

# BİL 3013 Veri Madenciliğine Giriş – Final Sınavı

Doç. Dr. Mete Eminağaoğlu

Bu sınav, iki bölümden oluşmaktadır. İlk bölümde, 30 puanlık bir adet soru sorulmuştur. İkinci bölümde ise 70 puanlık bir proje verilmiştir.

**Proje ve soruların yanıtları için son teslim tarihi: 17 Ocak 2022 Pazartesi, 20:00**

(ek süre kesinlikle verilmeyecektir. Herhangi bir nedenle zamanında iletilmeyen ödevler, hiçbir mazeret kabul edilmeden 0 (sıfır) olarak notlandırılacaktır.)

## Sınavın Teslim Şekli:

DEÜ Sakai sistemindeki ders sayfasında açılacak olan ödev yükleme (assignment) alanına; tüm dosyalar, rapor, vb. zip / rar sıkıştırılmış tek bir dosya olarak yüklenecektir.

Bu sınav tek kişiliktir. En ufak bir yardım, internet'ten kod kopyalanması, vb. eylemler kopya / intihal olarak değerlendirilecektir ve bunu yapan öğrenciler sınavdan ya da sınavın ilgili bölümünden 0 (sıfır) alacaktır.

## Final – 1. Bölüm (30 puan):

Take-home şeklindeki ara sınavın ilk bölümünde, aşağıda bir adet soru verilmiştir. Bu sorunun çözümü net, ayrıntılı ve düzgün bir biçimde, bir Word belgesi içerisine yazılacaktır (Eğer, öğrenciler çözümleri el yazısıyla kâğıda çözerse, scanner ile taranıp ya da akıllı telefonla fotoğrafı çekip, vb. Word içinde resim olarak kaydedip de iletebilir).

Aşağıda verilen veri seti için, **hiyerarşik aglomeratif kümeleme (hierarchical agglomerative clustering)** algoritması ile aşağıda verilen iki farklı yöntemle kümeleme işlemlerini yapacaksınız. **Her aşamada hangi kayıtların hangi kümeye atandığını da yazıp en sonunda da küme ağacını (dendrogram) çizeceksiniz.**

**DİKKAT - HATIRLATMA:** Öncelikle, bu veri setindeki değişkenlerde gerekli bir veri ön işleme / dönüştürme yapmanız da gerekmektedir.

**1.a. (15 puan) Single-Linkage** yöntemi ile hiyerarşik aglomeratif kümeleme

**1.b. (15 puan) Complete-Linkage** yöntemi ile hiyerarşik aglomeratif kümeleme

Kayıt Numarası	Miktar	Adet	Katsayı
1	2000	50	1.5
2	3000	12	3
3	4000	30	2.8
4	1800	100	4.3
5	500	180	3.2
6	1500	45	5
7	2500	10	4
8	900	64	2

## Final – 2. Bölüm (70 puan):

Bu final sınavının olduğu “Ödevler” kısmında yer alan **veri kümesini (“Final-data.txt”)** kullanılarak, **k-means** algoritması ile kümeleme (clustering) analizi yapan bir program geliştireceksiniz. **Ekrandan girilen “k” değerine göre**, k adet kümeye ayıracak ve veri dosyasındaki her bir kaydı ilgili kümeye hesaplayarak atayacaktır.

Programınız her bir kaydın hangi küme (cluster) ‘a atandığı ve genel sonucu programınız tarafından oluşturulacak bir **“sonuc.txt” dosyasına kaydedecektir. Ayrıca, bu dosyanın sonunda WCSS (Within Cluster Sum of Squares), BCSS (Between Cluster Sum of Squares) ve Dunn Index değerleri de hesaplanıp yazılacaktır.**

Örnek bir "sonuc.txt" dosyası içeriği (ekrandan k = 3 girildiği varsayılarak) aşağıda verilmiştir (sayılar gerçek değildir, fikir vermesi için yazılmıştır).

```
-----
Kayıt 1:      Küme 1
Kayıt 2:      Küme 1
Kayıt 3:      Küme 2
...
```

```
Kayıt 221:    Küme 3
Kayıt 222:    ...
...
...
Kayıt 249:    Küme 2
```

```
Küme 1: 208 kayıt
Küme 2:  35 kayıt
Küme 3:   6 kayıt
```

```
WCSS: 16.24
BCSS: 286.45
Dunn Index: 10.73
```

Programınızda **"veri görselleştirme"** kısmı da olacaktır. **"veri görselleştirme"** adlı bir seçenek / tuşa tıklandığında, k-means ile elde edilen kümelerin ekrana çizilebilmesi için x eksenini ve y eksenine hangi değişkenlerin geleceği seçilecek (veri setindeki "Sports, Religious, Nature, Theatre, Shopping, Picnic" değişkenlerinden ikisi), sonra da buna göre ekranda tüm kayıtlar ait olduğu küme farklı görsel (farklı renk veya şekil) olacak şekilde grafik çizimi yapılacaktır.

- (Her ne kadar önerilmese de), hazır kütüphane, hazır araç, vb. öğrenciler eğer istiyorsa kullanabilir.
- Kodlama kısmında, sadece, aşağıdaki programlama dillerinden birini kullanabilirsiniz, aşağıdakiler harici bir dil kullanılamaz:

**C, C++, C#, .Net, Java, Python.**

#### **Projede Teslim Edilecekler:**

- 1-Programın tüm kaynak kodları, bağlantılı kütüphane, dizinler, vb.
- 2-Kullanılan algoritmalar, vb. ile ilgili kısa bilgiler / notlar (istenirse kaynak kod içine de açıklamalar olarak eklenebilir).
- 3- sonuc.txt dosyası