Bilgi güvenliğini sağlamak için günümüzde çeşitli biyometrik doğrulama sistemleri kullanılmaktadır. Bunlardan biri de son zamanlarda popüler olan yüz tanıma biyometrisidir.

**LPP** (**Locality Preserving Projections**)

Bu çalışmada, Yerelliği Koruyan Projeksiyonları (LPP) yaklaşımların öğrenilmesi hedeflenmiştir. İlk olarak, ham kaydırma verilerinden elde edilen istatistiksel özelliklerin arttırılmasının performans üzerindeki etkisi araştırılmıştır. Daha sonra, KNN gibi yaygın kullanılan sınıflandırma yöntemlerinin üzerindeki etkisi incelenmiştir. Performans analizleri için literatürde yaygın kullanılan YALE veri setinden yararlanılmıştır. Yüksek boyutlu verilerin analizi birçok alanda karşımıza çıkmaktadır. Bu tür verilerin analizinde, daha az boyutla çalışmak için boyut indirgeme yöntemleri kullanılır.Genellikle lineer boyutluluk indirgeme yöntemleri daha çok tercih edilir. Bu çalışmada, popüler lineer boyutluluk indirgeme yöntemleri ve performansları araştırılmaktadır.Bu yöntem,yerelliği koruyan projeksiyon (LPP).

Boyut Azaltma (Dimensionality Reduction) veri bilimi için oldukça önemli bir yöntem. Başlıca sebepleri şöyle:

* Gerçek hayattaki veriler çok fazla boyuta (özniteliğe) sahip oluyor ve boyut büyüdükçe veri temizlemeden model kurmaya bütün süreçlerde harcamamız gereken zaman ve kaynaklar artıyor.
* Ne kadar çok boyuta sahip olursak görselleştirme de o kadar zor oluyor. 3 boyuttan sonrasını hayal etmek zor ama görsel olarak da yaptığımız çalışmanın bir karşılığı olsun insanlara anlatalım istiyoruz.
* Hemen her veri setinde bazı öznitelikler arasında yüksek korelasyon oluyor ve bu bizim gereksiz bilgiye sahip olmamıza ve modelimizde overfitting problemine sebep olabiliyor.

PCA (Principal Component Analysis) methodu yüksek boyutlu bir veri setinin boyutunu azaltmak için kullanılan en yaygın yöntemlerden biri.

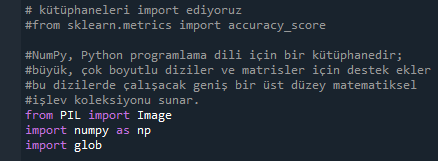
**Peki Nasıl?**

Boyut küçültmenin en kolay yolu verimizi en iyi tanımlayan öznitelikleri bulup diğerlerini atmaktır (**öznitelik seçimi — feature selection)**. Örneğin 100 boyuttan 10 tanesinin önemli olduğunu belirleyip kalan 90 özniteliği atmak ama bu tahmin edeceğiniz gibi **bilgi kaybına** sebep oluyor.

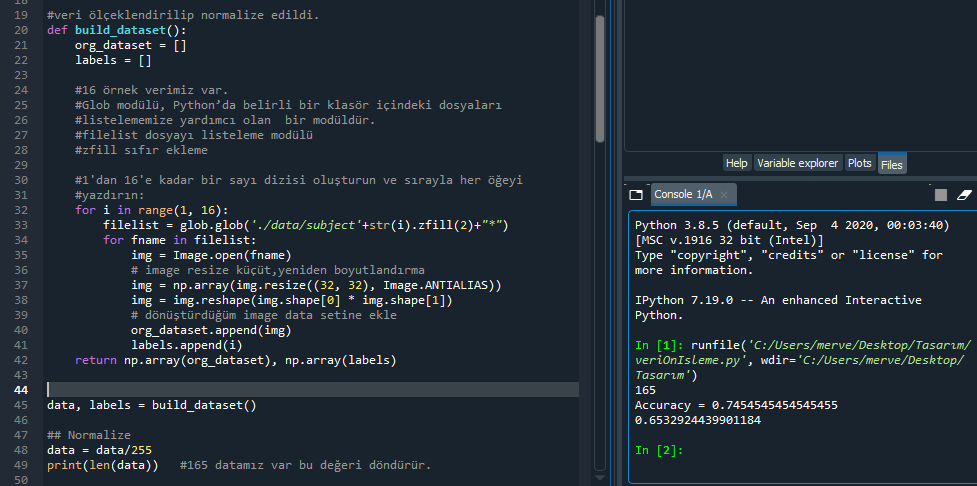
Bizim uğraşmamız gereken şey ise **öznitelik çıkarımı (feature extraction)** yapmak -en az bilgi kaybıyla boyut küçültmek-. Bunu yapmak için verideki **dağılımın maksimum varyansını-bilgisini tutan** minimum sayıda değişken oluşturuyoruz. **Eğer bir değişken her örnek için aynı değere sahip ise gereksiz bir değişkendir. Biz en yüksek varyansa sahip olan değişkenleri bulmalıyız.**

.

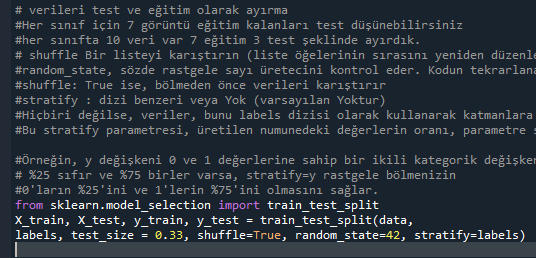
Şimdi öğrendiklerimizi makine öğrenmesinin en temel veri setlerinden biri olan yale üzerinde uygulayalım.



Kütüphaneler tamam şimdi yale veri setinde yer alan el yazısı veri setini yükleyelim ve X, y değişkenlerine bakalım.



Kütüphaneler tamam şimdi sklearn veri setinde yer alan el yazısı veri setini yükleyelim ve X, y değişkenlerine bakalım.



**Yale Veri Seti Üzerinde Kullanılan Özellikler**

Yale Yüz Veritabanı, 15 kişiden oluşan GIF formatında 165 gri tonlamalı görüntü içerir. Konu başına 11 görüntü, farklı yüz ifadesi veya konfigürasyonu başına bir tane vardır: merkez ışık, gözlüklü, mutlu, sol ışıklı, gözlüksüz, normal, sağ ışık, üzgün, uykulu, şaşırmış ve göz kırpma.

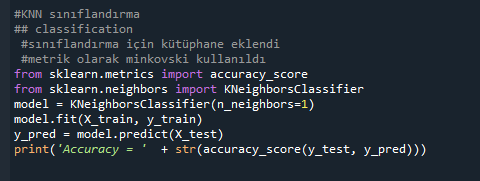
Boyut: Veri kümesinin boyutu 6.4 MB'dir ve her biri 576 görüntüleme koşulunda görülen 10 nesnenin 5760 tek ışık kaynağı görüntüsünü içerir.

Projeler: Veri seti yüz tanıma, doppelganger liste karşılaştırması vb. için kullanılabilir.Yayın Yılı: 2001

**Yale Veri Seti Üzerinde Sınıflandırma Çalışmaları**

Yale veri seti özellikleri çıkarılmış halde paylaşılmıştır. Eğitim ve test verisi oluşturulurken her sınıf için 7 görüntü eğitim kalanlarını test şeklide ayırdık. Her sınıfta 10 veri var 7 eğitim 3 test şeklinde ayırdık.Daha sonra random olarak %30 eğitim, %70 test için olmak üzere ayrılarak eğitim ve test verisi oluşturulmuştur. Sınıflandırmada KNN özellik çıkartma algoritmaları kullanılarak performansa katkıları da incelenmiştir.

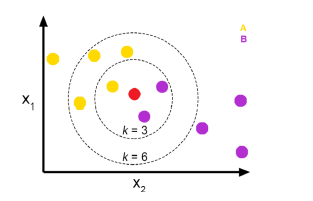
Sınflandırma



**K-En Yakın Komşu Algoritması**

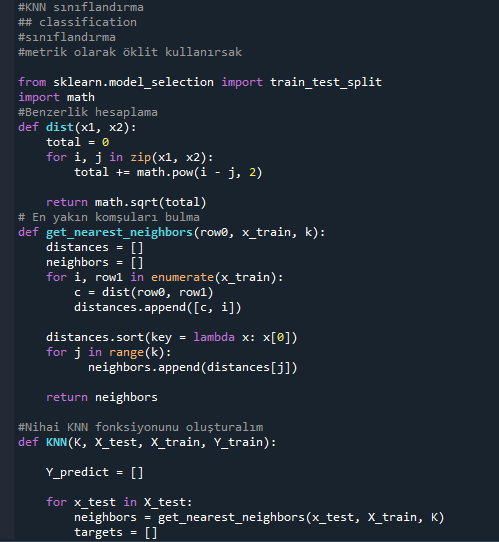
K-EYK yöntemi diğer sınıflandırma yöntemlerine göre basit ve anlaşılması daha kolay olan danışmanlı bir yöntemdir. Yöntem için tek parametre *k* parametresidir. K-EYK yöntemi sınıflandırılmak istenen bir veriyi daha önceki verilerle olan yakınlık ilişkisine göre sınıflandıran bir algoritmadır. Yeni gelen verinin eğitim verilerine uzaklıkları bulunur. En yakın *k* örnek içerisinde en çok hangi sınıfa ait veri varsa yeni gelen veri o sınıfa dahil edilir.

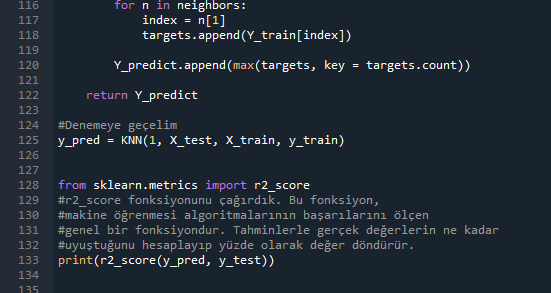
Aşağıdaki şekilde sarı ve mor dairelerden oluşan örnekler iki ayrı A ve B sınıfını ifade etsin. Merkezdeki kırmızı örnek sınıflandırılırken, *k* değerine göre çevresindeki örneklere olan uzaklıklara bakılır. Örneğin 𝑘=3 seçildiğinde en yakın 3 örneğin ikisi mor olduğundan kırmızı örnek B sınıfına dahil edilir. *k* değeri 6 seçildiğinde ise en yakın 6 örnekten 4 tanesi sarı olduğundan örnek A sınıfına dahil edilir. *k* değerinin küçük seçilmesi eğitilen modelin aykırı değerlere karşı daha duyarlı olmasına sebep olurken büyük seçilmesi de diğer sınıflardan çok sayıda örnek içermesine neden olacaktır.



Yöntem kullanılırken özellik vektörlerinin birbirlerine olan mesafelerinin hesaplanmasında çeşitli uzaklık ölçütleri kullanılmaktadır. En yaygın olanı Öklid uzaklığıdır. K-EYK yöntemi modifiye edilerek bu çalışmada kullanılan Ağırlıklı K-En Yakın Komşu (AK-EYK) yöntemi elde edilmiştir. AK-EYK yönteminde herhangi bir test örneğinin eğitim verisindeki en yakın *k* örneğe uzaklıkları bulunduktan sonra, her sınıf için ağırlıklar şu formülle hesaplanır: 𝑡𝑜𝑝𝑙𝑎𝑚𝑐= Σ1𝑢𝑧𝑎𝑘𝑙𝚤𝑘𝑖⁄𝑛𝑐𝑖=1. Burada C sınıf sayısını ifade ederken, nc de o sınıfa ait örnek sayısını ifade etmektedir. Son olarak test örneği ağırlığı en büyük olan sınıfa dahil edilir.

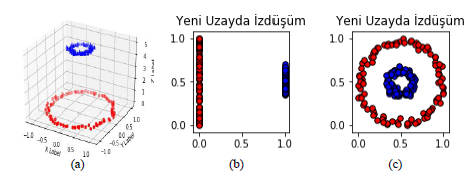






**Yerellik Koruyucu İzdüşüm**

Yerellik Koruyucu İzdüşüm (YKİ) yöntemi TBAyöntemine alternatif olarak ortaya konulan doğrusal bir boyut indirgeme yöntemidir [31]. TBA yönteminden farklı olarak, boyutu azaltılacak olan verinin komşuluk ilişkisini koruyacak şekilde izdüşüm yapan bir yöntemdir. Aşağıdaki şekilde de görüldüğü gibi 3 boyutlu ve iki sınıflı veri setine TBA ve YKİ özellik çıkartma algoritmaları uygulanarak, 2 boyutlu uzaydaki izdüşümleri gösterilmiştir. TBA verinin orjinal uzaydaki geometrik yapısını korumadan izdüşüm yaparken, YKİ yöntemi ise izdüşümü verinin geometrik yapısını koruyarak yapmaktadır.

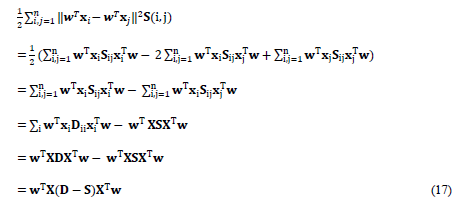


Şekil : (a) Üç boyutlu veri seti, (b) TBA izdüşümü ve (c) YKİ izdüşümü (c)

YKİ ve TBA arasındaki bir diğer fark aykırı değerlerin izdüşüm üzerindeki etkisidir. TBA varyans tabanlı bir yöntem olduğu için aykırı değerlerin varlığından güçlü şekilde etkilenmektedir. YKİ ise verinin komşuluk ilişkisini korumaya odaklandığı için aykırı değerlerin varlığından güçlü bir şekilde etkilenmemektedir. Yöntemin analizi yapılırken ilk olarak eşitlik (16) minimize edilmeye çalışılır.



Yukarıdaki denklemde 𝐲i ve 𝐲j, 𝐱i ve 𝐱j’nin yeni uzaya izdüşürülmüş biçimlerini ifade ederken 𝐒(i,j) ise verinin komşuluk ilişkisini ifade etmektedir. Komşuluk matrisi, YKİ yönteminin kullanım şekline göre farklılık gösterebilir. Etiket bilgisi kullanıldığında, 𝐱i ve 𝐱j aynı sınıfta ise 𝐒(i,j) yukarıdaki formüldeki gibi hesaplanır. Eğer aynı sınıfta değiller ise 𝐒(i,j) 0’a setlenir. Etiket bilgisi kullanılmadığında ise, 𝐱i 𝐱j’ye en yakın *k* komşudan biriyse 𝐒(i,j) yine formüldeki gibi bulunur. Aksi durumda ise 𝐒(i,j) 0’a setlenir. Minimize edilecek denklem daha da basitleştirilerek eşitlik (17) elde edilir.



Eşitlik (17)’deki **D** matrisi diagonal bir matristir ve **D**ii = Σ𝐒(i,j)j eşitliğiyle hesaplanır.

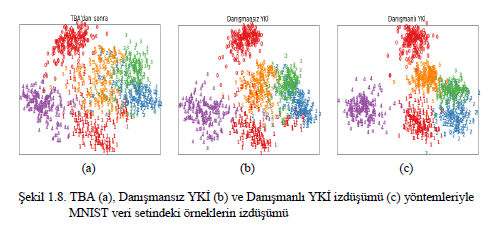
Minimizasyon problemi 𝐰T𝐗𝐃𝐗T𝐰=1 kısıt koşulu altında eşitlik (18)’deki gibi genelleştirilmiş özdeğer özvektör problemine dönüştürülür. En düşük özdeğerlere karşılık gelen özvektörler aranan sonuçlardır.

(18)

Eşitlik (18)’de, en küçük özdeğerlere karşılık düşen özvektörler, eşitlik (19)’daki en büyük özdeğerlere karşılık düşen özvektörlerle aynıdır.

(19)

YKİ yöntemi komşuluk matrisinin hesaplanma türüne göre danışmanlı ya da danışmansız olabilir. Aşağıdaki şekillerde MNIST [29] veri setine TBA, danışmansız ve danışmanlı YKİ yöntemleri ugulanarak iki boyutlu uzaydaki izdüşümleri gösterilmektedir.



Yukarıda da görüldüğü gibi YKİ yöntemi danışmansız olarak da kullanıldığında TBA’dan daha iyi sonuç vermektedir. Bunun nedeni YKİ’nin verinin komşuluk ilişkisini korumaya çalışmasından kaynaklanmaktadır.

YKİ yöntemi danışmansız olarak kullanıldığında komşuluk matrisi her veriye en yakın *k* tane örnek için denklem eşitlik (16)’daki gibi hesaplanır. En yakın *k* örnek içinde olmayanlar ise sıfıra setlenir. *k* parametresinin seçimi deneysel olarak belirlenmektedir. YKİ yöntemi danışmanlı olarak kullanıldığında ise sınıf içi geometrik yapı korunmaya çalışılır. Danışmanlı yöntemde bir örnek sadece aynı sınıftaki örneklerle eşitlik (16)’dan geçirilir. Sonuç olarak orjinal uzayda bir örnek aynı sınıftaki bir örnekle yakınsa yeni uzayda da yakın olması, farklı bir sınıftaki örneklerle ise uzak olması amaçlanır.

**Çekirdek Yerellik Koruyucu İzdüşüm**

YKİ doğrusal bir yöntemdir ve doğrusal olmayan ilişkileri temsil etmekte yetersizdir. Bu yüzden çekirdek kavramı kullanılarak yöntem doğrusal olmayan yapıya dönüştürülmüş ve Çekirdek Yerellik Koruyucu İzdüşüm (ÇYKİ) [37] adını almıştır.

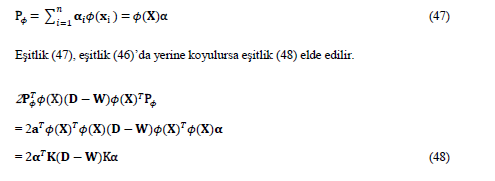
Giriş örnekleri **X** = [**x**1 ,**x**2,...,**x**n] olsun. İlk olarak doğrusal olmayan 𝜙 foksiyonuyla veri yüksek boyutlu *F* özellik uzayına haritalanır *F*: 𝜙(X) = [𝜙(x1 ),𝜙(x2),...,𝜙(xn)]. Daha sonra *F* özellik uzayında, örneklerin ağırlıklı mesafelerinin toplamını minimize ederek, 𝜙(**X**) verilerinin sınıf içi geometrik yapısını koruyabilen **P**𝜙 izdüşümü bulunur. İzdüşümü bulmak için ilk olarak eşitlik (45) minimize edilir.



Bu denklemdeki , **P**𝜙 üzerinde 𝜙(**x**i)’nin izdüşümünü ifade etmektedir. **W**(i, j) ise **x**i ve **x**j arasındaki komşuluk ilişkisini temsil etmektedir. Minimize edilecek denklem daha da basitleştirilerek eşitlik (46) elde edilir.



Eşitlik (46)’daki **D** matrisi diagonal bir matrisdir ve **D**ii = Σ𝐖(𝑖,𝑗)𝑗 eşitliğiyle hesaplanır.



Eşitlik (48)’de 𝐊(𝑖,𝑗)=𝜙(𝐱𝑖 )·𝜙(**x**j ) denklemiyle ifade edilir. Çekirdek kavramına göre, *F*’de iki vektörün noktasal çarpımı, doğrusal olmayan 𝜙 haritalamasını açıkça bilmeden bir çekirdek fonksiyonuyla *k*(x,y) = 𝜙(**x**)·𝜙(𝐲) hesaplanır.

Minimizasyon problemi kısıt koşulu altında eşitlik (49)’daki gibi genelleştirilmiş özdeğer problemine dönüştürülür. En düşük özdeğerlere karşılık gelen özvektörler aranan sonuçlardır.



Eşitlik (49)’da en küçük özdeğerlere karşılık düşen özvektörler ile eşitlik (50)’deki en büyük özdeğerlere karşılık düşen özvektörler aynıdır. Bu çalışmada özvektörler bulunurken eşitlik (50) kullanılmıştır.

TBA’da %95 kuralı, ÇTBA’da %99 kuralı, DDA ve ÇDA’da sınıf sayısının bir eksiği kuralı ile çıkış uzayının boyutu hesaplanırken YKİ ve ÇYKİ için kesin bir kural yoktur. Bu sebeple çıkış uzayının boyutu bu çalışmada deneysel çalışmalarla ile belirlenmiştir.

Aşağıdaki şekilde veri setine sırasıyla YKİ ve ÇYKİ yöntemleri uygulanıp, en küçük iki özdeğere karşılık gelen özvektörler alınarak iki boyutlu uzayda dağılımları gösterilmiştir. Şekilden de görüldüğü gibi çekirdek kavramının kullanılmasının sınıfları biribirinden ayırmada başarılı olduğu gözükmektedir.

