

**Ege Üniversitesi**

**Fen Bilimleri Enstitüsü**

**Bilgisayar Mühendisliği Anabilim Dalı**

**Prof. Dr. Aybars Uğur**

**2020-2021 Güz Yarıyılı Proje 1**

**Kalp Hastalıkları Sınıflandırılması**

**Yaren Gündüz**

**91200000186**

**SORU 1**

Veri Madenciliği ve Makine Öğrenmesinde Normalization nedir? Araştırınız. Kısaca tanımlayınız. Ne amaçla kullanıldığını açıklayınız. Bir normalizasyon kodu yazarak ne yaptığını anlatınız. İlgili kaynaklara atıf veriniz. (20 p)

**CEVAP 1**

Normalleştirme de asıl amaç veri kümesi içerisinde bulunan değerlerin aralılarındaki farklılığı bozmadan ortak bir paydada buluşturmaktır. Normalizasyonda parametrelerin aynı pozitif ölçeğe sahip olması yararlı olabilir. Bu durumda da outlier denilen uç değerlerin kaybolmasına da neden olabilir.

tablo içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Figür 1.1 Normalizasyon Açıklaması [1]*

Normalizasyonun ne olduğunu [1] kendi kaggle hesabımda açıkladığım gibi Figür 1.1’de görebiliriz.

Peki biz ne zaman normalizasyon kullanmalıyız. Normalleştirme aslında verilerin dağılımını bilmediğinizde veya dağılımın çan eğrisi olmadığını bilmediğimizde kullanmak iyi olur. Verilerimiz farklı ölçekte ve kullandığımız algoritma, KNN algoritmaları gibi veri dağılımı hakkında varsayımlar yapmadığında yararlıdır.

Normalleştirme 3 farklı şekilde olur.

* Min-max Normalizasyonu
* Z-score Normalizasyonu
* Ondalık Normalizasyonu

Minimum-maximum normalizasyonunda orijinal veriler minimum 0,maximum 1 aralığında yeni verilere dönüştürülür.

Z-score normalizasyonunda herhangi bir y değeri, değişkenin ortalaması ve standard sapmasına bağlı olarak Z -dönüşümü gerçekleştirir.

Ondalık normalizasyonunda , ondalık ölçekleme ile normalleştirmede ise ele alınan değişkenin değerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Figür 1.2 Normalizasyon Formülleri[2]*

Yukarıdaki görseldeki gibi normalizasyon formülleri [2] bu şekildedir.

Benim kendi kernelimde işlediğim normalizasyon ise min-max normalizasyonu idi. Başlangıçta mantıksal regresyon için oluşturduğum sonuç değerine ait değerleri içeren bir y değerim oldu ve x değişkenime ait verileri hedef kolonumu veri çerçevemden silerek x\_data değişkenimi yarattım.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Fig 1.3. Normalization Code[1]*

Daha sonradan yaptığım işlemde koduma normalizasyonu *Fig 1.3 deki gibi* entegre ettim.

**SORU 2**

1. **Problem ve Veri Seti (20 p)**

**Python scikit-learn kütüphanesini kullanarak bir makine öğrenmesi uygulaması yazınız.** Sınıflandırma ve tahminleme konularından istediğiniz birisini tercih edebilirsiniz. Alanda deneyiminiz yoksa sınıflandırma konusunu seçmenizi öneririm.

# UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets.html>

<https://machinelearningmastery.com/standard-machine-learning-datasets/>

<https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b>

# The 50 Best Public Datasets for Machine Learning:

https://medium.com/datadriveninvestor/the-50-best-public-datasets-for-machine-learning-d80e9f030279

bağlantılarındaki veri setlerinden **derste anlatılandan (Zambak) farklı** herhangi birisini tercih edebilirsiniz. Dileyenler kendileri de veriseti oluşturabilirler.

Seçtiğiniz problemi tanımlayınız (otomobillerin özelliklerine göre çok iyi, iyi, orta, kötü olarak etiketlenmesi… gibi). Probleminiz hangi makine öğrenmesi paradigması (sınıflandırma, kümeleme, diğer) kullanılarak çözülebilecektir? Belirtiniz.

**Verisetini inceleyip özet bilgileri rapora yazınız: Veri (örnek) sayısını, öznitelik (girdi) sayısını, özniteliklerin neler olduğunu, sınıf sayısını ve sınıfların neler olduğunu rapora yazınız. Örnek:** Araba sınıflandırması uygulaması için UCI ML Repository Data Folder’dan **Car Evaluation** veriseti -> 6 adet feature – öznitelik ve 4 adet kategori sınıf içeriyor : <https://archive.ics.uci.edu/ml/datasets/Car%20Evaluation>.

**CEVAP 2**

Ben çalışacağım (excel || csv ) dosyamı kaggle’da <https://www.kaggle.com/ronitf/heart-disease-uci> linkte bulunan veri kümesini seçtim. Bu projedeki asıl amacım kalp rahatsızlığı bulunan hastaları sınıflandırma üzerine bir çalıştım.

Seçtiğim problemi tanımlamak gerekirse hastaların yaşlarına, cinsiyetlerine, kalp atış hızlarına, kandaki açlık şekerine kadar birbirine bağlı değişenlere göre bireyin hasta olup olmadığını sınıflandırdım. Burada tek hedefe (kalp hastalığı var mı ya da yok mu?) bağlı 2 ayrı durum çıktımız olabilir ya hastadır, ya da hasta değildir.

Burada bu veri kümesi içerisinde bulunan hedef kolonunun dışında bulunan kolonlar benim aslında özniteliklerimdir, çünkü ben bunlara bakarak bir sınıflandırma yapacağım. Kendi kernelimde de belirttiğim gibi

Seçtiğim probleme bağlı olan veri kümesindeki kolonlar ve açıklamaları;

* *age -> yaşlar*
* *sex -> cinsiyetler(erkek=1,kadın=0)*
* *cp -> chest pain type - göğüs ağrı tipi*
* *trestbps -> resting blood pressure - istirahat tansiyonu*
* *chol -> serum cholestoral in mg/dl - mg/dl cinsinden serum kolestrol*
* *fbs -> fasting blood sugar > 120 mg/dl*
* *restecg -> resting electrocardiographic results (values 0,1,2) - açlık kan şekeri (varsa 1, yoksa 0)*
* *thalach -> maximum heart rate achieved - istirahat halindeki elektrokardiyografik sonuçlar*
* *exang -> exercise induced angina - egzersize bağlı anjin (varsa 1,yosa 0)*
* *oldpeak ->ST depression induced by exercise relative to rest - Dinlenmeye bağlı egzersizin neden olduğu ST depreyonu*
* *slope -> the slope of the peak exercise ST segment - ST segmentinin eğitinin tepe noktası eğimi*
* *ca -> number of major vessels (0-3) colored by flourosopy - florosopi ile renklendirilmiş ana damarların tal sayısı 0 ile 3 arasında veri alıyor.*
* *thal -> thal: 3 = normal; 6 = fixed defect; 7 = reversable defect - 3 normal,6 sabit kusut,7 tersine çevirilebilinir mi diye sınıflanırılmış bir usur olup olmadığı ile ilişkili.*
* *target -> 1 veya 0 - sonuç eğer kalp rahatsızlığı varsa 1,yoksa 0 diye veri kümesine girilmiş.*

Toplamda 303 tane veri kaydı var ve toplamda 13 tane öz nitelik değerlerim var.13 ayrı kolon değerine göre şekillendirilmiş 303 satırlık bir veri kümesi. Asıl amaç kalp rahatsızlığı olan hastalarla olmayanları sınıflandırmaya çalışmak. Bu sınıfta toplamda target değerine göre 2 kategorim var. Ya hasta olabilir yada olamaz ama bu duruma etken olan diğer 13 kolonda benim asıl özniteliklerimdir. Kategori sayısı 2 tanedir. Bu problem için hastalık varsa 1, yoksa 0 değerini alacaktır. Seçtiğim problem sınıflandırma yapılarak çözümlenecektir. Çünkü 2 kategorisi bulunan ve hedefi kalp hastalığı teşhisi olan bir problemde maximum verilebilenecek cevap hastalığın olma durumu ->1 veya hastalık bulunmama durumu ->0 ‘dır. Bu şekilde diğer öznitelikleri inceleyerek hedefte olan 2 kategoriye bağlı sınıflandırma yaptım.

**SORU 3**

**Yöntem Kodlama:** İki farklı sınıflandırıcı (MLP Classifier YSA, SVM, k-NN, Decision Tree, Random Forest …) / tahminleyici … kullanarak sınıflandırma / tahminleme … işlemini yapan Python kodunu yazınız. Rapora, seçtiğiniz yöntemlerin isimlerini yazınız ve kodu ekleyiniz **(20)**

**CEVAP 3**

Bu projemde DecisionTree ve RandomForest sınıflandırıcılarını seçtim. Bu iki farklı sınıflandırıcı ile işlemlerimi yaparken aslında RandomForest algoritmasında DecisionTree’ ye göre daha iyi doğruluk değerlerini çıkardığını gördüm

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Fig 3.1. ML Sınıflandırıcıları [1]*

Bu kodlarda asıl anlatılmak istenen sklearn kütüphanesine ait sınıflandırıcı algoritmalar kütüphaneden yüklendi ve çağırıldı, ardından kendim train verilerimi kullanarak modeli eğittim ve ardından test değerlerimi vererek test doğruluk değerini hesaplamaya çalıştım buradaki amaç ne kadar yüzdelikte sdoğru sınıflandırma yapabilmiş, bunu öğrenmeye açlıştım ve sonunda Random Forest’ın karar ağacına göre daha iyi öğrenip-test sonucu verdiğini gözlemledim.

**SORU 4**

**Deneysel Çalışma**: a) Verileri eğitim ve test verisi olarak ayırınız. Madde 3’te seçtiğiniz her 2 yöntem için de : a) Hata (Confusion) matrisini ve doğruluk (accuracy) değerini, b) 10-fold cross validation ile de doğruluk değerlerini (ve ortalamasını) elde ederek rapora ekleyiniz. http://scikit-learn.org/stable/modules/cross\_validation.html. Sonuçları yorumlayınız. **(20)**

**CEVAP 4**

Bu iki yöntem için aşağıda karmaşıklık matrisi verilmiştir. Karmaşıklık matrisi aslında bize veriler hakkında detaylı bir yol haritası çizer. Asıl amaç gerçek bulunan değerler ile tahminler arasında bir köprü kurmaktır.

TP(True Positive),TN( True Negative),FP(False Positive),FN(False Negative) değerlerini eğitim ve test süreçleri sonrasından elde edilen miktarlarını bize açıklamaya yarar sağlar. Böylelikle analiz edilen veri sonuçları aslında bize kalp hastalığı teşhisi gerçekte olan ve ne kadar gerçek teşhis konmuş, ne kadar hatalı teşhis konmuş ve hatta ne kadar doğru hastalık sınıflandırılması yapılmış bunu öğretir. metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Fig 4.1. Karmaşıklık Matrisi*

Ben bu veri kümem için 10 cross validation uyguladım. Bunların doğruluk değerleri aşağıdaki gibidir.

* **Decision Tree Doğruluk Değeri : 78.68852459016394**
* **Random Forest Doğruluk Değeri : 88.52%**

Bu iki yöntemin ortalama değeri ise 83% doğruluktadır diyebiliriz. Buna en yakın gerçek doğruluğu veren Random Forest algoritmasıdır.

**SORU 5**

Özdeğerlendirme Tablosu: (Aşağıdaki tablo da öğrenci tarafından doldurulmalıdır)

*Özdeğerlendirme tablosu teslim edilmeyen projeler değerlendirilmemektedir.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Madde Numarası** | **Puan** | **Var** | **Açıklama** | **Tahmini Puan** |
| 1. Normalizasyon | 20 | Var | Normalizasyonu python’a ait bir kod ile değil internetten araştırarak kendim formülize edip kodladım. | 20 |
| 1. Problem ve Veri Seti | 20 | Var | Problemim kalp hastalığı ile ilgili ve kaggle üzerinde bu veri kümemi buldum. Özelliklerini okudum ve kolonlarını ayrı ayrı inceleyip kodu nasıl yazacağımı belirledim. Kolonların hedefle ilişkisini inceledim. | 20 |
| 1. Yöntem Kodlama | 20 | Var | Yöntem kodlama olarak Logistic Regression kendim kodlamayı denedim ve rapora,kodlara yansıttığım kadarıyla Decision Tree ve Random Forest algoritmaları üzerinde de işlem yaptım. | 20 |
| 1. Deneysel Çalışma | 20 | Var | Deneysel Çalışmada yazdığım bütün kodların bilincinde olup artık verilerle konuşma kısmında gerekli analitik analizi yaptığımı düşünüyorum ve bu veri kümesi için en iyi oturan algoritmayı doğru tespit ettiğimin kanısındayım. | 20 |
| 1. Özdeğerlendirme Tablosu | 20 | Var | Öz değerlendirme tablosunda tamamen kendi irademle kendi yaptığım projemi güzel bir şekilde tamamladığımı ve doğru değerlendirdiğimi düşünüyorum | 20 |
| Toplam | 100 |  | | 100 |

**Teslim Tarihi ve Şekli:** Kaynak kod ve rapor paketlenerek son teslim tarihine kadar EGE Derse yüklenmelidir. Süre : 2 hafta, Teslim Tarihi : 23 Aralık 2020

Reference :

1 - <https://www.kaggle.com/yarengndz/kalp-hastal-klar-s-fland-r-lmas>

2 - <http://www.ttbilgin.com/2018-2019-guz/veri-bilimine-giris/hafta3/veri_on_isleme.pdf>

NOTE \*\* Kodlarımı Google Colab’da ve Kaggle’da yazıp-gerçekledim.

.ipynb dosyası ve .pynb dosyası olarak indirdim ve ekteti zipte kodları bulabilirsiniz. Eğer herhangi bir şekilde erişemezseniz aşağıdaki link üzerinden erişebilirsiniz.

<https://www.kaggle.com/yarengndz/kalp-hastal-klar-s-fland-r-lmas>