**İstatistikçiler Dergisi: İstatistik & Aktüerya**

Journal of Statisticians: Statistics and Actuarial Sciences

**IDIA 16**, 2023, 2, 39-56

**Geliş**/Received:05.10.2023, **Kabul**/Accepted: 04.12.2023

**Araştırma Makalesi /** Research Article

# Katkısı belirli emeklilik planlarında kayıp tutumuna göre optimal yatırım stratejisinin belirlenmesi

Murat Kırkağaç*
*Kütahya Dumlupınar Üniversitesi, Kütahya Uygulamalı Bilimler Fakültesi, Sigortacılık ve Risk Yönetimi Bölümü*
*43100-Merkez, Kütahya, Türkiye*
murat.kirkagac@dpu.edu.tr
ORCID:0000-0002-2703-8768

Yasemin Saykan
*Hacettepe Üniversitesi, Fen Fakültesi, Aktüerya Bilimleri Bölümü*
*06800-Beytepe, Ankara, Türkiye*
yasemins@hacettepe.edu.tr
ORCID:0000-0002-8916-8509

**Öz**

Dünyada olduğu gibi ülkemizde de son yıllarda faydası belirli emeklilik planlarından katkısı belirli emeklilik planlarına geçiş oldukça yaygınlaşmıştır. Katkısı belirli emeklilik planlarında yatırım riski katılımcı üzerinde olduğu için optimal yatırım stratejisinin belirlenmesi oldukça önemlidir. Katkısı belirli emeklilik planlarında optimal yatırım stratejisinin belirlendiği çalışmalarda genellikle, klasik bir yaklaşım olan beklenen faydanın maksimizasyonu kullanılmıştır. Fakat beklenen faydanın maksimizasyonu gerçek dünyayı, özellikle birey kayıptan kaçınan bir birey olduğunda iyi yansıtmamaktadır. Bununla birlikte, yatırımcıların çoğu da aslında kayıptan kaçınan bireylerdir. Bu nedenle katkısı belirli emeklilik planlarında optimal yatırım stratejisinin kayıptan kaçınan bireyler için belirlenmesi oldukça önemlidir. Hedeflenen fon ile gerçekleşen fon büyüklüğü arasındaki farkın minimizasyonuna dayanan bir diğer yöntem ise, dönem sonu hedef fon büyüklüğü ve ara dönem fon hedeflerini belirleyerek, hedeflenen fon büyüklüğü ile gerçekleşen fon büyüklüğü arasındaki farkın karesi olarak tanımlanan maliyet fonksiyonlarının iskontolu toplamını minimize edecek şekilde optimal yatırım stratejisinin belirlenmesidir. Bu çalışmada kayıptan kaçınan bireyler için elde edilen sonuçlar, maliyet fonksiyonunun kullanıldığı modelden elde edilen sonuçlar ile karşılaştırmalı olarak elde edilmiştir. Optimal yatırım stratejisi belirlenirken her iki modelde de dinamik programlama yöntemi kullanılmıştır. Her iki modelde elde sonuçlar incelendiğinde ise sonuçların birbirine çok yakın gerçekleştiği görülmüştür. Optimal yatırım stratejisi birikim döneminin başında fonun tamamının riskli yatırım aracında değerlendirilmesi, birikim döneminin ilerleyen yaşlarında fonun riskli yatırım aracında değerlendirilen oranının azaltılarak, risksiz yatırım aracında değerlendirilen oranının artırılması, birikim döneminin sonunda ise fonun büyük bir kısmının risksiz yatırım aracında değerlendirilmesi biçimindedir. Bununla birlikte kayıptan kaçınan bireyin daha uzun bir süre, daha az risk alarak daha tutucu bir yatırım stratejisi izlediği görülmektedir.

***Anahtar sözcükler:*** Katkısı belirli emeklilik planı, Kayıptan kaçınma, Maliyet fonksiyonu, Optimal yatırım stratejisi, Dinamik programlama.

* Bu çalışma, birinci yazarın, ikinci yazarın danışmanlığında hazırladığı doktora tezinden üretilmiştir.

**Abstract**

***Determining the optimal investment strategy according to the loss attitude in defined contribution pension plans***

*In recent years, as in the rest of the world, the transition from defined benefit pension plans to defined contribution pension plans has become quite common in our country as well. Because the investment risk is on the participant, it is very important to determine the optimal investment strategy in defined contribution pension plans. Studies that determine the optimal investment strategy in defined contribution pension plans generally use the classical approach of maximizing expected utility. However, maximizing expected utility does not reflect the real world well, especially when the individual is loss-averse. Besides, most investors are actually loss-averse. Therefore, it is crucial to determine the optimal investment strategy for loss-averse individuals in defined contribution pension plans. Another method based on minimizing the difference between the target fund and the actual fund size is to determine the optimal investment strategy by minimizing the discounted sum of cost functions defined as the square of the difference between the target fund size and the actual fund size, by specifying end-of-period target fund size and interim fund targets. In this study, the results obtained for loss-averse individuals are compared with the results obtained from the model using the cost function. Dynamic programming methods are used in both models to determine the optimal investment strategy. When the results obtained in both models were examined, it was observed that the results were very close to each other. The optimal investment strategy is to use the entire fund in a risky investment asset at the beginning of the accumulation period, to decrease the ratio of the fund used in the risk-free investment asset in the later years of the accumulation period, to increase the ratio used in the risk-free investment asset, and to use a large part of the fund in the risk-free investment asset at the end of the accumulation period. Furthermore, it is observed that a loss-averse individual follows a more conservative investment strategy over a longer period, taking less risk.*

***Keywords:*** *Dynamic programming, Defined contribution pension plan, Loss aversion, Cost function, Optimal investment strategy.,*

## 1. Giriş

Emeklilik planı, katılımcıların aktif çalışma süreleri boyunca yaptıkları katkıları uzun vadeli yatırıma yönlendirerek, emeklilik dönemlerinde çalışma dönemindeki tüketim seviyesini devam ettirebilmelerini sağlayabilecekleri bir gelir elde etmelerini sağlamayı ve bu sayede emeklilik döneminde başkalarına muhtaç olmalarını engellemeyi amaçlayan, plan katılımcılarının sahip olduğu hakları ve sorumluluklarını düzenleyen sözleşmelerdir [1].

Emeklilik planlarının dünyada en yaygın olarak kullanılan iki türü, faydası belirli emeklilik planları ve katkısı belirli emeklilik planlarıdır.

Faydası belirli emeklilik planları, katılımcının emeklilik döneminde hak kazanacağı gelirinin tam değerinin önceden belirlenememesine rağmen, çeşitli yöntemlerle hesaplanabildiği planlardır. Çoğunlukla kamu sektöründe rastlanılan bu planlarda, katılımcının yapacağı katkı oranı önceden planlanmaktadır [2].

Son yıllarda faydası belirli emeklilik planlarından katkısı belirli emeklilik planlarına geçiş oldukça yaygınlaşmış ve katkısı belirli emeklilik planları sosyal güvenlik sisteminde önemli bir rol oynamaya başlamıştır. Literatür incelendiğinde de katkısı belirli emeklilik planları üzerine çalışmaların son yıllarda oldukça arttığı görülmektedir.

Katkısı belirli emeklilik planları, katılımcı tarafından yapılacak katkı oranının daha önceden belirli olduğu emeklilik planlarıdır. Emeklilik döneminde katılımcı tarafından biriktirilen fon miktarı, birikim döneminde yapılan katkıların, katkı yapacağı sürenin yani emeklilik yaşının ve yatırım getirisinin bir fonksiyonudur [2]. Katkısı belirli emeklilik planlarında, katkıların yapıldığı birikim dönemi ve yapılan katkılar sonucu oluşan fonun emeklilik geliri olarak alındığı dağıtım dönemi olmak üzere iki dönem vardır. Faydası belirli emeklilik planlarında yatırım riski plan sponsoru tarafından üzerine alınırken, katkısı belirli emeklilik planlarında ise bu risk katılımcının üzerindedir. Dolayısıyla optimal yatırım stratejisinin belirlenmesi katılımcı için oldukça önemlidir.

Literatürde katkısı belirli emeklilik planlarında optimal yatırım stratejisine ilişkin çalışmalar incelendiğinde, bu çalışmaların çok eski yıllara dayandığı görülmektedir. Bu konuda yapılan ilk çalışmalar Samuelson [3] ve Merton'a [4], [5] ait olup, katkısı belirli emeklilik planlarında optimal yatırım stratejisinin belirlenmesine ilişkin çalışmalar günümüzde hala yapılmaktadır.

Bodie, Merton ve Samuelson [6], Cairns [7], Owadally [8] 2000'li yıllara kadar optimal yatırım stratejisi üzerine yapılan bazı diğer çalışmalardır.

Katkısı belirli emeklilik planlarında optimal yatırım stratejisi üzerine yapılan çalışmaların 2000'li yılların başında oldukça arttığı görülmektedir. Bu çalışmaların başlıcaları şunlardır: Vigna ve Haberman [9], Blake, Cairns ve Dowd [10], Haberman ve Vigna [11], Gerrard, Haberman ve Vigna [12], Cairns, Blake ve Dowd [13], Battocchio, Menoncin ve Scaillet [14], Yang ve Huang [15], Blake, Wright ve Zhang [16], Chen, Haberman ve Thomas [17].

Katkısı belirli emeklilik planlarında optimal yatırım stratejisinin belirlendiği çalışmalarda genellikle, klasik bir yaklaşım olan beklenen faydanın maksimizasyonu kullanılmıştır. Fakat Rabin ve Thailer [18] beklenen fayda kriterinin çoğu risk davranışı için uygun olmadığını belirtmişlerdir. Beklenen faydanın maksimizasyonu gerçek dünyayı, özellikle birey kayıptan kaçınan bir birey olduğunda iyi yansıtmamaktadır. Bununla birlikte yatırımcıların çoğu aslında kayıptan kaçınan bireylerdir. Bu nedenle kayıptan kaçınan bireyler için optimal yatırım stratejisinin belirlenmesi oldukça önemlidir.

Kayıptan kaçınma toplam varlığın kesin değerindeki değişimden ziyade, önceden tanımlanmış bir referans noktası veya gelire göre varlıktaki kayıp veya kazanç ile tanımlanır. Kayıptan kaçınma kavramı ilk olarak Kahneman ve Tversky [19] tarafından davranışsal finansın temel taşı olan "beklenti teorisinin" içinde tanımlanmıştır. Beklenti teorisi temel olarak bireyin davranışlarını belirleyen motivasyonun, bu davranış sonucundaki beklentiler olduğunu iddia eden teoridir. Bu teoriye göre kayıplar kazançlara göre yatırımcıları duygusal olarak daha fazla etkilemektedir, bir başka deyişle kayıp yatırımcıların gözünde kazanca göre daha önemlidir.

Katkısı belirli emeklilik planlarında kayıptan kaçınan bireyler için optimal yatırım stratejisinin belirlendiği başlıca çalışmalar, Berkelaar, Kouwenberg ve Post [20], Gomes [21], Blake, Wright ve Zhang [22]'e ait çalışmalardır. Berkelaar, Kouwenberg ve Post [20] ve Gomes [21] optimal yatırım stratejisini sürekli zamanda elde ederken, Blake, Wright ve Zhang [22] kesikli zamanda elde etmiştir. Bu çalışmalarda beklenen faydanın maksimize edilmesi yerine kayıptan kaçınan bireyler için hedeflenen fon ile gerçekleşen fon büyüklüğü arasındaki farkın minimize edilmesine olanak sağlayan beklenti teorisi kullanılmıştır. Hedeflenen fon ile gerçekleşen fon büyüklüğü arasındaki farkın minimizasyonu aslında literatürde yeni bir fikir değildir. Vigna ve Haberman [9] ve Haberman ve Vigna [11] dönem sonu hedef fon büyüklüğü ve ara dönem fon hedeflerini belirleyerek, hedeflenen fon büyüklüğü ile gerçekleşen fon büyüklüğü arasındaki farkın karesi olarak tanımlanan maliyet fonksiyonlarının iskontolu toplamını minimize edecek şekilde optimal yatırım stratejisini belirlemişlerdir. Ancak bu çalışmada negatif sapmaların yanı sıra hedeften pozitif sapmalar da aynı oranda cezalandırılmaktadır. Hedeflenen fon büyüklüğünden pozitif sapma istenilen bir durum olduğu için bu çalışmada; Blake, Wright ve Zhang [22]'de olduğu gibi sadece negatif sapmalar cezalandırılacak şekilde optimal strateji belirlenmiş, elde edilen sonuçlar maliyet fonksiyonunun kullanıldığı sonuçlar ile karşılaştırmalı olarak verilmiştir. Blake, Wright ve Zhang [22] bireyin kayıptan kaçınan bir birey olduğu düşüncesiyle, yatırımın ve gelirin stokastik olduğu durumda sadece birikim dönemini dikkate alarak optimal yatırım stratejisini belirlemişlerdir. Optimal yatırım stratejisi belirlenirken zamanın kesikli olduğu varsayılmış ve stokastik model kullanılmıştır.

Bu çalışmada yer alacak diğer bölümler şu şekilde oluşturulmuştur: İkinci bölümde maliyet fonksiyonunun kullanıldığı model tanıtılmış ve optimizasyon probleminin çözümü elde edilmiştir. Üçüncü bölümde kullanılacak stokastik model tanıtılmış, model varsayımları ve kullanılan parametre değerleri verilmiş, kurulan iki model için optimizasyon problemlerinin çözümü elde edilmiştir. Dördüncü bölümde optimal yatırım stratejisi kurulan her iki model için elde edilmiş, sonuçlar karşılaştırmalı olarak verilmiştir. Beşinci ve son bölümde ise elde edilen sonuçlar özetlendikten sonra tartışma ve öneriler ile çalışma sonlandırılmıştır.

## 2. Maliyet Fonksiyonunun Kullanıldığı Model

$\hat{r}$ finansal danışman tarafından yapılan getiri tahminini, $\ddot{s}_{T\neg}$ bu tahmin ile hesaplanan her dönem başında yapılan bir birimlik ödemenin *T* dönem sonundaki birikimli değerini, *f(T)* dönem sonunda hedeflenen fon büyüklüğünü göstermek üzere, katılımcı tarafından her dönem başında yapılacak sabit katkı miktarı *(C),*

$$C = \frac{f(T)}{\ddot{s}_{\overline{T}|\hat{r}}} \tag{2.1}$$

biçiminde hesaplanır. Fonun *t+1* anındaki değeri özyineli olarak,

$$F_{t+1} = (F_t + C)[(1-y_t)e^{\mu_t} + y_t e^{\lambda_t}] \tag{2.2}$$

eşitliğinden hesaplanır. Eşitlik 2.2'de:

$F_{t+1}$ : Fonun *t+1* anındaki değerini,
$F_t$ : Fonun *t* anındaki değerini,
$C$ : Sabit katkı miktarını,
$y_t$ : Fonun yüksek riskli yatırım aracında değerlendirilen oranını,
$(1-y_t)$ : Fonun düşük riskli yatırım aracında değerlendirilen oranını,
$\mu_t$ : [*t,t+1*] zaman aralığında sabit olan, düşük riskli yatırım aracı için anlık faiz oranını,
$\lambda_t$ : [*t,t+1*] zaman aralığında sabit olan, yüksek riskli yatırım aracı için anlık faiz oranını,

göstermektedir. Sabit katkılı bireysel emeklilik planlarında fonun %50'si düşük riskli, %50'si yüksek riskli yatırım araçlarında değerlendirildiği varsayıldığında Eşitlik 2.2,

$$F_{t+1} = (F_t + C)[\frac{e^{\mu_t} + e^{\lambda_t}}{2}] \tag{2.3}$$

biçiminde ifade edilebilir. Yatırım getirisinin tahmin edilenden farklı olması, vade sonunda gerçekleşen fon büyüklüğünün hedeflenen fon büyüklüğünden farklı olmasına yani vade sonunda açığın oluşmasına neden olur. Bu açık miktarı, hedeflenen fon büyüklüğü ile gerçekleşen fon büyüklüğü arasındaki farka eşittir:

$$D_T = f(T) - F_T \tag{2.4}$$

Katılımcı için risk, bu açığın yüksek olmasıdır. Bu açığın negatif olması ise fazlalık olarak adlandırılır. Fonda açık da fazlalık da istenmeyen durumlardır [23].

### *2.1. Optimal yatırım stratejisi*

Bireysel emeklilik planlarında hedef fon büyüklüğüne ulaşmanın diğer bir yolu sabit katkılardan oluşan fonun optimal yatırım stratejisi ile değerlendirilmesidir.

Optimal yatırım stratejisi belirlenirken, yatırım getirisinin zamanla değişmesinden kaynaklanan dinamik yapısı nedeniyle Dinamik Programlama (DP) yöntemi kullanılmıştır. DP yöntemi, Richard Ernest Bellman tarafından 1950 yılında isimlendirilmiştir. Başlangıçta yalnızca bir ekonomik sistemin zaman içindeki durumunun incelenmesinde kullanılan bu yöntem, günümüzde zamanla ilgili olan süreçlerin yanı sıra, farklı nitelikteki süreçlerin incelenmesinde de yaygın olarak kullanılmaktadır [24].

Bellman'ın Optimalite İlkesi: " Başlangıç koşulu ve başlangıç kararı ne olursa olsun geri kalan kararlar verilen ilk kararın sonucuna göre optimal bir politika oluşturmalıdırlar [25]" biçiminde tanımlanmaktadır.

DP problemlerinin çözümü, uygun bir matematiksel modelin kurulması ile başlanmaktadır. Bu bölümde, Vigna ve Haberman [18] tarafından oluşturulan Eşitlik 2.2'de verilen model kullanılmıştır.

Optimal yatırım stratejisi belirlenirken, dönem sonu fon büyüklüğü hedefinin yanı sıra ara dönem hedeflerinin de belirlenmesi gerekmektedir. t=1 anındaki hedef fon, t=0 anında Eşitlik 2.1'den hesaplanan sabit katkı miktarı C'nin bir dönem ileri çekilmesi ile bulunur:

$$f(1) = Ce^{i_A} \tag{2.5}$$

Diğer ara dönemlerde hedef fon büyüklüklerini bulmak için, hedef fon büyüklüğünün f(1)'den f(T)'ye doğrusal olarak arttığı varsayılmıştır.

Ara dönemlerde hedeflenen fon büyüklüğü ile gerçekleşen fon büyüklüğü arasındaki fark için maliyet fonksiyonu:

$$M(t) = \theta_1 D_t^2 = \theta_1 (F_t - f_t)^2 \tag{2.6}$$

ve dönem sonu için maliyet fonksiyonu:

$$M(T) = \theta_0 D_T^2 = (F_T - f_T)^2 \tag{2.7}$$

biçiminde tanımlanmıştır. Eşitlik 2.6 ve Eşitlik 2.7'de $\theta_0$ ve $\theta_1$ sabit birer katsayı olup, vade sonunda hedef fon büyüklüğüne ulaşılması ara dönem hedeflerine ulaşılmasından daha önemli olduğundan dolayı: $\theta_0$, $\theta_1$'den daha büyük seçilir.

Vade sonuna kadar her dönem oluşacak maliyetlerin t anındaki değeri:

$$G_t = \sum_{s=t}^{T} \gamma^{s-t} M(s) \tag{2.8}$$

biçiminde olup, bu eşitlikte verilen γ iskonto faktörüdür. Optimizasyonda hedeflenen fon büyüklüğü ile gerçekleşen fon büyüklüğü arasındaki farkın minimize edilmesi amaçlandığından Gt fonksiyonu minimize edilmiştir. Yapılan optimizasyon sonucunda optimal yatırım oranı:

$$y_t^* = -\frac{M_t}{2L_t} \tag{2.9}$$

olarak elde edilir. Burada:

$$L_t = P_{t+1}(F_t + C)^2 (e^{2\mu+\sigma_1^2} + e^{2\lambda+\sigma_2^2} - 2e^{\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{2}}) \tag{2.10}$$

$$M_t = P_{t+1}(F_t + C)^2 (-2e^{2\mu+\sigma_1^2} + 2e^{\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{2}}) - 2Q_{t+1}(F_t + C)(-e^{\mu+\frac{\sigma_1^2}{2}} + e^{\lambda+\frac{\sigma_2^2}{2}}) \tag{2.11}$$

biçimindedir. $L_t$ ve $M_t$ içinde bulunan $P_t$ ifadesi, $P_T=\theta$ son değerinden başlanarak;

$$P_t = 1 + P_{t+1} * [\frac{e^{2\mu+2\lambda+\sigma_1^2+\sigma_2^2}(e^{\sigma_1^2+\sigma_2^2} - 1)}{e^{2\mu+\sigma_1^2} + e^{2\lambda+\sigma_2^2} - 2e^{\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{2}}}] \tag{2.12}$$

formülü ile öz yineli olarak, $Q_t$ ifadesi, $Q_T = \theta f(T)$ son değerinden başlanarak;

$$Q_t = f(T) - [v * C * \frac{e^{2\mu+2\lambda+\sigma_1^2+\sigma_2^2}(e^{\sigma_1^2+\sigma_2^2}-1)}{e^{2\mu+\sigma_1^2}+e^{2\lambda+\sigma_2^2}-2e^{\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{2}}} * P_{t+1}] - [v * \frac{e^{\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{2}}(e^{\mu+0,5\sigma_1^2}+e^{\lambda+0,5\sigma_2^2}-e^{\mu+1,5\sigma_1^2}-e^{\lambda+1,5\sigma_1^2})}{e^{2\mu+\sigma_1^2}+e^{2\lambda+\sigma_2^2}-2e^{\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{2}}} * Q_{t+1}] \quad (2.13)$$

formülü ile özyineli olarak elde edilir [9].

## 3. Kayıptan Kaçınan Bireyler İçin Kullanılan Model

Çalışmanın bu bölümünde kayıptan kaçınan bireyler için kullanılan stokastik model tanıtılacaktır. Kullanılan modeller belirlenirken Blake, Wright ve Zhang [22] temel alınmıştır. Kullanılan modellerin varsayımları

- Fonun biri yüksek riskli diğeri risksiz olmak üzere 2 yatırım aracında değerlendirildiği,
- Katkıların yıllık olarak her yılın başında yapıldığı,
- Değerlendirmenin yıllık bazda yapıldığı ve zamanın kesikli olduğu,
- Tüm bireylerin sisteme 20 yaşında girdiği ve 65 yaşında emekli olduğu,
- Dönem sonu hedeflenen fon büyüklüğünün "2/3 yerine koyma modeli" dikkate alınarak belirlendiği

biçimindedir. Bu bölümde finansal araçların getirileri ve gelire ilişkin eşitliklere değinildikten sonra hedef fon büyüklükleri ve amaç fonksiyonu belirlenecektir.

### *3.1. Finansal Araçlar*

Yatırımın tahvil gibi biri risksiz, hisse senedi gibi biri yüksek riskli olmak üzere iki yatırım aracında değerlendirileceği varsayılsın.

*r,* risksiz yatırım aracının yıllık getirisini göstermek üzere,

*x* ile *x+1* yaşları arasında yüksek riskli yatırım aracının yıllık getirisi $R_x$:

$$R_x = r + \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma Z_x \quad (3.1)$$

biçimindedir. Burada $\mu$ yüksek riskli yatırım aracının yıllık risk primini, $\sigma$ standart sapmasını göstermekte olup $\{Z_x\}$ Standart Normal dağılıma sahip bağımsız raslantı değişkenleridir.

### *3.2. Gelir*

Bu çalışmada gelir için Cairns, Blake ve Dowd [13] tarafından oluşturulan model stokastik şok bileşenleri olmadan kullanılmıştır. $l_x$ gelirdeki büyüme oranını göstermekte olup:

$$l_x = \eta + \frac{S_x - S_{x-1}}{S_{x-1}} \quad (3.2)$$

biçimindedir. Burada $\eta$ ulusal kazancın ortalama yıllık büyüme oranını, *Sx, x* yaşındaki genel maaş profilini göstermekte olup:

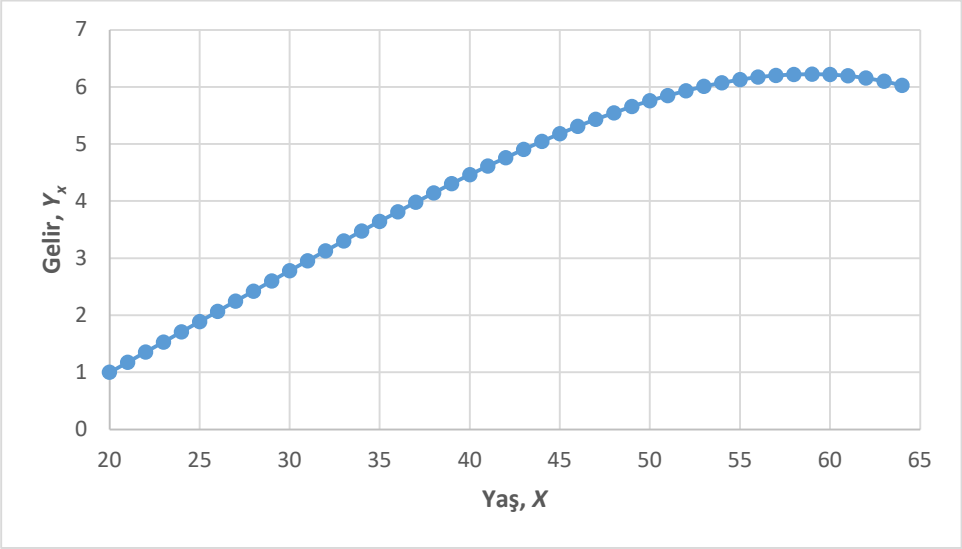
$$S_x = 1 + h_1\left[-1 + \frac{(x-20)}{45}\right] + h_2\left[-1 + \frac{4(x-20)}{45} - \left\{\frac{\sqrt{3}(x-20)}{45}\right\}^2\right]; x = 20,21,\ldots,65 \quad (3.3)$$

biçimindedir. Bu eşitlikte $h_1$=-0,1865, $h_2$=0,7537 olarak alınmıştır. Bu parametre tahminleri Cairns, Blake ve Dowd [13] tarafından, İngiltere'de erkeklere ilişkin 2005 yılına ait veriden, en küçük kareler yöntemi kullanılarak elde edilen değerlerdir.

Dolayısıyla, bireyin $x$ yaşında beklenen geliri $Y_{20}=1$ değerinden başlayarak iterasyon yoluyla:

$$Yx = Y_{x-1}exp(lx) ; \quad x = 21,22,\ldots,65 \tag{3.4}$$

biçiminde elde edilir. İşe başlama yaşı olan 20 yaştan, emeklilik yaşı olan 65 yaşa kadar beklenen gelirdeki değişim Şekil 3.1'de gösterilmektedir:




**Şekil 3.1:** Beklenen gelirin yaşa bağlı değişimi

Şekil 3.1'den beklenen gelirin birikim döneminin başı olan 20 yaştan itibaren doğrusal bir biçimde arttığı, emekliliğe yaklaştığında ise artış hızının azaldığı görülmekte olup, bu durum katılımcıların genel maaş profiline uygundur.

### *3.3. Emeklilik Fonunun Birikimi ve Hedefler*

Bu bölümde birikim dönemindeki herhangi bir zamanda gerçekleşen ve hedeflenen fon büyüklüklerinin nasıl elde edildiği verilmiştir. $\theta_{x-1}$ *x-1* yaşında fonun yüksek riskli yatırım aracında değerlendirilen oranını göstermek üzere $x$ yaşında fonun getirisi

$$Fon\ Getirisi = exp[\theta_{x-1}R_x + (1-\theta_{x-1})r]$$

$$= exp\left[\theta_{x-1}r + \theta_{x-1}\left(\mu - \frac{1}{2}\sigma^2\right) + \theta_{x-1}\sigma Z_x + r - \theta_{x-1}r\right]$$

$$=exp\left[r + \theta_{x-1}\left(\left(\mu - \tfrac{1}{2}\sigma^2\right) + \sigma Z_x\right)\right] \tag{3.5}$$

biçimindedir. $Y_x$, *x-1* yaşındaki geliri, π sabit katkı oranını göstermek üzere, fonun $x$ yaşında gerçekleşen değeri olan $F_x$, $F_{20}=0$ değerinden itibaren birikimli olarak

$$F_x = (F_{x-1} + \pi_x Y_{x-1})\exp\left[r + \theta_{x-1}\left(\left(\mu - \tfrac{1}{2}\sigma^2\right) + \sigma Z_x\right)\right]; x = 21,22,\ldots,65 \tag{3.6}$$

biçiminde elde edilir. Gerçekleşen fon büyüklüğünün yanı sıra, hedeflenen fon büyüklüğünün de belirlenmesi gerekmektedir. Dönem sonu hedeflenen fon büyüklüğü "2/3 yerine koyma modeli" kullanıldığında,

$$f(65) = \frac{2}{3} * Y_{65} * \ddot{a}_{65} \tag{3.7}$$

biçiminde elde edilir. Ara dönem hedeflenen fon büyüklükleri dönem sonu hedeflenen fon büyüklüğü *f(65)* değeri kullanılarak geriye doğru:

$$f(x) = \frac{f(x+1)}{\exp(r^*)} - \pi * Y_x;\ x = 64{,}63, \dots 20 \tag{3.8}$$

biçiminde elde edilir. Burada *r** ara dönem fon hedefleri belirlenirken kullanılan iskonto oranını göstermekte olup, risksiz yatırım aracının getirisine bağlı olarak $r^* = r + 0{,}011$ biçiminde alınmıştır [22].

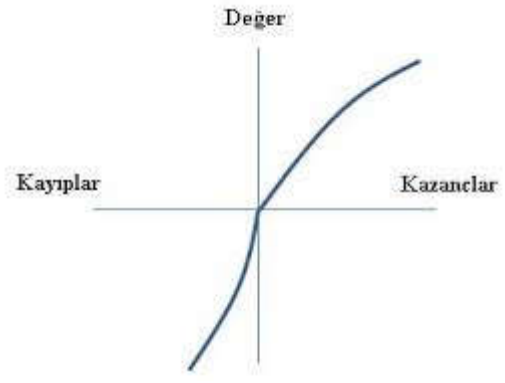
### *3.4. Amaç Fonksiyonunun Belirlenmesi*

Beklenti teorisi kullanıldığında işe giriş yaşından emekliliğe kadar her yaş için fayda fonksiyonu:

$$U_x(F_x) = \frac{(F_x - f(x))^{v_1}}{v_1};\ F_x \geq f(x)$$

$$= -\lambda \frac{(f(x) - F_x)^{v_2}}{v_2};\ F_x < f(x) \tag{3.9}$$

biçimindedir. Burada $F_x$ gerçekleşen fon büyüklüğünü, *f(x)* hedeflenen fon büyüklüğünü, $v_1$ kazanç için eğim parametresini, $v_2$ kayıp için eğim parametresini, $\lambda$ kayıptan kaçınma oranını göstermektedir. $\lambda$ kayıptan kaçınma oranı bireyin kayba karşı kazanca olduğundan kaç kat daha duyarlı olduğunu göstermekte olup $\lambda > 1$ olarak alınır. $v_1$ ve $v_2$ eğim parametreleri ise (0,1) aralığında değerler almaktadır. $v_1$ ve $v_2$'nin bu aralıkta değer alması bireyin fonda kayıp olması durumunda risk arayışında olduğunu, fonda kazanç olması durumunda da riskten kaçındığını göstermektedir. Bu durum, standart fayda fonksiyonlarındaki içbükey şeklin aksine fayda fonksiyonunun kayıp durumunda içbükey, kazanç durumunda dışbükey olarak s şeklinde olmasına yol açmaktadır [26]. $\lambda > 1,\ 0 < v_1 < 1$ ve $0 < v_2 < 1$ durumlarında beklenti teorisi fayda fonksiyonunu genel şekli Şekil 3.2'deki gibidir:




**Şekil 3.2:** Beklenti teorisi fayda fonksiyonu

Katılımcı açısından dönem sonu fon hedefine ulaşmak, ara dönem fon hedeflerine ulaşmaktan daha önemli olduğu için, ara dönem fon hedeflerine $w < 1$ olacak şekilde daha düşük bir ağırlık verilir. Dolayısıyla *x* yaşından emekliliğe kadar iskonto edilmiş fayda fonksiyonu toplamı

$$V_x = [\sum_{s=0}^{65-x-1} \beta^s w U_{x+s}(F_{x+s})] + \beta^{65-x} U_{65}(F_{65}) = w U_x(F_x) + \beta V_{x+1} \tag{3.10}$$

biçiminde belirlenir. Burada $\beta$ iskonto faktörüdür.

*3.5. Optimizasyon*

Fonun yüksek riskli yatırım aracında değerlendirilen optimal yatırım oranı $\theta_x$ bu toplamın beklenen değeri maksimize edilerek belirlenir. Optimizasyon problemi

$$\max_{\theta_x} E_x(V_x) = \max_{\theta_x} E_x[\{\sum_{s=0}^{65-x-1} \beta^s w U_{x+s}(F_{x+s})\} + \beta^{65-x} U_{65}(F_{65})] \quad (3.11)$$

biçimindedir. Bu optimizasyon problemine ilişkin kısıtlayıcı koşullar:

- $F_x = (F_{x-1} + \pi \quad Y_{x-1}) \exp\left[r + \theta_{x-1}\left(\left(\mu - \frac{1}{2}\sigma^2\right) + \sigma Z_x\right)\right] \geq 0; x = 21,22,\ldots,65$
- *Yx=Yx-1exp(lx)*
- $0 \leq \theta_x \leq 1$ (3.12)

biçimindedir.

*3.6. Optimal Yatırım Stratejisinin Elde Edilmesi*

Verilen optimizasyon probleminin çözümü için Eş. 3.11'de verilen $E_x(V_x)$ beklenen değerini maksimize eden $\theta_x$'lerin birikim döneminin başlangıcı olan 20 yaştan, birikim döneminin sonu olan 64 yaşa kadar her bir yaş için elde edilmesi gerekmektedir. Bu optimizasyon probleminin çözümünde, optimal yatırım stratejisinin her bir yaşta ayrı ayrı belirlenmesi ve problemin özyineli olarak birbirleriyle bağlantılı alt problemlere ayrışan yapısı nedeniyle dinamik programlama yöntemi kullanılmıştır.

Dinamik Programlama yönteminde problemler çözülürken ileriye ve geriye doğru yineleme yöntemleri kullanılabilir. Geriye doğru yineleme yöntemi ile öncelikle $E_{64}(V_{64})$ beklenen değerini maksimize eden $\theta_{64}$ değeri, sonra özyineli olarak 20 yaşa kadar $\theta_x$ değerleri elde edilmiştir. *x=64* yaş için:

$$\max_{\theta_{64}} E_{64}(V_{64}) = \max_{\theta_{64}} E_{64}[\{\sum_{s=0}^{0} \beta^s w U_{64+s}(F_{64+s})\} + \beta U_{65}(F_{65})]$$

$$\max_{\theta_{64}} E_{64}(V_{64}) = \max_{\theta_{64}} E_{64}[wU_{64}(F_{64}) + \beta U_{65}(F_{65})] \quad (3.13)$$

biçiminde yazılır. Burada fayda fonksiyonları:

$$U_{64}(F_{64}) = \frac{(F_{64} - f(64))^{v_1}}{v_1}; F_{64} \geq f(64)$$

$$= -\lambda \frac{(f(64) - F_{64})^{v_2}}{v_2}; F_{64} < f(64) \quad (3.14)$$

$$U_{65}(F_{65}) = \frac{(F_{65} - f(65))^{v_1}}{v_1}; F_{65} \geq f(65)$$

$$= -\lambda \frac{(f(65) - F_{65})^{v_2}}{v_2}; F_{65} < f(65) \quad (3.15)$$

biçiminde olup, $v_1$, $v_2$ *ve* $\lambda$ değerleri bilinen kayıptan kaçınma parametreleridir. Fayda fonksiyonunun ve diğer fonksiyonların değerleri elde edilirken kullanılan parametre değerleri Blake, Wright ve Zhang (2013)'de kullanılan değerler olup bu değerler ve Çizelge 3.1'de verilmektedir.

**Çizelge 3.1:** Parametre değerleri

| **Kayıptan Kaçınma Parametreleri** | | **Gelir Parametreleri** | |
|---|---|---|---|
| Kayıptan kaçınma oranı $\lambda$ | 4,50 | $r_1$ | 0,02 |
| Kazanç için eğim parametresi $v_1$ | 0,44 | $\sigma_1$ | 0,05 |
| Kayıp için eğim parametresi $v_2$ | 0,88 | $h_1$ | -0,1865 |
| | | $h_2$ | 0,7537 |
| **Varlık Getirileri** | | **Diğer Parametreler** | |
| Risksiz getiri oranı $r$ | 0,02 | Katkı oranı $\pi$ | 0,15 |
| Riskli yatırım aracının yıllık risk primi $\mu$ | 0,04 | Ara hedefler için ağırlık $w$ | 0,5 |
| Riskli yatırım aracının volatilitesi $\sigma$ | 0,18 | c | 0,13 |
| İskonto faktörü $\beta$ | 0,96 | h | 0,55 |
| | | k | 0,29 |

Dönem sonu hedeflenen fon büyüklüğü değerleri olan *f(65)* değeri, PMA92 (1992 yılı Emekli Erkekler için Mortalite Tablosu) kullanılarak Eş. 3.7'den *f(65)*=62,66, ara dönem hedeflenen fon büyüklüklerinden biri olan *f(64)* değeri ise Eş. 3.8'den *f(64)*=59,84 olarak elde edilir. Bilinmeyen değerler $F_{64}$ *ve* $F_{65}$ değerleridir. Bu değerler Çizelge 3.1'de verilen parametre değerlerinin Eş 3.6'da yerine konulmasıyla;

$$F_{65} = [F_{64} + (0{,}15 * 6{,}0254)]\, exp[0{,}02 + \theta_{64}(0{,}0238 + 0{,}18 * Z_x)]$$

$$F_{64} = [F_{63} + (0{,}15 * 6{,}0983)]\, exp[0{,}02 + \theta_{63}(0{,}0238 + 0{,}18 * Z_x)] \qquad (3.16)$$

olarak elde edilmiştir. Dolayısıyla fayda fonksiyonları da Eş. 3.9'dan

$$U_{64}(F_{64}) = \frac{(F_{64} - 59{,}84)^{0{,}44}}{0{,}44}; F_{64} \geq 59{,}84$$

$$= -4{,}5 * \frac{(59{,}84 - F_{64})^{0{,}88}}{0{,}88}; F_{64} < 59{,}84$$

$$U_{65}(F_{65}) = \frac{\{(F_{64} + 0{,}90)\exp[0{,}02 + \theta_{64}(0{,}0238 + 0{,}18 * Z_x)] - 62{,}66\}^{0{,}44}}{0{,}44}; F_{65} \geq 62{,}66$$

$$= -4{,}5 * \frac{\{62{,}66 - (F_{64} + 0{,}90)\exp[0{,}02 + \theta_{64}(0{,}0238 + 0{,}18 * Z_x)]\}^{0{,}88}}{0{,}88}; F_{65} < 62{,}66 \qquad (3.17)$$

biçiminde elde edilmiştir. Bu problemin stokastik dinamik programla yöntemi ile çözülebilmesi için beklenen değer ifadesinin kesikli hale getirilmesi gerekmektedir. Bu amaçla 64 yaşında gerçekleşen fon büyüklüğü olan $F_{64}$ değeri olası değer aralığı olan [0,200] arasında 201 parçaya bölünerek her bir değer için sonuçlar elde edilir. $F_{64}$ değeri için [0,200] aralığındaki değerler kullanıldığında $U_{64}(F_{64})$ değeri her bir aralık için sabit bir sayıya eşit olacağından Eş. 3.13 ile verilen fonksiyonda beklenen değer dışına sabit olarak çıkacaktır. Bu durumda beklenen değeri alınması gereken tek ifade $U_{65}(F_{65})$ değeri olur. Yani Eş. 3.13'teki beklenen değer ifadesi sabit sayıların dışarı alınması ile

$$\max_{\theta_{64}} E_{64}[U_{65}(F_{65})]$$

biçiminde sadeleşir. Buradaki beklenen değer

$$E_{64}[U_{65}(F_{65})] = \int_{-\infty}^{\infty} U_{65}(F_{65}) f(z) dz$$

biçimindedir. Bu beklenen değer standart normal dağılımın olasılık yoğunluk fonksiyonunu içerdiği için açık çözümü yoktur. Bu amaçla integralin yaklaşık değerinin elde edilmesi için Sayısal İntegrasyon yöntemlerden olan, Gauss-Hermite Kareleştirme yöntemi kullanılmıştır.

$U_{65}(F_{65})$ fayda fonksiyonunun açık hali yazıldıktan sonra, işlem kolaylığı için $U_{65}(F_{65})$ fayda fonksiyonundaki sabitler beklenen değer dışına çıkarılır ve beklenen değer içindeki ifade $g(z)$ fonksiyonu olarak tanımlanırsa birinci aralık ($F_x \geq f(x)$) için

$$g_1(z) = [(F_{64} + 0{,}90) exp(0{,}02 + 0{,}0238\theta_{64} + 0{,}18\theta_{64} Z_x) - 62{,}66]^{0{,}44} \quad (3.18)$$

ikinci aralık ($F_x < f(x)$) için ise

$$g_2(z) = [62{,}66 - (F_{64} + 0{,}90) exp(0{,}02 + 0{,}0238\theta_{64} + 0{,}18\theta_{64} Z_x)]^{0{,}88} \quad (3.19)$$

biçiminde elde edilir.

Eş 3.18 ve Eş 3.19'un beklenen değeri

$$E_{64}[g_j(z)] = \int_{-\infty}^{\infty} g_j(z) f(z) dz = \int_{-\infty}^{\infty} g_j(z) \frac{1}{\sqrt{2\Pi}} \exp\left(-\frac{z^2}{2}\right) dz \; ; j = 1{,}2$$

biçimindedir. $\frac{z}{\sqrt{2}} = x$ dönüşümü yapılırsa

$$E_{64}[g_j(z)] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\Pi}} \exp(-x^2)\, g_j(\sqrt{2}x) dx; j = 1{,}2$$

biçiminde yazılır. Gauss-Hermite yönteminde $\int_{-\infty}^{\infty} f(x) \exp(-x^2) \approx \sum_{i=1}^{n} w_i f(x_i)$ olduğundan

$$E_{64}[g_j(z)] \approx \frac{1}{\sqrt{\Pi}} \sum_{i=1}^{n} w_i\, g_j(\sqrt{2}x_i); i = 1{,}2, \dots, n, j = 1{,}2$$

olarak elde edilir. Burada düğüm sayısının $(n)$ belirlenmesi gerekmektedir. Düğüm sayısı ne kadar büyük alınırsa sonuca daha fazla yaklaşılmakla birlikte işlem yükü de artmaktadır. Bu nedenle düğüm sayısı 2'den başlanarak 9'a kadar alınıp işlemler tekrarlanmıştır. Düğüm sayısı 7 ve 9 alındığında elde edilen sonuçların birbirine çok yakın olduğu görüldüğünden dolayı düğüm sayısı 7 olarak belirlenmiştir.

Düğüm sayısı $n=7$ için Gauss-Hermite katsayıları ve ağırlıkları:

$x_{1,2} = \pm 2{,}6520$, $x_{3,4} = \pm 1{,}6735$, $x_{5,6} = \pm 0{,}8162$, $x_7 = 0$

$w_{1,2} = 9{,}7178$, $w_{3,4} = 0{,}0545$, $w_{5,6} = 0{,}4256$, $w_7 = 0{,}8103$

olduğundan beklenen değer ifadesi:

$$E_{64}[g_j(z)] \approx \frac{1}{\sqrt{\Pi}} \sum_{i=1}^{7} w_i\, g_j(\sqrt{2}x_i); i = 1{,}2, \dots, 7; j = 1{,}2$$

$$E_{64}\left[g_j(z)\right] \approx \frac{1}{\sqrt{\Pi}}\Big[9{,}7178\left(g_j(3{,}75)+g_j(-3{,}75)\right)+0{,}0545\left(g_j(2{,}37)+g_j(-2{,}37)\right)+$$
$$0{,}4256\left(g_j(1{,}15)+g_j(-1{,}15)\right)+0{,}8103g_j(0)\Big]; j=1{,}2 \qquad (3.20)$$

biçimindedir. $g_j(z)$; $j=1,2$ fonksiyonunun değeri Eş 3.18 ve Eş 3.19'dan elde edildikten sonra Eş 3.20'de yerine konulduğunda $E_{64}\left[g_j(z)\right]; j=1{,}2$ sadece $F_{64}$ ve $Q_{64}$'e bağlı bir fonksiyona dönüşür. Optimizasyon problemindeki amaç Eş 3.20 ile verilen beklenen değeri maksimize eden $\theta_{64}$'ü bulmak olduğu için gerçekleşen fon büyüklüğü $F_{64}$, olası değer aralığı olan [0,200] arasında 201 parçaya bölünerek her bir değer için sonuçlar elde edilir. $F_{64}$ değeri için [0,200] aralığındaki değerler kullanıldığında $U_{64}(F_{64})$ değeri her bir aralık için sabit bir sayıya eşit olacak $U_{65}(F_{65})$ değeri ise sadece $\theta_{64}$'e bağlı olacaktır. Böylelikle amaç fonksiyonu sadece $\theta_{64}$'e bağlı bir fonksiyon olacağından kolaylıkla maksimize edilebilir.

[0,200] aralığındaki tüm $F_{64}$ değerleri için amaç fonksiyonunu maksimize eden $\theta_{64}$ değerleri elde edildikten sonra bu $\theta_{64}$ değerleri kullanılarak $E_{63}(V_{63})$ değerini maksimize eden $\theta_{63}$ değerleri bulunur. Bu işlem özyineli olarak 20 yaşa kadar devam ederek en son $E_{20}(V_{20})$ değerini maksimize eden $\theta_{20}$ değerleri elde edilir.

## 4. Bulgular

### *4.1. Maliyet Fonksiyonunun Kullanıldığı Model İçin Optimal Yatırım Stratejisi*

Bu uygulamada, 100.000 tekrarlı benzetim çalışması ile sabit ve değişken katkılı bireysel emeklilik planları ile optimal yatırım stratejisinin kullanıldığı bireysel emeklilik planları sonucunda oluşan dönem sonu açık miktarları karşılaştırılarak; hangi planın daha riskli olduğuna, hangi planın getiriye ilişkin yapılan tahmine karşı daha duyarlı olduğuna karar verilecek, hangi planda hedeflenen fon büyüklüğüne daha çok yaklaşıldığı belirlenecektir.

$\hat{r}$ getiri tahminini göstermek üzere, kayıptan kaçınma parametreleri olan $\mu$=0,02, $\sigma_1$ =0 ve $\lambda$=0,04, $\sigma_2$=0,18 değerleri için getiri tahmini yaklaşık

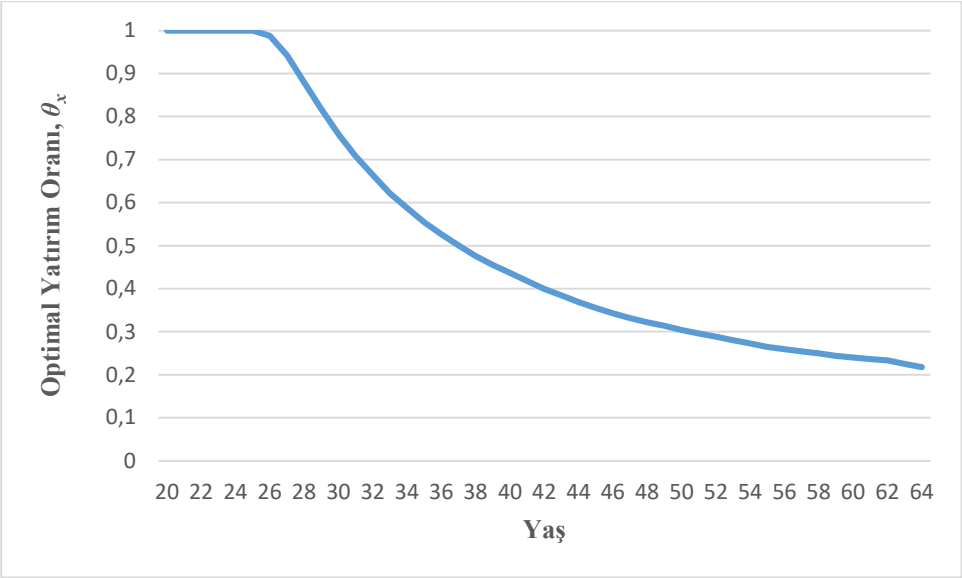
$$\hat{r} = \exp 0{,}5\left(\mu+\lambda+\frac{\sigma_1^2+\sigma_2^2}{4}\right) = \exp 0{,}5\ (0{,}02+0{,}04+\frac{0+0{,}18^2}{4}) \approx 0{,}03$$

olarak elde edilir. Finansal danışman vadenin başında fonun yıllık ortalama getirisine ilişkin bir tahmin yapmalıdır. Yapılan bu tahminin tam olarak gerçekleşmesi az rastlanan bir durumdur. Gerçekleşecek olan yatırım getirisinin üstünde veya altında bir tahminde bulunulabilir. Gerçekleşecek olan yatırım getirisi ile tahmin edilen yatırım getirisi arasındaki fark "yatırım getirisindeki tahmin hatası" olarak adlandırılmaktadır.

Yatırım getirisindeki tahmin hatasının 0 olduğu, yani yatırım getirisinin doğru tahmin edildiği durumda, yıllar itibariyle oluşan optimal yatırım oranları Çizelge 4.1 ve Şekil 4.1'de yer almaktadır.

**Çizelge 4.1:** Maliyet Fonksiyonun Kullanıldığı Model İçin Optimal Yatırım Oranları

| Yaş | Optimal Yatırım Oranı |
|---|---|
| 20 | 1 |
| 25 | 0,9998 |
| 30 | 0,7583 |
| 35 | 0,5543 |
| 40 | 0,4367 |
| 45 | 0,3553 |
| 50 | 0,3039 |
| 55 | 0,2649 |
| 60 | 0,2398 |
| 64 | 0,2178 |

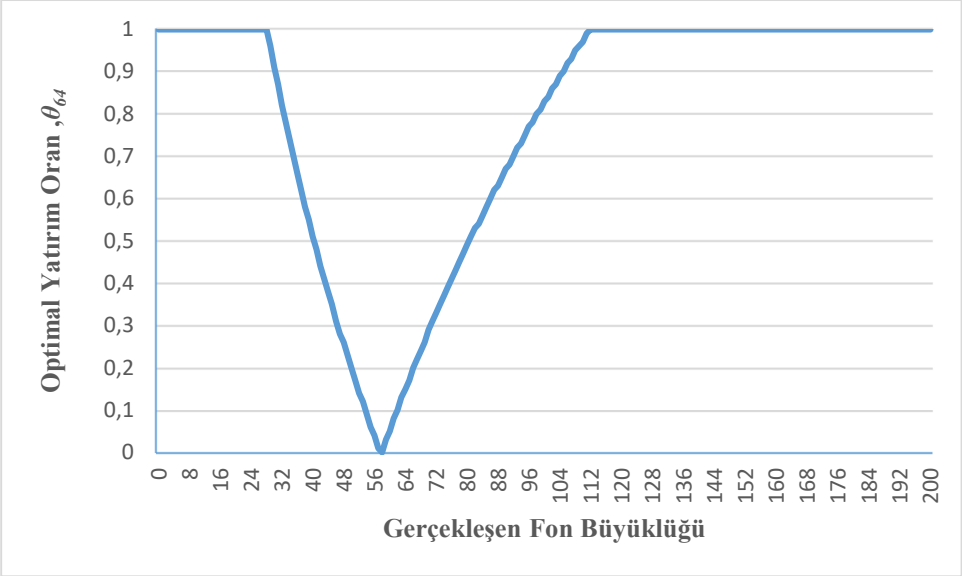



**Şekil 4.1:** Maliyet Fonksiyonun Kullanıldığı Model İçin Optimal Yatırım Stratejisi

Çizelge 4.1 ve Şekil 4.1'den ulaşılan bir diğer önemli sonuç fonun yüksek riskli yatırım aracında değerlendirilen oranı $y_t^*$'ın vadenin başında 1 değerini aldığı ve yıllar içinde azalarak vade sonunda yaklaşık 0,2332 değerine düştüğü, dolayısıyla fonun düşük riskli yatırım aracında değerlendirilen oranı $(1-y_t^*)$'ın vade başında 0 değerini aldığı ve yıllar içinde artarak vade sonunda yaklaşık 0,7668 değerine ulaştığıdır. Bu yatırım stratejisi aslında tüm emeklilik planlarında kabul edilmiş ve tavsiye edilen bir yatırım stratejisidir. Bu stratejiye göre birikim döneminin başlarında fonun büyük bir kısmı hisse senedi gibi yüksek riskli ve getirili yatırım araçlarında değerlendirilmeli, vade ilerledikçe fonun yüksek riskli yatırım araçlarında değerlendirilen oranı azaltılarak devlet tahvili gibi düşük riskli ve getirili yatırım araçlarında değerlendirilen oranı artırılmalı ve emeklilik dönemine yaklaşıldığında ise fonun büyük bir kısmı düşük riskli ve getirili yatırım araçlarında değerlendirilmelidir.

### *4.2. Kayıptan Kaçınan Bireyler İçin Optimal Yatırım Stratejisi*

Kayıptan kaçınan bireyler için kullanılan modelde, 64 yaş için amaç fonksiyonunu maksimize eden yatırım oranının gerçekleşen fon büyüklüğüne bağlı değişimi Şekil 4.2'de gösterilmektedir.

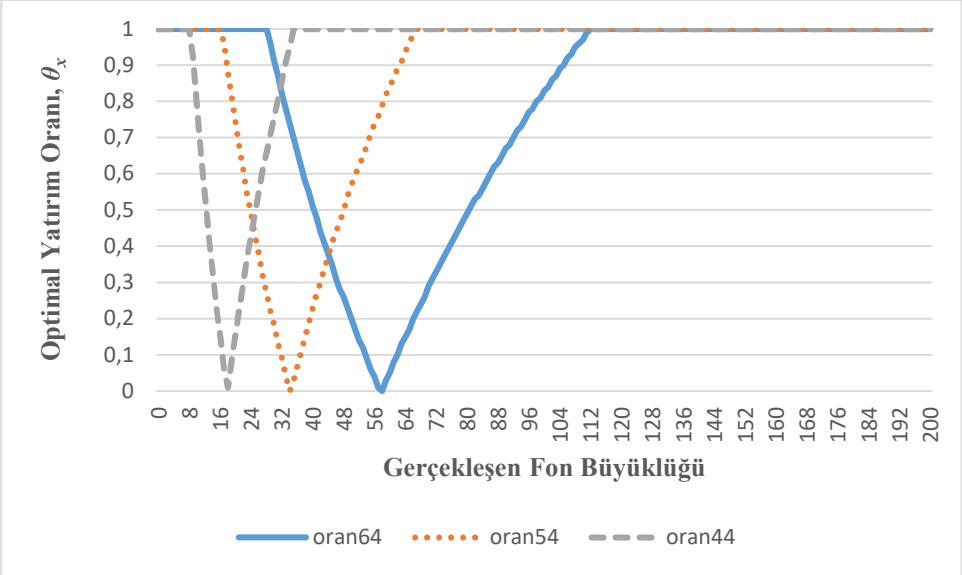


**Şekil 4.2:** 64 yaş için optimal yatırım oranının gerçekleşen fon büyüklüğüne bağlı değişimi

Şekil 4.2'den 64 yaşında gerçekleşen fon büyüklüğünün, hedeflenen fon büyüklüğü olan $f(64)$=59,84 değerine yakın olması halinde fonun yüksek riskli yatırım aracında değerlendirilen oranının 0'a yakın olduğu; yani bireyin gerçekleşen fon büyüklüğü hedeflenen fon büyüklüğüne yakın iken risk almayarak daha tutucu bir yatırım stratejisi izlediği, fakat gerçekleşen fon büyüklüğü hedeflenen fon büyüklüğünden uzaklaştıkça (bu uzaklaşmanın hem pozitif hem de negatif yönlü olması durumunda) fonun yüksek riskli yatırım aracında değerlendirilen oranının artarak 1'e ulaştığı; yani gerçekleşen fon büyüklüğü hedeflenen fon büyüklüğünden uzaklaştıkça bireyin hedeflenen fon büyüklüğüne ulaşmak için fonun neredeyse tamamını yüksek riskli yatırım aracında değerlendirerek, daha agresif bir yatırım stratejisi izlediği görülmektedir.

Şekil 4.2'den ayrıca hedeflenen fon büyüklüğünden küçük olan gerçekleşen fon büyüklüğü değerlerinde optimal oranın 0'a yaklaşma hızının, hedeflenen fon büyüklüğünden büyük olan gerçekleşen fon büyüklüğü değerlerindeki optimal oranın 1'e yaklaşma hızından daha fazla olduğu görülmektedir. Bu durum kayıptan kaçınma teorisinin de savunduğu gibi kayıpların bireyi kazanca göre daha fazla etkilediği sonucunu doğrulamaktadır. Birey gerçekleşen fon büyüklüğünün hedeflenen fon büyüklüğünden düşük olması durumunda daha agresif bir yatırım stratejisi izlerken, yüksek olması durumunda daha tutucu bir yatırım stratejisi izlemektedir.

Şekil 4.3'te optimal yatırım oranının 44, 54 ve 64 yaşlarındaki değişimi gösterilmektedir.




**Şekil 4.3:** 44,54 ve 64 yaşları için optimal yatırım oranının değişimi

Şekil 4.3'ten 44 ve 55 yaşlarında elde edilen optimal yatırım oranının 64 yaşa benzer bir değişim gösterdiği görülmektedir. Katılımcının tüm yaşlar için hedeflenen fon büyüklüğü gerçekleşen fon büyüklüğüne yaklaştıkça risksiz yatırım aracını tercih ettiği, hedeflenen fon büyüklüğü gerçekleşen fon büyüklüğünden uzaklaştıkça riskli yatırım aracını tercih ettiği, yaş arttıkça fonun düşük riskli yatırım aracında değerlendirilen oranının her iki yöne doğru arttığı aralığın da genişlediği görülmektedir. Şekil 4.3'ten ayrıca yaş ne olursa olsun hedeflenen fon ile gerçekleşen fon büyüklükleri birbirinden uzaklaştıkça fonun tamamının yüksek riskli yatırım araçlarında değerlendirilmesi gerektiği görülmektedir.

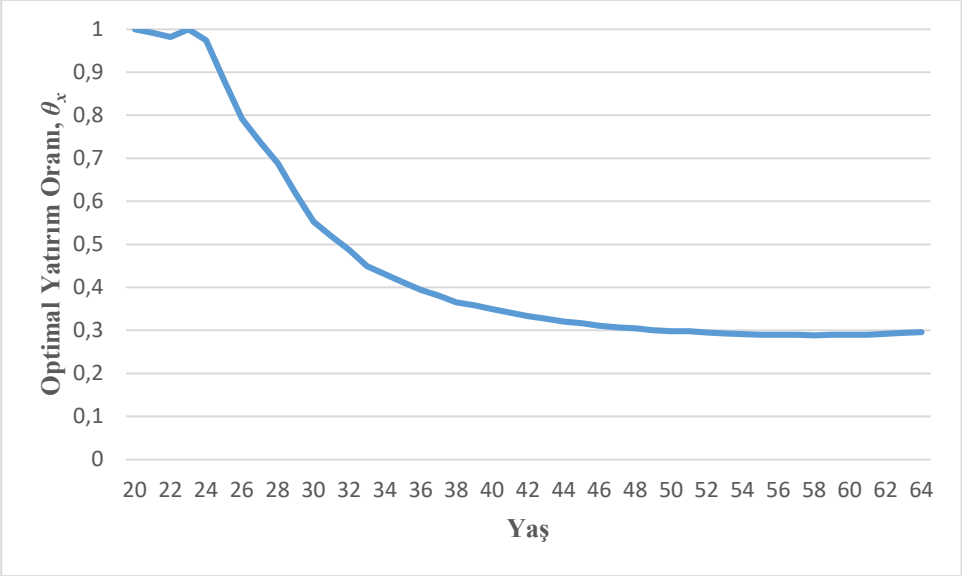
Birikim döneminin başlangıcı olan 20 yaştan, birikim döneminin sonu olan 64 yaşa kadar tüm yaşlar için optimal yatırım oranları olan $\theta_x$ değerleri, gerçekleşen fon büyüklüğünün aralığı olarak alınan [0,200] aralığındaki her bir fon değerine karşılık gelen 201 değer için elde edilmiştir. Burada her yaş için tek bir optimal yatırım oranının elde edilebilmesi için ise tüm yaşlardaki gerçekleşen fon büyüklüklerinin bilinmesi gerekir. Gerçekleşen fon büyüklüğü belirlenirken riskli yatırım aracının getirisine, yani optimize edilmeye

çalışılan $\theta_x$ değerlerine ihtiyaç vardır. $\theta_x$ değerlerine ilişkin önsel bir bilgi olarak Kırkağaç ve Gençtürk [27]'te elde edilen $\theta_x$ değerleri kullanılarak gerçekleşen fon büyüklükleri belirlenmiştir.

Optimal yatırım oranları sadece [0,200] aralığındaki 201 noktada elde edildiği için, fon büyüklüğünün tam sayı olmayan değerleri için optimal oranlar interpolasyon yöntemiyle bulunmuştur. $\{Z_x\}$ raslantı değişkenine bağlı olarak 100.000 benzetim yapılarak gerçekleşen fon büyüklüğüne ilişkin olası senaryolar üretilmiştir. Bu senaryolarda elde edilen her fon büyüklüğü için bir optimal yatırım oranı bulunmuş ve bu yatırım oranların ortalaması nihai optimal yatırım oranı olarak belirlenmiştir. Belirlenen nihai optimal yatırım oranının birikim döneminin başı olan 20 yaştan, birikim döneminin sonu olan 64 yaşa kadar değişimi Çizelge 4.2 ve Şekil 4.4'te gösterilmektedir.

**Çizelge 4.2:** Kayıptan Kaçınan Bireyler için Optimal Yatırım Oranları

| Yaş | Optimal Yatırım Oranı |
|---|---|
| 20 | 1 |
| 25 | 0,8810 |
| 30 | 0,5525 |
| 35 | 0,4118 |
| 40 | 0,3497 |
| 45 | 0,3164 |
| 50 | 0,2977 |
| 55 | 0,2901 |
| 60 | 0,2899 |
| 64 | 0,2959 |




**Şekil 4.4:** Kayıptan Kaçınan Bireyler için Optimal yatırım stratejileri

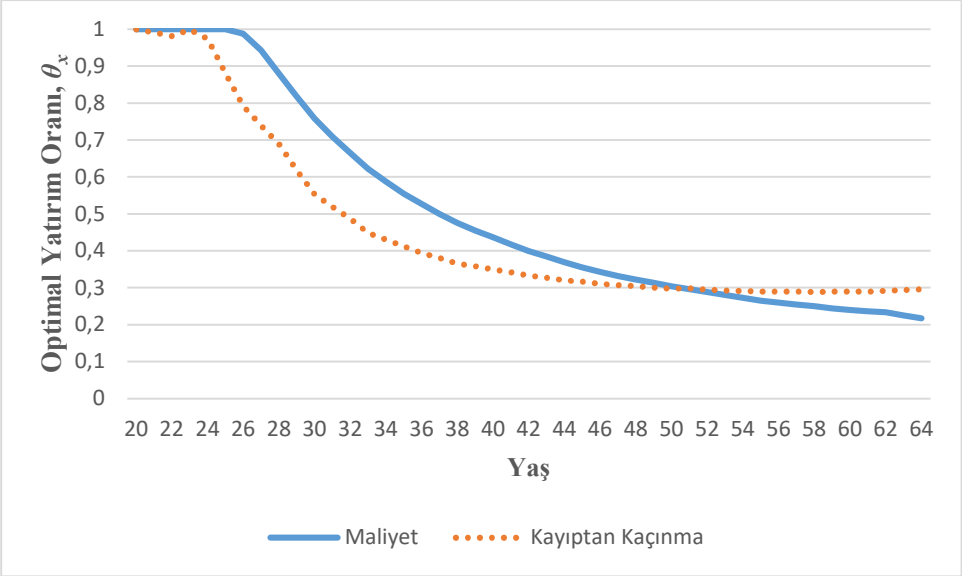
Çizelge 4.2 ve Şekil 4.4'ten elde edilen optimal yatırım stratejisi fonun yüksek riskli yatırım aracında değerlendirilen oranının vadenin başında 1 değerini almakta ve vade sonuna doğru giderek azalmaktadır. Elde edilen strateji aslında birçok emeklilik planında kabul görmüş ve yaygın olarak kullanılan "geleneksel yaşam tarzı stratejisine" benzemektedir. Elde edilen bu stratejiye göre birikim döneminin başında fonun yüksek riskli yatırım aracında değerlendirilen oranı 1 değerini almakta, yani fonun tamamı yüksek riskli yatırım aracında değerlenmektedir. Vade ilerledikçe fonun yüksek riskli yatırım aracında değerlendirilen oranı azalmakta, düşük riskli yatırım aracında değerlendirilen oranı artmaktadır. Birikim döneminin sonunda ise fonun büyük bir kısmı düşük riskli yatırım aracında değerlenmektedir.

*4.3. Maliyet Fonksiyonunun Kullanıldığı Model ile Kayıptan Kaçınan Bireyler İçin Optimal Yatırım Stratejilerinin Karşılaştırılması*

Bu bölümde maliyet fonksiyonunun kullanıldığı model ile kayıptan kaçınan bireyler için elde edilen optimal yatırım stratejilerinin karşılaştırılmasına yer verilmiştir. Her iki yöntemle elde edilen optimal yatırım stratejileri karşılaştırmalı olarak Çizelge 4.3 ve Şekil 4.5'te verilmektedir.

**Çizelge 4.3:** Optimal Yatırım Oranları

| Yaş | Maliyet | Kayıptan Kaçınma |
|---|---|---|
| 20 | 1 | 1 |
| 25 | 0,9998 | 0,8810 |
| 30 | 0,7583 | 0,5525 |
| 35 | 0,5543 | 0,4118 |
| 40 | 0,4367 | 0,3497 |
| 45 | 0,3553 | 0,3164 |
| 50 | 0,3039 | 0,2977 |
| 55 | 0,2649 | 0,2901 |
| 60 | 0,2398 | 0,2899 |
| 64 | 0,2178 | 0,2959 |




**Şekil 4.5:** Optimal yatırım stratejileri

Çizelge 4.3 ve Şekil 4.5'ten elde edilen optimal yatırım stratejisinin genel şeklinin maliyet fonksiyonunun kullanıldığı model ile elde edilen optimal yatırım stratejisi ile uyumlu olduğu görülmektedir. Her iki modelde de fonun yüksek riskli yatırım aracında değerlendirilen oranının vadenin başında 1 değerini almakta ve vade sonuna doğru giderek azalmaktadır. Her iki modelde de elde edilen strateji aslında birçok emeklilik planında kabul görmüş ve yaygın olarak kullanılan "geleneksel yaşam tarzı stratejisine" benzemektedir.

Çizelge 4.3 ve Şekil 4.5'ten ayrıca maliyet fonksiyonunun kullanıldığı modelde fonun riskli yatırım aracında değerlendirildiği yıl sayısının daha uzun olduğu, bununla birlikte birikim döneminin sonunda fonun daha büyük bir kısmının düşük riskli yatırım aracında değerlendirildiği görülmektedir. Bu bulgulara göre kayıptan kaçınan bireyin maliyet fonksiyonun kullanıldığı modele göre birikim döneminin başlarındaki uzun sürede daha az riskli bir yatırım stratejisi izlerken, birikim döneminin son yıllarındaki kısa sürede ise daha agresif bir yatırım stratejisi izlediği söylenebilir.

## 5. Sonuç ve öneriler

Bu çalışmada katkısı belirli emeklilik planlarında kayıptan kaçınan bireyler için ve gelecekte ortaya çıkacak hedef fondan sapmaları minimize eden optimal yatırım stratejisi belirlenmiş, optimal yatırım oranının gerçekleşen fon büyüklüğüne ve yaşa bağlı değişimi incelenmiştir.

Kayıptan kaçınan birey gerçekleşen fon büyüklüğünün hedeflenen fon büyüklüğüne yakın olması durumunda tutucu bir yatırım stratejisi izlerken, gerçekleşen fon büyüklüğü hedeflenen fon büyüklüğünden hem pozitif hem de negatif yönlü uzaklaşırken agresif bir yatırım stratejisi izlemektedir. Gerçekleşen ve hedeflenen fon büyüklüğü arasındaki fark arttıkça optimal yatırım stratejisi daha agresif hale gelmektedir. Kayıptan kaçınan birey için optimal yatırım stratejisi bir çok emeklilik planında kabul gören geleneksel yaşam tarzı stratejisi ile benzer olarak elde edilmiştir. Elde edilen bu stratejiye göre birikim döneminin başında fonun tamamı riskli yatırım aracında değerlenmeli, birikim döneminin ilerleyen yaşlarında fonun riskli yatırım aracında değerlendirilen oranı azaltılarak, risksiz yatırım aracında değerlendirilen oranı artırılmalı, birikim döneminin sonunda ise fonun büyük bir kısmı risksiz yatırım aracında değerlendirilmelidir.

Kayıptan kaçınan birey için elde edilen optimal yatırım stratejisi maliyet fonksiyonunun kullanıldığı model ile elde edilen strateji ile karşılaştırıldığında ise sonuçların birbirine çok yakın gerçekleştiği görülmüştür. Bununla birlikte kayıptan kaçınan bireyin beklenildiği gibi daha az risk alarak daha tutucu bir yatırım stratejisi izlediği görülmektedir. Kayıptan kaçınan bireyin riski tercih etmediği gerçeği göz önünde bulundurulduğunda sonuçların bu anlamda tutarlı olduğu söylenebilir.

Bu çalışmada elde edilen sonuçlar birçok yönden geliştirilmeye açıktır.

Kullanılan modelde yıllık logaritmik getirilerin dağılımının Normal dağılıma uyduğu ve getirilerin birbirinde bağımsız olduğu varsayılmıştır. Fakat gerçekte yıllık getiriler bu varsayıma uymadığından, optimal yatırım stratejisi yatırım getirilerinin dağılımının Normal dağılıma uymaması durumunda belirlenebilir. Bu durumda optimize edilmeye çalışılan beklenen değer içindeki ifadenin çözümü sayısal yöntemlerle elde edilemeyebileceğinden dolayı aktüerya literatüründe çok yaygın kullanım alanı olmayan, büyük boyutlu optimizasyon problemleri için kabul edilebilir sürede optimuma yakın çözümler verebilen sezgisel optimizasyon algoritmaları kullanılabilir. Yatırım getirilerinin birbirine bağımlı olması durumu ise kopulalar kullanılarak analiz edilebilir.

## Kaynaklar

[1] Blake, D., Annuities in Pension Plans, In Commentary at World Bank Annuities Workshop, 7-8 June, United Kingdom, 1999, p. 7.
[2] Aitken, W. H., A Problem-Solving Approach to Pension Funding and Valuation, Actex Publications, 1996.
[3] Samuelson, P. A., Lifetime Portfolio Selection by Dynamic Stochastic Programming, The Review of Economics and Statistics, (1969) 239.
[4] Merton, R. C., Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case, The Review Of Economics And Statistics, (1969) 247.
[5] Merton, R. C., Optimum Consumption and Portfolio Rules in a Continuous-Time Model, Journal of Economic Theory, 3(4) (1971) 373.
[6] Bodie, Z., Merton, R. C. and Samuelson, W. F., Labor Supply Flexibility and Portfolio Choice in a Life Cycle Model, Journal of Economic Dynamics and Control, 16(3-4) (1992) 427.
[7] Cairns, A. J. G., An Introduction to Stochastic Pension Fund Management, Working Paper 9607, Pensions Institute, 1996.
[8] Owadally, M. I., The Dynamics and Control of Pension Funding, Doctoral Dissertation, City University, London, 1998.
[9] Vigna, E. and Haberman, S., Optimal Investment Strategy for Defined Contribution Pension Schemes, Insurance: Mathematics and Economics, 28(2) (2001) 233.
[10] Blake, D., Cairns, A. J. and Dowd, K., Pensionmetrics: Stochastic Pension Plan Design and Value-At-Risk During The Accumulation Phase, Insurance: Mathematics and Economics, 29(2) (2001) 187.

[11] Haberman, S. and Vigna, E., Optimal Investment Strategies and Risk Measures in Defined Contribution Pension Schemes, Insurance: Mathematics and Economics, 31(1) (2002) 35.
[12] Gerrard, R., Haberman, S. and Vigna, E., Optimal Investment Choices Post-Retirement in a Defined Contribution Pension Scheme, Insurance: Mathematics and Economics, 35(2) (2004) 321.
[13] Cairns, A. J., Blake, D. and Dowd, K., Stochastic Lifestyling: Optimal Dynamic Asset Allocation for Defined Contribution Pension Plans, Journal of Economic Dynamics and Control, 30(5) (2006) 843.
[14] Battocchio, P., Menoncin, F. and Scaillet, O., Optimal Asset Allocation for Pension Funds Under Mortality Risk During The Accumulation and Decumulation Phases, Annals of Operations Research, 152(1) (2007) 141.
[15] Yang, S. S. and Huang, H. C., The Impact of Longevity Risk on the Optimal Contribution Rate and Asset Allocation for Defined Contribution Pension Plans, The Geneva Papers on Risk and Insurance Issues and Practice, 34(4) (2009) 660.
[16] Blake, D., Wright, D. and Zhang, Y., Age-Dependent Investing: Optimal Funding and Investment Strategies in Defined Contribution Pension Plans When Members Are Rational Life Cycle Financial Planners, Journal of Economic Dynamics and Control, 38 (2014) 105.
[17] Chen, A., Haberman, S. and Thomas, S., Optimal Decumulation Strategies During Retirement with Deferred Annuities, Available at SSRN 2911959, 2017.
[18] Rabin, M. and Thaler, R. H., Anomalies: Risk Aversion, The Journal of Economic Perspectives, 15(1) (2001) 219.
[19] Kahneman, D. and Tversky, A., Prospect Theory: An Analysis of Decision under Risk, Econometrica, 47 (1979) 263.
[20] Berkelaar, A. B., Kouwenberg, R. and Post, T., Optimal Portfolio Choice Under Loss Aversion. Review of Economics and Statistics, 86(4) (2004) 973.
[21] Gomes, F. J., Portfolio Choice and Trading Volume with Loss-Averse Investors, The Journal of Business, 78(2) (2005) 675.
[22] Blake, D., Wright, D. and Zhang, Y., Target-Driven Investing: Optimal Investment Strategies in Defined Contribution Pension Plans Under Loss Aversion, Journal of Economic Dynamics and Control, 37(1) (2013) 195.
[23] M. I. Owadally, S. Haberman, D. G. Hernández, 2013, A Savings Plan with Targeted Contributions, The Journal of Risk and Insurance, 80(4), 975-1000.
[24] H. K. Sezen, 2007, Yöneylem Araştırması, 2. Baskı, Ekin Basım Yayın Dağıtım, Bursa.
[25] R. Bellman, 1957, Dynamic Programming, Princeton University Press, New Jersey.
[26] Tversky, A. and Kahneman, D., Advances in Prospect Theory: Cumulative Representation of Uncertainty, Journal of Risk and uncertainty, 5(4) (1992) 297.
[27] Kırkağaç, M., Gençtürk, Y. (2016). Bireysel emeklilik planlarında hedef fon büyüklüğüne ulaşmak için değişken katkı ve optimal yatırım stratejisi. İstatistikçiler Dergisi: İstatistik ve Aktüerya, 9(2), 54-65.

**İstatistikçiler Dergisi: İstatistik & Aktüerya**

Journal of Statisticians: Statistics and Actuarial Sciences

**IDIA 16**, 2023, 2, 57-80

**Geliş** / Received: 25.11.2023, **Kabul** / Accepted: 28.12.2023

**Araştırma Makalesi /** Research Article

# Karma veri içeren çok yanıtlı problemlerde NSGA-II'ye dayalı bir optimizasyon yaklaşımı

Gözde Karakoç[1]

*Ankara Üniversitesi, Fen Bilimleri Enstitüsü,*
*İstatistik Anabilim Dalı*
*Ankara, Türkiye*
gozcirpan@ankara.edu.tr
ORCID: 0000-0001-9334-765X

Özlem Türkşen

*Ankara Üniversitesi, Fen Fakültesi,*
*İstatistik Bölümü*
*Ankara, Türkiye*
turksen@ankara.edu.tr
ORCID: 0000-0002-5592-1830

**Öz**

Bir optimizasyon probleminin matematiksel modeli, kesikli ve sürekli değer alan girdi ve/veya yanıt değişkenlerini içermesi durumunda problem, karma veri içeren optimizasyon problemi olarak adlandırılır. Bu çalışmada, girdi değişkenleri bakımından karma veri içeren çok yanıtlı problemlerin modelleme ve optimizasyon aşamaları ele alınmıştır. Modelleme aşamasında Genelleştirilmiş Lineer Modeller (GLM) kullanılarak tahmini yanıt fonksiyonları elde edilmiştir. Optimizasyon aşamasında ise elde edilen tahmini yanıt fonksiyonları bir amaç fonksiyonu olarak dikkate alınıp problem, eş anlı optimizasyonu istenilen çok amaçlı optimizasyon (ÇAO) problemi biçiminde ifade edilmiştir. Çalışmada, ÇAO'da sıklıkla kullanılan yapay zeka optimizasyon algoritmalarından biri olan NSGA-II (Non-dominated Sorting Genetic Algorithm-II)'ye dayalı yeni bir çözüm algoritması önerilmiştir. NSGA-II'de, değişken gösterimi, başlangıç popülasyonu oluşturma ve genetik operatörlerin uygulanması aşamalarında çeşitli uyarlamalar yapılarak hazırlanan bu algoritma, çalışma kapsamında MDNSGA-II (Mixed Data NSGA-II) olarak adlandırılmıştır. MDNSGA-II'de, her bir kesikli değişken değerine bir pozitif tam sayı değeri atanarak, kesikli değişken değerleri için bir tam sayı indekslemesi yapılmıştır. Yapılan indeksleme işlemiyle kesikli değişkenin tanım kümesinden değerler alması sağlanmıştır. Çalışmanın uygulama kısmında, UCI Repository veri tabanından enerji verimliliği konulu karma veri seti ve gıda alanında literatürde mevcut olan deneysel karma veri seti kullanılarak önerilen MDNSGA-II ile Pareto çözümlerin elde edilebilir olduğu gösterilmiştir.

***Anahtar sözcükler:*** Çok amaçlı optimizasyon, Çok yanıtlı problem, Karma veri, MDNSGA-II, NSGA-II

[1] Bu çalışma, birinci yazarın, ikinci yazarın danışmanlığında hazırladığı doktora tezinden üretilmiştir.

**Abstract**

***An optimization approach based on NSGA-II for multi-response problems with mixed data***

*If the mathematical model of an optimization problem contains input and/or response variables that take discrete and continuous values, the problem is called as a mixed data optimization problem. In this study, the modeling and the optimization phases of multi-response problems with mixed data in terms of input variables are discussed. In the modeling phase, estimated response functions are obtained using Generalized Linear Models (GLM). In the optimization phase, the estimated response functions are considered as an objective function and the problem is expressed as a multi-objective optimization (MOO) problem with simultaneous optimization. In this study, a new solution algorithm based on the Non-dominated Sorting Genetic Algorithm-II (NSGA-II), one of the most frequently used artificial intelligence optimization algorithms in MOO, is proposed. This algorithm, which was prepared by making various adaptations in the variable representation, initial population generation and application of genetic operators stages of the NSGA-II, is called MDNSGA-II (Mixed Data NSGA-II) in this study. In MDNSGA-II, each discrete variable value is assigned a positive integer value and an integer indexing is performed for the discrete variable values. It is ensured that the discrete variable takes values in support set of variables by using the indexing approach. In the application part of the study, it is shown that Pareto solutions can be obtained with the proposed MDNSGA-II by using the mixed data set on energy efficiency from the UCI Repository database and the experimental mixed data set available in the literature in the field of food.*

***Keywords:*** *Multi-objective optimization, Multi-response problem, Mixed data, MDNSGA-II, NSGA-II*

## 1. Giriş

Fen, mühendislik, sağlık, sosyal, finans vb. alanlarda yapılan araştırmalarda kullanılacak verilerin toplanması veri-bilgi keşfi sürecinde büyük önem taşımaktadır. Günümüzde gelişen teknolojiye bağlı olarak veri toplama yöntemleri hızlı ve pratik bir hal alsa da anketler, görüşmeler, gözlemler ve deneysel çalışmalar veri toplama yöntemleri başlığında yer alan dört temel yöntem olarak değerlendirilebilir. Toplanan verilerin analizinden önce veri yapısının araştırmacı tarafından iyi anlaşılmış olması gerekir. Gözlem ve deney yolu ile elde edilmiş bir veri seti birden fazla yanıt (bağımlı, açıklanan) ve girdi (bağımsız, açıklayıcı) değişkeni içerebilir.

Değişkenler sürekli, kesikli ve/veya tam sayı değerli olabilir. Veri setindeki değişkenlerin (yanıt ve/veya girdi değişkenlerinin) bazılarının sürekli bazılarının kesikli değerli olması durumunda veri, karma veri olarak tanımlanır. Karma veri ile ilgili modelleme ve optimizasyon çalışmaları disiplinler arası çalışmalarda oldukça önemli yer tutar. Optimizasyon sonuçlarının güvenilirliği, veriye ilişkin oluşturulan istatistiksel modelin güvenilirliği ile doğru orantılıdır.

Karma veri setindeki yanıt değişkeninin kesikli ya da sürekli değerli olmasına göre oluşturulacak model farklılık gösterir. Sürekli değerli yanıt değişkeni ile girdi değişkenleri arasındaki fonksiyonel ilişkinin tanımlanmasında çoklu doğrusal regresyon analizi uygulanırken, yanıt değişkeninin kesikli değerli olması durumunda dağılım yapısına uygun olarak lojistik, binom veya poisson regresyon analizleri uygulanabilir. Uygulanan doğrusal regresyon modelleri, Genelleştirilmiş Lineer Modeller (Generalized Linear Models - GLM)'in özel halleridir. GLM, kitle ortalamasının bir bağ (link) fonksiyonu ile doğrusal tahmin ediciye bağlı olmasına izin veren lineer modellerin genelleştirilmiş halidir [1]. GLM, genellikle biyolojik analizlerde, çeşitli uygulamalı biyomedikal alanlarda, güvenilirlik ve hayatta kalma analizlerinde elde edilen veri setlerinden istatistiksel sonuçlar çıkarmak için kullanılır.

Literatürde, GLM kullanılarak farklı alanlarda karma veri setleri için yapılan çalışmalar mevcuttur. GLM, HIV virüsü riski ile partnerlerle temas sayısı arasındaki ilişkinin incelenmesinde [2], entomolojide (böcek bilimi) böcek davranışındaki değişiklikleri bir bitki özütünün kimyasal bileşimindeki değişikliklerle

ilişkilendirmek amacıyla [3], klinik çalışmalarda hastanelerde hastalara uygulanan tedavilerin etkilerine ilişkin tahminlerin elde edilmesinde [4], önemli bir ağaç türünün mekansal desenini incelemede [5], klimatolojide (iklim bilimi) belirli bölgelerdeki temel klimatolojik modeli ve günlük maksimum rüzgar hızındaki eğilimlerin belirlenmesinde [6] kullanılmıştır.

GLM ile tahmini yanıt fonksiyonları oluşturulduktan sonra her bir yanıt fonksiyonu bir amaç fonksiyonu olarak ele alınıp çok amaçlı optimizasyon (ÇAO) aşamasına geçilir. ÇAO'da çözümler üstel hesaplama gerektirebilir. Klasik optimizasyon yöntemleri, bu tür zor optimizasyon problemlerini çözmede yetersiz kalmaktadır. Bu nedenle, zor optimizasyon problemlerinin çözümünde makul zamanda optimal sonuca yakın çözümler üreten yapay zeka optimizasyon algoritmaları önemli yer tutar. Karma veri içeren çok yanıtlı optimizasyon problemlerini çözmek için uygun yöntemin seçimi kesikli değer alan değişkenlerin tipine ve amaç fonksiyonunun yapısına bağlıdır. Literatürde, karma veri içeren ÇAO problemlerini çözmek için farklı alanlarda yapılan çalışmalar mevcuttur. Rajeev ve Krishnamoorthy [7], Lin ve Hajela [8] karma veri içeren optimizasyon problemlerinin çözümü için Genetik Algoritma (GA)'nın uygun olduğunu gösteren çalışmalar yapmış ve çalışmalarında ikili (0-1) kodlama kullanmışlardır. Wang ve ark. [9], bir bina tasarımında, tasarımcılara yardımcı olabilecek ÇAO problemini hem ekonomik hem de çevresel kriterler bakımından değerlendirerek Pareto optimal çözümün belirlenmesi için çok amaçlı GA kullanmışlardır. Rao ve Xiong [10], karma kesikli bulanık çok amaçlı programlama problemlerini çözmek için bulanık λ formülasyonu ve oyun teorisi tekniklerinin karma kesikli hibrit GA ile birleştirildiği yeni bir yöntem sunmuşlardır. Sundukları yöntemin kesin olmayan bir ortamda daha gerçekçi ve tatmin edici sonuçlar elde etmek için çeşitli mühendislik tasarım problemlerine esnek ve etkili bir biçimde uygulanabildiğini göstermişlerdir. Ahmadi ve ark. [11], havza ölçeğinde tarımsal koruma uygulamalarının hedeflenen şekilde uygulanması için karma karar değişkenlerine sahip NSGA-II kullanmışlardır. El-Kribi ve ark. [12], sürekli ve kesikli değişkenlere sahip dört çubuklu bir sistemin mekatronik tasarımını eş anlı optimize etmek için NSGA-II'yi kullanarak farklı tasarım koşulları için analiz etmişlerdir. Tong ve ark. [13], karma kesikli problemler için daha önce geliştirilen ve erken parçacık kümelenmesini önlemek için özel bir çeşitlilik koruma tekniği içeren karma-kesikli Parçacık Sürü Optimizasyonu algoritmasını çok amaçlı problemleri çözmek için kullanmış ve NSGA-II sonuçları ile karşılaştırmışlardır. Holzman ve Smith [14], kesikli ÇAO problemine tam verimli bir çözüm kümesi üretmek için modifiye edilmiş arttırılmış ağırlıklı Tchebychev normunu sunmuşlardır. Kullandıkları algoritmanın çalışma sürelerini literatürde önerilen algoritmaların çalışma süreleriyle karşılaştırmışlardır. Guangyong ve ark. [15], belirsizlikler içeren mühendislik yapılarının tasarımı için yeni çok amaçlı kesikli robust optimizasyon algoritması önermişlerdir. Çalışmalarında çok kriterli karar verme tekniğini Taguchi yöntemine dahil ederek çok amaçlı kesikli sağlam tasarımı ele almayı amaçlamışlardır. Önerdikleri algoritmanın NSGA-II ile elde edilen Pareto sınırlarına yakın olduğunu belirtmişlerdir. Roy ve ark. [16], iki dallı bir genetik algoritmayı küresel bir arama aracı olarak yerel arama için gradyan tabanlı bir yaklaşımla birleştiren, kısıtlı çok amaçlı karma kesikli doğrusal olmayan programlama problemine çözüm bulmak için yeni bir hibrit yaklaşım önermişlerdir.

Çoğu zaman karma veri içeren çok yanıtlı problemlerin optimizasyonu için uygun bir yöntemin seçimi, optimizasyon problemindeki amaç ve kısıt fonksiyonlarının yapısı ile kesikli değişkenlerin aldığı değerlere bağlıdır. Literatürde mevcut olan ÇAO yaklaşımlarında kesikli değişken gösterimi aşamasında ikili kodlamanın kullanıldığı görülmüştür. Bu kodlama tipi kolaylıkla uygulanabilir olsa da kesikli değişken sayısının ve kesikli değişkenlerin aldığı değerlerin fazla olması durumunda hesaplama süresi arttığından farklı kodlama yaklaşımları geliştirilmiştir. Yapılan bu çalışmayla, kesikli değişken gösterimi için ikili kodlama yerine değer kodlama yaklaşımı kullanılarak hesaplama süresinin kısaltılması hedeflenmiştir. NSGA-II'nin değişken gösterimi adımında yapılan bu değer kodlaması uyarlaması ile kesikli değişkenlerin aldığı değerlerin optimizasyon sürecine dahil edilmesi sağlanmıştır.

Bu çalışmada, kesikli ve sürekli değer alan değişkenlere sahip karma verilerin modellenmesi ve bu modellerin tanımlanan amaca yönelik ÇAO kapsamında Pareto çözüm sonuçlarının belirlenmesi amaçlanmıştır. Çok yanıtlı problemlerin optimizasyonunda etkin bir yapay zeka optimizasyon algoritması olan NSGA-II yönteminde, değişken gösterimi, başlangıç popülasyonu oluşturulması ve genetik operatörlerin belirlenmesi aşamalarında uyarlamalar yapılarak karma veri içeren yanıt fonksiyonlarının eş

anlı optimizasyonu sağlanmıştır. Uyarlanan NSGA-II, MDNSGA-II olarak adlandırılmıştır. Çözümlerin hesaplama süresini kısaltması nedeniyle çalışmada değer kodlama kullanılmıştır. Değer kodlaması ile ele alınan ÇAO probleminde kesikli değişken değerleri için pozitif tam sayı indekslemesi yapılmıştır. Yapılan indeksleme sonucunda kesikli değişkenlerin indeks değerleri dikkate alınarak karma veri içeren çok yanıtlı problemler için Pareto çözüm kümesi elde edilmiştir. Çalışmanın ikinci bölümünde, çok yanıtlı karma verilerin GLM ile modellenmesi hakkında kısa bilgi verilerek, çalışmada kullanılan GLM modelleri sunulmuştur. Üçüncü bölümde NSGA-II yönteminin bazı adımlarında uyarlamalar yapılarak oluşturulan MDNSGA-II detaylı olarak açıklanmıştır. Dördüncü bölümde, UCI Repository veri tabanından enerji verimliliği konulu karma veri seti ve literatürde mevcut olan gıda alanından deneysel karma veri seti uygulamalarına yer verilmiştir. Çalışmanın sonuç bölümünde, önerilen MDNSGA-II ile elde edilen Pareto çözüm kümesi değerlendirilerek, modelleme ve optimizasyon aşamalarında yapılması planlanan sonraki çalışmadan söz edilmiştir.

## 2. Çok yanıtlı karma verilerin modellenmesi

Birçok disiplinde, yapılan çalışmalardan elde edilen verilerin modellenmesi oldukça önem taşımaktadır. Güvenilir bir model kurma, veriyi iyi anlayıp doğru analizlerin yapılması ile mümkündür. Bu nedenle modelleme aşamasına geçmeden önce veri yapısı incelenmelidir. Veriler, girdi değişkenleri ve yanıt değişken değerlerinden oluşur. Yanıt değişkenlerinin birden fazla olması durumunda çok yanıtlı verilerin analizinde çok değişkenli yapının göz önünde bulundurulması gerekir. Yapılan modelleme çalışmalarında girdi değişkenlerinin $(X_i,\ i=1,2,...,p)$ ilgilenilen yanıt değişkenleri $(Y_j,\ j=1,2,...,r)$ üzerindeki etkisi incelenir. Yanıt değişkeni ile girdi değişkenleri arasındaki fonksiyonel ilişki, GLM ile elde edilir. GLM, kesikli ve sürekli yanıt değişkenleri için birleştirilmiş regresyon modelleri sınıfıdır [17]. Çizelge 1'de, $p$ tane girdi değişkeni, $r$ tane yanıt değişkenine sahip $n$ gözlemli çok yanıtlı karma veri seti yer almaktadır. Burada, $\mathbf{X}, n\times p$ boyutlu girdi değişkenlerin tasarım matrisi ve $\mathbf{Y}=[\boldsymbol{Y}_1\ \boldsymbol{Y}_2...\boldsymbol{Y}_r]^{\mathrm{T}}$, her bir bileşeni $\boldsymbol{Y}_i$, $i=1,2,...,r$ , $n\times 1$ boyutlu olan yanıt değişkenleri vektörüdür.

**Çizelge 1.** Çok yanıtlı karma veri seti

| | ***Girdi Değişkenleri*** | | | | ***Yanıt Değişkenleri*** | | | |
|---|---|---|---|---|---|---|---|---|
| ***No*** | $X_1$ | $X_2$ | **…** | $X_p$ | $\boldsymbol{Y_1}$ | $\boldsymbol{Y_2}$ | **…** | $\boldsymbol{Y_r}$ |
| ***1*** | $x_{11}$ | $x_{12}$ | $\cdots$ | $x_{1p}$ | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1r}$ |
| ***2*** | $x_{21}$ | $x_{22}$ | $\cdots$ | $x_{2p}$ | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2r}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| ***n*** | $x_{n1}$ | $x_{n2}$ | $\cdots$ | $x_{np}$ | $y_{n1}$ | $y_{n2}$ | $\cdots$ | $y_{nr}$ |

GLM, yanıt değişkeni dağılımının Normal dağılım varsayımını sağlamadığı durumları da göz önüne alan hem doğrusal hem de doğrusal olmayan regresyon modellerinin bir birleşimidir [18]. Başka bir deyişle GLM, kitle ortalamasının bir bağ fonksiyonu ile doğrusal tahmin ediciye bağlı olmasına izin veren lineer modellerin genelleştirilmiş halidir [1]. Nelder ve Wedderburn [19] tarafından ilk olarak kullanılan GLM, üç temel bileşenden meydana gelmektedir. Bu bileşenler, yanıt değişkeninin dağılımı, lineer tahmin edicilerin bulunduğu sistematik kısım ve bağ fonksiyonudur. Modelin sistematik kısmı, $X_1, X_2, \ldots, X_p$ girdi değişkenlerini içerir. GLM'de yanıt değişkeninin dağılımı, Normal, Poisson, Binom, Üstel ve Gamma dağılımlarını içeren üstel ailenin bir üyesi olmalıdır. Yanıt değişkeninin sürekli olduğu durumlarda Gamma, Ters Gauss dağılımı ve Normal dağılım, kesikli olduğu durumlarda Poisson, Bernoulli ve Binom dağılımı üstel aile örnekleridir [20]. GLM'de, $Y$'nin dağılımı üstel ailenin genel formunda

$$f(y|\theta,\phi)=\exp\left\{\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\right\} \tag{1}$$

biçiminde yazılır [21]. Burada, $a(.)$, $b(.)$ ve $c(.)$ bazı özel fonksiyonlardır. Dağılımın ortalamasının $(\mu)$ bir fonksiyonu olan $\theta$, kanonik parametre olarak adlandırılır ve dağılımın konumu hakkında bilgi içerir. $\phi$, yayılım veya ölçek parametresi ve $c(y,\phi)$, gözlemler ile yayılım parametresinin bir fonksiyonudur. Üstel dağılım ailesinin olabilirlik fonksiyonunun logaritmasının, $\theta$ 'ya göre birinci ve ikinci türevleri alınıp sıfıra eşitlendiğinde, *Y*'nin ortalama ve varyansı sırasıyla

$$E(Y) = \mu = b'(\theta) \tag{2}$$

ve

$$Var(Y) = b''(\theta)a(\phi) \tag{3}$$

olur. GLM'de temel amaç, yanıt değişkeninin beklenen değerinin uygun bir fonksiyonu için bir model geliştirmektir. Bağ ya da link fonksiyonu olarak adlandırılan bu fonksiyon, yanıt değişkeninin ortalaması ile doğrusal tahmin ediciler arasında ilişki kurulmasını sağlar ve

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = \boldsymbol{X^T \beta} \tag{4}$$

biçiminde ifade edilir. Eşitlik (4) ile ifade edilen bağ fonksiyonu $g(\bullet)$, monoton ve türevlenebilir olmalıdır. Bu fonksiyonun 1-1 olduğu anlamına gelir. Bu nedenle bağ fonksiyonunun ters dönüşümü vardır. Ters fonksiyon $g^{-1}(\bullet)$, ortalama fonksiyon olarak da adlandırılıp, $\mu = g^{-1}(\eta) = g^{-1}(\boldsymbol{X^T \beta})$ 'dır [22]. Çizelge 2'de, GLM ile kullanılan yaygın üstel aile dağılımları ve bu dağılımlarda kullanılan bağ fonksiyonları verilmiştir.

**Çizelge 2.** GLM ile kullanılan yaygın üstel aile dağılımları, bağ ve varyans fonksiyonları [23]

| ***Dağılım*** | ***Bağ fonksiyonu*** | ***Varyans fonksiyonu*** |
|---|---|---|
| Normal | $\eta = \mu$ (özdeş bağ) | 1 |
| Binom | $\eta = \log(\frac{\mu}{1-\mu})$ (logit bağ) | $\mu(1-\mu)$ |
| Poisson | $\eta = \log(\mu)$ (log bağ) | $\mu$ |
| Gamma | $\eta = \mu^{-1}$ (ters bağ) | $\mu^2$ |
| Ters Normal | $\eta = \mu^{-2}$ (ters kare bağ) | $\mu^3$ |

Bağ fonksiyonunun özdeş olması durumunda dağılım Normal olur ve değişkenler arasındaki ilişki klasik bilinen regresyon analizi ile oluşturulur. Çoklu doğrusal regresyon modelinin en genel hali

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon \tag{5}$$

biçiminde tanımlanır. Burada $\beta_j$, $j = 1,2,\ldots,p$ parametreleri regresyon katsayıları ve $\varepsilon$, hata terimidir. Polinom regresyon, çoklu doğrusal regresyonun özel bir durumudur. Polinom modelleri etkili ve esnek bir eğri uydurma tekniğidir [24]. *k* değişkenli ikinci dereceden bir polinom modeli

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k} \beta_{ii} X_i^2 + \sum_{i=1}^{k} \sum_{i<j}^{k} \beta_{ij} X_i X_j + \varepsilon \ , \ i = 1,2,...,k, \ j = 1,2,...,k \ , \ i \neq j \tag{6}$$

biçiminde verilen doğrusal regresyon modelidir. Eşitlik (6)'daki model parametreleri En Küçük Kareler (EKK) (Ordinary Least Squares-OLS) yöntemi kullanılarak tahmin edilebilir. Parametreler tahmin edildiğinde elde edilen tahmini yanıt modeli

$$\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i X_i + \sum_{i=1}^{k} \hat{\beta}_{ii} X_i^2 + \sum_{i=1}^{k} \sum_{i<j}^{k} \hat{\beta}_{ij} X_i X_j \quad , i = 1,2,...,k, \; j = 1,2,...,k \; , \; i \neq j \tag{7}$$

olur.

## 3. Karma veri modelleri için çok amaçlı optimizasyon

İki ve ikiden fazla sayıda amaç fonksiyonunun eş anlı optimal değerinin elde edilmesini hedefleyen optimizasyon modeli ÇAO modeli olarak adlandırılır. ÇAO'da amaç fonksiyonlarının hepsini aynı anda optimize etmek mümkün değildir. Amaç fonksiyonlarının birbiriyle çelişmesi durumunda etkin, uzlaşık, baskın çözümler olarak adlandırılan Pareto çözümler çok amaçlı modelin optimizasyonunu karakterize eder [25].

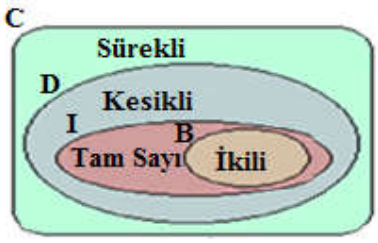
ÇAO problemi genel halde

$$\begin{aligned} min \,/\, max \;\; & \mathbf{f} = [f_1(\boldsymbol{x}) \; f_2(\boldsymbol{x}) \ldots f_r(\boldsymbol{x})] \\ & h_i(\boldsymbol{x}) = 0 \quad , \quad i = 1,2,\ldots,k \\ & g_j(\boldsymbol{x}) \{\leq, \geq\} 0 \quad , \quad j = 1,2,\ldots,m \\ & \boldsymbol{x} \in \mathrm{R}^n \end{aligned} \tag{8}$$

biçiminde tanımlanır. Burada, $\boldsymbol{x} = [x_1 x_2 \ldots x_n]$ biçiminde tanımlı girdi değişkenleri vektörü, $\mathbf{f} = [f_1 \; f_2 \; ... \; f_r]$ amaç fonksiyonu vektörü, $h_i, i = 1,2,...,k$ eşitlik ve $g_j, j = 1,2,...,m$ eşitsizlik biçiminde tanımlı kısıt fonksiyonlarıdır. Optimizasyon problemlerinde değişkenlerin bazılarının sürekli ve bazılarının kesikli değerli olması durumunda problem, karma kesikli optimizasyon problemi olarak adlandırılır. Genel kısıtlı karma kesikli ÇAO problemi

$$\begin{aligned} min \,/\, max \;\; & [f_1(\boldsymbol{x}) \; f_2(\boldsymbol{x}) \ldots f_r(\boldsymbol{x})] \\ & h_i(\boldsymbol{x}) = 0 \quad , \quad i = 1,2,\ldots,k \\ & g_j(\boldsymbol{x}) \{\leq, \geq\} 0 \quad , \quad j = 1,2,\ldots,m \\ & \boldsymbol{x} \in \mathrm{R}^{\mathrm{B}} \cup \mathrm{R}^{\mathrm{I}} \cup \mathrm{R}^{\mathrm{D}} \cup \mathrm{R}^{\mathrm{C}} \end{aligned} \tag{9}$$

biçiminde olur. Burada, $\mathrm{R}^{\mathrm{B}}, \mathrm{R}^{\mathrm{I}}, \mathrm{R}^{\mathrm{D}}$ ve $\mathrm{R}^{\mathrm{C}}$ sırasıyla ikili, tam sayılı, kesikli ve sürekli değişkenlerin destek kümeleridir. Hem $\mathrm{R}^{\mathrm{D}}$ hem de $\mathrm{R}^{\mathrm{C}}$ mevcut olduğunda problem karma kesikli optimizasyon problemine dönüşür. Şekil 1'de karma veriler için ilişki şeması yer almaktadır. Şekil 1'den görüldüğü gibi tam sayılı ve ikili değişkenler, kesikli değişkenlerin bir alt kümesi olarak kabul edilir. Bu nedenle, hem tam sayılı hem de kesikli değişkenleri ifade etmek için genel olarak kesikli değişkenler terimi kullanılır.




**Şekil 1.** Karma veriler için ilişki şeması

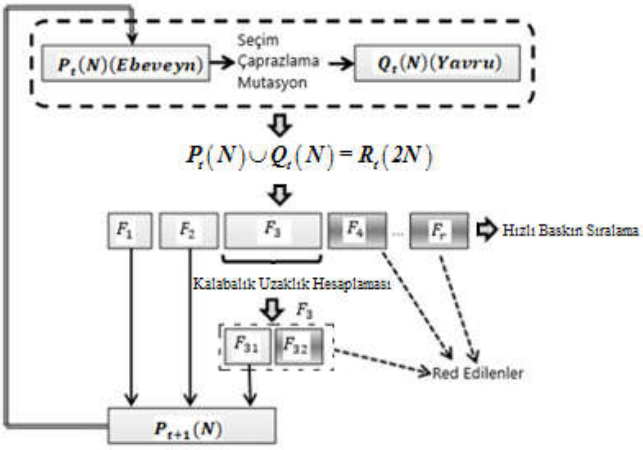
*3.1. NSGA-II*

ÇAO problemlerinde bütün amaç fonksiyonlarına aynı anda optimal değer veren bir çözüm vektörü bulunamayacağından ÇAO'da optimal çözüm yerine etkin, uzlaşık ve baskın çözümler olarak adlandırılan Pareto çözümler önem kazanır. Pareto çözüm kümesi, amaç fonksiyonları için uzlaşık çözümler içerir. ÇAO problemlerinin çözümü için birçok yapay zeka optimizasyon algoritması geliştirilmiştir. NSGA-II, en çok kullanılan popülasyon tabanlı bir yapay zeka optimizasyon algoritmasıdır. İlk olarak Goldberg [26] tarafından önerilen NSGA-II'nin Srinivas ve Deb'in [27] uygulamaları ile eksiklikleri giderilerek Deb ve ark. [28] tarafından geliştirilip gerçek Pareto yüzey yakınında çok daha iyi çözüm dağılımı veren bir algoritma olarak literatüre kazandırılmıştır. NSGA-II, ayarlanabilir parametre değerleri belirlendikten sonra bir başlangıç çözüm popülasyonu ile aramaya başlar. Uzlaşık çözümleri bulmak için genetik operatörleri kullanarak arama uzayında olasılıksal bir keşif uygular. Ayarlanabilir parametre değerlerinin seçimi ve kullanıcı tarafından seçilen operatörler, GA'nın performansı üzerinde oldukça etkilidir. Bu parametrelerin her kombinasyonu farklı optimum çözümlerle sonuçlanabilir. Ayarlanabilir parametrelerin farklı kombinasyonlarının uygun biçimde belirlenmesi amacıyla Türkşen ve Akgün [29] çalışmalarında Taguchi tasarımını kullanmışlardır. Ardışık nesillerdeki bireysel çözümlerin uygunluğu seçim, çaprazlama ve mutasyon yoluyla artırılır [26]. Çizelge 3'te GA'da kullanılan bazı terimlerin optimizasyon terminolojisindeki karşılıkları verilmiştir.

**Çizelge 3.** GA'da kullanılan terimlerin optimizasyondaki karşılıkları [30]

| **Biyoloji / Genetik** | **Optimizasyon** |
|---|---|
| Popülasyon | Aday çözümler kümesi |
| Birey / Kromozom | Kodlanmış aday çözüm |
| Gen | Tasarım / Karar değişkeni |
| Uygunluk fonksiyon değeri | Amaç fonksiyon değeri |
| Çevre | Kısıtlar |
| Kuşak / Generasyon / Nesil | Döngü / İterasyon / Yineleme |

NSGA-II'nın diğer çok amaçlı GA'lara göre üstünlüğü, hızlı baskın sıralama ve kalabalık uzaklığı yaklaşımlarıdır. Bu yaklaşımlarla sıralı baskın yüzeyler oluşturularak, baskın çözümler kümesi olarak bilinen Pareto çözüm kümesinde farklı seçenek çözümler elde edilir. NSGA-II'de Pareto çözümlerin sıralanmasında, her bir çözüm için baskınlık sayacı ve ilgili çözümün baskın olduğu çözüm kümesi kullanıldığından çözümlerin sıralanması NSGA'ya göre daha hızlı olmaktadır. Bu nedenle tanımlanan sıralama algoritması hızlı baskın sıralama algoritması olarak adlandırılır. Pareto çözüm kümesi içerisinde çözümlerin dağılımı ve çeşitliliği için kalabalık uzaklığı yaklaşımı kullanılır. Bir çözümün kalabalık uzaklığı, o çözümün bulunduğu yüzeyde ilgili çözüm ile komşu çözümleri arasındaki uzaklıktır [31]. Şekil 2'de, NSGA-II'nin hızlı baskın sıralama ve kalabalık uzaklık işleyişi görülmektedir. Mevcut popülasyon ($P_t(N)$) ile oluşturulan yavru popülasyon ($Q_t(N)$) birleştirilir, $P_t(N) \cup Q_t(N) = R_t(2N)$. Oluşturulan $R_t(2N)$ popülasyonundaki tüm çözümlerin birlikte değerlendirilmesi ile seçkinlik (elitizm) işlemi gerçekleştirilir. Hızlı baskın sıralama algoritması kullanılarak, $R_t(2N)$'deki baskın $F_1, F_2, ..., F_r$ yüzeyleri belirlenir. Tüm yüzeylerde sıralanan çözümlerin kalabalık uzaklığı hesaplanır. İkili turnuva seçim yöntemi kullanılarak $R_t(2N)$'deki ilk $N$ birey seçilerek yeni nesil ebeveyn popülasyonu $P_{t+1}(N)$ oluşturulur [32].




**Şekil 2.** NSGA-II'nin hızlı baskın sıralama ve kalabalık uzaklık hesaplama işleyişi

ÇAO problemlerinin NSGA-II ile çözümü için adımları aşağıdaki gibidir.

***Adım 0:*** *NSGA-II'nin ayarlanabilir parametrelerinin belirlenmesi*

NSGA-II'nin ayarlanabilir parametre değerleri girdi değişken sayısı ( $p$ ), popülasyon büyüklüğü ( $N$ ), çaprazlama olasılığı ( $Pr_{\text{cr}}$ ), çaprazlama indeksi ( $\eta_{cr}$ ), mutasyon olasılıkları ( $Pr_{\text{mut}}$ ) ve mutasyon indeksi ( $\eta_{mut}$ ), yineleme sayısı ( $t = 1, 2, ..., n_{\text{gen}}$ ) uzman görüşüne göre tanımlanır. Burada, yineleme sayısı NSGA-II için durdurma koşulu olarak kullanılır.

***Adım 1:*** *Başlangıç popülasyonunun oluşturulması ve uygunluk fonksiyonu değerlerinin hesaplanması*

Değişkenler reel değerli alınarak başlangıç popülasyonu rastgele oluşturulur ve bireylerin uygunluk fonksiyon değerleri hesaplanır.

***Adım 2:*** *Uygunluk fonksiyon değerlerine göre bireylerin sıralanması ve kalabalık uzaklık değerlerinin hesaplanması*

Bireyler uygunluk fonksiyon değerlerine göre sıralanır ve kalabalık uzaklık değerleri hesaplanır. Bireyler sıralamaya ve kalabalık uzaklığa göre seçildiğinden popülasyondaki tüm bireylere bir kalabalık uzaklık değeri atanır.

***Adım 3:*** *Genetik operatörlerin uygulanması*

Genetik operatör uygulamaları seçim, çaprazlama ve mutasyon aşamalarından oluşur. *Adım 0*'da belirtilen ayarlanabilir parametre değerleri kullanılarak genetik operatör işlemleri uygulanır.

***Adım 3.1:*** Baskınlık kriterine göre sıralanan bireylere kalabalık uzaklık değeri atandıktan sonra seçim, bir kalabalık karşılaştırma operatörü $(\prec)$ kullanılarak gerçekleştirilir. Bireyler, ikili turnuva seçimi kullanılarak seçilir. Seçim, iki kritere göre yapılır. Birincisi ve en önemlisi, çözümlerin bulunduğu yüzey veya sıradır. Daha düşük sıralamaya sahip bireyler seçilir. İkinci olarak, iki bireyin sıralaması aynı ise, kalabalık uzaklığı karşılaştırılır. Kalabalık uzaklığı daha büyük olan bireyler seçilir.

***Adım 3.2:*** Seçilen ebeveynlerden Simüle Edilmiş İkili (SBX) çaprazlama operatörü kullanılarak yavrular üretilir.

***Adım 3.3:*** SBX çaprazlama operatörü ile üretilen bazı yavrulara polinomsal mutasyon operatörü uygulanır.

***Adım 4:*** *Yeni popülasyonun oluşturulması*

$N$ ve $2N$ birey sayısı içeren yeni popülasyonlar oluşturulur.

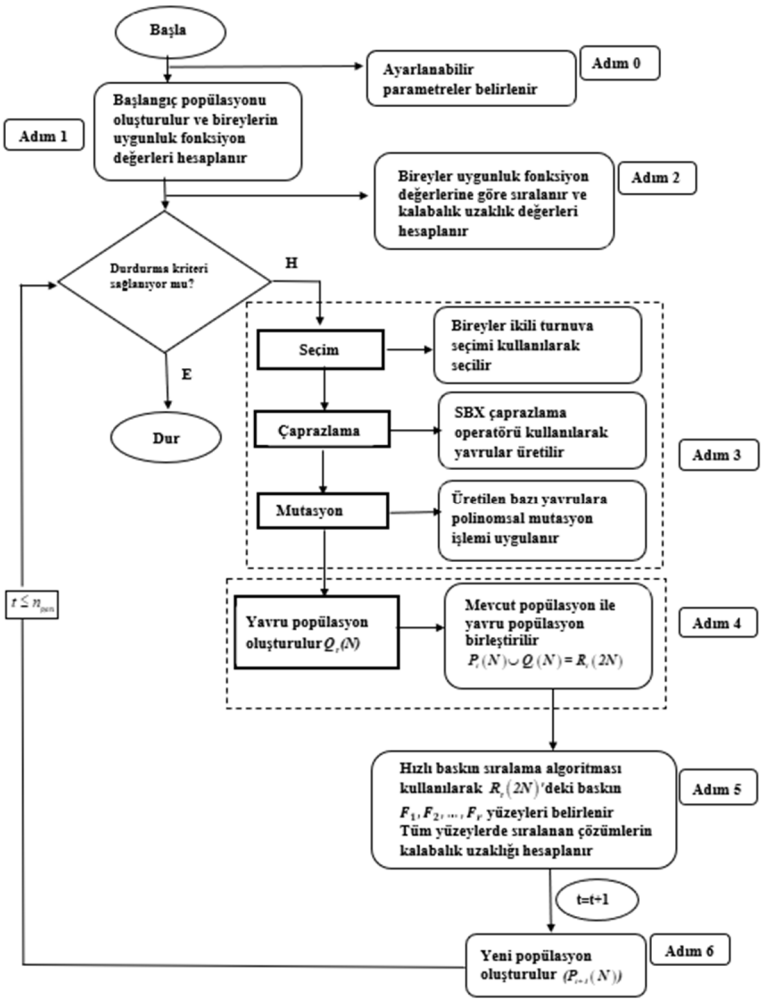
***Adım 4.1:*** Yeni $Q_t(N)$ (yavru) popülasyonu oluşturulur.

***Adım 4.2:*** Mevcut popülasyon ile oluşturulan yavru popülasyon birleştirilir, $P_t(N) \cup Q_t(N) = R_t(2N)$. Oluşturulan $R_t(2N)$ popülasyonundaki tüm çözümlerin birlikte değerlendirilmesi ile seçkinlik işlemi gerçekleştirilir.

***Adım 5:*** *Baskın yüzeylerin oluşturulması ve oluşturulan yüzeylerde sıralanan çözümlerin kalabalık uzaklıklarının hesaplanması*

Hızlı baskın sıralama algoritması kullanılarak, $R_t(2N)$'deki baskın $F_1, F_2, \ldots, F_r$ yüzeyleri belirlenir. Tüm yüzeylerde sıralanan çözümlerin kalabalık uzaklığı hesaplanır. Kalabalık uzaklık değerlerine göre başlangıçta belirlenen popülasyon sayısı ($N$) elde edilecek biçimde sıralı yüzeylerden bireyler seçilerek yeni popülasyon oluşturulur.

***Adım 5.1:*** İkili turnuva seçim yöntemi kullanılarak $R_t(2N)$'deki ilk $N$ birey seçilerek yeni nesil ebeveyn popülasyon $P_{t+1}(N)$ oluşturulur. Başlangıçta belirlenen yineleme sayısına ulaşılmış ise algoritma sonlandırılır. Aksi halde, Adım 3'e gidilir. Şekil 2'de hızlı baskın sıralama ve kalabalık uzaklık hesaplama işleyişi gösterilmiştir. NSGA-II'nin akış şeması Şekil 3'te verilmiştir.




**Şekil 3.** NSGA-II akış şeması

### *3.2. MDNSGA-II*

NSGA-II'nin, değişken gösterimi, başlangıç popülasyonu ve genetik operatörlere ilişkin adımlarında uyarlamalar yapılarak karma veri içeren çok yanıtlı problemler için Pareto çözüm kümesi elde edilmeye çalışılmıştır. Yapılan uyarlamalarla algoritma, MDNSGA-II olarak adlandırılmıştır.

#### *3.2.1 Değişkenlerin gösterimi*

Değişkenlerin gösteriminde, problemin doğasına ve verilerin yapısına göre kodlama türü seçilir. Başlıca kodlama türleri, ikili, sekizli, onaltılı, gri, permütasyon kodlama, değer kodlaması ve ağaç kodlamasıdır. Literatürde karma veri içeren problemler için ikili kodlama tercih edilmiştir. Çözümlerin hesaplama süresini kısaltması ve karma veri içeren problemler için uyarlanabilmesi sebebiyle bu çalışmada değer kodlaması kullanılmıştır. Değer kodlamasında sürekli ve tam sayılı değişkenlerin kendi değeriyle çalışılır. Fakat, değişkenin kesikli reel değer alması durumunda değişken gösteriminde uyarlama yapılması gerekir. Kesikli değişkenin tanım kümesindeki her bir değer küçükten büyüğe sıralanır. Sıralanan değerler sıra numarası ile indekslenir. Böylece kesikli değişkenin her bir değeri bir indeks değeriyle ifade edilmiş olur. Yapılan bu uyarlama, kesikli reel değer alan değişkenlerin indekslenmesi olarak tanımlanır.

Örneğin, kesikli bir değişkenin tanım kümesi $S \subset \mathrm{R}^{\mathrm{D}}$, kesikli değişkenin aldığı değer kümesi $S = \{s_1, s_2, ..., s_q\}$ olmak üzere, kesikli değişkenin her bir değerinin $I$ tam sayılar kümesinde pozitif bir tam sayıya eşlenmesi

$$\begin{array}{l} S = \{s_1, s_2, s_3, ..., s_q\} \\ \quad\quad \downarrow \downarrow \downarrow \ldots \downarrow \\ I = \{1,\ 2,\ 3, \ldots, q\} \end{array} \tag{10}$$

biçiminde gösterilir. Buna göre, $s_1$ için 1, $s_2$ için 2, …, $s_q$ için q indekslemesi yapılır. Böylece kesikli değişkenin aldığı değerler yerine değişkenin indeksleri kullanılır.

#### *3.2.2 Başlangıç popülasyonunun oluşturulması*

MDNSGA-II için başlangıç popülasyonu rastgele sayı üreteci kullanılarak oluşturulur. $\alpha$ , [0,1] aralığından üretilen bir rastgele sayı olmak üzere, *lb*, sürekli veya tam sayılı değişkenin alt sınırı ve *ub*, sürekli veya tam sayılı değişkenin üst sınırı olsun. Sürekli ve tam sayılı değişkenler için belirlenen [*lb*, *ub*] tanım aralıklarında sırasıyla

$$x = lb + \alpha(ub - lb) \tag{11}$$

ve

$$x = lb + round(\alpha(ub - lb)) \tag{12}$$

biçiminde tanımlı eşitliklerle başlangıç popülasyonu bileşenleri elde edilir. İkili (0,1) değerli değişkenler için ise

$$x = round(\alpha) \tag{13}$$

eşitliği ile başlangıç popülasyonu bileşen değerleri elde edilir. Kesikli değişkenler için başlangıç popülasyonu oluşturulurken değişken gösteriminde yapılan indeksleme işlemine göre uyarlama yapılır.

Kesikli değişken değerlerinin indeksleri için rastgele sayı üretilir. $r$ sıra numarası ve tam sayı eşlemesi yapılan kesikli değişkenin tanım aralığı [1, $q$] olmak üzere kesikli değişken için üretilecek rastgele sayı

$$x^{(r)} = round(1 + \alpha(q-1)) \tag{14}$$

biçiminde elde edilir. $S = \{s_1, s_2, ..., s_q\}$ kesikli değişken değerleri kümesi olmak üzere #($S$)=$q$'dur. Buna göre [1, $q$] aralığından $q$ sayıda rastgele indeks numaraları ($I_q$) üretilir. Üretilen $I_q$ indeks numaralarına göre, $S$ kümesinde karşılık gelen kesikli değer seçilerek yeni bir kesikli değer kümesi oluşturulur.

*3.2.3 Genetik operatörlerin uyarlanması*

Genetik operatörler seçim, çaprazlama ve mutasyon aşamalarından oluşur. MDNSGA-II'de seçim operatörü adımında herhangi bir uyarlama yapmaya gerek yoktur. Bu çalışmada, uyarlanan NSGA-II genetik operatörleri olarak Turnuva seçimi, SBX çaprazlama ve Polinomsal mutasyon kullanılmıştır. Genetik operatörlerin çaprazlama ve mutasyon aşamalarında yapılan uyarlamalar aşağıda verilmiştir.

***SBX çaprazlama operatörü*:** Bu operatör reel değerli değişkenler için oluşturulmuştur. $t$. yinelemede mevcut $x^{(1,t)}$ ve $x^{(2,t)}$ çözümleri kullanılarak yapılan hesaplamalarla iki yeni çözüm elde edilir. $U \in [0,1]$ aralığından rastgele bir sayı oluşturulur. Burada $\eta_{cr}$, belirlenen bir aralıkta tanımlı, pozitif reel sayı değerli çaprazlama dağılım indeksi olmak üzere

$$\beta_q = \begin{cases} (2U)^{\frac{1}{\eta_{cr}+1}}, & U \le 0.5 \\ \left(\frac{1}{2(1-U)}\right)^{\frac{1}{\eta_{cr}+1}}, & d.y. \end{cases} \tag{15}$$

ile tanımlı fonksiyona göre $\beta_q$ hesaplanır. Yavru çözümler ise

$$\begin{aligned} x^{(1,t+1)} &= 0.5\left[\left(1+\beta_q\right)x^{(1,t)} + \left(1-\beta_q\right)x^{(2,t)}\right] \\ x^{(2,t+1)} &= 0.5\left[\left(1-\beta_q\right)x^{(1,t)} + \left(1+\beta_q\right)x^{(2,t)}\right] \end{aligned} \tag{16}$$

biçiminde elde edilir. Sürekli değişkenler için Eşitlik (16) kullanılarak yavru çözümler oluşturulur.

MDNSGA-II'de çaprazlama operatörü kullanılarak oluşturulan yavru çözümler, tam sayı ve kesikli değişkenin tanım kümesi dışında değerler alabilir. Bu nedenle, yavru çözüm oluşturulurken çaprazlama operatöründe yapılacak değişikliklerle değişkenin tanım kümesi içinde değer alması sağlanır. Tam sayılı değişkenler için yavru çözümler

$$\begin{aligned} x^{(1,t+1)} &= round\left(0.5\left[\left(1+\beta_q\right)x^{(1,t)} + \left(1-\beta_q\right)x^{(2,t)}\right]\right) \\ x^{(2,t+1)} &= round\left(0.5\left[\left(1-\beta_q\right)x^{(1,t)} + \left(1+\beta_q\right)x^{(2,t)}\right]\right) \end{aligned} \tag{17}$$

biçiminde oluşturulur. Burada, tam sayılı değerlerin ardışık olması durumunda Eşitlik (17)'nin kullanımı uygundur. Aksi halde kesikli reel sayı değerli değişkenler için önerilen uyarlamanın yapılması uygun olur. Eşitlik (16)'da SBX çaprazlama operatörünün uygulanmasıyla elde edilen yavru çözümler kesikli değişkenlerin tanım kümesinde yer almayabilir. Bu durumda elde edilen yavru çözüm değerlerinin

$(x^{(1,t+1)}$ ve $x^{(2,t+1)})$, $S=\{s_1,s_2,...,s_q\}$ kümesindeki elemanlardan biri olması istenir. Eğer, $x^{(1,t+1)} \in [s_i, s_{i+1}], i=1,2,...,q-1$ ise $s_i' = \frac{s_i + s_{i+1}}{2}$ hesaplanır. $s_i'$ değeri ile Eşitlik (16) ile elde edilen yavru çözüm değerleri karşılaştırılır. Yapılan karşılaştırmada yavru çözüm değerinin $s_i$ ya da $s_{i+1}$, $i=1,2,...,q-1$ değerlerinden hangisine daha yakın olduğu belirlenir ve yavru çözüm değerlerine yakın olan kesikli değişken değeri alınır. Böylece, çaprazlama aşaması sonrasında elde edilen yavru çözüm

$$x^{(1,t+1)} = \begin{cases} s_i & , \quad 0.5\left[\left(1+\beta_q\right)x^{(1,t)} + \left(1-\beta_q\right)x^{(2,t)}\right] \le s_i' \\ s_{i+1} & , \quad d.y. \end{cases} \tag{18}$$

biçiminde elde edilir. $x^{(2,t+1)}$'de Eşitlik (18)'e benzer olarak hesaplanır.

***Polinomsal mutasyon operatörü:*** SBX ile elde edilen yavru çözümlere mutasyon aşaması uygulanarak çözüm çeşitliliği sağlanır. Mevcut çözüm $x^{(t)}$'den, yeni çözüm

$$x^{(t+1)} = x^{(t)} + (ub - lb)\delta \tag{19}$$

biçiminde elde edilir. Burada, $lb$ ve $ub$, sırasıyla çözümler için belirlenen alt ve üst sınır değerleri, $r \in [0,1]$ aralığından rastgele bir sayı ve $\eta_{mut}$ mutasyon dağılım indeksi olmak üzere, bir polinomsal dağılım kullanılarak Eşitlik (19)'da belirtilen değişim

$$\delta = \begin{cases} (2r)^{\frac{1}{\eta_{mut}+1}} - 1 \ , \ r \le 0.5 \\ 1 - \left[2\left(1-r\right)\right]^{\frac{1}{\eta_{mut}+1}}, \ d.y. \end{cases} \tag{20}$$

biçiminde hesaplanır [30]. Sürekli değişkenler için Eşitlik (19) kullanılarak yeni çözüm elde edilir.

MDNSGA-II'de mutasyon aşaması uygulanarak oluşturulan yeni çözümlerde yapılacak değişikliklerle değişkenin tanım kümesi içinde değer alması sağlanır. Çaprazlamada yapılan uyarlamalar benzer şekilde mutasyon aşamasında da yapılarak tam sayılı değişkenler için yeni çözüm

$$x^{(t+1)} = round(x^{(t)} + (ub - lb)\delta) \tag{21}$$

ve kesikli değişkenler için $x^{(t+1)} \in [s_i, s_{i+1}], i=1,2,...,q-1$ ise $s_i' = \frac{s_i + s_{i+1}}{2}$ hesaplanarak mutasyon aşaması sonrasında elde edilen yeni çözüm

$$x^{(1,t+1)} = \begin{cases} s_i & , \quad x^{(t)} + (ub - lb)\delta \le s_i' \\ s_{i+1} & , \quad d.y. \end{cases} \tag{22}$$

biçiminde elde edilir. Şekil 3'te verilen NSGA-II akış şemasında Adım 0 (Değişkenlerin gösterimi), Adım 1 (Başlangıç popülasyonunun oluşturulması) ve Adım 3'te (Genetik operatörler) yapılan uyarlamalarla MDNSGA-II oluşturulur.

## 4. Uygulama

Karma veri içeren çok yanıtlı optimizasyon problemlerinin çözümü için bu çalışmada önerilen MDNSGA-II yöntemi, literatürden çok iyi bilinen iki veri seti (enerji ve gıda alanı) üzerinde uygulanmıştır. İlk veri seti, optimizasyon alanındaki çalışmalarda kabul gören veri tabanından alınan enerji verimliliği ile ilgilidir

[33]. Bu veri setinde, binaların enerji verimliliğine ilişkin ısıtma yükü ($Y_1$) ve soğutma yükü ($Y_2$) gereksinimlerinin, bina özniteliklerinin bir fonksiyonu olarak değerlendirilmesi araştırılmıştır. İkinci veri seti ise, gıda alanında literatürde tanımlı çok yanıtlı deneysel çalışma ile edilen bir veri setidir. Veri setleri için GLM ve SUR yöntemleri ile uygun tahmini yanıt modelleri oluşturularak, karma veriler için modelleme aşaması sağlanmıştır. Veri setlerinde yer alan yanıt değişkenlerinin modellenmesinde RStudio ve IBM SPSS Statistics 24 programları kullanılmıştır. Optimizasyon aşamasında, MDNSGA-II yöntemi ile MATLAB R2023a programı kullanılarak veri setleri için Pareto çözüm kümesi elde edilmiştir.

### *4.1. Enerji verimliliği veri seti*

Veri kümesi, iki gerçek değerli yanıta (ısıtma yükü ve soğutma yükü) ilişkin 768 gözlem ve 8 özellik (girdi değişkenleri) içermektedir. Girdi değişkenleri ile yanıtlar arasındaki fonksiyonel ilişkinin belirlenmesi ve yanıt değişkenlerini eş anlı minimum yapacak girdi değişken değerlerinin belirlenmesi istenmektedir. Çizelge 4'te, 8 girdi değişkeni ve 2 yanıt değişkeninden oluşan 768 gözlemli karma veri seti örneği yer almaktadır.

**Çizelge 4.** Enerji verimliliği veri seti

| *No* | ***Girdi Değişkenler*** | | | | | | | | ***Yanıt Değişkenleri*** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $\boldsymbol{Y_1}$ | $\boldsymbol{Y_2}$ |
| **1** | 0.98 | 514.5 | 294 | 110.25 | 7 | 2 | 0 | 0 | 15.55 | 21.33 |
| **2** | 0.98 | 514.5 | 294 | 110.25 | 7 | 3 | 0 | 0 | 15.55 | 21.33 |
| **3** | 0.98 | 514.5 | 294 | 110.25 | 7 | 4 | 0 | 0 | 15.55 | 21.33 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **766** | 0.62 | 808.5 | 367.5 | 220.5 | 3.5 | 3 | 0.4 | 5 | 16.44 | 17.11 |
| **767** | 0.62 | 808.5 | 367.5 | 220.5 | 3.5 | 4 | 0.4 | 5 | 16.48 | 16.61 |
| **768** | 0.62 | 808.5 | 367.5 | 220.5 | 3.5 | 5 | 0.4 | 5 | 16.64 | 16.03 |

Çizelge 5'te girdi ve yanıt değişkenlerinin değişken türleri ve tanım aralığı görülmektedir. Girdi değişkenlerinden yapı yüksekliği ve yönelimi ile cam alanı ve cam alan dağılımı açısından bina özniteliklerinin farklılık gösterdiği görülmektedir. Yapı yüksekliği ($X_5$), sürekli bir değişken olmasına rağmen bu veri setinde tanım kümesinin $\{3.5, 7\}$ değerlerinden oluştuğu görülmektedir. Bu durum yapı yüksekliğinin bu veri seti için kesikli değişken olarak ele alındığını göstermektedir. Bina tasarımında yapı yüksekliği sadece iki değerden birini alabilmektedir. $X_6$ değişkeni yapının hangi cephede (kuzey, güney, doğu, batı) yer aldığını belirtmektedir. $X_5, X_6, X_7$ ve $X_8$ değişkenlerinin tanım kümelerine bakıldığında kesikli ve/veya tam sayılı değişkenlerden oluştuğu görülmektedir.

**Çizelge 5.** Enerji verimliliği veri setinin girdi değişkenleri, değişken türü ve tanım kümesi

| | ***Değişkenler*** | ***Değişken Türü*** | ***Tanım Kümesi*** |
|---|---|---|---|
| Girdi | Bağıl kompaktlık ($X_1$) | Sürekli | [0.62, 0.98] |
| | Yüzey alanı ($X_2$) | Sürekli | [514.5, 808.5] |
| | Duvar alanı ($X_3$) | Sürekli | [245, 416.5] |
| | Çatı alanı ($X_4$) | Sürekli | [110.25, 220.50] |
| | Yapı yüksekliği ($X_5$) | Kesikli | {3.5, 7} |
| | Yapı yönelimi ($X_6$) | Kesikli, Tam sayı | {2, 3, 4, 5} |
| | Cam alanı ($X_7$) | Kesikli | {0, 0.10, 0.25, 0.40} |
| | Cam alan dağılımı ($X_8$) | Kesikli, Tam sayı | {0, 1, 2, 3, 4, 5} |
| Yanıt | Isıtma yükü ($Y_1$) | Sürekli | $Y_1 \geq 0$ |
| | Soğutma yükü ($Y_2$) | Sürekli | $Y_2 \geq 0$ |

Binaların enerji tüketimi göz önünde bulundurulduğunda enerjinin kullanımı için verimli bina tasarımı önemlidir. İhtiyaç duyulan ısıtma ve soğutma ekipmanının özelliklerini belirlemek için ısıtma ve soğutma yüklerinin hesaplanması gerekir. Isıtma ve soğutma yüklerinin enerji verimliliği bakımından minimum olması istenir. Bu çalışmada, Tsanas ve Xifara'nın [34] çalışmasında uygulanan IRLS (Iteratively Reweighted Least Squares) yöntemi ile elde edilen regresyon modeli kullanılmıştır. Varyans homojenliği varsayımının sağlanamadığı durumlarda IRLS'nin kullanımı daha uygundur. IRLS ile regresyon katsayılarındaki ağırlıklar ayarlanarak regresyon eğrisi oluşturulurken aykırı değerlerin etkisi azaltılır. Böylece daha geliştirilmiş bir en küçük kareler tahmini sağlanır.

Çizelge 6'da, $Y_1$ ve $Y_2$ yanıtlarının normallik varsayımı için Kolmogrov-Smirnov testi sonuçları görülmektedir. %95 güven düzeyinde (*p-değeri=0.001<0.05*) $Y_1$ ve $Y_2$ yanıtlarına ilişkin gözlem değerlerinin Normal dağılım göstermediği söylenir.

**Çizelge 6.** Yanıtlar için Kolmogrov-Smirnov testi sonuçları

| ***Yanıt Değişkenleri*** | ***p-değeri*** |
|---|---|
| $Y_1$ | .001 |
| $Y_2$ | .001 |

$Y_1$ ve $Y_2$ yanıtları için elde edilen tahmini yanıt modelleri sırasıyla

$$\hat{Y}_1 = -4.75X_1 - 0.03X_2 + 0.07X_3 - 3.44X_5 - 0.01X_6 + 18.13X_7 + 0.09X_8 \tag{23}$$

ve

$$\hat{Y}_2 = -9.02X_1 - 0.01X_2 + 0.04X_3 - 4.30X_5 - 0.12X_6 + 14.49X_7 + 0.03X_8 \tag{24}$$

dır. $\hat{Y}_1$ ve $\hat{Y}_2$ yanıtlarını eş anlı minimize eden optimal girdi değerlerini elde edebilmek için optimizasyon aşamasında MDNSGA-II uygulanmıştır. Her bir yanıt fonksiyonu bir amaç fonksiyonu olarak değerlendirildiğinde çok yanıtlı optimizasyon problemi

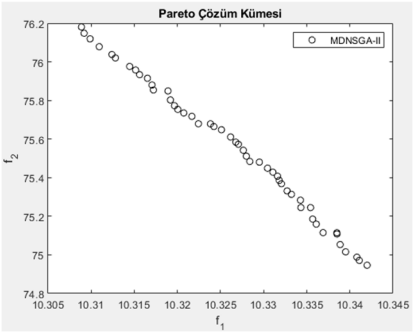
$$\begin{aligned} &min\ f_1 = \hat{Y}_1(\boldsymbol{x}) \\ &min\ f_2 = \hat{Y}_2(\boldsymbol{x}) \\ &\quad \boldsymbol{x} \in S \end{aligned} \tag{25}$$

biçiminde tanımlı ÇAO problemine dönüşecektir. Burada, $\{X_1, X_2, ..., X_8\}$ değişkenlerinin aldığı değerler vektörü $\boldsymbol{x}$ olarak belirtilmiştir. $f_1$ ve $f_2$ amaç fonksiyonları, sırasıyla $\hat{Y}_1$ ve $\hat{Y}_2$ tahmini yanıt fonksiyonlarına karşılık gelmektedir. Çizelge 7'de, bu çalışmada uygulanan MDNSGA-II'nin ayarlanabilir parametreleri ve parametrelerin aldığı değerler yer almaktadır.

**Çizelge 7.** MDNSGA-II'nin ayarlanabilir parametreleri ve parametre değerleri

| ***Ayarlanabilir parametreler*** | ***Parametre değerleri*** |
|---|---|
| Girdi değişken sayısı ($p$) | 8 |
| Popülasyon büyüklüğü ($N$) | 50 |
| Yineleme sayısı ($n_{gen}$) | 100 |
| Çaprazlama operatörü | SBX |
| Mutasyon operatörü | Polinomsal |
| Seçim operatörü | Turnuva |
| Çaprazlama olasılığı ($Pr_{cr}$) | 0.90 |
| Çaprazlama indeksi ($\eta_{cr}$) | 20 |
| Mutasyon olasılığı ($Pr_{mut}$) | 1/8 |
| Mutasyon indeksi ($\eta_{mut}$) | 20 |

Şekil 4'te enerji veri seti için elde edilen Pareto çözüm kümesi görülmektedir. Karma veri içeren çok yanıtlı enerji veri seti için MDNSGA-II ile çözüm çeşitliliğinin sağlandığı görülmektedir. Enerji veri seti için $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin Pareto çözüm değerleri Ek-1'de verilmiştir. Ek-1'de görüldüğü gibi, MDNSGA-II ile $\{X_5, X_6, X_7, X_8\}$ değişkenlerinin tanım kümesinden kesikli değerler alarak Pareto çözümlerin elde edilmesi sağlanmıştır.




**Şekil 4.** Enerji veri seti için $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin Pareto çözüm kümesi

*4.2. Gıda veri seti*

Bu uygulamada veri seti, Schmidt ve ark. [35] çalışmalarında kullandığı gıda alanından seçilmiştir. Karma veri modelleri için deneyin optimal değeri elde edilmeye çalışılmıştır. Çok yanıtlı deneysel karma veri seti 2 girdi değişkeni ve 4 yanıt değişkeni içermektedir. Sistein jel dokusu $(X_1)$ ve bir tuz çeşidi olan kalsiyum klorür (CaCl2) $(X_2)$ maddelerinin peynir altı suyu konsantresinin yapısal özellikleri ve su tutma özelliği üzerine etkisini incelemek amacıyla bir deney düzenlenmiştir. Burada, sistein ve kalsiyum klorür girdi değişkenleri; sertlik, yapışkanlık, esneklik yapısal özellikleri ile peynir altı suyunda tutulan sıkıştırılabilir su yanıt değişkenleridir. Peynir altı suyu proteini konsantresi sistemlerinin yanıt değişkenleri üzerindeki etkilerini ölçmek için çoklu doğrusal regresyon analizi kullanılmıştır. Çok yanıtlı deneysel karma veri seti için yanıtlar arasında doğrusal ilişkili olması durumunda yanıtlar, SUR (Seemingly Unrelated Regression) yöntemi ile modellenebilir [36]. Çizelge 8'de deneysel çalışma sonucunda elde edilen kodlanmış girdi değişken değerleri ile gözlenen yanıt değişkenlerine ait değerler yer almaktadır.

**Çizelge 8.** Deneysel çalışma sonucu elde edilen kodlanmış girdi değişken değerleri ve gözlenen yanıt değişken değerleri

| ***No*** | $X_1$ | $X_2$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ |
|---|---|---|---|---|---|---|
| 1 | -1 | -1 | 2.48 | 0.55 | 1.95 | 0.22 |
| 2 | 1 | -1 | 0.91 | 0.52 | 1.37 | 0.67 |
| 3 | -1 | 1 | 0.71 | 0.67 | 1.74 | 0.57 |
| 4 | 1 | 1 | 0.41 | 0.36 | 1.20 | 0.69 |
| 5 | -1.414 | 0 | 2.28 | 0.59 | 1.75 | 0.33 |
| 6 | 1.414 | 0 | 0.35 | 0.31 | 1.13 | 0.67 |
| 7 | 0 | -1.414 | 2.14 | 0.54 | 1.68 | 0.42 |
| 8 | 0 | 1.414 | 0.78 | 0.51 | 1.51 | 0.57 |
| 9 | 0 | 0 | 1.50 | 0.66 | 1.80 | 0.44 |
| 10 | 0 | 0 | 1.66 | 0.66 | 1.79 | 0.50 |
| 11 | 0 | 0 | 1.48 | 0.66 | 1.79 | 0.50 |
| 12 | 0 | 0 | 1.41 | 0.66 | 1.77 | 0.43 |
| 13 | 0 | 0 | 1.58 | 0.66 | 1.73 | 0.47 |

Çizelge 9'da gıda veri setinin değişken türü ve tanım aralığı açıklamaları görülmektedir. Burada, $X_1$ kesikli değişken olup $X_2$ sürekli değişken olarak tanımlanmıştır.

**Çizelge 9.** Gıda veri seti için girdi değişkenleri, değişken türü ve tanım kümesi

| *Girdi Değişkenleri* | *Değişken Türü* | *Tanım Kümesi* |
|---|---|---|
| Sistein jel dokusu $(X_1)$ | Kesikli | {-1.414, -1, 0, 1, 1.414} |
| Kalsiyum klorür $(X_2)$ | Sürekli | [-1.414, 1.414] |

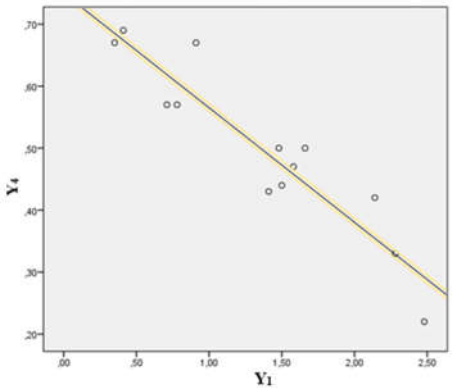
Yanıtlar arasındaki doğrusal ilişkinin incelenmesinden önce yanıtların Normal dağılım varsayımı test edilmiştir. Bu amaçla, Shapiro-Wilk (küçük örneklemlerde) testi uygulanmıştır. Shapiro-Wilk testi sonucu Çizelge 10'da özetlenmiştir. Çizelge 10'daki *p* değerlerine bakıldığında %95 güven düzeyinde $Y_1$ ve $Y_4$ yanıtlarının Normal dağılımlı olduğu (*p*-değeri>0.05), $Y_2$ ve $Y_3$ yanıtlarının dağılımının ise Normal dağılım göstermediği söylenir (*p*-değeri<0.05).

**Çizelge 10.** Gıda veri seti için yanıt değişkenlerinin Normallik testi

| *Yanıt Değişkenleri* | *Shapiro Wilk p-değeri* |
|---|---|
| $Y_1$ | **0.549** |
| $Y_2$ | 0.010 |
| $Y_3$ | 0.023 |
| $Y_4$ | **0.652** |

$Y_1 - Y_4$ yanıt değişkenleri için Pearson korelasyon katsayısı -0.932 olup bu yanıtlar arasında ters yönlü, güçlü ve anlamlı bir ilişki olduğu söylenir. Spearman korelasyon testine göre, $Y_1 - Y_3$, $Y_2 - Y_3$ ve $Y_3 - Y_4$ yanıtlarının da sırasıyla, 0.674, 0.69 ve -0.702 büyüklüklerinde doğrusal ilişkili olduğu söylenir. $Y_1 - Y_2$ ve $Y_2 - Y_4$ yanıtlarının ise doğrusal ilişkisiz olduğu görülmüştür. Bu çalışmada Normallik varsayımını sağladığı ve aralarında anlamlı ve güçlü bir ilişki olduğu için $Y_1 - Y_4$ yanıtları ile çalışılmıştır. Ayrıca, $Y_2 - Y_3$ yanıtları arasında anlamlı, pozitif yönlü doğrusal bir ilişki olması nedeniyle bu yanıtların da eş anlı optimizasyonu ile ilgilenilmiştir.

$Y_1 - Y_4$ yanıtlarına ilişkin saçılım grafiği Şekil 5'te verilmiştir. Yanıtlar arasındaki ilişki, ikinci dereceden polinomsal fonksiyonlar kullanılarak SUR yöntemi ile modellenmiştir. Çizelge 11'de, $Y_1 - Y_4$ yanıtlarının SUR yöntemi ile elde edilen parametre tahminleri yer almaktadır.




**Şekil 5.** $Y_1 - Y_4$ yanıt değişkenine ait saçılım grafiği

**Çizelge 11.** $Y_1$ ve $Y_4$ yanıtlarının SUR yöntemi ile parametre tahminleri

| ***Model Terimi*** | **SUR** | | | |
|---|---|---|---|---|
| | $Y_1$ | $p$-değeri | $Y_4$ | $p$-değeri |
| *Sabit* | 1.526 (.065) | **.001** | 0.468 (.013) | **.001** |
| $X_1$ | -0.575 (.051) | **.001** | 0.131 (.010) | **.001** |
| $X_2$ | -0.524 (.051) | **.001** | 0.073 (.010) | **.001** |
| $X_1^2$ | -0.171 (.055) | **.018** | 0.026 (.011) | .055 |
| $X_2^2$ | -0.098 (.055) | .12 | 0.024 (.011) | .076 |
| $X_1X_2$ | 0.318 (.073) | **.003** | -0.083 (.015) | **.001** |

*Standart hatalar parantez içinde belirtilmiştir.

Çizelge 11'den, $Y_1$ ve $Y_4$ yanıtları için elde edilen tahmini yanıt modelleri sırasıyla

$$\hat{Y}_1 = 1.526 - 0.575X_1 - 0.524X_2 - 0.171X_1^2 + 0.318X_1X_2 \tag{26}$$

ve

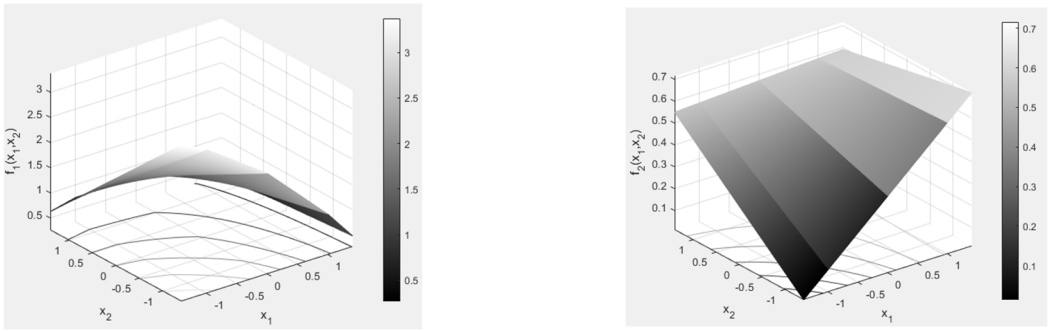
$$\hat{Y}_4 = 0.468 + 0.131X_1 + 0.073X_2 - 0.083X_1X_2 \tag{27}$$

dır. Optimal değeri elde edilmek istenilen her bir yanıt fonksiyonu, bir amaç fonksiyonu olarak değerlendirildiğinde çok yanıtlı optimizasyon problemi

$$\begin{aligned} &max\, f_1 = \hat{Y}_1(\boldsymbol{x}) \\ &min\, f_2 = \hat{Y}_4(\boldsymbol{x}) \\ &\boldsymbol{x} \in S \end{aligned} \tag{28}$$

biçiminde ÇAO problemine dönüşecektir. Burada, $\{X_1, X_2\}$ değişkenlerinin aldığı değerler vektörü $\boldsymbol{x}$ olarak belirtilmiştir. $f_1$ ve $f_2$ amaç fonksiyonları, sırasıyla $\hat{Y}_1$ ve $\hat{Y}_4$ tahmini yanıt fonksiyonlarına karşılık gelmektedir. Elde edilen tahmini yanıt fonksiyonları için yüzey grafikleri Şekil 6.(a)-(b)'de görülmektedir. Buna göre, Şekil 6. (a)'da bir maksimizasyon problemi ve Şekil 6.(b)'de bir minimizasyon problemi ile ilgilenildiği açıktır.

(a) (b)



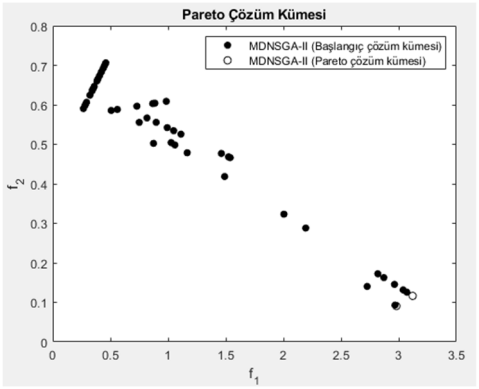

**Şekil 6.** (a) $f_1$ amaç fonksiyonu, (b) $f_2$ amaç fonksiyonu için yüzey grafikleri

Çizelge 12'de, bu çalışmada uygulanan MDNSGA-II'nin ayarlanabilir parametreleri ve parametrelerin aldığı değerler yer almaktadır.

**Çizelge 12.** MDNSGA-II'nin ayarlanabilir parametreleri ve parametre değerleri

| *Ayarlanabilir parametreler* | *Parametre değerleri* |
|---|---|
| Girdi değişken sayısı ($p$) | 2 |
| Popülasyon büyüklüğü ($N$) | 50 |
| Yineleme sayısı ( $n_{gen}$ ) | 100 |
| Çaprazlama operatörü | SBX |
| Mutasyon operatörü | Polinomsal |
| Seçim operatörü | Turnuva |
| Çaprazlama olasılığı ( $Pr_{cr}$ ) | 0.90 |
| Çaprazlama indeksi ( $\eta_{cr}$ ) | 20 |
| Mutasyon olasılığı ( $Pr_{mut}$ ) | 1/2 |
| Mutasyon indeksi ( $\eta_{mut}$ ) | 20 |

ÇAO probleminin $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin MDNSGA-II ile elde edilen başlangıç ve Pareto çözüm kümeleri Şekil 7'de verilmiştir.




**Şekil 7.** $f_1$ ve $f_2$ amaç fonksiyonları için başlangıç ve Pareto çözüm kümeleri

Çizelge 13'te amaç fonksiyonlarını eş anlı minimize eden optimal girdi değişken değerleri yer almaktadır.

**Çizelge 13.** $f_1$ ve $f_2$ amaç fonksiyonlarını eş anlı minimize eden optimal girdi değişken değerleri

| *Girdi Değişkenleri* | | *Amaç fonksiyonları* |
|---|---|---|
| $X_1$ | $X_2$ | $[f_1 \quad f_2]$ |
| -1.0000 | -1.4140 | [3.1206 0.1164] |
| -1.4140 | -1.0086 | [2.9791 0.0908] |

$Y_2$ ve $Y_3$ yanıtları için SUR yöntemi ile elde edilen parametre tahminleri Çizelge 14'te verilmiştir.

**Çizelge 14.** $Y_2$ ve $Y_3$ yanıtlarının SUR yöntemi ile parametre tahminleri

| ***Model Terimi*** | **SUR** | | | |
|---|---|---|---|---|
| | $Y_2$ | $p$-değeri | $Y_3$ | $p$-değeri |
| *Sabit* | 0.66 (.007) | **.001** | 1.776 (.016) | **.001** |
| $X_1$ | -0.092 (.005) | **.001** | -0.25 (.013) | **.001** |
| $X_2$ | -0.010 (.005) | .106 | -0.078 (.013) | **.001** |
| $X_1^2$ | -0.096 (.006) | **.001** | -0.156 (.014) | **.001** |
| $X_2^2$ | -0.058 (.006) | **.001** | -0.078 (.014) | **.001** |
| $X_1X_2$ | -0.070 (.007) | **.001** | 0.01 (.018) | .602 |

*Standart hatalar parantez içinde belirtilmiştir.

Çizelge 14'ten, $Y_2$ ve $Y_3$ yanıtları için elde edilen tahmini yanıt modelleri sırasıyla

$$\hat{Y}_2 = 0.660 - 0.092X_1 - 0.096X_1^2 - 0.058X_2^2 - 0.070X_1X_2 \tag{29}$$

ve

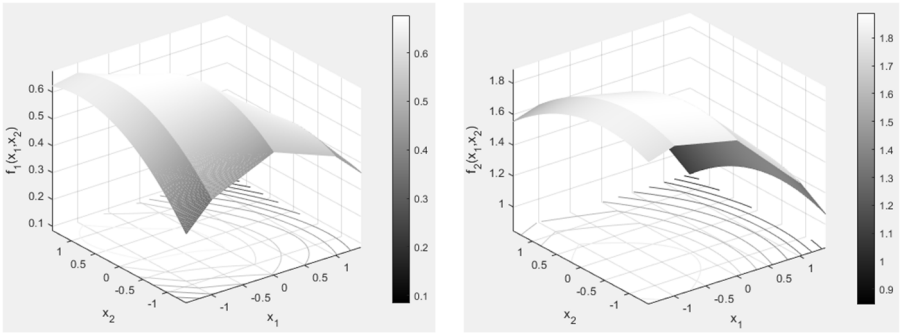
$$\hat{Y}_3 = 1.776 - 0.250X_1 - 0.078X_2 - 0.156X_1^2 - 0.079X_2^2 \tag{30}$$

dır. Optimal değeri elde edilmek istenilen her bir yanıt fonksiyonu, bir amaç fonksiyonu olarak değerlendirildiğinde çok yanıtlı optimizasyon problemi

$$\begin{aligned} &max\, f_1 = \hat{Y}_2(\boldsymbol{x}) \\ &max\, f_2 = \hat{Y}_3(\boldsymbol{x}) \\ &\quad \boldsymbol{x} \in S \end{aligned} \tag{31}$$

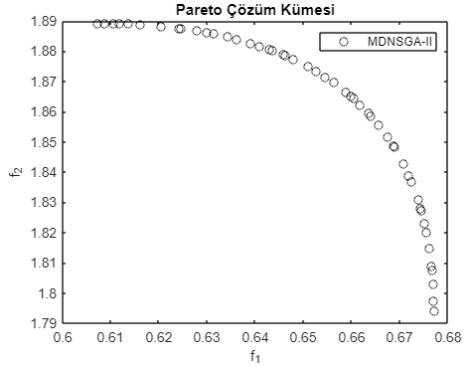
biçiminde ÇAO problemine dönüşecektir. Burada, $\{X_1, X_2\}$ değişkenlerinin aldığı değerler vektörü $\boldsymbol{x}$ olarak belirtilmiştir. $f_1$ ve $f_2$ amaç fonksiyonları, sırasıyla $\hat{Y}_2$ ve $\hat{Y}_3$ tahmini yanıt fonksiyonlarına karşılık gelmektedir. Elde edilen tahmini yanıt fonksiyonları için yüzey grafikleri Şekil 8.(a)-(b)'de görülmektedir. Şekil 8.(a)-(b)'ye bakıldığında maksimizasyon problemleriyle ilgilenildiği görülmektedir.

(a) (b)




**Şekil 8.** (a) $f_1$ amaç fonksiyonu (b) $f_2$ amaç fonksiyonu için yüzey grafikleri

ÇAO probleminin MDNSGA-II ile elde edilen Pareto çözüm kümesi Şekil 9'da verilmiştir. Gıda veri seti için $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin Pareto çözüm değerleri Ek-2'de verilmiştir. Ek-2'de görüldüğü gibi MDNSGA-II ile $\{X_1\}$ değişkeni tanım kümesinden kesikli değerler alarak Pareto çözümlerin elde edilmesi sağlanmıştır.




**Şekil 9.** Gıda veri seti için $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin Pareto çözüm kümesi

## 5. Sonuç ve öneriler

Bu çalışmada, karma veri içeren çok yanıtlı problemlerin modellenmesi ve optimizasyonu ile ilgilenilmiştir. Yanıt değişkenlerinin modellenmesinde GLM ve SUR modeller kullanılmıştır. Karma veri içeren çok yanıtlı modellerin tahmini yanıt fonksiyonları, amaç fonksiyonları olarak ele alınıp problem ÇAO problemi biçiminde değerlendirilmiştir. ÇAO için bir yapay zeka optimizasyon algoritması olan NSGA-II'ye dayalı algoritma önerilmiştir. Bu amaçla, NSGA-II'nin değişken gösterimi, başlangıç popülasyonunun oluşturulması ve genetik operatörlerin uygulanması aşamalarında indeksleme yapılarak kesikli değer alan değişkenlerle optimizasyon yapabilmek için NSGA-II modifiye edilmiştir. Önerilen algoritma, çalışma kapsamında MDNSGA-II olarak adlandırılmıştır. Çalışmada, kesikli değişkenlerin indeks değerleri dikkate alınarak karma veri içeren çok yanıtlı problemler için MDNSGA-II ile Pareto çözüm kümesinin elde edilebilir olduğu gösterilmiştir. Uygulamada kullanılan UCI Repository veri tabanından enerji verimliliği veri seti ile literatürde tanımlı gıda alanından deneysel karma veri seti için elde edilen sonuçlar Ek-1 ve Ek-2'de verilmiştir. Ek-1 ve Ek-2 incelendiğinde, uygulamada kullanılan veri setleri için MDNSGA-II ile Pareto çözüm kümesinin elde edilebildiği görülmüştür. Bununla birlikte, Enerji veri setinde $\hat{Y}_1$ - $\hat{Y}_2$ ve gıda veri setinde $\hat{Y}_2$ - $\hat{Y}_3$ tahmini yanıt fonksiyonları için Pareto çözüm kümesinde çözüm çeşitliliği sağlanırken, gıda veri setinin $\hat{Y}_1$ - $\hat{Y}_4$ tahmini yanıt fonksiyonları için Pareto çözüm kümesinde çözüm çeşitliliği yeterince sağlanamamıştır. Pareto çözüm kümesinde çözüm çeşitliliğini artırmak için referans noktalarına dayanan NSGA-III'ün uyarlanmasıyla bu sorunun giderilmesi sonraki çalışma planı olarak öngörülmektedir.

**Ek-1** Enerji veri seti için $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin Pareto çözüm kümesi

| *N* | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $f_1$ | $f_2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.62 | 610.6571 | 380.9550 | 113.8490 | 3.5 | 3 | 0.40 | 4 | 9.8763 | 83.3668 |
| 2 | 0.62 | 610.9978 | 408.3545 | 115.5357 | 3.5 | 3 | 0.40 | 4 | 10.1452 | 68.0712 |
| 3 | 0.62 | 612.4556 | 393.1433 | 115.6923 | 3.5 | 3 | 0.40 | 4 | 9.9712 | 76.7932 |
| 4 | 0.62 | 610.7042 | 390.4315 | 113.9442 | 3.5 | 3 | 0.40 | 4 | 9.9704 | 78.0666 |
| 5 | 0.62 | 610.6187 | 381.6393 | 113.8251 | 3.5 | 3 | 0.40 | 4 | 9.8837 | 82.9782 |
| 6 | 0.62 | 610.8081 | 386.8392 | 113.9304 | 3.5 | 3 | 0.40 | 4 | 9.9329 | 80.0928 |
| 7 | 0.62 | 610.6168 | 385.6920 | 113.9586 | 3.5 | 3 | 0.40 | 4 | 9.9243 | 80.7085 |
| 8 | 0.62 | 610.7253 | 384.3239 | 113.8813 | 3.5 | 3 | 0.40 | 4 | 9.9090 | 81.4898 |
| 9 | 0.62 | 612.3562 | 394.9505 | 115.7381 | 3.5 | 3 | 0.40 | 4 | 9.9908 | 75.7672 |
| 10 | 0.62 | 612.4023 | 401.1725 | 115.4784 | 3.5 | 3 | 0.40 | 4 | 10.0523 | 72.2893 |
| 11 | 0.62 | 610.8283 | 388.5117 | 113.9288 | 3.5 | 3 | 0.40 | 4 | 9.9493 | 79.1590 |
| 12 | 0.62 | 611.0447 | 406.3511 | 115.5784 | 3.5 | 3 | 0.40 | 4 | 10.1245 | 69.1996 |
| 13 | 0.62 | 611.4364 | 401.3444 | 115.4862 | 3.5 | 3 | 0.40 | 4 | 10.0685 | 72.0578 |
| 14 | 0.62 | 612.3345 | 397.0036 | 115.4902 | 3.5 | 3 | 0.40 | 4 | 10.0116 | 74.6144 |
| 15 | 0.62 | 612.2230 | 394.1300 | 115.7882 | 3.5 | 3 | 0.40 | 4 | 9.9846 | 76.2081 |
| 16 | 0.62 | 610.9330 | 387.8107 | 113.9694 | 3.5 | 3 | 0.40 | 4 | 9.9407 | 79.5663 |
| 17 | 0.62 | 612.3493 | 396.3216 | 115.5053 | 3.5 | 3 | 0.40 | 4 | 10.0046 | 74.9984 |
| 18 | 0.62 | 610.7242 | 385.0456 | 114.0231 | 3.5 | 3 | 0.40 | 4 | 9.9162 | 81.0855 |
| 19 | 0.62 | 610.6901 | 382.5880 | 113.6679 | 3.5 | 3 | 0.40 | 4 | 9.8921 | 82.4570 |
| 20 | 0.62 | 612.3872 | 395.5429 | 115.7546 | 3.5 | 3 | 0.40 | 4 | 9.9962 | 75.4398 |
| 21 | 0.62 | 611.3824 | 403.5865 | 115.3821 | 3.5 | 3 | 0.40 | 4 | 10.0917 | 70.7947 |
| 22 | 0.62 | 611.0526 | 405.6636 | 115.5859 | 3.5 | 3 | 0.40 | 4 | 10.1175 | 69.5858 |
| 23 | 0.62 | 610.7707 | 383.1573 | 113.9089 | 3.5 | 3 | 0.40 | 4 | 9.8966 | 82.1494 |
| 24 | 0.62 | 612.3377 | 397.4010 | 115.5058 | 3.5 | 3 | 0.40 | 4 | 10.0155 | 74.3923 |
| 25 | 0.62 | 611.0458 | 404.8979 | 115.5219 | 3.5 | 3 | 0.40 | 4 | 10.1099 | 70.0136 |
| 26 | 0.62 | 610.8196 | 387.4233 | 113.9322 | 3.5 | 3 | 0.40 | 4 | 9.9385 | 79.7673 |
| 27 | 0.62 | 610.7164 | 389.9160 | 113.9293 | 3.5 | 3 | 0.40 | 4 | 9.9650 | 78.3570 |
| 28 | 0.62 | 610.7170 | 389.5886 | 113.9400 | 3.5 | 3 | 0.40 | 4 | 9.9617 | 78.5404 |
| 29 | 0.62 | 611.0235 | 406.7547 | 115.6112 | 3.5 | 3 | 0.40 | 4 | 10.1288 | 68.9707 |
| 30 | 0.62 | 612.2376 | 395.8547 | 115.7415 | 3.5 | 3 | 0.40 | 4 | 10.0016 | 75.2442 |
| 31 | 0.62 | 611.1244 | 402.5708 | 115.3929 | 3.5 | 3 | 0.40 | 4 | 10.0855 | 71.3277 |
| 32 | 0.62 | 611.0496 | 405.4035 | 115.5863 | 3.5 | 3 | 0.40 | 4 | 10.1149 | 69.7310 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 33 | 0.62 | 612.4173 | 393.7951 | 115.7586 | 3.5 | 3 | 0.40 | 4 | 9.9783 | 76.4228 |
| 34 | 0.62 | 610.7554 | 385.3684 | 113.9496 | 3.5 | 3 | 0.40 | 4 | 9.9190 | 80.9091 |
| 35 | 0.62 | 612.1254 | 397.7169 | 115.3760 | 3.5 | 3 | 0.40 | 4 | 10.0219 | 74.1857 |
| 36 | 0.62 | 612.2661 | 399.6804 | 115.6109 | 3.5 | 3 | 0.40 | 4 | 10.0394 | 73.1058 |
| 37 | 0.62 | 611.4462 | 402.0085 | 115.3470 | 3.5 | 3 | 0.40 | 4 | 10.0750 | 71.6873 |
| 38 | 0.62 | 610.7435 | 389.0879 | 113.9321 | 3.5 | 3 | 0.40 | 4 | 9.9563 | 78.8245 |
| 39 | 0.62 | 610.7011 | 383.6942 | 113.8565 | 3.5 | 3 | 0.40 | 4 | 9.9030 | 81.8390 |
| 40 | 0.62 | 610.7108 | 383.7182 | 113.7821 | 3.5 | 3 | 0.40 | 4 | 9.9031 | 81.8269 |
| 41 | 0.62 | 611.3962 | 403.7842 | 115.3813 | 3.5 | 3 | 0.40 | 4 | 10.0935 | 70.6859 |
| 42 | 0.62 | 611.0798 | 403.8909 | 115.4860 | 3.5 | 3 | 0.40 | 4 | 10.0993 | 70.5823 |
| 43 | 0.62 | 612.4827 | 400.2912 | 115.5557 | 3.5 | 3 | 0.40 | 4 | 10.0423 | 72.7941 |
| 44 | 0.62 | 612.5187 | 400.9726 | 115.5398 | 3.5 | 3 | 0.40 | 4 | 10.0485 | 72.4176 |
| 45 | 0.62 | 612.3157 | 398.7933 | 115.3863 | 3.5 | 3 | 0.40 | 4 | 10.0298 | 73.6096 |
| 46 | 0.62 | 612.4558 | 398.2394 | 115.6069 | 3.5 | 3 | 0.40 | 4 | 10.0222 | 73.9394 |
| 47 | 0.62 | 612.5211 | 399.6069 | 115.5796 | 3.5 | 3 | 0.40 | 4 | 10.0349 | 73.1827 |
| 48 | 0.62 | 611.0621 | 404.0529 | 115.4698 | 3.5 | 3 | 0.40 | 4 | 10.1012 | 70.4891 |
| 49 | 0.62 | 611.1210 | 404.5703 | 115.4520 | 3.5 | 3 | 0.40 | 4 | 10.1055 | 70.2076 |
| 50 | 0.62 | 612.3504 | 400.3724 | 115.4825 | 3.5 | 3 | 0.40 | 4 | 10.0451 | 72.7301 |

**Ek-2** Gıda veri seti için $f_1$ ve $f_2$ amaç fonksiyonlarına ilişkin Pareto çözüm kümesi

| *N* | *X1* | *X2* | *f1* | *f2* |
|---|---|---|---|---|
| 1 | -1 | 0.6034 | 0.6771 | 1.7942 |
| 2 | -1 | -0.4937 | 0.6073 | 1.8893 |
| 3 | -1 | -0.3945 | 0.6194 | 1.8885 |
| 4 | -1 | -0.4679 | 0.6105 | 1.8892 |
| 5 | -1 | 0.2400 | 0.6695 | 1.8467 |
| 6 | -1 | -0.3555 | 0.6238 | 1.8877 |
| 7 | -1 | 0.2790 | 0.6710 | 1.8421 |
| 8 | -1 | -0.3324 | 0.6263 | 1.8872 |
| 9 | -1 | 0.3576 | 0.6736 | 1.8320 |
| 10 | -1 | 0.0050 | 0.6563 | 1.8696 |
| 11 | -1 | 0.1756 | 0.6665 | 1.8539 |
| 12 | -1 | -0.3004 | 0.6297 | 1.8863 |
| 13 | -1 | 0.2136 | 0.6683 | 1.8497 |
| 14 | -1 | -0.2504 | 0.6348 | 1.8846 |
| 15 | -1 | 0.3787 | 0.6742 | 1.8291 |
| 16 | -1 | 0.1448 | 0.6649 | 1.8570 |
| 17 | -1 | 0.1933 | 0.6674 | 1.8520 |
| 18 | -1 | -0.0680 | 0.6510 | 1.8749 |
| 19 | -1 | 0.0409 | 0.6588 | 1.8667 |
| 20 | -1 | 0.0995 | 0.6624 | 1.8615 |
| 21 | -1 | -0.0238 | 0.6543 | 1.8718 |
| 22 | -1 | 0.4018 | 0.6748 | 1.8259 |
| 23 | -1 | -0.2750 | 0.6324 | 1.8855 |
| 24 | -1 | 0.0804 | 0.6613 | 1.8632 |
| 25 | -1 | 0.3267 | 0.6727 | 1.8361 |
| 26 | -1 | -0.2304 | 0.6368 | 1.8838 |
| 27 | -1 | 0.4253 | 0.6753 | 1.8225 |
| 28 | -1 | -0.0911 | 0.6491 | 1.8764 |
| 29 | -1 | -0.1434 | 0.6448 | 1.8796 |
| 30 | -1 | -0.1522 | 0.6440 | 1.8800 |
| 31 | -1 | 0.4833 | 0.6763 | 1.8139 |
| 32 | -1 | -0.2185 | 0.6379 | 1.8833 |
| 33 | -1 | 0.0643 | 0.6603 | 1.8647 |
| 34 | -1 | 0.6022 | 0.6771 | 1.7944 |

| | | | | |
|---|---|---|---|---|
| 35 | -1 | 0.4615 | 0.6760 | 1.8172 |
| 36 | -1 | -0.1727 | 0.6422 | 1.8811 |
| 37 | -1 | 0.5783 | 0.6771 | 1.7985 |
| 38 | -1 | -0.0456 | 0.6527 | 1.8734 |
| 39 | -1 | -0.1226 | 0.6465 | 1.8784 |
| 40 | -1 | 0.5224 | 0.6767 | 1.8077 |
| 41 | -1 | -0.0535 | 0.6521 | 1.8739 |
| 42 | -1 | 0.0270 | 0.6578 | 1.8678 |
| 43 | -1 | 0.4473 | 0.6757 | 1.8193 |
| 44 | -1 | 0.4751 | 0.6762 | 1.8151 |
| 45 | -1 | 0.5642 | 0.6770 | 1.8009 |
| 46 | -1 | -0.1019 | 0.6483 | 1.8771 |
| 47 | -1 | -0.1198 | 0.6468 | 1.8782 |
| 48 | -1 | 0.5032 | 0.6765 | 1.8107 |
| 49 | -1 | 0.2967 | 0.6717 | 1.8399 |
| 50 | -1 | 0.5038 | 0.6765 | 1.8106 |

**Kaynaklar**


[1] J. Garrido, J. Zhou, 2009, Full Credibility with Generalized Linear and Mixed Models, *ASTIN Bulletin*, 39(1), 61-80.

[2] N. P. Jewell, S. Shiboski, 1990, Statistical analysis of HIV infectivity based on partner studies, *Biometrics*, 46, 1133-1150.

[3] A. Hern, S. Dorn, 2001, Statistical modelling of insect behavioral responses in relation to the chemical composition of test extracts, *Physiological Entomology*, 26, 381-390.

[4] Y. J. Lee, J. A. Nelder, 2002, Analysis of ulcer data using hierarchical generalized linear models, *Statistics in Medicine*, 21, 191-202.

[5] M. P. Diaz, A. H. Barchuk, S. Luque, C. Oviedo, 2002, Generalized linear models to study spatial distribution of tree species in Argentinean arid Chaco, *Journal of Applied Statistics*, 29, 5, 685-694.

[6] Z. W. Yan, S. Bate, R. E. Chandler, V. Isham, H. Wheater, 2002, An analysis of daily maximum wind speed in northwestern Europe using generalized linear models, *Journal of Climate*, 15, 2073-2088.

[7] S. Rajeev, C. S. Krishnamoorthy, 1992, Discrete Optimization of Structures Using Genetic Algorithms, *J. Struct. Eng.*, 118(5), 1233-1250.

[8] C. Y. Lin, P. Hajela, 1992, Genetic algorithms in optimization problems with discrete and integer design variables, *Engineering Optimization*, 19, 4, 309- 327.

[9] W. Wang, R. Zmeureanu, H. Rivard, 2005, Applying multi-objective genetic algorithms in green building design optimization, *Building and Environment* 40, 1512–1525.

[10] S. S. Rao, Y. Xiong, 2005, A hybrid genetic algorithm for mixed-discrete design optimization, *ASME Journal of Mechanical Design*, 127, 1100–1112.

[11] M. Ahmadi, M. Arabi, D. L. Hoag, B. A. Engel, 2013, A Mixed Discrete-Continuous Variable Multiobjective Genetic Algorithm For Targeted Implementation of Nonpoint Source Pollution Control Practices, *Water Resources Research*, 49, 8344–8356.

[12] B. El-Kribi, A. Houidi, Z. Affi, L. Romdhane, 2013, Application of multi-objective genetic algorithms to the mechatronic design of a four bar system with continuous and discrete variables, *Mechanism and Machine Theory*, 61, 68–83.

[13] W. Tong, S. Chowdhury and A. Messac, 2014, A new multi-Objective Mixed-Discrete Particle Swarm Optimization Algorithm, *Proceedings of the ASME 2014 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference Buffalo*, New York, USA.

[14] T. Holzmann, J. C. Smith, 2018, Solving discrete multi-objective optimization problems using modified augmented weighted Tchebychev scalarizations, *European Journal of Operational Research*, 271,436–449.

[15] S. Guangyong, Z. Huile, F. Jianguang, L. Guangyao, L. Qing, 2018, A new multi-objective discrete robust optimization algorithm for engineering design, *Applied Mathematical Modelling*, 53, 602-621.

[16] S. Roy, W. A. Crossley, S. Jain, 2021, *A Hyrid Approach for Solving Constrained Multi-Objective Mixed-Discrete Nonlinear Programming Engineering Problems*, Books, Engineering Problems - Uncertainties, Constraints and Optimization Techniques.

[17] I. Khuri, B. Mukherjee, B. K. Sinha, M. Ghosh, 2006, Design Issues for Generalized Linear Models: A Review, *Statistical Science*, 21(3), 376-399.

[18] D. Collins, 2008, *The performance of estimation methods for generalized linear mixed models, Doctor of Philosophy thesis*, University of Wollongong, School of Mathematics and Applied Statistics- Faculty of Informatics, 223, Australia.

[19] J. A. Nelder, R. W. M. Wedderburn, 1972, Generalized linear models, *Journal of the Royal Statistical Society A–General*, 135, 370–384.

[20] C. J. Anderson, J. Verkuilen, T. R. Johnson, 2012. *Applied Generalized Linear Mixed Models: Continuous and Discrete Data.*, Springer.

[21] P. McCullagh, J. A. Nelder, 1989, *Generalized Linear Models, Second Edition*, Chapman and Hall/CRC, 511, London.

[22] M. Friendly, D. Meyer, 2015, *Discrete Data Analysis with R:Visualization and Modeling Techniques for Categorical and Count Data*, Chapman and Hall/CRC Published.

[23] J. J. Faraway, 2006, *Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC Taylor & Francis Group.

[24] E. Ostertagová, 2012, *Modelling Using Polynomial Regression*, Procedia Engineering, 48, 500-506.

[25] Ö. Türkşen, 2023, *Optimizasyon Yöntemleri ve Matlab, Python, R Uygulamaları*, Nobel, 1.Basım, 448, Ankara.

[26] D. E. Goldberg, 1989, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, ABD.

[27] N. Srinivas, K. Deb, 1994, Mulltiobjective Optimization Using Non-Dominated Sorting in Genetic Algorithms, *Evolutionary Computation*, 2, 221-248.

[28] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, 2002, A fast and elitist multiobjective genetic algortihm: NSGA-II, *IEEE Transactions on Evolutionary Computation*, 6, 2.

[29] Ö. Türkşen, F. Akgün, 2018, Genetik-Simpleks hibrit algoritması ile doğrusal olmayan regresyon model parametrelerinin nokta tahmini, *İstatistikçiler Dergisi: İstatistik & Aktüerya*, 2, 81-92.

[30] Z. Cebeci, 2021, *R ile Genetik Algoritmalar ve Optimizasyon Uygulamaları*, Nobel, 535, Ankara.

[31] Ö. Türkşen, 2011, *Çok Yanıtlı Yüzey Problemlerinin Çözümüne Bulanık ve Sezgisel Yaklaşım*, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, İstatistik Anabilim Dalı.

[32] B. El-Kribi, A. Houidi, Z. Affi, L. Romdhane, 2013, Application of multi-objective genetic algorithms to the mechatronic design of a four bar system with continuous and discrete variables, *Mechanism and Machine Theory*, 61, 68–83.

[33] A. Asuncion, D. Newman, UCI Machine Learning Repository. Available online:https://archive.ics.uci.edu/dataset/242/energy+efficiency.

[34] A. Tsanas, A. Xifara, 2012, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, *Energy and Buildings*, 49, 560-567.

[35] R. H. Schmidt, R, B. L. Illingworth, J. C. Deng, J. A. Cornell, 1979, Multiple Regression and Response Surface Analysis of the Effects of Calcium Chloride and Cysteine on Heat-Induced Whey Protein Gelation*, J. Agrie. Food Chem.*, 27(3), 529–532.

[36] S. Tunçel, 2022, *Çok Yanıtlı Deneysel Verilerin Görünüşte İlişkisiz Regresyon Analizi ile Modellenmesi ve Optimal Değişken Değerlerinin Belirlenmesi*, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstatistik Anabilim Dalı.

Aktüerya Derneği

**İstatistikçiler Dergisi: İstatistik & Aktüerya**

Journal of Statisticians: Statistics and Actuarial Sciences

**IDIA 16**, 2023, 2, 81-99

**Geliş** / Received:04.12.2023, **Kabul** / Accepted: 28.12.2023

**Araştırma Makalesi /** Research Article

# Regression Tree Approach to Estimation of Health Insurance Premium

Başak Bulut Karageyik
*Hacettepe University*
*Department of Actuarial Sciences, Ankara, Türkiye*
*basakbulut@hacettepe.edu.tr*
ORCID:0000-0003-4080-9165

**Abstract**

This paper proposes an approach to predicting insurance premiums in health insurance by combining traditional generalized linear models (GLM) with advanced machine learning-driven regression tree analysis. The study first uses GLM on real complementary health insurance data to examine the importance of variables, focusing on those variables that have a large impact on premium estimates. Subsequently, it is investigated whether the variables identified as significant by GLM can also be identified as significant by regression tree analysis. In the application of machine learning, the effect of stratified sampling in accordance with the data structure in terms of the risk variables considered in premium forecasts is also analyzed. This study contributes to the actuarial understanding of premium estimation and provides insurers with a concrete framework to help them negotiate the complex world of health insurance data. By integrating the advantages of GLM and regression trees, this study provides a comprehensive comparison for insurers to adapt to changing risk factors. This study represents a innovative attempt to incorporate a regression tree methodology, providing a novel and accurate estimation of premium amounts in the realm of insurance analysis.

***Keywords:*** Actuarial premium estimation, Regression tree, Machine learning techniques, Generalized linear models,

## Öz

### Sağlık Sigortası Primi Tahmininde Regresyon Ağacı Yaklaşımı

*Bu çalışma, geleneksel genelleştirilmiş doğrusal modelleri (GLM) gelişmiş makine öğrenimi odaklı regresyon ağacı analizi ile birleştirerek sağlık sigortasında sigorta primlerini tahmin etmeye yönelik bir yaklaşım önermektedir. Çalışmada ilk olarak değişkenlerin önemini incelemek için gerçek tamamlayıcı sağlık sigortası verileri üzerine GLM uygulanmakta ve prim tahminleri üzerinde büyük etkisi olan değişkenlere odaklanılmaktadır. Daha sonra, GLM tarafından önemli olarak tanımlanan değişkenlerin regresyon ağacı analizi ile de önemli olarak tanımlanıp tanımlanamayacağı araştırılmaktadır. Makine öğrenmesi uygulamasında, prim tahminlerinde dikkate alınan risk değişkenleri açısından veri yapısına uygun olarak tabakalı örneklemenin etkisi de analiz edilmektedir. Bu çalışma, prim tahminine ilişkin aktüeryal anlayışa katkıda bulunmakta ve sigortacılara sağlık sigortası verilerinin karmaşık dünyasında müzakere etmelerine yardımcı olacak somut bir çerçeve sunmaktadır. GLM ve regresyon ağaçlarının avantajlarını bir araya getiren bu çalışma, sigortacıların değişen risk faktörlerine uyum sağlamaları için kapsamlı bir karşılaştırma sunmakta ve sigorta analizi alanında prim tutarlarının yeni ve doğru bir şekilde tahmin edilmesini sağlayan bir regresyon ağacı metodolojisini içeren yenilikçi bir çalışmayı temsil etmektedir.*


***Anahtar sözcükler:*** *Aktüeryal prim tahmini, Regresyon ağacı, Makine öğrenme teknikleri, Genelleştirilmiş doğrusal modeller*

## 1. Introduction

Actuarial science brings essential insights with a statistical, demographic, and social perspective into the analysis of risk factors that influence complicated insurance occurrences. Hence, an actuarial perspective enhances the ability to manage the complex world of risk. In the ever-evolving landscape of actuarial science, the estimation of insurance premiums stands as intricate with mathematical precision, statistical insight, and an acute understanding of risk dynamics. The premium estimation is the most important and lies at the heart of insurance pricing, financial sustainability, and risk management strategies. The estimation of insurance premiums holds paramount importance within the domain of actuarial science and the broader insurance industry.

The significance of premium estimation is multifaceted, encompassing financial stability, risk management, market competitiveness, and the overall sustainability of insurance operations. Premium estimation in actuarial science involves the use of various techniques and models to assess risk and determine the appropriate pricing for insurance coverage. In premium estimation techniques, as a combination of traditional and modern methodologies, frequency and severity models, generalized linear models (GLM), the loss ratio method, credibility theory, Bayesian methods, time series analysis, and extreme value theory (EVT) are commonly used in actual science. Among these methods, GLMs is one of the most preferred because it extends traditional linear models to handle non-normally distributed response variables. Actuaries use GLMs to model relationships between premiums and risk factors, incorporating link functions that account for the specific distribution of the response variable (e.g., Poisson or Gamma distributions).

GLMs play a pivotal role, especially in non-life insurance, in assessing and pricing risks, as well as in estimating more accurate reserves. Actuarial science often involves the application of statistical models, including GLMs, for analyzing and modeling insurance-related data. The GLM is developed as actuarial illustrations in the standard text by McCullagh and Nelder [1]. They provide numerous instances of how GLMs have been fitted to other kinds of data, such as average claim costs from a portfolio of auto insurance. Then Renshaw [2] and Renshaw and Verall [3] made the first studies in the actuarial field. In 1996, Haberman and Renshaw [4] analyzed in detail the use of GLMs in actuarial data analysis and demonstrated the use of GLMs in insurance claim frequency and severity. Over the years, GLM has been

applied in the calculation of loss resezrves, credibility, and mortality forecasting. These references cover the theoretical foundations as well as practical applications of GLMs in the actuarial sciences: Dobson [5], Anderson et al. [6], Antonio and Beirlant [7], De Jong and Heller [8], Wüthrich and Merz [9], Ohlsson et al. [10], and Frees [11].

A decision tree is a graphical representation and predictive modeling tool used in machine learning and data analysis. Decision trees are particularly useful for classification and regression tasks, as they help break down complex decision-making processes into a series of simpler, interpretable steps based on the input features of the data. CART is a versatile type of decision tree that can be used for both classification and regression tasks. It recursively splits the dataset based on the most significant attribute at each node. According to its purpose, CART is divided into two parts: firstly, to classify the data into discrete classes or categories, and secondly, to predict numerical values, making it suitable for the regression task.

Regression trees are an important instrument in the actuarial toolbox since they offer a special perspective for understanding and evaluating intricate risk dynamics. Regression trees excel at identifying distinct segments within a dataset, enabling actuaries to tailor risk assessment strategies to specific groups. Quan [12] summarized the advantages of the tree-based model that are important for the analysis of actuarial and insurance data in five points: Tree-based models are considered as nonparametric models and therefore do not require distributional assumptions, tree-based models can be used as a practical algorithm that can handle missing data and categorical variables in a natural way, tree-based models can automatically detect non-linear effects and potential effects, and they are easy to interpret by visualizing the tree structure in a graph, especially for smaller size trees. These advantages are particularly useful for reporting models used in actuarial and insurance data analysis.

Regression trees are employed in estimating insurance premiums by capturing the non-linear relationships between policyholder attributes and expected claim amounts. This aids insurers in setting accurate premium rates based on a nuanced understanding of risk factors. Regression trees are especially ideally suited for situations where the impact of variables is not constant because, in contrast to typical linear models, they are able to capture non-linear correlations in data. Regression trees' intuitive design makes interpretation simple, which makes it easier to communicate findings. Actuaries can better prioritize elements that have a major impact on the outcomes of interest by using regression trees, which offer insights into the relative relevance of various variables.

The combination of regression trees and machine learning has become a revolutionary force in the dynamic field of actuarial science, revolutionizing the way actuaries approach risk assessment and predictive modeling.

Machine learning augments the predictive power of regression trees, enabling the model to capture intricate relationships and dependencies within the data. This enhanced modeling capability is particularly valuable in estimating premium, predicting claim occurrences and assessing severity with a higher degree of accuracy. This study examines the mutually beneficial relationship that exists between regression trees and machine learning, highlighting the special advantages and potential uses that result from this potent union. The combination of regression trees and machine learning is set to define the forefront of data-driven decision-making in actuarial practice as the discipline continues to embrace technological breakthroughs.

Due to its increasing impact and importance in recent years, there have been many studies on classification and regression trees driven by machine learning. However, few papers can be found in the insurance literature related to regression trees and machine learning. Gardner et al. [13] use regression trees and two-stage screening were assessed by contrasting their accuracy with traditional actuarial techniques. Steadman et al. [14] proposed that a classification tree approach and two decision thresholds can enhance the use of actuarial violence risk assessment tools in clinical practice. Guelman [15] compares gradient-boosted trees with GLMs to forecast the cost of vehicle accident losses for at-fault claims. William [16] suggests a two-phase modeling process that expands on previous statistical tools like classification and regression trees, generalized linear mixed models, and actuarial methods from conventional insurance claim cost modeling. Wuthrich and Buser [17] applied various statistical methods and machine learning techniques for non-life

insurance pricing, including regression trees, bagging, random forest, boosting, and support vector machines. Diao and Weng [18] merge machine learning methods with credibility theory and suggest an approach based on regression trees to incorporate covariate data into the estimation of the credibility premium. Baillargeon [19] presents a neural architecture that can predict actuarial risk factors in accident descriptions using dense embeddings, yielding more performing and interpretable models than traditional actuarial data mining methods. Tober [20] focuses on creating and assessing three tree-based machine learning models to forecast the frequency of claims, advancing from straightforward decision trees to more complex ensemble techniques like random forests and gradient boosting machines. Henckaerts et al. [21] concentrate on using machine learning techniques to create comprehensive tariff plans based on the severity and frequency of claims. Rokicki [22] proposed the modified actuarial credibility approach, which provides accurate initial cost estimates for transport infrastructure projects, outperforming more complex methods like regression analysis and machine learning. Richman [23, 24] looks into the potential evolution and adaptation of actuarial science to include machine learning. Wong [25] provides the state of the art in ratemaking and reserving and examines how machine learning is being applied to the field of actuarial science. Quag [26] examines the various applications of tree-based models in insurance and actuarial science.

Resampling techniques play a pivotal role in machine learning, offering a strategic approach to mitigate bias, enhance model robustness, and provide a more accurate assessment of a model's performance. Among resampling processes, the "stratified random sampling" method is superior to the balanced (representative) sample when used appropriately. Stratification is the process of dividing the population into homogeneous subgroups prior to sampling.

Health insurance is more sensitive to individual characteristics, causing concentrations or infrequent conditions to be observed in the relevant risk factors and sub-fractures. For this reason, the risk factors and the probability of observation in subcategories should be taken into consideration in order to better represent the whole data in the analysis. In this study, stratified sampling was used because a non-homogeneous structure was also observed in the subgroups of various risk factors in the data examined. The theoretical background regarding stratified sampling can be found in these studies: Neyman [27], Neyman and Pearson [28], Singh and Mangat [29], and Parsons [30]. The references regarding the application of machine learning to stratified sampling are located in Liberty et al. [31], Ye et al. [32], Yu et al. [33], and Lu et al. [34].

Although actuarial science has a wealth of traditional approaches, there is a notable lack of comprehensive research on regression trees and more general machine learning applications. The absence of research in this area is especially noteworthy considering the opportunity these methods offer to improve the accuracy and flexibility of premium estimating algorithms. The new research aims to bridge this gap by delving into the unexplored area of regression trees and machine learning applications within actuarial science. Therefore, this study aims to reflect the differences between regression trees for premium forecasting from a general perspective and when they are used for forecasting purposes in the light of prior information obtained from GLM.

The remainder of the paper is organized as follows into five sections: Section 1 provides a brief introduction and a concise literature review on generalized linear models (GLM), regression trees, and the broader landscape of machine learning. Section 2 delves into the intricacies of GLM, shedding light on its mathematical foundations and highlighting its applications in actuarial science. Section 3 shifts focus to regression trees, providing a nuanced discussion on both standard regression trees (CART) and regression trees integrated with machine learning techniques. Section 4 presents a numerical analysis of premium estimation on health insurance data in the context of GLM, regression trees, and machine learning. Section 5 summarizes the results of this work and draws conclusions.

## 2. Generalized linear models

Generalized linear models (GLMs) are a class of statistical models that expand the linear model framework to handle a wider range of data distributions and connections. Conventional linear models make the assumption that the response variable, also known as the dependent variable, is normally distributed and that there is a linear relationship between the predictors, or independent variables, and the response. The response variable may have any of the exponential family distributions—normal, binomial, Poisson, gamma, and so on—according to GLMs, which loosen these requirements. A GLM's essential elements consist of the random component, the systematic component (linear predictor), and the link function. The response variable's probability distribution is described by the random component. It is a member of the exponential distribution family. The linear combination of the predictor variables is represented by the systemic component, also known as the linear predictor. A connection function connects it to the random component. The systematic component is connected to the expected value of the response variable by a link function. It guarantees that the model accurately captures the connection between the predictors and the distribution mean. The choice of link function depends on the distribution of the response variable. For each ith of n independently collected observations, a random component that, given the values of the explanatory variables in the model, specifies the conditional distribution of the response variable, Yi. A distribution like the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian families of distributions, or the $Y_i$ distribution, was included in the initial definition of GLMs. A linear predictor, or a linear function of regressors,

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

An invertible and smooth linearizing link function, g(.) is used to convert the response variable's expectation, $\mu_i = E(Y_i)$, into a linear predictor:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

where g(.) is the link function; $\mu_i$ is the expected value of the response variables; $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients of the model and $x_{i1}, x_{i2}, \ldots x_{ik}$ are the predictor variables [35].

Common link functions include the identity link for Gaussian distribution, the logit link for binomial distribution, and the log link for Poisson distribution. The random component describes the response variable's probability distribution. The distribution in question is a member of the exponential family. GLMs are especially well-suited for actuarial applications where risk events frequently follow non-normal distributions, as, in contrast to linear models, they can incorporate a range of probability distributions and accommodate non-normal distributions of response variables.

GLMs are employed in this study to determine the importance and influence of the variables in our dataset. Beyond traditional linear models, GLMs offer a robust analytical framework that can handle a wide range of data distributions and capture complex interactions between variables. The variables determined to be important by GLM will be used in establishing the regression tree model, and it will be examined whether the criteria that are important in the predictions obtained according to the regression tree are similar to the variables obtained by GLM.

## 3. Decision Trees

A decision tree is a data analysis and machine learning tool for graphical representation and predictive modeling. It resembles an inverted tree, with each node standing for a choice or test on a certain attribute, each branch for the decision's result, and each leaf node for the outcome that was ultimately expected or the class label. Decision trees are especially helpful for tasks involving classification and regression because

they assist in decomposing intricate decision-making procedures into a number of easier to understand steps that are dependent on the data's input attributes. Because they can manage both numerical and categorical data while promoting transparency in the decision-making process, they are extensively used in a variety of industries, such as marketing, finance, and healthcare. The references on decision analysis and decision trees can be found in Magee[36], Murthy[37], Keeney [38], Tjen-Sien et. al [39] and Kotsiantis[40].

There are several types of decision trees, and their variations are often designed to address specific challenges or data characteristics. CART (Classification and Regression Trees), a decision tree that can be used for both classification and regression tasks [41]; ID3 (Iterative Dichotomiser 3), which uses entropy and information gain measures to decide the best attribute to split the dataset [42]; C4.5, which is an extension of ID3 and handles both continuous and discrete data using information gain [43] , CHAID (Chi-square Automatic Interaction Detector), which uses chi-square tests to identify significant relationships between variables, used for categorical target variables [44], Random Forest, an ensemble learning method that builds multiple decision trees, combines their predictions, helps to increase accuracy and reduce overfitting [45]. In this study, numerical analysis is performed on the CART algorithm.

### *3.1. CART (Classification and Regression Tree - C&RT) Algorithm*

The Classification and Regression Tree Analysis (CART) algorithm was developed in 1984 by Breiman, Freidman, Olshen, and Stone [41]. CART is a straightforward but effective analytical approach that assists in identifying the most "important" (based on explanatory power) variables in a given dataset. CART algorithm divide the predictor space recursively into subsets where the distribution of y is progressively more homogeneous using a binary tree [46].

In the non-parametric regression-type CART algorithm, the data is divided into nodes based on conditional binary answers to questions containing the predictor variable y for predicting continuous dependent variables with categorical and/or continuous predictor variables.

CART can statistically show which variables, in terms of variance and explanatory power, are most significant in a model or relationship. In this sense, CART offers a complex overview of the relationships between the variables in the data and can be employed as an initial stage in building a useful model or a final representation of significant correlations. CART's visual bridge between statistical rigor and interpretation helps to make relevant and valid model creation easier [47].

The CART method creates an algorithm to predict the target values by extracting decision rules from characteristics, much like other decision tree algorithms. Both qualitative and numerical data may be included in the characteristics. Breiman et al. [41], Chipman et al. [46], Verbyla [48], Clark and Pregibon [49] are recommended readings for a comprehensive overview of the CART algorithm.

#### *3.1.1. Regression trees*

Although classification and regression in tree analysis use relatively similar statistical techniques, it's crucial to understand the differences between the two. It is desired to classify the response variable, which is often binary (0–1), in order to divide the dataset into groups. Regression trees will be used when our response variable is numeric or continuous and we want to use the data to predict the outcome. In essence, a classification tree divides the data according to homogeneity; categorizing according to similar data and filtering out the "noise" makes it more "pure"—hence the idea of a purity criterion [41]. The separations in the regression tree are performed according to the "reduction of the squares of the residuals algorithm", which means that the total variance estimated for the two resulting nodes must be minimized [41], [50]. Figure 1 is a valuable illustration of this procedure [41].

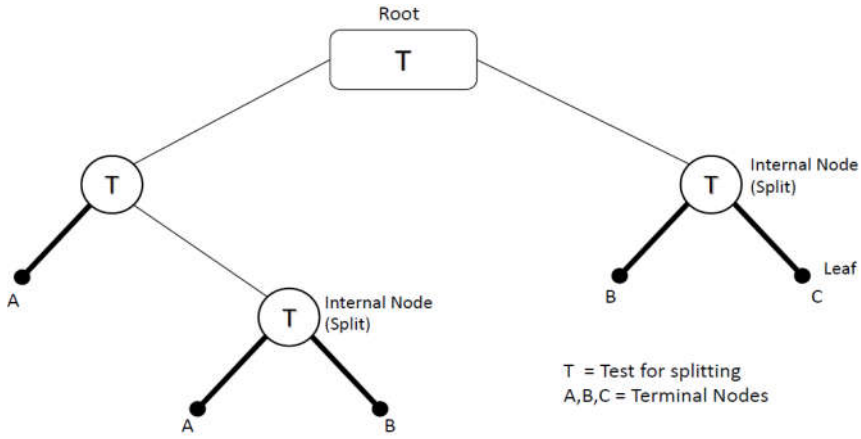



**Figure 1.** The structure of CART

Regression trees are an essential part of statistical modeling and machine learning, and actuarial science has greatly benefited from their capacity to reveal subtle patterns in large, complicated datasets. A tree-like model known as a regression tree divides the dataset into homogenous subsets recursively according to the values of predictor variables. Regression trees are especially useful for interpretation and prediction since each terminal node, or leaf, indicates a predicted outcome. The decision-making paths can be transparently visualized due to the tree structure's simplicity. Since CART is a non-parametric method for estimating the continuous dependent variable with categorical predictor variables and is appropriate for prediction with the variable set chosen for this investigation, it was decided to work with regression trees.

In order to improve the model's performance, adaptability, and interpretability, machine learning is integrated into regression trees through the use of sophisticated machine learning ideas and methods. In this study, it is aimed at the integration of machine learning with regression trees using splitting for training and testing data. A basic machine learning technique for assessing a model's performance is to divide the data into training and testing sets. The machine learning model (the regression tree) is trained on the training set, and its performance on untested data is assessed on the testing set. The training dataset is fed into the regression tree algorithm as part of the training process. Recursively dividing the data according to features yields decision nodes in the tree that forecast the target variable, which in regression is a continuous variable. The performance of the model must be assessed once the regression tree has been trained using the training set of data. The testing set is useful in this situation. Regression tree generalization to new, unknown data is evaluated using the testing set, which the model has not seen during training. The effectiveness of the regression tree on the testing set can be evaluated using a variety of indicators. Mean Squared Error (MSE), Mean Absolute Error (MAE), or R-squared are often used metrics in regression tasks. These metrics measure how much the actual values in the testing set depart from the projected values.

## 4. Application on Insurance Data

In this section, we employ a comprehensive analysis of complementary health insurance data utilizing both GLM and regression trees within a machine learning framework to evaluate the risk factors involved in the estimation of premium amounts for an insurance company.

We aim to enhance the accuracy and reliability of premium estimations, thereby catching more effective risk assessment in the estimation by using this integrated approach, harnesses the strengths of both GLM and regression trees, leveraging machine learning techniques.

Before implementation, the data set was preprocessed. Specifically, it was focused on duplicated and inconsistent data. In particular, incorrect information regarding impossible situations for employment and

the marital status of age groups was eliminated or corrected. After all corrections and data pre-processing, it was decided to conduct an examination on a sample set that would explain the entire portfolio.

All analyses were performed with the relevant packages within R programming [51].

### *4.1. Data*

The data used in this study is complementary health insurance data from an insurance company that operates in Turkey for the period 2019-2023. The data sample was requested by the private insurance company for study purposes only. Many variables, both continuous and categorical, are included in the data sample that has been used for the analysis, as per policy. The categorical variables used in this study include the region, employment status, age group, BMI group, marital status, and gender.

Since the data set included both category and numerical variables, the features were displayed independently. Categorical features were found for every subcategory, while numerical features were represented using minimum, maximum, median, mean, and standard deviation values.

Table 1 displays the categories of categorical variables and the circumstances that were considered while assigning the classes.

**Table 1.** The categories of categorical variables

| Category | Sub-Category | Descriptions |
|---|---|---|
| Region | Aegean, Black Sea, Central Anatolia, Eastern, Marmara, Mediterranean, Southeast | It is classified according to 7 main regions in Turkey |
| Employment status | Infant, Student, Teacher , Officer, Blue Collar, White Collar Retired, Unemployment , Other | Ages 0-6 are called infants.<br>The majority of the ages between 7 and 20 are students. |
| Age group | 00-06 ages; 07-20 ages; 21-25 ages; 26-30 ages; 31-35 ages; 36-40 ages; 41-45 ages; 46-50 ages; 51-55 ages; 56-60 ages; 61-65 ages; 65+ | For a more specific analysis, age groups were divided into 12. |
| Body mass index (BMIgroup) | Infant; Underweight, Normal weight, Overweight, Obesity, High obesity | BMI groups are divided into the following categories according to the BMI range - kg/m2, World Health Organization (WHO).<br>BMI range < 18.5 : Underweight<br>18.51<BMI range<24.99 : Normal weight<br>25<BMI range <29.99 : Overweight<br>30<BMI range<34.99 : Obesity<br>BMI range > 35 : High Obesity |
| Marital status | Child, Single, Married, Divorced, Widow | |
| Gender | Female<br>Male | |

The two most important variables in premium estimation, claim amount and claim number, were included in the analysis as continuous variables. The premium amounts that were planned to be estimated and are currently used by the company were included in the analysis as dependent variables. The number and type of categorical variables on gender-based differences from the dataset are shown in the Appendix, Table A.1. The statistics of the premium amount according to the type of categorical variables based on gender are shown in Appendix A.2.

The majority of machine learning algorithms proceed on the assumption that the predictor variables are independent of each other. Mutlicollinearity, or the removal of strongly correlated predictors, is an excellent way to make an analysis robust. The correlation matrix of the continuous variables is given in Figure 2. According to Table, although there is a higher relationship between the number of claims and claim amount variables than with all other continuous variables, it is obtained that none of the variables have a significant relationship with each other.

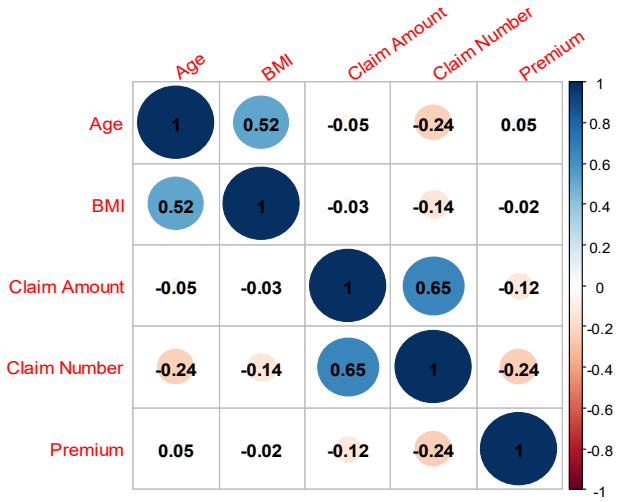



**Figure 2**. The correlation matrix of continous variables

*4.2. Generalized Linear Model Analysis*

In the GLM analysis, for the assumption of family and link functions, the premium amounts, which are the dependent variables, are visually shown to be suitable for certain distributions. In deciding which of the three available graphs is appropriate for the distribution of premium amounts, the visual consistency shown in Figure 3 is utilized. Among the default Weibull, gamma, and lognormal distributions, the gamma distribution, which is frequently used in the literature, was used together with the log link function.

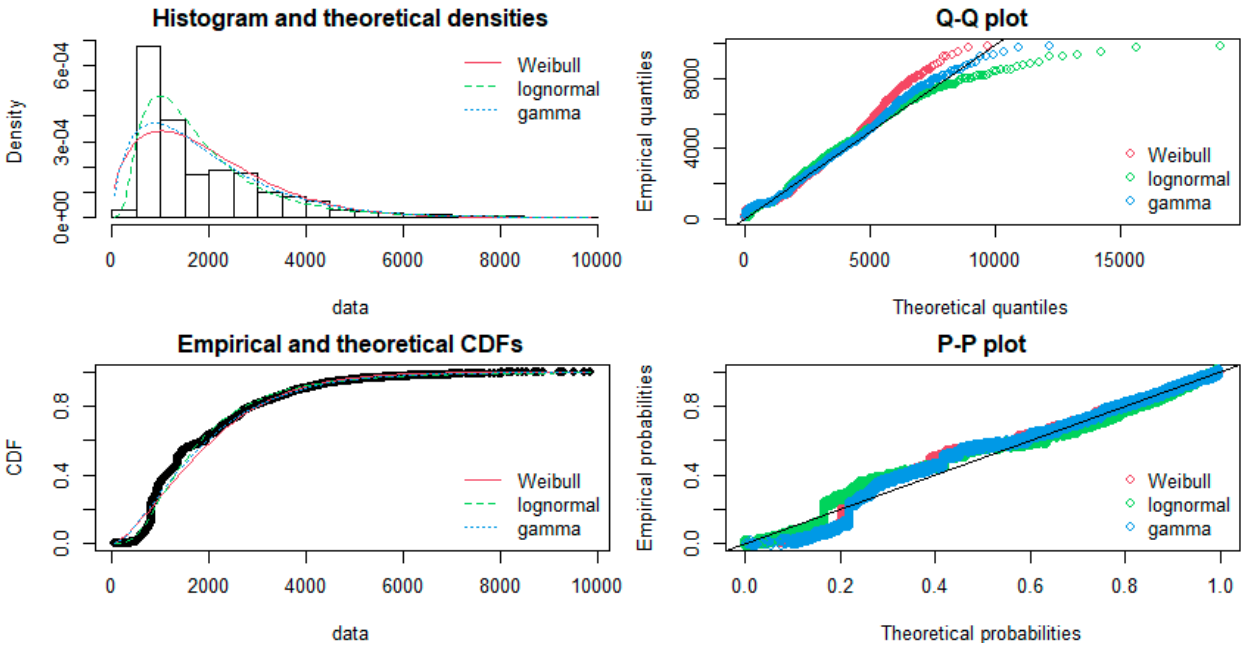



**Figure 3.** Histogram-density, P-P plot, Q-Q plot, and theoretical and empirical CDFs of premium amount

The GLM analysis results for the case where the premium amount is the dependent variable and all other variables are independent are shown in Table 2.

**Table 2**. Results of the GLM analysis

| Variable | Variable level | Estimate | Std. Error | t value | p |
|---|---|---|---|---|---|
| Renewal | Renewal | 0.3728 | 0.0253 | 14.7610 | < 2e-16 |
| Region | Black Sea | 0.0359 | 0.0480 | 0.7490 | 0.4541 |
| | Central Anatolia | -0.1025 | 0.0453 | -2.2650 | 0.0235 |
| | Eastern | -0.0643 | 0.0728 | -0.8830 | 0.3770 |
| | Marmara | 0.2081 | 0.0409 | 5.0900 | 0.0000 |
| | Mediterranean | 0.1587 | 0.0597 | 2.6580 | 0.0079 |
| | Southeast | 0.0238 | 0.0535 | 0.4440 | 0.6571 |
| Employment status | Infant | -2.1390 | 0.5798 | -3.6890 | 0.0002 |
| | Officer | -0.2930 | 0.0422 | -6.9370 | 0.0000 |
| | Other | -0.3204 | 0.0291 | -11.0230 | < 2e-16 |
| | Retired | -0.0333 | 0.1348 | -0.2470 | 0.8050 |
| | Student | -0.2739 | 0.0711 | -3.8500 | 0.0001 |
| | Teacher | -0.1813 | 0.0764 | -2.3730 | 0.0177 |
| | Unemployment | 0.1671 | 0.0907 | 1.8430 | 0.0654 |
| | White Collar | -0.5380 | 0.0299 | -18.0270 | < 2e-16 |
| Age | Age | -0.0100 | 0.0052 | -1.9190 | 0.0550 |
| Age Group | 07-20 ages | -1.8080 | 0.5181 | -3.4910 | 0.0005 |
| | 21-25 ages | -1.6720 | 0.5066 | -3.3000 | 0.0010 |
| | 26-30 ages | -1.5430 | 0.4953 | -3.1160 | 0.0018 |
| | 31-35 ages | -1.5200 | 0.4858 | -3.1300 | 0.0018 |
| | 36-40 ages | -1.4570 | 0.4773 | -3.0530 | 0.0023 |
| | 41-45 ages | -1.3930 | 0.4704 | -2.9620 | 0.0031 |
| | 46-50 ages | -1.1680 | 0.4650 | -2.5120 | 0.0120 |
| | 51-55 ages | -1.0230 | 0.4610 | -2.2190 | 0.0265 |
| | 56-60 ages | -1.0310 | 0.4597 | -2.2420 | 0.0250 |
| | 61-65 ages | -0.5294 | 0.5221 | -1.0140 | 0.3106 |
| BMI | BMI | -0.0186 | 0.0041 | -4.4930 | 0.0000 |
| BMI Group | Normal weight | -0.3237 | 0.1819 | -1.7790 | 0.0753 |
| | Obesity | 0.0729 | 0.1938 | 0.3760 | 0.7071 |
| | Overweight | -0.2454 | 0.1782 | -1.3770 | 0.1687 |
| | Underweight | -0.3528 | 0.1922 | -1.8350 | 0.0666 |
| Gender | Male | -0.0987 | 0.0192 | -5.1530 | 0.0000 |
| Marital Status | Divorced | 0.0090 | 0.1116 | 0.0810 | 0.9356 |
| | Married | 0.0154 | 0.0957 | 0.1610 | 0.8719 |
| | Single | -0.0171 | 0.0870 | -0.1970 | 0.8440 |
| | Widow | 0.2184 | 0.1603 | 1.3630 | 0.1731 |
| Claim | Claim Amount | 0.0000 | 0.0000 | 5.1400 | 0.0000 |
| Claim | Claim Number | -0.0756 | 0.0039 | -19.2490 | < 2e-16 |

The GLM, which are displayed in Table 1, indicate that the variables of employment status and region were ranked in order of significance in their respective subcategories, while the type of renewal, age group (apart from 61–65 years old), BMI, gender, claim amount, and claim number were found to be significant along with all of their subcategories. The variables age and BMI group were not found to be statistically significant, which is among the unexpected findings. The study will continue to determine whether the variables identified by GLM as significant have importance in the regression tree analysis. When the important variables are common, it will also look at how significant they are, how they are categorized into smaller groups, and how this classification affects the estimated premium amounts.

*4.3.Implementation of the CART algorithm*

*4.3.1. General Perspective using Regression Tree*

In the implementation of CART, two approaches were used in modeling the regression tree. First of all, a regression tree was created with only the variables whose importance was determined in GLM. In the second approach, modeling was applied depending on all variables in the data. Fortunately, the same results were obtained with both approaches. A visual representation of the decision tree obtained according to regression tree modeling is shown in Figure 1.

In modeling *mini split*, which refers to the minimum number of observations that are required at each node to split further, *maxdepth*, which is described as the length of the longest path from the tree root to a leaf, and the *complexity parameter (cp),* which is the minimum improvement in the model needed at each node, are applied as the control criteria. However, even possible changes in the control variables did not cause a change in the resulting regression tree.

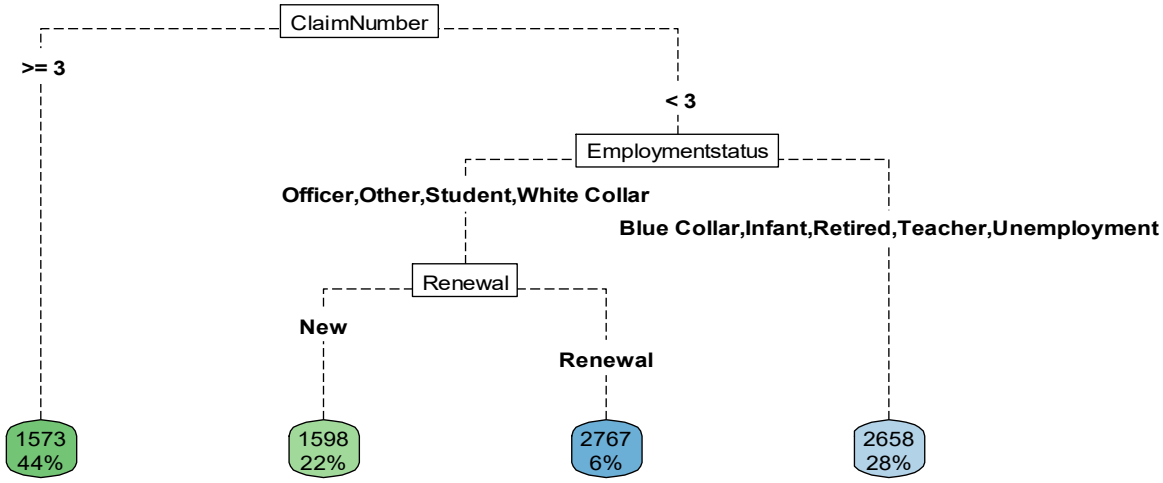



**Figure 4**. A visual representation of the regression tree for estimating premium amount

It can be clearly seen from Figure 4 that the premium is determined according to three important criteria in the regression tree. These variables are claim number, employment status, and renewal. These three variables are determined to be important in the GLM. However, it appears that not all variables determined to be important in GLM are taken into account in the regression tree classification.

*4.3.2. Application of Regression Tree for Prediction*

In the third part of the analysis, regression analysis is studied for prediction. In the analysis, first, the determination of the variable that will be the basis of the resampling method and the decision on the division ratios of the train and test data were examined.

Stratified sampling is a sampling technique used in statistical research in which the population is divided into subgroups, or "strata," based on certain characteristics, and then samples are randomly selected from each stratum. The goal is to ensure that each subgroup is represented in the sample proportionally to its presence in the overall population. Instead of separating the data into train and test as standard, we also used stratified sampling, in which the percentage of the number of observations in the categorical variables and subgroups that are important in the whole data is preserved in the selected sample. Stratified sampling is particularly useful when there are significant variations within the population and you want to ensure that each subgroup is adequately represented in the sample. This can lead to more accurate and reliable statistical

analysis, especially when dealing with diverse populations. When deciding which categorical variable to take into account in stratified sampling, the four most important categorical variables obtained from the GLM-age group, gender, employment status, and region- are taken into account. The observation and percentage densities of these four variables are also shown in Table 3.

**Table 3.** The observation and percentage densities of these four variables

| **AgeGroup** | | |
|---|---|---|
| **Sub-Category** | **freq** | **prob** |
| **00-06 ages** | **1023** | **20.50%** |
| 07-20 ages | 706 | 14.10% |
| 21-25 ages | 242 | 4.84% |
| 26-30 ages | 574 | 11.50% |
| 31-35 ages | 828 | 16.60% |
| 36-40 ages | 612 | 12.20% |
| 41-45 ages | 424 | 8.48% |
| 46-50 ages | 277 | 5.54% |
| 51-55 ages | 196 | 3.92% |
| 56-60 ages | 110 | 2.20% |
| 61-65 ages | 6 | 0.12% |
| 65+ | 2 | 0.04% |

| **Gender** | | |
|---|---|---|
| **Sub-Category** | **freq** | **prob** |
| **Female** | **2849** | **57%** |
| Male | 2151 | 43% |

| **Employment Status** | | |
|---|---|---|
| **Sub-Category** | **freq** | **prob** |
| **Blue Collar** | **1428** | **28.60%** |
| Infant | 1023 | 20.50% |
| Officer | 308 | 6.16% |
| Other | 760 | 15.20% |
| Retired | 24 | 0.48% |
| Student | 626 | 12.50% |
| Teacher | 74 | 1.48% |
| Unemployment | 53 | 1.06% |
| White Collar | 704 | 14.10% |

| **Region** | | |
|---|---|---|
| **Sub-Category** | **freq** | **prob** |
| Aegean | 276 | 5.52% |
| Black Sea | 539 | 10.80% |
| Central Anatolia | 854 | 17.10% |
| Eastern | 110 | 2.20% |
| **Marmara** | **2711** | **54.20%** |
| Mediterranean | 200 | 4% |
| Southeast | 310 | 6.20% |

Figure 5 shows the density of changes in premium amounts in relation to age groups for the variables employment status and region.

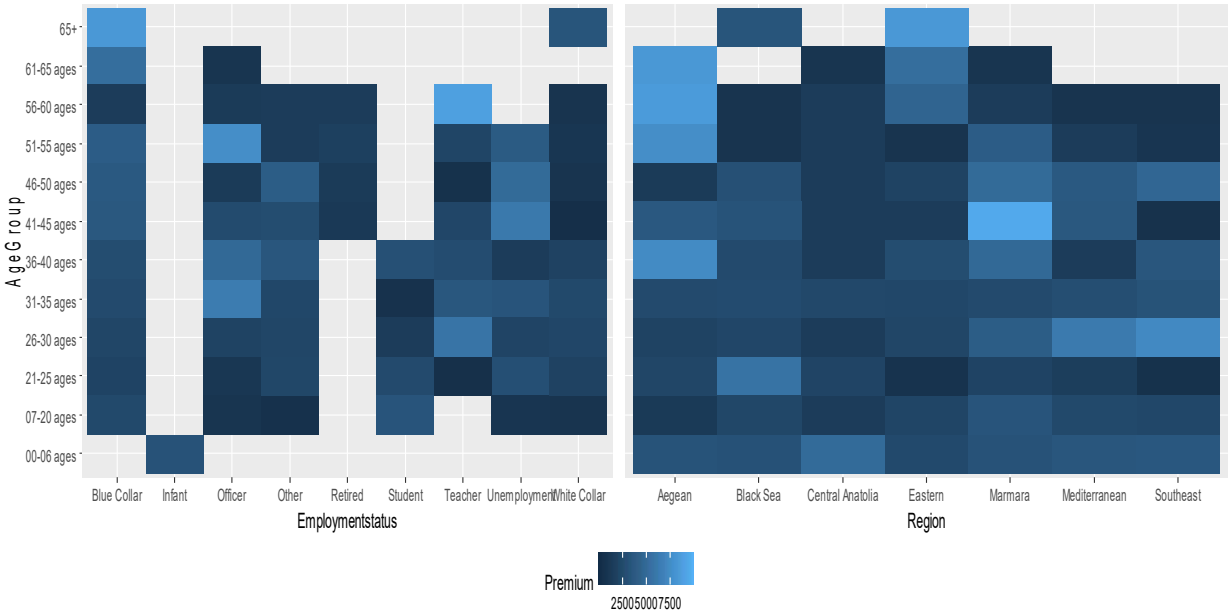



**Figure 5** Changes in premium amounts in relation to age groups for the variables employment status and region.

Employment status and region have been determined to be the categorical variables with the highest degree of significance, and the effect whose percentage distribution within the sample was wanted to be explored in stratified sampling.

In regression tree algorithms, the observations should be split into training and testing data to prevent overfitting. The percentage of data used for training and testing for the validity of the model is based on the size of the dataset, the complexity of the model, and the desired performance metrics. The training dataset should be larger to have a better machine learning rate than the test dataset. Any train-test split that has more data in the training set will most likely give better accuracy as calculated on the test set. Unlike the previous standard perspective on regression trees, this section analyses train and test data in order to make predictions.

In deciding the split ratios of train and test data, the RMSE and MAE values of the prediction values were examined. One of the most successful approaches to determining the most appropriate regression tree is to decide on the appropriate split percentage by comparing the MAE and RMSEs depending on the different split ratios of the train and test data, respectively. When the performance metrics for the most commonly used split ratios of (70%–30%), (80%–20%), and (90%–10%) are compared, it is assumed that the split ratio of 70%–30%, which gives the minimum value in test errors, is appropriate and sampling is performed.

In stratified sampling, where employment status is selected as the strata, the results of the regression tree with machine learning for the estimation of premium amounts in line with the 70%–30% split ratio are displayed in Figure 6.

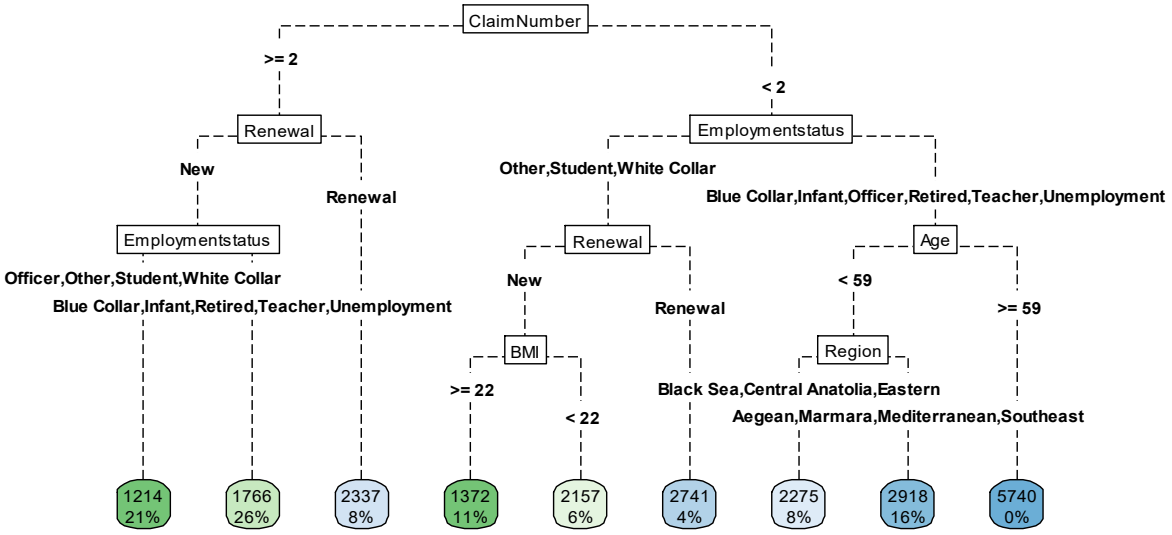



**Figure 6.** A visual representation of the regression tree with machine learning for estimating premium amount (Employment Status as Strata)

According to the results of the regression tree analysis, the claim number variable is the most prioritized variable, as in the general perspective using regression tree analysis. The first criterion is that the claim number is greater than or equal to 2. Then, unlike the standard perspective of regression analysis, it is determined that changes in premium amounts should be observed depending on the subcategories of renewal and employment status. Compared to the general perspective using regression tree, it is seen that BMI, age, and region variables, which are also found to be important in the GLM analysis but not included in the standard perspective regression analysis, are also important variables in the estimation of premium amounts. The results obtained show that machine learning-based regression tree analysis provides a much more comprehensive and detailed analysis by giving importance to different risk variables than general perspective using regression tree analysis.

Figure 7 presents the results of the regression tree used to estimate premium amounts in accordance with the 70%–30% split ratio in stratified sampling, where the region has been chosen as the strata.

Similar to the results in the regression tree obtained for employment status, the variability starts according to the claim number. After the claim number, employment status, renewal, age, and BMI are taken into account as variables, respectively. Surprisingly, region is not selected as a prior criterion, even in the analysis where region is selected as the base layer. The region variable was also determined as the lowest level criterion in the stratified analysis according to employment status. In fact, this analysis shows that even a variable that is determined to be important in the GLM and has high heterogeneity due to its subcategories can be evaluated after other variables in premium estimation. The result of the analysis may also vary depending on the sample size, the type of insurance, and the risk factors considered.

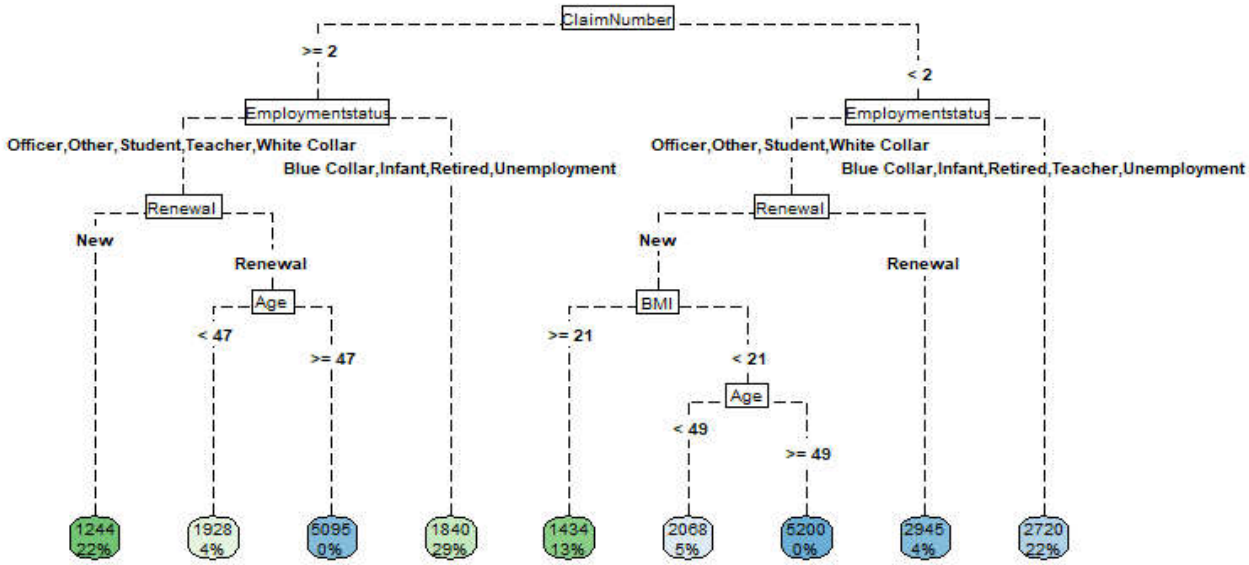



**Figure 7.** A visual representation of the regression tree with machine learning for estimating premium amount (Region as Strata)

A thorough investigation of the differences between the application of regression tree for prediction and general perspective using regression trees has shown that there are notable disparities in their predictive capacities. After a careful analysis, it is clear that prediction with machine learning, combining its sophisticated algorithms and methods, performs better than traditional regression trees in terms of accuracy and the capacity to pinpoint important factors.

The analysis conducted on these models indicates that machine learning-based regression trees perform more effectively by providing more accurate predictions and effectively detecting common important variables, akin to GLMs. This implies that machine learning techniques have a distinct advantage in identifying complex relationships and patterns in the data, which standard perspective regression models may struggle to discern.

One noteworthy finding is the machine learning models' proficiency in capturing variables that standard perspective regression models might overlook. The adaptability and flexibility of machine-learning algorithms enable them to handle intricate relationships and nonlinearities present in real-world datasets, leading to more nuanced and accurate predictions.

## 5. Conclusions and Recommendations

In the context of the non-life and health insurance sectors, the implications of adopting machine learning methodologies are particularly significant. Insurance data is often characterized by its complexity, with numerous variables interplaying to determine outcomes. Machine learning and regression trees, through their ability to handle diverse and intricate datasets, prove invaluable in accurately modeling and predicting outcomes in this sector.

This study represents a comprehensive investigation that attempts to decipher the complexity involved in risk assessment and premium calculation, starting with the fundamental GLM and ending with the creative incorporation of machine learning-driven regression trees. To begin our investigation, we first carefully looked at variables of importance using GLM, which provided a formal framework for comprehending the complex network of factors affecting health insurance premiums. The knowledge gathered from this first stage not only laid a strong foundation for further investigations, but it also highlighted how crucial variable priority is to the actuarial decision-making process. The variable importance comparison between regression trees and GLM supplied insightful information on the different viewpoints that each methodology presented, paving the way for a more in-depth comprehension of the variables influencing premium estimates. We took a step farther in terms of technological innovation and included machine learning into our regression tree method. This combination improved our model's ability to adjust to the changing and dynamic health insurance market while also improving the precision of our premium estimation.

Furthermore, this study underscores the potential for further improvement by expanding the scope of data and considering different types of insurance. The application of machine learning perspectives can be extended to other insurance sectors, fostering a more comprehensive understanding of the intricacies involved. By incorporating diverse datasets and varying insurance contexts, researchers can refine their models to enhance predictive accuracy and relevance across a broader spectrum. Regression trees and machine learning are expected to play an increasingly important role in actuarial science as technology advances, providing fresh perspectives and creative approaches to the problems associated with risk assessment and management.

## Kaynaklar


[1] P. McCullagh, J. A. Nelder,1989, *Generalized Linear Models 2nd ed.*. London: Chapman and Hall.

[2] A. E. Renshaw ,1991, Actuarial graduation practice and generalized linear and non-linear models. J Inst. Act., 118, 295-312.

[3] A. E. Renshaw, P. Verrall, 1994, A Stochastic Model Underlying The Chain Ladder Technique. In *Proceedings of the XXV ASTIN Colloquium, Cannes*.

[4] S. Haberman, A. E. Renshaw, 1996, *Generalized Linear Models and Actuarial Science*. *Journal of the Royal Statistical Society. Series D The Statistician*, 454, 407–436. https://doi.org/10.2307/2988543

[5] A. J. Dobson, 2002, *An Introduction to Generalized Linear Models Second Edition*. London: Chapman and Hall/CRC.

[6] D. Andersen, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi, 2005, *A Practitioner's Guide to Generalized Linear Models Second Edition*. CAS Study Note.

[7] K. Antonio, J. Beirlant, 2007, Actuarial statistics with generalized linear mixed models. *Insurance Mathematics & Economics*, 40, pp. 58-76. https://doi.org/10.1016/J.INSMATHECO.2006.02.013.

[8] P. De Jong, G. Heller, 2008, *Generalized Linear Models for Insurance Data International Series on Actuarial Science*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511755408

[9] M. V. Wüthrich, M. Merz, 2008, *Stochastic claims reserving methods in insurance*. John Wiley & Sons.

[10] E. Ohlsson, B. Johansson, 2010, *Non-Life Insurance Pricing with Generalized Linear Models*. Springer.

[11] E. W. Frees, 2015, *Analytics of insurance markets*. *Annual Review of Financial Economics*, 7, 253–77

[12] Z. Quan, Insurance Analytics with Tree-Based Models. PhD thesis, University of Connecticut, 2019.

[13] W. Gardner, C. Lidz, E. Mulvey, E. C. Shaw, 1996, A comparison of actuarial methods for identifying repetitively violent patients with mental illnesses. *Law and Human Behavior*, 20, 35-48.

[14] H. Steadman, E. Silver, J. Monahan, P. Appelbaum, P. Robbins, E. Mulvey, T. Grisso, L. Roth, S. Banks, 2000, A Classification Tree Approach to the Development of Actuarial Violence Risk Assessment Tools. *Law and Human Behavior*, 24, 83-100. https://doi.org/10.1023/A:1005478820425.

[15] L. Guelman, 2012, Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, 393, 3659–67.

[16] J. William, M. Martin, C. Chojenta, D. Loxton, 2018, *An actuarial investigation into maternal hospital cost risk factors for public patients*. *Annals of Actuarial Science*, 12, 106 - 129. https://doi.org/10.1017/S174849951700015X.

[17] M. V. Wuthrich, C. Buser, 2023, *Data Analytics for Non-Life Insurance Pricing*. Swiss Finance Institute Research Paper No. 16-68. Available at SSRN: https://ssrn.com/abstract=2870308 or http://dx.doi.org/10.2139/ssrn.2870308

[18] L. Diao, C. Weng, 2019, *Regression Tree Credibility Model*. *North American Actuarial Journal*, 232, 169-196. DOI: 10.1080/10920277.2018.1554497

[19] J. Baillargeon, L. Lamontagne, É. Marceau, 2020, *Mining Actuarial Risk Predictors in Accident Descriptions Using Recurrent Neural Networks*. *Risks*. https://doi.org/10.3390/risks9010007.

[20] S. Tober, 2020, *Tree-based Machine Learning Models with Applications in Insurance Frequency Modelling Dissertation*. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-276233

[21] R. Henckaerts, M.-P. Côté, K. Antonio, R. Verbelen, 2021, *Boosting Insights in Insurance Tariff Plans with Tree-Based Machine Learning Methods*. *North American Actuarial Journal*, 252, 255-285. DOI: 10.1080/10920277.2020.1745656

[22] B. Rokicki, K. Ostaszewski, 2022, *Actuarial Credibility Approach in Adjusting Initial Cost Estimates of Transport Infrastructure Projects*. *Sustainability*. https://doi.org/10.3390/su142013371.

[23] R. Richman, 2021a, *AI in actuarial science—a review of recent advances—part 1*. *Ann. Actuar. Sci.*, 152, 207–29

[24] R. Richman, 2021b, *AI in actuarial science—a review of recent advances—part 2*. *Ann. Actuar. Sci.*, 152, 230–58

[25] B. Wong, J. Christopher, H. Cossette, L. Lamontagne, E. Marceau, 2021, *Machine Learning in P&C Insurance: A Review for Pricing and Reserving*. *Risks*, 91, 4. https://doi.org/10.3390/risks9010004

[26] Z. Quan, 2019, *Insurance Analytics with Tree-Based Models Doctoral Dissertations No. 2374*. Retrieved from https://digitalcommons.lib.uconn.edu/dissertations/2374

[27] J. Neyman, 1934, On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625

[28] J. Neyman, E. S. Pearson, 1933, On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289–337. http://www.jstor.org/stable/91247

[29] R. Singh, N. S. Mangat, 1996, *Stratified Sampling*. In: *Elements of Survey Sampling*, Vol. 15. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-1404-4_5

[30] V. L. Parsons, 2014, *Stratified sampling*. *Wiley StatsRef: Statistics Reference Online*, 1-11.

[31] E. Liberty, K. Lang, K. Shmakov, 2016, June. *Stratified sampling meets machine learning*. In *International conference on machine learning* pp. 2320-2329. PMLR.

[32] Y. Ye, Q. Wu, J. Z. Huang, M. K. Ng, X. Li, 2013, *Stratified sampling for feature subspace selection in random forests for high dimensional data*. *Pattern Recognition*, 463, 769-787.

[33] T. Yu, X. Zhai, S. Sra, 2019, *Near Optimal Stratified Sampling*. *ArXiv, abs/1906.11289*.

[34] Y. Lu, Y. Park, L. Chen, Y. Wang, C. De Sa, D. Foster, 2021, July. *Variance reduced training with stratified sampling for forecasting models*. In *International Conference on Machine Learning* pp. 7145-7155. PMLR.

[35] J. Fox, 2008, *Applied Regression Analysis and Generalized Linear Models*, 2nd Edn. Thousand Oaks, CA: Sage.

[36] J.F. Magee, 1964, *Decision trees for decision making, Harvard Business Review, pp. 126-138.*

[37] S.K. Murthy, 1998, Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Mining Knowl Discovery* 2(4):345–389

[38] R.L. Keeney, 1982, Decision Analysis: An Overview. *Operations Research*, *30*(5).
[39] L.Tjen-Sien, L. Wei-Yin, S.Yu-Shan, 2000, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Mach Learn* 40:203–228
[40] S.B. Kotsiantis, 2013, Decision trees: a recent overview. *Artif Intell Rev* 39, 261–283
[41] L. Breiman, J. Friedman, R. Olshen, C. J. Stone, 1984, *Classification and regression Trees*. Wadsworth, Belmont, CA.
[42] J.R. Quinlan, 1986, Induction of decision trees. *Mach Learn 1*, 81–106.
[43] J.R. Quinlan, 1993, *C4.5: programs for machine learning*. Morgan Kaufmann, San Francisco
[44] G.V. Kass, 1980, "An Exploratory Technique for Investigating Large Quantities of Categorical Data". *Applied Statistics.* 29 (2): 119–127
[45] J. Gehrke, R. Ramakrishnan, V. Ganti, 2000, RainForest: a framework for fast decision tree construction of large datasets. *Data Mining Knowl Discovery* 4(2–3):127–162
[46] H. A. Chipman, E. I. George, R. E. McCulloch, 1998, *Bayesian CART model search*. *Journal of the American Statistical Association*, 93443, 935-960 pp.
[47] J. Morgan, 2014, *Classification and regression tree analysis*. Boston: Boston University, 298.
[48] D. L. Verbyla, 1987, *Classification trees: a new discrimination tool*. *Canadian Journal of Forest Research,* 17, 9, 1150–1152.
[49] L. A. Clark, D. Pregibon, 1992, *Tree-based models*. In: *Statistical models* Eds. Chambers JM, Hastie TJ. Pacific Grove, CA: Wadsworth, p 377–419.
[50] G. De'ath, K. E. Fabricius, 2000, *Classification and Regression Trees: A Powerful yet Simple* Technique for Ecological Data Analysis. Ecology, 81, 3178-3192
[51] R Core Team , 2021, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. RStudio 2023.09.1

# Appendix

**Table A.1** The number and the type of categorical variables on gender-based

| Categories | Sub-categories | Gender | | | | |
|---|---|---|---|---|---|---|
| | | Female | | Male | | Total |
| | | Number of Observations | Percentage (%) | Number of Observations | Percentage (%) | Number of Observations |
| Renewal | New | 2423 | 57,4% | 1796 | 42,6% | 4219 |
| | Renewal | 426 | 54,5% | 355 | 45,5% | 781 |
| Region | Aegean | 154 | 55,8% | 122 | 44,2% | 276 |
| | Black Sea | 317 | 58,8% | 222 | 41,2% | 539 |
| | Central Anatolia | 506 | 59,3% | 348 | 40,7% | 854 |
| | Eastern | 49 | 44,5% | 61 | 55,5% | 110 |
| | Marmara | 1560 | 57,5% | 1151 | 42,5% | 2711 |
| | Mediterranean | 113 | 56,5% | 87 | 43,5% | 200 |
| | Southeast | 150 | 48,4% | 160 | 51,6% | 310 |
| Employment statue | Blue Collar | 947 | 66,3% | 481 | 33,7% | 1428 |
| | Infant | 489 | 47,8% | 534 | 52,2% | 1023 |
| | Officer | 188 | 61,0% | 120 | 39,0% | 308 |
| | Other | 445 | 58,6% | 315 | 41,4% | 760 |
| | Retired | 10 | 41,7% | 14 | 58,3% | 24 |
| | Student | 311 | 49,7% | 315 | 50,3% | 626 |
| | Teacher | 56 | 75,7% | 18 | 24,3% | 74 |
| | Unemployment | 33 | 62,3% | 20 | 37,7% | 53 |
| | White Collar | 370 | 52,6% | 334 | 47,4% | 704 |
| Age group | 00-06 ages | 489 | 47,8% | 534 | 52,2% | 1023 |
| | 07-20 ages | 347 | 49,2% | 359 | 50,8% | 706 |
| | 21-25 ages | 163 | 67,4% | 79 | 32,6% | 242 |
| | 26-30 ages | 403 | 70,2% | 171 | 29,8% | 574 |
| | 31-35 ages | 540 | 65,2% | 288 | 34,8% | 828 |
| | 36-40 ages | 335 | 54,7% | 277 | 45,3% | 612 |
| | 41-45 ages | 250 | 59,0% | 174 | 41,0% | 424 |
| | 46-50 ages | 154 | 55,6% | 123 | 44,4% | 277 |
| | 51-55 ages | 110 | 56,1% | 86 | 43,9% | 196 |
| | 56-60 ages | 54 | 49,1% | 56 | 50,9% | 110 |
| | 61-65 ages | 3 | 50,0% | 3 | 50,0% | 6 |
| | 65+ | 1 | 50,0% | 1 | 50,0% | 2 |
| BMI group | High obesity | 7 | 50,0% | 7 | 50,0% | 14 |
| | Infant | 489 | 47,8% | 534 | 52,2% | 1023 |
| | Normal weight | 1827 | 64,1% | 1022 | 35,9% | 2849 |
| | Obesity | 15 | 29,4% | 36 | 70,6% | 51 |
| | Overweight | 329 | 43,8% | 422 | 56,2% | 751 |
| | Underweight | 182 | 58,3% | 130 | 41,7% | 312 |
| Marital status | Child | 783 | 48,0% | 849 | 52,0% | 1632 |
| | Divorced | 96 | 73,3% | 35 | 26,7% | 131 |
| | Married | 1639 | 62,0% | 1003 | 38,0% | 2642 |
| | Single | 307 | 54,0% | 262 | 46,0% | 569 |
| | Widow | 24 | 92,3% | 2 | 7,7% | 26 |
| Gender | Male | 0 | 0,0% | 2151 | 100,0% | 2151 |
| | Female | 2849 | 100,0% | 0 | 0,0% | 2849 |

**Table A.2** The statistics of the premium amount according to type of categorical variables on gender-based

| | | Gender | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Female | | | | Male | | | |
| | | Minimum | Maximum | Mean | Standard Deviation | Minimum | Maximum | Mean | Standard Deviation |
| Region | Aegean | 408,87 | 8706,76 | 1990,77 | 1817,32 | 421,00 | 8383,63 | 1712,23 | 1291,11 |
| | Black Sea | 207,04 | 9229,54 | 1899,28 | 1471,04 | 381,94 | 5279,54 | 1591,23 | 1043,38 |
| | Central Anatolia | 58,92 | 8451,76 | 1580,52 | 1268,81 | 170,65 | 9742,52 | 1537,22 | 1154,88 |
| | Eastern | 509,29 | 6255,22 | 1775,35 | 1145,54 | 502,81 | 8171,71 | 1857,61 | 1416,85 |
| | Marmara | 128,45 | 9846,17 | 2163,27 | 1544,13 | 121,44 | 8720,90 | 2051,50 | 1395,82 |
| | Mediterranean | 279,91 | 7656,97 | 2392,99 | 1857,63 | 400,00 | 6753,05 | 1849,52 | 1334,71 |
| | Southeast | 400,00 | 7170,24 | 2092,06 | 1561,73 | 422,81 | 9301,10 | 1607,51 | 1334,53 |
| Employment status | Blue Collar | 162,34 | 9846,17 | 2495,51 | 1776,09 | 381,94 | 8171,71 | 2136,76 | 1392,46 |
| | Infant | 80,60 | 8021,25 | 2016,56 | 1353,37 | 157,55 | 9301,10 | 2052,79 | 1405,56 |
| | Officer | 197,67 | 8706,76 | 1726,85 | 1593,68 | 170,65 | 8383,63 | 1399,38 | 1004,46 |
| | Other | 66,33 | 8803,85 | 1686,88 | 1356,41 | 148,09 | 8187,70 | 1707,33 | 1291,06 |
| | Retired | 1211,75 | 5072,30 | 2764,98 | 1462,63 | 800,00 | 9742,52 | 2756,14 | 2489,18 |
| | Student | 58,92 | 4921,72 | 1706,66 | 1005,75 | 121,44 | 5597,49 | 1767,53 | 1075,69 |
| | Teacher | 306,99 | 7971,17 | 1872,25 | 1570,70 | 732,76 | 8720,90 | 2481,27 | 1873,05 |
| | Unemployment | 549,32 | 6121,14 | 2355,29 | 1516,25 | 1123,09 | 6017,00 | 2903,40 | 1281,82 |
| | White Collar | 532,97 | 9508,34 | 1590,98 | 1259,57 | 400,00 | 8041,72 | 1383,25 | 1149,38 |
| Age group | 00-06 ages | 80,60 | 8021,25 | 2016,56 | 1353,37 | 157,55 | 9301,10 | 2052,79 | 1405,56 |
| | 07-20 ages | 58,92 | 4921,72 | 1664,83 | 993,24 | 121,44 | 5597,49 | 1770,35 | 1074,25 |
| | 21-25 ages | 162,34 | 9229,54 | 2085,46 | 1634,46 | 400,00 | 4187,46 | 1434,33 | 869,77 |
| | 26-30 ages | 197,67 | 8538,31 | 2084,77 | 1766,13 | 170,65 | 4621,44 | 1467,33 | 980,19 |
| | 31-35 ages | 66,33 | 8706,76 | 1917,77 | 1433,89 | 148,09 | 5614,11 | 1597,07 | 1030,86 |
| | 36-40 ages | 155,91 | 9508,34 | 2016,92 | 1512,32 | 563,75 | 5801,39 | 1671,15 | 1153,47 |
| | 41-45 ages | 595,00 | 9290,45 | 2147,73 | 1530,49 | 155,43 | 8187,70 | 1837,66 | 1400,53 |
| | 46-50 ages | 207,04 | 8803,85 | 2239,93 | 1709,49 | 595,00 | 6790,11 | 2286,24 | 1579,41 |
| | 51-55 ages | 400,00 | 8536,70 | 2505,87 | 2114,11 | 595,00 | 8041,72 | 2405,58 | 1822,79 |
| | 56-60 ages | 800,00 | 9846,17 | 2377,62 | 2110,55 | 344,20 | 9742,52 | 2495,62 | 2287,57 |
| | 61-65 ages | 850,00 | 8164,46 | 3288,15 | 4223,01 | 4382,26 | 5586,93 | 5065,45 | 618,40 |
| | 65+ | 3372,89 | 3372,89 | 3372,89 | . | 8171,71 | 8171,71 | 8171,71 | . |
| BMII group | High obesity | 408,87 | 4313,85 | 1923,88 | 1559,13 | 1315,24 | 5381,07 | 2408,08 | 1455,94 |
| | Infant | 80,60 | 8021,25 | 2016,56 | 1353,37 | 157,55 | 9301,10 | 2052,79 | 1405,56 |
| | Normal weight | 128,45 | 9508,34 | 1987,61 | 1546,90 | 121,44 | 9742,52 | 1745,83 | 1278,02 |
| | Obesity | 472,77 | 7663,38 | 2905,90 | 2227,08 | 589,22 | 6790,11 | 2374,88 | 1635,28 |
| | Overweight | 66,33 | 9846,17 | 2209,16 | 1755,73 | 155,43 | 8720,90 | 1854,93 | 1364,45 |
| | Underweight | 58,92 | 9290,45 | 1939,44 | 1289,66 | 421,00 | 5279,54 | 1724,93 | 1085,18 |
| Marital status | Child | 58,92 | 8021,25 | 1897,84 | 1247,26 | 121,44 | 9301,10 | 1943,42 | 1291,46 |
| | Divorced | 595,00 | 8472,84 | 2249,49 | 1682,69 | 525,68 | 5283,10 | 1892,55 | 1261,13 |
| | Married | 66,33 | 9508,34 | 2043,88 | 1631,43 | 148,09 | 9742,52 | 1833,54 | 1420,07 |
| | Single | 480,32 | 8803,85 | 2032,34 | 1414,23 | 400,00 | 5802,25 | 1630,47 | 1042,95 |
| | Widow | 595,00 | 9846,17 | 3270,39 | 2748,35 | 1478,53 | 5897,23 | 3687,88 | 3124,49 |
| Gender | Female | 58,92 | 9846,17 | 2019,76 | 1532,21 | . | . | . | . |
| | Male | . | . | . | . | 121,44 | 9742,52 | 1854,86 | 1331,16 |

**İstatistikçiler Dergisi: İstatistik & Aktüerya**

Journal of Statisticians: Statistics and Actuarial Sciences

**IDIA 16**, 2023, 2, 100-115

**Geliş** / Received: 05.12.2023, **Kabul** / Accepted: 29.12.2023

**Araştırma Makalesi /** Research Article

# Veri madenciliği yöntemleri ile bir melez sınıflandırma yaklaşımı ve uygulaması

Gözde Ulu Metin[1]

*Ankara Üniversitesi, Fen Bilimleri Enstitüsü,*
*İstatistik Anabilim Dalı,*
*Ankara, Türkiye*
*ulumetin@ankara.edu.tr*
ORCID: 0000-0003-0384-9504

Özlem Türkşen

*Ankara Üniversitesi, Fen Fakültesi,*
*İstatistik Bölümü,*
*Ankara, Türkiye*
*turksen@ankara.edu.tr*
ORCID: 0000-0002-5592-1830

**Öz**

Son yıllarda hızla artan büyüklükteki veri setlerinden bilgi keşfetmek oldukça değerlidir. Veri madenciliği yöntemleri, sınıflandırma problemlerinde, büyük ve karmaşık veri setlerindeki gizli örüntünün ortaya çıkarılarak verilerin belli bir sınıfa atanması amacıyla kullanılır. Bu çalışmada, kurumların başarım değerlendirilmesi sürecine istatistiksel bakış açısı kazandırmak amacıyla veri madenciliği yöntemleri ile Analitik Hiyerarşi Süreci (AHP) ve CODAS yöntemleri kullanılarak bir melez sınıflandırma yaklaşımı önerilmiştir. Uygulama amacıyla bir kurum verisi ele alınmıştır. Veri seti ön işleme aşamasından geçirilerek, veri setindeki değişkenler, uzman bilgisi dikkate alınarak AHP yöntemi ile ağırlıklandırılmıştır. Ağırlıklandırılmış gerçek veri setine, veri madenciliği sınıflandırma yöntemlerinden Lojistik Regresyon (LR), K-En Yakın Komşu (KNN) algoritması, Destek Vektör Makineleri (SVM) ve Rastgele Orman (RF) algoritması uygulanmıştır. Sınıflandırma yöntemleri, 5-kat çapraz doğrulama sonucu elde edilen doğruluk, kesinlik, duyarlılık ve $F_1$-skor performans ölçütlerine göre hesaplanmıştır. Elde edilen performans ölçütleri, çok ölçütlü karar verme yöntemi olan CODAS'a göre değerlendirilmiştir. Yapılan melez sınıflandırma yaklaşımına göre, Ar-Ge ve Tasarım merkezlerinin faaliyetlerinin değerlendirilmesi konusunda RF yönteminin daha iyi sınıflandırma performansına sahip olduğu görülmüştür.

***Anahtar sözcükler:*** AHP, CODAS, Çok ölçütlü karar verme, Melez sınıflandırma, Performans ölçütleri, Veri madenciliği yöntemleri, Veri ön işleme

[1] Bu çalışma, birinci yazarın, ikinci yazarın danışmanlığında hazırladığı doktora tezinden üretilmiştir.

**Abstract**

***A hybrid classification approach with data mining methods and an application***

*In recent years, it is very valuable to discover information from data sets of rapidly increasing size. Data mining methods are used in classification problems to assign data to a certain class by revealing the hidden pattern in large and complex data sets. In this study, a hybrid classification approach is proposed by using data mining methods with Analytic Hierarchy Process (AHP) and CODAS methods in order to gain a statistical perspective on the performance evaluation process of the instutitions. An instutition data is taken as a basis for the application. The data set is preprocessed and the variables in the data set are weighted by AHP method by taking into account expert knowledge. Logistic Regression (LR), K-Nearest Neighbour (KNN) algorithm, Support Vector Machines (SVM) and Random Forest (RF) algorithm, data mining classification methods, were applied to the weighted real data set. The classification methods were calculated according to the accuracy, precision, sensitivity and $F_1$-score performance measures obtained from 5-fold cross-validation. The obtained performance criteria were evaluated according to the CODAS, a multi-criteria decision making method. As a result of the hybrid classification approach, it was seen that the RF method has better classification performance about the evaluation of the activities of R&D and Design centers.*

***Keywords:*** *AHP, CODAS, Multi criteria decision making, Hybrid classification, Performance metrics, Data mining methods, Data preprocessing*

## 1. Giriş

Günümüzde depolanmış veri setlerinde mevcut olan ve saklı kalan bilgileri ortaya çıkarmak oldukça kritik bir rol oynar. Fakat, oluşan büyük veri yığınlarında geleneksel veri analizi yöntemleri yetersiz kalmaktadır. Veri madenciliği, ham veride bulunan örüntüleri ortaya çıkarmak ve keşfetmek adına, özellikle sınıflandırma problemleri üzerinde etkili bir biçimde kullanılır. Veri madenciliği yöntemleri ile sınıflandırma yapılmadan önce keşfedici veri analizi ile veri ön işleme aşamalarının uygulanması gerekir. Çetin ve Yıldız [1] çalışmalarında, literatürde bulunan çok sayıda veri ön işleme yöntemleri ve algoritmaları üzerine kapsamlı bir inceleme yapmışlardır. Emeç ve Özcanhan [2] çalışmalarında, veri ön işleme yöntemlerini ayrıntılı olarak ele alıp yapılan uygulama sonucunda, veri ön işlemenin karar vermede daha doğru sonuçlar elde edilmesine yardımcı olduğunu belirtmişlerdir.

Veri ön işleme aşamasında, veri setinin yapısını anlamak, boyut azaltmak ya da gruplamak amacıyla farklı yöntemler kullanılabilir. Veri setindeki değişkenlerin önem ağırlıklarının hesaplanması ve subjektif değerlendirmelerden yararlanılması amacıyla Çok Ölçütlü Karar Verme (Multi Criteria Decision Making-MCDM) yöntemlerinin kullanılması da veri ön işleme sürecine dahil edilebilir. Bu çalışmada, bir MCDM yöntemi olan Analitik Hiyerarşi Süreci (Analytic Hierarchy Process-AHP) yöntemi ile uzman görüşü dikkate alınarak değişkenlerin önemine göre değişken ağırlıkları belirlenmiştir. Böylece, veri ön işleme sürecinde, MCDM yöntemleri kullanılarak veri setindeki değişkenlerin önemi hakkında önsel bilgi elde edinilmesi sağlanmıştır.

Veri madenciliği, veri yığınlarında veriye dayalı derinlemesine keşifler yapmayı, istatistiksel yöntemler ile örtülü bilgileri, veriden çıkarmayı amaçlar [3]. Han vd. [4] çalışmalarında, veri madenciliği kavramı ve yöntemleri detaylı olarak anlatılarak örneklerle açıklanmıştır. Çınar ve Silahtaroğlu [5] çalışmalarında, bir anket veri seti üzerinde veri madenciliği yöntemlerini uygulayarak, gizli kalmış örüntü ve nedenleri keşfetmişlerdir.

Veri madenciliği yöntemleri, denetimli öğrenme (supervised learning) ve denetimsiz öğrenmeden (unsupervised learning) oluşur. Denetimli öğrenme, sınıflandırma yöntemlerini, denetimsiz öğrenme de kümeleme ve birliktelik kurallarını içermektedir. Sınıflandırma çalışması yapılması istenen ön işlemesi yapılmış veri setinde, veri madenciliğinin denetimli öğrenme başlığı altında sınıflandırma algoritmaları kullanılır. Sınıflandırma performansının ölçülmesinde, Doğruluk (Accuracy), Duyarlılık (Sensivity),

Kesinlik (Precision) ve $F_1$-Skor ($F_1$-Score) ölçütleri hesaplanır. Nieto vd. [6] çalışmalarında, stratejik karar vermede denetimli sınıflandırma yöntemlerini kullanmışlardır. Gerçek veri seti ile yapılan çalışmada, performans ölçütlerine göre RF algoritmasının diğer yöntemlere göre daha iyi performansa sahip olduğu görülmüştür. Öztürk Zan [7] çalışmasında, gerçek veri seti üzerinde denetimli öğrenme algoritmasını uygulamıştır. 5-kat çapraz doğrulama (Cross Validation-CV) sonucu elde edilen performans ölçütlerine göre RF algoritmasının en iyi performansı gösterdiği belirtilmiştir. Yavuz vd. [8] çalışmalarında, karar destek sisteminin geliştirilmesi amacıyla, sınıflandırma yöntemlerinden Naive Bayes, KNN, Karar Ağacı ve RF kullanılarak performansları değerlendirilmiştir.

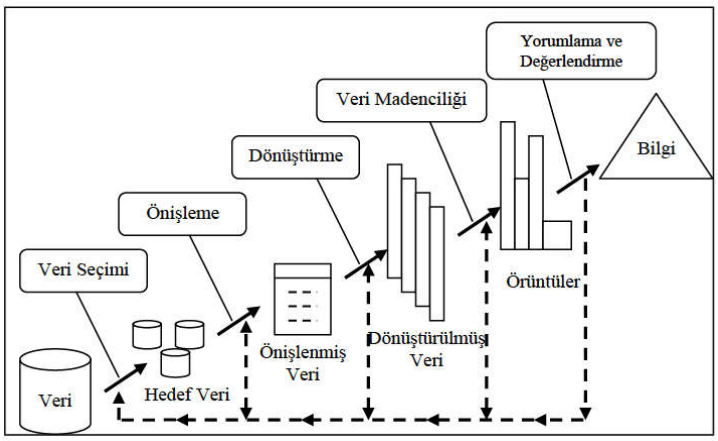
Yapılan çalışmalarda, çok ölçüte sahip karar seçenekleri arasından birine karar vermek oldukça zordur. MCDM, belirlenen bir amacı gerçekleştirmek için mevcut seçenekler arasından, belirlenen ölçütler dikkate alınarak en uygun olanın seçimine karar verme sürecidir. Karar verme sürecinde, MCDM yöntemlerinden CODAS (Combinative Distance-based Assesment) uygulanarak, sınıflandırma yöntemlerinin öncelikli sıralaması yapılır. Ulutaş [9] çalışmasında, bir tekstil şirketi için lojistik sağlayıcı seçiminde AHP ve CODAS yöntemlerini birlikte kullanmıştır. AHP ile elde edilen kriter ağırlıkları CODAS yöntemi ile birleştirilerek alternatifler sıralanmıştır. Can vd. [10] çalışmalarında, sağlık sektöründe altı sigma projelerinin önceliklendirilmesi ve seçimi için AHP ve CODAS yöntemlerini entegre eden bir hibrit karar verme modeli önermektedir. AHP yöntemi ile elde edilen kriter ağırlıkları kullanılarak CODAS uygulanmıştır.

Bu çalışmada, sınıflandırma yöntemlerinin MCDM yöntemleri ile ağırlıklandırılarak performans ölçütlerinin hesaplandığı ve performans ölçütlerinin değerlendirilmesinde MCDM yöntemi kullanılarak karar verilmesi konusunda, veri madenciliği yöntemleri ile MCDM yöntemleri kullanılarak bir melez sınıflandırma yaklaşımı önerilmiştir. Çalışmanın ikinci bölümünde, veri ön işleme ile veri setini etkileyen değişken ağırlıklarının hesaplanmasında kullanılan AHP yöntemine yer verilmiştir. Çalışmanın üçüncü bölümünde, veri madenciliği ile sınıflandırma başlığında Lojistik Regresyon (LR), K-En Yakın Komşu (KNN) algoritması, Destek Vektör Makineleri (SVM) ve Rastgele Orman (RF) algoritması açıklanmıştır. Sınıflandırma performans ölçütleri tanımlanarak performans ölçütlerinin karar verilmesinde kullanılan CODAS yöntemi anlatılmıştır. Çalışmanın dördüncü bölümünde, bir gerçek veri seti üzerinde uygulama yapılmıştır. Çalışmanın beşinci bölümünde ise sonuçlara yer verilmiştir.

## 2. Veri ön işleme

Bilgi teknolojisinin gelişimi ile verinin depolanma kapasitesi artarak büyük veri (big data) yapıları oluşmuştur. Elde edilen her yeni veri saklanmakta fakat, oluşan veri yığınından anlamlı bilgi çıkarmak zorlaşmaktadır. Veriden bilgi elde edilmesi süreci, bilgi keşfi olarak adlandırılır. Fayyad [11] çalışmasında, bilgi keşfi sayesinde oluşan veri yığınlarının etkili bir biçimde kullanılmasıyla değer elde edilmesinin önemi vurgulanmıştır. Kavurkacı vd. [12] çalışmalarında büyük veri işlemede kullanılan yöntemlere genel bir bakış açısı sunmuştur.

Veri madenciliği kavramı, yüksek kapasiteli verinin içerisindeki keşfedilmemiş bilgiyi ortaya çıkarmayı hedefler [13]. Veriden bilgi keşfi süreci olarak adlandırılan bu süreç aşamaları Şekil 1'de özetlenmiştir. Birinci aşamada araştırma konusuna yönelik elde edilen ham veriden, hedeflenen veri seçimi yapıldığı Şekil 1'den açıkça görülmektedir. Verideki bilgi keşfi sürecinin ön işleme aşamasında (veri temizleme, veri bütünleştirme, boyut azaltma, veri seçme) istatistiksel yaklaşımlara dayalı veri analizi yapılarak, hedef veride, eksik değer tamamlanarak hatalı, anlamsız değerler çıkartılır. Eğer, veri dönüşüm gerektiriyorsa, dönüştürme işlemi yapılarak dönüştürülmüş verilere ulaşılır.




**Şekil 1.** Veri-Bilgi Keşfi Süreci [18]

Çizelge 1'de verilen veri setindeki değişkenlerin ölçüldüğü birimler arasındaki farklılıkların giderilmesi amacıyla Z-skor, Min-Max gibi standartlaştırma yöntemleri kullanılarak veri seti, ölçü biriminden bağımsız hale getirilir. Veri hazırlama aşamasından sonra büyük verilerin analizini kolaylaştıran, gizli örüntü keşfini sağlayan temeli istatistiksel yöntemlere dayalı MCDM yöntemleri kullanılır. Keleş ve Tunca [14] çalışmalarında, bir Ar-Ge firmasının kuruluş aşamasında, işletmelerin görüşüne göre önemli değişkenlerin AHP yöntemi ile ağırlıklarını belirlemişlerdir. Arslan ve Belgin [15] çalışmalarında, AHP yöntemini, imalat sanayisindeki öncelikli teknoloji alanlarını etkileyen değişkenlerin ağırlıklandırılmasında kullanmıştır. Çalışma sonucunda elde edilen sıralamaya göre ilgili sektörlere sağlanan Ar-Ge, yenilik ve girişimcilik desteklerinde öncelik tanınması önerilmiştir. Güryeli [16] çalışmasında, Bilim, Sanayi ve Teknoloji Bakanlığı (günümüzde Sanayi ve Teknoloji Bakanlığı) tarafından yürütülen Teknolojik Ürün Yatırım Destek Programı desteği için sunulan Ar-Ge projelerinin seçim sürecini incelemiştir. Bu süreçte, alanında uzman akademisyenlerin görüşlerine göre AHP uygulanarak Ar-Ge projelerinin değerlendirilmesinde dikkate alınan değişkenlerin göreceli önem seviyeleri elde edilmiştir. Subjektif değerlendirmelerden yararlanılarak, veri setindeki değişkenlerin önem ağırlıklarının hesaplanması amacıyla AHP yöntemi uygulanır.

**Çizelge 1.** Çok değişkenli veri seti

| | ***Bağımsız değişkenler*** | | | | ***Bağımlı değişken*** |
|---|---|---|---|---|---|
| **No.** | $X_1$ | $X_2$ | ... | $X_m$ | $Y$ |
| 1 | $x_{11}$ | $x_{12}$ | ... | $x_{1m}$ | $Y_1$ |
| 2 | $x_{21}$ | $x_{22}$ | ... | $x_{2m}$ | $Y_2$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $n$ | $x_{n1}$ | $x_{n2}$ | ... | $x_{nm}$ | $Y_n$ |

AHP yöntemi, değişkenlerin ikili kıyaslamasını yaparken sözel ifadeleri sayısal değerler kullanarak ifade edip karşılaştırma yapan MCDM yöntemlerinden biridir [17]. AHP'de, Saaty tarafından önerilmiş olan karşılaştırma ölçeği ile subjektif değerlendirmeler, matrisler yardımıyla matematiksel olarak ifade edilir [18]. AHP uygulamak için Çizelge 1'de görülen $m$ sayıda değişken, Saaty [18] çalışmasında önerilen karşılaştırma ölçeğiyle birbirlerine göre önem değerleri dikkate alınarak

$$A=\begin{bmatrix} 1 & a_{12} & \cdots & a_{1m} \\ 1/a_{12} & 1 & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1/a_{1m} & 1/a_{2m} & \cdots & 1 \end{bmatrix} \tag{1}$$

biçiminde bir karşılaştırma matrisi oluşturulur. Eşitlik (1) ile tanımlı karşılaştırma matrisinin her bir elemanı

$$a'_{ij} = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}, \ i = 1, 2, \ldots, m \tag{2}$$

olacak biçimde normalleştirilir**.** Normalleştirilmiş karşılaştırma matrisinin satır ortalamaları değişkenlerin önem ağırlıkları olup

$$w_j = \frac{1}{m} \sum_{j=1}^{m} a'_{ij}, \ i = 1, 2, \ldots, m \tag{3}$$

olur. Elde edilen değişken ağırlıkları ile veri seti Çizelge 2'deki biçimde ağırlıklandırılır.

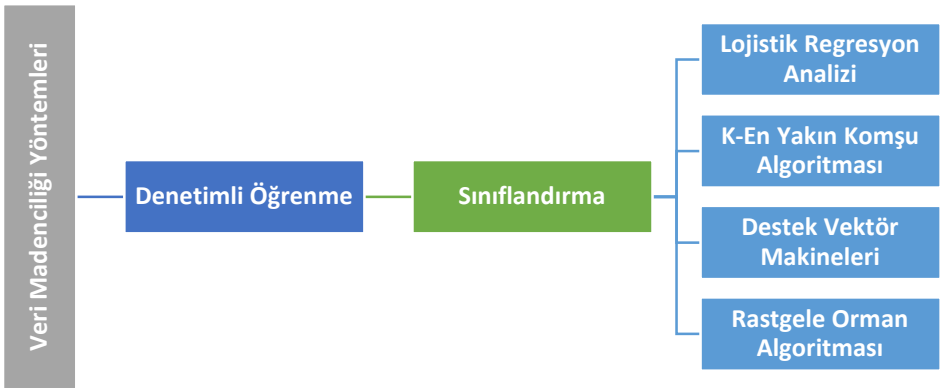
**Çizelge 2.** AHP ile ağırlıklandırılmış çok değişkenli bir veri seti

| | ***Bağımsız değişkenler*** | | | | ***Bağımlı değişken*** |
|---|---|---|---|---|---|
| **No.** | $w_1X_1$ | $w_2X_2$ | ... | $w_mX_m$ | $Y$ |
| 1 | $w_1x_{11}$ | $w_2x_{12}$ | ... | $w_mx_{1m}$ | $Y_1$ |
| 2 | $w_1x_{21}$ | $w_2x_{22}$ | ... | $w_mx_{2m}$ | $Y_2$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $n$ | $w_1x_{n1}$ | $w_2x_{n2}$ | ... | $w_mx_{nm}$ | $Y_n$ |

Ağırlıklandırılmış veri setinin elde edilmesi ile veri ön işleme süreci tamamlanır. İstatistiksel bakış açısı ile daha esnek hesaplama kolaylığı sağlayan veri madenciliği sınıflandırma yöntemleri uygulanarak veri yapısına uygun biçimde analiz yapılır.

## 3. Veri madenciliği sınıflandırma yöntemleri

Veri madenciliği sınıflandırma yöntemleri, verideki gizli örüntülerin ortaya çıkarılarak gözlemlerin hangi sınıfa ait olduğunun tahmin edilmesini sağlayan denetimli öğrenme yöntemleridir. Bağımlı değişkenin kategorik ve bağımsız değişkenlerin sürekli ve kategorik değişkenlerden oluştuğu veri setleri için Şekil 3'te verilen veri madenciliği sınıflandırma yöntemleri kullanılmaktadır.




**Şekil 3.** Veri madenciliği sınıflandırma yöntemleri

Veri madenciliği sınıflandırma yöntemlerinin uygulanabilmesi için, yöntemlere ilişkin ayarlanabilir parametrelerin (tuning parameters) ilgilenilen veri setine yönelik uygun biçimde belirlenmiş olması gerekir. Ayarlanabilir parametrelerin seçimi yöntemlerin sınıflandırma performansında etkindir. Uzman görüşü alınarak da belirlenen bu parametrelerin seçimi önemlidir.

### 3.1. Lojistik Regresyon

Lojistik regresyon (LR), veri setindeki gözlemlerin bir sınıfa ait olup olmama olasılığını hesaplayarak yeni gelecek gözlemleri sınıflandıran bir veri madenciliği yöntemidir. $Y_i \in \{0,1\}$ olmak üzere $\{(X_i, Y_i)\}_{i=1}^{n}$, etiketli bir veri seti olsun. Bağımlı değişken kategorik olduğundan dolayı bağımlı değişkenin olasılığı hesaplanıp lojit dönüşüm yapılarak lojistik regresyon modeli

$$f(x_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{X}_i \boldsymbol{\beta})}},\ i = 1, 2, \ldots, n \tag{4}$$

biçiminde elde edilir. Burada

$$\beta_0 + \boldsymbol{X}_i \boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{nm},\ i = 1, 2, \ldots, n \tag{5}$$

dir. Model parametrelerinin tahmin edilmesinde maksimum olabilirlik yöntemi kullanılarak tahmin olasılıklarının maksimize edilmesi istenir [19]. Eşitlik (4) bir olasılık değeri olup, eşik değeri $0.5$ alındığında sınıflandırma

$$\begin{aligned} f(x_i) \geq 0.5,\ &\Rightarrow \hat{Y}_i = 1 \\ f(x_i) < 0.5,\ &\Rightarrow \hat{Y}_i = 0 \end{aligned} \tag{6}$$

biçiminde yapılır. Lojistik regresyonun ayarlanabilir parametreleri, kısıtlama yapılacak yöntemi (Lasso, Ridge, None) belirleyen ceza değeri (penalty) ve kısıtlama oranı ($c$)'dır.

### 3.2. K-En Yakın Komşu Algoritması

*K*-En Yakın Komşu Algoritması (KNN), uygulaması kolay ve anlaşılır sınıflandırma yöntemlerinden biridir. KNN yönteminde, sınıfı belli olan gözlemlerden yararlanılarak yeni katılacak gözlemin hangi sınıfa ait olup olmadığı belirlenir. Bu yöntemde, gözlemlerin her biri ile yeni gelecek gözlem arasındaki uzaklıklar hesaplanarak en küçük uzaklığa sahip *k* sayıda gözlemin seçimi yapılır. Hesaplanan uzaklık değerleri arasında en çok tekrar eden sınıf, yeni gözlem değerinin sınıfıdır. Gözlemler arasındaki uzaklıklar hesaplanırken yaygın olarak Öklid uzaklık formülü kullanılır. *m* değişken sayılı veri setindeki *i*. ve *j*. gözlemler arasındaki uzaklık

$$d_{ij} = \sqrt{\sum_{K=1}^{m} (x_{ik} - x_{jk})^2},\ \ i, j = 1, 2, \ldots, m,\ i \neq j \tag{7}$$

biçiminde hesaplanır. Burada, *k* ayarlanabilir parametresinin belirlenmesi önemlidir. En küçük uzaklığın elde edilmesi amaçlanır.

### 3.3. Destek Vektör Makineleri

Destek vektör makineleri (SVM) veri madenciliği yöntemlerinden biridir [20]. Bu yöntem, veriyi doğrusal olarak iki sınıfa ayırabilmek için en uygun fonksiyonun (hiperdüzlemin) tahmin edilmesi esasına dayanır. $Y_i \in \{-1,1\}$ olmak üzere $\{(X_i, Y_i)\}_{i=1}^{n}$, etiketli bir veri seti olsun. Nokta çarpımları Eşitlik (8) biçiminde ifade edilir.

$$< \boldsymbol{\beta},\ \boldsymbol{X} > = \boldsymbol{\beta} * \boldsymbol{X} = \boldsymbol{\beta}^T \boldsymbol{X} = \sum_{j=1}^{m} \beta_j X_j \tag{8}$$

İki sınıf arasındaki karar sınırı olan hiperdüzlem denklemi

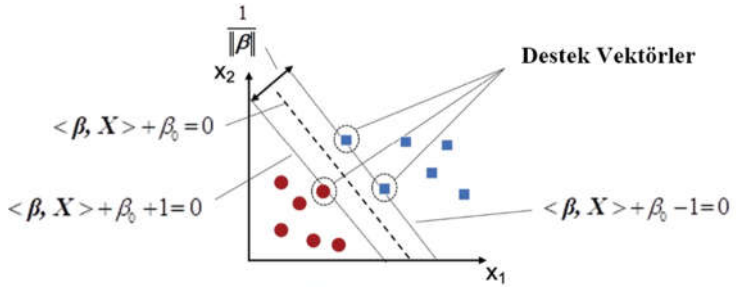
$$H\ düzlemi: <\boldsymbol{\beta},\ \boldsymbol{X}> + \beta_0 = 0 \tag{9}$$

biçiminde yazılır.

Şekil 4'te verilen $\boldsymbol{\beta}$ ağırlık vektörü, hiperdüzleme dik yönde olup eğimi belirlemektedir. Burada, $\beta_0$ sabit terimdir. $H$ düzlemi optimal hiperdüzlem olup $H_1$ düzlemi ve $H_2$ düzlemi

$$\begin{aligned} H_1\ düzlemi &: <\boldsymbol{\beta},\ \boldsymbol{X}> + \beta_0 - 1 = 0 \\ H_2\ düzlemi &: <\boldsymbol{\beta},\ \boldsymbol{X}> + \beta_0 + 1 = 0 \end{aligned} \tag{10}$$

biçiminde olur. $H_1$ düzlemi ve $H_2$ düzlemi üzerindeki gözlemler, sınırları belirleyen destek vektörleri (support vectors) olarak tanımlanmaktadır. Destek vektörlerinin seçimi yapılırken $H_1$ düzlemi ve $H_2$ düzlemleri arasındaki mesafenin en büyük olması istenmektedir. Bu mesafeye kenar payı (marjin) adı verilmekte olup kenar payı $2d = \frac{2}{\|\boldsymbol{\beta}\|}$ değerinin en büyük yapılması $\min \frac{1}{2}\|\boldsymbol{\beta}\|^2$ amaçlanmaktadır.




**Şekil 4.** İki sınıflı veri setinin SVM yöntemi ile sınıflandırılması

SVM problemi

$$\begin{aligned} &\min_{\beta}\ f(\boldsymbol{\beta}) = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle \\ &g(\boldsymbol{\beta},\ \beta_0) = Y_i\left(<\boldsymbol{\beta}, \boldsymbol{X}_i> + \beta_0\right) - 1 \geq 0,\ \ i = 1,2,..,n \end{aligned} \tag{11}$$

biçimde eşitsizlik kısıtlı optimizasyon problemi yazılır. Yeni gelen gözlemin sınıflandırılmasında

$$Y' = \text{sgn}(\boldsymbol{\beta}^T \boldsymbol{X} + \beta_0) \tag{12}$$

ifadesi kullanılır [21,22]. Doğrusal sınıflamanın mümkün olmadığı ya da değişken sayısının fazla olduğu durumlarda, SVM yöntemi çekirdek fonksiyonları kullanarak sınıflandırma yapar [23]. Çekirdek fonksiyonu

$$K\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = <\phi_{X_i}, \phi_{X_j}> = \left(\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle\right)^2 \tag{13}$$

biçiminde tanımlanır. Herhangi bir doğrusal sınıflandırma probleminin amaç fonksiyonunda $\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle$ biçiminde vektörlerin iç çarpımı yer alıyorsa, bu ifade yerine uygun bir $K\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$ çekirdek fonksiyonu yazılarak problem güncellenir. Çekirdek fonksiyonların seçimi, SVM performansını etkilemektedir. Yaygın olarak kullanılan çekirdek fonksiyonları Çizelge 3'te verilmiştir.

**Çizelge 3.** Yaygın olarak kullanılan çekirdek fonksiyonları

| |
|---|
| *Doğrusal*: $K\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\left(\left\langle \boldsymbol{X}_i,\boldsymbol{X}_j\right\rangle\right)$ |
| *Polinom*: $K\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\left(\left\langle \boldsymbol{X}_i,\boldsymbol{X}_j\right\rangle+1\right)^m$ |
| *Dairesel* (*Radyal*) *Tabanlı*: $K(\boldsymbol{X}_i,\boldsymbol{X}_j)=\exp(-\gamma\left\|\boldsymbol{X}_i-\boldsymbol{X}_j\right\|^2)$ |
| *Sigmoid*: $K(\boldsymbol{X}_i,\boldsymbol{X}_j)=\tanh(\gamma\boldsymbol{X}_i^T\boldsymbol{X}_j+r)$ |

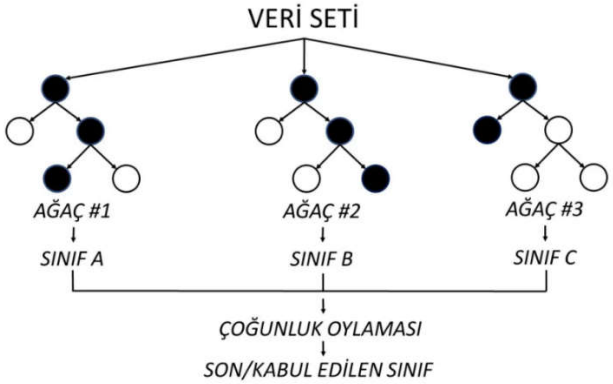
SVM için ayarlanabilir parametreler Çizelge 4'te verilmiştir.

**Çizelge 4.** SVM'de kullanılan ayarlanabilir parametreler

| |
|---|
| *c*: Kısıtlama oranı |
| *Kernel*: Çekirdek fonksiyon (*Polynomial*, *Rbf*, *Sigmoid*, *Linear*) |
| *Gamma*: *c* oranını düzenleyici değer |

*3.4. Rastgele Orman Algoritması*

Rastgele Orman Algoritması (RF), verilen veri kümesinin çeşitli alt kümelerinde birden çok sayıda karar ağacı içeren sınıflandırma yöntemidir [23]. Bu yöntem, veri kümesinin tahmin doğruluğunu iyileştirmek, varyansı ve yanlılığı azaltmak amacıyla kullanılır. RF, tek bir karar ağacına güvenmek yerine Şekil 5'teki gibi her ağaçtan tahminleri toplar ve tahminlerin çoğunluğuna dayanarak nihai sonucu tahmin eder.




**Şekil 5.** RF Algoritması [24]

Bootstrap yöntemi ile eğitim verisinin 2/3'ü (In-bag(IB)) ile örneklemlerden ağaçlar oluşturulur. Eğitim veri setinin geriye kalan 1/3'ü (Out-of-bag(OOB)), ağaçların performans değerlendirmesi için hataların kestirim hesabında kullanılır. Her bir ağaç ikili bölünme yapısı ile alt düğümlere ayrılır. Oluşturulacak ağaç sayısı araştırmacı tarafından belirlenmektedir. Her bir örneklem için her düğümde *m* değişken arasından $s=\sqrt{m}$ adet değişken belirlenir. Her eğitim setinden elde edilen tahminlerin çoğunluk oylamasına göre sınıflandırma yapılır [21]. $B$ oluşturulan ağaç sayısı olmak üzere $(b=1,2,\ldots,B)$, $b$. ağacın sınıf tahmini

$$\hat{C}_{RF}^{B}(x)=\text{çoğunluk oylaması}\left\{\hat{C}_b(x)\right\}_1^B \tag{14}$$

biçiminde hesaplanır. Performansın değerlendirilmesinde OOB hata oranı

$$E_{OOB}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(Y_i\neq\hat{C}_{RF}^{B}(x)) \tag{15}$$

dir. OOB hata oranının minimum yapılması istenir. RF için ayarlanabilir parametreler Çizelge 5'te verilmiştir.

**Çizelge 5.** RF'de kullanılan ayarlanabilir parametreler

| |
|---|
| *Maksimum derinlik* (*Max dept*): Ardışık soru sayısı |
| *Maksimum değişken* (*Max features*): Her düğümde değerlendirilen değişken sayısı |
| *Minumum yaprak örneği* (*Min samples leaf* ): Bir yapraktaki minimum gözlem sayısı |
| *Minimum örneklem* (*Min samples split*): Bir düğüm bölünmeden önce gerekli gözlem sayısı |
| *n tahmin edici* (*n estimators*): Oluşan ağaç sayısı (*B*) |
| *Kriter* (*Criterion*): İndeks hesaplama yöntemi (Gini, Entropy) |

Uygulanan sınıflandırma yöntemlerinde, algoritmanın performansının karşılaştırılabilmesi için performans ölçütleri bulunmaktadır. Sınıflandırma performans ölçütleri temel alınarak MCDM ile sınıflandırma yöntemine karar verilir. Sınıflandırma yöntemlerinde, performans ölçütleri hesaplanmadan önce, veri seti, eğitim seti ve test seti (genellikle %80 eğitim seti ile %20 test seti) olmak üzere ikiye ayrılır. Veri seti *k* sayıda parçaya bölünür. Bölünen her bir parçadan birisi test diğer *k*-1 parça eğitim verisi olarak kullanılır. Eğitilen her bir parça test edilerek performans ölçütü hesaplanır. Hesaplanan *k* tane performans ölçütünün ortalamasıyla, *k*-kat Çapraz Doğrulama (*k*-fold Cross Validation-CV) ile performans ölçütleri elde edilir. Sınıflandırma yöntemlerinin performans ölçütlerinin hesaplanabilmesi amacıyla Çizelge 6'da verilen karışıklık matrisi kullanılmaktadır.

**Çizelge 6.** Karışıklık matrisi

| | | **Tahmin Sınıfı** | |
|---|---|---|---|
| | **Sınıf** | **Pozitif** | **Negatif** |
| **Gerçek Sınıf** | **Pozitif** | Doğru Pozitif (DP) | Yanlış Negatif (YN) |
| | **Negatif** | Yanlış Pozitif (YP) | Doğru Negatif (DN) |

**Doğruluk (Accuracy):** Doğru olarak sınıflandırılmış örneklerin toplam örnek sayısına oranı

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \tag{16}$$

ile hesaplanır.

**Duyarlılık (Sensivity):** Doğru olarak sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranı olarak

$$Duyarlılık = \frac{DP}{DP + YN} \tag{17}$$

biçiminde hesaplanır.

**Kesinlik (Precision)**: Pozitif olarak tahmin edilen değerlerin ne oranda doğru olarak tahmin edildiği ölçütü

$$Kesinlik = \frac{DP}{DP + YP} \tag{18}$$

ile hesaplanmaktadır.

**$F_1$-skor ($F_1$-score ):** Kesinlik ve duyarlılık ölçütünün harmonik ortalaması olan ölçüt

$$F_1\text{-}skor = 2\frac{(Duyarlılık)\,x\,(Kesinlik)}{(Duyarlılık)+(Kesinlik)} \tag{19}$$

biçiminde hesaplanır. Sınıflandırma yöntemlerine karar verilebilmesi için sınıflandırma performans ölçütlerine göre MCDM yöntemlerinden CODAS uygulanacaktır.

CODAS yöntemi, 2016 yılında geliştirilen MCDM yöntemlerinden biridir [25]. Bu yöntemde alternatiflerin değerlendirilmesi negatif ideal çözüme olan uzaklıklara dayanır. Alternatiflerin negatif ideal çözüme olan uzaklıklarının hesaplanmasında Öklid ve Taxicab uzaklıkları kullanılır. CODAS işleyiş adımları aşağıdaki gibi tanımlanır.

**Adım 1:** CODAS yönteminde $n$ tane sınıflandırma yöntemi (seçenekler), $m$ tane performans ölçütleri (ölçütler) olacak şekilde $n \times m$ boyutlu bir karar matrisi Eşitlik (20)'de verilen biçimde belirlenir.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \tag{20}$$

**Adım 2:** Normalleştirilmiş matris Eşitlik (21) kullanılarak hesaplanır.

$$d'_{ij} = \frac{d_{ij}}{maxd_{ij}},\ i = 1,2,\ldots,n,\quad j = 1,2,\ldots,m \tag{21}$$

Eşitlik (22)'de verilen $V$ normalleştirilmiş karar matrisi elde edilir.

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix} \tag{22}$$

Performans ölçütlerini en büyük yapacak negatif ideal çözüm $V^- = \left\{v_1^-, v_2^-, \ldots, v_m^-\right\}$, $V$ matrisinin sütunlarının en küçük değerleri ile elde edilir.

**Adım 3:** Öklid uzaklığı kullanılarak her bir yönteme ilişkin Eşitlik (23)'teki gibi negatif ideal çözüm değerleri uzaklıkları hesaplanır.

$$E_i = \sqrt{\sum_{j=1}^{m}(v_{ij} - v_j^-)^2},\ i = 1,2,\ldots,n \tag{23}$$

**Adım 4:** Taxicab uzaklığı kullanılarak her bir yönteme ilişkin Eşitlik (24)'teki gibi negatif ideal çözüm değerleri uzaklıkları hesaplanır.

$$T_i = \sum_{j=1}^{m}\left|v_{ij} - v_j^-\right|,\ i = 1,2,\ldots,n \tag{24}$$

**Adım 5:** Göreli değerlendirme matrisi Eşitlik (25)'teki gibi elde edilir.

$$h_{ik} = (E_i - E_k) + \varphi(E_i - E_k)x(T_i - T_k) \tag{25}$$

$\varphi$ ile iki alternatif arasındaki Öklid uzaklık değeri için bir eşik fonksiyonu Eşitlik (26) ile elde edilir.

$$\varphi(x)=\begin{cases}0, & |x|<0.2\\ 1, & |x|\geq 0.2\end{cases} \tag{26}$$

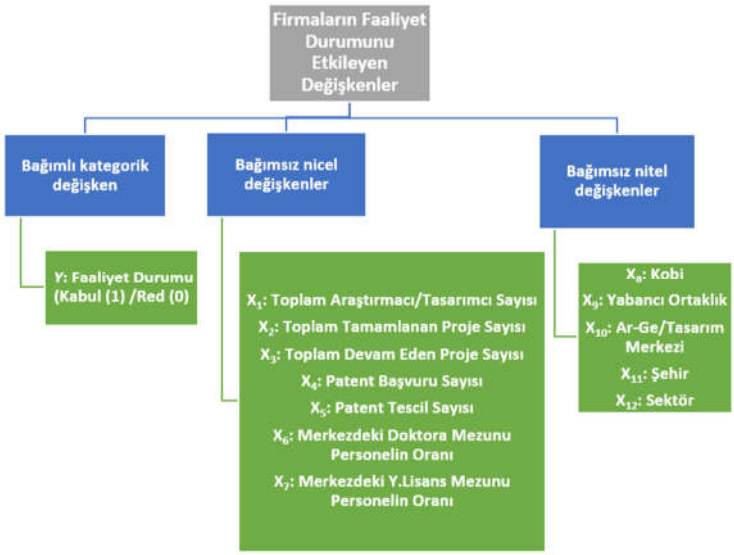
**Adım 6:** Değerlendirme puanlarının hesaplanması için Eşitlik (27) kullanılır.

$$H_i=\sum_{k=1}^{m} h_{ik} \tag{27}$$

Değerlendirme puanı büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

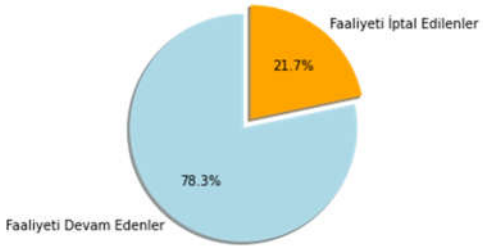
## 4. Sayısal uygulama

Bu çalışmada, Ar-Ge ve Tasarım merkezi belgesi alan firmaların çalışmalarına göre faaliyetlerinin devam ettirilmesi ya da iptal edilmesi (belge iptali) durumlarına (faaliyet sonucuna) göre sınıflandırılması istenmiştir. Buna göre, yetkililerin bilgisi dahilinde Ar-Ge ve Tasarım Merkezleri Daire Başkanlığı'ndan veri seti talep edilmiştir. Veri seti, 2008-2021 yılları arasında faaliyeti devam eden 2334 adet Ar-Ge ve Tasarım Merkezleri'ni kapsamakta olup uygun biçimde veri tabanından temin edilmiştir. Veri setinde her bir bilginin değerli olması ve bilgi kaybının istenmemesi nedeniyle veri setinde örnekleme yapılmadan verinin tamamı değerlendirilmiştir. Çalışma kapsamında, ilgilenilen veri setine, öncelikli olarak veri ön işleme uygulanmıştır. Aynı bilgiyi içeren değişkenler birleştirilip bazı değişkenlerin daha anlamlı olması adına oransal değişkenlere dönüştürülerek değişken seçimi ve boyut indirgeme yapılmıştır. Çalışmada yapılan analizler için Python 3.11.3 programı ve kütüphaneleri kullanılmıştır. İlgilenilen veri setine yönelik, bağımlı ve bağımsız değişkenler belirlenerek, değişkenlerin aldığı değerlere ilişkin kategorileri Şekil 6'da detaylı biçimde belirtilmiştir.



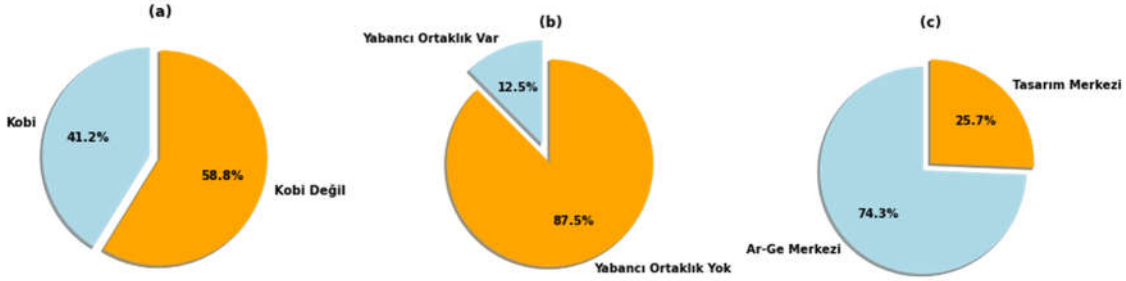

**Şekil 6.** Firmaların faaliyet durumunu etkileyen değişkenler

Şekil 6'dan, veri setinin bir kategorik bağımlı değişken, yedi nicel ve beş nitel değişken olmak üzere on iki (12) bağımsız değişkenden oluştuğu görülmektedir. Betimsel istatistikler kullanılarak veri seti hakkında özet bilgiler elde edilmiştir.



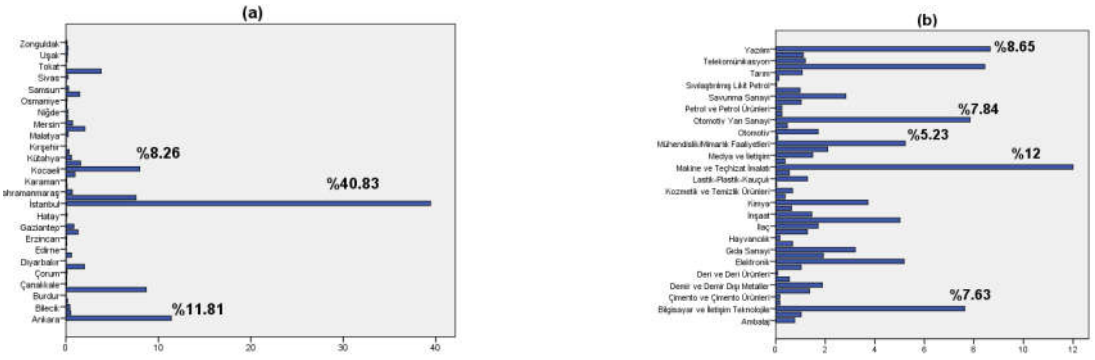

**Şekil 7.** Veri setinde kategorik bağımlı değişkene ait pasta grafiği

Şekil 7'ye göre Ar-Ge ve Tasarım merkezlerinin yaklaşık %78'inin faaliyetinin devam ettiği, yaklaşık %22'sinin ise faaliyeti iptal edilmiştir. Buna göre yanıt değişkeninin $Y$=1 olma olasılığı $p$=0.78 olarak da ifade edilebilir.




**Şekil 8.** (a) Kobi, (b) Yabancı Ortaklık ve (c) Ar-Ge/Tasarım Merkezi bilgisine ait pasta grafiği

Şekil 8'e göre, Ar-Ge ve Tasarım merkezlerinin %58.83'ü büyük ölçekli işletme sınıfında yer alırken, %41.17'si Kobi'dir. Ar-Ge ve Tasarım merkezlerinin %87.53'ünün yabancı ortaklığı bulunmazken, %12.47'sinin yabancı ortaklığı bulunmaktadır. Firmaların %73.34'ü Ar-Ge merkezi iken, %25.66'sı Tasarım merkezidir.




**Şekil 9.** Ar-Ge ve Tasarım merkezlerinin (a) Şehir ve (b) Sektör değişkenlerine göre yüzdelik değerleri

Şekil 9'a göre, Ar-Ge ve Tasarım merkezi kuran firmaların bulunduğu şehirlerin %40.83'ü İstanbul'da, %11.81'i Ankara'da, %8.26'sı Kocaeli'nde bulunmaktadır. Ar-Ge ve Tasarım merkezlerinin %12'si makine ve teçhizat imalatı sektöründe bulunurken, %8.65'i yazılım, %7.84'i ise otomotiv yan sanayi, %7.63'ü bilgisayar ve iletişim teknolojileri, %5.23 mühendislik/mimarlık sektöründe faaliyet göstermektedir.

**Çizelge 8.** Veri setindeki nicel değişkenlere ilişkin betimsel istatistikler

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| *n* | 2334 | 2334 | 2334 | 2334 | 2334 | 2334 | 2334 |
| *min* | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $Q_1$ | 11 | 5 | 4 | 0 | 0 | 0 | 0.030 |
| $Q_2$ | 15 | 12 | 6 | 0 | 0 | 0 | 0.080 |
| $Q_3$ | 24 | 26 | 11 | 2 | 1 | 0 | 0.140 |
| *max* | 2710 | 1056 | 485 | 2127 | 693 | 0.43 | 0.710 |

Nicel değişkenlere ait minimum, maksimum ve çeyreklik değerleri Çizelge 8'de verilmiştir. Ar-Ge ya da Tasarım merkezi olan firmaların yarısında 15'ten fazla Araştırmacı/Tasarımcı personelin bulunduğu, en fazla Araştırmacı/Tasarımcı personeli bulunan firmada ise bu sayının 2710 olduğu, toplam tamamlanan ve toplam devam eden proje sayısı sıfır olan firmaların bulunduğu fakat bu firmaların faaliyette olmadığı, en fazla 2127 patent başvurusu yapıldığı, en fazla 693 patente tescil alındığı, en fazla doktoralı oranının 0.43, en fazla yüksek lisanslı oranının 0.71 olduğu görülmektedir.

Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkileyen değişkenlerin, Ar-Ge ve Tasarım Merkezi Dairesi'nde çalışan uzmanların görüşleri dikkate alınarak melez bir sınıflandırma yaklaşımı uygulanması istenmiştir. Bu amaçla, nitel ve nicel değişkenleri değerlendirebilen hem objektif hem subjektif bilgileri karar sürecine dahil edebilen AHP yöntemi seçilmiştir. Ar-Ge ve Tasarım Merkezlerinin faaliyetlerini etkileyen değişkenlerin önemlerine göre ağırlıklarının hesaplanması istenmiştir. AHP yönteminde her bir değişkenin birbirine göre doğrudan öneminin belirlenebilmesi amacıyla on beş uzmanın görüşü alınmıştır. Buna göre, Eşitlik (1)'de hesaplaması gösterilen karşılaştırma matrisi

$$A = \begin{bmatrix}
1.0000 & 2.0000 & 1.0000 & 2.0000 & 1.0000 & 0.5000 & 1.0000 & 9.0000 & 3.0000 & 4.0000 & 5.0000 & 4.0000 \\
0.5000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 6.0000 & 3.0000 & 2.0000 & 4.0000 & 3.0000 \\
1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 5.0000 & 3.0000 & 2.0000 & 4.0000 & 2.0000 \\
0.5000 & 1.0000 & 1.0000 & 1.0000 & 0.5000 & 0.3333 & 0.3333 & 6.0000 & 3.0000 & 2.0000 & 4.0000 & 2.0000 \\
1.0000 & 1.0000 & 1.0000 & 2.0000 & 1.0000 & 1.0000 & 1.0000 & 7.0000 & 3.0000 & 3.0000 & 6.0000 & 3.0000 \\
2.0000 & 1.0000 & 1.0000 & 3.0000 & 1.0000 & 1.0000 & 5.0000 & 9.0000 & 5.0000 & 4.0000 & 6.0000 & 5.0000 \\
1.0000 & 1.0000 & 1.0000 & 3.0000 & 1.0000 & 0.2000 & 1.0000 & 8.0000 & 4.0000 & 3.0000 & 5.0000 & 3.0000 \\
0.1111 & 0.1667 & 0.2000 & 0.1667 & 0.1429 & 0.1111 & 0.1250 & 1.0000 & 0.3333 & 0.5000 & 1.0000 & 1.0000 \\
0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.2000 & 0.5000 & 3.0000 & 1.0000 & 1.0000 & 2.0000 & 1.0000 \\
0.2500 & 0.5000 & 0.5000 & 0.5000 & 0.3333 & 0.2500 & 0.6667 & 2.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\
0.2000 & 0.2500 & 0.2500 & 0.2500 & 0.1667 & 0.1667 & 0.2000 & 1.0000 & 0.5000 & 1.0000 & 1.0000 & 1.0000 \\
0.2500 & 0.3333 & 0.5000 & 0.5000 & 0.3333 & 0.4000 & 0.3333 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000
\end{bmatrix}$$

biçiminde oluşturulmuştur.

**Çizelge 9.** AHP ile belirlenen değişken ağırlıkları

| | **Değişkenler** | $w$ |
|---|---|---|
| **Nicel** | Araştırmacı/Tasarımcı Sayısı $(X_1)$ | 0.1310 |
| | Toplam Tamamlanan Proje Sayısı $(X_2)$ | 0.1020 |
| | Toplam Devam eden Proje Sayısı $(X_3)$ | 0.1026 |
| | Patent Başvuru Sayısı $(X_4)$ | 0.0800 |
| | Patent Tescil Sayısı $(X_5)$ | 0.1218 |
| | Merkezdeki Doktora Mezunu Personelin Oranı $(X_6)$ | 0.1835 |
| | Merkezdeki Y. Lisans Mezunu Personelin Oranı $(X_7)$ | 0.1189 |
| **Nitel** | Sektör $(X_8)$ | 0.0373 |
| | Şehir $(X_9)$ | 0.0249 |
| | Kobi $(X_{10})$ | 0.0186 |
| | Ar-Ge/Tasarım Merkezi $(X_{11})$ | 0.0404 |
| | Yabancı Ortaklık $(X_{12})$ | 0.0390 |

Eşitlik (2) ve (3) kullanılarak hesaplanan değişken ağırlıkları Çizelge 9'da verilmiştir. Buna göre Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkilemede en önemli değişkenlerin Merkezdeki Doktora Mezunu Personelin Oranı ile Araştırmacı/Tasarımcı Sayısı olduğu uzmanlar tarafından düşünülmektedir. Merkeze ait Sektör, Şehir, Kobi, Ar-Ge/Tasarım Merkezi ve Yabancı Ortaklık bilgilerinin Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkilemede en az öneme sahip olduğu görülmektedir. AHP sonucunda en az öneme sahip olduğu düşünülen merkezlere ait bilgileri içeren değişkenler (Kobi, Yabancı Ortaklık Ar-Ge/Tasarım Merkezi, Şehir, Sektör) veri madenciliği sınıflandırma yöntemlerinde dikkate alınmamıştır. AHP sonucu elde edilen değişken ağırlıkları kullanılarak veri seti Çizelge 2'deki biçimde ağırlıklandırılır. Veri ön işleme sonucunda veri seti, yedi bağımsız nicel değişken ve bir kategorik bağımlı değişken ile analize hazır hale getirilmiştir. Çalışmanın bir sınıflandırma problemi olması sebebiyle veri madenciliği sınıflandırma yöntemlerinden LR, KNN, SVM ve RF kullanılmıştır. Sınıflandırma yöntemlerinin performansının ölçülebilmesi amacıyla veri seti, %80 eğitim seti ile %20 test seti olmak üzere ayrılmıştır. Python 3.11.3 sürümünde Scikit-learn kütüphanesinde sınıflandırma yöntemlerinin performansı 5-kat CV kullanılarak test edilmiş, ızgara arama yöntemi ile optimal parametreler elde edilmiştir.

**Çizelge 10.** Izgara arama ile belirlenen ayarlanabilir parametre değerleri

| ***Yöntemler*** | ***Ayarlanabilir Parametreleri*** |
|---|---|
| **LR** | *c* = 1e-05, *Penalty* = None |
| **KNN** | *Metric* = Euclidean, *Neighbors* (k) = 5 |
| **SVM** | *c* = 100, *Gamma* = 10, *Kernel* = Rbf' |
| **RF** | *Bootstrap* = True, *Criterion* = Gini, *Max Depth* = 3, *Max Features* = 3, *Min Samples Leaf* = 5, *Min Samples Split* = 3, *n Estimator* = 250 |

Her bir sınıflandırma yöntemi için kullanılan ayarlanabilir parametre değerleri Çizelge 10'da verilmiştir. Sınıflandırma yöntemlerini değerlendirmek için performans ölçütlerinden doğruluk, kesinlik, duyarlılık ve $F_1$-skor test veri seti için hesaplanmıştır. Test veri setine ait performans ölçütleri Çizelge 11'de verilmiştir.

**Çizelge 11.** Test veri seti sonuçları

| | **Performans Ölçütleri** | | | |
|---|---|---|---|---|
| ***Yöntemler*** | ***Doğruluk*** | ***Kesinlik*** | ***Duyarlılık*** | ***$F_1$-skor*** |
| **LR** | 0.85 | 0.89 | 0.92 | 0.90 |
| **KNN** | 0.83 | 0.87 | 0.91 | 0.89 |
| **SVM** | 0.85 | 0.90 | 0.90 | 0.90 |
| **RF** | 0.86 | 0.89 | 0.93 | 0.91 |

Çizelge 11'de görülen değerlere göre performans ölçütleri bakımından sınıflandırma yöntemlerini karşılaştırmak zordur. Objektif karşılaştırma yapılabilmesi için CODAS çok ölçütlü karar verme yöntemi uygulanmıştır. Çizelge 11'de verilen sonuçlar kullanılarak Eşitlik (20)'ye göre karar matrisi

$$D = \begin{bmatrix} 0.85 & 0.89 & 0.92 & 0.90 \\ 0.83 & 0.87 & 0.91 & 0.89 \\ 0.85 & 0.90 & 0.90 & 0.90 \\ 0.86 & 0.89 & 0.93 & 0.91 \end{bmatrix} \tag{28}$$

biçiminde oluşturulur. Eşitlik (21-27) kullanılarak CODAS yöntemine göre, değerlendirme puanı büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

**Çizelge 12.** Test veri seti için CODAS'a göre sıralama

| *Yöntemler* | *Değerlendirme Puanları* | *Sıralama* |
|---|---|---|
| **LR** | 0.0129 | 3 |
| **KNN** | -0.1069 | 4 |
| **SVM** | 0.0202 | 2 |
| **RF** | 0.0821 | 1 |

Çizelge 12'ye göre, RF sınıflandırma yöntemi öncelikli tercih edilir. Buna göre, sınıflandırma yöntemlerinin öncelikli tercih sıralamasının RF >> SVM >> LR >> KNN biçiminde olduğu söylenir.

## 5. Sonuç

Bu çalışmada, sınıflandırma problemlerinin çözümünde, karar verme sürecine katkı sağlayacağı düşünülen bir melez sınıflandırma yaklaşımı önerilmiştir. Oluşturulan melez sınıflandırma yaklaşımında, veriden bilgi elde etmeye yönelik subjektif değerlendirmenin AHP yöntemi ile dikkate alınmasının yanı sıra veri madenciliği yöntemleri ile sınıflandırma yaparken MCDM yöntemi ile objektif olarak sınıflandırma yöntemlerinin performansına karar verilmiştir. Titizlikle veri ön işleme aşamasından geçirilen veri setindeki değişkenler için Ar-Ge ve Tasarım Merkezi Dairesi'nde çalışan uzmanların görüşleri alınarak AHP yöntemi değişkenlerin ağırlıkları belirlenmiştir. Ağırlıklandırılmış veri setine, LR, KNN, SVM ve RF sınıflandırma yöntemleri uygulanarak yöntemlerin performans ölçütleri hesaplanmıştır. Elde edilen sınıflandırma performans değerleri bakımından karar verilmesi çok ölçütlü bir karar verme problemi olduğundan, objektif karar verebilmek için performans ölçütü hesaplama sonuçları bir karar matrisi olarak ele alınıp CODAS uygulanmıştır. CODAS sonucuna göre sınıflandırma yöntemlerinden RF'nin öncelikli tercih edilebileceği kararı elde edilerek, Ar-Ge ve Tasarım merkezlerinin faaliyetlerinin değerlendirilmesinde RF yönteminin SVM, KNN ve LR yöntemlerine göre daha iyi sınıflandırma performansı gösterdiği sonucuna ulaşılmıştır.

## Kaynaklar

[1] V. Çetin ve O. A. Yıldız, 2022, A Comprehensive review on data preprocessing techniques in data analysis, *Pamukkale University Journal of Engineering Sciences*, 28(2), 299-312.

[2] M. Emeç ve M. H. Özcanhan, 2023, Veri Ön İşleme ve Öznitelik Mühendisliğinin Yapay Zekâ Yöntemlerine Uygulanması, *Mühendislikte Öncü ve Çağdaş Çalışmalar*, 33-54.

[3] A. Burkov, "*The Hundred-Page Machine Learning Book*" kitabından çeviri, Çeviren: A. Okatan, T. Karatekin ve K. Okatan, 2021 ,100 Sayfada Makine Öğrenmesi Kitabı, (1), *Papatya Yayıncılık Eğitim*, İstanbul.

[4] J. Han, M. Kamber and J. Pei, 2012, Data mining concepts and techniques, University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

[5] A. Çınar ve G. Silahtaroğlu, 2012, Veri madenciliği teknikleri ile müşteri memnuniyetine etki eden gizli nedenlerin keşfi, *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 33(2), 309-330.

[6] Y. Nieto, V. Gacía-Díaz, C. Montenegro, C. C. González and R. G. Crespo, 2019, Usage of machine learning for strategic decision making at higher educational institutions, *IEEE Access*, 7, 75007-75017.

[7] Ç. Öztürk Zan, 2021, Prediction of Soil Radon Gas Using Meteorological Parameters with Machine Learning Algorithms, *M.Sc Thesis*, Dokuz Eylül University Graduate School of Natural and Applied Sciences.

[8] Ö. Ç. Yavuz, E. Karaman ve C. Yeşilyaprak, 2022, Makine öğrenmesi algoritmalarıyla astronomik gözlem kalitesi tahminine yönelik karar destek sistemi geliştirilmesi ve uygulanması, *Trends in Business and Economics*, 36 (3), 289-303.

[9] A. Ulutaş, 2019, Third-Party Logistics Provider Selection By Using AHP and CODAS Methods, *SETSCI Conference Proceedings*, 4 (8), 36-38.

[10] G. F. Can, P. Toktaş ve F. Pakdil, 2021, Six Sigma Project Prioritization and Selection Using AHP–CODAS Integration: A Case Study in Healthcare Industry, *IEEE Transactions on Engineering Management*, 70 (10), 3587-3600.

[11] U. Fayyad, 1997, Knowledge discovery in databases: An overview, *In International Conference on Inductive Logic Programming*, 1-16, Berlin, Heidelberg: Springer Berlin Heidelberg.

[12] Ş. Kavurkacı, Z. K. Aydın ve R. Şamlı, 2011, Büyük ölçekli veri tabanlarında bilgi keşfi, *Akademik Bilişim Konferansları*, 2-4.

[13] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, 1996, From data mining to knowledge discovery in databases, *AI magazine*, 17(3), 37-37.

[14] K. Keleş ve P. Z. Tunca, 2015, Hiyerarşik Electre Yönteminin Teknokent Seçiminde Kullanımı Üzerine Bir Çalışma, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 20 (1), 199-223.

[15] S. Arslan ve Ö. Belgin, 2020, Yüksek ve Orta-Yüksek Teknoloji Alanındaki Sektörlerin Çok Kriterli Karar Verme Teknikleri ile Önceliklendirilmesi, *Verimlilik Dergisi*, (4), 7-23. DOI: 10.51551/verimlilik.556526.

[16] M. Güryeli, 2016, Ar-Ge Projeleri Seçim Probleminin AHP Yöntemi ile İncelenmesi: Kamu Destekli Teknolojik Ürün Yatırım Destek Programı Üzerine Bir Uygulama", *Yüksek Lisans Tezi*, Adnan Menderes Üniversitesi, Sosyal Bilimler Enstitüsü.

[17] T. L. Saaty, 2008, Decision making with The Analytic Hierarchy Process, *International Journal Services Sciences*, 1(1), 83-98.

[18] T. L. Saaty, 1990, The Analytıc Hıerarchy Process In Conflıct Management, *International Journal of Conflict Management*, 1(1), 47-68. https://doi.org/10.1108/eb022672

[19] M. Ö. Dolgun, T. G. Özdemir ve D. Oğuz, 2009, Veri madenciliğinde yapısal olmayan verinin analizi: Metin ve web madenciliği. *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, *2*(2), 48-58.

[20] C. Cortes and V. Vapnik, 1995, Support-vector networks, *Machine learning*, 20, 273–297. https://doi.org/10.1007/BF00994018

[21] N. Bayram Arlı, M. Engin ve S. Gürsakal, 2022, Random Forest. Supervised Machine Learning Algorithms R and Python Applications, *Nobel Yayınevi*, Ankara.

[22] A. Géron, 2019, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, *O'Reilly Media,* Sebastopol, CA.

[23] L. Breiman, 2001, Random Forest, *Machine learning*, 45, 5-32.https://doi.org/10.1023/A:1010933404324

[24] M. Öztürk, Python ile Sınıflandırma Analizleri – Rastgele Orman (Random Forest) Algoritması- Miraç ÖZTÜRK (miracozturk.com), Erişim tarihi:04.10.2023.

[25] M. K. Ghorabaee, E. K. Zavadskas, Z. Turskis and J. Antucheviciene, 2016, A new combinative distance-based assessment (CODAS) method for multi-criteria decision-making. *Economic Computation and Economic Cybernetics Studies and Research*, 50, 25–44.

**İstatistikçiler Dergisi: İstatistik & Aktüerya**

Journal of Statisticians: Statistics and Actuarial Sciences

**IDIA 16**, 2023, 2, 116-146

**Geliş** / Received:13.11.2023, **Kabul** / Accepted: 31.12.2023

**Araştırma Makalesi /** Research Article

# The novel kumaraswamy extended garima distribution, statistical properties and its application

Ayşe Metin Karakaş

*Bitlis Eren University*
*Department of Statistics*
*Bitlis, Turkey*
*akarakas@beu.edu.tr*
ORCID:0000-0003-3552-0105

Murat Karakaş

*Bitlis Eren University*
*Department of Mathematics*
*Bitlis, Turkey*
*mkarakas@beu.edu.tr*
ORCID:0000-0002-5174-0282

Mine Doğan

*Firat University*
*Department of Statistics*
*Elazığ, Turkey*
*mine.dogan@firat.edu.tr*
ORCID:0000-0002-2745-9909

## ABSTRACT

This essay aims to present a novel Kumaraswamy Extended Garima distribution family. We obtain a cumulative distribution function, the failure rate, the risk rate, the inverse risk function, the odd function, the cumulative risk function, the moment r-th, the characteristic function of the moment generating function, the moments, the mean and the variance, Lorenz and Bonferroni curves, order statistics, MLE, mean time between failures (MTBF), Renyi and Tsallis entropies. The MLE technique estimates the parameters of the new Kumaraswamy Extended Garima distribution. The parameters of the MLE technique are derived using a nonlinear system of equations and the Symmetric Information Matrix. Furthermore, the consequences are analogous to others because they are probability distributions. According to the findings, the proposed distribution fits these data sets better than existing probability distributions.

**Key words:** Garima distribution, Kumaraswamy distribution, Cumulative distribution function, Characteristic function, Symmetric information matrix.

**ÖZ**

***Yeni Kumaraswamy genişletilmiş Garima dağılımı istatistiksel özellikleri ve uygulaması***

*Bu makale yeni bir Kumaraswamy Genişletilmiş Garima dağılım ailesini sunmayı amaçlamaktadır. Kümülatif bir dağılım fonksiyonu, başarısızlık oranı, risk oranı, ters risk fonksiyonu, tek fonksiyon, kümülatif risk fonksiyonu, moment r-th, moment üreten fonksiyonun karakteristik fonksiyonu, momentler, ortalama ve varyans, Lorenz ve Bonferroni eğrileri, sıra istatistikleri, MLE, arızalar arasındaki ortalama süre (MTBF), Renyi ve Tsallis entropileri gibi özellikleri elde ettik. MLE tekniği, yeni Kumaraswamy Extended Garima dağılımının parametrelerini tahmin eder. MLE tekniğinin parametreleri, doğrusal olmayan bir denklem sistemi ve Simetrik Bilgi Matrisi kullanılarak türetilir. Ayrıca sonuçlar olasılık dağılımları olduğundan diğerlerine benzer. Bulgulara göre, bu veri setlerine önerilen dağılım, mevcut olasılık dağılımlarından daha iyi uymaktadır.*

***Anahtar Kelimeler:*** *Garima dağılımı, Kumaraswamy dağılımı, Kümülatif dağılım fonksiyonu, Karakteristik fonksiyon, Simetrik bilgi matrisi.*

## 1. Introduction

Kumaraswamy [15] developed the Kumaraswamy probability distribution primarily for hydrological applications. Garg [11], Jones [14], Mitnik [18], and Nadarajah [21] investigated the Kumaraswamy distribution in theoretical terms. Mitnik [19] demonstrated that Kumaraswamy variables exhibit closeness under exponentiation and linear transformation, and he also introduced some of the distribution's limiting distributions and an analytical expression for the mean absolute deviation around the median as a distribution parameter function. Again, the author provided some boundaries for this dispersion measure and the variance in this investigation. Tahir et al. [27] presented a new Kumaraswamy generalized (G) distribution family via a novel generator that could be a replacement for the Kumaraswamy-G family. Cordeiro and de Castro [7] defined a new family of generalized distributions to extend the normal, Weibull, gamma, Gumbel, and inverse Gaussian distributions, among others, and discussed some special distributions in the new family, such as the Kw-normal, Kw-Weibull, Kw-gamma, Kw-Gumbel, and Kw-inverse Gaussian distributions. Carrasco et al. [5] demonstrated the log-Kumaraswamy MW regression model for censored data analysis. Asiribo et al. [2] defined the Lomax-Kumaraswamy distribution with four parameters and offered various statistical features. Wang et al. [11] investigated the estimation of points and intervals for the Kumaraswamy distribution. As a specific model, Gomes et al. [12] introduced the Kumaraswamy - Kumaraswamy (KW-KW) distribution. Kumaraswamy class generalized (KW-G) distributions. El-Sherpieny and Ahmed [13] introduced the Kumaraswamy GR (KwGR) distribution for analyzing lifespan data, as well as a linear log KwGR regression model for analyzing data with real support in order to expand some of the current regression models. Tahir et al. [26] developed a new extension of the Kumaraswamy distribution by using the Weibull link function to add a shape and a scale parameter to the Kumaraswamy distribution. Yang [29] proposed a generalized inverse Weibull distribution that incorporates both the proportional inverse hazard and the Kumaraswamy generalized

inverse Weibull distributions. Carrasco et al. [6] used a definite probability integral transform to propose and test a new five-parameter continuous distribution over a unit interval. Dey et al. [8] explored from a different perspective several estimating methods of uncertain parameters of the two-parameter Kumaraswamy distribution and they handled the estimate of the Kumaraswamy distribution's unknown parameters using simple random sampling (SRS). They ordered cluster sampling (RSS), as well as maximum probability estimation and Bayesian estimation approaches. They created a new distribution known as the generalized inverted Kumaraswamy (GIKum). The authors' main purpose in Iqbal et al. [13] is to improve a common structure for the inverse Kumaraswamy (IKum) distribution that is more flexible than the IKum distribution and all its related and submodels. Salman [25] analyzes various strategies for calculating scale and form parameters. Mohiuddin et al. [20] created the Transmuted Garima Distribution and investigated its features and applications. We propose a new model of Kumaraswamy Extended Garima distribution in our work. Several aspects of this novel model are explored and computed, including reliability functions, apparent assertions of the moments, mean deviations, Lorenz and Bonferroni curves, and Renyi and Shannon entropies. The maximum likelihood method was used to estimate the model parameters. We use this method to obtain parameter estimates and then conduct a simulation exercise to determine the performance of the maximum likelihood estimators. Furthermore, we use four real-world data sets to demonstrate the use and significance of the new family of distributions. Finally, we show that our new distribution model outperforms the well-known distributions.

## 2. Materials and Methods

### Definition 2.1. The Kumaraswamy Distribution

Kumaraswamy presented a two-parameter distribution on (0,1). Let 'Kw' represent the abbreviated name of this distribution. Its cdf is as follows:

$$G(x;\alpha,\beta) = 1-(1-x^{\alpha})^{\beta}, x \in (0,1) \tag{1}$$

as well as the probability density function

$$g(x;\alpha,\beta) = \alpha\beta x^{\alpha-1}(1-x^{\alpha})^{\beta-1}, x \in (0,1) \tag{2}$$

where $\alpha > 0$ and $\beta > 0$ denote shape parameters. Kumaraswamy [15] proposed the Kumaraswamy-G (Kw-G) distribution with the following pdf "f(x)" and cdf "F(x)" for any baseline cumulative function G(x).

$$f(x) = \alpha\beta g(x)G^{\alpha-1}(x)\left((1-G(x)^{\alpha})\right)^{\beta-1} \tag{3}$$

and

$$F(x) = 1 - (1 - G(x)^{\alpha})^{\beta} \tag{4}$$

g(x) = dG(x)/dx, etc. The parameters of the Kw-G and G distributions are identical. If X is a random variable with a density function (3), it may be represented as X Kw G(a, b) [11].

**Definition 2.2. Garima Distribution**

Assume X is a random variable with the Garima distribution and parameter. As a result, probability density is represented by;

$$g(x;\theta) = \frac{\theta}{\theta + 2}(1 + \theta + \theta x)e^{-\theta x}, x > 0, \theta > 0. \tag{5}$$

The pertinent cumulative distribution function (c.d.f.) is as follows:

$$G(x;\theta) = 1 - \left[1 + \frac{\theta x}{\theta + 2}\right]e^{-\theta x}, x > 0, \theta > 0. \tag{6}$$

**3. The Novel Kumaraswamy Extended Garima Distribution**

**Definition 3.1.** Let X $G(x,\theta)$ represent the cdf of the Garima distribution provided by (6). The novel three-parameter cdf Kumaraswamy Substituting (6) into equation (4) yields the Extended Garima (KwEG) distribution.

$$F(x) = 1 - \left(1 - \left\{1 - \left[1 + \frac{\theta x}{\theta + 2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}. \tag{7}$$

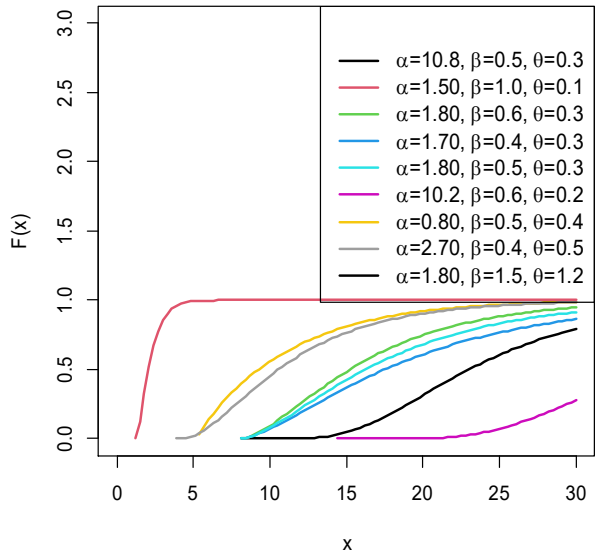



**Figure 1.** The cdf's of Kw Extended Garima distributions

The probability density function corresponding to f(x) is given by

$$f(x)=\alpha\beta\frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}. \quad (8)$$

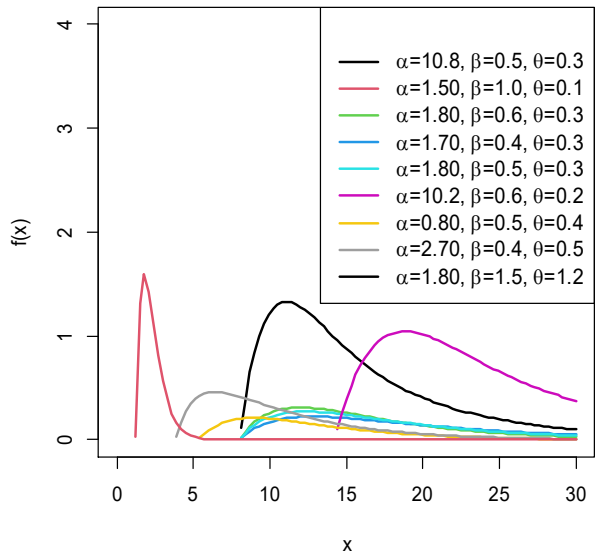


Figure 2. The pdf's of Kw- Extended Garima distributions

**Lemma 3.1.** When a=1 and b=1, the Kw- Extended Garima distribution in Equation (8) is reduced to the Garima distribution in Equation (5) with parameter.

**Lemma 3.2.** When b=1, the Kw- Extended Garima distribution is reduced to the generalized exponentiated Garima distribution, with parameters shape a and scale $\theta$ .

**Theorem.3.1.** If X follows the probability density function (8) for all $\alpha, \beta, \theta > 0$ , then X's quantile function is

$$Q_u = -\frac{2}{\theta} - 1 - \frac{1}{\theta} W_{-1}\left( -(\theta+2)e^{-(\theta+2)}\left(1-\left(1-(1-u)^{1/\beta}\right)\right)^{1/\alpha} \right).$$

$W_{-1}$ denotes the negative branch of the Lambert W function in this case.

**Proof.** If $F_x$ is a continuous and strictly increasing function, then the quantile function $Q_u$ of X is defined as follows:

$$Q(u) = F^{-1}(u), u \in (0,1) \tag{9}$$

We have an equation to solve here, derived from (7) and (9).

$$u = 1 - \left( 1 - \left\{ 1 - \left[ 1 + \frac{\theta Q(u)}{\theta + 2} \right] e^{-\theta Q(u)} \right\}^{\alpha} \right)^{\beta}.$$

We can deduce from this equation,

$$\left(\theta + 2 + \theta Q(u)\right) e^{-\theta Q(u)} = \left(\theta + 2\right)\left(1 - \left(1 - (1-u)^{1/\beta}\right)^{1/\alpha}\right) \tag{10}$$

and

$$-\left(\theta + 2 + \theta Q(u)\right) e^{-\left(\theta Q(u) + 2 + \theta\right)} = e^{-(\theta+2)} \left(\theta + 2\right)\left(1 - \left(1 - (1-u)^{1/\beta}\right)^{1/\alpha}\right) \tag{11}$$

The equation's solution is then

$$\left(\theta + 2 + \theta Q(u)\right) = W_{-1}\left( e^{-(\theta+2)} \left(\theta + 2\right)\left(1 - \left(1 - (1-u)^{1/\beta}\right)^{1/\alpha}\right)\right) \tag{12}$$

From (12), we obtained

$$Q_u = -\frac{2}{\theta} - 1 - \frac{1}{\theta} W_{-1}\left( -(\theta+2) e^{-(\theta+2)} \left( 1 - \left( 1 - (1-u)^{1/\beta} \right) \right)^{1/\alpha} \right). \tag{13}$$

**Collary 3.1.** Setting u to 0.5 in equality (13), we get the median (M) of X as;

$$M = -\frac{2}{\theta} - 1 - \frac{1}{\theta} W_{-1}\left( -(\theta+2) e^{-(\theta+2)} \left( 1 - \left( 1 - (0.5)^{1/\beta} \right) \right)^{1/\alpha} \right).$$

However, by setting u to 0.25 and 0.75 inequality (9), the 25th and 75th percentiles for the random variable X are obtained as;

$$Q_1 = -\frac{2}{\theta} - 1 - \frac{1}{\theta} W_{-1}\left( -(\theta+2) e^{-(\theta+2)} \left( 1 - \left( 1 - (0.75)^{1/\beta} \right) \right)^{1/\alpha} \right).$$

$$Q_3 = -\frac{2}{\theta} - 1 - \frac{1}{\theta} W_{-1}\left( -(\theta+2) e^{-(\theta+2)} \left( 1 - \left( 1 - (0.25)^{1/\beta} \right) \right)^{1/\alpha} \right).$$

The quantile function yields the Bowley's skewness as

$$S_k = \frac{Q_{0.75} - 2Q_{0.50} + Q_{0.25}}{Q_{0.75} - Q_{0.25}}.$$

The kurtosis of the Moor is written as

$$M_k = \frac{Q_{0.875} - Q_{0.025} - Q_{0..375} + Q_{0.125}}{Q_{0.75} - Q_{0.25}}.$$

We can easily generate X by treating u as a uniform random variable in the range (0,1).

**Theorem.3.2.** The r-th moment $E(X^r)$ of the Kw- Extended Garima distributed random variable X is given by Theorem.3.2.

$$E\left(X^r\right) = \alpha\beta \frac{\theta}{\theta+2} \left( \frac{\theta}{\theta+2} \right)^{aj+i} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{t=0}^{\infty} (-1)^{\alpha-1-i} (-1)^{\beta-1-j} (-1)^{\alpha j-k} \binom{\alpha-1}{i} \binom{\beta-1}{j} \binom{\alpha j}{k} \binom{\alpha j+i}{t}$$
$$\theta\left(\theta\left(1+i+\alpha j\right)\right)^{-2-i-\alpha j-r} \left( \left(2+\theta\right)\left(1+i+\alpha j\right)+r \right) \Gamma\left[1+i+\alpha j+r\right].$$

**Proof.** $\mu_r^{'} = E(X^r) = \int_0^\infty x^r f(x)dx$

$$= \int_0^\infty x^r \left\{ \alpha\beta \frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}\right\} dx$$

and

$$= \alpha\beta \frac{\theta}{\theta+2}\int_0^\infty x^r \left\{ (1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}\right\} dx. \quad (14)$$

Using binomial expansions of

$$\left(1-e^{-\theta x}\left(1+\frac{\theta x}{\theta+2}\right)\right)^{\alpha-1} = \sum_{i=0}^{\infty}\binom{\alpha-1}{i}\left(e^{-\theta x}\left(1+\frac{\theta x}{\theta+2}\right)\right)^{i}(-1)^{\alpha-1-i}$$

$$\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1} = \sum_{j=0}^{\infty}\binom{\beta-1}{j}\left(\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{j}(-1)^{\alpha-1-j},$$

Then, equation (14) becomes,

$$= \alpha\beta \frac{\theta}{\theta+2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}(-1)^{a-1-i}(-1)^{\beta-1-j}\binom{\alpha-1}{i}\binom{\beta-1}{j}\int_0^\infty x^r \left\{(1+\theta+\theta x)e^{-\theta x}\left(\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{i}\left(\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{j}\right\} dx. \quad (15)$$

if the following equation is written in equation (15),

$$\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha j} = \sum_{k=0}^{\infty}\binom{\alpha j}{k}\left(\left(1+\frac{\theta x}{\theta+2}\right)e^{-\theta x}\right)^{aj}(-1)^{aj-k},$$

Then, there's

$$\alpha\beta \frac{\theta}{\theta+2}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\int_0^\infty x^r(1+\theta+\theta x)e^{-\theta x(i+\alpha j+1)}\left(1+\frac{\theta x}{\theta+2}\right)^{i+\alpha j} dx.$$

As a result, we have

$$\alpha\beta\frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{\alpha j+i}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{b-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{b-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t}\int_0^{\infty}x^{r}(1+\theta+\theta x)e^{-\theta x(i+\alpha j+1)}x^{aj+i}dx. \tag{16}$$

We can deduce from equation (16), that

$$\alpha\beta\frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{aj+i}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t}\theta\left(\theta\left(1+i+\alpha j\right)\right)^{-2-i-\alpha j-r}$$
$$\left(\left(2+\theta\right)\left(1+i+\alpha j\right)+r\right)\Gamma\left[1+i+\alpha j+r\right]. \tag{17}$$

All of the moments exist because the series in (17) is convergent.

**Theorem.3.3.** . Assume X has the Kw- Extended Garima distribution. Then, say $M_X(t)$, the moment generating function of X.

$$\phi(e^{mt})=\alpha\beta\frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{\alpha j+i}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}\sum_{s=0}^{\infty}\frac{t^s}{s!}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t}$$
$$\theta\left(\theta\left(1+i+\alpha j\right)\right)^{-2-i-\alpha j-ms}\left(\left(2+\theta\right)\left(1+i+\alpha j\right)+ms\right)\Gamma\left[1+i+\alpha j+ms\right].$$

**Proof.**

$$M_X(t)=E(e^{tx})=\int_0^{\infty}e^{tx}f(x)dx. \tag{18}$$

When we apply Taylor's series to equation (18), we get

$$M_X(t)=\int_0^{\infty}\left(1+tx+\frac{(tx)^2}{2!}+\ldots\right)f(x)dx. \tag{19}$$

From equality (19), we can write

$$M_X(t)=\int_0^{\infty}\sum_{s=0}^{\infty}\frac{t^s}{s!}x^s f(x)dx. \tag{20}$$

From equality (20), we have

$$M_X(t) = \sum_{s=0}^{\infty} \frac{t^s}{s!} \mu_s'$$

$$= \alpha\beta \frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{\alpha j+i} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}\sum_{s=0}^{\infty} \frac{t^s}{s!}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t}$$
$$\theta\left(\theta\left(1+i+\alpha j\right)\right)^{-2-i-\alpha j-s}\left(\left(2+\theta\right)\left(1+i+\alpha j\right)+s\right)\Gamma\left[1+i+\alpha j+s\right].$$

The characteristic function of X is then calculated as follows:

$$\phi(e^{mt}) = \alpha\beta \frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{\alpha j+i} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}\sum_{s=0}^{\infty} \frac{t^s}{s!}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t}$$
$$\theta\left(\theta\left(1+i+\alpha j\right)\right)^{-2-i-\alpha j-ms}\left(\left(2+\theta\right)\left(1+i+\alpha j\right)+ms\right)\Gamma\left[1+i+\alpha j+ms\right].$$

(21)

**Theorem.3.4.** Assume X has the Kw- Extended Garima distribution. Renyi Entropy of X is thus given by

$$= \frac{1}{1-\gamma}\log\left(\begin{array}{l} \alpha\beta \frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{\alpha j+i} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t} \\ \frac{1}{\Gamma\left[-\gamma\right]}\left(1+\theta\right)^{\gamma}\left(\frac{\theta}{1+\theta}\right)^{-1-i-\alpha j}\Gamma\left[1+i+\alpha j\right]\Gamma\left[-1-i-\alpha j-\gamma\right] \\ Hypergeometric1F1\left[1+i+\alpha j, 2+i+\alpha j+\gamma, \alpha\left(1+\theta\right)\left(j+\gamma\right)\right] \\ +\left(\frac{\theta}{1+\theta}\right)^{\gamma}\left(\alpha\theta\left(j+\gamma\right)\right)^{-1-i-\alpha j-\gamma}\Gamma\left[-\gamma\right]\Gamma\left[1+i+\alpha j+\gamma\right] \\ Hypergeometric1F1\left[-\gamma, -i-\alpha j-\gamma, \alpha\left(1+\theta\right)\left(j+\gamma\right)\right] \end{array}\right).$$

The hypergeometric series, which covers many other special functions as specific or limiting cases, is used to show the hypergeometric1F1 function here.

**Proof**.

$$e(\gamma) = \frac{1}{1-\gamma}\log\left(\int_0^{\infty} f^{\gamma}(x)dx\right)$$

where $\gamma > 0$ and $\gamma \neq 1$. As a result, we have

$$e(\gamma)=\frac{1}{1-\gamma}\log\left(\int_0^{\infty}\left\{\alpha\beta\frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{a-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{a}\right)^{\beta-1}\right\}^{\gamma}dx\right)$$

$$=\frac{1}{1-\gamma}\log\left(\int_0^{\infty}\left\{(\alpha\beta)^{\gamma}\left(\frac{\theta}{\theta+2}\right)^{\gamma}(1+\theta+\theta x)^{\gamma}e^{-\gamma cx}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{(\alpha-1)\gamma}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{(\beta-1)\gamma}\right\}dx\right).$$

(22)

If the expressions below are written in equality (22),

$$\left(1-e^{-\theta x}\left(1+\frac{\theta x}{\theta+2}\right)\right)^{\gamma(\alpha-1)}=\sum_{i=0}^{\infty}\binom{\gamma(\alpha-1)}{i}\left(e^{-\theta x}\left(1+\frac{\theta x}{\theta+2}\right)\right)^{i}(-1)^{\gamma(\alpha-1)-i}$$

,

$$\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\gamma(b-1)}=\sum_{j=0}^{\infty}\binom{\gamma(\beta-1)}{j}\left(\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{j}(-1)^{\gamma(\beta-1)-j}$$

$$\left(\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{j}=\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha j}\sum_{k=0}^{\infty}\binom{\alpha j}{k}\left(\left(1+\frac{\theta x}{\theta+2}\right)e^{-\theta x}\right)^{\alpha j}(-1)^{\alpha j-k}$$

We then write, correspondingly,

$$=\frac{1}{1-\gamma}\log\left(\begin{array}{l}(\alpha\beta)^{\gamma}\left(\frac{\theta}{\theta+2}\right)^{\gamma}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=o}^{\infty}(-1)^{\gamma(\alpha-1)-i}(-1)^{\gamma(\beta-1)-j}(-1)^{\alpha j-k}\binom{\gamma(\alpha-1)}{i}\binom{\gamma(\beta-1)}{j}\binom{\alpha j}{k}\\ \int_0^{\infty}\left\{(1+\theta+\theta x)^{\gamma}e^{-\gamma\theta x}\left(\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{(a-1)\gamma}\left(\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{aj}\right\}dx\end{array}\right)$$

$$=\frac{1}{1-\gamma}\log\left(\begin{array}{l}\alpha\beta\frac{\theta}{\theta+2}\left(\frac{\theta}{\theta+2}\right)^{\alpha j+i}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t}\\ \int_0^{\infty}(1+\theta+\theta x)^{\gamma}e^{-\gamma\theta x-\theta x(\alpha-1)\gamma-\theta x\alpha j}x^{\alpha j+i}dx\end{array}\right).$$

When we integrate the above equation, we get the following equation. This concludes the evidence.

$$= \frac{1}{1-\gamma}\log\left(\begin{array}{l} \alpha\beta\dfrac{\theta}{\theta+2}\left(\dfrac{\theta}{\theta+2}\right)^{\alpha j+i}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t} \\ \dfrac{1}{\Gamma[-\gamma]}(1+\theta)^{\gamma}\left(\dfrac{\theta}{1+\theta}\right)^{-1-i-\alpha j}\Gamma[1+i+\alpha j]\Gamma[-1-i-\alpha j-\gamma] \\ Hypergeometric1F1\left[1+i+\alpha j, 2+i+\alpha j+\gamma, \alpha(1+\theta)(j+\gamma)\right] \\ +\left(\dfrac{\theta}{1+\theta}\right)^{\gamma}\left(\alpha\theta(j+\gamma)\right)^{-1-i-\alpha j-\gamma}\Gamma[-\gamma]\Gamma[1+i+\alpha j+\gamma] \\ Hypergeometric1F1\left[-\gamma, -i-\alpha j-\gamma, \alpha(1+\theta)(j+\gamma)\right] \end{array}\right).$$

**Theorem.3.5. .** Suppose X has the Kw- Extended Garima distribution. Tsallis Entropy of X is thus given by

$$= \frac{1}{1-\lambda}\left(\begin{array}{l} \alpha\beta\dfrac{\theta}{\theta+2}\left(\dfrac{\theta}{\theta+2}\right)^{\alpha j+i}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{t=0}^{\infty}(-1)^{\alpha-1-i}(-1)^{\beta-1-j}(-1)^{\alpha j-k}\binom{\alpha-1}{i}\binom{\beta-1}{j}\binom{\alpha j}{k}\binom{\alpha j+i}{t} \\ \dfrac{1}{\Gamma[-\lambda]}(1+\theta)^{\lambda}\left(\dfrac{\theta}{1+\theta}\right)^{-1-i-\alpha j}\Gamma[1+i+\alpha j]\Gamma[-1-i-\alpha j-\lambda] \\ Hypergeometric1F1\left[1+i+\alpha j, 2+i+\alpha j+\lambda, \alpha(1+\theta)(j+\lambda)\right] \\ +\left(\dfrac{\theta}{1+\theta}\right)^{\lambda}\left(\alpha\theta(j+\lambda)\right)^{-1-i-\alpha j-\lambda}\Gamma[-\lambda]\Gamma[1+i+\alpha j+\lambda] \\ Hypergeometric1F1\left[-\lambda, -i-\alpha j-\lambda, \alpha(1+\theta)(j+\lambda)\right] \end{array}\right)$$

The hypergeometric series, which covers many other special functions as specific or limiting cases, is used to present the hypergeometric1F1 function here.

**Proof.** Tsallis Entropy for the Kw-Extended Garima Distribution;

$$S_{\lambda} = \frac{1}{1-\lambda}\left(1-\int_0^{\infty} f^{\lambda}(x)dx\right)$$

$$= \frac{1}{1-\lambda}\left(1-\int_0^{\infty}\left\{\begin{array}{l} \alpha\beta\dfrac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x} \\ \left(1-\left[1+\dfrac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\dfrac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1} \end{array}\right\}^{\lambda} dx\right).$$

The Tsallis entropy is obtained by integrating the above equation. This brings the proof to a close.

## 4. Reliability Analysis

The reliability or survival function for the Kw-Extended Garima distribution is stated as (7) in R(t), and Figure 3 depicts the reliability function of the Kw-Extended Garima distribution for different parameter values:

$$R(t)=\left(1-\left\{1-\left[1+\frac{\theta t}{\theta+2}\right]e^{-\theta t}\right\}^{\alpha}\right)^{\beta}. \tag{23}$$

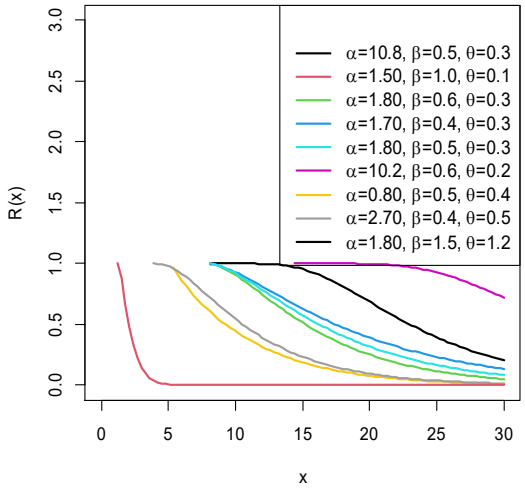



**Figure 3.** The reliability function of Kw- Extended Garima distributions

The hazard rate function, defined h as the event at time t conditional on survival until time t, is the other function. Assume that an item has survived for time t and state the likelihood,

$$h(t)=\lim_{dt\to 0}\frac{P\left(t\le T\le t+dt)\right)}{dtR(t)}=\frac{f(t)}{R(t)}=-\frac{R'(t)}{R(t)}. \tag{24}$$

The hazard rate for the Kw- Extended Garima distribution can be calculated by combining (8) and (23) and defining it as follows;

$$h(t)=\frac{\alpha\beta\dfrac{\theta}{\theta+2}(1+\theta+\theta t)e^{-\theta t}\left(1-\left[1+\dfrac{\theta t}{\theta+2}\right]e^{-\theta t}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\dfrac{\theta t}{\theta+2}\right]e^{-\theta t}\right\}^{\alpha}\right)^{\beta-1}}{\left(1-\left\{1-\left[1+\dfrac{\theta t}{\theta+2}\right]e^{-\theta t}\right\}^{\alpha}\right)^{\beta}}. \quad (25)$$

**Figure 4**. Depicts the hazard rate function of the Kw-Extended Garima distribution for various parameter values.

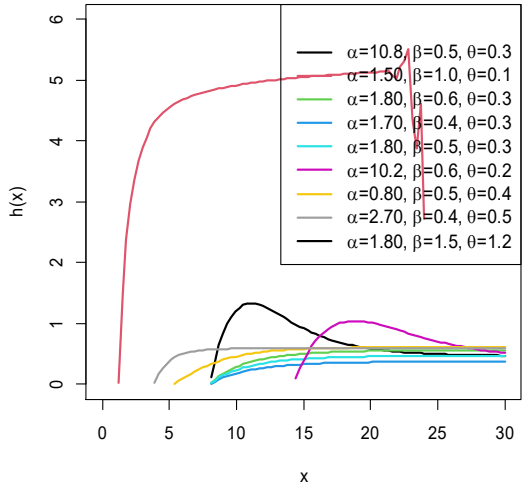



**Figure 4.** The hazard rate function of Kw- Extended Garima distributions

The reversed hazard rate is defined as follows;

$$r(x)=\frac{f(x)}{F(x)}. \quad (26)$$

Setting (7) and (8) in (26) yields the reversed hazard rate for the Kw- Extended Garima distribution, which has the following format. Figure 5 depicts the reverse hazard rate function of the Kw- Extended Garima distribution for different parameter values.

$$r(x)=\frac{\alpha\beta\dfrac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\dfrac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\dfrac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}}{1-\left(1-\left\{1-\left[1+\dfrac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}}.$$

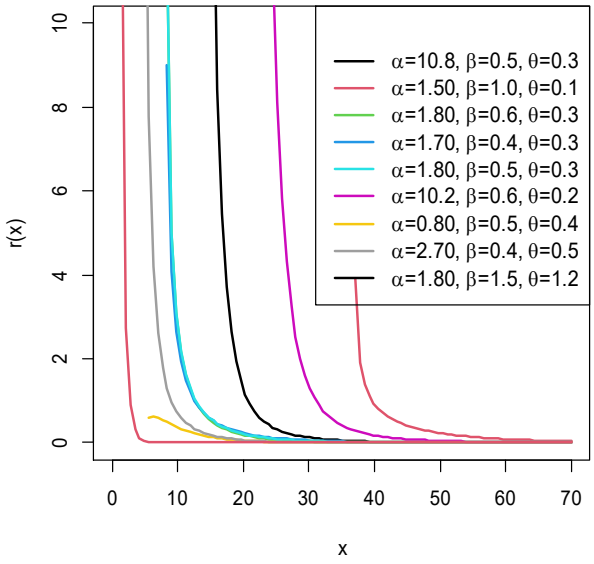



**Figure 5.** The reversed hazard rate function of Kw- Extended Garima distributions

The following format describes the odds function;

$$O(x)=\frac{F(x)}{R(x)}. \tag{27}$$

Setting (8) and (10) in (27) yields the odds function for the Kw- Extended Garima distribution, which has a following statement.

$$O(x)=\frac{1-\left(1-\left\{1-\left[1+\dfrac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}}{\left(1-\left\{1-\left[1+\dfrac{\theta t}{\theta+2}\right]e^{-\theta t}\right\}^{\alpha}\right)^{\beta}}. \tag{28}$$

Figure 6 depicts the odds function of the Kw- Extended Garima distribution for various parameter values.

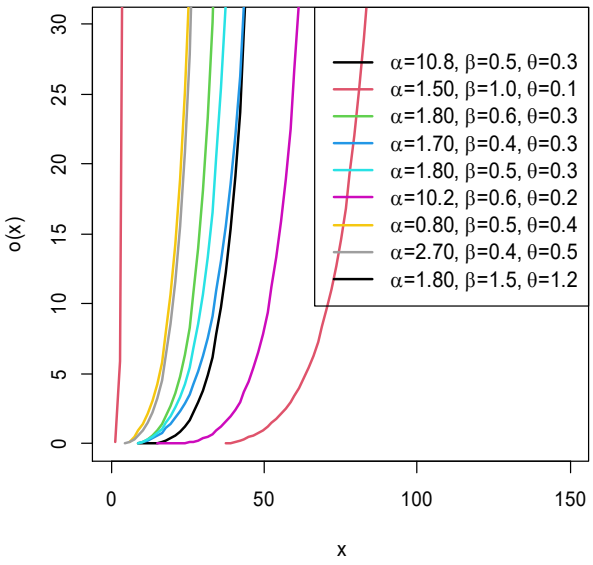



Figure 6. The odds function The Kw- Extended Garima distributions

## 5. Order Statistics

The jth order statistics of the Kw- Extended Garima distribution's pdf and cdf are as follows:

$$f_{j:n}(x)=\frac{n!}{(j-1)!(n-i)!}\left[F(x)\right]^{j-1}\left[1-F(x)\right]^{n-j}f(x)$$

$$f_{X_{(j)}}(x)=\frac{n!}{(j-1)!(n-j)!}\alpha\beta\frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}$$

$$\left(1-\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}\right)^{j-1}\left(1-\left(1-\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}\right)\right)^{n-j}.$$

$$F_{j:n}(x)=\sum_{r=j}^{n}\left(F(x)\right)^{r}\left(1-F(x)\right)^{n-r}$$

$$=\binom{n}{j}\left(1-\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)\right)^{i}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{n-i}$$

For the largest order statistic, the pdf is

$$f_{X_{(n)}}(x) = n\alpha\beta\frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}$$

$$\left(1-\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}\right)^{n-1},$$

and here is the pdf for the smallest order statistic:

$$f_{X_{(1)}}(x) = n\alpha\beta\frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta-1}$$

$$\left(1-\left(1-\left(1-\left\{1-\left[1+\frac{\theta x}{\theta+2}\right]e^{-\theta x}\right\}^{\alpha}\right)^{\beta}\right)\right)^{n-1}.$$

## 6. Maximum Likelihood Estimation (MLE)

This function of Kw- Extended Garima is described in the following format:

$$\phi(X) = \alpha^n\beta^n\left(\frac{\theta}{\theta+2}\right)^n e^{-\theta\sum_{i=1}^{n}x_i}\prod_{i=1}^{n}(1+\theta+\theta x_i)\left(1-\left[1+\frac{\theta x_i}{\theta+2}\right]e^{-\theta x_i}\right)^{\alpha-1}\left(1-\left\{1-\left[1+\frac{\theta x_i}{\theta+2}\right]e^{-\theta x_i}\right\}^{\alpha}\right)^{\beta-1}.$$

(29)

The function is as follows:

$$l_n = \log(\phi) = n\log\alpha + n\log\beta + n\log\theta - n\log(\theta+2) - \theta\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\log(1+\theta+\theta x_i) + (\alpha-1)\sum_{i=1}^{n}\log\left(1-\left[1+\frac{\theta x_i}{\theta+2}\right]e^{-\theta x_i}\right)$$

$$+(\beta-1)\sum_{i=1}^{n}\log\left(1-\left\{1-\left[1+\frac{\theta x_i}{\theta+2}\right]e^{-\theta x_i}\right\}^{\alpha}\right).$$

(30)

Now setting,

$$\frac{\partial l_n}{\partial\alpha}=0,\frac{\partial l_n}{\partial\beta}=0,\frac{\partial l_n}{\partial\theta}=0.$$

We have obtained,

$$\frac{n}{\alpha}+\sum_{i=1}^{n}\log\left[1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]+\left(-1+\beta\right)\sum_{i=1}^{n}-\frac{\log\left[1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}{1-\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}=0$$

$$\frac{n}{\beta}+\sum_{i=1}^{n}\log\left[1-\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}\right]=0$$

$$\frac{n}{\theta}-\frac{n}{2+\theta}-\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\frac{1+x_i}{1+\theta+\theta x_i}+\left(-1+\alpha\right)\sum_{i=1}^{n}\frac{-\mathrm{e}^{-\theta x_i}\left(-\frac{\theta x_i}{\left(2+\theta\right)^2}+\frac{x_i}{2+\theta}\right)+\mathrm{e}^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)}{1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)}$$

$$+\left(-1+\beta\right)\sum_{i=1}^{n}-\frac{\alpha\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{-1+\alpha}\left(-\mathrm{e}^{-\theta x_i}\left(-\frac{\theta x_i}{\left(2+\theta\right)^2}+\frac{x_i}{2+\theta}\right)+\mathrm{e}^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)\right)}{1-\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}=0$$

Estimates are derived by partially differentiating the equation (30) with respect to $\alpha,\beta$ and $\theta$, but the EM method is the best strategy for estimating the parameters. By solving this nonlinear system of equations, the MLE $\left(\hat{\alpha},\hat{\beta},\hat{\theta}\right)$ of $\left(\alpha,\beta,\theta\right)$ is obtained. To numerically optimize the sample likelihood function given in (29), it is usually more convenient to use nonlinear optimization algorithms such as the quasi-Newton approach. The MLE $\hat{\gamma}=\left(\hat{\alpha},\hat{\beta},\hat{\theta}\right)$ can be approximated as tri-variate normal with mean $\hat{\eta}$ and variance-covariance matrix equal to the inverse of the expected information matrix. $I^{-1}(\gamma)$ denotes the variance-covariance matrix of $\hat{\gamma}$. The members of the three-dimensional matrix $I(\gamma)$ can be estimated using $\sqrt{n}\left(\hat{\gamma}-\gamma\right)\rightarrow N_3\left(0,nI^{-1}(\gamma)\right)$.

Where, $I^{-1}(\gamma)$ is the variance-covariance matrix of $\hat{\gamma}$ .The elements of the 3 × 3 matrix $I(\gamma)$ can be estimated by

$$I^{-1} = -E\begin{bmatrix} \frac{\partial^2 \log L}{\partial \alpha^2} & \frac{\partial^2 \log L}{\partial \alpha \partial \beta} & \frac{\partial^2 \log L}{\partial \alpha \partial \theta} \\ \frac{\partial^2 \log L}{\partial \beta \partial \alpha} & \frac{\partial^2 \log L}{\partial \beta^2} & \frac{\partial^2 \log L}{\partial \beta \partial \theta} \\ \frac{\partial^2 \log L}{\partial \theta \partial \alpha} & \frac{\partial^2 \log L}{\partial \theta \partial \beta} & \frac{\partial^2 \log L}{\partial \theta^2} \end{bmatrix} = \begin{bmatrix} I_{11}^{-1} & I_{12}^{-1} & I_{13}^{-1} \\ I_{21}^{-1} & I_{22}^{-1} & I_{23}^{-1} \\ I_{31}^{-1} & I_{32}^{-1} & I_{33}^{-1} \end{bmatrix}$$

The Hessian matrix entries that correspond to the elements in Equation (29),

$$\frac{\partial^2 \log L}{\partial \alpha^2} = -\frac{n}{\alpha^2} + \left(-1+\beta\right)$$
$$\sum_{i=1}^{n}\left( -\frac{\log\left[1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]^2 \left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{2\alpha}}{\left(1-\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}\right)^2} - \frac{\log\left[1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]^2 \left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}{1-\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}} \right)$$

$$\frac{\partial^2 \log L}{\partial \alpha \partial \beta} = \sum_{i=1}^{n} - \frac{\log\left[1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}{1-\left(1-\mathrm{e}^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}$$

$$\frac{\partial^2 \log L}{\partial\alpha\partial\theta} = \sum_{i=1}^{n} \frac{-e^{-\theta x_i}\left(-\frac{\theta x_i}{(2+\theta)^2}+\frac{x_i}{2+\theta}\right)+e^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)}{1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)}+(-1+\beta)$$

$$\sum_{i=1}^{n}\left(\begin{array}{l} -\dfrac{\alpha\log\left[1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{-1+2\alpha}\left(-e^{-\theta x_i}\left(-\frac{\theta x_i}{(2+\theta)^2}+\frac{x_i}{2+\theta}\right)+e^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)\right)}{\left(1-\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}\right)^2} \\ -\dfrac{\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{-1+\alpha}\left(-e^{-\theta x_i}\left(-\frac{\theta x_i}{(2+\theta)^2}+\frac{x_i}{2+\theta}\right)+e^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)\right)}{1-\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}} \\ -\dfrac{\alpha\log\left[1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right]\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{-1+\alpha}\left(-e^{-\theta x_i}\left(-\frac{\theta x_i}{(2+\theta)^2}+\frac{x_i}{2+\theta}\right)+e^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)\right)}{1-\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}} \end{array}\right)$$

$$\frac{\partial^2 \log L}{\partial\beta^2} = -\frac{n}{\beta^2}$$

$$\frac{\partial^2 \log L}{\partial\beta\partial\theta} = \sum_{i=1}^{n} -\frac{\alpha\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{-1+\alpha}\left(-e^{-\theta x_i}\left(-\frac{\theta x_i}{(2+\theta)^2}+\frac{x_i}{2+\theta}\right)+e^{-\theta x_i}x_i\left(1+\frac{\theta x_i}{2+\theta}\right)\right)}{1-\left(1-e^{-\theta x_i}\left(1+\frac{\theta x_i}{2+\theta}\right)\right)^{\alpha}}$$

$$\frac{\partial^2 \log L}{\partial\theta^2} = -\frac{n}{\theta^2}+\frac{n}{(2+\theta)^2}+\sum_{i=1}^{n}-\frac{(1+x_i)^2}{(1+\theta+\theta x_i)^2}+(-1+\alpha)A-(1-\beta)\sum_{i=1}^{n}B+C+D$$

$$A = \sum_{i=1}^{n} \left( -\frac{\left( -\mathrm{e}^{-\theta x_i}\left( -\frac{\theta x_i}{(2+\theta)^2} + \frac{x_i}{2+\theta} \right) + \mathrm{e}^{-\theta x_i} x_i \left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^2}{\left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^2} + \frac{-\mathrm{e}^{-\theta x_i}\left( \frac{2\theta x_i}{(2+\theta)^3} - \frac{2x_i}{(2+\theta)^2} \right) + 2\mathrm{e}^{-\theta x_i} x_i \left( -\frac{\theta x_i}{(2+\theta)^2} + \frac{x_i}{2+\theta} \right) - \mathrm{e}^{-\theta x_i} x_i^2 \left( 1 + \frac{\theta x_i}{2+\theta} \right)}{1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right)} \right)$$

$$B = \frac{\alpha^2 \left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^{-2+2\alpha} \left( -\mathrm{e}^{-\theta x_i}\left( -\frac{\theta x_i}{(2+\theta)^2} + \frac{x_i}{2+\theta} \right) + \mathrm{e}^{-\theta x_i} x_i \left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^2}{\left( 1 - \left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^{\alpha} \right)^2},$$

$$C = \frac{(-1+\alpha)\alpha \left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^{-2+\alpha} \left( -\mathrm{e}^{-\theta x_i}\left( -\frac{\theta x_i}{(2+\theta)^2} + \frac{x_i}{2+\theta} \right) + \mathrm{e}^{-\theta x_i} x_i \left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^2}{1 - \left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^{\alpha}},$$

$$D = \frac{\alpha \left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^{-1+\alpha} \left( -\mathrm{e}^{-\theta x_i}\left( \frac{2\theta x_i}{(2+\theta)^3} - \frac{2x_i}{(2+\theta)^2} \right) + +2\mathrm{e}^{-\theta x_i} x_i \left( -\frac{\theta x_i}{(2+\theta)^2} + \frac{x_i}{2+\theta} \right) - \mathrm{e}^{-\theta x_i} x_i^2 \left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)}{1 - \left( 1 - \mathrm{e}^{-\theta x_i}\left( 1 + \frac{\theta x_i}{2+\theta} \right) \right)^{\alpha}}$$

The approximate two-sided confidence intervals for $\alpha, \beta, \theta$ are as follows:

$$\widehat{\alpha} \pm z_{\alpha/2}\left( I_{11}^{-1}(\widehat{\gamma}) \right)^{1/2} \quad \widehat{\beta} \pm z_{\alpha/2}\left( I_{22}^{-1}(\widehat{\gamma}) \right)^{1/2} \text{ and } \quad \widehat{\theta} \pm z_{\alpha/2}\left( I_{22}^{-1}(\widehat{\gamma}) \right)^{1/2}$$

where $z_{\alpha}$ denotes the upper $\alpha$ th quantile of the standard normal distribution.

## 7. Application

### 7.1. Simulation Studies

We ran a simulation to explore the flexibility and competency of the Kumaraswamy Extended Garima distribution class. Eghwerido et al. [9], Team, R.C [28] used R for computing. The simulation was investigated further below;

- The quantile function of the Kumaraswamy Extended Garima distribution was used to create the data, as shown in equation (9)
- The sample sizes of 5, 10, 20, 30, 50, 80 and 100, for, $\alpha = 0.5$ , $\beta = 2$ , $\theta = 0.5$ ,The MSE of the parameters $\gamma = (\alpha, \beta, \theta)$ .The simulation study investigated the mean estimates, variance, MSE (Mean Square Error).

- $$MSE = \sum_{i=1}^{1000} \frac{\left(\hat{\gamma} - \gamma\right)^2}{1000}.$$

**Table 1.** Monte Carlo simulation study for the Kumaraswamy Garima Distrubution

| | Parameter | Variance | MSE |
|---|---|---|---|
| 5 | $\alpha = 0.5$ | 0.098978 | 0.1572 |
| | $\beta = 2$ | 0.097515 | 0.1516 |
| | $\theta = 0.5$ | 0.099191 | 0.1338 |
| 10 | $\alpha = 0.5$ | 0.8425 | 0.1226 |
| | $\beta = 2$ | 0.089813 | 0.1377 |
| | $\theta = 0.5$ | 0.080249 | 0.1096 |
| 20 | $\alpha = 0.5$ | 0.089228 | 0.0753 |
| | $\beta = 2$ | 0.076534 | 0.0832 |
| | $\theta = 0.5$ | 0.072762 | 0.0804 |
| 30 | $\alpha = 0.5$ | 0.080662 | 0.0716 |

| | | | |
|---|---|---|---|
| | $\beta = 2$ | 0.090617 | 0.0685 |
| | $\theta = 0.5$ | 0.084565 | 0.0738 |
| 50 | $\alpha = 0.5$ | 0.093203 | 0.0700 |
| | $\beta = 2$ | 0.085992 | 0.0694 |
| | $\theta = 0.5$ | 0.084664 | 0.0704 |
| 80 | $\alpha = 0.5$ | 0.07661 | 0.0692 |
| | $\beta = 2$ | 0.08511 | 0.0681 |
| | $\theta = 0.5$ | 0.070994 | 0.0677 |
| 100 | $\alpha = 0.5$ | 0.074335 | 0.0652 |
| | $\beta = 2$ | 0.070458 | 0.0586 |
| | $\theta = 0.5$ | 0.070579 | 0.0590 |

In our simulation analysis, the experiment from Table 1 was repeated 1000 times, and the MSE values fell as the sample size increased. This demonstrates that the parameter estimate is consistent with the asymptotic theory or big sample theory.

### 7.2. Real Data Analysis

We employed four real lifespan data sets in Kumaraswamy's Extended Garima distribution and compared the model to exponentiated Garima (Rather and Subramanian [24]), Garima, Exponential distribution, Generalized Exponential, and Exponentiated Weibull. To compare the Kumaraswamy Extended Garima, exponentiated Garima distribution with Garima, Exponential distribution, Generalized Exponential, and Exponentiated Weibull distributions, these distributions are provided in the following order:

(i) Exponential distribution

$$f(x;\theta) = \theta e^{-\theta x}, \mathrm{x} \geq 0, \theta > 0$$

(ii) Garima distribution

$$f(x;\theta)=\frac{\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x},x>0,\theta>0.$$

(iii) Exponentiated Garima

$$\mathrm{f}(x;\theta,\alpha)=\frac{\alpha\theta}{\theta+2}(1+\theta+\theta x)e^{-\theta x}\left(1-\left(1+\frac{\theta x}{\theta+2}\right)e^{-\theta x}\right)^{\alpha-1},x>0,\alpha,\theta>0.$$

(iv) Generalized Exponential

$$\mathrm{f}(x;\theta,\alpha)=\alpha\theta e^{-\theta x}\left(1-e^{-\theta x}\right)^{\alpha-1},x>0,\alpha,\theta>0.$$

(v) Exponentiated Weibull

$$\mathrm{f}(x;\theta,\alpha,\beta)=\frac{\alpha\theta}{\beta^{\alpha}}x^{\alpha-1}e^{-(x/\beta)^{\alpha}}\left(1-e^{-(x/\beta)^{\alpha}}\right)^{\theta-1},$$

$$x>0,\alpha,\theta>0,\beta>0.$$

We take into account metrics such as the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC), the Hannan Quinn (HIC), and -2 logL. The model utilized is best if the values of AIC, BIC, HIC, AICC, and - 2 log L are as low as possible. The following formulas can be used to calculate AIC, BIC, HIC, AICC, and -2 logL;

$$AIC=2k\ -\ 2\log L$$

$$BIC=k\log n\ -\ 2\log L$$

$$HIC=2\mathrm{kln}(\mathrm{In}(\mathrm{n}))-\ 2\log L$$

$$AICC=AIC+\frac{2k(k+1)}{n-k-1}$$

**Data set 1**: The data set exhibits the tensile strength, measured in GPa, of n=69 carbon fibers tested under tension at gauge lengths of 20mm, as reported by (see M. Bader and Priest [3], Almanjahie et al. [1], Mead et al. [17]). The information is as follows:

1.901, 2.132, 2.203, 2.228, 2.257, 2.350, 2.361, 2.396, 2.397, 2.445, 2.454, 2.474, 2.518, 2.522, 2.525, 2.532, 2.575, 2.614, 2.616, 2.618, 2.624, 2.659, 2.675, 2.738, 2.740, 2.856, 2.917, 2.928, 2.937, 2.937, 2.977, 2.996, 3.030, 3.125, 3.139, 3.145,3.220, 3.223, 3.235, 3.243, 3.264, 3.272, 3.294, 3.332,

3.346, 3.377, 3.408, 3.435, 3.493, 3.501, 3.537, 3.554, 3.562, 3.628, 3.852, 3.871, 3.886, 3.971, 4.024, 4.027 ,4.225, 4.395, 5.020.

**Data set 2**: To define the novel findings presented in this paper, we applied the kw extended garima distribution to the data set utilized by Nichols and Padgett [22], which included 100 research on the fracture stress of carbon fibers (in Gba).

3.7, 2.74, 2.73, 2.5, 3.6, 3.11, 3.27, 2.87, 1.47, 3.11, 4.42, 2.41, 3.19, 3.22, 1.69, 3.28, 3.09, 1.87, 3.15, 4.9, 3.75, 2.43, 2.95,2.97, 3.39, 2.96, 2.53, 2.67, 2.93, 3.22, 3.39, 2.81, 4.2, 3.33, 2.55, 3.31, 3.31, 2.85, 2.56, 3.56, 3.15, 2.35, 2.55, 2.59, 2.38, 2.81, 2.77, 2.17, 2.83, 1.92, 1.41, 3.68, 2.97, 1.36, 0.98, 2.76, 4.91, 3.68, 1.84, 1.59, 3.19, 1.57, 0.81, 5.56, 1.73, 1.59, 2, 1.22,1.12, 1.71, 2.17, 1.17, 5.08, 2.48, 1.18, 3.51, 2.17, 1.69, 1.25, 4.38, 1.84, 0.39, 3.68, 2.48, 0.85, 1.61, 2.79, 4.7, 2.03, 1.8, 1.57,1.08, 2.03, 1.61, 2.12, 1.89, 2.88, 2.82, 2.05, 3.65.

**Data Set 3.** Bjerkedalen [4] investigated and informed on the survival periods in days of 72 guinea pigs infected with virulent tubercle bacilli.

0.1, 0.33, 0.44, 0.56, 0.59, 0.72, 0.74, 0.77, 0.92, 0.93, 0.96, 1, 1, 1.02, 1.05, 1.07, 1.07, 1.08, 1.08, 1.08,1.09, 1.12, 1.13, 1.15, 1.16, 1.2, 1.21, 1.22, 1.22, 1.24, 1.3, 1.34, 1.36, 1.39, 1.44, 1.46, 1.53, 1.59, 1.6,1.63, 1.63, 1.68, 1.71, 1.72, 1.76, 1.83, 1.95, 1.96, 1.97, 2.02, 2.13, 2.15, 2.16, 2.22, 2.3, 2.31, 2.4, 2.45,2.51, 2.53, 2.54, 2.54, 2.78, 2.93, 3.27, 3.42, 3.47, 3.61, 4.02, 4.32, 4.58, 5.55.

**Data Set 4.** The data shown below represent the failure times of 84 aircraft windshields as reported by Cordeiro, G. M., and Castro, M. A [7]. The data sets are listed below.

0.040, 1.866, 2.385, 3.443, 0.301, 1.876, 2.481, 3.467, 0.309, 1.899, 2.610, 3.478, 0.557, 1.911, 2.625, 3.578, 0.943, 1.912, 2.632, 3.595, 1.070, 1.914, 2.646, 3.699, 1.124, 1.981, 2.661, 3.779, 1.248, 2.010, 2.688, 3.924, 1.281, 2.038, 2.82, 3, 4.035, 1.281, 2.085, 2.890, 4.121, 1.303, 2.089, 2.902, 4.167, 1.432, 2.097, 2.934, 4.240, 1.480, 2.135, 2.962, 4.255, 1.505, 2.154, 2.964, 4.278, 1.506, 2.190, 3.000, 4.305, 1.568, 2.194, 3.103, 4.376, 1.615, 2.223, 3.114, 4.449, 1.619, 2.224, 3.117, 4.485, 1.652, 2.229, 3.166, 4.570, 1.652, 2.300, 3.344, 4.602, 1.757, 2.324, 3.376, 4.663.

**Table 2.** MLEs and the AIC, BIC and Logl statistics.

| **Data Set** | **Distribution** | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | **-2 logL** | **AIC** | **BIC** | **HIC** | **AICC** |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Kw Extended Garima | 0.39 | 7.66 | 0.1148 | 61.205 | 67.205 | 66.721 | 69.864 | 67.574 |
| | Exponentiated Garima | | - | 2.1527 | 112.92 | 116.92 | 115.94 | 118.04 | 117.10 |
| | Garima | - | - | 0.4784 | 256.32 | 258.32 | 258.15 | 259.20 | 258.38 |
| | Exponential | - | - | 0.3266 | 266.89 | 268.89 | 268.73 | 269.77 | 268.95 |
| | Generalized Exponantial | 218.23 | - | 1.9458 | 113.03 | 115.03 | 116.71 | 118.80 | 115.23 |
| | Exponantiated Weibull | 0.81 | 1.50 | 31.802 | 116.62 | 118.62 | 117,66 | 125.28 | 118.99 |
| 2 | Kw Extended Garima | 0.13 | 2.90 | 0.0363 | 111.87 | 117.87 | 117.87 | 121.03 | 118.12 |
| | Exponentiated Garima | 6.68 | - | 1.2109 | 240.49 | 244.49 | 244.49 | 246.60 | 245.05 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Garima | - | - | 0.5485 | 379.41 | 381.41 | 381.41 | 382.47 | 381.46 |
| | Exponential | - | - | 0.3814 | 392.74 | 394.74 | 394.74 | 395.79 | 394.78 |
| | Generalized Exponantial | 7.78 | - | 1.0131 | 292.36 | 294.34 | 296.36 | 298.47 | 294.46 |
| | Exponantiated Weibull | 0.37 | 2.40 | 1.3198 | 286.91 | 288.91 | 292.91 | 296.07 | 289.16 |
| 3 | Kw Extended Garima | 0.08 | 1.90 | 0.1781 | 103.74 | 109.74 | 107.46 | 112.46 | 110.09 |
| | Exponentiated Garima | 3.22 | - | 1.3626 | 188.13 | 192.13 | 191.85 | 193.94 | 192.3 |
| | Garima | - | - | 0.78350 | 220.05 | 222.05 | 221.90 | 221.50 | 222.1 |
| | Exponential | - | - | 0.56554 | 226.07 | 228.07 | 227.93 | 228.98 | 228.1 |
| | Generalized Exponantial | 3.62 | - | 1.12713 | 188.47 | 190.47 | 192.18 | 194.28 | 190.6 |
| | Exponantiated Weibull | 0.88 | 1.16 | 2.6523 | 192.1 | 194.1 | 197.73 | 218.32 | 194.5 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4 | Kw Extended Garima | 6 | 2.43 | 0.5 | 92.25 | 98.25 | 98.026 | 101.18 | 98.55 |
| | Exponentiated Garima | 3.22 | - | 0.9508 | 278.6 | 282.6 | 282.54 | 284.65 | 282.8 |
| | Garima | - | - | 0.5585 | 319.6 | 321.6 | 321.57 | 322.62 | 321.7 |
| | Exponential | | | 0.3902 | 329.9 | 331.9 | 331.89 | 332.95 | 332.0 |
| | Generalized Exponantial | 3.59 | - | 0.7594 | 282.7 | 284.7 | 286.64 | 288.7 | 284.9 |
| | Exponantiated Weibull | 0.25 | 5.82 | 0.28168 | 261.58 | 263.58 | 267.358 | 270.51 | 263.88 |

Table 2 clearly shows that the Kumaraswamy Garima distribution has values of AIC, BIC, HIC, AICC, and -2 logL when compared to exponentiated Garima distribution, Garima, Exponential distributions, Generalized Exponential, and Exponentiated Weibull. As a result, the Kumaraswamy Garima distribution fits better than the exponentiated Garima, Garima, Exponential, Generalized Exponential, and Exponentiated Weibull distributions.

**8. Conclusion**

In our research, we developed a novel distribution known as the Kw- Extended Garima, where Garima is a single-parameter life distribution. We've also gotten the probability and cumulative distribution functions for this new distribution. The reliability function, hazard rate, failure rate, inverse hazard function, odd function, cumulative hazard function, r-th moment, moment generating function, characteristic function, moments, mean and variance, Bonferroni and Lorenz curves, order statistics,

MLE, and mean time between failures (MTBF) were then obtained based on these functions. We plotted these functions in the R program and obtained some results in the Mathematica program. We use maximum likelihood to forecast the parameters of the innovative model. In addition, we create the information matrix. With the help of simulation analysis, our distribution supported the findings of the asymptotic theory. We discovered that our unique model outperformed previous models when applied to four real-world data sets. The Kumaraswamy Garima distribution has been shown to be superior to the other distributions.

**Data Availability**

There is data supporting the findings of this study, where it was taken from is clearly stated in the text, it can be sent again by the relevant author if desired.

**Conflict of Interest**

The article's authors declare that there is no conflict of interest between them.

**Author's Contributions**

The contribution of the authors is equal.

**REFERENCES**

[1] Almanjahie I. M., Dar J. G., Laksaci, A., Ahmad, I., A new probability model for modeling of strength of carbon fiber data: properties and applications, Environmental and Ecological Statistics. 28(3), 523-547, (2021).

[2] Asiribo O.E., Mabur T.M., Soyinka A.T., On the Lomax-Kumaraswamy distribution, Benin Journal of Statistics, Vol, 2, 107-120, (2019).

[3] Bader M.G., Priest A.M., Statistical aspects of fibre and bundle strength in hybrid composites. Progress in science and engineering of composites, 1129-1136, (1982).

[4] Bjerkedal T., Acquisition of Resistance in Guinea Pies infected with Different Doses of Virulent Tubercle Bacilli, American Journal of Hygiene, 72(1), 130-48, (1960).

[5] Carrasco J.M., Ferrari, S.L., Cordeiro G.M., A new generalized Kumaraswamy distribution. arXiv preprint arXiv:1004.0911, (2010).

[6] Carrasco J.M., Cordeiro G.M., An extension of the Kumaraswamy distribution, International Journal of Statistics and Probability, 6(3), 61, (2017).

[7] Cordeiro G.M., de Castro M. A new family of generalized distributions, Journal of statistical computation and simulation, 81(7), 883-898, (2011).

[8] Dey S., Mazucheli J., Nadarajah S., Kumaraswamy distribution: different methods of estimation, Computational and Applied Mathematics, 37, 2094-2111, (2018).

[9] Eghwerido J.T., Ogbo J.O., Omotoye A. E., The Marshall-Olkin Gompertz distribution: properties and applications. Statistica, 81(2), 183-215, (2021).

[10] El-Sherpieny E.S.A., Ahmed M.A., On the kumaraswamy Kumaraswamy distribution, International Journal of Basic and Applied Sciences, 3(4), 372, (2014).

[11] Garg M., On Distribution of Order Statistics from Kumaraswamy Distribution, Kyungpook mathematical journal, 48(3), (2008).

[12] Gomes A.E., da-Silva C.Q., Cordeiro G.M., Ortega E.M., A new lifetime model: the Kumaraswamy generalized Rayleigh distribution, Journal of statistical computation and simulation, 84(2), 290-309, (2014).

[13] Iqbal Z., Tahir M.M., Riaz N., Ali S.A., Ahmad M., Generalized inverted kumaraswamy distribution: properties and application, Open Journal of Statistics, 7(4), 645-662, (2017).

[14] Jones M.C.,Kumaraswamy's distribution: A beta-type distribution with some tractability advantages, Statistical methodology, 6(1), 70-81, (2009).

[15] Kumaraswamy P.A., Generalized probability density function for double-bounded random processes, Journal of hydrology, 46(1-2), 79-88, (1980).

[16] Linhart H., Zucchini W., Model Selection, Wiley. New York, (1986).

[17] Mead M.E., Afify A.Z., Hamedani G.G., Ghosh I., The beta exponential Fréchet distribution with applications, Austrian Journal of Statistics, 46(1), 41-63, (2017).

[18] Mitnik P.A., Baek S., The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation, Statistical Papers, 54, 177-192, (2013).

[19] Mitnik P.A., New properties of the Kumaraswamy distribution, Communications in Statistics-Theory and Methods, 42(5), 741-755, (2013).

[20] Mohiuddin M., Rather A.A., Subramanian C., Dar S.A., Transmuted Garima Distribution: Properties and Applications, Journal of Xidian University, 14(3), (2020).

[21] Nadarajah S., On the distribution of Kumaraswamy, Journal of Hydrology, 348(3), 568-569, (2008).

[22] Nichols M.D., Padgett W.J., A bootstrap control chart for Weibull percentiles, Quality and reliability engineering international, 22(2), 141-151, (2006).

[23] Paranaiba P.F., Ortega E.M., Cordeiro G.M., Pascoa M.A D., The Kumaraswamy Burr XII distribution: theory and practice, Journal of Statistical Computation and Simulation, 83(11), 2117-2143, (2013).

[24] Rather A.A., Subramanian C., A New Exponentiated Distribution with Engineering Science Applications, J. Stat. Appl., Pro, 9, 127-137, (2020).

[25] Salman M.S., Comparing Different Estimators of two Parameters Kumaraswamy Distribution, Journal of Babylon University, Pure and Applied Sciences, 25(2), 395-402, (2017).

[26] Tahir M.H., Zubair M., Mansoor M., Cordeiro G.M., Alizahdehk M., Hamedani G., A new Weibull-G family of distributions, Hacettepe Journal of Mathematics and statistics, 45(2), 629-647, (2016).

[27] Tahir M.H., Hussain M.A., Cordeiro G.M., El-Morshedy M., Eliwa M.S., A new Kumaraswamy generalized family of distributions with properties, applications, and bivariate extension, Mathematics, 8(11), 1989, (2020).

[28] Team R.C., R: A language and environment for statistical computing computer program, version 3.6. 1. R Core Team, Vienna, Austria (2019).

[28] Team R.D.C., R: A language and environment for statistical computing, (2010).

[29] Yang T., Statistical properties of Kumaraswamy generalized inverse Weibull distribution, (2012).

[30] Wang B.X., Wang X.K., Yu K., Inference on the Kumaraswamy distribution, Communications in Statistics-Theory and Methods, 46(5), 2079-2090, (2017).