağ cezalandırılmaktadır ve kayıp fonksiyonu olarak ortalama kare hata kullanılmaktadır. Ancak ağın cezalandırılması sadece kullanıcının derecelendirdiği ürünler için gerçekleştirilmelidir. Bu nedenle eğitim aşamasında otokodlayıcının ürettiği öneriler *map*'in ilgili kısmı ile çarpılarak orijinal halinde 0 olan veriler yeniden 0 haline getirilmektedir. Bu sayede ağın eğitiminde ağırlıklar güncellenirken boş olan değerleri göz ardı edilebilmektedir. Eğitim setinde bulunan her bir maskelenmiş derecelendirme vektörü için bu süreç devam eder ve sürecinin ardından elde edilen eğitilmiş otokodlayıcıya eğitim verileri beslenerek ağın öğrendiği gizli özelliklere göre tahmin üretebilme yeteneği incelenir. Eğitilmiş otokodlayıcıya beslenen eğitim verileri de boş değerler için de öneriler içerecek şekilde yeniden oluşturulduğu için çıktısı test kümesine ait *map* ile çarpılır ve sadece derecelendirilmiş ürünler için üretilen öneriler görülür. Bu aşamada elde edilen öneriler maskelenmiş derecelendirme değerleri halindedir ve veri setinin orijinal derecelendirme aralığında ifade edilebilmesi için denormalizasyon işlemi yapılmalıdır. Denormalizasyon işlemi, öneri üretilen kullanıcının standart sapma ve ortalama verileri kullanılarak gerçekleştirilir. Maskelenmiş derecelendirmeler içeren vektör standart sapma değeri ile çarpılır ve ortalama değeri eklenerek [1,10] aralığında ifade edilmesi sağlanır. Veri setinin değerlendirme ölçeği dahiline getirilen derecelendirme değerleri için MAE ve RMSE hata metrikleri kullanılarak elde edilen öneri doğruluğu incelenir.

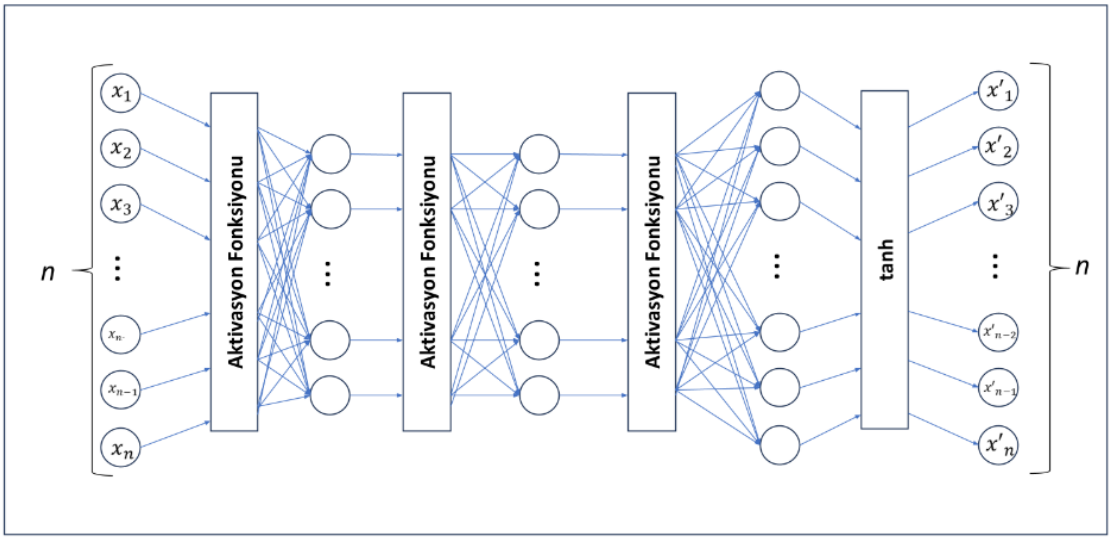
4.2.1. Otokodlayıcı mimarisi ve deneysel metodoloji

Otokodlayıcı eğitiminde Bayesian normalleştirme geri yayılımı kullanılmıştır (Batmaz ve Kaleli, 2019). Deneylerde 5 kat çapraz doğrulama kullanılmıştır. Otokodlayıcıların eğitimi için (1,10] aralığında değişen σ değerleri ile maskelenen BXM veri seti 5 gruba ayrılmıştır ve 4 parçası eğitim 1 parçası test seti olacak şekilde veri setinin tamamı hem eğitim hem de test

için kullanılmıştır. Bu sayede eğitim ve test verilerinde çeşitlilik sağlanarak farklı kombinasyonlardaki kümelerle deneyler yapılması sağlanmıştır. Deneyler her bir σ değeriyle maskelenen veri setleri için ayrı ayrı gerçekleştirilmiş ve RKT ile maskelenen rastgele sayı vektöründeki rastgelelikten kaynaklanabilecek sapmaların öneri doğruluğu üzerindeki etkisini azaltmak için 5 kez tekrar edilmiştir. Her bir tekrar için elde edilen hata oranlarının aritmetik ortalaması alınarak her σ değeri için tek bir hata değeri nihai sonuç olarak sunulmuştur. Ayrıca aşırı uyma problemini engellemek amacıyla L2 düzenlileştirmesi kullanılmıştır.

Otokodlayıcı her bir kullanıcı derecelendirme vektörünü girdi olarak almaktadır ve kodlayıp yeniden oluşturarak eğitilmektedir. Bu eğitim sürecinde otokodlayıcının ağırlık değerlerinin güncellenmesi sadece kullanıcının derecelendirme verdiği değerler üzerinden hesaplanmalıdır, vektörde herhangi bir derecelendirme değeri olmadığından boş olan hücreler ağırlıkların güncellenmesinde etkili olmamalıdır. Bu nedenle oluşturulan otokodlayıcının eğitiminde dolu değerlerin daha etkili olması gerektiği için özel bir kayıp fonksiyonuna ihtiyaç duyulmaktadır. Batmaz ve Kaleli tarafından sunulan yaklaşımda kullanılan kayıp fonksiyonu ve otokodlayıcı mimarisi yapılan çalışmada referans alınmıştır (Batmaz ve Kaleli, 2019).

Otokodlayıcının iç katmanlarında hiperbolik tanjant (*tanh*), üstel doğrusal birim (*elu*), ölçeklendirilmiş üstel doğrusal birim (*selu*) ve doğrusal (*linear*) aktivasyon fonksiyonları kullanılmıştır. Çıkış katmanında *tanh* aktivasyon fonksiyonu sabit tutulmuştur. Kullanılan otokodlayıcının yapısı Şekil 4.2'de verilmiştir.



Şekil 4.2. Kullanılan Otokodlayıcının Yapısı

Deneysel çalışmalarda kullanılan 2 gizli katman içeren otokodlayıcı mimarisi Şekil 4.2'de görülmektedir. Her bir kullanıcının derecelendirme değerleri vektörü giriş katmanından

otokodlayıcıya beslenir. Giriş katmanından alınan derecelendirme vektörü öncelikle aktivasyon fonksiyonuna girer. Aktivasyon fonksiyonu ile belirli bir aralığa sınırlandırılan veriler kodlayıcı katmana iletilir. Bu katman giriş katmanına göre daha az sayıda düğüm içermektedir, veri setinin özelliklerinin çıkarılması ve boyutunun indirgenmesi bu katmanda gerçekleşmektedir. Boyutu indirgenen BXM veri seti bu katmanın çıktısıdır ve kod çözücü katmana aktarılmadan önce tekrar aktivasyon fonksiyonuna girmektedir. Kod çözücü katmanda BXM veri seti yeniden oluşturulmaktadır. Bu nedenle çıkış katmanındaki nöron sayısı giriş katmanındaki nöron sayısına eşittir.

4.3. Aktivasyon Fonksiyonları ve Öneri Doğruluklarına Etkileri

Bu çalışmada, iki gizli katmanlı otokodlayıcı mimarisinin iç katmanlarında *tanh, elu, selu* ve *linear* aktivasyon fonksiyonları test edilmektedir. Çıktı katmanında iç katmanlardaki aktivasyon fonksiyonlarından bağımsız olarak *tanh* aktivasyon fonksiyonu sabit tutulmuştur. *tanh* aktivasyon fonksiyonu girdileri [-1,1] aralığına sıkıştırarak çıktı veren bir aktivasyon fonksiyonudur (Dubey vd., 2022). Kullanılan veri setine yapılan ön işlemlerde z-skor normalizasyonu da bulunduğundan giriş verileri [-1,1] aralığına normalize edilmektedir ve *tanh* fonksiyonu için uygun değer aralığındadırlar. Bu nedenle çıktı katmanında *tanh* aktivasyon fonksiyonu sabit tutulmuştur, iç katmanlardaki aktivasyon fonksiyonları değiştirilerek öneri doğruluğuna etkileri gözlemlenmiştir. Aktivasyon fonksiyonu, kodlayıcı katmanda elde edilen çıktıları sınırlar ve bu sayede elde edilen özelliklerin belirli bir aralıkta kalmasını sağlar.

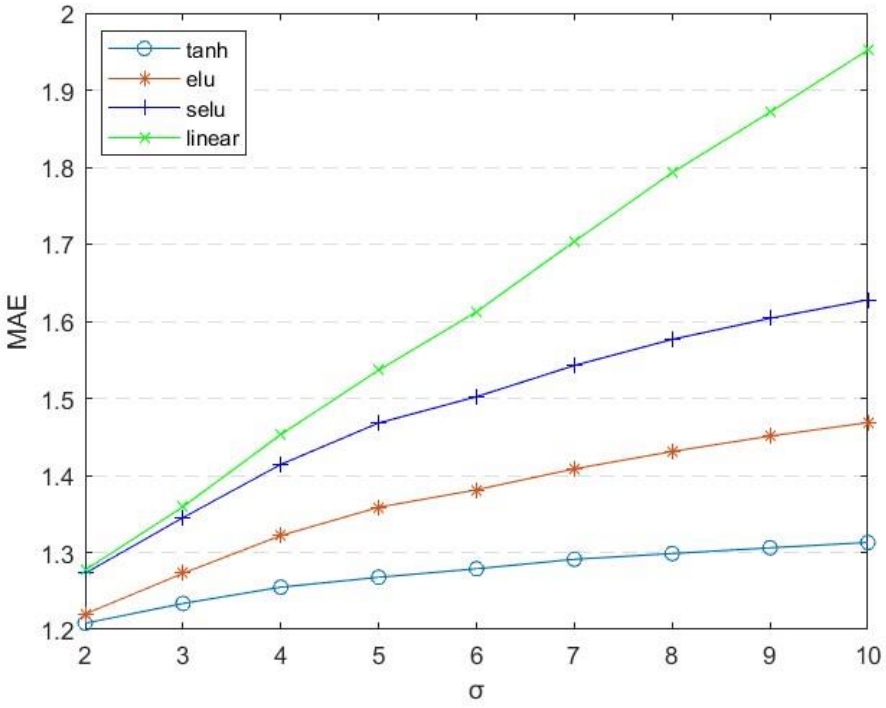
elu aktivasyon fonksiyonu, negatif değerlerin temsilini destekleyen ve veri setindeki aykırı değerleri tolere edebilen bir aktivasyon fonksiyonudur. Bu nedenle ÖS'nde kullanılan veri setlerine uygun olabilmekte ve eğitim sürecini olumlu etkileyebilmektedir. Çıkış değerleri giriş değerlerinin aralığına göre değişkenlik göstermekle birlikte negatif değerlerdeki girişler 0'a yaklaşmaktadır (Nwankpa vd., 2018; Dubey vd., 2022). *selu* aktivasyon fonksiyonu *elu*'nun özel bir versiyonudur ve hız bakımından performansı yükseltmek amacıyla geliştirilmiştir (Klambauer vd., 2017). DÖ mimarilerinde hızlı yakınsama ve sürekli öğrenme özelliği ile öne çıkar (Ding, Qian ve Zhou, 2018). Çıkış aralığı giriş değerlerine bağlı olarak değişkenlik göstermekle birlikte negatif değerler daha yüksek değerlere doğru kaymaktadır. *linear* aktivasyon fonksiyonu ise giriş değerini doğrudan çıkış değeri olarak aktarmaktadır. Otokodlayıcı tabanlı yaklaşımda giriş verisinin çıktı katmanında yüksek doğrulukta yeniden elde edilmesi amaçlandığı için verilerin sahip oldukları doğrusal olmayan özelliklerin tespit edilmesi önemlidir. İç katmanlarda kullanılan *linear* aktivasyon fonksiyonu ağın öğrenme kapasitesini sınırlar ve karmaşık yapıları öğrenmesini engeller (Batmaz ve Kaleli, 2019). Bu

durum göz önünde bulundurulduğunda *linear* aktivasyon fonksiyonunun yüksek öneri doğruluğu sağlayamayacağı öngörülebilir ancak diğer aktivasyon fonksiyonlarını sağladıkları öneri doğruluğu performansını ve aktivasyon fonksiyonlarını öneri doğruluğu üzerindeki etkisini gözlemlemek için bir referans noktası sağlayacaktır.

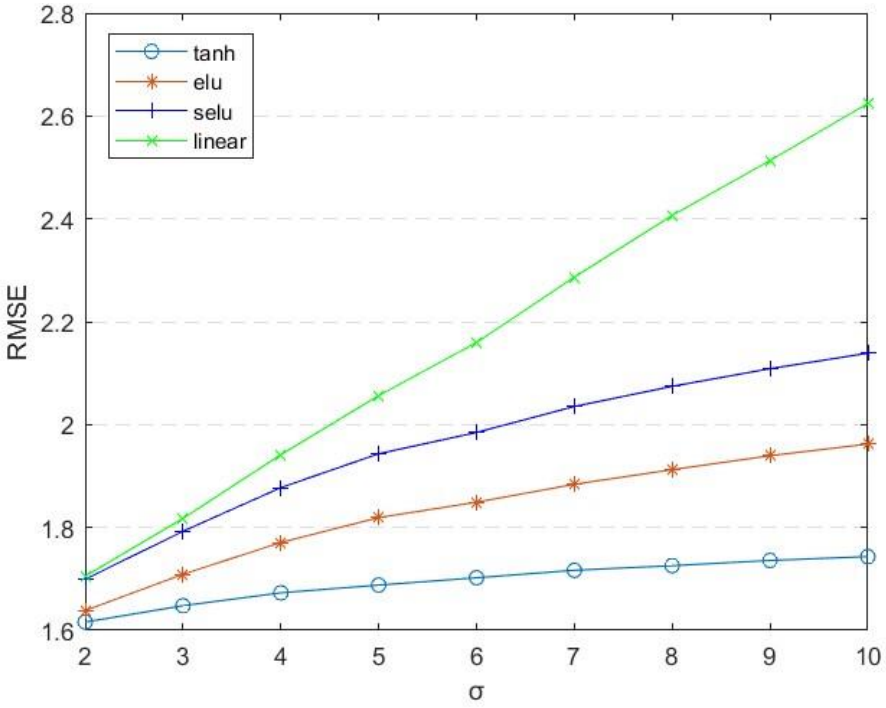
Tablo 4.2. Farklı σ Değerleriyle Maskelenmiş Veri Setlerindeki Öneri Doğruluğunda Farklı Aktivasyon Fonksiyonlarıyla 2 Katmanlı Mimaride Gözlemlenen Hata Değerleri

	tanh		***elu***		***selu***		***linear***	
	MAE	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
$\boldsymbol{\sigma = 2}$	1,208	1,616	1,220	1,639	1,273	1,699	1,277	1,705
$\boldsymbol{\sigma = 3}$	1,234	1,648	1,273	1.709	1,345	1,793	1,359	1,818
$\boldsymbol{\sigma = 4}$	1,255	1,673	1,322	1,779	1,414	1,878	1,453	1,942
$\boldsymbol{\sigma = 5}$	1,268	1,688	1,359	1,819	1,468	1,943	1,537	2,056
$\boldsymbol{\sigma = 6}$	1,279	1,702	1,381	1,849	1,502	1,985	1,612	2,159
$\boldsymbol{\sigma = 7}$	1,291	1,717	1,409	1,884	1,543	2,035	1,704	2,287
$\boldsymbol{\sigma = 8}$	1,299	1,726	1,431	1,912	1,577	2,074	1,791	2,407
$\boldsymbol{\sigma = 9}$	1,306	1,736	1,451	1,940	1,604	2,109	1,871	2,514
$\boldsymbol{\sigma = 10}$	1,313	1,743	1,469	1,962	1,628	2,139	1,952	2,624

İç katmanlarda değişen aktivasyon fonksiyonları kullanılarak 2 gizli katmanlı otokodlayıcı mimarisi ile yapılan deneylerde elde edilen hata değerleri Tablo 4.2'de verilmiştir. Tablo 4.2'de, en iyi öneri doğruluğunun *tanh* ile en düşük öneri doğruluğunun ise *linear* aktivasyon fonksiyonu kullanıldığında elde edildiği görülmektedir. Bu durum, *tanh* aktivasyon fonksiyonunun veri setindeki değerlerle uyumlu aralıkta çıkış üretmesi ile açıklanabilir. Giriş değerlerine uyumlu bir aralıkta sınırlandırılan çıkış değerleri daha yüksek doğrulukla tahmin edilebilmekte, gerçek değerlerden daha az sapma göstermektedir. *linear* aktivasyon fonksiyonu ise verileri belirli bir aralıkta sınırlamaz ve veri setindeki gizli özelliklerin çıkarılmasını destekleyici bir avantaj sağlamaz. Ağın eğitiminde kullanılan geri yayılım algoritması türeve dayalı çalışmaktadır, doğrusal bir fonksiyon kullanıldığında türev sabit kaldığından dolayı ağın öğrenmesi durumu söz konusu olmamaktadır (Apicella vd., 2021). *elu* aktivasyon fonksiyonu ise negatif değerler de üretebilmektedir, bu nedenle linear aktivasyon fonksiyonuna göre daha hızlı yakınsama sağlar (Clevert, 2015). Bu nedenle *elu* ve *selu* aktivasyon fonksiyonları *linear* aktivasyon fonksiyonuna göre daha yüksek öneri doğruluğu sağlayabilmiş ancak *tanh* aktivasyon fonksiyonu giriş verilerine daha uyumlu bir aralık sağladığı için hala daha düşük hatayla tahmin üretmektedir. Farklı aktivasyon fonksiyonlarının neden olduğu hata değerleri Şekil 4.3'te MAE cinsinden, Şekil 4.4'te RMSE cinsinden görülmektedir.



Şekil 4.3. Farklı Aktivasyon Fonksiyonları Kullanılarak Yapılan Deneylerde Her Bir σ Seviyesi İçin Elde Edilen MAE



Şekil 4.4. Farklı Aktivasyon Fonksiyonları Kullanılarak Yapılan Deneylerde Her Bir σ Seviyesi İçin Elde Edilen RMSE

Şekil 4.3 ve Şekil 4.4'te görüldüğü gibi kullanılan aktivasyon fonksiyonları üretilen tahminlerin doğruluğunda önemli rol oynamaktadır. Grafikte görüldüğü üzere, en düşük σ seviyesinde *tanh* aktivasyon fonksiyonu kullanıldığında en yüksek öneri doğruluğu elde edilirken, en yüksek σ seviyesinde *linear* aktivasyon fonksiyonu kullanıldığında en düşük öneri doğruluğu elde edilmiştir ve bu ikisi arasındaki fark MAE metriğine göre ~%61 ve RMSE metriğine göre ~%62 düzeylerindedir. Tüm aktivasyon fonksiyonları ile yapılan deneylerde hem MAE hem de RMSE değerlerinde σ değeri arttıkça daha yüksek hata değerlerine ulaşıldığı görülmüştür. Bu durum en fazla *linear* aktivasyon fonksiyonu kullanıldığında belirgin hale gelirken, σ değerinin artışına bağlı olarak hatadaki yükselme oranı en az *tanh* aktivasyon fonksiyonu kullanıldığında görülmüştür. Kullanılan aktivasyon fonksiyonları ağın problemi öğrenmesinde ve problemi çözme yeteneğini geliştirmesinde önemli role sahiptir. Artan σ değeri ile birlikte daha geniş bir aralıkta dağılım gösteren verilerden öğrenilen özelliklere göre üretilen öneriler gerçek verilerden mutlak mesafe olarak daha uzak olacaktır. Bu durum σ değeri arttıkça MAE ve RMSE değerinin artış göstermesinin nedeni olarak gösterilebilmektedir.

4.4. Katman Sayısının Öneri Doğruluğuna Etkileri

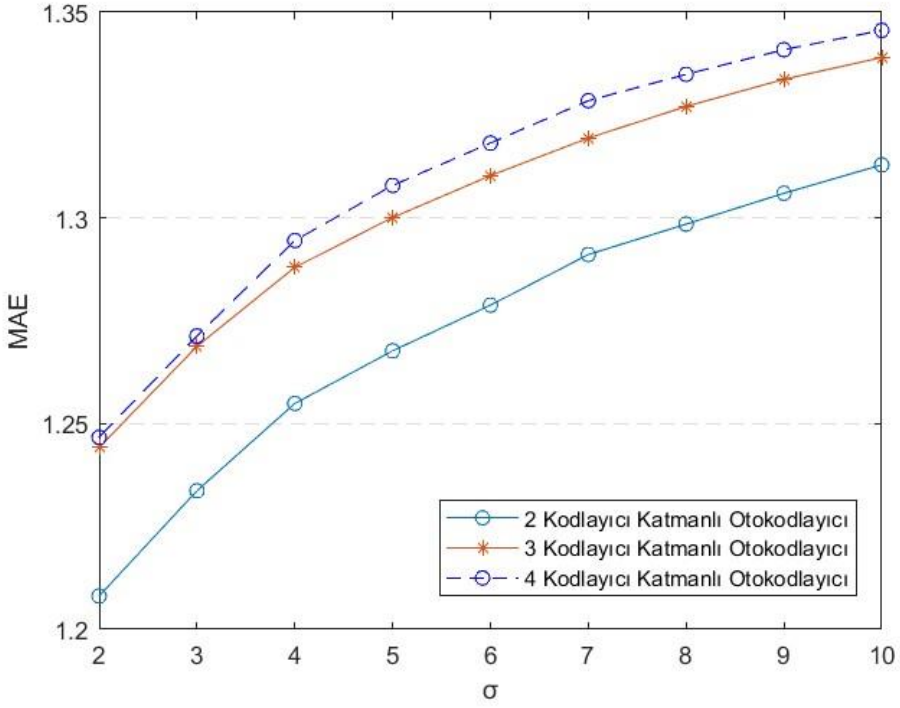
Kodlayıcı katman sayısındaki değişikliklerin öneri doğruluğuna etkilerini gözlemlemek için 2, 3 ve 4 kodlayıcı katman içeren üç farklı otokodlayıcı mimarisi ile deneyler yapılmıştır. Kodlayıcı katman sayıları arttıkça her bir katmanın düğüm sayısı dıştan içeri doğru sırasıyla *n/5, n/8* ve *n/12* olarak azalmaktadır (Batmaz, 2019). Kodlayıcı katman sayısı 2 olan otokodlayıcı mimarisi kullanılarak farklı aktivasyon fonksiyonlarının sağladığı öneri doğrulukları bir önceki bölümde incelenmiş ve en yüksek öneri doğruluğunun *tanh* aktivasyon fonksiyonu ile elde edildiği görülmüştür. Bu nedenle 3 ve 4 katmanlı otokodlayıcılarda yapılan deneylerde *tanh* aktivasyon fonksiyonu kullanılmıştır. Tablo 4.3'de, 3 ve 4 kodlayıcı katman bulunduran otokodlayıcıların farklı σ seviyelerinde maskelenmiş veriler kullanılarak öneri üretirken elde edilen RMSE ve MAE değerleri görülmektedir.

Tablo 4.3. Değişen Kodlayıcı Katman Sayılarının Farklı σ Seviyelerindeki Verilerde Öneri Doğruluğuna Etkisi

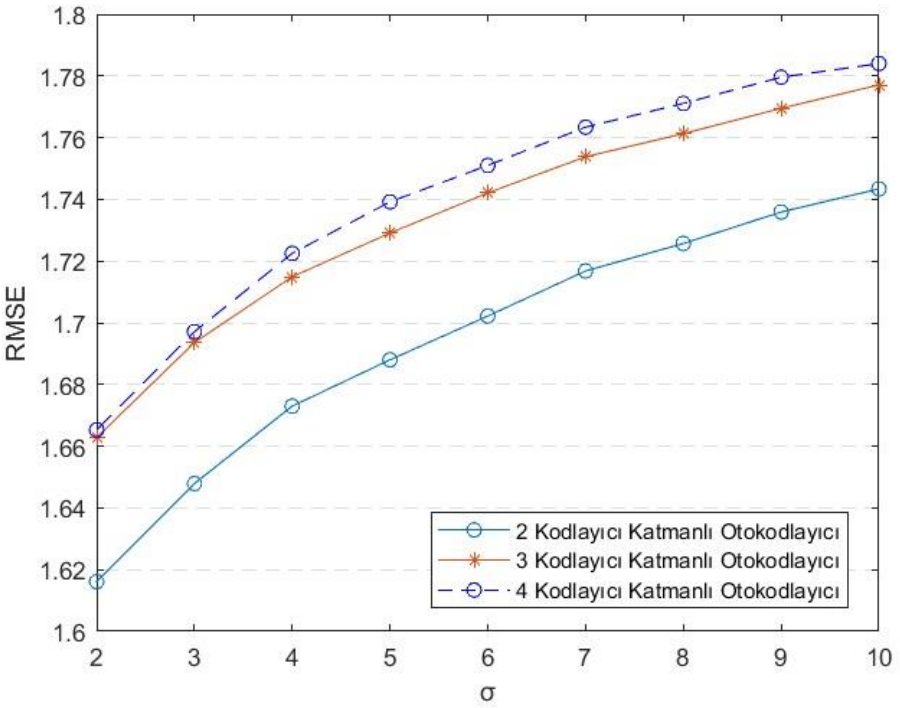
	3 Kodlayıcı Katman		**4 Kodlayıcı Katman**	
	MAE	**RMSE**	**MAE**	**RMSE**
$\boldsymbol{\sigma = 2}$	1,2442	1,6628	1,2467	1,6653
$\boldsymbol{\sigma = 3}$	1,2688	1,6936	1,2713	1,6971
$\boldsymbol{\sigma = 4}$	1,2880	1,7149	1,2945	1,7225
$\boldsymbol{\sigma = 5}$	1,3001	1,7291	1,3079	1,7392
$\boldsymbol{\sigma = 6}$	1,3103	1,7421	1,3182	1,7510
$\boldsymbol{\sigma = 7}$	1,3194	1,7538	1,3285	1,7634
$\boldsymbol{\sigma = 8}$	1,3271	1,7613	1,3349	1,7711
$\boldsymbol{\sigma = 9}$	1,3337	1,7695	1,3409	1,7796
$\boldsymbol{\sigma = 10}$	1,3390	1,7771	1,3456	1,7840

Hesaplanan MAE ve RMSE değerleri incelendiğinde, hem σ değerinin artmasıyla hem de katman sayısının artmasıyla hata değerinin yükseldiği görülmektedir. Bu durum şu şekilde yorumlanabilir; modelin öğrenme kapasitesi modelin katman sayısına bağlı olarak artmaktadır ancak bu durum beraberinde aşırı uyum problemini de getirmektedir. Öğrenme performansı artan katman sayısıyla yükseltilebilir ancak modelin eğitim verilerini ezberleyerek eğitim sürecini tamamlaması da ihtimal dahilindedir. Otokodlayıcı aşırı öğrenme problemine bağlı olarak test verilerindeki aykırı değerler için daha yüksek hata değeri hesaplamaktadır. Bir başka sebep, eğitim verisi sayısının katman sayısı arttıkça yetersiz kalması olabilmektedir.

Sonuç olarak, kullanılan BXM veri seti için katman sayısının artması öneri doğruluğu üzerinde olumlu bir etki yaratmamıştır. Şekil 4.5 ve Şekil 4.6'da verilmiş olan grafiklerde, artan kodlayıcı katman sayılarıyla ve σ değerleriyle birlikte MAE ve RMSE'de ortaya çıkan artış görülmektedir. Grafik incelendiğinde 2 kodlayıcı katman bulunduran otokodlayıcı mimarisi kullanılarak üretilen önerilerin 3 ve 4 kodlayıcı katman bulunduran otokodlayıcı mimarisiyle üretilen önerilere göre daha yüksek öneri doğruluğu sağladığı görülmektedir.



Şekil 4.5. Farklı Katman Sayılarındaki Otokodlayıcıların Her Bir σ Seviyesindeki MAE



Şekil 4.6. Farklı Katman Sayılarındaki Otokodlayıcıların Her Bir σ Seviyesindeki RMSE

4.5. Önerilen Yaklaşımın Geleneksel GKOF Sistemleri ile Karşılaştırılması

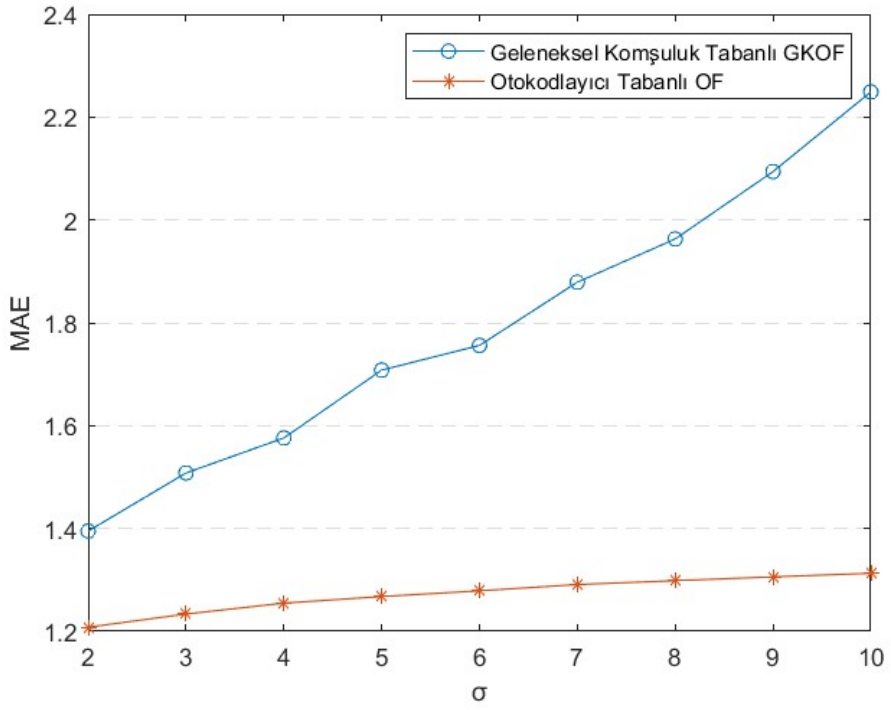
Kullanılan otokodlayıcı tabanlı yaklaşım geleneksel komşuluk tabanlı OF sistemleriyle öneri doğruluğu bakımından karşılaştırılmıştır. Otokodlayıcı tabanlı OF sisteminde, RKT ile maskelenen kullanıcı derecelendirme değerleri otokodlayıcının girişine beslenir ve çıkış katmanında yeniden oluşturulur. Otokodlayıcının kodlayıcı kısmında BXM veri setindeki özellikler doğrusal olmayan bir aktivasyon fonksiyonu olan *tanh* kullanılarak çıkarılmaktadır ve veri seti daha düşük boyutlu hale getirilmektedir. Otokodlayıcının giriş verisini yeniden oluşturmasındaki başarısı ise öneri doğruluğunu ifade etmektedir. Deneysel çalışmalarda görüldüğü üzere, otokodlayıcı tabanlı yaklaşımla BXM veri seti için en yüksek öneri doğruluğu, 2 kodlayıcı katman içeren ve *tanh* aktivasyon fonksiyonunu kullanan otokodlayıcı mimarisi ile elde edilmiştir. Tablo 4.4'te, geleneksel komşuluk tabanlı OF sistemi ile elde edilen hata değerleri ve en iyi sonucu veren otokodlayıcı mimarisi ile elde edilen hata değerleri sunulmuştur.

Tablo 4.4. Geleneksel GKOF ve Otokodlayıcı Tabanlı ÖS'nin Öneri Doğrulukları

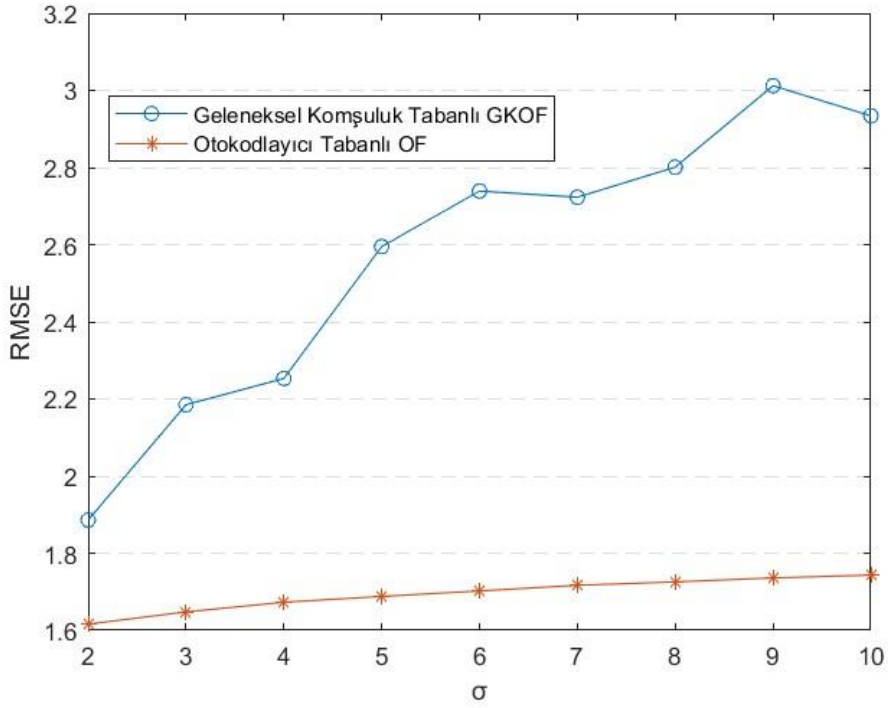
	Geleneksel GKOF		**Otokodlayıcı Tabanlı ÖS**	
	MAE	**RMSE**	**MAE**	**RMSE**
$\boldsymbol{\sigma = 2}$	1,395	1,886	1,208	1,616
$\boldsymbol{\sigma = 3}$	1,508	2,185	1,234	1,648
$\boldsymbol{\sigma = 4}$	1,576	2,253	1,255	1,673
$\boldsymbol{\sigma = 5}$	1,708	2,595	1,268	1,688
$\boldsymbol{\sigma = 6}$	1,756	2,739	1,279	1,702
$\boldsymbol{\sigma = 7}$	1,879	2,726	1,291	1,717
$\boldsymbol{\sigma = 8}$	1,963	2,801	1,299	1,726
$\boldsymbol{\sigma = 9}$	2,094	3,012	1,306	1,736
$\boldsymbol{\sigma = 10}$	2,249	2,934	1,313	1,743

Tablo 4.4'te görüldüğü gibi, otokodlayıcı tabanlı yaklaşım geleneksel komşuluk tabanlı OF sistemine göre daha yüksek doğrulukta öneriler üretmektedir. Veri setinin mahremiyet seviyesi yükseldikçe hem geleneksel komşuluk tabanlı OF sisteminde hem de otokodlayıcı tabanlı OF sisteminde öneri doğruluğu olumsuz etkilenmektedir. En yüksek öneri doğruluğuna ulaşan otokodlayıcı mimarisinde, en düşük gizlilik seviyesinde ve en yüksek gizlilik seviyesinde maskelenmiş veri setleri için üretilen önerilerin doğrulukları arasında, MAE'ye göre ~%9 fark, RMSE'ye göre ~%7,8 fark bulunmaktadır. Aynı otokodlayıcı mimarisi ile elde edilen öneri doğrulukları geleneksel komşuluk tabanlı OF sistemleri ile elde edilen öneri doğrulukları ile karşılaştırıldığında ise hem MAE hem de RMSE metriğine göre ve tüm gizlilik seviyelerinde otokodlayıcı tabanlı OF sisteminin daha başarılı sonuçlar verdiği görülmektedir.

Otokodlayıcı tabanlı OF sistemi en düşük gizlilik seviyesinde geleneksel komşuluk tabanlı OF sistemine göre ~%13,4 daha düşük MAE ve ~%14,3 daha düşük RMSE ile öneri üretmektedir. Aradaki fark en yüksek gizlilik seviyesinde ise MAE için ~%41,6 ve RMSE için ~%40,6 seviyesine kadar yükselmektedir. Bu veriler göz önünde bulundurulduğunda, otokodlayıcı tabanlı OF sistemi ile öneri üretildiğinde kullanıcıya daha yüksek gizlilik seviyeleri sağlanırken geleneksel komşuluk tabanlı OF sistemine göre daha belirgin bir farkla yüksek doğrulukta öneriler üretilebildiği görülmektedir. Tablo 4.4'te görülen hata değerlerine göre dikkat çeken bir diğer nokta, otokodlayıcı tabanlı OF sistemi tarafından en yüksek gizlilik seviyesinde maskelenen veriler kullanılarak üretilen öneriler için MAE değeri 1,313 ve RMSE değeri 1,743 olarak hesaplanırken, geleneksel komşuluk tabanlı OF sistemi tarafından en düşük gizlilik seviyesinde maskelenen veriler kullanılarak üretilen önerilerin hata değerlerinin MAE için 1,395 ve RMSE için 1,886 olarak hesaplanarak daha düşük doğruluğa sahip olduğunun görülmesidir. Bu veriler referans alınarak otokodlayıcı tabanlı OF sisteminin en yüksek gizlilik seviyesinde dahi sunduğu öneri doğruluğunun geleneksel komşuluk tabanlı OF sisteminin en düşük gizlilik seviyesinde sunduğu öneri doğruluğundan daha yüksek olduğu söylenebilmektedir. Otokodlayıcı tabanlı OF sisteminin hata değerleri kendi içinde değerlendirildiğinde ise en düşük gizlilik seviyesinde elde edilen öneri doğruluğu ve en yüksek gizlilik seviyesinde elde edilen öneri doğruluğu arasında MAE için ~%8,7 ve RMSE için ~%7,3 seviyelerinde kayıp görülmektedir. Aynı durum geleneksel komşuluk tabanlı OF sistemi için MAE değerinde ~%38 ve RMSE değerinde ~%35,7 öneri doğruluğu kaybı olarak görülmektedir. Bu verilere bakılarak, otokodlayıcı tabanlı OF sisteminin yükselen gizlilik seviyelerine bağlı olarak veri setinde ortaya çıkan bozulmaları daha iyi tolere edebildiği, kullanıcıya yüksek mahremiyet seviyeleri sağlandığında da yüksek doğrulukta öneriler üretebildiği söylenebilmektedir. Otokodlayıcı tabanlı OF sisteminin öneri üretme mekanizması veri setindeki gizli özelliklerin öğrenilmesine dayalı olduğundan, yükselen gizlilik seviyesine bağlı olarak verilerin dağılımı daha geniş bir aralıkta olsa bile öneri doğruluğundaki kayıp daha az olmaktadır.



Şekil 4.7. Geleneksel GKOF Sistemi ve Otokodlayıcı Tabanlı ÖS ile Elde Edilen MAE Değerleri



Şekil 4.8. Geleneksel GKOF Sistemi ve Otokodlayıcı Tabanlı ÖS ile Elde Edilen RMSE Değerleri

Şekil 4.7'de ve Şekil 4.8'de geleneksel komşuluk tabanlı OF sistemi ve otokodlayıcı tabanlı OF sisteminde elde edilen hata değerlerinin σ seviyesine göre gösterdiği değişiklik görülmektedir. Otokodlayıcı tabanlı yaklaşımda hata miktarının σ seviyesi ile birlikte gösterdiği artışın geleneksel komşuluk tabanlı OF yaklaşımına kıyasla daha az olduğu oldukça belirgin şekilde görülebilmektedir. Kullanılan yaklaşım, geleneksel komşuluk tabanlı OF yaklaşımına göre öneri doğruluğunu yükselterek, daha yüksek mahremiyet seviyelerinde daha başarılı tahminlerde bulunulmasını sağlamıştır. Şekil 4.7 ve Şekil 4.8'de sırasıyla gösterilen geleneksel komşuluk tabanlı OF sisteminin MAE ve RMSE grafikleri incelendiğinde, MAE grafiğinin RMSE grafiğine göre daha az kırılımla arttığı ve daha düşük seviyelerde seyrettiği görülmektedir. Bu farkın temel nedeni, RMSE hata metriğinde kullanılan kare alma işlemidir. Bu işlem, hataların büyüklüğünde üstel bir artışa neden olmaktadır ve hataları daha belirgin hale getirmektedir. MAE ise bu tür bir artış göstermemekte, tahmin edilen değerlerin gerçek değerlere olan mutlak uzaklığına göre hesaplanmaktadır. RMSE hata metriği MAE hata metriğine göre büyük hataları daha vurgulu bir şekilde ortaya koyduğundan grafikte hem daha yüksek değerler görülmekte hem de farklı noktalarda farklı büyüklükte hatalar görülebileceğinden daha kırılımlı bir eğri oluşmaktadır. Sonuç olarak, otokodlayıcı tabanlı OF sistemi ile elde edilen kazançlar aşağıdaki gibi sıralanabilir;

- Kullanıcılara tüm gizlilik seviyelerinde daha yüksek öneri doğrulukları sunulmuştur.
- Artan gizlilik seviyesine bağlı olarak RKT'nin neden olduğu artan öneri doğruluğu kaybı azaltılmıştır.
- Kullanıcılara, geleneksel komşuluk tabanlı OF sistemlerine kıyasla öneri doğruluğundan çok daha az ödün verilerek yüksek gizlilik seviyeleri sağlanmıştır.

5. SONUÇLAR

Yapılan tez çalışmasında, RKT ile maskelenmiş BXM veri seti kullanılarak geleneksel komşuluk tabanlı OF sistemi ve otokodlayıcı tabanlı bir OF sistemi oluşturulmuştur ve her iki sistemin de değişen mahremiyet seviyelerinde sağladığı öneri doğrulukları incelenmiştir. Otokodlayıcı tabanlı yaklaşım kullanılarak, geleneksel komşuluk tabanlı OF sisteminde üretilen önerilerde görülen ve RKT'nin neden olduğu öneri doğruluğu kayıplarının azaltılması amaçlanmıştır. Her iki yaklaşımın da sağladığı öneri doğruluğunun ayrı ayrı görülmesi ve karşılaştırılabilmesi için çalışma iki aşamada gerçekleştirilmiştir.

Geleneksel komşuluk tabanlı OF sistemi ile BXM veri seti kullanılarak üretilen önerilerin doğrulukları ve RKT ile sağlanan mahremiyet seviyesi Bölüm 3'te sunulmuştur. RKT'nin OF sistemlerinin mahremiyet risklerini azaltmakta ve kullanıcı mahremiyetini sağlamaktaki etkisi hesaplanmıştır. Hesaplanan mahremiyet seviyesinin σ değerine bağlı olarak yükseldiği görülmüştür. OF sistemlerinin mahremiyet risklerini azaltmak için kullanılan RKT'nin neden olduğu öneri doğruluğu kaybının, mahremiyet seviyesinin yükselmesi ile birlikte arttığı görülmüştür. Kullanıcıların mahremiyeti korunsa bile yeterince doğru öneriler üretemeyen bir ÖS kullanıcılara iyi bir deneyim sunmayacağı için tercih edilmeyecektir. RKT ile mahremiyet sağlanırken öneri doğruluğundan daha az taviz verilerek öneri üretilebilecek bir yaklaşım olarak otokodlayıcı tabanlı OF kullanılmıştır.

Bölüm 4'te, kullanıcıların mahremiyet riskleri hafifletilirken daha yüksek öneri doğruluğu sağlamak amacıyla otokodlayıcı tabanlı bir yaklaşım kullanılmıştır. Bu yaklaşımda, bir önceki bölümde kullanılan RKT ile maskelenmiş BXM veri seti kullanılmıştır ve öneri üretme prosedürü otokodlayıcı kullanılarak gerçekleştirilmiştir. Ayrıca farklı aktivasyon fonksiyonları ve kodlayıcı katman sayılarının öneri doğruluğuna etkilerini gözlemleyebilmek için *tanh*, *elu*, *selu* ve *linear* aktivasyon fonksiyonlarıyla 2, 3 ve 4 gizli katman içeren farklı otokodlayıcı mimarileriyle deneyler yapılmıştır. Öncelikle 2 gizli katman içeren otokodlayıcı mimarisi kullanılarak değişen aktivasyon fonksiyonlarının öneri doğruluğu üzerindeki etkisi incelenmiştir. Deneylerde, otokodlayıcının iç katmanlarındaki aktivasyon fonksiyonları sonuca olan etkisi görülmek üzere değiştirilmiştir. Bu sırada giriş verileriyle uyumlu bir çıkış aralığı sağlamak amacıyla çıkış katmanındaki aktivasyon fonksiyonu *tanh* olarak sabit bırakılmıştır. Deneysel çalışmalar sonucunda en yüksek öneri doğruluğu sağlayan aktivasyon fonksiyonu *tanh* olarak belirlenirken en düşük öneri doğruluğu *linear* ile elde edilmiştir.

Otokodlayıcılarda giriş ve çıkış katmanlarının uyumlu aralıkta olmasının modelin genel performansı üzerinde kritik bir etkisi olduğu tespit edilmiştir. Hem iç katmanlarda hem de dış katmanlarda *tanh* aktivasyon fonksiyonunun kullanılması, eğitimin giriş verileriyle uyumlu bir şekilde ilerlemesine imkan tanımış ve bu durum, BXM veri seti için öneri doğruluğunu olumlu yönde etkilemiştir. Bununla birlikte *linear* aktivasyon fonksiyonunun verileri bir aralığa sıkıştırmaması ve doğrusal bir fonksiyon olması nedeniyle, türeve dayalı öğrenme sağlayan Bayesian geri yayılım algoritmasında öğrenmeye destekleyici bir avantaj sağlamaması, daha düşük doğrulukta öneriler üretilmesine yol açmıştır. Giriş ve çıkış verilerinin birbiriyle uyumlu olması, öğrenilen temsillerin ve bu temsillerin yeniden oluşturulması sürecinin öneri doğruluğu üzerinde belirleyici bir rol oynadığı görülmüştür. Eğer giriş ve çıkış katmanındaki veriler uyumsuz olursa, otokodlayıcının veriyi doğru bir şekilde yeniden oluşturması zorlaşmakta ve bu durum öneri doğruluğunda kayıplara neden olmaktadır. t*anh* aktivasyon fonksiyonu, veriyi [-1,1] aralığına sıkıştırmaktadır ve bu, z-skor ile normalize edilen giriş verileri için ideal bir aralık sağlamaktadır. Bu yaklaşım, otokodlayıcının daha tutarlı temsiller öğrenmesini desteklemekte ve öneri doğruluğunu artırmaktadır. Sonuç olarak, deneyler aktivasyon fonksiyonunun seçiminin öneri doğruluğu üzerinde belirleyici olduğunu ortaya koymuştur.

Katman sayısının öneri doğruluğu üzerindeki etkisinin görülmesi amacıyla değişen gizli katman sayılarıyla yapılan deneylerde ise, BXM veri seti için, artan katman sayısının öneri doğruluğu üzerinde olumlu bir etki yaratmadığı belirlenmiştir. Artan katman sayısıyla, modelin karmaşıklığında meydana gelen artış, özellikle sınırlı giriş verileri olduğunda, verinin yetersiz kalabileceği bir duruma yol açabilmektedir. Ayrıca, çok sayıda katmanın kullanılması, aşırı öğrenme problemine neden olabilmekte ve bu da genelleme kabiliyetinin azalmasına yol açmaktadır. Giriş verilerinin yetersiz kaldığı ve aşırı öğrenmenin tetiklendiği durumlarda, öneri doğruluğunun beklenenden daha düşük olabildiği görülmüştür. Dolayısıyla, deneylerde katman sayısının öneri doğruluğu üzerindeki etkisinin sadece modelin kapasitesini artırmakla sınırlı olmadığı, aynı zamanda modelin genelleme yeteneği ve öğrenme dinamikleri üzerinde de belirleyici olduğu sonucuna varılmıştır.

Geleneksel GKOF sistemi ve önerilen otokodlayıcı tabanlı gizliliği koruyan ÖS öneri doğruluğu bakımından karşılaştırılmış ve önerilen yöntemin daha yüksek öneri doğruluğuna ulaşarak etkili bir yaklaşım olduğu görülmüştür. Bu sayede ÖS'lerin temel problemlerinden olan mahremiyet riskleri hafifletilirken, daha yüksek öneri doğruluğuna ulaşılabilmesi sağlanmıştır. Özetle, geleneksel komşuluk tabanlı OF sistemlerinde RKT ile mahremiyet

sağlanırken öneri doğruluğunda görülen kaybı azaltmak amacıyla yapılan bu çalışma neticesinde elde edilen sonuçlar şöyle listelenebilmektedir:

- Geleneksel komşuluk tabanlı OF sistemlerinde, RKT ile maskelenmiş BXM veri seti kullanılarak öneriler üretilmiştir.
- RKT ile farklı σ seviyelerinde maskelenen BXM veri setinde ulaşılan mahremiyet seviyesi tespit edilmiştir.
- Öneri üretme süreci, veri setinin farklı gizlilik seviyelerinde maskelenmiş versiyonlarında gerçekleştirilmiş olup, RKT ile sağlanan mahremiyet seviyesine göre geleneksel komşuluk tabanlı OF sistemlerinde üretilen önerilerin doğruluğundaki değişim görülmüştür.
- RKT ile daha yüksek seviyelerde mahremiyet sağlandığında, öneri doğruluğu üzerinde olumsuz etkisi olduğu görülmüştür.
- RKT'nin öneri doğruluğu üzerindeki olumsuz etkisinin azaltılması ve daha yüksek gizlilik seviyelerinde daha yüksek öneri doğruluğu elde edilebilmesi için otokodlayıcı tabanlı OF kullanılmış ve bu yaklaşım ile elde edilen öneri doğrulukları incelenmiştir.
- Otokodlayıcılarda kullanılan farklı aktivasyon fonksiyonlarının öneri doğruluğu üzerinde kritik etkisi olduğu görülmüştür.
- Otokodlayıcı mimarisindeki gizli katman sayısını arttırmanın mevcut veri seti için avantaj sağlamadığı ve öneri doğruluğu üzerinde olumsuz etki yarattığı görülmüştür.
- Artan σ seviyesinin hem geleneksel komşuluk tabanlı OF sisteminde, hem de otokodlayıcı tabanlı OF sisteminde öneri doğruluğunu olumsuz etkilediği sonucuna ulaşılmıştır. Ancak gizlilik seviyesinin artmasıyla birlikte öneri doğruluğunda meydana gelen kaybın geleneksel komşuluk tabanlı OF sisteminde otokodlayıcı tabanlı OF sistemine göre çok daha yüksek olduğu görülmüştür.
- Sonuç olarak, kullanılan otokodlayıcı tabanlı yaklaşım ile kullanıcıya daha yüksek gizlilik seviyesinde daha yüksek öneri doğruluğu sağlanabilmiştir.

Yapılan çalışmada maskelenmiş veri seti üzerinde DÖ tabanlı yaklaşımların tahmin doğruluğu üzerine olumlu etkisi olduğu görülmektedir. Bu doğrultuda, otokodlayıcıya ek olarak, farklı DÖ yaklaşımlarının RKT ile maskelenmiş veri setlerinde öneri doğruluğu üzerine etkisinin incelenmesi hedeflenmektedir.

KAYNAKÇA

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). Privacy in e-commerce: examining user scenarios and privacy preferences. *In Proceedings of the 1st ACM Conference on Electronic Commerce* 1-8.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6), 734-749.

Afoudi, Y., Lazaar, M., & Al Achhab, M. (2021). Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, 113, 102375.

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

Al-Mani, I. A., Al-Sabaawi, A. M. A., & Hussien, M. H. (2022). A Review Paper of Model Based Collaborative Filtering Techniques. In 2022 *International Conference on Data Science and Intelligent Computing (ICDSIC)* 52-57. IEEE.

AL SBOU, A. M., & Abd Rahim, N. H. (2022). Performance comparison of three different types of autoencoders using recommendation systems. *Journal of Theoretical and Applied Information Technology*, 100(5).

Althbiti, A., Alshamrani, R., & Ma, X. (2020). A Literature Review of Data Mining Techniques Used in Collaborative Filtering Recommender Systems. In 2020 *International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 424-430). IEEE.

Apicella, A. vd. (2021). A survey on modern trainable activation functions. *Neural Networks*, 138, 14-32.

Bathla, G., Aggarwal, H., & Rani, R. (2020). AutoTrustRec: Recommender system with social trust and deep learning using autoEncoder. *Multimedia Tools and Applications*, 79, 20845-20860.

Batmaz, Z. (2019). Derin Öğrenme Yaklaşımları ile Çoklu-Kriterli Öneri Sistemlerinin Problemlerini Çözmek, *Doktora Tezi*. Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Eskişehir.

Batmaz, Z., & Kaleli, C. (2019). AE-MCCF: an autoencoder-based multi-criteria recommendation algorithm. *Arabian Journal for Science and Engineering*, 44, 9235- 9247.

Batmaz, Z., & Polat, H. (2016). Randomization-based privacy-preserving frameworks for collaborative filtering. *Procedia Computer Science*, 96, 33-42.

Bergner, Y. vd. (2012). Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*.

Berkovsky, S. vd. (2007). Enhancing privacy and preserving accuracy of a distributed collaborative filtering. In Proceedings of the 2007 ACM conference on Recommender systems 9-16.

Bilge, A. vd. (2013). A survey of privacy-preserving collaborative filtering schemes. *International Journal of Software Engineering and Knowledge Engineering*, 23(08), 1085–1108.

Bobadilla, J. vd. (2013). Recommender systems survey. *Knowledge-based systems*, 46, 109-132.

Book-Crossing.(2004). Book-Crossing Veri Seti [Erişim: 05.04.2023, https://grouplens.org/datasets/book-crossing/]

Casino, F. vd. (2013, September). On privacy preserving collaborative filtering: Current trends, open problems, and new issues. *In 2013 IEEE 10th International Conference on e-Business Engineering* 244-249. IEEE.

Chaabane, A., Acs, G., & Kaafar, M. A. (2012). You are what you like! information leakage through users' interests. In Proceedings of the 19th annual network & distributed system security symposium (NDSS). Citeseer.

Clevert, D.-A., Unterthiner, T. and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus), arXiv preprint arXiv:1511.07289.

Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*.

Ding, B., Qian, H., & Zhou, J. (2018, June). Activation functions and their characteristics in deep neural networks. In 2018 *Chinese control and decision conference (CCDC)* (1836-1841). IEEE.

Dwivedi, P., & Islam, B. (2023, March). An Item-based Collaborative Filtering Approach for Movie Recommendation System. *In 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)* 153-158. IEEE.

Fkih, F. (2022). Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison. *Journal of King Saud University- Computer and Information Sciences*, 34(9), 7645-7669.

Georgiev, K., & Nakov, P. (2013). A non-iid framework for collaborative filtering with restricted boltzmann machines. *In International conference on machine learning* 1148-1156. PMLR.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning. MIT press*.

Haghighi, P. S., Seton, O., & Nasraoui, O. (2019). An explainable autoencoder for collaborative filtering recommendation. arXiv preprint arXiv:2001.04344.

Heidari, N., Moradi, P., & Koochari, A. (2022). An attention-based deep learning method for solving the cold-start and sparsity issues of recommender systems. *Knowledge-Based Systems*, 256, 109835.

Ibrahim, M. vd. (2023). An Intelligent Hybrid Neural Collaborative Filtering Approach for True Recommendations. IEEE Access.

Karatzoglou, A., & Hidasi, B. (2017). Deep learning for recommender systems. In *Proceedings of the eleventh ACM conference on recommender systems* 396-397.

Khan, Z. Y. vd. (2021). Deep learning techniques for rating prediction: a survey of the state-of-the-art. *Artificial Intelligence Review*, 54, 95-135.

Kiran, R., Kumar, P., & Bhasker, B. (2020). DNNRec: A novel deep learning based hybrid recommender system. *Expert Systems with Applications*, 144, 113054.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning, Nature, 521(7553), 436-444

Liu, X. vd. (2017). When differential privacy meets randomized perturbation: a hybrid approach for privacy-preserving recommender system. In Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22 576-591. *Springer International Publishing*.

Ning, X., Desrosiers, C., & Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook*, 37-76.

Nwankpa, C. vd. (2018). Activation functions: Comparison of trends in practice and research for deep learning. arXiv preprint arXiv:1811.03378.

Phaisangittisagul, E. (2016). An analysis of the regularization between L2 and dropout in single hidden layer neural network. In 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS) 174-179. IEEE.

Polat, H., & Du, W. (2003). Privacy-preserving collaborative filtering using randomized perturbation techniques. *In Third IEEE international conference on data mining* 625-628. IEEE.

Polat, H., & Du, W. (2005). Privacy-preserving collaborative filtering. *International journal of electronic commerce*, 9(4), 9-35.

Polatidis, N. vd. (2017). Privacy-preserving collaborative recommendations based on random perturbations. *Expert Systems with Applications*, 71, 18-25.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: introduction and challenges. *Recommender systems handbook*, 1-34.

Sarwar, B. vd. (2001, April). Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th international conference on World Wide Web* 285-295.

Sedhain, S. vd. (2015). Autorec: Autoencoders meet collaborative filtering. *In Proceedings of the 24th international conference on World Wide Web* 111-112.

Shyong, K., Frankowski, D. ve Riedl, J. (2006). Do you trust your recommendations? An exploration of security and privacy issues in recommender systems. *In Emerging trends in information and communication security, Springer, Berlin, Heidelberg*, 14-29.

Strub, F., Gaudel, R., & Mary, J. (2016). Hybrid recommender system based on autoencoders. *In Proceedings of the 1st workshop on deep learning for recommender systems* 11-16.

Strub, F., Mary, J., & Philippe, P. (2015). Collaborative filtering with stacked denoising autoencoders and sparse inputs. *In NIPS workshop on machine learning for eCommerce*.

Thorat, S. A., Ashwini, G., & Seema, M. (2023). Survey on Collaborative and Content-based Recommendation Systems. I*n 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* 1541-1548. IEEE.

Unger, M. (2015, September). Latent context-aware recommender systems. *In Proceedings of the 9th ACM Conference on Recommender Systems* 383-386.

Unger, M. vd. (2016). Towards latent context-aware recommendation systems. *Knowledge-Based Systems*, 104, 165-178.

Unger, M. vd. (2018). Inferring contextual preferences using deep encoder-decoder learners. New Review of Hypermedia and Multimedia, 24(3), 262-290.

Wei, R., Tian, H., & Shen, H. (2018). Improving k-anonymity based privacy preservation for collaborative filtering. *Computers & Electrical Engineering*, 67, 509-519.

Wei, Y. vd. (2022). Heterogeneous graph neural network for privacy-preserving recommendation. arXiv preprint arXiv:2210.00538.

Weinsberg, U. vd. (2012). BlurMe: Inferring and obfuscating user gender based on ratings. *In Proceedings of the sixth ACM conference on Recommender systems* 195-202.

Yalcin, E., & Bilge, A. (2023). Popularity bias in personality perspective: An analysis of how personality traits expose individuals to the unfair recommendation. *Concurrency and Computation: Practice and Experience*, 35(9), e7647.

Yang, C. vd. (2017). Bridging collaborative filtering and semi-supervised learning: a neural approach for poi recommendation. *In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* 1245-1254.

Yargıç, A., & Açıl, E. T. (2022). Privacy-preserving collaborative filtering system for book-crossing dataset. VI.- International European Conference on Interdisciplinary Scientific Research, 289-295

Yargıç, A., & Bilge, A. (2017). Privacy Risks for Multi-Criteria Collaborative Filtering Systems, 26th International Conference on Computer Communication and Networks (ICCCN), Vancouver, BC, Canada, 1-6.

Yargıç, A., & Bilge, A. (2019). Privacy-preserving multi-criteria collaborative filtering. Information Processing & Management, 56(3), 994-1009.

Zhang, G., Liu, Y., & Jin, X. (2020). A survey of autoencoder-based recommender systems. Frontiers of Computer Science, 14, 430-450.