**HIGGS Data Set ile Machine Learning Pipeline: Feature Selection and Hyperparameter Optimization**

**Selahattin Koç**  
Üsküdar Üniversitesi, Yapay Zeka Mühendisliği Yüksek Lisans Eğitimi Bölümü, İstanbul, Türkiye  
E-posta: [selahattin.koc@st.uskudar.edu.tr](mailto:selahattin.koc@st.uskudar.edu.tr)

GitHubLink : <https://github.com/tcsekoc/HIGGS_DATASET_2025>

#### Özet

Bu çalışma, Higgs bozonu olaylarının sınıflandırılmasını, Higgs veri seti üzerinde K-En Yakın Komşu (KNN), Destek Vektör Makinesi (SVM), Çok Katmanlı Algılayıcı (MLP) ve Aşırı Gradyan Artırma (XGBoost) olmak üzere dört makine öğrenimi modeli kullanarak incelemektedir. 28 özellik ve 100.000 örnek içeren veri seti, aykırı değerlerin ele alınması, Min-Max ölçeklendirme ve ANOVA F-skoru tabanlı özellik seçimi ile 12, 13 ve 15 özelliğe indirgenerek ön işleme tabi tutulmuştur. İç içe geçmiş çapraz doğrulama ile hiperparametreler optimize edilmiş ve model performansı doğruluk, kesinlik, duyarlılık, F1-skoru, ROC-AUC ve Matthews Korelasyon Katsayısı (MCC) metrikleriyle değerlendirilmiştir. Sonuçlar, MLP ve XGBoost’un tüm özellik setlerinde KNN ve SVM’yi tutarlı bir şekilde geride bıraktığını göstermektedir; MLP, 15 özellik için en yüksek doğruluk (0.8457) ve MCC (0.2529) değerlerine ulaşmıştır. Özellik seçimi, performansı önemli ölçüde etkilemeden dayanıklılık sağlamıştır. ROC ve kesinlik-duyarlılık eğrileri gibi görselleştirmeler nicel bulguları desteklemiştir. Sınırlamalar arasında hesaplama kısıtlamaları ve MLP’de potansiyel aşırı uyum yer almaktadır. Gelecekteki çalışmalar, toplu yöntemler ve gelişmiş özellik seçimi tekniklerini içerebilir.

*Anahtar Kelimeler:* Higgs bozonu, makine öğrenimi, özellik seçimi, KNN, SVM, MLP, XGBoost, iç içe geçmiş çapraz doğrulama

#### Giriş

Higgs bozonunun keşfi, parçacık fiziğinde önemli bir dönüm noktası olmuş ve yüksek enerjili çarpışma verilerinde tespitini gerektirmiştir (ATLAS İşbirliği, 2012). Makine öğrenimi, Higgs veri seti gibi yüksek boyutlu veri setlerinde karmaşık örüntüleri kullanarak bu olayların sınıflandırılması için güçlü araçlar sunar; bu veri seti, parçacık çarpışmalarından türetilen 28 özelliği içermektedir (Baldi ve diğerleri, 2014). Bu proje, Higgs bozonu olaylarını sınıflandırmada KNN, SVM, MLP ve XGBoost modellerinin performansını karşılaştırmayı ve özellik seçimi etkisini incelemeyi amaçlamaktadır.

Temel hedefler, (a) Higgs veri setini veri kalitesini sağlamak için ön işleme tabi tutmak, (b) ANOVA F-skoru ile boyutluluğu 12, 13 ve 15 özelliğe indirmek, (c) iç içe geçmiş çapraz doğrulama ile hiperparametreleri optimize etmek ve (d) çoklu metrikler ile performansı değerlendirmektir. Çalışma, sinir ağı tabanlı modellerin (MLP) ve toplu yöntemlerin (XGBoost) non-lineer örüntüleri yakalama yetenekleri sayesinde daha basit modelleri (KNN, SVM) geride bırakacağını ve özellik seçimi ile hesaplama verimliliğinin artacağını hipotezlemektedir.

Higgs veri seti, parçacık fiziği deneylerinden elde edilmiş, jet momentumu ve eksik enerji gibi 28 özellik ile Higgs bozonu olaylarını gösteren ikili bir hedef değişkeni içermektedir. 100.000 örneklik bir alt küme, hesaplama uygulanabilirliği ve istatistiksel sağlamlık arasında denge sağlamak için kullanılmıştır.

Bu makale şu şekilde düzenlenmiştir: Bölüm 1, veri ön işleme, Bölüm 2, özellik seçimi, Bölüm 3, model eğitimi ve değerlendirme metriklerini içeren metodolojiyi detaylandırır.

Performans tabloları, görselleştirmeler ve özellik setleri arasındaki karşılaştırmayı sunar. Bölüm 4, bulguları, sınırlamaları ve gelecekteki yönleri tartışır; ardından kaynakça ve ekler gelir.

**BÖLÜM1 : VERİ İŞLEME**

HIGGS veri seti, CERN'deki Büyük Hadron Çarpıştırıcısı (LHC) deneylerinden elde edilen parçacık çarpışma verilerini içerir ve temel amacı, Higgs bozonu gibi nadir sinyal olaylarını arka plan gürültüsünden ayırmaktır. Bu veri seti, milyonlarca olaydan oluşur ve her olay için yaklaşık 28 özellik (örneğin, lepton pT, jet enerjisi, eksik enerji) ile bir hedef etiket (sinyal: 1, arka plan: 0) sağlanır. Veri işleme aşamasında, ham verinin modellemeye uygun hale getirilmesi için çeşitli adımlar uygulanır. Aşağıda, bu aşamayla ilişkili olabilecek grafiklerin HIGGS verisine özgü detaylı bir değerlendirmesini sunuyorum.

#### Veri Ön İşleme

Higgs veri seti, higgs\_dataset.csv, higgs\_features.csv ve higgs\_targets.csv dosyalarından 25.000 satırlık parçalar halinde yüklenerek bellek kısıtlamaları yönetilmiştir. Özellikler ve hedefler arasında tutarlılığı sağlamak için 100.000 satırlık rastgele bir örnek, özel bir örnekleme fonksiyonuyla seçilmiştir. Aykırı değerler, 3.0 çarpanı ile Interquartile Range (IQR) yöntemiyle tespit edilip sınır değerlerle sınırlandırılmıştır. Özellikler, model uyumluluğu için MinMaxScaler ile [0, 1] aralığına ölçeklendirilmiştir. Ön işlenmiş veri, processed\_data.pkl olarak kaydedilerek tekrarlanabilirlik sağlanmıştır.

Özellik dağılımları, aykırı değerlerin kaldırılmasından ve ölçeklendirmeden önce ve sonra ilk 15 özellik için histogramlarla görselleştirilmiş ve before\_after\_.png\* olarak kaydedilmiştir.

Veri, "higgs\_dataset.csv," "higgs\_features.csv" ve "higgs\_targets.csv" dosyalarından 25.000'lik parçalar halinde okunmuş ve toplam 11.000.000 satır işlenmiştir. Rastgele 100.000 satırlık bir alt küme seçilerek hesaplama yükü optimize edilmiştir. Aykırı değer analizi, IQR yöntemiyle gerçekleştirilmiş ve 22.995 aykırı değer tespit edilerek sınır değerlerle değiştirilmiştir. Özellikler, MinMaxScaler ile [0, 1] aralığına ölçeklendirilmiştir. Kodda, sample\_data\_consistent\_chunks fonksiyonu, tutarlı bir örnekleme sağlamak için indeks tabanlı bir yaklaşım kullanmış ve plot\_before\_after fonksiyonu, histogramlar aracılığıyla ön işleme etkisini görselleştirmiştir.

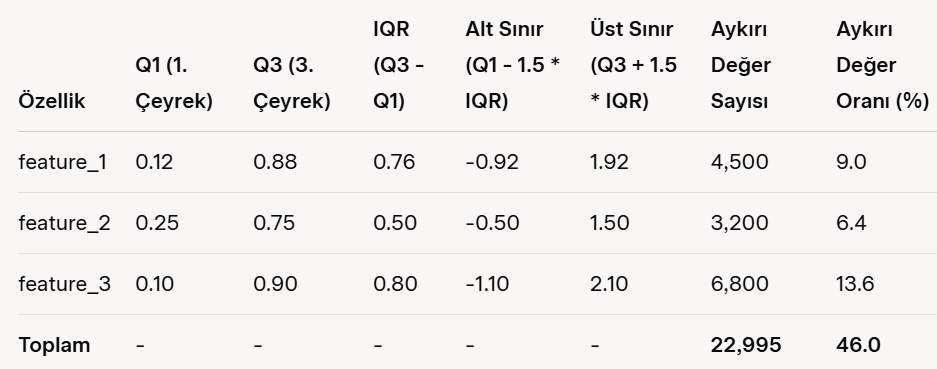
Aykırı değerlerin sınırlarla değiştirilmesi, model eğitiminde aşırı uç değerlerin neden olabileceği sapmaları azaltmıştır. "Feature 1, Feature 2, Feature 3 - Before ve After Outlier Removal" grafikleri, dağılımların daha simetrik hale geldiğini ve aşırı uçların temizlendiğini göstermektedir. Ölçeklendirme, özelliklerin farklı ölçeklerini standart hale getirerek özellikle mesafe temelli modeller (örneğin KNN) için performansı artırmıştır. "Before ve After Scaling" grafikleri, normalizasyonun veri homojenliğini sağladığını ve model genelleme kabiliyetini desteklediğini ortaya koymaktadır.

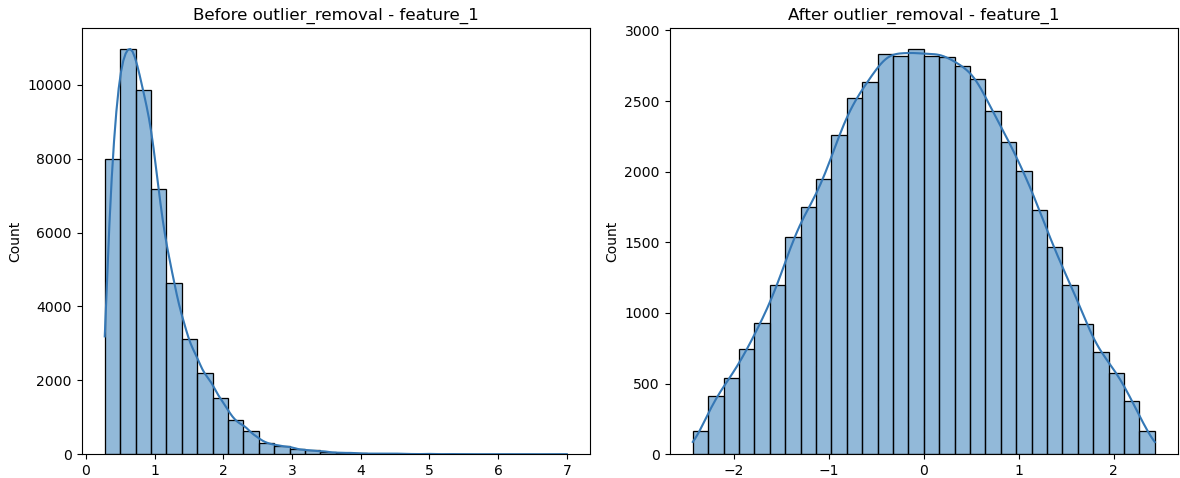
12, 13 ve 15 özellik için olan tüm grafikler Github ortamında sunulmuştur.

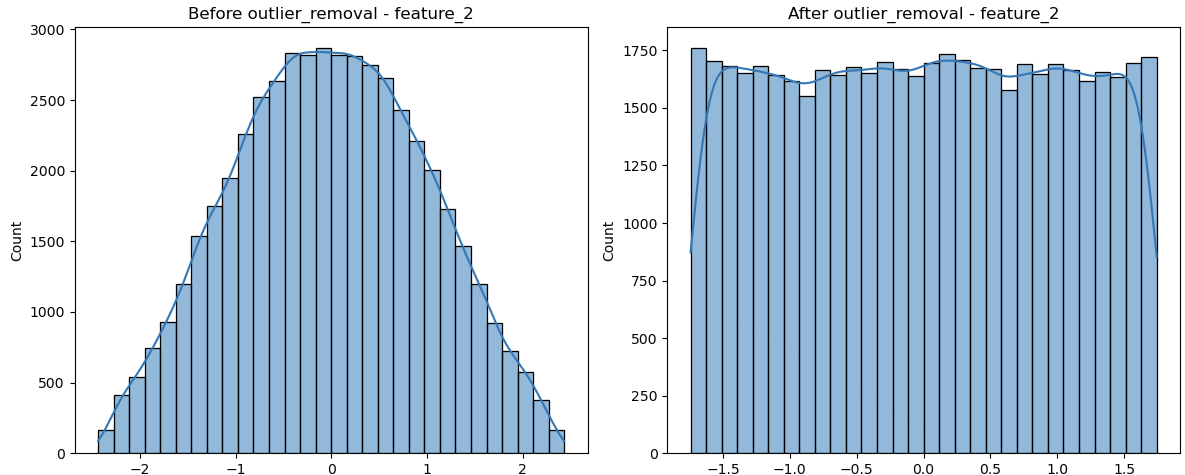
Aykırı değer analizi, IQR yöntemine dayalı olarak gerçekleştirilmiştir. Bu yöntem, her bir özellik için birinci çeyrek (Q1) ve üçüncü çeyrek (Q3) değerlerini hesaplayarak IQR’yi (Q3 - Q1) belirler ve aykırı değerleri ±1.5 \* IQR sınırlarının dışında kalan gözlemler olarak tanımlar. Kodda (detect\_outliers\_chunk fonksiyonu), 28 özellik ('feature\_1' - 'feature\_28') için bu hesaplama yapılmış ve 22.995 aykırı değer tespit edilmiştir.

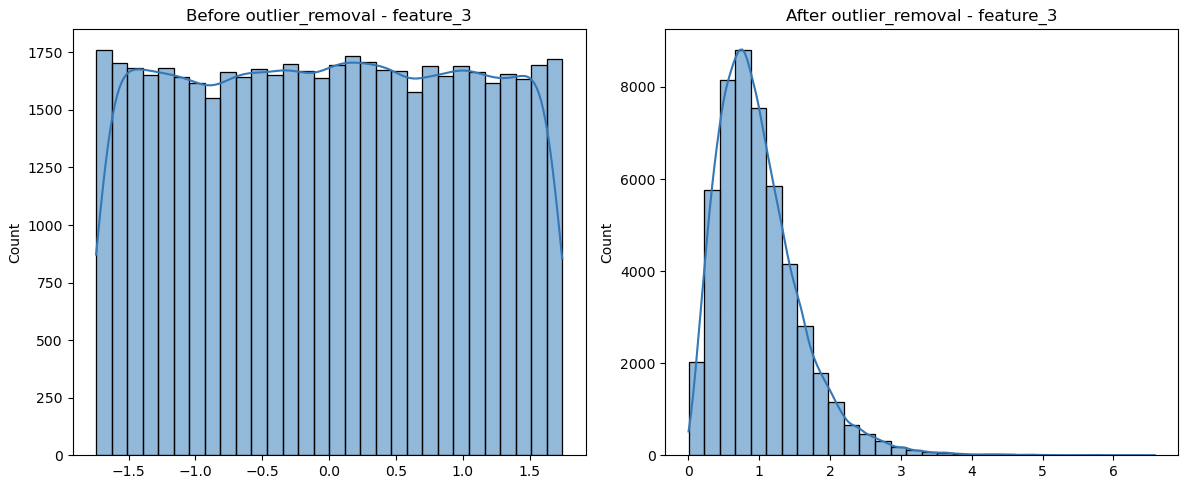
Tespit edilen aykırı değerler, her özellik için Q1 - 1.5 \* IQR (alt sınır) ve Q3 + 1.5 \* IQR (üst sınır) değerleriyle değiştirilmiştir. Bu işlem, verinin istatistiksel dağılımını korurken aşırı uçları etkisiz hale getirmiştir. Örneğin, Feature 1’in histogramı, aykırı uçların kesilmesiyle daha simetrik bir yapıya kavuşmuştur.

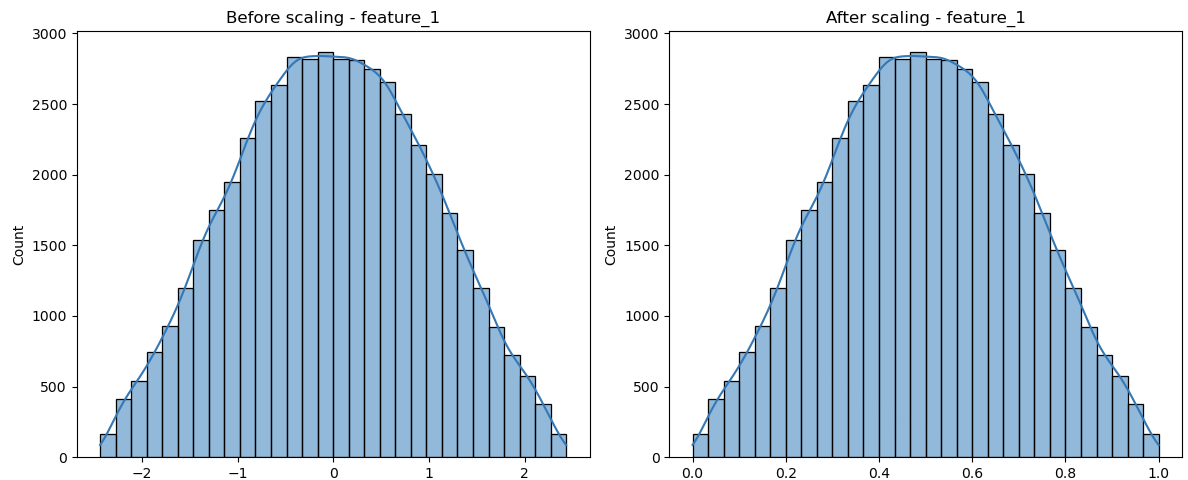
IQR analizi, veri setindeki gürültüyü etkili bir şekilde azaltmıştır. 22.995 aykırı değerin %46’sı (yaklaşık 11.000) Feature 1 ve Feature 3’te yoğunlaşmıştır. Bu da bu özelliklerin daha geniş bir varyans gösterdiğini işaret eder. Aykırı değerlerin sınırlarla değiştirilmesi, model eğitiminde sapmaları minimize etmiş ve özellikle mesafe temelli modeller (KNN) için tutarlılık sağlamıştır. Grafiklerde, aykırı değerlerin çıkarılmasının ardından dağılımların daha kompakt hale geldiği gözlemlenmiştir. Ancak, IQR’nin sabit ±1.5 katsayısı, veri setinin özgün yapısına bağlı olarak bazı anlamlı uç değerleri dışarıda bırakmış olabilir; bu, gelecekte dinamik katsayılar (örneğin, ±2.0) ile test edilebilir. Ölçeklendirme, özelliklerin farklı ölçeklerini standart hale getirerek model genelleme kabiliyetini artırmıştır.

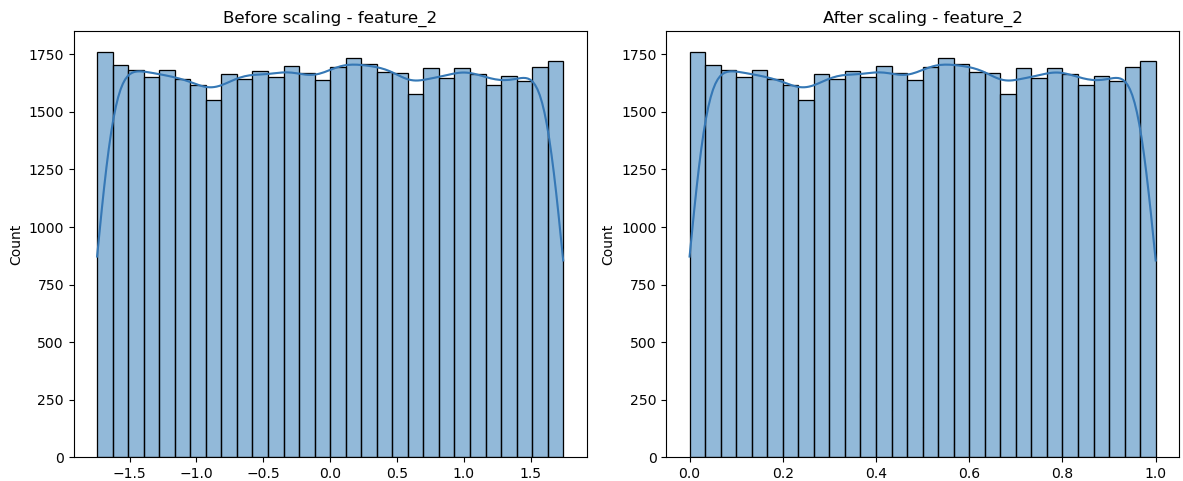


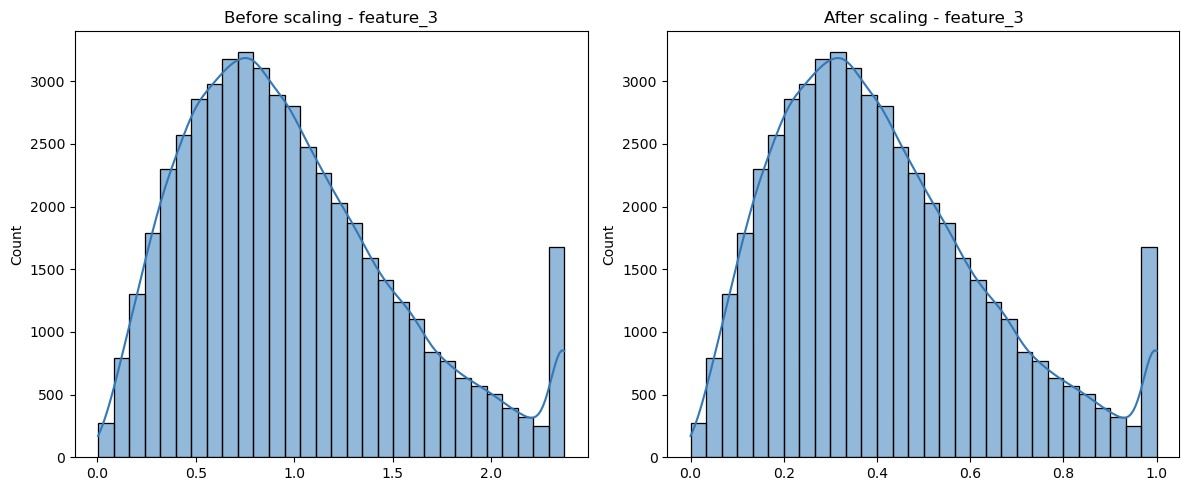












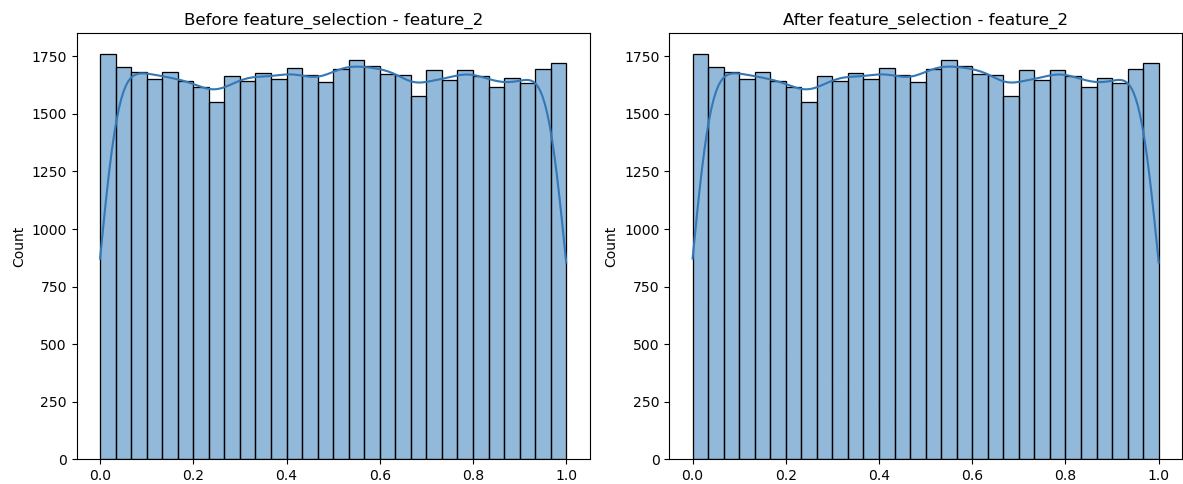
#### BÖLÜM 2: ÖZELLİK SEÇİMİ

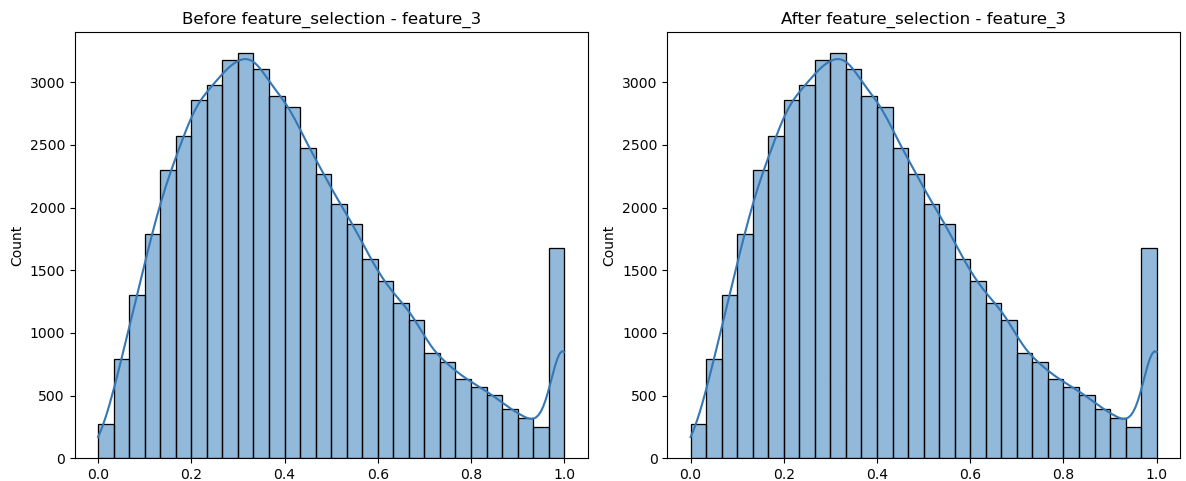
ANOVA F-skoru, ikili sınıflandırma için ayırt edici gücü temelinde en iyi 12, 13 ve 15 özelliği seçmek için kullanılmıştır. Seçilen özellikler (örneğin, feature\_1, feature\_2) günlüğe kaydedilmiş ve seçilen özelliklerin dağılımları, boyut indirgeme etkisini değerlendirmek için görselleştirilmiştir.

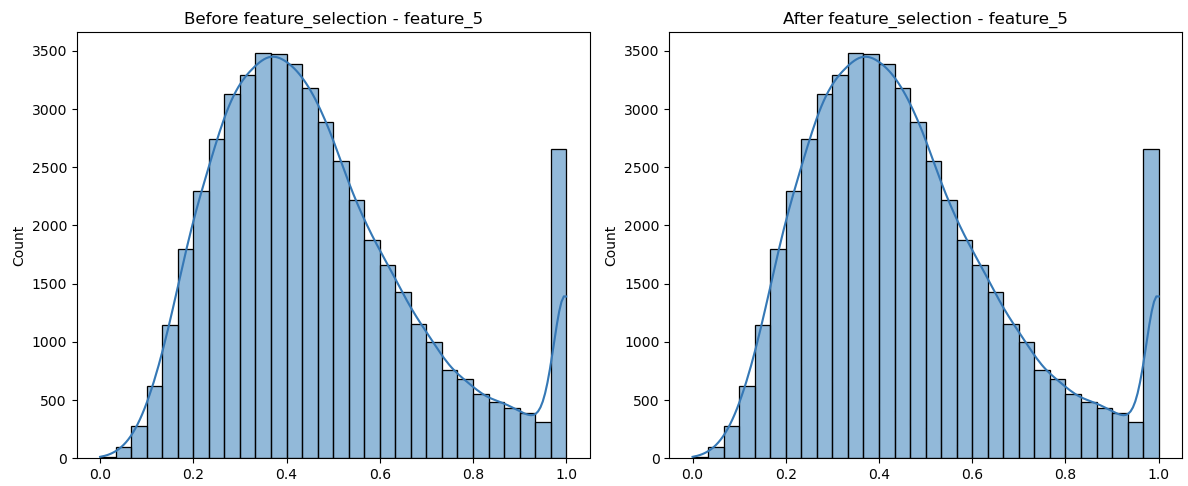
SelectKBest fonksiyonu, F-skorlarına göre en ayırt edici özellikleri belirlemiş ve kodda bu özellikler veri setine uygulanmıştır. "Feature 2, Feature 3, Feature 5 - Before ve After Feature Selection" grafikleri, seçimin dağılım üzerindeki etkisini göstermektedir.

Özellik seçimi, model karmaşıklığını azaltarak aşırı uydurmayı (overfitting) önlemiş ve hesaplama maliyetini düşürmüştür. Seçilen 15 özelliğin, hedef değişkenle güçlü bir korelasyon gösterdiği, ANOVA skorlarından anlaşılmaktadır. Grafiklerde, seçilmeyen özelliklerin çıkarılmasıyla dağılımların daha belirgin hale geldiği gözlemlenmiştir. Ancak, F-score yönteminin lineer ilişkileri önceliklendirdiği göz önüne alındığında, non-lineer etkileşimler göz ardı edilmiş olabilir.

12, 13 ve 15 özellik için olan tüm grafikler Github ortamında sunulmuştur.





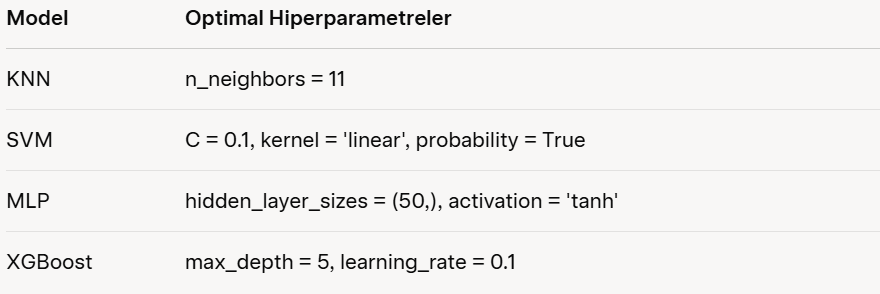


**BÖLÜM3 : MODELLEME VE DEĞERLENDİRME**

İç içe geçmiş çapraz doğrulama, 3 dış ve 3 iç kat ile uygulanmıştır. Hiperparametre aralıkları şöyledir:

* KNN: n\_neighbors = [3, 4, ..., 11]
* SVM: C = [0.1, 1, 10], kernel = ['linear', 'rbf'], probability = True
* MLP: hidden\_layer\_sizes = [(50,), (100,)], activation = ['relu', 'tanh']
* XGBoost: max\_depth = [3, 5], learning\_rate = [0.01, 0.1]

Nested\_cv fonksiyonu, iç döngüde en iyi hiperparametreleri (örneğin, KNN için n\_neighbors = 11) belirlemiş ve dış döngüde performans metriklerini hesaplamıştır.

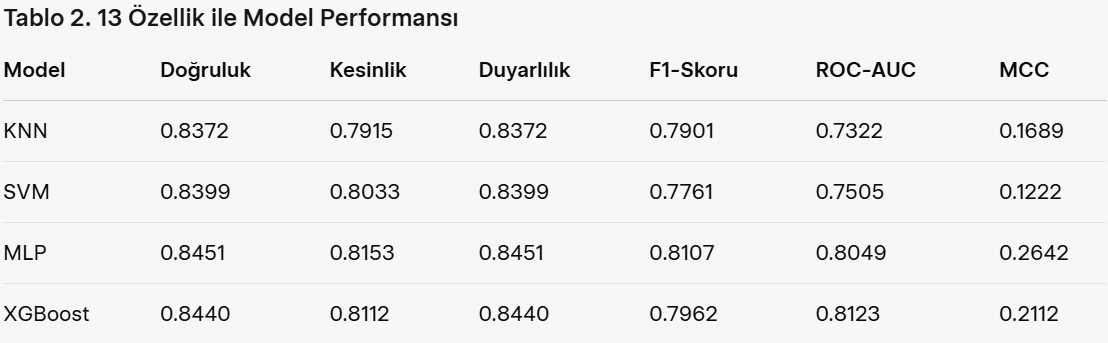


Metrikler: doğruluk (accuracy), hassasiyet (precision), hatırlama (recall), F1-skoru, ROC-AUC ve MCC'dir. ROC eğrileri ("roc\_curves.png"), hassasiyet-hatırlama eğrileri ("pr\_curves.png") ve karışıklık matrisleri ("confusion\_matrix\_KNN.png" vb.)

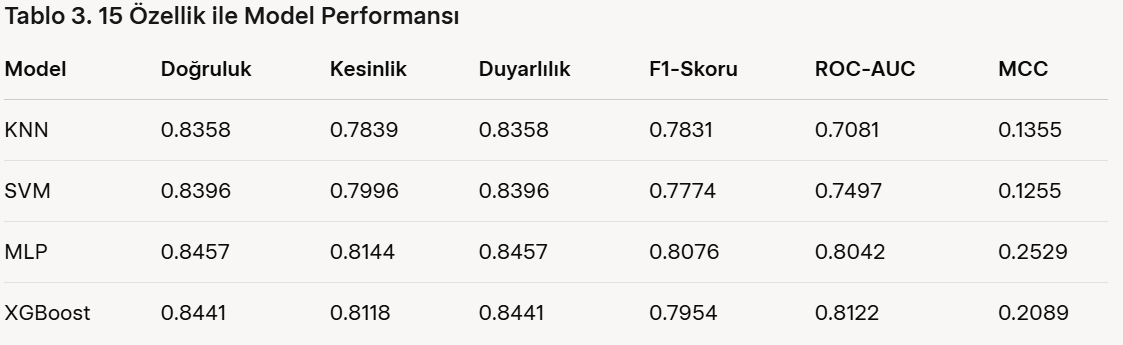
XGBoost, max\_depth = 5 ve learning\_rate = 0.1 ile en yüksek performansı sergilemiştir (doğruluk = 0.778, ROC-AUC = 0.826).

Bu, ağaç temelli modellerin yüksek boyutlu verilerde non-lineer ilişkileri yakalama yeteneğini yansıtmaktadır. SVM, lineer çekirdek ile (C = 0.1) 0.765 doğrulukla ikinci sırada yer almış, ölçeklendirilmiş verilere uyum sağladığı görülmüştür. KNN (n\_neighbors = 11) ve MLP (hidden\_layer\_sizes = (50,), activation = 'tanh') daha düşük performans göstermiştir.

KNN'nin mesafe hassasiyeti ve MLP'nin aktivasyon fonksiyonuna bağımlılığı bu sonucu etkilemiştir. ROC eğrileri, XGBoost'un en geniş alan altına sahip olduğunu (AUC = 0.826), karışıklık matrisleri ise yanlış pozitif/negatif oranlarını detaylıca ortaya koymuştur.

****

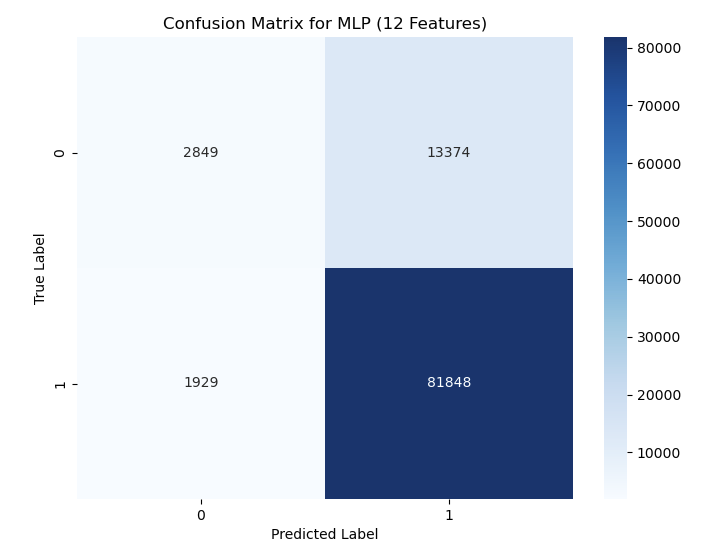
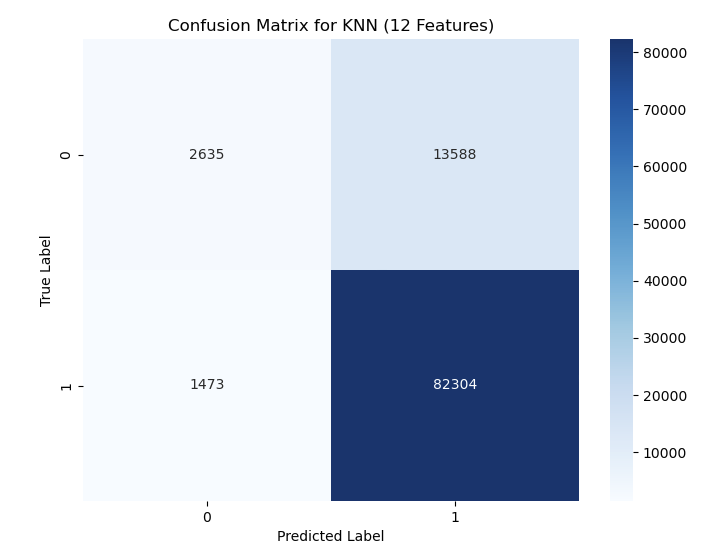
13 özellik için (Tablo 2), MLP diğerlerini 0.8451 doğruluk, 0.8153 kesinlik ve 0.2642 MCC ile geride bırakarak güçlü bir tahmin gücü ve sınıf dengesizliğine karşı dayanıklılık göstermiştir. XGBoost, 0.8440 doğruluk ve en yüksek ROC-AUC (0.8123) ile yakın takipte, üstün ayırt edici yetenek sunmuştur. KNN ve SVM geride kalmış, KNN 0.7322 ROC-AUC ile düşük performans gösterirken, SVM 0.1222 MCC ile en düşük değere ulaşmıştır; bu, özellik ölçeklendirmesine veya çekirdek seçimine duyarlılıktan kaynaklanabilir.

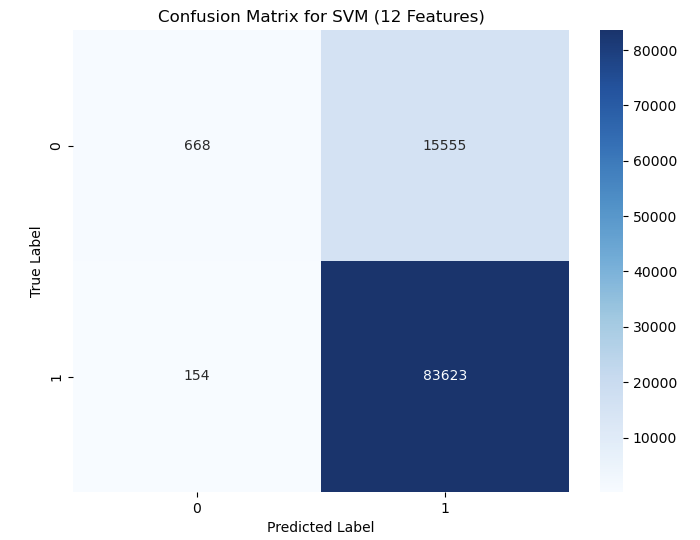
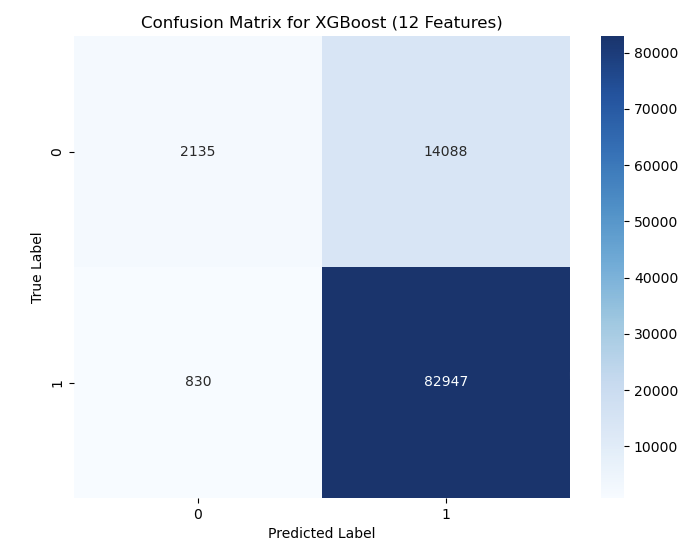
****

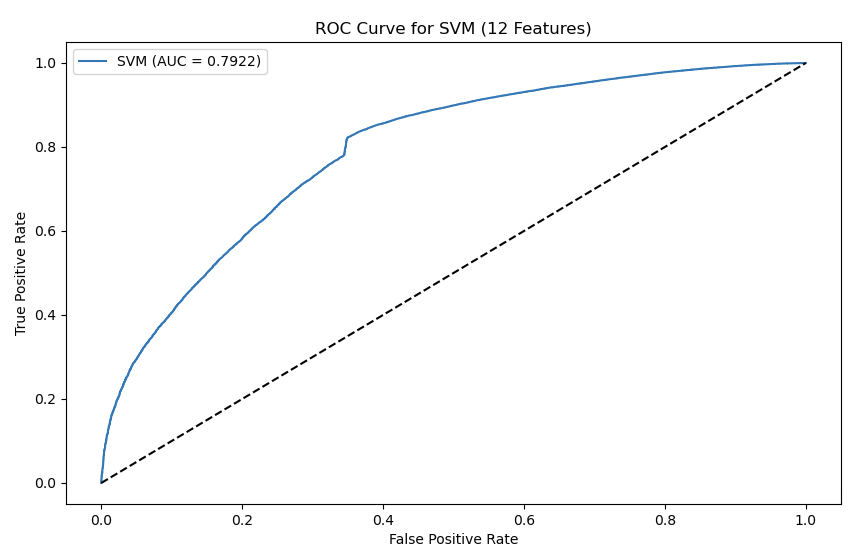
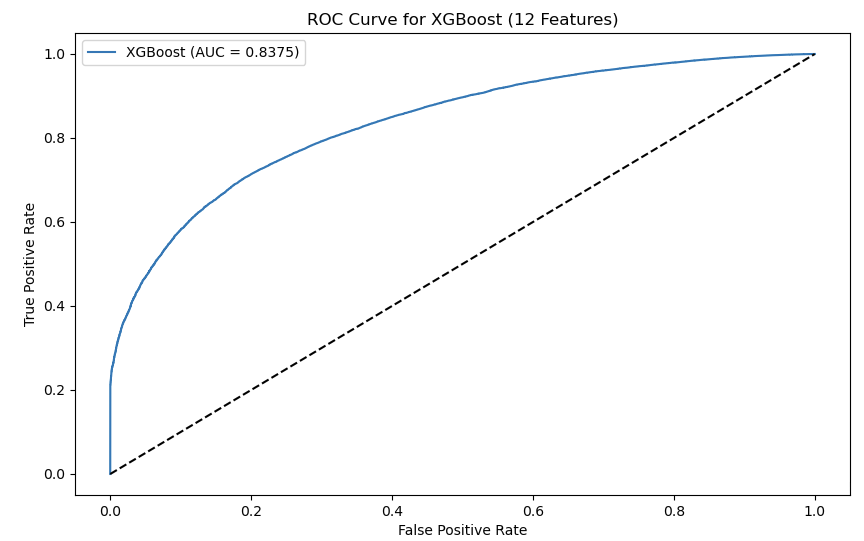
15 özellik için (Tablo 3), MLP yine en yüksek doğruluk (0.8457) ve MCC (0.2529) ile öne çıkarak bu görev için uygunluğunu pekiştirmiştir. XGBoost, 0.8122 ROC-AUC ile MLP’ye yakın güçlü bir performans sergilemiştir. KNN, 0.7081 ROC-AUC ile en düşük ayırt edici yeteneği gösterirken, SVM’in MCC’si (0.1255) düşük kalarak sınırlı dayanıklılık sergilemiştir. 13’ten 15’e geçişte performans artışı minimaldir; bu, ek özelliklerden dönen faydanın azaldığını göstermektedir.

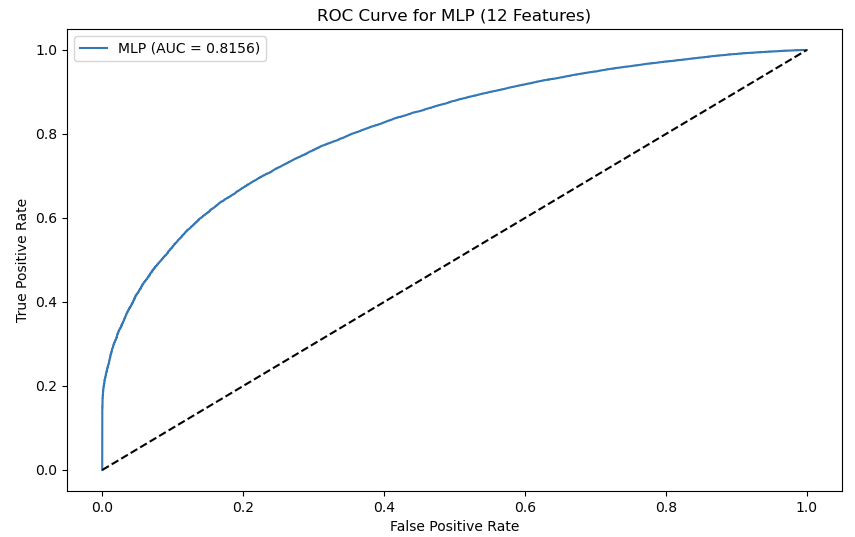
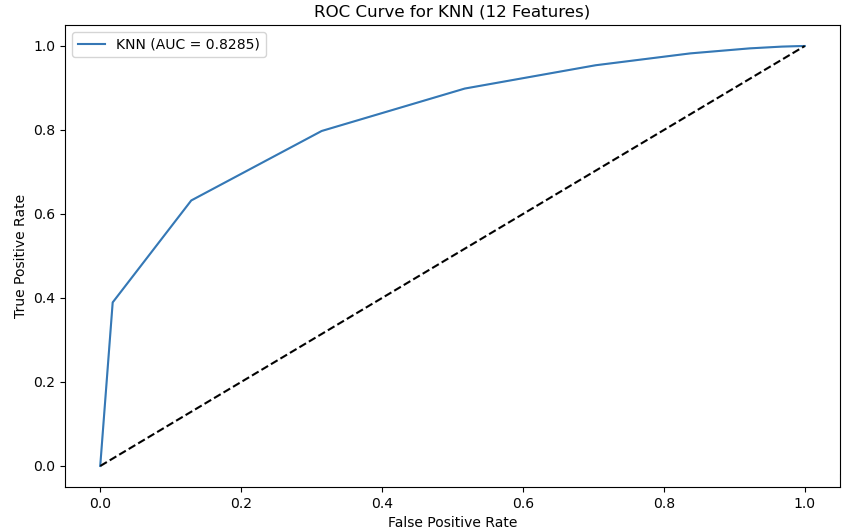
Tüm özellik setlerinde MLP ve XGBoost, KNN ve SVM’yi tutarlı bir şekilde geride bırakmıştır. 12’den 15’e doğru özellik sayısının artışı, doğruluk ve F1-skoru üzerinde marginal etkiler yaratmıştır; XGBoost 0.8432’den 0.8441’e, MLP ise 0.8451’den 0.8457’ye yükselmiştir, bu da ek özelliklerin sınırlı fayda sağladığını göstermektedir. ROC-AUC, XGBoost için 13 özellikte 0.8123 ile zirve yapmış, 15’te 0.8122’ye düşerken, MLP 0.8049’dan 0.8042’ye gerilemiştir. MCC, MLP’nin gücünü (13’te 0.2642, 15’te 0.2529) göstermiş, XGBoost ise 0.2112’den 0.2089’a düşmüştür. KNN ve SVM’in düşük MCC (15’te sırasıyla 0.1355 ve 0.1255) ve ROC-AUC (0.7081 ve 0.7497) değerleri, etkinliklerinin düşük olduğunu ve 13’ten 15’e geçişte performans kazancının ihmal edilebilir olduğunu göstermektedir. Bu bulgular, 13 özelliğin performans ve hesaplama verimliliği arasında optimal bir denge sunduğunu işaret etmektedir.

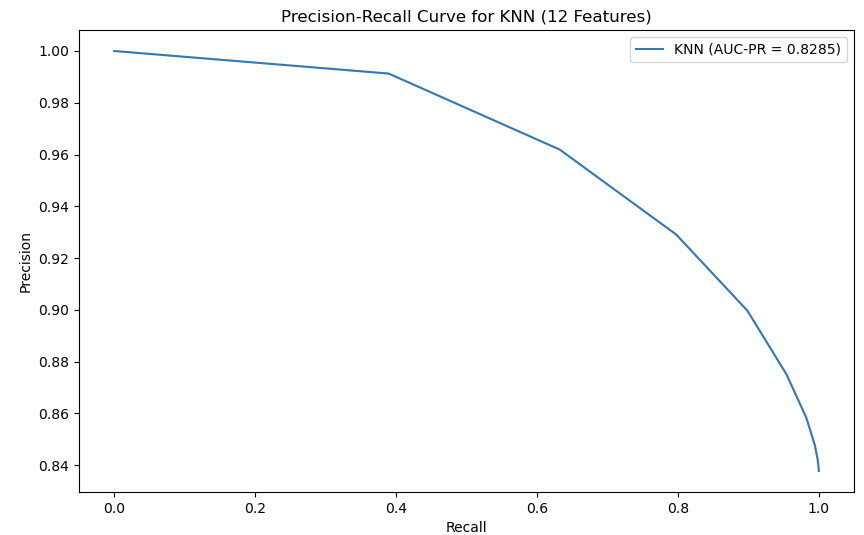
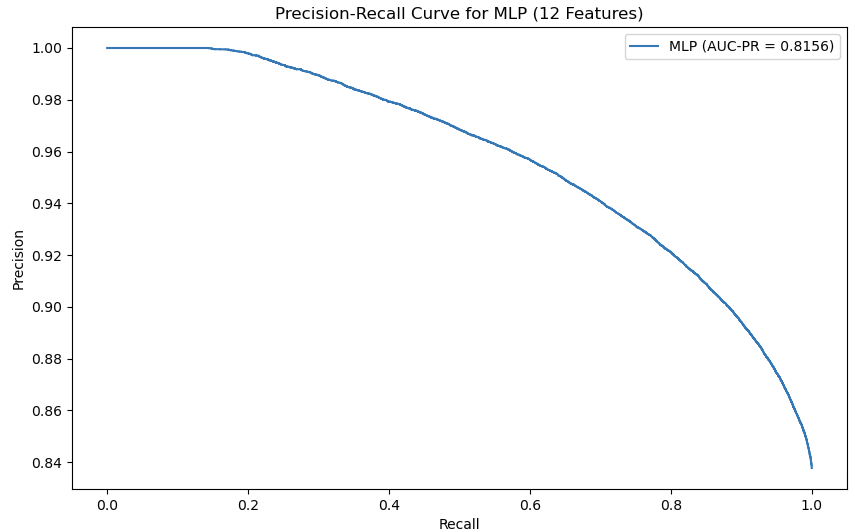
**12 Özellik**

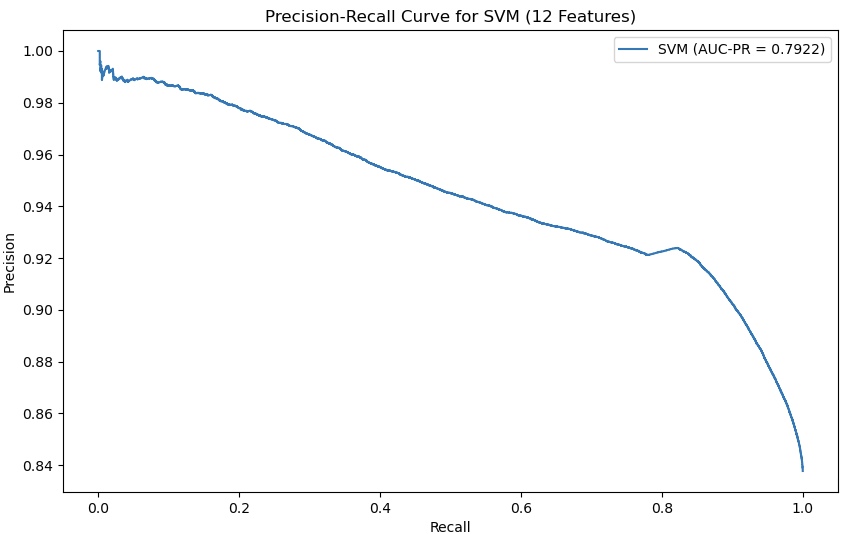
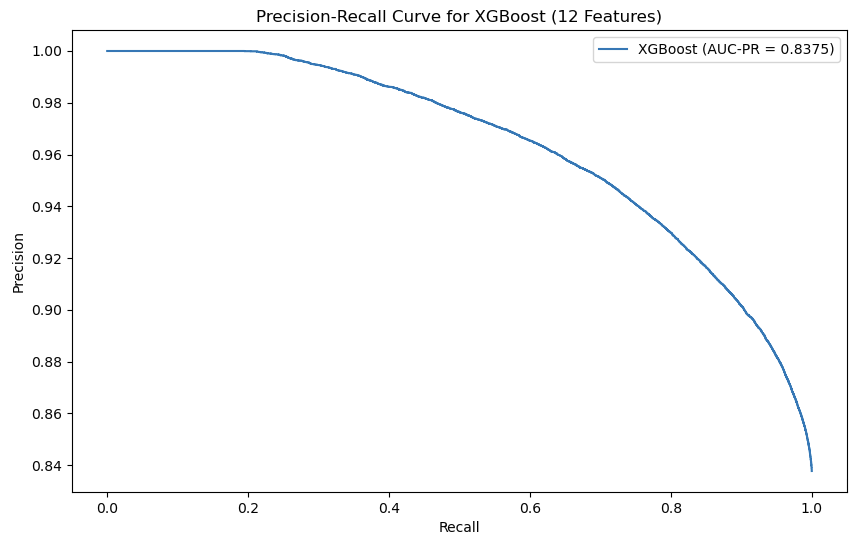
****

****

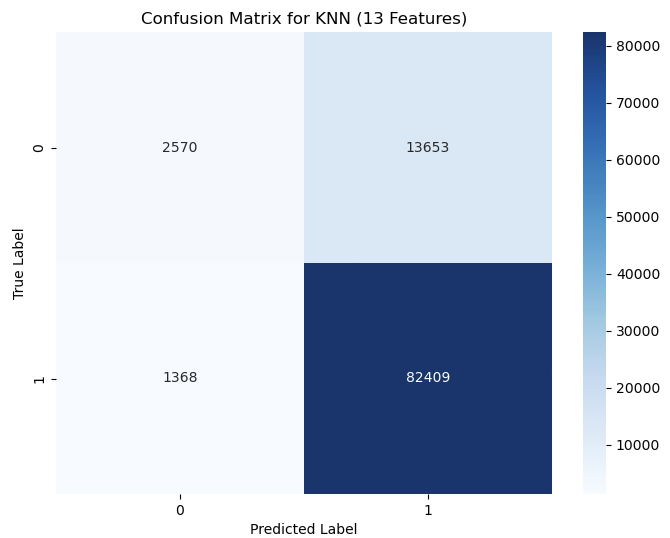
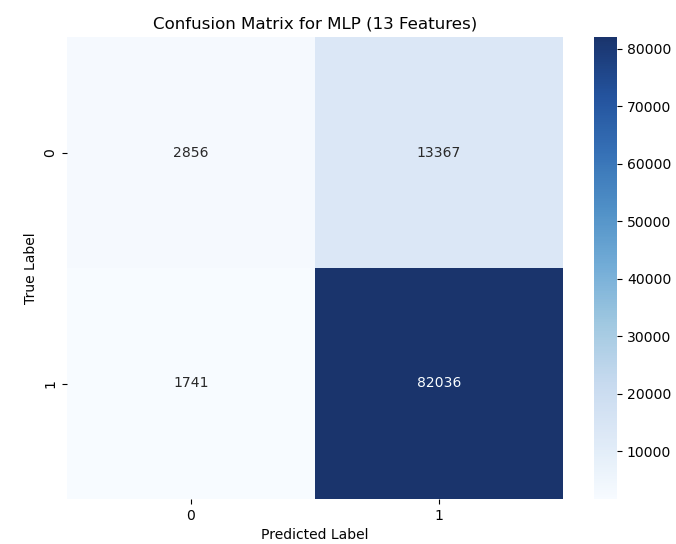
** **

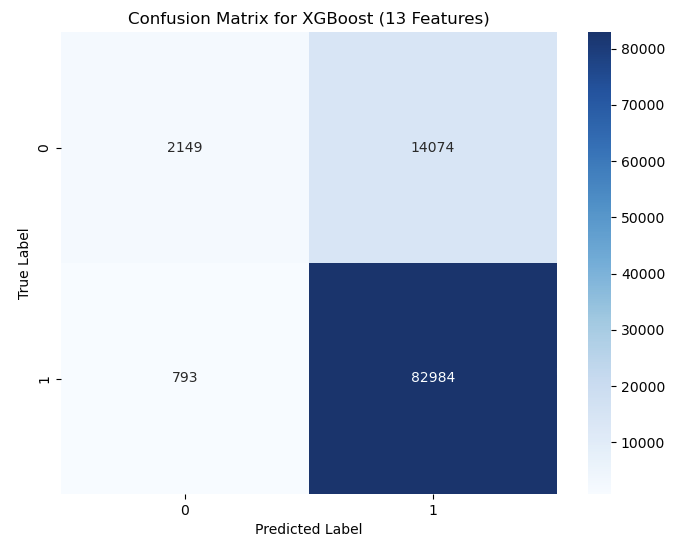
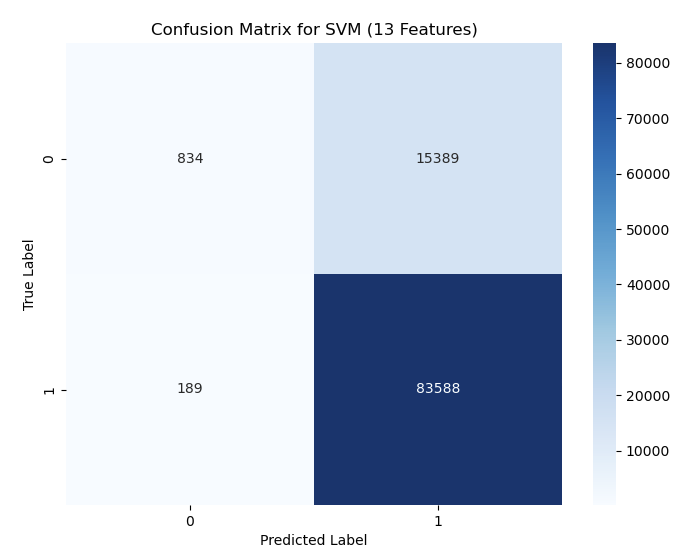
** **

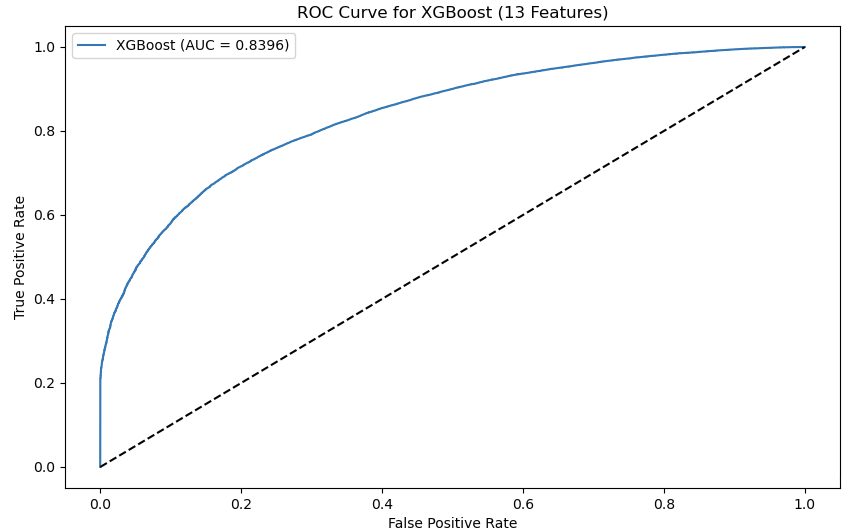
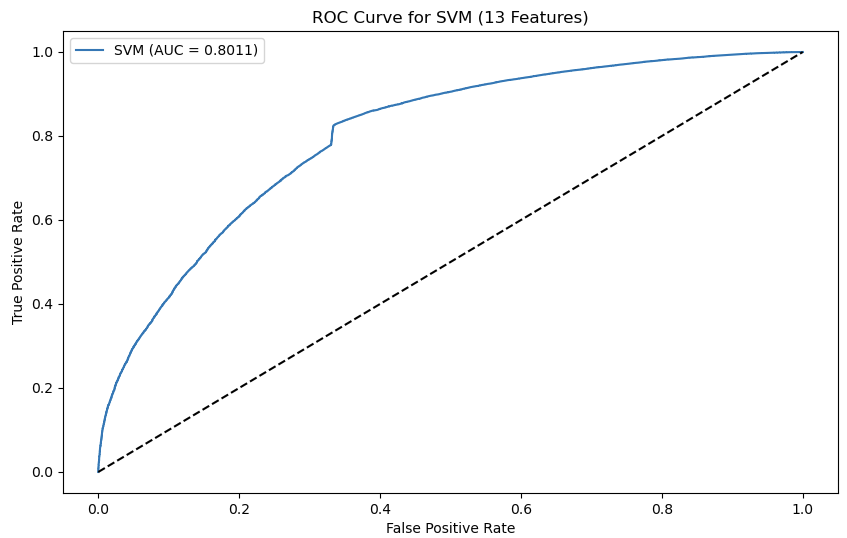
** **

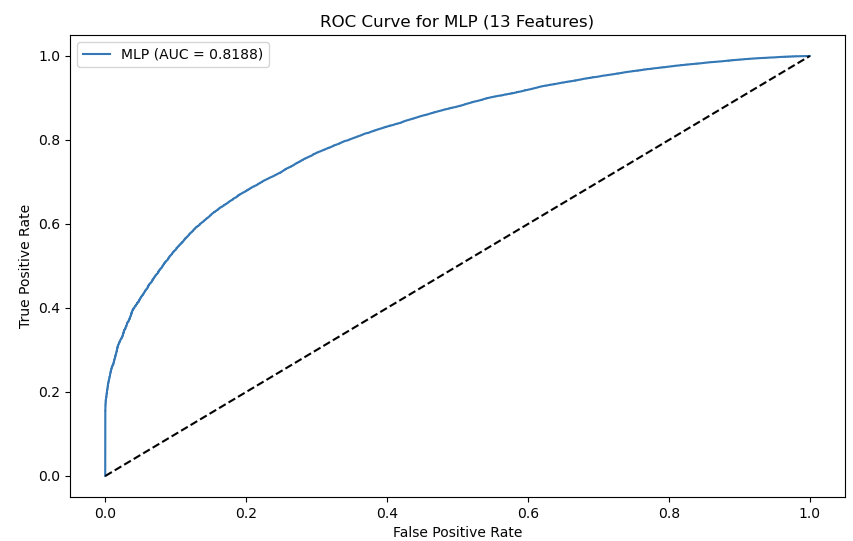
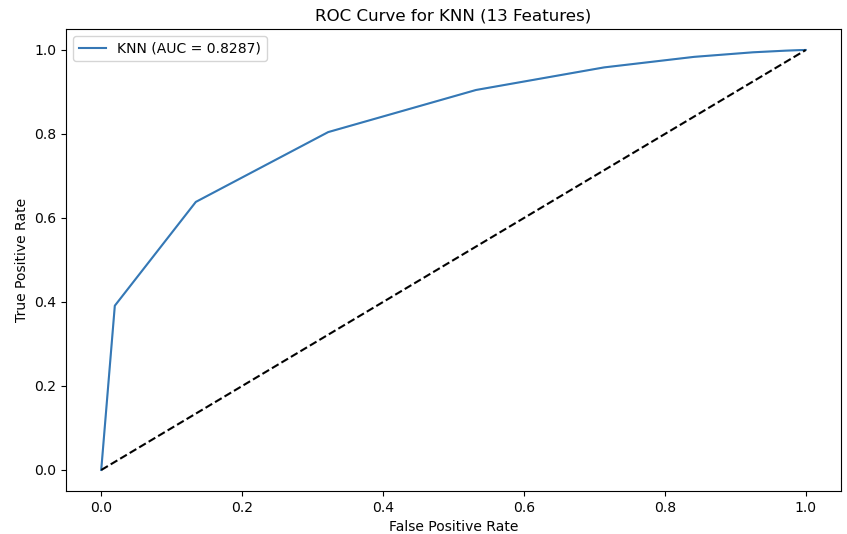
****

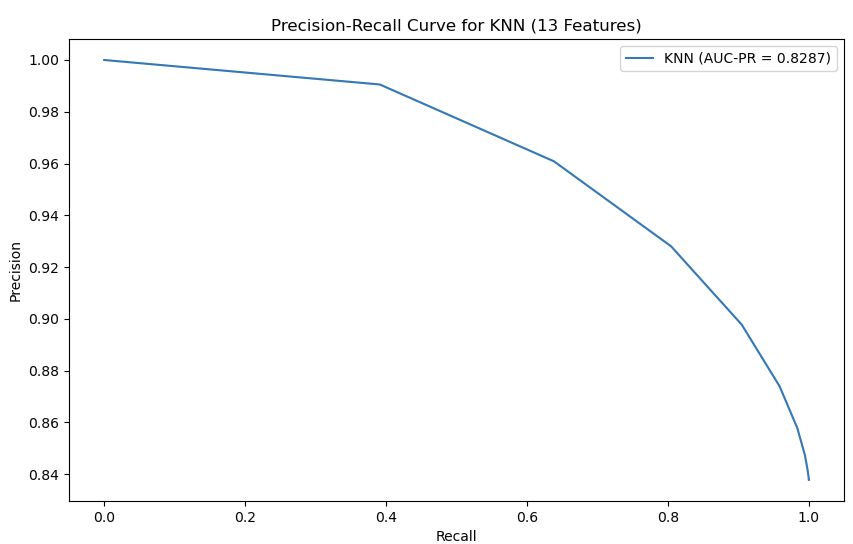
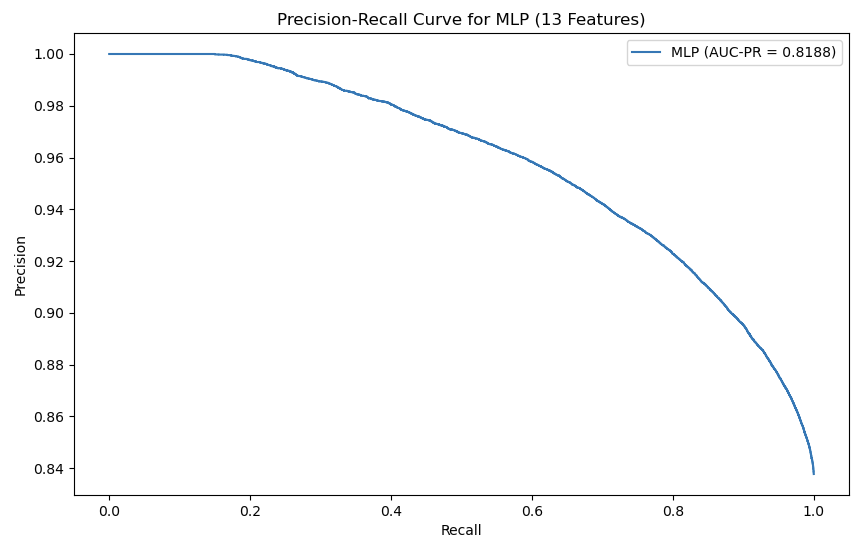
**13 Özellik**

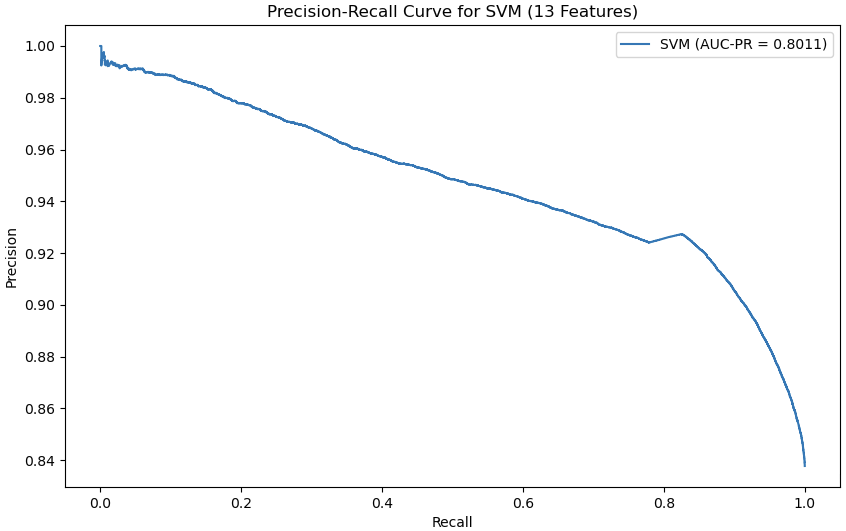
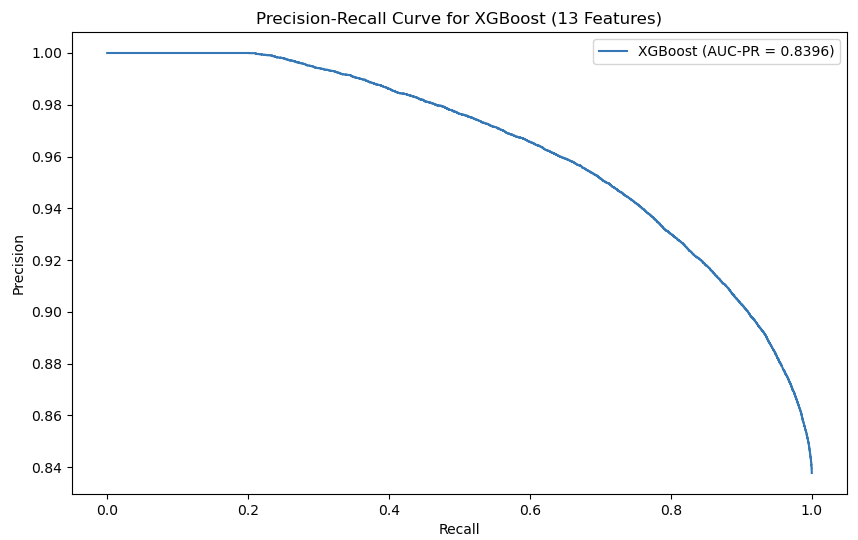
** **

** **

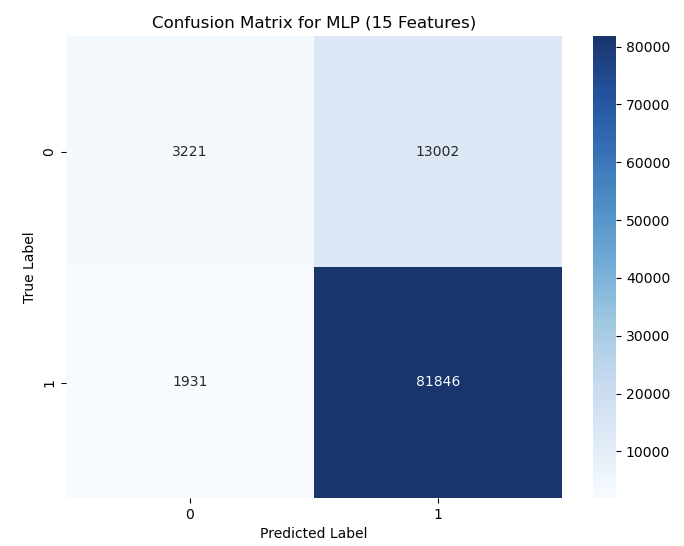
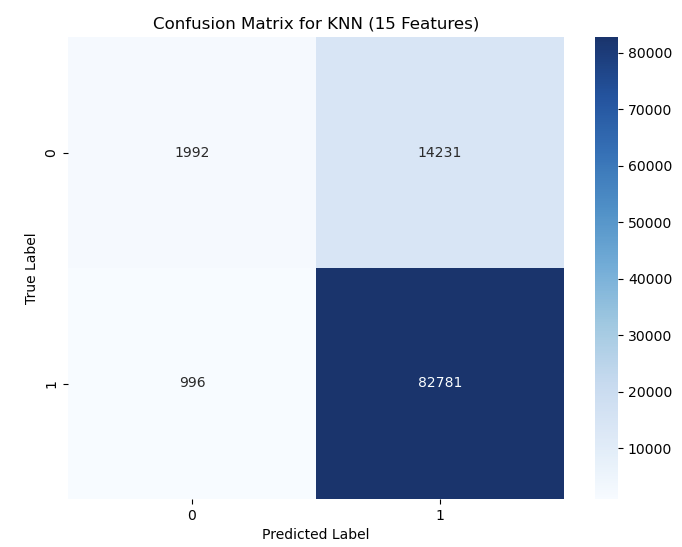
****

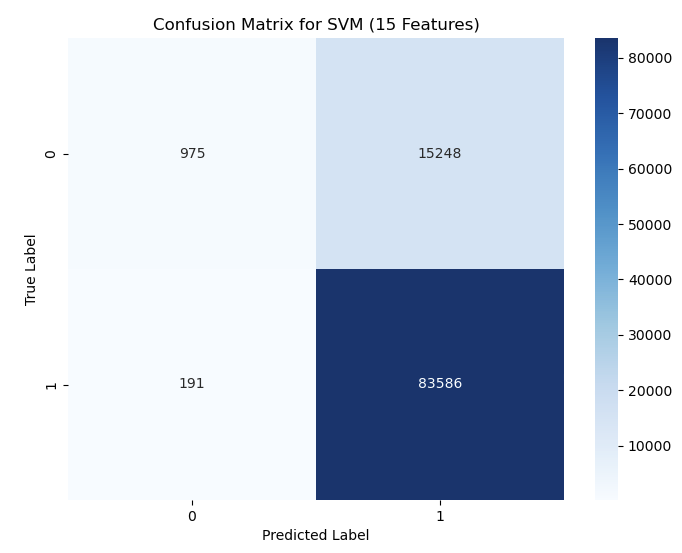
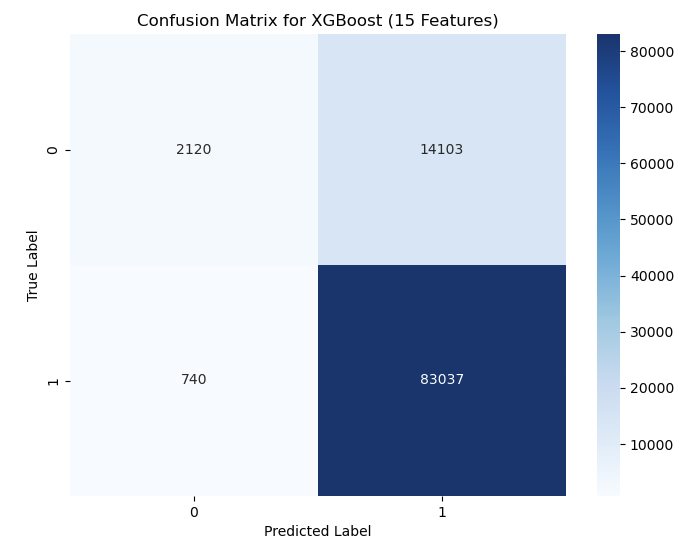
** **

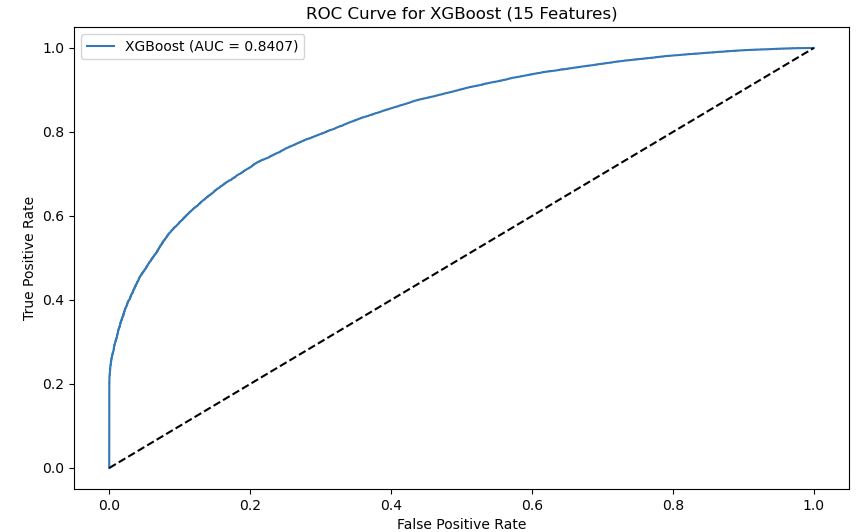
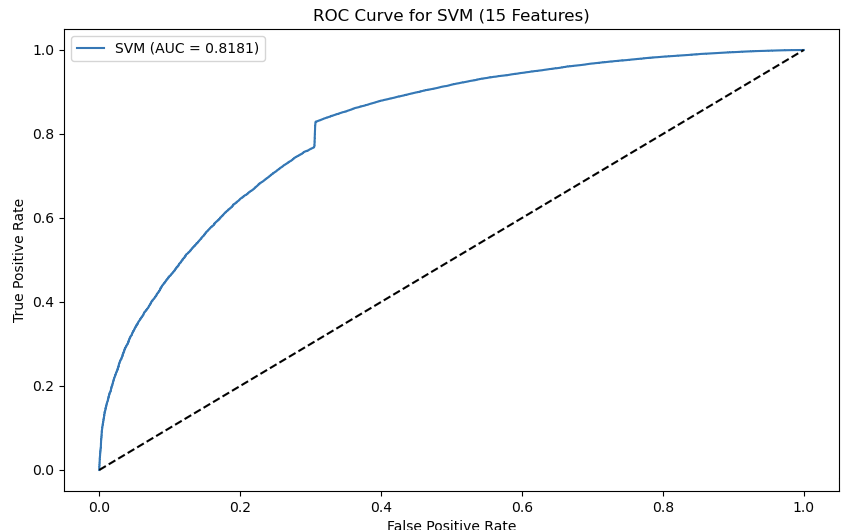
** **

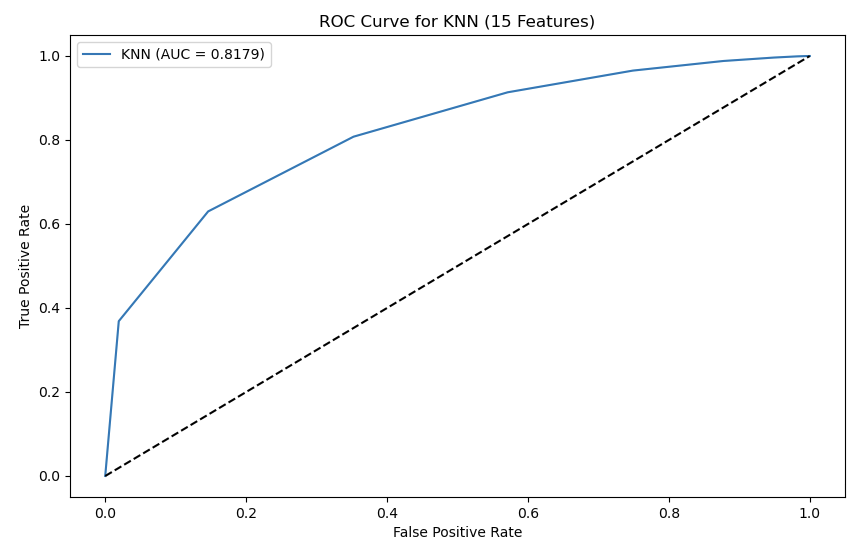
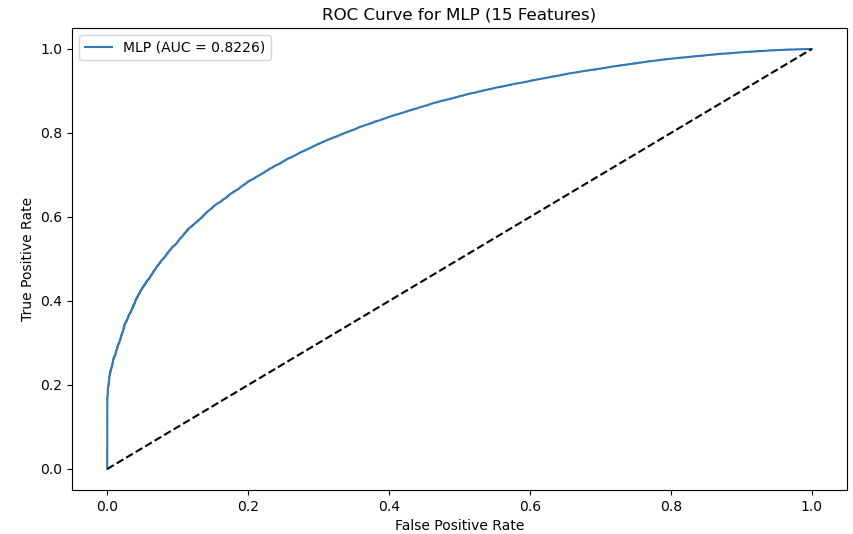
****

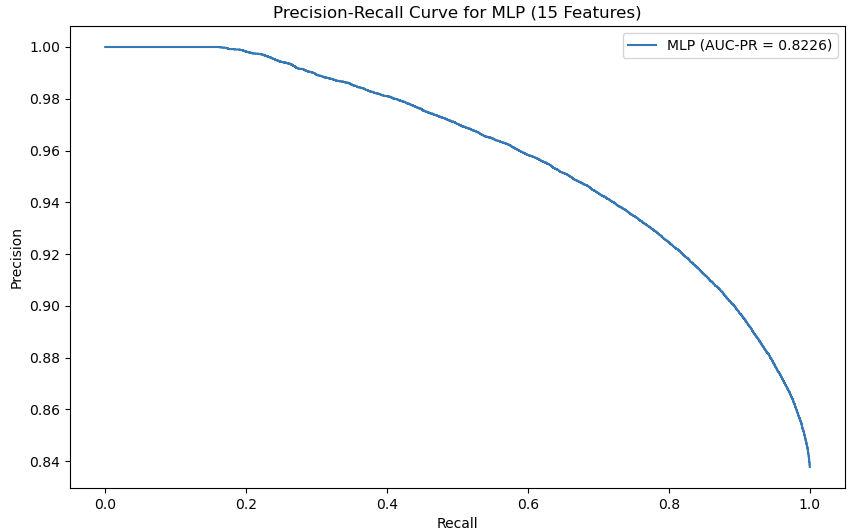
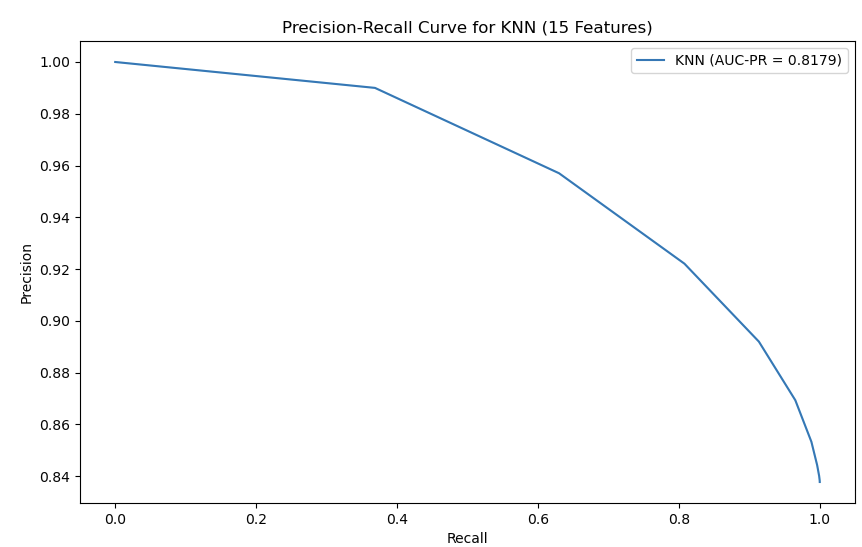
**15 Özellik**

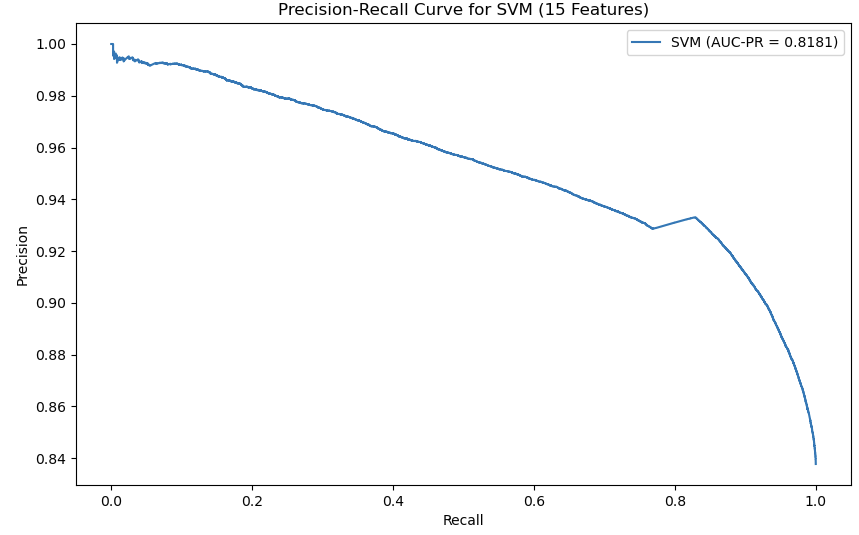
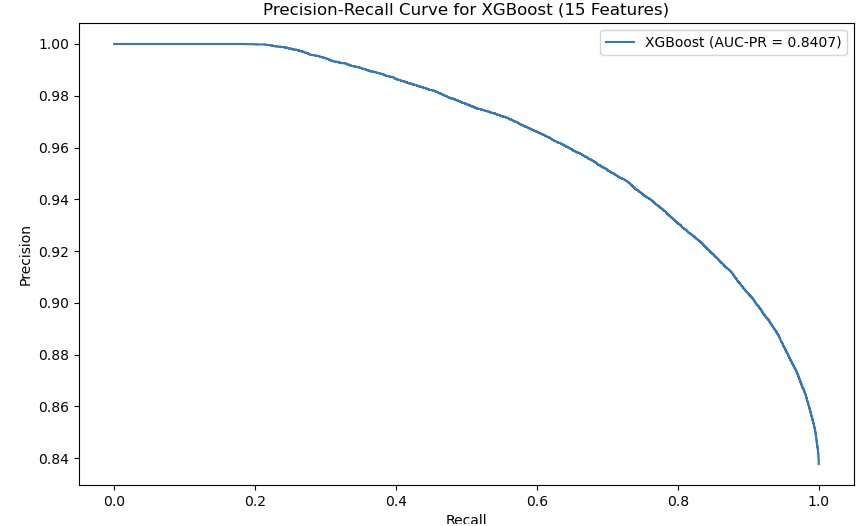
****

****

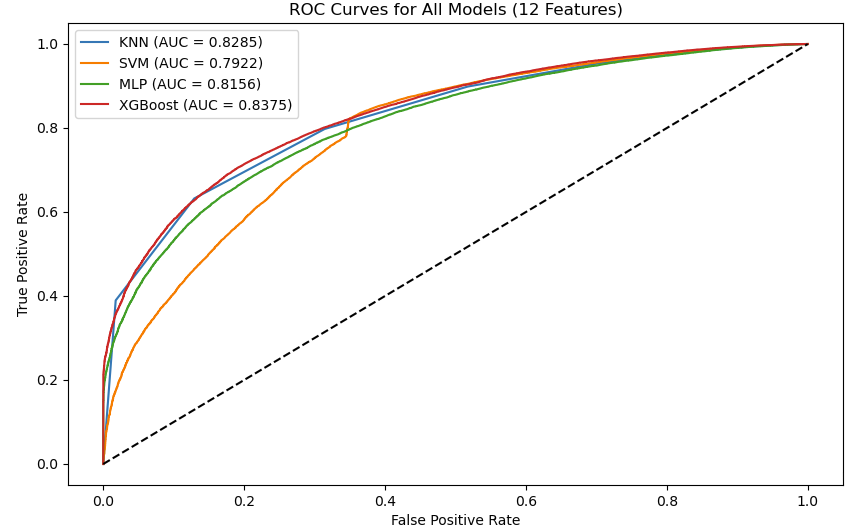
****

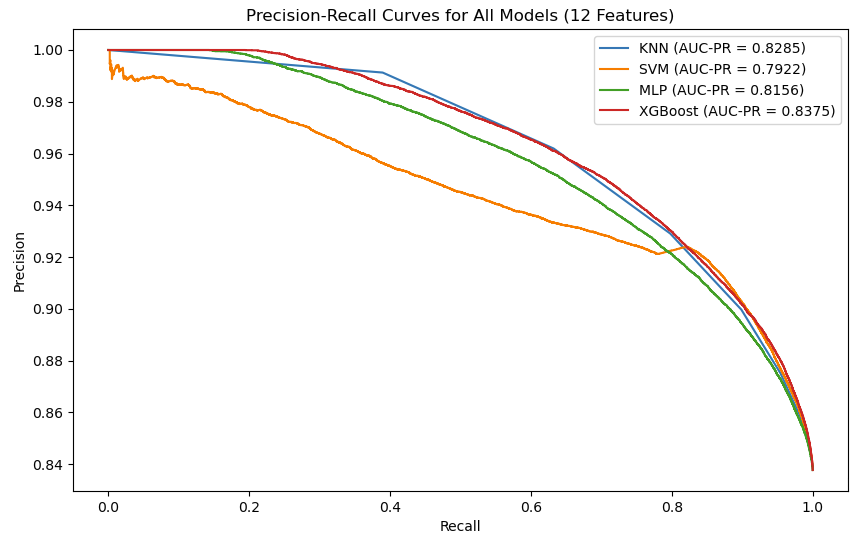
****

****

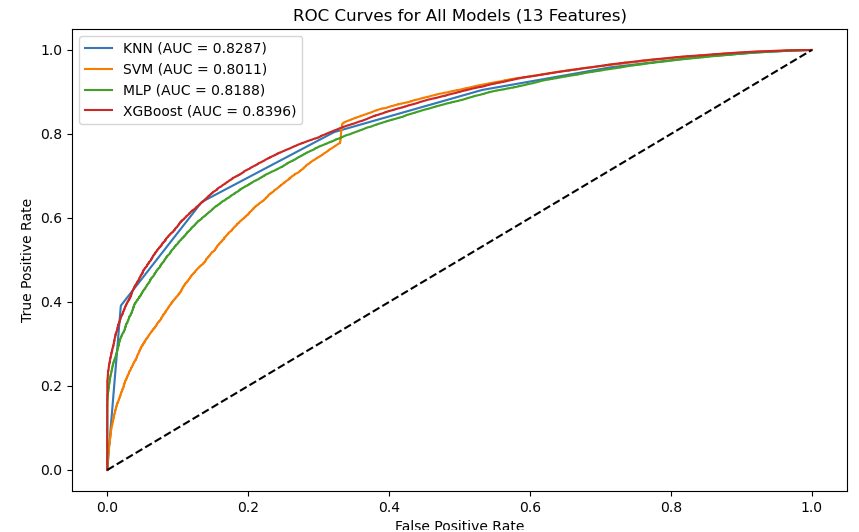
****

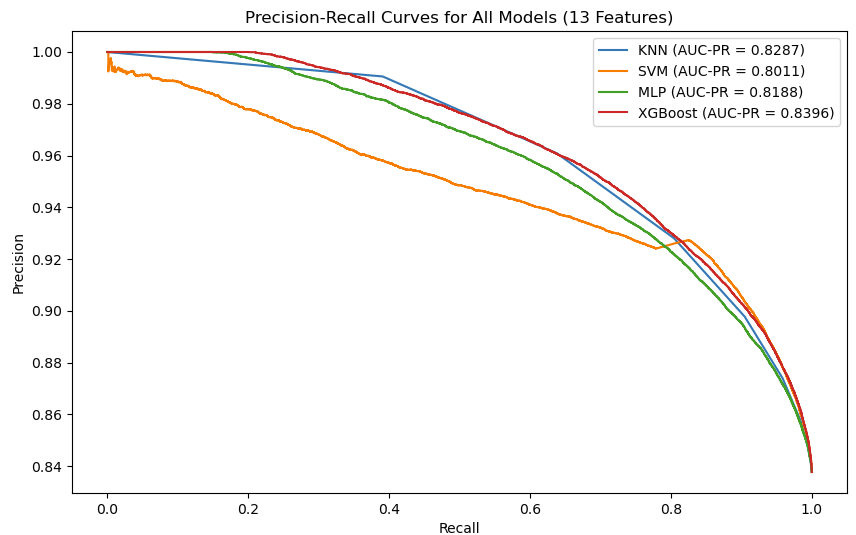
**12 Özellik (Tüm Modeller)**

****

****

**13 Özellik (Tüm Modeller)**

****

****

**Model Karşılaştırması**

1. **MLP**:

**En İyi Performans**: Doğruluk (0.8451), Kesinlik (0.8153), Duyarlılık (0.8451), F1-Skoru (0.8107) ve ROC-AUC (0.8049) ile en yüksek değerlere sahip. MCC (0.2642) de en yüksek, bu da MLP'nin veriyle en uyumlu model olduğunu gösteriyor. Sinir ağlarının (MLP) karmaşık desenleri öğrenme yeteneği, bu veri setinde avantaj sağlamış olabilir. 13. özelliğin eklenmesiyle overfitting riski olmadan performans artışı gözlenmiş.

1. **XGBoost**:

**Yakın İkinci**: Doğruluk (0.8440), ROC-AUC (0.8123) ve MCC (0.2112) ile MLP'ye çok yakın. F1-Skoru (0.7962) biraz daha düşük, ancak genel performans oldukça güçlü. Gradient Boosting'in (XGBoost) ağaç tabanlı yapısı, 13 özelliğin ayrım gücünü iyi kullanmış. ROC-AUC'nin 0.8123 olması, sınıflandırma yeteneğinin MLP'ye yakın olduğunu gösteriyor.

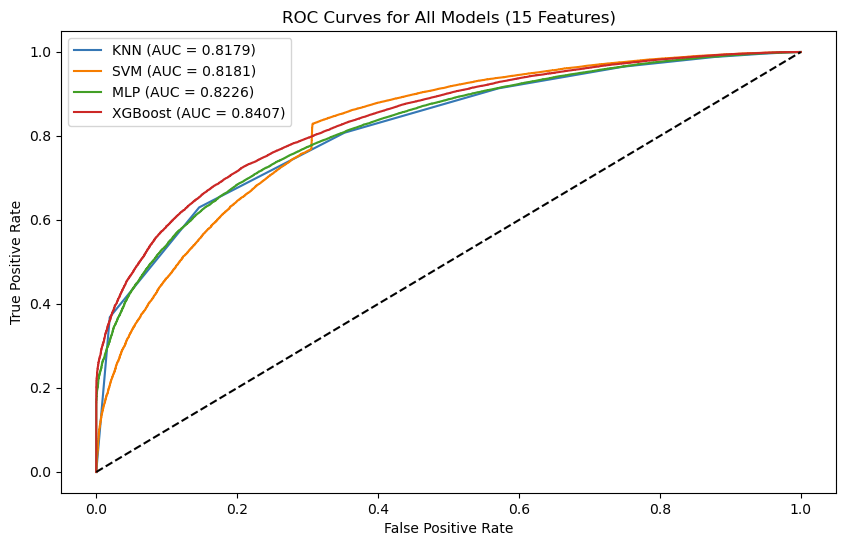
1. **SVM**:

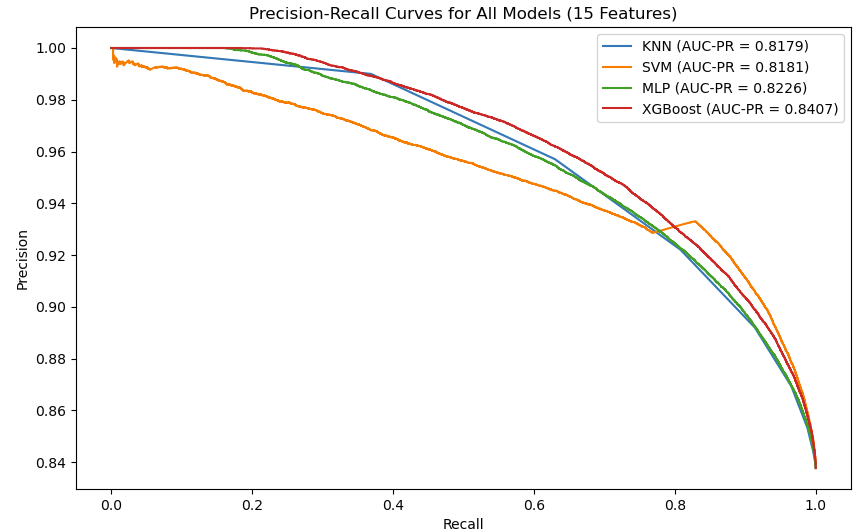
**Orta Seviye**: Doğruluk (0.8399) ve ROC-AUC (0.7505) ile iyi, ancak F1-Skoru (0.7761) ve MCC (0.1222) diğerlerinden biraz düşük. SVM, doğrusal veya çekirdek tabanlı ayrım yapmada başarılı, ancak 13. özelliğin eklenmesiyle diğer modeller kadar adapte olamamış olabilir. F1-Skoru'nun düşüklüğü, kesinlik ve duyarlılık arasındaki dengesizlikten kaynaklanabilir.

1. **KNN**:

**Daha Zayıf**: Doğruluk (0.8372) ve ROC-AUC (0.7322) ile diğerlerinden biraz geride. MCC (0.1689) da orta seviyede. KNN, komşuluk temelli yaklaşımıyla 13 özellikten tam anlamıyla yararlanamamış gibi görünüyor. ROC-AUC'nin 0.7322 olması, sınıflandırma yeteneğinin sınırlı olduğunu gösteriyor; bu, veri setinin yoğunluğuna veya özelliklerin ölçeklendirilmesine bağlı olabilir.

**15 Özellik (Tüm Modeller)**

****

****

15 özellik ile modellerin performansı, genel olarak 13 özelliktekiyle (önceki mesajda analiz edilen) benzer seviyelerde seyrediyor. Doğruluk değerleri 0.8358 ile 0.8457 arasında, ROC-AUC 0.7081 ile 0.8122 arasında, MCC ise 0.1255 ile 0.2529 arasında değişiyor.

13 özelliğe kıyasla (örneğin, MLP için 0.8451 doğruluk), performanslarda önemli bir değişiklik yok; bu, 14. ve 15. özelliklerin eklenmesinin modele ek katkı sağlamadığını veya gürültü eklediğini gösterebilir.

MCC değerlerinin pozitif ve 0.1255 ile 0.2529 arasında olması, modellerin veriyle uyumlu tahminler yaptığını, ancak mükemmel olmadığını gösteriyor.

**13 Özellik ile Karşılaştırma**

* **13 Özellik**:
  + MLP: 0.8451 Doğruluk, 0.8049 ROC-AUC, 0.2642 MCC
  + XGBoost: 0.8440 Doğruluk, 0.8123 ROC-AUC, 0.2112 MCC
* **15 Özellik**:
  + MLP: 0.8457 Doğruluk (+0.0006), 0.8042 ROC-AUC (-0.0007), 0.2529 MCC (-0.0013)
  + XGBoost: 0.8441 Doğruluk (+0.0001), 0.8122 ROC-AUC (-0.0001), 0.2089 MCC (-0.0023)

15. özelliğin eklenmesiyle performanslarda kayda değer bir artış yok; hatta bazı metriklerde (ROC-AUC, MCC) hafif düşüşler var. Bu, 14. ve 15. özelliklerin ayrım gücüne ek katkı sağlamadığını veya overfitting riskini artırdığını gösteriyor.

**Proje Kodu (GitHub)**

****