BİRLEŞMİŞ MİLLETLER KALKINMA PROGRAMI BEŞERİ KALKINMA ENDEKSİ VERİLERİNİ KULLANARAK DİSKRİMİNANT ANALİZİ VE LOJİSTİK REGRESYON ANALİZİNİN SINIFLANDIRMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Serhat BURMAOĞLU[1]
Erkan OKTAY[2]
Üstün ÖZEN[3]

ÖZET

Sınıflandırma gerçek hayatta birçok alanda farklı yöntemler kullanılarak yapılmaktadır. Bu çalışmada çok değişkenli istatistiksel sınıflandırma yöntemlerinden diskriminant analizi ve lojistik regresyon analizi incelenmiştir. Çalışmanın amacı iki yöntemin kullanımını metodolojik olarak göstermek ve sınıflandırma başarısı sonuçlarını karşılaştırmaktır. Uygulama verisi olarak Birleşmiş Milletler Kalkınma Programının Beşeri Kalkınma endeksi 2007/2008 verileri kullanılmıştır. Analizler sonrasında Diskriminant analizinde %92,5'lik ve Lojistik Regresyon Analizinde %100'lük sınıflandırma başarısı elde edilmiştir.

Anahtar Kelimeler: Diskriminant Analizi, Lojistik Regresyon , Logit, Sınıflandırma

COMPARING CLASSIFICATION SUCCESS OF DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION ANALYSIS USING UNITED NATIONS DEVELOPING PROGRAMME'S HUMAN DEVELOPMENT INDEX

ABSTRACT

Classification has been made on many fields with various techniques in real life. In this study discriminant analysis and logistic regression analysis are scrutinized as statistical classification methods. Purpose of this study is to show the methodology of two-techniques' usage and compare classification success results. United Nations Developing Programme's Human Development Index 2007/2008 data have been used as application data. After making analysis, classification success of Discriminant Analysis has been found as %92,5 and classification success of Logistic Regression Analysis has been found as %100.

Keywords: Discriminant Analysis, Logistic Regression, Logit, Classification

[1] Kara Harp Okulu Sistem Yönetim Bilimleri Bölümü Öğretim Elemanı, Bakanlıklar, Ankara. sburmaoglu@kho.edu.tr
[2] Prof. Dr., Atatürk Üniversitesi İİBF İşletme Bölümü Sayısal Yöntemler Anabilim Dalı Öğretim Üyesi, Erzurum. erkanoktay@hotmail.com
[3] Doç. Dr., Atatürk Üniversitesi İİBF İşletme Bölümü Sayısal Yöntemler Anabilim Dalı Öğretim Üyesi, Erzurum. uozen@atauni.edu.tr

Makalenin geliş tarihi: 11.09.2009 ***Kabul tarihi:*** 19.10.2009

GİRİŞ

Çok değişkenli istatistiksel analizlerde sıklıkla karşılaşılan problemlerden birisi sınıflandırma sorunudur. Araştırmacı farklı yığınlardan gelen bireylerin p sayıdaki özelliğini ölçtüğünde elindeki bireyin hangi gruptan geldiğini merak edebilir. Bu durumda sınıflandırma problemi, bireyin p sayıda özelliğini inceleyerek hangi gruptan geldiğine karar verme problemi olarak nitelendirilebilir.

Sınıflandırma problemi stokastik istatistiksel bir karar verme sürecidir. Bu süreçte araştırmacı, bireyin hangi gruptan geldiğine karar vermelidir. Bazı durumlarda grupların olasılık dağılımları ve bu dağılımların parametreleri bilinmektedir. Ancak uygulamada genellikle her grubun p değişkene ilişkin bir dağılıma sahip olduğu varsayılır ve bu dağılımın parametreleri seçilen örnek aracılığıyla tahmin edilir. Ardından karar verme problemi çözülmeye çalışılır. Bu düzeyde, araştırmacı için iki karar verme konusu bulunmaktadır. Birincisi grubun ayırt edici özelliklerini araştırarak ayırt edicilikte etkili olan değişkenleri belirlemek, ikincisi bu ayırt edici fonksiyonlar yardımıyla bireyleri gruplara sınıflandırmaktır.

Sınıflandırma yöntemleri incelendiğinde iki grubun oluştuğu görülmektedir. Sınıfların önceden bilinen gruplar olması veya önceden grupların bilinmemesi durumuna göre sınıflandırma teknikleri kendi içlerinde ikiye ayrılmaktadır. Sınıfların önceden bilinmemesi durumuna göre sınıflandırmada çok boyutlu ölçekleme analizi ve kümeleme analizi kullanılırken, sınıfların önceden bilinmesi durumunda ise diskriminant analizi ve lojistik regresyon analizi kullanılmaktadır.

Bu çalışmanın amacı, geleneksel çok değişkenli istatistiksel yöntemlerden Diskriminant Analizi ve Lojistik Regresyon Analizinin sınıflandırma performansının karşılaştırılmasıdır. Yapılacak uygulama ile sınıflandırma başarıları karşılaştırılarak başarı yüzdeleri ve analiz tekniği hakkında yorum yapılmıştır. Bu çalışmada önceden bilinen ve Birleşmiş Milletler Kalkınma Programı tarafından yapılan çok gelişmiş ve orta düzeyde gelişmiş ülke sınıflandırması dikkate alındığından yeniden sınıflandırma için diskriminant analizi ve lojistik regresyon analizi yöntemleri kullanılmıştır.

Çalışma dört bölümden oluşmaktadır. Birinci bölümde diskriminant analizinden ve ikinci bölümde lojistik regresyon analizinden bahsedilmiştir. Üçüncü bölümde sınıflandırma yöntemlerinin birlikte kullanıldığı bilimsel makaleler incelenmiş, diskriminant ve logit analizleri arasındaki benzerlikler

ve farklılıklar gösterilmeye çalışılmıştır. Dördüncü bölümde örnek bir veri seti üzerinde analizler uygulanmış ve sonuçları karşılaştırılmıştır.

1. DİSKRİMİNANT ANALİZİ

Diskriminant analizi, bir araştırmacının aynı anda çeşitli değişkenlere göre iki veya daha fazla örnek grup arasındaki farklılıklar üzerinde çalışmasına olanak sağlayan bir istatistiksel tekniktir. Genel olarak birimlerin gruplanmasında bazı matematiksel eşitliklerden faydalanılır. Diskriminant fonksiyonu olarak adlandırılan bu eşitlikler, birbirine en çok benzeyen grupları belirlemeye olanak sağlayacak şekilde grupların ortak özelliklerini belirlemek amacıyla kullanılmaktadır. Grupları ayırmak amacıyla kullanılan karakteristikler ise diskriminant değişkenleri olarak adlandırılmaktadır. Kısaca, diskriminant analizi, iki veya daha fazla sayıdaki grubun farklılıklarının diskriminant değişkenleri vasıtasıyla ortaya konması işlemidir. Birbiriyle yakından ilişkili birkaç istatistiksel yaklaşımı kapsayan geniş bir kavramdır (Klecka 1980).

Diskriminant analizi aracılığıyla elde edilen diskriminant (ayırıcı) fonksiyonları, tahmin değişkenlerinin doğrusal bileşenlerinden oluşur. Diskriminant fonksiyonları gruplar arası farklılığa etki eden tahmin değişkenlerinin hangileri olduğunu ortaya çıkarır. Gruplar arası farklılığa etki eden bu değişkenlere diskriminant (ayırıcı) değişkenler denir. Diskriminant analizinin bir diğer işlevi ise, gruplardan herhangi birisine ait olan fakat hangi gruptan geldiği bilinmeyen bir birimin ait olduğu grubu en az hata ile saptamaktır.

Diskriminant analizi, farklılığın en fazla hangi değişkenlerde yoğunlaştığının belirlenmesi ve böylece grupların farklılaşmasında etkili olan faktörlerin saptanmasını da sağlar. Analiz sonucunda yapılan sınıflama ile orijinal grup üyeliklerinin karşılaştırılması, bilinen fonksiyonun yeterli olup olmadığını test etmeye olanak sağlar (Erçetin 1993).

Diskriminant analizi, Çok Değişkenli Varyans Analizi (MANOVA) yönteminde olduğu gibi grupları ortalamalarına (ortalama vektörlerine) göre ortak ortalamadan (ortalama vektöründen) farklı olmalarını sağlayacak bir ayırma kriteri geliştirmeyi amaçlayan bir yöntemdir. Bu nedenle veri setlerine diskriminant analizi uygulanabilmesi için veri setlerinin aşağıdaki varsayımları taşıması gereklidir.

- X veri matrisi çok değişkenli normal dağılım göstermelidir.

- Değişkenlerin varyans ve kovaryansları homojen olmalıdır. X matrisinde yer alan değişkenler ortak kovaryans matrisine sahip çok değişkenli ana kütleden çekilmiş örnekler olmalıdır.
- Değişkenlerin ortalamaları ve varyansları arasında bir korelasyon bulunmamalıdır.
- Değişkenler arasında çoklu bağlantı (multicollinearity) bulunmamalıdır.
- X matrisi grupların birbirinden ayrılmasında rol oynamayacak gereksiz değişken içermemeli, grupların birbirinden ayrılmasını sağlayacak kadar doğru ve gerekli değişkenleri içermelidir.

Bazı araştırmacılar diskriminant analizinde diskriminant fonksiyonu katsayılarının hesaplanmasında başvurulan yöntemlere göre Diskriminant Analizi isminin başına getirilen ek sözcüklere göre Fisher'in Doğrusal Diskriminant Analizi, Kernel Tabanlı Kümeleme ile Diskriminant Analizi (Kernel Based Discriminant Analysis), En Büyük Benzerlik Diskriminant Analizi (Maximum Likelihood Discriminant Analysis), Bayes Diskriminant Analizi (Bayesian Discriminant Analysis), Laplacian Doğrusal Diskriminant Analizi (Laplacian Linear Diskriminant Analysis) gibi isimlerle anmayı uygun bulmaktadırlar (Tang vd 2005; Liang ve Shi 2004; Lu vd 2005; Zheng 2005; Srivastava vd 2007). Bu çalışmada Kuadratik diskriminant analizi kullanıldığından diğer yöntemlere yer verilmeden yalnızca bu konu ile ilgili matematiksel altyapıdan bahsedilecektir.

Doğrusal diskriminant fonksiyonunun normallikten uzaklaşmayı engellemede kuvvetli, fakat eğik dağılımlarda kullanılamayacağı bilinmektedir. Bu varsayımların bozulduğu durumlarda alternatif fonksiyonlar kullanılır. Kuadratik diskriminant fonksiyonu verilerin normal dağıldığı ancak grupların varyans-kovaryans matrislerinin farklı olmaları durumunda kullanılan fonksiyondur. Kovaryans matrislerinin eşitliği varsayımı nadiren görülebilen bir durumdur (Lachenbruch, 1975: 20).

Kuadratik diskriminant analizinde katsayıların hesaplanmasında ortak kovaryans matrisi yerine (S) grupların kovaryans matrislerinin farkları alınır.

$$Q(x) = \frac{1}{2}\log\frac{|S_j|}{|S_i|} - \frac{1}{2}(x^{-(i)}S_i^{-1}x^{-(i)} - x^{-(j)}S_j^{-1}x^{-(j)} + x(S_i^{-1}x^{-(i)} - S_j^{-1}x^{-(j)})) - \frac{1}{2}x(S_i^{-1} - S_j^{-1})x \quad (1)$$

Başlangıçta iki grup için geliştirilen bu fonksiyon ikişerli alınarak çok grup olma durumu için de kullanılır. Fonksiyonda S_i ve S_j sırasıyla i'nci ve

j'nci gruba ilişkin varyans-kovaryans matrisleridir. $S_i=S_j=S$ alınırsa; karesel fonksiyon doğrusal fonksiyona eşit olacaktır.

Fonksiyon değeri $Q(x)\geq 0$ ise bireyin R_i bölgesine, değilse R_j bölgesine sınıflandığı bu yöntemde, hatalı sınıflandırma olasılığı:

$$R_{Q(x)} = \left[1+\exp(Q(x)-\log(\hat{q}_j/\hat{q}_i))\right]^{-1} \qquad (2)$$

eşitliği ile ifade edilir.

Kovaryans matrislerinin eşit olmaması durumunda bir önceki işlemlere ilave olarak ($\Sigma_1 \neq \Sigma_2$) ise sınıflandırma bölgeleri R_1 ve R_2 şu şekilde hesaplanmaktadır:

$k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}(\mu_1{}'\Sigma_1^{-1}\mu_1 - \mu_2{}'\Sigma_2^{-1}\mu_2)$ olmak üzere,

$$R_1 = -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x + (\mu_1{}'\Sigma_1^{-1}-\mu_2{}'\Sigma_2^{-1})x - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (3)$$

$$R_2 = -\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x + (\mu_1{}'\Sigma_1^{-1}-\mu_2{}'\Sigma_2^{-1})x - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (4)$$

olur. Sınıflandırma bölgeleri x'in kuadratik fonksiyonu olarak tanımlanmaktadır. Kovaryans matrislerinin eşit olması durumunda $\Sigma_1 = \Sigma_2$ olacağından $-\frac{1}{2}x'(\Sigma_1^{-1}-\Sigma_2^{-1})x$ kuadratik terimi yok olacaktır ve sınıflandırma bölgeleri kovaryans matrislerinin eşitliğinde olduğu gibi hesaplanabilecektir.

Şayet π_1 ve π_2 yığınları çok değişkenli normal yoğunluk fonksiyonuna sahiplerse ve ortalama ve kovaryans matrisleri $\mu_1;\Sigma_1$ ve $\mu_2;\Sigma_2$ olarak kabul edilirse x_0'ın π_1 yığınına tahsis edilmesi şayet,

$$-\frac{1}{2}x_0'(\Sigma_1^{-1}-\Sigma_2^{-1})x_0 + (\mu_1{}'\Sigma_1^{-1}-\mu_2{}'\Sigma_2^{-1})x_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (5)$$

şartı sağlanırsa yapılabilecektir. Aksi takdirde x_0, π_2 yığınına tahsis edilecektir.

2. LOJİSTİK REGRESYON ANALİZİ

Lojistik Regresyon Analizi kategorik verileri analiz etmeye yarayan ve sıklıkla araştırmalarda kullanılan bir yöntemdir. Sosyal Bilimlerde yapılan araştırmalardan sağlık bilimlerinde yapılan araştırmalara, ekonomiden pazarlama ve bankacılık alanına kadar çok geniş bir alanda ilişkisel analiz yapılmasına imkân sağlar.

Çok değişkenli istatistiksel verilerin sınıflandırılmasında kullanılan yöntemlerden biri olan lojistik regresyon analizinde verilerin yapısındaki grup sayısı bilinmekte ve bu verilerden hareketle bir ayrımsama modeli oluşturulmaktadır (Ulupınar, 2007: 39).

Lojistik Regresyon Analizinde Diskriminant Analizinde belirtilen varsayımların olmaması ve bağımsız değişkenlerin kategorik olabilmesi bu tekniğin kullanımını kolaylaştırmaktadır.

Lojistik Regresyon Analizinin temel amacı diğer regresyon yöntemlerinde olduğu gibi bağımsız değişkenler ile bağımlı değişken arasındaki nedensellik ilişkisini incelemektir. Başka bir deyişle amaç en az değişken ile sonuç değişkeni ve açıklayıcı değişkenler arasındaki ilişkiyi tanımlayan kabul edilebilir modeli kurmaktır. Lojistik regresyon yönteminde bağımlı değişkenin sürekli olması gibi bir varsayım yoktur, özellikle bağımlı değişkenin iki veya daha çok kalitatif değer aldığı durumlarda kullanılır (Ulupınar, 2007: 39).

Sınıflayıcı değişkenin ölçeğine göre üç tip lojistik regresyon analizi söz konusudur:

- İkili (Binary) Lojistik Regresyon
- Sıralı (Ordinal) Lojistik Regresyon
- İsimsel (Nomial ve Multinomial) Lojistik Regresyon.

İkili Lojistik Regresyon yönteminde sınıflayıcı değişken iki sonuçludur. Bu değişken sayısal veya kısa alfanümerik bir değişken olabilir. Analizde sınıflayıcı değişken bağımlı değişken olarak referans kabul edilir ve bağımsız değişkenlerle olan ilişkisi incelenerek sınıflandırmada kullanılacak tahmini regresyon denklemi kurulur. Kurulan denklem yardımıyla sınıfların tahminine çalışılır.

Sıralı Lojistik Regresyon bağımlı değişkenin üç veya daha fazla cevaplı olması durumunda uygulanan bir yöntemdir. Ayrıca cevaplar arasında sıralı (ordinal) bir ilişki de olması gerekir.

İsimsel Lojistik Regresyon yöntemi ise Sıralı Lojistik Regresyona benzer ancak burada bağımlı değişkenin aldığı cevapların sıralı olması şartı aranmamaktadır.

Bu çalışmada İkili Lojistik regresyon yöntemi kullanıldığından bahse konu tekniğin matematiksel altyapısı bu bölümde izah edilecektir.

Çeşitli gösterim biçimleri olan genel doğrusal regresyon modeli,

$$E(y_{.} / x_{i1},...,x_{ip}) = \sum_{k=0}^{p} \beta_k x_{ik} ; i = 1,...,n \text{ için} \quad (6)$$

biçiminde koşullu beklenen değer olarak da yazılması mümkündür. Bu modelde açıklayıcı değişkenler üzerinde kısıt yok iken sabit varsayılır, onun için y bağımlı değişkeninin sürekli olması koşulu vardır. Herhangi bir i'inci gözlem için,

$$y_i = \sum_{k=0}^{p} \beta_k x_{ik} + u_i \quad (7)$$

biçiminde ifade edilen modelde açıklayıcı değişkenler üzerinde bir kısıt olmadığından y_i sonuç değeri $-\infty$ ile $+\infty$ arasında tüm değerleri alabilmektedir. Bağımlı değişkenin 0, 1 gibi değerler aldığı durumda bu kural bozulmakta ve $P(y_i=1)$, i'inci gözlemin 1 değerini alma olasılığı olmak üzere, beklenen değer,

$$E(y_i) = P(y_i = 1) = \sum_{k=0}^{p} \beta_k x_{ik} \quad (8)$$

olarak bulunur. Sol tarafı 0 ile 1 arasında değerleri alan bu denkleme doğrusal olasılık modeli adı verilmektedir (Tatlıdil, 1996: 290).

Açıklayıcı değişkenlerin sınırsız değerler alması nedeniyle söz konusu eşitlik her zaman sağlanmaktadır. Bu sebeple çeşitli dönüşümler yapılmaktadır. Bu dönüşümlerden en yaygın olarak kullanılan iki tanesi logit ve probit dönüşümlerdir.

Logit dönüşümde doğrusal olasılık modelinde olasılık değerleri üzerinde P/(1-P) dönüşümü yapılarak sonuç değişkeninin sınırları 0, $+\infty$ yapılmakta, daha sonra ise bu oran değerinin doğal logaritması alınarak sonuç değişkenin sınırları $-\infty$, $+\infty$ yapılmaktadır. Bu dönüşümlerden sonra elde edilen yeni fonksiyon,

$$E(y_i) = L_i = \log(P_i/(1-P_i)) = \sum_{k=0}^{p} \beta_k x_{ik} \qquad (9)$$

olarak yazılmaktadır. Lojistik model ya da kısaca logit olarak bilinen bu modelde P_i olasılık değeri,

$$P_i = \exp(\sum_{k=0}^{p} \beta_k x_{ik})/(1+\exp(\sum_{k=0}^{p} \beta_k x_{ik})) \qquad (10)$$

biçiminde tanımlanmakta ve lojistik fonksiyon adını almaktadır. Bu modelde sonuç değişkeninin iki değer alması nedeni ile hata terimi sıfır ortalama ve P(1-P) varyanslıdır. Hata terimi bu parametrelerle binom dağılımlı olup, analiz bu teorik temele dayanmaktadır.

Logit fonksiyonu aynı zamanda şu şekilde de gösterilmektedir:

$$P_i = \frac{1}{1+e^{-(\alpha+\beta x)}} \qquad (11)$$

Bu eşitliğe lojistik dağılım fonksiyonu adı verilir. $\alpha + \beta x = Z$ olarak kabul edilirse bu durumda,

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i} \qquad (12)$$

eşitliğine ulaşılır. Bu eşitlik odds (bahis) oranı olarak adlandırılır. Odds oranı daha özet bir ifadeyle olayın gerçekleşme olasılığının olayın gerçekleşmeme olasılığına olan oranını ifade etmektedir. Odds oranından genellikle ikili değişken arasındaki ilişkinin ölçümlenmesinde yararlanılır. Etki katsayısı veya etki büyüklüğü olarak tanımlanan Exp(β), aynı zamanda Odds oranını vermektedir ve bu değer açıklayıcı değişkenlerin etkisinin kolayca yorumlanabilmesi açısından önemlidir.

Odds oranının doğal logaritması alınırsa Logite ulaşılır. Yani odds oranının logaritması katsayı tahminleri bakımından yalnız X'e göre değil ana kütle katsayılarına göre de doğrusaldır (Gujarati, 2001: 555). Ayrıca odds oranları, x'in arttığı her birim için e^{β}'nın katları kadar artar.

Böylece doğrusal olmayan ilişki logit fonksiyonu yardımıyla doğrusal hale getirilmiştir.

3. SINIFLANDIRMA VE TAHMİN YÖNTEMLERİNİN KULLANILDIĞI BİLİMSEL ÇALIŞMALAR

Diskriminant Analizinin sınıflandırma ve tahmin etmede kullanıldığı ikinci bölümde izah edilmişti. Bu bölümde değişik alanlarda diskriminant analizinin uygulamaları özet bir şekilde ifade edilecektir.

Balcaen ve Ooghe (2005) iş yaşamındaki başarısızlıkların sınıflandırılmasında son 35 yılda kullanılan istatistiksel teknikler ve bu tekniklere ilişkin problemleri yaptıkları çalışmada ele almışlardır. Yaptıkları çalışmada Çoklu Diskriminant Analizi, Logit Modeller Şartlı Olasılık Modelleri ve tek değişkenli analiz yöntemlerini karşılaştırmışlardır.

Sueyoshi (2004) Diskriminant analizi ile standart tam sayılı programlama modelleri ve iki aşamalı tam sayılı programlama modellerini kullanarak sınıflandırma başarılarını incelemiştir. Japon bankalarından elde ettiği veriler üzerinde de uygulamasını yapmıştır.

Berg (2007) doğrusal diskriminant analizi, genelleştirilmiş doğrusal modeller ve yapay sinir ağlarını kullanarak firmaların iflas tahminlerini yapmaya çalışmıştır.

Çilan v.d. (2009) Avrupa Birliği üyesi olan ve olmayanlar arasındaki dijital ayrımın analizini Diskriminant analizi ile yapmışlarıdır. Yaptıkları analizde sınıflandırma başarısı %74,1 ile başarılı bulunmuştur. Analiz öncesi normallik varsayımının testi yapılmıştır.

Bosse (2008) çoklu diskriminant analizi ile küçük firmaların borç alırken kredibilitesinin ayrıştırılmasını modellemiş ve %86,6'lık bir sınıflandırma başarısı elde etmiştir.

Wu v.d.(2008) Çin kamu şirketlerinin finansal olumsuzluklarının analizini olasılıklı yapay sinir ağları ve diskriminant analizi kullanarak yapmıştır. Kısa dönem tahminlerde çoklu diskriminant analizi ile %81,25, uzun dönem tahminlerde %56,25'lik bir sınıflandırma başarısı elde etmiştir. Buna karşılık yapay sinir ağları ile yapılan analiz sonucunda kısa dönemde %87,5'lik, uzun dönemde %81,25'lik bir sınıflandırma başarısı elde etmişlerdir.

Chen v.d. (2008) iletişim aracının seçimi ile ilgili olarak belirlenen kriterleri diskriminant analizi ile analiz etmiş ve tahmin modeli geliştirerek değişkenler arasındaki ilişkiyi incelemiştir.

Pompe ve Bilderbeek (2005) küçük ve orta ölçekli sanayi firmalarının iflasını tahmin etmede çoklu diskriminant analizi ile geliştirdiği diskriminant modelini kullanmıştır.

4. ARAŞTIRMA

Bu çalışmanın amacı çok değişkenli istatistiksel sınıflandırma ve tahmin yöntemlerinden ikisini karşılaştırılarak kullanılan veri setine göre en iyi yöntemin belirlenmesidir. Bu çalışmanın literatüre olan en önemli katkısı diskriminant analizi ve lojistik regresyon analizinin metodolojik olarak gösterilmesi suretiyle araştırmacılara ışık tutabilecek olmasıdır.

4.1. Araştırmanın Kapsamı

Araştırmada 155 ülkenin değerleri kullanılarak diskriminant analizi ve lojistik regresyon analizi yapılmış ve sonuçlar elde edilmiştir. Değişken değerleri dikkate alındığında toplam 155 ülkenin 35'i kayıp verilerden dolayı işleme dâhil edilmemiş ve 120 ülke ile sınıflandırma süreci yürütülmüştür.

Analizde bağımlı değişken olarak çok gelişmiş (1) ve orta düzeyde gelişmiş (2) ülke sınıflandırması kullanılmıştır. Başlangıçta 28 değişkene göre analiz çalışmalarına başlanmış ancak bu değişkenlerin kullanılması ile ülke sayısı 74'e düştüğünden 12 değişken analiz dışı bırakılarak 120 ülkenin analize alınmasına çalışılmıştır. İnsani Kalkınma Endeksinde özellikle gelir, eğitim, yaşam beklentisi değişkenleri olmazsa olmaz olarak kullanıldığından 16 değişkenden daha fazla ödün verilmemesi gerektiği düşünülmüş ve analizde bu 16 değişkenin kullanımına karar verilmiştir. Bağımsız değişkenler Tablo 1'de görülmektedir.

Ayrıca 120 ülke içerisinde çok gelişmiş 56 ülke ve orta düzeyde gelişmiş 64 ülke bulunduğundan çok gelişmiş ülkelerin grup önceliği %46,7 ve orta düzeyde gelişmiş ülkelerin grup önceliği %53,3'tür. Bu olasılıklar daha sonra sınıflandırma oranının değerlendirilmesinde kullanılacaktır. Lojistik Regresyon Analizinde ihtiyaç olmamasına karşın diskriminant analizi ile ilgili olarak analize başlamadan önce Normallik varsayımı, kovaryans matrislerinin eşitliği varsayımı ve çoklu bağlantı varsayımı incelenecek, daha sonra sınıflandırma sonuçları dikkate alınacaktır.

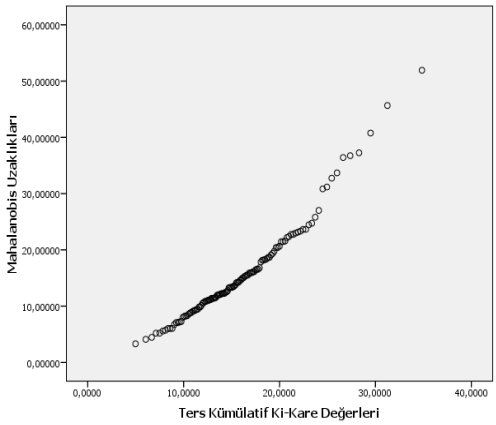
Tablo 1: Seçilen Beşeri Kalkınmışlık Sınıflandırma Değişkenleri

	Etiket	N
Toplam Nüfus(2005)	IV1	154
Kırsal Nüfus(2005)	IV2	155
Kadın Parlamenter Oranı (Toplamın Yüzdesi)	IV3	150
Sağlık Harcamaları Kamu (GSYİH'nın yüzdesi)(2007)	IV4	154
Sağlık Harcamaları Özel (GSYİH'nın yüzdesi)(2007)	IV5	154
Sağlık Harcamaları Kişi Başına (Satın Alma Gücü Paritesine göre US$)(2007)	IV6	152
Doğumda Yaşam Beklentisi(2002-2005)	IV7	151
İlköğretime net kayıt oranı	IV8	141
1000 kişiye düşen telefon hattı sayısı (2005)	IV9	152
1000 kişiye düşen cep telefonu abonesi sayısı (2005)	IV10	154
1000 kişiye düşen internet kullanıcısı sayısı (2005)	IV11	153
GSYİH (Milyar Dolar) (2005)	IV12	152
İthal Edilen Mallar (GSYİH %'si olarak) (2005)	IV13	146
İhraç Edilen Mallar (GSYİH %'si olarak) (2005)	IV14	147
Elektrik Tüketimi (Kw-H olarak)(2004)	IV15	151
Hapiste Bulunan Kişi Sayısı (2007)	IV16	155
Geçerli N		120

4.2. Varsayımların Test Edilmesi

Diskriminant analizi ile ilgili literatürde de bahsedildiği gibi çok önemli üç temel varsayım analiz öncesi araştırılmakta, elde edilen değerlere göre analiz yapılmamakta veya farklı yöntemler kullanılarak analize devam edilmektedir. Bu varsayımların başında çok değişkenli normallik, kovaryans matrislerinin eşitliği ve çoklu bağlantı varsayımı gelmektedir. Varsayımların sağlanamamasının elde edilecek sınıflandırma sonuçları açısından sorun yaratacağı ve arzu edilen yüksek oranlarda sınıflandırma yapılamayacağı literatürde ifade edilmektedir.

Öncelikle tek değişkenli normallik testleri yapılmış ve normal dağılmayan değişkenler logaritmik dönüşüme tabi tutularak normal hâle getirilmiştir. Bilahare çok değişkenli normallik testi Sharma (1996, 380-382)'nın ifade ettiği gibi Mahalanobis uzaklıkları kullanılarak yapılmıştır. Sonuçta ters kümülatif ki-kare değerleri ile Mahalanobis uzaklıkları arasında 0,979'luk yüksek bir korelasyon olduğu tespit edilmiştir. Korelasyonu gösteren serpilme diyagramı Şekil 1'de görülmektedir.



Şekil 1: Mahalanobis Uzaklıkları ve Ki-Kare Değerleri Korelâsyon Diyagramı

Şekil 1'den de görülebileceği gibi korelasyonun 0,979 olması grupların dağılımının çok değişkenli normalliğe uyduğunu göstermektedir. Kovaryans matrislerinin eşit olması durumunda Doğrusal Diskriminant Analizi yapılabilirken kovaryans matrislerinin eşit olmaması durumunda Kuadratik Diskriminant analizi yapılarak sınıflandırma sonuçları elde edilebilmektedir.

Kovaryans matrislerinin eşitliği için Box's M testi kullanılmıştır. Grup içi kovaryans matrisleri (within Groups) seçeneği kullanıldığında kovaryans matrisi eşitliği sağlanamamıştır ($p<0,05$). Bu sebeple Kuadratik Diskriminant Analizi kullanılması için kovaryans matrislerinin ayrı gruplar olması gerektiği (seperate-groups) seçeneği işaretlenerek tekrarlanmış ve anlamlılık değeri 0,703 bulunmuştur. Bulunan değerler Tablo 2'de bulunmaktadır.

Tablo 2: Box's M Test Sonuçları

Box's M		**0,147**
F	Yaklaşık.	0,146
	Sd1	1,000
	Sd2	41262,347
	Anl.	0,703

Ayrı gruplar için kovaryans matrislerinin eşitliği Box's M testi ile sınanmıştır anlamlılık değeri 0,05'ten ($p>0,05$) büyük olduğu için sıfır hipotezi kabul edilerek ayrı gruplar için kovaryans matrislerinin eşit olduğu

tespit edilmiştir. Kovaryans matrislerinin eşitliği gruplar arasında sağlanamadığından kuadratik değerler dikkate alınmıştır.

Çoklu doğrusallık testinde VIF ve Tolerans değerlerinin incelenmesinde VIF değerlerinin 10'dan küçük olduğu ve Tolerans değerlerinin 0,30'un üzerinde olduğu gözlenmektedir. Bu durum çoklu doğrusal ilişkinin olmadığı yönünde yorumlanabilmektedir. Ayrıca t değerlerinin çok küçük değer almasının da çoklu doğrusallık sorununa işaret ettiği bazı yazarlarca ifade edilmektedir. Yapılan analiz neticesinde t değerlerinden 0'a çok yakın değerler bulunmadığı da ayrıca gözlenmiştir.

4.3. SPSS Kullanılarak Elde Edilen Araştırma Sonuçları

4.3.1. Diskriminant Analizi Sonuçları

Başlangıçta belirlenen iki grup (Çok Gelişmiş ve Orta Düzeyde Gelişmiş) olduğu için 1 Diskriminant fonksiyonu türetilmiştir. Özdeğerin (Eigenvalue) büyük olması bağımlı değişkendeki varyansın daha büyük bir kısmının elde edilen fonksiyon tarafından açıklanabildiğini göstermektedir. Kesin bir değer olmamakla birlikte 0,40'ın üzerindeki değerler iyi olarak kabul edilmektedir. Tablo 3'te görülebileceği gibi modelde özdeğer 2,385 bulunmuş ve varyansın %100'ünü açıklamaktadır. Ayrıca Kanonik Korelâsyon Katsayısı 0,839 olarak bulunmuştur. Katsayının karesi (r^2) 0,704'dür. Bağımsız değişkenlerin bağımlı değişkeni %70,4 oranında açıkladığı söylenebilir.

Tablo 3: Özdeğerler Çizelgesi

Fonksiyon	Özdeğer	% Varyans	Kümülatif %	Kanonik Korelâsyon
1	2,385[a]	100,0	100,0	0,839

Tablo 4: Wilk's Lambda Değeri

Test Edilen Fonksiyon	Wilks' Lambda	Ki-Kare	Sd	Anl.
1	0,295	134,142	16	0,000

Wilk's Lambda istatistiği, diskriminant skorlarının gruplar arasındaki toplam varyansın gruplar arasındaki farklar tarafından açıklanamayan kısmını (oranını) göstermektedir. Modelde 0,295 yani toplam varyansın %29,5'u gruplar arasındaki farklar tarafından açıklanamamaktadır.

Elde edilen bir adet fonksiyon ve fonksiyon içerisinde bulunan değişkenlerin katsayıları Tablo 5'te bulunmaktadır.

Tablo 5: Standartlaştırılmamış Diskriminant Fonksiyonu Katsayıları

	Fonksiyon
	1
LogIV1	0,922
LogIV6	2,676
LogIV9	-0,528
LogIV11	0,160
LogIV12	-0,495
LogIV13	-0,565
LogIV14	0,385
LogIV15	0,440
LogIV16	-0,424
IV2	-0,002
IV3	0,001
IV4	-0,127
IV5	-0,189
IV7	0,045
IV8	-0,019
IV10	0,002
(Constant)	-7,119

Ülkelerin gelişmişlik düzeylerinin belirlenmesinde kullanılan diskriminant fonksiyonu Z gelişmişlik düzeyini belirlemek üzere (1,2) şu şekilde oluşturulmuştur:

Z*=-7,119**+**0,922(LogIV1)+**2,676***(LogIV6)**-0,002***(IV2)+**0,01***(IV3)**-0,127***(IV4)-**0,189***(IV5)+**0,045***(IV7)**-0,019***(IV8)-**0,528***(IV9)**+0,002***(IV10)+**0,160***(LogIV11)-**0,495***(LogIV12)-**0,565***(LogIV13)+**0,385***(LogIV14)+**0,440***(LogIV15)**–0,424***(LogIV16)

Modelden görülebileceği gibi 1 birimlik artış ile bağımlı değişken üzerinde en büyük etki yaratan değişken LogIV6-Kişi Başına Düşen Satın Alma Gücü Paritesine göre Sağlık Harcamaları'dır. 1 birimlik artışla 2.676'lık pozitif bir etki yaratmaktadır. LogIV13-İthalatın yüksek oluşunun negatif etkisi olduğu LogIV14-ihracatın ise pozitif etki yarattığı

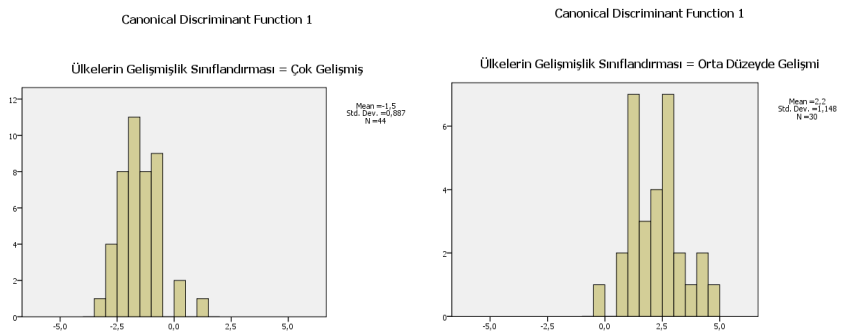
görülmektedir. Ayrıca elektrik tüketiminin ve iletişim değişkenlerinin de pozitif etki yarattığı söylenebilir.

Tablo 6: Sınıflandırma Sonuçları

		Ülkelerin Gelişmişlik Sınıflandırması	Tahmin Edilen Grup Üyeliği		
			Çok Gelişmiş	Orta Düzeyde Gelişmiş	Toplam
Orjinal	Sayılan	Çok Gelişmiş	51	5	56
		Orta Düzeyde Gelişmiş	4	60	64
	%	Çok Gelişmiş	91,1	8,9	100,0
		Orta Düzeyde Gelişmiş	6,2	93,8	100,0

Oluşturulan modelin Tablo 6'da görülebileceği gibi %92,5'lik toplam sınıflandırma oranı ile başarılı bir sınıflandırma yaptığı söylenebilir. Ancak bu sınıflandırmanın doğruluğunun test edilmesi maksadıyla nispi şans kriteri ve maksimum şans kriterinin hesaplanarak karşılaştırılması gerekmektedir. Hesaplamaya alınan örneklem büyüklüğü 120'dir. Dolayısıyla Çok Gelişmiş grup örneklemin %59'unu, orta gelişmiş grup ise %41'ini oluşturmaktadır. Şans değeri çok gelişmiş grubun seçilme yüzdesi yani 0,59 ve orta gelişmiş grubun seçilme ihtimali yani 0,41'dir. Burada maksimum şans kriteri 0,59'dur. Nispi şans kriteri ise $(0,59)^2+(0,41)^2=0,5162$'dir.Diskriminant analizi sonucunda elde edilen sınıflandırma oranı bu değerlerin çok üzerindedir.

Sınıflandırmada hatalı sınıflandırılan 9 ülke bulunmaktadır. Çok gelişmiş ülke grubunda iken elde edilen diskriminant modeli ile orta gelişmiş düzeyde olarak 5 ülke, Uruguay, Meksika, Panama, Beyaz Rusya ve Arnavutluk bulunmuştur. Orta gelişmiş ülkeler arasında olup ta diskriminant analizinin çok gelişmiş ülke grubuna ise 4 ülkeyi, Türkiye, Kolombiya, Tunus ve Jamaika'yı atadığı gözlenmektedir. Değişkenlerin sıra numaraları aynı zamanda Birleşmiş Milletler Kalkınma Programında belirlenen gelişmişliğe bağlı sıra numaralarıdır. Bu durumda hatalı sınıflandırılan ülkelerin çok gelişmiş ülkeler ile orta düzeyde gelişmiş ülkelerin sınır noktasına yakın olmasının yapılan analizin Birleşmiş Milletler Kalkınma Programınca yapılan sınıflandırmaya çok yakın bir çalışma olduğunu göstermektedir. Analizde Türkiye'nin belirlenen bağımsız değişken değerlerine göre Çok Gelişmiş olarak sınıflandırılırken Birleşmiş Milletler Kalkınma Programı tarafından Orta Gelişmişlik düzeyinde sınıflandırılmasının da anlamlı bir sonuç olabileceği düşünülmektedir.



(a) (b)

Şekil 2: Çok Gelişmiş (a) ve Orta Düzeyde Gelişmiş (b) Ülke Gruplarının Dağılım Grafiği

4.3.2. Lojistik Regresyon Sonuçları

Lojistik regresyon analizinde literatürde de belirtildiği gibi varsayımlar yoktur. Bu sebeple analize doğrudan hiçbir varsayım araştırması yapılmadan başlanacaktır. Yapılacak analizde Lojistik Regresyon modeline değişkenlerin alınış yöntemlerine göre En Küçük Kareler yöntemi kullanılmıştır. Sonrasında modelde kullanılan değişkenlerin önem derecesinin tespit edilebilmesi ve daha az değişkenle sınıflandırma yapabilecek modelin elde edilebilmesi için ileri ve geri adımsal olabilirlik oranı (likelihood ratio) yöntemleri kullanılmıştır. Sonuçlar ayrı ayrı gösterilerek yorumlanmıştır.

4.3.2.1. EKK Yöntemi ile Yapılan Lojistik Regresyon Analizi

Kullanılan 155 ülkelik örneklemin Diskriminant Analizinde olduğu gibi kayıp verilerden dolayı 120'si Lojistik Regresyon Analizinde kullanılmıştır. Lojistik Regresyon Analizinde de bağımlı değişken olarak çok gelişmiş ve orta düzeyde gelişmiş ülke sınıflandırması kullanılmıştır. Ayrıca bağımsız değişken olarak ta 16 sürekli bağımsız değişken analize alınmıştır.

Başlangıç durumunda Lojistik Regresyon analizinde referans olarak Çok Gelişmiş grup için 0 ve Orta Düzeyde gelişmiş için 1 kodlanmıştır. Bunun sebebi orta düzeyde gelişmiş ülkelere ait şans kriterinin yüksek oluşudur. Bu sayede Orta düzeyde gelişmiş ülkelerin referans olarak

seçilmesi ile sınıflandırma başarısı analiz başlangıcından itibaren yüksek tutulabilecektir.

Tablo 7: İterasyon Geçmişi

İterasyon		-2 Log likelihood	Katsayılar
			Sabit
Adım 0	1	165,822	0,133
	2	165,822	0,134

4.3.2.2. Analiz sonucu elde edilen lojistik regresyon modeli

Lojistik Regresyon Analizi Sonucunda aşağıdaki denklem elde edilmiştir:

ln(Z)= 1585,67-0,268*Toplam Nüfus(2005)+ 0,810*Kırsal Nüfus(2005)-0,634*Kadın Parlamenter Oranı (Toplamın Yüzdesi)-10,510*Sağlık Harcamaları Kamu (GSYİH'nın yüzdesi)+16,708*(2004 Sağlık Harcamaları Özel (GSYİH'nın yüzdesi) (2004) -0,053*Sağlık Harcamaları Kişi Başına (Satın Alma Gücü Paritesine göre US$)(2004)-14,665*Doğumda Yaşam Beklentisi(2002-2005)-4,683*İlköğretime net kayıt oranı-0,154*1000 kişiye düşen telefon hattı sayısı (2005) +0,006*1000 kişiye düşen cep telefonu abonesi sayısı (2005)+0,021*1000 kişiye düşen internet kullanıcısı sayısı (2005)+ 0,088*GSYİH (Milyar Dolar) (2005) –0,683*İthal Edilen Mallar (GSYİH %'si olarak) (2005)+0,705*İhraç Edilen Mallar (GSYİH %'si olarak) (2005)- 0,039*Elektrik Tüketimi (Kw-H olarak)(2004)+0,000*Hapiste Bulunan Şahıs Sayısı

Burada en önemli etkinin özel sağlık harcamaları, kamu sağlık harcamaları ve doğumda yaşam beklentisi tarafından yapıldığı dikkat çekicidir. Bazı değişkenlerin katsayıları ise ihmal edilecek derecede düşük hesaplanmıştır.

4.3.2.3. Modelin Anlamlılığının Test Edilmesi

Geleneksel Ki-Kare metodu kullanılarak modelin anlamlılığı bu aşamada test edilmektedir. Bağımlı değişkenin bağımsız değişkenler tarafından hep birlikte kullanılması ile test edilebilmesi durumu da burada test edilmektedir. Tablo 12'deki değerler incelendiğinde analizin doğrudan enter yöntemi ile yapılması nedeniyle Adım, Blok ve Model Ki-kare

değerlerinin aynı olduğu görülebilmektedir. Şayet Adımsal (Stepwise) yöntem kullanılsaydı bu durum her adım için değişebilecekti. Yapılan analiz neticesinde önem derecelerinin 0,05'ten küçük olması (p<0,05) nedeniyle modelin anlamlı olduğu söylenebilir.

Tablo 12: Model Katsayıları için Omnibus Testi

		Ki-Kare	df	P
Adım 1	Adım	165,822	16	0,000
	Blok	165,822	16	0,000
	Model	165,822	16	0,000

Ayrıca Modelin uygunluğunun test edilmesinde Hosmer ve Lemeshow testi de kullanılmaktadır. Tablo 13'te elde edilen değerler görülebilmektedir.

Tablo 13: Hosmer ve Lemeshow Uyum İyiliği Testi Sonuçları

Adım	Ki-Kare	Sd	P
1	0,000	5	1,000

Hosmer ve Lemeshow testinde modelde tahmin edilen değerlerle gerçekte gözlenen değerler arasında fark yoktur sıfır hipotezi test edilmektedir.

H_0 : Teorik model verileri iyi temsil etmektedir.

H_A : Teorik model verileri iyi temsil etmemektedir.

Elde edilen önem derecesinin 1 olması sebebi ile (p>0,05) modelin tahmin ettiği verilerin istenen anlamlılık düzeyinde kabul edilebilir olduğu söylenebilir. Hosmer ve Lemeshow Uyum İyilik testi model ile elde edilen tahminlerin gerçek gruplardan anlamlı bir fark oluşturmadığını göstermektedir. Bu bağımsız değişkenin varyans açıklama yüzdesini göstermez ancak en azından açıklamanın anlamlı olduğunu ifade eder. Örneklem büyüklüğü arttıkça Hosmer ve Lemeshow Uyum İyiliği testi daha hassas değerlerle daha küçük farkları bulabilir.

Tablo 14: Modelin Özeti

Adım	-2 Log likelihood	Cox & Snell R Kare	Nagelkerke R Kare
1	0,000[a]	0,749	1,000

Tablo 14'te de görülebileceği gibi 25'inci iterasyonda en uygun model bulunmuştur. Ancak bu sonucun tek olduğunun düşünülmemesi gerektiği SPSS tarafından uyarı olarak verilmektedir. –2 Log Likelihood istatistiği modelin ne kadar güçlü veya zayıf kararlar verebileceğini ifade etmektedir. Yüksek değerler alması durumunda modelin zayıf olduğu, çok küçük değerler aldığında ise modelin iyi olduğu belirtilmektedir. Çalışılan modelde –2 Log Likelihood istatistiği 0 olduğundan modelin iyi olduğu söylenebilir. Cox ve Snell R Kare istatistiği regresyondaki R kare değeri ile aynı yorumlanabilir. Yani bağımsız değişkenlerin bağımlı değişkeni %74,9 açıklayabildiği söylenebilir. Ancak Cox ve Snell R Kare istatistiği asla 1 değerini alamaz, bu durum da yorum yapmayı güçleştirebilir. Nagelkerke R Kare istatistiği ise Cox ve Snell R Kare istatistiğinin modifiye edilmiş halidir. Nagelkerke R Kare istatistiği genelde Cox ve Snell R Kare istatistiğinden yüksek bir değer almaktadır. Modelde Nagelkerke R Kare istatistiği 1'dir. Yani bağımsız değişkenlerin bağımlı değişkeni %100 açıklayabildiği söylenebilir.

4.3.2.4. Sınıflandırma Sonuçları

Tablo 15: Hosmer ve Lemeshow Kontenjans Tablosu

		Ülkelerin Gelişmişlik Sınıflandırması = Çok Gelişmiş		Ülkelerin Gelişmişlik Sınıflandırması = Orta Düzeyde Gelişmiş		Toplam
		Gözlenen	Beklenen	Gözlenen	Beklenen	
Adım 1	1	12	12,000	0	0,000	12
	2	12	12,000	0	0,000	12
	3	12	12,000	0	0,000	12
	4	12	12,000	0	0,000	12
	5	8	8,000	4	4,000	12

	6	0	0,000	8	8,000	8
	7	0	0,000	52	52,000	52

Tablo 15'de Hosmer ve Lemeshow Kontenjans Tablosunda birinci adımda örneklemin 7 grup halinde oluşturularak yapılan tahminler görülebilmektedir.

Tablo 16: Lojistik Regresyon Analizi Sınıflandırma Sonucu

		Frekans	Yüzde	Geçerli Yüzde	Kümülatif Yüzde
Geçerli	Çok Gelişmiş	56	36,1	46,7	46,7
	Orta Düzeyde Gelişmiş	64	41,3	53,3	100,0
	Toplam	120	77,4	100,0	
Kayıp	Sistem	35	22,6		
Toplam		155	100,0		

Tablo 16'da EKK yöntemi ile değişkenlerin modele alındığı Lojistik Regresyon Analizinde %100'lük bir sınıflandırma başarısı olduğu görülebilmektedir. Ayrıca uç değerlerin (outliers) analizi yapılmış ve uç değer bulunmadığı için örneklemden herhangi bir denek çıkarılmamıştır.

4.3.3. İleri Adımsal Olabilirlik Yöntemi ile Lojistik Regresyon Analizi

Lojistik Regresyon analizi ile yüksek bir sınıflandırma oranı yakalanmasının yanı sıra modele etki eden önemli değişkenlerin neler olduğu ve bu değişkenler yardımıyla sınıflandırma başarısının ne olabileceği konusu yapılacak analizle ortaya konmaya çalışılacaktır. Sonuçta lojistik regresyon analizinde değişken sayısının azaltılarak modelin yorumlanabilirliğinin kolaylaştırılması da ayrı bir amaç olarak değerlendirilmektedir.

İleriye adımsal olabilirlik yönteminde değişkenlerin modele alınmasında 0,15 ve çıkarılmasında 0,25 olasılığı kullanılmıştır. Bu analizde de yalnızca modele ilave edilen değişkenler ve sınıflandırma sonucu burada gösterilecektir.

Lojistik Regresyon Analizi sonucunda Tablo 17'de bulunan değişkenler modelde kullanılmıştır. Denklemde kırsal nüfus (kırsalnüf), Sağlık Harcamaları Kamu (GSYİH'nın yüzdesi)(2007) (Saglık1), Doğumda Yaşam Beklentisi (YasamBek), İlköğretime net kayıt oranı (Egitim3), 1000 kişiye düşen telefon hattı sayısı (2005) (İletisim1) ve Elektrik Tüketimi (elektriktuk) değişkenleri kullanılmıştır.

Tablo 17: Modelde Kullanılan Değişkenler

								95,0% C.I. EXP(B)	
		B	S.E.	Wald	Df	Anl.	Exp(B)	Alt	Üst
Adım 9	IV2	1,655	39,444	0,002	1	0,967	5,231	0,000	1,965E34
	IV4	-31,080	675,622	0,002	1	0,963	0,000	0,000	.
	IV7	-27,656	499,266	0,003	1	0,956	0,000	0,000	.
	IV8	-9,243	171,574	0,003	1	0,957	0,000	0,000	1,072E142
	IV9	-0,342	9,688	0,001	1	0,972	0,710	0,000	1,252E8
	IV15	-0,068	1,203	0,003	1	0,955	0,935	0,088	9,881
	Sabit	3076,479	54108,286	0,003	1	0,955	.		

Kullanılan değişkenler sonucunda aşağıdaki regresyon denklemi elde edilmiştir:

ln(Z)= 3076,479+1,655*IV2(Kırsal Nüfus)-**31,080***IV4(GSYİH'nın yüzdesi olarak Kamu Sağlık Harcamaları(2005))-**27,656***IV7(Doğumda Yaşam Beklentisi(2002-2005))-**9,243***IV8(İlköğretime Net Kayıt Oranı)-**0,342***IV9(1000 kişiye düşen telefon hattı sayısı (2005)) -**0,068***IV15(Elektrik Tüketimi (Kw-H olarak)(2004))

Denkleme 9'uncu adımda değişkenler dâhil edilmiş ve bu değişkenler kullanılarak sınıflandırma sonucu bulunmuştur. Değişkenler incelendiğinde kırsal nüfusun pozitif bir etki yarattığı, diğer sağlık, eğitim, iletişim ve enerji değişkenlerinin negatif etki yarattığı gözlenmektedir. Başlangıçta sayıca çok olan grubun şans kriteri yüksek olduğu için 1 olarak belirlendiği ifade edilmişti. Yani yüksek gelişmiş ülke grubunda model 0 ve 0'a yakın sonuçları dikkate alacak, orta düzeyde gelişmiş grupta ise 1 ve 1'e yakın değerleri dikkate alacaktır. Bu açıklama ile denklem katsayılarının

özellikle negatif etki yaratanların çok gelişmiş ülke grubuna sınıflandırmada etkili olduğu söylenebilir. En yüksek etkinin sağlık ve yaşam beklentisi değişkenleri tarafından olduğu görülebilmektedir.

Yapılan incelemede kurulan modelin uyum iyiliği ve değişkenlerin anlamlılığı test edilmiş sonuçları olumlu çıkmıştır. Sınıflandırma sonucunda ise Tablo 18'de görülebileceği gibi %100'lük bir başarı elde edilmiştir.

Tablo 18: İleri Adımsal Olabilirlik Yöntemi Sınıflandırma Sonucu

			Tahmin Edilen		
			Ülkelerin Gelişmişlik Sınıflandırması		
	Gözlenen		Çok Gelişmiş	Orta Düzeyde Gelişmiş	Yüzdesi Doğru
Adım 9	Ülkelerin Gelişmişlik Sınıflandırması	Çok Gelişmiş	56	0	100,0
		Orta Düzeyde Gelişmiş	0	64	100,0
		Toplam Yüzdesi			100,0

SONUÇ

Sınıflandırma sosyal bilimlerde yapılan araştırmalarda birçok metot kullanılarak yapılmakta ve yapılan bu sınıflandırma sonuçlarına göre istatistiksel olarak çıkarımlar ortaya konmaktadır. Kullanılan veri seti, ülkelerin İnsani Kalkınmışlık Endeksine göre Birleşmiş Milletler tarafından yapılan çok gelişmiş ve orta düzeyde gelişmiş ülke sıralamasında yararlandığı verilerdir. Bu çalışmada endeks değerleri yerine endeksleri oluşturan ham veriler kullanılmış ve ülkeler yeniden sınıflandırılmıştır.

Diskriminant Analizinde normallik varsayımı, kovaryans matrislerinin eşitliği varsayımı ve çoklu bağlantı varsayımı sınanmıştır. Normallik varsayımının sınanmasında bazı değişkenlerin tek değişkenli normal dağılım göstermediği için normallik dönüşümleri yapılmış ve bilahare yapılan testlerde çok değişkenli normal dağılım ortaya konulmuştur. Kovaryans matrisleri eşitliği varsayımı sağlanamadığından Doğrusal Diskriminant Analizi yerine Kuadratik Diskriminant Analizi kullanılmıştır.

Çoklu bağlantı probleminin olmadığı yapılan analizler neticesinde gösterilmiştir.

Diskriminant analizi sonucunda bağımsız değişkenlerin bağımlı değişkeni %70,39 açıklayabildiği görülmüştür. Kullanılan bağımsız değişkenlerin önem değerlendirmesinde iletişim verilerinin, doğumda yaşam beklentisinin, GSYİH'nın yüzdesi olarak ihraç edilen malların, kamu sağlık harcamalarının çok gelişmiş ülke olabilmede pozitif etki yarattığı tespit edilmiştir. Ülkelerin gelişmişliklerinin belirlenmesinde bu durumun gerçekçi olduğu düşünülmektedir. Zira gelişmiş ve üretim yapan ülkelerde ihracatın yüksek olması, iletişimin daha iyi tesis edilmiş olması, sağlık yatırımlarının ve buna bağlı olarak doğumda yaşam beklentisinin üst seviyede görülmesi olağan bir durumdur. Kırsal nüfus, ithal edilen mallar ve ilköğretime net kayıt oranı değişkenlerinin ise modele negatif etkide bulundukları da tespit edilmiştir. Bu durumun da rasyonel olduğu değerlendirilmektedir. Çok gelişmiş ülke konumunda bir ülkenin sınıflandırılmasında diskriminant analizi sonuçları dikkate alındığında ihracat, iletişim ve sağlık yatırımlarının artırılmasının önemli olduğu açıktır. Yani sağlıklı bireylerin koordineli bir şekilde el ele vererek üretime dönük çalışması bir ülkeyi çok gelişmiş ülke kategorisine taşıyabilecektir.

Diskriminant analizi ile yapılan sınıflandırmada %92,5'lik bir sınıflandırma başarısı elde edilmiştir. Hatalı sınıflandırılan ülkeler incelendiğinde bu ülkelerin Birleşmiş Milletler tarafından yapılan sıralamada çok gelişmiş ülkeler ve orta düzeyde gelişmiş ülkeler sınırına yakın ülkeler arasında olduğu görülmektedir. Bu durum hazırlanan modelin Birleşmiş Milletler Kalkınma Programı tarafından yapılan sıralamaya çok aykırı sonuçlar ortaya koymadığına da işaret etmektedir. Hatalı sınıflandırılan ülkelerden birisi de Türkiye'dir. Kullanılan ham verilere göre kurulan model Birleşmiş Milletler tarafından orta düzeyde gelişmiş ülkeler kategorisinde sınıflandırılan Türkiye'yi çok gelişmiş kategoride sınıflandırmaktadır. Bu durumunda önemli olduğu düşünülmektedir.

Lojistik Regresyon Analizinde ise başlangıçta tüm bağımsız değişkenlerin modele dahil edilmesi ile Diskriminant Analizi ile yapılacak karşılaştırmada sonucun daha tutarlı yorumlanabileceği düşünülmüştür. Lojistik Regresyon Analizi ile amaçlanan daha az değişken kullanılarak sınıflandırmanın yapılabilmesi olduğundan İleri Adımsal Olabilirlik Oran yöntemi kullanılarak yeni regresyon modelleri oluşturulmuştur. Kurulan bu adımsal modeller yardımıyla bağımsız değişken sayısının indirgenmesi ve modelin kolayca yorumlanması sağlanmıştır.

Tüm verilerin kullanıldığı regresyon modelinde özel sağlık harcamaları pozitif etki gösterirken, kamu sağlık harcamaları, yaşam beklentisi ve ilköğretime net kayıt oranı değişkenleri negatif etki göstermişlerdir. Başlangıçta nisbi şans kriterinin yüksek tutulması için orta düzeyde gelişmiş grup "1" ile çok gelişmiş grup ise "0" ile kodlanmıştır. Yani Yaşam beklentisi, ilköğretime net kayıt oranı ve kamu sağlık harcamaları değeri 0'a yaklaştırarak ülkenin çok gelişmiş sınıfına dahil edilmesini sağlamaktadır.

İleri Adımsal regresyon modelinde 7 değişken seçilmiştir. Bu değişkenler genel olarak incelendiğinde modelin sağlık, eğitim, enerji ve iletişim değişkenlerinden oluştuğu gözlenmektedir. Ayrıca kırsal nüfusunda önemli ancak ters yönlü bir etkisi olduğu da gözlenmiştir. Yedi değişken ile de yapılan sınıflandırma neticesinde %100'lük bir sınıflandırma başarısı elde edilmiştir.

Lojistik Regresyon Analizi ile Diskriminant Analizi sonucunda elde edilen değişkenlerin gelişmişliğe olan pozitif ve negatif etkileri karşılaştırmalı olarak Tablo 19'da gösterilmektedir.

Tablo 19: Analiz Sonuçları Karşılaştırma Çizelgesi

Değişken Etiketi	Diskriminant Analizi Sonucunda Değişkenin Gelişmişliğe Etkisi	Lojistik Regresyon Analizi Sonucunda Değişkenin Gelişmişliğe Etkisi
IV1	Pozitif	Pozitif
IV2	Negatif	Negatif
IV3	Pozitif	Pozitif
IV4	Negatif	Pozitif
IV5	Negatif	Negatif
IV6	Pozitif	Pozitif
IV7	Pozitif	Pozitif
IV8	Negatif	Pozitif
IV9	Negatif	Pozitif
IV10	Pozitif	Negatif

IV11	Pozitif	Negatif
IV12	Negatif	Negatif
IV13	Negatif	Pozitif
IV14	Pozitif	Negatif
IV15	Pozitif	Pozitif
IV16	Negatif	Negatif

Tablo 19'da değişkenlerin pozitif etkileri ülkeleri çok gelişmiş ülke sınıfına taşırken negatif etkiler ülkelerin orta düzeyde gelişmiş ülkeler grubunda sınıflandırılmalarına sebep olmaktadır. Tablo 19'da Kamu Sağlık Harcamaları (IV4), Eğitim (IV8), İletişim (IV9,IV10,IV11), İthalat (IV13) ve İhracat (IV14) değerleri iki analizde farklı etkiler göstermektedir. Bu değişkenlerden IV9,IV11,IV13 ve IV16 diskriminant analizinde normallik dönüşümüne tabi tutulmuştur. Lojistik Regresyon Analizinde hiçbir değişken herhangi bir dönüşüme tabi tutulmadan kullanılmıştır.

Sınıflandırma sonuçları incelendiğinde diskriminant analizinde %92,5 ve lojistik regresyon analizinde %100'lük bir başarı söz konusudur. Bu durum metrik verilerin kullanıldığı durumlarda lojistik regresyon analizi gibi hiçbir varsayım gerektirmeyen bir yöntemi birçok varsayımın karşılanmasını gerekli kılan diskriminant analizine göre üstün hale getirmektedir. Diskriminant analizinden elde edilen başarı yüzdesi aslında çok yüksek bir başarı yüzdesidir. İncelenen çalışmalarda %65'in üzerindeki başarı yüzdelerinin kabul edilebilir olduğu gözlenmiştir.

Bu çalışma ile iki istatistiksel sınıflandırma tekniği uygulamalı olarak karşılaştırılmıştır. Bu karşılaştırma sonucunda diskriminant analizinin varsayımlarının fazla olmasının analizci açısından zaman problemi yaratacağı aşikardır. Ayrıca kullanılacak bağımsız değişkenler ölçülebilir ise ve bağımlı değişken ikili (binary) olarak kurgulanmışsa bu durumda lojistik regresyon analizinin kullanılması ile daha etkin bir fonksiyon bulunabilecektir. Bu durumun grupların örneklem hacmi arttıkça diskriminant analizi lehine gelişeceği değerlendirilmektedir. Bu konunun ileriki çalışmalarda bu kapsamda değerlendirilerek daha büyük örnek hacimleri kullanılarak test edilebileceği ve karşılaştırmaların değerlendirilebileceği düşünülmektedir.

KAYNAKÇA

Balcaen, S., H. Ooghe (2006); "35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems" ***The British Accounting Review***, 38, 63-93.

Berg, D. (2004); "Bankruptcy Prediction by Generalized Additive Models" ***Applied Stochastic Models in Business and Industry***, 23, 129-143.

Bosse, D. A. (2008); "Bundling Governance Mechanisms to Efficiently Organize Small Firm Loans" ***Journal of Business Venturing***, 24, 183-195.

Chen, K., D. C. Yen, S. Hung, A. H. Huang (2008); "An Exploratory Study of the Selection of Communication Media: The Relationship Between Flow and Communication Outcomes", ***Decision Support Systems***, 45, 822-832.

Cheng, B., D. M. Titterington (1994); "Neural Networks: A Review From a Statistical Perspective", ***Statistical Science***, 9(1), 2-30.

Çilan, Ç. A., B. A. Bolat, E. Coşkun (2009); "Analyzing Digital Divide Within and Between Member and Candidate Countries of European Union", ***Government Information Quarterly***, 26, 98-105.

Erçetin, Y. (1993); ***Diskriminant Analizi ve Bankalar Üzerine Bir Uygulama***, Türkiye Kalkınma Bankası A.Ş., APM/28 (KİG-26), 1-2.

Gujarati, D. N. (2001); ***Temel Ekonometri***, Çev: Ümit Şenesen, Gülay G. Şenesen, İstanbul.

Klecka, W. (1980); ***Discriminant Analysis***, Sage Publications, London.

Lachenbruch, P. A. (1975); ***Discriminant Analysis***, Hafner Press, London.

Liang, Z., P. Shi (2004); "Kernel Discriminant Analysis and Its Theoretical Foundation", ***The Journal of The Pattern Recognition Society***, 38, 445-447.

Lu, J., K. N. Plataniotis, A. N. Venetsanapoulos, J. Wang (2005); "An Efficent Kernel Discriminant Analysis Method", ***The Journal of The Pattern Recognition Society***, 38, 1788-1790.

Pompe, P. P. M., J. Bilderbeek (2005); "The Prediction of Bankruptcy of Small-and-Medium Sized Industrial Firms", ***Journal of Business Venturing***, 20, 847-868.

Sharma, S. (1996); ***Applied Multivariate Techniques***, John Wiley and Sons Inc., Canada.

Srivastava, S., M. Gupta, B. Frigyik (2007); "Bayesian Quadratic Discriminant Analysis", ***Journal of Machine Learning Research***, 8, 1277-1305.

Sueyoshi, T. (2004); "A Methodological Comparison Between Standard and Two Stage Mixed Integer Approaches for Discriminant Analysis", ***Asia-Pacific Journal of Operations Research***, 4, 513-528.

Tang, H., T. Fang, P. Shi (2005); "Laplacian Discriminant Analysis", ***The Journal of The Pattern Recognition Society***, 39, 136-139.

Tatlıdil, H. (1996); ***Uygulamalı Çok Değişkenli İstatistik Teknikleri***, Cem Ofset Ltd.Şti., Ankara.

Ulupınar, S. D. (2007); ***2001 Kriz Dönemi, Öncesi ve Sonrasında Türk Ticari Bankalarının Karlılıklarının Lojistik Regresyon Analizi ile İncelenmesi***, İstatistik Bilim Dalı Yüksek Lisans Tezi, Marmara Üniversitesi, İstanbul.

Wu, D. D., L. Liang, Y. Zijiang (2008); "Analyzing Financial Distress of Chinese Public Companies Using Probabilistic Neural Networks and Multivariate Discriminate Analysis", ***Socio Economic Planning Sciences***, 42, 206-220.