NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2023; 12(1), 064-071

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Şeker hastalığı teşhisi ve önerilen modellerinin karşılaştırılması

Diagnosis of diabetes and comparison of proposed models

Merve Korkmaz[1], Kaplan Kaplan[2,*]

[1] *Kocaeli Üniversitesi, Bilgisayar Mühendisliği Bölümü, 41580, Kocaeli, Türkiye*
[2] *Kocaeli Üniversitesi, Yazılım Mühendisliği Bölümü, 41580, Kocaeli, Türkiye*

Öz

Şeker hastalığı insanlarda kan şekeri seviyesinin anormal değerlere ulaştığı kronik bir rahatsızlıktır. Şeker hastalığının erken teşhisi, bu hastalığın sebep olabileceği daha büyük hastalıkların önlenmesi ve gerekli tedavi planlamasının zamanında gerçekleştirilmesi açısından önemlidir. Bu çalışma kapsamında şeker hastalığı çeşitli modeller ile teşhis edilerek, bu problem için kullanılabilecek en uygun model belirlenmeye çalışılmıştır. Çalışmada Lojistik Regresyon, k-En Yakın Komşuluk, CART (Sınıflandırma ve Regresyon Ağacı), Rastgele Orman, Destek Vektör Makinesi, XGBoost ve LightGBM sınıflandırıcı modelleri kullanılmıştır. 10 katlı çapraz doğrulama yöntemi kullanılarak performans ölçütleri elde edilmiştir. Modellerin doğruluk oranları sırası ile %84.58, %84.59, %85.02, %88.29, %84.73, %89.29 ve %88.72 olarak elde edilmiştir. Modeller arasında en iyi üç doğruluk oranını veren Rastgele Orman, XGBoost ve LightGBM yöntemlerinde hiper-parametre ayarlaması gerçekleştirilerek en iyi parametreler belirlenmiştir. Bu parametreler ile final modellerinin doğruluk oranları sırasıyla %89.30 , %90.01 ve %90.01 olarak elde edilmiştir. Sonuç olarak XGBoost ve LightGBM modellerinin final teşhis modelleri olarak kullanılabileceği gözlemlenmiştir.

Anahtar kelimeler: Şeker hastalığı, Teşhis, Sınıflandırma, Makine öğrenmesi, Topluluk öğrenimi

Abstract

Diabetes is a chronic disease in which blood sugar levels reach abnormal values in humans. Early diagnosis of diabetes is important in terms of preventing larger diseases that this disease may cause and realizing the necessary treatment planning in a timely manner. Within the scope of this study, diabetes was diagnosed with various models and the most suitable model that could be used for this problem was tried to be determined. In this study, Logistic Regression, k-Nearest Neighborhood, CART (Classification and Regression Tree), Random Forest, Support Vector Machine, XGBoost and LightGBM classifier models were used. Performance metrics were obtained using the 10-fold cross-validation method. The accuracy rates of the models were obtained as 84.58%, 84.59%, 85.02%, 88.29%, 84.73%, 89.29% and 88.72%, respectively. The best parameters were determined by performing hyper-parameter tuning in Random Forest, XGBoost and LightGBM methods, which gave the three best accuracy rates among the models. With these parameters, the accuracy rates of the final models were 89.30%, 90.01% and 90.01%, respectively. As a result, it has been observed that XGBoost and LightGBM models can be used as final diagnostic models.

Keywords: Diabetes mellitus, Diagnosis, Classification, machine learning, Ensemble learning

1 Giriş

Şeker hastalığı vücutta yer alan pankreas organının yeterince insülin hormonu üretememesi veya vücudun ürettiği insülin hormonunun yetersizliğine bağlı olarak gelişen ve günümüzde insanlarda en sık görülen kronik hastalıklardan bir tanesidir [1]. Ağız kuruluğu, noktüri, poliüri, polidipsi, iştahsızlık, halsizlik veya çabuk yorulma gibi klasik semptomlar ile ya da bulanık görme, açıklanamayan kilo kaybı, kaşıntı, tekrarlayan mantar enfeksiyonları ya da inatçı enfeksiyonlar gibi daha az görülen semptomlar ile belirti gösteren bu hastalık [2] hipoglisemi, diyabetik ketoasidoz, hiperglisemik hiperozmolar nonketotik koma, koroner arter hastalığı, serebrovasküler hastalık, periferik arter hastalığı gibi makrovasküler komplikasyonlara ya da retinopati, nefropati, nöropati veya diyabetik ayak gibi mikrovasküler komplikasyonlara sebep olması nedeni ile morbidite ve mortalite oranı yüksek kronik metabolik hastalıklardan biridir [3].

2021 yılı itibari ile dünyada 537 milyon şeker hastası olduğu ve sadece 2021 yılında 6.7 milyon kişinin şeker hastalığı sebebi ile hayatını kaybettiği bilinmektedir [1]. Bu sebep ile şeker hastalığının erken teşhis edilmesi, şeker hastalığının sebep olabileceği komplikasyonların önüne geçilebilmesi ve şeker hastalığına bağlı gelişen komplikasyonlar sebebi ile oluşabilecek organ kaybı ya da insan ölümlerinin engellenebilmesi açısından büyük önem taşımaktadır.

Şeker hastalığının teşhisi insan sağlığı uzmanları tarafından elle yapılan muayeneler sonucunda ya da hastalardan alınan kan numunelerinin laboratuvar ortamında tıbbi bir cihaz yardımıyla incelenmesi sonunda yapılabilir. Fakat şeker hastalığı çok fazla belirti göstermeden ilerleyen bir hastalık olması sebebi ile alanında uzman olan doktorlar tarafından bile net olarak teşhis edilemeyebilir [4].

* Sorumlu yazar / Corresponding author, e-posta / e-mail: kaplan.kaplan@kocaeli.edu.tr (K. Kaplan)
Geliş / Recieved: 14.08.2022 Kabul / Accepted: 19.12.2022 Yayımlanma / Published: 15.01.2023
doi: 10.28948/ngumuh.1161768

Teknolojik gelişmelerin hızla ilerlemesiyle birlikte dünya nüfusunu büyük oranda etkileyen bu gibi hastalıkların erken teşhis edilebilmesi için çoğu araştırmacı insan sağlığı uzmanlarının teşhis süresini en aza indirmek ve teşhislerin doğruluk oranını arttırmak amacıyla makine öğrenmesi, yapay zekâ ve derin öğrenme alanlarında çalışmalar yürütmektedir.

Bugüne kadar gerçekleştirilen ve aynı veri kümesinin kullanıldığı çalışmalarda veri kümesinde bulunan eksikliklerin göz ardı edilmesi ve aykırı değerlerin incelenmemesi model performanslarında önemli derecede azalmaya neden olmuştur. Aynı zamanda hesaplanan özniteliklerin karakteristik ve tanımlayıcı olmaması ve hiper parametre ayarlamasının deneme yanılma ile yapılması, diğer çalışmaların performanslarını sınırlandırmıştır.

Literatürde yer alan çalışmalara kıyasla yapılmış olan bu çalışmanın keşifsel veri analizi aşamasında kullanılan yöntemlerin çeşitliliği sayesinde veri kümesindeki dengesizlikler belirlenerek veri kümesinde yer alan eksik ve aykırı değer sorunları çözülerek veri kümesi sınıflandırma modelleri için daha uygun bir hale getirildi. Ardından "Glucose", "BMI" ve "İnsulin" özelliklerinden yeni özelliklerin oluşturulması ile kullanılan sınıflandırma modellerinde özelliklerin modellerdeki özellik önem düzeylerini de belirlememize katkı sağladı. Bu sayede final modellerinin hepsi için sınıflandırmadaki en önemli özelliğin "Insulin" olduğu belirlendi. Bu çalışma kapsamında veri kümesi üzerinde dışarıda tutma ve 10 katlı çapraz doğrulama yöntemleri ayrı ayrı kullanıldı ve en iyi doğruluk veren yöntemin 10 katlı çapraz doğrulama yöntemi olduğu belirlendi. Final modellerinin oluşturulması belirlenen bu yöntem üzerinden gerçekleştirildi.

Gerçekleştirilen bu çalışmada ise Kaggle platformunda yer alan ve literatürde; "Pima Indians Diabetes" ismi ile bilinen açık kaynak veri kümesi kullanılarak şeker hastalığının analizi, belirlenmesi ve sınıflandırılmasına yönelik makine öğrenmesi sınıflandırma modelleri kullanılarak kullanılan modellerin performanslarının karşılaştırılması ve özellik önem düzeylerinin belirlenmesi gerçekleştirilmiştir. En iyi doğruluk veren modellerde hiper parametre ayarlaması gerçekleştirilerek teşhis aşamasında kullanılabilecek final modeller belirlenmiştir.

Bu çalışmanın 2. bölümünde literatürde yer alan çalışmalara değinilerek bir literatür özeti verilmiştir. 3. bölümde çalışma kapsamında kullanılan yaklaşımlardan, veri kümesinden ve veri kümesinin uygulanacak olan modellere uygun hale getirilmesi aşamasına kadar yapılan işlemlerden bahsedilmiştir. 4. bölümde ise çalışmanın deneysel sonuçlarına değinilmiştir. Son olarak ise 5. bölümde çalışmanın sonuçları verilerek çalışma sonlandırılmıştır.

2 Literatür taraması

Günümüzde teknolojinin hızla gelişmesiyle birlikte birçok alanda teknolojik yenilikler gerçekleşmiştir ve bu yenilikler gün geçtikçe artmaktadır. Günümüzde teknolojik yeniliklerin en önemli kullanım alanlarından bir tanesi de sağlık alanıdır. Sağlık alanındaki verimliliğin arttırılması, tedavi planlamasının zamanında gerçekleştirilmesi, doğru ve hızlı bir şekilde hastalık teşhislerinin yapılabilmesi için yapay zekâ teknolojileri sıkça tercih edilen yöntemler arasında yer almaktadır [5-6]. Literatürde yer alan çalışmalarda incelendiğinde sağlık alanında hastalıkların teşhis edilmesi, sınıflandırılması ve hastalıklara bağlı ölüm risk oranlarının tahmin edilmesi üzerine birçok çalışma yapıldığı görülmektedir. Literatürdeki çalışmalar incelendiğinde şeker hastalığı tahmin veya teşhisi için yapılmış birçok çalışmanın var olduğu ve farklı veri kümelerinin kullanıldığı görülmektedir. Veranyurt ve ark. gerçekleştirdikleri çalışmada birçok hastalığın oluşumuna sebep olması ve görülme sıklığının giderek artması sebebi ile şeker hastalığı üzerine çalışmışlardır. Çalışmalarında hastalığın erken teşhis edilebilmesinde büyük öneme sahip değişkenlerin olduğu açık kaynak bir veri kümesini seçerek bu veri kümesi üzerinde Rastgele Orman, K-En Yakın Komşu ve Adaboost modellerini kullanarak şeker hastalığını sırasıyla %92.30, %92.3 ve %90.59 başarı oranları ile tespit etmişlerdir [7]. Özkan ve ark. [8] çalışmalarında şeker hastalığı tanısının konulmasında veri kümesinde yer alan hangi değişkenin daha önemli olduğunu belirlemek için sınıflandırma modellerini iki farklı yaklaşım üzerine kurmuşlardır. Ayrıca kullandıkları veri kümesini bir devlet hastanesinden 18 yaşından büyük hastalardan topladıkları veriler ile elde etmişlerdir. Çalışmalarında göze çarpan bir diğer detay ise veri kümelerinde yer alan verilerin sadece şeker hastası olan ve şeker hastası olmayan hastalardan oluşturmaktan ziyade şeker hastası olan, gizli şeker hastalığı (prediyabet) olan, şeker hastalığı olmayan ve gizli şeker hastalığı olmayan hastalardan elde ederek gerçekleştirmişlerdir. Çalışmalarında kullandıkları veri kümesinde yedisi kategorik olmak üzere 39 bağımsız değişken yer almaktadır [8]. Yaptıkları bu çalışmada şeker hastalığının tahmini için denetimli öğrenme tekniklerinin performanslarını karşılaştırmışlardır. Çalışma sonucunda en iyi performansa sahip algoritmanın Rastgele Orman algoritması olduğunu ve modelin doğruluğunun %84,48 olduğu belirtmişlerdir [8]. Açık kaynak veri kümelerinden bir başkasının kullanıldığı Akyol ve ark. [4] şeker hastalığının erken tanısı gerçekleştirmek için yaptıkları bu çalışmada şeker hastalığının belirtileri olan poliüri, polidipsi ve polifaji gibi özniteliklerin olduğu bir veri kümesini kullanmışlardır. Gerçekleştirdikleri çalışmada Topluluk Oylama, Gradyan Artırma, Rastgele Orman, K-En Yakın Komşu ve Derin Sinir Ağı modellerini kullanmışlardır. Beş katlı çapraz doğrulama yöntemi sonucunda modellerin doğruluk oranlarını sırası ile %97.31, %95.38, %96.54, %85.58 ve %95.19 olarak bulduklarını belirtmişlerdir [4]. Fakat çalışmada veri kümesinde yer alan özniteliklerden hangisinin ya da hangilerinin önemli olduğuna değinilmediğini ilerleyen çalışmalarında bunun tespiti ve sınıflandırma modellerinin optimizasyonu üzerine çalışmayı hedeflediklerini belirtmişlerdir [4]. Şeker hastalığının erken dönemde teşhis edilebilmesi için yapılan farklı bir çalışmada ise açık kaynak bir veri kümesi kullanılmıştır. Kullanılan veri kümesi şeker hastalığı semptomlarını gösteren şeker hastası tanısı konmuş ve şeker hastalığı tanısı konulmamış ama şeker hastalığı semptomlarını gösteren bireylerden doğrudan yapılan bir anket çalışması ile oluşturulmuştur [9].

Bu çalışmada üç, beş ve 10 katlı çapraz doğrulama yöntemleri kullanılmıştır. 10 katlı çapraz doğrulama sonucunda K-En Yakın Komşu modeli %99.81 doğruluk oranı vermiştir. Ayrıca çalışma kapsamında veri kümesinde yer alan şeker hastalığı semptomlarının yer aldığı küçük bir anket tarzı bilgisayar arayüzü hazırlanarak şeker hastalığının ön tahminin yapılması gerçekleştirilmiştir [9].

Literatürde yer alan çalışmalar incelendiğinde şeker hastalığının makine öğrenmesi yöntemleri kullanılarak birçok sınıflandırma modeli önerildiği görülmektedir. Şeker hastalığının sınıflandırılması için kullanılan en yaygın açık veri kümelerinden bir tanesi "Pima Indians Diabetes" (PID) isimli veri kümesidir. Bu veri kümesinin dahil edildiği çalışmalardan birisini Tigga ve ark. gerçekleştirmiştir [10]. Çalışmalarında kullandıkları diğer bir veri kümesini 18 yaş üstü 952 katılımcıdan topladıkları veriler ile oluşturmuşlardır [10]. Çalışmalarında kullanılan makine öğrenmesi algoritmaları Lojistik Regresyon, K-En Yakın Komşu, Destek Vektör Makinesi, Naive Bayes, Karar Ağacı ve Rastgele Orman algoritmalarıdır. Bu sınıflandırma algoritmaları sırası PID veri kümesi üzerinde %74.4, %70.8, %74.4, %68.9, %69.7 ve %75 doğruluk değerlerini vermiştir. 10 katlı çapraz doğrulama sonuçları ise sırasıyla %77, %74.2, %77, %75.6, %74.9 ve %77.4 şeklindedir [10]. Çalışmalarında kullanmak üzere kendilerinin oluşturdukları daha kapsamlı olan veri kümesinde ise aynı modeller üzerindeki sonuçlar PID veri kümesinin sonuçlarına kıyasla daha iyi sonuçlar vermiştir [10]. Bu sonuçta göstermektedir ki kullanılan veri kümesinin içeriği ne kadar kapsamlı olursa modeller o kadar iyi sonuç vermektedir. Bir diğer çalışmada ise [11] makine öğrenmesi sınıflandırma algoritmaları kullanılarak yapılan şeker hastalığı tahmini çalışmasında yine aynı veri kümesi kullanılmıştır ve modellemelerden önce iki yeni özellik çıkarımı yapılmıştır. Bunun sonucunda kan basıncı değeri 80'in üzerinde ve glikoz seviyesi 105'in üzerinde olan herkes şeker hastası olarak tanımlanmıştır [11]. Yapılan çalışmada Rastgele Orman, K-En Yakın Komşu, Destek Vektör Makinesi, Karar Ağacı ve YSA (Yapay Sinir Ağı) algoritmaları ile tahmin modelleri oluşturulmuştur. Bu algoritmalar sonucunda %88.31, %81, %77, %84 ve %86 doğruluk oranları elde edilmiştir [11]. Aynı veri kümesinin kullanıldığı diğer bir çalışma da Chang ve ark. makine öğrenmesi algoritmalarına dayalı şeker hastalığı sınıflandırma çalışmasında J.48, Naive Bayes ve Rastgele Orman algoritmalarını kullanmışlardır [12]. Yapılan bu çalışmada araştırmacılar özellik seçiminin sınıflandırma modelleri üzerindeki etkisini incelemek için modelleri özellik seçimi olmadan, üç faktörlü ve beş faktörlü özellik seçimine bağlı olarak gösterdikleri sonuçları incelemişlerdir [12]. Çalışmada kullanılan algoritmalar arasında Rastgele Orman algoritmasının özellik seçimine tabi tutulmadığı durumda %79.57 doğruluk oranı ile diğer iki algoritmadan ve özellik seçimlerine tabi tutulan modellerden daha iyi sonuç verdiğini gözlemlemişlerdir [12]. Joshi ve ark. yaptığı PID veri kümesinin kullanıldığı diğer bir çalışmada ise [13] WEKA yazılım aracı kullanılarak şeker hastalığının sınıflandırılmasında kullanılan farklı yöntemlerin performans analizinin incelenmesi gerçekleştirilmiştir. Bu çalışmada Bayesnet, Naive Bayes, J48, Rastgele Orman, Rastgele Ağaç, REP Ağacı, CART (Sınıflandırma ve Regresyon Ağacı) ve K-En Yakın Komşu algoritmalarını kullanmışlardır. Bu algoritmaların sırasıyla şeker hastalığını doğru sınıflandırma oranlarının %78.25, %76.30, %84.11, %98.43, %100, %83.07, %77.21 ve %100 gözlemlenmiştir [13]. Aynı veri kümesi kullanılarak yapılan diğer bir çalışmada da Harman [14] Destek Vektör Makinesi ve Naive Bayes algoritmaları kullanılarak şeker hastalığı sınıflandırması yapmıştır. Yapılan çalışmada Harman veri kümesindeki dengesiz sınıf problemini çözmek için SMOTE tekniğini kullanmıştır [14]. Çalışma sonucunda Destek Vektör Makinesi %90 doğruluk, Naive Bayes algoritması ise %77 doğruluk oranı vermiştir [14]. Aynı veri kümesinin kullanıldığı farklı bir çalışmada Gua ve ark. şeker hastalığının tahmini için yalnızca Bayes ağlarını kullanmışlardır [15]. Çalışmanın sonucunda Naive Bayes ağı %71.5 doğruluk oranı verirken önermiş oldukları "Byes Network" isimli metot ile %72.3 doğruluk oranı elde etmişlerdir [15]. Farklı bir çalışmada ise şeker hastalığının tahmini için Er ve ark. ESA (Evrişimli Sinir Ağı) ve LSTM (Uzun/Kısa Süreli Bellek) modellerini kullanılmışlardır [16]. Araştırmacılar veri kümesi üzerinde iki deney yapmışlardır. İlk deneyde veri kümesinin %70 eğitim, %30 test verisi olarak bölmüşlerdir. İlk deneyin sonucunda ESA modelinde %82.47 doğruluk LSTM modelinde %83.77 doğruluk elde etmişlerdir. İkinci deneylerinde ise veri kümesini %80 eğitim, %20 test verisi olacak şekilde bölmüşlerdir ve bu deney sonucunda ESA modelinde %83.25 doğruluk LSTM modelinde ise %85.21 doğruluk elde etmişlerdir [16]. Aynı zamanda önermiş oldukları ESA+LSTM hibrit modeli veri kümesinin %70 eğitim, %30 test verisi olarak ayrıldığı deneyde %85.21 doğruluk oranı %80 eğitim, %20 test verisi olarak ayrıldığı deneyse ise %86.45 doğruluk oranı vermiştir [16]. PID veri kümesinin çalışmaya dâhil edildiği bir diğer çalışmada ise XGBoost ve Karar Ağacı tabanlı algoritmalar kullanılmıştır [17]. Yapılan bu çalışmada Türkiye'de on yıllık çalışma sonucunda oluşturulan bir şeker hastalığı veri kümesi de kullanılmıştır [17]. Karar Ağacı, Rastgele Orman, Gradient Boosting ve XGBoost algoritmalarının kullanıldığı bu çalışmada PID veri kümesine uygulanan algoritmalar daha yüksek doğruluk oranı vermiştir. Kullanılan algoritmaların PID veri kümesi üzerinde göstermiş oldukları doğruluk oranları sırasıyla %75.82, %81.05, %81.70 ve %82.35 şeklindedir [17]. Aynı veri kümesinin kullanıldığı farklı bir çalışma da ise Karegowda ve ark. alışılmışın dışında veri kümesini %60 eğitim ve %40 test verisi olacak şekilde bölerek kullanmışlardır. Yapılan bu çalışmada şeker hastalığının kural tabanlı sınıflandırmasında K Ortalama Kümeleme ve Karar Ağacı C4.5 algoritmalarını kullanmışlardır [18]. Yapılan çalışmanın sonucunda bu iki algoritmanın oluşturduğu hibrit modeli kullanılmışlardır ve önerilen bu model %93.33 doğruluk oranı vermiştir [18]. PID veri kümesinin çalışmanın bir kısmında ele alındığı bir diğer çalışmada Maniruzzaman ve ark. şeker hastalığının sınıflandırılması için karşılaştırılmalı yaklaşımlar önermişlerdir [19]. Önerdikleri bu yaklaşımlardan birisi olan Gauss süreci (Gaussian Process) tabanlı model PID veri kümesi üzerinde %81.97 doğruluk oranı vermiştir [19]. Deperlioğlu ve ark. yaptığı bir çalışma ise aynı veri kümesi

üzerinde derin sinir ağları tabanlı Oto Kodlayıcı Sinir Ağları (OKSA) kullanılmıştır [20]. Kullandıkları bu OKSA modeli ile %97.30 doğruluk oranı elde edilmiştir [20]. Çalışmalarında OKSA modelini tercih etmelerinin nedeni ise herhangi bir optimizasyon işlemi yapmadan sınıflandırma işleminin başarısının arttırılabileceğini göstermektir [20]. Karşılaştırmalı olarak gerçekleştirilen bir başka çalışmada Cihan ve ark. Lojistik Regresyon, K-En Yakın Komşu, Destek Vektör Makinesi, Gauss Naive Bayes, Karar Ağacı, Rastgele Orman ve Yapay Sinir Ağı modellerini kullanılmışlardır [21]. 10 katlı çapraz doğrulama yöntemi ile modellerin değerlendirilmesi sonucunda Lojistik Regresyon modelinin diğer modeller arasındaki en iyi sonucu verdiği görülmüştür. Modelin performans ölçütleri olarak kesinlik, duyarlılık, ROC ve PRC değerleri ölçülmüştür. Bu değerler sırası ile 0.76, 0.77, 0.83 ve 0.83 şeklindedir [21]. Aynı veri kümesinin kullanıldığı farklı bir çalışmada ise [22] Kumar ve ark. sınıflandırma tahmini için Derin Sinir Ağı (DNN) sınıflandırıcısı kullanmıştır. Kullanılan modelin, %98.16 doğruluk oranı ile literatürde yer alan çalışmalara göre daha iyi sonuç verdiğini fakat hesaplama süresinin çalışmanın ana sınırlanması olduğunu bundan dolayı ilerleyen çalışmalarda hesaplama süresi için optimizasyon çalışmaları ile hesaplama süresinin iyileştirilebilmesi üzerine çalışmalar yapılmasının çalışmayı daha etkin bir hale getireceğini belirtmişlerdir [22]. PID veri kümesinin kullanıldığı son çalışmalardan birisini Yıldırım ve ark. gerçekleştirmiştir. Bu çalışmada [23] Apache Spark teknolojisinden faydalanarak Yapay Sinir Ağları, Destek Vektör Makinesi, Lojistik Regresyon, Rastgele Orman ve Naive Bayes sınıflandırma algoritmalarını kullanmışlardır. Kullandıkları sınıflandırma modellerini literatürde yer alan çalışmalardan farklı olarak kesinlik, duyarlılık, negatif tahmin değeri, f ölçüsü, doğruluk, özgünlük ve hata oranı olmak üzere yedi farklı ölçüm değeri üzerinden değerlendirmişlerdir [23]. Kullandıkları sınıflandırma modelleri arasında en yüksek ortalama doğruluk oranını veren model %85.04 ile Rastgele Orman modeliyken en düşük ortama doğruluk oranını veren sınıflandırma modeli ise %75.87 doğruluk oranı ile Naive Bayes modeli olmuştur [23].

3 Materyal ve metot

Bu bölümde ise çalışmanın gerçekleştirilmesinde kullanılan yaklaşımlardan ve kullanılan veri kümesinden bahsedilecektir. Çalışma esnasında "Pima Indians Diabetes" veri kümesi hakkında detaylı bilgi sahibi olabilmek amacıyla eksik değer analizi, eşsiz değer analizi, sayısal özellik analizi, box plot analizi, swarm plot analizi, cat plot analizi ve korelasyon analizi gerçekleştirilmiştir. Veri analizi sonrasında veri kümesinin tahmin modelleri için hazır hale getirilmesi amacıyla eksik değerler ortalama değer ile değiştirilmiş, aykırı değerler veri kümesinden çıkarılmış ve var olan örneklerden özellik çıkarımı yapılmıştır. One Hot Encoding işlemi ile kategorik veriler sayısal hale dönüştürülmüştür. Son olarak ise veriler üzerinde standardizasyon işlemi uygulanmıştır. Bu işlem adımlarını takiben sınıflandırma aşamasına geçilmiştir. Modelleme esnasında dışarıda tutma (Hold-Out , %80 eğitim - %20 test) ve 10 katlı çapraz doğrulama yöntemleri kullanılmıştır. Çalışma kapsamında Lojistik Regresyon, K-En Yakın Komşu, CART (Sınıflandırma ve Regresyon Karar Ağacı), Destek Vektör Makinesi, Rastgele Orman, XGBoost ve LightGBM modelleri kullanılmıştır.

3.1 Veri kümesi

Gerçekleştirilen çalışma kapsamında kullanılan veri kümesi Kaggle platformunda açık kaynak olarak yer almaktadır. Amerika Birleşik Devletleri Ulusal Sağlık Enstitülerinin bir parçası olan Ulusal Diyabet Sindirim ve Böbrek Enstitüsü'nden (NIDDK) elde edilen bu veriler 21 yaş ve üstü kadınlardan toplanmıştır [24]. Veri kümesinde 768 hasta kadına ait veriler yer almaktadır. Tablo 1 veri kümesinde yer alan sekiz adet özellik ve bir adet hedef değişkeni göstermektedir.

Tablo 1. Veri kümesinde yer alan özellikler, açıklamaları ve değer aralıkları

Özellikler	Açıklama	Değer Aralığı
Pregnancies (Gebelik)	Hamile kalma Sayısı	0-17 arası
Glucose (Glukoz)	Kan glukoz değeri (2 saatlik şeker yükleme testi)	0-199 arası
Blood Pressure (Kan Basıncı)	Kan Basıncı	0-122 (mm/Hg) arası
Skin Thickness (Cilt Kalınlığı)	Deri Kalınlığı	0-99 (mm) arası
Insulin (İnsülin)	2 saatlik kan insülin serum değeri	0-846 (mu U/ml) arası
BMI (Vücut Kitle İndeksi)	Vücut Kitle İndeksi	(kg ve m2)
Diabetes Pedigree Function (Diyabet Soyağacı Fonksiyonu)	Kişinin şeker hastalığına genetik olarak yatkınlık durumu	0.078-2.42 arası
Age (Yaş)	Kişinin yaşı	21-81 yıl arası
Outcome (Çıktı)	Hedef Değişken	0 veya 1

3.2 Veri kümesi analizi

Veri kümesi hakkında daha detaylı bilgi sahibi olabilmek amacıyla kayıp değer analizi, eşsiz değer analizi, sayısal özellik analizi, box plot analizi, swarm plot analizi, cat plot analizi ve korelasyon analizi gibi analizler gerçekleştirilmiştir. Yapılan analizler sonucunda veri kümesinde eksik değer olmadığı fakat bazı özelliklerin değerlerinde sorunların olduğu gözlemlenmiştir. Örneğin "Insulin" özelliğinde değerlerin çoğunun sıfır (0) olduğu görülmüştür. Fakat veri kümesinde "Pregnancies" ve "Outcome" özellikleri dışında yer alan özelliklerin sıfır (0) değerini alması mümkün olamayacağından dolayı bu durum kullanılan veri kümesinde eksik değerlerin olduğunu ve bu

eksik değerlerin giderilmesi amacıyla eksik değerler yerine sıfır (0) değerlerinin girildiği belirlenmiştir. Ayrıca box plot analizi ve swarm plot analizi sonrasında veri kümesinde aykırı değerler olduğu belirlenmiştir. Bu durumun giderilmesi için veri kümesinde "Pregnancies" ve "Outcome" özellikleri dışında yer alan özelliklerdeki sıfır (0) değerleri Outcome sınıf değişkenine göre ortalama değer ile değiştirilmiştir.

3.3 Veri kümesini hazırlama

Veri analizi bölümünde bahsedilen analiz yöntemleri sonrasında analiz sırasında fark edilen problemlerin giderilmesi ve veri kümesinin makine öğrenmesi modellerinin uygulanması için uygun hale getirilmesi bu aşamada gerçekleştirilmiştir.

3.3.1 Eksik değerlerin ortalama değer ile değiştirilmesi

Veri kümesinde yer alan eksik değerlerin giderilmesi için veri kümesindeki "Pregnancies" ve "Outcome" özellikleri dışında yer alan tüm özelliklerdeki sıfır (0) değerleri Outcome sınıf değişken kırılımına göre ortalama değer ile değiştirilmiştir.

3.3.2 Aykırı değerlerin çıkarılması

Box plot analizi ve swarm plot analizinde veri kümesinde bazı aykırı değerler olduğu görüldüğü için veri kümesindeki alt sınır ve üst sınır değer dışında yer alan aykırı değerler veri kümesinden çıkarılarak veri kümesi (768,9) boyutundan (701,9) boyutuna getirildi.

3.3.3 Özellik çıkarımı

Son olarak ise var olan "Glucose", "BMI" ve "Insulin" özelliklerinden özellik çıkarımı yapılarak yeni özellikler oluşturuldu. Tablo 2, Tablo 3 ve Tablo 4'te oluşturulan yeni özellikler ve değer aralıkları yer almaktadır.

Tablo 2. Glucose özelliğinden çıkarılan özellikler

Glukoz
Glukoz değeri 70' e eşit veya 70'den küçük ise "Low"
Glukoz değeri 70'den büyük, 99'a eşit veya 99'dan küçük ise "Normal"
Glukoz değeri 99'dan büyük, 126'ya eşit veya 126'dan küçük ise "Secret"
Glukoz değeri 126'dan büyük ise "High"

Tablo 3. Insulin özelliğinden çıkarılan özellikler

Insulin
İnsülin değeri 16'ya eşit veya büyükse ve 166'ya eşit ve küçük ise küçük "Normal"
İnsülin değeri 166'dan büyük ise "Abnormal"

Tablo 4. BMI özelliğinden çıkarılan özellikler

BMI
Vücut kütle indeksi 18.5'ten küçük ise "Underweigh"
Vücut kütle indeksi 18.5'ten büyük ve 24.9'a eşit veya küçük ise "Normal"
Vücut kütle indeksi 24.9'dan büyük ve 29.9'a eşit veya küçük ise "Overweight"
Vücut kütle indeksi 29.9'dan büyük ve 34.9'a eşit veya küçük ise "Obesity 1"
Vücut kütle indeksi 34.9'dan büyük ve 39.9'a eşit veya küçük ise "Obesity 2"
Vücut kütle indeksi 39.9'dan büyük ise "Obesity 3"

3.3.4. One hot encoding yöntemi

Oluşturulan bu özellikler arasındaki kategorik değişkenler One-Hot-Encoding yöntemi ile sayısal değerler haline getirilmiştir.

3.3.5. Standardizasyon

Şeker hastalığının analiz edilmesi için oluşturulacak olan tahmin modellerinden önce verilere standardizasyon işlemi uygulanarak, veriler sıfır (0) ortalamalı ve bir (1) standart sapmalı hale dönüştürülmüştür.

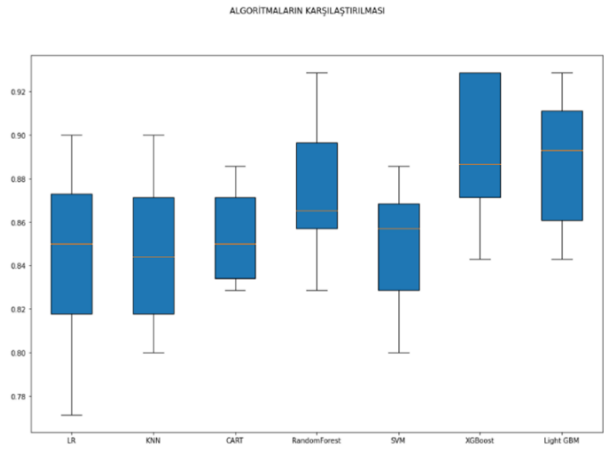
4. Deneysel sonuçlar

Bu çalışma kapsamında veri kümesi %80 eğitim ve %20 test veri kümesi şeklinde bölünerek modellerde iki farklı yöntem ile doğruluk skor değerleri incelenmiştir. Kullanılan tahmin modellerin, dışarıda tutma yöntemi ve 10 katlı çapraz doğrulama sonucunda oluşan doğruluk değerleri Tablo 5'deki gibidir.

Tablo 5. Veri kümesinde yer alan özellikler, açıklamaları ve değer aralıkları

Modeller	**Dışarıda Tutma Yöntemi Doğruluk Oranları**	**10 Katlı Çapraz Doğrulama Doğruluk Oranları**
Lojistik Regresyon (LR)	%84.39	%84.58
K En Yakın Komşu (KNN)	%82.97	%84.59
Sınıflandırma ve Regresyon Ağacı (CART)	%80.14	%85.02
Rastgele Orman(RF)	%82.97	%88.29
Destek Vektör Makinesi (SVM)	%79.43	%84.73
XGBoost	%83.68	%89.29
LightGBM	%85.10	%88.72

10 katlı çapraz doğrulama sonucunda elde edilen doğruluk oranları dikkate alındığında en iyi doğruluk oranını veren modellerin Rastgele Orman, XGBoost ve LightGBM modelleri olduğu görülmüştür. Bu sebep ile bu çalışma kapsamında bu üç modelin performansını artırmaya yönelik Hiper-parametre ayarlaması Pyhton Sklearn kütüphanesi GridSearchCV yöntemi ile gerçekleştirilmiştir. Tablo 6, Tablo 7 ve Tablo 8'de final modeli oluşturulacak modeller için ayarlanan parametreler ve parametre değerleri yer almaktadır.



Şekil. 1. 10 Katlı çapraz doğrulama sonucunda modellerin karşılaştırılması

Tablo 6. Rastgele orman modeli için GridSearchCV ile bulunan en iyi parametreler

Parametreler	Değerler
n_estimators	100, 200, 500, 1000
max_features	3, 5, 7
n_samples_split	2, 5, 10, 30
max_depth	3, 5, 8, None

Tablo 7. XGBoost modeli için GridSearchCV ile bulunan en iyi parametreler

Parametreler	Değerler
learning_rate	0.01, 0.1, 0.2, 1
min_samples_split	0.1, 0.5, 3
max_depth	3, 5, 8
subsample	0.5, 0.9, 1.0
n_estimators	100, 500

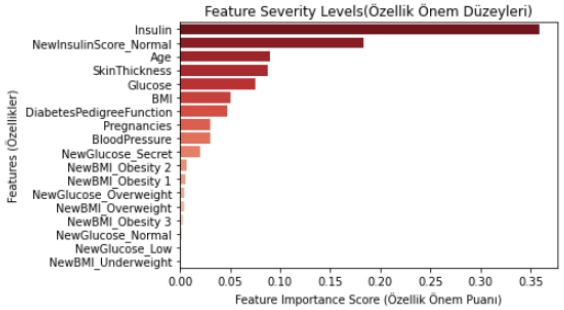
Tablo 8. LightGBM modeli için GridSearchCV ile bulunan en iyi parametreler

Parametreler	Değerler
learning_rate	0.01, 0.03, 0.05, 0.1, 0.5
n_estimators	500, 1000, 1500
max_depth	3, 5, 8

GridSearchCV yönteminin 10 katlı çapraz doğrulama sonrasında final modelleri için elde edilen en iyi parametreleri Tablo 9, Tablo 10 ve Tablo 11'de gösterilmiştir. Şekil 2, Şekil 3 ve Şekil 4' de ise final modellerinin özellik önem düzeyleri gösterilmiştir.

Tablo 9. Rastgele orman modeli için GridSearchCV ile bulunan en iyi parametreler

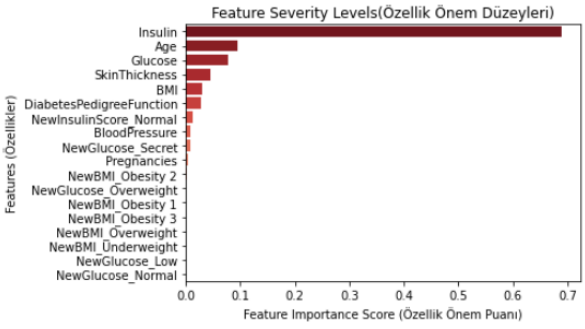
Parametreler	Değerler
n_estimators	100
max_features	7
min_samples_split	2
max_depth	8



Şekil. 2. Rastgele Orman modeli özellik önem düzeyi

Tablo 10. XGBoost modeli için GridSearchCV ile bulunan en iyi parametreler

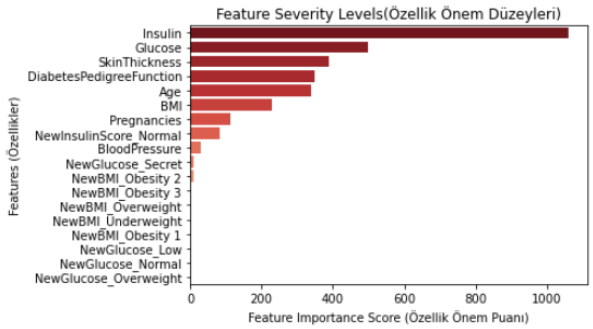
Parametreler	Değerler
learning_rate	0.01
min_samples_split	0.1
max_depth	5
subsample	1.0
n_estimators	500



Şekil. 3. XGBoost modeli özellik önem düzeyi

Tablo 11. LightGBM modeli için GridSearchCV ile bulunan en iyi parametreler

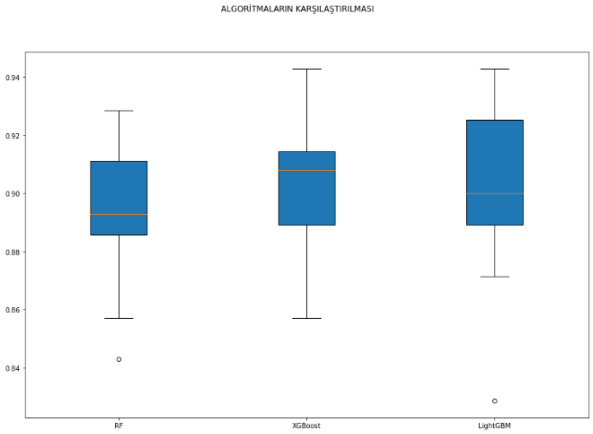
Parametreler	Değerler
learning_rate	0.01
n_estimators	500
max_depth	3



Şekil. 4. LightGBM modeli özellik önem düzeyi

Tablo 12. Final modelleri ve doğruluk oranları

Modeller	Doğruluk
Rastgele Orman	% 89.30
XGBoost	% 90.01
LightGBM	% 90.01



Şekil. 5. Final modellerinin karşılaştırılması

5 Sonuçlar

Bu çalışma kapsamında şeker hastalığının doğru teşhisinde kullanılmak üzere en uygun model belirlenmeye çalışılmıştır. Bu amaçla açık kaynak veri kümesi olan "Pima Indians Diabetes" isimli veri kümesi kullanılmıştır. Çalışma kapsamında ilk önce kayıp değer analizi, eşsiz değer analizi, sayısal özellik analizi, box plot analizi, swarm plot analizi, cat plot analizi ve korelasyon analizi gibi analiz yöntemleri ile analizler gerçekleştirilerek veri kümesi incelenmiştir. Veri kümesinde eksik değer olmadığı fakat bazı özelliklerin değerlerinde sorunların olduğu gözlemlenmiştir. Aynı zamanda aykırı değerlerin çıkarılarak veri kümesi daha temiz hale dönüştürülmüştür. Analizler tamamlandıktan sonra var olan "Glucose", "BMI" ve "Insulin" isimli özelliklerinden özellik çıkarımı yapılarak yeni özellikler oluşturulmuştur. Elde edilen öznitelik değerleri "one hot encoding" işlemi ile sayısal işleme çevrilmiş ve veri kümesi analiz ve iyileştirme işlemleri tamamlanmıştır. Son olarak sınıflandırma aşamasında Lojistik Regresyon, K-En Yakın Komşu, CART (Sınıflandırma ve Regresyon Ağacı), Rastgele Orman, Destek Vektör Makinesi, XGBoost ve LightGBM sınıflandırıcı modelleri kullanılmıştır. Kullanılan bu modeller 10 katlı çapraz doğrulama yöntemi ile test edilmiş ve performans ölçütleri hesaplanmıştır. Modeller arasında en iyi üç doğruluk oranını veren Rastgele Orman, XGBoost ve LightGBM modellerinde hiperparametre ayarlaması yapılarak elde edilen parametreler ile final modelleri belirlenmiştir. Final modellerinin doğruluk oranları sırasıyla %89.30, %90.01 ve %90.01 olarak gözlemlenmiştir. XGBoost ve LightGBM modellerinin umut vaat eden sonuçlar verdiği gözlemlenmiş ve final modelleri olarak kullanılabileceği önerilmiştir.

Çalışmamız diğer çalışmalar ile kıyaslandığında daha iyi sonuçlar elde edildiği gözlemlenebilmektedir. Aynı veri kümesinin kullanıldığı çalışmalarda elde edilen sonuçların farklı olması, veri kümesinde yer alan verilerden eksik olanların giderilmesi, aykırı olan verilerin ne kadarının veri kümesinden çıkarıldığına veya veri kümesinde yer alan özelliklerden yeni özelliklerin oluşturulmasına bağlı olan ya da modellerin hiper parametre ayarlarının yapılmasındaki oluşturduğu farklar olarak değerlendirilebilir.

Yapmış olduğumuz bu çalışmada literatür de yer alan çalışmalardan farklı olarak eksik değerlerin sınıf çıktı değişkeninin ortalaması ile değiştirilmesi sonucunda modellerin performansının büyük oranda arttığı gözlemlenmiştir. Aynı zamanda box plot analizi ve swarm plot analizinde görülen bazı aykırı değerler veri kümesinden çıkarılarak veri kümesinin daha temiz veri kümesi olmasına katkı sağlamıştır.

Gelecek çalışmalarda bu çalışma kapsamında kullanılan veri kümesindeki eksik değer sorunun giderilmesi için yeni yaklaşımlar geliştirilerek sınıflandırma modellerinin performansının artırılması planlanmaktadır.

Çıkar çatışması

Yazarlar çıkar çatışması olmadığını beyan etmektedir.

Benzerlik oranı (iThenticate): %11

Kaynaklar

[1] B. Ö. Başer, M. Yangın, ve E. S. Sarıdaş, Makine öğrenmesi teknikleriyle diyabet hastalığının sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(1), 112–120, 2021. https://doi.org/10.19113/sdufenbed.842460.

[2] Diabetes mellitus ve komplikasyonlarının tanı, tedavi ve izlem kılavuzu, https://file.temd.org.tr/Uploads/publications/guides/documents/diabetes-mellitus_2022.pdf, Accessed 09 May, 2022.

[3] N. Eroğlu, Diabetes Mellitus'un komplikasyonları. İzmir Demokrasi Üniversitesi Sağlık Bilimleri Dergisi, 1(2), 6-12, 2018.

[4] K. Akyol ve A. Karacı, Diyabet hastalığının erken aşamada tahmin edilmesi için makine öğrenme algoritmalarının performanslarının karşılaştırılması. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 9(6), 123–134, 2021, https://doi.org/10.1016/10.29130/dubited.1014508.

[5] F. Al-Areqi and M. Z. Konyar, Effectiveness evaluation of different feature extraction methods for classification of Covid-19 from computed tomography

images: A high accuracy classification study. Biomedical Signal Processing and Control, 76, 2022, https://doi.org/10.1016/j.bspc.2022.103662.

[6] F. Al-Areqi and M. Z. Konyar, transfer öğrenme mimarileri kullanılarak bilgisayarlı tomografi görüntülerinden Covid-19'un yüksek doğrulukla sınıflandırılması. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, 13(3), 457-466, 2022, https://doi.org/10.24012/dumf.1129870.

[7] Ü. Veranyurt, A. F. Deveci, ve M. F. Esen, Makine öğrenmesi teknikleriyle hastalık sınıflandırması: Random Forest, K-Nearest Neighbour ve Adaboost algoritmaları uygulaması. Uluslararası Sağlık Yönetimi ve Araştırmaları Dergisi, 6(2), 275–286, 2020.

[8] Y. Özkan, B. S. Yürekli, ve A. Suner, Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması. Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 12(1), 211–226, 2021, https://doi.org/10.17714/gumusfenbil.820882.

[9] G. Bilgin, Makine öğrenmesi algoritmaları kullanarak erken dönemde diyabet hastalığı riskinin araştırılması. Zeki Sistemler Teori ve Uygulamaları Dergisi, 4(1), 55–64, 2021, https://doi.org/10.38016/jista.877292.

[10] N. P. Tigga and S. Garg, Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science., 167, 706–716, 2020, https://doi.org/10.1016/j.procs.2020.03.336.

[11] S. Nahzat ve M. Yağanoğlu, Diabetes prediction using machine learning classification algorithms. European Journal of Science and Technology, 24, 53–59, 2021, https://doi.org/10.31590/ejosat.899716.

[12] V. Chang, J. Bailey, Qianwen, A. Xu, and Z. Sun, Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms, https://doi.org/10.1007/s00521-022-07049-z.

[13] S. Joshi and S. R. P. Shetty, Performance analysis of different classification methods in data mining for diabetes dataset using WEKA tool. International Journal on Recent and Innovation Trends in Computing and Communication, 3(3), 1168–1173, 2015, https://doi.org/10/10.17762/ijritcc2321-8169.150361.

[14] G. Harman, Prediction of Diabates Mentilus by using SVM and Naive Bayes classification algorithms. European Journal of Science and Technology, 32, 7–13, 2021, https://doi.org/10.31590/ejosat.1041186.

[15] Y. Guo, G. Bai, and Y. Hu, Using Bayes Network for prediction of type-2 diabetes. International Conference for Internet Technology and Secured Transactions (ICITST 2012), pp. 471–476, London, England, 2012.

[16] M. B. ER ve İ. Işık, LSTM tabanlı derin ağlar kullanılarak diyabet hastalığı tahmini. Türk Doğa ve Fen Dergisi, 10(1), 68–74, 2021, https://doi.org/10.46810/tdfd.818528.

[17] G. Yangın, XGboost ve Karar Ağacı tabanlı algoritmaların diyabet veri setleri üzerine uygulaması. Yüksek Lisans Tezi, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü, Türkiye, 2019.

[18] A. G. Karegowda, V. Punya,, M. A. Jayaram, and A. S. Manjunath, Rule based classification for diabetic patients using Cascaded K-Means and Decision Tree C4 . 5. International Journal of Computer Applications, 45(12), 45–50, 2012, https://doi.org/10.5120/6836-9460.

[19] M. Maniruzzaman, N. Kumar, M. M. Abedin, M. S. Islam, H. S. Suri, A.s El-Baz, J. S. Suri, Comparative approaches for classification of Diabetes Mellitus data: Machine learning paradigm. Computer Methods and Programs in Biomedicine. 152, 23–34, 2017, https://doi.org/10.1016/J.CMPB.2017.09.004.

[20] Ö. Deperlioğlu ve U. Köse, Derin Sinir Ağları kullanarak diabet teşhisi., 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1–4, Ankara, Türkiye, 2018.

[21] P. Cihan and H. Coskun, Performance comparison of machine learning models for diabetes prediction, 29th Signal Processing and Communications Applications Conference (SIU'2021), pp. 26–30, İstanbul, Türkiye, 2021,

[22] P. B. M. Kumar, R. S. Perumal, R. K. Nadesh, and K. Arivuselvan, Type 2: Diabetes Mellitus prediction using Deep Neural Networks classifier. International Journal of Cognitive Computing in Engineering, 1, 55–61, 2020, https://doi.org/10.1016/j.ijcce.2020.10.002.

[23] E. Yıldırım and A. Çalhan, Machine learning supported diabetes prediction with Apache Spark. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 10(3), 1107–1117, 2022, https://doi.org/10.29130/dubited.999048.

[24] Pima Indians Diabetes Database | Kaggle, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/, Accessed 09 May, 2022.