Güncel İstatistik Uygulamaları

Editör

Levent ÖZBEK

BİDGE

BİDGE Yayınları

Güncel İstatistik Uygulamaları

Editör: Doç. Dr. Levent Özbek

ISBN: 978-625-372-438-2

1. Baskı
Sayfa Düzeni: Gözde YÜCEL
Yayınlama Tarihi: 25.12.2024
BİDGE Yayınları

Bu eserin bütün hakları saklıdır. Kaynak gösterilerek tanıtım için yapılacak kısa alıntılar dışında yayıncının ve editörün yazılı izni olmaksızın hiçbir yolla çoğaltılamaz.

Sertifika No: 71374
Yayın hakları © BİDGE Yayınları
www.bidgeyayinlari.com.tr - bidgeyayinlari@gmail.com
Krc Bilişim Ticaret ve Organizasyon Ltd. Şti.
Güzeltepe Mahallesi Abidin Daver Sokak Sefer Apartmanı No: 7/9 Çankaya / Ankara

İçindekiler

Karma Yöntem Araştırmaları 4
Havva Ezgi ALBAYRAK 4
Alper PERÇİN 4
Multinomial Lojistik Regresyonda Farklı Kestirim Yöntemlerinin İçilişkili ve İçilişkisiz Veriler Üzerinde Karşılaştırılması 16
Nurkut Nuray URGAN 16
Ekrem BOZBEY 16
Bayesci İstatistiğe Giriş: Ibm Spss ile Bayesci T-testleri 46
Hamza ERDOĞDU 46
Harika Gözde GÖZÜKARA BAĞ 46
Ibm Spss Kullanarak Bayesci Regresyon Analizi 62
Hamza ERDOĞDU 62
Harika Gözde GÖZÜKARA BAĞ 62
Trafik Kazalarının Analizi: Poisson Regresyon ve Sağlam Poisson Regresyon Yöntemleri ile Risk Faktörlerinin Belirlenmesi 79
Barış ERGÜL 79
Arzu ALTIN YAVUZ 79
Yaralanmalı Kaza Sayılarının İzlenmesi İçin Regresyon Kontrol Grafikleri 98
Barış ERGÜL 98
Arzu ALTIN YAVUZ 98
Genelleştirilmiş Pareto-Poisson Süreci ve bir uygulama 112
Gamze ÖZEL KADILAR 112
Bir Veri Dengeleme Tekniği Olarak SMOTEENN 139
Gökhan KORKMAZ 139

BÖLÜM I

Karma Yöntem Araştırmaları

Havva Ezgi ALBAYRAK[1]
Alper PERÇİN[2]

Giriş

Sağlık, eğitim ve sosyal araştırmalarda "karma yöntem" yaklaşımının önemli bir dönüm noktası olduğu savunulmaktır. Karma yöntem araştırmaları, araştırmacının bir çalışma veya birbirini izleyen çalışmalar içerisinde nitel ve nicel yöntem, yaklaşım ve kavramları birleştirmesi olarak tanımlanır (Creswell, 2003; Tashakkori & Teddlie, 1998; Johnson & Turner, 2003). Karma yöntemle araştırma yapmak ise çeşitli yöntemler kullanılarak olayları bir çerçeve içerisinde sunma, analiz etme ve bir araya

[1] Öğr.Gör., Avrasya Üniversitesi, Sağlık Hizmetleri Meslek Yüksekokulu, Terapi ve Rehabilitasyon Bölümü, Trabzon/Türkiye, Orcid: 0000-0002-4595-2043, ezgii.karas@gmail.com
[2] Dr..Öğr. Üyesi., Avrasya Üniversitesi, Sağlık Bilimleri Fakültesi, Ergoterapi Bölümü, Trabzon/Türkiye, Orcid: 0000-0001-7974-6945, fztalperpercin@gmail.com

getirmektir. Yapılan bir çalışmada karma araştırmanın temel ilkesi "araştırmacı farklı strateji, yöntem ve yaklaşımları kullanarak çoklu veriler toplamalı" olarak belirtilmiştir (Johnson & Turner, 2003).

Farklı bir çalışmada ise karma yöntemin önermesi için "nicel ve nitel yaklaşımları birlikte kullanmak, her iki yaklaşımı da tek başına kullanmaya oranla araştırma problemlerini daha iyi anlamızı sağlar" şeklinde vermiştir (Creswell, 2006). Bunun paralelinde yukarıda verilen tanıma benzer olarak karma yöntem çalışmalarının bir araştırma programı kapsamında yapılan tek bir çalışma veya çoklu çalışmalar içerisinde, nitel ve nicel verilerin toplanması ve analiz edilmesini kapsadığını söylenmektedir.

Tseng ve diğerleri (Tseng & ark., 2023) tarafından kullanılan karma yöntem araştırması sıklıkla üçüncü araştırma hareketi olarak anılır ve genellikle insanların günlük yaşamlarını daha doğal bir şekilde yansıtan sezgisel bir yaklaşım olarak tanımlanır (Teddlie & Tashakkori, 2009). Örneğin, sağlık profesyonlleri yöntemleri karıştırma konusunda halihazırda bir miktar deneyime sahiptir çünkü teşhis ve tedavi kararlarını vermek için hasta görüşmeleri, fiziksel muayeneler ve teşhis testleri yoluyla düzenli olarak hasta bilgilerini toplarlar. Cresswell ve Plano Clark ve diğerleri, karma yöntemi niteliksel ve niceliksel yaklaşımlardan elde edilen verilerin toplanmasını ve analizini, bu yaklaşımların belirli tasarımlar kullanılarak düzenlenmesini, bulguların bütünleştirilmesini ve bu prosedürlerin teorik ve felsefi perspektiflerle çerçevelenmesini içerecek şekilde tanımlamaktadır (Creswell & Plano Clark, 2018).

Karma yöntem araştırması, niceliksel ve niteliksel verileri bir çalışma içinde bütünleştirir ve araştırmacıların çoğulcu bir bakış

açısıyla nedensel çıkarımı geliştirmede önemli olduğu öne sürülür (Krieger & Smith, 2016; Schwartz, Gatto & Campbell, 2016; Vandenbroucke, Broadbent & Pearce, 2016). Popülasyonun içgüdüsünü hipotez oluşturmaya getirtebilen ve bağlamı nedensel yapılara dahil edebilir.

Niteliksel ve niceliksel araştırma, temel bilimsel paradigmaları bakımından farklılık gösterir (Vandenbroucke, Broadbent & Pearce, 2016; Johnson, Russo & Schoonenboom, 2019). Köklerini pozitivizm ve ampirizme dayanan niceliksel paradigma, olguları gerçeği temsil eden ampirik göstergelere indirger (Johnson, Russo & Schoonenboom, 2019). Bunun aksine nitel araştırma paradigması, gerçekliğin sosyal olarak inşa edildiği ve sürekli değiştiği yapılandırmacılığa dayanmaktadır (Johnson, Russo & Schoonenboom, 2019; Creswell & Plano Clark, 2018).

Yaklaşım açısından niceliksel yöntemler öncelikle tümdengelimlidir; teoriden yukarıdan aşşağıya, hipotezin formülasyonuna ve ardından bireysel gözlemlerle doğrulama veya reddetmeye doğru ilerler. Niteliksel yöntemler ise aksine, öncelikle tümevarıma dayalıdır; belirli gözlemlerden, kalıplara, hipotezlerin formülasyonuna ve ardından teorilere aşşağıdan yukarıya doğru hareket ederler (Sale, Lohfeld & Brazil, 2002). Ayrıca araştırmacılar niceliksel verileri genellikle etik veya harici bir perspektiften yorumlarlar.

Öte yandan nitel araştırma, geleneksel olarak takip edilen, katılımcıları ve onların görüşlerini araştırmanın merkezine koyan bir bakış açısıyla karakterize edilen yaklaşıma bağlı kalır (Curry, Nembhard & Bradley, 2009; Di Giacomo, 1999). Örneğin, beslenme

epidemiyolojisinde nicel görüş, besin seviyesi verilerini elde etmek için yöntemler kullanılabilirken, nitel görüş yemeklerle ilgili kültürel uygulamaları anlamak için yöntemler kullanabilir.

Nicel ve nitel veri toplama ve analizi de farklılık gösterir. Bunlardan ilki, niceliksel yöntemler genellikle kapalı uçlu ve yapılandırılmış anketlerden, kamuya açık veri kaynaklarından, klinik kayıtlardan veya biyolojik ölçümlerden sayısal verileri toplar; nitel bilgi kaynakları ise belgelerden, transkripsiyonlardan veya alan notlarından elde edilen metin ve görüntüleri içerir. (derinlemesine görüşmeler, odak grupları, katılımcı gözlemleri vb). Bir diğeri ise, niceliksel veri toplama genellikle ayrı olarak ve veri analizinden önce gerçekleşir. Buna karşılık niteliksel veri toplama daha tekrarlı olma eğilimdedir (Bogner & ark., 2008).

Niteliksel araştırmacılar, doygunluğu değerlendirmeye yardımcı olmak için topladıkları verileri, yani ek katılımcılardan yeni temalar ortaya çıkmadığında görüşme kılavuzlarını geliştirebilir ve verileri analiz edebilirler. Üçüncüsü, niceliksel veri toplama genelleştirilebilir olma hedefiyle olasılıklı bir örneklemden üretilme eğilimdeyken, niteliksel veri toplama derilemesine bilgi elde etmek için amaçlı bir örnekleme stratejisi izler. Ister amaçlı ister olasılıksal olsun, her iki örnekleme stratejisi de benzerlik ve farklılık unsurlarını yakarlar (Johnson & Onwuegbuzie, 2004). Ayrıca gözlemsel çalışmalar çoğunlukla niceliksel verileri olasılıksaldan daha az ve uygun örneklerden daha fazla toplar. Bazı araştırmacılar karşıt paradigmaların niceliksel ve niteliksel yaklaşımları ayrı tutmayı haklı çıkardığını öne sürmektedir (Palinkas, Mendon & Hamilton, 2019).

Nicel ve niteliksel yöntemlerin aynı anda kullanımı ile karma yöntem araştırmaları, her iki tekniğin avantajlı taraflarını kuvvetlendirmektedir. Karma yöntemleri kullanan araştırmacıların belirledikleri araştırma sorularıyla ilişkili olan yöntem ve yaklaşımları seçme şansları daha fazladır. Karma yöntem araştırmaları, araştırmacının seçeneklerini kısıtlamak yerine araştırma sorularına cevap arar. Karma yöntemin amacı pek çok durumda bir fikri doğrulamak ya da desteklemek değil, kişinin olayla ilgili anlayışını genişletmek ve farklı bir perspektif sağlamaktır (Bazeley, 2004).

Karma yöntem; kapsamlı, çoğulcu ve araştırmacıya yöntem seçimi ve araştıram hakkında tasarlama yapması için seçmeci bir bakış açısı önerir. Birçok araştırma sorusu veya soruları karma yöntemin sunduğu çözüm yolları ile tamamen cevaplandırılabilir (Onwuegbuzie & Leech, 2004). Yapılan bir çalışmada; tek bir araştırma içerisinde nitel ve nicel yöntemleri birleştirmenin daha bütüncül bir bakış açısı sağlayarak, daha iyi bilgilendirilmiş eğitim politikaları oluşturarak, araştırma konusunun farklı yönlerini açıklamaya yardımcı olduğunu belirtilmiştir (Onwuegbuzie & Leech, 2004).

Karma yöntemlerin kullanım koşullarını belirlemek, çalışmanın genel amacına neden daha iyi hizmet edebileceği konusunda aydınlatma gerektirir (Greene, 2017). Bazı yaygın nedenler arasında üçgenleme, daha eksiksiz bulgulara duyulan ihtiyaç, ilk bulguları açıklama ihtiyacı veya önce araştırma yapma ve daha sonra bu bulguların daha geniş bir popülasyondaki düzenliliğini inceleme ihtiyacı yer alır (Cresswell & Plano Clark, 2018). Karma yöntem kullanım nedenleri aşağıdadır:

- Üçgenleme; bir araştırmacının amacı, bulguları doğrulamak ve çıkarımların geçerliliğini arttırmak için önyargıları dengeleyen farklı araştırma yaklaşımları kullanmak olduğunda seçilir. Araştırmacılar bir olgunun daha derin, daha zengin, daha eksiksiz bir tanımı veya anlayışını arayabilir. Burada araştırmacılar, aksi takdirde tek yöntemli bir yaklaşımla erişelemeyecek bulguları belirlemek amacıyla bir olgunun farklı yönlerini incelemek için kasıtlı olarak birden fazla araştırma biçimi kullanırlar.
- Araştırmacılar araştırma yöntemini genişletmek için karma yöntemi kullanabilirler. Örneğin, nicel yaklaşımlar değişkenler arasındaki ilişkilerin belirlenmesine yardımcı nitel yaklaşımlar bu ilişkilerin altında yatan bağlamlar veya süreçler hakkında daha fazla ayrıntıyı ortaya çıkarmak için daha uygundur.
- Daha az bilinen olguları araştırırken, araştırmacılar hangi soruların değerli olduğunu daha iyi anlamak için keşfedici bir yaklaşımla başlayabilirler. Daha sonra daha büyük bir popülasyondaki düzenliliği belirlemek amacıyla araştırmanın kapsamını genişletmek için bu bulguları kullanabilirler (Yan Ran & ark., 2023).

Herhangi bir çalışmayı tasarlamadan önce, karma yöntem araştırmalarına neden ihtiyaç duyulduğunu açıklamak için, çalışma kapsamındaki olguyu ve çalışma için kullanılacak amacı veya araştırma sorusunu tanımlamak önemlidir. Bu araştırmacının çalışmanın felsefi temelini araştırma tasarımına bağlanmasını sağlar (Hesse-Biber, 2010).

Niteliksel verilerin rolü, ağırlıklı olarak niceliksel bir tasarımı desteklemek için olabilir; örneğin bir klinik araştırmanın yanı sıra niteliksel bir süreç değerlendirmesi. Benzer şekilde, nicel verilerin rolü ağırlıklı olarak nitel verilerin savunulması olabilir; örneğin keşif amaçlı bir vaka çalışmasında nitel bulguları zenginleştirmek için daha fazla bilgi sağlamak amacıyla nicel bir ölçümün kullanılması. Karma yöntem araştırmasının birçok türü vardır. Bunlar aşağıdaki maddelerden oluşur:

- Başlangıçta nitel veri toplama ve analizi ile karakterize edilebilen, ardından iki ayrı veri derlemesinden verilerin entegrasyonuna veya bağlanmasına yol açan nicel veri toplama analiziyle karakterize edilebilen sıralı keşfedici karma yöntem tasarımı
- Niceliksel sonuçları nitel bulgular kullanarak açıklamak için genellikle niceliksel ve ardından nitel verilerin birbirini takip eden iki aşamada toplanmasını, analiz edilmesini ve birleştirilmesini, bulguların tek bir çalışma içinde bütünleştirilmesini içeren sıralı açıklayıcı karma yöntem tasarımı.
- Farklı yöntemlerden ve veri kaynaklarından elde edilen sonuçların geçerliliğini doğrulayan bileşen kanıtları tanımlamayı amaçlayan nicel ve nitel veriler içerek yakınsak karma yöntemli çalışma tasarımı.
- Sonuçların kalitesini arttırmak için niceliksel ve niteliksel bileşenlerin yan yana yer aldığı, ancak bir bileşenin açıkça baskın olduğu ve diğerinin iç içe geçtiği veya gömülü olduğu iç içe geçmiş karma yöntem çalışması.

Hangi karma yöntem araştırma tasarımı kullanılırsa kullanılsın, en yüksek kaliteye ulaşmak yapılan şeyin (neden, nasıl, ne zaman) ayrıntılı olarak tanımlanmasında gözükmektedir (Barret, Kajamaa, & Johnston, 2020). Başka bir deyişle çalışmanın sınırlılıklarına karşı duyarlı olmak ve bunları raporlamak önemlidir.

Sonuç

Bu derleme makalesinde karma yöntem hakkında genel bilgiler verilmiş, yöntemin tanımı, karma yöntemle araştırma yapmanın gerekçeleri ile avantaj ve dezavantajlarına değinilmiştir. Karma yöntem araştırması, karmaşık olguları ve durumları araştırmak için iyi bir örnektir. Araştırmacılara belirli olgular hakkında tek yöntemlerin kullanımına göre daha ayrıntılı bir anlayış sağlayabilir.

Sağlık ve sosyal bakımdaki karmaşık süreçlerin ve sistemlerin araştırılması için hem niceliksel hem niteliksel yaklaşımlardan yararlanan güçlü bir yöntemdir. Bu konuda yazılan kaynakların ve yapılan çalışmaların artıyor olması zamanla yöntem hakkındaki belirsizliklerin ortadan kalkacağına ve daha sağlam temelli karma yöntem araştırmalarının yapılacağına olan umudu kuvvetlendirmektedir.

Kaynaklar

Barrett, A., Kajamaa, A., & Johnston, J. (2020). How to … be reflexive when conducting qualitative research. The clinical teacher, 17(1), 9–12.https://doi.org/10.1111/tct.13133

Bazeley, P. (2004). Issues in mixing qualitative and quantitative approaches to research. Buber R, Gadner j, Richards L, eds. In: Applying qualitative methods to marketing management research. Palgrave Macmillan. p:141–156.

Bogner, H. R., Dahlberg, B., de Vries, H. F., Cahill, E., & Barg, F. K. (2008). Older patients' views on the relationship between depression and heart disease. Family medicine, 40(9), 652–657. PMCID: PMC2825038 PMID: 18830841

Creswell, J. W. (2003). Research design: Qualitative, quantitative, and mixed methods approaches (2nd ed.). Thousand Oaks, CA: Sage

Creswell, J.W. (2006). Understanding Mixed Methods Research, (Chapter 1). Available at: http://www.sagepub.com/upm-data/10981_Chapter_1.pdf

Cresswell, J.W., Plano Clark VL. 2018a. Core mixed methods designs. In: Designing and conducting mixed methods research. 3rd ed. Thousand Oaks (CA): Sage Publications, Inc.; p. 51–99.

Curry, L. A., Nembhard, I. M., & Bradley, E. H. (2009). Qualitative and mixed methods provide unique contributions to outcomes research. Circulation, 119(10), 1442–1452. https://doi.org/10.1161/CIRCULATIONAHA.107.742775

Di Giacomo SM. Can there be a “cultural epidemiology”? Med Anthropol Q. 1999;13:436–457. doi: 10.1525/maq.1999.13.4.436. https://doi.org/10.1525/maq.1999.13.4.436

Greene, J.C. (2007). Mixing methods on purpose. In: Mixed methods in social inquiry. San Francisco (CA): John Wiley & Sons; p. 95–111.

Hesse-Biber, S. (2010). Qualitative Approaches to Mixed Methods Practice. Qualitative Inquiry, 16(6), 455-468. https://doi.org/10.1177/1077800410364611

Johnson, B., & Turner, L. A. (2003). Data collection strategies in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), Handbook of mixed methods in social and behavioral research (pp. 297-319). Thousand Oaks, CA: Sage.

Johnson, R. B., Russo, F., & Schoonenboom, J. (2019). Causation in Mixed Methods Research: The Meeting of Philosophy, Science, and Practice. Journal of Mixed Methods Research, 13(2), 143-162. https://doi.org/10.1177/1558689817719610

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. Educational Researcher, 33(7), 14-26. https://doi.org/10.3102/0013189X033007014

Krieger, N., & Davey Smith, G. (2016). The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. International journal of epidemiology, 45(6), 1787–1808. https://doi.org/10.1093/ije/dyw114

Palinkas, L. A., Mendon, S. J., & Hamilton, A. B. (2019). Innovations in Mixed Methods Evaluations. Annual review of public health, 40, 423–442. https://doi.org/10.1146/annurev-publhealth-040218-044215

Onwuegbuzie, A. J., & Leech, N. L. (2004). "Enhancing the Interpretation of “Significant” Findings: The Role of Mixed Methods Research". The Qualitative Report, 9(4): 770-792.

Onwuegbuzie, A. J., & Leech, N. L. (2006). Linking research questions to mixed methods data analysis procedures. The Qualitative Report, 11 (3), 474-498. Retrieved [10.06.2011], from http://www.nova.edu/ssss/QR/QR11-3/onwuegbuzie.pdf

Sale JEM, Lohfeld LH, Brazil K. Revisiting the Quantitative-Qualitative Debate: Implications for Mixed-Methods Research. Vol 36. Qual Quant; 2002.

Schwartz, S., Gatto, N. M., & Campbell, U. B. (2016). Causal identification: a charge of epidemiology in danger of marginalization. Annals of epidemiology, 26(10), 669–673. https://doi.org/10.1016/j.annepidem.2016.03.013

Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches. Applied Social Research Methods Series (Vol.46). Thousand Oaks, CA: Sage.

Teddlie, C., Tashakkori, A. (2009). Mixed Methods as the third research community. In Foundations of mixed methods research: integrating quantitative and qualitative approaches in the social and behavioral sciences. Thousand Oaks (CA): Sage Publications, Inc.; p. 3–18.

Tseng, A. S., Edwards, C., Rawls, M., McGinn, M., Wieghard, N., Santen, S. A., & Deiorio, N. M. (2023). A mixed methods exploration of the emergence of goal orientation in medical students' individualized learning plans. Medical teacher, 45(6), 588–595. https://doi.org/10.1080/0142159X.2023.2169117

Vandenbroucke, J. P., Broadbent, A., & Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. International journal of epidemiology, 45(6), 1776–1786. https://doi.org/10.1093/ije/dyv341

Li, Y.R., Zhang, Z.H., Li, W. Wang, P., Li, S.W., Su, D., Zang, T. (2023). Effectiveness and learning experience from undergraduate nursing students in surgical nursing skills course: a quasi- experimental study about blended learning. BMC Nurs 22, 396. https://doi.org/10.1186/s12912-023-01537-w

BÖLÜM II

Multinomial Lojistik Regresyonda Farklı Kestirim Yöntemlerinin İçilişkili ve İçilişkisiz Veriler Üzerinde Karşılaştırılması

Nurkut Nuray URGAN[1]
Ekrem BOZBEY[2]

1.Giriş

Araştırmacılar, bir yanıt değişkenin bir ya da birden fazla açıklayıcı değişken yardımıyla tahmin edildiği araştırmalarda sıklıkla regresyon analizine başvurur. Yanıt değişkenin sürekli olduğu verilerde doğrusal regresyon analizi, kategorik olduğu verilerde ise lojistik regresyon analizi kullanılabilir. Lojistik regresyon analizinde yanıt değişken nominal veya ordinal olabilir.

[1] Dr. Öğrt.Üyesi, Tekirdağ Namık Kemal Üniversitesi, Fen Edebiyat Fakültesi. Matematik Bölümü, e-mail: nurgan@nku.edu.tr, ORCID: 0000-0002-8691-4692.
[2] Öğretmen, Tekirdağ BİLSEM, e-mail: krmbzby@gmail.com, ORCID: 0009-0004-1399-1170.

Yanıt değişkenin iki kategorili olduğu verilerde Binomial/Binary Lojistik Regresyon, üç veya daha fazla kategorili olduğu verilerde ise Multinomial Lojistik Regresyon uygulanır.

Lojistik regresyon analizinde, temel olarak maksimum olabilirlik kestirimi kullanılmakla birlikte, bazı durumlarda modelin doğruluğunun ve analiz gücünün arttırılması için farklı kestirim yöntemleri kullanılabilir.

Bu çalışmada, Multinomial Lojistik Regresyon (MLRA) için maksimum olabilirlik yöntemine alternatif olarak kullanılabilecek dört kestirim yöntemi incelenmiştir.

2. Lojistik Regresyon

Lojistik regresyon analizi kullanmanın amacı, istatistikte kullanılan herhangi bir model oluşturma tekniğiyle aynıdır. Bu amaç, bir yanıt değişkeni ile bir açıklayıcı değişkenler kümesi arsındaki ilişkiyi tanımlamak için en uygun ve en ekonomik, aynı zamanda en makul modeli oluşturmaktır (Hosmer ve Lemeshow, 2000). Lojistik fonksiyon,

$$f(x) = \frac{1}{1+e^{-u(x)}} \tag{1.1}$$

şeklindedir.

2.1. Maksimum Olabilirlik Yöntemi

Lojistik model, denklem (1.1)'deki lineer olmayan lojistik fonksiyon yardımıyla yanıt değişkenin hangi kategoriye ait olduğuna yönelik olasılıkları hesaplar. Bu model, maksimum olabilirlik yöntemi ile elde edilebilir. Maksimum olabilirlik kestirimi, veri setindeki açıklayıcı değişkenlerin katsayıları için yansız seçilmiş değerler ile işe başlayan, log-olabilirlik ifadesinde gözlenen

frekansları elde etme olabilirliğini en yüksek düzeye çıkaracak katsayıların yönü ve büyüklüğü hakkında karar vermeye yarayan iteratif bir süreçtir (Tabachnick ve Fidel, 2013).

$$L(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \log (P_{ij}) \quad (1.2)$$

2.2. Ridge Yöntemi

Hoerl ve Kennard tarafından tarafından önerilen Ridge yönteminin amacı, X matrisinde çeşitli sütunlar arasındaki ilişkinin, X'X matrisinin tekile yakın olmasına neden olduğu ve büyük standart hatalarla birlikte kararsız parametre tahminlerine yol açtığı durumların üstesinden gelmektir. Bunun için, doğrusal regresyon analizinde hata kareler toplamına (SSE) L2 normuyla bir λ ceza parametresi uygulanırken, bu ceza lojistik regresyon analizinde log-olabilirlik fonksiyonuna (1.3)'teki gibi uygulanır.

$$\sum_{i=1}^{N} L(\beta) - \lambda \sum_{j=1}^{p} {\beta_j}^2 \quad (1.3)$$

2.3. LASSO Yöntemi

Tibshirani tarafından önerilen LASSO yönteminde, yalnızca en önemli değişkenler seçilerek yanıt değişken tahminlenirken, diğer değişkenlerin katsayıları 0 yapılarak modelden çıkartılır. Bu sayede, özellikle fazla sayıda açıklayıcı değişkenin olduğu veri setlerinde otomatik değişken seçimi yapılarak, modelin aşırı öğrenme (overfitting) riskinden kaçınılması ve model isabetinin arttırılması hedeflenir. Lojistik regresyon analizinde LASSO yöntemi kullanılırken, L1 normuyla λ ceza terimi log-olabilirlik fonksiyonuna (1.4)'daki gibi uygulanır.

$$\sum_{i=1}^{N} L(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \quad (1.4)$$

2.4. Elastic-net Yöntemi

LASSO yöntemi, değişken seçiminde sık kullanılmakla birlikte bazı dezavantajları vardır. 2005 yılında Zou ve Hastie, LASSO yönteminin aralarında yüksek korelasyon olan bir grup değişkenden yalnızca bir tanesini seçme eğiliminde olduğunu ve yüksek korelasyona sahip açıklayıcı değişkenlerin baskın olduğu veri setlerinde LASSO modelin tahmin performansının Ridge modelin performansının gerisinde kaldığını vurgulayarak, her iki modelin avantajlarını bünyesinde toplamayı amaçlayan, L1 ve L2 normlarını birlikte kullanan Elastic-net yöntemini önerdiler. Lojistik regresyon analizinde Elastic-net yöntemi kullanılırken, λ ceza terimi log-olabilirlik fonksiyonuna (1.5)'deki gibi uygulanır.

$$\sum_{i=1}^{N} L(\beta) - \lambda_1 \sum_{j=1}^{p} {\beta_j}^2 - \lambda_2 \sum_{j=1}^{p} |\beta_j| \qquad (1.5)$$

2.5. Temel Bileşenler Analizi (PCA)

Temel Bileşenler Analizi, orijinal veri setindeki değişkenlerin lineer kombinasyonlarıyla oluşturulan bileşenler yardımıyla modelin kurulduğu, orijinal veri setindeki bilgiyi koruyan ve boyut azaltmaya yarayan bir yöntemdir. Bu sayede daha sade ve daha kolay yorumlanabilir modeller elde edilir. Birinci temel bileşen, orijinal veri setindeki tüm gözlemlere olabilecek en yakın doğruyu temsil eder. İkinci temel bileşen, birinci temel bileşene diktir. Üçüncü temel bileşen, ilk iki temel bileşene dik olur. Bu şekilde, birbiriyle ilişkili değişkenlerin lineer kombinasyonlarından, ortogonal bileşenler seti elde edilir.

3. Multinomial Lojistik Regresyonda Kestirimler

Bu çalışmada, yanıt değişkenin üç veya dört kategorili, açıklayıcı değişkenlerin tümünün sürekli olduğu Kaggle

platformundan alınan iki farklı gerçek veri setinde MLRA yapılmış, Ridge, LASSO, Elastic-net ve Temel Bileşenler Analizi yöntemleri uygulanmış ve bu yöntemlerle oluşturulan modellerin maksimum olabilirlik yöntemine göre elde edilen modele göre performansları, test veri setindeki tahmin isabet oranlarına göre kıyaslanmıştır. Veri setlerinden ilki çoklu iç ilişkisi yüksek açıklayıcı değişkenlerden oluşurken, ikinci veri setinde değişkenler arasında çoklu iç ilişki yoktur. Bu ilişki düzeyini belirlemek adına, açıklayıcı değişkenler arasındaki pearson korelasyonu ölçüt alınmıştır. Korelasyonların iç ilişkiyi işaret etmediği veri setlerinde VIF değerlerine de bakılmıştır. Diğer taraftan, p adet açıklayıcı değişkenli veri setinden belirlenmiş N (gözlem sayısı) değerlerine göre rastgele örneklemler alınarak N/p oranı kontrollü bir şekilde küçültülmüş ve bu örneklemler için ilgili yöntemlerin model performansları ayrıca kıyaslanmıştır. Model performansları, gözlem sayısının azalması ve çoklu iç ilişki kriterlerine göre irdelenmiştir. Veri setleri %80 eğitim (train) seti ve %20 test (validation) seti olarak ayrılmış ve eğitim veri seti üzerinden modeller oluşturulmuştur. Oluşturulan modellere ait katsayılar, standart hatalar, sapma ve AIC istatistikleri, Ridge, LASSO, Elastic-net modelleri için log lambda-sapma ve log lambda-katsayı değişim grafikleri incelenmiştir. Her bir modelin performansı, test veri setindeki tahmin isabet oranına bakarak kıyaslanmıştır. Veri setlerinde N/p oranlarını birbirine yakın tutmak için, anemi veri setinden 150 ve 50 gözlemli rastgele örnekler alınmış, deniz kulağı veri setinden ise 100 ve 40 gözlemli rastgele örnekler alınmıştır.

Çalışmada yapılan analizlerin tamamı için açık lisanslı R programı kullanılmıştır. Veri manipülasyonu için dplyr, örneklem

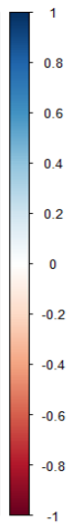
seçimi ve train-test ayrımı için rsample, model oluşturmak için nnet ve glmnet paketleri kullanılmıştır.

3.1. Denizkulağı Veri Seti

Yetişkin erkek, yetişkin dişi ve yavru denizkulaklarına ait 4177 gözlem kullanılmıştır. MLRA ile 4177 denizkulağına ait 8 farklı değişkenden yetişkin cinsiyet (M/F) veya yavru(I) tahmini yapılmıştır. Veri setinin çoklu iç ilişki düzeyini belirlemek için, açıklayıcı değişkenler arasındaki korelasyonlara bakılmıştır. Değişkenler arasındaki korelasyonlar Tablo 1' de gösterilmiştir.

Tablo 1: Denizkulağı veri seti korelasyon matrisi

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
Length	1.00	0.99	0.83	0.93	0.90	0.90	0.90	0.56
Diameter	0.99	1.00	0.83	0.93	0.89	0.90	0.91	0.57
Height	0.83	0.83	1.00	0.82	0.77	0.80	0.82	0.56
Whole weight	0.93	0.93	0.82	1.00	0.97	0.97	0.96	0.54
Shucked weight	0.90	0.89	0.77	0.97	1.00	0.93	0.88	0.42
Viscera weight	0.90	0.90	0.80	0.97	0.93	1.00	0.91	0.50
Shell weight	0.90	0.91	0.82	0.96	0.88	0.91	1.00	0.63
Rings	0.56	0.57	0.56	0.54	0.42	0.50	0.63	1.00



Denizkulağı veri setindeki 28 açıklayıcı değişken ikilisinden 21 tanesinde pearson korelasyonu 0.70'den büyük bulunmuştur. Veri setinin, iç ilişkisi yüksek değişkenlerden oluştuğu

görülmektedir. Dişi denizkulağı (F), Tablo 2' de referans kategori olarak alınarak ML yöntemiyle oluşturulan modelde, sapma istatistiği 5695.2 ve AIC istatistiği 5731.2 olarak bulunmuştur.

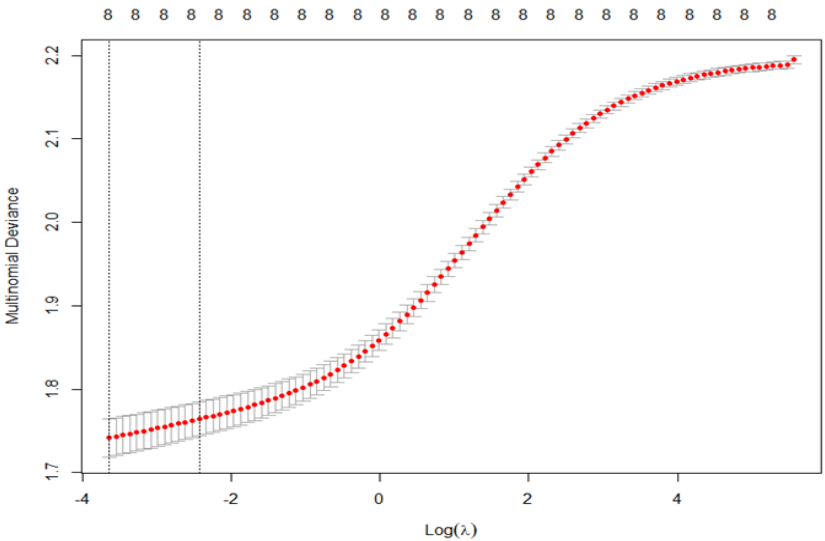
Tablo 2: Denizkulağı veri seti için maksimum olabilirlik yöntemi ile oluşturulan model

```
Call:
multinom(formula = Sex ~ ., data = trainset)

Coefficients:
  (Intercept)    Length   Diameter    Height `Whole weight` `Shucked weight` `Viscera weight`
I    2.550622 17.567074 -11.731078 -5.254056     -7.7174331         5.804126       -12.214507
M    2.375080 -0.537803  -4.977035 -3.644638     -0.1306915         2.989846        -2.225402
  `Shell weight`       Rings
I      7.2402665 -0.22664017
M      0.1744038  0.01300875

Std. Errors:
  (Intercept)   Length Diameter   Height `Whole weight` `Shucked weight` `Viscera weight` `Shell weight`
I   0.5572313 3.245307 4.005253 3.668387      1.9550401        2.2522866         2.946214       2.778817
M   0.4762063 2.163942 2.625750 2.244510      0.8022258        0.9660145         1.422334       1.241646
       Rings
I 0.03018655
M 0.01746852

Residual Deviance: 5695.167
AIC: 5731.167
```

Test veri setindeki 837 gözlemden 468 tanesi, ML yöntemiyle oluşturulan model tarafından doğru tahmin edilmiştir. ML modelinin test veri setindeki tahmin isabet oranı %55.9 olmuştur. Modelin, kategorilere göre tahminleri Tablo 3' de verilmiştir.

Tablo 3: Denizkulağı seti ML modelinin test veri seti için kategori tahminleri

	testing_y		
predabalone	F	I	M
F	84	12	68
I	56	218	72
M	122	39	166



Şekil 1: Ridge λ ceza parametresi-sapma değişim grafiği

Ridge için 10 tekrarlı CV yöntemiyle bulunan minimum lambda değeri 0.026'dır. Şekil 1'deki soldaki kesikli düşey çizgi, en küçük sapmayı veren bu lambda değerine karşılık gelmektedir.

Tablo 4: Denizkulağı seti Ridge modelin test veri seti için kategori tahminleri

	testing_y		
predablnridge	F	I	M
F	73	10	57
I	44	210	65
M	145	49	184

Ridge modelin test veri setindeki tahmin isabet oranı %55.7 olmuştur. Tablo 4 'de Ridge modelin test veri setinde kategorilere göre tahminleri verilmiştir.

LASSO için 10 tekrarlı CV yöntemiyle bulunan minimum lambda değeri 0.00014'tür.

Tablo 5: Denizkulağı seti LASSO modelin test veri seti için kategori tahminleri

	testing_y		
predablnlasso	F	I	M
F	85	12	69
I	55	218	70
M	122	39	167

LASSO modelin test veri seti için tahmin isabet oranı %56.1 olmuştur. Tablo 5' de LASSO modelin test veri setinde kategorilere göre tahminleri gösterilmiştir.

Elastic-net için 10 tekrarlı CV yöntemiyle bulunan minimum lambda değeri 0.00008'dir. Elastic-net yöntemi, 8 açıklayıcı değişkenden en az 1 tanesini modelden çıkartmıştır.

Tablo 6: Denizkulağı seti Elastic-net modelin test veri seti için kategori tahminleri

	testing_y		
predablnelastic	F	I	M
F	84	12	69
I	56	218	70
M	122	39	167

Elastic-net modelin test veri seti için tahmin isabet oranı %56.0 olmuştur. Tablo 6' da Elastic-net modelin test veri setinde kategorilere göre tahminleri gösterilmiştir.

Tablo 7: Denizkulağı seti PCA yöntemiyle oluşturulan temel bileşenler ve bu bileşenlerin verideki varyansı açıklama yüzdesi

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.5862 0.83755 0.52517 0.40618 0.29377 0.25229 0.11246 0.08108
Proportion of Variance 0.8361 0.08769 0.03448 0.02062 0.01079 0.00796 0.00158 0.00082
Cumulative Proportion  0.8361 0.92376 0.95823 0.97885 0.98964 0.99760 0.99918 1.00000
```

PCA için birinci temel bileşen tek başına verideki varyansın %83'ünü temsil etmektedir. Tüm bileşenlerin verideki bilgiyi temsil etme oranları Tablo 7' de gösterilmiştir. Bu bileşenlere, veri setindeki orijinal değişkenlerin katkısı ise Tablo 8' de görülmektedir.

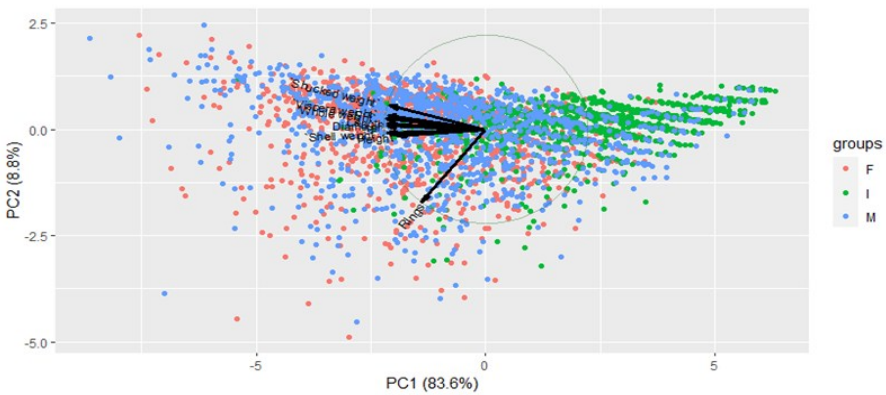
Tablo 8: Denizkulağı setinde her bir değişkenin temel bileşenlere katkısı

```
Standard deviations (1, .., p=8):
[1] 2.58622524 0.83754556 0.52516714 0.40618215 0.29376797 0.25229280 0.11246098 0.08107557

Rotation (n x k) = (8 x 8):
                       PC1         PC2           PC3         PC4          PC5         PC6          PC7
Length          -0.3728487  0.06924442 -0.0081830209 -0.60556939  0.004148367  0.04657920  0.697576169
Diameter        -0.3739334  0.03959268  0.0001843853 -0.58715619 -0.060816371  0.01695082 -0.714011063
Height          -0.3364932 -0.08089504  0.9165381523  0.19356876  0.044225694 -0.02621897  0.008387652
`Whole weight`  -0.3789238  0.13858083 -0.1860756573  0.25737529 -0.004851467 -0.12139020 -0.016939059
`Shucked weight` -0.3631166  0.29786846 -0.1855741082  0.19766554  0.417051671 -0.61228242 -0.001814284
`Viscera weight` -0.3691843  0.17783822 -0.1688992441  0.27332137  0.293618511  0.77544807 -0.026085721
`Shell weight`  -0.3710400 -0.04662457 -0.1568970287  0.26090441 -0.826249975 -0.05098910  0.049266505
Rings           -0.2419201 -0.91943628 -0.1941269879  0.05454419  0.226873893 -0.05753837  0.010366362
```

Veri setindeki tüm açıklayıcı değişkenler PC1'e negatif yönlü ve Rings değişkeni hariç neredeyse eşit düzeyde katkı yapmışlardır. Verideki varyansın yaklaşık %83'ünü temsil eden birinci bileşen ile model kurulmuştur. Bu modele ait katsayılar, standart hatalar, sapma ve AIC istatistikleri Tablo 9' da görülmektedir.

Tablo 9: Denizkulağı seti PCA modelin katsayıları, standart hataları, sapma ve AIC değerleri

```
Call:
multinom(formula = Sex ~ PC1, data = trn)

Coefficients:
  (Intercept)        PC1
I  -0.3593468 0.76269051
M   0.2487784 0.08812892

Std. Errors:
  (Intercept)        PC1
I  0.05946442 0.02827746
M  0.04720101 0.01967765

Residual Deviance: 5957.735
AIC: 5965.735
```



Şekil 2: Eğitim seti verilerinin PC1-PC2 uzayında kategorilerine göre dağılımı

Yatay eksen 1.temel bileşeni, düşey eksen 2.temel bileşeni temsil etmek üzere, model kurmak için kullanılan eğitim veri setindeki 3340 adet gözlemin kategorilere göre dağılımını gösteren grafik Şekil 2' de gösterilmiştir.

Tablo 10: Denizkulağı seti PCA modelin test veri seti için kategori tahminleri

p1	F	I	M
F	60	3	56
I	46	203	61
M	156	63	189

PCA modelin test veri setindeki tahmin isabet oranı %54.0 olmuştur. PCA modelin test veri setinde kategorilere göre tahminleri Tablo 10' da gösterilmiştir.

Aynı yöntemler, 4177 gözlemli denizkulağı veri setinden rastgele alınan N=40 ve N=100 gözlemli örneklemler için de uygulanmıştır. Bu örneklemler için kurulan modellerin test veri setlerindeki performansları Tablo 11 ve Tablo 12' de paylaşılmıştır.

Tablo 11: Denizkulağı seti N=40 için 5 modelin test veri setindeki tahminleri

	testing_y		
predabalone40	F	I	M
F	0	1	1
I	0	2	1
M	2	0	2

ML

	testing_y		
predabln40ridge	F	I	M
F	1	0	1
I	0	2	1
M	1	1	2

RIDGE

	testing_y		
predabln40lasso	F	I	M
F	1	0	1
I	0	2	1
M	1	1	2

LASSO

	testing_y		
predabln40elastic	F	I	M
F	1	0	1
I	0	2	1
M	1	1	2

ELASTIC NET

p1	F	I	M
F	1	0	1
I	0	2	1
M	1	1	2

PCA

Tablo 12: Denizkulağı seti N=100 için 5 modelin test veri setindeki tahminleri

```
                 testing_y
predabalone100 F I M
             F 2 0 4
             I 2 5 1
             M 2 1 4
```

ML

```
                    testing_y
predabln100ridge F I M
               F 2 0 2
               I 1 5 1
               M 3 1 6
```

RIDGE

```
                    testing_y
predabln100lasso F I M
               F 2 0 2
               I 2 4 1
               M 2 2 6
```

LASSO

```
                      testing_y
predabln100elastic F I M
                 F 2 0 1
                 I 2 4 1
                 M 2 2 7
```

ELASTIC NET

```
p1  F I M
  F 2 0 1
  I 2 4 1
  M 2 2 7
```

PCA

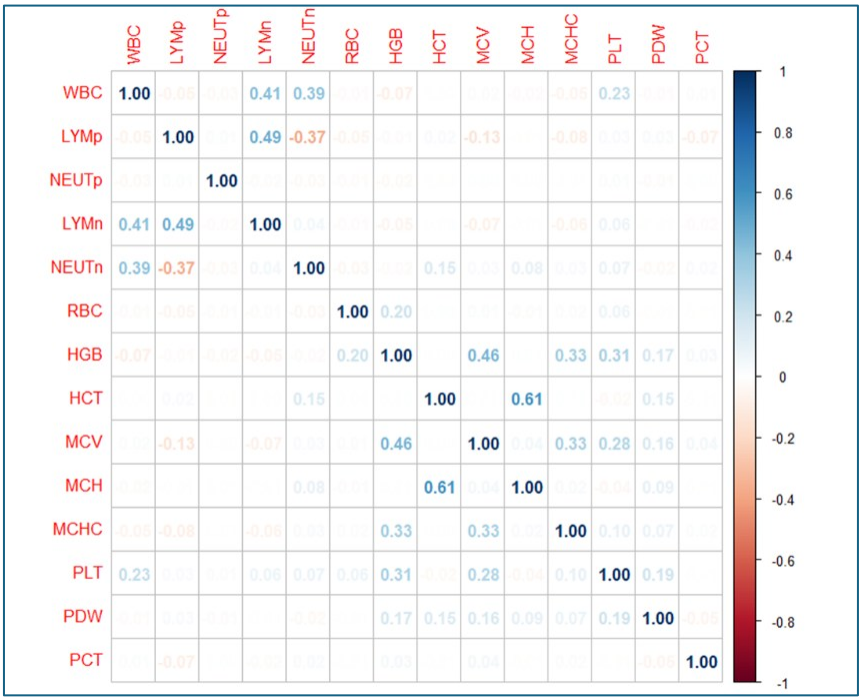
Tablo 13: Denizkulağı veri seti için farklı N değerlerine göre test setinde model tahmin isabet oranları

GÖZLEM SAYISI	**MODEL TAHMİN İSABET ORANI (%)**				
	ML	**Ridge**	**LASSO**	**Elastic-net**	**PCA**
N=4177	55.9	55.7	56.1	56.0	54.0
N=100	52.3	61.9	57.1	61.9	61.9
N=40	44.4	55.5	55.5	55.5	55.5

3.2. Anemi Veri Seti

336'sı sağlıklı, 189'u demir eksikliği anemisi, 279'u hypochromic anemi ve 269'u normochromic anemi olan toplam 1073 kişiye ait kan ölçüm sonuçları kullanılmıştır. Veri setinin çoklu iç ilişki düzeyini belirlemek için, açıklayıcı değişkenler arasındaki korelasyonlara bakılmıştır. Değişkenler arasındaki korelasyonlar Tablo 14' de görülmektedir.

Tablo 14: Anemi veri seti korelasyon matrisi



Anemi veri setindeki açıklayıcı değişkenler arasında korelasyon düşüktür. Pozitif yönlü ilişkinin en çok olduğu değişkenler 0.61 ile MCH ve HCT, negatif yönlü ilişkinin en çok olduğu değişkenler ise -0.37 ile NEUTn ve LYMp'dir. Hiçbir değişken ikilisi arasındaki korelasyon 0.70'ten büyük veya -0.70'ten küçük değildir. Anemi veri setinin iç ilişkisi yüksek olmayan değişkenlerden oluştuğu görülmektedir.

Anemi veri setinde pearson korelasyonun yanı sıra, çoklu iç ilişkiyi araştırmak adına VIF değerlerine bakılmıştır. En yüksek VIF değeri LYMp için 1.71, en düşük VIF değeri PCT için 1.01 olarak

bulunmuştur. VIF testine göre de Anemi veri setinde çoklu iç ilişki tespit edilmemiştir.

Tablo 15: Anemi veri seti için maksimum olabilirlik yöntemi ile oluşturulan model katsayıları, standart hataları, sapma ve AIC değerleri

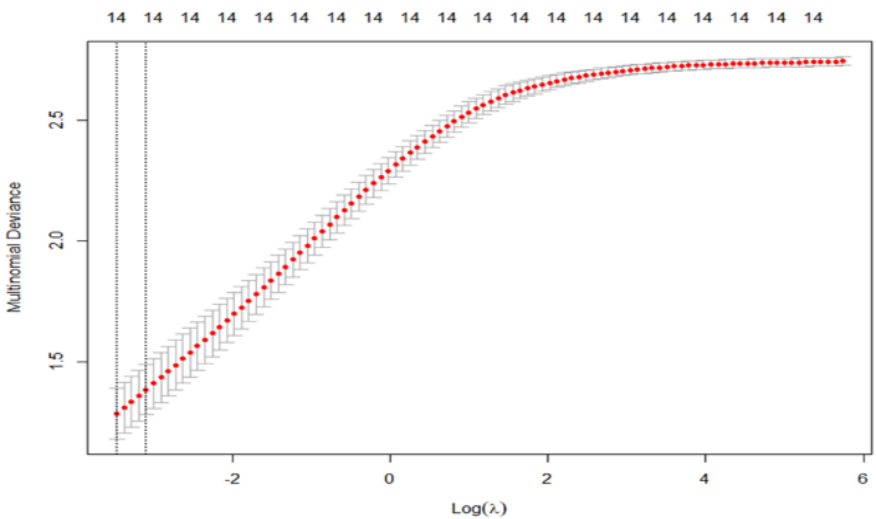
```
Call:
multinom(formula = Diagnosis ~ ., data = trainset)

Coefficients:
                              (Intercept)        WBC      LYMp       NEUTp         LYMn    NEUTn
Iron deficiency anemia           544.0630 -0.7971906 0.1502002 -0.04729316  1.47680681 2.215936
Normocytic hypochromic anemia    328.2129 -0.4630850 0.2381281  0.03587150  0.59323131 1.952104
Normocytic normochromic anemia   260.3418 -0.6599813 0.2099891  0.03569937 -0.05047731 2.338993
                                       RBC       HGB        HCT         MCV        MCH      MCHC
Iron deficiency anemia         -0.24216537 -28.36294 -0.4655952 -1.8770781 -0.1164136 -0.4670801
Normocytic hypochromic anemia   0.04372841 -27.51521 -0.5231175  0.3716185  1.1971757 -0.9338837
Normocytic normochromic anemia -2.43458297 -26.18405 -0.4398786  0.4400625  1.2146968  0.5980332
                                       PLT         PDW        PCT
Iron deficiency anemia         -0.02865797 -0.07212814 -0.5012513
Normocytic hypochromic anemia  -0.02677232  0.13561400 -1.1804290
Normocytic normochromic anemia -0.01661374  0.37693697 -1.1803382

Std. Errors:
                              (Intercept)       WBC      LYMp     NEUTp      LYMn     NEUTn       RBC
Iron deficiency anemia         0.05536112 0.6480548 0.3028508 0.2752444 0.8283829 0.9070096 0.4461309
Normocytic hypochromic anemia  0.03533083 0.4194748 0.2874287 0.2871904 0.4988985 0.8321777 0.1141167
Normocytic normochromic anemia 0.07466738 0.4199552 0.2888805 0.2872417 0.6618159 0.8347233 0.7832583
                                      HGB       HCT       MCV       MCH      MCHC         PLT       PDW
Iron deficiency anemia         0.7236602 0.4851656 0.2994723 1.3187698 1.1525942 0.010349790 0.4079059
Normocytic hypochromic anemia  0.8597196 0.5649763 0.2163550 0.6880801 0.4553819 0.008937538 0.3580335
Normocytic normochromic anemia 0.9134533 0.5650162 0.2163777 0.6879997 0.4483305 0.008946756 0.3512764
                                       PCT
Iron deficiency anemia         1.146942
Normocytic hypochromic anemia  1.791687
Normocytic normochromic anemia 1.805411

Residual Deviance: 331.7673
AIC: 421.7673
```

Sağlıklı (Healthy), referans kategori olarak alınarak ML yöntemiyle oluşturulan modelde, sapma istatistiği 331.8 ve AIC istatistiği 421.8 olarak bulunmuştur. Bu istatistiklerle beraber model katsayıları ve standard hataları Tablo 15’ de gösterilmiştir.

Tablo 17: Anemi seti ML modelinin test veri seti için kategori tahminleri

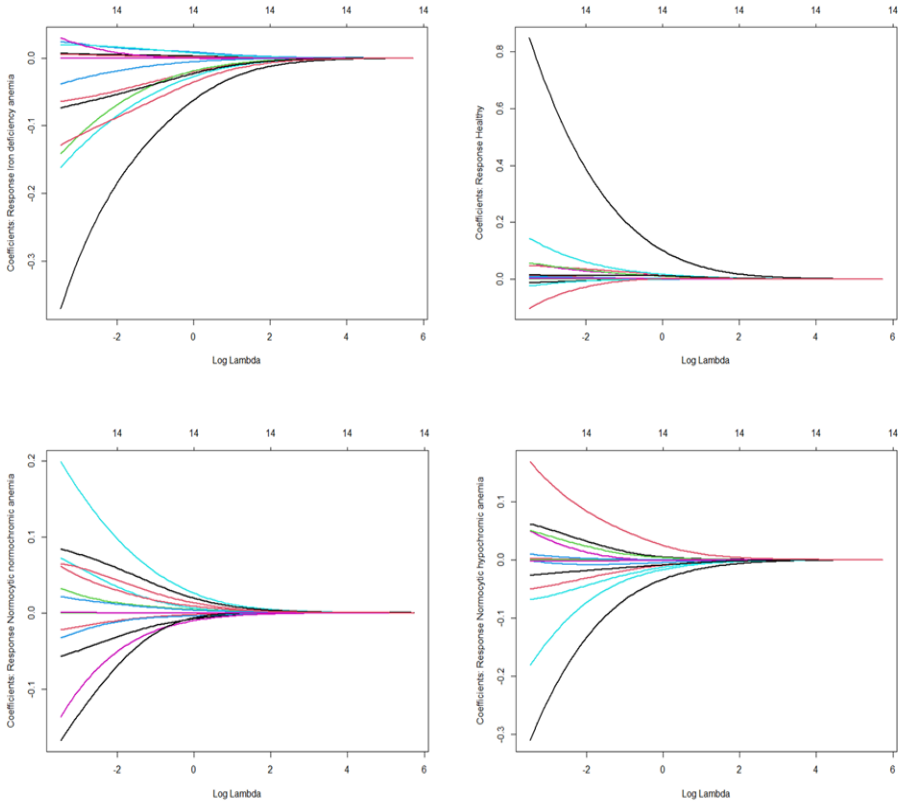
predanemia	testing_y Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	67	3	7	0
Iron deficiency anemia	0	33	2	0
Normocytic hypochromic anemia	0	2	42	1
Normocytic normochromic anemia	1	0	5	53

ML modelinin test veri seri için tahmin isabeti %90.2 olmuştur. Modelin test veri setindeki kategorilere göre tahminleri Tablo 17' de gösterilmiştir.



Şekil 3: Ridge λ ceza parametresi-sapma değişim grafiği

Ridge için 10 tekrarlı CV yöntemiyle bulunan minimum lambda değeri 0.031'dir. Şekil 3'deki soldaki kesikli düşey çizgi, en küçük sapmayı veren bu lambda değerine karşılık gelmektedir.



Şekil 4: Ridge Log λ – Model katsayıları değişim grafikleri

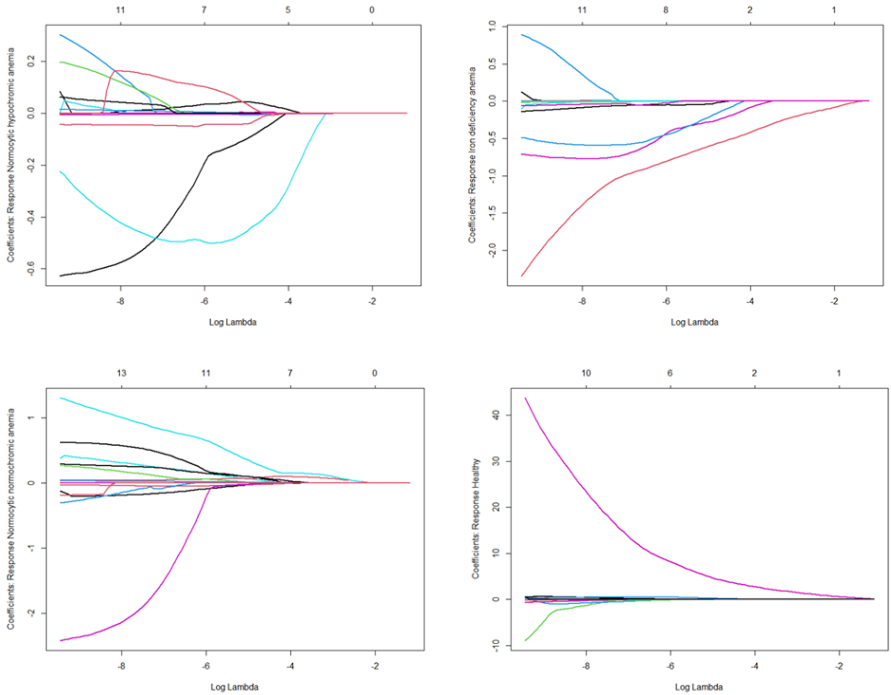
Veri setindeki dört kategorik değişkene ait model katsayılarının farklı lambda değerlerine göre değişimi Şekil 4'de gösterilmiştir.

Tablo 18: Anemi seti Ridge modelin test veri seti için kategori tahminleri

	testing_y			
predanemiaridge	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytıc normochromıc anemıa
Healthy	66	1	5	4
Iron deficiency anemia	0	33	0	0
Normocytic hypochromic anemia	1	3	44	2
Normocytic normochromic anemia	1	1	7	48

Ridge modelin test veri setindeki tahmin isabet oranı %88.4 olmuştur. Tablo 18, modelin test veri setinde kategorilere göre tahminlerini göstermektedir.

LASSO için 10 tekrarlı CV yöntemiyle bulunan minimum lambda değeri 0.00016'dır. Farklı kategorilere ait model katsayılarının ne olduğu ve farklı lambda değerlerine göre nasıl değiştiği Şekil 5'de görülmektedir.



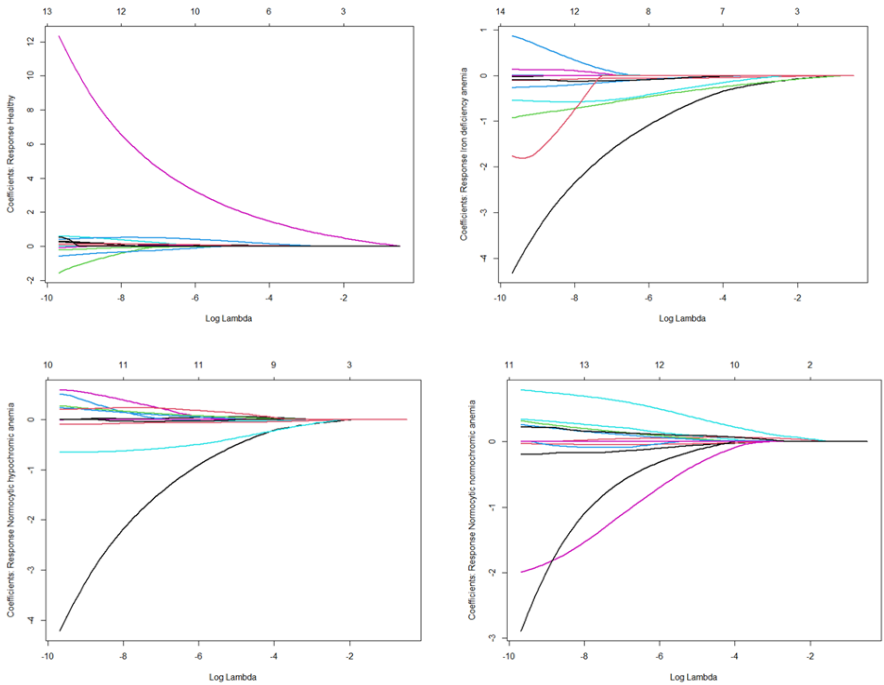
Şekil 5: LASSO Log λ – Model katsayıları değişim grafikleri

Tablo 19: Anemi seti LASSO modelinin test veri seti için kategori tahminleri

	testing_y			
predanemialasso	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	66	2	5	0
Iron deficiency anemia	0	34	0	0
Normocytic hypochromic anemia	0	2	45	1
Normocytic normochromic anemia	2	0	6	53

LASSO modelin test veri seti için tahmin isabet oranı %91.6 olmuştur. Modelin test veri setinde kategorilere göre tahminleri Tablo 19' da verilmiştir.

Elastic-net için 10 tekrarlı CV yöntemiyle bulunan minimum lambda değeri $6{,}22.10^{-5}$ 'dir. Farklı kategorilere ait model katsayılarının ne olduğu ve farklı lambda değerlerine göre nasıl değiştiği Şekil 6'da görülmektedir.



Şekil 6: Elastic-net Log λ – Model katsayıları değişim grafiği

Elastic-net modelin test veri seti için tahmin isabet oranı %91.2 olmuştur. Modelin test veri setinde kategorilere göre tahminleri Tablo 20'de gösterilmiştir.

Tablo 20: Anemi seti Elastic-net modelinin test veri seti için tahminleri

	testing_y			
anemiaelastic	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	65	2	4	0
Iron deficiency anemia	0	34	0	0
Normocytic hypochromic anemia	1	2	45	1
Normocytic normochromic anemia	2	0	7	53

Tablo 21: Anemi seti PCA yöntemiyle oluşturulan temel bileşenler ve bu bileşenlerin verideki varyansı açıklama yüzdesi

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.6338	1.3205	1.2648	1.02627	1.02210	1.00374	0.98779	0.93818	0.87943	0.79243
Proportion of Variance	0.1907	0.1245	0.1143	0.07523	0.07462	0.07196	0.06969	0.06287	0.05524	0.04485
Cumulative Proportion	0.1907	0.3152	0.4295	0.50471	0.57933	0.65129	0.72099	0.78386	0.83910	0.88395

	PC11	PC12	PC13	PC14
Standard deviation	0.73294	0.66315	0.60230	0.53377
Proportion of Variance	0.03837	0.03141	0.02591	0.02035
Cumulative Proportion	0.92233	0.95374	0.97965	1.00000

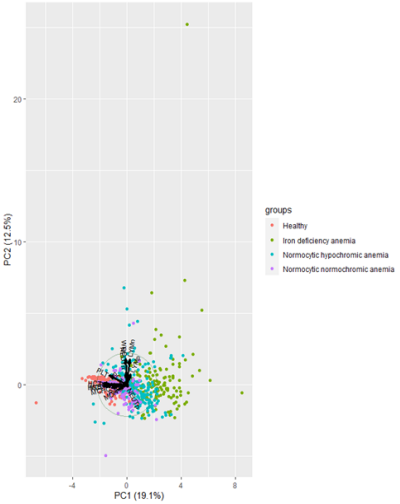
PCA için birinci temel bileşen verideki varyansın %19'unu, ikinci temel bileşen %12'sini açıklamaktadır. Verideki varyansın %84'ü, ilk 9 temel bileşen tarafından temsil edilmektedir. Tüm bileşenlerin verideki bilgiyi temsil etme yüzdesi tek tek ve kümülatif olarak Tablo 21' de gösterilmiştir. Bu bileşenlerden ilk 7'si için, veri setindeki orijinal değişkenlerin katkısı Tablo 22' de görülmektedir.

Tablo 22: Anemi setinde her bir değişkenin temel bileşenlere katkısı

```
Standard deviations (1, .., p=14):
 [1] 1.6338064 1.3204649 1.2647957 1.0262711 1.0221035 1.0037419 0.9877903 0.9381830 0.8794342 0.7924271
[11] 0.7329433 0.6631528 0.6023016 0.5337682

Rotation (n x k) = (14 x 14):
              PC1          PC2          PC3          PC4         PC5          PC6          PC7
WBC   -0.02810224  0.624362108 -0.256186944 -0.009176034  0.02132864  0.032926327  0.1278848332
LYMp   0.07460037  0.252920104  0.648770997 -0.135395776 -0.00907535 -0.004641790 -0.0620293965
NEUTp  0.02007197 -0.044131622  0.052497115 -0.107155514  0.05330677 -0.867961642  0.4587246780
LYMn   0.05447440  0.612495664  0.288744447 -0.179030058 -0.04336723  0.080161966  0.0888449968
NEUTn -0.02283222  0.298781563 -0.595453582  0.075256262  0.08867512  0.001651352  0.0803029967
RBC   -0.08349523 -0.041082011 -0.005184455  0.153778843 -0.83875811  0.109445870  0.3073169061
HGB   -0.49010221 -0.034544303  0.093224516 -0.047528743 -0.18034752  0.012513235 -0.0001479604
HCT   -0.49339941  0.034713422  0.051716622  0.097646565 -0.04890620 -0.082720395 -0.1602452155
MCV   -0.44183076 -0.030059208 -0.038181123 -0.134058590  0.11418582 -0.003404457  0.0483467479
MCH   -0.18003124 -0.121368258  0.003251783 -0.431989661  0.28629767  0.276634349  0.3845724186
MCHC  -0.30634309 -0.070493400 -0.022866827 -0.356544787  0.04374391  0.146821724  0.1364407129
PLT   -0.36147034  0.228110204  0.024809087  0.282021971  0.01168358 -0.169432545 -0.0369146953
PDW   -0.21069333  0.007880323  0.179637233  0.509733662  0.32777709 -0.025081209 -0.1298655616
PCT   -0.05113271  0.055216507 -0.162793068 -0.475256449 -0.20964805 -0.304465697 -0.6691995791
```

Verideki varyansın yaklaşık %84'ünü temsil eden ilk 9 temel bileşen ile model kurulmuştur.

Fasulye veri setinde 16 açıklayıcı değişkenin sakladığı bilginin %97'si yalnızca 3 bileşen tarafından temsil edilirken, anemi veri setinde 14 açıklayıcı değişkenin sakladığı bilginin ancak %84'ünün 9 bileşen tarafından temsil edildiği görülmektedir. Bu farklılığın temel sebebinin, veri setlerindeki değişkenler arasındaki iç ilişki düzeylerinin farklılığı olduğu öne sürülebilir.



Şekil 7: Eğitim seti verilerinin PC1-PC2 uzayında kategorilerine göre dağılımı

Yatay eksen 1.temel bileşeni, düşey eksen 2.temel bileşeni temsil etmek üzere, model kurmak için kullanılan eğitim veri setindeki 857 adet gözlemin kategorilere göre dağılımını gösteren grafik Şekil 7' de gösterilmiştir. Eğitim veri setinde uç değerler (outliers) olduğu görülmektedir.

Tablo 23: Anemi seti PCA modelin test veri seti için tahminleri

p1	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	64	1	3	7
Iron deficiency anemia	0	35	0	0
Normocytic hypochromic anemia	2	2	43	1
Normocytic normochromic anemia	2	0	10	46

PCA kullanılarak kurulan modelin test veri seti için tahmin isabet oranı %87.0 olarak bulunmuştur. Modelin test veri setinde kategorilere göre tahminleri Tablo 23' de gösterilmiştir.

Aynı yöntemler, 1073 gözlemli anemi veri setinden rastgele alınan N=50 ve N=150 gözlemli örneklemler için de uygulanmıştır. Bu örneklemler için kurulan modellerin test veri setlerindeki performansları Tablo 24 ve Tablo 25' de paylaşılmıştır.

Tablo 24: Anemi seti N=50 için 5 modelin test veri setindeki tahminleri

	testing_y			
predanemia50	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	2	0	0	0
Iron deficiency anemia	0	2	0	0
Normocytic hypochromic anemia	0	0	1	0
Normocytic normochromic anemia	1	0	2	4

ML

	testing_y			
predanemia50ridge	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	2	0	0	2
Iron deficiency anemia	0	1	0	0
Normocytic normochromic anemia	1	1	3	2

RIDGE

	testing_y			
predanemia50lasso	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	3	1	0	2
Iron deficiency anemia	0	1	0	0
Normocytic normochromic anemia	0	0	3	2

LASSO

	testing_y			
pred_anemia50elastic	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	3	1	0	2
Normocytic normochromic anemia	0	1	3	2

ELASTIC NET

p1	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	2	0	0	0
Iron deficiency anemia	0	0	0	0
Normocytic hypochromic anemia	0	0	2	0
Normocytic normochromic anemia	1	2	1	4

PCA

Tablo25: Anemi seti N=150 için 5 modelin test veri setindeki tahminleri

	testing_y			
predanemia150	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	8	0	5	2
Iron deficiency anemia	1	5	0	0
Normocytic hypochromic anemia	1	0	2	0
Normocytic normochromic anemia	0	0	3	5

ML

	testing_y			
predanemia150ridge	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	9	0	0	0
Iron deficiency anemia	1	5	0	0
Normocytic hypochromic anemia	0	0	7	1
Normocytic normochromic anemia	0	0	3	6

RIDGE

	testing_y			
predanemia150lasso	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	9	0	0	1
Iron deficiency anemia	1	5	0	0
Normocytic hypochromic anemia	0	0	9	0
Normocytic normochromic anemia	0	0	1	6

LASSSO

	testing_y			
predanemia150elastic	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	9	0	0	0
Iron deficiency anemia	1	5	0	0
Normocytic hypochromic anemia	0	0	8	1
Normocytic normochromic anemia	0	0	2	6

ELASTIC NET

pl	Healthy	Iron deficiency anemia	Normocytic hypochromic anemia	Normocytic normochromic anemia
Healthy	9	0	0	0
Iron deficiency anemia	1	5	0	0
Normocytic hypochromic anemia	0	0	5	2
Normocytic normochromic anemia	0	0	5	5

PCA

Tablo 26: Anemi veri seti için farklı N değerlerine göre test setinde model tahmin isabet oranları

GÖZLEM SAYISI	MODEL TAHMİN İSABET ORANI (%)				
	ML	**Ridge**	**LASSO**	**Elastic-net**	**PCA**
N=1073	90.2	88.4	91.6	91.2	87.0
N=150	62.5	84.3	90.6	87.5	75.0
N=50	75.0	41.6	50.0	41.6	66.6

4.Sonuç

Tablo 13 ve Tablo 26' da her bir veri seti için ve farklı gözlem sayılarına sahip örneklemler için oluşturulan 5 modelin tahmin isabet oranları paylaşılmıştır.

Tüm veri setleri ve bu veri setlerinden alınan farklı gözlem sayısına sahip örneklemler için kurulan 5 farklı modelden hangisinin ya da hangilerinin en iyi performansı gösterdiği ise Tablo 27' de gösterilmiştir. Test veri setinde en yüksek performans gösteren modellerden yüzdesel olarak eşit tahmin isabeti gösteren birden çok model varsa, bu modellerin tümüne Çizelge 3.1'de yer verilmiştir.

Tablo 27: Tüm veri setleri için en yüksek performans gösteren modeller

	DENİZKULAĞI	**ANEMİ**
TÜM VERİ	LASSO	LASSO
N=150 **N=100**	RIDGE, ELASTIC PCA	LASSO
N=50 **N=40**	RIDGE, LASSO ELASTIC, PCA	ML

Açıklayıcı değişken ikililerinin %75'inde yüksek korelasyon olan, 8 açıklayıcı değişken ve 4177 gözlemden oluşan denizkulağı veri setinin tamamı için oluşturulan modellerin birbirine çok yakın ve aynı zamanda düşük performanslar gösterdiği görülmektedir. Test veri setindeki tahmin isabet oranlarına bakıldığında en yüksek tahmin isabet oranı %56.1 ile LASSO modeline, en düşük tahmin isabet oranı ise %54.0 ile PCA modeline aittir.

Denizkulağı veri setinden rastgele alınan N=100 ve N=40 gözlemli örneklemler için regülarizasyon ve boyut azaltma teknikleriyle oluşturulan modellerin performansları stabil kalmıştır. ML yöntemi ile oluşturulan modeller ise N=100 için %52.3, N=40 için ise %44.4 tahmin isabet oranı göstermiştir. Denizkulağı veri

setinde de fasulye veri setinde olduğu gibi ML yöntemi ile oluşturulan modellerin performansının, gözlem sayısı düştükçe düştüğü görülmektedir.

Açıklayıcı değişken ikililerinin hiçbirisinde yüksek korelasyon bulunmayan, 14 açıklayıcı değişken ve 1073 gözlemden oluşan anemi veri setinin tamamı için oluşturulan modellerin hepsi yüksek tahmin performansı göstermiştir. Test veri setindeki tahmin isabet oranlarına bakıldığında en yüksek tahmin isabet oranı %91.6 ile LASSO modeline, en düşük tahmin isabet oranı ise %87.0 ile PCA modeline aittir.

N=150 için Ridge, LASSO ve Elastic net modelleri yine yüksek tahmin oranlarını korurken, PCA modelin performansı %75'e, ML modelinin performansı ise %62.5'e kadar düşmüştür. Ancak, N=150 örneklemi için en kötü performansı gösteren ML yöntemi, N=50 değeri için diğer tüm yöntemlerden daha iyi performans göstermiştir. Burada, regülarizasyon yöntemlerinin performansı kritik bir şekilde düşerken, PCA yönteminin regülarizasyon yöntemlerine göre daha iyi performans gösterdiği görülmektedir.

Sonuç itibariyle, iç ilişkisi yüksek açıklayıcı değişkenlerden oluşan Denizkulağı veri setinden alınan az gözlemli örneklemler için oluşturulan modellerin tamamında ML yönteminin tahmin performansı, diğer tüm modellerin gerisinde kalmıştır. İç ilişkisi yüksek olmayan anemi veri setinde ise ML yöntemi N=150 için tüm modeller arasında en düşük tahmin isabet oranına sahipken, gözlem sayısı 50'ye düştüğünde ise tüm modeller arasında en yüksek tahmin isabet oranına sahip olmuştur. Anemi veri seti için Ridge, LASSO,

Elastic net ve PCA modelleri N=50 için oldukça düşük performans göstermiştir.

Bu çalışmada, multinomial lojistik regresyon analizinde tüm açıklayıcı değişkenlerin sürekli olduğu durumlar için, veri setlerindeki iç ilişki düzeyine ve gözlem sayısına göre 5 farklı kestirim yönteminin performanslarına odaklanılmıştır. Başka çalışmalarda yine bu yöntemler, açıklayıcı değişkenlerin tamamının kesikli olduğu veri setleri için uygulanıp model performansları kıyaslanabilir. Bu çalışmada ele alınan yöntemler haricinde başka kestirim yöntemleri, veri setlerinin iç ilişki düzeyine göre ve açıklayıcı değişkenlerin sürekli/kesikli olma durumlarına göre gruplandırılarak uygulanabilir.

KAYNAKLAR

Aboelnaga, E., *Anemia types classification*, (20/04/2024 tarihinde https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification adresinden ulaşılmıştır.)

Hoerl, A.E., Kennard, R.W. (1970a). Ridge Regression: Biased estimation for Nonorthogonal problems. *Technometrics*, 12, pp. 55-67.

Hoerl, A.E., Kennard, R.W. (1970b). Ridge Regression: Applications Nonorthogonal problems. *Technometrics,* 12, pp. 69-82.

Mendes, R., *Abalone dataset.* (19/05/2024 tarihinde https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset adresinden ulaşılmıştır.)

Nibbering, D. and Hastie, T.J. (2022). Multiclass-penalized logistic regression, *Computational Statistics & Data Analysis,* 169, 107414.

Tabachinck, B.G. and Fidel, L.S. (2013). *Using Multivariate Statistics* (6th ed.). Pearson Education, Inc.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via Lasso, *Wiley for the Royal Statistical Society*, 58, 267-288.

Zou, H., Hastie, T. (2005). Regularization and Variable Selection via Elastic-net, *Royal Statistical Society*, 67, 301-320.

BÖLÜM III

Bayesci İstatistiğe Giriş: Ibm Spss ile Bayesci T-testleri

Hamza ERDOĞDU[1]
Harika Gözde GÖZÜKARA BAĞ[2]

GİRİŞ

Medical Outreach Team of the Drug Information Association Bayesian Scientific Working Group tarafından; klinik araştırmacı (istatistikçi olmayan) toplulukta Bayes yöntemlerini uygulamadaki algılanan en büyük engelin ne olduğu ile Bayesci ve geleneksel istatistiksel sonuçların yorumlanmasındaki mevcut anlayışın değerlendirildiği 22 soruluk bir anket düzenlemiş ve bu anket sonuçları iki ayrı makalede yayınlanmıştır (Clark, ve diğerleri, 2022); (Bray R, 2022). 323 katılımcı klinisyenin

[1] Asst. Prof., Harran University, Biostatistics and Medical Informatics, 0000-0002-5025-2367
[2] Prof., Inonu University, Biostatistics and Medical Informatics, 0000-0003-1208-4072

katıldığı ankete göre; klinik denemeler için Bayes yaklaşımlarına ilişkin yetersiz bilgi en büyük engel olarak kabul edilirken, ikinci olarak Bayesci yöntemlerin uygulamada düzenleyici rehberliğinin eksikliği ikinci engel olarak değerlendirilmiştir. İlave olarak, klinisyenlerin çoğunun hem Bayesci hem de geleneksel istatistiksel çıktıların sonuçlarını yorumlamaya çalışırken belirsizlikle karşı karşıya kaldığı sonucuna varılmıştır. Sözkonusu çalışmanın bir benzerinin Türkiye'de yapıldığına dair literatürde bir bulguya rastlanılmamakla birlikte klinik araştırmacılar arasında Bayes yöntemlerine dair Türkiye'de sağlık ve tıp alanında yapılan çalışmalar incelendiğinde, Bayes yöntemlerinin uygulamada sınırlı sayıda kullanıldığı görülmektedir.

Bu çalışmanın amacı; tıp ve sağlık alanında çalışan profesyonellerin topladıkları verileri analiz etmek için kullandıkları geleneksel yöntemlerin Bayesci karşılıklarını (bağımsız ve eşleştirilmiş örneklem t-testlerini) yorumlama ve uygulama konusunda bilgi ve becerilerini geliştirmek, böylece akademik çalışmalarında bu yöntemleri daha çok tercih etmelerine yardımcı olmaktır. Daha sonra, istatistiksel analizleri için çoğu araştırmacı tarafından tercih edilen ve kullanıcı dostu olan SPSS uygulamaları, klasik istatistiksel yöntemlerle karşılaştırmalı olarak ele alınacaktır.

Bu çalışmanın organizasyonu şu şekildedir: Öncelikle Bayesci çıkarımın temel bileşenlerinin genel hatlarıyla tanımını yapılmaktadır. Sonrasında sonsal dağılımın elde edilmesinden sonra, model parametresi θ 'nın Bayes nokta tahminlerinin nasıl hesaplandığı gösterilmektedir. Daha sonra Bayesci çıkarımı örneklemek için, Bağımsız Örneklem t-testi ile anneleri sigara içen bebeklerin doğum ağırlıklarını, anneleri sigara içmeyen bebeklerin

doğum ağırlıklarıyla karşılaştırarak ve Bağımlı Örneklem ile kan kolesterol seviyelerinin iki farklı laboratuvar sonuçları analiz edilmektedir. Son kısımda ise çalışmadan beklenen sonuçların değerlendirmesi yapılmaktadır.

Bayesci İstatistiğin Bileşenleri

Bayes istatistiklerinin temelinde yatan üç temel bileşen vardır (Rens van de Schoot, Neyer, Kaplan, Jens B. Asendorpf, & van Aken, 2014): Kısaca, bu bileşenler aşağıdaki gibi tanımlanabilir.

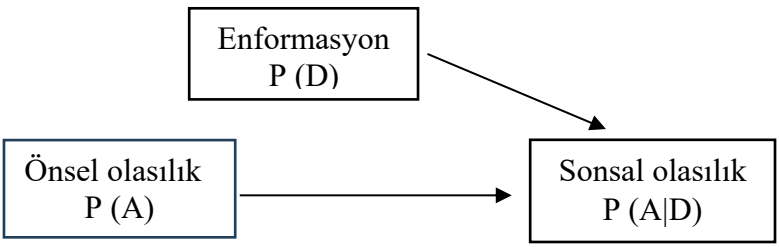
1. İlk bileşen, test edilen modelin parametrelerine ilişkin arka plan bilgisidir. *Önsel dağılım* (prior distribution) olarak adlandırılan bu ilk bileşen, verileri görmeden önce mevcut olan tüm bilgileri ifade eder, örneğin normal dağılım gibi. Bu önsel dağılımın varyansı, ilgi duyulan parametrenin popülasyon değeri hakkındaki belirsizlik seviyesini yansıtır: Varyans ne kadar büyükse, o kadar belirsizlik sözkonudur. Önsel varyans, basitçe varyansın tersi olan *kesinlik* (precision) olarak ifade edilir. Önsel varyans ne kadar küçükse, kesinlik o kadar yüksek olur ve ilk ortalamanın popülasyon ortalamasını yansıttığına dair güven o kadar artar.

2. İkinci bileşen, verilerin kendisindeki enformasyondur. Parametreler verildiğinde verilerin *olasılık fonksiyonu* (likelihood function) açısından ifade edilen gözlenen kanıttır.

3. Üçüncü bileşen, ilk iki bileşenin birleşimine dayanır ve *sonsal dağılım* (posterior distribution) adını alır.

Önceki dağılım ve mevcut veriler elde edildiğinde, bunlar daha sonra Bayes teoremi aracılığıyla birleştirilerek sonsal dağılım oluşturulur. Özellikle, bu durumda, Bayes teoremi, ön bilgimizin mevcut verilerle güncellenerek sonsal dağılım biçiminde güncellenmiş bilgi elde edildiğini belirtir.

Yukarıda açıklanan bileşenler birlikte ele alındığında Bayesci çıkarımı aşağıda verilen Şekil 1'deki gibi görselleştirmek mümkündür.



Şekil 1- Bayesci Çıkarım

Bayesci Nokta Tahmini

Bir Bayesci modelinin temel bileşenleri şu şekilde formülüze edilebilir (Puza, 2015):

- y; data
- θ; parametre
- $f(y|\theta)$ veya $F(y|\theta)$'nın veya $(y|\theta)$'nın dağılımı için seçilen, *model dağılımı*
- $f(\theta)$ veya $F(\theta)$'nın veya (θ)'nın dağılımı için seçilen, *önsel dağılımı*

göstermektedir. Burada, F, kümülatif dağılım fonksiyonunu (cdf),

f, olasılık yoğunluk fonksiyonunu (pdf) (sürekli bir rastgele değişkene uygulandığında) veya olasılık fonksiyonunu (pmf) (kesikli bir rastgele değişkene uygulandığında) belirten genel bir semboldür.

Bayesçi çıkarım, θ 'nın *sonsal olasılık dağılımı*nın belirlenmesini gerektirir. Bu aynı zamanda, θ'nın *sonsal pdf*'sini bulmaya eşdeğerdir ve aşağıdaki denklem kullanılarak yapılabilir,

$$f(\theta|y) = \frac{\mathrm{f}(\theta)f(y|\theta)}{\mathrm{f}(y)}, \qquad (1.1)$$

Burada, $\mathrm{f}(y)$: y 'nin koşulsuz (veya önsel) pdf'sini

$$\mathrm{f}(y) = \int f(y|\theta)dF(\theta) = \begin{cases} \int f(\theta)f(y|\theta)d(\theta), & \theta\ sürekli\ ise \\ \sum_{\theta}(\theta)f(y|\theta), & \theta\ kesikli\ ise. \end{cases}$$

$\mathrm{f}(y)$ 'nin Bayes denkleminde θ 'ye göre bir sabit olduğunu gözlemlendiğinde,

$f(\theta|y) = \mathrm{f}(\theta)f(y|\theta)/\mathrm{f}(y)$ denklemi,

$f(\theta|y) \propto \mathrm{f}(\theta)f(y|\theta)$ şeklinde yazılabilir, $\propto$: orantılılık işareti.

Son denklem kelimelerle şu şekilde de ifade edilebilir:

$Sonsal\ olasılık\ dağılımı\ \propto$ $önsel\ dağılım * olabilirlik\ fonksiyonu$

Sonsal dağılım veya $f(\theta|y)$ yoğunluk fonksiyonu elde edildikten sonra, model parametresi θ 'nın Bayes nokta tahminleri hesaplanabilir. En sık kullanılan üç nokta tahmini aşağıdaki gibidir:

- θ 'nın *sonsal ortalaması*,

$$E(\theta|y) = \int \theta dF(\theta|y) = \begin{cases} \int \theta f(\theta|y) d\theta, & \theta \text{ sürekli ise} \\ \sum_{\theta} \theta f(\theta|y), & \theta \text{ kesikli ise.} \end{cases}$$

- θ 'nın *sonsal modu*,

$$Mod(\theta|y) = f(\theta = m|x);\ m \in \Re$$

- θ 'nın *sonsal medyanı*,

$$Medyan(\theta|y) = P(\theta \leq m|y) \geq 1/2 \ ve \ P(\theta \geq m|y) \geq 1/2$$

Bayes yaklaşımı, istatistiksel çıkarımın yaygın sorunlarına, yüksek boyutlu veri kümelerinin ortaya çıkması nedeniyle yüksek boyutlu veri analizinin yeni sorunlarına ve gerçek hayatın karmaşık karar sorunlarına oldukça açık çözümler sunmaktadır. Buna ilave olarak temel Bayes fikirlerinin ve ölçümlerinin yorumlanması dolayısıyla aktarılması da kolaydır (Ghosh, Delampady , & Samanta, 2006).

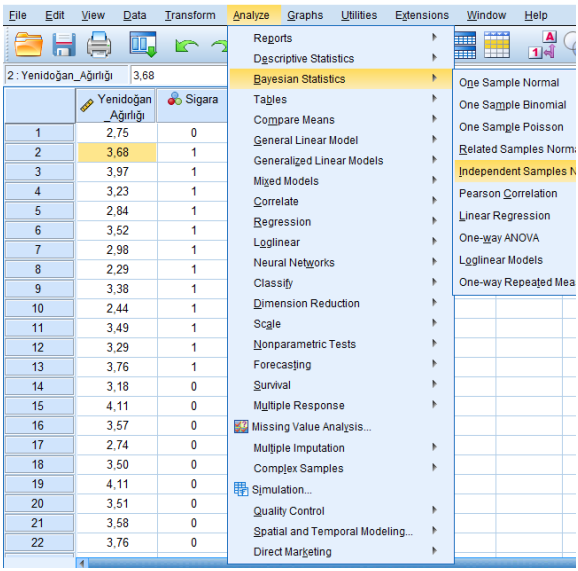
Bayesci Bağımsız Örnekler T-testleri

Birbirinden bağımsız iki grubu tanımlamak için bir grup değişkeni kullanan ve iki grup ortalamasının farkı üzerinde Bayesci çıkarım yapma seçenekleri sunan IBM SPSS Bayesci bağımsız örnek çıkarım prosedürü ile aynı zamanda Bayes Faktörlerini farklı yaklaşımlar kullanarak tahmin etmek ve varyansların bilindiği veya bilinmediği durumlar için istenen sonsal dağılımı karakterize etmek mümkündür. Bayesci Bağımsız Örnekler T-testi prosedürünün (Saneii & Doosti, 2024)'den alınan bebeklerin doğum ağırlıkları ve

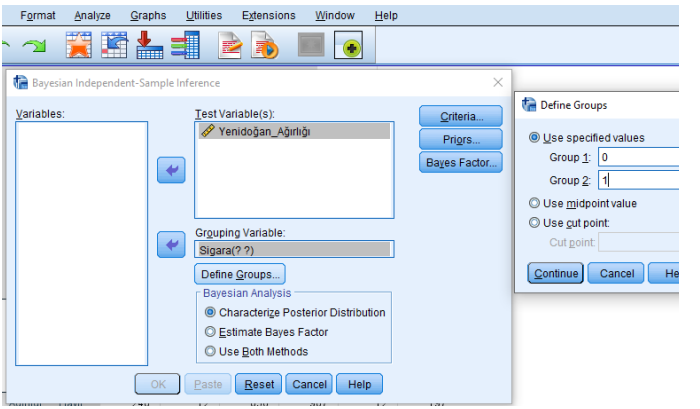
annelerin sigara kullanımı verilerine uygulaması aşağıda açıklamalı olarak verilmektedir.

Uygulama 1. Anneleri sigara içen bebeklerin doğum ağırlıklarını, anneleri sigara içmeyen bebeklerin doğum ağırlıklarıyla karşılaştırmak için bir çalışma yapıldı. 12 sigara içen anneden oluşan bir örneklemde örnek ortalaması ve standart sapma 3,24 ve 0,516 iken 12 sigara içmeyen anneden oluşan bir örneklemde ise 3,51 ve 0,444 idi. İki grubun doğum ağırlıkları eşit olabilir mi?

Bayesci bağımsız t-testini gerçekleştirmek için öncelikle bebeklerin doğum aralıkları ve grup değişkenlerinin değerleri SPSS'e girilir. Ardından **Analyze\Bayesian Statistics\Independent Samples Normal** yolu seçilir.



Böylece aşağıdaki pencerede, önce sol kutudan ilgilenilen değişken seçilir. Sonrasında **Grouping Variable** kısmında grup değişkeninin değerleri tanıtılarak **Continue** düğmesi tıklanır.



Son olarak **Use Both Methods** seçeneği işaretlendikten sonra SPSS'in Bayesci çıktıları aşağıdaki gibi elde edilecektir.

1. Çıktı

Group Statistics

	Sigara	N	Mean	Std. Deviation	Std. Error Mean
Yenidoğan_Ağırlığı	= Hayır	12	3,5117	,44380	,12811
	= Evet	12	3,2392	,51611	,14899

2. Çıktı

Bayes Factor Independent Sample Test (Method = Rouder)[a]

	Mean Difference	Pooled Std. Error Difference	Bayes Factor[b]	t	df	Sig.(2-tailed)
Yenidoğan_Ağırlığı	-,2725	,19650	1,600	-1,387	22	,179

a. Assumes unequal variance between groups.

b. Bayes factor: Null versus alternative hypothesis.

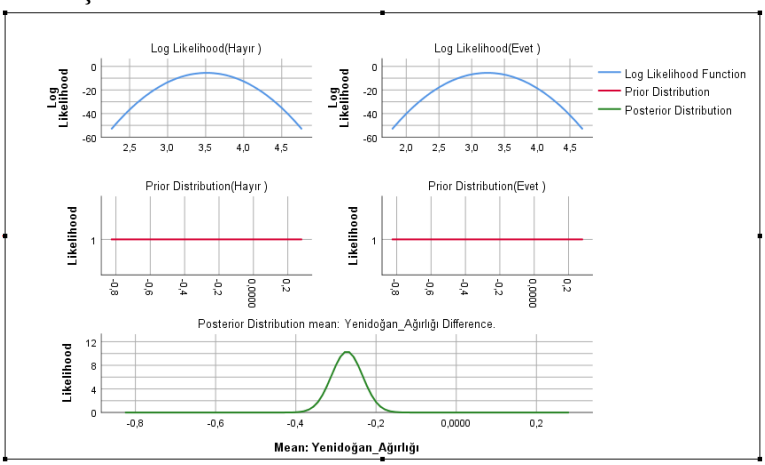
3. Çıktı

Posterior Distribution Characterization for Independent Sample Mean[a]

	Posterior			95% Credible Interval	
	Mode	Mean	Variance	Lower Bound	Upper Bound
Yenidoğan_Ağırlığı	-,2725	-,2725	,039	-,6575	,1125

a. Prior for Variance: Diffuse. Prior for Mean: Diffuse.

4. Çıktı



İlk çıktı, iki grup için ortalama değerlerin ne olduğunu söyleyen tanımlayıcı bir istatistik tablosudur. Sigara içmeyen kadınların bebeklerinin doğum ağırlıkları ortalaması 3,51 iken diğer grup ortalaması 3,24. Aradaki 0,27'lik fark büyük olmayan bir fark olarak nitelendirilebilir.

İkinci çıktı, ilk Bayesci hesaplama olarak Bayes Faktör değeri üreten tablodur. Örneğimizde, 1,60'lık bir değer görülmektedir.

Bunun yorumu; sıfır hipotezi altında alternatif hipotez altında olduğundan 1,6 kat daha büyük bir fark görme olasılığının olduğu şeklindedir. Yani bu, sıfır hipotezinin doğru olduğuna dair nispeten daha fazla kanıt olduğu anlamına gelmektedir. Yine de, Bayes faktörünün her iki hipotezden biri için hala güçlü bir destek sağlamadığını söylemek gerekmektedir.

Üçüncü çıktı, Bayesci hesaplamalardan biri olan Güvenilir Aralık (Credible Interval) değerinin olduğu tablodur. Güvenilir Aralık parametrenin kendisi hakkında bir olasılık ifadesi olarak yorumlanır. Örneğimiz için, %95 olasılıkla gerçek parametre değerinin, -0,6575 ile 0,1125 arasında herhangi bir yerde olabileceğini ifade etmektedir.

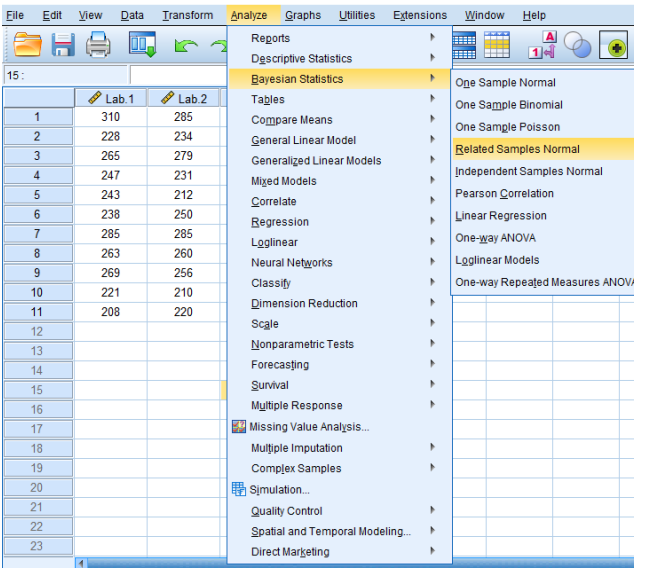
Dördüncü çıktı, mevcut olan bilgi düzeyinin yeni kanıtlara göre güncellemesine olanak tanıyan Bayesçi çıkarımın temel bileşenleri olan önsel olasılık, olabilirlik ve sonsal olasılıkların dağılımlarını göstermektedir. Yenidoğan ağırlıkları farkının sonsal dağılım değerlerinin [-0,4 ile -0,2] aralığında olduğu görülmektedir. Bu aralık değerlerinin olduğu sonsal dağılımının ortalaması hesaplandığında ise bir önceki çıktıda bulunan -0,2725 sonsal ortalama değeri bulunur.

Bayesci Bağımlı Örnekler T-Testi

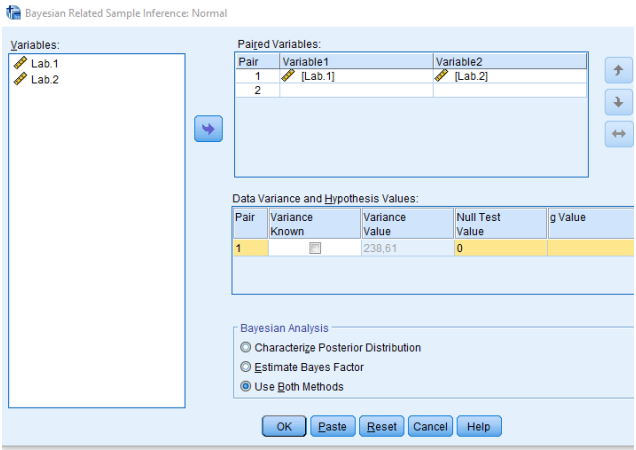
Bayesci ilişkili örnek çıkarım tasarımı ile çiftler halinde belirtilen değişkenlerin ortalama farkı üzerinde Bayes analizi yapılabilir. Bayesci Bağımlı Örnekler Testi prosedürünün (Saneii & Doosti, 2024)'den alınan kan kolestrol seviyelerinin iki farklı laboratuvar sonuçlarının olduğu uygulama açıklamalı olarak aşağıda verilmektedir.

Uygulama 2. Kan kolesterol seviyelerini belirlemek için 11 kişiden iki kan örneği alındı ve iki farklı laboratuvara gönderildi. Bu iki laboratuvarın sonuçları eşit kabul edilebilir mi?

SPSS'te Bayesci bağımlı örnek t-testini uygulamak için alınan laboratuvar verilerinin ayrı bir sütuna (örneğin, laboratuvar 1 ve laboratuvar 2) girilmesi gerektiğine dikkat edilmelidir. Ardından aşağıdaki pencereyi açmak için **Analyze\Bayesian Statistics\Related Samples Normal** yolunu seçilir.



Bu pencerede, önce sol kutudan eşleştirilmiş değişkenleri seçilir ve **Paired Variables** adı verilen sağ kutuya girin. Ardından, **OK** düğmesine basılarak test yürütülür.



Bayesci Bağımlı (Eşleştirilmiş) t-testi için aşağıdaki SPSS çıktıları elde edilir:

1. Çıktı

Bayes Factor for Related-Sample T Test

	N	Mean Difference	Std. Deviation	Std. Error Mean	Bayes Factor	t	df	Sig.(2-tailed)
Lab.1 - Lab.2	11	5,00	15,447	4,657	2,665	1,074	10	,308

Bayes factor: Null versus alternative hypothesis.

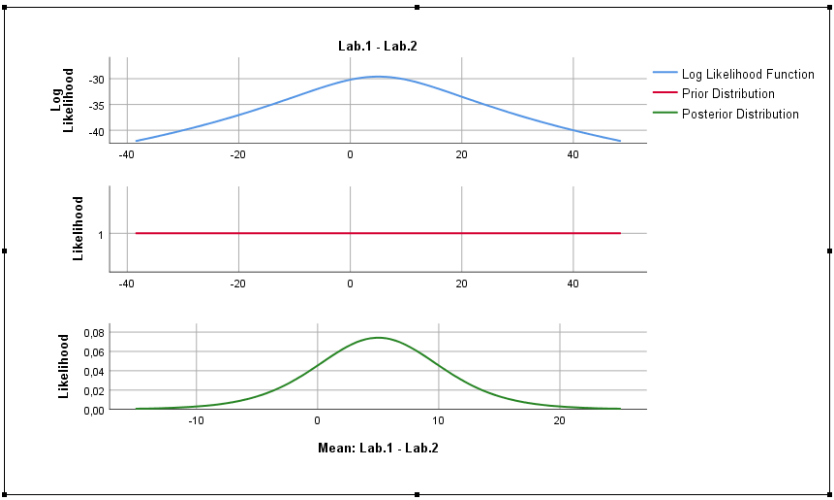
2. Çıktı

Posterior Distribution Characterization for Related-Sample Mean Difference

	N	Posterior			95% Credible Interval	
		Mode	Mean	Variance	Lower Bound	Upper Bound
Lab.1 - Lab.2	11	5,00	5,00	36,152	-7,01	17,01

Prior on Variance: Diffuse. Prior on Mean: Diffuse.

3. Çıktı



İlk çıktı, Bayesci bağımlı örnek t-testinin Bayes Faktör değerini üreten hesaplayan tablodur. Örneğimizde, Bayes Faktör=2,665 olarak bulunmuştur. Yorumu; sıfır hipotezi altında alternatif hipotez altında olduğundan yaklaşık 2,7 kat daha büyük bir fark görme olasılığının olduğudur. Sıfır hipotezinin doğru olduğuna dair görece daha fazla kanıt olduğu söylenebilir.

İkinci çıktıya göre, kan kolestrol seviyelerinin Lab1 ve Lab2 laboratuvar sonuçlarının farkının Güvenilir Aralık değerinin alt sınırı -7,01 iken üst sınır 17,01'dir. Bu değer, %95 olasılıkla gerçek parametre değerinin bu aralık arasında herhangi bir yerde olabileceğini ifade etmektedir.

Üçüncü çıktı, kan kolestrol seviyelerinin Lab1 ve Lab2 laboratuvar sonuçlarının farkı sonsal dağılım değerlerinin [-10 ile 20] aralığında olduğu görülmektedir. Bu aralık değerlerinin olduğu sonsal dağılımının sonsal ortalama değeri 5,0 bulunur.

Bayesci analizin uygulamasına hızlı bir giriş niteliği taşıyan bu çalışmada teorik bilgi gerektiren önseller, yakınsama (convergence) ve tohumların (seeds) SPSS paket programındaki diğer seçenek ve uygulamalarına dair birçok sorun kapsanamamıştır. Bunlar ve daha ileri konuların ele alındığı daha fazla okuma için bakınız (Berry, 2005), (Kruschke, 2010), (Hoff, 2009), (McElreath, 2020).

Bu uygulamalar ile birlikte tıp ve sağlık alanında çalışan profesyonellere, Bayesci bağımsız ve eşleştirilmiş örneklem t-testlerini SPSS ile uygulama ve yorumlama konusunda yeni bilgiler sağlanmaktadır. Bunun bir sonucu olarak akademik çalışmalarda sözkonusu yöntemleri daha çok tercih edilmesi beklenmektedir.

KAYNAKÇA

Berry, D. A. (2005). Berry, D.A. (2005). Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis (Cilt 2). Clinical Trials.

Bray R, H. A. (2022). Why are there not more Bayesıan clinical trials? Perceptions and interpretation of Bayesian and conventional statistics arnong medical researchers involved drug development. THER INNOV REGUL SCI. doi:https://doi.org/10.1007/ s43441-022-00482-1

Clark J, M. N. (2022). Why are not there more Bayesian clinical trials? Perceived barriers and educational preferences among medical researchers involved in drug development. THER INNOV REGUL SCI.

Hoff, P. D. (2009). A First Course in Bayesian Statistical Methods. Springer Texts in Statistics. doi:doi:10.1007/978-0-387-92407-6

Jayanta K. Ghosh, M. D. (2006). An Introduction to Bayesian Analysis Theory and Practice. Springer Texts in Statistics .

Kruschke, J. (2010). Doing Bayesian Data Analysis: A Tutorial with R and BUGS (Cilt 7). Europe's Journal of Psychology.

McElreath, R. (2020). Statistical Rethinking: A Bayesian Course with Examples in R and STAN (2 b.). Chapman and Hall/CRC. doi: https://doi.org/10.1201/9780429029608

Puza, B. (2015). Bayesian methods for statistical analysis. ANU eView.

Rens van de Schoot, J. D. (2014). A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. Child Development. doi: https://doi.org/10.1111/cdev.12169

Saneii, S. H. (2024). Practical Biostatistics for Medical and Health Sciences. Springer. doi:https://doi.org/10.1007/978-981-97-3083-4

BÖLÜM IV

Ibm Spss Kullanarak Bayesci Regresyon Analizi

Hamza ERDOĞDU[1]
Harika Gözde GÖZÜKARA BAĞ[2]

GİRİŞ

Klasik (Sıklıkçı) istatistik ve Bayesci istatistik, olasılıksal süreçlerden istatistiksel bilgi üretirken kullanılan iki yaklaşımdır. Klasik yaklaşımda ve istatistiksel tahmin süreçlerinde elde edilen sonuçlar yalnızca mevcut verilerden türetilirken, Bayesci yaklaşımda, bunlar ön bilgi veya beklentiler ile mevcut verilerin bir karışımından türetilmektedir.

Türkiye'de sağlık ve tıp alanlarında çalışan araştırmacı ve akademisyenlerin çoğu, lisans biyoistatistik eğitimlerinde klasik

[1] Asst. Prof., Harran University, Biostatistics and Medical Informatics, 0000-0002-5025-2367
[2] Prof., Inonu University, Biostatistics and Medical Informatics, 0000-0003-1208-4072

istatistik tabanlı testler (bağımsız t-testleri, eşleştirilmiş t-testleri, tek yönlü varyans analizi, korelasyon analizi ve doğrusal regresyon analizi) konusunda eğitim almaktadır. Bayesci istatistik konuları ve uygulamaları lisans biyoistatistik müfredatında yer almamaktadır. Bunun bir sonucu olarak, yapılan çalışmalarda ve danışmanlık yapılan tezlerde ele alınan veri kümelerinin analizinde klasik istatistiksel yöntemler tercih edilirken, Bayesci yöntemlerin tercih edildiği akademik çalışmaların oldukça nadir olduğu görülmektedir.

Bu çalışma, araştırmacıların ve uygulayıcıların klasik istatistikteki temel kavramların Bayesci eşdeğerlerini teorik olarak açıkladıktan sonra, Bayesci yöntemler kullanılarak veri analiz etme aşamalarını anlamalarına yardımcı olmayı amaçlamaktadır. Bu amaçla, regresyon analizi ve korelasyon analizleri için Bayesci istatistiksel çıkarım prosedürlerini açıklamaktadır. Ayrıca, çalışma, uygulamalar sonunda elde edilen Bayesci istatistik sonuçlarının nasıl yorumlanacağını göstermektedir.

Çalışmanın tasarımı şöyledir: İlk olarak Bayesci paradigmanın prensipleri ve çalışma şekli anlatılmaktadır. Sonrasında sıklıkçı ve Bayesci yaklaşımlar arasındaki farklılıklar özet bir tablo şeklinde verilmiştir. Daha sonra regresyon ve korelasyon analizleri hakkında temel bilgiler verildikten sonra uygulamalara geçilmiştir. Son olarak uygulamadan elde edilen Bayesci istatistik sonuçlarının yorumları yapılmış ve değerlendirme kısmı ile çalışma sonlandırılmıştır.

YÖNTEM

Sıklıkçı ve Bayesci çıkarımın her ikisi de olasılık teorisine dayanmakla birlikte olasılığın yorumlanması iki düşünce okulu için çok farklıdır (Yang & Novick, 2019); sıklıkçı paradigmada,

parametreler sabit kabul edilip istatistiksel çıkarım, verilerin örnekleme dağılımına dayanarak yapılmakta ve istatistiksel bir prosedürün performansı, aynı koşullar altında tekrarlanan örnekleme ile karakterize edilirken; Bayesci paradigma bilinmeyen parametreleri sabit olarak görmeyip olayın olası olduğuna dair önceden bir inançla başlar ve ardından mevcut deneyden gelen veriler kullanılabilir hale geldiğinde bilgileri günceller:

i. θ, y ve $\tilde{y}$ sırasıyla ilgilenilen parametreyi, gözlemlenebilir verileri ve gelecekteki verileri göstersin. Bayes yöntemleri, θ ve y 'yi $\pi(\theta)$ parametreleri için bir önsel dağılım ve $\pi(y \mid \theta)$ verileri için bir koşullu olasılık fonksiyonu ile tanımlayarak başlar. İkincisi, y gözlendikten sonraki olasılık fonksiyonu olarak da adlandırılır ve genellikle $L(\theta|y)$olarak gösterilir. Sıklıkçı paradigma, bilinmeyen parametre θ ile ilgili tüm bilgilerin deney tamamlandıktan sonra olasılık $\pi(y|\theta)$ içinde yer aldığına inanır ve dolayısıyla θ ile ilgili tüm istatistiksel çıkarımları yalnızca $\pi\ (\ y \mid \theta\ )$ üzerine temellendirir.

ii. Bayesci yaklaşımda, önsel bilgi Bayes Teoremi ile güncellenerek bir sonsal dağılım elde edilir:

$$\pi(\theta \mid y) = \frac{\pi(\theta)\pi(y|\theta)}{\pi(y)}, \qquad (1.1)$$

öyle ki $\pi(y) = \int \pi(\theta)\, \pi(y|\theta) d\theta$.

iii. Sonsal dağılımın iki temel bileşeni vardır: $\pi(\theta)$ ve $\pi(y|\theta)$. θ 'nin önsel dağılımı olarak adlandırılan ilki,

mevcut veriler gözlenmeden önceki parametrelere ilişkin bilgileri temsil ederken, ikincisi ise θ'ya bağlı olarak mevcut verileri gözlemleme olasılığıdır. Birlikte, mevcut verilere kadar parametrenin toplam bilgisini temsil ederler. Bayes Teoremi ayrıca, mevcut sonsalı gelecekteki sonsal dağılımın önsel değeri olarak kullanarak yeni veriler mevcut oldukça sonsal dağılımı güncellemek için bir araç sağlar.

Klasik ve Bayesci Yaklaşımlar Arasındaki Farklılıklar

Sıklıkçı ve Bayesci yaklaşımlar arasındaki felsefi ve metodolojik farklılıklar (Goligher, Heath, & Harhay, 2024) ve (van de Schoot, ve diğerleri, 2014)'den derlenerek aşağıda Tablo 1'de verilmiştir.

Tablo 1. *Sıklıkçı ve Bayesci Yaklaşım Arasındaki Farklılıklar*

	Sıklıkçı istatistik	**Bayesci istatistik**
Olasılık teriminin anlamı	Olayların beklenen sıklığı	Bir hipoteze olan inanç derecesi
Temel yaklaşım	Sıfır hipotezi altında gözlemlenen verilerin beklenen sıklığına odaklanarak bilimsel çıkarımı öznel bir inanç derecesinden ziyade sözde nesnel verilere dayandırır	Bir hipotez hakkında mevcut bilgileri ve belirsizliği (yani inanç derecesini) olasılık dağılımlarıyla nicel olarak temsil eder
Çıkarım için önceden tanımlanmış olasılık eşiği	Genellikle, $p<0{\cdot}05$	Bir hipoteze olan inancı tanımlayan tek bir sonsal olasılık değeri yoktur

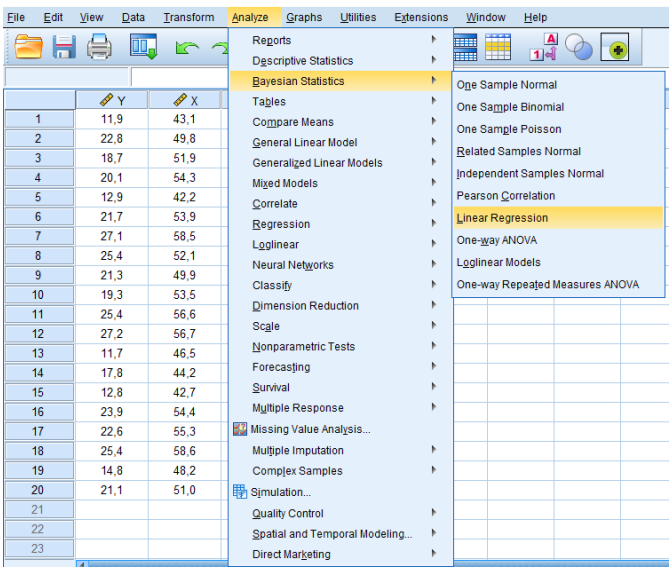
Önceki bilgilerin dikkate alınması	Hiçbir değerlendirme yapılmaz (çıkarımın yalnızca çalışma verilerine dayanması gerektiği konusunda ısrar eder)	Sonsal olasılığı tahmin etmek için önsel olasılığın belirtilmesini gerektirir
p-değerinin tanımı	Yokluk (Null) hipotezinin popülasyonda doğru olduğu varsayılarak aynı veya daha uç verilerin gözlemlenme olasılığı	Yokluk hipotezinin olasılığı
Büyük örneklere ihtiyaç duyulur mu?	Genellikle, normal teori tabanlı yöntemler kullanıldığında	Gerek yok
Ön bilginin dahil edilmesi mümkün mü?	Hayır	Evet
Modeldeki parametrelerin doğası?	Bilinmeyen ancak sabit	Bilinmeyen ve bu nedenle rastgele
Populasyon parametresi	Gerçek değerde	Belirsizliği yansıtan değerlerin dağılımı
Belirsizlik şu şekilde tanımlanır:	Sonsuz tekrarlanan örnekleme fikrine dayanan örnekleme dağılımı	Populasyon parametresi için olasılık dağılımı
Tahmin aralıkları	Güven aralığı (Confidence interval): Popülasyondan alınan sonsuz sayıda örnekten, bunların %95'i gerçek popülasyon değerini içerir	Güvenilirlik aralığı (Credibility interval): Popülasyon değerinin aralığın sınırları içinde olma olasılığı %95'tir

BULGULAR

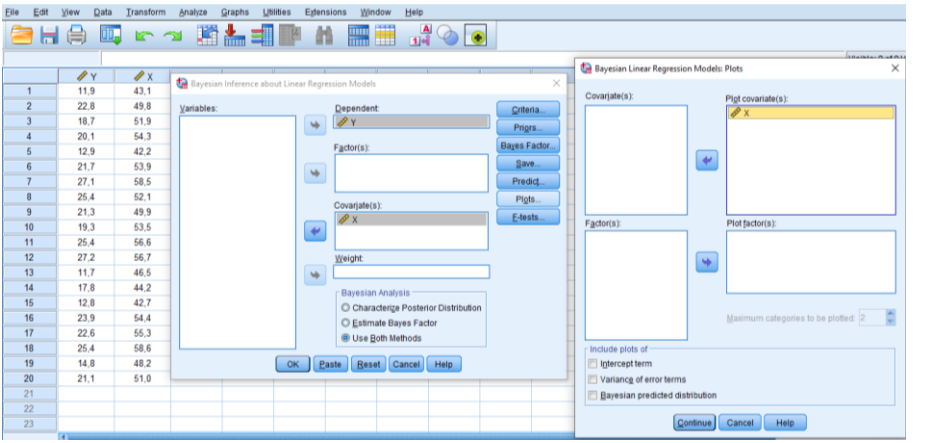
Regresyon Analizi

Regresyon analizi, tıbbi verilerin analizi için önemli bir istatistiksel yöntemdir. Birden fazla faktör arasındaki ilişkilerin belirlenmesini ve nitelendirilmesini sağlar. Ayrıca prognostik olarak ilgili risk faktörlerinin tanımlanmasına ve bireysel prognoz için risk puanlarının hesaplanmasına olanak verir (Schneider, Hommel, & Blettner, 2010).

Uygulama 1. Bir çalışmada, elde edilmesi zor olan sporcuların vücut yağ yüzdesi (Y) ve uyluk çevresi (X) gibi kolay elde edilen antropometrik ölçümle kestirilmek istenmektedir. 20 sporcuya ilişkin ölçümler verilmiştir. Bayesci Doğrusal Regresyon prosedürünün (Alpar, 2017)'den belli kısmı alınan vücut yağ yüzdesi verilerine uygulanması açıklamalı bir şekilde aşağıda verilmiştir.



Bayesci regresyon analizi için öncelikle vücut yağ yüzdesi (Y) ve uyluk çevresi (X) değişkenlerinin SPSS'e girilmesinin ardından **Analyze\Bayesian Statistics\Linear Regression** yolu seçilir.



Sonrasında sol kutudan vücut yağ yüzdesi (Y) değişkeni bağımlı değişken olarak **Dependent** kutucuğuna, uyluk çevresi (X) değişkenini de sürekli bir açıklayıcı değişken olduğu için **Covariate(s)** kutucuğuna alınır. Son olarak **Use Both Methods** seçeneği işaretlendikten sonra aşağıda verilen Bayesci regresyon analizi sonuçlar elde edilir.

1. Sonuç

Bayes Factor Model Summary[a,b]

Bayes Factor[c]	R	R Square	Adjusted R Square	Std. Error of the Estimate
55922,587	,878	,771	,758	2,510

a. Method: JZS

b. Model: (Intercept), X

c. Bayes factor: Testing model versus null model (Intercept).

2. Sonuç

Bayesian Estimates of Coefficients[a,b,c]

Parameter	Posterior Mode	Posterior Mean	Posterior Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
(Intercept)	-23,634	-23,634	36,007	-35,520	-11,749
X	,857	,857	,014	,625	1,088

a. Dependent Variable: Y

b. Model: (Intercept), X

c. Assume standard reference priors.

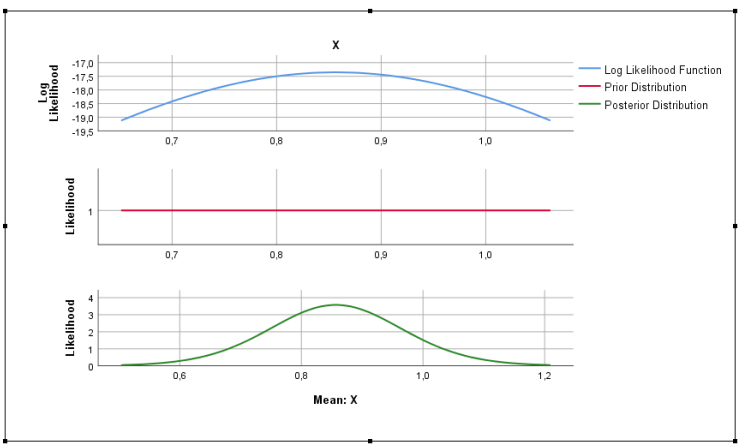
3. Sonuç

Bayesian Estimates of Error Variance[a]

Parameter	Posterior Mode	Posterior Mean	Posterior Variance	95% Credible Interval Lower Bound	95% Credible Interval Upper Bound
Error variance	5,671	7,089	7,179	3,598	13,780

a. Assume standard reference priors.

4. Sonuç



Yukarıda izlenen aşamalar itibariyle IBM SPSS Bayesci regresyon analizi dört sonuç üretmektedir. İlk olarak Bayes Faktörü=

55922,587 değerinin hesaplandığı sonuç tablosu verilmektedir. Bu faktör, sadece sabitin yer aldığı null (yokluk) modeline göre uyluk çevresi değişkeninin yer aldığı alternatif modelde fark görme olasılığının yaklaşık 55923 kat daha fazla olduğunu söylemektedir. Diğer bir ifadeyle, alternatif modelin çok büyük oranda yokluk modeline göre daha fazla kanıtlandığını göstermektedir.

İkinci sonuç tablosuna bakıldığında regresyon katsayılarının Bayesci tahmincilerinin sonsal istatistiklerinin ve Güvenilir Aralık (Credible Interval) hesaplandığı görülmektedir. Sabit terim katsayısı için sonsal ortalamalası -23,634 ve Güvenilir Aralık alt ve üst sınır değerleri [-35,520 ile -11,749] olarak hesaplanmıştır. Benzer şekilde eğim katsayısı için sonsal ortalama değeri 0,857 ve Güvenilir Aralık sınır değerleri [0,625 ile 1,088] olarak bulunmuştur. Buna göre sabit terim ve eğim katsayılarının gerçek parametre değerlerinin %95 olasılıkla yukarıda verilen aralıkların herhangi bir yerinde olacağı şeklinde yorumlanır.

Üçüncü olarak, regresyon analizinde önemli bir yeri olan hata varyansı (error variance)'nın sonsal istatistiği ve Güvenilir Aralığının hesaplanan değerleri bağımsız bir tablo olarak verilmektedir. Hata varyansının sonsal ortalama değeri 7,089 olarak bulunurken, hata varyansının gerçek tahmininin [3,598 ile 13,780] aralığı içinde yer alacağı olasılığı %95'tir.

Dördüncü sonuç tablosunda eğim katsayısının ortalamasının sonsal dağılımının elde edilmesinde kullanılan önsel olasılık ile log olabilirlik fonksiyonun dağılımları verilmektedir. Sonsal dağılımın ortalamasının değerinin 0,8 ile 1,0 arasında olup 0,8'e daha yakın

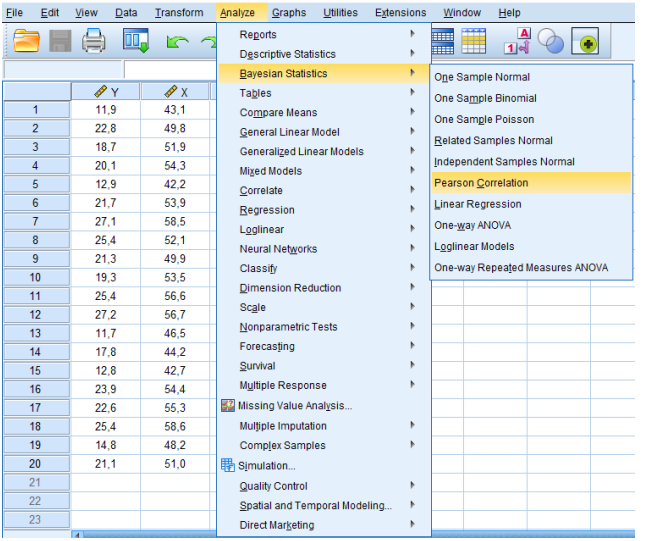
olduğu görülmektedir. Spesifik olarak bu sayı, ikinci tabloda verilen eğim katsayısının sonsal ortalama değeri olan 0,857'dir.

Korelasyon Analizi

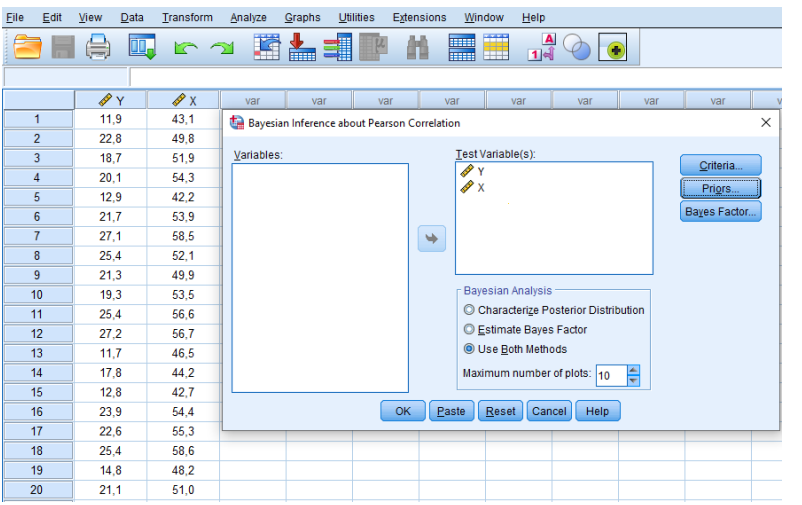
Korelasyon en geniş anlamıyla değişkenler arasındaki ilişkinin bir ölçüsüdür. Korelasyonlu verilerde, bir değişkenin büyüklüğündeki değişim, aynı (pozitif korelasyon) veya zıt (negatif korelasyon) yönde başka bir değişkenin büyüklüğündeki değişimle ilişkilidir. Çoğu zaman, korelasyon terimi iki sürekli değişken arasındaki doğrusal bir ilişki bağlamında kullanılır ve Pearson ürün-moment korelasyonu olarak ifade edilir. Pearson korelasyon katsayısı genellikle ortak normal dağılımlı veriler (iki değişkenli normal dağılım izleyen veriler) için kullanılır (Schober, Boer, & Schwarte, 2018).

Uygulama 2.

IBM SPSS'de Pearson korelasyon katsayısı hakkındaki Bayesci çıkarımın tasarımı, Bayes faktörlerini tahmin ederek ve sonsal dağılımları karakterize ederek Bayesci çıkarımlar yapabilme olanağı sunmaktadır. Bu amaçla yukarıda verilen 20 sporcunun vücut yağ yüzdesi (Y) ve uyluk çevresi (X) ölçümleri veri seti kullanılarak uygulanan Bayesci Pearson Korelasyon analizinin aşamaları ve elde edilen sonuçlar aşağıdadır:



Sporcuların uyluk çevresi (X) değişkenin büyüklüğündeki değişimin vücut yağ yüzdesi (Y) değişkeninin büyüklüğündeki değişimle ilişkisinin Bayesci korelasyon analizi için **Analyze\Bayesian Statistics\Pearson Correlation** yolu seçilir.



Sonrasında korelasyon ilişkileri araştırılan vücut yağ yüzdesi (Y) ve uyluk çevresi (X) değişkenleri **Test Variable(s)** kutucuğuna alınır. Son olarak **Use Both Methods** seçeneği işaretlenerek aşağıdaki üç sonuç tablosu elde edilir.

1. Sonuç

Bayes Factor Inference on Pairwise Correlations[a]

		Y	X
Y	Pearson Correlation	1	,878
	Bayes Factor		,000
	N	20	20
X	Pearson Correlation	,878	1
	Bayes Factor	,000	
	N	20	20

a. Bayes factor: Null versus alternative hypothesis.

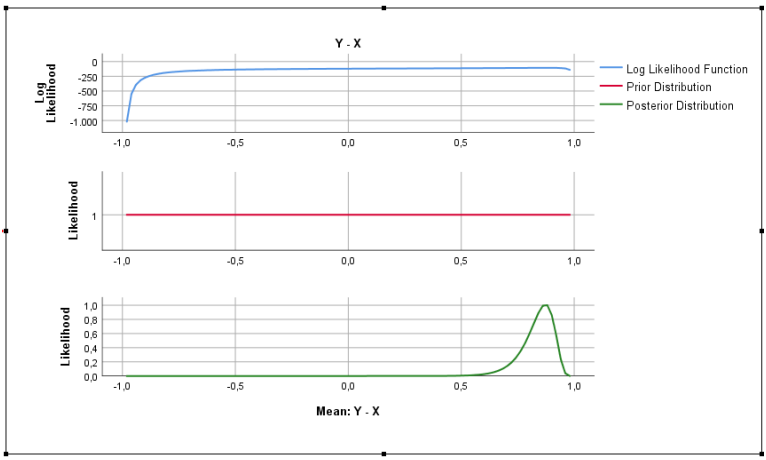
2. Sonuç

Posterior Distribution Characterization for Pairwise Correlations[a]

			Y	X
Y	Posterior	Mode		,873
		Mean		,838
		Variance		,005
	95% Credible Interval	Lower Bound		,699
		Upper Bound		,950
	N		20	20
X	Posterior	Mode	,873	
		Mean	,838	
		Variance	,005	
	95% Credible Interval	Lower Bound	,699	
		Upper Bound	,950	
	N		20	20

a. The analyses assume reference priors (c = 0).

3. Sonuç



IBM SPSS Bayesci korelasyon analizi üç adet sonuç tablosu üretmektedir. İlk tablo Bayes Faktörü çıkarımına dair bir tablo olup, Bayes Faktörü= 0,000 olarak hesaplanmıştır. Bu küçük bir değer, sıfır hipotezinin doğru olma şansının alternatif hipotezle karşılaştırıldığında oldukça küçük olduğu anlamına gelmektedir.

İkinci sonuç tablosu, vücut yağ yüzdesi (Y) ve uyluk çevresi (X) değişkenleri arasındaki korelasyon katsayısının sonsal dağılım karakteristiklerinin hesaplandığı tablodur. Bayesci korelasyon katsayısının sonsal istatistiklerinin ve Güvenilir Aralığının büyüklükleri hesaplanmıştır. Korelasyon katsayısı için sonsal mod değeri 0,873, sonsal ortalama değeri 0,838 bulunmuştur. Bunun yanında Güvenilir Aralık sınır değerleri ise [0,699 ile 0,950] hesaplanmıştır. Buna göre korelasyon katsayısının gerçek parametre değerinin %95 olasılıkla 0,699 – 0,950 aralığının içinde olacağı şeklinde yorumlanır.

Üçüncü sonuç tablosu, vücut yağ yüzdesi (Y) ve uyluk çevresi (X) değişkenleri arasındaki korelasyon katsayı ortalamasının sonsal dağılımı elde edilirken hangi önsel dağılım ve log olabilirlik fonksiyonun dağılımının kullanıldığını göstermektedir. Vücut yağ yüzdesi (Y) ve uyluk çevresi (X) değişkenlerinin korelasyon katsayısının sonsal dağılım ortalaması değeri 0,5 ile 1,0 arasında olup 1,0'a daha yakın olan 0,833'dür.

SONUÇ

(Depaoli & Van de Schoot, 2017); Bayes istatistiklerini kullanmanın çok cazip olmakla birlikte, bu konuda yeterli bilgi olmadan uygulamanın üç sebepten dolayı tehlikeli olabileceğini ifade etmektedir: önsellerin potansiyel etkisi, Bayes özelliklerinin ve sonuçlarının yanlış yorumlanması ve Bayes sonuçlarının uygunsuz şekilde raporlanması.

Bayesci istatistiğe dair temel bilgiler ve Bayes paradigması altında yürütülen klinik denemelerin mevcut en son klasik klinik deneme metodolojisinden nasıl farklı olduğu konusunda açık ve basit açıklamalar, tıp fakültesinin erken temel bilim yıllarında, öğrencilerin inancı bilgiden ayırmayı öğrenebilecekleri zaman başlayabilir (istatistikteki ilk klasik ders genellikle tıp fakültesinin ilk iki yılında sunulur). Ayrıca, Bayes paradigması araştırma asistanlarının ve klinik deneme proje yöneticilerine aldıkları biyoistatistik eğitimleri sırasında gösterilebilir. Böylece erken bir tanıtım, yeni başlayan öğrencilere sorunsuz bir öğrenme eğrisi sağlayabilir ve deneyimli Bayes klinik araştırmacıları olma yollarını kolaylaştırabilir (Moyé, 2007).

Eğitim ve öğretim ile, belirsizlik ölçülerini tahmin etme, anlama ve yorumlama ile ilgili bilgi ve beceriler geliştirmek ve bunların klasik veya Bayesci yaklaşımlar altında kullanımına yönelik engelleri azaltmak mümkün olabilir (Hespanhol, Vallio, Costa, & Saragiotto, 2019).

KAYNAKÇA

Alpar, R. (2017). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler* (5 b.). Ankara: Detay Yayıncılık.

Depaoli, S., & Van de Schoot, R. (2017). Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist. *Psychological Methods, 22*(2), s. 240-261. doi:http://dx.doi.org/10.1037/met0000065

Goligher, E. C., Heath, A., & Harhay, M. O. (2024). Bayesian statistics for clinical research. *Lancet, 404*, s. 1067–1076.

Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy, 23*(4), s. 290-301. doi:https://doi.org/10.1016/j.bjpt.2018.12.006

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2 b.). Chapman and Hall/CRC. doi: https://doi.org/10.1201/9780429029608

Moyé, L. A. (2007). *Elementary Bayesian Biostatistics.* Houston: Chapman & Hall/CRC.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt international, 107*(44), s. 776-782. doi:https://doi.org/10.3238/arztebl.2010.0776

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia and

analgesia. *126*(5), s. 1763–1768. doi:https://doi.org/10.1213/ANE.0000000000002864

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. G. (2014). van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B.A gentle introduction to bayesian analysis: applications to developmental research. *van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to bayesian analysis: Child development, 85*(3), s. 842-860. doi:van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. G. (2014). A gentle introduction to bayesian anhttps://doi.org/10.1111/cdev.1

Yang, H., & Novick, S. (2019). *Bayesian Analysis with R for Drug Development: Concepts, Algorithms, and Case Studies* (1 b.). Chapman and Hall/CRC. doi:https://doi.org/10.1201/9781315100388

BÖLÜM V

Trafik Kazalarının Analizi: Poisson Regresyon ve Sağlam Poisson Regresyon Yöntemleri ile Risk Faktörlerinin Belirlenmesi

Barış ERGÜL[1]
Arzu ALTIN YAVUZ[2]

Giriş

Trafik kazaları, günümüz dünyasının önemli bir sağlık ve güvenlik sorunu olarak öne çıkmaktadır. Dünya Sağlık Örgütü (WHO), her yıl yaklaşık 1.35 milyon insanın trafik kazalarında hayatını kaybettiğini belirtmektedir. Bu kazaların dünya genelinde 20-29 yaş arası bireyler için en yaygın ölüm nedeni olduğunu vurgulamaktadır (WHO, 2021). Kazaların temel nedenleri arasında dikkatsizlik, aşırı hız, alkol tüketimi ve yeterli olmayan yol alt yapısı

[1] Dr. Öğr. Üy., Eskişehir Osmangazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir/TÜRKİYE, Orcid: 0000-0002-1811-5143, bergul@ogu.edu.tr
[2] Prof. Dr., Eskişehir Osmangazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir/TÜRKİYE, Orcid: 0000-0002-3277-740X, aaltin@ogu.edu.tr

gibi faktörler yer almaktadır. Özellikle şehir içi trafiğinde artan araç sayısı, sürücülerin dikkatini dağıtan unsurların çoğalmasına ve bu sebeple trafik kazalarına yol açmasına neden olmaktadır. Sürücülerin eğitimine ve trafik kurallarına uyum sağlamalarına yönelik çalışmalar, kazaların önlenmesinde kritik bir rol oynamaktadır. Araştırmalar, sürücülerin trafik güvenliği bilincinin artırılması için düzenlenen eğitim programlarının etkili olduğunu göstermektedir (Sharma & ark., 2023). Bu eğitimler, sürücülerin riskli davranışlarını azaltmalarına ve daha dikkatli bir şekilde araç kullanmalarına yardımcı olmaktadır.

Kavşaklar, trafik kazalarının en sık yaşandığı alanlar arasındadır. Kavşaklarda meydana gelen kazaların çoğu, sürücülerin birbirlerini görememesi veya dikkatsizlik nedeniyle yanlış kararlar almalarından kaynaklanmaktadır (Lee & Abdel-Aty, 2005). Özellikle ışıkların yetersiz olduğu veya işaretlerin belirsiz olduğu kavşaklarda, sürücülerin karar verme süreçleri olumsuz etkilenmektedir. Ayrıca, yayaların da yoğun olarak bulunduğu bu bölgelerde, sürücülerin dikkatini dağıtan faktörler kazaların ciddiyetini artırabilmektedir. Kavşakların tasarımı ve düzenlemesi, trafik güvenliği açısından hayati bir öneme sahiptir. Bu durum için önerilen çözümler arasında, döner kavşaklar, trafik ışıkları ve yayalar için güvenli geçiş yollarının oluşturulması yer almaktadır. Bu tür düzenlemeler, kazaların azaltılmasında etkili olmaktadır.

Trafik kazalarının sonuçları yalnızca fiziksel yaralanmalarla sınırlı kalmamaktadır. Kazaya karışan bireyler, travma sonrası stres bozukluğu gibi psikolojik sorunlarla da karşılaşabilmektedir. Ehlers & Clark (2000), trafik kazalarının psikolojik etkilerinin, kazadan sonra bireylerin yaşam kalitesini olumsuz yönde etkilediğini ortaya

koymaktadır. Bu durum, hem bireylerin sosyal ilişkilerini hem de iş hayatlarını etkileyebilir. Ayrıca, kazaların yarattığı maddi kayıplar, sağlık hizmetleri üzerinde baskı oluşturarak toplumun genel refahını tehdit etmektedir. Araştırmalar, trafik kazalarının sadece fiziksel yaralanmalar değil, aynı zamanda psikolojik travmalar da yarattığını göstermektedir. Bu nedenle, trafik kazalarının sonuçlarının çok boyutlu olarak ele alınması ve hem fiziksel hem de psikolojik destek hizmetlerinin sağlanması gerekmektedir. Kazaların ardından, bireylerin rehabilitasyon süreçlerinde psikolojik destek sunulması, onların topluma yeniden entegre olmalarını kolaylaştıracaktır.

Kazaların önlenmesi için toplumsal bir bilinç oluşturulması gerekmektedir. Eğitim kampanyaları, sürücülerin trafik kurallarına uyumunu artırmak için önemli bir rol oynamaktadır (Pajkovic & Grdinic, 2013). Bu kampanyalar, özellikle genç sürücüler arasında trafik güvenliği bilincini artırmayı hedeflemektedir. Genç sürücüler, deneyimsizlikleri nedeniyle daha fazla risk altındadır ve bu nedenle özel eğitim programlarına ihtiyaç duyulmaktadır. Devletin yol güvenliği standartlarını yükseltmesi ve altyapı yatırımlarını artırması da kazaların azaltılması için elzemdir. Araştırmalar, etkili yol tasarımı ve düzenlemelerinin kaza oranlarını önemli ölçüde azaltabileceğini göstermektedir (Elvik, 2009). Altyapı iyileştirmeleri, yol işaretlerinin belirginliğini artırmak, güvenli geçiş noktaları oluşturmak ve kavşak düzenlemelerini optimize etmek gibi unsurları içermektedir. Bu tür önlemler, sürücülerin güvenliğini artırmakta ve kazaların sıklığını azaltmaktadır.

Trafik kazalarının azaltılması için sürekli bir çaba içinde olunması ve toplumun tüm kesimlerinin bu mücadeleye dahil edilmesi gerekmektedir. Bireysel sorumlulukların yanı sıra, devlet

ve yerel yönetimlerin de trafik güvenliği konusunda aktif rol alması büyük önem taşımaktadır. Sadece bireylerin değil, tüm toplumun trafik güvenliği konusunda bilinçlenmesi, kazaların önlenmesinde etkili bir strateji olacaktır. Bu nedenle, trafik kazalarının azaltılması için multidisipliner bir yaklaşım benimsenmeli ve tüm paydaşlar iş birliği içinde çalışmalıdır. Ayrıca, trafik güvenliği konusundaki araştırmaların desteklenmesi ve yeni teknolojilerin uygulanması, gelecekte daha güvenli bir trafik ortamı oluşturulmasına katkı sağlayacaktır.

Trafik kazaları için literatür incelendiğinde, çok fazla sayıda çalışma olduğu görülmektedir. Son yıllarda poisson regresyon ile kaza analizi yapılan çalışmalar oldukça dikkat çekicidir. Bunun nedeni daha önceki regresyon çalışmalarının genellikle kaza sınıflandırmaya yönelik olması, poisson regresyonun ise ölümlü/yaralanmalı kaza sayısını tahmin etmede kullanılabilmesidir. Poisson regresyon, modelin kesikli dağılım içermesi ve tahmin edilen değerin negatif olmayan tam sayı olması ile ilgili olarak diğer regresyon analizlerine göre avantajlı bir yapıya sahiptir (Glenberg, 1996). Trafik kaza sayıları ile ilgili yapılan istatistiksel analizlerde, tüm araştırmacılar normalde Poisson regresyonunu analiz için bir başlangıç noktası olarak uygulamıştırlar (Lord & Mannering, 2010; Sonia, 2012). Başlangıç aşamasında doğru sonuçlar elde edilmesi çalışmanın devamındaki doğruluğu arttıracaktır. Ulaştırma sektöründe zaman serisi kullanan çalışmalar da bulunmaktadır (Neslihanoğlu & Aksoy, 2020).

Bu çalışmada, Ankara İli için 2017-2018 yılları arasında meydana gelen trafik kazası tutanakları yardımıyla trafik kazalarının analizinin yapılması amaçlanmıştır. Kullanılacak veriler Emniyet

Genel Müdürlüğü'nden elde edilmiştir. Çalışmada ilk olarak poisson regresyon modeli tanıtılmıştır. Daha sonra, poisson regresyon modelinin kullanılabilmesi için gerekli model varsayımlarının sağlanıp sağlanmadığı kontrol edilmiştir. Model varsayımlarının sağlanamadığı durumlarda alternatif olarak sağlam poisson regresyon modeli kullanılmıştır. Elde edilen tüm sonuçlar karşılaştırmalı olarak verilmiştir.

Materyal ve Metot

Poisson regresyon, olay sayılarının modellemesi için yaygın olarak kullanılan bir istatistiksel yöntemdir. Özellikle belirli bir zaman diliminde meydana gelen olayların sayısını tahmin etmek için uygun bir modelleme sağlamaktadır. Bu model, bağımlı değişkenin Poisson dağılımına sahip olduğunu varsayar ve bağımsız değişkenlerin bu olaylar üzerindeki etkisini değerlendirir. Poisson regresyon modeli Eşitlik (1)'de görüldüğü gibi ifade edilmektedir: (Cameron & Trivedi, 2013).

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + e \qquad (1)$$

Poisson regresyon modeli, sağlık bilimlerinden sosyal bilimlere, pazarlamadan çevre bilimlerine kadar geniş bir uygulama alanına sahiptir.

Poisson regresyonunun önemli bir varsayımı, bağımlı değişkenin ortalaması ile varyansının eşit olmasıdır. Ancak, gerçek dünyada genellikle bu varsayım ihlal edilmektedir. Aşırı saçılım durumu, varyansın ortalamadan daha büyük olduğu durumları ifade

eder ve bu, modelin tahmin gücünü olumsuz olarak etkilemektedir. Bu tür durumlarda, Negative Binomial regresyon, Quasi-Poisson regresyon veya Sağlam Poisson Regresyon gibi alternatif modeller kullanılabilir (Hilbe, 2011). Bu alternatif modeller, aşırı saçılımın etkilerini düzeltmek ve daha güvenilir tahminler yapmak için tasarlanmıştır. Doğru bir şekilde uygulandığında, Poisson regresyon, olayların sayısını etkileyen faktörleri anlamak ve tahmin etmek için güçlü bir araçtır. Poisson regresyonunun uygulanmasında dikkat edilmesi gereken bir diğer önemli nokta, modelin doğruluğunu artırmak için bağımsız değişkenlerin dikkatli bir şekilde seçilmesidir. Bağımsız değişkenler, olayların sayısını etkileyen faktörleri temsil eder ve bu değişkenlerin doğru bir şekilde belirlenmesi, modelin geçerliliğini artırır. Eksik veriler, aykırı değerler ve yanlış ölçümler, modelin güvenilirliğini azaltabilir. Bu nedenle, veri ön işleme adımları, modelin başarısı için kritik öneme sahiptir (Long & Freese, 2006).

Sağlam Poisson regresyon, klasik poisson regresyonunun varsayımlarının ihlal edildiği durumlarla başa çıkmak için geliştirilmiş bir istatistiksel yöntemdir. Klasik Poisson regresyon, bağımlı değişkenin Poisson dağılımına sahip olduğunu varsayar ve bu dağılımın ortalaması ile varyansının eşit olduğunu kabul etmektedir. Matematiksel olarak, bu durum şu şekilde ifade edilebilir:

$$Y_i \sim Poisson(\lambda_i) \tag{2}$$

Burada; Y_i : olay sayısını, λ_i : olayların beklenen sayısını belirtmektedir.

Poisson regresyonda, bağımlı değişkenin varyansı ortalamaya eşittir. Ancak, gerçek verilerde bu varsayım çoğu zaman geçerli olmaz ve aşırı saçılım gözlenir. Sağlam poisson modelleri, bu durumu ele almak için klasik poisson regresyon modeli kullanılarak tahmin edilir. Bu süreçte, varyansın fazla olmasına sebep olan gözlemler için penalizasyon veya ağırlık değeri uygulanır, alternatif olarak negatif binom dağılımı gibi alternatif modeller kullanılarak tahmin süreci iyileştirilir (Hilbe, 2014).

Aşırı saçılım durumu olduğunda, klasik poisson regresyon modelinin güvenirliği azalır ve sağlam poisson regresyon modeli bu durumu ele alacak şekilde modellenir. Varyansı aşırı saçılma parametresi olan ϕ ile ağırlıklandırır:

$$\mathrm{Var}(Y) = \phi\lambda \tag{3}$$

Burada; $\phi > 1$: aşırı saçılmanın şiddetini belirtmektedir ve genel olarak aykırı değerlerin varlığı durumunda ortaya çıkmaktadır.

Klasik Poisson regresyon modelinde, aykırı değerler tahminlerin değerlerini farklı hale getirmektedir ve modelin genel performansı düşmektedir. Örneğin, bir gözlem grubu beklenenden çok daha fazla olay sayısı içeriyorsa, bu durum tahmin edilen parametrelerin yanlış hesaplanmasına yol açmaktadır. Sağlam Poisson regresyon, bu tür sorunları ele almak için ağırlıklı en çok

olabilirlik tahmin yöntemlerini kullanır. Geleneksel Poisson regresyonu, aykırı değerlerden önemli ölçüde etkilenebilir. Sağlam Poisson regresyonu, aykırı gözlemleri modelin sonuçlarını fazla etkilemeyecek şekilde ele alır. Böylece daha güvenilir sonuçlar elde edilir. Poisson regresyonu için (4) no'lu eşitlikte verilen denklemlerin çözülmesi ile elde edilir (İdriss & Cheng, 2023):

$$\sum_{i=1}^{n} w_i \left[y_i log(\lambda_i) - \lambda_i - log(y_i!)\right] \tag{4}$$

Uygulamada ağırlık fonksiyonu olarak Huber ve Tukey ağırlık fonksiyonları yaygın bir şekilde kullanılmaktadır. Sağlam Poisson Regresyon modeli, aşırı saçılmanın yanı sıra aykırı değerlerin etkisini de azaltmaktadır. Sağlam Poisson Regresyon yöntemi, genellikle M-tahminci veya bir tür iteratif yeniden ağırlıklandırılmış en küçük kareler (IRLS) algoritmasını kullanmaktadır (Cantoni & Ronchetti, 2001).

Sağlam Poisson Regresyon, özellikle aşırı saçılma ve aykırı değerlerin mevcut olduğu durumlarda Klasik Poisson regresyon modellerine göre önemli avantajlar sunar. Bu modeller, varyansı açıklayıcı değişkenlerin doğrusal etkisinin ötesine genişletebildiği için düzensiz olay verilerini modellemede sıkça tercih edilmektedir. Ayrıca, sağlam poisson regresyon, eksik veya yanlış tanımlanmış modellerin neden olduğu sistematik hataları azaltır ve sonuçların daha güvenilir olmasını sağlar (Cantoni & Ronchetti, 2001; Hilbe, 2014).

Bulgular

Bu çalışmada Ankara İli için 2017-2018 yılları arasında meydana gelen trafik kazaları ele alınmıştır. Veri seti, kazaya karışan araç cinsi, emniyet kemeri kullanma durumu, yol tipi, hava durumu, trafik levhası durumu, trafik lamba durumu, aydınlatma durumu ve trafik kazaları sayılarını içermektedir. İlgili veri seti çalışmaya hazır hale getirilmek için ön işlemeden geçirilmiş ve nihai olarak 3503 trafik kazası çalışmaya dahil edilmiştir.

Burada, bağımlı değişken olarak trafik kaza sayıları ele alınmış ve kaza sayısının modellenmesi üzerinde durulmuştur. Kazaya karışan araç cinsi, emniyet kemeri kullanma durumu, yol tipi, hava durumu, trafik levhası durumu, trafik lamba durumu, aydınlatma durumu değişkenleri ise, trafik kaza sayısını etkileyen bağımsız değişkenler olarak belirlenmiştir.

$\mathrm{Y} = \text{Trafik kaza sayısı}$

$X_1 = Araç\ Cinsi$

$X_2 = Emniyet\ Kemeri\ Durumu$

$X_3 = Yol\ Tipi$

$X_4 = Hava\ Durumu$

$X_5 = Trafik\ Levhası\ \ Durumu$

$X_6 = Trafik\ Lamba\ Durumu$

$X_7 = Aydınlatma\ Durum$

Bağımlı ve bağımsız değişkenler belirlendikten sonra veri analizi aşamasına geçilmiştir. İlk aşamada, bağımlı değişken değerlerinden oluşan veri setinin ortalama ve varyansı hesaplanmıştır ve Tablo 1'de gösterilmiştir.

Tablo 1: Trafik Kazası veri seti için Ortalama ve Varyans

	Ortalama	**Varyans**
Trafik Kaza Sayısı	1,673	1,849

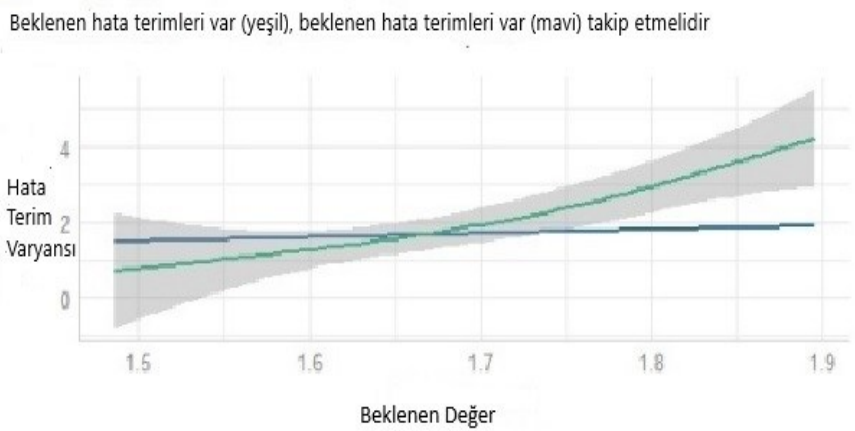
Veri analizi için R programından yararlanılmıştır. R programında glm paketi kullanılarak klasik poisson regresyon analizi uygulanmış ve elde edilen sonuçlar Tablo 2'de özetlenmiştir. Tablo 2'den görülebileceği gibi sabit terim ve aydınlatma durumu değişkenlerinin katsayı değeri için elde edilen p değeri 0,05'ten küçük olduğundan katsayıların anlamlı olduğuna karar verilmiştir. Aydınlatma durumu için parametre tahmin değeri, 0,0983 olarak bulunmuştur. Bunun anlamı, aydınlatma olmaması, beklenen trafik kazası (logaritma) sayısını 0,0983 birim arttırmaktadır. Klasik Poisson Regresyon modeli ile kurulan model için bulunan AIC değeri, 10619 olarak bulunmuştur.

Tablo 2: Trafik Kazası veri seti için Klasik Poisson Regresyon Analizi Sonuçları

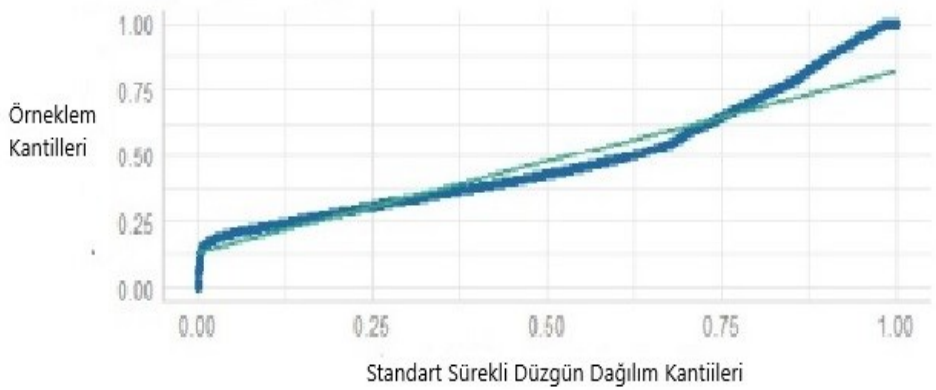
Bağımsız değişkenler	**Katsayı**	**Standart Hata**	**z**	**p**	**AIC**
Sabit Terim	0,3537	0,1397	2,532	**0,011**	10619
X_1	-0,0044	0,0078	-0,569	0,569	
X_2	0,0110	0,0197	0,558	0,576	
X_3	-0,0255	0,0153	-1,675	0,094	
X_4	-0,0102	0,0209	-0,487	0,626	
X_5	-0,0003	0,0307	-0,011	0,991	
X_6	0,0375	0,0344	1,088	0,277	
X_7	0,0983	0,0321	3,064	**0,002**	

Sonraki aşamada, Klasik Poisson regresyon modeli için varsayımlar araştırılmıştır. Varsayımların araştırılması için R programı içinde yer alan DHARMa paketi kullanılmıştır.

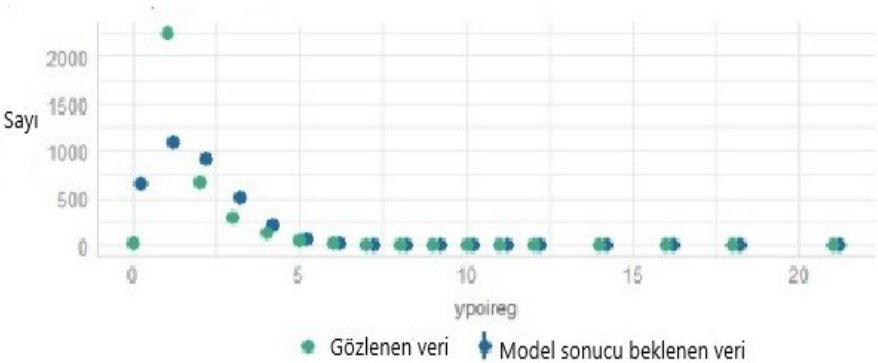
Öncelikle aşırı saçılma durumu teşhis edilmiştir. Bu iş için bağımlı değişken veri setinin ortalama ve varyansı hesaplanmıştır. İlgili sonuç Tablo 1'de görüldüğü gibidir. Tablo 1'e göre, trafik kazaları sayısı varyansının ortalamadan büyük olduğu görülmektedir. Bu durum ise aşırı saçılım durumu olduğunun göstergesidir. Diğer yandan, DHARMa paketinden alınan grafik, Şekil 1'de görülmektedir. Şekil 1'e göre, gözlenen hata terimlerinin varyansı, beklenen hata terimi varyansını takip etmedği gözlenmiştir. Bu durumda, aşırı saçılım durumunun varlığından söz edilebilir. Ek olarak, hata terimlerinin tek düze olup olmadığına ait grafik Şekil 2'de gösterilmiştir. Şekil 2'ye göre, hata terimlerinin düzgün dağılmadığı tespit edilmiştir. Çünkü, veri noktaları 45^0 'lik çizgiden saptıkları görülmektedir. Klasik Poisson regresyon modeli uygulanmış veri noktalarının, gözlenen ve beklenen veri sayıları Şekil 3'te belirlendiği gibidir.



Şekil 1: Poisson Regresyon İçin Saçılım Grafiği



Şekil 2: Poisson Regresyon İçin Hata Terimleri Sürekli Düzgün Dağılım Grafiği



Şekil 3: Poisson Regresyon İçin Gözlenen ve Beklenen Veri Noktaları Sayısına Ait Grafik

Trafik kazaları veri setinin aşırı saçılım gösterdiği göz önüne alınarak sağlam poisson regresyon analizi uygulanmış ve ilgili sonuçlar Tablo 3'te verilmiştir. Tablo 3'den görülebileceği gibi sabit terim, emniyet kemeri durumu, yol tipi, trafik lamba durumu ve aydınlatma durumu değişkenlerinin katsayı değeri için elde edilen p değeri 0,05'ten küçük olduğundan ilgili katsayıların anlamlı olduğuna karar verilmiştir. Emniyet kemeri durumu için parametre tahmin değeri,-0,04646 olarak bulunmuştur. Bunun anlamı, emniyet kemeri takıyor olmak, beklenen trafik kaza (logaritma) sayısını 0,04646 birim azaltmaktadır. Yol tipi için parametre tahmin değeri, -0,00091'dir. Bölünmüş yolda seyrediyor olmak, beklenen trafik kazası (logaritması) sayısını 0,00091 birim azaltmaktadır. Trafik lambası durumu için bulunan parametre tahmini değeri, -0,00306'dır. Yol üzerinde trafik lambası olması, beklenen trafik kazası (logaritma) sayısı 0,00306 birim azaltmaktadır. Aydınlatma durumu değişkeni için tahmin edilen paramtre tahmin değeri,

0,00175'tir. Aydınlatma olmaması, beklenen trafik kazası (logaritma) sayısını 0,00175 birim arttırmaktadır. Sağlam Poisson Regresyon modeli ile kurulan model için bulunan AIC değeri, 10572 olarak bulunmuştur.

Tablo 3: Trafik Kazası veri seti için Sağlam Poisson Regresyon Analizi Sonuçları

Bağımsız değişkenler	**Katsayı**	**Standart Hata**	**z**	**p**	**AIC**
Sabit Terim	1,23700	0,00006	19330,625	**0,001**	10572
X_1	0,00011	0,00013	0,842	0,399	
X_2	-0,04646	0,00032	-145,588	**0,001**	
X_3	-0,00091	0,00025	-3,689	**0,001**	
X_4	-0,00032	0,00035	-0,921	0,357	
X_5	0,00032	0,00051	0,639	0,523	
X_6	-0,00306	0,00056	-5,467	**0,001**	
X_7	0,00175	0,00055	3,186	**0,001**	

Sonuç ve Öneriler

Trafik kazaları sayılarını etkileyen risk faktörlerinin belirlenmesinin amaçlandığı bu çalışmada, poisson regresyon analizi kullanılarak ilgili veri setinden çıkarsamalar yapılmıştır.

Bu çalışma için Ankara İli için 2017-2018 yılları arasında meydana gelen trafik kazaları göz önünde bulundurulmuştur. Veri seti; kazaya karışan araç cinsi, emniyet kemeri takma durumu, yol tipi, hava durumu, trafik levhası durumu, trafik lamba durumu, aydınlatma durumu ve trafik kaza sayıları gibi değişkenleri içermektedir. İlgili veri seti, çeşitli düzenlemeler yapıldıktan sonra, 3503 trafik kazası tutanaklarından alınan verilerden oluşmaktadır.

İlgilenilen veri setine öncelikle, klasik poisson regresyon analizi uygulanmıştır. Aydınlatma durumu değişkeni için bulunan katsayı değerinin anlamlı olduğu belirlenmiştir. Klasik poisson regresyon analizi için AIC değeri, 10619 olarak bulunmuştur.

Sonrasında, klasik poisson regresyon analizi için gerekli olan varsayımların sağlanıp sağlanmadığı kontrol edilmiştir. Analiz sonucunda, klasik poisson regresyon analizi varsayımlarından aşırı saçılma problemi ile karşılaşıldığı tespit edilmiştir.

Son aşamada, varsayımların sağlanmaması durumunda kullanılan sağlam poisson regresyon analizi uygulanmıştır. Sağlam poisson regresyon analizi sonucunda, emniyet kemeri durumu, yol tipi, trafik lamba durumu ve aydınlatma durumu değişkenlerinin katsayılarının anlamlı olduğuna karar verilmiştir. Gaber & ark., (2023), hava ve yol koşullarının trafik kazalarını etkilediği sonucuna

varmışlardır. Bu çalışmada, benzer sonuçlar ortaya konulmuştur. Sağlam poisson regresyon analizi için AIC değeri, 10572 olduğu tespit edilmiştir. AIC değerleri karşılaştırıldığında, sağlam poisson regresyon analizi için bulunan AIC değerinin klasik poisson regresyon analizine göre daha küçük olduğu görülmektedir. Bu durumda, sağlam poisson regresyon analizinin klasik poisson regresyon analizine göre daha güvenilir sonuçlar verdiği ortaya konulmuştur.

Bu çalışmada, klasik ve sağlam poisson regresyon analizleri için örnek bir uygulama yapılmıştır. Bundan sonra yapılacak trafik kazaları ile ilgili çalışmalarda, başlangıç modeli olan poisson regresyon analizi için temel varsayımların kontrol edilmesi ve buna bağlı olarak parametre tahminlerinin yapılması, elde edilecek sonuçların ve çalışmanın kalan kısmının geçerlilik ve güvenilirliğini arttıracaktır.

Kaynakça

World Health Organization (WHO). (2021). *Global Status Report on Road Safety 2021.*

Sharma, V. K., Upadhyay, R. K., & Kumar, S. (2023). A Systematic Review On Road Traffic Accident: Causes and Control Measures. *IJCRT, 11* (4), 291-300.

Lee, J., & Abdel-Aty, M. (2005). Comprehensive Analysis of Vehicle-Pedestrian Crashes At Intersections In Florida. *Accident Analysis & Prevention, 37* (4)*,* 775-786.

Ehlers, A., & Clark, D. M. (2000). A Cognitive Model of Post-Traumatic Stress Disorder. *Behaviour Research and Therapy, 38* (4), 319-345.

Pajkovic, V., & Grdinic, M.. (2013). An Analysis Of Young Driver Accidents In Road Traffic Using In-Depth Crash Investigation Data. *International Conference on Accomplishmants in Electrical and Mechanical Engineering and Information Tehnology*, Banja Luka, May, 2013, 1-6.

Elvik, R. (2009). *Handbook of Road Safety Measures.* Emerald Group Publishing.

Glenberg, A. (1996). *Learning from Data: An Introduction to Statistical Reasoning.* 2nd Edition, Lawrence Erlbaum Associates, Mahwah.

Lord, D., & Mannering, F. (2010). The statistical Analysis of Crash Frequency Data: A Review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention ,44*, 292-303.

Sonia, R. (2012). Development of an Accident Prediction Model for Intersections of Dhaka City, Bangladesh. *International Journal of Computer Applications, 47* (16), 10-16.

Neslihanoğlu, S. & Aksoy, T. (2020). Bist'teki UlaştırmavSektörü Firmalarının Verilerinin Modellenmesi ve Gelecek Tahmini İçin Doğrusal Piyasa Modeli Yeterli Mi?, *Yönetim ve Ekonomi Araştırmaları Dergisi, 18* (4), 54-72.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data.* Cambridge University Press.

Hilbe, J. M. (2011). *Negative Binomial Regression.* Cambridge University Press.

Long, J. S., & Freese, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata*. Stata Press.

İdriss, I. A., & Cheng, W. (2023). Robust Estimators for Poisson Regression. *Open Journal of Statistics, 13* (1), 112-118.

Cantoni, E., & Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association, 96,* 1022-1030.

Hilbe, J. M. (2014). *Modeling Count Data.* Cambridge University Press.

Gaber, M., Diab, A., Othman, A., & Wahaballa, M. (2023). Analysis and Modeling of Rural Roads Traffic Safety Data. *Sohag Engineering Journal (SEJ), 3* (1), 57-67.

BÖLÜM VI

Yaralanmalı Kaza Sayılarının İzlenmesi İçin Regresyon Kontrol Grafikleri

Barış ERGÜL[1]
Arzu ALTIN YAVUZ[2]

Giriş

Kontrol grafikleri, süreç değişkenliklerini ölçmek ve süreç istikrarını değerlendirmek için kullanılan istatistiksel araçlardan bir tanesidir. Walter A. Shewhart (1931) tarafından geliştirilen bu grafikler, süreç performansının zaman içindeki değişimini izlemek için kullanılan temel yöntem olarak ön plana çıkmaktadır (Montgomery, 2013). Kontrol grafiklerinin temel amacı, süreçteki rassal değişkenlik ile özel nedenlerden kaynaklanan değişkenliği ayırt etmektir.

[1] Dr. Öğr. Üy., Eskişehir Osmangazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir/TÜRKİYE, Orcid: 0000-0002-1811-5143, bergul@ogu.edu.tr
[2] Prof. Dr., Eskişehir Osmangazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir/TÜRKİYE, Orcid: 0000-0002-3277-740X, aaltin@ogu.edu.tr

Kontrol grafiklerinin yaygın kullanılan türleri arasında $\bar{X}$, R, p, np, c ve u grafikleri yer almaktadır. $\bar{X}$ ve R grafikleri, bir sürecin ortalaması ve değişimini zaman içinde izlemek amacıyla kullanılan istatistiksel yöntemlerdir (Montgomery, 2013). p ve np grafikleri, belirli bir süreçte kusurlu oranlarını veya kusurlu birim sayısını değerlendirerek kalite performansını ölçmeyi hedefler (Evans ve Lindsay, 2014). Öte yandan, c ve u grafikleri, birim başına veya belirli bir alanda ortaya çıkan kusur sayılarının kontrol edilmesini sağlamak için kullanılır ve özellikle kusurların nadiren görüldüğü süreçlerde etkilidir (Woodall & Montgomery, 1999).

Regresyon kontrol grafikleri (RKG) ise, geleneksel kontrol grafiklerinden farklı olarak bağımlı ve bağımsız değişkenler arasındaki ilişkiyi ön planda tutarak, süreçteki karmaşık ilişkileri analiz etmeye olanak sağlamaktadır. Regresyon kontrol kartları, süreç değişkenleri arasındaki ilişkilerin zamanla istikrarlı olup olmadığını inceleyerek süreç kontrolünün sağlanmasına yardımcı olur (Montgomery, 2013). Bu kontrol kartları, özellikle bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkiyi değerlendiren regresyon analizine dayalı artıkları kullanır (Montgomery ve Runger, 2018). Regresyon kontrol kartları, proses verilerinde öngörülemeyen değişimlerin tespit edilmesine ve bu değişimlerin nedenlerinin analiz edilmesine olanak tanır (Woodall, 2000). Standart Shewhart kontrol kartlarının genişletilmiş bir uygulaması olan regresyon kontrol kartları, süreçteki özel nedenli değişimlerin erken aşamada belirlenmesinde oldukça etkilidir (Alwan & Roberts, 1988). Süreç kontrolünde yalnızca ortalama ve varyansa odaklanmak yerine, regresyon kontrol kartları değişkenler arasındaki doğrusal ilişkiyi de dikkate alarak daha derinlemesine

analiz sağlar (Montgomery, 2013). Regresyon kontrol kartları, özellikle kalite kontrol süreçlerinde çoklu değişkenlerin kontrol altında tutulmasını ve süreç performansının istikrarlı hale getirilmesini destekler (Pan & Jarrett, 2004). Kalite kontrol çalışmalarında çoğu zaman kullanılan zaman serisi şeklindeki verilere otoregresif model uygulama çalışmaları da mevcuttur (Neslihanoğlu & Paker, 2021).

Kontrol grafiklerindeki temel bileşenler, merkezi çizgi (ortalama), üst kontrol limiti (UCL) ve alt kontrol limiti (LCL) olmak üzere üç ana unsurdan oluşur. Eğer gözlem değerleri, kontrol limitleri içinde kalıyorsa süreç istikrarlı kabul edilir; aksi takdirde, özel nedenlere bağlı bir sorun olduğu düşünülür (Ryan, 2011). RKG, klasik kontrol grafiklerine benzer şekilde merkezi çizgi ve kontrol limitlerinden oluşmaktadır. Ancak bu grafiklerde, merkezi çizgi bir regresyon modelinden elde edilen tahmin değerlerini temsil etmektedir. Bağımsız değişkenlerin sürece etkileri, regresyon modeliyle belirlenmektedir ve gözlem değerleri, tahmin edilen değerlerle karşılaştırılmaktadır.

Kontrol grafiklerinin sağladığı avantajlar arasında, temel istatistik bilgisiyle kolayca uygulanabilir olması, süreçte meydana gelen özel nedenlere bağlı sapmaların etkin bir şekilde tespit edilmesi (Ryan, 2011) ve farklı süreç türleri için uygun kontrol grafiklerinin kullanılabilir olması yer almaktadır. Ancak geleneksel kontrol grafiklerinin önemli sınırlılıkları bulunmaktadır. Bu grafikler genellikle yalnızca tek bir değişkeni izlemek amacıyla tasarlanmış olup, değişkenler arasındaki ilişkileri dikkate almaz (Woodall, 2000). Bu durum, karmaşık ve çok değişkenli süreçlerin izlenmesinde yetersizliğe yol açmaktadır. Bu bağlamda regresyon

kontrol kartlarına olan ihtiyaç ön plana çıkmaktadır. Regresyon kontrol kartları, süreç değişkenleri arasındaki doğrusal veya doğrusal olmayan ilişkileri göz önünde bulundurarak artıkların (residuals) analizine dayanır ve çoklu değişken içeren süreçlerde özel neden kaynaklı değişimleri tespit etmede daha etkili bir yöntem sunar (Alwan & Roberts, 1988; Montgomery, 2013). Özellikle süreçteki bağımlı ve bağımsız değişkenler arasındaki ilişkinin kontrol edilmesi gerektiği durumlarda, regresyon kontrol kartları geleneksel kontrol grafiklerinin sınırlılıklarını aşarak daha derinlemesine analiz imkanı sağlar.

Literatürde yapılan çalışmalar, bu yöntemlerin doğru bir şekilde kullanıldığında süreç iyileştirme ve kalite yönetiminde önemli katkılar sağladığını göstermektedir. Ancak, her iki yöntem de uygulama sırasında dikkatli bir şekilde seçilmeli ve doğru bir şekilde yorumlanmalıdır. Woodall (2000), kontrol grafiklerinin sınırlamalarını ele aldığı çalışmasında, RKG'nin bağımsız değişkenler arasındaki ilişkileri analiz etme konusundaki avantajlarını detaylandırmaktadır. Özellikle karmaşık süreçlerin kontrolünde RKG'nin önemine işaret etmiştir. Vargas (2003), çok değişkenli kontrol grafiklerini ele aldıkları çalışmalarında, RKG'nin bağımsız değişkenler arasındaki etkileşimleri dikkate alarak daha iyi sonuçlar verdiğini belirtirmektedir. Zou & Qiu (2009), zaman serisi verilerinde RKG'nin süreç izleme ve kontrol için nasıl etkili bir araç olduğunu gösterilmişlerdir ve dinamik süreçlerde RKG'nin avantajlarını vurgulamışlardır.

Bu çalışmada, RKG'nin özellikleri, avantajları, sınırlamaları ve kullanım alanları tanıtılarak, ulaştırma alanında bir uygulaması sunulacaktır.

Materyal ve Metot

Regresyon analizi, bir asırdan uzun süredir tarım, biyolojik bilimler, tıp, deney tasarımı, endüstriyel ve hizmet süreçleri de dahil çok geniş bir uygulama alanı bulmuştur. Doğrusal regresyon analizinde amaç, iki değişken arasındaki ilişkiyi belirlemek ve modellemektir. Regresyon doğrusu, $y = b_0 + b_1 x$ genel olarak tahmin edilir. Regresyon analizinde, iki değişken karşılaştırıldığında, ilişki ile nedensellik arasında bir ayrım yapılması gerekmektedir. İlişki söz konusu olduğunda, bir dağılım grafiği kullanılarak, iki değişkenin eşlenmiş değerlerinin aynı yönde hareket edip etmediği araştırılmaktadır. Bununla birlikte, nedensellik yalnızca görsel ya da matematiksel bir ilişki değildir ve çok daha fazla araştırma ve konu uzmanlığı gerektirmektedir (Mandel, 1969).

Regresyon kontrol grafiği için regresyon analizi, kontrol grafiği teorileri ve uygulamaları birleştirilerek, süreç analizi için etkili bir yöntem oluşturulmaktadır. Bağımsız değişkenlerdeki değişimin, incelenen bağımlı değişkenin özellikleri üzerindeki etkisini araştırmak kalitenin güvence altında olması açısından önem kazanmaktadır. Regresyon kontrol kartları, zaman serisi verilerinin ve bağımsız değişkenlerin doğrusal ilişkisini modelleyen bir istatistiksel yöntemdir. Temelde, bir regresyon modelinin hata terimleri izlenerek, modeldeki hataların istatistiksel olarak anlamlı sapmalarını tespit etmek amaçlanır (Stewhart, 1931). Bu kartlar, genellikle "tahmin değerleri" ve "gerçek değerler" arasındaki farkların izlenmesine dayanır. Hata terimi, gözlemler ile model tarafından tahmin edilen değerler arasındaki farktır. Hedef, bu farkların zamanla kontrol altında tutulmasını sağlamaktır.

Regresyon kontrol kartı, regresyon modeline dayalı olarak hesaplanan hata terimlerinin zamanla izlenmesini sağlar. Bu kart, geleneksel kontrol kartlarından farklı olarak, modelin hata terimlerinin ortalama değeri ve dağılımına dayanır. Hata terimlerinin belirli bir kontrol sınırları içerisinde olup olmadığı izlenir. Shewhart'ın (1931) kontrol grafiği teorisi kullanılarak kontrol sınırları hesaplanır. Tahmin standart hatası (σ_ε) belirlenir. Regresyon doğrusundan $\pm 3\sigma_\varepsilon$ mesafeye paralel kontrol sınırları yerleştirilir; bu sınırlar $\bar{e} \pm 3\sigma_\varepsilon$ olarak ifade edilir. Kontrol kartının sınırları şu şekilde hesaplanabilir:

- **Üst Kontrol Limiti (UCL)**: $\bar{e} + 3\sigma_\varepsilon$ (1)
- **Alt Kontrol Limiti (LCL)**: $\bar{e} - 3\sigma_\varepsilon$ (2)

Burada, $\bar{e}$ hata terimlerinin ortalamasıdır ve σ_ε hata terimlerinin standart sapmasıdır.

Bu sınırlar arasındaki alan, bağımsız değişkenden kaynaklanan kabul edilebilir değişim olarak değerlendirilmektedir. Kontrol sınırları içinde kalan noktalar, genel değişim olarak kabul edilmektedir. Kontrol sınırlarının üstünde veya altında yer alan değerler ise, özel nedenler olarak değerlendirilmektedir ve bu nedenle araştırılması gerekmektedir (Hayati, 2017). Eğer hata terimi, belirlenen üst ve alt kontrol limitlerini aşarsa, bu, modeldeki bir sapmanın veya süreçteki bir anormalliğin göstergesi olarak kabul edilir.

Regresyon kontrol kartının uygulanması için izlenecek genel adımlar şunlardır:

Veri Toplama: Süreçten veya gözlemlerden elde edilen veriler toplanır ve modelin bağımsız ve bağımlı değişkenleri belirlenir.

Regresyon Modeli Kurma: Verilere dayalı olarak bir regresyon modeli oluşturulur.

Tahmin ve Hata Hesaplama: Regresyon modeline dayalı olarak her bir gözlem için tahmin edilen değerler hesaplanır ve gerçek gözlemler ile tahminler arasındaki fark (hata) hesaplanır.

Kontrol Kartı Oluşturma: Hata terimleri izlenerek kontrol kartı çizilir. Üst ve alt kontrol limitleri hesaplanır.

İzleme ve Değerlendirme: Regresyon kontrol kartı zamanla izlenir ve hata terimleri kontrol sınırlarının dışında kaldığında müdahale edilir.

Genelleştirilmiş Doğrusal Model (GLM) tabanlı kontrol grafiği, regresyon tabanlı kontrol grafikleriyle aynıdır ve bağımlı değişken verileri sayma verileri olduğunda geçerlidir. Bağımlı değişken değerleri sayma verileri olduğunda ve dağılımı Poisson dağılımı olduğu zaman iyi sonuç vermektedir. Hata terimleri, poisson regresyon modelinden elde edilir, kontrol grafiği için merkez çizgi, UCL ve LCL değerleri hata terimleri kullanılarak hesaplanır (Mammadova & Özkale, 2018).

Uyum elde etmek için gerçek ve tahmin edilmiş değerleri göz öüne alan Pearson Ki-Kare istatistiği kullanılır. Poisson regresyon modelinden elde edilen Pearson hata terimleri aşağıdaki gibi hesaplanır: (Russo v.d., 2008).

$$e_i = \frac{x_i - \lambda_i}{\sqrt{\lambda_i}} \quad (3)$$

Bulgular

Bu çalışmada Türkiye'de 2002-2023 yılları arasında meydana gelen trafik kazaları incelenmiştir. Veriler, İçişleri Bakanlığı Emniyet Genel Müdürlüğü Trafik Başkanlığından elde edilmiştir. Çalışmada kaza sayısı değişkeni bağımsız değişken, trafik kazasında yaralanma sayıları ise bağımlı değişken olarak incelenmiştir. Tablo 1 ve Tablo 2'de, 2002-2023 dönemi için trafik kazaları ve yaralanma kayıtlarının özetini sunmaktadır.

Tablo 1: Trafik Kaza Sayısı için Özetleyici İstatistikler

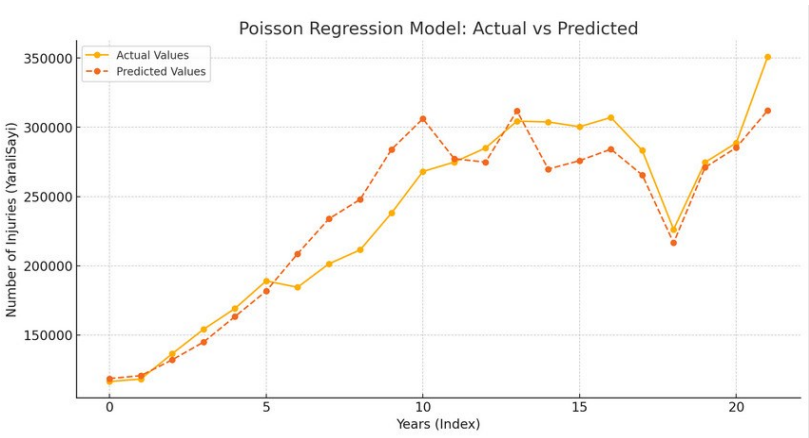
Minimum	**1. Kantil**	**Medyan**	**Ortalama**	**3. Kantil**	**Maksimum**
439777	801359,5	1175317,5	1021036	1229037	1314136

Tablo 2: Trafik Yaralanma Sayısı için Özetleyici İstatistikler

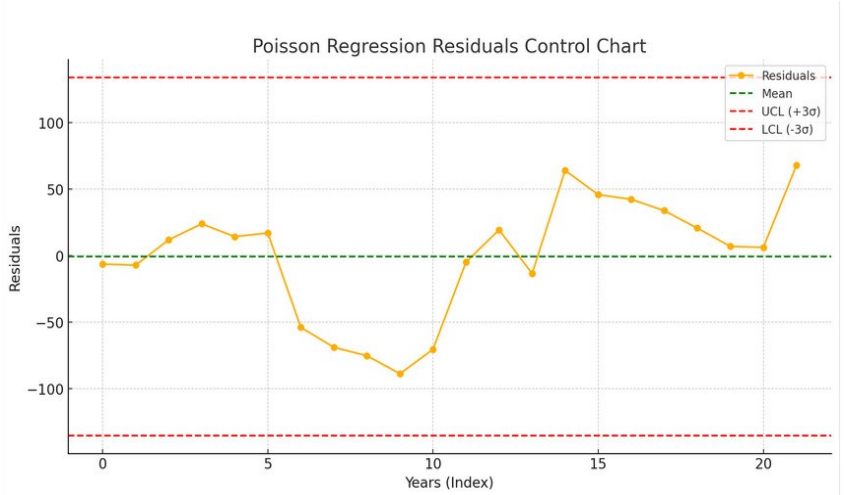
Minimum	**1. Kantil**	**Medyan**	**Ortalama**	**3. Kantil**	**Maksimum**
116412	180621	253076,5	235728	291617,75	350855

Poisson regresyon modeli, kaza sayıları ile yaralı sayıları arasındaki ilişkiyi incelemek için kullanılmıştır. İlgili veri setine ilk olarak poisson regresyon analizi uygulanmış ve gerçek değerler ile

tahmini değerlerin yer aldığı grafik Şekil 1'de gösterildiği gibi bulunmuştur. Şekil 1'e göre, modelin tahmini değerleri ile gerçek gerçek değerleri karşılaştırılmış ve genel olarak uyum sağladığı görülmüştür. Ancak bazı yıllar için, model tahmin değerlerinin gerçek değerlerden sapma gösterdiği fark edilmektedir. Bu sapmalar, model eksikliklerini veya veri setindeki anormallikleri işaret eder. Ek olarak, bağımsız değişken olan kaza sayısı değişkeni anlamlı olarak bulunmuştur ($p<0,05$). Ancak modelin log-bağlantı fonksiyonu, tahmin edilen değerlerin pozitif kalmasını sağlamıştır ve en çok olabilirlik yöntemi ile parametre tahminleri elde edildiğinden, modele daha başka bağımsız değişkenlerin eklenmesi ya da model yapısının geliştirilmesi gereği göz önünde bulundurulmalıdır.



Şekil 1: Trafik Kazaları Poisson Regresyon Model ile Gözlenen ve Beklenen Değerler Grafiği



Şekil 2: Trafik Kazaları Poisson Regresyon Model ile Belirlenen Yaralı Sayıları Hata Terimleri İçin Kontrol Grafiği

Poisson regresyon modeli kullanılarak oluşturulan merkez çizgi, UCL ve LCL değerleri Tablo 3'te özetlenmiştir.

Tablo 3: Trafik Yaralanma Sayısı için Kontrol Limitleri İstatistikleri

LCL	**Merkez Çizgi**	**UCL**	**Standart Sapma**
-135,220	-0,590	134,040	44,880

Kontrol sınırları, merkezi çizgi doğrusuna paralel olarak belirlenmiştir. Dağılım grafiğinde üç çizgi çizilmiştir: merkezi çizgi (en iyi uyum doğrusu), üst kontrol sınırı (UCL) ve alt kontrol sınırı (LCL). Bu çizgiler, regresyon doğrusunun etrafındaki değişimi izlemektedir ve sürecin normal veya özel nedenlerle sapmalarını

belirlemek için kullanılmaktadır. UCL ve LCL, veri noktalarının kabul edilebilir değişim sınırlarını tanımlarken, uyarı sınırları, olası sorunların erken bir uyarısı olarak işlev görmektedir. Yaralı sayısı için elde edilen poisson regresyon modeli ile oluşturulan hata terimleri kullanılarak belirlenmiş RKG, Şekil 2'de gösterildiği gibidir. Şekil 2'de, tüm değerlerin kontrol altında olduğu ve Türkiye için trafik kazası yaralanmaları için özel bir nedenin bulunmadığı belirtilmektedir. Bu durumda, regresyon kontrol grafiği bir performans göstergesi olarak kullanılmaktadır ve üç sigma sınırları, kabul edilebilir değişim aralığını göstermektedir.

Sonuç ve Öneriler

Bu çalışmada, trafik kazası sırasında meydana gelen yaralanmalar ile kaza sayısı arasındaki ilişki incelenmiştir.

RKG'ni oluşturmak için, önce regresyon denklemi poison regresyon modeli ile bulunmuş ve tahmin standart hatası hesaplanmıştır. Kontrol sınırları ise, üç sigma ($\pm 3\sigma$) ve paralel çizgiler (kırmızı kesikli) olarak çizilmiştir. Üç sigma seviyesinde, tüm değerlerin üst kontrol sınırı (UCL) ve alt kontrol sınırının (LCL) içinde kaldığı görülmüştür. Sonuç olarak, Türkiye için trafik kazalarındaki yaralı sayısı kontrol altındadır denilebilir.

İki değişken arasındaki ilişkiyi incelerken, bir değişkenin kontrol grafiği kontrol altında olabilirken, regresyon kontrol grafiği korelasyon ve diğer değişkenin etkisi nedeniyle kontrol dışında olabilir. Ayrıca, bir değişkenin kontrol grafiği sınırların dışında değerler gösterirken, regresyon kontrol grafiği, diğer değişkenin katkısından dolayı varyasyonun haklı bir şekilde kontrol altında olduğunu gösterebilir (Hayati, 2017).

Genel olarak, bir çok yazar, diğer faktörlerini dikkate almadan özel nedenleri incelemektedir (Hayati, 2017). Trafik kazası yaralanmalarını, diğer faktörleri dikkate almadan izlemek, böyle bir incelemeyi yetersiz kılmaktadır. Kaza sayısındaki değişim, trafik kazası yaralanmaları üzerinde önemli bir etkiye sahip bir değişimin kaynağıdır. Bu nedenle, RKG, kaza sayısının analizde yer aldığı durumlar için uygun bir yöntem olarak karşımıza çıkmaktadır.

RKG, diğer değişkenlerden gelen kabul edilebilir değişimi hesaba katmanın etkili bir yolu olmakla birlikte, geleneksel kontrol grafiğine göre birçok avantaj sunmaktadır. Bu çalışmada, RKG, trafik kazası yaralanmaları için sunulmuştur. Bu durum, trafik kazası sayısındaki değişimin etkileri dikkate alınarak yapılmıştır. Üç sigma seviyesinde, RKG tüm değerlerin kontrol sınırları içinde olmadığını göstermiştir. RKG, trafik kaza yaralanmaları sayısını değerlendirirken kaza sayısından gelen değişimin etkilerini de göze almıştır. RKG'nin, süreç analizi ve kalite iyileştirme için sağlam bir yöntem olarak hizmet, üretim, araştırma ve sağlık hizmetleri süreçlerinde yaygın olarak kullanılmasının önerildiği sonucuna varılmıştır.

Kaynakça

Shewhart, W. A. (1931). *Economic control of quality of manufactured product.* London: MacMillan and Co.

Montgomery, D. C. (2013). *Introduction to Statistical Quality Control (7th ed.).* John Wiley & Sons.

Evans, J.R., & Lindsay, W.M. (2014). *Managing for Quality and Performance Excellence.* Cengage Learning.

Montgomery, D.C., & Runger, G.C. (2018). *Applied Statistics and Probability for Engineers*. Wiley.

Woodall, W.H., & Montgomery, D.C. (1999). Research Issues and Ideas in Statistical Process Control. *Journal of Quality Technology, 31* (4), 376–386.

Woodall, W. H. (2000). Controversies and Contradictions in Statistical Process Control. *Journal of Quality Technology*, 32 (4), 341-350.

Alwan, L.C., & Roberts, H.V. (1988). Time-Series Modeling for Statistical Process Control. *Journal of Business & Economic Statistics, 6* (1), 87–95.

Pan, X., & Jarrett, J. (2004). Applying State Space Model with Regression Adjustment to SPC. *International Journal of Production Economics, 91*(1), 173–184.

Neslihanoğlu, S., & Paker, M. (2021). Beta Risklerinin Modellenmesi ve Tahmini: Türkiye'deki Döviz Portföyü Örneği. *Çankırı Karatekin İİBF Dergisi, 11* (2), 467-491.

Ryan, T. P. (2011). *Modern Regression Methods.* Wiley.

Vargas, F. A. (2003). Robust estimation in multivariate control charts forindividual observations. *Journal of Quality Technology,* 35 (4), 367-376.

Zou, C., & Qiu, P. (2009). Multivariate Statistical Process Control Using LASSO. *Journal of American Statistical Association,* 104 (488), 1586-1596.

Mandel, B.J. (1969). The regression control chart. *Journal of Quality Technology*, 1, 1–9.

Shewhart, W. (1931). *The Application of Statistical Methods to Quality Control.* McGraw-Hill.

Hayati, F. (2017). Regression control chart for two related variables: a forgotten lesson. *Int. J. Modelling in Operations Management,* 6 (4), 262-279.

Mammadova, U., & Özkale, M. R. (2018). Poisson Yanıt Değişkenli Genelleştirilmiş Lineer Modellere Dayalı Kontrol Grafikleri. *Ç.Ü Fen ve Mühendislik Bilimleri Dergisi, 35*(2), 95-104.

Russo, S., Camargo, M. E., & Samohyl, R. W. (2008). Contol Chart Based On The Poisson Regression Models Residuals. *ABEPRO, 8*(4), 1-10.

BÖLÜM VII

Genelleştirilmiş Pareto-Poisson Süreci ve bir uygulama

Gamze ÖZEL KADILAR[1]

Giriş

Poisson süreci, stokastik süreçler teorisinde temel bir yapı taşıdır ve özellikle rastgele olayların belirli bir zaman diliminde meydana gelme sıklığını modellemek için kullanılmaktadır. Bu süreç, belirli bir zaman aralığında meydana gelen olayların sayısının Poisson dağılımına ve olaylar arasında geçen sürenin ise üstel dağılıma uyduğu varsayımına dayanmaktadır (Özel Kadılar, 2023). Poisson süreci, olayların bağımsız ve belirsiz bir şekilde gerçekleştiği birçok veriyi modellemek için kullanılmaktadır. Telekomünikasyon alanında, arama sayıları veya veri paketlerinin belirli bir zaman diliminde bir ağa gelme oranlarını modellemek için

[1] Prof. Dr., Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Ankara/Türkiye, Orcid: 0000-0003-3886-3074, gamzeozl@hacettepe.edu.tr

kullanılmaktadır. Sigortacılıkta, belirli bir süre içinde bir sigorta şirketine bildirilen hasar sayısını tahmin etmek için kullanılmaktadır. Ayrıca biyolojik bilimlerde, radyasyon kaynaklı hücre mutasyonlarının zaman içindeki dağılımını modellemede de Poisson sürecinden faydalanılmaktadır. Finansal piyasalarda ise Poisson süreci, yüksek frekanslı ticaret işlemleri ve bu işlemlerin yoğunluğu üzerine analizler yapmak için kullanılmaktadır. Jeolojide Poisson süreci, depremlerin meydana gelme sıklığını modellemek için yaygın olarak kullanılan bir yöntemdir. Bu süreç, belirli bir bölgede ve zaman aralığında depremlerin geliş zamanlarının bağımsız ve rastgele olduğu varsayımına dayanmaktadır. Poisson sürecinde depremlerin belirli bir zaman aralığında sabit bir ortalama oranla meydana geldiği ancak bir depremin diğerini tetikleme veya etkileme olasılığı yoktur. Yani, her bir deprem olayı bağımsız olarak değerlendirilmektedir (Özel, 2013). Ancak, depremler genellikle birbirini tetikleyen olaylar şeklinde gerçekleşir; bu nedenle, özellikle artçı depremler ve deprem kümelenmeleri gibi durumlarda, daha karmaşık süreçlerin kullanılması gerekebilir.

Poisson sürecinin temelleri üzerine geliştirilen birleşik Poisson süreci, olayların sayısına ek olarak, her bir olayın büyüklüğünü de modellemeye olanak tanımaktadır. Birleşik Poisson sürecinde, her bir olay, bir rasgele değişken ile ilişkilendirilir ve bu rasgele değişkenler birbirinden bağımsız ve özdeş dağılımlıdır (Bak vd., 2002). Aynı zamanda, bu rasgele değişkenler, temel Poisson sürecinden bağımsızdır. Sonuç olarak, belirli bir zaman diliminde meydana gelen olayların büyüklüklerinin toplamı, Poisson dağılımlı rasgele bir toplam oluşturmaktadır (Dutfoy, 2019).

Birleşik Poisson süreci, hem olayların sıklığının hem de büyüklüğünün rasgele olduğu verilerin incelenmesinde kullanılmaktadır. Birleşik Poisson süreci, klasik Poisson sürecinin bir genellemesidir. Klasik Poisson sürecinde, belirli bir zaman diliminde meydana gelen olayların sayısı Poisson dağılımına uyarken, birleşik Poisson sürecinde, her bir olayın (bu bağlamda deprem) yanında bir büyüklük veya değer de bulunur. Bu büyüklükler, belirli bir dağılıma göre rastgele olarak belirlenir ve bu süreçte meydana gelen tüm olayların toplam büyüklüğü incelenir. Sigortacılıkta toplam tazminat taleplerini modellemek için kullanılırken, finansal risk yönetiminde, risklerin hem gerçekleşme sıklığını hem de potansiyel kayıpların büyüklüğünü değerlendirmek için uygulanmaktadır (Ferrares, 2003). Ayrıca, bu süreç, biyolojik ve epidemiyolojik modellemede, belirli bir süre içinde ortaya çıkan vakaların sayısını ve bu vakaların şiddetini incelemek için de kullanılmaktadır. Birleşik Poisson süreci, hem depremlerin meydana gelme olasılığını hem de bu depremlerin büyüklüğünü birlikte değerlendirmeye olanak sağlamaktadır (Özel, 2011a). Bu da, belirli bir bölgede, belirli bir süre içinde beklenen toplam sismik enerji miktarını veya potansiyel hasarın tahmini açısından önemlidir.

Türkiye, Alp-Himalaya deprem kuşağı üzerinde yer alarak, dünyadaki en aktif sismik bölgelerden birinde konumlanmaktadır. Ülkemizin yüzölçümünün %42'si birinci derece deprem kuşağına dâhil olup, özellikle Ege Bölgesi, aktif tektonik yapısı ile öne çıkmaktadır. Bu bölge, Gediz graben sisteminin batı ucunda yer almakta ve Batı Anadolu Açılma Rejimi'nin etkisi altında bulunmaktadır. Bu nedenle, tarih boyunca birçok yıkıcı depreme maruz kalmış ve ciddi zararlar görmüştür. Bölgede son olarak

gerçekleşen ve hasara neden olan Seferihisar (İzmir) depremi (30 Ekim 2020, Mw 6.6) gibi olaylar, Ege Bölgesi'nin yüksek deprem potansiyelinin daha ayrıntılı bir şekilde incelenmesi gerektiğini bir kez daha ortaya koymuştur.

Bu çalışmada, Ege Bölgesi'nde meydana gelen yıkıcı depremlerin hem sıklığını hem de büyüklüğünü tahmin etmek için birleşik Poisson süreci kullanılmıştır. Bu model, bölgedeki depremlerin yalnızca meydana gelme sıklığını değil, aynı zamanda her bir depremin büyüklüğünü de dikkate alarak, gelecekteki sismik tehlikelerin ve potansiyel zararların daha doğru bir şekilde öngörülmesine olanak tanımaktadır. Çalışmanın amacı, bu süreç üzerinden elde edilen sonuçlarla, Ege Bölgesi'nin sismik risk profilini daha kapsamlı bir şekilde ortaya koymak ve bu risklere karşı alınabilecek önlemlerin geliştirilmesine katkıda bulunmaktır.

2. POISSON SÜREÇLERİ

2.1. Klasik Poisson Süreci

Stokastik süreçler teorisinde önemli bir yere sahip olan Poisson süreçlerinin biyoloji, tıp, jeoloji, sismoloji, meteoroloji, endüstri, finans, sigortacılık gibi birçok alanda nadir olay sayılarının zamana bağlı incelenmesinde uygulamaları bulunmaktadır. Poisson süreçleri taşıdıkları özelliklere göre, homojen (homogeneous) Poisson süreci, homojen olmayan Poisson süreci ve birleşik (compound) Poisson süreci adını almaktadır. Depremlerin zaman içinde gösterdikleri rassal dağılımın modellenmesi için olasılıksal sismik tehlike analizi çalışmalarının çoğunluğunda depremlerin zaman içindeki oluşumları homojen Poisson süreci ile modellenmektedir. Poisson modelinde belirli bir büyüklüğün

üzerindeki deprem sıklığının birbirlerinden bağımsız oldukları varsayılmaktadır (Yazdani ve Kowsari, 2011).

Klasik (ordinary) Poisson süreci, belirli bir zaman aralığında meydana gelen olayların sayısını modelleyen bir stokastik süreçtir. Bu süreçte olaylar bağımsız olarak ve sabit bir ortalama oranla meydana gelmektedir Bir Poisson sürecinde, belirli bir zaman diliminde gerçekleşen olayların sayısı, λ oranıyla bir Poisson dağılımına uymaktadır. Bir sismik kaynak içerisinde (0, t] zaman aralığında ortaya çıkan deprem sayısı $\{N_t, t \geq 0\}$ ile gösterilmek üzere, t zamanına dek i tane depremin ortaya çıkma olasılığı aşağıdaki gibi elde edilir:

$$P(N_t = i) = p_i(t) = e^{-\lambda t} \frac{(\lambda t)^i}{i!}, \quad \boldsymbol{i = 0, 1, 2, \ldots}$$

Burada λ, incelenen bölgede birim zamanda (genellikle bir yıl) ortaya çıkması beklenen olay sayısıdır ve ortalama oran adını alır. Yukarıda elde edilen bu fonksiyon Poisson olasılık fonksiyonudur. λ, ise, zaman içerişinde değişmeyen sabit bir değere eşittir. Bu nedenle, $\{N_t, t \geq 0\}$ sürecine homojen Poisson süreci adı verilir (Özel Kadılar, 2020). Sürecin beklenen değer ve varyansı sırasıyla aşağıdaki gibidir:

$$E(N_t) = \lambda t, \quad Var(N_t) = \lambda t$$

Bu çalışmada deprem sıklıklarının tahmininde homojen Poisson sürecinden yararlanılmıştır. Nt raslantı değişkeni, (0, t] zaman aralığında meydana gelen deprem sayısını göstermektedir. Poisson modeline göre bir sonraki depremin oluşması için geçen bekleme zamanının dağılımı, bir önceki depremin oluşundan itibaren geçen zamandan etkilenmez ve önceki çalışmalar homojen Poisson

sürecinin magnitüdü yüksek depremler için geçerli olduğunu göstermektedir. Ayrıca küçük alanlar göz önüne alındığında verilerin Poisson modele uymadığı görülmüştür. Deprem olaylarının zaman uzayında bir Poisson sürecine uyabilmeleri için aşağıdaki koşulları sağlaması gerekir:

• Herhangi bir yıl içerisinde olacak deprem, takip eden yılda olacak depremi etkilemez. Yani depremler zamanda bağımsızdır,

• Herhangi bir kaynakta oluşan deprem başka bir kaynakta olacak depremi etkilemez. Yani depremler uzayda bağımsızdır,

• Aynı anda ve aynı yerde iki ayrı depremin olma olasılığı sıfırdır.

Bu bağımsızlık koşulunu sağlama açısından öncü ve artçı depremlerin sismik tehlike analizinin dışında tutulması gerekmektedir. Bu proje çalışmasında da öncü ve artçı depremler ayıklanmış, kümesizleştirilmiş ana deprem sıklıklarının homojen Poisson süreci ile modellenmesi sağlanmıştır. Bu bağımsızlık koşulunu sağlama açısından öncü ve artçı depremlerin sismik tehlike analizinin dışında tutulması gerekmektedir. Bu çalışmada öncü ve artçı depremler ayıklanmış, kümesizleştirilmiş ana deprem sıklıklarının homojen Poisson süreci ile modellenmesi sağlanmıştır (Özel ve İnal, 2008).

2.2. Birleşik Poisson Süreci

t zamanına dek ortaya çıkan deprem sayısı $\{N_t, t \geq 0\}$ biçimindeki Poisson sürecine uyduğu gösterildikten sonra deprem sayısının az olmaması için 5.0 ve üzerindeki deprem kümesi ile çalışmaya devam edilmesine karar verilmiştir. t zaman doğrusu

boyunca ortaya çıkan i. depremin büyüklüğü (M_w) ile biçimindeki sürekli raslantı değişkeni ile gösterilmek üzere,

$$\boldsymbol{S_t = \sum_{i=1}^{N_t} X_i}$$

biçiminde tanımlanan sürecine birleşik Poisson süreci adı verilir. Bu süreç ile ilgili temel özellikler aşağıda özetlenmiştir (Özel, 2011a):

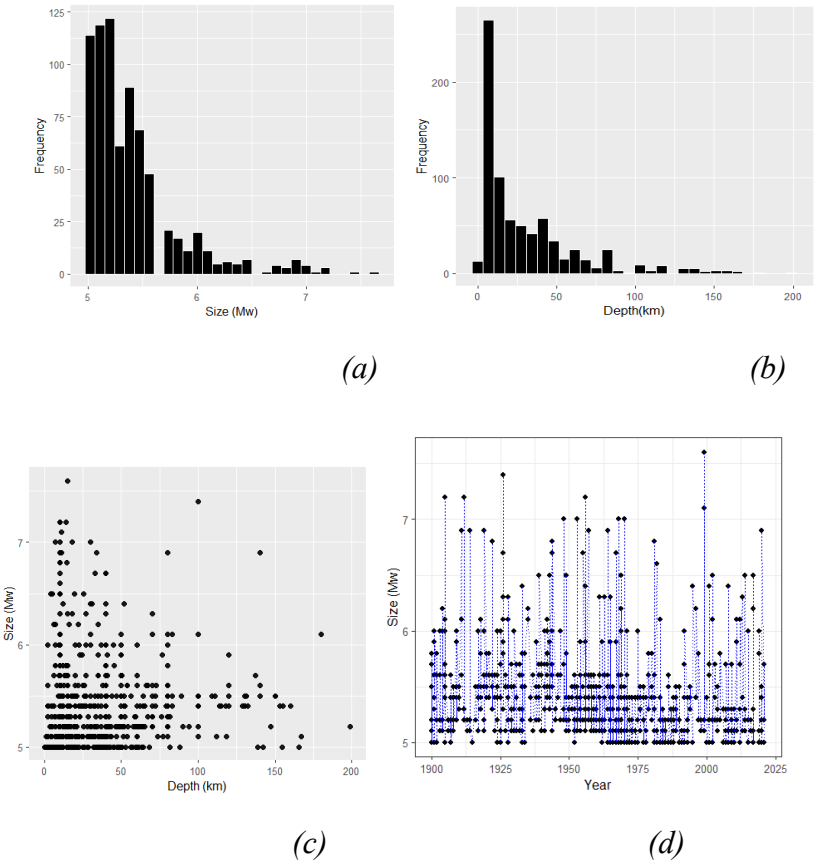
- N_t ve X_i bağımsız raslantı değişkenleridir.
- $E(S_t) = E(N_t)\, E(Y) = \lambda t$
- $V(S_t) = E(N_t)V(Y) + V(N_t)(E(Y))^2$
- Birleşik Poisson süreci sıçramalı bir süreçtir. Sürekli değişimler yerine belirli zaman noktalarında sıçramalar (olaylar) gerçekleşir.

Birleşik Poisson sürecinde X_i raslantı değişkenlerinin dağılımları belirlenerek farklı uygulama alanlarına ait modeler geliştirilebilir. Birleşik Poisson sürecinin analitik çözümleme yapılabilmesi için X_i'lerin dağılımına bağlı olarak çeşitli yaklaşımlar kullanılabilir. Ancak S_t 'nin dağılımı genellikle kapalı formda bulunamaz; bu durumda sayısal simülasyonlar veya moment tahminleri kullanılır (Özel, 2011b).

3. UYGULAMA

Bu çalışmada, Türkiye'nin aktif tektonik bölgelerinden biri olan Ege Bölgesi'nin deprem sıklığı ve büyüklükleri arasındaki

ilişkiler Poisson süreçleri yardımıyla incelenmiştir. Söz konusu çalışma, TÜBİTAK 1001 kapsamında tamamlanan “Yapay Zeka ve Olasılıksal Model Tabanlı Deprem Tehlike Haritası” (Proje No: 121F208) adlı projeye dayanmaktadır. Çalışma alanı, Batı Anadolu Açılma Rejimi'nin etkisi altında bulunan Gediz graben sisteminin batı ucunda yer almaktadır. Bölgedeki ana grabenler ve aktif fay sistemleri detaylı şekilde ele alınarak, bölgenin tektonik özellikleri, deprem aktivitesi ve alüvyon ovalarının oluşumu üzerinde durulmuştur. Özellikle Büyük Menderes ve Gediz grabenleri ile bölgenin diğer önemli fay hatlarının jeolojik ve neotektonik yapıları incelenmiştir. Bu kapsamda, bölgenin deprem potansiyeli ve tektonik dinamikleri analiz edilmiştir. hareketler sonunda bazı alanlar yükselmiş ve bugünkü dağlık alanları meydana getirmiştir. Bazı alanlarda ise çökmüş (graben) ile çöküntü alanları oluşmaktadır. Bu çöküntü sonunda alanlarının akarsular tarafında alüvyon setlerle doldurulması nedenleriyle yer alan ovalar oluşmuştur. Birleşik Poisson süreci ile modelleme çalışmasına başlamadan önce moment büyüklüğü 5.0 ve üzerinde olan depremlerin büyüklük (a), derinlik (b) sıklık, derinlik-büyüklük saçılım (c) ve zamana bağlı büyüklük çizgi (d) grafikleri RStudio paket programında “ggplot2” kütüphanesi ile elde edilmiş ve sırasıyla Grafik 1.a ile Grafik 1.b arasında sunulmuştur:



Grafik 1. Büyüklüğü 5.0 ve üzerinde olan depremlerin büyüklük (a), derinlik (b) sıklık, derinlik-büyüklük saçılım (c) ve zamana bağlı büyüklük çizgi (d) grafikleri

Grafik 1 incelendiğinde, deprem büyüklüklerinin ve derinliklerinin sağa çarpık dağıldıkları görülmektedir. Türkiye'de meydana gelen depremlerin çoğunlukla sığ depremler olduğu ve bu nedenle yüzeye yakın depremlerin yarattığı hasarın büyük olduğu bilinmektedir. Poisson süreci, yapısı gereği nadir olayların zamana

bağlı olarak meydana gelişlerinin incelenmesinde kullanılmaktadır. Deprem sıklıklarının Poisson süreci ile incelenebilmesi için Poisson dağılımının nadir olayların dağılımı olduğu düşünüldüğünde büyüklüğü en az 5.0 ve üzerindeki depremlerin Poisson dağılımına uyumunun test edilmesine karar verilmiştir. 1975 öncesi ve sonrası depremler arasında özellikle 5.0 ve altındaki büyüklükteki depremlerde gözlenen bu farklılık deprem kayıtlarının eski yıllarda özellikle küçük ölçeklerde kataloglarda doğru biçimde yer almamasından kaynaklanmaktadır. N_t raslantı değişkeni $(0, t]$ zaman aralığında meydana gelen deprem sayısını göstermektedir. Poisson modeline göre bir sonraki depremin oluşması için geçen bekleme zamanının dağılımı bir önceki depremin oluşundan itibaren geçen zamandan etkilenmez ve önceki çalışmalar homojen Poisson sürecinin magnitüdü yüksek depremler için geçerli olduğunu göstermektedir. Ayrıca küçük alanlar göz önüne alındığında verilerin Poisson modele uymadığı görülmüştür. Deprem olaylarının zaman uzayında bir Poisson sürecine uyabilmeleri için aşağıdaki koşulları sağlaması gerekir:

- Herhangi bir yıl içerisinde olacak deprem takip eden yılda olacak depremi etkilemez. Yani depremler zamanda bağımsızdır.
- Herhangi bir kaynakta oluşan deprem başka bir kaynakta olacak depremi etkilemez. Yani depremler uzayda bağımsızdır.
- Aynı anda ve aynı yerde iki ayrı depremin olma olasılığı sıfırdır.

Bu bağımsızlık koşulunu sağlama açısından öncü ve artçı depremlerin sismik tehlike analizinin dışında tutulması gerekmektedir. Çalışmada öncelikle öncü ve artçı depremler ayıklanmış, kümesizleştirilmiş ana deprem sıklıklarının homojen Poisson süreci ile modellenmesi sağlanmıştır. Tek örneklem K-S (Kolmogorov-Smirnov) testi bir veri setinin belirli bir teorik dağılıma uyumunu değerlendirmek için kullanılan bir istatistiksel testtir. Poisson süreci için birim zaman yıl olarak tanımlanmış, 5.0 ve üzerindeki depremler beş sınıfa ayrılarak her sınıftaki deprem sıklığının K-S testi ile sürece uyumu incelenmiştir. Elde edilen sonuçlar Tablo 1'de verilmiştir:

Tablo 1. Tek Örneklem Kolmogorov-Smirnov Testi Sonuçları

Deprem Büyüklük Sınıfı	5.0 ve üzeri		5.5 ve üzeri	6.0 ve üzeri	6.5 ve üzeri	7.0 ve üzeri
λ	Ortalama	6.1475	2.0082	0.6475	0.2623	0.0820
Z Değeri	0.768		0.608	0.164	0.043	0.055
p-değeri	0.597		0.853	1.000	1.000	1.000

Tablo 1'e göre, her deprem sınıfının aslında p-değerinin 0.05'ten büyük olduğu için ana deprem sıklıklarının Poisson dağılımına uyduğu sonucuna varılmıştır. Tablo 2'de süreçlere ait λ

değerlerine de yer verilmiştir. Tablo 1'e göre, 5.0 ve üzerinde bir deprem yılda ortalama olarak yaklaşık 6 kez görülmektedir. λ değerleri sırasıyla 6.1475, 2.0082, 0.6475, 0.2623 ve 0.0820 alınarak farklı zaman aralıklarında en az bir deprem olma olasılıklarını R kullanarak hesaplanmış ve sonuçlar Tablo 2'de verilmiştir:

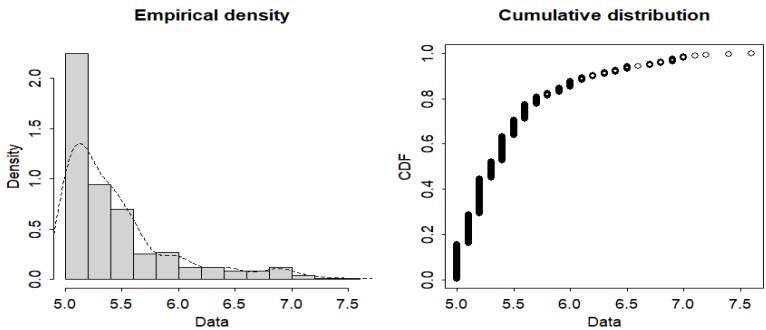
Tablo 2. Her deprem sınıfı için deprem tekrarlanma olasılıkları

Deprem Büyüklük Sınıfı	Yıl	1	2	3	4
	λ	Olasılıklar			
5.0 ve üzeri	6.1475	0.98223	0.99867	0.99983	0.99998
5.5 ve üzeri	2.0082	0.86578	0.98127	0.99714	0.99960
6.0 ve üzeri	0.6475	0.51971	0.76500	0.89535	0.96034
6.5 ve üzeri	0.2623	0.76759	0.93515	0.98727	0.99751
7.0 ve üzeri	0.0820	0.33907	0.54017	0.67886	0.78107

Deprem tekrarlanma olasılıkları, belirli büyüklüklerdeki depremlerin belirli zaman dilimlerinde meydana gelme olasılıklarını ifade etmektedir. Tablo 2'de belirli büyüklüklerdeki depremlerin belirli zaman aralıklarında tekrarlanma olasılıkları verilmiştir. Farklı

büyüklük sınıfları için çeşitli zaman dilimlerinde (1 yıl, 2 yıl, 3 yıl, 4 yıl) depremlerin tekrarlanma olasılıkları yıllık frekans (λ) kullanılarak hesaplanmıştır. 5.0 ve üzerindeki bir depremin yıllık tekrarlanma olasılığı %98.2 olup, bu büyüklükteki depremin bölgede 5.0 ve üzerinde hiç deprem olmaması durumunda 4 yıl içinde gerçekleşme olasılığı %99.9'dur; bu durum depremlerin sıkça yaşandığını işaret etmektedir. 5.5 ve üzeri depremler için yıllık olasılık %86.6 ve yine 5.5 ve üzerinde herhangi bir deprem olmaması durumunda4 yıllık tekrarlanma olasılığı %99.96'dır. 6.0 ve üzeri depremlerin yıllık tekrarlanma olasılığı %52.0, 4 yıllık tekrarlanma olasılığı ise %96.0346'dur. 6.5 ve üzeri depremlerin yıllık tekrarlanma olasılığı %76.8 ve 4 yıllık tekrarlanma olasılığı %99.7519'dur. 7.0 ve üzeri depremler için yıllık tekrarlanma olasılığı %33.9, 4 yıllık olasılık ise %78.1079 olup, bu büyüklükteki depremlerin daha nadir ancak yine de ciddi bir risk oluşturabileceğini göstermektedir. Dikkat edilecek önemli bir nokta depremin büyüklük sınıfının artması durumunda tekrarlanma olasılığının azalması ve araştırma alanında 5.0 ve üzerinde depremin dört yılda bir tekrarlanma olasılığının yüksek olmasıdır. Depremlerin tekrarlanma olasılıkları, biriken enerjinin serbest bırakılması ile doğrudan ilişkilidir. 5.0 ve üzeri büyüklükte bir deprem araştırma sahasında meydana gelmedikçe yıllar içinde tekrarlanma olasılığı giderek artmaktadır. Bu bilgiler, deprem hazırlık ve risk değerlendirmelerinde kritik öneme sahiptir. t zamanına dek ortaya çıkan deprem sayısı $\{N_t, t\geq 0\}$ biçimindeki Poisson sürecine uyduğu gösterildikten sonra deprem sayısının az olmaması için 5.0 ve üzerindeki deprem kümesi ile çalışmaya devam edilmesine karar verilmiştir. t zaman doğrusu boyunca ortaya çıkan

i. depremin büyüklüğü (Mw) $X_i, i = 1, 2, \ldots$ ile biçimindeki sürekli raslantı değişkeni ile gösterilmek üzere, $\{S_t, t \geq 0\}$ sürecine birleşik Poisson süreci ile zamana bağlı toplam deprem büyüklükleri incelenmiştir. Grafik 2'de t zaman doğrusu boyunca ortaya çıkan i. depremin büyüklüğü (Mw) $X_i, i = 1, 2, \ldots$ raslantı değişkenine ait deneysel yoğunluk ve birikimli dağılım fonksiyonu grafikleri sırasıyla Grafik 2.a ve Şekil 2.b'de görülmektedir:



a. Deprem Büyüklüklerine ait Yoğunluk Fonksiyonu

b. Deprem Büyüklüklerine ait Birikimli Dağılım Yoğunluk Fonksiyonu

Grafik 2. Deprem Büyüklüklerine Ait Grafikler

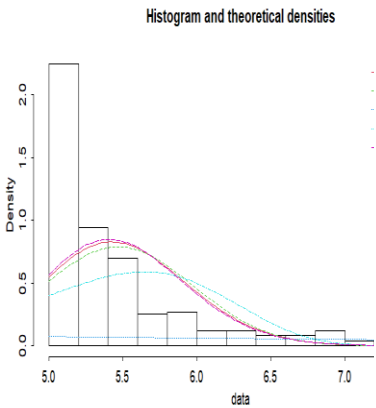
Tablo 3'te deprem büyüklüklerine ait temel istatistikler elde edilmiştir:

Tablo 3. 5.0 ve üzerindeki deprem büyüklüklerine ait temel istatistikler

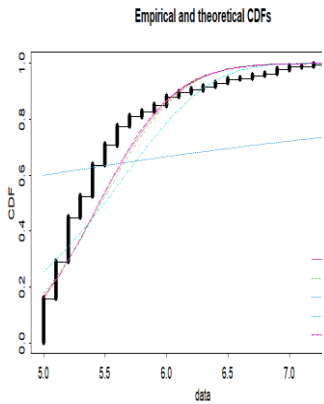
Minimum	Q_1	Medyan	Ortalama	Q_3	Maksimum
5.0	5.10	5.30	5.47	5.60	7.60

Tablo 3'e göre, en küçük deprem büyüklüğü 5.0 ve en büyük deprem büyüklüğü 7.6 olarak belirlenmiştir. Medyan değeri 5.3 iken ortalama deprem büyüklüğü değer 5.5 olarak hesaplanmıştır. Bu istatistikler, veri setinin merkezi eğilimini ve dağılımını özetlerken, 1. çeyrek değeri 5.1 ve 3. çeyrek değeri 5.6 olarak belirlenmiştir. Bu değerler veri kümesinin alt ve üst yarısının ortalamasının altında ve üstünde olan değerleri ifade etmektedir. R programında ayrıca "D'Agostino çarpıklık testi" de yapılmıştırve çarpıklık değeri 1.666, Z-değeri 10.141 ve p-değeri: < 2.2e-16 (veya yaklaşık olarak 0.00000000000000022) olarak elde edilmiştir. Çarpıklık, bir veri dağılımının simetrik olup olmadığını gösteren bir ölçüdür. Pozitif çarpıklık, sağa çarpıktır (uzun kuyruk sağda), negatif çarpıklık ise sola çarpıktır (uzun kuyruk solda). Burada, 1.666 olan çarpıklık değeri pozitif olduğundan, 5.0 ve üzerindeki deprem büyüklüklerinin beklenildiği gibi sağa çarpık olduğu gösterilmiştir. R paket programında ayrıca Anscombe-Glynn basıklık (kurtosis) testi yapılmış ve basıklık değeri 5.5211, Z-değeri 5.3562 ve p-değeri 8.499e-08 (veya yaklaşık olarak 0.00000008499) olarak elde edilmiştir. Basıklık, bir veri dağılımının kuyruk ağırlığının ve tepe noktasının ölçüsüdür. Basıklık değeri 3'e eşitse, veri normal dağılıma sahiptir. Bu testte, basıklık değeri 5.5211 olduğu için, 5.0

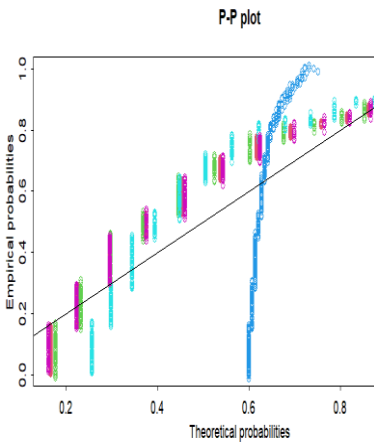
ve üzerindeki deprem büyüklüklerinin normal dağılıma sahip olmadığı sonucuna ulaşılmıştır. 5.0 ve üzerindeki deprem büyüklüklerinin sahip olabileceği dağılımlar sağa çarpık dağılımlar (gama, Weibull, log-normal) dağılımları arasından denenmiştir. Ayrıca doğruluğun sınanması açısından üstel ve normal dağılımlar da denenmiştir. Ancak bu dağılımlara ait Şekil Grafik 3'te verilen uyum grafikleri ve Tablo 6.4'te K-S İstatistiği, Cramer-von Mises (CVM) İstatistiği, Anderson-Darling (AD) İstatistiği, Akaike Bilgi Kriteri (AIC), Bayesian Bilgi Kriteri (BIC) sonuçlarına göre belirlenen dağılımlar ile deprem büyüklük dağılımı arasında uyum olmadığını göstermiştir.

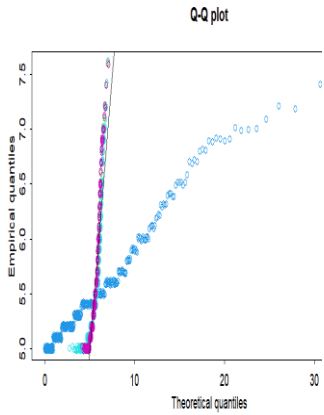


a. Deneysel ve Kuramsal Yoğunluk Fonksiyonları



b. Kuramsal ve Deneysel Dağılım Fonksiyonları





c. P-P Plot Grafiği *d. Q-Q Plot Grafiği*

Grafik 3. 5.0 ve Üzerindeki Depremlerin Büyüklük Dağılımlarına Ait Uyum İyiliği Grafikleri

K-S testi, bir veri setinin belirli bir teorik dağılıma uygun olup olmadığını değerlendirmek için kullanılan bir istatistik testidir. Tablo 4'te elde edilen sonuçlar incelendiğinde K-S testi p-değerlerine göre bu dağılımların deprem büyüklük verisine uymadığı belirlenmiştir.

Tablo 4. Uyum İyiliği İstatistik ve Kriterleri

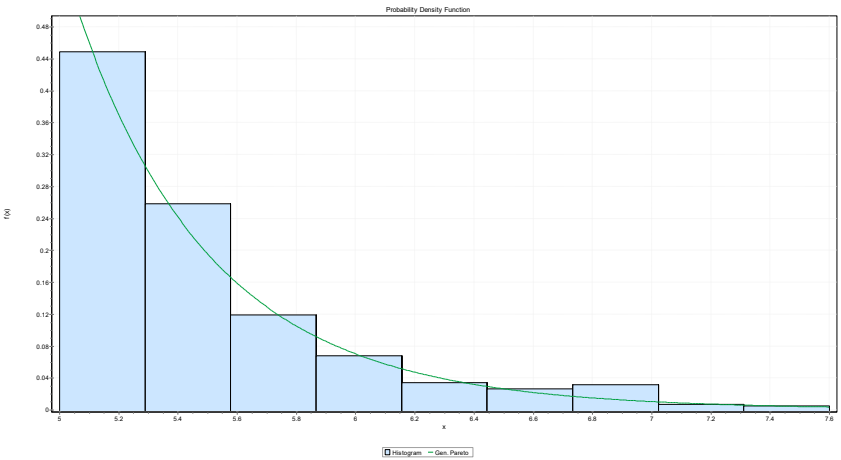
Uyum İyiliği İstatistikleri	Gamma	Normal	Üstel	Weibull	Log-Normal
K-S İstatistiği	0.182	0.191	0.599	0.255	0.177
K-S testi p-değeri	2.911e-12	1.87e-13	< 2.2e-16	< 2.2e-16	1.248e-11
CVM İstatistiği	3.60913	4.14385	34.8548	6.3387	3.35243
AD İstatistiği	21.449	24.222	160.033	35.191	20.116
Uyum İyiliği Kriterleri	Gamma	Normal	Üstel	Weibull	Log-Normal
AIC	569.762	608.047	2215.33	765.857	551.980
BIC	577.795	616.080	2219.35	773.890	560.013

Tablo 5’te KS ve AD test sonuçlarına göre belirlenen en iyi beş dağılıma yer verilmiştir:

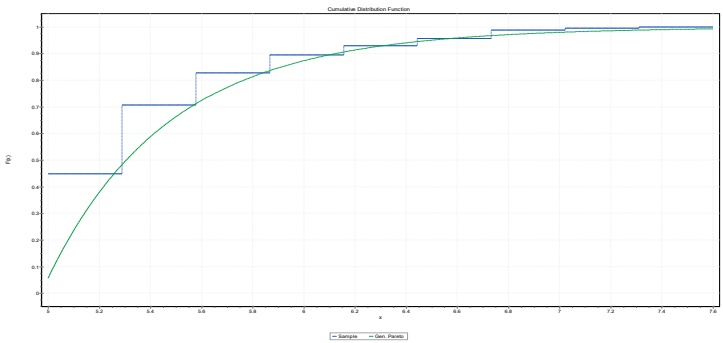
Tablo 5. Uç Değer Dağılımlara Ait Uyum İyiliği İstatistikleri

Dağılım	Kolmogorov-Smirnov		Anderson-Darling	
	İstatistik	Sıra	İstatistik	Sıra
Genelleştirilmiş Pareto	0.10179	1	4.4547	1
Frechet	0.10677	2	5.5678	3
Genelleştirilmiş Uç Değer	0.10857	3	4.5258	2
Pearson-5	0.10866	4	5.6258	4
Ters Normal	0.11696	5	6.5183	5

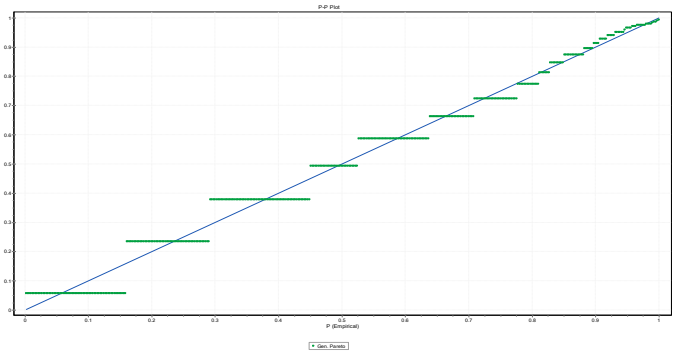
Tablo 5 incelendiğinde genelleştirilmiş Pareto dağılımının 5.0 ve üzerindeki deprem büyüklüklerine en iyi uyumu gösterdiği sonucuna ulaşılmıştır. Genelleştirilmiş Pareto dağılımına ait kuramsal eğri ile deprem büyüklüklerine ait histogram grafiği Grafik 4'te, birikimli dağılım fonksiyonu grafiği Grafik 5'te, P-Plot grafiği Grafik 6 ve Q-Q Plot grafiği Grafik 7'de verilmiştir:



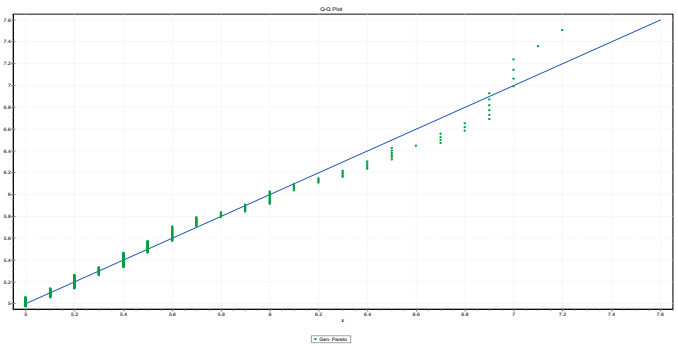
Grafik 4. Genelleştirilmiş Pareto Dağılım Eğrisi ve Deprem Büyüklüklerine Ait Histogram



Grafik 5. Genelleştirilmiş Pareto Dağılım Eğrisi ve Deprem Büyüklüklerine Ait Dağılım Fonksiyonu



Grafik 6. Deprem Büyüklük Değerleri ve P-P Plot Grafiği



Grafik 7. Deprem Büyüklük Değerleri ve Q-Q Plot Grafiği

Grafik 4 ile Grafik 7 arasında verilen grafikler incelendiğinde deprem büyüklüğü için genelleştirilmiş Pareto dağılımının oldukça uyumlu olduğu görülmektedir. Tablo 6'da genelleştirilmiş Pareto dağılımı için KS ve AD test sonuçları verilmiştir:

Tablo 6. Genelleştirilmiş Pareto Dağılımı İçin K-S ve AD test Sonuçları

Kolmogorov-Smirnov Testi Sonuçları					
İstatistik	0.10179				
p-değeri	3.7557E-4				
α	0.2	0.1	0.05	0.02	0.01
Kritik Değer	0.05299	0.0604	0.06707	0.07497	0.08045
Anderson-Darling Testi Sonuçları					
İstatistik	4.4547				
α	0.2	0.1	0.05	0.02	0.01
Kritik Değer	1.3749	1.9286	2.5018	3.2892	3.9074

Tablo 6 incelendiğinde genelleştirilmiş Pareto dağılımının 5.0 ve üzerindeki deprem büyüklüklerine uygun olduğu sonucuna ulaşılmıştır. Bu dağılıma ait en çok olabilirlik parametre tahmin

değerleri ise, şekil paremetresi k= 0.04704, ölçek parametresi σ=0.04704; konum parametresi μ=4.9723 olarak bulunmuştur. t zamanına dek ortaya çıkan 5.0 ve üzeri büyüklükteki deprem sayısı $\{N_t, t \geq 0\}$ biçimindeki Poisson süreci, t zaman doğrusu boyunca ortaya çıkan i. depremin büyüklüğü (Mw) $X_i, i = 1, 2, \ldots$ ile biçimindeki genelleştirilmiş Pareto dağılımlı sürekli raslantı değişkeni ile gösterilmek üzere, $\{S_t, t \geq 0\}$ sürecine genelleştirilmiş Pareto dağılımlı birleşik Poisson süreci adı verilmiştir. Poisson süreci ve genelleştirilmiş Pareto dağılımı kullanarak S_t için beklenen değeri hesaplamak için öncelikle deprem büyüklük dağılımı olan genelleştirilmiş Pareto dağılımının beklenen değeri E(X)=μ+(σ-1-k) ve 5.0 ve üzeri büyüklükteki deprem sayısını gösteren homojen Poisson sürecinin beklenen değeri $E(N_t) = \lambda t$ olarak alınarak t zamanına dek meydana gelen toplam deprem büyüklüğünü gösteren $E(S_t)$ değerleri deprem sınıflarına ait farklı λ değerleri ve t =0.5 (6 ay) için R kodu yardımıyla hesaplanmıştır. Tablo 7'de 6 ay içerisinde bölgede meydana gelebilecek farklı büyüklerdeki depremlere ait ortalama deprem sayısı ve toplam ortalama deprem büyüklükleri verilmiştir:

Tablo 7. 6 ay içerisinde meydana gelebilecek farklı büyüklerdeki depremlere ait ortalama deprem sayısı ve toplam ortalama deprem büyüklükleri

Deprem Sınıfı	λ	$E(N_t)$	$E(S_t)$
5.0 ve üzeri	6.15	6.15	30.9
5.5 ve üzeri	2.01	2.01	10.1
6.0 ve üzeri	0.648	0.648	3.25
6.5 ve üzeri	0.262	0.262	1.32
7.0 ve üzeri	0.082	0.082	0.412

Tablo 7 incelendiğinde depremlerin büyüklüğü arttıkça beklenen ortalama deprem sayısının azaldığı görülmektedir. Burada S_t, belirli bir zaman aralığında (t) meydana gelen depremlerin etkisinin toplamını temsil etmektedir. N_t, Poisson süreci, belirli bir zaman aralığında meydana gelen 5.0 ve üzerindeki deprem sayısını göstermektedir. Birleşik Poisson sürecinde depremlerin zamanlaması rastgeledir. X_i ise genelleştirilmiş Pareto dağılımına sahip rastgele değişkeni olup, i-inci depremin büyüklüğünü veya etkisini ifade etmektedir. Bu durumda, S_t toplam etkiyi diğer bir

deyişle, belirli bir zaman aralığında meydana gelen tüm depremlerin büyüklüklerinin toplamını ifade etmektedir.

KAYNAKLAR

Bak, P., Christensen, K., Danon, L., & Scanlon, T. (2002). Unified scaling law for earthquakes. Physical Review Letters, 88(17), 178501-1–178501-4. https://doi.org/10.1103/PhysRevLett.88.178501

Dutfoy, A. (2019). Estimation of tail distribution of the annual maximum earthquake magnitude using extreme value theory. Pure and Applied Geophysics, 176(2), 527–540. https://doi.org/10.1007/s00024-018-1965-8

Ferraes, S. G. (2003). The conditional probability of earthquake occurrence and the next large earthquake in Tokyo, Japan. Journal of Seismology, 7(2), 145–153. https://doi.org/10.1023/A:1022919706712

Özel, G., & İnal, C. (2008). The probability function of the compound Poisson process and an application to aftershock sequences. Environmetrics, 19(1), 79–85. https://doi.org/10.1002/env.867

Özel, G. (2011a). A bivariate compound Poisson model for the occurrence of foreshock and aftershock sequences in Turkey. Environmetrics, 22(7), 847–856. https://doi.org/10.1002/env.1126

Özel, G. (2011b). On certain properties of a class of bivariate compound Poisson distributions and an application to earthquake data. Revista Colombiana de Estadística, 34(3), 545–566.

Yazdani, A., & Kowsari, M. (2011). Statistical prediction of the sequence of large earthquakes in Iran. International Journal of Engineering, 24(4), 325–336.

Özel Kadılar, G. (2020). Stokastik Süreçler ve Uygulamaları. Hacettepe Üniversitesi Yayınları.

Özel Kadılar, G. (2023). Stokastik Modelleme ve Risk Analizi. İstanbul Medeniyet Üniversitesi Yayınları.

Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK). (2024). 121F208 – Yapay Zeka ve Olasılıksal Model Tabanlı Deprem Tehlike Haritası.

BÖLÜM VIII

Bir Veri Dengeleme Tekniği Olarak SMOTEENN

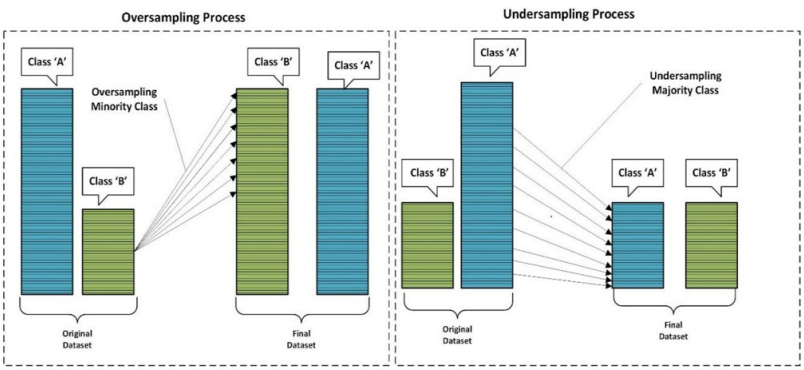
Gökhan KORKMAZ

1. Giriş

Çoğunlukla, gerçek dünya veri kümeleri doğası gereği sınıf dengesizliğine sahiptir ve dengesiz sınıf dağılımları sınıflandırıcıların eğitiminde çeşitli sorunlara yol açar. Bu durum, sınıflandırıcılar için düşük doğruluk, kesinlik, geri çağırma ve yüksek derecede yanlış sınıflandırma vb. sorunlarına yol açar (Kumar & ark., 2022).

Sınıflandırma problemlerinin çözüm süreçlerinde kullanılan veri kümesinin dengesi, sağlıklı sınıflandırma sonuçları elde edilmesinde çok önemlidir. Eğer veri kümesi yeterince dengeli değilse veri kümesinin dengeleme yöntemlerinin kullanımı sınıflandırma algoritmalarının performansını artıracaktır. Sınıflandırma süreçlerinde performans ölçümleri doğru sınıflandırılanları ve yanlış sınıflandırılanları bulmak için eğitilir,

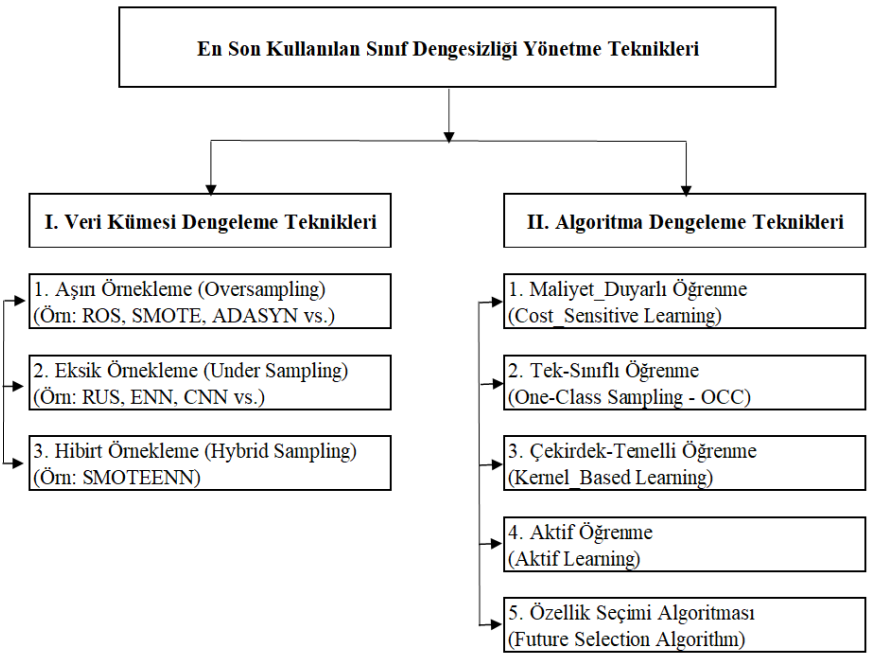
ardından bunların oranı sınıflandırıcının verimliliğini bulmak için kullanılır. Sayıları az olsa bile, tüm önemli sınıflar yanlış sınıflandırılmış olsa bile yüksek bir doğruluk oranı elde etmek mümkündür. Bu durumu aşmak için veriler dengelenmelidir. Bunun için kullanılabilecek olan yöntemler yetersiz örnekleme veya aşırı örnekleme yöntemleri olabilir. Rastgele aşırı örnekleme Şekil 1'de (Kumar & ark., 2022) görüldüğü üzere azınlık kategorisinin doğru versiyonlarını ürettiği için aşırı uyumu artırıcı bir etki yapar. Rastgele düşük örnekleme, önemli ve de muhtemelen yararlı olan bilgileri atabilmektedir. Her ikisinin kombinasyonu ve bilhassa SMOTEENN, bu dezavantajların üstesinden gelebilmektedir (Manju & Nair, 2019).



Şekil 1: Rastgele Aşırı Örnekleme ve Az Örnekleme Süreçleri

Şekil 2'de de görüleceği üzere veri kümesi dengeleme çözümleri, sınıf dengesizliğini azaltmayı amaçlar. Hibrit örnekleme teknikleri aşırı örnekleme ve az örnekleme tekniklerini birleştirir. algoritma dengeleme çözümleri, maliyet veya ağırlık fonksiyonu vb. kullanarak altta yatan sınıflandırma algoritmalarının davranışını değiştirir, veri dengeleme yaklaşımları, tahmin sonuçlarını daha

güvenilir ve doğru hale getirerek çoğunluk sınıfına yönelik önyargıyı düzeltmeyi hedeflerken, maliyet duyarlı sınıflandırma gibi algoritma dengeleme yaklaşımları, yanlış sınıflandırma maliyetlerini çoğunluk ve azınlık sınıfları için eşit olduğunu varsaymak yerine sınıf ağırlığına göre ayarlamaya odaklanır. Ancak, bu yaklaşımların her birinin avantajları ve dezavantajları vardır. Aşırı örnekleme teknikleri aşırı uyuma ve yüksek hesaplama ve bellek maliyetine yol açabilir ve. Az örnekleme teknikleri, bazı değerli bilgilerin yetersiz uyuma ve olası bilgi kaybına yol açabilir (Srivastava & Sharan, 2023).



Şekil 2. Veri Dengeleme Yöntemleri

SMOTEENN, aşırı örnekleme "Sentetik Azınlık aşırı örnekleme Tekniği" (Synthetic Minority Over-sampling Technique

- SMOTE) ve az örnekleme “Düzenlenmiş En Yakın Komşu (Edited Nearest Neighbors - ENN)” tekniklerinin birleşiminden oluşmaktadır. SMOTE, KNN yöntemi ile en yakın k komşuyu seçer, bunları birbirine bağlar ve uzayda yeni örnekler oluşturulur. ENN yöntemi aşırı örneklemeli verileri temizler (Manju & Nair, 2019).

Her örnek x için, k-en yakın komşuları içindeki farklı sınıflardan gelen örnek sayısı hesaplanır. x'in sınıfı k-en yakın komşuları arasında çoğunluktaysa, örnek tutulur; aksi takdirde kaldırılır. 1. ve 2. Adımlar, istenen sınıf dengesi ve örnek kalitesi elde edilene kadar tekrarlanır. Her yinelemede, SMOTE azınlık sınıfı örnekleri arasında interpolasyon yaparak yeni sentetik örnekler üretir ve azınlık sınıfı örneklerinin sayısını artırır. Bu, azınlık sınıfının karar sınırını genişletmeye ve sınıflandırıcının azınlık sınıfı örneklerini tanıma yeteneğini iyileştirmeye yardımcı olur. ENN ise çoğunluk sınıfından bazı örnekleri kaldırarak çoğunluk sınıfı örneklerinin sayısını azaltır. Kaldırılan örnekler genellikle sınıflandırıcının performansını düşürebilecek gürültü veya sınır örnekleridir. ENN böylelikle gürültüyü ve sınır örneklerini kaldırarak veri setini kademeli olarak optimize eder. SMOTE'nin aşırı örnekleme oranını ve ENN'nin alt örnekleme oranını ayarlayarak, son veri setindeki her sınıftaki örnek sayısı kontrol edilebilir ve böylece sınıf dengesi elde edilebilir (Yu & ark., 2024).

SMOTEENN yönteminin, yetersiz örnekleme ve aşırı örneklemenin mevcut diğer kombinasyonlarından daha iyi olduğu kanıtlanmıştır. SMOTEENN'in ayrıca ENN-SMOTE'dan da daha iyi olduğu da bulunmuştur (Manju & Nair, 2019).

Veri örnekleme algoritmalarında kullanılan parametreler ve kullanılan yaklaşık değerler tablo 1'de gösterilmektedir (Srivastava & Sharan, 2023).

Tablo 1. Veri Örnekleme Algoritmalarında Kullanılan Parametreler ve Kullanılan Yaklaşık Değerler

Veri Örnekleme Yöntemi	Parametreler	Değer	Tanım
SMOTE	Örneklem Stratejisi (sampling_strategy)	Otomatik	Azınlık Sınıfını Yeniden Örnekle.
	Tesadüfi durum (random_state)	42	Rastgele sayı üreteci tarafından kullanılan tohum
	k-komşular (k_Neighbours)	5	Yeni bir sentetik örnek oluşturmak için 5 En Yakın Komşuyu göz önünde bulundurun.
	N_işler (n_Jobs)	-	Çapraz doğrulama sırasında kullanılan CPU Çekirdek Sayısı.
	N_özellikler (n_features)	30	Giriş verilerindeki özellik sayısı.
ENN	Örnekleme stratejisi (sampling_strategy)	Otomatik	Çoğunluk sınıfını yeniden örnekle.
	n komşular (n_neighbours)	3	En Yakın Komşuları hesaplamak için mahalle boyutu =3'tür.
	Özellik seçimi (kind_sel)		Bir örneği hariç tutmak için komşuların Çoğunluk Oyu'nu kullanın.
	n-işler (n_jobs)	-	
SMOTEENN	Örneklem Stratejisi (sampling_strategy)	Otomatik	Çoğunluk sınıfını yeniden örnekle
	Rasgele durum	42	Rastgele sayı üreteci tarafından kullanılan tohum
	SMOTE	-	Parametrenin varsayılan değerleriyle SMOTE nesnesini kullanın.
	ENN	-	kind_sel='all' ve parametrenin varsayılan değerleriyle ENN nesnesini kullanın.
	n-işler (n-jobs)	-	Çapraz doğrulama sırasında kullanılan CPU Çekirdek Sayısı

2. Literatür Taraması

SMOTEEN, bir sınıflandırma yöntemi olmadığı ve bir veri ön işleme tekniği olduğu için doğrudan literatür taraması gerektiren bir konu gibi algılanmayabilir. Fakat yöntemin, kullanışlı olup olmadığı, nerelerde kullanılmasının ve nasıl kullanılmasının faydalı olabileceğini anlamak adına literatürde ne gibi yerlerde ve nasıl kullanıldığı ve sonuçlara olan etkisi konusunda bir bilinç geliştirebilmek için belli başlı çalışmalara burada yer verilmesi gerektiğini düşünüyoruz.

Manju & Nair (2019), kardiyak aritmilerin erken tespiti, EKG sinyalini kardiyak aritmilerden birine analiz ederek sınıflandırma problemini çözerlerken veri kümesi oldukça dengesiz olduğundan, SMOTEENN kullanılarak aşırı örnekleme ve yetersiz örneklemenin bir kombinasyonunu uyguladıklarını belirtmişlerdir. Bu çalışmada sınıflandırıcı yöntemlerin SMOTEENN kullanmadan önceki ve kullandıktan sonraki performansları Tablo 2'de ve yine aynı sınıflandırıcıların sonuç performanslarına SMOTEENN dışındaki veri dengeleme ya da veri ön işleme yöntemleri diyebileceğimiz yöntemlerin etkisini mukayeseli bir şekilde Tablo 3'de görülmektedir.

Tablo 2. Yöntemlerin SMOTEENN Kullanmadan Önceki ve Kullandıktan Sonraki Performansları

Sınıflandırıcı	SMOTEENN Kullanmadan (%)	SMOTEENN Kullanarak (%)
KA	60,82	93,67
KYK	59,39	92,74
DVM	68,42	97,35
RO	59,93	93,16

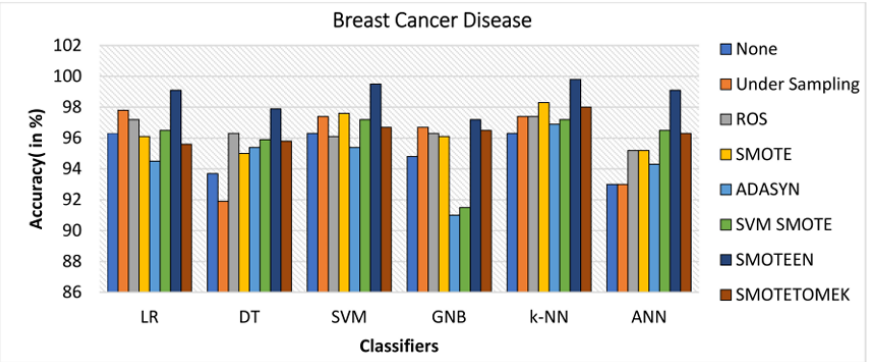
Tablo 3. Yöntemlerin Veri Ön İşleme Yöntemleri Sonrasındaki Sınıflandırma Performansları

Sınıflandırıcı	Tomek Links	SMOTE ENN	Cluster Centroit	Near Miss
KA	67,2	93,67	46	33
KYK	72	92,7	40	26
DVM	74	97,3	42	45
RO	70	93,16	33	13

Li & Wu (2022), çalışmalarında, bankacılık sektöründe, müşterilerin vadeli mevduatlara olan aboneliklerinin doğru bir şekilde tahmin edilmesinin, banka verimliliğini artırmak için hedeflenen müşteri hizmetlerine harcanan gereksiz zaman ve enerjiyi azaltabileceğini belirtmişlerdir. Geleneksel tahmin yöntemlerinin dengesiz veri sorununu çok iyi ele alamadığını, dengesiz verilerden kaynaklanan etkileri en aza indirmek için, verileri mümkün olduğunca dengeli hale getirmek gerektiğini ve bu amaçla SMOTEENN kullanmak suretiyle XGBoost yöntemiyle elde ettikleri sonuçların en iyi sonuçlar olduğunu belirtmişlerdir.

Kumar & ark. (2022), çalışmalarında beş dengesiz klinik veri kümesi: Meme Kanseri Hastalığı, Koroner Kalp Hastalığı, Hint Karaciğer Hastası, Pima Kızılderilileri Diyabet Veri tabanı ve Koroner Böbrek Hastalığı üzerinde, KA, K-En Yakın Komşu (KYK), LR, YSA, DVM ve Gauss (NB) olmak üzere altı sınıflandırıcının ampirik performans değerlendirmesini, Alt Örnekleme, Rastgele Aşırı Örnekleme, SMOTE, ADASYN, SVM-SMOTE, SMOTEEN ve SMOTETOMEK olmak üzere yedi farklı sınıf dengeleme tekniği ile dengelenmiş ve sonuç olarak SMOTEEN dengeleme yönteminin genellikle tüm diğer altı veri dengeleme tekniğine kıyasla tüm altı sınıflandırıcı ve tüm beş klinik veri kümesi

için daha iyi performans elde ettiklerini belirtmişlerdir. Diğer altı dengeleme tekniğinin neredeyse eşit ancak SMOTEEN'den orta derecede daha düşük performans gösterdiğini eklemişlerdir. Bu sonuçlar veri kümelerinden birisi olan Meme Kanseri adına Şekil 3'te gösterilmektedir.



Şekil 3. Meme Kanseri Verisine Uygulanan Veri Dengeleme Yöntemlerinin Performansları

Mbunge & ark. (2023), yaptıkları çalışmada, ishal tedavisi süreçlerine makine öğrenmesi yöntemlerini entegre etmeye çalışmışlar ve bunun için ayrıca makine öğrenmesi yöntemlerine SMOTEEN sınıf dengesizliği yaklaşımlarını uyguladıklarını ve bu yöntem ve türevlerinin makine öğrenmesi yöntemlerinin performansını önemli ölçüde artırdığını belirtmişlerdir.

Wee & ark. (2023), makine öğrenimi tabanlı diyabet sınıflandırma modellerinde önerilen bir SMOTEENN tabanlı tek değişkenli özellik seçimi yöntemi önerilmiş, daha düşük boyutlu girdiyle sınıflandırma oranını iyileştirmek için SMOTEENN aşırı örnekleme ve tek değişkenli özellik seçiminin avantajlarını birleştirilmiş ve sonuçta bu yöntemin yüksek sınıflandırma

doğruluğu elde etmede yararlı olduğu aynı zamanda sınıflandırma görevleri için gereken hesaplama maliyetini ve süreyi de azalttığını raporlamışlardır.

Srivastava & Sharan (2023), SMOTEENN'i sırasıyla Aşırı Gradyan Artırıcı (eXtreme Gradient Boosting - XGB), Rasgele Orman (RO)(Random Forest – RF), Lojistik Regresyon (LR) ve Destek Vektör Makineleri (DVM) (Support Vector Machine – SVM) sınıflandırıcılarıyla birleştirerek SMOTEENN-XGB, SMOTEENN-RF, SMOTEENN-LR ve SMOTEENN-SVM adlı dört yeni Kimlik Avı Web Sitesi Sınıflandırma modeli önerdiklerini belirtmişlerdir. Sonuç olarak veri dengeleme için önerdikleri yeni SMOTEENN hibrit yeniden örnekleme yönteminin diğer üç yaklaşımdan daha iyi performans gösterdiğini, ek olarak önerdikleri modeller arasından SMOTEENN-XGB modelinin geri kalan üç modeli, yani SMOTEENN-RF, SMOTEENN-LR ve SMOTEENN-SVM'yi ROC_AUC, PR_AUC, F1, GMean ve Accuracy gibi tüm performans ölçütlerinde tutarlı bir şekilde geride bıraktığını, ardından SMOTEEN-RF, sonra SMOTEENN-LR ve SMOTEENN-SVM'nin geldiğini gösterdiklerini beyan etmişlerdir. Srivastava & Sharan, çalışmalarında SMOTEENN yönteminin XGBosst, LR, RO ve DVM yöntemlerinin hepsinde ve tüm mukayese kriterlerinde (ROC eğri altında kalan alan, PR-AUC, Geometrik Ortalama, Doğruluk ve F1 skoru) en iyi sonuçları veren veri dengeleme yöntemi olduğunu ayrıca sonuç rakamları doğrultusunda raporlamışlardır.

Zhang & ark. (2023) çalışmalarında, ağ ihlali veri dengesizliği ve tahmin doğruluğu sorunlarını ele almak için geliştirilmiş CatBoost algoritmasına dayalı bir ağ ihlali tespit modeli

önerilmiştir ve SMOTEENN örneklemesi veri dengesizliği sorununu çözmek için kullanmışlar sonuçta ise geliştirdikleri modelin optimal olduğunu belirtmişlerdir.

Ako & ark. (2024), müşteri kaybını çalıştıkları veri kümesini SMOTEENN ile dengelemişler ve uygulamanın, SMOTEENN yeniden örneklemenin kullanımının diğer şemalardan daha iyi performans gösterdiğini desteklediğini belirtmişlerdir.

Yu & ark. (2024), çalışmalarında, araştırma verisi olarak ticari bir banka tarafından sağlanan 40.000'den fazla kayıttan oluşan bir veri kümesini kullanmışlar ve bir dizi araştırma ve işlemeden sonra, SMOTEENN'i bu tekniklerle birleştirerek mükemmel araştırma sonuçları elde ettiklerini raporlamışlardır.

Literatürü özetlemek gerekirse SMOTEENN güncel bir veri dengeleme yöntemi olup incelenen çalışmaların tamamında en iyi veri dengeleme yöntemi olduğu, sonuçların açık ara daha iyi olduğu ve ayrıca zaman maliyetini azalttığı belirtilmiştir. Bu arada SVM SMOTE vb. çok fazla türevi geliştirilmiş olup bu türevler arasında yine en iyi yöntemin hibrit SMOTEENN yöntemi olduğu çeşitli çalışmalarda belirtilmektedir.

Literatürden anlaşıldığı kadarıyla yöntemin avantaj ve dezavantajları aşağıdaki gibi sıralanabilir.

3. Avantajları

- Dengesiz veri setlerini dengelemekte ve daha iyi sınıflandırıcı sonuçları elde edilebilmektedir.

- Daha az gürültülü bir veri kümesi oluşturulmasını sağlar. ENN, gürültülü ve yanlış etiketli örnekleri temizlediği için modelin doğruluğunu artırmaktadır.
- Hem azınlık sınıfı örneklerini artırır hem de veri kümesindeki potansiyel problemleri azaltan esnek bir yöntemdir.

4. Dezavantajları

- Hesaplama maliyeti yüksek olabilmektedir. SMOTE ve ENN işlemleri birleştirildiği için özellikle büyük veri setlerinde zaman alıcı olabilmektedir.
- SMOTE ile oluşturulan sentetik örnekler bazen aşırı uyuma neden olabilmektedir.
- ENN, bazı önemli örnekleri yanlışlıkla temizleyebilir ve sınıf sınırlarını bozabilir.

5. Alternatifleri

Literatürde adı geçen ve SMOTEEN yerine kullanılabilecek olan alternatif yöntemler aşağıda listelenmiştir:

- SMOTE
- ENN
- Tomek Links
- Cluster Centroit
- Near Miss
- Under Sampling
- ADASYN

- ROS
- SVM SMOTE
- SMOTETOMEK
- SMOTE-RO
- SMOTE-LR
- SMOTE-XGB

6. Sonuçlar

SMOTE aşırı örnekleme yöntemi, aşırı uyuma, ENN ise tam tersi şekilde eksik uyum neden olduğu için her iki yöntemi birleştiren ve öncelikle aşırı örnekleme ile azınlık veri kümesinin boyutunu artırıp sonrasında ENN ile aşırı örnekleri budayarak daha dengeli ve daha optimal bir veri kümesi üreten hibrit SMOTEENN yöntemi sınıflandırıcılar için oldukça kullanışlı bir yöntem olarak literatürde kullanımı giderek artmaktadır. Ancak, karmaşıklığı ve hassas veri noktalarını yanlışlıkla temizleme riski ve aşırı uyum ihtimali nedeniyle dikkatli bir şekilde uygulanmalıdır. Sınıflandırıcıların performansı için SMOTEENN yönteminin parametrelerini doğru ayarlayabilmek önem arz etmektedir.

Literatürde yapılan ve atıf sayılarından dolayı burada ele çalışmaların istisnasız tamamı, dezavantajlarını göz önünde bulundurmak kaydı ile SMOTEENN yönteminin veriyi çok iyi dengelediğini ve sınıflandırıcıların performansını artırdığını belirtmişlerdir.

SMOTEENN, çok kuvvetli bir ihtimalle, yakın gelecekte, veri ön işleme diye tabir ettiğimiz sürecin önemli bir parçasına

dönüşebilir. Çünkü hemen her uygulamada kullanılan normalizasyon teknikleriyle benzer bir misyonu paylaşmaktadır. Dolayısıyla bundan sonraki çalışmalar için rutin bir sürece dönüşebilir.

KAYNAKÇA

Ako, R. E., Aghware, F., Okpor, M. D. and Akazue, Maureen, A., Yoro, R. E., Ojugo, A. A.,Setiadi, D. R. I. M., Odiakaose, C. C. Abere, Reuben A. Emordi, F. U., and Geteloma, V. & Ejeh, P. (2024). Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost. *Journal of Computing Theories and Applications, 2* (1). 86-101. DOI: 10.62411/jcta.10562.

Kumar, V. Lalotra, G.S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., Shorfuzzaman, M., Alsufyani, A. & Uddin, M. (2022). Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Technique. *Healthcare (MDPI), 10*, 1293. 1-28. https://doi.org/10.3390/ healthcare10071293.

Li, Y. & Wu, Z. (2022). Prediction of Customers' Subscription to Time Deposits Based on SMOTEENN-XGBoost Model. *WHICEB 2022 Proceedings*. *61*. https://aisel.aisnet.org/whiceb2022/61.

Manju, B. R. & Nair, A.R. (2019). Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms. *9th International Symposium on Embedded Computing and System Design (ISED) IEEE,* Aralık 2019, India.

Mbunge, E., Sibiya, M. N., Takavarasha, S., Millham, R. C., Chemhaka, G. & Muchemwa, B. (2023). Implementation of ensemble machine learning classifiers to predict diarrhoea with SMOTEENN, SMOTE, and SMOTETomek class imbalance

approaches.*2023 Conference on Information Communications Technology and Society (ICTAS) IEEE,* Durban, South Africa. DOI: 10.1109/ICTAS56421.2023.10082744.

Srivastava, J. & Sharan, A. (2022). SMOTEEN Hybrid Sampling Based Improved Phishing Website Detection. *Communication, Networking And Broadcast Technologies, TechRxiv,* DOI: 10.36227/techrxiv. 20207765.v1.

Wee, B. F., Hwong, H. H. Sivakumar, S., Lim, K. H., Wong, W.K. & Chew, I. M. (2023). Optimizing Diabetes Classification with Support Vector Machine and SMOTEENN-based Feature Selection. 2023 International Conference on Digital Applications, Transformation & Economy (ICDATE), 14-16 Temmuz 2023, Miri, Sarawak, Malaysia.

Yu, C., Jin, Y., Xing, Q., Zhang, Ye, Guo, S. & Meng, S. (2024). Advanced User Credit Risk Prediction Model using LightGBM, XGBoost and Tabnet with SMOTEENN. *Computer Science,* 1-8. https://doi.org/10.48550/arXiv.2408.03497.

Zhang, H., Zhang, Y., Lu, P. & Wang, C. (2023). Research On Network Intrusion Detection Based On SMOTEENN And Improved Catboost Algorithm. *Proceedings Volume 12800, Sixth International Conference on Computer Information Science and Application Technology (CISAT 2023)*; 128001U. Hangzhou, China. https://doi.org/10.1117/12.3003918.

BIDGE