YAP470 Ödev Raporu

1- Veri Seti Bulguları

Veri setinde tahmin edilmek istenen “Attrition” değeri dengesiz şekilde veriye yayılmıştır. Veri setinde kayıp değerler yoktur. Veride sadece tek değer alan 3 tane column tespit edilip ve oluşturacağımız modelde bir işlevi olmayan çalışan id columnları öncelikle veri setinden çıkarılmıştır. Veri setinde diğer featureler “Attrition” üzerindeki dağılımlarına bakılıp çoğu dengeli sayılabilecek dağılıma sahiptir. Ayrıca Veri Setinde yüksek korelasyon içeren featureler vardır. Bunların hepsi şirkette toplam çalıştığı yıl ile orantılıdır. Model eğitirken bunlar düşürülebilirdi yönergede belirtilmediği için veri ön işleme aşamasında düşürülmemiştir. Fakat lineer regression bunları regulize ederek kendisi düşürmüştür. Diğer modeller bunları tam olarak başaramamıştır. Veri seti %80 train, %10 validation, %10 test olarak 3 parçaya ayrılıp MinMax ile normalize edilmişitir. Ayrından model eğitim aşamasına geçilmiştir.

2- Model Bulguları

Lineer Regression, Random Forest ve Support Vector Machine olmak üzere 3 farklı train datası üzerinde model eğitilip GridSearch ile HyperParametre optimizasyonu yapılmıştır. Hyperparametre optimizasyonu lineer regression için f1 skoru üzerinde pek etki göstermeyip diğer modellerin gelişmesinde yardımcı olmuştur. Sonuç olarak tüm aşamalarda en iyi model Lineer Regression olarak gözlemlenmiştir. Bunun sebebi Feature Selection işlemini otomatik olarak yaptığından kaynaklanıyor olabilir. Ayrıca en hızlı eğitilen model de Lineer Regression modelidir.