



**TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ**



GELENEKSEL MAKİNE ÖĞRENMESİ ve KARIŞIK ETKİLİ MAKİNE ÖĞRENMESİ MODEL PERFORMANSLARININ BENZETİM ÇALIŞMASI ile DEĞERLENDİRİLMESİ

Ebru TURGAL

**BİYOİSTATİSTİK ANABİLİM DALI
DOKTORA TEZİ**

**DANIŞMAN
Doç. Dr. Beyza DOĞANAY ERDOĞAN**

**ANKARA
2022**

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

GELENEKSEL MAKİNE ÖĞRENMESİ ve KARIŞIK ETKİLİ
MAKİNE ÖĞRENMESİ MODEL PERFORMANSLARININ
BENZETİM ÇALIŞMASI ile DEĞERLENDİRİLMESİ

Ebru TURGAL

BİYOİSTATİSTİK ANABİLİM DALI

DOKTORA TEZİ

DANIŞMAN

Doç. Dr. Beyza DOĞANAY ERDOĞAN

ANKARA

2022

Ankara Üniversitesi

Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Doktora tezi olarak hazırlayıp sunduğum “Geleneksel Makine Öğrenmesi ve Karışık Etkili Makine Öğrenmesi Model Performanslarının Benzetim Çalışması ile Değerlendirilmesi” başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan çalışma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir. Yukarıda belirtilen hususların doğruluğunu beyan ederim.

Öğrencinin Adı Soyadı: Ebru TURGAL

Tarih: 04/03/2022

İmza:

Ankara Üniversitesi Sağlık Bilimleri Enstitüsü
Biyoistatistik Anabilim Dalında
Ebru TURGAL tarafından hazırlanan
Geleneksel Makine Öğrenmesi ve Karışık Etkili Makine Öğrenmesi Model
Performanslarının Benzetim Çalışması ile Değerlendirilmesi adlı tez çalışması aşağıdaki
jüri tarafından DOKTORA TEZİ olarak OY BİRLİĞİ ile kabul edilmiştir.

Tez Savunma Tarihi: 04/03/2022

Prof. Dr. Atilla Halil ELHAN

Ankara Üniversitesi Ankara Üniversitesi
Jüri Başkanı

Doç. Dr. Derya GÖKMEN
Ankara Üniversitesi

Doç. Dr. Beyza DOĞANAY ERDOĞAN
Ankara Üniversitesi

Doç. Dr. Mustafa Agah TEKİNDAL
İzmir Katip Çelebi Üniversitesi

Doç. Dr. Can ATEŞ
Aksaray Üniversitesi

Tez hakkında alınan jüri kararı, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu tarafından onaylanmıştır.

Prof. Dr. Fügen AKTAN

Sağlık Bilimleri Enstitüsü Müdürü

İÇİNDEKİLER

Etik Beyan	ii
Kabul ve Onay	iii
İçindekiler	iv
Önsöz	v
Simgeler ve Kısaltmalar	vi
Şekiller	viii
Çizelgeler	ix
1. GİRİŞ	1
1.1. Araştırmanın Konusu, Amacı ve Önemi	1
1.2. Karışık Etkili Makine Öğrenmesinin Tarihçesi	1
1.3. Doğrusal Karışık Etkili Model	6
1.4. Karışık Etkili Makine Öğrenmesi (MEM)	8
1.4.1. Karışık Etkili Makine Öğrenmesi Formülasyonu	8
1.5. Karışık Etkili Regresyon Ağacı (MERT)	9
1.6. Karışık Etkili Random Forest (MERF)	10
1.7. Topluluk Yöntemleri	11
1.7.1. Bagging (Torbalama)	11
1.7.2. Boosting (Arttırma)	11
1.8. Gauss Süreci	12
1.8.1. Gauss Süreci Boosting	15
1.8.2. Gauss Süreci ile Karışık Etkili Model	15
2. GEREÇ VE YÖNTEM	18
2.1. Benzetim Çalışması	18
2.1.1. Kullanılan Sistem Kapasitesi ve Programlar	19
2.1.2. GPBoost için Model Parametreleri	20
2.1.2.1. GPBoost için Ayarlama Parametreleri	21
2.2. Hata Ölçüm Parametreleri	22
2.2.1. Hata Kareler Ortalamasının Karekökü (RMSE)	22
3. BULGULAR	24
4. TARTIŞMA	39
5. SONUÇ VE ÖNERİLER	41
ÖZET	43
SUMMARY	44
KAYNAKLAR	44
ÖZGEÇMİŞ	49

ÖNSÖZ

Bu tez çalışmasında, Geleneksel Makine Öğrenmesi ve Karışık Etkili Makine Öğrenmesi Model performansları incelenmiş ve benzetim çalışması ile değerlendirilmiştir.

Doktora eğitimim ve tez çalışması süresince ilgisini, bilgisini ve desteğini esirgemeyen değerli danışman hocam Doç. Dr. Beyza DOĞANAY ERDOĞAN'a,

Tanıdığım günden itibaren her konuda bana destek olan, ilgi ve alakasını esirgemeyen değerli hocalarım Prof. Dr. Atilla Halil ELHAN'a, Prof. Dr. Yasemin YAVUZ'a, Doç.Dr. Serdal Kenan KÖSE'ye ve Doç. Dr. Derya Gökmen'e sadece doktora hayatımda öğrettikleri için değil, hayatımın seyrini değiştirdikleri için, ayrıca ilgi ve anlayışlarından dolayı tüm Biyoistatistik Anabilim Dalı asistan arkadaşlarıma ve bölüm çalışanlarına,

Eğitim sürecimde maddi-manevi yanımda olan, tüm başarılarımın en büyük mimarı anneme, babama, kardeşime, kuzenime, teyzeme ve anneanneme tez çalışmamın her anında bana destek olan Hazal Doğaroğlu'na teşekkürlerimi sunarım.

SİMGELER VE KISALTMALAR

$\hat{f}(x_{ij})$	i Kümesine Ait Yeni bir j Gözleminin Ortalama Popülasyon Ağaç (Orman) Tahmini
CART	Sınıflandırma ve Regresyon Ağaçları
EM	Beklenti Maksimizasyonu
GAM	Genelleştirilmiş Toplamsal Modeller
GBM	Gradyan Destekli Makine
GLL	Genelleştirilmiş Log-olabilirlik
GLM	Genelleştirilmiş Doğrusal Model
GLMM	Genelleştirilmiş Doğrusal Karışık Etkili Model
GMERT	Genelleştirilmiş Karışık Etkili Regresyon Ağacı
GPBoost	Gauss Süreci Boosting
HKT	Hata Kareler Toplamı
LME	Doğrusal Karışık Etkili Model
m	Küme Sayısı
MEml	Karışık Etkili Makine Öğrenmesi Yöntemi Paketi
MERF	Karışık Etkili Random Forest
MERT	Karışık Etkili Random Tree
ML	Makine Öğrenmesi
NLMM	Doğrusal Olmayan Karışık Etkiler Modeli
RE-EM ağacı	Rastgele Etki Beklenti Maksimizasyonu Ağacı
RF	Random Forest

RMSE	Hata Kareler Ortalamasının Karekökü
SVM	Destek Vektör Makinesi
XGBoost	Extreme Gradient Boosting



ŞEKİLLER

Şekil 1.1: Sınıflandırıcı karşılaştırma çıktıları	13
Şekil 2.1: Eğitim seti için 4 kat çapraz doğrulama kullanılarak oluşturulan parametre ayarı	22
Şekil 3.1: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (n=5000)	30
Şekil 3.2: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (n=10000)	30
Şekil 3.3: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (n=15000)	31
Şekil 3.4: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (n=5000)	31
Şekil 3.5: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (n=10000)	32
Şekil 3.6: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (n=15000)	32
Şekil 3.7: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (n=5000)	33
Şekil 3.8: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (n=10000)	33
Şekil 3.9: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (n=15000)	34
Şekil 3.10: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (n=5000)	34
Şekil 3.11: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (n=10000)	35
Şekil 3.12: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (n=15000)	35
Şekil 3.13: 'doğrusal' fonksiyon için gerçek ve tahmin edilen sabit etkilerin karşılaştırılması ($\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1$, n=5000, küme sayısı=250)	36
Şekil 3.14: 'friedman3' fonksiyonu (Eşitlik 2.2 için) için SHAP değerleri	37
Şekil 3.15: 'friedman3' fonksiyonu (Eşitlik 2.2 için) SHAP özet grafiği	37
Şekil 3.16: 'friedman3' fonksiyonu (Eşitlik 2.2 için) SHAP etkileşim grafiği	38

ÇİZELGELER

Çizelge 1.1: Bagging ve boosting süreçleri	12
Çizelge 1.2: Doğrusal ve doğrusal olmayan karışık etkili model yöntemleri için özet bilgi	17
Çizelge 3.1: Kümelenmiş rastgele etkiler modeli ve ortalama $F = \text{'doğrusal'}$ fonksiyonu için sonuçlar	25
Çizelge 3.2: Kümelenmiş rastgele etkiler modeli ve ortalama $F = \text{'doğrusal'}$ fonksiyonu için sonuçlar	25
Çizelge 3.3: Kümelenmiş rastgele etkiler modeli ve ortalama $F = \text{'doğrusal'}$ fonksiyonu için sonuçlar	26
Çizelge 3.4: Kümelenmiş rastgele etkiler modeli ve ortalama $F = \text{'friedman3'}$ fonksiyonu için sonuçlar	26
Çizelge 3.5: Kümelenmiş rastgele etkiler modeli ve ortalama $F = \text{'friedman3'}$ fonksiyonu için sonuçlar	27
Çizelge 3.6: Kümelenmiş rastgele etkiler modeli ve ortalama $F = \text{'friedman3'}$ fonksiyonu için sonuçlar	27

1. GİRİŞ

1.1. Araştırmanın Konusu, Amacı ve Önemi

Makine öğrenmesinin gelişimi son yıllarda hızla artmış ve bu konuda pek çok algoritma geliştirilmiştir. Şimdiye kadar, kümelenmiş verilerde karışık etkili makine öğrenmesi modellerinin performanslarını inceleyen yalnızca bir çalışmaya rastlanmıştır (Ngufor ve ark., 2019). Ngufor ve ark. (2019) çalışması hariç olmak üzere, bu çalışma literatürde karışık etkili makine öğrenmesi algoritmasını farklı senaryolar altında karşılaştıran ilk çalışma olacaktır.

Bu tez çalışmasında, kümelenmiş verilerin sınıflamasında kullanılan geleneksel modellerden doğrusal karışık etkili model (LME), random forest ile yeni bir yaklaşım olan karışık etkili makine öğrenmesi modelinin etkinlikleri karşılaştırılacaktır.

1.2. Karışık Etkili Makine Öğrenmesinin Tarihçesi

Gauss süreci ve karışık etkili modellerin çoğunluğu, ortalama fonksiyonunun doğrusal bir regresyon fonksiyonu olduğunu varsayar. Ancak bu doğrusallık varsayımı genelde ya sağlanamaz ya da varsayım göz ardı edilmek zorunda kalınır. Ortalama fonksiyonunu doğrusal bir şekilde değil de doğrusal olmayan bir şekilde modelleyen veri madenciliği yöntemlerine literatürde az rastlanmıştır. Ağaç arttırma veya random forest gibi modern denetimli makine öğrenimi tekniklerini karışık etkili modellerle ve özellikle Gauss süreçleriyle birleştirme konusunda çok az araştırma yapılmıştır.

Hajjem ve ark. (2011), Sela ve Simonoff (2012), Fu ve Simonoff (2015) ortalama fonksiyon için regresyon ağaçlarını kullanmışlardır. Hajjem ve ark. (2014) ise boylamsal verilerde kullanılan kümelenmiş rastgele etki modelleri için random forest kullanarak parametrik olmayan, makine öğrenimine dayalı yaklaşımlar önermiştir.

Hajjem ve ark. (2011, 2014) karışık etkili modeller için bir beklenti maksimizasyonu (EM) algoritmasından (Dempster ve ark. 1977) esinlenen bir yaklaşım önermektedir (Laird ve ark., 1982). Bu yöntem ağacın bölünme sürecinde farklı zaman dilimlerinde tekrarlı olarak ölçülen ortak değişkenleri kullanır ve bu ortak değişkenlerle ilgili uygun rastgele etkileri ele alır.

Ayrıca, Sela ve Simonoff (2012), Fu ve Simonoff (2015) ilk olarak, ağaç algoritması kullanarak bölünmeyi öğrenen ve ikinci olarak, ağaç katsayılarını ve kovaryans parametrelerini tahmin eden geleneksel doğrusal karışık etkiler model yöntemini içeren yinelemeli bir algoritma kullanmıştır. Bu yöntemlerin ortak özelliği, iteratif olarak (i) bir makine öğrenimi tekniği kullanarak ortalama fonksiyonu öğrenmeleri, (ii) rastgele etkilerin tahminlerini hesaplamaları ve (iii) kovaryans parametrelerini tahmin etmeleridir.

Özellikle ortalama fonksiyon, tekrar tekrar tahmin edilmesi gereken random forest gibi karmaşık bir modelden oluştuğunda bu yaklaşımlar hesaplama açısından zorlu hale gelebilir. Ek olarak, bu algoritmaların yakınsama garantisi yoktur. Sela ve Simonoff (2012), Fu ve Simonoff (2015), Hajjem ve ark. (2011), Hajjem ve ark. (2014), uygun bir E-adımı içermediklerinden karışık etkili modeller için doğru şekilde belirlenen EM algoritmalarına karşılık gelmez (Laird ve ark., 1982) ve bu nedenle bu algoritmaların yakınsayıp yakınsamadığı ve hangi değerlere yaklaştığı açık değildir (Sigrist, 2020).

Karışık etkili modeller için EM algoritmasındaki E-adım, ilk olarak, mevcut kovaryans parametrelerindeki çok değişkenli normal olabilirliğin ortalama fonksiyonu için bir en çoklayıcı (maksimizer) belirleyerek, rastgele etkiler için tahmin elde eder ve sonra bu ikisini tüm verinin log-olabilirlik beklentisini hesaplamak için kullanır (Laird ve ark., 1982).

Bununla birlikte, Hajjem ve ark. (2011), Hajjem ve ark. (2014) sabit kovaryans parametreleri için çok değişkenli normal olabilirliğin en çoklayıcısı olarak ortalama fonksiyonun optimize edicisini bulmaz. Bunun yerine, ilk olarak, ortalama fonksiyonu tahmin etmek için rastgele etkilerin tahmin edilen değerlerini yanıt değişkeninden çıkardıktan sonra elde edilen bağımsız bir normal olasılığı kullanılır; ikinci olarak, rastgele etkiler için tahminler elde edilir, üçüncü olarak; kovaryans parametrelerinin tahminleri bulunur (M- adım) ve bu üç adım yakınsama sağlanana kadar yinelenir.

Sela ve Simonoff (2012), Fu ve Simonoff (2015) ilk olarak yanıt değişkeninden rastgele etkilerin tahmin değerlerinin çıkarılmasından elde edilen bağımsız normal olabilirliği kullanarak ağaç yapısının tahminleri ile, ikinci olarak, klasik doğrusal karışık etkiler modeli kullanarak yaprak değerler ve kovaryans parametreleri arasındaki tahminleri yinelemişlerdir. Aslında bunun tam bir EM algoritması olduğu söylenemez, çünkü log-olabilirlik beklentisini hesaplayan bir E-adımı içermemektedir (Sigrist, 2020).

Bassetti ve ark. 2015 yılında siyasal şiddetin bazı ekonomik göstergelerle ilişkili olduğunu göstermek için, Genelleştirilmiş Karışık Etkili Regresyon Ağacı (GMERT) analizini kullanmışlardır. Bu yöntem, boylamsal veriler için rastgele etki model yapısını bir ağaç regresyon yönteminin esnekliğiyle birleştirmiştir. Bu yöntem regresyon fonksiyonunu tahmin etmek için parametrik olmayan bir yöntemdir. 2002–2007 döneminde siyasal şiddete etki eden faktörler belirlenmiş ve bu etkenlerin makine öğrenmesi yöntemi kullanılarak hem sınıflama yapısını göz önünde bulundurmuş hem de öldürülen ve yaralanan kişilerin tahminleri elde edilmiştir. Terör saldırıları, düşük

eđitim endeksli (0,52'den az) ve kiři bařına yksek gelir (1817 doların zerinde) ile karakterize edilen lkelerde yođunluk gsterirken, eđitim endeksi 0,52'den yksek olan lkelerde yılda ortalama 47,33 saldırı gerekleřmiřtir.

Natarajan (2019) alıřmasında NBA oyuncu seimi tahmini iin karıřık etkili model oluřturmuřtur. ne ıkan basketbol oyuncularının szleřmelerini etkileyen faktrler karıřık etkili model ile sunulmuřtur. Natarajan (2019), řut sayısı, top kullanma oranı, artı eksi istatistiđi ve her oyunda alınan puan gibi parametreleri kullanarak en iyi oyuncuları tahmin eden modeli sunmuřtur. Bulgulara gre Karıřık Etkili Random Forest (MERF) modelinin rneklem ii ve rneklem dıřı performansı random forest ve gradyan destekli makine đrenme yntemlerine gre yksek ıkmıřtır.

Karıřık Etkili Random Forest ynteminin ele alındıđı bir bařka alıřma ise 2019 yılında yayımlanan Galeshchuk ve Qiu'nun (2019) makalesidir. Twitter'ın, fikirlerin hızlı ve net bir řekilde paylařılmasına yardımcı olan sosyal bir platform olduđu ve bu platform, kısa metinler yayınlamak iin eřit haklar sunsa da, bu mesajların sıklıkla ektiđi ilginin, kullanıcının gerek dnyadaki durumuna bađlı olduđu belirtilmiřtir. Dolayısıyla, gerek hayattaki yksek profilli fenomenlerin tweet'lerinin, toplumun ilgisini artırma olasılıđı, yapay tabanlardan gelen mesajlara gre daha yksek olacaktır. Makale, bu tr fikir oluřturucularını Twitter'da otomatik olarak algılayan geliřmiř sınıflandırıcayı ayrıntılı olarak ele almıř olup, yaklařım yntemi olarak Karıřık Etkili Random Forest yntemini kullanır. nerilen tekniđin dođruluđu ve kesinliđi ile birlikte, rneklem dıřı verilerde diđer makine đrenimi sınıflandırıcılarının sonularından daha iyi performans gsterdiđi aıklanmıřtır. Dođruluk oranı %80,7 bulunarak, lojistik regresyon, karar destek makineleri, radyal baz fonksiyonu, random forest ve gradyan destekli makine modellerini geride bırakmıřtır.

Beltempo ve ark. (2020) makine đrenimi modellerinden GMERT analizini kullanarak yeni dođan yođun bakım biriminde nozokomiyal enfeksiyonların ve tıbbi

kazaların başlangıcını tahmin etmeyi amaçlamıştır. Model, önemli değişkenleri belirlemede kullanılmıştır ve sonuç olarak; teşhis ilişkili gruplar, normal çalışma saatleri, fazla mesai, kabul oranları, doğum ağırlığı ve meslek oranları ana parametreleri oluşturmuştur.

Cochrane ve ark. (2021) yayımladıkları çalışmada, geleneksel Doğrusal Karışık Etkiler modelini, MERF adı verilen yarı parametrik, doğrusal olmayan bir modelle birleştiren bir Topluluk Karışık Etkiler Modeli sunmaktadır. Bu modeli kullanarak, laboratuvar deneyleri sırasında incelenen 44 uyku kaybı yaşayan katılımcıdan demografik, uyku ve bilişsel test verilerini içeren bir veri ile sürekli dikkat testi olan Psikomotor Vigilance Görevi (PVT) üzerindeki performansı tahmin edilmek istenmiştir. Bir bireyin uyku kaybı sonrası bilişsel performansının nasıl üstesinden geldiğini tahmin eden bir topluluk modeli geliştirmişlerdir. MERF, kümelenmiş veriler için karışık modelleme yaklaşımlarının başarısını random forest algoritmasının güçlü yönleriyle birleştirmiştir.

Zhou ve ark. (2021) monoton otomatik sürüşte sürücü yorgunluğuna etki eden değişkenleri tahmin etmek için GPBoost ve SHAP modellemesini kullanmışlardır. Ayrıca; doğrusal regresyon, doğrusal, karesel, gaussçu destek vektör makinesi, fine (ince) ağaç, random forest, XGBoost ve GPBoost modellerinin uygulamalı olarak karşılaştırmasını göstermişlerdir.

Levy ve ark. (2021), Nanostring GeoMx üzerinde Dijital Mekansal Profilci kullanılarak toplanan uzamsal protein belirteçlerinin değerlendirilmesi ve kolon metastaz tahmini için Karışık Etkili Makine Öğrenmesi yöntemlerini karşılaştırmalı olarak göstermişlerdir.

Fontana ve ark. (2021), Sela ve Simonoff'un (2012) geliştirdiği RE-EM ağacını genelleştirmeyi amaçlamışlardır. Ayrıca, bu yöntem kümelenmiş veri yapısını

geleneksel çok düzeyli modeller gibi ele alır. GMET adını verdikleri bu model, yanıt değişkenini bir Genelleştirilmiş Doğrusal Model (GLM) aracılığıyla (yanıt ailesi dağılımına bağlı olan uygun bağlantı fonksiyonlarını kullanarak) tahmin eder. Bağımlı değişken olarak tahmin edilen yanıt değişkenini kullanarak bir regresyon ağacı oluşturur ve ağaç tarafından tahmin edilen sabit etkiler kısmını sapma olarak kullanır ve karışık etkiler modeli ile rastgele etkiler kısmını tahmin eder. GMET'in performansının, diğer ağaç tabanlı karışık etkili modellere göre ((GLMM), BiMM (Speiser ve ark., 2020), GLMERT (Fokkema ve ark., 2018), GMERT (Hajjem ve ark., 2017)) veriler doğrusal bir yapıya sahip olduğunda daha iyi ve yakınsama süresinde daha hızlı olduğunu göstermiştir.

Tezin karışık etkili makine öğrenimi ile birleşecek olan Gauss süreçleri, son on yılda makine öğrenimi topluluğundan büyük ilgi görmüştür. İlk olarak bu terim 1940'larda Robert Winner ve Kolmogorov tarafından ortaya atılmış ve zaman serileri çözümlemesinde açıklanmıştır (Whittle, 1963). Ardından 1950'lerde Güney Afrika'daki Witwatersrand'daki altın yataklarının modellenmesi üzerinde çalışan Danie Krige tarafından geliştirilmiştir. Formülasyon, 1963 yılında Fransız matematikçi Georges Matheron tarafından resmileştirilmiştir (Matheron, 1963). Geliştirilen yöntem kriging olarak biliniyordu ve bu terim istatistik literatüründe kullanılmaya devam etmektedir (Ashcroft, 2021).

1.3. Doğrusal Karışık Etkili Model

Genel olarak, bir doğrusal karışık etkili model aşağıdaki koşulları sağlayan herhangi bir model olabilir.

$$y_i = X_i\beta + Z_ib_i + \varepsilon_i,$$

$$b_i \sim N(0, D),$$

$$\varepsilon_i \sim N(0, \sigma^2 I_{n_i}),$$

$$\text{Cov}(b_i, \varepsilon_i) = 0,$$

$$i = 1, 2, \dots, N.$$

(Eşitlik 1.1)

Modelde y_i i. birey için n_i boyutlu yanıt vektörü, $1 \leq i \leq N$, N: birey sayısı, X_i ve Z_i sırasıyla $(n_i \times p)$ ve $(n_i \times q)$ boyutlu bilinen açıklayıcı değişkenler matrisleri, β sabit etkileri içeren p-boyutlu vektör, b_i rastgele etkileri içeren q-boyutlu vektör, ε_i hata bileşenlerini içeren n_i -boyutlu vektördür. Son olarak, D, $d_{ij} = d_{ji}$ olmak üzere, $(q \times q)$ boyutlu genel kovaryans matrisi, ve $\sigma^2 I_{n_i}$ ($n_i \times n_i$) boyutlu kovaryans matrisidir (Verbeke ve Molenberghs, 2000). Rastgele etkilerin sıfır ortalama ve varyans kovaryans matrisi D ile normal dağıldığı varsayılır ve ε_i hata terimlerinin bağımsız olduğu varsayılır.

β sabit etkilerin yorumu, basit doğrusal regresyon modelindeki gibidir. İlgili değişkendeki bir birim artışın y_i ortalamadaki değişimi gösterir. b_i rastgele etkiler i. denek için regresyon parametrelerinin bir alt kümesi olarak popülasyonda diğerlerinden nasıl saptığı şeklinde yorumlanır.

Karışık etkili modellerin avantajı, hem popülasyonda değişkenin ortalama yanıtının nasıl değiştiğini açıklayan parametreleri hem de zamanla bireysel trendlerin nasıl değiştiğini tahmin eder (Pinheiro ve Bates, 2000; Verbeke ve Molenberghs, 2000).

1.4. Karışık Etkili Makine Öğrenmesi (MEml)

1.4.1. Karışık Etkili Makine Öğrenmesi Formülasyonu

MEml, güçlü bir makine öğrenme algoritması kullanarak Eşitlik 1.2'deki sabit etkiler bileşenini $f(x_i)$ ve GLMM (Genelleştirilmiş Karışık Etkili Model) kullanarak b_i rastgele etkiler değerini tahmin eder.

$$\eta_{it} = f(x_i) + b_i^T z_i \quad (\text{Eşitlik 1.2})$$

$$y_i^* = f(x_i) + b_i^T z_i + \varepsilon_i^* \quad (\text{Eşitlik 1.3})$$

f : bilinmeyen fonksiyondur.

Eşitlik 1.3'de f ağaç tabanlı makine öğrenme algoritmaları kullanılarak tahmin edilir. Bunlar; random forest (Breiman ve ark., 1984), gradyan destekli makine (GBM) (Friedman, 2001), model tabanlı özyinelemeli bölümlleme ve koşullu çıkarım ağaçlarıdır (Zeileis ve ark, 2008). Ağaç tabanlı öğrenme algoritmaları kolay yorumlanabilir olduğu için daha sık tercih edilmektedir. RF (Random Forest) ve GBM, bir model topluluğu oluşturan ve daha sonra tahminleri birleştiren topluluk makine öğrenme yöntemleridir.

RF'de eğitim setindeki gözlemlerin rastgele örnekleri ortalama (avareging) ve oy çokluğu (majority vote) ile birleştirilen tahminleri, karar ağaçları tarafından öğrenilir. Oy çokluğu yöntemi ise, bir modelde yanlış tahmin edilen imgenin başka modellerde doğru tahmin edilmesiyle yanlış tahminin düzeltilebilmesini sağlar. Ortalama yöntemi regresyon için, oy çokluğu yöntemi ise sınıflamada kullanılır (Karakaş, 2019).

1.5. Karışık Etkili Regresyon Ağacı (MERT)

Hajjem ve ark. 2011 yılında MERT ile, rastgele etkileri EM algoritması kullanarak elde ettikten sonra, sabit etkileri standart bir regresyon ağacı yöntemiyle kestirmiştir. Bu yaklaşım, yanıt değişkeni ile sabit etkiler arasında doğrusallık varsayımını gerektirmez, ancak rastgele etkilerle ilgili olarak doğrusal karışık etkili modellerle aynı varsayımların sağlanması gerekir.

Karışık etkili regresyon ağacı modeli aşağıdaki gibidir:

$$y_i = f(X_i) + Z_i b_i + \varepsilon_{it},$$

$$b_i \sim N(0, D),$$

$$\varepsilon_i \sim N(0, \sigma^2 I_{n_i}),$$

$$i = 1, 2, \dots, N.$$

(Eşitlik 1.4)

Sabit kısım hariç tüm nicelikler klasik doğrusal karışık etkili modeldeki gibi gösterilmiştir. $f(X_i)$, artık $X_i \beta$ doğrusal kısım yerine geçer. Bu kısım, ağaç ya da orman ile kestirilecektir. $Z_i b_i$ rastgele etkilerin doğrusallık varsayımını sağlaması gerekir. ML-tabanlı EM-algoritması için akış diyagramı Wu ve Zhang'da (2006) açıklanmıştır. Yanıt değişkenini tahmin etmek için MERT / MERF modelini kestirmede kullanılan i kümesine ait yeni bir j gözleminin ortalama popülasyon ağaç / orman tahmini $\hat{f}(x_{ij})$ ve kendi kümesindeki $Z_i \hat{b}_i$ rastgele kısmın tahmini kullanılır. Modeli eğitmek için kullanılan örnekte yer almayan bir kümeye ait yeni bir gözlem için, yalnızca karşılık gelen $\hat{f}(x_{ij})$ ortalama popülasyon ağaç / orman) tahmini alınabilir. Bu nedenle,

$$1. \text{ Bilinen bir küme için tahmin} = \hat{f}(x_{ij}) + Z_i \hat{b}_i$$

2. Yeni küme için tahmin = $\hat{f}(x_{ij})$

şeklindedir.

1.6. Karışık Etkili Random Forest (MERF)

Random forest yönteminin kümelenmiş veri için kullanılan bir uzantısı olan karışık etkili random forest yöntemi 2014 yılında Hajjem ve ark. tarafından geliştirilmiştir. MERT'in bir uzantısı olarak ortaya konmuştur. Yöntem 2017 yılında MERF adlı bir Python paketinde uygulanmıştır (Hajjem ve ark., 2014). MERF algoritması, MERT ile aynı şekilde çalışır ve rastgele etkiler üzerinde aynı varsayımlara sahiptir. Bununla birlikte, sabit etkileri tahmin etmek için tek bir regresyon ağacı kullanmak yerine random forest kullanılır. Rastgeleleştirme yoluyla ilişkili hale getirilmiş birçok regresyon ağacının toplanması daha doğru tahmin elde edilmesini sağlar.

Eşitlik 1.1'de gösterilen LME modelinde olduğu gibi, bu modeldeki terimler neredeyse aynıdır, sadece $X_i\beta$ 'nın doğrusal sabit etkisi doğrusal olmayan $f(\cdot)$ fonksiyonuyla yer değiştirir. Doğrusal karışık etkili modelde sabit etkiler, basit parametrik bir biçimde modellenir. Oysa çok sayıda ortak değişken (sabit etki) olduğu durumda, modele eklenmesi gereken sabit etkilerin sayısı ve birbirleri ile olan etkileşimleri de artacak ve bu karmaşık yapının modellenmesi güçleşecektir. Ayrıca sabit etkinin sayıca fazla olduğu durumda model seçimi de zorlaşacaktır. Doğrusal modeller, kayıp veri içeren değişkenleri veri madenciliği yöntemleri kadar kolay ele alamamaktadır. Sabit etkiler ile yanıt değişkeni arasındaki doğrusal olmayan ilişki random forest yöntemi ile öğrenilmektedir. Bu regresyon modeli gradyan destekli ağaç veya derin sinir ağı gibi doğrusal olmayan herhangi bir regresyon modeli olabilir.

1.7. Topluluk Yöntemleri

1.7.1. Bagging (Torbalama)

Bu algoritma, bootstrap ve aggregating (toplama) teriminin kısaltılmış hali olup, Breiman tarafından 1996 yılında ortaya atılmıştır (Yangın, 2019). Torbalama algoritmasının amacı, eğitim veri kümesi kullanarak rastgele yeni veri kümeleri oluşturup farklılıkların oluşmasını ve böylece sınıflandırma başarısını arttırmaktır. Ayrıca, eğitim veri kümesinden, yerine koyarak rastgele seçim yöntemi ile yeni eğitim kümeleri türeterek sınıflandırma modelini yeniden eğitmeyi amaçlamaktadır. Torbalama yönteminde yapılması gereken ilk adım, veri kümesini eğitim kümesi ve test kümesi olarak bölmektir. N tane örneklem içeren eğitim kümesinden, yerine koyarak rastgele seçim yöntemiyle yine n tane örnekleme sahip bir veya birden fazla yeni bir eğitim kümesi üretilir. Torbalama yöntemiyle oluşan topluluktaki her bir sınıflandırıcı, oluşturulmuş farklı örneklem içeren eğitim kümesi ile eğitilir. Her bir sınıflandırıcının sonucu çoğunluk oylaması ile birleştirilir (Yangın, 2019; Yılmaz, 2014; Çoşkun, 2020; Demirsöz ve ark., 2021).

1.7.2. Boosting (Arttırma)

Arttırma algoritması Schapire tarafından 1999 yılında geliştirilmiştir. Arttırma algoritması, yavaş öğrenmeye dayalı, sıralı bir yöntemdir ve hatadan öğrenmeyi amaçlamaktadır. Bu algoritmalar, yüksek hassasiyetli modeller oluşturmak için düşük hassasiyetli birkaç modeli birleştirmektedir. Amaç, her iterasyonda elde edilen modelleri, belirli kurallar çerçevesinde birleştirerek güçlü bir model elde etmeye çalışmaktır (Yangın, 2019; Yılmaz, 2019).

Çizelge 1.1: Bagging ve boosting süreçleri

Bagging	Boosting
1. Orjinal eğitim setinden N örnek rastgele seçilerek yeniden örneklenmiş eğitim setleri oluşturulur.	1. Orjinal eğitim setinden rastgele yeniden örneklendirilmiş bir alt eğitim seti oluşturulur.
2. Oluşturulan her bir eğitim seti seçilen öğrenim algoritmasına eklenir.	2. Oluşturulan eğitim seti öğrenim algoritmasına sokulur.
3. Tüm sınıflandırıcılardan elde edilen sonuçların ortalaması alınarak tek bir model elde edilir.	3. Tahminlerde hatalı olan örneklerin ağırlıkları daha yüksek verilerek ikinci bir yeniden örneklendirilmiş eğitim seti oluşturulur.
	4. 3. adım M kere gerçekleştirilerek tüm modellerden elde edilen sonuçların ağırlıklı kombinasyonu alınarak tek bir model elde edilir.

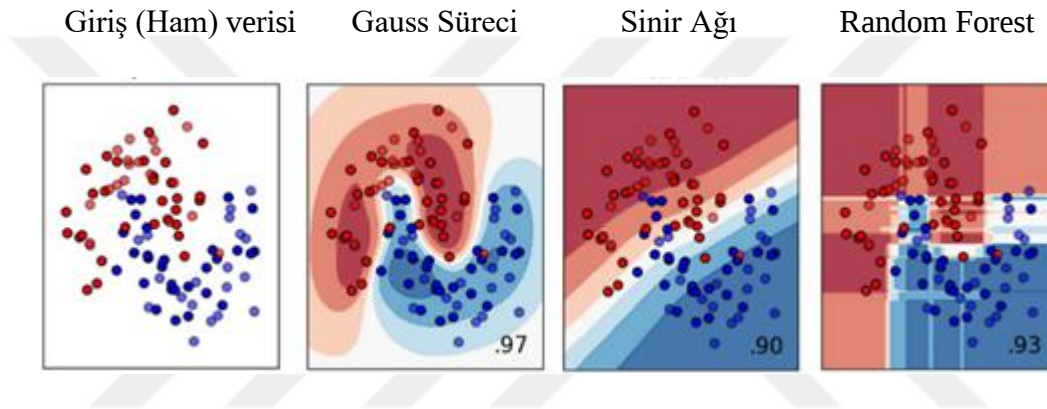
(Yıldırım, 2017).

1.8. Gauss Süreci

Gauss süreçleri (Williams ve Rasmussen, 2006) son teknoloji tahmin doğruluğuna ulaşan ve olasılıksal tahminler yapmaya olanak tanıyan esnek parametrik olmayan fonksiyon modelleridir (Gneiting ve ark, 2007). Olasılıksal gözetimli makine öğrenme yöntemlerinden olan bu süreç, hem regresyon hem de sınıflama problemlerinde kullanılmaktadır.

Gauss süreci, fonksiyonları yaklaştırmak için kullanılan ve parametrik olmayan Bayesci bir regresyon yaklaşımı (Ateş, 2020) olan bu Bayesci yaklaşım, olası fonksiyonlar üzerinden bir olasılık dağılımı belirler (Knagg, 2020). Bu sürecin algoritması, rastgele değişken çiftleri arasındaki doğrusal olmayan ilişkileri modellemek için güçlü bir algoritmadır. Gerçek fonksiyonel ilişki konusundaki belirsizliği göstermek için uygulanabilecek fonksiyonlar üzerinde bir dağılım tanımlar (Williams ve Rasmussen, 2006).

Bu yaklaşımda parametreler bağımsız özdeşçe dağıtılmış rastgele değişkenler olarak görülmektedir. Diğer denetimli öğrenme yöntemleri de her parametre için kesin değerleri öğrenir. Gauss süreci regresyonu verilere uygun tüm kabul edilebilir fonksiyonlar üzerindeki olasılık dağılımını hesaplar. Önsel nokta belirlenir, eğitim verileri kullanılarak sonsal nokta hesaplanır ve ilgili noktalardaki tahmini sonsal ile karşılaştırılır. Bu yöntemde $\sum(y - \hat{y})^2$ sonucunun minimum olması gerekir (Ateş, 2020).



Şekil 1.1: Sınıflandırıcı karşılaştırma çıktıları

Şekil 1.1’de mavi ve kırmızı noktaları ayırmak için kullanılan farklı yöntemlerle öğrenilen sınıflandırma fonksiyonları gösterilmektedir. Yaygın olarak kullanılan ve güçlü yöntemler olan sinir ağı ve random forest, eğitim verilerinden uzakta tahminler üretmektedir. Gauss süreci, model çıktısını daha kesin bir doğrulukla elde eder, bu yapı özellikle kimlik doğrulama ve güvenlik açısından kritik kullanımlarda önemlilik göstermektedir (Knagg, 2019).

Gauss süreci, parametrik olmayan regresyon, zaman serilerinin modellenmesi (Shumway ve Stoffer, 2017), uzaysal (Banerjee ve ark., 2014) ve mekan-zamansal (Cressie ve Wikle, 2015) verileri, bilgisayar deneyleri bayes kalibrasyonu (Kennedy ve O’Hagan, 2001), pahalı kara kutu fonksiyonlarının optimizasyonu (Jones ve ark., 1998) ve makine öğrenimi modellerinde parametre ayarlama (Snoek ve ark., 2012) gibi

alanlarda da kullanılır. Ek olarak robotik alanında hareket planlama, konum belirleme, lokalizasyon ve eşleme konularında oldukça yaygın bir şekilde kullanılmaktadır (Mukadam ve ark., 2016; Yangın, 2016).

Gauss süreci regresyonu, insanların tercihleri üzerinden model kurulması gereken alanlarda (kullanıcı tercihlerinin anlaşılması gibi) kullanılmaktadır. Günümüzde insan yürüyüşü, dinazor yürüyüşü, yangın, patlama gibi dinamik animasyonların modellenmesinde kullanılmaktadır. Çözünürlüğü düşük animasyonlar insanlara izletilerek puanlandırmaları istendikten sonra insanların gerçeklik algısı puanlanarak model parametreleri son halini almaktadır. Gauss regresyon süreci insanların tercihlerini birer fonksiyon olarak modellemektedir. Süreç Bayesci olarak başlatıldığı için ilk başta öncül bilgilerle başlayan model sonuçta gerçeğe yakın olmaktadır. Böylece, yüksek maliyetli bir iş düşük işlemci maliyeti ile çözülmüş olmaktadır.

Gauss süreçleri sadece regresyon probleminde değil, sınıflandırma problemlerinde de kullanılmaktadır. Örnek olarak Google'ın sınıflandırma problemleri verilebilir. Google'da 'kedi' kelimesiyle resim aratıldığında arama motoru kedi olduğunu düşündüğü birçok fotoğrafı sonuç olarak vermektedir. Kediye en çok benzeyen tıklandığında bu fotoğrafların içinden çoğunlukla kullanıcının ilk olarak basmayı tercih ettiği fotoğraf kediye en çok benzeyen fotoğraftır diye kabul edilir. En çok tıklanan yüksek bir oylama oranı ile 'kedi' olarak sınıflandırır. Daha sonra bu öncül bilgiyi kullanılarak 'kedi' sınıflandırma probleminde kestirim başarımını artırarak daha sonraki kullanıcıların kedi aramasında arama motoru kediye daha çok benzeyen fotoğrafları öne çıkarmaktadır.

1.8.1. Gauss Süreci Boosting

Model, GPBoost algoritması kullanılarak eğitilir. Bu eğitim, rastgele etkilerin kovaryans parametrelerini ve ortalama $F(X)$ fonksiyonunu bir ağaç topluluğu modeli ile eğiterek gerçekleşir. Kullanılan algoritma, kovaryans parametrelerini yinelemeli olarak öğrenen ve bir gradyan veya Newton artırma adımı kullanarak ağaç topluluğuna ağaç ekleyen bir artırma algoritmasıdır. Kovaryans parametreleri (Nesterov hızlandırılmış) gradyan inişi veya Fisher skorlaması kullanılarak öğrenilebilir (Sigrist, 2020).

Sonlu örnekler için, artırma aşırı uyma eğilimindedir ve bunu önlemek için, özellikle regresyon görevlerinde erken durdurma uygulanmalıdır.

1.8.2. Gauss Süreci ile Karışık Etkili Model

Kümelenmiş veriler aşağıdaki şekillerdeki gibi modellenenebilir;

1. Grup yapısı göz ardı edilebilir
2. Her grup (örn. her öğrenci ya da müşteri) ayrı modellenenebilir
3. Modele grupta değişkeni (örn. öğrenci veya müşteri kimliği) dahil edilir ve kategorik bir değişken olarak ele alınır. Bu uygulanabilir bir yaklaşım olsa da bazı dezavantajlara sahiptir. Çoğu zaman, grup başına ölçüm sayısı (örneğin, öğrenci başına test sayısı, müşteri başına işlem sayısı) küçüktür ve farklı grupların sayısı fazladır (örneğin, öğrenci sayısı, müşteri sayısı, vb.). Bu durumda, modelin öğrenmeyi verimsiz hale getirebilecek daha az veriye dayalı birçok parametreyi (her grup için bir tane) öğrenmesi gerekir. Ayrıca, fazla sayıda alınan kategorik değişkenler ağaçlar için problem teşkil edebilir.

4. Gruplama deęiřkeni rastgele etkiler olarak ele alınır ve Gauss sũreci ile karıřık etkili model yaklařımı kullanılabilir.

Gauss sũreci ile karıřık etkili modellerde, ilk momentin genellikle sıfır veya ortak deęiřkenlerin doęrusal bir fonksiyonu olduęu varsayılır. Yapılandırılmıř varyasyon hatası (residual structured variation), daha sonra sıfır ortalamalı Gauss sũreci ve / veya kũmelenmiř rastgele etkiler modeli kullanılarak modellenir. Bununla birlikte, hem sıfır ortalama hem de doęrusallık varsayımı genellikle gerçekçi deęildir ve bu varsayımlar esnetilerek daha yũksek tahmin doęruluęu elde edilebilir (Sigrist, 2020). Ayrıca, bir Gauss sũreci modelinin ortalama fonksiyonu yanlış tanımlanırsa, bũyle bir yanlış tanımlanmıř modelin kovaryans fonksiyonu, gerçek kovaryans fonksiyonu artı ortalama fonksiyonun hata karesine eřit olduęu iin sahte ikinci dereceden duraęanlık meydana gelebilir (Fuglstad ve ark., 2015; Schmidt ve Guttorp, 2020). Bu nedenle, potansiyel ikinci derece duraęan olmayan artıęı hesaba katmaya alıřmadan ũnce, ilk olarak ortalama fonksiyonu doęru řekilde modellemek ũnemlidir. En son teknolojiye sahip denetimli makine ũęrenimi algoritmalarında, ũzellikle arttırma (boosting) ile esnek ve potansiyel olarak karmařık bir fonksiyonun bir dizi tahmin deęiřkenini bir yanıt deęiřkeniyle iliřkilendirdięi varsayılır. Tahmin edici deęiřkenlere baęlı olarak, yanıt deęiřkeninin gũzlemler arasında baęımsız olduęu varsayılır. Bu, potansiyel artık korelasyonunun, yani regresyon fonksiyonu tarafından aıklanmayan korelasyonun gũz ardı edildięi anlamına gelir. Bũyle bir korelasyonu modellemek sadece regresyon fonksiyonunun daha iyi ũęrenilmesini saęlamaz, aynı zamanda olasılıklı tahminler ve gruplar ũzerinden ortalamaları veya toplamları tahmin etmek iin de ũnemlidir (Sigrist, 2020).

Kısaca bu yaklařım, ortalama fonksiyonu ve rastgele etkilerin kovaryans yapısının ortalama fonksiyon ile birlikte tahmin edilen parametrelerini, arttırma kullanarak ařamalı bir řekilde ũęrenen regresyon aęaları (Breiman ve ark., 1984) gibi temel ũęrenicilerden oluřan bir toplulukla modellemeyi ũnermektedir.

Makine öğrenmesi literatüründe Gauss süreci ile karışık etkili modellerde $F(X) = 0$ olduğu varsayılırken, Gauss süreci ile karışık etkili modellerde önsel dağılım fonksiyonunun Eşitlik 1.5’deki gibi olduğu varsayılır.

$$F(X) = X^T \beta, \beta \in \mathbb{R}^P \text{ ortak değişkenler vektörü} \quad (\text{Eşitlik 1.5})$$

Bu yaklaşımın amacı doğrusallık varsayımını ya da sıfır önsel ortalama varsayımını esnetmektir.

Algoritmanın temelinde tüm modeller eğitim verileri üzerinde eğitilir ve test verileri üzerinde değerlendirilir. Her bir adımda benzetim, kısaca “interpolasyon” ve “ekstrapolasyon” test setleri olarak gösterilen iki test veri seti aşağıdaki gibi oluşturulur. Kümelenmiş rastgele etkiler modeli için, “interpolasyon” test veri seti, eğitim verilerinde olduğu gibi aynı gruplar için rastgele etkilerden oluşur. “Ekstrapolasyon” test verisi, eğitim veri setinde yer almayan yeni gruplar için m bağımsız rastgele etkileri içermektedir.

Çizelge 1.2: Doğrusal ve doğrusal olmayan karışık etkili model yöntemleri için özet bilgi

Yöntem	Sabit etkiler/önsel ortalama fonksiyon $F(\cdot)$	Rastgele etkiler (Zb)
Kümelenmiş rastgele etkiler modeli	Doğrusal $F(\cdot)$	Kümelenmiş rastgele etkiler
Gauss süreci modeli	Doğrusal $F(\cdot)$ /yok	Gauss süreci
Klasik boosting	Boosting ile öğrenilmiş doğrusal olmayan $F(\cdot)$	-
MERT	Random tree ile öğrenilmiş doğrusal olmayan $F(\cdot)$	Kümelenmiş rastgele etkiler
MERF	Random forest ile öğrenilmiş doğrusal olmayan $F(\cdot)$	Kümelenmiş rastgele etkiler
GPBoost	Boosting ile öğrenilmiş doğrusal olmayan $F(\cdot)$	Gauss süreçleri ve/veya rastgele etkiler

2. GEREÇ VE YÖNTEM

Bu tez çalışmasının amacı kümelenmiş verilerin sınıflamasında kullanılan geleneksel modellerden doğrusal karışık etkili model ile yeni birer yaklaşım olan karışık etkili makine öğrenmesi modellerinin etkinliklerini incelemektir. Tez kapsamında; veri türetme örneklem büyüklüğü, küme sayısı, rastgele etki varyansı ve hata varyansına göre farklı senaryolarda yapılmıştır.

Benzetim tekniği için örneklem büyüklükleri 5000, 10000 ve 15000; küme sayıları 250, 500, 1000 alınmıştır.

2.1. Benzetim Çalışması

Benzetim çalışması kapsamında doğrusal karışık etkili model, MERF ve GPBoost yöntemlerinin RMSE değeri bakımından karşılaştırılması iki fonksiyon üzerinde gerçekleştirmiştir. Bu fonksiyonlardan ilki doğrusal bir fonksiyondur ve Eşitlik 2.1’de gösterilmiştir. Ayrıca tek bir gruplama değişkeni kullanılmıştır. Ancak hiyerarşik olarak iç içe geçmiş rastgele etkiler ve çaprazlanmış rastgele etkiler kullanılabilir. Çalışmada rastgele etki varyansı 1, hata varyansı 1 ve 4 olarak ayrı ayrı alınmıştır. Her senaryo 100 kere tekrarlanarak gerçekleştirilmiştir.

$$F(x) = C_1 * x + 2 \quad (\text{Eşitlik 2.1})$$

İkinci olarak, doğrusal olmayan bir fonksiyon olan 'Friedman3' fonksiyonu da benzetim çalışmasına dahil edilmiştir. 'Friedman3' fonksiyonu ilk olarak Friedman'da (1991) yer almıştır ve parametrik olmayan regresyon modellerini karşılaştırmak için sıklıkla kullanılmaktadır (Eşitlik 2.2). Bu fonksiyon 4 tahmin edici içeren doğrusal

olmayan bir fonksiyondur ve çok boyutlu bir yapıya sahip olduğundan dolayı genellikle makine öğrenmesi kapsamında bu fonksiyon kullanılarak y yanıt değişkenine en yakın tahmin değeri elde edilmeye çalışılmaktadır.

$$F(x) = C_2 * \arctan\left(\frac{x_2x_3 - \frac{1}{x_2x_4}}{x_1}\right) \quad (\text{Eşitlik 2.2})$$

C sabitleri, Eşitlik 2.1 ve 2.2’de gösterildiği üzere $F(X)$ ’in varyansı 1’e eşit olacak, yani $F(X)$ rastgele etkiler ile aynı güce sahip olacak şekilde seçilmiştir.

$$y = F(X) + Zb + \varepsilon, b \sim N(0, \sigma_b^2), \varepsilon_i \sim N(0, \sigma^2) \quad (\text{Eşitlik 2.3})$$

2.1.1. Kullanılan Sistem Kapasitesi ve Programlar

Araştırmanın tüm aşamalarında, Intel(R) Xeon(R) Gold 6136 3 GHz 16 çekirdekli işlemci özelliğine ve 256 GB rastgele erişimli belleğe (RAM) sahip bir sunucu kullanılmıştır. Ayrıca, çalışmadaki analizlerde Python Jupyter Notebook 6.4.8 kullanılmıştır.

Çalışma süresince oluşturulan ve çalıştırılan kodlar, talep üzerine ilgili yazardan istenebilir.

2.1.2. GPBoost için Model Parametreleri

Max_Depth: Ağacın dallarının aşağı doğru uzamasının değeridir. Diğer bir deyişle ağacın derinliğidir. Aşırı öğrenmeden kaçınmak için optimum seviyeye getirilmelidir. Çok dallanma aşırı öğrenmeye, az dallanma eksik öğrenmeye neden olmaktadır. Varsayılan parametre -1 olarak alınmıştır.

Learning_rate: Ağaçları ölçeklendirmek için 0-1 arasında verilen bir değerdir. Bu değer küçük olması daha iyi tahmin gücüne yardımcı olmaktadır. Fakat; öğrenim süresini ve aşırı öğrenme olasılığını arttıracaktır. Varsayılan parametre 0,1 olarak alınmıştır.

Iterations: Oluşturulacak ağaç sayısıdır. Farklı algoritmalarda “num_boost_round”, “n_estimators”, “num_trees”, “num_iterations” isimleri ile de kullanılır. Az olması eksik öğrenmeye, fazla olması aşırı öğrenmeye neden olabilmektedir. Ayrıca sayının artması eğitim süresini de arttırmaktadır. Varsayılan parametre 32 olarak alınmıştır.

Early_stopping_rounds: Aşırı öğrenmeyi engellemek için kullanılan parametredir. En uygun adımı bulduktan sonra kaç kez deneme yapılacağı belirtilmelidir. Bu değer 100 alındığında model en uygun olduğu andan sonra 100 iterasyon daha yapar ve hedef parametreler yakalanmasa bile model öğrenmeyi durdurur. Örneğin başlangıç parametresinde iterasyon sayısı 2000 verilmiş olsun, en uygun ana 1000. iterasyonda ulaşırsa model burada duracaktır (Muratlar, 2020). Varsayılan parametre 5 olarak alınmıştır.

Verbosity: Her iterasyonda çıktı olarak modelin öğrenim durumu, toplam süre ve kalan süre bilgisi gelmektedir. Bu çıktı çok iterasyonun olduğu durumda ekranda çok fazla alan kaplamaktadır. Varsayılan parametre 0 olarak alınmıştır.

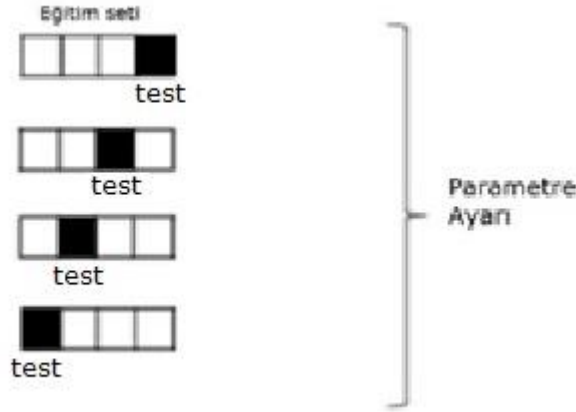
Feature_fraction: 1,0'dan küçükse, her yinelemede (ağaç) rastgele bir öznelik (parametre) alt kümesi seçer. Örneğin, 0,8 olarak alınırsa, algoritma her bir ağacı eğitmeden önce parametrelerin %80'ini seçecektir. Varsayılan parametre 1,0 olarak alınmıştır.

Subsample: Eğitim örneğinin alt örnek oranıdır. Ağaçlara verilen örnekleri kontrol etmektedir. Örneğin 0,5 değerine ayarlanırsa, algoritmanın ağaçları büyütmeden önce eğitim verilerinin yarısını rastgele örneklemesine izin vermek demektir ve bu da aşırı öğrenmeyi önleyecektir. Varsayılan parametre 1,0 olarak alınmıştır.

Kullanılan karar ağaçları algoritması ile arttırma algoritmalarının karşılaştırılmasının objektif bir şekilde yapılabilmesi amacıyla algoritmalarda kullanılan temel parametreler her bir algoritma için aynı girilmiştir. Aynı değerlere sahip parametreler sırasıyla; 'n_estimators', 'learning_rate' ve 'max_depth' dir. Yüksek max_depth değeri daha kompleks bir model yaratır ve aşırı uyum olasılığını arttırır (Coşkun, 2020).

2.1.2.1. GPBoost için Ayarlama Parametreleri

Arttırma algoritmalarından GPBoost, ayar (tuning) parametreleri, ölçüt olarak RMSE ile eğitim verileri üzerinde 4 kat çapraz doğrulama kullanılarak ve GPBoost model tahminleri için rastgele etkiler göz ardı edilerek seçilmiştir.



Şekil 2.1: Eğitim seti için 4 kat çapraz doğrulama kullanılarak oluşturulan parametre ayarı (Ek, 2020)

2.1.3. MERF için Model Parametreleri

MERF algoritması için MERF Python paketi (versiyon 0.3) kullanılmıştır ve MERF algoritmasının yineleme sayısı 100 olarak alınmıştır. Bununla birlikte, MERF için, ağaç sayısını $\{100, 300, 500\}$, ağaçların maksimum derinliğini $\in \{1, 5, 10\}$ ve genelleştirilmiş log-olabilirlik (GLL) kriterini minimum yapma amacıyla, bölünmeler $\in \{0,5, 0,75, 1\}$ şeklinde ele alınmıştır. Maksimal bir ağaç derinliği limiti koyulmayıp ağaç sayısı 300'e ayarlanmıştır. Bunlar MERF paketinin varsayılan değerleridir (Hajjem ve ark., 2014).

2.2. Hata Ölçüm Parametreleri

2.2.1. Hata Kareler Ortalamasının Karekökü (RMSE)

Hata Kareler Ortalamasının Karekökü (RMSE), artıkların (tahmin hataları) standart sapmasıdır. Artıklar, regresyon çizgisi veri noktalarından ne kadar uzakta olduğunun bir ölçüsüdür. RMSE, bu artıkların ne kadar yayıldığının bir ölçüsüdür.

Başka bir deyişle, verilerin en uygun çizgi etrafında ne kadar yoğun olduğunu söyler. Ortalama karekök hata, deneysel sonuçları doğrulamak için klimatoloji, tahmin ve regresyon analizinde yaygın olarak kullanılır. RMSE aşağıdaki eşitlikle ifade edilir (Glen, 2021).

$$\text{RMSE} = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}} \quad (\text{Eşitlik 2.4})$$



3. BULGULAR

Örneklem büyüklüğünün 5000, 10000 ve 15000, küme sayısının 250, 500 ve 1000, rastgele etki varyansının 1 ve hata varyansının 1 ve 4 olduğu durumlara göre farklı toplamda 108 senaryo oluşturulmuştur. Bu senaryolar için makine öğrenmesi yöntemlerine ait sonuçlar çizelgeler ve şekillerle verilmiştir. Tüm senaryolar için sonuçlar RMSE, RMSE'ye ait standart sapma ve zaman (saniye) kriterleri çizelge olarak, 'friedman3' modeline ait başlıca özellik seçimi için önemlilik sıralaması SHapley Additive exPlanations (SHAP) tekniği ile şekil üzerinde yapılmıştır.

Çizelge 3.1: Kümelenmiş rastgele etkiler modeli ve ortalama F = 'doğrusal' fonksiyonu için sonuçlar¹

Örneklem sayısı	Küme sayısı	σ_b^2	σ_ε^2	Doğrusal ME			MERF			GPBoost			p değeri	
				RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	Doğrusal ME - GPBoost	MERF - GPBoost
5000	250	1	1	1,254	0,0212	<0,001	1,260	0,0212	172,01	1,264	0,0209	0,813	8,67E-59	9,41E-38
	500			1,246	0,0178	0,016	1,252	0,0179	188,58	1,253	0,0179	0,053	1,55E-62	1,89E-43
	1000			1,264	0,0134	<0,001	1,268	0,0135	283,73	1,271	0,0131	0,228	1,98E-66	2,42E-38
5000	250	1	4	2,139	0,0226	<0,001	2,149	0,0228	131,162	2,161	0,0225	0,098	9,63E-58	1,47E-36
	500			2,134	0,0310	<0,001	2,147	0,0235	186,91	2,145	0,0254	0,094	0,0105	2,86E-10
	1000			2,252	0,0245	0,016	2,183	0,0209	283,36	2,187	0,0321	0,141	0,0173	5,83E-40

Çizelge 3.2: Kümelenmiş rastgele etkiler modeli ve ortalama F = 'doğrusal' fonksiyonu için sonuçlar¹

Örneklem sayısı	Küme sayısı	σ_b^2	σ_ε^2	Doğrusal ME			MERF			GPBoost			p değeri	
				RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	Doğrusal ME - GPBoost	MERF - GPBoost
10000	250	1	1	1,250	0,0217	<0,001	1,252	0,0216	154,57	1,254	0,0217	0,047	1,85E-68	1,79E-44
	500			1,246	0,0157	<0,001	1,249	0,0158	211,10	1,251	0,0157	0,098	2,80E-63	3,75E-43
	1000			1,244	0,0109	0,016	1,246	0,0108	329,77	1,248	0,0109	0,047	5,93E-62	2,79E-41
10000	250	1	4	2,142	0,0184	<0,001	2,144	0,0185	153,73	2,147	0,0186	0,173	1,65E-67	4,32E-40
	500			2,135	0,0162	0,016	2,140	0,0163	212,83	2,144	0,0163	0,176	1,04E-60	3,14E-40
	1000			2,149	0,0166	<0,001	2,152	0,0143	319,77	2,156	0,0143	0,141	1,61E-14	1,34E-34

¹: σ_b^2 : rastgele etkiler varyansı σ_ε^2 : hata varyansı, SS: standart sapma

Çizelge 3.3: Kümelenmiş rastgele etkiler modeli ve ortalama F = 'doğrusal' fonksiyonu için sonuçlar

Örneklem sayısı	Küme sayısı	σ_b^2	σ_ε^2	Doğrusal ME			MERF			GPBoost			p değeri	
				RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	Doğrusal ME - GPBoost	MERF - GPBoost
15000	250	1	1	1,242	0,0193	<0,001	1,244	0,0192	225,95	1,246	0,0192	0,113	9,37E-68	1,16E-42
	500			1,253	0,0144	<0,001	1,254	0,0143	216,71	1,255	0,0144	0,109	1,40E-63	8,49E-38
	1000			1,243	0,0106	0,016	1,245	0,0106	328,19	1,246	0,0105	0,098	6,23E-65	2,04E-43
15000	250	1	4	2,127	0,0152	0,016	2,132	0,0152	228,20	2,134	0,0152	0,237	1,73E-67	5,23E-43
	500			2,141	0,0139	<0,001	2,143	0,0138	216,34	2,146	0,0142	0,191	1,28E-60	1,20E-38
	1000			2,144	0,0124	0,016	2,147	0,0125	327,82	2,150	0,0124	0,238	1,24E-69	1,53E-45

Çizelge 3.4: Kümelenmiş rastgele etkiler modeli ve ortalama F = 'friedman3' fonksiyonu için sonuçlar²

Örneklem sayısı	Küme sayısı	σ_b^2	σ_ε^2	Doğrusal ME			MERF			GPBoost			p değeri	
				RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	Doğrusal ME - GPBoost	MERF- GPBoost
5000	250	1	1	1,395	0,0214	<0,001	1,256	0,0237	133,599	1,251	0,0229	0,063	9,66E-118	0,05130
	500			1,391	0,0187	<0,001	1,248	0,0171	197,464	1,248	0,0167	0,083	4,37E-111	0,00032
	1000			1,405	0,0163	<0,001	1,255	0,0154	303,210	1,258	0,0149	0,063	1,19E-116	0,30796
5000	250	1	4	2,221	0,0241	<0,001	2,147	0,0254	133,345	2,150	0,0242	0,129	5,82E-94	2,22E-34
	500			2,222	0,0345	<0,001	2,140	0,0222	200,187	2,145	0,0260	0,123	3,03E-53	9,10E-10
	1000			2,275	0,0252	<0,001	2,152	0,0223	315,551	2,214	0,0297	0,121	1,95E-53	2,57E-48

². σ_b^2 : rastgele etkiler varyansı σ_ε^2 : hata varyansı, SS: standart sapma

Çizelge 3.5: Kümelenmiş rastgele etkiler modeli ve ortalama F = 'friedman3' fonksiyonu için sonuçlar³

Örneklem sayısı	Küme sayısı	σ_b^2	σ_ε^2	Doğrusal ME			MERF			GPBoost			p değeri	
				RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	Doğrusal ME - GPBoost	MERF- GPBoost
10000	250	1	1	1,395	0,0194	<0,001	1,246	0,0215	161,466	1,246	0,0211	0,083	6,27E-129	9,49E-15
	500			1,399	0,0150	<0,001	1,243	0,0143	216,028	1,238	0,0144	0,079	1,65E-128	3,74E-29
	1000			1,414	0,0126	0,008	1,256	0,0128	358,004	1,251	0,0126	0,113	2,67E-133	3,08E-32
10000	250	1	4	2,238	0,0187	<0,001	2,156	0,0192	157,910	2,161	0,0190	0,145	2,08E-111	6,71E-30
	500			2,258	0,0180	0,016	2,162	0,0166	213,937	2,164	0,0166	0,145	1,49E-105	5,82E-27
	1000			2,280	0,0258	0,016	2,181	0,0172	473,814	2,187	0,0217	0,177	2,023E-70	1,65E-05

Çizelge 3.6: Kümelenmiş rastgele etkiler modeli ve ortalama F = 'friedman3' fonksiyonu için sonuçlar³

Örneklem sayısı	Küme sayısı	σ_b^2	σ_ε^2	Doğrusal ME			MERF			GPBoost			p değeri	
				RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	RMSE	SS	Zaman (s)	Doğrusal ME - GPBoost	MERF- GPBoost
15000	250	1	1	1,376	0,0170	0,016	1,241	0,0192	215,126	1,229	0,0189	0,114	6,00E-145	1,09E-27
	500			1,382	0,0132	<0,001	1,235	0,0142	222,246	1,229	0,0143	0,129	4,70E-141	1,51E-47
	1000			1,385	0,0108	0,016	1,248	0,0117	335,205	1,242	0,0120	0,141	1,62E-145	1,38E-53
15000	250	1	4	2,220	0,0168	<0,001	2,141	0,0170	220,922	2,142	0,0171	0,206	3,09E-124	1,74E-20
	500			2,234	0,0143	0,016	2,148	0,0136	221,315	2,151	0,0138	0,240	1,06E-119	5,40E-20
	1000			2,238	0,0130	<0,001	2,162	0,0136	344,900	2,160	0,0140	0,252	3,84E-124	8,66E-26

³: σ_b^2 : rastgele etkiler varyansı σ_ε^2 : hata varyansı, SS: standart sapma

Çizelge 3.1’de örneklem sayısı 5000, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı ve hata varyansı 1 olarak alındığında doğrusal karışık etkili modelin hata kareler ortalamasının karekökünün MERF ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir. Ek olarak, küme sayısı sırasıyla 250, 500 olduğunda rastgele etkiler varyansı 1 ve hata varyansı 4 olarak alındığında doğrusal karışık etkili modelin hata kareler ortalamasının karekökünün MERF ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir. Küme sayısı 1000 olduğunda MERF modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir.

Çizelge 3.2’de örneklem sayısı 10000, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı ve hata varyansı 1 olarak alındığında doğrusal karışık etkili modelin hata kareler ortalamasının karekökünün MERF ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir. Ek olarak, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı 1 ve hata varyansı 4 olarak alındığında doğrusal karışık etkili modelin hata kareler ortalamasının karekökünün MERF ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir.

Çizelge 3.3’de örneklem sayısı 15000, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı ve hata varyansı 1 olarak alındığında doğrusal karışık etkili modelin hata kareler ortalamasının karekökünün MERF ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir. Ek olarak, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı 1 ve hata varyansı 4 olarak alındığında doğrusal karışık etkili modelin hata kareler ortalamasının karekökünün MERF ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir.

Çizelge 3.4’te örneklem sayısı 5000, küme sayısı sırasıyla 250, 500 olduğunda rastgele etkiler varyansı ve hata varyansı 1 olarak alındığında GPBoost modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve MERF modeline göre daha

küçük olduğu gözlemlenmiştir. Küme sayısı 1000 olduğunda MERF modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir. Ek olarak, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı 1 ve hata varyansı 4 olarak alındığında MERF modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir.

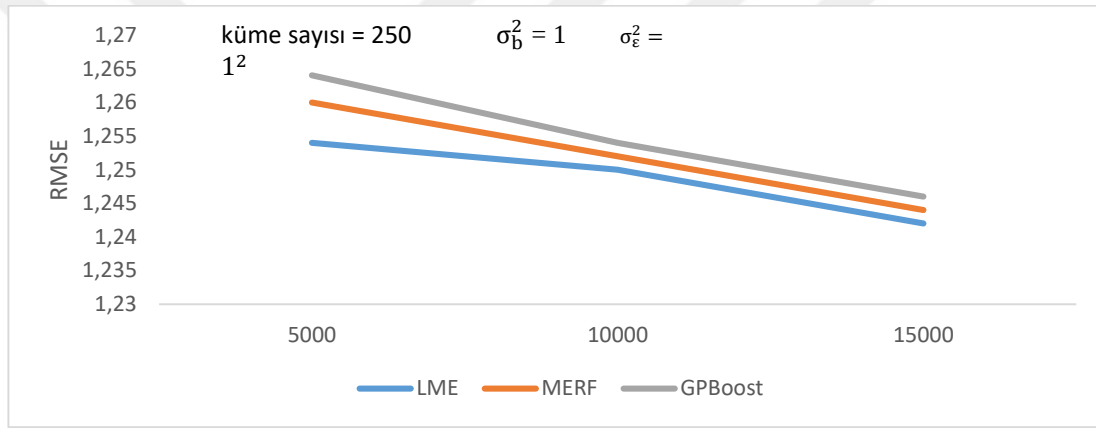
Çizelge 3.5'te örneklem sayısı 10000, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı ve hata varyansı 1 olarak alındığında GPBoost modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve MERF modeline göre daha küçük olduğu gözlemlenmiştir. Ek olarak, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı 1 ve hata varyansı 4 olarak alındığında MERF modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir.

Çizelge 3.6'da örneklem sayısı 15000, küme sayısı sırasıyla 250, 500, 1000 olduğunda rastgele etkiler varyansı ve hata varyansı 1 olarak alındığında GPBoost modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve MERF modeline göre daha küçük olduğu gözlemlenmiştir. Ek olarak küme sayısı sırasıyla 250, 500 olduğunda rastgele etkiler varyansı 1 ve hata varyansı 4 olarak alındığında MERF modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve GPBoost modeline göre daha küçük olduğu gözlemlenmiştir. Küme sayısı 1000 olduğunda GPBoost modelinin hata kareler ortalamasının karekökünün doğrusal karışık etkili ve MERF modeline göre daha küçük olduğu gözlemlenmiştir.

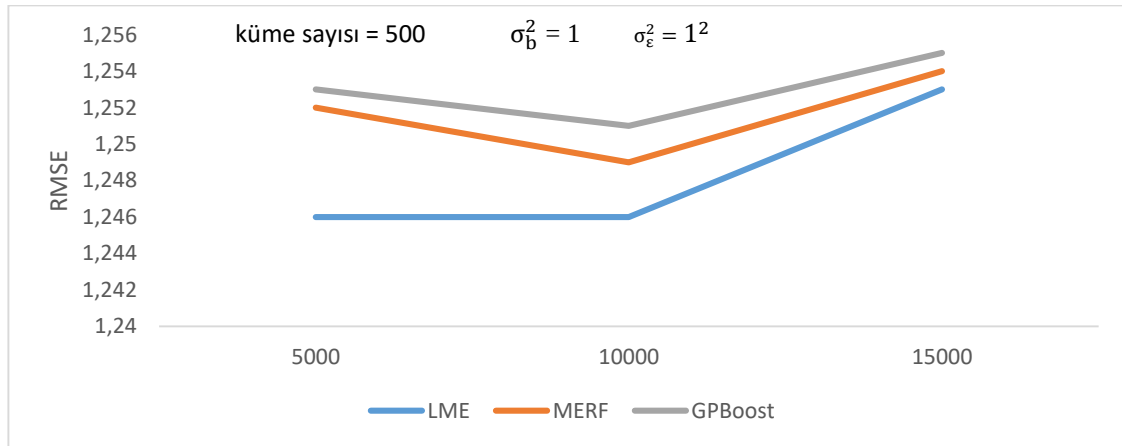
Beklenen bir şekilde, doğrusal model GPBoost algoritmasından daha hızlıdır. Ayrıca, GPBoost algoritmasının 100 benzetim için MERF algoritmasından 4000 kat daha hızlı çalıştığı da görülmektedir.

GPBoost algoritmasını RMSE değerlerini diğer yaklaşımlardan elde edilen RMSE değerleriyle karşılaştıran bağımlı gruplarda t-testleri için p değerleri de hesaplanmıştır. Sonuçlar, RMSE açısından GPBoost ile LME ve GPBoost ile MERF yöntemleri arasındaki farkın istatistiksel olarak önemli olduğunu göstermiştir ($p < 0,05$) (Çizelge 3.1 - 3.6).

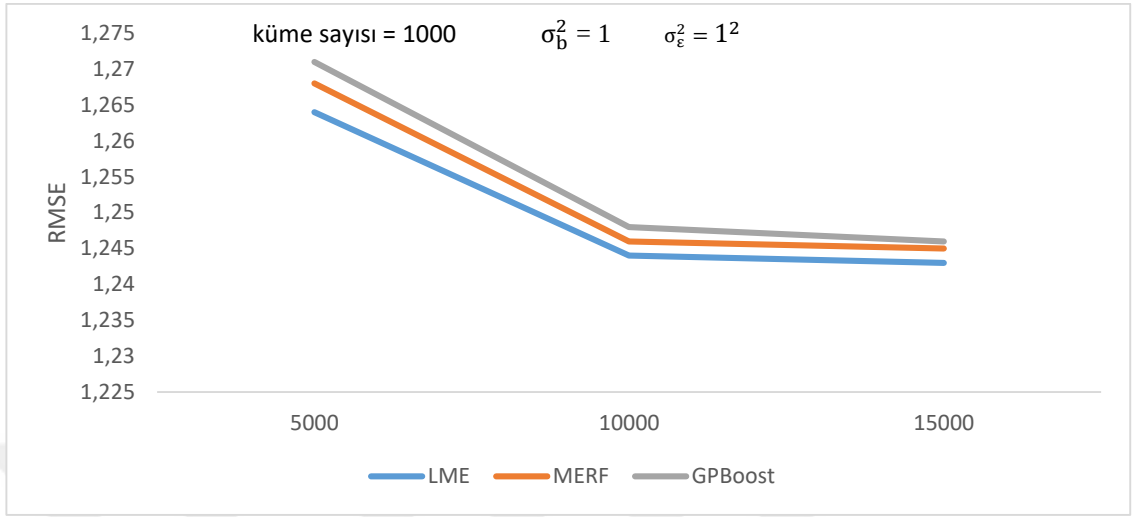
Çizelgelerde doğrusal model dışında, GPBoost algoritmasının diğer tüm yaklaşımlardan önemli ölçüde üstün olduğu görülmektedir.



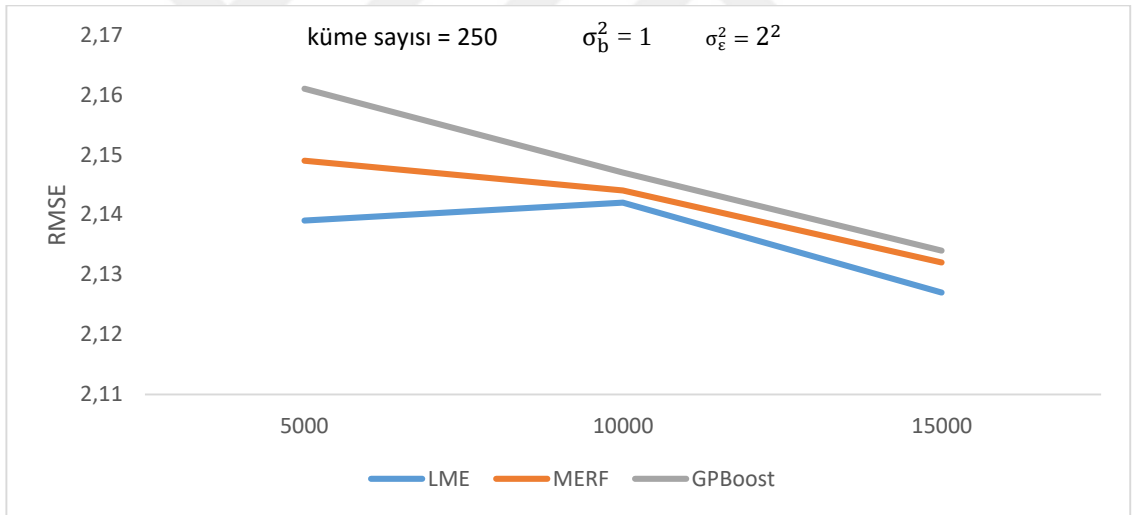
Şekil 3.1: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması ($m = 250$)



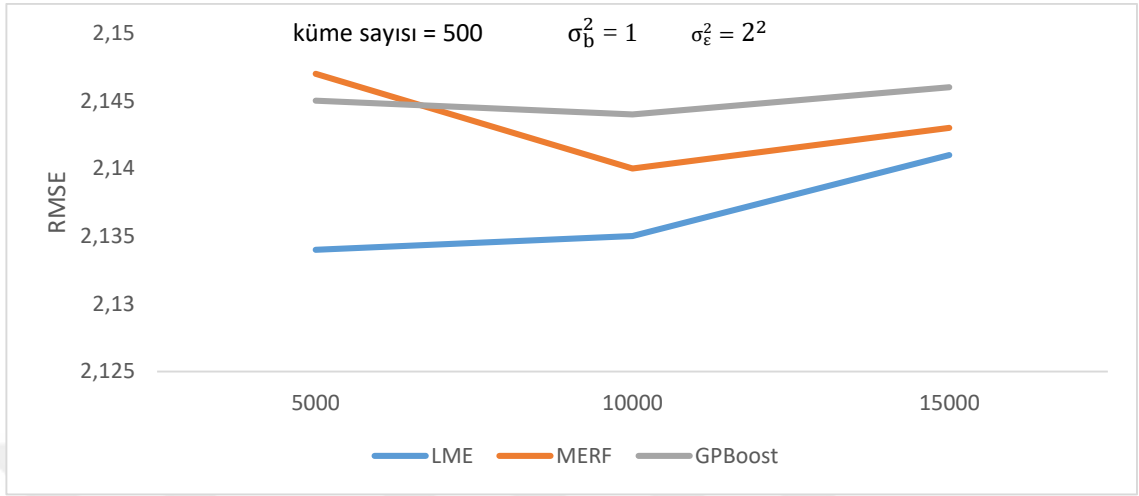
Şekil 3.2: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması ($m = 500$)



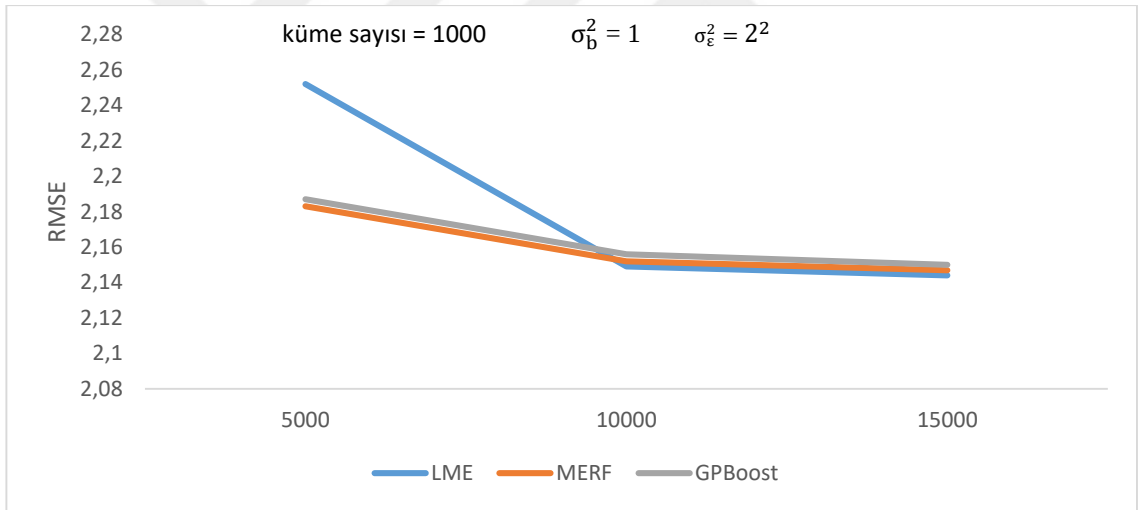
Şekil 3.3: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (m = 1000)



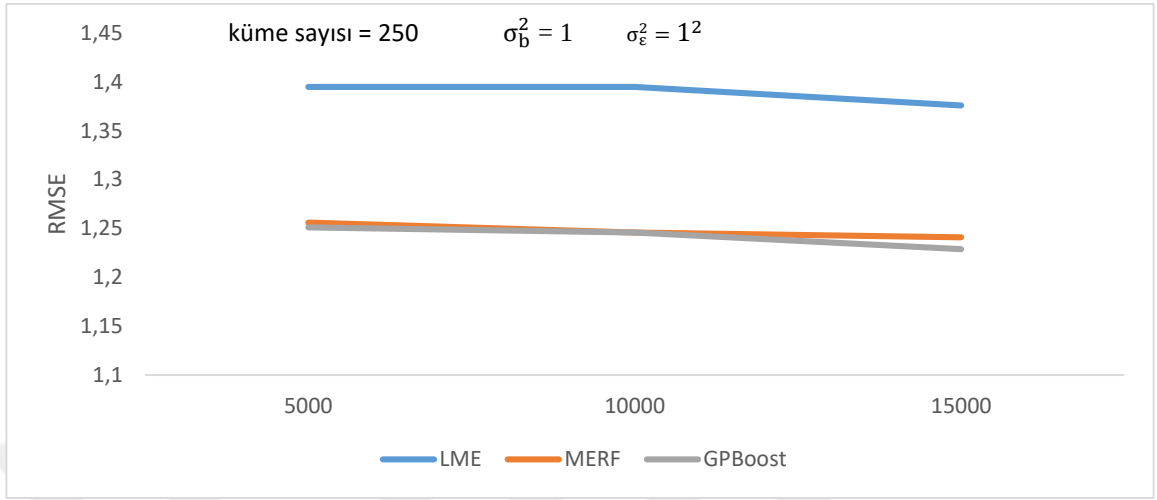
Şekil 3.4: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (m = 250)



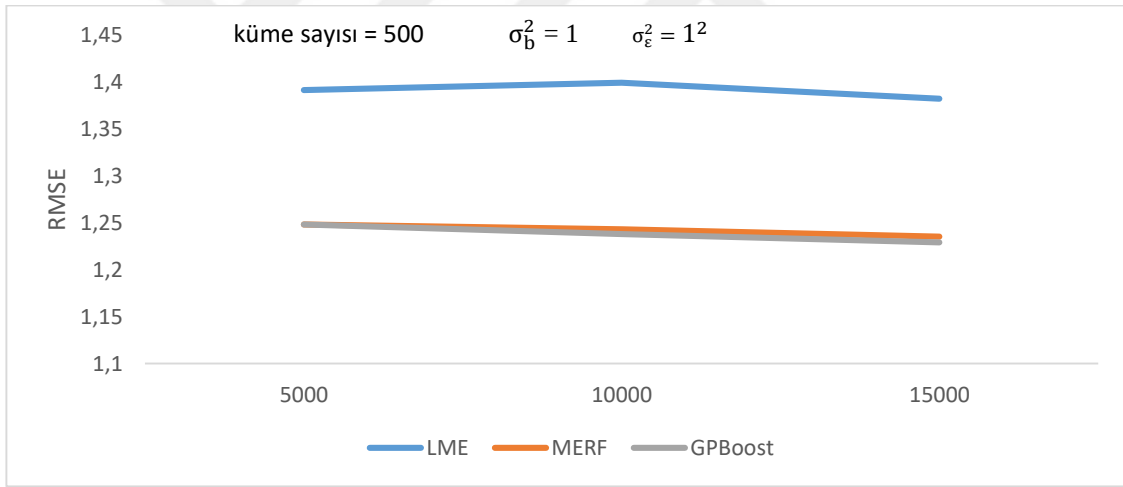
Şekil 3.5: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (m = 500)



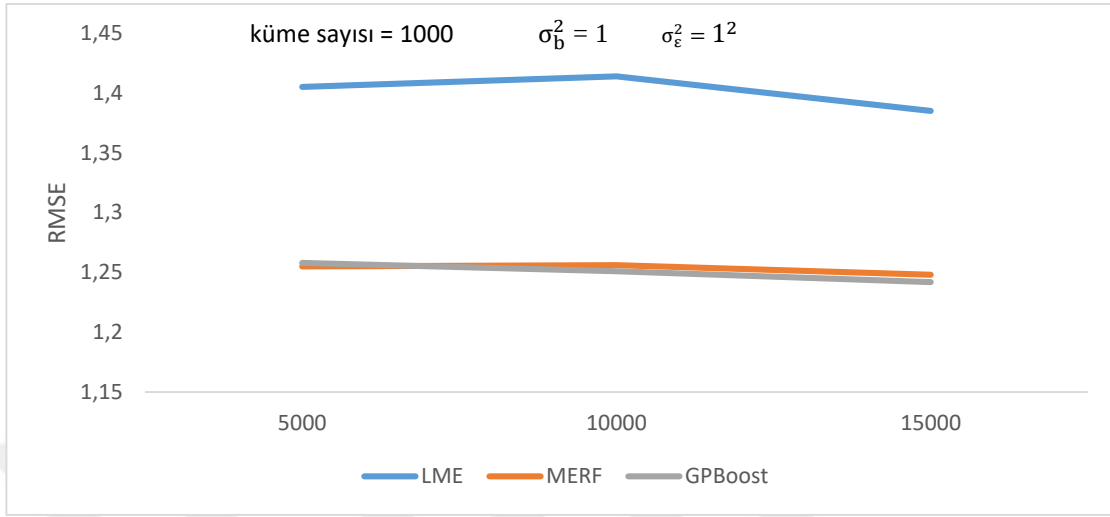
Şekil 3.6: 'doğrusal' fonksiyon için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (m = 1000)



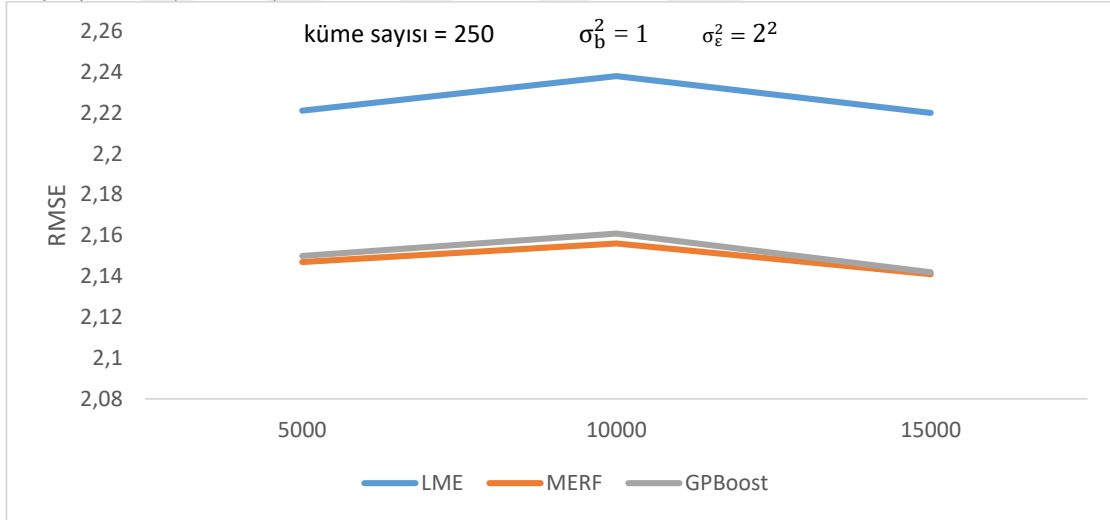
Şekil 3.7: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (m = 250)



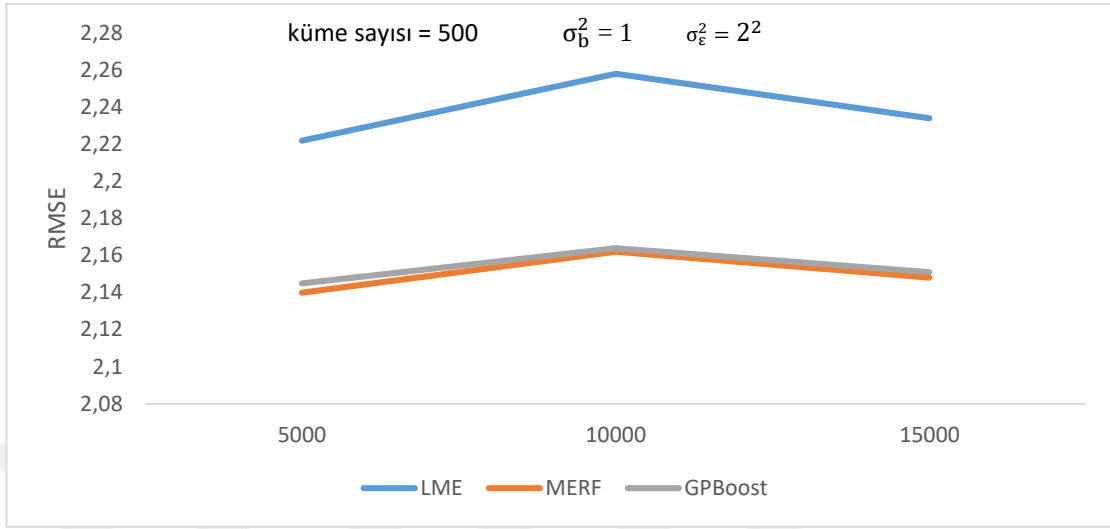
Şekil 3.8: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1^2$ için yöntem karşılaştırması (m =500)



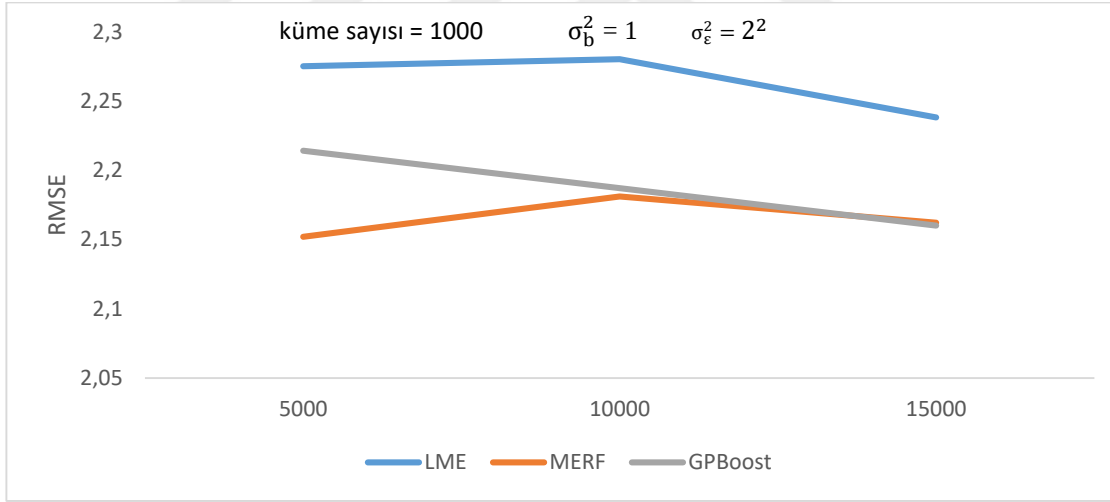
Şekil 3.9: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\epsilon^2 = 1^2$ için yöntem karşılaştırması (m=1000)



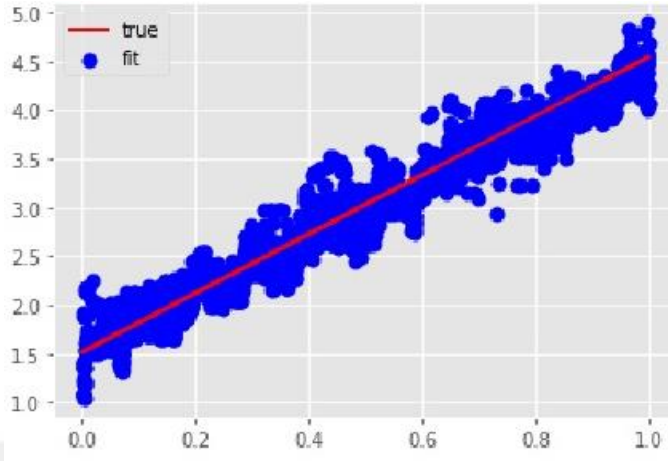
Şekil 3.10: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\epsilon^2 = 2^2$ için yöntem karşılaştırması (m=250)



Şekil 3.11: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (m=500)



Şekil 3.12: 'friedman3' fonksiyonu için RMSE değerlerine göre $\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 2^2$ için yöntem karşılaştırması (m = 1000)

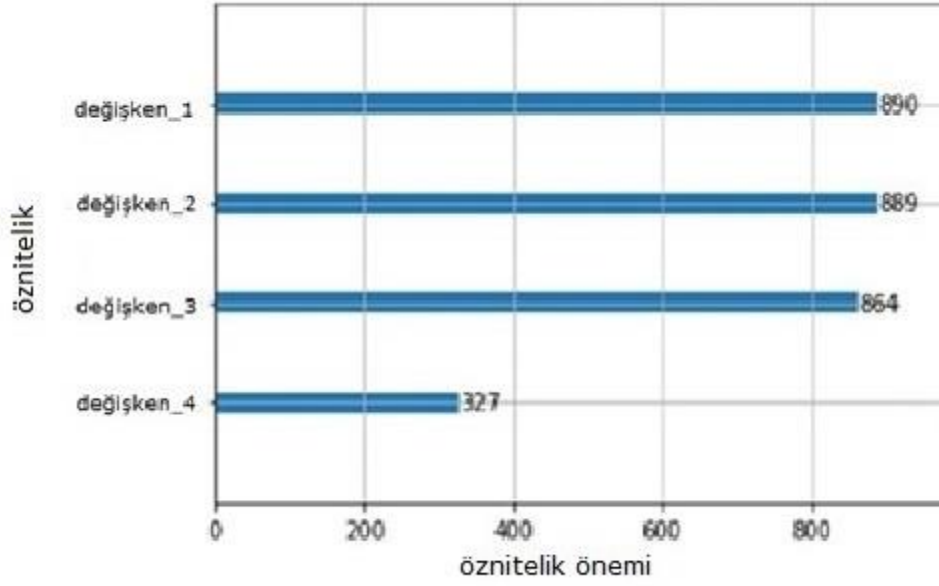


Şekil 3.13: 'doğrusal' fonksiyon için gerçek ve tahmin edilen sabit etkilerin karşılaştırılması ($\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1$, $n=5000$, küme $m = 250$)

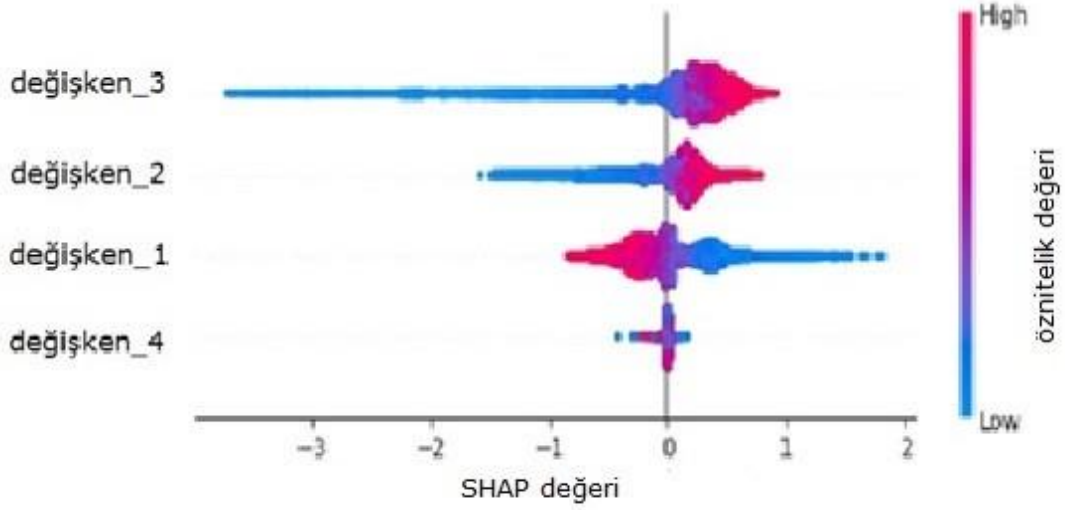
Şekil 3.13'te 'doğrusal' fonksiyon için gerçek ve tahmin edilen sabit etkilerin karşılaştırılması ($\sigma_b^2 = 1$ ve $\sigma_\varepsilon^2 = 1$, $n=5000$, küme sayısı=250) görülmektedir.

3.1. SHAP tekniği (SHapley Additive exPlanations)

Yanıt değişkeninin tahmin edilmesinde her bir özelliğin önemini değerlendirmek için SHAP tekniği kullanılmıştır (Lundberg ve Lee, 2017). SHAP tekniğinde her özelliğin tahmine katkısını açıklayan Shapley değerlerine dayalı olarak model yorumlanabilmektedir (Rodríguez-Pérez ve Bajorath, 2020). Şekil 3.14'e göre "değişken1" niteliğinin y yanıt değişkeninin belirlenmesinde tahmin modeline en fazla katkıda bulunduğu görülmektedir.



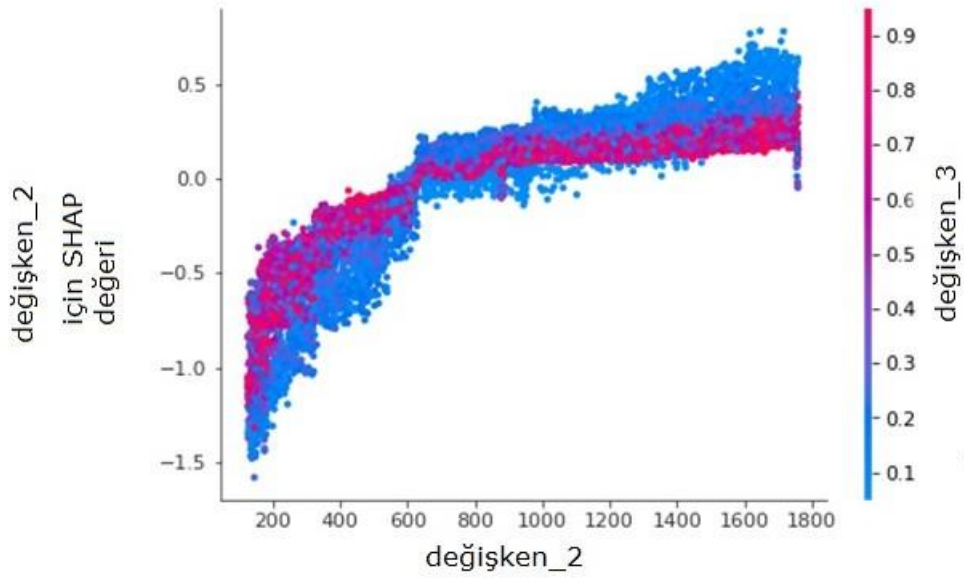
řekil 3.14: 'friedman3' fonksiyonu için (Eřitlik 2.2 için) SHAP deęerleri



řekil 3.15: 'friedman3' fonksiyonu (Eřitlik 2.2 için) SHAP özet grafięi

Bir deęiřkenin yüksek SHAP deęerine (negatif yönde olduęu için) sahip olması tahmini üzerinde yüksek ve pozitif deęerlerle iliřkili olması anlamı taşımaktadır (řekil 3.15)

Şekil 3.15'te 4 değişkenin önem sıralaması gösterilmiştir. SHAP özet grafiği, en önemli özellikleri ve bunların veri kümesi üzerindeki etkilerini göstermektedir. Bu grafik, 'friedman3' fonksiyonu kullanıldığında 'değişken 2' ve 'değişken 3' ile yanıt değişkeni arasında pozitif bir ilişki olduğunu göstermektedir. 'değişken 1' ile yanıt değişkeni arasında negatif bir ilişki olduğunu göstermektedir.



Şekil 3.16: 'friedman3' fonksiyonu (Eşitlik 2.2 için) SHAP etkileşim grafiği

Şekil 3.16'de 'değişken 2'nin 'değişken 3' ile sıklıkla etkileşime girdiği gösterilmektedir.

4. TARTIŞMA

Karışık etkili modellerde doğrusallık varsayımını esnetmek için genelleştirilmiş karışık etkili modeller kullanılması önerilmektedir. Bu, modelin hatalı kurulmasına neden olabilir. Bu nedenle kümelenmiş veya boylamsal veriler için gruplandırılmış rastgele etki modelleri, parametrik olmayan, makine öğrenmesine dayalı yaklaşımlar önerilmiştir. MERF modeli Manifold AI tarafından, GPBoost modeli ise Sigrist (2020) tarafından kullanılmıştır. GPBoost makine öğrenimi algoritması, hem Gauss süreci ve karışık etkiler modellerindeki doğrusallık varsayımını esnetmek için, hem de boosting konusundaki bağımsızlık varsayımını esnetmek için etkili bir yöntemdir. Karışık etkili makine öğrenmesi modeli doğrusal bir fonksiyon içeriyorsa çoğunlukla hem RMSE açısından hem de zaman(s) açısından LME yönteminin kullanımı GPBoost ve MERF yöntemine göre daha verimli olacaktır, ancak doğrusal olmayan bir fonksiyon içeriyorsa genellikle GPBoost yönteminin performansı diğer modellere kıyasla daha iyidir.

Elde ettiğimiz sonuçlara bakarsak doğrusal fonksiyon için hata varyans değeri 4, örneklem sayısı az ve küme sayısı fazla olduğunda, doğrusal model yerine MERF modelinin RMSE değeri daha küçük ve bu fark istatistiksel olarak anlamlı bulunmuştur. Bunun haricindeki tüm senaryolarda doğrusal modelin RMSE değerinin küçük ve bu farkın istatistiksel olarak anlamlı olduğu görülmüştür. Sigrist ise 2020 yılında yaptığı çalışmada küme sayısının az ve örneklem sayısının düşük olduğu her benzetim modelinde LME modelinin RMSE değerini daha küçük ve bu fark istatistiksel olarak anlamlı bulmuştur.

Doğrusal olmayan fonksiyon için hata varyans değeri 1, örneklem sayısının düşük ve küme sayısının yüksek olduğu durumlarda, GPBoost model yerine MERF modelinin RMSE değeri daha küçük ve bu fark istatistiksel olarak anlamlı bulunmuştur. Bunun haricindeki tüm senaryolarda (hata varyansı 1 iken) GPBoost modelinin RMSE

değerinin küçük, LME ile arasında farkın istatistiksel olarak anlamlı, MERF ile arasındaki farkın istatistiksel olarak anlamlı olmadığı görülmüştür. Sigrist ise 2020 yılında yaptığı çalışmada küme sayısının az ve örneklem sayısının düşük olduğu her benzetim modelinde GPBoost modelinin RMSE değerini küçük ve bu fark istatistiksel olarak anlamlı bulmuştur.

Doğrusal olmayan fonksiyon için hata varyans değeri 4, örneklem ve küme sayısının yüksek olduğu durumlarda, GPBoost model yerine MERF modelinin RMSE değeri daha küçük ve bu fark istatistiksel olarak anlamlı bulunmuştur. Bunun haricindeki tüm senaryolarda (hata varyansı 4 iken) GPBoost modelinin RMSE değeri küçük ve bu fark istatistiksel olarak anlamlı bulunmuştur.

Gerçek hayat verilerie uygulanabilmesi için benzetim çalışması ile örneklem büyüklüğü, küme sayısı ve hata varyans değerlerinin farklı şekillerde ele alındığı çalışmalar yapılmasına alanyazında ihtiyaç vardır.

5. SONUÇ VE ÖNERİLER

Gauss süreci ile karışık etkili modellerde, kümelenmiş etkiler için olasılıksal modeller oluşturulur, böylece tahmindeki belirsizliği açıklayan olasılıklı tahminler üretir (Sigrist, 2020).

GPBoost algoritmasını RMSE açısından diğer yaklaşımlarla karşılaştıran bağımlı gruplar için t-testlerinden p değerleri de hesaplanmıştır. Sonuçlar, yöntemler arasındaki farkın istatistiksel olarak önemli olduğunu göstermiştir ($p < 0,001$).

Zaman açısından bakıldığında MERF algoritmasının hızının diğerlerine göre çok yavaş olduğu gözlemlenmektedir, buna, düzgün tanımlanmış bir EM algoritması olmamasının neden olduğu düşünülmektedir (Sigrist, 2020).

Çalışmanın benzetim basamağında doğrusal olmayan bir fonksiyon için GPBoost, MERF ve LME yöntemine göre RMSE ve zaman açısından daha iyi bir performans göstermiştir. Ancak doğrusal bir fonksiyon ele alındığında LME daha iyi bir sonuç vermektedir.

Bu yöntem, tıp alanında; gen ve protein belirteçlerinin lokalizasyonu, hastalık patogenezi ve etiyolojisi hakkında ek çözüm, kısaca daha bilgilendirici ve umut verici tanısal karar vermeye katkı sağlar. Robotik alanında; hareket planlama, konum belirleme, lokalizasyon ve eşleme konularında etkin bir performans gösterebilir. Spor alanında; NBA free agency tahmini için geçmiş yıllardaki kullanımı sürdürülebilir (Natarajan, 2019).

ÖZET

Geleneksel Makine Öğrenmesi ve Karışık Etkili Makine Öğrenmesi Model Performanslarının Benzetim Çalışması ile Değerlendirilmesi

Kümelenmiş verilerin analizi için yaygın olarak klasik yaklaşım (genelleştirilmiş) doğrusal karışık etkili modeller (LME, GLMM) kullanılmaktadır. Bununla birlikte, doğrusal parametrik modellerin varsayımlarını sağlamak oldukça zordur. Klasik yöntemler, boylamsal verilerin alt grup gözlemlerinin yüksek derecede korelasyon sergilediği çalışmalarda düşük performansa yol açmaktadır. Ayrıca, bu yöntemler, verilerdeki doğal korelasyon yapısını hesaba katmadan kullanıldığından çıkarımların yanıltıcı olmasına yol açar.

Ağaç tabanlı algoritmaları kümelenmiş verilere genişletmek için makine öğrenme yöntemleri uygulanabilir. Bunlarda biri karışık etkili makine öğrenmesi modelidir. Bu model, iki regresyon ağacı yöntemini, doğrusal olmayan karışık etkili modelinin (NLMM) sabit ve rastgele etki bileşenlerini tahmin etmek için birleştirir. Hajjem (2011) tarafından önerilen genelleştirilmiş karışık etkiler regresyon ağaçları (GMERT) yapısı ve Sela (2012) tarafından önerilen rastgele etkiler beklenti maksimizasyonu (RE-EM) ağaç yapısı birlikte ele alınır. Türetilmiş veriler üzerinde geleneksel makine öğrenmesi modelleri ile yeni bir yaklaşım olarak makine öğrenmesi araçlarından karışık etkili makine öğrenmesi modelinin performansları benzetim uygulaması ile incelenecektir ve uygulama için örneklem büyüklükleri 5000, 10000, 15000 olarak belirlenmiştir. Gerçekleştirilecek işlemler için R programından veya Python programlama dilinden ve var olan hazır kütüphanelerden yararlanılacak, gerekli yerlerde yeni kod yazılacaktır.

Anahtar Kelimeler: Makine Öğrenmesi, Rastgele Etki, Benzetim, Kümelenmiş Veri

SUMMARY

Assessment of the Performance of Classic Standard Machine Learning Models and Mixed Effect Machine Learning Models via Simulation Study

Classical (generalized) linear mixed effect models (LME, GLMM) are widely used for the analysis of grouped data. However, it is quite difficult to provide assumptions of linear parametric models. Classical methods lead mediocre performance in studies where subgroup observations of longitudinal data exhibit high degree of correlations. Also, these methods are used for analysis of longitudinal data without accounting for the inherent correlation structure in the data often leading to potential for misleading inference.

Machine learning methods can be applied to extend tree-based algorithms to grouped data. One of them is the Mixed Effect Machine Learning model. This model combines the generalized mixed-effects regression trees (GMERT) structure for modeling general types of outcomes and the expectation maximization structure of the random-effects expectation maximization (RE-EM) tree for estimating the fixed and random-effects components of nonlinear mixed effect model (NLMM). The performances of the standard machine learning models and the mixed effects machine learning models will be analyzed by simulation study on the generated data. The sample sizes for the application were determined as 500, 1000, 15000. The R program and Python programming language and present libraries will be used for the analysis to be performed, and new codes will be written where necessary.

Keywords: Machine Learning, Random Effects, Simulation, Grouped Data

KAYNAKLAR

ABAR H (2020). XGBOOST ve MARS Yöntemleriyle Altın Fiyatlarının Kestirimi. *EKEV Akademi Dergisi*: 24(83).

ASHCROFT M. Advanced Machine Learning: Gaussian Processes. Erişim Adresi: [https://futurelearn.com/info/courses/advanced-machine-learning/0/steps/49565]. Erişim Tarihi: 07/04/2021.

ATEŞ B (2020). Gemi Yapılarında Gerilme Yığılması Öngörülerinin Kaba Ağ Yapısı ve Makine Öğrenmesi ile Gerçekleştirilmesi. İstanbul Teknik Üniversitesi. Gemi ve Deniz Teknolojileri Mühendisliği Ana Bilim Dalı Gemi ve Deniz Teknolojisi Mühendisliği Programı, Yüksek Lisans Tezi.

BANERJEE S, CARLIN BP, GELFAND AE (2003). *Hierarchical Modeling and Analysis For Spatial Data*. Chapman and Hall/CRC.

BASSETTI T, CARUSO R, SCHNEIDER F (2018). The Tree of Political Violence: A GMERT Analysis. *Empirical Economics*, **54** (2), 839-850.

BELTEMPO M, BRESSON G, LACROIX G (2020). Using Machine Learning to Predict Nosocomial Infections and Medical Accidents in a NICU.

BREIMAN L (1998). Arcing Classifiers. *Annals of Statistics*, 801-824.

BREIMAN L (2001). Random Forests. *Machine Learning*, **45** (1), 5-32.

BREIMAN L, FRIEDMAN J, STONE CJ, OLSHEN RA (1984). Classification and Regression Trees, CRC Press.

COCHRANE C, BA D, KLERNAN EB, HILAIRE MAS (2021). An Ensemble Mixed Effects Model of Sleep Loss and Performance. *Journal of Theoretical Biology*, 509, 110497.

COŞKUN K (2020). Ağ Saldırılarının Sınıflandırılmasında Karar Ağaçlarına Dayalı Arttırma (Boosting) Algoritmalarının Karşılaştırılması. Muğla Sıtkı Koçman Üniversitesi. Fen Bilimleri Enstitüsü, Bilişim Sistemleri Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi.

CRESSIE N, WIKLE CK (2015). Statistics for Spatio-Temporal Data. John Wiley & Sons.

DE'ATH G (2002). Multivariate Regression Trees: A New Technique for Modeling Species–Environment Relationships. *Ecology*, **83** (4), 1105-1117.

DEBRUINE LM, BARR DJ (2020). Understanding Mixed Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*.

DEMPSTER AP, LAIRD NM, RUBIN DB (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39** (1), 1-22.

DEMİRSÖZ M, ÖZEL Z, YONAR H, YONAR A, TEKİNDAL MA (2021). Karar Ağacına Dayalı Yöntemler. *Sağlık Bilişimi ve Veri Analitiği JAMOVİ ve JASP Uygulamaları*. Akademisyen Kitabevi A.Ş., Editör: TEKİNDAL MA, BURMAOĞLU S, 1, s:147 -177.

ESCABIAS CB (2017). Tree Boosting Data Competitions with XGBoost. Universitat Politècnica de Catalunya-Universitat de Barcelona, Yüksek Lisans Tezi. Erişim adresi: [http://hdl.handle.net/2117/100293].

- FOKKEMA M, SMITS N, ZEILEIS A, HOTHORN T, KELDERMAN H (2018). Detecting Treatment-Subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees. *Behavior Research Methods*, **50** (5), 2016-2034.
- FONTANA L, MASCI C, IEVA F, PAGANONI AM (2021). Performing Learning Analytics via Generalised Mixed-Effects Trees. *Data*, **6** (7), 74.
- FRIEDMAN JH (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 1-67.
- FRIEDMAN JH (2001). Greedy Function Approximation: a Gradient Boosting Machine, *Annals of Statistics*, 1189-1232.
- FU, W, SIMONOFF JS (2015). Unbiased Regression Trees for Longitudinal and Clustered Data. *Computational Statistics & Data Analysis*, 88, 53-74.
- FUGLSTAD GA, SIMPSON D, LINDGREN F, RUE H (2015). Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields? *Spatial Statistics*, **14**, 505-531.
- GALESHCHUK S, QIU J (2019). Identification of Opinion Makers on Twitter. International Conference on Data Science and Social Research. Springer, Cham, 187-198.
- GAMZE EK (2020). Prediction of Druggability Properties of Chemicals Using Machine Learning Techniques. Boğaziçi Üniversitesi. Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi.
- GEÇALP E (2020). Atm Demand Prediction with Machine Learning Methods. Galatasaray Üniversitesi. Fen Bilimleri Enstitüsü, Endüstri Anabilim Dalı, Yüksek Lisans Tezi.
- GLEN S. Root Mean Square Error, Erişim Adresi: [<https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/>]. Erişim Tarihi: 10.03.2021.
- GNEITING T, BALABDAOUI F, RAFTERY AE (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of The Royal Statistical Society: Series B (Statistical Methodology)*, **69** (2), 243-268.
- GÜVEN, CT (2020), Yapay Zekâ Yöntemleri Kullanılarak Akıllı Ev Sistemi Geliştirilmesi. Mersin Üniversitesi. Fen Bilimleri Enstitüsü. Bilgisayar Mühendisliği Ana Bilim Dalı, Yüksek Lisans Tezi.
- HAJJEM A, BELLAVANCE F, LAROCQUE D (2011). Mixed Effects Regression Trees for Clustered Data. *Statistics & Probability Letters*, **81** (4), 451-459.
- HAJJEM A, BELLAVANCE F, LAROCQUE D (2014). Mixed-Effects Random Forest for Clustered Data. *Journal of Statistical Computation and Simulation*, **84** (6), 1313-1328.
- HAJJEM A, LAROCQUE D, BELLAVANCE F (2017). Generalized Mixed Effects Regression Trees. *Statistics & Probability Letters*, **126**, 114-118.
- HOTHORN T, BÜHLMANN P, KNEIB T, SCHMID M, HOFNER B (2010). mboost: Model-Based Boosting, URL <http://r-forge.r-project.org/projects/mboost>. R package version 2.1-0.
- HOTHORN T, BÜHLMANN P, KNEIB T, SCHMID M., HOFNER B. (2010). Model-Based Boosting 2.0. *Journal of Machine Learning Research*, **11**, 2109-2113.
- JONES DR, SCHONLAU M, WELCH WJ (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, **13** (4):455-492.
- KARAKAŞ, E (2019). Derin Sinir Ağlarının Füzyonu ile Yüz İmgelerinden Yaş Grubu ve Cinsiyet Sınıflandırma. Atatürk Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği ABD, Yüksek Lisans Tezi.

KENNEDY MC, O'HAGAN A (2001). Bayesian Calibration of Computer Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63** (3), 425-464.

KHAIR U, FAHMI H, AL HAKIM S, RAHIM R (2017). Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error. In *Journal of Physics: Conference Series* (Vol. 930, No. 1, p. 012002). IOP Publishing.

KNAGG O (2019). An Intuitive Guide to Gaussian Processes. Erişim Adresi: [https://towardsdatascience.com/an-intuitive-guide-to-gaussian-processes-ec2f0b45c71d]. Erişim Tarihi: 07/12/2020.

KUBO Y, WATANABE S, NAKAMURA A, MCDERMOTT E, KOBAYASHI T (2010). A Sequential Pattern Classifier Based on Hidden Markov Kernel Machine and its Application to Phoneme Classification. *IEEE Journal of Selected Topics in Signal Processing*, **4** (6), 974-984.

KÜKNER CU (2020). A Comparative Analysis of LSTM and XGBOOST Methods for day Ahead Electricity Price Forecasting. İstanbul Teknik Üniversitesi. Enerji Enstitüsü, Enerji Bilim ve Teknoloji Anabilim Dalı, Yüksek Lisans Tezi.

LAIRD NM, WARE JH (1982). Random-effects Models for Longitudinal Data. *Biometrics*, 963-974.

LEVY J, BOBAK C, NASIR-MOIN M, VEZIROGLU EM, PALISOUL SM, BARNEY RE, SALAS AL, CHRISTENSEN BC, TSONGALIS GJ, VAICKUS LJ (2021). Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction Using Spatially Localized Immuno-Oncology Markers. *BioRxiv*.

LUNDBERG SM, LEE SI (2017). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768-4777.

MATHERON G (1963). Principles of Geostatistics. *Economic Geology*, **58** (8), 1246-1266.

merf package (v0.3; Manifoldai, 2019).

MUKADAM M, YAN X, BOOTS B (2016). Gaussian Process Motion Planning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 9-15, IEEE.

MURATLAR ER (2020). CatBoost Nedir? Diğer Boosting Algoritmalarından Farkı Nelerdir? Erişim Adresi: [https://www.veribilimiokulu.com/catboost-nedir-diger-boosting-algoritmalarindan-farki-nelerdir/]. Erişim Tarihi: 07/12/2020.

NATARAJAN SS (2019). A Mixed Effects Modeling Approach to Predicting NBA Free Agency. Erişim Adresi: [https://hockeygraphsdotcom.files.wordpress.com/2019/09/mixedfa_model_deck-senthil-v2.pdf]. Erişim Tarihi: 07/09/2020.

NGUFOR C, VAN HOUTEN H, CAFFO BS, SHAH ND, MCCOY RG (2019). Mixed Effect Machine Learning: a Framework for Predicting Longitudinal Change in Hemoglobin A1c. *Journal of Biomedical Informatics*, **89**, 56-67.

PINHEIRO J, BATES D (2000). *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag, New York.

R Core Team (2013). R: A Language and Environment for Statistical Computing, <http://www.r-project.org/>.

RODRIGUEZ-PEREZ R, BAJORATH, J (2020). Interpretation of Machine Learning Models Using Shapley Values: Application to Compound Potency and Multi-Target Activity Predictions. *Journal of Computer-Aided Molecular Design*, **34** (10), 1013-1026.

SAİTOĞLU SY (2015). Sınıflama ve Regresyon Ağaçları. Marmara Üniversitesi. Sosyal Bilimler Enstitüsü. Ekonometri ABD, Doktora Tezi.

- SAYIN G (2019). Halka Açık Finans Dışı Şirketlerde Süreklilik Riskinin Makine Öğrenmesi ile Öngörülmesi. İstanbul Aydın Üniversitesi. Sosyal Bilimler Enstitüsü, Muhasebe ve Finansal Yönetim ABD, Doktora Tezi.
- SCHAPIRE RE (1999). A Brief Introduction to Boosting. *In Ijcai* (Vol. 99, 1401-1406).
- SCHMIDT AM, GUTTORP P (2020). Flexible Spatial Covariance Functions. *Spatial Statistics*, 100416.
- SEGAL MR (1992). Tree-structured Methods for Longitudinal Data, *Journal of the American Statistical Association*, **87** (418) 407-418.
- SELA RJ, SIMONOFF JS (2009). RE-EM trees: a New Data Mining Approach for Longitudinal Data. *NYU Stern Working Paper* SOR-2009-03.
- SELA RJ, SIMONOFF JS (2011). REEMtree: Regression Trees with Random Effects. R Package version 0.90, **3**, 741-749.
- SELA RJ, SIMONOFF JS (2012). RE-EM trees: a Data Mining Approach for Longitudinal and Clustered Data. *Machine Learning*, **86** (2) 169-207.
- SHUMWAY RH, STOFFER DS (2000). Time Series Analysis and its Applications (Vol. 3). New York: Springer.
- SHUMWAY RH, STOFFER DS (2017). Time Series Analysis and its Applications: with R Examples. Springer.
- SIGRIST F (2020). Gaussian Process Boosting. arXiv preprint arXiv:2004.02653.
- SNOEK J, LAROCHELLE H, ADAMS RP (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *In Advances in Neural Information Processing Systems*: 2951-2959.
- SPEISER JL, WOLF BJ, CHUNG D, KARVELLAS CJ, KOCH DG, DURKALSKI VL (2020). BiMM tree: a Decision Tree Method for Modeling Clustered and Longitudinal Binary Outcomes. *Communications in Statistics-Simulation and Computation*, **49** (4), 1004-1023.
- STROUP WW (2012). Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC Press.
- ŞAHINARSLAN, FV (2019). Makine Öğrenmesi Algoritmaları ile Nüfus Tahmini: Türkiye Örneği. İstanbul Teknik Üniversitesi. Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, Yüksek Lisans Tezi.
- UYSAL G, MONTERO AR, ŞORMAN AA, ŞENSOY A (2019). Dağlık Bir Havza için Hidrolojik Modelleme ile Elde Edilen Akım Tahminlerinin Ensemble Kalman Filtresi ile İyileştirilmesi. Muğla Sıtkı Koçman Üniversitesi. 10. Ulusal Hidroloji Kongresi, Cilt II, 949-957.
- VERBEKE G, MOLENBERGHS G (2000). Linear Mixed Models for Longitudinal Data. New York John Wiley Sons.
- WHITTLE P (1963). Prediction and Regulation by Linear Least-Square Methods. English Universities Press.
- WILLIAMS CK, RASMUSSEN CE (2006). Gaussian Processes for Machine Learning (Vol. 2, No.3, p. 4). Cambridge, MA: MIT press.
- WU H, ZHANG JT (2006). Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-effects Modeling Approaches. Wiley, New York.
- YANGIN G (2019). XGBoost ve Karar Ağacı Tabanlı Algoritmanın Diyabet Veri Setleri Üzerine Uygulaması. Mimar Sinan Güzel Sanatlar Üniversitesi. Fen Bilimleri Enstitüsü. Yüksek Lisans Tezi.

YILDIRIM MZ (2017). Makine Öğrenmesi Yöntemleri ile Network Üzerinde Saldırı Tespiti. Karabük Üniversitesi. Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Yüksek Lisans Tezi.

YILMAZ H (2014). Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama. Eskişehir Osmangazi Üniversitesi. Sağlık Bilimleri Enstitüsü, Biyoistatistik ABD, Yüksek Lisans Tezi.

ZEILEIS A, HOTHORN T, HORNIK K (2008). Model-based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17** (2), 492–514.

ZHANG H (1998). Classification Trees for Multiple Binary Responses, *Journal of the American Statistical Association*, **93** (441), 180-193.

ZHOU F, ALSAID A, BLOMMER M, CURRY R, SWAMINATHAN R, KOCHHAR D, TIJERINA L (2021). Predicting Driver Fatigue in Monotonous Automated Driving with Explanation using GPBoost and SHAP. *International Journal of Human–Computer Interaction*, 1-11.