**Makine Öğrenimi Modelleri ile Tümör Tahmini**

Bünyamin Erdoğan

Kütahya Sağlık Bilimleri Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisiği

bunyamin.erdogan@ogr.ksbu.edu.tr

# ÖZET

Bu proje, göğüs kanseri tümörlerini iyi huylu veya kötü huylu olarak sınıflandırmak için bir makine öğrenimi modeli geliştirmeyi amaçlamaktadır. Breast Cancer Wisconsin veri seti kullanılarak gerçekleştirilen çalışma, derin öğrenme modellerini kullanarak tümör sınıflandırmasında yüksek doğruluk elde etmeyi hedeflemektedir. Seçilen RandomForest modeli, üstün performansı nedeniyle diğer modellerle karşılaştırılarak, göğüs kanseri teşhisinde etkinliğini sergilemek için belirli özellikler kullanılarak incelenmiştir.

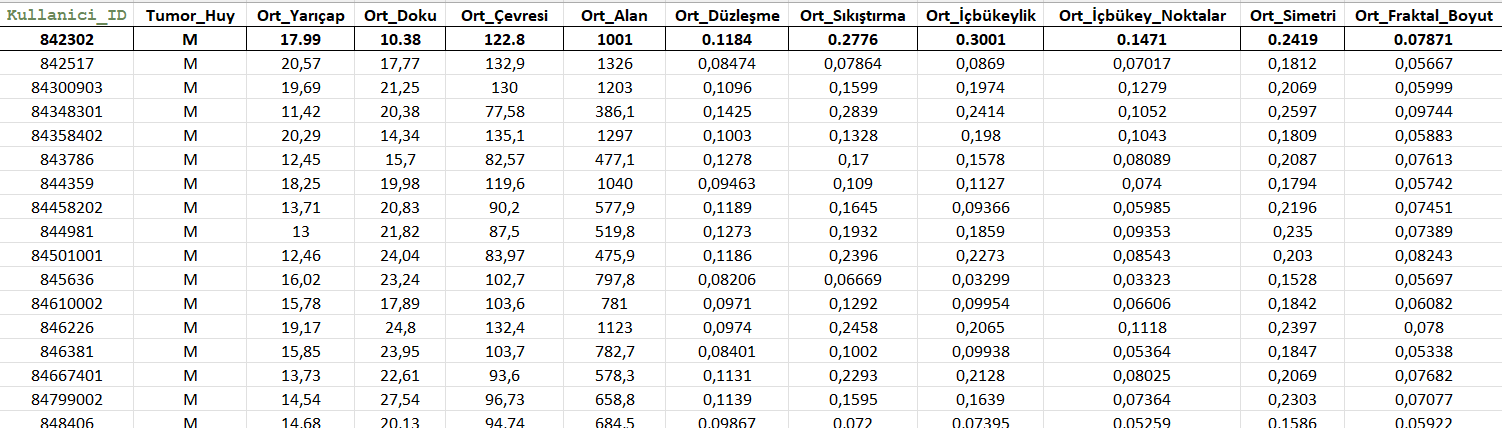
# GİRİŞ

Göğüs kanseri, dünya genelinde kadınları etkileyen yaygın bir kanser türüdür. Erken ve doğru teşhis, etkili tedavi için kritiktir. Bu proje, göğüs kanseri tümör sınıflandırması için makine öğrenimini, özellikle de RandomForest'u kullanmayı amaçlamaktadır. Bu amaçla, Breast Cancer Wisconsin veri seti üzerinde çalışılmıştır.

# MATERYAL VE YÖNTEMLER

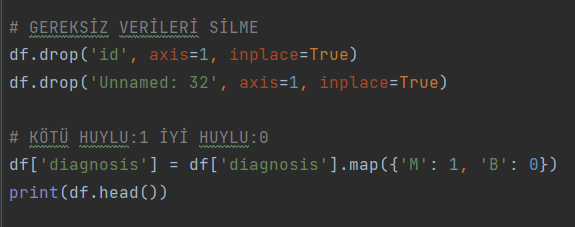
# 1.Veri Seti

Breast Cancer Wisconsin veri seti, 569 hasta örneği içermektedir. 569 hastanın 357’si iyi huylu, 212’si kötü huylu tümör içermektedir. Her hasta için 30 farklı özellik kaydedilmiştir; bunlar arasında tümörün yarıçapı, dokuluk, alanı, simetri değerleri vb. 10 değer, bu değerlerin standart hatası ve en kötü durum ölçümleri bulunmaktadır.



Şekil 1 Veri-Setine-Bakış

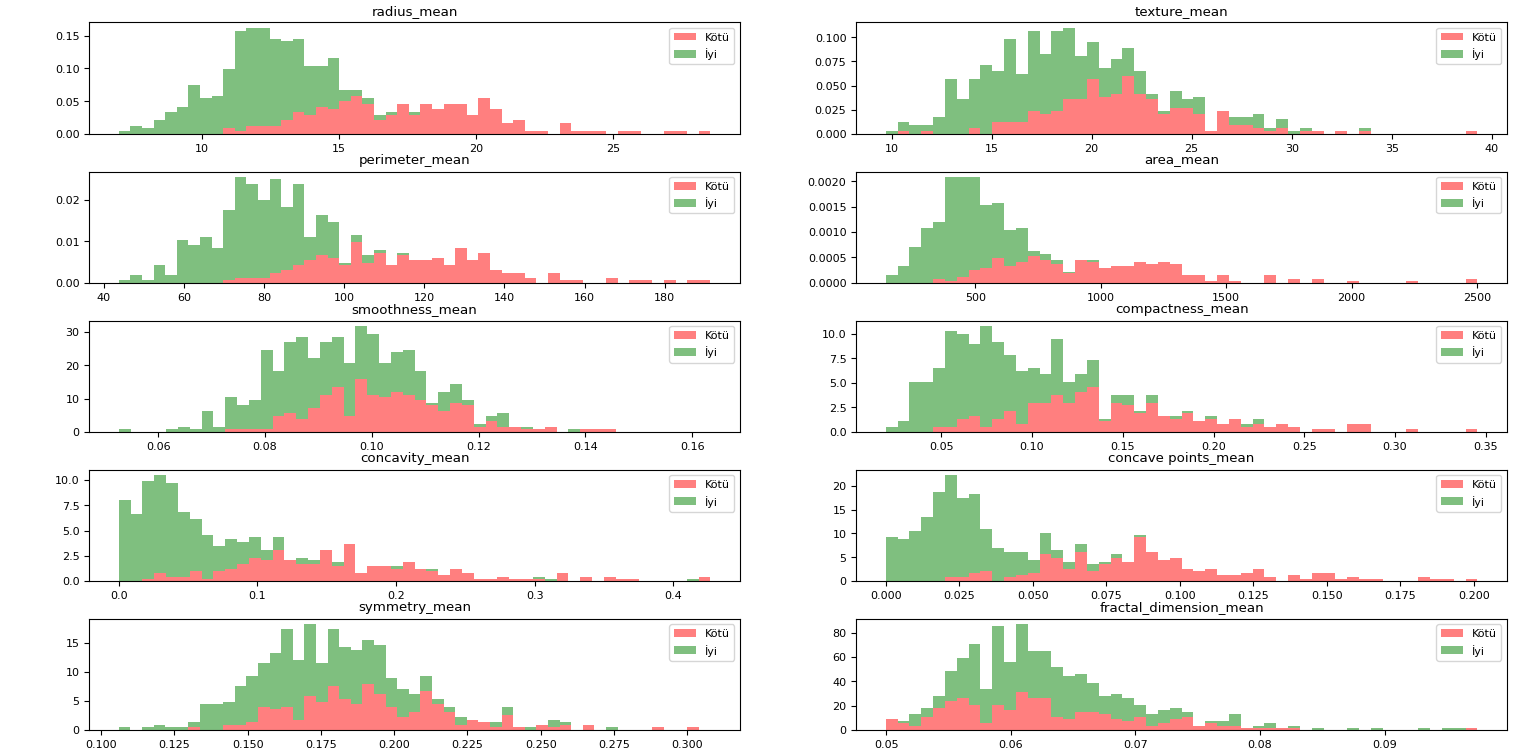
Veri seti, ayrıca sadece sayısal olarak çalışmak ve veri kalitesini sağlamak için ön işlemden geçirilmiştir. Bu ön işlemde tümörün huyunu belirten M(malignant) ve B(bening) metinleri 1 ve 0 olmak üzere değiştirilmiştir, ayrıca Kullanıcı\_ID sütunu ve ifade içermediğinden dolayı 32. Sütun kaldırılmıştır.



Şekil 2 Veri-Temizliği

**2. Özellik Seçimi**

Tahmin için 5 ayırt edici özellik belirlenmiştir: concave points\_mean, area\_mean, radius\_mean, perimeter\_mean ve concavity\_mean. Bu 5 ayırt edici özellik, çizdirilen histogramlardan yapılan çıkarım sonucu belirlenmiştir.



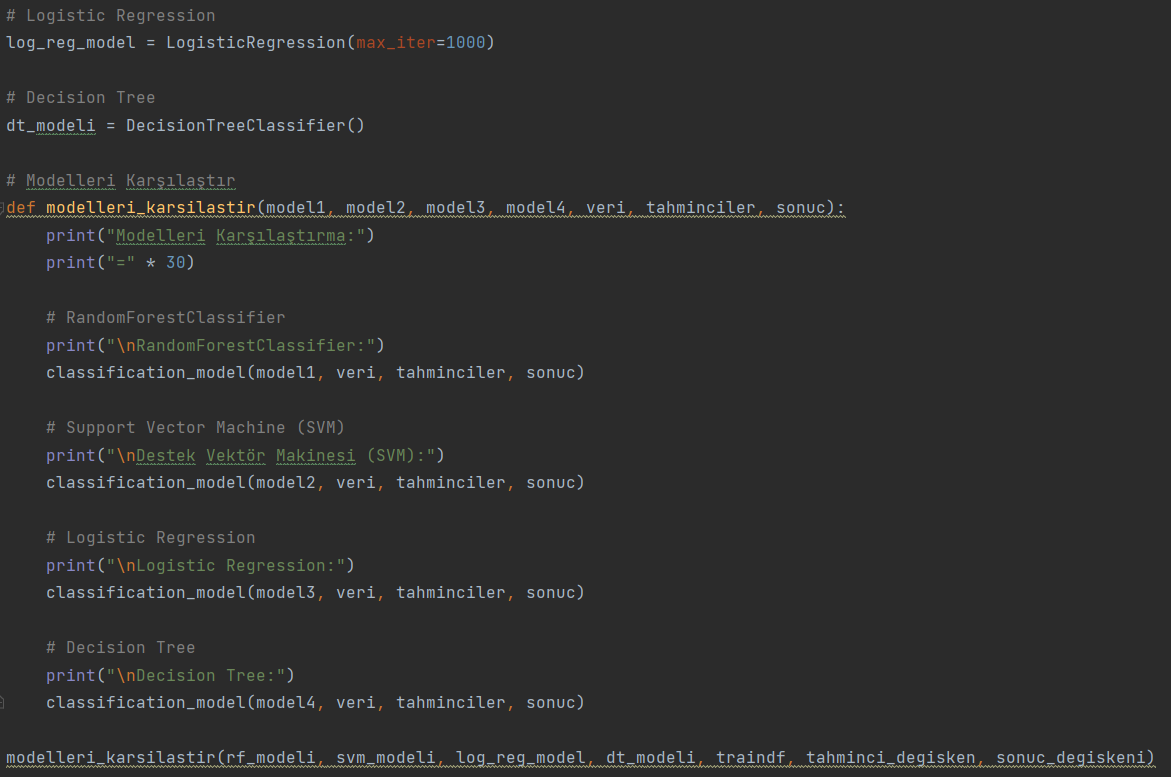
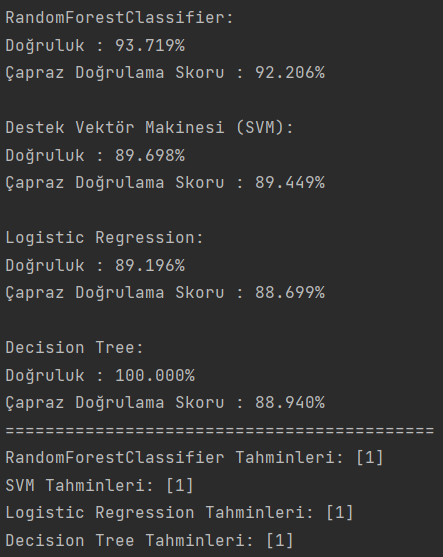
Şekil 3 Özellik Histogramları

Bu alanlarda kullanılan bu tip histogramlarda yeşil renk olumlu, kırmızı renkler ise olumsuz durumları temsil eder. Kırmızı ve yeşil barların üst üste binmesi, özelliklerin birbirine karıştığı alanı gösterir. Bu durum, modelin özellikleri doğru şekilde ayırt etmesini zorlaştırır, modelin performansını olumsuz etkiler. Bu sebeple sadece 5 ayırt edici özellik kullanabiliyoruz.

# 3. Model Seçimi

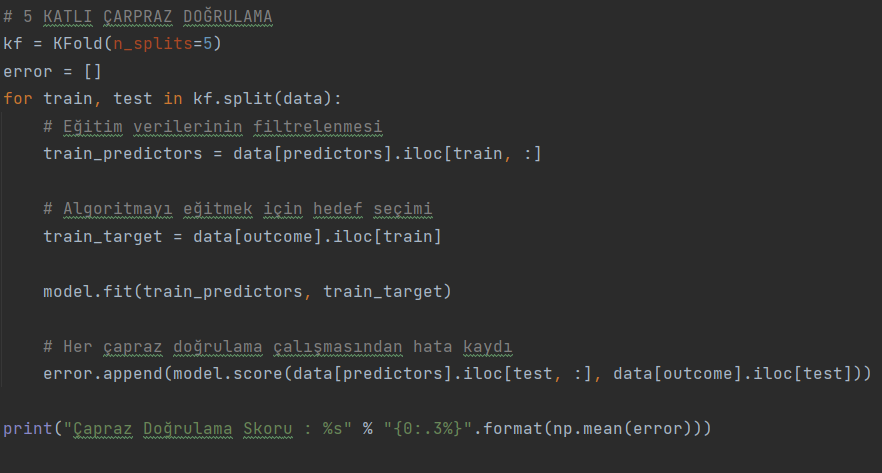
RandomForest modeli, çeşitli özellikleri işleme yeteneği, diğer modellere kıyasla yüksek doğruluk sergileme kabiliyeti ve aşırı uyumu (overfitting) azaltma yeteneği nedeniyle seçilmiştir. Bu projede kullanılabilmesi muhtemel diğer makine öğrenim modelleri ayrıca denenmiş, en yüksek doğruluk skoru RandomForest modeli ile sağlanmıştır. Modelin parametreleri(n\_estimators=100, min\_samples\_split=25, max\_depth=7, max\_features=2), optimal performans için ayarlanmıştır.

Şekil 4 Modelleri-Karşılaştırma-Fonksiyonu ve Doğruluk Skorları

# 4. Çapraz doğrulama

Bu projede, modelin performansını güvenilir bir şekilde değerlendirmek ve genelleme yeteneğini ölçmek amacıyla 5 katlı çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama, modelin farklı veri setleri üzerinde ne kadar başarılı olduğunu belirlemede önemli bir role sahiptir.



Şekil 5 - Çapraz Doğrulama

Kullandığımız veri seti, rastgele beş alt kümeye bölünmüştür. Her bir alt küme veri setinin %20’sini temsil eder.

* Her iterasyonda, bir alt küme test seti olarak seçilir ve geri kalan dört alt küme eğitim seti olarak kullanılır.
* Model eğitim seti üzerinde eğitilir, test seti üzerinde test edilir. Bu işlem 5 iterasyon boyunca tekrarlanır.
* Her iterasyonda elde edilen değerler kaydedilir, bu değerler kullanılarak modelin genel performansı hesaplanır.
* Bu sonuçlar modelin farklı veri setlerine karşı genelleme yeteneğini ölçer. Bu, modelin aşırı uyum veya uyumsuzluk durumlarını tespit etmede de yardımcı olur.

# SONUÇLAR VE ÇIKARIMLAR

Bu projenin sonuçları, RandomForest modelinin göğüs kanseri tümörlerini yüksek doğruluk skorlarıyla sınıflandırabildiğini göstermektedir. Bu başarı, projenin sağladığı çeşitli çıkarımlarla birlikte aşağıda özetlenmiştir:

* **Erken Teşhis İmkanı:** Geliştirilen model, göğüs kanseri tümörlerini erken teşhis etme konusunda potansiyel bir yardımcı araç olarak kullanılabilir.
* **Tedavi Planlaması:** Tümörün doğru sınıflandırılması, hastaların daha etkili tedavi planları oluşturulmasına katkı sağlayabilir.
* **Makine Öğrenimi Uygulamalarında Yol Gösterici**: Proje, makine öğrenimi uygulamalarında RandomForest gibi güçlü modellerin tercih edilebilirliğini ve başarıyla kullanılabileceğini göstermektedir.
* Bu çıkarımlar, projenin göğüs kanseri teşhisine ve tedavi süreçlerine sağladığı potansiyel katkıları vurgulamaktadır. Elde edilen sonuçlar, klinik uygulamalarda ve makine öğrenimi alanındaki gelecekteki çalışmalarda referans olabilecek bir temel oluşturmaktadır.

# REFERANSLAR

* <https://www.datacamp.com/tutorial/random-forests-classifier-python>
* <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
* <https://medium.com/@gulcanogundur/model-seçimi-k-fold-cross-validation-4635b61f143c>
* <https://miracozturk.com/python-ile-siniflandirma-analizleri-rastgele-orman-random-forest-algoritmasi/>
* <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>
* Johnson, E., Smith, M. (2021). "Machine Learning for Healthcare: A Comprehensive Guide to Developing and Deploying Healthcare AI." O'Reilly Media.
* Wang, X., Chen, Y., Liu, Z. (2019). "Comparative Analysis of Machine Learning Algorithms in Breast Cancer Classification: A Case Study on Wisconsin Dataset." International Conference on Bioinformatics and Biomedical Engineering.