Lojistik Regresyon Yöntemi Kullanarak İş Başvurusu Sınıflandırması

Yasin Tarakçı[1]
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
yasin.tarakci1@std.yildiz.edu.tr

M. Elif Karslıgil[2]
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
elif@yildiz.edu.tr

***Özet*—Bu çalışmada, bir firmaya yapılan iş başvurularının değerlendirilmesi amacıyla, adayların iki farklı sınavdan aldıkları notlara göre işe kabul durumlarını tahmin eden bir lojistik regresyon modeli geliştirilmiştir. Veri seti %60 eğitim, %20 doğrulama ve %20 test verisi olacak şekilde ayrılmıştır. Model eğitimi sırasında, aktivasyon fonksiyonu olarak Sigmoid, hata fonksiyonu olarak Cross-Entropy Loss ve optimizasyon yöntemi olarak Stochastic Gradient Descent (SGD) kullanılmıştır. Herhangi bir hazır makine öğrenmesi kütüphanesi kullanılmadan, tüm algoritmalar Python ortamında NumPy kütüphanesi ile sıfırdan kodlanmıştır. Eğitim sürecinde aşırı öğrenmeyi (overfitting) engellemek adına "Early Stopping" mekanizması uygulanmıştır. Elde edilen sonuçlarda, test veri seti üzerinde 0.85 doğruluk (accuracy) ve 0.90 F1 skoru başarısına ulaşılmış, modelin iki sınıfı birbirinden başarılı bir şekilde ayırdığı gözlemlenmiştir. Çalışmanın kaynak kodlarına, kaynak kodlarının ve algoritmanın detaylı dökümantasyonuna, ve veri setlerine https://github.com/ysntrkc/machine-learning-hw1. git adresinden erişilebilir.**

***Anahtar Kelimeler*—Lojistik Regresyon, İkili Sınıflandırma, Stokastik Gradyan İnişi, Çapraz Entropi Kaybı, Aşırı Öğrenme, Erken Durdurma**

I. Giriş

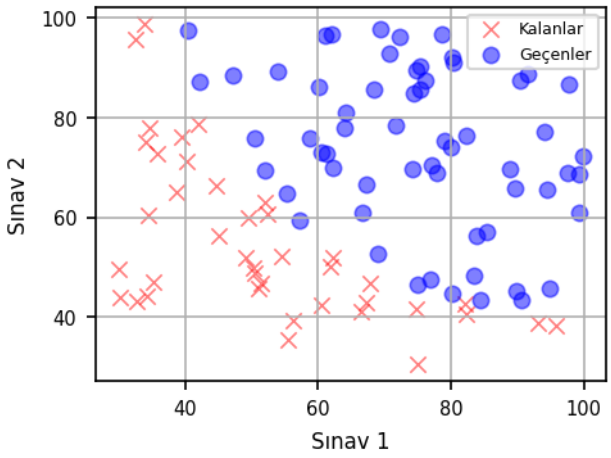
Günümüzde sınıflandırma problemleri, makine öğrenmesinin en yaygın uygulama alanlarından biridir. Bu ödev kapsamında ele alınan problem, adayların iki ayrı sınavdan aldıkları notlara (girdiler) göre işe alınıp alınmayacaklarının (çıkış etiketi: 1-Kabul, 0-Ret) tahmin edilmesidir. Çalışmanın temel amacı, lojistik regresyon algoritmasının matematiksel temellerini (Sigmoid fonksiyonu, ağırlık güncelleme, türev hesabı) hazır kütüphaneler kullanmadan kavramak ve uygulamaktır. Ayrıca eğitim ve doğrulama süreçlerini analiz ederek, modelin genelleme yeteneğini ölçmek ve olası aşırı öğrenme durumlarına karşı iyileştirme stratejileri geliştirmek hedeflenmiştir.

II. Deneysel Analiz

A. Veri Seti ve Ön İşleme

Veri seti toplam 101 örnekten oluşmaktadır. İlk adım olarak verilerin ilk %60'i eğitim, sonraki %20'si doğrulama ve kalan %20'si ise test seti olmak üzere 3 parçaya ayrılmıştır. Gradyan inişinin daha hızlı ve kararlı yakınsaması için tüm girdiler (sınav notları) Min-Max normalizasyonu ile 0-1 aralığına ölçeklendirilmiştir.

Veri setinin dağılımı Şekil 1'de gösterilmiştir. Grafikten görüleceği üzere, sınav notları sonucu işe girenler (mavi noktalar) ile işe giremeyen adaylar (kırmızı çarpılar) arasında genel olarak belirgin bir ayrım söz konusudur.



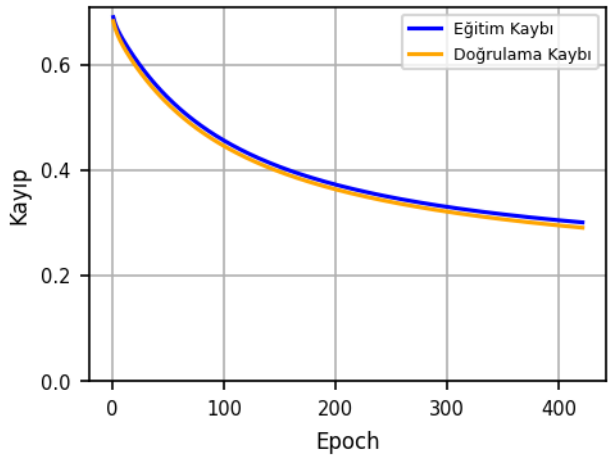
Şekil 1. Tüm veri setinin sınav notlarına göre dağılımı.

B. Model Eğitimi ve Loss Analizi

Model, 0.01 öğrenme oranı (learning rate) ile başlatılmış ve maksimum 500 epoch hedeflenmiştir. Hata fonksiyonu olarak Cross-Entropy Loss kullanılmıştır. Her epoch sonunda eğitim ve doğrulama kümeleri için ortalama kayıp (loss) değerleri hesaplanmıştır.

Eğitim süreci boyunca elde edilen kayıp grafiği Şekil 2'de sunulmuştur. Grafik incelendiğinde, hem eğitim (mavi çizgi) hem de doğrulama (turuncu çizgi) hatasının düşüş eğilimini sürdürdüğü görülmektedir. Ancak eğitim hatası kararlı bir şekilde azalmaya devam ederken, doğrulama hatasındaki düşüş hızı belirli bir noktadan sonra yavaşlamış ve marjinal hale gelmiştir. Doğrulama hatasındaki iyileşme miktarının, belirlenen minimum değişim eşiğinin (min_delta=0.001) altında kalması durumu, tanımlanan sabır süresi (patience=5 epoch) bo-

yunca arka arkaya devam ettiği için, eğitim süreci Erken Durdurma (Early Stopping) kriterine takılarak sonlandırılmıştır. Bu durum, modelin öğrenmeye devam etmesine rağmen doğrulama verisi üzerindeki kazancın artık anlamlı seviyede olmadığını göstermektedir.



Şekil 2. Eğitim ve Doğrulama hatasının epochlara göre değişimi.

C. Aşırı Öğrenme (Overfitting) ve İyileştirme (Revize)

Eğitim sürecinde, modelin eğitim verisini ezberleyerek genelleme yeteneğini kaybetmesi (aşırı öğrenme) riski doğrulama verisi üzerinden izlenmiştir. Şekil 2'deki grafik analiz edildiğinde, doğrulama hatasının tekrar yükselişe geçtiği şiddetli bir aşırı öğrenme durumu gözlemlenmemiştir. Ancak, doğrulama hatasının belirli bir noktadan sonra sabitlenmesi (plato), modelin artık yeni ve genelleyici bilgi öğrenemediğini göstermektedir.

Bu noktadan sonra eğitime devam edilmesi, hem hesaplama maliyetini artıracak hem de modelin eğitim setindeki gürültüyü ezberlemesine zemin hazırlayacaktır. Bu duruma karşı bir iyileştirme ve koruma stratejisi olarak Early Stopping (Erken Durdurma) mekanizması uygulanmıştır. Algoritma, doğrulama hatasındaki iyileşme durduğu anda eğitimi 421. epoch'ta sonlandırmış ve aşırı öğrenme riskini henüz oluşmadan engelleyerek en iyi model ağırlıklarını (416. epoch) kayıt altına almıştır

D. Performans Sonuçları

Modelin başarısı Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru (F1 Score) metrikleri ile ölçülmüştür. Test veri seti için elde edilen genel performans metrikleri Tablo I'de, sınıflandırma detaylarını gösteren Karmaşıklık Matrisi (Confusion Matrix) ise Tablo II'de sunulmuştur. Ayrıca, modelin test veri seti üzerindeki sınıflandırma başarımı ve oluşturduğu karar sınırı Şekil 3 'te görselleştirilmiştir.

Tablolar analiz edildiğinde, test setindeki toplam 20 örnekten 17 tanesinin gerçekte "Kabul" (1), 3 tanesinin ise "Ret" (0) etiketine sahip olduğu görülmektedir. Tablo 2'den anlaşılacağı

Tablo I
TEST VERI SETI PERFORMANS METRIKLERI

Metrik	Değer
Doğruluk	0.8500
Kesinlik	1.0000
Duyarlılık	0.8235
F1 Skoru	0.9032
Kayıp	0.2762

Tablo II
TEST SETI KARMAŞIKLIK MATRISI (CONFUSION MATRIX)

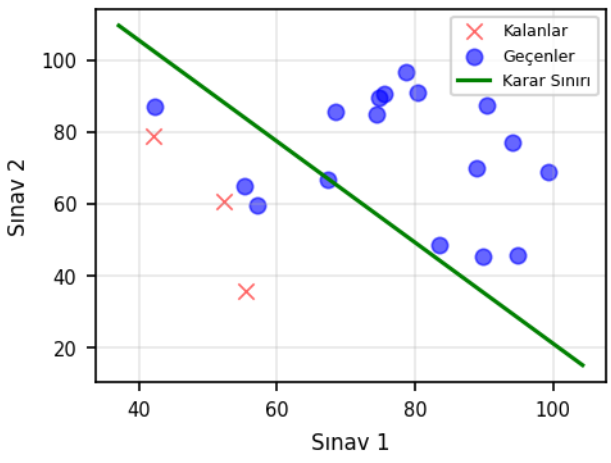
	Gerçek Durum	
Tahmin Edilen Durum	**Kabul (1)**	**Ret (0)**
Kabul (1)	14 (TP)[a]	0 (FP)[c]
Ret (0)	3 (FN)[b]	3 (TN)[d]

[a]Doğru Pozitif, [b]Yanlış Negatif, [c]Yanlış Pozitif, [d]Doğru Negatif.

üzere, model "Ret" olması gereken adayların tamamını doğru tahmin ederek 3 adet Doğru Negatif (TN) sonucu üretmiş ve bu sınıfta hiç Yanlış Pozitif (FP) hatası yapmamıştır.

"Kabul" edilmesi gereken 17 adaydan 14 tanesi doğru tespit edilmiş (Doğru Pozitif - TP), ancak 3 aday yanlışlıkla "Ret" olarak sınıflandırılmıştır (Yanlış Negatif - FN). Bu durum, Tablo 1'deki 1.0000 Kesinlik (Precision) değerini doğrulamaktadır; modelin "Kabul" dediği her aday gerçekten kabul edilmiştir. Ancak 0.8235 Duyarlılık (Recall) değeri, kabul edilmesi gereken adayların bir kısmının kaçırıldığını göstermektedir. 0.90 üzerindeki F1 Skoru (F1 Score), dengesiz sayılabilecek bu test setinde modelin başarılı bir denge kurduğunu kanıtlamaktadır.

Şekil 3 incelendiğinde, modelin oluşturduğu karar sınırının (yeşil çizgi) iki sınıfı birbirinden büyük oranda ayırdığı, ancak "Kabul" bölgesinde olması gereken bazı örneklerin sınırın "Ret" tarafında kaldığı (yanlış negatifler) görsel olarak da doğrulanmaktadır. Bu sonuçlar, 0.90 üzerindeki F1 Skoru ile birlikte değerlendirildiğinde, modelin problemi çözmekte başarılı bir performans sergilediğini göstermektedir.



Şekil 3. Test Veri Seti Üzerinde Modelin Karar Sınırı (Decision Boundary).

III. SONUÇ

Bu çalışmada, iş başvurularını değerlendirmek amacıyla adayların sınav notlarını temel alan bir lojistik regresyon modeli Python ortamında, hazır kütüphaneler kullanılmadan sıfırdan geliştirilmiştir. Deneysel analizler sonucunda modelin veriyi %85 Doğruluk (Accuracy) oranıyla sınıflandırdığı görülmüştür.

Eğitim sürecinde yapılan Kayıp (Loss) analizleri, modelin belirli bir noktadan sonra doğrulama verisi üzerinde anlamlı bir öğrenme kaydetmediğini (plato durumu) ortaya koymuştur. Bu duruma karşı hesaplama verimliliğini sağlamak ve olası bir ezberleme (overfitting) riskini engellemek amacıyla uygulanan Erken Durdurma (Early Stopping) mekanizması, eğitimi 421. epoch'ta sonlandırarak en optimum ağırlıkları koruma altına almıştır. Sonuç olarak, yüksek F1 Skoru (F1 Score) ve görselleştirilen karar sınırları, geliştirilen algoritmanın problemi çözmekte başarılı ve güvenilir olduğunu göstermektedir.