Türkiye Diyabet ve Obezite Dergisi / *Turkish Journal of Diabetes and Obesity*
Özgün Araştırma / Original Article

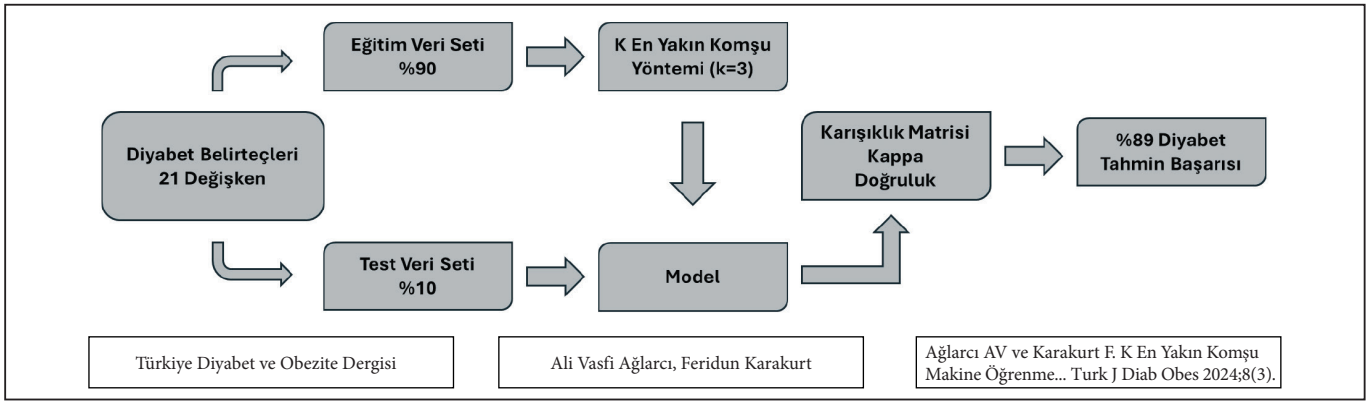
K En Yakın Komşu Makine Öğrenme Algoritmasına Dayalı Diabetes Mellitus Tahmini

Ali Vasfi AĞLARCI[1], Feridun KARAKURT[2]

[1]Kastamonu Üniversitesi, Tıp Fakültesi Biyoistatistik Anabilim Dalı, Kastamonu, Türkiye
[2]Necmettin Erbakan Üniversitesi, Tıp Fakültesi Dahili Tıp Bilimleri Bölümü, İç Hastalıkları Anabilim Dalı, Konya, Türkiye

Bu makaleye yapılacak atıf: Ağlarcı AV ve Karakurt F. K en yakın komşu makine öğrenme algoritmasına dayalı diabetes mellitus tahmini. Turk J Diab Obes 2024;8(3): 265-276.

GRAFİKSEL ÖZET



ÖZ

Amaç: Çalışmamızın amacı dünya çapında giderek artan ve önemli bir halk sağlığı sorunu hâline gelen diabetes mellitus hastalığının makine öğrenme yöntemi ile tahmin edilmesidir.

Gereç ve Yöntemler: Çalışmada diabetes mellitus sağlık göstergelerini içeren ve kaggle veri tabanından elde edilen 253.680 örnek hacmine sahip veri kayıtları kullanılmıştır. K en yakın komşu yöntemi ile hastaların diabetes mellitus durumları makine öğrenme yaklaşımıyla tahmin edilmeye çalışılmıştır. Tüm işlemler R programı ile gerçekleştirilmiştir.

Bulgular: Kişilerin yaklaşık %15,8'i preDM ya da diabetes mellitus tanılıdır, %42,9'unde yüksek tansiyon, %42,4'ünde yüksek kolesterol bulunmaktadır. Sigara içenlerin oranı %44,3, ağır alkol tüketenlerin oranı ise %5,6'dır. Kalp hastalığı/krizi geçirenleri oranı ise %9,4, yürüyüşte zorluk çektiğini bildirenlerin oranı ise %16,8'dir. Fiziksel aktivitesi bulunmayanların oranı %24,4'tür. Diabetes mellitus tanısı olmayanların BMI ortalaması 27,74±6,26 iken diyabet hastası olanların BMI ortalaması 31,94±7,36 olarak bulunmuştur. K en yakın komşu yöntemi ile yapılan uygulamada diabetes mellitus tahmini en iyi eğitim ve test verisinin %90,0-%10,0 olarak ayrıldığı ve K komşuluk değerinin 3 (üç) alındığı durumda elde edilmiştir. İlgili belirteçler kullanılarak %97,2 doğruluk ve %88,9 kappa başarı değeri ile diabetes mellitus hastalığına sahip kişiler doğru tahmin edilebilmiştir.

Sonuç: Makine öğrenme yöntemlerinin son yıllarda birçok alanda kullanımının yaygınlaştığı ve başarılı sonuçlar verdiği literatürde bildirilmektedir. Bu araştırmada da makine öğrenme yaklaşımıyla diabetes mellitus tahmininin yüksek başarı oranı ile gerçekleştirildiği uygulamalı olarak gösterilmiştir. Diabetes mellitus hastalığının sessiz ve artan sayıda ilerlediği bilindiğinden erken tanı hayati öneme sahiptir. K en yakın komşu yönteminin kolay uygulanabilirliği ve yüksek sınıflama performansı gibi avantajlarından dolayı diabetes mellitus hastalığının erken tanı ve tedavisi için sağlık hizmeti sağlayıcıları tarafından kullanılması önerilmektedir.

Anahtar Sözcükler: *Diabetes mellitus, Sağlık ve hastalık, Makine öğrenmesi, Tahmin, Sağlık uygulamaları, Akıllı sistem*

ORCID: Ali Vasfi Ağlarcı / 0000-0002-9010-4537, Feridun Karakurt / 0000-0001-7629-9625

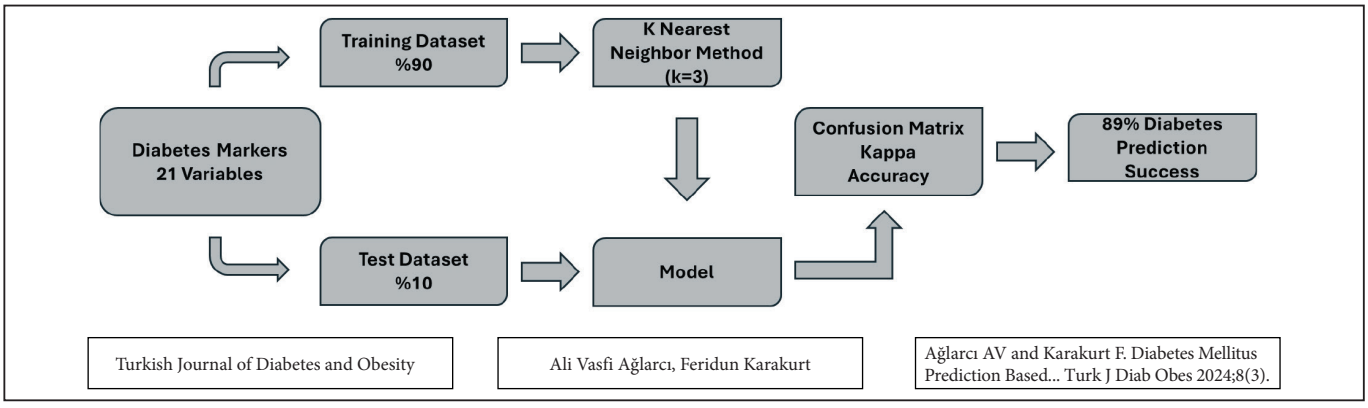
Yazışma Adresi / Correspondence Address:
Ali Vasfi AĞLARCI
Kastamonu Üniversitesi, Tıp Fakültesi Biyoistatistik Anabilim Dalı, Kastamonu, Türkiye
E-posta: avaglarci@kastamonu.edu.tr

DOI: 10.25048/tudod.1549498

Geliş tarihi / *Received* : 13.09.2024
Revizyon tarihi / *Revision* : 16.10.2024
Kabul tarihi / *Accepted* : 19.12.2024

Diabetes Mellitus Prediction Based on K Nearest Neighbor Machine Learning Algorithm

GRAPHICAL ABSTRACT



ABSTRACT

Aim: The aim of our study is to predict diabetes mellitus, which is increasing worldwide and has become an important public health problem, with machine learning method.

Material and Methods: In the study, data records containing diabetes mellitus health indicators with a sample size of 253,680 obtained from the kaggle database were used. K nearest neighbor method was used to predict the diabetes mellitus status of the patients with a machine learning approach. All operations were performed with the R program.

Results: Approximately 15.8% of the individuals were diagnosed with preDM or diabetes mellitus, 42.9% had high blood pressure and 42.4% had high cholesterol. 44.3% were smokers and 5.6% were heavy alcohol consumers. The rate of those who have had heart disease/crisis is 9.4%, and the rate of those who reported having difficulty in walking is 16.8%. The rate of those with no physical activity was 24.4%. The mean BMI of those without diabetes mellitus was 27.74±6.26, while the mean BMI of those with diabetes mellitus was 31.94±7.36. In the application with the k nearest neighbor method, the best prediction of diabetes mellitus was obtained when the training and test data were separated as 90.0%-10.0% and the k neighborhood value was 3 (three). Using the relevant markers, people with diabetes mellitus disease were correctly predicted with 97.2% accuracy and 88.9% kappa success value.

Conclusion: It is reported in the literature that machine learning methods have been widely used in many fields in recent years and have yielded successful results. In this study, it has been demonstrated that the prediction of diabetes mellitus with machine learning approach is realized with a high success rate. Since diabetes mellitus is known to progress silently and in increasing numbers, early diagnosis is of vital importance. Due to the advantages of K nearest neighbor method such as easy applicability and high classification performance, it is recommended to be used by healthcare providers for early diagnosis and treatment of diabetes mellitus.

Keywords: *Diabetes mellitus, Health and disease, Machine learning, Prediction, Healthcare applications, Intelligent system*

GİRİŞ

Diabetes mellitus, pankreastan insülin üretiminin yetersiz olması ya da periferik dokularda insülin etkisine karşı direnç gelişmesi nedeniyle ortaya çıkan ve kan dolaşımında yüksek glikoz seviyeleri ile karakterize kronik bir metabolizma hastalığıdır. Önemli bir halk sağlığı sorunu olan diabetes mellitus, akut ve kronik komplikasyonlarla seyreden, sürekli tıbbi ve öz bakım gerektiren küresel yaygınlığa sahip kronik bir hastalıktır (1). Hem gelişmiş hem de gelişmekte olan ülkelerde önde gelen bir ölüm nedeni hâline gelmiş ve dünya çapında giderek artan sayıda kişiyi etkilemektedir (2). Ayrıca diabetes mellitus, hastanın ölümüne yol açabilecek birkaç başka ciddi hastalığa da yol açabilir (3). Modern yaşam tarzı diabetes mellitus hastalığının görülme sıklığını önemli ölçüde artırmıştır (4). Dünya yetişkin nüfusunun on birinden biri (yaklaşık 537 milyon) diabetes mellitus tanısı almıştır. Tanı konulmamış diyabetli sayısının yaklaşık 240 milyon olduğu tahmin edilmektedir. Bu da yetişkinlerin hastalık durumlarından farkında olmadığını göstermektedir. Bazı çalışmalar bu yüzden hastalığı sessiz katil olarak tanımlamıştır (5). Diabetes mellitus kaynaklı ölümlerin yıllık 6,7 milyona ulaştığı bildirilmiştir. Diabetes mellitus tanılı birey sayısının yıllar itibariyle artacağı, 2030 yılında 643 milyon, 2045 yılında 783 milyon olacağı tahmin edilmektedir. Türkiye'de ise

diabetes mellitus tanılı yetişkin bireylerin sayısı 2021 yılında 9 milyon olup, 2045 yılında yaklaşık 13,4 milyon olacağı ilgili raporlarda belirtilmiştir (1,6).

Diabetes mellitus ile ilgili risk faktörlerinin belirlenmesi, hastalığın önlenmesi ve tedavilerin geliştirilmesi amacıyla epidemiyolojik çalışmalar yürütülmektedir. Son yıllarda bu çalışmalar arasında ön plana çıkanlar ise teşhis ve tedavinin erken tahminine olanak sağlayan makine öğrenmesi çalışmalarıdır. Makine öğrenmesi, bilgisayar programının girilen verilerden öğrendiği ve sonrasında yeni gözlemleri sınıflandırmak için bu öğrenmeyi kullandığı öğrenme yaklaşımı olarak tanımlanır (7,8). Yapay zekânın alt kümesi olan makine öğreniminin sağlamış olduğu öngörü, tahmin, kümeleme ve sınıflama başarısı, son yıllarda çeşitli alanlarda (tıp, mühendislik, ziraat vb.) kullanımını artırmıştır (9).

Makine öğrenme yöntemleri, hekimlere hastalığı erken teşhis etmek, zamanında müdahalelerde bulunmak ve hasta sonuçlarını iyileştirmek için değerli araçlardır (2). Bu çerçevede diabetes mellitus hastalığının erken teşhisi ve önlenmesi de hayati önem taşır (10). Diabetes mellitus oluşumunun tahmini, risk altında olan bir kişinin hastalığın başlangıcını önleyebilecek veya ilerlemesini geciktirebilecek eylemlerin bulunmasını sağlar (11). Makine öğrenmesi erken tespit için iyi bir tahmin yöntemidir. Literatürdeki çalışmalar incelendiğinde makine öğrenme yöntemleri ile farklı değişkenler aracılığıyla diabetes mellitus hastalığının erken tanı ve teşhisine yönelik çalışmalar yürütülmüştür (12-15).

Diabetes mellitus tanısının pozitif mi negatif mi olduğunun belirlenmesine yönelik yapılan ikili sınıflandırma çalışmasında destek vektör makineleri ve yapay sinir ağları kullanılmış ve %94,87'lik bir tahmin doğruluğuna ulaşıldığı bildirilmiştir (16). Diabetes mellitus risk tahmini amacıyla yürütülen araştırmada makine öğrenme yöntemleri kullanılmış ve 16 değişken kullanılarak rastgele orman yöntemiyle yüksek tahmin performansı (%98,6 doğruluk) elde edildiği bildirilmiştir (4). Diabetes mellitus gelişiminin tahminini konu alan çalışmada makine öğrenme yöntemleri aracılığıyla %81,0 doğru tahmin gücüne ulaşıldığı belirtilmiştir. Çalışmada ayrıca diabetes mellitus hastalığının gelecekteki gelişimiyle yüksek oranda ilişkili olan en iyi özellikleri bulmak hedeflenmiştir (10). Erken evre diabetes mellitus risk tahmininin yapıldığı başka bir çalışmada farklı makine öğrenme yöntemlerinin kullanıldığı ve sinir ağları modeliyle %99,2 doğru tahmin performansı elde edildiği ifade edilmiştir (5). Makine öğrenmesine dayalı tip 2 diabetes mellitus tahminine yönelik yapılan çalışmada Kore elektronik sağlık kayıtları kullanılarak rastgele orman, destek vektör makinesi, rastgele orman, XGBoost gibi makine öğrenme algoritmaları ile kişilerin diabetes mellitus hastalık sonuçları tahmin edilmiştir. Modellerin üstün performans gösterdiği, klinisyenlere ve hastalara tip 2 diabetes mellitus geliştirme olasılığı hakkında değerli tahmin bilgisi sağladığı belirtilmiştir (11).

Tıbbi teşhislerde makine öğrenimi tekniklerini kullanan birçok çalışma olmasına rağmen, özellikle diabetes mellitus hastalığının uzun vadeli tahmini konusunda çok az çalışma yapıldığı ve bu tür çalışmaların sayısının artırılması gerektiği ifade edilmektedir (10).

Bu çalışmada ise makine öğrenme yöntemlerinden olan K en yakın komşu algoritması ile diabetes mellitus sağlık göstergeleri aracılığıyla diabetes mellitus tahmini yapılması amaçlanmıştır. Demografik bilgiler, laboratuvar test sonuçları ve her hasta için anket sorularına ilişkin veriler kullanılmıştır. K en yakın komşu yönteminden detaylı bahsedilerek işlem basamakları ve uygulama adımları gösterilmiştir. Bu araştırma ayrıca K en yakın komşu algoritmasının sağlık alanında uygulamasını göstererek yaygın kullanımını amaçlamıştır.

GEREÇ ve YÖNTEMLER

Diyabet Veri Seti

Çalışmanın uygulama kısmında kullanılacak olan veri seti kaggle veri tabanından temin edilmiş olup, UCI makine öğrenimi deposunun bir parçasıdır. Kastamonu Bilimsel Araştırmalar ve Yayın Etiği Kurulundan onay alınmıştır (Tarih:04.09.2024, karar no:6). Açık erişimli "Diabetes Health Indicators" isimli veri seti 21 açıklayıcı değişken, bir hedef değişkeni ve 253.680 veri içermektedir (17). Veriler 13.04.2024 tarihinde ilgili veri tabanından indirilmiştir. Veri seti ABD'de yaşam tarzı ile diabetes mellitus arasındaki ilişkiyi daha iyi anlamak için oluşturulmuş örnek bir veri setidir. Bu veri seti diabetes mellitus tanısıyla birlikte genel olarak kişiler hakkında sağlık istatistikleri ve yaşam tarzı anketi bilgileri içerir. 21 özellik, bazı demografik bilgiler, laboratuvar test sonuçları ve her hasta için anket sorularına verilen yanıtlardan oluşur. Sınıflandırma için hedef değişkeni kişilerde DM yok, preDM ya da DM bilgisini sunar. DM yok diabetes mellitus tanısı olmayan sağlıklı kişileri, preDM ise kan şekeri düzeyinin normalden yüksek olmasına rağmen diabetes mellitus tanısı koymak için yeterli yükseklikte olmayan kişileri ifade eder. DM sınıfı ise diabetes mellitus tanısı almış kişileri göstermektedir (17). Veri seti üç sürekli, 19 kategorik değişkenden oluşmaktadır. Değişkenlere ilişkin açıklamalar Tablo 1'de verilmiştir.

Tablo 1'de açıklaması yapılan veri seti farklı oranlarda eğitim ve test verisi olarak ayrılarak sınıflandırmadaki değişimler incelenmiştir. Eğitim verisi sırasıyla %50,0, %60,0, %75,0 ve %90,0 olarak ayrılmıştır. Eğitim amaçlı ayrılan kısım K en yakın komşu yöntemi ile öğrenme amaçlı, test kısmı ise diabetes mellitus hastalığının sınıflandırma başarısını

Tablo 1: Kullanılan değişkenler ve özellikleri

Değişken Adı	Türü	Değişken Açıklaması
Diabetes Mellitus	Kategorik	*Hedef değişken*: 0 = DM yok 1 = preDM 2 = DM
Yüksek Tansiyon	Kategorik	0 = yüksek tansiyon yok 1 = yüksek tansiyon var
Yüksek Kolesterol	Kategorik	0 = yüksek kolesterol yok 1 = yüksek kolesterol yok var
Kolesterol Kontrolü	Kategorik	0 = 5 yıl içinde kolesterol kontrolü yok 1 = 5 yıl içinde kolesterol kontrolü var
BMI	Sürekli	Vücut kütle indeksi
Sigara	Kategorik	Hayatınız boyunca en az 100 sigara içtiniz mi? [Not: 5 paket = 100 sigara] 0 = hayır 1 = evet
Felç	Kategorik	(Hiç) felç geçirdiğinizi söylediler mi? 0 = hayır 1 = evet
Kalp Hastalığı veya Krizi	Kategorik	Koroner kalp hastalığı veya miyokard enfarktüsü 0 = hayır 1 = evet
Fiziksel Aktivite	Kategorik	Geçtiğimiz 30 gündeki fiziksel aktivite (iş hariç) 0 = hayır 1 = evet
Meyveler	Kategorik	Günde 1 veya daha fazla kez meyve tüketimi 0 = hayır 1 = evet
Sebzeler	Kategorik	Günde 1 veya daha fazla kez sebze tüketimi 0 = hayır 1 = evet
Ağır Alkol Tüketimi	Kategorik	Ağır içiciler (yetişkin erkeklerin haftada 14'ten fazla içki içmesi ve yetişkin kadınların haftada 7'den fazla içki içmesi) 0 = hayır 1 = evet
Sağlık Sigortası	Kategorik	Herhangi bir sağlık sigortası kapsamına sahip misiniz? 0 = hayır 1 = evet
Sağlık Maliyeti	Kategorik	Geçtiğimiz 12 ayda doktora görünmeniz gerektiği ancak maliyet nedeniyle gidemediğiniz bir zaman oldu mu? 0 = hayır 1 = evet
Genel Sağlık	Kategorik	Genel olarak sağlığınızın şu şekilde olduğunu söyler misiniz: 1-5 ölçeğinde 1 = mükemmel 2 = çok iyi 3 = iyi 4 = orta 5 = kötü
Mental Sağlık	Sürekli	Şimdi stres, depresyon ve duygu sorunları gibi zihinsel sağlığınızı düşündüğünüzde, son 30 gün içinde zihinsel sağlığınız kaç gün boyunca iyi değildi? 1-30 gün ölçeği
Fiziksel Sağlık	Sürekli	Şimdi fiziksel hastalık ve yaralanmaları da içeren fiziksel sağlığınızı düşündüğünüzde, son 30 gün içinde fiziksel sağlığınız kaç gün iyi değildi? 1-30 gün ölçeği
Yürüyüşte Zorluk	Kategorik	Yürüme veya merdiven çıkma konusunda ciddi zorluk çekiyor musunuz? 0 = hayır 1 = evet
Cinsiyet	Kategorik	0 = kız 1 = erkek
Yaş	Kategorik	13 seviyeli yaş kategorisi: 1 = 18-24 9 = 60-64 13 = 80 veya üzeri
Eğitim	Kategorik	Eğitim seviyesi 6 kategori: 1 = Hiç okula gitmemiş veya sadece anaokuluna gitmiş 2 = 1. sınıftan 8. sınıfa kadar (İlkokul) 3 = 9. sınıftan 11. sınıfa kadar (Bazıları lise) 4 = 12. sınıf veya Lise mezunu 5 = 1 ila 3 yıl üniversite (Bazıları kolej veya teknik okul) 6 = 4 yıl veya daha fazla kolej (Üniversite mezunu)
Gelir	Kategorik	Gelir ölçeği 8 kategori: 1 = 10.000$'dan az 5 = 35.000$'dan az 8 = 75.000$ veya daha fazla

test etme amaçlı kullanılmıştır. R programı R3.6.0 versiyonu (18) kullanılarak yapılan sınıflandırma çalışmasında K en yakın komşu yöntemi için "class" paket kullanılmıştır. Uzaklık ölçüsü olarak Öklid uzaklık ölçüsü alınmıştır. Farklı k değerleri denenerek sınıflandırma başarısındaki değişimler gösterilmiştir. Diabetes mellitus hastalık durumunun sınıflandırma başarısını ölçmek için hata matrisi ("confusion matrix") aracılığıyla hesaplanan doğruluk, kappa ve diğer performans değerleri kullanılmıştır.

Hata matrisi ("confusion matrix") makine öğrenimi alanında ve istatistiksel sınıflandırma problemlerinde yöntemin performansını değerlendirmek için kullanılır. Kullanılan yöntemin öğrenme performansını görselleştirme imkânı verir. Tahmin edilen sınıf değeri ile gerçek değerler karşılaştırılır. Hata matrisi sınıf değişkeni kategori sayısınca satır ve sütun içerir (19,20).

Doğruluk, kappa, kesinlik (precision), duyarlılık (recall) ve F1 skor değerleri hata matrisi aracılığıyla hesaplanan ve yöntemin sınıflandırma başarısını ölçen performans metrikleridir. Doğruluk, incelenen toplam vaka içerisinden doğru tahminlerin oranıdır. Örnek olarak bir sınıflandırıcı on tahmin yaparsa ve bunların yedisi doğruysa doğruluk %70,0 olarak belirlenmektedir (21).

Kappa katsayısı ise iki değerlendirici arasındaki uyumu ölçen bir istatistik olduğu gibi bir sınıflandırıcının sınıflandırma başarısı hakkında da bilgi vermektedir. Kappa katsayısı 0 ile 1 arası değerler alır, 1'e yaklaştıkça uyumun arttığı yani sınıflandırma başarısının arttığı ifade edilmektedir (22).

Kesinlik pozitif olarak tahmin edilen örneklerin ne kadarının gerçekten pozitif olduğunu ölçer. Modelin pozitif sınıf olarak tahmin ettiği örneklerin doğruluğunu gösterir. Duyarlılık gerçek pozitiflerin ne kadarını doğru tahmin edebildiğimizi ölçer. Modelin pozitif sınıfa ait örnekleri ne kadar iyi yakaladığını gösterir. F1 skor ise kesinlik ve duyarlılık değerlerinin harmonik ortalamasını göstermektedir.

Hata matrisi Tablo 2'de, performans metrik değerlerine ilişkin formüller ise eşitlik (1), (2), (3), (4) ve (5) ile gösterilmiştir.

K En Yakın Komşu Yöntemi (Knn)

Makine öğrenme yöntemlerinden olan Knn yöntemi ilk defa 1951 yılında Fix ve Hodges tarafından tanıtılmış parametrik olmayan bir yöntemdir. Knn algoritması, kolay anlaşılır bir çalışma prensibine sahip olup başarılı sonuçlar vermesi ile ön plana çıkmaktadır. Veri setindeki eksik gözlemlerden etkilenmeyen, kategorik değişkenler için eksik gözlem değerlendirmesi yapabilen ve varsayımlarının az olması gibi avantajlara sahiptir. Daha çok sınıflandırma problemlerinin çözümünde kullanılmakla birlikte regresyon problemlerinde de kullanılabilmektedir. Veri setindeki gözlemlerin birbirine olan uzaklık ve benzerliklerine göre tahmin işlemi gerçekleştirmektedir. Algoritmanın genel amacına baktığımızda gözlemleri kendine ait özelliklere göre önceden belirlenen sınıflara atamaktır. Bunun yanında yeni bir gözlemin sınıflandırılması da sağlanır. Sınıflandırılmak istenen yeni gözlem, öğrenme veri seti yardımıyla en yakın k gözlem ile aynı gruba sınıflandırılır (23). Knn yöntemi sınıflandırmayı belirli işlem basamaklarına göre gerçekleştirir:

* İlk olarak veri ön işlemden geçirilerek kontrol edilir.
* Sınıflandırılacak olan gözlemin veri setindeki bütün gözlemlere olan uzaklığı hesaplanır.
* Daha sonra hesaplanan uzaklık değerleri sıralanır.
* K sayıda en az uzaklığa sahip (en yakın) gözlem belirlenir. K gözlem arasından en fazla olan sınıf yeni gözlemin sınıfı olarak atanır.

Burada belirlenecek olan k sayısı oldukça önemlidir. Belirlenecek olan k sayısı sınıflandırma başarısına doğrudan etki edecektir. Örneğin k sayısı 1 olarak belirlenirse yeni gözlem kendine en yakın komşusuna ait sınıfa atanacaktır. Farklı k değerleri denenerek yöntemin sınıflandırma başarısı artırılabilir (23).

Knn algoritmasının diğer yöntemlere göre matematiksel karmaşıklığının düşük olması, kolay anlaşılır ve uygulanabilir olması, başarılı sonuçlar vermesi sebebiyle bu çalışmada tercih edilmiştir. Knn, parametrik olmayan bir modeldir. Bu, verilerin belirli bir dağılım varsayımı gerektirmediği anlamına gelir. Verilerin doğrusal olup olmamasına bakılmaksızın, farklı veri yapıları üzerinde çalışabilir. Veri sayısının çok olması yöntemin performansını artırmaktadır. Bu çalışmada da 250 binden fazla örneğin bulunduğu veri seti kullanılmıştır. Hekim ve sağlık profesyonellerinin sade olan bu yöntemi kolayca uygulaması amacıyla çalışma prensibi ve uygulama kodları adım adım paylaşılmıştır. Yöntemin ayrıca diğer algoritmalara göre daha hızlı olması tercih nedenlerinden olmuştur (24). Yönetimin zayıf yönlerine bakacak olursak; veri boyutunun (değişken sayısının) fazla olması

$$kappa = \frac{P_0 - P_e}{1 - P_e} \quad P_0 = \frac{D_a + D_b + D_c}{GT} \quad P_e = \left(\frac{G_a}{GT} x \frac{T_a}{GT}\right) + \left(\frac{G_b}{GT} x \frac{T_b}{GT}\right) + \left(\frac{G_c}{GT} x \frac{T_c}{GT}\right) \quad (1)$$

$$Doğruluk = (D_a + D_b + D_c) / GT \quad (2)$$

$$Duyarlılık = Doğru\ Pozitif / (Doğru\ Pozitif + Yanlış\ Negatif) \quad (3)$$

$$Kesinlik = Doğru\ Pozitif / (Doğru\ Pozitif + Yanlış\ Pozitif) \quad (4)$$

$$F1\ Skor = 2\ X\ (Kesinlik\ X\ Duyarlılık) / (Kesinlik + Duyarlılık) \quad (5)$$

Tablo 2: Üç kategorili hedef değişkeni için "confusion matrix"

		Gerçek Sınıf			
		A	**B**	**C**	**Toplam**
Tahmin Sınıf	**A**	Da	Yab	Yac	Ta=Da+Yab+Yac
	B	Yba	Db	Ybc	Tb=Yba+Db+Ybc
	C	Yca	Ycb	Dc	Tc=Yca+Ycb+Dc
Toplam		Ga=Da+Yba+Yca	Gb=Yab+Db+Ycb	Gc=Yac+Ybc+Dc	GT=Ta+Tb+Tc
					GT=Ga+Gb+Gc

yöntemin performansını olumsuz etkilemektedir. Fakat bu çalışmada 21 belirteç kullanılmıştır. Bir diğer zayıf yönü k değerinin yanlış belirlenmesi sonucu kötü sonuçlar vermesidir. Bu çalışmada farklı k değerleri denenerek en yüksek performans elde edilmeye çalışılmıştır.

İstatistiksel Analiz

Veri ön işleme, ham verilerin analiz, modelleme veya makine öğrenimi algoritmaları için uygun hale getirilmesi sürecidir. Bu çalışmada kullanılan veri seti, veri tabanından indirildikten sonra birtakım ön işlemden geçirilmiştir. Eksik, hatalı ve aykırı değerler bakımından süzgeçten geçirilmiştir. Değişkenler kategorik ve sürekli olarak tanımlanarak dağılımları kontrol edilmiştir. Toplamda 21 belirteç kullanılmıştır. Veri kaynağında ilgili değişkenlerin diabetes mellitus belirteçleri olduğu belirtilmiştir ve bu sebeple öznitelik seçimine gidilmemiştir. Veri seti model eğitimi için eğitim seti ve modelin performansını değerlendirmek için test seti olarak ikiye ayrılmıştır.

Knn yönteminin R programında uygulaması aşağıdaki basamaklara göre yapılmıştır. Öncelikle "install.package" ve "library" komutları ile ilgili paketler yüklenmiş ve çağrılmıştır. "Diabetes" isimli veri seti R programına yüklendikten sonra df (dataframe) isimli değişkene atandı. Daha sonra yüklenen veri seti data frame'e ("as.data.frame" komutu ile) dönüştürüldü. Veri seti içerisindeki kategorik değişkenleri tanımlamak için "as.factor", sürekli değişkenleri tanımlamak için "as.numeric" komutu kullanıldı. Veri setinin eğitim (train) ve test olarak ayrılması için "caret" paketi içerisinde yer alan "createDataPartition" komutu kullanıldı. "dplyr" paketi ile sınıf değişkeni ve diğer bağımsız değişkenlerin ayrılması sağlandı. Eğitim veri seti ile "class" paketi içerisinde yer alan "knn" fonksiyonu kullanılarak öğrenme işlemi gerçekleştirildi. Test veri seti aracılığıyla sınıflandırma başarısı değerlendirilirken "confusionMatrix" fonksiyonu ile hata matrisi sonucu elde edildi. Herhangi bir k değeri için sınıflandırma on kez tekrardandı ve elde edilen performans değerlerinin ortalaması alındı. On tekrarın her birinde veri seti içerisindeki eğitim kısım rastgele seçildi, geri kalan kısım test verisi olarak kullanıldı. Bir satır on sütunluk doğruluk ve kappa isimli iki matris oluşturuldu ve elde edilen performans değerleri bu matrislere kaydedilerek en son ortalaması alındı. R programı kod satırları Tablo 3'te gösterilmiştir. Tablo 3, veri setinin %75,0 eğitim, %25,0 test olarak ayrıldığı ve k değerinin 1 (bir) olarak kullanıldığı örneği içermektedir.

BULGULAR

Diabetes mellitus tahmini için kullanılan ve diabetes mellitus belirteçlerini içeren 22 değişkenli veri setine ilişkin tanımlayıcı istatistikler Tablo 4'te verilmiştir. Veri seti içerisinde kayıp veri ve aykırı gözlem bulunmamaktadır. Sınıf değişkenine (diabetes mellitus durumu) ilişkin dağılım ilgili tabloda paylaşılmıştır. ABD'de yaşayan bireylerden toplanan verilerde kişilerin yaklaşık %15,8'i diabetes mellitus tanılıdır, %42,9'unde yüksek tansiyon, %42,4'ünde yüksek kolesterol bulunmaktadır.

Tablo 3: Uygulama kısmı için R program kodları

```
#install.packages("class")
library(class)#knn icin
#install.packages("caret")
library(caret)
#install.packages("tidyverse")
library(tidyverse)#dplyr için
Diabetes <- read_excel("C:/Users/hp/OneDrive/Masaüstü/ Diabetes.xls")
df=Diabetes
df=as.data.frame(df)
df$Diabetes_012=as.factor(df$Diabetes_012)
df$HighBP=as.factor(df$HighBP)
df$HighChol=as.factor(df$HighChol)
df$CholCheck=as.factor(df$CholCheck)
df$Smoker=as.factor(df$Smoker)
df$Stroke=as.factor(df$Stroke)
df$HeartDiseaseorAttack=as.factor(df$HeartDiseaseorAttack)
df$PhysActivity=as.factor(df$PhysActivity)
df$Fruits=as.factor(df$Fruits)
df$BMI=as.numeric(df$BMI)
#diğer değişkenler de aynı şekilde kategorik veya sürekli olarak tanımlanır
dogruluk <- c(1:10)
kappa <- c(1:10)
s=1
for (i in 1:10) {
 train_indeks <- createDataPartition(df$ Diabetes_012,
p = 0.75, list = FALSE, times = 1)
  train <- df[train_indeks,]
  test <- df[-train_indeks,]
  train_x <- train %>% dplyr::select(-Diabetes_012)
  train_y <- train$ Diabetes_012
  test_x <- test %>% dplyr::select(-Diabetes_012)
  test_y <- test$ Diabetes_012
   knn_fit <- knn(train = train, test = test, cl = train_y,
k = 1)
  sonuc<-confusionMatrix(knn_fit,test_y)
  d<-sonuc$overall
  dogruluk[s] <- d['Accuracy']
  kappa[s] <- d["Kappa"]
  s=s+1
}
mean(dogruluk)
mean(kappa)
```

Tablo 4: Değişkenlere ilişkin tanımlayıcı istatistikler

Değişkenler, n (%)*	Sonuç (n=253680)
Diabetes Mellitus	
DM yok	213703 (84,2)
preDM	4631 (1,8)
DM	35346 (13,9)
Yüksek Tansiyon	
Yüksek tansiyon yok	144851 (57,1)
Yüksek tansiyon var	108829 (42,9)
Yüksek Kolesterol	
Yüksek kolesterol yok	146089 (57,6)
Yüksek kolesterol var	107591 (42,4)
Kolesterol Kontrolü	
5 yıl içerisinde yok	9470 (3,7)
5 yıl içerisinde var	244210 (96,3)
Sigara	
Sigara hayır	141257 (55,7)
Sigara evet	112423 (44,3)
Felç	
Felç geçirmeyen	243388 (95,9)
Felç geçiren	10292 (4,1)
Kalp Hastalığı/Krizi	
Kalp hastalığı/krizi geçirmeyen	229787 (90,6)
Kalp hastalığı/krizi geçiren	23893 (9,4)
Fiziksel Aktivite	
Fiziksel aktivite yok	61760 (24,4)
Fiziksel aktivite var	191920 (75,7)
Meyveler	
Meyve tüketimi yok	92782 (36,6)
Meyve tüketimi var	160898 (63,4)
Sebzeler	
Sebze tüketimi yok	47839 (18,9)
Sebze tüketimi var	205841 (81,1)
Ağır Alkol Tüketimi	
Ağır alkol tüketimi yok	239424 (94,4)
Ağır alkol tüketimi var	14256 (5,6)
Sağlık Sigortası	
Sağlık sigortası yok	12417 (4,9)
Sağlık sigortası var	241263 (95,1)
Sağlık Maliyeti	
Maliyet zorluğu yok	232326 (91,6)
Maliyet zorluğu var	21354 (8,4)
Yürüyüşte Zorluk	
Zorluk çekme yok	211005 (83,2)
Zorluk çekme var	42675 (16,8)
Genel Sağlık	
Mükemmel	45299 (17,9)
Çok İyi	89084 (35,1)
İyi	75646 (29,8)
Orta	31570 (12,4)
Kötü	12081 (4,8)
Cinsiyet	
Kadın	141974 (56,0)
Erkek	111706 (44,0)
Yaş	
18-24	5700 (2,3)
25-29	7598 (3,0)
30-34	11123 (4,4)
35-39	13823 (5,5)
40-44	16157 (6,4)
45-49	19819 (7,8)
50-54	26314 (10,4)
55-59	30832 (12,2)
60-64	33244 (13,1)
65-69	32194 (12,7)
70-74	23533 (9,3)
75-79	15980 (6,3)
80 ve üzeri	17363 (6,4)
Eğitim	
Hiç okula gitmemiş	174 (0,1)
1. sınıftan 8. sınıfa kadar (İlkokul)	4043 (1,6)
9. sınıftan 11. sınıfa kadar	9478 (3,7)
12. sınıf veya Lise mezunu	62750 (24,7)
1 ila 3 yıl üniversite	69910 (27,6)
Üniversite mezunu	107325 (42,3)
Gelir	
10.000$'dan az	9811 (3,9)
15.000$'dan az	11783 (4,6)
20.000$'dan az	15994 (6,3)
25.000$'dan az	20135 (7,9)
35.000$'dan az	25883 (10,2)
50.000$'dan az	36470 (14,4)
75.000$'dan az	43219 (17,0)
75.000$ ve üzeri	90385 (35,6)
	Ort±SS (Medyan)
BMI	28,38±6,61 (27)
Mental Sağlık	3,18±7,41 (0)
Fiziksel Sağlık	4,24±8,72 (0)

*Veriler n (%) olarak gösterilmiştir.

Sigara içenlerin oranı %44,3, ağır alkol tüketenlerin oranı ise %5,6'dır. Felç geçirenlerin oranı %4,1, kalp hastalığı/krizi geçirenleri oranı ise %9,4, yürüyüşte zorluk çektiğini bildirenlerin oranı ise %16,8'dir. Fiziksel aktivitesi bulunmayanların oranı %24,4'tür. Bireylerin %36,6'sı meyve, %18,9'u sebze tüketmemektedir. Genel sağlık durumunu kötü olarak tanımlayanlar, katılımcıların yaklaşık %4,8'ini oluşturmaktadır. Çalışmaya katılanların demografik özelliklerine bakıldığında %44,0'ı erkek, yaklaşık %36,0'ı 50-65 yaş aralığında ve %42,3'ü üniversite mezunudur. Vücut kütle indeksi (BMI) ortalaması 28,38±6,61, mental sağlığının iyi olmadığı ortalama gün sayısı 3,18±7,41, zihinsel sağlığının iyi olmadığı ortalama gün sayısı 4,24±8,72 olarak bulunmuştur. Diabetes mellitus olmayanların BMI ortalaması 27,74±6,26 iken diabetes mellitus hastası olanların BMI ortalaması ise 31,94±7,36 olarak bulunmuştur.

K en yakın komşu algoritması ile Tablo 1'de verilen belirteçler (değişkenler) kullanılarak diabetes mellitus tahmini için makine öğrenme sınıflandırması yapılmıştır. Bu belirteçler yardımıyla kişilerin DM yok, preDM ya da DM olma durumları tahmin edilmeye çalışılmıştır. Yapılan diabetes mellitus tahmini sonucuna ilişkin performans değerleri Tablo 5'te gösterilmiştir. %97,2 doğruluk ve %88,9 kappa başarı değeri ile diabetes mellitus tahmin edilmiştir. Veri setinin eğitim ve test ayrım oranları ile k değerindeki değişimin sınıflandırma başarısını etkilediği görülmektedir. Farklı k değerleri için sınıflandırma sonuçları incelenmiş ve tüm eğitim test ayrım oranlarında en iyi performans k=3 (üç) değerinde elde edilmiştir. Bunun yanında k=3 (üç) değerine sahipken her bir kategori için kesinlik, duyarlılık ve F1 skor değerleri de hesaplanmıştır. DM yok grubu için sırasıyla 0,97, 0,99 ve 0,99 değerleri bulunmuştur. PreDM grubu için sırasıyla 0,34, 0,04, ve 0,06 değerleri bulunmuştur. DM grubu için sırasıyla 0,97, 0,92 ve 0,95 değerleri bulunmuştur. Ayrıca eğitim veri setinin hacminin artırılması yöntemin öğrenme başarısını ve buna bağlı olarak sınıflandırma başarısını artırdığı görülmüştür. k değerlerine göre diabetes mellitus tahmini performans değişimleri Şekil 1 ve 2'deki grafiklerde görülmektedir. Ulaşılan sonuçlar K en yakın komşu yönteminin kişilerin diabetes mellitus hastalık durumunu yüksek başarı ile sınıflandırabildiğini göstermektedir.

TARTIŞMA ve SONUÇ

Diabetes mellitus dünya çapında artan sayıda kişiyi etkileyen bir küresel sağlık sorunu hâline gelmiştir. Dünyada önde gelen ölüm nedenleri arasına girmiştir (2). Modern yaşam tarzı da diabetes mellitus görülme sıklığını önemli ölçüde artırmıştır (4). Bu nedenle, hastalığın erken teşhisi bir zorunluluk hâline gelmiştir. Bu doğrultuda diyabet geliştirme riski daha yüksek olan kişileri doğru bir şekilde belirleyebilen sistemlere ihtiyaç duyulmaktadır. Bu da diabetes mellitus hastalığının gelecekteki gelişimiyle yüksek oranda ilişkili olan en iyi özellikleri bulmayı başararak gerçekleşebilir (10). Ekonomik açıdan bakıldığında da diabetes mellitus en maliyetli hastalıklardan biridir, ayrıca diabetes mellitus hastalığına sahip yetişkinlerin yüksek bir yüzdesi düşük ve orta gelirli ülkelerde yaşamaktadır ve bu da bu ülkeler için daha fazla ekonomik sıkıntıya neden olmaktadır (5). Hastalığın yaygınlığı artmaya devam ettikçe, araştırmacılar doğru diabetes mellitus tahmini için gayretle çalışmaktadırlar (2). Hekimler hastaların diabetes mellitus hastalığının farkında olmadıklarından vakaların çoğunun teşhis edilemediği ve buna bağlı olarak önleme süreci gecikmeyle beraber karmaşıklaştığını belirtmektedir. Çalışmada kullanılan veriler incelendiğinde bu sonucu desteklemektedir. DM ve preDM tanılı kişilerin oranı %15,8 iken, genel sağlığını iyi, çok iyi ve mükemmel olarak belirten kişilerin oranı ise %82,8'dir. Sadece %4,8'i genel sağlığının kötü olduğunu belirtmiştir. Bu

Tablo 5: k değerlerine göre performans metriklerindeki değişim

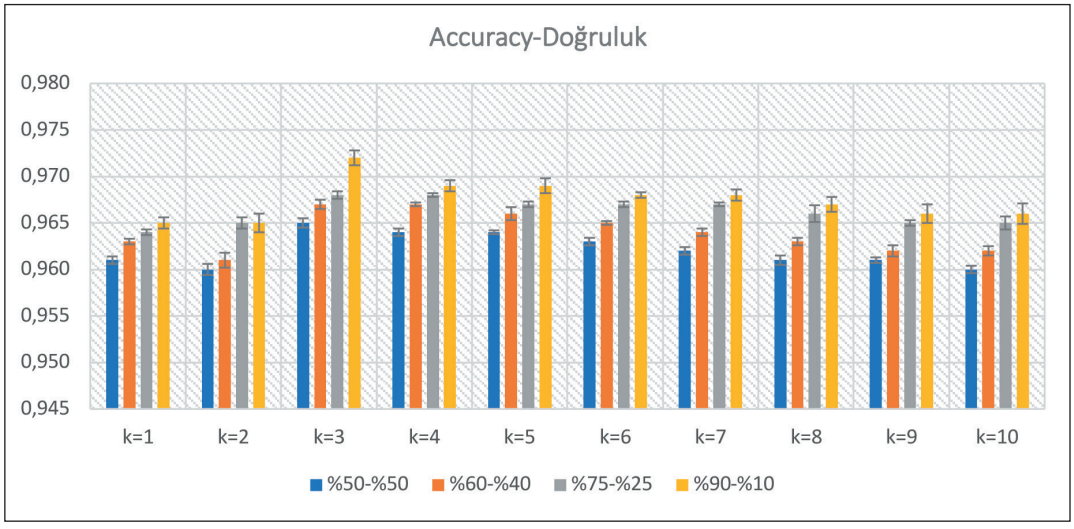
Eğitim-Test	%50-%50		%60-%40		%75-%25		%90-%10	
k değerleri	**Doğruluk**	**Kappa**	**Doğruluk**	**Kappa**	**Doğruluk**	**Kappa**	**Doğruluk**	**Kappa**
k=1	0,961	0,847	0,963	0,857	0,964	0,862	0,965	0,865
k=2	0,960	0,842	0,961	0,848	0,965	0,862	0,965	0,865
k=3	**0,965**	**0,861**	**0,967**	**0,868**	**0,968**	**0,875**	**0,972**	**0,889**
k=4	0,964	0,857	0,967	0,866	0,968	0,872	0,969	0,877
k=5	0,964	0,856	0,966	0,862	0,967	0,869	0,969	0,875
k=6	0,963	0,850	0,965	0,859	0,967	0,866	0,968	0,870
k=7	0,962	0,847	0,964	0,854	0,967	0,868	0,968	0,871
k=8	0,961	0,843	0,963	0,850	0,966	0,862	0,967	0,868
k=9	0,961	0,840	0,962	0,847	0,965	0,856	0,966	0,861
k=10	0,960	0,838	0,962	0,847	0,965	0,859	0,966	0,864

nedenle, diabetes mellitus hastalığının ilk evre tanısı, klinik olarak anlamlı sonuçları mümkün kılmak için önemli bir faktördür (13,15).

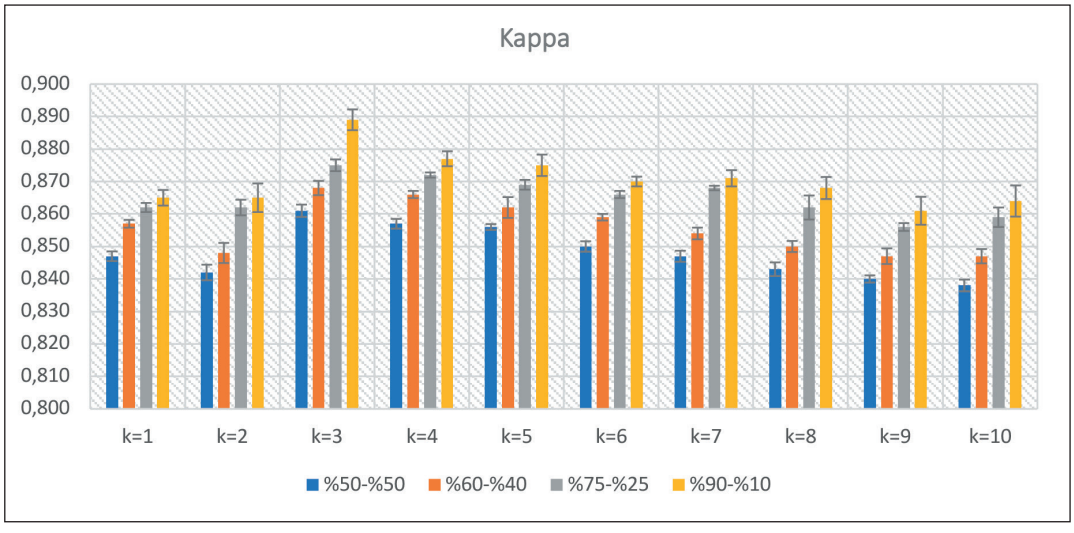
Makine öğrenmesi, çeşitli durumların risk tahmini, prognozu, tedavisi ve yönetimi için etkili araçlar geliştirmedeki yüksek potansiyeli nedeniyle sağlık hizmeti sağlayıcıları ve doktorlar arasında büyük popülerlik kazanmıştır (4). Makine öğrenmesi gibi popüler veri analizi araçları, sağlık hizmetlerinde devrim yaratabilecek, kişiye özel yönetim ve erken keşif olanağı sağlar (14).

Bu araştırmada son yıllarda popülerliği artan makine öğrenme algoritmalarından K en yakın komşu algoritması kullanılarak diabetes mellitus hastalığının erken tanı ve teşhisine yönelik bir uygulama çalışması gerçekleştirilmiştir. Araştırmanın uygulama kısmında kullanılan veri seti ABD'de yaşayan bireylerden toplanan ve UCI veri tabanına kayıtlı verilerden oluşmaktadır. Bu veri setinde diabetes mellitus tanısı için 21 belirteç (değişken) yer almaktadır. Bu belirteçler ve özellikleri Tablo 1'de gösterilmiştir. K en yakın komşu algoritmasından detaylı bahsedilerek R programı ile yöntemin nasıl uygulandığı basamaklar hâlinde gösterilmiştir.

Öncelikle veri seti R programına aktarılmıştır. Burada değişkenlerin kategorik ve sürekli olmasına göre türleri tanımlanmıştır. Daha sonra veri seti eğitim ve test olarak ikiye ayrılmıştır (%50,0, %60,0, %75,0, %90,0). Eğitim ve test veri seti içerisindeki hedef değişken (diabetes mellitus durum değişkeni) veri setinden ayrılmıştır. Belirlenen k değeri için knn fonksiyonu ile eğitim veri seti kullanılmış ve test veri seti için diabetes mellitus tahminleri oluşturulmuştur. Hata



Şekil 1: k değerlerine göre doğruluk (accuracy) değişimi



Şekil 2: k değerlerine göre kappa değişimi

matrisi aracılığıyla tahmin edilen diabetes mellitus durum kategorileri ile test veri setindeki hedef değişkeni kategorileri için çapraz tablo oluşturulmuştur. Daha sonra diabetes mellitus tahmin başarısını ölçmek için doğruluk, kappa ve diğer performans değerleri hata matrisinden hesaplanmıştır. Bu işlem on kez tekrarlanarak (on kat çapraz geçerlilik) performans metriklerinin ortalaması alınmıştır. Farklı k değerleri için hesaplanan tahmin performans değerleri Tablo 5'te gösterilmiştir. En iyi performans k=3 (üç) komşuluk değeri için elde edilmiştir. Tablo 1'de verilen belirteçler kullanılarak K en yakın komşu yöntemi sayesinde %97,2 doğruluk ve %88,9 kappa performansı ile başarılı bir şekilde diabetes mellitus hastalığının tahmin edilebildiği görülmüştür.

Literatürde makine öğrenme yöntemlerini kullanarak diabetes mellitus tahminine yönelik çalışmalar mevcuttur. Kullanılan belirteçlerin (değişkenlerin) farklı olduğu bu çalışmalardan bazıları kişilerin diabetes mellitus olup olmadığını tahmin eden ikili sınıflandırma çalışmalarıdır. Bazıları ise preDM sınıfını da içeren üçlü sınıflandırma çalışmalarını içermektedir. Bu çalışmalarda benzer şekilde makine öğrenme yöntemlerinin yüksek doğruluk değerleri ile tahmin yapabildiği belirtilmektedir.

Glikoz, gebelikler, kan basıncı, cilt kalınlığı, insülin, vücut kütle indeksi, yaş, diabetes mellitus öyküsü gibi belirteçlerin kullanıldığı diabetes mellitus tahminine yönelik ikili sınıflandırma çalışmasında destek vektör makineleri ve yapay sinir ağları makine öğrenme yöntemleri ile %95 doğruluk elde edilmiştir (25). Aynı veri setinin kullanıldığı başka bir çalışmada farklı makine öğrenme yöntemleri kullanılarak %90,0-%93,0 arası doğruluk değerleri elde edilmiştir (2). Bangladeş'teki Sylhet Hastanesi tarafından derlenen (University of California Irvine) UCI Makine Öğrenimi deposundan alınan 520 örnek sayısına ve 16 belirtece sahip veri seti ile diabetes mellitus tanısının pozitif mi negatif mi olduğunu belirlemeye yönelik yapılan makine öğrenmesi çalışmasında %94,9 tahmin doğruluğu elde edilmiştir ve diabetes mellitus hastalarının erken tespiti için bir çerçeve önerilmiştir (16). Diabetes mellitus tahmini için 16 belirteçli aynı veri setini ve farklı makine öğrenme yöntemini kullanan başka bir çalışmada ise %98,1 doğruluk elde edilmiştir (3). Belirteç sayısının 12, denek sayısının 253.395 olduğu diabetes mellitus tahminine yönelik yapılan araştırmada çeşitli makine öğrenme yöntemleri ile %71,0-%73,0 doğruluk başarısı elde edilmiştir. Diabetes mellitus tanısının 232 hasta bireyle tahminlenmesinde farklı yöntemlerin kullanıldığı araştırmada en yüksek başarının rastgele orman algoritması ile elde edildiği ve diabetes mellitus hastalığına sahip kişilerin %81,9-%84,5 doğruluk değerleri ile tahmin edilebildiği ifade edilmiştir. Knn yönteminin de kullanılarak sınıflama yapılan bu çalışmada %60,0-%63,0 doğruluk değerleri elde edilmiştir (24). Sınıf dengesizliği durumlarında diabetes mellitus tahmini ve hastalıkların belirlenmesinde öznitelik seçimi konulara da literatürde değinen çalışmalar olmuştur (26, 27). 18 yaşından büyük 185 hasta ile yapılan çalışmada sınıf dengesizliği durumlarında diabetes mellitus tahmini için çeşitli makine öğrenme yöntemleri denenmiş ve veri setlerini yeniden örnekleme yöntemlerine tabi tutarak veriyi dengeledikten sonra sınıflandırma algoritmalarının kullanılması önerilmiştir (26). Öznitelik seçimi öneminin vurgulandığı çalışmada farklı veri setleri ve farklı sınıflandırma algoritmaları kullanılmış ve çok boyutlu veri setlerinde öznitelik seçimi ile düşük boyutlu veri kullanılması sınıflandırma başarısını artırabileceği ifade edilmiştir (27). Bu çalışmada veri boyutu çok büyük olmadığı için ve veri kaynağında değişkenlerin diabetes mellitus belirteçleri olduğu bildirildiği için öznitelik seçimine gidilmemiştir.

Görüldüğü üzere tanı ve teşhis için makine öğrenimini kullanan ve yüksek doğruluk başarısına ulaşan çalışmalar yapılmıştır. Fakat bu çalışmalarda sınıflandırma başarısının doğruluk değeri üzerinden yapıldığı görülmektedir. Sadece doğruluk değeri üzerinden yapılan başarı yorumlaması yanıltıcı olabilmektedir. Sadece bir sınıfın yüksek oranda doğru tahminlenmesi doğruluk değerinin yüksek bulunmasına sebep olabilir ve diabetes mellitus için doğru olmayan belirteçlerin diabetes mellitus hastalığını tahmin edebildiği gibi bir sonuca varılabilir. Literatürde sadece doğruluk derecesine bakılarak sınıflandırma başarısı hakkında yorum yapılmasının özellikle dengesiz ve çok sınıflı veri setleri için yanıltıcı olacağı bildirilmiştir (28). Yapılan bu çalışmada 21 belirteç kullanarak diabetes mellitus hastalığının çoklu sınıflandırması gerçekleştirilmiştir. Sınıflandırma (tahmin) sonucu hem doğruluk hem kappa hem de diğer performans değerlerine dikkate alınarak performans değerlendirmesi yapılmıştır. Kappa performans metriği özellikle çok sınıflı sınıflandırma problemlerinde performans ölçüsü olarak kullanılmaktadır. Tablo 5 incelendiğinde doğruluk değerlerinin hep yüksek çıkma eğiliminde olduğu görülecektir. Örnek hacminin büyük olması Knn yönteminin başarısını artırdığı literatürde bildirilmektedir. Bu yönüyle bakıldığında 232 hasta ve 18 belirteç kullanılarak Knn yöntemi ile diabetes mellitus tahmininde %63,0 doğruluk elde edilirken (24), 250 binden fazla hasta ve 21 belirteç kullanarak Knn yöntemi ile diabetes mellitus tahmininde %97,2 doğruluk elde edilmiştir. Büyük örnek hacimlerinde Knn yönteminin başarılı sonuçlar verdiği bilgisi desteklenmiştir.

Sonuç olarak Tablo 1'de verilen belirteçler kullanılarak K en yakın komşu yöntemi ile diabetes mellitus tahmini %88,9 kappa başarısı ile tahmin edilmiştir. Makine öğrenimindeki gelişmeler diabetes mellitus hastalığının önceden belirlenmesine katkı sağlayacaktır. Literatürde geliştirilen makine

öğrenmesi tahmin modelleri ve bu çalışmanın bulguları hem klinisyenler hem de hastalar için yararlı olacaktır. Çalışma sonuçları uygulayıcılar için klinik karar alma ve hasta danışmanlığında uygulanabilir bir destek olarak kullanılabilir. Ayrıca hastalığın erken tahmini, diabetes mellitus hastalarının ve diabetes mellitus riski taşıyanların hastalığın ilerlemesini ve yaşamı tehdit eden komplikasyonlarını geciktirebilecek önleyici tedbirler almasını sağlayabilir.

Bunun yanında diabetes mellitus tahmininde uygulaması kolay, hızlı ve başarılı sonuç veren Knn yönteminin kullanılması hekime önemli bir yol gösterici olarak hizmet edebilir. Erken teşhis, risk faktörlerinin tespiti, kişiselleştirilmiş tedavi, sürekli izleme ve komplikasyon tahmini gibi unsurlar, diyabetin daha etkin yönetilmesine yardımcı olabilir. Belirtilen pozitif katkıları ile hekimin kararlarını desteklerken, aynı zamanda hasta sonuçlarını iyileştirme potansiyeline sahiptir.

Ayrıca diabetes mellitus hastalığının doğru tahminlenmesinde belirteçlerin oldukça önemli olduğu görülmektedir. Bu çalışmada diabetes mellitus tahmini için kullanılan veriler ABD'de yaşayan bireylere ilişkin verilerdir. Aynı değişkenlerin ülkemizdeki bireyler için de diabetes mellitus belirteçleri olarak kullanılıp kullanılamayacağına yönelik bir çalışma gerçekleştirilebilir.

Son olarak bu çalışmanın, Knn makine öğrenme yöntemine yönelik uygulamalı anlatımıyla diabetes mellitus hastalık tahmini araştırmacılarına ve sağlık hizmeti sağlayıcılarına yararlı bir kaynak olacağı düşünülmektedir.

Teşekkür
Çalışmanın uygulama kısmı için veri sağlayan "UC Irvine Machine Learning Repository" veri tabanına teşekkürlerimi sunarım.

Çıkar çatışması
Çıkar çatışması bulunmamaktadır.

Yazar Katkı Beyanı
Çalışma tasarımı: **Ali Vasfi Ağlarcı, Feridun Karakurt,** Yöntem: **Ali Vasfi Ağlarcı,** Araştırma: **Ali Vasfi Ağlarcı, Feridun Karakurt,** veri analizi: **Ali Vasfi Ağlarcı,** Yazma: **Ali Vasfi Ağlarcı,** Yazma-gözden geçirme-düzenleme: **Ali Vasfi Ağlarcı, Feridun Karakurt.**

Finansal destek
Finansal destek bulunmamaktadır.

Etik Kurul Onayı
Kastamonu Bilimsel Araştırmalar ve Yayın Etiği Kurulundan onay alınmıştır (Tarih:04.09.2024, karar no:6).

Hakemlik Süreci
Kör hakemlik süreci sonrası yayına uygun bulunmuştur.

KAYNAKLAR

1. Kır Biçer E, Çekiç M, Ayvazoğlu G. Üniversite Çalışanlarında Tip 2 Diyabet Riskinin ve İlişkili Faktörlerin Değerlendirilmesi. IGUSABDER. 2024;253–272.
2. Oliullah K, Rasel MH, Islam, MM. et al. A stacked ensemble machine learning approach for the prediction of diabetes. J Diabetes Metab Disord 23, 603–617 (2024). https://doi.org/10.1007/s40200-023-01321-2
3. Elsayed N, ElSayed Z and Ozer M. "Early Stage Diabetes Prediction via Extreme Learning Machine," SoutheastCon 2022, Mobile, AL, USA, 2022, pp. 374-379, doi: 10.1109/SoutheastCon48659.2022.9764032.
4. Dritsas E, Trigka M. Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. Sensors. 2022; 22(14):5304. https://doi.org/10.3390/s22145304
5. Al-Haija QA, Smadi M, Al-Bataineh OM. Early Stage Diabetes Risk Prediction via Machine Learning. In: Abraham, A., et al. Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021) (2022). Lecture Notes in Networks and Systems, vol 417. Springer, Cham. https://doi.org/10.1007/978-3-030-96302-6_42.
6. International Diabetes Federation. IDF Diabetes Atlas: 10th edition 2021. https://diabetesatlas.org/data/en/country/203/tr.html Erişim Tarihi:07.07.2024.
7. Bishop CM. Pattern Recognition and Machine Learning. Springer, ISBN: 0-387- 31073-8 (2007).
8. Alpaydin E. Introduction to Machine Learning. London: The MIT Press (2010).
9. Khakurel U, Abdelmoumin G, Bajracharya A, Rawat DB. "Exploring bias and fairness in artificial intelligence and machine learning algorithms", Proc. SPIE 12113, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, 1211324 (6 June 2022), https://doi.org/10.1117/12.2621282.
10. Islam MS, Qaraqe MK, Abbas HT, Erraguntla M and Abdul-Ghani M. "The Prediction of Diabetes Development: A Machine Learning Framework," 2020 IEEE 5th Middle East and Africa Conference on Biomedical Engineering (MECBME), Amman, Jordan, 2020, pp. 1-6, doi: 10.1109/MECBME47393.2020.9292043.
11. Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. International Journal of Environmental Research and Public Health. 2021; 18(6):3317. https://doi.org/10.3390/ijerph18063317
12. Singh Y, Tiwari M. Revolutionizing Diabetes Disease Prediction Through Novel Machine Learning Techniques. Nano. 2024;19(4). https://doi.org/10.1142/S179329202350056X
13. Islam MS, Minul Alam M, Ahamed A and Ali Meerza SI. "Prediction of Diabetes at Early Stage using Interpretable Machine Learning," SoutheastCon 2023, Orlando, FL, USA, 2023, pp. 261-265, doi: 10.1109/SoutheastCon51012.2023.10115152.

14. Bassam G, Rouai A, Ahmad R and Khan MA. "Diabetes Prediction Empowered with Multi-level Data Fusion and Machine Learning" International Journal of Advanced Computer Science and Applications(IJACSA), 14(10), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0141062
15. Abnoosian K, Farnoosh R and Behzadi MH. Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics 24, 337 (2023). https://doi.org/10.1186/s12859-023-05465-z
16. Ahmed U et al. "Prediction of Diabetes Empowered With Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.
17. UC Irvine Machine Learning Repository. CDC Diabetes Health Indicators. https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators. Erişim Tarihi: 13.04.2024. DOI 10.24432/C53919
18. R3.6.0, https://cran.r-project.org/bin/windows/base/old/
19. Powers DMW. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2011;2 (1): 37–63.
20. Stehman SV. "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 1997;62 (1): 77–89. doi:10.1016/S0034-4257(97)00083-7
21. Metz CE. "Basic principles of ROC analysis" (PDF). Semin Nucl Med.1978;8 (4): 283–98. doi:10.1016/s0001-2998(78)80014-2.
22. Sim J, Wright CC. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements" in Physical Therapy. 2005;85, 257-268.
23. Cunningham P, Delany SJ. K-Neighbor Classifiers. J Multiple Classifier Syst. 2007;34(8):1-17.
24. Özkan Y, Sarer Yürekli B, Suner A. Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması. Gümüşhane Üniversitesi Fen Bilimleri Dergisi, 2022;12(1), 211-226. https://doi.org/10.17714/gumusfenbil.820882
25. Nadeem MW, Goh HG, Ponnusamy V, Andonovic I, Khan MA, Hussain M. A Fusion-Based Machine Learning Approach for the Prediction of the Onset of Diabetes. Healthcare. 2021; 9(10):1393. https://doi.org/10.3390/healthcare9101393.
26. Turhan S, Özkan Y, Yürekli BS, Suner A, Doğu E. Sınıf Dengesizliği Varlığında Hastalık Tanısı için Kolektif Öğrenme Yöntemlerinin Karşılaştırılması: Diyabet Tanısı Örneği. Turkiye Klinikleri J Biostat. 2020;12(1):16-26. DOI: 10.5336/biostatic.2019-66816
27. Demirarslan M, Suner A. Sağlık Veri Setlerinde Öznitelik Seçiminin Sınıflandırma Performansına Etkisi. JAIHS 2021; 1(1):6-11. DOI 10.52309/jai.2021.2
28. Ağlarcı AV, Bal C. Effect of various factors on classification performance of ordinal logistic regression. International Journal of Data Mining, Modelling and Management. 2024;16(2):196-208. https://doi.org/10.1504/IJDMMM.2024.138813.