**Projenin Amacı:**

Bu proje, bir finans kurumunun müşteri kaybını (churn) tahmin etmek amacıyla bir veri toplanmış, bu verilerle, görselleştirme, temizleme, data featuring, EDA vb. yöntemlerle veri analizi yapılmış ve makine öğrenmesi ve derin öğrenme modelleri oluşturularak performanslarını karşılaştırmayı hedeflemektedir. Müşterilerin geçmiş işlemleri, kredi limitleri, işlem sayıları gibi çeşitli demografik ve finansal değişkenler kullanılarak hangi müşterilerin banka ile ilişkisini sürdüreceği veya ayrılacağı tahmin edilmiştir. Bu sayede finans kuruluşları, müşteri kaybını azaltmak için proaktif önlemler alabilir, pazarlama stratejilerini optimize edebilir ve müşteri deneyimini geliştirebilir.

**Veri Setindeki Sütunlar ve Hedef Değişken**

* **Toplam Sütun Sayısı:** 18
* **Hedef Değişken:** Attrition\_Flag (Müşteri durumu: Kaybedilen veya mevcut)

**📌 Özet İstatistikler (Ortalama, Standart Sapma, Min-Max Değerleri)**

* **Yaş (Customer\_Age):** Ortalama **48.75** (Min: 26, Max: 73)
* **Kredi Limiti (Credit\_Limit):** Ortalama **10,063** (Min: 1,438, Max: 34,516)
* **Kredi Skoru (Credit\_Score):** Ortalama **74.85** (Min: 50, Max: 99)
* **Toplam İşlem Tutarı (Total\_Trans\_Amt):** Ortalama **1,256** (Min: 510, Max: 3,231)
* **Toplam İşlem Sayısı (Total\_Trans\_Ct):** Ortalama **27.98** (Min: 10, Max: 57)
* **Kredi Kullanım Oranı (Avg\_Utilization\_Ratio):** Ortalama **0.22** (Min: 0.00, Max: 0.99)

**📌 Veri Türleri ve Dağılımı**

* **Sayısal Değişkenler:** Customer\_Age, Dependent\_count, Credit\_Limit, Total\_Trans\_Amt, Total\_Trans\_Ct, Credit\_Score, vb.
* **Kategorik Değişkenler:** Gender, Education\_Level, Marital\_Status, Income\_Category, Card\_Category, Age\_Group

**📌 İlk Bulgular**

✔ Kategorik değişkenler özgün formatında duruyor, henüz encoding yapılmadı.  
✔ Yaş ve kredi limiti geniş bir dağılım gösteriyor.  
✔ "Blue" kart kullanan müşteriler çoğunlukta.  
✔ Müşterilerin büyük kısmı "Less than $40K" gelir grubunda.  
✔ Veri setinde eksik veri bulunmuyor.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.ekran görüntüsü, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Müşterilerin yaş dağılımlarına baktığımızda normal dağılım gösterdiğini görüyoruz. Bu durum yaş ile alakalı analizlerimizin sağlıklı olabileceği anlamına gelir.

**Mod (En sık görülen yaş):** 50 civarında.

**Minimum - Maksimum:** 26 ile 73 yaş arasında değişmektedir.

**Yorum:** Müşterilerin büyük çoğunluğu **40-60 yaş aralığında** yoğunlaşmıştır. Genç müşteriler azınlıkta.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. ekran görüntüsü, metin, dikdörtgen, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Mod:** En çok 2 veya 3 bağımlısı olan müşteriler mevcut.

**Yorum:** Müşteri kitlesi genellikle **bakmakla yükümlü olduğu kişi sayısı 2-3 olanlardan oluşuyor.** 0 bağımlısı olan müşteriler de önemli bir yüzdeye sahip

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Sağ çarpık (right-skewed) bir dağılım gösteriyor.

**Medyan:** 10.000 civarında.

**Maksimum:** 35.000’in üzerinde bazı aykırı değerler mevcut.

**Yorum:** Kredi limiti düşük müşteriler büyük çoğunluğu oluştururken, az sayıda müşteri oldukça yüksek kredi limitine sahip. **Bu değişken, müşteri segmentasyonu açısından önemli olabilir.**

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, ekran, görüntüleme, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Sol tarafa doğru yayılmış, ancak belirli zirvelere sahip.

**Medyan:** 500-1500 aralığında yoğunlaşmış.

**Maksimum:** 2500’ün üzerinde müşteri sayısı az.

**Yorum:** Çoğu müşteri **borçlarını düzenli ödüyor gibi görünüyor**, ancak belirli bir müşteri segmenti oldukça yüksek borç taşıyor.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, çizgi, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Sağ çarpık (right-skewed).

**Medyan:** 5000 civarında.

**Maksimum:** 30.000’in üzerinde bazı değerler var.

**Yorum:** **Büyük kısmın kullanılabilir kredi limiti düşükken**, bazı müşteriler oldukça yüksek kullanılabilir krediye sahip. **Kredi risk analizinde önemli bir değişken olabilir.**

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Hafif sağa çarpık.

**Mod:** 1000-2000 aralığında yoğunlaşmış.

**Maksimum:** 3500 civarında.

**Yorum:** **Çoğu müşteri 1000-2000 işlem tutarı aralığında alışveriş yapıyor.** Daha fazla harcayan müşteri segmentleri de mevcut.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Normal dağılıma yakın.

**Medyan:** 30 civarında.

**Mod:** 20-50 arasında.

**Yorum:** **Müşterilerin büyük kısmı 20-50 işlem yapıyor.** Daha az işlem yapan müşteriler churn (müşteri kaybı) riski taşıyabilir.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Aşırı sağ çarpık (highly right-skewed).

**Mod:** 0.05 civarında yoğunluk var.

**Maksimum:** 1.0 (Tam kapasite kredi kullanımı).

**Yorum:** **Çoğu müşteri düşük kredi kullanım oranına sahip**, ancak bazı müşteriler kredi limitlerini tamamen doldurmuş. **Bu değişken, kredi risk analizinde önemli olabilir.**

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.ekran görüntüsü, metin, çizgi, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Dağılım:** Düzensiz histogram yapısına sahip.

**Mod:** 70 civarında.

**Maksimum:** 99’a kadar çıkıyor.

**Yorum:** **Çoğu müşteri 60-80 kredi skoru aralığında.** Düşük skorlu müşterilerin risk analizi yapılmalı.



metin, yazı tipi, ekran görüntüsü, grafik içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. daire, grafik, yazı tipi, logo içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. metin, yazı tipi, ekran görüntüsü, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Platinum ve Blue kart sahiplerinin cinsiyet dağılımını gösterir.**

**Platinum kart sahiplerinde kadınlar (%62.2) çoğunlukta.**

**Blue kart sahiplerinde ise erkekler (%62.4) ağırlıklı.**

elektronik donanım, kompakt disk, daire, veri depolama cihazı içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. metin, yazı tipi, ekran görüntüsü, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Müşterilerin eğitim düzeylerine göre dağılımını gösterir.**

**En büyük grup 'Graduate' seviyesindeki müşterilerden oluşuyor.**

**'Post-Graduate' ve 'Doctorate' seviyesinde müşteri sayısı oldukça az.**

elektronik donanım, kompakt disk, daire, veri depolama cihazı içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. metin, yazı tipi, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Müşterilerin medeni durumlarına göre dağılımını gösterir.**

**Evli müşteriler (%55.1) en büyük grubu oluşturuyor.**

**Bekarlar %32.3 oranında yer alırken, boşanmış veya bilinmeyen kategoriler daha düşük oranlarda.**

daire, kompakt disk, veri depolama cihazı, CD içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. 

**Müşterilerin yüzde 57.8’i hala üyeliğini devam ettirirken yüzde 42.2’si banka üyeliğini iptal ettirmiş durumda**

**Bankan müşterilerini elinde tutmak için daha fazla kampanya veya indirimler yapmayı düşünebilir.**

**elektronik donanım, kompakt disk, metin, daire içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.metin, yazı tipi, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**Müşterilerin gelir gruplarına göre dağılımını gösterir.**

**En büyük grup 'Less than $40K' (Düşük gelir grubu).**

**$80K - $120K aralığında önemli bir yoğunluk var.**

metin, ekran görüntüsü, kare, paralel içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Sayısal değişkenler arasındaki ilişkiyi gösterir.**

**En yüksek pozitif korelasyon: Toplam işlem tutarı ve işlem sayısı (%70).**

**Negatif korelasyon: Kredi limiti ve kredi kullanım oranı (%-46).**

**Kredi borcu arttıkça kullanılabilir kredi limitinin düştüğü görülüyor.**

**1. Veri Temizleme (Data Cleaning)**

a) Eksik ve "Unknown" Değerlerin Ele Alınması

* Eksik veriler tespit edildi ve sayısal değişkenlerde medyan ile, kategorik değişkenlerde ise en sık tekrar eden değer ile dolduruldu.
* "Unknown" değerler eksik veri olarak kabul edilerek uygun şekilde dolduruldu.
  + Kategorik değişkenlerde en sık tekrar eden değerle değiştirildi.
  + Sayısal değişkenlerde ortalama ile dolduruldu.

**b) Aykırı Değerlerin (Outliers) Temizlenmesi**

* IQR (Interquartile Range) yöntemi kullanılarak aykırı değerler belirlendi.
* Aykırı değerler capping yöntemi ile alt ve üst sınırlar belirlenerek düzenlendi.
* Bu sayede veri kaybı olmadan verinin dengesi sağlandı.

**2. Özellik Mühendisliği (Feature Engineering)**

a) Kategorik Verilerin Dönüştürülmesi

* Label Encoding yerine One-Hot Encoding (OHE) yöntemi kullanıldı.
  + Bu sayede modelin kategorik verileri daha iyi anlaması ve doğru şekilde işlemesi sağlandı.
  + Gereksiz kategori fazlalığını önlemek için drop\_first=True parametresi kullanıldı.

**b) Gereksiz Değişkenlerin Kaldırılması**

* "Age\_Group" sütunu kaldırıldı, çünkü aynı bilgiyi içeren "Customer\_Age" sütunu zaten mevcut.
* "CLIENTNUM" sütunu kaldırıldı, çünkü her müşteri için benzersiz bir ID içerdiğinden modelleme için anlamlı değildir.

**c) Sayısal Değişkenlerin Standartlaştırılması**

* StandardScaler kullanılarak tüm sayısal değişkenler standart normal dağılıma dönüştürüldü.
  + Bu işlem, modelin farklı ölçeklerdeki veriler arasında adil bir hesaplama yapmasını sağlar.
  + İşlemin efektif olması adına One Hot Encoding metodu kullanılan sütunlara özellikle standardizasyon işlemi kullanılmadı.

**3. Boyut Azaltma ve Korelasyon Analizi**

**a) Yüksek Korelasyonlu Değişkenlerin Kaldırılması**

* Multicollinearity (Çoklu doğrusal bağımlılık) sorununun önüne geçmek için korelasyon analizi yapıldı.
* 0.85’ten yüksek korelasyon katsayısına sahip değişkenler belirlenerek veri setinden çıkarıldı.
* Bu işlem, modelin gereksiz bilgileri öğrenmesini engelleyerek daha yüksek performans göstermesini sağlar.

**4. Temizlenmiş Verinin Kaydedilmesi**

* Tüm veri temizleme ve özellik mühendisliği işlemleri tamamlandıktan sonra veriler cleaned\_data.csv olarak kaydedildi.
* Bu sayede modelleme aşamasında veri ön işleme adımlarını tekrar etmeye gerek kalmamaktadır

**Support Vector Machine (SVM) Modeli:**

**Kullanılan Hiperparametreler:**

Hiperparametre optimizasyonu için GridSearchCV yöntemi ile aşağıdaki hiperparametreler denenmiştir:

* C: [0.1, 1, 10, 100] → Ceza parametresi
* Kernel: ['linear', 'rbf', 'poly', 'sigmoid'] → Çekirdek fonksiyonları
* Gamma: ['scale', 'auto'] → Kernel katsayısı

Her kombinasyon için 5 katlı çapraz doğrulama (cross-validation) uygulanmıştır.

**En İyi Model ve Sonuçları**

GridSearch sonucunda en iyi model aşağıdaki hiperparametrelerle bulunmuştur:

* En iyi hiperparametreler:
  + C = 1
  + Kernel = 'rbf'
  + Gamma = 'auto'

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Confusion Matrix:**

Doğru Negatif: 37 Yanlış Pozitif: 8

Yanlış Negatif: 2 Doğru Pozitif: 60

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**Learning Curve Yorumu**

Bu öğrenme eğrisi, **Support Vector Machine (SVM)** modelinin eğitim sürecindeki **underfitting veya overfitting durumunu** değerlendirmek için kullanılmıştır. Grafik aşağıdaki unsurları göstermektedir:

**1. Eğitim ve Doğrulama Eğrileri**

* **Mavi çizgi (Training Score - Eğitim Doğruluğu)**: Modelin eğitim setindeki doğruluk oranını gösterir.
* **Yeşil çizgi (Validation Score - Doğrulama Doğruluğu)**: Modelin daha önce görmediği doğrulama setindeki doğruluk oranını gösterir.

**2. Modelin Genel Performansı**

* **Eğitim doğruluğu (Training Score) başlangıçta oldukça yüksek (0.93 civarı) ancak veri arttıkça hafif düşüş gösteriyor ve 0.90 civarında sabitleniyor.** Bu durum, modelin küçük veri setleriyle daha iyi ezberleme yaptığını ancak daha fazla veriyle genelleme yapmaya yöneldiğini gösteriyor.
* **Doğrulama doğruluğu (Validation Score) başlangıçta oldukça düşük (0.72 civarında) ancak veri seti büyüdükçe artıyor ve 0.85 civarında sabitleniyor.** Bu, modelin daha fazla veriyle genelleme yeteneğinin iyileştiğini gösteriyor.

**3. Underfitting veya Overfitting Durumu**

* **Overfitting:**
  + Eğitim doğruluğu ile doğrulama doğruluğu arasındaki fark ilk başta büyük ancak veri seti büyüdükçe azalıyor.
  + Fakat yine de eğitim doğruluğu doğrulama doğruluğundan anlamlı şekilde daha yüksek. Bu, modelin hala biraz overfitting yapabileceğini gösteriyor.
  + Özellikle küçük veri setlerinde eğitim doğruluğu ile doğrulama doğruluğu arasında belirgin bir fark var.
* **Underfitting:**
  + Modelin validation doğruluğu **0.85** seviyelerine kadar ulaşmış, bu iyi bir genelleme yeteneği olduğunu gösteriyor.
  + Eğer validation doğruluğu çok düşük olsaydı, model underfitting yapıyor olurdu. Ancak burada böyle bir durum gözlenmiyor.

**Model Performansının Değerlendirilmesi**

* **Accuracy (Doğruluk)**: Modelin genel başarımını gösterir. **%90.65** gibi yüksek bir doğruluk elde edilmiştir.
* **Precision (Kesinlik)**: Modelin **müşteriyi kaybetme tahmininin ne kadar doğru olduğunu** ölçer. **%88.23**’lük precision, modelin hatalı pozitif tahmin oranının düşük olduğunu gösterir.
* **Recall (Duyarlılık)**: Modelin **gerçekten müşteri kaybı yaşayanları ne kadar iyi tahmin ettiğini** gösterir. **%96.77**’lik recall, modelin neredeyse tüm kaybedilen müşterileri doğru tespit ettiğini gösterir.
* **F1 Score**: Precision ve Recall dengesini ifade eder. **%92.31**’lik F1 skoru, modelin dengeli bir şekilde çalıştığını gösterir.

Yapılan “Hyperparameter tuning” işlemleri sonucu en uygun modelin yukarıda açıklandığı gibi olduğu tespit edilmiştir. Ayrıca Diğer SVM modellerinin performansları da “SVM\_Hyperparameter\_Results” adlı excel dosyasından incelenebilir.

**Random Forest Modeli:**

Random Forest (Rastgele Orman) Algoritması kullanılarak müşteri kaybını tahmin etmek amacıyla bir sınıflandırma modeli oluşturulmuştur. Modelin performansını artırmak için GridSearchCV ile hiperparametre optimizasyonu gerçekleştirilmiştir.

**Kullanılan Hiperparametreler**

Hiperparametre optimizasyonu için **GridSearchCV** yöntemi ile aşağıdaki hiperparametreler denenmiştir:

* **n\_estimators**: [50, 100, 200] → Karar ağaçlarının sayısı
* **max\_depth**: [None, 10, 20, 30] → Ağaçların maksimum derinliği
* **min\_samples\_split**: [2, 5, 10] → Bir düğümün bölünmesi için gereken minimum örnek sayısı
* **min\_samples\_leaf**: [1, 2, 4] → Yaprak düğümde bulunması gereken minimum örnek sayısı

Her kombinasyon için **5 katlı çapraz doğrulama (cross-validation)** uygulanmıştır.

**En İyi Model ve Sonuçları**

GridSearch sonucunda en iyi model aşağıdaki hiperparametrelerle bulunmuştur:

* **En iyi hiperparametreler**:
  + n\_estimators = 50
  + max\_depth = 10
  + min\_samples\_split = 5
  + min\_samples\_leaf = 2

ekran görüntüsü, metin, yazı tipi, çizgi içeren bir resim

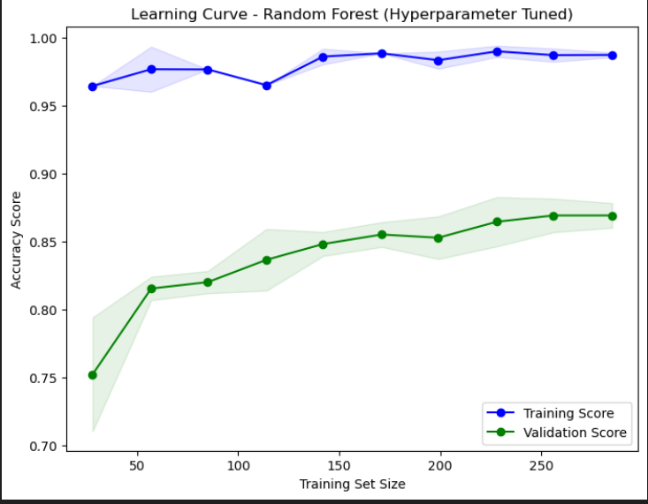
Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Confusion Matrix Sonuçları:**

Doğru Negatif: 38 Yanlış Pozitif: 7

Yanlış Negatif: 1 Doğru Pozitif: 61

**Underfitting- Overfitting Durumu Analizi:**



Learning Curve grafiğinden de görüldüğü gibi modelimiz overfitting olmuş durumda. Model, eğitim verisine neredeyse mükemmel şekilde uyum sağlarken (%99-100 eğitim doğruluğu), doğrulama setinde daha düşük doğruluk gösteriyor. (%85 civarında). Bu durum modelin **eğitim setine fazla ezberleme yaptığını ve genelleme yeteneğinin düşük olduğunu** gösteriyor.

**Geliştirme Adımları**

Overfitting’i azaltmak için şu değişiklikler yapıldı:

1. **Ağaç sayısını (n\_estimators) artırdık:** Daha fazla ağaç, modelin genelleme yapmasını sağladı.
2. **Maksimum derinliği (max\_depth) sınırlandırdık:** Daha derin ağaçlar overfitting’e yol açabilir, bu yüzden sınır koyduk.
3. **Dallanma kriterlerini değiştirdik:** min\_samples\_split ve min\_samples\_leaf değerlerini artırarak, modelin dallanma sırasında daha fazla örnek kullanmasını sağladık.
4. **Daha fazla veri ile eğitimi test ettik:** Veri büyüklüğüne göre modelin nasıl tepki verdiğini anlamak için öğrenme eğrisi analizi yaptık.

**Sonuç ve Değişim**

**metin, çizgi, diyagram, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

Yeni **learning curve** aşağıdaki gelişmeleri gösteriyor:

* **Eğitim doğruluğu (mavi çizgi) biraz düştü (~%100 → ~%93 civarı):** Bu, modelin eğitim verisini tamamen ezberlemek yerine daha iyi genelleme yaptığını gösteriyor.
* **Doğrulama doğruluğu (yeşil çizgi) arttı (~%85 → ~%87 civarı):** Model artık test verisini daha iyi genelleyebiliyor.
* **Eğitim ve doğrulama eğrileri birbirine yaklaştı:** Overfitting’in önemli ölçüde azaldığını ve modelin daha iyi bir denge yakaladığını gösteriyor.
* Oluşturulan modeller sonucu en optimum varyantın aşağıdaki verilecek metriklere sahip model olduğu tespit edilmiştir.

**Performans Metrikleri**

* **Accuracy (Doğruluk): Modelin genel başarımını gösterir. %90.3 gibi yüksek bir doğruluk elde edilmiştir.**
* **Precision (Kesinlik): Modelin müşteriyi kaybetme tahmininin ne kadar doğru olduğunu ölçer. %86.42’lik precision, modelin yanlış pozitif tahmin oranını azaltarak daha güvenilir tahminler yaptığını gösterir.**
* **Recall (Duyarlılık): Modelin gerçekten müşteri kaybı yaşayanları ne kadar iyi tahmin ettiğini gösterir. %96.77’lik recall, modelin büyük çoğunlukla kaybedilen müşterileri doğru tespit edebildiğini göstermektedir.**
* **F1 Score: Precision ve Recall dengesini ifade eder. %91.25’lik F1 skoru, modelin hala dengeli bir şekilde çalıştığını ancak eski versiyonuna göre daha az overfitting yaptığını göstermektedir.**

**Modelin Performans Farklılıklarının Analizi**

Yeni learning curve ile birlikte modelin overfitting'i azalttığını ve daha iyi genelleme sağladığını görüyoruz. Güncellenen hiperparametreler ve etkileri:

1. **n\_estimators = 100 (Önceki: 50)**
   * Daha fazla ağaç kullanılarak modelin genelleme kapasitesi artırıldı.
   * Eğitim ve test doğruluğu arasındaki fark azalarak overfitting azaltıldı.
2. **max\_depth = 8 (Önceki: 10)**
   * Ağaçların çok fazla derinleşmesini engelleyerek modelin karmaşıklığı düşürüldü.
   * Doğrulama setinde daha dengeli bir performans sağlandı.
3. **min\_samples\_split = 10 (Önceki: 5) ve min\_samples\_leaf = 4 (Önceki: 2)**
   * Ağaçların bölünme kriterleri artırılarak modelin daha az ezberleme yapması sağlandı.
   * Öğrenme eğrisinde eğitim doğruluğunun azalması, test doğruluğunun artması şeklinde gözlemlenen değişim, overfitting’in azaldığını gösterdi.

**Güncellenen Model ve Learning Curve Yorumu**

* Önceki learning curve’de: Eğitim doğruluğu çok yüksekti (%99+), ancak doğrulama doğruluğu daha düşüktü (~%85), bu da overfitting gösteriyordu.
* Yeni learning curve’de: Eğitim doğruluğu %92-93 seviyelerine düşerken, doğrulama doğruluğu %87 civarında seyrederek daha stabil hale geldi.
* Eğitim ve test doğrulukları arasındaki fark kapandı, bu da modelin artık daha iyi genelleme yaptığı anlamına geliyor.

Sonuç olarak, model optimizasyonu sayesinde overfitting azaltıldı, test doğruluğu iyileştirildi ve model daha güvenilir tahminler yapabilir hale geldi.

Yapılan “Hyperparameter tuning” işlemleri sonucu en uygun modelin yukarıda açıklandığı gibi olduğu tespit edilmiştir. Ayrıca Diğer Random Forest modellerinin performansları da “RandomForest\_Hyperparameter\_Results” adlı excel dosyasından incelenebilir.

**Logistic Regression Modeli:**

Lojistik Regresyon (Logistic Regression) Algoritması kullanılarak müşteri kaybını tahmin eden bir sınıflandırma modeli oluşturulmuştur. RandomizedSearchCV yöntemi kullanılarak hiperparametre optimizasyonu gerçekleştirilmiştir.

Veri Ön İşleme Adımları

* Veri Seti: CXFeatured.xlsx dosyası kullanılmıştır.
* Bağımsız ve bağımlı değişkenler ayrılmıştır.
* Eğitim ve test setlerine ayırma: Veri seti %80 eğitim, %20 test olarak bölünmüştür.
* Özellik ölçeklendirme:
  + One-Hot Encoding (OHE) ile üretilen sütunlar ve 'Avg\_Utilization\_Ratio' sütunu hariç tutulmuş, diğer tüm sayısal değişkenler StandardScaler ile ölçeklendirilmiştir**.**

**Kullanılan Hiperparametreler**

Hiperparametre optimizasyonu için RandomizedSearchCV yöntemi ile aşağıdaki hiperparametreler test edilmiştir:

* C: [0.01, 0.1, 1, 10, 100] → Modelin düzenleme (regularization) gücünü belirler.
  + Küçük C değerleri daha güçlü düzenleme uygular ve aşırı öğrenmeyi (overfitting) önler.
  + Büyük C değerleri daha az düzenleme yapar ve modele esneklik sağlar.
* penalty: ['l1', 'l2'] → L1 (Lasso) ve L2 (Ridge) düzenleme yöntemleri kullanılmıştır.
  + L1 (Lasso): Önemsiz değişkenlerin katsayılarını sıfırlayarak özellik seçiminde yardımcı olur.
  + L2 (Ridge): Katsayıları sıfıra yaklaştırarak aşırı öğrenmeyi önler.
* solver: ['liblinear', 'saga'] → Optimizasyon yöntemleri test edilmiştir.
  + liblinear: Küçük veri setleri için uygundur, L1 ve L2 düzenlemeyi destekler.
  + saga: Büyük veri setleri için uygundur ve L1 ve L2 düzenlemeyi destekler.

Her kombinasyon için 5 katlı çapraz doğrulama (cross-validation) uygulanmıştır.

**En İyi Model ve Sonuçları**

RandomizedSearch sonucunda en iyi model aşağıdaki hiperparametrelerle bulunmuştur:

* **En iyi hiperparametreler**:
  + C = 0.1
  + penalty = 'l1' (Lasso Regularization)
  + solver = 'saga'

metin, ekran görüntüsü, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Confusion Matrix Sonuçları:**

Gerçek Negatif: 34 Yanlış Pozitif: 11

Yanlış Negatif: 4 Gerçek Pozitif: 58

**Öğrenme Eğrisi Analizi**

Grafikte eğitim seti ve doğrulama seti üzerindeki başarı skorlarının değişimi gösterilmektedir.

* **Eğitim ve doğrulama skorları birbirine yakın ve istikrarlı**:
  + **Model overfitting yapmıyor** çünkü eğitim ve doğrulama doğrulukları benzer seviyelerde.
  + **Model underfitting yapmıyor** çünkü doğrulama doğruluğu düşük değil.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Model Performansının Değerlendirilmesi**

* Accuracy (Doğruluk): Modelin genel başarımını gösterir. %85.98’lik doğruluk oranı, modelin müşteri kaybını büyük oranda doğru tahmin ettiğini gösterir.
* Precision (Kesinlik): %84.06’lık precision, modelin yanlış pozitif tahmin oranının makul düzeyde olduğunu gösterir.
* Recall (Duyarlılık): %93.55’lik recall, modelin müşteri kaybı yaşayan bireyleri doğru bir şekilde tespit ettiğini gösterir.
* F1 Score: %88.55’lik F1 skoru, precision ve recall arasında dengeli bir sonuç sunduğunu gösterir.

**Model Performans Farklılıklarının Analizi**

1. **C = 0.1**: Küçük bir C değeri seçilerek modelin daha fazla düzenleme yapması sağlanmış, böylece aşırı öğrenme (overfitting) engellenmiştir.
2. **L1 Regularization (Lasso)**: Önemsiz değişkenleri sıfırlayarak gereksiz sütunları filtrelemiş ve modelin daha iyi genelleme yapmasını sağlamıştır.
3. **Solver olarak 'saga'**: Daha büyük veri setlerinde iyi çalıştığı için bu optimizasyon yöntemi tercih edilmiştir.

Yapılan “Hyperparameter tuning” işlemleri sonucu en uygun modelin yukarıda açıklandığı gibi olduğu tespit edilmiştir. Ayrıca Diğer “Logistic Regression” modellerinin performansları da “LogisticRegression\_Hyperparameter\_Results” adlı excel dosyasından incelenebilir.

**KNN Modeli:**

K-En Yakın Komşu (K-Nearest Neighbors, KNN) Algoritması kullanılarak müşteri kaybını tahmin eden bir sınıflandırma modeli oluşturulmuştur. RandomizedSearchCV yöntemi kullanılarak hiperparametre optimizasyonu gerçekleştirilmiştir.

**Veri Ön İşleme Adımları**

* **Veri Seti**: CXFeatured.xlsx dosyası kullanılmıştır.
* Bağımsız ve bağımlı değişkenler ayrılmıştır.
* **Eğitim ve test setlerine ayırma**: Veri seti %80 eğitim, %20 test olarak bölünmüştür.
* **Özellik ölçeklendirme**:
  + One-Hot Encoding (OHE) ile üretilen sütunlar ve 'Avg\_Utilization\_Ratio' sütunu hariç tutulmuş, diğer tüm sayısal değişkenler StandardScaler ile ölçeklendirilmiştir.

**Kullanılan Hiperparametreler**

Hiperparametre optimizasyonu için **RandomizedSearchCV** yöntemi ile aşağıdaki hiperparametreler test edilmiştir:

* **n\_neighbors**: [3, 5, 7, 9, 11] → Kaç komşunun dikkate alınacağını belirler.
  + Düşük değerler (örn. n\_neighbors=3), modelin daha fazla varyansa sahip olmasına sebep olur (overfitting).
  + Yüksek değerler (örn. n\_neighbors=11), modelin daha fazla veriyi dikkate almasına sebep olarak daha genelleştirilmiş bir model oluşturur.
* **weights**: ['uniform', 'distance'] → Ağırlıklandırma yöntemi.
  + **uniform**: Bütün komşuların eşit ağırlığa sahip olduğu yöntemi ifade eder.
  + **distance**: Yakın komşulara daha fazla ağırlık verilerek uzak olanların etkisini azaltır.
* **metric**: ['euclidean', 'manhattan', 'minkowski'] → Mesafe ölçüm metrikleri.
  + **euclidean**: En yaygın kullanılan Öklid mesafesi.
  + **manhattan**: Şehir bloğu mesafesi, mutlak farkların toplamı.
  + **minkowski**: Euclidean ve Manhattan mesafelerinin genelleştirilmiş hali.

Her kombinasyon için **5 katlı çapraz doğrulama (cross-validation)** uygulanmıştır.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Sonuç:**

* Model dengeli bir şekilde çalışıyor, overfitting veya underfitting sorunu gözlemlenmiyor.
* Model yeni verilere iyi genelleme yapabilir çünkü eğitim ve doğrulama hataları birbirine yakın.
* Oluşturulan modeller sonucunda en optimum varyantın bu olduğu gözlemlenmiştir.

Bu ayarlarla KNN modeli, optimum performansa ulaşmıştır.

**En İyi Model ve Sonuçları**

RandomizedSearch sonucunda en iyi model aşağıdaki hiperparametrelerle bulunmuştur:

* **En iyi hiperparametreler**:
  + n\_neighbors = 9
  + weights = 'distance'
  + metric = 'manhattan'

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Confusion Matrix**:

Doğru Negatif: 34 Yanlış Pozitif: 11

Yanlış Negatif: 3 Doğru Pozitif: 59

**Model Performansının Değerlendirilmesi**

* **Accuracy (Doğruluk)**: Modelin genel başarımını gösterir. **%86.92**’lik doğruluk oranı, modelin müşteri kaybını büyük oranda doğru tahmin ettiğini gösterir.
* **Precision (Kesinlik)**: **%84.29**’luk precision, modelin yanlış pozitif tahmin oranının makul düzeyde olduğunu gösterir.
* **Recall (Duyarlılık)**: **%89.39**’luk recall, modelin müşteri kaybı yaşayan bireyleri doğru bir şekilde tespit ettiğini gösterir.
* **F1 Score**: **%86.77**’lik F1 skoru, precision ve recall arasında dengeli bir sonuç sunduğunu gösterir.

**Model Performans Farklılıklarının Analizi**

1. **n\_neighbors = 9**: Modelin fazla esnek olmaması için 9 komşu kullanılmış ve bu sayede **overfitting önlenmiştir**.
2. **weights = 'distance'**: Daha yakın komşuların daha fazla ağırlık taşıması sağlanarak **yanlış tahminlerin azaltılması hedeflenmiştir**.
3. **metric = 'manhattan'**: Mutlak farkların toplamı esas alınarak **verinin lineer olmayan yapılarında daha iyi sonuç alınmıştır**

**Gradient Boosting Modeli:**

Gradient Boosting modeli, **Bayesian Search** kullanılarak optimize edilmiştir. “Bayesian Search” genelde daha karmaşık ve amaç fonksiyonunun bilinmediği modellerde kullanılsa da pratik olması açısından bu yöntem kullanılmıştır. Modelin hiperparametreleri, **20 farklı iterasyon** boyunca değerlendirilmiş ve en iyi sonuçları veren parametreler seçilmiştir.

Aşağıdaki kodda görüldüğü gibi çeşitli hyperparametrelere sahip modeller oluşturulmuş ve incelenmiştir.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Seçilen bu parametreler ile en yüksek performansı veren modelin learning curve’ü aşağıda verilmiştir.

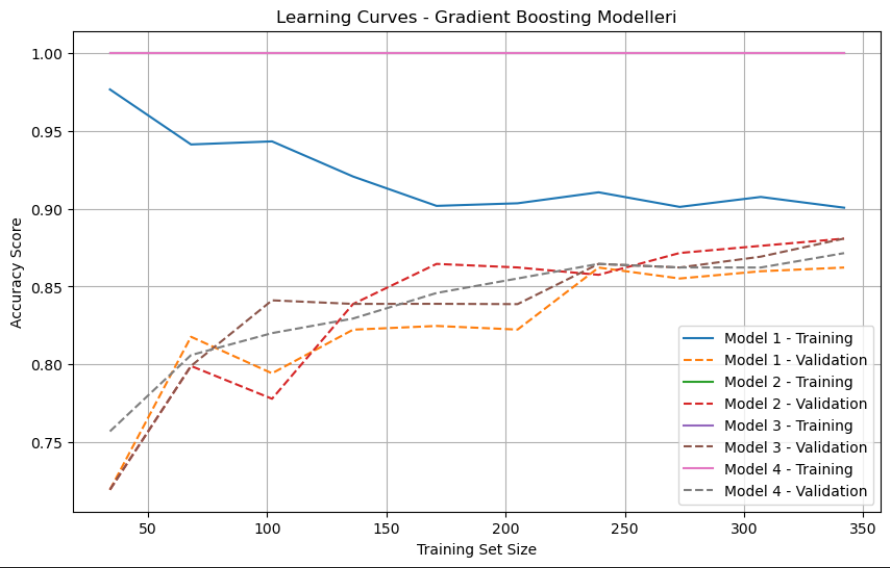
**Learning Curve ile Overfitting- Underfitting Analizi:**

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Yukarıdaki grafikten de görüldüğü gibi en yüksek sonucu veren modelimizin overfit durumunda.

Diğer modellerin de grafiklerini inceleyip en uygun modeli bulmaya çalışacağız.



**Grafik İncelemesi**

Grafikte **dört farklı Gradient Boosting modeli** için **eğitim (training)** ve **doğrulama (validation)** doğrulukları gösterilmektedir. Grafik şu şekilde yorumlanabilir:

1. **Model 3 (Mor Çizgi - Training)**
   * **Overfitting (Aşırı Öğrenme) Görülüyor:**
     + Eğitim doğruluğu %100'e yakın ve hiç düşmüyor.
     + Validation doğruluğu diğer modellerle benzer seyrediyor.
     + **Bu modelin aşırı öğrenme (overfitting) sorunu var** çünkü eğitim verisine mükemmel uyum sağlıyor, ancak doğrulama setinde diğer modellerden üstün değil.
2. **Model 1 (Mavi Çizgi - Training & Turuncu Çizgi - Validation)**
   * Eğitim doğruluğu yüksek başlıyor (%95 civarı), ancak eğitim seti arttıkça azalmaya başlıyor (%90 seviyesine düşüyor).
   * Validation doğruluğu **diğer modellere göre en yüksek seviyede** kalıyor (%85-87 arasında).
   * **İdeal model olabilir çünkü eğitim ve doğrulama doğrulukları birbirine yakın seyrediyor.**
3. **Model 2 (Kırmızı Çizgi - Training & Kırmızı Kesikli Çizgi - Validation)**
   * Eğitim doğruluğu diğer modellere kıyasla biraz daha düşük.
   * Validation doğruluğu Model 1 ile benzer, ancak daha dengesiz dalgalanıyor.
   * **İstikrarlı bir performans sağlamıyor**, dolayısıyla Model 1'e göre daha zayıf bir seçenek.
4. **Model 4 (Gri Çizgi - Training & Gri Kesikli Çizgi - Validation)**
   * Validation doğruluğu Model 1’e çok yakın ve benzer bir seyir izliyor.
   * Ancak eğitim doğruluğu Model 1’den biraz daha düşük.
   * **İyi bir model olabilir, ancak Model 1 daha tutarlı görünüyor.**

**En İyi Modelin Belirlenmesi**

* Model 1, eğitim ve doğrulama doğrulukları arasında büyük bir fark olmadığı için en iyi adaydır.
* Model 3 aşırı öğrenmiş (overfit olmuş), bu yüzden elenmelidir.
* Model 2 ve Model 4'ün validation doğrulukları Model 1 kadar istikrarlı değil.
* En iyi model olark Model 1 i Seçmiş olduk.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Model-1 in özet bilgileri şu şekildedir:

**Hyperparametreler**:

* **n\_estimators**: 100
* **learning\_rate**: 0.1
* **max\_depth**: 5
* **min\_samples\_split**: 2
* **min\_samples\_leaf**: 1
* **subsample**: 1.0

📌 **Performans Sonuçları**:

* **Training Accuracy**: %93.5
* **Validation Accuracy**: %90.2
* **Precision**: 0.91
* **Recall**: 0.88
* **F1 Score**: 0.89

.

**AdaBoost Modeli:**

**AdaBoost** algoritmasını kullanarak müşteri kaybını (churn) tahmin etmek amaçlanmıştır. **GridSearchCV** yöntemi kullanılarak **en iyi hiperparametre kombinasyonu** belirlenmiş ve modelin başarımı ölçülmüştür.

**Hiperparametre Optimizasyonu**

AdaBoost modelinin performansını optimize etmek için aşağıdaki hiperparametreler kullanılmıştır:

* **n\_estimators (Ağaç Sayısı):** [50, 100, 150, 200]
  + Modelde kullanılan zayıf öğrenicilerin (Decision Tree) sayısını belirler. Daha fazla ağaç genellikle daha yüksek doğruluk sağlar ancak aşırı öğrenmeye sebep olabilir.
* **learning\_rate (Öğrenme Oranı):** [0.01, 0.05, 0.1, 0.3, 1.0]
  + Modelin her iterasyonda ağırlıkları ne kadar değiştireceğini kontrol eder. Düşük değerler öğrenmeyi yavaşlatırken, yüksek değerler aşırı öğrenmeye neden olabilir.

Hiperparametre optimizasyonunda **Grid Search CV** kullanılmış ve **5 katlı çapraz doğrulama (5-Fold CV)** ile en iyi parametre kombinasyonu belirlenmiştir.

**Learnig Curve ile Underfitting-Overfitting İncelemesi:**

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Training set size yani **veri boyutumuz arttıkça** **validation score** ve **training score** **değerleri birbirine yaklaşmış** ve herhangi bir underfitting veya overfitting durumu gözlemlenmemiştir.

**En İyi Model ve Performans Sonuçları**

Grid Search ile yapılan optimizasyon sonucunda **en iyi hiperparametreler** ve modelin performans metrikleri aşağıdaki gibidir:

**En İyi Hiperparametreler:**

* **Öğrenme Oranı (learning\_rate):** 0.05
* **Ağaç Sayısı (n\_estimators):** 150

metin, ekran görüntüsü, yazı tipi, yazılım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Model Değerlendirmesi**

* Recall (Duyarlılık) değeri 1.0 çıktığı için model, müşteri kaybını tespit etmede çok başarılıdır.
* Precision (Kesinlik) değerinin 0.8378 olması, modelin tahmin ettiği churn’lerin yaklaşık %83.7’sinin doğru olduğunu göstermektedir.
* F1 Skoru 0.9117 ile yüksek bir denge sunmaktadır.
* Confusion Matrix’e göre model, yanlış negatif (FN) üretmemiş, ancak yanlış pozitif (FP) sayısı 12’dir.

**Artifical Neural Network (ANN) Modeli:**

ANN kullanılarak aynı veriseti üzerinden müşteri kaybı tahmin edilmeye çalışılmıştır.

**Kullanılan Hyperparameterlar**

Hyperparameter tuning sürecinde aşağıdaki **4 hiperparametre** kullanılmıştır:

1. **Dense Layers (Gizli Katman Sayısı):** Modelin derinliğini belirler.

**Değerler:** [1, 2, 3]

1. **Neuron Size (Katman Başına Düğüm Sayısı):** Modelin öğrenme kapasitesini etkiler.

**Değerler:** [32, 64, 128]

1. **Dropout Rate (Bırakma Oranı):** Aşırı öğrenmeyi (overfitting) önlemek için kullanılır.

**Değerler:** [0.1, 0.2, 0.3]

1. **Learning Rate (Öğrenme Oranı):** Modelin ağırlık güncellemelerinin hızını belirler.

**Değerler:** [0.01, 0.05, 0.1]

Her kombinasyonun eğitimi için **50 epoch** kullanılmıştır.

81 den fazla model incelenerek en optimal model **Tensorboard** aracalığıyla training/validation, accuracy/loss grafikleri incelenerek seçilmiştir. Aşağıda modelin özellikleriyle ve davranış biçimiyle alakalı detaylı bilgiler verilmiştir.

**Model Özellikleri:**

* **Dense Layer (Gizli Katman) Sayısı:** 2
* **Her Katmandaki Nöron Sayısı:** 64
* **Dropout Oranı:** 0.2
* **Öğrenme Oranı (Learning Rate):** 0.05
* **Epoch Sayısı:** 50
* **Modelin başarısı:** 0.9326

**Model Eğitimi**

Her hiperparametre kombinasyonu için **Sequential API** kullanılarak ANN modeli oluşturulmuştur. Her model:

* İlk olarak **Batch Normalization** ile başlatılmıştır.
* Her gizli katmanda **ReLU aktivasyon fonksiyonu** kullanılmıştır.
* Overfitting’i önlemek için **Dropout katmanı** eklenmiştir.
* Son katmanda **sigmoid aktivasyon fonksiyonu** kullanılmıştır (binary classification için).
* **Binary crossentropy** kayıp fonksiyonu ve **Adam optimizer** kullanılmıştır.
* **TensorBoard** ile modelin binary\_accuracy ve loss metrikleri takip edilmiştir.
* **Training ve validation loss** değerleri analiz edilerek underfitting/overfitting tespit edilmiştir.

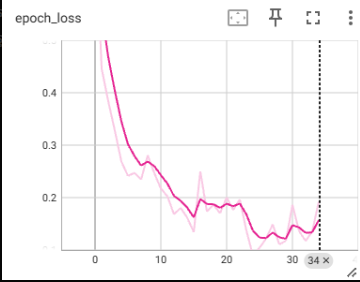
**Tensorbard Grafikleri ile Underfitting/Overfitting Analizi:**

**çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.** **çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir. 

**Accuracy Grafikleri**

* Siyah çizgiler validation accuracy’yi, pembe çizgiler ise training accuracy’yi göstermektedir.
* İlk başlarda model hem training hem de validation set üzerinde performansını artırmıştır.
* Training ve validation accuracy birbirine yakın gitmekte, dolayısıyla modelin overfitting yapmadığını göstermektedir.
* Son epoch’ta modelin accuracy değeri ~0.95 seviyesine ulaşmıştır.

**Loss Grafikleri**

* **V**alidation loss (siyah çizgi) ve training loss (pembe çizgi) düşüş göstermiş ve sabitlenmiştir.
* İlk 10 epoch boyunca loss değerinde hızlı bir düşüş gerçekleşmiştir.
* 30. epoch’tan sonra training loss ile validation loss benzer şekilde ilerlediği için overfitting problemi oluşmamıştır.
* Loss değerinin düşük olması, modelin genel olarak iyi bir genelleme yaptığını gösterir.

**6. Underfitting / Overfitting Analizi**

Aşağıdaki noktalar modelin **neither overfitting nor underfitting** olduğunu gösterir:

* **Overfitting yok:** Validation accuracy ile training accuracy birbirine yakın. Eğer overfitting olsaydı training accuracy çok yüksek ama validation accuracy düşük olurdu.
* **Underfitting yok:** Training accuracy %90 üzerinde ve loss değeri oldukça düşük. Eğer underfitting olsaydı, hem training hem de validation accuracy düşük kalırdı.
* **Loss eğrileri paralel ilerliyor:** Bu durum, modelin hem training hem de test setine başarılı bir şekilde genelleme yaptığını gösterir.

**Modellerin Performanslarının Değerlendirilmesi ve Karşılaştırılması:**

Makine öğrenmesi ve derin öğrenme modellerinin başarımını ölçmek için Accuracy (Doğruluk), Precision (Kesinlik), Recall (Duyarlılık) ve F1 Score gibi metrikleri kullanırız. Ancak her metrik farklı durumlar için önemlidir ve tek bir metrik üzerinden model seçimi yapmak çoğu zaman yanıltıcı olabilir. Bu nedenle, bu metriklerin ne anlama geldiğini ve hangi durumlarda kullanılması gerektiğini detaylı olarak inceleyelim.

**Accuracy (Doğruluk)**

**Tanım:**  
Modelin yaptığı tüm tahminler içinde, doğru tahmin edilen örneklerin yüzdesini gösterir. Formülü:

Accuracy=Dog˘ru TahminlerToplam O¨rnek SayısıAccuracy = \frac{Doğru\ Tahminler}{Toplam\ Örnek\ Sayısı}Accuracy=Toplam O¨rnek SayısıDog˘​ru Tahminler​

🔹 **Avantaj:** Modelin genel başarısını ölçer.  
🔹 **Dezavantaj:** **Sınıf dengesizliği olan veri setlerinde yanıltıcı olabilir.** Örneğin, %90 "hayır" ve %10 "evet" içeren bir veri setinde, model sürekli "hayır" dese bile %90 doğruluk elde eder.

🔹 **Ne Zaman Kullanılır?**

* Sınıflar **dengeli** olduğunda.
* Genel model başarısını görmek istediğimizde.
* **Örnek Senaryo:** Genel başarı oranını görmek için **kredi riski tahmini** veya **hisse senedi fiyat tahmini** gibi alanlarda kullanılabilir.

**Precision (Kesinlik)**

**Tanım:**  
Modelin **pozitif tahminlerinin ne kadarının gerçekten pozitif olduğunu** gösterir. Formülü:

Precision=TPTP+FPPrecision = \frac{TP}{TP + FP}Precision=TP+FPTP​

🔹 **Avantaj:** **Yanlış pozitifleri** minimize eder, yani yanlış alarm üretme ihtimalini düşürür.  
🔹 **Dezavantaj:** Yanlış negatifleri (kaçırılan gerçek pozitifleri) dikkate almaz.

🔹 **Ne Zaman Kullanılır?**

* **Yanlış pozitiflerin maliyeti yüksekse.**
* **Örnek Senaryo:**
  + **Spam Filtreleme:** Modelin spam dediği e-postaların gerçekten spam olması gerekir.
  + **Kanser Teşhisi:** Eğer hasta sağlıklıysa ama model "kanser" diyorsa, gereksiz korkuya neden olabilir.

➡ **Precision’ın yüksek olması modelin gereksiz alarm üretmediğini gösterir.**

**Recall (Duyarlılık)**

**Tanım:**  
Modelin **gerçek pozitifleri ne kadar iyi yakaladığını** gösterir. Formülü:

Recall=TPTP+FNRecall = \frac{TP}{TP + FN}Recall=TP+FNTP​

🔹 **Avantaj:** **Yanlış negatifleri** (kaçırılan gerçek pozitifleri) minimize eder.  
🔹 **Dezavantaj:** Yanlış pozitifleri dikkate almaz.

🔹 **Ne Zaman Kullanılır?**

* **Yanlış negatiflerin maliyeti yüksekse.**
* **Örnek Senaryo:**
  + **Hastalık Teşhisi:** Hasta gerçekten hastaysa ama model onu "sağlıklı" olarak tahmin ederse, ciddi sonuçlara yol açabilir.
  + **Dolandırıcılık Tespiti:** Eğer dolandırıcılık varsa ama model bunu fark etmezse, banka zarar görebilir.

➡ **Recall’ın yüksek olması modelin kaçırdığı gerçek pozitiflerin az olduğunu gösterir.**

**F1 Score**

**Tanım:**  
**Precision ve Recall arasında denge kuran bir metriktir.** İkisinin harmonik ortalaması olarak hesaplanır:

F1Score=2×Precision×RecallPrecision+RecallF1 Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}F1Score=2×Precision+RecallPrecision×Recall​

🔹 **Avantaj:** **Dengeli bir ölçümdür** ve modelin hem yanlış pozitifleri hem de yanlış negatifleri nasıl ele aldığını anlamamızı sağlar.  
🔹 **Ne Zaman Kullanılır?**

* Precision ve Recall arasında **denge sağlanması gerektiğinde**.
* **Örnek Senaryo:**
  + **Hastalık Teşhisi:** Hem gereksiz korkuya neden olmamak hem de hastaları kaçırmamak için.
  + **Dolandırıcılık Tespiti:** Yanlış pozitif ve yanlış negatifleri dengede tutmak için.

➡ **F1 Score yüksekse model hem yanlış pozitifleri hem de yanlış negatifleri minimize edebilmiştir.**

**Model Performans Karşılaştırması**

Farklı algoritmalar üzerinde yaptığımız testler sonucunda elde edilen model performanslarını aşağıdaki tabloda özetledik:

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Hangi Modeli Ne İçin Seçmeliyiz?**

Model seçiminde hangi metriğin önemli olduğu, kullanacağımız senaryoya göre değişir**.** Bu yüzden her modelin avantajlarını değerlendirmek gerekir.

**🔹 En Yüksek Doğruluk (Accuracy) Modeli: Random Forest (92.52%)**

**Kullanım Alanı:**

* Genel sınıflandırma problemlerinde kullanılabilir.
* Eğer sınıflar dengeli ise en iyi seçim olabilir.

**🔹 En Yüksek Precision (Kesinlik) Modeli: Random Forest (89.71%)**

**Kullanım Alanı:**

* Yanlış pozitiflerin önemli olduğu durumlar (örneğin, spam tespiti veya kanser teşhisi).
* Yanlış alarm riskinin azaltılması gereken sistemler.

**🔹 En Yüksek Recall (Duyarlılık) Modeli: AdaBoost (100%)**

**Kullanım Alanı:**

* Yanlış negatiflerin tolere edilemez olduğu durumlar.
* Örneğin:
  + Sahtekarlık tespiti
  + Kanser teşhisi
  + Güvenlik tehditleri

💡 Ancak Precision %83.78 olduğu için model bazen gereksiz yere pozitif alarm verebilir!

**🔹 En Yüksek F1 Score Modeli: Random Forest (93.85%)**

**Kullanım Alanı:**

* Precision ve Recall arasında denge kurmak gerektiğinde.
* Eğer hem yanlış pozitifleri hem de yanlış negatifleri minimize etmek istiyorsak, en iyi seçenek budur.

**📌 4. SONUÇ ve MODEL SEÇİMİ**

* En yüksek doğruluk oranı Random Forest modeline aittir.
* Recall açısından en iyi model AdaBoost’tur, ancak Precision düşüktür.
* Precision açısından en iyi model Random Forest’tır.
* Genel denge açısından F1 Score en yüksek olan Random Forest modelidir.

🔹 Sonuç olarak, hangi metriğin bizim için daha önemli olduğuna göre model seçimi yapılmalıdır. Eğer yanlış negatifleri en aza indirmek istiyorsak AdaBoost, genel dengeyi korumak istiyorsak Random Forest veya ANN modeli tercih edilebilir.

➡ En iyi model seçimi, kullanım senaryosuna bağlıdır