logo, yazı tipi, simge, sembol, grafik içeren bir resim

Açıklama otomatik olarak oluşturuldu

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

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

**BİTİRME PROJESİ 2. ARA RAPORU**

Bilgisayar Mühendisliği Bölümü

**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, xxx tarihinde savunulmuş ve jüri üyeleri tarafından başarılı bulunmuştur.

**Jüri Üyeleri**

|  |  |  |
| --- | --- | --- |
| Dr. Öğr. Üyesi xxx xxx  Marmara Üniversitesi | **(Danışman)** | (İMZA)………….. |
| Prof. Dr. Xxx xxx  Marmara Üniversitesi | (Üye) | (İMZA)………….. |
| Prof. Dr. Xxx xxx  Marmara Üniversitesi | (Üye) | (İMZA)………….. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**ÖNSÖZ**

Proje çalışmamız süresince karşılaştığım bütün problemlerde, sabırla yardım ve bilgilerini esirgemeyen, tüm desteğini sonuna kadar yanımda hissettiğim değerli hocalarım, sayın Dr. Öğr. Üyesi Xxx xxx ve sayın Prof. Dr. Xxx xxx’ a en içten teşekkürlerimi sunarım.

Bu proje çalışması fikrinin oluşması ve ortaya çıkmasındaki önerisi ve desteğinden dolayı değerli hocam Dr. Öğr. Üyesi Xxx xxx’ a teşekkür ederim.

Proje çalışmam sırasında maddi ve manevi desteklerini esirgemeyen okul içerisinde ve okul dışında her zaman yanımda olan değerli çalışma arkadaşlarım ve hocalarım Doç. Dr. Xxx xxx ve Dr. Öğr. Üyesi ’ xxx xxx a sonsuz teşekkürlerimi sunarım.

**İÇİNDEKİLER**

[1. GİRİŞ 1](#_Toc193719797)

[2. LİTERATÜR TARAMASI 4](#_Toc193719798)

[3. KULLANILAN YÖNTEMLER 7](#_Toc193719799)

[3.1 Veri Setleri 7](#_Toc193719800)

[3.1.1 Diabetes Health Indicators Dataset 8](#_Toc193719801)

[3.1.2 Heart Attack Dataset 12](#_Toc193719802)

[3.2 Prototip Uygulama 16](#_Toc193719803)

[3.2.1 İstemciler 17](#_Toc193719804)

[3.2.2 API 17](#_Toc193719805)

[3.2.2 Ana Sunucu 17](#_Toc193719806)

[4. DENEYSEL SONUÇLAR 18](#_Toc193719807)

**ÖZET**

**FEDERE ÖĞRENME İLE SAĞLIK VERİLERİNİ KULLANARAK MODEL EĞİTİMİ VE GERÇEK DÜNYA ANALİZİ**

Makine öğrenimi, büyük veri kümelerinden anlamlı bilgiler çıkarma ve karmaşık sorunları çözme yeteneği nedeniyle sağlık sektöründe büyük önem kazanmıştır. Ancak sağlık verilerinin mahremiyeti ve gizliliği, bu verilerin merkezi sunucularda toplanmasını zorlaştırmakta ve geleneksel makine öğrenimi yöntemlerinin uygulanabilirliğini sınırlamaktadır. Bu araştırmada, sağlık verileri üzerinde hastalık tahmini için federe öğrenme mimarisi kapsamında geliştirilen Veriden Habersiz İlişkili Kurallara Dayalı Sınıflandırma (duCBA) adlı bir federe birleştirme algoritması kullanılmıştır. Model eğitimi, Kaggle platformundan elde edilen Diabetes Health Indicators ve Heart Attack Dataset veri kümeleri kullanılarak gerçekleştirilmiştir. Bu veri kümelerine Mutual Information, Chi-Square ve Principal Component Analysis (PCA) gibi farklı öznitelik seçme yöntemleri uygulanarak üç farklı ön-işlenmiş veri kümesi oluşturulmuştur. Bu veri kümeleri ile oluşturulan sistem üzerinde deneyler gerçekleştirilmiştir. Sistem merkezi bir sunucu, bir API sunucusu ve istemcilerden oluşan bir yapı ile tasarlanmıştır. İstemciler, verilerini paylaşmadan kendi cihazlarında yerel olarak modelleri eğitirken, API sunucusu bu modelleri toplayarak merkezi sunucuya iletmiştir. Merkezi sunucu daha sonra duCBA algoritmasını kullanarak yerel modelleri bir araya getirmiş ve güncellenmiş global modeli istemcilere geri dağıtmıştır. Bu makalede duCBA algoritmasının performansı analiz edilmiştir. Araştırma, duCBA'nın sağlık sektöründeki yapay zeka uygulamalarını geliştirme potansiyelini vurgulamaktadır. Bu nedenle duCBA algoritması, Federated Averaging (FedAvg) adı verilen bir federasyon toplama algoritması ile karşılaştırılmıştır. Sonuçlar duCBA'nın FedAvg ile benzer performans gösterdiğini ortaya koymaktadır. Ayrıca duCBA algoritması, Explainable Artificial Intelligence (XAI) ilkelerine uygun olarak, modelin karar verme sürecini şeffaf bir şekilde takip etmeyi sağlar. Farklı sayıda istemci ve özellik seçim yöntemleri ile yapılan testlerde duCBA algoritmasının özellikle az sayıda istemci içeren sistemlerde kararlı ve dengeli sonuçlar verdiği ve %67-71 doğruluk oranı ile tutarlı bir performans sergilediği görülmüştür.

**Abdülbaki DEMİR**

**Ahmet Yasir KULAKSIZ**

**Mart, 2025**

**Abdülbaki DEMİR**

**Ahmet Yasir KULAKSIZ**

**ABSTRACT**

**MODEL TRAİNİNG AND REAL WORLD ANALYSİS USİNG HEALTH DATA WİTH FEDERATED LEARNİNG**

Machine learning has gained great importance in the healthcare industry due to its ability to extract meaningful information from large datasets and solve complex problems. However, the privacy and confidentiality of health data makes it difficult to collect this data on centralized servers and limits the applicability of traditional machine learning methods. In this research, a federated aggregation algorithm called Data Unaware Classification Based on Association (duCBA), developed within the scope of federated learning architecture, is used for disease prediction on health data. Model training was performed using Diabetes Health Indicators and Heart Attack datasets obtained from the Kaggle platform. Three different pre-processed datasets were created by applying different feature selection methods such as Mutual Information, Chi-Square and Principal Component Analysis (PCA) to these datasets. Experiments were conducted on the system created with these datasets. The system is designed with structure consisting of a central server, an API server and clients. Clients trained models locally on their own devices without sharing their data, while the API server aggregated these models and transmitted them to the central server. The central server then aggregated the local models using the duCBA algorithm and distributed the updated global model back to the clients. In this paper, the performance of the duCBA algorithm is analysed. The research highlights the potential of duCBA to enhance artificial intelligence applications in the healthcare sector. Therefore, the duCBA algorithm is compared with a federated aggregation algorithm called Federated Averaging (FedAvg). The results show that duCBA performs similarly to FedAvg. Furthermore, the duCBA algorithm allows to follow the decision-making process of the model in a transparent way, in accordance with the principles of Explainable Artificial Intelligence (XAI). In tests with different numbers of clients and feature selection methods, it has been observed that the duCBA algorithm provides stable and balanced results, especially in systems with a small number of clients, and shows a consistent performance with an accuracy rate of 67-71%.

**Abdülbaki DEMİR**

**Ahmet Yasir KULAKSIZ**

**March, 2025**

**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**

[Şekil 1: Federe Öğrenme Akışı Diyagramı 2](#_Toc193719820)

[Şekil 2: Bağımsız Değişkenlerin Diyabet\_binary ile Korelasyonu 8](#_Toc193719821)

[Şekil 3: Diabetes\_binary için Mutual Informatıon Skorları 9](#_Toc193719822)

[Şekil 4: Diabetes\_binary için Chi-Square Skorları 10](#_Toc193719823)

[Şekil 5: Diabetes\_binary için PCA Katkısı 10](#_Toc193719824)

[Şekil 6: Bağımsız Değişkenlerin HeartDisease ile Korelasyonu 12](#_Toc193719825)

[Şekil 7: Bağımsız Değişkenlerin HeartDisease ile Korelasyonu 13](#_Toc193719826)

[Şekil 8: HeartDisease için Chi-Square Skorları 14](#_Toc193719827)

[Şekil 9: HeartDisease için PCA Katkısı 14](#_Toc193719828)

[Şekil 10: duCBA Akışı Diyagramı 16](#_Toc193719829)

**TABLO LİSTESİ**

[Tablo 1: Dıabetes Health Indıcators Dataset için Özellı̇k Seçı̇mı̇ 11](#_Toc193719838)

[Tablo 2: Heart Attack Dataset için Özellı̇k Seçı̇mı̇ 15](#_Toc193719839)

[Tablo 3: Dıabetes Health Indıcators Dataset için Sınıflandırma Raporu 19](#_Toc193719840)

[Tablo 4: Heart Attack Dataset için Sınıflandırma Raporu 20](#_Toc193719841)

# GİRİŞ

Veri odaklı teknolojilerin hızla gelişmesiyle birlikte makine öğrenmesi modelleri birçok alanda önemli bir rol oynamaktadır [1]. Bu veriler kullanıcı hareketlerinin analiz edilmesinden çevresel faktörlerin izlenmesine kadar birçok farklı alanda değerlendirilmektedir [2]. Ayrıca makine öğrenmesi algoritmaları ile oluşturulan öngörüsel modeller sayesinde çeşitli sistemler geliştirilmektedir [3]. Bu modeller büyük miktarda veriyi analiz ederek karmaşık örüntüleri ve ilişkileri ortaya çıkarabilir, gelecekteki sonuçları tahmin edebilir ve karar destek sistemleri için kritik bilgiler sağlayabilir. Ancak makine öğrenmesi modellerinin etkin bir şekilde çalışabilmesi için büyük ve çeşitli veri kümelerine ihtiyaç duyulmaktadır. Bu durum, veri boyutunun artmasıyla birlikte donanım kapasitesinin zorlanması, eğitim süresinin uzaması ve veri güvenliği gibi sorunları da beraberinde getirmektedir [5].

Özellikle sağlık sektörü gibi hassas verilerin bulunduğu alanlarda veri gizliliği ve güvenliği kritik öneme sahiptir [6-7]. Sağlık verilerinin merkezi bir sunucuda toplanması hem yasal düzenlemeler hem de güvenlik riskleri nedeniyle sınırlamalarla karşılaşmaktadır [8]. Örneğin, birçok ülkedeki kişisel verilerin korunması düzenlemeleri, sağlık verilerinin paylaşılmasını ve merkezi olarak işlenmesini kısıtlamaktadır. Bu durum geleneksel makine öğrenmesi yöntemlerinin uygulanabilirliğini sınırlandırmakta ve yeni yaklaşımların geliştirilmesini zorunlu kılmaktadır.

Bu bağlamda, federe öğrenme, merkezi veri toplama ihtiyacını ortadan kaldıran ve verilerin yerel cihazlarda işlenmesini sağlayan yenilikçi bir çözüm olarak öne çıkıyor. Federe öğrenme, kullanıcı verilerinin gizliliğini korurken makine öğrenimi modellerinin dağıtık bir şekilde eğitilmesine olanak tanır [9]. Bu yöntem, özellikle sağlık sektörü gibi veri gizliliğinin ön planda olduğu alanlarda güçlü bir alternatif sunmaktadır. Federe öğrenme ile hasta verileri yerel cihazlarda işlenir ve sadece model güncellemeleri merkezi sunucuya iletilir. Böylece iletişim ve depolama maliyetleri azalır [10]. Ayrıca federe öğrenme yöntemi merkezi yapıya veri göndermediği için yerel cihazlardaki kritik bilgilerin güvenliği de sağlanmış olur [11].

Federe öğrenmenin bir diğer önemli avantajı ise veri çeşitliliğini artırarak daha genelleştirilebilir modellere olanak sağlamasıdır. Geleneksel yöntemlerde farklı kaynaklardan gelen verilerin bir araya getirilmesi yasal ve teknik zorluklar yaratırken, federe öğrenme bu verilerin yerel olarak işlenmesini sağlayarak bu engellerin üstesinden gelir. Ayrıca, veri sahiplerinin kontrolü elinde tutmasını sağlayarak güveni artırır ve daha geniş katılımı teşvik eder.

metin, diyagram, ekran görüntüsü, plan içeren bir resim

Açıklama otomatik olarak oluşturuldu

Şekil : Federe Öğrenme Akışı Diyagramı

Şekil 1’de federe öğrenme mimarisinin akış diyagramını göstermektedir. Akış aşağıdaki gibi çalışır:

1. Her istemci kendi verileriyle yerel bir modeli eğitir.
2. Eğitilen yerel modeller merkezi bir sunucuya gönderilir. Ancak veriler değil, sadece modellerin güncellenmiş parametreleri paylaşılır.
3. Merkezi sunucu, tüm istemcilerden gelen yerel modelleri toplayarak global bir model oluşturur. Bu süreç, istemcilerden gelen bilgilerin küresel bir modelde toplanmasını sağlar.
4. Sunucu, oluşturulan global modeli tüm istemcilere yeniden gönderir. İstemciler bu modeli alır ve bir sonraki turda kendi yerel verileriyle yeniden eğitir.

Federe öğrenme mimarileri eğitimden sağlığa, güvenlikten finansa kadar birçok alanda kullanılmaktadır. Özellikle hasta mahremiyetinin korunması açısından sağlık alanındaki araştırmalar önemli ve güncel olmalıdır. Bu araştırmada, federe öğrenme mimarisi kullanılarak sağlık verilerinde hastalık tahmini üzerine bir uygulama geliştirilmiştir. Diyabet dünya çapında giderek artan bir sağlık sorunudur ve erken teşhisi büyük önem taşımaktadır [12]. Ayrıca American Heart Association (AHA) Başkanlık Danışmanlığı, kardiyovasküler hastalıkların 2050 yılına kadar %11,3'ten %15'e yükseleceği ve 45 milyon ABD'li yetişkini etkileyeceği konusunda uyarmaktadır [13]. Bununla birlikte, tıbbi verilerin gizliliği ve güvenliği, geleneksel makine öğrenimi yaklaşımlarının uygulanmasının önünde büyük bir engeldir. Federe öğrenme ile hasta verileri merkezi bir sunucuya aktarılmadan yerel cihazlarda işlenir ve böylece hasta mahremiyeti korunur.

Bu çalışmanın temel katkıları aşağıdaki gibi özetlenebilir:

* Federe Öğrenme ile Yenilikçi Entegrasyon Önerisi: Bu çalışmada, federe birleştirme algoritmalarından biri olan duCBA algoritması kullanılmış ve bir sağlık hizmeti veri kümesine uygulanmıştır. Gizliliği koruyan hastalık tahminine odaklanan yaklaşım, model verimliliği ve güvenliği sağlarken dağıtılmış sağlık verileri analizindeki temel zorlukları ele almaktadır.
* Sağlık Hizmetleri Alanında Uygulama ve Doğrulama: Araştırma, önerilen çerçevenin uygulanabilirliğini iki kritik sağlık veri kümesi üzerinde göstermektedir: Diabetes Health Indicators Dataset ve Heart Attack Dataset. Her iki veri kümesinde de elde edilen tutarlı sonuçlar, duCBA algoritmasının risk analizi ve erken hastalık tespiti için uyarlanabilirliğinin ve sağlamlığının altını çizmektedir.
* Gelişmiş Özellik Seçimi ile Tahmine Dayalı Performansı Artırma: Gelişmiş özellik seçimi yöntemlerinden yararlanan çalışma, üstün tahmin doğruluğu elde etmek için model girdilerini iyileştiriyor. Bu, sağlık uygulamalarında kritik bir faktör olan veri gizliliğini korurken hastalık tahminlerinin güvenilirliğini sağlar.
* Ölçeklenebilir ve Gizliliği Koruyan Bir Çözüm Sağlama: Veri güvenliği ve federe model ölçeklenebilirliği gibi zorlukları ele alan bu çalışma, sağlık hizmetleri gibi hassas alanlarda federe öğrenmenin uygulanmasına yönelik pratik bir çözüm sunarak daha fazla gerçek dünya uygulamasının önünü açıyor.

Araştırmanın sonraki bölümleri aşağıdaki konuları incelemektedir. Bölüm 2, harmanlanmış öğrenme için yapılan çalışmaları gözden geçirmektedir. Bölüm 3, geliştirilen uygulama tekniklerini detaylandırmaktadır. Bölüm 4, uygulamadan elde edilen deneysel sonuçları sunmaktadır.

# LİTERATÜR TARAMASI

Bu bölümde, federe öğrenme mimarisi ve sağlık alanındaki uygulamaları üzerine yapılan araştırmalar incelenmektedir.

Google tarafından 2016 yılında veri gizliliğini korumak amacıyla geliştirilen federe öğrenme mimarisi çeşitli avantajlar sunmaktadır. Bu mimari sayesinde model eğitimi sırasında veri gizliliği sağlanırken, eğitim süreçleri de hızlandırılıyor. Veri yerine modelin istemciler ve sunucu arasında gidip gelmesi bant genişliği ve iletişim gecikmesi gibi sorunları ortadan kaldırıyor, ayrıca merkezi sunucuda veri depolama ihtiyacını azaltarak maliyetleri düşürüyor. Ancak, federe öğrenmenin bu avantajlarının yanı sıra, dört ana başlık altında ele alınan önemli sorunlar da vardır:

* İletişim maliyeti
* Sistem heterojenliği
* İstatistiksel heterojenlik
* Gizlilik endişeleri.

Yerel cihaz sayısı arttıkça model gönderme sürecini yönetmek giderek zorlaşır ve bu durum iletişim maliyeti olarak adlandırılır [14]. Sistem heterojenliği, model eğitimi için kullanılan cihazların donanım ve yazılımlarındaki farklılıklardan kaynaklanan performans değişikliklerini ifade eder [15]. İstatistiksel heterojenlik, cihazlar tarafından toplanan veri dağılımlarındaki farklılıklardan kaynaklanır ve bu da modelin bazı cihazlarda iyi performans gösterirken diğerlerinde kötü performans göstermesine neden olabilir [16]. Federe öğrenme yaklaşımında hassas bilgilerin gizliliği, herhangi bir verinin eğitim için merkezi bir sunucuya aktarılmaması ve bunun yerine yalnızca yerel olarak eğitilmiş modellerin paylaşılmasıyla korunur. Ancak, paylaşılan modellerin parametreleri aracılığıyla kritik bilgilere erişme potansiyeli konusunda endişeler devam etmektedir. Bu gizlilik endişelerini gidermek için çeşitli algoritmalar geliştirilmiştir [15-17].

Literatürde federe öğrenme mimarisi için bazı yöntemler geliştirilmiştir.

* **FedAvg:** Federe öğrenme alanında öne çıkan bir yöntemdir ve 2017 yılında Brendan McMahan ve arkadaşları tarafından önerilmiştir. FedAvg, istemciler üzerinde yerel olarak eğitilen modellerin ağırlıklarının merkezi bir sunucu üzerinde ortalamasının alınmasına dayanır. Bu yaklaşım, merkezi veri toplama ihtiyacını ortadan kaldırır ve veri gizliliğini sağlarken iletişim maliyetlerini önemli ölçüde azaltır [18].
* **FedGiA:** İletişim maliyeti ve istatistiksel heterojenlik sorunlarını hibrit bir yaklaşımla çözmeyi amaçlamaktadır [19].
* **FedAvu:** Katılımcıların kendi hızlarında çalışmalarını sağlayarak sistem heterojenliği sorununu ele alır [20].
* **FairFed:** Farklı gruplar arasında dengeli model performansı sağlamaya odaklanarak istatistiksel heterojenlik sorununu ele alır [21].
* **RoFL:** Güvenli gradyan toplama ve şifreleme teknikleri aracılığıyla veri gizliliğini ve güvenliğini sağlamaya çalışır [22].

Federe öğrenme mimarisi için bu çalışmada duCBA yöntemi kullanılacaktır. duCBA, CBA algoritması ile eğitilen modellerin birleştirilmesi için kullanılan bir yöntemdir [23]. Bu süreçte her bir kuralın destek ve güven değerleri yeniden hesaplanır ve kurallar sıralanır. Farklı modellerden gelen aynı etiketlemeye sahip kurallar için her bir kuralın veri kümesinde bulunma oranı kontrol edilerek destek ve güven değerleri buna göre güncellenir. Aynı içeriğe sahip ancak farklı etiketlere sahip kurallar için destek değeri kontrol edilir ve destek değeri yüksek olan kural listeye dahil edilir. Bu işlem sonunda oluşan kural listesi nihai modelin temelini oluşturur. duCBA algoritması ilişkisel sınıflandırma yaklaşımı temel alınarak geliştirilmiş olup verilerin yerel cihazlarda eğitilmesi ve merkezi bir sunucuda toplanması ihtiyacını ortadan kaldırır. Bu şekilde, veri gizliliği korunurken ağ trafiği ve veri aktarımı en aza indirilir.

Federe öğrenme, tıbbi verilerin işlenmesinde gizlilik ve güvenlik standartlarını yükselten yenilikçi bir yaklaşımdır. Bu yöntem, verileri yerel cihazlarda işleyerek hasta bilgilerinin merkezi sunuculara taşınması ihtiyacını ortadan kaldırır. Böylece bireysel veri gizliliği korunurken, hasta bilgilerinin olası güvenlik açıklarından etkilenme riski de en aza indirilmiş olur [6-8].

Sağlık hizmetlerinde federe öğrenmenin etkinliğini gösteren birçok çalışma, bu yaklaşımın potansiyel faydalarını vurgulamaktadır. Bu çalışmalardan bazıları aşağıda özetlenmiştir:

* **Differential privacy-enabled federated learning for sensitive health data (2019):**Choudhury ve arkadaşları, diferansiyel gizlilikle desteklenen federe öğrenme modellerinin hassas sağlık verilerinin yerel olarak işlenmesini sağladığını, böylece merkezi sistemlerde veri toplama ihtiyacını ortadan kaldırdığını ve hasta verilerinin sızma riskini önemli ölçüde azalttığını belirtmiştir [24].
* **Exploratory Analysis of Federated Learning Methods with Differential Privacy on MIMIC-III (2023):** Horvath ve arkadaşları, MIMIC-III veri kümesi üzerinde diferansiyel gizlilik ve birleştirilmiş öğrenmenin etkinliğini incelemiştir. Bu araştırmada hasta verilerinin gizliliği korunarak başarılı modeller geliştirilebileceği ve dağıtık öğrenme süreçleri ile tıbbi verilerin güvenli bir şekilde paylaşılabileceği belirtilmiştir [25].
* **Federated learning model for healthchain system (2021):** Durga ve arkadaşları, federe öğrenme modelinin blok zinciri tabanlı güvenlik önlemleriyle entegre edildiği Healthchain sistemini ele almış ve bu sistemin merkezi bir aracıya ihtiyaç duymadan güvenli veri işleme ve paylaşımı sağladığını vurgulamıştır [26].
* **Comparison of machine learning algorithms and feature visualisation analysis for diabetes risk prediction (2023):** Diyabet tahmininde makine öğrenimi algoritmalarının doğruluk performansı karşılaştırılmış ve bazı algoritmaların risk tahmininde daha başarılı olduğu ve özellik görselleştirme analizinin de bu başarıyı desteklediği görülmüştür [27].
* **Diabetes type 2 classification using machine learning algorithms with up-sampling technique (2023):** Diyabet sınıflandırmasında dengesiz veri kümelerini dengelemek için yukarı örnekleme tekniğinin kullanılmasının makine öğrenmesi algoritmalarının doğruluğunu artırdığı gösterilmiştir [28].
* **Heart Disease Prediction Using Machine Learning (2022):** Kalp hastalığı tahmininde Random Forest, Logistic Regression ve Support Vector Machines gibi algoritmalarla yüksek doğruluk oranları elde etmiş ve bu modellerin yaş, kan basıncı ve kolesterol seviyeleri gibi risk faktörlerinin analizinde etkili olduğunu belirtmiştir [29].
* **Heart Disease Prediction Using Machine Learning (2022):** UCI Machine Learning Repository'deki veri kümelerini kullanarak federatif bir öğrenme ortamında kalp hastalığını tahmin etmiş ve kardiyovasküler risk faktörlerini tahmin etmede yüksek doğruluk bildirmiştir [30].

Bu çalışmalar, federe öğrenmenin sağlık alanında önemli bir uygulama alanına sahip olduğunu ve bu alanda aktif bir araştırma sürecinin devam ettiğini göstermektedir. Özellikle veri gizliliği, güvenlik ve sistem verimliliği gibi kritik sorunlara çözüm arayan federe öğrenme, tıbbi veriler üzerinde güçlü ve etkili makine öğrenmesi modellerinin eğitilmesine olanak sağlamaktadır. Bu bağlamda mevcut literatür, federe öğrenmenin tıbbi veriler üzerinde gizliliği ve verimliliği koruma potansiyelini ortaya koyuyor.

Literatürde, federe öğrenme yöntemleri genellikle merkezi olmayan veri analizi gerçekleştirirken veri gizliliği ve sistem verimliliği gibi kritik sorunları ele almıştır. Ancak, mevcut yöntemler çoğunlukla model güncellemelerindeki heterojenlik, iletişim maliyet ve istatistiksel heterojenlik sorunlarını ele alırken, ilişkisel sınıflandırmaya dayalı bir federe öğrenme mimarisine odaklanmamıştır. duCBA birleştirme algoritması, özellikle CBA algoritmasını federe öğrenme ortamına entegre ederek ilişkisel sınıflandırmaya odaklanan bir model sunmaktadır. Bu algoritma, farklı istemcilerden gelen kuralların destek ve güven değerlerini güncelleyerek daha optimize ve birleşik bir model oluşturmayı amaçlamaktadır. Oluşturulan kurallar hem yorumlanabilen hem de analiz edilebilen bir yapıya sahiptir. Bu sayede sadece mevcut durumları değerlendirmek değil, gelecekteki eğilimleri ve riskleri tahmin etmek de mümkün olmaktadır. Literatürdeki boşluk, ilişkisel sınıflandırma kurallarının federe öğrenme mimarisine uygun olarak güncellenmesi ve gizlilik korunarak paylaşılmasıdır. duCBA bu boşluğu doldurarak hem gizliliği ön planda tutan hem de veri transferini en aza indiren yenilikçi bir çözüm sunmaktadır. Bu yönüyle diğer yöntemlerden farklı olarak hem model doğruluğunu hem de sistem verimliliğini artırmayı hedeflemektedir.

# KULLANILAN YÖNTEMLER

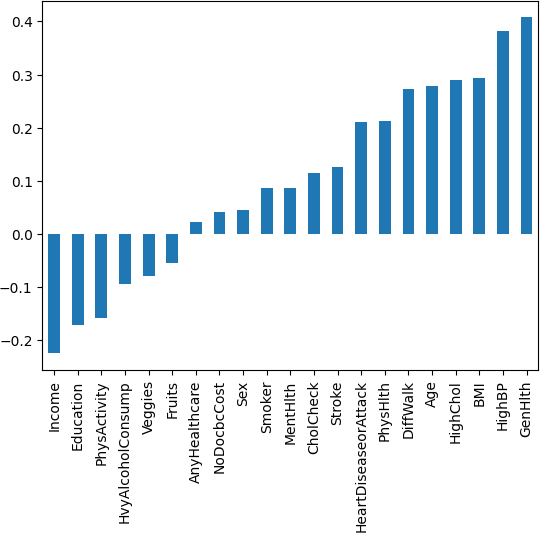
Araştırmanın bu bölümünde model eğitimi için kullanılan veri seti, geliştirilen uygulama için kullanılan teknolojiler ve yöntemler detaylı olarak anlatılmaktadır.

## 3.1 Veri Setleri

Bu araştırmada modellerin eğitilmesi ve değerlendirilmesi için Kaggle platformunda bulunan ‘Diabetes Health Indicators Dataset’ [31] ve ‘Heart Attack Dataset’ [32] veri setleri kullanılmıştır. Bu veri setleri üzerinde literatürde yaygın ve etkili olan üç farklı özellik seçim algoritması PCA, Chi-Square ve Mutual Information algoritmaları uygulanmıştır. PCA, boyut azaltmaya ve özelliklerin varyansını analiz etmeye odaklanmaktadır. Chi-Square, istatistiksel bağımsızlık testi temelinde önemli özellikleri seçer. Mutual Information, bilgi teorisine dayalı bir yaklaşım sağlayarak hedef değişken ile özellikler arasındaki ilişkinin gücünü değerlendirir. Kullanılan algoritmalar, özellik seçim tekniklerinin farklı perspektiflerini temsil ettikleri için seçilmiştir. Bu çeşitlilik, farklı yöntemlerin model üzerindeki etkisinin dengeli bir şekilde değerlendirilmesini sağlamıştır.

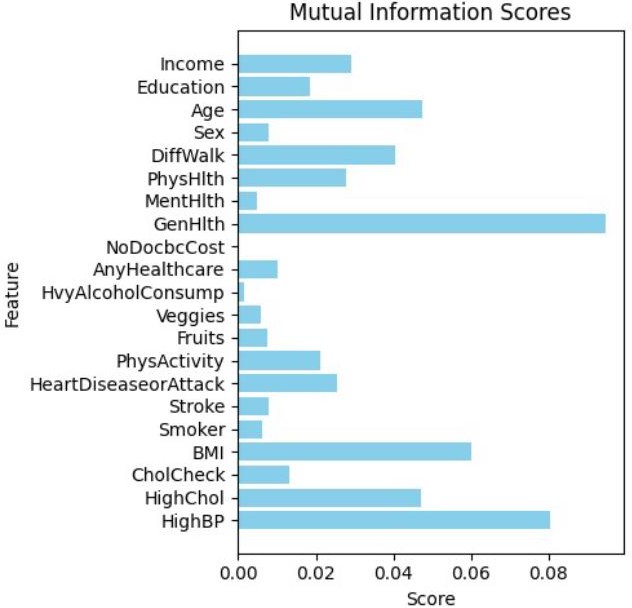
3.1.1 Diabetes Health Indicators Dataset

The Diabetes Health Indicators Dataset, US Centres for Disease Control and Prevention (CDC) tarafından yürütülen Behavioural Risk Factor Surveillance System (BRFSS) anketinden elde edilen verilere dayanmaktadır. Bu veri setinin diyabet ve bireylerin genel sağlığını etkileyebilecek diğer faktörler hakkında bilgiler içermektedir. Veri seti her iki sınıf içinde dengelenmiştir ve 70.692 satır ve 22 sütundan oluşmaktadır. Bağımlı değişken olan ‘Diabetes\_binary’ sütunu katılımcının diyabet durumu hakkında ikili bilgi sağlamaktadır. Diğer 21 bağımsız değişken bireylerin yaşam tarzlarını, sağlık davranışlarını ve genel sağlık durumlarını temsil eden çeşitli faktörleri kapsamaktadır. Bu değişkenler arasında yüksek tansiyon, kolesterol seviyesi, vücut kitle endeksi (BMI), fiziksel aktivite seviyesi, sigara kullanımı, kalp hastalığı geçmişi, genel sağlık durumu ve gelir seviyeleri yer almaktadır. Veri serileri analiz edildikten sonra, veri seti üzerinde herhangi bir ön işlem yapılmamasına karar verilmiştir.



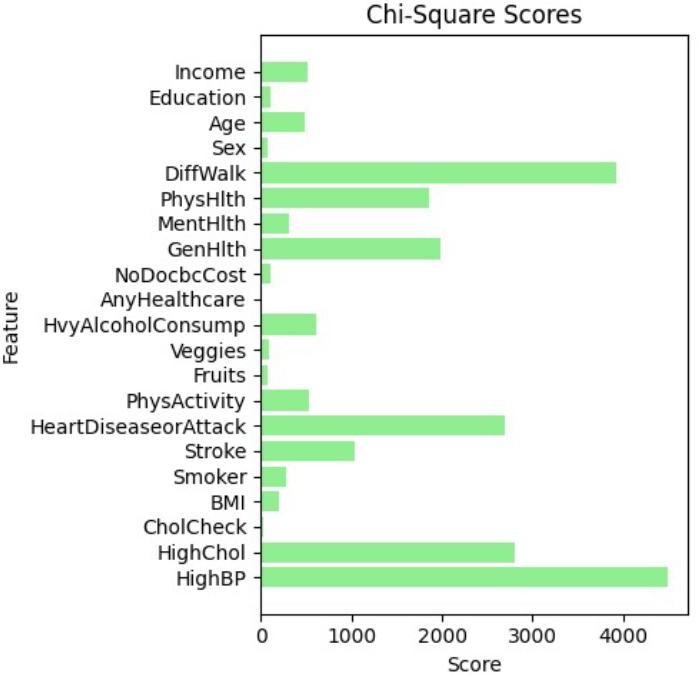
Şekil : Bağımsız Değişkenlerin Diyabet\_binary ile Korelasyonu

Şekil 2, veri setindeki bağımsız değişkenleri ve bunların diyabet üzerindeki korelasyon grafiğini göstermektedir. Bu grafiğe göre bağımsız değişkenlerin diyabet üzerindeki etkisi pozitif ya da negatif olabilmektedir. Pozitif değerler diyabet riskini artırırken, negatif değerler riski azaltmaktadır. Şekil 2'ye göre, ‘GenHlth’ ve ‘HighBP’ diyabetle güçlü bir şekilde ilişkilidir ve bu da diyabet riskini artırabilir. Öte yandan, ‘Gelir’ gibi değişkenler diyabet riskiyle negatif ilişkilidir, bu da daha yüksek gelir düzeyinin diyabet riskini azaltabileceğini düşündürmektedir.



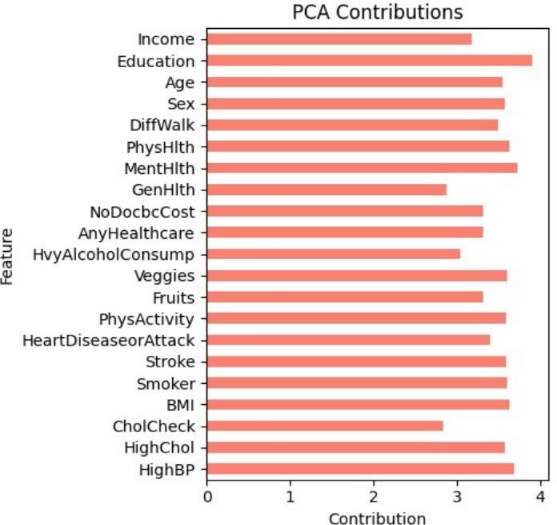
Şekil : Diabetes\_binary için Mutual Informatıon Skorları

Şekil 3'te gösterildiği gibi, özelliklerin hedef değişkenle bilgi paylaşımını temsil eden Mutual Information skorları değerlendirilmiştir. GenHlth, HighBP ve BMI yüksek skorlarla hedef değişkenle güçlü bir bilgi paylaşımı göstermektedir. Öte yandan, CholCheck, Smoker, Stroke, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, MentHlth, Sex, Education düşük skorlara sahiptir. Bu, bu özelliklerin hedef değişken üzerinde önemli bir etkiye sahip olmadığını ve modelden çıkarılabileceğini göstermektedir.



Şekil : Diabetes\_binary için Chi-Square Skorları

Şekil 4'te gösterildiği üzere, Chi-Square analizi kategorik değişkenlerin hedef değişkenle olan bağımlılık ilişkisini ortaya koymaktadır. HighBP, DiffWalk ve HeartDiseaseorAttack en yüksek Ki-Kare skorlarına sahiptir ve hedef değişkenle güçlü bir ilişki göstermektedir. Ancak CholCheck, Veggies, Fruits, AnyHealthcare ve Sex gibi düşük skorlu özellikler hedef değişkenle zayıf bir ilişki göstermektedir. Bu durum, bu özelliklerin hedef değişken üzerinde önemli bir etkiye sahip olmadığını ve modelden çıkarılabileceğini göstermektedir.



Şekil : Diabetes\_binary için PCA Katkısı

Şekil 5'te gösterildiği gibi, PCA analizi her bir özelliğin toplam varyansı açıklamadaki katkısını değerlendirmektedir. Fruits, MentHlth, ve Veggies PCA'da en yüksek katkıyı sağlamaktadır. Tüm özelliklerin katkısı önemli olduğu için PCA sonuçlarına göre çıkarılması gereken bir özellik bulunmamaktadır. PCA analizi boyut azaltma ve varyans açıklama açısından önemli bilgiler sağlamıştır ancak bu veri setinde herhangi bir özelliğin çıkarılmasına gerek duyulmamıştır.

Tablo : Dıabetes Health Indıcators Dataset için Özellı̇k Seçı̇mı̇

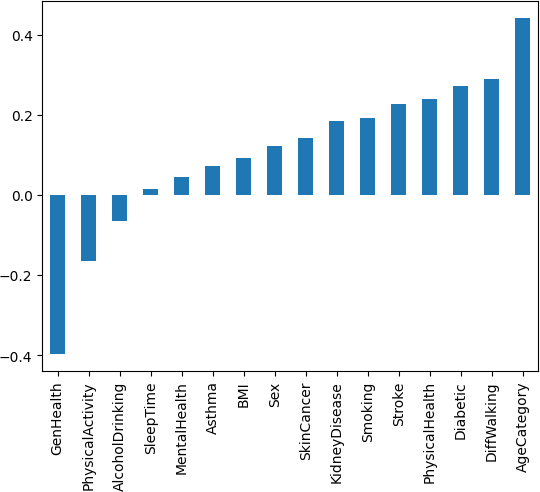
|  |  |  |  |
| --- | --- | --- | --- |
|  | Mutual Information | Chi Square | PCA |
| HighBP | + | + | + |
| HighChol | + | + | + |
| CholCheck | - | - | + |
| BMI | + | + | + |
| Smoker | - | + | + |
| Stroke | - | + | + |
| HeartDiseaseorAttack | + | + | + |
| PhysActivity | - | + | + |
| Fruits |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

Tablo 1, Diabetes Health Indicators Dataset üzerinde kullanılan üç farklı özellik seçme yöntemi sonucunda hangi özelliklerin seçildiğini göstermektedir.

3.1.2 Heart Attack Dataset

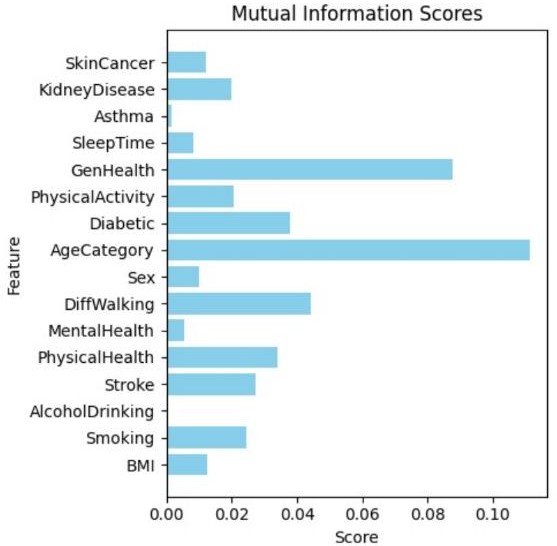
Bu araştırmada Kaggle'dan elde edilen Heart Attack Dataset kullanılmıştır. Veri seti 319795 satırdan oluşmaktadır ve HeartDisease, BMI, Smoking, Alcohol Drinking, Stroke, PhysicalHealth, MentalHealth, Walking Differently, Gender, AgeCategory, Diabetic, PhysicalActivity, GeneHealth, SleepTime, Asthma, KidneyDisease ve SkinCancer özelliklerini içermektedir. Hedef değişken, bireyin kalp hastalığı olup olmadığını gösterir (Evet: Sağlıklı, Hayır: Hastalık mevcut). Bu özellikler kalp hastalığının tahmini için anlamlı ve zengin bir veri kaynağı sağlamaktadır. Bu veri kümesini kullanmadan önce, veri kümesi üzerinde aşağıdaki ön işleme işlemlerini gerçekleştirilmiştir.

* Veri kümesindeki tüm evet değerleri 1 ve hayır değerleri 0 olarak değiştirildi.
* Veri kümesindeki tüm erkek değerleri 1 ve kadın değerleri 0 olarak değiştirildi.
* AgeCategory özelliği 0-12 aralığında yeniden kategorize edildi (0 = 18-24, 12 = 80 veya üzeri).
* GenHealth özelliği 0-4 aralığında yeniden kategorize edildi (0 = zayıf, 4 = mükemmel).
* Race özelliği kaldırılmıştır.
* Veri kümesi, HeartDisease özelliği eşit sayıda 0 ve 1 değerine sahip olacak şekilde dengelenmiştir.



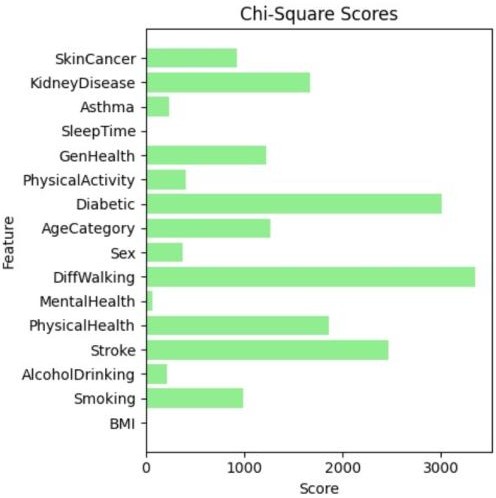
Şekil : Bağımsız Değişkenlerin HeartDisease ile Korelasyonu

Şekil 6'daki korelasyon grafiği ‘HeartDisease’ değişkeni ile diğer değişkenler arasındaki korelasyonu göstermektedir. En yüksek pozitif korelasyon yaş kategorisi ile görülmekte, bu da kalp hastalığı riskinin yaş ilerledikçe arttığını göstermektedir. Diabetes, Smoking, Stroke ve BMI gibi faktörler de pozitif korelasyon göstererek kalp hastalığı riskini artıran önemli değişkenler arasında yer almaktadır. Öte yandan genel sağlık durumu ve fiziksel aktivite negatif korelasyon gösteriyor, yani bu değişkenler arttıkça kalp hastalığı riski azalıyor. Ayrıca SleepTime ve MentalHealth gibi değişkenlerin daha zayıf ilişkilere sahip olduğu görülmektedir.



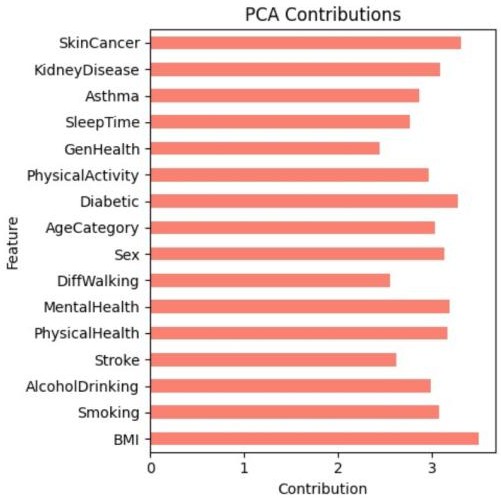
Şekil : Bağımsız Değişkenlerin HeartDisease ile Korelasyonu

Şekil 7'de gösterildiği gibi, özelliklerin hedef değişkenle bilgi paylaşımını ölçen Mutual Information skorları analiz edilmiştir. Bu analiz, AgeCategory ve GenHealth gibi değişkenlerin hedef değişken üzerinde en etkili özellikler olduğunu ortaya koymaktadır. PhysicalHealth, DiffWalking ve Stroke gibi değişkenler de kayda değer katkılar sağlarken, Smoking ve SleepTime gibi faktörlerin daha az etkili olduğu görülmektedir. Bu tür bir analiz, modelleme ve tahmin süreçlerinde hangi değişkenlere öncelik verilmesi gerektiğini belirlemeye yardımcı olabilir. Analiz sonucunda, hedef değişken üzerinde düşük bilgi skorlarına sahip BMI, AlcoholDrinking, MentalHealth ve Asthma özellikleri modelden çıkarılmıştır.



Şekil : HeartDisease için Chi-Square Skorları

Şekil 8'de gösterildiği üzere, özelliklerin hedef değişken ile bilgi paylaşımını ölçen Chi-Square Skorları analiz edilmiştir. Analiz sonucunda, hedef değişken üzerinde düşük Chi-Square skorlarına sahip olan BMI, MentalHealth ve SleepTime özellikleri modelden çıkarılmıştır. Bu özellikler hedef değişkenin tahminine önemli bir katkı sağlamadığı için modelin performansını artırmak ve gereksiz karmaşıklığı azaltmak amacıyla veri setinden çıkarılmaları uygun görülmüştür. Bu süreç, modelin daha güçlü ilişkilere sahip değişkenlere odaklanmasını sağlayarak daha etkili ve anlamlı bir analiz yapılmasına olanak tanımıştır. Analiz sonucunda özellikle AgeCategory, DiffWalking ve Diabetic özeliklerinin hedef değişken üzerinde güçlü etkileri olduğu tespit edilmiş ve model bu değişkenlere odaklanmıştır.



Şekil : HeartDisease için PCA Katkısı

Şekil 9'da gösterildiği gibi, PCA analizi her bir özelliğin toplam varyansı açıklamadaki katkısını değerlendirmektedir. Analiz sonucunda hiçbir özellik atlanmamıştır. Sonuçlar özellikle AgeCategory, Diabetic ve DiffWalking özeliklerinin modelin genel başarısı için kritik olduğunu göstermektedir. Diğer özelliklerin katkısı daha az olsa da, verilerin bütünlüğünü korumak ve daha geniş bir analiz perspektifi sağlamak için hepsini dahil etmek önemlidir. Bu sonuçlar, PCA'nın tüm özellikleri dikkate alan varyans açıklamalarına dayalı modelin nasıl yapılandırılacağı konusunda faydalı ipuçları sağlamaktadır.

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 | + | + | + |

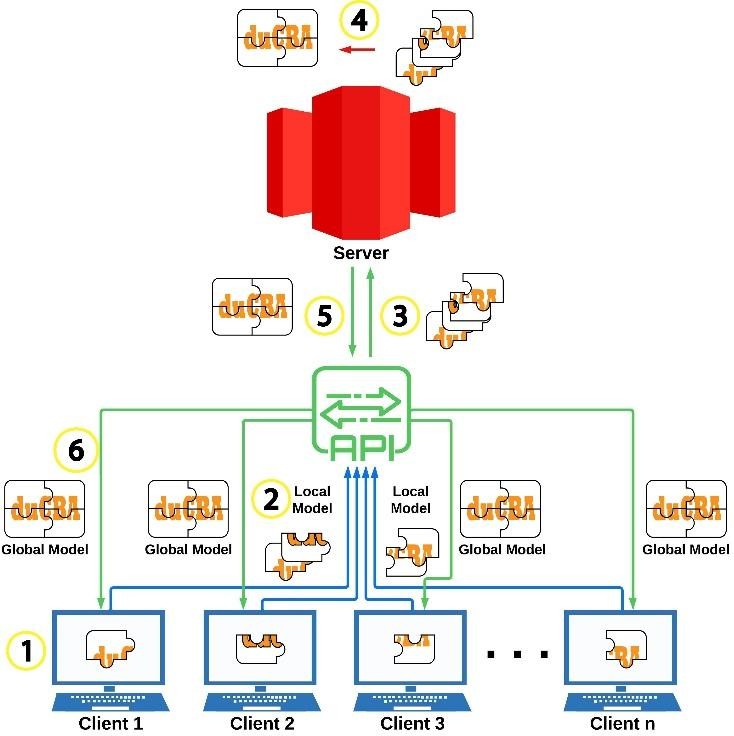
Tablo 2, Heart Attack Dataset üzerinde kullanılan üç farklı özellik seçme yöntemi sonucunda hangi özelliklerin seçildiğini göstermektedir.

## 3.2 Prototip Uygulama

Araştırma kapsamında geliştirilen prototip uygulamada üç farklı ortam kullanılmıştır.

* İstemciler
* API
* Ana Sunucu

Şekil 10'da prototip uygulamanın çalışma prensibi gösterilmektedir. İlk adımda yerel istemciler kendi verileri üzerinde CBA yöntemini kullanarak modelleri eğitmekte ve bu modelleri API sunucusuna göndermektedir. API sunucusu yerel modelleri toplar ve ana sunucuya iletir. Merkezi sunucu, gelen modelleri duCBA birleştirme algoritması ile birleştirerek güncellenmiş bir model oluşturur [23]. Bu güncellenmiş model API sunucusuna geri gönderilir. API sunucusu bu güncellenmiş modeli istemcilere dağıtarak her bir istemcinin daha sağlam ve küresel bir model kullanmasını sağlar. Şekil 10 bu süreci 1 tur olarak göstermektedir. Federe öğrenmenin yapısı gereği tur sürekli devam eder ve daha fazla veri ile eğitilen modeller eğitilir. Bu süreç, verilerin gizliliğini koruyarak yerel veriler üzerinden öğrenmeyi ve bu şekilde küresel bir model eğitmeyi amaçlayan bir yaklaşım sunmaktadır.



Şekil : duCBA Akışı Diyagramı

3.2.1 İstemciler

İstemciler için Google Cloud Platform üzerinde özdeş sanal makineler oluşturulmuştur. X86/64 mimarisini kullanan e2-medium makine tipi kullanılmıştır. Model eğitimi için Python 3.8.10 sürümü ve numpy, pandas, scikit-learn, pickle, pyfim, requests, pyarc kütüphaneleri kullanılmıştır. Model eğitim sürecinde kullanılan veri kümelerinin %20'si test için ayrılmış, kalan kısım ise istemci sayısına eşit olarak dağıtılmıştır. Model eğitimi için CBA algoritması kullanılmıştır. CBA için kullanacağımız support ve confidence değerleri sırasıyla 0,2 ve 0,5 olarak belirlenmiştir. Bu değerler önceden test edilmiş ve onaylanmış değerlerdir [21]. Bu ayarlarla oluşturulan ve eğitilen modeller HTTP isteği ile json formatında API'ye gönderilir. API bu modelleri toplanması için ana sunucuya gönderir. Ana sunucuda toplanan modeller API tarafından istemcilere gönderilir ve istemciler global model ile eğitimlerine devam ederler.

3.2.2 API

API için Google Cloud Platform üzerinde e2-medium makine tipi kullanılmıştır. Python 3.8.10 sürümü ve flask, numpy, pandas, scikit-learn, pickle, pyfim, websockets, requests, pyarc kütüphaneleri kullanılmıştır. API'ye gelen istekler 5000 portu kullanılarak alındı. Prototip uygulama testlerinde istemcilerden gelen modeller API ile 10 kez toplanarak ana sunucuya gönderilmiştir. Ana sunucuya websockets ile gönderilen bu modeller ana sunucudan alınarak istemcilere API ile dağıtılmıştır.

3.2.2 Ana Sunucu

Ana sunucu için Google Cloud Platform üzerinde e2-medium makine tipi kullanılmıştır. Python 3.8.10 sürümü ve numpy, pandas, scikit-learn, pickle, pyfim, websockets, requests, pyarc kütüphaneleri kullanılmıştır. Websocket bağlantısı 8000 portu kullanılarak gerçekleştirilmiştir. API'den gelen modelleri birleştirmek için duCBA birleştirme algoritması kullanılmıştır. Algoritma, her bir istemci üzerinde CBA yöntemi ile eğitilen yerel modelleri sunucu üzerinde toplamayı amaçlamaktadır. Birleştirme işlemi sırasında, her bir istemciden birleştirilen kurallar, support ve confidence değerleri yeniden hesaplanarak kapsamlı bir şekilde değerlendirilir ve birleştirilir [23]. Bu değerlendirme işlemi, farklı istemcilerden benzer veya aynı kurallar geldiğinde en uygun sonucu elde etmek için gerçekleştirilir. Birleştirme sırasında, veri kümesinde aynı etiketle işaretlenmiş ancak farklı istemcilerden gelen kaç kural olduğu kontrol edilerek support ve confidence değerleri güncellenir. Böylece her bir kuralın tüm sistem için geçerliliği daha doğru bir şekilde hesaplanır. Aynı içeriğe sahip ancak farklı etiketlerle işaretlenmiş kurallar arasında görülme sıklığı (support değeri) karşılaştırılır ve support değeri yüksek olan kural seçilerek nihai listeye eklenir. Güncellemeden sonra kurallar yeni confidence değerlerine göre sıralanır ve confidence değerlerinin eşit olması durumunda support değeri dikkate alınır. Support değerlerinin aynı olması durumunda listedeki ilk kural daha üst sırada yer alır.

Bu kapsamlı değerlendirme ve sıralama süreci, istemcilerden gelen modellerin katkıda bulunduğu tüm kuralların entegre edildiği birleştirilmiş bir nihai modelin oluşturulmasıyla sonuçlanır. İstemciler tarafından ana sunucudaki API aracılığıyla oluşturulan yerel modeller, daha önce belirtildiği gibi duCBA birleştirme algoritması kullanılarak birleştirilir. Ortaya çıkan global model daha sonra API aracılığıyla istemcilere geri gönderilerek federe öğrenme döngüsü tamamlanır.

# DENEYSEL SONUÇLAR

Bu bölümde, çalışmanın amacı doğrultusunda hipotezler ve araştırma sorularıyla ilgili verilere odaklanarak analiz sonuçlarını ayrıntılı olarak sunuyoruz.

Modelin precision, recall, F1-score ve accuracy değerlerini 2, 4 ve 8 istemcili dağıtık ortamlarda iki farklı veri kümesi kullanılarak değerlendirilmiştir. Testler sırasında üç farklı özellik seçim algoritması uygulanmış ve istemciler modelleri Yapay Sinir Ağı (YSA) ve CBA algoritmaları ile yerel veriler üzerinde eğitmiştir. YSA modelinin ilk katmanında 64 nöronlu bir Dense katmanı ve girdi boyutuna uygun bir ReLU aktivasyon fonksiyonu kullanılarak özellikler doğrusal olmayan bir şekilde dönüştürülürken, tek nöronlu ve sigmoid aktivasyon fonksiyonlu çıktı katmanında modelin çıktısı iki sınıf arasındaki olasılık olarak hesaplanmıştır. Optimizasyon için model, varsayılan öğrenme katsayısı 0.001 olan Adam optimizasyon algoritmasını kullanmış; sınıflandırma hataları kayıp fonksiyonu olarak binary\_crossentropy ile minimize edilmiş ve performansı bir doğruluk metriği ile değerlendirilmiştir. Model, 5 epok boyunca iterasyon başına 32 örnekten oluşan minibatch'lerle eğitilmiştir. FedAvg ve duCBA olmak üzere iki model toplama algoritması, merkezi sunucuda global modeli oluşturmak için kullanılmıştır. İstemci sayısının, özellik seçim yöntemlerinin ve model birleştirme algoritmalarının performans üzerindeki etkisini analiz etmek için tüm modeller 10 iterasyon boyunca eğitilmiştir.

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 eğitimi için veri kümesi olarak Diabetes Health Indicators Dataset kullanılarak elde edilen sonuçları göstermektedir. Bu sonuçlar üç farklı özellik seçim algoritmasının ve iki farklı toplama algoritmasının performansını 2, 4 ve 8 istemci ile karşılaştırmaktadır. Genel olarak FedAvg, özellikle istemci sayısı arttıkça daha yüksek F1-Skoru ve Precision değerleri ile öne çıkarken, duCBA daha az istemci ile daha istikrarlı ve dengeli sonuçlar sağlamaktadır. Bu sonuçlar duCBA'nın küçük ve orta ölçekli sistemlerde kararlı öğrenme sağlayabileceğini göstermektedir. Özellikle Mutual Information, Chi Square ve PCA gibi farklı özellik seçim algoritmaları ile benzer sonuçlar elde edilmesi, yöntemin çeşitli veri ön işleme yöntemleri ile uyumlu olduğunu göstermektedir. Bununla birlikte, istemci sayısındaki artış bazı durumlarda küçük dalgalanmalara neden olsa da, duCBA genel doğruluk açısından 0,69-0,70 arasında sabit değerler sağlayarak veri gizliliği gerektiren dağıtık sistemlerde güvenilir bir yöntem olarak öne çıkmaktadır. Bu sonuçlar duCBA'nın daha kararlı ve güvenilir performans gerektiren uygulamalar için uygun bir yöntem olduğunu göstermektedir.

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 eğitimi için veri kümesi olarak Heart Attack Dataset kullanılarak elde edilen sonuçları göstermektedir. Bu sonuçlar, üç farklı özellik seçim algoritması ve iki farklı toplama algoritmasının performansını 2, 4 ve 8 istemci ile karşılaştırmaktadır. Genel olarak FedAvg özellikle istemci sayısının artmasıyla birlikte daha yüksek F1-Score ve Precision değerleriyle öne çıkarken duCBA daha az istemciyle dengeli ve tutarlı sonuçlar vermektedir. duCBA, Mutual Information, Chi Square ve PCA gibi farklı özellik seçim algoritmalarıyla da benzer performans göstermiş ve bu yöntemle birlikte kullanılan veri ön işleme algoritmalarının etkisinin sınırlı olduğu gözlemlenmiştir. İstemci sayısındaki artış bazı metriklerde dalgalanmalara neden olsa da duCBA'nın doğruluk açısından sonuçları 0,67-0,68 arasında sabit kalarak yöntemin küçük ve orta ölçekli sistemler için güvenilir bir çözüm sunduğunu göstermiştir.

Elde edilen sonuçlar değerlendirildiğinde duCBA algoritmasının özellikle küçük ve orta ölçekli sistemlerde kararlı ve dengeli sonuçlar verdiği görülmektedir. duCBA, istemci sayısına bağlı olarak dalgalanmalar göstermeden kararlı ve güvenilir bir performans sergileyerek veri gizliliği gerektiren senaryolarda etkili bir yöntem olduğunu kanıtlamıştır. Ayrıca, özellik seçim algoritmalarından bağımsız olarak elde edilen tutarlı sonuçlar, yöntemin farklı veri ön işleme yöntemleriyle iyi çalıştığını göstermektedir. FedAvg, daha fazla istemciye sahip sistemlerde F1-Score ve Precision metriklerinde üstün performans sağlarken, duCBA düşük istemcili ve hassas veri senaryolarında daha etkilidir. Bu bulgular, duCBA'nın dağıtık sistemlerde güvenilirlik ve kararlılık hedefleyen uygulamalar için güçlü bir alternatif olduğunu göstermektedir.

Çalışmanın bundan sonraki kısmı için “Tartışma” ve “Sonuç” başlıkları eklenecektir.

**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