YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 12, Sayı 1, Ocak 2024

SAĞLIKTA DİJİTAL İKİZ (DIGITAL TWIN IN HEALTHCARE)

ELANUR İMİRGİ

Ondokuz Mayıs Üniversitesi, Bilgisayar Mühendisliği, elanurimirgi@gmail.com

Özet

Bu makale, sağlık sektöründe dijital ikiz kavramının nasıl kullanılabileceğini ve bu kullanımın diyabet gibi hastalıkların teşhisi üzerindeki etkisini ele almaktadır. Dijital ikiz, fiziksel dünyadaki sağlık verilerinin dijital bir kopyasını oluşturarak, hastalıkların nasıl etkilendiğini ve çeşitli faktörlerin hastalıklarla nasıl ilişkilendirildiğini incelemek için kullanılır. Bu makale, CRISP-DM (Cross-Industry Standard Process for Data Mining) metodolojisi kullanılarak Gradient Boosting Regresyon, SVM (Support Vector Machine), KNN (K Nearest Neighbors), Lojistik Regresyon, Random Forest Classifier, Naive Bayes, XGBoost, Decision Tree gibi makine öğrenimi algoritmalarıyla ANN (Multi-Layer Perceptron), LSTM (Long Short-Term Memory), MLP (Multi-Layer Perceptron) gibi derin öğrenme algoritmalarının nasıl uygulanabileceğini ve sonuçlarını açıklamaktadır. Makale, uyku kalitesi, yaş, beden kitle indeksi, stres seviyesi, uyku süresi gibi faktörlerin diyabet riski üzerindeki etkisini araştırarak, bu faktörlere dayalı olarak kişilerin diyabet hastası olup olmadığını tespit etmeyi amaçlar. Ayrıca, bu algoritmalar arasındaki farkları vurgulayarak gelecekte bu alanda yapılacak çalışmalara ilham kaynağı olmayı hedefler.

Anahtar Kelimeler: Dijital İkiz, Sağlık, Makine Öğrenmesi, Derin Öğrenme

Abstract

This article explores the use of the digital twin concept in the healthcare sector and its impact on the diagnosis of diseases such as diabetes. The digital twin aims to create a digital copy of physical-world health data, allowing for an examination of how diseases are affected and how various factors are related to them.The article applies machine learning algorithms such as Gradient Boosting Regression, SVM, KNN, Logistic Regression, Random Forest Classifier, Naive Bayes, XGBoost, and Decision Tree, as well as deep learning algorithms like Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and Multi-Layer Perceptron (MLP) using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. It explains how these algorithms are implemented and presents their results.Furthermore, the article investigates the impact of factors such as sleep quality, age, body mass index, stress levels, and sleep duration on the risk of diabetes. Its aim is to determine whether individuals are diabetic or not based on these factors. Additionally, the article highlights the differences between these algorithms, intending to inspire future research in this field.

Keywords: Digital Twin, Health, Machine Learning, Deep Learning

1 GİRİŞ

Dijital ikiz maddesel bir ürün veya hizmetin gerçek dünyadaki bir yansıması olarak karşımıza çıkmaktadır. Bu kavram hizmet yada ürünün sanal bir modelidir. Gittikçe popülerleşen yapay zekayla beraber her alanda mevcut sorunlara çözüm aranmaya başlanmasıyla ortaya atılan fikirlerden biri olan dijital ikiz nesnelerin dijital ortamda birebir aynılarının var edilmesine denir. Gerçek hayatta kullanılan verilerin dijital dünyada birer kopyalarının oluşturulmasıyla denenebilecek çözüm yöntemlerinin önce dijital ortamda denenerek hem maliyet hem zaman tasarrufu yapılabilmesi hem de gerçek zamanlı olarak modellerin izlenebilmesi sayesinde günümüzde oldukça popüler bir kavramdır. Biraz daha detaylandırmak gerekirse dijital ikiz gerçek zamanlı alınan verilerin sensörlerle ve tıbbi nesnelerin interneti ile yapay zeka modelleri kullanılarak gerçek zamanlı incelenerek bulut tabanlı bir sisteme aktarılıp izlenmesiyle yapılır [1,2].

Özellikle son dönemde adı çok duyulan dijital ikiz kavramının tarihçesi 1960’lara kadar uzanmaktadır. O yıllarda NASA uzay teknolojilerinin uzay roketlerinin sorunlarını giderme amacıyla bu kavrama ilişkin çalışmaları başlatmıştır. Fakat toplumun kavramı tanıması yaklaşık olarak 42 yıl sonra Michigan Üniversitesi’nde yapılan sunum sayesinde olmuştur. Duyulduğundan bu yana üzerinde pek çok çalışma yapılan dijital ikiz, günümüzde makine öğrenmesi, Internet of Everything (IoE), blokzincir, bulut bilişim, artırılmış gerçeklik (AR), haberleşme teknolojileri, otomotiv ve havacılık, tedarik zinciri ve perakende, akıllı şehirler, akıllı bina ve işletmeler, sağlık hizmetleri [3] gibi pek çok alanda 3 boyutlu bina modelleri, bireylerin genetik ve fizyolojik yapıları, şehir planlamalarında enerji tasarrufu, araba yarışlarına performans iyileştirme gibi amaçlarla kullanılmaktadır [4].

Yazılan makale özelinde sağlık alanında makine öğrenmesi; yaygın olarak hastalık tanısı, hastalık sonrası oluşabilecek komplikasyonlar, hastalık tahmini gibi alanlarda kullanılmakta ve aynı zamanda iş yükü, zaman tasarrufu ile hastaların kaliteli sağlık hizmeti alabilmesi amacıyla kullanılmaktadır. Sağlıkta makine, öğrenimi; diyabeti erken teşhis ederek diyabete bağlı ortaya çıkabilecek diğer hastalıkların önüne geçmek, ayrıca diyabet ve buna bağlı hastalıklardan dolayı oluşacak maliyetin yüksek oranda önüne geçilmesi amacıyla kullanılmaktadır.

Konu hakkında birçok çalışma mevcuttur. Bu çalışmalardan birinde diyabet hastalığının makine öğrenmesi yoluyla eldeki doğruluk tespit oranını arttırma amacıyla büyük veri kullanma yoluna gidilmiştir. Herhangi bir hastanın eldeki veri kümesi kullanılan büyük veri kümesindeki parametrelerle çoğaltılarak hastalık tanısının konmasındaki doğruluk oranı arttırılmıştır. Farklı bir çalışmada hastalığın tahmini, hastalık sonrası ortaya çıkan komplikasyonlar, hastaların tedavi yönetim süreçleriyle genetik geçmişleri, yapılan farklı çalışmalar aracılığıyla yorumlanmış ve buradan çıkarılan sonuçlardan hareketle makine öğrenimi algoritmaları ağırlıklı sınıflandırma ile beraber tahmin algoritmaları da kullanılmıştır. Yapılan farklı çalışmada örnek olarak farelerin beyinlerindeki zedelenme verileri kullanılarak makine öğrenimi ve derin öğrenme algoritmaları kullanılmış en iyi sonuç CNN (Convolutional Neural Network) ile elde edilmiştir. Yine farklı bir çalışmada ise solunum fonksiyon testlerine bağlı olarak hastaların akciğer kanseri tipleri tahmini yapılıp en iyi sonucu KNN (K-Nearest Neighbour) algoritmasının verdiği görülmüştür [5].

Sağlık alanındaki dijital ikiz kullanımını detaylandırmak gerekirse başlıca kullanılan 3 türü vardır. İlk türü hastanenin ölçüm yapılması istenen bölümlerine yerleştirilen sensörler ile dijital ikizinin tasarlanmasıdır. İkincisi insanların sağlık takibini yapabilmek için toplanan sağlık verilerinden oluşturulan dijital ikizin tasarlanmasıdır. Üçüncüsü ise hastalıklı bölgelerin takibini sağlamak ve bu bölgelere yapılacak operasyonların veya verilecek ilaçların etkilerini test için tasarlanan dijital ikizlerdir [6].

Sağlık alanında ayna dünyaları (mirror worlds) teorisiyle ajan tabanlı dijital ikizi kullanarak yapılan çalışmalar incelendiğinde, Valk ve ekibinin travma yönetimi için geliştirdikleri yöntem öne çıkmaktadır. Bu çalışmada, ajan tabanlı dijital ikizler, fiziksel varlıkları ve süreçleri dijital olarak temsil eden ve bu temsilleri izleyen yazılım araçları olarak kullanılmıştır. Bu ajanlar, dijital ikizlerle etkileşim kurarak verileri güncellemekte ve travma dokümantasyonunu

yönetmektedirler. Bu yaklaşım, sağlık profesyonellerine destek sağlamak ve dijital ikizlerin verilerini anlamlı bilgilere dönüştürmek amacıyla kullanılmıştır [7].

Wickramasinghe ve arkadaşları ise dijital ikizlerin kullanımını, herhangi bir hastanın mevcut bir rahatsızlığına dayalı olarak Internet of Things (IOT) cihazlarından elde edilen verileri kullanarak gerçekleştirmiştir. Bu yöntemde, siyah kutu (black-box) modeli kullanılarak hastaya en yakın dijital ikiz elde edilmiştir. Bu sayede kişiye özgü tedavi yöntemlerinin uygulanabileceği düşünülmüştür [8].

Gilette ve ekibi, insan kalbinin dijital ikizini elektrofizyoloji sinyalleri aracılığıyla oluşturarak daha güvenli ve ekonomik müdahale yöntemlerinin geliştirilebileceğini savunmuşlardır [9].

Multipl skleroz hastalığının tedavisi konusunda Voigt ve arkadaşları, yapay zeka ile birleştirilen dijital ikiz kullanarak hastanın hastalık verileri ve tıbbi kayıtlarının analizini yapmış ve hastaya özgü tedavi yöntemleri uygulamışlardır. Bu çalışmada, Support Vector Machine (SVM) modelinin Lojistik Regresyon modelinden daha etkili sonuçlar verdiği gözlemlenmiştir [10].

Son olarak, Chakshu ve ekibi yoğun bakım ünitelerindeki zatürre hastalarının hastalık boyutunu öğrenmek için derin öğrenme modellerini önermişlerdir. Bu yöntem, daha kritik durumdaki hastaların tedavilerinin önceliklendirilmesine yardımcı olabilir. Ayrıca, yapılan deneyler sonucunda MultiLayer Perceptron (MLP) algoritmasının en iyi sonuçları verdiği gözlemlenmiştir [11].

Hussain ve arkadaşları ise elektroensefalografi (EEG) verileri yardımıyla inme tespitinde makine öğrenmesi modellerinden SVM (Support Vector Machine) ile %76 accuracy, %73 sensitivity, %79 specificity, %77 precision, %84 AUC; Random Tree ile %70 accuracy, %74 sensitivity, %66 specificity, %68 precision, %78 AUC; Lojistik Regresyon ile %69 accuracy, %67 sensitivity, %71 specificity, %69 precision, %76 AUC metrikleri ile sonuçlar bulunmuş ve en iyi sonucu %84 ile SVM modelinin AUC metriğinin verdiği görülmüştür [12]. Martinez-Velazquez ve arkadaşları kardiyovasküler dijital ikiz mimarisi kullanarak kalbin güncel durumu hakkında bilgi sahibi olunabileceğini tavsiye etmişlerdir. Farklı platformlardan (sensörler, tıbbi belgeler gibi) alınan veriler birleştirilip standartlaştırılarak analiz edilir ve daha sonra 3 evrişim havuzlama katmanı, düzleştirilmiş bir tabaka, relu ve sigmoidden oluşan bir adet çıkış nöronundan oluşan bir CNN modeli ile accuracy metriğininde %85,81, precision metriğinde %86,29, specifity metriğinde ise %83,87 oranlarında doğruluk elde edilir. Makalenin sonunda bu doğrulukların daha fazla sensör verisi kullanılarak daha da iyileştirilebileceği konusunda tavsiye verilir [13]. Xu ve arkadaşları mevcut hastaların hastalıklarını daha hassas olarak tedavi amacıyla derin öğrenme modeli olan Deep Transfer Learning (DFDD) modelini önerdiler. Çünkü birazdan açıklanacak olan denedikleri diğer modeller yetersiz veriden DFDD’ye göre daha çok etkilenmişlerdir. DFDD ile %97,96, DNNV (Deep Neural Network Verification) ile %74,74, DNNP (Deep Neural Network Prediction) ile %91,54 ortalama doğruluk değerleri elde etmişlerdir. Derin transfer öğrenimi yaklaşımı kullanılarak eğitilmiş tanı modeli hastanın izlenmesi için kullanılır. Ancak bu teknoloji hala gelişim aşamasında olduğundan bu yaklaşımla oluşturulan modeller hızla değişen veri durumları için uygun olmayabilir [14]. Jun Zhang ve arkadaşları sağlıkta kullanılan dijital ikizin güvenlik açıklarını önleme amaçlı siber dayanıklılık temelli zafiyet tespiti yapmak için LSTM’in çift yönlü çeşidi olan Bi-LSTM ile kendi oluşturdukları DeepVR algoritmalarını ve C / C++ ile yazılmış olan güvenlik açığı tespiti bulma amaçlı kullanılan Flawinder uygulamasını, hazır bulunan SARD veri kümesi ve açık kaynak hazır veri kümeleri ile precision, recall, f1 score kullanarak karşılaştırmışlardır. Açık kaynaklı veri kümesinde Bi-LSTM ile %61 precision, %63 recall, %62 f1 score; DeepVR ile %79 precision, %77 recall, %78 f1 score sonuçları; Flawinder ile %30 precision, %28 recall, %29 f1 score elde edilmiştir. SARD veri kümesinde Bi-LSTM ile %92 precision, %100 recall, %96 f1 score; DeepVR ile %99 precision, %97 recall, %96 f1 score; Flawinder ile %59 precision, %52 recall, %56 f1 score sonuçları elde edilmiştir. Bu çalışmada kendi kendine dikkat mekanizmasına sahip ve veriler arasında çift yönlü ilişki avantajından ötürü makale

sonunda Bi-LSTM modeli önerilmiştir [15]. Subramanian ve arkadaşları güncel olarak kişinin kendisinin dijital ikizini bir web kamerası kullanarak anlık görüntüleyen bir duygu tanıma (ER) sistemi önermişlerdir. Bu sistemi makine öğrenmesi ve derin öğrenme yöntemlerini kullanarak oluşturmuşlardır. Sistemde algoritma olarak ilerleyen cümlede sıralanan algoritmalar ve doğruluk metriği olarak accuracy kullanmışlardır. Lojistik Regresyon ile %99,1, Random Forest ile %99,6, Gradient Boosting ile %99,9, KNN ile %99,7, SVM ile %98,1, Desicion Tree Classifer ile %98,6, Naive Bayes ile %76,5, Ridge Regresyon ile %98,1 doğruluk değerlerini elde etmişlerdir. Bu denenen algoritmalardan en yüksek sonucu vermesinden ötürü Gradient Boosting algoritmasını önermişlerdir [16].

Sonuç olarak, yukarıdaki paragraf sağlık sektöründe dijital ikizlerin kullanımına yönelik bazı araştırmaları özetlemektedir. Sağlığın teknolojiyle beraber daha iyi duruma gelmesiyle hastaların tedavi süreçlerinin kolaylaşması, sağlık hizmetlerinin iyileştirilmesi bahsi geçen çalışmaların amacıdır. Fakat veri güvenliği, gizliliği, maliyeti ve karmaşıklığı bu teknolojide hala tam olarak çözülemeyen bir konudur. Bu nedenle bu alana olan ilgi fazlalaştırılmalıdır. Yapılan ve yapılacak ilerlemelerle sağlık sektörünün daha iyi hizmet sunma kapasitesi artırılarak hasta deneyimleri iyileştirilebilir.

Yukarıdaki metinsel ifadenin kavranabilmesi adına aşağıda tablo gösterimi bulunmaktadır.

Tablo 1: Yapılan Çalışmalar ile Kullanılan Algoritma ve Metrikler Tablosu

Çalışmalar ve Gerçekleştirenler	Algoritmalar	Accuracy	Sensitivity	Specificity	Precision	AUC
İnme tespiti (Hussain ve arkadaşları)	SVM (Support Vector Machine)	76	73	79	77	84
Kardiyovasküler risk (Martinez-Velazquez ve arkadaşları)	CNN modeli	85,81	NaN	83,87	86,29	NaN
Hastalıkların daha hassas tedavisi (Xu ve arkadaşları)	DFDD	97,96	NaN	NaN	NaN	NaN
Siber dayanıklılık temelli veri güvenliği sağlanması (Jun Zhang ve arkadaşları)	Bi-LSTM	61	63	62	NaN	NaN
Siber dayanıklılık temelli veri güvenliği sağlanması (Jun Zhang ve arkadaşları)	DeepVR	79	77	78	NaN	NaN
Siber dayanıklılık temelli veri güvenliği sağlanması (Jun Zhang ve arkadaşları)	Flawinder	30	28	29	NaN	NaN
Siber dayanıklılık temelli veri güvenliği sağlanması (Jun Zhang ve arkadaşları)	Bi-LSTM (SARD)	92	100	96	NaN	NaN
Siber dayanıklılık temelli veri güvenliği sağlanması (Jun Zhang ve arkadaşları)	DeepVR (SARD)	99	97	96	NaN	NaN
Siber dayanıklılık temelli veri güvenliği sağlanması (Jun Zhang ve arkadaşları)	Flawinder (SARD)	59	52	56	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	Lojistik Regresyon	99,1	NaN	NaN	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	Random Forest	99,6	NaN	NaN	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	Gradient Boosting	99,9	NaN	NaN	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	KNN	99,7	NaN	NaN	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	SVM	98,1	NaN	NaN	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	Decision Tree Classifier	98,6	NaN	NaN	NaN	NaN

Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	Naive Bayes	76,5	NaN	NaN	NaN	NaN
Web kamerasıyla kişinin anlık dijital ikiz görüntülenmesi (Subramanian ve arkadaşları)	Ridge Regresyon	98,1	NaN	NaN	NaN	NaN

Bu konuya ilişkin çalışmalar yapan girişim ve şirketlere örnek olarak; “Philips Kalbi simüle ederek doktorların tedavi süreci başlamadan bu doğrultuda karar verebilmelerini kolaylaştırmak amacıyla hasta verilerini dijital ortama ultrason görüntülerini yapay zeka algoritmalarıyla eğitmiş ve kalbin anatomik yapısını buna göre oluşturmuştur [17]. Siemens kişiye özel tedavi yöntemleri, hastalıkların ne gibi durum ve faktörlerde ortaya çıktığı veya mevcut hasta olan insanların hastalıklarından nasıl en kısa sürede kurtularak normal yaşamlarına geri dönebilecekleri amacıyla kişinin dijital ikizini oluşturmayı hedeflemekte, buna yönelik çalışmalarını sürdürmektedir [18]. Fransız girişim olan Nurea kardiyovasküler hastalıkların tanısı için yapay zeka algoritmaları kullanarak cerrahların tanı kararlarını daha erken ve etkili verebilmelerini sağlamak için 3d görüntüleme teknikleriyle dijital ikizi kullanır [19]. Amerikan girişimi olan Predictiv bireylerin DNA'sını kullanarak bireyin genetik risk değerlendirmelerini dijital ikiz teknolojileriyle açıklayarak her birey için kişisel sağlık planlaması hazırlar [20]. Alman startup olan Virtonomy.io, hayvan ve insanlar için istatistik, veri analizi, simülasyon teknikleriyle 3 boyutlu görüntülerini sanal ortama aktarıp bu verilerin dijital ikizini oluşturur [21].

Her ne kadar farklı alanlarda çalışıyorlar gibi görünseler de yukarıdakiler ve onlar gibi şirket ve girişimlerin ortak amacı dijital ikizin sağlıkta kullanılmasını yaygınlaştırmak bu alana olan ilgi ve dikkati arttırarak yapılan çalışma sayılarını fazlalaştırmaktır” gösterilebilir.

Ayrıca yapılan literatür araştırmasına ve bu araştırmadan yola çıkılarak çalışmada kullanılması uygun olduğu düşünülen algoritmalardan makalede yer verilecek olan algoritmalar aşağıda açıklanacaktır.

Gradient Boosting Regresyon

Gradient Boosting Regresyon, veriler arasındaki bağlantıları kavrayıp hataları düzelterek ilerleyip olabilecek en doğru sonucu vermektir. Bu model diğer zayıf tahmin algoritmalarının birleştirilmesiyle (örneğin karar ağaçları) daha güçlü bir algoritma oluşturmayı amaçlar [22].

Support Vector Machine (SVM)

SVM sınıflandırma ve regresyon modeli olarak makine öğrenmesinde sıkça kullanılan bir algoritma olup verileri kategorilere ayırmak veya bir değeri tahmin etmek amacıyla kullanılır ayrıca SVM doğrusal ve doğrusal olmayan problemler üzerinde durarak genellikle büyük verilerle kullanılır [23].

K Nearest Neighbors (KNN)

K-En Yakın Komşu (K-Nearest Neighbors veya KNN), sınıflandırma ve regresyon problemleri için veri madenciliği ve makine öğrenmesi alanlarında kullanılır. Bu algoritma genel olarak veri noktalarının benzerlikleri üzerine kurulu olup bu noktaların uzaydaki yönlerine göre gruplandırılmıştır [24].

Lojistik Regresyon

Genelde ikili sınıflandırmada (örneğin 0 ile 1) kullanılan bir istatistiksel modeldir. Bağımsız değişkenler girdi olarak verilip bağımlı değişkenlerin tahmin edilmesi yöntemiyle çalışır.

Aşağıdaki denklem lojistik regresyonun matematiksel ifadesini temsil etmektedir [25].

$$z=w_0x_0+w_1x_1+w_2x_2+\cdots+w_nx_n$$

Random Forest Classifier

Random Forest Classifier, makine öğrenmesinde sınıflandırma problemleri için oluşturulan ensemble bir algoritmadır. Ensemble algoritmalar birden çok modelin (örneğin karar ağaçları) birleştirilmesi ile daha doğru tahminler yapabilmek için aşırı öğrenme (overfitting) az öğrenme (underfitting) gibi durumlara dirençli tasarlanmıştır. [26].

Naive Bayes

Temel amacı veri analizi sayesinde verileri gruplandırmak olup makine öğrenmesinde sınıflandırma problemlerini çözmek için kullanılan istatistiksel bir modeldir. Algoritma matematikteki naive bayes mantığından esinlenilerek oluşturulmuştur.[27].

Multilayer Perceptron (MLP)

MLP (Multilayer Perceptron), derin öğrenmede çok sık kullanılan bir yapay sinir ağı modelidir. Tipik olarak giriş katmanı, gizli katman, çıkış katmanı adı verilen 3 katmandan oluşur. Gizli katman birden fazla sayıda olabilir. Bu fazlalık modelin karmaşık problemlerdeki başarı performansını yükseltir [28].

Extreme Gradient Boosting (XGBoost)

Temel olarak zayıf tahmin algoritmalarının (örneğin karar ağaçları) birleştirerek güçlü bir tahmin algoritması oluşturmayı amaçlar. Sınıflandırma ve tahmin algoritması olarak makine öğreniminde kullanılan Gradient Boosting yöntemlerinin gelişmiş versiyonudur [29].

Artificial Neural Network (ANN)

ANN (Artificial Neural Network), insan beyninden esinlenerek oluşturulan bir yapay sinir ağı modelidir. Girdi verilerini alarak bu veriler üzerinden birtakım matematiksel işlemler yapar ve çıktıları üretir [30].

Long Short-Term Memory (LSTM)

LSTM zaman içindeki bilgileri saklayabilen bir bellek mekanizması sağlayarak uzun vadeli bağımlılıkları koruyabilen bir yapıya sahiptir. Bu anlamda da geleneksel yapay sinir ağı yapılarından farklıdır. LTSM kullanılmayan modeller zaman içindeki ilişkileri yakalamakta zorlanabileceği veya yanıltıcı sonuçlar üretebileceği için bu tür modellerde LTSM kullanılması mantıklı sonuçlar almaya yardımcı olur [31].

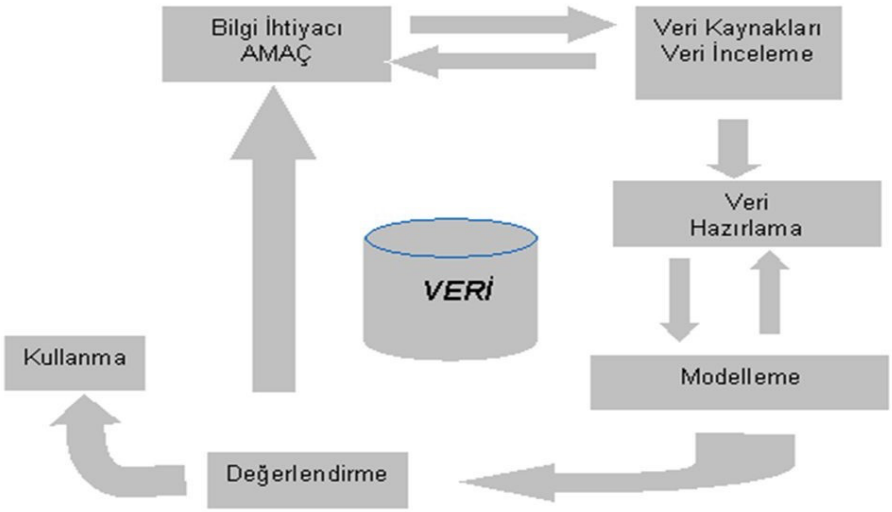
Decision Tree

Hem sınıflandırma hem regresyon problemlerinin çözümü için kullanılan Decision Tree veri kümesindeki veri noktalarını sınıflandırmak, değer tahmini yapmak için kullanılan ağaca benzer bir algoritmadır. Her iç düğüm belirli bir özniteliği temsil ederken yaprak düğümleri ise sonuç veya tahminleri temsil eder. Düşük hesaplama karmaşıklığı hem sayısal hem sınıfsal verilerle kullanılabilmesi, hızlı veri ön işleme, kolay anlaşılabilirlik gibi avantajlarının yanında aşırı öğrenme, model karmaşıklığı, veri dengesizliği gibi dezavantajlara da sahiptir. [32].

Bu algoritmalar çeşitli yöntemler ve parametreler kullanılarak veya mevcut parametreler üzerinde değişiklik yapılarak (çapraz doğrulama, epoch değerleri, train ve test verilerinin dağılım oranları, optimizer ve loss fonksiyonları seçimi, kullanılan doğruluk metrikleri gibi) olabilecekleri en iyi hale deney, gözlem yoluyla getirildi.

2 METODOLOJİ

Veri biliminde kullanılan 3 adet yaklaşım vardır: CRISP – DM (Cross-Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Databases) ve SEMMA (Sample, Explore, Modify, Model, Assess). KDD ve SEMMA daha çok büyük veri projelerinde kullanılırken CRISP – DM ise veri analizi ve veri madenciliği projelerinde kullanılır [33]. Bu projenin kapsamı incelendiğinde proje ihtiyaçlarından ötürü CRISP – DM yaklaşımının kullanılması kararlaştırılmıştır.



Şekil-1: CRISP – DM Yaklaşımının aşamalı şema gösterimi

Şeki-1 de görüldüğü üzere CRISP – DM Yöntemi Veri Kaynakları ve Veri İnceleme, Veri Hazırlama, Modelleme, Değerlendirme, Kullanma olarak 5 adımlı bir süreç olarak karşımıza çıkmaktadır [34].

3 PROJE UYGULANMASI

Bu adımda veri kümeleri tablolaştırılıp “Ortalama, Standart Sapma, En küçük değer, %25 (Alt çeyrek), %50 (Medyan), %75 (Üst çeyrek), En büyük değer” sütunlarına bağlı olarak incelenecektir.

3.1 Veri Kaynakları ve Veri İnceleme

Projenin genel amacı ve iş gereksinimleri açıkça anlaşılır. Beklenen sonuçlar ve projeden elde edilmesi beklenen değer belirlenir. Bu adım, projenin başlangıcında bir çerçeve oluşturarak, veri bilimcilerin ve projenin paydaşlarının projeyi nasıl yönlendireceğini anlamalarına yardımcı olur [35].

Projenin kodlama aşamasından önce projenin asıl hedefi olan dijital ikizin sağlıktaki makine, öğrenimi ve derin öğrenme modellerinde kullanımı hakkında önceden yazılmış makaleler, konu hakkında yazı yazılmış internet siteleri detaylıca incelenmiştir. Yapılan bu araştırmalar sonucu sağlıkta dijital ikizin kullanımıyla ilgili Support Vector Machine (SVM), K Nearest Neighbors (KNN), Lojistik Regresyon, Random Forest Classifier, Naive Bayes, Multilayer Perceptron (MLP), Gradient Boosting Regresyon, Extreme Gradient Boosting (XGBoost), Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Decision Tree gibi makine öğrenmesi ve derin öğrenme modellerin uygulandığı görülmüştür. Ayrıca literatürde kullanılan modellerle ilgili kodlar incelenerek yapılacak olan bu projenin temel amacına katkıda bulunulmuştur.

3.2 Veri Hazırlama

Projede kullanılacak veri kaynaklarını ve veri kümesini anlamak için ayrılmış bir süreçtir. Bu adım, projenin veriye dayalı analizlerine temel oluşturarak, projenin geri kalan kısımlarında daha iyi bir şekilde veriye yaklaşmayı sağlar. Bu

aşamada, veri bilimciler projenin gereksinimlerini karşılamak için hangi veri kaynaklarının kullanılacağını belirlerler. Bu kaynaklar genellikle işletme sistemlerinden, veri tabanlarından veya dış kaynaklardan gelir [35].

Bu projede yukarıda da bahsedildiği üzere bu şartlara uygun veri kümeleri olarak “Diabetes Health Indicators Dataset” [36] ve “Sleep Health and Lifestyle Dataset”[37] veri kümeleri kullanılmıştır. İlk veri kümesi 253.681 veri satırı ve 10 adet sütun değişkeni, ikinci veri kümesi ise 375 veri satırı ve 13 veri sütunu içermektedir. tablo 1 ve tablo 2 de veri kümelerinin “ortalama, standart sapma, en küçük değer, %25 (alt çeyrek), %50 (medyan), %75 (üst çeyrek), en büyük değer” şeklinde rakamsal detaylı bilgileri verilmiştir.

Tablo 2: Diyabet verileri

Diyabet veri sütunları	Ortalama	Standart Sapma	En küçük değer	%25 (Alt çeyrek)	%50 (Medyan)	%75 (Üst çeyrek)	En büyük değer
Diabetes_012 (Diyabet)	0,29	0,69	0	0	0	0	2
HighBP (Yüksek Tansiyon)	0,42	0,49	0	0	0	1	1
HighChol (Yüksek Kolestrerol)	0,42	0,49	0	0	0	1	1
CholCheck (Kolesterol Kontrolü)	0,96	0,18	0	1	1	1	1
BMI (Beden Kitle İndeksi)	28,38	6,60	12	24	27	31	98
Smoker (Sigara İçen)	0,44	0,49	0	0	0	1	1
Stroke (İnme)	0,04	0,19	0	0	0	0	1
HeartDiseaseorAttack (Kalp Hatalığı veya Atak)	0,09	0,29	0	0	0	0	0
PhysActivity (Fiziksel Aktivite)	0,75	0,42	0	1	1	1	1
Fruits (Meyveler)	0,63	0,48	0	0	1	1	1
Veggies (Sebzeler)	0,81	0,39	0	1	1	1	1
HvyAlcoholConsump (Ağır Alkol Tüketimi)	0,05	0,23	0	0	0	0	1
AnyHealthCare (Herhangi Sağlık Hizmeti)	0,95	0,21	0	1	1	1	1

NoDocbcCost (Doktor Maliyeti)	0,08	0,27	0	0	0	0	1
GenHlth (Gen Sağlığı)	2,51	1,06	1	2	2	3	5
MentHlth (Ruh Sağlığı)	3,18	7,41	0	0	0	2	30
PhysHlth (Fiziksel Sağlık)	4,24	8,71	0	0	0	3	30
DiffWalk (Diferansiyel Yürüyüş)	0,16	0,37	0	0	0	0	1
Sex (Cinsiyet)	0,44	0,49	0	0	0	1	1
Age (Yaş)	8,03	3,05	1	6	8	10	13
Education (Eğitim)	5,05	0,98	1	4	5	6	6
Income (Gelir)	6,05	2,07	1	5	7	8	8

Şekilde verilen tabloda birinci veri kümesine ait 8 sütun diyabeti etkileyen faktörler, veri ortalamaları, verilerin standart sapması, veri kümesindeki en küçük değer, veri kümesi küçükten büyüğe doğru sıralandığında alt çeyrek bölümü (%25), veri kümesi küçükten büyüğe doğru sıralandığında veri kümesinin yarısı (%50), veri kümesi küçükten büyüğe doğru sıralandığında verilerin üst çeyrek bölümü (%75) (%25 olan kısım, verilerin baştan %25 ‘lik kısmı olarak düşünülürse %75 olan kısım, verilerin sondan %25 ‘lik kısmı gibi düşünülebilir.) ve veri kümesindeki en büyük değer olmak üzere diyabet durumuna, yüksek tansiyon durumuna, Yüksek kolesterol durumuna, kolesterol kontrol durumuna, beden kitle indeksine, sigara içme durumuna, inme durumuna, kalp hastalığı veya atak durumuna, fiziksel aktivite durumuna, meyve yeme durumuna, sebze yeme durumuna, alkol tüketim durumuna, herhangi bir sağlık hizmeti alıp almama durumuna, doktor kontrol maliyet durumuna, gen sağlık durumuna, ruh sağlık durumuna, fiziksel sağlık durumuna, diferansiyel yürüyüş durumuna, cinsiyete yaşa, eğitim durumuna, gelir durumuna göre verilmiştir.

Tablo 3: Uyku verileri

Uyku veri sütunları	Ortalama	Standart Sapma	En küçük değer	%25 (Alt çeyrek)	%50 (Medyan)	%75 (Üst çeyrek)	En büyük değer
Person ID (Kişi numarası)	187,5	108,10	1	94,25	187,5	280,75	374
Age (Yaş)	42,18	8,67	27	35,25	43	50	59
Sleep Duration (Uyku süresi)	7,13	0,79	5,8	6,4	7,2	7,8	8,5
Quality of Sleep (Uyku kalitesi)	7,31	1,19	4	6	7	8	9
Physical Activity Level (Fiziksel	59,17	20,83	30	45	60	75	90

aktivite seviyesi)							
Stress Level (Stres seviyesi)	5,38	1,7	3	4	5	7	8
Heart Rate (Kalp atış hızı)	70,16	4,13	65	68	70	72	86
Daily Steps (Günlük adım)	6816,84	1617,91	3000	5600	7000	8000	10000

Şekilde verilen tabloda ikinci veri kümesine ait 8 sütun uykuyu etkileyen faktörler, veri ortalamaları, verilerin standart sapması, veri kümesindeki en küçük değer, veri kümesi küçükten büyüğe doğru sıralandığında alt çeyrek bölümü (%25), veri kümesi küçükten büyüğe doğru sıralandığında veri kümesinin yarısı (%50), veri kümesi küçükten büyüğe doğru sıralandığında verilerin üst çeyrek bölümü (%75) (%25 olan kısım, verilerin baştan %25 ‘lik kısmı olarak düşünülürse %75 olan kısım, verilerin sondan %25 ‘lik kısmı gibi düşünülebilir.) ve veri kümesindeki en büyük değer olmak üzere kişiye, yaşa, uyku süresine, uyku kalitesine, fiziksel aktivite seviyesine, stres seviyesine, kalp atış hızına, günlük adım sayısına bağlı olarak verilmiştir.

Buna karşın verilerin makine öğrenimi ile derin öğrenme algoritmaları kullanılacak kısımları, kullanılan 2 veri kümesindeki ortak ve birbiriyle ilişkili olduğu düşünülen sütunlar tespit edilerek bu konumlardaki değerler int ve float tipinde verildiğinden ötürü herhangi bir ayıklama veya dönüştürme işlemi yapılmadan projede incelenecek ilgili sütunlar birleştirilerek aşağıdaki tabloda verilmiştir.

Tablo 4: Birleştirilmiş veriler

Birleştirilmiş Veriler	Ortalama	Standart Sapma	En Küçük	25%	50%	75%	En Büyük	Veri Tipleri
BMI (Beden Kitle İndeksi)	28,65	5,83	16	24	28	32	55	float64
Age (Yaş)	8,85	2,62	1	7	9	11	13	float64
PyhsActivity (Fiziksel Aktivite)	0,58	0,49	0	0	1	1	1	float64
Fruits (Meyveler)	0,55	0,49	0	0	1	1	1	float64
Veggies (Sebzeler)	0,75	0,42	0	1	1	1	1	float64
Diabetes_012 (Diyabet)	0,47	0,84	0	0	0	1	2	float64
Sleep Duration (Uyku Süresi)	7,13	0,79	5,8	6,4	7,2	7,8	8,5	float64
Physical Activity Level (Fiziksel Aktivite Düzeyi)	59,17	20,83	30	45	60	75	90	int64

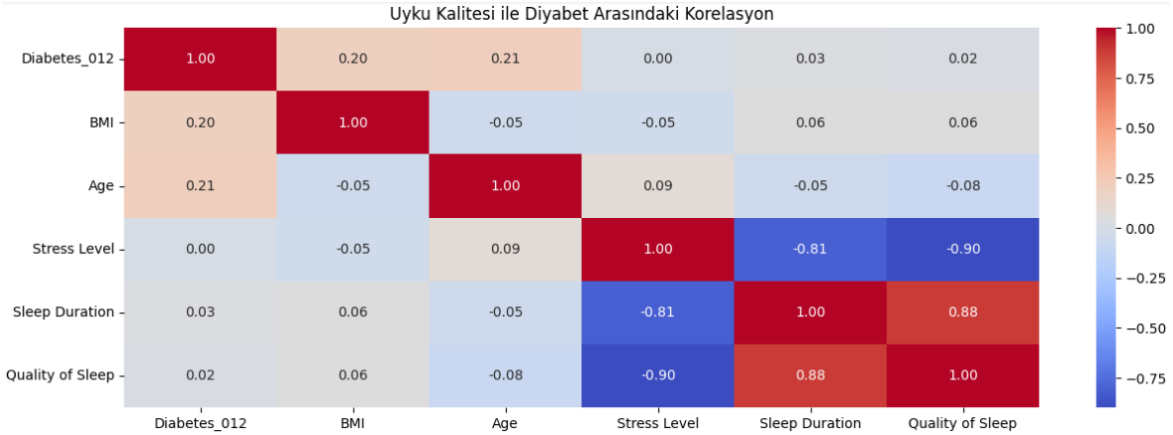
Stress Level (Stres Düzeyi)	5,38	1,77	3	4	5	7	8	int64
Quality of Sleep (Uyku Kalitesi)	7,31	1,19	4	6	7	8	9	int64

Şekilde verilen tabloda tablo2 ve tablo3 ‘teki veri kümelerine ait 9 sütun birleştirilmiş veriler, veri ortalamaları, verilerin standart sapması, veri kümesindeki en küçük değer, veri kümesi küçükten büyüğe doğru sıralandığında alt çeyrek bölümü (%25), veri kümesi küçükten büyüğe doğru sıralandığında veri kümesinin yarısı (%50), veri kümesi küçükten büyüğe doğru sıralandığında verilerin üst çeyrek bölümü (%75) (%25 olan kısım, verilerin baştan %25 ‘lik kısmı olarak düşünülürse %75 olan kısım, verilerin sondan %25 ‘lik kısmı gibi düşünülebilir.) ve veri kümesindeki en büyük değer, birleştirilmiş verilerin veri tipleri olmak üzere verilmiştir.

Tablo 5: Veriler arasındaki ilişkinin korelasyon tablosu

Özellikler	Sleep Duration (Uyku Süresi)	Stress Level (Stres Seviyesi)	Quality of Sleep (Uyku Kalitesi)	Diabetes_012 (Diyabet)	BMI (Beden Kitle İndeksi)	Age (Yaş)
Sleep Duration (Uyku Süresi)	1	-0,81	0,88	0,03	0,05	-0.04
Stress Level (Stres Seviyesi)	-0,81	1	-0,89	0,001	-0,04	0.09
Quality of Sleep (Uyku Kalitesi)	0,88	-0,89	1	0,02	0,05	-0.08
Diabetes_012 (Diyabet)	0,03	0,001	0,02	1	0,19	0.21
BMI (Beden Kitle İndeksi)	0,05	-0,04	0,05	0,19	1	-0.04
Age (Yaş)	-0.04	0.09	-0.08	0.21	-0.04	1

Tablo 5 ‘te verilen korelasyon matrisinden anlaşılacağı üzere uyku süresi ile stres seviyesi arasında negatif güçlü bir ilişki, stres seviyesi ile uyku süresi ve uyku kalitesi arasında negatif güçlü bir ilişki, uyku kalitesi ile uyku süresi arasında pozitif güçlü bir ilişki, uyku kalitesi ile stres seviyesi arasında negatif güçlü bir ilişki olduğu; uyku süresi ile diyabet ve beden kitle indeksi arasında pozitif zayıf bir ilişki, stres seviyesi ile diyabet arasında pozitif zayıf bir ilişki, stres seviyesi ile beden kitle arasında negatif zayıf bir ilişki, uyku kalitesi ile diyabet ve beden kitle indeksi arasında pozitif zayıf bir ilişki, diyabet ile beden kitle indeksi arasında pozitif zayıf bir ilişki, yaş ile uyku süresi, uyku kalitesi, beden kitle indeksi arasında negatif zayıf bir ilişki, yaş ile stres seviyesi, diyabet arasında pozitif zayıf bir ilişki olduğu görülmektedir.



Şekil-2: Uyku kalitesi, Yaş, Beden Kitle İndeksi, Stres Seviyesi, Uyku Süresi ile Diyabet arasındaki ilişkiyi temsil eden korelasyon ısı grafiği

Tablo 5 ve şekil-2 bulunan sonuçları destekler niteliktedir.

Tüm bu inceleme, birleştirme, korelasyon işlemlerinden sonra veriler eğitim ve test kümeleri olarak ayrılmış olup modelleme aşamasına hazırlık yapılmıştır.

3.3 Modelleme

Bu adım veri analizi aşamasının merkezinde yer alır ve projenin amacına yönelik tahmin modelleri veya sınıflandırıcılar gibi analitik yöntemlerin oluşturulduğu aşamayı ifade eder. Bu adım, veri kümesinin kullanılabilir bilgiye dönüştürüldüğü ve sonuçların tahmin edildiği kritik bir aşamadır [35].

CRISP-DM adımlarından ilki olan ve bölüm 3.1’de açıklanan veri kaynakları ve veri inceleme aşamasında da bahsedildiği üzere yapılan detaylı araştırma ve incelemelerden sonra bu aşamada kullanılacak olan model ve algoritmalara karar verilmiştir. Makine öğrenmesinin tahmin algoritmalarından Gradient Boosting Regresyon; sınıflandırma algoritmalarından SVM, KNN, Lojistik Regresyon, Random Forest Classifier, Naive Bayes, XGBoost, Decision Tree; derin öğrenme algoritmalarından ANN, LSTM, MLP algoritmaları kullanılmıştır.

3.3.1Gradient Boosting Regresyon

Projede amaç uyku kalitesi, yaş, beden kitle indeksi, uyku süresi özelliklerinden hareketle diyabet ile uyku kalitesi, yaş, beden kitle indeksi, uyku süresi arasındaki ilişkiyi bulmaya çalışmaktır. Bu amaçla karmaşık veriler (bu projede 2 veri kümesi üzerinde çalışıldı) arasındaki yüksek tahmin ve performans göstermesinden dolayı Gradient Boosting Regresyon kullanılmıştır.

3.3.2Support Vector Machine (SVM)

Projedeki amaç doğrultusunda karmaşık veriler üzerinde (bu projede 2 veri kümesi üzerinde çalışıldı.) aşırı uyum kontrolü (overfitting) ve dengesiz veriler (herhangi bir veri kümesindeki örnek sayısı diğer veri kümesindeki örnek sayısından daha fazla veya daha az) üzerindeki işlevselliğinden dolayı tercih edilmiştir.

3.3.3K Nearest Neighbors (KNN)

KNN yakınlık ilkesini benimseyen bir algoritma olduğundan dolayı bu projede uyku kalitesi, yaş, beden kitle indeksi, uyku süresi ile diyabet arasındaki benzerlikler üzerinden sınıflandırma yapmak için kullanılmıştır.

3.3.4Lojistik Regresyon

Lojistik regresyon anlaşılır ve yorumlanabilir bir algoritma olduğundan ötürü ayrıca dengesiz (veri kümesindeki örnek sayısı diğer veri kümesindeki örnek sayısından daha fazla veya daha az) veri kümelerine olan uygunluğundan dolayı bu projede tercih edilmiştir.

3.3.5Random Forest Classifier

Bu projede veri kümeleri arasındaki karmaşıklıktan ötürü birden çok karar ağacının harmanlanmasıyla oluşan bu modelin proje gereksinimlerini karşılaması gerekçesiyle ve ayrıca aşırı öğrenme (overfitting) sorununa karşın kullanılması uygun görülmüştür.

3.3.6Naive Bayes

Naive bayes olası sonuçların olasılığına dayanan istatistiksel bir algoritmadır. Bu proje gereksinimine bağlı olarak hızlı ve basit olduğundan, düşük boyutlu veri kümelerinde iyi performans gösterdiğinden ve etkileşimini basit bir şekilde ele aldığından dolayı tercih edilmiştir.

3.3.7Multilayer Perceptron (MLP)

Projedeki veri kümelerinden yola çıkarak non-lineer (doğrusal olmayan) ilişkileri iyi çözümleyebilmesi, veri kümesindeki karmaşık ilişkileri modelleme yeteneği ve genel uygulanabilirliği ayrıca orta düzeyde veri boyutlarına ve karmaşıklığa sahip veri kümelerinde iyi performans gösterdiğinden dolayı bu projede kullanılmıştır.

3.3.8Extreme Gradient Boosting (XGBoost)

Bu projede aşırı öğreme (overfitting) ve az öğrenme (underfitting) dirençli olduğundan dengesiz (veri kümesindeki örnek sayısı diğer veri kümesindeki örnek sayısından daha fazla veya daha az) veri kümeleri gibi zorlu problemlerde iyi sonuçlar elde edilmesi mümkün olduğundan, yüksek performanslı olduğundan ve öznitelik önem sıralamalarını sağladığından dolayı kullanılmıştır.

3.3.9Decision Tree

Bu projede Decision Tree algoritması kullanılma nedeni, veri kümesinin sınıflandırma bir problem olduğu ve bu problemi çözmek için bir makine öğrenimi modeli gerektiği içindir. Decision Tree, sınıflandırma problemleri için oldukça yaygın bir şekilde kullanılan bir algoritmadır ve veri kümesindeki özniteliklerin sınıflandırma kararlarını açıkça görselleştiren bir ağaç yapısı oluşturur.

3.3.10 Artificial Neural Network (ANN)

Bu projede ANN modeli kullanım amacı, geniş veri kümelerini analiz etmek, özellikler arasındaki karmaşık ilişkileri yakalamak, sınıflandırma veya regresyon görevlerini gerçekleştirmek gibi işlevlerinden ve bu işlevler projedeki veri kümesine uygun olduğundan projede kullanılması uygun görülmüştür.

3.3.11 Long Short-Term Memory (LSTM)

Bu projede LSTM, karmaşık veri yapılarını öğrenme yeteneğine sahipliğinden, uzun vadeli bağımlılıklarını başarılı bir şekilde modelleyip ve tahmin edebildiğinden dolayı kullanılmıştır.

3.4 Değerlendirme

Bu adım oluşturulan modellerin ve analizlerin performansının değerlendirildiği kritik bir aşamadır. Bu adım, modelleme ve analiz aşamalarının ardından elde edilen sonuçların gerçek dünya performansını anlamayı sağlar. Değerlendirme aşamasında, oluşturulan modellerin iş gereksinimlerini ne kadar iyi karşıladığı, tahminlerin ne kadar doğru olduğu ve modelin genel performansı analiz edilir [35].

Kullanılan CRISP – DM metodolojisine göre model performanslarının değerlendirilmesi kısmında modelin ele alınan problem türüne göre değerlendime yöntemleri mevcuttur. Yapılan literatür taramasında regresyon modellerinde performans metriği olarak Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-Squared (R^2), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE) gibi metrikler; sınıflandırma modellerinde ise Accuracy (Doğruluk), Precision (Kesinlik), Recall (Hassasiyet, Geri Çağırma), F1-Score, Area Under the ROC Curve (AUC-ROC), Log Loss (Logaritmik Kayıp), Confusion Matrix (Karmaşıklık Matrisi) gibi metrikler kullanılmaktadır [38,39,40].

Bu projede bahsi geçen metrikler tek tek proje üzerinde denenerek en iyi sonuç verenlerin regresyon modelleri için MSE, MAE, RMSE; Sınıflandırma modelleri için Accuracy, Presicion, Recall olduğu görülmüş, dolayısıyla uygulanmıştır.

Tablo 6: Regresyon Modelleri – Performans Metrikleri

REGRESYON MODELLERİ	MSE	MAE	RMSE	MEAN MSE
GRADIENT BOOSTİNG REGRESYON	0.77	0.69	0.88	nan
CROSS – VAL GRADIENT BOOSTING REGRESYON	nan	nan	nan	0.91
ANN	0.61	0.63	0.78	nan
CROSS – VAL ANN (MEAN)	0.85	0.76	0.92	nan
LSTM	0.65	0.67	0.81	nan
CROSS – VAL LSTM (MEAN)	0.76	0.75	0.86	nan

Tablo-6 ‘da görüldüğü üzere regresyon modellerinden en iyi sonucu ANN modelinin verdiği görülmüştür.

Tablo 7: Sınıflandırma Modelleri – Performans Metrikleri

SINIFLANDIRMA MODELLERİ	ACCURACY	PRECİSİON	RECALL
SVM	0.79	0.63	0.79
CROSS-VAL SVM (MEAN)	0.73	nan	nan
KNN	0.70	0.68	0.70
MODEL ENSEMBLE KNN	0.69	0.67	0.69

LOGISTIC REGRESSION	0.80	0.83	0.80
RANDOM FOREST CLASSİFİER	0.71	0.68	0.71
NAİVE BAYES	0.79	0.77	0.79
MLP	0.79	0.63	0.79
CROSS-VAL MLP	0.67	0.83	0.80
XGBOOST	0.65	0.65	0.65
CROSS-VAL XGBOOST	0.59	0.58	0.59
DECISION TREE	0.63	0.67	0.63

Tablo-7 ‘de görüldüğü üzere Sınıflandırma modellerinden en iyi sonucu lojistik regresyon modelinin verdiği görülmüştür.

3.5 Modelin Çalıştırılması

Bu adım oluşturulan model veya analiz sonuçlarının gerçek dünya uygulamalarına dönüştürüldüğü ve değer sağladığı aşamayı ifade eder. Bu aşama, projenin hedeflenen kazançların veya sonuçların elde edilmesini amaçlar [35].

%80 üstü başarı sağlaması ve literatürde çokça tercih edilmesi sebebiyle bir sınıflandırma algoritması olan lojistik regresyon modelinin uyku kalitesi, yaş, beden kitle indeksi, stres seviyesi, uyku süresi gibi faktörlere bağlı diyabet verisini kullanılan algoritmalar içinde en iyi tahmin eden algoritma olduğu gözlemlenmiştir.

Bu algoritmanın başarısı kullanılan parametrelerin değiştirilmesi veya yaş, beden kitle indeksi faktörlerin belirli aralıklara sınırlandırılması, veri toplama gibi pek çok faktöre bağlıdır. Gelecek çalışmalarda bu gibi faktörler veya kullanılan algoritmalar değiştirilerek daha yüksek başarı değerleri çıkarılabilir ayrıca bu makalede kullanılan ve en doğru sonucu verdiği düşünülen lojistik regresyon algoritması da kullanılarak diyabet tespiti yapılabilir.

4 SONUÇ

Proje kapsamında sağlık ve diyabet verileri, CRISP – DM metodolojisi üzerinden veri ön işleme, analiz, görselleştirme ve çıkarım yapılmıştır. Projenin daha iyi kavranıp uygulanması amacıyla detaylı bir araştırma ve literatür taramasına gerek duyulmuş ve buna yönelik hareket edilmiştir. Yapılan araştırma sonuçlarından elde edilen bilgilere göre, regresyon algoritmalarından Gradient Boosting Regresyon, yapay sinir ağı olup aynı zamanda regresyon amacıyla kullanılan algoritmalardan ANN, MLP ve LSTM; sınıflandırma algoritmalarından SVM, KNN, Lojistik Regresyon, Random Forest Classifier, Naive Bayes, XGBoost, Decision Tree kullanılması uygun görülmüştür. Ayrıca yine literatürde bahsi geçen performans metriklerinden regresyon için MSE, MAE, RMSE; sınıflandırma için Accuracy, Precision, Recall metriklerinin kullanımı uygun görülmüştür.

Bu modeller hiperparametre ayarları, Cross Validation işlemleri, birden çok performans metriği uygulanması, yapay sinir ağı modellerinde epoch değerleri gibi birçok farklı deneme sürecinden sonra en iyi sonucu verecek şekilde makalede toplanmıştır.

Aşağıda makale bulgularının daha iyi algılanabilmesi adına bu makaledeki veri kümelerinin ayrı ayrı farklı makine öğrenmesi algoritmalarıyla kullanımlarında verdikleri sonuçlar ile makale özelinde bulunan sonuçların karşılaştırmalı tabloları (tablo 8, tablo 9) verilmiştir.

Tablo 8: Diyabet Veri Kümesi ile Uygulanan Algoritmaların Accuracy, Precision, Recall Değerleri Tablosu

Model	(Literatür) Accuracy - (Bulunan) Accuracy	(Literatür) Precision - (Bulunan) Precision	(Literatür) Recall - (Bulunan) Recall
SVM	0.88 - 0.79 [41]	0.87 - 0.63 [41]	0.81 - 0.79 [41]
Random Forest Classifier	0.77 - 0.71 [41]	0.76 - 0.68 [41]	0.77 - 0.71 [41]
Gaussian Naive Bayes	0.77 - 0.79 [41]	0.76 - 0.77 [41]	0.77 - 0.79 [41]
XGBoost	0.83 - 0.65 [42]	NaN – 0.65	NaN - 0.65
Logistic Regression	0.84 - 0.80 [42]	NaN – 0.83	NaN – 0.80
Decision Tree Classifier	0.78 - 0.63 [43]	0.72 – 0.67 [43]	0.76 – 0.63 [43]

Tablo 8’de makalede kullanılan diyabet veri kümesi üzerinde makine öğrenmesi kullanılarak literatürdeki sonuçlar ve makale özelinde bulunan sonuçların karşılaştırılmış tablosu verilmiştir.

Tablo 9: Uyku Veri Kümesi ile Uygulanan Algoritmaların Accuracy Değerleri Tablosu

Model	(Literatür) Accuracy - (Bulunan) Accuracy
SVM	0.84 - 0.79 [44]
Decision Tree Classifier	0.95 - 0.63 [45]
Gaussian Naive Bayes	0.75 - 0.79 [45]
KNN	0.90 - 0.70 [45]
Logistic Regression	0.92 - 0.80 [46]
XGBoost	0.82 - 0.65 [46]

Tablo 9’da makalede kullanılan uyku veri kümesi üzerinde makine öğrenmesi algoritmaları kullanılarak literatürdeki sonuçlar ve makale özelinde bulunan sonuçların karşılaştırılmış tablosu verilmiştir.

Tablo 8 ile tablo 9’da literatürde olan ve makalede bulunan sonuçlar arasındaki farkların nedeni makalede iki veri kümesinin (uyku ve diyabet) birlikte kullanılmasından ötürü olduğu ayrıca kullanılan sütunların da farklılığından ötürü olduğu öngörülmektedir.

Sağlıkta dijital ikiz, Hastalık tahmini, hastalık teşhisi, Tedavi ve ilaç optimizasyonu, kişiselleştirilmiş tıp, hastalık modelleri ve epidemiyoloji gibi amaçlarla kullanılmaktadır. Burada verilen amaçlardan bu makalede sağlıkta dijital ikizin hastalık teşhisi amacına bağlı olarak uyku kalitesi, yaş, beden kitle indeksi, stres seviyesi, uyku süresi gibi faktörlerle diyabet arasındaki ilişki incelenmiştir.

5 KAYNAKÇA

[1] KUMAŞ, Esra, and E. R. O. L. Serpil. "Endüstri 4.0’da anahtar teknoloji olarak dijital ikizler." Politeknik Dergisi 24.2 (2021): 691-701.

[2] Zerrin Ayvaz Reis, Nilgün Bozbuğa. “Dijital İkiz ve Tıbbi Sistemlerin Modellenmesi”. Cdn.istanbul.edu.tr .D158A1B0877147B7A575405823C27F40 (istanbul.edu.tr) . 23 Aralık 2022.

[3] Ayyüce kızrak. “Yolların Kesişimi: Dijital İkiz”. medium.com . Yolların Kesişimi: Dijital İkiz. “Hayal dünyası ile gerçek dünya… | by Ayyüce Kızrak, Ph.D. | Medium . 12 Kasım 2022.

[4] AYNACI, İffet. "DİJİTAL İKİZ VE SAĞLIK UYGULAMALARI." İzmir Kâtip Çelebi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 3.1 (2020): 70-79.

[5] Veranyurt, Ülkü, et al. "MAKİNE ÖĞRENMESİ TEKNİKLERİYLE HASTALIK SINIFLANDIRMASI: RANDOM FOREST, K-NEAREST NEIGHBOUR VE ADABOOST ALGORİTMALARI UYGULAMASI." Uluslararası Sağlık Yönetimi ve Stratejileri Araştırma Dergisi 6.2 (2020): 275-286.

[6] İsmail Gürbüz. “Dijital İkiz Nedir?”. medium.com . Dijital İkiz Nedir?. Ülkemizde yeni yeni popüler olmaya… | by İsmail GÜRBÜZ | Bi’ Dünya İçerik | Medium . 12 Aralık 2020.

[7] Croatti, Angelo, et al. "On the integration of agents and digital twins in healthcare." Journal of Medical Systems 44 (2020): 1-8.

[8] N. Wickramasinghe, P. P. Jayaraman, A. R. M. Forkan, N. Ulapane, R. Kaul, S. Vaughan, and J. Zelcer, “A vision for leveraging the concept of digital twins to support the provision of personalized cancer care,” IEEE Internet Computing, vol. 26, no. 5, pp. 17–24, 2021.

[9] Gillette, Karli, et al. "A framework for the generation of digital twins of cardiac electrophysiology from clinical 12-leads ECGs." Medical Image Analysis 71 (2021): 102080.

[10] Voigt, Isabel, et al. "Digital twins for multiple sclerosis." Frontiers in immunology 12 (2021): 669811.

[11] Chakshu, Neeraj Kavan, and Perumal Nithiarasu. "An AI based digital-twin for prioritising pneumonia patient treatment." Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine 236.11 (2022): 1662-1674.

[12] Hussain, Iqram, Md Azam Hossain, and Se-Jin Park. "A healthcare digital twin for diagnosis of stroke." 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON). IEEE, 2021.

[13] Martinez-Velazquez, Roberto, Rogelio Gamez, and Abdulmotaleb El Saddik. "Cardio Twin: A Digital Twin of the human heart running on the edge." 2019 IEEE international symposium on medical measurements and applications (MeMeA). IEEE, 2019.

[14] Xu, Yan, et al. "A digital-twin-assisted fault diagnosis using deep transfer learning." Ieee Access 7 (2019): 19990-19999.

[15] Zhang, Jun, et al. "Cyber resilience in healthcare digital twin on lung cancer." IEEE Access 8 (2020): 201900-201913.

[16] B. Subramanian, J. Kim, M. Maray, and A. Paul, “Digital twin model: A real-time emotion recognition system for personalized healthcare,” IEEE Access, vol. 10, pp. 81155–81165, 2022.

[17] Philips. “How a virtual heart could save your real one”. philips.com . How a virtual heart could save your real one - Blog | Philips . 12 Kasım 2018.

[18] Siemens. “What is a digital patient twin”. Siemens.com . What is a digital patient twin? (siemens-healthineers.com) . 21 Mart 2023.

[19] Nurea. “You are interested by our software for vascular diseases quantification ?”. Decision support software for vascular diseases (nurea-soft.com) . 2023.

[20] Predictiv. “DNA-based Digital Twin”. The Process | Predictiv (predictivcare.com) . 2023.

[21] Virtonomy.io. “End-to-end digital twin solution”. In silico trials with Virtonomy's product v-Patients - Virtonomy.io . 2022.

[22] Alex Rogozhnikov. “Gradient Boosting explained”. arogozhnikov.github.io. Gradient Boosting explained [demonstration] (arogozhnikov.github.io) . 24 Haziran 2016.

[23] Aytaç, Muhammed Bilgehan, and Hasan Şakir Bilge. "TELE PAZARLAMA VERİLERİNİN BİRLİKTELİK KURALLARIYLA VE CRISP-DM YÖNTEMİYLE ANALİZ EDİLMESİ." Aksaray Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 5.2 (2013): 25-40.

[24] Shang, Wenqian, et al. "An improved kNN algorithm–fuzzy kNN." Computational Intelligence and Security: International Conference, CIS 2005, Xi’an, China, December 15-19, 2005, Proceedings Part I. Springer Berlin Heidelberg, 2005.

[25] Zou, Xiaonan, et al. "Logistic regression model optimization and case analysis." 2019 IEEE 7th international conference on computer science and network technology (ICCSNT). IEEE, 2019.

[26] Rodriguez-Galiano, Victor Francisco, et al. "An assessment of the effectiveness of a random forest classifier for land-cover classification." ISPRS journal of photogrammetry and remote sensing 67 (2012): 93-104.

[27] Zhang, Harry. "The optimality of naive Bayes." Aa 1.2 (2004): 3.

[28] Taud, Hind, and J. F. Mas. "Multilayer perceptron (MLP)." Geomatic approaches for modeling land change scenarios (2018): 451-455.

[29] Chen, Minghua, et al. "XGBoost tabanlı algoritma yorumlaması ve güç sisteminin arıza sonrası geçici kararlılık durumu tahmini üzerine uygulama." IEEE Erişim 7 (2019): 13149-13158.

[30] Agatonovic-Kustrin, S., and Rosemary Beresford. "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research." Journal of pharmaceutical and biomedical analysis 22.5 (2000): 717-727.

[31] Zhu, Yu, et al. "What to Do Next: Modeling User Behaviors by Time-LSTM." IJCAI. Vol. 17. 2017.

[32] Şadi Evren Şeker. “Karar Ağacı Öğrenmesi (decision tree learning)”. bilgisayarkavramlari.com . Karar Ağacı Öğrenmesi (decision tree learning) – Bilgisayar Kavramları (bilgisayarkavramlari.com). 11 Nisan 2012.

[33] Shafique, Umair, and Haseeb Qaiser. "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)." International Journal of Innovation and Scientific Research 12.1 (2014): 217-222.

[34] Şeker, Şadi Evren. "CRISP-DM: Endüstriler Arası Standart İşleme–Veri Madenciliği için (Cross Industry Standard Processing–Data Mining)." YBS Ansiklopedi 5.2 (2018).

[35] Dolgun, Muhsin Özgür, and Derya Ersel. "Doğrudan pazarlama stratejilerinin belirlenmesinde veri madenciliği yöntemlerinin kullanımı." İstatistikçiler Dergisi: İstatistik ve Aktüerya 7.1 (2014): 1-13.

[36] Alex Teboul. “Diabetes Health Indicators Dataset”. kaggle.com. Diabetes Health Indicators Dataset | Kaggle .

[37] Laksika Tharmalingam. “Sleep Health and Lifestyle Dataset”. kaggle.com. Sleep Health and Lifestyle Dataset | Kaggle .

[38] Syafrudin, Muhammad, et al. "Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing." Sensors 18.9 (2018): 2946.

[39] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science 7 (2021): e623.

[40] Li, Mingqi, Xiaoyang Fu, and Dongdong Li. "Diabetes prediction based on XGBoost algorithm." IOP conference series: materials science and engineering. Vol. 768. No. 7. IOP Publishing, 2020.

[41] HARMAN, Güneş. "Destek vektör makineleri ve naive bayes sınıflandırma algoritmalarını kullanarak diabetes mellitus tahmini." Avrupa Bilim ve Teknoloji Dergisi 32 (2021): 7-13.

[42] Korkmaz, Merve, and Kaplan Kaplan. "Şeker hastalığı teşhisi ve önerilen modellerinin karşılaştırılması." Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi 12.1 (2023): 1-1.

[43] CİHAN, Pınar, and Hakan COŞKUN. "Diyabet Tahmini için Makine Öğrenmesi Modellerinin Performans Karşılaştırılması Performance Comparison of Machine Learning Models for Diabetes Prediction."

[44] Nazli, Bahar. "Evaluation of different machine learning algorithms for classification of sleep apnea." 2021 29th Signal Processing and Communications Applications Conference (SIU). IEEE, 2021.

[45] Memiş, Gökhan, and Mustafa Sert. "Classification of Obstructive Sleep Apnea using Multimodal and Sigma-based Feature Representation." 2019 27th Signal Processing and Communications Applications Conference (SIU). IEEE, 2019.

[46] KARADÖL, İsrafil. "OBSTRÜKTİF UYKU APNESİ TESPİTİNDE POLİSOMNOGRAFİYE ALTERNATİF YENİ YÖNTEMLER." Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi 26.1 (2023): 295-307.