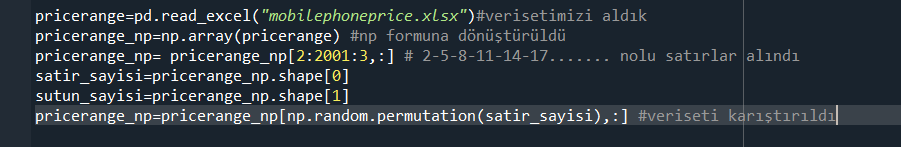
**VERİ MADENCİLİĞİ UYGULAMASI RAPORU**

**-Samed Baskın**

**-www.linkedin.com/in/sbsamed/**

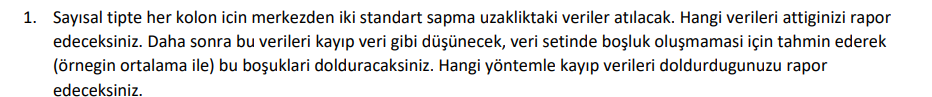
**Kullanılan Veri Seti Hakkında**

‘mobilephoneprice.xlsx’ isimli veri setinde 2,5,8,11,14…. numaralı satırlar ile çalıştım .Aşağıdaki kodlar ile veri setimi oluşturdum.



**Görevler:**

1. **STANDART SAPMA VE ARİTMETİK ORTALAMAYA GÖRE VERİ DEĞİŞTİRME**



Burada ilk başta sayısal kolonları içeren bir veri seti oluşturmamız gerekiyor. Bana verilen “mobilephoneprice.xlsx“ isimli veri seti pythonda okunduktan sonra “np.delete()” ile kategorik tipte olan kolonlar silindi.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Daha sonra bu veri setinde her kolona özgü standart sapma ve aritmetik ortalama değerleri hesaplandı. Bu değerler kolon indis numaralarına göre dizilerde tutuldu. Mesela 2.kolona özgü aritmetik ortalama değeri “aritmetikortalama[]” dizisinin 2.indisinde yer alıyor. 5.kolona özgü standart sapma ise “s\_sapma[5]” de yer alıyor. Bence burada dikkat edilmesi gereken ise sınıf bilgilerine ait kolonların da aritmetik ortalama veya standart sapmasının hesaplanıp bu kolona ait satırlarında işleme tutulmamasıdır. Sınıf bilgilerinin olduğu kolona hiçbir işlem yapmamız gerekiyor.

Bu hesaplamalardan sonra ise her kolona özgü satırlardaki değerler belirli bir işleme göre kontrol edildi. Eğer burada ilgili değer o kolona ait aritmetik ortalama değerinden en az 2\*standart sapma kadar fazla ise bu değer o kolona ait **aritmetik ortalama ile değiştirildi**. Yine aynı şekilde ilgili değer o kolona ait aritmetik ortalama değerinden en az 2\*standart sapma kadar az ise yine bu değer **aritmetik ortalama ile değiştirildi**

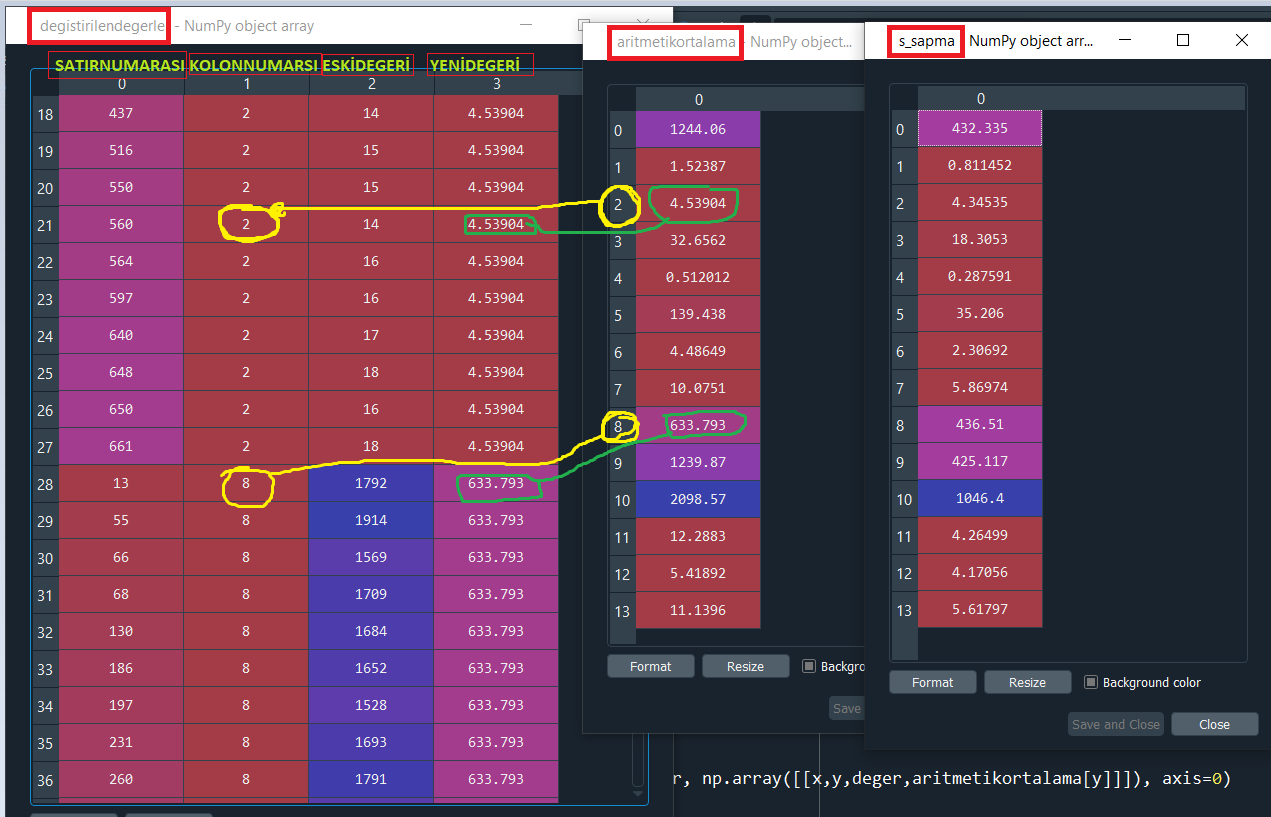
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Bu işlemler yapılırken değiştirilen değerler, bu değerlerin yeni değerleri ve değerlere ait satır ve sütun numaralarını ise iki boyutlu ayrı bir dizide tutuldu.**

🡪 degistirilendegerler [satırnumarası,kolonnumarası,eskidegeri,yenidegeri]

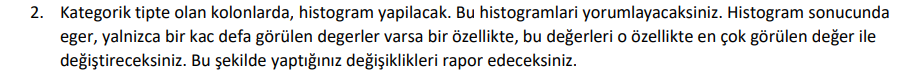
Aşağıdaki şekilde güncelleme yapılan değerlere ait bazı örnek raporlar bulunmaktadır.



Yukarıdaki görseli incelersek mesela degistirilendegerler[21,:]. satıra bakalım veri setimizdeki 2.kolona ait 560 numaralı satırdaki 14 degeri ilgili kolonun işlemlerine göre 4,53904+2\*434535=13,22974’dan büyük olduğu için bu deger o kolonun aritmetik ortalaması olan 4,53904 ile değiştirildi.

* Toplam 89 değer değiştirildi değiştirilen değerler sadece 2,8 ve 12. kolonlarda yer alıyor.
* Bu kolonlardaki özellikler sırasıyla; clock\_speed\_mikrioişlemci hizi, m\_width\_telefonun genişliği (mm cinsinden), ram\_megabyte

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**2-HİSTOGRAM ÇİZİLMESİ**

Burada ilk başta kategorik tipte olan özellikler belirlendi. Bunlar benim veri setimde bulunan 1,3,5,17,18 numaralı kolonlardı. Ancak benim veri setimdeki kategorik tipteki verilerimin tamamı 0 ve 1 (var-yok)‘lerden oluşuyor ve iki değer alıyor. Aldıkları bu değerlerin kolonlara özgü toplam adetleri ise neredeyse yarı yarıya o yüzden histogram çizdirdiğim zaman birkaç defa görülen bir değer yok. Örnek olarak indis numarası 1 olan bluetooth sütununu ele alalım ve histogramı çizdirelim.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Burada sinifözellikiliskisayisi[0,4]=bluetooth(0) dan kaç adet olduğunu [1,4] ise bluetooth(1) adedini belirtiyor. Bu matrisin tam olarak ne anlama geldiğini bir sonraki bölümde detaylıca anlattım.

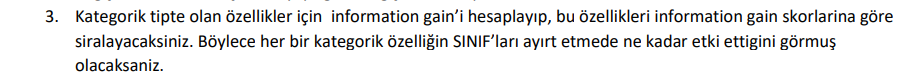
metin, ekran, ekran görüntüsü, vitrin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Görüldüğü gibi burada bluetooth(0) ve bluetooth(1)’in sayıları birbirine çok yakın. Diğer kategorik tipteki kolonlar içinde buna benzer görüntüler oluşuyor. Bu yüzden bu bölümde kategorik tipteki özelliklerde herhangi bir veri atması/değişimi yapmadım.

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**3-İNFORMATİON GAİN HESAPLAMA**

****

Burada ilk başta kategorik tipte olan özellikler belirlendi. Bunlar benım veri setimde bulunan 1,3,5,17,18 numaralı kolonlardı. Bunları bir dizide tuttum ve döngü ile ilgili kolonlara dizinin indisini belirterek erişim sağladım. Bu kategorik tipteki kolonların hepsi iki değer alıyordu 1 ve 0. Yani bluetooth var-yok, wifi var-yok , dokunmatikekran var-yok gibi..

İnformation gain hesaplanırken ilk başta sistemin genel entropisi yani sınıf kolonunun entropisini hesaplamam gerekiyordu. Bunun için ilk başta hangi sınıftan ne kadar oldugunu python kodları ile buldum. Daha sonra bunu “sinifsayisi[]” dizisinde sınıf numaraları indis numaraları olacak şekilde tuttum. mesela 2 numaralı sınıftan toplam 174 adet bulunuyor ve bu sinifsayisi[2] de yer alıyor.(burada python kodları diğer hesaplamalarla da ilişkili olduğu için daha sonra aşağıda hepsini ekledım)

metin, ekran içeren bir resim

Açıklama otomatik olarak oluşturuldu SINIF SAYILARI

Daha sonra aşağıdaki kod ile sistem entropisi hesaplandı. Burada bulunan 666 sayısı satırların toplam sayısını 168,173,174,151 sayıları ise sırayla 0,1,2,3 numaralı sınıflara ait adet bilgisidir. Bu bilgilere yine sinifsayisi[] dizisinden ulaşıldı.



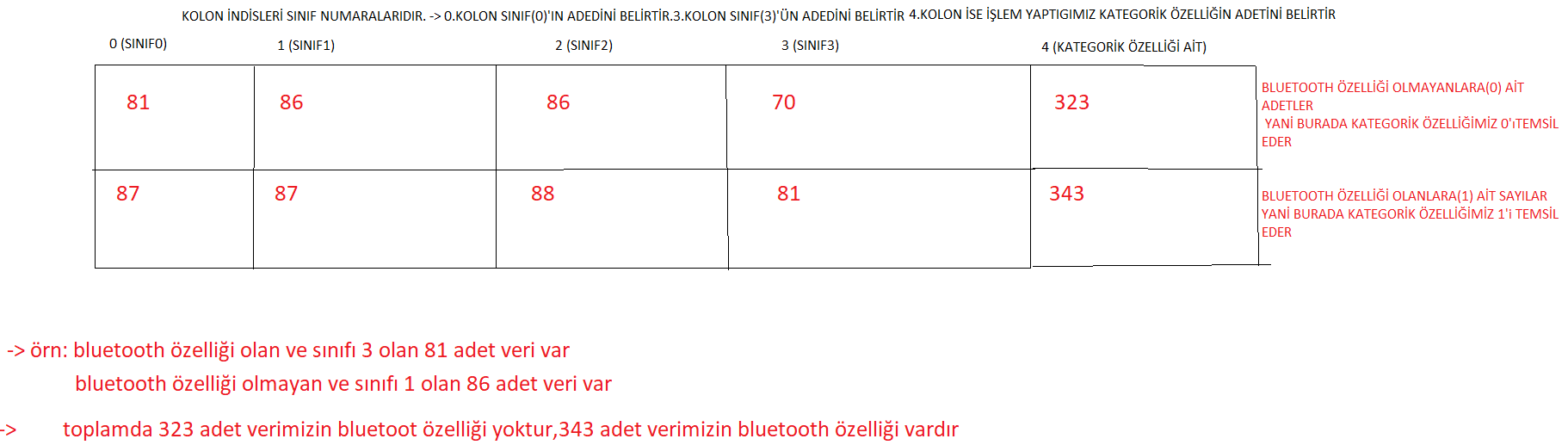


**-( 168/666\*log(168/666) + 173/666\*log(173/666) + 174/666\*log(174/666) +151/666\*log(151/666))**

--------------------------------------------------------------------------------------------------------

Sistemin entropisi hesaplandıktan sonra her kolona ait ağırlıklı entropiler bulunması gerekiyordu. Bunun için ilk başta seçilen kolon üzerindeki tüm satırlarda işlemler yapıldı. Yani örnek verirsek kategorik tipte olan kolonlardan 1.kolonu inceleyelim.(bluetooth var-yok).Bu kolonda ilk başta kaç adet bluetooth(0) ve bluetooth(1) olduğunu buldum. Şimdi buradaki bluetooth(0) özelliği üzerinden işlemimize devam edelim. Daha sonra sınıf kolonuma gittim ve bu sınıflardan hangisinde kaç adet bluetooth(0) olduğuna baktım ve buradan gelen sayıları diziye kaydettim. Bu işlem yine bluetooth(1) için de tekrarlandı.

Yukarıda anlattığım işlemlerden elde ettiğim verileri “sinifözellikiliskisayisi[]” adlı 2 boyutlu bir dizide tutum. Bu dizi kategorik özelliklerimizle alakalı adet bilgilerini tutmaktadır. Bu dizide kolonlar sınıfları, satırlar ise seçtiğimiz kategorik özelliğimizin aldığı değeri temsil eder. Anlatımı biraz zor olduğu için dizide yaptığım modellemeyi aşağıdaki görselde anlattım.

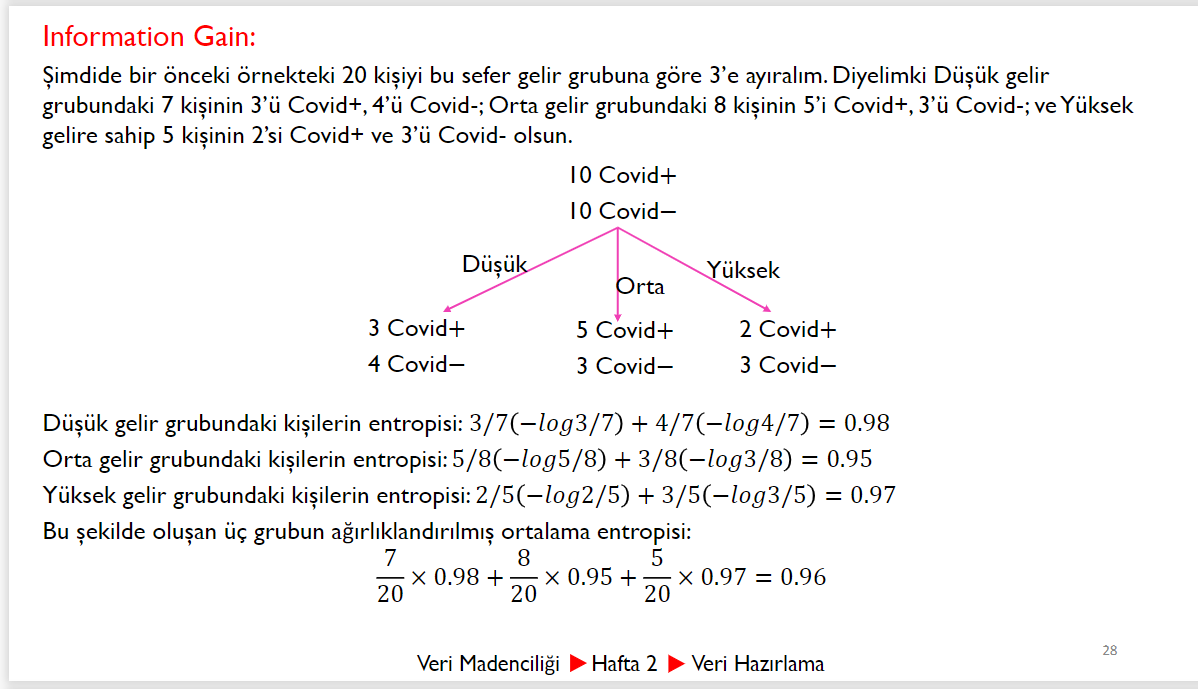


metin, ekran, elektronik eşyalar içeren bir resim

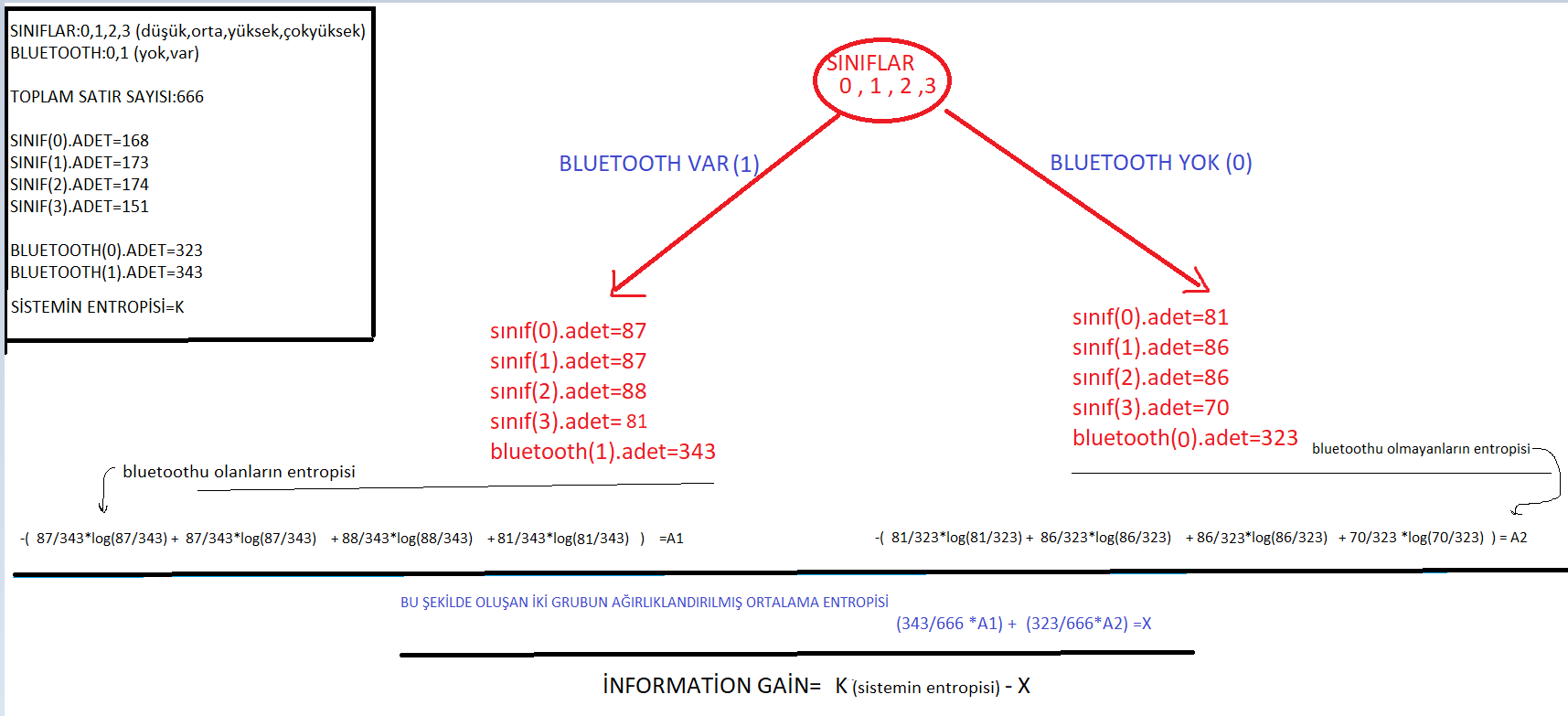
Açıklama otomatik olarak oluşturuldu 1 indisindeki kolona ait bilgilertablo içeren bir resim

Açıklama otomatik olarak oluşturuldu 17 indisindeki kolona ait bilgiler

Tüm kategorik kolonlar için bu veriler elde edildikten sonra sıra bu kolonların tek tek ağırlıklı entropilerini hesaplamaya geldi. Bunlar hesaplanırken dersin anlatım föyünde bulunan covid +- sınıflaması örnek alındı.



Ağırlıklı entropiyi hesaplama aşamasını yine bluetooth özelliği üzerinden anlatırsak bu kolonun ağırlıklı entropisinin hesaplanması ve information gain’in bulunması aşağıdaki görseldeki gibi yapılmıştır. Buradaki sayısal değerlere sinifözellikiliskisayisi[] dizisinden ulaşıldı. Bu hesaplamalar diğer kolonlar içinde benzer şekilde tekrar eder.



Yukarıdaki görselde information gain hesaplamam ile ilgili her şeyi anlattığımı düşünüyorum. 😊

Bu hesaplamalar sonucunda sistemimin çıktısı aşağıdaki gibidir.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Bu değerlere bakacak olursak bilgi kazançlarının çok düşük olduğunu görüyoruz. Ben bunun sebebini kategorik özelliklerimin sadece 0 ve 1 lerden oluşmasından ötürü olduğunu düşünüyorum. Çünkü elimde 4 adet sınıf var ama kategorik özelliklerim sadece 2 adet değerden oluşuyor. Rastgele seçilen bir kategorik özellikteki verinin sınıfının ne olduğunu yorumlamak bu yüzden zor olacaktır. Ayrıca yukarıdaki adet sayılarının oldugu tablodaki değerlere de bakarsak adet sayıları birbirine çok yakın. Sırf bu yüzden de direk olarak sınıfını kestirmek çok zor görünüyor.**

* **Kodlar 4.bölümün sonunda yer almaktadır**

**/\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**4-Entropi Hesaplama**

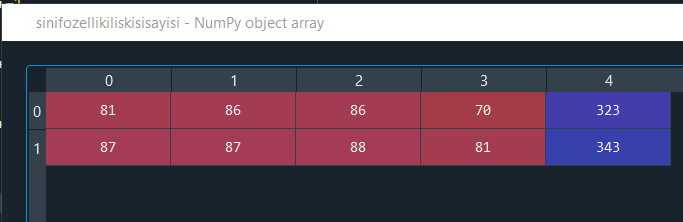


Kategorik özellikteki verilerimin entropilerini hesaplarken yine dersin anlatım föyündeki örneklerden yararlandım.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Burada kategorik tipte olan kolona eriştikten sonra bu kolon içerisinde bulunan 0 ve 1 değerlerinden ne kadar olduğu bulunur ve entropi için gerekli işlemler yapılıp kolonun entropi değeri bulunur. Bu işlem tüm kategorik kolonlar için tekrar edilir. Örnek olarak yine bluetooth özelliğinin olduğu kolonu ele alalım.



Bu tabloya bakacak olursak burada 343 adet bluetooth(1) degeri ve 323 adet bluetooth(0) degeri bulunmaktadır. Bu verilere nasıl eriştiğimi 3.bölümde anlatmıştım

**Bluetooth özelliği için Entropi= - ( 343 /666\*log(343 /666) + 323 /666\*log(323 /666) )**

Bu formül diğer kolonlara da uyarlanarak yapıldıktan sonra sistemin çıktısı aşağıdaki gibi olmaktadır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Burada dikkat edersek entropiler 1’e çok yakın. Bunun nedeni ise kategorik özelliklerimiz sadece 1 ve 0’lardan oluşuyor ve ilgili kolonlardaki 0 ve 1 değerlerinin toplam sayıları birbirine çok yakın. Bu yüzden belirsizlik çok fazla.**

**3 ve 4. Bölüme Ait Kodlar :**

**metin içeren bir resim

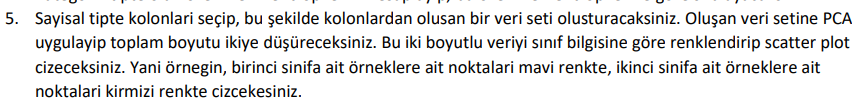
Açıklama otomatik olarak oluşturuldu**

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**5-SAYISAL TİPTEKİ ÖZELLİKLERE PCA UYGULANMASI**



Kategorik tipteki kolonlar “np.delete()” ile silindikten sonra sadece sayısal kolonların olduğu veri setini elde etmiş olduk. Buradan sonra hazır kütüphane içinde bulunan PCA işlemine ait fonksiyonlar ile özellik sayımızı ikiye düşürüp bu özellikteki değerleri kendi içinde düzenleyebiliyoruz. Ancak ben bu işlemlerden önce KNN algoritmasındaki başarı oranını artıracağını düşündüğümden raporun en başında anlattığım 1.görevdeki işlemleri de uyguladım. Yani ilgili sayısal kolondaki merkezden 2 standart sapma uzaklıktaki verilerin değiştirilmesinden bahsediyorum .Aşağıdaki kodda ortalamaya ve standart sapmaya göre değerler değiştirildi.

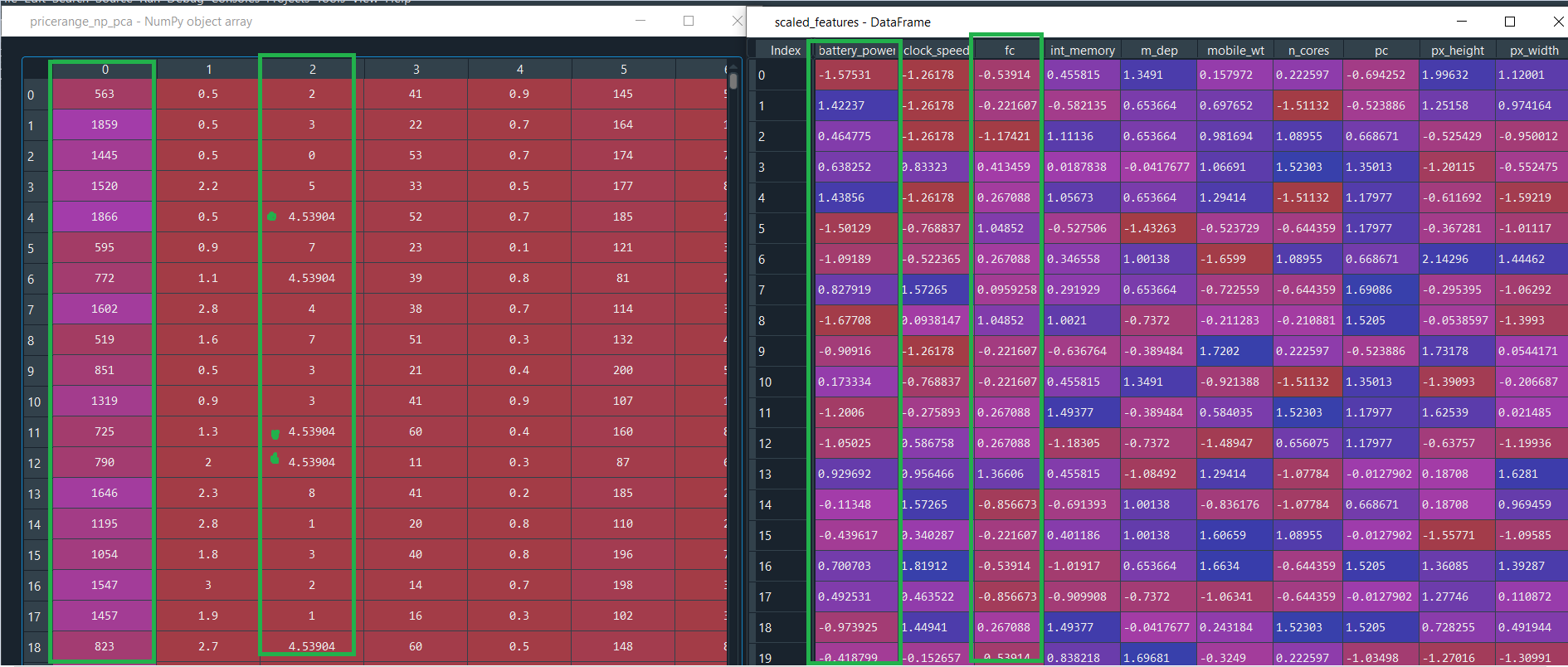
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu değerler değiştikten sonra veri setimden sınıfların bulunduğu sütunu ayırdım ve geriye kalan sayısal sütunlara ölçeklendirme işlemi yapıldı. Bu işlem kütüphane içerisinde bulunan fonksiyonlarla (StandardScaler) gerçekleştirildi.

metin, ekran görüntüsü, ekran içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu kod bloğundan sonra değişikler aşağıdaki gibidir (ölçeklendirme)

Bu işlemlerden sonra aşağıdaki fonksiyonlar yardımıyla özellik sayımız ikiye düşürüldü ve sınıf kolonu veri setimize dahil edildi.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Buradan elde ettiğimiz 2 özelliği bulunan veri setimdeki değerlerin bir kısmı aşağıdaki görseldeki gibidir.

metin, dizüstü, ekran görüntüsü, dolap içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran, vitrin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Daha sonra aşağıdaki kodlar ile bu veri setindeki sınıflar renklendirilip scatter plot çizildi.

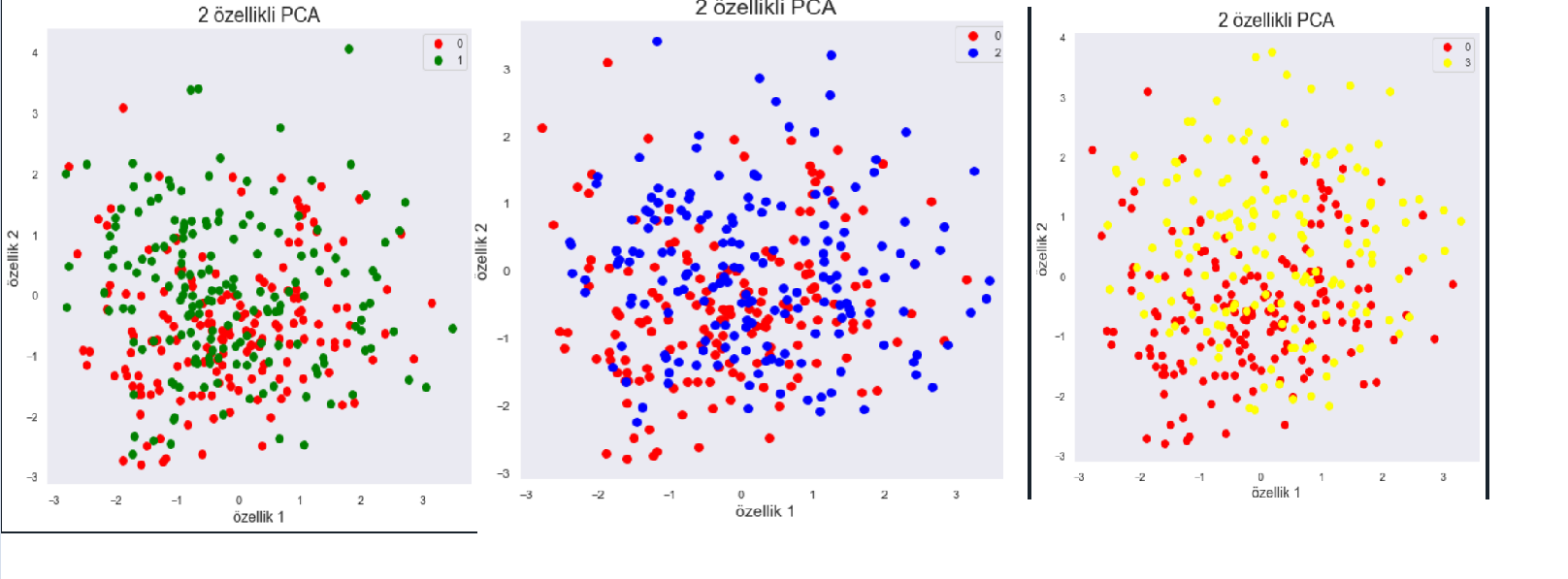
metin içeren bir resim

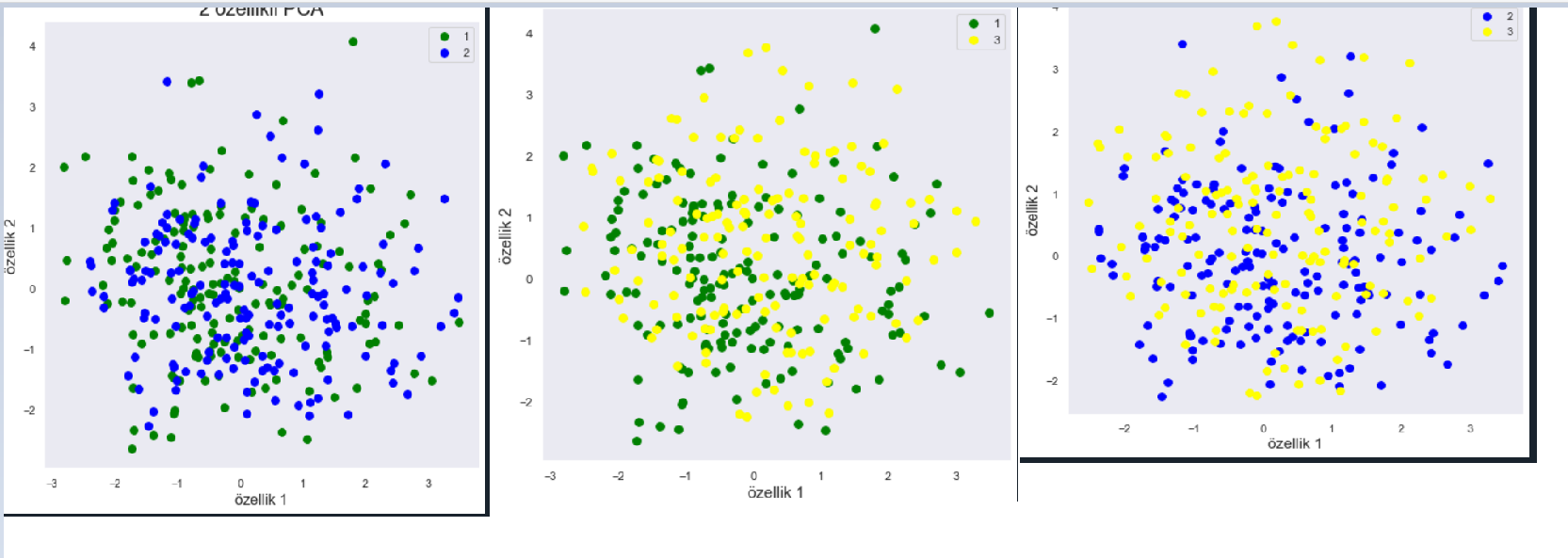
Açıklama otomatik olarak oluşturuldu

Bu kodlar sonrasında elde ettiğim scatter plot çizimleri ise aşağıdaki gibidir. Toplam 4 sınıf ilk başta bir arada daha sonra ise 2’li kombinasyonlar şeklinde bir arada gösterilmiştir.

metin, elektronik eşyalar, vitrin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

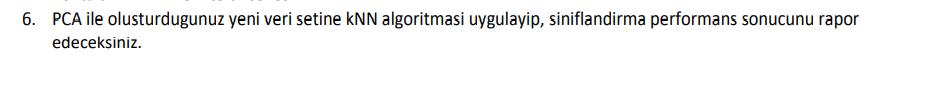




* **PCA sonucunda özelliklerin olduğu sütun sayısı 2 ye düştü ve satır sayım 666’da kaldı. Bu sınıf tahmin algoritmaları için çok iyi bir şey. Grafiklere de gelecek olursak internette araştırma yaptığım zaman genelde daha küçük bir veri seti olan Iris Dataset üzerinde PCA örnekleri var. Iris Dataset’te bu grafikler daha net. Ancak benim grafik çıktılarım tam olarak net bir belirginlik sunamıyor. Buna neden olan iki sebebin olduğunu düşünüyordum. İlk olarak veri setimde sınıf sayım 4 adet olmasına yorumladım ama bu da pek mantıklı gelmedi. Bir diğeri de veri setimin tüm satırlarıyla çalışmadım bana ayrılan satırlar ile çalıştım(2-5-8-11-..) belki burada bir veri kaybım olabilir diyordum ancak tüm veri setimi dahil ettiğim zaman da yine benzer sonuçlar alıyorum yine ayırt edebileceğimiz bir belirginlik yok .Çoğu fonksiyonu algoritmayı da direk olarak kütüphanelerden çağırarak kullandım... Belki de çıktılar doğrudur ama veri setimden dolayı sonuçlar iyi değildir.** ☹

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**6-PCA UYGULANAN VERİ SETİNE KNN ALGORİTMASININ UYGULANMASI**

****

K-en yakın komşu algoritmasında sınıfı öğrenilmek istenilen test setindeki verilerin eğitim setindeki verilere uzaklığı (öklid) hesaplanır ve bu uzaklıklardan en yakın k tanesinin sınıfı ne ise test setindeki o verinin de sınıfı o olur. Öklid uzaklığı hesaplanırken bazı değerlerin baskın gelmemesi için (örn maaş verisinin, yaş verisine olan baskınlığı) her sütundaki tüm değerler normalizasyona tabi tutulur. Benim veri setimde battery\_power, px\_height, px\_width ve ram kesinlikle normalizsasyona tabi tutulması gerekiyordu. Bu işlem başarı oranına direk etki ettiğinden tüm sütunlar için bu işlemi gerçekleştirdim. Daha sonra veri setimi 0.5 eğitim seti 0.25 test seti ve 0.25 doğrulama seti olarak ayırdım. Doğrulama testindeki verilerine knn uyguladım ve burada en çok başarı gösteren k değerini referans alıp test setine ‘k’ kadar uzaklıktaki verilerde en çok görülen sınıfı test setimdeki verinin sınıfı olarak belirledim.

* 1.bölümde sayısal kolonlarda ortalama ve standart sapmaya göre bazı işlemler yapmıştık bu işlemleri de tekrar yaptıktan sonra PCA ile iki sütuna düşürdüğüm veri setime derste yazılan KNN algoritmasını uyguladım. Derste yazılan KNN algoritmasını kullandığım için algoritmadan detaylıca bahsetmedim.

a)veri seti alındı,ortalama ve standart sapmaya göre veri düzeltme yapıldı

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

b) veri setimize PCA uygulandı

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

c)derste yazılan knn algoritmasının uygulanması

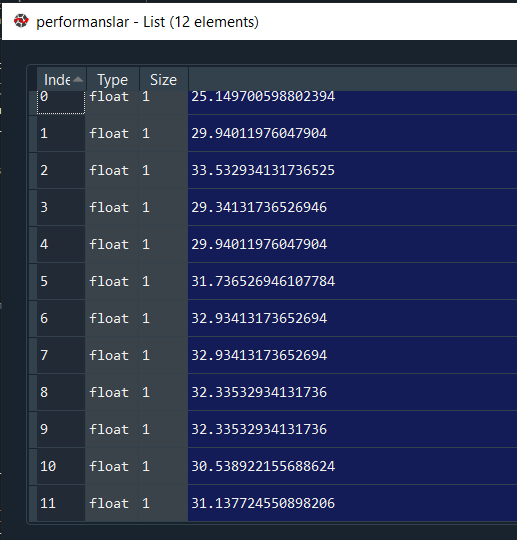
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

* **KNN Başarı Oranı:** 5.bölümde scatter plot ile çizdirdiğim grafikte sınıflar net olarak ayırt edilemiyordu. İşte o net olarak ayırt edememe sorunu direk olarak knn algoritmasının başarı oranına etki etti ve %30-%35 civarında bir başarı elde ediyoruz.Verileri karıştırıp yaptığımız için bu başarı oranı her seferinde değişiyor.Önceki denemelerimde en yükse %46’yı gördüm ☹

****

* **Bu sonuçlardan sonra algoritmayı kendi veri setime uyarlayamadım diye düşündüm ve sonra hazır kütüphaneden KneighborsClassifier’ı kullandım fakat sonuç burada da değişmedi.**

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, elektronik eşyalar içeren bir resim

Açıklama otomatik olarak oluşturuldu**

* Hazır kütüphane fonksiyonlarını da kullanınca elde ettiğim başarı bu şekilde ☹

/\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*/