**T.C.**

**HALİÇ ÜNİVERSİTESİ**

**MÜHENDİSLİK FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**K- En Yakın Algoritması**

**BİM458 VERİ MADENCİLİĞİ PROJESİ**

**Hazırlayan**

**Kürşad KARAGÖZ**

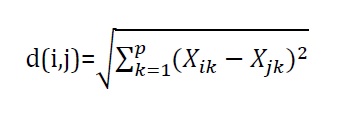
**Karagoz29kursat@gmail.com**

**İstanbul - 2018**

**EN YAKIN K KOMŞU ALGORİTMASI İLE SINIFLANDIRMA**

Bu yöntem sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak , örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır.

Söz konusu bu yöntem örnek kümesindeki gözlemlerin herbirinin , sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır.Uzaklıkların hesaplanmasında , i ve j noktaları için genellikle kullanılan uzaklık çeşiti olarak öklid uzaklığı kullanılabilir.



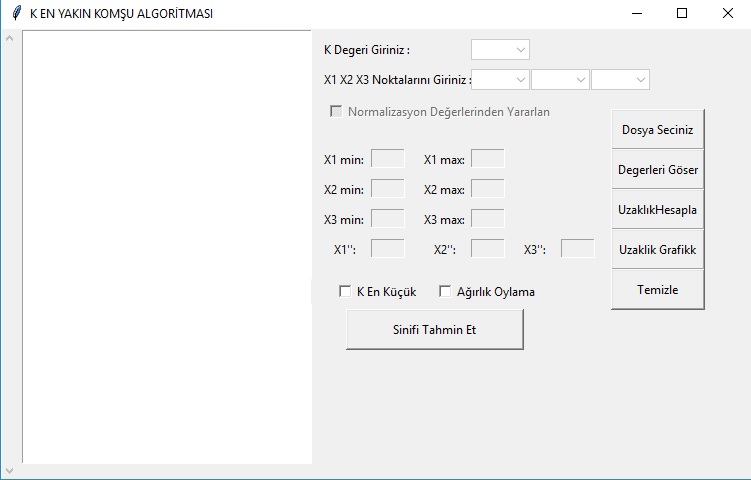
***En Yakın k-Komşu Algoritması Adımları:***

1. K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır.
2. Bu algoritma verilen bir noktaya en yakın komşuları belirleyeceği için söz konusu nokta ile diğer tüm noktalar arası uzaklıklar tek tek hesaplanır
3. Yukarıda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir.
4. Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrarlanan kategori değeri seçilir.
5. Seçilen kategori tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir.

**~~DİKKAT~~**:

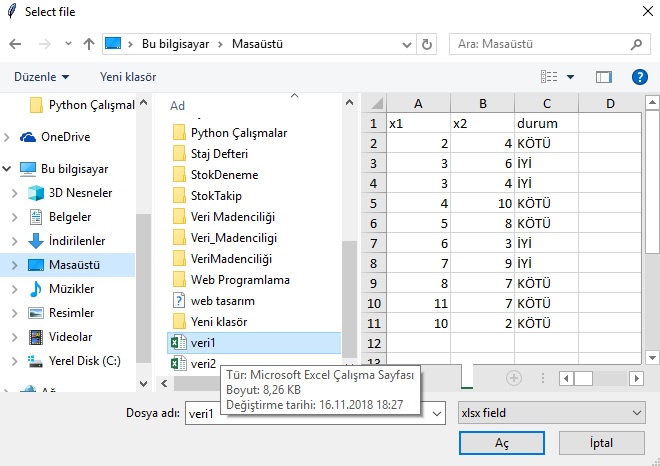
K en yakın komşu algoritması veri madenciliğinden kullanılabilecek algoritmalardan yalnızca bir tanesidir hangi algoritma kullanılacak olursa olsun öncelikle veri setine normalizasyon ya da Z-score dönüşümü yapılmalı ve elde edilen değerler sonucunda bu algoritmalar uygulanmalıdır.

**EN YAKIN K- KOMŞU ALGORİTMASI ÖRNEK**



Yukarıda görüldüğü üzere programı çalıştırmak için öncelikle programa veri seti “DosyaSeciniz” butonuna gömülü select file widgetinden dahil edilmelidir. Aksi taktirde herhangi bir işlem yapılamaz.

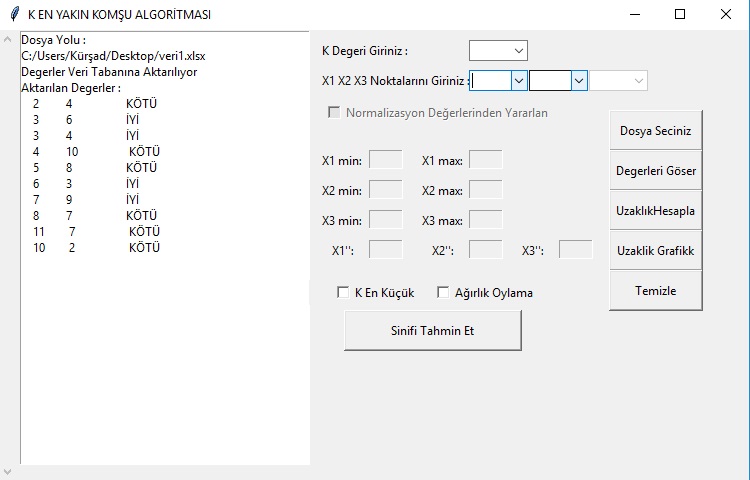
Bu programda veri seti için excele gömülü verilere kullanılmıştır.



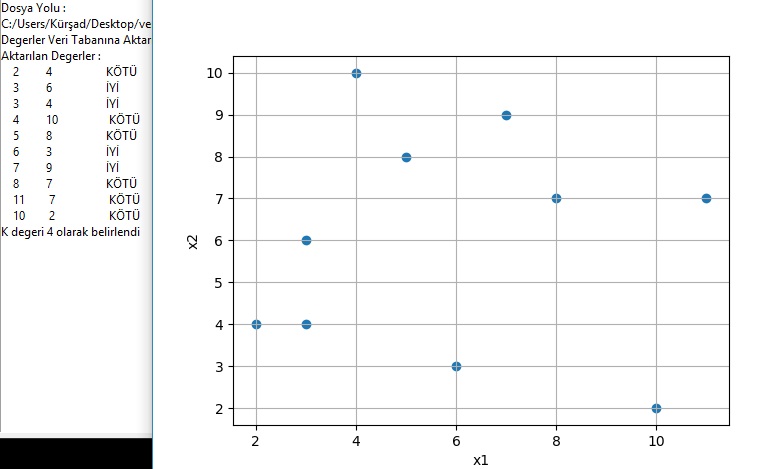
Örnekte görüldüğü gibi x1, x2 sütunları (nitelikler) altında değerler ve durum sütunu altında ise bu değerlere karşılık durumlar (sınıf değerleri) yer almaktadır.

Bu programın çalışabilmesi için excelde nitelikler ve sınıfın “durum” sütunu altında olması zorundadır çünkü okuma işlemi için “durum” referans alınmaktadır.

“Veri1” excel tablosu seçilerek programa veri seti dahil edilir.

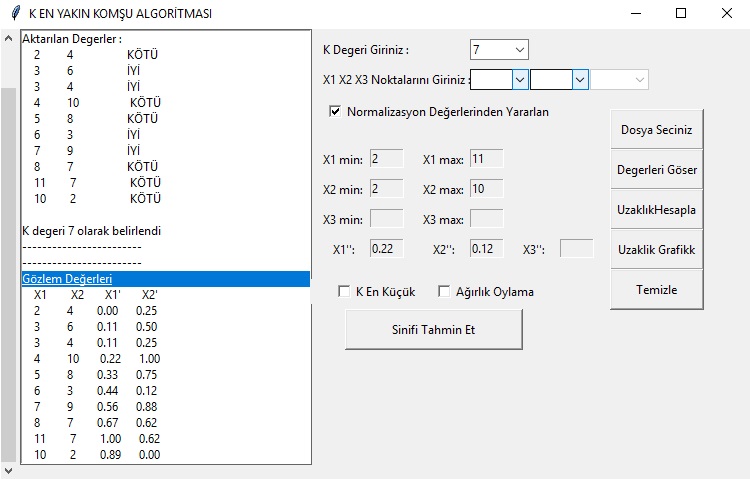


Tablo seçildikten sonra sol tarafta yer alan Listbox içerisine değerler aktarılmaktadır , Aynı şekilde bu değerler MYSQL veritabanında’da tutulmaktadır.



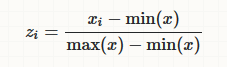
İstenildiği taktirde “DeğerleriGöster” butonuna basılarak x1 ve x2 değerleri görülebilir.

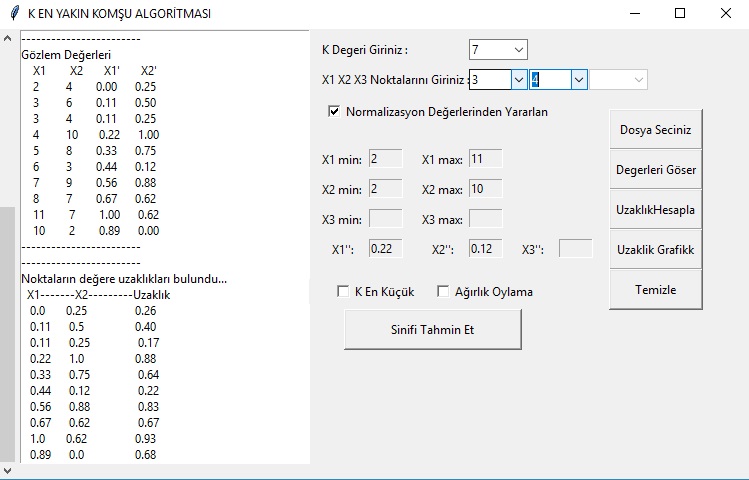
* **Veri seti programa dahil edildiğine göre ,**  adımlara uyarak algoritmaya başlanabilir. Bunun için öncelik K değeri belirlenir.



* Daha sonra eğer x1 veya x2 üzerine normalizasyon yöntemi uygulanacak ise “Normalizasyon değerinden yararlan” adlı checkbox’a tıklanarak normalizasyon sonucu değerler elde edilir.

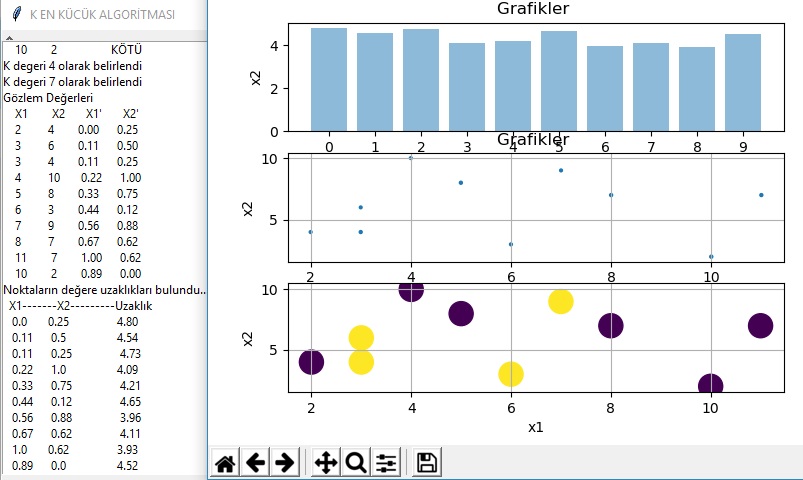
Normalizasyon için kullanılan formül:





* Daha sonra her bir verinin öklid uzaklığını hesaplamak için gözlem değerleri girilir ve ardından “Uzaklık Hesapla” butonuna tıklanır.

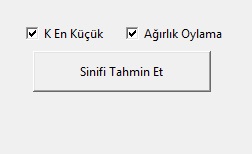
Soldaki listboxda görüldüğü üzere sayfanın başında verdiğimiz formülden yararlanılarak tüm uzaklıklar bulunmuştur.



İstenildiği taktirde uzaklıklara ait grafikler görülebilir.

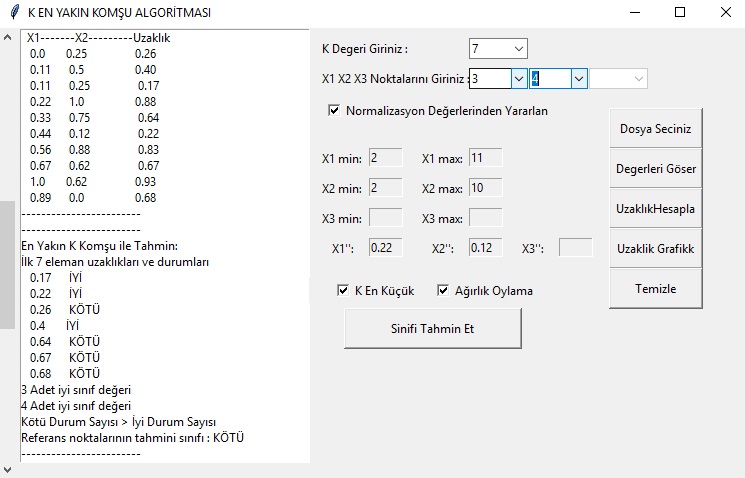
1. Grafikte uzaklık değerleri kolonlar halinde görülmektedir.
2. Grafikte her x1…. Değerlerinin uzaklıklarına göre ait nesnesinde doğru orantılı bir büyüme gösteren anlatım görülmektedir.
3. Grafikte ise x1…Değerlerinin kordinatları durumlarına göre bu örnekte durumlar : iyi veya kötü olmak üzere gruplara ayrılmış halde gösterimi söz konusudur

Morlar: Kötü , Sarılar : İyi



* Son adım olarak gözlem değerlerinin hangi sınıfa ait olduğunu belirlemek amacıyla iki tahmin yönteminden birini veya ikisini bir kullanma işlemine geldik.

Öncelik olarak aşağıda k en yakın algoritması ile sınıf tahminin sonucuna bakalım



Yukarıda görüldüğü üzere ilk k adet (7) veri küçükten büyüğe sıralanarak durumları gösterilmiştir ve bu durumları kıyasladığımızda kötü sınıf değerinin iyi sınıfından daha fazla olduğu belirlenmiştir ve sınıfa yeni katılacak gözlem değerinin tahmini sınıfı “KÖTÜ” olarak belirlenmiştir.

**Ağırlık Oylama ile Sınıflandırma:**

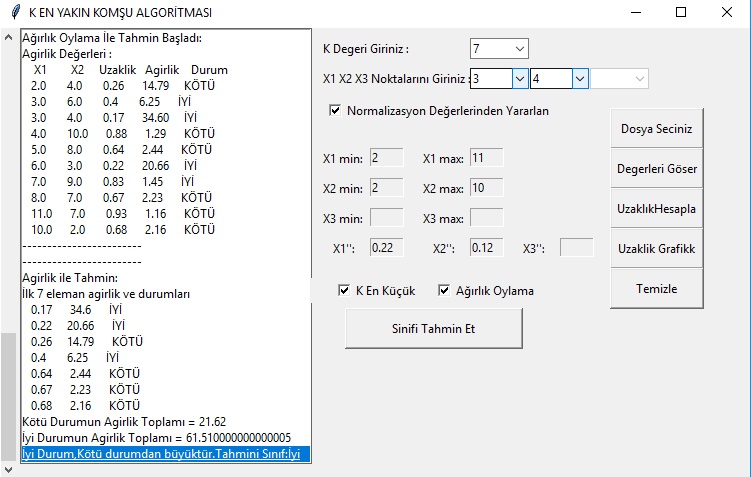
K –En Yakın Komşu algoritması, yukarıda gösterilenlerde söz edildiği üzere verilen bir gözleme en yakın k komşunun belirlenmesi ve sınıfı bilinmeyen yeni bir gözlem değeri için , bu k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanıyordu. Ancak seçilen busınıf , sadece komşu sayısının göz önüne alınması nedeniyle herzaman uygun olmayabilir.

K komşu arasında en çok tekrarlanan sınıfı seçme yöntemi yerine komşuların uzaklıklarını da göz önüne alan ağırlıklı oylama yöntemi geliştirilmiştir.

Bunun ilçin kullanılan formül:

**d(i,j) ‘=1/d(i,j)2**

Ağırlık oylama ile sınıf tahmini sonucuna bakacak olursak:



Ağırlık değerlerini uzaklık değerinden yararlanılarak her birisi için bulunmuştur.

Daha sonra k adet ağırlık değerinin küçükten büyüğe sıralanarak durum değerlerine göre grup ayrılması ve toplamlarının kıyaslanması sonucunda büyük gelenin sınıfının yeni gözlem değerlerinin sınıfı ilan edilmesi söz konusudur.

**Veri Tabanı Tasarımı**

****