



CEZALANDIRILMIŞ LOJİSTİK REGRESYON YÖNTEMLERİNİN
KARŞILAŞTIRILMASI

Pınar KILINÇ ÖZARSLAN

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HAZİRAN 2024

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Pınar KILINÇ ÖZARSLAN

28/06/2024

CEZALANDIRILMIŞ LOJİSTİK REGRESYON YÖNTEMLERİNİN
KARŞILATIRILMASI
(Yüksek Lisans Tezi)

Pınar KILINÇ ÖZARSLAN

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Haziran 2024

ÖZET

İstatistiksel çalışmalarda değişkenler arasındaki matematiksel ilişkiyi modellemek için kullanılan yöntemlerden biri çoklu doğrusal regresyon analizidir. Bağımlı değişkenin iki kategorili olduğu durumlarda ise ikili lojistik regresyon analizi kullanılır. Bağımsız değişkenler arasında yüksek derecede korelasyon olması durumunda çoklu bağlantı sorunu ortaya çıkabilir. Bu sorunun üstesinden gelebilmek ve tutarlı değişken seçimi yapabilmek için L_1 ve L_2 düzenleme tekniklerini kullanarak elde edilen cezalandırılmış yöntemler geliştirilmiştir. Bu yöntemlerin lojistik regresyon modeline uyarlanmasıyla cezalandırılmış lojistik regresyon yöntemleri elde edilir. Bu çalışmada cezalandırılmış lojistik regresyon yöntemlerinden Ridge, Lasso, Elastik-Ağ, Uyarlanabilir Lasso ($\gamma=1$), Uyarlanabilir Lasso ($\gamma=2$), Uyarlanabilir Elastik-Ağ ($\gamma=1$), Uyarlanabilir Elastik-Ağ ($\gamma=2$), Esnek Uyarlanabilir Lasso yöntemleri ele alınmıştır. Bu tez çalışmasının amacı, literatürde yer alan Esnek Uyarlanabilir Lasso regresyon yönteminin lojistik regresyon yöntemine uyarlanması ve değişken seçimi yapabilen en iyi cezalandırılmış lojistik regresyon yöntemini bulmaktır. Bu amaçla, ele alınan cezalandırılmış lojistik regresyon yöntemlerinde farklı örneklem çapları ve değişken sayıları ele alınarak, farklı korelasyon yapıları için bir simülasyon çalışması dizayn edilmiştir. Tüm yöntemlerin performanslarını değerlendirmek amacıyla yöntemlere ilişkin ortalama doğru sınıflandırma oranına, standart sapma değerine, tahmin edilen regresyon katsayısının sıfırdan farklı olanlarının sayısına ve R-Studio’da geçirdikleri süreye bakılmıştır. Ayrıca bu yöntemlerin performanslarını karşılaştırmak amacıyla iki farklı gerçek veri seti de ele alınmıştır. Çalışma sonucunda küçük gözlem sayısına veya küçük değişken sayısına sahip olan gerçek veri seti ve simülasyon çalışmasında ortak olarak görülen çoklu bağlantı sorunu varlığında Elastik-Ağ lojistik regresyon yöntemi en yüksek ortalama doğru sınıflandırma oranını vererek en iyi yöntem olarak kabul edilmiştir. Gözlem sayısı ve değişken sayısı arttıkça Esnek Uyarlanabilir Lasso lojistik regresyon yöntemi en yüksek ortalama doğru sınıflandırma oranını vererek en iyi yöntemler arasında yer almıştır.

Bilim Kodu : 20513
Anahtar Kelimeler : Lojistik regresyon, çoklu bağlantı, cezalandırılmış, Ridge, Lasso, Elastik-Ağ, Uyarlanabilir, Esnek Uyarlanabilir
Sayfa Adedi : 69
Danışman : Prof. Dr. Hülya OLMUŞ

COMPARISON OF PENALIZED LOGISTIC REGRESSION METHODS

(M. Sc. Thesis)

Pınar KILINÇ ÖZARSLAN

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2024

ABSTRACT

One of the methods used to model the mathematical relationship between variables in statistical studies is multiple linear regression analysis. When the dependent variable has two categories, binary logistic regression analysis is used. If there is a high degree of correlation between independent variables, multicollinearity problems may arise. To overcome this problem and make consistent variable selection, penalized methods using L_1 and L_2 editing techniques have been developed. Penalized logistic regression methods are obtained by adapting these methods to the logistic regression model. In this study, penalized logistic regression methods are Ridge, Lasso, Elastic-Net, Adaptive Lasso ($Y=1$), Adaptive Lasso ($Y=2$), Adaptive Elastic-Net ($Y=1$), Adaptive Elastic-Net ($Y=2$), Relaxed Adaptive Lasso methods are discussed. The aim of this thesis is to adapt the Relaxed Adaptive Lasso regression method in the literature to the logistic regression method and to find the best penalized logistic regression method that can perform variable selection. For this purpose, a simulation study was designed for different correlation structures by considering different sample sizes and variable numbers with the penalized logistic regression methods. In order to evaluate the performance of all methods, the average correct classification rate of the methods, the standard deviation value, the number of non-zero predicted regression coefficients and the time spent in R-Studio were examined. In addition, two different real data sets were considered to compare the performances of these methods. As a result of the study, in the presence of a real data set with a small number of observations or a small number of variables and multicollinearity problems, which are common in simulation studies by giving the highest average correct classification rate, the Elastic-Net logistic regression method was accepted as the best method. As the number of observations and the number of variables increased, the Relaxed Adaptive Lasso logistic regression method became one of the best methods by giving the highest average correct classification rate.

Science Code : 20513
Key Words : Logistic regression, multicollinearity, penalized, Ridge, Lasso, Elastic-Net, Adaptive, Relaxed Adaptive
Page Number : 69
Supervisor : Prof. Dr. Hülya OLMUŞ

TEŞEKKÜR

Bu tezin hazırlanmasında göstermiş olduğu ilgi, görüş ve eleştirileri için danışman hocam Sayın Prof. Dr. Hülya OLMUŞ'a, bana her zaman inanan, bana güç veren sevgili annem Yıldız KILINÇ ve babam Ali KILINÇ'a, destekleri ve sevgileri ile her zaman yanımda olan sevgili kardeşlerim Damla KILINÇ ve Saadet Irmak KILINÇ'a, çalışmamın başından sonuna kadar yanımda olan ve göstermiş olduğu sabır ve emeği için sevgili eşim Emirhan ÖZARSLAN'a sonsuz teşekkürlerimi sunarım.



İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
RESİMLERİN LİSTESİ	xii
SİMGELER VE KISALTMALAR.....	xiii
1. GİRİŞ.....	1
2. YÖNTEM.....	7
2.1. Lojistik Regresyon	7
2.1.1. İkili (Binary) lojistik regresyon.....	7
2.2. Çoklu Bağlantı Sorunu.....	10
2.3. Cezalandırılmış (Penalized) Lojistik Regresyon.....	12
2.3.1. Düzenleme (Reguralization) teknikleri	12
2.3.2. Ayarlama (Tuning) parametresi	12
2.4. Çapraz Doğrulama Yöntemi	13
2.5. Cezalandırılmış (Penalized) Lojistik Regresyon Yöntemleri	14
2.5.1. Ridge lojistik regresyon	14
2.5.2. Lasso lojistik regresyon	16
2.5.3. Elastik-Ağ lojistik regresyon.....	18
2.5.4. Uyarlanabilir (Adaptive) Lasso lojistik regresyon.....	20
2.5.5. Uyarlanabilir (Adaptive) Elastik-Ağ lojistik regresyon.....	21
2.5.6. Esnek (Relaxed) Uyarlanabilir (Adaptive) Lasso lojistik regresyon	23
3. SİMÜLASYON ÇALIŞMASI.....	25
3.1. Simülasyon Tasarımı.....	25

	Sayfa
3.2. Simülasyon Sonuçları.....	27
3.2.1. Birinci senaryo sonuçları	27
3.2.2. İkinci senaryo sonuçları	37
3.2.3. Senaryo sonuçlarına dair özet tablolar	48
4. UYGULAMA.....	51
4.1. Veri Setleri ve Sonuçları	51
4.1.1. Pirinç türü sınıflandırma	51
4.1.2. Parkinson.....	54
5. SONUÇ VE ÖNERİLER	57
KAYNAKLAR	63
ÖZGEÇMİŞ	69

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. Olasılık değerleri	7
Çizelge 2.2. Karışıklık matrisi	9
Çizelge 3.1. $n=20(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	28
Çizelge 3.2. $n=50(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	29
Çizelge 3.3. $n=100(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	30
Çizelge 3.4. $n=500(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	31
Çizelge 3.5. $n=50(p=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	32
Çizelge 3.6. $n=100(p=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	33
Çizelge 3.7. $n=500(p=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	34
Çizelge 3.8. $n=100(p=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	35
Çizelge 3.9. $n=500(p=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri i	36
Çizelge 3.10. $n=500(p=100)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri	37

Çizelge	Sayfa
Çizelge 3.11. $p=20(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	38
Çizelge 3.12. $p=50(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	39
Çizelge 3.13. $p=100(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	40
Çizelge 3.14. $p=500(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	41
Çizelge 3.15. $p=50(n=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	42
Çizelge 3.16. $p=100(n=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	43
Çizelge 3.17. $p=500(n=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	44
Çizelge 3.18. $p=100(n=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	45
Çizelge 3.19. $p=500(n=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	46
Çizelge 3.20. $p=500(n=100)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, STD, SS, ortalama DSO değerleri.....	47
Çizelge 3.21. Lojistik regresyon yöntemlerinin numaralandırılması	48
Çizelge 3.22. Birinci senaryo sonuçlarına ait özet tablosu.....	49
Çizelge 3.23. İkinci senaryo sonuçlarına ait özet tablosu	49
Çizelge 4.1. Pirinç türü sınıflandırma veri seti için özet bilgiler tablosu	53

Çizelge	Sayfa
Çizelge 4.2. Pirinç türü sınıflandırma veri seti için tüm lojistik regresyon yöntemlere ait DSO ve SS değerleri.....	53
Çizelge 4.3. Parkinson veri seti için ilk 14 değişkene ait özet bilgiler tablosu	55
Çizelge 4.4. Parkinson veri seti için tüm lojistik regresyon yöntemlere ait DSO ve SS değerleri.....	56



RESİMLERİN LİSTESİ

Resim	Sayfa
Resim 2.1. k -katlı çapraz doğrulama yöntemi gösterim şeması.....	13
Resim 2.2. Ridge ve Lasso regresyon yöntemlerinin tahminleri	17
Resim 2.3. Ridge, Lasso, Elastik-Ağ regresyon yöntemlerinin tahminleri	19
Resim 3.1. 10-katlı çapraz doğrulama yöntemi gösterim şeması	26
Resim 4.1. Pirinç türü sınıflandırma veri setinin değişkenlerine ait korelasyon tablosu.....	52
Resim 4.2. Parkinson veri seti için ilk 14 değişkene ait korelasyon tablosu	55

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklamalar

R^2

Belirtme Katsayısı

p

Değişken Sayısı

λ

Ayarlama Parametresi

ε

Hata terimi

φ

Gevşeme Parametresi

n

Gözlem Sayısı

σ^2

Yığın Varyansı

ρ

Korelasyon

K

Koşul Katsayısı

β

Regresyon Katsayıları

Kısaltmalar

Açıklamalar

BIC

Bayesian Bilgi Kriteri

EKK

En Küçük Kareler

DSO

Doğru Sınıflandırma Oranı

GEO

Gen Veri Merkezi

HKT

Hata Kareler Toplamı

SS

Tahmin Edilen β Katsayılarının Sıfırdan Farklı Olanlarının Sayısı

STD

Standart Sapma Değeri

VIF

Varyans Şişirme Değer

1.GİRİŞ

Regresyon analizi, bağımlı değişken ve bağımsız değişkenler arasındaki ilişkinin tahmin edilmesini amaçlamaktadır (Dhhan, Rana ve Midi, 2016). Bağımlı değişkenin ikili veya daha fazla kategorik değerler aldığı regresyon analizine lojistik regresyon analizi denilmektedir. Lojistik regresyon analizinde de doğrusal regresyon analizinde olduğu gibi değişkenler arasındaki nedensellik ilişkisi tespit edilmeye çalışılmaktadır. Ancak lojistik regresyonda değişkenler arasındaki ilişkinin doğrusal olduğu varsayılmamaktadır. Lojistik regresyon varsayımlarına göre bağımlı değişkenin alacağı değerler karşılıklı zıt olaylar olmalı ve gerçekleşme olasılıkları toplamı 1 olmalıdır (Aldrich ve Nelson, 1984). Bağımlı değişkenin sadece iki kategoriden oluşması ya da 0 ve 1 değerlerini almasıyla ikili (binary) lojistik regresyon analizi elde edilmektedir.

Bağımsız değişkenler arasında ilişki olma durumu yani herhangi iki bağımsız değişken arasında yüksek derecede korelasyon olması durumuna çoklu bağlantı sorunu denilmektedir. Bu sorun belirtme katsayısının büyük çıkmasına, ikinci tip hatanın artmasına, katsayıların tahmininde yanlış işaret ve büyüklüğe yol açarak bağımlı ve bağımsız değişkenler arasındaki ilişki hakkında yanlış sonuçlara varılmasına neden olabilmektedir. Çoklu bağlantı sorununun üstesinden gelebilmek için L_1 ve L_2 düzenleme teknikleri kullanılarak cezalandırılmış regresyon yöntemleri geliştirilmiştir (James vd. 2014, Bruce ve Bruce, 2017). Cezalandırılmış regresyon yöntemlerinde katsayılar üzerinde ceza terimleri kullanılarak regresyon katsayıları tahmin edilir ve çoğunlukla regresyon katsayılarına bir miktar yavaşlık eklenerek katsayıların varyansının azaltılması amaçlanmaktadır.

İlk geliştirilen cezalandırılmış regresyon yöntem Ridge regresyon yöntemidir. 1970 yılında Hoerl ve Kennard tarafından geliştirilmiştir. L_2 düzenleme tekniğini kullanarak değişkenlerin tamamını modele dahil eden regresyon modeline Ridge regresyon denir. Ridge regresyon lojistik regresyon analizinde kullanılmaktadır. Lojistik regresyonun negatif log-olabilirlik fonksiyonuna Ridge regresyonun ceza teriminin eklenmesiyle Ridge lojistik regresyon yöntemi elde edilmektedir.

Ridge regresyon yönteminden sonra Lasso regresyon yöntemi Tibshirani (1996) tarafından geliştirilmiştir. L_1 düzenleme tekniğini kullanarak değişkenlerden bazılarını modele dahil eden regresyon modeline Lasso regresyon denir. Ridge regresyon yönteminde katsayılar sıfıra indirgenemediğinden değişken seçimi yapılamaz.

Lasso regresyon yönteminde ise katsayılar sıfıra indirgenebilmektedir. Bu durum Lasso regresyon yönteminin değişken seçimi yapmasına sade ve anlaşılır modeller sunmasına neden olmaktadır. Lojistik regresyonun negatif log-olabilirlik fonksiyonuna Lasso regresyonun ceza teriminin eklenmesiyle Lasso lojistik regresyon yöntemi elde edilmektedir. Ancak Lasso regresyon yönteminin dezavantajları bulunmaktadır. Birinci dezavantaj; Lasso regresyon yöntemi çoklu bağlantı sorunu varlığında yüksek derecede ilişkili değişkenlerden birini rastgele seçmektedir. Bu durumda modelde olması gereken bir değişkenin çoklu bağlantı nedeniyle modelden çıkarılmış olması sorunuyla karşılaşmaktadır. İkinci dezavantaj; değişken sayısının gözlem sayısından büyük olduğu ($p > n$) yüksek boyutlu durumlarda Lasso regresyon yöntemi en fazla n tane değişkeni modele dâhil edebilmektedir. Bu durum, özellik seçimi için sınırlayıcı olmaktadır. Gözlem sayısının değişken sayısından büyük olduğu ($n > p$) küçük boyutlu durumlarda, bağımsız değişkenler arasında yüksek korelasyon varlığında Lasso regresyon yönteminin Ridge regresyon yönteminden daha iyi bir yöntem olduğu söylenebilir (Tibshirani, 1996).

Lasso regresyon yönteminin dezavantajlarının azaltılması amacıyla Zou ve Hastie (2005) L_1 ve L_2 düzenleme tekniklerini birleştiren Elastik-Ağ regresyon yöntemini önermişlerdir. Elastik-Ağ regresyon yöntemi değişken seçiminde Lasso regresyon yönteminin, yüksek korelasyon sorunuyla başa çıkmak için ise Ridge regresyon yönteminin cezasını kullanmaktadır. Elastik-Ağ regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla yani lojistik regresyonun negatif log-olabilirlik fonksiyonuna Elastik-Ağ regresyonun ceza teriminin eklenmesiyle Elastik-Ağ lojistik regresyon yöntemi elde edilmektedir.

Uyarlanabilir (Adaptive) Lasso regresyon yöntemi, Elastik-Ağ regresyon yönteminde olduğu gibi Lasso regresyon yönteminin belirtilen dezavantajlarının azaltılması amacıyla Zou (2006) tarafından Lasso regresyon yöntemine bir ağırlık terimi eklenmesiyle geliştirilmiştir (Zhang, Zhao, Lu ve Xu, 2022). Bu yöntemde, Oracle özelliklerinden yararlanılmaktadır (Algarnal ve Lee, 2015). Oracle özelliklerine göre, tahmin edici olasılığı

bire yaklaşan bir modelde sıfır olmayan katsayıları doğru şekilde seçebilmelidir ve sıfır olmayan katsayılar asimptotik olarak normal dağılmaktadır. Oracle özellikleri Uyarlanabilir Lasso regresyon yöntemini daha optimal hale getirmektedir. Uyarlanabilir Lasso regresyon yönteminin Lasso regresyon yönteminden daha üstün olması Oracle özelliklerini kullanmasıdır. Böylece Lasso regresyon yöntemine göre daha hızlı ve daha tutarlı değişken seçimi yapabilir. Lasso lojistik regresyon yöntemine ağırlık teriminin eklenmesiyle Uyarlanabilir Lasso lojistik regresyon yöntemi elde edilmiştir.

Uyarlanabilir Elastik-Ağ regresyon yöntemi, Zou ve Zhang (2009) tarafından $p > n$ olduğu yüksek boyutlu durumlarda değişken seçimi ve tahmini yapabilmek için Elastik-Ağ regresyon yönteminden geliştirilmiştir (Zou ve Zhang, 2009). Uyarlanabilir Elastik-Ağ regresyon yönteminde, Elastik-Ağ regresyon yöntemi ile Uyarlanabilir Lasso regresyon yöntemi birleştirilmektedir. Uyarlanabilir Elastik-Ağ regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla Uyarlanabilir Elastik-Ağ lojistik regresyon yöntemi elde edilmektedir.

Uyarlanabilir Lasso regresyon yönteminin özel hali ise, Meinshausen tarafından geliştirilen Esnek (Relaxed) Lasso regresyon yöntemidir (Meinshausen, 2006). Esnek Lasso regresyon yönteminde amaç, her düzenleme paramatresi için değişkenleri herhangi bir cezalandırma olmadan seçebilmektir. Esnek Lasso regresyon yöntemi, modelde olan önemsiz değişkenleri silerek modelin doğruluğunu arttırmayı amaçlamaktadır. Uyarlanabilir Lasso regresyon yöntemi ve Esnek Lasso regresyon yöntemlerinin avantajlarını göz önünde bulundurarak bu iki yöntemi birleştiren Esnek Uyarlanabilir Lasso regresyon yöntemi geliştirilmiştir. Bu yöntem, Esnek Lasso regresyon yöntemine Uyarlanabilir Lasso regresyon yönteminin ağırlık vektörünün eklenmesiyle elde edilmektedir.

Bu tez çalışmasında, Esnek Uyarlanabilir Lasso regresyon yönteminin ceza teriminin lojistik regresyonun negatif log-olabilirlik fonksiyonuna eklenmesiyle Esnek Uyarlanabilir Lasso lojistik regresyon yöntemi elde edilmiştir.

Cezalandırılmış regresyon ve cezalandırılmış lojistik regresyon yöntemlerine ilişkin literatürde farklı alanlarda birçok çalışma yer almaktadır. Pregibon (1981), ikili lojistik regresyon yönteminde; etkinliği, aykırı gözlemleri ve belirleme ölçülerini incelemiştir.

Bonney (1987), lojistik regresyon modelinin geliştirilmesi için çalışmalar yapmıştır. Aguilera, Escabias ve Valderrama (2006), yüksek derecede çoklu bağlantı durumunda ve çoklu doğrusallık altında lojistik model parametrelerinin tahminini geliştirmek ve sürekli ortak değişkenlerle problemin boyutunu azaltmak için, lojistik modelin ortak değişkenleri olarak orijinal tahmin edicilerin optimum temel bileşenlerinin azaltılmış bir kümesinin kullanılmasını önermişlerdir. Önerilen temel bileşende, lojistik regresyon modelinin performansını gösteren ve optimum temel bileşenlerin seçimine yönelik farklı yöntemlerin karşılaştırıldığı bir simülasyon çalışmasını geliştirerek analiz etmişlerdir. Bielza ve diğerleri (2011), mikrodizi verileriyle üç kanser sınıflandırma probleminde önerilen yöntemde Lasso lojistik regresyon ve Ridge lojistik regresyonun davranışını karşılaştırmışlar ve bu karşılaştırmaya dayalı deneysel sonuçları göstermişlerdir. Arı ve Önder (2013), lojistik regresyon ve diğer farklı regresyon modellerini karşılaştırarak bu regresyon modellerinin hangi durumlarda kullanılabileceğini incelemişler bunları histogram grafikleriyle göstermişlerdir. Ertuğrul (2018), farklı tuz oranlarına sahip göllerde, popülasyonun büyüklüğünün alan ve zaman olarak modellenmesi için lojistik regresyon ve Ridge lojistik regresyon yöntemlerini kullanmıştır. Araveeporn (2021), tarafından $p > n$ olduğu yüksek boyutlu durumlar için ikili lojistik regresyon yöntemine uygun olacak şekilde gerçek veri seti ve simülasyon çalışmasını ele alarak Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon yöntemini ve Uyarlanabilir Elastik-Ağ lojistik regresyon yöntemlerine ait performansları değerlendirip karşılaştırma yapmıştır. Kullandığı gerçek veri seti ise akut miyeloid lösemili ve akut lenfoblastik lösemili 72 hastadan oluşmaktadır. Çalışmanın sonuçlarına göre, Uyarlanabilir Elastik-Ağ lojistik regresyon ve Elastik-Ağ lojistik regresyon yönteminin verinin dağılımına dayalı gen seçiminde etkili olduğu, değişken sayısı arttıkça yani yüksek boyutlu veri setlerinde Elastik-Ağ lojistik regresyonun diğer yöntemlere göre daha etkili olduğu gösterilmiştir. Lee ve diğerleri (2021), DNA mikrodizi verilerinde gen katsayılarını eşzamanlı olarak tahmin etmek, gen seçimini gerçekleştirmek ve değişkenlerin sınıflandırmasında Uyarlanabilir Lasso lojistik regresyon yönteminin performansını gözlemek istemişlerdir. Zhang, Zhao, Lu ve Xu (2022), $n > p$ olduğu yani düşük boyutlu durumlar için gerçek veri seti ve simülasyon çalışmasını ele alarak Lasso regresyon, Esnek Lasso regresyon, Uyarlanabilir Lasso regresyon, Esnek Uyarlanabilir Lasso regresyon yöntemlerinin performanslarını karşılaştırmışlardır. Kullandıkları gerçek veri seti borsada işlem gören bir şirketin 2021 yılı mali verilerinden oluşmuştur. Çalışmanın sonucuna göre, Esnek Uyarlanabilir Lasso regresyon ile Esnek Lasso regresyonun aynı yakınsama oranına

sahip olduđu ve düşük boyutlu veriler için her ikisinin de Uyarlanabilir Lasso ve Lasso'dan daha hızlı olduđu kanısına varılmıştır.

Bu tez çalışmasının amacı, literatürde yer alan Esnek Uyarlanabilir Lasso regresyonun lojistik regresyon yöntemine uyarlanması ve değişken seçimi yapabilen en iyi cezalandırılmış lojistik regresyon bulmaktır.

Bu bilgiler ışığında, bu tez çalışmasında cezalandırılmış lojistik regresyon yöntemlerinden Ridge lojistik, Lasso lojistik, Elastik-Ağ lojistik, Uyarlanabilir Lasso lojistik ($Y=1$), Uyarlanabilir Lasso lojistik ($Y=2$), Uyarlanabilir Elastik-Ağ lojistik ($Y=1$), Uyarlanabilir Elastik-Ağ lojistik ($Y=2$), Esnek Uyarlanabilir Lasso lojistik regresyon yöntemleri ele alınarak farklı örneklem çapları, farklı değişken sayıları ele alınarak, farklı korelasyon yapıları için bir simülasyon çalışması dizayn edilmiştir. Tüm yöntemlerin performanslarını değerlendirmek amacıyla yöntemlere ilişkin; ortalama doğru sınıflandırma oranına (DSO), standart sapma değerine (STD), tahmin edilen regresyon katsayısının sıfırdan farklı olanlarının sayısına (SS) ve R-Studio'da geçirdikleri süreye bakılmıştır. Ayrıca bu yöntemlerin performanslarını karşılaştırmak amacıyla iki farklı gerçek veri seti de ele alınmıştır. Birinci veri seti; gözlem sayısının değişken sayısından büyük olduđu, düşük boyutlu durumda, Kaggle veri merkezinden çekilen Pirinç türü sınıflandırma (Rice type classification) veri setidir. İkinci veri seti; değişken sayısının gözlem sayısından büyük olduđu, yüksek boyutlu durumda, GSE68719 koduyla GEO veri tabanında bulunan Parkinson veri setidir. Her iki veri setinde de simülasyon çalışmasında kullanılan cezalandırılmış lojistik regresyon yöntemleri ele alınarak tüm yöntemlerin SS ve ortalama DSO değerine bakılmıştır.

Çalışmanın ikinci bölümünde sıklıkla kullanılan ikili lojistik regresyon ve cezalandırılmış lojistik regresyon yöntemleri sunularak bu yöntemlere dair teorik bilgilere yer verilmiştir. Üçüncü bölümde simülasyon tasarımı detaylı olarak verilerek elde edilen sonuçlara yer verilmiştir. Dördüncü bölümde Kaggle veri merkezinden çekilen Pirinç türü sınıflandırma veri seti ve GSE68719 koduyla GEO veri tabanında bulunan Parkinson veri setleri tanıtılarak, veri setlerine ilişkin elde edilen sonuçlara yer verilmiştir. Beşinci bölümde çalışmadan elde edilen bulgular irdelenerek tez çalışmasına yönelik genel sonuçlar ve bu sonuçlara ilişkin yorumlar verilmiştir.

2. YÖNTEM

Bu bölümde ikili lojistik regresyon ve cezalandırılmış lojistik regresyon yöntemleri sunularak, bu yöntemlere dair teorik bilgilere yer verilmiştir.

2.1. Lojistik Regresyon

Doğrusal regresyon yönteminde, bağımlı değişken sürekli veri yapısına sahiptir. Bağımlı değişkenin kategorik veri yapısına sahip olduğu durumlarda çoklu doğrusal regresyon analizi yerine lojistik regresyon analizi kullanılmaktadır. Lojistik regresyon analizi, sınıflama ve değişken seçimi yapmaya yardımcı olan bir regresyon yöntemidir. Bağımlı değişkenin iki kategorili olduğu ya da 0 ve 1 değerlerini aldığı durumlarda uygulanan yöntem ikili lojistik regresyon analizidir.

2.1.1. İkili (Binary) lojistik regresyon

Lojistik regresyon analizinin amacı; modele en çok katkı sağlayan değişkenlerin seçilmesi, modele ait sınıflama doğruluğunun artırılması ve modelin açıklanabilir olmasıdır. Lojistik regresyon modeli ortalama, bağımlı değişkenin bir fonksiyonu olarak ifade edilir. Bu nedenle lojistik regresyon modeli için dağılımlar Bernoulli dağılımından gelmektedir (Myers, Montgomery, Goffery Vinnig, Robinson, 2012). Bağımlı değişkene ait olasılık değerleri Çizelge 2.1’de gösterilmektedir.

Çizelge 2.1. Olasılık değerleri

y_i	Olasılık Değerleri
0	$P(y_i=0) = 1 - \pi_i$
1	$P(y_i=1) = \pi_i$

İkili lojistik regresyonda bağımlı değişken $Y=1$, bağımsız değişken vektörü $X=(X_1, X_2, \dots, X_k)$ ve lojistik regresyon yönteminde $E(Y/X=x_i) = \pi(x_i)$ koşullu olasılığı olmak üzere lojistik regresyon yönteminde bağımlı değişkene ait koşullu olasılık fonksiyonu Eşitlik (2.1)’de gösterilmektedir.

$$E(Y=1/x_i) = \pi(x_i) = (e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} / 1) + e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}) \quad (2.1)$$

Lojistik regresyon yönteminde kullanılan terimlerden biri Odds oranıdır. Odds bir olayın olma olasılığının, olayın olmama olasılığına bölümüdür. Odds oranı 0 ile sonsuz arasında değer almaktadır. Eşitlik (2.2)'de Odds oranı denklemi gösterilmektedir.

$$OR = \frac{\pi}{1-\pi} \quad (2.2)$$

Odds oranı denklemi Eşitlik (2.3)'de gösterilmektedir.

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} \quad (2.3)$$

Bağımlı değişkenin alacağı değerlerin (0,1) aralığında olması için dönüşüm yapılması gerekmektedir. $\pi(x) / [1-\pi(x)]$ dönüşümü bağımlı değişkenin sınırlarını (0, ∞) aralığına taşımaktadır. Bağımlı değişkenin sınırlarının $(-\infty, +\infty)$ aralığında olması için ise $\pi(x) / [1-\pi(x)]$ oranının logaritmasının alınması gerekmektedir. Bu dönüşüm yardımıyla elde edilen yeni bağımlı değişken bağımsız değişkenin doğrusal bir fonksiyonu olmaktadır ve bu dönüşüme 'lojit dönüşüm' denilmektedir. Lojit dönüşüm regresyon katsayılarının doğrusal fonksiyonuna ulaşılmasına yardımcı olmaktadır. Eşitlik (2.4)'de lojistik regresyon yöntemine ilişkin lojit dönüşüm verilmiştir (Ekstrand, Westergren, Årestedt, Hellström, Hagell, 2022).

$$\text{lojit}\pi(x) = \ln[\pi_i / (1 - \pi_i)] = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (2.4)$$

Lojistik regresyon yönteminde parametre tahmini için en çok olabilirlik yöntemi kullanılmaktadır. Bu yöntem Eşitlik (2.5)'de verilmiştir.

$$L(\beta, y_i) = \prod_{(i=1)}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \quad (2.5)$$

Eşitlik (2.5)'deki denklemin logaritması alınarak log-olabilirlik fonksiyonu elde edilir. Bu fonksiyon en çok olabilirlik yönteminin tahmin edicisidir. Bu fonksiyon Eşitlik (2.6)'da gösterilmektedir.

$$\ln(\beta, y_i) = \sum_{i=1}^n \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - (\pi(x_i)))\} \quad (2.6)$$

n arttıkça en çok olabilirlik yönteminin tahmin edicisi yeterli, tutarlı ve asimptotik olarak normal dağılmaktadır. Veride çoklu bağlantı sorunu olduğu durumlarda ya da değişken sayısı gözlem sayısından büyük olduğu durumlarda da en çok olabilirlik yöntemi en iyi tahmin edici değildir (Hilbe, 2009).

Lojistik regresyon analizinde bağımsız değişkenin elde edilen model içinde ne kadar iyi açıklandığı model uyum iyiliği ile test edilmektedir. Model uyum iyiliği Pearson Ki-kare testi, Hosmer-Lameshow testi, sınıflandırma tablosu, sapma istatistikleri ve ROC Eğrisi altında kalan alan yöntemleriyle incelenmektedir. Pearson Ki-kare testi ve sapma istatistikleri gözlenen değer ve beklenen değer arasındaki farkla modelin uyum olup olmadığını test etmektedir. Hosmer ve Lameshow testi ise tahmin edilen olasılıkları küçükten büyüğe doğru sıralayıp bu olasılıklara göre bireyleri k tane alt gruba bölmektedir. Her bir alt grup için gözlenen değeri ve beklenen değeri hesaplanarak Ki-kare testi uygulamaktadır. Sınıflandırma tablosu ise lojistik regresyon sonucunda elde edilen karışıklık matrisi ile hesaplanmaktadır. Çizelge 2.2'de karışıklık matrisi verilmiştir.

Çizelge 2.2. Karışıklık matrisi

	Gerçek Durum	
	Pozitif	Negatif
Tahmin Edilen		
Pozitif Tahmin	Doğru Pozitif (TP)	Yanlış Negatif (FN)
Negatif Tahmin	Yanlış Pozitif (FP)	Doğru Negatif (TN)

Duyarlılık (Doğru Pozitif Oranı); gerçekte pozitif olan gözlemlerden ne kadarının pozitif olarak tahmin edildiğini,

Özgüllük (Doğru Negatif Oranı); negatif tahmin edilen gözlemlerden ne kadarının gerçekten negatif olduğunu

1-Duyarlılık (Yanlış Negatif Oranı); gerçekte pozitif olanların yanlış şekilde negatif tahmin edilme olasılığını,

1-Özgüllük (Yanlış Pozitif Oranı); gerçekte negatif olanların yanlış şekilde pozitif tahmin edilme olasılığını gösteren ölçüttür (Metz, 1978).

Doğru Sınıflandırma Oranı (DSO); $\frac{TP+TN}{TP+FN+FP+TN}$ ile hesaplanmaktadır ve doğru tahmin edilen gözlemlerin yüzdesini göstermektedir.

Yanlış Sınıflandırma Oranı; $\frac{FN+FP}{TP+FN+FP+TN}$ ile hesaplanmaktadır ve yanlış tahmin edilen gözlemlerin yüzdesini göstermektedir.

ROC eğrisi, duyarlılık ve özgüllük değerlerini kullanan bir ölçüttür. ROC eğrisinin dikey ekseninde duyarlılık, yatay ekseninde ise 1-özgüllük değeri bulunmaktadır. Bu değerler bir grafik üzerinde ROC eğrisi adı verilen bir eğriyi oluşturmaktadırlar. Oluşan bu eğri altında kalan alan hesaplanmaktadır, eğrinin altında kalan alan yani hesaplanan alan pozitif ve negatif sonuçların doğru sınıflandırma oranını göstermektedir. DSO'nun 1'e yakın olması model uyumunun iyi olduğunu belirtmektedir (Hajian-Tilaki, 2013).

2.2. Çoklu Bağlantı Sorunu

Bağımsız değişkenler arasında ilişki olma durumu yani herhangi iki bağımsız değişken arasında yüksek derecede korelasyon olması durumuna çoklu bağlantı sorunu denilmektedir. Çoklu bağlantı sorunu çoğunlukla aşağıdaki gibi nedenlerden dolayı ortaya çıkmaktadır;

- Yığından çekilen örneklerin yığını yeterince temsil etmemesi,
- Bağımsız değişken sayısının gözlem sayısından büyük olması,
- Araştırmacıdan kaynaklanan hatalar.

Çoklu bağlantı sorunu olduğu durumlarda bağımsız değişkenlerle en iyi tahmin değerleri elde edilse belirtme katsayısı (R^2) güvenilir bir şekilde yorumlanamaz.

Çoklu bağlantı sorunu; R^2 'nin çok büyük çıkmasına, ikinci tip hatanın artmasına ve katsayıların tahmininde yanlış işaret ve büyüklüğe yol açarak bağımlı ve bağımsız değişkenler arasındaki ilişki hakkında yanlış sonuçlara varılmasına sebep olabilmektedir.

Literatürde birden fazla çoklu bağlantı sorunu göstergesi vardır. Bunlar;

- Korelasyon Matrisi
- R^2
- Öz değer ve Öz vektörler
- Koşul Katsayısı (K)
- Koşul İndeksi
- VIF (Varyans Şişirme Değeri);

VIF ; bağımsız değişkenlere ilişkin korelasyon matrisinin köşegen elamanlarından hesaplanan değere denilmektedir. Bu değer Eşitlik (2.7)'de verilmektedir.

$$VIF_i = (1 - R_i^2) \quad (2.7)$$

VIF değeri 10 ve üzerindeyse çoklu bağlantı sorunu varlığından bahsedilmektedir. Bu tez çalışmasında sadece VIF değerine bakılmıştır.

Çoklu Bağlantı Sorununu Giderme Yöntemleri;

Değişkenler arasındaki ilişkiyi belirleyen katsayı tahminlerinin standart hatasını küçültmek ve daha tutarlı tahmin yapabilmek için çoklu bağlantı sorunun giderilmesi gerekmektedir. Bunun için aşağıdaki yöntemler önerilmektedir;

- Gözlem sayısı artırılarak parametre tahminlerine ait varyansların küçülmesi sağlanabilir,

- Bağımlı değişken üzerinde daha az etkili olan bağımsız değişkenler regresyon modelinden çıkarılabilir,
- Birbiriyle bağlantılı olan bağımsız değişkenlerden bazıları birleştirilerek tek bir değişken olarak modele konulabilir.

Bağımsız değişkenler arasında ilişki olması durumunda en küçük kareler tahmin edicisi en küçük varyansa sahip tahmin edici özelliğini kaybetmektedir. Böyle durumlarda regresyon katsayısı (β), $\hat{\beta}$ yansız tahmin edicisinden daha küçük varyansa sahip olan $\hat{\beta}^*$ yanlı (cezalandırılmış) tahmin edicisi ile tahmin edilmektedir. Cezalandırılmış tahmin yöntemlerinde amaç yanlılık terimini kullanarak varyansı daha küçük bir tahmin edici bulmaktır.

2.3. Cezalandırılmış (Penalized) Lojistik Regresyon

Cezalandırılmış lojistik regresyon yöntemlerinde katsayılar üzerinde ceza terimi terimleri kullanılarak regresyon katsayıları tahmin edilir ve çoğunlukla regresyon katsayılarına bir miktar yanlılık eklenerek katsayıların varyansının azaltılması amaçlanmaktadır. Veri setlerinde çoklu bağlantı sorunu olduğunda veya yüksek boyut ($p > n$) durumu varlığında klasik tahmin yöntemlerine alternatif olarak cezalandırılmış regresyon yöntemleri yaygın olarak kullanılmaktadır. Bu yöntemler büzülme veya düzenleme yöntemleri olarak da bilinmektedir (James, Witten, Hastie ve Tibshirani, 2013; Bruce ve Bruce, 2017).

2.3.1. Düzenleme (Reguralization) teknikleri

Düzenleme teknikleri L_1 ve L_2 düzenleme teknikleri olarak 2'ye ayrılmaktadır.

- L_1 düzenleme terimi, katsayıların (ağırlıkların) büyüklüğünün mutlak değerlerine, L_2 düzenleme terimi, katsayıların (ağırlıkların) büyüklüğünün karesine eşittir.
- L_1 düzenleme tekniğinde değişkenlerin çoğu önemsizdir ancak L_2 düzenleme tekniğinde değişkenlerin tamamı önemlidir.
- L_1 düzenleme tekniğinde λ (düzenleme parametresi) arttıkça katsayıların büyüklüğü 0'ya eşit çıkabilir, L_2 düzenleme tekniğinde ise 0'a yaklaşabilir ancak eşit çıkamaz.

L_1 düzenleme tekniği Eşitlik (2.8), L_2 düzenleme tekniği Eşitlik (2.9)'da gösterilmektedir.

$$\sum_{i=1}^n (y_i \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (2.8)$$

$$\sum_{i=1}^n (y_i \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (2.9)$$

2.3.2. Ayarlama (Tuning) parametresi

Cezalandırılmış lojistik regresyon yöntemlerinde, en iyi sonucun elde edilebilmesi için uygun bir değerinin seçilmesi gerekmektedir. Ayarlama parametresinin (λ) seçimi için çeşitli yöntemler bulunmaktadır. Tibshirani (2005), model seçimi için literatürde de yaygın olarak kullanılmakta olan çapraz doğrulama (cross validation) yöntemini önermiştir.

2.4. Çapraz Doğrulama Yöntemi

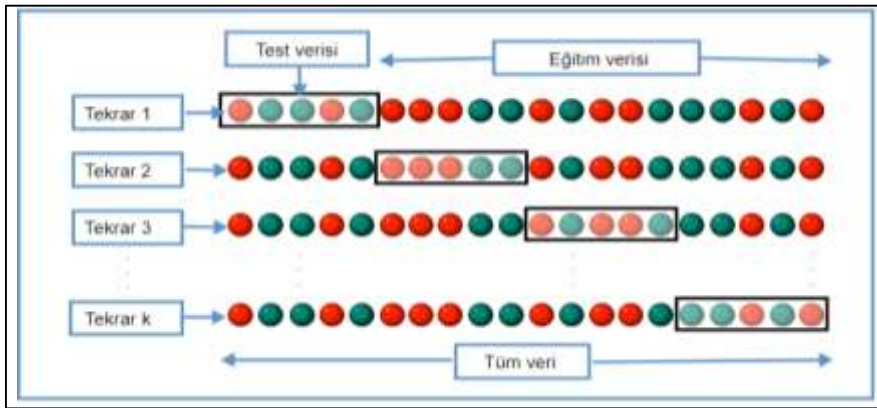
Çapraz doğrulama katsayısı her defasında bir gözlem noktasını dışarıda bırakarak geriye kalan veri noktalarıyla uygun bir model oluşturarak dışarıda bırakılan gözlem noktalarıyla ortalama tahmin hatasını minimum yapmaya çalışmaktadır. Çapraz doğrulama yöntemi, her modeli mevcut verilerin bir alt kümesine (eğitim seti) yerleştirip ardından modellerin tahmin kapasitelerini (kayıp) verilerin geri kalan kısmında (test seti) karşılaştırarak çalışmaktadır (Yates, Aandhal, Richards, Brook, 2022). Çapraz doğrulama katsayısı Eşitlik (2.10)'da gösterilmektedir.

$$CV(\lambda) = 1 / n \sum_{i=1}^n (y_i - \hat{y}^{-i})^2 \quad (2.10)$$

k -katlı çapraz doğrulama yönteminde ise veri seti k alt kümelere eşit olarak bölünür. Veri setindeki k adet bölünmüş olan gruplardan her seferde bir grup test grubu olarak ayrı tutulur ve diğer $k-1$ adet grup ise eğitim seti olarak kullanılır. Bu durum k adet grup için sırasıyla test grubu olarak değerlendirilecek şekilde tekrarlanmaktadır. Oluşturulmuş $k-1$ modelle ayrı tutulan test grubu için beklenen tahmin hatası bulunmaktadır. Sonuç olarak

bulunmuş beklenen tahmin hatalarından minimum olan model seçilir ve en doğru tahmin sonucunu veren değeri elde edilmektedir. k -katlı çapraz doğrulama katsayısı Eşitlik (2.11)'de gösterilmektedir. k -katlı çapraz doğrulama yöntemi gösterim şeması Resim 2.1'de verilmiştir.

$$CV(\lambda) = 1/n \sum_{k=1}^k \sum_{i \in F_k} (y_i - f_{\lambda}^{-k}(x_i))^2 \quad (2.11)$$



Resim 2.1. k -katlı çapraz doğrulama yöntemi gösterim şeması

2.5. Cezalandırılmış (Penalized) Lojistik Regresyon Yöntemleri

2.5.1. Ridge lojistik regresyon

L_2 düzenleme tekniğini kullanarak değişkenlerin tamamını modele dahil eden regresyon modeline Ridge regresyon denir.

Regresyon modelinde çoklu bağlantı sorunu durumunda regresyon katsayılarının varyans ve kovaryanslarının artması ile katsayıların standart hataları da artmaktadır. Bu durumda model iyiliğini ölçmek için yapılan testler yanlış sonuç vermektedir bu nedenle modele değişken eklemek veya çıkarmak gerekmektedir. Bu durumda en küçük kareler (EKK) yönteminde modele katkı sağlamayan değişkenlerin modele dâhil edilmesi, bağımsız değişkenlerin katsayılarının yanlış hesaplanması gibi sorunlarla karşılaşılmaktadır. Bu sorunların ortadan kaldırılması için 1970 yılında Hoerl ve Kennard tarafından Ridge regresyon yöntemi geliştirilmiştir. Hoerl ve Kennard, Ridge regresyon yöntemi sayesinde EKK tahmincisine oranla daha küçük hata kareler toplamının (HKT) ortalamasına sahip

regresyon katsayıları elde etmişlerdir. Çoklu bağlantı sorunu sonucunda ortaya çıkan bağımsız değişkenlere ilişkin korelasyon matrisinin birim matrsten uzaklaşması, parametre tahminlerine ait varyans değerlerinin yükselmesi ile EKK yöntemi tahmincisi ile elde edilen sonuçlar güvenilirliğini kaybetmektedir. Yanlı tahmin yöntemi olmasına karşın Ridge regresyonunu EKK yönteminden üstün kılan etkiler aşağıdaki gibidir;

1. Bağımsız değişkenlerdeki çoklu bağlantı sorununu gidermek,
2. Regresyonda yanlılık karesiyle varyansı değiştirerek, hata kareler ortalamasını azaltmaktır (İpek, 2000).

Ridge regresyon yönteminde, parametre tahminlerinin varyanslarını küçültmek ve daha güvenilir tahminler elde etmek için bağımsız değişkenlere ilişkin matrise bir ceza parametresi eklenmektedir. Bu sayede matrisin sütunlarındaki doğrusal bağımlılık azaltılmış olacaktır. Ridge regresyon yanlı bir tahmin edici olmasına rağmen katsayıların tahminleri EKK yönteminden daha küçük olarak hesaplanmaktadır (Lemein, 2010). Hoerl ve Kennard'a göre Ridge regresyon üç amaçla önerilmiştir (Hoerl ve Kennard, 1970; Montgomery ve Peck, 1991);

1. Yüksek dereceli çoklu bağlantı sorunu varlığında, katsayılarda meydana gelen kararsızlıkların grafik üzerinde gösterilmesinde,
2. Çoklu doğrusal regresyon modelinde bağımsız değişkenlerin birbirleriyle ilişkili oldukları durumlarda EKK yönteminin tahmininden daha küçük varyanslı tahminlerin elde edilmesinde,
3. Modeldeki gereksiz değişkenlerin çıkartılmasında kullanılabilir.

L_2 düzenlemesi olarak da bilinen Ridge regresyon yöntemi Eşitlik (2.12)'de verilmektedir.

$$\begin{aligned}
HKT &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\
\beta_{Ridge} &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \\
\beta_{Ridge} &= HKT + \lambda \sum_{j=1}^p \beta_j^2
\end{aligned} \tag{2.12}$$

Ridge regresyon yönteminin bir başka gösterimi ise Eşitlik (2.13)'de verilmektedir.

$$\beta_{Ridge} = \sum_{i=1}^n (y_i - y_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \tag{2.13}$$

Bağımlı değişkenin iki düzeyli kategorik veya 0 ve 1 değerlerini aldığı durumlarda lojistik regresyon yöntemi kullanılmaktadır. Ridge lojistik regresyon yöntemi ise 1989 yılında Duffy ve Santer tarafından önerilmiş bir cezalandırılmış regresyon yöntemi olarak ortaya çıkmıştır. Eşitlik (2.14)'de lojistik regresyonun negatif log-olabilirlik fonksiyonu gösterilmektedir.

$$\underset{\beta}{argmin} [\{-\sum_{i=1}^n y_i (\ln(\pi)) + (1 - y_i)(\ln(1 - \pi_i))\}] \tag{2.14}$$

Lojistik regresyonun negatif log-olabilirlik fonksiyonuna Ridge regresyonun ceza teriminin eklenmesiyle Ridge lojistik regresyon yöntemi elde edilir ve Eşitlik (2.15)'de gösterilmektedir.

$$\beta_{log-Ridge} = \underset{\beta}{argmin} [\{-\sum_{i=1}^n y_i (\ln(\pi)) + (1 - y_i)(\ln(1 - \pi_i))\} + \lambda \sum_{j=1}^p \beta_j^2] \tag{2.15}$$

2.5.2. Lasso lojistik regresyon

L_1 düzenleme tekniğini kullanarak değişkenlerden bazılarını modele dahil eden regresyon modeline Lasso regresyon denir.

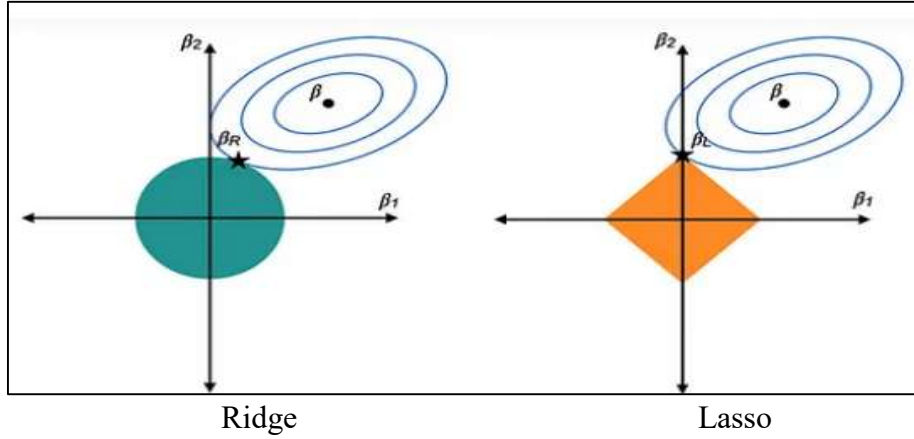
Lasso (Least Absolute Shrinkage and Selection Operator) regresyon yöntemi Tibshirani tarafından (1996) geliştirilmiştir (Tibshirani, 1996). Katsayıların daraltılması ve otomatik değişken seçiminde iyi bir performans sergilemektedir. Lasso regresyon ile benzer olarak Lasso regresyonda ortalama hata kareler yönteminin en küçük değerini elde etmeyi amaçlamaktadır (Flexeder, 2010). L_1 düzenlemesi olarak da bilinen Lasso regresyon yöntemi Eşitlik (2.16)'da verilmektedir.

$$\begin{aligned}
 HKT &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\
 \beta_{Lasso} &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \\
 \beta_{Lasso} &= HKT + \lambda \sum_{j=1}^p |\beta_j|
 \end{aligned} \tag{2.16}$$

Lasso regresyon yönteminin bir başka gösterimi ise Eşitlik (2.17)'de verilmektedir.

$$\beta_{Lasso} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| \tag{2.17}$$

Ridge regresyon yöntemi katsayıları sıfıra indirgeyemez böylece değişken seçimi yapmamış olur. Lasso regresyon yönteminde ise katsayılar sıfıra indirgenebilmektedir. Bu durum Lasso regresyon yönteminin değişken seçimi yapmasına sade ve anlaşılır modeller sunmasına neden olur. Lasso regresyon, bir tahmin alt kümesini içeren daha basit ve daha yorumlanabilir modeller üretmesi açısından Ridge regresyona göre büyük bir avantaja sahiptir. Ancak tüm değişkenlerin mutlaka kullanılması gereken durumda Ridge regresyon daha avantajlı olacaktır.



Resim 2.2. Ridge ve Lasso regresyon yöntemlerinin tahminleri

Ridge regresyon ve Lasso regresyon yöntemlerinin tahminleri Resim 2.2’de grafiksel olarak gösterilmiştir. Şekilde sarı ve yeşil ile gösterilen alanlar kısıt bölgelerini belirtmektedir. İlk grafikteki yeşil alan Ridge regresyon yönteminin tahmin edicisine, ikinci grafikteki sarı alan ise Lasso regresyon yönteminin tahmin edicisine ait olmaktadır. İki grafikte de bulunan ve mavi ile belirtilen daireler EKK tahmin edicisinin yayılımını simgelemektedir. Lasso regresyon yöntemine ait sarı alan konvektir ve konveks fonksiyonun kesildiği noktada fonksiyon minimum değerini almaktadır. Lasso regresyon yönteminde tam olarak sıfıra indirgenebilen yani değişken seçimi yapabilen regresyon katsayıları elde edilirken Ridge regresyon yönteminde ise EKK tahmin edicisi belirli bir noktaya kadar daralabilmesi sebebiyle katsayılar tam olarak sıfıra indirgenememektedir yani değişken seçimi yapılamamaktadır (Hastie, Tibrishani ve Friedman, 2008).

Lasso regresyon, model seçimi ve parametrelerin tahmininde yüksek boyutlu veriler için oldukça kullanışlıdır. Ancak Lasso regresyon yönteminin verinin yapısına bağlı olarak bazı problemleri çözmede dezavantajları vardır. Bunlar;

- Lasso regresyon yöntemi çoklu bağlantı sorunu varlığında, yüksek derecede ilişkili değişkenlerden birini rastgele seçer. Bu durumda, modelde olması gereken bir değişkenin çoklu bağlantı nedeniyle modelden çıkarılmış olması sorunuyla karşılaşmaktadır.
- Değişken sayısının gözlem sayısından büyük olduğu ($p > n$) durumlarda Lasso regresyon yöntemi, en fazla n tane değişkeni modele dâhil edebilmektedir. Bu durum, özellik seçimi için sınırlayıcı olmaktadır.

- Gözlem sayısının değişken sayısından büyük olduğu ($n > p$) durumlarda, bağımsız değişkenler arasında yüksek korelasyon varsa Lasso regresyon yöntemi Ridge regresyon yönteminden daha iyidir (Tibshirani, 1996).

Lojistik regresyonun negatif log-olabilirlik fonksiyonuna Lasso regresyonun ceza teriminin eklenmesiyle yani Lasso regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla Lasso lojistik regresyon yöntemi elde edilmiştir ve Eşitlik (2.18)'de gösterilmektedir.

$$\beta_{\log-Lasso} = \underset{\beta}{\operatorname{argmin}} \left[\left\{ -\sum_{i=1}^n y_i (\ln(\pi)) + (1 - y_i) (\ln(1 - \pi_i)) \right\} + \lambda \sum_{j=1}^p |\beta_j| \right] \quad (2.18)$$

2.5.3. Elastik-Ağ lojistik regresyon

Lasso regresyon yönteminin belirtilen dezavantajlarının azaltılması amacıyla Zou ve Hastie (2005) L_1 ve L_2 düzenleme tekniklerini birleştiren Elastik-Ağ regresyon yöntemini önermişlerdir. Elastik-Ağ regresyon yöntemi değişken seçimi ve grup etkisinin gözlemlenmesi için daha etkili bir yöntem elde edilmesi amacıyla geliştirilmiştir. Elastik-Ağ regresyon yöntemi Eşitlik (2.19)'da verilmektedir.

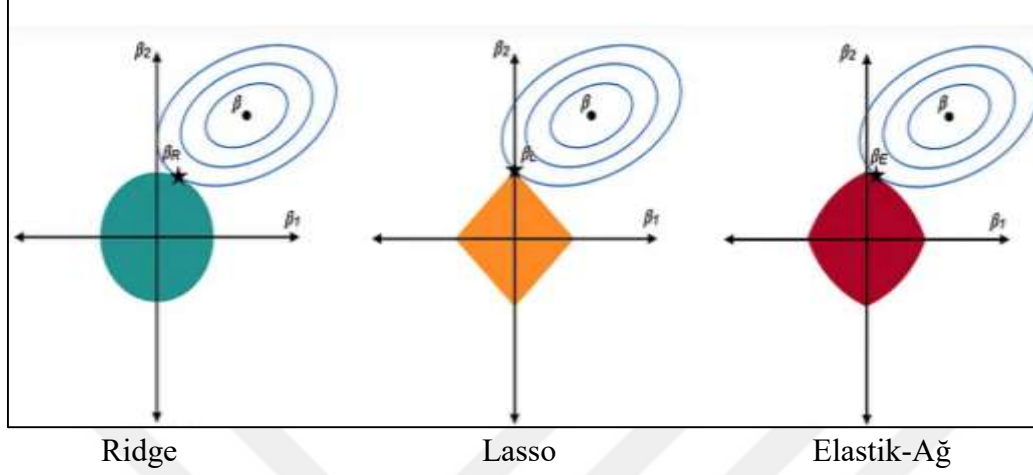
$$\begin{aligned} HKT &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\ \beta_{\text{Elastik-Ağ}} &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \\ \beta_{\text{Elastik-Ağ}} &= HKT + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \end{aligned} \quad (2.19)$$

Elastik-Ağ regresyon yönteminin bir başka gösterimi ise Eşitlik (2.20)'de verilmektedir.

$$\beta_{\text{Elastik-Ağ}} = \sum_{i=1}^n (y_i - y_i)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \quad (2.20)$$

λ_1 sıfır olduğunda Ridge regresyon yöntemini, λ_2 sıfır olduğunda ise Lasso regresyon yöntemini vermektedir. Elastik-Ağ regresyon yöntemi için değerleri 0 ile 1 arasında değer

almaktadır. Ridge, Lasso, Elastik-Ağ regresyon yöntemlerinin tahminleri Resim 2.3'te gösterilmiştir.



Resim 2.3. Ridge, Lasso, Elastik-Ağ regresyon yöntemlerinin tahminleri

Ridge regresyon, Lasso regresyon ve Elastik-Ağ regresyon yöntemlerinin tahminleri Resim 2.3'te grafiksel olarak gösterilmiştir. Şekilde sarı ve yeşil ile gösterilen alanlar kısıt bölgelerini belirtmektedir. İlk grafikteki yeşil alan Ridge regresyon yönteminin tahmin edicisine, ikinci grafikteki sarı alan ise Lasso regresyon yönteminin tahmin edicisine, üçüncü grafikteki bordo alan ise Elastik-Ağ regresyon yöntemine ait olmaktadır. Üç grafikte de bulunan ve mavi ile belirtilen daireler EKK tahmin edicisinin yayılımını simgelemektedir. Lasso regresyon yönteminde tam olarak sıfıra indirgenebilen regresyon katsayıları elde edilirken, Ridge regresyon yönteminde EKK tahmin edicisi belirli bir noktaya kadar daralabilmesi sebebiyle katsayılar tam olarak sıfıra indirgenememektedir (Hastie, Tibrişani ve Friedman, 2008). Elastik-Ağ regresyon modelinde EKK tahmin edicisinin daraldığı alan Ridge ve Lasso regresyon yöntemlerine göre daha fazladır. Ayrıca yapısı bakımında Ridge ve Lasso regresyon yöntemlerinin birleşimi olduğu da açıkça görülmektedir. Elastik-Ağ regresyon yöntemi Lasso regresyon yönteminin aksine değişken sayısının gözlem sayısından büyük olduğu ($p > n$) durumlarda yüksek boyutta olan tüm değişkenleri seçebilmektedir. Değişken seçiminde Lasso regresyon yönteminin, yüksek korelasyon sorunuyla başa çıkmak için ise Ridge regresyon yönteminin cezasını kullanmaktadır.

Lojistik regresyonun negatif log-olabilirlik fonksiyonuna Elastik-Ağ regresyonun ceza teriminin eklenmesiyle yani Elastik-Ağ regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla Elastik-Ağ lojistik regresyon yöntemi elde edilmiştir ve Eşitlik (2.21)'de gösterilmektedir.

$$\beta_{\log-\text{Elastik}-\text{Ağ}} = \underset{\beta}{\operatorname{argmin}} \left[\left\{ -\sum_{i=1}^n y_i (\ln(\pi)) + (1 - y_i) (\ln(1 - \pi_i)) \right\} + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right] \quad (2.21)$$

2.5.4. Uyarlanabilir (Adaptive) Lasso lojistik regresyon

Uyarlanabilir (Adaptive) Lasso regresyon yöntemi Elastik-Ağ regresyon yönteminde olduğu gibi Lasso regresyon yönteminin belirtilen dezavantajlarının azaltılması amacıyla Zou (2006) tarafından Lasso regresyon yöntemine bir ağırlık terimi eklenmesiyle geliştirilmiştir (Zhang, Zhao, Lu ve Xu, 2022). Bu yöntemde Oracle özelliklerinden de yararlanılmaktadır (Algama ve Lee, 2015). Oracle Özellikleri;

- Tahmin edici olasılığı bire yaklaşan bir modelde sıfır olmayan katsayıları doğru şekilde seçebilmelidir.
- Sıfır olmayan katsayılar asimptotik olarak normal dağılmalıdır.

Oracle özellikleri Uyarlanabilir Lasso regresyon yöntemini daha optimal hale getirmektedir. Uyarlanabilir Lasso regresyon yöntemi Eşitlik (2.22)'de gösterilmektedir.

$$\begin{aligned} HKT &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\ \beta_{\text{Uyar.Lasso}} &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \\ \beta_{\text{Uyar.Lasso}} &= HKT + \lambda \sum_{j=1}^p w_j |\beta_j| \end{aligned} \quad (2.22)$$

Uyarlanabilir Lasso regresyon yönteminin bir başka gösterimi ise Eşitlik (2.23)'de verilmektedir.

$$\beta_{Uyar.Lasso} = \sum_{i=1}^n (y_i - y_i)^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \quad (2.23)$$

Eşitlik (2.22) ve Eşitlik (2.23)'de görülen $\tilde{w}_j$ ağırlık vektörüdür. Ağırlık vektörü her katsayı için farklı bir düzenleme yapabilir yani bu vektör cezayı her katsayı için farklı şekilde ayarlayabilmektedir. Ağırlık vektörü denklemini Eşitlik (2.24)'de verilmektedir.

$$w_j = 1 / (|\beta_j^{ini}|)^r \quad (2.24)$$

$|\beta_j^{ini}|$ cezalandırılmış regresyon yöntemlerden elde edilen katsayıdır. r ise ağırlık vektörünün ayarlanması için pozitif bir sabittir. Bu sabit genellikle 1,2 veya 0,5 olarak alınabilmektedir. Uyarlanabilir Lasso regresyon yönteminin temel prensibi büyük katsayılara küçük ağırlık, küçük katsayılara büyük ağırlık vererek yanlılığı azaltıp Lasso regresyon yöntemine kıyasla daha tutarlı tahminler yapmaktır.

Lojistik regresyonun negatif log-olabilirlik fonksiyonuna Lasso regresyonun ceza teriminin eklenmesiyle yani Lasso regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla Lasso lojistik regresyon yöntemi elde edilmiştir. Eşitlik (2.18)'de gösterilmektedir. Lasso lojistik regresyon yöntemine ağırlığın eklenmesiyle Uyarlanabilir Lasso lojistik regresyon yöntemi elde edilmiştir. Uyarlanabilir Lasso lojistik regresyon yöntemi Eşitlik (2.25)'de gösterilmektedir.

$$\hat{\beta}_{log-Uyar.Lasso} = \underset{\beta}{argmin} \left[\left\{ -\sum_{i=1}^n y_i (\ln(\pi_i)) + (1 - y_i) (\ln(1 - \pi_i)) \right\} + \lambda \sum_{j=1}^p w_j |\beta_j| \right] \quad (2.25)$$

Uyarlanabilir Lasso lojistik regresyon yöntemindeki ağırlık vektöründe verilen $|\hat{\beta}_j^{ini}|$ katsayısı Lasso lojistik regresyon yönteminden elde edilmektedir (Araveeporn, 2021).

2.5.5. Uyarlanabilir (Adaptive) Elastik-Ağ lojistik regresyon

Uyarlanabilir (Adaptive) Elastik-Ağ regresyon yöntemi yüksek boyutlu verilerde değişken seçimi ve tahmini yapabilmek için Elastik-Ağ regresyon yönteminden geliştirilmiştir. Bu

yöntemi Zou ve Zhang (2009) önermişlerdir. Uyarlanabilir Elastik-Ağ regresyon yönteminde Elastik-Ağ regresyon yöntemi ile Uyarlanabilir Lasso regresyon yöntemi birleştirilmektedir ve Eşitlik (2.26)'da gösterilmektedir.

$$\begin{aligned}
 HKT &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 \\
 \beta_{Uyar.Elastik-Ağ} &= \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda_1 \sum_{j=1}^p w_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \\
 \beta_{Uyar.Elastik-Ağ} &= HKT + \lambda_1 \sum_{j=1}^p w_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2
 \end{aligned} \tag{2.26}$$

Uyarlanabilir Elastik-Ağ regresyon yönteminin bir başka gösterimi ise Eşitlik (2.27)'de verilmektedir.

$$\beta_{Uyar.Elastik-Ağ} = \sum_{i=1}^n (y_i - y_i)^2 + \lambda_1 \sum_{j=1}^p w_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \tag{2.27}$$

Uyarlanabilir Elastik-Ağ regresyon yönteminde düzenleme parametresinin optimal değerini seçme yöntemi olan Bayesian Bilgi Kriteri (BIC) kullanılarak ayar parametresi kontrol edilir. BIC 1978 yılında ortaya atılan bir ölçüttür. Farklı parametre sayılarına sahip parametrik gerçek modeller sınıfı arasında bir model seçimi türüdür. Eşitlik (2.26) ve Eşitlik (2.27)'de görülen $\tilde{w}_j$ ağırlık vektörüdür. Ağırlık vektör denklemi Eşitlik (2.28)'de verilmektedir.

$$w_j = (|\beta_j^{ini}| + 1/n)^{-Y} \quad Y > 0 \tag{2.28}$$

Uyarlanabilir Elastik-Ağ regresyon yönteminde λ_2 sıfır değerini aldığında Uyarlanabilir Lasso regresyon yöntemi, λ_1 sıfır değerini aldığında Ridge regresyon yöntemi elde edilmektedir. Uyarlanabilir Lasso regresyon yöntemi ve Elastik-Ağ regresyon yönteminin birleşmesiyle Uyarlanabilir Elastik-Ağ regresyon yöntemi elde edilmiştir ve Eşitlik (2.26) ve Eşitlik (2.27)'de gösterilmektedir. Uyarlanabilir Elastik-Ağ regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla Uyarlanabilir Elastik-Ağ lojistik regresyon yöntemi elde edilmektedir. Eşitlik (2.29)'da gösterilmektedir.

$$\beta_{\log-Uyar.Elastik-Ağ} = \underset{\beta}{argmin} [\{-\sum_{i=1}^n y_i (\ln(\pi)) + (1-y_i)(\ln(1-\pi_i))\} + \lambda_1 \sum_{j=1}^p w_j |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2] \quad (2.29)$$

Uyarlanabilir Elastik-Ağ lojistik regresyon yöntemindeki ağırlık vektöründe verilen $|\hat{\beta}_j^{ini}|$ katsayısı Elastik-Ağ lojistik regresyon yönteminden elde edilmektedir (Araveeporn, 2021).

2.5.6. Esnek (Relaxed) Uyarlanabilir (Adaptive) Lasso lojistik regresyon

Uyarlanabilir Lasso regresyon yönteminin Lasso regresyon yönteminden üstün olması Oracle özelliklerini kullanmasıdır. Böylece Lasso regresyon yöntemine göre daha hızlı ve daha tutarlı değişken seçimi yapılabilir. Uyarlanabilir Lasso regresyon yönteminin özel hali ise Meinshausen tarafından geliştirilen Esnek (Relaxed) Lasso regresyon yöntemidir (Meinshausen, 2006).

Esnek Lasso regresyon yönteminde amaç her düzenleme parametresi için değişkenleri herhangi bir cezalandırma olmadan seçebilmektir. Bunun için de bir gevşeme parametresi kullanmaktadır. Gevşeme parametresi φ ile gösterilmektedir. Gevşeme parametresi Esnek Lasso regresyon yönteminin performansını belirtir. Parametre uygun olarak seçildiğinde Lasso regresyon sağladığı seyrek çözümü aşırı büzülme yaşamamasına katkı sağlar. Modelin Esnek Lasso regresyon yöntemini kullanarak genişletmenin nedeni de budur (Zhang, Zhao, Lu ve Xu, 2022). Esnek Lasso regresyon yöntemi Eşitlik (2.30)'da gösterilmektedir.

$$\beta_{Esnek.Lasso} = \underset{\beta}{argmin} n^{-1} \sum_{i=1}^n (y_i - x_i^T \{\beta \cdot 1_{M_\lambda}\})^2 + \varphi \lambda \|\beta\|_1$$

$$\varphi \in (0,1] \quad (2.30)$$

$$\in [0, \infty]$$

$$\{\beta \cdot 1_M\}_k = \{0, \quad k \notin M, \beta_k, \quad k \in M\}, \quad k \in \{1, \dots, p\} \quad M \text{ gösterge işlevidir.}$$

Bir başka gösterimi ise Eşitlik (2.31)'de verilmektedir.

$$\beta_{Esnek.Lasso} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \varphi \sum_{j=1}^p |\beta_j| \quad (2.31)$$

Esnek Lasso regresyon modelde olan önemsiz değişkenleri silerek modelin doğruluğunu arttırmayı amaçlamaktadır. Uyarlanabilir Lasso regresyon yöntemi ve Esnek Lasso regresyon yöntemlerinin avantajlarını göz önünde bulundurarak bu iki yöntemi birleştiren Esnek (Relaxed) Uyarlanabilir (Adaptive) Lasso regresyon yöntemi geliştirilmiştir. Bu yöntem Esnek Lasso regresyon yöntemine Uyarlanabilir Lasso regresyon yönteminin ağırlık vektörünün eklenmesiyle elde edilir ve Eşitlik (2.32)'de gösterilmektedir.

$$\beta_{Esnek.Uyar.Lasso} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \varphi \sum_{j=1}^p w_j |\beta_j| \quad (2.32)$$

Esnek Uyarlanabilir Lasso regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla yani Esnek Uyarlanabilir Lasso regresyon yönteminin ceza teriminin lojistik regresyonun negatif log-olabilirlik fonksiyonuna eklenmesiyle Esnek Uyarlanabilir Lasso lojistik regresyon yöntemi elde edilmiştir. Bu yöntem literatüre ilk defa bu çalışma ile girmiştir ve Eşitlik (2.33)'de gösterilmektedir.

$$\beta_{log.Esnek.Uyar.Lasso} = \underset{\beta}{argmin} [\{-\sum_{i=1}^n y_i (\ln(\pi)) + (1 - y_i) (\ln(1 - \pi_i))\} + \lambda \varphi \sum_{j=1}^p w_j |\beta_j|] \quad (2.33)$$



3. SİMÜLASYON ÇALIŞMASI

Bu bölümde Ridge lojistik regresyon, Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($\Upsilon=1$), Uyarlanabilir Lasso lojistik regresyon ($\Upsilon=2$), Uyarlanabilir Elastik-Ağ lojistik regresyon ($\Upsilon=1$), Uyarlanabilir Elastik-Ağ lojistik regresyon ($\Upsilon=2$), Esnek Uyarlanabilir Lasso lojistik regresyon yöntemleri ele alınarak farklı örneklem çapları, farklı değişken sayıları ve düşük, orta ve yüksek olmak üzere farklı korelasyon yapıları için bir simülasyon çalışması dizayn edilmiştir.

3.1. Simülasyon Tasarımı

Simülasyon tasarımı aşamaları Zhang ve diğerleri (2022)'nin yaptığı çalışma referans alınarak aşağıdaki gibi düzenlenmiştir;

1. Gerçek regresyon katsayısı $\beta \in \mathbb{R}^p$, tüm $j \leq q$ için β_j sıfırdan farklı ve ilk $q=10$ ($q \leq p$) için 0,5-10 arasında sabit aralıkta rastgele değişkenler olarak, geri kalan $p-q$ kadar katsayı ise sıfır olarak belirlenmiştir.

2. Tasarım matrisi $x \in \mathbb{R}^p$, normal dağılımdan $N(0, \sigma^2)$ üretilmiştir. Burada kovaryans matrisi $i=j$ ise $\Sigma = \text{cov}(x) = (c_{ij})_{p \times p}$ $c_{ij} = 1$, $i=j=1, \dots, p$, $i \neq j$ ise $c_{ij} = \rho^{|i-j|}$ dir. Bağımsız değişkenler arasındaki korelasyon $\rho = 0,20$ (düşük), $\rho = 0,50$ (orta), $\rho = 0,90$ (yüksek) olarak ayarlanmıştır. Bağımlı değişkenler $\hat{\pi}(x_i) \geq 0,5$ için $y_i = 1$, $\hat{\pi}(x_i) \leq 0,5$ için $y_i = 0$ olarak tahmin edilmiş ve bağımlı değişken eşit sayıda 0 ile 1'lerden oluşturulmuştur.

3. Model için en uygun düzenleme parametresinin seçimi çok önemlidir. Bunun seçimi için literatürde birçok farklı yöntem vardır. Bu yöntemler içinde en çok kullanılan yöntem çapraz doğrulama yöntemidir. Bu çalışmada k-katlı çapraz doğrulama yöntemi kullanılmıştır. Yöntemde $k=10$ olarak alınmıştır. Yöntemin aşamaları aşağıdaki gibidir;

- n sayıda gözlem içeren veri seti 10 sayıda yaklaşık olarak eşit parçalara ayrılır.
- 10 parçaya bölünen veri seti 9 parça eğitim verisi, geri kalan 1 parça test verisi olarak seçilir. Bu işlem 10 adım tekrar eder ve her adımda farklı bir test verisi seçilir.

- Her adımda ortalama kareler hatası elde edilir. Hesaplanan tüm tahminlerin ortalaması alınarak çapraz doğrulama hatası eğrisi elde edilir.

Model için en uygun düzenleme parametresi yani optimal ceza parametresi, çapraz doğrulama hata eğrisinin en küçük değerine ulaştıran değerdir. 10-katlı çapraz doğrulama yöntemi gösterim şeması Resim 3.1’de gösterilmektedir.

	1. Parça	2. Parça	3. Parça	4. Parça	5. Parça	6. Parça	7. Parça	8. Parça	9. Parça	10. Parça
1. Adım	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
2. Adım	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
3. Adım	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
4. Adım	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
5. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim
6. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim	Eğitim
7. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim	Eğitim
8. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim	Eğitim
9. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test	Eğitim
10. Adım	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Eğitim	Test

Resim 3.1. 10-katlı çapraz doğrulama yöntemi gösterim şeması

Simülasyon çalışmasında ele alınan senaryolar aşağıda verilmiştir.

Senaryo 1: $n > p$ olduğu düşük boyutlu durumlar,

- $n=20$ ($p=15$)
- $n=50$ ($p=15, 25$)
- $n=100$ ($p=15, 25, 50$)
- $n=500$ ($p=15, 25, 50, 100$)

Senaryo 2: $p > n$ olduğu büyük boyutlu durumlar;

- $p=20$ ($n=15$)
- $p=50$ ($n=15, 25$)
- $p=100$ ($n=15, 25, 50$)
- $p=500$ ($n=15, 25, 50, 100$)

Her iki senaryoda tüm yöntemler için korelasyon;

- $\rho=0,20$ (düşük),
- $\rho=0,50$ (orta),
- $\rho=0,90$ (yüksek) durumları ele alınmıştır.

Veri seti %80 eğitim, %20 test verisi olarak ayrılmıştır ve tüm lojistik regresyon yöntemlerin RStudio ortamında geçen toplam süresine, yöntemlere ait ortalama doğru sınıflandırma oranına (DSO), standart sapma değerlerine (STD) ve tahmin edilen regresyon katsayılarının sıfırdan farklı olanlarının sayısına (SS) bakılmıştır. Simülasyon çalışmasında 1000 tekrar yapılmıştır.

Yöntemler arasında ortalama DSO değerlerine ilişkin fark olup olmadığını belirlemek amacıyla bağımlı iki örneklem için t istatistiğine dair hipotezler aşağıdaki gibi kurulmuştur.

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

Yöntemlerin ortalama DSO değerleri arasında fark olması durumunda en yüksek DSO değerini veren yöntem, fark olmaması durumunda en yüksek aynı DSO'yu veren yöntemler arasında en düşük STD'yi veren yöntem /ler belirlenmiştir.

3.2. Simülasyon Sonuçları

Simülasyon sonuçlarına ilişkin her bir senaryoda tüm yöntemlere dair farklı korelasyon yapıları için sonuçlar ve yorumlamalar yapılmıştır.

3.2.1. Birinci senaryo sonuçları

Birinci senaryoda; $p=15$ ve $n=20, 50, 100, 500$ için elde edilen sonuçlar Çizelge 3.1-3.4; $p=25$ ve $n=50, 100, 500$ için elde edilen sonuçlar Çizelge 3.5-3.7; $p=50$ ve $n=100, 500$ için elde edilen sonuçlar Çizelge 3.8-3.9; $p=100$ ve $n=500$ için elde edilen sonuçlar Çizelge 3.10'da verilmiştir.

Çizelge 3.1. $n=20(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=20(p=15)$	$\rho=0,20$ 671,58 sn			$\rho=0,50$ 605,70 sn			$\rho=0,90$ 554,44 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	93,900	16,97286	15	96,975	12,08161	15	96,850	10,43726	15
Lasso	93,600	17,80047	1	95,950	14,31896	1	95,800	12,19667	4
Elastik-Ağ	93,725	17,32916	1	96,200	13,11419	1	96,575	10,73590	8
Uyarlanabilir Lasso ($Y=1$)	94,100	16,96991	1	95,700	14,02522	1	96,025	11,45450	5
Uyarlanabilir Elastik- Ağ ($Y=1$)	94,125	16,99701	1	96,025	13,46312	7	96,650	9,81702	8
Uyarlanabilir Lasso ($Y=2$)	93,850	17,02843	1	95,700	13,98054	4	95,575	12,39750	5
Uyarlanabilir Elastik- Ağ ($Y=2$)	93,950	17,02755	1	95,675	14,21694	6	96,250	11,17194	7
Esnek Uyarlanabilir Lasso	94,050	17,09941	1	95,675	14,21694	3	96,075	11,47175	5

Çizelge 3.1 sonuçlarına göre; $n=20(p=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini Oracle prensibini kullanan Uyarlanabilir Lasso lojistik regresyon ($Y=1$), orta ve yüksek korelasyon düzeyinde ise Ridge lojistik regresyon vermiştir. Orta düzey korelasyon düzeyinde, Ridge lojistik regresyondan sonra en yüksek ortalama DSO değerine sahip olan Elastik-Ağ lojistik regresyon, yüksek korelasyon düzeyinde ise Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) yönteminin olduğu gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin SS artmış, RStudio'da geçen süre ve STD'si azalmıştır. Ayrıca seçilen yöntemlerin STD'sinin diğer yöntemlere göre daha düşük olduğu gözlenmiştir.

Çizelge 3.2. $n=50(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=50(p=15)$	$\rho =0,20$ 14,15 sn			$\rho =0,50$ 32,44 sn			$\rho =0,90$ 135,58 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,520	4,51552	15	99,310	4,45462	15	99,160	3,75612	15
Lasso	99,500	4,60121	9	99,170	4,49302	9	98,980	4,26502	5
Elastik-Ağ	99,460	4,65634	8	99,310	4,22394	10	99,140	3,77818	8
Uyarlanabilir Lasso ($Y=1$)	99,510	5,05311	13	99,100	5,56935	6	98,470	5,63834	5
Uyarlanabilir Elastik-Ağ ($Y=1$)	99,470	4,94404	12	99,160	5,22701	9	98,610	5,17635	8
Uyarlanabilir Lasso ($Y=2$)	99,470	5,024375	13	99,060	5,86095	6	98,560	5,56393	5
Uyarlanabilir Elastik-Ağ ($Y=2$)	99,470	5,024375	14	99,150	5,31090	9	98,550	5,49797	8
Esnek Uyarlanabilir Lasso	99,480	4,91464	11	99,040	6,07579	6	98,570	5,26144	5

Çizelge 3.2 sonuçlarına göre; $n=50(p=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Düşük ve yüksek korelasyon düzeyinde, en yüksek ortalama DSO değerini Ridge lojistik regresyon vermiştir. Orta korelasyon düzeyinde ise Ridge lojistik regresyon ve Elastik-Ağ lojistik regresyon yöntemlerinin ortalama DSO değerleri eşit çıkmıştır. Ancak Ridge lojistik regresyondan sonra en yüksek ortalama DSO'nun düşük korelasyon düzeyinde, Oracle prensibini kullanan Uyarlanabilir Lasso lojistik regresyon ($Y=1$), orta ve yüksek korelasyon düzeyinde ise Elastik-Ağ lojistik regresyon yönteminde olduğu gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin SS ve ortalama DSO değerleri azalmış, RStudio'da geçen süre ise artmıştır.

Çizelge 3.3. $n=100(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=100(p=15)$	$\rho =0,20$ 50,51 sn			$\rho =0,50$ 302,29 sn			$\rho =0,90$ 1741,6 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,680	2,89931	15	98,730	4,72328	15	95,395	5,01439	15
Lasso	99,620	3,28883	14	98,540	5,36163	12	93,600	5,76795	7
Elastik-Ağ	99,640	3,15124	14	98,585	5,21055	12	94,850	5,07969	9
Uyarlanabilir Lasso ($\Upsilon=1$)	99,650	3,18396	15	98,425	5,70762	12	90,745	6,84228	7
Uyarlanabilir Elastik Ağ ($\Upsilon=1$)	99,640	3,26052	15	98,380	5,85746	12	91,100	6,79235	9
Uyarlanabilir Lasso ($\Upsilon=2$)	99,635	3,30192	15	98,410	5,72316	12	90,730	6,84572	7
Uyarlanabilir Elastik-Ağ ($\Upsilon=2$)	99,630	3,33530	15	98,350	5,88745	12	90,855	6,86229	9
Esnek Uyarlanabilir Lasso	99,630	3,35027	15	98,405	5,74578	12	90,825	6,90956	7

Çizelge 3.3 sonuçlarına göre; $n=100(p=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Tüm korelasyon düzeylerinde, en yüksek ortalama DSO değerini Ridge lojistik regresyon vermiştir ancak düşük korelasyon düzeyinde, Ridge lojistik regresyondan sonra gelen en yüksek ortalama DSO değerine Uyarlanabilir Lasso lojistik regresyon ($\Upsilon=1$), orta ve yüksek korelasyon düzeyinde ise Elastik-Ağ lojistik regresyon yönteminin verdiği gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin SS ve ortalama DSO değerleri azalmış, RStudio'da geçen süre ve STD artmıştır. Orta korelasyon düzeyinde, Ridge lojistik regresyon dışındaki diğer tüm yöntemlerde SS aynı çıkmıştır.

Çizelge 3.4. $n=500(p=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=500 (p=15)$	$\rho=0,20$ 1724,56 sn			$\rho=0,50$ 1726,96 sn			$\rho=0,90$ 1881,82 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	79,072	3,95733	15	87,030	3,51759	15	96,376	1,85366	15
Lasso	78,971	4,03440	14	87,703	3,48840	14	95,798	2,03976	12
Elastik-Ağ	78,948	4,03429	14	86,739	3,50395	15	96,239	1,89670	12
Uyarlanabilir Lasso ($Y=1$)	78,998	4,01668	15	86,600	3,48139	15	94,765	2,26117	12
Uyarlanabilir Elastik- Ağ ($Y=1$)	78,746	4,01925	15	86,491	3,48339	15	94,684	2,28897	12
Uyarlanabilir Lasso ($Y=2$)	78,886	4,03337	15	86,496	3,48599	15	94,806	2,28858	12
Uyarlanabilir Elastik- Ağ ($Y=2$)	78,787	4,04845	15	86,221	3,50206	15	94,552	2,31535	12
Esnek Uyarlanabilir Lasso	78,912	4,01526	15	86,591	3,49169	15	94,765	2,28626	12

Çizelge 3.4 sonuçlarına göre; $n=500(p=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Tüm korelasyon düzeylerinde, en yüksek ortalama DSO değerini Ridge lojistik regresyon vermiştir ancak yine düşük korelasyon düzeyinde, Ridge lojistik regresyondan sonra gelen en yüksek ortalama DSO değerine Uyarlanabilir Lasso lojistik regresyon ($Y=1$), orta ve yüksek korelasyon düzeyinde ise Elastik-Ağ lojistik regresyon yönteminin verdiği gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin ortalama DSO değerleri ve RStudio'da geçen süre artmış, STD ve SS azalmıştır. Düşük korelasyon düzeyinde, Lasso lojistik regresyon ile Elastik-Ağ lojistik regresyonun SS'leri ve diğer tüm lojistik regresyon yöntemlerin SS'i ise kendi aralarında eşit çıkmıştır. Orta korelasyon düzeyinde ise Lasso lojistik regresyon ve yüksek korelasyon düzeyinde Ridge lojistik regresyon dışındaki yöntemlerin SS'i eşit çıkmıştır.

Çizelge 3.5. $n=50(p=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=50(p=25)$	$\rho =0,20$ 18,88 sn			$\rho =0,50$ 27,87 sn			$\rho =0,90$ 36,89 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,950	1,04814	25	99,920	1,26301	25	99,920	1,48182	25
Lasso	99,950*	1,30353	4	99,920	1,26335	1	99,930*	1,44817	1
Elastik-Ağ	99,940	1,34096	1	99,930*	1,22335	1	99,930*	1,44817	5
Uyarlanabilir Lasso ($Y=1$)	99,950*	1,30353	6	99,930*	1,22335	1	99,930*	1,44817	5
Uyarlanabilir Elastik-Ağ ($Y=1$)	99,950*	1,30353	5	99,930*	1,22335	2	99,930*	1,44817	7
Uyarlanabilir Lasso ($Y=2$)	99,920	1,78795	4	99,930*	1,22335	15	99,920	1,61127	5
Uyarlanabilir Elastik-Ağ ($Y=2$)	99,920	1,78795	6	99,930*	1,22335	15	99,930*	1,44817	7
Esnek Uyarlanabilir Lasso	99,950*	1,30353	6	99,930*	1,22335	16	99,930*	1,44817	7

Çizelge 3.5-Çizelge 3.6’da her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre bazı yöntemler arasında %95 güven düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiştir.

Çizelge 3.5 sonuçlarına göre; $n=50(p=25)$ için düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini Ridge lojistik regresyon, Lasso lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($Y=1$), Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), Esnek Uyarlanabilir Lasso lojistik regresyon vermiştir. Orta korelasyon düzeyinde, Ridge lojistik regresyon ve Lasso lojistik regresyon dışındaki tüm lojistik regresyon yöntemlerin, yüksek korelasyon düzeyinde ise Ridge lojistik regresyon ve Uyarlanabilir Lasso lojistik regresyon ($Y=2$) dışındaki tüm lojistik regresyon yöntemlerin verdiği gözlenmiştir. Düşük korelasyon düzeyinde, en düşük STD’yi Ridge lojistik regresyon dışındaki tüm yöntemler vermiştir.

Çizelge 3.6. $n=100(p=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=100(p=25)$	$\rho=0,20$ 5,38 sn			$\rho=0,50$ 12,11 sn			$\rho=0,90$ 53,81 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,915	1,60328	25	99,845	1,91170	25	99,835	2,01291	25
Lasso	99,920	1,41265	4	99,840	2,01952	1	99,840	1,95658	8
Elastik-Ağ	99,915	1,52318	4	99,830	2,13908	4	99,840	1,94375	8
Uyarlanabilir Lasso ($Y=1$)	99,925*	1,49052	11	99,860	1,85306	3	99,860*	1,86651	8
Uyarlanabilir Elastik- Ağ ($Y=1$)	99,925*	1,49052	11	99,865*	1,77762	6	99,860*	1,86651	8
Uyarlanabilir Lasso ($Y=2$)	99,925*	1,49052	10	99,865*	1,77762	3	99,860*	1,86651	7
Uyarlanabilir Elastik- Ağ ($Y=2$)	99,925*	1,49052	11	99,865*	1,77762	6	99,860*	1,86651	8
Esnek Uyarlanabilir Lasso	99,925*	1,49052	11	99,865*	1,77762	6	99,860*	1,86651	8

Çizelge 3.6 sonuçlarına göre; $n=100(p=25)$ için düşük ve yüksek korelasyon düzeyinde en yüksek ortalama DSO değeri ve en düşük STD'yi Ridge lojistik regresyon, Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon dışındaki tüm yöntemler, orta korelasyon düzeyinde ise Ridge lojistik regresyon, Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($Y=1$) dışındaki tüm lojistik regresyon yöntemlerin verdiği gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin RStudio'da geçen süre artmış ve ortalama DSO azalmıştır. Uyarlanabilir Lasso lojistik regresyon ($Y=2$) korelasyon durumlarının tamamında SS değerinin diğer yöntemlerin SS değerinden daha az değerde olduğu gözlenmiştir. Korelasyon yapısına ilişkin seçilen tüm lojistik regresyon yöntemlerin STD'si aynı çıkmıştır.

Çizelge 3.7. $n=500(p=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=500(p=25)$	$\rho =0,20$ 4,83 sn			$\rho =0,50$ 10,01 sn			$\rho =0,90$ 1941,46 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,943	1,08023	25	99,867	1,60183	25	90,742	10,61679	25
Lasso	99,944	1,03727	22	99,854	1,77082	19	90,377	11,02236	11
Elastik-Ağ	99,945	1,01932	22	99,851	1,80067	19	90,417	10,96756	11
Uyarlanabilir Lasso ($Y=1$)	99,956*	0,99953	24	99,878	1,58228	15	90,570	10,82258	11
Uyarlanabilir ElastikAğ ($Y =1$)	99,956*	0,99953	23	99,981	1,54117	16	90,646	10,73424	11
Uyarlanabilir Lasso ($Y=2$)	99,956*	0,99953	24	99,876	1,61034	15	90,447	10,95122	11
Uyarlanabilir ElastikAğ ($Y =2$)	99,956*	0,99953	24	99,976	1,60848	16	90,489	10,91407	11
Esnek Uyarlanabilir Lasso	99,956*	0,99953	24	99,878	1,58228	15	90,566	10,81594	11

Çizelge 3.7’de düşük korelasyon düzeyinde ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre bazı yöntemler arasında %95 güven düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiştir.

Çizelge 3.7 sonuçlarına göre; $n=500(p=25)$ için düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini ve en düşük STD’yi Ridge lojistik regresyon, Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon dışındaki tüm lojistik regresyon yöntemler vermiştir. En yüksek ortalama DSO değerini, orta korelasyon düzeyinde Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), yüksek korelasyon düzeyinde ise Ridge lojistik regresyon yöntemi vermiştir. Ancak Ridge lojistik regresyondan sonra gelen en yüksek ortalama DSO’ya sahip yöntemin Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) yönteminin olduğu gözlenmiştir. Düşük korelasyon düzeyinde, Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$)’nin SS’si diğer yöntemlerin SS’inden düşük olduğu ancak STD

hepsinde eşit olduğu gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin STD ve RStudio’da geçen süre artmış, ortalama DSO değeri ve SS azalmıştır.

Çizelge 3.8. $n=100(p=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=100(p=50)$	$\rho=0,20$ 4,5 sn			$\rho=0,50$ 17,61 sn			$\rho=0,90$ 27,4 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,980	0,49984	50	99,985	0,35341	50	99,930	1,11639	50
Lasso	99,980	0,49984	1	99,985	0,35341	1	99,930	1,11639	3
Elastik-Ağ	99,980	0,49984	1	99,985	0,34341	2	99,930	1,11639	7
Uyarlanabilir Lasso ($Y=1$)	99,985*	0,47434	2	99,990*	0,31622	7	99,940*	1,07123	9
Uyarlanabilir Elastik-Ağ ($Y=1$)	99,985*	0,47434	2	99,990*	0,31622	7	99,940*	1,07123	7
Uyarlanabilir Lasso ($Y=2$)	99,985*	0,47434	1	99,990*	0,31622	3	99,940*	1,07123	5
Uyarlanabilir Elastik-Ağ ($Y=2$)	99,985*	0,47434	2	99,990*	0,31622	7	99,940*	1,07123	9
Esnek Uyarlanabilir Lasso	99,985*	0,47434	5	99,990*	0,31622	9	99,940*	1,07123	10

Çizelge 3.8-Çizelge 3.9’da her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre bazı yöntemler arasında %95 güven düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiştir.

Çizelge 3.8 sonuçlarına göre; $n=100(p=50)$ için tüm korelasyon düzeylerinde, en yüksek ortalama DSO değerini ve en düşük STD’yi, Ridge lojistik regresyon, Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon dışındaki tüm yöntemler vermiştir. Tüm korelasyon yapılarında, seçilen yöntemlerin SS’si en yüksek olarak Esnek Uyarlanabilir

Lasso lojistik regresyon yönteminde gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin RStudio’da geçen süre artmıştır.

Çizelge 3.9. $n=500(p=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=500(p=50)$	$\rho=0,20$ 18,98 sn			$\rho=0,50$ 40,27 sn			$\rho=0,90$ 17,83 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,975	0,56982	50	99,957	0,71809	50	99,938	0,79924	50
Lasso	99,975	0,56982	8	99,956	0,72566	11	99,936	0,82618	11
Elastik-Ağ	99,975	0,56982	10	99,956	0,72566	14	99,937	0,80974	12
Uyarlanabilir Lasso ($Y=1$)	99,990*	0,31622	17	99,971*	0,56610	12	99,946*	0,75872	9
Uyarlanabilir Elastik-Ağ ($Y=1$)	99,990*	0,31622	14	99,971*	0,56610	13	99,946*	0,75872	8
Uyarlanabilir Lasso ($Y=2$)	99,990*	0,31622	13	99,971*	0,56610	13	99,946*	0,75872	6
Uyarlanabilir Elastik-Ağ ($Y=2$)	99,990*	0,31622	18	99,971*	0,56610	11	99,946*	0,75872	9
Esnek Uyarlanabilir Lasso	99,990*	0,31622	19	99,971*	0,56610	14	99,946*	0,75872	10

Çizelge 3.9 sonuçlarına göre; $n=500(p=50)$ için tüm korelasyon düzeylerinde, en yüksek ortalama DSO değerini ve en düşük STD’yi Ridge lojistik regresyon, Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon dışındaki tüm yöntemler vermiştir. Tüm korelasyon yapılarında, seçilen yöntemlerin SS’i yine en yüksek olarak Esnek Uyarlanabilir Lasso lojistik regresyon yönteminde gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin ortalama DSO değeri ve SS azalmış, STD ise artmıştır.

Çizelge 3.10. $n=500(p=100)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$n=500(p=100)$	$\rho =0,20$ 48,19 sn			$\rho =0,50$ 49,82 sn			$\rho =0,90$ 48,23 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,994	0,13409	100	99,989	0,24680	100	99,987	0,29819	100
Lasso	99,994	0,13409	13	99,989	0,24680	7	99,986	0,31607	19
Elastik-Ağ	99,994	0,13409	13	99,989	0,24680	7	99,985	0,33598	19
Uyarlanabilir Lasso ($\gamma=1$)	99,997*	0,09486	2	99,995*	0,11067	11	99,992*	0,16799	18
Uyarlanabilir Elastik-Ağ ($\gamma=1$)	99,997*	0,09486	3	99,995*	0,11067	14	99,992*	0,16799	20
Uyarlanabilir Lasso ($\gamma=2$)	99,997*	0,09486	2	99,995*	0,11067	11	99,992*	0,16799	18
Uyarlanabilir Elastik-Ağ ($\gamma=2$)	99,997*	0,09486	3	99,995*	0,11067	14	99,992*	0,16799	20
Esnek Uyarlanabilir Lasso	99,997*	0,09486	3	99,995*	0,11067	14	99,992*	0,16799	20

Çizelge 3.10’da her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre bazı yöntemler arasında %95 güven düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiştir.

Çizelge 3.10 sonuçlarına göre; $n=500(p=100)$ için tüm korelasyon düzeylerinde, en yüksek ortalama DSO değeri ve en düşük STD’yi, Ridge lojistik regresyon, Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon dışındaki tüm yöntemler vermiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin SS ve STD artmış, ortalama DSO değeri ise azalmıştır.

3.2.2. İkinci senaryo sonuçları

İkinci senaryoda; $n=15$ ve $p=20, 50, 100, 500$ için elde edilen sonuçlar Çizelge 3.11-3.14; $n=25$ ve $p=50, 100, 500$ için elde edilen sonuçlar Çizelge 3.15-3.17; $n=50$ ve $p=100, 500$ için elde edilen sonuçlar Çizelge 3.18-3.19; $n=100$ ve $p=500$ için elde edilen sonuçlar Çizelge 3.20’de verilmiştir.

Çizelge 3.11. $p=20(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=20(n=15)$	$\rho =0,20$ 45,48 sn			$\rho =0,50$ 52,24 sn			$\rho =0,90$ 228,42 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	98,333	9,63694	20	97,800	11,63165	20	96,100	13,05818	20
Lasso	98,500	8,88482	9	98,333	9,52083	6	96,267	11,60868	8
Elastik-Ağ	98,333	9,16367	9	98,300	9,80285	5	96,493	11,33586	6
Uyarlanabilir Lasso ($Y=1$)	98,566	8,89583	5	98,333	9,28425	5	96,100	12,53672	3
Uyarlanabilir Elastik- Ağ ($Y=1$)	98,500	9,25275	9	98,267	9,85356	9	96,533	11,70505	6
Uyarlanabilir Lasso ($Y=2$)	98,467	9,18694	5	98,200	9,72791	8	96,000	12,63785	3
Uyarlanabilir Elastik- Ağ ($Y=2$)	98,567	8,82761	10	98,333	9,21083	6	96,367	11,98370	6
Esnek Uyarlanabilir Lasso	98,566	8,89583	7	98,300	9,33794	8	96,200	12,25376	3

Çizelge 3.11 sonuçlarına göre; $p=20(n=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$), yüksek korelasyon düzeyinde ise Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) yöntemi vermiştir. Orta korelasyon düzeyinde ise en yüksek ortalama DSO değerini Lasso lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($Y=1$),

Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$) yöntemleri vermiştir ancak yöntemler arasındaki en düşük STD Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$) yönteminde gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin STD, RStudio’da geçen süre artmış, ortalama DSO değeri ise azalmıştır.

Çizelge 3.12. $p=50(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=50(n=15)$	$\rho=0,20$ 348,69 sn			$\rho=0,50$ 226,45 sn			$\rho=0,90$ 110,01 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	84,300	27,80597	50	89,967	25,12208	50	98,400	8,41898	50
Lasso	83,966	28,33061	1	89,267	24,36614	1	97,767	10,36058	10
Elastik-Ağ	84,500	27,39526	3	89,333	24,36149	1	97,333	10,01367	16
Uyarlanabilir Lasso ($Y=1$)	84,167	28,00954	1	89,100	24,77995	13	97,767	10,14361	9
Uyarlanabilir Elastik- Ağ ($Y=1$)	84,766	27,34195	3	89,600	23,87861	15	97,767	10,39215	16
Uyarlanabilir Lasso ($Y=2$)	84,133	28,24776	1	89,300	24,27824	3	97,667	9,49336	7
Uyarlanabilir Elastik- Ağ ($Y=2$)	84,267	27,84708	5	89,567	23,84073	4	97,900	9,60894	15
Esnek Uyarlanabilir Lasso	84,067	27,85307	1	89,533	23,84943	1	97,900	9,93668	12

Çizelge 3.12 sonuçlarına göre; $p=50(n=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir Düşük korelasyon düzeyinde, en yüksek ortalama DSO değeri Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), orta ve yüksek korelasyon düzeyinde ise Ridge lojistik regresyon yönteminde gözlenmiştir. Ancak Ridge lojistik regresyondan sonra en yüksek ortalama DSO değerini orta korelasyon düzeyinde, Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), yüksek korelasyon düzeyinde ise Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$) yöntemi vermiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin STD, RStudio’da geçen süre azalmış, ortalama DSO değeri ise artmıştır.

Çizelge 3.13. $p=100(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=100(n=15)$	$\rho =0,20$ 221,43 sn			$\rho =0,50$ 194,97 sn			$\rho =0,90$ 242,50 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	97,763	11,71417	100	97,867	11,83357	100	98,700	8,53389	100
Lasso	97,400	12,99864	2	97,600	13,12212	11	98,167	10,36058	8
Elastik-Ağ	97,400	13,08392	8	97,867	11,92719	3	98,333	9,63694	12
Uyarlanabilir Lasso ($\Upsilon=1$)	97,500	12,53956	2	97,633	13,00047	4	98,300	9,455630	6
Uyarlanabilir Elastik- Ağ ($\Upsilon=1$)	97,533	12,1248	3	97,533	13,10974	4	98,233	10,01561	12
Uyarlanabilir Lasso ($\Upsilon=2$)	97,633	11,92975	3	97,533	13,44481	6	98,333	9,28425	4
Uyarlanabilir Elastik- Ağ ($\Upsilon=2$)	97,733	11,71417	8	97,600	13,12212	3	98,467	8,94153	14
Esnek Uyarlanabilir Lasso	97,533	12,32255	3	97,633	12,91464	2	98,400	9,17558	5

Çizelge 3.13 sonuçlarına göre; $p=100(n=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Düşük ve yüksek korelasyon düzeyinde, en yüksek ortalama DSO değeri Ridge lojistik regresyon yönteminde gözlenmiştir, orta korelasyon düzeyinde ise Ridge lojistik regresyon ve Elastik-Ağ lojistik regresyon yöntemlerinin DSO değerleri eşit çıkmıştır. Ancak Ridge lojistik regresyondan sonra en yüksek ortalama DSO değeri düşük ve yüksek korelasyon düzeyinde, Uyarlanabilir Elastik-Ağ lojistik regresyon ($\Upsilon=2$), orta korelasyon düzeyinde ise Elastik-Ağ lojistik regresyon yönteminde gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin ortalama DSO değeri ise artmıştır.

Çizelge 3.14. $p=500(n=15)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=500(n=15)$	$\rho =0,20$ 549,25 sn			$\rho =0,50$ 217,72 sn			$\rho =0,90$ 67,77 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	96,300	15,01414	500	98,300	9,91561	500	99,600	4,69939	500
Lasso	96,033	15,67234	2	98,233	10,55266	3	99,600	4,69933	10
Elastik-Ağ	96,100	15,83023	3	98,267	10,29263	5	99,667	4,20527	20
Uyarlanabilir Lasso ($Y=1$)	96,133	15,73272	2	98,067	10,89050	2	99,533	5,56095	5
Uyarlanabilir Elastik-Ağ ($Y=1$)	96,167	15,49161	2	98,067	11,19269	4	99,600	4,45644	21
Uyarlanabilir Lasso ($Y=2$)	96,367	14,88169	2	98,267	10,40516	2	99,433	6,56165	6
Uyarlanabilir Elastik-Ağ ($Y=2$)	96,167	15,27471	2	98,167	11,05969	5	98,467	6,64293	19
Esnek Uyarlanabilir Lasso	96,100	15,97013	2	98,033	11,03673	2	98,600	4,69939	6

Çizelge 3.14 sonuçlarına göre; $p=500(n=15)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini Uyarlanabilir Lasso lojistik regresyon ($Y=2$) yüksek korelasyon düzeyinde, Elastik-Ağ lojistik regresyon ve orta korelasyon düzeyinde ise Ridge lojistik regresyon vermiştir. Ridge lojistik regresyondan sonra en yüksek ortalama DSO değeri, Uyarlanabilir Lasso lojistik regresyon ($Y=2$) ve Elastik-Ağ lojistik regresyon yöntemlerinde eşit çıkmıştır. Ancak düşük STD'yi Elastik-Ağ lojistik regresyon vermiştir.

Çizelge 3.15. $p=50(n=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=50(n=25)$	$\rho =0,20$ 26,41 sn			$\rho =0,50$ 65,98 sn			$\rho =0,90$ 143,13 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,585	3,71987	50	99,120	5,14136	50	99,060	5,54501	50
Lasso	99,560	3,87574	2	98,900	6,16971	2	98,900	6,26334	2
Elastik-Ağ	99,580	3,77716	12	98,980	5,66471	2	98,960	6,04606	2
Uyarlanabilir Lasso ($\gamma=1$)	99,440	4,61602	1	98,660	7,74195	18	98,520	8,01709	2
Uyarlanabilir Elastik- Ağ ($\gamma=1$)	99,480	4,55332	9	98,900	6,35299	5	98,640	7,57679	2
Uyarlanabilir Lasso ($\gamma=2$)	99,340	5,65653	3	98,480	7,14218	1	98,360	8,70552	2
Uyarlanabilir Elastik- Ağ ($\gamma=2$)	99,440	4,78636	12	98,600	7,22151	2	98,360	8,88759	2
Esnek Uyarlanabilir Lasso	99,440	4,70196	2	98,640	7,57067	2	98,560	7,82225	2

Çizelge 3.15 sonuçlarına göre; $p=50(n=25)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Tüm korelasyon düzeylerinde, en yüksek ortalama DSO değerini Ridge lojistik regresyon vermiştir ancak Ridge lojistik regresyondan sonra en yüksek ortalama DSO'ya sahip olan yöntem Elastik-Ağ lojistik regresyon olarak gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin ortalama DSO değeri ve SS azalmış, RStudio da geçen süre ve STD ise artmıştır.

Çizelge 3.16. $p=100(n=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=100(n=25)$	$\rho=0,20$ 30,92 sn			$\rho=0,50$ 83,15 sn			$\rho=0,90$ 101,31 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,940	1,41364	100	99,820	3,02932	100	99,620	3,61506	100
Lasso	99,940	1,41364	2	99,800	2,67715	2	99,580	4,12803	10
Elastik-Ağ	99,960*	1,26491	2	99,800*	2,25775	2	99,580*	3,92925	11
Uyarlanabilir Lasso ($Y=1$)	99,960*	1,26491	2	99,700	3,94024	2	99,520	4,87784	9
Uyarlanabilir Elastik-Ağ ($Y=1$)	99,960	1,36591	12	99,720	3,99215	2	99,580	4,31766	3
Uyarlanabilir Lasso ($Y=2$)	99,920	1,78795	2	99,620	4,50287	1	99,500	4,91672	1
Uyarlanabilir Elastik-Ağ ($Y=2$)	99,940	1,41364	5	99,660	4,13539	1	99,520	4,95925	3
Esnek Uyarlanabilir Lasso	99,960	1,36591	2	99,720	3,89059	1	99,540	4,67014	1

Çizelge 3.16’da için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde düşük ve orta korelasyon düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden tüm korelasyon düzeylerinde aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiştir.

Çizelge 3.16 sonuçlarına göre; $p=100(n=25)$ düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($Y=1$), Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), Esnek Uyarlanabilir Lasso lojistik regresyon yöntemi vermiştir. Yöntemler arasında en düşük STD’yi Elastik-Ağ lojistik regresyon ve Uyarlanabilir Lasso lojistik regresyon ($Y=1$) yöntemi vermiştir. Orta ve yüksek korelasyon düzeyinde, en yüksek ortalama DSO değerini ise Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon vermiştir ve yöntemler arasında en düşük STD

Elastik-Ağ lojistik regresyon yönteminde gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin ortalama DSO değeri azalmış, STD ve RStudio’da geçen süre ise artmıştır.

Çizelge 3.17. $p=500(n=25)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=500(n=25)$	$\rho =0,20$ 17,69 sn			$\rho =0,50$ 43,48 sn			$\rho =0,90$ 41,14 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,790	3,52619	500	99,680	3,21987	500	99,440	4,34802	500
Lasso	99,780	3,15619	15	99,520	4,62504	15	99,360	4,94113	15
Elastik-Ağ	99,740	3,62567	20	99,600	3,66789	20	99,400	4,86453	2
Uyarlanabilir Lasso ($\gamma=1$)	99,760	3,45750	22	99,380	4,98401	22	99,260	5,64661	10
Uyarlanabilir Elastik-Ağ ($\gamma=1$)	99,740	3,62567	1	99,480	4,62938	1	99,280	5,39543	11
Uyarlanabilir Lasso ($\gamma=2$)	99,640	4,80559	1	99,240	5,74938	1	99,160	6,27166	1
Uyarlanabilir Elastik-Ağ ($\gamma=2$)	99,660	4,41632	1	99,260	5,57525	1	99,140	6,23693	1
Esnek Uyarlanabilir Lasso	99,720	3,89059	20	99,440	4,701969	20	99,300	5,36085	2

Çizelge 3.17 sonuçlarına göre; $p=500(n=25)$ için her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir. Tüm korelasyon düzeylerinde, en yüksek ortalama DSO değeri Ridge lojistik regresyon yönteminde gözlenmiştir. Düşük korelasyon düzeyinde, en yüksek ortalama DSO değerini Ridge lojistik regresyondan sonra Lasso lojistik regresyon, orta ve yüksek korelasyon düzeyinde ise yine Elastik-Ağ lojistik regresyon yöntemi vermiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin STD genel olarak artmış, ortalama DSO değeri ise azalmıştır.

Çizelge 3.18. $p=100(n=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=100(n=50)$	$\rho =0,20$ 194,35 sn			$\rho =0,50$ 67,6 sn			$\rho =0,90$ 55,58 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,730	2,05702	100	99,920	1,26301	100	99,950	0,83558	100
Lasso	99,730*	2,05702	2	99,920	1,26301	1	99,950	0,83558	1
Elastik-Ağ	99,730*	2,05702	2	99,920	1,26301	9	99,950	0,83558	1
Uyarlanabilir Lasso ($\gamma=1$)	99,590	3,39756	2	99,910	1,37612	1	99,910	1,51465	4
Uyarlanabilir Elastik- Ağ ($\gamma=1$)	99,590	3,39756	2	99,900	1,41139	1	99,920	1,33992	6
Uyarlanabilir Lasso ($\gamma=2$)	99,590	3,39756	2	99,880	1,66984	9	99,920	1,33992	3
Uyarlanabilir Elastik- Ağ ($\gamma=2$)	99,570	3,33560	2	99,850	2,20961	9	99,850	3,08009	4
Esnek Uyarlanabilir Lasso	99,560	3,55206	2	99,910	1,44706	8	99,910	1,51465	3

Çizelge 3.18’de her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde orta ve yüksek korelasyon düzeyinde yöntemler arasında fark olduğu tespit edilmiş ancak düşük korelasyon düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden düşük korelasyon düzeyinde en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiş, orta ve yüksek korelasyon düzeyinde ise yöntemler arasında en yüksek ortalama DSO değerini veren yöntemler belirlenmiştir.

Çizelge 3.18 sonuçlarına göre; $p=100(n=50)$ tüm korelasyon düzeylerinde, Ridge lojistik regresyon, Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon yöntemleri en yüksek ortalama DSO değerini veren yöntemler olarak gözlenmiş ayrıca düşük korelasyon düzeylerinde bu yöntemlerin STD’si en düşük olarak belirlenmiştir. Tüm korelasyon düzeylerinde, seçilen yöntemlerin SS aynı iken tek fark orta düzey korelasyon durumunda

gözlenmiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin ortalama DSO değeri artmış, STD ve RStudio’da geçen süre azalmıştır.

Çizelge 3.19. $p=500(n=50)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=500(n=50)$	$\rho =0,20$ 81,96 sn			$\rho =0,50$ 28,56 sn			$\rho =0,90$ 16,03 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,800	1,88774	500	99,990	1,26158	500	99,930	1,13859	500
Lasso	99,800*	1,88774	1	99,990	1,26158	1	99,940*	0,99869	1
Elastik-Ağ	99,800*	1,88774	1	99,990	1,26158	1	99,890	2,62579	5
Uyarlanabilir Lasso ($\gamma =1$)	99,800*	1,88774	1	99,850	1,75515	1	99,920	1,51465	1
Uyarlanabilir Elastik-Ağ ($\gamma =1$)	99,800*	1,88774	1	99,860	1,66828	1	99,940*	0,99869	4
Uyarlanabilir Lasso ($\gamma =2$)	99,780	1,93787	1	99,820	1,99288	1	99,930	1,13859	3
Uyarlanabilir Elastik-Ağ ($\gamma =2$)	99,720	3,36642	1	99,840	1,94375	1	99,930	1,13859	5
Esnek Uyarlanabilir Lasso	99,800*	1,88774	1	99,980	1,54531	1	99,930	1,13859	4

Çizelge 3.19’da her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde orta korelasyon düzeyinde yöntemler arasında fark olduğu tespit edilmiş ancak düşük ve yüksek korelasyon düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden düşük ve yüksek korelasyon düzeyinde aynı ortalama DSO değerini veren yöntemler arasında en düşük STD’yi veren yöntem/ler belirlenmiş, orta korelasyon düzeyinde ise yöntemler arasında en yüksek ortalama DSO değerini veren yöntemler belirlenmiştir.

Çizelge 3.19 sonuçlarına göre; $p=500(n=50)$ düşük korelasyon düzeyinde, Uyarlanabilir Lasso ($Y=2$) ve Esnek Uyarlanabilir Lasso lojistik regresyon dışındaki tüm yöntemler, orta korelasyon düzeyinde, Ridge lojistik regresyon ve Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon yöntemler, yüksek korelasyon düzeyinde ise Lasso lojistik regresyon ve Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) en yüksek ortalama DSO değerini veren yöntemler olarak gözlenmiştir. Ayrıca düşük ve yüksek korelasyon düzeyinde en düşük STD'yi de en yüksek ortalama DSO'yu veren yöntemler vermiştir.

Çizelge 3.20. $p=500(n=100)$; tüm lojistik regresyon yöntemlerin tüm korelasyon yapıları için RStudio ortamında geçen toplam süre, SS, STD, ortalama DSO değerleri

$p=500(n=100)$	$\rho=0,20$ 19,31 sn			$\rho=0,50$ 48,23 sn			$\rho=0,90$ 74,56 sn		
	DSO	STD	SS	DSO	STD	SS	DSO	STD	SS
Ridge	99,965	0,61167	500	99,960	0,49864	500	99,920	0,80264	500
Lasso	99,965*	0,61167	1	99,960*	0,49864	1	99,920*	0,80264	1
Elastik-Ağ	99,965*	0,61167	1	99,960*	0,49864	1	99,920*	0,80264	1
Uyarlanabilir Lasso ($Y=1$)	99,965*	0,61167	1	99,960*	0,49864	1	99,920*	0,80264	1
Uyarlanabilir Elastik-Ağ ($Y=1$)	99,895	2,71502	1	99,865	3,04401	13	99,720	4,53786	1
Uyarlanabilir Lasso ($Y=2$)	99,955	0,75732	1	99,935	0,75587	7	99,850	1,47637	1
Uyarlanabilir Elastik-Ağ ($Y=2$)	99,960	0,70632	1	99,950	0,58973	11	99,865	1,26822	1
Esnek Uyarlanabilir Lasso	99,965*	0,61167	1	99,960*	0,49864	10	99,920*	0,80264	1

Çizelge 3.20'de her bir yöntemde ve farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre bazı yöntemler arasında %95 güven düzeyinde fark olmadığı tespit edilmiştir ve tablolarda * ile gösterilmiştir. Bu yüzden en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD'yi veren yöntem/ler belirlenmiştir.

Çizelge 3.20 sonuçlarına göre; $p=500(n=100)$ tüm korelasyon düzeylerinde, en yüksek DSO değerini ve en düşük STD değerlerini, Ridge lojistik regresyon, Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($\gamma=1$) ve Esnek Uyarlanabilir Lasso lojistik regresyon yöntemleri vermiştir. Korelasyon seviyesi arttıkça yöntemlere ilişkin RStudio’da geçen süreler artmış, ortalama DSO değeri ise azalmıştır. Seçilen yöntemlerin SS’i düşük ve yüksek korelasyonda aynı, orta korelasyon düzeyinde ise SS’lerde artış olduğu gözlenmiştir.

3.2.3. Senaryo sonuçlarına dair özet tablolar

Çizelge 3.21’de lojistik regresyon yöntemlerin numaralandırılması, Çizelge 3.22’de birinci senaryo sonuçlarına ait özet tablosu, Çizelge 3.23’te ikinci senaryo sonuçlarına ait özet tablosu verilmiştir.

Çizelge 3.21. Lojistik regresyon yöntemlerinin numaralandırılması

Ridge lojistik regresyon	Metot1 (M1)
Lasso lojistik regresyon	Metot2 (M2)
Elastik-Ağ lojistik regresyon	Metot3 (M3)
Uyarlanabilir Lasso ($\gamma=1$) lojistik regresyon	Metot4 (M4)
Uyarlanabilir Elastik-Ağ ($\gamma=1$) lojistik regresyon	Metot5 (M5)
Uyarlanabilir Lasso ($\gamma=2$) lojistik regresyon	Metot6 (M6)
Uyarlanabilir Elastik-Ağ ($\gamma=2$) lojistik regresyon	Metot7 (M7)
Esnek Uyarlanabilir Lasso lojistik regresyon	Metot8 (M8)

Çizelge 3.22. Birinci senaryo sonuçlarına ait özet tablosu

	Korelasyon Yapıları			
		Düşük	Orta	Yüksek
$p=15$	$n=20$	M4	M1,M3	M1,M5
	$n=50$	M1,M4	M1,M3	M1,M3
	$n=100$	M1,M4	M1,M3	M1,M3
	$n=500$	M1,M4	M1,M3	M1,M3
$p=25$	$n=50$	M1,M2*,M4*,M5*,M8*	M3*,M4*,M5*,M6*,M7*,M8*	M2*,M3*,M4*,M5*, M7*,M8*
	$n=100$	M4*,M5*,M6*,M7*,M8*	M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*, M8*
	$n=500$	M4*,M5*,M6*,M7*,M8*	M5	M1,M5
$p=50$	$n=100$	M4*,M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*, M8*
	$n=500$	M4*,M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*, M8*
$p=100$	$n=500$	M4*,M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*,M8*	M4*,M5*,M6*,M7*, M8*

Çizelge 3.23. İkinci senaryo sonuçlarına ait özet tablosu

	Korelasyon Yapıları			
		Düşük	Orta	Yüksek
$n=15$	$p=20$	M7	M2,M4,M7	M5
	$p=50$	M5	M1,M5	M1,M7
	$p=100$	M1,M7	M1,M3	M1,M7
	$p=500$	M6	M1,M3,M6	M3
$n=25$	$p=50$	M1,M3	M1,M3	M1,M3
	$p=100$	M3*,M4*,M5,M8	M2,M3	M2,M3
	$p=500$	M1,M2	M1,M3	M1,M3
$n=50$	$p=100$	M1,M2*,M3*	M1,M2,M3	M1,M2,M3
	$p=500$	M1,M2*,M3*,M4*,M5*,M8*	M1,M2,M3	M2*,M5*
$n=100$	$p=500$	M1,M2*,M3*,M4*,M8*	M1,M2*,M3*,M4*,M8*	M1,M2*,M3*,M4*,M8*

Simülasyon çalışmasında, farklı korelasyon yapılarında ortalama DSO değerlerine ilişkin bağımlı iki örneklem için t istatistiğine göre bazı yöntemler arasında %95 güven düzeyinde fark olmadığı tespit edilmiştir bu yüzden en yüksek aynı ortalama DSO değerini veren yöntemler arasında en düşük STD'yi veren yöntem/ler belirlenmiştir. Belirlenen yöntemler Çizelge 3.22 ve Çizelge 3.23'te * ile gösterilmiştir.



4. UYGULAMA

Çalışmada iki gerçek veri seti ele alınmıştır. Birinci veri seti, gözlem sayısının değişken sayısından büyük olduğu düşük boyutlu, Kaggle veri merkezindeki Pirinç türü sınıflandırma (Rice type classification) veri seti ele alınmıştır. İkinci veri seti, değişken sayısının gözlem sayısından büyük olduğu durum için kullanılan GSE68719 koduyla GEO veri tabanında bulunan Parkinson veri setidir. Her iki veri setinde de simülasyon çalışmasında kullanılan cezalandırılmış lojistik regresyon yöntemleri ele alınarak tüm yöntemlerin ortalama doğru sınıflandırma oranı (DSO) ve tahmin edilen regresyon katsayılarının sıfırdan farklı olanlarının sayısına (SS) bakılmıştır. Ayrıca iki veri seti de %80 eğitim, %20 test verisi olarak ayrılmıştır.

4.1. Veri Setleri ve Sonuçları

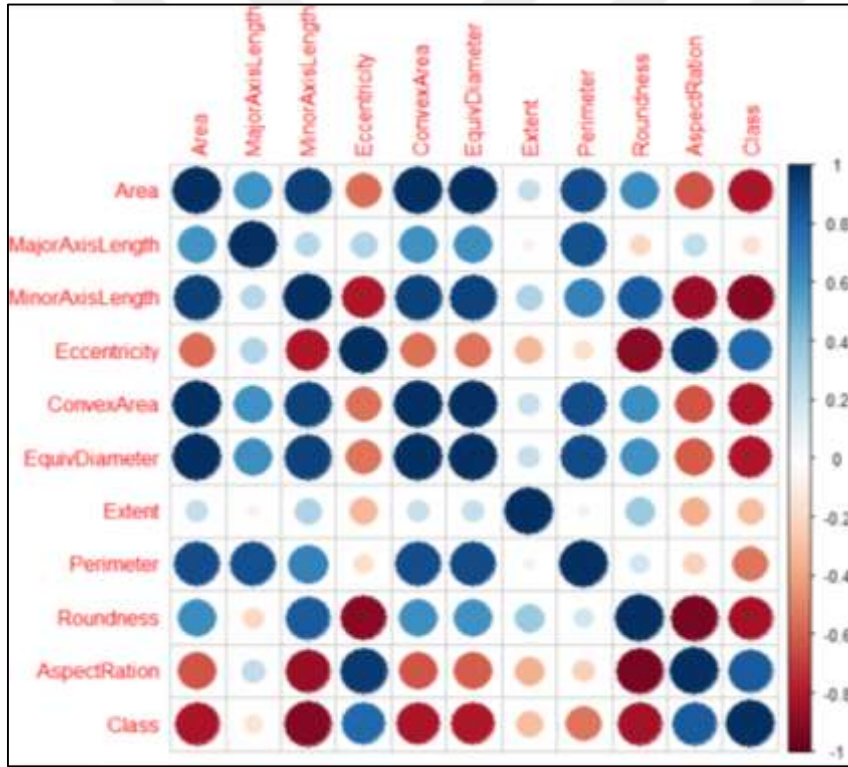
4.1.1. Pirinç türü sınıflandırma

Birinci veri seti, gözlem sayısının değişken sayısından büyük olduğu düşük boyutlu durumlar için kullanılan Kaggle veri merkezinden çekilen Pirinç türü sınıflandırma veri setidir. Pirinç türleri sınıflandırılması üzerine hazırlanmış, 2 sene önce güncelenmiş, lojistik regresyona uygun olan bir veri setidir. Veri setine <https://www.kaggle.com/datasets/mssmartypants/rice-type-classification> adresinden ulaşılabilir. Kaggle, veri merkezi 2010 yılında kurulmuştur. Kullanıcıların veri kümelerini bulmasını, yayınlamasına veya kullanmasına olanak tanıyan bir platformdur. Veri setinde gözlem sayısı 18185, değişken sayısı ise 12'dir ancak veri setinde ilk olarak veri manipülasyonu yapılmıştır. Bağımsız değişken sayısı 11 olarak ve bağımlı değişkenler olarak Yasemin pirinci 1, Gönen pirinci 0 olarak tanımlanmıştır. Bağımsız değişkenler ise;

- Alan (Area)
- Ana Eksen Uzunluğu (MajorAxisLength)
- Küçük Eksen Uzunluğu (MinorAxisLength)
- Eksantriklik (Eccentricity)
- Dış Bükey Alan (ConvexArea)
- Eşdeğer Çap (EquivDiameter)

- Kapsam (Extent)
- Çevre (Perimeter)
- Yuvarlaklık (Roundness)
- En Boy Oranı (AspectRatio)
- Sınıf (Class) olarak tanımlanmıştır.

Ayrıca değişkenler arasında çoklu bağlantı sorunu varlığına bakılmış *VIF* değeri 10,65 olarak bulunmuştur. Yani değişkenler arasında, yüksek derecede korelasyon olduğu gözlemlenmiştir. Resim 4.1.'de Pirinç türü sınıflandırma veri setinin değişkenlerine ait korelasyon tablosu verilmiştir.



Resim 4.1. Pirinç türü sınıflandırma veri setinin değişkenlerine ait korelasyon tablosu

Çizelge 4.1'de Pirinç türü sınıflandırma veri seti için tüm değişkenlere ait özet bilgiler ve Çizelge 4.2'de ise tüm lojistik regresyon yöntemlere ait DSO değerleri ve SS değerleri verilmiştir.

Çizelge 4.1. Pirinç türü sınıflandırma veri seti için özet bilgiler tablosu

Değişkenler	Minimum Değer	Ortalama	Maksimum Değer	STD
Alan	2522	6660	10210	1467,197
Ana Eksen Uzunluğu	74,13	153,88	183,21	12,3764
Küçük Eksen Uzunluğu	34,41	55,72	82,55	10,0615
Eksantriklik	0,6766	0,9233	0,9668	0,0305
Dış Bükey Alan	2579	6843	11008	1502,007
Eşdeğer Çap	56,67	92,09	114,02	9,9062
Kapsam	0,3832	0,6012	0,8866	0,1043
Çevre	197	353,1	508,5	29,500
Yuvarlaklık	0,1746	0,7019	0,9047	0,0673
En Boy Oranı	1,358	2,603	3,912	0,4348

Çizelge 4.2. Pirinç türü sınıflandırma veri seti için tüm lojistik regresyon yöntemlere ait DSO ve SS değerleri

Lojistik Regresyon Yöntemler	DSO (SS)
Ridge	98,474 (11)
Lasso	98,845 (9)
Elastik-Ağ	98,913 (10)
Uyarlanabilir Lasso ($\gamma=1$)	98,907 (10)
Uyarlanabilir Elastik-Ağ ($\gamma=1$)	98,900 (11)
Uyarlanabilir Lasso ($\gamma=2$)	98,908 (10)
Uyarlanabilir Elastik-Ağ ($\gamma=2$)	98,907 (11)
Esnek Uyarlanabilir Lasso	98,901 (9)

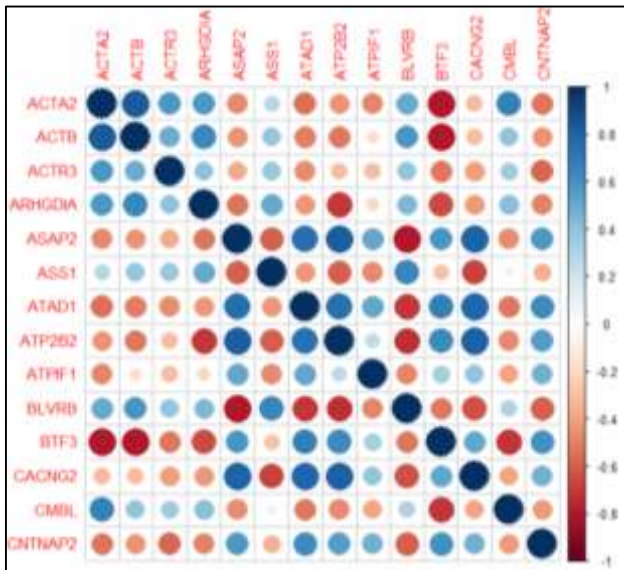
- Birinci satırdaki değerler DSO'yu, parantez içindekiler SS'i göstermektedir.
- Pirinç türü sınıflandırma veri setinde, Elastik-Ağ lojistik regresyon en yüksek ortalama DSO değerini veren yöntem olarak gözlenmiştir. Bu yöntemde sıfırdan farklı regresyon katsayısı 10 olarak belirlenmiştir.

4.1.2. Parkinson

İkinci veri seti, değişken sayısının gözlem sayısından büyük olduğu yüksek boyutlu durum için kullanılan GSE68719 koduyla GEO veri tabanında bulunan Parkinson veri setidir. GEO, veri gönderimlerini destekleyen kamuya açık işlevsel bir genomik veri deposudur. Veri setlerine NCBI tarafından hizmete sunulan GEO veri tabanı üzerinden erişim sağlanabilmektedir.

Veri setine <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68719> adresinden ulaşılabilir. Veri seti Amerika Birleşik Devletleri, Boston Üniversitesi, Nöroloji Bilim Dalı, Alexandra Dumitriu tarafından veri tabanına teslim edilmiştir. Veri setinin son güncellenme tarihi 15 Mayıs 2019'dur.

Veri seti, Parkinson hastalığına sahip insanların ölüm sonrası beyin dokusu bozukluğu için hazırlanmış bir setidir. Bu veri setinden sadece erkek bireylere ilişkin veri seti çekilmiş olup, 24 erkek bireye ilişkin 68 gen incelemeye alınmıştır. Bağımlı değişken, Parkinson hastalığına hastalığa sahip erkek bireyler için 1, sahip olmayanlar ise 0 olarak tanımlanmıştır. Resim 4.2'de GSE68719 veri seti için ilk 14 bağımsız değişkene ait korelasyon tablosu verilmiştir. Çizelge 4.3'te Parkinson veri seti için ilk 14 değişkene ait özet bilgiler Çizelge 4.4'de ise tüm lojistik regresyon yöntemlere ilişkin DSO ve SS değerleri verilmiştir.



Resim 4.2. Parkinson veri seti için ilk 14 değişkene ait korelasyon tablosu

Çizelge 4.3. Parkinson veri seti için ilk 14 değişkene ait özet bilgiler tablosu

Değişkenler	Minimum Değer	Ortalama	Maksimum Değer	STD
ACTA2	0,9138	0,9785	1,1423	0,0593
ACTB	0,9107	0,9889	1,1219	0,0534
ACTR3	0,8378	1,0140	1,1389	0,0700
ARHGDA	0,9009	0,9996	1,1256	0,0631
ASAP2	0,7754	0,989	1,2688	0,1244
ASS1	0,7619	0,9813	1,2998	0,1209
ATAD1	0,7200	1,0084	1,2100	0,1195
ATP2B2	0,6759	1,0130	1,2444	0,1518
ATPIF1	0,3931	1,0147	1,5778	0,3061
BLVRB	0,7345	0,9697	1,3571	0,1539
BTF3	0,8488	0,9964	1,1280	0,0695
CACNG2	0,5656	1,0210	1,2863	0,1841
CMBL	0,8494	1,0038	1,1638	0,0798
CNTNAP2	0,9226	0,9983	1,1406	0,0559

Çizelge 4.4. Parkinson veri seti için tüm lojistik regresyon yöntemlere ait DSO ve SS değerleri

Lojistik Regresyon Yöntemler	DSO (SS)
Ridge	98,300 (68)
Lasso	91,360 (11)
Elastik-Ağ	98,900 (13)
Uyarlanabilir Lasso ($\gamma = 1$)	90,620 (12)
Uyarlanabilir Elastik-Ağ ($\gamma = 1$)	94,200 (14)
Uyarlanabilir Lasso ($\gamma = 2$)	90,080 (13)
Uyarlanabilir Elastik-Ağ ($\gamma = 2$)	93,220 (14)
Esnek Uyarlanabilir Lasso	90,600 (11)

- Birinci satırdaki değerler DSO'yu, parantez içindekiler SS'i göstermektedir.

Parkinson veri setinde, Elastik-Ağ lojistik regresyon en yüksek ortalama DSO deęerini veren yöntem olarak gözlenmiştir. Bu yöntemde sıfırdan farklı regresyon katsayısı 13 olarak belirlenmiştir.



5. SONUÇ VE ÖNERİLER

Bu tez çalışmasında cezalandırılmış lojistik regresyon yöntemlerinden Ridge lojistik, Lasso lojistik, Elastik-Ağ lojistik, Uyarlanabilir Lasso lojistik ($Y=1$), Uyarlanabilir Lasso lojistik ($Y=2$), Uyarlanabilir Elastik-Ağ lojistik ($Y=1$), Uyarlanabilir Elastik-Ağ lojistik ($Y=2$), Esnek Uyarlanabilir Lasso lojistik regresyon yöntemleri ele alınarak; farklı örneklem çapları, farklı değişken sayıları için $n > p$ (düşük boyutlu) ve $p > n$ (yüksek boyutlu) durumlar ele alınarak; düşük, orta ve yüksek olmak üzere farklı korelasyon yapıları için bir simülasyon çalışması dizayn edilmiştir. Tüm lojistik regresyon yöntemlerin RStudio ortamında geçen toplam süresine, yöntemlere ait ortalama doğru sınıflandırma oranı (DSO), standart sapma değerlerine (STD) ve tahmin edilen regresyon katsayılarının sıfırdan farklı olanlarının sayısına (SS) bakılmıştır. Çalışmada iki gerçek veri seti ele alınmıştır. Birinci veri seti gözlem sayısının değişken sayısından büyük olduğu Kaggle veri merkezinden çekilen Pirinç türü sınıflandırma veri seti, ikinci veri seti değişken sayısının gözlem sayısından büyük olduğu GSE68719 koduyla GEO veri tabanında bulunan Parkinson veri setidir. Her iki veri setinde de simülasyon çalışmasında kullanılan cezalandırılmış lojistik regresyon yöntemleri ele alınarak tüm yöntemlerin SS ve DSO değerine bakılmıştır.

Ridge lojistik regresyon yöntemi değişkenlerin tamamını modele dahil eder yani değişken seçimi yapamaz. Bu yüzden hem simülasyon çalışması hem gerçek veri setleriyle olan çalışmada Ridge lojistik regresyon hesaplanmış ancak yorumlara dahil edilmemiştir.

Yapılan simülasyon çalışmasında, iki farklı senaryo uygulanmıştır. Birinci senaryo gözlem sayısı değişken sayısından büyük yani düşük boyutlu, ikinci senaryo ise değişken sayısı gözlem sayısından büyük yani yüksek boyutlu olarak tasarlanmıştır.

Düşük boyutlu senaryoda $p=15$ iken, DSO değerlerine ait verilerde t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir ancak değişken sayısı arttıkça yöntemler arasındaki farkın azaldığı hatta $p=50,100$ için farkın olmadığı gözlenmiştir. Bu durumda en yüksek ortalama DSO ve en düşük STD'yi veren yöntemler belirlenmeye çalışılmıştır ve * ile gösterilmiştir. (*) ise bahsedilen yöntemler dışındaki

tüm yöntemlerin en yüksek ortalama DSO ve en düşük STD'ye sahip olduğunu belirtmektedir.

- $p=15$ iken düşük ve orta korelasyon düzeyinde; $n=100$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $n=500$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Yüksek korelasyon düzeyinde ise $n=50$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $n=100$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Düşük korelasyon düzeyinde; n arttıkça en yüksek ortalama DSO değeri değişmemiş ve Oracle prensibini kullanan Uyarlanabilir Lasso lojistik regresyon ($Y=1$), orta korelasyon düzeyinde ise Elastik-Ağ lojistik regresyon olmuştur. Yüksek korelasyon düzeyinde; $n=20$ için en yüksek ortalama DSO değeri Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) yönteminde ancak n arttıkça Elastik-Ağ lojistik regresyon yönteminde gözlenmiştir.
- $p=25$ iken düşük ve orta korelasyon düzeyinde; $n=50$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $n=100$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Yüksek korelasyon düzeyinde ise $n=50$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $n=500$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Düşük korelasyon düzeyinde; $n=50$ için Lasso lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($Y=1$), Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$)*, Esnek Uyarlanabilir Lasso lojistik regresyon*, $n=100, 500$ için Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon dışındaki tüm yöntemler (*) en yüksek ortalama DSO değerini vermiştir. Orta korelasyon düzeyinde; $n=50$ için ise Lasso lojistik regresyon dışındaki tüm yöntemler (*), $n=100$ için Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ($Y=1$) lojistik regresyon dışındaki tüm yöntemler (*) ve $n=500$ için Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) en yüksek ortalama DSO değerini vermiştir. Yüksek korelasyon düzeyinde; $n=50$ için Uyarlanabilir Lasso lojistik regresyon ($Y=2$) dışındaki tüm yöntemler (*), $n=100$ için Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon dışındaki tüm yöntemler (*) ve $n=500$ için sadece Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) en yüksek ortalama DSO değerini vermiştir.
- $p=50$ iken düşük ve orta korelasyon düzeyinde; $n=100$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $n=500$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Yüksek korelasyon düzeyinde ise $n=500$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $n=100$ için ise tüm yöntemlerin ortalama DSO değeri

en düşük olduğu gözlenmiştir. $p=100$ iken düşük korelasyon düzeyinde; en yüksek ortalama DSO diğer korelasyon düzeylerine göre daha büyük olarak gözlenmiştir. $p=50, 100$ iken gözlem sayısı arttıkça en yüksek ortalama DSO değerini veren yöntemler değişmemiştir ve Uyarlanabilir Lasso lojistik regresyon ($Y=1$)*, Uyarlanabilir Lasso lojistik regresyon ($Y=2$)*, Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$)*, Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$)*, Esnek Uyarlanabilir Lasso lojistik regresyon* yöntemleri en yüksek ortalama DSO değerlerini vermişlerdir. STD ise farklı korelasyon durumlarında tüm yöntemlerde eşit olarak gözlenmiştir. Yöntemler arasında fark gözlenmemiştir.

Değişken sayısı arttıkça en yüksek DSO değerini veren yöntemler artmış ve bu yöntemler arasında literatür yer alan Esnek Uyarlanabilir Lasso regresyon yönteminin lojistik regresyon yöntemine uyarlanan Esnek Uyarlanabilir Lasso lojistik regresyonda yer almıştır.

Yüksek boyutlu senaryoda $n=15$ iken DSO değerlerine ait verilerde t istatistiğine göre tüm yöntemler arasında %95 güven düzeyinde fark olduğu tespit edilmiştir ancak gözlem sayısı arttıkça yöntemler arasındaki farkın azaldığı hatta $n=100$ için farkın olmadığı gözlenmiştir. Bu durumda en yüksek ortalama DSO ve en düşük STD'yi veren yöntemler belirlenmeye çalışılmıştır ve * ile gösterilmiştir. (*) ise bahsedilen yöntemler dışındaki tüm yöntemlerin en yüksek ortalama DSO ve en düşük STD'ye sahip olduğunu belirtmektedir.

- $n=15$ iken düşük korelasyon düzeyinde; $n=100$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=50$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Orta korelasyon düzeyinde; $p=500$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=50$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Yüksek korelasyon düzeyinde; $p=500$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=20$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Düşük korelasyon düzeyinde; $p=20, 100$ için Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$), $p=50$ için Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), $p=500$ için Uyarlanabilir Lasso lojistik regresyon ($Y=2$) en yüksek ortalama DSO değerini veren yöntemler olarak gözlenmiştir. Orta korelasyon düzeyinde; $p=50$ için Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), $p=20$ için Uyarlanabilir Lasso lojistik regresyon ($Y=1$) ve Uyarlanabilir Elastik-Ağ lojistik

regresyon ($Y=2$), $p=100$ için Elastik-Ağ lojistik regresyon, $p=500$ için Elastik-Ağ lojistik regresyon ve Uyarlanabilir Lasso lojistik regresyon ($Y=2$) en yüksek ortalama DSO değerini veren yöntemler olarak gözlenmiştir. Yüksek korelasyon düzeyinde; $p=50$, 100 için Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$), $p=20$ için Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), $p=500$ Elastik-Ağ lojistik regresyon en yüksek ortalama DSO değerini veren yöntemler olarak gözlenmiştir.

- $n=25$ iken orta ve yüksek korelasyon düzeyinde; $p=100$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=50$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Düşük korelasyon düzeyinde ise $p=500$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=50$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. $p=50$ için tüm korelasyon düzeylerinde; en yüksek ortalama DSO değeri Elastik-Ağ lojistik regresyon, $p=100$ için düşük korelasyon düzeyinde; Elastik-Ağ lojistik regresyon*, Uyarlanabilir Lasso lojistik regresyon ($Y=1$)*, Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$), Esnek Uyarlanabilir Lasso lojistik regresyon, orta ve yüksek korelasyon düzeyinde ise Lasso lojistik regresyon ve Elastik-Ağ lojistik regresyon seçilmiştir. $p=500$ için düşük korelasyon düzeyinde; Lasso lojistik regresyon, orta ve yüksek korelasyon düzeylerinde ise Elastik-Ağ lojistik regresyon en yüksek ortalama DSO değerini veren yöntemler olarak gözlenmiştir.
- $n=50$ iken düşük ve orta korelasyon düzeyinde; $p=500$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=100$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. Yüksek korelasyon düzeyinde ise $p=100$ için tüm yöntemlerin ortalama DSO değeri en yüksek, $p=500$ için ise tüm yöntemlerin ortalama DSO değeri en düşük olduğu gözlenmiştir. $p=100$ için tüm korelasyon düzeylerinde; Lasso lojistik regresyon* ve Elastik-Ağ lojistik regresyon*, $p=500$ için düşük korelasyon düzeyinde; Uyarlanabilir Lasso lojistik regresyon ($Y=2$), Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$) dışındaki tüm yöntemler (*), orta korelasyon düzeyinde; Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, yüksek korelasyon düzeyinde ise Lasso lojistik regresyon* ve Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$)*, en yüksek ortalama DSO değerini veren yöntemler olarak seçilmiştir.
- $n=100$ iken düşük korelasyon düzeyi diğer korelasyon yapılarına göre ortalama DSO değerleri daha yüksektir ayrıca gözlem sayısı arttıkça en yüksek ortalama DSO değerini veren yöntemler değişmemiştir ve Lasso lojistik regresyon*, Elastik-Ağ lojistik regresyon*, Uyarlanabilir Lasso lojistik regresyon ($Y=1$)*, Esnek Uyarlanabilir Lasso lojistik regresyon* yöntemleri en yüksek ortalama DSO değerlerini vermişlerdir.

STD ise farklı korelasyon durumlarında tüm yöntemlerde eşit olarak gözlenmiştir. Yöntemler arasında fark gözlenememiştir.

Simülasyon çalışmasında, düşük boyutlu senaryoda gözlem sayısı arttıkça Uyarlanabilir Lasso lojistik regresyon, Uyarlanabilir Elastik-Ağ lojistik regresyon ve Esnek Uyarlanabilir Lasso lojistik regresyon yöntemlerinin, yüksek boyutlu senaryoda ise değişken sayısı arttıkça Lasso lojistik regresyon, Elastik-Ağ lojistik regresyon, Uyarlanabilir Lasso lojistik regresyon ve Esnek Uyarlanabilir Lasso lojistik regresyon yöntemlerinin iyi performans gösterdiği gözlemlenmiştir. Düşük boyutlu senaryoda, genellikle değişken sayısı arttıkça SS değerinin de arttığı gözlenmiştir. Yüksek boyutlu senaryoda ise genellikle gözlem sayısı arttıkça SS değerinin de azaldığı gözlenmiştir. RStudio'da geçen süre her çalışmada farklılık göstermiş ancak genel olarak korelasyon arttıkça bu sürede de artış gözlenmiştir.

- Gerçek veri setleriyle yapılan çalışmalarda her iki farklı senaryodaki simülasyon çalışmasındaki gözlem ve değişken sayılarına bakarak eş bir sonuç vermiş her iki gerçek veri setinde de en yüksek ortalama DSO değerini Elastik-Ağ lojistik regresyon vermiştir. SS'i en yüksek yöntemler olarak Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=1$) ve Uyarlanabilir Elastik-Ağ lojistik regresyon ($Y=2$) gözlenmiş bu iki yöntemin SS'leri eşit çıkmıştır.

Çalışma sonucunda, küçük gözlem sayısına veya küçük değişken sayısına sahip olan gerçek veri seti ve simülasyon çalışmasında ortak olarak görülen yüksek korelasyon düzeyinde yani değişkenler arasında çoklu bağlantı varlığında Elastik-Ağ lojistik regresyon yöntemi en yüksek ortalama DSO değerini vererek değişken seçimi yapan en iyi yöntem olarak kabul edilmiştir. Gözlem sayısı ve değişken sayısı arttıkça en yüksek ortalama DSO'ya sahip olan yöntemler artmış yani birbiri yerine kullanılabilen, aynı STD'ye sahip ve değişken seçimi yapan cezalandırılmış lojistik regresyon yöntemlerin olduğu gözlenmiştir. Ayrıca literatürde yer alan Esnek Uyarlanabilir Lasso regresyon yönteminin lojistik regresyon yöntemine uyarlanmasıyla elde edilen Esnek Uyarlanabilir Lasso lojistik regresyon yöntemi de bu yöntemler arasında yer almıştır.

Bu tez çalışması farklı korelasyon yapılarında farklı cezalandırılmış lojistik regresyon yöntemlerin performanslarını karşılaştırmış, literatüre Esnek Uyarlanabilir Lasso lojistik

regresyon yöntemini kazandırmaya çalışmıştır ve tez çalışmasının bu alanda yapılan araştırmalara ışık tutacağı düşünülmektedir. Bunun yanı sıra incelenen cezalandırılmış lojistik regresyon farklı senaryolarda, farklı korelasyon yapılarında ve gerçek veri setleriyle yapılacak çalışmalara daha çok ihtiyaç olduğu düşünülmektedir.



KAYNAKLAR

- Aguilera, A. M., Escabias, M. and Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8), 1905-1924.
- Aldrich, J.H. and Nelson, F.D. (1984). *Linear probability, logit and probit model*. New York: Sage, 241-265.
- Algarni, Z. Y. and Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42(23), 9326-9332.
- Araveeporn, A. (2021). The higher-order of adaptive lasso and elastic net methods for classification on high dimensional data. *Mathematics*, 9(10), 1091.
- Ari, A. ve Önder, H., 2013, Farklı veri yapılarında kullanılabilecek regresyon yöntemleri. *Anadolu Tarım Bilimleri Dergisi*, 28(3), 168-174.
- Bielza, C., Robles, V. and Larrañaga, P. (2011). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications*, 38(5), 5110-5118.
- Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics*, 43(4), 951- 973.
- Bruce, P. and Bruce, A. (2017). *Practical statistics for data scientists*. Kaliforniya: O'Reilly Media, 363.
- Darlington, R. B. and Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation*. New York: Guilford Publications, 125-141.
- Dhhan, W., Rana, S. and Midi, H. (2016). A high breakdown, high efficiency and bounded influence modified GM estimator based on support vector regression. *Journal of Applied Statistics*, 44(4), 700-714.
- Duffy, D. E. and Santner, T. J. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics-Theory and Methods*, 18(3), 959-980.
- Ekstrand, J., Westergren, A., Årestedt, K., Hellström, A. and Hagell, P. (2022). Transformation of Rasch model logits for enhanced interpretability. *BMC Medical Research Methodology*, 22(1), 332.
- Ertuğrul, S., (2018). *Farklı tuzluluğa sahip göllerde popülasyon büyüklüğünün alansal ve zamansal modellenmesi*. Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 145.

- Flexeder, C., (2020). *Generalized lasso regularization for regression models*. Doctoral dissertation, Ludwig-Maximilians-Universität Institut für Statistik, München: 147.
- Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4(2), 627.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: prediction, inference and data mining*. New York: Springer-Verlag, 200.
- Hilbe, J. M. (2011). Logistic regression. *International Encyclopedia of Statistical Science*, 1, 15-32.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- Işık, A., (2006). *Uygulamalı istatistik II*. İstanbul: Beta Basım Yayım, 690-700.
- İnternet: 10-katlı çapraz doğrulama yöntemi gösterim şeması. URL: <https://www.ahmetcevahircinar.com.tr/2017/03/25/k-kez-capraz-dogrulama-yontemi/> Son Erişim Tarihi: 21.11.2023.
- İnternet: k -katlı çapraz doğrulama yöntemi gösterim şeması; URL: https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29, Son Erişim Tarihi: 21.11.2023.
- İnternet: Ridge ve Lasso regresyon yöntemlerinin tahminleri: Ridge, lasso ve Elastik-Ağ regresyon yöntemlerinin tahminleri. URL: <https://blog.jovian.com/regularization-techniqueslasso-and-ridge-90d3cc73ca4c>, Son Erişim Tarihi: 17.12.2023.
- İpek, O., (2000). Ridge regresyon üzerine bir çalışma. Yayınlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 42-53.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer. 112, 18.
- Lemein, J., (2010). *fMRI-based image reconstruction using the elastic net*. Bachelor Thesis, Radboud Universiteit, Nijmegen: 28.
- Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1), 374-393.
- Metz, C. E. (1978, October). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4), 283-298.
- Montgomery, D. and Peck, E. (1991). *Introduction to linear regression analysis*. NewYork, Wiley Series, 6, 441.

- Myers, R. H., Montgomery, D. C., Vining, G. G. and Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences*. New York: John Wiley & Sons. 123-127.
- Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4), 705-724.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1), 267-288.
- Yates, L. A., Aandahl, Z., Richards, S. A. and Brook, B. W. (2023). Cross validation for model selection: a review with examples from ecology. *Ecological Monographs*, 93(1), e1557.
- Zhang, R., Zhao, T., Lu, Y. and Xu, X. (2022). Relaxed adaptive lasso and its asymptotic results. *Symmetry*, 14(7), 1422.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301-320.
- Zou, H., and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4), 1733.





Gazili olmak ayrıcalıktır