1 BIL 366 Data Mining: Homework-2

1.0.1 Bilinmesi gerekenler

distance measures, preprocessing (aggregation, cleaning, sampling), pandas,numpy, matplotlib.pyplot ### Öğrenme Hedefleri Pandas kullanılarak distance measureler kullanılarak data benzerliğinin ölçülmesi ve outlier detection, sampling ile sample size'ın küçültülmesi ve ne kadar sample size'ın datayı ne kadar temsil ettiğinin gözlemlenmesi.

1.0.2 Giriş

Covid süresince Türkiye location datasından Google'in elde etmiş olduğu mobility datası https://www.google.com/covid19/mobility/ sitesi üzerinden yayınlanmıştır. Bu ödevde sizden bu datayla ilgili önanaliz yapıp raporlamanız istenmektedir. Datayı indirme ve yüklemede bir bütünlük olması açısından bu datayı drive üzerinden erişime açtım, aşağıdaki şekilde erişebilirsiniz.

[ ]:
```
import pandas as pd
from scipy.spatial import distance

#https://www.google.com/covid19/mobility/
url = 'https://drive.google.com/file/d/18gyHbx6rfogq3yQ-GR9COjcGgyYlCnBZ/view?usp=sharing'
url2020 = 'https://drive.google.com/uc?id=' + url.split('/')[-2]
url = 'https://drive.google.com/file/d/1Eg8Lffm49bc-bGFkv_4ddrQw8U8WE6P4/view?usp=sharing'
url2021 = 'https://drive.google.com/uc?id=' + url.split('/')[-2]

df20 = pd.read_csv(url2020)
df20.info()

df21 = pd.read_csv(url2021)
df21.info()
```

1.0.3 Yapılacaklar

1. Yukarıdaki data framede içerisinde data olmayan (tüm sütun null), sütunları çıkarınız.
2. Mahalanobis distance distributiondan uzaklığı ölçtüğü için outlier belirlenmesinde kullanılabilir.2020 ve 2021 her iki datayıda aylara göre gruplandırdıktan(mean kullanabilirsiniz) sonra (aggregation) her bir satır ile data (tüm sütun) arasındaki Mahalanobis distance'ı hesaplayarak yeni bir sütun olarak ekleyiniz ve buradaki en büyük elemanın outlier olduğunu 2020 ve 2021 yılları için ayrı ayrı gösteriniz.

[ ]:
```
#Distance ölçümü için
#covariance matrisini numpy.cov() fonksiyonu,
#inverse (tersini) numpy.linalg.inv() fonksiyonu
#ve aşağıdaki scipy fonksiyonunu kullanabilirsiniz:
#https://docs.scipy.org/doc/scipy/reference/spatial.distance.html
import numpy as np #np.cov(), np.linalg.inv()
from scipy.spatial import distance
```

3. 2020 ve 2021 datalarını aylara göre grupladıktan sonra (mean değerleri ile) en az iki adet fark/benzerlik ölçümü (**slaylatlardaki similarity measures**) kullanarak 2020 ve 2021 datalarının 9-14 sütun verilerinin **aynı aylarda** birbirlerine ne oranda benzediğini bulunuz.
4. 2020 datasından (50-1000) aralığında farklı büyüklüklerde samplelar oluşturarak aylık mean değerlerin ortalama ne kadar değiştiğini grafikle gösteriniz (x sample size, y ortalama değişim):
   Açıklama→ Tüm datanın aylık ortalama değerleri ile sample datanın aylık ortalama değerleri arasındaki farkların mutlak değerlerini toplayarak ortalamasını almanız gerekiyor. Bu şekilde her bir sample için bir hata datası elde etmiş oluyorsunuz. Sonra bunları x axisde sample size, y axisde hata olacak şekilde grafiklemeniz istenmektedir.
5. 2020 datasından her aydan (50-1000) aralığında olacak şekilde samplelar oluşturarak aylık mean değerlerin ortalama ne kadar değiştiğini grafikle gösteriniz (x sample size, y ortalama değişim):
   Açıklama→ Bu soruda her bir aydan eşit miktarda sampller alarak sample oluşturmanız (mesela 50 için her bir aydan 50şer satır alarak, aysayısı x 50 büyüklüğünde bir sample elde etmiş oluyorsunuz) ve 4.sorudaki gibi ortalama hatayı bularak yine sample size'a göre grafiklemeniz istenmektedir.

1.0.4 Teslim

Her sorunun altında **hem kodu ve hemde çıktısını** içeren Jupiter notebook dökümanını **pdfe** çevirerek classroom üzerinden teslim ediniz(colab de direk print ile pdf alabilirsiniz). Çıktının güzel görünmesi için latex çıktı alarak oradan düzenleyebilirsiniz. Yada html aldıktan sonra print yapabilirsiniz.

1.0.5 Değerlendirme

Her soru 20 puandır.