Lojistik Regresyon Yöntemiyle İkili Sınıflandırma

Mehmet Taştan   
Bilgisayar Mühendisliği  
Yıldız Teknik ÜniversitesiAnkara,Türkiye  
mehmet.tastan@std.yildiz.edu.tr

*Özet*— Bu çalışma, bir lojistik regresyon modeli kullanarak iş başvurusunda bulunan kişilerin sınav notlarına göre kabul veya reddedilme durumlarını sınıflandırmayı hedeflemektedir. Çalışmada, temel hesaplamalar için önceden oluşturulmuş makine öğrenmesi kütüphaneleri kullanılmadan, Python NumPy ile manuel olarak geliştirilen bir lojistik regresyon modeli uygulanmıştır. Veriseti, iki sınav notu ve işe alım kararını gösteren ikili bir çıktıdan oluşmaktadır. Metodoloji, veriseti %60 eğitim, %20 doğrulama ve %20 test olacak şekilde bölmeyi; parametre optimizasyonu için stokastik gradyan inişi (SGD) algoritmasını ve model eğitimi için çapraz entropi kayıp fonksiyonunu içermektedir. Model performansı doğruluk, kesinlik, geri çağırma ve F1-skora dayalı olarak değerlendirilmiştir. Veri dağılımı ve öğrenme eğrilerini görselleştirmek için Matplotlib ve mplcursors araçları kullanılmıştır. Eğitim ve doğrulama kayıp eğrileri incelenerek aşırı öğrenme (overfitting) önlenmiş ve nihai model, test verisinde tatmin edici bir performans göstermiştir.

Anahtar Kelimeler—Lojistik Regresyon, İkili Sınıflandırma, Makine Öğrenmesi, Çapraz Entropi Kaybı, Stokastik Gradyan İnişi (SGD), Model Değerlendirme, Performans Metrikleri, Veri Görselleştirme, Aşırı Öğrenme

# Giriş

Günümüzde şirketler, işe alım süreçlerini daha hızlı ve objektif hale getirebilmek için veri odaklı yaklaşımlar benimsemektedir. Özellikle, adayların sınav notları ve diğer değerlendirme kriterlerine dayalı karar alma süreçlerinde makine öğrenmesi yöntemleri etkili araçlar sunmaktadır. Bu bağlamda, ikili sınıflandırma problemleri için yaygın olarak kullanılan lojistik regresyon yöntemi, hem doğrusal sınırlayıcılar hem de olasılık tahminleri sağlaması nedeniyle tercih edilen bir yöntemdir.

Bu çalışmada, bir firmanın iş başvuru sürecinde adayların iki sınav notuna dayalı olarak işe kabul veya reddedilme durumlarının tahmin edilmesi hedeflenmiştir. Lojistik regresyon modeli, sınav sonuçlarına dayalı olarak işe alım kararını temsil eden bir olasılık değeri üretmektedir. Bu değer, 0.50 eşik değeri kullanılarak "kabul" veya "ret" olarak sınıflandırılmaktadır.

Çalışmanın katkıları şunlardır: (1) Model geliştirme sürecinde NumPy kullanılarak özelleştirilmiş bir lojistik regresyon modeli uygulanmıştır. (2) Model performansı, doğruluk, kesinlik, geri çağırma ve F1-skoru gibi metriklerle değerlendirilmiş ve görselleştirme araçları ile analiz edilmiştir. (3) Eğitim ve doğrulama kayıpları karşılaştırılarak aşırı öğrenme durumları incelenmiş ve modelin genelleme kabiliyeti optimize edilmiştir.

# Deneysel analiz

Deney bölümünde toplamda 0.01, 0.001, 0.0001, 0.00001, 0.000001 öğrenme oranları için 200, 500, 1000, 2000, 5000, 10000, 20000, 50000, 100000,200000,500000 epoklar için deney yapılmıştır. Toplamda 55 deney yapılmıştır.

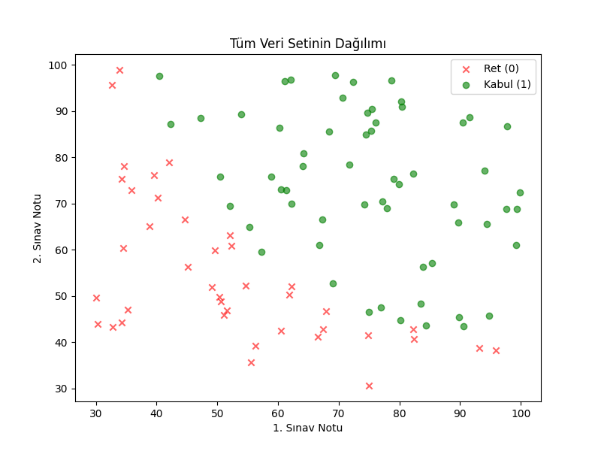
## Deneylerin Genel Açıklaması

Bu çalışmada, modelin farklı öğrenme oranları (learning rate) ve epok sayıları kullanılarak performansı değerlendirilmiştir. Toplamda 55 farklı deney gerçekleştirilmiş olup, detaylı sonuçlar learning\_rate\_epoks\_table.xlsx dosyasında tablo halinde sunulmuştur. Bu dosyada her bir deney için eğitim kaybı, doğrulama kaybı, eğitim süresi, doğruluk, kesinlik, geri çağırma ve F1-skoru gibi metrikler detaylı bir şekilde yer almaktadır. Bu analiz, modelin genelleme kapasitesi, öğrenme eğilimleri ve olası aşırı öğrenme (overfitting) durumlarını değerlendirmek amacıyla yapılmıştır.

## Veri Setlerinin Deneysel Dağılımı

Bu çalışma için kullanılan veri seti, toplam 100 örnekten oluşmaktadır ve aşağıdaki şekilde bölünmüştür. Örneklerin ilk %60’ını eğitim, sonraki %20’sini doğrulama, kalan %20’sini test için. için kullanılmıştır.

#### Şekil 1: Tüm Veri Setinin Sınıf Dağılımı



#### Şekil 2: Test Veri Setinin Sınıf Dağılımı

ekran görüntüsü, metin, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Test veri setini incelediğimizde sadece 3 adet veri ret olarak sınıflandırıldığı için model hiç eğitilmeden hepsine kabul verse bile %85 başarıya ulaşacağı için başarı kriterlerinde sadece test başarısına bakmak mantıklı olmayacaktır.

## Performans Metriklerinin Genel Özeti

Gerçekleştirilen deneylerin sonuçları, modelin farklı öğrenme oranları ve epok sayıları üzerindeki etkisini analiz etmek için aşağıdaki tabloda özetlenmiştir. Her bir kombinasyon için eğitim kaybı, doğrulama kaybı, eğitim süresi, eğitim doğruluğu, doğrulama doğruluğu ve test doğruluğu gibi kritik performans metrikleri verilmiştir.

Bu deneylerin tam ve detaylı sonuçları, learning\_rate\_epoks\_table.xlsx dosyasında yer almakta olup, bu dosya sistematik analizler ve ileride yapılacak çalışmalara yönelik kapsamlı bir referans sunmaktadır.

1. **Tablo I: Eğitim ve Doğrulama Kayıpları (Farklı Learning Rate ve Epok Çiftleri)**

|  |  |  |  |
| --- | --- | --- | --- |
| Learning Rate | Epok | Son Eğitim Kaybı | Son Doğrulama Kaybı |
| 0.01 | 1000 | 9.4029 | 9.7718 |
| 0.01 | 10000 | 5.2723 | 1.0503 |
| 0.01 | 100000 | 2.8330 | 1.3179 |
| 0.001 | 1000 | 1.7793 | 1.2336 |
| 0.001 | 10000 | 0.8225 | 0.1843 |
| 0.001 | 100000 | 0.6256 | 0.3633 |
| 0.0001 | 1000 | 0.7016 | 0.6701 |
| 0.0001 | 10000 | 0.5168 | 0.4487 |
| 0.0001 | 100000 | 0.2824 | 0.2506 |
| 0.00001 | 1000 | 0.6643 | 0.7097 |
| 0.00001 | 10000 | 0.6390 | 0.6769 |
| 0.00001 | 100000 | 0.6318 | 0.6801 |

1. Tablo II: Doğruluk Oranları (Farklı Learning Rate ve Epok Çiftleri)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Learning Rate | Epok | Eğitim Doğruluğu | Doğrulama Doğruluğu | Test Doğruluğu |
| 0.01 | 1000 | 53.33% | 55.00% | 85.00% |
| 0.01 | 10000 | 90.00% | 95.00% | 90.00% |
| 0.01 | 100000 | 90.00% | 90.00% | 85.00% |
| 0.001 | 1000 | 53.33% | 55.00% | 85.00% |
| 0.001 | 10000 | 85.00% | 95.00% | 95.00% |
| 0.001 | 100000 | 85.00% | 90.00% | 85.00% |
| 0.0001 | 1000 | 53.33% | 55.00% | 85.00% |
| 0.0001 | 10000 | 80.00% | 90.00% | 90.00% |
| 0.0001 | 100000 | 88.33% | 90.00% | 85.00% |
| 0.00001 | 1000 | 63.33% | 50.00% | 85.00% |
| 0.00001 | 10000 | 71.67% | 65.00% | 85.00% |
| 0.00001 | 100000 | 71.67% | 60.00% | 85.00% |

Yukarıdaki tablolar, farklı öğrenme oranları ve epok sayılarının modelin performansı üzerindeki etkilerini detaylı şekilde göstermektedir. Tablo I, eğitim ve doğrulama kayıplarının (loss) incelendiği bir özet sunarken, Tablo II ise eğitim, doğrulama ve test doğruluklarını karşılaştırmaktadır. Tablolara göre belirgin olan bazı kritik bulgular aşağıdaki gibi özetlenebilir:

Öğrenme Oranı ve Epokların Eğitim/Doğrulama Kaybına Etkisi: Öğrenme oranı 0.01 iken, yüksek epok sayılarında doğrulama kaybında azalma gözlenirken, daha küçük öğrenme oranlarında (örneğin 0.00001) doğrulama kaybının belirgin bir şekilde azalmadığı ve modelin öğreniminin yavaşladığı fark edilmiştir. Özellikle 0.0001 öğrenme oranı ile 100.000 epokluk durumda eğitim ve doğrulama kaybı minimum seviyeye ulaşmış, bu da optimal öğrenmenin gerçekleştiğini göstermektedir. Bununla birlikte, çok yüksek epoklar için düşük öğrenme oranları aşırı uyum (overfitting) riskine işaret edebilir.

Eğitim ve Doğrulama Doğruluğu Karşılaştırması: Tablo II, doğruluk oranları açısından dengeli bir performans sergileyen kombinasyonları açıkça göstermektedir. Örneğin, 0.001 öğrenme oranı ile 10.000 epok kullanıldığında hem eğitim hem doğrulama doğruluğunun %95 seviyesine ulaştığı görülmüştür. Ancak düşük öğrenme oranlarında (0.00001), doğrulama doğruluğundaki düşüş modelin öğrenme kapasitesinin sınırlandığını veya yetersiz eğitim yapılmış olabileceğini göstermektedir.

Underfitting ve Overfitting Durumları: Daha düşük epok ve daha yüksek öğrenme oranı kombinasyonlarında (örneğin, 0.01 öğrenme oranı ve 1000 epok) underfitting eğilimleri gözlemlenmiştir. Bu durum, modelin eğitim setinden yeterince öğrenemediğine işaret etmektedir. Öte yandan, 0.0001 öğrenme oranı ve 100.000 epok kombinasyonunda doğrulama doğruluğu ve eğitim doğruluğunun yakın olması, modelin dengeli bir şekilde öğrendiğini göstermektedir.

Test Doğruluğu Analizi: Test doğruluğu değerleri, genellikle eğitim ve doğrulama doğruluklarına benzer seviyelerde sabit kalmıştır. Bununla birlikte, bazı durumlarda (örneğin 0.01 öğrenme oranı ve 1000 epok), eğitim ve doğrulama doğruluklarının düşük olduğu durumlarda test doğruluğunun %85 seviyesinde sabit kalması dikkat çekicidir. Bu durum, modelin genelleme kapasitesinin iyi olduğunu, ancak daha iyi öğrenme parametreleriyle performansının artabileceğini göstermektedir.

Bu analizlere dayanarak, 0.001 öğrenme oranı ile 10.000 epok kombinasyonu, hem kayıp değerleri hem de doğruluk oranları açısından en iyi performansı sunan kombinasyon olarak öne çıkmaktadır. Diğer yandan, düşük öğrenme oranları (0.00001 gibi) daha uzun eğitim süreçleri gerektirmekte, ancak model performansında belirgin bir iyileşme sağlayamamaktadır. Bu bulgular, gelecekteki model optimizasyonları için önemli bir rehber olarak değerlendirilebilir.

## Öğrenme Oranı ve Epokların Çapraz Entropi Kayıplarına Etkisi

Bu bölümde, farklı öğrenme oranları (learning rate) ve epok sayılarıyla eğitim sırasında kaydedilen çapraz entropi kaybı (Cross-Entropy Loss) grafiklerinin detaylı analizine odaklanılacaktır. Analiz edilen üç farklı grafik, modelin performansını görsel olarak ortaya koymakta ve öğrenme oranı ile epok sayısının model doğruluğu üzerindeki etkilerini net bir şekilde göstermektedir.

#### **Grafik: Learning Rate: 0.01, Epok Sayısı: 1000**

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Bu grafik, eğitim ve doğrulama kayıplarının istikrarlı bir şekilde azalmadığını, aksine dalgalı bir seyir izlediğini göstermektedir. Özellikle doğrulama kaybındaki keskin düşüşler, modelin öğrenme sürecinde istikrarsızlık yaşadığını ifade etmektedir.

Öğrenme oranının (0.01) yüksek olması, modelin aşırı büyük adımlarla parametreleri optimize etmesine neden olmuş olabilir. Bu durum, dalgalı bir seyirle overfitting veya underfitting sorunlarını işaret eder. Eğitim ve doğrulama kayıpları arasındaki uyumsuzluk, modelin yeterince öğrenemediğini veya veri dağılımını kötü genelleştirdiğini gösterir.

#### **Grafik: Learning Rate: 0.001, Epok Sayısı: 10000**

metin, çizgi, diyagram, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Eğitim ve doğrulama kayıplarındaki düzenli azalma, modelin istikrarlı bir öğrenme süreci izlediğini göstermektedir. Özellikle doğrulama kaybının belirgin bir şekilde azalması, modelin test verisinde genelleme başarısının yüksek olduğunu göstermektedir.

Orta düzeyde bir öğrenme oranı (0.001) ve epok sayısının (10,000) iyi dengelenmiş olması, modelin doğru bir şekilde optimize edilmesine olanak tanımış olabilir. Bu grafik, iyi bir genelleme ve düşük overfitting riski ile başarılı bir öğrenme süreci örneği sunmaktadır.

#### **Grafik: Learning Rate: 0.0001, Epok Sayısı: 100000**

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Eğitim ve doğrulama kayıplarındaki paralel azalma, modelin hem eğitim hem de doğrulama veri setlerinde tutarlı bir şekilde iyileştiğini ortaya koymaktadır. Doğrulama kaybı, eğitim kaybından biraz daha yüksek bir seviyede kalsa da, bu fark oldukça düşüktür.

Düşük öğrenme oranı (0.0001) ve yüksek epok sayısı (100,000), modelin daha küçük adımlarla ancak daha fazla tekrar ile optimize edilmesini sağlamıştır. Bu ayar, aşırı öğrenme (overfitting) riskini azaltarak, eğitim ve doğrulama performansı arasındaki dengeyi güçlendirmiştir. Ancak, eğitim süresinin uzamış olması bu kombinasyonun maliyetini artırabilir.

# Sonuç

Bu çalışma, lojistik regresyon modeli kullanılarak öğrenme oranı (learning rate) ve epok sayılarının model performansı üzerindeki etkilerini analiz etmek amacıyla 55 farklı kombinasyonu değerlendirmiştir. Deneysel analizlerde eğitim ve doğrulama kayıpları, doğruluk oranları, eğitim süresi gibi performans metrikleri üzerinden değerlendirme yapılmıştır.

## Genel Bulgular

### Optimal Öğrenme Oranı ve Epok Sayısı: Learning Rate=0.001 ve Epochs=10,000 kombinasyonu, hem eğitim hem de doğrulama setlerinde en iyi sonuçları sağlamıştır. Bu kombinasyonda doğruluk oranı %95’e ulaşırken, eğitim ve doğrulama kayıplarındaki azalma modelin genelleme başarısını ortaya koymaktadır.

### Underfitting Durumları: Yüksek öğrenme oranları (lr=0.01) ve düşük epok sayıları, modelin veri setinden yeterince öğrenememesine neden olmuştur. Örneğin, lr=0.01,epochs=200 kombinasyonu, yüksek eğitim ve doğrulama kayıpları ile belirgin underfitting göstermiştir.

### Overfitting Durumları: Düşük öğrenme oranları (lr=0.00001) ve yüksek epok sayıları, eğitim setinde başarılı sonuçlar üretirken doğrulama setlerinde daha düşük performansa neden olmuş ve modelin aşırı öğrenim gösterdiğini işaret etmiştir.

### Dengeli Kombinasyonlar: Orta düzey öğrenme oranları ve yüksek epok sayıları, genellikle modelin dengeli bir şekilde öğrenmesini sağlamıştır. lr=0.0001,epochs=100,000 bu duruma bir örnek olup, eğitim ve doğrulama kayıplarındaki paralel azalma modelin dengeli ve genelleme başarısı yüksek bir performans sunduğunu göstermiştir

## Çapraz Entropi Kayıpları

Eğitim ve doğrulama kayıp grafikleri incelendiğinde şu önemli noktalar gözlemlenmiştir:

* Yüksek öğrenme oranları dalgalı kayıp eğrilerine neden olmuş, bu da modelin istikrarsız bir öğrenme süreci geçirdiğini göstermiştir.
* Düşük öğrenme oranları, kayıpların düzenli azalmasını sağlamış ancak eğitim süresini önemli ölçüde uzatmıştır.
* Orta düzey öğrenme oranları (lr=0.001) ve yeterli epok sayıları (epochs=10,000), modelin hem eğitim hem de doğrulama kayıplarını etkili bir şekilde minimize etmiştir.

## Eğitim ve Doğrulama Doğrulukları

Tablo II’de özetlenen doğruluk oranları, farklı öğrenme oranları ve epok sayılarına göre modelin genelleme kapasitesini değerlendirmek için kullanılmıştır:

* lr=0.001, epochs=10,000 kombinasyonu hem eğitim (%95) hem de doğrulama (%95) setlerinde en yüksek doğruluğa ulaşmıştır.
* Daha düşük öğrenme oranları (lr=0.00001) ve yüksek epok sayıları, doğrulama doğruluğunda belirgin bir iyileşme sağlayamamıştır.
* Test doğrulukları, genellikle eğitim ve doğrulama doğrulukları ile tutarlı sonuçlar vermiş, ancak bazı durumlarda (%85 seviyesinde sabit kalan) beklenmedik farklar gözlemlenmiştir.

## Performans Metrikleri Üzerine Detaylı Değerlendirme

* Eğitim Süresi: Düşük öğrenme oranları ve yüksek epok sayıları, eğitim sürelerini belirgin şekilde artırmıştır. Örneğin, lr=0.0001,epochs=500,000 kombinasyonu oldukça uzun bir süre gerektirmiştir.
* Eğitim ve Doğrulama Kaybı Dengesi: Eğitim kaybı ve doğrulama kaybı arasındaki farklar, modelin genelleme kapasitesini belirlemek için önemli bir gösterge olmuştur. Düşük farklar, genellikle başarılı bir öğrenme sürecine işaret ederken, büyük farklar aşırı öğrenme riskine işaret etmiştir.

## Confusion Matrix ile değerlendirme

En iyi sonucumuz olan lr=0.001, epochs=10,000 kombinasyonu incelersek;

#### Şekil 3: Eğitim Confusion Matrix

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

Açıklama otomatik olarak oluşturuldu

Modelin eğitim verilerinde sadece 9 kere sınıflandırmada red olan bir kişiye Kabul demiştir.

#### Şekil 4: Doğrulama Confusion Matrix

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

Açıklama otomatik olarak oluşturuldu

Modelin doğrulama esnasında sadece 1 kere sınıflandırmada red olan bir kişiye Kabul demiştir.

#### Şekil 4: Test Confusion Matrix

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

Açıklama otomatik olarak oluşturuldu

Model test esnasında sadece 1 kere sınıflandırmada red olan 1 kişiye Kabul demiştir.

Bu üç matris bize gösteriyor ki modelimiz gerçekte kabul alması gereken bir kişiyi reddetmiyor. Veri setimiz sınırlı olduğu için kesinlikle modelimiz Kabul olan kişiyi reddetmiyor kanısına varamyabiliriz ama bu çalışma için modelimiz sadece reddedilmesi gereken 10 kişiyi de Kabul etmiştir. Aslında firma mülakat süreçlerine bakacak olursak Kabul alması gereken kişinin red alması daha büyük problemdir. Bu yüzden modelimiz kabul alması gereken kişileri reddetmediği için iyi bir sonuç gösteriyor. Diğer yandan reddetmesi gerekirken kabul ettiği 10 kişi ekstra mülakatta elenebilir.

## Çalışılan Veri üzerine değerlendirme

Çalışılan verinin 100 adet’datadan oluşması ve özellikle ilk 60’ının eğitim sonraki 20 verinin doğrulama sonraki 20’nin de testte kullanılması hem data azlığı nedeniyle hem dataların karıştırılamamsı ve ikili sınıflandırma sonuçlarının dengeli dağılmaması sebebiyle belirli sınırlamalar olmuştur. Örneğin test verisinde sadece 3 adet negatif sınıfı olduğu için modeli hiç eğitmeden hepsine 1 dediğimizde %85 başarı elde edebiliyoruz. O yüzden metrikleri değerlendirirken bunu da göz önüne alarak sadece test başarısı yerine tüm başarıların en iyi olduğu öğrenme oranı epok sayısı çifti seçilmiştir yoksa %100 test başarısına ulaşan da öğrenme sayısı, epok çifti vardır fakat hem grafiklerinden hem de eğitim ve doğrulama başarı parametrelerinden bunun en iyi olmadığını gözlemleyebiliriz.

## Sonuçların Özetlenmesi

En iyi performans, lr=0.001 ve epochs=10,000 kombinasyonunda elde edilmiştir.

Hem eğitim hem de doğrulama setlerinde düzenli kayıp azalması ve yüksek doğruluk oranları, modelin genelleme başarısını doğrulamıştır.

Overfitting ve underfitting gibi sorunlar, öğrenme oranı ve epok sayılarının doğru ayarlanması ile minimize edilmiştir.