CSE 555-Final Projesi

Son Gönderim Tarihi: 22.05.2024 Saat 23:59

Kullanılacak VeriSeti: Dry Bean Dataset

(https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset)

Bu veri setinde 7 farklı fasülye sınıfına (Şeker, Barbunya, Bombay, Çalı, Dermosan, Horoz ve Sira) ait veriler bulunmaktadır. Her kayıtta bir örneğe dair 16 öznitelik bulunmaktadır.

İstenilen: Bu veri setini kullanarak aşağıda belirtilen maddeleri python ortamında gerçeklemeniz ve her madde sonucunda elde ettiğiniz çıktıları yorumlamanız beklenmektedir. **Proje neticesinde hazırlayacağınız raporda akademik yazım kurallarına uymanız beklenmektedir (format puanlanacaktır)**. Proje sonucunda raporu ve projeyi hazırlayacağınız Jupiter Notebook çalışmanızı paylaşmanız beklenmektedir.

1- Her özniteliğin dağılımını boxplot ile çizerek yorumlayınız.
2- Verileri z-score ile normalize edip, her özniteliğin Fisher Uzaklığını (sınıflar arası) hesaplayarak analiz ediniz. Veri setinde 7 farklı sınıf olduğundan bu Fisher uzaklığını 7 farklı sınıfı dikkate alarak nasıl adapte ettiğinizi yazınız.
3- PCA ile veri setini dönüştürüp her Eigen vektör üzerindeki izdüşümünü bir öznitelik olarak değerlendirin ve her özniteliğin Fisher Uzaklığını hesaplayın ve 2. maddede elde edilen Fisher Uzaklıkları ile kıyaslayınız. Eigen değerleri (Eigen Value) ile izdüşümü alınarak edilen özniteliklerin Fisher mesafeleri arasında bir ilişki var mıdır? Yorumlayınız.
4- Verinin en önemli iki Eigen vektör üzerindeki izdüşümünü scatter plot ile çizerek yorumlayınız. Scatter plot işleminde her sınıfa ait örneği farklı renkte gösterin. Aynı işlemi en önemsiz 2 Eigen vektör için tekrarlayınız.
5- Veri Setini LDA algoritması ile 2 boyuta indirgeyiniz ve scatter diagramında gösteriniz. Bu gösterimde de her sınıfı farklı renkte gösteriniz ve bulduğunuz diagramı 4. maddede elde ettiğiniz diagram ile kıyaslanız. PCA ile elde edilen ve LDA ile elde edilen diagramın sınıflar arası ayrımsallığı ortaya çıkarmadaki performansını değerlendirin. (Bunu bir metrik ile hesaplamanız bonus olarak değerlendirilecektir)
6- Veri setini K-Means, DBSCAN, t-SNE ve SOM algoritması ile gruplayınız ve sonuçları gösteriniz. K-Means ve DBSCAN algoritmasında oriinal veri yerine PCA algoritması ile elde ettiğiniz ilk 2 bileşendeki izdüşümlerin kullanabilirsiniz. t-SNE SOM algoritmasında ise orijinal veri setini kullanınız. Bu veri görselleştirme işlemlerinde her sınıfa ait örneği farklı renkte gösteriniz.
7- Belirlemiş olduğunuz bir outlier tespit yöntemini uygulayarak, veri temizliği yapınız ve sonuçları yorumlayınız.

NOT: Projede alternatif olarak analiz etmek istediğiniz çok sınıflı (>2) ve çok öznitelikli (>10) bir veriseti de kullabailirsiniz. Kendi veri setinizi kullanmanız durumunda da yukarıda istenilen tüm maddeleri gerçeklemeniz gerekecektir. Kendi veri setinizi kullanmanız durumunda veri setinizi detaylıca anlatan bir bölüm de eklemeyi unutmayınız.