

**T.C.**

**SİVAS CUMHURİYET ÜNİVERSİTESİ**

**MÜHENDİSLİK FAKÜLTESİ**

**BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ**

**DÖNEM ÖDEVİ**

**VERİ MADENCİLİĞİ DERSİ ÖDEV 1**

**HAZIRLAYAN**

**AHMET UTKU ELİK**

**5518123001**

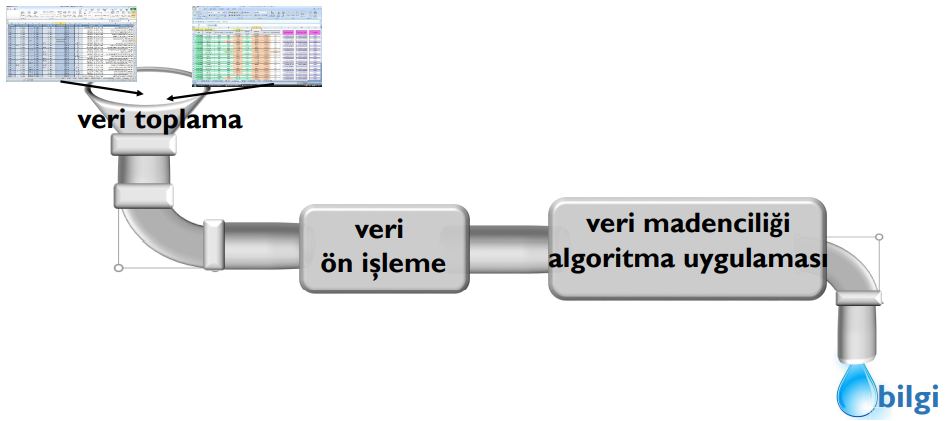
**DERS HOCASI**

**Dr. Öğr. Üyesi Fırat İSMAİLOĞLU**

**Sivas 2021**

**ÖZET**

Veri Madenciliği dersi 1. Dönem ödevi kapsamında hazırlanan bu raporda, veri toplama aşaması tamamlanmış hazır veri setleri üzerinde geliştirilen veri ön işleme ve sınıflandırma algoritmaları incelenecek ve sonuçları raporlanacaktır.



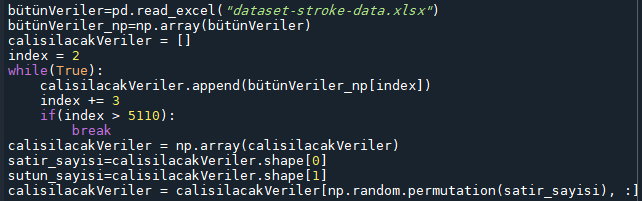
**Şekil-1** Bilgi Keşfi Süreci (Knowledge Discovery)

**Çalışılan Veri Seti:** Bu veri seti yaş, cinsiyet, sahip olunan çeşitli hastalıklar gibi bilgilerin kalp krizine etki edip etmedigini tahmin etmek için toplanmıştır.

**Veri Setindeki Özellikler (Kolon Bilgileri):**

* **gender:** cinsiyet
* **age:** yas (integer)
* **hypertension:** 1 ise kişinin hiper tansiyonu vardir, 0 ise yoktur.
* **heart\_disease:** 1 ise kişinin herhangi bir kalp rahatsizligi vardir, 0 ise yoktur.
* **ever\_married:** kişinin daha önce evlenip, evlenmedigi (Yes/No)
* **work\_type:** calisma türü:"children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
* **Residence\_type:** kişinin kirsalda yada şehirde yaşadiği: "Rural" or "Urban"
* **avg\_glucose\_level:** kandaki ortalama glikoz seviyesi
* **bmi:** vucut kitle indeksi
* **smoking\_status:** sigara icme durumu: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
* **stroke:** SINIF (tahmin etmek istediğimiz kolon): 1 ise kişi kalp krizi geçirmiştir, 0 ise geçirmemiştir.

İlk olarak excel dosyasında olan bu veri setini veri tipi Data Frame olan bir değişkende depolamamız ve arından öğrenci numaramızın 3’e bölümünden kalan sayının değerine göre çalışacak olduğumuz veri setinde karşılık gelen index değerlerindeki verileri (satırları) asıl çalışacak doluğumuz array veri tipindeki değişkene aktarmamız gerekmektedir. Öğrenci numaram 5518123001 olduğu için 3’e bölümünden kalan yani başlangıç index değerimiz 2 olacaktır. Şekil-2’de bu işlemi gerçekleştiren kod yapısı verilmiştir.

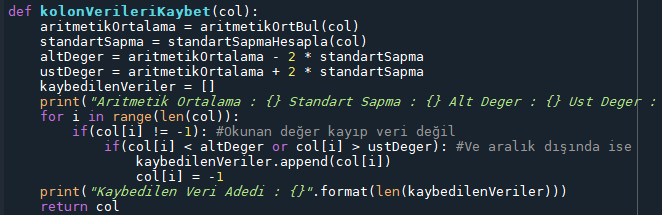


**Şekil-2** Çalışılacak Olan Veri Setini Oluşturma

**Veri Seti Oluşturulduktan Sonra Yapılacak Görevler:**

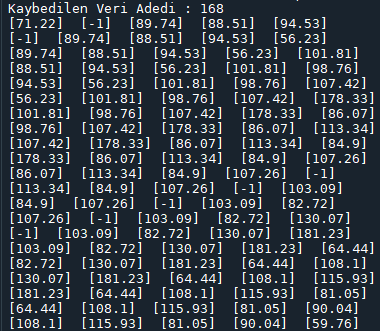
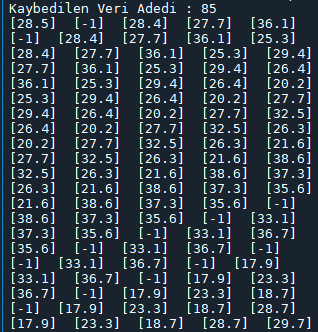
**1.** Sayısal tipte olan her kolon icin merkezden iki standart sapma uzakliktaki veriler atılacak. Hangi verileri attiginizi rapor edeceksiniz. Daha sonra bu verileri kayıp veri gibi düşünecek, veri setinde boşluk oluşmamasi için tahmin ederek (örnegin ortalama ile) bu boşlukları dolduracaksiniz. Hangi yöntemle kayıp verileri doldurdugunuzu rapor edeceksiniz.

Stroke (dataset\_1) veri setindeki sayısal kolonlar gender, avg\_glucose\_level, bmi özellikleridir ve bu kolonlara karşılık gelen indis değerleri ise (1,7,8)’dir. Bu sayısal kolonlar üzerinde daha rahat matematiksel işlem yapabilmek için “nan” veya “null” değerleri (–1) değeri ile sayısallaştırılmıştır. Arından her bir sayısal tipteki kolon için aritmetik ortalama ve standart sapma değerleri hesaplanıp, merkezden iki standart sapma uzaklıkta olan değer aralığı belirlenip altDeger = aritmetikOrtalama – 2 \* standartSapma, ustDeger = aritmetikOrtalama + 2 \* standartSapma şeklinde bu aralık değişkenlerde saklanmıştır. Her bir sayısal kolon için hesaplanan alt ve üst değerler kullanılarak merkezden iki standart sapma uzaklıktaki veriler atılıp kayıp veri olarak sayılmıştır(Değerleri -1 olarak güncellenir). Bu işlemi gerçekleştiren fonksiyon Şekil-3’te verilmiştir.



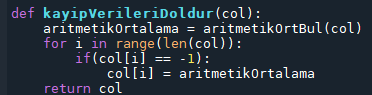
**Şekil-3** kolonVerileriKaybet() Metodu

Merkezden iki standart sapma uzaklıktaki verileri kaybetme algoritmamız sonucunda gender kolonundan 0, avg\_glucose\_level kolonundan 168 ve smoking\_status kolonundan ise 85 adet veri silinmiştir. Gender özelliği için belirlenen aralık [0, 87], avg\_glucose\_level için [18, 194] ve smoking\_status kolonu için [15, 43] şeklinde belirlenmiştir. Kaybedilen bu verilerden bir kısmını raporlayacak olursak avg\_glucose\_level kolonundan silinen veriler Şekil-4’te smoking\_status kolonundan silinen veriler ise Şekil-5’te verilmiştir.

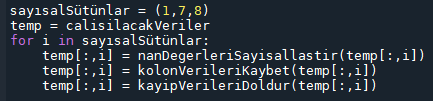
**Şekil-4** avg\_glucose\_level **Şekil-5** smoking\_status

Bu işlemlerden sonra veri setinde oluşan ve kayıp veri olarak kabul ettiğimiz boşlukları ortalama atama yöntemiyle doldurmaya çalışacağız. Artık sayısal kolonlardaki kayıp veriler -1 ile ifade edildiği için kolon üzerinde dolaşıp -1 değerini gördüğümüz noktalara o kolonun kendine ait airtmetik ortalama değerini atıyarak kayıp verileri doldurmuş ve veri setinde boşluk bırakmamış oluruz. Kayıp verileri bulan ve ortalama metodu ile dolduran metot Şekil-6’te verilmiştir.

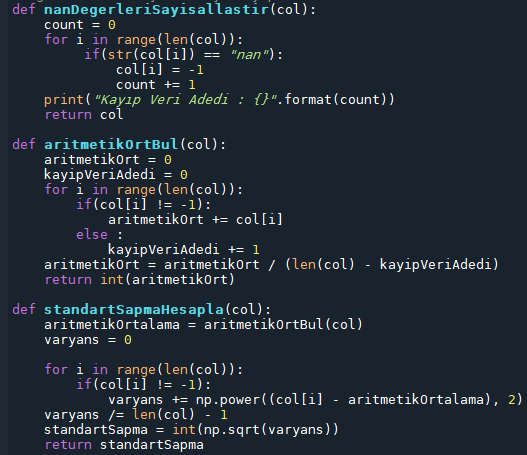


**Şekil-6** kayipVerileriDoldur() Metodu

Sayısal kolonlara (1, 7, 8) veri kaybetme ve doldurma işlemi Şekil-7’te verilen kod yapısı ile ve bu işlemleri yaparken kullanılan yardımcı metotlar ise Şekil-8’da verilmiştir.



**Şekil-7** Sayısal Kolonlar İçin Veri Kaybetme ve Kayıp Verileri Doldurma

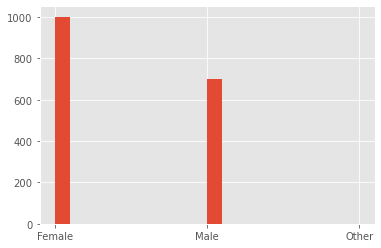


**Şekil-8** Yardımcı Metotlar

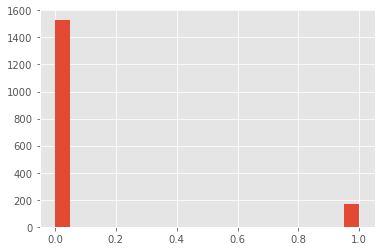
**2.** Kategorik tipte olan kolonlarda, histogram yapılacak. Bu histogramlari yormlayacaksınız. Histogram sonucunda eğer, yalnızca bir kaç defa görülen değerler varsa bir özellikte, bu değerleri o özellikte en çok görülen değer ile değiştireceksiniz. Bu şekilde yaptığınız değişiklikleri rapor edeceksiniz.

Stroke veri setindeki kategorik kolonlar gender, hypertension, heart\_disease, ever\_married, work\_type, Residence\_type ve smoking\_status özellikleridir. Özelliklerin karşılık geldiği indis değerleri ise (0, 2, 3, 4, 5, 6, 9)’dır. Bu özelliklerden bazıları 2 adet (gender, hypertension, heart\_disease, ever\_married ve Residence\_type) bazıları 4 adet (smoking\_status) bazıları ise 5 adet (work\_type) kategorik tip içermektedir. İşlem kolaylığı sağlaması için bu kategorit tipler enumerate edilerek sayısallaştırılmıştır. Örneğin Gender özelliği için 0 -> Kadın 1 -> Erkek, smoking\_Status özelliği için 0-> formerly smoked 1-> never smoked 2-> smokes 3-> Unknown olarak sayısallaştırılmıştır.

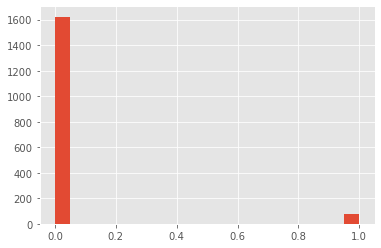
Her bir kategorik kolonun histogram grafiği aşşağıdaki şekillerde verilmiştir.



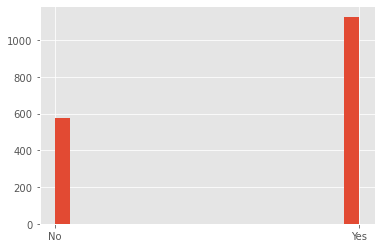
**Şekil-9** Gender Histogram Grafiği (Indis : 0)



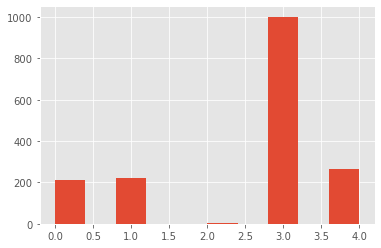
**Şekil-10** Hypertension Histogram Grafiği (Indis : 2)



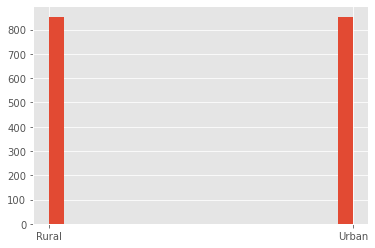
**Şekil-11** Heart\_disease Histogram Grafiği (Indis : 3)



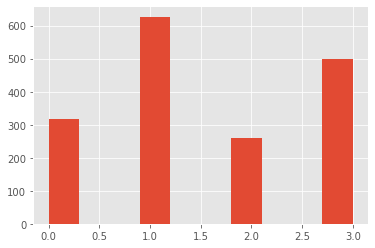
**Şekil-12** Ever\_married Histogram Grafiği (Indis : 4)



**Şekil-13** Work\_type Histogram Grafiği (Indis : 5)



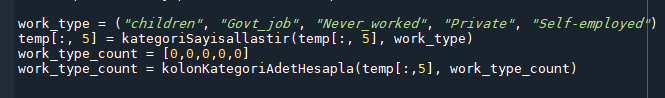
**Şekil-14** Residence\_type Histogram Grafiği (Indis : 6)



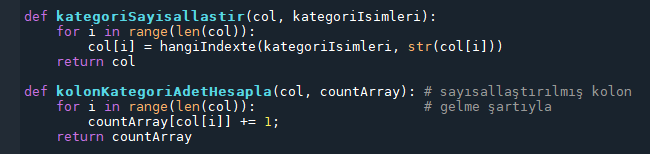
**Şekil-15** Smoking Status Grafiği (Indis : 9)

Bu histogram eğrilerinden yola çıkarak Hypertension, Heart\_disease ve Work\_type kategorik kolonlarının histogram eğrileri diğer kolonlardan farklı olarak kategorik veri tipi sayıları arasında ciddi bir fark vardır. Teorik olarak düşünürsek bu 3 kolondaki az rastlanan kategorik tip değerleri yerine o kolonun çok rastlanan kategorik tip değerini atamamız gerekir. Fakat bu elemeyi yaparken sonuca yani kalp krizi geçirme olasılığına direk tesir edebilecek Hypertension ve Heart\_disease özellikleri bu konuyla ilgili uzman olan ve alan bilgisine sahip kişilere danışılarak yapılması gerekir. Bu yüzden bu aşamada diğer 2 özelliğe kıyasla sonuca etki etme olasılığı daha düşük olan Work\_type özelliğinde az rastlanan kategorik tipe çok rastlanan kategorik tip değeri yani 2 yerine 3 değeri atanacaktır.

Work\_type kolonuna yani 5. İndisteki özelliğe ilk önce sayısallaştırma işlemi yapılmaktadır ardından work\_type\_count dizisinde 0 ile sayısallaştırılan değerin geçme adedi bu dizinin 0. İndisinde, aynı şekilde 1 ile numaralandırılan kategorik tipin geçme sayısıda aynı dizinin 1. İndisinde saklanıcak böylelikle bir değişken ile tüm kategorik tiplerin geçme adetlerini elde etmiş olacağız. Temel algoritma Şekil-16’da kullanılan metot ise Şekil-17’de verilmiştir.

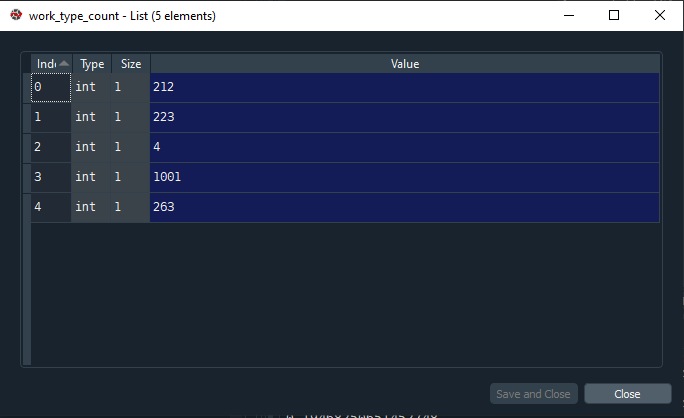


**Şekil-16** Work\_type\_count Algoritma



**Şekil-17** kolonKategoriAdetHesapla() ve Sayısallaştırma Metotları

Bu işlem sonucunda work\_type\_count dizisinin değeri Şekil-18’de verilmiştir.

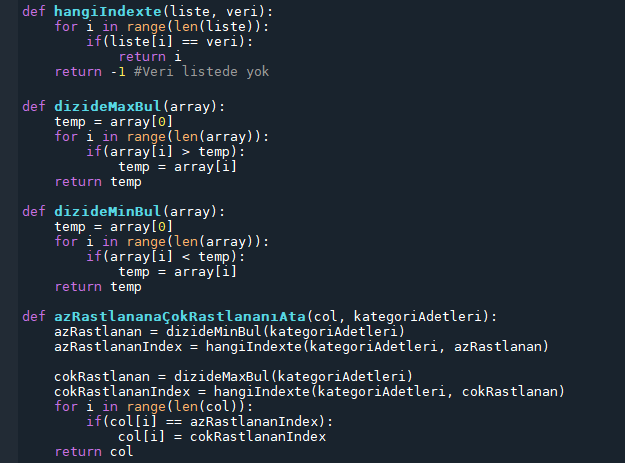


**Şekil-18** work\_type\_count Dizisinin Değerleri

Şekil-13’teki work\_type Histogram grafiği ve work\_type\_count dizisindeki değerler incelenecek olursa 2 ile numaralandırılan kategorik tip (Never\_worked) yerine 3 ile numaralandırılan kategorik tip olan (Private) değeri atanmalıdır. Bu işlemi gerçekleştirecek olan kod satırları ise Şekil-19 ve metodu ise Şekil-20’de verilmiştir.

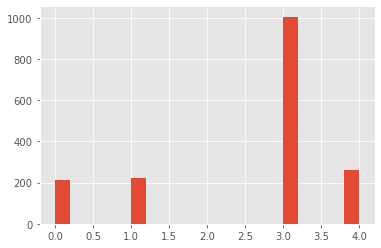


**Şekil-19**

****

**Şekil-20** Kategorik Tip Değiştirme Metotu ve Yardımcı Metotlar

Work\_type kolonu için yapılan kategori değiştirme algoritmalarından sonra Histogram grafiği Şekil-21’de work\_type\_count dizisinin değeri ise Şekil-22’de verilmiştir.



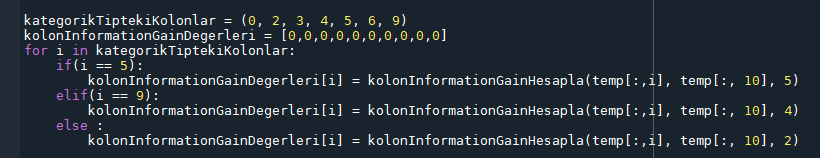
**Şekil-21** Yeniwork\_type Histogram Grafiği

****

**Şekil-22** Yeni work\_type\_count Dizi Değerleri

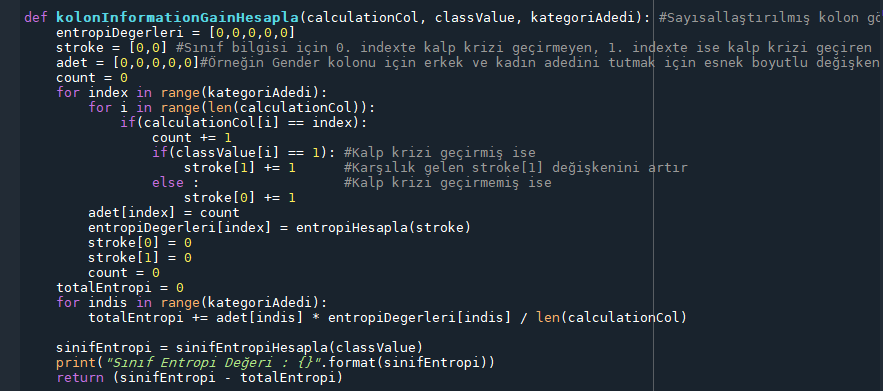
**3.** Kategorik tipte olan özellikler için information gain’i hesaplayıp, bu özellikleri information gain skorlarina göre siralayacaksiniz. Böylece her bir kategorik özelliğin SINIF’ları ayırt etmede ne kadar etki ettigini görmuş olacaksaniz.

Önceki madde de belirtildiği gibi kategorik tipteki kolonlar gender, hypertension, heart\_disease, ever\_married, work\_type, Residence\_type ve smoking\_status özellikleridir. Özelliklerin karşılık geldiği indis değerleri ise (0, 2, 3, 4, 5, 6, 9)’dır. Her bir kategorik kolon için information gain değerlerini hesaplayan kod yapısı Şekil-23’te ve kullanılan fonksiyonlar Şekil-24’te verilmiştir.



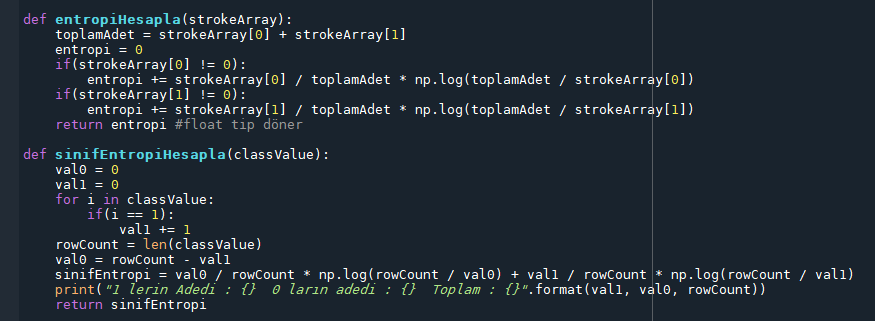
**Şekil-23** Information Gain Hesaplama Kod Yapısı

Kod yapısı incelenecek olursa, kategorik tipteki kolonların indis değerleri listede saklanmaktadır. KolonInformationGainDegerleri dizisinde ise kategorik tipteki kolonun indis değerine karşılık gelecek şekilde information gain değerleri saklanacaktır. Örnek verecek olursak work\_type kolonu 5 indise karşılık geldiği için kolonInformationGainDegerleri dizisindeki 5 numaralı indiste work\_type kolonunun information gain değeri saklanacaktır

.

**Şekil-24** kolonInformationGainHesapla() Metodu

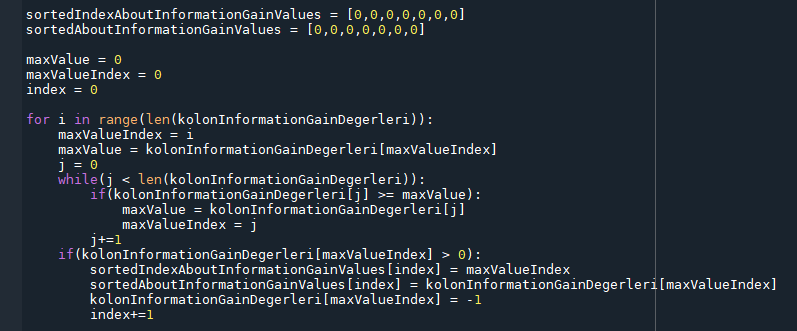
Bu metodu inceleyecek olursak entropiDegerleri adlı dizi değişkeninde kategorik tipteki kolon içerisinde bulunan veri tiplerinin ayrı ayrı entropi değerlerini saklar. Mesela gender kolonu için 0. indiste kadınların 1. İndiste ise erkeklerin entropi değerlerini saklar. Eleman sayısının 5 olmasının nedeni ise work\_type özelliğinde 5 adet kategorik tip bulunmasıdır. Stroke isimli dizi değişkeninin 0. indisinde kalp kirizi geçirmeyenleri 1. indisinde ise kalp krizi geçirenlerin adedini saklamaktadır iki indis değerinin toplamı 1703 olan çalışma veri adedini (satır sayısını) vermektedir. Adet isimli dizi değişkeninde ise her bir kategorik tipin adedini saklamaktadır. Bunun nedeni toplam entropiyi bulurken kullanılan formülden dolayıdır. Son olarak sınıf veya sistem entropisinden hesaplanması istenilen kolonun kendi entropi değeri çıkartılarak o kolonun Information Gain değeri hesaplanır. Bu metotta kullanılan diğer yardımcı entropiHesapla() ve sinifEntropiHesapla() metotları Şekil-25’te verilmiştir.



**Şekil-25** entropHesapla() ve sinifEntropiHesapla() Metotları

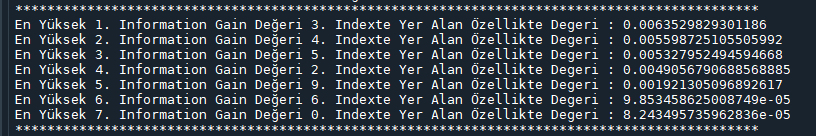
1703 adet veriden 83 tanesi kalp krizi geçiren geriye kalan 1620 adet verinin kalp krizi geçirmeyen ile sınıflandığı için sınıf yani sistem entropi değeri çok küçük çıkmakta. Sınıf Entropi Değeri : 0.19478104110077757

Şekil-26’da görülen kod bloğu ile hesaplanan kolonInformationGainDegerleri dizisi üzerinde işlemler yapılarak sortedIndexAboutInformationGainValues dizi değişkeni üzerinden Information Gain değerleri büyükten küçüğe doğru kolonların karşılık geldiği indis değerlerine göre sıralanmaktadır. SortedAboutInformationGainValues dizisinde ise Information Gain değerleri büyükten küçüğe doğru sıralanmıştır.

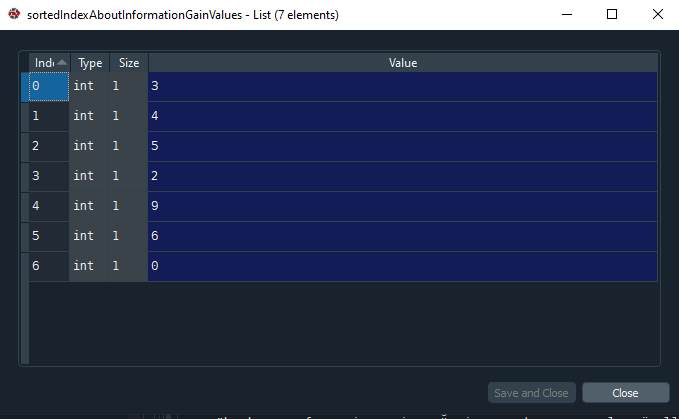


**Şekil-26** Kolonların Information Gain Değerlerini Sıralı Hesaplayan Kod Yapısı

Kategorik tipteki kolonların Information Gain değelerine göre sıralanmış hali Şekil-27’de gösterilmiştir.



**Şekil-27** Sıralı Olarak Kolon Information Gain Değeleri

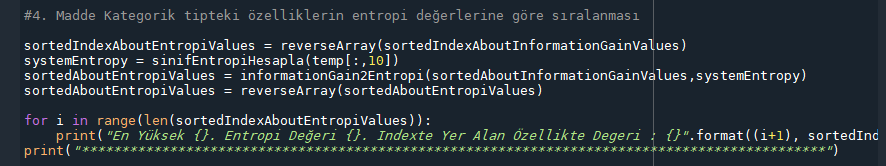


**Şekil-28** sortedIndexAboutInformationGainValues Dizi Değişkeni Değerleri

Information Gain değerlerine göre kategorik kolonları sıralayacak olursak büyükten küçüğe doğru sıralama şöyledir; heart\_disease > ever\_married > work\_type > hypertension > smoking\_status > Residence\_type > gender

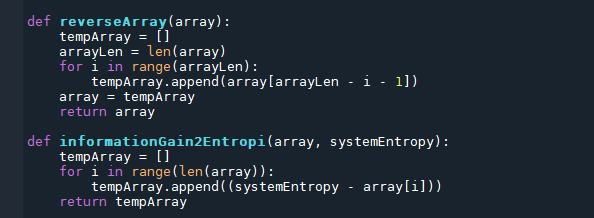
**4.** Kategorik tipte olan özelliklerin entropilerini hesaplayip, bu özellikleri entropilerine göre siralayacaksiniz.

Sınıf yani sistem entropi değeri sabit olduğu için (0.19478…) bu değerden önceki madde de hesaplanan Information Gain değelerini çıkartarak kategorik tipteki kolonların Entropi değelerini elde etmiş oluruz. Bunun için kullanılan kod yapısı aşşağıda verilmiştir.



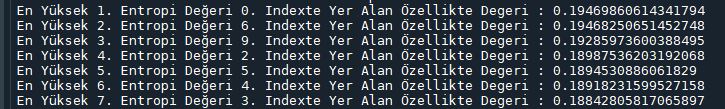
**Şekil-29** Entropi Kod Yapısı

Büyükten küçüğe doğru sıralanmış bir diziden tümleyen değerini çıkartırsak küçükten büyüğe sıralanmış halini elde ederiz. Bu nedene kategorik tipteki kolonların Entropi değerlerinin index sıralaması Information Gain değerlerinin index sıralamasının tam tersi olacaktır. Bu yüzden sortedIndexAboutInformationGainValues dizisinin tersini alırsak Entropi değerlerinin sıralı index hallerini elde ederiz. Ayrıca sortedAboutEntropiValues dizisini Sistem Entropisi – Kolonların Information Gain Degerleri = Kolonların Entropi Değerleri olarak hesaplayabiliriz bu işlemleri gerçekleştiren metotlar ise Şekil-30’da verilmiştir.

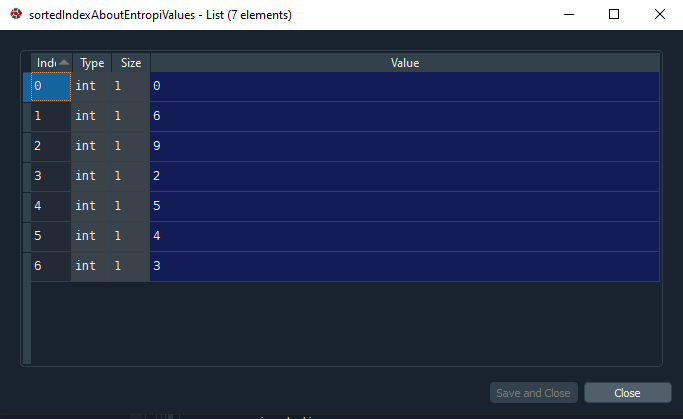


**Şekil-30** Information Gain Üzerinden Entropi Hesaplama Metotları

Bu işlemler sonucunda kategorik tipteki kolonların Entropi değerlerinin sıralı hali ve sortedIndexAboutEntropiValues dizi değişkeni değeri aşşağıdaki şekillerde verilmiştir.



**Şekil-31** Sıralı Olarak Kolon Entropi Değeleri



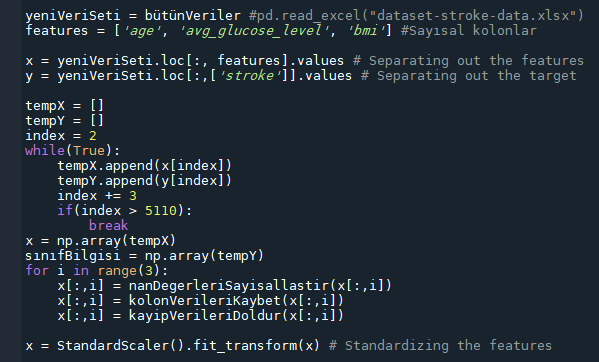
**Şekil-32** sortedIndexAboutEntropiGainValues Dizi Değişkeni Değerleri

Entropi değerlerine göre kategorik kolonları sıralayacak olursak büyükten küçüğe doğru sıralama şöyledir; gender > Residence\_type > smoking\_status > hypertension > work\_type > ever\_married > heart\_disease

**5.** Sayisal tipte kolonlari seçip, bu şekilde kolonlardan oluşan bir veri seti oluşturacaksınız. Oluşan veri setine PCA uygulayip toplam boyutu ikiye düşüreceksiniz. Bu iki boyutlu veriyi sınıf bilgisine göre renklendirip scatter plot çizeceksiniz. Yani örnegin, birinci sinifa ait örneklere ait noktalari mavi renkte, ikinci sinifa ait örneklere ait noktalari kirmizi renkte cizcekesiniz.

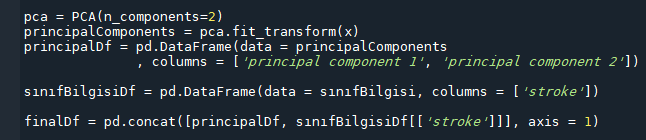
Sadece sayısal kolonları içeren bir veri seti oluşturmak için 2 yöntem vardır. Bunlardan ilki verimizin kaynağı olan “dataset-stroke-data.xlsx” excel dökümanındaki kategorik tipteki kolonları silip yeni bir excel dökümanına çevirip (Örneğin sayısalKolonlar.xlsx) sonrasında pandas.read\_excel ile okuyabiliriz. Fakat bu yöntem daha çok kategorik tipte kolon içeren veri setlerinde kullanışsız olacak ve uzun zaman alacaktır. Diğer bir yöntem ise Data Frame veri tiplerine uygulayabileceğimiz DataFrame.loc[‘kolonIsmi’].values komutu ile istenilen kolonların isimlerini girip sonucunda dizi formatında veri setimizi elde edebiliriz. Veri setini elde etme ve ardından merkezden iki standart sapma uzaklıktaki verileri kaybedip sonrasında kayıp verileri doldurma işlemini gerçekleştirdiğimiz kod yapısı Şekil-33’te verilmiştir

.



**Şekil-33** Yeni Veri Seti Oluşturma Kod Yapısı

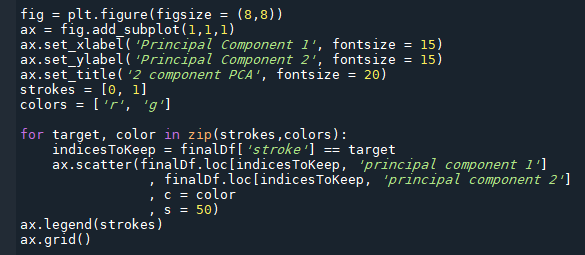
Veri setimizi elde ettikten sonra PCA(Temel Bileşenler Analizi) uygulayıp toplam boyut 2’ye düşürülmüştür. Bunu gerçekleştiren kod yapısı Şekil-34’te verilmiştir.



**Şekil-34** PCA Kod Yapısı

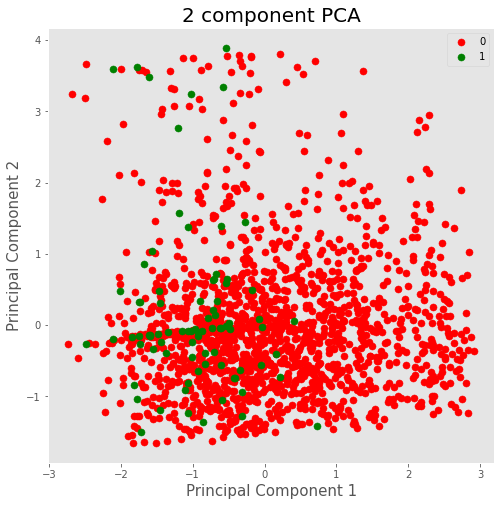
Bu kod yapısında PCA ile düşürmek istediğimiz boyut adedi girildikten sonra principalDf değişkeninde 2 boyuta düşürülmüş sayısal kolonlar bulunmaktadır ve sınıfBilgisiDf değişkeninde ise bu değerlere karşılık gelen Sınıf bilgileri yer almaktadır. Sonrasında pd.concat() metodu ile bu iki Data Frame tipindeki değişkeni tek bir Data Frame olarak birleştiriyoruz.

Elde edilen finalDf Data Frame’inin Scatter Plot grafiğini ise aşşağıdaki kod yapısı ile elde etmekteyiz.

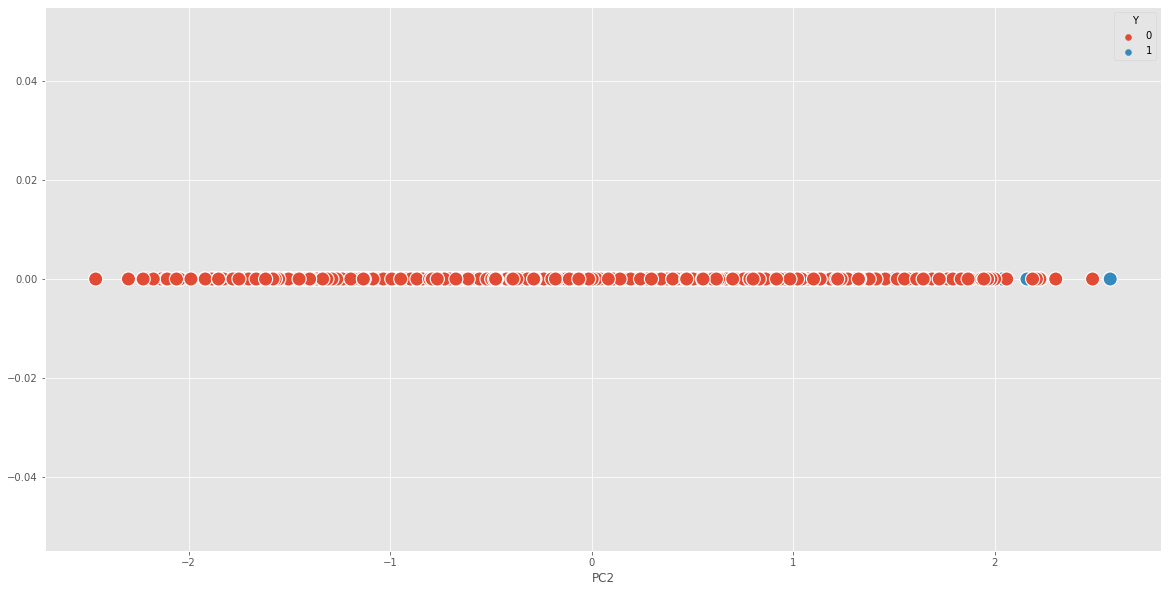


**Şekil-35** Scatter Plot Kodları

Çizim sonucunda elde edilen Scatter Plot grafiği ise şöyledir;



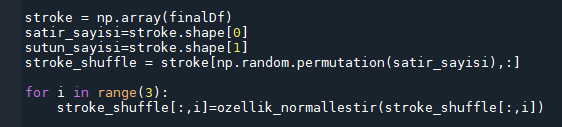
**Şekil-36** PCA Scatter Plot Grafiği



**Şekil-37** Seaborn Scatter Plot Grafiği

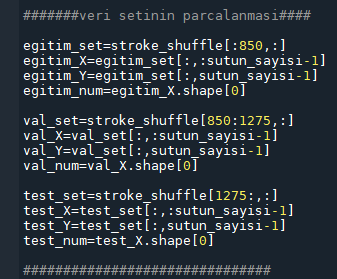
**6.** PCA ile oluşturduğunuz yeni veri setine kNN algoritmasi uygulayıp, sınıflandırma performans sonucunu rapor edeceksiniz.

PCA ile oluşturduğumuz finalDf veri setini okuma ve ardından özelliklerini normalleştirip kNN algoritmasından maksimum verim almak için shuffle işlemini aşşağıdaki kod blokları ile yapılmıştır.



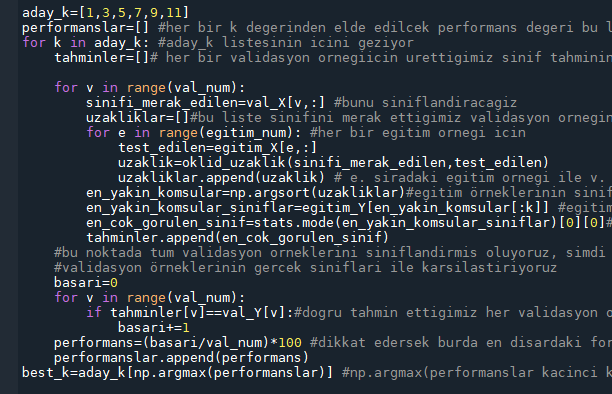
**Şekil-38** kNN İçin Veri Seti

Sonrasında algoritmamızı eğitebilmek ve test edebilmek için parçalara ayırdık. Eğitim Seti %50, Değer Seti %25 ve Test Seti %25 orantılı olacak şekilde satırlar parçalandı.

****

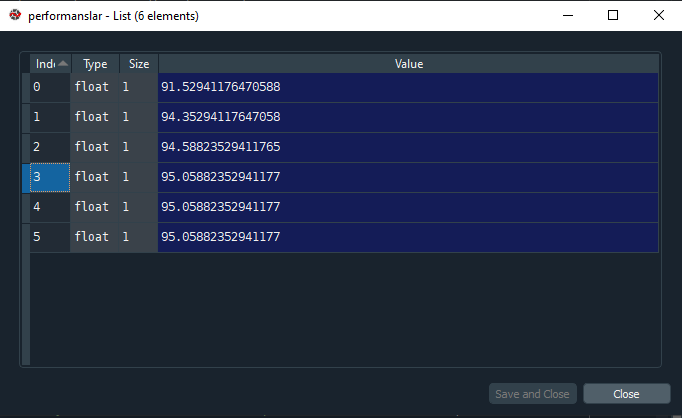
**Şekil-39** Veri Setinin Parçalanması

Performans değerlerini tespit ederkende Şekil-40’daki kod yapısı kullanılmıştır.



**Şekil-40** Performans Değeri Tespit Etme

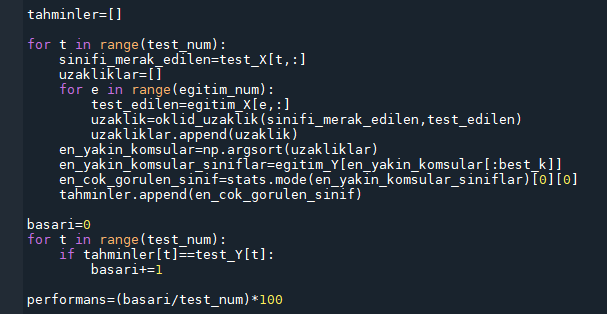
En yakın komşu algoritmasını (1,3,5,7,9,11) değerleri için test edip performanslar dizisinde verimleri toplandı. Elde edilen sonuçlar Şekil-41’de verilmiştir.



**Şekil-40** (1,3,5,7,9,11) kNN Performans Değerleri

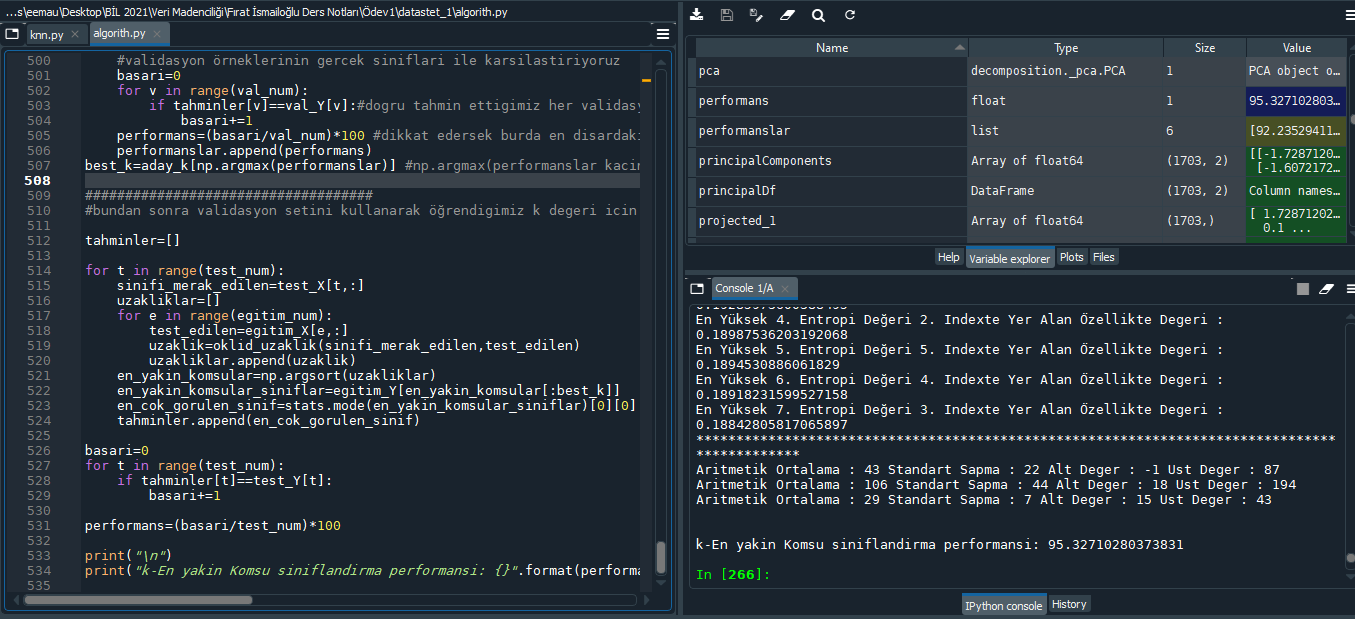
Bu dizi değerlerini inceleyecek olursak best\_k değeri 3. indiste olan 7 değeri olacaktır.

Validasyon setini kullanarak öğrendiğimiz best\_k değerini (k = 7) kullanarak test örneklerini sınıflandırıp tahminlerimle eşleşme durumuna göre en yakın komşu sınıflandırma performansını elde etmekteyiz. Bunu gerçekleştiren kod yapısı Şekil-41’de verilmiştir.



**Şekil-41** kNN Performans Hesaplama

Sonuç olarak **%95.327**’lik bir performans elde edilmiştir.



**Şekil-42** kNN Performans Çıktı

**EKLER**

**Kaynak Kod Linki :** https://github.com/Rexoes/Veri-Madenciligi-Donem-Odevi