Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Araştırma Makalesi / Research Article

Anne Sağlığı Riski İçin Makine Öğrenmesi Modellerinin Performans Karşılaştırması

Performance Comparison Of Machine Learning Models For Maternal Health Risk

Fatih Şahin[1*], Şeyma Karaca[2], Gökalp Tulum[3]

[1] İstanbul Topkapı Üniversitesi, Yazılım Mühendisliği Bölümü, fatihsahin@topkapi.edu.tr
ORCID: https://orcid.org/0000-0002-8036-3156

[2] Üsküdar Üniversitesi , Fen Bilimleri Enstitüsü, seymakaraca1997@gmail.com
ORCID: https://orcid.org/ 0000-0002-7790-7896

[3] İstanbul Topkapı Üniversitesi, Elektrik Elektronik Mühendisliği Bölümü, gokalptulum@topkapi.edu.tr
ORCID :https://orcid.org/0000-0003-1906-0401

MAKALE BİLGİLERİ

Makale Geçmişi:

Geliş 10 Temmuz 2023
Revizyon 4 Ekim 2023
Kabul 25 Ekim2023
Online 31 Aralık 2023

Anahtar Kelimeler:

Anne Ölümleri, Anne Sağlığı Riski Tahmini, Makine Öğrenmesi, Sınıflandırma Algoritmaları Karşılaştırması, Yapay Zekâ

ÖZ

Sağlık sektöründe hastalıkların teşhisi için yapay zekânın alt dallarından olan makine öğrenmesi oldukça yaygın kullanılmaktadır. Çalışmada anne sağlığı riski üzerine bir veri seti kullanılarak hamilelikte risk üzerine sınıflandırma çalışması yapılması amaçlanmıştır. Çalışmada, makine öğrenmesi algoritmalarından lineer regresyon, destek vektör makineleri, karar ağacı algoritması, rastgele orman algoritması, çok katmanlı algılayıcı, naive bayes sınıflandırıcısı, k-en yakın komşu algoritması ve XGBoost sınıflandırıcısı kullanılmıştır. Aynı zamanda her bir algoritma için temel bileşenler analizi (PCA) ve lineer diskriminant analizi (LDA) uygulanmış olup oluşturulan modellerin doğruluk oranlarına bakılarak tahminde bulunulmuştur. Yapılan tahmin sonucunda en yüksek doğruluk oranı %84 ile rastgele orman algoritmasından, PCA dönüşümü kullanılarak yapılan tahmin sonucuna göre en yüksek doğruluk oranı %82 ile rastgele orman algoritması ve K-en yakın komşu algoritmasından ve LDA dönüşümü kullanılarak yapılan tahmin sonucuna göre ise en yüksek doğruluk oranı %85 ile karar ağacı algoritması ve K-en yakın komşu algoritmasından elde edilmiştir. Sınıflandırma işleminde LDA dönüşümünün daha yüksek sonuç elde ettiği görülmektedir.

ARTICLE INFO

Article history:

Received 10 July 2023
Received in revised form 4 October 2023
Accepted 25 October 2023
Available online 31 December 2023

Keywords:

Maternal Mortality, Maternal Health Risk Estimation, Machine Learning, Classification Algorithms Comparison, Artificial Intelligence)

Doi: 10.24012/dumf.1325431

ABSTRACT

Machine learning, one of the sub-branches of artificial intelligence, is widely used in the health science for the diagnosis of diseases. In this study, it was aimed to perform a classification study on risk in pregnancy using a data set on maternal health risk. In this study, linear regression, support vector machines, decision tree algorithm, random forest algorithm, multilayer perceptron, naive bayes classifier, k-nearest neighbor algorithm and XGBoost classifier were used as a classifier. At the same time, principal components analysis (PCA) and linear discriminant analysis (LDA) were applied for each algorithm, and predictions were calculated as the accuracy of the models. As a result of the classification, the highest accuracy was 84% for the random forest algorithm. With PCA transform, the highest accuracy was calculated 82% for the random forest algorithm and K-nearest neighbor algorithm. The best accuracy for LDA transform was obtained 85% for decision tree algorithm and K-nearest neighbor algorithm. It is seen that the LDA transformation achieves higher accuracy for the classification process.

* Sorumlu Yazar

Giriş

Her kadının dünyaya yeni bir hayat getirdiğinde hissettiği mutluluk paha biçilemezdir. Elbette bu sevinci yaşamak her annenin en doğal hakkıdır. Ancak bazı sebeplerden dolayı dünyada ki birçok kadın için bu durum korkutucu hale gelebilmektedir. Hamilelik sırasında annenin yaşadığı bazı problemler hem anne hem de bebeğin sağlığı için riskli durumlar oluşturabilir. Aniden ortaya çıkabilecek komplikasyonlar, şiddetli kanama riski, enfeksiyonlar veya artan kan basıncı gibi nedenler anne ve bebeğin ölümüne yol açabilecek etkenler olarak sıralanabilir. Anne ölümlerinin yaklaşık %94'ünün gelişmekte olan ülkelerde meydana geldiği görülmektedir ve bu ölümlerin neredeyse %75'i önlenebilir ölümlerdir [1]. Anne ölümlerinin önüne geçilmesi açısından erken teşhis önemli bir rol oynamaktadır. Bu sebepten dolayı teşhis yöntemlerinin de önemi büyüktür. Teknolojinin gelişmesiyle birlikte teşhis yöntemleri de gelişmiş olup yapay zekâ bu yöntemlerden biri olmuştur. Yapay zekâ, hem büyük veri setleri ile rahatça çalışabildiği için hem de hızlı bir şekilde çözüm üretebildiği için son yıllarda çokça tercih edilen yöntemler arasında yer almaktadır. Makine öğrenmesi yapay zekânın alt dallarından biri olup veri setlerinin eğitilerek tahminlerde bulunulmasını sağlamaktadır.

Marzia Ahmed ve Mohammod Abul Kashem, yaptıkları çalışmada veri setinden anne sağlığı riskini tahminlemek için makine öğrenmesi yöntemlerinden karar ağacı, rastgele orman, destek vektör makineleri, sıralı minimum optimizasyon, lojistik regresyon, naive bayes sınıflandırıcısı, Ibk ve lojistik model ağacı algoritmalarını kullanmışlardır. Öznitelik olarak hiper parametre ayarlama yöntemi olan GridSearchCV kullanarak en iyi sonucu %98.51 ile karar ağacı ile elde etmişlerdir [2]. Lokesh Pawar ve arkadaşları yaptıkları çalışmada anne sağlığı riskini tahminlemek için makine öğrenmesi sınıflandırıcılarından karar ağacı, naive bayes sınıflandırıcısı, çok katmanlı algılayıcı, J48, lojistik model ağacı, rastgele orman, Rep ağacı ve bagging algortimalarını kullanmışlardır. Gini indeksini kullanarak veri seti üzerinde özellik seçimi yapmışlardır ve en iyi sonucu %70.21 ile rastgele orman ile elde etmişlerdir [3]. Yu Mu ve arkadaşları, yaptıkları çalışmada makine öğrenmesi yöntemlerinden 5 seviye k-en yakın komşu ve karar ağacı algoritmalarını kullanmışlardır. 5 kat çapraz doğrulama ve eğitim doğrulama test veri kümelerinin oranını 3:1:1 yaparak en iyi sonucu %85 ile 5 seviye k-en yakın komşu ile elde etmişlerdir. Karar ağacından ise %79 sonucunu elde etmişlerdir [4]. Zahra Hoodboy ve arkadaşları yaptıkları çalışmada makine öğrenmesi yöntemlerinden destek vektör makinesi, k-en yakın komşu, XGBoost sınıflandırıcısı, AdaBoost sınıflandırıcısı, rastgele orman, lojistik regresyon, Gauss naive bayes sınıflandırıcı ve karar ağacı yöntemlerini kullanmışlardır. k-fold çapraz doğrulama ve sentetik azınlık aşırı örnekleme dengeleme tekniğini kullanarak en iyi sonucu %93 olarak XGBoost sınıflandırıcısı ile elde etmişlerdir [5]. Shuojia Wang ve arkadaşları, yaptıkları çalışmada makine öğrenmesi yöntemlerinden lojistik regresyon, destek vektör makinesi, karar ağacı, naive bayes sınıflandırıcısı, XGBoost sınıflandırıcısı ve rastgele orman yöntemlerini kullanmışlardır. 10 kat çapraz doğrulama ve çoklu doğrusallığı önlemek için değişkenlerin Pearson korelasyonunu test ederek en iyi sonucu %79 ile destek vekör makinesi ile elde etmişlerdir [6]. Prajina Edayath, yaptığı çalışmada makine öğrenmesi yöntemlerinden lojistik regresyon, naive bayes sınıflandırıcısı ve rastgele orman kullanmıştır. Alakasız özellikleri azaltmak için RFE yöntemini uygulayarak %62 ile uygulanan tüm yöntemlerin aynı sonucu verdiğini gözlemlemiştir [7]. Silas S. L. Pereira ve arkadaşları, yaptıkları çalışmada makine öğrenmesi yöntemlerinden rastgele orman, destek vektör makinesi, çok katmanlı algılayıcı, AdaBoost sınıflandırıcısı, karar ağacı ve Gauss naive bayes sınıflandırıcısını kullanmışlardır. Çalışmada RFE yöntemini ve temel bileşen analizini (PCA) uygulayarak en iyi sonucu %98 ile rastgele orman ile elde etmişlerdir [8]. Literatürde gerçekleştirilen çalışmaların listesi Tablo I'de verildiği gibidir.

Bu çalışmanın amacı; hamilelik esnasında ortaya çıkabilecek kanama veya enfeksiyon gibi risklerin basit tetkik ve ölçümlerden oluşturulabilecek verisetleri ve akıllı sınıflandırıcılar ile, bu komplikasyonlar yaşanmadan önce belirlenebilmesi için makine öğrenme yöntemlerinin performanslarının karşılaştırılmasıdır. Bu neden ile anne sağlığı riski ile litertürde var olan bir veri seti ve denetimli öğrenme yöntemleri olarak da bilinen sınıflandırıcılar kullanarak hamilelikte anne sağlığı riski üzerine bir tahminleme yapmaktır. Bunun yanı sıra çalışmada sınıflandırıcıların performanslarının karşılaştırması hedeflenmiştir. En iyi performansı veren yöntemlerin belirlenmesi, yapılan çalışmalar üzerinde doğru sonuçlar elde ederek hayatımıza yön veren veren durumlar için etki etmesi açısından önem arz etmektedir.

Anne sağlığı üzerine literatürde gerçekleştirilen çalışmalar incelendiğinde dönüşüm algoritmaları olarak PCA ve LDA dönüşümü seçiminin çok tercih edilmediği söylenebilmektedir. Bu çalışmada PCA ve LDA dönüşümünün kullanımının modellerin performansı üzerinde olumlu bir etkiye sahip olduğu görülebilmektedir. Özellikle LDA dönüşümünün kullanımı en yüksek doğruluk oranını elde etmeye yardımcı olmuştur. Bu oran %85 ile karar ağacı ve K-en yakın komşu ile elde edilmiştir. Bu sebeple LDA dönüşümünün sınıflandırıcı yöntemleri ile birleştirilerek sınıflandırma başarıların arttırılabileceği bu çalışma ile literatüre katkı olarak sunulmuştur..

Tablo 1. Literatür araştırması özet tablo

SAYI	YAYINLAYICI	VERİ SETİ	ALGORİTMA	ÖZNİTELİK	SONUÇ
1	Marzia Ahmed ve Mohammod Abul Kashem, 2021	-Dakka'daki hastaneler ve doğum kliniklerinden elde edilen bir veri seti	-Karar ağacı -Rastgele orman -Destek vektör makineleri -Sıralı minimum optimizasyon -Lojistik regresyon -Naive bayes sınıflandırıcısı -Ibk algoritması -Lojistik model ağacı	-GridSearchCV yöntemi -Ki-kare testi -Bilgi kazanımı -Kazanç oranı -15 çapraz katlama doğrulaması	-%98.51 ile Karar ağacı en iyi sonucu vermiştir.
2	Lokesh Pawar ve Arkadaşları, 2022	-Bangladeş' deki hastaneler ve doğum kliniklerinden elde edilen toplam 1014 veri	-Karar ağacı -Naive bayes sınıflandırıcısı -Çok katmanlı algılayıcı -J48 -Lojistik model ağacı -Rastgele orman -REP ağacı -Bagging algoritması	-Gini indeksini ile özellik seçimi -k-kat çapraz doğrulamaya dayanan Robust Modeli -%70 eğitim %30 test ve %60 eğitim %40 test olarak iki senaryo	-%70.21 ile Rastgele orman en iyi sonucu vermiştir.
3	Yu Mu ve Arkadaşları, 2018	-75542 çiftin çok boyutlu gebelik öncesi sağlık verileri	-5 layer k-en yakın komşu -Karar ağacı	-5 kat çapraz doğrulama -Eğitim doğrulama test veri kümelerinin oranı 3:1:1'dir.	-%85 ile 5-layer k-en yakın komşu en iyi sonucu vermiştir.
4	Zahra Hoodbhoy ve Arkadaşları, 2019	-California Üniversitesi Irvine Machine Learning Repository'den elde edilen gebeliğin üçüncü trimesterinde olan 2126 gebe kadından oluşa veri seti	-Destek vektör makinesi -K-en yakın komşu -XGBoost sınıflandırıcı -AdaBoost sınıflandırıcı -Rastgele orman -Lojistik regresyon -Gauss Naive bayes sınıflandırıcısı -Karar ağacı	-K-Fold Çapraz Doğrulama -Sentetik Azınlık Aşırı Örnekleme Dengeleme Tekniği (SMOTE)	-%93 ile XGBoost sınıflandırıcı en iyi sonucu vermiştir.
5	Shuojia Wang ve Arkadaşları, 2019	-2015 – 2017 yılları arasında Weill Cornell Medicine ve NewYork Presbyterian Hospital'den alınan veri seti	-Lojistik regresyon -Destek vektör makinesi -Karar ağacı -Naive bayes sınıflandırıcısı -XGBoost sınıflandırıcı -Rastgele orman	-10 kat çapraz doğrulama -Doğrusallığı önlemek için Pearson korelasyonu	-%79 ile destek vektör makinesi en iyi sonucu vermiştir.
6	Prajina Edayath ve Arkadaşları, 2022	-NIH All of Us'un elektronik sağlık kayıtlarında hamile kadınlardan alınan veriler	-Lojistik regresyon -Naive Bayes sınıflandırıcısı -Rastgele orman	-RFE yöntemi -Boyutsallık indirgemesi -Veri seti 60'a 40 olarak ayrılmış	-%62 ile bütün aynı sonucu vermiştir.
7	Silas S. L. Pereira ve arkadaşları	-Anonimleş tirilmiş bir veri seti	-Rastgele orman -Destek vektör makinesi -Çok katmanlı algılayıcı -AdaBoost sınıflandırıcısı -Karar ağacı -Gauss Naive bayes sınıflandırıcı	-RFE yöntemi -Temel Bileşen Analizi (PCA)	-%98 ile rastgele orman en iyi sonucu vermiştir.

Materyal ve Metot

Veri Setinin Elde Edilmesi

Çalışmada Marzia Ahmed isimli araştırmacının sunduğu The UCI (University of California, Irvine) Machine Learning Repository web sitesinden alınan "Maternal Health Risk Dataset" [9] adlı veri seti kullanılmıştır. Veriler, Bagladeş'in kırsal bölgelerinden IoT tabanlı risk izleme sistemi aracılığıyla farklı hastanelerden, toplum kliniklerinden ve anne sağlığı hizmetlerinden toplanmıştır. Veri seti 1014 örnek ve 7 öznitelikten oluşmaktadır. Veri setinde üç etiket bulunmaktadır. Bunlar yüksek risk, düşük risk ve orta risk şeklindedir. Tablo 2'de, veri setinde kullanılan öznitelikler ve açıklamaları gösterilmiştir.

Tablo 2. Öznitelik tablosu

Sıra No	Öznitelik	Açıklama	Değer
1	Age (Yaş)	Annenin hamilelik yaşını belirtir.	Nümerik
2	SistolikBP (Büyük Tansiyon)	Kan basıncının (tansiyon) mmHg cinsinden üst değerini belirtir.	Nümerik
3	DiastolikBP (Küçük Tansiyon)	Kan basıncının (tansiyon) mmHg cinsinden alt değerini belirtir.	Nümerik
4	BS (Kan Şekeri)	Kan şekeri seviyelerini molar konsantrasyon (mmol/L) cinsinden belirtir.	Nümerik
5	HeartRate (Kalp Atımı)	Dakikadaki atış cinsinden normal dinlenme kalp atış hızı.	Nümerik
6	BodyTemp (Vücut Sıcaklığı)	Anne vücut sıcaklığını belirtir. Fahrenheit cinsinden verilmiştir.	Nümerik
7	Risk Level (Risk Düzeyi)	Önceki özellikleri dikkate alınarak gebelik sırasında öngörülen risk yoğunluğu düzeyi	Kategorik

Öznitelik Seçimi

Makine öğrenmesi yöntemlerini etkin bir şekilde kullanabilmek için ön verilerin işlenmesi esastır. Öznitelik seçimi, veri ön analizinde en sık kullanılan ve önemli tekniklerden biridir [10]. Bu çalışmada dönüşüm olarak Temel Bileşenler Analizi ve Doğrusal Diskriminant Analizi kullanılmıştır. PCA veri kümelerinin boyutunu azaltarak yorumlanabilirliği arttırmak ve aynı zamanda bilgi kaybını en aza indirmek amacıyla kullanılan bir tekniktir. Varyansı art arda maksimize eden yeni ilintisiz değişkenler oluşturarak bilgi kaybını en aza indirir. LDA veri setindeki değişkenlerin iki veya daha fazla gerçek gruba ayrılmasını sağlamak için kullanılan analizdir. Bu yöntemi kullanarak, p özelliği bilinen birimleri gerçek gruplarına optimal düzeyde atayacak fonksiyonlar bulunabilir [11]. Çalışmada PCA ve LDA dönüşümleri kullanılarak sınıflandırma performansları karşılaştırılmıştır.

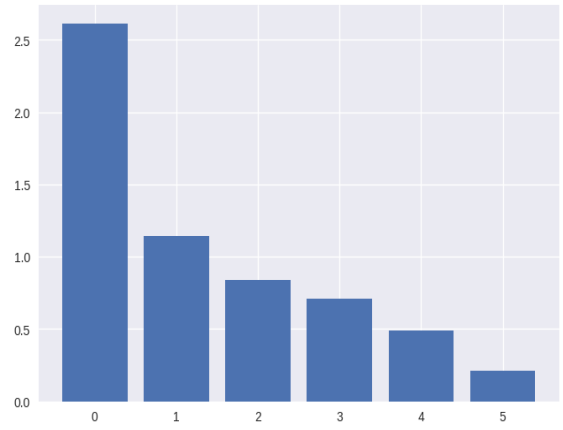
Sınıflandırıcıların Uygulanması

Çalışmada denetimli öğrenme sınıflandırıcılarından lineer regresyon, destek vektör makineleri, karar ağacı, rastgele orman, çok katmanlı algılayıcı, naive bayes sınıflandırıcısı, k-en yakın komşu ve XGBoost sınıflandırıcısı kullanılmıştır ve aşağıda tanımlamaları verilmiştir.

Lineer regresyon, x ve y değişkenleri arasındaki ilişkinin doğrusal bir şekilde yayıldığı istatistiksel bir yöntemdir. Bu yöntemde giriş verisi bağımsız x değişkeni ve çıkış verisi ise x'e bağımlı bir y değişkeni olarak adlandırılır. Yalnızca bir x değişkenine bağlı olan y değişkeni, bağımlı değişken olarak bilinir [12]. Destek vektör makineleri, sınıflandırma ve regresyon analizi yapabilen denetimli öğrenme modellerinden biridir. Eğitim veri setindeki her öğe, bir vektörün bir tarafında veya diğer tarafında işaretlenir. Eğitim, yeni bir örnek için iki sınıftan birini belirleyen bir model oluşturur. Bu sınıflandırıcının temel amacı ikili bir doğrusal sınıf oluşturmaktır [13]. Karar ağacı, bir kök düğümü ve her biri bir girdi alan iç düğümlerden oluşan yönlü bir ağaçtır. İç düğümler, girdi özelliklerine göre veriyi bölme kararlarını alır. Test düğümleri, çıktıları bir başka düğüme girdi olarak alan ve belirli bir kurala göre veriyi sınıflandıran veya tahmin eden düğümlerdir. Yaprak düğümleri ise sonuç çıktıları veren ve başka bir düğüme bağlı olmayan düğümlerdir [14]. Rastgele Orman bir ağaç oluştururken her düğümde tüm değişkenler arasından en iyi dalı kullanmak yerine, o düğüm için rastgele bir alt küme seçerek en iyi dala göre dallara ayırma işlemini gerçekleştirir. Bu şekilde, her ağaç farklı bir alt küme ile oluşturulur ve sonuçta oluşturulan tüm ağaçların tahminleri birleştirilerek sınıflandırma yapılır [14]. Çok katmanlı algılayıcılar üç farklı katmandan oluşmaktadır. Bunlar, gelen bilgilerin öğrenme işlemi için kullanıldığı ve gizli katmana iletildiği girdi katmanı, öğrenme işleminin gerçekleştirildiği bir veya birden çok gizli katman ve bilgi çıkışının sağlandığı çıkış katmanıdır. Bu katmanlar ileri beslemeli-geri yayılımlı algoritmaların etkili bir şekilde kullanılmasına olanak tanır [15]. Naive Bayes Sınıflandırıcı, veri setindeki verilerin sınıflandırılmasını olasılık hesapları yardımıyla gerçekleştirir. Her bir veri elemanı için ayrı ayrı tüm olasılıkları hesaplar ve en yüksek olasılık değerine sahip sınıfa göre sınıflandırma yapar [16]. K-En Yakın Komşu, bir verinin kendisine en yakın komşularının kimliğini kullanarak sınıflandırma veya regresyon yapar. K, komşu sayısını belirten bir parametredir ve doğru bir sınıflandırma için doğru K değerinin seçilmesi önemlidir [17]. XGBoost sınıflandırıcısı, ağaç güçlendirme mantığına dayanmaktadır. Hızlı işlem yeteneği ve yüksek performansı nedeniyle yüksek başarımda sonuçlar veren bir makine öğrenmesi yöntemidir[18].

Uygulamada bütün sınıflandırma işlemleri için Google Colaboratory [19] üzerinden Python [20] kodları kullanılmıştır. Veri seti eğitim ile test olmak üzere tüm modeller için %75'e %25 olarak ayrılmıştır ve her zaman aynı ayrımı yapması için verisetindeki eğitim ve test verileri sabit tutulmuştur. Daha sonra min-max standardizasyon işlemi uygulanmıştır. Eğitim ve test veri setleri standardize edildikten sonra, denetimli öğrenme sınıflandırıcıları kullanılmıştır.

Veri setine uygulanacak olan dönüşümler uygulanmadan önce boyut sayısı belirleme işlemi yapılmıştır. Temel bileşenler analizi için açıklanan varyans değerlerine bakılarak boyut sayısı 2 olarak belirlenmiş ve LDA dönüşümü için de aynı değer kullanılmıştır. Şekil 1'de en yüksek varyans değerine sahip ilk 5 öznitelik gösterilmiştir. PCA ve LDA dönüşümleri tarafından dönüştürülen veri setinde; eğitim veri seti 760 satır ve 2 sütun, test veri seti de 254 satır ve 2 sütundan oluşmaktadır. Uygulamada önce dönüşümsüz model performansları belirlenmiş daha sonra PCA ve LDA dönüşümleriyle elde edilen model performansları belirlenerek karşılaştırma yapılmıştır.



Şekil 1. Açıklanan varyans değerleri

Sonuçlar

Çalışmada, her bir yöntem için önce dönüşüm yapılmadan doğruluk oranına daha sonra da PCA ve LDA dönüşümü uygulanarak doğruluk oranları hesaplanmıştır. Kullanılan yöntemlerde optimum hiperparametreler deneysel olarak belirlenmiş ve sınıflandırıcıların başarıları karşılaştırılmıştır.

Anne sağlığı riski veri setine; lineer regresyon (LR) uygulandığında, doğruluk oranı %62 olarak elde edilmiştir. PCA dönüşümü kullanıldığında doğruluk oranı %60 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %61 olarak elde edilmiştir. Destek vektör makineleri (DVM) uygulandığında, kernel fonksiyonu olarak rbf modeli ile gamma = 31, C = 9 değerleri seçildiğinde en iyi doğruluk oranı %78 olarak bulunmuştur. Aynı parametreler ile PCA dönüşümü kullanıldığında doğruluk oranı %77 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %77 olarak elde edilmiştir. Karar ağacı (KA) uygulandığında, "gini" parametresi kullanılarak en iyi doğruluk oranı %79 olarak bulunmuştur. Aynı parametre ile PCA dönüşümü kullanıldığında doğruluk oranı %81 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %85 olarak elde edilmiştir. Rastgele orman (RO) uygulandığında, ağaç sayısı 31 seçilerek en iyi doğruluk oranı %84 olarak bulunmuştur. Aynı parametre ile PCA dönüşümü kullanıldığında doğruluk oranı %82 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %84 olarak elde edilmiştir. Çok katmanlı algılayıcılar (ÇKA) uygulandığında, 3 tane gizli katman ile aktivasyon fonksiyonu "relu" ve ağırlık optimizasyonu "adam" seçildiğinde doğruluk oranı %78 olarak bulunmuştur. Aynı parametreler ile PCA dönüşümü kullanıldığında doğruluk oranı %67 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %72 olarak elde edilmiştir. Naive Bayes (NB) sınıflandırıcısı uygulandığında, doğruluk oranı %62 olarak elde edilmiştir. PCA dönüşümü kullanıldığında doğruluk oranı %55 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %64 olarak elde edilmiştir. K-en yakın komşu (KNN) uygulandığında, k=1 seçildiğinde en iyi doğruluk oranı %76 olarak bulunmuştur. Aynı parametre ile PCA dönüşümü kullanıldığında doğruluk oranı %82 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %85 olarak elde edilmiştir. XGBoost (XGB) sınıflandırıcı uygulandığında, değerlendirme ölçütü olarak "mlogloss" seçildiğinde en iyi doğruluk oranı %78 olarak bulunmuştur. Aynı parametre ile PCA dönüşümü kullanıldığında doğruluk oranı %67 olarak bulunmuştur. LDA dönüşümü kullanıldığında ise doğruluk oranı %72 olarak elde edilmiştir. Elde edilen sonuçlar Tablo 3'te verildiği gibidir.

Tablo 3. Sınıflandırıcıların test veri seti doğruluk oranları

Model	Doğruluk	Doğruluk (PCA)	Doğruluk (LDA)
LR	%62	%60	%61
DVM	%78	%77	%77
KA	%79	%81	**%85**
RO	**%84**	**%82**	%84
ÇKA	%78	%67	%72
NB	%62	%55	%64
KNN	%76	**%82**	**%85**
XGB	%78	%67	%72

Çalışmada kullanılan sınıflandırıcılar arasında uygulanan t-testi sonuçlarında p değeri 0.05'den küçük hesaplanmıştır. Bu sınıflandırıcıların doğruluk oranlarının birbirleri ile arasında istatistiksel olarak anlamlı olduğunu göstermektedir ($p < 0,05$).

Tartışma ve Öneriler

Sınıflandırıcıların performansları incelendiğinde, rastgele orman %84 ile en yüksek doğruluk oranına sahip olduğu görülmektedir. Benzer bir çalışma yapan Lokesh Pawar ve

arkadaşları [3] en iyi sonucu %70.21 ile rastgele orman ile elde etmişlerdir. Aynı veri seti ile çalışma yapan Marzia Ahmed ve Mohammod Abul Kashem [2] ise en iyi sonucu %98.51 ile karar ağacı ile elde etmişlerdir. Aynı veri setine aynı sınıflandırıcılar uygulandığında elde edilen sonuçların farklı olmasının en önemli nedeni eğitim ve test verilerinin bölünmesinde farklı yöntemlerin kullanılmış olmasıdır.

Sınıflandırıcıların test veri seti doğruluk oranlarına bakıldığında destek vektör makinasının en yüksek %78 oranını verdiği görülmektedir. Shuojia Wang ve arkadaşları [6] en yüksek doğruluk oranını %79 ile destek vektör makinesi ile elde etmişlerdir.

PCA dönüşümü kullanılarak yapılan tahminleme sonuçlarına göre en yüksek doğruluk oranının %82 ile rastgele orman ve K-en yakın komşu olduğu görülmektedir. Aynı dönüşümü kullanarak çalışma yapan Silas S. L. Pereira ve arkadaşları [8] en yüksek doğruluk oranını %98 ile rastgele orman ile elde etmişlerdir. Burada öznitelik seçimi sonrasında elde edilen öznitelik sayıları sınıflandırma başarısını etkileyen faktörlerin başında gelmektedir.

LDA dönüşümü kullanılarak yapılan tahminleme sonuçlarına göre ise en yüksek doğruluk oranının %85 ile karar ağacı ve K-en yakın komşu ile elde edildiği görülmektedir. Yu Mu ve arkadaşları [4] yaptıkları çalışmada en iyi sonucu %85 ile K-en yakın komşu ile elde etmişlerdir. Zahra Hoodboy ve arkadaşları [5] yaptıkları çalışmada en iyi sonucu %93 olarak XGBoost sınıflandırıcısından elde etmişlerdir. Bu çalışma için doğruluk oranlarına bakıldığında ise XGBoost sınıflandırıcısının en yüksek %78 oranını verdiği görülmektedir.

Çalışmada test veri seti doğruluk oranlarına göre lojistik regresyon %62 ile en düşük doğruluk değerini vermiştir. Prajina Edayath ve arkadaşları [7] lojistik regresyon ile %62 oranında bir doğruluk sağlamışlardır.Bu çalışmaya ve benzer veri setleri kullanılarak daha önce yapılmış çalışmalara bakıldığında doğruluk oranları göz önüne alınırsa rastgele orman, karar ağacı ve k-en yakın komşu yöntemlerinin yüksek performanslar elde edildiği görülmektedir. Bu durum benzer veri setleri kullanılarak gelecekte yapılması hedeflenen çalışmalara yön gösterebilmektedir. Doğru sınıflandırıcı seçimi, doğru performans değerleri elde ettiğinden çalışmaların amacına uygun kullanımlara hazır hale gelebilmesi açısından önem arz etmektedir.

PCA ve LDA gibi dönüşüm yöntemleri ile öznitelik boyutu indirgeme işlemlerinde, varyansın maksimize edilmesine odaklanıldığı için bazı durumlarda orijinal veri seti üzerindeki ayırt edici bilgilerin kaybolması olası durumlardandır. Bu durumu etkileyen önemli faktörlerden biri PCA ve LDA'nın veri dağılımının varsayımlarına dayalı olarak çalışmasıdır. Bu durum öznitelik indirgeme işlemleri sonucunda sınıflandırma performansında iyileşmenin minimum değerlerde olmasına neden olabilmektedir. Gerçekleştirilen çalışmada öznitelik indirgeme işlemleri olarak kullanılan LDA ve PCA ile elde edilmiş veri setleri için elde edilen sınıflandırma performansı ile orijinal veri seti ile gerçekleştirilmiş sınıflandırma performanslarının yakın sonuçlar vermesinin nedeni olarak açıklanabilir.

Anne sağlığı riski veri seti kullanılarak yapılan çalışmada elde edilen modeller üzerinden daha güncel bir veri seti kullanılarak doğruluk tahminleri yapılabilir. Aynı zamanda birden fazla veri seti birleştirilerek hibrit bir çalışma denenebilir. Bu çalışmada LDA dönüşümü fayda sağlamıştır fakat daha farklı öznitelikler de kullanılarak en iyi sonuçlar elde edilmeye çalışılabilir. Daha farklı sınıflandırıcıların kullanımı da en iyi modeli bulmada etkili olabilmektedir.

Kaynakça

[1] Demir Yıldırım, A. ve Hotun Şahin, N. (2022) Anne Ölümlerinin Önlenmesi: Uluslararası Bakım ve İzlem Modelleri, Üsküdar Üniversitesi, Sağlık Bilimleri Fakültesi, Ebelik Bölümü, Jinekoloji – Obstetrik Neonatoloji Tıp Dergisi, 19(1), DOI: 10.38136/jgon.842685.

[2] Ahmed M. ve Kashem M. A. (2020) IoT Based Risk Level Prediction Model For Maternal Health Care In The Context Of Bangladesh, 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), (19.12.2020-20.12.2020)

[3] Pawar, L., Arora, D., Malhotra, J., Vaidya, D., Sharma, A. (2022) A Robust Machine Learning Predictive Model for Maternal Health Risk, Proceedings of the Third International Conference on Electronics and Sustainable Communication Systems, ISBN: 978-1-6654-7971-4, DOI: 10.1109/ICESC54411.2022.9885515.

[4] Mu Y., Feng K., Yang Y. ve Wang J. (2018) Applying Deep Learning for Adverse Pregnancy Outcome Detection With Pre-pregnancy Health Data, MATEC Web of Conferences 189, 10014.

[5] Hoodbhoy Z., Noman M., Shafique A., Nasim A., Chowdhury D. ve Hasan B. (2019) Use of Machine Learning Algorithms for Prediction of Fetal Risk using Cardiotocographic Data, International Journal of Applied and Basic Medical Research, The Aga Khan University, Department of Paediatrics and Child Health, Department of Artificial Intelligence, 9:226-30.

[6] Wang S., Pathak J. ve Zhang Y. (2019) Using Electronic Health Records and Machine Learning to Predict Postpartum Depression, Cornell University, International Medical Informatics Association (IMIA) and IOS Press, doi:10. 3233/SHTI190351

[7] Edayath P. (2022) Analysis Of Factors Affecting Maternal Health Using Data Mining Techniques, Master Thesis, Master's Program in Industrial Engineering, University of Texas at El Paso.

[8] Pereira S. L. S., Filho R. V. C., Ramos R., Oliveira M., Moreira M. W. L., Rodrigues J. J. P. C. ve Solic P. (2020) Improving Maternal Risk Analysis in Public

Health Systems, 5th International Conference on Smart and Sustainable Technologies (SpliTech), (23-26. 09. 2020).

[9] Ahmed M. (2020) "Maternal Health Risk Data Set" UCI Repository of Machine Learning Databases, University of California, Irvine. Erişim adresi: https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set

[10] Ertuğrul S. (2022) Öznitelik Seçimi ve Makine Öğrenimi Kullanılarak Enerji İletim, Kontrol ve Yönetim Sistemlerinde Siber Güvenlik Analizi, Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı, İstanbul Arel Üniversitesi.

[11] Saeed M. T. M. (2022) Temel Bileşenler Analizi ve Yapay Sinir Ağları Kullanılarak Turbofan Motorunun Kalan Faydalı Ömür Tahmini, Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı, İstanbul Üniversitesi.

[12] Doğan G. (2022) Makine Öğrenmesi Algoritmaları ile Betonarme Kirişlerin Burulma Momenti Tahmini, El-Cezerî Fen ve Mühendislik Dergisi, 9(2); 912-924.

[13] Batı F. (2020) Makine Öğrenmesi Sınıflandırma Algoritmaları Kullanılarak Meme Kanseri Tahmini, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul Aydın Üniversitesi.

[14] Özgür E. (2022) Lenf Kanseri Görüntülerinin Makine Öğrenmesi Yöntemleri İle Sınıflandırılması, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı, Tekirdağ Namık Kemal Üniversitesi.

[15] Kaya M. S. ve İnce K. (2021) Nesnelerin İnternetinde Çok Katmanlı Algılayıcı Kullanarak Zamanlama Analizi Saldırısı ile Özel Anahtar Tahminlemesi, Bilgisayar Bilimleri Dergisi, İnönü Üniversitesi, 385-390.

[16] Görgün M. (2020) Makine Öğrenmesi Yöntemleri İle Kalp Hastalığının Teşhis Edilmesi, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul Aydın Üniversitesi.

[17] Keskinbıçak F. (2023) Makine Öğrenmesi İle Nohutta Verim ve Tür Tahmini, Yüksek Lisans Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı, Harran Üniversitesi.

[18] Azizoğlu F. (2023) Makine Öğrenmesi Yöntemleriyle Kalp Hastalıklarının Sağkalım Tahmini, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Ana Bilim Dalı, Sivas Cumhuriyet Üniversitesi.

[19] Google (2017). [https://globalaihub.com/google-colab-nedir-ve-nasil-kullanilir], Erişim tarihi: 09.02.2023

[20] Rossum G (1990). [https://tr.wikipedia.org/wiki/Python#], Erişim tarihi: 09.02.2023.

[21] Kılınç D., Borandağ E., Yücalar F., Tunalı V., Şimşek M. ve Özçift A. (2016) KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi, Marmara Fen Bilimleri Dergisi, Hasan Ferdi Turgutlu Teknoloji Fakültesi, Celal Bayar Üniversitesi, 3:89-94.

[22] Gürsoy G. (2022) Makine Öğrenmesi Algoritmaları İle Kalp Hastalığı Tahmini, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Anabilim Dalı, Maltepe Üniversitesi.

[23] Korkmaz A. ve Büyükgöze S. (2019) Sahte Web Sitelerinin Sınıflandırma Algoritmaları İle Tespit Edilmesi, Avrupa Bilim ve Teknoloji Dergisi, Fen Bilimleri Enstitüsü, İstanbul Üniversitesi, 16:826-833.

[24] Durak M. N. (2017) Makine Öğrenmesi Sınıflandırma Yöntemleri İle Meme Kanserinin Erken Teşhisi, Yüksek Lisans Tezi, İstatistik Anabilim Dalı, Yıldız Teknik Üniversitesi.