27 Ocak 2021

ÖMER FARUK YILDIZ-2028142006

Karabük Üniversitesi

Biyomedikal Mühendisliğinde Yapay Zeka ve Uygulamaları

Makine öğrenmesi ile iyi huylu-kötü huylu meme kanseri tahmini

İçindekiler

[**Özet** 2](#_Toc62652033)

[**1.** **Giriş** 2](#_Toc62652034)

[**1.1.** **Veri Setine Genel Bir Bakış** 2](#_Toc62652035)

[**1.2.** **Veri Hazırlık Aşaması** 3](#_Toc62652036)

[**2.** **Çok Katmanlı Algılayıcılar (MLP) ile Model Oluşturma** 4](#_Toc62652037)

[**3.** **Keras ile Model Oluşturma** 5](#_Toc62652038)

# **Özet**

Birçok ülkedeki farklı sağlık sistemlerinden elde edilebilen gerçek zamanlı hasta verileriyle gömülü ML tabanlı uygulamalarla karşılaşmak oldukça yaygınlaşmıştır. Bu nedenle hastalığın tanısı ve tedavisinde yapay zekâ uygulamaları oldukça artmıştır. Kullanılan meme kanseri veri seti, kanser hücrelerinin iyi huylu veya kötü huylu olup olmadığını tahmin edilmesi ile ilgilidir. Yapılan çalışmada iyi huylu ve kötü huylu kanser hücrelerine ait 10 özellik kullanılarak Makine Öğrenmesi modeli oluşturulmuş ve %96 oranında başarı elde edilmiştir.

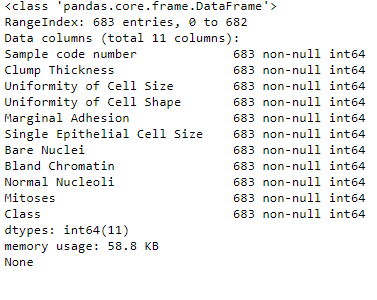
# **Giriş**

Meme kanserinin tespit edilmesi için farklı hastalardan alınan ve kanserli hücrelere ait 10 özelliğin kullanıldığı veri setine kaggle.com adlı internet sitesinden ulaşılmıştır (Kodlama, 2021). Kullanılan kanser veri setinde belirtilen 10 özellik şu şekildedir:

* **Örnek kod numarası**: id numarası
* **Küme Kalınlığı**: 1-10
* **Hücre Boyutunun Tekdüzeliği**: 1-10
* **Hücre Şekli Tekdüzeliği**: 1-10
* **Marjinal Yapışma**: 1-10
* **Tek Epitel Hücre Boyutu**: 1- 10
* **Çıplak Çekirdek:** 1-10
* **Yumuşak Kromatin:** 1-10
* **Normal Nükleol**: 1-10
* **Mitoz:** 1-10
* **Öngörülen sınıf:** 2 =iyi huylu ve 4= kötü huylu

## **Veri Setine Genel Bir Bakış**

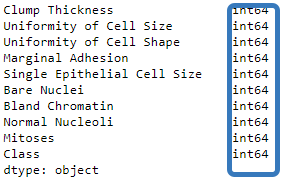
Meme kanserine ait veri setinde kullanılmak üzere 10 farklı özellik bulunmaktadır. Bu 10 özelliğin her birinden toplamda 683 veri toplanmıştır. Hazırlanan bu verilerin tamamı sayısal değer olarak tutulmaktadır ve bütün veri setine bakıldığı zaman kayıp verilerin olmadığı görülmektedir (Şekil-1). Bu yüzden herhangi bir veri tipi dönüşümü yapılmasına ve eksik verilerin giderilmesi için çalışma yapılmasına ihtiyaç duyulmamaktadır.



Şekil 1 Data sete ait bilgiler

## **Veri Hazırlık Aşaması**

Makine Öğrenmesi uygulamalarında sağlıklı bir model oluşturabilmek için kullanılacak veri setinde bazı öz hazırlıklar yapılması gerekmektedir. Meme kanseri ile ilgili veri setine bakıldığı zaman   
683 veri arasında kayıp veri bulunmamaktadır. Makine Öğrenmesinde model oluşturulurken sayısal değerlerin kullanılması gerekmektedir bu yüzden veri setlerindeki obje türündeki değerler sayısal verilere dönüştürülmesi gerekmektedir. Bu çalışmada kullanılacak olan veri setine bakıldığı zaman bütün veriler int olarak saklandığı görülmekte (Şekil-2) bu yüzden herhangi bir işleme gerek duyulmamaktadır.



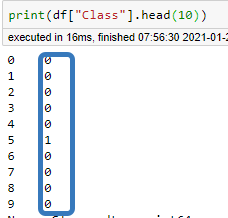
Şekil 2 Veri setine ait veri tipleri

Makine öğrenmesi algoritmalarında gereksiz verilerin sistemi olumsuz olarak etkilememesi için kullanılmayacak olan verilerin veri setinden kaldırılması sağlanmalıdır. Veri setine bakıldığı zaman alınan verilerin id bilgilerini saklayan bir sütun görülmektedir (Şekil-3). Bu bilgi, model içerisinde kullanıldığında sistemi yanlış yönlendirecektir bu yüzden id bilgisinin bulunduğu sütun veri setinden kaldırılmıştır.



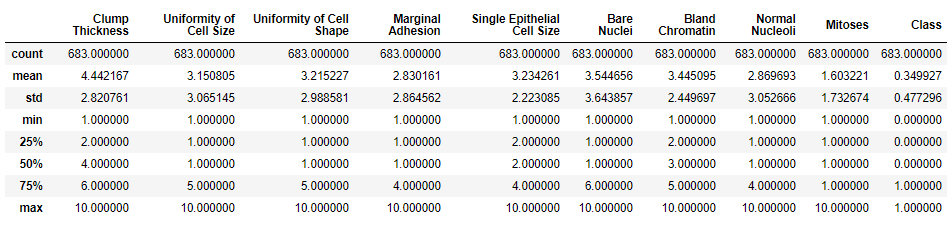
Şekil 3 Id bilgisini tutan değiken

Bu veri setinde yapılan diğer bir işlem ise tahmin etmeye çalıştığımız yani hedef sütununda kullanılan veriler için olmuştur. “Class” adındaki bu sütunda benign değerleri temsil etmek için 2 değeri ve malignant değerlerini belirtmek için ise 4 değeri kullanılmaktadır. Bu tür iyi huylu ve kötü huylu verilerin 0 ve 1 şeklinde yaygın kullanımı vardır. Bu yüzden “Class” sütunundaki 2 ve 4 değerleri 0 ve 1 olarak değiştirilmiştir (Şekil-4).



Şekil 4 Class adındaki sütun

Bu tip çalışmalarda model oluşturulurken aşırı verilerin veri setinde bulunmamasına veya düzeltilmesine özen gösterilmelidir. Meme kanserine ait bu veri setinin genel özelliklerine bakıldığında, tahmin edilecek olan hedef sütun dışındaki bütün veriler 0 ile 10 değerleri arasında olduğu görülmektedir. Bu sütunların max ve min değerlerine bakıldığında sütunda aşırı verilerin olmadığı görülmektedir (Şekil-5).



Şekil 5 Veri setinin genel özellikleri

# **Çok Katmanlı Algılayıcılar (MLP) ile Model Oluşturma**

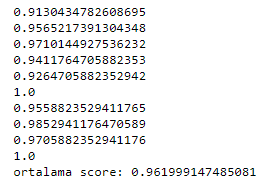
Makine Öğrenmesi yöntemlerinde model oluşturmak için iki seçenek bulunmaktadır. Birinci seçenek gizli katman kullanmadan model oluşturmak, ikinci seçenek ise modele ara katmanlar ekleyerek modeli geliştirmek. Yapılan bu çalışmada sisteme gizli katmanlar eklenerek modelin başarısı daha yükseğe çıkarılmaya çalışılmıştır.

Model oluşturulmadan önce verilerin düzenli bir dağılıma sahip olması amacıyla normalizasyon işlemi gerçekleştirilmiştir. MLP modeli oluştururken modele ait hangi parametrenin daha iyi olduğunu öğrenebilmek amacıyla Grid Search işlemi yapılmıştır. Yapılan işlemde sisteme aktivasyon fonksiyonu olarak: "relu","logistic","tanh","identity” seçenekleri, alpha değeri için: 0.1,0.01,0.001,0.0001 seçenekleri, gizli katman için (10,10), (3,3,3), (100,100), (3,5,3) seçenekleri ve solver olarak: "lbfgs","adam","sgd" seçenekleri sunulmuştur. Bu seçenekler sistem tarafından değerlendirilerek en iyi accuricy oranını veren değerler belirlenmiştir. Bu veri seti için en iyi parametreler Şekil-6’ da gösterilmektedir.



Şekil 6 MLP için en iyi parametreler

Yapılan Grid Search işleminden sonra modelde ezberlemeyi veya öğrenememeyi ortadan kaldırmak için K-Fold işlemi yapılmıştır. Bu işlem yapılırken Grid Search ile elde edilen parametreler kullanılmıştır. K-Fold işlemi için veri seti 10 farklı parçaya bölünmüş ve sonuçlar bu şekilde üretilmiştir. Bu işlemlerin sonucunda başarı oranı %96 olarak elde edilmiştir (Şekil-7).

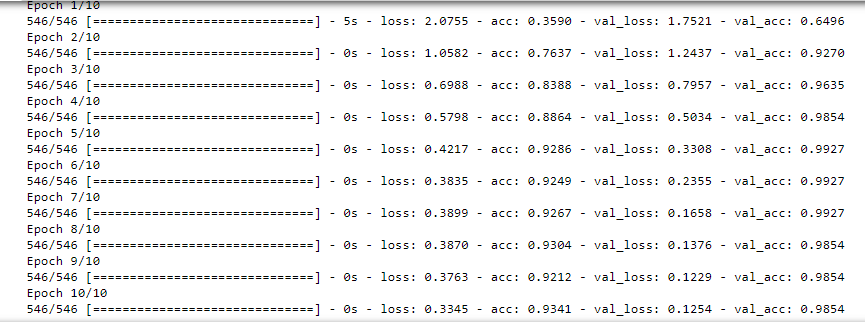


Şekil 7 Accuricy oranları

# **Keras ile Model Oluşturma**

Çok katmanlı algılayıcılar ile çalışılırken belirlenen gizli katmanlara erişilememekte ve her gizli katmana ait farklı özellikler seçilememektedir. Bu problemi kaldırmak amacıyla model oluşturmak için Keras kütüphanesi kullanılmaktadır. Kullanılan bu veri setiyle Keras aracılığıyla bir model oluşturulmuştur. Bu modelde 5 farklı katman belirlenmiş ve bu katmanlardaki nöron sayısı 15, 30, 60, 120 ve 240 olarak belirlenmiştir.

Modelin kaç defa eğitileceği (epoch) CPU ile çalışılırken 10 olarak belirlenmiştir. Optimizer olarak “adam”, loss fonksiyonu olarak ise “spars categorical crossentropi ” kullanılmıştır. Model eğitilmiş ve 10. eğitimde %93 accuricy oranı elde edilmiştir.

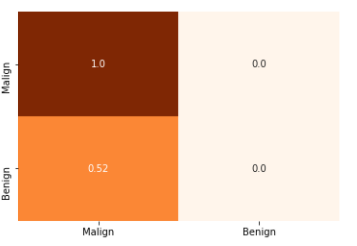


Şekil 8 Keras eğitim sonuçları

# **Değerlendirme**

Meme kanserinde iyi huylu veya kötü huylu kanserin tespiti için oluşturulan YSA modelinde tahmin başarısı %65 olarak ortaya çıkmaktadır. Bu veriden yola çıkarak bu başarı oranı ile bu modelin biyomedikal alanda kullanılması uygun olmadığı görülmektedir.

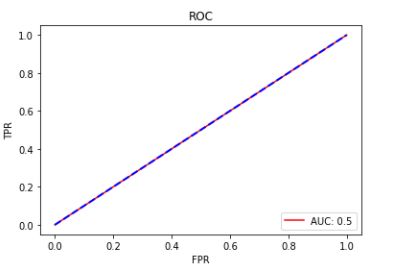
Çalışmaya ait confusion matrix değerlendirildiğinde, modelin iyi huylu kanseri kötü huylu kansere göre daha iyi bir skorla tahmin ettiği görülmektedir. Fakat kötü huylu tümörün tahmininde başarısız olduğu ortaya çıkmaktadır (Şekil-9).



Şekil 9 Confusion Matrix

F1 skoruna bakıldığı zaman 0 olarak karşımıza çıkmaktadır. Bunun sebebi, modelin kötü huylu kanseri tahmin edememesinden kaynaklanmaktadır.

Bu çalışmaya ait auc ve roc eğrisine bakacak olursak sistemin %50 oranında başarışı olduğu görülmektedir (Şekil-10). Veri seti biraz daha genişletilir ve ön hazırlık aşaması daha titiz bir şekilde oluşturulursa bu oranlar daha da yukarıya çekilebilir.



Şekil 10 Roc ve Auc