T.C

BEYKENT ÜNİVERSİTESİ

MÜHENDİSLİK MİMARLIK FAKÜLTESİ

LİSANS PROGRAMI

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

**K-NN Yöntemiyle Şeker Hastalığı Teşhisi**

Lisans Projesi, Makine Öğrenmesi

Hazırlayan

**BETÜL TUĞÇE DİKDOĞMUŞ**

Danışman

**DR. ÖĞR. ÜYESİ ATINÇ YILMAZ**

İstanbul, 2021

**ÖZET**

Makine öğrenmesi insan kaynakları, finans, lojistik ve üretim gibi birçok alanda kullanılmaktadır. Hastalıkları daha doğru teşhis etmek, insan yükünü hafifletmek ve sağlık risklerini değerlendirmek amacıyla sağlık hizmetleri de makine öğrenmesinden yararlanmaktadır. Diyabet hastalığının teşhisi çalışmalarının temelini sınıflandırma çalışmaları oluşturmaktadır. Bu çalışmanın amacı, hastaların bilgilerine göre diyabet hastası olup olmadıklarının teşhis edilmesidir. K En Yakın Komşu (KNN) yöntemi kullanılmıştır. Elde edilen sınıflandırma doğruluğu %87,66’dır. Bu durum, hastalıkların teşhis ve tedavi edilme süreçlerinde insan kaynaklı hataları minimize etmek ve medikal karar süreçlerine destek vermek için makine öğrenmesinden faydalanılabileceğini ortaya koymaktadır.

**Anahtar Kelimeler:** Diyabet hastalığı, k-nn, sağlıkta makine öğrenmesi

**1. GİRİŞ**

Sağlık hizmetlerinde hastalıkların tanı ve teşhisinin yalnızca insan kontrolü ile yönetilmesi mümkün değildir. Hastalıkların erken teşhisi, tedavi edilmesi ve toplumun sağlık düzeyinin iyileştirilmesi için makine öğrenmesi sıkça kullanılan yöntemlerden biridir. Aynı zamanda maliyeti azaltmak, verimliliği artırmak, insan kaynaklı hataları en aza indirmek, hastalıkları doğru ve zamanında teşhis edebilmek için makine öğrenmesinden yararlanılmaktadır.

Diyabet dünyada sıkça rastlanan ve hızla artış göstermekte olan hastalıklardan biridir. Ayrıca kalp rahatsızlığı, böbrek hastalıkları, körlük gibi çeşitli hastalıklara sebep olabilmektedir. Diyabetin erken teşhisi başka hastalıkların oluşumuna sebep olmasını önlemek adına çok önemlidir. Bu bağlamda, birçok komplikasyonun oluşumuna sebep olan diyabet için sınıflandırma tercih edilmiştir. Şeker hastalığı olarak bilinen diyabet, insülin hormonunun, vücutta bulunmaması veya az seviyede bulunmasından kaynaklanan bir hastalık olarak tanımlanmaktadır. Bundan dolayı kandaki glikoz miktarı normalin üzerine çıkmakta ve normalde şeker içermemesi gereken idrarda şekere rastlanmaktadır. Uluslararası Diyabet Federasyonu’nun (IDF) sağladığı istatistiki verilere göre her 11 yetişkinden biri diyabet hastalığına sahiptir. Her 6 saniyede bir kişinin diyabet kaynaklı sebeplerden ötürü öldüğü belirtilmektedir. 11 yetişkinden biri diyabetle savaşırken, teşhis konulmadığı için 2 diyabetli yetişkinden biri (%46,5) diyabetli olduğunu bilmemektedir. Uluslararası Diyabet Federasyonu’nun tahminleri, diyabetin 2040 yılında 10 kişiden birinde görüleceği yönündedir.

Utku Köse çalışmasında, diyabet teşhisine odaklanmıştır. Amacı doğrultusunda, zeki optimizasyon tabanlı Destek Vektör Makineleri (DVM) yaklaşımına dayalı bir süreç izlemiştir. Çalışmasını Pima Yerlileri veri seti üzerinde uygulamıştır. Zeki optimizasyon kapsamında seçilen algoritmaların diyabet teşhisi için performansları değerlendirildiğinde bütün algoritmaların istendik seviyelerde başarılı olduğu görülmüştür [1]. Güler ve Übeyli çalışmalarında, dört farklı algoritma ile eğitilen çok katmanlı perseptron sinir ağlarını diyabet teşhisinde kullanmışlar ve en başarılı algoritmayı belirlemişlerdir. Performans belirleyiciler ve istatiksel ölçümler ile çok katmanlı perseptron sinir ağlarının değerlendirilmesi sonucunda teşhis için hızlı yayılım algoritmasının en başarılı algoritma olduğu görülmüştür [2]. Veranyurt, Deveci vd. çalışmalarında, RF, KNN ve Adaboost algoritmalarını kullanarak diyabet hastalığı teşhisi için sınıflandırma performanslarını kıyaslamışlardır. Çalışma sonucunda RF ve KNN algoritmaları %92,30 oranında başarıya ulaşmıştır. Adaboost algoritması için başarı oranının %90,59 olduğu gözlemlenmiştir. Sonuç olarak RF ve KNN algoritmalarının hastalık sınıflandırma başarısı daha yüksek bulunmuştur [3]. Deperlioğlu ve Köse diyabet hastalığının teşhisi için, Oto Kodlayıcı Sinir Ağları (OKSA) ile UCI makine öğrenme laboratuvarındaki Pima Indian diyabet veri setinden faydalanarak sınıflandırma yapmışlardır. %97,3’lük bir sınıflandırma doğruluğuna ulaşmışlardır [4]. Bir diğer çalışmada ise Çerkezi, göz hastalığı diyabetik retinopati için bir karar destek sistemi tasarlamıştır. Çalışmasında KNN ve bayes veri madenciliği sınıflandırma yöntemlerini kullanmıştır. KNN yönteminin daha iyi doğruluk oranı verdiği görülmüştür [5].

**2. YÖNTEM**

Yapılan çalışmalara göre makine öğrenmesi, hastalıkların erken tanısında en çok kullanılan yöntemlerdendir. İnsan yaşamını birçok alanda kolaylaştıran bilgisayarların, günümüzde insan gibi düşünmesini sağlamak için yapılan çalışmalar, makine öğrenimini ortaya çıkarmıştır [6]. Makine öğrenmesi, verilen bir problemi probleme ilişkin ortamdan edinilen veriye göre modelleyen bilgisayar algoritmalarının genel adıdır. Amacı, daha önce yaşanmış bir durumu tecrübe haline getirmek, problemi genelleştirmek ve sistemi karşısına gelebilecek olan yeni durumlara karşı cevap verebilecek hale getirmektir. Sınıflandırma veri setindeki noktaları, sınıf bilgisindeki benzerliklere göre gruplama işlemidir. Amaç, sınıfı belli olmayan bir veriyi modele göre sınıflandırmaktır. Bu çalışmada, bir sınıflandırma algoritması olan KNN algoritması kullanılarak sınıflandırma yapılmıştır.

**2.1. K En Yakın Komşu (KNN)**

KNN algoritmasının adından da anlaşılacağı gibi amacı, verileri sınıflandırdıktan sonra istenilen veriye en yakın K komşuyu bulmak ve buna göre yeni verinin hangi sınıfa ait olduğunu ortaya koymaktır.

Algoritma adımları şu şekildedir:

1. K değeri belirlenir.
2. Tüm verilerin hedef veriye olan uzaklıkları hesaplanır.
3. Uzaklıklar mesafesi az olandan çok olana doğru sıralanır.
4. En yakın komşu kategorileri toplanır.
5. En uygun komşu kategorisi seçilir.

**2.1.1. Mesafe Ölçümleri**

Uzaklık hesaplamak için farklı yöntemler kullanılabilmektedir.

**2.1.1.1. Şehir Mesafe Uzaklığı (Manhattan Uzaklığı)**

Adını Manhattan şehrinden almıştır. Bu yöntemde dikey ve yatay uzaklıkların mutlak değer toplamlarına bakılmaktadır. Hesaplama işlemi denklem (1) ile gerçekleştirilmektedir.

 **(1)**

**2.1.1.2. Öklid Uzaklığı (Euclidean Uzaklığı)**

İki nokta arasındaki doğrusal uzaklıktır. A ve B noktaları arasındaki Öklid uzaklığı denklem (2) ile hesaplanmaktadır.

 **(2)**

Bağıntı genelleştirildiğinde denklem (3) elde edilmektedir.

 **(3)**

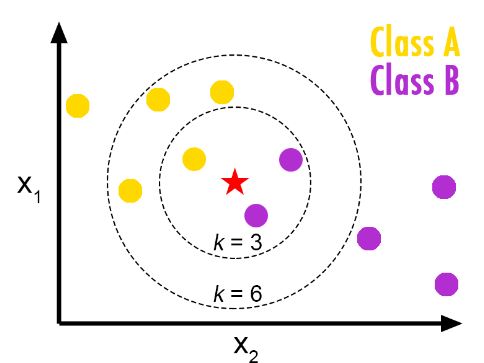
**2.1.1.3. Minkowski Uzaklığı**

Öklid uzaklığı ve Manhattan uzaklığı gibi uzaklık ölçütlerinin genelleştirilmiş halidir. Uzaklık denklem (4) ile hesaplanmaktadır.

 **(4)**

**2.1.2. Komşu Sayısı (K)**

K değeri 1 olarak alınırsa hedef veri en yakın olduğu sınıfa, 1’den büyük alınırsa yakın olduğu sınıflar içinden en fazla bulunan sınıfa dahil olmaktadır. K sayısının değeri, genelde üzerinde çalışılan veri setine göre değişiklik göstermektedir. Geçerliliği net olan bir K değeri yoktur ancak, tek sayıların kullanılması önerilmektedir. Çift sayı alınırsa iki farklı sınıf arasında eşitlik olması gibi durumlar söz konusu olabilmektedir. Şekil 1’de sınıfı belli olmayan verinin k değeri 3 alındığında B sınıfına ait olduğu görülmektedir. Ancak k değeri 6 alındığında A sınıfına ait olmaktadır.



**Şekil 1.** KNN Algoritmasında K Değeri Etkisi

**2.1.3. KNN Algoritması Avantajları ve Dezavantajları**

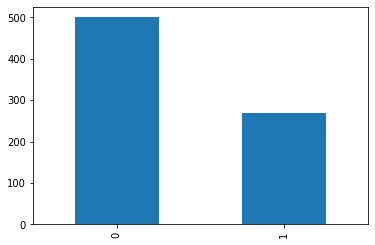
KNN algoritmasının uygulanması basittir. Algoritmanın kullanılabilmesi için eğitim setinin tamamının, algoritmanın çalıştırılacağı her zaman bulunması şarttır, çünkü bellek tabanlı bir sınıflayıcıdır. Eğitim setinin çok olması durumunda etkili sonuçlar üretmektedir. Veri seti çok büyük olduğu zaman eğitim setindeki her bir veriye olan uzaklık hesaplandığından işlem yükü oluşmaktadır ve hesaplama maliyeti yükselmektedir.

**2.2. Veri Seti ve Uygulama**

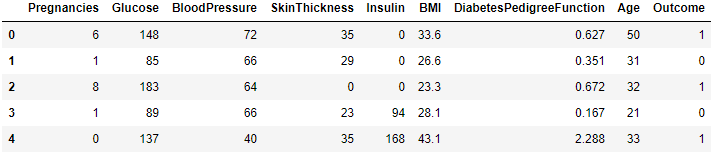
Bu çalışmada, diyabetli hastaları teşhis etmek için KNN yöntemiyle sınıflandırma yapılmıştır. Pima Indian diyabet veri seti kullanılmıştır. Veri seti 8 bağımsız değişkenden ve 1 sınıf değişkeninden oluşmaktadır. Veri setindeki değişkenler şu şekilde listelenmektedir:

1. Pregnancies: Hamile Kalma Sayısı
2. Glucose: Plazma Glikoz Konsantrasyonu
3. BloodPressure: Diyastolik Kan Basıncı
4. SkinThickness: Triceps Cilt Kıvrım Kalınlığı
5. Insulin: Serum İnsülini
6. BMI: Vücut Kitle İndeksi
7. DiabetesPedigreeFunction: Diyabet Soyağacı Fonksiyonu
8. Age: Yaş
9. Outcome: Sınıf Değişkeni (0 veya 1)

Veri setinde 768 gözlem bulunmaktadır. Hasta ve sağlıklı bireylere ilişkin bar grafiği Şekil 2’de gösterilmektedir.

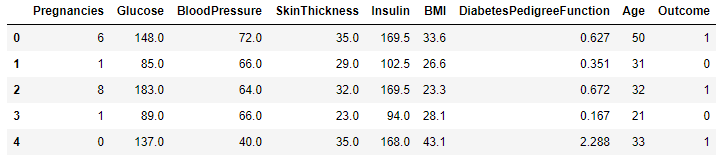


**Şekil 2.** Diyabet Hastası ve Sağlıklı Bireyler



**Tablo 1.** Veri Setinin İlk 5 Satırı (Kayıp Veriler ile)

Tablo 1’de görüldüğü üzere veri setindeki ilk 3 verinin insülin sütunundaki değerleri eksiktir. Bu gibi kayıp verileri çözümlemek için silme ya da doldurma işlemleri uygulanmaktadır. Bu çalışmada, kayıp verileri çözümlemek için medyan doldurma işlemi gerçekleştirilmiştir ve Tablo 2 elde edilmiştir.



**Tablo 2.** Veri Setinin İlk 5 Satırı (Çözümlenmiş)

Veri setinin %80’i eğitim, %20’si ise test verisi olarak ayrılmıştır. 10 katlı çapraz doğrulama işlemi yapılmıştır ve komşu sayısı 12 olarak belirlenmiştir. Bu bağlamda, modelin doğruluk oranı, eğitim ve test hataları hesaplanmıştır.

**3. SONUÇ**

Diyabet oldukça tehlikeli ve farklı rahatsızlıklara sebep olabilen bir hastalıktır. Hasta sayısı hızla artış göstermektedir. Bu bağlamda, diyabet hastalarını ve sağlıklı bireyleri KNN yöntemiyle sınıflandırma çalışması yapılmıştır. 768 gözlem bulunan Pima Indian diyabet veri seti kullanılmıştır. Veri setindeki kayıp verileri çözümlemek için medyan doldurma işlemi gerçekleştirilmiştir. Veri setinin %80’i eğitim, %20’si ise test verisi olarak ayrılmıştır. Modelin tahmin performansını iyileştirme çalışmaları yapılmıştır. 10 katlı çapraz doğrulama işlemi yapılmıştır ve komşu sayısı 12 olarak belirlenmiştir. Eğitim hatası 0,34 ve test hatası 0,35 olarak elde edilmiştir. Modelin sınıflandırma doğruluğu %87,66’dır. Bu durum modelin başarılı olduğunu göstermekte ve hastalıkların teşhis edilme süreçlerinde makine öğrenmesinden faydalanılabileceğini ortaya koymaktadır.

**KAYNAKÇA**

1. Köse U. (2019). Zeki Optimizasyon Tabanlı Destek Vektör Makineleri ile Diyabet Teşhisi, Politeknik Dergisi, 557-566.
2. Güler İ., Übeyli E.D. (2006). Çok Katmanlı Perseptron Sinir Ağları ile Diyabet Hastalığının Teşhisi, Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, Cilt 21, No 2, 319-326.
3. Veranyurt Ü., Deveci A.F., Esen M.F., Veranyurt O. (2020). Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-Nearest Neighbour ve Adaboost Algoritmaları Uygulaması, Uluslararası Sağlık Yönetim ve Stratejileri Araştırma Dergisi, Cilt 6, 275-286.
4. Deperlioğlu Ö., Köse U. (2018). Derin Sinir Ağları Kullanarak Diabet Teşhisi, Tıp Teknolojileri Kongresi, 66-69.
5. Çerkezi M. (2013). Veri Madenciliği Yöntemlerini Kullanarak Diyabetik Retinopati Hastalığının Teşhisi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi.
6. Yılmaz A., Kaya U., Dikmen Y. (2019). Sağlık Alanında Kullanılan Derin Öğrenme Yöntemleri, Avrupa Bilim ve Teknoloji Dergisi, Sayı 16, 792-808.
7. Altay O. (2016). Genetik ve Genetik Olmayan Faktörlere Bağlı Olarak Türk Hastalarda Varfarin Dozajını Tahmin Eden Bir Uzman Sistem Geliştirilmesi, Fırat Üniversitesi Fen Bilimleri Enstitüsü Yüksek Lisans Tezi.