

**MARMARA ÜNİVERSİTESİ**

**TEKNOLOJİ FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**BİTİRME PROJESİ**

Federe Öğrenme ile Sağlık Verilerini Kullanarak Model Eğitimi ve Gerçek Dünya Analizi

**PROJE YAZARI**

Abdülbaki Demir

170421003

Ahmet Yasir Kulaksız

170421041

**DANIŞMAN**

Doc. Dr. Kazım YILDIZ

**İSTANBUL, 2025**

**MARMARA ÜNİVERSİTESİ**

**TEKNOLOJİ FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

Marmara Üniversitesi Teknoloji Fakültesi Bilgisayar Mühendisliği Öğrencileri Abdülbaki DEMİR ve Ahmet Yasir KULAKSIZ tarafından “Federe Öğrenme ile Sağlık Verilerini Kullanarak Model Eğitimi ve Gerçek Dünya Analizi” başlıklı proje çalışması, 19.06.2025 tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

**Jüri Üyeleri**

Doç. Dr. Kazım YILDIZ (Danışman)

Marmara Üniversitesi ......................................................................... (İMZA) ..................

Dr. Öğr. Üyesi Ali SARIKAŞ (Üye)

Marmara Üniversitesi ......................................................................... (İMZA) ..................

**İÇİNDEKİLER**

Sayfa

[ÖNSÖZ iii](#_Toc200814092)

[KISALTMALAR iv](#_Toc200814093)

[ŞEKİL LİSTESİ v](#_Toc200814094)

[TABLO LİSTESİ vi](#_Toc200814095)

[ÖZET vii](#_Toc200814096)

[1. Giriş 1](#_Toc200814097)

[2. Literatür Taraması 4](#_Toc200814098)

[3. Kullanılan Yöntemler 6](#_Toc200814099)

[3.1. Veri Setleri 6](#_Toc200814100)

[3.1.1 Diabetes Health Indicators Dataset 7](#_Toc200814101)

[3.1.2 Heart Attack Dataset 11](#_Toc200814102)

[3.2 Prototip Uygulama 15](#_Toc200814103)

[3.2.1 İstemciler 17](#_Toc200814104)

[3.2.2 API 17](#_Toc200814105)

[3.2.2 Ana Sunucu 17](#_Toc200814106)

[4. DENEYSEL SONUÇLAR 19](#_Toc200814107)

[5. TARTIŞMA 23](#_Toc200814108)

[6. SONUÇ 27](#_Toc200814109)

ÖNSÖZ

Bu proje çalışmamız süresince bizlere her zaman destek olan, bilgi ve deneyimlerini bizden esirgemeyen değerli hocalarımız Doc. Dr. Kazım Yıldız, Arş. Gör. Büşra Büyüktanır, ve Dr. Öğr. Üyesi Gözde Karataş Baydoğmuş’a en içten teşekkürlerimizi sunarız.

Proje fikrinin şekillenmesinden uygulama sürecine kadar karşılaştığımız her türlü sorunda bizlere yol gösteren, sabırla yanımızda olan ve desteğini her daim hissettiren hocalarımızın katkıları, çalışmamızın başarıyla tamamlanmasında büyük rol oynamıştır.

Ayrıca proje sürecinde bizlere moral ve motivasyon sağlayan, fikir alışverişlerinde bulunduğumuz tüm çalışma arkadaşlarımıza ve desteklerini hissettiğimiz diğer akademisyenlerimize de teşekkür ederiz.

KISALTMALAR

**API :** application programming ınterface

**BRFSS :** behavioral risk factor surveillance system

**CBA :** classification based on association rules

**CDC :** centers for disease control and prevention

**duCBA :** data unaware classification based on association

**FairFed :** fair federated learning

**FedAvg :** federated averaging

**FedAvu :** federated asynchronous update

**FedGiA :** federated gradient ınformation aggregation

**HTTP :** hyper text transfer protocol

**IoT :** internet of things

**JSON :** javascript object notation

**PCA :** principal component analysis

**RoFL :** robustness of secure federated learning

**XAI :** explainable artificial ıntelligence

ŞEKİL LİSTESİ

Sayfa

[Şekil 1: Federe Öğrenme Akışı Diyagramı 2](#_Toc200719266)

[Şekil 2: Bağımsız Değişkenlerin Diyabet\_binary ile Korelasyonu 7](#_Toc200719267)

[Şekil 3: Diabetes\_binary için Mutual Informatıon Skorları 8](#_Toc200719268)

[Şekil 4: Diabetes\_binary için Chi-Square Skorları 9](#_Toc200719269)

[Şekil 5: Diabetes\_binary için PCA Katkısı 9](#_Toc200719270)

[Şekil 6: Bağımsız Değişkenlerin HeartDisease ile Korelasyonu 12](#_Toc200719271)

[Şekil 7: Bağımsız Değişkenlerin HeartDisease ile Korelasyonu 12](#_Toc200719272)

[Şekil 8: HeartDisease için Chi-Square Skorları 13](#_Toc200719273)

[Şekil 9: HeartDisease için PCA Katkısı 14](#_Toc200719274)

[Şekil 10: duCBA Akışı Diyagramı 16](#_Toc200719275)

TABLO LİSTESİ

Sayfa

[Tablo 1: Dıabetes Health Indıcators Dataset için Özellı̇k Seçı̇mı̇ 10](#_Toc200719276)

[Tablo 2: Heart Attack Dataset için Özellı̇k Seçı̇mı̇ 15](#_Toc200719277)

[Tablo 3: Dıabetes Health Indıcators Dataset için Sınıflandırma Raporu 20](#_Toc200719278)

[Tablo 4: Heart Attack Dataset için Sınıflandırma Raporu 21](#_Toc200719279)

ÖZET

Bu çalışma, sağlık alanında kişisel veri gizliliğinin korunması gerekliliğinden hareketle, merkezi veri toplama gerektirmeden etkili tahmin modelleri geliştirebilen federe öğrenme mimarisini ele almaktadır. Özellikle diyabet hastalığının erken teşhisi için, federe öğrenme altyapısı üzerinde geliştirilmiş duCBA algoritması kullanılarak, dağıtık bir makine öğrenmesi modeli oluşturulmuştur. Projede, katılımcıların verileri yerel ortamlarında işlenmiş ve yalnızca model parametreleri merkezi sunucuya iletilmiştir. Böylece, hem veri gizliliği sağlanmış hem de merkezi depolamanın yol açabileceği güvenlik riskleri azaltılmıştır. Model eğitimi için kamuya açık bir sağlık verisi seti üzerinde çeşitli özellik seçimi yöntemleri kullanılmış ve elde edilen veriler farklı istemciler arasında paylaştırılmıştır. Geliştirilen sistem, ana sunucu, arayüz sunucusu ve istemcilerden oluşan bir mimari üzerinde test edilmiştir. Deneysel sonuçlar, önerilen yaklaşımın diyabet tahmininde istikrarlı ve tatmin edici doğruluk oranlarına ulaştığını göstermiştir. Ayrıca, modelin şeffaf bir şekilde izlenebilmesi ve anlaşılabilir kurallar üretmesi sayesinde, elde edilen sonuçlar hem sağlık personeli hem de son kullanıcılar için yorumlanabilir niteliktedir. Sonuç olarak, bu çalışma, veri gizliliği hassasiyeti yüksek olan sağlık uygulamaları için etkili ve güvenli bir makine öğrenimi çözümü sunmaktadır.

**ABSTRACT**

This study addresses the necessity of protecting personal data privacy in the healthcare domain by employing a federated learning architecture that enables the development of effective prediction models without the need for centralized data collection. Focusing on the early diagnosis of diabetes, a distributed machine learning model was created using the duCBA algorithm, which is specifically designed for federated learning environments. In this project, participant data were processed locally, and only the model parameters were transmitted to the central server, thus ensuring data privacy and minimizing the security risks associated with centralized storage. Publicly available health data was used for model training, and various feature selection methods were applied before distributing the processed data across different clients. The developed system was tested on an architecture consisting of a main server, an interface server, and clients. Experimental results demonstrated that the proposed approach achieved stable and satisfactory accuracy rates in diabetes prediction. Furthermore, the model's ability to generate interpretable rules and offer transparent decision-making makes the results understandable for both healthcare professionals and end users. In conclusion, this study presents an effective and secure machine learning solution for healthcare applications where data privacy is of utmost importance.

1. Giriş

Veri odaklı teknolojilerdeki hızlı ilerleme, makine öğrenmesi modellerinin pek çok alanda temel bir unsur haline gelmesine yol açmıştır [1]. Elde edilen bu veriler; kullanıcı davranışlarının incelenmesinden, çevresel parametrelerin takibine kadar geniş bir yelpazede analiz edilmektedir [2]. Makine öğrenmesi algoritmalarının sunduğu öngörü gücü sayesinde, çeşitli alanlarda kullanılmak üzere çok sayıda sistem geliştirilmektedir [3]. Bu tür modeller, büyük hacimli veri kümelerini değerlendirerek karmaşık ilişkileri ve örüntüleri ortaya koymakta, geleceğe yönelik tahminlerde bulunmakta ve karar destek süreçlerine önemli katkı sunmaktadır. Ancak bu başarı, modellerin yeterli düzeyde etkin çalışabilmesi için büyük ve çeşitli veri setlerine ihtiyaç duymasıyla yakından ilişkilidir. Veri miktarının artması ise donanım kaynaklarının zorlanmasına, model eğitim süresinin uzamasına ve beraberinde veri güvenliği gibi bazı sorunların gündeme gelmesine neden olmaktadır [5].

Özellikle sağlık sektörü gibi hassas verilerin işlendiği ortamlarda, veri güvenliği ve gizliliği her zamankinden daha önemli bir hâl almaktadır [6-7]. Sağlık verisinin merkezi bir sunucuda toplanması hem yasal düzenlemeler hem de çeşitli güvenlik riskleri nedeniyle ciddi kısıtlamalarla karşı karşıya kalmaktadır [8]. Birçok ülkede yürürlükte olan kişisel verilerin korunması yasaları, sağlık verilerinin paylaşımını ve merkezi olarak işlenmesini önemli ölçüde sınırlandırmaktadır. Bu tür yasal ve etik sınırlamalar, klasik makine öğrenmesi uygulamalarının yaygınlaştırılmasını zorlaştırmakta ve yeni çözüm yolları geliştirilmesini gerektirmektedir.

Bu doğrultuda öne çıkan federe öğrenme yaklaşımı, verilerin merkezi bir noktada toplanması gereksinimini ortadan kaldıran ve verilerin yerel olarak, bulundukları cihazlarda işlenmesini sağlayan yenilikçi bir makine öğrenimi yöntemidir. Federe öğrenme, kullanıcıların verilerinin gizliliğini koruyarak, dağıtık bir şekilde model eğitimi yapılmasına imkân verir [9]. Özellikle veri gizliliğinin kritik olduğu sağlık sektörü gibi alanlarda, federe öğrenme etkili bir alternatif olarak ön plana çıkmaktadır. Bu yöntem sayesinde, hastalara ait veriler bulundukları cihazda işlenirken, yalnızca model güncellemeleri merkezi sunucuya iletilmektedir. Böylece iletişim ve depolama maliyetleri önemli ölçüde düşerken [10], merkezi yapıya veri aktarımı gerekmeyeceği için yerel ortamlardaki hassas bilgilerin güvenliği de artırılmış olur [11].

Federe öğrenmenin öne çıkan bir diğer avantajı da, veri çeşitliliğini artırarak geliştirilen modellerin genelleme kabiliyetini yükseltmesidir. Geleneksel yöntemlerle farklı kaynaklardan veri toplamak yasal ve teknik olarak pek çok güçlüğe yol açarken, federe öğrenme sayesinde bu veriler doğrudan kaynaklarında işlenebilmekte ve söz konusu engeller aşılabilmektedir. Ayrıca, veri sahipleri üzerinde kontrolün kalması, sisteme olan güveni artırırken daha fazla paydaşın sürece dâhil olmasını teşvik etmektedir.

metin, diyagram, ekran görüntüsü, plan içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.

Şekil : Federe Öğrenme Akışı Diyagramı

Şekil 1’de federe öğrenme mimarisinin temel işleyiş sürecini özetleyen bir akış diyagramı yer almaktadır. Bu yapı şu şekilde çalışır:

1. Her bir istemci, kendi elindeki verileri kullanarak yerel bir makine öğrenmesi modeli oluşturur ve eğitir.
2. Yerelde eğitilen modellerin yalnızca güncellenmiş parametreleri merkezi sunucuya gönderilir; verilerin kendisi hiçbir zaman paylaşılmaz.
3. Merkezi sunucu, farklı istemcilerden topladığı bu yerel model parametrelerini bir araya getirerek bütüncül bir global model oluşturur. Böylece, her bir istemcinin katkısı küresel modelde temsil edilmiş olur.
4. Global model, tekrar tüm istemcilere dağıtılır. İstemciler bu modeli alarak kendi verileriyle yeniden eğitir ve süreç döngüsel olarak devam eder.

Federe öğrenme, eğitimden sağlık uygulamalarına, güvenlikten finans sektörüne kadar birçok alanda etkili şekilde kullanılmaktadır. Özellikle hasta mahremiyetinin ve veri güvenliğinin ön planda tutulduğu sağlık sektöründe, federe öğrenme tabanlı araştırmalar hem güncelliğini hem de önemini korumaktadır. Bu tez kapsamında, federe öğrenme mimarisiyle sağlık verilerinin yerel cihazlarda işlenmesi sağlanarak, merkezi veri paylaşımına gerek duyulmadan hastalık tahmini üzerine bir uygulama geliştirilmiştir.

Diyabet, dünya genelinde hızla artan bir sağlık problemi olup, erken teşhis büyük bir öneme sahiptir [12]. Öte yandan, American Heart Association’ın güncel verilerine göre, 2050 yılında kardiyovasküler hastalıkların ABD’de 45 milyon yetişkini etkileyerek, yaygınlığının %11,3’ten %15’e yükseleceği öngörülmektedir [13]. Ancak, tıbbi verilerin gizliliği ve güvenliği, klasik makine öğrenimi yaklaşımlarının pratikte uygulanmasını sınırlayan başlıca sorunlardan biri olmayı sürdürmektedir. Federe öğrenme, bu zorluğu aşarak, hasta verilerinin yerel olarak analiz edilmesini ve veri mahremiyetinin korunmasını mümkün kılmaktadır.

Bu çalışmanın öne çıkan katkıları aşağıdaki gibi sıralanabilir:

* Federe Öğrenme ile Yenilikçi Yaklaşım: Bu araştırmada, federated learning ortamlarında model birleştirme algoritmalarından biri olan duCBA yöntemi kullanılmış ve sağlık verisi üzerinde uygulanmıştır. Geliştirilen sistem, veri gizliliğini ön planda tutarken hastalık tahmininde yüksek başarı sağlamakta; ayrıca, dağıtık sağlık verilerinin analizinde karşılaşılan temel zorluklara çözüm getirmektedir.
* Sağlık Alanında Uygulama ve Geçerlilik: Sunulan yöntem, iki farklı ve kritik sağlık veri seti üzerinde sınanmıştır: Diabetes Health Indicators Dataset ve Heart Attack Dataset. Elde edilen benzer ve tutarlı sonuçlar, duCBA algoritmasının risk analizi ve erken hastalık tespiti konularında esnek ve güvenilir şekilde uyarlanabileceğini göstermektedir.
* Gelişmiş Özellik Seçimi ile Model Performansını Yükseltme: Çalışmada uygulanan gelişmiş özellik seçimi teknikleri sayesinde, modelin giriş verileri iyileştirilmiş; bu da hastalık tahminlerinde doğruluk oranının artmasını sağlamıştır. Aynı zamanda, sağlık uygulamalarında büyük önem taşıyan veri gizliliği ilkesinden ödün verilmemiştir.
* Ölçeklenebilir ve Güvenli Çözüm: Çalışma kapsamında geliştirilen çözüm hem veri güvenliği hem de federated model ölçeklenebilirliği gibi önemli problemlere pratik ve uygulanabilir yaklaşımlar sunmaktadır. Bu yönüyle, özellikle hassasiyet gerektiren sağlık hizmetlerinde federated learning tabanlı sistemlerin gerçek dünya uygulamalarının önünü açmaktadır.

Araştırmanın devamında; ikinci bölümde federated learning’e dair mevcut çalışmalar ele alınmakta, üçüncü bölümde geliştirilen uygulama detaylandırılmakta, dördüncü bölümde deneysel sonuçlar sunulmakta, beşinci bölümde bulgular tartışılmakta ve altıncı bölümde genel bir değerlendirme yapılmaktadır.

1. Literatür Taraması

Bu bölümde, federated learning (federe öğrenme) mimarisinin genel ilkeleri ve sağlık alanındaki uygulamalarına ilişkin güncel literatür incelenmektedir.

Federe öğrenme, 2016 yılında Google tarafından veri gizliliğini ön planda tutmak amacıyla geliştirilmiş olup, çok sayıda avantajı beraberinde getirmiştir. Bu mimari sayesinde, model eğitimi sırasında verilerin gizliliği korunurken; eğitim süreleri de kısalmakta, model güncellemeleri veri yerine istemciler ve sunucu arasında aktarılmaktadır. Böylece bant genişliği ve iletişim gecikmesi sorunları ortadan kaldırılmakta ve merkezi sunucuda veri depolama gereksinimi azalarak maliyetler düşmektedir. Bununla birlikte, federe öğrenmenin önünde halen çözülmesi gereken dört temel zorluk bulunmaktadır: iletişim maliyetleri, sistem heterojenliği, istatistiksel heterojenlik ve gizlilik kaygıları.

Yerel cihaz sayısı arttıkça model güncellemelerinin merkezi sunucuya iletilmesi ve yönetimi karmaşıklaşmakta, bu da iletişim maliyeti sorununu doğurmaktadır [14]. Sistem heterojenliği, eğitimde kullanılan cihazlar arasındaki donanım ve yazılım farklarının performans üzerinde oluşturduğu etkiyi tanımlar [15]. Cihazların topladığı verilerdeki dağılım farklılıkları ise istatistiksel heterojenlik olarak bilinir ve bu durum bazı cihazlarda model başarısının düşmesine neden olabilmektedir [16]. Her ne kadar federe öğrenme sayesinde veriler merkezi olarak toplanmasa da paylaşılan model parametreleri üzerinden hassas bilgilere ulaşılabileceği endişesi sürmektedir. Bu nedenle, çeşitli gizlilik artırıcı algoritmalar geliştirilmiştir [15-17].

Literatürde federe öğrenme mimarisi için farklı çözüm yaklaşımları öne çıkmıştır.

* FedAvg: 2017’de Brendan McMahan ve ekibi tarafından önerilen bu yöntem, istemcilerde eğitilen modellerin ağırlıklarının ortalamasının alınması esasına dayanır. FedAvg, merkezi veri toplama ihtiyacını ortadan kaldırırken, iletişim maliyetlerini azaltır ve veri gizliliğini korur [18].
* FedGiA: İletişim yükünü ve istatistiksel heterojenlik sorununu hibrit bir yaklaşımla azaltmayı hedefler [19].
* FedAvu: Katılımcıların kendi hızlarında model eğitebilmesine olanak tanıyarak sistem heterojenliğini adresler [20].
* FairFed: İstatistiksel heterojenlik sorununa çözüm getirerek farklı gruplar arasında dengeli model performansı sunmayı amaçlar [21].
* RoFL: Güvenli gradyan toplama ve şifreleme teknolojileriyle veri gizliliği ve güvenliğini artırır [22].

Bu çalışmada ise, federated learning ortamına özgü birleştirme yöntemi olan duCBA algoritması tercih edilmiştir. duCBA, CBA algoritması ile yerelde eğitilen modellerden elde edilen kuralları merkezi sunucuda birleştirerek destek ve güven değerlerini yeniden hesaplar; aynı etikete sahip kuralların katkı oranını göz önünde bulundurur ve nihai modelin temeli olacak bir kural listesi oluşturur [23]. Bu sayede, veri merkezi olarak paylaşılmadan hem gizlilik korunur hem de ağ trafiği ile veri transferi minimize edilmiş olur.

Federe öğrenme, sağlık verilerinin işlenmesinde güvenlik ve gizlilik standartlarını artıran yenilikçi bir yaklaşım olarak öne çıkmaktadır. Bu yöntemle, hasta verileri yerel cihazlarda analiz edilir, böylece bilgilerin merkezi sunuculara taşınmasına gerek kalmaz; bu da bireysel veri gizliliğini korurken olası güvenlik açıklarından etkilenme riskini azaltır [6-8].

Federe öğrenmenin sağlık hizmetlerindeki etkinliğine dair çok sayıda bilimsel çalışma yapılmıştır.

* Choudhury ve arkadaşları, diferansiyel gizlilikle desteklenen federated learning modellerinin, hassas sağlık verilerinin yerel olarak işlenmesini sağlayarak veri sızıntısı riskini önemli ölçüde azalttığını raporlamıştır [24].
* Horvath ve ekibi, MIMIC-III veri setinde, dağıtık öğrenme süreçleri ve diferansiyel gizlilik uygulamalarının, hasta verisi gizliliğini koruyarak etkili modeller üretmeye imkan verdiğini göstermiştir [25].
* Durga ve arkadaşları, blockchain tabanlı güvenlik önlemleriyle entegre edilen federe öğrenme sistemlerinin, merkezi bir aracıya ihtiyaç olmadan güvenli veri paylaşımını mümkün kıldığını vurgulamıştır [26].
* Chen’in çalışmasında, diyabet tahmini için farklı makine öğrenmesi algoritmalarının doğruluk performansları kıyaslanmış, özellik görselleştirme analizinin model başarısına katkısı gözlemlenmiştir [27].
* Hama Saeed ise, tip 2 diyabet sınıflandırmasında dengesiz veri kümelerinin up-sampling tekniğiyle dengelenmesinin model doğruluğunu artırdığını göstermiştir [28].
* Kumar ve Choudhury’nin çalışmasında, kalp hastalığı tahmininde Random Forest, Lojistik Regresyon ve Destek Vektör Makineleri gibi algoritmalarla yüksek başarı elde edilmiş ve bu modellerin risk faktörlerinin analizinde etkinliği gösterilmiştir [29]
* Gupta ve ekibi, UCI veri setleriyle federated learning ortamında kalp hastalığı tahmininin yüksek doğrulukla yapılabildiğini göstermiştir [30].

Bu araştırmaların ortak noktası, federated learning’in sağlık sektöründe veri gizliliği, güvenlik ve sistem verimliliği gibi temel ihtiyaçlara etkin çözümler sunmasıdır. Mevcut literatür, federe öğrenmenin tıbbi verilerde hem gizliliği hem de verimliliği artırıcı etkisini net biçimde ortaya koymaktadır.

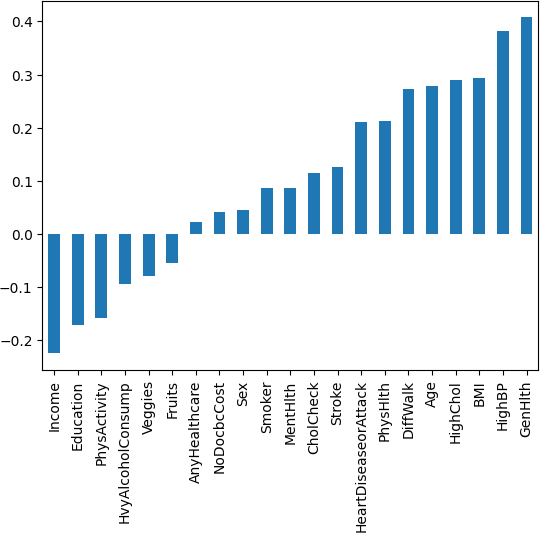
Yine de literatürdeki mevcut federated learning yaklaşımları çoğunlukla model güncellemeleriyle ilgili heterojenlik, iletişim maliyeti ve istatistiksel dengesizlik gibi sorunlara odaklanmıştır. İlişkisel sınıflandırmaya dayalı federe öğrenme mimarileri ise yeterince ele alınmamıştır. duCBA algoritması, özellikle CBA tabanlı ilişkisel sınıflandırmayı federated learning ortamına entegre ederek hem açıklanabilir hem de optimize bir model elde etmeyi amaçlamaktadır. Farklı istemcilerden gelen kuralların destek ve güven değerleri güncellenerek, daha bütünleşik ve analiz edilebilir bir yapı oluşturulmaktadır. Böylece, sadece mevcut durumların değil, aynı zamanda gelecekteki eğilimlerin ve risklerin de öngörülmesi mümkün olmaktadır. Literatürde ilişkisel kuralların gizlilikten ödün verilmeden federe öğrenmeye uyarlanması konusunda önemli bir boşluk bulunmakta; duCBA algoritması bu ihtiyaca yenilikçi ve etkili bir çözüm sunmaktadır. Bu yönüyle duCBA, hem model doğruluğu hem de sistem verimliliği açısından literatürdeki pek çok yöntemden ayrışmaktadır.

1. Kullanılan Yöntemler
   1. Veri Setleri

Bu çalışmada, modellerin hem eğitimi hem de değerlendirilmesi amacıyla Kaggle platformunda yayımlanan iki farklı veri setinden yararlanılmıştır: Diabetes Health Indicators Dataset [31] ve Heart Attack Dataset [32]. Her iki veri seti üzerinde de, literatürde yaygın şekilde başvurulan ve etkili sonuçlar veren üç ayrı özellik seçim yöntemi kullanılmıştır: Principal Component Analysis (PCA), Chi-Square ve Mutual Information. PCA algoritması, boyut indirgeme sağlarken veri özelliklerinin toplam varyansını analiz etmeye odaklanır. Chi-Square yöntemi, istatistiksel bağımsızlık testini esas alarak hedef değişken ile anlamlı ilişkisi olan nitelikleri belirler. Mutual Information ise, bilgi teorisine dayalı olarak, özelliklerle hedef değişken arasındaki karşılıklı bilginin gücünü ölçer. Bu algoritmaların her biri, farklı bakış açıları sunduğu için tercih edilmiş; bu sayede, değişik özellik seçim tekniklerinin model performansına etkisi dengeli biçimde karşılaştırılabilmiştir.

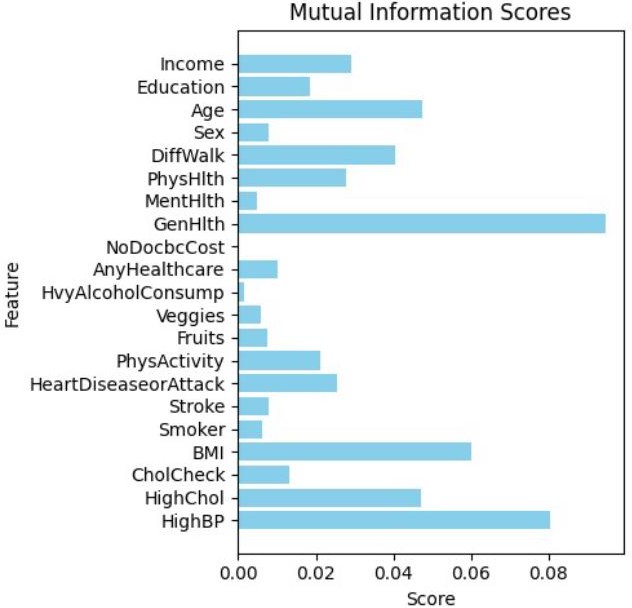
3.1.1 Diabetes Health Indicators Dataset

Diabetes Health Indicators veri seti, ABD Hastalık Kontrol ve Önleme Merkezleri (CDC) tarafından yürütülen Behavioural Risk Factor Surveillance System (BRFSS) anketine dayanmaktadır. Bu veri seti, diyabet hastalığı ve bireylerin genel sağlık durumunu etkileyen çeşitli parametreler hakkında bilgiler içermektedir. Her iki sınıf açısından dengeli olacak şekilde hazırlanmış olan veri seti, toplamda 70.692 satır ve 22 sütundan oluşmaktadır. Bağımlı değişken olarak belirlenen ‘Diabetes\_binary’ sütunu, katılımcının diyabet olup olmadığını ikili biçimde göstermektedir. Kalan 21 bağımsız değişken ise, katılımcıların yaşam alışkanlıkları, sağlık davranışları ve genel sağlık durumu gibi birçok faktörü kapsamaktadır. Bu değişkenler arasında yüksek tansiyon, kolesterol, vücut kitle endeksi (BMI), fiziksel aktivite düzeyi, sigara kullanımı, kalp rahatsızlığı geçmişi, genel sağlık değerlendirmesi ve gelir seviyesi gibi parametreler yer almaktadır. Veri serileri üzerinde yapılan incelemeler sonucunda, veri setinin ek bir ön işleme gerektirmediğine karar verilmiştir.



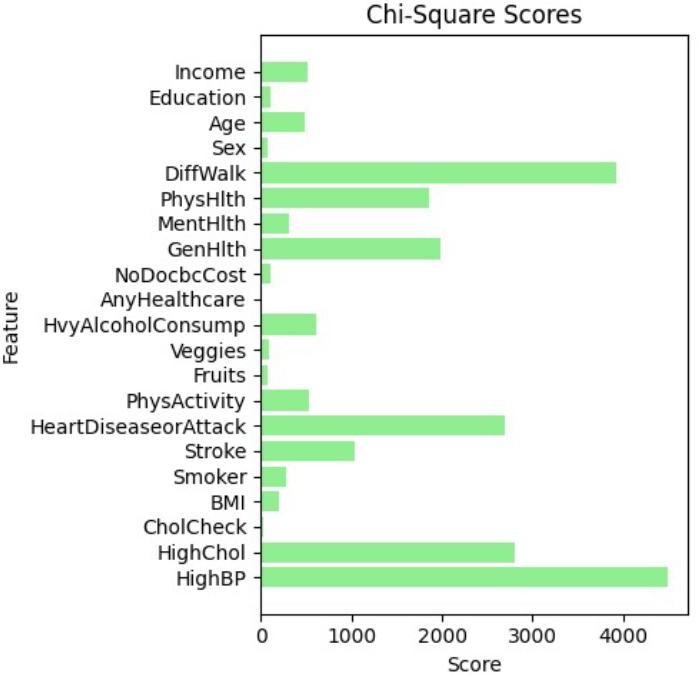
Şekil : Bağımsız Değişkenlerin Diyabet\_binary ile Korelasyonu

Şekil 2'de, veri setinde yer alan bağımsız değişkenlerin diyabet ile olan korelasyonları görselleştirilmiştir. Grafikte, her bir değişkenin diyabet üzerindeki etkisi pozitif veya negatif yönde olabilmektedir. Pozitif korelasyon değerleri, ilgili özelliğin diyabet riskini yükselttiğini; negatif değerler ise riskin azaldığını göstermektedir. Örneğin, ‘GenHlth’ ve ‘HighBP’ değişkenlerinin diyabet ile arasında güçlü bir ilişki olduğu görülmekte ve bu durum, bu faktörlerin diyabet riskini artırabileceğini ortaya koymaktadır. Buna karşılık, ‘Gelir’ gibi değişkenler diyabetle negatif bir korelasyon gösterdiğinden, yüksek gelir seviyesine sahip bireylerde diyabet riskinin daha düşük olabileceği sonucu çıkarılabilir.



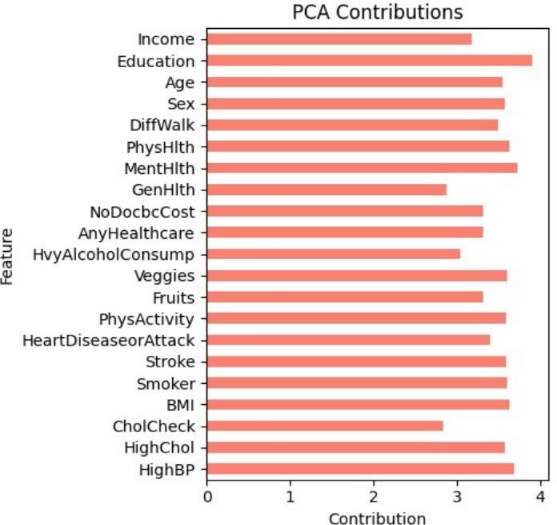
Şekil : Diabetes\_binary için Mutual Informatıon Skorları

Şekil 3’te, özelliklerin hedef değişkenle olan bilgi paylaşım düzeyini gösteren Mutual Information skorları incelenmiştir. Sonuçlara göre, GenHlth, HighBP ve BMI gibi değişkenler, yüksek skorlarıyla hedef değişken üzerinde güçlü bir bilgi aktarımı sağlamaktadır. Buna karşılık, CholCheck, Smoker, Stroke, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, MentHlth, Sex ve Education gibi değişkenler düşük skorlar almıştır. Bu durum, söz konusu değişkenlerin modelde anlamlı bir katkı sunmadığını ve çıkarılmalarının model performansı açısından uygun olabileceğini göstermektedir.



Şekil : Diabetes\_binary için Chi-Square Skorları

Şekil 4'te sunulan Chi-Square analizi, veri setindeki kategorik değişkenlerin hedef değişkenle olan istatistiksel bağımlılık düzeyini ortaya koymaktadır. Analiz sonuçlarına göre, HighBP, DiffWalk ve HeartDiseaseorAttack değişkenleri en yüksek Ki-Kare skorlarına sahip olup, hedef değişkenle güçlü bir ilişki sergilemektedir. Buna karşın, CholCheck, Veggies, Fruits, AnyHealthcare ve Sex gibi değişkenler düşük Ki-Kare değerleriyle, hedef değişkenle aralarındaki ilişkinin oldukça zayıf olduğunu göstermektedir. Bu durum, söz konusu değişkenlerin modelleme sürecinde sınırlı katkı sunduğunu ve modelin performansını olumsuz etkilemeden çıkarılabileceklerini düşündürmektedir.



Şekil : Diabetes\_binary için PCA Katkısı

Şekil 5’te yer alan PCA (Principal Component Analysis) sonuçları, her bir değişkenin toplam varyansın açıklanmasına olan katkısını değerlendirmektedir. Analiz bulgularına göre, Fruits, MentHlth ve Veggies değişkenleri, en yüksek katkıyı sağlayan öznitelikler olarak öne çıkmaktadır. Elde edilen PCA skorları, tüm değişkenlerin anlamlı düzeyde varyansa katkıda bulunduğunu göstermekte olup, bu veri seti özelinde herhangi bir özelliğin çıkarılmasını gerektirecek düşük etkiye sahip bir değişken belirlenmemiştir. Bu doğrultuda PCA, özellikle boyut indirgeme ve veri yapısının daha iyi anlaşılması açısından faydalı bir analiz sunmuş; ancak sonuçlar doğrultusunda veri setinden herhangi bir öznitelik elenmemiştir.

Tablo : Dıabetes Health Indıcators Dataset için Özellı̇k Seçı̇mı̇

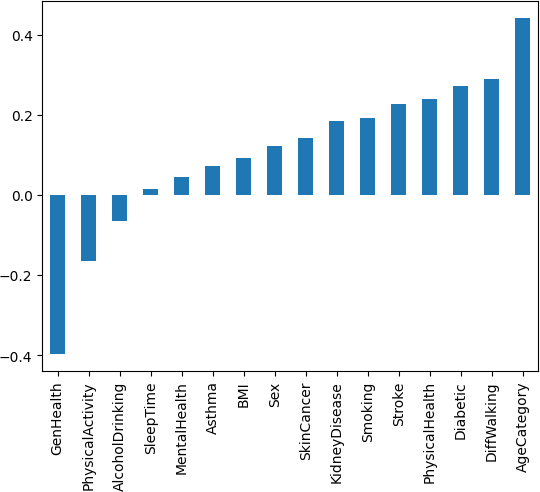
|  |  |  |  |
| --- | --- | --- | --- |
|  | Mutual Information | Chi Square | PCA |
| HighBP | + | + | + |
| HighChol | + | + | + |
| CholCheck | - | - | + |
| BMI | + | + | + |
| Smoker | - | + | + |
| Stroke | - | + | + |
| HeartDiseaseorAttack | + | + | + |
| PhysActivity | - | + | + |
| Fruits | - | - | + |
| Veggies | - | - | + |
| HvyAlcoholConsump | - | + | + |
| AnyHealthcare | - | - | + |
| NoDocbcCost | - | + | + |
| GenHlth | + | + | + |
| MentHlth | - | + | + |
| PhysHlth | + | + | + |
| DiffWalk | + | + | + |
| Sex | - | - | + |
| Age | + | + | + |
| Education | - | + | + |
| Income | + | + | + |

Tablo 1, Diabetes Health Indicators Dataset üzerinde uygulanan üç farklı özellik seçim yönteminin (PCA, Chi-Square ve Mutual Information) sonucunda hangi değişkenlerin öne çıktığını ve modelde kullanılmak üzere seçildiğini göstermektedir.

3.1.2 Heart Attack Dataset

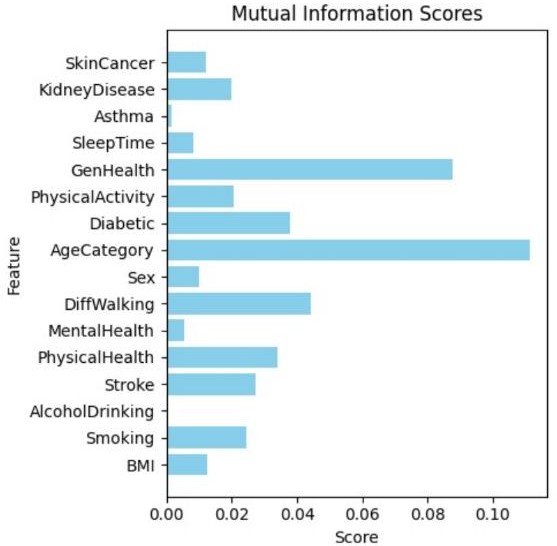
Bu çalışmada, model geliştirme sürecinde Kaggle üzerinden temin edilen Heart Attack Dataset kullanılmıştır. Veri seti, toplamda 319.795 satırdan oluşmakta olup; HeartDisease, BMI, Smoking, Alcohol Drinking, Stroke, PhysicalHealth, MentalHealth, Walking Differently, Gender, AgeCategory, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease ve SkinCancer gibi çeşitli öznitelikler içermektedir. Veri setinde yer alan hedef değişken, bireyin kalp hastalığına sahip olup olmadığını ifade etmektedir (Evet: sağlıklı birey, Hayır: kalp hastalığı mevcut). Bu değişkenlerin tamamı, kalp rahatsızlıklarının erken teşhisinde kullanılabilecek zengin ve anlamlı bir veri kaynağı sunmaktadır. Model eğitimi öncesinde, veri seti üzerinde aşağıda sıralanan ön işleme adımları uygulanmıştır:

* “Evet” ve “Hayır” şeklindeki kategorik değerler, sırasıyla 1 ve 0 olarak sayısal biçime dönüştürülmüştür.
* Cinsiyet bilgisi de benzer şekilde kodlanmış; erkek = 1, kadın = 0 olacak şekilde düzenlenmiştir.
* AgeCategory özelliği, yaş aralıklarına göre 0 ile 12 arasında yeniden sınıflandırılmıştır (örneğin, 0 = 18–24 yaş, 12 = 80 yaş ve üzeri).
* GenHealth (Genel Sağlık Durumu) özelliği de 0–4 arasında puanlanarak yeniden kodlanmıştır (0 = zayıf, 4 = mükemmel).
* Model performansına anlamlı katkı sunmadığı değerlendirilen Race (Irk) değişkeni veri setinden çıkarılmıştır.
* Ayrıca, HeartDisease değişkeninin sınıf dağılımını dengelemek amacıyla veri seti, pozitif ve negatif sınıflar eşit olacak şekilde yeniden örneklenerek dengelenmiştir.



Şekil : Bağımsız Değişkenlerin HeartDisease ile Korelasyonu

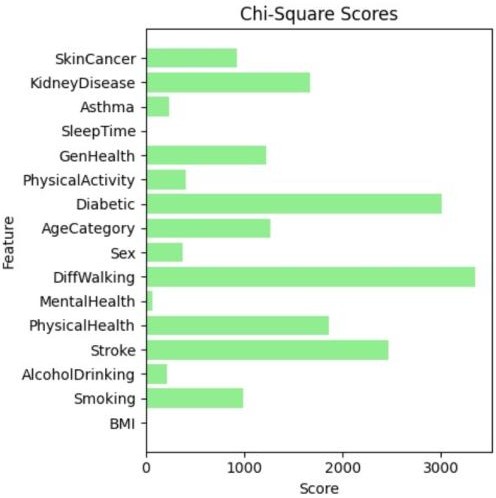
Şekil 6, HeartDisease hedef değişkeni ile veri setindeki diğer öznitelikler arasındaki korelasyonları görsel olarak sunmaktadır. Analiz sonuçlarına göre, kalp hastalığı ile en güçlü pozitif korelasyon, AgeCategory (yaş kategorisi) değişkeniyle gözlemlenmiştir; bu durum, yaş ilerledikçe kalp hastalığı riskinin de arttığını göstermektedir. Ayrıca, Diabetes, Smoking, Stroke ve BMI gibi değişkenler de pozitif yönde korelasyon sergileyerek kalp hastalığı açısından önemli risk faktörleri olarak öne çıkmaktadır. Buna karşın, GenHealth (genel sağlık durumu) ve PhysicalActivity (fiziksel aktivite) değişkenleri, hedef değişkenle negatif korelasyon göstermekte; bu da bu faktörlerdeki artışın kalp hastalığı riskini azaltabileceğini işaret etmektedir. Diğer yandan, SleepTime ve MentalHealth gibi bazı özniteliklerin, hedef değişkenle olan ilişkilerinin daha zayıf düzeyde olduğu gözlemlenmiştir.



Şekil : Bağımsız Değişkenlerin HeartDisease ile Korelasyonu

Şekil 7, veri setindeki değişkenlerin hedef değişkenle olan bilgi paylaşım düzeyini ölçmek amacıyla gerçekleştirilen Mutual Information analizini göstermektedir. Bu değerlendirme, AgeCategory ve GenHealth değişkenlerinin, kalp hastalığı tahmini açısından en yüksek etkiye sahip öznitelikler olduğunu ortaya koymaktadır. Ayrıca PhysicalHealth, DiffWalking ve Stroke gibi faktörler de önemli katkılar sunarak modelin bilgi kazanımına anlamlı ölçüde katkı sağlamaktadır. Buna karşılık, Smoking ve SleepTime değişkenlerinin hedef değişkenle olan bilgi paylaşımı daha düşük düzeydedir ve bu özelliklerin modeldeki etkisi sınırlı kalmaktadır.

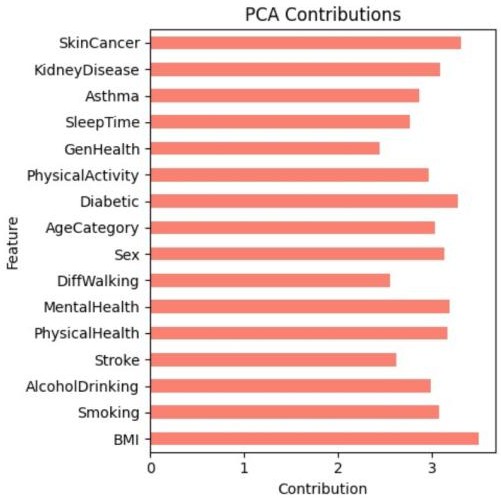
Bu tür analizler, modelleme sürecinde hangi değişkenlerin daha belirleyici olduğunu saptamak ve özellik önceliklendirmesi yapmak açısından kritik öneme sahiptir. Elde edilen bulgular doğrultusunda, bilgi skorları düşük çıkan BMI, AlcoholDrinking, MentalHealth ve Asthma değişkenleri modelleme sürecinden çıkarılmıştır. Böylece modelin gereksiz girdi yükünden arındırılması ve daha etkili sonuçlar üretmesi hedeflenmiştir.



Şekil : HeartDisease için Chi-Square Skorları

Şekil 8, hedef değişken ile öznitelikler arasındaki bilgi ilişkisini değerlendirmek amacıyla uygulanan Chi-Square Skorları analizini göstermektedir. Yapılan analiz, bazı değişkenlerin hedef değişken üzerindeki etkisinin sınırlı olduğunu ortaya koymuştur. Özellikle BMI, MentalHealth ve SleepTime değişkenleri, düşük Chi-Square skorlarına sahip oldukları için modelleme sürecinden çıkarılmıştır. Bu özniteliklerin tahmin gücüne anlamlı katkı sunmaması, modelin hem doğruluğunu artırmak hem de yapısal karmaşıklığını azaltmak adına bu kararın alınmasında etkili olmuştur.

Bu özellik azaltma süreci sayesinde, modelin güçlü ilişkiler barındıran değişkenler üzerine odaklanması sağlanmış ve böylece daha verimli ve anlamlı bir analiz ortamı oluşturulmuştur. Analiz sonucunda öne çıkan AgeCategory, DiffWalking ve Diabetic değişkenlerinin hedef değişken üzerinde yüksek düzeyde etkili olduğu belirlenmiş ve modelin öğrenme süreci bu özniteliklerin önceliğiyle şekillendirilmiştir.



Şekil : HeartDisease için PCA Katkısı

Şekil 9’da sunulan PCA (Principal Component Analysis) sonuçları, her bir değişkenin toplam varyans üzerindeki etkisini değerlendirerek modelin genel yapısına katkı düzeylerini ortaya koymaktadır. Analiz bulgularına göre, herhangi bir değişkenin ihmal edilmesine gerek duyulmamış; tüm öznitelikler modelde tutulmuştur. Özellikle AgeCategory, Diabetic ve DiffWalking değişkenleri, varyansın açıklanmasında öne çıkan ve model performansı açısından kritik öneme sahip öznitelikler olarak belirlenmiştir.

Diğer değişkenlerin katkı düzeyleri bu üç değişken kadar yüksek olmasa da, genel veri yapısının korunması ve analiz sürecinde geniş bir perspektif sunulabilmesi açısından veri setinin bütününün kullanılması tercih edilmiştir. PCA'nın sunduğu bu analiz çerçevesi, özelliklerin varyansa dayalı katkılarının modellenmesi konusunda önemli yol gösterici bilgiler sağlamış ve modelin yapılandırılmasına yön vermiştir.

Tablo : Heart Attack Dataset için Özellı̇k Seçı̇mı̇

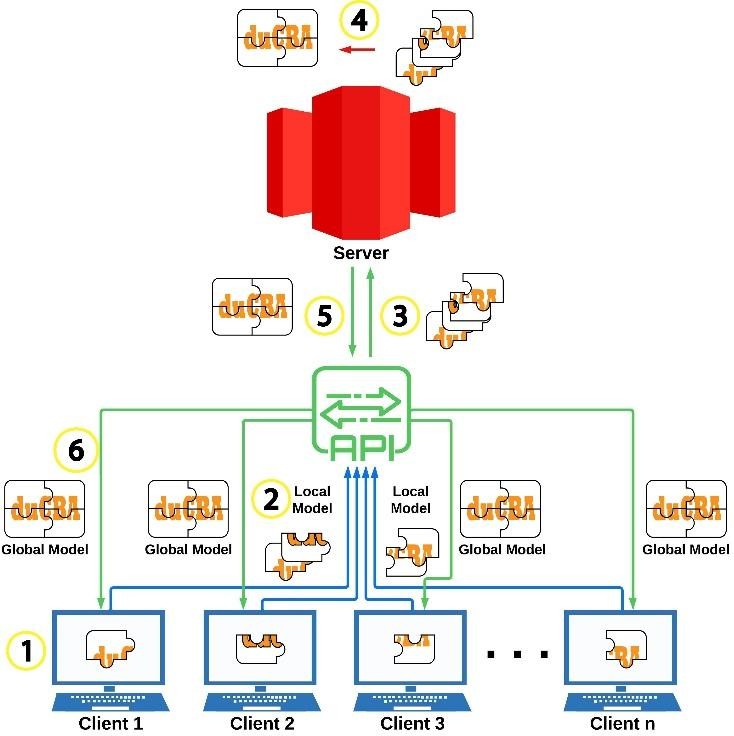
|  |  |  |  |
| --- | --- | --- | --- |
|  | Mutual Information | Chi Square | PCA |
| BMI | - | - | + |
| Smoking | + | + | + |
| AlcoholDrinking | - | + | + |
| Stroke | + | + | + |
| PhysicalHealth | + | + | + |
| MentalHealth | - | - | + |
| DiffWalking | + | + | + |
| Sex | + | + | + |
| AgeCategory | + | + | + |
| Diabetic | + | + | + |
| PhysicalActivity | + | + | + |
| GenHealth | + | + | + |
| SleepTime | + | - | + |
| Asthma | - | + | + |
| KidneyDisease | + | + | + |
| SkinCancer | + | + | + |

Tablo 2, Heart Attack Dataset üzerinde uygulanan üç farklı özellik seçme yönteminin (PCA, Chi-Square ve Mutual Information) sonuçlarını karşılaştırmalı olarak sunmakta olup, her yöntemin hangi öznitelikleri modelleme sürecine dahil ettiğini göstermektedir.

3.2 Prototip Uygulama

Bu araştırma kapsamında geliştirilen prototip sistem, üç temel bileşenden oluşan bir mimari üzerinde yapılandırılmıştır:

* İstemciler (Clients)
* API Sunucusu
* Merkezi Sunucu (Ana Sunucu)



Şekil : duCBA Akışı Diyagramı

Şekil 10, geliştirilen uygulamanın temel işleyişini ve bileşenler arasındaki veri akışını görsel olarak göstermektedir. Uygulamanın ilk aşamasında, her istemci kendi yerel verisini kullanarak CBA (Classification Based on Association) yöntemi ile bir model oluşturur. Eğitim sürecinin ardından istemciler, elde ettikleri yerel modelleri API sunucusuna iletir. API sunucusu ise bu modelleri toplayarak merkezi sunucuya aktarır.

Merkezi sunucu, farklı istemcilerden gelen yerel modelleri duCBA algoritması yardımıyla birleştirir ve güncellenmiş bir küresel model üretir [23]. Ardından, oluşturulan bu güncel model yeniden API sunucusuna gönderilir. API sunucusu, güncellenmiş modeli ilgili istemcilere dağıtarak, her bir istemcinin daha güçlü ve genel geçer bir modeli yerel olarak kullanabilmesini sağlar.

Şekil 10’da gösterilen bu süreç, federe öğrenme sisteminde bir turu temsil etmektedir. Federe mimarinin doğası gereği, bu tur çok sayıda kez tekrar edilmekte ve her döngüde model daha fazla veriyle eğitilerek gelişmektedir. Böylelikle, veriler fiziksel olarak asla merkezi bir noktaya taşınmadan; veri sahipliği korunarak, dağıtık öğrenme gerçekleştirilmektedir. Bu yapı, özellikle veri gizliliği gerektiren sağlık gibi alanlarda etkili ve güvenli modelleme imkânı sunmaktadır.

3.2.1 İstemciler

Prototip sistemde istemciler, Google Cloud Platform üzerinde oluşturulan eş yapıdaki sanal makineler kullanılarak yapılandırılmıştır. Her bir istemci, x86/64 mimarisi üzerinde çalışan e2-medium türünde sanal makinelerdir. Model eğitimi süreci, Python 3.8.10 sürümü ve ilgili kütüphaneler kullanılarak gerçekleştirilmiştir. Bu kütüphaneler arasında numpy, pandas, scikit-learn, pickle, pyfim, requests ve pyarc yer almaktadır.

Model eğitimi sırasında kullanılan veri kümeleri, %20 oranında test verisi olarak ayrılmış; geri kalan %80’lik bölüm ise mevcut istemci sayısına göre eşit şekilde dağıtılmıştır. Yerel modellerin eğitimi için CBA (Classification Based on Association) algoritması tercih edilmiştir. Bu algoritma için support ve confidence eşik değerleri sırasıyla 0.2 ve 0.5 olarak belirlenmiş ve bu değerler daha önce yapılan testler sonucunda uygunluğu onaylanmış ayarlardır [21]. İstemciler tarafından eğitilen modeller, JSON formatında HTTP üzerinden API sunucusuna iletilmektedir. API sunucusu ise bu modelleri toplar ve daha sonra merkezi sunucuya yönlendirir. Global model oluşturulduktan sonra, istemcilere yeniden dağıtılır ve böylece yerel modeller küresel modelle uyumlu hale getirilerek bir sonraki eğitim turuna geçilir.

3.2.2 API

API sunucusu da, istemcilerde olduğu gibi Google Cloud Platform üzerinde konumlandırılmış e2-medium türünde bir sanal makine olarak yapılandırılmıştır. Yazılım ortamı yine Python 3.8.10 olup; Flask, numpy, pandas, scikit-learn, pickle, pyfim, websockets, requests ve pyarc gibi kütüphaneler kullanılmıştır. API sunucusu, istemcilerden gelen talepleri 5000 portu üzerinden dinlemekte ve işlemektedir.

Prototip sistemin test aşamasında, istemcilerden gelen yerel modeller API aracılığıyla 10 tur boyunca toplanmış ve her turda merkezi sunucuya iletilmiştir. Modeller, websockets protokolü aracılığıyla ana sunucuya aktarılmış, orada birleştirildikten sonra API tarafından tekrar istemcilere yönlendirilmiştir.

3.2.2 Ana Sunucu

Sistem mimarisinin merkezinde yer alan ana sunucu, tüm istemcilerden gelen model çıktılarının toplanıp işlenerek küresel bir modelin oluşturulmasından sorumludur. Bu bileşen, Google Cloud Platform üzerinde yapılandırılmış e2-medium türünde sanal bir makine üzerinde çalıştırılmıştır. Ana sunucunun yazılım ortamı, istemci ve API sunucularıyla tutarlı olacak şekilde Python 3.8.10 sürümünde hazırlanmıştır. Uygulamada kullanılan temel Python kütüphaneleri şunlardır: numpy, pandas, scikit-learn, pickle, pyfim, websockets, requests ve pyarc. Ana sunucunun API ile veri alışverişi, 8000 portu üzerinden gerçekleştirilen WebSocket protokolü aracılığıyla sağlanmıştır.

Ana sunucunun temel işlevi, istemcilerde CBA algoritması ile eğitilen yerel modellerin çıktılarını alarak bunları duCBA (Distributed and Unified Classification Based on Association Rules) algoritması aracılığıyla birleştirmek ve optimize bir küresel model üretmektir [23]. Bu süreç, yalnızca modellerin entegrasyonu değil, aynı zamanda her bir kuralın güvenilirlik düzeyinin yeniden değerlendirilmesi gibi ayrıntılı bir işlemi de kapsamaktadır.

Birleştirme işlemi başlatıldığında, farklı istemcilerden gelen kural setleri toplanır. Her bir kuralın support (destek) ve confidence (güven) değerleri tekrar hesaplanarak merkezi olarak analiz edilir. Özellikle aynı etikete sahip fakat farklı istemcilerden gelen kuralların tekrarlılığı dikkate alınır. Böylece, aynı içeriğe sahip kuralların sistem genelinde ne kadar yaygın olduğu gözlemlenir ve bu da her bir kuralın genellenebilirliğini değerlendirmede kullanılır. Bu noktada support değeri öne çıkmakta; hangi kural daha fazla veri örneğinde yer alıyorsa, o kural daha yüksek önem kazanarak listeye dahil edilmektedir.

Buna ek olarak, aynı yapıya sahip ancak farklı etiketlerle sınıflandırılmış kurallar arasında tercih yapılırken önce support, ardından eşitlik durumunda confidence değeri dikkate alınır. Eğer hem support hem de confidence değerleri eşitse, kural sıralaması veri akışına göre yapılır ve listedeki ilk kural daha üst sırada yer alır. Bu sıralama mantığı, modelin tutarlı ve genellenebilir olmasını sağlarken, farklı kaynaklardan gelen bilgilerin dengeli bir biçimde küresel modele yansımasına olanak tanır.

Bu kapsamlı analiz ve sıralama sürecinin sonunda, her istemciden gelen katkıları içeren birleştirilmiş ve optimize edilmiş bir global model oluşturulur. Bu model, daha önce belirtildiği gibi API sunucusu aracılığıyla tüm istemcilere tekrar dağıtılır. Böylece, her istemci bir sonraki tur eğitimine daha güçlü, daha genel ve diğer istemcilerin bilgisini de içeren bir modelle başlar. Bu süreç döngüsel olarak devam eder ve her turda model daha fazla veriye dayalı hale gelerek gelişir.

Bu yapının en önemli avantajlarından biri, veri gizliliğinin korunmasıdır. Veriler fiziksel olarak merkezi sunucuya gönderilmediği için, her istemci yalnızca kendi verisini yerel olarak işler ve sisteme yalnızca model çıktısını aktarır. Bu yaklaşım, özellikle sağlık gibi gizliliğin ön planda olduğu alanlarda federe öğrenme mimarisinin güvenli ve ölçeklenebilir bir çözüm sunduğunu göstermektedir.

1. DENEYSEL SONUÇLAR

Bu bölümde, araştırmanın temel amacı doğrultusunda belirlenen hipotezler ve araştırma soruları çerçevesinde elde edilen analiz sonuçları detaylı biçimde sunulmaktadır. Federe öğrenme mimarisi kullanılarak geliştirilen sistem, iki farklı sağlık veri seti üzerinde, farklı istemci sayıları ve çeşitli algoritmalarla değerlendirilmiştir. Model başarımı, precision, recall, F1-score ve accuracy gibi yaygın sınıflandırma ölçütleriyle ölçülmüştür.

Deneysel çalışmada, sistem 2, 4 ve 8 istemcili olmak üzere üç farklı dağıtık konfigürasyonda test edilmiştir. Her bir dağıtımda, istemciler modellerini kendi yerel verileri üzerinde eğitmişlerdir. Özellik seçimi için Principal Component Analysis (PCA), Chi-Square ve Mutual Information algoritmaları ayrı ayrı uygulanmış; böylece farklı boyut indirgeme tekniklerinin model başarımına etkisi analiz edilmiştir.

Model eğitimi, iki farklı makine öğrenme yöntemi ile gerçekleştirilmiştir: Yapay Sinir Ağları (YSA) ve CBA (Classification Based on Association) algoritması. YSA modeli, giriş katmanında 64 nöronlu bir Dense katman içermekte ve bu katmanda ReLU aktivasyon fonksiyonu kullanılarak doğrusal olmayan dönüşüm sağlanmaktadır. Çıktı katmanında ise tek bir nöron ve sigmoid aktivasyon fonksiyonu kullanılarak, modelin çıktısı iki sınıf arasında olasılık değeri şeklinde hesaplanmaktadır. Optimizasyon sürecinde, öğrenme katsayısı 0.001 olan Adam optimizasyon algoritması kullanılmış; sınıflandırma hataları ise binary\_crossentropy kayıp fonksiyonu ile minimize edilmiştir. Modelin başarımı, doğruluk metriği ile takip edilmiştir. Eğitim süreci, 5 epok boyunca, her bir iterasyonda 32 örnekten oluşan minibatch’ler ile gerçekleştirilmiştir.

Global modelin oluşturulması sürecinde iki farklı model birleştirme algoritması test edilmiştir:

* FedAvg: İstemcilerde eğitilen modellerin ağırlıklı ortalamasını alarak merkezi modeli günceller.
* duCBA: İlişkisel kurallara dayalı olarak istemcilerden gelen kuralları değerlendirip güvenilirliklerine göre yeniden sıralayarak nihai modeli oluşturur.

Her bir test konfigürasyonu (istemci sayısı × özellik seçimi × modelleme yöntemi × birleştirme algoritması) için modeller 10 iterasyon boyunca eğitilmiş ve değerlendirilmiştir. Bu yapı sayesinde istemci sayısının, seçilen özellik kümesinin ve global model oluşturma yönteminin model başarımı üzerindeki etkisi karşılaştırmalı olarak incelenmiştir.

Tablo : Dıabetes Health Indıcators Dataset için Sınıflandırma Raporu

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | Özellik Seçme Algoritmaları | | | | | | | | | | | | |
| Mutual Information | | | | Chi Square | | | | PCA | | | | |
| Precision | Recall | F1- Score | Samples | Precision | Recall | F1- Score | Samples | Precision | Recall | F1- Score | Samples |
| 2 İstemci | duCBA | 0 | 0.70 | 0.69 | 0.69 | 7090 | 0.69 | 0.70 | 0.70 | 7090 | 0.75 | 0.59 | 0.66 | 7090 |
| 1 | 0.69 | 0.70 | 0.70 | 7049 | 0.69 | 0.69 | 0.69 | 7049 | 0.66 | 0.80 | 0.72 | 7049 |
| Accuracy |  | | 0.70 | 14139 |  | | 0.69 | 14139 |  | | 0.70 | 14139 |
| Macro Avg. | 0.70 | 0.70 | 0.70 | 14139 | 0.69 | 0.69 | 0.69 | 14139 | 0.71 | 0.70 | 0.69 | 14139 |
| Weighted Avg. | 0.70 | 0.70 | 0.70 | 14139 | 0.69 | 0.69 | 0.69 | 14139 | 0.71 | 0.70 | 0.69 | 14139 |
| FedAvg | 0 | 0.88 | 0.43 | 0.57 | 7090 | 0.86 | 0.50 | 0.63 | 7090 | 0.79 | 0.67 | 0.72 | 7090 |
| 1 | 0.62 | 0.94 | 0.75 | 7049 | 0.65 | 0.91 | 0.76 | 7049 | 0.71 | 0.82 | 0.76 | 7049 |
| Accuracy |  | | 0.68 | 14139 |  | | 0.71 | 14139 |  | | 0.74 | 14139 |
| Macro Avg. | 0.75 | 0.68 | 0.66 | 14139 | 0.75 | 0.71 | 0.70 | 14139 | 0.75 | 0.74 | 0.74 | 14139 |
| Weighted Avg. | 0.75 | 0.68 | 0.66 | 14139 | 0.75 | 0.71 | 0.70 | 14139 | 0.75 | 0.74 | 0.74 | 14139 |
| 4 İstemci | duCBA | 0 | 0.70 | 0.69 | 0.69 | 7090 | 0.70 | 0.70 | 0.70 | 7090 | 0.68 | 0.73 | 0.71 | 7090 |
| 1 | 0.69 | 0.70 | 0.70 | 7049 | 0.70 | 0.70 | 0.70 | 7049 | 0.71 | 0.66 | 0.68 | 7049 |
| Accuracy |  | | 0.70 | 14139 |  | | 0.70 | 14139 |  | | 0.69 | 14139 |
| Macro Avg. | 0.70 | 0.70 | 0.70 | 14139 | 0.70 | 0.70 | 0.70 | 14139 | 0.70 | 0.69 | 0.69 | 14139 |
| Weighted Avg. | 0.70 | 0.70 | 0.70 | 14139 | 0.70 | 0.70 | 0.70 | 14139 | 0.70 | 0.69 | 0.69 | 14139 |
| FedAvg | 0 | 0.80 | 0.66 | 0.72 | 7090 | 0.78 | 0.68 | 0.73 | 7090 | 0.79 | 0.65 | 0.72 | 7090 |
| 1 | 0.71 | 0.83 | 0.76 | 7049 | 0.71 | 0.81 | 0.76 | 7049 | 0.70 | 0.83 | 0.76 | 7049 |
| Accuracy |  | | 0.74 | 14139 |  | | 0.74 | 14139 |  | | 0.74 | 14139 |
| Macro Avg. | 0.75 | 0.74 | 0.74 | 14139 | 0.75 | 0.74 | 0.74 | 14139 | 0.75 | 0.74 | 0.74 | 14139 |
| Weighted Avg. | 0.75 | 0.74 | 0.74 | 14139 | 0.75 | 0.74 | 0.74 | 14139 | 0.75 | 0.74 | 0.74 | 14139 |
| 8 İstemci | duCBA | 0 | 0.67 | 0.75 | 0.71 | 7090 | 0.68 | 0.73 | 0.71 | 7090 | 0.68 | 0.68 | 0.68 | 7090 |
| 1 | 0.72 | 0.63 | 0.67 | 7049 | 0.71 | 0.65 | 0.68 | 7049 | 0.68 | 0.64 | 0.67 | 7049 |
| Accuracy |  | | 0.69 | 14139 |  | | 0.69 | 14139 |  | | 0.69 | 14139 |
| Macro Avg. | 0.69 | 0.69 | 0.69 | 14139 | 0.69 | 0.69 | 0.69 | 14139 | 0.69 | 0.69 | 0.69 | 14139 |
| Weighted Avg. | 0.69 | 0.69 | 0.69 | 14139 | 0.69 | 0.69 | 0.69 | 14139 | 0.69 | 0.69 | 0.69 | 14139 |
| FedAvg | 0 | 0.79 | 0.63 | 0.70 | 7090 | 0.79 | 0.66 | 0.71 | 7090 | 0.73 | 0.71 | 0.72 | 7090 |
| 1 | 0.69 | 0.84 | 0.76 | 7049 | 0.70 | 0.82 | 0.76 | 7049 | 0.72 | 0.73 | 0.72 | 7049 |
| Accuracy |  | | 0.73 | 14139 |  | | 0.74 | 14139 |  | | 0.72 | 14139 |
| Macro Avg. | 0.74 | 0.73 | 0.73 | 14139 | 0.74 | 0.74 | 0.74 | 14139 | 0.72 | 0.72 | 0.72 | 14139 |
| Weighted Avg. | 0.74 | 0.73 | 0.73 | 14139 | 0.74 | 0.74 | 0.74 | 14139 | 0.72 | 0.72 | 0.72 | 14139 |

Tablo 3, model eğitiminin Diabetes Health Indicators Dataset kullanılarak gerçekleştirildiği senaryolarda elde edilen performans sonuçlarını sunmaktadır. Bu tabloda, üç farklı özellik seçim yöntemi (Mutual Information, Chi-Square, PCA) ile iki farklı model toplama algoritmasının (FedAvg ve duCBA) 2, 4 ve 8 istemcili federe ortamlardaki başarımları karşılaştırılmıştır.

Elde edilen bulgulara göre, FedAvg, özellikle istemci sayısının arttığı durumlarda daha yüksek F1-Skoru ve Precision değerleriyle öne çıkmaktadır. Buna karşın, duCBA algoritması düşük ve orta ölçekli istemci sayılarına sahip senaryolarda daha istikrarlı ve dengeli sonuçlar üretmiştir. Bu durum, duCBA’nın sınırlı sayıda katılımcı bulunan dağıtık sistemlerde kararlı öğrenme sağlayabildiğini göstermektedir.

Farklı özellik seçim algoritmalarıyla elde edilen benzer performans değerleri, duCBA algoritmasının ön işleme çeşitliliğine karşı toleranslı ve uyumlu bir yapıya sahip olduğunu ortaya koymaktadır. İstemci sayısındaki artış bazı senaryolarda minör dalgalanmalara neden olmuş olsa da, duCBA algoritması genel doğruluk açısından yaklaşık %69 ile %70 arasında sabit kalan değerlerle dikkat çekmektedir. Bu istikrar, veri gizliliği ön planda olan federe öğrenme ortamlarında güvenilirlik kriteri açısından önemli bir avantaj sunmaktadır.

Sonuçlar, duCBA’nın özellikle kararlılık ve güven gerektiren dağıtık sistem uygulamaları için tercih edilebilecek etkili bir yöntem olduğunu desteklemektedir.

Tablo : Heart Attack Dataset için Sınıflandırma Raporu

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | Özellik Seçme Algoritmaları | | | | | | | | | | | | |
| Mutual Information | | | | Chi Square | | | | PCA | | | | |
| Precision | Recall | F1- Score | Samples | Precision | Recall | F1- Score | Samples | Precision | Recall | F1- Score | Samples |
| 2 İstemci | duCBA | 0 | 0.69 | 0.64 | 0.67 | 5296 | 0.69 | 0.64 | 0.67 | 5296 | 0.69 | 0.64 | 0.67 | 5296 |
| 1 | 0.67 | 0.72 | 0.69 | 5295 | 0.67 | 0.71 | 0.69 | 5295 | 0.67 | 0.71 | 0.69 | 5295 |
| Accuracy |  | | 0.68 | 10591 |  | | 0.68 | 10591 |  | | 0.68 | 10591 |
| Macro Avg. | 0.68 | 0.68 | 0.68 | 10591 | 0.68 | 0.68 | 0.68 | 10591 | 0.68 | 0.68 | 0.68 | 10591 |
| Weighted Avg. | 0.68 | 0.68 | 0.68 | 10591 | 0.68 | 0.68 | 0.68 | 10591 | 0.68 | 0.68 | 0.68 | 10591 |
| FedAvg | 0 | 0.76 | 0.73 | 0.74 | 5296 | 0.79 | 0.70 | 0.74 | 5296 | 0.77 | 0.73 | 0.75 | 5296 |
| 1 | 0.74 | 0.76 | 0.75 | 5295 | 0.73 | 0.81 | 0.77 | 5295 | 0.74 | 0.78 | 0.76 | 5295 |
| Accuracy |  | | 0.75 | 10591 |  | | 0.76 | 10591 |  | | 0.76 | 10591 |
| Macro Avg. | 0.75 | 0.75 | 0.75 | 10591 | 0.76 | 0.76 | 0.76 | 10591 | 0.76 | 0.76 | 0.76 | 10591 |
| Weighted Avg. | 0.75 | 0.75 | 0.75 | 10591 | 0.76 | 0.76 | 0.76 | 10591 | 0.76 | 0.76 | 0.76 | 10591 |
| 4 İstemci | duCBA | 0 | 0.69 | 0.64 | 0.66 | 5296 | 0.68 | 0.65 | 0.66 | 5296 | 0.68 | 0.65 | 0.66 | 5296 |
| 1 | 0.66 | 0.70 | 0.68 | 5295 | 0.66 | 0.70 | 0.68 | 5295 | 0.66 | 0.70 | 0.68 | 5295 |
| Accuracy |  | | 0.67 | 10591 |  | | 0.67 | 10591 |  | | 0.67 | 10591 |
| Macro Avg. | 0.68 | 0.67 | 0.67 | 10591 | 0.67 | 0.67 | 0.67 | 10591 | 0.67 | 0.67 | 0.67 | 10591 |
| Weighted Avg. | 0.68 | 0.67 | 0.67 | 10591 | 0.67 | 0.67 | 0.67 | 10591 | 0.67 | 0.67 | 0.67 | 10591 |
| FedAvg | 0 | 0.74 | 0.76 | 0.75 | 5296 | 0.75 | 0.71 | 0.73 | 5296 | 0.75 | 0.75 | 0.75 | 5296 |
| 1 | 0.75 | 0.73 | 0.74 | 5295 | 0.73 | 0.77 | 0.75 | 5295 | 0.75 | 0.75 | 0.75 | 5295 |
| Accuracy |  | | 0.75 | 10591 |  | | 0.74 | 10591 |  | | 0.75 | 10591 |
| Macro Avg. | 0.75 | 0.75 | 0.75 | 10591 | 0.74 | 0.74 | 0.74 | 10591 | 0.75 | 0.75 | 0.75 | 10591 |
| Weighted Avg. | 0.75 | 0.75 | 0.75 | 10591 | 0.74 | 0.74 | 0.74 | 10591 | 0.75 | 0.75 | 0.75 | 10591 |
| 8 İstemci | duCBA | 0 | 0.64 | 0.80 | 0.71 | 5296 | 0.63 | 0.83 | 0.72 | 5296 | 0.63 | 0.81 | 0.71 | 5296 |
| 1 | 0.73 | 0.55 | 0.63 | 5295 | 0.75 | 0.52 | 0.61 | 5295 | 0.73 | 0.54 | 0.62 | 5295 |
| Accuracy |  | | 0.67 | 10591 |  | | 0.67 | 10591 |  | | 0.67 | 10591 |
| Macro Avg. | 0.69 | 0.67 | 0.67 | 10591 | 0.69 | 0.67 | 0.66 | 10591 | 0.68 | 0.67 | 0.66 | 10591 |
| Weighted Avg. | 0.69 | 0.67 | 0.67 | 10591 | 0.69 | 0.67 | 0.66 | 10591 | 0.68 | 0.67 | 0.66 | 10591 |
| FedAvg | 0 | 0.79 | 0.68 | 0.73 | 5296 | 0.79 | 0.68 | 0.73 | 5296 | 0.77 | 0.70 | 0.73 | 5296 |
| 1 | 0.72 | 0.82 | 0.82 | 5295 | 0.72 | 0.82 | 0.77 | 5295 | 0.72 | 0.79 | 0.75 | 5295 |
| Accuracy |  | | 0.75 | 10591 |  | | 0.75 | 10591 |  | | 0.74 | 10591 |
| Macro Avg. | 0.75 | 0.75 | 0.75 | 10591 | 0.76 | 0.75 | 0.75 | 10591 | 0.75 | 0.74 | 0.74 | 10591 |
| Weighted Avg. | 0.75 | 0.75 | 0.75 | 10591 | 0.76 | 0.75 | 0.75 | 10591 | 0.75 | 0.74 | 0.74 | 10591 |

Tablo 4, model eğitiminin Heart Attack Dataset kullanılarak gerçekleştirildiği deneylerde elde edilen performans sonuçlarını sunmaktadır. Bu sonuçlar, üç farklı özellik seçim algoritması (Mutual Information, Chi-Square, PCA) ile iki model birleştirme yaklaşımının (FedAvg ve duCBA) 2, 4 ve 8 istemcili dağıtık yapıdaki sistemlerde gösterdiği performansı karşılaştırmalı olarak ortaya koymaktadır.

Elde edilen bulgulara göre, FedAvg, özellikle istemci sayısının arttığı durumlarda F1-Score ve Precision metriklerinde daha yüksek değerler üretmiştir. Ancak duCBA algoritması, daha az sayıda istemcinin bulunduğu konfigürasyonlarda dengeli ve tutarlı bir performans sergileyerek öne çıkmaktadır. Ayrıca duCBA’nın, farklı özellik seçim algoritmaları ile benzer performans sonuçları vermesi, bu yöntemin veri ön işleme sürecine karşı esnek ve dayanıklı olduğunu göstermektedir.

Bazı test senaryolarında istemci sayısındaki artışa bağlı olarak metriklerde küçük dalgalanmalar gözlemlenmiş olsa da, duCBA algoritması genel doğruluk değerlerini %67–68 bandında sabit tutarak, özellikle küçük ve orta ölçekli sistemlerde güvenilir bir çözüm olarak öne çıkmıştır. Bu durum, duCBA'nın sistem boyutundan bağımsız olarak istikrarlı bir şekilde çalışabildiğini kanıtlamaktadır.

Genel değerlendirme açısından bakıldığında, duCBA algoritması, düşük istemcili ve veri gizliliği öncelikli senaryolarda kararlı ve güvenilir sonuçlar sağlamıştır. Özellik seçim yöntemi fark etmeksizin elde edilen benzer skorlar, algoritmanın çeşitli ön işleme teknikleriyle başarılı bir şekilde entegre olabildiğini ortaya koymaktadır. Öte yandan, FedAvg, yüksek istemci sayısına sahip dağıtık yapılarda öne çıkarken; duCBA, daha az istemcinin bulunduğu, hassasiyet gerektiren uygulamalar için tercih edilebilir bir alternatif sunmaktadır.

Bu sonuçlar, duCBA’nın özellikle gizlilik odaklı, küçük ve orta ölçekli federe sistemlerde kullanılabilecek istikrarlı, esnek ve güvenilir bir öğrenme mekanizması olduğunu açıkça ortaya koymaktadır.

1. TARTIŞMA

Bu bölümde, yürütülen çalışmada elde edilen deneysel bulgulara dair yorumlar sunulmakta ve elde edilen sonuçlar üzerinden çıkarımlar yapılmaktadır.

Gerçekleştirilen çalışmada, iki farklı sağlık veri kümesi kullanılarak federe öğrenme mimarisi ile model eğitimi gerçekleştirilmiştir. İstemcilerde yerel olarak eğitilen modeller, merkezi sunucuda iki farklı federe birleştirme algoritması yardımıyla birleştirilmiştir. Bu algoritmalardan ilki olan FedAvg, istemcilerde eğitilen modellerin parametrelerini alarak bunların ortalamasını hesaplamakta ve merkezi bir model oluşturmaktadır. Federated learning alanında en yaygın kullanılan yöntemlerden biri olan FedAvg, ağırlıklı ortalama yaklaşımıyla çalıştığı için yapısal olarak sade ve etkili bir yöntemdir.

İkinci olarak kullanılan yöntem ise, bu çalışmada önerilen duCBA algoritmasıdır. duCBA, ilişkisel sınıflandırma kuralları temelinde çalışan, şeffaflığı ve açıklanabilirliği esas alan bir toplulaştırma yöntemidir. Her istemcide CBA algoritması ile elde edilen kural kümeleri, merkezi sunucuya aktarılmakta; burada kurallar support (destek) ve confidence (güven) değerlerine göre yeniden değerlendirilmektedir. Kurallar arasında önceliklendirme ve sıralama işlemleri uygulanarak nihai model oluşturulmaktadır. Bu sayede, model birleştirme sürecinin her adımı tanımlı, izlenebilir ve anlaşılabilir hale gelmekte; böylece modelin karar mekanizması açıklanabilir yapıya kavuşmaktadır. Bu yönüyle duCBA, XAI (Explainable Artificial Intelligence) yaklaşımı kapsamında önemli bir katkı sunmaktadır. Buna karşın, FedAvg algoritması ile elde edilen nihai modelin, karar sürecine dair şeffaflık ve yorumlanabilirlik açısından sınırlı bir yapıya sahip olduğu görülmüştür.

Bu çalışmada geliştirilen duCBA algoritması, özellikle açıklanabilirlik gerektiren senaryolarda etkili bir çözüm sunmaktadır. Her bir veri kümesi özelinde elde edilen kurallar ayrı ayrı analiz edilmiş; support ve confidence değerlerine göre sıralanan en güçlü kurallar detaylı biçimde yorumlanmıştır.

Aşağıda, Diabetes Health Indicators Dataset için duCBA ile oluşturulan nihai modelde öne çıkan kuralların yorumu sunulmaktadır:

**A. Genel Sağlık Durumu ve Diyabet Riski**

* Diyabet riski düşük bireylerde, belirli sağlık göstergelerinin normal aralıkta olduğu ve ciddi sağlık problemlerinin bulunmadığı gözlemlenmiştir:
* Yüksek Tansiyonun Bulunmaması (HighBP = 0.0): Normal tansiyon seviyeleri, insülin direnci ve metabolik sendromla ilişkili risklerin azalmasına katkı sağlar.
* Kolesterol Düzeylerinin Normal Olması (HighChol = 0.0): Düşük kolesterol, genel metabolik sağlığın iyi olduğunu ve dolaylı olarak diyabet riskinin azaldığını gösterir.
* Yürüme Zorluğunun Olmaması (DiffWalk = 0.0): Hareket kabiliyetinin korunmuş olması, daha aktif bir yaşam tarzını işaret eder.
* İnme ve Kalp Hastalığı Öyküsünün Olmaması (Stroke = 0.0, HeartDiseaseorAttack = 0.0): Kardiyovasküler rahatsızlıkların yokluğu, diyabetin ilerlemiş komplikasyonlarının görülmediğini gösterir.

Bu bulgular, yaşam tarzına bağlı sağlık alışkanlıklarının (dengeli beslenme, egzersiz, stresten uzak yaşam) diyabetten korunmada kritik rol oynadığını göstermektedir.

**B. Fiziksel Aktivite ve Diyetin Koruyucu Etkisi**

Sağlıklı beslenme ve düzenli fiziksel aktivitenin diyabet riski üzerindeki olumlu etkisi açıkça görülmektedir.

Örnek Kural: {PhysActivity = 1.0, Veggies = 1.0, DiffWalk = 0.0, HighBP = 0.0} → Diabetes\_binary = 0.0 (%79.4 güven düzeyi)

Bu kural, sebze ağırlıklı diyetin ve aktif yaşam tarzının insülin hassasiyetini artırabileceğini ve kilo kontrolünü destekleyerek kan şekeri düzeylerini düzenleyebileceğini ortaya koymaktadır.

**C. Sağlık Hizmetlerine Erişim ve Diyabet İlişkisi**

Sağlık hizmetlerine kolay erişim, bireylerin diyabet tanısı ve takibi açısından avantaj sağlamaktadır.

Örnek Kural: {HighBP = 0.0, AnyHealthcare = 1.0, NoDocbcCost = 0.0} → Diabetes\_binary = 0.0 (%81.6 güven düzeyi)

Düzenli sağlık kontrolleri ve maddi engellerin olmaması, diyabetin erken tespiti ve yönetimi açısından önemlidir. Bu bulgu, sağlık politikalarının erişilebilirlik temelinde şekillendirilmesi gerektiğini desteklemektedir.

**D. Çoklu Risk Faktörlerinin Yokluğu**

Örnek Kural: {HighChol = 0.0, Stroke = 0.0, DiffWalk = 0.0, HeartDiseaseorAttack = 0.0, HighBP = 0.0} → Diabetes\_binary = 0.0 (%84.3 güven düzeyi)

Bu kural, tek bir faktöre değil, bireyin genel sağlık profiline odaklanmanın diyabetten korunmada daha etkili olduğunu göstermektedir.

Sonuç olarak, bu çalışmada elde edilen kural tabanlı analizler, diyabet üzerinde etkili olan başlıca faktörlerin yaşam tarzı, sağlık hizmetlerine erişim, fiziksel aktivite düzeyi ve kişisel sağlık geçmişi olduğunu açıkça ortaya koymuştur. Elde edilen bulgular, diyabetin önlenmesi açısından bireylerin yaşam alışkanlıklarının iyileştirilmesinin ve düzenli sağlık takibinin önemini vurgulamaktadır. Sağlıklı beslenme, egzersiz alışkanlığı, stres yönetimi ve düzenli sağlık kontrolleri gibi koruyucu sağlık davranışları, diyabet riskini önemli ölçüde azaltmaktadır.

Bununla birlikte, sağlık hizmetlerinin erişilebilirliği yalnızca bireysel değil, toplumsal düzeyde de koruyucu sağlık stratejilerinin etkinliğini artıran temel bir unsurdur. Sağlık hizmetlerine maddi veya yapısal engel olmaksızın ulaşabilen bireylerin, erken tanı ve müdahale avantajı sayesinde hastalık riskleri azaltılabilmektedir.

Sonuçlar, veri temelli sağlık politikalarının geliştirilmesi açısından güçlü bir temel sunmakta ve daha hedef odaklı önleyici stratejilerin tasarlanmasının mümkün olduğunu göstermektedir. Bu bağlamda, hem toplumsal farkındalık düzeyinin artırılmasına hem de bireysel sağlık yönetimi bilincinin güçlendirilmesine yönelik programların uygulanması, halk sağlığının iyileştirilmesine doğrudan katkı sağlayacaktır.

Heart Attack Dataset için elde edilen kuralların yorumları aşağıdaki gibidir:

**A. Sağlık Koşulları ile Kalp Hastalığı İlişkisi**

Kalp hastalığı bulunmayan bireylerin çoğunlukla çeşitli kronik hastalık risk faktörlerinden arınmış olduğu gözlemlenmiştir. Özellikle:

* Diyabetin Bulunmaması (Diabetic = 0.0): Kalp-damar hastalıklarıyla doğrudan ilişkili olan diyabetin yokluğu, kalp hastalığı riskinin azalmasında önemli rol oynamaktadır.
* Yürüme Güçlüğü Olmaması (DiffWalking = 0.0): Fiziksel hareketliliğin korunmuş olması, hem kalp sağlığı hem de metabolik denge açısından koruyucu bir etkendir.
* Sigara Kullanmama (Smoking = 0.0): Tütün kullanımının olmaması, kalp-damar sisteminin bütünlüğünü koruyan başlıca faktörlerden biridir
* Örnek Kural: {Diabetic = 0.0, DiffWalking = 0.0, KidneyDisease = 0.0, Stroke = 0.0, SkinCancer = 0.0, Smoking = 0.0} → HeartDisease = 0.0 (%76.2 doğruluk)

Bu kural, çoklu risk faktörlerinin ortadan kaldırılmasının kalp hastalığı riskini anlamlı düzeyde azaltabileceğini göstermektedir.

**B. Fiziksel Aktivitenin Koruyucu Rolü**

Düzenli fiziksel aktivitenin, kalp sağlığı üzerinde belirgin olumlu etkileri olduğu elde edilen kurallarla desteklenmiştir.

Örnek Kural: {PhysicalActivity = 1.0, KidneyDisease = 0.0, Smoking = 0.0, Diabetic = 0.0, Stroke = 0.0} → HeartDisease = 0.0 (%76.6 doğruluk)

Bu sonuç, fiziksel aktivite düzeyi yüksek bireylerin, kalp hastalığı açısından daha düşük risk taşıdığını ve dolayısıyla düzenli egzersizin koruyucu bir önlem olarak değerlendirilmesi gerektiğini vurgulamaktadır.

**C. Çoklu Risk Faktörlerinin Olmaması**

Birden fazla risk faktörünün aynı anda bulunmaması, kalp hastalığı riskinin belirgin şekilde düştüğünü ortaya koymaktadır.

Örnek Kural: {Diabetic = 0.0, DiffWalking = 0.0, KidneyDisease = 0.0, Stroke = 0.0, PhysicalActivity = 1.0, Smoking = 0.0} → HeartDisease = 0.0 (%74.9 doğruluk)

Bu kural, bireysel risk faktörlerinden ziyade genel sağlık profilinin bütüncül olarak değerlendirilmesinin daha anlamlı sonuçlar verebileceğini göstermektedir.

**D. Böbrek Sağlığı ile Kalp Hastalığı Arasındaki İlişki**

Veriler, böbrek hastalığı bulunmayan bireylerde kalp hastalığı riskinin daha düşük olduğunu ortaya koymaktadır.

Örnek Kural: {KidneyDisease = 0.0, SkinCancer = 0.0, Smoking = 0.0, Diabetic = 0.0, Stroke = 0.0} → HeartDisease = 0.0 (%77.2 doğruluk)

Bu bulgu, kronik hastalıkların sistemik etkilerini ve organlar arası ilişkilerin kalp hastalığı riskine olan etkisini gözler önüne sermektedir.

Elde edilen kurallar, kalp hastalığı riskinin; fiziksel aktivite eksikliği, sigara kullanımı, diyabet, böbrek hastalığı ve hareket kısıtlılığı gibi birden fazla değişkenin bileşkesiyle şekillendiğini göstermektedir. Bu nedenle, bireysel değil, bütünsel risk değerlendirme yaklaşımlarının benimsenmesi gerektiği ortaya konmaktadır.

Bu doğrultuda geliştirilecek olan farkındalık artırıcı halk sağlığı programları, düzenli sağlık taramaları, yaşam tarzı değişikliklerini destekleyen uygulamalar ve önleyici tıp politikaları, toplum genelinde kalp hastalıklarının önlenmesinde etkili stratejiler sunabilir.

1. SONUÇ

Bu çalışma, sağlık alanında hassas veri işleme gereksinimlerini karşılamak üzere tasarlanmış bir federe öğrenme mimarisi çerçevesinde, özellikle model birleştirme sürecine odaklanarak gerçekleştirilmiştir. Temel amaç, istemcilerde yerel olarak eğitilen modellerin merkezi sunucuda etkili bir şekilde birleştirilmesini sağlayan duCBA algoritmasının sağlık verileri üzerindeki performansını değerlendirmektir. Bu amaçla, iki farklı sağlık veri seti — Diabetes Health Indicators ve Heart Attack Dataset — kullanılarak deneysel çalışmalar yürütülmüş; farklı istemci sayıları, üç ayrı özellik seçimi yöntemi (Mutual Information, Chi-Square ve PCA) ve iki farklı federe toplama algoritması (FedAvg ve duCBA) karşılaştırmalı olarak analiz edilmiştir. Sonuçlar, duCBA algoritmasının çeşitli özellik seçimi teknikleriyle sorunsuz çalıştığını ve özellikle az sayıda istemciyle yürütülen sistemlerde %67 ile %71 arasında değişen doğruluk oranlarıyla dengeli ve güvenilir çıktılar sunduğunu ortaya koymuştur.

duCBA’nın dikkat çekici yönlerinden biri, yalnızca model çıktılarının doğruluğuna odaklanmakla kalmayıp, aynı zamanda elde edilen kuralların açıklanabilirliğini ön planda tutmasıdır. İstemcilerde CBA algoritması ile elde edilen ilişkilendirme kuralları, merkezi sunucuda support ve confidence değerlerine göre yeniden değerlendirilmekte; bu süreç sonunda oluşan nihai model, karar mekanizması izlenebilir ve yorumlanabilir bir yapıya kavuşmaktadır. Bu da, duCBA’nın sadece teknik doğruluğu değil, aynı zamanda şeffaflığı ve kullanıcıya sunduğu açıklama kabiliyetiyle de öne çıkmasını sağlamaktadır. Özellikle sağlık gibi kararların yüksek etkiler doğurabileceği bir alanda, modelin neden ve nasıl belirli çıktılar ürettiğini açıklayabilmek hem etik hem de uygulama açısından kritik bir gerekliliktir. Bu noktada, duCBA algoritması, geleneksel kara kutu modellerin sınırlılıklarını aşarak, XAI (Açıklanabilir Yapay Zekâ) ilkeleri doğrultusunda anlamlı bir çözüm sunmaktadır.

Uygulamalı sonuçlar, duCBA algoritmasının sadece akademik bir çözüm olarak değil, aynı zamanda pratik sağlık uygulamaları için de uygun olduğunu göstermektedir. Özellikle diyabet ve kalp hastalığı gibi yaygın ve kronik rahatsızlıkların tahmininde elde edilen kurallar, bireysel düzeyde önleyici stratejilerin geliştirilmesine olanak tanımaktadır. Örneğin, fiziksel aktivite, sigara kullanımı, diyabet geçmişi, böbrek hastalıkları ve sağlık hizmetlerine erişim gibi çoklu değişkenlerin birlikte ele alınmasıyla oluşturulan kurallar, klinik karar destek sistemlerinde kullanılabilecek anlamlı içgörüler sunmaktadır. Bu bağlamda, duCBA algoritması yalnızca model başarımı değil, aynı zamanda klinik geçerlilik açısından da güçlü bir adaydır.

Bununla birlikte, bu çalışma mevcut veri setleri ve sınırlı istemci sayıları ile yürütülmüş olup, gelecekte daha büyük ve çeşitli veri kümeleri üzerinde yapılacak çalışmalarla duCBA’nın ölçeklenebilirliği, genel performansı ve farklı veri tipleriyle (örneğin görüntü, zaman serisi, metin) uyumu detaylı şekilde test edilebilir. Ayrıca, duCBA’nın diferansiyel gizlilik veya güvenli çok taraflı hesaplama gibi ileri düzey gizlilik teknikleri ile entegre edilmesi, hem güvenlik düzeyini artıracak hem de uygulama alanlarını genişletecektir.

Sonuç olarak, bu çalışma ile duCBA algoritması, sağlık verileri üzerinde hem veri gizliliğini koruyan hem de açıklanabilir kararlar sunan bir model birleştirme çözümü olarak öne çıkmıştır. Modelin karar mantığını kullanıcıya sunabilme özelliği, sadece teknik bir kazanım değil, aynı zamanda kullanıcı güvenini artıran ve yasal-etik standartları destekleyen bir avantaj olarak değerlendirilebilir. duCBA, federe öğrenme alanına yalnızca bir algoritma değil, açıklanabilir ve güvenilir bir yaklaşım olarak katkı sağlamaktadır.

**KAYNAKLAR**

[1] Sasaki, Y., A survey on IoT big data analytic systems: Current and future. IEEE Internet of Things Journal, 9(2), 1024–1036, 2022.

[2] John Dian, F., Vahidnia, R., & Rahmati, A., Wearables and the Internet of Things (IoT), applications opportunities and challenges: A survey, IEEE Access, 8, 69200–69211, 2020.

[3] Bian, J., Arafat, A. A., Xiong, H., Li, J., Li, L., Chen, H., Wang, J., Dou, D., & Guo, Z., Machine learning in real-time Internet of Things (IoT) systems: A survey. IEEE Internet of Things Journal, 9(11), 8364–8386, 2022.

[4] Telikani, A., Tahmassebi, A., Banzhaf, W., & Gandomi, A. H., Evolutionary machine learning: A survey. ACM Computing Surveys, 54(8), Article 161, 2021.

[5] Paleyes, A., Urma, R.-G., & Lawrence, N. D., Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), Article 114, 1–29, 2022

[6] Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., & Das, A., Differential privacy-enabled federated learning for sensitive health data, 2019

[7] Zhou, S., & Li, G. Y., FedGiA: An efficient hybrid algorithm for federated learning, IEEE Transactions on Signal Processing, 71, 1493-1508, 2023.

[8] Sun, Y., Li, H., Shen, Y., Xie, J., Zhao, Y., Gao, X., & Si, N., An Asynchronous Federated Learning Algorithm Based on a Backup Update of Model Version Parameters, 2023 3rd International Conference on Electrical Engineering and Control Science (IC2ECS), 1538-1544, 2023.

[9] Raj, A., Sharma, V., & Shanu, A. K., Comparative analysis of security and privacy techniques for federated learning in IoT- based devices, In 2022 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 1-5, 2022.

[10] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V., Federated Learning: Challenges Methods and Future Directions, IEEE Signal Processing Magazine, 37(3), 50-60, 1, May, 2020.

[11] Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., & Zhang, W., A survey on federated learning: Challenges and applications, International Journal of Machine Learning and Cybernetics, 14(2), 513-535, 2023.

[12] Ali, M. K., Pearson-Stuttard, J., Selvin, E., & Gregg, E. W., Interpreting global trends in type 2 diabetes complications and mortality. Diabetologia, 65(1), 3–13, 2022.

[13] American Heart Association, AHA names biggest advances in cardiovascular research for 2024, [Online].Available: https://www.heart.org/en/around-the-aha/aha-names-biggest-advances-in cardiovascular-research-for-2024

[14] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V., Federated Learning: Challenges Methods and Future Directions, IEEE Signal Processing Magazine, 37(3), 50-60, 1, May, 2020.

[15] Wen, J., Zhang, Z., Lan, Y., Cui, Z., Cai, J., & Zhang, W., A survey on federated learning: Challenges and applications, International Journal of Machine Learning and Cybernetics, 14(2), 513-535, 2023.

[16] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y., A survey on federated learning, Knowledge-Based Systems, 216, 106775, 2021.

[17] Kaur, H., Rani, V., Kumar, M., Sachdeva, M., Mittal, A., & Kumar, K., Federated learning: A comprehensive review of recent advances and applications, Multimedia Tools and Applications, 83(18), 54165-54188 , 2024.

[18] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas,B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).

[19] Zhou, S., & Li, G. Y., FedGiA: An efficient hybrid algorithm for federated learning, IEEE Transactions on Signal Processing, 71, 1493-1508, 2023.Sun, Y., Li, H., Shen, Y., Xie, J., Zhao, Y., Gao, X., & Si,

[20] N., An Asynchronous Federated Learning Algorithm Based on a Backup Update of Model Version Parameters, 2023 3rd International Conference on Electrical Engineering and Control Science (IC2ECS), 1538-1544, 2023.

[21] Ezzeldin, Y. H., Yan, S., He, C., Ferrara, E., & Avestimehr, A. S., Fairfed: Enabling group fairness in federated learning, In Proceedings of the AAAI conference on artificial intelligence, 37(6), 7494-7502, June, 2023.

[22] Lycklama, H., Burkhalter, L., Viand, A., Küchler, N., & Hithnawi, A., RoFL: Robustness of Secure Federated Learning, 2023 IEEE Symposium on Security and Privacy (SP), 453-476, 2023.

[23] Büyüktanır, B., Yıldız, K., Ülkü, E. E., & Bütüktanir,T., Du-Cba: Data-Agnostic And İncremental Classification-Based Association Rules Extraction Architecture Du-Cba: Veriden Habersiz Ve Artirimli Siniflandirmaya Dayali Birliktelik Kurallari Çikarma Mimarisi. Journal Of The Faculty Of Engineering And Architecture Of Gazi University , Vol.38, No.3, 1919-1929, 2023

[24] Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., & Das, A., Differential privacy-enabled federated learning for sensitive health data, arXiv preprint arXiv:1910.02578, 2019.

[25] Horvath, A. N., Berchier, M., Nooralahzadeh, F., Allam, A., & Krauthammer, M., Exploratory Analysis of Federated Learning Methods with Differential Privacy on MIMIC-III, arXiv preprint arXiv:2302.04208, 2023.

[26] Durga, R., & Poovammal, E., Federated learning model for healthchain system, In 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 1-6, 2021.

[27] Chen, S., Comparison of machine learning algorithms and feature visualization analysis for diabetes risk prediction,Journal of Physics: Conference Series, 2646(1), 012013, 2023.

[28] Hama Saeed, M. A., Diabetes type 2 classification using machine learning algorithms with up-sampling technique, Journal of Electrical Systems and Information Technology, 10(1), 8, 2023.

[29] Kumar, A., & Choudhury, T.,Heart Disease Prediction Using Machine Learning, 2022 2nd International Conference on Sustainable Technologies for Computational Intelligence (ICTSCI), 1-6, 2022.

[30] Gupta, S., Kumar, P., Srivastava, N. V., Kumar, A., & Chaurasia, B. K., Heart Disease Prediction Using Federated Learning, In S. C. Satapathy, V. Bhateja, & S. Das (Eds.), Proceedings of International Conference on Recent Innovations in Computing ICRIC 2023, pp. 35-47, 2024.

[31] Centers for Disease Control and Prevention (CDC), “Diabetes Health Indicators Dataset”, 2024, Kaggle. [Online].Available:http://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data

[32] O. Fayez, “Heart Attack Data,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/omarfayez/heart-attack-data