Kopula Temelli Değişken Kümeleme Tekniklerinin İncelenmesi ve Mortalite Tahmini Uygulaması

Zeynep İlhan

DOKTORA TEZİ

İstatistik Anabilim Dalı

Ekim 2019

Analysis of Copula Based Variable Clustering Techniques and Application of Mortality Estimation

Zeynep Ilhan

DOCTORAL DISSERTATION

Department of Statistics

October 2019

Kopula Temelli Değişken Kümeleme Tekniklerinin İncelenmesi ve Mortalite Tahmini Uygulaması

Zeynep İlhan

Eskişehir Osmangazi Üniversitesi
Fen Bilimleri Enstitüsü
Lisansüstü Yönetmeliği Uyarınca
İstatistik Anabilim Dalı
Risk Analizi Bilim Dalında
DOKTORA TEZİ
Olarak Hazırlanmıştır

Danışman: Prof. Dr. Veysel Yılmaz

Ekim 2019

ONAY

İstatistik Anabilim Dalı Doktora öğrencisi Zeynep İlhan'ın DOKTORA tezi olarak hazırladığı "Kopula Temelli Değişken Kümeleme Tekniklerinin İncelenmesi ve Mortalite Tahmini Uygulaması" başlıklı bu çalışma, jürimizce lisansüstü yönetmeliğin ilgili maddeleri uyarınca değerlendirilerek oybirliği ile kabul edilmiştir.

Danışman **:** Prof. Dr. Veysel Yılmaz

İkinci Danışman **:** Prof. Dr. Ş.Kasırga Yıldırak

Doktora Tez Savunma Jürisi:

Üye : Prof.Dr.Veysel YILMAZ

Üye : Prof.Dr.Özlem ALPU

Üye : Prof.Dr.Zeki YILDIZ

Üye : Doç.Dr. Kadir Özgür PEKER

Üye : Doç. Dr. Erkan ARI

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun ..... tarih ve ..... sayılı kararıyla onaylanmıştır.

Prof. Dr. Hürriyet ERŞAHAN
Enstitü Müdürü

ETİK BEYAN

Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna göre, Prof.Dr.Veysel YILMAZ ve Prof.Dr.Ş.Kasırga YILDIRAK danışmanlığında hazırlamış olduğum "Kopula Temelli Değişken Kümeleme Tekniklerinin İncelenmesi ve Mortalite Tahmini Uygulaması" başlıklı DOKTORA tezimin özgün bir çalışma olduğunu; tez çalışmamın tüm aşamalarında bilimsel etik ilke ve kurallara uygun olarak elde ettiğimi; tez çalışmamda yararlandığım eserlerin tümüne atıf yaptığımı ve kaynak gösterdiğimi ve bilgi, belge ve sonuçları bilimsel etik ilke ve kurallara göre sunduğumu beyan ederim. 16/10/2019

Zeynep İLHAN

ÖZET

İstatistik biliminin önemli konularından birisi bağımlılık yapısının belirlenmesidir. Doğrusal olmayan bağımlılıklarla ilgilenen kopulalar ise bağımlılık yapılarının tespiti için kullanılan yeni ve popüler araçlardandır. Kopula temelli kümeleme tekniklerinden CoClust ve kuyruk bağımlılığı ile kümeleme teknikleri bağımlılık yapıları gösteren değişkenlerin kümelemesinde yardımcı olmaktadır. Bu tez çalışmasında, iki kümeleme tekniği aracılığıyla değişkenler arasındaki bağımlılık yapısı incelenmiştir. Böylelikle, tekniklerden elde edilen sonuçlar karşılaştırılarak, bağımlılık yapısı gösteren değişkenler birlikte değerlendirilmiştir. Bağımlı olduğu belirlenen değişkenler mortalitenin tespitinde kullanılmıştır. Mortalite tahmini, binlerce hasta hakkında bilgi içeren MIMIC veri tabanının son versiyonu olan MIMIC-III kullanılarak modellenmiştir. Mortalitenin tespiti bağımlı değişkenlerin Lojistik Regresyon Analizi ile modellenmesiyle sağlanmıştır. Belirlenen modellerin geçerlilikleri hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi ile incelenmiştir. Her iki tekniğin de anlamlı ve geçerli modeller verdiği tespit edilmiştir.

Anahtar Kelimeler: Kopula, CoClust, Kuyruk Bağımlılığı İle Kümeleme, Lojistik Regresyon Analizi, Hata Matrisi, Çapraz Geçerlilik Ölçütü, ROC Eğrisi, Mortalite Tahmini

SUMMARY

One of the important issues of statistical science is to determine the structure of dependence. Copulas dealing with nonlinear dependencies are new and popular tools used for the detection of addiction structures. CoClust and tail dependency and clustering techniques, which are copula based clustering techniques, help clustering variables showing dependency structures. In this thesis, the dependency structure between the variables is examined through two clustering techniques. Thus, the results obtained from the techniques are compared and the variables showing dependency structure are evaluated together. Then, dependent variables were used to determine mortality. Mortality estimation was modeled using MIMIC-III, the latest version of the MIMIC database, which contains information about thousands of patients. The determination of mortality is provided by modeling dependent variables with Logistic Regression Analysis. The validity of the determined models was examined with error matrix, cross validity criterion and ROC curve. Both techniques are found to provide meaningful and valid models.

Keywords: Copula, CoClust, Clustering With Tail Dependence, Logistic Regression Analysis, Error Matrix, Cross-Validation Criterion, ROC Curve, Mortality Estimation

TEŞEKKÜR

Akademik hayata ilk adımımı attığım günden itibaren bilgisi ve deneyimiyle ihtiyaç duyduğum her an bana destek olan, bu tez çalışmasında da bana sabırla yol gösteren danışmanım Prof.Dr.Veysel YILMAZ'a teşekkürü bir borç bilirim.

Kapısını çaldığım ilk günden itibaren tereddüt etmeden bana yardımcı olan, bilgi birikimini ve akademik tecrübelerini benimle paylaşarak beni cesaretlendiren, ikinci danışmanlığımı kabul ederek önemli katkı ve desteklerini sunan saygıdeğer hocam Prof.Dr.Ş.Kasırga YILDIRAK'a teşekkürlerimi sunarım.

Bu uzun süreçte desteklerini esirgemeden, güleryüzle bana yol gösteren sevgili hocalarım Prof.Dr.Özlem ALPU'ya ve Doç.Dr.K.Özgür PEKER'e teşekkür ederim.

Kendi ayakları üzerinde durabilen, güçlü bir kadın olmayı kendisinden öğrendiğim, hayattaki ilk öğretmenim olan annem Güllü İLHAN'a, varlığı bana her zaman güç kaynağı olan sevgisini ve desteğini hep üzerimde hissettiğim canım kardeşim Adil İLHAN'a teşekkür borçluyum.

Ve teşekkürün en büyüğü babam Hakkı İLHAN'a. Bana olan inancıyla bugünü görmeyi arzulayan, attığım her adımda inançla ve gururla yanımda olup bana destek olan, onurlu bir yaşam sürmeyi bana öğreten babam. Teşekkür ederim. Şimdi ve sonsuza dek benimle olduğunu biliyorum.

İÇİNDEKİLER

Sayfa

İÇİNDEKİLER (devam)

Sayfa

İÇİNDEKİLER (devam)

Sayfa

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

ÇİZELGELER DİZİNİ

Çizelge **Sayfa**

ÇİZELGELER DİZİNİ (devam)

Çizelge **Sayfa**

ÇİZELGELER DİZİNİ (devam)

Çizelge **Sayfa**

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar	Açıklama
APACHE II	Acute Physiology and Chronic Health Evaluation II
SAPS II	Simplified Acute Physiology Score II
SOFA	The Sequential Organ Failure Assessment
MIMIC	Medical Information Mart for Intensive Care
ROC	Receiver Operating Characteristic
RMD	Riske Maruz Değer
IBNR	Gerçekleşmiş Fakat Rapor Edilmemiş Hasar
ÜFE	Üretici Fiyatları Endeksi
TÜFE	Tüketici Fiyatları Endeksi
FGM	Farlie-Gumbel-Morgenstern
BIC	Bayesyen Bilgi Kriteri
AIC	Akaike Bilgi Kriteri
MIT	Massachusetts Institute Of Technology
MSVD	Eksik Değer Tekil Değer Ayrışımı
VIM	Visualization and Imputation of Missing Values
MICE	Multivariate Imputation by Chained Equations
HMISC	Harrell Miscellaneous
VIF	Variance Inflation Factor

SİMGELER VE KISALTMALAR DİZİNİ (devam)

Kısaltmalar	**Açıklama**
NHANES	National Health and Nutrition Examination Survey
VAR	Varyans
COV	Kovaryans
χ^2	Ki-Kare

1. GİRİŞ VE AMAÇ

İstatistik biliminin en temel ve en önemli konularından birisi bağımlılık yapılarının incelenmesidir. Değişkenler arasında ilişkinin olup olmadığı, varsa ilişkinin yönünün ve şiddetinin tespiti pek çok istatistiksel yöntem için oldukça önemlidir. Doğrusal Regresyon Analizi, Lojistik Regresyon Analizi gibi bağımlı değişken yapısını açıklama tekniklerinde bağımlılık yapısı kullanılmakla birlikte; Faktör Analizi ve Kümeleme Analizi gibi boyut indirgeme ve kümeleme tekniklerinin de temeli, değişkenler arası bağımlılık yapısının doğru tespitidir.

Öte yandan, literatürün yeni konularından birisi olan kopula aileleri ise değişkenler arasındaki doğrusal olmayan bağımlılığı modellemek amacıyla kullanılmaktadır. Kopulalar, parametrik olmayan bağımlılık ölçülerini geliştirmede kullanılan popüler bir yöntemdir. Böylelikle, çok değişkenli bağımlılık yapılarının araştırılmasında kullanılarak literatüre farklı bir bakış açısı getirilmektedir. Aktüerya bilimlerinde mortalitenin modellemesinden, finansta kredi değerlemesine, mühendislikte çok değişkenli süreç kontrolüne kadar pek çok konuda etkin bir şekilde kopulalar tercih edilmektedir.

Kümeleme teknikleri ise veri setinde bulunan bağımlılıklar incelenerek sınıflama yapmaktadır. Kümelerin oluşturulmasında karşılıklı ilişki durumu göz önüne alınmaktadır. İlk kez 2008 yılında önerilen CoClust Tekniği ise doğrusal bağımlılık kısıtlarının üstesinden gelerek klasik kümeleme tekniklerine alternatif olmaktadır. Bu teknik, çok değişkenli bağımlılık yapısını inceleyerek kümeleme yapmak amacıyla kopulaları kullanmaktadır. Kümeler arası çok değişkenli bağımlılığın gücü ve türü sırasıyla, bir kopula fonksiyonu ve kopulanın bağımlılık parametresi ile modellenmektedir.

Kopulaların kullanıldığı kümeleme tekniklerinden bir diğeri ise kuyruk bağımlılığı aracılığıyla kümeleme tekniğidir. İki rassal değişken arasındaki ilişkiye dayanan benzerlik ilişkisini incelemek yerine, değişkenlerin dağılım kuyruğundaki birlikte hareketleri üzerinden kümeleme tekniğidir. Değişkenlerin bu birlikte hareketi, kopulalar aracılığıyla modellenmektedir. Benzemezlik üzerine kurulan teknikte kopulalar aracılığıyla elde edilen alt ve üst kuyruk bağımlılık katsayıları oldukça önemlidir.

Çalışmanın temel amacı, bağımlılık yapılarında doğrusallık gibi kısıtları aşmaya yardımcı olan kopulaları ve kopula ailelerini inceleyerek, bunlar aracılığıyla elde edilen kümeleme tekniklerinin karşılaştırılarak incelenmesidir. Kopulalar aracılığıyla bağımlılık yapısının incelenmesi ve ilgili kümeleme tekniklerinin kullanımı ile klasik kümeleme tekniklerinin dışına çıkılması hedeflenmektedir.

Kopulaların kullanıldığı kümeleme tekniklerinden CoClust ve kuyruk bağımlılığı tekniklerinden elde edilen değişkenler kullanılarak büyük bir veri setinde farklı mortalite sonuçlarında tekniklerin karşılaştırılması sağlanmıştır. Böylelikle tekniklerin sağladığı bağımlılık sonuçlarının modellenmesi ile yeni tekniklerin literatürdeki gelişimine katkı sağlamak çalışmanın hedeflenen sonuçlarındandır. Değişkenler arasındaki bağımlılık yapılarının tespiti için verimli bir teknik önerisi getirebilmek çalışmanın amaçlarındandır.

İstatistik ve tıp alanının ortak çalışma alanlarından bir diğeri ise mortalitenin tespitidir. Bu tespit pek çok istatistiki çalışmaya konu olmakla birlikte elde edilen sonuçlar tıp alanı için önemli olduğu kadar istatistik ve sigortacılık alanları için de önemlidir.

Bir yoğun bakım hastasının hastaneye başvurusu sırasında ve bakımı süresince hastanın olası mortalitenin tespiti, ilk müdahalede ve tedavi süresince oldukça önemlidir. Gerek klinik çalışmalarda gerekse sigortacılık çalışmalarında hastaların mortalitenin tespitinin farklı durumlar için önemli olması nedeniyle, hem istatistik ve aktüerya hem de klinik literatürde bu konuyla ilgili pek çok çalışma yapılmıştır. Yapılan çalışmalar, mortalitenin oldukça yüksek olduğu yoğun bakım ünitesinde yatan hastalar üzerinde yoğunlaşmıştır. Hastanın tedaviye kabul anında ve ilerleyen süreçte, mortalitenin tespit edilerek ilerlenmesi tedavi sürecinin önemli bir aşamasıdır.

Bu kapsamda, yoğun bakım hastalarında mortaliteyi tespit etmek amacıyla kullanılan çeşitli skorlar bulunmaktadır. Bunların en sık tercih edilenleri SAPS II (Simplified Acute Physiology Score), APACHE II (Acute Physiology and Chronic Health Evaluation) ve SOFA (The Sequential Organ Failure Assessment) skorlarıdır. Sağlık çalışanları, hastanın yoğun bakım sürecinde bu skorlar üzerinden hastanın mortaliteyi takip etmektedir.

Literatürde mortalite tahmininde kullanılan değişkenler ilgili skorlardan yola çıkılarak seçilebildiği gibi, araştırmacının tecrübesine dayanarak belirlenmiştir. Yoğun bakım hastalarında kullanılan skorlarda ise fizyolojik değişkenler kullanılırken hayati

değişkenler göz ardı edilmektedir. Öte yandan, mortalite tahmini için sıklıkla kullanılan SOFA skoru ise organ yetmezliği riskine odaklanmaktadır. APACHE II ve SAPS II skorları ise akut fizyolojik skorlar olarak albümin, hemoglobin gibi hayati değişkenleri yine göz ardı etmektedir. Skorlar hasta üzerinde belirli durumlara odaklanmışken, kullanım alanları yoğun bakıma ulaşan tüm hastalar olmak üzere oldukça geneldir. Dolayısıyla hayati değişkenlerin de dâhil olduğu ve bu genel yaklaşıma uygun olarak kullanılabilecek bir mortalite tahmin modeli hastaların doğru değerlendirilmesi için oldukça önemlidir.

Ancak literatürde yapılan mortalite tahmini çalışmalarında, skorların kullanımı haricinde görece küçük veri setlerinin kullanımı önemli sorunlardan birisini oluşturmaktadır. Yapılan tahminler isabetli olsa da, mortalite tahminlerinde fark yaratacak bir ilerlemeye neden olamamaktadır. Bu nedenle, büyük bir veri setinin kullanımının literatürde yaratacağı fark dikkat çekmiştir.

Kapsamlı ve görece büyük bir veri setinin incelenecek modeller için verimli bir sonuç sağlayacağı düşünülmektedir. Bu nedenle, hem fazla sayıda hastaya sahip olması hem de hastalara dair demografik bilgiler, laboratuvar sonuçları, görüntüleme sonuçları gibi pek çok bilgiyi içeren MIMIC (Medical Information Mart for Intensive Care) veri tabanının kullanımının yapılacak çalışmaya verimli bir ortam sağlayacağı düşünülmektedir. Dünya çapındaki araştırmacılar için ücretsiz ulaşılabilir olması ve geniş hasta popülasyonu nedeniyle son dönemde veri madenciliği, sağlık alanlarında sıklıkla kullanılmaktadır.

Literatürde yürütülen çalışmalarda, MIMIC veri tabanında ve diğer veri setleriyle yapılan mortalite tahmini çalışmalarında Lojistik Regresyon Analizi dikkat çekmektedir. İkili sonuç vermesi nedeniyle ve ilgilenilen risk açısından verimli bir yöntemdir. Ancak incelenen çalışmalarda, yapılacak analiz için özel bir değişken seçimi yoluna gidilmemiştir. Kullanılan değişkenler genellikle üzerinde çalışılan veri setinin sağladığı olanaklar çerçevesinde seçilmiştir. Bu durum, çalışma kapsamında veri seti kaynaklı dışarıdan bir kısıt getirmektedir.

Regresyon analizi temelinde, değişkenlerin birbirleriyle doğrudan ilişkili olması beklenmese de birbirleriyle anlamlı ilişkiler kurulabiliyor olması önemlidir. Bu nedenle, bu tez çalışmasında birbirleriyle ilişkili olabilecek, kendi içinde homojen değişkenlerin birlikte incelenmesi ile mortalite üzerinde hedefe daha uygun sonuçlar elde edilmesi

hedeflenmiştir. Değişkenlerin homojen kümelerde incelenmesi aşamasında doğrusallık ve parametrik bağımlılık gibi kısıtları aşma amacıyla kopulalardan yararlanılmıştır. Bu bağlamda klasik kümeleme ve bağımlılık ölçütlerinin doğrusallık ve parametrik yaklaşımlarına alternatif bir yoldan ilerlenmiş olacaktır.

Yoğun bakım ünitelerinde kullanılan skorlarda kullanılan ortak değişkenler olmakla birlikte her skoru diğerlerinden ayıran farklı değişkenler de bulunmaktadır. Yoğun bakım hastalarına ait bu değişkenler doğrudan mortalitenin tespitinde kullanılmaktadır. Bu tez çalışmasında, mortaliteyi verecek kendi içinde homojen değişkenlerin bir araya getirilmesi hedeflenmiştir. Bu değişkenlerin kümelemesini klasik kümeleme teknikleri yerine, literatürde bağımlılık konusunda dikkat çeken bir konu olan kopulalar aracılığıyla gerçekleştirilmesi, tezin önemli amaçlarından bir başkasıdır.

Kopulaların kullanıldığı kümeleme tekniklerinin en yenisi olan CoClust kümeleme tekniğinin öne sürüldüğü ilk çalışma da yine klinik bir çalışmadır. Kuyruk bağımlılığı aracılığıyla yapılan kümeleme tekniği, literatürde nispeten daha eski bir yöntem olmakla birlikte, bu yöntem ile CoClust'ın birlikte incelenmesi iki kümeleme yöntemini karşılaştırma olanağı da sağlayacaktır. Böylelikle, klasik kümeleme yöntemlerinin dışına çıkılarak bağımlılık yapısının kopulalarla belirlendiği iki farklı kümeleme tekniği karşılaştırılarak incelenmiş olacaktır.

Çalışmanın bulgular bölümünde elde edilen kümelerde bulunan değişkenler aracılığıyla mortalitenin tahmini ise literatürde sıklıkla tercih edilen Lojistik Regresyon Analizi ile yapılmıştır. CoClust ve kuyruk bağımlılığı ile elde edilmiş kümelerde bulunan değişkenler bağımsız değişkenler olarak kullanılarak mortalite Lojistik Regresyon Analizi ile modellenmiştir.

Buradan yola çıkarak, öncelikle MIMIC-III gibi geniş bir veri setiyle çalışmak özellikle sağlık verilerinde ortaya çıkan küçük veri seti sorununun üstesinden gelmeye yardımcı olacaktır. Mortalite tahmininde geniş veri setiyle çalışarak elde edilen modellerin tahmin gücünün artırılması hedeflenmiştir. İlgili veri setinde bulunan değişkenlerin seçimi ile ilgili doğrudan müdahalede bulunmayarak mortalite tespitinde kullanılan skorlardan yola çıkılmıştır. Bu skorların hesaplanmasında kullanılan değişkenlerin kullanımıyla insan faktörünün neden olabileceği hatalar bertaraf edilmesi hedeflenmiştir.

Modellerde kullanılacak değişkenlerden oluşturulan havuzdan hangi değişkenlerin kullanılacağı ise değişkenler arasındaki ilişki gözetilerek belirlenmiştir. Bu ilişkinin, çoklu bağlantı sorunu olmaksızın, değişkenlerde homojenliğin yakalanması ile ortaya konulması hedeflenmiştir. Değişkenler arasındaki bağımlılık yapısının incelenmesinde, doğrusallık ve parametrik yaklaşım beklentisi olan klasik bağımlılık modellerinin üstesinden gelmek için de kopulalar tercih edilmiştir. Bu nedenle, kopulalar aracılığıyla kümeleme tekniği kullanılarak değişkenler arası homojenliğin yakalanması beklenmiştir. Bu kapsamda, kopulalar aracılığıyla kümeleme teknikleri olan CoClust ve kuyruk bağımlılığı teknikleri ile kümeleme yapılmıştır. Böylelikle, hem kopulaların efektif kullanımı hem de kümeleme tekniklerinin ilgili modellemedeki verimlilikleri karşılaştırılmıştır. Bu iki tekniğin karşılaştırılması ile literatürde yeni olan CoClust tekniği görece daha eski olan kuyruk bağımlılığı yöntemi ile birlikte değerlendirilmiş olacaktır.

Kümelerde elde edilen değişkenler aracılığıyla mortalite modellenmiştir. Belirlenen modellerin ise APACHE II, SAPS II ve SOFA skorlarının belirli durumları kapsaması sorununu aşması beklenmektedir. Yoğun bakıma ulaşan hastalara hayati değişkenleri de kapsayacak şekilde uygulanabilecek model ya da modellerin geliştirilmesi çalışmanın bir başka amacıdır.

Belirlenen tüm modellerin yalnızca anlamlılık ve uygunluk bakımından yeterli olmasının modelin geçerliliği bakımından, yeterli olacağı düşünülmemiştir. Anlamlı ve uygun modellerin geçerlilik ve güvenirlikleri hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi ile belirlenmiştir. Buradan elde edilen sonuçlara göre en iyi modelin seçilmesi çalışmanın ana hedeflerinden birisidir.

Bu kapsamda birinci bölümde, çalışmaya giriş yapılarak çalışmanın amacı açıklanmıştır.

Çalışmanın ikinci bölümünde, gerek veri tabanı gerekse kullanılacak teknikler ile ilgili detaylı bir literatür taraması yapılmıştır.

Üçüncü bölümde, kopula aileleri ve bunlara dayalı kümeleme teknikleri, Lojistik Regresyon Analizi ve model geçerlilik değerlendirme ölçütleri açıklanmıştır.

Dördüncü bölümde, açıklanan teknik yöntemler MIMIC-III veri tabanından elde edilen hasta popülasyonuna uygulanarak gerekli kümeleme ve modelleme sonuçları elde edilerek yorumlanmıştır.

Sonuç ve öneriler başlıklı beşinci bölümde ise çalışma boyunca hedeflenenler ile elde edilen sonuçlar yorumlanarak literatüre katkısı tartışılmış ve ilerleyen çalışmalar için önerilerde bulunulmuştur.

2. LİTERATÜR ARAŞTIRMASI

Çalışmanın temelini oluşturan kopulalar, ilk kez 1959'da Abe Sklar tarafından kullanılmasına rağmen aktif ve yoğun kullanımı 2000'li yıllarda başlamıştır.

Sigortacılık alanında ilk kullanımlarından birisi Albers (1999)'in makalesinde karşımıza çıkmaktadır. Toplam hasar fazlası primleri, genellikle, sigortalıların yaşamlarının bağımsız olduğu varsayımı altında hesaplanmaktadır ancak, bu konuda bağımsızlık varsayımı gerçekçi değildir. Bu çalışmada, Edgeworth genişlemelerini kullanarak, bağımlılık parametrelerinin hangi biçimlerinin stop-loss primlerinde önemli sapmalara neden olabileceği açıklanmaktadır.

Kopula modellerinin finans alanındaki öncü çalışması ise Embrechts vd. (1999) tarafından yapılmıştır. Küresel ve eliptik dağılımlar üzerinden doğrusal korelasyon tartışılarak kopula uygulamaları yapılmıştır.

Bir sigorta portföyünde önemli olan bir başka konu ise, hasarın ortaya çıkışı ile hasar büyüklüğünün bağımsızlığıdır. Song (2000) çalışmasında Poisson, Normal, Gamma gibi modellerdeki bağımlılık yapısı için önerilen Gaussian kopula ailesinini kullanmıştır. Hasar sayıları ve hasar büyüklerinin modellemesi Czado vd. (2012) tarafından da değerlendirilmiştir. Parametre tahminleri Maximization by Parts'ın uyarlanmış bir sürümü ile yapılmıştır. Tahmin yönteminin performansı benzetim çalışmasıyla doğrulanmıştır.

Hasar büyüklüğü ve hasarların ortaya çıkışı arasındaki bağımlılık yapısını inceleyen üç farklı çalışma Cossette vd. tarafından 2002, 2008 ve 2009'da Cook–Johnson ve Farlie-Gumbel-Morgenstern kopula aileleri kullanılarak yapılmıştır. Heilpern (2014) ise üstel hasar büyüklükleri için birleşik Poisson risk modeli için iflas olasılığı hesaplamalarında hasar büyüklüğü ve hasarların ortaya çıkışı arasındaki bağımlılık yapısını Spearman kopula yöntemiyle irdelemiştir. Bu konudaki çalışmalar aynı ve farklı kopula aileleri ile tekrarlanarak sürmüştür.

2006 yılında Wüthrich, alandaki önemli çalışmalardan birisini yaparak, Uç Değer Teorisini ünlü Fisher-Tippett Teoremi ile bağımlı rassal değişkenlerle kopulalar

aracılığıyla genelleştirmiştir. Böylelikle, bağımlı rassal dizilerin davranışını incelemek mümkün olmuştur.

Nikoloulopoulos ve Karlis (2009) tarafından çok değişkenli kesikli veri modellemesinde kullanılabilen ve çözülmesinde esnek bir bağımlılık yapısı sunan yeni bir kopula ailesi önerilmiştir. Önerilen yeni kopula ailesinin özellikleri altı epidemik hastalık arasındaki bağımlılık üzerinden incelenmiş ve uygulanabilirliğini göstermek için somut bir uygulama yapılmıştır.

Şirketlerin yaşam süresini etkileyen önemli konulardan birisi uygun şekilde rezerv tutulmasıdır. Gerçekleşmiş fakat rapor edilmemiş hasar (IBNR) karşılığı tutulan rezerv tahmini, sigorta şirketlerinin diğer yükümlülükleri açısından önemli bir başlıktır. Son zamanlarda, toplu hasar kayıp modellerinin eksiklerinin üstesinden gelen bireysel hasar kayıp modelleri aktüeryal literatürde büyük ilgi görmektedir. Zhaoa ve Zhouc (2010) çalışmasında, bireysel hasar kayıp modellerinde olay zamanlarının bağımlılık yapısına uyması için yarı rekabetçi risk kopulası ve yarı sağkalım kopulası kullanımını önermektedir.

Genel sigortacılıkta, sigorta hasarları modelleme ve saf prim tespiti için kullanışlı bir araç olan Tweedie dağılımı kopula çalışmalarının da ilgi odağı olmuştur. Shi (2016), Massachusetts otomobil sigortası poliçelerinden oluşan portföyde, hasar modellemesine kopula temelli çok değişkenli Tweedie regresyonu önermiştir.

Literatür özetinde görüldüğü üzere kopulalar bilim dünyasının son 20 yılda ilgi odaklarından birisi olmuştur ve bu ilgi geliştirilerek sürdürülmektedir. Alan uygulamalarında görülen eksikler kolektif olarak doldurulmaktadır.

Ülkemizde de son 15 yılda dikkati çeken kopulalar üzerine çalışmalar hızla sürmektedir. Ancak yine de uygulama alanı oldukça geniş olan bu bağımlılık yapısı inceleme aracı, pek çok alanda kendisine yer bulabilmektedir.

Türkiye'de kopulalar kullanılarak hazırlanan ilk tez çalışması Karadağ (2003) tarafından hazırlanmıştır. Hisse senedi portföylerinin risklerini hesaplamak amacıyla kopulalar kullanılmıştır. Çalışmada New York Hisse Senedi Borsası'ndan 15 adet hisse senedi seçilerek oluşturulan portföyler oluşturulmuş, getirilere en uygun kopula olan t-kopula uygulamada tercih edilmiştir.

Ülkemiz literatüründe ikinci tez çalışmasında Özbakış (2006), İMKB ve São Paulo Borsası arasındaki bağımlılık yapısını kopula tahmin yöntemiyle incelemiştir ve bu yapıya en uygun kopula ailesinin Ali-Mikhail-Haq kopula ailesi olduğunu belirlemiştir.

İlerleyen zamanlarda finans ve ekonomi alanında kopula çalışmaları Büyükyılmaz (2011) ve Avutman (2011) tarafından yapılmıştır. Büyükyılmaz (2011) Üretici Fiyatları Endeksi (ÜFE) ve Tüketici Fiyatları Endeksi (TÜFE) arasındaki bağımlılık yapısını Ali-Mikhail-Haq, Clayton, Frank, Gumbel Hougaard Kopula ailelerini kullanarak incelerken; Avutman (2011) Finansbank'a ait iki tip ve iki farklı stratejide oluşturulan yatırım fonunun arasındaki bağımlılık yapısı Clayton, Gumbel, Frank kopula aileleri aracılığıyla gözlemiştir.

Sigortacılık ve risk analizi konusunda ilk kopula çalışması ise Kızılok (2010) tarafından Riske Maruz Değer (RMD) ve Koşullu Riske Maruz Değer konuları üzerinde yapılmıştır. Diğer RMD çalışmalarından farkla risklerin bağımlı olduğu durum değerlendirilmiştir. Bu bağımlılık yapısı ise Farlie-Gumbel-Morgenstern (FGM) kopula ailesi ile değerlendirilmiştir. Bu çalışmada uygulama benzetim çalışması ile gerçekleştirilmiştir.

Hayat ve hayat dışı sigortalar alanında ise Sarıdaş (2012) ve Karagül (2013)'ün çalışmaları bulunmaktadır. Sarıdaş (2012)'ın çalışmasında evli çiftlerde gelecek yaşam sürelerinin bağımsız kabul edilmesinden doğan yüksek ve düşük fiyatlandırma durumlarının önüne geçme amacıyla Frank kopula fonksiyonu ile bağımlılık yapısı incelenmiştir.

Karagül (2013) ise tez çalışmasında kopula aileleri aracılığıyla iki benzetim çalışması yapmıştır. Birinci benzetim çalışmasında yatırımlar arasında, hasarlar arasında ve yatırımlarla hasarlar arasında bağımlılığın olduğu dört boyutlu bir bağımlılık yapısı üzerinde çalışılırken; ikinci benzetim çalışmasında ise iki farklı sigorta dalı arasında bağımlılık yapısı incelenmiştir.

Sigortacılık ve finans alanı için de oldukça önemli olan klinik çalışmalardaki kopula uygulamaları ise dikkat çekmektedir. Çünkü klinik alanlardaki ölüm ve uzun yaşam riski özellikle fiyatlama konularında sigortacılık sektörünü doğrudan etkilemektedir. Bir sağlık çalışanı için hastanın yaşaması veya ölmesi ne kadar önemliyse, sigorta sektörü çalışanı için de finansal açıdan oldukça önemli olmaktadır.

Kopulalar klinik uygulamalarda da sıklıkla tercih edilen araçlardan birisi olagelmiştir. Kumar ve Shoukri (2007) herhangi bir, çok değişkenli klinik uygulamasında risk tahmini konusunda kopula temelli tahmin yöntemleri doğrusal korelasyon temelli klasik yöntemlerle karşılaştırmıştır. Bu uygulama aort yetersizliği ve düzeltici operasyonlar üzerinde Arşimedyan kopula aileleri aracılığıyla gerçekleştirilmiştir. Farklı marjinal dağılımlarla simüle edilerek elde edilen verilerde kopulaların avantajları açıklanmıştır.

Sinir sistemi görüntüleme yöntemlerinin incelemesinde ise Ince vd. (2017) Gaussian (Normal) kopula uygulaması ile kapalı form çözümü geliştirmişlerdir. Elde edilen sonuç, davranışsal ve beyin tepkilerini karşılaştırma yöntemi için verimli, esnek ve sağlam bir, çok değişkenli istatistiksel çerçeve olarak yorumlanmıştır.

Klinik çalışmalarla birlikte, mortalite ve uzun ömürlülük riski tahminlerinde de kopulalara ihtiyaç duyulmuştur. Chen vd. (2015), çoklu popülasyonlarda eş hareketlerin modellenmesi konusuna vurgu yaparak zaman serileri analizi ve faktör kopula yaklaşımı ile iki aşamalı bir yöntem önermişlerdir. Artık riskleri, çok boyutlu veride kullanımı uygun olması nedeniyle tek faktörlü kopula ile modellenmiştir. Buradan yola çıkarak mortalite modeli ve maksimum entropi ile mortalite risk fiyatlaması açıklanmıştır.

İstatistik alanında çalışmaların pek çoğunun temelini bağımlılık yapıları oluşturmaktadır. Özellikle kümeleme ve gruplama tekniklerinde bağımlılık yapılarını doğru şekilde çözümlemek oldukça önemlidir. Bu tekniklerde kullanılan klasik bağımlılık ölçütleri yerine kopulalar aracılığıyla bağımlılık yapısı incelemesi yapabilmek alana farklı bir bakış açısı katmıştır.

CoClust Tekniği ise kümeleme çalışmalarına kopula yaklaşımı kazandıran yine oldukça yeni bir uygulamadır. Tekniği literatüre ilk kazandıran çalışma Lascio (2008)'nun doktora tez çalışmasıdır. Teknik, klinik mikro dizi veri analizi üzerinde geliştirilmiştir. Çok değişkenli veride kullanılan kümeleme tekniklerine kopula kullanımıyla farklı bir yaklaşım getirmiştir. Çalışma aynı zamanda genler arasındaki ilişkinin kümelemesinin literatürdeki eksikliğini vurgulamaktadır. Kopula fonksiyonlarının çok değişkenli bağımlılık yapılarında etkili olması nedeniyle bağımlılık modellemesinde tercih edilmiştir. Bu çalışmada, tekniğin teorik ve uygulama adımları detaylı şekilde açıklanmıştır, R paketi geliştirilerek gerçek bir mikrodizi veri setine uygulanmıştır.

Lascio ve Giannerini (2019), tekniğin geliştirilmiş sürümünü yaptıkları çalışma ile sunmuşlardır. Monte Carlo benzetimi ile eski sürüm ile yeni sürüm arasındaki gelişim incelenmiştir ve yeni sürümün özellikle Gaussian dağılımlarda olmak üzere daha verimli olduğu belirtilmiştir. Uygulanan üç senaryoda elde edilen sonuç tatmin edici olarak ifade edilmiştir. Son olarak elde edilen son sürüm, meme tümörü ile gerçek veri seti üzerine ikinci bir uygulama olarak gerçekleştirilmiştir.

Kırk Avrupa ülkesinin kırk yıllık gözlemlerini, Avrupa sağlıklı beslenme kurallarına göre gelişimini Lascio ve Disegna (2017) CoClust ile incelemeyi hedeflemiştir. Elde edilen sonuçlara göre Orta ve Doğu Avrupa ülkeleri sağlıksız beslenme eğiliminde kümelenirken, geri kalan Avrupa ülkelerinin ulusal politikalara uygun şekilde sağlıklı ve dengeli beslenen kümelerde yer aldığı tespit edilmiştir.

Chessa vd. (2014) ise CoClust'a oldukça benzer kopula temelli kümeleme tekniği önermişlerdir. Bu yöntemin CoClust'tan dikkat çekici farkı ise yalnızca ikili matrislere uygulanabiliyor olmasıdır. Kopula temelli kümeleme uygulamalarının dikkat çekici olduğu da aşikârdır.

Bu tez çalışmasında, veri kısıtını aşmak için yaklaşık 10 yılda binlerce yoğun bakım ünitesi hastasından elde edilen laboratuvar ölçümlerinden, görüntüleme bilgilerine dair pek çok değişkene sahip MIMIC-III (Medical Information Mart for Intensive Care III) veri tabanı tercih edilmiştir. Böylelikle, varsayımsal veri üzerinden değil doğrudan hasta bilgileri üzerinden mortalite tahmini yapılmıştır.

Literatür incelemesine göre, mortalitenin tespiti de sıklıkla incelenen ve kopulaların kullanıldığı bir konudur. Hastaların gerek hastaneye başvuru aşamasında gerek de tedavi süreleri boyunca mortalite tahminleri oldukça önemlidir. Alan çalışmalarında, bu tahminin tespiti için Lojistik Regresyon Analizi sıklıkla tercih edilse de kullanılacak bağımsız değişkenlerin tespiti konusunda pek çok çalışmada özel bir yol izlenmemiştir. Mortalitenin belirlenmesinde en önemli eksikliklerden birisi budur. Kullanılan değişkenler eldeki veri seti üzerinden tercih edilerek ilerlenmiştir. Ancak, mortalitenin ortaya çıkmasında doğrudan etkili olabilecek değişkenlerin seçimi oldukça önemlidir. Yine kullanılacak değişkenlerin seçiminde birbiriyle ilişkili olmaları da önemli bir detaydır.

Başvuran hastalarda, mortalite konusunda, hekimlere yol göstermesi için kullanılan skorlarda kullanılan değişkenler bu çalışmada değişken seçimi konusunda çıkış noktasıdır.

Bu değişkenleri çoklu bağlantı sorunu olmaksızın bir araya getirerek birbirleriyle ilişkili değişkenleri kullanarak mortalite olasılığını net olarak ortaya koymak hedeflenmiştir.

Bu kapsamda, Liang ve Zeger (1993) regresyon analizinde kullanılacak bağımsız değişkenler arasında korelasyon olmaması özelliğini açıklayarak bunun özellikle tıp alanındaki çalışmalarda her zaman anlamlı olamayacağına dikkat çekmiştir. Sağlık alanında değişkenlerin yapısı gereği birbiriyle ilişkili olacağını ve bunun kaçınılmaz olduğunu vurgulamıştır. Bu nedenle, değişkenler arasında çoklu bağlantı sorununun olmamasının önemini belirterek değişkenler arasındaki ilişkinin kaçınılmaz ve aksine tıp alanında ilişkilerin incelenmesinin daha verimli sonuçlar vereceğini belirtmiştir. Doğrudan kümeleme yoluna gidilmese de değişkenler arasındaki ilişki gözetilerek halk sağlığı üzerine Lojistik Regresyon Analizi çalışılmıştır.

Hanley vd. (2003) ise kümelenmiş ilişkili veri setinin incelenmesinde bir regresyon tekniği olan genel tahmin eşitliklerini önermekle birlikte, geniş veri setlerinde yapılan kümeleme ile Lojistik Regresyon Analizi'nin de etkin sonuç vereceğini ileri sürmüştür.

Bühlmann vd. (2013) kanonik korelasyonları kullanarak hiyerarşik kümeleme ile değişkenleri kümeleme ile çalışmanın ilk adımını atmaktadır. Birbiriyle ilişkili değişkenlerin regresyon modellerindeki etkisini araştırmıştır. Değişken seçiminde bağımlıkların etkisini vurgulayarak seçim yaparken kullanılabilecek kümeleme teknikleri önerisinde bulunmuştur. Çalışmada, Kendall (1957), Hastie vd. (2000), Dettling ve Bühlmann (2004) ve Bondell ve Reich (2008)'e atıf yaparak Temel Bileşenler Regresyonunda da değişken seçiminde yüksek korelasyonun dikkate alındığını ve değişkenlerin kümeleme tekniği ile seçildiğini belirtmiştir. Regresyon analizinde elde edilen bağımlı değişken ile de bağımlılık ile elde edilen değişkenlerin denetlenebildiğini belirtmiştir.

Hastie vd. (2000)'nin yürüttüğü çalışmada DNA'dan elde edilen geniş bir veri setinde bulunan genleri kümelemek için hiyerarşik kümelemeye alternatif 'gen traşlama' isimli yeni bir kümeleme tekniği önermiştir. Bu teknik ile elde edilen kümelerin bir sonuç çıktısıyla ölçülebileceğini belirtmiştir. Burada da sonuç çıktısı olarak hastaların yaşam durumu kümelenen genler aracılığıyla Cox regresyon aracılığıyla modellenmiştir.

Thompson (2009) değişken seçiminin modellemenin önemli bir adımı olduğunu vurgulamıştır. Sonuç değişkenindeki değişiklikleri en iyi açıklayan değişkenlerin seçiminin

önemini açıklayarak Bootstrap, Ridge regresyon, Lasso ve Bayesyen model ortalamaları ile değişkenler arasındaki bağımlılık yapısını inceleyerek koroner kalp hastalarında iki sonuçlu durum için elde edilen sonuçları değerlendirmiştir. 20 diyet değişkeninin koroner kalp hastalığı üzerindeki etkisi çalışmada incelenmiştir. İkili sonuç değişkeni için ise Lojistik Regresyon Analizi kullanılmıştır.

Bununla birlikte, 2000'li yılların başında MIMIC veri tabanı gerek veri madenciliği bakımından gerek klinik araştırmalar bakımından gerekse de istatistiksel çalışmalar açısından dikkat çekmeye başlamıştır. Sunduğu kapsamlı demografik bilgiler laboratuvar sonuçları, görüntüleme sonuçları gibi değişkenler açısından araştırmacılar için oldukça verimli bir kaynaktır. Bu ilgi, MIMIC-II çalışmalarıyla artmışken MIMIC-III ile daha da dikkat çekici olmuştur.

Özellikle ölüm ve yaşam riski için çalışılan veri seti önem kazanmaktadır. Elde edilecek sonuçların geçerliği ve güvenirliği açısından kapsamlı bir veri setiyle çalışmak oldukça önemlidir. MIMIC veri tabanı, bu konuda araştırmacılara büyük alan tanımaktadır.

Mandelbaum vd. (2013), MIMIC-II veri tabanından elde edilen 14526 hasta üzerinde yapılan çalışmada çok değişkenli lojistik regresyon ile kreatinin ve üre çıkışı gibi böbrek fonksiyonlarının mortalite ile ilişkisi incelemiş ve yüksek kreatinin değerinin mortaliteyi artırdığı tespit etmişlerdir.

Danziger vd. (2016), yine MIMIC-II veri tabanından alınan yoğun bakım hastaları üzerinde yapılan çalışmada obezite ve akut böbrek hasarı değişkenleri arasındaki ilişki ile bu değişkenlerin mortalite ile arasındaki etki durumu lojistik regresyon ve Cox regresyon analizi ile incelemiştir. Obezitenin mortalite ile arasındaki ilişki gösterilmiştir.

Marshall vd. (2017) ise yaptıkları çalışmada, serum sodyumun yoğun bakıma başvuran hastalarda sık görülen bir elektrolit sorunu olduğunu vurgulayarak, yoğun bakım hastalarında disnatremi ve serum sodyum dalgalanması ile mortaliteyi incelemeyi hedeflemiştir. Bu araştırma tek ve çok değişkenli lojistik regresyonu ile yürütülmüştür. Sodyum değerleri ile mortalite arasındaki ilişki anlamlı olarak tespit edilmiştir.

2002-2012 yılları arasında yoğun bakımda tedavi görmüş hastalar arasından seçilmiş 8429 hasta ile yapılan çalışmada Stretch vd. (2018) yoğun bakımda tedavi görmüş durumu kritik hastalara odaklanmıştır. Yarı parametrik iki değişkenli probit tahmini

kullanmışlardır. Eğilim puanı uyumu ve lojistik regresyon sağlamlık kontrolü için kullanılmıştır.

Serpa Neto vd. (2018), mekanik solunum ile 30 günde ölüm, yoğun bakımda ölüm ve hastanede ölüm ile ilişkisinin önemini vurgulayarak mortaliteyi lojistik regresyon ile incelerken, Feng vd. (2018) MIMIC-III veritabanı aracılığıyla transtorasik ekokardiyografinin yoğun bakım ünitelerinde kullanımı üzerine yoğunlaşmıştır. Popülasyonda 28 günde ölüm ile transtorasik ekokardiyografinin ilişkisi çok değişkenli regresyon ve ikiye katlamalı sağlam regresyon gibi istatistik yöntemlerle incelemiştir.

Değişkenlerin bağımlılıklarının incelenmesi konusunda ise klasik kümeleme tekniklerinin haricinde literatürde son yıllarda sıklıkla kullanılan kopulalar tercih edilmiştir. Kopulaların verdiği bağımlılık ölçüsü temelinde yapılan ve literatürde nispeten eski bir yöntem olan kümeleme tekniklerinden kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniği ve yeni bir yöntem olan CoClust Tekniği kullanılmıştır. Böylece iki tekniğin de avantajları kullanılarak bir yaklaşım geliştirilmiştir.

Elde edilen kümelerde bulunan değişkenler Lojistik Regresyon Analizi aracılığıyla modellenerek mortalite tahmin edilmiştir. Tahmin edilen modellerin geçerlilik ve güvenirliklerinin tespiti ise hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi kullanılarak belirlenmiştir.

3. MATERYAL VE YÖNTEM

Tez çalışması kapsamında kullanılacak ana yöntemler kopula ailelerine dayanan kümeleme teknikleri ve Lojistik Regresyon Analizidir. Öncelikle bu yöntemler olmak üzere, bağımlılık ölçütleri, kopula aileleri, kopulalar aracılığıyla kümeleme teknikleri, hata matrisi, çapraz geçerlilik ölçütü ve ROC (Receiver Operating Characteristic) eğrisi tanıtılarak uygulama için gerekli teorik alt yapı sunulmuştur.

3.1. Bağımlılık Ölçütleri

İki ölçüm arasında bir ilişki olup olmadığını, varsa bu ilişkinin yönünü ve şiddetini belirlemek istatistiksel olarak önemli bir adımdır. Bağımlılık ölçüsü negatif ise iki değişken arasında ters yönde ilişki vardır, yani "değişkenlerden birinin değeri artarken diğerinin değeri de azalmaktadır"; pozitif ise "değişkenlerden birinin değeri artarken diğerinin de değeri artmaktadır veya tam tersi birinin değeri azalırken diğerinin değeri de azalmaktadır" yorumu yapılabilir. Bağımlılık ölçütlerinden doğrusal korelasyon katsayısı, sıra korelasyon katsayısı ve kuyruk bağımlılığı katsayısı bu bölümde tanıtılmıştır.

3.1.1. Doğrusal korelasyon katsayısı (Pearson korelasyon katsayısı)

Pearson korelasyon katsayısı olarak da bilinen doğrusal korelasyon katsayısı ρ, en çok bilinen ve uygulamalarda sıklıkla kullanılan bir bağımlılık ölçüsüdür. Bu bağımlılık ölçüsü değişkenler arasındaki ilişkinin doğrusal olduğunu ve değişkenlerin normal dağılıma uygun olduğu varsayımlarına dayanmaktadır.

X ve Y rassal değişkenleri ele alındığında, $Var(X)$ ve $Var(Y)$ değerleri değişkenlerin varyansını, $Cov(X,Y)$ ise değişkenlerin kovaryansını ifade etmektedir. Doğrusal Korelasyon Katsayısı Eşitlik 3.1'deki gibi ifade edilebilir.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{3.1}$$

$Cov(X,Y)$ değeri ise $Cov(X,Y)=E[XY]-E[X]E[Y]$ eşitliği ile hesaplanmaktadır.

Örneklem doğrusal korelasyon katsayısı kopulalar aracılığıyla Eşitlik 3.2'deki gibi ifade edilebilir. Bu eşitlikte F_X ve F_Y sırasıyla X ve Y rassal değişkenlerinin dağılım fonksiyonlarını, C ise rassal değişkenlerin kopula fonksiyonunu göstermektedir ve $u=F(x)$ ile $v=F(y)$ eşitlikleri sağlanmaktadır (Schweizer ve Wolff, 1981).

$$r = \frac{1}{\sqrt{Var(X)Var(Y)}} \int_0^1 \int_0^1 [C(u,v) - uv] dF_X^{-1}(u) dF_Y^{-1}(v) \tag{3.2}$$

Doğrusal korelasyon katsayısı ρ'nun özellikleri aşağıdaki gibidir.

- [-1,1] aralığında değer almaktadır; $-1 \le \rho(X,Y) \le 1$.
- Değişkenlerin monoton değişimi altında doğrusal korelasyon katsayısının sabit olması durumu ise Eşitlik 3.3'te görülmektedir.

$$\left.\begin{array}{l} X' = aX + b \\ Y' = cY + d \end{array}\right\} \Rightarrow \rho(X',Y') = sign(ac)\rho(X,Y) \tag{3.3}$$

Bağımlılık yapısının incelenmesinde doğrusallık varsayımı söz konusu olduğunda doğrusal korelasyon katsayısı tercih edilirken, ikili sıralama durumları veya normal dağılıma uygunluk varsayımının sağlanamaması durumlarında diğer bağımlılık katsayıları kullanılmaktadır.

3.1.2. Sıra korelasyon katsayısı

Sıra korelasyon katsayısı, farklı değişkenlerin sıralaması veya aynı değişkenin farklı sıralaması gibi, bir sıralı ilişkiyi ölçen istatistik ölçüsüdür. İki sıralama arasındaki benzerliğin derecesini ölçer ve aralarındaki ilişkinin önemini değerlendirmek için kullanılmaktadır. Bu bölümde, sıra korelasyon katsayılarından Kendall'ın Tau katsayısı ve Spearman'ın Rho katsayısı incelenecektir.

3.1.2.1 <u>Kendall'ın Tau katsayısı</u>

$\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ sürekli rassal değişkenlerin bir vektörü olan *(X,Y)*'den alınmış n tane gözlemi ifade etmektedir. Burada, $\binom{n}{2}$ kadar farklı sayıda uyumlu veya

uyumsuz gözlem çifti vardır. Bunların bir kısmı uyumlu, bir kısmı ise uyumsuz çiftlerdir. *c* tane uyumlu, *d* tane de uyumsuz çift olduğu varsayımıyla örneklem için Kendall Tau Eşitlik 3.4'deki gibi ifade edilebilir. *c* uyumlu çift sayısı $[(x_i > x_j)$ ve $(y_i > y_j)]$ veya $[(x_i < x_j)$ ve $(y_i < y_j)]$ durumunu sağlayan her bir çift için +1 eklenmesiyle; *d* uyumsuz çift sayısı $[(x_i > x_j)$ ve $(y_i < y_j)]$ veya $[(x_i < x_j)$ ve $(y_i > y_j)]$ durumunu sağlayan her bir çift için -1 eklenmesiyle belirlenir (Kruskal, 1958; Hollander ve Wolfe, 1973; Lehmann, 1975).

$$t = \frac{c-d}{c+d} = \frac{c-d}{\binom{n}{2}} \tag{3.4}$$

Eşitlik 3.4'de uyumlu ve uyumsuz çift sayısı aracılığıyla elde edilen *t* değeri, örneklemden rassal olarak seçilen (x_i,y_i) ve (x_j,y_j) gözlem çiftleri bakımından uyum ve uyumsuzluk olasılıkları kullanılarak da elde edilebilir. Buradan yola çıkarak Kendall'ın Tau katsayısı anakitle açısından ortak dağılım fonksiyonu *H* olan (X,Y) için benzer şekilde ifade edilebilir. *H* ortak dağılım fonksiyonuna sahip (X_1,Y_1) ve (X_2,Y_2) bağımsız ve aynı dağılıma sahip rassal değişkenlerdir. Anakitle için Kendall'ın Tau katsayısı Eşitlik 3.5'deki gibi gösterilmektedir (Nelsen, 2006).

$$\tau = \tau_{X,Y} = P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} \tag{3.5}$$

(X, Y) sürekli rassal değişkenlerinin kopulası *C* ile gösterilmesi durumunda, *X* ve *Y* için $u=F(x)$ ve $v=G(y)$ göstermek üzere Kendall'ın Tau katsayısı Eşitlik 3.6'daki gibi hesaplanmaktadır (Nelsen, 1999).

$$\tau_{X,Y} = \tau_C = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1 = 1 - 4\int_0^1\int_0^1 \frac{\partial}{\partial u}C(u,v)\frac{\partial}{\partial v}C(u,v)dudv \tag{3.6}$$

Ortak dağılımı *H* ve sürekli marjinal dağılım fonksiyonları *F* ve *G* olan *X* ve *Y* sürekli rassal değişkenler olsun. Bu durumda aşağıdaki eşitlikler geçerlidir.

1) $\rho_S(X,Y) = \rho_S(Y,X)$, $\tau(X,Y) = \tau(Y,X)$
2) X ve Y bağımsız ise, $\rho_S(X,Y) = \tau(X,Y) = 0$
3) $-1 \le \rho_S, \tau \le +1$

X'in değer kümesinde yapılan $T: R \to R$ kesin monoton dönüşümü için ρ_S ve τ'nun her ikisi de Eşitlik 3.7'deki bağıntıyı sağlamaktadır (De Matteis, 2001).

$$\rho(T(X),Y) = \begin{cases} \rho(X,Y) & \text{T artan ise,} \\ -\rho(X,Y) & \text{T azalan ise.} \end{cases} \tag{3.7}$$

Uyumluluk ve uyumsuzluk ölçütü olarak kullanılabilen Kendall'ın Tau katsayısı'nın belirtilen özellikleri taşıması beklenirken, alternatifi olarak Spearman'ın Rho katsayısı da literatürde tercih edilmektedir.

3.1.2.2. Spearman'ın Rho katsayısı

Çok değişkenli normal dağılıma uymayan rassal değişkenler arasında doğrusal bir ilişki yoksa, değişkenler eşit aralıklı ölçülmemişse veya değişkenlerin varyansları bilinmediğinde, Kendall'ın Tau katsayısı gibi uyumluluk ve uyumsuzlukla ilgili bir uyum ölçüsü olan Spearman'ın rho katsayısı tercih edilir.

Spearman'ın rho katsayısı bağımlılık ölçümü, Kendall'ın Tau katsayısında olduğu gibi uyumluluk ve uyumsuzlukla ilişkilidir. Aynı bileşik dağılım fonksiyonuna sahip üç bağımsız vektör olarak (X_1,Y_1), (X_2,Y_2) ve (X_3,Y_3) çiftleri ele alınsın. Spearman'ın Rho katsayısının anakitle için uygun şekli (X_1,Y_1) ve (X_2,Y_3) çiftleri için uyumluluğun olasılığı ile uyumsuzluğun olasılığı arasındaki fark ile orantılıdır ve Eşitlik 3.8'deki gibi ifade edilebilir (Nelsen, 2006).

$$\begin{aligned} &\rho_S(X,Y) = \rho_S(Y,X) \\ &\rho_S = \rho_S(X,Y) = 3\Big(P\big[(X_1 - X_2)(Y_1 - Y_3) > 0\big] - P\big[(X_1 - X_2)(Y_1 - Y_3) < 0\big]\Big) \end{aligned} \tag{3.8}$$

Karadağ (2003)'e göre, (X, Y) sürekli rassal değişkenlerinden n gözlemli bir rassal örneklemin $\{(x_1, y_1),(x_2, y_2),...(x_n, y_n)\}$ şeklinde çekilmesiyle Spearman'ın Rho katsayısı tahmin edicisi Eşitlik 3.9'daki gibi gösterilmektedir.

$$\hat{\rho}_S = \hat{\rho}_S(X,Y) = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \left(rank(X_i) - \frac{n+1}{2} \right) \left(rank(Y_i) - \frac{n+1}{2} \right) \tag{3.9}$$

(X_1, Y_1) sürekli rassal değişkenlerinin kopulası *C* ile, *(X_2, Y_3)* kopulasının Π ile gösterilmesiyle, kopulaya dayalı Spearman'ın Rho katsayısı Eşitlik 3.10'da gösterilmektedir (Cherubini vd. 2004).

$$\begin{aligned}\rho_{X,Y} = \rho_C &= Q(C,\Pi) \\ &= 12\iint uvdC(u,v) - 3 = 12\iint C(u,v)dudv - 3\end{aligned} \tag{3.10}$$

Kendall'ın Tau ve Spearman'ın Rho katsayılarının benzerlikleri ve farklılıkları devam eden bölümde detaylı açıklanacaktır.

3.1.2.3. Kendall'ın Tau ve Spearman'ın Rho katsayısının benzerlikleri ve farklılıkları

Kendall'ın Tau katsayısı ve Spearman'ın Rho katsayısı arasındaki benzerlikler aşağıdaki gibi ifade edilebilir.

- *(X, Y)* gibi sürekli rassal değişkenler için tanımlanırlar.
- Her iki katsayı da simetriktir (Eşitlik 3.11 ve Eşitlik 3.12).

$$\rho_S(X,Y) = \rho_S(Y,X) \tag{3.11}$$

$$\tau(X,Y) = \tau(Y,X) \tag{3.12}$$

- Her iki katsayı da [-1,1] arasında değer almaktadır (Eşitlik 3.13).

$$-1 \le \rho_S(X,Y), \tau(X,Y) \le 1 \tag{3.13}$$

- *X* ve *Y* bağımsızsa katsayılar sıfıra eşittirler.
- Eğer *(X, Y)* aynı yönlü harekete sahip ise $\rho_S(X,Y) = \tau(Y,X) = 1$, zıt yönlü harekete sahip ise $\rho_S(X,Y) = \tau(Y,X) = -1$ olur (Malevergne ve Sornette, 2006).

Her iki ilişki katsayısı da rassal değişkenler arasındaki uyumluluk ve uyumsuzluğu ifade etse de hesaplama yöntemlerinin farklılığı nedeniyle büyüklükleri aynı değildir ve farklı şekilde yorumlanırlar.

X ve *Y* sürekli rassal değişkenler ve τ ve ρ_s sırasıyla Kendall'ın Tau ve Sperman'ın Rho katsayılarını göstermek üzere Eşitlik 3.5 ve Eşitlik 3.7'den yola çıkarak Eşitlik 3.14 ve Eşitlik 3.15 elde edilmektedir. (Nelsen, 2006).

$$-1 \le 3\tau - 2\rho_s \le 1 \tag{3.14}$$

$$\begin{aligned} \frac{1+\rho_s}{2} &\ge \left(\frac{1+\tau}{2}\right)^2 \\ \frac{1-\rho_s}{2} &\ge \left(\frac{1-\tau}{2}\right)^2 \end{aligned} \tag{3.15}$$

Kendall'ın Tau katsayısı ve Spearman'ın Rho katsayısı arasındaki farklılıklar bu şekilde ifade edilebilir.

3.1.3. Kuyruk bağımlılığı katsayısı

Kuyruk bağımlılığı katsayısı ise iki değişkenli dağılımın kuyruğundaki bağımlılık durumunu açıklamaktadır. Bir başka ifadeyle, kuyruk bağımlılığı iki değişkenli bir dağılımın sağ (üst) veya sol (alt) köşesindeki bağımlılık derecesini gösterir. Son zamanlarda, Embrechts vd. (2003) ve Hauksson vd. (2001) çalışmalarında olduğu gibi finansal uygulamalarda piyasa ve kredi riskiyle alakalı olarak kullanılmaktadır. Özellikle, varlık portföyleri için Riske Maruz Değer tahmini yapmak amacıyla kuyruk bağımlı dağılımlar tercih edilmektedir, çünkü bu dağılımlar ile büyük hasarlarla farklı varlıklar arasındaki bağımlılıklar modellenebilir (Cizek vd., 2005).

Çok değişkenli rassal vektörler üzerinde tanımlanan kuyruk bağımlılıkları bu değişkenlerin iki değişkenli marjinal dağılım fonksiyonları ile ilişkilidir. Kuyruk bağımlılıkları bir marjinalin belirli bir eşiği aştığı bilindiği durumda diğer marjinalin de bu eşiği aşma durumunun sınırlandırıcı oranını vermektedir.

$X = (X_1, X_2)^T$ iki boyutlu rassal vektörü, X_1 ve X_2'nin genelleştirilmiş ters dağılım fonksiyonları F_1^{-1}, F_2^{-1} ile ifade edilirken, λ_U üst kuyruk dağılımını göstermektedir. Eşitlik 3.16'daki eşitlik sağlanırsa *X*, üst kuyruk bağımlıdır denilebilirken, $\lambda_U = 0$ eşitliği sağlanırsa üst kuyruk için bağımsızdır yorumu yapılabilir (Cizek vd., 2005).

$$\lambda_U = \lim_{\upsilon \to 1^-} P\left\{ X_1 > F_1^{-1}(v) \mid X_2 > F_2^{-1}(v) > 0 \right\} \quad (3.16)$$

Benzer şekilde λ_L, X'in alt kuyruk bağımlılık katsayısını göstermek üzere, alt kuyruk bağımlılığı Eşitlik 3.17'de ifade edilmektedir (Cizek vd., 2005).

$$\lambda_L = \lim_{\upsilon \to 0^+} P\left\{ X_1 \le F_1^{-1}(v) \mid X_2 \le F_2^{-1}(v) \right\} \quad (3.17)$$

Kuyruk bağımlılığı katsayıları kopula temelli kümeleme tekniklerinde kullanılacak önemli katsayılardandır.

3.2. Kopulalarda Önemli Kavramlar

Çalışmanın temelini oluşturan kopulalar, Nelsen (2006) tarafından "çok değişkenli dağılım fonksiyonlarını kendi tek boyutlu marjinal dağılım fonksiyonlarına bağlayan fonksiyondur" şeklinde tanımlanmıştır.

Kopula fonksiyonu bağımlı tek değişkenli marjinalleri çok değişkenli dağılımlarına bağlayan bir fonksiyon olarak olasılıklı metrik uzaylar kapsamında ilk olarak 1959 yılında Abe Sklar tarafından ele alınmıştır (Frees ve Valdez, 1998).

X ve *Y* rassal değişken çiftinin dağılım fonksiyonları $F(x)=P(X \le x)$ ve $G(y)=P(Y \le y)$ ve bunların ortak dağılım fonksiyonu $H(x, y)=P(X \le x, Y \le y)$ olsun. Her (x,y) reel sayı çifti $F(x)$, $G(y)$ ve $H(x, y)$ olmak üzere üç fonksiyonla ilişkilendirilebilir. Bu reel sayı çiftlerinin her biri [0,1] aralığında yer alır. Diğer bir deyişle her (x, y) reel sayı çifti [0,1]×[0,1] birim karesinde yer alan bir $(F(x), G(y))$ noktasını belirtir, ve her biri [0,1] aralığında bir $H(x, y)$ sayısına karşılık gelir. Bu ortak dağılım fonksiyonunun değerini ayrı ayrı dağılım fonksiyonlarının değerlerinden oluşan düzenli çiftlere atayan fonksiyonlara kopulalar adı verilmektedir (Nelsen, 2006).

Son yıllarda kullanım alanı hızla genişleyen kopulalar, basitçe ifade edilecek olursa, rassal değişkenler arasındaki doğrusal olmayan bağımlılık yapısını inceleme amacıyla kullanılmaktadır.

Kopulalar, çok değişkenli dağılımların dağılım fonksiyonlarını yine aynı değişkenlerin tek değişkenli marjinal dağılım fonksiyonlarına bağlayan çok değişkenli dağılım fonksiyonlarıdır. Bu tek değişkenli marjinal dağılım fonksiyonları [0,1] aralığında

düzgün dağılıma sahiptir. Parametrik olmayan bağımlılık ölçüleri ile çalışırken, çok değişkenli bağımlılık yapılarının incelenmesinde ve Markov süreçlerini incelemede kopulalar sıklıkla tercih edilmektedir (Çelebioğlu, 2007).

Doğrusal olmayan ve parametrik olmayan bağımlılıkların incelenmesinde, kalın kuyruklu dağılımlar için bağımlılıkların incelenmesinde tercih edilen kopulalar, değişkenler arasındaki bağımlılık yapısının araştırılmasında birçok ayrı sebepten dolayı kullanılmaktadır. İki adımlı tahmin süreci nedeniyle hızlı olarak tanımlanan kopulalar, farklı bağımlılık yapılarının asimptotik özelliklerini incelemeye aracılık etmektedir (Karagül, 2013).

Kopulalar, ayrıca aktüerya bilimlerinde bağımlı mortalitelerin ve hasarların modellenmesinde sıklıkla kullanılmaktadır (Frees vd.,1996; Frees ve Valdez 1998; Frees ve Wang 2005).

Bouyé vd. (2000), Embrechts vd. (2003) ve Cherubini vd. (2004) çalışmalarına göre, finansta; varlık paylaşımında, kredi değerlemesinde, risk modellemesinde ve risk yönetiminde kullanılırken; Wang ve Wells (2000) ve Escarela ve Carriere (2003)'e göre biyomedikal çalışmalarda; ilişkili olay zamanlarının ve yarışan risklerin modellenmesinde tercih edilmektedir. Mühendislikte ise çok değişkenli süreç kontrolü ve hidrolojik modellemede kullanımı Yan (2007) ve Genest ve Favre (2007) çalışmalarıyla gösterilmiştir.

İlerleyen bölümlerde anlatılacak olan Sklar teoremi, değişmezlik teoremi ve Fréchet-Hoeffding sınırları kopulaların önemli başlıklarıdır.

3.2.1. Sklar teoremi

Sklar teoremi kopulalar teorisinin yapı taşıdır. Sklar teoremi olmadan kopula kavramı ortak dağılım fonksiyonunun zengin bir üyesi olacaktı (Piotr vd., 2009). Bu teorem, çok boyutlu dağılım fonksiyonlarının tek değişkenli marjinalleri ile aralarındaki ilişkide kopulaların ne işe yaradığını açıklamaktadır (Nelsen, 2006).

Tanım 4.1.

Tanım kümesi $\bar{R}$ olan bir F fonksiyonu bir dağılım fonksiyonudur öyle ki;

F azalmayan bir fonksiyondur,

F (- ∞)= 0 ve F(∞) =1.

Tanım 4.2.

H; tanım kümesi $\bar{R}^2$ olan ortak dağılım fonksiyonu olsun, öyle ki;

H, artan bir fonksiyondur,

H (x,- ∞) = H(-∞ ,y) = 0 ve H(-∞,∞) = 1

Teorem 4.1.

H, marjinalleri *F* ve *G* olan bir dağılım fonksiyonu olsun. Her *x, y* $\in \bar{R}$ için bir *C* kopulası Eşitlik 3.18'deki gibi tanımlanır (Nelsen, 2006).

$$H(x, y) = C(F(x), G(y)) \tag{3.18}$$

Eğer F ve G sürekli ise C kopulası tektir, tam tersi durumda yani sürekli olmadığı durumda ise C kopulası F ve G'nin değer kümelerinin Kartezyen çarpımları üzerinde tek olarak tanımlanmıştır. Tam tersine, eğer *C* bir kopula ve *F* ve *G* dağılım fonksiyonları ise, bu durumda *H* fonksiyonu marjinalleri *F* ve *G* ile gösterilen bir ortak dağılım fonksiyonu olur.

Bu teoremden, F^{1}, G^{-1} sırasıyla *F* ve *G* fonksiyonlarının tersi olmak üzere, *C* kopulası için Eşitlik 3.19 biçiminde yazılabilir (Nelsen, 2006).

$$C(x, y) = F(F^{-1}(x), G^{-1}(y)) \tag{3.19}$$

F(x) ve *G(y)* fonksiyonları sürekli ise Eşitlik 3.2'deki sonuç kopulalar için de sağlanır ve böylelikle ortak dağılım fonksiyonlarından kopulaları oluşturan bir yöntem sağlamaktadır (Nelsen, 2006).

3.2.2. Değişmezlik teoremi

Rassal değişkenlerin doğrusal olmayan dönüşümleri üzerine geliştirilen değişmezlik teoremi kopulaların temel konularındandır.

$X_1, X_2, ..., X_n$ rassal değişkenleri C kopulasına sahipse, Y_i de X_i'nin artan fonksiyonu olmak şartıyla $Y_1 = h_1(X_1), ..., Y_n = h_n(X_n)$ 'nin de aynı C kopulasına sahip olduğu Eşitlik 3.20'de görülmektedir (Malevergne ve Sornette, 2006).

$$C(F_1(x_1), ..., F_n(x_n)) = C(h_1(F_1(x_1)), ..., h_n(F_n(x_n))) \quad (3.20)$$

Değişmezlik teoreminden görülebileceği gibi kopulalar rassal değişkenlerin doğrusal olmayan dönüşümlerinden etkilenmemektedir (Malevergne ve Sornette, 2006).

3.2.3. Fréchet-Hoeffding sınırları

Fréchet-Hoeffding sınırları kopulaların tam negatif ve tam pozitif bağımlılıklarını tanımlamaktadır. Bir kopulayı gösteren C, $(u,v) \in I^2$ için Eşitlik 3.21'deki eşitsizliği sağlamaktadır (Nelsen, 1999).

$$\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v) \quad (3.21)$$

$W(u,v) = \max(u + v - 1, 0)$ ve $M(u,v) = \min(u,v)$ ile gösterilirse Eşitlik 3.21'deki teorem Eşitlik 3.22'deki gibi ifade edilebilir. Bu eşitlikte $W(u,v)$ Fréchet-Hoeffding alt sınırını, $M(u,v)$ ise Fréchet-Hoeffding üst sınırını göstermektedir (Nelsen, 1999).

$$W(u, v) \leq C(u, v) \leq M(u, v) \quad (3.22)$$

Burada n=2 için alt ve üst sınırların kendileri birer kopuladır ve W tam negatif bağımlılığı ve M de tam pozitif bağımlılığı belirtmektedir (Nelsen, 2006).

3.2.4. Yaşam (Sağkalım) kopulası

Yürütülen pek çok çalışmada, yaşam süreleri rassal değişkenler olan sistemler ilgi çekmektedir. Bir bireyin veya sistemin x gibi bir zamandan daha uzun yaşama olasılığı $\bar{F}_i(x) = 1 - F_i(x)$ yaşam (sağkalım) fonksiyonu olarak ifade edilmektedir. ($X_1, X_2, ..., X_n$) için çok değişkenli yaşam (sağkalım) fonksiyonu $\bar{F}(x_1, x_2, ..., x_n)$ şeklinde gösterilebilir. Bu durumda yaşam (sağkalım) kopulası Eşitlik 3.23'de ifade edildiği gibidir (Kaishev vd., 2007).

$$\bar{F}(x_1, x_2, ..., x_n) = \bar{C}(\bar{F}_1(x_1), \bar{F}_2(x_2), ..., \bar{F}_n(x_n)) \quad (3.23)$$

Yaşam kopulası aracılığıyla özellikle çoklu yaşam durumlarında ortaya çıkan birleşik olasılıklar incelenebilmektedir.

3.3. Bazı Önemli Kopula Aileleri

Çalışmalarda kullanılan çok çeşitli kopula aileleri bulunmaktadır. Ancak, belli başlı temel bazı kopula çeşitleri bulunmaktadır. İlerleyen bölümlerde aşağıda bahsi geçen kopula aileleri incelenecektir.

a) Eliptik Kopulalar

i. Gaussian (Normal) Kopula

ii. Student-t Kopula

b) Arşimedyan Kopulalar

i. Gumbel Kopula

ii. Clayton Kopula

iii. Frank Kopula

iv. Ali-Mikhail-Haq Kopula

c) Farlie-Gumbel-Morgenstern (FGM) Kopulalar

3.3.1. Eliptik kopulalar

Eliptik kopulalar çok değişkenli eliptik dağılımlardan elde edilmektedir. En önemli eliptik kopula aileleri Student-t kopula ve Gaussian (Normal) kopula aileleridir.

Malevergne ve Sornette (2006) eliptik kopulaları sayısal olarak kolay sentezlenebilir olarak tanımlarken, benzetimlerde bu nedenle daha kullanışlı bulmaktadır. Kolay sentezlenebilme sebepleri olarak, Gaussian ve t dağılımlı rassal değişkenlerin kolayca üretilmesi gösterilmektedir. Değişkenlerin monoton değişimleri ile kopulanın değişmezliği korunurken, bu dağılımların da doğru marjinal dağılımlar verdiği belirtilmiştir.

3.3.1.1. Gaussian (Normal) kopula

Gaussian veya normal kopula Sklar teoreminden türemiştir. Asyalı bankacı David X. Li tarafından ortaya atılan kopula Eşitlik 3.24 ile tanımlanmıştır.

$$C(u_1,u_2)=\Phi_2(\Phi^{-1}(u_1),\Phi^{-1}(u_2),\rho),\quad -1\le\rho\le 1 \tag{3.24}$$

Φ_2 Eşitlik 3.3'de ifade edilen korelasyon katsayısının iki boyutlu normal dağılım fonksiyonudur ve Φ^{-1} ise tek değişkenli normal dağılım fonksiyonunun tersidir (Li, 2000).

Normallik varsayımı nedeniyle Gaussian kopula sıklıkla tercih edilmekle birlikte Normal dağılımla yakından ilişkili olan Student-t dağılımından elde edilen Student-t kopula da çalışmalarda yer bulmaktadır.

3.3.1.2. Student-t kopula

Normal kopulanın iki değişkenli Normal dağılımdan üretildiği gibi Student-t kopulası da iki değişkenli Student-t dağılımından elde edilir. *m* serbestlik dereceli tek değişkenli Student-t dağılımının dağılım fonksiyonu Eşitlik 3.25'deki gibidir (Nelsen, 1999).

$$t_m(x)=\int_{-\infty}^{\infty}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)}\left(1+\frac{x^2}{m}\right)^{-\frac{m+1}{2}}dx \tag{3.25}$$

Student-t kopulası ise

$$\begin{aligned}C_{m,\rho}(u_1,u_2)&=t_{m,\rho}(t_m^{-1}(u_1),t_m^{-1}(u_2))\\&=\int_{-\infty}^{t_m^{-1}(u_1)}\int_{-\infty}^{t_m^{-1}(u_2)}\frac{1}{2\pi\sqrt{1-\rho^2}}\left(1+\frac{x_1^2+x_2^2-2\rho x_1x_2}{m(1-\rho^2)}\right)^{-\frac{m+2}{2}}dx_1x_2\quad,-1\le\rho\le 1\end{aligned} \tag{3.26}$$

Eşitlik 3.26'da görüldüğü gibi tanımlanır. $m\to\infty$ için Student-t kopulası Normal dağılıma yaklaşır. Sınırlı sayıda serbestlik dereceleri için iki kopulanın eğimleri oldukça farklıdır.

3.3.2. Arşimedyan kopulalar

İlk defa 1965 yılında Ling tarafından isimlendirilen Arşimedyan kopulaların t norm'da geçerliliği Sklar ve Schweizer tarafından ileri sürülmüştür (Cherubini, 2004). Oluşturulmalarının kolay olması ve parametrik kopulaların birçoğunun Arşimedyan kopula olması nedeniyle uygulama alanları oldukça geniştir (Yaprakçı, 2007). Diğer kopulaların aksine arşimedyan kopulalar Sklar teoremini kullanan çok değişkenli dağılım fonksiyonlarından türememiştir (Embrechts, 2001).

Başlıca Arşimedyan kopulalar, Gumbel, Clayton, Frank, Ali-Mikhail-Haq kopuladır.

3.3.2.1. Gumbel kopula

İki değişkenli Gumbel kopulası veya Gumbel-Hougaard kopulası asimetrik bir Arşimedyan kopuladır. Eşitlik 3.27 biçiminde tanımlanan Gumbel kopula ailesi üst (sağ) kuyruk bağımlılığı göstermektedir.

$$C_{\theta}^{Gu}(u_1,u_2)=\exp\left[-((-\ln u_1)^{\theta}+(-\ln u_2)^{\theta})^{\frac{1}{\theta}}\right],\quad \theta\in[1,\infty) \tag{3.27}$$

İlgili kopulada üretici fonksiyonu, $\varphi(t)=(-\ln t)^{\theta}$ eşitliği ile elde edilir. θ bağımlılık yapısını göstermek üzere, θ parametresinin artmasıyla gözlemler arasındaki bağımlılık artar. $\theta=1$ için $\lim_{n\to\infty} C_{\theta}(u_1,u_2)=u_1u_2=\Pi(u_1,u_2)$ eşitliği ile bağımsızlık durumunu ifade ederken, $\theta\to\infty$ için maksimum pozitif bağımlılık vardır (Nelsen, 2006).

3.3.2.2. Clayton kopula

İlk olarak Kimeldorf ve Sampson (1975) tarafından bulunmasına rağmen Clayton (1978)'a atfedilmiştir. Bu dağılımlara ilişkin ilk sistematik çalışma ise Cook ve Johnson (1981) tarafından yapılmıştır. Clayton kopula ailesi alt (sol) kuyrukta bağımlılık inceleyen asimetrik bir Arşimedyen kopuladır.

$\theta \in [-1, \infty) / \{0\}$ olmak üzere Clayton kopula Eşitlik 3.28'deki gibi tanımlanır.

$$C_\theta(u_1, u_2) = \max(\left[u_1^{-\theta} + u_2^{-\theta} - 1 \right]^{-1/\theta}, 0) \tag{3.28}$$

Üretici fonksiyonu, Eşitlik 3.29'da görüldüğü gibidir.

$$\varphi(t) = \frac{1}{\theta}(t^{-\theta} - 1) \tag{3.29}$$

θ parametresinin artmasıyla gözlemler arasındaki bağımlılık artar. $\theta \to 0$ iken, yani $\lim_{n\to\infty} C_\theta(u_1, u_2) = u_1 u_2 = \Pi(u_1, u_2)$ eşitliği ile bağımsızlık durumunu ifade ederken, $\theta \to \infty$ için maksimum pozitif bağımlılık vardır (Nelsen, 2006).

3.3.2.3. Frank kopula

Frank kopula ailesi simetrik bir Arşimedyen kopuladır. Bağımlılık her iki kuyrukta incelenmektedir. $\theta \in \mathbb{R} / \{0\}$ olmak üzere Frank kopula Eşitlik 3.30'daki gibi tanımlanır.

$$C_\theta(u_1, u_2) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{(e^{-\theta} - 1)} \right] \tag{3.30}$$

Üretici fonksiyonu ise Eşitlik 3.31'de ifade edilmektedir.

$$\varphi(t) = -\ln \frac{e^{-\theta t} - 1}{e^{-\theta} - 1} \tag{3.31}$$

$\theta \to 0$ iken, yani $\lim_{n\to\infty} C_\theta(u_1, u_2) = u_1 u_2 = \Pi(u_1, u_2)$ eşitliği ile bağımsızlık durumunu ifade ederken, $\theta \to \infty$ için Fréchet-Hoeffding üst sınırına eşit olur.

Frank kopula ailesi, diğer bazı kopulalardan farklı olarak marjinaller arasında negatif bağımlılığa izin vermesi ve bağımlılığın kuyrukların ikisinde de olmasından dolayı, yani simetrik olması nedeniyle, sıklıkla tercih edilmektedir (Trivedi ve Zimmer, 2005).

3.3.2.4. Ali-Mikhail-Haq kopula

X ve *Y* tesadüfi değişkenleri için Gumbel kopulanın iki değişkenli lojistik dağılımı $H(X, Y)=(1+e^{-X}+e^{-Y})^{-1}$ şeklindedir.

Ali, Mikhail ve Haq ortak dağılım fonksiyonunu Eşitlik 3.32'deki gibi tanımlayarak doğrulamışlardır.

$$H_\theta(X,Y) = (1+e^{-X}+e^{-Y}+(1+\theta)e^{-X-Y})^{-1}, \qquad \theta \in [-1,1] \tag{3.32}$$

Olasılık dönüşümleri ve matematiksel yöntemler kullanılarak, AMH kopula elde edilir ve Eşitlik 3.33 ile ifade edilir.

$$\varphi(t) = \ln \frac{1-\theta(1-t)}{t} \tag{3.33}$$

Eşitlik 3.33'deki üretici fonksiyonu dikkate alınarak AMH kopula eşitliği Eşitlik 3.34'deki gibi elde edilir.

$$C_\theta(u_1,u_2) = u_1u_2\left[1-\theta(1-u_1)(1-u_2)\right]^{-1} \tag{3.34}$$

[-1, 1] aralığında değer alan θ parametresi pozitif ve negatif bağımlılığa izin verir. $\theta=0$ iken, yani Eşitlik 3.35'teki bağımsızlık kopulasına eşit olur (Kumar, 2010).

$$\lim_{n\to\infty} C_\theta(u_1,u_2) = u_1u_2 = \Pi(u_1,u_2) \tag{3.35}$$

Sıklıkla tercih edilen Arşimedyan kopulaların ardından Farlie-Gumbel-Morgenstern (FGM) kopulalar açıklanacaktır.

3.3.3. Farlie-Gumbel-Morgenstern (FGM) kopulalar

Farlie-Gumbel-Morgenstern (FGM) kopulaları, Cauchy marjinalleri kullanılarak ilk olarak Morgenstern (1956), Gumbel (1960) ve Farlie (1960)'nin çalışmaları ile ilerlemiştir.

Rassal değişkenler arasındaki bağımlılık durumu incelendiğinde, Lai ve Xie (2000) pozitif kadran bağımlı iki değişkenli dağılımları ele almıştır. *X* ve *Y* pozitif kadran bağımlı değişkenler, yani $\forall x,y$ için $P\{X \le x, Y \le y\} \ge P\{X \le x\}\, P\{Y \le y\}$ olduğu durumda; *F(x,y), X*

ve *Y*'nin ortak dağılım fonksiyonu, $F_X(x)$ ve $F_Y(y)$ sürekli marjinal dağılımları olmasıyla *X* ve *Y* 'ye ait ortak dağılım fonksiyonu Eşitlik 3.36'da görüldüğü gibi yazılabilir. $\forall x,y$ için $w(x,y)$ negatif olmayan bir fonksiyon ise $F(x, y)$ pozitif bağımlı bir dağılım fonksiyonudur (Sevindik, 2009).

$$F(x,y) = F_X(x)F_Y(y) + w(x,y) \tag{3.36}$$

Eşitlik 3.36'dan yola çıkarak, (X, Y) mutlak sürekli rassal değişkenler olmak üzere $x \to 1$ iken $\lim_{x\to 1} A(x) = 0$ ve $\lim_{x\to 1} B(x) = 0$, $F(x)$ ve $G(y)$ mutlak sürekli marjinal dağılım fonksiyonları olmak üzere FGM kopula ailesinin genel gösterimi Eşitlik 3.37'de ifade edildiği gibidir (Bairamov vd., 2001).

$$C_a(x,y) = F(x)G(y)\{1 + \alpha A(F(x))B(G(y))\} \tag{3.37}$$

Bairamov vd. (2001) FGM kopulalarına dair yeni bir modifikasyon elde etmiş ve ilişki parametresine ait sınırları tanıtılmıştır ve Eşitlik 3.38'de gösterilmektedir.

$$\begin{aligned} F(x,y) = xy\left\{1+\alpha\left(1-x^{p_1}\right)^{q_1}\left(1-y^{p_2}\right)^{q_2}\right\};\ p_1, p_2 \geq 1,\ q_1, q_2 \geq 1,\ 0 \leq x, y \leq 1 \\ -\min\left\{1, \frac{1}{p_1 p_2}\left(\frac{1+p_1q_1}{p_1(q_1-1)}\right)^{q_1-1}\left(\frac{1+p_2q_2}{p_2(q_2-1)}\right)^{q_2-1}\right\} \leq \alpha \leq \\ \min\left\{\frac{1}{p_1}\left(\frac{1+p_1q_1}{p_1(q_1-1)}\right)^{q_1-1}, \frac{1}{p_2}\left(\frac{1+p_2q_2}{p_2(q_2-1)}\right)^{q_2-1}\right\} \end{aligned} \tag{3.38}$$

FGM kopulaları için son modifikasyon 2004 yılında Rodriguez ve Flores tarafından ileri sürülmüştür. Bu dağılım kopulası ve ilişki parametresine ait sınırlar Eşitlik 39'da ifade edilmektedir.

$$C(u,v) = uv + \lambda u^a v^b (1-u)^c (1-v)^d;\ a,b,c,d \geq 1$$

$$-\frac{1}{\max\{\alpha\gamma, \beta\delta\}} \leq \lambda \leq -\frac{1}{\min\{\alpha\gamma, \beta\delta\}}$$

$$\alpha = -\left(\frac{a}{a+c}\right)^{a-1}\left(1+\sqrt{\frac{c}{a(a+c-1)}}\right)^{a-1}\left(\frac{c}{a+c}\right)^{c-1}\left(1-\sqrt{\frac{a}{c(a+c-1)}}\right)^{c-1}\sqrt{\frac{ac}{a+c-1}}$$

$$\beta = \left(\frac{a}{a+c}\right)^{a-1}\left(1-\sqrt{\frac{c}{a(a+c-1)}}\right)^{a-1}\left(\frac{c}{a+c}\right)^{c-1}\left(1+\sqrt{\frac{a}{c(a+c-1)}}\right)^{c-1}\sqrt{\frac{ac}{a+c-1}} \qquad (3.39)$$

$$\gamma = -\left(\frac{b}{b+d}\right)^{b-1}\left(1+\sqrt{\frac{d}{b(b+d-1)}}\right)^{b-1}\left(\frac{d}{b+d}\right)^{d-1}\left(1-\sqrt{\frac{b}{d(b+d-1)}}\right)^{d-1}\sqrt{\frac{bd}{b+d-1}}$$

$$\delta = \left(\frac{b}{b+d}\right)^{b-1}\left(1-\sqrt{\frac{d}{b(b+d-1)}}\right)^{b-1}\left(\frac{d}{b+d}\right)^{d-1}\left(1+\sqrt{\frac{b}{d(b+d-1)}}\right)^{d-1}\sqrt{\frac{bd}{b+d-1}}$$

Bu bölümde (X, Y) rasgele değişken çifti arasındaki bağımlılık yapısı C kopula parametre ailesi ile modellenmiştir. *(X, Y)*'den alınan rasgele örneklem $(X_1, Y_1), ..., (X_n, Y_n)$ için belirlenen kopula modeline ait Θ parametresi için "parametrik, yarı parametrik ve parametrik olmayan yöntemler" alt başlıklarında tahmin etme yöntemleri hakkında bilgiler verilmiştir (Genest ve Favre, 2007).

3.4. Kopula Tahmin Yöntemleri

Kopulaların parametre tahminlerinde diğer istatistik yöntemleri gibi pek çok yöntem bulunmaktadır. Bu tahmin yöntemleri ise aşağıdaki gibi gösterilmektedir.

a) Parametrik Yöntemler
   i. Tam En Çok Olabilirlik Yöntemi (MLE)
   ii. Marjinallere İlişkin Çıkarsama Fonksiyonu (IFM)
b) Yarı Parametrik Yöntemler
   i. Sözde En Çok Olabilirlik Yöntemi (PMLE)
c) Parametrik Olmayan Yöntemler

3.4.1. Parametrik yöntemler

Parametrik yöntemler, değişkenlere ait marjinal olasılık yoğunluk fonksiyonları ve kopula fonksiyonları ile ilgili bilgi sahibi olunması durumunda kullanılan tahmin yöntemleridir. Tam En Çok Olabilirlik yöntemi ve Marjinallere İlişkin Çıkarsama Fonksiyonu en sık tercih edilenleridir.

3.4.1.1. Tam En Çok Olabilirlik Yöntemi (MLE)

Bu tahmin yöntemini açıklamadan önce, kanonik gösterim olarak adlandırılan olabilirlik fonksiyonunun ifade edilmesi gerekmektedir. Kanonik gösterim Eşitlik 3.40'da görülmektedir (Cherubini vd., 2004).

$$f(x_1,...,x_n) = c(F_1(x_1),...,F_n(x_n))\prod_{j=1}^{n} F_j(x_j) \tag{3.40}$$

Kanonik gösterimde $c(F_1(x_1),...,F_n(x_n))$ c kopula yoğunluk fonksiyonunu ifade etmek üzere, n. mertebeden kısmi türevi Eşitlik 3.41'de ifade edilmektedir.

$$c(F_1(x_1),...,F_n(x_n)) = \frac{\partial^n C(F_1(x_1),...,F_n(x_n))}{\partial F_1(x_1)\partial F_2(x_2)...\partial F_n(x_n)} \tag{3.41}$$

Örnek veri matrisi $S = \{x_{1t},...,x_{nt}\}_{t=1}^{T}$ olsun. Buna göre log-olabilirlik fonksiyonu Eşitlik 3.42'de görüldüğü biçimde ifade edilir.

$$l(\theta) = \sum_{t=1}^{T} \ln c\left(F_{1t}(x_{1t}),...,F_{nt}(x_{nt})\right) + \sum_{t=1}^{T}\sum_{j=1}^{n} \ln f_j(x_{jt}) \tag{3.42}$$

Eşitlik 3.36'da kopula ve marjinallere ilişkin tüm parametrelerin kümesi Θ ile ifade edilmek üzere en çok olabilirlik tahmin edicisi $\theta_{MLE} = \max_{\theta \in 0} l(\theta)$ olacaktır (Cherubini vd., 2004).

3.4.1.2. Marjinallere ilişkin çıkarsama fonksiyonu (IFM)

En çok olabilirlik yöntemi uygulamalarında marjinal dağılım parametreleri ve kopula yoğunluk fonksiyonu bilinmelidir. Buna rağmen, log-olabilirlik fonksiyonu pozitif iki terimden oluşmaktadır. İlk terim kopula yoğunluğunu ve parametrelerini ifade ederken, diğer terim marjinalleri ve kopula yoğunluğunun parametrelerini içermektedir. Joe (1997) bu parametrelerin tahmin edilmesi için iki adımdan oluşan süreci önermektedir.

Adım 1: Tek değişkenli marjinal dağılımlar ile marjinal parametreleri θ_1 Eşitlik 3.43'de görüldüğü şekilde tahmin edilir.

$$\hat{\theta}_1 = ArgMax_{\theta_1} \sum_{t=1}^{T} \sum_{j=1}^{n} \log f_i(x_{jt}; \theta_1) \tag{3.43}$$

Adım 2: Elde edilen $\hat{\theta}_1$ log-olabilirlik fonksiyonunda yerine konularak kopula parametresi Eşitlik 3.44'de görüldüğü şekilde tahmin edilir.

$$\hat{\theta}_2 = ArgMax_{\theta_2} \sum_{t=1}^{T} \ln c\left(F_1(x_{1t}), F_2(x_{2t}), ..., F_n(x_{nt}); \theta_2, \hat{\theta}_1\right) \tag{3.44}$$

Marjinallere ilişkin çıkarsama yönteminin tahmin edicisi $\hat{\theta}_{IFM} = \left(\hat{\theta}_1, \hat{\theta}_2\right)'$ eşitliği ile ifade edilir.

l, log-olabilirlik; l_j, j'inci marjinal log-olabilirlik ve l_c, kopulanın log-olabilirlik fonksiyonunu göstermek üzere, IFM tahmin edicisi Eşitlik 3.45'de gösterilen eşitliğin çözümüdür.

$$\left(\frac{\partial l_1}{\partial \theta_{11}}, \frac{\partial l_2}{\partial \theta_{12}}, ..., \frac{\partial l_n}{\partial \theta_{1n}} \frac{\partial l_c}{\partial \theta_2}\right) = 0' \tag{3.45}$$

En çok olabilirlik tahmin edicisi ise Eşitlik 3.46'da gösterilen eşitliğin çözümüdür.

$$\left(\frac{\partial l}{\partial \theta_{11}}, \frac{\partial l}{\partial \theta_{12}}, ..., \frac{\partial l}{\partial \theta_{1n}} \frac{\partial l}{\partial \theta_2}\right) = 0' \tag{3.46}$$

IFM tahmin edicisi MLE tahmin edicisine göre daha kolay hesaplanmasına rağmen bu iki tahmin edicinin sonuçları genellikle eşit çıkmamaktadır. Joe (1997)'de IFM tahmin

edicisinin MLE tahmin edicisine göre asimptotik olarak daha etkin bir tahmin edici olduğu gösterilmiştir (Joe, 1997 ve Cherubini, 2004).

3.4.2. Yarı parametrik yöntemler

Kopula fonksiyonuna ait yoğunluk fonksiyonunun bilindiği durumlarda kullanılan yöntemlerdir. Bu bölümde Sözde En Çok Olabilirlik Yöntemi (PMLE) anlatılacaktır.

Sözde En Çok Olabilirlik Yöntemi, bağımlılık parametresinin sıra sayılarına dayalı olduğu bir tahmine uyarlanmasıdır. Söz konusu uyarlama genel hatlarıyla Oakes (1994) tarafından tanımlanmıştır.

c_θ yoğunluğuna sahip C_θ kopulasının kesinlikle sürekli olmasını gerektiren sözde en çok olabilirlik yöntemi, basit sıra sayılarına dayalı log-olabilirlik fonksiyonunu maksimum hale getirecek parametrenin tahminini Eşitlik 3.47'de görüldüğü gibi elde edilir.

$$l(\theta) = \sum_{i=1}^{n} \log\left\{ c_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) \right\} \tag{3.47}$$

Eşitlik 3.48'de belirtilen klasik log-olabilirlik fonksiyonunda bilinmeyen *F* ve *G* marjinal dağılım fonksiyonlarının yerine Eşitlik 3.49 ve Eşitlik 3.50 kullanılır.

$$l(\theta) = \sum_{i=1}^{n} \log\left\{ c_\theta\left(F(X_i), G(Y_i) \right) \right\} \tag{3.48}$$

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} I(X_i \le x) = \frac{R_i}{n+1} \tag{3.49}$$

$$G_n(y) = \frac{1}{n+1} \sum_{i=1}^{n} I(Y_i \le y) = \frac{S_i}{n+1} \tag{3.50}$$

R_i ve S_i, X_i ve Y_i rassal değişkenlerinin gözlem değerlerinden elde edilen sıra sayı değerlerini göstermektedir. En çok sözde olabilirlik yöntemi, daha çok sayısal işlem içerdiği ve c_θ yoğunluğunun var olmasını gerektirdiği için Kendall'ın Tau (τ)'sunu ve Spearman'ın Rho (ρ_S)'sunu içeren yöntemlere göre daha az tercih edilmektedir. Öte

yandan, bu yöntem bağımlılık parametresi θ'nın gerçek sayı olması koşulunu öne sürmediğinden, diğer yöntemlere göre daha fazla uygulanabilir (Genest ve Favre, 2007).

3.4.3. Parametrik olmayan yöntemler

Parametrik olmayan yöntemler, marjinal dağılımlar ve parametreler ile ilgili bilgi sahibi olunamadığı durumlarda kullanılan tahmin yöntemleridir. Bu yöntemlerde sıra korelasyon katsayıları olan Kendall'ın Tau'suna dayalı yöntem ve Spearman'ın Rho'suna dayanan yöntem bağımlılık ölçüleri olarak kullanılmaktadır. Elde bulunan veri kümesinden örneklem çekilerek bağımlılık ölçülerine ait değerlerin hesaplanmasının ardından, kopula ailesi tahmin edilir.

3.5. Kopula Temelli Kümeleme Teknikleri

Kümeleme analizi, bir grup nesnenin kümeler denilen gruplara atanmasıyla veri içindeki bir yapıyı tespit etmeyi amaçlayan denetimsiz bir sınıflama yöntemidir. Klasik anlamda, aynı kümedeki nesnelerin farklı kümelere atanan nesnelerden daha yakın şekilde birbirleriyle ilişkili olduğu yorumu yapılabilir.

Kümeleme yöntemleri üzerine literatür çok geniş ve bu yöntemleri düzenlemek ve sunmak için farklı ölçütler göz önünde bulundurulmaktadır. Kümeleme yöntemlerinin matematiksel yanlarından birisini ifade eden bir ölçüt, bir olasılık modeline karşı bir uzaklık veya benzemezlik ölçüsüdür. Veri matrisinin belirli bir veri üretme sürecine göre oluşturulduğunu varsayan model tabanlı kümeleme yöntemlerini ele aldığımızda, çok değişkenli normal dağılım gibi, karışık çok değişkenli olasılık dağılımlarına dayanmakta olduğu görülmektedir. Bununla birlikte, bu yaklaşım nesneler arasındaki doğrusal bağımlılık üzerine kuruludur ve doğrusal korelasyon katsayısının sınırlamalarına maruz kalmaktadır.

Kopula temelli kümeleme tekniklerinden birisi CoClust iken diğeri benzemezlik matrisine dayanan kuyruk bağımlılığı aracılığıyla kümeleme tekniğidir. Doğrusal olmayan bağımlılık yapısı üzerine kurulu bu iki teknik bu bölümün temel başlıklarıdır.

3.5.1. CoClust tekniği ile kopulalar aracılığıyla kümeleme

Kopula temelli kümeleme tekniği, bir kümeleme algoritmasına kopula ailelerini dahil etmeye yardımcı olur. Kopula temelli kümeleme tekniği (CoClust), Lascio tarafından 2008 yılında doktora tezi aracılığıyla tanıtılmış, 2016 yılında geliştirilmiş ve Lascio ve Giannerini (2019) tarafından son hali sunulmuştur. Veri işleme sürecinin karmaşık çok değişkenli bağımlılık yapısına göre kümeleme olanağı vermektedir. Bu nedenle, diğer avantajlarının yanı sıra, CoClust tekniği, doğrusal iki değişkenli bağımlılıklarla ilgilenen klasik yaklaşımların sınırlamalarının üstesinden gelmektedir.

Lascio (2008) çalışmasında ilk önce, kümeleme algoritmasını açıklamaktadır. Daha sonra, kümeleme algoritmasını R paketi aracılığıyla açıklamaktadır. Lascio (2008) ana R komutlarının çok değişkenli bağımlı verilerin tam gelişmiş kümelenmesini gerçekleştirmek için nasıl kullanılabileceğini gösteren sayısal örnekleri bu çalışmada sunmaktadır.

Bu nedenle, veri matrisinin K-boyutlu kopulalar aracılığıyla üretildiğini varsayan kümeleme yöntemlerine odaklanmaktadır. Öyle ki, K kümelerinin her biri sürekli tek değişkenli yoğunluk fonksiyonu ile temsil edilir ve kümeler arasındaki karmaşık çok değişkenli ilişki kopula ve kopulanın bağımlılık parametresi tarafından ifade edilmektedir. Kopula temelli kümeleme tekniği Lascio (2008) tarafından tanımlanmış CoClust yaklaşımı Chessa vd.'nin (2014) çalışmasına ilham olmakla birlikte, Durante vd. (2014), Durante vd. (2015), De Luca ve Zuccolotto (2011), De Luca ve Zuccolotto (2014) ve De Luca ve Zuccolotto (2017)'da ifade edildiği gibi kopula temelli farklı kümeleme teknikleri geliştirilmiştir.

Lascio (2008) çok değişkenli gözlemleri karmaşık bir bağımlılık yapısı ile kümeleyebilen CoClust adında bir küme algoritması önermektedir. CoClust kavramı, olabilirlik kopulasına dayanan, çok değişkenli bağımlı değişkenlerin kümelenmesini ifade etmektedir. Bu kümelemeyi gerçekleştirebilmek için, CoClust verinin parametrelerinin kümeleri temsil eden ve her bir kümenin tek değişkenli yoğunluk fonksiyonu (marjinal fonksiyon) ile temsil edildiği bilinen, çok değişkenli kopula fonksiyonu tarafından türetildiğini varsaymaktadır. Kümeler arası çok değişkenli bağımlılığın gücü ve türü sırasıyla, bir kopula fonksiyonu ve kopulanın bağımlılık parametresi ile modellenir. Kopula tabanlı olan CoClust, kopula fonksiyonlarının tüm avantajlarını kullanır ve veri üretme sürecinin çok değişkenli karmaşık bağımlılık yapısı kümeleme analizini

gerçekleştirmek için dikkate alınabilir. Bununla birlikte, CoClust tekniğinin ilk taslağında önemli sınırlamalar mevcuttu. Örneğin, tüm gözlemleri ilgisiz gözlemleri atmadan kümelere otomatik olarak ayırıyordu ve bu durum uzun hesaplama süreci getirmektedir. Lascio ve Giannerini (2019) bu sınırlamaları aşacak şekilde yeni bir CoClust algoritması önermektedir. R paketinin CoClust uygulamasında uygulanan CoClust algoritmasının en son versiyonunu açıklamaktadır.

Algoritmanın başlangıç noktası (n x p) veri matrisi X'dir. Eşitlik 3.51 ile ifade edilmektedir.

$$X = \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i'1} & \dots & x_{i'j} & \dots & x_{i'p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \tag{3.51}$$

Kümelemenin amacı nxp boyutlu veri setini, K grupta gruplandırmaktır. CoClust, analiz amacına uygun olarak, satır veya sütun veri matrisine uygulanabilir. Her iki durumda da, CoClust vektörlerle çalışır ve X veri matrisinin her bir satırını (sütununu) bir kümeye ayrılacak tek bir öğe olarak ele alır. Bir satır (veya sütun) vektöründeki değerler, aynı yoğunluk fonksiyonunun bağımsız fonksiyonlarıdır, böylece her bir kümedeki gözlemler aynı dağılımdandır. Burada algoritma, veri matrisinin satırlara uygulandığı şekilde tarif edilmektedir (Lascio, 2008).

CoClust, kopula fonksiyonlarına dayanarak, marjinaller üzerinde herhangi bir varsayım yapmadan çok değişkenli bağımlılık yapısına göre gözlemlerin kümelenmesini sağlar. CoClust ardındaki temel fikir, satır veri matrisinin K grubunu bir seferde ayırmasıdır, yani her küme için p boyutlu vektörü ayıran ileri düzey bir prosedür oluşturmaktadır. Satırların her bir K grubunun ayrımı ile ilgili karar kopulanın veriye uygunluğuna göre log-olabilirlik değerine dayanmaktadır. Bu olabilirlik değeri, halihazırda ayrılmış K grupları kullanarak hesaplanır ve $x_{i'} = \left(x_{i'1}, \dots, x_{i'j}, \dots, x_{i'p}\right)$ gibi ayrıştırılacak olan aday, veriye uygun kopulayı maksimize eden kombinasyonu tespit etmek amacıyla $x_{i'}$ 'deki değişkenlerin permütasyonları aracılığıyla elde edilir (Lascio, 2008).

Lascio ve Giannerini (2019), CoClust tekniğinin iki özelliğinin tartışmaya açık olduğunu belirtmektedir. Bunlardan ilki, ayrıştırılmaya aday K grup adayının yapılandırılması, ikincisi ise K küme sayılarının seçimidir. Ayrıştırılmaya aday satırların K grubu Eşitlik 3.52'de ifade edilen H(.) fonksiyonu temel alınarak yapılandırılmaktadır, ilgili fonksiyon Spearman'ın ρ korelasyon katsayısına dayanan çok değişkenli bir ilişki ölçüsüdür.

$$H\left(\Lambda_2 \mid \Lambda_1\right) = \max_{i' \in \Lambda_2} \left\{ \underset{i \in \Lambda_1}{\psi} \left(\rho\left(x_i, x_{i'}\right) \right) \right\} \tag{3.52}$$

Eşitlik 3.52'de, Λ_1 K grubu oluşturmak için hâlihazırda seçilmiş satır dizin vektörlerinin alt kümesi, Λ_2 kalan aday satır dizin vektörlerinin alt kümesi ψ ortalama, medyan veya maximum kümeleme fonksiyonudur. K'nın seçimine gelince, CoClust'ın klasik kümeleme teknikleriyle ilgili avantajlarından biri, küme sayısının otomatik olarak seçilmesidir. Aslında, CoClust kullanıcı tarafından verilen tüm olasılıkları araştırır ve K'yı araştırmacı tarafından önceden tanımlanmış adımına kadar ayrılan k gruplarının alt kümelerinde tahmin edilen kopulanın log-olabilirlik olasılığına dayanarak seçer (Lascio, 2008).

n satır veri matrisini kümeleme için gerekli CoClust algoritmasının ana adımları aşağıdaki gibi açıklanmıştır (Lascio ve Giannerini, 2019).

1. Kullanıcı tarafından tanımlanan olasılık kümesinde k kümelerinin sayısı değiştirilir, öyle ki $2 \leq k \leq n$,

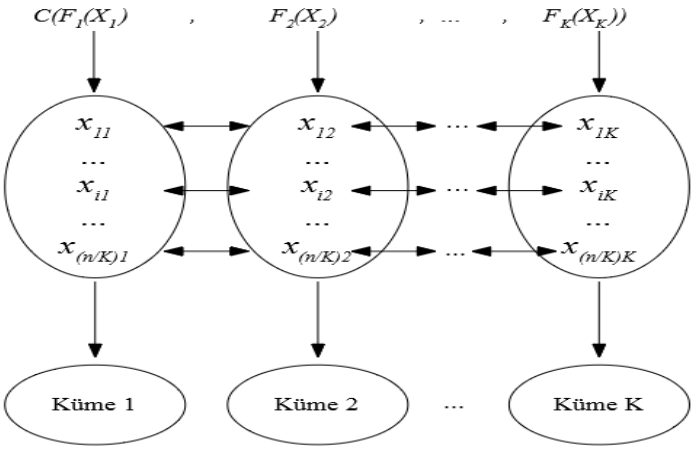
a) Eşitlik 3.52'deki ölçüm temelinde, Eşitlik 3.51'de ifade edilen veri matrisinde satırların k-pletlerinin, n_k alt kümesi seçilir;

b) Kopula modeli satırların k gruplarına, n_k'ya yarı parametrik tahmin modelleri aracılığıyla uygulanır.

2. Satırların k gruplarının, n_k alt kümesi seçilir, kopulayı maksimize eden log-olabilirlik değeri K küme sayısıdır, yani kopulanın boyutudur ve otomatik olarak seçilir. K gruba da n_K denilmektedir ve ilgili kümelere ayrıştırılır.
3. Eşitlik 3.52'de bulunan ölçümü kullanarak, kalanlar arasından yeni bir K grup seçilir ve kopulaların K! permütasyonu hâlihazırda kümelenmiş gözlemleri kullanarak tahmin edilir.

4. Seçilen K grubun permütasyonu kümeleme için, kopula uyumunun log-olabilirliğini artırıyorsa her bir satır ilgili kümeye atanacak şekilde ayrıştırılır, değilse bırakılır.
5. Tüm gözlemler değerlendirilene, yani ya kümelenene ya da atılana kadar 3. ve 4. adımları tekrarlanır.

Algoritmanın sonunda, her biri maksimum (n/K)p adet bağımsız gözlemi içeren kümeler elde edilmiş olur. Böylelikle, kümeler arası bağımlılık ilişkisi ortaya çıkar. Bu nedenle, çok değişkenli ilişkilerin buradaki yapılandırılması klasik kümeleme yöntemlerindeki kümeler içi ilişkilere dayanmamaktadır. CoClust kümelemesinin son hali Şekil 1'de görüldüğü gibi ifade edilmektedir. Her bir küme aynı marjinal dağılımdan elde edilen bağımsız özdeş dağılımlar olmakla birlikte, kümeler arası gözlemler aynı çok değişkenli bağımlılık yapısını paylaşmaktadır (Lascio, 2008).



Şekil 3. 1. CoClust algoritmasının temeli

Algoritmanın her adımında iç içe olmayan modeller test edilmesi sebebiyle, yani tek değişkenli bağımlılık parametresi kullanılarak kopula modelleri ile çalışılması durumu, tanımlanmış log-olabilirlik temelli ölçüt, Bayesyen bilgi kriteri ve Akaike bilgi kriterine eşdeğerdir. Son olarak, K küme sayılarının seçimi n_k gözlemlerinin temsili bir alt kümesine dayanmakta olduğu unutulmamalıdır. Bundan dolayı, algoritma K küme sayısını $\sum_{k=K_{\min}}^{K_{\max}} \binom{n_k}{k}$ gerekli uyumluluğunu tahmin ederek seçmektedir. Burada $[K_{min}, K_{max}]$ aralığı küme sayılarını ifade ederken, $K_{min} \geq 2$ ve n_k ise $n_k \geq p$ eşitsizliklerini sağlamaktadır. Bu durum, örneklem büyüklüğüne bağlı olmaksızın hesaplama karmaşıklığını kontrol altında tutmaya yardımcı olur (Lascio, 2008).

Lascio (2008)'e göre, CoClust algoritması kopula modelinin seçimini otomatik olarak gerçekleştirmek amacıyla kullanılmamaktadır, bir bilgi kriterine *sonsal bilgiye* ihtiyaç duymaktadır. Bayesyen bilgi kriteri K boyutlu kopula modeli *m* için Eşitlik 3.53'de olduğu gibi tanımlanmaktadır.

$$BIC_{K,m} = -2\log\prod_{i=1}^{n} c_m\left\{\hat{F}_1(X_{1i}),\ldots,\hat{F}_k(X_{ki}),\ldots,\hat{F}_K(X_{Ki});\hat{\theta}\right\} + s\log\left((n/K)\,p\right) \quad (3.53)$$

Eşitlik 3.53'de $\hat{\theta}$ Eşitlik 3.54'de ifade edilen toplam ile elde edilmekle birlikte, kümelenmiş gözlemlerin sayısı üzerinden toplamı ifade etmektedir. Bu değer (n/K)p maksimum değerine eşittir. *s* değeri ise parametre sayısını göstermektedir (Lascio ve Giannerini, 2019).

$$\hat{\theta} = \arg\ \max_{\theta}\sum_{i=1}^{n}\log c\left[F_1\left(X_{1i};\hat{\beta}_1\right),\ldots,F_k\left(X_{ki};\hat{\beta}_k\right),\ldots,F_K\left(X_{Ki};\hat{\beta}_K\right);\theta\right] \quad (3.54)$$

Buna göre, BIC değerini minimize eden kopula modeli seçilir. Benzer şekilde, Akaike bilgi kriteri (AIC) Eşitlik 3.55 ile ifade edilir ve kopula modelinin seçiminde kullanılır (Lascio, 2008).

$$AIC_{K,m} = -2\log\prod_{i=1}^{n} c_m\left\{\hat{F}_1(X_{1i}),\ldots,\hat{F}_k(X_{ki}),\ldots,\hat{F}_K(X_{Ki});\hat{\theta}\right\} + 2s \quad (3.55)$$

CoClust tekniğinin gücünün değerlendirilmesi ise verilerin temelindeki gerçek çok değişkenli kümelenme yapısını bulmada CoClust algoritmasının uyum iyiliğinin araştırılmasıdır. Özellikle, CoClust çalışmasının ilk versiyonunda (Lascio, 2008) farklı senaryolarda simüle edilmiş veri üzerinde test edilmiştir ve model tabanlı kümeleme ile karşılaştırılmıştır. Bu karşılaştırma gösteriyor ki, CoClust kümelerin doğru sayısını ve büyüklüğünü pek çok durumda doğru tanımlayabilmektedir. Dahası, model tabanlı kümeleme ile karşılaştırıldığında, CoClust bağımlı verinin modellemesinde daha başarılı bulunmuştur. Lascio ve Giannerini (2019) çalışmasında, CoClust algoritmasının güncel versiyonunun yeni özelliklerini araştıran daha karmaşık bir Monte Carlo çalışması gerçekleştirmiştir. Bu çalışmada, algoritmanın küme sayısını doğru tespit edebilmesi, kopulanın boyutunun değişmesi durumunda k grupların doğru yapılandırılması, ψ kümeleme fonksiyonunun belirlenmesi için kopula modelinin tatmin edici olduğu belirlenmiştir.

Gerçek veri uygulamalarına gelince, CoClust algoritması çeşitli veri setlerine başarıyla uygulanmıştır. Lascio vd. (2017)'de ise tümörlerden ve kanser hücre

çizgilerinden organ tipini belirlemeye çalışmıştır. Biyomedikal uygulamalarla ilgili olarak, Lascio ve Giannerini (2019) çalışmasında genler arası olası fonksiyonel ilişkiyi hipotezlerle formüle edebilmek için CoClust algoritması uygulamıştır. Lascio vd. (2017) çalışması diğer alanlardaki uygulamalara örnek olarak verilebilir. Bu çalışmada amaç, sağlık diyetleri rehberliğinde ve yaygın Avrupa politikaları kapsamında AB ülke diyetlerindeki değişiklikleri analiz etmeyi amaçlamaktadır ve Lascio vd. (2017) çalışmasında yağış ölçümlerinin coğrafik dağılımını araştırmak amacıyla kullanılmıştır.

Bu tekniği benzerlerinden ayıran bir diğer önemli özellik ise, elde bulunan değişkenlerin hepsini kümelemeye tabi tutmamasıdır. Lascio (2008)'in açıkladığı üzere, teknik yalnızca aralarında ilişki olduğunu tespit ettiği değişkenleri kümelemeye tabi tutmaktadır. Yani, elde bulunan tüm değişkenler kümelere yerleştirilmemektedir. İlişkisiz görülen değişkenler kümelerin dışında kalmaktadır. Bu yönüyle kuyruk bağımlılığı tekniğinden de ayrılmaktadır.

Tekniğin bir diğer önemli farkı ise, marjinal ve küme kavramını birbiri yerine kullanmakla birlikte, her bir kümenin aynı marjinal dağılımdan geldiğini, kümeler arası gözlemlerin ise aynı çok değişkenli bağımlılık yapısından geldiğini vurgulamasıdır. Bir başka deyişle, kümelerde bulunan aynı satırdaki değişkenler birbirleriyle ilişkilidir.

Kopula temelli kümeleme algoritması prosedürü R paketi CoClust'ta uygulanmıştır (Lascio, 2008). Buradan yola çıkarak ilgili paket indirilir;

```
R> install.packages(CoClust)
```

ve devamda bulunan kod aracılığıyla yüklenmelidir

```
R> library(CoClust)
```

CoClust paketinin kodları tamamen R ortamında yazılmıştır ve açık kaynak kod sistemi nedeniyle kolaylıkla ulaşılabilirdir.
CoClust paketinin ana fonksiyonu CoClust ( ) kopula tabanlı kümelemeyi uygulamaktadır. Bu fonksiyon temel olarak bazı seçenekler sunmaktadır (Lascio, 2008).

- Çeşitli kopula modellerini (kopula fonksiyonunu kullanarak), marjinaller ve kopulalar için (method.ma ve method.c kullanılır sırasıyla) farklı tahmin prosedürleri ile veriye uyumlu hale getirir. Özellikle, Eliptik ve Arşimedyan tüm

kopula modelleri, iki farklı varyans tahmin edicisine dayanan En Çok Sözde Olabilirlik tahmin yöntemleri aracılığıyla ve Kendall'ın τ tahmincisinin ve Spearman'ın ρ tahmin edicisinin tersine çevrilmesi ile R paketinde uygulanır. Marjinallere gelince ise, parametrik ve parametrik olmayan olmak üzere iki farklı tahmin yöntemi uygulanmaktadır. Bu yöntemler En Çok Olabilirlik Yöntemi ve Ampirik Kümülatif Dağılım Fonksiyonudur.

- Kopula modeli için, boyut setini veya aralığı ayarlanır, yani küme sayısı belirlenir (dimset aracılığıyla).
- Küme sayısını seçmek için kullanılan örneklem birimlerinin boyutu belirlenir (noc aracılığıyla).
- Ortalama, medyan veya maksimum arasından k-pletleri seçmek için kullanılan Eşitlik 2'deki Spearman'ın ρ korelasyonunun ikili kombinasyon fonksiyonları seçilir (fun aracılığıyla).
- AIC, BIC ve log-olabilirlik kriteleri kullanılarak küme sayısı belirlenir (penalty aracılığıyla).

"En iyi" modelin seçimine gelince, CoClust ilgilenilen modellerin türünü değiştirerek çalıştırılabilir ve belirlenen kriterlerden birine göre "sonsal" duruma en uygun olan seçilir (Lascio ve Giannerini, 2019).

Tipik bir CoClust fonksiyonunun kullanımı aşağıdaki gibidir.

CoClust(m, dimset = 2:5, noc = 4, copula = "frank", fun = median, method.ma = c("empirical", "pseudo"), method.c = c("ml", "mpl", "irho", "itau"), dfree = NULL, writeout = 5, penalty = c("BICk", "AICk", "LL"), …)

Burada "m" tüm veri matrisini ifade etmektedir ve "writeout" komutu ayrıştırma sürecini göstermektedir, çünkü her yeni ayrıştırılan gözlem hakkında bilgi vermektedir.

CoClust fonksiyonu üzerinden elde edilen çıktılar aşağıdaki gibi ifade edilebilir (Lascio, 2008).

1. Küme Sayısı: Seçilmiş ve tanımlanmış K küme sayısı
2. İndeks Matrisi: *n.göz x (K+1)*matrisidir. *n.göz* ise her bir kümeye konulan gözlemlerin sayısıdır. İndeks matrisi gözlemlerin veri matrisinin satır indekslerini içermektedir ve son sütun kopulanın log-olabilirliğini içermektedir.

3. Veri Kümeleri: Verinin kümelenmiş halidir; her bir sütun kümelenmiş gözlemleri içerir.
4. Bağımlılık: Aşağıdaki listeyi içeren bir çıktı oluşmaktadır.
   a. Model: Kümeleme için kullanılan kopula modeli
   b. Param: Kümeler arasındaki tahmini bağımlılık parametresi
   c. Std.Err: Param çıktısının standart hatası
   d. P.val: $H_0 : \theta = 0$ hipotezi ile ilişkili p-değeri
5. LogLik: Uyumlu kopulanın maksimum log-olabilirlik değeri
6. Est.Method: Kopulanın uyumlu hale getirilebilmesi için kullanılan tahmin yöntemi
7. Opt.Method: Kopulanın uyumlu hale getirilebilmesi için kullanılan optimizasyon yöntemi
8. LLC: dimset 'te her bir k için log-olabilirlik kriteri değeri
9. Index.dimset: dimset 'te her bir k için, küme sayısını seçmek için kullanılan n_k başlangıç setinin indeks matrisini içermektedir. Uyumlu kopulanın maksimum log-olabilirlik değerini de içerir.

Doğrusal olmayan ve parametrik olmayan bağımlılık yapısı üzerine kurulu olan CoClust tekniği algoritma ve uygulaması tanıtılmıştır. Benzemezlikler üzerine kurulu kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniği izleyen bölümde tanıtılmıştır.

3.5.2. Kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme

Kuyruk bağımlılığı kavramı, olasılık teorisinde değişkenlerin dağılım kuyruğunda birlikte hareketlerini incelemektedir. Özellikle Uç Değer Teorisi'nde sıklıkla tercih edilmektedir (Hartmann, 2004).

İki rassal değişken arasındaki ilişkiye dayanan benzerlik oluşturmak yerine, bir ortak dağılım fonksiyonunun, kuyruk hareketi gibi belirli özellikleri dikkate alınabilir. Bu yaklaşıma, Luca ve Zuccolotto (2011) ile giriş yapılarak yine Luca ve Zuccolotto'nun 2014 ve 2017 yıllarındaki çalışmaları ve Durante vd. (2015)'nin çalışması ile geliştirilmiştir. Bir *C* kopulası ile ilişkili kuyruk bağımlılık katsayılarına (alt λ_L ve üst λ_U) dayanmaktadır. Alt kuyruk bağımlılığı Eşitlik 3.56 ve üst kuyruk bağımlılığı Eşitlik 3.57 ile tanımlanmaktadır (Lascio vd., 2017).

$$\lambda_L = \lim_{u\to 0^+} \frac{\delta_C(u)}{u} = \lim_{u\to 0^+} \frac{P(X \le F_X^{-1}(u), Y \le F_Y^{-1}(u))}{P(X \le F_X^{-1}(u))} = \lim_{u\to 0^+} \frac{C(u,u)}{u} \tag{3.56}$$

$$\lambda_U = \lim_{u\to 1^-} \frac{1-2u+\delta_C(u)}{1-u} = \lim_{u\to 1^-} \frac{P(X > F_X^{-1}(u), Y > F_Y^{-1}(u))}{P(X > F_X^{-1}(u))} = \lim_{u\to 1^-} \frac{C^*(1-u,1-u)}{1-u} = \lim_{u\to 1^-} \frac{1-2u+C(u,u)}{1-u} \tag{3.57}$$

$\delta_C(t):[0,1]\to[0,1]$ olmak üzere $\delta_C(t)=C(t,t)$ eşitliği C kopulasının köşegenini ifade etmektedir ve üst limit söz konusudur. Bu durumda yaygın yaklaşım, formül aracılığıyla X ve Y rassal değişkenleri arasındaki benzerliği (benzemezliği) tanımlamaktır. Benzemezlik matrisi Eşitlik 3.58 ile elde edilmektedir (Lascio vd., 2017).

$$diss(X,Y) = -\log(\lambda_L) \text{ veya } diss(X,Y) = -\log(\lambda_U) \tag{3.58}$$

Bu nedenle, eğer X ve Y birlikte monoton harekete sahip ise $\lambda_L=\lambda_U=1$ ve $diss(X,Y)=0$'dır ancak X ve Y asimptotik olarak bağımsız ise $diss(X,Y)=+\infty$ ve $\lambda_L=\lambda_U=0$ 'dır.

Kuyruk bağımlılık katsayıları, kopulanın kuyruk davranışının asimptotik yaklaşımını vermektedir. Birim karenin köşelerine yaklaşan bazı noktalarda kuyruk bağımlılığını dikkate almak da bu anlamda önemli olabilir. Bu konuda Sweeting ve Fotiou (2013) ve Durante vd. (2015) incelenebilir.

Bu amaçla, $q_C:(0,1)\to[0,1]$ verildiğinde Eşitlik 3.59'daki gibi tanımlanmış sözde *kuyruk yoğunluk fonksiyonu* dikkate alınabilir (Lascio vd., 2017).

$$q_C(t) = \frac{\delta_C(t)}{t}.1_{(0,0.5]} + \frac{1-2t+\delta_C(t)}{1-t}.1_{(0.5,1)} \tag{3.59}$$

Burada 1_S, S kümesinin gösterge fonksiyonunu göstermektedir. Durante vd. (2015)'de birlikte monoton harekete sahip kopulaların kuyruk yoğunluk fonsiyonu ve q_C arasındaki uygun uzaklıklar için kullanılmaktadır.

Bir başka benzer yol ise, kuyruk bağımlılık katsayısının bileşik dağılımın tanım kümesindeki kuyruk bölgelerine odaklanan yerel bir uyum ölçütü ile değiştirilmesidir.

Örneğin Durante vd. (2014)'de, *(X,Y)* ikilisinin Spearman'ın koşullu bağımlılık değeri ρ kullanır ve her ikisi de küçük değerler almak şartıyla, *X* ve *Y* rassal değişkenleri arasındaki bağımlılığın derecesi olarak yorumlanabilir tespiti yapılmıştır (Lascio vd., 2017).

Teknik uygun bir uzaklık matrisi $\Delta = \delta_{ij}$ oluşturmaya dayanmaktadır.

- Her i, j için $\delta_{ij} \geq 0$
- Her i, j için $\delta_{ij} = \delta_{ji}$
- Her i için $\delta_{ii} = 0$

CoClust'tan farklı olarak kopula ailesi seçiminde bilgi kriteri değil, araştırmacının tecrübesi ve veri yapısı önem kazanmaktadır. Ancak, popüler ve iyi çalışılmış kopula ailesi olan Gaussian kopula ile birlikte sıklıkla tercih edilmektedir (MacKenzie ve Spears, 2014).

Benzemezlik matrisi elde edildikten sonra, Kaufman ve Rousseeuw (1990) tarafından tarif edildiği gibi, hiyerarşik kümeleme (R: hclust) veya "bulanık" kümeleme (R: fanny) gibi standart kümeleme teknikleri uygulanabilir (Lascio vd., 2017).

Algoritma:

1) Her bir değişkenin marjinal davranışının etkisi olmadan, ilgilenilen değişkenler arasındaki bağa dikkat ederek uygun kopula modeli seçilir.
2) Farklı değişkenler arasındaki kuyruk bağımlılığı için uzaklık matrisi tespit edilir.
3) Uygun bir küme algoritması ile kümeleme uygulanır.

Kaufman ve Rousseeuw (1990) kopula temelli kuyruk bağımlılığı için önerdiği kümeleme yöntemlerden en sık kullanılanı hiyerarşik kümeleme yöntemidir.

Hiyerarşik kümeleme adından anlaşılacağı üzere bir kümeleme yöntemidir. K-ortalama temelli kümeleme algoritmalarında araştırma başlangıcında küme sayısının belirlenmesi önemlidir. Buna karşılık, hiyerarşik kümeleme yöntemlerinde böyle bir şart bulunmamaktadır. Bunun yerine, değişkenler arasındaki ikili benzemezliklere dayanan benzemezlik matrisinin elde edilmesi önemlidir. Adından da anlaşılacağı üzere, her seviyedeki kümeleri bir alt seviyedeki kümelerle birleştirerek hiyerarşik gösterim elde edilmektedir. En yüksek seviyede tüm değişkenleri içeren sadece bir küme bulunurken, en düşük seviyede her bir küme yalnızca bir değişken içermektedir (Friedman vd. 2009).

Hiyerarşik kümeleme yöntemleri iki temel başlığa ayrılmaktadır. İlki aşağıdan yukarı ilerleyen *yığınsal* (agglomerative), ikincisi ise yukarıdan aşağıya ilerleyen *ayrıştırıcı* (divisive) yöntemdir. Yığınsal yöntemde, her bir değişken bir kümeyi oluşturmaktadır. Birbirine yakın kümeler birleştirilerek yeni bir küme oluşturulur. Bu işlem, sistem kararlı oluncaya dek tekrarlanır. Ayrıştırıcı yöntemde ise, yığınsal yöntemin tersine tüm değişkenler tek bir küme oluşturmaktadır ve uzaklık durumuna göre küme parçalanarak alt kümelere ayrıştırılmaktadır (Friedman vd. 2009).

Hem yığınsal hem de ayrıştırıcı yöntem monotonluk özelliğine sahiptir. Yani, küme kapsamı genişledikçe uzaklık oranı artmaktadır. Bu durum gruplar arası ilişkiyi gösteren ağaç grafik yöntemiyle ifade edilmektedir. Bu ağaç grafiğine *dendrogram* denilmektedir.

Yığınsal yöntem her bir değişkenin birer kümeyi oluşturmasıyla başlamaktadır. Adımların her birinde en yakın iki küme (en az farklı) bir kümede birleştirilerek, sonraki adıma daha az kümeyle geçilir. Bu nedenle iki küme arasında bir uzaklık ölçüsü tanımlanmalıdır. Örneğin *G* ve *H* iki grubu tanımlasın. *G* ve *H* arasındaki uzaklık *d(G, H),* değişkenlere ait *i* ve *i'* gözlemleri incelenerek elde edilmektedir. Bu uzaklık matrisinin elde edilmesi için kullanılan aşağıdaki gibi uzaklık formülleri mevcuttur (Friedman vd., 2009).

-Tek Bağlantı (Single Linkage): İki küme arasındaki en yakın mesafeyi hesaplayarak kümelemeyi hedeflemektedir. Eşitlik 3.60 ile gösterilmektedir.

$$d_{ij} = \min_{ij}\left(X_i, X_j\right) \tag{3.60}$$

-Tam Bağlantı (Complete Linkage): İki küme arasındaki en uzak mesafeyi hesaplayarak kümelemeyi hedeflemektedir. Eşitlik 3.61 ile gösterilmektedir.

$$d_{ij} = \max_{ij}\left(X_i, X_j\right) \tag{3.61}$$

-Ortalama Bağlantı (Average Linkage): İki küme arasındaki ortalama mesafeyi hesaplayarak kümelemeyi hedeflemektedir. Eşitlik 3.62 ile gösterilmektedir.

$$d_{ij} = \frac{1}{k.l}\sum_{i=1}^{k}\sum_{j=1}^{l} d\left(X_i, X_j\right) \tag{3.62}$$

-Ağırlık Merkezi (Centroid Method): İki küme arasındaki ağırlık merkezlerinin uzaklığı incelenerek kümelemeyi hedeflemektedir. Eşitlik 3.63 ile gösterilmektedir.

$$d_{ij} = \left\| \overline{X}_i - \overline{X}_j \right\|_2 \tag{3.63}$$

-Ward Yöntemi (Ward's Method): Küme içi toplam varyansı en küçük hale getirerek kümelemeyi hedeflemektedir. Eşitlik 3.64 ile gösterilmektedir.

$$d_{ij} = d\left(\{X_i\}, \{X_j\}\right) = \left\| X_i - X_j \right\|^2 \tag{3.64}$$

Hiyerarşik kümelemenin programlama algoritması aşağıdaki gibi ilerlemektedir (Hardle ve Simar, 2009).

1. Uzaklık matrisi D hesaplanır.
2. En yakın mesafedeki iki küme bulunur.
3. Bulunan bu iki küme tek bir kümeye yerleştirilir.
4. Yeni oluşturulan gruplar arası uzaklıklar hesaplanır ve küçültülmüş uzaklık matrisi D elde edilir.
5. Tüm kümeler yığınlaştırılıncaya dek devam edilir.

Buradan yola çıkarak, uzaklık matrisleri kullanılarak yığınsal yöntem aracılığıyla kümeleme yöntemi sonuçlandırılmaktadır.

Bağımlılık ölçüleri, kopulalar, kopula tahmin yöntemleri ve kopula temelli kümeleme tekniklerinin incelenmesinin ardından modelleme aşamasında kullanılacak olan Lojistik Regresyon Analizi bir sonraki bölümde tanıtılacaktır.

3.6. Lojistik Regresyon Analizi

Regresyon analizi, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi inceleyerek bağımlı değişkeni değerlendirmede en sık kullanılan yöntemlerden birisidir. Bağımsız değişkenler sürekli ve kesikli yapıda olabilirken bağımlı değişken iki veya ikiden fazla kategorili ise doğrusal regresyon analizine alternatif olarak Lojistik Regresyon Analizi önerilmektedir.

Türker (2016), Lojistik Regresyon Analizinin ilk kez 19. yüzyılda Quetelet ve Verhulst tarafından nüfus değişim hızını bulmak için kullanıldığını ifade etmektedir. 1920'de ise Pearl ve Reed "uzun zaman önce unutulmuş" diyerek Verhulst'un çalışmasını tekrarlayarak Lojistik Regresyon Analizini tekrar gündeme getirmiştir. Ancak lojistik ismi

1925'te Yule'un çalışması ile canlandırılmıştır. Son olarak, Cox ve Snell (1989), Lojistik Regresyon Analizinin Diskriminant Analizi ve Loglineer model ile benzerliğini açıklayarak, geçmişe ait vaka kontrollerinde kullanılabileceğini makalelerinde açıklamıştır.

Lojistik Regresyon Analizinde bağımlı değişken ikili olarak ifade edilmekte ve bağımsız değişken kategorik veya nümerik olabilmektedir. Bağımlı değişkenin ikiden fazla seçenekli olduğu durumda ise çoklu lojistik regresyon analizi tercih edilmektedir (Agresti, 1990; Hosmer ve Lemeshow, 2000).

"Lojistik Regresyon Analizi, çok değişkenli normallik ve varyansların homojenliği gibi varsayımları olan diskriminant analizine alternatif bir yöntem olmaktadır. Ayrıca çoklu doğrusal regresyonda, elde edilen çoklu doğrusal denklemindeki bağımsız değişkenlere ilişkin katsayıları/ağırlıkları yardımıyla bağımlı değişkenin gerçek değeri kestirilirken, lojistik regresyonda bağımlı değişken kategorilerinden birine atanma olasılığı elde edilir. Dolayısıyla bağımlı değişkene ilişkin olasılıklar 0-1 arasında değişir." (Alpar, 2011).

Takip eden bölümlerde Lojistik Regresyon modeli, anlamlılık için kullanılan istatistikler, modellerin değerlendirilmesi ve değişken seçimi adımları açıklanacaktır.

3.6.1. Lojistik regresyon modeli

Basit doğrusal regresyon modeli Eşitlik 3.65'te ifade edilmektedir. Eşitlik 3.65'te Y_i bağımlı değişkeni, X_i bağımsız değişkeni ve e_i rassal hata terimini göstermektedir.

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, 2, ..., n \tag{3.65}$$

Bağımlı değişkenin 0 ve 1 değerlerini alarak Bernoulli dağılımına uyduğu durumlar incelenmiştir. Lojistik modele ulaşmak amacıyla doğrusal modelin beklenen değeri Eşitlik 3.66'da gösterilmektedir. Rassal hata teriminin beklenen değeri ise 0'dır.

$$E(Y_i) = E(\beta_0 + \beta_1 X_i) + E(e_i) = \beta_0 + \beta_1 X_i \tag{3.66}$$

$Y_i=1$ olma olasılığı $\pi(x_i)$ ise $Y_i=0$ olma olasılığı $1-\pi(x_i)$ olmaktadır. Buradan beklenen değer Eşitlik 3.67 haline dönüşür.

$$E(Y_i) = 1.\pi(x_i) + 0.(1-\pi(x_i)) = \pi(x_i) \tag{3.67}$$

Eşitlik 3.66 ve Eşitlik 3.67'den, Eşitlik 3.68 elde edilir.

$$\pi(x_i) = \beta_0 + \beta_1 X_i \tag{3.68}$$

Y_i, 0 ve 1 değerlerini alırken, Y_i'nin beklenen değeri P(Y_i=*1)*'e eşittir. X=x değeri bilindiğinde Y'nin koşullu beklenen değeri Eşitlik 3.69'a eşit olmaktadır (Hosmer ve Lemeshow, 2000).

$$E(Y \mid x) = \beta_0 + \beta_1 X_i = \pi(x) \tag{3.69}$$

Lojistik regresyon modelinin özel formu ise Eşitlik 3.70'de gösterilmektedir (Hosmer ve Lemeshow, 2000).

$$\pi(x_i) = E(Y \mid x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \tag{3.70}$$

Lojistik regresyon çalışmasının merkezini $\pi(x_i)$ *lojit dönüşümü* oluşturmaktadır. Bu dönüşüm $\pi(x_i)$ açısından Eşitlik 3.71 ile gösterilmektedir (Hosmer ve Lemeshow, 2000).

$$\begin{aligned} g(x_i) &= \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) \\ &= \ln(\pi(x_i)) - \ln(1-\pi(x_i)) \\ &= \ln\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) - \ln\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right) \\ &= \ln\left(e^{\beta_0+\beta_1 x_i}\right) \\ &= \beta_0 + \beta_1 x_i \end{aligned} \tag{3.71}$$

Eşitlik 3.71'de oranı bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı ile aralarındaki ilişkiyi göstermektedir. Bu orana *odds oranı* denilmektedir. Referans duruma göre ilgilenilen durumun olasılığı ile ilgili bilgi vermektedir (Hosmer ve Lemeshow, 2000).

Çoklu lojistik regresyon modeli ise bağımsız değişken sayısının birden fazla olduğu durumlarda kullanılmaktadır ve süreç basit lojistik regresyon adımlarına benzer olarak ilerlemektedir (Hosmer ve Lemeshow, 2000).

$x_i = (x_{i1}, x_{i2}, ..., x_{ip}),\ i = 1, 2, ..., n$ olmak üzere model Eşitlik 3.72 ile gösterilebilir.

$$\pi(x_i) = E(Y \mid x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}}} \tag{3.72}$$

Modelin lojit dönüşümü ise Eşitlik 3.73 ile ifade edilmektedir.

$$\begin{aligned} g(x_i) &= \ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \ln\left(\frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}}}\right) - \ln\left(\frac{1}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}}}\right) \\ &= \ln\left(e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \end{aligned} \tag{3.73}$$

Modelin genel tanımı ve elde edilişinin ifade edilmesinin ardından bir diğer önemli adım model anlamlılığının doğru şekilde tespit edilmesidir. Belirlenen modelin kullanımı anlamlılığın ölçülmesiyle doğrudan ilişkilidir. Bu nedenle, modelin belirlenmesi kadar önemli adımlardandır.

3.6.2. Modellerin değerlendirilmesi

Elde edilen anlamlı modellerin hangisinin verileri daha iyi özetlediğini tespit etmek amacıyla kullanılacak ölçütlere ihtiyaç vardır. Doğrusal regresyon analizinde kullanılan R^2 istatistiğine benzeyen bir istatistik Lojistik Regresyon Analizinde bulunmamaktadır.

Türker (2016)'da, R^2'nin bağımlı değişkenin açıklanan varyansının yüzdesini gösterdiğini belirterek, bağımlı değişkenin varyansının bu değişkenin olasılık dağılımına bağlı olduğunu ifade etmiştir. İki sonuçlu bağımlı değişkenin varyansının grup frekansları eşit olduğu zaman maksimum olacağını, bu nedenle regresyon analizindeki R^2 değerinden farklı olduğunu açıklamıştır.

Öte yandan, lojistik regresyonda model uygunluğunu tespit etmek için kullanılan R^2 değerleri de bulunmaktadır. Bu istatistikler aracılığıyla, modelin veriyi özetleme becerisi hakkında bilgi sahibi olunmaktadır.

i. Nagelkerke R^2: Cox ve Snell R^2 istatistiğinin [0,1] aralığında değerler alarak yorumlanmasını kolaylaştırmak amacıyla geliştirilmiştir.

ii. Cox ve Snell R^2: Olabilirlik esasına dayanmaktadır. Maksimum değerinin genellikle 1'den küçük olması nedeniyle yorumlanması zor olmaktadır.

Elde edilen modellerin anlamlılığı ve geçerliliğinin değerlendirilmesi ile Nagelkerke R^2 değeri araştırmacıya modelin kullanımı ile ilgili yol gösterici sonuçlardan birisidir. Bu nedenle, Lojistik Regresyon modellemesinde sıklıkla kullanılmaktadır.

3.6.3. Değişken seçimi

Lojistik Regresyon modellemesinde en önemli adımlardan bir diğeri ise modele değişken seçimidir. Doğrusal regresyon analizinde olduğu gibi bütün katsayıların sıfıra eşit olup olmadığı hakkında yargıya varmak için ileriye doğru, geriye doğru ve adımsal yöntem Lojistik Regresyon Analizinde de uygulanmaktadır.

Burada değişken seçimi χ^2'ye uygun Wald test istatistikleri aracılığıyla yapılmaktadır. Wald değerine bakarak önemsiz olduğuna karar verilen değişkenler modelden çıkartılarak kalan değişkenlerden oluşan modelin istatistiğine tekrar bakarak kesin karar verilir.

3.6.3.1. <u>İleriye doğru seçim yöntemi (Forward stepwise)</u>

Modelde yalnızca sabit terimin bulunmasıyla incelemeye başlanır. Değişken eklenerek ilgili istatistik aracılığıyla değişkenin modele ayırıcı gücü artırma bakımından katkısı incelenir. Bu şekilde değişkenler eklenip incelenerek, eklenecek değişken kalmayana dek tekrarlanır.

3.6.3.2. <u>Geriye doğru seçim yöntemi (Backward stepwise)</u>

İleriye doğru seçim yönteminin tersi şekilde başlangıçta bütün değişkenler modelde bulunur. Her adımda modelin ayırıcı gücünü en az şekilde azaltacak şekilde değişken çıkartılarak ilgili istatistik aracılığıyla incelenir. Daha fazla değişken çıkartılamayana dek yöntem tekrarlanır.

3.6.3.3. Adımsal yöntem

Adımsal yöntem ise ileriye doğru seçim ve geriye doğru seçim yönteminin bileşimi şeklinde ilerlemektedir. Modelde değişken bulunmazken başlar, ilk adımda değişken eklenir ve ilerleyen her adımda ya değişken eklenir ya da çıkartılır.

Lojistik regresyon modellemesinde, denkleme eklenecek bir değişken ile istatistik ölçütlerine göre ayırıcı güçte anlamlı bir artış oluşmazsa eşitlikten çıkartılır. Türker (2016) bu durumu, "Eğer ilgili adımda herhangi bir değişken atılmazsa o zaman istatistik ölçüte göre anlamlı en fazla ayırıcı gücü ekleyen değişken lojistik regresyon denklemine dâhil edilir. Eğer verilen adımda lojistik regresyon denklemine değişken eklenmiyor veya çıkarılmıyorsa işlem durur" şeklinde açıklamaktadır.

Lojistik Regresyon analizinde değişken seçimi ile modelin oluşturulması, model anlamlılıklarının belirlenmesi kullanılacak modelin seçimi ile ilgili oldukça fazla bilgi vermektedir. Ancak, belirlenen modelin çeşitli yollarla geçerliliklerinin değerlendirilmesi model kullanımında daha güvenli yol açmaktadır. Hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi model seçiminde önemli yol göstericilerdir.

3.7. Modellerin Geçerliliklerinin Değerlendirmesi

Model değerlendirme aşaması, model geliştirme sürecinin önemli ve dikkatle incelenmesi gereken adımlarından birisidir. Verilerimizi temsil eden en iyi modeli ve seçilen modelin gelecekte ne kadar iyi çalışacağına dair yorum yapmada model değerlendirme yöntemleri önem arz etmektedir.

Modellerin elde edilme aşamasında kullanılan anlamlılık değeri, Nagelkerke R^2 veya Hosmer-Lemeshow testleri ile incelenerek ilerlenir. Elde edilen bu değerler model hakkında anlamlılık, uygunluk gibi fikirler verirken elde edilen modelin geçerliliği ile ilgili hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi değerleriyle tespit edilmektedir (Trehan ve Joshi, 2018).

3.7.1. Hata matrisi

Yapay zekâ alanında ve istatistiksel sınıflama problemlerinde hata matrisi, bir modelin performansının incelenmesi için yardımcı olan özel bir tablodur (Stehman, 1997).

Matrisin satırları model aracılığıyla elde edilen tahmin sonuçlarını gösterirken, sütunları gözlem sonucundaki elde edilen sonuçları göstermektedir veya tam tersidir (Powers, 2011).

Hata matrisi, ROC eğrisi başlığında ifade edilen tahmin sonucu elde edilen ayırıcı tablodur. İlgili tablo hata matrisi olarak da kullanılmaktadır ve Çizelge 3.1'de gösterilmektedir (Powers, 2011).

Çizelge 3.1. Hata matrisi

		Gözlenen		**Toplam**
		H+	**H-**	
Tahmin Edilen	**H+**	Gerçek Pozitif (GP = A)	Yanlış Pozitif (YP = B)	T(T+)
	H-	Yanlış Negatif (YN = C)	Gerçek Negatif (GN = D)	T(T-)
Toplam		G (H+) = A+C	G (H-) = B+D	N

Doğru sınıflandırma oranı olarak tanımlanabilecek olan *doğruluk,* toplam tahminler içerisinde pozitif ve negatif tahminlerin ayırıcı gücünü göstermektedir ve Eşitlik 3.74 ile ifade edilmektedir.

$$Doğruluk(Accuracy) = \frac{A+D}{N} = \frac{GP+GN}{Toplam} \tag{3.74}$$

Yanlış sınıflandırma oranı olarak tanımlanabilecek olan *hata oranı*, gözlenen pozitif iken tahminin negatif; gözlenen negatif iken tahminin pozitif olması durumundaki hata oranını göstermektedir. Eşitlik 3.75 ile hesaplanmaktadır.

$$Hata\ Oranı(Error\ Rate) = \frac{B+C}{N} = \frac{YP+YN}{Toplam} \tag{3.75}$$

Modele dair doğruluk, özgüllük gibi değerleri vererek model geçerliliğini sorgulamaya yardımcı olan hata matrisi değerlendirmesinden sonra bir diğer önemli geçerlilik ölçütü çapraz geçerlilik ölçütüdür.

3.7.2. Çapraz geçerlilik ölçütü

Rotasyon aracılığıyla tahmin yöntemi olarak da adlandırılan çapraz geçerlilik ölçütü, örneklem dışı istatistiksel analiz yöntemlerinin geçerliliğini inceleyen çeşitli yöntemlerden birisidir (Geisser, 1993; Kohavi, 1995; Devijver ve Kittler, 1982).

Çapraz geçerlilik ölçütü model geçerliliklerinin değerlendirilmesi için en çok kullanılan tekniklerden birisidir ve artık tabanlı ölçümlerden daha iyi bir teknik olarak kabul edilmiştir. Benzetimin ve olasılıklıklı örneklemenin çalışmacıya daha fazla yol sunduğu kabul edilmiştir (Ramasubramanian ve Singh, 2016).

Ramasubramanian ve Singh (2016) aynı çalışmada K-katlı çapraz geçerlilik ölçütünün yapay zekâda oldukça önemli olduğunu vurgulamaktadır. Büyük sayılar kanununa benzeterek, kat sayısının artmasıyla yorumlamanın daha isabetli olacağını belirtmiştir. Elde edilen tahminin varyansı, K arttıkça azalmaktadır. Çapraz geçerlilik ölçütünün adımları aşağıdaki gibidir.

Adım 1: Veri seti K alt gruba ayrılır.

Adım 2: K-1 alt gruba model uygulanır.

Adım 3: Kalan bir alt gruba da model uygulanır ve hata tespit edilir.

Adım 4: Tüm alt kümeler için model kaydırılarak uygulanacak şekilde 1’den 3’e kadar tüm adımlar tekrarlanır.

Adım 5: Simülasyon hataları ortalaması alınarak çapraz geçerlilik sonucu elde edilir.

Elde edilen sonuçlar kapsamında belirlenen doğruluk ve Kappa değerleri üzerinden modelin geçerliliği incelenmektedir. Kappa veya Cohen’ın kappa katsayısı, gözlemlenen doğruluk ile beklenen doğruluk arasındaki ilişkiyi ölçen bir istatistiktir. Jacob Cohen, Kappa'yı 1960 yılında Eğitim ve Psikolojik Ölçüm Dergisi'nde yayınlanan bir makalede

tanıtmıştır. Benzer bir istatistik Pi ismiyle 1955'te Scott tarafından tanıtılmıştır ancak olasılıkların hesaplanması bakımından farklıdır. Ancak bu yöntem, değerlendiriciler arası anlaşmalarda önemli bir yer bulmuştur. Kappa ölçüt sınırları Çizelge 3.2'de ifade edilmektedir (Ramasubramanian ve Singh, 2016).

Çizelge 3.2. Kappa ölçüt sınırları

0 - 0,20	Zayıf
0,20 - 0,40	Vasat
0,40 - 0,60	Orta
0,60 – 0,80	İyi
0,80 – 1,00	Çok İyi

Kappa ölçütü ile modelin gücü hakkında fikir edinmekle birlikte, modelin ayırım gücüyle ilgili olarak özellikle klinik çalışmalarda en önemli adımlardan birisi ROC eğrisi analizidir.

3.7.3. ROC eğrisi analizi

Alıcı işletim karakteristik eğrisi (A receiver operating characteristic curve), yani ROC eğrisi, ikili sınıflayıcının tanılama yeteneğini gösteren çizgisel bir grafiktir.

İstatistiksel karar kuramı ile sinyal algılama kuramının temel ilkelerine dayanarak ortaya çıkmış bir yöntemdir. İkinci Dünya Savaşı sürecinde radar sinyallerinin isabetli ve hatalı olanlarını ayırt edebilme amacıyla geliştirilmiş bir analiz yöntemidir (Egan, 1975; Green ve Swets, 1988).

Daha sonra psikoloji, tıp, radyoloji, biometrik, doğal afetlerin öngörüsü, meteoroloji alanlarında model değerlendirilmesinde sıkça tercih edilmekle birlikte, veri madenciliği ve yapay öğrenmede giderek önem kazanmıştır (Murphy, 1996; Peres ve Cancelliere, 2014; Peres vd., 2015).

ROC eğrisi, çeşitli kesim noktalarına göre *gerçek pozitif orana* (TP) karşılık *yanlış pozitif oranının* (FP) çizilmesiyle elde edilmektedir. Gerçek pozitif orana *duyarlılık* veya *belirleme olasılığı* denilmektedir. Yanlış pozitif oran ise yanlış sonucun olasılığı olarak ifade edilebilir ve (*1-özgüllük*) ile hesaplanmaktadır (Giancristofaro ve Salmaso, 2003).

Özdamar (2003)'te ROC eğrisi grafiğinin düşey ekseninde *duyarlılık,* yatay ekseninde *özgüllük* değerlerinin bulunduğunu belirtmektedir. Seçilen kesim noktalarına göre tespit edilen farklı duyarlılık ve özgüllük değerlerine göre ROC eğrileri oluşturulmaktadır. Keçeoğlu vd. (2016) ise, ROC eğrisinin en yüksek doğruluk veren kesim noktasını (cut-off) belirlediğini ifade etmektedir. Eğri aracılığıyla duyarlılık ve özgüllük arasında optimal bir ilişkinin sağlanması ile kesme değerinin belirlenmesi hedeflenmektedir.

Bu yöntemde, araştırmacıya yol gösteren en önemli kavram eğri altında kalan alandır. Eğri altında kalan alan, testin bireyleri ayırt etme derecesini bir sayısal sonuçla özetleyerek modelin gücü ile ilgili olarak çalışmacıya yol göstermektedir. Bu alanın alabileceği en küçük değer 0,50 iken en büyük değer 1,00'dır. Alanın 1,00 olması, seçilen kesim noktası ile gerçek sonucun %100 uyumlu olduğunu göstermektedir (İyisoy, 2014).

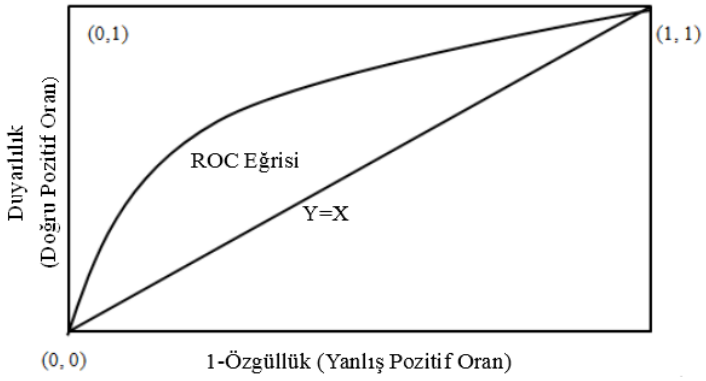
Araştırmanın amacına göre ROC eğrisi için kestirimlerin farklı adlandırılması söz konusudur. Bu adlandırma Çizelge 3.3'de ifade edilmektedir.

Çizelge 3.3. ROC eğrisi bileşenleri için adlandırma

Bağımsız değişken	**Bağımlı değişken**	**Bağımlı değişkenin kestirim / öngörü değerleri**
Kestirici Belirtici Skor Öngörü	Sonuç Durum Altın standart Gösterge	Gözlem - Kontrol Hasta - Sağlıklı Pozitif – Negatif Var - Yok İsabetli - İsabetsiz

Kaynak: Mithat Gönen, Analyzing Receiver Operating Characteristic Curves Using SAS, Cary, NC: SAS Press, 2007.

ROC eğrisi görselleştirilmesinde karar verici için önemli noktalardan birisi ROC eğrisi altında kalan alandır. Eğri altında kalan alan modelin geçerliliği ve ayırıcı başarısı konusunda çalışmacıya ışık tutmaktadır. Örnek bir ROC eğrisi Şekil 3.2'de gösterilmektedir.



Şekil 3. 2. ROC Eğrisi

ROC eğrisi altında kalan alan (AUC) modelin ayırıcı gücü ile ilgili bilgi vermektedir. Genel kural Çizelge 3.4 'de ifade edildiği gibidir (Hosmer ve Lemeshow, 2000).

Çizelge 3.4. AUC değerlerine göre modelin ayırım gücü

AUC ≥ 0,9	Üstün ayırım gücü
0,9 > AUC ≥ 0,8	Mükemmel ayırım gücü
0,8 > AUC ≥ 0,7	Kabul edilebilir ayırım gücü
0,7 > AUC > 0,5	Zayıf ayırım gücü
AUC = 0,5	Ayırım gücü yok

Kaynak: Özdamar, 2003

Elde edilen modelin ayırıcı gücünü gösteren bir tahmin tablosu oluşmaktadır. Bu tahmin tablosu aracılığıyla ROC eğrisine bağlı duyarlılık ve özgüllük değerleri hesaplanabileceği gibi hata matrisi adımında kullanılarak doğruluk ve hata oranı da hesaplanabilmektedir.

Duyarlılık, pozitif gözlenen durumun sonuçlarının tahmin sonucu hangi oranda saptanabildiğini belirten olasılıktır ve Eşitlik 3.76 ile tespit edilmektedir (Özdamar, 2003).

$$P(D) = \frac{A}{A+C} = \frac{GP}{(GP+YN)} \tag{3.76}$$

Özgüllük ise Eşitlik 3.77 ile ifade edilen, modelin gerçekten negatif sonuçları ayırabilme yeteneğini belirten orandır (Özdamar, 2003).

$$P(\ddot{O}) = \frac{D}{B+D} = \frac{GN}{(YP+GN)} \tag{3.77}$$

Anlamlı ve uygun bulunan modellerin geçerliliği tüm bu ölçütler kullanılarak daha net olarak gözlenebilmektedir ve araştırmacıya tahmin yapmada yüksek doğruluk olanağı sağlamaktadır. Buradan yola çıkarak, hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi analizi geçerlilik ölçütü olarak kullanılarak model doğruluğu artırılması hedeflenmiştir. Modelin sağladığı anlamlılık ve uygun bilgisi yanı sıra bu geçerlilik yöntemlerinin kullanımı bulgular açısından oldukça önemlidir.

4. BULGULAR VE TARTIŞMA

Çalışmanın ana hedeflerinden olan mortalite tahmininde kullanılacak değişkenlerin birbirleriyle ilişkili olması amacıyla değişkenler kümelenmiştir. Kümeleme yöntemlerinde doğrusallık gibi kısıtları aşmak amacıyla kopulalar kullanılmış ve kopulalar aracılığıyla kümeleme yapılması amaçlanmıştır. Böylelikle, ortaya konulan modellerde tahmin gücünün artması beklenmiştir. Öte yandan, geniş bir veri setiyle çalışmanın avantajları da çalışmanın önemli adımlarındandır.

Elde edilecek mortalite tahmin modelleri ile yalnızca fizyolojik değişkenlerin kullanımından çıkılarak, hayati değişkenlerin de mortalite üzerindeki etkisinin görülmesi hedeflenmiştir. Böylelikle, özel olarak hesaplanan mortalite skorlarının genel kullanımı için de farklı bir yol açılacaktır.

Materyal ve yöntem başlığında detaylı şekilde açıklanan yöntemlerin uygulamaları adım adım gerçekleştirilecektir. Öncelikle, çalışmanın en önemli kısımlarından birisi olan veri seti detaylı olarak açıklanmıştır.

MIMIC-III yoğun bakım veri tabanı detaylı şekilde açıklanarak, veri tabanından çekilen değişkenlerin seçim ölçütleri anlatılmıştır. Belirtilen değişkenler kapsamında elde edilen veri setinin düzenlenmesi ve eksik gözlemlerin tahmini de önemli bir adımdır.

Veri setinin tanımlayıcı istatistikleri verilmesinin ardından, kopulalar aracılığıyla kümeleme teknikleri uygulanarak değişkenler kümelenmiştir. İlgili kümelerdeki değişkenler kullanılarak Lojistik Regresyon Analizi ile mortalite modellenmiştir.

Elde edilen modellerin geçerlilik ve güvenirliklerinin sınanması son adımda oldukça önemlidir. Bu nedenle, hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi kullanılmıştır.

Veri setinin düzenlenmesi Microsoft Excel, eksik veri tahminleri ve bağımlılık yapılarının incelenmesi R programı, Lojistik Regresyon Analizi ise SPSS 23.0 ile gerçekleştirilmiştir. Model geçerlilikleri yine R progrmaı ve SPSS 23.0 ile araştırılmıştır.

4.1. MIMIC-III Yoğun Bakım Veri Tabanı ve Özellikleri

Bu tez çalışmasında kullanılan veriler, MIT Hesaplamalı Fizyoloji Laboratuvarı tarafından geliştirilen ve kullanıma açık bir veri tabanı olan MIMIC-III'ten alınmıştır (Johnson vd., 2016).

MIMIC, Beth Israel Deaconess Tıp Merkezi'nin yoğun bakım ünitelerinde 2001 ve 2012 tarihleri arasında tedavi gören 40 binden fazla hastaya ait büyük ve ücretsiz ulaşılabilir bir veri tabanıdır. Bu veri tabanı, demografik bilgiler, laboratuvar test sonuçları, uygulanan işlemleri, ilaçları, bakıcı notları, görüntüleme raporları, başucunda not edilen hayati işaret ölçümleri (saat başı) ve ölüm durumu (hastane içinde veya dışında) değişkenlerini içermektedir (Johnson vd., 2016).

MIMIC, epidemiyoloji, klinik karar iyileştirme ve elektronik geliştirmelerini kapsayan çok çeşitli alanlarda analitik çalışmaları desteklemektedir. Şu üç durum dikkat çekmektedir.

• Dünya çapındaki araştırmacılar için ücretsiz ulaşılabilir.

• Farklı ve çok geniş bir hasta popülasyonunu kapsar.

• Laboratuvar sonuçları dahil olmak üzere yüksek geçici çözünürlük verilerini, elektronik belgeleri ve başucu monitör verilerini içermektedir.

MIMIC-III veri tabanı daha fazla veri elde edildikçe, veri toplama yöntemleri geliştikçe ve ilgili veri tabanı ile ilgili geri bildirimler aldıkça güncellenmektedir.

MIMIC'in son versiyonu MIMIC-III 38645'i yetişkin, 7875'i yeni doğan olmak üzere 58000 hastanın hastane kayıtlarını içermektedir. Güncellenen veriler Haziran 2001 ile Ekim 2012 arasını kapsamaktadır. Veri tabanı, tanımlanmamış olmasına rağmen, hastaların klinik bakımı ile ilgili detaylı bilgiler içermektedir.

MIMIC son güncelleme ile aşağıda bulunan başlıklarda bilgiler içermektedir.

• Hastanede yatan ve ayakta bakım gören hastaların laboratuvar ölçümleri,

• Hastanın yoğun bakım ünitesinde kaldığı sürede yapılan gözlem değerleri,

• Hastanın hastanede yatış sürecinde hemşirelerin ve doktorların aldığı notlar dahil olmak üzere, taburcu özeti gibi bilgileri,

• Ekokardiyografi raporları,

• Elektrokardiyogram raporlarından on iki önemli rapor bilgileri bulunmaktadır.

İlerleyen zamanlarda, ilgili güncellemelerle MIMIC'in aşağıdaki bilgilerle geliştirilmesi hedeflenmektedir.

• Hastanın hastanede kaldığı süre boyunca yapılan tıbbi görüntüleme sonuçları,

• Ameliyat esnasında kaydedilen klinik ölçümler,

• Acil servis ölçümleri bilgilerinin de eklenmesi hedeflenmektedir.

Veri setinin elde edilmesinin ardından modellemede verinin etkili bir biçimde kullanılabilmesi için değişken seçimi ve verinin düzenlenmesi adımları oldukça önemlidir. Model oluşumunda kullanılacak değişkenlerin belirlenmesi ve veri setinin düzenlenmesi izleyen iki bölümün konularıdır.

4.2. Değişken Seçimi

Pek çok mortalite tahmin yöntemi, bir sonucu tahmin etmek için hastalık şiddetini değerlendiren skor temelli modeller olarak kabul edilir. Bu modeller, hasta demografisini ve yoğun bakım ünitesine girişinden sonraki ilk 12 ila 24 saat içinde toplanan yaş, sıcaklık ve kalp atış hızı gibi fizyolojik değişkenleri yoğun bakım ünitesi performansını değerlendirmek için kullanmaktadır (Sadeghi vd., 2018).

Yoğun bakım ünitesinde yatan hastaların mortalite tahminleri için sıklıkla tercih edilen skorlama yöntemleri SOFA (Sequential Organ Failure Assessment-Sıralı Organ Yetmezliği Değerlendirmesi), APACHE (Acute Physiology and Chronic Health Evaluation-Akut Fizyoloji ve Kronik Sağlık Değerlendirmesi) ve SAPS (Simplified Acute Physiology Score - Basitleştirilmiş Akut Fizyoloji Skoru)'tır (Waudby-Smith vd., 2018; Johnson ve Mark, 2017). Bu skorlar uygulanan tıbbi yöntemi etkilemek için tasarlanmamıştır, hastanın seyri ve iyileşme şansı ile ilgili olarak aileye bilgi vermek, klinik değerlendirme yapmak ve kalite değerlendirmesi yapmak amacıyla kullanılmaktadır.

Hastaya yapılan müdahalenin başarısını veya başarısızlığını tespit etmek için kullanılmamaktadır (Vincent vd., 1996; Knaus vd., 1985; Le Gall vd., 1993). Bu skorlar, hastanın demografik verileri, laboratuvar test sonuçları ve elektronik sağlık kaydında bulunan hayati bulguları işlemektedir (Waudby-Smith vd., 2018).

SOFA, APACHE ve SAPS skorları ilgili değişkenlerin en kötü değerleri kullanılarak hesaplanmaktadır (Vincent vd., 1996; Knaus vd., 1985; Le Gall vd., 1993). SOFA ve APACHE II Skorları gibi klinik tahmin araçları, keskinlik düzeyine ve mortaliteye karar vermek için yoğun bakım ünitesine başvuran tüm hastalarda ölçülebilir (Vincent vd., 1996; Knaus vd., 1985).

APACHE II skoru, Knaus vd. (1985) tarafından, 13 hastaneden 5815 hasta ile yaptıkları çalışma ile hastanın durumunun ciddiyetini kavrama aracı olarak mortaliteyi tahmin etme amacıyla geliştirilmiştir. Bu puanlama endeksi, hastane kaynaklarının kullanımını değerlendirmek ve yoğun bakımın farklı hastanelerde veya zaman içindeki etkinliğini karşılaştırmak için kullanılabilir.

APACHE II skoru, 12 rutin fizyolojik ölçüm değişkeni ile birlikte demografik değişkenlerle birlikte ölçülmektedir. Bu değişkenler Çizelge 4.1'de görülmektedir.

Çizelge 4.1. APACHE II skorunda kullanılan değişkenler

Organ Yetmezliği Geçmişi veya İmmün Sistemi Baskılanması
Yaş
Vücut Sıcaklığı
Ortalama Atardamar Basıncı
Anyon Açığı
Kalp Atış Hızı
Solunum Hızı
Sodyum
Potasyum
Kreatinin
Akut Böbrek Yetmezliği
Kırmızı Kan Hücresi
FiO_2

SAPS II skoru, Le Gall vd. (1993) tarafından, 13152 hasta üzerinde yapılan bir çalışma ile SAPS skorunu (Basitleştirilmiş Akut Fizyoloji Skoru)'nu geliştirip, doğrulayarak, mortalite skoruna dönüştürmek amacıyla geliştirilmiştir.

Uluslararası geniş bir hasta örneklemine dayanarak ortaya çıkartılan SAPS II, temel bir teşhis belirtmek zorunda kalmadan mortalitenin tahmin edilmesini sağlar. Bu, yoğun bakım ünitelerinin etkinliğinin gelecekte yapılacak değerlendirilmesi için bir başlangıç noktası olarak kullanılmaktadır (Le Gall vd., 1993).

SAPS II skoru, 12 fizyoloji değişkeni olmak üzere 16 adet değişkeni kullanmaktadır. Bu değişkenler Çizelge 4.2'de görülmektedir.

Çizelge 4.2. SAPS II skorunda kullanılan değişkenler

Yaş
Kalp Atış Hızı
Sistolik ve Diastolik Kan Basıncı (Büyük ve Küçük Tansiyon)
Vücut Sıcaklığı
Glasgow Koma Skalası
PaO_2/FiO_2 Oranı
Mekanik Solunum
Kan Üre Azotu
Üre Çıkışı
Sodyum
Potasyum
Bikarbonat
Bilirubin (Karaciğer)
Akyuvar Sayısı
Kronik Hastalıklar
Yoğun Bakım Ünitesi Kabul Tipi

SOFA Skoru ise Vincent vd., (1996) tarafından yoğun bakım ünitesine başvuran tüm hastalarda kullanılmak üzere geliştirilmiştir. SOFA, İngilizce The Sequential Organ Failure Assessment; Türkçe Sıralı Organ Yetmezliği Değerlendirmesi demektir. SOFA

Skoru altı organ sisteminin işlevsizliğinin derecesine dayanan bir ölüm tahmin skorudur. Skor, giriş sırasında ve taburcu edilmeden bir gün öncesine kadar 24 saatte bir hesaplanır.

Bu skor aracılığıyla, yoğun bakım ünitesi hastalarının durumunun, skorun hesaplanması için kullanılan verilerin kabul değerleri ile sınırlı olmadığı göz önüne alındığında mortalitenin iyi bir şekilde sınıflandırıldığına inanılmaktadır. Bu değişkenler Çizelge 4.3'te görülmektedir.

Çizelge 4.3. SOFA skorunda kullanılan değişkenler

PaO_2/FiO_2 Oranı
Mekanik Solunum
Trombositler
Glasgow Koma Skalası
Bilirubin (Karaciğer)
Ortalama Atardamar Basıncı veya Gerekli Vazoaktif Ajanların Uygulanması
Kreatinin (Böbrek Yetmezliği)

SOFA, APACHE II ve SAPS II skorları tüm hastalar için ölçülebilir ve birbirleriyle karşılaştırılabilir (Vincent vd., 1996; Knaus vd., 1985; Le Gall vd., 1993).

Öte yandan, bu modeller yeterli sonuçlar vermesine rağmen, yoğun bakım ünitesi hastaları çeşitlidir ve çoklu hastalıklara maruz kalabilirler. Bu nedenle, hemen yoğun bakım ünitesine başvuran özel bir hasta için doğru modeli seçmek zordur (Sadeghi vd., 2018).

Bu skorlar, yoğun bakım ünitesi girişinden sonraki ilk saatlerde toplanan belirli klinik kayıtlara dayanarak düzenlenir. Laboratuvar test sonuçları bu skorların tahminlerinde önemli rol oynamaktadır. Ancak, kan basıncı, vücut sıcaklığı, solunum gibi hayati sinyaller, mortalite ile güçlü bir ilişkiye sahip olduğu kanıtlanmıştır ve hekime sayısız bilgi sağlayabilir (Zhang vd., 2015). Bu nedenle, hayati sinyal dalgalanmaları, mortaliteyi klinik temelli yöntemlerden daha doğru ve daha hızlı tahmin etmek için yüksek yetenek sağlayabilir (Sadeghi vd., 2018).

Hug ve Szolovits (2009) tarafından MIMIC II veri tabanı ile yapılan mortalite çalışmasında, değişken seçiminde SAPS II skorundan yola çıkarak literatür incelemesiyle ilerlenmiştir.

Johnson vd. (2017), MIMIC-III veri tabanında mortaliteyi incelerken literatürde kullanılan değişkenlerden yola çıkmışlardır. Waudby-Smith vd. (2018) ise, MIMIC-III veri tabanını kullandıkları çalışmalarında hasta seçimini koroner bakım ünitesi, kalp cerrahisi kurtarma ünitesi, tıbbi yoğun bakım ünitesi, cerrahi yoğun bakım ünitesi ve travma/cerrahi yoğun bakım ünitesinde hizmet alanlar olarak belirleyerek değişken seçimini yine literatür üzerinden seçim yaparak sonuçlandırmışlardır.

Kaji vd. (2019) ise MIMIC-III veri tabanında yaptıkları çalışmada değişkenleri laboratuvar bulguları, hayati göstergeler, ilaç tedavisi başlıkları altında topluca değerlendirerek seçim yapmışlardır.

Vincent vd. (2018) yoğun bakım hastalarında ortalama arter basıncı ve dolaşım sorunu yaşayanları tercih ederek, değişken seçimini biraz daha kısıtlayarak yine MIMIC-III veri tabanı üzerinde çalışmışlardır. Böylelikle, hastalık ölçütü ile doğal bir seçilim ortaya çıkmış olmaktadır.

Literatürde yapılan çalışmalarda, değişken seçimi konusunda araştırmacıların tecrübesi ile ilerlenmesinin yaygın olarak tercih edildiği görülmektedir. Mortalite tanımları üzerinden değişken seçimleri gerçekleştirilmiş olduğu tespit edilmiştir. Bir başka seçim yolu olarak da, tedavi gören hastaların hastalıkları veya servis birimleri üzerinde kısıtlama yapılarak ilerlenmesidir. Belli hastalıklarla başvuran hastalar veya belirli servislerde hizmet alan hastalar üzerinde çalışarak değişkenler üzerinde doğal bir seçim yoluna gidilmiştir.

Bu çalışmada ise, hasta mortalite tespitinde kullanılan skorlardan yola çıkılarak değişkenlerde bir kısıtlamaya veya seçime gidilmeden ilerlenmiştir. APACHE II, SAPS II, SOFA skorlarında kullanılan değişkenler temel alınarak Johnson vd. (2017) ve Yılmaz vd. (2014) çalışmalarında olduğu gibi literatürde mortalite ile ilişkilendirilmiş olan albümin, hemoglobin, glikoz gibi hayati değişkenler de eklenerek değişken havuzu oluşturulmuştur.

Buradan yola çıkarak iki tanesi ölüm durumunu gösteren değişkenler olmak üzere 27 değişken bir araya getirilmiştir. Bu değişkenler, mortalite durumu ortaya çıkmasına

neden olan fizyolojik ve hayati değişkenler olarak belirlenmiştir. Hedeflenen mortalite modellemesi için *hastanede ölüm* ve *24 saatte ölüm* seçenekleri tercih edilmiştir. Kullanılan değişkenler Çizelge 4.4'de gösterilmektedir.

Çizelge 4.4. Mortalitenin tahmininde kullanılan değişkenler

1	Cinsiyet	Kategorik
2	Yaş	Sürekli
3	Karaciğer	mg/dL
4	Mekanik Solunum	Kategorik
5	Böbrek Yetmezliği	mg/dL'den kategorik
6	Nefes	mmHg
7	Pıhtılaşma	$\times 10^3/\mu L$
8	Kardiyovasküler	mmHg'den kategorik
9	Merkezi Sinir Sistemi	Skor
10	Kalp Atış Hızı	Kesikli
11	Vücut Sıcaklığı	ºC
12	Sistolik Kan Basıncı	mm/Hg
13	Albümin	g/L
14	Sodyum	mEq/L
15	Potasyum	mEq/L
16	Bikarbonat	mEq/L
17	Klorür	mEq/L
18	Laktat	mEq/L
19	Kan Üre Azotu	mg/dL
20	Beyaz Kan Hücresi	$10^3/mm^3$
21	Glikoz	mg/dL
22	Kırmızı Kan Hücresi	%
23	Hemoglobin	g/dL
24	Anyon Açığı	mEq/L
25	Olgunlaşmamış Nötrofil Hücreler	>10% olması ile
26	24 Saatte Ölüm	Kategorik
27	Hastanede Ölüm	Kategorik

Kategorik değişkenlerin değer aralıkları Çizelge 4.5'te gösterilmiştir (Vincent vd., 1996).

Çizelge 4.5. Mortalitenin tahmininde kullanılan kategorik değişkenlerin değer aralıkları

Cinsiyet	1: Kadın 2: Erkek
Karaciğer	En yüksek bilirubin ≥ 12,0 ise 4 En yüksek bilirubin ≥ 6,0 ise 3 En yüksek bilirubin ≥ 2,0 ise 2 En yüksek bilirubin ≥ 1,2 ise 1 Aralık dışında ise 0
Mekanik Solunum	1: Müdahale var (Kanül vb,) 0: Müdahale yok
Böbrek Yetmezliği	En yüksek kreatinin ≥ 5,0 ise 4 En yüksek kreatinin ≥ 3,5 ve En yüksek kreatinin < 5,0 ise 3 En yüksek kreatinin ≥ 2,0 ve En yüksek kreatinin < 3,5 ise 2 En yüksek kreatinin ≥ 1,2 ve En yüksek kreatinin < 2,0 ise 1 Aralık dışında ise 0
Nefes	100 > PaO_2 / FiO_2 > 0 ise 4 199 ≥ PaO_2 / FiO_2 ≥ 100 ise 3 299 ≥ PaO_2 / FiO_2 ≥ 200 ise 2 399 ≥ PaO_2 / FiO_2 ≥ 300 ise 1 PaO_2 / FiO_2 ≥ 400 ise 0
Pıhtılaşma	En düşük trombosit < 20 ise 4 20 < En düşük trombosit < 49 ise 3 50 < En düşük trombosit < 99 ise 2 100 < En düşük trombosit < 150 ise 1 150 ≤ En düşük trombosit ise 0
Kardiyovasküler	15 < Dopamin; 0,1 < Epinefrin veya 0,1 < Norepinefrin ise 4 5 < Dopamin; 0,1 ≥ Epinefrin veya 0,1 ≥ Norepinefrin ise 3 5 ≥ Dopamin veya herhangi bir doz Dobutamin ise 2 Ortalama Arter Basınç < 70 ise 1 Aralık dışında ise 0
Merkezi Sinir Sistemi	En düşük koma skoru < 6 ise 4 6 ≤ En düşük koma skoru ≤ 9 ise 3 10 ≤ En düşük koma skoru ≤ 12 ise 2 13 ≤ En düşük koma skoru ≤ 14 ise 1 Aralık dışında ise 0
Hastanede Ölüm	1: Ölüm var 0: Ölüm yok
24 Saatte Ölüm	1: Ölüm var 0: Ölüm yok

Değişkenlerin elde edildiği ölçümler aşağıdaki gibi açıklanabilir.

• *Cinsiyet*, kadın-erkek olarak iki kategorili olarak belirlenmiştir.

• *Yaş*, sürekli değişken olarak veri setinden alınmıştır.

• *Karaciğer*, kandaki bilirubin değeri kullanılarak kategorik olarak kullanılır.

• *Mekanik solunum*, kanül vb. ile müdahale edilip edilmediğine göre kategorilendirilir.

• *Böbrek yetmezliği*, kandaki kreatinin yükseliğine göre kategorilendirilerek kullanılır.

• *Nefes*, kandaki oksijen miktarının alınan havadaki oksijen yüzdesine oranı hesaplanarak kategorilendirilir.

• *Pıhtılaşma*, kanda bulunan en düşük trombosit değeri üzerinden hesaplanır.

• *Kardiyovasküler*, dopamin, epinefrin ve norepinefrin kullanımı ile kategorilendirilir.

• *Merkezi sinir sistemi,* Glasgow Koma Skoru aracılığı ile kategorilendirilir.

• *Kalp atış hızı*, hastanın nabız sayısı ile belirlenir.

• *Vücut sıcaklığı*, hastanın ºC ile ölçülen vücut sıcaklığıdır.

• *Sistolik kan basıncı*, hastanın büyük kan basıncıdır.

• *Albümin*, kandaki albümin miktarıdır.

• *Sodyum*, kandaki sodyum miktarıdır.

• *Potasyum*, kandaki potasyum miktarıdır.

• *Bikarbonat*, kandaki bikarbonat miktarıdır.

• *Klorür*, kandaki klorür miktarıdır.

• *Laktat*, kandaki klorür miktarıdır.

• *Kan üre azotu*, kanda sabit olarak bulunan üre azotunun ölçümü ile belirlenir.

• *Beyaz kan hücresi*, kanda beyaz kan hücresi sayımı ile belirlenir.

• *Glikoz,* kan glikoz seviyesidir.

• *Kırmızı kan hücresi*, kanda kırmızı kan hücresi sayımı ile belirlenir

• *Hemoglobin,* kandaki hemoglobin seviyesidir.

• *Anyon açığı*, anyonlar=katyonlar eşitliği üzerinden idrardan belirlenir.

• *Olgunlaşmamış nötrofil hücreler*, mikroorganizmalarla savaşması gereken ancak normal işlevlerini yerine getiremeyen hücre miktarıdır.

• *24 saatte ölüm* ve *hastanede ölüm* değişkenleri ölüm var ve ölüm yok şeklinde kategorilendirilerek kullanılır.

Modellemede kullanılacak değişkenlerin belirlenmesinin ardından, veri setinin literatüre ve sağlık sistemi kayıt yöntemlerine uygun şekilde düzenlenmesi oldukça önemlidir. Böylelikle, yanlış ve mükerrer kayıtlar, eksik bilgiler gibi sorunların oluşmaması hedeflenmiştir.

4.3. Veri Düzenleme

Çalışma kapsamında incelenen çalışmalarda veri setinden hasta seçiminde pek çok ölçüt bulunmaktadır. Kullanılan yöntemlerin gerektirdiği özel kısıtlar dışında çalışmalarda kullanılan ortak ölçütler mevcuttur.

Öncelikle, MIMIC-III'te Amerika'daki Sağlık Sigortası Taşınabilirlik ve Sorumluluk Yasası gereği 89 yaşından büyük hastaların gizliliğinin korunması adına doğum tarihi bilgileri değiştirilmiştir (MIT, 2016). Bu nedenle doğum tarihi tahmini yapılsa da 89 yaş üzeri hastalar modele dâhil edilmemiştir.

Yaş değişkeninde bir diğer önemli durum ise, sadece yetişkin hastaları kapsamasıdır (Waudby-Smith vd., 2018, Johnson vd., 2017). Çalışma kapsamında yalnız 15 yaş ve üzeri hastalar veri setinde değerlendirilmiştir.

Grafiksel gözlem bilgileri olmayan, kalp atış hızı ölçümleri, yoğun bakım ünitesi kabulü ve taburcu bilgileri olmayan hastalar veri setinden çıkartılmıştır. Bu tip eksiklikler kayıt hatalarına karşılık gelir sayılmıştır (Johnson vd., 2017).

"Yeniden kabul edilme" şeklinde kaydedilen ve 4 saatten az süre yoğun bakım ünitesinde kalan hastalar çıkartılmıştır. Bu kalışlar mortalite tahminlerinde küçük bir etkiye

neden olmaktadır. Ayrıca hastanın birden çok kalması durumunda son yatış verileri dikkate alınmıştır (Johnson vd., 2017).

Belirtilen tüm düzenlemeler sonucunda toplam 38015 hasta ile çalışma yürütülmüştür. Hastalara ait değişkenlerde bulunan eksik gözlemler tahmin edilerek çalışmaya devam edilmiştir.

4.3.1. Eksik gözlemlerin tahmin edilmesi

MIMIC-III ve mortalite tahmin çalışmalarında eksik verilerin tahmininde farklı yollar kullanılmıştır. Sadeghi vd. (2018) eksik olan değerleri bir önceki bilinen değerle değiştirerek, Kaji vd. (2019) bilinen değerlerin medyanını kullanarak, Alistair ve Mark (2017) eksik verilerin bulunduğu hastaları çıkararak mortalite tahmini yapmışlardır.

Che vd. (2018) eksik gözlem tahmini için tekrarlayan sinir ağları yöntemlerinden önerilerde bulunmuştur. Kim vd. (2014) ise interpolasyon, 5 ortalama, Eksik Değer Tekil Değer Ayrışımı (MSVD) ve Beklenti En Büyükleme Tabanlı Temel Bileşenler Analizi'ni karşılaştırmıştır. Abdala ve Saeed (2004) ağırlıklı K-En Yakın Komşular Algoritması ile MIMIC-II üzerinden APACHE skoru aracılığıyla eksik gözlem tahmini üzerinde durmuştur.

MIMIC veri tabanında yapılan çalışmalarda yazarlar farklı teknikler kullanarak eksik verileri değerlendirme yoluna gitmişlerdir. Yadav ve Roychoudhury (2018) ise R paket programının eksik veri değerlendirmede kullanılan paketlerini karşılaştırarak incelemiştir. Farklı boyutta ve farklı sayıda eksik veriye sahip veri setleri üzerinde VIM, MICE, MissForest ve HMISC paketleri değerlendirilmiştir.

Bu paket ve yöntemlerin veri seti büyüklüğü üzerindeki etkisini gözlemleyebilmek için %10, %20, %30 ve %40 kadar eksik veriler oluşan 10000, 15000, 20000, 50000 ve 100000 satırlık veri setleri üzerinde çalışılmıştır. İlgili veri setlerinde ilgili paketlerin performansı zaman, doğruluk, etkinlik ve orijinal varyansa göre ortaya çıkan değişimler üzerinden incelenmiştir. Bir eksik veri tahminleme paketi, daha az sürede yüksek doğruluk sağlıyorsa verimli olarak kabul edilir. Ayrıca varyans açısından, tahminleme paketinin değişkenlerin orijinal varyansını korumasını sağlaması beklenmektedir (Yadav ve Roychoudhury, 2018).

Yadav ve Roychoudhury (2018) çalışması sonucunda, Yapay Sinir Ağları ile büyük veri setlerinde ve fazla sayıda eksik veri olması durumunda iyi bir performans gösterirken, küçük veri setlerinde düşük performans gösterdiği tespit edilmiştir. Varyans yüzdesi, kullanılan tahmin modeli için performansın mutlak ölçütü olmamakla birlikte, modelin doğruluğunu gösteren önemli bir ölçüttür. Genel olarak, varyans yüzdesi VIM paketinde eksik veri arttıkça artış göstermektedir. Varyans yüzdesi MICE paketinde az sayıda eksik veride Lojistik Regresyon gibi tahminlerde ve çok sayıda eksik veri durumunda yapay sinir ağları kullanımında düşüktür. Küçük veri setleri için VIM paketi uygunken daha büyük veri setlerinde MICE ve HMISC paketlerinin kullanımı önerilmektedir.

MICE paketi, Zincirleme Denklemler ile Çok Değişkenli tahminde (Royston, 2004) ve rassal tahminlerle rassal olmayan tahminlerde kullanılabilir (Buuren ve Groothuis-oudshoorn, 2011). MICE farklı değişken tipleri için farklı değerleme modellerine destek olur. Sayısal değişkenler için Tahmini Ortalama Eşleştirme Yöntemi, ikili değişkenler için Lojistik Regresyon Yöntemi, faktör değişkenleri için Bayesçi Regresyon, Orantılı Olasılık modeli için kullanılabilir (Buuren ve Groothuis-oudshoorn, 2011; Horton vd. 2001).

Bu nedenle, %40 kadar eksik veriye sahip veri setinde, eksik verilerin değerlendirilmesi MICE isimli R paketi aracılığıyla gerçekleştirilmiştir. İlgili pakette, Bayesçi Regresyon yöntemini kullanarak eksik gözlemlerin tahmini yapılmıştır.

Kullanılan veri setinin düzenlenmesi ve değişken seçimi tamamlanmasının ardından eksik gözlemlerin tahmini ile veriler, bağımlılık yapısının kopulalar aracılığıyla incelenmesine hazır duruma gelmiştir. Kümeleme tekniklerinin ve lojistik modellerin oluşturulmasına başlanmadan önce veriyi tanıtmak adına tanımlayıcı istatistikler tablolaştırılarak açıklanmıştır.

4.4. Tanımlayıcı İstatistikler

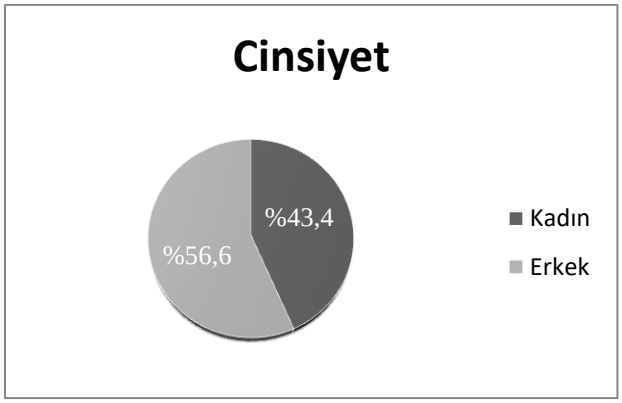
Çalışmada kullanılan ilgili veri tabanında bulunan değişkenler öncelikle tanımlayıcı istatistikler bazında incelenmiştir. Hastaların cinsiyet ve yaş dağılımları ile klinik değişkenler hakkında ön bilgi edinmek çalışmanın yönünü de doğrudan etkileyecektir.

Veri setinde bulunan hastaların yaklaşık %57'si erkek iken %43'ünü kadınlar oluşturmaktadır. *Cinsiyet* değişkenine dair sıklıklar Çizelge 4.6'da gösterilmektedir.

Çizelge 4.6. Cinsiyet değişkeninin sıklık ve yüzdeleri

Cinsiyet		
	Sıklık	**Yüzde**
Kadın	16493	43,4
Erkek	21522	56,6
Toplam	38015	%100

Cinsiyet değişkenine ait pasta grafiği Şekil 4.1'de ifade edilmektedir.



Şekil 4. 1. Cinsiyet değişkenine ait pasta grafiği

Hastaların karaciğer değerleri incelendiğinde yalnızca %1'inin ilgili *bilirubin* değerlerinin yüksek olduğu, yaklaşık %90'ının ise düşük bilirubin değerine sahip olduğu gözlemlenmiştir. Hastaların yüksek kreatinin nedeniyle *böbrek yetmezliği* görülme sıklığı %8,5 iken düşük kreatinin görülen hasta sıklığı %66,3'tür. *Nefes* ile ilgili hayati sorun yaşayan hastaların oranı %10,2 iken, sorun yaşamayan hastaların oranı ise %57 olarak belirlenmiştir. Pıhtılaşmanın az olduğu, kanama riski olan hastaların yüzdesi yaklaşık %1 iken pıhtılaşmanın çok olduğu hastaların oranı ise yaklaşık %90'dır. Dopamin ve epinefrin uygulamalarının yapıldığı ve kardiyovasküler riski olan hastaların yüzdesi yaklaşık %11,8 iken riskin olmadığı hastaların oranı %3,1'tir. Koma skoruna göre kategorilendirilen *merkezi sinir sistemi* değişkeninde koma durumu gözlenen hastaların oranı %2,5 iken riskin azaldığı hastaların oranı %82,2'dir. Kategorik değişkenler için sıklık ve yüzde değerleri Çizelge 4.7'de ifade edilmiştir.

Çizelge 4.7. Kategorik değişkenlerinin sıklık ve yüzdeleri

		0	**1**	**2**	**3**	**4**	**Toplam**
Karaciğer	**Sıklık**	17219	16882	3548	349	17	38015
	%	45,3	44,4	9,3	0,9	0,1	100%
Böbrek Yetmezliği	**Sıklık**	10017	15179	9597	2581	641	38015
	%	26,4	39,9	25,2	6,8	1,7	100%
Nefes	**Sıklık**	4355	17305	12470	3539	346	38015
	%	11,5	45,5	32,8	9,3	0,9	100%
Pıhtılaşma	**Sıklık**	17684	16637	3374	302	18	38015
	%	46,5	43,7	8,9	0,8	0,1	100%
Kardiyovasküler	**Sıklık**	1167	24842	7508	3963	535	38015
	%	3,1	65,3	19,8	10,4	1,4	100%
Merkezi Sinir Sistemi	**Sıklık**	14348	16909	5800	844	114	38015
	%	37,7	44,5	15,3	2,2	0,3	100%

Hastaların yaklaşık %53'üne kanül vb. mekanik solunum cihazları uygulaması yapılmışken, yaklaşık %47'sine *mekanik solunum* uygulaması yapılmamıştır. İlgili değerler Çizelge 4.8'de gösterilmiştir.

Çizelge 4.8. Mekanik solunum değişkeninin sıklık ve yüzdeleri

Mekanik Solunum		
	Sıklık	**Yüzde**
0	17706	46,6
1	20309	53,4
Toplam	38015	%100

Hastaların *yaş* değişkeninin, *kalp atış hızı*, *vücut sıcaklığı*, *sistolik kan basıncı* gibi hayati değişkenlerinin, laboratuvar sonuçlarını gösteren *albümin, sodyum, potasyum, bikarbonat, klorür, laktat* ve *glikoz* değişkenlerinin ve kan değerlerini gösteren *kan üre azotu, hemoglobin, kırmızı kan hücresi, beyaz kan hücresi, anyon açığı* ve *olgunlaşmamış nötrofil hücreler* değişkenlerinin en düşük, en yüksek ve ortalamaları gösteren betimsel istatistikler Çizelge 4.9'da ifade edilmiştir.

Çizelge 4.9. Hayati değişkenlerin betimsel istatistikleri

	En Düşük	En Yüksek	Ortalama	Standart Sapma
Yaş	15	89	63,84	17,52
Albümin	0,4	5,7	3,19	0,56
Sodyum	68	176	138,58	4,27
Potasyum	2	15,5	4,25	0,59
Bikarbonat	5	52	24,23	4,32
Klorür	56	154	104,76	5,51
Laktat	0,30	24,6	2,27	1,42
Glikoz	26	0,935	145,15	56,66
Kalp Atış Hızı	33	173	87,64	15,35
Vücut Sıcaklığı	26,85	40,83	36,78	0,62
Sistolik Kan Basıncı	27	214	121,02	16,65
Kan Üre Azotu	1	232	26,41	20,14
Hemoglobin	3,40	20,80	10,99	1,86
Kırmızı Kan Hücresi	10,40	64	32,67	5,26
Beyaz Kan Hücresi	0,10	247,90	12,03	7,39
Anyon Açığı	2	48	14,41	3,554
Olg. Nötrofil Hücreler	1	69	9,75	4,31

Ölüm gerçekleşme oranlarını gösteren değerler Çizelge 4.10‘da ifade gösterilmiştir. *Hastanede ölüm* durumunun gerçekleşme oranı %17,5 iken *24 saatte ölüm* değişkeninin gerçekleşme oranı %1,8’dir.

Çizelge 4.10. Hastanede ve 24 saatte ölüm değişkenlerinin sıklık ve yüzdeleri

	Hastanede Ölüm		24 Saatte Ölüm	
	Sıklık	**Yüzde**	**Sıklık**	**Yüzde**
Ölüm Yok	31375	82,5	37331	98,2
Ölüm Var	6640	17,5	684	1,8
Toplam	38015	%100	38015	%100

Modellemede kullanılacak değişkenlerin birbirileriyle ilişkili olması hedeflenmesi nedeniyle, tanımlayıcı istatistiklerle tanıtılan değişkenlerin kümeleme teknikleriyle incelenerek modellenmesi ilerleyen bölümlerde gerçekleştirilecektir.

4.5. Kopulalar Aracılığıyla Kümeleme Tekniklerinin İncelenmesi

Doğrusallık ve parametrik bağımlılık kısıtlarını aşmak amacıyla kullanılan kopulalar aracılığıyla yapılan kümeleme tekniklerinden CoClust ve Kuyruk Bağımlılığı uygulamalarının sonuçları incelenmiştir.

CoClust aracılığıyla Frank, Gumbel ve Clayton kopulalar kullanılarak kümeleme yapılırken; kuyruk bağımlılığı ile hiyerarşik kümeleme tekniği kullanılarak Ward ve Tam bağlantı formülleri ile kümeler elde edilmiştir.

4.5.1. CoClust Tekniği ile kopulalar aracılığıyla kümeleme

Kopulalar aracılığıyla kümeleme tekniklerinden Coclust ile yapılan ilk kümeleme sonucu Frank kopula ile elde edilmiştir. Frank kopula aracılığıyla yapılan CoClust ile kümeleme sonucu değişkenler iki kümeye ayrılmıştır. Çizelge 4.11'de gösterilmektedir. Lascio (2008)'in açıkladığı üzere CoClust Tekniği ile elde edilen kümelerde satırlarda bulunan değişkenlerin ilişkili olduğu görülmektedir. Buradan yola çıkarak, *hemoglobin* ve *kırmızı kan hücresi*, *klorür* ve *sodyum*, *mekanik solunum* ve *nefes, potasyum ve kan üre azotu* değişkenleri birbiriyle ilişkilidir. CoClust Tekniği gereği satırların birbiriyle ilişkili olmasından yola çıkarak Frank kopuladan elde edilen bu kümeleme sonucu Lojistik Regresyon Analizinde kullanılmamıştır.

Çizelge 4.11. Frank kopuladan CoClust ile elde edilen kümeler

Küme 1	**Küme 2**
Hemoglobin	Kırmızı Kan Hücresi
Klorür	Sodyum
Mekanik Solunum	Nefes
Potasyum	Kan Üre Azotu
Vücut Sıcaklığı	Kalp Atış Hızı
Böbrek	Kardiyovasküler

Gumbel kopula aracılığıyla yapılan kümeleme sonucu değişkenler üç kümeye ayrılmıştır. Çizelge 4.12 incelendiğinde, Lascio (2008)'in açıkladığı gibi satır değişkenlerinin ilişkili olması nedeniyle *klorür, sodyum* ve *bikarbonat* birbiri ile ilişkili

değişkenler iken; öte yandan *böbrek, karaciğer* ve *pıhtılaşma* adı verilen kendi aralarında ilişkili değişkenlerdir. Mortalite modellemesinde buradan yola çıkarak satır değişkenleri ile ilerlenmiştir.

Çizelge 4.12. Gumbel kopuladan CoClust ile elde edilen kümeler

Küme 1	**Küme 2**	**Küme 3**
Hemoglobin	Kırmızı Kan Hücresi	Albümin
Klorür	Sodyum	Bikarbonat
Anyon Açığı	Kan Üre Azotu	Potasyum
Kardiyovasküler	Nefes	Mekanik Solunum
Böbrek	Karaciğer	Pıhtılaşma

Clayton kopula aracılığıyla yapılan kümeleme sonucu değişkenler beş kümeye ayrılmıştır. Çizelge 4.13 incelendiğinde, Lascio (2008)'in açıkladığı gibi satır değişkenlerinin ilişkili olması nedeniyle *nefes, kardiyovasküler, mekanik solunum, pıhtılaşma* ve *böbrek* değişkenleri birbiri ile ilişkili iken; öte yandan *hemoglobin, kırmızı kan hücresi, albümin, kan üre azotu* ve *anyon açığı* kendi aralarında ilişkili değişkenlerdir. Mortalite modellemesinde buradan yola çıkarak satır değişkenleri ile ilerlenmiştir.

Çizelge 4.13. Clayton kopuladan CoClust ile elde edilen kümeler

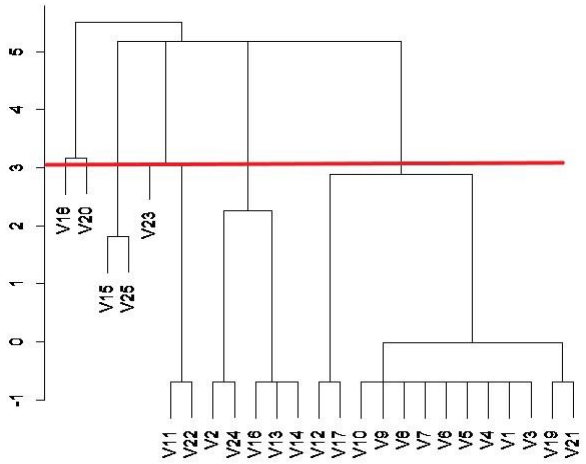
Küme 1	**Küme 2**	**Küme 3**	**Küme 4**	**Küme 5**
Hemoglobin	Kırmızı Kan Hücresi	Albümin	Kan Üre Azotu	Anyon Açığı
Bikarbonat	Klorür	Olgunlaşmamış Nötrofil Hücreler	Sodyum	Sistolik Kan Basıncı
Nefes	Kardiyovasküler	Mekanik Solunum	Pıhtılaşma	Böbrek
Beyaz Kan Hücresi	Potasyum	Laktat	Glikoz	Kalp Atış Hızı

Simetrik bir kopula ailesi olan Frank kopula ailesi ile yapılan kümelemede iki küme elde edilmesi nedeniyle Lojistik Regresyon uygulaması yapılsa modele iki değişkenle başlanacağı için modelleme konusunda verimli yol alınamamıştır ve modellemeye dahil edilmemiştir. Clayton ve Gumbel kopula gibi negatif ve pozitif kuyrukta bağımlılık yapısını inceleyen asimetrik kopulalarda detaylı bir kümeleme sonucu elde edilmiştir. Bu ailelerden elde edilen kümelerde bulunan değişkenler mortalite tahmininde kullanılmıştır.

Literatürde yeni bir yöntem olan CoClust'ın ardından görece daha eski bir teknik olan kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniği yine asimetrik bir yaklaşımla bağımlılık yapısını inceleyerek kümeleme olanağı sağlamaktadır. Bu yöntemden elde edilen kümeler takip eden bölümde tanıtılmıştır.

4.5.2. Kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme

Kuyruk bağımlılığı ile kümeleme yapılırken, *tam bağlantı formülü* ile yığınsal şekilde yapılan kümeleme sonucu tekniğin sağladığı *dendrogram* grafiği elde edilmiştir. Kümeleme sonucu bu grafik üzerinden yorumlanmaktadır. Dendrogram sonucu incelendiğinde, değişkenler altı kümeye ayrılmıştır. Tam bağlantı formülü için elde edilen dendrogram grafiği Şekil 4.1'de gösterilmektedir.



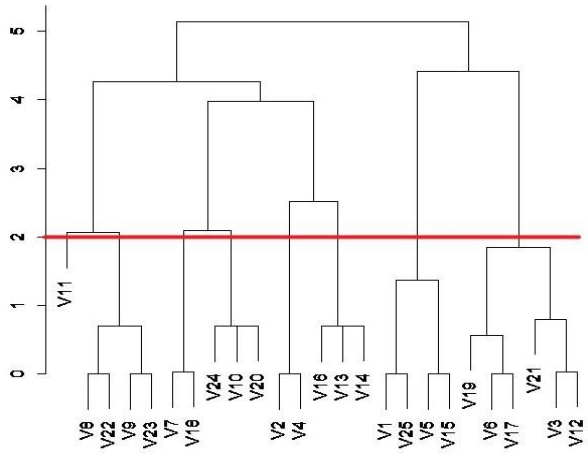
Şekil 4.1. Tam bağlantı formülü sonucu kümeleme dendrogramı

Kümeleme sonucu elde edilen *Cinsiyet, karaciğer, mekanik solunum, böbrek, nefes, pıhtılaşma, kardiyovasküler, merkezi sinir sistemi, kalp atış hızı, sistolik kan basıncı, klorür, kan üre azotu* ve *glikoz* değişkenleri birbiri ile ilişkili iken, *vücut sıcaklığı, kırmızı kan hücresi* ve *hemoglobin* kendi aralarında ilişkili değişkenlerdir. *Laktat* ve *beyaz kan hücresi* değişkenlerinin herhangi bir kümeye yerleştirilemediği, ayrı birer küme olarak bulundukları kümeleme sonucu Çizelge 4.14'de gösterilmiştir.

Çizelge 4.14. Tam bağlantı formülü ile elde edilen kümeler

Küme 1	Küme 2	Küme 3	Küme 4	Küme 5	Küme 6
Cinsiyet Karaciğer Mekanik Solunum Böbrek Nefes Pıhtılaşma Kardiyovasküler Merkezi Sinir Sistemi Kalp Atış Hızı Sistolik Kan Basıncı Klorür Kan Üre Azotu Glikoz	Yaş Albümin Sodyum Bikarbonat Anyon Açığı	Vücut Sıcaklığı Kırmızı Kan Hücresi Hemoglobin	Potasyum Olg. Nötrofil Hücreler	Laktat	Beyaz Kan Hücresi

Kuyruk bağımlılığı ile kümeleme yapılırken, *Ward formülü* ile yığınsal şekilde yapılan kümeleme sonucu tekniğin sağladığı *dendrogram* grafiği elde edilmiştir. Kümeleme sonucu bu grafik üzerinden yorumlanmaktadır. Dendrogram sonucu incelendiğinde, değişkenler sekiz kümeye ayrılmıştır. Ward formülü için elde edilen dendrogram grafiği Şekil 4.2'de gösterilmektedir.



Şekil 4.2. Ward formülü sonucu kümeleme dendrogramı

Cinsiyet, böbrek, potasyum ve olgunlaşmamış nötrofil hücreler değişkenleri birbiri ile ilişkili iken, kalp *atış hızı, beyaz kan hücresi* ve *anyon açığı* kendi aralarında ilişkili değişkenlerdir. Ward formülü ile elde edilen kümeler Çizelge 4.15'te ifade edilmektedir.

Çizelge 4.15. Ward formülü ile elde edilen kümeler

Küme 1	**Küme 2**	**Küme 3**	**Küme 4**	**Küme 5**	**Küme 6**	**Küme 7**	**Küme 8**
Cinsiyet Böbrek Potasyum Olg. Nötrofil Hücreler	Yaş Mekanik Solunum	Karaciğer Nefes Sistolik Kan Basıncı Klorür Kan Üre Azotu Glikoz	Pıhtılaşma Laktat	Kardiyo. Merkezi Sinir Sistemi Kırmızı Kan Hücresi Hemo.	Kalp Atış Hızı Beyaz Kan Hücresi Anyon Açığı	Albümin Sodyum Bikarb.	Vücut Sıc.

CoClust ve kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme teknikleri ile elde edilmiş kümelerde bulunan değişkenler kullanılarak mortalite tahmin Lojistik Regresyon Analizi aracılığıyla modellenmiştir.

4.6. Kümeleme Sonuçlarının Lojistik Regresyon Analizi Aracılığıyla Modellenmesi

CoClust Tekniği ve kuyruk bağımlılığı ile elde edilen kümelerdeki değişkenler kullanılarak hastaların mortalite Lojistik Regresyon Analizi aracılığıyla modellenmiştir. İkili lojistik model kullanımıyla ölüm var-yok şeklinde ortaya konan kategorik bağımlı değişken modellenmiştir. Böylelikle Diskriminant Analizi'nde mevcut olan bağımsız değişkenlerin normal dağılıma uyması ve bağımsız değişkenlerin grup düzeyinde kovaryanslarının eşitliği gibi kısıtları aşılmıştır. Diğer yandan, kategorik bağımlı değişken nedeniyle doğrusal regresyonda parametre tahmini için kullanılan en küçük kareler yönteminden yararlanılamamaktadır. Bu nedenle de, Lojistik Regresyon Analizi'nin kullanımı çalışma kapsamında verimli bir tercihtir.

Yapılan modellemeden sonra anlamlı modellerde bulunan anlamsız değişkenler çıkartılarak, tüm değişkenler anlamlı olana kadar süreç yenilenmiştir. Elde edilen modellerin uygunluğu ise Hosmer-Lemeshow'a göre değerlendirilmiştir. Aynı zamanda

modellerin çoklu bağlantı durumu da varyans şişkinlik faktörü (VIF, Variance Inflation Factor) aracılığıyla incelenmiştir.

İncelenen lojistik modellerde CoClust tekniği ile elde edilmiş kümelerde satır değişkenlerine uygulama yapılmıştır. Kopula aileleri kümeleri ile ilgili kopula ailesinden elde edilen kümelerin satırlarında bulunan değişkenler ifade edilmektedir.

Bu amaçla öncelikle *24 saatte ölüm* durumu modellenmiştir. İlk olarak CoClust aracılığıyla elde edilen kümeler incelenirken, ardından kuyruk bağımlılığı ile elde edilmiş kümeler modellenmiştir.

Clayton kopula aracılığıyla elde edilen kümelerden *ilki* ile yapılan modelleme sonucu %95 güven düzeyinde anlamlı bir model elde edilmiştir ($p<\alpha=0,05$) Çizelge 4.16'da gösterilmektedir.

Çizelge 4.16. Clayton kopula ilk kümesindeki değişkenlerin anlamlılıkları

Albümin	p < 0,05 *****
Kan Üre Azotu	p = 0,106
Kırmızı Kan Hücresi	p < 0,05 *****
Hemoglobin	p = 0,096
Anyon Açığı	p < 0,05 *****

Clayton kopula kümelemesinden elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenerek Çizelge 4.17'de ifade edilmektedir. Elde edilen bu model de anlamlı olarak tespit edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,932 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir.

Çizelge 4.17. Clayton kopula ilk küme anlamlılıklarının yinelenmesi

Albümin	p < 0,05 *****
Kırmızı Kan Hücresi	p < 0,05 *****
Anyon Açığı	p < 0,05 *****

Clayton kopula aracılığıyla elde edilen kümelerden *ikincisi* ile yapılan modelleme sonucu anlamlı bir model elde edilememiştir ($sig>\alpha=0,05$).

Clayton kopula aracılığıyla elde edilen kümelerden *üçüncüsü* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$). Elde edilen lojistik modelin değişkenleri Çizelge 4.18'de gösterilmektedir.

Çizelge 4.18. Clayton kopula üçüncü kümesindeki değişkenlerin anlamlılıkları

Kardiyovasküler 1	p = 0,092
Kardiyovasküler 2	p = 0,060
Kardiyovasküler 3	p < 0,05 *****
Kardiyovasküler 4	p = 0,256
Nefes 1	p < 0,05 *****
Nefes 2	p < 0,05 *****
Nefes 3	p < 0,05 *****
Nefes 4	p < 0,05 *****
Mekanik Solunum	p < 0,05 *****
Pıhtılaşma 1	p = 0,060
Pıhtılaşma 2	p = 0,060
Pıhtılaşma 3	p = 0,060
Pıhtılaşma 4	p = 0,060
Böbrek 1	p < 0,05 *****
Böbrek 2	p = 0,719
Böbrek 3	p = 0,474
Böbrek 4	p < 0,05 *****

Elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. İkinci adımda tespit edilen model anlamlı olarak belirlenememiştir ($sig>\alpha=0,05$).

Clayton kopula aracılığıyla elde edilen kümelerden *dördüncüsü* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$) ve Çizelge 4.19'da ifade edilmektedir.

Çizelge 4.19. Clayton kopula dördüncü kümesindeki değişkenlerin anlamlılıkları

Potasyum	p < 0,05 *****
Laktat	p = 0,256
Beyaz Kan Hücresi	p = 0,886
Glikoz	p = 0,067
Kalp Atış Hızı	p < 0,05 *****

Clayton kopula dördüncü kümesinden elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. Elde edilen bu model de anlamlı olarak

tespit edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,933 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir. Yinelenen modelin değişkenleri Çizelge 4.20'de gösterilmektedir.

Çizelge 4.20. Clayton kopula dördüncü küme anlamlılıklarının yinelenmesi

Potasyum	p < 0,05 *****
Kalp Atış Hızı	p < 0,05 *****

Gumbel kopula aracılığıyla elde edilen *ilk* kümede bulunan değişkenler Clayton kopula aracılığıyla elde edilen ilk kümeden yola çıkarak yenilenen modelde bulunan değişkenlerdir. *Albümin, kırmızı kan hücresi, hemoglobin* değişkenleri bulunmaktadır modelde. Bu nedenle Clayton kopula modeli kullanılmaya devam edilmiştir.

Gumbel kopula aracılığıyla elde edilen kümelerden *ikincisi* ile yapılan modelleme sonucu ise anlamlı bir model elde edilememiştir ($sig>\alpha=0,05$).

Gumbel kopula aracılığıyla elde edilen kümelerden *üçüncüsü* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$). Çizelge 4.21'de gösterilen model aynı zamanda uygun olarak da tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,902 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir.

Çizelge 4.21. Gumbel kopula üçüncü kümesindeki değişkenlerin anlamlılıkları

Potasyum	p < 0,05 *****
Kan Üre Azotu	p < 0,05 *****
Anyon Açığı	p < 0,05 *****

CoClust kümelerinin ardından kuyruk bağımlılığı kümelerinin lojistik modelleri incelenmiştir. Bunlardan ilki *tam bağlantı formülü* elde edilen altı adet kümeden *ilki* ile yapılan lojistik modelleme Çizelge 4.22'dedir. Bu modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$).

Çizelge 4.22. Tam bağlantı formülünün ilk kümesindeki değişkenlerin anlamlılıkları

Cinsiyet	p = 0,649	Glikoz	p = 0,536
Mekanik Solunum	p < 0,05 *	Pıhtılaşma 1	p = 0,345
Karaciğer 1	p = 0,804	Pıhtılaşma 2	p = 0,545
Karaciğer 2	p = 0,766	Pıhtılaşma 3	p = 0,992
Karaciğer 3	p = 0,080	Pıhtılaşma 4	p = 0,398
Karaciğer 4	p = 0,510	Kardiyovasküler 1	p < 0,05 *
Böbrek 1	p < 0,05 *	Kardiyovasküler 2	p < 0,05 *
Böbrek 2	p = 0,126	Kardiyovasküler 3	p = 0,051
Böbrek 3	p = 0,752	Kardiyovasküler 4	p = 0,145
Böbrek 4	p = 0,267	Merkezi Sinir Sistemi 1	p < 0,05 *
Nefes 1	p < 0,05 *	Merkezi Sinir Sistemi 2	p = 0,341
Nefes 2	p < 0,05 *	Merkezi Sinir Sistemi 3	p = 0,782
Nefes 3	p < 0,05 *	Merkezi Sinir Sistemi 4	p = 0,256
Nefes 4	p = 0,554	Sistolik Kan Basıncı	p = 0,936
Kalp Atış Hızı	p = 0,579	Klorür	p < 0,05 *
		Kan Üre Azotu	p = 0,855

Tam bağlantı formülü ile kuyruk bağımlılığı kümelemesinden elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. Elde edilen bu model de anlamlı olarak tespit edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,939 olarak tespit edilmiştir. Çizelge 4.23'te ifade edilen modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir.

Çizelge 4.23. Tam bağlantı formülünün ilk küme anlamlılıklarının yinelenmesi

Mekanik Solunum	p < 0,05 *
Böbrek	p < 0,05 *
Böbrek 1	p < 0,05 *
Nefes	p < 0,05 *
Nefes 1	p < 0,05 *
Nefes 2	p < 0,05 *
Nefes 3	p < 0,05 *
Kardiyovasküler	p < 0,05 *
Kardiyovasküler 1	p < 0,05 *
Kardiyovasküler 2	p < 0,05 *
Merkezi Sinir Sistemi	p < 0,05 *
Merkezi Sinir Sistemi 1	p < 0,05 *
Klorür	p < 0,05 *

Tam bağlantı formülü ile belirlenen kümelerden *ikincisi, üçüncüsü* ve *dördüncüsü* ile yapılan modellemede anlamlı bir model elde edilememiştir (sig>α=0,05).

Tam bağlantı formülü ile elde edilen altı adet kümeden *beşincisi* ve *altıncısı* yapılan lojistik modelleme sonucu anlamlı modeller elde edilmiştir (p<α=0,05). Ancak tek değişkenli modeller olduğu için ilerleyen adımlarda göz ardı edilmiştir. Ayrıca, Çizelge 4.24 ve Çizelge 4.25'de gösterilen modellerin, tek değişkenli olmaları nedeniyle, VIF değerleri de incelenmemiştir.

Çizelge 4.24. Tam bağlantı formülünün beşinci kümesindeki değişkenlerin anlamlılıkları

Laktat	p < 0,05 *****

Çizelge 4.25. Tam bağlantı formülünün altıncı kümesindeki değişkenlerin anlamlılıkları

Beyaz Kan Hücresi	p < 0,05 *****

Ward formülü ile elde edilen kümeler aracılığıyla *24 saatte ölüm* değişkeni için anlamlı bir lojistik model elde edilememiştir (sig>α=0,05).

24 saatte ölüm değişkeni modellemesinden sonra *hastanede ölüm* değişkeni modellenmiştir. İlk olarak CoClust aracılığıyla elde edilen kümeler incelenirken, ardından kuyruk bağımlılığı ile elde edilmiş kümeler incelenmiştir. CoClust ile yine Clayton kopula ve Gumbel kopuladan elde edilen bağımlılık yapıları modellenmiştir.

Clayton kopula aracılığıyla elde edilen kümelerden *ilki* ve *ikincisi* ile yapılan modelleme sonucu anlamlı bir model elde edilememiştir (sig>α=0,05).

Clayton kopula aracılığıyla elde edilen kümelerden *üçüncüsü* ile yapılan modelleme sonucu %95 güven düzeyinde anlamlı bir model elde edilmiştir (p<α=0,05) ve Çizelge 4.26'da gösterilmektedir.

Çizelge 4.26. Clayton kopula üçüncü kümesindeki değişkenlerin anlamlılıkları

Kardiyovasküler 1	p < 0,05 *	Pıhtılaşma 1	p < 0,05 *
Kardiyovasküler 2	p < 0,05 *	Pıhtılaşma 2	p < 0,05 *
Kardiyovasküler 3	p < 0,05 *	Pıhtılaşma 3	p = 0,324
Kardiyovasküler 4	p < 0,05 *	Pıhtılaşma 4	p < 0,05 *
Nefes 1	p < 0,05 *	Böbrek 1	p < 0,05 *
Nefes 2	p < 0,05 *	Böbrek 2	p = 0,719
Nefes 3	p < 0,05 *	Böbrek 3	p = 0,474
Nefes 4	p < 0,05 *	Böbrek 4	p < 0,05 *
Mekanik Solunum	p < 0,05 *		

Clayton kopula ile yapılan kümelemede elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. Elde edilen bu model de anlamlı olarak tespit edilmiştir ($p<\alpha=0,05$). Ancak model uygun olarak tespit edilmemiştir ($p<\alpha=0,05$). Nagelkerke R^2 değeri ise 0,516 olarak tespit edilmiştir. Bu model Çizelge 4.27'de gösterilmektedir.

Çizelge 4.27. Clayton kopula üçüncü küme anlamlılıklarının yinelenmesi

Kardiyovasküler 2	p < 0,05 *	Pıhtılaşma 1	p < 0,05 *
Kardiyovasküler 3	p < 0,05 *	Pıhtılaşma 2	p < 0,05 *
Kardiyovasküler 4	p < 0,05 *	Pıhtılaşma 4	p < 0,05 *
Nefes 1	p < 0,05 *	Böbrek 1	p < 0,05 *
Nefes 4	p < 0,05 *	Böbrek 4	p < 0,05 *
Mekanik Solunum	p < 0,05 *		

Clayton kopula aracılığıyla elde edilen kümelerden *dördüncüsü* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$) ve Çizelge 4.28'de ifade edilmiştir.

Çizelge 4.28. Clayton kopula dördüncü kümesindeki değişkenlerin anlamlılıkları

Potasyum	p < 0,05 *
Laktat	p = 0,256
Beyaz Kan Hücresi	p = 0,886
Glikoz	p = 0,067
Kalp Atış Hızı	p < 0,05 *

Clayton kopula ile yapılan kümelemede elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. Elde edilen bu model de anlamlı olarak

tespit edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,688 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir Çizelge 4.29'da gösterilmiştir.

Çizelge 4.29. Clayton kopula dördüncü küme anlamlılıklarının yinelenmesi

Potasyum	p < 0,05 *****
Kalp Atış Hızı	p < 0,05 *****

Gumbel kopula aracılığıyla elde edilen kümelerden *birincisi* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$). Fakat model uygun olarak tespit edilmemiştir ($p<\alpha=0,05$). Nagelkerke R^2 değeri ise 0,486 olarak tespit edilen bu model Çizelge 4.30'da gösterilmiştir.

Çizelge 4.30. Gumbel kopula birinci kümesindeki değişkenlerin anlamlılıkları

Albümin	p < 0,05 *****
Kırmızı Kan Hücresi	p < 0,05 *****
Hemoglobin	p < 0,05 *****

Gumbel kopula aracılığıyla elde edilen kümelerden *ikincisi* ile yapılan modelleme sonucu Çizelge 4.31'de gösterilmiştir ve model anlamlı olarak tespit edilmiştir ($p<\alpha=0,05$).

Çizelge 4.31. Gumbel kopula kümesindeki değişkenlerin anlamlılıkları

Sodyum	p < 0,05 *****
Bikarbonat	p = 0,337
Klorür	p = 0,313

İkinci küme aracılığıyla elde edilen modelde iki adet anlamlı olmayan değişken modelden çıkartıldığında tek değişken kalacağından bu modelde yineleme yapılmamıştır.

Gumbel kopula aracılığıyla elde edilen kümelerden *üçüncüsü* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,578 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir. Elde edilen model Çizelge 4.32'de görülmektedir.

Çizelge 4.32. Gumbel kopula üçüncü kümesindeki değişkenlerin anlamlılıkları

Potasyum	p < 0,05 *****
Kan Üre Azotu	p < 0,05 *****
Anyon Açığı	p < 0,05 *****

Gumbel kopula aracılığıyla elde edilen kümelerden *dördüncüsü* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$) ve Çizelge 4.33'te ifade edilmektedir.

Çizelge 4.33. Gumbel kopula dördüncü kümesindeki değişkenlerin anlamlılıkları

Kardiyovasküler 1	p < 0,05 *****
Kardiyovasküler 2	p = 0,913
Kardiyovasküler 3	p < 0,05 *****
Kardiyovasküler 4	p = 0,258
Nefes 1	p < 0,05 *****
Nefes 2	p < 0,05 *****
Nefes 3	p < 0,05 *****
Nefes 4	p = 0,756
Mekanik Solunum	p < 0,05 *****

Gumbel kopula ile yapılan kümelemede elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. Elde edilen bu model de anlamlı olarak tespit edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,610 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir. Yinelenen ve anlamlı olarak belirlenen model Çizelge 4.34'de gösterilmiştir.

Çizelge 4.34. Gumbel kopula dördüncü küme anlamlılıklarının yinelenmesi

Kardiyovasküler	p < 0,05 *****
Kardiyovasküler 1	p < 0,05 *****
Kardiyovasküler 3	p < 0,05 *****
Nefes	p < 0,05 *****
Nefes 1	p < 0,05 *****
Mekanik Solunum	p < 0,05 *****

Gumbel kopula aracılığıyla elde edilen kümelerden *beşincisi* ile yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$) ve Çizelge 4.35'te gösterilmektedir.

Çizelge 4.35. Gumbel kopula beşinci kümesindeki değişkenlerin anlamlılıkları

Karaciğer 1	p = 0,576
Karaciğer 2	p = 0,943
Karaciğer 3	p = 0,631
Karaciğer 4	p = 0,985
Pıhtılaşma 1	p < 0,05 *
Pıhtılaşma 2	p < 0,05 *
Pıhtılaşma 3	p = 0,326
Pıhtılaşma 4	p < 0,05 *
Böbrek 1	p < 0,05 *
Böbrek 2	p < 0,05 *
Böbrek 3	p < 0,05 *
Böbrek 4	p < 0,05 *

Gumbel kopula ile yapılan kümelemede elde edilen modelde anlamsız değişkenler modelden çıkartılarak modelleme yenilenmiştir. Elde edilen bu model de anlamlı olarak tespit edilmiştir ($p<\alpha=0,05$). Ancak Çizelge 4.36'da gösterilen bu model, uygun olarak tespit edilmemiştir ($p<\alpha=0,05$). Nagelkerke R^2 değeri ise 0,532 olarak tespit edilmiştir.

Çizelge 4.36. Gumbel kopula beşinci küme anlamlılıklarının yinelenmesi

Pıhtılaşma 1	p < 0,05 *****
Pıhtılaşma 2	p < 0,05 *****
Pıhtılaşma 4	p < 0,05 *****
Böbrek 1	p < 0,05 *****
Böbrek 2	p < 0,05 *****
Böbrek 3	p < 0,05 *****
Böbrek 4	p < 0,05 *****

CoClust kümelerinin ardından kuyruk bağımlılığı kümelerinin lojistik modelleri incelenmiştir. Bunlardan ilki *tam bağlantı formülü* ile elde edilen üç adet kümeden ikincisi ile yapılan lojistik modellemedir. İlk ve üçüncü kümeden gelen değişkenlerden elde edilen modeller anlamlı olarak belirlenmemiştir ($sig>\alpha=0,05$).

İkinci kümeyle yapılan modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,687 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir. Bu model Çizelge 4.37'de gösterilmektedir.

Çizelge 4.37. Tam bağlantı formülünün ikinci kümesindeki değişkenlerin anlamlılıkları

Potasyum	p < 0,05 *****
Olgunlaşmamış Nötrofil Hücreler	p < 0,05 *****

Çizelge 4.38'de ifade edilen lojistik model *Ward formülü* elde edilen sekiz adet kümeden *ilki* ile yapılan lojistik modelleme sonucudur. Bu modelleme sonucu anlamlı bir model elde edilmiştir ($p<\alpha=0,05$). *İlk* ve *altıncı* kümeler dışındaki kümelerden gelen değişkenlerden elde edilen modeller anlamlı olarak belirlenmemiştir ($sig>\alpha=0,05$).

Çizelge 4.38. Ward formülünün ilk kümesindeki değişkenlerin anlamlılıkları

Cinsiyet	p = 0,531
Böbrek 1	p < 0,05 *****
Böbrek 2	p < 0,05 *****
Böbrek 3	p < 0,05 *****
Böbrek 4	p < 0,05 *****
Potasyum	p = 0,962
Olgunlaşmamış Nötrofil Hücreler	p = 0,244

Ward formülünden gelen *ilk* modelin yinelenmesi sonucu elde edilen ikinci model anlamlı olarak belirlenmemiştir ($sig>\alpha=0,05$).

Altıncı kümeye yapılan modelleme sonucu anlamlı bir model elde edilmiştir ve Çizelge 4.39'da ifade edilmektedir ($p<\alpha=0,05$).

Çizelge 4.39. Ward formülünün altıncı kümesindeki değişkenlerin anlamlılıkları

Anyon Açığı	p < 0,05 *****
Beyaz Kan Hücresi	p = 0,437
Kalp Atış Hızı	p < 0,05 *****

Ward formülünden gelen modelin yinelenmesi sonucu elde edilen *ikinci* model yine anlamlı olarak belirlenmiştir ($p<\alpha=0,05$). Model aynı zamanda da uygun olarak tespit edilmiştir ($sig>\alpha=0,05$). Nagelkerke R^2 değeri ise 0,786 olarak tespit edilmiştir. Modelin VIF değerleri ise 1'e yakın olması nedeniyle çoklu bağlantı sorununun olmadığı tespit edilmiştir. Anlamsız değişkenleri çıkartarak yinelenen modelin son hali Çizelge 4.40'da gösterilmektedir.

Çizelge 4.40. Ward formülünün altıncı küme anlamlılıklarının yinelenmesi

Anyon Açığı	p < 0,05 *****
Kalp Atış Hızı	p < 0,05 *****

Lojistik Regresyon Analizi sonuçlarına göre *24 saatte ölüm* bağımlı değişkeni için dört adet model anlamlı ve uygun olarak belirlenmiştir. Bunlardan üçü CoClust Tekniği ile yapılan kümeleme sonucunda Clayton kopula ile yapılan kümelemeden elde edilen *birinci* ve *dördüncü* küme ile Gumbel kopula ile yapılan kümelemeden elde edilen *üçüncü* kümeden gelen değişkenlerle yapılan modellerdir. *24 saatte ölüm* değişkeni için anlamlı ve uygun olarak belirlenen son model kuyruk bağımlılığı ile yapılan kümelemede Tam bağlantı formülünden gelen *ilk* kümedir.

Hastanede ölüm bağımlı değişkeni için ise beş adet model anlamlı ve uygun olarak belirlenmiştir. Bunlardan üçü CoClust Tekniği ile yapılan kümeleme sonucunda Clayton kopula ile yapılan kümelemeden elde edilen *dördüncü* küme ile Gumbel kopula ile yapılan kümelemeden elde edilen *üçüncü* ve *dördüncü* kümeden gelen değişkenlerle yapılan modellerdir. Kuyruk bağımlılığı ile yapılan kümeleme sonucu elde edilen anlamlı ve uygun modeller ise Tam bağlantı formülünden gelen *ikinci,* Ward formülünden gelen *altıncı* kümedir.

Bağımlı değişkenler için elde edilen anlamlı ve uygun modellerin geçerlilikleri hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi ile incelenecektir.

4.7. Elde Edilen Modellerin Geçerliliklerinin İncelenmesi

Lojistik regresyon aracılığıyla elde edilen modellerde anlamlı ve uygun modeller seçilerek geçerlilikleri hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi ile incelenmiştir.

24 saatte ölüm durumu için anlamlı ve uygun modeller Clayton kopula ailesinden elde edilen *birinci* ve *dördüncü* kümeler, Gumbel kopula ailesinden elde edilen *üçüncü* küme ve kuyruk bağımlılığı yönteminde Tam bağlantı formülünden elde edilen *ilk* kümedir. Bu kümeler Çizelge 4.41'de gösterilmektedir.

Çizelge 4.41. 24 saatte ölüm değişkeni için anlamlı ve uygun modeller

Clayton 1	Clayton 4	Gumbel 3	Tam bağlantı 1
Albümin Kırmızı Kan Hücresi Anyon Açığı	Potasyum Kalp Atış Hızı	Potasyum Kan Üre Azotu Anyon Açığı	Mekanik Solunum Böbrek 1 Nefes 1 Nefes 2 Nefes 3 Kardiyovasküler 1 Kardiyovasküler 2 Merkezi Sinir Sistemi 1 Klorür

Hastanede ölüm durumu için anlamlı ve uygun modeller Clayton kopula ailesinden elde edilen *dördüncü* küme, Gumbel kopula ailesinden elde edilen *üçüncü* ve *dördüncü* kümeler ile kuyruk bağımlılığı yönteminde Tam bağlantı formülünden elde edilen *ikinci* küme ve Ward formülünden elde edilen *altıncı* kümedir. Bu kümeler Çizelge 4.42'de belirtilmiştir.

Çizelge 4.42. Hastanede ölüm değişkeni için anlamlı ve uygun modeller

Clayton 4	Gumbel 3	Gumbel 4	Tam bağlantı 2	Ward 6
Potasyum Kalp Atış Hızı	Potasyum Kan Üre Azotu Anyon Açığı	Kardiyovasküler 1 Kardiyovasküler 3 Nefes 1 Mekanik Solunum	Potasyum Olg. Nötrofil Hücreler	Anyon Açığı Kalp Atış Hızı

Anlamlı ve uygun modeller incelendiğinde kuyruk bağımlılığı ile kopulalar aracılığıyla elde edilen kümelerde tüm değişkenler kümelenmiştir. Ancak CoClust tekniği aracılığıyla elde edilen kümelerde tekniğin doğası gereği tüm değişkenler kümelenmemektedir. Yalnızca, ilişkili değişkenler kümelere atanmaktadır.

CoClust tekniği ile elde edilen kümelerde *yaş* ve *cinsiyet* gibi değişkenler kümelerin ilk halinde dahi kümelere atanmamıştır. Bir başka deyişle, diğer değişkenlerle ilişkisiz bulunmuştur. Kuyruk bağımlılığı ile kopulalar aracılığıyla yapılan kümeleme sonucu ise bu değişkenler belirli kümelere atanmıştır. Bu değişkenlerin bulunduğu kümeler aracılığıyla elde edilen modeller anlamlı ve uygun bulunsa da bu değişkenler anlamlı bulunmayarak model dışında bırakılmıştır.

İlgili teknikler aracılığıyla elde edile anlamlı ve uygun modellere göre, CoClust tekniğinin yalnızca ilişkili bulduğu değişkenleri kümeleme özelliğinin modelleme hızını artırmakla birlikte, çalışmalarda olumlu bir sonuç yaratabileceği görülmüştür. Bu kapsamda, literatürde yeni olan CoClust tekniğinin avantajlarından birisi tespit edilmiştir.

Lojistik modeller tarafından anlamlı uygun modellerin geçerlilikleri hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi aracılığıyla ilerleyen bölümlerde incelenecektir.

4.7.1. Hata matrisi sonuçları

Oluşturulan kümelerden yola çıkarak *24 saatte ölüm* değişkeni için elde edilen lojistik modellerin hata matrisi Çizelge 4.43'te gösterilmiştir. Sütunlarda veri setinden elde edilen gözlemler bulunurken, satırlarda elde edilen modeller aracılığıyla elde edilen tahminler bulunmaktadır.

Çizelge 4.43. 24 saatte ölüm değişkeni için hata matrisi

			Gözlenen	
			0	**1**
Tahmin Edilen	**Clayton 1**	**0**	37176	328
		1	155	356
	Clayton 4	**0**	37179	308
		1	152	376
	Gumbel 3	**0**	37196	409
		1	135	275
	Tam bağlantı 1	**0**	37175	299
		1	156	385

Elde edilen hata matrisinden yola çıkılarak elde edilen doğruluk, duyarlılık ve özgüllük değerleri Çizelge 4.44'te gösterilmektedir. Buna göre Clayton ile gelen *ilk* kümeden elde edilen modelin doğruluk oranı yaklaşık %98'dir. Modelin ölen hastaları ölen, yaşayan hastaları yaşayan olarak belirleme yetkinliğini doğruluk oranından gözleyebiliriz. Gumbel ile gelen *üçüncü* kümeden elde edilen lojistik modelin duyarlılığı ise %99,64 olarak belirlenmiştir. Yaşayan hastaların yaşayan olarak belirlenmesini ise duyarlılık ile belirleyebiliriz. Modeller arasında özgüllük değeri en düşük olan model Gumbel kopuladan gelen *üçüncü* küme elde edilen lojistik modeldir. Ölen hastaların ölen olarak tahmin edilmesindeki başarıyı gösteren özgüllük değeri, burada başarı oranındaki düşüklüğü ifade etmektedir.

Çizelge 4.44. Elde edilen hata matrisine dair bilgiler

	Clayton 1	**Clayton 4**	**Gumbel 3**	**Tam bağlantı 1**
Doğruluk	0,9873	0,9879	0,9857	0,9880
Hata Oranı	1-0,9873	1-0,9879	1-0,9857	1-0,9880
Duyarlılık	0,9958	0,9959	0,9964	0,9958
Özgüllük	0,5205	0,5497	0,4020	0,5629

Oluşturulan kümelerden yola çıkarak *hastanede ölüm* değişkeni için elde edilen lojistik modellerin hata matrisi Çizelge 4.45'te gösterilmiştir. Sütunlarda veri setinden elde edilen gözlemler bulunurken, satırlarda elde edilen modeller aracılığıyla elde edilen tahminler bulunmaktadır.

Çizelge 4.45. Hastanede ölüm değişkeni için hata matrisi

			Gözlenen	
			0	**1**
Tahmin Edilen	**Clayton 4**	**0**	30618	4087
		1	757	2553
	Gumbel 3	**0**	30554	3992
		1	821	2648
	Gumbel 4	**0**	30730	2884
		1	645	3756
	Tam bağlantı 2	**0**	30449	3790
		1	926	2850
	Ward 6	**0**	30749	3803
		1	626	2837

Hastanede ölüm değişkeni için elde edilen hata matrisinden yola çıkılarak elde edilen doğruluk, duyarlılık ve özgüllük değerleri Çizelge 4.46'da gösterilmektedir. Buna göre Clayton ile gelen *ilk* kümeden elde edilen modelin doğruluk oranı yaklaşık %87,26'dır. Modelin ölen hastaları ölen, yaşayan hastaları yaşayan olarak belirleme yetkinliğini doğruluk oranından gözleyebiliriz. Gumbel ile gelen *dördüncü* kümeden elde edilen lojistik modelin duyarlılığı ise %97,94 olarak tespit edilmiştir. Yaşayan hastaların yaşayan olarak belirlenmesini ise duyarlılık ile belirleyebiliriz. Modeller arasında özgüllük değeri en düşük olan model Clayton kopuladan gelen *dördüncü* küme elde edilen lojistik modeldir. Ölen hastaların ölen olarak tahmin edilmesindeki başarıyı gösteren özgüllük değeri, burada başarı oranındaki düşüklüğü ifade etmektedir.

Çizelge 4.46. Elde edilen hata matrisine dair bilgiler

	Clayton 4	**Gumbel 3**	**Gumbel 4**	**Tam bağlantı 2**	**Ward 6**
Doğruluk	0,8726	0,8734	0,9072	0,8759	0,9097
Hata Oranı	1-0,8726	1-0,8734	1-0,9072	1-0,8759	1-0,9097
Duyarlılık	0,9759	0,9738	0,9794	0,9705	0,9796
Özgüllük	0,3845	0,3988	0,5657	0,4292	0,5795

Elde edilen lojistik modelleri değerlendirilmesinde önemli olan ilk adım hata matrisi değerlendirilmesinin ardından çapraz geçerlilik ölçütü ile devam edilecektir. Elde edilen modellerin ölçütle değerlendirilmesinden sonra ROC eğrisi incelemelerine geçilecektir.

4.7.2. Çapraz geçerlilik ölçütü sonuçları

Oluşturulan kümelerden yola çıkarak *24 saatte ölüm* değişkeni için elde edilen lojistik modellerin çapraz geçerlilik ölçütü Kappa değerleri Çizelge 4.47'de gösterilmiştir. Elde edilen değerlere göre Tam bağlantı formülünün *birinci* kümesinden ve Clayton kopulanın *dördüncü* kümesinden elde edilen lojistik modeller *iyi* olarak değerlendirilir. Gumbel kopulanın *üçüncü* kümesinden elde edilen model ise *orta* olarak değerlendirilmektedir.

Çizelge 4.47. 24 saatte ölüm değişkeni için Kappa değerleri

	Kappa Değerleri
Clayton 1	0,5894991
Clayton 4	0,6144173
Gumbel 3	0,4959443
Tam bağlantı 1	0,6225732

Elde edilen kümelerden yola çıkarak *hastanede ölüm* değişkeni için elde edilen lojistik modellerin çapraz geçerlilik ölçütü Kappa değerleri Çizelge 4.48'de gösterilmiştir. Sonuçlar incelendiğinde Gumbel kopuladan gelen *dördüncü* küme ile Ward formülünden gelen *altıncı* küme aracılığıyla elde edilen lojistik modeller *iyi* olarak değerlendirilirken, diğer modeller *orta* olarak belirlenmiştir.

Çizelge 4.48. Hastanede ölüm değişkeni için Kappa değerleri

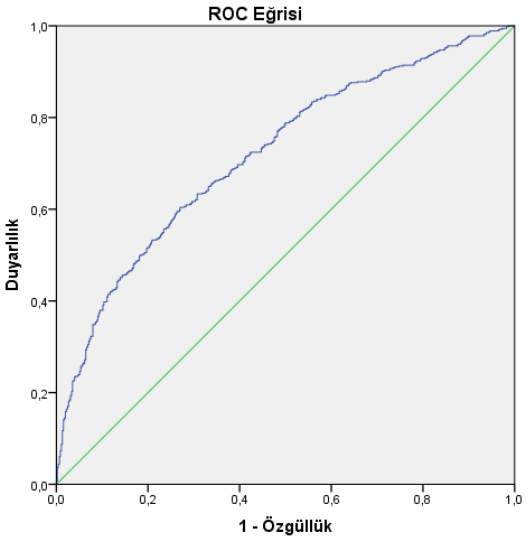
	Kappa Değerleri
Clayton 4	0,4491510
Gumbel 3	0,4590405
Gumbel 4	0,6286660
Tam bağlantı 2	0,4815821
Ward 6	0,6409308

Çapraz geçerlilik ölçütünün ardından bir diğer önemli geçerlilik inceleme tekniği ROC eğrisi sonuçları izleyen bölümde gösterilmektedir.

4.7.3. ROC eğrisi sonuçları

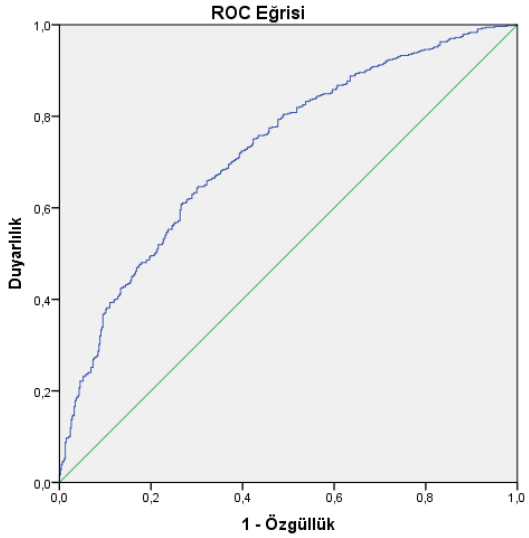
Modellerin geçerliliklerinin sınanmasında ilk adım ROC eğrisi aracılığıyla değerlendirmektir. Eğri altındaki alan ve alanın anlamlılığı üzerinden modeller değerlendirilmiştir ve Çizelge 3.4'te gösterilen ayırım gücü değerlerine göre yorumlanmıştır.

24 saatte ölüm değişkeni için anlamlı ve uygun modellerden Clayton kopula ailesinden elde edilen *birinci* küme için ROC eğrisi altında kalan alan 0,720 ve istatistiksel olarak anlamlıdır ($p<\alpha=0,05$). Şekil 4.3'te görülen eğrinin altında kalana göre model ayırım gücü *kabul edilebilir* olarak değerlendirilmiştir.



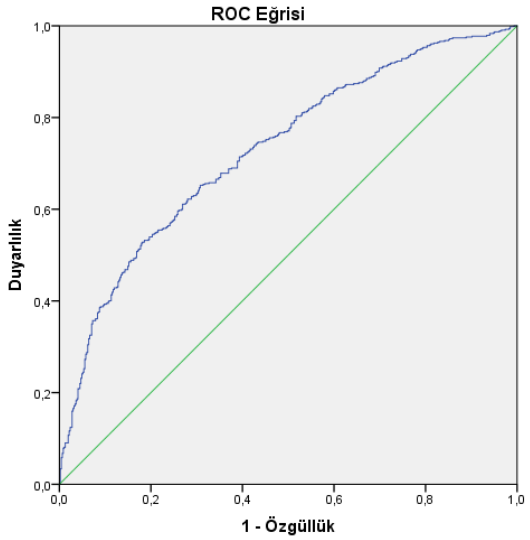
Şekil 4.3. Clayton kopula ailesi birinci kümesi için ROC eğrisi

Clayton kopula ailesinden elde edilen *dördüncü* küme için ROC eğrisi Şekil 4.4'te gösterilmektedir ve altında kalan alan 0,725. Eğrinin altında kalana göre model ayırım gücü *kabul edilebilir* olarak değerlendirilmiştir ve istatistiksel olarak anlamlıdır ($p<\alpha=0,05$).



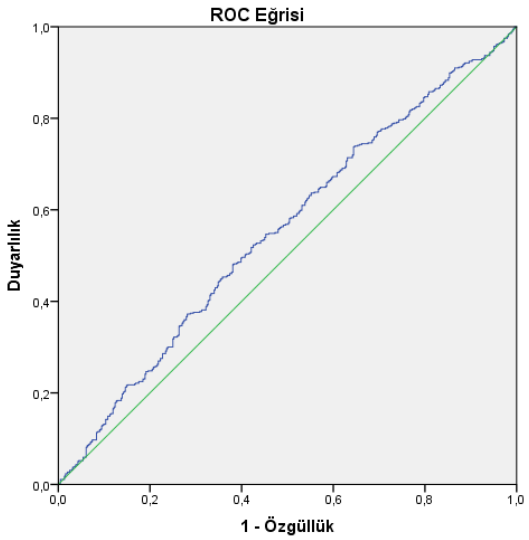
Şekil 4.4. Clayton kopula ailesi dördüncü kümesi için ROC eğrisi

Gumbel kopula ailesinden elde edilen *üçüncü* küme için ROC eğrisi altında kalan alan 0,729 ve istatistiksel olarak anlamlıdır ($p<\alpha=0,05$). Şekil 4.5'te görülen eğrinin altında kalana göre model ayırım gücü *kabul edilebilir* olarak belirlenmiştir.



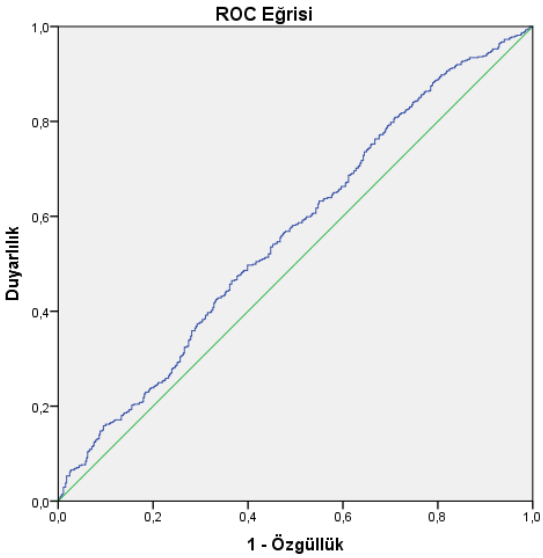
Şekil 4.5. Gumbel kopula ailesi üçüncü kümesi için ROC eğrisi

Kuyruk bağımlılığı yönteminde Tam bağlantı formülü birinci küme için ROC eğrisi altında kalan alan 0,553 ve istatistiksel olarak anlamlıdır ($p<\alpha=0,05$). Ancak, Şekil 4.6'da görülen eğrinin altında kalana göre model ayırım gücü zayıf olarak değerlendirilmiştir.



Şekil 4.6. Tam bağlantı formülü birinci kümesi için ROC eğrisi

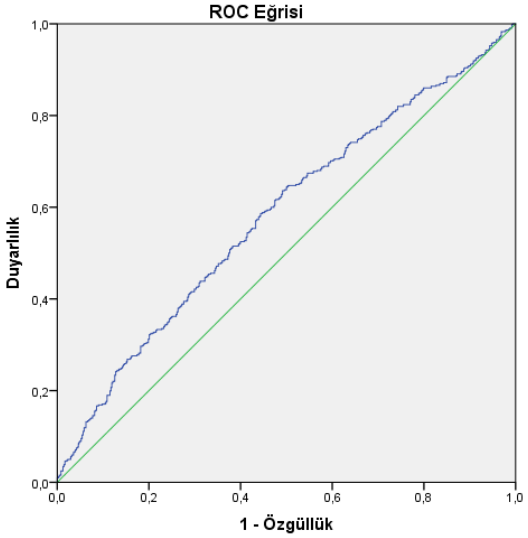
Hastanede ölüm tahmini için anlamlı ve uygun modellerden Clayton kopula ailesinden elde edilen *dördüncü* küme için ROC eğrisi altında kalan alan 0,563 ve istatistiksel olarak anlamlıdır ($p<\alpha=0,05$). Ancak, Şekil 4.7'de görülen eğrinin altında kalana göre model ayırım gücü *zayıf* olarak belirlenmiştir.



Şekil 4.7. Clayton kopula ailesi dördüncü kümesi için ROC eğrisi

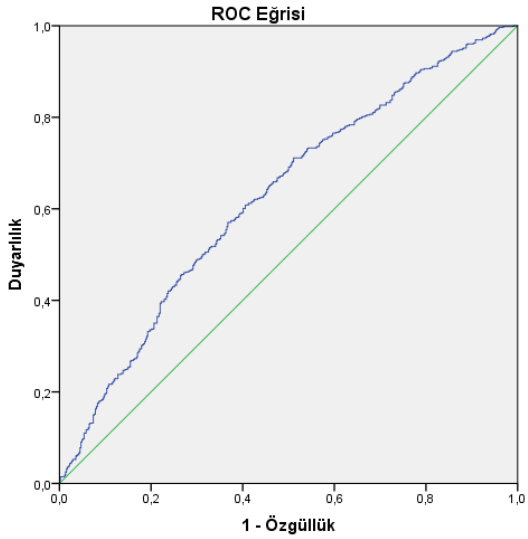
Gumbel kopula ailesinden elde edilen *üçüncü* kümesi için ROC eğrisi Şekil 4.8'de gösterilmiştir. Eğrinin altında kalan alan 0,582 ve istatistiksel olarak anlamlıdır

(p<α=0,05). Ancak, eğrinin altında kalana göre model ayırım gücü *zayıf* olarak değerlendirilmiştir.



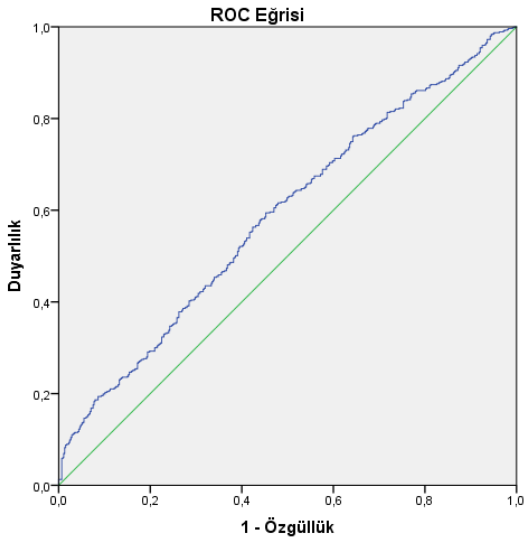
Şekil 4.8. Gumbel kopula ailesi üçüncü kümesi için ROC eğrisi

Gumbel kopula ailesinden elde edilen *dördüncü* kümesi için ROC eğrisi altında kalan alan 0,704 ve istatistiksel olarak anlamlıdır (p<α=0,05). Şekil 4.9'da ifade edilen eğrinin altında kalana göre model ayırım gücü *kabul edilebilir* olarak belirlenmiştir.



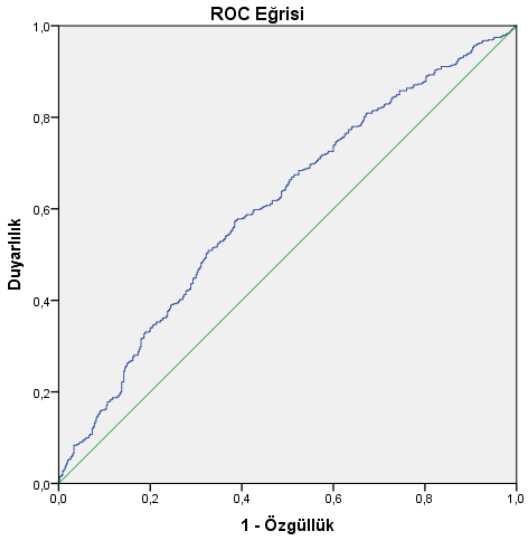
Şekil 4.9. Gumbel kopula ailesi dördüncü kümesi için ROC eğrisi

Kuyruk bağımlılığı yönteminde Tam bağlantı formülü *ikinci* küme için ROC eğrisi Şekil 4.10'da gösterilmiştir. Eğrinin altında kalan alan 0,599 ve istatistiksel olarak anlamlıdır (p<α=0,05). Ancak eğrinin altında kalana göre model ayırım gücü *zayıf* olarak değerlendirilmiştir.



Şekil 4.10. Tam bağlantı formülü ikinci kümesi için ROC eğrisi

Ward formülü *altıncı* küme için ROC eğrisi altında kalan alan 0,707 ve istatistiksel olarak anlamlıdır ($p<\alpha=0,05$). Şekil 4.11'de ifade edilen eğrinin altında kalana göre model ayırım gücü *kabul edilebilir* olarak tespit edilmiştir.



Şekil 4.11. Ward formülü ikinci kümesi için ROC eğrisi

Lojistik Regresyon Analizi sonuçlarına göre *24 saatte ölüm* bağımlı değişkeni için model anlamlı ve uygun olarak dört adet model ROC eğrisi ile değerlendirilmiştir. CoClust Tekniği ile yapılan kümeleme sonucunda Clayton kopula ile yapılan kümelemeden elde edilen *birinci* ve *dördüncü* kümeden üretilen lojistik model ile Gumbel kopula ile yapılan kümelemeden elde edilen *üçüncü* kümeden elde edilen lojistik model *kabul edilebilir*

olarak belirlenirken; Tam bağlantı formülünden gelen *ilk* kümeden elde edilen lojistik modelin ROC eğrisine göre değerlendirilmesi *zayıf* olarak tespit edilmiştir.

Hastanede ölüm bağımlı değişkeni için ise beş adet model anlamlı ve uygun olarak belirlenmişti. CoClust Tekniği ile yapılan kümeleme sonucunda Clayton kopula ile yapılan kümelemeden elde edilen *dördüncü* küme ile Gumbel kopula ile yapılan kümelemeden elde edilen *üçüncü* kümeden yola çıkılarak üretilen lojistik modeller ROC eğrisine göre değerlendirilmesi *zayıf* olarak değerlendirilirken ve Gumbel kopuladan gelen *dördüncü* küme ile elde edilen gelen lojistik modelin ROC eğrisine göre değerlendirilmesi *kabul edilebilir* olarak yorumlanmıştır. Kuyruk bağımlılığı ile yapılan kümeleme sonucu elde edilen anlamlı ve uygun modeller değerlendirildiğinde ise Tam bağlantı formülünden gelen *ikinci* küme aracılığıyla elde edilen lojistik model *zayıf;* Ward formülünden gelen *altıncı* küme ile edilen lojistik model ise *kabul edilebilir* olarak değerlendirilmiştir.

4.8. İstatistiksel Olarak Anlamlı ve Geçerli Modellerin İncelenmesi

Lojistik Regresyon Analizi ile elde edilen modellere dair anlamlılık, uygunluk ve Nagelkerke R^2 değerleri sonuçları değerlendirmek açısından oldukça önemlidir. Ancak bu sonuçlar modellerin kullanılabilirliği konusunda yeterli bilgi sağlamamaktadır. Bu nedenle ilgili modeller hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi ile incelenmiştir.

Modellerin değerlendirilmesi açısından, hata matrisinden elde edilen yüksek doğruluk ve duyarlılık değerleri ile yine yüksek Kappa ve AUC değerleri oldukça önemlidir.

Hata matrisine göre, *24 saatte ölüm* değişkeni için tüm modellerin doğruluk ve duyarlılık değerleri oldukça yüksek olmakla birlikte en yüksek özgüllük değeri %56,29 ile Tam bağlantı formülünden gelen *birinci* küme elde edilen lojistik modele aittir. *Hastanede ölüm* değişkeni için en yüksek doğruluk değeri yaklaşık %90 ile Gumbel kopuladan gelen *dördüncü,* Ward formülünden gelen *altıncı* kümeden elde edilen lojistik modellere aittir. Özgüllük değerleri incelendiğinde ise yine en yüksek değerler bu modellere aittir.

Çapraz geçerlilik ölçütüne göre ise, *24 saatte ölüm* değişkeni için en yüksek Kappa değerleri Clayton kopula *dördüncü* ve Tam bağlantı formülü *birinci* kümeye aittir, bu modeller *iyi* olarak değerlendirilmektedir. Diğer modeller *orta* olarak ifade edilebilir.

Bununla birlikte, *hastanede ölüm* değişkeni için *iyi* olarak değerlendirilen modeller Gumbel kopuladan gelen *dördüncü* model ile Ward formülünden gelen *altıncı* modeldir.

Son olarak, geçerlilik konusunda kullanılan ölçüt ROC eğrisidir. *24 saatte ölüm* değişkeni için oluşturulan eğrilerin incelenmesi sonucu Clayton kopuladan gelen *birinci* ve *dördüncü* model ile Gumbel kopuladan gelen *üçüncü* model için ROC eğrisi altında kalan alana göre modeller *kabul edilebilir* olarak belirlenmiştir. *Hastanede ölüm* değişkeni için incelenen ROC eğrilerine göre ise *kabul edilebilir* olarak belirlenen modeller Gumbel kopuladan gelen *dördüncü* model ile Ward formülünden gelen *altıncı* modeldir.

Geçerlilik sonuçlarına göre, anlamlı ve uygun modeller arasından her bir bağımlı değişken için geçerlilik oranı yüksek modeller seçilmiştir. Bağımlı değişkenler buna göre değerlendirilmiştir.

Modellerin değerlendirilmesinde referans kategori 0 ile kodlanan "*ölüm yok*" kategorisi olarak belirlenmiştir. Anlamlı ve uygun modellerde kullanılan diğer kategorik değişkenler için de referans kategori 0 ile kodlanan kategoriler olarak belirlenmiştir. Öte yandan *24 saatte ölüm* değişkeni Y_1 ile *hastanede ölüm* değişkeni ise Y_2 ile gösterilmektedir.

24 saatte ölüm bağımlı değişkeni için belirlenen modeller Clayton kopuladan elde edilen *birinci* ve *dördüncü* model ile Gumbel kopuladan elde edilen *üçüncü* modeldir.

Hastanede ölüm bağımlı değişkeni için belirlenen modeller ise Gumbel kopuladan elde edilen *dördüncü* model ile Ward formülünden elde edilen *altıncı* modeldir.

24 saatte ölüm bağımlı değişkeni için Clayton kopuladan elde edilen *birinci* küme için lojistik model Çizelge 4.49'da gösterilmektedir.

Çizelge 4.49. 24 saatte ölüm için Clayton kopula birinci kümesinin lojistik modeli

		B	**Standart Hata**	**Exp(B)**
1	Albümin	0,363	0,075	1,439
2	Kırmızı Kan Hücresi	0,056	0,007	1,058
3	Anyon Açığı	0,074	0,011	1,077

Clayton kopuladan elde edilen *birinci* kümenin modellenmesi ile oluşturulan modelde bulunan değişkenlerden *kırmızı kan hücresi* ve *anyon açığı* değişkenleri

APACHE II skorundan gelirken, *albümin* ise değişkeni literatürden elde edilen hayati değişkenlerdendir.

Çizelge 4.49'da ifade edilen model Eşitlik 4.1 ile ifade edilmektedir. Buna göre kırmızı kan hücresi değişkenindeki değişim, ölüm olasılığını 1,058 kat artırırken, anyon açığındaki değişim 1,077 kat artırmaktadır.

$$P\left(Y_1 / X_i\right) = 0,363X_1 + 0,056X_2 + 0,074X_3 \qquad (4.1)$$

Yılmaz vd. (2014)'nin laktat, glukoz, PaO_2/FiO_2, albümin ve CRP düzeylerinin mortaliteye etkisini Lojistik Regresyon Analizi aracılığıyla araştırdıkları çalışmada albumin değerinin düşmesiyle ölüm riskinin arttığı vurgulanmıştır. Ahn vd. (2014) ise yaşlı hastalar üzerinde yaptığı çalışmada anyon açığı değişkenindeki artışın mortaliteyi artırdığını tespit etmiştir. Yine, Verma ve Qayyum (2017)'de NHANES (National Health and Nutrition Examination Survey) isimli veri tabanında bulunan kanser hastaları ile yürüttüğü çalışmada anyon açığı değişkenindeki artışın mortaliteyi artırdığını belirlemiştir. Öte yandan, albümin ve anyon açığı değişkeninin mortalite üzerinde etkisini birlikte değerlendiren Abramowitz vd. (2012) çalışmasında hem albümin değişkenindeki hem de anyon açığı değişkenindeki artışın mortalitede artışa neden olduğu belirlenmiştir. Barış vd. (2016), kırmızı kan hücrelerindeki artışın hem kalp sorunlarını hem de mortaliteyi artırdığını ifade etmiştir.

Albümin değişkeni için literatürde farklı hastalık tipleri için ve farklı değişkenlere çalışılması durumunda farklı yorumlar söz konusu olsa da *anyon açığı* ile yapılan çalışmaların sonucu elde ettiğimiz modelin sonucuyla örtüşmektedir. *Kırmızı kan hücresi* ve *anyon açığı* değişkenleri için elde edilen sonuçlar literatürle uyuşmaktadır.

24 saatte ölüm bağımlı değişkeni için Clayton kopuladan elde edilen *dördüncü* küme için lojistik model Çizelge 4.50'de ifade edilmektedir.

Çizelge 4.50. 24 saatte ölüm için Clayton kopula dördüncü kümesinin lojistik modeli

		B	**Standart Hata**	**Exp(B)**
1	Potasyum	0,019	0,002	1,019
2	Kalp Atış Hızı	0,549	0,043	1,730

Clayton kopuladan elde edilen *dördüncü* kümenin modellenmesi ile oluşturulan modelde bulunan *potasyum* ve *kalp atış hızı* değişkenleri APACHE II ve SAPS II skorlarından gelmektedir.

Çizelge 4.50'de ifade edilen model Eşitlik 4.2 ile ifade edilmektedir. Modele göre potasyumdaki değişimin ölüm olasılığını 1,019 kat artırdığı gözlenirken kalp atış hızındaki artışın ölüm olasılığını artış miktarı ise 1,730 kattır.

$$P\left(Y_1 / X_i\right) = 0,019X_1 + 0,549X_2 \tag{4.2}$$

Kjeldsen (2010) ve Nakhoul vd. (2015) potasyum değerlerindeki artışın kalpte aritmi sorununa neden olmakla birlikte, mortaliteyi de artırdığını göstermiştir. Zhang vd. (2016) yaptıkları çalışmada aritminin, yani kalp atış hızındaki hem artışın hem de azalışın mortaliteyi artırdığını tespit etmiştir. Buradan, *potasyum* ve *kalp atış hızı* arasındaki ilişki görülmekle birlikte, Clayton kopula aracılığıyla elde edilen modelin sonuçlarının literatürle benzerlik gösterdiği görülmektedir.

24 saatte ölüm bağımlı değişkeni için Gumbel kopuladan elde edilen *üçüncü* küme için lojistik model Çizelge 4.51'de belirtilmiştir.

Çizelge 4.51. 24 saatte ölüm için Gumbel kopula üçüncü kümesinin lojistik modeli

		B	**Standart Hata**	**Exp(B)**
1	Potasyum	0,684	0,038	1,980
2	Kan Üre Azotu	-0,007	0,002	0,993
3	Anyon Açığı	0,092	0,012	1,096

Çizelge 4.51'de ifade edilen model Eşitlik 4.3 ile ifade edilmektedir. Model incelendiğinde, ölüm olasılığını *anyon açığının* 1,096 kat, potasyum değişkeninin 1,980 kat artırdığı ortaya çıkmaktadır. *Kan üre azotunun* artışının ise mortalite olasılığında azalışa neden olduğu söylenebilir.

$$P\left(Y_1 / X_i\right) = 0,684X_1 - 0,007X_2 + 0,092X_3 \tag{4.3}$$

Gumbel kopuladan elde edilen *üçüncü* kümenin modelinde bulunan *kan üre azotu* SAPS II skorunda bulunurken, *anyon açığı* değişkeni APACHE II skorundan gelirken *potasyum* değişkeni her iki skorda da mortalite tespitinde kullanılmaktadır.

Wernly vd. (2018) yaptıkları çalışmada kan üre azotunun mortalite üzerindeki etkisini incelemiştir. Laktat, kreatinin gibi değişkenlerle mortalite arasındaki ilişkiyi Lojistik Regresyon Analizi aracılığıyla inceledikleri çalışmada kan üre azotu ile mortalite arasında aynı yönlü bir ilişki olduğu belirlenmiştir. Lee vd. (2016) ise yürüttükleri çalışmada kan üre azotu ve anyon açığı arasında pozitif yönlü bir ilişki tespit ederken, mortalite ile de doğrudan ilişkilerini açıklamıştır. Kan üre azotu ve anyon açığı birlikte arttıkça mortalite olasılığını da artırmaktadır. Potasyum, kreatinin ve kan üre azotunun birlikte mortalite üzerindeki etkisinin incelendiği Fang vd. (2000) çalışmasında potasyum ve kan üre azotu arttıkça mortalite olasılığının arttığı tespit edilmiştir.

Gumbel kopula aracılığıyla elde edilen model ile *kan üre azotu* ile mortalite arasındaki ters yönlü ilişki incelenen çalışmalarla uyuşmamakla birlikte, *potasyum* ve *anyon açığı* değişkenindeki artışların mortalitede artışa neden olması literatürle uyuşmaktadır.

Hastanede ölüm bağımlı değişkeni için Gumbel kopuladan elde edilen *dördüncü* küme için lojistik model Çizelge 4.52'de gösterilmiştir.

Çizelge 4.52. Hastanede ölüm için Gumbel kopula dördüncü kümesinin lojistik modeli

		B	**Standart Hata**	**Exp(B)**
1	Kardiyovasküler 1	-0,067	0,030	0,934
2	Kardiyovasküler 3	0,961	0,036	2,617
3	Nefes 1	0,134	0,030	1,143
4	Mekanik Solunum	-0,165	0,028	0,983

Çizelge 4.52'de ifade edilen model Eşitlik 4.4 ile ifade edilmektedir. İlgili modelde *kardiyovasküler* değişken incelendiğinde arter basıncın <70 olacak şekilde artması durumunda mortalite olasılığında azalış gözlenirken, dopamin ve epinefrin uygulamalarının artması ile mortalite olasılığı 2,617 kat artmaktadır. *Nefes* değişkeni incelendiğinde ise PaO_2 / FiO_2 oranının 400'den fazla olması durumuna göre 399-300 aralığında olması durumunda ise olasılık 1,143 artmaktadır. Kanül vs gibi *mekanik solunum* müdahaleleri ise mortalite olasılığında azalışa neden olmaktadır.

$$P\left(Y_2 / X_i\right) = -0,067X_1 + 0,961X_2 + 0,134X_3 - 0,165X_4 \qquad (4.4)$$

Gumbel kopuladan elde edilen *dördüncü* kümeden elde edilen modelde *kardiyovasküler* değişkeni SOFA skorda kullanılırken, *nefes* ve *mekanik solunum* değişkeni hem SAPS II skorunda hem de SOFA'da kullanılmaktadır.

Kardiyovasküler değişkenini etkileyen vazoaktif ajanların (epinefrin, dopamin vb.) kullanımı ile ilgili olarak vazoaktif ajan kullanımının artışı ile mortalitede de artış gözlendiği Kellum ve Decker (2001), Marik (2002) ve Rauch vd. (2006) çalışmalarında gösterilmiştir. Nefes ve mekanik solunum ile ilgili olarak Fialkow vd. (2016) yürüttükleri çalışmada, nefeste oluşan sorunların mortalite olasılığını artırdığı belirlenirken, kanül gibi mekanik solunum desteklerinin kullanımıyla mortalitenin azaldığı belirtilmiştir.

Gumbel kopula ile elde edilen model *nefes* değişkenini oluşturan PaO_2 / FiO_2 oranındaki düşüş ile mortalitenin artması ve kanül gibi *mekanik solunum* cihazlarının kullanımıyla mortalite olasılığının düştüğü tespit edilmiştir. Buradan elde edilen sonucun literatürle uyuşması ile birlikte *kardiyovasküler* değişkende kısmen benzer sonuç elde edilmiştir. Modele göre vazoaktif ajanların kullanımının bir miktar artışı mortaliteyi azaltmakla birlikte ajan kullanımının artışı ile mortalitede artış gözlenmesi incelenen çalışmalarla benzer sonucu vermektedir.

Hastanede ölüm bağımlı değişkeni için Ward formülünün *altıncı* kümesi aracılığıyla elde edilen lojistik model Çizelge 4.53'te belirtilmiştir.

Çizelge 4.53. Hastanede ölüm için Ward formülünün altıncı kümesinin lojistik modeli

		B	**Standart Hata**	**Exp(B)**
1	Anyon Açığı	0,011	0,001	1,011
2	Kalp Atış Hızı	0,041	0,003	1,041

Çizelge 4.53'te ifade edilen model Eşitlik 4.5 ile ifade edilmektedir. Lojistik modelde *anyon açığı* değişkeninin ölüm olasılığını 1,011 kat artırdığı görülürken, *kalp atış hızının* ise 1,041 kat artırdığı gözlenmiştir.

$$P\left(Y_2 / X_i\right) = 0,011X_1 + 0,041X_2 \tag{4.5}$$

Ward formülünün *altıncı* kümesi aracılığıyla elde edilen modelde anyon açığı değişkeni APACHE II skorundan gelirken, kalp atış hızı değişkeni hem APACHE II ve SAPS II skorunda kullanılmaktadır.

Ahn vd. (2014), Abramowitz vd. (2012), Lee vd. (2016) ve Verma ve Qayyum (2017)'un çalışmalarının anyon açığındaki artışın mortaliteyi artırdığı sonucu bilinmektedir. Ward formülü ile belirlenen modelin anyon açığı değişkeni bazında literatürle uyumlu olduğu görülmektedir. Kjeldsen (2010), Nakhoul vd. (2015) ve Zhang vd. (2016)'nun ise kalp atış hızı ile ilgili olarak aritmi vurgusu dikkat çekicidir. Belirlenen modelde yüksek kalp atışının mortaliteyi artırdığı yorumu yapılabilir.

Elde edilen modeller incelendiğinde *24 saatte ölüm* riskini tespit etmek için APACHE II ve SAPS II skorundan gelen değişkenlerin birlikte verimli çalıştığı tespit edilmiştir. *Hastanede ölüm* riskinin tespitinde ise kuyruk bağımlılığı ile kümeleme aşamasında Ward formülünde yine APACHE II ve SAPS II skorlarının birlikte çalıştığı belirlenmiştir. Öte yandan, SOFA skoru ise SAPS II ile birlikte çalışmasını *hastanede ölüm* riskinin tespitinde CoClust tekniğinde gerçekleştirmektedir.

Buradan ilk 24 saatte ölüm riskinin tespiti için fizyolojik skorların daha etkili olduğu, yoğun bakımda tedavi süresince ölüm durumunun tespitinde ise organ yetmezliği skorunun ve fizyoloji skorunun etkili olduğu yorumu yapılabilir. Öte yandan, literatürde hayati değişken olarak belirlenen değişkenlerde yalnız *albümin* değişkeninin mortalitede etkili olduğu belirlenmiştir. Mortalitenin tespitinde skorlardan yola çıkarak değişken seçiminin daha isabetli olacağı böylelikle söylenebilir.

5. SONUÇ ve ÖNERİLER

Kopula ailelerinde kümeleme tekniklerini incelemek amacıyla yola çıkılan bu çalışmada geniş kapsamlı bir veri seti ile çalışılmıştır. İlgili veri setinden elde edilen değişkenlerin bağımlılık yapısı kopulalar aracılığıyla uygulanan kümeleme teknikleri kullanılarak belirlenmiştir. Teknikler kullanılarak belirlenen ilişkili değişkenler kullanılarak tekniklerin verimlilikleri de incelenmiştir.

Veri setinden elde edilen değişkenler yoğun bakım ünitesinde kullanılan skorlar üzerinden seçilmiştir. Bu skorlarda kullanılan ortak değişkenlerle birlikte farklı değişkenlerin kullanımı da söz konusudur. Bu değişkenlerden bir havuz oluşturularak bağımlılıklar üzerinden değişken seçiminin yapılması hedeflenmiştir. Böylelikle hem MIMIC veri tabanında hem de diğer mortalite tahmin çalışmalarında izlenen yollardan farklı bir yol izlenerek değişken seçimine gidilmiştir.

Değişken seçimi için yoğun bakım hastalarında kullanılan çeşitli skorlama yöntemlerinden en yaygın kullanılanlarından SOFA, APACHE ve SAPS tercih edilmiştir. Bu skorlarda kullanılan değişkenler bir araya getirilerek aralarındaki bağımlılık durumundan yola çıkılarak kümeleme yoluyla değişken seçimine gidilmiştir. Değişken seçimi konusunda kümeleme tekniklerinin belirlemesi dışında çalışmacı tarafından doğrudan müdahalede bulunulmamıştır.

Çalışma, MIT Hesaplamalı Fizyoloji Laboratuvarı tarafından geliştirilen ve araştırmacılara açık olan MIMIC-III veri tabanından elde edilen veri seti ile yürütülmüş ve 38015 hasta ile çalışılmıştır. Elde edilen veri seti literatürden yola çıkılarak düzenlenmiştir. Sadece yetişkin hastalar çalışmaya dahil edilmekle birlikte, Amerikan Sağlık Sigortası Taşınabilirlik ve Sorumluluk Yasası düzenlemesi nedeniyle 89 yaş üzeri hastalar çalışmadan çıkartılmıştır. Büro hatalarına karşılık gelen eksikliklerin olduğu, birden fazla kaydedilme durumunun gözlendiği gözlemler de veri setinden çıkartılmıştır. Eksik gözlemler ise Bayesçi Regresyon aracılığıyla MICE adlı R paketi ile tahmin edilmiştir.

Veride ilgili düzenlemeler yapıldıktan sonra, kopulalar aracılığıyla değişkenler arası bağımlılık yapısı incelenmiştir. Bağımlılık yapısının incelenmesi ise değişkenlerin kümelenmesi ile yapılmıştır. Bu konuda, literatürde oldukça yeni bir yöntem olan CoClust

ile nispeten daha eski ve sıklıkla kullanılan kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniği kullanılmıştır. Böylelikle kopulaların literatüre katkısı olan kümeleme teknikleri birlikte incelenmiş olacaktır.

CoClust ile yapılan kümeleme ile Arşimedyan kopulalar aracılığıyla kümeler elde edilmiştir. Frank kopula ile *iki* küme, Gumbel kopula ile *üç* küme, Clayton kopula ile *beş* küme elde edilmiştir. Ancak burada klasik kümeleme tekniklerinden farklı olarak kümeler kendi içlerinde heterojen bir yapıya sahiptir. Küme satırları arasında bir ilişkiden söz edilebilir. Bu nedenle Frank kopula aracılığıyla elde edilen kümeler değerlendirmeye alınmamıştır.

Öte yandan, kuyruk bağımlılığı aracılığıyla kümeleme tekniği sonucu Ward ve Tam bağlantı formülü ile elde edilen kümeler değerlendirilmiştir. Tam bağlantı formülü ile *altı* küme, Ward formülü ile *sekiz* küme elde edilmiştir. Burada kümeler kendi içlerinde homojen bir yapı oluşturmaktadır.

Elde edilen kümeler, mortalite tahmininde sıklıkla tercih edilen Lojistik Regresyon Analizi ile modellenmiştir. *24 saatte ölüm* ve *hastanede ölüm* değişkenleri için önce CoClust tekniği ile elde edilen kümeler, ardından kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniğinden elde edilen kümeler incelenmiştir.

24 saatte ölüm değişkeni için CoClust tekniğinden elde edilen kümelerden Clayton kopula ile gelen kümelerden *birinci* ve *dördüncü* kümedeki değişkenler aracılığıyla elde edilen modeller ve Gumbel kopula ile gelen kümelerden *üçüncüsü* ile tespit edilen model anlamlı ve uygun olarak belirlenmiştir. Kuyruk bağımlılığı ile kopulalar aracılığıyla yapılan kümelemede ise yalnızca Tam bağlantı formülünden anlamlı bir model elde edilebilmiştir. Bu teknikten gelen kümelerden *birincisi* ile elde edilen model anlamlı ve uygundur.

Hastanede ölüm değişkeni için ise CoClust tekniğinden elde edilen kümelerden Clayton kopula ile gelen kümelerden *dördüncüsü* ile elde edilen model ve Gumbel kopuladan gelen kümelerden *üçüncüsü* ve *dördüncüsü* ile elde edilen modeller anlamlı ve uygundur. Kuyruk bağımlılığı ile kopulalar aracılığıyla yapılan kümelemede Tam bağlantı formülünden gelen *ikinci* küme ile Ward formülünden gelen *altıncı* küme ile elde edilen modeller anlamlı ve uygundur.

Modellerin anlamlılık ve uygunluk yorumu Lojistik Regresyon Analizi üzerinden yapılmıştır. Ancak anlamlı ve uygun modellerin kullanımı ile ilgili son kararı vermek için modellerin geçerlilikleri incelenmiştir. Bu nedenle geçerlilik ve güvenilirlikler ile ilgili olarak hata matrisi, çapraz geçerlilik ölçütü ve ROC eğrisi kullanılmıştır. Buradan tespit edilen sonuçlara göre model katsayıları incelenmiştir.

Hata matrisleri incelendiğinde *24 saatte ölüm* değişkeni için en yüksek doğruluk ve özgüllük değerleri Tam bağlantı formülünden elde edilen *birinci* küme ile elde edilen model için belirlenmiştir. *Hastanede ölüm* değişkeni için ise Ward formülünden elde edilen *altıncı* küme için en yüksek doğruluk ve özgüllük değerleri elde edilmiştir. CoClust tekniğinden gelen kümeler için yapılan modellemelerde yüksek doğruluk oranlarına ulaşılsa da özgüllük oranları kuyruk bağımlılığı ile kopulalar aracılığıyla elde edilen kümelerin özgüllük oranlarına ulaşamamaktadır.

Diğer geçerlilik inceleme yolu olan çapraz geçerlilik ölçütü ile modellerin incelemesi ise Kappa değerleri üzerinden yapılmıştır. Modellerde yüksek Kappa değeri elde etme beklentisi bulunmaktadır. Buradan yola çıkarak, *24 saatte ölüm* değişkeni için en yüksek Kappa değerleri Clayton kopula *dördüncü* küme ve Tam bağlantı formülü *birinci* küme için elde edilmiştir. Bu modeller iyi olarak değerlendirilirken, Gumbel kopuladan gelen *üçüncü* model orta olarak belirlenmiştir. *Hastanede ölüm* değişkeni için ise, Kappa değerlerine göre iyi olarak belirlenen modeller Gumbel kopula *dördüncü* model ve Ward formülü *altıncı* modeldir.

Modeller ile ilgili olarak geçerlilikleri konusunda ROC eğrisi incelemesi ile son karar verilmiştir. *24 saatte ölüm* değişkeni için Tam bağlantı formülü ile elde edilen model zayıf kabul edilirken, CoClust aracılığıyla oluşturulan modeller kabul edilebilir olarak belirlenmiştir. *Hastanede ölüm* değişkeni için elde edilen modeller incelendiğinde, Gumbel kopula *dördüncü* küme ve Ward formülü *altıncı* küme ile yapılan modeller kabul edilebilir olarak tespit edilmiştir.

Lojistik Regresyon Analizi aracılığıyla yapılan modellemede yalnızca analizin sağladığı anlamlılık ve uygunluk incelemesi ile hem *24 saatte ölüm* hem de *hastanede ölüm* değişkeni için CoClust ve kuyruk bağımlılığı teknikleri ile anlamlı ve uygun modeller elde edilmiştir. CoClust tekniği ile verimli modeller Clayton ve Gumbel kopula gibi asimetrik kopula ailelerinden elde edilmiştir. Öte yandan, modellemede kullanılmasa da

Frank kopula gibi simetrik kopula ailesi de küçük de olsa homojen kümeler yaratmaya yardımcı olmuştur. Buradan, parametrik kopula ailelerinin mortalite konusunda verimli bir ortam sağladığı söylenebilir.

İlgili değişkenler için kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniği de anlamlı ve uygun modeller sağlamıştır. Benzerlik oluşturmak yerine benzemezlik tanımlanarak da anlamlı modellere ulaşılabileceği görülmüştür. Yine, bu yaklaşımda alt (sol) ve üst (sağ) kuyruk olmak üzere asimetrik bir yaklaşım söz konusudur. Mortalite tahmininde asimetrik yaklaşımın verimli sonuç verdiği açıkça görülmektedir. Öte yandan, CoClust tekniğinin hem asimetrik hem de simetrik yaklaşımla çözüm olanağı sunması ile kuyruk bağımlılığı tekniğinden ayrıldığını göstermektedir. Bu nedenle, kuyruklarda bağımlılık araştırmasında CoClust'ın daha verimli bir çalışma ortamı sağladığı düşünülmektedir.

Kuyruk bağımlılığı ile kopulalar aracılığıyla kümeleme tekniğinde benzemezlik matrisinin oluşturulmasında iki küme arasındaki en uzak mesafeyi dikkate alan *Tam bağlantı formülü* ve küme içi toplam varyansı en küçük hale getirmeyi hedefleyen *Ward formülü* anlamlı ve uygun modeller sağlamıştır.

24 saatte ölüm değişkeni için elde edilen modellerde bulunan değişkenler incelendiğinde APACHE II ve SAPS II skorlarından gelen değişkenlerin bulunduğu gözlenmiştir. *Hastanede ölüm* değişkeni için elde edilen model incelendiğinde ise, CoClust tekniğinden elde edilen modelde bulunan değişkenler SAPS II ve SOFA'dan gelirken, kuyruk bağımlılığı ile Ward formülü ile elde edilen modelin değişkenleri APACHE II ve SOFA skorundan gelmektedir. Hasta kabulünden itibaren 24 saat içindeki olası mortalite tespiti için APACHE II ve SAPS II skorlarının verimli sonuç verdiği söylenebilir. Yine, APACHE II ve SAPS II skorlarında kullanılan değişkenlerin birlikte iyi çalıştığı yorumu da yapılabilir.

Elde edilen modeller incelendiğinde, *24 saatte ölüm* bağımlı değişkeni için elde edilen modellerde kullanılan değişkenler literatürle karşılaştırılmıştır ve bulgularda ifade edilen karşılaştırmalara aşağıdaki gibi özetlenebilir.

- Clayton kopula birinci kümesinden elde edilen modelde bulunan *albümin*, *kırmızı kan hücresi* ve *anyon açığı* değişkenlerindeki artışın mortaliteyi artırdığının tespit edilmesi literatürde bulunan Abramowitz vd. (2012),

Yılmaz vd. (2014) ve Barış vd. (2016)'nin elde ettiği sonuçlarla uyuşmaktadır.

- Clayton kopula dördüncü kümesinden elde edilen modelde bulunan *potasyum* ve *kalp atış hızı* değişkenlerindeki artışın mortaliteyi artırdığının tespit edilmesi literatürde bulunan Kjeldsen (2010) ve Nakhoul vd. (2015) çalışmalarından elde edilen potasyum sonuçlarıyla uyuştuğu görülmektedir. Ancak, Zhang vd. (2016) incelendiğinde kalp atış hızındaki hem artış hem azalış mortaliteyi artırmaktadır. Burada da kısmen benzer bir sonuçtan söz edilebilir.
- Gumbel kopula üçüncü kümesinden elde edilen modelde bulunan *potasyum* ve *anyon açığı* değişkenlerindeki artışın mortaliteyi artırdığının tespit edilmesi literatürde Lee vd. (2016) ve Fang vd. (2000) çalışmalarında tespit edilen sonuçlarla uyuşmaktadır. Ancak, *kan üre azotu* değişkenin mortalite ile arasındaki ters yönlü ilişki incelenen Wernly vd. (2018) ve Lee vd. (2016) çalışmalarıyla uyuşmamaktadır.

Hastanede ölüm bağımlı değişkeni için elde edilen modellerde kullanılan değişkenler literatürle karşılaştırılmıştır ve aşağıdaki sonuçlara ulaşılmıştır.

- Gumbel kopula dördüncü kümesinden elde edilen modelde bulunan *kardiyovasküler*, *nefes* ve *mekanik solunum* değişkenlerinin mortalite üzerindeki etkisinin tespiti literatürde bulunan Kellum ve Decker (2001), Marik (2002), Rauch vd. (2006) ve Fialkow vd. (2016) çalışmalarının sonuçlarıyla uyuşmaktadır.
- Ward formülünün altıncı kümesinden elde edilen modelde bulunan *anyon açığı* ve *kalp atış hızı* değişkenlerindeki artışın mortaliteyi artırdığının tespit edilmesi literatürde incelenen Ahn vd. (2014), Abramowitz vd. (2012), Lee vd. (2016) ve Verma ve Qayyum (2017)'un çalışmalarının sonuçları ile uyumludur.

24 saatte ölüm bağımlı değişkeni için yalnızca CoClust tekniğinden elde edilen kümelerden gelen değişkenlerle oluşturulan modeller hem anlamlılık ve uygunluk hem de geçerlilik bakımından incelenerek kullanılabilir olarak belirlenmişken, *hastanede ölüm* bağımlı değişkeni için ise CoClust tekniğinden elde edilen bir kümeden ve kuyruk

bağımlılığı ile kümeleme tekniğinden bir kümeden yola çıkılarak oluşturulan model kullanılabilir olarak değerlendirilmiştir.

Her iki bağımlı değişken için de CoClust tekniğinin önemli bir sonuç vererek kullanılabilir bir model sunması, tekniğin yeni olmasına rağmen dikkat çekici olduğu ve gelişime açık olduğunu göstermektedir. Farklı çalışmalarda tekniğin kullanılması ve karşılaştırılması gelişimi açısından önem göstermektedir. Lascio (2008) tarafından yürütülen klinik çalışmada Normallik varsayımıyla yürütülen Gaussian (Normal) kopula ve simetrik kopula ailesi olan Frank kopula ile modelleme yapılmıştır. Ancak, tekniğin güncel sürümü olan Lascio ve Giannerini (2019)'da asimetrik kopula ailesi olan Clayton ve Gumbel aileleri simülasyon sonuçlarında anlamlı modeller elde edilmiştir. Bu kapsamda, hem bu tez çalışmasında elde edilen sonuçlara göre hem de Lascio (2008) ve Lascio (2019)'a göre farklı kopula aileleri ile anlamlı ve uygun sonuçlar elde edilmesi tekniğin gelişime açık olduğunu göstermektedir. Tekniğin çıkış çalışmasının klinik veri olması nedeniyle de bu alanda farklı veri setleri ve tahmin yöntemleri ile bu konuda gelişimin süreceği düşünülmektedir.

CoClust tekniği ile ilgili bir diğer dikkat çekici sonuç ise, tekniğin doğası gereği yalnızca ilişkili olan değişkenleri kümelemesi ile elde edilmiştir. Tekniğin, ilişkisiz bularak küme dışında bıraktığı *yaş*, *cinsiyet* gibi değişkenler kuyruk bağımlılığı aracılığıyla belirli kümelere atanmıştır. Bahsi geçen değişkenlerin atandığı kümeler aracılığıyla elde edilen modeller anlamı ve uygun bulunsa da bu değişkenler anlamlı olarak tespit edilememiş ve model dışında kalmıştır. Buradan CoClust tekniğine dair önemli bir avantaj tespit edilmiştir.

Geçerlilik sonuçlarına göre verimli beş adet model belirlenmiştir. Bu modellerin dördü CoClust tekniği aracılığıyla elde edilen modellerdir. Asimetrik yaklaşımla CoClust'ın oldukça iyi sonuç verdiği böylelikle görülmektedir. Asimetrik yaklaşımın sağladığı çözüm ile mortalite tespiti için kullanılabilecek birden fazla modelin bulunması tekniğin gelişime açık olduğunu göstermektedir.

Diğer yandan, tekniğin mortalite tespitinde uygun modeller vermesi sağlık alanında kullanılabilecek yeni ve uygun bir teknik olduğunu da göstermektedir. Bu kapsamda, mortalite tespitinde skorların da desteğiyle ilgili teknik kullanılarak yeni ve kapsamlı modelleirn gelişiminin önünün açık olduğu görülmektedir.

Bu bağlamda, tekniğin yerel ve global skor çalışmalarında tekrarlanarak kullanımı mortalite tespiti için kullanılacak modellerin gelişiminin önünü açacaktır. Belirlenen modellerle hastalara yönelik kullanımda daha özel skorlama yöntemleri belirlenebilecektir.

KAYNAKLAR DİZİNİ

Abdala, O.T., Saeed, M., 2004, Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted k-nearest neighbors algorithm, Computers in Cardiology, 693-696.

Abramowitz, M.K., Hostetter, T.H., Melamed, M.L., 2012, The serum anion gap is altered in early kidney disease and associates with mortality, Kidney International, 82:6, 701-709.

Agresti, A., 1990, Categorical Data Analysis, Wiley & Sons, New York, p. 734.

Ahn, S.Y., Ryu, J., Baek, S.H., Han, J.W., Lee, Ahn, S., Kim, K., Chin, H.J., Na, K.Y., Chae, D.W., Kim, K.W., Kim, S., 2014, Serum anion gap is predictive of mortality in an elderly population, Experimental Gerontology, 50, February, 122-127.

Albers, W., 1999, Stop-loss premiums under dependence, Insurance: Mathematics and Economics, 24- 3, 173-185.

Alpar, R., 2011, Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Detay Yayıncılık, Ankara, s. 511.

Avutman, Ö., 2011, Yatırım fonu stratejileri arasındaki bağımlılığın copula ile modellenmesi ve bir uygulama, Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, 93 s.

Bairamov, I., Kotz, S., Bekci, M., 2001, New Generalized Farlie-Gumbel- Morgenstern distributions and concomitants of order statistics, Journal of Applied Statistics, Vol. 28, No 5, 521-536.

Barış, S.A., Önyılmaz, T., Uçar, E.K., Çiftçi, T., Başyiğit, İ., Boyacı, H., Yıldız, F., 2016, Serum RDW düzeyinin pulmoner tromboemboli tanılı hastalarda klinik özellikler ve mortalite üzerine etkisi, Kocaeli Medical J, 5;3:18-24.

Bondell, H., Reich, B., 2008, Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR, Biometrics, 64,115–123.

Bouyé, E., Durrleman, V., Bikeghbali, A., Riboulet, G., Roncalli, T., 2000, Copulas for finance–a reading guide and some applications, Working paper, Goupe de Recherche Opérationnelle, Crédit Lyonnais.

Buuren, S.V., Groothuis-oudshoorn, K., 2011, Mice : multivariate imputation by chained, J.Stat. Softw., 45 (3).

Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.H., 2013, Correlated variables inregression: Clustering and sparse estimation, Journal of Statistical Planning and Inference, Volume 143, Issue 11, Pages 1835-1858.

Büyükyılmaz, A., 2011, Bazı kapula tahmin yöntemleri ve Üfe-Tüfe arasındaki bağımlılık yapısı üzerine bir uygulama, Yüksek Lisans Tezi, Akdeniz Üniversitesi, Sosyal Bilimler Enstitüsü, Ekonometri Ana Bilim Dalı, 78 s.

KAYNAKLAR DİZİNİ (devam)

Che, Z., Purushotham, S., Cho, K., Sontag D., Liu, Y., 2018, Recurrent Neural Networks for multıvarıate tıme serıes wıth mıssıng values, Scientific Reports, volume 8, Article Number: 6085, 1-15.

Chen, H., MacMinn, R., Sun, T., Multi-population mortality models: A factor copula approach, Insurance: Mathematics and Economics, Volume 63, July 2015, 135-146.

Cherubini, U., Luciano, E., Vecchiato, W., 2004, Copula Methods in Finance, John Wiley & Sons, p. 310.

Chessa, A., Crimaldi, I., Riccaboni M., Trapin L., 2014, Cluster analysis of weighted bipartite networks: a new copula-based approach, PLoS ONE 9(10):e109507.

Cizek, P., Hardle, W., Weron, W., 2005, Statistical Tools in Finance and Insurance, Chapter 3, Springer, p. 424.

Clayton, D.G., 1978, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika, Volume 65, Issue 1, 141-151.

Cohen, J., 1960, A coefficient of agreement for nominal scales, Educational and Psychological Measurement, Vol 20, Issue 1, 37-46.

Cook , R.D., Johnson, M.E., 1981, A family of distributions for modelling non-elliptically symmetric multivariate data, Journal of the Royal Statistical Society, Series B (Methodological), Volume 43, No. 2, 210-218.

Cossette, H., Gaillardetz, P., Marceau, E., Rioux, J., 2002, On two dependent individual risk models, Insurance: Mathematics and Economics, 30, 153–166.

Cossette, H., Marceau, E., Marri, F., 2008, On the compound Poisson risk model with dependence based on a generalized Farlie-Gumbel-Morgenstern copula, Insurance: Mathematics and Economics, 444-455.

Cossette, H., Marceau, E., Marri, F., 2009, Analysis of ruin measures for the classical compound Poisson risk model with dependence, Scandinavian Actuarial Journal, 3, 221-245.

Cox D.R., Snell E.J., 1989, Analysis of Binary Data, Chapman & Hall, London, p. 240.

Czado, C., Kastenmeıer, R., Brechmann, E.C., Mın, A., 2012, A mixed copula model for insurance claims and claim sizes, Scandinavian Actuarial Journal , 4, 278-305.

Çelebioğlu, S., 2007, Üretici Fiyat Endeksi ve Tüketici Fiyat Endeksi arasındaki bağımlılık yapısı üzerine bir çalışma, 16. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, 56-66.

Danziger, J., Chen, K.P., Lee, J., Feng, M., Mark, R.G., Celi, L.A., Mukamal, K.J., 2016, Obesity, acute kidney ınjury, and mortality in critical ıllness, Crit Care Med., Feb, 44(2), 328-34.

De Luca, G., Zuccolotto, P., 2011, A tail dependence-based dissimilarity measure for financial time series clustering, Adv. Data Anal. Classif. 5(4), 323–340.

KAYNAKLAR DİZİNİ (devam)

De Luca, G., Zuccolotto, P., 2014, Time series clustering on lower tail dependence for portfolio selection. In: M. Corazza, C. Pizzi (eds.) Mathematical and Statistical Methods for Actuarial Sciences and Finance, 131–140.

De Luca, G., Zuccolotto, P., 2017, A double clustering algorithm for financial time series based on extreme events, Statistics & Risk Modeling, Volume 34, Issue 1-2, 1-12.

De Luca, G., Zuccolotto, P., 2017, Dynamic tail dependence clustering of financial time series, Statistical Papers, Volume 58, Issue 3, 641–657.

De Matteis, R., 2001, Fitting copulas to data, Diploma Thesis, Institute of Mathematics of the University of Zurich, 98 p.

Dettling, M., Bühlmann, P., 2004, Finding predictive gene groups from microarray data, Journal of Multivariate Analysis, 90, 106–131.

Devijver, P.A., Kittler, J., 1982, Pattern Recognition: A Statistical Approach, London, GB: Prentice-Hall, 448 p.

Durante, F., Pappada, R., Torelli, N., 2014, Clustering of financial time series in risky scenarios, Adv. Data Anal. Classif. 8, 359–376.

Durante, F., Fernandez-Sanchez, J., Pappada, R., 2015, Copulas, diagonals and tail dependence, Fuzzy Sets and Systems 264, 22–41.

Durante, F., Pappada, R., Torelli, N., 2015, Clustering of time series via non–parametric tail dependence estimation, Statist. Papers, 56(3), 701–721.

Embrechts, P., McNeil, A., Straumann, D., 1999, Correlation: pitfalls and alternatives, Risk, 5, 69–71.

Embrechts, P., Lindskog, F., McNeil, A., 2001, Modelling dependence with copulas and applications to risk management, Zurich: Department of Mathematics, "Handbook of Heavy Tailed Distribution in Finance," Elsevier, 329–384.

Embrechts, P., Lindskog, F., McNeil, A., 2003, Modelling dependence with copulas and applications to risk management, In S Rachev (ed.), "Handbook of Heavy Tailed Distribution in Finance," Elsevier, 329–384.

Escarela, G., Carriére, J.F., 2003, Fitting competing risks with an assumed copula, Statistical Methods in Medical Research, 12(4), 333–349.

Farlie, D.J.G., 1960, The performance of some correlation coeficients for a general bivariate distribution, Biometrika, 47, 307- 323.

Fang, J., Madhavan, S., Cohen, H., 2000, Serum Potassium and cardiovascular mortality, J Gen Intern Med, Dec; 15(12): 885–890.

Feng, M., McSparron, J.I., Kien, D.T., Stone, D.J., Roberts, D.H., Schwartzstein, R.M., Vieillard-Baron, A., Celi, L.A., 2018, Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database, Intensive Care Med., Jun;44(6):884-892.

KAYNAKLAR DİZİNİ (devam)

Fialkow, L., Farenzena, M., Wawrzeniak, I.C., Brauner, J.S., Vieira, S.R.R., Vigo, A., Bozzetti, M.C., 2016, Mechanical ventilation in patients in the intensive care unit of a general university hospital in southern Brazil: an epidemiological study, Clinics (Sao Paulo), 71(3): 145–151.

Frees, E.W., Carriere, J., Valdez, E., 1996, Annuity valuation with dependent mortality, Journal of Risk and Insurance 63, 229-261.

Frees, E.W., Valdez, E. A., 1998, Understanding relationships using copulas, North American Actuarial Journal 2(3), 143-149.

Frees, E.W., Wang, P., 2005, Credibility using copulas, North American Actuarial Journal, 9(2), 31-48.

Friedman, J., Hastie, T., Tibshirani, R., 2009, The Elements of Statistical Learning, Data Mining Inference and Prediction, Springer, 745 p.

Geisser, S., 1993, Predictive Inference, New York, NY: Chapman and Hall, p. 240.

Genest, C., Favre, A.C., 2007, Everything you always wanted to know about copula modeling but were afraid to ask, Journal of Hydrologic Engineering, 12, 347-368.

Giancristofaro, R.A., Salmaso, L., 2003, Model performance analysis and model validation in logistic regression, STATISTICA, anno LXIII, n. 2.

Gönen, M., 2007, Analyzing Receiver Operating Characteristic Curves Using SAS, Cary, NC: SAS Press, 152 p.

Green, D.M., Swets, J.A., 1988, Signal Detection Theory and Psychophysics, Reprint Edition, Los Altos, CA:Peninsula Publishing, 521 p.

Gumbel, E.J., 1960, Bivariate exponential distributions, Journal of American Statistical Association, 55, 698-707.

Hanley, J.A., Negassa, A., deb Edwardes, M.D., Forrester, J.E., 2003, Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation, American Journal of Epidemiology, 157, 4, 364-375.

Hardle, W., Simar, L., 2009, Applied Multivariate Statistical Analysis, Springer, 580 p.

Hartmann, P., Straetmans, S.T.M., De Vries, C.G., 2004, Asset market linkages in crisis periods, Review of Economics and Statistics, 86 (1): 313–326.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P., 2000, 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, Genome Biology 1, 1–21.

Hauksson, H., Dacorogna, M., Domenig, T., Mueller, U., Samorodnitsky, G., 2001, Multivariate extremes, aggregation and risk estimation, Quantitative Finance, 1: 79–95.

KAYNAKLAR DİZİNİ (devam)

Heilpern, S., 2014, Ruin measures for a compound Poisson risk model with dependence based on the Spearman copula and the exponential claim sizes, Insurance: Mathematics and Economics, 251-257.

Hollander, M., Wolfe, D. A., 1973, Non-parametric statistical methods., 46(4), 488-489.

Horton, N.J., Lipsitz, S.R., Orton, N.J.H., Ipsitz, S.R.L., 2001, Multiple imputation in practice: comparison of software packages for regression models with missing variables, Am. Stat. 55 (3), 244–254.

Hosmer, D., Lemeshow, S., 2000, Applied Logistic Regression, Wiley & Sons, New York, p. 528.

Hug, C.W., Szolovits, P., 2009, ICU acuity: real-time models versus daily models, AMIA Symposium Proceedings, 260-264.

Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., Schyns, P.G., 2017, A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula, Hum Brain Mapp., 38(3): 1541–1573.

İyisoy, M.S., 2014, Tanı Test Ölçütlerinde ROC Eğrisi ve Sınıflama Analizlerinin Karşılaştırılmasında Kullanımı, Yüksek Lisans Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüsü, 53 s.

James P.E., 1975, Signal Detection Theory and ROC Analysis, Academic Press, New York, p.30.

Joe, H., 1997, Multivariate Models and Multivariate Dependence Concepts, Chapman and Hall, London, p. 424.

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G., 2016, MIMIC-III, a freely accessible critical care database, Scientific Data, Volume 3, Article number: 160035.

Johnson, A.E.W., Mark, R.G., 2017, Real-time mortality prediction in the Intensive Care Unit, AMIA Annu Symp Proc., 994–1003.

Johnson, A.E.W., Pollard, T.J., Mark, R.G., 2017, Reproducibility in critical care: a mortality prediction case study, Proceedings of the 2nd Machine Learning for Healthcare Conference, in PMLR 68:361-376.

Kaishev, V.K., Dimitrova, D.S., Haberman, S., 2007, Modeling the joint distribution of competing risks survival times using copula functions, Insurance Mathematics and Economics, 41, 339–361.

Kaji, D.A., Zech, J.R., Kim, J.S., Cho, S.K., Dangayach, N.S., Costa, A.B., Oermann, E.K., 2019, An attention based deep learning model of clinical events in the intensive care unit, PLoS ONE 14(2): e0211057.

Karadağ, D.T., 2003, Portfolio Risk Calculation and Stochastic Portfolio Optimization by a Copula Based Approach, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, 216 s.

KAYNAKLAR DİZİNİ (devam)

Karagül, B.Z., 2013, Hayat Dışı Sigortalarda Doğrusal Olmayan Bağımlılığın Kopulalar İle Dinamik Finansal Analizi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, 80 s.

Kaufman, L., Rousseeuw, P.J., 1990, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Series in Probability and Statistics, 342 p.

Keçeoğlu, Ç.R., Gelbal, S., Doğan, N., 2016, Roc eğrisi yöntemi ile kesme puanının belirlenmesi, The Journal of Academic Social Science Studies, Number: 50, 553-562.

Kellum, J.A., Decker J.M., 2001, Use of dopamine in acute renal failure: A meta-analysis, Crit Care Med, 29:1526-31.

Kendall, M., 1957, A Course in Multivariate Analysis, Griffin, London, 185p.

Kızılok, E., 2010, Çok Değişkenli Bağımlı Risklerin Modellenmesi Ve Optimal Aktueryal Kararlar, Ankara Üniversitesi Fen Bilimleri Enstitüsü, 142 s.

Kim, S.H., Yang, H.J., Kim, S.H., Lee, G.S., 204, Physiocover: recovering the missing values ın physiological data of Intensive Care Units, International Journal Of Contents, Vol.10, No.2, Jun, 47-58.

Kimeldorf, G., Sampson, A.R., 1975, Uniform representation of bivariate distributions, Communications in Statistics 4, 617-627.

Kjeldsen, K., 2010, Hypokalemia and sudden cardiac death, Experimental and Clinical Cardiology, 15(4): e96–e99.

Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E., 1985, Apache II: a severity of disease classification system, Critical care medicine 13(10), 818–829.

Kohavi, R., 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 1137–1143.

Kruskal, W.H., 1958, Ordinal Measures of Association, Journal of the American Statistical Association, 53(284), 814-861.

Kumar, P., Shoukri, M.M., 2007, Copula based prediction models: an application to an aortic regurgitation study, BMC Med Res Methodol, 7-21.

Kumar, P., 2010, Probability distributions and estimation of Ali-Mikhail-Haq Copula, Pranesh, Applied Mathematical Sciences, Vol. 4, No.14, 657-666.

Lai, C.D., Xie, M., 2000, A new family of positive quadrant dependent bivariate distributions, Statist. Probab. Letters, 46, 359-364.

Lascio, F.M.L.D., 2008, Analyzing the Dependence Structure of Microarray Data: a Copula–Based Approach, PhD. Thesis, Department of Statistical Sciences, University of Bologna, Italy, 142 p.

Lascio, F.M.L.D., Disegna, M., 2017, A copula-based clustering algorithm to analyse EU country diets, Knowledge-Based Systems 132, 72–84.

KAYNAKLAR DİZİNİ (devam)

Lascio, F.M.L.D., Durante, F., Pappada, R., 2017, Copulas and Dependence Models with Applications, Manuel Úbeda Flores, vd. (eds), Springer, p. 258.

Lascio, F.M.L.D., Giannerini, S., 2019, Clustering dependent observations with copula functions, Statistical Papers February, Volume 60, Issue 1, pp 35–51.

Le Gall, J.R., Lemeshow, S., Saulnier, F., 1993, A new simplified acute physiology score (Saps II) based on a European/North American multicenter study, Jama 270(24), 2957–2963.

Lee, S.W., Kim, S., Na, K.Y., Cha, R., Kang, S.W., Park, C.W., Cha, D.R., Kim, S.G., Yoon, S.A., Han, S.Y., Park, J.H, Chang, J.H., Lim, C.S., Kim, Y.S., 2016, Serum anion gap predicts all-cause mortality in patients with advanced chronic kidney disease: a retrospective analysis of a randomized controlled study, PLoS One, 11(6): e0156381.

Lehmann, E.L., 1975, Nonparametrics: Statistical Methods Based on Ranks, Springer Verlag New York, p. 464.

Li, D.X., 2000, On Default Correlation: A copula function approach, The RiskMetrics Group Working Paper Number 99-07, 43-54.

Liang, K.Y., Zeger, S.L., 1993, Regression Analysis For Correlated Data, Annu. Rev. Pub. Health, 14, 43-68.

MacKenzie, D., Spears, T., 2014, A device for being able to book P&L': The organizational embedding of the Gaussian copula, Social Studies of Science 44 (3), 418–440.

Malevergne, Y., Sornette, D., 2006, Extreme Financial Risks: From Dependence to Risk Management, Springer, 312 p.

Mandelbaum, T., Lee, J., Scott, D.J., Mark, R.G., Malhotra, A., Howell, M.D., Talmor, D., 2013, Empirical relationships among oliguria, creatinine, mortality, and renal replacement therapy in the critically ill. Intensive Care Medicine, Volume 39, Issue 3, 414–419.

Marik, P.E., 2002, Low-dose dopamin: A systematic review, Intensive Care Med, 28, 877-83.

Marshall, D.C., Salciccioli, J.D., Goodson, R.J., Pimentel, M.A., Sun, K.Y., Celi, L.A., Shalhoub, J., 2017, The association between sodium fluctuations and mortality in surgical patients requiring intensive care, J Crit Care,Aug,40, 63-68.

MIT, 2016, Secondary Analysis of Electronic Health Records, Massachusetts Institute of Technology Cambridge, 435 p.

Morgenstern, D., 1956, Einfache Beispiele zweidimensionaler Verteilungen, Mitteilungsblatt für Mathematische Statistik, 8, 234-235.

Murphy, A.H., 1996, The Finley Affair: A Signal Event in the History of Forecast Verification, Weather and Forecasting, 11, (1), 3–20.

KAYNAKLAR DİZİNİ (devam)

Nakhoul, G.N., Huang, H., Arrigain, S., Jolly, S.E., Schold, J.D., Nally, V.J., Navaneethan, S.D., 2015, Serum Potassium, End-Stage Renal Disease and Mortality in Chronic Kidney Disease, American journal of nephrology, 41(6), 456–463.

Nelsen, R.B., 1999, An Introduction to Copulas, Springer-Verlag, New York, 250 p.

Nelsen, R.B., 2006, An Introduction to Copulas, Springer Verlag, New York, 272 p.

Nikoloulopoulos, A.K., Karlis, D., 2009, Finite normal mixture copulas for multivariate discrete data modeling, Journal of Statistical Planning and Inference, 3878-3890.

Oakes, D., 1994, Multivariate survival distributions, J. Nonparametr. Stat., 3, 343–354.

Özbakış, Y.G., 2006, Bazı Kapula Tahmin Yöntemleri Ve Bir Uygulama, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 83 s.

Özdamar, K., 2003, SPSS ile Biyoistatistik. Eskişehir: Kaan Kitabevi, s.498.

Pearl, R., Reed, L.J., 1920, On the rate of growth of the population of the united states since 1790 and its mathematical representation, PNAS June 1, 6, (6), 275-288.

Peres, D.J., Cancelliere, A., 2014, Derivation and evaluation of landslide-triggering thresholds by a Monte Carlo approach, Hydrol. Earth Syst. Sci., 18, 12, 4913-4931.

Peres, D.J., Iuppa, C., Cavallaro, L., Cancelliere, A., Foti, E., 2015, Significant wave height record extension by neural networks and reanalysis wind data, Ocean Modelling, 94, 128–140.

Piotr, J., Fabrizio, D., Wolfgang, H., Tomasz, R., 2009, Copula Theory And Its Applications, Springer, 25-26.

Powers, D.M.W., 2011, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies, 2, (1), 37–63.

Ramasubramanian, K., Singh, A., 2016, Machine Learning Using R, APress, 1 st edition, December 24, 592 p.

Rauch, H., Motsch, J., Böttiger, B.W., 2006, Newer approaches to the pharmacological management of heart failure, Curr Opin Anaesthesiol, 19, 75-81.

Rodríguez, J.A, Flores, M., 2004, A New Class of Bivariate Copulas, Statistical & Probability Letters, Vol. 66, 315-325.

Royston, P., 2004, Multiple imputation of missing values, Stata J., 4, (3), 227–241.

Sadeghi, R., Banerjee, T., Romine, W., 2018, Early hospital mortality prediction using vital signals, Smart Health, Volumes 9–10, 265-274.

Sarıdaş, E.S., 2012, Bağımlı Yaşam Sürelerinin Modellenmesi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, 85 s.

KAYNAKLAR DİZİNİ (devam)

Schweizer, B., Wolff, E.F., 1981, On nonparametric measures of dependence for random variables, Ann. Statist., 9: 870-885.

Scott, W.A., 1955, Reliability of content analysis: The case of nominal scaling, Public Opinion Quarterly, 19(3), 321–325.

Serpa Neto, A., Deliberato, R.O., Johnson, A.E.W., Bos, L.D., Amorim, P., Pereira, S.M., Cazati, D.C., Cordioli, R.L., Correa, T.D., Pollard, T.J., Schettino. G.P.P., Timenetsky, K.T., Celi, L.A., Pelosi, P., Gama de Abreu, M., Schultz, M.J., 2018, Mechanical power of ventilation is associated with mortality in critically ill patients: an analysis of patients in two observational cohorts, Intensive Care Med., Nov, 44(11), 1914-1922.

Sevindik, S., 2009, Farlie-Gumbel-Morgenstern Kapulaları Ve Onların Modifikasyonu, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, Uygulamalı İstatistik Anabilim Dalı, 81 s.

Shi, P., 2016, Insurance ratemaking using a copula-based multivariate Tweedie model, Scandinavian Actuarial Journal, 198-215.

Sklar, A., 1959, Fonctions de repartition á n dimensions et leurs marges, Publ. Inst. Statistique Univ. Paris 8, 229-231.

Song, P., 2000, Multivariate dispersion models generated from Gaussian copula, Scandinavian Journal of Statistics, 27, 305–320.

Stehman, S.V., 1997, Selecting and interpreting measures of thematic classification accuracy, Remote Sensing of Environment, 62, (1), 77–89.

Stretch, R., Della, P.N., Celi, L.A., Landon, B.E., 2018, Effect of boarding on mortality in ICUs, Crit Care Med., Apr, 46(4), 525-531.

Sweeting, P., Fotiou, F.:, 2013, Calculating and communicating tail association and the risk of extreme loss, British Actuarial Journal 18, 13–72.

Thompson, W.R., 2009, Variable Selection of Correlated Predictors in Logistic Regression: Investigating the Diet-Heart Hypothesis, PhD Thesis, Florıda State Unıversıty, College Of Arts And Scıences, 105p.

Trehan, S., Joshi, R.M., 2018, Building and evaluating logistic regression models for explaining the choice to adopt MOOCs in India, International Journal of Education and Development using Information and Communication Technology (IJEDICT), Vol. 14, Issue 1, 33-51.

Trivedi, P. K., Zimmer, D. M., 2005, Copula modeling: an ıntroduction for practitioners, foundations and trends in econometrics, Volume 1, Issue 1, 1-115.

Türker, H., 2016, Lojistik Regresyon Ve Uygulamaları, Yüksek Lisans Tezi, Çanakkale Onsekiz Mart Üniversitesi Fen Bilimleri Enstitüsü, Matematik Anabilim dalı, 120 s.

Verma, A., Qayyum, R., Anion gap and cancer mortality: Insight from NHANES database, Journal of Clinical Oncology, 35, e13068.

KAYNAKLAR DİZİNİ (devam)

Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C.K., Suter, P.M., Thijs, L.G., 1996, The SOFA (Sepsis Related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive care medicine 22(7), 707–710.

Vincent, J.L., Nielsen, N.D., Shapiro, N.I., Gerbasi, M.E., Grossman, A., Doroff, R., Zeng, F., Young, P.J., Russell, J.A., 2018, Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the MIMIC-III database, Ann Intensive Care, Nov, 8(1), 107.

Wang, W., Wells, M.T., 2000, Model selection and semiparametric ınference for bivariate failure-time data, Journal of the American Statistical Association, 95(449), 62–72.

Waudby-Smith, I.E.R., Tran, N., Dubin, J.A., Lee, J., 2018, Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients, PLoS ONE 13(6): e0198687.

Wernly, B., Lichtenauer, M., Vellina, N.A.R., Boerma, E.C., Ince, C., Kelm, M., Jung, C., 2018, Blood urea nitrogen (BUN) independently predicts mortality in critically ill patients admitted to ICU: A multicenter study, Clin Hemorheol Microcirc, 69, 123 -131.

Wüthrıch, M.V., 2006, Extreme value theory and Archimedean copulas, Scandinavian Actuarial Journal, 211-228.

Yadav, M.L., Roychoudhury, B., 2018, Handling missing values: A study of popular imputation packages in R, Knowledge-Based Systems, 160, 104–118.

Yan, J., 2007, Enjoy the joy of copulas: with a package copula, Journal of Statistical Software, Vol.21, Issue 4, 1-21.

Yaprakçı, G., 2007, Kopulalar Teorisinin Finansta Uygulamalar, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, 88 s.

Yılmaz, E., Bor, C., Uyar, M., Demirağ, K., Çankayalı, İ., 2014, Travma hastalarının yoğun bakıma kabulündeki laktat, albümin, C-reaktif protein, PaO_2/FiO_2 ve glukoz düzeylerinin mortaliteye etkisi, Türk Yoğun Bakım Derneği Dergisi, 12: 82-85.

Yule, G.U., 1925, A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. Roy. Soc. Lond., B 213, 21–87.

Zhang, Q., Xiao, M., Singh, V.P., 2015, Uncertainty evaluation of copula analysis of hydrological droughts in the East River basin, China, Global and Planetary Change, 129, 1–9.

Zhang, D., Shen, X., Qi, X., 2016, Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis, Canadian Medical Association Journal, 188(3), e53-e63.

Zhaoa, X., Zhouc, X., 2010, Applying copula models to individual claim loss reserving methods, Insurance: Mathematics and Economics, 46, 290-299.

ÖZGEÇMİŞ

Zeynep İLHAN 1988 yılında Ankara'da doğmuştur. 2006 yılında Hacettepe Üniversitesi Fen Fakültesi Aktüerya Bilimleri Bölümü'nde lisans öğrenimine başlamıştır. 2011 yılında lisans öğrenimini bitirmesinin ardından aynı yıl Eskişehir Osmangazi Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü'nün Olasılık Teorisi Ve Olasılık Süreçleri Anabilim Dalı'na Araştırma Görevlisi olarak atanarak aynı anabilimdalında yüksek lisans eğitimine başlamıştır. Takiben 2014 yılında doktora eğitimine başlamıştır. Halen aynı bölümde çalışmalarına devam etmektedir.