

# İK Analitiği Projesi: Çalışan İstifası Tahmini Raporu

Ad Soyad / No: Alaattin Buğra DURMUŞ / 231118013

## Özet ve Problem Tanımı

Bu proje, IBM'in anonim İK veri setini kullanarak çalışanların işten ayrılma riskini (% Attrition) doğru bir şekilde tahmin etmeyi amaçlamaktadır. Projenin temel hedefi, sadece tahmin yapmak değil, aynı zamanda istifaya neden olan faktörleri (Feature Importance) belirleyerek şirkete somut stratejik öneriler sunmaktır.

## Model Performansı ve Kıyaslama Sonuçları

Dört farklı makine öğrenmesi sınıflandırma algoritması (Logistic Regression, SVM, KNN ve Random Forest) kullanılarak performans testleri yapılmıştır. Elde edilen ham doğruluk oranları aşağıdaki gibidir:

| Model Adı                    | Ham Doğruluk Oranı (Accuracy) | Yorum   |
|------------------------------|-------------------------------|---|
| Support Vector Machine (SVM) | %87.53                        | En yüksek tahmin doğruluğu.                         |
| K-Nearest Neighbors (KNN)    | %87.30                        | Yüksek doğrulukta, ancak hesaplama maliyeti yüksek. |
| Logistic Regression          | %86.62                        | Basit ve hızlı temel çizgi (Baseline) modeli.       |
| Random Forest                | %85.71                        | En düşük ham doğruluk oranı.                        |

## Doğruluk Tuzağı ve Dengesiz Veri (Imbalanced Data)

Sonuçlara göre **SVM**, **%87.53** ile en yüksek doğruluk oranını elde etmiştir. Ancak veri setimizde çalışanların büyük çoğunluğu (%84) istifa etmediği için, sadece **No Attrition** (İstifa Yok) tahmin eden bir model bile yüksek doğruluk oranına ulaşabilir.

Bu tür dengesiz verilerde asıl önemli olan, **küçük olan sınıfı (yani istifa edenleri)** ne kadar başarılı yakaladığımızı gösteren **Recall (Hassasiyet)** ve **F1-Score** metrikleridir.

## Nihai Model Seçimi: Random Forest

Ham doğruluk oranı en düşük olmasına rağmen, projenin nihai tahmin modeli olarak **Random Forest** seçilmiştir. Bunun temel nedenleri teknik sağlamlık ve iş dünyası için sunduğu yorumlanabilirliğidir:

### 1. Yorumlanabilirlik (Interpretability) Avantajı

SVM ve KNN gibi modeller birer **Kara Kutu (Black Box)** gibi çalışır; yüksek doğruluk sağlar ancak "neden" sorusuna cevap veremezler. İK yöneticilerinin sadece "Bu kişi gidecek" tahminine değil, "**Bu kişi şu sebeplerle (Fazla Mesai, Düşük Zam Oranı vb.) gidiyor**" açıklamasına ihtiyacı vardır.

Random Forest, diğer algoritmala göre en iyi **Özellik Önem Düzeyi (Feature Importance)** çıktısını sağlar. Bu, projenin asıl değerini oluşturan, istifanın arkasındaki en kritik 10 nedeni bilimsel olarak listeleye imkanı sunar.

### 2. Dengesiz Veriye Dayanıklılık

Random Forest, birden fazla karar ağacını bir araya getirdiği için, veri setindeki gürültüye ve dengesizliklere karşı daha dirençlidir. Gerçek dünyadaki iş uygulamalarında genellikle SVM ve KNN'e göre **daha tutarlı ve genelleştirilebilir** performans sergiler.

## İş Çıktısı ve Öneriler

Seçilen Random Forest modeli üzerinden yapılan Özellik Önem Düzeyi analizi sonucunda, şirketin çalışan kaybını azaltmak için aşağıdaki aksiyonları alması önerilir:

- Fazla Mesai (OverTime):** İstifa etme riskini en çok artıran faktör olduğu tespit edilmiştir. Bu durum, maaş zammından bile daha güçlü bir belirleyicidir. Fazla mesai politikaları acilen gözden geçirilmelidir.
- Maaş Seviyesi (Job Role/Salary Level):** En düşük seviyede maaş alan çalışanlar risk grubundadır.
- İşe Uzaklık (DistanceFromHome):** Evden uzak çalışanların istifa eğilimi daha yüksektir. Uzaktan çalışma (Remote Work) imkanları bu riski azaltabilir.

**Sonuç:** Projenin amacı göz önüne alındığında, **%85.71 Doğruluk** ile çalışan **Random Forest** modeli, sunduğu **nedensellik** bilgisi ile **%87.53 Doğruluk** sunan SVM modelinden operasyonel açıdan daha değerlidir.