Obezite Tahmininde Makine Öğrenimi ve AutoML Modellerinin 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

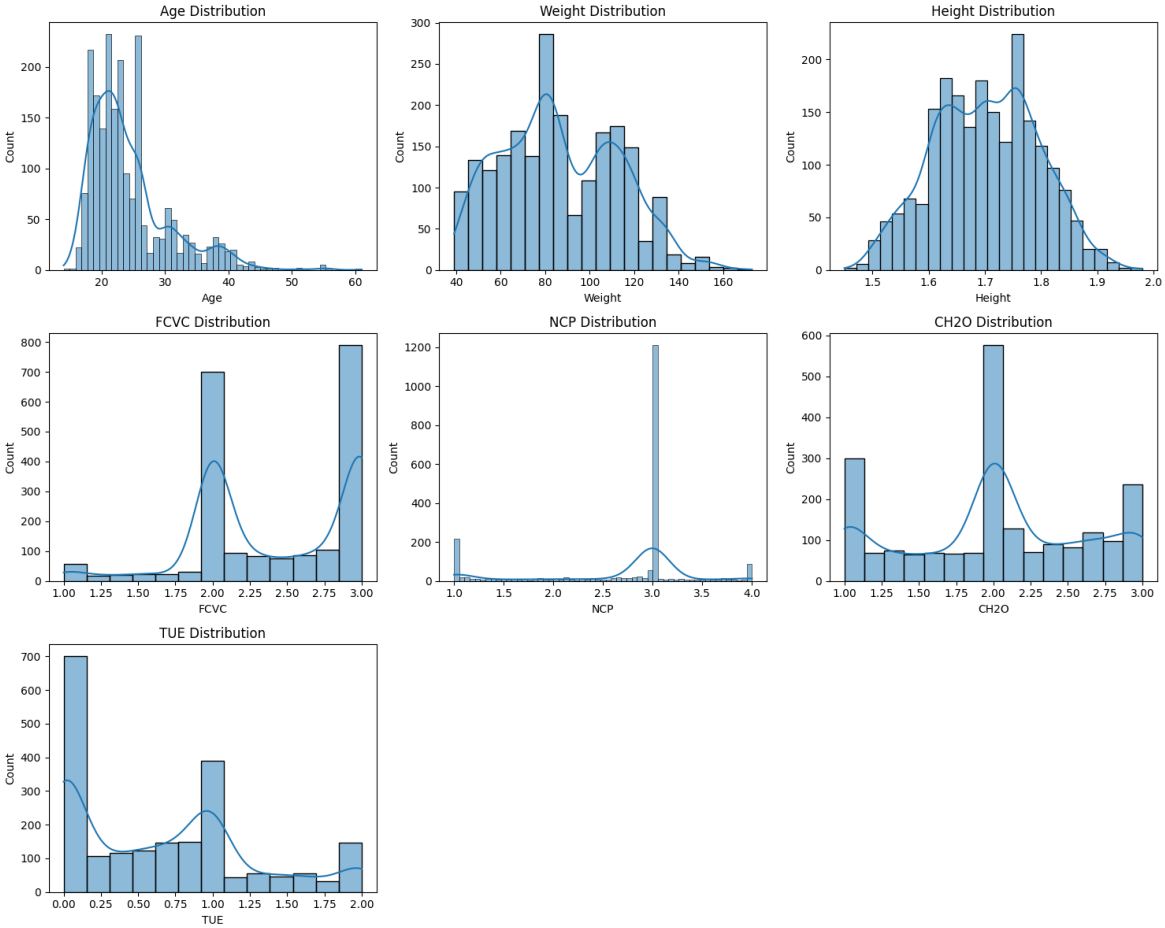
Çalışmada kullanılan veri seti, obezite durumu ile ilişkili birçok değişkeni kapsamaktadır. Aşağıda veri setinde yer alan temel özellikler ve açıklamaları verilmiştir:

- **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

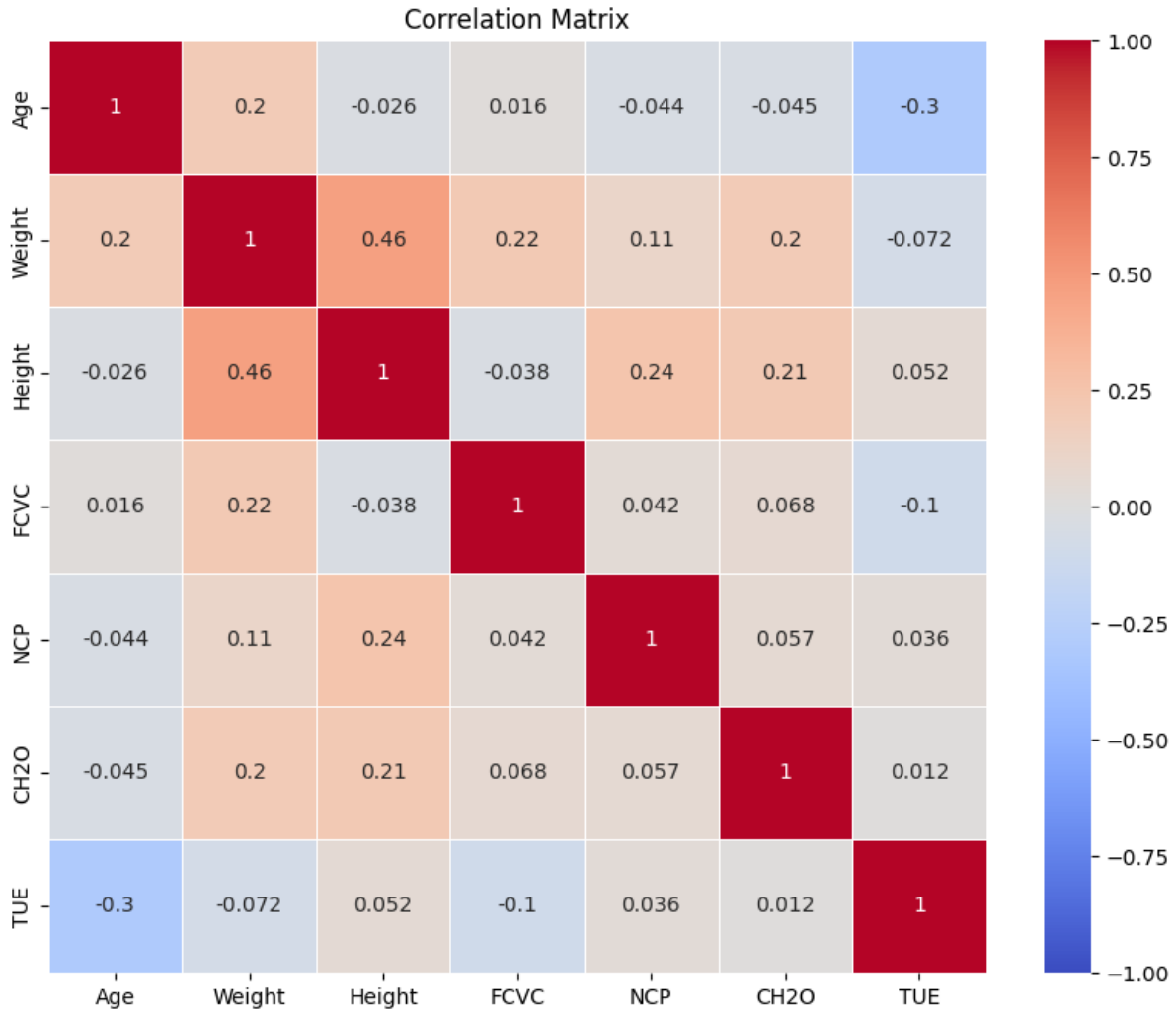
Histogramlar ve Dağılımlar :



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. Makine Öğrenimi Modellerinin Obezite Tahminindeki Performans Karşılaştırması

Kullanılan Modeller

Çalışmada, çeşitli denetimli makine öğrenimi algoritmaları kullanılmıştır. Bunlar arasında **Random Forest**, **Gradient Boosting**, **AdaBoost**, **SVC (Destek Vektör Makineleri)**, **Lojistik Regresyon**, **K-Nearest Neighbors (KNN)**, **Decision Tree**, **Gaussian Naive Bayes**, **Extra Trees Classifier** ve **MLP Classifier** (Çok Katmanlı Algılayıcı) bulunmaktadır.

Her bir model, eğitim verileri üzerinde eğitilmiş ve test verileri üzerinde tahmin yapılmıştır. Performans sonuçları aşağıdaki gibi dört temel metrik kullanılarak değerlendirilmiştir:

1. **Accuracy (Doğruluk):** Modelin doğru sınıflandırdığı örneklerin tüm örneklere oranıdır.
2. **Precision (Kesinlik):** Modelin pozitif sınıflandırdığı örneklerin ne kadarının doğru olduğunun ölçüsüdür.
3. **Recall (Geri Çağırma):** Pozitif sınıflandırılması gereken örneklerin ne kadarının doğru bir şekilde sınıflandırıldığını gösterir.
4. **F1 Score:** Precision ve Recall'un harmonik ortalamasıdır.

Modellerin Performans Sonuçları

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.9432	0.9452	0.9432	0.9437
Gradient Boosting	0.9479	0.9492	0.9479	0.9478
AdaBoost	0.3659	0.5303	0.3659	0.2595
SVC	0.8722	0.8765	0.8722	0.8725
Logistic Regression	0.8580	0.8613	0.8580	0.8543
K-Nearest Neighbors	0.8044	0.8023	0.8044	0.7958
Decision Tree	0.9148	0.9147	0.9148	0.9145
Gaussian Naive Bayes	0.6057	0.6148	0.6057	0.5867
Extra Trees Classifier	0.9353	0.9364	0.9353	0.9354
MLP Classifier	0.9259	0.9260	0.9259	0.9259

Sonuçlar ve Yorumlar

1) **Gradient Boosting** ve **Random Forest** modelleri, en yüksek doğruluk ve F1 skorlarına sahip olup, obezite sınıflandırma probleminin çözümünde en başarılı sonuçları

vermiştir. Gradient Boosting modeli, %94.79 doğruluk oranı ile en iyi performansa ulaşmıştır.

2) **AdaBoost**, diğer modellerden belirgin şekilde düşük performans göstermiştir. Bu durum, algoritmanın sınıflar arasındaki dengesizlikle başa çıkmakta zorlandığını gösteriyor olabilir.
3) **SVC**, **Lojistik Regresyon** ve **K-Nearest Neighbors (KNN)** modelleri orta seviyede doğruluk değerleri sağlamıştır. Bu modeller, özellik uzayındaki karmaşıklıkla başa çıkmakta sınırlı kalmış olabilir.
4) **Gaussian Naive Bayes**, doğruluk ve diğer metriklerde düşük performans göstermiştir. Naive Bayes'in, özellikle sınıflar arası bağımsızlık varsayımı nedeniyle karmaşık veri yapılarında düşük performans gösterebileceği bilinmektedir.
5) **MLP Classifier** (Çok Katmanlı Algılayıcı), %92.59 doğruluk oranı ile güçlü bir performans sergilemiş ve karmaşık ilişkileri başarılı bir şekilde öğrenmiştir.

Sonuç olarak, Gradient Boosting ve Random Forest modelleri, obezite sınıflandırma probleminin çözümünde en etkili algoritmalar olarak öne çıkmıştır. Bu modeller, verilerdeki karmaşıklık ve sınıflar arasındaki ilişkileri daha iyi kavrayarak, obezite durumunu yüksek doğrulukla tahmin edebilmiştir. Toplu yöntemler (ensemble methods) ve derin öğrenmeye dayalı yaklaşımlar, klasik yöntemlere kıyasla daha yüksek performans sağlamış, özellikle karmaşık veri yapılarını öğrenme yetenekleri sayesinde başarılı sonuçlar elde edilmiştir.

Buna karşın, AdaBoost ve Gaussian Naive Bayes modelleri, veri setinin doğası gereği daha düşük performans göstermiştir. Gelecekteki çalışmalarda, model seçiminde sınıf dengesizliklerinin göz önünde bulundurulması ve daha gelişmiş optimizasyon tekniklerinin kullanılması önerilmektedir.

5. TPOT ile Obezite Tahmini Süreci

TPOT (Tree-based Pipeline Optimization Tool), makine öğrenimi modellemelerini optimize etmek ve hiperparametreleri ayarlamak için genetik algoritmalar kullanan güçlü bir AutoML aracıdır. Bu çalışmada, TPOT ile obezite tahmini gerçekleştirilmiş ve çeşitli makine öğrenimi modelleri otomatik olarak test edilerek en yüksek doğruluk sağlayan model seçilmiştir. Projede kullanılan veri seti, bireylerin demografik bilgileri ve yaşam tarzlarına ilişkin verilerden oluşmaktadır. TPOT sürecinde aşağıdaki adımlar uygulanmıştır:

1. Veri Hazırlığı

Veri seti, yaş, cinsiyet, boy, kilo gibi demografik özelliklerin yanı sıra, beslenme alışkanlıkları, fiziksel aktivite düzeyi ve sigara-alkol tüketimi gibi yaşam tarzı faktörlerini içermektedir. Veri temizleme adımları tamamlandıktan sonra, değişkenler normalleştirilmiş ve uygun veri dönüşümleri uygulanmıştır.

2. TPOT Modeli

TPOTClassifier modeli şu şekilde yapılandırılmıştır:

```
tpot = TPOTClassifier(verbosity=2, generations=5,population_size=50, random_state=42)
```

- **Generations=5:** Genetik algoritma beş jenerasyon boyunca çalıştırılarak en uygun modeller denenmiştir.
- **Population Size=50:** Her jenerasyonda 50 farklı model kombinasyonu test edilmiştir.

TPOT, çeşitli makine öğrenimi algoritmalarını (Random Forest, Logistic Regression, SVC vb.) optimize etmek ve performans açısından en iyi sonucu bulmak için hiperparametre ayarlamaları yapmıştır.

3. Model Optimizasyonu ve Performans

TPOT'un genetik algoritması, beş jenerasyon boyunca model performansını optimize etmiş ve içsel çapraz doğrulama (CV) skorlarına göre en iyi modeller belirlenmiştir. Süreçte, model doğruluk oranı jenerasyonlar boyunca artmış ve nihai jenerasyonda en yüksek performansa ulaşılmıştır:

- **Generation 1:** 0.9549
- **Generation 3:** 0.9609
- **Generation 5:** 0.9680

Nihai model, %95.98 doğruluk oranıyla test verileri üzerinde oldukça başarılı sonuçlar vermiştir. Ek olarak, sınıflandırma raporunda obezite durumlarını tahmin etmede yüksek precision, recall ve F1 score değerleri gözlemlenmiştir.

6. Genel Sonuç: AutoML ve Makine Öğrenimi Modellerinin Karşılaştırılması

Bu çalışmada, TPOT kullanarak gerçekleştirilen AutoML süreci, klasik makine öğrenimi algoritmalarıyla elde edilen sonuçlarla karşılaştırılmıştır. TPOT, genetik algoritmalar kullanarak model seçiminde ve hiperparametre ayarlamalarında manuel süreçlere göre belirgin avantajlar sunmuştur.

AutoML'nin Avantajları:

TPOT ile elde edilen sonuçlar, manuel olarak ayarlanmış makine öğrenimi modellerine kıyasla daha yüksek doğruluk oranları ve performans metrikleri göstermiştir. TPOT'un optimize ettiği modelin doğruluk oranı **%95.98** iken, manuel olarak ayarlanan en başarılı model olan **Gradient Boosting** %94.79 doğruluk oranına ulaşmıştır. Diğer modellerin doğruluk oranları ise daha düşük kalmıştır (Random Forest: %94.32, SVC: %87.22, Logistic Regression: %85.80).

Performans Karşılaştırması:

AutoML ve manuel modellerin performansı karşılaştırıldığında, TPOT'un genetik algoritmalarla optimize ettiği model, özellikle karmaşık veri yapılarını daha iyi öğrenmiş ve obezite tahmininde daha başarılı olmuştur. Precision, recall ve F1 score gibi diğer performans metriklerinde de TPOT'un modelleri, manuel yöntemlere kıyasla daha yüksek sonuçlar vermiştir.

- **TPOT Modeli:** Accuracy: 0.9598, Precision: 0.96, Recall: 0.96, F1 Score: 0.96
- **Gradient Boosting:** Accuracy: 0.9479, Precision: 0.9491, Recall: 0.9479, F1 Score: 0.9478
- **Random Forest:** Accuracy: 0.9432, Precision: 0.9451, Recall: 0.9432, F1 Score: 0.9436

AutoML, daha karmaşık ve zaman alıcı manuel modelleme süreçlerini optimize ederek hem zaman kazandırmış hem de daha yüksek doğruluk oranlarıyla sonuçlanmıştır.

Genel Değerlendirme:

TPOT, makine öğrenimi süreçlerinde insan müdahalesini minimuma indirirken, hiperparametre optimizasyonu ve model seçimi gibi karmaşık görevleri otomatikleştirerek etkileyici sonuçlar ortaya koymuştur. Obezite tahmininde TPOT'un sunduğu AutoML yaklaşımı, manuel olarak oluşturulan modellerden daha başarılı olmuştur. Bu, AutoML'nin gelecekte sağlık gibi kritik alanlarda daha yaygın kullanılabileceğini göstermektedir.