# K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi

**Erdal Taşcı[1], Aytuğ Onan[2]**

[1] Ege Üniversitesi, Bilgisayar Mühendisliği Bölümü, İzmir

[2] Celal Bayar Üniversitesi, Bilgisayar Mühendisliği Bölümü, Manisa

arif.erdal.tasci@ege.edu.tr, aytug.onan@cbu.edu.tr

**Özet:** K-en yakın komşu algoritması (K-NN), gerçekleştiriminin basit ve kolay, öğrenme sürecinin güçlü ve kullanışlı olmasından dolayı sınıflandırmada yaygın biçimde kullanılmaktadır. Makine öğrenmesi, veri madenciliği gibi çok çeşitli alanlarda uygulanmaktadır. Bu çalışmada, UCI Machine Learning Repository'de bulunan 6 farklı gerçek dünya veri seti için K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisi incelenmiş ve tartışılmıştır. Çalışma kapsamında, k komşu sayısı, uzaklık ve ağırlıklandırma ölçütlerini içeren parametreler kullanılmış, sınıflandırıcının farklı parametre değerlerinde doğruluk oranı ölçülerek performansı test edilmiştir.

**Anahtar Sözcükler:** K-En Yakın Komşu, Sınıflandırma, Parametre, Performans, Makine Öğrenmesi, Veri Madenciliği.

**The Investigation of Performance Effects of K-Nearest Neighbor Algorithm Parameters on Classification**

**Abstract:** K-nearest neighbor algorithm (K-NN) is widely used in classification owing to its simplicity to implement and learn and its robustness and usefulness in the learning process. K-NN is applied in a variety of fields, such as machine learning, data mining. In this study, the classification performance of K-nearest neighbor algorithm is investigated and discussed based on the different parameters of the algorithm by using 6 real-world datasets on UCI Machine Learning Repository. In the study, number of neighbors (*k*), distance functions and weighting functions are utilized as comparative parameters. Based on the parameter values, the predictive performance (in terms of classification accuracy) of classification algorithms is performed.

**Keywords:** K-Nearest Neighbor, Classification, Parameter, Performance, Machine Learning, Data Mining.

## 1. Giriş

K-NN algoritması, T. M. Cover ve P. E. Hart tarafından önerilen, örnek veri noktasının bulunduğu sınıfın ve en yakın komşunun, k değerine göre belirlendiği bir sınıflandırma yöntemidir [1]. Bu algoritma, en iyi bilinen, eski, basit ve etkili örüntü sınıflandırma yöntemlerinden biridir ve makine öğrenme algoritmaları arasında popüler olarak kullanılmaktadır [2, 3, 4]. Nesnelerin sınıflandırılması önemli bir araştırma alanıdır ve örüntü tanıma, veri madenciliği, yapay zekâ, istatistik, bilişsel psikoloji, tıp, biyoinformatik gibi çok çeşitli alanlarda uygulanmaktadır [5, 6].

K-NN algoritması, eğitiminin olmaması, gerçekleştiriminin kolay, analitik olarak izlenebilir, yerel bilgilere uyarlanabilir, paralel gerçekleştirime uygun, gürültülü eğitim verilerine karşı dirençli olması gibi avantajları ile sınıflandırma uygulamalarında özellikle tercih edilmektedir [2]. Bu avantajlara rağmen, yüksek miktarda bellek alanına gereksinim duyması, veri seti ve öznitelik boyutu arttıkça işlem yükünün ve maliyetin önemli ölçüde yükselmesi, performansın k komşu sayısı, uzaklık ölçütü ve öznitelik sayısı gibi parametre ve özelliklere bağlı olarak etkilenmesi gibi birtakım dezavantajları da beraberinde getirmektedir [2, 7].

Bu çalışmada, K-NN algoritmasında k komşuluk sayısı, uzaklık ve ağırlıklandırma ölçütleri gibi parametrelerin sınıflandırma performansını nasıl etkilediği incelenmiştir. Farklı veri setleri üzerinde değişen parametre değerlerine göre sınıflandırma yapılarak performans sonuçları değerlendirilmiş ve tartışılmıştır.

Çalışmanın geri kalan bölümleri şu şekilde organize edilmiştir: İkinci bölümde, K-NN algoritması tanımlanmış ve bu algoritmanın genel işleyişi, avantajları, dezavantajları hakkında bilgi verilmiştir. Üçüncü bölümde, K-NN parametreleri açıklanmıştır. Dördüncü bölümde, çalışma kapsamında kullanılan materyal ve yöntemler açıklanmış, deneysel sonuçlar sunulmuştur. Son bölümde ise sonuç, tartışma ve önerilere yer verilmiştir.

## 2. K-NN

K-NN algoritması, en temel örnek tabanlı öğrenme algoritmaları arasındadır. Örnek tabanlı öğrenme algoritmalarında, öğrenme işlemi eğitim setinde tutulan verilere dayalı olarak gerçekleştirilmektedir. Yeni karşılaşılan bir örnek, eğitim setinde yer alan örnekler ile arasındaki benzerliğe göre sınıflandırılmaktadır [8]. K-NN algoritmasında, eğitim setinde yer alan örnekler *n* boyutlu sayısal nitelikler ile belirtilir. Her örnek *n* boyutlu uzayda bir noktayı temsil edecek biçimde tüm eğitim örnekleri *n* boyutlu bir örnek uzayında tutulur. Bilinmeyen bir örnek ile karşılaşıldığında, eğitim setinden ilgili örneğe en yakın *k* tane örnek belirlenerek yeni örneğin sınıf etiketi, *k* en yakın komşusunun sınıf etiketlerinin çoğunluk oylamasına göre atanır [9]. Algoritma 1'de K-NN algoritmasının genel işleyişi özetlenmektedir:

**Eğitim Algoritması**

- Eğitim setinde yer alan her bir örneği ($x$, $f(x)$) eğitim örnekleri listesine ekle.

**Sınıflandırma Algoritması**

- Sınıflandırılmak üzere verilen $x_q$ örneğini aşağıdaki kurala göre sınıfla:
  - Eğitim örnekleri arasında yer alan $x_1$, ..., $x_k$, $x_q$ örneğine en yakın *k* tane örneği temsil etmek üzere, $x_q$ örneğinin sınıf etiketinin belirlenmesi:

$$f(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$$

  - Burada, *a* ve *b* eşit olduğu takdirde $\delta(a,b)=1$ olarak, aksi takdirde $\delta(a,b)=0$ olarak alınacaktır.

**Algoritma 1:** K-NN algoritmasının genel işleyişi [8]

K-NN algoritmasının performansı için kritik öneme sahip noktalardan birisi örnekler arası

yakınlığın nasıl ölçümleneceğidir. Yakınlık, Öklid uzaklığı ya da bir başka uzaklık ölçütü kullanılarak hesaplanabilir. Şekil 1'den de görülebildiği gibi, K-NN algoritması basit bir yapıya sahiptir ve az sayıda parametre gerektirmektedir. Temel K-NN algoritmasında, sınıf etiketinin çoğunluk oylamasına dayalı olarak belirlenmesi, simetrik olmayan dağılıma sahip veri setlerinde sıklıkla görülen sınıfların, yeni örneklerin sınıf etiketlerinin belirlenmesinde daha baskın bir role sahip olmalarına neden olmaktadır [10]. Bu nedenle, temel K-NN algoritmasının uzaklık ölçütünün etki değerine farklı şekillerde ağırlık değeri atayan yöntemler bulunmaktadır [11].

K-NN algoritması, büyük eğitim setlerinin varlığında, oldukça etkin sonuçlar verebilmektedir. K-NN algoritması, ilgisiz özniteliklerin varlığında da sınıflandırma modeli oluşturabilmektedir. Böyle durumlarda eğitim için gereken süre oldukça artmaktadır [12]. K-NN algoritması basit yapısına karşın, yüksek bir hesaplama maliyetine sahiptir. Sınıf etiketi belirlenmek istenen örneğin, veri setinde yer alan örnekler ile arasındaki uzaklığın belirlenmesi, özellikle büyük eğitim veri setleri için oldukça maliyetli olabilmektedir. Bu maliyeti ortadan kaldırmak için, K-NN algoritması temel bileşenler analizi gibi boyut azaltma yöntemleri ile ya da arama ağaçları gibi daha güçlü veri yapıları ile birlikte kullanılabilmektedir [13]. Bunun yanı sıra, K-NN algoritması, çok boyutlu veri setlerinde etkin değildir, yüksek bellek gereksinimlerine sahiptir, komşu sayısı, uzaklık ölçütü gibi parametrelere duyarlıdır [14].

## 3. K-NN Parametreleri

K-NN algoritmasının performansında etkili ve önemli parametreler uzaklık ölçütü, komşu sayısı ($k$) ve ağırlıklandırma yöntemidir. Alt bölümlerde bu parametreler açıklanmaktadır.

### 3.1 Uzaklık Ölçütleri

Uzaklık ölçütleri olarak, Minkowski, Öklid, Manhattan, Chebyschev ve Dilca uzaklığı kullanılmaktadır.

#### 3.1.1 Minkowski Uzaklığı

Minkowski uzaklığı, Öklid uzayında tanımlı bir dizidir. Sınıflandırma, kümeleme gibi makine öğrenmesi, veri madenciliği uygulamalarında sıklıkla kullanılan Öklid uzaklığı, Manhattan uzaklığı gibi uzaklık ölçütlerinin genelleştirilmiş halidir. Herhangi iki nokta $P$ ve $Q$ arasındaki Minkowski uzaklığı $P=(x_1, x_2, \ldots, x_n)$ ve $Q=(y_1, y_2, \ldots, y_n)$ olmak üzere, Eşitlik 1'e göre hesaplanır:

$$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \quad (1)$$

Minkowski uzaklığı, genel bir formül ile ifade edilmekte olup $p$'nin farklı değerleri için çeşitli uzaklık ölçütlerini tanımlamak amacıyla da kullanılmaktadır. Minkowski ölçütünün $p=2$ olduğu özel durumu, Öklid uzaklığını, $p=1$ olduğu özel durumu Manhattan uzaklığını ve $n\to\infty$ olduğu özel durum, Chebyschev uzaklığını vermektedir [15].

#### 3.1.2 Öklid Uzaklığı

Öklid uzaklığı, sınıflandırma ve kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür. Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklık olup herhangi iki nokta, $P$ ve $Q$ arasındaki Öklid uzaklığı $P=(x_1, x_2, \ldots, x_n)$ ve $Q=(y_1, y_2, \ldots, y_n)$ olmak üzere, Eşitlik 2'ye göre hesaplanır [15]:

$$\left(\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right) \quad (2)$$

Öklid uzaklığı, K-ortalama kümeleme algoritması, temel K-NN algoritması gibi sınıflandırma ve kümeleme algoritmalarında yakınlığın ölçümlenmesi için kullanılan temel uzaklık ölçütüdür.

#### 3.1.3 Manhattan Uzaklığı

Manhattan uzaklığı, $n$ boyutlu iki nokta arasındaki farkların mutlak değerlerinin toplamıdır. Herhangi iki nokta, $P$ ve $Q$ arasındaki Manhattan uzaklığı $P=(x_1, x_2, \ldots,$

$x_n$) ve $Q=(y_1, y_2, \ldots, y_n)$ olmak üzere, Eşitlik 3'e göre hesaplanır [15]:

$$\left(\sum_{i=1}^{n}|x_i - y_i|\right) \quad (3)$$

### 3.1.4 Chebyschev Uzaklığı

Chebyschev uzaklığı (maksimum değer uzaklığı), Minkowski uzaklığının, $n\to\infty$ olduğu özel durum olup, iki nokta arasındaki farkların mutlak değerlerinin maksimumu olarak tanımlanmaktadır. Herhangi iki nokta, *P ve Q* arasındaki Chebyschev uzaklığı $P=(x_1, x_2, \ldots, x_n)$ ve $Q=(y_1, y_2, \ldots, y_n)$ olmak üzere, Eşitlik 4'e göre hesaplanır [16].

$$\lim_{p\to\infty}\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} = \max_{i=1}^{n}|x_i - y_i| \quad (4)$$

### 3.1.5 Dilca Uzaklığı

Dilca (Distance Learning in Categorical Attribute) uzaklığı, kategorik öznitelik değerleri arasındaki uzaklığı ölçümlemek için kullanılan iki aşamalı bir ölçüttür [17]. Bu ölçütte öncelikle, simetrik belirsizlik katsayısı yöntemi kullanılarak öznitelik seçimi işlemi gerçekleştirilerek eş-oluşum tablosu oluşturulmaktadır. Ardından, eş-oluşum tablosu üzerinde koşullu olasılık ve Öklid uzaklığına dayalı hesaplama gerçekleştirilerek uzaklık ölçümlenmektedir. Bilgi kazancı, yüksek değer içeren özniteliklere karşı taraflıdır. Simetrik belirsizlik katsayısı (*symmetrical uncertainty*) (*SU*), bilgi kazancının (*information gain*) (*IG*) bu problemini ortadan kaldırmak için, bilgi kazancının *X* ve *Y* özniteliklerinin entropi değerleri toplamına bölünmesi ile belirler. Simetrik belirsizlik katsayısı Eşitlik 5'e göre hesaplanır [18]:

$$SU = 2\times\left[\frac{IG}{H(Y)+H(X)}\right] \quad (5)$$

Öznitelik seçiminin ardından, uzaklık hesaplaması Eşitlik 6'da belirtilen formüle göre gerçekleştirilir [17]:

$$d(x_i, x_j) = \sqrt{\sum_{Y\in baglam(X)}\sum_{y_k\in Y}(P(x_i|y_k) - P(x_j|y_k))^2} \quad (6)$$

Burada, $x_i$ ve $x_k$ incelenmekte olan özniteliğin aldığı değer çiftleri ve $x_i, x_j \in X$ dir. Her bir *Y* bağlam özniteliği için $x_i$ ve $x_k$ değerlerine dayalı koşullu olasılık hesaplanıp ardından Öklid uzaklığı alınmaktadır. Her bir öznitelik için bağlam öznitelikleri, simetrik belirsizlik katsayısına dayalı olarak belirli bir sezgisel değerlendirme ölçütü aracılığıyla yapılmaktadır [17].

### 3.2 Komşu Sayısı (*k*)

K-NN algoritmasında, komşu sayısı (*k*) parametresinin değerine dayalı olarak sınıflandırma yapılmaktadır. Sınıflandırma sürecinde, $k=1$ için, sadece en yakın komşunun bulunduğu sınıfa atanırken, *k* sayısı örnek sayısına (*N*) yaklaştıkça veri setinde yer alan tüm veriler dikkate alınmakta ve oylamaya göre seçim yapılmaktadır.

### 3.3 Ağırlıklandırma

Komşular için ağırlık değerleri atanması ile sınıflandırılmakta olan örneğe daha yakın olan komşu örneklerin, çoğunluk oylamasına daha fazla katkı koyması amaçlanır. En çok kullanılan ağırlık değeri atama yöntemleri, her bir komşunun ağırlığının, *d*, komşular arası uzaklık olmak üzere, $1/d$ ya da $1/d^2$ şeklinde alınmasıdır [19].

## 4. Materyal, Yöntem ve Sonuçlar

Bu bölümde, kullanılan materyallere, parametrelere ve yöntemlere ilişkin bilgiler verilmiş, deneysel sonuçlar sunulmuştur.

### 4.1. Veri Setleri

Bu çalışmada, K-NN algoritmasının farklı parametre türlerine ve değerlerine ilişkin sınıflandırma performansının incelenmesi için, makine öğrenmesi alanında ve sınıflandırma problemlerinde yaygın olarak kullanılan altı farklı veri setinden yararlanılmıştır. Bu kapsamda, UCI Machine Learning Repository'de yer alan, "Breast

Cancer Wisconsin", "Cardiotocography", "Ionosphere", "Leaf", "Parkinsons" ve "Thoracic Surgery" veri setleri kullanılmıştır [20]. Veri setlerine ilişkin temel özellikler, Tablo 1'de özetlenmiştir.

**Tablo 1:** Veri setlerine ilişkin temel özellikler

| Veri Seti | Örnek Sayısı | Öznitelik Sayısı | Sınıf Sayısı |
|---|---|---|---|
| Breast Cancer Wisconsin | 699 | 10 | 2 |
| Cardiotocography | 2126 | 23 | 3 |
| Ionosphere | 351 | 34 | 2 |
| Leaf | 340 | 16 | 30 |
| Parkinsons | 197 | 23 | 2 |
| Thoracic Surgery | 470 | 17 | 2 |

### 4.2. Deneysel Süreç

Deneysel çalışmalar, WEKA 3.7.11 yazılımı kullanılarak gerçekleştirilmiştir. Bu çalışmada, K-NN algoritmasının, sınıflandırma performansındaki farklı parametre değerlerinin etkisini incelemek amacıyla, her bir veri seti için algoritmanın komşu sayısı parametresi ($k$), uzaklık ölçütü parametresi ve yönteme ağırlık değeri atama yöntemi ($1/d$) uygulanıp uygulanmayacağı şeklinde 100'er farklı K-NN yapılandırması ele alınmıştır. Komşu sayısı parametresi için 1'den 10'a kadar olan değerler kullanılmıştır. Uzaklık ölçütü kapsamında, Minkowski, Öklid, Manhattan, Chebyschev ve Dilca uzaklıklarından oluşan 5 farklı durum incelenmiştir. Veri setlerinin eğitim ve test setleri olarak ayrılmasında ise, 10-kat çapraz geçerleme yöntemi kullanılmış, sınıflandırıcının genelleştirme performansı hesaplanmıştır.

**Tablo 2:** Hata matrisi [21]

| | T / G | Gerçek Sınıf | |
|---|---|---|---|
| | | p | n |
| Tahminlenen Sınıf | Y | Gerçek Pozitif (GP) | Yanlış Pozitif (YP) |
| | N | Yanlış Negatif (YN) | Gerçek Negatif (GN) |
| Sütun Toplamı | | P | N |

$$ACC = \frac{GP + GN}{P + N} \qquad (7)$$

Farklı parametre değerlerinin sınıflandırma performanslarına etkisinin incelenmesinde, sınıflandırma doğruluk oranı (Accuracy) (ACC) kullanılmıştır. Sınıflandırma doğruluk oranı, Tablo 2'de sunulan hata matrisi değerleri doğrultusunda Eşitlik 7'ye göre hesaplanmaktadır.

### 4.3 Deneysel Sonuçlar

Komşu sayısı (k) ve 5 farklı uzaklık ölçütüne göre ağırlık ataması ($1/d$) uygulanmadığında elde edilen deneysel sonuçlar Tablo 3'te gösterilmiştir. Bu tabloya göre, k değeri 1'e eşit olduğunda en iyi sonuçların, çoğunlukla Manhattan uzaklığı için sağlandığı, Öklid veya Minkowski uzaklıklarının ise komşu sayısı arttıkça daha etkili olduğu gözlemlenmektedir. Ayrıca, öznitelik sayısının artmasına bağlı olarak Minkowski uzaklığının daha uygun sonuçlar verdiği belirlenmiştir.

Komşu sayısı (k) ve 5 farklı uzaklık ölçütüne göre ağırlık ataması ($1/d$) uygulandığında elde edilen deneysel sonuçlar ise Tablo 4'te sunulmuştur. Bu tabloya göre, Breast Cancer Wisconsin ve Cardiotocography verisetleri için elde edilen en iyi sonuçların ağırlık atama yöntemi uygulandığında yükseldiği gözlemlenmiştir. Tablo 3 ve Tablo 4 birlikte incelendiğinde k=1,2 değerleri için daha başarılı sonuçların alındığı, k değerinin 5'den itibaren artmasıyla, en iyi sonuçların önemli bir kısmında sınıflandırma performansının azaldığı görülmektedir. Ayrıca, verisetinde kullanılan sınıf etiketi sayısının artmasının performansı önemli ölçüde etkilediği gözlemlenmiştir. Leaf verisetinde yer alan 30 sınıf etiketi için tüm durumlarda en iyi %66.26'lık doğruluk oranına erişilebilmiştir.

**Tablo 3:** Ağırlık ataması (1/*d*) uygulanmadığında, komşu sayısı (k) ve uzaklık ölçütlerine göre elde edilen sınıflandırma performansı

| Uzaklık | Veriseti | Sınıflandırma Doğruluk Oranı (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1-NN | 2-NN | 3-NN | 4-NN | 5-NN | 6-NN | 7-NN | 8-NN | 9-NN | 10-NN |
| Dilca | ionosphere | 89,77 | 89,91 | 90,45 | **90,74** | 90,00 | 90,57 | 89,38 | 90,46 | 89,21 | 89,95 |
| Manhattan | ionosphere | **90,74** | 90,45 | 88,80 | 90,09 | 88,49 | 89,69 | 88,35 | 88,95 | 87,12 | 88,09 |
| Minkowski | ionosphere | 87,10 | **89,77** | 86,02 | 87,21 | 85,10 | 85,76 | 84,30 | 85,22 | 84,30 | 84,87 |
| Euclidean | ionosphere | 87,10 | **89,77** | 86,02 | 87,21 | 85,10 | 85,76 | 84,30 | 85,22 | 84,30 | 84,87 |
| Chebyshev | ionosphere | 87,10 | **87,69** | 82,25 | 80,43 | 78,69 | 79,41 | 78,83 | 79,40 | 79,18 | 79,78 |
| Dilca | breast-cancer-wisconsin | 93,46 | 91,43 | 94,12 | 93,52 | **94,66** | 94,28 | 94,45 | 93,98 | 94,22 | 94,02 |
| Manhattan | breast-cancer-wisconsin | 96,44 | 95,29 | **96,65** | 96,12 | 96,10 | 95,90 | 96,24 | 96,38 | 96,44 | 96,41 |
| Minkowski | breast-cancer-wisconsin | 95,35 | 94,56 | 96,45 | 96,14 | **96,94** | 96,60 | 96,65 | 96,64 | 96,77 | 96,60 |
| Euclidean | breast-cancer-wisconsin | 95,35 | 94,56 | 96,45 | 96,14 | **96,94** | 96,60 | 96,65 | 96,64 | 96,77 | 96,60 |
| Chebyshev | breast-cancer-wisconsin | 94,75 | 94,74 | 95,85 | 95,91 | 96,31 | **96,50** | 96,44 | **96,50** | 96,41 | 96,28 |
| Dilca | cardiotocographt-3class | 98,96 | 98,97 | **99,10** | 98,90 | 98,97 | 98,83 | 98,84 | 98,80 | 98,80 | 98,74 |
| Manhattan | cardiotocographt-3class | 99,09 | **99,13** | 99,11 | 99,07 | 99,08 | 99,05 | 99,07 | 99,04 | 99,06 | 98,94 |
| Minkowski | cardiotocographt-3class | 99,02 | 99,02 | **99,15** | 99,00 | 99,05 | 99,03 | 99,00 | 98,93 | 98,95 | 98,80 |
| Euclidean | cardiotocographt-3class | 99,02 | 99,02 | **99,15** | 99,00 | 99,05 | 99,03 | 99,00 | 98,93 | 98,95 | 98,80 |
| Chebyshev | cardiotocographt-3class | 98,89 | 98,81 | **98,90** | 98,80 | 98,62 | 98,50 | 98,47 | 98,40 | 98,40 | 98,34 |
| Dilca | leaf | **58,38** | 57,26 | 56,97 | 54,76 | 56,29 | 53,88 | 54,06 | 53,56 | 52,18 | 51,47 |
| Manhattan | leaf | **66,26** | 61,79 | 60,68 | 62,97 | 61,91 | 61,76 | 62,38 | 61,68 | 61,82 | 61,18 |
| Minkowski | leaf | **62,62** | 57,47 | 58,50 | 58,03 | 56,68 | 56,35 | 56,85 | 57,56 | 57,91 | 56,32 |
| Euclidean | leaf | **62,62** | 57,47 | 58,50 | 58,03 | 56,68 | 56,35 | 56,85 | 57,56 | 57,91 | 56,32 |
| Chebyshev | leaf | **58,74** | 53,74 | 54,26 | 49,47 | 50,65 | 50,65 | 50,65 | 49,29 | 49,74 | 48,74 |
| Dilca | parkinsons | **89,05** | 86,81 | 88,34 | 88,09 | **89,05** | 88,29 | 87,94 | 87,23 | 86,77 | 87,07 |
| Manhattan | parkinsons | 93,81 | 91,80 | **94,78** | 93,76 | 93,92 | 92,69 | 91,35 | 91,66 | 90,17 | 90,02 |
| Minkowski | parkinsons | **95,91** | 93,49 | 93,48 | 93,65 | 92,73 | 91,67 | 92,17 | 91,09 | 91,14 | 90,51 |
| Euclidean | parkinsons | **95,91** | 93,49 | 93,48 | 93,65 | 92,73 | 91,67 | 92,17 | 91,09 | 91,14 | 90,51 |
| Chebyshev | parkinsons | **92,88** | 89,58 | 90,73 | 90,11 | 90,89 | 91,24 | 90,29 | 90,20 | 89,85 | 89,44 |
| Dilca | thoracic-surgery | 78,66 | 76,00 | 82,60 | 81,11 | 83,19 | 82,79 | 84,45 | 84,43 | **84,98** | 84,81 |
| Manhattan | thoracic-surgery | 77,02 | 72,04 | 83,00 | 79,85 | 84,64 | 83,53 | 84,64 | 84,21 | **84,70** | 83,96 |
| Minkowski | thoracic-surgery | 77,02 | 71,72 | 82,81 | 80,38 | 84,79 | 83,77 | **84,91** | 84,43 | 84,87 | 84,28 |
| Euclidean | thoracic-surgery | 77,02 | 71,72 | 82,81 | 80,38 | 84,79 | 83,77 | **84,91** | 84,43 | 84,87 | 84,28 |
| Chebyshev | thoracic-surgery | 78,51 | 77,34 | 84,21 | 84,38 | 84,87 | 84,87 | 85,09 | 85,09 | **85,11** | 85,09 |

**Tablo 4:** Ağırlık ataması (1/*d*) uygulandığında, komşu sayısı (k) ve uzaklık ölçütlerine göre elde edilen sınıflandırma performansı

| Uzaklık | Veriseti | Sınıflandırma Doğruluk Oranı (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1-NN | 2-NN | 3-NN | 4-NN | 5-NN | 6-NN | 7-NN | 8-NN | 9-NN | 10-NN |
| Dilca | ionosphere | 89,77 | 89,77 | **90,48** | 89,91 | 90,03 | 90,20 | 89,38 | 89,69 | 89,21 | 89,49 |
| Manhattan | ionosphere | **90,74** | **90,74** | 88,86 | 89,49 | 88,95 | 88,61 | 88,69 | 88,18 | 87,75 | 87,49 |
| Minkowski | ionosphere | 87,10 | **87,41** | 86,02 | 86,02 | 85,33 | 85,19 | 84,30 | 84,56 | 84,30 | 84,27 |
| Euclidean | ionosphere | 87,10 | **87,41** | 86,02 | 86,02 | 85,33 | 85,19 | 84,30 | 84,56 | 84,30 | 84,27 |
| Chebyshev | ionosphere | **87,10** | 86,56 | 81,88 | 79,49 | 78,72 | 78,64 | 78,83 | 78,92 | 79,18 | 79,09 |
| Dilca | breast-cancer-wisconsin | 93,46 | 93,46 | 94,15 | 94,46 | 94,41 | **94,71** | **94,71** | 94,51 | 94,59 | 94,51 |
| Manhattan | breast-cancer-wisconsin | 96,44 | 96,47 | 96,78 | **96,81** | 96,57 | 96,75 | 96,57 | 96,73 | 96,64 | 96,50 |
| Minkowski | breast-cancer-wisconsin | 95,35 | 95,47 | 96,45 | 96,25 | 96,95 | **97,04** | 96,74 | 96,97 | 96,78 | 96,77 |
| Euclidean | breast-cancer-wisconsin | 95,35 | 95,47 | 96,45 | 96,25 | 96,95 | **97,04** | 96,74 | 96,97 | 96,78 | 96,77 |
| Chebyshev | breast-cancer-wisconsin | 94,75 | 95,09 | 96,10 | 96,30 | 96,41 | **96,55** | 96,52 | 96,48 | 96,38 | 96,22 |
| Dilca | cardiotocographt-3class | 98,96 | 98,98 | 99,03 | **99,06** | 98,93 | 98,99 | 98,86 | 98,89 | 98,84 | 98,85 |
| Manhattan | cardiotocographt-3class | 99,09 | 99,09 | 99,13 | **99,15** | 99,05 | 99,01 | 99,01 | 98,99 | 99,02 | 98,99 |
| Minkowski | cardiotocographt-3class | 99,02 | 99,02 | 99,20 | **99,23** | 99,05 | 99,05 | 99,00 | 99,00 | 98,96 | 99,01 |
| Euclidean | cardiotocographt-3class | 99,02 | 99,02 | 99,20 | **99,23** | 99,05 | 99,05 | 99,00 | 99,00 | 98,96 | 99,01 |
| Chebyshev | cardiotocographt-3class | 98,89 | 98,90 | **99,04** | 98,97 | 98,71 | 98,65 | 98,61 | 98,51 | 98,43 | 98,42 |
| Dilca | leaf | 58,38 | 58,38 | 58,35 | **58,56** | 56,97 | 56,82 | 56,09 | 55,35 | 55,03 | 54,59 |
| Manhattan | leaf | **66,26** | **66,26** | 64,88 | 65,41 | 64,06 | 64,68 | 64,03 | 64,59 | 64,76 | 64,38 |
| Minkowski | leaf | **62,62** | **62,62** | 61,32 | 60,79 | 59,88 | 61,26 | 61,44 | 62,03 | 62,59 | 61,82 |
| Euclidean | leaf | **62,62** | **62,62** | 61,32 | 60,79 | 59,88 | 61,26 | 61,44 | 62,03 | 62,59 | 61,82 |
| Chebyshev | leaf | **58,74** | **58,74** | 58,56 | 57,00 | 55,47 | 56,59 | 56,24 | 55,62 | 55,50 | 54,94 |
| Dilca | parkinsons | 89,05 | 89,10 | 88,34 | 88,85 | **89,16** | 89,15 | 88,29 | 88,14 | 87,54 | 87,83 |
| Manhattan | parkinsons | 93,81 | 93,81 | 94,83 | **95,35** | 94,43 | 94,43 | 92,53 | 92,67 | 92,27 | 92,42 |
| Minkowski | parkinsons | **95,91** | **95,91** | 94,51 | 94,26 | 93,34 | 92,52 | 93,50 | 92,06 | 93,34 | 92,26 |
| Euclidean | parkinsons | **95,91** | **95,91** | 94,51 | 94,26 | 93,34 | 92,52 | 93,50 | 92,06 | 93,34 | 92,26 |
| Chebyshev | parkinsons | **92,88** | **92,88** | 91,76 | 92,42 | 92,22 | 92,16 | 91,58 | 90,56 | 90,77 | 90,46 |
| Dilca | thoracic-surgery | 78,66 | 78,85 | 81,28 | 81,74 | 82,15 | 82,49 | 82,87 | 83,23 | 83,40 | **83,57** |
| Manhattan | thoracic-surgery | 77,02 | 77,02 | 81,11 | 80,91 | 82,57 | 82,74 | 83,60 | 83,87 | 83,94 | **84,49** |
| Minkowski | thoracic-surgery | 77,02 | 77,02 | 81,32 | 81,30 | 82,89 | 82,72 | 83,77 | 83,66 | **84,51** | 84,40 |
| Euclidean | thoracic-surgery | 77,02 | 77,02 | 81,32 | 81,30 | 82,89 | 82,72 | 83,77 | 83,66 | **84,51** | 84,40 |
| Chebyshev | thoracic-surgery | 78,51 | 80,11 | 83,47 | 84,40 | 84,00 | 84,34 | 84,91 | 84,91 | 84,94 | **85,02** |

## 5. Sonuç, Tartışma ve Öneriler

Bu çalışmada, K-NN algoritmasının komşu sayısı (*k*), uzaklık ve ağırlık fonksiyonlarına ilişkin parametrelerinin, sınıflandırma performansını ne ölçüde etkilediği incelenmiştir. *k* değerinin uygun değerde seçilmesi oldukça önem taşımaktadır. *k* değeri büyüdükçe daha düzgün karar sınırları oluşmasına karşın hesaplama yükü artacak, k değeri küçüldükçe ise K-NN gürültülü veriye daha hassas olacak fakat hızlı çalışacaktır. Uzaklık fonksiyonları da örneklerin dağılımına uygun olarak seçilmelidir.

Parametrelerin etkisi dışında, kullanılan veri setindeki öznitelik sayısı, özniteliklerin ayırt edicilik kriteri gibi etkenlere de bağlı olarak sınıflandırma performansının önemli oranda değişebileceği sonucuna varılmıştır. Öznitelik sayısının artış göstermesiyle boyut artış göstermekte ve boyutun kapladığı bölgeye düşen nokta sayısı azalacaktır. Bu durum sınıflandırma başarısının ciddi oranda düşmesine sebep olacaktır. İlişkisiz özniteliklerin artması, ayırt edicilik kriteri yüksek olan özniteliklerin bilgilerini etkisiz hale getirecek ve K-NN algoritması parametrelerinin güvenilirliğini azaltacaktır. Bu kapsamda, parametrelerin daha etkili sonuçlar vermesi açısından özniteliklerin seçimi, sınıflandırma öncesi önişleme aşaması olarak kullanılmalıdır. Ayrıca, farklı ağırlıklandırma ölçütleri geliştirilerek veya parametre eniyilemesi yapılarak sınıflandırma performansı artırılabilir.

## 6. Teşekkür

2211-Yurt İçi Doktora Burs Programı kapsamında sağladığı destekten ötürü TÜBİTAK Bilim İnsanı Destekleme Daire Başkanlığı birimine teşekkür ederiz.

## 7. Kaynaklar

[1] Cover, T.M. and Hart, P.E., "Nearest neighbor pattern classification". **IEEE Transactions on Information Theory,** IT-13(1):21–27 (1967).

[2] Bhatia, N. and Vandana, "Survey of nearest neighbor techniques", **International Journal of Computer Science and Information Security**, 8(2):302-305 (2010).

[3] Qiu, X.Y., Kang, K. and Zhang, H.X., "Selection of kernel parameters for K-NN", **IEEE International Joint Conference on Neural Networks (IJCNN),** 61-65 (2008).

[4] Batista, G.E.A.P.A. and Silva, D.F., "How k-nearest neighbor parameters affect its performance", **Simposio Argentino de Inteligencia Artificial (ASAI 2009),** 95–106 (2009).

[5] Keller, J. M., Gray, M.R. and Givens, J.A., "A fuzzy K-nearest neighbor algorithm", **IEEE Transactions on Systems, Man and Cybernetics**, SMC-15(4):580-585 (1985).

[6] Mao, C., Hu, B., Wang, M. and Moore, P., "Learning from neighborhood for classification with local distribution characteristics", **IEEE International Joint Conference on Neural Networks (IJCNN),** 1-8 (2015).

[7] Liu, H. and Zhang, S., "Noisy data elimination using mutual k-nearest neighbor for classification mining", **Journal of Systems and Software,** 85(5):1067-1074 (2012).

[8] Mitchell, T., "Machine Learning", **McGraw Hill,** New York, (1997).

[9] Han, J. and Kamber, M., "Data mining: concepts and techniques", **Morgan Kaufmann Publishers,** Burlington, (2006).

[10] Coomans, D and Massart, D.L., "Alternative k-nearest neighbour rules in supervised pattern recognitition: Part 1. K-nearest neighbour classification by using

alternative voting rules", **Analytica Chimica Acta,** 136: 15-27 (1982).

[11] Gartner, T., Lloyd, J.W. and Flach, P.A., "Kernels and Distances for Structured Data", **Machine Learning,** 57(3):205-232 (2004).

[12] Aha, D.W., Kibler, D. and Goldstone, R.L., "Instance-based learning algorithms", **Machine Learning,** 6:37-66 (1991).

[13] Shmueli, G., Patel, N.R. and Bruce, P.C., "Data mining: for Business Intelligence", **John Wiley & Sons,** New Jersey, (2010).

[14] Duda, R.O., Hart, P.E. and Stork, D.G., "Pattern Classification", **John Wiley & Sons,** New Jersey, (2000).

[15] Kresse, W. and Danko, D.M., "Springer Handbook of Geographic Information", **Springer-Verlag,** Berlin, (2012).

[16] Xu, G., Zong, Y. and Yang, Z., "Applied Data Mining", **CRC Press,** New York, (2013).

[17] Ienco, D., Pensa, G. and Meo, R., "From context to distance: learning dissimilarity for categorical data clustering", **ACM Transactions on Knowledge Discovery,** 6(1): 1-27 (2012).

[18] Hall, M.A. and Holmes, G., "Benchmarking attribute selection techniques for discrete class data mining", **IEEE Transactions on Knowledge and Data Engineering,** 15(6) 1437-1447 (2003).

[19] Doad, P.K. and Bartere, M.M., "A Review : Study of Various Clustering Techniques", **International Journal of Engineering Research & Technology,** 2(11):3141-3145 (2013).

[20] Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], **Irvine, CA: University of California, School of Information and Computer Science**, (2013).

[21] Taşcı, E., "Akciğer tomografileri kullanılarak yapay zeka ve görüntü işleme tekniklerine dayalı otomatik nodül bölge tespit yöntemi geliştirilmesi", **Yüksek Lisans Tezi, Ege Üniversitesi,** 104p (Yayınlanmamış) (2013).