UZUN SÜRELİ VERİLERİN ANALİZİNDE KULLANILAN MAKİNE ÖĞRENMESİ ALGORİTMALARI

MACHINE LEARNING ALGORITHMS FOR LONGITUDINAL DATA ANALYSIS

CAN DEMİRCİGİL

DOÇ. DR. MELİKE BAHÇECİTAPAR
Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin
İstatistik Anabilim Dalı için Öngördüğü
YÜKSEK LİSANS TEZİ
olarak hazırlanmıştır.

2024

ÖZET

UZUN SÜRELİ VERİLERİN ANALİZİNDE KULLANILAN MAKİNE ÖĞRENMESİ ALGORİTMALARI

Can DEMİRCİGİL

Yüksek Lisans, İstatistik Bölümü
Tez Danışmanı: Doç. Dr. Melike BAHÇECİTAPAR

Mayıs 2024, 76 sayfa

Aynı birimlerin zaman boyunca takip edilerek ölçümlerin tekrarlı olarak alınması ile elde edilen veriler "uzun süreli veriler" (boylamsal veriler) olarak adlandırılmaktadır. Tıp, psikoloji, sosyoloji, çevre bilimi vb. alanlarda toplanan uzun süreli veriler, zaman serileri ve klasik regresyon analizlerinden ziyade karma etkiler modelleri gibi özel istatistiksel yöntemlerle analiz edilmektedir.

Son yıllarda popülerliği giderek artan makine öğrenmesi algoritmaları uzun süreli veri kümeleri için de kullanılabilir hale gelmiştir. Bu noktada bilgi teknolojilerinin de etkisiyle R ve Python gibi yazılımlar kullanılarak bu algoritmalar paketler halinde kullanılabilmektedir. Uzun süreli verilerin analizinde başvurulan makine öğrenmesi algoritmaları karma etkiler modellerinin sabit etki parametrelerinin tahmin edilmesi için yararlanılmakta olup bu yöntemler nitel veya nicel sonuçlar ve hayatta kalma süreleri gibi farklı cevapları ele alabilmektedir. Ayrıca, bu yöntemler herhangi bir varsayım gereksinimini çeşitli ölçeklerdeki veya dağılımlardaki değişkenlerle çalışabilmekte ve açıklayıcı değişken sayısının gözlem sayısından daha fazla olduğu çok boyutlu veri kümeleri için de uygundur.

Bu tez çalışmasında, Türkiye'de trafiğe kayıtlı olan motorlu taşıtlara ilişkin araç muayene istasyonlarında derlenen idari kayıt verileri kullanılarak 2013-2023 yılları arasında her iki yılda bir düzenli olarak muayeneye gelen 1569 adet araca ilişkin 5 farklı zaman noktasında ölçümler içeren dengeli bir uzun süreli veri kümesi istatistiksel model olarak karma etkiler modelleri ve yapay zekâ dallarından biri olan makine öğrenmesi algoritmaları ile incelenmiştir. Sağa çarpık bir dağılım gösteren ölçümlere sahip olan araçların yıllara göre katettikleri mesafeler üzerinde yıl, araç cinsi, araçların yakıt türü ve kullanım amacı açıklayıcı değişkenlerinin etkileri istatistiksel ve makine öğrenmesi yöntemleri ile karma etkili modeller oluşturularak incelenmiştir. İstatistiksel yöntemlerden Genelleştirilmiş Doğrusal Karma Etki Modelleri (GDKEM) ile makine öğrenmesi yöntemlerinden Karma Etkili Rastgele Ağaç/Orman, Rastgele Etki Beklenti Maksimizasyonu Ağacı/Ormanı, GDKEM Ağacı ve Gauss Süreci Güçlendirmesi yöntemleri uzun süreli veri kümesi üzerinde farklı bağ fonksiyonları ve kovaryans yapıları düşünülerek uygulanmış ve modeller performans değerlendirme ölçütlerine göre karşılaştırılmıştır. Çalışma sonucunda, AR(1) varyans-kovaryans yapısına sahip Karma Etkili Rastgele Orman algoritmasının tüm istatistiksel ve makine öğrenmesi modelleri içerisinde HKO, HKOK ve OMH model performans değerlendirme ölçütlerine göre en iyi sonuç veren model olduğu sonucuna ulaşılmıştır.

Anahtar Kelimeler: Hibrit yöntemler, Karma etkiler modelleri, Makine öğrenmesi, Uzun süreli veriler.

ABSTRACT

MACHINE LEARNING ALGORITHMS FOR LONGITUDINAL DATA ANALYSIS

Can DEMİRCİGİL

Master of Science, Department of Statistics
Supervisor: Assoc. Prof. Dr. Melike BAHÇECİTAPAR

May 2024, 76 pages

Data obtained by tracking the same units over time and taking measurements repeatedly are called "longitudinal data". Longitudinal data collected in fields such as medicine, psychology, sociology, environmental science, etc. are analysed with special statistical methods such as Mixed Effects Models rather than time series and classical regression analyses.

Machine learning algorithms, which have become increasingly popular in recent years, have become available for longitudinal datasets. At this point, with the effect of information technologies, these algorithms can be used as packages using software such as R and Python. Machine learning algorithms utilized in the analysis of longitudinal data are used to estimate the fixed effect parameters of Mixed Effects Models, and these methods can handle different responses such as categorical or quantitative results and survival times. Furthermore, these methods can work with variables of various scales or distributions without any assumption requirements and are also suitable for multidimensional datasets where the number of explanatory variables is greater than the number of observations.

In this thesis, a balanced longitudinal dataset containing measurements at 5 different time points for 1569 vehicles that regularly come for inspection every two years between 2013 and 2023 using administrative records compiled at vehicle inspection stations for motor vehicles registered to traffic in Türkiye is analysed with Mixed Effects Models as statistical models and machine learning algorithms, one of the branches of artificial intelligence. The effects of the explanatory variables of year, vehicle type, fuel type and purpose of use on the distances travelled by the vehicles according to years, which have measurements showing a right-skewed distribution, are examined by creating Mixed Effects Models with statistical and machine learning methods. Generalized Linear Mixed Effects Models (GLMM) from statistical methods and Mixed Effects Random Tree/Forest, Random Effects Expectation Maximisation Tree/Forest, GLMM Tree and Gaussian Process Boosting methods from machine learning methods were applied on the longitudinal dataset considering different link functions and covariance structures and the models were compared according to performance evaluation criteria. As a result of this study, it is concluded that the Mixed Effect Random Forest algorithm with AR(1) variance-covariance structure is the best model among all statistical and machine learning models according to the MSE, RMSE and MAE model performance evaluation criteria.

Keywords: Hybrid methods, Longitudinal data, Machine learning, Mixed effect models.

TEŞEKKÜR

Çalışma konumun belirlenmesinde ve çalışmamın hazırlanma sürecinin her evresinde değerli emek ve katkıları ile bana destek olan tez danışmanım Sayın Doç. Dr. Melike BAHÇECİTAPAR'a,

Engin bilgi ve tecrübeleriyle kıymetli zamanlarını ayırıp bu çalışmanın geliştirilmesinde bana yardımcı olan değerli hocalarım Sayın Prof. Dr. Serpil AKTAŞ ALTUNAY ve Sayın Doç. Dr. Hatice Tül Kübra AKDUR'a,

Çalışma koşullarının sağlanması noktasında desteklerini esirgemeyerek bana olan inançlarını içtenlikle hissettiğim Sayın Abdullah YÜKSEL ve Sayın Hasan Evrim PINAR'a,

Çalışmamın başından sonuna kadar her ihtiyaç duyduğumda büyük bir özveriyle yardımıma koşan sevgili arkadaşlarım Gökhan ELYILDIRIM, Nuray TUNCER, Elif DUMAN, Levent KARAKAYA ve Gonca BUDAK UÇAN'a,

Beni bugünlere getiren, her türlü maddi ve manevi olanaklarını esirgemeyen, bana her zaman inanan ve cesaretlendiren canım aileme sonsuz teşekkürlerimi sunarım.

Can DEMİRCİGİL
Mayıs 2024, Ankara

İÇİNDEKİLER

ÇİZELGELER DİZİNİ

ŞEKİLLER DİZİNİ

SİMGELER VE KISALTMALAR

AIC	Akaike Information Criteria
BIC	Bayesian Information Criteria
CART	Classification and Regression Tree
DKEM	Doğrusal Karma Etki Modelleri
DVM	Destek Vektör Makineleri
EÇO	En Çok Olabilirlik
EM	Expectation Maximization
GBDT	Gradient Boosting Decision Tree
GDKEM	Genelleştirilmiş Doğrusal Karma Etki Modelleri
GEE	Generalized Estimating Equations
GKERA	Genelleştirilmiş Karma Etkili Regresyon Ağacı
GLMM	Generalized Linear Mixed Models
GMERT	Generalized Mixed Effect Random Tree
GP	Gauss Process
GS	Gauss Süreci
GTD	Genelleştirilmiş Tahmin Denklemleri
HKO	Hata Kareler Ortalaması
HKOK	Hata Kareler Ortalamasının Karekökü
KEÇO	Kısıtlı En Çok Olabilirlik
KEM	Karma Etki Modelleri
KERA	Karma Etkili Regresyon Ağacı
KERO	Karma Etkili Rasgele Orman
LMM	Linear Mixed Effects Models
MAE	Mean Absolute Error
ME-LS-SVM	Mixed Effect Least Squares Support Vector Machine
MEML	Mixed Effect Machine Learning
MERF	Mixed Effect Random Forest
MERT	Mixed Effect Random Tree
ML	Maximum Likelihood
MOB	Model-based-recursive
MSE	Mean Squared Error

OMH	Ortalama Mutlak Hata
REML	Restricted Maximum Likelihood
RE-EM	Random Effect Expectation Maximization
RF	Random Forest
RMSE	Root Mean Squared Error
RO	Rastgele Orman
SMERF	Stochastic Mixed Effect Random Forest
SMERT	Stochastic Mixed Effect Random Tree
SRE-EM	Stochastic Random Effect Expectation Maximization
SVM	Support Vector Machine
TÜİK	Türkiye İstatistik Kurumu
YSA	Yapay Sinir Ağı

1. GİRİŞ

Uzun süreli veriler, zaman boyunca aynı deney/gözlem birimlerinin tekrarlı olarak ölçülmesi ile elde edilmektedir. Uzun süreli çalışmalar deneklerden yalnızca bir kez ölçüm alınarak yürütülen kesitsel verilere dayalı çalışmalardan farklılık göstermektedir [1]. Uzun süreli veriler, özellikle klinik deney çalışmaları, tıp bilimi ve ekonomi alanlarında yürütülen araştırmaların artması sonucu son yıllarda giderek popüler hale gelmiştir [1], [2]. Uzun süreli veri düzeni, her bir birim için birden fazla zaman noktasında ölçülen ölçümlerin mevcut olduğu durum olup tekrarlı ölçümlü veri düzeninin özel bir halidir.

Uzun süreli çalışmaların birçok avantajları vardır. Uzun süreli veri düzeni, bireysel değişimin ölçülebilmesine olanak sağlayan tek tasarımdır. Benzer özelliklere sahip deneklerin zaman boyunca farklı davranabildiği dikkate alınmaktadır. Zamanla değişen açıklayıcı değişkenlerin cevap değişkeni üzerindeki etkisi incelenebilmektedir. Ölçümler arasındaki değişime neden olan faktörler kontrol edilebilmektedir. Buna karşın, uzun süreli çalışmalarda bazı dezavantajlar da söz konusudur. Zaman noktası sayısı çoğunlukla üç veya üçten fazla olduğu için bu çalışmalar kesitsel çalışmalara göre daha fazla emek, maliyet ve zaman gerektirir. Uzun süreli çalışmaların özelliğinden dolayı her bir denekteki ölçümler arasında korelasyon söz konusudur. Bu korelasyon varlığının analizlerde dikkate alınması gerekir. Bu korelasyon ihmal edildiğinde, regresyon parametreleri yanlış tahmin edilmekte ve ölçümlerdeki zaman boyunca değişimin doğru ve etkin bir şekilde tespit edilebilmesi zorlaşmaktadır. Bir diğer dezavantaj ise; kayıp verilerin varlığıdır. Ölçüm sayıları her bir denek için farklılık gösterebilmekte ve bu durum daha karmaşık analizlerin kullanılmasını gerektiren dengeli olmayan uzun süreli veri kümesine neden olmaktadır. Sonuç olarak; uzun süreli veri analizinde, avantajlar ve dezavantajlar düşünülerek uygun analiz yöntemlerine başvurulması gerekmektedir [3].

Uzun süreli veriler birim-yatay (person-level) veya birim-dönem (person-period) veri düzenine göre iki farklı şekilde düzenlenmektedir [4]. Çizelge 1.1'de gösterilen birim-dönem uzun süreli veri düzeninde zaman noktaları, cevap ve açıklayıcı değişkenler sütun olarak yer aldığı için ölçümlerin zaman boyunca nasıl değiştiği detaylı incelenebilmektedir. Bu nedenle, birim-dönem veri düzeni en sık tercih edilen uzun süreli veri düzenidir.

Çizelge 1.1. Birim-dönem uzun süreli veri düzeni

Birim Numarası	Zaman Noktaları	Cevap Değişkeni Ölçümleri	Açıklayıcı Değişkenlerin Ölçümleri		
1	1	y_{11}	x_{111}	…	x_{11p}
1	2	y_{12}	x_{121}	…	x_{12p}
	.		.	…	.
1	n	y_{1n}	x_{1n1}	…	x_{1np}
.	.	.	.	…	.
.	.	.	.	…	.
.	.	.	.	…	.
m	1	y_{m1}	x_{m11}	…	x_{m1p}
m	2	y_{m2}	x_{m21}	…	x_{m2p}
.	.		.	…	.
m	n	y_{mn}	x_{mn1}	…	x_{mnp}

Çizelge 1.1'de, her bir birim için ölçüm zamanlarının aynı olduğu düşünüldüğünde y_{ij} ($i = 1,....,m$ ve $j = 1,....,n$ olmak üzere), i. birimin j. zaman noktasındaki ölçüm değerini, m birim sayısını, n zaman noktası sayısını göstermektedir. Açıklayıcı değişken sayısı p olmak üzere, x_{ijk}, i. birimin j. zaman noktasındaki k. açıklayıcı değişkene ait ölçümü göstermektedir.

Avantajlar ve dezavantajlar dikkate alınarak uzun süreli verilerin analizlerine ilişkin son yıllarda yeni yöntemler geliştirilmiştir. Bu yöntemler popülerliği giderek artan makine öğrenmesi algoritmalarının ve istatistiksel analiz yöntemlerinin birlikte kullanıldığı hibrit (melez) yöntemlerdir.

1.1. Uzun Süreli Veriler için İstatistiksel Yöntemler

Uzun süreli verilerin analizi için marjinal (marginal), geçiş (transition) ve rastgele etkiler (random effects) modellerinden yararlanılmaktadır [5], [6].

Marjinal modeller kitle ortalamasına dayanmaktadır. Kitle ortalamalı parametreler, açıklayıcı değişkenlerin kitle ortalaması üzerindeki etkilerini ifade etmektedir. Birim içi korelasyonun yapısı regresyon katsayılarının yorumlarını değiştirmemektedir. Buna karşın, geçiş ve rastgele etkiler modellerinde regresyon parametreleri tahmin edilirken birim içi korelasyonlar hesaba katılmaktadır [3].

Geçiş modellerinde, bir zaman noktasından diğerine cevaplara ne olduğu ile ilgilenilir ve cevap değişkeninin stokastik bir süreci takip ettiği varsayılmaktadır [3].

Karma modelin özel bir durumu olan rastgele etkiler modelleri, zaman boyunca elde edilen cevaplar arasındaki korelasyon ve birimler arasındaki heterojenliği ifade eden rastgele etkileri birlikte ele alabilmektedir [3].

Korelasyonlu uzun süreli cevaplar için, yarı-olabilirlik yöntemin uzantısı ve bir marjinal model yöntemi olan Genelleştirilmiş Tahmin Denklemleri (GTD) yöntemi kullanılmaktadır [7]. Bu yöntemin avantajı, korelasyon yapısı doğru belirlenmediğinde bile tutarlı ve asimptotik olarak normal tahminlerin elde edilebilmesidir. Varyans-kovaryans veya korelasyon yapısı “sandviç (sandwich)” veya “sağlam (robust)” tahmin edicisi ile tahmin edilmektedir. GTD tahminleri, iteratif bir yarı puanlama algoritması kullanılarak elde edilmektedir. GTD, farklı yapıları barındırabilen esnek bir modeldir. Yaygın olarak kullanılan korelasyon yapıları arasında birinci-dereceden otoregresif (AR(1)), yapısal olmayan (unstructured), bileşik simetri (compound symmetry) ve toeplitz yapıları yer almaktadır [8].

Uzun süreli verilerin dağılım varsayımlarına göre farklı istatistiksel analizler yapılabilmektedir. Hem sabit etkileri hem de rastgele etkileri içeren Karma Etki Modelleri (KEM) (Mixed Effects Models) uzun süreli veriler arasındaki korelasyonu ele alabildiğinden popüler bir analiz yöntemidir [9]. Doğrusal Karma Etkiler Modelleri (DKEM) (Linear Mixed Effects Models, LMM), aynı birimin ölçümleri arasındaki korelasyonun yanı sıra her

bir birime ait rastgele etkilerin de hesaba katılması ile oluşan basit doğrusal modelin bir uzantısıdır [9].

DKEM'de normal dağılımlı sürekli cevap değişkeni analiz edilmektedir. Ancak uzun süreli verilerin doğası gereği her zaman bu varsayımlar sağlanmayabilir. Cevap değişkeninin normal dağılmadığı, kesikli veya kategorik olduğu durumlar ile karşılaşılabilir. Bu durumda DKEM, cevap değişkenini açıklamada yeterli olmamaktadır. DKEM'nin kısıtlayıcı bir varsayımı olan cevapların normal dağılım göstermesi varsayımını gerektirmeyen Genelleştirilmiş Doğrusal Karma Etkiler Modelleri (GDKEM) (Generalized Linear Mixed Effects Models, GLMM) geliştirilmiştir [10]. Özellikle tıp, eczacılık, ekonomi ve çevrebilim alanlarında sık rastlanan dengesiz uzun süreli veriler için sunduğu analiz kabiliyetinden dolayı GDKEM sık tercih edilmektedir [11].

Uzun süreli veriler için KEM'de En Çok Olabilirlik (EÇO) veya Kısıtlı En Çok Olabilirlik (KEÇO) tahmini için beklenti maksimizasyonu (expectation maximization, EM) algoritması veya Newton-Raphson algoritması kullanılmaktadır. Newton-Raphson algoritmasının daha tutarlı tahminler verdiği bulunmuştur [12].

1.2. Makine Öğrenmesi Yöntemleri

Makine öğrenmesi (machine learning), sınıflandırma ve regresyon ağaçları, destek-vektör makineleri ve yapay sinir ağları gibi birçok yöntemi kapsamaktadır. Makine öğrenmesinde veriler arasındaki karmaşık yapıları anlamak ve bir tahmin elde etmek hedeflenmektedir. Denetimli makine öğrenmesi (supervised machine learning), tahmin için bir model oluşturmayı amaçlar. Denetimsiz makine öğrenmesinde (unsupervised machine learning) ise girdiler vardır ancak çıktılar yoktur. En basit ve popüler yöntem, nicel değerleri tahmin etmek için kullanılan doğrusal regresyondur. Doğrusal diskriminant ve lojistik regresyon analizleri nitel ölçümleri tahmin etmek için kullanılmaktadır [13], [14].

Ağaç tabanlı makine öğrenmesi yöntemleri, en yaygın kullanılan denetimli makine öğrenmesi yöntemleri arasındadır. Burada ağacı oluşturan iki özellik dallar ve düğümlerdir. Ağaçların özelliği verilerin grafiksel olarak kolay bir şekilde anlaşılabilmelerini ve yorumlanabilmelerini sağlamaktır. Ağaç tabanlı makine öğrenmesi yöntemleri, bir veri kümesini özyinelemeli (recursion) olarak en etkili şekilde farklı özelliklere göre böler. Bölme işlemi, veri setinden çıkarılan basit karar kurallarının öğrenilmesine dayanır. Karar

Ağaçları (Decision Trees), Torbalama (Bagging), Rastgele Orman (RO) (Random Forest, RF) ve Güçlendirme (Boosting) yöntemleri ağaç tabanlı yöntemlerdir [15], [16], [17]. Sınıflandırma ve Regresyon Ağacı algoritması (Classification and Regression Tree, CART) oldukça sık kullanılmaktadır [15].

Ağaç tabanlı yöntemlerin birçok avantajı vardır: Fazla sayıda açıklayıcı değişken içeren büyük veri kümelerini, çoklu bağlantı sorunlarını dikkate alma zorunluluğu olmadan analiz edebilmektedir. Ayrıca, açıklayıcı değişkenler ile cevap değişkeni arasındaki doğrusal olmayan ilişkileri herhangi bir dağılım varsayımına dayalı olmadan da ele alabilmektedir.

Torbalama yöntemi, regresyon ya da sınıflama problemlerinde çok sayıda bootstrap örneklemlerinin bir araya gelmesiyle oluşan bir topluluk öğrenmesi yöntemidir. Torbalama yöntemi, karar ağaçlarında tahmin varyansını düşürme özelliğine sahiptir [16].

RO, çok sayıda ağaçtan oluşan topluluğun torbalama temeline dayanarak bunların ortalamasını alan popüler bir ağaç tabanlı topluluk öğrenmesi yöntemidir [17].

Destek-vektör Makineleri (DVM) (Support Vector Machine, SVM) denetimli bir makine öğrenmesi yöntemi olup, sınıflandırma ve regresyon problemlerini çözmek için kullanılmaktadır [18]. Bir sınıflandırma ve regresyon tahmin yöntemi olan DVM, en doğru şekilde tahmin yapmak için makine öğrenmesi teorisine dayalıdır [18].

Yukarıda anlatılan makine öğrenmesi yöntemleri, aynı birimlere ait ölçümlerin ilişkili ve farklı birimlerin bağımsız olduğu varsayımına dayanan yalnızca uzun süreli verileri analiz etmek için tasarlanmamıştır. Uzun süreli verilerdeki korelasyonları ve kayıp ölçümleri ele almak için karma etkili hibrit makine öğrenmesi (hybrid mixed effect machine learning) modelleri ortaya çıkmıştır. Karma etkili makine öğrenmesinin ana fikri, sabit etkileri makine öğrenmesi algoritmalarıyla herhangi bir kısıtlayıcı model varsayımı olmaksızın tahmin etmektir [19], [20], [21], [22], [23], [24].

1.3. Uzun Süreli Veriler için Hibrit Yöntemler

Hibrit yöntemler, KEM ve makine öğrenmesi yöntemlerinin birlikte kullanıldığı yöntemlerdir. Hibrit yöntemlere ilişkin yeni çalışmalar yürütülmektedir. Uzun süreli veriler için yapılan analizlerde, KEM veya makine öğrenmesinin birlikte kullanılmasının tek başına kullanılmalarından daha iyi sonuçlar verdiği kanıtlanmıştır. Temel olarak, sabit etkileri tahmin etmek için makine öğrenmesi yöntemleri ve modelin rastgele etkilerini tahmin etmek için ise istatistiksel yöntemler kullanılmaktadır. Hem makine öğrenmesi hem de istatistiksel yöntemlerin avantajları birleştirilerek tahmin ve sınıflandırma performansının iyileştirildiği gösterilmiştir [25], [26], [23], [27], [21], [28].

Uzun süreli verilerin analizinde hibrit makine öğrenmesi yöntemlerinin, ölçüm sayısından daha fazla sayıda açıklayıcı değişkenin modelde yer alması veya açıklayıcı değişkenlerin çok düzeyli kategorik veri yapısına sahip olması durumunda klasik istatistiksel yöntemlere göre daha iyi sonuç verdiği görülmüştür [28]. Ayrıca, kayıp verilerin olması durumunda da bazı hibrit yöntemler kayıp verilerden etkilenmeden yüksek performansta tahminler gerçekleştirebilmektedirler [29].

Uzun süreli verilerin hibrit yöntemler ile analizinde birçok algoritma kullanılmaktadır: Karma Etkili Makine Öğrenmesi (Mixed Effect Machine Learning, MEML) [26], Rastgele Etki Beklenti Maksimizasyonu Ağaçları (RE-EM Ağaçları) (Random Effects Expectation Maximization Trees, RE-EM Trees) [25], Karma Etkili En Küçük Kareler Destek Vektör Makinesi (Mixed Effect Least Squared Support Vector Machine, ME-LS-SVM) [20], Karma Etkili Regresyon Ağacı (Mixed Effect Regression Tree, MERT) [27], Karma Etkili Rastgele Orman (Mixed Effect Random Forest, MERF) [21], Genelleştirilmiş Karma Etkili Regresyon Ağaçları (Generalized Mixed Effect Regression Trees, GMERT) [30], Stokastik Karma Etkili Rastgele Ormanlar (Stokastic Mixed Effect Random Forest, SMERF), Stokastik Karma Etkili Regresyon Ağacı (Stokastic Mixed Effect Regression Tree, SMERT), Stokastik Rastgele Etkili Beklenti Maksimizasyon Ağaçları (Stokastic Random Effect Expectation Maximization Trees, SREEM Trees), Stokastik Rastgele Etkili Beklenti Maksimizasyon Ormanı (Stokastic Random Effect Expectation Maximization Forest, SREEM Forest) [28], Genelleştirilmiş Doğrusal Karma Etki Modeli Ağacı (GDKEM Ağacı) (Generalized Linear Mixed Effect Model Tree, GLMM Tree) [31] ve Gauss Süreci Güçlendirmesi (Gauss Process Boosting, GP Boosting) [29].

1.4. Literatür Taraması

Regresyon ağaçları uzun süreli veriler için ilk kez Segal [24] tarafından uygulanmıştır. Segal [24], uzun süreli verilerde karşılaşılan korelasyon yapılarını ele alarak, kayıp veriler ve zamanla değişen açıklayıcı değişkenler söz konusu olduğunda uygun ağaç boyutunun belirlenmesi gibi veri analitiği konularını incelemiştir.

Cho [20], uzun süreli veriler için ME-LS-SVM yöntemini geliştirmiş ve Luts ve ark. [23], karma etkili DVM sınıflandırıcısı yöntemini tanıtmıştır.

Hajjem ve ark. [27], regresyon ağaçlarını kayıp verilerin söz konusu olduğu uzun süreli çalışmalar için kullanmışlar ve regresyon ağacı yönteminde EM algoritmasını uyarlamışlar ve standart regresyon ağaçları (CART) yöntemlerinin kapsamlı bir versiyonu olan MERT yöntemini tanıtmışlardır. MERT algoritmasının arkasındaki temel mantık, sabit etkiyi tahmin etmek için standart regresyon ağaçlarını kullanmak ve EM algoritmasını kullanarak rastgele etkiyi tahmin etmek amacıyla ağacın her bir düğümü için DKEM kullanmaktır. Böylece, DKEM'yi ve makine öğrenmesi yaklaşımını birleştirmişlerdir. MERT algoritması, rastgele etki göz ardı edilemediğinde diğer yaklaşımlardan önemli ölçüde daha iyi bir performans göstermiştir.

Sela ve Simonoff [25] korelasyonlu uzun süreli verilerde RE-EM Ağacı algoritmasını geliştirmişlerdir. Ortalama fonksiyonunu tahmin etmek için bir ağaç yapısı kullanmışlar ve modele birimlerin rastgele etkilerini dâhil etmişlerdir.

Hajjem ve ark. [21], MERT algoritmasını rastgele ağaçlar yerine RO yöntemini uyarlayarak geliştirmişler ve buna MERF adını vermişlerdir. RO, bootstrap örnekleri aracılığıyla oluşturulmaktadır. MERF algoritması, yakınsayana kadar farklı algoritmalar kullanarak sabit ve rastgele etkileri tahmin eden MERF algoritması ile aynı fikre sahiptir.

Hajjem ve ark. [30], iki düzeyli veya sayılarak elde edilen verileri tahmin etmek için GMERT yöntemini geliştirmiştir. Tasarlanan algoritma kayıp veriler için de çalışabilmektedir. GMERT, RO ve GDKEM gibi diğer modellere göre daha üstün bir performans göstermiştir.

Fokkema ve ark. [31], uzun süreli verilerin analizi için modele dayalı özyinelemeli ayrıştırma (model based recursive partitioning) (MOB) ile GDKEM yöntemlerini birlikte barındıran bir tür GMERT yöntemi önermişlerdir. Yapılan benzetim çalışmasında önerdikleri GMERT yönteminin klasik GDKEM'ye göre daha iyi performans gösterdiği görülmüştür.

Ngufor ve ark. [26], MEML yöntemini önermiştir. MEML kullanılarak tip-2 diyabetli yetişkinlerin glisemik endeksindeki zaman boyunca değişim ölçülmüştür. Bu yöntem, modeldeki sabit etkiler için ağaç yapılarını ve rastgele etkiler için yakınsama olana kadar veya maksimum iterasyon sayısına ulaşana kadar GDKEM yaklaşımını kullanmayı temel almaktadır. GDKEM'nin rastgele etki yapısını hesaba katabilen karma etkili makine öğrenmesi algoritmasını (MEML) kullanmışlardır. MEML'nin klasik GDKEM yöntemi ve diğer makine öğrenmesi algoritmalarından daha iyi sonuç verdiği görülmüştür.

Sigrist [29], uzun süreli verilerde KEM ile Gauss Süreci Güçlendirme, diğer adıyla GS Güçlendirmesi (Gauss Process Boosting, GP Boosting) yöntemini birleştirerek yeni bir model önermiştir. Önerilen bu model ayrıca yüksek düzeyli kategorik verileri de inceleyebilmektedir. Benzetim çalışmalarında ve gerçek veriler üzerinde yapılan uygulamalarda bu modelin mevcut modellere göre daha iyi performans gösterdiği görülmüştür.

Capitaine ve ark. [28], stokastik etkiler içeren uzun süreli veriler için geliştirdikleri RO algoritmalarını tanıtmış ve simülasyon çalışmalarını yapmışlardır. Hajjem ve ark. [27], [21] tarafından önerilen MERT ve MERF algoritmaları ile Sela ve Simonoff [25] tarafından önerilen RE-EM Ağacı algoritmalarına stokastik etkileri de ilave etmiş ve ayrıca stokastik etkileri de barındırabilen yeni bir algoritma olan RE-EM Ormanı algoritmasını geliştirmişlerdir.

Vuuren ve ark. [32], intihar öyküsü olan öğrencilerin gelecekteki intihar eğilimlerini tahmin edebilmek amacıyla RO ve Lasso Regresyon modellerini kullanmışlardır. Çalışma sonunda RO algoritmasının söz konusu veriler üzerinde daha iyi performans gösterdiği görülmüş ve ilk kez bu alanda makine öğrenmesi modellerinden yararlanıldığı vurgulanmıştır.

Cao [33], GDKEM ve makine öğrenmesi modellerini beraber kullanmıştır. GDKEM, RO ve gradyan artırılmış karar ağacı (Gradient Boosting Decision Tree, GBDT) modellerini uygulamış ve makine öğrenmesi modellerinin klasik GDKEM'ye göre daha iyi sonuç verdiğini göstermiştir.

Mens ve ark. [34], zaman boyunca takip edilen psikiyatrik risk verilerinin analizi için lojistik regresyon, K-en yakın komşu (K-Nearest Neigbourhood), sınıflandırma ağacı, RO, gradyan artırma (gradiant boosting) ve DVM algoritmalarından yararlanmışlardır. Tüm modellerin benzer ölçüde performans gösterdiği ancak RO algoritmasının en uygun performans ölçütlerine sahip olduğu sonucuna varmışlardır.

Hu [3], DKEM ile ağaca dayalı yöntemlerden, RO, torbalama, güçlendirme, DVM ve YSA gibi makine öğrenmesi yöntemlerini uzun süreli verilere uygulamıştır. Modellerin ortalama fonksiyonu doğru belirlendiğinde klasik istatistiksel yöntemlerin makine öğrenmesi yöntemlerine göre daha iyi sonuç verdiğini göstermiştir.

Erduran [35], iki farklı uzun süreli veri kümesine DKEM ve hibrit makine öğrenmesi yöntemlerini birlikte uygulamış ve hibrit yöntemlerin klasik yöntemlere göre daha iyi sonuç verdiğini göstermiştir.

Çakar [36], bilişsel veriler üzerinde farklı istatistiksel ve makine öğrenmesi yöntemlerinden olan DKEM, GDKEM Ağacı, RE-EM Ağacı, yansız RE-EM Ağacı, LongCART ve GP Boosting algoritmalarını uygulamış ve model sonuçlarını karşılaştırmıştır. GDKEM Ağacı yönteminin en iyi sonuç verdiği ve hesaplama hızının diğer yöntemlere göre daha düşük olduğu görülmüştür.

Erdoğan [37], Türkiye'deki tüm illere ilişkin 2012-2019 yıllarını kapsayan uzun süreli verileri kullanarak intihar oranlarına etki edebilecek sosyal ve ekonomik faktörleri sabit etki, rastgele etki ve geçiş modelleri gibi klasik istatistikler yöntemler ile MERT, MERF ve RE-EM Ağacı gibi hibrit yöntemleri kullanarak incelemiştir. Çalışma sonucunda hibrit modellerin klasik yöntemlerden daha iyi sonuç verdiği görülmüştür.

1.5. Tezin Amacı

Bu tez çalışmasının amacı, uzun süreli veriler üzerinde son yıllarda birçok alanda popülerliği giderek artan karma etkili makine öğrenmesi algoritmaları ile uzun yıllardır literatürde yer alan klasik istatistiksel yöntemlerin detaylı olarak tanıtılması ve bu yöntemlerin gerçek hayata ilişkin bir uzun süreli veri kümesi üzerinde uygulamasının yapılarak makine öğrenmesi algoritmalarının klasik istatistiksel yöntemlere göre sahip olduğu avantajlarının gösterilmesidir.

Çalışmanın ikinci bölümünde, uzun süreli veriler için kullanılan klasik istatistiksel yöntemlere ilişkin bilgiler verilmiş ve üçüncü bölümde ise karma etkili makine öğrenmesi algoritmaları tanıtılmıştır. Dördüncü bölümde, uzun süreli veriler için kullanılan klasik istatistiksel yöntemler ve karma etkili makine öğrenmesi algoritmaları gerçek bir uzun süreli veri kümesi üzerinde uygulanmış ve beşinci bölümde elde edilen sonuçlar yorumlanmıştır. En iyi model çeşitli performans ölçütlerine göre seçilmiştir.

2. UZUN SÜRELİ VERİLERİN ANALİZİ İÇİN İSTATİSTİKSEL YÖNTEMLER

2.1. Doğrusal Karma Etki Modelleri (DKEM)

DKEM, uzun süreli veriler için geliştirilmiş bir doğrusal regresyon modelidir. DKEM, sadece sabit etkileri içeren ve bağımsız hatalar varsayımını gerektiren klasik doğrusal regresyon modellerinin aksine korelasyonlu verileri ele alabilen ve hem rastgele hem de sabit etkileri içeren bir modeldir [9].

DKEM, Eşitlik (2.1)'deki gibi ifade edilir [9]:

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \boldsymbol{\varepsilon_i}. \tag{2.1}$$

Eşitlik (2.1)'de n_i, i. birimin ölçümlerinin alındığı zaman noktası sayısı ve p açıklayıcı değişken sayısı olmak üzere; $\mathbf{Y_i}$, i. birim için $n_i \times 1$ boyutlu cevap değişkeni vektörünü, $\mathbf{X_i}$, $n_i \times (p+1)$ boyutlu açıklayıcı değişken matrisini, $\boldsymbol{\beta}$, her bir birim için aynı olduğu varsayılan $(p+1)\times 1$ boyutlu bilinmeyen kitle parametreleri vektörünü, $\mathbf{Z_i}$, $n_i \times q$ boyutlu rastgele etkiler matrisini, $\mathbf{b_i}$, q sayıda bilinmeyen rastgele etki parametre vektörünü ve $\boldsymbol{\varepsilon_i}$ $n_i \times 1$ boyutlu birim içi hata terimini ifade etmektedir.

DKEM'de hatalar ve rastgele etkiler birbirinden bağımsızdır ve $\boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \mathbf{R_i})$, $\mathbf{b_i} \sim N(\mathbf{0}, \mathbf{D})$ olduğu varsayılmaktadır. Hataların $\mathbf{0}$ ortalamalı ve $\mathbf{R}_i$ varyans-kovaryans yapısına sahip normal dağılım gösterdiği kabul edilir. Çalışmada rastgele etkilerin ($\mathbf{b_i}$), $\mathbf{0}$ ortalama ve modelde korelasyonlu iki adet rastgele etkili açıklayıcı değişken olması durumunda, $\mathbf{D} = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} \\ \sigma_{b_1 b_0} & \sigma_{b_1}^2 \end{bmatrix}$ varyans-kovaryans matrisi ile normal dağılım gösterdiği varsayılmaktadır.

$\mathbf{b_i}$ rastgele etkilerin ve $\boldsymbol{\varepsilon_i}$ hataların bağımsız olduğu, normal dağıldığı ve DKEM'de doğrusal olduğu varsayımı altında cevap değişkeni vektörü $\mathbf{Y_i}$'nin DKEM için,

$$\mathbf{Y_i} \sim N(\mathbf{X_i\beta}, \mathbf{V_i}) \tag{2.2}$$

şeklinde $\mathbf{X_i\beta}$ ortalamalı ve

$$\mathbf{V_i} = \mathbf{Z_i D Z_i'} + \mathbf{R_i} \tag{2.3}$$

varyans-kovaryans matrisli normal dağılımdan geldiği varsayılır.

EÇO tahmini ve KEÇO yöntemleri, sabit etki parametrelerini ($\boldsymbol{\beta}$) tahmin etmek için kullanılır. Varyans-kovaryans matrislerindeki parametreler $\boldsymbol{\theta}$ ile gösterilmiştir [38].

Çok değişkenli normal dağılımlı $\mathbf{Y}_i$ cevap değişkeninin marjinal fonksiyonu Eşitlik (2.4)'te,

$$f(\mathbf{Y_i} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = (2\pi)^{-n_i/2} \det(\mathbf{V_i})^{-1/2} \exp(-0,5(\mathbf{Y_i} - \mathbf{X_i\beta})^T \mathbf{V_i}^{-1}(\mathbf{Y_i} - \mathbf{X_i\beta})) \tag{2.4}$$

olarak ifade edilir. Eşitlik (2.4)'te, $\boldsymbol{\theta}$ parametresi, $\mathbf{D}$ ve $\mathbf{R_i}$ matrislerinde yer alan varyans-kovaryans parametrelerini içermektedir.

Eşitlik (2.4)'ün olabilirlik fonksiyonu Eşitlik (2.5)'te verilmiştir:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = (2\pi)^{-n_i/2} \det(\mathbf{V_i})^{-1/2} \exp(-0,5(\mathbf{y_i} - \mathbf{X_i\beta})^T \mathbf{V_i}^{-1}(\mathbf{y_i} - \mathbf{X_i\beta})) \,. \tag{2.5}$$

Tüm m sayıda birim dikkate alındığında, ortaya çıkan olabilirlik fonksiyonu her bir birimin ($i = 1, ..., m$) katkısını içeren çarpım fonksiyonu Eşitlik (2.6)'daki gibi oluşturulur:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_i L_i(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_i [(2\pi)^{-n_i/2} \det(\mathbf{V_i})^{-1/2} \exp(-0,5(\mathbf{y_i} - \mathbf{X_i\beta})^T \mathbf{V_i}^{-1}(\mathbf{y_i} - \mathbf{X_i\beta}))] \,. \tag{2.6}$$

Eşitlik (2.6)'nın logaritması alındığında ve $n = \sum_{i=1}^{m} n_i$ toplam gözlem sayısını gösterdiğinde log-olabilirlik fonksiyonu Eşitlik (2.7)'deki gibi olur [38]:

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = \ln L(\boldsymbol{\beta},\boldsymbol{\theta}) = \prod_i L_i(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2}\sum_i \ln(\det(\mathbf{V_i})) - \frac{1}{2}\sum_i (\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})^T \mathbf{V}_i^{-1} (\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta}) . \quad (2.7)$$

DKEM'de EÇO tahmini için iteratif bir algoritma olan EM algoritması kullanılır [39]. EM algoritması "Beklenti" (Expectation) ve "Maksimizasyon" (Maximization) olmak üzere iki adımdan oluşur. E adımında algoritma, mevcut parametre tahminlerini göz önünde bulundurarak tüm verilerin logaritmik olasılıklarının koşullu beklenen değerini hesaplar. M adımı, güncellenmiş parametre tahminlerini elde etmek için E adımında türetilen koşullu beklentinin maksimize edilmesine dayanır. Algoritma bilinmeyen parametreler için başlangıç değerleri ile başlar. E ve M adımlarındaki iterasyonlar yakınsama sağlanana kadar devam eder. Sonuç olarak, EM algoritması iterasyonlar sona erdiğinde, bilinmeyen parametreler için elde edilen değerlerin olabilirlik fonksiyonunu maksimize ettiği varsayılır [40].

r, $r = 0,1,....$ iterasyon sayısını ve $\boldsymbol{\eta} = (\boldsymbol{\beta},\boldsymbol{\theta})$, DKEM'deki tüm bilinmeyen parametreleri içeren bir matris olarak tanımlansın. EM algoritmasının E adımının $r.$ iterasyonunda hesaplanan değeri Eşitlik (2.8)'de verilen fonksiyon ile hesaplanır:

$$Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(r)}) = E(\log L(\boldsymbol{\eta} \mid \mathbf{y}, \mathbf{b}) \mid \mathbf{y}, \boldsymbol{\eta}^{(r)}) = E\left(\sum_{i=1}^{n} \left[\log(f(\mathbf{y}_i \mid \mathbf{b}_i, \boldsymbol{\beta}, \mathbf{R}_i) + \log f(\mathbf{b}_i \mid \mathbf{D}) \mid \mathbf{y}_i, \boldsymbol{\eta}^{(r)} \right] \right). \quad (2.8)$$

Eşitlik (2.8)'de koşullu beklenen değer $f(\mathbf{b}_i \mid \mathbf{y}_i, \boldsymbol{\eta}^{(r)})$ koşullu dağılım fonksiyonundan elde edilir. E adımında DKEM'deki $\mathbf{D}$ ve $\mathbf{R}_i$ varyans-kovaryans matrislerindeki parametreler için yeterli istatistikler sırasıyla Eşitlik (2.9) ve Eşitlik (2.10) ile hesaplanır:

$$\sum_{i=1}^{n} E(\boldsymbol{\varepsilon}_i^T \boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\eta}}^{(r)}) = \sum_{i=1}^{n} \left[\hat{\boldsymbol{\varepsilon}}_i^{(r)T} \hat{\boldsymbol{\varepsilon}}_i^{(r)} + iz(Cov(\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\eta}}^{(r)})) \right], \quad (2.9)$$

$$\sum_{i=1}^{n} E(\mathbf{b}_i \mathbf{b}_i^T \mid \mathbf{y}_i, \hat{\boldsymbol{\eta}}^{(r)}) = \sum_{i=1}^{n} \left[\hat{\mathbf{b}}_i^{(r)} \hat{\mathbf{b}}_i^{(r)T} + iz(Cov(\mathbf{b}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\eta}}^{(r)})) \right]. \quad (2.10)$$

Eşitlik (2.9) ve Eşitlik (2.10)'da,

$$\hat{\boldsymbol{\varepsilon}}_i^{(r)} = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)} - \mathbf{Z}_i\hat{\mathbf{b}}_i^{(r)},$$

$$\hat{\mathbf{b}}_i^{(r)} = \mathbf{D}(\hat{\boldsymbol{\eta}}^{(r)})\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\hat{\boldsymbol{\eta}}^{(r)})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)}) \text{ ve}$$

$$\mathbf{V}(\hat{\boldsymbol{\eta}}^{(r)}) = \mathbf{Z}_i\mathbf{D}(\hat{\boldsymbol{\eta}}^{(r)})\mathbf{Z}_i^T + \mathbf{R}_i(\hat{\boldsymbol{\eta}}^{(r)})$$

olarak ifade edilir.

Eşitlik (2.11)'de $\boldsymbol{\beta}$, Eşitlik (2.12)'de σ ve Eşitlik (2.13)'te $\mathbf{D}$ matrisi tahminleri verilmektedir:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left[\sum_{i=1}^{n}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}(\boldsymbol{\eta}^{(r)})\mathbf{X}_i\right]^{-1}\sum_{i=1}^{n}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{-1}(\boldsymbol{\eta}^{(r)})\mathbf{y}_i\,, \tag{2.11}$$

$$\hat{\sigma}^{(r+1)2} = \frac{\sum_{i=1}^{n}E(\boldsymbol{\varepsilon}_i^T\boldsymbol{\varepsilon}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\eta}}^{(r)})}{\sum_{i=1}^{m}n_i}\,, \tag{2.12}$$

$$\hat{\mathbf{D}}^{(r+1)} = \frac{\sum_{i=1}^{n}E(\mathbf{b}_i\mathbf{b}_i^T \mid \mathbf{y}_i\hat{\boldsymbol{\eta}}^{(r)})}{n},\ r = 0,1,.... \tag{2.13}$$

Yukarıda belirtilen matematiksel iterasyonlar yakınsama sağlanana kadar devam eder ve yakınsama gerçekleştiğinde $\boldsymbol{\eta}$'nın EÇO'su elde edilmiş olur [40]. Veri kümesi elde edildikten sonra, Eşitlik (2.14)'te belirtilen ampirik Bayes tahminleri rastgele etkilerin tahminlerini elde etmek için kullanılır:

$$\hat{\mathbf{b}}_i = E(\boldsymbol{\beta}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\eta}}) = \hat{\mathbf{D}}(\hat{\boldsymbol{\eta}})\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\hat{\boldsymbol{\eta}})(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})\,. \tag{2.14}$$

Eşitlik (2.14)'te verilen formülle elde edilen rastgele etkilerin tahminlerini kullanarak konuya özgü çıkarımlar yapmak mümkündür. Örneğin, i. birimin cevap tahmini Eşitlik (2.15) ile elde edilir [40]:

$$\hat{\mathbf{y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \mathbf{Z}_i\hat{\mathbf{b}}_\mathbf{i}\,. \tag{2.15}$$

2.2. Genelleştirilmiş Doğrusal Karma Etki Modelleri (GDKEM)

DKEM'de sürekli cevap değişkeninin normal dağılıma sahip olduğu varsayılmaktadır. Ancak, uzun süreli verilerde bu varsayım her zaman sağlanamaz. Cevap değişkeninin kesikli veya normal dağılım göstermeyen sürekli olduğu durumlarda DKEM cevap değişkenini açıklamada yeterli olmamaktadır. GDKEM, DKEM'nin geliştirilmiş hali olarak ortaya çıkmıştır [10]. GDKEM, genel olarak Eşitlik (2.16)'daki gibi verilir:

$$h^{-1} = \left\{E(\mathbf{Y}_i \mid \mathbf{b}_i, \mathbf{X}_i, \mathbf{Z}_i)\right\} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i\,. \tag{2.16}$$

Eşitlik (2.16)'da, h^{-1}, vektörel bir bağ fonksiyonunu göstermektedir. Eşitlik (2.16)'da yer alan değişkenler ise DKEM'dekiyle aynıdır. DKEM, rastgele etkilerin çok değişkenli normal dağılımdan geldiğini varsaymaktadır. Cevap değişkeni için koşullu bağımsızlık varsayımı söz konusudur. Cevap değişkeninin Eşitlik (2.17)'deki üstel dağılım ailesinden olduğu varsayılmaktadır:

$$f(\mathbf{y}_{ij} \mid \mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \exp\left[\phi^{-1}\left\{\mathbf{y}_{ij}\boldsymbol{\theta}_{ij} - \psi(\boldsymbol{\theta}_{ij})\right\} + c(\mathbf{y}_{ij}, \phi)\right]. \tag{2.17}$$

Eşitlik (2.17)'de, ϕ dağılım parametresini ve $\boldsymbol{\theta}_{ij}$, Eşitlik (2.18) ile verilen $\boldsymbol{\eta}_{ij}$'nin fonksiyonun kanonik ya da doğal parametresini ifade eder. $\psi(.)$ ve $c(.)$ ise bilinen birer fonksiyondur.

$$\boldsymbol{\eta}_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i\,. \tag{2.18}$$

Buna göre, koşullu ortalama Eşitlik (2.19)'da, $\psi'(\boldsymbol{\theta}_{ij})$'nin birinci türevi ve koşullu varyans Eşitlik (2.20) ile verilmiştir:

$$\boldsymbol{\mu}_{ij} = h(\boldsymbol{\eta}_{ij}) = \psi'(\boldsymbol{\theta}_{ij}) = \frac{\partial\psi(\boldsymbol{\theta}_{ij})}{\partial(\boldsymbol{\theta}_{ij})}\,, \tag{2.19}$$

$$Var(\mathbf{Y}_{ij} \mid \boldsymbol{\eta}_{ij}) = \phi\psi''(\boldsymbol{\theta}_{ij}) = \phi V(\boldsymbol{\mu}_{ij}) \,. \tag{2.20}$$

Eşitlik (2.20)'de $V(\boldsymbol{\mu}_{ij})$ varyans fonksiyonunu ifade etmektedir [41].

GDKEM, cevap değişkeninin yapısına göre değişmektedir. Cevap değişkeninin sürekli ve normal dağılımlı olduğu düşünüldüğünde, GDKEM, DKEM ile aynı yapıya sahiptir. Bu modellerde açıklayıcı değişkenler ile cevap değişkeni arasında logaritmik bağ fonksiyonu kullanılır [42].

Sabit ve rastgele etkilerin ayrı olarak tahmin edildiği GDKEM'de EÇO tahmin yöntemi en sık kullanılan tahmin yöntemidir.

2.2.1. Sabit Etkilerin En Çok Olabilirlik Tahmini

GDKEM'de marjinal logaritmik olabilirlik fonksiyonu Eşitlik (2.21) ile verilmiştir:

$$l(\boldsymbol{\beta}, \mathbf{G}) = \ln \prod_{i=1}^{N} \int \varphi(\mathbf{b}_i; \mathbf{0}, \mathbf{G}) \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{b}_i) db_i \,. \tag{2.21}$$

Eşitlik (2.21)'de, $\varphi(\mathbf{b}_i, \mathbf{0}, \mathbf{G})$, $\mathbf{0}$ ortalama ve $\mathbf{G}$ kovaryans matrisi ile çok değişkenli normal dağılımın yoğunluk fonksiyonudur. Değişkenler için standart normal dağılımlı rastgele etkiler $\mathbf{v}_i$ kullanıldığında integral işleminin kolay hesaplanması söz konusu olacaktır. Bunun için $\mathbf{G}$ kovaryans matrisinin Cholesky ayrışımı (Q) kullanılır ve $\mathbf{b}_i = Q\mathbf{v}_i$ olarak ifade edilir. Logaritmik olasılık fonksiyonu Eşitlik (2.22) ile yeniden düzenlenir:

$$l(\boldsymbol{\beta}, \mathbf{G}) = \ln \prod_{i=1}^{N} \int_{-\infty}^{+\infty} \phi(\mathbf{v}_{iq}) \ldots \left[\int_{-\infty}^{+\infty} \phi(\mathbf{v}_{i1}) \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{v}_i) d\mathbf{v}_{i1} \right] \ldots d\mathbf{v}_{iq} \,. \tag{2.22}$$

Eşitlik (2.22)'de $\phi(.)$ tek değişkenli standart normal yoğunluk fonksiyonudur ve Eşitlik (2.22)'deki integral işlemi Eşitlik (2.23)'teki gibidir [41]:

$$\int_{-\infty}^{+\infty} \phi(\mathbf{v}_{i1}) \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{v}_i) dv_i \approx \sum_{r=1}^{R} p_r \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid (a_r, v_{i2}, ..., v_{iq})) . \quad (2.23)$$

Eşitlik (2.23)'te $\sqrt{\pi}\, p_r$ ve $a_r / \sqrt{2}$ ifadeleri ağırlıkları ve $(2R-2)$ dereceden Gauss-Hermite tümlev kuralının konumlarını ifade eder. Gauss-Hermite tümlevi düşük dereceli polinomlar için iyi çalışmaktadır. Korelasyon yüksek, cevap değişkeni Normal ya da Poisson dağılımından ya da n_i büyük olduğu durumlarda, yüksek tümlev noktalarına ihtiyaç duyulmaktadır. Bu problemin ortadan kaldırılması için uyarlanabilir tümlev (adaptive quadrature) kullanılmaktadır [41]. Eşitlik (2.23)'e uyarlanabilir tümlev kuralı uygulanması sonucunda Eşitlik (2.24)'teki integral işlemi elde edilir:

$$\int_{-\infty}^{+\infty} \phi(\mathbf{v}_i) \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{v}_i) dv_i = \int_{-\infty}^{+\infty} \varphi(\mathbf{v}_i; \boldsymbol{\mu}_i, \tau_i^2) \left[\frac{\phi(\mathbf{v}_i) \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{v}_i)}{\varphi(\mathbf{v}_i; \boldsymbol{\mu}_i, \tau_i^2)} \right] dv_i . \quad (2.24)$$

Eşitlik (2.24)'te, $\varphi(v_i; \mu_i, \tau_i^2)$, μ_i ortalama ve τ_i^2 varyansı ile normal olasılık fonksiyonunu ifade etmektedir. Yaklaştırma işlemi Eşitlik (2.25)'teki gibi gösterilmektedir:

$$\int_{-\infty}^{+\infty} \phi(\mathbf{v}_i) \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{v}_i) dv_i \approx \sum_{i=1}^{R} p_{ir} \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid a_{ir}) \quad . \quad (2.25)$$

Eşitlik (2.25)'te, $a_{ir} \equiv \tau_i a_r + \mu_i$ ifadesi, birim-özel konumu ve $p_{ir} \equiv \sqrt{2\pi}\tau_i \exp(a_r^2 / 2) \phi(\tau_i a_r + \mu_i)\, p_r$ ifadesi ise ağırlıkları ifade etmektedir.

Uyarlanmış tümlevin kullanıldığı GDKEM için EÇO tahmin yöntemi kullanılabilmektedir.

2.2.2. Rastgele Etkilerin Tahmini

Rastgele etkiler için EÇO tahmini, cevap değişkeninin birleşik dağılımının Eşitlik (2.26)'daki gibi maksimize edilmesiyle elde edilir:

$$\prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \hat{\boldsymbol{\beta}}, \hat{\mathbf{G}}) . \tag{2.26}$$

DKEM için EÇO tahmini Eşitlik (2.27) ile verilir:

$$\hat{\mathbf{b}}_i = (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) . \tag{2.27}$$

Eşitlik (2.27) en küçük kareler (EKK) tahminidir. EÇO tahmini ise Eşitlik (2.28) ile verilmektedir:

$$\hat{\mathbf{b}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}}) . \tag{2.28}$$

Rastgele etkiler ampirik Bayes yöntemiyle de tahmin edilmektedir. $\boldsymbol{\beta}$, ϕ ve $\mathbf{G}$ parametrelerinin EÇO tahminleri verildiğinde, Eşitlik (2.29)'daki gibi rastgele etkiler tahmin edilir:

$$\hat{\mathbf{b}}_i = E(\mathbf{b}_i \mid \mathbf{Y}_i, \hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\mathbf{G}}) . \tag{2.29}$$

$\boldsymbol{b_i}$'nin ampirik Bayes tahmini, $\mathbf{Y}_i$, $\hat{\boldsymbol{\beta}}$ ve $\hat{\mathbf{G}}$ verildiğinde i. birim ($i = 1, \ldots, m$) için tahmin edilen rastgele etkilerdir. Bayes tahmini rastgele etkilerin en iyi doğrusal yansız tahminidir [42].

2.3. Uzun Süreli Verilerde Varyans-Kovaryans Yapıları

Dengeli uzun süreli verilerde, $\mathbf{Y}_i$ cevap vektörü ve $\mathbf{R}_i$ varyans-kovaryans matrisi tüm i'ler için aynı boyuttadır. Tüm i'ler için aynı varyans-kovaryans matrisi $\mathbf{R}_i$ için yapısal olmayan, AR(1) ve Toeplitz kovaryans yapıları örnek verilebilir [43], [4], [44].

2.3.1. Yapısal Olmayan Varyans-Kovaryans Yapısı

Varyansların ve kovaryansların biribirine eşit olmadığı yapısal olmayan (unstructured) varyans-kovaryans matrisinde, her bir zaman noktası için bir tane olmak üzere n tane farklı varyans ve her bir zaman çifti için farklı kovaryanslar söz konusudur. Toplam $n+n(n-1)/2 = n(n+1)/2$ adet varyans-kovaryans parametresi söz konusudur:

$$\boldsymbol{R_i} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \vdots & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \vdots & \sigma_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \vdots & \sigma_{nn} \end{bmatrix}$$

2.3.2. AR(1) Varyans-Kovaryans Yapısı

Gözlem zamanlarının eşit aralıklı olduğu uzun süreli veri kümesinde AR(1) yapısı sık karşılaşılan bir varyans-kovaryans yapısıdır. AR(1) korelasyon yapısında zaman boyunca ölçümler arasındaki korelasyonun (ρ ($-1 < \rho < 1$)) azaldığı ifade edilir. n tane zaman noktası için korelasyon matrisi aşağıdaki gibi verilir:

$$\Gamma = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \vdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & \rho & 1 \end{bmatrix}$$

Uzun süreli verilerde, zaman noktaları arasındaki süre arttıkça korelasyonların azalması beklendiğinden, AR(1) en iyi uyan varyans-kovaryans yapısıdır [45].

Heterojen AR(1) varyans-kovaryans matrisi, her bir zaman noktasında varyansların farklı olabildiği durum için tercih edilir:

$$\boldsymbol{R_i} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \rho^3\sigma_1\sigma_4 & \rho^4\sigma_1\sigma_5 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho^2\sigma_2\sigma_4 & \rho^3\sigma_2\sigma_5 \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 & \rho^2\sigma_3\sigma_5 \\ \rho^3\sigma_1\sigma_4 & \rho^2\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 & \rho\sigma_4\sigma_5 \\ \rho^4\sigma_1\sigma_5 & \rho^3\sigma_2\sigma_5 & \rho^2\sigma_3\sigma_5 & \rho\sigma_4\sigma_5 & \sigma_5^2 \end{bmatrix}$$

2.3.3. Toeplitz Varyans-Kovaryans Yapısı

Bu varyans-kovaryans yapısında sadece art arda gelen ölçümlerin ilişkili olduğu belirtilmektedir. Art arda gelen herhangi iki ölçüm arasındaki korelasyon $-1 < \rho < 1$ ve her zaman noktasında varyans aynıdır ($\sigma_j^2 = \sigma^2$). Art arda gelen cevaplar için bağımlılık ihmal edilmediğinden, Toeplitz yapısı, bir-bağımlı (one dependent) korelasyon yapısı veya sınırlandırılmış (banded) Toeplitz matrisi olarak da adlandırılmaktadır. $n = 5$ için bir-bağımlı korelasyon matrisi aşağıdaki gibi verilir:

$$\Gamma = \begin{bmatrix} 1 & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & \rho & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{bmatrix}$$

Toeplitz yapısı, eşit aralıklı zaman noktaları için anlamlıdır. Her bir zaman noktası için varyansların aynı ($\sigma_j^2 = \sigma^2$) olduğu durumda, bir-bağımlı korelasyon matrisine karşılık gelen varyans-kovaryans matrisi aşağıda verilmiştir:

$$\boldsymbol{R_i} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & 0 & 0 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & 0 & 0 \\ 0 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & 0 \\ 0 & 0 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ 0 & 0 & 0 & \rho\sigma^2 & \sigma^2 \end{bmatrix} = \sigma^2\Gamma$$

3. UZUN SÜRELİ VERİ ANALİZİ İÇİN HİBRİT MAKİNE ÖĞRENMESİ YÖNTEMLERİ

3.1. Karma Etkili Regresyon Ağacı (MERT)

Hajjem ve ark. [27] tarafından geliştirilen MERT algoritmasının temel mantığı sabit etkileri rastgele etkilerden ayırmaktır. Bu algoritmada sabit etkileri modellemek için standart bir regresyon ağacı (CART) [15] ve rastgele etkileri modellemek için ağacın her bir terminal düğümünde düğümle değişmeyen doğrusal bir yapı kullanılmaktadır. Yöntem, beklenti maksimizasyonu (EM) [39], [46] algoritması çerçevesinde standart bir ağaç algoritması kullanılarak uygulanmaktadır. Bir diğer ifadeyle, DKEM'deki [47], [9] sabit etki bileşeninin doğrusal tahmini, standart bir regresyon ağacı algoritması ile geliştirilmiştir.

MERT için model denklemi Eşitlik (3.1) ile verilmiştir:

$$\mathbf{y}_i = f(\mathbf{X}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_\mathbf{i}.\ i = 1,...,m. \tag{3.1}$$

Eşitlik (3.1)'de, $\mathbf{y}_i = [y_{i1},....,y_{in_i}]^T$ ifadesi $i.$ birimin $n_i \times 1$ boyutlu cevap değişkeni vektörünü, $f(\mathbf{X}_i)$ standart ağaç algoritması ile tahmin edilen sabit etki fonksiyonunu, $\mathbf{X}_i = [x_{i1},....,x_{in_i}]^T$ ifadesi $n_i \times p$ boyutlu sabit etkili açıklayıcı değişken matrisini, $\mathbf{Z}_i = [z_{i1},....,z_{in_i}]^T$ ifadesi $n_i \times q$ boyutlu rastgele etkili açıklayıcı değişken matrisini, $\boldsymbol{\varepsilon}_i = [\varepsilon_{i1},....,\varepsilon_{in_i}]^T$, $n_i \times 1$ boyutlu hata vektörünü ve $\mathbf{b}_i = (b_{i1},....,b_{iq})^T$ $i.$ birimin $q \times 1$ boyutlu bilinmeyen rastgele etkiler vektörünü belirtmektedir. Toplam gözlem sayısı ise $N = \sum_{i=1}^{m} n_i$ olarak ifade edilmektedir. $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$ ve $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i)$ olduğu varsayılmaktadır.

MERT algoritmasının adımları, r $(r = 0,1,2,...)$ iterasyon sayısı olmak üzere aşağıda verilmiştir:

0. Adım: $r=0$, $\hat{b}_{i(0)}=0$, $\hat{\sigma}^2_{(0)}=1$, $\hat{\mathbf{D}}_{(0)}=\mathbf{I}_q$ olsun.

1. Adım: $r=r+1$, $\mathbf{y}^*_{i(r)}$, $\hat{f}(\mathbf{X}_i)_{(r)}$, $\hat{\mathbf{b}}_{i(r)}$ olacak şekilde güncellenir:

i) $\hat{\mathbf{y}}_{i(r)}=\mathbf{y}_i-\mathbf{Z}_i\hat{\mathbf{b}}_{i(r-1)}$, $i=1,...,m$ (3.2)

ii) $\hat{f}(\mathbf{X}_i)_{(r)}$, $f(\mathbf{X}_i)$'nin standart ağaç algoritmasından elde edilen tahmini olsun. Bu adımda, $\mathbf{y}^*_{i(r)}$ cevap değişkenini, $\mathbf{X}_i$ ise açıklayıcı değişkenleri göstermektedir.

iii) $\hat{\mathbf{b}}_{i(r)}=\hat{\mathbf{D}}_{(r-1)}\mathbf{Z}_i^T\hat{\mathbf{V}}^{-1}_{i(r-1)}(\mathbf{y}_i-\hat{f}(\mathbf{X}_i)_{(r)})$, $i=1,....,m$ (3.3)

Eşitlik (3.3)'te,

$$\hat{\mathbf{V}}^{-1}_{i(r-1)}=\mathbf{Z}_i\hat{\mathbf{D}}_{(r-1)}\mathbf{Z}_i^T+\hat{\sigma}^2_{(r-1)}\mathbf{I}_{n_i},\ i=1,....,m$$

olarak ifade edilir.

2. Adım: $\hat{\sigma}^2_{(r)}$ ve $\hat{\mathbf{D}}_{(r)}$ terimleri sırasıyla Eşitlik (3.4), Eşitlik (3.5) ve Eşitlik (3.6)'daki gibi güncellenir:

$$\hat{\sigma}^2_{(r)}=N^{-1}\sum_{i=1}^{m}\left\{\hat{\boldsymbol{\varepsilon}}^T_{i(r)}\boldsymbol{\varepsilon}_{i(r)}+\hat{\sigma}^2_{(r-1)}\left[n_i-\hat{\sigma}^2_{(r-1)}iz(\hat{\mathbf{V}}_{i(r-1)})\right]\right\}, \quad (3.4)$$

$$\hat{\mathbf{D}}_{(r)}=m^{-1}\sum_{i=1}^{m}\left\{\hat{\mathbf{b}}_{i(r)}\hat{\mathbf{b}}^T_{i(r)}+\left[\hat{\mathbf{D}}_{(r-1)}-\hat{\mathbf{D}}_{(r-1)}\mathbf{Z}_i^T\mathbf{V}^{-1}{}_{i(r-1)}\mathbf{Z}_i\hat{\mathbf{D}}_{(r-1)}\right]\right\}, \quad (3.5)$$

$$\hat{\boldsymbol{\varepsilon}}_{i(r)}=\mathbf{y}_i-\hat{f}(\mathbf{X}_i)_{(r)}-\mathbf{Z}_i\hat{\mathbf{b}}_{i(r)}. \quad (3.6)$$

3.Adım: Yakınsama gerçekleşene kadar 1.Adım ve 2. Adım tekrarlanır.

MERT algoritması 0. Adımda $\hat{\mathbf{b}}_i$, $\hat{\sigma}^2$ ve $\hat{\mathbf{D}}$ için varsayılan başlangıç değerleriyle başlar. 1. Adımda, ilk olarak $\mathbf{y}_\mathbf{i}^*$ cevap değişkeninin sabit etki kısmı rastgele etki kısmının etkisi katılmadan hesaplanır. Ardından, sabit etkili bileşen olan $\hat{f}(\mathbf{X}_i)$ standart ağaç algoritması kullanılarak $\mathbf{y}_\mathbf{i}^*$ cevapları, $\mathbf{X}_i$ açıklayıcı değişkenler ile tahmin edilir. Daha sonra, $\hat{\mathbf{b}}_i$ parametresi güncellenir. 2.Adımda, bir önceki adımda tahmin edilen sabit etki bileşeni $\hat{f}(\mathbf{X}_i)$ ham veriden $(\mathbf{y}_i)$ çıkarıldıktan sonra $\hat{\sigma}^2$ ve $\hat{\mathbf{D}}$ varyans bileşenleri hatalar göz önünde bulundurularak güncellenir. Yakınsama gerçekleşene kadar iterasyonlar 1. ve 2.Adımda devam eder.

MERT algoritmasının yakınsaması, her iterasyonda Eşitlik (3.7)'deki genelleştirilmiş log-olabilirlik (GLL) fonksiyonu ile hesaplanarak takip edilir:

$$GLL(f,\mathbf{b}_i \mid \mathbf{y}) = \sum_{i=1}^{m} \left\{ \left[\mathbf{y}_i - f(\mathbf{X}_i) - \mathbf{Z}_i\mathbf{b}_i\right]^T \mathbf{R}_i^{-1} \left[\mathbf{y}_i - f(\mathbf{X}_i) - \mathbf{Z}_i\mathbf{b}_i\right] + \mathbf{b}_i^T\mathbf{D}^{-1}\mathbf{b}_i + \log|\mathbf{D}| + \log|\mathbf{R}_i| \right\}. \quad (3.7)$$

3.2. Karma Etkili Rastgele Orman (MERF)

Hajjem ve ark. [21] geliştirdikleri MERT algoritmasının her iterasyonundaki standart ağaç regresyon algoritması yerine rastgele orman yaklaşımını [17], [16] kullanarak Karma Etkili Rastgele Orman (MERF) algoritmasını geliştirmişlerdir.

MERF modeli Eşitlik (3.8)'deki gibi tanımlansın:

$$\mathbf{y}_i = f(\mathbf{X}_i) + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1,...,m. \quad (3.8)$$

Eşitlik (3.8)'de, $\mathbf{y}_i = [y_{i1},....,y_{in_i}]^T$ ifadesi i. birimin $n_i \times 1$ boyutlu cevap değişkeni vektörünü, $f(\mathbf{X}_i)$ standart regresyon ağaçları ormanı algoritması ile tahmin edilen fonksiyonu, $\mathbf{X}_i = [x_{i1},....,x_{in_i}]^T$ ifadesi $n_i \times p$ boyutlu sabit etkili açıklayıcı değişken matrisini, $\mathbf{Z}_i = [z_{i1},....,z_{in_i}]^T$ ifadesi $n_i \times q$ boyutlu rastgele etkili açıklayıcı değişken matrisini, $\boldsymbol{\varepsilon}_i = [\varepsilon_{i1},....,\varepsilon_{in_i}]^T$ ifadesi $n_i \times 1$ boyutlu hata vektörünü ve $\mathbf{b}_i = (b_{i1},....,b_{iq})^T$ ifadesi i. birime ilişkin $q \times 1$ boyutlu rastgele etki katsayıları vektörünü belirtmektedir. Toplam

gözlem sayısı $N = \sum_{i=1}^{m} n_i$ olarak ifade edilmektedir. $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$ ve $\boldsymbol{\varepsilon}_i \sim N_{n_i}(\mathbf{0}, \mathbf{R}_i)$ olduğu varsayılmaktadır. Modelin rastgele etkili kısmı olan $\mathbf{Z}_i\mathbf{b}_i$'nin doğrusal olduğu varsayılmaktadır. Ayrıca $\mathbf{b}_i$ ve $\boldsymbol{\varepsilon}_i$ terimlerinin birbirinden bağımsız olacak şekilde normal dağılım gösterdiği ve birimler arası ölçümlerin de bağımsız olduğu varsayılmaktadır. i. birimin $\mathbf{y}_i$ cevap değişkeni vektörünün kovaryans matrisi Eşitlik (3.9)'da verilmiştir:

$$\mathbf{V}_i = Cov(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i . \tag{3.9}$$

Eşitlik (3.9)'da, $\mathbf{V} = Cov(\mathbf{y}) = diag(\mathbf{V}_1, ..., \mathbf{V}_m)$ ve $\mathbf{y} = [\mathbf{y_1}^{\mathbf{T}}, ..., \mathbf{y}_m^{\mathbf{T}}]^T$ olarak ifade edilmektedir.

MERF algoritması temel olarak MERT algoritmasındaki modelin sabit etkilerinin standart rastgele orman algoritması kullanılarak tahmin edildiği bir algoritmadır. MERF algoritmasının adımları aşağıda verilmiştir:

0. Adım: $r = 0$, $\hat{\mathbf{b}}_{i(0)} = 0$, $\hat{\sigma}^2_{(0)} = 1$, $\hat{\mathbf{D}}_{(0)} = \mathbf{I}_q$ olsun.

1. Adım: $r = r + 1$, $\mathbf{y}^*_{i(r)}$, $\hat{f}(\mathbf{X}_i)_{(r)}$, $\hat{\mathbf{b}}_{i(r)}$ olarak güncellenir.

i) $\hat{\mathbf{y}}_{i(r)} = \mathbf{y}_i - \mathbf{Z}_i\hat{\mathbf{b}}_{i(r-1)}$, $i = 1, ..., m$ (3.10)

ii) $\hat{f}(\mathbf{X}_i)_{(r)}$, $\mathbf{y}^*_{i(r)}$'nin cevap değişkeni olduğu durumda $f(\mathbf{X}_i)$'nin standart rastgele orman algoritmasından elde edilen torba dışı (out of bag) tahmini olsun. $\mathbf{X}_i$ açıklayıcı değişkenleri göstermekte olup, $(i = 1, ..., m)$, $(\mathbf{y}^*_{i(r)}, \mathbf{X}_i)$ ikilileri bootstrap yöntemi ile istenen sayıda çekilen örneklemlerle modele girdi sağlanır.

iii) $\hat{\mathbf{b}}_{i(r)} = \hat{\mathbf{D}}_{(r-1)}\mathbf{Z}_i^T\hat{\mathbf{V}}^{-1}_{i(r-1)}(\mathbf{y}_i - \hat{f}(\mathbf{X}_i)_{(r)})$, $i = 1, ..., m$ (3.11)

Eşitlik (3.11)'de,

$\hat{\mathbf{V}}_{i(r-1)}^{-1} = \mathbf{Z}_i \hat{\mathbf{D}}_{(r-1)} \mathbf{Z}_i^T + \hat{\sigma}_{(r-1)}^2 \mathbf{I}_{n_i}$, $i = 1, ..., m$ olarak ifade edilmektedir.

2. Adım: $\hat{\sigma}_{(r)}^2$ ve $\hat{\mathbf{D}}_{(r)}$ Eşitlik (3.12) ve Eşitlik (3.13)'teki gibi güncellenir:

$$\hat{\sigma}_{(r)}^2 = N^{-1} \sum_{i=1}^{m} \left\{ \hat{\boldsymbol{\varepsilon}}_{i(r)}^T \boldsymbol{\varepsilon}_{i(r)} + \hat{\sigma}_{(r-1)}^2 \left[n_i - \hat{\sigma}_{(r-1)}^2 iz(\hat{\mathbf{V}}_{i(r-1)}) \right] \right\}, \tag{3.12}$$

$$\hat{\mathbf{D}}_{(r)} = m^{-1} \sum_{i=1}^{m} \left\{ \hat{\mathbf{b}}_{i(r)} \hat{\mathbf{b}}_{i(r)}^T + \left[\hat{\mathbf{D}}_{(r-1)} - \hat{\mathbf{D}}_{(r-1)} \mathbf{Z}_i^T \mathbf{V}^{-1}{}_{i(r-1)} \mathbf{Z}_i \hat{\mathbf{D}}_{(r-1)} \right] \right\}. \tag{3.13}$$

Eşitlik (3.12)'de,

$\hat{\boldsymbol{\varepsilon}}_{i(r)} = \mathbf{y}_i - \hat{f}(\mathbf{X}_i)_{(r)} - \mathbf{Z}_i \hat{\mathbf{b}}_{i(r)}$ olarak ifade edilir.

3. Adım: Yakınsama gerçekleşene kadar 1.Adım ve 2. Adım tekrarlanır.

MERF algoritmasının çalışma adımları şu şekildedir: Algoritma ilk olarak başlangıç adımında $\hat{\mathbf{b}}_i$, $\hat{\sigma}^2$ ve $\hat{\mathbf{D}}$ için varsayılan başlangıç değerleriyle başlar. 1. Adımda ilk olarak cevap değişkeninin sabit etkili kısmı rastgele kısmın etkisi katılmadan hesaplanır. Daha sonra bootstrap ile çekilen $(\mathbf{y}_{i(r)}^*, \mathbf{X}_i)$ örneklemleri ile rastgele ormanlar oluşturulur. Aşırı öğrenme (overfitting) riskini azaltmak için i. birimdeki j. ölçümün sabit etkisinin tahmini bu gözlemin yer almadığı bootstrap örnekleri ile oluşturulan ormanlarda yer alan ağaç topluluklarından elde edilir. Bu yöntem out-of-bag (torba dışı, OOB) tahmin yöntemi olarak adlandırılır. Ardından $\hat{\mathbf{b}}_i$ güncellenir. 2.Adımda, tahmin edilen sabit etki bileşeni $\hat{f}(\mathbf{X}_i)$ ham veriden $(\mathbf{y}_i)$ çıkarıldıktan sonra $\hat{\sigma}^2$ ve $\hat{\mathbf{D}}$ varyans bileşenleri hatalar göz önünde bulundurularak güncellenir. 1. ve 2.Adımdaki iterasyonlar yakınsama gerçekleşene kadar devam eder.

Algoritmanın yakınsaması, her iterasyonda Eşitlik (3.14)'te verilen genelleştirilmiş log-olabilirlik (GLL) fonksiyonu ile hesaplanarak takip edilir:

$$GLL(f, \mathbf{b}_i \mid \mathbf{y}) = \sum_{i=1}^{m} \left\{ \left[\mathbf{y}_i - f(\mathbf{X}_i) - \mathbf{Z}_i \mathbf{b}_i \right]^T \mathbf{R}_i^{-1} \left[\mathbf{y}_i - f(\mathbf{X}_i) - \mathbf{Z}_i \mathbf{b}_i \right] + \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i + \log |\mathbf{D}| + \log |\mathbf{R}_i| \right\}. \quad (3.14)$$

3.3. Rastgele Beklenti Maksimizasyonu Ağacı (RE-EM Ağacı)

RE-EM Ağacı algoritması Sela ve Simonoff [25] tarafından geliştirilmiştir. En genel DKEM formu, sabit etki terimleri ve cevaplar arasındaki ilişki için fonksiyonel bir bağlantı kurar. Hata terimi ve rastgele etkiler üzerindeki dağılım varsayımlarıyla birlikte temel model Eşitlik (3.15) ile ifade edilir:

$$\mathbf{y}_{it} = \mathbf{Z}_{it} \mathbf{b}_{\mathbf{i}} + f(x_{it1}, ..., x_{itK}) + \boldsymbol{\varepsilon}_{it}. \quad (3.15)$$

Eşitlik (3.15)'te, i, uzun süreli veri kümesindeki birimleri ($i = 1, 2, ..., m$) belirtirken, t, ($t = 1, 2, ..., T_i$) zaman boyunca alınan ölçümleri temsil eder. f fonksiyonu, sabit etkiler ve sayısal cevap değişkeni arasındaki ilişkinin bir fonksiyonudur. $\mathbf{Z}_{it}$, rastgele etki değişkenlerine karşılık gelen tasarım matrisini ve $\mathbf{b}_i$, birime özgü etkileri temsil eden bilinmeyen rastgele etki katsayı vektörünü belirtmekte olup $\boldsymbol{\varepsilon}_{it} \sim N(\mathbf{0}, \mathbf{R}_i)$ ve $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ olduğu varsayılmaktadır.

RE-EM Ağacı modeli için EM algoritması ile EÇO yöntemi kullanılabilmektedir. Ancak, analiz edilecek uzun süreli veri kümesinin yapısına ve özelliklerine göre, bu geleneksel yöntemleri kullanmak bazı sorunlara yol açabilmektedir. Örneğin, DKEM, f fonksiyonu için parametrik bir form varsayar ($f = \mathbf{X}\boldsymbol{\beta}$). Ancak, f genellikle bilinmediği için bu varsayım kısıtlayıcı olabilir ve verilere doğrusal bir model uydurmak en iyi seçim olmayabilir. Sela ve Simonoff [25] tarafından uzun süreli verilerin analizi için geliştirilen RE-EM Ağacı algoritmasında, EM algoritmasından esinlenilmiş ve rastgele etkiler ($\mathbf{b}_i$) modele dahil edilerek f değeri tahmin edilebilmektedir. Bu yöntemde, aynı birime karşılık gelen farklı gözlemsel birimler farklı düğümlere dahil edilir. RE-EM Ağacının avantajı,

geleneksel DKEM yaklaşımına kıyasla sabit etkiler için esnek bir yapıya sahip olmasıdır [25].

RE-EM Ağacı algoritmasının adımları aşağıda verilmiştir:

0. Adım: $\hat{\mathbf{b}}_i = 0$ olsun.

1.Adım: Rastgele etki tahminleri yakınsayana kadar aşağıda yer alan a ve b adımları tekrarlanır. Bu adımda yakınsama işlemi belirlenen tolerans değerinden daha küçük bir değere ulaşacak EÇO ya da KEÇO'daki değişime göre izlenir.

a) Cevap değişkeninin $\mathbf{y}_{it} - \mathbf{Z}_{it}\hat{\mathbf{b}}_i$ olduğu ve açıklayıcı değişkenlerin ($i = 1, 2, ..., m$) ve t ($t = 1, 2, ..., T_i$) için $\mathbf{x}_{it} = \left(x_{it1}, \ldots, x_{itK}\right)$ olarak ifade edildiği göz önüne alındığında, f fonksiyonuna yaklaşan bir regresyon ağacı tahmin edilir. Tahmin edilen regresyon ağacını kullanarak g_p'nin tüm terminal düğümleri arasında değiştiği bir gösterge değişkenleri kümesi $\mathbf{I}$ $(x_{it} \in g_p)$ elde edilir.

b) $\mathbf{y}_{it} = \mathbf{Z}_{it}\mathbf{b}_i + \mathbf{I}(\mathbf{X}_{it} \in g_p)\boldsymbol{\mu}_p + \boldsymbol{\varepsilon}_{it}$ karma etki modeli tahmin edilir ve $\hat{\mathbf{b}}_i$ elde edilir.

2.Adım: Tahmin edilen ağacın her bir terminal düğümündeki tahmin edilen cevap değeri Adım 1b'deki karma etkili modelden elde edilen tahmini ortalama $\boldsymbol{\mu}_p$ ile değiştirilir.

3.4. Yarı Parametrik Stokastik Karma Etki Modelleri

Capitaine [28], karma etkili uzun süreli veriler için Hajjem [27], [21] tarafından geliştirilen MERT ve MERF algoritması ile Sela ve Simenoff [25] tarafından geliştirilen RE-EM Ağacı algoritmalarına stokastik süreç parametresini ekleyerek mevcut yöntemler olan MERT, MERF ve RE-EM Ağacı algoritmalarını geliştirmiştir.

i. birimin n_i sayıda ölçüme sahip olduğu N sayıda gözlem biriminden oluşan t_{ij} zaman noktasındaki cevap ölçüm değeri y_{ij} ($i = 1, ..., m$ ve $j = 1, ..., n_i$) olduğu uzun süreli veri modeli Eşitlik (3.15)'te verilmiştir:

$$\mathbf{y}_{ij} = f(\mathbf{X}_{ij}) + \mathbf{Z}_{ij}\mathbf{b}_i + w_i(\mathbf{t}_{ij}) + \boldsymbol{\varepsilon}_{ij} . \quad (3.15)$$

Eşitlik (3.15)'te, $\mathbf{X}_{ij}$ $p \times 1$ boyutlu açıklayıcı değişken vektörünü, f $\mathbb{R}^p \to \mathbb{R}$ bilinmeyen ortalama fonksiyonunu, $\mathbf{b}_i$ $1 \times q$ boyutlu $\mathbf{Z}_{ij}$ rastgele etki açıklayıcı değişkenlerin $q \times 1$ boyutlu katsayı vektörünü, $w_i(\mathbf{t})$ seri korelasyonların modellenmesinde kullanılan stokastik süreç parametresini ve $\boldsymbol{\varepsilon}_{ij}$ hata terimini belirtmektedir. $\mathbf{b}_i$, $w_i(\mathbf{t})$ ve $\boldsymbol{\varepsilon}_{ij}$ terimlerinin bağımsız olduğu varsayılmaktadır. Ayrıca, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{B})$ olduğu ve $w_i(\mathbf{t})$ merkezi Gauss süreci $\Gamma_i(s,t;\gamma^2) = \text{cov}(w_i(\mathbf{t}), w_i(s))$ ve $\boldsymbol{\varepsilon}_{ij} \sim N(\mathbf{0}, \sigma^2)$ olduğu varsayılmaktadır.

Parametre tahmini için Eşitlik (3.15) vektörel olarak Eşitlik (3.16)'daki gibi yazılabilir:

$$\mathbf{Y}_i = \mathbf{f}_i + \mathbf{Z}_i\mathbf{b}_i + \mathbf{w}_i + \boldsymbol{\varepsilon}_i , \; i = 1, ..., m . \quad (3.16)$$

Eşitlik (3.16)'da $\mathbf{f}_i = (f(\mathbf{X}_{i1}), ..., f(\mathbf{X}_{in_i}))^T$, $\mathbf{Y}_i = (\mathbf{Y}_{i1}, ..., \mathbf{Y}_{in_i})^T$, $\mathbf{Z}_i = [\mathbf{Z}_{i1}, ..., \mathbf{Z}_{in_i}]^T$, $\mathbf{w}_i = (w(\mathbf{t}_{i1}), ..., w(\mathbf{t}_{in_i}))^T$ ve $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, ..., \boldsymbol{\varepsilon}_{in_i})^T$ olarak belirtilmektedir. Ayrıca, $(\boldsymbol{\Gamma}_i(t_{ij}, t_{il}; \gamma^2))_{1 \le j,l \le n_i} = \gamma^2 \mathbf{K}_i(t_{ij}, t_{il})$. K_i $(1 \le j, l \le n_i$, $i = 1, ..., m)$ yalnızca zamana bağlı ölçümlere ilişkin pozitif tanımlı matrisi göstermektedir.

Parametre tahmini için Hajjem [21] tarafından yapılan çalışmada izlenen yöntemdeki gibi bir EM algoritması kullanılmakta olup algoritma adımları aşağıda verilmiştir:

0. Adım: r = 0, $\hat{\mathbf{b}}_{i(r)} = 0_q$, $\hat{\mathbf{w}}_{i(r)} = 0_{n_i}$, $\hat{\mathbf{B}}_r = \mathbf{I}_q$, $\hat{\gamma}_{(r)}{}^2 = 1$ ve $\hat{\sigma}^2_{(r)} = 1$ olsun.

1. Adım: $r = r + 1$. $\mathbf{f}$, Eşitlik (3.17)'de göstertildiği gibi standart regresyon ile elde edilir:

$$\hat{\mathbf{y}}_{ij(r-1)} = f(\mathbf{X}_{ij}) + \boldsymbol{\varepsilon}_{ij} . \quad (3.17)$$

Eşitlik (3.17)'de,

$\hat{\mathbf{Y}}_{ij(r-1)} = \mathbf{Y}_{ij} - \mathbf{Z}_{ij}\hat{\mathbf{b}}_{i(r-1)} - \hat{\mathbf{w}}_{ij(r-1)}$ olarak ifade edilmektedir.

Daha sonra $\hat{\mathbf{b}}_{i(r)}$ ve $\hat{\mathbf{w}}_{i(r)}$ parametreleri, verilen $\hat{\mathbf{B}}_{(r-1)}$, $\hat{\gamma}^2_{(r-1)}$ ve $\hat{\sigma}^2_{(r-1)}$ ile tahmin edilir.

2. Adım: $\hat{\mathbf{B}}_{(r)}$, $\hat{\gamma}^2_{(r)}$ ve $\hat{\sigma}^2_{(r)}$ terimleri $\hat{\mathbf{f}}$, $\hat{\mathbf{b}}_{i(r)}$, $\hat{\mathbf{w}}_{i(r)}$ ile güncellenir.

3. Adım: Yakınsama gerçekleşene kadar önceki adımlar tekrarlanır.

Algoritmanın 1.Adımında, bilinen ve önceki iterasyonun tahminleriyle verilen varyans parametreleri dikkate alınmaktadır. Ortalama fonksiyonu f, herhangi bir regresyon yöntemi ile tahmin edilebilir. f, CART ile tahmin edildiğinde, Hajjem ve ark. [27] tarafından geliştirilen MERT algoritmasına atıfta bulunur. CART [15], tahmin için en iyi bölümü elde etmek üzere açıklayıcı değişken uzayının özyinelemeli bir şekilde bölümlenmesinden oluşur. Bölümlemenin her adımında, uzay iki alt parçaya ayrılır. Dolayısıyla, elde edilen bölüm CART Ağacı olarak adlandırılan ikili bir ağaçla ilişkilendirilir. Ayrıca, her bir bölünme tüm açıklayıcı değişkenler arasında optimize edilir ve CART algoritması iki adımda çalışır. Bu adımlar, tahmin hatası açısından en iyi tahmin ediciyi vermek için budama adımı ve maksimum ağaç oluşturma adımıdır.

Hajjem ve ark. [21] tarafından geliştirilen algoritmada f, RO ile tahmin edildiğinde MERF'e atıfta bulunmaktadır. RO, birden fazla rastgele CART ağacının bir araya getirilmesinden oluşur. Burada, bir araya getirme işlemi, bireysel ağaç tahminlerinin ortalamasının birleştirilmesi anlamına gelir. İlk olarak, öğrenme kümesinin bir bootstrap örneği üzerine inşa edilir ve ikinci olarak, bölümlemenin her adımında, en iyi bölünme, rastgele çekilmiş bir açıklayıcı değişken alt kümesi arasında optimize edilir. Genellikle "mtry" olarak adlandırılan değişken alt kümesinin boyutu, yöntemin en önemli parametresidir. RO doğal olarak tahmin hatasını Torba Dışı (Out of Bag, OOB) hata ile tahmin eder. Cevap tahmini için öğrenme kümesinin belirli bir ölçüm değeri için, yalnızca bootstrap örnekleri üzerine inşa edilmiş ağaçlar bu gözlemi içeren veriler toplanır. Ayrıca, OOB örnekleri (gözlemlerden oluşan bootstrap örneklerinde seçilmeyen) bir değişken önem skorunu (variable importance, VI) hesaplamak için de kullanılır. Sabit etkili bir değişken için, değişken önem skoru ilgili ağaçtaki ortalama OOB hatalarının artışı olarak tanımlanır.

Bu bilgiler ışığında; MERF modeline stokastik süreç parametresi eklendiğinde SMERF modeli olarak adlandırılmaktadır [28].

Ortalama fonksiyonu bir CART ağacı T ile tahmin edildiğinde, Sela ve Simonoff [25], ayırma işlemini DKEM ile uygulamayı önermişlerdir.

Φ^i 'nin, $\mathbf{\Phi}^i_{j,l} = \mathbf{I}_{\{x_{ij} \in g_l\}}$ olarak ifade edildiği ve g_l 'nin T ağacının l. yaprağını belirttiği model Eşitlik (3.18) ile verilmiştir:

$$\mathbf{Y}_i = \mathbf{\Phi}^i \boldsymbol{\mu}_T + \mathbf{Z}_i \mathbf{b}_i + \mathbf{w}_i + \boldsymbol{\varepsilon}_i \,. \tag{3.18}$$

1.Adımda T ağacının yapraklarına ilişkin μ_T değerlerinin tahmini Eşitlik (3.19)'da verilmiştir:

$$\hat{\boldsymbol{\mu}}_T = \Big(\sum_{1 \le i \le N} (\mathbf{\Phi}^i)^T \mathbf{V}_i^{-1} \mathbf{\Phi}^i \Big)^{-1} \Big(\sum_{1 \le i \le N} (\mathbf{\Phi}^i)^T \mathbf{V}_i^{-1} \mathbf{Y}_i \Big) \,. \tag{3.19}$$

Eşitlik (3.19)'da,

$\mathbf{V}_i = Var(\mathbf{Y}_i) = \mathbf{Z}_i \mathbf{B} \mathbf{Z}_i^T + \gamma^2 \mathbf{K}_i + \sigma^2 \mathbf{I}_{n_i}$, $i = 1, ..., m$ olarak ifade edilir.

Bu yöntemle, T 'nin yapraklarıyla ilişkili değerleri, yapraktaki değerlerin basit ortalamasını dikkate almak yerine birim içi kovaryans matrisi olan $\mathbf{V}_i$ dikkate alınarak hesaplanır. Modellenen ağaç olan $\hat{\mathbf{f}}_i = \mathbf{\Phi}^T \boldsymbol{\mu}_T$ RE-EM Ağacı olarak adlandırılır.

3.5. Rastgele Etki Beklenti Maksimizasyonu Ormanı (RE-EM Ormanı)

Capitaine [28], Hajjem [27], [21] ile Sela ve Simonoff [25] tarafından geliştirilen mevcut algoritmalara stokastik süreç parametrelerini dahil ettikten sonra, bir dizi RE-EM Ağacı topluluğundan oluşan RE-EM Ormanı algoritmasını geliştirmiştir.

RE-EM Ormanı modeli Eşitlik (3.20)'de verilmiştir:

$$\mathbf{Y}_i = \mathbf{\Phi}^{i,l}\boldsymbol{\mu}_{T_l} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{w}_i + \boldsymbol{\varepsilon}_i . \quad (3.20)$$

Eşitlik (3.20)'de; L rastgele ağaçlar $T_1,....,T_L$, $\mathbf{\Phi}^{i,l}$ l. rastgele ağacın T_l'nin göstergeler matrisi, $\hat{\boldsymbol{\mu}}_{T_l}$ DKEM ile tahmin edilen T_l'nin yapraklarını ifade etmektedir.

RE-EM Ormanı tahmin edicisi L sayıda modellenen RE-EM Ağaçlarının ortalaması olacak şekilde Eşitlik (3.21)'de verilmiştir:

$$\hat{\mathbf{f}}_i = \frac{1}{L}\sum_{l=1}^{L}\mathbf{\Phi}^{i,l}\hat{\boldsymbol{\mu}}_{T_l} . \quad (3.21)$$

Eşitlik (3.21)'de, $\hat{\mathbf{f}}_i$ hesaplandıktan sonra $\mathbf{b}_i$ rastgele etki tahminleri ve $\mathbf{w}_i$ stokastik süreç değerleri $\mathbf{B}$, γ^2 ve σ^2 bilinen parametreler ile sırasıyla Eşitlik (3.22) ve Eşitlik (3.23) ile tahmin edilir:

$$\hat{\mathbf{b}}_i = \mathbf{B}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \hat{\mathbf{f}}_i), \quad (3.22)$$

$$\hat{\mathbf{w}}_i = \gamma^2\mathbf{K}_i\mathbf{V}_i^{-1}(\mathbf{Y}_i - \hat{\mathbf{f}}_i). \quad (3.23)$$

3.6. Genelleştirilmiş Doğrusal Karma Etki Modeli Ağacı (GDKEM Ağacı)

GDKEM Ağacı yöntemi Fokkema ve ark. [31] tarafından gruplanmış verilerde tedavi alt grupları arasındaki etkileşimleri tespit edebilmek için geliştirilmiştir. Bu model global ve yerel olmak üzere iki kısımdan oluşmaktadır. GDKEM'nin global kısmı, tüm gözlemlerle birlikte rastgele etki terimlerini içerir. Yerel kısım ise sabit etkilerden oluşur. Algoritma veri kümesini ayırma değişkenleri olarak bilinen ek değişkenlere göre bölümlere ayırır. Ardından, bölümlere ayrılan her bir ayrı veri kümesinde sabit etkiler tahmin edilir. GDKEM Ağacı algoritması alt grupları tespit etme esnekliğine sahiptir çünkü Zeileis ve ark. [48] tarafından önerilen modele dayalı özyinelemeli ayrıştırma (model-based recursive partitioning, MOB) yöntemine dayalı olarak çalışmaktadır. MOB, Genelleştirilmiş Doğrusal Modeller (GDM) gibi tek bir modelin uzun süreli veri kümesini tam anlamıyla açıklayamayacağını varsayar.

3.6.1. Modele Dayalı Özyinelemeli Ayrıştırma (MOB)

MOB'nin temel mantığı, verilerin bazı durumlarda GDM gibi tek bir global model tarafından iyi tanımlanamayabileceği ve verilerin diğer ek değişkenlerle bölümlendirilebileceğidir. Bu gibi durumlarda, veriler için her bir bölümde daha iyi uyumlar bulmak mümkün olabilir. Örneğin, cevap üzerindeki tedavi etkisini incelemek için global bir GDM oluşturulabilir, ancak bu model sabit etkili değişkeninin cevap değişkeni üzerindeki etkisi tespit edilmek istendiğinde bu sabit etkinin katsayısı aynı katsayıyı gösterebileceğinden bu etki tüm gözlemler için aynı olacaktır. Diğer yandan, farklı birimlerin farklı özellikler göstermesi durumunda, verideki farklı gözlem kümeleri ile veri bölümlemesi diğer açıklayıcı değişkenlerle birlikte ayrıştırılması gerekebilir.

$\mathbf{x}_i$ açıklayıcı değişken matrisi olmak üzere, beklenen cevap $\mathbf{y}_i$'yi modelleyebilecek tek bir global GDM Eşitlik (3.24)'teki gibi bir bağ fonksiyonu aracılığıyla verilmektedir:

$$g(\boldsymbol{\mu}_i) = \mathbf{x}_i^T \boldsymbol{\beta} . \tag{3.24}$$

Eşitlik (3.24)'te, $\mathbf{x}_i$, i. birimin açıklayıcı değişkenler matrisini, $\boldsymbol{\beta}$, sabit etki regresyon katsayıları vektörünü, $\boldsymbol{\mu}_i$, $g()$ bağ fonksiyonu vasıtasıyla $\mathbf{x}_i\boldsymbol{\beta}$ ile doğrusal olarak modellenen beklenen cevapları göstermektedir.

En iyi uyum, özellikle bölümleme değişkenleri mevcut olduğunda, Eşitlik (3.24)'teki gibi tek bir modelle elde edilemeyebilir. Bu model, tüm veri noktaları için cevaplar üzerinde aynı etkiyi/katsayıyı dikkate aldığından, potansiyel bölümleme değişkenlerinin mevcut olması durumunda verilere iyi uyum sağlayamaz. MOB algoritması, veri bölümlerini bulurken diğer açıklayıcı değişkenleri de hesaba katabilir ve daha iyi yerel modellere uyum sağlayabilir. MOB algoritması bunu başarmak için bir dizi bölümleme değişkeni üzerinde parametre kararlılık testlerini kullanır.

3.6.2. Rastgele Etkilerin Modele Katılması

Uzun süreli veri yapısının analizi için karma etki modelinin kullanılması daha uygundur. Bu tür verilerin analizi için Eşitlik (3.24)'ten genişletilen GDM Eşitlik (3.25)'teki gibi verilebilir:

$$g(\boldsymbol{\mu}_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{Z}_i^T \mathbf{b} . \quad (3.25)$$

Eğer model yalnızca rastgele bir kesişimden (random intercept) oluşuyorsa, $\mathbf{z}_i$ M elemanlı bir birim vektörünü belirtir ve bunun m. girişi ($m = 1, ..., M$) 1 iken diğer girişleri 0 değerini alır. m, i birimin kümesini temsil eder. Ayrıca, $\mathbf{b}$, M boyutlu rastgele etki vektörünü belirtir. Burada, m. küme için her bir m. elemanın rastgele kesişimi yer alır.
Eşitlik (3.25)'te verilerin kümelenmiş yapısı rastgele kısım tarafından açıklanabilse de, global sabit etkiler kısmı $\mathbf{x}_i^T \boldsymbol{\beta}$ verilere iyi uyum sağlayamayabilir. Karma etki modelinin bu sınırlamasını göz önünde bulunduran Fokkema ve ark. [31] GDKEM Ağacı yöntemini geliştirmişlerdir. Önerdikleri yöntemde, sabit etkili katsayılar da verileri bölebilmekte olup model denklemi Eşitlik (3.26) ile verilmiştir:

$$g(\boldsymbol{\mu}_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}_j + \mathbf{Z}_i^T \mathbf{b} . \quad (3.26)$$

Eşitlik (3.26)'da, sabit etki katsayıları $\boldsymbol{\beta}_j$ yerel kısımdır ve bu terimlerin değerleri j terminal düğümüne bağlıyken rastgele etkiler ($\mathbf{b}$) algoritmanın global kısmına karşılık gelir.

Eşitlik (3.26)'nın tahmin edilmesine yönelik GDKEM Ağacı algoritma adımları aşağıda verilmiştir:

0. Adım: $r = 0$, $\hat{\mathbf{b}}_{(r)} = 0$ olsun.

1. Adım: $r = r + 1$, $\mathbf{Z}_i^T \hat{\mathbf{b}}_{(r-1)}$ kullanarak GDM Ağacı modeli oluşturulur.

2. Adım: $g(\boldsymbol{\mu}_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}_j + \mathbf{Z}_i^T \mathbf{b}$ karma etki modeli $j(r)$ terminal düğümü ile 1.Adımdaki GDM Ağacı ile tahmin edilir.

3. Adım: Yakınsama gerçekleşene kadar 1. ve 2. Adımı tekrarlanır.

Rastgele etkiler başlangıçta bilinmediğinden, algoritma bunları 0'a eşitleyerek başlar. Her iterasyonda, algoritma GDM Ağacını 1. Adımda özyinelemeli olarak uyarlar, böylece sabit

ve rastgele etkiler 2. Adımda yeniden tahmin edilir. Rastgele etkiler üzerinde herhangi bir bölümleme yoktur, ancak bunlar global olarak tahmin edilir. Öte yandan, algoritma sabit etkileri her bir bölüm hücresinde yerel olarak tahmin eder.

3.7. Gauss Süreci Güçlendirmesi (GP Boosting)

Güçlendirme (Boosting) [49], bir makine öğrenmesi algoritmasının tahmin yeteneğini geliştirebilen genel bir tekniktir. Algoritmanın tahmin performansını artırmanın yanı sıra, temel öğreniciler olarak ağaçları kullanan boosting, çoklu bağlantı, doğrusal olmama durumu, kesikli veri durumu ve yüksek dereceli etkileşimlerle başa çıkabilmektedir. Ayrıca, güçlendirme yöntemi, veri kümesinde kayıp gözlemlerin ve aykırı değerlerin olması durumunda herhangi bir bilgi kaybı olmadan performans gösterebilir [29]. Öte yandan, farklı veri noktalarında birimler arası bağımsızlık varsayımı gibi bazı sınırlamalar vardır. Hatalar bir korelasyon sergilediğinde algoritma doğru bir şekilde çalışamaz. Uzun süreli verilerdeki korelasyon yapısının yanı sıra, güçlendirme çok düzeyli kategorik değişkenlerde zorluk yaşamaktadır.

Rasmussen [50] Gauss Sürecini "herhangi bir sonlu sayıda (tutarlı) ortak Gauss dağılımına sahip rastgele değişkenler topluluğu" olarak tanımlamaktadır. Bu süreçler, esnek parametrik olmayan modellerin üstün tahmin doğruluğu sağlamasına ve olasılıksal tahminler yapmasına olanak tanımaktadır.

Karma etki modelleri uzun süreli veriler için yaygın olarak kullanılan tekniklere sahiptir. Bu tür veri kümelerinin gruplama yapısı nedeniyle ölçümler korelasyon sergileyebilir. Bu modeller yüksek düzeylere sahip kategorik değişkenleri de ele alabilmektedir.

Gauss süreci ve karma etkiler regresyon modellerinde ortalamanın ya sıfır ya da verilen açıklayıcı değişkenlerin doğrusal bir fonksiyonu olduğu varsayılır. Yapılandırılmış hata varyansı, sıfır ortalamalı bir Gauss süreci ve bir karma etki modeli kullanılarak modellenebilir. Ancak hem sıfır ortalama hem de doğrusallık varsayımı genellikle gerçekçi değildir ve yüksek tahmin performansı elde etmek için bu varsayımların gevşetilmesi gerekebilir [29].

Sigrist [29], ağaç güçlendirmeyi (tree boosting) Gauss süreci ve karma etki modeli ile birleştirebilen bir yöntem geliştirmiştir. Sabit etkiler ve cevaplar arasındaki ilişkiyi temsil etmek için esnek bir fonksiyonel forma sahip karma etki modeli Eşitlik (3.27) ile verilmiştir:

$$\mathbf{y} = F(\mathbf{X}) + \mathbf{Zb} + \boldsymbol{\varepsilon}. \tag{3.27}$$

Eşitlik (3.27)'de, $\mathbf{y} = (\mathbf{y}_1,....,\mathbf{y}_n)^T \in \mathbb{R}^n$ cevap değişkenini, $\mathrm{F}(\mathbf{X}) \in \mathbb{R}^n$ sabit etkileri, $\mathbf{b} \in \mathbb{R}^m \sum \in \mathbb{R}^{m \times m}$ kovaryans matrisli rastgele etkileri ve $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1,....,\boldsymbol{\varepsilon}_n)^T \in \mathbb{R}^n$ hata terimini belirtmektedir. n gözlem sayısını, m rastgele etki boyutunu, p sabit etki sayısını göstermektedir. Gruplandırılmış rastgele etkiler durumunda, Eşitlik (3.27)'deki $\mathbf{Z}$ matrisi bir insidans matrisi $\mathbf{Z} \in \{0,1\}^{n \times m}$ ve bu matris birim düzeyindeki rastgele etkileri veri noktalarını ifade eder. Ayrıca $\mathbf{b} \sim N(\mathbf{0}, \sum)$ ve $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ olduğu varsayılmaktadır.

Eşitlik (3.27)'de verilen model için risk fonksiyonu olarak da bilinen olabilirlik fonksiyonu Eşitlik (3.28) ile verilmiştir:

$$p(\mathbf{y} \mid F, \boldsymbol{\theta}) = \int p(\mathbf{y} \mid \mathbf{b}, F, \boldsymbol{\theta})\, p(\mathbf{b} \mid \boldsymbol{\theta}) db. \tag{3.28}$$

Eşitlik (3.28)'de,

$p(\mathbf{b} \mid \boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\mathbf{b}^T \sum^{-1} \mathbf{b}\right) \mid \sum \mid^{-1/2} (2\pi)^{-m/2}$ olarak belirtilir. $F = F(\mathbf{X})$ ve $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ tüm varyans ve kovaryans parametrelerini belirtmektedir. $F(.)$ fonksiyonu, $\mathbf{X}$'te değerlendirilen F fonksiyonunu belirtir. Ayrıca, θ_1, $\boldsymbol{\theta}$'nın ilk elemanı olup σ^2 varyansına eşittir [29].

Cevap değişkeninin marjinal dağılımı Eşitlik (3.29)'da verilmiştir:

$$\mathbf{y} \sim N(F(\mathbf{X}), \boldsymbol{\Psi}),\ \boldsymbol{\Psi} = \mathbf{Z} \sum \mathbf{Z}^T + \sigma^2 \mathbf{I}_n. \tag{3.29}$$

Bu modelin negatif log en çok olabilirlik fonksiyonu Eşitlik (3.30) ile verilmiştir:

$$L(\mathbf{y}, F, \boldsymbol{\theta}) = \frac{1}{2}(\mathbf{y} - F)^T \boldsymbol{\Psi}^{-1}(\mathbf{y} - F) + \frac{1}{2}\log\det(\boldsymbol{\Psi}) + \frac{n}{2}\log(2\pi). \tag{3.30}$$

Eşitlik (3.31)'de $\boldsymbol{\Psi}^{\dagger} = \boldsymbol{\Psi} / \sigma^2$ olacak şekilde yeniden düzenleme yapılmıştır:

$$\mathbf{y} \sim N(F(\mathbf{X}), \sigma^2 \boldsymbol{\Psi}^{\dagger}),\ \sum{}^{\dagger} = \sum / \sigma^2. \tag{3.31}$$

Amaç, verilen bu negatif log-olabilirlik için risk fonksiyonunu $L(y; F; \theta)$'yi minimize etmektir:

$$R(F(.), \boldsymbol{\theta}) :\ (F(.), \boldsymbol{\theta}) \rightarrow L(\mathbf{y}, F, \boldsymbol{\theta})_{F = F(\mathbf{X})}$$

$F(.)$, $\mathbf{X}$'te bir fonksiyondur ve fonksiyon uzayı H ile ifade edilebilir. Sigrist [29] tarafından geliştirilen algoritmaya göre, Eşitlik (3.32) ve Eşitlik (3.33)'teki ortak en küçükleyen değer, Gauss sürecini DKEM ile birleştirerek elde edilebilir:

$$(\hat{F}(.), \hat{\boldsymbol{\theta}}) = \arg\min_{(F(.), \boldsymbol{\theta}) \in (H, \Theta)} R(F(.), \boldsymbol{\theta}), \tag{3.32}$$

$$p(\mathbf{y} \mid \mathbf{b}, F, \boldsymbol{\theta}) = \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - F - Z\mathbf{b})^T (\mathbf{y} - F - \mathbf{Z}\mathbf{b}) \right)(2\pi\sigma^2)^{-n/2}. \tag{3.33}$$

θ **sabit olduğunda güçlendirme:** Sabit varyans-kovaryans matrisi $\boldsymbol{\theta}$ verildiğinde, güçlendirme algoritması Eşitlik (3.34)'teki gibi $F_{m-1}(.)$'nin son değerine $f_m(.)$ değeri eklenip güncellenmesiyle elde edilir:

$$F_m(.) = F_{m-1}(.) + f_m(.),\ f_m \in S,\ m = 1, ..., M\ . \tag{3.34}$$

Gradyan güçlendirme yönteminde $f_m(.)$, Eşitlik (3.35)'te verilen en küçük kareler yakınsamasıyla (Least Squared Approximiation) elde edilir:

$$f_m(.) = \arg\min_{f(.)\in S} \| \mathbf{\Psi}^{-1}(\mathbf{y} - F_{m-1} - f \|^2 . \quad (3.35)$$

Eşitlik (3.35)'te, $\mathbf{f} = (f(\mathbf{X}_1),...., f(\mathbf{X}_M))^T$ olarak ifade edilir.

Sabit bir $\boldsymbol{\theta}$ için, Newton güçlendirme yöntemi Eşitlik (3.36)'da belirtildiği gibi ikinci dereceden Taylor fonksiyonu yakınsamasının en küçükleyenini bulmaktadır:

$$f_m(.) = \arg\min_{f(.)\in S} (\mathbf{y} - F_{m-1} - \mathbf{f})^T \mathbf{\Psi}^{\dagger^{-1}} (\mathbf{y} - F_{m-1} - \mathbf{f}) . \quad (3.36)$$

Bu iki yöntemi birleştiren hibrit bir gradyan Newton güçlendirme yöntemi de bulunmaktadır. Bu durumda, temel öğrenicilerin Eşitlik (3.37)'deki gibi olduğu varsayılır:

$$f(.) = h(.;\alpha)^T \Upsilon , \; h(.;\alpha), \Upsilon \in \mathbb{R}^L, \alpha \in \mathbb{R}^Q . \quad (3.37)$$

Eşitlik (3.37)'de, α ve Υ temel öğrenici parametrelerini, $h(.;\alpha) : \mathbb{R}^p \to \mathbb{R}^L$ belirtmektedir. $h(.;a)$, tahmin edicileri ağacın terminal düğümleriyle eşleyen bir fonksiyonu ifade eder. Böyle bir durumda, bu güçlendirme yöntemi başlangıçta a_m'yi Eşitlik (3.35)'te tanımlanan gradyan güçlendirme kullanarak öğrenir ve daha sonra Eşitlik (3.36)'da tanımlanan Newton güçlendirme adımını kullanarak Υ_m'yi öğrenir.

Genelleştirilmiş en küçük kareler çözümü Eşitlik (3.38) ile verilmiştir:

$$\Upsilon_m = (h_{\alpha_m}^T \mathbf{\Psi}^{\dagger^{-1}} h_{\alpha_m})^{-1} h_{\alpha_m}^T \mathbf{\Psi}^{\dagger^{-1}} (\mathbf{y} - F_{m-1}) . \quad (3.38)$$

Eşitlik (3.38)'de, $h_{\alpha_m} \in \mathbb{R}^{n\times L}$ $(h_{\alpha_m})_{il} = (h(\mathbf{X}_i;\alpha_m))_l$, $i = 1,...., m$, $l = 1,...., L$ elemanlı bir matristir.

Eşitlik (3.34)'teki güncelleme, daha yüksek tahmin performansı elde etmek için bir ν faktörü ile değiştirilerek Eşitlik (3.39) ile verilebilir:

$$F_m(.) = F_{m-1}(.) + \nu\, f_m(.)\,,\ \nu > 0\,. \tag{3.39}$$

Eşitlik (3.39)'da ν, öğrenme oranını ifade eder.

3.8. Model Performans Ölçütleri

Farklı makine öğrenmesi algoritmalarının cevap değişkeni sürekli olan bir uzun süreli veri kümesi üzerinde uygulanmasının ardından en iyi modelin seçilebilmesi için bazı model performans ölçütlerinin hesaplanması gerekmektedir.

Bu kapsamda üç adet model performans ölçütü olan Hata Kareler Ortalaması (HKO) (Mean Squared Error, MSE), HKO Kare Kökü (HKOK) (Root Mean Squared Error, RMSE) ve Ortalama Mutlak Hata (OMH) (Mean Absolute Error, MAE) sırasıyla Eşitlik (3.40), Eşitlik (3.41) ve Eşitlik (3.42)'de verilmiştir:

$$HKO = \frac{1}{n}\sum_{i=1}^{m}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2\,, \tag{3.40}$$

$$HKOK = \sqrt{\frac{1}{n}\sum_{i=1}^{m}(\hat{\mathbf{y}}_i - \mathbf{y}_i)^2}\,, \tag{3.41}$$

$$OMH = \frac{1}{n}\sum_{i=1}^{m}|\,\hat{\mathbf{y}}_i - \mathbf{y}_i\,| \tag{3.42}$$

Yukarıdaki eşitliklerde $\mathbf{y}_i$ ve $\hat{\mathbf{y}}_{\hat{i}}$ sırasıyla n birim üzerinden gözlenen ve tahmin edilen cevap değerlerini göstermektedir.

HKO gözlenen ve tahmin edilen cevap değerleri arasındaki ortalama hata karelerini gösterirken, HKOK ise ortalama sapmaları göstermektedir. HKOK değeri cevap değeri ölçü birimi ile aynı birime sahip olduğundan HKO'ya göre daha karşılaştırılabilirdir. OMH ise hataların mutlak değer ortalamalarını belirtir ve veri kümesinde aykırı değerlerin olması durumunda diğer yöntemlere göre daha sağlamdır.

4. UYGULAMA

Bu çalışmada, sağa çarpık bir dağılıma sahip sürekli yapıdaki uzun süreli veri yapısı için hem klasik bir istatistiksel yöntem olan GDKEM'yi hem de modern analiz yöntemlerinden makine öğrenmesi algoritmalarının incelenmesi amaçlanmıştır. Türkiye İstatistik Kurumu (TÜİK)'ndan temin edilen Türkiye'de trafiğe kayıtlı motorlu araçlara ilişkin 2013-2023 yıl aralığını kapsayan araç muayene verileri kullanılmıştır. Türkiye'de trafiğe kayıtlı motorlu araçların yıllara göre katettikleri mesafelere etki eden faktörlerin araştırılmasına ilişkin hazırlanmış uzun süreli veriler, istatistiksel bir yöntem olan GDKEM ve makine öğrenmesi algoritmaları kullanılarak analiz edilmiş ve performans ölçütlerine göre en iyi sonuç veren model seçilmiştir.

Bu çalışma, uzun süreli veri analizi literatüründe sağa çarpık uzun süreli verilerin makine öğrenmesi algoritmaları ile incelendiği ve bu algoritmaların klasik bir istatistiksel yöntem olan GDKEM ile karşılaştırmalarının yapıldığı ilk çalışma olma özelliğine sahiptir.

Çalışmada sonuçlar, R programının "*glmmTMB"* [51], "*LongituRF"* [52], "*gpboost"* [53] ve "*glmertree"* [54] paketlerinden yararlanılarak elde edilmiştir.

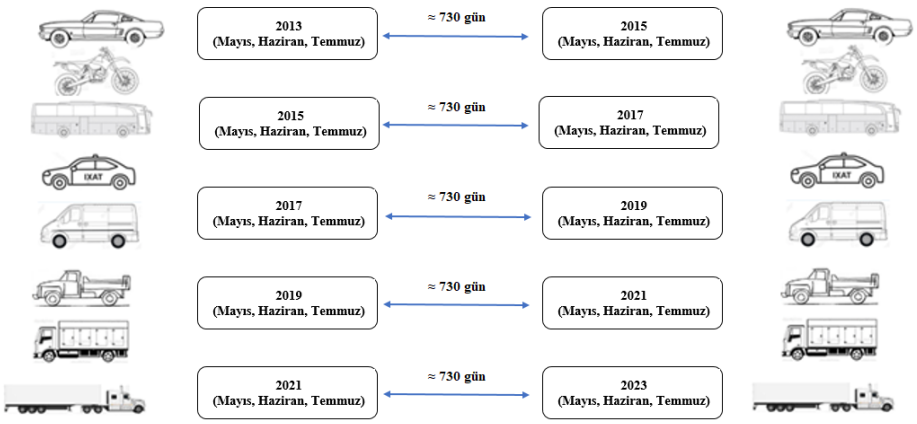
Türkiye'de trafiğe kayıtlı motorlu araçlara ilişkin zorunlu araç muayene sürecinde araçlara ilişkin şu veriler toplanabilmektedir: araçların cinsi, kullandıkları yakıt türü, kullanım amacı, model yılı gibi teknik bilgiler ile muayene esnasında ölçülen km saati değeri (odometre değeri) ve muayene tarihleri. Her bir araç cinsi için T.C. Ulaştırma ve Altyapı Bakanlığı tarafından belirlenen zorunlu muayene periyotları bulunmakta ve araçların bu periyotlara uyacak şekilde muayene istasyonlarına gelerek zorunlu muayenelerini yaptırmaları beklenmektedir. Çizelge 4.1'de araç cinslerine göre zorunlu muayene periyotları verilmiştir [55]:

Çizelge 4.1. Araçların zorunlu muayene periyotları

Araç cinsi	**Muayene periyodu**
Hususi/Resmi otomobiller ve her türlü römorkları	İlk üç yaş ve sonrası her iki yılda bir
Taksi, kamyon, otobüs, kamyonet vb. ticari araçlar	İlk bir yaş ve sonrası her yıl
İki veya üç tekerlekli araçlar ve her türlü römorkları	İlk üç yaş ve sonrası her iki yılda bir

Çizelge 4.1'de verilen bilgiler doğrultusunda uzun süreli verilerin hazırlanmasında tüm hedef araç gruplarının analizde temsil edilebilmesi için iki yıllık zaman aralığına göre ölçümlerin oluşturulmasına karar verilmiş ve araçların 2013, 2015, 2017, 2019, 2021 ve 2023 yıllarındaki ölçüm değerleri dikkate alınmıştır. Araçların belirtilen bu yıllarda muayeneye geliş tarihleri yılın geneline yayılmakta, dolayısıyla farklılık göstermektedir. Örneğin, bir araç 21 Şubat 2023 tarihinde muayeneye girebileceği gibi bir diğer araç 20 Aralık 2023 tarihinde muayeneye girebilmektedir. Bununla birlikte, spesifik bir araç zaman boyunca takip edildiğinde aracın ikinci ve daha sonraki muayene tarihleri genel olarak ilk muayene tarihlerine benzerdir. Bunun temel nedenlerinden birisi maddi kaygıdır. Araçlar belirtilen muayene periyotlarının dışına çıktığı zaman gecikilen gün sayısı üzerinden cezai yaptırıma tabi tutulmaktadır. Tam tersi düşünüldüğünde, bir araç bir sonraki dönemde gerçekleşmesi gereken muayene tarihinden daha önce bir tarihte muayeneye maruz kalırsa bu durumda ödenen muayene ücretinin bir kısmı önemsiz bir hale gelmiş olacaktır.

Tüm bu bilgiler göz önünde bulundurularak hazırlanan uzun süreli veri kümesinde sağlıklı analizlerin yapılabilmesi için araçların muayene tarihlerinin standart bir zaman aralığında olması gerekir. Şekil 4.1'de araçların bu gereksinime uygun şekilde nasıl seçildiği gösterilmiştir:



Şekil 4.1. Araçların seçimi

Şekil 4.1'de, her iki muayene arasında yaklaşık 730 gün (iki takvim yılı) olan ve bu koşulu 2013 yılından 2023 yılına kadar sağlayan araçlar tespit edilerek dengeli bir uzun süreli veri kümesi oluşturulmuştur.

Her bir aracın zaman boyunca katettiği mesafe bilgisi ardışık iki muayenede kaydedilen km saati değerlerinin farkı alınarak hesaplanabilir. Hesaplamalarda cevap değişkeni olarak kullanılacak olan bu değişken standart bir ölçeğe sahip olarak yüzde fark şeklinde hesaplanmıştır.

Analizde kullanılan uzun süreli verilerin kaynağı idari kayıtlardır ve idari kayıtlar, veri giriş hataları veya bazı diğer nedenlerden dolayı aykırı değerler barındırabilmektedir. Çalışmada aykırı değerlere sahip olan araçlar veri kümesinden çıkarılmıştır. Elde edilen veri kümesi 1569 adet farklı aracın 5 farklı dönemde (2013-2015, 2015-2017, 2017-2019, 2019-2021, 2021-2023) ölçümlerini içeren dengeli bir uzun süreli veri kümesidir ve toplam 7845 adet ölçümü içermektedir. Çizelge 4.2'de, çalışmada kullanılan uzun süreli veri kümesinde yer alan değişkenlere ait tanımlar verilmiştir:

Çizelge 4.2. Modelde kullanılan değişken tanımları

Değişken	**Değişken açıklaması**
kmFark	Cevap değişkeni: Araçların bir önceki yıla göre katettiği mesafenin yüzde (%) artışı
yil	Açıklayıcı değişken: Ölçüm yapılan yıl (0, 2, 4, 6, 8)
cinsi	Açıklayıcı değişken: Araçların cinsi (Düzeyler: Otomobil, Minibüs, Otobüs, Kamyonet, Kamyon, Motosiklet, Özel amaçlı) Referans düzey: Kamyon
yakit_turu	Açıklayıcı değişken: Araçların kullandıkları yakıt türü (Düzeyler: Benzinli, Dizel, Lpg) Referans düzey: Benzinli
kullanim_amaci	Açıklayıcı değişken: Araçların kullanım amaçları (Düzeyler: Hususi, Resmi, Ticari) Referans düzey: Hususi

Çalışmada kullanılan uzun süreli veri kümesinde yer alan araçların sayısı 5 farklı zaman noktası için araçların cinsi Çizelge 4.3'te, yakıt türü Çizelge 4.4 'te ve kullanım amacına göre, Çizelge 4.5'te verilmiştir. Veri kümesinde toplam 1569 adet farklı araca ait 5 farklı zaman noktasında dengeli olacak şekilde toplam 7845 adet gözlem bulunmaktadır.

Çizelge 4.3. Araç cinsi ve zaman noktalarına göre araç sayıları

Araç cinsi	**Yıl**					**Toplam**
	0	**2**	**4**	**6**	**8**	
Kamyon	220	221	220	219	219	1099
Kamyonet	310	310	310	310	310	1550
Minibüs	146	148	149	151	152	746
Motosiklet	177	177	177	177	177	885
Otobüs	46	44	42	39	38	209
Otomobil	644	644	644	644	644	3220
Özel Amaçlı	26	25	27	29	29	136
Toplam	1569	1569	1569	1569	1569	7845

Çizelge 4.4. Yakıt türü ve zaman noktalarına göre araç sayıları

Yakıt türü	Yıl					Toplam
	0	2	4	6	8	
Benzinli	314	303	295	289	282	1483
Dizel	949	952	952	952	951	4756
Lpg	306	314	322	328	336	1606
Toplam	1569	1569	1569	1569	1569	7845

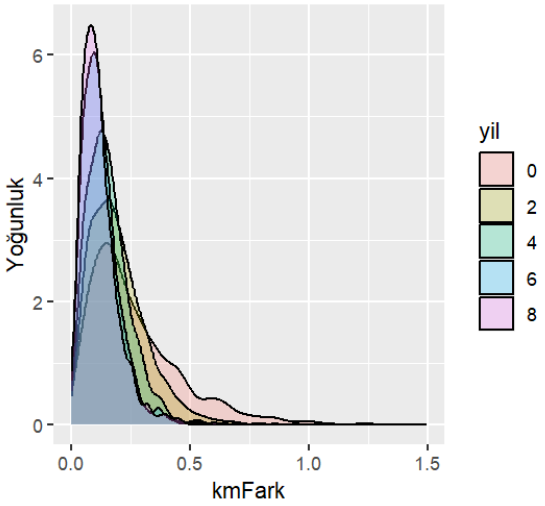
Çizelge 4.5. Kullanım amacı ve zaman noktalarına göre araç sayıları

Kullanım amacı	Yıl					Toplam
	0	2	4	6	8	
Hususi	1048	1051	1054	1056	1069	5278
Resmi	134	135	135	134	132	670
Ticari	387	383	380	379	368	1897
Toplam	1569	1569	1569	1569	1569	7845

Çizelge 4.6'da kmFark cevap değişkenine ilişkin tanımlayıcı istatistikler her bir farklı zaman noktası için verilmiş ve Şekil 4.2'de kmFark değişkeninin zaman noktaları için dağılım grafiği gösterilmiştir:

Çizelge 4.6. kmFark değişkenine ilişkin tanımlayıcı istatistikler

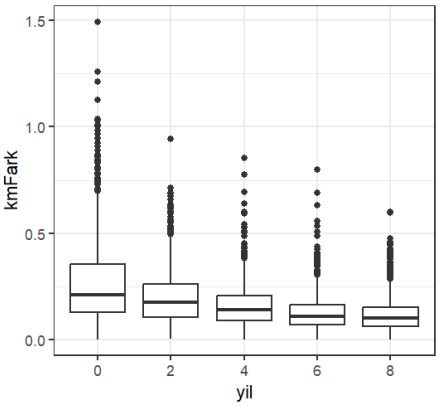
İstatistik	Yıl				
	0	2	4	6	8
Minimum ölçüm değeri	0,000018	0,003181	0,000540	0,000808	0,000014
1. çeyrek değeri	0,127002	0,106366	0,088845	0,069418	0,063632
Ortanca	0,210478	0,175041	0,142372	0,109626	0,101831
Ortalama	0,264081	0,195978	0,156880	0,125086	0,117272
3. çeyrek değeri	0,353216	0,260057	0,205632	0,162794	0,151779
Maksimum ölçüm değeri	1,494547	0,942628	0,854299	0,799856	0,600931
Varyans	0,037408	0,014273	0,008934	0,006566	0,005931
Standart sapma	0,193413	0,119471	0,094521	0,081028	0,077013



Şekil 4.2. kmFark değişkeninin farklı zaman noktalarındaki dağılım grafiği

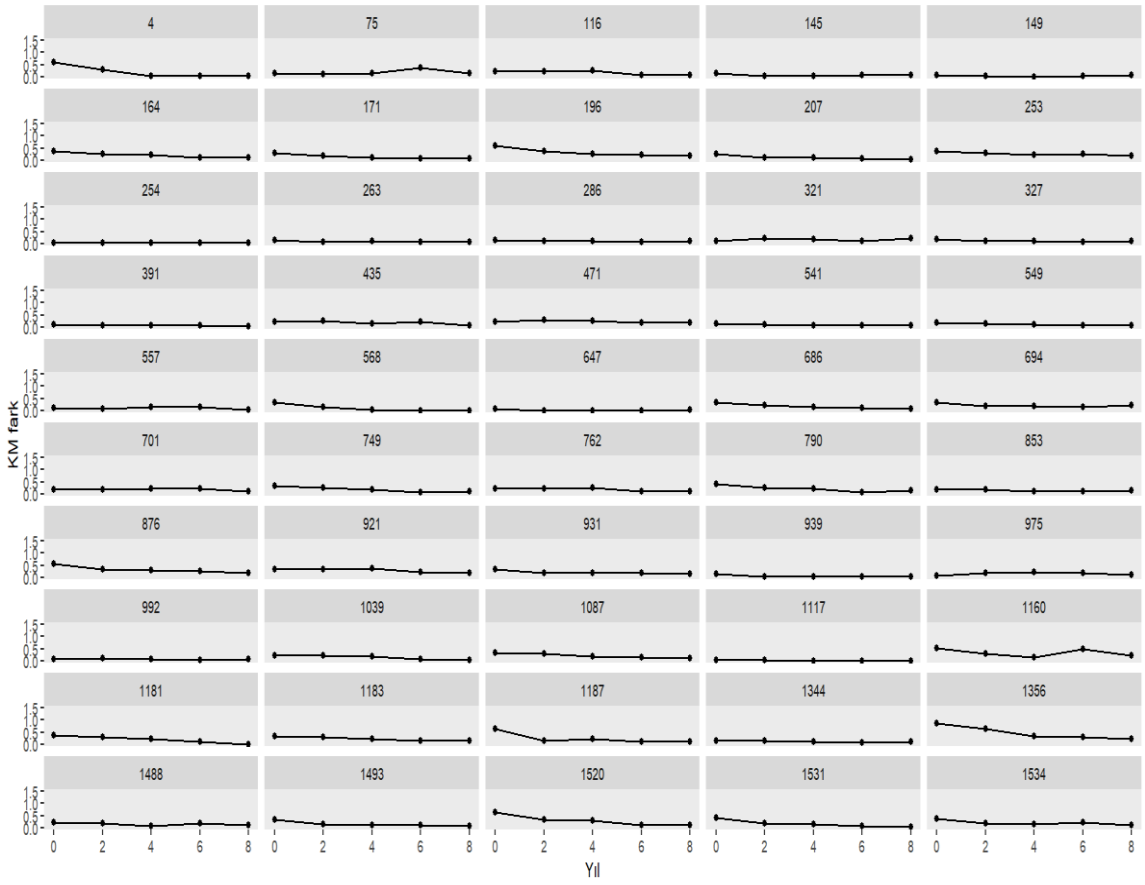
Şekil 4.2'de, kmFark değişkeninin her bir zaman noktasında simetrik bir dağılıma sahip olmayıp sağa çarpık bir dağılıma sahip olduğu görülmektedir.

kmFark değişkenin her bir zaman noktasındaki box-plot grafiği Şekil 4.3'te gösterilmiştir:

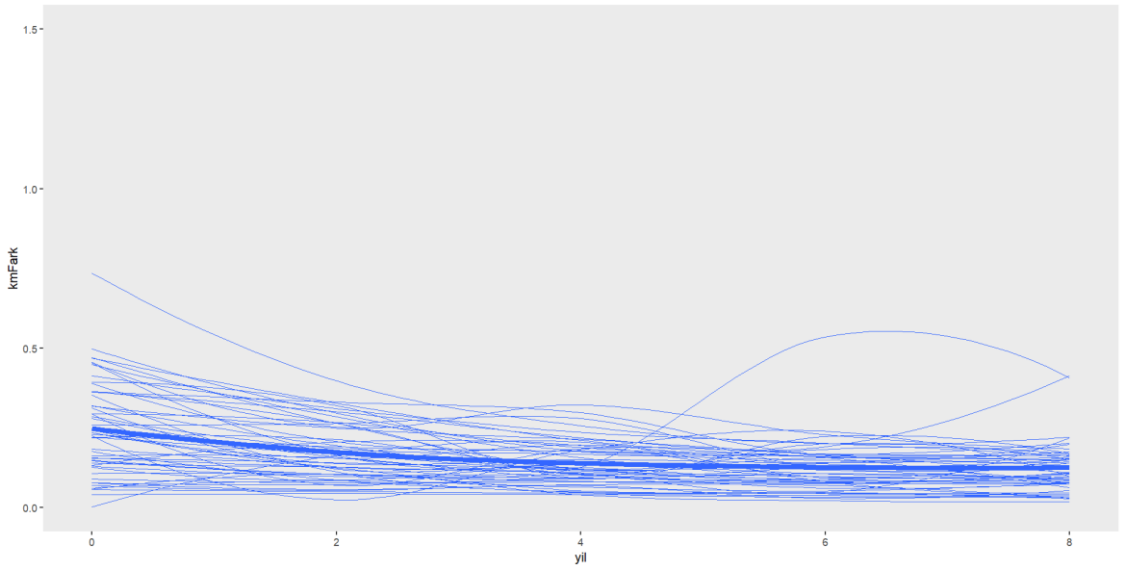


Şekil 4.3. kmFark değişkeninin zaman noktalarına göre box-plot grafiği

Şekil 4.4'te ve Şekil 4.5'te uzun süreli veri kümesinden rastgele seçilen 50 araca ilişkin kmFark değişkenlerinin zamana bağlı olarak değişimi gösterilmiştir. kmFark değerlerinin genel olarak zaman boyunca azalan bir trende sahip olduğu Şekil 4.4'te görülmektedir. Şekil 4.5'teki spagetti grafiğinde mavi kalın çizgi rastgele seçilen 50 aracın 5 farklı zaman noktasındaki ortalama kmFark değerlerinin zamana göre değişimini belirtmektedir. Burada araçların kmFark değerlerinin zaman boyunca azaldığı görülmektedir.



Şekil 4.4. Rastgele seçilen 50 araç için kmFark değerinin zaman noktalarına göre değişimi



Şekil 4.5. Rastgele seçilen 50 adet araç için spagetti grafiği

4.1. İstatistiksel ve Makine Öğrenmesi Yöntemlerinin Uygulanması

Bu bölümde oluşturulan uzun süreli veri kümesi üzerinde karma etkili istatistiksel ve makine öğrenmesi yöntemleri uygulanmıştır. Analizlerde kmFark değişkeni cevap değişkeni, yil, cinsi, yakit_turu ve kullanim_amaci değişkenleri açıklayıcı değişkenlerdir. Analizler "12th Gen Intel(R) Core(TM) i7-12700 2.10 GHz" işlemcili ve 16 Gb geçici bellekli masaüstü bilgisayarda R programı kullanılarak yapılmıştır.

4.1.1. GDKEM'ye Göre Analiz Sonuçları

kmFark cevap değişkeni sağa çarpık bir dağılıma sahip olduğu için Eşitlik (4.1)'deki GDKEM kullanılmış ve analizler R programında *"glmmTMB"* [51] paketi kullanılarak yapılmıştır.

$$\begin{aligned} h\{E(\mathbf{Y}_i \mid \mathbf{b}_i, \mathbf{X}_i, \mathbf{Z}_i)\} &= \beta_0 + \beta_1 \times yil + \beta_2 \times cinsiKamyonet + \beta_3 \times cinsiMinibüs \\ &+ \beta_4 \times cinsiMotosiklet + \beta_5 \times cinsiOtobüs + \beta_6 \times cinsiOtomobil + \beta_7 \times cinsiÖzelAmaçlı \\ &+ \beta_8 \times yakit\_turuDizel + \beta_9 \times yakit\_turuLpg + \beta_{10} \times kullanim\_amaciResmi \\ &+ \beta_{11} \times kullanim\_amaciTicari + b_{0i} + b_{1i} \times yil \end{aligned} \tag{4.1}$$

Eşitlik (4.1)'de; h, logaritmik bağ fonksiyonu, β_0 sabit etki, β_1 yil değişkeni için sabit etki, $\beta_2, ..., \beta_7$ cinsi değişkeninin düzeyleri için sabit etki β_8 ve β_9 yakit_turu değişkeni için sabit etki β_{10} ve β_{11} ise kullanim_amaci değişkeni düzeyleri için sabit etkilerdir. b_{0i} ve b_{1i}, i. birime ait rastgele etkilerdir.

Eşitlik (4.1)'deki GDKEM'de cinsi değişkeni için "Kamyon", yakit_turu değişkeni için "Benzinli" ve kullanim_amaci değişkeni için "Hususi" düzeyleri referans düzeylerdir.

GDKEM'de zaman noktaları arasında yapısal olmayan, AR(1), Toeplitz ve Gauss varyans-kovaryans yapıları ayrı ayrı hesaba katılmıştır. Yapısal olmayan, AR(1), Toeplitz ve Gauss varyans-kovaryans yapıları kullanılarak elde edilen model sonuçları ve parametre tahminleri sırasıyla Çizelge 4.7, Çizelge 4.8, Çizelge 4.9 ve Çizelge 4.10'da verilmiştir:

Çizelge 4.7. Yapısal olmayan varyans-kovaryans yapısı kullanılarak elde edilen GDKEM sonuçları

Rastgele etkiler:				
Parametre	**Sabit terim**	**yil**	**Hata terimi**	
Varyans	0,3946	0,0060	0,1423	
Standart sapma	0,6282	0,0776	0,3773	
Korelasyon		-0,53		
Sabit etkiler:				
Parametre	**Parametre tahmini**	**Standart hata**	**Z değeri**	**p değeri**
Sabit terim	-2,2042	0,0731	-30,17	<0,001
yil	-0,0991	0,0025	-39,93	<0,001
cinsiKamyonet	0,3857	0,0573	6,73	<0,001
cinsiMinibüs	0,0242	0,0601	0,40	0,687
cinsiMotosiklet	0,5618	0,0839	6,69	<0,001
cinsiOtobüs	0,1206	0,0878	1,37	0,169
cinsiOtomobil	0,2975	0,0564	5,27	<0,001
cinsiÖzelAmaçlı	0,1877	0,1028	1,83	0,068
cinsiOtomobil	0	.	.	.
yakit_turuDizel	0,2600	0,0586	4,44	<0,001
yakit_turuLpg	0,3228	0,0513	6,29	<0,001
yakit_turuBenzinli	0	.	.	.
kullanim_amaciResmi	0,3594	0,0560	6,42	<0,001
kullanim_amaciTicari	0,4615	0,0387	11,93	<0,001
kullanim_amaciHususi	0	.	.	.

Çizelge 4.8. AR(1) varyans-kovaryans yapısı kullanılarak elde edilen GDKEM sonuçları

Rastgele etkiler:				
Parametre	**yil**	**Hata terimi**		
Varyans	0,3567	0,1287		
Standart sapma	0,5972	0,3587		
Korelasyon	0,90			
Sabit etkiler:				
Parametre	**Parametre tahmini**	**Standart hata**	**Z değeri**	**p değeri**
Sabit terim	-2,2257	0,0741	-30,05	<0,001
yil	-0,0986	0,0022	-45,46	<0,001
cinsiKamyonet	0,3968	0,0584	6,80	<0,001
cinsiMinibüs	0,0338	0,0612	0,55	0,580
cinsiMotosiklet	0,5787	0,0853	6,78	<0,001
cinsiOtobüs	0,1356	0,0888	1,53	0,127
cinsiOtomobil	0,3007	0,0575	5,23	<0,001
cinsiÖzelAmaçlı	0,1816	0,1046	1,74	0,083
cinsiOtomobil	0	.	.	.
yakit_turuDizel	0,2706	0,0596	4,54	<0,001
yakit_turuLpg	0,3364	0,0518	6,49	<0,001
yakit_turuBenzinli	0	.	.	.
kullanim_amaciResmi	0,3837	0,0568	6,75	<0,001
kullanim_amaciTicari	0,4834	0,0390	12,41	<0,001
kullanim_amaciHususi	0	.	.	.

Çizelge 4.9. Toeplitz varyans-kovaryans yapısı kullanılarak elde edilen GDKEM sonuçları

Rastgele etkiler:				
Parametre	**Sabit terim**	**yil**	**Hata terimi**	
Varyans	0,3946	0,0060	0,1423	
Standart sapma	0,6282	0,0776	0,3773	
Korelasyon		-0,53		
Sabit etkiler:				
Parametre	**Parametre tahmini**	**Standart hata**	**Z değeri**	**p değeri**
Sabit terim	-2,2042	0,0731	-30,17	<0,001
yil	-0,0991	0,0025	-39,93	<0,001
cinsiKamyonet	0,3857	0,0573	6,73	<0,001
cinsiMinibüs	0,0242	0,0601	0,40	0,687
cinsiMotosiklet	0,5618	0,0839	6,69	<0,001
cinsiOtobüs	0,1206	0,0878	1,37	0,169
cinsiOtomobil	0,2975	0,0564	5,27	<0,001
cinsiÖzelAmaçlı	0,1877	0,1028	1,83	0,068
cinsiOtomobil	0	.	.	.
yakit_turuDizel	0,2600	0,0586	4,44	<0,001
yakit_turuLpg	0,3228	0,0513	6,29	<0,001
yakit_turuBenzinli	0	.	.	.
kullanim_amaciResmi	0,3594	0,0560	6,42	<0,001
kullanim_amaciTicari	0,4615	0,0387	11,93	<0,001
kullanim_amaciHususi	0	.	.	.

Çizelge 4.10. Gauss varyans-kovaryans yapısı kullanılarak elde edilen GDKEM sonuçları

Rastgele etkiler:				
Parametre	**yil**	**Hata terimi**		
Varyans	0,3474	0,1374		
Standart sapma	0,5894	0,3707		
Korelasyon	0,96			
Sabit etkiler:				
Parametre	**Parametre tahmini**	**Standart hata**	**Z değeri**	**p değeri**
Sabit terim	-2,1898	0,0730	-30,00	<0,001
yil	-0,0986	0,0024	-40,33	<0,001
cinsiKamyonet	0,3942	0,0573	6,88	<0,001
cinsiMinibüs	0,0407	0,0601	0,68	0,498
cinsiMotosiklet	0,5600	0,0838	6,68	<0,001
cinsiOtobüs	0,1298	0,0876	1,48	0,139
cinsiOtomobil	0,2900	0,0564	5,15	<0,001
cinsiÖzelAmaçlı	0,1710	0,1029	1,66	0,097
cinsiOtomobil	0	.	.	.
yakit_turuDizel	0,2473	0,0587	4,21	<0,001
yakit_turuLpg	0,3156	0,0516	6,12	<0,001
yakit_turuBenzinli	0	.	.	.
kullanim_amaciResmi	0,4050	0,0558	7,26	<0,001
kullanim_amaciTicari	0,5002	0,0386	12,97	<0,001
kullanim_amaciHususi	0	.	.	.

Çizelge 4.11'de farklı varyans-kovaryans yapılarına göre uygulanan GDKEM'lere ilişkin model performans değerlendirme ölçüleri verilmiştir:

Çizelge 4.11. Farklı varyans-kovaryans yapılarına göre uygulanan GDKEM'ler için model performans değerlendirme ölçüleri

Kovaryans yapıları	**Model performans göstergeleri**				
	AIC	**BIC**	**HKO**	**HKOK**	**OMH**
Yapısal olmayan	-19273,2	-19161,8	0,00307	0,05542	0,03460
AR(1)	-19059,6	-18955,1	**0,00242**	**0,04923**	**0,02947**
Toeplitz	-19273,2	-19161,8	0,00307	0,05542	0,03460
Gauss	-19269,3	-19164,8	0,00298	0,05458	0,03345

Çizelge 4.11'de, AR(1) varyans-kovaryans yapısına sahip GDKEM'nin diğer GDKEM'lere göre daha düşük HKO, HKOK ve OMH değerlerine sahip olduğu görülmektedir. İncelenen GDKEM'ler arasında, AR(1) varyans-kovaryans yapısına sahip GDKEM HKO, HKOK ve OMH model performans değerlendirme ölçütlerine göre en iyi modeldir. Bu modelde yil, cinsiKamyonet, cinsiMotosiklet, cinsiOtomobil, yakit_turuDizel, yakit_turuLpg, kullanim_amaciResmi ve kullanim_amaciTicari değişkenleri 0,05 anlamlılık düzeyinde istatistiksel olarak anlamlı bulunmuş, cinsiMinibüs, cinsiOtobüs ve cinsiÖzelAmaçlı değişkenleri ise istatistiksel olarak anlamlı bulunmamıştır. Çizelge 4.8'de verilen parametre tahminlerine göre yil değişkeni için $\beta_1 = -0{,}0986$ ($p<0{,}001$)'dır. Buna göre zaman ilerledikçe kmFark azalmaktadır.

AR(1) varyans-kovaryans yapısına sahip GDKEM ile cinsi, yakit_turu ve kullanim_amaci değişkenlerinin düzeyleri için hesaplanan odds oranları Çizelge 4.12'de verilmiştir:

Çizelge 4.12. AR(1) varyans-kovaryans yapısına sahip GDKEM için odds oranları

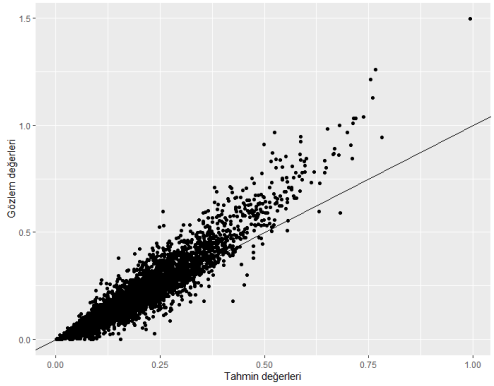
Karşılaştırma	**Odds oranı**	**Standart hata**	**Z değeri**	**p değeri**
cinsi				
Kamyon / Kamyonet	0,672	0,039	-6,80	<0,001
Kamyon / Minibüs	0,967	0,059	-0,55	0,998
Kamyon / Motosiklet	0,561	0,048	-6,79	<0,001
Kamyon / Otobüs	0,873	0,078	-1,53	0,729
Kamyon / Otomobil	0,740	0,043	-5,23	<0,001

Çizelge 4.12. AR(1) varyans-kovaryans yapısına sahip GDKEM için odds oranları (devamı)

Karşılaştırma	Odds oranı	Standart hata	Z değeri	p değeri
Kamyon / ÖzelAmaçlı	0,834	0,087	-1,74	0,592
Kamyonet / Minibüs	1,438	0,083	6,28	<0,001
Kamyonet / Motosiklet	0,834	0,067	-2,28	0,256
Kamyonet / Otobüs	1,299	0,119	2,85	0,066
Kamyonet / Otomobil	1,101	0,058	1,83	0,531
Kamyonet / ÖzelAmaçlı	1,240	0,133	2,01	0,407
Minibüs / Motosiklet	0,580	0,050	-6,32	<0,001
Minibüs / Otobüs	0,903	0,081	-1,13	0,920
Minibüs / Otomobil	0,766	0,046	-4,42	<0,001
Minibüs / ÖzelAmaçlı	0,863	0,093	-1,37	0,819
Motosiklet / Otobüs	1,558	0,173	4,00	<0,001
Motosiklet / Otomobil	1,320	0,084	4,39	<0,001
Motosiklet / ÖzelAmaçlı	1,488	0,185	3,20	0,024
Otobüs / Otomobil	0,848	0,077	-1,82	0,537
Otobüs / ÖzelAmaçlı	0,955	0,119	-0,37	0,998
Otomobil / ÖzelAmaçlı	1,127	0,119	1,13	0,920
yakit_turu				
Benzinli / Dizel	0,763	0,045	-4,54	<0,001
Benzinli / Lpg	0,714	0,037	-6,49	<0,001
Dizel / Lpg	0,936	0,048	-1,29	0,399
kullanim_amaci				
Hususi / Resmi	0,681	0,039	-6,75	<0,001
Hususi / Ticari	0,617	0,024	-12,41	<0,001
Resmi / Ticari	0,905	0,053	-1,70	0,205

Çizelge 4.12'ye göre; motosikletlerin kamyonlara göre kmFark değerleri 1,8 ($p<0,001$), minibüslerin motosikletlere göre 1,7 ($p<0,001$) ve kamyonetlerin minibüslere göre 1,4 ($p<0,001$) kat daha fazladır. Ayrıca, Lpg yakıtlı araçların benzin yakıtlı araçlara göre kmFark değerleri 1,4 kat ve Ticari araçların Hususi araçlara göre kmFark değerleri 1,6 ($p<0,001$) kat daha fazladır.

Şekil 4.6'da, AR(1) varyans-kovaryans yapısına sahip GDKEM ile elde edilen tahmin değerleri ve gözlem değerlerine ilişkin saçılım grafiği verilmiştir. Şekil 4.6'da AR(1) kovaryans yapısı kullanılan GDKEM ile elde edilen tahmin değerlerinin, gözlem değerleri ile büyük ölçüde paralel olduğu görülmektedir.



Şekil 4.6. AR(1) varyans-kovaryans yapısına sahip GDKEM ile elde edilen tahmin değerleri ve gözlem değerlerine ilişkin saçılım grafiği

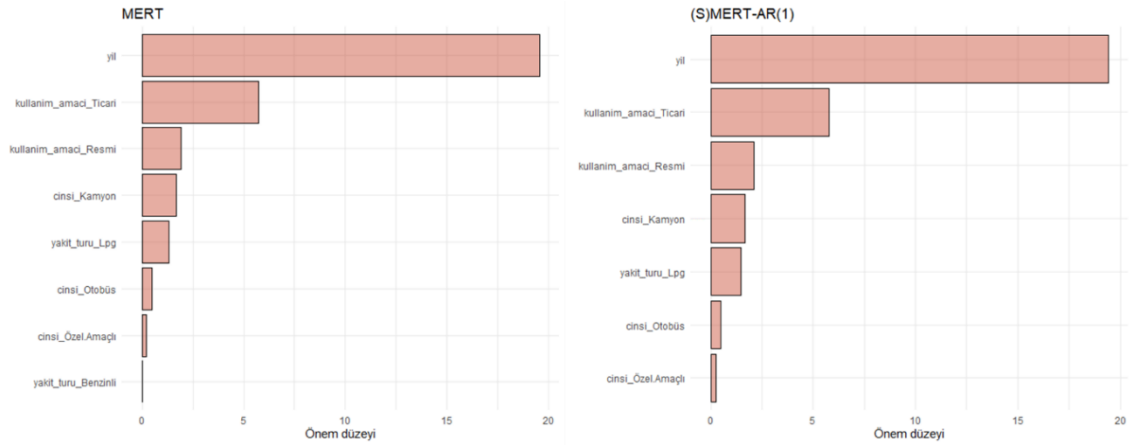
4.1.2. Karma Etkili Regresyon Ağaçları ve Rastgele Orman Algoritmalarına Göre Analiz Sonuçları

Çalışmada uzun süreli veri kümesi için karma etkili (S)MERT, (S)MERF, (S)RE-EM Ağacı ve (S)RE-EM Ormanı makine öğrenmesi algoritmaları uygulanmıştır. Analizlerde yapısal olmayan ve AR(1) varyans-kovaryans yapıları düşünülmüştür. Analizler R programında "*LongituRF"* [52] paketi kullanılarak yapılmıştır.

4.1.2.1. MERT ve (S)MERT Algoritmaları Kullanılarak Elde Edilen Analiz Sonuçları

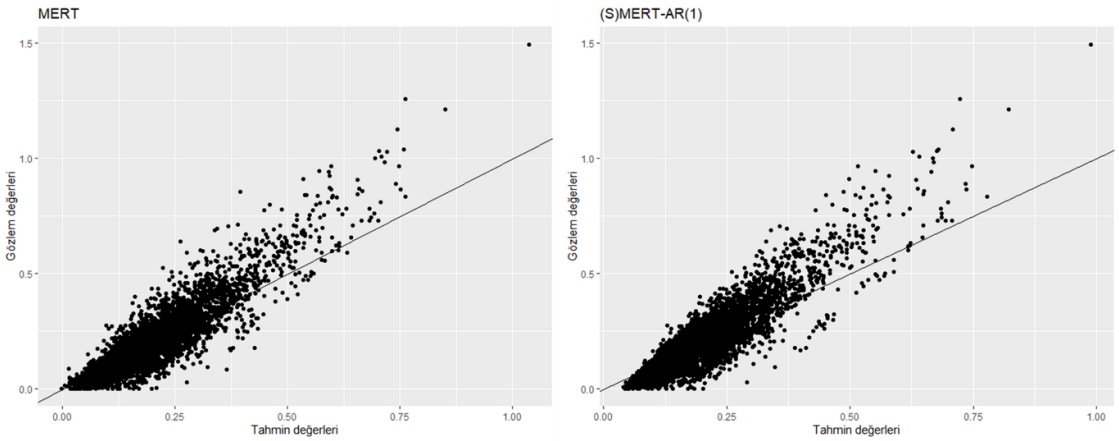
MERT algoritması AR(1) varyans-kovaryans yapısı dikkate alınarak "(S)MERT" ve herhangi özel bir kovaryans yapısı dikkate alınmadan "MERT" şeklinde adlandırılarak iki farklı şekilde uygulanmıştır.

Şekil 4.7'de MERT ve (S)MERT modellerinde açıklayıcı değişkenlerin önem düzeyleri her iki model için verilmiştir. Şekil 4.7'de MERT ve (S)MERF modellerinde yil değişkeninin en yüksek öneme sahip olduğu görülmektedir.



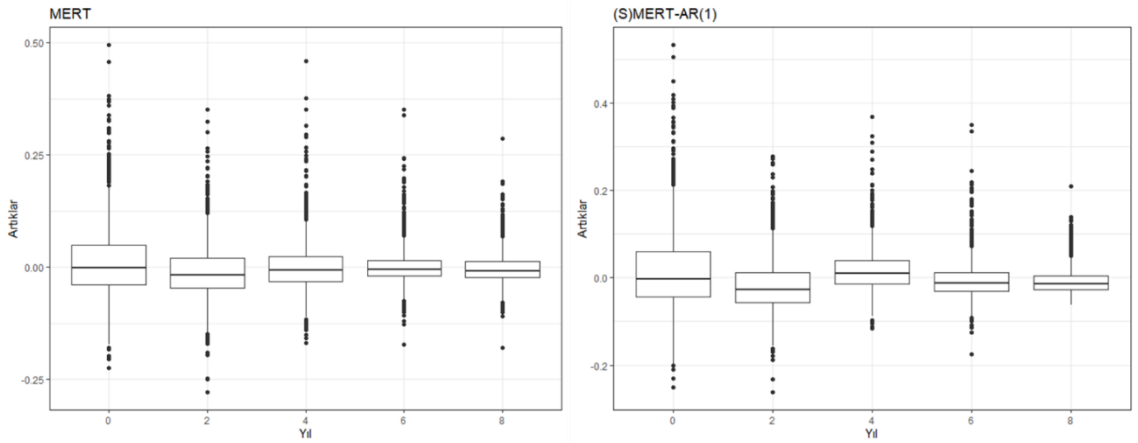
Şekil 4.7. MERT ve (S)MERT modelleri için açıklayıcı değişken önem düzeyleri

Şekil 4.8'de MERT ve (S)MERT modelleri için tahmin ve gözlem değerlerine ilişkin saçılım grafikleri verilmiştir. Şekil 4.8'de hem MERT hem de (S)MERT modellerinin tahmin performanslarının her iki model için de benzer olduğu görülmektedir.



Şekil 4.8. MERT ve (S)MERT modelleri için tahmin edilen ve gözlem değerlerine ilişkin saçılım grafikleri

Şekil 4.9'da MERT ve (S)MERT modellerine göre artıkların yıllar bakımından box-plot grafikleri verilmiştir:



Şekil 4.9. MERT ve (S)MERT modellerine göre artıkların yıllar bakımından box-plot grafikleri

Şekil 4.9'da iki model ile ortaya çıkan artıkların farklı zaman noktalarında benzer dağılımlara sahip olduğu görülmektedir.

MERT ve (S)MERT modellerine ilişkin model performans değerlendirme ölçütleri Çizelge 4.13'te verilmiştir:

Çizelge 4.13. MERT ve (S)MERT modellerine ilişkin performans değerlendirme ölçüleri

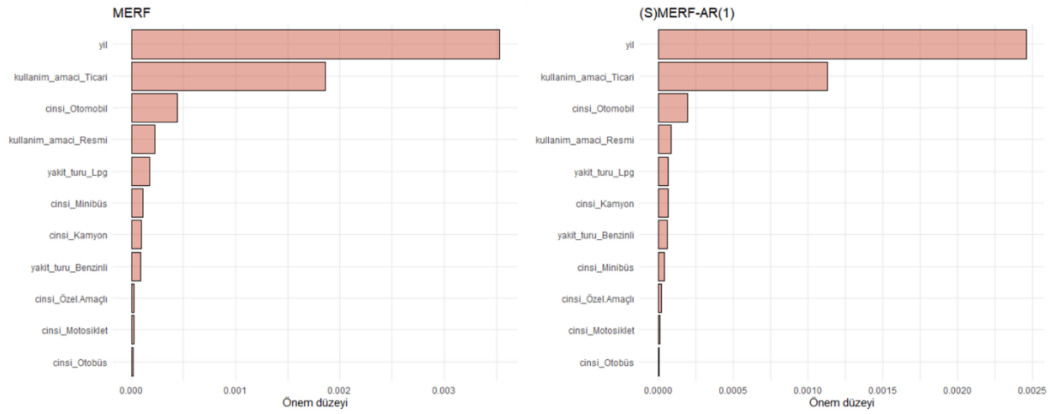
Performans ölçütü / Model	**MERT**	**(S)MERT**
HKO	**0,00348**	0,00363
HKOK	**0,05900**	0,06024
OMH	**0,04029**	0,04101

Çizelge 4.13'te MERT ve (S)MERT modelleri için verilen performans değerlendirme ölçüleri incelendiğinde, yapısal olmayan varyans-kovaryans yapısına sahip MERT modelinin AR(1) varyans-kovaryans yapısına sahip (S)MERT modeline göre daha küçük HKO, HKOK ve OMH değerlerine sahip olduğu görülmektedir. MERT modelinin (S)MERT modeline göre daha iyi olduğu söylenebilir.

4.1.2.2. MERF ve (S)MERF Algoritmaları Kullanılarak Elde Edilen Analiz Sonuçları

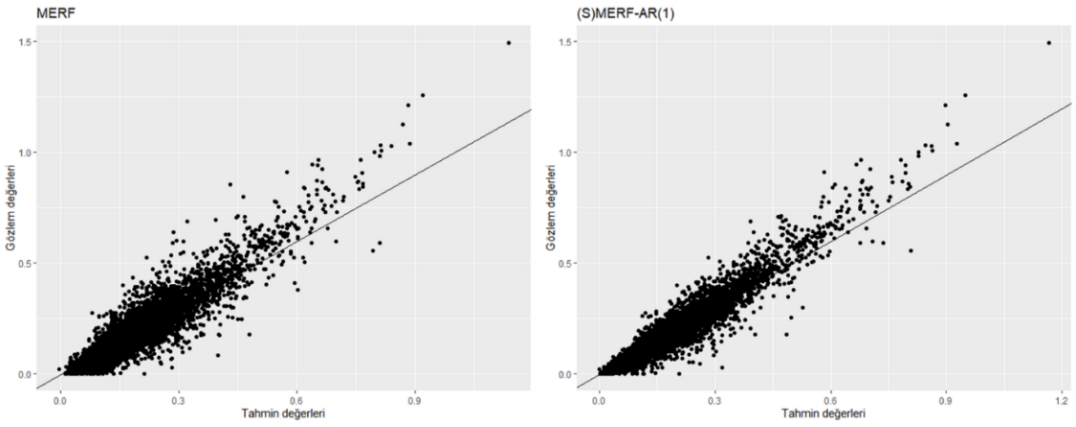
Çalışmada MERF algoritması AR(1) varyans-kovaryans yapısı dikkate alınarak "(S)MERF" ve herhangi özel bir varyans-kovaryans dikkate alınmadan "MERF" şeklinde adlandırılarak iki şekilde uygulanmıştır.

Şekil 4.10'da açıklayıcı değişkenlerin önem düzeyleri her iki model için verilmiştir. Şekil 4.10'da hem MERF hem de (S)MERF modellerinde yil açıklayıcı değişkeninin en yüksek önem düzeyine sahip olduğu görülmektedir.



Şekil 4.10. MERF ve (S)MERF modelleri için açıklayıcı değişken önem düzeyleri

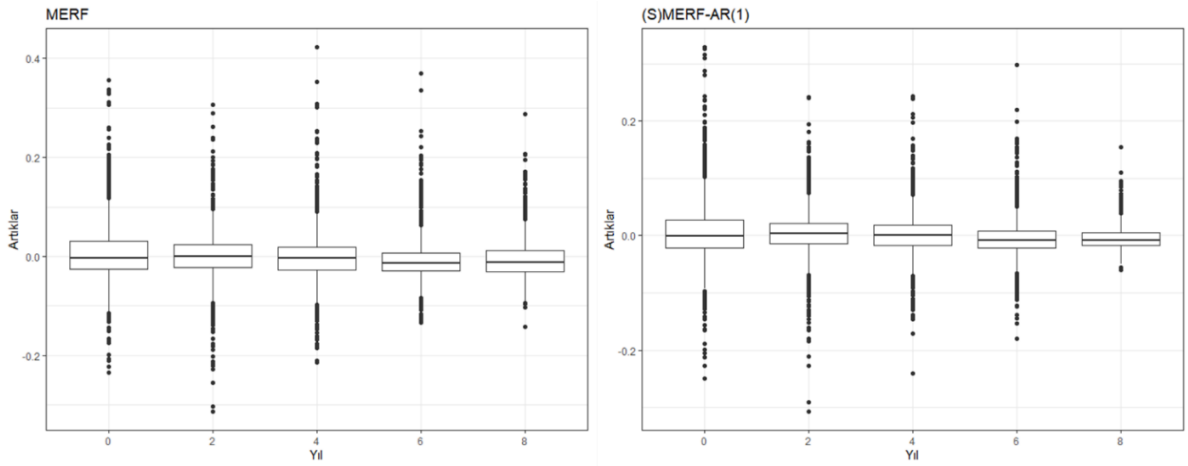
MERF ve (S)MERF modelleri ile elde edilen tahmin değerleri ve gözlem değerlerine ilişkin saçılım grafiği Şekil 4.11'de verilmiştir:



Şekil 4.11. MERF ve (S)MERF modelleri için tahmin ve gözlem değerlerine ilişkin saçılım grafikleri

Şekil 4.11’de, (S)MERF modelinin MERF modeline göre tahmin performansının daha iyi olduğu görülmektedir.

Şekil 4.12’de MERF ve (S)MERF için artıkların yıllara göre box-plot grafikleri verilmiştir:



Şekil 4.12. MERF ve (S)MERF modelleri ile ortaya çıkan artıkların yıllara göre box-plot grafikleri

Şekil 4.12’ye göre, her iki model için de artıklar benzer bir trende sahiptir.

MERF ve (S)MERF modellerine ilişkin model performans ölçütleri Çizelge 4.14’te verilmiştir:

Çizelge 4.14. MERF ve (S)MERF modellerine ilişkin performans değerlendirme ölçüleri

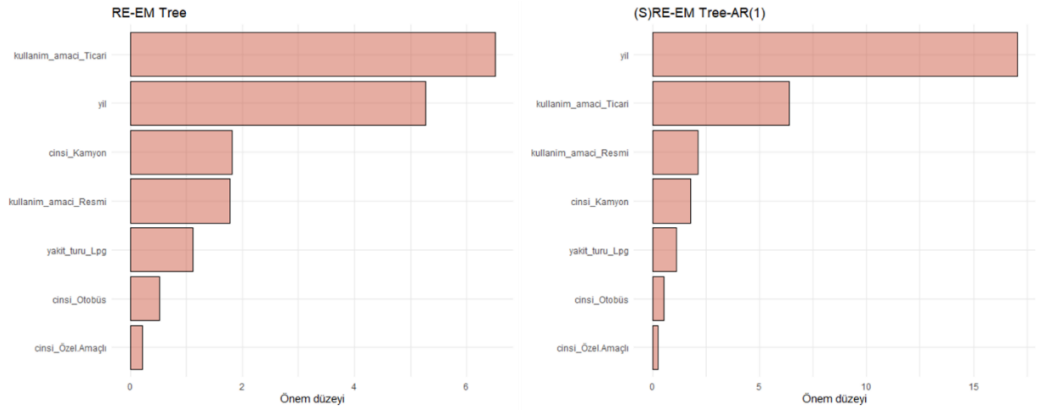
Performans ölçütü / Model	**MERF**	**(S)MERF**
HKO	0,00268	**0,00169**
HKOK	0,05175	**0,04112**
OMH	0,03542	**0,02681**

Çizelge 4.14’te MERF ve (S)MERF için elde edilen performans değerlendirme ölçüleri incelendiğinde, (S)MMERF modelinin MERF modeline göre daha düşük HKO, HKOK ve OMH değerlerine sahip olduğu görülmektedir. Buna göre (S)MERF modeli MERF modeline göre daha iyi bir modeldir.

4.1.2.3. RE-EM ve (S)RE-EM Ağacı Algoritmaları Kullanılarak Elde Edilen Analiz Sonuçları

RE-EM Ağacı algoritması AR(1) varyans-kovaryans yapısı dikkate alınarak "(S)RE-EM Ağacı" ve herhangi özel bir varyans-kovaryans yapısı dikkate alınmadan "RE-EM Ağacı" şeklinde adlandırılarak iki şekilde uygulanmıştır.

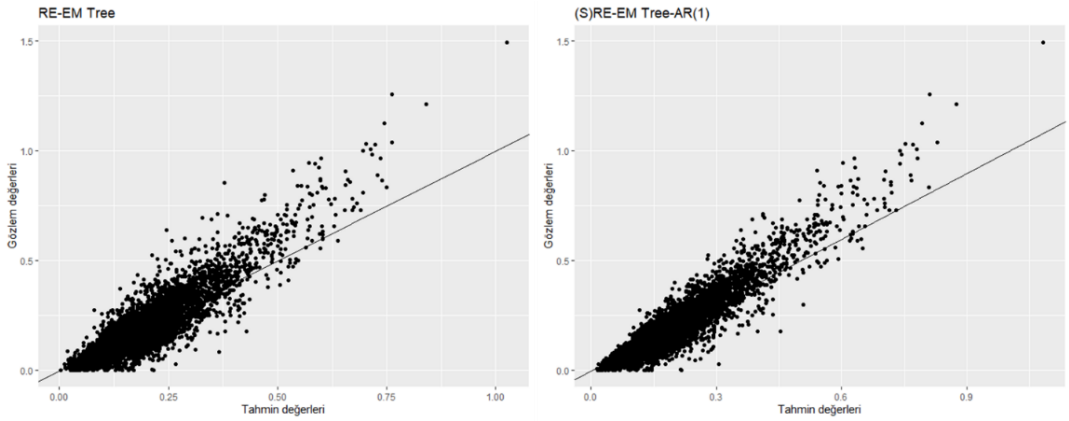
Şekil 4.13'te RE-EM Ağacı ve (S)RE-EM Ağacı modellerinde yer alan açıklayıcı değişkenlerin önem düzeyleri verilmiştir:



Şekil 4.13. RE-EM Ağacı ve (S)RE-EM Ağacı modelleri için açıklayıcı değişken önem düzeyleri

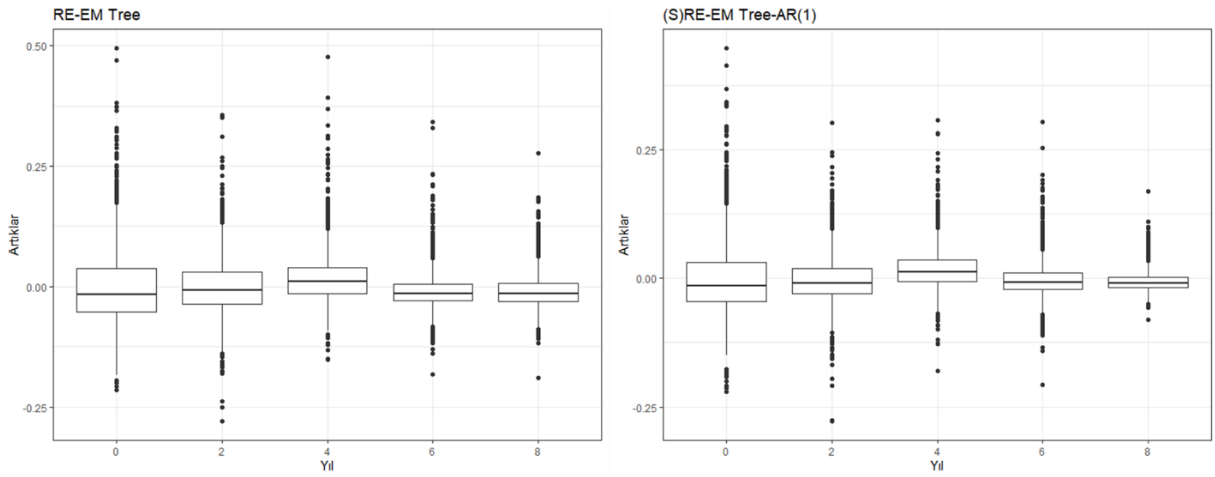
Şekil 4.13'te, RE-EM Ağacı modelinde kullanim_amaci açıklayıcı değişkeninin "Ticari" düzeyinin en yüksek öneme sahip olduğu görülürken, (S)RE-EM Ağacı modelinde ise yil değişkeninin en yüksek öneme sahip olduğu görülmektedir.

Şekil 4.14'te RE-EM Ağacı ve (S)RE-EM Ağacı için tahmin ve gözlem değerlerine ilişkin saçılım grafikleri verilmiştir. (S)RE-EM Ağacı modelinin RE-EM Ağacı modeline göre tahmin performansı açısından daha iyi sonuç verdiği görülmektedir.



Şekil 4.14. RE-EM Ağacı ve (S)RE-EM Ağacı modelleri için tahmin edilen ve gözlem değerlerine ilişkin saçılım grafikleri

Şekil 4.15'te RE-EM Ağacı ve (S)RE-EM Ağacı modellerine göre artıkların yıllara göre box-plot grafikleri verilmiştir:



Şekil 4.15. RE-EM Ağacı ve (S)RE-EM Ağacı modelleri ile ortaya çıkan artıkların yıllara göre box-plot grafikleri

Şekil 4.15'te RE-EM Ağacı ve (S)RE-EM Ağacı modelleri ile elde edilen tahmin değerlerinden kaynaklanan artıkların yıllara göre dağılımı incelendiğinde her iki modelin benzer sonuçlar verdiği görülmektedir. Her iki modelde de 4 numaralı zaman noktasında ortaya çıkan artıkların diğer zaman noktalarından farklılık gösterdiği görülmektedir.

RE-EM Ağacı ve (S)RE-EM Ağacı modellerine ilişkin elde edilen model performans ölçütleri Çizelge 4.15'te verilmiştir:

Çizelge 4.15. RE-EM Ağacı ve (S)RE-EM Ağacı modellerine ilişkin performans değerleri

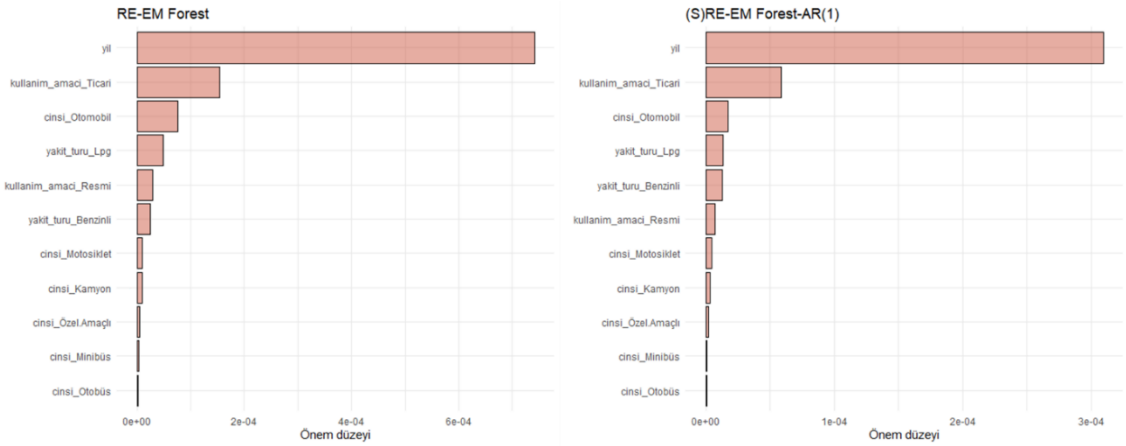
Performans ölçütü / Model	**RE-EM Ağacı**	**(S)RE-EM Ağacı**
HKO	0,00351	**0,00240**
HKOK	0,05926	**0,04896**
OMH	0,04099	**0,03246**

Çizelge 4.15'te RE-EM Ağacı ve (S)RE-EM Ağacı modelleri için verilen performans değerlendirme ölçüleri incelendiğinde (S)RE-EM Ağacı modeli RE-EM Ağacı modeline göre daha düşük HKO, HKOK ve OMH değerlerine sahiptir. Buna göre (S)RE-EM Ağacı modeli, RE-EM Ağacı modeline göre daha iyi modeldir.

4.1.2.4. RE-EM ve (S)RE-EM Ormanı Algoritmaları Kullanılarak Elde Edilen Analiz Sonuçları

RE-EM Ormanı algoritması AR(1) varyans-kovaryans yapısı dikkate alınarak "(S)RE-EM Ormanı" ve herhangi özel bir varyans-kovaryans dikkate alınmadan "RE-EM Ormanı" şeklinde adlandırılarak iki şekilde uygulanmıştır.

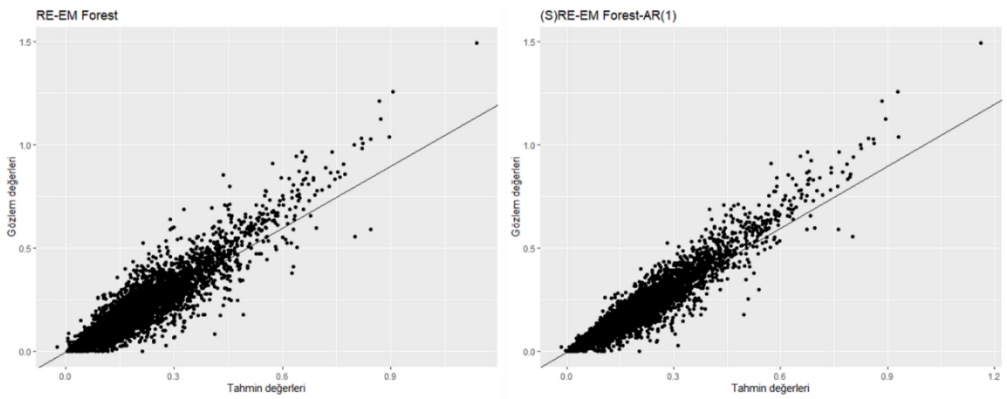
Şekil 4.16'da RE-EM Ormanı ve (S)RE-EM Ormanı modellerinde yer alan açıklayıcı değişkenlerin önem düzeyleri verilmiştir:



Şekil 4.16. RE-EM Ormanı ve (S)RE-EM Ormanı modelleri için açıklayıcı değişken önem düzeyleri

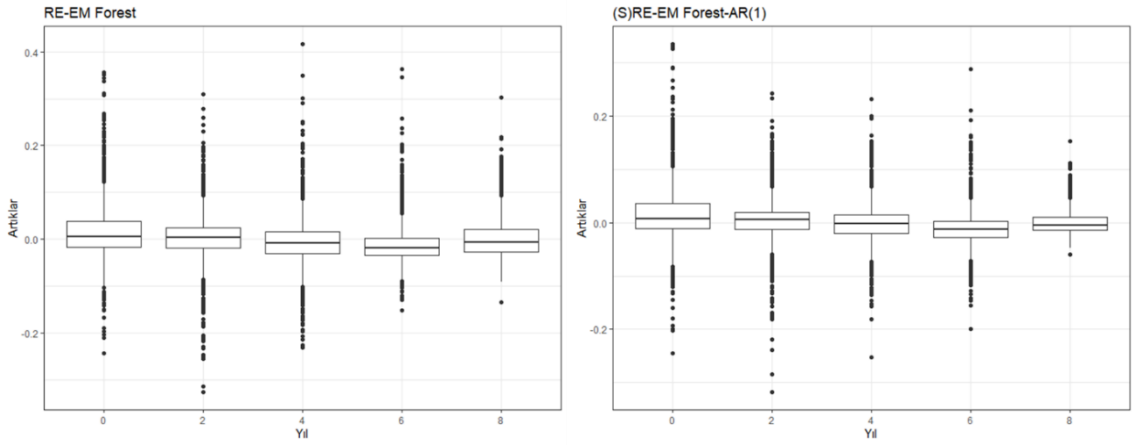
Şekil 4.16'da yil değişkeninin hem RE-EM Ormanı hem de (S)RE-EM Ormanı modelinde en yüksek öneme sahip olduğu görülmektedir.

Şekil 4.17'de RE-EM Ormanı ve (S)RE-EM Ormanı modelleri için tahmin edilen ve gözlem değerlerine ilişkin saçılım grafikleri verilmiştir:



Şekil 4.17. RE-EM Ormanı ve (S)RE-EM Ormanı modelleri için tahmin edilen ve gözlem değerlerine ilişkin saçılım grafikleri

Şekil 4.17'de (S)RE-EM Ormanı modelinin RE-EM Ormanı modeline göre daha iyi tahmin performansına sahip olduğu görülmektedir.



Şekil 4.18. RE-EM Ormanı ve (S)RE-EM Ormanı modelleri için artıkların yıllara göre box-plot grafikleri

Şekil 4.18'de RE-EM Ormanı ve (S)RE-EM Ormanı modelleri ile elde edilen tahmin değerlerinden kaynaklanan artıkların zaman noktalarına göre dağılımı incelendiğinde, her iki modelde de artıkların benzer bir trende sahip olduğu görülmektedir.

RE-EM Ormanı ve (S)RE-EM Ormanı modellerine ilişkin model performans ölçütleri Çizelge 4.16'da verilmiştir:

Çizelge 4.16. RE-EM Ormanı ve (S)RE-EM Ormanı modellerine ilişkin performans değerleri

Performans ölçütü / Model	**RE-EM Ormanı**	**(S)RE-EM Ormanı**
HKO	0,00280	**0,00177**
HKOK	0,05273	**0,04205**
OMH	0,03605	**0,02729**

Çizelge 4.16'da RE-EM Ormanı ve (S)RE-EM Ormanı modelleri için verilen performans değerlendirme ölçüleri dikkate alındığında, (S)RE-EM Ormanı modelinin RE-EM Ormanı modeline göre daha iyi olduğu görülmektedir.

4.1.3. Gauss Süreci Güçlendirme (GP Boosting) Algoritmasına Göre Analiz Sonuçları

Bu bölümde uzun süreli veriler üzerinde GP Boosting algoritmaları uygulanmıştır. Analizler R programında "*gpboost"* [53] paketi kullanılarak yapılmıştır. Analizlerde, Gauss ve AR(1) varyans-kovaryans yapıları düşünülmüş ve gamma dağılım fonksiyonu kullanılmıştır.

Çizelge 4.17 ve Çizelge 4.18'de GP Boosting algoritması sırasıyla AR(1) ve Gauss varyans-kovaryans yapıları kullanılarak elde edilen parametre tahminleri verilmiştir:

Çizelge 4.17. AR(1) varyans-kovaryans yapısına göre GP Boosting parametre tahminleri

Parametre	**Parametre tahmini**	**Standart sapma**	**Z değeri**	**p değeri**
Sabit terim	-2,1975	0,0740	-29,68	<0,001
yil	-0,0986	0,0022	-45,48	<0,001
cinsiKamyonet	0,3840	0,0583	6,58	<0,001
cinsiMinibüs	0,0228	0,0612	0,37	0,710
cinsiMotosiklet	0,5475	0,0853	6,42	<0,001
cinsiOtobüs	0,1263	0,0888	1,42	0,155

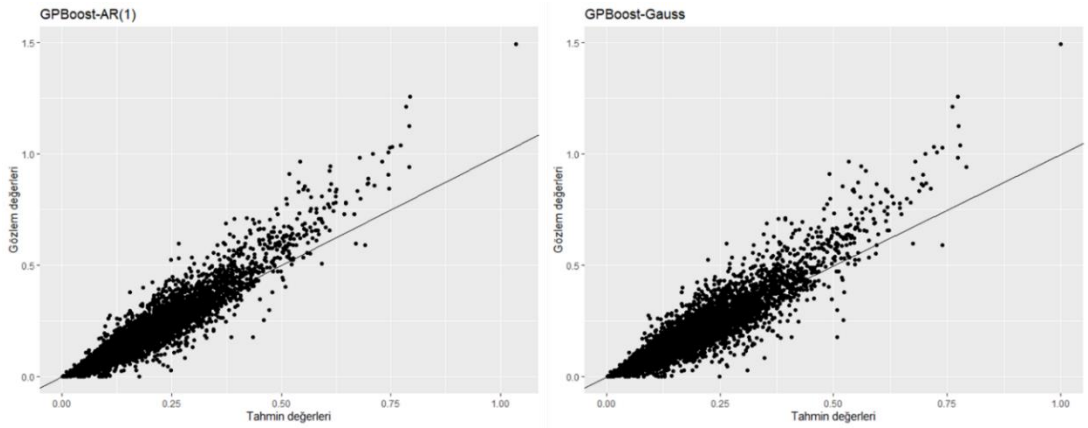
Çizelge 4.17. AR(1) varyans-kovaryans yapısına göre GP Boosting parametre tahminleri (devamı)

Parametre	Parametre tahmini	Standart sapma	Z değeri	p değeri
cinsiOtomobil	0,2816	0,0575	4,90	<0,001
cinsiÖzelAmaçlı	0,1726	0,1046	1,65	0,099
cinsiKamyon	0	.	.	.
yakit_turuDizel	0,2544	0,0595	4,27	<0,001
yakit_turuLpg	0,3281	0,0517	6,34	<0,001
yakit_turuBenzinli	0	.	.	.
kullanim_amaciResmi	0,3809	0,0568	6,70	<0,001
kullanim_amaciTicari	0,4789	0,0389	12,31	<0,001
kullanim_amaciHususi	0	.	.	.

Çizelge 4.18. Gauss varyans-kovaryans yapısına göre GP Boosting parametre tahminleri

Parametre	Parametre tahmini	Standart sapma	Z değeri	p değeri
Sabit terim	-2,1614	0,0730	-29,61	<0,001
yil	-0,0986	0,0024	-40,35	<0,001
cinsiKamyonet	0,3814	0,0572	6,66	<0,001
cinsiMinibüs	0,0298	0,0600	0,50	0,620
cinsiMotosiklet	0,5286	0,0838	6,31	<0,001
cinsiOtobüs	0,1207	0,0876	1,38	0,168
cinsiOtomobil	0,2710	0,0564	4,81	<0,001
cinsiÖzelAmaçlı	0,1622	0,1029	1,58	0,115
cinsiKamyon	0	.	.	.
yakit_turuDizel	0,2308	0,0587	3,93	<0,001
yakit_turuLpg	0,3068	0,0515	5,96	<0,001
yakit_turu_Benzinli	0	.	.	.
kullanim_amaciResmi	0,4021	0,0557	7,21	<0,001
kullanim_amaciTicari	0,4958	0,0385	12,89	<0,001
kullanim_amaciHususi	0	.	.	.

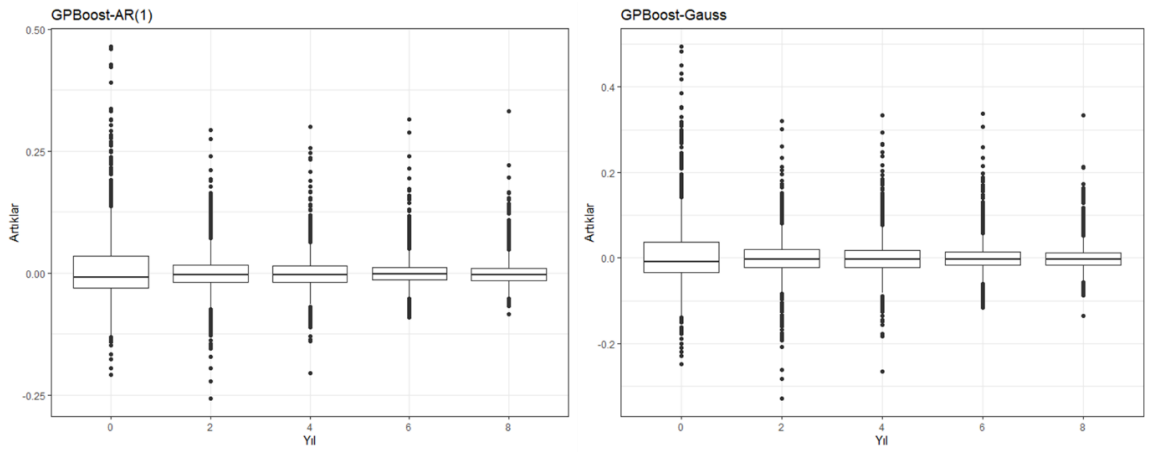
Şekil 4.19'da GP Boosting algoritması kapsamında AR(1) ve Gauss varyans-kovaryans yapıları kullanılarak elde edilen tahmin değerleri ile gözlem değerlerine ilişkin saçılım grafikleri verilmiştir:



Şekil 4.19. GP Boosting modelleri için tahmin edilen ve gözlem değerlerine ilişkin saçılım grafikleri

Şekil 4.19'da hem AR(1) hem de Gauss varyans-kovaryans yapıları ile oluşturulan modellerin benzer tahmin performansı gösterdiği görülmektedir.

Şekil 4.20'de GP Boosting modelleri ile ortaya çıkan artıkların yıllara göre box-plot grafikleri verilmiştir:



Şekil 4.20. GP Boosting modelleri için artıkların yıllara göre box-plot grafikleri

Şekil 4.20'de AR(1) ve Gauss varyans-kovaryans yapılarına sahip modellerden elde edilen artıkların dağılımlarının yıllara göre benzer olduğu görülmektedir.

GP Boosting modellerine ilişkin model performans ölçütleri Çizelge 4.19'da verilmiştir:

Çizelge 4.19. GP Boosting modellerine ilişkin performans değerleri

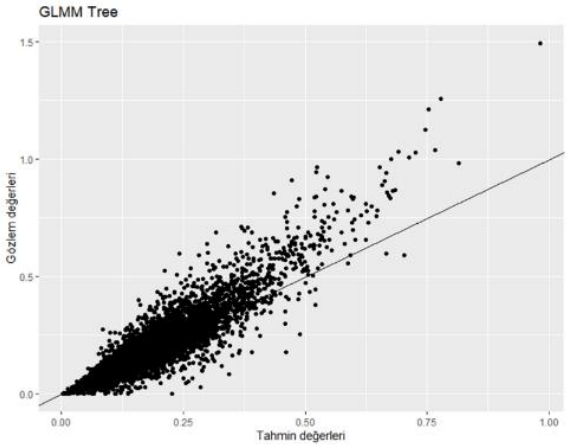
Performans ölçütü / Model	**GP Boosting-AR(1)**	**GP Boosting-Gauss**
HKO	**0,00229**	0,00285
HKOK	**0,04780**	0,05334
OMH	**0,02930**	0,03326

Çizelge 4.19'da, AR(1) varyans-kovaryans yapılı modelin Gauss varyans-kovaryans yapısına sahip modele göre daha düşük HKO, HKOK ve OMH değerlerine sahip olduğundan, AR(1) varyans-kovaryans yapılı GP Boosting modelinin daha iyi uyum sağladığı görülmektedir. Ayrıca Çizelge 4.17'de görüleceği üzere bu modelde yil, cinsiKamyonet, cinsiMotosiklet, cinsiOtomobil, yakit_turuDizel, yakit_turuLpg, kullanim_amaciResmi ve kullanim_amaciTicari değişkenleri 0,05 önem düzeyinde istatistiksel olarak anlamlı bulunmuş, cinsiMinibüs, cinsiOtobüs ve cinsiÖzelAmaçlı değişkenleri ise istatistiksel olarak anlamlı bulunmamıştır.

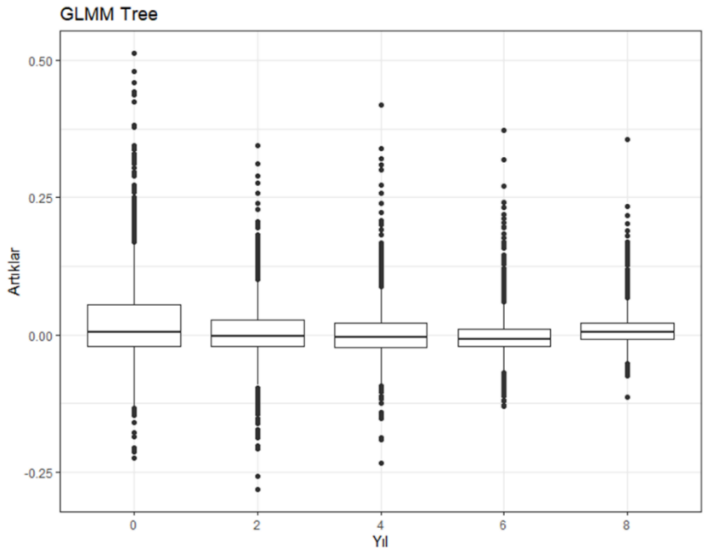
4.1.4. GDKEM Ağacı Algoritmasına Göre Analiz Sonuçları

Bu bölümde uzun süreli veri üzerinde uygulanan GDKEM Ağacı algoritması ile elde edilen analiz sonuçları verilmiştir. Analizler R programında "*glmertree"* [54] paketi kullanılarak yapılmıştır. Analizlerde logaritmik bağ fonksiyonu kullanılmıştır.

Şekil 4.21'de GDKEM Ağacı algoritması kullanılarak elde edilen tahmin değerleri ile gözlem değerlerine ilişkin saçılım grafiği, Şekil 4.22'de GDKEM Ağacı algoritması ile ortaya çıkan artıkların yıllara göre box-plot grafiği ve Şekil 4.23'te GDKEM Ağacı algoritması ile elde edilen ağaç yapısı gösterilmiştir. GDKEM Ağacı algoritması için performans değerlendirme ölçüleri ise Çizelge 4.20'de verilmiştir:

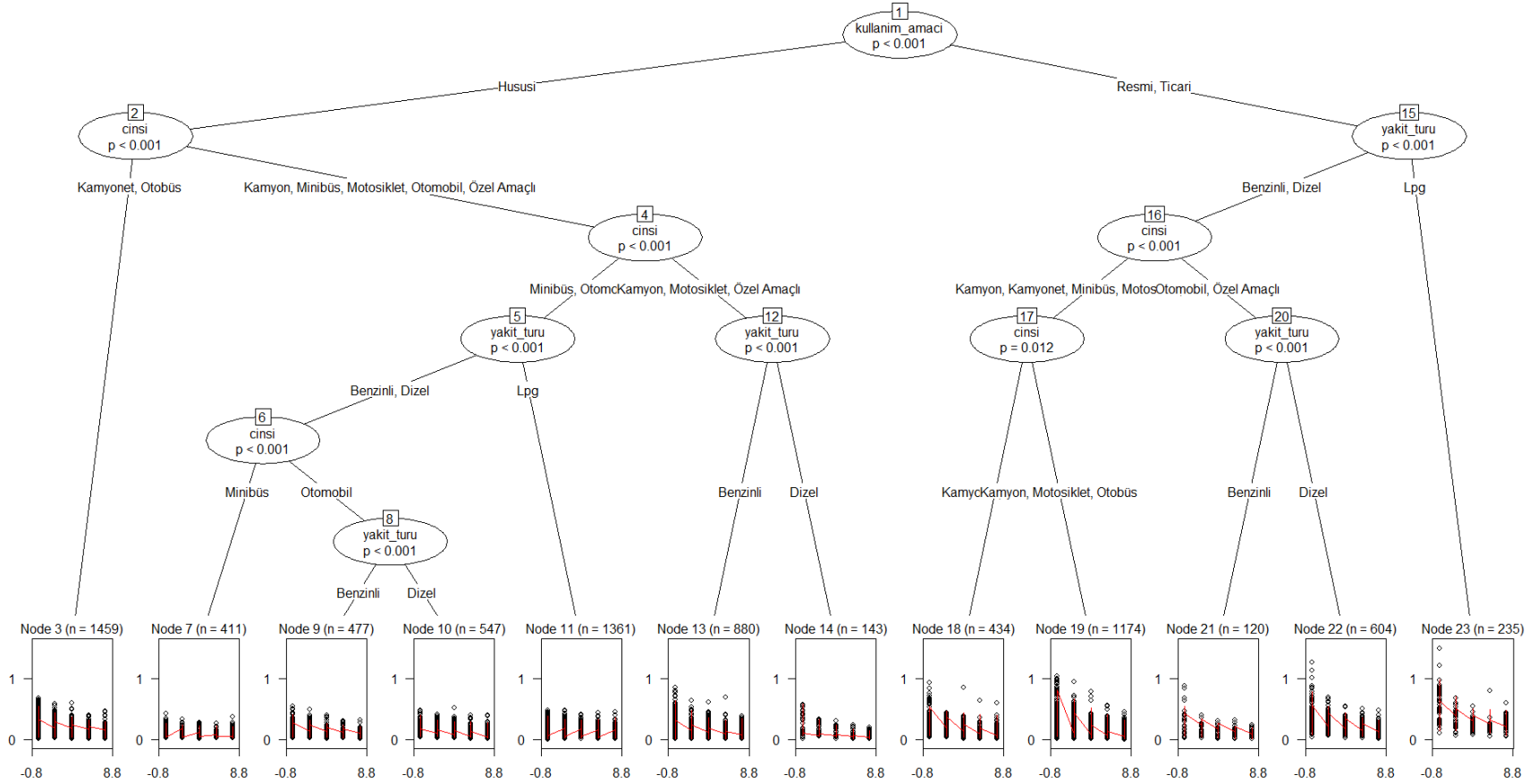


Şekil 4.21. GDKEM Ağacı modeli için tahmin edilen ve gözlem değerlerine ilişkin saçılım grafiği



Şekil 4.22. GDKEM Ağacı algoritması için artıkların yıllara göre box-plot grafiği

Şekil 4.23'te verilen GDKEM Ağacında cinsi, yakit_turu ve kullanim_amaci değişkenleri ağaç dallarının oluşturulması için kullanılmıştır. Ağaç ilk olarak kullanım_amaci kök düğümü ile başlamaktadır. Daha sonra cinsi ve yakit_turu düğümleri ile ikili bir şekilde dallanarak en son 23 numaralı terminal düğümü ile son bulur. kmFark değişkeninde yıllara göre en fazla farklılığın 18, 19, 22 ve 23 numaralı terminal düğümlerindeki taşıtlarda olduğu söylenebilir. Örneğin 23 numaralı terminal düğümünde 235 adet resmi ve ticari LPG yakıtlı taşıt bulunmakta ve bunların zaman boyunca kat ettikleri mesafeler benzer özellik göstermektedir.



Şekil 4.23. GDKEM Ağacı algoritması ile oluşturulan ağaç yapısı

Çizelge 4.20. GDKEM Ağacı modeline ilişkin performans değerleri

Performans ölçütü / Model	**GDKEM Ağacı**
HKO	0,00320
HKOK	0,05658
OMH	0,03542

4.2. Modellerin Karşılaştırılması

Bu bölümde, Bölüm 4.1'de uygulanan istatistiksel ve makine öğrenmesi yöntemlerine ilişkin model performans değerlendirme ölçüleri ve algoritmaların hesaplama süreleri verilerek en iyi model incelenmiş ve sonuçlar Çizelge 4.21'de verilmiştir:

Çizelge 4.21. İstatistiksel ve makine öğrenmesi modellerinin performans değerleri

Model / Performans Ölçütü	**HKO**	**HKOK**	**OMH**	**Süre (dk)**
(S)MERF - AR(1)	**0,00169**	**0,04112**	**0,02681**	1,500
(S)RE-EM Ormanı - AR(1)	0,00177	0,04205	0,02729	43,880
GP Boosting - AR(1)	0,00229	0,04780	0,02930	7,460
(S)RE-EM Ağacı - AR(1)	0,00240	0,04896	0,03246	0,078
GDKEM - AR(1)	0,00242	0,04923	0,02947	0,057
MERF	0,00268	0,05175	0,03542	0,653
GP Boosting - Gauss	0,00285	0,05334	0,03326	7,980
RE-EM Ormanı	0,00278	0,05273	0,03605	28,040
GDKEM - Yapısal olmayan	0,00307	0,05542	0,03460	0,045
GDKEM Ağacı	0,00320	0,05658	0,03542	1,220
MERT	0,00348	0,05900	0,04029	0,037
RE-EM Ağacı	0,00351	0,05926	0,04099	0,046
(S)MERT - AR(1)	0,00363	0,06024	0,04101	0,062

Çizelge 4.21'de, bu çalışmada incelenen uzun süreli veri kümesi için tüm modellere ilişkin HKO, HKOK, OMH değerleri hesaplama süreleri ile birlikte verilmiştir. AR(1) varyans-kovaryans yapısına sahip (S)MERF modelinin HKO, HKOK ve OMH değerlerinin sırasıyla 0,00169, 0.04112 ve 0.02681 olduğu görülmektedir. AR(1) varyans-kovaryans yapısına sahip (S)MERF modeli diğer modellere göre en düşük model performans değerlendirme ölçülerine sahiptir. Bu nedenle, çalışmada analiz edilen uzun süreli veri kümesi için uygulanan tüm modeller arasında AR(1) varyans-kovaryans yapısına sahip (S)MERF modelinin HKO, HKOK ve OMH model performans değerlendirme ölçütlerine göre en iyi model olduğu sonucuna varılmaktadır.

5. SONUÇLAR VE TARTIŞMA

Çalışmada, uzun süreli veriler için geliştirilmiş karma etkili makine öğrenmesi algoritmaları ile klasik istatistiksel yöntemler tanıtılmış, bu yöntemler gerçek hayata dair bir uzun süreli veri kümesi üzerinde uygulanmış ve söz konusu veri kümesi üzerinde model performans ölçütlerine göre en iyi sonucu veren model seçilmiştir.

Çalışmada, klasik istatistiksel yöntemlerden GDKEM, uzun süreli veriler için geliştirilen karma etkili makine öğrenmesi yöntemlerinden ise (S)MERT, (S)MERF, (S)RE-EM Ağacı, (S)RE-EM Ormanı, GDKEM Ağacı ve Gauss Süreci Güçlendirme algoritmaları detaylı bir şekilde tanıtılarak uzun süreli veri kümesi için uygulanmıştır.

Çalışmada, Türkiye'de trafiğe kayıtlı olan motorlu taşıtlara ilişkin araç muayene istasyonlarında derlenen idari kayıt verileri kullanılmıştır. 2013-2023 yılları arasında her iki yılda bir düzenli olarak muayeneye gelen 1569 adet araca ilişkin 5 farklı zaman noktasında dengeli olarak, araçların katettikleri mesafeler üzerinde yıl, araç cinsi, araçların yakıt türü ve kullanım amacı açıklayıcı değişkenlerinin etkisini inceleyebilmek amacıyla karma etki modelleri oluşturulmuştur.

İlk olarak, model oluşturma sürecinin öncesinde cevap değişkenin 5 farklı zaman noktasındaki dağılımı incelenmiştir. Yapılan analizlerde kmFark değişkeninin tüm zaman noktalarında sağa çarpık bir dağılıma sahip olduğu görülmüştür.

Cevap değişkeninin dağılımına karar verildikten sonra oluşturulan uzun süreli veri kümesi için yapısal olmayan, AR(1), Toeplitz ve Gauss gibi varyans-kovaryans yapılarını içeren farklı GDKEM'ler logaritmik bağ fonksiyonu kullanılarak uygulanmıştır. Uygulanan GDKEM'ler arasında AR(1) varyans-kovaryans yapısına sahip modelin diğer varyans-kovaryans yapılı modellere göre daha düşük HKO, HKOK ve OMH değerlerine sahip olduğu ve bu nedenle diğer modellere göre tahmin gücünün daha yüksek olduğu sonucuna ulaşılmıştır. AR(1) varyans-kovaryans yapısına sahip GDKEM için HKOK değeri 0,04923 ve OMH değeri ise 0,02947 olarak hesaplanmıştır. Bu modelde, araç cinslerinden otobüs, minibüs ve özel amaçlı olanları 0,05 anlamlılık düzeyinde istatistiksel olarak anlamlı bulunmamıştır.

Elde edilen GDKEM sonuçlarının ardından karma etkili makine öğrenmesi algoritmalarından (S)MERT, (S)MERF, (S)RE-EM Ağacı ve (S)RE-EM Ormanı algoritmaları aynı uzun süreli veri kümesi için uygulanmıştır. Bu makine öğrenmesi algoritmalar cevap değişkeninin dağılımı için herhangi bir koşul gerektirmediğinden cevap değişkenin dağılımı için herhangi bir ön tanımlı bilgi modelde yer almamıştır. Bununla birlikte algoritmalar her bir model için yapısal olmayan ve AR(1) varyans-kovaryans yapısı içerecek şekilde ayrı ayrı uygulanmıştır. Açıklayıcı değişkenlerin önem düzeyleri incelendiğinde, yapısal olmayan varyans-kovaryans yapılı MERT, MERF ve RE-EM Ormanı modelleri ile AR(1) varyans-kovaryans yapılı MERT, MERF, RE-EM Ağacı ve RE-EM Ormanı modellerinde yıl değişkenin en yüksek öneme sahip olduğu, yapısal olmayan varyans-kovaryans yapılı RE-EM Ağacı modelinde ise ticari kullanım amacı düzeyinin en yüksek öneme sahip olduğu görülmüştür. Oluşturulan modeller içerisinde AR(1) varyans-kovaryans yapısına sahip MERF modelinin diğer modellere göre daha düşük HKOK ve OMH değerlerine sahip olduğu saptanmış olup, bu değerler sırasıyla 0,04112 ve 0,02681 olarak hesaplanmıştır. Bu nedenle AR(1) varyans-kovaryans yapılı MERF modelinin söz konusu uzun süreli veri kümesi için tahmin performansının MERT, RE-EM Ağacı ve RE-EM Ormanı modellerinden daha iyi olduğu sonucuna ulaşılmıştır.

Karma etkili makine öğrenmesi algoritmalarından bir diğeri olan Gauss süreci güçlendirme algoritması uzun süreli veri kümesi için AR(1) ve Gauss varyans-kovaryans yapılarını içerecek şekilde iki farklı biçimde uygulanmıştır. Bu algoritma cevap değişkenin sahip olduğu dağılım bilgisine modelde yer verebildiğinden oluşturulan modellerde logaritmik bağ fonksiyonu kullanılmıştır. AR(1) varyans-kovaryans yapısına sahip modelin Gauss varyans-kovaryans yapısına sahip modele göre daha düşük HKOK ve OMH değerlerine sahip olmasından dolayı, çalışmadaki uzun süreli veri kümesi için tahmin performansı açısından daha uygun olduğu görülmüştür.

Çalışmada uzun süreli veri kümesi için uygulanan karma etkili makine öğrenmesi algoritmalarının sonuncusu ise GDKEM Ağacı algoritmasıdır. Bu algoritma ile oluşturulan modelde cevap değişkenin dağılım bilgisine yer verilerek logaritmik bağ fonksiyonu kullanılmış ancak varyans-kovaryans yapısına ilişkin bir bilgi modele eklenememiştir. GDKEM Ağacı algoritması ile elde edilen model performans değerleri ise HKOK ve OMH için sırasıyla 0,05658 ve 0,03542 olarak hesaplanmıştır.

Uzun süreli veri kümesi için oluşturulan tüm klasik istatistiksel yöntemler ve makine öğrenmesi yöntemleri birlikte düşünüldüğünde AR(1) varyans-kovaryans yapısına sahip MERF modelinin en iyi model performans değerlerine sahip olduğu ve kullanılan uzun süreli için tahmin performansı açısından HKO, HKOK ve OMH değerleri bakımından en iyi model olduğu sonucuna varılmıştır. Varyans-kovaryans yapısı dikkate alınmadığında, HKOK ve OMH değeri ile yine MERF modelinin diğer yöntemlere göre üstün performans gösterdiği görülmüştür.

Algoritmaların çalışma süreleri incelendiğinde, uzun süreli veri kümesi için en uygun model olarak belirlenen AR(1) varyans-kovaryans yapısına sahip MERF modelinin çalışma süresinin 1,5 dk olduğu görülmüştür. Bununla birlikte, AR(1) varyans-kovaryans yapısına sahip RE-EM Ormanı modelinin 43,9 dk ile en yüksek çalışma süresine sahip olmasına rağmen, tüm modeller içerisinde HKO, HKOK ve OMH değerleri bakımından en iyi ikinci model olduğu göze çarpmıştır.

Çalışmanın sonucunda, rastgele etki parametreleri için önceden belirlenmiş bir varyans-kovaryans yapısının ve cevap değişkeninin sahip olduğu dağılım bilgisinin karma etki modellerine ilave edilmesi durumunda uzun süreli verilerin analizi için karma etkili makine öğrenmesi algoritmalarının klasik istatistiksel yöntemlere göre model tahmin performansı açısından daha iyi sonuç verdiği anlaşılmıştır. Buna ek olarak, uygulamada HKO, HKOK ve OMH model performans değerlendirme ölçütlerine göre en iyi performans gösteren (S)MERF modelinin sabit etki parametreleri için herhangi dağılımsal bir varsayıma ihtiyaç duymaması da makine öğrenmesi algoritmalarının uzun süreli verilerdeki kullanım esnekliğini ön plana çıkarmıştır.

İleriki çalışmalarda, karma etkili makine öğrenmesi algoritmalarının sahip olduğu avantajlar düşünülerek, birçok farklı alanda cevap değişkeninin normal dağılıma sahip olmadığı ve kayıp ölçümlerin olduğu büyük hacimli uzun süreli veri kümelerinde karma etkili makine öğrenmesi algoritmaları kullanılarak gerçek hayata ilişkin tahminler yapılabilir.

KAYNAKLAR

[1] P. Diggle, Analysis of Longitudinal Data, New York: Oxford University Press, **2002**.

[2] J. J. Heckman and B. Singer, Econometric analysis of longitudinal data, Handbook of econometrics, p. 1689–1763, **1986**.

[3] S. Hu, Statistical Modelling and Machine Learning in Longitudinal Data Analysis, Doktora Tezi, Queensland University of Technology, **2021**.

[4] J. D. Singer and J. B. Willett, Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence, New York, Oxford University Press, p. 644, **2003**.

[5] R. W. Wedderburn, Quasi-likelihood functions, generalized linear models, and the gauss—newton method, Biometrika, no. 61(3), p. 439-447, **1974**.

[6] S. L. Zeger and K.Y. Liang, An overview of methods for the analysis of longitudinal data, Statistics in medicine, no. 1(14-15), p. 1825–1839, **1992**.

[7] K. Y. Liang and S. L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika, no. 73(1), p. 13-22, **1986**.

[8] M. Wang, Generalized estimating equations in longitudinal data analysis: a review and recent developments, Advances in Statistics, **2014**.

[9] N. M. Laird and J. H. Ware, Random-effects models for longitudinal data, Biometrics, p. 963–974, **1982**.

[10] G. Fitzmaurice and G. Molenberghs, Advances in Longitudinal Data Analysis: An Historical Perspective, Harvard University, School of Public Health, **2008**.

[11] J. C. Pinheiro, D. M. Bates and M. J. Lindstorm, Model Building for Nonlinear Mixed Effects Models, **1994**.

[12] M. J. Lindstrom and D. M. Bates, Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data, Journal of the American Statistical Association, no. 83(404), p. 1014–1022, **1988**.

[13] D. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of eugenics, no. 7(2), p. 179–188, **1939**.

[14] J. S. Cramer, The origins of logistic regression, Tinbergen Institute Working Paper, **2002**.

[15] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and regression trees, **1984**.

[16] L. Breiman, Bagging predictors, Machine Learning, no. 24, pp. 123-140, **1996**.

[17] L. Breiman, Random forests, Machine Learning, no. 45(1), pp. 5-32, **2001**.

[18] V. Vapnik, S. Golowich and A. Smola, Support vector method for function approximation, regression estimation, and signal processing, Advances in Neural Information Processing Systems, no. 9, pp. 281-287, **1997**.

[19] M. Berger and G. Tutz, Tree-structured clustering in fixed effects models, Journal of computational and graphical statistics, no. 27(2), pp. 380-392, **2018**.

[20] D. Cho, Mixed-effects ls-svm for longitudinal data, Journal of the Korean Data & Information Science Society, no. 21, p. 363–369, **2010**.

[21] A. Hajjem, F. Bellavance and D. Larocque, Mixed-effects random forest for clustered data, Journal of Statistical Computation and Simulation, no. 84(6), p. 1313–1328, **2014**.

[22] M. G. Kundu and J. Harezlak, Regression trees for longitudinal data with baseline covariates, Biostatistics & epidemiology, no. 3(1), pp. 1-22, **2019**.

[23] J. Luts, G. Molenberghs, G. Verbeke, S. Van Huffel and J. A. Suykens, A mixed effects least squares support vector machine model for classification of longitudinal data, Computational Statistics & Data Analysis, no. 56(3), p. 611–628, **2012**.

[24] M. R. Segal, Tree-structured models for longitudinal data, Journal of the American Statistical Association, no. 87(418), pp. 407-418, **1992**.

[25] R. J. Sela and J. S. Simonoff, RE-EM trees: a data mining approach for longitudinal data, Machine Learning, pp. 169-207, **2012**.

[26] C. Ngufor, H. V. Houten, B. Caffo, N. Shah and R. McCoy, Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin A1c, Journal of Biomedical Informatics, pp. 56-67, **2019**.

[27] A. Hajjem, F. Bellavance and D. Larocque, Mixed effects regression trees for clustered data, Statistics and Probability Letters, pp. 451-459, **2011**.

[28] L. Capitaine, R. Genuer and R. Thiebaut, Random forests for high-dimensional data, Statistical Methods in Medical Research, pp. 166-184, **2021**.

[29] F. Sigrist, Gaussian process boosting, arXiv preprint arXiv:2004.02653, **2020**.

[30] A. Hajjem, D. Larocque and F. Bellavance, Generalized mixed effects regression trees, Statistics and Probability Letters, pp. 114-118, **2017**.

[31] M. Fokkema, N. Smits, A. Zeileis, T. Hothorn and H. Kelderman, Detecting treatment-subgroup interactions in clustered data with generalized linear mixed effects model trees, Behavior Research Methods, no. 50(5) , pp. 2016–2034, **2018**.

[32] C. L. Vuuren, Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey, Journal of Affective Disorders, no. 295, pp. 1415-1420, **2020**.

[33] W. Cao, Longitudinal Data Prediction in EHR: Comparison of GLMM and Machine Learning Methods, Yüksek Lisans Tezi, University of Rhode Island, **2019**.

[34] K. V. Mens, C. Schepper, B. Wijnen and S. J. Koldijk, Predicting future suicidal behaviour in young adults, with different machine learning techniques: A population-based longitudinal study, Journal of Affective Disorders, no. 271, pp. 169-177, **2020**.

[35] I. H. Erduran, Adapting a Robust Model into Hybrid Implementations of Machine Learning Algortihms and Statistical Methods for Longitudinal Data,Yüksek Lisans Tezi, Middle East Technical University, Ankara, **2021**.

[36] S. Çakar, Longitudinal Data Analysis with Statistical and Machine Learning Methonds in Neuroscience,Yüksek Lisans Tezi, Middle East Technical University, Ankara, **2022**.

[37] S. Erdoğan, Statistical and Machine Learning Modelling of Suicide Rate with Respect Provinces in Turkey, Yüksek Lisans Tezi, Middle East Technical University, Ankara, **2022**.

[38] W. T. West, K. B. Welch and G. A. T. , Linear mixed models: a practical guide using statistical software, Chapman and Hall/CRC, **2006**.

[39] A. P. Dempster, L. N. M. Laird and R. D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society:Series B (Methodological), no. 39(1), pp. 1-22, **1977**.

[40] L. Wu, Mixed effects models for complex data, CRC press, **2009**.

[41] G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs, Longitudinal Data Analysis, Chapman & Hall/CRC Press, New York, **2009**.

[42] S. Kılıç, Genelleştirilmiş Tahmin Denklemlerinde Çalışan Korelasyon Yapısına Entropi Yaklaşımı, Marmara Üniversitesi SBE, İstanbul, **2012**.

[43] R. I. Jennrich and M. D. Schluchter, Unbalanced Repeated Measures Models with Structured Covariance Matrices, Biometrics, no. 42, pp. 805-820, **1986**.

[44] M. Davidian, ST732-Applied Longitudinal Data Analysis, http://www.stat.ncsu.edu/people/davidian/st732/., (Erişim tarihi: **12 Şubat 2024**).

[45] T. Louis, General Methods for Analysing Repeated Measures, Statistics in Medicine, no. 7, pp. 29-45, **1988**.

[46] G. J. McLachlan and K. T., The EM algorithm and extensions, Wiley, New York, **1997**.

[47] D. A. Harville, Maximum likelihood approaches to variance component estimation and to related problems, Journal of the American Statistical Association, no. 72, pp. 320-338, **1977**.

[48] A. Zeileis, T. Hothorn and K. Hornik, Model-based recursive partitioning, Journal of Computational and Graphical Statistics, no. 17(2), p. 492–514, **2008**.

[49] R. R. Schapire and Y. Freund, Boosting: Foundations and algorithms, Kybernetes, **2013**.

[50] C. E. Rasmussen, Gaussian processes in machine learning, in Summer School on Machine Learning, Lecture Notes in Computer Science, no. 3176, pp. 63-71, **2003**.

[51] M. Brooks, K. Kristensen, K. van Benthem, A. Magnusson, C. Berg, A. Nielsen, H. Skaug, M. Maechler and B. Bolker, glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling, The R Journal, no. 9(2), p. 378–400, **2017**.

[52] L. Capitaine, R Package 'LongituRF': Random Forests for Longitudinal Data, **2020**.

[53] F. Sigrist, R Package gpboost: Combining Tree-Boosting with Gaussian Process and Mixed Effects Models, **2022**.

[54] M. Fokkema and A. Zeileis, R Package 'glmertree', **2023**.

[55] Tüvtürk, Araçların Muayene Periyodları, https://www.tuvturk.com.tr/arac-muayene-periyodlari.aspx. (Erişim tarihi: **20 Aralık 2023**).