Matlab ile KNN

K-Nearest Neighborhood

Halil UĞUR

Öğrenci No: 130304015

Proje Tanımı

KNN sınıflandırıcısı oldukça güçlü ve önü açık bir yöntemdir. Elimizde bulunan bazı verileri sınıflandırmak istersek bu algoritma oldukça iş görecektir. Bildiğiniz üzere yapay sinir ağları ile sayıları tanımlamayı gerçekleştirmiştik ve oldukça başarılı bir sonuç elde ettik. Aynı veri setini kullanılarak KNN sınıflandırma yöntemini kullanacağız.

Projenin Amacı

MATLAB ile KNN sınıflandırma yöntemini kullanarak 0 – 9 arasındaki sayıları sınıflara ayırıp verilerin en iyi şekilde doğru sonuç vermesini amaçlamaktayız. Eğitim ve test verilerimizi MNIST veri tabanından yararlanarak yapacağız.

Proje Raporu

Öncelikle şu bilinmelidir ki KNN sınıflandırması oluşturulurken; komşu sayısına, uzaklık ölçme yöntemine ve sınıflandırma kurallarına bağlı olarak değişiklikler gösterebilir bu nedenle testler birbirinden farklı üç özellik ile yapılarak toplamda 27 test yapılacaktır.

Komşu sayısı (k): Yeni veriye en yakın elemanları kontrol etme sayısı.

Uzaklık ölçme tipi (t): Euclidean, Minkowski ve Mahalanobis

Sınıflandırma kuralları (s): Nearest, Random ve Consensus

Not: Mavi grafikler doğru sayısını, Kırmızı grafik ise yanlış bulunan sayıların sayısını göstermektedir.

Birinci Deneme

KNN de Kullanılan özellikler: (k = 3) (t = Minkowski) (s = Consensus)

Eğitimde kullanılan veri sayısı: 20000

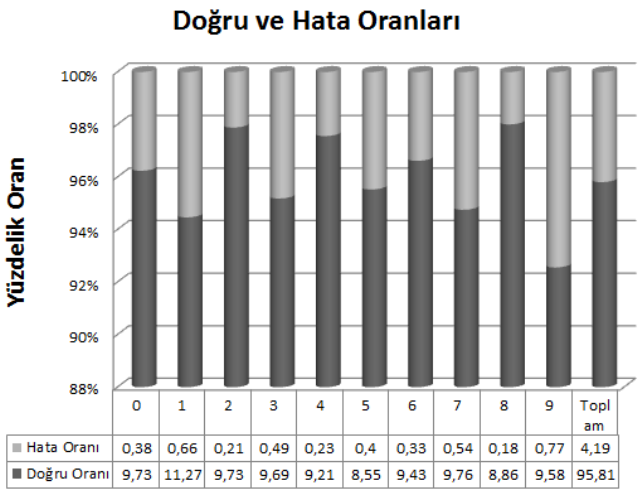
Testte kullanılan veri sayısı: 10000

Eğitimde ve Testte geçen süre: 4 dakika 50 saniye

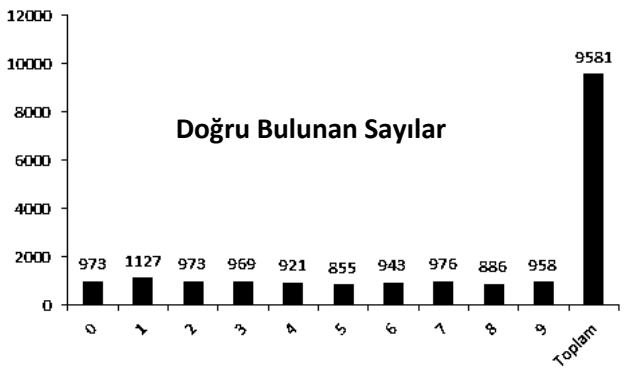
Doğruluk Oranı: %95.81

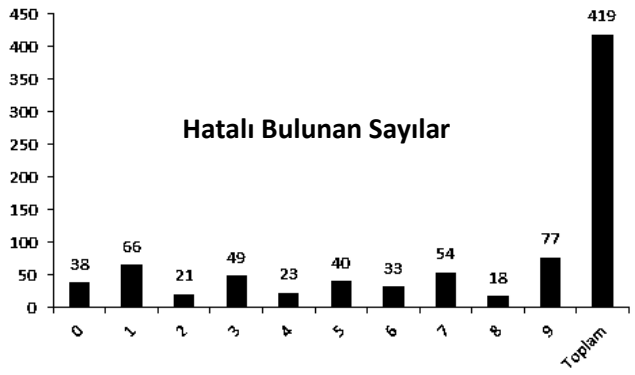
Hata Oranı: %4.19

Komşu elemanı Öklid uzaklık yöntemi kullanılarak sınıflandırma işlemi yapılmıştır. İstenilen başarı %96 ve üstü olduğundan bu deneme başarısız sayılmıştır.



	0	1	2	3	4	5	6	7	8	9	Toplam
Hata Oranı	0,38	0,66	0,21	0,49	0,23	0,4	0,33	0,54	0,18	0,77	4,19
Doğru Oranı	9,73	11,27	9,73	9,69	9,21	8,55	9,43	9,76	8,86	9,58	95,81





İkinci Deneme

KNN de Kullanılan özellikler: (k = 7) (t = Mahalanobis) (s = Random)

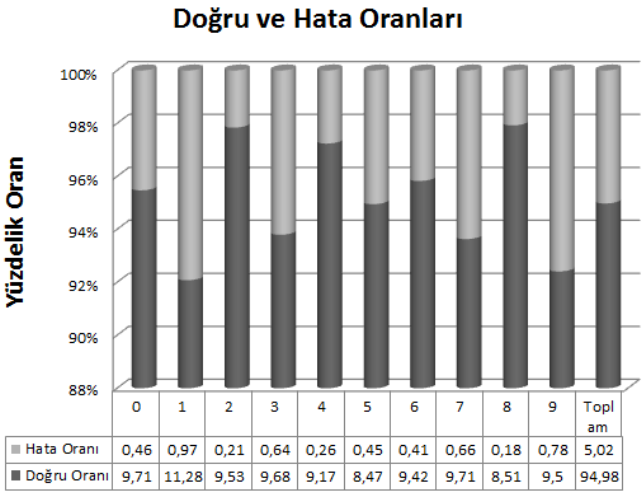
Eğitimde kullanılan veri sayısı: 20000

Testte kullanılan veri sayısı: 10000

Eğitimde ve Testte geçen süre: 4 dakika 44 saniye

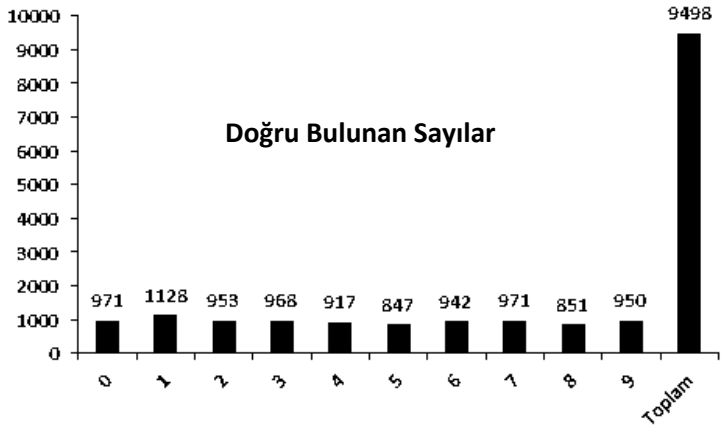
Doğruluk Oranı: %94.98

Hata Oranı: %5.02

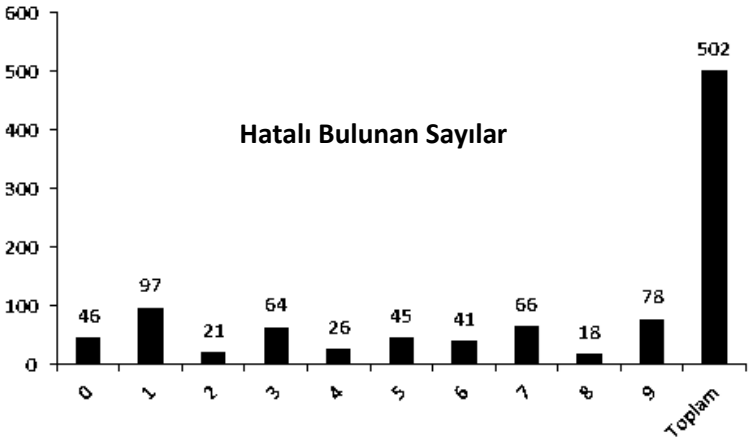


	0	1	2	3	4	5	6	7	8	9	Toplam
Hata Oranı	0,46	0,97	0,21	0,64	0,26	0,45	0,41	0,66	0,18	0,78	5,02
Doğru Oranı	9,71	11,28	9,53	9,68	9,17	8,47	9,42	9,71	8,51	9,5	94,98

Bu testte görüldüğü gibi başarı oranı düşmüştür. Bunun nedeni komşuluğun fazlaca bakılmasıdır. Bundan önceki testte oran %95 gibi bir değere sahipti ancak k değerini değiştirdikten sonra başarıda azalma görüldü. Bu nedenle komşuluk değerini artırma her zaman iyi bir sonuç üretemeyebilir. Aynı zamanda komşular arasındaki uzaklık faktörü de oldukça etkili durumda Öklid yöntemi kullanıldığında %95,5 gibi bir başarı sağladı. Yukarıda kullanılan uzaklık yöntemi ise bize %94,98 başarı getirmekte.



Sınıflandırmada kullanılan üç öznitelik için bütün denemeler yapılmış ve başarısız sonuçlanmıştır. 27 testten sonra elde edilen veriler yetersiz kaldı ve %96 değerini aşamadı bu yüzden veri setinin yarısını kullanarak daha etkili sonuç elde etmeyi deneyeceğiz.



Eğitim ve testte kullanılan uzaklık (Euclidean, Minkowski, Mahalanobis) ölçüleri sınıflandırmada oldukça etkili rol oynamakta bunun yanı sıra k sayısı da başarı oranını etkileyebilmektedir. Sınıflandırma tipi olan Nearest, Random ve Consensus, sınıflandırmada çok fazla bir etki göstermese de +-%0.5 değerlerinde oynama yapabilmektedir. Bu nedenle varsayılan olarak Nearest seçilmiştir.

Üçüncü Deneme Başarılı Sonuç

KNN de Kullanılan özellikler: (k = 1) (t = Euclidean) (s = Nearest)

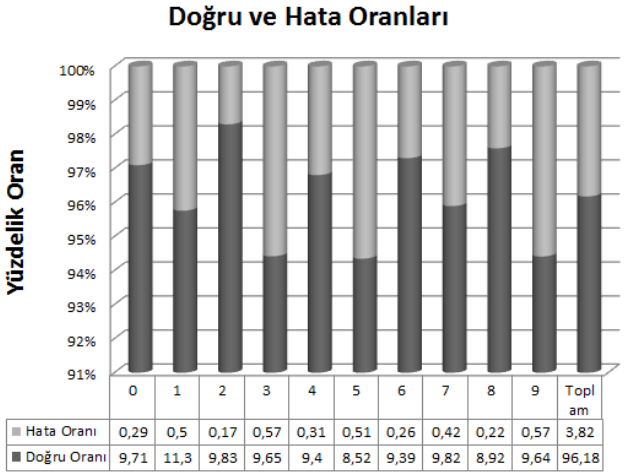
Eğitimde kullanılan veri sayısı: 30000

Testte kullanılan veri sayısı: 10000

Eğitimde ve Testte geçen süre: 7 dakika 32 saniye

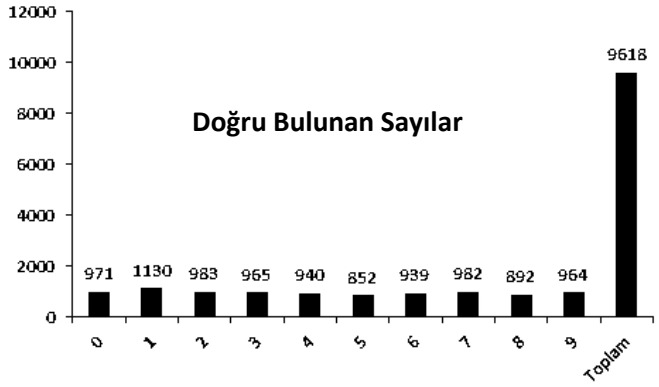
Doğruluk Oranı: %96.18

Hata Oranı: %3.82

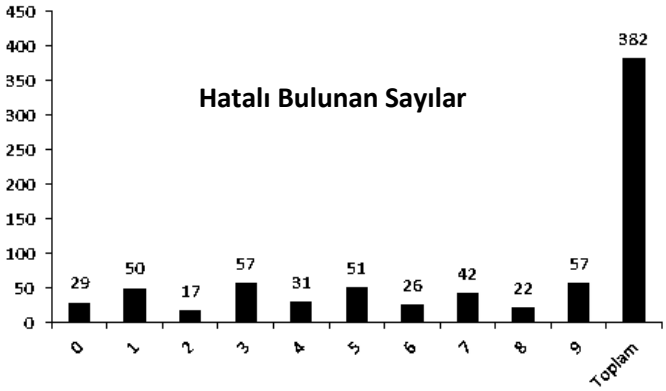


	0	1	2	3	4	5	6	7	8	9	Toplam
Hata Oranı	0,29	0,5	0,17	0,57	0,31	0,51	0,26	0,42	0,22	0,57	3,82
Doğru Oranı	9,71	11,3	9,83	9,65	9,4	8,52	9,39	9,82	8,92	9,64	96,18

Verinin artması ile birlikte elde edilecek olan doğru tahmin durumu da artabilmektedir. Verinin arama şekli kendine en çok yakın olan komşu veya komşuları değerlendirerek doğru bir başarımı yakalayabilmektedir. Son deneme veri setinde elde edilen %96.18 sonucu oldukça başarılı bir sonuç kabul edilmiş ve son durum paylaşılmıştır.



KNN algoritmasının genel yapısından bahsedecek olursak; k komşuluk durumuna göre yeni gelen bir veriye en yakın uzaklık mesafelerinin hesaplanmasından elde edilen sınıflama algoritmasıdır. Bu algoritmada veri setinin fazla olması başarıyı yukarıya taşımanın kilit anahtarıdır, k komşusu ve uzaklık ölçüleri de önemli rol oynamaktadır.



Eğitimde ve testlerde analiz yapılırken birkaç sonuca varmak mümkün oldu bunlardan birincisi; k durumunun her zaman yüksek olması başarıyı getiremeyebilir. İkincisi k değerinin her zaman tek sayı seçilmesinin en önemli nedeni, sınıflandırma sırasında aynı uzaklıkta iki farklı sınıfın eşit olması durumu söz konusu olmasıdır buda sınıflandırmanın ikilemde kalmasına neden olmaktadır.
Uzaklık ölçülerinden en etkilisi Euclidean ölçüsüdür, diğer ölçümlere göre daha etkili bir sonuç üretmektedir. Ancak şu da unutulmamalıdır ki sınıflandırılacak veriye göre uzaklık ölçüsü de değişebilmektedir. Bu veri setinde ise en etkili çalışan Euclidean bağlantısıdır.