

**Gölhisar Uygulamalı Bilimler Yüksekokulu**

**NESNE TABANLI PROGRAMLAMA II DERSİ**

**PROJE KONUSU: Kanser Teşhisi ve Sınıflandırması**

**Öğrenci Ad-Soyad/Numara**

**1- Yiğit Faruk Demir -2112903026**

**2-Hilmi Can Ürlü -2212903003**

**3- Vagıf Beyoğlu-2212903201**

**4- Mert Samet Çeliker-211290357**

**5- Gül Karaman-2212903009**

MAYIS 2024 BURDUR

# GİRİŞ

Bu proje XGBoost algoritması kullanılarak göğüs kanseri tanısına yardımcı makine öğrenmesi modeli oluşturmayı amaçlamaktadır.Yazılım dili olarak Python kullanılmıştır. Veri işlemesi için Pandas ,Veritabanındaki boşlukların NaN değerleri için Numpy,Eğitim ve test verilerini bölmek için Sckitlearn ,Veriyi görselleştirmek için ise Seaborn ve Mathplotlib kütüphanelerinden destek alınmıştır.

XGBoost (Extreme Gradient Boosting), sınıflandırma ve regresyon problemleri için kullanılan, gradient boosting framework'üne dayanan güçlü bir makine öğrenimi algoritmasıdır. XGBoost, hızlı ve verimli olması, ayrıca büyük veri kümeleri ve yüksek boyutlu verilerle başa çıkabilmesi nedeniyle yaygın olarak tercih edilir.

Kullanılan veritabanı Wisconsin Breast Canser veritabanıdır.Bu veritabanından alınmış ve projemizde kullandığımız .csv uzantılı dosyamız 683 satır ve 11 sütundan oluşmaktadır.Kolonlar Sample\_code\_number , Clump\_thickness, Uniformity\_of\_cell\_size, Uniformity\_of\_cell\_shape, Marginal\_adhesion, Single\_epithelial\_cell\_size, Bare\_nuclei, Bland\_chromatin, Normal\_nucleoli, Mitoses, Class isimlerine sahiptir.Bu kolonlardan ilki ID numarası ve sonuncusu sınıf kolonu iken diğer 9 kolon veri kümesinin 9 özelliğini(feature) temsil eder.

Wisconsin Breast Cancer Veri Kümesi, meme kanseri teşhisinde kullanılan yaygın bir veri setidir. Bu veri seti, meme kanseri hücrelerinin özelliklerini içeren ve benign (iyi huylu) veya malignant (kötü huylu) olarak sınıflandırılmış örnekler içerir. Dr. William H. Wolberg tarafından Madison, Wisconsin'deki University of Wisconsin Hastanesi'nde toplanmıştır. Bu veri seti, hem klinik araştırmalar hem de makine öğrenimi modellerinin eğitimi için oldukça değerli bir kaynaktır.

# GEREÇ VE YÖNTEM

XGBoost, yani Extreme Gradient Boosting, 2016 yılında Tianqi Chen ve Carlos Guestrin tarafından geliştirildi ve yayımlandı. XGBoost, Gradient Boosting Machine (GBM) algoritmasının optimizasyonu olarak tanımlanabilir. GBM, hataları minimize etmek için bir dizi zayıf tahmincinin (genellikle karar ağaçları) ardışık olarak eğitildiği bir makine öğrenimi tekniğidir. XGBoost, bu süreci daha verimli ve hızlı hale getiren çeşitli geliştirmeler içerir. Tarihsel olarak, XGBoost, Kaggle yarışmalarında ve çeşitli makine öğrenimi yarışmalarında en üst sıralarda yer alan modellerde kullanılarak popülaritesini hızla artırmıştır.

XGBoost algoritması Gradient Boosting algoritmasının optimize edilmiş bir türüdür. Önceki versiyonlara göre sağladığı avantajları XGBoost kullanımının yaygınlaşmasındaki en önemlinedendir. XGBoost, ağacı oluştururken maksimum derinlik değerini kullanır. Oluşturulan ağaç aşağı yönde aşırı ilerleme gösterirse, budama gerçekleştirilir. Aşırı öğrenmenin önüne geçilir. Gradient Boosting algoritması, kayıp fonksiyonun hesaplanmasında birinci dereceden fonksiyon kullanırken,XGBoost bu hesaplamaları ikinci dereceden fonksiyonlar kullanarak gerçekleştirir. Paralel çalışma özelliği, diğer algoritlamalara göre sonuca daha kısa sürede ulaşılmasını sağlar.

XGBoost, farklı alanlarda geniş bir uygulama yelpazesine sahiptir. Finansal modelleme, pazarlama analitiği, sağlık hizmetleri, biyoinformatik ve müşteri davranış analizi gibi pek çok alanda kullanılmıştır. Örneğin, kredi kartı sahtekarlığı tespiti, müşteri segmentasyonu ve ürün öneri sistemlerinde başarıyla uygulanmıştır. Sağlık hizmetlerinde, özellikle hastalık teşhisi ve prognozu için biyomedikal veri analizi gibi kritik uygulamalarda önemli sonuçlar elde etmiştir. Bu başarıların ardında, XGBoost'un hızlı hesaplama yetenekleri, paralel işlem kapasitesi ve overfitting (aşırı öğrenme) gibi yaygın model problemlerine karşı dirençli olması yatmaktadır.

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu"Gradient boosting" terimi, kolektif olarak güçlü bir model oluşturmak amacıyla tek bir zayıf modeli bir dizi diğer zayıf modelle birleştirerek yükseltme veya geliştirme fikrinden gelir. Gradyan artırma, hataları en aza indirmek amacıyla bir sonraki model için hedeflenen sonuçları belirler. Her bir durum için hedeflenen sonuçlar, tahmine göre hatanın derecesine dayanmaktadır.

## Karşılaştırmalar

Çizelge 1. Projemizin aynı veritabanı üzerinde yapılmış çalışmalar ile performans karşılaştırması

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Referans | Yıl | ML Teknikleri | Accuracy | Precision |
| *Azar and El-Said* | *2013* | *Artificial Neural Networks* | *97,66%* | *97,77%* |
| *Wang and Feng* | *2018* | *K-Nearest Neighbors* | *95,80%* | *-* |
| *Sivapriya J, Aravind Kumar V* | *2019* | *Linear Regression* | *99.06%* | *-* |
| *Sulyman Age Abdulkareem* | *2021* | *Random Forest* | *97,07%* | *97%* |
| *BİZİM PROJEMİZ* | *2024* | *XGBoost* | *94,16%* | *96,29%* |

# BULGULAR

## Grafikler

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

**(1) Örnek şekil 1**

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

**(2) Örnek şekil 2**

Şekil 1 ve 2 de veritabanında bulunan class kısmındaki veri dağılımını görüyoruz benign (0 ve sağılıklı isimleri ile temsil edildi) malignant(1 ve hastalıklı isimleri ile temsil edildi)

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

**(3) Örnek şekil 3**

Şekil 3 karmaşıklık matrisinin görselleştirilmesidir. Karmaşıklık matrisin sol sütunu algoritmamız sonucunda tahmin edilen false(0-sağlıklı) değeridir sağ sütun algoritmamız tarafından tahmin edilen true(1-hastalıklı) değeridir.Karmaşıklık matrisinin ilk satırı veritabanındaki false(0-sağlıklı) değeridir.İkinci satır ise true(1-hastalıklı)değerini temsil eder

TP (True Positive-matrisimizin sağ alt köşesi): 1 olarak sınıflandırdığımız ve gerçekten de 1 olan değerlerin sayısıdır.

FP (False Positive-matrisimizin sağ üst köşesi): 1 olarak sınıflandırdığımız fakat gerçekte 0 olan değerlerin sayısıdır.

TN (True Negative-matrisimizin sol üst köşesi): 0 olarak sınıflandırdığımız ve gerçekten de 0 olan değerlerin sayısıdır.

FN (False Negative-matrisimizin sol alt köşesi): 0 olarak sınıflandırdığımız fakat gerçekte 1 olan değerlerin sayısıdır.

**metin, yazı tipi, ekran görüntüsü, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**(4) Örnek şekil 4**

Şekil 4 Accuracy (Doğruluk) ,F1 Score (F1 Puanı), Precision and Recall (Hassasiyet ve Geri Çağırma) performans metriklerinin hesaplamasından elde edilen sonuçlardır.

metin, diyagram, plan, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldu

**(5) Örnek şekil 5**

metin, diyagram, ekran görüntüsü, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu

**(6) Örnek şekil 6**

Şekil 5 ve 6 veritabanındaki 9 özelliğin (feature) dağılımını gösteren grafiklerdir.

Şekil 5 bir histogram grafiğidir.

Şekil 6 bir keman grafiğidir.

# KAYNAKLAR

<https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original>

<https://www.nvidia.com/en-us/glossary/xgboost/>

MQTT Trafiğinde DoS Saldırılarının Makine Öğrenmesi ile Sınıflandırılması ve Modelin SHAP ile Yorumlanması AC Kelle, H Yüce Journal of Materials and Mechatronics: A, 2022•dergipark.org.tr

RR Kadhim, MY Kamil - Int J Reconfigurable & Embedded Syst ISSN, 2022 - academia.edu

An Evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique Sulyman Age Abdulkareema , Zainab Olorunbukademi Abdulkareemb

Breast Cancer Prediction using Machine Learning Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S

Y. Wang and L. Feng, "Hybrid feature selection using component co-occurrence based feature relevance measurement," Expert Systems with Applications, vol. 102, pp. 83-99, 2018, doi: 10.1016/j.eswa.2018.01.041.

A. T. Azar and S. A. El-Said, "Probabilistic neural network for breast cancer classification," Neural Computing and Applications, vol. 23, no. 6, pp. 1737-1751, 2013, doi: 10.1007/s00521-012-1134-8

**KAYNAK KOD ADRESİ**

<https://github.com/YigitFarukDemir/Breast_Cancer_Prediction_using_XGBoost>