İSTATİSTİKSEL ÖĞRENMEDE DOĞRUSAL SINIFLANDIRMA TEKNİKLERİ VE DİYABET VERİSİ ÜZERİNE BİR UYGULAMA

Gizem UYLU

Yüksek Lisans Tezi

İstatistik Anabilim Dalı

Uygulamalı İstatistik Bilim Dalı

Danışman: Doç. Dr. Kadir Özgür PEKER

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Ocak 2020

JÜRİ VE ENSTİTÜ ONAYI

Gizem UYLU'nun İSTATİSTİKSEL ÖĞRENMEDE DOĞRUSAL SINIFLANDIRMA TEKNİKLERİ VE DİYABET VERİSİ ÜZERİNE BİR UYGULAMA başlıklı tezi 17/01/2020 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Eskişehir Teknik Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği"nin ilgili maddeleri uyarınca, İstatistik Anabilim dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Jüri Üyeleri	**Unvan Adı Soyadı**	**İmza**
Üye (Tez Danışmanı)	: Doç. Dr. Kadir Özgür PEKER	
Üye	: Prof. Dr. Fikret ER	
Üye	: Dr. Öğr. Üyesi Levent TERLEMEZ	

Prof. Dr. Murat TANIŞLI

Lisansüstü Eğitim Enstitüsü Müdürü

ÖZET

İSTATİSTİKSEL ÖĞRENMEDE DOĞRUSAL SINIFLANDIRMA TEKNİKLERİ VE DİYABET VERİSİ ÜZERİNE BİR UYGULAMA

Gizem UYLU

İstatistik Anabilim Dalı

Uygulamalı İstatistik Bilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Ocak 2020

Danışman: Doç. Dr. Kadir Özgür PEKER

Günümüzde kaydedilen teknolojik gelişmelere paralel olarak, veri madenciliği, istatistiksel öğrenme ve makine öğrenme gibi konulara olan ilgi giderek artmaktadır. Özellikle mühendislik ve sağlık bilimleri ile finans ve endüstri gibi alanlarda istatistiksel öğrenme önemli bir rol oynamaktadır. Bu çalışmada, öncelikle istatistiksel öğrenme teorisine ilişkin genel bilgiler verilmiş, ardından istatistiksel sınıflandırma analizlerinden olan diskriminant analizi ve lojistik regresyon analizine ait teorik bilgiler ayrıntılı olarak incelenmiştir. Bu yöntemler kullanılarak elde edilen modellerin sınıflandırma sonuçlarını karşılaştırmak amaçlanmıştır. Kurulan bu modeller yardımıyla, bağımsız değişkenlerde meydana gelen değişimlerin yanıt değişkenini nasıl etkilediğine ilişkin yorum yapılabilmektedir. Çalışmanın amacı doğrultusunda gerçekleştirilen uygulamada, 21-81 yaş aralığındaki 768 adet Pima Yerlisi kadın bireye ait veriler kullanılmıştır. Bu veriler için kurulan diskriminant analizi ve lojistik regresyon analizi modellerinin sınıflandırma başarıları değerlendirilmiş ve elde edilen sonuçlar yorumlanmıştır.

Anahtar Sözcükler: İstatistiksel öğrenme, Diskriminant analizi, Lojistik regresyon analizi, Sınıflandırma.

ABSTRACT

LINEAR CLASSIFICATION TECHNIQUES IN STATISTICAL LEARNING AND AN APPLICATION TO DIABETES DATA

Gizem UYLU

Department of Statistics

Programme in Applied Statistics

Eskişehir Technical University, Institute of Graduate Programs, January 2020

Supervisor: Assoc. Prof. Dr. Kadir Özgür PEKER

In parallel with current technological developments, interest in topics such as data mining, statistical learning and machine learning is gradually increasing. Statistical learning plays an important role particularly in areas such as engineering and health sciences with finance and industry. In this study, firstly general information about statistical learning theory is given and then the theoretical information about discriminant analysis and logistic regression analysis which are the statistical classification methods are examined in detail. It is aimed to compare the classification results of the models obtained using these methods. It can be interpreted how the changes in the independent variables affect the response variable by means of these models. Data set of 768 Pima Native female individuals aged 21-81 years are used in the application which is performed for the purpose of the study. The classification successes of the models for discriminant and logistic regression analysis, which are set up for this data set, are evaluated and the results are interpreted.

Keywords: Statistical learning, Discriminant analysis, Logistic regression analysis, Classification.

TEŞEKKÜR

Bu çalışmanın her aşamasında değerli katkı ve deneyimleriyle bana yön veren, her türlü desteği ve anlayışı gösteren, sabırla bu süreci bitirmemi bekleyen danışman hocam Sayın Doç. Dr. Kadir Özgür PEKER'e, tezin gerçekleşmesinde tüm içtenlikleriyle değerli bilgilerini benimle paylaşan Sayın Prof. Dr. Fikret ER ve Sayın Dr. Öğr. Üyesi Levent TERLEMEZ hocalarıma ve tezin yazımı sırasında manevi destekleriyle yanımda olan tüm bölüm hocalarıma teşekkürlerimi sunarım.

Öğrenim hayatım boyunca her zaman yanımda olan bana güvenen ve hiçbir zaman desteğini esirgemeyen sevgili babam H. Adnan BULUT'a, benimle beraber uykusuz kalan bana ve bebeğime sabırla bakan canım annem Nurdan BULUT'a ve çalışma boyunca tüm teknik desteğimi sağlayan dert ortağım biricik kardeşim Berkay BULUT'a çok teşekkür ederim.

Çalışma dönemimin başından sonuna kadar manevi desteklerini esirgemeyen tüm dostlarım ve en büyüğünden en küçüğüne kadar sahip olduğum kocaman aileme, ben evden uzaktayken benimle beraber tüm zorluklarına sabırla katlanan, desteğini ve ilgisini hiç esirgemeyen sevgili eşim Emre UYLU'ya teşekkür ederim. En özel teşekkürüm ise yokluğuma sabrettiği için moral kaynağım minik kızım Doğa UYLU'ya onlara sahip olduğum için çok şanslı olduğumu belirterek canım çekirdek aileme en içten dileklerimle teşekkür ederim.

Gizem UYLU

17/01/2020

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Eskişehir Teknik Üniversitesi tarafından kullanılan "bilimsel intihal tespit programı"yla tarandığını ve hiçbir şekilde "intihal içermediğini" beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Gizem UYLU

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

SİMGELER DİZİNİ

x_i	:	Girdi Değişkeni
y_i	:	Yanıt Değişkeni
n_i	:	Her Bir Gruptaki Gözlem Sayısı
B	:	Gruplar Arası Varyans Matrisi
W	:	Grup İçi Varyans Matrisi
z_k	:	Kritik Değer
Λ	:	Wilks' Lambda Katsayısı
$\ell(\beta)$	:	Olabilirlik Fonksiyonu
$L(\beta)$	:	Logaritmik Olabilirlik Fonksiyonu
w_i	:	Ağırlık Değeri
$I(\beta)$	:	Bilgi Matrisi
D	:	Sapma Ölçütü
G	:	Olabilirlik Oranı Testi

1. GİRİŞ

İstatistik biliminde *veri*, genellikle tecrübe, gözlem, deney ya da ölçüm sonucunda elde edilen bir veya birden fazla sayı, kelime veya görsellerden oluşan bir kümeyi ifade eder.

İstatistiksel öğrenme (Statistical Learning) teorisi istatistik ve fonksiyonel analiz alanlarından makine öğrenme tasarımı için temel bir yapı oluşturur ve veriye dayalı tahmini bir fonksiyon bulma problemi ile ilgilenir (Hastie, Tibshirani and Friedman, 2008). Öğrenme bilimi, öncelikle mühendislik ve diğer disiplinlerle birlikte kullanılan istatistik, veri madenciliği ve yapay zeka gibi alanlarda kilit rol oynamaktadır. Aynı zamanda istatistiksel öğrenme, finans, endüstri ve sağlık gibi birçok bilim alanında da aktif olarak rol oynamaktadır.

İstatistiksel öğrenme, "veriden öğrenme" olarak da ifade edilebilir. Eldeki nicel ya da nitel olarak ölçülmüş veriyi kullanır ve verideki birtakım öznitelikleri kullanarak tahminde bulunmaya çalışır (Hastie, Tibshirani and Friedman, 2008).

Tüm istatistiksel öğrenme problemleri beklenen hatayı en aza indirecek şekilde yapılandırılabilir. Öğrenme algoritmalarında ilk olarak verilerden rassal bir örneklem seçilir. Model kurmada kullanılan bu örneklem eğitim verisidir. Bu veriler değişken ve yanıt arasındaki ilişkiyi geliştirmek için kullanılır ve model parametreleri bu verilere dayanarak tahmin edilir. Model performansını değerlendirmek için ise verilerin geri kalanı kullanılır. Bu bölüme de test verileri adı verilir. Test verileri, yalnızca birkaç güçlü aday model arasından bir model kesinleştiğinde kullanılır. İstatistiksel öğrenme de modellere ve modellerin yorumlarına odaklanılmakla birlikte, tutarlılık ve belirsizlik gibi kavramlarla da ilgilenilmektedir.

Model kullanılan bağımsız değişkenlerde meydana gelen değişimlerin yanıt değişkeninde meydana getirmesi beklenen değişimler ile ilgili yorum yapılabilir. Yorum yapılabilmesinin nedeni, oluşturulan tahmin fonksiyonunun doğrusal olmasıdır. Doğrusal olması yorumlama yapmaya imkan tanımış olsa da gerçek hayatta doğrusallık çok kolay sağlanamamakta ve beraberinde fazlaca varsayım gerektirmektedir.

Öğrenme probleminin modellenmesi oldukça yaygındır. Ancak üç temel problem tipi vardır;

i. Sınıflandırma

ii. Regresyon

iii. Yoğunluk tahmini.

Sınıflandırma, öğrenme sürecinin bir parçası olarak gösterilebilir. Günlük hayattaki gibi bilimsel araştırmalarda da sorunların çözümünde sağladığı fayda nedeniyle oldukça sık başvurulur.

Kullanıldığı alanların bazıları;

- Dolandırıcılık, sahtekarlık gibi olayları tahmin etmede,
- İnternet ortamında web sayfalarının sınıflandırılmasında,
- E-postaların spam olup olmadığını belirlemede,
- Tıbbi teşhislerde, DNA dizilerine dayalı hastalık tiplerini belirlemede,
- Sigortacılık ve bankacılık hesaplamalarında,
- Görüntü kümesinden kimlik tespitinde ve parmak izi uygulamalarında.

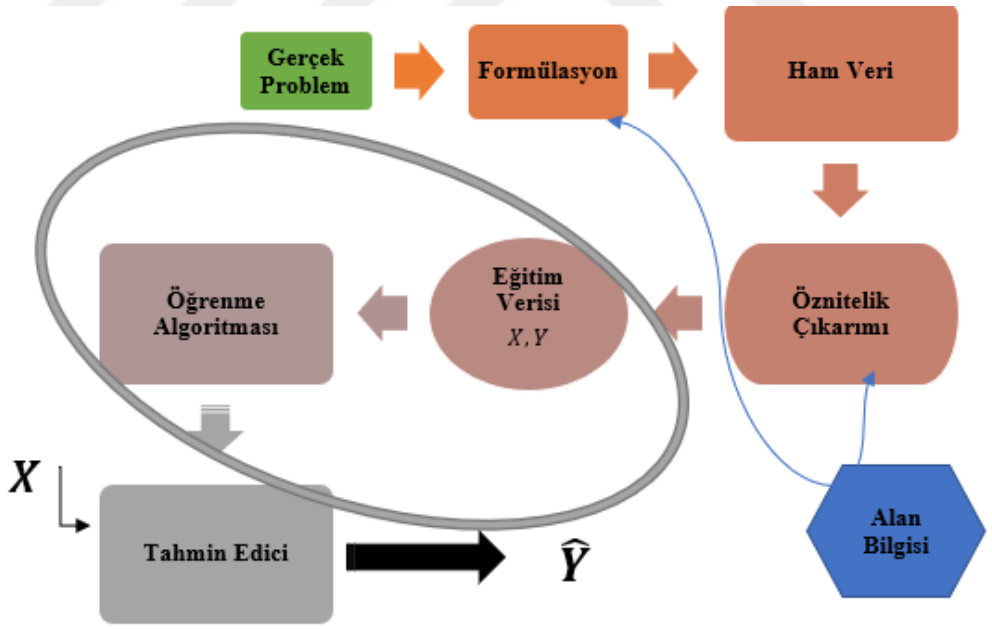
Sınıflandırmanın ilk adımı, veri tabanında bulunan değişkenlere uygun bir modelin oluşturulmasıdır ve bunu yaparken verilere ait özellikler kullanılır. Sınıflandırma modeli işleminin gerçekleşmesi için verilerin bir kısmı eğitim kümesi olarak kullanılır. Rasgele seçilen bu verilere algoritma uygulanır ve sınıflama modeli elde edilir. İkinci adımda modelin doğruluğu test kümesi ile belirlenir. Son adımda ise elde edilen model sınıflandırma için kullanılır.

Sınıflandırma, veri dizisinin istatistik veya makine öğrenimi yöntemleri kullanılarak önceden belirlenen sınıflara atanması işlemidir ve çeşitli sınıflandırma yöntemleri geliştirilmiştir. Bu çalışma, istatistiksel öğrenme teorisi hakkında kısa bir bilgi verildikten sonra istatistiksel sınıflandırma tekniklerinden lojistik regresyon analizi ile diskriminant analizi ayrıntılı olarak ele alınacak ve 768 Pima Yerlisinin kadın bireylerinden alınan veriler ile sınıflandırma başarıları gözlemlenecektir.

2. İSTATİSTİKSEL ÖĞRENME TEORİSİ

İstatistiğin temelleri 200 yıl öncesine dayanmaktadır. Ancak problemlerin çözümünde kullanılan sistemli istatistiksel yöntemler 1920'lerin sonlarından bu yana çalışılmaya başlanmıştır (Vapnik, 1998, s.2). Fakat klasik istatistiksel dağılımların yetersiz kalmaya başladığı bu dönemde istatistiksel öğrenme teorisi ortaya çıkmıştır. İstatistiksel öğrenme teorisi 1960'ların sonlarında Vladimir Vapnik ve Alexey Chervonenkis tarafından tanıtılmıştır.

Veriden öğrenme biliminden de genellikle nicel veya kategorik çıktı ölçümü olarak ifade edilen bir Y değişkeni vardır ve bu çıktı, bir dizi özellik incelenerek tahmin edilir. Burada özellik, X değişkeni olarak gösterilirken girdi ölçümleri olarak adlandırılır. Girdi ve çıktı ölçümlerinin gözlemleri arasındaki ilişkiyi geliştiren eğitim verileri vardır. Bu verilerine verilere dayanarak model parametreleri tahmin edilir ve öğrenme algoritmaları kurulur. İstatistiksel öğrenme modelinde algoritmalar tarafından üretilen tahminlemeler üzerinde çalışılır.



Şekil 2.1. *İstatistiksel Öğrenme Modeli.*

Şekil 2.1.'de görüldüğü gibi, herhangi bir modeli oluşturma sürecinde ilk olarak, verinin grafiksel ve analitik olarak anlaşılması gerekir. Veriler karmaşık ise açıklayıcı veri analizi teknikleri yardımıyla görsel ve analitik süreçlerin birleşiminden en iyi sonucu

alınır. Daha sonra model oluşturmak için verilerden rassal bir örneklem seçilir. Modelin başarısını değerlendirmek için ise verinin geri kalanı kullanılır. Model oluşturmak için kullanılan örnekleme *eğitim seti,* geri kalan verilere ise *test seti* adı verilir.

Öğrenme algoritmasında amaç, geçmiş verileri kullanarak gelecek veriler için genel bir algoritma oluşturmaktır. Bunun için birçok model içinden veriye uyan ve en etkili sonuca ulaştıracak olan modeli seçmek gerekir.

İstatistiksel öğrenme teorisi, modelleme problemine yeni bir bakış açısı getirmektedir. Bir modelin uygun olduğuna veya başka bir modelin daha iyi sonuç verip vermeyeceğine nasıl karar verilir? Klasik istatistik doğru modelin formunun bilindiğini varsayıp, modelin parametrelerini belirlemeyi amaçlarken, istatistiksel öğrenme teorisi modelin formunun bilinmediğini kabul ederek doğru olabilecek modeller arasında en iyi modelin bulunmasını hedeflemektedir (Tolun, 2008, s.19).

İstatistiksel Öğrenme modelin sahip olması gereken bazı özellikler vardır;

- Modelin eğitim setine iyi bir şekilde uyması,
- Modelin mümkün olduğunca robust olması.

Basit bir modelin karmaşık bir modele göre robust olma olasılığı daha fazladır. Dolayısıyla daha başarılı tahminlemeler yapması beklenir. Sağlam bir algoritmanın test verileri de daha iyi sonuç verecektir. Karmaşık bir modelle çalışıldığında, eğitim setine daha uyumlu olmakla birlikte test setindeki performansı daha düşük olabilir ya da tam tersi basit bir model eğitim setine tam olarak uymayabilir. Dolayısıyla eğitim ve test setlerinin arasında bir denge mekanizması olduğu ortaya çıkar. Aynı şekilde test hatası tahmin edicilerin yeni veriler üzerinde çalışıp çalışmadığını gösterirken, eğitim hatası verilerin iyi uyup uymadığını göstermektedir **(http-1).**

Aşırı uyumlu bir model eğitim verileri ile yakın ilişkilidir. Sapma düşük olabilir, ancak varyansı yüksek olacaktır. Bu, tahmin edicinin eğitim verileri üzerinde çok iyi çalıştığını, ancak test verilerinde önemli ölçüde daha kötü bir sonuç verdiğini gösterir. Sapma, bir modelin gerçeğe ne kadar benzediğinin bir ölçüsüdür. X ve Y arasındaki ilişki ikinci dereceden olduğunda doğrusal bir model önerilirse, önerilen model yanlı olacaktır. Aynı öğrenme algoritması çoklu bağımsız eğitim verilerine uygulanırsa, farklı bir tahmin ortaya çıkar. Bu tahminlerin ortalaması incelenen istatistiğin gerçek değeri ile aynı ise

tahmin yansız olur. Bir model daha fazla parametre ve karmaşık ilişki içeriyorsa sapmalar daha küçük olma eğilimindedir.

Varyans ise farklı eğitim verileri kullanıldığında tahminin ne kadar değiştiğinin bir ölçüsüdür. Varyans daha karmaşık veriler için daha yüksek olma eğilimindedir. Sapma ve varyans arasındaki dengeyi bulmak, optimum tahmin modelinin geliştirilme amacıdır. Çünkü bir modelin doğruluğu her ikisinden de etkilenir**.**

2.1. Denetimli ve Denetimsiz Öğrenme

Öğrenme algoritmaları, ilgilenilen veri setini araştırarak girdi-çıktı ölçümleri arasında bir ilişki ortaya koymaktadır. Bu ise denetimli öğrenme olarak adlandırılır. Öğrenme sürecinde eğitim verilerinde bir çıktı değişkeninin varlığı söz konusu ise bu durumda öğrenme yöntemi denetimlidir denilir. Ancak denetimsiz öğrenme yönteminde sadece girdi değişkeni gözlenmektedir. Çıktı ölçümü yoktur ya da göz ardı edilebilmektedir.

Denetimli ve denetimsiz ifadeleri birbirine zıt anlamda kullanılmaktadır. Denetimli ve denetimsiz yöntemleri sürecin bütünü açısından değerlendirmek gerekirse;

- Denetimsiz yöntemler daha çok veriyi anlamaya, tanımaya, keşfetmeye yönelik olarak kullanılmakta ve sonraki uygulanacak yöntemler için fikir vermeyi amaçlamaktadır,
- Denetimli yöntemler ise veriden bilgi ve sonuç çıkarmaya yönelik kullanılmaktadır

denilebilir (Koyuncugil ve Özgülbaş, 2009, s.21).

Sınıflandırma yöntemleri de denetimli öğrenme modelinin bir parçasıdır.

2.2. Sınıflandırma

İstatistiksel karar verme süreçlerinden biri olan sınıflandırma probleminde bir gözlemin hangi gruba atanacağına karar verilmelidir. Veri setinde bulunan her gözlemin bir özelliği vardır ve bu özellikler sınıflara ayrılır. Hangi sınıfa ait olduğu belirlenen gözlemlerle bir model oluşturur. Oluşturulan modelin eğitim setinde yer almayan gözlemleri ile başarısı ölçülür. Bir sınıf algoritmasında özellikleri bilinen yeni

gözlemlerin atanacağı sınıfları tahmin edebilecek bir modelin belirlemesi için mevcut gözlemlerden faydalanılabilir. Tahmin edilecek hedef değişkenler sonlu ve kategoriktir.

Sınıflandırma yöntemlerinde oldukça farklı algoritmalar kullanılmaktadır. Bu algoritmaların hangi alanlarda hangi tür değişkenlerde daha etkin sonuçlar verdiğinin bilinmesi yöntemlerin başarısını artırmaktadır. Bu nedenle sınıflandırma yöntemlerinin karşılaştırılarak uygulanması önemlidir. Çok sayıda algoritma olması, her algoritmanın kendi içinde farklı parametrelerle çalışması ve birden çok versiyonunun bulunması, çalışılan algorimaların farklı amaçlara yönelik olması, kullanılan veri kaynağının farklı olması, algoritmaların farklı veri tipini desteklemesi ve veri üzerinde ön işlemlerin uygulayıcıya bağlı olması gibi sebeplerle farklı sonuçlar elde edilmiştir (Kuyucu, 2012).

İstatistiksel öğrenme teorisinin amaçlarından biride, dağılımdan bağımsız yöntemler ile sınıflandırma probleminin çözümüne yönelik modeller oluşturmaktır.

İstatistiksel dağılımın bilindiği ideal durumlarda sınıflandırma probleminin çözümü kolaylıkla bulunabilir ve sonuçları temel istatistiklerle hesaplanabilir. Ancak gerçek hayat uygulamalarında verilerin dağılımı genellikle bilinmemektedir. Böyle bir durumda daha önceden sınıflandırılmış verileri tekrar düzenleyerek bu bilgi eksikliğinin üstesinden gelinmeye çalışılır. Bu tip öğrenme problemine, istatistiksel öğrenme, örneklerden öğrenme, denetimli öğrenme, istatistiksel örüntü tanıma ve istatistiksel örüntü sınıflandırması adları verilir ve bu, makine öğrenmesi yöntemlerinden yalnızca bir tanesidir (Kulkarni and Harman, 2011, s.6).

2.3. İstatistiksel Öğrenme İçin Model Oluşturma

Sınıflandırma bir denetimli öğrenme problemidir ve kurulumu aşağıdaki adımlarla mümkündür.

i. $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), \quad i = 1, .., n$ Eğitim verileri,

ii. $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ vektörü, p bağımsız değişkenlerin sayısı olmak üzere her bir $X_j,\ j = 1, \dots, p$ birbirinden bağımsız ve niceldir,

iii. Her özellik vektörü X için tanımlanmış Y kategorik yanıt değişkeni,

iv. X'e dayalı Y'yi en iyi şekilde tahmin eden

$$g(x) = f(x, a), a \in \Theta \quad \text{g: X} \rightarrow \text{Y ve g} \in \text{G fonksiyonu}$$

olarak gösterilir **(http-1).**

Öğrenme probleminde amaç; oluşturulan $g(x)$ fonksiyonlar kümesinden test seti kullanılarak yanıt değişkenini en iyi şekilde tahmin eden fonksiyonu oluşturabilmektir. $g(x)$ fonksiyonunun seçilebilmesi için bir ölçüte ihtiyaç duyulur. Bu ölçüt hata oranının, dolayısıyla $P(g(X) \neq Y)$ olasılığının düşük olmasıdır (Bousquet, Boucheron and Lugosi, 2004, s.178).

İkili sınıflandırma problemi ele alındığında Y yanıt değişkeni 0 ve 1 olarak tanımlanabilir. Bayes sınıflandırma kuralı $y \in \{1, -1\}$ ve $\boldsymbol{X} = (x_1, x_2, \dots, x_P)$ özellik vektörüne göre etiketlendiğinde ve eğitim verilerinin $p(x, y)$ koşullu dağılımından çekildiği göz önüne alındığında,

$$p(x, y) = p(y|x)\, p(x) \qquad (x_i, y_i), \qquad i = 1,2,..,n \tag{2.1}$$

biçimde gösterilir.

Öğrenme yöntemleriyle hesaplanan yanıt değişkeni ile gerçek y arasındaki farkı minimum yapan en iyi yöntem seçilir ve bu farkı hesaplayan fonksiyona kayıp fonksiyonu adı verilir. Kayıp fonksiyonu, $L(y, g(x))$ ile ifade edilir.

L'nin, 0-1 kayıp fonksiyonu olduğu varsayımı altında beklenen kayıp, bir modelin tüm verileri için ortaya çıkması beklenen hataların sayısıdır ve risk olarak adlandırılır.

$$R(g) = E[L(y, g(x))] = \int L\big(y, g(x)\big) dP(x, y) \tag{2.2}$$

Buna göre risk, Eşitlik (2.2)'deki fonksiyon ile ifade edilir.

İstatistiksel öğrenmede $R(g)$'nin *(beklenen riskin)* minimizasyonu amaçlanmaktadır. Dolayısıyla, oluşturulan G fonksiyonlar kümesinden risk fonksiyonunu en aza indiren $g(x)$ fonksiyonu bulunmaya çalışılır.

$$\hat{g} = argmin_{\mathrm{g}\in \mathrm{G}}\, E[L(y, g(x))] \tag{2.3}$$

Eşitlik (2.3)'e Bayes sınıflandırma kuralı adı verilir. 0-1 kayıp fonksiyonu ise,

$$L(y, g(x)) = \begin{cases} 1 & eğer\ \ y \neq g(x) \\ 0 & eğer\ \ y = g(x) \end{cases} \tag{2.4}$$

$$E[L(y, g(x))] = 1 - P(\mathrm{G} = g|X = x) \tag{2.5}$$

Eşitlik (2.4) ve (2.5) ile ifade edilir.

$$\hat{g} = argmin_{\mathrm{g}\in\mathrm{G}}\, E\big[L\big(y, g(x)\big)\big]$$

$$= argmax_{\mathrm{g}\in\mathrm{G}}\, P(\mathrm{G} = g|X = x) \qquad (2.6)$$

Eşitlik (2.6) ile gösterilen çözüm, Bayes sınıflandırıcısı olarak bilinir ve kesikli $P(\mathrm{G}|X)$ dağılımı kullanılarak en uygun sınıflara ayrılır (Hastie, Tibshirani and Friedman, 2008, s.21).

2.4. Sınıflandırma İçin Doğrusal Yöntemler

Girdi alanı, sınıflandırmaya göre etiketlenen bir bölgeler kümesi biçiminde her zaman bölünebilir. Bu bölgenin sınırları, tahmin fonksiyonuna bağlı olarak, düz ya da eğri olabilir. Doğrusal sınıflandırma yöntemleriyle anlatılmak istenen, belli sayıda işlemlerden sonra, bu karar sınırlarının doğrusal olduğudur (Hastie, Tibshirani and Friedman, 2008, s.101).

İki sınıf problemi:

- İki sınıf arasındaki karar sınırı öznitelik vektör uzayında bir hiper düzlemdir.
- *P* boyutlu giriş alanındaki bir hiper düzlem kümesidir.

$$\left\{x: a_0 + \sum_{j=1}^{p} a_j x_j = 0\right\}$$

$$\left\{x: a_0 + \sum_{j=1}^{p} a_j x_j > 0\right\}\left\{x: a_0 + \sum_{j=1}^{p} a_j x_j < 0\right\} \qquad (2.7)$$

Eşitlik (2.7) ile bir hiper düzlemle ayrılmış iki bölge gösterilmektedir.

İkiden çok sınıf problemi:

- Herhangi bir *k* ve *l* sınıfları arasındaki karar sınırları bir hiperdüzlemdir.

Hiper düzleme karar vermek için örnek doğrusal yöntemler; doğrusal diskriminant analizi ve lojistik regresyon analizleridir **(http-1).**

3. DİSKRİMİNANT ANALİZİ

Diskriminant analizi, ayırma analizi olarak da bilinmektedir ve sınıflandırma işleminde kullanılan çok değişkenli istatistiksel yöntemdir. İstatistiksel karar teorisine dayanan bu analiz kullanılarak sınıfları birbirinden ayıracak en iyi fonksiyonu bulmak amaçlanmaktadır.

Diskriminant analizi ile ilgili ilk çalışmalar genellikle biyoloji ve davranış bilimlerinde görülmektedir. Son yıllarda ise Finans ve Matematik bilimlerinde yaygın olarak kullanılmaktadır.

Diskriminant analizi bir kategorik bağımlı değişken ile sayısal değerler alan bağımsız değişkenlerin arasında yapılır. Diskriminant analizi bağımsız değişkenlerin bağımlı değişkenleri etkilemelerine göre ya aynı ya da farklı gruplara göre sınıflandırılmasını sağlar **(http-2).**

1850'li yıllardan beri kullanılan diskriminant analizi kullanım kolaylığı, istatistiksel yollarla ayrıştırma ve sınıflandırma yaparak hata payını minimuma indirmesi nedeniyle sıklıkla tercih edilmektedir. Sınıflandırma yöntemlerinde kullanılan en yaygın analiz türü olmakla beraber bazı varsayımsal sınırlandırmalardan dolayı alternatif yöntemler geliştirilmiştir. Diskriminant Analizi diğer analiz türleriyle beraber de uygulanabilmesinden dolayı geliştirilmeye de açık bir analiz türüdür.

3.1. Diskriminant Analizi Tarihçesi

Diskriminant analizine ilişkin ilk çalışmalar çok eski olmakla beraber bilinen ilk kullanımı 1851 yılında James Joseph Sylvester tarafından önerilmiştir.

Sir Ronald Aylmer Fisher'in (1936), yılında yayımladığı *The Use of Multiple Measurements in Taxonomic Problems* isimli makalesinde tanımlanan doğrusal diskriminant analizi, çok değişkenli bir istatistik analiz yöntemidir.

Roy (1939), yılında ilk kez iki normal dağılım için kovaryans matrisinin eşitliğini test etmiştir. Günümüzde Diskriminant Analizi yaparken gruplar arasında kovaryans eşitliği olup olmadığı kontrol edilir.

Rao (1948), Doğrusal Diskriminant Analizini ikiden fazla grup için geliştirmiştir.

Cavalli (1945), Penrose (1947), Smith (1947), von Mises (1954), verileri tek değişkenli ya da çok değişkenli iki gruba ayırma problemi üzerine çalışmışlardır.

Anderson ve Bahadur (1962), farklı ortalamalı, farklı varyans covaryans matrisli çok değişkenli normal dağılımlı verilerin sınıflara ayrılması işlemini çözmüşlerdir.

Bartlett ve Please (1963), aynı ortalamalı, farklı varyans kovaryans matrisli grupların sınıflandırılmasını araştırmışlardır.

Smith (1947), Cooper (1963) ve Bunke (1964) kuadratik diskriminant fonksiyonlarını incelemişlerdir.

Lachenburch (1975), çoklu normal dağılım ve eşit kovaryans varsayımına uymamanın Diskriminant analizi sonuçlarını büyük oranda etkilemeyeceğini göstermiştir.

Klecka (1980), normal dağılım göstermeyen değişkenlerin Diskriminant Analizi sonuçlarını değiştiremeyeceğini ortaya koymuştur.

Carl J. Huberty, Mohamed H. Hussein (2003), yaptıkları çalışmada diskriminant analizi kullanırken raporlama problemini incelemişlerdir.

Kim Fung Lam, Jane W. Moy (2003), parçalı doğrusal programlama ve Fisher'in doğrusal diskriminant analizinin benzer sonuçlar verdiğini ortaya koymuştur.

Sueyoshi (2004), Diskriminant analizi ile standart tam sayılı programlama modelleri ve iki aşamalı tam sayılı programlama modellerini kullanarak bunların sınıflandırma işleminde gösterdikleri başarıları araştırmıştır.

Sugnet Gardner, Niel J. le Roux (2005), diskriminant analizlerinin biplot metodolojisiyle nasıl çalışacağı üzerinde inceleme yapmışlardır.

Michael J. Brusco, Douglas Steinley (2011), doğrusal diskriminant analizinin değişkenleri belirleyebilmesi için hazırlanan kesin ve yaklaşık algoritmaları incelemişlerdir.

Cesar J. Rebellon (2012), deneysel özellikteki testlerin diskriminantın geçerlilik ve seçimlilik üzerindeki etkisini araştırmıştır.

3.2. Diskriminant Analizi Varsayımları ve Amaçları

Diskriminant analizi, iki veya daha fazla grubun çok sayıda değişkene bağlı olarak karşılaştırılmasını sağlayan bir yöntemdir. Analizde grup sayısı bilinmekte ve bu sayı

analiz devam ettiği sürece değişmemektedir. Analizin amacı, grupların hangi değişkenler açısından birbirinden farklılaştığının ortaya çıkarılmasıdır. Diğer bir ifadeyle, grupların ayırıcı özelliklerinin belirlenmesidir (Oktay Fırat ve Demirhan, 2003, s.18).

Diskriminant analizi bağımlı değişkenin kategorik olduğu ve bağımsız değişkenlerin sürekli sayısal değerler aldığı durumlarda kullanılan bir tekniktir. Tüm analizlerde olduğu gibi sonuçlardaki hata oranının en aza indirilmesi için belli varsayımların gerçekleşmesine ihtiyaç vardır. Bu varsayımlar kısaca aşağıdaki gibi sıralanabilir.

- *Bağımsız değişkenler çok değişkenli normal dağılıma sahiptir.*

Bu varsayım bağımsız değişkenlerin her birinin ayrı ayrı normal dağılması gerektiğini belirtmektedir. Eğer bu varsayım gerçekleşmezse diskriminant fonksiyonunda hata oluşmaktadır. Değişkenlerin normal dağılıp dağılmadığı çizilecek histogram grafiği yardımıyla incelenebilir. Normal olmama durumu, basıklıktan değil de çarpıklıktan kaynaklanıyorsa, önemli sorunlar ortaya çıkarır. Bu durumda Lojistik Regresyon Analizi tercih edilmelidir **(http-3).**

- *Bütün gruplar için korelasyon matrisleri eşittir.*

Veri kümesindeki değişkenlerin ortak varyans-kovaryans matrisine sahip çok değişkenli anakütleden çekildiği varsayılmaktadır.

Bağımlı değişken tarafından oluşturulan her bir grupta bağımsız değişkenlerin varyans ve ortalamaları aynıdır (Börüban, 2009, s. 27).

Varyansların homojenliği serpilme diyagramları ile incelenebilir. Homojenliği bozan sapan değerler olabilir ve diskriminant analizi bu değere karşı oldukça hassastır. Böyle bir durumda değerlere değişken dönüştürme tekniği uygulanabilir. Homojen olmama durumu bir ya da daha fazla grupta sapan değerler olduğunu gösteriyor olabilir.

- *Bağımsız değişkenler en azından eşit aralıklı ölçek kullanılarak ölçülmektedir.*

Bağımlı değişkenleri gruplara ayırmak için bağımsız değişkenler kullanılır. Bu değişkenler ayırıcı değişkenlerdir ve en az aralık ölçekle ölçülmelidir.

- *Bir bağımsız değişken ile diğer bir bağımsız değişkenin çoklu doğrusal bağımlılığı olmamalıdır.*

Bu varsayım gerçekleşmediğinde analiz sürecinde yapılan hesaplamalar güvenirliliğini kaybeder. Tolerans düzeyi, yapılan hesaplamaların güvenirliliğinin arzu edilen düzeyde tutulmasını sağlamaktadır. Diğer bir deyişle, tolerans düzeyi kabul edilebilecek çoklu bağlantının derecesini belirlemektedir. R_i^2, diskriminant fonksiyonundaki değişkenlerden biri ile geriye kalan diğer bağımsız değişkenler arasındaki açıklama katsayısını gösterdiğinde, bir değişkenin toleransı, $1 - R_i^2$'ye eşittir. Bir bağımsız değişkenle analizdeki diğer değişkenler arasında daha yüksek çoklu bağlantı, daha düşük tolerans değeri anlamına gelmektedir (Albayrak, 2006).

- *En küçük grubun toplam gözlem sayısı bağımsız değişkenlerin sayısından fazla olmalı ve bağımsız değişkenlerin sayısı toplam gözlem sayısının en fazla iki eksiği olmalıdır.*

Diskriminant analizi iki adımdan meydana gelen bir süreçtir. Birinci adımda diskriminant fonksiyonu tahmin edilirken, ikinci adımda hesaplanan fonksiyon değerlerine göre nesnelerin olasılıklar dahilinde sınıflandırılması gerçekleştirilecektir **(**Akpınar, 2014, s.191**)**.

- *Veri kümesi grupların birbirinden ayrılmasını sağlayacak kadar doğru ve gerekli değişkenleri içermelidir.*
- *Grup sayısı iki ya da ikiden fazla olmalıdır.*

Diskriminant analizi, gruplar arasındaki farklılıkları inceler ve bu farklılıkların hangi değişkenler üzerinde yoğunlaştığını belirler. Böylece grupları farklılaştıran etkenler açıklanmış olur.

- *Her bir grup için en az iki gözlem olmalıdır.*

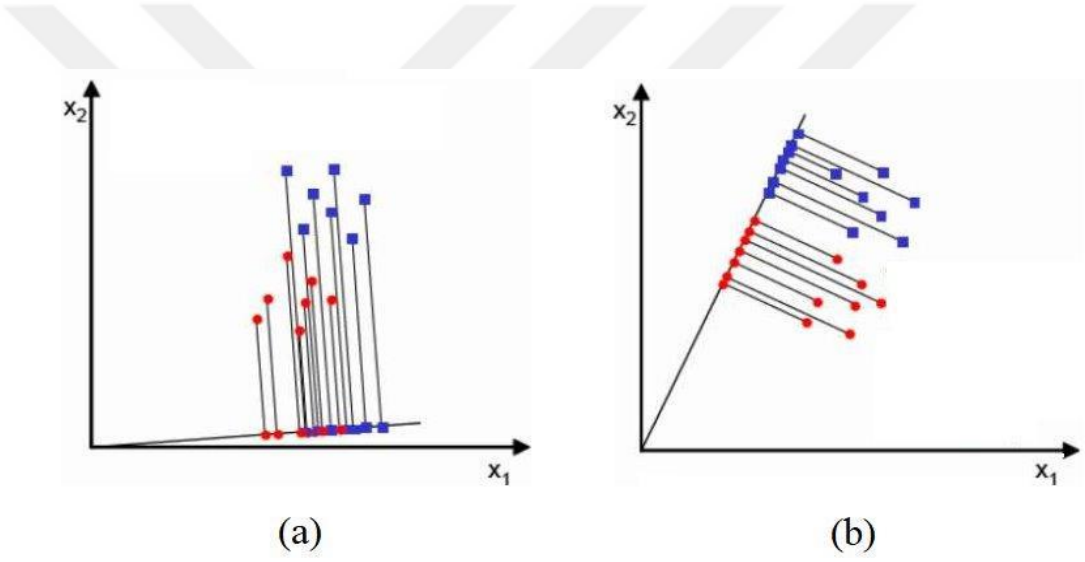
Diskriminant analizi birçok alana hizmet etse de, temelde iki amacı olduğu savunulmaktadır. Bu amaçlardan ilki, gruplar arasındaki farkın hangi değişkenden kaynaklandığını belirlemektir. Buda, grupları ayıran özellikleri ayrıntılı biçimde incelemek anlamına gelmektedir. Bu inceleme sırasında hangi değişkenin gruplar arasındaki farklılığa ne derecede etki ettiği belirlenebilir. Gruplar arasındaki ayrımı gerçekleştirmede kullanılan özellikler ayırt edici değişkenler adını alan bağımsız değişkenlerdir (Burmaoğlu, 2009).

Bu ayrımı en rahat biçimde yapabilmek için bir yöntem belirlemek de bu amaç doğrultusunda gerçekleşir.

İkincisi ise analizde ele alınan gözlemlerin hangi gruba ait olduğunu belirlemektir. Her bir yeni gözlemin sınıflandırma hatası en az olacak şekilde grup üyeliklerinin tayin edilmesi amaçlanmaktadır. Gözlem değerlerini diskriminant değerlerine göre sınıflayabilmek için yöntemler geliştirilir. Herhangi bir durumda yanlış sınıflandırma maliyeti önceden hesaplanır.

Diskriminant analizi, gözlemleri en az hatayla gruplara ayırma amacı taşır. Aynı zamanda hangi gözlemlerin ait oldukları gruplarda uç değer olduğunu belirlemekte de diskriminant analizi kullanılmaktadır.

Doğrusal diskriminant analizinin genel görünümü Şekil 3.1.'de verilmiştir.



Şekil 3.1. *Doğrusal Diskriminant Analizi Genel Görünüm Grafiği (http-4)*

Şekil 3.1.(a)'da görülen şekilde noktaların doğru üzerindeki izdüşümlerine bakıldığında sınıfların net bir ayrımı yapılamamaktadır. (b)'de görülen doğruda ise sınıfların net bir şekilde ayrılabildiği görülmektedir ve burada problemin boyutu (x_1, x_2) olarak verilen iki özellikten tek bir skaler y değerine indirgenmektedir.

3.3. İki Grup Olması Durumunda Doğrusal Diskriminant Analizi

Gözlem birimlerinden birden çok değişkenin ölçümlendiği ve bunların sınıflandırılmasında değişkenler arası ilişkileri göz önünde bulunduran fonksiyona *Diskriminant Fonksiyonu* adı verilir.

Her bir grup için yazılabilecek diskriminant fonksiyonu Eşitlik (3.1)'de görülmektedir.

$$y_i = a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \cdots + a_p x_{ip} \tag{3.1}$$

Burada, $x_1, x_2, \ldots, x_p$ bağımsız değişkenleri $a_1, a_2, \ldots, a_p$ bu değişkenlerin katsayılarını, y_i ise gruplarda gözlenen yanıt değişkenini belirtmektedir.

Diskriminant fonksiyonu bulunurken, Eşitlik (3.2)'de görülen gruplar arası varyansın grup içi varyansa oranının maksimum yapılması gerekmektedir (Tatlıdil, 1996).

$$F = max\left(\frac{Gruplar\ Arası\ Varyans}{Grup\ \text{İçi}\ Varyans}\right) \tag{3.2}$$

Diskriminant fonksiyonu gruplar arası farklılığı maksimize edecek şekilde oluşturulur. Böylece grupların en iyi şekilde ayrımı söz konusu olacaktır.

$$\gamma = \frac{\boldsymbol{a'Ba}}{\boldsymbol{a'Wa}} = \frac{\boldsymbol{a'}\left(\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'\right)\boldsymbol{a}}{\boldsymbol{a'}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'\right)\boldsymbol{a}} \tag{3.3}$$

Katsayıların bulunmasında ve gruplar arası farklılığı belirlemek için kullanılan Eşitlik (3.3), R. A. Fisher (1936) tarafından geliştirilmiştir ve bu şekilde tanımlanan bir fonksiyonun en büyüklenmesi amacını taşır. Ayrıca bu yöntemde ayırıcı değişkenlere doğrusal dönüşüm uygulanır ve bu değerler iyi ayırım yapan yeni değerler olarak ortaya çıkarlar.

Bu eşitlikte;

n_i: her bir gruptaki gözlem sayısını,

x_{ij}: gözlem değerlerini,

$\bar{\boldsymbol{x}}$: $p \times 1$ boyutlu genel ortalama vektörünü,

$\bar{\boldsymbol{x}}_{\boldsymbol{i}}$: $p \times 1$ boyutlu her bir gruba ait ortalama vektörünü

$\boldsymbol{a}$: $p \times 1$ boyutlu katsayılar vektörünü

belirtmektedir.

$\boldsymbol{B}: p \times p$ boyutlu gruplar arası varyans matrisidir ve bu matris genel ortalama vektöründen her bir grup için hesaplanmış ortalama vektörlerinin çıkartılmasıyla oluşturulur.

$\boldsymbol{W}: p \times p$ boyutlu grup içi varyans matrisidir ve bu matris her gruptaki gözlemlerin kendi ortalama vektörlerinin farklarının alınmasıyla elde edilir.

$\boldsymbol{a}$ vektöründeki katsayıları bulabilmek için fonksiyonun bu elemanlanlara göre kısmi türevi alınarak sıfıra eşitlenir. Bu işlem gerçekleştirildikten sonra,

$$(\boldsymbol{Ba} - \boldsymbol{\gamma W a}) = 0 \tag{3.4}$$

elde edilir. Bu eşitlikte her iki taraf W^{-1} ile çarpıldığında,

$$\boldsymbol{W^{-1}Ba} - \boldsymbol{W^{-1}W a\gamma} = 0$$

$$\boldsymbol{W^{-1}Ba} - \boldsymbol{a\gamma} = 0$$

$$\boldsymbol{a}(\boldsymbol{W^{-1}B} - \boldsymbol{\gamma I}) = 0 \tag{3.5}$$

olarak yazılabilir. Burada $\boldsymbol{a} = 0$ olduğu görülür. Eğer $(\boldsymbol{W^{-1}B} - \boldsymbol{\gamma I})$ matrisinin tersi alınıyor ise sonuç 0 olur fakat tersi alınmadığında determinantının sıfır olması gerekir. Başka bir deyişle $\boldsymbol{a}$ vektörünün gerçek bir çözümü olması için

$$\left|\boldsymbol{W^{-1}B} - \boldsymbol{\gamma I}\right| = 0 \tag{3.6}$$

olması gerekmektedir. Bu eşitlikten, $\boldsymbol{W^{-1}B}$ matrisine karşılık gelen γ_i $(i = 1,2, \dots, r)$ özdeğerleri ve onlara karşılık gelen özvektörler elde edilir. k grup sayısını, p değişken sayısını gösterdiğinde, diskriminant fonksiyonu sayısı; $r = \min(k-1, p)$ olmak üzere elde edilen r tane özvektör olacaktır.

İki grup olması durumunda diskriminant analizinde sadece bir diskriminant fonksiyonu elde edilerek, gruplar arasındaki fark tek bir sistem içinde incelenebilmektedir.

$\bar{x}_J^{(1)}$, birinci grubun j' inci değişken ortalaması $\bar{x}_J^{(2)}$ ise ikinci grubun j' inci değişken ortalaması iken,

$$\boldsymbol{d}' = \left(d_1, d_2, \dots, d_p\right) \tag{3.7}$$

$$\boldsymbol{d_j} = \left(\bar{x}_J^{(1)} - \bar{x}_J^{(2)}\right) \tag{3.8}$$

vektörleridir. Eşitlik (3.8) ile iki grup ortalamaları arasındaki fark ifade edilir.

$$\boldsymbol{C} = \frac{1}{n_1 + n_2 - 2}\boldsymbol{W} \tag{3.9}$$

olarak tanımlandığında, Eşitlik (3.3) aynı zamanda,

$$\gamma = \frac{n_1 n_2}{n_1 + n_2}\frac{\boldsymbol{a'dd'a}}{\boldsymbol{a'Ca}} \tag{3.10}$$

biçiminde de yazılabilir.

Eşitlik (3.9)'daki C, grup içi kovaryans matrisinden elde edilmiş bir matristir. Bu durumda (3.10) eşitliği p bilinmeyenli denklem sistemidir ve iki grubun grup içi varyanslarının eşitliği varsayıldığında,

$$\boldsymbol{a} = \boldsymbol{C^{-1}d} \tag{3.11}$$

ile hesaplanır. Bu durumda diskriminant fonksiyonu değerlerinin, grup içi kovaryans matrisinin tersi ve iki grup ortalamaları arasındaki fark ile doğru orantılı olduğu görülmektedir.

Eşitlik (3.1)'deki y_i'nin diskriminant fonksiyonu olması için standartlaştırılması gerekir. Birinci grubun diskriminant fonksiyonu ortalaması, x_i'lerin yerine grup ortalama değerleri koyularak hesaplanır. Öncelikle birinci grup y_1 için ortalama ve varyans hesaplanmak istensin.

$$y_i = a_i x \qquad i = 1,2,\dots,p \tag{3.12}$$

$$\bar{y}^{(1)} = a_1\bar{x}_1^{(1)} + a_2\bar{x}_2^{(1)} + \cdots + a_p\bar{x}_p^{(1)} \tag{3.13}$$

olarak alındığında birinci grup için beklenen değer,

$$E\left(y^{(1)}\right) = E\left(a^{(1)}x\right) = a^{(1)}E(x) \tag{3.14}$$

olur. Standartlaştırılan her bir değişkenin ortalaması sıfır olduğundan, bütün gözlemlere göre standartlaştırılmış diskriminant fonksiyonunun ortalaması 0 ve her bir y_i nin standart sapması 1 olarak hesaplanır. Bunun için, bir gözlemin y_i değeri sıfırdan büyükse sınıflandırılabilir fakat y_i değeri sıfırdan küçükse başka bir gruba sınıflandırılır.

İki grubun varyans kovaryans matrislerinin ortak olması durumunda, bulunan diskriminant fonksiyonunun varyansı da ortak olacaktır. Fonksiyonun varyansı karesel biçimde ve matris notasyonu ile,

$$Var(y) = \sum_{i=1}^{p}\sum_{j=1}^{p} c_{ij} a_i\, a_j = \boldsymbol{a'Ca} \tag{3.15}$$

olarak yazılabilir (Tatlıdil, 1996, s.259).

$$\begin{aligned} Var(y) &= \boldsymbol{d'C^{-1}CC^{-1}d} \\ &= \boldsymbol{d'C^{-1}d} \end{aligned} \tag{3.16}$$

$\boldsymbol{d} = \boldsymbol{Ca}$ ve $\boldsymbol{d} = (\bar{x}^{(1)} - \bar{x}^{(2)})$ eşitliklerinden,

$$\begin{aligned} Var(y) &= \boldsymbol{d'C^{-1}Ca} = \boldsymbol{d'a} \\ &= \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)'\boldsymbol{a} = \bar{x}^{(1)'}\boldsymbol{a} - \bar{x}^{(2)'}\boldsymbol{a} \\ &= \bar{y}^{(1)} - \bar{y}^{(2)} \end{aligned} \tag{3.17}$$

sonuçlarına ulaşılır (Tatlıdil, 1996, s.260).

Birinci grup için varyans hesapladığında,

$$Var\left(y^{(1)}\right) = a^{(1)'}Ca^{(1)} = a^{(1)'}\left(\frac{1}{n-1}\boldsymbol{W}\right)a^{(1)} \tag{3.18}$$

sonucuna ulaşılır ve bulunan ortalama ve varyanslara dayanarak, γ_1, özdeğer ve $a^{(1)}$ona karşılık gelen özvektör ise birinci diskriminant fonksiyonu Eşitlik (3.19) ile elde edilir.

$$y^{(1)} = \boldsymbol{a_1^{(1)}}x_1 + \boldsymbol{a_2^{(1)}}x_2 + \cdots + \boldsymbol{a_p^{(1)}}x_p \tag{3.19}$$

Diskriminant fonksiyonu katsayıları gruplar arası ayırımları en büyükleştiren kısmi katsayılardır. Ayırım değişkeninde meydana gelebilecek değişimler gözlem değerinde de değişimi yansıtır. Değişkenlerin sınıflandırmaya olan katkı miktarlarının karşılaştırılabilmesi için katsayıların standartlaştırılması gerekir. Bu standartlaştırmayı yaparken Eşitlik (3.20) kullanılır.

$$\boldsymbol{a_i}^{(j)} = \sqrt{\boldsymbol{a_i}^{(j)}\boldsymbol{W_{ii}}} \tag{3.20}$$

Bu eşitlikte $i = 1,2,\ldots,p$ değişken sayısını gösterirken, $j = 1,2,\ldots,r$ grup sayısını belirtmektedir.

Standartlaştırılmış diskriminant fonksiyonu katsayıları, değişkenlerin modeldeki mutlak önemini göstermektedir. Değişkenlerin model üzerinde ne kadar farklılaşma

yarattığını anlamamıza yardımcı olur. Katsayıların değeri, ilgili değişkenlerin gruplara ayırımına katkısını göstermektedir. Ayrıca katsayıları standartlaştırmanın en önemli sebeplerinden biri de bağımsız değişkenlerin sahip olduğu farklı ortalama ve standart sapmalardır. Bu nedenle değişkenlerdeki artış ve azalışlar aynı derecede etki etmeyeceğinden standartlaştırılmış katsayılar tercih edilir. Her bir değişkenin diskriminant fonksiyonuna olan katkısını ölçmek için standartlaştırılmış diskriminant fonksiyon katsayıları kullanılırken, diskriminant fonksiyonu ve bağımsız değişken arasındaki ilişkiyi anlamlandırmak için yapısal katsayılar kullanılır.

Yapı katsayıları diskriminant fonksiyonu ile ayırıcı değişkenlerin birbirlerine ne kadar benzediklerini öğrenmek için kullanılan katsayılardır. Fonksiyonun oluşmasında her bir değişkenin ne derecede katkıda bulunduğunu gösterir. Yapısal katsayının mutlak değerinin büyük olması ya da 1'e yakın olması, fonksiyon ile değişken yapısının aynı olduğunu, mutlak değerinin küçük olması ya da 0'a yakın olması ise değişken ile fonksiyon arasındaki ilişkinin az olduğunu göstermesi açısından önemlidir (Çinko, 2003, s.13).

3.4. Diskriminant Analizinde Yeni Gözlemleri Sınıflama İşlemi

Sınıflama işlemleri, gözlemleri sınıflandırmak ya da yeni gözlemlerin gruplarının belirlenmesi için uygulanan yöntemlerdir. Oluşturulan modeller her yeni gözlemi bir gruba atar ve bu işlemin sonucunda oluşan matrise *sınıflama matrisi* adı verilir. Örneğin, iki gruba ayrışmanın söz konusu olduğu bir modelde sınıflama matrisi, gerçek gözlem değeri birinci grupta olan gözlemlerin ne kadarının, model tarafından da doğru olarak ilk gruba ayrıştırıldığını ya da hatalı olarak ikinci gruba ayrıştırıldığını gösterir (Çinko, 2003).

Sınıflama işleminin uygulandığı yöntemlerden ilki *Sınır (kritik) Değer Yöntemi*dir. Bu yönteme göre sınıflandırılma yapıldığında her bir gözlem için bir diskriminant skoru hesaplanır ve bu skorlar birbirine yakın olduğunda aynı grupta olma olasılıkları maksimumdur. Eğer gruplardaki gözlem sayıları eşit ise, kritik değer (z_k) iki grup merkezinin ortalaması olur. Gözlem sayıları farklı olduğunda ise kritik değer grupların diskriminant değerlerinin ağırlıklı ortalamasıdır.

Sınıflandırma işleminde hatalı sınıflandırma olasılığını en aza indirgemek için kullanılan diğer bir yöntem *İstatistiksel Karar Teorisi*dir. Yeni bir gözlemin, ilgili iki grup

arasından hangisine atanacağı önsel olasılıklarla belirlenir. Önsel olasılıklar eşit ya da farklı olabilir. Bu olasılıklar ve her bir gözlemin hatalı sınıflandırma olasılığı dikkate alınarak bir diskriminant değeri hesaplanır ve kritik z_k değeri ile kıyaslanır.

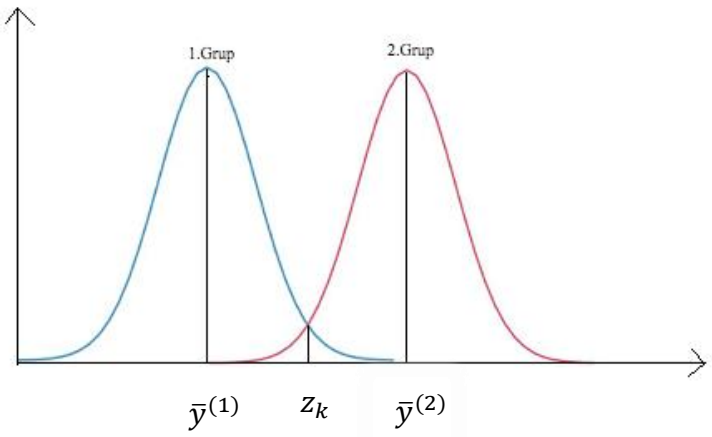
Eğer gruplardaki gözlem sayıları eşit ise ilgili gözlemlerin birinci gruba ait olma olasılığı P_1, ikinci gruba ait olma olasılığı da P_2 ile gösterilir. Gruplardaki gözlem sayılarının farklı olması durumunda ise ilk olarak gruplara ait önsel olasılıklar bulunur. Birinci gruba ait önsel olasılık q_1, ikinci gruba ait önsel olasılık q_2 ile belirtilebilir. Bu durumda hatalı sınıflandırma olasılığını minimize etmek amacıyla kritik değer, Eşitlik (3.21) ile hesaplanır.

$$z_k = \frac{\bar{y}^{(1)} + \bar{y}^{(2)}}{2} + \ln\left(\frac{q_2}{q_1}\right) \tag{3.21}$$

Eşitlik (3.21) kullanılarak iki gruplu diskriminant analizinde yeni bir gözlemin değeri hesaplanır.

Eğer, $y \geq z_k$ ise gözlem birinci gruba aittir.

Eğer, $y < z_k$ ise gözlem ikinci gruba aittir.



Şekil 3.2. *Hatalı Sınıflandırma Durumu*

Şekil 3.2.'de z_k kritik değer olmak üzere, iki grup olması durumunda hatalı sınıflandırma gösterilmiştir. Hatalı sınıflandırma büyük çoğunlukla her iki grup ortalamalarına aynı uzaklıkta bulunan gözlemlerde görülür. Eğer bir gözlem z_k noktasında yer almış ise hatalı sınıflandırma oranı büyük, z_k noktasının sağında ya da solunda yer almış ise her iki grup için hatalı sınıflandırma olasılıkları farklı olacaktır.

Gruplardaki gözlem sayıları eşit olduğunda ise her grup için hatalı sınıflandırma olasılığı aynı olacaktır.

Bu bilgilerin yardımı ile hatalı sınıflandırma maliyeti matrisi aşağıdaki gibi oluşturulur.

Tablo 3.1. *Hatalı Sınıflandırma Maliyetleri.*

	Sınıflandırma	
Tahmin Edilen Grup	Grup 1	Grup 2
Grup 1	0	$C(1\mid 2)$
Grup 2	$C(2\mid 1)$	0

Tablo 3.1'de $C(1\mid 2)$, gözlemin ikinci gruba ait olduğu bilindiği halde, yanlışlıkla birinci gruba atanmasının maliyetini, $C(2\mid 1)$ ise gözlemin birinci gruba ait olduğu bilindiği halde, yanlışlıkla ikinci gruba atanmasının maliyetini göstermektedir. Kovaryans matrisleri eşit değilken gözlemin 1. gruba ait olması durumunda hatalı sınıflandırma maliyetini minimize eden bölge (3.22) ile, 2. gruba ait olması durumunda hatalı sınıflandırma maliyetini minimize eden bölge ise (3.23) ile gösterilebilir.

$$\frac{P_1(x)}{P_2(x)} \geq \ln\left(\frac{q_2}{q_1}\right)\left(\frac{C(1\mid 2)}{C(2\mid 1)}\right) \tag{3.22}$$

$$\frac{P_1(x)}{P_2(x)} < \ln\left(\frac{q_2}{q_1}\right)\left(\frac{C(1\mid 2)}{C(2\mid 1)}\right) \tag{3.23}$$

i'inci gruba ait yoğunluk fonksiyonu $P_i(x)$ ile gösterilir ve Eşitlik (3.24) ile ifade edilir.

$$P_i(x) = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(x-\mu^{(i)}\right)'\Sigma^{-1}\left(x-\mu^{(i)}\right)\right] \tag{3.24}$$

Önsel olasılıklar kullanılarak gözlemlerin gruplara ait olma olasılıkları Eşitlik (3.25) ve (3.26) ile hesaplanır.

$$P_1' = \frac{q_1P_1(x)}{q_1P_1(x)+q_2P_2(x)} \tag{3.25}$$

$$P_2' = 1 - P_1' = \frac{q_2P_2(x)}{q_1P_1(x)+q_2P_2(x)} \tag{3.26}$$

Diğer bir sınıflandırma işlemi *Mahalanobis D^2 uzaklığı*dır. Mahalanobis D^2 gruplar arası karesel uzaklığı verir. Burada gözlemler, hesaplanan uzaklık değerine göre yakın olan gruba atanmaktadır. Mahalanobis uzaklığı ne kadar küçükse gözlem grup ortalamasına o kadar yakındır ve gözlemin ilgili gruba dahil edilme olasılığı o derecede anlamlıdır. Normal dağılım ve ortak varyans kovaryans varsayımlarını sağlayan iki grup için D^2 uzaklığı F dağılımlıdır ve Eşitlik (3.27) ile hesaplanır.

$$F = \frac{(n_1 + n_2 - p - 1)n_1 n_2}{p(n_1 + n_2 - 2)(n_1 + n_2)} D^2 \tag{3.27}$$

Burada $D^2 = \bar{y}^{(1)} - \bar{y}^{(2)}$ ile verilir ve bu eşitlik kullanılarak elde edilen fonksiyonun, grupları birbirinden ayırmada etkin rol oynayıp oynamadığının testinde, p ve $(n_1 + n_2 - p - 1)$ serbestlik derecelerindeki tablo değeri kritik değer olarak kullanılır ve F'in önemliliği, dağılımın kritik değeri $[F(\alpha, p, (n_1 + n_2 - p - 1))]$ kullanılarak belirlenir. Eşitlik (3.27) kullanılarak elde edilen değer ilgili tablo değeriyle karşılaştırıldığında, hesaplanan F değeri tablo değerinden büyükse ilgili diskriminant fonksiyonunun grupları sınıflandırmada önemli bir etkinliğe ulaştığı sonucuna varılır.

3.5. İkiden Çok Grup Olması Durumunda Doğrusal Diskriminant Analizi

İki grup için geliştirilen diskriminant analizinin genişletilmiş biçimi, p değişkenli sonlu sayıda ikiden fazla grup için kullanılır. İkiden çok grup olması durumunda amaç; gruplar arasındaki sınıflandırmayı sağlayacak diskriminant fonksiyonu sayısının minimum olacak şekilde belirlenmesidir $(\min(k-1, p))$. Çünkü çok gruplu diskriminant analizinde grupları birbirinden ayırabilmek için birden fazla diskriminant fonksiyonu oluşabilir. Bulunan diskriminant fonksiyonlarından ilki gruplar arasında en iyi sınıflandırmayı sağlayan eşitlik olacaktır.

Çok grup olması durumunda da Eşitlik (3.2)'de olduğu gibi varyans oranlarının maksimize edilmesi amaçlanmaktadır. Bu amaçla oluşturulan diskriminant fonksiyonunun ayırıcı özelliğinin önemli olup olmadığının testinde Wilks tarafından geliştirilen Λ kullanılır. Wilks' lambda katsayısı genelleştirmiş varyans olarak da bilinir. Lambda'nın 1 olması, gözlenen grup ortalamalarının farklı olmadığını, 1'den küçük değerler alması ise grup ortalamalarının farklı olduğunu belirtir (Kamışlı ve Girginer, 2010, s.14).

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|} = \frac{|\boldsymbol{W}|}{|\boldsymbol{W}+\boldsymbol{B}|} \tag{3.28}$$

Eşitlik (3.28)'de $\boldsymbol{W}$, grup içi varyans matrisini, $\boldsymbol{B}$ ise gruplar arası varyans matrisini göstermektedir. Dolayısıyla, Wilks' Lambda istatistiği, grup içi kareler toplamının genel kareler toplamına bölünmesi ile bulunur. Λ değerinin küçük bulunması (0'a yaklaşması) gruplar arası farklılığın önemli olduğunu göstermektedir. Başka bir deyişle, lambda istatistiği ne kadar küçükse, değişken gruplar arası ayrımda o kadar etkilidir denilir. Bu katsayının test edilebilmesi için öncelikle F ya da ki-kare dağılımına uygun hale getirilmesi gerekmektedir.

Gruplardaki gözlem sayısının yeterli sayıda olması durumunda Eşitlik (3.29) ile verilen test istatistiği bulunur.

$$X^2 = -\left[n - \left(\frac{p+k}{2}\right) - 1\right] \log(\Lambda) \sim \chi^2{}_{p(k-1);\alpha} \tag{3.29}$$

Tatsuoka (1971), Cooley ve Lohnes (1973), Λ oranının diskriminant fonksiyonlarının sayısını belirlemede kullanılabileceğini göstererek bir yöntem oluşturmuşlardır. Bu yöntem Eşitlik (3.30)'da verilmektedir

$$\begin{aligned} \frac{1}{\Lambda} = \frac{|\boldsymbol{T}|}{|\boldsymbol{W}|} &= \left|\boldsymbol{W^{-1}}\right| |\boldsymbol{T}| = \left|\boldsymbol{W^{-1}T}\right| \\ &= \left|\boldsymbol{W^{-1}}(\boldsymbol{W}+\boldsymbol{B})\right| = \left|\boldsymbol{W^{-1}W} + \boldsymbol{W^{-1}B}\right| \\ &= \left|1 + \boldsymbol{W^{-1}B}\right| \end{aligned} \tag{3.30}$$

(Cangül, 2006, s.64).

$\boldsymbol{W^{-1}B}$ matrisinin özvektörler matrisi kullanılarak Eşitlik (3.31) elde edilir.

$$\frac{1}{\Lambda} = \prod_{i=1}^{r} 1 + \lambda_i \ , \Lambda = \prod_{i=1}^{r} \frac{1}{1+\lambda_i} \tag{3.31}$$

Bu eşitliklerden yararlanarak

$$\begin{aligned} log\Lambda &= -log\left(\frac{1}{\Lambda}\right) = -log\left(\prod_{i=1}^{r}(1+\lambda_i)\right) \\ &= -[(1+\lambda_1) + (1+\lambda_2) + \cdots + (1+\lambda_r)] \end{aligned} \tag{3.32}$$

olmak üzere, Eşitlik (3.29) kullanılarak, Eşitlik (3.33) elde edilir.

$$X_i^2 = -\left[n - \left(\frac{p+k}{2}\right) - 1\right] \sum_{i=1}^{r} \log(1+\lambda_i) \tag{3.33}$$

Her bir λ_i için hesaplanan X_i^2'lerin $X_i^2 \sim \chi^2_{(p+k-2i)}$ dağılımına sahip olacağı düşünülerek,

$X^2 = X_1^2 + X_2^2 + \cdots + X_r^2$ biçiminde oluşturulan diskriminant fonksiyonlarının kaç tanesinin sınıflandırma için kullanılacağını ve ihmal edilenleri ayırmada önemli bir etkisinin olup olmadığını test etmek için hipotezler kurulur.

H_0: Oluşturulan diskriminant fonksiyonlarının etkisi önemli değildir.

H_1: Oluşturulan diskriminant fonksiyonlarından en az 1 tanesinin etkisi önemlidir.

Bu testte, $X_i^2 \sim \chi^2_{(1-a;p+k-2i)}$ ise H_0 reddedilir.

r tane diskriminant fonksiyonundan ilk m tanesinin önemli olduğu varsayımıyla, geriye kalan fonksiyonların önemliliğinin testinde Eşitlik (3.34) kullanılır.

$$X_{r-m}^2 = X^2 - \sum_{i=1}^{m} X_i^2 \sim \chi^2_{p(k-1)-\sum_{i=1}^{m}(p+k-2i)} \text{ , } m<r \tag{3.34}$$

Burada $m = n - 1 - (p+k)/2$ ile belirlenir ve $X_{r-m}{}^2 \geq \chi^2_{p(k-1)-\sum_{i=1}^{m}(p+k-2i)}$ ise H_0 reddedilir.

Diskriminant analizinde verinin normal dağılımlı bir matris olması ve grupların ortak varyans-kovaryans matrisine sahip olması varsayımları altında çalışılır ve gözlemlerin ikiden fazla gruptan elde edilmesi durumunda, Anderson tarafından geliştirilen bir yöntem kullanılır. Belirtilen bu yöntem, Fisher'ın iki grup olması durumunda kullanılan diskriminant analizine benzer niteliktedir.

k tane grubun ayırıcı özelliklerinin birbiriyle karşılaştırıldığı bu yöntemde, $k(k-1)/2$ tane diskriminant fonksiyonunun elde edilmesi gerekmektedir. Eşitlik (3.24)'e göre i ve j gibi iki grubun karşılaştırılması amacıyla, Eşitlik (3.35)'te verilen fonksiyon kullanılmaktadır (Tatlıdil, 1996, s.270).

$$f_{ij} = \frac{P_i(x)}{P_j(x)} = \frac{exp\left[-\frac{1}{2}\left(x-\mu^{(i)}\right)'\mathbf{\Sigma}^{-1}\left(x-\mu^{(i)}\right)\right]}{exp\left[-\frac{1}{2}(x-\mu^{(j)})'\mathbf{\Sigma}^{-1}(x-\mu^{(j)})\right]}$$

$$= x'\Sigma^{-1}\left(\mu^{(i)}-\mu^{(j)}\right) - \frac{1}{2}\left(\mu^{(i)}-\mu^{(j)}\right)'\mathbf{\Sigma}^{-1}\left(\mu^{(i)}-\mu^{(j)}\right) \quad (3.35)$$

Grup parametreleri tahmin edilerek yeni oluşturulan diskriminant fonksiyonu, Eşitlik (3.36) ile verilir.

$$f_{ij}(x) = x'\mathbf{C}^{-1}\left(\bar{x}^{(i)}-\bar{x}^{(j)}\right) - \frac{1}{2}\left(\bar{x}^{(i)}+\bar{x}^{(j)}\right)'\mathbf{C}^{-1}\left(\bar{x}^{(i)}-\bar{x}^{(j)}\right) \quad (3.36)$$

Burada C, Eşitlik (3.10)'da görüldüğü gibi iki grup olması durumunda oluşturulan tanımı dikkate alındığında Eşitlik (3.37) biçiminde oluşturulur.

$$\boldsymbol{C} = \frac{1}{(\sum_{i=1}^{k} n_i)-k}\boldsymbol{W} \quad (3.37)$$

3.6. Özel Durumlarda Kullanılan Diskriminant Analizi Teknikleri

Diskriminant analizi konusunun tamamında verilerin normal dağıldığı ve grupların ortak varyans-kovaryans matrisine sahip olduğu varsayımıyla çalışılmaktadır. Fisher yöntemi ve onun genelleştirilmiş şekli olan Anderson yöntemi Eşitlik (3.38) biçimindeki doğrusal diskriminant fonksiyonuna dayanmaktadır.

$$f = f_{ij} = L(x) = x' - \frac{1}{2}\left(\bar{x}^{(i)}+\bar{x}^{(j)}\right)'\mathrm{S}^{-1}\left(\bar{x}^{(i)}-\bar{x}^{(j)}\right) \quad (3.38)$$

Fakat belirtilen varsayımların sağlanmaması durumunda bu yöntemler yeterince iyi sonuçlar vermemektedir. Böyle durumlar için çeşitli yöntemler geliştirilmiştir.

3.6.1. Çoklu karesel diskriminant analizi

Verilerin normal dağıldığı ancak varyans kovaryans matrislerinin birbirine eşit olmadığı durumlarda kullanılır. Bu yöntem için oluşturulan fonksiyon Eşitlik (3.39)'da görüldüğü gibidir.

$$Q(x) = \frac{1}{2}log\frac{|S_j|}{|S_i|} - \frac{1}{2}\left(\bar{x}^{(i)'}S_i^{-1}\bar{x}^{(i)} - \bar{x}^{(j)'}S_j^{-1}\bar{x}^{(j)} + x'\left(S_i^{-1}\bar{x}^{i} - S_j^{-1}\bar{x}^{j}\right)\right) - \frac{1}{2}x'\left(S_i^{-1}-S_j^{-1}\right)X \quad (3.39)$$

Bu yöntem ilk olarak iki grup olduğu durumlarda kullanılmak üzere geliştirilmiştir. S_i ve S_j sırasıyla i-inci ve j-inci gruplara ait varyans-kovaryans matrisleri olduğunda Eşitlik (3.39)'da $S_i = S_j = S$ alınırsa, belirtilen karesel fonksiyonun doğrusal bir fonksiyona eşit hale dönüştüğü görülür. Çok gruplu durumlarda ise gruplar ikişerli alındığında bu fonksiyon kullanılabilir.

$Q(x) < 0$ olduğunda gözlemin R_j bölgesine,

$Q(x) \geq 0$ olduğunda ise gözlemin R_i bölgesine sınıflandırıldığı yöntemde hatalı sınıflandırılma olasılığı Eşitlik (3.40) ile verilir.

$$R_Q(x) = \left(1 + \exp\left(Q(x) - \log\left(\frac{\hat{q}_j}{\hat{q}_i}\right)\right)\right)^{-1} \tag{3.40}$$

Sınıflandırılma bölgeleri x'in karesel fonksiyonu olarak tanımlanır (Burmaoğlu, 2009).

3.6.2. Yüksek derece terimli doğrusal diskriminant analizi

Bazı değişkenlerin sürekli ve normal dağılımlı olmasının yanında bazılarının ise kesikli durumda olması başka bir deyişle veriler karışık bir şekilde bir araya gelmiş olması durumunda doğrusal diskriminant fonksiyonuna bazı karesel terimler eklenerek yeni bir fonksiyon oluşturulur

$$LQ(x) = b_0 + b_1 w_1 + \cdots + b_r w_r \qquad i = 1,2,..,r \tag{3.41}$$

(Tatlıdil, 1996).

Eşitlik (3.41)'de w_i değerleri bağımsız değişkenleri göstermektedir ve katsayılar birinci ve ikinci dereceden değişkenlere aittir. Hatalı sınıflandırma olasılığı da karesel diskriminant analiziyle benzer şekilde Eşitlik (3.42) ile ifade edilmektedir.

$$R_{LQ}(x) = \left(1 + \exp\left(LQ(x) - \log\left(\frac{\hat{q}_j}{\hat{q}_i}\right)\right)\right)^{-1} \tag{3.42}$$

Yüksek derece terimli diskriminant analizinin kullanılmasına sebebiyet veren karışık değişkenlerin olması durumunda bu yönteme alternatif olarak logistik regresyon analizi önerilmektedir.

4. LOJİSTİK REGRESYON ANALİZİ

Diskriminant analizinin temel varsayımları olan gözlemlerin çok değişkenli normal dağılımdan gelmesi ve değişkenlerin ortak varyansa sahip olması çoğu zaman sağlanamadığından, bu varsayımların bozulması durumunda lojistik regresyon analizi sıklıkla kullanılmaya başlanmıştır.

Bu analiz çeşitli varsayım (normallik, ortak kovaryansa sahip olma gibi) bozulumları durumunda, diskriminant analizi ve çapraz tablolara bir alternatif olurken, bağımlı değişkenin 0,1 gibi ikili ya da ikiden çok düzey içeren kesikli değişken olması durumunda normallik varsayımı kısıtı olmaması nedeniyle kullanım rahatlığının yanı sıra çözümlemeden elde edilen modelin matematiksel olarak çok esnek olması, kolay yorumlanabilir olması yönteme olan ilgiyi arttırmaktadır (Tatlıdil, 1996, s.289).

Son yıllarda, lojistik regresyon analizinin diğer istatistiksel tekniklere göre sıklıkla tercih edilmesinin nedenleri aşağıdaki gibi sıralanabilir:

i. Gözlemlerin normal olmayıp grup büyüklüklerinin farklı olduğu durumlarda da kullanılabilir.
ii. Sonsal olasılıkların tahminine yönelik araştırmalarda kolay uygulanıp, anlamlı sonuçlar vermektedir.
iii. Bilgisayar yazılımlarındaki ilerlemeler sonucunda çok sayıda lojistik regresyon analizi yapabilen paket program geliştirilmiştir.
iv. Model parametrelerinin tahmin edilmesi ve yorumlanmasında kolaylık sağlar.

Lojistik regresyon analizi çoğunlukla sosyal bilimlerde, istatistiksel ve makine öğrenimi uygulamalarında, veri madenciliğinde ve sınıflandırma modellerinde kullanılmaktadır. Diskriminant analizinde olduğu gibi, gözlemleri bu analizde de ait oldukları sınıflara en uygun şekilde atayabilecek model oluşturulur.

Lojistik regresyon analizinin temel amacı, bağımlı değişkenin kategorik, bağımsız değişkenlerin ise hem kategorik hem de sürekli olabildiği durumda belli bir ön koşula bağlı kalmadan bağımlı değişken ile bağımsız değişken arasındaki ilişkiyi tanımlamak amacıyla modelleme yapmaktır. Bir diğer deyişle bağımlı değişken ve bağımsız değişkenler arasındaki neden sonuç ilişkisini araştıran bir analiz yöntemidir. Değişkenler arasındaki ilişkinin doğrusal olmasına gerek yoktur. Üstel ya da binom dağılımlı bir ilişki de olabilir. Dolayısıyla lojistik regresyonda, doğrusallık olmasa bile değişkenler arasındaki ilişkiye zarar vermeden lojit bir ilişki olduğu varsayımıyla çeşitli logaritmik

dönüşümler yapılır. Bu dönüşümler sonucunda değişkenler arasındaki ilişkinin formunun doğrusal hale getirilmesi amaçlanır.

Kısıtlı ön koşulların çoğu lojistik regresyon analizinde aranmazken, bazı temel varsayımlara dikkat edilmesi gerekir;

- *Gerekli tüm bağımsız değişkenler modele dahil edilirken, gerekli olmayan bağımsız değişkenler modelden çıkartılmalıdır. Diğer bir deyişle model net bir şekilde oluşturulmalıdır.*

Bazı değişkenlerin modele dahil edilmemesi modeli yetersiz kılarken bazı değişkenlerin dahil edilmesi modelin yorumlamasında zorluklar oluşturabilir.

- *Bağımsız değişkenlerde ölçüm hatası küçük olmalıdır.*

Veri setinde kayıp (eksik) veri olmamalıdır. Kayıp veriden doğacak hatalar, katsayıların tahmininde yanlılığa ve modelin yetersizliğine neden olur **(http-5).**

- *Bağımsız değişkenler arasında çoklu bağlantı problemi olmamalıdır.*

Gözlemleri sınıflara ayırma probleminde, çoklu bağlantı sorunu gözardı edilebilir. Herhangi bir değişken, diğer değişkenlerin doğrusal birleşimi şeklinde yazılmamalıdır. Değişkenler arasında çoklu bağlantı problemi olduğunda, Lojistik regresyon analizinde katsayıların standart hatalarının yüksek çıkması ve modeli tahmin gücünün azalması gibi sorunlar ortaya çıkabilir. Böyle bir durum olduğu tespit edildiğinde çeşitli düzenlemelerin yapılması gereklidir.

- *Gözlem sayısı yeterli olmalıdır.*

Veri seti az sayıda gözlem içeriyorsa çoklu bağlantı problemi ortaya çıkma olasılığı yükselir. Ayrıca gözlem sayısı yeterli olmadığında tahminin güvenilirliği azalır.

Analizde tek bir bağımsız değişkene yer verildiğinde lojistik regresyon, birden fazla bağımsız değişkene yer verildiğinde ise çoklu lojistik regresyon söz konusudur. Öte yandan, bağımlı değişken sadece iki kategoriye sahip olduğunda ikili lojistik regresyon, buna karşılık sınıflayıcı ölçme düzeyinde ölçülmüş ikiden fazla kategoriye sahip olduğunda multinominal lojistik regresyondan söz edilir (Bayram, 2004, s.62).

4.1. Lojistik Regresyon Analizi Tarihçesi

İstatistiksel verileri sınıflandırmaya yönelik yöntemlerden biri olan Lojistik regresyon analizi son yıllarda sosyoekonomik araştırmalarda yoğun olarak kullanılmaktadır.

Analizin kullanımı çok eskiye dayanmamakla birlikte, ilk kez Berkson (1944) tarafından önerilmiştir. Berkson biyolojik deneylerin analizlerinde kullanmaya başlamıştır.

Cornfield (1962), lojistik regresyondaki katsayıların tahmininde ilk kez diskriminant fonksiyonunu kullanmıştır.

Cox (1970), lojistik regresyon modelini geliştirmeye yönelik çalışmalar yapmıştır.

Pregibon (1981), lojistik regresyon modelinde sapan ve bağımlı değişken üzerinde etkili olan gözlemler üzerinde araştırmalar yapmıştı.

Aranda- Ordaz (1981), gözlemlerin lojistik regresyon modeline uyumu ile ilgili çalışmalarda bulunmuşlardır.

Begg ve Gray (1984), çok gruplu lojistik modeller üzerinde araştırmalar yapmışlar ve bununla ilgili temel bilgiler vermişlerdir.

Carroll ve ark. (1984), ikili lojistik regresyon modellerinde bağımlı değişkenin hatalı sonuçları durumunda kullanılabilecek tahmin yöntemleri üzerinde durmuşlardır.

Bonney (1987), katsayıların tahmininde en çok olabilirlik yöntemini paket programlar vasıtasıyla uygulamıştır.

Robert ve ark. (1987), anket denemelerinde lojistik regresyon analizinin kullanılabileceğini göstermişlerdir.

Hosmer ve Lemesdow (1989), basit lojistik regresyon modelinin tanıtımı, katsayıların önemi ve tahmin seti, çoklu lojistik regresyon modeli, modelde yer alan katsayıların yorumlanması, uygun modeli kurma aşamaları, uyum iyiliği testleri, vaka-kontrol çalışmalarında lojistik regresyon modelinin kullanımı, cevap değişkeninin ikiden çok seviye içerdiği durumlar için kurulan lojistik regresyon modelinin analizi ve hayatta kalma verilerine lojistik regresyon analizinin uygulanmasına ilişkin temel teorik ve uygulamalı açıklamalarda bulunmuştur.

Cox ve Snell (1989), ikili yanıt değişkeninin özellikleri üzerinde durmuşlardır.

Duffy (1990), lojistik regresyonda hata terimleri ve katsayı değerleri ile gerçek değerler arasındaki ilişkiyi incelemiştir.

Hsu ve Leonard (1997), Lojistik regresyon fonksiyonlarında Bayes tahmin edicilerinin elde edilmesini araştırmışlardır.

Kaygın, Tazegül ve Yazarkan (2016), veri madenciliği ve lojistik regresyon analizlerinin sınıflama tahminini karşılaştırmışlardır.

4.2. Lojistik Regresyon Modeli

Doğrusal regresyon analizi model oluşturmada kullanılan bir yöntem olmakla birlikte, birçok istatistiksel yöntem ve yazılımlarında Lojistik regresyon analizi de sıklıkla uygulanmaktadır.

İki analiz arasındaki en önemli farklardan biri bağımlı değişken doğrusal regresyon analizinde sürekliyken, lojistik regresyon analizinde ikili ya da çokludur. Diğer bir fark doğrusal regresyon analizinde bağımlı değişken değeri tahmin edilirken, lojistik regresyon analizinde bağımlı değişkenin aldığı değerlerden birinin gerçekleşme olasılığı tahmin edilir.

Bu bölümde lojistik regresyon modeli oluşturulmadan önce doğrusal regresyon analizine kısaca değinilecektir.

Doğrusal regresyon modelinde y_i yanıt *(bağımlı)* değişkenini, x_i girdi *(bağımsız)* değişkenini gösterdiğinde, verilen bir x_i değeri için y_i'nin beklenen değeri, diğer bir deyişle koşullu olasılığı Eşitlik (4.1) ile belirtilir.

$$E\left(y_i \middle| x_{i1}, x_{i2}, \dots, x_{np}\right) = \sum_{k=0}^{p} \beta_k x_{ik} \tag{4.1}$$

Burada $i = 1,2, \dots, n$ iken, β_k ile belirtilen değer, modeldeki bağımsız değişkenlerin bağımlı değişkeni ne düzeyde açıkladığını ifade etmektedir. Bağımsız değişkenler için herhangi bir kısıt olmamakla birlikte bağımlı değişkenin sürekli olduğu varsayılır. Başka bir ifadeyle, bağımsız değişkenler $-\infty$ ile $+\infty$ arasında değerler alabildiğinden, bağımlı değişken $-\infty$ ile $+\infty$ arasındaki bütün değerleri alabilir. Ancak bağımlı değişken kategorik olduğunda bu kural bozulur.

Herhangi bir i'inci gözlem değeri Eşitlik (4.2) ile ifade edilir.

$$y_i = \sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i \tag{4.2}$$

İki kategoriye sahip veriler için koşullu ortalama en az 0, en fazla 1 olmaktadır.

$0 \leq E(y_i|x_i) \leq 1$. Bağımlı değişken iki kategoriye sahip olduğunda, $E(y_i|x_i)$ için bir model ortaya koymada pek çok birikimli dağılımdan yararlanılabilir (Bayram, 2004, s.62).

i'inci gözlemin 1 değerini alma olasılığı ya da bir olayın gerçekleşme olasılığı $P(y_i = 1)$, olayın gerçekleşmeme olasılığı ise $P(y_i = 0)$ ile belirtildiğinde beklenen değer Eşitlik (4.3) ile verilir;

$$E(y_i) = 0 \times P(y_i = 0) + 1 \times P(y_i = 1) = P$$

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik} \tag{4.3}$$

Sol tarafı $[0 - 1]$ aralığında olasılık değerleri alan bu denkleme *Doğrusal Olasılık Modeli* adı verilmektedir (Tatlıdil, 1996, s.290).

y_i'nin varyansı ise Eşitlik (4.4) olarak elde edilir.

$$Var(y_i) = E({y_i}^2) - [E(y_i)]^2 = P(y_i = 1)[1 - P(y_i = 1)] = P(1 - P) \tag{4.4}$$

Ayrıca büyük ve küçük x değerleri y_i'nin 0-1 aralığı dışına çıkmasına neden olabilecektir. Diğer taraftan hata terimi de var-yok gibi iki değer alabildiğinden normal dağılım değil binom dağılım gösterecek ve değişen varyansa sahip olacaktır (Agresti, 2002, s.120). Eğer bağımlı değişkenin ikili olduğu düşünülürse, Eşitlik (4.2)'nin hata terimleriyle denklemler Eşitlik (4.5)'teki gibi yazılabilir.

$$y_i = 1 \text{ ise } \textstyle\sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i = 1 \Rightarrow \varepsilon_i = 1 - \sum_{k=0}^{p} \beta_k x_{ik}$$

$$y_i = 0 \text{ ise } \textstyle\sum_{k=0}^{p} \beta_k x_{ik} + \varepsilon_i = 0 \Rightarrow \varepsilon_i = -\sum_{k=0}^{p} \beta_k x_{ik} \tag{4.5}$$

Hata terimlerinde beklenen değerin 0, varyansın ise değişken olması beklenmektedir. Buna göre,

$$E(\varepsilon_i) = P(y_i = 0)\left(-\sum_{k=0}^{p} \beta_k x_{ik}\right) + P(y_i = 1)\left(1 - \sum_{k=0}^{p} \beta_k x_{ik}\right)$$

$$= -[1 - P(y_i = 1)]\left(-\sum_{k=0}^{p} \beta_k x_{ik}\right) + P(y_i = 1)\left(1 - \sum_{k=0}^{p} \beta_k x_{ik}\right) = 0$$

$$Var(\varepsilon_i) = E({\varepsilon_i}^2) = P(y_i = 0)\left(-\sum_{k=0}^{p}\beta_k x_{ik}\right)^2 + P(y_i = 1)\left(1-\sum_{k=0}^{p}\beta_k x_{ik}\right)^2$$

$$= \left(\sum_{k=0}^{p}\beta_k x_{ik}\right) + \left(1-\sum_{k=0}^{p}\beta_k x_{ik}\right) \tag{4.6}$$

olacağından tahmin edilen β_k katsayıları yansız olacak, ancak hata terimlerinin sabit bir varyansı olmayacağından β_k katsayıları en iyi olmayacaktır. Varyansın sabit olmaması nedeniyle hipotez testleri ve güven aralıkları geçerliliklerini kaybedeceklerdir. Bu durumda Goldberger tarafından önerilen ağırlıklı tahmin yöntemi kullanılabilir. Varyansın sabit hale getirilmesi için gözlem değerleri

$$w_i = \left(\left[\sum_{k=0}^{p}\beta_k x_{ik}\right]\left[1-\sum_{k=0}^{p}\beta_k x_{ik}\right]\right)^{1/2} \tag{4.7}$$

Eşitlik (4.7) ile belirtilen katsayıya bölünür (Çinko, 2003, s.24).

Eşitlik (4.2)'de verilen model ağırlıklandırıldığında yeni değerler ile regresyon modeli oluşturulur ve bulunan yeni hata terimi sabit varyanslı ve yansız olacaktır.

4.3. İkili Sınıflandırmaya Dayalı Lojistik Regresyon Modeli

Bağımlı değişken iki kategorili olduğunda doğrusal olasılık modelinde çeşitli sorunlar ortaya çıkar. Bu sorunlar genellikle olasılıkların 0 ve 1 arasında değer almasına karşın doğrusal fonksiyonun $-\infty$ ile $+\infty$ arasında değer alabilmesinden kaynaklıdır. Bir başka deyişle olasılık değeri ve bağımsız değişkenler arasında doğrusal bir ilişki olmadığında sorunlar oluşur. Bu nedenle Eşitlik (4.3) her zaman sağlanamamaktadır. Böyle bir durumu engellemek için olasılık değerini $-\infty$ ile $+\infty$ arasında tanımlı hale getirmek en iyi çözüm olacaktır.

Bunun için geliştirilen ve en çok kullanılan uygulamalar Lojit ve Probit dönüşümleridir. Örneklemde yeterli sayıda gözlem var ise her iki yöntem de birbirine yakın sonuçlar vermektedir. Fakat matematiksel olarak kolay kullanıma sahip olması nedeniyle Eşitlik (4.3)'e lojit dönüşüm uygulanır.

İlk olarak olasılık modeli üzerinde $P_i/1-P_i$ dönüşümü yapılır. Bu dönüşüm olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı olarak ifade edilir. Bu oranın bağımsız değişkenlerle ilişkisi üsteldir. Başka bir ifadeyle, yanıt değişkenleri 0 ile $+\infty$ arasında değerler alır. Bundan dolayı kullanılan bu oranın doğal logaritması alınır ve yanıt

değişkenlerinin sınırları $-\infty$ ile $+\infty$ arasında tanımlanır. Böylece elde edilen modelin bağımsız değişkenlerle olan ilişkisi doğrusallaşmış olur. Uygulanan dönüşüm Eşitlik (4.8) ile gösterilebilir.

$$L_i = log\left(\frac{P_i}{1-P_i}\right) = \sum_{k=0}^{p} \beta_k x_{ik} \tag{4.8}$$

Buradaki $P_i/1-P_i$ oranına aynı zamanda *odds değeri* ya da *odds oranı* da denilmektedir. Odds orası iki grubun sahip oldukları olasılık değerlerinin birbirine olan oranıdır. Lojit dönüşüm P_i'nin 0-1 arasında kalmasını sağlamaktadır.

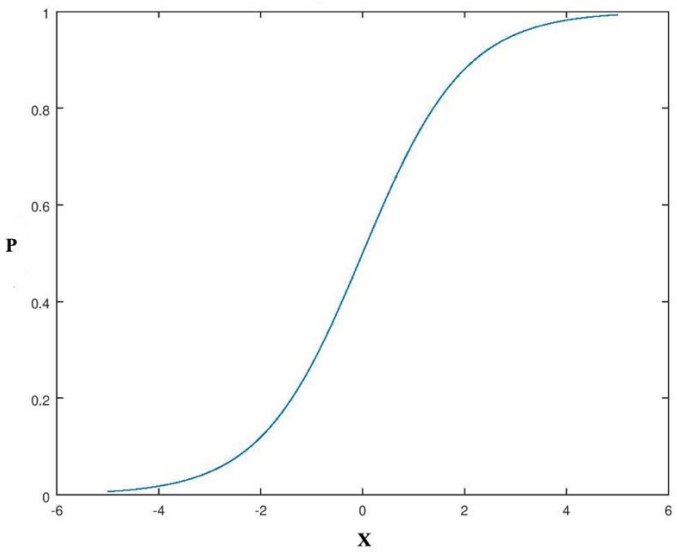
$P_i \to 0$'a yaklaştığında $L_i \to -\infty$'a, $P_i \to 1'$e yaklaştığında $L_i \to +\infty$'a doğru gitmektedir.

Tüm bu dönüşümlerden sonra elde edilen lojistik regresyon modeli Eşitlik (4.9) biçiminde tanımlanır.

$$P_i = \frac{e^{\sum_{k=0}^{p} \beta_k x_{ik}}}{1 + e^{\sum_{k=0}^{p} \beta_k x_{ik}}} \tag{4.9}$$

Bu fonksiyon aynı zamanda Eşitlik (4.10) ile de gösterilebilir.

$$P_i = \frac{1}{1 + e^{-\sum_{k=0}^{p} \beta_k x_{ik}}} \tag{4.10}$$



Şekil 4.1. *Lojistik Regresyon Eğrisi*

Şekil 4.1 de görüldüğü gibi, olasılıklar sıfırdan bire doğru büyürken lojistik fonksiyonu $-\infty$ ile $+\infty$ arasında değerler almaktadır.

Lojit dönüşümünün bazı özellikleri;

i. P, 0-1 arasında değerler alırken, *lojit* P $-\infty$ ile $+\infty$ arasında değişir.
ii. Lojit modeli, x'e göre doğrusal olmakla beraber olasılıkları doğrusal değildir (Albayrak, 2006, s.447).
iii. $P < 0{,}5$ olduğunda $logit\ P < 0$, $P \geq 0$ olduğunda ise $logit\ P > 0$ olur.
iv. P arttıkça $logit\ P$'de artar (Tatlıdil, 1996, s.293).

İkili lojistik regresyon modelinin varsayımları aşağıdaki gibidir.

- $y_i \in (0,1) \qquad i = 1,2,\dots,n$
- $P(y_i = 1/x_i) = P_i = \frac{e^{\sum_{k=0}^{p} \beta_k x_{ik}}}{1+e^{\sum_{k=0}^{p} \beta_k x_{ik}}}$
- $y_1, y_2, \dots, y_n$ istatistiksel olarak bağımsızdır.
- Bağımsız değişkenler birbirinden bağımsızdır.

Bağımsız değişkenlerin üzerinde herhangi bir kısıt olmamasından dolayı, lojistik regresyon modeli değişkenlerin sürekli ve kesikli olma durumlarına göre değişir.

4.4. Katsayı Tahmin Yöntemleri

Lojistik regresyon modelinin amaçlarından biri de her bir gözleme ait katsayı değişiminin, bağımlı değişkenin olasılık değeri üzerindeki etkisini belirleyebilmektir. Bu amaçla kullanılan çözüm yöntemleri En çok olabilirlik, yeniden ağırlıklandırılmış iteratif en küçük kareler ve minimum lojit ki-kare yöntemleridir.

4.4.1. En çok olabilirlik yöntemi

Doğrusal regresyon analizinde katsayıların tahmininde kullanılan birçok yöntem bulunmaktadır. Fakat doğrusal regresyon analizinin gerektirdiği varsayımlar bozulduğunda bu yöntemler iyi sonuçlar vermemektedir. Bu nedenle lojistik regresyon analizinde katsayı tahmini için çeşitli yöntemler geliştirilmiştir. Uygulamada en çok kullanılan yöntem en çok olabilirlik yöntemidir.

En çok olabilirlik yönteminin lojistik regresyon analizinde kullanılmasındaki en temel amaç gruplardaki gözlemlerin olabilirliğini maksimize edecek katsayıları en iyi

şekilde tahmin edebilmektir. Diğer bir deyişle, yöntemde logaritmik olasılık değerinin maksimize edilmesi amaçlanır. Bu amaç için öncelikle en çok olabilirlik fonksiyonu oluşturulur.

İki gruplu bağımlı değişken olması durumunda;

$(x_i, y_i), \quad i = 1,2, \dots, n$ ile gösterilen n tane gözlem olduğu varsayımı altında, $y_i = 1$ ise olabilirlik olasılığı P_i, $y_i = 0$ ise $1 - P_i$ olur ve i'inci gözlem için olasılık, Eşitlik (4.11) ile belirtilir.

$$P_i^{y_i}(1 - P_i)^{1-y_i} \tag{4.11}$$

Gözlemlerin birbirinden bağımsız ve aynı dağılıma sahip olduğu varsayımı sağlandığında, n tane sayıda gözlem için olabilirlik fonksiyonu Eşitlik (4.12)'deki gibi elde edilir.

$$\ell(\beta) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i} \tag{4.12}$$

En çok olabilirlik yöntemi Eşitlik (4.12)'yi maksimize eden β'ların tahmin edilmesini amaçlamaktadır. $\ln(\ell(\beta))$ fonksiyonu ℓ'ye göre artan bir fonksiyon olduğundan çoğunlukla $\ell(\beta)$ yerine $\ln(\ell(\beta))$ kullanılır.

$$\begin{aligned} \mathrm{L}(\beta) = \ln\big(\ell(\beta)\big) &= \sum_{i=1}^{n} \{y_i ln P_i + (1 - y_i)\ln(1 - P_i)\} \\ &= \sum_{i=1}^{n} ln 1 - P_i + \sum_{i=1}^{n} y_i \, ln(P_i/1 - P_i) \\ &= \sum_{i=1}^{n} -ln 1 + e^{\beta_0 + \beta_1 x_i} + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) \end{aligned} \tag{4.13}$$

Eşitlik (4.13)'e *logaritmik olabilirlik fonksiyonu* denir. L(β)'yı maksimum yapan $\beta_0$ve β_1 katsayılarına göre kısmi türev alınır.

$$\sum_{i=1}^{n} [y_i - P_i] x_i = 0$$

$$\sum_{i=1}^{n} [y_i - P_i] = 0 \tag{4.14}$$

Buradan eşitlik (4.14) ile gösterilen olabilirlik denklemleri elde edilir. P_i ile katsayılar arasında doğrusal bir ilişki bulunmadığından, lojistik regresyon analizi için bu ifadeler doğrusal olmayan denklemlerdir. Bu durumda, Eşitlik (4.14)'teki denklemler çözülerek β değerlerinin bulunabilmesi için özel iterasyon gerektiren yöntemler kullanılır. Bu yöntemler amaçları gereği tekrarlıdır. İlgili β değerlerine bir başlangıç değeri verilir ve elde edilen ilk tahminden, δ kadar eksiltme veya arttırma yapılarak türev alınır. Bu süreç sonuca ulaşıncaya kadar tekrar edilir. Gözlemler birbirinden çok ayrık olduğunda herhangi bir sonuca varılamaz. Burada esas ilgi katsayı tahminlerinin yorumundan ziyade, Eşitlik (4.9)'da verilen modelde yer alan sınıf olasılıklarının tahmini ve verilen bir x değişkenine göre yeni bir gözlemin sınıflandırılmasıdır (Güneri ve Aydın, 2017, s.50).

4.4.2. Yeniden ağırlıklandırılmış iteratif en küçük kareler yöntemi

Sınıflandırılmış verilerde her bir bağımsız değişkenin aldığı değerler için olayın gerçekleşme olasılığı $P_j = {}^{n_j}/_{N_j}$ olur. Yanıt değişkeninin ikili değerler aldığı durumda gözlemler binom dağılımlı olduğundan değişen varyanslılık söz konusudur. Bu nedenle lojit P_j değerinin bağımsız değişkenler üzerindeki ağırlık değerlerinden yararlanarak ağırlıklandırılmış regresyon modeli kullanılır. Buradaki ağırlık değeri Eşitlik (4.7)'den yola çıkılarak,

$$w_j = \left[N_j P_j \left(1 - P_j\right)\right]^{1/2} \tag{4.15}$$

olarak alınır. Ancak w_j değerleri P_j olasılığının bir fonksiyonu olduğu için, en küçük kareler yöntemi yinelemeli olarak uygulanabilecektir. Ayrıca ağırlıklandırılmış modelde hata terimi $w_j \varepsilon_j$'nin varyansı sabit olacağından, hata teriminin tahmini en küçük kareler yöntemiyle yapılabilir.

4.4.3. Minimum lojit ki-kare yöntemi

Berkson (1955), tarafından geliştirilen ve gözlemlerin tekrar etmesi durumunda kullanılan bu yöntem, ağırlıklandırılmış en küçük kareler yönteminin özel bir halidir.

Yöntemde $(2 \times j)$ çapraz tablolarındaki beklenen ve gözlenen lojit değerler arasındaki fark kullanılır. Ağırlıklandırılmış en küçük kareler yönteminde tanımlanan ve üzerinde lojit dönüşüm yapılan olasılık değeri yanıt değişkeni olarak alınır.

Yöntem, lojit değeri olarak tanımlanan yanıt değişkeninin bağımsız değişkenler üzerindeki ağırlık değeri ile ağırlıklandırılmış regresyonundan en küçük kareler tahminlerini elde etmeye dayanmaktadır. Buradan tek adımda bulunan ağırlıklı en küçük kareler tahminleri *minimum lojit ki-kare* adını almaktadır (Şensoy, 2009, s.11).

4.5. Lojistik Regresyon Modelinin Uyum İyiliği ve Katsayı Testleri

Katsayıların tahmin edilmesinden sonra oluşturulan modelin gözlemlerle olan uyumunun değerlendirilmesi oldukça önemlidir. Lojistik regresyon modelinin uyum iyiliğini test etmek için belli teknikler kullanılır. Uyum iyiliği testinde, modelin yanıt değişkenini tanımlamakta ne kadar etkili olduğu test edilir.

Lojistik regresyon modelinin normallik varsayımı olmadığından uyum iyiliği testlerinde t ve F testleri gibi parametrik testler kullanılmamaktadır. Bunların yerine çoğunlukla parametrik olmayan testler kullanılmaktadır (Kara, 2015).

4.5.1. Pearson Ki-kare istatistiği ve sapma ölçütü

Lojistik regresyon modelinin genel anlamlılığının test edilmesinde kullanılan en fazla bilinen parametrik olmayan ölçüttür.

Pearson Ki-kare istatistiği Eşitlik (4.16) biçiminde belirtilir.

$$\chi^2 = \sum_{i=1}^{J} \frac{\left(y_i - n_i \hat{P}_i\right)^2}{n_i P_i \left(1 - \hat{P}_i\right)} \tag{4.16}$$

Burada $\hat{P}_i$; tahmin edilen olasılık değeri, n_i ise i-inci grubun toplam gözlem değeridir.

Tahmin edilen model, model üzerinde sadece etkili anlamlı değişkenleri içeren model, doymuş model ise değişken sayısı kadar parametre içeren model olarak alındığında, sapma ölçütü Eşitlik (4.17) biçiminde tanımlanır.

$$D = -2ln\left(\frac{Tahmin\ edilen\ modelin\ olabilirliği}{Doymuş\ modelin\ olabilirliği}\right) \tag{4.17}$$

Eşitlik (4.17)'de kullanılan kesirli ifadeye *olabilirlik oranı* denilmektedir. Bu eşitlik üzerinde logaritmik olabilirlik fonksiyonunun yazılması ile Eşitlik (4.18) elde edilir.

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{P}_i}{y_i}\right) + (1 - y_i) ln\left(\frac{1 - \hat{P}_i}{1 - y_i}\right)\right] \tag{4.18}$$

Modeldeki parametre sayısı k olduğundan, sapma $n-k$ serbestlik dereceli ki-kare tablo değeriyle karşılaştırılır. Sapma, ki-kare dağılımlı olduğundan doğrusal regresyon analizindeki hata kareleri toplamına benzer ve uyum iyiliğine karar vermede kullanılır.

Lojistik model verileri tam olarak temsil ediyorsa, diğer bir ifadeyle, modelin uyum iyiliği mükemmel ise olabilirlik oranı 1 ve D istatistiği 0 değerini alır. Modelin doymuş modelle tam olarak uyum sağlaması durumunda bu sonuca varılır. Dolayısıyla, modeldeki D istatistiği minimum olduğunda en iyi model ortaya çıkar.

4.5.2. Hosmer-Lemeshow G istatistiği

G istatistiği lojistik modelin genel anlamlılığını test eden bir ölçüttür ve gözlenen ile tahmin edilen değerler arasında fark olmadığı hipotezini araştırır.

Hosmer ve Lemeshow, tahmin edilen olasılık değerlerinin gruplandırılmasını önermişlerdir. Burada gruplama yapmaktaki amaç; daha düzgün tahmin edilen değerler oluşturulmasıyla mevcut dağılımı ki-kare dağılımına yaklaştırarak anlamlı ve yorumlanabilir bir model elde etmektir (Şensoy, 2009, s.27).

Hosmer ve Lemeshow tarafından uyum iyiliğini test etmek için geliştirilmiş birçok istatistik bulunmaktadır. Ancak geliştirilen istatistiklerden sadece iki yöntem kullanılmaktadır. Bu yöntemlerden ilki, tahmin edilen olasılıkları dikkate alarak gözlemleri 10 gruba ayırır ve dağılımın ki-kare dağılımına yakınsadığı düşünülür. g, grup sayısı olarak tanımlandığında $G^2 \sim \chi^2_{(g-2)}$ olarak alınır.

Diğer yöntemde ise gözlemler belirlenen kesim noktasına göre gruplandırılır. Tahmin edilen olasılıklara göre gruplar oluşturulur ve gözlemlerin bu gruplarda yer alması sağlanır. Gruplara ayrılan gözlemlere daha sonra Pearson ki-kare uyum iyiliği testi uygulanır. Böylece tahmin edilen ve gözlenen değerler karşılaştırılır. Test sonucunun anlamlı bulunmaması, modelin veriye iyi uyduğunu gösterir. G istatistiği, yeterli gözlem sayısına sahip olunduğunda uygulanabilir. Eğer örneklem büyüklüğü yeterli değilse test sonucu anlamsızlaşır.

Hosmer ve Lemeshow istatistiği doğrusal regresyon analizindeki F istatistiği ile aynı amaca yönelik olarak kullanılmaktadır.

4.5.3. R^2 istatistikleri

Lojistik regresyon modellerin uyum iyiliği testinde olabilirlik oranına ve hata kareler toplamına dayalı R^2 ölçütleri de kullanılmaktadır. Bu ölçütler, doğrusal regresyon analizinde kullanılan R^2 belirtme katsayısına benzetilse de, modelin gücünü açıklamada çoğunlukla onun gibi yorumlanmamaktadır.

Modelin uyum iyiliğini belirlemede kullanılan ölçütlerden birincisi Cox ve Snell R^2 istatistiğidir. Olabilirlik esasına göre çok değişkenli regresyon analizinde kullanılan R^2 istatistiğine benzemektedir.

$$R^2{}_{c\&s} = 1 - \left(\frac{L_0}{L}\right)^{2/n} \tag{4.19}$$

Eşitlik (4.19)'la hesaplanan Cox ve Snell R^2 istatistiğinde n gözlem sayısını belirtir. L_0 sadece sabit içeren modelin olabilirliğini, L ise bağımsız değişkenlerin tümünü içeren modelin olabilirliğini göstermektedir. Eşitlik (4.19)'dan anlaşılacağı gibi Cox ve Snell R^2 istatistiğinin maksimum değeri 1'den küçük bir değer alır. Bu nedenle sonucun yorumlanması güçleşmektedir.

Diğer bir R^2 istatistiği ise Nargelkarke R^2 istatistiğidir. Cox ve Snell R^2 istatistiğinin dezavantajını ortadan kaldırmak için geliştirilmiştir. Diğer bir deyişle, Cox ve Snell R^2 istatistiğinin 0-1 arasında değerler alacak şekilde geliştirilmesiyle elde edilmiştir.

$$R^2{}_N = \frac{1 - \left(\frac{L_0}{L}\right)^{2/n}}{1 - (L_0)^{2/n}} \tag{4.20}$$

Eşitlik (4.20)'nin 0-1 arasında değerler alıyor olması modelin yorumlanmasını kolaylaştırır. Nargelkarke R^2 istatistiğinin 1'e eşit olması, değişkenler arasındaki uyumun iyi olduğunu ve tüm modelin sonucu iyi bir şekilde tahmin ettiğini göstermektedir.

Bir diğer ölçüt, Mcfadden R^2 istatistiğidir. Burada sadece sabit içeren modelin olabilirliği $-2lnL_0$ ve tüm bağımsız değişkenleri içeren modelin olabilirliği $-2lnL$ olarak alınır. Bu durumda Mcfadden R^2 istatistiği Eşitlik (4.21) ile hesaplanır.

$$R^2{}_{McF} = 1 - \left(\frac{-2lnL}{-2lnL_0}\right) = 1 - \left(\frac{lnL}{lnL_0}\right) \tag{4.21}$$

Eşitlikte lnL ile gösterilen ve tüm bağımsız değişkenleri içeren modelin olabilirliği hata kareler toplamına, lnL_0 ile gösterilen ve sadece sabit içeren modelin olabilirliği ise kareler toplamına karşılık gelmektedir. (Kara, 2015, s.47)

4.5.4. Sınıflama Tabloları

Lojistik regresyon analizi gözlemlerin sınıflandırılması amacıyla kullanıldığından, modelin uyum iyiliği ölçütlerinden biri de sınıflandırma oranıdır. En doğru sınıflandırma oranını belirlemek ve model sonuçlarını özetlemek için sınıflama tablolarından yararlanılır. Sınıflama tabloları gruplardaki gözlem sayısına duyarlıdır. Gözlem sayısı ne kadar büyükse doğru sınıflandırma olasılığı o kadar artmaktadır.

Sınıflama tabloları, tahmin edilen olasılıklardan türetilen değerlerin çapraz sınıflama yapılarak gruplara ayrılması ilkesine dayanmaktadır. Yanıt değişkenin gözlenen gerçek değerleri ile tahmin edilen değerleri çapraz sınıflama yapılarak gruplara ayrılır (Kara, 2015, s.48).

Gruplama işleminin yapılabilmesi için ilk olarak bir kesim noktası belirlenir ve tahmin edilen her olasılık değeri bu kesim noktası ile karşılaştırılır. Eğer tahmin edilen olasılık değeri kesim noktasından küçük ise $y = 0$, büyük ise $y = 1$ atanır. Kesim noktasının değeri çoğunlukla 0,5 olarak alınır.

4.6. Model Katsayılarının Anlamlılığının Test Edilmesi

Lojistik Regresyon analizinde, doğrusal regresyon analizinde olduğu gibi, modelin bütün olarak anlamlılığının test edilmesinin ardından, sabit terim ve bağımsız değişkenlerin de tek tek anlamlılığının test edilmesi gerekmektedir (Börüban, 2009, s.47).

Modeldeki değişkenlerin katsayılarının tahmin edilmesinin ardından, bağımsız değişkenlerin yanıt değişkeniyle anlamlı bir ilişkisi olup olmadığı belirlenir. Bağımsız değişken modelde olduğunda ve olmadığında elde edilen tahmin değerleriyle bağımlı değişkene ait gözlenen değerler karşılaştırılır ve bunun sonucunda ilgili değişkeni içeren model, içermeyen modelden daha doğru ise o değişkenin anlamlı olduğu sonucuna varılır.

Lojistik modele dahil edilmesi düşünülen bağımsız değişkene ait katsayının anlamlılığını test etmek için kullanılan temel testler aşağıdaki gibi verilmektedir;

- Olabilirlik Oran Testi,
- Wald Testi,
- Skor Testi.

4.6.1. Olabilirlik oran testi

Lojistik regresyon modelinin uyum iyiliğinin değerlendirilmesinde önemli bir ölçüt olan D istatistiği, bağımsız değişkenin önemini değerlendirmede de kullanılır. Modelde bağımsız değişkenin olduğu ve olmadığı eşitliklerin D değerleri karşılaştırılır.

$$G = D_{değişken\ içermeyen\ model} - D_{değişken\ içeren\ model} \tag{4.22}$$

Dolayısıyla Eşitlik (4.22) kullanılarak bağımsız değişkenin istatistiksel anlamlılığı test edilmiş olur.

Bulunan G istatistiği doğrusal regresyonda kullanılan F testindeki pay kısmı ile benzer özellik göstermektedir. Bu istatistiği hesaplamak için kullanılan her iki D değeri için de doymuş modelin olabilirliği aynıdır. Bu durumda G, Eşitlik (4.17) ele alındığında, Eşitlik (4.23)'teki gibi bulunur.

$$G = -2ln\left(\frac{Değişken\ içermeyen\ modelin\ olabilirliği}{Değişken\ içeren\ modelin\ olabilirliği}\right) \tag{4.23}$$

Bu teste *olabilirlik oran testi* adı verilir ve asimptotik olarak ki-kare dağılımına sahiptir. Bu testte;

$H_0: \beta_i = 0$

$H_1: \beta_i \neq 0$

hipotezleri test edilmektedir. χ^2 dağılımı için serbestlik derecesi ise iki modelde de tahmin edilen parametre sayıları arasındaki farka eşittir. H_0 reddedilir ise modele eklenen bağımsız değişkenin modelin tahmininde önemli olduğu sonucuna varılır.

4.6.2. Wald testi

Wald istatistiği doğrusal regresyonda kullanılan t istatistiğinin karesine benzemektedir. Wald istatistiği lojistik regresyon modeli oluşturulurken, ilgili değişkenin tahmin edilen β katsayısının en çok olabilirlik değerinin, β'nın standart hata tahminine bölünmesi ile Eşitlik (4.24)'teki gibi elde edilir.

Bu oran, $H_0: \beta_i = 0$ hipotezi altında standart normal dağılım göstermektedir.

$$W = \frac{\widehat{\beta_\iota}}{\widehat{SH}(\widehat{\beta_\iota})} \tag{4.24}$$

Bulunan W istatistiği tablo değerinden büyük ise $H_0: \beta_i = 0$ hipotezi reddedilir ve ilgili katsayının anlamlı olduğu sonucuna varılır.

W testi daha farklı bir biçimde de daha yazılabilir. Çünkü, standart normal dağılıma sahip W istatistiğinin karesinin alınması sonucu elde edilen W^2 istatistiği 1 serbestlik derecesiyle χ^2 dağılır. Bu durumda Eşitlik (4.25) elde edilir.

$$W^2 = \left(\frac{\widehat{\beta_\iota}}{\widehat{SH}(\widehat{\beta_\iota})}\right)^2 \sim \chi^2_{(1-a,1)} \tag{4.25}$$

(Vupa, 2004, s.16).

Wald istatistiğinin uygulanmasının kolay olmasına rağmen güvenirliliği sorgulanmaktadır. W testi büyük örneklem durumunda daha iyi sonuçlar vermekle birlikte küçük örneklemlerde hatalı sonuçlar üretebilmektedir. Wald istatistiğinin diğer bir dezavantajı da; lojistik regresyon katsayıları mutlak olarak büyüdükçe standart hataların tahmini de büyür. Bu da daha küçük W istatistiği demektir. W değerinin küçük çıkması H_0 hipotezinin reddedilmesi gerektiği halde kabul edilmesine ve bağımsız değişkenin yanlışlıkla modelde kalmasına neden olabilir.

Küçük örneklemlerde ve katsayılar mutlak olarak büyüdüğünde olabilirlik oran testi genellikle daha güçlü ve daha güvenilir sonuçlar vermektedir (Agresti, 1996, s.109).

4.6.3. Skor testi

Olabilirlik oran ve Wald istatistiklerinin tersine Skor testinde katsayıların en çok olabilirlik tahmininin hesaplanması gerekmemektedir.

Skor testi için test istatistiği Eşitlik (4.26)'da verildiği gibidir.

$$ST = \frac{\sum_{i=1}^{n} x_i\,(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{4.26}$$

Bu eşitlikten de anlaşılacağı üzere, yapılan test olabilirlik denklemlerine dayanır ve test istatistiği Wald istatistiği gibi standart normal dağılım gösterir.

Skor testinde yapılan hesaplamalar diğer testlere göre daha kolay olmasına rağmen, paket programlarda yer almadığı için pek tercih edilmemektedir.

4.7. Çok Gruplu Lojistik Regresyon Modeli

Çok gruplu lojistik regresyon modeli, ikiden fazla yanıt değişkeni bulunan lojistik regresyon modelinin genelleştirilmesiyle oluşturulur. Diğer çok değişkenli regresyon modelleriyle benzer özellikler taşıyan çoklu lojistik regresyon modelinin en belirgin farkı katsayıların tahmin edilmesi ve yorumlanmasındadır. Katsayıların ait oldukları bağımsız değişkenler sürekli olmayıp, kesikli ve sınıflayıcı ölçekle ölçülmüş değişkenler olabilir. Bu tür değişkenleri modele dahil etmek için tasarım değişkenler kullanılır.

Bir başka deyişle çok gruplu lojistik regresyon modeli, ikiden fazla yanıt değişkeninin ikiden fazla grup içerdiği ve değerlerinin sınıflayıcı ölçekle elde edildiği durumlarda yanıt değişkenleri ve bağımsız değişkenler arasındaki ilişkiyi ortaya koyan bir modeldir.

Çok gruplu lojistik regresyon modeli, Eşitlik (4.27) ile ifade edilir.

$$P(y=j)=\frac{e^{\sum_{k=1}^{K}\beta_{jk}x_k}}{1+\sum_{j=1}^{J-1}e^{\sum_{k=1}^{K}\beta_{jk}x_k}} \qquad j=1,2,\dots,J-1 \tag{4.27}$$

Burada $P(y=j)$; yanıt değişkeninin j'inci gruba sınıflanması olasılığını belirtmektedir. Modelde yer alan β_{jk} katsayısındaki k indisi bağımsız değişkeni belirtirken, j indisi yanıt değişkeninin grubunu belirtmektedir (Liao, 1994, s.48). Sınıflayıcı ölçekle ölçülmüş J gruplu yanıt değişkenine sahip olan çoklu lojistik regresyon modelinin, $J-1$ adet lojistik regresyon modeline sahip olması gerekir.

Yanıt değişkeni ikiden çok gruplu olan modellerde çözümleme yapmak için bir referans grubu tanımlanır. J referans grubu olarak seçilirse, yanıt değişkeninin referans grubuna düşme olasılığı Eşitlik (4.28) yardımıyla hesaplanabilir.

$$P(y=J)=\frac{1}{1+\sum_{j=1}^{J-1}e^{\sum_{k=1}^{K}\beta_{jk}x_k}} \qquad j=1,2,\dots,J-1 \tag{4.28}$$

Bunun yerine diğer olasılıklar bilindiğinde, bu olasılıklar yardımıyla Eşitlik (4.29) ile de hesaplanabilir

$$P(y = J) = 1 - [P(y = 1) + P(y = 2) + \cdots + P(y = J - 1)] \qquad (4.29)$$

(Liao, 1994, s.49).

İki gruplu lojistik modeller, çok gruplu lojistik regresyon modellerinde de kullanılabilir. Çok gruplu modellerde lojit dönüşüm referans grubu seçilerek odds oranlarının logaritması alınarak yapılır. 0, 1 ve 2 gibi yanıt değişkeninin üç gruplu olduğu bir örnek için $y = 0$ grubu referans grubu olarak seçildiğinde lojit dönüşümleri Eşitlik (4.30) ve (4.31) ile elde edilmektedir.

$$ln\left[\frac{P(y = 1|x_1)}{P(y = 0|x_1)}\right] = \beta_{10} + \beta_{11}x_1 \qquad (4.30)$$

$$ln\left[\frac{P(y = 1|x_2)}{P(y = 0|x_2)}\right] = \beta_{20} + \beta_{21}x_1 \qquad (4.31)$$

Böylece çok gruplu lojistik modeli 0 ile 1 ve 0 ile 2'den oluşan iki farklı lojistik modele sahiptir. Dolayısıyla yanıt değişkeni üç gruba sahip olan bir model için 2 adet odds oranı elde edilir ve bu oranlar ile her bir grup için karşılaştırma yapılır. Eşitlik (4.30) ve (4.31)'de görüldüğü gibi, odds oranlarının logaritmaları alınarak lojistik modeller elde edilir ve model doğrusal hale getirilir.

Üç grup olması durumunda yanıt değişkenleri için koşullu olasılıklar, verilen bilgiler ile Eşitlik (4.32) biçiminde genelleştirilebilir.

$$ln\left[\frac{P(y = j)}{P(y = J)}\right] = \sum_{k=1}^{K} \beta_{jk}x_k \qquad j = 1,2, \dots, J - 1 \qquad (4.32)$$

Olabilirlik fonksiyonunu oluşturmak için grup üyeliğini belirlemede üç adet iki değerli değişkenden yararlanılmaktadır. Bu değişkenler;

$y = 0$ için $y_0 = 1, y_1 = 0, y_2 = 0$

$y = 1$ için $y_0 = 0, y_1 = 1, y_2 = 0$

$y = 2$ için $y_0 = 0, y_1 = 0, y_2 = 1$

olmak üzere tasarım değişkenler biçiminde kodlanmaktadır. y'nin tüm değerleri için $\sum_j y_i = 1$ iken n bağımsız gözlemli örneklem için koşullu olabilirlik fonksiyonu Eşitlik (4.33) biçimindedir

$$L(\beta) = \prod_{i=1}^{n} P_0(x_i)^{y_{0i}} P_1(x_i)^{y_{1i}} P_2(x_i)^{y_{2i}} \quad (4.33)$$

(Tatlıdil, 1996, s.304).

4.7.1. Çok gruplu lojistik regresyon modelinde tahmin yöntemleri

İki gruplu lojistik regresyon modellerinde olduğu gibi çok gruplu lojistik regresyon modellerinde de katsayıların tahmin edilmesi uygun modelin elde edilmesi için önemlidir. Bu amaca yönelik çok gruplu lojistik regresyon modellerinde kullanılan yöntemler, iki gruplu lojistik regresyon modellerde kullanılan yöntemlerin genelleştirilmiş halidir.

$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_p})$ iken çok değişkenli durumda, katsayıların tahmini için en çok olabilirlik yöntemi kullanılır. $(\tilde{x}_i, y_i)$, $i = 1,2,..,n$; n bağımsız gözlemli örneklem olarak ifade edilir (Vupa, 2004, s.17).

$\widetilde{\boldsymbol{x}} = (\boldsymbol{x_1}, \boldsymbol{x_1}, \dots, \boldsymbol{x_p})$ bağımsız değişkenlerin oluşturduğu matris iken olabilirlik modeli üzerinde yapılan lojit dönüşümle çok gruplu lojistik regresyon modeli Eşitlik (4.34) ve (4.35)'teki gibi ifade edilebilir

$$P(\tilde{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} = \frac{\exp(g(\tilde{x}))}{1 + \exp(g(\tilde{x}))} \quad (4.34)$$

$$g(\tilde{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (4.35)$$

(Vupa, 2004, s.17).

Böylece Eşitlik (4.12)'den yararlanılarak çok gruplu lojistik modelinin en çok olabilirlik fonksiyonu Eşitlik (4.36) ile verilir.

$$\ell(\tilde{\beta}) = \prod_{i=1}^{n} P(\tilde{x}_i)^{y_i} (1 - P(\tilde{x}_i))^{1-y_i} \quad (4.36)$$

Logaritmik olabilirlik fonksiyonunun, $p + 1$ adet katsayıya göre kısmi türevlerinin alınmasıyla, Eşitlik (4.37)'de verilen olabilirlik denklemleri elde edilir.

$$\sum_{i=1}^{n} [y_i - P(\tilde{x}_i)] = 0$$

$$\sum_{i=1}^{n} [y_i - P(\tilde{x}_i)] x_{ij} = 0 \qquad j = 1,2,\dots,p \quad (4.37)$$

$\hat{\beta}$, bu olabilirlik denklemlerinin sonuçlarını göstersin. Dolayısıyla çoklu lojistik regresyon modeli için uygun değerler, Eşitlik (4.34)'teki ifadede $\hat{\beta}$ ve x_i kullanılarak $\hat{P}(\tilde{x}_i)$ biçiminde bulunur (Ürük, 2007).

Oluşan yeni çok gruplu lojistik regresyon modeli Eşitlik (4.38) ile ifade edilebilir.

$$\hat{P}(\tilde{x}_i) = \frac{\exp(\hat{g}(\tilde{x}))}{1 + e\exp(\hat{g}(\tilde{x}))} \tag{4.38}$$

Çok gruplu lojistik modelin logaritmik olabilirlik fonksiyonunun ikinci kısmi türevleri, en çok olabilirlik tahmin edicisinin varyans kovaryans matrisini ve bilgi matrisini bulmaya yardımcı olur. $P(\tilde{x}_i) = \tilde{P}_i$ olmak üzere bu kısmi türevler alındığında Eşitlik (4.39) ve (4.40) ortaya çıkar.

$$\frac{\partial^2 L(\tilde{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \tilde{P}_i\left(1 - \tilde{P}_i\right) \tag{4.39}$$

$$\frac{\partial^2 L(\tilde{\beta})}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu}\, \tilde{P}_i\left(1 - \tilde{P}_i\right) \quad j, u = 0,1, \dots, p \tag{4.40}$$

$(p+1) \times (p+1)$ boyutlu matris, bu eşitliklerden elde edilen değerlerin negatiflerini içerir ve $I(\beta)$ ile gösterilir. Bu matrise *gözlemlenen bilgi matrisi* adı verilir. Tahmin edilen katsayıların varyans ve kovaryansları ise gözlemlenen bilgi matrisinin tersidir ve $Var(\beta) = I(\beta)^{-1}$ biçiminde gösterilir (Hosmer ve Lemeshow, 2000, s.34). Matrisin j. köşegen elemanı $Var(\beta_j)$'dir. Köşegen dışındaki matris elemanları ise Cov (β_j, β_u) ile gösterilir. Varyans ve kovaryansların tahminleri olan $Var(\beta)$, $\hat{\beta}$ kullanılarak bulunur (Şensoy, 2009, s.15).

Tahmin edilen katsayıların standart hata tahminlerinin kullanılması gerekmektedir. Bu standart hata tahminleri Eşitlik (4.41)'de verilmektedir

$$\widehat{SH}(\hat{\beta}_j) = \left[\widehat{Var}(\hat{\beta}_j)\right]^{1/2}, \quad j = 0,1, \dots, p \tag{4.41}$$

(Ürük, 2007).

Tahmin edilen modelin uygunluğunun değerlendirilmesinde kullanılan bilgi matrisi $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}) = [\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}]_{(p+1)\times(p+1)}$ biçiminde elde edilir. Buradaki $\boldsymbol{X}$ ve $\boldsymbol{V}$ matrisleri sırasıyla Eşitlik (4.42) ve (4.43)'te verildiği gibidir.

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}_{n\times(p+1)} \tag{4.42}$$

$$\boldsymbol{V} = \begin{bmatrix} \hat{P}_1(1-\hat{P}_1) & 0 & \dots & 0 \\ 0 & \hat{P}_2(1-\hat{P}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{P}_n(1-\hat{P}_n) \end{bmatrix}_{n\times n} \tag{4.43}$$

4.7.2. Çok gruplu lojistik regresyon modelinde katsayıların anlamlılık testleri

Lojistik regresyon analizinde katsayıların anlamlılıkları için kullanılan testler ve uyum iyiliği ölçütleri çok gruplu lojistik regresyon analizinde de kullanılır.

Anlamlılık testlerinde, olabilirlik oran testinin çok gruplu lojistik regresyon modeli için genelleştirilmiş hali kullanılmaktadır. Olabilirlik oran testi G istatistiğine dayanır. Aralarındaki fark ise çoklu lojistik regresyonda $p+1$ katsayılı modelde $\hat{\hat{\beta}}$ vektörünün $\hat{\hat{P}}$ değeridir.

$$G = -2ln\left(\frac{kısaltılmış\ modelin\ olabilirliği}{tüm\ modelin\ olabilirliği}\right) \tag{4.44}$$

Burada test edilecek hipotezler,

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$

$H_1: En\ az\ bir\ \beta_p \neq 0$

biçiminde gösterilebilir. p adet *eğim katsayısı*nın sıfıra eşit olması hipotezi altında G istatistiği p serbestlik derecesiyle ki-kare dağılımına sahiptir. İlgili örneklem yeterince büyük olduğunda, H_0 hipotezi doğru ise G istatistiği ${\chi^2}_{(1-a,(s_2-s_1))}$ dağılımına sahip olur.

Dolayısıyla bu istatistik, ${\chi^2}_{(1-a,(s_2-s_1))}$ değerinden büyük ise H_0 hipotezi reddedilir. H_0 hipotezi reddedildiğinde en az bir veya birden fazla katsayının anlamlı olduğu sonucuna varılır.

Çok gruplu regresyon modelinde katsayıların anlamlı olup olmadığına karar vermeden önce kullanılan bir diğer istatistik de Wald test istatistiğidir. Bu istatistik, ikili

lojistik regresyon modelinde kullanılan Wald istatistiği ile aynı işlevi görür ve standart normal dağılıma yakınsar. Bu test istatistiği Eşitlik (4.45)'deki gibi elde edilir.

$$W = \widehat{\widehat{\boldsymbol{\beta}}'} \left[\widehat{\boldsymbol{Var}}\,(\widehat{\widehat{\boldsymbol{\beta}}})\right]^{-1} \widehat{\widehat{\boldsymbol{\beta}}}$$

$$= \widehat{\widehat{\boldsymbol{\beta}}'}[\boldsymbol{X'VX}]\widehat{\widehat{\boldsymbol{\beta}}} \qquad (4.45)$$

Burada W, H_0 hipotezi altında $\chi^2_{(p+1)}$ dağılımına sahiptir. $\hat{\beta}_0$ katsayısının analizden, dolayısıyla $\boldsymbol{X'VX}$ matrisinin satır ve sütunundan çıkarılmasıyla W test istatistiği p serbestlik derecesiyle χ^2 dağılımına sahip olur. Bu testin değerlendirilmesinde, yukarıda verilen vektör-matris işlemlerinin gerçekleştirilmesi ve $\widehat{\widehat{\boldsymbol{\beta}}}$'nın elde edilmesi gerektiğinden, modelin anlamlılığı için gerçekleştirilen olabilirlik oran testinin bir yararı yoktur (Homer ve Lemeshow, 2000, s.39).

İki gruplu lojistik regresyon modelinde kullanılan diğer bir test de Skor testidir. $L(\tilde{\beta})$'nın $\tilde{\beta}$' ya göre p adet türevinin koşullu dağılımı dikkate alınarak hesaplanır (Vupa, 2004, s.25) Bu testte Wald testi gibi çok gruplu modelin anlamlılığında kullanışlı değildir. Bu nedenle çok gruplu modelde olabilirlik oran testi kullanılması tercih edilir.

5. UYGULAMA

Diskriminant analizi ve lojistik regresyon analizi tekniklerinin sınıflandırma başarılarının karşılaştırılması amacıyla, kullanılan veri seti www.kaggle.com adlı web sitesinin veri tabanından sağlanmıştır **(http-6)**. Kaggle, veri analizi alanında çalışan kişilerin oluşturduğu bir veri bilimi platformudur. İstatistiksel analiz tekniklerinde kullanılabilecek çok sayıda veri seti bu sitede kullanıcılara açık kaynak olarak sunulmaktadır.

Bu çalışmada, Pima yerlilerinden (orta ve güney Arizona bölgesinde yaşayan bir grup Amerikan Yerlisi) 21-81 yaş aralığındaki 768 kadın diyabet hastasına ilişkin veriler incelenmiştir. Bu veri setinin kullanıcılara sunulma amacı, bir hastanın belirli teşhis ölçümlerine dayanarak diyabet hastası olup olmadığını araştırmaktır. Orijinal verilere *Ulusal Diyabet, Sindirim ve Böbrek Hastalıkları Enstitüsü* (*NIDDK*)'den de ulaşılabilmektedir. Çalışmamızın amacı doğrultusunda iki gruplu olarak verilen diyabet yanıt değişkenini etkileyen 8 tane bağımsız değişken ele alınmıştır. Tablo 5.1'de ele alınan tüm değişkenlerin listesi görülmektedir.

Tablo 5.1. *Veri Setine Ait Değişkenler Listesi.*

	Değişkenler	**Açıklama**
y	Diyabet	0: Diyabet hastalığı yok, 1: Diyabet hastalığı var
x_1	Gebelik (GB)	Yaşanılan gebelik sayısı
x_2	Glukoz (GL)	2 saatlik plazma glukoz konsantrasyonu
x_3	Kan Basıncı (mm Hg) (KB)	Diastolik kan basıncı (Küçük tansiyon)
x_4	Cilt Kalınlığı (mm) (CK)	Triseps cilt kıvrım kalınlığı
x_5	İnsülin (mu U / ml) (INS)	2 saatlik serum insülini
x_6	BMI (kg / m^2) (BMI)	Vücut Kitle İndeksi
x_7	Diyabet Geçmişi Fonksiyonu (DG)	
x_8	Yaş (Yıl) (YŞ)	

Çalışmada ele alınan 768 hastaya ilişkin diyabet bulgularının sıklık ve oransal dağılımları Tablo 5.2'de verilmektedir.

Tablo 5.2. *Hastaların Diyabet Bulgularına Göre Dağılımı*.

Yanıt Değişkeni	Sıklık	Oran (%)
Diyabet Var (1)	268	34,9
Diyabet Yok (0)	500	65,1
Toplam	768	100

Bu çalışmada, istatistiksel öğrenmede sıklıkla kullanılan sınıflandırma yöntemlerine ilişkin analizlerin yapılması ve performanslarının değerlendirilmesi için SPSS Version 20.0 (Statistical Package for the Social Sciences) paket programı kullanılmıştır. SPSS, anket analizleri başta olmak üzere, sağlık bilimleri, sosyal bilimler ve fen bilimlerinde elde edilen veri setlerinin analizinde kullanılan bir istatistik paketidir.

5.1. Diskriminant Analizi ile Sınıflandırma Uygulaması

Diskriminant analizi, gözlemlerin hangi gruba atanacağına karar vermede ve bağımsız değişkenlere dayanarak grupları birbirinden ayırmada kullanılan bir çok değişkenli istatistiksel analiz tekniğidir. Diskriminant analizi çoğunlukla büyük veri setlerinde bağımsız değişkenler ve yanıt değişkenleri arasındaki ilişkinin incelenmesinde kullanılmaktadır. Bu çalışmada, 768 Pima Yerlisinden alınan bilgiler kullanılarak diskriminant analizi yapılmış ve sonuçlar elde edilmiştir.

Diskriminant analizinin uygulanabilmesi için, çalışmanın üçüncü bölümünde değinildiği gibi bazı temel varsayımlar test edilmelidir. Elde edilen sonuçlara göre analiz sonlandırılmakta ya da farklı yöntemler kullanılarak analiz edilmektedir. Varsayımların sağlanamaması durumunda diskriminanat analizi sınıflandırma için uygun bir analiz olmayacaktır.

Diskriminant analizi varsayımlarından ilki değişkenlerin çok değişkenli normal dağılıma sahip olmalarıdır. Değişkenlerin normalliklerinin test edilmesi amacıyla, tanımlayıcı istatistiklere bakılmış ve Kolmogorov-Smirnov normallik sınaması yapılmıştır.

Pima Yerlilerinin diyabet olma ya da olmama bakımından sınıflandırılması amacıyla yapılan çalışmada kullanılan bağımsız değişkenlere ilişkin tanımlayıcı

istatistikler Tablo 5.3'te verilmiştir. Değişkenlerin çoğunluğunun normal dağılıma uyum gösterdiği belirlenmiştir.

Tablo 5.3. *Değişkenlere İlişkin Tanımlayıcı İstatistikler.*

Değişken	**Ortalama**	**Varyans**	**Medyan**
Gebelik	3,850	11,354	3,000
Glukoz	120,890	1022,248	117,000
Kan Basıncı	69,110	374,647	72,000
Cilt Kalınlığı	20,540	254,473	23,000
İnsülin	79,800	13281,180	30,500
BMI	31,967	62,292	32,000
Diyabet Geçmişi	0,471	0,110	0,372
Yaş	33,240	138,303	29,000

Diskriminant analizi uygulanmadan önce sınanması gereken bir diğer varsayım ise kovaryans matrislerinin eşitliğidir. Bu varsayımın sınanması için Box's M testi yapılmıştır. Teste ait hipotezler,

H_0: Grup kovaryans matrisleri eşittir.

$H_{1:}$ Grup kovaryans matrisleri eşit değildir.

şeklinde verilmiştir.

Tablo 5.4. *Box's M Testi Sonuçları.*

Box's M Değeri	229,880
F Hesap Değeri	6,306
Sd1	36
Sd2	1049414,373
p	0,000

Tablo 5.4'te görüldüğü gibi 0,05 anlamlılık düzeyinde H_0 hipotezi reddedilmiştir. Bu durumda grup kovaryans matrislerinin eşit olmadığı sonucuna varılır. Bu nedenle çoklu karesel diskriminant analizi kullanılmıştır ve tekrar edilen Box's M testi sonucunda yeni oluşan anlamlılık düzeyi 0,275 olarak bulunmuştur. Buna göre gruplar arası kovaryans matrislerinin eşit olduğu sonucuna varılır.

Bir diğer varsayım olan çoklu doğrusal bağlantı olmaması varsayımını test etmek amacıyla değişkenler arası korelasyon matrisi oluşturulmuştur.

Tablo 5.5. *Değişkenler Arası Korelasyon Matrisi*.

	Gebelik	Glukoz	Kan Basıncı	Cilt Kalınlığı	İnsülin	BMI	Diyabet Geçmişi	Yaş
Gebelik	1,000	0,030	0,130	-0,101	-0,106	-0,048	-0,077	0,519
Glukoz	0,030	1,000	0,138	0,025	0,308	0,101	0,064	0,177
Kan Basıncı	0,130	0,138	1,000	0,204	0,081	0,275	0,031	0,231
Cilt Kalınlığı	-0,101	0,025	0,204	1,000	0,432	0,388	0,173	-0,136
İnsülin	-0,106	0,308	0,081	0,432	1,000	0,169	0,165	-0,076
BMI	-0,48	0,101	0,275	0,388	0,169	1,000	0,095	-0,034
Diyabet Geçmişi	-0,077	0,064	0,031	0,173	0,165	0,095	1,000	-0,008
Yaş	0,519	0,177	0,231	-0,136	-0,076	-0,034	-0,008	1,000

Tablo 5.5'te korelasyon katsayılarının oldukça küçük olduğu görülmektedir. Bu durumda çoklu doğrusal bağlantı yoktur denilebilir. Çoklu bağlantının sorununun olup olmadığını belirlemek amacıyla kullanılan diğer yöntem doğrusal regresyon analizi ile çoklu ilişki istatistiklerinin belirlenmesidir.

Tablo 5.6. *Çoklu Bağlantı İncelemesi Sonuçları.*

	Tolerans	VIF
Gebelik	0,699	1,431
Glukoz	0,770	1,299
Kan Basıncı	0,847	1,181
Cilt Kalınlığı	0,664	1,507
İnsülin	0,701	1,427
BMI	0,771	1,298
Diyabet Geçmişi	0,938	1,066
Yaş	0,630	1,588

Tablo 5.6'da görülen çoklu bağlantı incelemesi sonuçlarına göre modelde yer alan bağımsız değişkenlerin Varyans Şişkinlik Faktörü (VIF) değerleri 1'e oldukça yakın ve Tolerans değerleri 0,60'dan küçük olmadığı için değişkenler arasında çoklu bağlantının olmadığı sonucuna varılır.

5.1.1. Diskriminant analizi fonksiyonlarının önemliliği

Diskriminant analizinde fonksiyonun önemliliğini belirlemek amacıyla Özdeğer, Kanonik Koreslasyon ve Wilks' Lambda istatistikleri kullanılır. Bu istatistikler Tablo 5.7 ve Tablo 5.8 ile gösterilmiştir.

Tablo 5.7. *Özdeğerler Tablosu.*

Fonksiyon	Özdeğer	Varyans %	Kümülatif %	Kanonik Korelasyon
1	0,436	100,0	100,0	0,551

Tablo 5.8. *Wilks' Lambda Tablosu*.

Fonksiyonun Testi	Wilks' Lambda	Ki-Kare	Serbestlik Derecesi	*p*
1	0,697	275,581	8	0,000

Diyabet verisinde diyabet olma ve olmama olarak belirlenmiş iki grup olduğundan yalnızca bir diskriminant fonksiyonu üretilmiştir. Özdeğerin büyük olması yanıt değişkenindeki varyansın daha büyük bir bölümünün elde edilen fonksiyon tarafından açıklanacağını göstermektedir. Kesin olmamakla beraber literatürde 0,40 üzerinde değerler alan özdeğerlerin iyi olduğu kabul edilmektedir. Tablo 5.7'de görüldüğü gibi özdeğer 0,436 ile varyansın %100'ünü açıklamaktadır.

Kanonik korelasyon 0 ve 1 arasında değerler alır ve diskriminant skorları ile gruplar arası korelasyonu ölçmektedir. Kanonik korelasyon katsayısının yorumlanabilmesi için karesinin alınması gerekmektedir. Tablo 5.7'de görüldüğü gibi, 0,551 bulunan katsayının karesi alındığında, elde edilen modelin gruplar arası ilişkinin %30,4'ünü açıkladığı söylenebilir.

Diskriminant skorlarındaki toplam varyansın gruplar arasındaki farklar tarafından açıklanamayan kısmını gösteren Wilks' Lambda istatistiği Tablo 5.8'de görüldüğü gibi 0,697 olarak bulunmuştur. 0,697 değerine karşılık gelen ki-kare tablo değeri 275,851 olup varyansın %69,7'si gruplar arası farklar tarafından açıklanamamaktadır. Ayrıca Tablo 5.8'e göre fonksiyonun ayırma gücünün önemli düzeyde yüksek olduğu sonucuna ulaşılır.

5.1.2. Diskriminant analizinde bağımsız değişkenlerin önemliliği

Diskriminant analizinde model oluşturulduktan sonra bağımsız değişkenlerin önemliliğinin test edilmesi gerekmektedir. Bu amaçla diskriminant fonksiyonu katsayıları ve yapı matrisi incelenir.

Yapı matrisi her bir bağımsız değişkenin diskriminant fonksiyonu ile olan korelasyonunu göstermektedir.

Tablo 5.9. *Yapı Matrisi.*

	Fonksiyon 1
Gebelik	0,345
Glukoz	0,799
Kan Basıncı	0,099
Cilt Kalınlığı	0,114
İnsülin	0,199
BMI	0,467
Diyabet Geçmişi	0,265
Yaş	0,372

Tablo 5.9 ile verilen yapı matrisi incelendiğinde, model ile en yüksek korelasyona sahip olan değişken Glukoz iken en düşük korelasyona sahip değişken ise Kan Basıncı olarak görülür.

Tablo 5.10. *Standartlaştırılmış Kanonik Diskriminant Fonksiyonu Katsayıları.*

	Fonksiyon 1
Gebelik	0,308
Glukoz	0,763
Kan Basıncı	-0,205
Cilt Kalınlığı	0,110
İnsülin	-0,094
BMI	0,457
Diyabet Geçmişi	0,218
Yaş	0,136

Tablo 5.10'da standartlaştırılmış kanonik diskriminant fonksiyonu katsayıları verilmektedir. Tabloda görülen standartlaştırılmış diskriminant analizi skorları modeldeki bağımsız değişkenlerin grupların ayrılmasında ne kadar önemli olduğunu gösteren değerlerdir. İşaret gözetmeksizin incelenen diskriminant skorlarında görünen en büyük katsayı ilgili bağımsız değişkenin modeldeki önemliliğini göstermektedir. Tablo 5.10'da görüldüğü gibi en büyük katsayının Glukoz değişkenine, en küçük katsayısının ise İnsülin değişkenine ait olduğu görülmektedir. Bu durumda grupları birbirinden ayırmada en fazla katkıyı Glukoz değişkeni sağlarken, en düşük katkıyı İnsülin değişkeninin sağladığı söylenebilir.

Standartlaştırılmamış kanonik diskriminant fonksiyonu katsayıları ise Tablo 5.11'de verilmiştir.

Tablo 5.11. *Standartlaştırılmamış Kanonik Diskriminant Fonksiyonu Katsayıları.*

	Fonksiyon 1
Gebelik	0,094
Glukoz	0,027
KanBasıncı	-0,011
CiltKalınlığı	0,001
İnsülin	-0,001
BMI	0,061
Diyabet Geçmişi	0,667
Yaş	0,012
Sabit	-5,480

Tablo 5.11'deki katsayılar kullanılarak diyabet olup olmama durumu için diskriminant fonksiyonu oluşturulur.

$$Z = -5{,}480 + 0{,}094\ (\text{Gebelik}) + 0{,}027\ (\text{Glukoz}) - 0{,}011\ (\text{KanBasıncı}) + 0{,}001\ (\text{CiltKalınlığı}) - 0{,}001\ (\text{insülin}) + 0{,}061\ (\text{BMI}) + 0{,}667\ (\text{DiyabetGeçmişi}) + 0{,}012\ (\text{Yaş})$$

Modelden anlaşılacağı gibi yanıt değişkeni üzerinde en büyük etkiyi 1 birimlik artışta 0,667 değerinde pozitif etki yaratan Diyabet Geçmişi değişkeni göstermektedir. İnsülin ve Kan Basıncı değişkenlerinin modelde negatif yönde ve çok düşük düzeyde etkisi olduğu söylenebilir.

Analiz sonucunda ulaşılan her gruba ait ortalama diskriminant fonksiyonu değerleri ise Tablo 5.12'de verilmiştir. Bu çalışmada Diyabet rahatsızlığı olmayan bireylerin oluşturduğu grup '0' iken, Diyabet rahatsızlığı olan bireylerin oluşturduğu grup '1' olarak ifade edilir. Analiz sonuçlarına göre diyabet olmama ve diyabet olma gruplarının ortalama diskriminant değerleri sırasıyla –0,483 ve 0,900'dür. Bu değerler, diyabet olmayan grubun (0) ve diyabet olan grubun (1) ortalama değerlerinin fonksiyona olan uzaklıklarını belirtir. İki grup arasındaki bu fark, fonksiyonun ayırma gücünün oldukça yüksek olduğu biçiminde yorumlanabilir.

Tablo 5.12. *Grupların Ortalama Diskriminant Değerleri*.

Diyabet	Fonksiyon
	1
0	-0,483
1	0,900

5.1.3. Diskriminant analizinde sınıflandırma sonuçları

Tablo 5.13'de sınıflandırma fonksiyonu katsayıları verilmiştir. Katsayılar kullanılarak 0 ve 1 grupları için sınıflandırma fonksiyonları oluşturulur.

Tablo 5.13. *Sınıflandırma Fonksiyonu Katsayıları.*

	Diyabet	
	0	**1**
Gebelik	-0,057	0,073
Glukoz	0,116	0,153
KanBasıncı	0,093	0,079
CiltKalınlığı	0,005	0,006
İnsülin	-0,010	-0,011
BMI	0,440	0,523
Diyabet Geçmişi	2,813	3,736
Yaş	0,163	0,180
Sabit	-19,392	-27,883

Buna göre *'Diyabet yok'* grubu için sınıflandırma fonksiyonu,

$$Y_0 = -19{,}392 - 0{,}057\,(\text{Gebelik}) + 0{,}116\,(\text{Glukoz}) + 0{,}093\,(\text{KanBasıncı}) + 0{,}005\,(\text{CiltKalınlığı}) - 0{,}010\,(\text{insülin}) + 0{,}440\,(\text{BMI}) + 2{,}813\,(\text{DiyabetGeçmişi}) + 0{,}163\,(\text{Yaş})$$

'*Diyabet var'* grubu için sınıflandırma fonksiyonu,

$$Y_1 = -27{,}883 - 0{,}073\,(\text{Gebelik}) + 0{,}153\,(\text{Glukoz}) + 0{,}079\,(\text{KanBasıncı}) + 0{,}006\,(\text{CiltKalınlığı}) - 0{,}011\,(\text{insülin}) + 0{,}523\,(\text{BMI}) + 3{,}736\,(\text{DiyabetGeçmişi}) + 0{,}180\,(\text{Yaş})$$

olarak elde edilir.

Elde edilen diskriminant analizi modelinin başarısının değerlendirilmesi için sınıflandırma tablosu oluşturulur.

Tablo 5.14. *Diskriminant Analizi Sınıflandırma Sonuçları.*

			Tahmin Edilen		Toplam
			Diyabet Yok (0)	Diyabet Var (1)	
Gözlenen	Sayı	0	446	54	500
		1	113	155	268
	%	0	89,2	10,8	100,0
		1	42,2	57,8	100,0

Tablo 5.14'e göre, oluşturulan model %78,3 $((446 + 155)/768)$'lük bir doğru sınıflandırma oranı ile oldukça başarılı bir sınıflandırma gerçekleştirmiştir. 0 ile gösterilen diyabet olmama durumunda 446 gözlem doğru sınıflandırılırken 54 gözlem yanlış sınıflandırılmıştır. Diyabet olmama durumunun yanlış sınıflandırılma oranının %10,8 olduğu görülmektedir. 1 ile gösterilen diyabet olma durumunda ise 155 gözlem doğru sınıflandırılırken 113 gözlem yanlış sınıflandırılmıştır. Buna göre diyabet olma durumunun yanlış sınıflandırma oranı %42,2'dir.

5.2. Lojistik Regresyon ile Sınıflandırma Uygulaması

Lojistik regresyon analizi ile gözlemlerin hangi gruba ait olduğunu bulmaya yönelik bir model oluşturulmaya çalışılır. Lojistik regresyon analizinin amaçlarından biri sınıflandırma yapmak iken, diğeri yanıt değişkeni ile bağımsız değişkenler arasındaki ilişkileri incelemektir. Bu analiz yardımıyla, Pima Yerlilerinin kadın bireylerinin çeşitli tıbbi etkenler göz önüne alındığında diyabet olup olmadığına ilişkin sonuçlar ikili lojistik regresyon analiziyle değerlendirilmiştir.

Lojistik regresyon analizinde diskriminant analizinde olduğu gibi analize başlamadan önce çeşitli varsayımların incelenmesine gerek yoktur. Bundan dolayı analize hiçbir varsayım araştırması yapılmadan başlanmıştır.

Lojistik regresyon modeline tüm değişkenler dahil edilmiş ve kullanılan değişkenlerin önemliliğinin değerlendirilmesinde olabilirlik oran testi kullanılmıştır.

5.2.1. Model anlamlılığının test edilmesi

Tablo 5.15. *Model Katsayıları İçin Omnibus Testi.*

		Ki-Kare	Serbestlik Derecesi	*p*
Adım 1	Adım	270,376	8	0,000
	Blok	270,376	8	0,000
	Model	270,376	8	0,000

Tablo 5.15'te yapılan analiz sonucunda anlamlılık değerlerinin 0,05'den küçük olması sebebiyle ikili lojistik modelin anlamlı olduğu söylenebilir. Dolayısıyla bağımsız değişkenlerden en az birinin yanıt değişkeniyle anlamlı derecede ilişkili olduğu ortaya çıkmıştır. Sadece sabit içeren modelin *-2 Log olabilirlik* değeri ile bağımsız değişkenleri içeren modelin *-2 Log olabilirlik* değeri arasındaki fark 270,376 olarak bulunmuştur.

Modelin uygunluğunun test edilmesinde Hosmer ve Lemeshow testi de kullanılmaktadır. Hosmer ve Lemeshow testinde, tahmin edilen değerlerle gözlenen değerler arasında fark bulunmadığı biçiminde kurulan sıfır hipotezi test edilmektedir. Uygulanan bu test sonucu Tablo 5.16'da görülmektedir.

Tablo 5.16. *Hosmer ve Lemeshow Testi.*

Adım	Ki-Kare	Serbestlık Derecesi	*p*
1	8,336	8	0,401

Test sonucunda anlamlılık değeri 0,05'den büyük bulunduğundan model ile verinin uyumunun yeterli düzeyde olduğu söylenebilir.

Tablo 5.17. *Lojistik Regresyon Modelinin Özeti.*

Adım	-2 Log olabilirlik	Cox & Snell R Kare	Nagelkerke R Kare
1	723,108	0,297	0,409

Tablo 5.17'de görüldüğü gibi, modele ait Cox ve Snell R kare istatistiği 0,297 olarak bulunmuştur. Bu durumda bağımsız değişkenlerin yanıt değişkenini %29,7 oranında açıklayabildiği söylenebilir. Nagelkerke R kare istatistiği ise genel olarak Cox ve Snell R kare istatistiğine göre daha yüksek değerler almaktadır. R kare istatistikleri çoğunlukla küçük çıkma eğiliminde olduğundan, model uyumunun incelenmesinden ziyade model oluşturma aşamasında farklı modellerin performanslarını karşılaştırmak amacıyla kullanılmaktadır.

Tablo 5.18. *İkili Lojistik Regresyon Analizi Sonuçları.*

	β Katsayısı	Standart Hata	Wald İstatistiği	Serbestlik Derecesi	p	Exp (β)
(GB)	0,123	0,032	14,714	1	0,000	1,131
(GL)	0,035	0,004	89,779	1	0,000	1,036
(KB)	-0,013	0,005	6,478	1	0,011	0,987
(CK)	0,001	0,007	0,008	1	0,930	1,001
(İNS)	-0,001	0,001	1,767	1	0,184	0,999
(BMI)	0,090	0,015	35,711	1	0,000	1,094
(DG)	0,939	0,299	9,853	1	0,002	2,559
(YŞ)	0,015	0,009	2,519	1	0,113	1,015
Sabit	-8,409	0,716	137,755	1	0,000	0,000

Tablo 5.18'de yapılan ikili lojistik regresyon analizi sonuçları verilmektedir. Tablonun son sutününda görülen Exp (β) değerleri odds oranlarını belirtmektedir. Bu çalışmada odds oranı, diyabet olma olasılığının olmama olasılığına oranını ifade etmektedir ve bağımsız değişkenlerdeki değişikliklerin yanıt değişkeni üzerindeki etkilerini göstermektedir. Gebelik dikkate alındığında diyabet olma olasılığı diyabet olmama olasılığına göre 1,131 kat daha fazladır. Glukoz değişkenindeki bir birim artış diyabet görülme olasılığını 1,036 kat, Cilt Kalınlığı değişkenindeki bir birimlik artış diyabet görülme olasılığını 1,001 kat, BMI değişkenindeki bir birim artış diyabet görülme olasılığını 1,094 kat, Diyabet Geçmişi değişkenindeki bir birimlik artış diyabet görülme olasılığını 2,559 kat, Yaş değişkenindeki bir birim artış diyabet görülme olasılığını 1,015 kat arttırmaktadır.

İnsülin değişkeninin Exp (β) değeri 1'e çok yakın olmakla birlikte, Kan Basıncı değişkenine ilişkin Exp (β) değerinin 1'den bir miktar küçük çıkması nedeniyle, kan basıncındaki bir birimlik artış diyabet görülme olasılığını az da olsa düşürmektedir.

Analiz sonucunda %5 anlamlılık düzeyinde Cilt Kalınlığı, İnsülin ve Yaş değişkenlerinin anlamlı olmadıkları görülmüştür. Dolayısıyla istatistiksel olarak anlamlı bulunan değişkenler ile lojistik regresyon modeli aşağıdaki gibi kurulmuştur.

$$Z = \ln(p/1-p) = -8{,}409 + 0{,}123(\text{Gebelik}) + 0{,}035(\text{Glukoz}) - 0{,}013(\text{KanBasıncı}) + 0{,}090\ (\text{BMI}) + 0{,}939\ (\text{DiyabetGeçmişi})$$

Lojistik regresyon modelini kullanarak diyabet olma olasılıklarını bulmak amacıyla her bir birey için Z_i değerleri bulunur. Elde edilen bu değerler daha sonra p_i lojistik regresyon fonksiyonunda yerine konularak bireylerin diyabet olma olasılıkları

bulunabilir. Modelde diyabet olma ve olmama durumları 0 ve 1 ile ifade edildiği için, bu değerlerin ortalaması olan 0,5 kopuş değeri olarak alınır. p_i değeri 0,5'ten küçük olan bireylerde diyabet rahatsızlığının olduğu, 0,5'ten büyük olan bireylerde ise diyabet rahatsızlığının olmadığı kabul edilir.

$$p_i = \frac{1}{1 + e^{-z_i}} = \frac{1}{1 + e^{-(-8{,}409+0{,}123(GB)+0{,}035(GL)-0{,}013(KB)+0{,}090(BMI)+0{,}939(DG)}}$$

5.2.2. Lojistik regresyon analizinde sınıflandırma sonuçları

Tablo 5.19'da, gerçekleştirilen iki gruplu lojistik regresyon analizine ilişkin sınıflandırma sonuçları görülmektedir.

Tablo 5.19. *Lojistik Regresyon Analizi Sınıflandırma Sonuçları.*

			Tahmin Edilen		
			Diyabet Yok (0)	Diyabet Var (1)	Toplam
Gözlenen	Sayı	0	445	55	500
		1	112	156	268
	%	0	89,0	11,0	100,0
		1	41,8	58,2	100,0

Tahmin edilen modelin sınıflandırma başarısı %78,3 ((445+156)/768)'tür. Diyabet olmama durumunda 500 gözlemden 445 tanesi doğru sınıflandırılırken 55 gözlem yanlış sınıflandırılmıştır. Diyabet olma durumunda ise 268 gözlemden 156 tanesi doğru sınıflandırılırken 112 adedi yanlış sınıflandırılmıştır. Diyabetli bireylerin doğru sınıflandırılma oranı %58,2 iken diyabetli olmayan bireylerin doğru sınıflandırma oranının ise %89,0 olduğu görülmektedir.

6. SONUÇ VE ÖNERİLER

Son yıllarda teknolojinin hızla gelişmesiyle veri miktarında büyük artışlar meydana gelmiştir. Bu verilerin bilgiye dönüştürülmesi ise günümüzün en temel problemlerinden ve zorunluluklarından biridir. İstatistiksel öğrenme bilimi kısaca veriden öğrenme olarak da ifade edilebilir. İstatistiksel öğrenmenin en önemli amaçlarından biri sınıflandırmadır. Modelde kullanılan büyük veri yığınlarının sınıflandırılması istatistiksel analizlerin önemli bir bölümünü oluşturarak, başta sağlık ve teknoloji olmak üzere çeşitli bilim dallarında çok geniş bir alanda kullanılmaktadır.

Bu çalışmada, istatistiksel sınıflandırma tekniklerinden diskriminant analizi ve lojistik regresyon analizi ayrıntılı olarak incelenmiştir. Pima Yerlisi kadın bireylerin diyabet rahatsızlığı olup olmamasını etkilediği düşünülen, gebelik, glukoz, kan basıncı, cilt kalınlığı, insülin, BMI, diyabet geçmişi ve yaş değişkenleri kullanılarak diskriminant analizi ve iki gruplu lojistik regresyon analizi modelleri kurulmuş ve bu modeller yardımıyla gözlemlerin doğru sınıflandırılmasına çalışılmıştır.

Diskriminant analizi ile kurulan fonksiyonunun anlamlılığının testinde kullanılan istatistikler incelendiğinde, fonksiyonun grupları ayırıcılık değeri olan özdeğerinin 0,436 ile kabul edilebilir bir değerde olduğu görülmüş, kanonik korelasyon değeri ile modelin gruplar arası ilişkinin % 30,4'ünü açıklayabildiği ve varyansın %69,7'sinin gruplar arası farklar tarafından açıklanamadığı sonucuna varılmıştır. Grupların ayrılmasında modeldeki bağımsız değişkenlerden en önemli katkıyı Glukoz değişkeni sağlamış, en az katkıyı ise İnsülin değişkeni vermiştir. Diyabet olma ve olmama yanıt değişkeni üzerinde en etkili değişken Diyabet Geçmişi iken, İnsülin ve Kan Basıncı değişkenleri ters etkilidir.

Çalışmada çeşitli tıbbi değişkenler yardımıyla diyabet olma olasılığı tahmin edilirken kullanılan diğer bir analiz ise lojistik regresyon analizidir. Lojistik regresyon analizinde modelin anlamlılığını test etmek için Hosmer-Lemeshow istatistiği kullanılmış ve testin sonucunda model ile verinin uyumunun yeterli düzeyde olduğu görülmüştür. Analizde hesaplanan odds oranları incelendiğinde Diyabet Geçmişinin bireylerin diyabet olma olasılığını en fazla etkileyen değişken olduğu gözlemlenmiş ve diskriminant analizi sonuçlarına benzer şekilde İnsülin ile Kan Basıncı değişkenlerinin yanıt değişkenini ters yönde etkiledikleri sonucuna ulaşılmıştır. Lojistik regresyon modelinde diskriminant analizi modelinden farklı olarak Cilt Kalınlığı, İnsülin ve Yaş değişkenlerinin anlamlı olmadığı sonucuna varılmıştır.

Kurulan modellerde değişkenlerin katsayılarının farklılık göstermesi, istatistiksel analiz tekniklerinin sınıflandırma problemini çözme biçimleri düşüldüğünde beklenen bir durumdur. Sonuçta uygulanan iki analizin de başarılı olduğu söylenebilir. Verilerin doğru sınıflandırılma oranları her iki analiz için de %78,3 olarak bulunmuştur. İki istatistiksel sınıflandırma analizinin de aynı sonucu vermesi, diyabet rahatsızlığı olup olmamayı etkileyen değişkenler bilindiğinde, modele yeni alınan bireyleri gruplara atayabilmek için kurulan diskriminant analizi ve lojistik regresyon analizi modellerinin birbirlerine alternatif olarak kullanılabileceğini göstermektedir.

Gelecekteki çalışmalarda modelde yer alan tüm değişkenlerin anlamlılığının sağlanabilmesi, kullanılabilecek değişkenlerin çeşitlendirilebilmesi ve bu değişkenler arasındaki ilişkiler detaylandırılarak daha fazla veriyle Pima Yerlisi kadın bireylerin diyabet rahatsızlığı geçirme olasılıkları belirlenmeye çalışılabilir. Böylelikle, bu rahatsızlıkla ilişkin alınabilecek önlemlerle ilgili önerilerde bulunulabilecektir.

KAYNAKÇA

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

Agresti A. (2002). *Categorical Data Analysis*. New York: John Wiley & Sons.

Akpınar, H. (2014). *DATA Veri Madenciliği Veri Analizi.* İstanbul: Papatya Yayıncılık Eğitim.

Albayrak, A. S. (2006). *Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım.

Anderson, T. W. ve Bahadur, R. R. (1962). Classification into Two Multivariate Normal Distributions With Different Covariance Matrices. *The Annals of Mathematical Statistics.* 33 (2), 420-431.

Aranda ve Ordaz, F.J. (1981). On Two Families of Transformations to Additivity for Binary Responce Data, *Biometrika,* 68 (2), 357-363.

Bartlett, M. S. ve Please, N. W. (1963). Discrimination in The Case of Zero Mean Differences. *Biometrika,* 50 (1-2), 17-21.

Bayram, N. (2004). Multinominal Lojistik Regresyon Analizinin İstihdamdaki İşgücüne Uygulanması. *İstanbul Üniversitesi İktisat Fakültesi Mecmuası,* 54 (2), 61-75.

Begg, C. B. ve Gray, R. (1984). Calculation of Polychotomous Logistic Regression Parameters using Individualized Regressions. *Biometrika,* 71 (1), 11-18.

Berkson, J. (1944). Application of The Logistic Function to Bio-Assay. *Journal of The American Statistical Association,* 39 (227), 357-365

Bonney, G. E. (1987). Logistic Regression for Dependent Binary Observations, *Biometrics,* 43 (4), 951-973.

Bousquet, O., Boucheron, S. ve Lugosi, G. (2004). *Introduction to Statistical Learning Theory, Advanced Lectures on Machine Learning.* Berlin Heidelberg: Springer.

Börüban, C. (2009) *Firmaların Mali Başarısızlıklarının Öngörülmesinde Diskriminant Analizi Ve Lojistik Regresyon Analizi Yöntemlerinin Karşılaştırılması.* Yüksek Lisans Tezi. İstanbul: Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

Bunke, O. (1964). Uber Optimale Verfahren der Discriminanzanalyse, *Abl. Deutsch. Akad. Wiss. Klasse Math. Phys. Tech,* 4, 35-41.

Burmaoğlu, S. (2009). *Birleşmiş Milletler Kalkınma Programı Beşeri Kalkınma Endeksi Verilerini Kullanarak Diskriminant Analizi, Lojistik Regresyon Analizi Ve Yapay Sinir Ağlarının Sınıflandırma Başarılarının Değerlendirilmesi.* Doktora Tezi. Erzurum: Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı.

Cangül, O. 2006. *Dikriminant Analizi ve Bir Uygulama Denemesi*. Yüksek Lisans Tezi. Bursa: Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

Carroll, R. J., Spiegelman, C. H., Gordon K. K., Bailey, K. T. and Abbott, R. D., (1984). On Errors in Variables for Binary Regression Models, *Biometrika*, 71 (1), 19-25.

Cavalli, L. L. (1945). Alumni Problemi Della Analisi Biometrica di Popolazioni Naturali, *Mem. Ist. Idrobiol.,* 2, 301-323.

Cooley, W. W. ve Lohnes, P. R. (1973). Multivariate Data Analysis. *Journal of The Royal Statistical Society-Series B*, 136 (1), 101-103.

Cooper, P. W. (1963). Statistical Classification with Quadratic Forms. *Biometrika.* 50 (3-4), 439-448.

Cornfield, J. (1962). Joint Dependence of the Risk of Coronary Heart Disease on serum Cholesterol and Sistolic Blood Pressure: A Diskrimant Function Analysis, *Federation Proceedings,* 21, 58-6.

Cox, D. R. ve Snell, E. J. (1989). *Analysis of Binary Data: 2nd Edition.* London: Chapman and Holl/ CRC.

Çinko, M. (2003). *Diskriminant ve Lojistik Regresyon Analizlerinde Bootstrap Tekniğinin Kullanımı Ve Kredi Riski Modeli Oluşturulması.* Doktora Tezi. İstanbul: Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

Duffy, D.E. (1990). On Continuity-Corrected Residuals in Logistic Regression. *Biometrika,* 77 (2), 287-293.

Demirhan, A. ve Oktay Fırat, Ü. (2003). Ticaret Bankalarının 1999 ve 2000 Yıllarındaki Finansal Performanslarının Faktör Analizi ve Diskriminant Analizi Kullanılarak Araştırılması. *İstanbul Üniversitesi İşletme Fakültesi Dergisi,* 32 (2), 9-27.

Fisher R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics,* 7 (7), 179-188.

Gardner, S. ve Le Roux, N. J. (2005). Extensions of Biplot Methodology to Discriminant Analysis. *Journal of Classification,* 22 (1), 59-86.

Goldberger, A. S. (1964). *Econometric Theory*. New York: John Wiley & Sons.

Güneri, Ö. ve Aydın, D. (2017). Grup Üyelerini Belirlemede İstatistiksel Sınıflandırma Yöntemleri: Karşılaştırmalı Bir Çalışma. *Türkiye Klinikleri J Biostat*. 9 (1). 45-67.

Hastie T., Tibshirani, R. ve Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. California: Springer.

Hosmer, D.W. ve Lemeshow, S. (2000). *Applied Logistic Regression.* New York: John Wiley & Sons.

Hsu, J. S. J. ve Leonard, T. (1997). Hierarchical Bayesian Semiparametric Procedures for Logistic Regression. *Biometrika*, 84, 85-93.

Huberty, C. J. ve Hussein, M. H. (2003). Some Problems in reporting use of discriminant Analyses. *The Journal of experimental Education,* 71 (2), 177-192.

Kamışlı, M. ve Girginer, N. (2010). İşlem Bazlı Manipülasyonun İstatistiksel Sınıflandırma Analizleriyle Belirlenmesi. *İstanbul Üniversitesi İktisat Fakültesi Ekonometri Ve İstatistik Dergisi,* 11, 1-30.

Kara, Ö. S. (2015). *Lojistik Regresyon Analizi ve Kadın İşgücü Üzerine Bir Uygulama*. Yüksek Lisans Tezi. Bursa: Uludağ Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı.

Kaygın, C., Tazegül, A. ve Yazarkan, H. (2016). Estimation Capability of Financial Failures and Successes of Enterprises Using Data Mining and Logistic Regression Analysis. *Ege Akademik Bakış Dergisi,* 16 (1), 147-159.

Klecka, W.R. (1980). *Discriminant Analysis. Quantitative Applications in Social Sciences*. USA: Sage Publications.

Koyuncugil, A. ve Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. *Bilişim Teknolojileri Dergisi*, 2 (2), 21-31.

Kulkarni, S. ve Harman, G. (2011). *An Elementary Introduction to Statistical Learning Theory.* NJ: John Wiley & Sons.

Kuyucu, Y. (2012). *Lojistik Regresyon Analizi (LRA), Yapay Sinir Ağları (YSA) ve Sınıflandırma ve Regresyon Ağaçları (C&RT) Yöntemlerinin Karşılaştırılması ve Tıp Alanında Bir Uygulama.* Yüksek Lisans Tezi. Tokat: Gaziosmanpaşa Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı.

Lachenbruch, P. A. (1975). *Discriminant Analysis*. New York: Hafner.

Lam, K. F. ve Moy, J. W. (2003). A Simple Weighting Scheme for Classification in Two Group Discriminant Problems. *Computers and Operations Research,* 30 (1), 155-164.

Liao, T.F. (1994). *Interpreting Probability Models. Logit, Probit and Other Generalized Linear Models. Quantitative Applications in Social Sciences*. USA: Sage Publications.

Penrose, L. S. (1947). Some Notes On Discrimination. *Annals of Eugenics*. 13 (1), 514-528.

Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics*, 9 (4), 705-724.

Rao, C. R. (1948). The Utiliztion of Multiple Measurements in Problems of Biologial Classification. *Journal of The Royal Society-Series B,* 10 (2), 159-203.

Rebellon, C. J. (2012). Differential Association and Substance Use: Addensing The Roles of Discriminant Validity, Socialization and Selection in Traditional Empirical Tests. *European Journal of Criminology,* 9 (1), 74-96.

Robert, G., Rao, J. N. K. ve Kumar, S. (1987). Logistic Regression Analysis of Sample Survey Data. *Biometrika*, 74 (1), 1-12.

Roy, S. N. (1939). *p*-Statistics or Some Generalisations in Analysis of Variance Appropriate to Mutlivariate Problems. *Sankhya*, 4, 381-396.

Steinley, D. ve Brusco, M. J. (2011). Choosing The Number of Clusters in K-means Clustering. *Psychological Methods,* 16 (3), 285-297.

Smith, C. A. B. (1947). Some Examples of Discrimination. *Annals of Eugenics,* 13(1), 272-282.

Sueyoshi, T. (2004). Mixed Integer Programming Approach of Extended DEA-Discriminant Analysis. *European Journal of Operational Research,* 152 (1), 45-55.

Şensoy, E. Z. (2009). *Nonlineer Lojistik Regresyon ve Uygulaması.* Yüksek Lisans Tezi. İstanbul: Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı.

Tatlıdil, H. (1996). *Uygulamalı Çok Değişkenli İstatistiksel Analiz.* Ankara: Akademi Matbaası. 54 (2), 61-75.

Tatsuoka, M. (1971). *Multivariate analysis. Techniques for Educational and Psychological Research.* New York: John Wiley & Sons.

Tolun, S. (2008). *Destek Vektör Makineleri: Banka Başarısızlığı Tahmini Üzerine Bir Uygulama.* Doktora Tezi. İstanbul: İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı.

Ürük, E. (2007). *İstatistiksel Uygulamalarda Lojistik Regresyon Analizi.* Yüksek Lisans Tezi. İstanbul: Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı.

Vapnik, V.N. (1998). *Statistical Learning Theory*. USA: John Wiley & Sons.

Von Mises, R. (1954). One the Classification of Observation Data into Distinct Groups. *The Annals of Mathematical Statistics.* 16, 68-73.

Vupa, Ö. (2004) *Model Building Of Logistic Regression Models.* Yüksek Lisans Tezi. İzmir: Dokuz Eylül Üniversitesi.

http-1: https://newonlinecourses.science.psu.edu/stat508/lesson/9

(Erişim tarihi: 19.11.2019)

http-2: https://www.academia.edu/34924130/diskriminant.pptx

(Erişim tarihi: 31.10.2019)

http-3: www2.chass.ncsu.edu/garson/pa765

(Erişim tarihi: 01.11.2019)

http-4: http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf

(Erişim tarihi: 13.12.2019)

http-5: http://78.189.53.61/-/bs/ess/k_sumbuloglu.pdf

(Erişim tarihi: 22.11.2019)

http-6: https://www.kaggle.com/uciml/pima-indians-diabetes-database

(Erişim tarihi: 17.11.2019)

ÖZGEÇMİŞ

Adı Soyadı : Gizem UYLU

Yabancı Dil : İngilizce

Doğum Yeri ve Yılı : Eskişehir / 1990

E-Posta : gibu26@gmail.com

Eğitim ve Mesleki Geçmişi:

- 05.2015- 02.2016, Gişe Yetkilisi, Akbank T.A.Ş.
- 05.2017- 09.2017, Denetmen, Seri Jakala.
- 2015-2020, Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstatistik Anabilim Dalı.
- 2009-2014, Anadolu Üniversitesi, Fen Fakültesi, İstatistik Bölümü.
- 2011-2014, Anadolu Üniversitesi, Açıköğretim Fakültesi, Dış Ticaret Bölümü.
- 2004-2008, H. Ahmed Yesevi Süper Lisesi.

Sanatsal Faaliyetleri:

- 2009-2014, Sanatsal, Aktif Dansçı, Anadolu Üniversitesi Halk Dansları Topluluğu, Eskişehir.
- 2013, Sanatsal, Dansçı, Terranostra Festival İnternazionale del Dolclore, Apiro, Italy.
- 2012, Sanatsal, Dansçı, Hello Schoten 54ste. WereldddansFESTİVAL, Schoten, Belgium.