Hacettepe Üniversitesi
YAPAY ZEKA
Topluluğu

Göğüs Kanseri Sınıflandırması
Proje Raporu

Hazırlayan
Nurşah Satılmış

İçindekiler

1.Giriş

Meme kanseri, dünya genelinde kadınlarda en sık görülen kanser türlerinden biridir. Erken teşhis ve doğru sınıflandırma, tedavi başarısını artırabilir ve hastalığın ilerlemesini engelleyebilir. Bu çalışmada, meme kanseri sınıflandırması için makine öğrenimi tekniklerini kullanarak, Wisconsin Üniversitesi'nde toplanan orijinal veri seti üzerinde çalıştık.

Veri setimiz, meme hücrelerinin çeşitli özelliklerini tanımlayan bir dizi sayısal özellikten oluşmaktadır. Bu özellikler kullanılarak hücrelerin iyi huylu (iyi) veya kötü huylu (kötü) olarak sınıflandırılması için bir model oluşturmayı amaçladık.

Çalışmamızda, veri setini keşfettik, ön işleme adımları gerçekleştirdik ve farklı makine öğrenimi algoritmalarını kullanarak modeller oluşturduk. Oluşturduğumuz modelleri değerlendirerek, meme kanseri sınıflandırması için etkili bir model geliştirmeyi hedefledik.

2.Veri Setinin Araştırılması

1990'ların başlarında University of Wisconsin Hastanelerinden Dr. William H. Wolberg tarafından elde edilen ve tıbbi teşhisler için kullanılan bir veri seti olan Wisconsin Breast Cancer Database, meme kanseri teşhisi için kullanılan özellikleri içermektedir. Bu veri seti, 699 örnekten oluşmaktadır ve örnekler benign ve malign olmak üzere iki sınıfa ayrılmıştır. Veri seti, William Wolberg tarafından 14 Temmuz 1992'de bağışlanmıştır.

Veri setinin kronolojik olarak gruplandırılması, örneklerin klinik vakaların rapor edilmesine dayandığını gösterir, bu da veri toplama sürecinin titiz ve sistematik olduğunu düşündürebilir. Veri seti, Creative Commons Attribution 4.0 International (CC BY 4.0) lisansına sahiptir.

Bu veri seti, bilimsel araştırmalarda da sıkça kullanılmaktadır. Örneğin, 2019 yılında Cancer Information dergisinde yayınlanan "On the Bias of Precision Estimation Under Separate Sampling" makalesinde, Data Science Journal'da yayınlanan "A Privacy-Preserving Data Mining Method Based on Singular Value Decomposition and Independent Component Analysis" makalesinde ve çeşitli sempozyum ve konferans makalelerinde kullanılmıştır. Bu çalışmalar, veri setinin geniş bir akademik ve bilimsel topluluk tarafından değerli bulunduğunu göstermektedir. Bu bilgiler ışığında veri setini güvenilir kabul edebiliriz.

3. Keşifçi Veri Analizi

Tahminci değişkenlerin kategorik değişken olup olmadığına bakmak için .info() ile değişkenlerin veri tipleri incelendi kategorik değişkene rastlanmadı. Tahmin edici değişkenler için :

3.1 Eksik Veri incelemesi

Eksik Veri incelemesi yapıldı , "Bare_nuclei" değişkeninde 16 eksik veri tespit edildi. Bu değişkeni rastgele değerle doldurma işlemine uygun olup olmadığına bakılması için görselleştirme yapıldı, eksik değerlere rastgele değerler atandı.

3.2 Aykırı değer analizi

Aykırı değer analizi yapıldı , data seti çok büyük olmadığı için shapiro-Wilk testi ile verinin normal dağılımlı olup olmadığına bakılarak Z-score veya IQR aykırı değer tespit yöntemleri ile aykırı değerler saptandı. Clipping yöntemi ile veri aykırı değerlerden arındırıldı.

3.3 Normalizasyon

Algoritmaların daha iyi çalışması için özelliklerin benzer ölçeklerde olmasını sağlamak, veri dağılımını düzeltmek ve bazı optimizasyon algoritmalarının daha hızlı ve daha iyi ilerlemesine yardımcı olmak amacıyla tahminciler için normalizasyon uygulandı.

Hedef Değişken için ;

- ✓ Eksik veriler kontrol edildi , eksik veriye rastlanmadı.
- ✓ "2" ve "4" değerlerinden oluştuğu gözlemlendi. Veri seti bilgilerinde 2= iyi huylu , 4= kötü huylu olarak verilmişti.
- ✓ 458 iyi huylu , 241 kötü huylu değer içerdiği gözlemlendi
- ✓ Genellikle, makine öğrenimi modelleri, sınıflandırma problemleri için etiketleri 1 ve 0 olarak kullanmayı tercih ettiğinden 2 ve 4 değerleri 1 ve 0 a dönüştürüldü.

3.4 Veri Görselleştirme

Veriyi daha iyi anlamak için veri hedef değişkene göre histogram kullanılarak görselleştirildi.

Heatmap ile tahmincilerin hedef değişken ile korolasyonları incelendi.

"Mitoses" değişkeni hariç diğer tshminci değişekenler ve hedef değişken arasında %70 ve üzeri korolasyon gözlemlendi.

4. Modellerin Kıyaslanması

Lojistik Regresyon:

- Çalışma Prensibi: Binary sınıflandırma için kullanılır. Bir doğrusal regresyon modeline benzer, ancak çıktıyı bir logit dönüşümü ile 0 ile 1 arasında sıkıştırır.
- Avantajları: Basit, yorumlanabilir, hızlı eğitim ve tahmin süreçleri sağlar.
- Dezavantajları: Düz çizgilerle sınırlı kalır, çok karmaşık ilişkilerde başarılı olamayabilir.

Random Forest:

- Çalışma Prensibi: Karar ağaçları (decision trees) temelli bir yöntemdir. Birçok karar ağacı oluşturarak her bir ağacın tahminini alır ve sonuçları birleştirir.
- Avantajları: Karmaşık ilişkileri öğrenebilir, veri üzerindeki gürültüye dayanıklıdır, ağaçları paralel olarak eğitebilir.
- Dezavantajları: Aşırı uyuma (overfitting) eğilimli olabilir, büyük veri kümelerinde eğitim süresi uzun olabilir.

Support Vector Classification (SVC):

- Çalışma Prensibi: Sınıflar arasındaki en geniş boşluğu (margin) bulmaya çalışır. Bu geniş boşluk, sınıfları birbirinden en iyi şekilde ayıran karar sınırlarını belirler.
- Avantajları: Veri boyutu (dimensionality) yüksek olduğunda iyi çalışabilir, aşırı uyuma karşı dirençlidir, farklı çekirdek fonksiyonları kullanarak esneklik sağlar.
- Dezavantajları: Büyük veri kümelerinde eğitim süresi uzun olabilir, bazı durumlarda çekirdek fonksiyonunun seçimi zor olabilir.

XGBoost:

- Çalışma Prensibi: Gradient Boosting algoritmasının bir uygulamasıdır. Zayıf öğrenicileri (genellikle karar ağaçları) bir araya getirerek güçlü bir model oluşturur.
- Avantajları: Yüksek performanslı, genellikle diğer modellere göre daha iyi sonuçlar verir, aşırı uyuma karşı dirençlidir.
- Dezavantajları: Hiperparametre ayarına duyarlı olabilir, eğitim süresi diğer modellere göre daha uzun olabilir.

CatBoost:

- Çalışma Prensibi: Kategorik değişkenleri doğrudan ele alarak, kategorik değişkenleri otomatik olarak kodlar ve bu kodlama işlemini öğrenme sürecine dahil eder.
- Avantajları: Kategorik değişkenleri kolayca işleyebilir, genellikle iyi bir genelleştirme yeteneğine sahiptir.
- Dezavantajları: Diğer algoritmalara göre eğitim süresi daha uzun olabilir, hiperparametre ayarına dikkat edilmesi gerekebilir.

Model	Accuracy	Precision	Recall	ROC AUC Score
XGBoost Classification	0.95	0.95	0.95	0,9877
Logistic Regression	0.9785	0.98	0.98	0.9914
Random Forest Classification	0.9642	0.96	0.96	0.9906
Support Vector Classification	0.9642	0.96	0.96	0.9911
Neural Network Classification	0.9285	0.94	0.93	0.9091
CATBoost Classificaiton	0.9714	0.97	0.97	0.9627

Bu çalışmada, farklı sınıflandırma modellerinin performansını doğruluk (Accuracy), hassasiyet (Precision), geri çağırma (Recall) ve ROC AUC puanı üzerinden değerlendirdik

Doğruluk (Accuracy): Doğru olarak sınıflandırılan örneklerin toplam örnek sayısına oranıdır. Yüksek doğruluk, modelin genel olarak doğru tahminlerde bulunduğunu gösterir. Ancak, dengesiz sınıflar (imbalance) durumunda doğruluk yanıltıcı olabilir.

Hassasiyet (Precision): Pozitif olarak tahmin edilen örneklerin gerçekten pozitif olma oranıdır. Yüksek hassasiyet, modelin yanlış pozitif tahminler yapma olasılığının düşük olduğunu gösterir. Özellikle yanlış pozitiflerin maliyeti yüksek olduğunda önemlidir.

Geri Çağırma (Recall): Gerçekten pozitif olan örneklerin ne kadarının doğru bir şekilde pozitif olarak tahmin edildiğini gösterir. Yüksek geri çağırma, modelin pozitifleri kaçırmama yeteneğini gösterir. Özellikle yanlış negatiflerin (gerçek pozitifleri kaçırma) maliyeti yüksek olduğunda önemlidir.

ROC AUC Puanı: ROC eğrisi (Receiver Operating Characteristic) altındaki alanı ifade eder. ROC eğrisi, hassasiyet ve geri çağırma arasındaki ilişkiyi gösteren bir grafiktir. ROC AUC puanı, modelin sınıflandırma performansını ölçmek için kullanılan genel bir metriktir. 1'e yaklaştıkça, modelin performansı daha iyidir
Sınıflandırma Modeli Performans Analizi

.

Logistic Regression modeli, en yüksek doğruluk (%97.85), hassasiyet (%98) ve geri çağırma (%98) puanlarını elde ederek dikkat çekmektedir. Ayrıca, ROC AUC puanı da (%0.9914) oldukça yüksektir.

XGBoost modeli de yüksek doğruluk (%95), hassasiyet (%95) ve geri çağırma (%95) puanlarına sahiptir. ROC AUC puanı (%0.9877) da oldukça iyidir.

Random Forest Classification, Support Vector Classification ve CATBoost Classification modelleri, benzer şekilde yüksek doğruluk, hassasiyet, geri çağırma ve ROC AUC puanlarına sahiptir. Bu modellerin performansları birbirine yakındır.

Neural Network Classification modeli ise diğer modellere kıyasla daha düşük performans sergilemektedir. Düşük ROC AUC puanı (%0.9091) ve diğer metriklerdeki (%92.85 doğruluk, %94 hassasiyet, %93 geri çağırma) daha düşük değerler, modelin iyileştirilmesi gerektiğini göstermektedir.

5.Sonuç

Sonuç olarak, veri setimiz için en uygun model Logistic Regression modelidir. Yüksek doğruluk, hassasiyet, geri çağırma ve ROC AUC puanlarıyla bu model, veri setimizdeki sınıflandırma görevini başarılı bir şekilde yerine getirmektedir.
Verinin karmaşıklaşması , dengesiz sınıflar veya daha büyük bir veri seti ile çalışmak gerekirse logistic regresyon aynı başarıyı gösteremeyebilir. Bu durumda XGBoost karmaşık ilişkileri ele alma yeteneğiyle genel olarak iyi performans sergilediği için ikinci en iyi sınıflandırma modeli olarak tercih edilebilir.