T.C.
YALOVA ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

VERİ MADENCİLİĞİ DERSİ

K-EN YAKIN KOMŞU İLE KÜMELEME ÖDEV RAPORU

HAZIRLAYAN

Muhammed Furkan Baltacı 17010106

NİSAN 2020

1.Veri seti hakkında,nereden alındı ? İçeriği ve amacı nedir ?

Veri seti Kaggle.com'dan alınmıştır. On üç özellik ve bir tane hedef sütunundan oluşan bu veri setinde, yaş,cinsiyet,göğüs ağrısı tipi (0,1,2,3 değerleri şiddete göre belirlenmiştir.), kan basıncı,kolestrol değeri,açlık kan şekeri (değeri 120 den büyükler için 1,küçükler için ise 0 kullanılmış.) , elektrokardiyografik sonuçları (0,1,2 değerleri kullanılmıştır.), maksimum kalp atış hızı, egzersize bağlı göğüs sıkışma değeri (0,1), depresyon, büyük damar sayısı ve tal değeri (3 normal; 6 sabit kusur; 7 tersinir kusur) özellikleri kullanılmıştır.

Veri setinin kullanılma amacı, birey üzerinde ki, kalp hastalığının tespitidir.

2. Kullanılan k-en yakın komşu algoritması nedir ve neden kullanılır ?

Veri noktaları arasındaki uzaklıkları kullanarak sınıflandırma yapan en popular yöntemlerden kabul edilir.Bu algoritma test verisini, eğitim verisindeki en yakın komşusunun sınıfı olarak tanımlar. Örnek veri setine katılacak olan yeni verinin,kullanılan verilere uzaklığı hesaplanarak n sayıda yakın komşuluğuna bakılır.Dezavantaj olarak,bütün durumları saklamasından dolayı,büyük veriler için kullanıldığında bellek alanına ihtiyaç duyar. Çalışmamızda uzaklık fonksiyonlarından Öklid, Minkowski ve Manhattan kullanılmıştır.

3.K-nn algoritması veri üzerinde nasıl kullanıldı ?

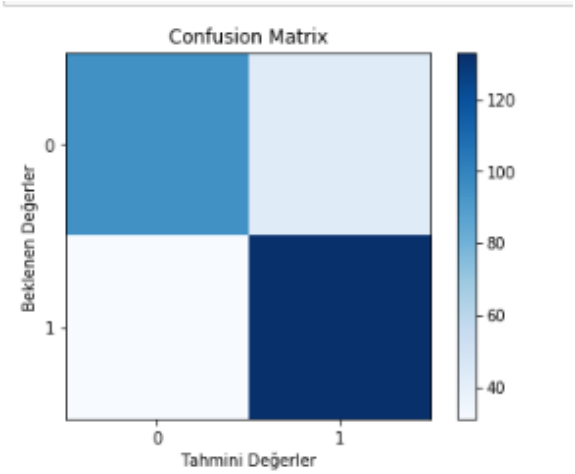
Öncelikle verileri almak için bir fonskiyon yazıldı ve veriler satır satır alınıp veriSeti isminde liste de tutulmuştur. İlk olarak uzaklıkların hesabı için fonksiyon yazıldı. Sonrasında komşuları belirlediğimiz ve komşulardan yola çıkarak tahmin ettireceğimiz fonksiyonlar yazıldı.Verileri alırken ise 'str' hatası ile karşılaşıldığı için,string değer tiplerinin integer ve float değerlerine dönüşümü için bir fonksiyon yazılmıştır.

Kodların çalışmasına gelirsek, k parametresi yani komşu sayımızısı(biz bütün mesafelerde 5 belirledik) belirledik, yeni verinin kullanılan verilere göre uzaklıkları, seçilen uzaklık fonksiyonlarımızdan biri ile tek tek hesaplanır. (komsuBelirlemeOklid fonksiyonu içerisinde) Bu hesaplamalar mesafeler adında bir liste objesinde tutulup,küçükten büyüğe doğru sıralanır ve en yakın belirlenen k tane komşu seçilir. Tahmin kısmına

gelince, karşılaştırma yapılarak en yüksek değerli sınıf belirlenip, yeni veri o sınıfa dahil edilir.

Analiz ve sonuç

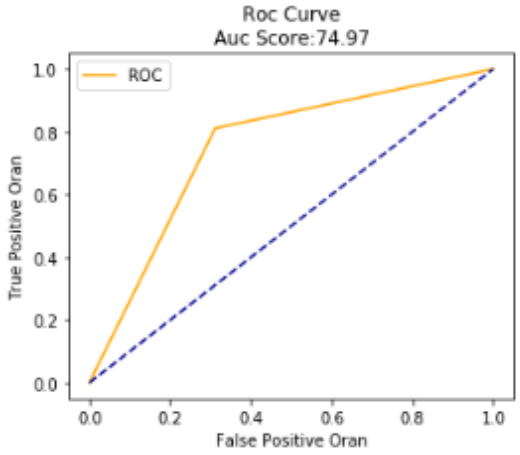
Çalışmamın analizlerini görselleştirmek için her bir uzaklık fonksiyonu sonucunda oluşan değerleri roc eğrisi ve hata(confusion) matrisi kullandım. Uzaklık fonskiyonlarından birinin sonucunu örnek göstererek raporumda şu şekilde anlatmak istiyorum.



Hata matrislerinde doğruya doğru (doğru), doğruya yanlış (yanlış), yanlışa doğru (yanlış), ve yanlışa yanlış (doğru), değerleri vardır. Çalışmamda görülüyorki %70'in üzerinde doğru değerler alınarak

0 ve 1 beklenen değerler çoğunluk olarak doğru belirlenmiştir. Ancak %20 nin üzerinde de hatanın olduğu görülmekte buda verinin azlığından ve ya başka model eksikliklerinden kaynaklanıyor olabilir.

Roc eğrisine gelirsek,



%74.97'lik bir score elde edilmiştir. Turuncu ile gösterilen kısım 1'e ne kadar yakın olursa doğruluk değeri o kadar yüksek olur.

Öklid mesafesinde;

0 tahmini bulma oranı %75
1 tahmini bulma oranı %77
Auc score ise 75.81 olarak bulunmuştur.

Minkowski mesafesinde;

0 tahmini bulma oranı %75
1 tahmini bulma oranı %76
Auc score ise 74.97 olarak bulunmuştur.

Manhattan mesafesinde;

0 tahmini bulma oranı %81
1 tahmini bulma oranı %80
Auc score ise 80.36 olarak bulunmuştur.

Doğruluklar sırasıyla;
Manhattan,Öklid ve Minkowski
Olarak belirlenmiştir.