Abd’ye ait kanser verilerinin regresyon yöntemleri ile tahmini

Ömer Faruk Yıldız

Biyomedikal Mühendisliği- Yüksek Lisans

2028142006

BMM728-Biyomedikal Mühendisliğinde Yapay Sinir Ağı Uygulamaları

İçindekiler

[**1.** **Giriş** 2](#_Toc58799501)

[**2.** **Kanser Veri Setine Genel Bir Bakış** 2](#_Toc58799502)

[**3.** **Özellik Seçimi** 3](#_Toc58799503)

[**4.** **Veri Seti Temizliği ve Dönüşümü** 4](#_Toc58799504)

[**5.** **Uygulama** 5](#_Toc58799505)

[**6.** **Sonuç** 6](#_Toc58799506)

# **Giriş**

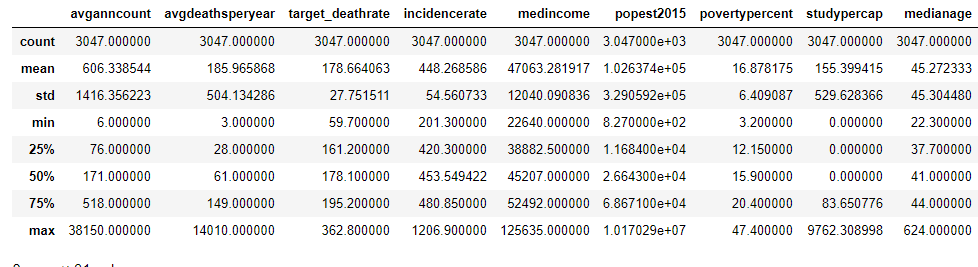
Yapılan bu çalışmada Amerika Birleşik Devleti’ndeki eyaletlerden kanserli bireylere ait bilgiler doğrultusunda makine öğrenmesi regresyon yöntemleri kullanılarak gelecek yıllara ait kanserli birey sayıları tahmin edilmeye çalışılmıştır. Tahmin yapılırken farklı algoritmalar kullanılmış ve bu algoritmalar belirli yöntemler aracılığıyla karşılaştırılmıştır.

# **Kanser Veri Setine Genel Bir Bakış**

Kullanılan bu veri seti 3047 satır ve 33 sütundan oluşmaktadır. Bu veri setindeki sütunların neleri ifade ettiği aşağıda maddeler halinde açıklanmıştır.

* **TARGET\_deathRate:** Bağımlı değişken. Kişi başına ortalama (100.000) kanser ölümleri (a)
* **avgAnnCount:** Yıllık olarak teşhis edilen kanser vakalarının ortalama sayısı (a)
* **avgDeathsPerYear:** Kansere bağlı olarak bildirilen ortalama ölüm sayısı (a)
* **incidenceRate:** Kişi başına ortalama (100.000) kanser teşhisi (a)
* **medianIncome:** İlçe başına ortalama gelir (b)
* **popEst2015:** İlçe nüfusu (b)
* **povertyPercent:** Yoksulluktaki nüfus yüzdesi (b)
* **studyPerCap:** *İlçe başına düşen kanserle ilgili klinik araştırma sayısı (a)*
* **binnedInc:** Kişi başına ortalama gelir ondalık dilim (b)
* **MedianAge:** İlçe sakinlerinin ortalama yaşı (b)
* **MedianAgeMale:** İlçe sakinlerinin erkek ortalama yaşı (b)
* **MedianAgeFemale:** İlçe sakinlerinin kadın ortalama yaşı (b)
* **Geography:** İlçe adı (b)
* **AvgHouseholdSize:** İlçenin ortalama hane büyüklüğü (b)
* **PercentMarried:** İlçe sakinlerinin evli olan yüzdesi (b)
* **PctNoHS18\_24:** İlçe sakinlerinin 18-24 yaşları arasındaki en yüksek eğitime ulaşılan yüzdesi: liseden az (b)
* **PctHS18\_24:** İlçe sakinlerinin 18-24 yaşları arasındaki en yüksek eğitime ulaşılan yüzdesi: lise diploması (b)
* **PctSomeCol18\_24:** İlçe sakinlerinin 18-24 yaşları arasındaki en yüksek eğitime ulaşılan yüzdesi: bazı üniversiteler (b)
* **PctBachDeg18\_24:** İlçe sakinlerinin 18-24 yaşları arasındaki en yüksek eğitime ulaşılan yüzdesi: lisans derecesi (b)
* **PctHS25\_Over:** İlçe sakinlerinin 25 yaş ve üzerindeki en yüksek eğitime sahip olan yüzdesi: lise diploması (b)
* **PctBachDeg25\_Over:** İlçe sakinlerinin 25 yaş ve üzerindeki en yüksek eğitime ulaşılan yüzdesi: lisans derecesi (b)
* **PctEmployed16\_Over:** 16 yaş ve üzerindeki ilçe sakinlerinin yüzdesi (b)
* **PctUnemployed16\_Over:** 16 yaş ve üstü işsizlerin yüzdesi (b)
* **PctPrivateCoverage:** Özel sağlık sigortası olan ilçe sakinlerinin yüzdesi (b)
* **PctPrivateCoverageAlone:** Sadece özel sağlık sigortası olan ilçe sakinlerinin yüzdesi (kamu yardımı yok) (b)
* **PctEmpPrivCoverage:** Çalışan tarafından sağlanan özel sağlık sigortasına sahip ilçe sakinlerinin yüzdesi (b)
* **PctPublicCoverage:** Devlet tarafından sağlanan sağlık sigortasına sahip ilçe sakinlerinin yüzdesi (b)
* **PctPubliceCoverageAlone:** Yalnızca devlet tarafından sağlanan sağlık sigortasına sahip ilçe sakinlerinin yüzdesi (b)
* **PctWhite:** Beyaz olarak tanımlanan ilçe sakinlerinin yüzdesi (b)
* **PctBlack:** Siyah olarak tanımlanan ilçe sakinlerinin yüzdesi (b)
* **PctAsian:** Asyalı olarak tanımlanan ilçe sakinlerinin yüzdesi (b)
* **PctOtherRace:** Beyaz, Siyah veya Asyalı olmayan bir kategoriyi belirleyen ilçe sakinlerinin yüzdesi (b)
* **PctMarriedHouseholds:** Evli hanelerin yüzdesi (b)
* **BirthRate:** İlçedeki kadın sayısına göre canlı doğum sayısı (b)

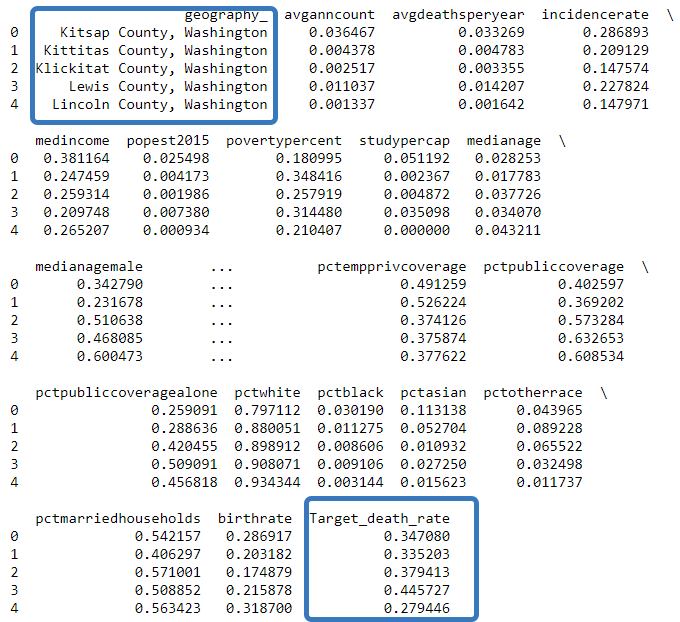
Bu veri setinin veri tiplerine bakıldığı zaman iki sütundaki değerlerin string türünde geriye kalan verilerin ise sayısal halde tanımlandığı görülmüştür. Bu veri setine ait genel bilgiler aşağıdaki tabloda gösterilmektedir.



Şekil 1 Veri setine ait genel bilgiler

# **Özellik Seçimi**

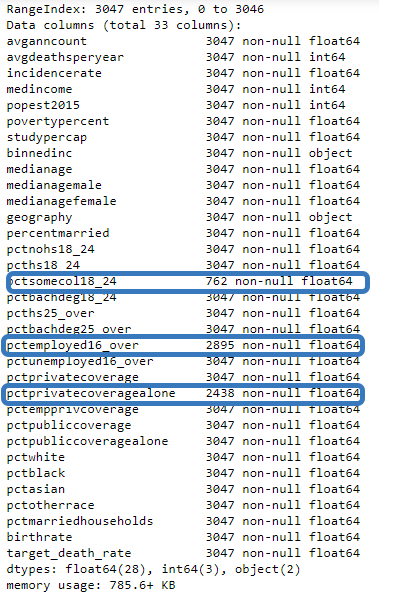
Veri seti incelendiğinde bazı verilerin yerlerinin değişmesinin tahmin yaparken indekslemede kolaylık sağlayacağı fark edilmiştir. Bu verilerden ilki ilçe adının tutulduğu “**Geography**” sütunudur. Bu sütun ilk başta veri setinin ortasında bulunmaktadır ve object veri tipi olarak tutulmaktadır. Bu sütun sayısal bir veri dönüştürülemeyeceğinden dolayı ileri aşamalarda indeksleme işlemi yapılırken veriyi bölmemek için veri setinin başına alınmıştır. Yeri değişmesinin fayda sağlayacağı düşünülen diğer bir sütun ise tahminin yapılacağı yani y sütunu olarak tutulacak olan “**TARGET\_deathRate**” verisidir. Bu sütun veri setinin son kısmına alınarak veri seti daha okunabilir bir hale getirilmiştir.



Şekil 2 Veri setinde sütunların yerinin değişmiş hali

# **Veri Seti Temizliği ve Dönüşümü**

Kullanılan veri seti info() komutu ile incelendiğinde bazı verilerin kayıp veriler olarak geldiği görülmüştür. Bu kayıp verilerin bazıları sütunun ortalaması ile doldurulmuş bazıları ise veri setini olumsuz etkileyebileceği düşünüldüğünden silinmiştir.



Veri setine bakıldığında “**PctSomeCol18\_24**” sütununda 2285 verinin kayıp olarak geldiği görülmüştür. Bu kayıp verilerin yerinin tek bir sabit veri ile doldurulması çok sayıda tekrarlı veriye yol açacağından bu veri sütunu silinmiştir.

Veri setinde “**pctsomecol18\_24**” sütununa bakıldığında 152 verinin kayıp veri olarak geldiği ve “**pctprivatecoveragealone**” sütununa bakıldığında 609 verinin eksik geldiği görülmüş ve bu sütunlardaki kayıp veriler sütunun ortalaması ile doldurulmuştur.

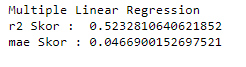
Kanser data setinde aşırı veri kontrolü yapılmak istendiğinde veri setinde birbirinden uzak verilerin oldukça fazla olduğu görülmüştür. Bu verilerin fazla olması anormal bir durum değil aksine veri setinin farklı değerleri bir arada bulundurabilmesinden kaynaklanmaktadır. Örneğin 16 yaş ve üstü işsizlerin yüzdesi oldukça düşük rakamlar alabilirken yüksek değerlere de ulaşabilir. Bu yüzden aşırı veri kontrolü sonucunda aşırı veri olmadığı kanaatine varılmıştır.

Makine öğrenmesi çalışmalarında regresyon analizleri yapılırken genelde veriler üzerine normalizasyon işlemi uygulanır. Bu çalışmada da verilerin daha sağlıklı sonuçlar vermesi amacıyla verileri 0-1 aralığına düşürerek normalizasyon uygulanmıştır.

# **Uygulama**

Kanser veri seti ile makine öğrenmesi çalışması kapsamında veriler multiple lineer regresyon, polinomal lineer regresyon, decision tree ve rondom forest algoritmaları yardımıyla skorlar elde edilmiştir.

Multiple lineer regresyon ile yapılan çalışmalarda r2 skoru ve Mean Absolute Error(MAE) hesaplanmıştır (Şekil-3). Skorların oldukça düşük olduğu görülmüş bu yüzden sonuçları iyileştirmek amacıyla train-test olarak veri seti bölünmüştür. Sonuçta skorlarda herhangi bir değişimin olmadığı gözlemlenmiştir.



Şekil 3 Multiple lineer regresyon r2 skoru ve mae skoru



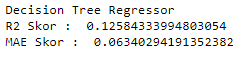
Şekil 4 Multiple lineer regresyon train-test ile r2 skoru ve mae skoru

Veri seti başka bir uygulama olarak decision tree algoritması ile eğitilmiştir. Bu veriler ilk aşamada herhangi bir bölünme yapılmadan test edilmiş ve r2 skoru ve MAE hesaplanmıştır (Şekil-5).



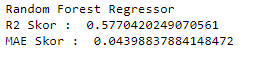
Şekil 5 Decision tree r2 ve MAE skoru

Bu sonuçlar analiz edildiğinde r2 skorun yüksek bir değerde olması algoritmanın oldukça başarılı olduğu görülmüş fakat eğitim sürecinde ezber şüphesini doğurmuştur. Bu yüzden veri seti eğitim ve test olarak ikiye bölünmüştür. Bu işlem sonucunda tekrar r2 skoru ve MAE hesaplandığında oldukça büyük bir düşüşün olduğu görülmektedir (Şekil-6). Bu sonuçlar önceki uygulamada bir ezber olduğunu ortaya koymaktadır.



Şekil 6 Decision tree train-test skoru

Kanser veri seti son olarak rondom forest algoritması ile denenmiştir. İlk olarak veriler herhangi bir ayrım işlemi olmadan denenmiştir. Bu işlem sonucunda sonuçlar oldukça başarılı çıkmıştır (Şekil-7).



Şekil 7 Rondom forest r2 skoru ve MAE

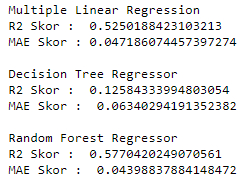
Çıkan sonuçlarda öğrenme aşamasında algoritmanın ezber yapıp yapmadığını görmek için veri seti test ve eğitim olarak iki parçaya ayrılmıştır. Algoritma sonucu olarak r2 skoru ve MAE hesaplanmış (Şekil-8) ve test ve eğitim olarak parçalanmadan yapılan işlemde sistemin ezber yaptığı görülmüştür.



Şekil 8 Rondom forest test-train ile r2 skoru ve MAE sonucu

# **Sonuç**

Yapılan çalışmada kanser veri seti multiple lineer regresyon, decision tree ve rondom forest algoritmaları ile analiz edilmiştir. Bu çalıma sonunda her yöntem için r2 skoru ve Mean Absolute Error(MAE) değeri hesaplanmıştır (Şekil-9).



Şekil 9 Sonuçların bir arada verildiği tablo

Bu sonuçlar doğrultusunda en iyi tahmin yapan algoritmanın rondom forest algoritması olduğu anlaşılmaktadır. En kötü tahmin yapan algoritmanın ise decision tree algoritması olduğu görülmektedir. Bu veri seti ile yapılan çalışmada en iyi sonuç veren algoritma ile yaklaşık %57 oranında başarı elde edilmiştir. Fakat bu oranın biyomedikal alanında güvenli bir sonuç olmadığı görülmektedir.