# AutoML Modellerinin Obezite Tahmininde Kıyaslanması

## 1. Giriş ve Projenin Amacı

Obezite, günümüzde giderek artan bir sağlık sorunu haline gelmiştir ve hem bireysel hem de toplumsal düzeyde ciddi sonuçlara yol açmaktadır. Obezite, yüksek tansiyon, diyabet, kalp hastalıkları gibi kronik rahatsızlıklara zemin hazırlarken, yaşam kalitesini de önemli ölçüde düşürmektedir. Bu doğrultuda, bireylerin obezite riski taşıyıp taşımadıklarının erken teşhisi büyük önem arz etmektedir.

Makine öğrenimi, bireylerin sağlık verilerini analiz ederek obezite tahmininde kullanılabilecek etkili bir araçtır. Bu çalışmada, çeşitli makine öğrenimi modelleri kullanarak obezite durumunu tahmin etmek amaçlanmıştır. Veri seti, bireylerin demografik özellikleri, yaşam tarzları ve beslenme alışkanlıklarını içermekte olup, bu faktörlerin obezite ile olan ilişkisi modeller aracılığıyla değerlendirilmiştir. Çalışmanın amacı, en iyi performansı gösteren makine öğrenimi modelini belirlemek ve bu modelin obezite tahmininde nasıl kullanılabileceğini ortaya koymaktır.

## 2. Veri Seti Değerlendirilmesi

| | Gender | Age | Height | Weight | family_history_with_overweight | FAVC | FCVC | NCP | CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 21.0 | 1.62 | 64.0 | yes | no | 2.0 | 3.0 | Sometimes | no | 2.0 | no | 0.0 | 1.0 | no | Public_Transportation | Normal_Weight |
| 1 | Female | 21.0 | 1.52 | 56.0 | yes | no | 3.0 | 3.0 | Sometimes | yes | 3.0 | yes | 3.0 | 0.0 | Sometimes | Public_Transportation | Normal_Weight |
| 2 | Male | 23.0 | 1.80 | 77.0 | yes | no | 2.0 | 3.0 | Sometimes | no | 2.0 | no | 2.0 | 1.0 | Frequently | Public_Transportation | Normal_Weight |
| 3 | Male | 27.0 | 1.80 | 87.0 | no | no | 3.0 | 3.0 | Sometimes | no | 2.0 | no | 2.0 | 0.0 | Frequently | Walking | Overweight_Level_I |
| 4 | Male | 22.0 | 1.78 | 89.8 | no | no | 2.0 | 1.0 | Sometimes | no | 2.0 | no | 0.0 | 0.0 | Sometimes | Public_Transportation | Overweight_Level_II |

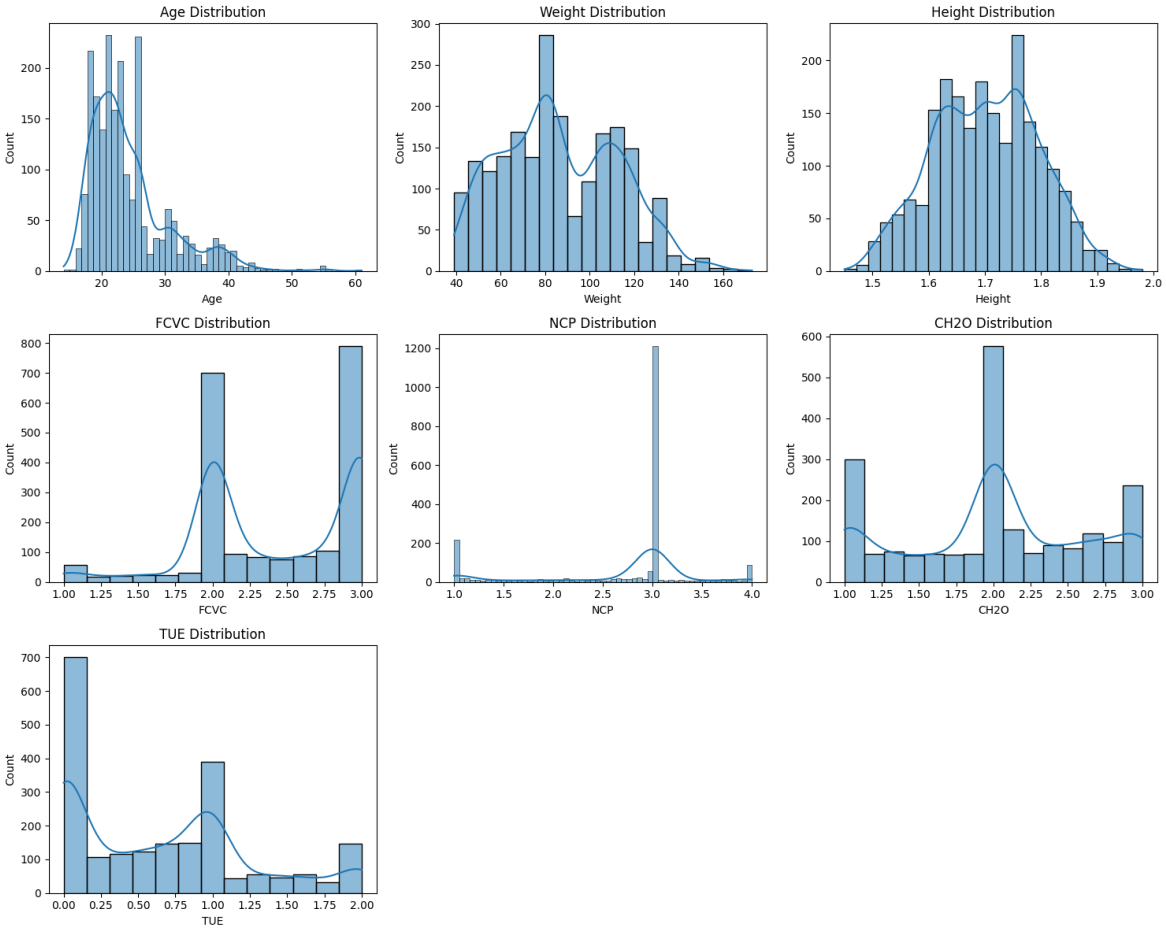
Çalışmada kullanılan veri seti, obezite durumu ile ilişkili birçok değişkeni kapsamaktadır. Aşağıda veri setinde yer alan temel özellikler ve açıklamaları verilmiştir:

- **Cinsiyet (Gender)**: Bireyin cinsiyeti (0 = Erkek, 1 = Kadın).
- **Yaş (Age)**: Bireyin yaşı.
- **Boy (Height)**: Bireyin metre cinsinden boy uzunluğu.
- **Ağırlık (Weight)**: Bireyin kilogram cinsinden vücut ağırlığı.
- **Ailede fazla kilo geçmişi (family_history_with_overweight)**: Bireyin ailesinde fazla kilo geçmişi olup olmadığı (1 = Evet, 0 = Hayır).
- **Yüksek kalorili yiyecek tüketimi (FAVC)**: Bireyin yüksek kalorili yiyecek tüketme durumu (1 = Evet, 0 = Hayır).
- **Sebze tüketim sıklığı (FCVC)**: Bireyin sebze tüketme alışkanlığı (0 = Hiç, 1 = Az, 2 = Orta, 3 = Çok).
- **Günlük öğün sayısı (NCP)**: Bireyin günlük ana öğün sayısı (1 = 1 öğün, 2 = 2 öğün, 3 = 3 öğün).
- **Ara öğün sıklığı (CAEC)**: Bireyin ara öğün tüketim sıklığı (0 = Hiç, 1 = Bazen, 2 = Sık sık, 3 = Her zaman).
- **Sigara içme durumu (SMOKE)**: Bireyin sigara içme durumu (1 = Evet, 0 = Hayır).

- **Günlük su tüketimi (CH2O)**: Bireyin günlük su tüketme miktarı (1 = Az, 2 = Orta, 3 = Çok).
- **Kalori alımını takip etme (SCC)**: Bireyin kalori alımını takip etme durumu (1 = Evet, 0 = Hayır).
- **Fiziksel aktivite sıklığı (FAF)**: Bireyin haftalık fiziksel aktivite sıklığı (0 = Hiç, 1 = Az, 2 = Orta, 3 = Sık).
- **Günlük cihaz kullanımı (TUE)**: Bireyin günlük teknolojik cihaz kullanma süresi (1 = Az, 2 = Orta, 3 = Çok).
- **Alkol tüketimi (CALC)**: Bireyin alkol tüketme durumu (0 = Hiç, 1 = Az, 2 = Orta, 3 = Sık).
- **Ulaşım şekli (MTRANS)**: Bireyin ulaşım tercihi (0 = Yürüyerek, 1 = Bisiklet, 2 = Kamu Taşımacılığı, 3 = Araç).
- **Obezite durumu (NObeyesdad)**: Bireyin vücut kitle indeksine (BMI) göre sınıflandırılan obezite durumu (0 = Zayıf, 1 = Normal, 2 = Fazla Kilolu, 3 = Obez, 4 = Aşırı Obez).

## 3. Grafiklerin Değerlendirilmesi

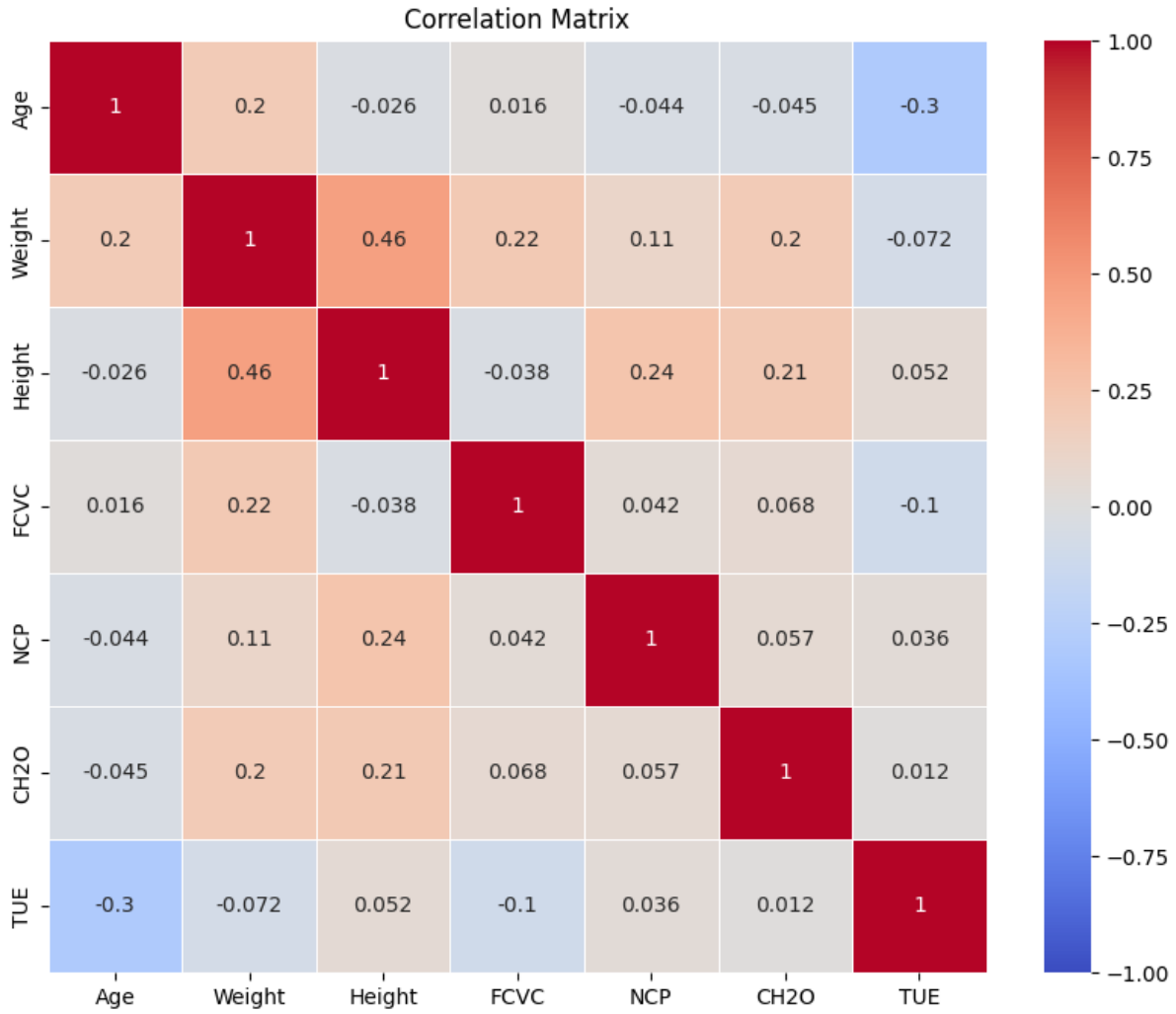
### Histogramlar ve Dağılımlar :




İlk görselde farklı değişkenlerin dağılımı gösterilmiştir:

- **Age Distribution (Yaş Dağılımı)**: Dağılım, genç bireylerin daha fazla olduğunu göstermektedir. 20-25 yaş arası bireyler en yoğun grubu oluşturmaktadır. Daha yaşlı bireylerin sayısı giderek azalmaktadır.
- **Weight Distribution (Ağırlık Dağılımı)**: Ağırlık dağılımı, 60 ile 90 kilogram arasındaki bireylerin yoğunlaştığını göstermektedir. Ağırlık artışı ile birey sayısında bir azalma gözlemlenmektedir. Ancak, 120 kg ve üzeri bireylerde hafif bir artış da gözlemlenmektedir.
- **Height Distribution (Boy Dağılımı)**: Boy dağılımı ortalamaya yakın bir eğilim sergilemektedir. Bireylerin çoğu 1.6 ile 1.8 metre boyunda olup, uç değerler daha az sayıda birey içermektedir.
- **FCVC (Sebze Tüketim Sıklığı)**: Çoğu birey "2" ve "3" değerlerine sahip, yani orta ve yüksek seviyede sebze tüketimi yaygın. Hiç veya çok az sebze tüketen bireyler daha az.
- **NCP (Günlük Ana Öğün Sayısı)**: Çoğu birey 3 ana öğün tüketmektedir. 1 veya 2 öğün tüketen bireyler daha azdır.
- **CH2O (Su Tüketim Miktarı)**: Su tüketiminde çoğunluk orta seviyededir (2). 1 ve 3 seviyesinde tüketim daha azdır.
- **TUE (Günlük Teknolojik Cihaz Kullanımı)**: Genellikle bireyler az ve orta seviyede teknolojik cihaz kullanmaktadır (1-2 saat). Çok fazla cihaz kullanan birey sayısı düşüktür.

## Korelasyon Matrisi :




İkinci görsel, değişkenler arasındaki korelasyonları göstermektedir. Korelasyon katsayıları -1 ile 1 arasında değişir; -1 ters ilişki, 1 ise doğrudan ilişki anlamına gelir. 0 ise ilişki olmadığını gösterir.

- **Age (Yaş)** ve **TUE (Teknolojik Cihaz Kullanımı)** arasında orta seviyede negatif bir ilişki var (-0.3). Yani yaş arttıkça cihaz kullanım süresi azalıyor.
- **Weight (Ağırlık)** ve **Height (Boy)** arasında pozitif bir ilişki var (0.46). Bu da, genel olarak daha uzun bireylerin daha ağır olduğunu gösterir.
- **Weight (Ağırlık)** ve **FCVC (Sebze Tüketim Sıklığı)** arasında pozitif bir korelasyon var (0.22). Bu, sebze tüketiminin ağırlıkla bir miktar ilişkili olduğunu gösterir.
- Diğer değişkenler arasında genellikle düşük korelasyonlar bulunmaktadır, yani güçlü bir doğrusal ilişki görülmemektedir.

### Sonuç:

Veriler, yaş, ağırlık, boy ve yaşam tarzı alışkanlıkları arasında bazı anlamlı ilişkiler olduğunu gösteriyor. Ancak, değişkenler arasındaki ilişkilerin çoğunluğu zayıf veya orta seviyede, bu da

kompleks modellerin gerekebileceğini işaret etmektedir. Özellikle yaş ve cihaz kullanımı gibi faktörler arasında dikkat çekici korelasyonlar bulunmuştur.

# 4.AutoML Modellerinin Kıyaslanması

## Kullanılan Yöntemler

#### 1) MLJAR AutoML

MLJAR, kullanıcıların çok az manuel müdahale ile yüksek performanslı makine öğrenme modelleri oluşturmasına olanak tanıyan bir otomatik modelleme aracıdır. Hem sınıflandırma hem de regresyon görevlerinde, birçok farklı model tipini deneyerek en iyi performans gösteren modeli bulur. MLJAR, modele dayalı eğitim süresi ve doğruluk gibi faktörleri optimize ederek zaman kazandırır. Bu çalışmada, en iyi model olarak **LightGBM** seçilmiştir.

#### 2) FLAML AutoML

FLAML (Fast and Lightweight AutoML), düşük maliyetli ve hızlı model eğitimi sağlamak için tasarlanmış hafif bir otomatik modelleme aracıdır. FLAML, hiperparametre optimizasyonunu otomatik olarak gerçekleştirir ve doğruluğu optimize eden modeller sunar. Bu çalışmada, **LightGBM** modeli FLAML tarafından optimize edilmiş ve en iyi model olarak belirlenmiştir.

#### 3) TPOT AutoML

TPOT, genetik algoritmalar kullanarak en iyi makine öğrenme modellerini ve pipeline'ları bulan bir AutoML kütüphanesidir. TPOT, çok sayıda model ve pipeline denemesi gerçekleştirir ve hiperparametre optimizasyonu yaparak en iyi performansı sunan pipeline'ı seçer. Bu çalışmada, TPOT ile en iyi model olarak **RandomForestClassifier** belirlenmiş ve hiperparametre optimizasyonu yapılmıştır. TPOT tarafından elde edilen nihai doğruluk oranı **%96.69** olarak hesaplanmıştır.

TPOT'nin arama süreci sırasında kullandığı hiperparametreler şu şekildedir:

1. **generations (nesiller)**: Genetik algoritmanın kaç jenerasyon boyunca çalışacağını belirler. Bu çalışmada 5 ve 10 jenerasyon seçenekleri kullanılmıştır. Daha fazla jenerasyon, daha fazla model denemesi anlamına gelir, ancak bu süreç daha uzun sürebilir.
2. **population_size (popülasyon büyüklüğü)**: Her jenerasyonda değerlendirilecek model sayısını ifade eder. Bu çalışmada 50 ve 100 birey popülasyon büyüklükleri denenmiştir. Daha büyük popülasyon, daha geniş bir model araması yapılmasına olanak sağlar, ancak yine daha fazla zaman gerektirir.

3. **verbosity (detay seviyesi)**: TPOT'nin çalışma sırasında ne kadar bilgi vereceğini ayarlar. Bu çalışmada verbosity 2 olarak ayarlanmış olup, TPOT işlemleri sırasında daha fazla detaylı çıktı üretmiştir.

Sonuç olarak, TPOT'nin seçtiği en iyi model olan **RandomForestClassifier**, aşağıdaki hiperparametrelerle çalıştırılmıştır:

- **bootstrap**: False
- **criterion**: entropy
- **max_features**: 0.6
- **min_samples_leaf**: 3
- **min_samples_split**: 2
- **n_estimators**: 100

Bu optimizasyonlar sonucunda TPOT, en uygun modeli oluşturarak %96.69 doğruluk oranına ulaşmıştır.

## Sonuçlar

#### MLJAR AutoML

MLJAR AutoML'de en iyi model olarak **LightGBM** seçilmiştir ve model doğruluk oranı %96.22 olarak elde edilmiştir.

#### FLAML AutoML

FLAML AutoML'de **LightGBM** modeli %96.93 doğruluk oranıyla en yüksek performansı göstermiştir. *(Parametreler : LGBMClassifier(colsample_bytree=0.763007791741338, learning_rate=0.16645809713264254, max_bin=1023, min_child_samples=6, n_estimators=1, n_jobs=-1, num_leaves=12, reg_alpha=0.0009765625, reg_lambda=0.10626868868028042, verbose=-1))*

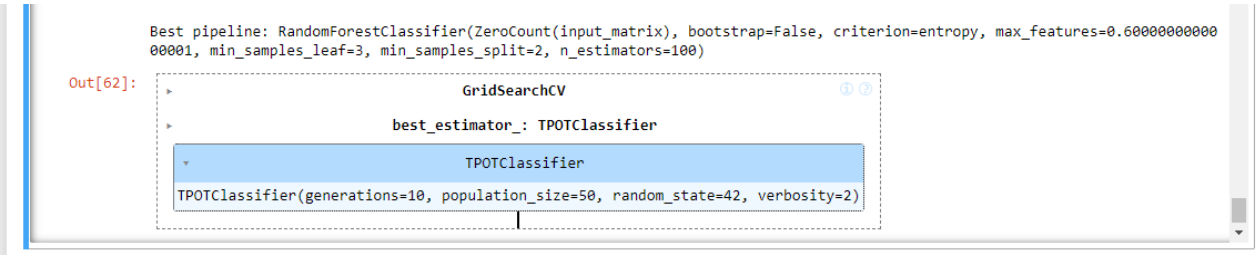
#### TPOT AutoML

TPOT AutoML, genetik algoritma ile optimize edilen **RandomForestClassifier** modeliyle %96.69 doğruluk oranı elde etmiştir.

(*Parametreler : Best pipeline: RandomForestClassifier(ZeroCount(input_matrix), bootstrap=False, criterion=entropy, max_features=0.6000000000000001, min_samples_leaf=3, min_samples_split=2, n_estimators=100)*)

### *Genel Değerlendirme:*

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| # FLAML (Fast and Lightweight AutoML): (LGBMClassifier) | 0.9692 | 0.97 | 0.97 | 0.97 |
| MLJAR AutoML¶ (LightGBM) | 0.9621 | 0.96 | 0.96 | 0.96 |
| TPOT (RandomForestClassifier) | 0.966 | 0.97 | 0.97 | 0.97 |




## Sonuç Raporu

Bu çalışma, obezite sınıflandırması için MLJAR, FLAML ve TPOT AutoML araçlarını kullanarak kapsamlı bir karşılaştırma yapmıştır. Sonuçlar, her üç aracın da yüksek doğruluk oranları ile başarılı sınıflandırma modelleri ürettiğini göstermektedir.

- **FLAML AutoML**, %96.93 doğruluk oranıyla en yüksek performansı göstermiştir.(**LightGBM**)
- **TPOT AutoML**, optimize edilen **RandomForestClassifier** ile %96.69 doğruluk oranına ulaşmıştır.
- **MLJAR AutoML**, %96.22 doğruluk oranıyla onu takip etmektedir. (**LightGBM**)

Bu üç yöntemin her biri, sınıflandırma problemlerine yönelik farklı yaklaşımlar sunarken, hepsi de güçlü sonuçlar elde etmiştir. Verilen süre ve kaynak göz önünde bulundurulduğunda, FLAML AutoML en hızlı ve en doğru sonuçları sağlamıştır. Ancak, her üç araç da model performansını artırmak için etkili seçenekler sunmaktadır.