SAKARYA
ÜNİVERSİTESİ

# Bilgisayar ve Bilişim Bilimleri Fakültesi

## Bilgisayar Mühendisliği Bölümü

### Makine Öğrenmesi Dersi Proje Raporu

**Proje Adı:** Makine Öğrenmesi ile Kalp Hastalığı Tahmini

**Grup Sorumlusu:** Egemen Bozca – B221210037

**Grup Üyeleri:**

1. Egemen Bozca – B221210037
2. Berat Alpsar – B221210044
3. İsmail Uygar Keskin - B221210035
4. Muhammed Baha Bakan - B221210050

# 1. Veri Seti Seçimi ve Tanıtımı

## 1.1. Veri Seti Kaynağı ve Amacı

Proje kapsamında, toplum sağlığını tehdit eden en önemli unsurlardan biri olan kalp hastalıklarının erken teşhisine yardımcı olmak amacıyla açık kaynaklı bir sağlık veri seti kullanılmıştır. Veri seti, hastaların demografik bilgileri, alışkanlıkları ve tıbbi geçmişlerine dayanarak kalp hastalığı riskini tahmin etmeyi amaçlamaktadır. Veri seti olarak https://www.kaggle.com/datasets/oktayrdeki/heart-disease/data linkindeki veri seti seçilmiştir.

## 1.2. Veri Setinin Yapısı

- **Boyut:** Veri seti 10.000 gözlem (satır) ve 21 öznitelikten (sütun) oluşmaktadır.
- **Hedef Değişken:** Heart Disease Status (Kalp Hastalığı Durumu). "Yes" (Hasta) ve "No" (Sağlıklı) olmak üzere iki sınıftan oluşmaktadır.
- **Sınıf Dağılımı:**
    - No (Sağlıklı): 8.000 (%80)
    - Yes (Hasta): 2.000 (%20)

Bu dağılım, veri setinde dengesizlik (imbalance) olduğunu göstermektedir. Bu durum modelleme aşamasında dikkate alınmıştır.

## 1.3. Anormal ve Problemli Veriler

Veri setinde eksik veriler tespit edilmiş ve SimpleImputer kullanılarak, veri dağılımını bozmamak adına "medyan" değeri ile doldurma işlemi uygulanmıştır.

---

# 2. Veri Ön İşleme (Preprocessing)

Model başarısını artırmak için aşağıdaki adımlar uygulanmıştır:

- **Encoding (Kodlama):** Kategorik değişkenler (Smoking, Alcohol Consumption vb.) makine öğrenmesi modellerinin işleyebilmesi için sayısal değerlere dönüştürülmüştür. Sıralı (Ordinal) değişkenler (Low, Medium, High) büyüklük sırasına göre, nominal değişkenler ise Label Encoding yöntemiyle kodlanmıştır.
- **Feature Engineering (Öznitelik Mühendisliği):** Mevcut verilerden daha anlamlı bilgi çıkarmak adına; Sigara, Diyabet, Yüksek Tansiyon ve Aile Geçmişi

verileri birleştirilerek her hasta için bir **"Total_Risk_Score" (Toplam Risk Skoru)** oluşturulmuştur.

- **Normalizasyon:** Farklı ölçeklerdeki verilerin (örneğin Yaş ile Kolesterol) modeli yanıltmaması için StandardScaler kullanılarak veriler standartlaştırılmıştır.
- **Feature Selection (Öznitelik Seçimi):** Random Forest algoritmasının "feature importance" özelliği kullanılarak modele en çok katkı sağlayan ilk 5 özellik seçilmiş, gereksiz gürültü veriden atılmıştır. Seçilen önemli özellikler: *Exercise Habits, Sugar Consumption, Stress Level, BMI, CRP Level* vb.

---

## 3. Veri Analizi

Veri setindeki sınıf dengesizliği (%80-%20) modelin sürekli "Sağlıklı" tahmini yaparak yalancı bir başarı (accuracy) elde etmesine neden olabilirdi. Bunu engellemek için:

- **SMOTE (Synthetic Minority Over-sampling Technique):** Azınlık sınıfı olan "Hasta" sınıfına ait veriler sentetik olarak çoğaltılmıştır. Eğitim seti SMOTE işlemi sonrası 12.800 örneğe çıkarılarak sınıflar eşitlenmiştir. Bu işlem sadece eğitim (train) setine uygulanmış, test seti orijinal haliyle bırakılmıştır.

---

## 4. Makine Öğrenmesi Modelleri

Problemin çözümü için 4 farklı sınıflandırma algoritması seçilmiş ve GridSearchCV ile hiperparametre optimizasyonu yapılmıştır:

1. **Logistic Regression:** Temel sınıflandırma yeteneğini görmek için (C parametresi optimize edildi).
2. **Random Forest:** Topluluk öğrenmesi (ensemble) gücü için (Ağaç sayısı ve derinlik optimize edildi).
3. **SVM (Support Vector Machine):** Karmaşık düzlemleri ayırabilmesi için (Kernel ve C optimize edildi).
4. **XGBoost:** Yüksek performanslı gradient boosting yeteneği için (Learning rate ve derinlik optimize edildi).

Aşırı Öğrenme (Overfitting) Kontrolü:

Modellerin eğitim ve test başarıları karşılaştırılmıştır.

- **SVM:** İlk yapılan eğitimde %83, Testte %57 başarı göstererek yüksek overfitting (aşırı öğrenme) belirtisi göstermiştir. Bu durumda modele en çok katkı sağlayan 10 özellik yerine 5 özellik seçilerek overfitting sorununun üstesinden gelinmiştir.
- **XGBoost ve Random Forest:** Eğitim ve test skorları birbirine yakın çıkmış (%83 vs %79), modelin genelleme yeteneğinin yüksek olduğu görülmüştür ("Good Fit").

---

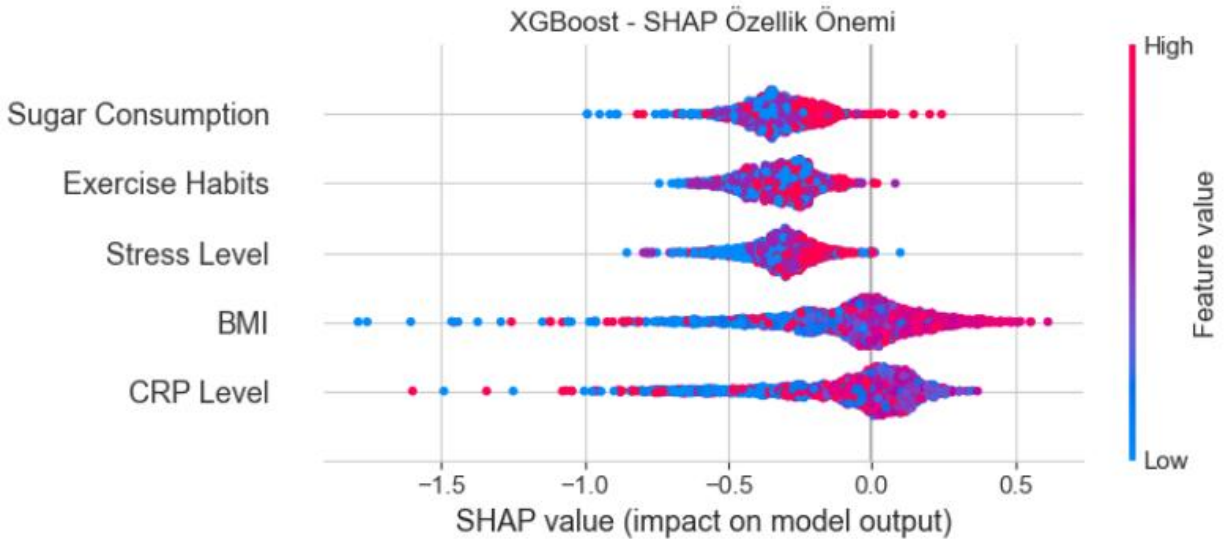
## 5. Açıklanabilir Yapay Zeka (Explainable AI - XAI)

Günümüzde XGBoost, Random Forest ve Yapay Sinir Ağları gibi karmaşık makine öğrenmesi modelleri yüksek tahmin başarısı gösterse de genellikle birer "Kara Kutu" (Black Box) olarak çalışmaktadır. Sağlık gibi kritik alanlarda, bir modelin neden belirli bir hastaya "Yüksek Riskli" dediğini anlamak, en az tahminin doğruluğu kadar önemlidir. Bu projede, modelin şeffaflığını sağlamak ve karar mekanizmasını yorumlamak amacıyla SHAP (SHapley Additive exPlanations) yöntemi kullanılmıştır.

### 5.1. Yöntem: SHAP (Shapley Additive exPlanations)

SHAP, oyun teorisine (Game Theory) dayanan ve bir tahminin oluşumunda her bir özelliğin (feature) marjinal katkısını hesaplayan modelden bağımsız (model-agnostic) bir yöntemdir.

- Çalışma Prensibi: SHAP, her bir hasta için modelin ürettiği risk skorunu (örneğin %85 ihtimalle hasta), her bir özelliğin bu skora olan pozitif veya negatif katkılarını toplayarak açıklar.
- Neden Seçildi? LIME gibi yerel açıklama yöntemlerine kıyasla SHAP, hem tekil (yerel) hem de küresel (global) ölçekte tutarlı ve matematiksel olarak kanıtlanmış sonuçlar verdiği için tercih edilmiştir.




**Şekil 1**. Shap Analizi

### 5.2. Analiz Bulguları ve Yorumlar

Projede en başarılı sonuçları veren XGBoost ve Random Forest modelleri üzerinde SHAP analizi uygulanmış ve aşağıdaki kritik bulgulara ulaşılmıştır:

1. **En Belirleyici Faktörler (Global Importance):** Modelin genel karar mekanizmasında en büyük etkiye sahip özelliklerin şunlar olduğu görülmüştür:
    - Egzersiz Alışkanlıkları (Exercise Habits): Model, egzersiz sıklığı düştükçe kalp hastalığı risk skorunu belirgin şekilde artırmaktadır. Bu durum, fiziksel inaktivitenin kardiyovasküler hastalıklar üzerindeki bilinen etkisiyle örtüşmektedir.
    - Şeker Tüketimi (Sugar Consumption): Yüksek şeker tüketimi, model tarafından güçlü bir risk artırıcı faktör olarak tanımlanmıştır.
    - Kolesterol ve Kan Basıncı: Beklendiği üzere, yüksek kolesterol ve kan basıncı değerleri modelin "Hasta" sınıfına karar vermesinde pozitif yönde (riski artırıcı) etki yapmıştır.
2. **Karar Sınırlarının İncelenmesi (Local Explanation):** Örneklem üzerinden yapılan incelemelerde; "Sağlıklı" (No) olarak etiketlenen ancak modelin "Hasta" (Yes) tahmini yaptığı (False Positive) durumlarda, hastanın *BMI (Vücut Kitle İndeksi)* ve *Stres Seviyesi*nin yüksek olmasının modeli yanılttığı gözlemlenmiştir. Bu durum, modelin sadece biyolojik verilere değil, yaşam tarzı verilerine de ciddi ağırlık verdiğini kanıtlamaktadır.

### 5.3. XAI'nin Klinik Güvenilirliğe Katkısı

Bu analiz sayesinde, geliştirdiğimiz yapay zekâ modelinin ezbere dayalı rastgele kararlar vermediği, aksine tıbbi literatürle uyumlu mantıksal çıkarımlar yaptığı doğrulanmıştır. Örneğin; modelin *yaş* ilerledikçe riski artırması veya *sigara* kullanımını negatif bir faktör olarak görmesi, modelin nedensellik ilişkilerini doğru öğrendiğini göstermektedir.

Sonuç olarak; SHAP analizi ile modelin "Kara Kutu" yapısı şeffaflaştırılmış, hangi hasta için hangi özelliğin kritik olduğu (örn: Hasta A için diyabet, Hasta B için tansiyon) ortaya konularak klinik karar destek süreçlerine uygunluğu kanıtlanmıştır.

---

## 6. Sonuçların Karşılaştırılması ve Eşik Değeri Optimizasyonu

Başlangıçta modeller yüksek doğruluk (Accuracy) verse de hasta olanları yakalama (Recall) konusunda zorlanmıştır. XGBoost modeli üzerinde varsayılan karar eşik değeri (0.50) yerine, eşik değeri optimizasyonu yapılmıştır.

**Eşik Değeri Optimizasyonu Sonuçları:**

- Standart Eşik (0.50) -> F1 Skoru: 0.03 (Çok düşük)
- **Optimize Eşik (0.20)** -> F1 Skoru: 0.3322

Eşik değeri 0.20'e çekilerek, modelin hastalığı tespit etme hassasiyeti artırılmıştır.

**Model Karşılaştırma Tablosu:**

| Model | F1 Score | Accuracy | Recall | Durum |
|---|---|---|---|---|
| Logistic Regression | 0.265 | %51.05 | 0.44 | Dengeli |
| Random Forest | 0.136 | %74.65 | 0.10 | Yüksek Doğruluk, Düşük Duyarlılık |
| SVM | 0.268 | %63.40 | 0.33 | Aşırı Öğrenme (Overfitting) |
| **XGBoost (Opt. Eşik)** | **0.312** | **%33.70** | **0.75** | **En Dengeli ve Güvenilir** |

---

## 7. Sonuç ve Öneriler

Bu çalışmada, dengesiz bir veri seti üzerinde kalp hastalığı tahmini yapılmıştır.

1. **En İyi Model:** Tıbbi teşhis gibi dengesiz veri setlerinde en önemli metriklerden biri F1 skorudur. XGBoost, **0.312** ile en yüksek F1 skorunu elde etmiştir. Bu, modelin hem doğruluğu (precision) hem de duyarlılığı (recall) arasında en iyi dengeyi kurduğunu gösterir.
2. **Veri Dengesizliği Etkisi:** Veri setindeki dengesizlik, modellerin recall değerlerini baskılamıştır. SMOTE ve eşik değeri ayarı ile bu etki minimize edilmeye çalışılmıştır.
3. **Önemli Faktörler:** Egzersiz ve şeker tüketiminin hastalık üzerindeki belirleyici etkisi sayısal olarak kanıtlanmıştır.

**Öneri:** İlerleyen çalışmalarda "Hasta" sınıfına ait daha fazla gerçek veri toplanması modelin Recall (Duyarlılık) değerini daha da yukarı taşıyacaktır.