BİL3003 Veri Madenciliği'ne Giriş – 2. Ödev

Dersin Öğr. Elemanı: Dr. Öğr. Üyesi Mete Eminağaoğlu

Ödev Konusu: Öğrenciler, aşağıda **kısaca açıklanan ödevi** kodlayacak ve raporla birlikte kaynak kodları teslim edecektir. Aşağıda özet bilgiler verilmiştir. Detaylı açıklamalar, sadece bir kere, ders saatinde yapılacaktır. O derse gelemeyen, vb. öğrencilerin ödevi yanlış / eksik vb. anladığı için yetersiz bir ödev iletmesinden sadece o öğrencinin kendisi sorumludur.

Ödevin Son Teslim Tarihi: 4 Ocak 2019, 23:00 (TSİ)

Ödevin Teslim Şekli:
CSC ÖBS (Moodle) sistemindeki ders sayfasında açılacak olan ödev yükleme (assignment) alanına; tüm dosyalar vb. **zip / rar sıkıştırılmış tek bir dosya olarak yüklenecektir**.

Bu ödev **tek** kişiliktir. **Her türlü yardım, vb. kopya / intihal olarak** değerlendirilecektir ve **yardım eden / yardım alan** öğrencilerin hepsi bu ödevden **0 (sıfır) alacaktır.**

Ödevin geç teslimi durumları:
Her ne nedenle olursa olsun, geç teslim edilen ödevler iletilmemiş olarak değerlendirilecek ve **0 olarak notlanacaktır.**

Ödev Konusu:
Bu dersin Moodle sistemindeki Ödev-2 kısmında size iletilmiş olan .csv uzantılı veri kullanılarak DBScan algoritması ile kümeleme (clustering) analizi yapılacaktır.

- **Veri Seti ile ilgili bilgiler:**
  - Ödev-2 kısmında ayrı bir .txt dosyasında verilmiştir.
  - Her değişken , (virgül işareti) ile ayrılmıştır.
- **DBScan, clustering algoritmalarına ilişkin hazır fonksiyon, kütüphane, hazır araç, vb. kullanımı kesinlikle yasaktır. Öte yandan veri görselleştirme, grafik çizim, vb için hazır araç, kütüphane, vb. kullanabilirsiniz.**
- Kodlama kısmında, **sadece aşağıdaki programlama** dillerinden istediğiniz birini seçip kullanabilirsiniz:
  C, C++, C#, .Net, Java, Python.
- **DİKKAT: Java programlama dilinde kodlayacak öğrenciler Eclipse üzerinde çalıştırmalıdır. Python programlama dilini kullanacaklar da Python 2.7 veya daha üst bir sürümü (version) kullanmak zorundadır.**

Veri setinde, öncelikle bazı çeşitli ve farklı veri düzenleme / düzeltme / temizleme (data pre-processing) işlemleri yapılmayacaktır.

Program, dosyayı okuduktan sonra, ekrandan **Epsilon** değeri ve **Minimum nokta sayısı (minimum number of points)** girilen değerlere göre, DBScan algoritması ile kayıtları kümeleyecek, ve her bir kaydın hangi küme (cluster) 'a atandığı ve varsa outlier (kümeye atanamayanları) ve genel sonucu programınız tarafından oluşturulacak bir **"sonuc.txt" dosyasına kaydedecektir.**

Örnek bir sonuc.txt dosyası içeriği aşağıda verilmiştir.

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

Kayıt 221:          Küme 1
Kayıt 222:          ?
...
...
```

```
Küme 1: 208 kayıt
Küme 2: 27 kayıt
Küme 3:  8 kayıt
Kümeye atanmayan (sapan değer): 6 kayıt
-----------------------------------------------
```

Programınız DBScan'i çalıştırmayı bitirip "sonuc.txt" dosyasını oluşturduktan sonra ekranda "dosya oluşturuldu" bilgisi çıkacaktır.

Ayrıca,**"veri görselleştirme"** adlı bir seçenek / tuşa tıklandığında, DBScan ile elde edilen kümelerin ekrana çizilebilmesi için x ekseni ve y eksenine hangi değişkenlerin geleceği seçilecek, sonra da buna göre ekranda tüm kayıtlar ait olduğu küme farklı görsel (farklı renk veya şekil) olacak şekilde grafik çizimi yapılacaktır. Hiç bir kümeye atanamayan sapan (outlier) kayıtlar da ayrıca farklı bir şekille aynı grafik üzerinde gösterilecektir. Ayrıca, grafik üzerinde herhangi bir kayda ait noktaya tıklandığında ekrana o kaydın kaçıncı kayıt olduğu ve hangi kümeye ait olduğu bilgisini ekrana yazacaktır.

Ödevde Teslim Edilecekler:

1-Programın tüm kaynak kodları, bağlantılı kütüphane, dizinler, vb.

2-Kullanılan yöntemler, işlemler, vb. ile ilgili kısa bilgiler / notlar (kaynak kod içine kısa açıklamalar olarak eklenmelidir).