İK Analitiği Projesi: Çalışan İstifası Tahmini Raporu

Ad Soyad / No: Alaattin Buğra DURMUŞ / 231118013
Kullanılan Model: 1D-Convolutional Neural Network (CNN)

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 matematiksel bir tahmin yapmak değil; istifaya neden olan kök nedenleri **(Feature Importance)** belirlemek ve **geliştirilen interaktif arayüz ile** bu teknik çıktıları İK departmanının kullanabileceği somut bir karar destek sistemine dönüştürmektir.

1. Veri Ön İşleme ve Mühendisliği (Data Preprocessing)

Veri seti üzerinde yapılan işlemler, ham verinin bir derin öğrenme modeline beslenecek formata getirilmesini kapsar.

- **Özellik Seçimi:** SHAP analizleri sonucunda model başarısını maksimize eden ve İK süreçlerinde en anlamlı olan **8 değişken** (OverTime, MonthlyIncome, TotalWorkingYears, DistanceFromHome, Age, MaritalStatus, StockOptionLevel, NumCompaniesWorked) seçilmiştir.
- **Encoding:** `OverTime` ve `Attrition` ikili (binary), `MaritalStatus` ise çoklu (label encoding) olarak sayısallaştırılmıştır.
- **Ölçeklendirme (Scaling):** CNN modelinin gradyan hesaplamalarında kararlı kalması için `StandardScaler` kullanılarak veriler ortalama 0 ve standart sapma 1 olacak şekilde normalize edilmiştir.

2. Model Mimarisi: 1D-CNN

Projede kullanılan `OptimizedCNN` sınıfı, tablosal verilerdeki gizli örüntüleri yakalamak üzere tasarlanmıştır.

Mimari Bileşenler:

- **1D Convolutional Layers:** Özellikler arasındaki yerel korelasyonları filtreler aracılığıyla öğrenir.
- **Batch Normalization:** Eğitim sırasında her katmanın girdilerini normalize ederek eğitimin hızlanmasını ve modelin daha kararlı olmasını sağlar.
- **Adaptive Average Pooling:** Verinin en önemli özelliklerini özetleyerek vektör boyutunu düşürür.
- **Dropout (%40):** Eğitim sırasında nöronların bir kısmını rastgele kapatarak modelin veriyi ezberlemesini (overfitting) engeller.
- **Sigmoid Aktivasyonu:** Çıkış katmanında 0 ile 1 arasında bir olasılık değeri üretir.

3. Eğitim Stratejisi

Modelin klasik yöntemlerden (SVM, RF vb.) daha iyi performans göstermesi için şu teknikler uygulanmıştır:

- **Optimizer:** `AdamW` (Ağırlık sönümlemeli Adam) kullanılarak modelin genelleme kapasitesi artırılmıştır.
- **Learning Rate Scheduler:** `ReduceLROnPlateau` ile eğitim tıkandığında öğrenme hızı otomatik olarak düşürülmüştür.
- **Early Stopping:** Modelin test başarısı 30 epoch boyunca artmadığında eğitim durdurularak en iyi ağırlıklar (`optimized_cnn.pt`) kaydedilmiştir.

4. Açıklanabilir Yapay Zeka (XAI) ve SHAP

Projenin en kritik bölümlerinden biri, modelin neden "İstifa" dediğini kanıtlamaktır.

- **TreeExplainer / DeepExplainer:** Random Forest ve CNN modelleri üzerinde çalıştırılarak öznitelik önem düzeyleri belirlenmiştir.
- **Bulgular:** Analiz sonucunda **Fazla Mesai (OverTime)** ve **Aylık Gelir (MonthlyIncome)** değişkenlerinin model kararlarında en yüksek ağırlığa sahip olduğu kanıtlanmıştır.

5. Uygulama Arayüzü (Gradio)

İK departmanının kullanımı için tasarlanan arayüz şu mantıkla çalışır:

1. **Girdi:** Kullanıcıdan 8 parametre alınır.
2. **İşleme:** Alınan veriler eğitimdeki `scaler` nesnesi ile ölçeklenir ve PyTorch tensörüne çevrilir.
3. **Tahmin:** Eğitilmiş CNN modeli `%` cinsinden bir risk puanı üretir.
4. **Karar Mekanizması:**
    - `> %60`: Kritik Risk (Kırmızı Alarm)
    - `%35 - %60`: Orta Risk (Turuncu Alarm)
    - `< %35`: Düşük Risk (Yeşil Işık)

Proje sonucunda ulaşılan **%86.39** doğruluk oranı, derin öğrenme mimarisinin tablosal verilerde doğru optimizasyon ile klasik modellerle yarışabileceğini göstermiştir. Özellikle sınıf dengesizliğine rağmen modelin istifaları yakalama (Recall) başarısı, İK süreçlerinde maliyet tasarrufu sağlama potansiyeline sahiptir.

Model Performansı ve Kıyaslama Sonuçları

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

Model Adı	**Ham Doğruluk Oranı (Accuracy)**	**Yorum**
Support Vector Machine (SVM)	%85.03	En yüksek tahmin doğruluğu.
K-Nearest Neighbors (KNN)	%82.65	Yüksek doğrulukta, ancak hesaplama maliyeti yüksek.
Logistic Regression	%85.37	Basit ve hızlı temel çizgi (Baseline) modeli.
Random Forest	%83.33	En düşük ham doğruluk oranı.
1D CNN	**%86.39**	**En yüksek genel doğruluk; derin öğrenme ile karmaşık örüntü yakalama başarısı.**