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 1. 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](#_Toc193191305)

[2. LİTERATÜR TARAMASI 3](#_Toc193191306)

[3. KULLANILAN YÖNTEMLER 6](#_Toc193191307)

[3.1 Veri Kümesi 6](#_Toc193191308)

[3.2 Uygulama 7](#_Toc193191309)

[3.2.1 İstemci 8](#_Toc193191310)

[3.2.2 API 9](#_Toc193191311)

[3.2.3 Ana Sunucu 9](#_Toc193191312)

[4. BULGULAR VE TARTIŞMA 10](#_Toc193191313)

**ÖZET**

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

Makine öğrenimi, büyük veri setlerinden anlamlı bilgiler çıkarma, karmaşık problemleri çözme ve gelecekle ilgili tahminler yapma yeteneği sayesinde birçok alanda ortaya çıkmıştır. Sağlık hizmetlerinde makine öğrenimi, hastalık teşhisi, tedavi planlaması, hasta takibi ve kişiselleştirilmiş sağlık hizmetleri gibi çeşitli uygulamalarda kullanılmaktadır. Ancak bu uygulamaların başarılı olabilmesi için büyük miktarlarda veriye ihtiyaç duyulmaktadır. Sağlık verilerinin gizliliği ve mahremiyeti konusundaki endişeler, bu verilerin merkezi sunucularda toplanmasını zorlaştırmakta ve veriye dayalı modellerin etkinliğini sınırlamaktadır. Mahremiyetin korunması ihtiyacı nedeniyle, bireysel hasta verilerinin merkezi bir sistemde toplanması ve analiz edilmesi büyük bir zorluk haline gelmektedir. Federe öğrenme, verilerin merkezi bir sunucuda toplanmadan yerel olarak işlenmesini ve merkezi sunucuya yalnızca model parametrelerinin gönderilmesini sağlayan bir makine öğrenimi yaklaşımıdır. Bu yöntem, verilerin gizliliğini korurken ağ trafiğini azaltarak güvenli model eğitimi sağlar. Bu makalede, federe öğrenme mimarisi için geliştirilen duCBA yöntemini kullanarak diyabet tahminine odaklanıyoruz. Model eğitimi için Kaggle platformundaki Diabetes Health Indicators adlı veri kümesi kullanılmıştır. Veri akışını yönetmek ve model eğitimini güvenli bir şekilde gerçekleştirmek için üç farklı ortam kullanılmıştır: bir ana sunucu, Flask tabanlı bir API sunucusu ve istemciler. İstemciler, verileri paylaşmadan kendi yerel verileri üzerinde dağıtılmış model eğitimi gerçekleştirerek güvenli veri gizliliği sağlar. API ve ana sunucu Google Cloud ortamında çalıştırılmıştır. API sunucusu, istemcilerden yerel model güncellemelerini toplar ve yeterli modele sahip olduğunda bunları WebSocket üzerinden ana sunucuya gönderir. Ana sunucu model birleştirme işlemini gerçekleştirdikten sonra güncellenmiş global modeli API sunucusuna iletir. API sunucusu daha sonra bu modeli istemcilere dağıtır. Bu yapı ile istemciler ana sunucuya doğrudan erişmeden API üzerinden merkezi bir global model eğitim sürecine katılmakta ve WebSocket protokolü ile iletişim güvenliği sağlanmaktadır. Prototipte yapılan testler, modelin geleneksel makine öğrenimi yöntemlerine benzer veya daha iyi performans gösterdiğini ortaya koymuştur. Bu durum, federe öğrenmenin ve duCBA algoritmasının veri gizliliğini korurken etkili bir tahmin modeli geliştirme potansiyelini ortaya koymaktadır. Bu sonuçlar, sağlık hizmetleri alanında federe öğrenmenin geniş uygulanabilirliğini desteklemektedir.

**Abdülbaki DEMİR**

**Ahmet Yasir KULAKSIZ**

**Mart, 2025**

**ABSTRACT**

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

Machine learning has emerged in many fields thanks to its ability to extract meaningful information from large data sets, solve complex problems, and make predictions about the future. In healthcare, machine learning is used in various applications such as disease diagnosis, treatment planning, patient monitoring, and personalized healthcare. However, for these applications to be successful, large size of data are needed. Concerns about the privacy and confidentiality of health data make it difficult to collect this data on centralized servers and limit the effectiveness of data-driven models. Because of the need to protect privacy, collecting and analyzing individual patient data in a centralized system becomes a major challenge. Federated learning is a machine learning approach that allows data to be processed locally without being collected on a central server, and only model parameters are sent to the central server. This method provides secure model training by reducing network traffic while maintaining the confidentiality of the data. In this paper, we focus on diabetes prediction using the duCBA method developed for federated learning architecture. For model training, a dataset called Diabetes Health Indicators on the Kaggle platform was used. To manage the data flow and perform model training securely, three different environments were used: a central server, a Flask-based API server, and clients. Clients provide secure data privacy by performing distributed model training on their local data without sharing the data. The API and the central server are run in the Google Cloud Platform environment. The API server collects local model updates from clients and sends them to the central server using WebSocket when it has enough models. After the central server performs the model merge, it sends the updated global model to the API server. The API server then distributes this model to the clients. With this structure, clients participate in a centralized global model training process through the API without directly accessing the central server, and communication security is ensured through the WebSocket protocol. As a result of the tests conducted in the prototype application, the accuracy value of the model was calculated as 70%. These results support the wide applicability of federated learning in the healthcare domain.

**Abdülbaki DEMİR**

**Ahmet Yasir KULAKSIZ**

**March, 2025**

**KISALTMALAR**

**IoT :** internet of things

**FedGiA :** federated gradient ınformation aggregation

**FedAvu :** federated asynchronous update

**FairFed :** fair federated learning

**RoFL :** robustness of secure federated learning

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

**CBA :** classification based on association rules

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

**BRFSS :** behavioral risk factor surveillance system

**API :** application programming ınterface

**HTTP :** hyper text transfer protocol

**JSON :** javascript object notation

**ŞEKİL LİSTESİ**

[Şekil 1: Bağımsız Değişkenlerin Bağımlı Değişkene Göre Korelasyonu 7](#_Toc192854177)

[Şekil 2: duCBA Çalışma Şeması 8](#_Toc192854178)

[Şekil 3: Global Model Karışıklık Matrisi 11](#_Toc192854179)

**TABLO LİSTESİ**

[Tablo 1: İterasyon Başına Accuracy Değerleri 10](#_Toc192854313)

[Tablo 2: duCBA Sınıflandırma Raporu 11](#_Toc192854314)

# GİRİŞ

Gelişen teknoloji ile IoT cihazların kullanımının artması beraberinde veri üretimininde artmasını sağlamıştır [1]. Bu cihazlar birçok farklı alanda kullanıldığı için çeşitli veri toplama imkanı sunmaktadır. Bu veriler kullanıcının davranışlarını analiz etmekten ortam koşullarını takip etmeye kadar birçok alanda kullanılmaktadır [2]. Bu veriler kullanılırken makine öğrenimi algoritmaları ile tahmin modelleri oluşturularak kullanılan sistemler oluşturulmaktadır [3].

Makine öğrenimi, verileri kullanarak öğrenebilen ve programlama ihtiyacı olmadan karmaşık örüntüleri ve ilişkileri belirleyen, daha önce görülmemiş verilere yönelik kararlar alan veya gelecekteki sonuçları tahmin eden bir yapay zeka türüdür [4]. Modelin etkili performans gösterebilmesi için büyük miktarda verinin analiz edilmesi gerekir. Ancak veri miktarının artması hem donanım kapasitesinin hem de modelin eğitim süresinin uzamasına yol açar [5]. Ayrıca, eğitim sürecinde verilerin tek bir sisteme iletilmesi zorunluluğu, depolama ve güvenlik gibi sorunları beraberinde getirir. Özellikle sağlık sektörü gibi büyük miktarda hassas verinin bulunduğu alanlarda, bu sorunlar daha da kritik hale gelir. Bu nedenle, sağlık verilerinin güvenli bir şekilde saklanması ve korunması büyük önem taşır [6-7]. Bu verilerin toplanması ve kullanımı konusunda birçok ülkede yasal düzenlemeler bulunmakta olup, kişisel verilerin gizliliğini korumak için teknolojik ve hukuki çözümler geliştirilmektedir [8]. Ancak kişisel verilerin korunması için oluşturulmuş bu yasalar makine öğrenimi projeleri için bazı zorluklar ortaya çıkarmıştır. Özellikle kişisel verilerin paylaşımı konusunda getirilen kısıtlamalar makine öğrenimi için gerekli olan veri setlerinin oluşturulamamasına yol açabilmektedir. Bu da makine öğrenimi için önemli bir sorundur.

Federe öğrenme, klasik makine öğrenmesi yöntemlerinden farklı bir yaklaşımdır. Bu yöntemde, model eğitimi için kullanılacak veriler merkezi bir sunucuya toplanmaz. Bunun yerine, modelin kullanılacağı ve veri üreten yerel ortamlarda modeller bağımsız olarak eğitilir. Eğitilen bu bağımsız modeller daha sonra başka bir ortamda birleştirilir. Böylece, genel bir model oluşturulmuş olur [9]. Federe öğrenmenin bu yapısının en önemli avantajları eğitilen modelin eğitim süresinin kısalması, veri gizliliğinin sağlanması ve merkezi yapı ile istemci arasındaki iletişimde kullanılan bant genişliğinin daha az kullanılmasıdır. Klasik model eğitim sürecinde eğitim verisi arttıkça eğitim süresi de artmaktadır. Ancak federe öğrenmede veriler küçük parçalara bölünerek eğitildiği için, eğitim süresi kısalır. Bununla beraber klasik model eğitim sürecinde merkezi bir sunucuya eğitim verilerinin gönderilmesi ve bu verilerin merkezi yapıda depolanması gerekmektedir. Federe öğrenme yönteminde eğitim verisi gönderimi yerine yerel cihazlarda eğitilen modeller gönderildiği için hem merkezi yapıda oluşan depolama sorunu hem de veri gönderimi sırasında oluşan bant genişliği sorununa bir çözüm getirmiş bulunmaktadır [10]. Ayrıca federe öğrenme yönteminde merkezi yapıya veri gönderilmediği için yerel cihazlarda bulunan kritik bilgilerin güvenliği de sağlanmış olmaktadır [11].

Tüm bu durumlar göz önünde bulundurulduğunda federe öğrenmenin avantajları şunlardır:

● Veri Gizliliği Sağlanır - Model eğitimi sürecinde yerel cihazlarda toplanan veriler merkezi bir sunucuya gönderilmeden eğitim yapılır. Bu sayede veriler paylaşılmadan model eğitimi sağlanmaktadır.

● Eğitim Süresi Kısalır - Tek seferde büyük miktarda işlemek yerine daha küçük miktardaki verilerin birçok cihazda işlenerek birleştirilmesi eğitim süresini kısaltmaktadır.

● Bant Genişliğinde Tasarruf Sağlanır- Merkezi sunucuya eğitilecek veriler yerine eğitilmiş modellerin gönderilmesi, bant genişliğinde tasarruf sağlamaktadır.

● Merkezi Sunucuda Depolama Problemleri Çözülür - Veriler merkezi sunucuda toplanmadığı için depolama alanı gereksinimi azalmaktadır.

● Yasal Düzenlemelere Uyum Sağlar - Kişilerin hassas verilerinin paylaşımıyla ilgili çeşitli yasal düzenlemeler bulunmaktadır. Federe öğrenme, verileri merkezi bir yapıya göndermeden işlediği için kullanıcı gizliliğini korumakta ve bu yasal düzenlemelerin getirdiği zorluklara karşı etkili bir çözüm sunmaktadır.

Federe öğrenme mimarisi eğitimden sağlığa, güvenlikten finansa birçok alanda kullanılmaktadır. Özellikle hasta mahremiyetini korumak adına sağlık alanında yapılan çalışmalar önemli ve günceldir. Bu çalışmada, federe öğrenme mimarisi kullanılarak diyabet tahmini üzerine bir uygulama geliştirilmiştir. Diyabet, dünya genelinde artan bir sağlık sorunu olarak karşımıza çıkmaktadır ve erken teşhisi büyük bir öneme sahiptir [12]. Ancak tıbbi verilerin gizliliği ve güvenliği, geleneksel makine öğrenimi yaklaşımlarının uygulanmasında önemli bir engel teşkil etmektedir. Federe öğrenme sayesinde, hasta verileri merkezi bir sunucuya aktarılmadan yerel cihazlarda işlenmiş, bu sayede hasta mahremiyeti korunmuştur.

Çalışmanın sonraki bölümlerinde şu konular incelenmiştir. Bölüm 2’de federe öğrenme için yapılan çalışmalar değerlendirilmiştir. Bölüm 3’de geliştirilen uygulama teknikleri detayları ile anlatılmıştır. Bölüm 4’de uygulama sonucu elde edilen bulgular sunulmaktadır.

# LİTERATÜR TARAMASI

Bu bölümde federe öğrenme mimarisi ve bu mimari kapsamında sağlık alanında yapılmış güncel çalışmalara yer verilmiştir.

2016 yılında veri mahremiyetini korumak adına Google tarafından ortaya çıkarılmış federe öğrenme mimarisinin birçok avantajı bulunmaktadır. Model eğitimi süresince veri gizliliği sağlanmakta, model eğitim süresi kısalmaktadır. İstemciler ve sunucular arasında veri yerine model dolaştığı için bant genişliği problemi ve iletişim gecikmesi sorunları ortadan kalkmaktadır. Ayrıca merkezi sunucuda veri yerine model toplandığı için depolama giderleri azalmaktadır. Federe öğrenmenin tüm bu avantajlarının yanı sıra, dört ana başlık altında toplanan bazı önemli sorunları bulunmaktadır. Bunlar; iletişim pahalılığı, sistem heterojenliği, istatistiksel heterojenlik ve gizlilik endişesidir. İletişim pahalılığı, yerel cihazların sayısının artmasıyla birlikte model gönderim sürecinin yönetiminin zorlaşmasıdır [13]. Sistem heterojenliği, model eğitiminin gerçekleştirildiği cihazların donanım ve yazılım farklarından kaynaklanan performans farklılıklarına yol açmasıdır [14]. İstatistiksel heterojenlik ise cihazların topladığı verilerin dağılımındaki farklılıklardan kaynaklanmaktadır. Bu durum modelin bazı cihazlarda iyi performans gösterip bazılarında zayıf kalmasına neden olabilmektedir [15]. Federe öğrenme yönteminde ana sunucuya model eğitimi için herhangi bir veri gönderilmeyip sadece yerelde eğitilen model gönderilmesiyle önemli bilgilerin gizliliği sağlanmış olsa da gönderilen modellerin parametrelerini kullanarak kritik bilgilere ulaşma endişesi hala bulunmaktadır. Bu durum gizlilik endişelerini devam ettirmektedir. Gizlilik endişelerine yönelik algoritmalar geliştirilmiştir [14-16].

Federe öğrenme mimarisi için geliştirilmiş olan birçok yöntem literatüre kazandırılmıştır. Federated Gradient Information Aggregation (FedGiA) algoritması, iletişim pahalılığı ve istatistiksel heterojenlik problemlerine çözüm sunarak hibrit bir yaklaşımla gradyan toplama ve model güncellemeyi birleştirmektedir. Bu sayede iletişim maliyetlerini düşürüp model doğruluğunu artırmayı hedeflemektedir [17]. Federated Asynchronous Update (FedAvu) yöntemi, sistem heterojenliği problemine asenkron model güncellemelerle çözüm getirerek katılımcıların kendi hızlarında çalışmasına izin vermektedir [18]. Fair Federated Learning (FairFed) algoritması ise farklı gruplar arasında dengeli model performansı sağlamaya odaklanarak istatistiksel heterojenlik problemini çözmektedir [19]. Gizlilik odaklı Robustness of Secure Federated Learning (RoFL) yöntemi ise kötü niyetli katılımcıların etkisini minimize eden güvenli gradyan toplama ve kriptografik tekniklerle veri gizliliğini ve güvenliği sağlamayı amaçlamaktadır [20]. Bu çalışma kapsamında kullanılacak olan ve literatüre kazandırılmış bir diğer federe öğrenme mimariside Veriden Habersiz İlişkili Kurallara Dayalı Sınıflandırma (Data Unaware Classification Based on Association, duCBA) yöntemidir [21]. duCBA uç cihazlarda Birliktelik Kurallarına Dayalı Sınıflandırma (Classification Based on Association Rules, CBA) [22] algoritması ile eğitilen modelleri birleştirmek için kullanılan bir yöntemdir. duCBA yönteminde modeller birleştirilirken kurallar önce kapsamlı bir değerlendirmeden geçirilir, ardından birleştirme süreci tamamlanır. Bu süreçte, her bir kuralın destek ve güven değerleri yeniden hesaplanarak kurallar sıralanır. Aynı etiketle işaretlenmiş kurallar farklı modellerden geldiğinde, her kuralın veri setinde ne kadar yer aldığı kontrol edilir ve bu doğrultuda destek ve güven değerleri güncellenir. Farklı etiketlerle işaretlenmiş ancak aynı içeriğe sahip kurallar için önce bulunma sıklığı karşılaştırılır; daha yüksek destek değeri olan kural seçilerek listede yer alır. Güncelleme işlemi sonrası, kurallar yeni güven değerlerine göre sıralanır. Eşit güven değerlerinde destek değerine bakılarak bir sıralama yapılır. Destek değerleri de aynıysa, listede önce gelen kural daha yüksek sıraya yerleştirilir. Tüm bu işlemler sonunda elde edilen kural listesi nihai modelin temelini oluşturur. Yani duCBA algoritması, ilişkisel sınıflandırma yaklaşımını temel alarak veri gizliliğini sağlamak amacıyla geliştirilmiştir. Bu algoritmanın temel amacı, merkezi bir veri toplama yerine istemci cihazlarda yerel olarak veriyle model eğitimi yaparak sadece eğitilmiş modelleri sunucuya göndermektir. Böylece, veri gizliliği korunurken ağ trafiği ve veri transferi minimize edilmektedir.

Federe öğrenme, tıbbi veriler üzerinde makine öğrenimi modelleri geliştirirken gizlilik ve güvenlik gereksinimlerini karşılamaya yönelik güçlü bir yöntem olarak öne çıkmaktadır. Bu yöntem, hasta verilerinin merkezi bir sunucuya taşınmasını gereksiz kılarak, verilerin yerel cihazlarda işlenmesiyle veri gizliliğini korur. Yapılan değerlendirmeler, federe öğrenmenin bu gizlilik ve güvenlik odaklı yapısının, tıbbi veriler üzerinde uygulanabilir olduğunu ve dağıtık veri işleme modelleri geliştirilirken hasta gizliliğinin korunmasına önemli katkılar sağladığını göstermektedir [6-8].

Choudhury ve arkadaşları (2019) [23] yaptıkları çalışmada, Diferansiyel gizlilik ile desteklenen federe öğrenme modellerinin, hassas sağlık verilerinin merkezi sistemlerde toplanmadan yerel olarak işlenmesini sağladığını ve böylece hasta verilerinin sızma riskini önemli ölçüde azalttığını ifade etmektedir. Bu yöntem sayesinde, sağlık verilerinin güvenli bir şekilde analiz edilmesi mümkün hale gelirken bireysel verilerin mahremiyeti korunmaktadır. Benzer şekilde, Horvath ve arkadaşlarının (2023) [24] çalışması da MIMIC-III veri kümesi üzerinde diferansiyel gizlilik ile federe öğrenmenin etkinliğini araştırmıştır. Bu çalışmada, hasta verilerinin gizliliğini koruyarak başarılı modeller oluşturulabileceği gösterilmiş ve merkezi olmayan öğrenme süreçlerinin tıbbi verilerin farklı kurumlar arasında güvenle paylaşılmasını sağladığı belirtilmiştir. Durga ve arkadaşlarının çalışmasında (2021) [25], federe öğrenme modelinin blockchain tabanlı güvenlik önlemleriyle entegre edildiği Healthchain sistemi ele alınmıştır. Bu sistemin, verilerin güvenli şekilde işlenmesine ve veri paylaşımının merkezi bir aracıya ihtiyaç duyulmadan gerçekleştirilmesine olanak tanıdığı ifade edilmiştir. Böylece federe öğrenme, tıbbi veriler üzerinde gizlilik ve güvenliği ön planda tutarak daha etkili ve güvenilir makine öğrenimi modelleri geliştirmenin anahtarı olarak değerlendirilmektedir.

Chen ve arkadaşlarının (2023) [26] makine öğrenimi algoritmalarının karşılaştırılması ve özellik görselleştirme analizi üzerine yaptıkları çalışmada, farklı algoritmaların diyabet tahmininde doğruluk performansları incelenmiş ve belirli algoritmaların risk tahmininde daha başarılı olduğu gösterilmiştir. Benzer şekilde, Hama Saeed'in (2023) [27] diyabet sınıflandırması için yaptığı başka bir çalışmada, tip 2 diyabet sınıflandırmasında, dengesiz veri setlerini dengelemek için up-sampling tekniği kullanılarak makine öğrenimi algoritmalarının doğruluğu artırılmıştır. Bu çalışmalar, makine öğrenimi algoritmalarının diyabet risk tahmini ve sınıflandırması konusunda güçlü araçlar sunduğunu ve verilerin dengelenmesi gibi tekniklerin performansı daha da iyileştirebileceğini göstermektedir.

Yapılan bu çalışmalar, federe öğrenmenin sağlık alanındaki önemli bir uygulama alanı olduğunu ve bu alanda aktif bir araştırma sürecinin devam ettiğini göstermektedir. Özellikle veri gizliliği, güvenlik ve sistem verimliliği gibi kritik sorunlara çözüm arayarak, federe öğrenme, tıbbi veriler üzerinde güçlü ve etkili makine öğrenimi modellerinin geliştirilmesine olanak tanımaktadır. Bu bağlamda, mevcut literatür, federe öğrenmenin tıbbi verilerdeki gizliliği koruma ve verimliliği artırma potansiyelini ortaya koymaktadır. Bu çalışma ile literatüre, federe öğrenme yöntemlerinin sağlık alanındaki uygulamaları konusunda yeni bir katkı sağlanması hedeflenmektedir.

# KULLANILAN YÖNTEMLER

Çalışmanın bu bölümünde model eğitimi için kullanılan veri seti, geliştirilen uygulama için kullanılan teknolojiler ve yöntemler detaylı bir şekilde açıklanmaktadır.

## 3.1 Veri Kümesi

Bu çalışmanın yürütülmesinde, modelleri eğitmek ve değerlendirmek için Kaggle platformunda bulunan "Diabetes Health Indicators Dataset"[28] veri seti kullanılmıştır. Veri seti, ABD Hastalık Kontrol ve Önleme Merkezleri (CDC) tarafından yürütülen Behavioral Risk Factor Surveillance System (BRFSS) anket verilerine dayanmaktadır. Bu veri seti, diyabet hastalığı ve bireylerin genel sağlık durumlarını etkileyebilecek diğer faktörler hakkında bilgiler içermektedir. Veri seti her iki sınıf içinde dengelenmiş olup 70.692 satır ve 22 sütundan oluşmaktadır. Bağımlı değişken olan "Diabetes\_binary" sütunu, katılımcının diyabet durumu hakkında ikili bir bilgi sunmaktadır. Diğer 21 bağımsız değişken, bireylerin yaşam tarzı, sağlık davranışları ve genel sağlık durumlarını temsil eden çeşitli faktörleri kapsamaktadır. Bu değişkenler arasında yüksek tansiyon, kolesterol seviyesi, vücut kitle indeksi (BMI), fiziksel aktivite durumu, sigara kullanımı, kalp rahatsızlığı geçmişi, genel sağlık durumu ve gelir seviyeleri gibi bilgiler yer almaktadır. Veri seti analizi ve görselleştirme için Python programlama dili kullanılmıştır. Ayrıca Pandas, Scikit-learn, Numpy Seaborn ve Matplotlib kütüphaneleri kullanılarak veri ön işleme ve değerlendirme aşamaları gerçekleştirilmiştir.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil : Bağımsız Değişkenlerin Bağımlı Değişkene Göre Korelasyonu

Şekil 1'de veri seti içerisinde bulunan bağımsız değişkenler ve bu değişkenlerin, diyabet üzerindeki korelasyon grafiği yer almaktadır. Bu grafiğe göre bağımsız değişkenlerin diyabet üzerindeki etkisi pozitif ya da negatif olabilir. Pozitif değerler diyabet riskini artırırken, negatif değerler riskini azaltmaktadır. Şekle göre 'GenHlth' ve 'HighBP' diyabetle güçlü bir şekilde ilişkilidir ve bu durum diyabet riskini artırabilir. Öte yandan, 'Gelir' gibi değişkenler diyabet riskiyle negatif ilişkilidir, bu da daha yüksek gelir seviyesinin diyabet riskini azaltabileceğini göstermektedir.

## 3.2 Uygulama

Çalışma kapsamında geliştirilen uygulamada üç farklı ortam kullanılmıştır.

● İstemci

● API

● Ana Sunucu

metin, diyagram, çizgi film, daire içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil : duCBA Çalışma Şeması

Şekil 2'de prototip uygulamanın çalışma prensibi gösterilmiştir. İlk adımda yerel istemciler, kendi verileri üzerinde CBA yöntemini kullanarak model eğitmektedir ve bu modelleri API sunucusuna göndermektedir. API sunucusu, yerel modelleri toplamakta ve merkezi sunucuya iletmektedir. Merkezi sunucu, gelen modelleri duCBA yöntemi ile birleştirerek güncellenmiş bir model oluşturmaktadır. Güncellenmiş bu model API sunucusuna geri gönderilmektedir. API sunucusu ise bu güncellenmiş modeli istemcilere dağıtarak her istemcinin daha güçlü ve ortak bir model kullanmasını sağlamaktadır. Şekil 2'de bu işlem süreci 1 tur olarak ifade edilmiştir. Federe öğrenme yapısı gereği tur sürekli devam ederek daha çok veriyle eğitilmiş modeller geliştirilmektedir. Bu süreç, verilerin gizliliğini koruyarak, yerel veri üzerinde öğrenme yapmayı ve bu şekilde ortak bir model oluşturmayı hedefleyen yaklaşımı sağlamaktadır.

### 3.2.1 İstemci

İstemciler için Google Cloud Platform üzerinden 25 adet eş özellikte sanal makine oluşturulmuştur. x86/64 mimarisi kullanılan 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 eğitimi sürecinde veri setinden rastgele 1000-3000 arasında veriler seçilerek model eğitilmiştir. Model eğitimi için CBA algoritması kullanılmıştır. CBA için kullanacağımız destek, güven değerleri sırasıyla 0,2 ve 0,5 olarak belirlenmiştir. Belirlenen bu değerler önceden test edilmiş ve kabul görmüş değerlerdir [21]. Bu ayarlar ile oluşturulan ve eğitilen modeller HTTP isteği ile json formatında API'ye gönderilmektedir. API bu modelleri birleştirmesi için ana sunucuya göndermektedir. Ana sunucuda birleştirilen modeller API aracılığıyla istemciye gönderilir ve istemcinin global modeli ile eğitimlerine devam etmektedir.

### 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 numaralı port üzerinden alınmıştır. Pilot uygulama testlerinde API ile 25 adet istemciden gelen modeller 10 kez toplanıp ana sunucuya iletilmiştir. Ana sunucuya websocket üzerinden iletilen bu modeller ana sunucudan birleştirilmiş şekilde alınarak istemcilere API üzerinden dağıtılmıştır.

### 3.2.3 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 numaralı port üzerinden gerçekleştirilmiştir. API'den gelen modelleri birleştirmek için duCBA algoritması kullanılmıştır. Algoritma, her istemcide CBA yöntemi ile eğitilmiş yerel modelleri sunarak birleştirmeyi amaçlamaktadır. Birleştirme süreci sırasında her bir istemciden elde edilen kurallar kapsamlı bir değerlendirmeden geçirilir ve destek (destek) ve güven (güven) değerleri yeniden hesaplanarak birleştirilir. Bu değerlendirme süreci, benzer veya aynı kuralların farklı istemcilerden geldiği durumlarda optimal bir sonuç elde edilmesi amacıyla gerçekleştirilir. Birleştirme sırasında, aynı etiketle işaretlenmiş fakat farklı istemcilerden gelen kuralların veri setinde ne kadar yer aldığı kontrol edilerek destek ve güven değerleri güncellenir. Böylece, her kuralın tüm sistem için geçerliliği daha doğru bir şekilde hesaplanır. Aynı içeriğe sahip ancak farklı etiketlerle işaretlenmiş kurallar arasında ise bulunma sıklığı (destek değeri) karşılaştırılır ve daha yüksek destek değeri olan kural seçilerek nihai listeye eklenir. Güncelleme sonrasında, kurallar yeni güven değerlerine göre sıralanır ve eşit güven değerlerinde destek değeri dikkate alınır. Destek değerleri de aynı şekilde, listede önce gelen kural daha yüksek sıraya yerleştirilir.

Bu kapsamlı değerlendirme ve sıralama işlemleri sonucunda, istemcilerden gelen tüm kuralların katkıda bulunduğu birleştirilmiş bir model elde edilmektedir.

Ana sunucu üzerinde API ile istemcinin oluşturduğu modeller yukarıda belirtildiği gibi duCBA algoritmasıyla birleştirilir ve oluşturulan global model tekrardan API aracılığıyla istemciye iletilir.

# BULGULAR VE TARTIŞMA

Bu Bu bölümde, araştırma kapsamında elde edilen verilerin analiz edilmesi sonucunda ortaya çıkan bulgular ayrıntılı olarak sunulmuştur. Bulgular, çalışmanın amacına uygun olarak belirlenen hipotezler ve araştırma sorularına yönelik verileri kapsamaktadır.

Prototip uygulamanın test sürecinde 25 adet istemcinin model eğitimi için toplam 10 iterasyonda birleştirilen global modellerin doğruluk değerleri. Tablo 1'de gösterilmektedir.

Tablo : İterasyon Başına Accuracy Değerleri

| **İterasyon** | **Accuracy** |
| --- | --- |
| 1 | 0.71 |
| 2 | 0.70 |
| 3 | 0.69 |
| 4 | 0.70 |
| 5 | 0.70 |
| 6 | 0.70 |
| 7 | 0.69 |
| 8 | 0.70 |
| 9 | 0.69 |
| 10 | 0.70 |

Tablo 1'de görülen bu değerler 0.69 ile 0.71 arasında değişmektedir. Genel olarak modelin doğruluğu, iterasyonlar boyunca küçük dalgalanmalar göstererek 0,69 ve 0,70 civarında sabit kalmıştır. Bu durum modelin tutarlı ancak belirgin bir gelişme göstermeyen bir performansa sahip olduğunu düşündürmektedir. Modelin performansını iyileştirmek için hiper parametre ayarlamaları yapmak, veri setini çeşitlendirmek veya erken durdurma gibi yöntemler denenebilir.

**metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

Şekil : Global Model Karışıklık Matrisi

Şekil 3'te duCBA algoritmasıyla birleştirilen 10 iterasyon sonunda elde edilen global modelin 14000 test verisi üzerinde çalıştırılmasıyla elde edilen sonuçlarla oluşturulan karışıklık matrisi bulunmaktadır. Bu matristeki veriler ile precision, recall ve f1-score değerleri hesaplanmıştır. Bu veriler tablo 2'de gösterilmektedir.

Tablo : duCBA Sınıflandırma Raporu

|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| 0 | 0.68 | 0.76 | 0.72 | 7000 |
| 1 | 0.73 | 0.64 | 0.68 | 7000 |
| Accuracy |  | | 0.70 | 14000 |
| Macro Avg. | 0.70 | 0.70 | 0.70 | 14000 |
| Weighted Avg. | 0.70 | 0.70 | 0.70 | 14000 |

Tablo 2'de, duCBA ile birleştirilen global modelin diyabet tahmini üzerindeki performansını değerlendiren metrikler görülmektedir. Precision, Recall ve F1-score gibi ölçütler, modelin her sınıf için nasıl performans gösterdiğini ortaya koymaktadır. Sınıf 0 için Precision 0.68, Recall 0.76 ve F1-score 0.72 olarak hesaplanmıştır. Bu durum, modelin sınıf 0'ı tahmin etmede nispeten daha fazla duyarlı olduğunu, ancak bazı yanlış pozitifler de ürettiğini göstermektedir. Sınıf 1 için Precision 0.73, Recall 0.64 ve F1-score 0.68 değerlerine sahiptir; bu da modelin sınıf 1'de daha kesin tahminler yaptığını, ancak gerçekte sınıf 1 olan örnekleri bulmada sınıf 0'a göre daha az başarılı olduğunu göstermektedir. Global modelin genel doğruluğu %70 olarak verilmiştir. Bu doğruluk oranı, modelin genel olarak dengeli bir performans sergilediğini ifade etmektedir. Makro ortalama ve ağırlıklı ortalama değerleri de her iki sınıfta benzer olarak 0,70 olarak hesaplanmıştır, bu da sınıflar arasında belirgin bir performans farkı olmadığını göstermektedir.

Çalışmanın bundan sonraki kısmında farklı birleştirme algoritmaları kullanılarak, duCBA algoritmasının diğer algoritmalara göre precision, recall, f1-score ve accuracy değerleri karşılaştırılacaktır. Ayrıca farklı özellik seçim algoritmaları ve veri setleri kullanılarak daha kapsamlı bir karşılaştırma yapılması hedeflenmektedir.

**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] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] BÜYÜKTANIR, B., YILDIZ, K., ÜLKÜ, E. E., & Bütüktanir, T., du-CBA: Data-agnostic and incremental 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

[22] Abouelmehdi, K., Beni-Hssane, A., Khaloufi, H., & Saadi, M., Big data security and privacy in healthcare: A review. Procedia Computer Science, 113, 73-80, 2017

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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