

PAPER DETAILS

TITLE: Vücut Yag Yüzdesi Tahmini İçin Özellik Seçim Yöntemlerinin Karsilastirilmesi

AUTHORS: Asude Altiparmak Bilgin,Burhan Barakli

PAGES: 2068-2093

ORIGINAL PDF URL: <https://dergipark.org.tr/tr/download/article-file/2423939>



Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Vücut Yağ Yüzdesi Tahmini İçin Özellik Seçim Yöntemlerinin Karşılaştırılması

 Asude ALTIPARMAK BİLGİN ^{a,*},  Burhan BARAKLI ^b

^a Elektrik Elektronik Mühendisliği Bölümü, Mühendislik Fakültesi, Sakarya Üniversitesi, Sakarya, TÜRKİYE

^b Elektrik Elektronik Mühendisliği Bölümü, Mühendislik Fakültesi, Sakarya Üniversitesi, Sakarya, TÜRKİYE

* Sorumlu yazarın e-posta adresi: asude6parmak@gmail.com

DOI:10.29130/dubited.1115703

Öz

Çağımızın yaygın olarak görülen sağlık problemlerinden biri olan obezite, kişinin yaşam kalitesine olumsuz etkisinin yanında birçok rahatsızlığa da sebep olmaktadır. Vücut yağ yüzdesi, obezitenin teşhis edilmesinde en önemli göstergedir. Vücut yağ yüzdesinin hızlı, kolay, maliyetsiz ve yüksek doğruluk ile belirlenmesi ise en az obezitenin teşhis edilebilmesi kadar önemlidir. Antropometrik verilerden hesaplanabilen vücut yağ yüzdesi değerini makine öğrenmesi algoritmaları ile güvenli bir şekilde hesaplamak mümkündür. Ancak yüksek boyutlu, alakasız ve gereksiz veriler makine öğrenmesi algoritmalarının doğruluğunu saptırmakta ve modelin eğitim süresini arttırmaktadır. Makine öğrenmesi algoritmalarını daha az özellik ile kullanarak daha yüksek doğruluğun elde edilmesini sağlayan özellik seçim algoritmaları bulunmaktadır. Bu çalışmada vücut yağ yüzdesi tahmini için yedi farklı özellik seçim algoritmaları karşılaştırılıp daha az özellik ile daha yüksek doğrulukta sonuçların elde edilmesi sağlanmıştır. Özellik seçim yöntemlerinin farklı modellere etkisini incelemek için dört makine öğrenmesi yöntemi kullanılmıştır. Bu makine öğrenmesi algoritmalarının eğitim süreleri karşılaştırılmıştır. Deneysel çalışmalar sonucunda özellik seçim yöntemleri kullanılarak daha az özellik ile modelin eğitimi için daha kısa süre harcanarak daha yüksek doğrulukta tahminler elde edilebileceği gösterilmiştir.

Anahtar Kelimeler: Özellik seçimi, Makine öğrenmesi, Vücut yağ yüzdesi

Comparison of Feature Selection Methods for Estimation of Body Fat Percentage

ABSTRACT

Obesity, which is one of the common health problems of our age, causes many discomforts as well as its negative impact on the quality of life of the person. Body fat percentage is the most important indicator in diagnosing obesity. Determining body fat percentage quickly, easily, inexpensively and with high accuracy is as important as diagnosing obesity. It is possible to reliably calculate the body fat percentage estimation, which can be calculated from anthropometric data with machine learning algorithms. However, high-dimensional, irrelevant and redundant data distort the accuracy of machine learning algorithms and increase the training time of the model. There are feature selection algorithms that provide higher success by using machine learning algorithms with fewer features. In this study, seven different feature selection algorithms for body fat percentage estimation have been compared and higher accuracy results have been obtained with fewer features. Four machine learning methods have been used to examine the effect of feature selection methods on different models. The training times of these machine learning algorithms have been compared. As a result of experimental studies, it has been shown that by using feature selection methods, higher accuracy predictions can be obtained by spending less time on training the model with fewer features.

Keywords: Feature selection, Machine learning, Body fat percentage

I. GİRİŞ

Obezite, önemli bir sağlık problemidir. Son yıllarda hızla yaygınlaşmakta ve sadece gelişmiş ülkelerde değil, gelişmekte olan ülkelerde de hızla yayılarak küresel bir salgın haline gelmiştir. ABD gibi Kuzey Amerika ülkelerinde yetişkin nüfusunun üçte birinin obez olduğu bildirilmiştir. Çin’de son 8 yılda obezite oranı erkeklerde 3 katına, kadınlarda da 2 katına çıkmıştır. Obezite, sosyal yaşama olan etkilerinin yanı sıra kalp-damar hastalıkları, kanser, hipertansiyon, diyabet, gluko/metabolic sendrom [1] gibi birçok farklı hastalık türüne de neden olmaktadır. İhracat maliyetini düşürmek, daha lezzetli hale getirmek gibi amaçlarla yiyeceklerin içerisine eklenen yan ürünler, olması gerekenden daha yüksek kalorili besinlerin tüketilmesine yol açmaktadır. Ayrıca teknolojinin gelişimi ile harcanan enerji miktarı ve günlük alınan kalori miktarı arasındaki dengesizlikte obezitenin yaygınlaşmasının önemli bir nedeni olmuştur. Çağımızın önemli hastalıkları arasında olan obezite, diyabet, hipertansiyon, kanser gibi insan hayatını olumsuz etkileyen birçok ölümcül hastalığın oluşmasına neden olan ciddi bir problemdir. Giderek yaygınlaşan obezite sadece yetişkinler için değil, çocuklar için de bir tehlike haline gelmektedir [2], [3].

Ağırlık vücut yağı hakkında bilgi sağlayan ölçülmesi kolay bir nicelik olsa da genel vücut yapısı hakkında ilgi olmadığında bir anlam ifade etmez. Dünya Sağlık Örgütü tarafından ağırlık ve boy bilgileri ile hesaplanan vücut kitle indeksinin (BMI) obeziteyi tanımlamada kullanılabileceği onaylanmıştır. BMI genel beslenme durumunun bir göstergesidir. BMI, vücut ağırlığının vücut boyunun karesine oranı olarak ifade edilir. BMI 25 kg/m^2 değerinden büyük olduğu durumlar aşırı kiloluğa karşılık gelirken 30 kg/m^2 olduğu durumlar obeziteye karşılık gelmektedir. Ancak çalışmalar gösteriyor ki vücut yağ yüzdesi (BFP) obezite ve obeziteden kaynaklanan sağlık problemleri hakkında BMI değerinden daha doğru bilgi sağlamaktadır. BFP yağ kütlesi hakkında bilgi verirken BMI vücut yağ kütlesinin bir göstergesi değildir [4], [5].

BFP, toplam vücut yağı ölçümüdür. Omurga ve kalça taramalarına dâhil edilen yumuşak doku bölgeleri için BFP değerlerinin ortalaması alınarak toplam vücut yağlanma indeksi elde edilir. BFP, obezitenin doğru teşhisi için gereklidir. Ayrıca yağ kütlesi hakkında bilgi verir ve kalp damar hastalıkları, diyabet, metabolik hastalıklar ile ilgili riskleri tahmin etmede BMI değerine göre daha iyi bir göstergedir [5], [6].

BMI, boy ve ağırlık ile değerlendirildiğinden vücut yağı ve ölüm oranı arasındaki ilişkiyi belirlemede yeterli bir ölçüt değildir. Yapılan çalışmalarda düşük vücut kitle endeksi olmasına rağmen yüksek BFP değerine sahip bir bireyin tüm nedenlere bağlı olarak ölüm oranının yüksek olduğu saptanmıştır. BFP değeri ve ölüm riski arasındaki ilişkiyi belirlemek için daha detaylı vücut ölçümleri gerekmektedir. Yani bir birey düşük BMI değerine sahip olsa bile ölüm riski yüksek olabilir [6].

BFP tahmini için antropometri (vücut kütlesi, vücuttaki belli bölgelerin çevresi, deri kıvrım kalınlığı vb.), su altı tartımı (UWW), X-ışını absorpsiyometrisi (DEXA), biyoelektrik empedans analizi, manyetik rezonans görüntüleme, havada yer değiştirme pletismografisi ve yakın kızılötesi etkileşim gibi sayısız teknik vardır. DEXA ve UWW teknikleri antropometrik ölçümlerden vücut yağı için daha doğru tahminler gerçekleştirir. Ancak bu ölçümleri gerçekleştirmek yüksek maliyetli ve kullanımları sınırlıdır. Çeşitli formüller ile yaş, cinsiyet ve BMI değerlerinden BFP tahmini hem düşük maliyetlidir hem de tıbbi bir cihaz gerektirmez. Bu yüzden geniş çapta kullanımı uygun görülmüştür [5].

BFP tahmini, makine öğrenimi (ML) algoritmaları ile gerçekleştirilebilmekte ve bir regresyon problemi olarak ele alınmaktadır [7]. Literatürde BFP tahmini ile ilgili çalışmalar, genellikle yaş, boy, ağırlık, çeşitli vücut ölçümleri gibi ölçülmesi kolay antropometrik verileri kullanarak ML metotları ile tahmin gerçekleştirmeyi amaçlamıştır [8]. Bu çalışmalar genellikle herhangi bir tıbbi cihaz kullanmadan düşük maliyet ile yüksek doğrulukta tahminler gerçekleştirmeyi amaçlamıştır.

Son dönemlerde yüksek veri boyutunun ML modellerinin performansını ve iş yükünü olumsuz etkilemesinden dolayı özellik seçim yöntemleri önemli bir hale gelmiştir. ML algoritmalarının

kullandığı verilerin boyutu tahmin doğruluğu hesaplama yükü gibi birçok açıdan önemlidir. Gereksiz, alakasız gürültülü verilerin varlığı model performansına olumsuz etkisinin yanı sıra hesaplama yükünü arttırmakta eğitim süresinin uzamasına neden olmaktadır. Bu yüzden özellik seçim (FS) algoritmalarının kullanımı önemli bir veri ön işlem aşamasıdır [9]. Literatürde FS yöntemleri kullanarak ML algoritmaları ile BFP tahmini gerçekleştiren çeşitli çalışmalar bulunmaktadır.

Baraklı B. ve Küçüker A. çalışmalarında FS yöntemlerinin ve özellik azaltımının BFP tahmini için regresyon yöntemlerinin başarısına olan katkısını incelemişlerdir. FS için tek değişkenli doğrusal regresyon ve özellik azaltımı için ise temel bileşenler analizi (PCA) yöntemlerini kullanmışlardır. Destek vektör makineleri (SVM), rastgele orman ağaçları (RF) makine öğrenmesi (ML) modellerini kullanarak BFP değerini tahmin etmişlerdir [7].

Uçar M. K. vd. çalışmalarında en az veri kullanarak yüksek doğruluğa sahip BFP tahmini hibrit ML modelleri kullanarak gerçekleştirmeyi amaçlamışlar. Uygulamada kullanılan PCA analizi temelli FS algoritması ile farklı özellik alt grupları oluşturarak önerilen hibrit modeller ile durumları test etmişlerdir. SVM, çok katmanlı ileri beslemeli sinir ağları, karar ağacı (DT) ML modellerini kullanmanın yanı sıra bu modellerden elde edilen 4 farklı hibrit model ile birlikte toplam 7 ML modeli kullanarak deneysel sonuç elde etmişlerdir [8].

Kupusinac A. vd. çalışmalarında daha yüksek doğruluğa sahip BFP tahmini amacıyla yapay sinir ağları tekniğini kullanarak yeni bir yaklaşım göstermişlerdir. BFP değerini cinsiyet, yaş ve BMI değerlerinden hesaplayan formüller ile karşılaştırdığında Kupusinac ve arkadaşlarının önerdiği metot, benzer maliyet ve komplekslik göstermiş olsa da daha yüksek doğrulukta BFP tahmini gerçekleştirmiştir [10].

Shao Y.E. çalışmasında BFP tahmin çalışmasında daha az değişken kullanarak modelinin daha iyi bir tahmin yapmasını amaçlayan yeni bir hibrit metot önermiştir. Hibrit modeli çoklu regresyon (MR), yapay sinir ağları, çok değişkenli uyarlanabilir regresyon eğrileri (MARS) ve destek vektör regresyon (SVR) tekniklerini içermektedir. Modelleme ilk aşamada daha önemli olan değişkenleri seçmek için MR ve MARS'ın kullanımını içermiştir. İkinci aşamada kalan önemli değişkenler diğer tahmin modellerinde BFP tahmini için kullanmıştır. Önerilen bu hibrit model diğer tek aşamalı modellere göre daha iyi tahmin yaptığını sonuçlarda göstermiştir [11].

Ferenci T. vd. çalışmalarında doğrusal regresyon, ileri beslemeli sinir ağı ve SVM tekniklerini kullanmışlardır. Önyükleme doğrulaması ile en uygun parametreleri seçmişlerdir. Sonuçlarda az bir fark ile SVM tekniğinin diğer iki tekniğe göre daha doğru tahmin yaptığını göstermişlerdir [12].

Keivanian F. Vd çalışmalarında, çok katmanlı algılayıcı (MLP) yapay sinir ağı (ANN) tahmin modeli kullanarak FS için yeni bir bulanık uyarlamalı ikili küresel öğrenme kolonizasyon yöntemi önermişlerdir. Önerilen metot ile karşılaştırma yapılması amacıyla MLP ile hibritleyerek bazı iyi bilinen meta-sezgisel yöntemin ikili versiyonları çalışmaya dâhil etmişlerdir. Çalışmalarına metotların çalışma sürelerini, seçilen özellik sayılarını ekleyip değerlendirmişlerdir [13].

Chiong R. vd. çalışmalarında, biri antropometrik ölçümleri içeren diğeri fiziksel inceleme ve laboratuvar ölçümleri içeren 2 ayrı veri seti kullanarak BFP tahmini için göreceli destek vektör makineleri (IRE-SVM) yaklaşımı geliştirmiştir. Önerilen metot yansız bir tahmin modeli elde etmek için amaç fonksiyonuna yanlış hata kontrol terimi eklemesini kapsamaktadır. Ayrıca anlamlı bilgiyi kaybetmeden gereksiz ve ilgisiz özelliklerin kaldırılmasını içeren FS tekniği uygulamışlardır. SVM, RE-SVM, ANN'nin bir tipi olan MLP, RF, aşırı gradyan artırma regresyon (XGBR) olmak üzere 5 farklı tahmin modelini çalışmada önerilen model ile karşılaştırmak için kullanmışlardır [14].

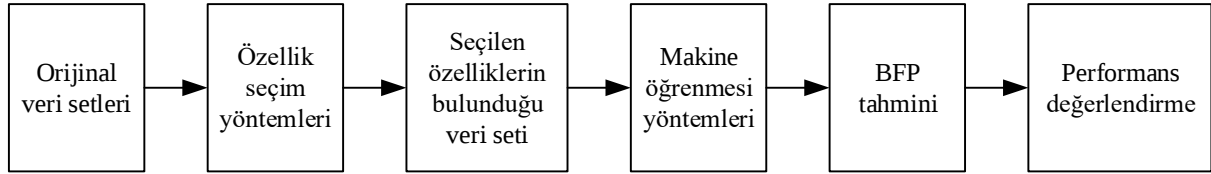
Hussain A. S vd. çalışmalarında, BFP tahmini için SVM ve duygusal yapay sinir ağları (EANN) metotuna dayalı hibrit bir model önermişlerdir. Bu hibrit model, SVM kullanarak özellik seçmekte ve EANN ile BFP tahmini gerçekleştirmektedir. Yüksek tahmin oranları elde etmek ve hedef değişkeni

etkileyen önemli faktörleri belirlemek için sinir ağının temel yapısını çeşitli duygusal işlevlerle birleştirmişlerdir [15].

Bu çalışmada literatürdeki benzer çalışmalardan da yola çıkarak BFP tahmini için karşılıklı bilgi (MI), sıralı ileriye doğru (SFS), sıralı geriye doğru (SBS), rastgele orman ağaçları makine öğrenim algoritması kullanılarak özelliklerin önemine dayalı (RFI), rastgele orman ağaçları makine öğrenim algoritması kullanılarak özelliklerin önemine dayalı yinelemeli (RRFI), rastgele karıştırma (RS), yinelemeli özellik eleme (RFE) özellik seçim yöntemleri olmak üzere 7 FS algoritmasının modelin performansına, eğitim süresine olan etkisi incelenmiştir. FS yöntemleri ile daha az sayıda özellik uzayına sahip olan yeni veri setlerine 4 regresyon modeli uygulanmıştır. Bunlar: RF, SVM, gradyan artırma regresyon (GBR) ve XGBR modelleridir. Performans değerlendirmesi için literatürde sıkça kullanılan ortalama karesel hata (MSE), belirleme katsayısı (R^2), ortalama mutlak yüzde hatası (MAPE) ve medyan mutlak hata (MAE) olmak üzere 4 performans metriği kullanılmıştır.

II. GEREÇ VE YÖNTEMLER

Bu çalışmada BFP tahmini için 2 veri setine çeşitli FS yöntemleri uygulanmıştır. Veri setlerine FS yöntemleri uygulanarak özellikler seçilmiştir. Seçilen özelliklere farklı regresyon yöntemleri uygulanarak BFP tahmin performansları incelenmiştir. Çalışmanın genel blok diyagramı Şekil 1 ile verilmiştir. Regresyon metodlarının daha doğru tahminler yapmasını sağlamak için parametre ayarlamaları yapılmıştır. En uygun parametrelerin bulunması ve doğrulanması amacıyla ızgara arama yöntemi ve k katmanlı çapraz doğrulama yöntemi kullanılmıştır.



Şekil 1. Özellik seçim işlemleri için temel adımlar

A. VERİ SETİ

Bu çalışmada 2 adet veri seti kullanılmıştır. Biri 248 kişiden toplanan 13 farklı antropometrik ölçümden elde edilen orijinal veri seti (VS1), diğeri ise VS1'deki özelliklere istatistiksel yöntemler kullanarak 25 özellik daha eklenmiş 38 özellikli ikinci veri seti (VS2) kullanılmaktadır.

VS1 veri setindeki özelliklerin ikili korelasyonunu gösteren renkli ısı haritası Şekil 2'de verilmiştir. Birbiri arasında yüksek ilişkiye sahip özellikler mevcuttur. Örneğin bağımsız özelliklerden göğüs çevresi ile abdomen çevresi, ağırlık ile kalça çevresi arasında diğerlerine göre daha yüksek ilişki bulunmaktadır. Bağımlı değişken BFP ile abdomen çevresi arasında yüksek ilişki vardır.

Şekil 3'te de VS1 veri setindeki özellikler arasındaki tüm ikili ilişkiler farklı bir açıdan verilmiştir. Bu gösterimden, özellikler arasında doğrusal ve doğrusal olmayan ilişkilerin bulunduğu görülmektedir. Örneğin abdomen çevresi ile kalça çevresi arasında doğrusal bir ilişki var iken boy, yaş gibi özelliklerin genellikle diğer özellikler ile doğrusal olmayan bir ilişkisi mevcuttur. Ayrıca, ağırlık özelliğinin diğer özelliklerle daha yüksek bir ilişkisi var iken boy özelliği diğer özelliklerle daha zayıf ilişki göstermektedir.

Tablo 1. VS2 veri setindeki 25 özelliğin BFP ile ilişki katsayıları

Ad	Özellik	BFP ile İlişkileri	Ad	Özellik	BFP ile İlişkileri
c14	Kurtosis - Basıklık	-0,789	c27	Merkezi Moment	-0,453
c15	Skewness - Çarpıklık	-0,795	c28	Ortalama değeri	0,684
c16	Çeyrekler arası genişlik	0,751	c29	Ortalama Eğri Uzunluğu	0,345
c17	Değişim Katsayısı	-0,603	c30	Ortalama Enerji	0,627
c18	Geometrik ortalama	0,68	c31	Ortalama Karakök RMS değeri	0,635
c19	Harmonik ortalama	0,629	c32	Standart hata	0,4
c20	Hijort Aktivite Katsayısı	0,397	c33	Standart Sapma	0,4
c21	Hijort Hareketlilik Katsayısı	-0,525	c34	Şekil Faktörü	0,518
c22	Hijort Karmaşıklık Katsayısı	-0,437	c35	Tekil Değer Ayrışımı	0,635
c23	Maksimum değeri	-0,048	c36	"%25" için kesilmiş ortalama değeri	0,729
c24	Medyan değeri	0,32	c37	"%50" için kesilmiş ortalama değeri	0,725
c25	Mutlak Sapma	0,598	c38	5 Ortalama Teager Enerjisi	0,22
c26	Minimum değeri	0,342	BFP	Vücut Yağ Yüzdesi	1

Şekil 2. Özellikler arasındaki korelasyon ilişkisini gösteren renkli ısı haritası

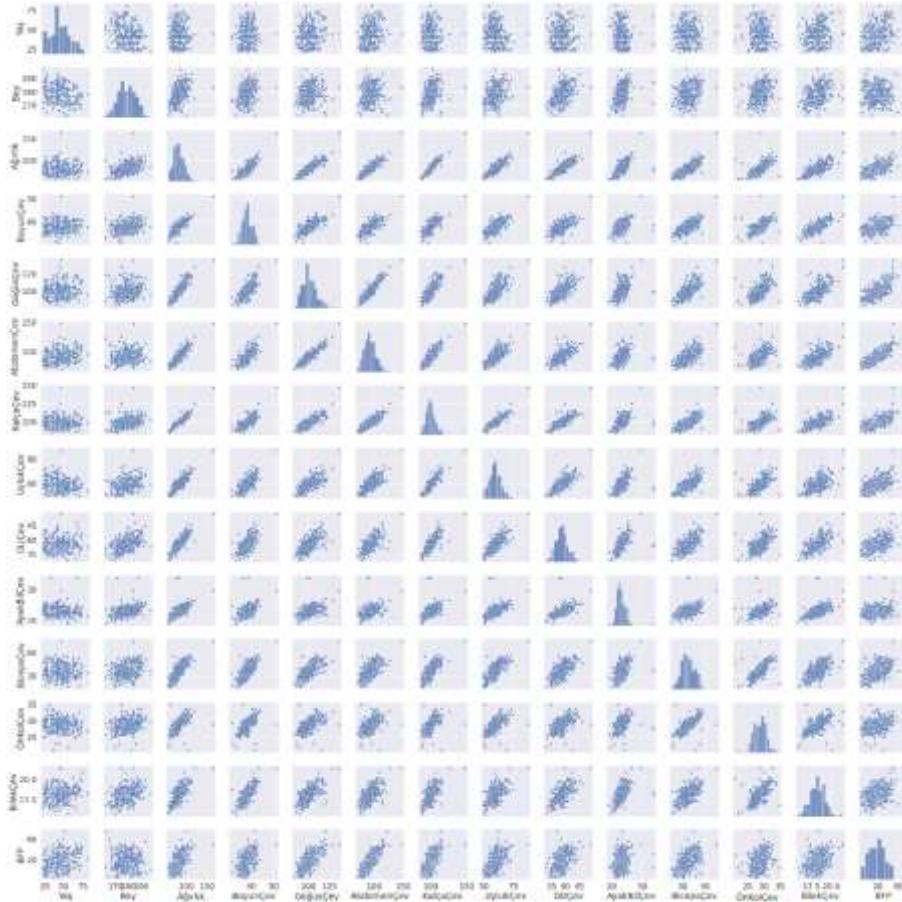


Tablo 1’de VS2 veri setindeki VS1 veri setinden istatistiksel yöntemlerle elde edilen 25 özelliğin BFP hedef özelliği ile olan ilişkisi gösterilmiştir. Verilen değerler korelasyon katsayılarıdır.

B. MAKİNE ÖĞRENMESİ İLE REGRESYON YÖNTEMLERİ

Makine öğrenmesi veriyi analiz etme, uyarılama, öğrenme, tahmin etme konusunda güçlü bir algoritma dizisi sunmaktadır. Böylelikle araştırma alanlarında makine öğrenmesi uygulamaları giderek artmaktadır [16]. Öğrenme süreci oldukça geniş kapsamlıdır. Temel olarak örüntü tanıma, regresyon tahmini ve yoğunluk tahmini problemleri gösterilebilir. Bu çalışmada BFP hesabı regresyon tahmini olarak ele alınmaktadır [17].

Regresyon problemlerinde model, gerçek değere en yakın tahmin değerini gerçekleştirmeyi amaçlar. Koşullu dağılım fonksiyonu $F(y|x)$ ’e göre her rassal giriş değişkeni x ‘e karşılık gerçek y değeri vardır. ML algoritması giriş değişkenleri dizisi için tahmin değeri oluşturan $f(x, \alpha), \alpha \in \Lambda$ fonksiyonunu kullanır. Burada Λ, α özelliğinin elemanı olduğu özellik dizisidir ve verilen giriş değeri x için ML modeli en iyi yaklaşımı gerçekleştiren $f(x, \alpha)$ fonksiyonu ile Eş. (1)’de verilen $\hat{y}$ tahmin değeri oluşturulur.



Şekil 3. VS1'deki özelliklerin ikili ilişkileri

$$\hat{y} = f(x, \alpha) \quad (1)$$

Regresyon problemlerinde amaç, Eş. (2)'de verilen risk fonksiyonunun minimum olmasıdır. Burada ortak olasılık dağılım fonksiyonu $F(x, y)$ için bilgiye eğitim veri setinden ulaşılabilir. Gerçek değer ile tahmin değeri arasındaki fark, Eş. (3)'te verilen kayıp fonksiyonu ile ifade edilir. Risk fonksiyonu minimum olması için kayıp fonksiyonunu minimum yapan Eş. (4)'te verilen $f(x, \alpha_0)$ değeri bulunmalıdır.

$$R(\alpha) = \int L(y, f(x, \alpha)) dF(x, y) \quad (2)$$

$$L(y, f(x, \alpha)) = (y - f(x, \alpha))^2 \quad (3)$$

$$f(x, \alpha_0) = \int y dF(y|x) \quad (4)$$

Regresyon tahmin problemi, ortak olasılık dağılım fonksiyonu $F(x, y)$ değerinin bilinmediği ancak veri setinin verildiği durumlarda kayıp fonksiyonu ile risk fonksiyonunu minimize etme problemidir.

Literatürde ML algoritmaları kullanılarak BFP tahmini gerçekleştirilen birçok regresyon yöntemi bulunmaktadır. Bu çalışmada BFP regresyon tahmini için RF, SVR, GBR, XGBR yöntemleri kullanılmıştır.

B. 1. Rastgele Orman Ağaçları

RF, bireysel karar ağaçlarından oluşan ağaç topluluğudur. Bu yüzden DT model yapısına benzemektedir. DT modeli düğümlerden ve o düğümdeki dallardan oluşan bir yapıya sahiptir. En üstteki düğüm kök düğümdür ve buradan sonra soru sorularak dallanmalara başlanır [18]. DT oluşturmak için en çok kullanılan ki-kare otomatik etkileşim algılama (CHAID) algoritması sadece kategorik değişkenler ile sınırlı iken hem sınıflandırma hem regresyon ağaçları (CART) algoritması ve C4.5/5.0 algoritması sürekli ve kategorik değişkenler ile kullanılabilir [19].

DT yaklaşık hedef değerlerini bir arada gruplamak için tekrarlı olarak özellik uzayını bölmelere ayırır. Bir düğüm eşik değerine göre aday bölücü tarafından sağ sol düğüm olarak bölünür ve bu durum için kayıp fonksiyonu hesaplanır. Tüm aday bölücüler içerisinde kayıp fonksiyonunu en minimize eden bölücü seçilir. Ağacın erişebileceği maksimum derinliğe kadar tüm düğümlerin bölünmesi devam eder [20].

Regresyon problemleri için RF tahmini toplanan ağaçların ağırlıksız ortalaması Eş. (5)' te verilmiştir.

$$\bar{h}(x) = (1/K) \sum_{k=1}^K h(x; \theta_k) \quad , \quad k = 1, 2, \dots, K \quad (5)$$

Burada K ağaç sayısıdır ve $h(x; \theta_k)$, x girişi için ağaçların tahmin değerini verir. Ağaç topluluğunun tahmin değeri $\bar{h}(x)$ ise ağaçların tahmin değerlerinin ortalaması ile hesaplanır. θ_k bağımsız ve özdeş dağıtılmış rastgele vektördür [21]. Ağaçların maksimum derinliğe kadar büyütülmesi önemlidir. Ancak çok fazla ağaç, kararsızlığa neden olabilir ve bu da tahmin hatalarını olumsuz olarak etkiler. Regresyon probleminde nihai tahmin değeri için tüm ağaçların çıkış değerlerinin ortalaması alınır [22].

B. 2. Destek Vektör Makineleri

SVM, maksimum marjini sağlayan en uygun hiper-düzlemi destek vektörleri ile temsil eder. Destek vektör regresyonu (SVR) büyük ve küçük yanlış tahminleri eşit şekilde cezalandıran asimetrik kayıp fonksiyonunu kullanarak eğitim sağlar. SVM'yi, SVR'ye genellemesi için fonksiyon etrafında Vapnik'in [17] $\mathcal{E}$ duyarsız yaklaşımı kullanılır. Tahmin fonksiyonu etrafında simetrik olarak minimum yarıçaplı esnek bir tüp oluşturur. Böylece belirlenen eşik değerinin altındaki hataların mutlak değeri tahminin hem üstünde hem de altında yok sayılır [23]. SVR'de bir doğrusal karar fonksiyonu Eş. (6)'da verildiği gibi ifade edilir.

$$f(x) = \langle w, x \rangle + b \quad (6)$$

Doğrusal SVR'de, Eş. (7)'de verilen bir kısıt altındaki kayıp fonksiyonunu, $\mathcal{E}$ esnek kenar payı değerlerini kullanarak ve model karmaşıklığını da azaltarak minimize etmeyi amaçlar [23]. SVM'de hatalar, esnek değişkenler ξ_i, ξ_i^* ile tolere edilebilir. C parametresi, optimizasyon problemlerinde model karmaşıklığı ile hataların tolere edilme derecesi arasındaki dengeyi belirler. C parametresi ile modele ceza uygulanarak tahmin fonksiyonun veriye uyması sağlanır [24]-[26].

SVR, doğrusal ve doğrusal olmayan problem için kullanılabilirler [27], [28]. Doğrusal olmayan SVR analizi, doğrusal durumlar için geçerli olan analize benzerdir. Ancak verilerin yüksek boyutlu Hilbert uzayına [17] haritalandırma problemi ile çözülür. Doğrusal olmayan problemlerde ikili çözüm yöntemi önerilir.

$$L_p = \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*)$$

$$Kısıt : \begin{cases} y_i - \langle w, x_i \rangle + b \leq \varepsilon + \xi_i & i = 1, 2, \dots, m \\ \langle w, x_i \rangle - b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad (7)$$

Eş. (8)'de ikili çözüm için kayıp fonksiyonu verilmiştir. Burada L_d bir lagrange fonksiyonudur ve $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ lagrange çarpanlarıdır. Bu çarpanların sıfırdan büyük ya da eşit olması gerekir. Eş. (8)'de verilen L_d fonksiyonundaki w, b, ξ_i, ξ_i^* değerleri için kısmi türev alınarak Eş. (9) elde edilir.

$$L_d = \begin{cases} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) - \sum_{i=1}^m (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ - \sum_{i=1}^m \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) - \sum_{i=1}^m \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b) \end{cases} \quad (8)$$

$$\max L_d = -\frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i - x_j \rangle - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*)$$

$$Kısıt : \sum_{i,j=1}^m (\alpha_i - \alpha_i^*) = 0 \quad \text{ve} \quad \alpha_i, \alpha_i^* \in [0, C] \quad (9)$$

Eş. (9) doğrusal problemler için uygundur. Doğrusal olmayan problemlerde yüksek uzaya haritalama $\Phi : R^n \rightarrow H$ ve kernel hilesi $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$ kullanılarak Eş. (10) yazılır. Böylece doğrusal olmayan sistemler için verilen optimizasyon problemi kullanılır [29].

$$\max L_d = -\frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i - x_j) - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*)$$

$$Kısıt : \sum_{i,j=1}^m (\alpha_i - \alpha_i^*) = 0 \quad \text{ve} \quad \alpha_i, \alpha_i^* \in [0, C] \quad (10)$$

B. 3. Gradyan Arttırma Regresyon

Arttırma metotları, sıra ile birden fazla temel modeller geliştirerek tahmin doğruluğunun arttırılmasını amaçlar. Her eklenen temel model bir önceki temel modelin yaptığı hatayı düzeltmeye çalışır. Zayıf bir öğrenici rastgele tahminden biraz daha doğru tahminler yaparken; güçlü öğrenici problem ile iyi bir ilişki kurup daha doğru tahminlerin yapılmasını sağlar. Güçlü bir modele göre zayıf bir model oluşturmak daha kolaydır. Bundan yola çıkarak birçok zayıf modelin birleştirilmesi ile tek ve yüksek doğrulukta tahmin modeli geliştirmek için arttırma metotları kullanılır [30].

Arttırma metotlarının geleneksel makine öğrenmesi algoritmalarından farkı, optimizasyon fonksiyon uzayında tutulmaktadır. Yani Eş. (11)'deki fonksiyon tahmini $\hat{f}(x)$, belirtildiği gibi eklemeli fonksiyonel formdadır. Burada M iterasyon sayısı, $\hat{f}_0$ başlangıç tahmin fonksiyonu, $\{\hat{f}_i\}_{i=1}^M$ fonksiyon artışlarıdır. Arttırma metotlarında her adımda kayıp fonksiyonunu en iyi şekilde azaltan bir temel model eklenerek optimize edilir [30], [31].

$$\hat{f}(x) = \hat{f}^M(x) = \sum_{i=0}^M \hat{f}_i(x) \quad (11)$$

Temel öğrenici tahmin fonksiyonu $h(x, \theta)$ ve temel öğreniciden elde edilen tahmin değeri $\hat{\theta}$ değeridir. t . iterasyondaki topluluk tahmin fonksiyonu Eş. (12)'de ve optimizasyon kuralı ise Eş. (13)'te verilmiştir [31], [32]. N veri sayısı olmak üzere fonksiyon tahmini için her iterasyonda optimizasyon kuralını minimum yapan adım sayısı ρ ve θ değerleri seçilmelidir.

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t) \quad (12)$$

$$(\rho_t, \theta_t) = \arg \min \sum_{i=1}^N L(y_i, \hat{f}_{t-1}) + \rho h(x_i, \theta) \quad (13)$$

B. 4. Aşırı Gradyan Arttırma Regresyon

Gradyan arttırmaya benzer yapıda çalışmaktadır. XGBR algoritmasının farkı, modelin karmaşıklığını cezalandıran bir düzenleme parametresinin eklenmesidir. Eş. (14)'te XGBR yönteminin kullandığı kayıp fonksiyonu verilmiştir. GBR'den farklı olarak düzenleme parametresi Ω vardır. Eğer Ω sıfıra ayarlanırsa model GBR metodu gibi çalışır [33]. Düzenleme parametresi Ω , birçok değişken için ağırlıkları sıfıra çekmeye çalışır ve böylece yüksek boyutlu problemlerde önemli bir rol oynayan özellik seçimini gerçekleştirir [34].

$$L(\phi) = \sum_i L(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad \text{burada } \Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (14)$$

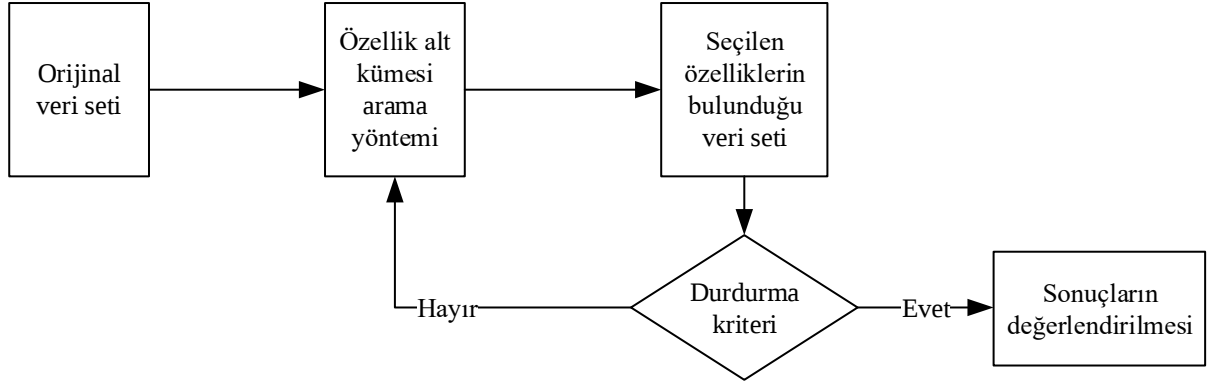
Burada ω , yapraklardaki skor vektörü, λ düzenleme parametresi ve γ yaprak düğümü bölmek için gereken minimum kayıptır.

XGBR, ağaç karmaşıklığı, öğrenme oranı, düzenleme parametresi gibi birçok hiper-parametre ayarlaması ile modelin aşırı öğrenmesinin önüne geçebilir. Boş değerler için hassastır. Bir veri setinde boş değerler varsa ilk tahminde varsayılan değer boş verilere atanır. Sonraki tahminde boş verilerden oluşan hatalar farklı dallara yerleştirilir ve oluşan kazanç değerine bakılarak kazancın yüksek olduğu dallara boş değerler atanır [33], [35].

C. ÖZELLİK SEÇİM YÖNTEMLERİ

Son yıllarda yüksek boyutlu verilerin varlığı ML algoritmalarının kullanımını zorlaştırmaktadır. ML algoritmalarını daha etkili şekilde kullanabilmek için daha az özellik ile çalışmayı sağlayan özellik seçim algoritmaları bulunmaktadır. Özellik seçim yöntemlerinin temel adımları Şekil 4'te verilmiştir. Bir veri setinde seçim gerçekleştirmek için özellikler bireysel olarak ya da özelliklerden oluşan bir alt küme olarak değerlendirilir. Bireysel değerlendirme, özelliklerin alaka derecesine göre yapılır ve özelliklere bireysel ağırlık atanır. Bu ağırlıklara göre özellikler önem sırasına göre sıralanır. Alt küme değerlendirmesi ise araştırma stratejilerine göre aday özellik alt kümeleri oluşturulur ve değerlendirilir. Genellikle FS süreci alt küme oluşturma, alt kümeyi değerlendirme, durdurma kriteri sağlayıp sağlamadığına bakma ve performansın değerlendirilmesi sırasıyla gerçekleşir [36].

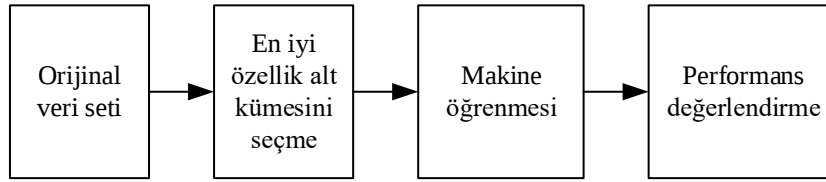
Özellik seçimi için 4 grup yaklaşım mevcuttur. Birincisi, Filtre FS yöntemi, eğitim verisinin genel karakteristiğini kullanır. İkincisi, Sarmal FS yöntemi, en uygun özellik alt küme seçimi ile uygunluk arasındaki ilişkiyi kullanır. Üçüncüsü, Gömülü FS yöntemi, eğitim esnasında özellik seçimi uygulayan ML algoritmaları ile yapılır. Son olarak Hibrit FS yöntemi, en uygun filtre ve sarmal metotların birleşiminden oluşmaktadır. İlk önce filtre metot uygulanarak özellik uzay boyutu azaltılır ve aday özellik alt kümeleri oluşturulur. Daha sonra, sarmal metot adaylar içerisinde en uygun özellik alt kümesini bulmaya çalışır. Hibrit yöntemler, filtre yöntemlerin yüksek etkililiğinden, sarmal yöntemlerin ise yüksek doğruluğundan faydalanır [36], [37].



Şekil 4. Özellik seçim işlemleri için temel adımlar

C. 1. Filtre Özellik Seçim Yöntemi

Filtre özellik seçim yöntemi, bir veri setindeki en ayırt edici özelliği seçme üzerine çalışmaktadır. Genellikle, ML yöntemlerinden önce özellik seçimi gerçekleşir ve iki aşamada strateji gerçekleşir. İlk olarak, tüm özellikler belirlenen bir kritere göre sıralanır ve ardından en yüksek sıralamaya sahip özellikler seçilir. Seçilen özellikler ile ML işlemi gerçekleştirilir. Sınıflandırma, regresyon işlemleri için birçok filtre tipi yöntemler mevcuttur. Şekil 5'te filtre özellik seçim yönteminin genel yapısı gösterilmiştir [38].



Şekil 5. Filtre FS modeli genel yapısı

Filtre metod tutarlılık, bağımlılık ve uzaklığa dayalı olarak genel karakteristikleri ölçerek giriş verilerinden ilgili özellikleri seçer. Pearson, Sperman, Kendall gibi korelasyon matrislerini kullanarak çıktı verisi ile yüksek korelasyona sahip verileri filtreler [39].

Bu çalışmada filtre özellik seçim yöntemlerinden karşılıklı bilgi yöntemi kullanılmıştır.

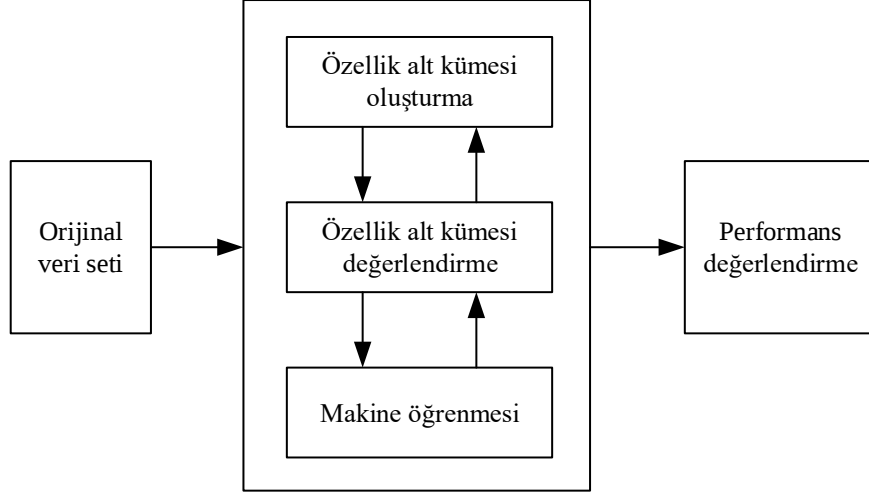
C.1.1. Karşılıklı Bilgi ile Özellik Seçimi

Regresyon problemlerinin ana amacı, hata kriterini mümkün olduğu kadar azaltmaktır. Hata değeri hesaplamada, MSE ve mutlak hata ortalaması (MAE) sıkça kullanılır. MI, iki rastgele değişken arasındaki bağılılığı ölçen bir niceliktir. MI, bir değişkenin değerleri bilinirken diğer değişkenin değerlerindeki entropi tarafından ölçülen belirsizliğin azalmasıdır. Bu yöntem bir FS kriteri olarak kullanılabilir ve değerlendirmek için mümkün olan özellik alt kümelerini seçebilir. Ayrıca, değişkenler arasında doğrusal olmayan ilişkiyi belirleme avantajına sahiptir. Diğer bilinen bazı kriterler, sadece doğrusal bağılıklar ile sınırlıdır [40].

MI yöntemi, doğrusal korelasyon katsayısının aksine kovaryansta kendini göstermeyen bağımlılıklara duyarlıdır. Eğer karşılıklı ilişki sıfır ise, iki rastgele değişken birbirinden tamamen bağımsızdır. Amaç, değişkenlerin olasılık yoğunlukları bilinmeden veri setinden karşılıklı bilgiyi tahmin etmektir [41].

C. 2. Sarmal Özellik Seçim Yöntemi

Filtre yaklaşım ile sarmal yaklaşım sadece değerlendirme kriteri noktasında birbirinden ayrılır. Sarmal yaklaşım alt küme değerlendirmesi için öğrenme algoritması kullanır. Alt küme oluşturma ve alt küme değerlendirme tekniklerini çeşitlendirerek farklı sarmal yaklaşımlar oluşturulabilir. Sarmal yaklaşım, öğrenme algoritmasına en iyi uygun alt kümeyi seçer. Bu yüzden genellikle sarmal yöntem filtre yöntemine göre daha iyi sonuçlar gösterir [36]. Şekil 6’da sarmal özellik seçim metodunun genel yapısı verilmiştir.



Şekil 6. Sarmal FS modeli genel yapısı [16]

Sıralı özellik seçim algoritmaları, başlangıç özellik uzayı boyutunu azaltmayı amaçlayan açgözlü olarak nitelendirilen bir araştırma metodudur. FS algoritmaları arkasındaki motivasyon, otomatik olarak problem ile en alakalı özellik alt kümesini seçmektir [42].

Sarmal özellik seçim yöntemi, ML algoritması ile farklı özellik alt kümelerini değerlendirmeye çalışır ve en iyi performans gösteren alt kümeyi seçer. En temel metod SFS yöntemidir. Boş veri seti ile sürece başlar ve birer birer özellikleri ekleyerek devam eder. Her adımda özellik alt kümesine eklendiğinde en iyi genelleme yapan özelliği ekler ve bir kez eklenen özellik bir daha kaldırılmaz. SBS ise tüm özellikler ile sürece başlayıp her adımda özellikleri kaldırarak devam eder. SBS’de, veri setinden çıkarıldığında en iyi genellemeyi veren özelliği kaldırılır. SFS’deki gibi SBS’de de bir özellik kaldırılınca geri eklenmez [43].

C.2.1. Sıralı İleriye Doğru Özellik Seçimi

SFS algoritması, d boyutlu X özellik kümesinden $x_1, x_2, \dots, x_k$ özelliklerini belli kritere göre seçerek X_k özellik alt kümesini oluşturur ($k < d$). Bu algoritma başlangıçta boş küme ile işleme başlar. Yani başlangıçta alt küme boyutu k sıfırdır. Sonraki adımda X_k özellik kümesine x^+ özelliği eklenir. x^+ , Eş. (15)’de gösterildiği gibi problemde kullanılan J kriter fonksiyonunu maksimum yapan özelliktir ve en iyi performansı göstermeyi amaçlar. Eş. (16)’de gösterildiği gibi kriteri sağlayan x^+ özelliği, özellik alt kümesine eklenir ve k değeri 1 artırılarak Eş. (15) ve Eş. (16) tekrarlanır. Bu işlemler belirlenen durdurma kriteri karşılanana kadar devam eder [44].

$$x^+ = \arg \max J(X_k + x), \text{ burada } x \in X - X_k \quad (15)$$

$$X_{k+1} = X_k + x^+ \quad (16)$$

C.2.2. Sıralı Geriye Doğru Özellik Seçimi

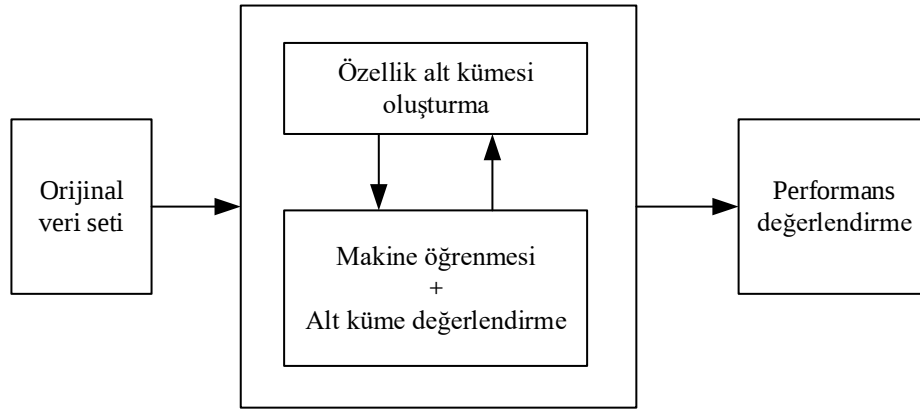
SBS algoritması, d boyutlu X özellik kümesinden $x_1, x_2, \dots, x_k$ özelliklerini belli kritere göre seçerek X_k özellik alt kümesini oluşturur ($k < d$). Bu algoritma başlangıçta tüm özelliklerin bulunduğu veri seti ile işleme başlar. Yani başlangıçta $k = d$. Seçilen özelliklerin bulunduğu k boyutlu bir özellik alt kümesi geri döndürür ($k < d$). Verilen özellik küme seti boyutu ($k = d$) ile algoritma başlar. Sonraki adımda X_k özellik kümesinden x^- özelliği çıkarılır. x^- , Eş. (17)'de gösterildiği gibi problemde kullanılan J kriter fonksiyonunu maksimum yapan özelliktir ve en iyi performansı göstermeyi amaçlar. Eş. (18)'te gösterildiği gibi kriteri sağlayan x^- özelliği, özellik alt kümesinden çıkarılır ve k değeri 1 azaltılarak Eş. (17) ve Eş. (18) tekrarlanır. Bu işlemler belirlenen durdurma kriteri karşılanana kadar devam eder [44].

$$x^- = \arg \max J(X_k - x), \text{ burada } x \in X_k \quad (17)$$

$$X_{k-1} = X_k - x^- \quad (18)$$

C. 3. Gömülü Özellik Seçim Yöntemi

Gömülü özellik seçim yöntemi, ML algoritmalarının dâhili olan bilgilerini kullanır. Örneğin, RF öğrenme algoritmasında özelliğin önemi belirlenir ve buna göre özellik seçimleri gerçekleşir. Gömülü FS yöntemi, genellikle modelin performansı ve hesaplama yükü arasında bir denge sağlamaya çalışır [45]. Gömülü FS yönteminin genel yapısı Şekil 7'de verilmiştir.



Şekil 7. Gömülü FS modeli genel yapısı

Gömülü FS yöntemi, sarmal FS yöntemlerine göre daha az hesaplama yüküne sahiptir. Sadece giriş özellikleri ile çıkış arasındaki ilişkiye bakmaz aynı zamanda giriş özelliklerinin arasındaki ilişkiyi de araştırır. Uygun olan alt kümeleri belirler ve bunlar arasından en uygun alt kümeyi ML algoritması kullanarak seçer. Ağaç ve kural tabanlı modeller, düzenleme modelleri gibi bazı modeller bu metota örnek gösterilebilir. En iyi tahminciyi ve en iyi sonuçlar veren bölme noktalarını araştırır. Eğer herhangi bir değişken bölme noktasında kullanılmıyorsa, bu tahmin eşitliğinde yer almaz, modelden çıkarılmış olduğu anlamına gelir. Ağaç toplulukları genelde benzer yapıdadırlar ancak RF gibi bazı algoritmalar bir ağaç oluşurken alakasız özellikler üzerinde bölmeleri zorlar [46].

C.3.1. Rastgele Orman Ağaçları Makine Öğrenim Algoritması Kullanılarak Özelliklerin Önemine Dayalı Özellik Seçimi

RF'de bir regresyon ağacı, yinelemeli olarak kök düğümü uç düğüme kadar homojen gruplara bölerek oluşturulur. Her bölme bir özelliğin değerine bağlı olarak gerçekleşir ve bölme kriterine göre seçilir.

Bir ağaç oluşturulduğunda herhangi bir gözlem değeri için cevap, bölme değişkenleri için gözlenen değerlere bağlı olarak kök düğümünden yaklaşık son düğüme kadar giden bir yol izlenerek tahmin edilebilir. Tahmin cevabı düğümdeki cevapların ortalaması alınarak elde edilir. Rastgele ormanlar, çok sayıda ağaçtan oluşur. Özellik önemi için de çok sayıda ağacın bulunması önemlidir. Özellik önemi, regresyon problemleri için ortalama azalan safsızlık ölçütüdür. Bölme değişkenlerinin seçimi için safsızlığın yanlılığından dolayı sonuç değişken önemi metriği de yanlıdır [47].

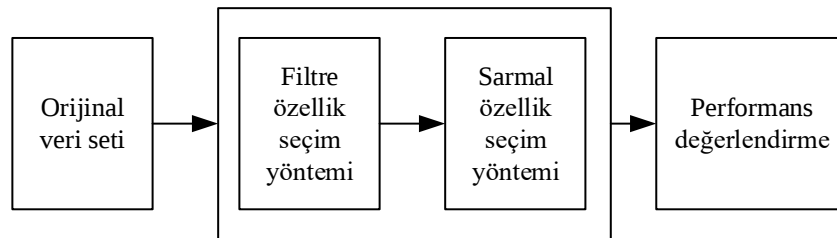
RFI, ağaçlar topluluğunda hesaplanan ve sadece ilişkili değişkenlere bağlı olan önemlilik değerine göre bir seçim gerçekleştirir. Regresyon problemlerinde bir ağaç eğitildiği zaman, her özelliğin safsızlığı yani varyansı ne kadar azalttığını hesaplamak mümkündür. Bir özellik safsızlığı ne kadar azaltırsa o özellik o kadar önemlidir. Her özellik için önem değeri RF tarafından oluşturulur [48]. İlişkisiz değişkenler sıfır öneme sahiptir ve ilişkili özelliklerin önemini etkilemez. Pratikte istenen durum, gürültüye bir önem verilmemesi ve gürültünün başka değişkeni daha fazla (veya daha az) önemli hale getirmemesidir [49]. Değişken önemi için eğitim esnasında, önyüklem örneğinde seçilmeyen verilerden oluşan ilgili torba dışı veriler seçilir ve bu verilerin tahmin hatası hesaplanır ve orman üzerinden ortalaması alınır. Eğitimden sonra belli bir özelliğin önemi ölçülmek istendiğinde, bu özelliğin değerleri rastgele permüte edilir ve tekrar tahmin hatası hesaplanır. O değişkenin önem değeri permütasyondan önceki ve sonraki hata farkının ortalaması alınarak hesaplanır [50].

C.3.2. Rastgele Orman Ağaçları Makine Öğrenim Algoritması Kullanılarak Özelliklerin Öneme Dayalı Yinelemeli Özellik Seçimi

RRFI, C.3.1’de anlatılan RF tarafından elde edilen özellik önemi bilgisine dayanarak özellik seçimi yapmaktadır. Tek fark, bu işlemler tekrarlı olarak yapılır ve her adımda özellik önemi hesaplanır. RF ilk koşuturulmasında başlangıç özellik önemleri belirlenir ve en düşük öneme sahip özellik veri setinden çıkartılarak performanstaki değişim kriteri değerlendirilir. Kriter sağlanıyorsa özellik kalıcı olarak veri setinden çıkartılır. Sonraki adımda kalan özellikler ile yeniden RF algoritması koşuturularak özellik önemleri belirlenir. Her adımda yeni bir önem sıralaması oluşturulur ve tekrar en düşük öneme sahip özellik çıkartılarak performans değerlendirilir. Belirlenen durdurma kriteri sağlanana kadar işlemler tekrarlı olarak devam eder. Böylelikle özellik önemleri, koşullar arasında değil sadece bir koşu için karşılaştırılmış olur [51].

C. 4. Hibrit Özellik Seçim Yöntemi

Hibrit metotlar, en uygun filtre ve sarmal metotların birleşiminden oluşmaktadır. İlk önce filtre metot uygulanarak özellik uzay boyutu azaltılarak aday özellik alt kümeleri oluşturulur. Daha sonra sarmal metot ise en iyi özellik alt kümesini bulmaya çalışır. Şekil 8’de hibrit özellik seçim yönteminin genel yapısı verilmiştir. Hibrit metotlar, filtre metotların yüksek etkililiğinden, sarmal metotların ise yüksek doğruluğundan faydalanır [37].



Şekil 8. Hibrit FS modeli genel yapısı

C.4.1. Rastgele Karıştırma ile Hibrit Özellik Seçimi

RS bir özelliğin gözlem değerlerini rastgele karıştırılmasına dayalı özellik seçimi yapmaktadır. Hibrit özellik seçim yöntemi, filtre metotlar gibi belli kriterlere göre özellikleri tek tek değerlendirir ve

sarmal yöntemler gibi aday özellik alt kümelerini belli bir öğrenme algoritması ile test ederek en uygun alt kümesini seçer. Bu yöntem tek tek özelliklerin gözlem değerlerini karıştırır ve bu karıştırmanın ML algoritmasının performansını nasıl etkilediğini araştırır. Eğer gözlem değerleri karıştırılan özellik, önemli ise kullanılan performans değerlendirme metriğine göre performansın düşmesi beklenir. Aksi halde performans üzerinde etkisinin ya çok az olması ya da hiç olmaması beklenir. Böylelikle karıştırıldıktan sonra performansta düşüşe neden olan özelliğin veri setinde kalması, aksi etki gösteren özelliğin ise veri setinden çıkartılması gerekir [52].

C.4.2. Yinelemeli Özellik Eleme ile Hibrit Özellik Seçimi

RFE, eğitim hatasından en az etkiye sahip olan yani en az öneme sahip özelliği veri setinden çıkartarak modelin genel performansını arttırmayı amaçlar. Zayıf ve gereksiz özelliklerin veri setinden çıkartılması performansı iyileştirebilir. Ancak tek başına yetersiz olan özellikler başka özellikler ile birlikte kullanıldığında önemli bir performans artışı sağlayabilirler. Bu nedenle, yinelemeli özellik eleme yöntemi, tekrarlı olarak her adımda en zayıf özelliği veri setinden çıkarır ve kalan özellikler ML algoritması tarafından performansa etkisi değerlendirilir. Eğer özellik çıkarıldıktan sonra performans düşerse özellik veri setinde kalmalıdır [53].

III. DENEYSEL ÇALIŞMALAR

Bu çalışmada BFP regresyon tahmini için 2 adet veri seti kullanılmıştır. Veri setlerine 4 grup özellik seçim yöntemlerinden toplam 7 FS algoritması kullanılarak özellik sayıları azaltılmıştır. Yeni oluşturulan özellik alt kümelerine 4 ML algoritması uygulanarak BFP tahmini gerçekleştirilmiştir. 13 özellik bulunan VS setinden 5, 8 ve 11 özellik seçilmiştir. 38 özellik bulunan VS2 setinden 8, 20 ve 32 özellik seçilmiştir. Seçilen özellikler ile ML modelleri ile gerçekleştirilen tahminler 4 karşılaştırma metriği kullanılarak değerlendirilmiştir. Ayrıca modellerin eğitim süreleri hesaplanarak karşılaştırılmıştır.

Regresyon yöntemlerinde kullanılan makine öğrenimi algoritmalarının parametre ayarları, modelin performansını etkileyen önemli bir adımdır. Bu nedenle en uygun regresyon parametreleri belirlenmelidir. Bu amaçla, ızgara arama ve 5 katmanlı çapraz doğrulama yöntemleri kullanılarak regresyonlara ait en iyi sonucu veren parametreler belirlenmiştir. Ayrıca veri setinin %80'i eğitim için, %20'si test ve değerlendirmeler için kullanılmıştır.

A. REGRESYON PARAMETRELERİ VE KARŞILAŞTIRMA METRİKLERİ

A.1. Parametre Ayarları

Bu çalışmada RF, SVR, GBR ve XGBR için kullanılan parametreler:

- RF için:
 - 13 özellik bulunan veri seti için: Ağaç sayısı 310, maksimum derinlik 45, minimum yapraktaki veri sayısı 5, minimum bölme veri sayısı 5.
 - 38 özellik bulunan veri seti için: Ağaç sayısı 170, maksimum derinlik 21, minimum yapraktaki veri sayısı 1, minimum bölme veri sayısı 13.
- SVR için:
 - 13 özellik bulunan veri seti için: RBF çekirdeği, regülasyon parametresi 490, gama parametresi 0.001.
 - 38 özellik bulunan veri seti için: Lineer çekirdeği, regülasyon parametresi 1.
- GBR için:
 - 13 özellik bulunan veri seti için: Öğrenme oranı 0.09, maksimum derinlik 3.
 - 38 özellik bulunan veri seti için: Öğrenme oranı 0.1, maksimum derinlik 3.

- XGBR için:
 - 13 özellik bulunan veri seti için: Öğrenme oranı 0.12, alfa parametresi 10^{-6} , gama parametresi 0.4, her ağaç oluşturulurken sütunların alt örnek oranı 0.6, eğitim örneklerinin alt örnek oranı 0.6.
 - 38 özellik bulunan veri seti için: Öğrenme oranı 0.12, bir çocukta ihtiyaç duyulan minimum örnek ağırlığı (kendi ağırlığı) toplamı 3, alfa parametresi 10^{-6} , gama parametresi 0.3.

A.2. Regresyon Başarımı Karşılaştırma Metrikleri

BFP tahmini için her özellik seçim algoritması uygulandıktan sonra seçilen özellikler ile yeni oluşturulan veri setleri makine öğrenim algoritmaları ile hedef değişkenin tahmini gerçekleştirildi. Çıkan sonuçları karşılaştırıp değerlendirmek için literatürde sıklıkla karşılaşılan MSE, R^2 , MAPE ve MAE metrikleri kullanıldı. Eş. (19), Eş. (20), Eş. (21) ve Eş. (22)'de n veri sayısı olmak üzere $\hat{y}_j$, j . bağımlı değişken y_j 'nin tahmin değeridir.

MSE, regresyon problemlerinde hedef değişkenin gerçek değeri ile tahmin edilen değeri arasındaki farkın karesinin ortalamasıdır. Eş. (19)'da verilen MSE, hatanın karesini alarak küçük hataları dahi cezalandırır. Sıfıra ne kadar yakın olursa o kadar iyi performans gösterdiği anlamına gelir.

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2 \quad (19)$$

R^2 , hataların karelerinin toplamı ile hesaplanır. Bağımlı değişkendeki toplam varyasyonun yüzde kaçının bağımsız değişkendeki varyasyon tarafından açıklandığını belirler. Eş. (20)'de $\bar{y}$ bağımlı değişkenin gerçek değerlerin ortalamasıdır.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2} \quad (20)$$

MAPE, bağımlı değişkenin gerçek değeri ile tahmin değeri arasındaki farkın yüzdelik cinsten ifadesi Eş. (21)'deki gibidir.

$$MAPE = \frac{1}{n} \sum_{j=1}^n \frac{|y_j - \hat{y}_j|}{y_j} \times 100 \quad (21)$$

MAE, hedef değişkenin gerçek değeri ile tahmin edilen değeri arasındaki farkın mutlak değerinin ortalamasıdır. MAE, Eş. (22)'de verildiği gibi ifade edilir.

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (22)$$

B. ÖZELLİK SEÇİMİNDE KULLANILAN REGRESYON YÖNTEMLERİ

Bu çalışmada filtre, sarmal, gömülü ve hibrit özellik seçim yöntemleri kapsamında 7 adet özellik seçim algoritması kullanılmıştır. Çalışmalar Python programlama dili kullanılarak gerçekleştirilmiştir.

ML ve FS işlemleri için bir Python kütüphanesi olan Scikit-learn kullanılmıştır. ML algoritmalarında modelin eğitilmesi ve başarımın test edilmesi amacıyla verilerin %80'ni eğitim için %20'si test için kullanılmıştır.

- MI filtre özellik yöntemi özelliklerin hedef özellik ile arasındaki bilgiyi değerlendirir. Durdurma kriterini sağlayana dek hedef özellik ile en çok ilişkili olan bağımsız özellikler seçilmiştir.
- SFS ve sıralı SBS sarmal FS yöntemleri, alt küme oluşturma ve alt küme değerlendirmesi için öğrenme algoritması kullanır. Bu özellik seçim yöntemleri için RF regresyon öğrenme algoritması kullanılmıştır.
- Gömülü FS yöntemleri, algoritma yapısında özellik seçimi gerçekleştiren makine öğrenim yöntemlerini kullanarak yapılan bir yöntemdir. Bu çalışmada gömülü FS yöntemleri için RF regresyon öğrenme algoritması kullanılarak özelliklerin önem değerleri belirlenip bu değerlere dayalı bir seçim gerçekleştirilmiştir.
- Hibrit özellik seçim yöntemleri, uygun filtre ve sarmal metotları birlikte kullanan bir yöntemdir. Bu çalışmada RS hibrit FS yöntemi için SVR öğrenme algoritması kullanılmıştır. RFE hibrit FS yöntemi için GBR öğrenme algoritması kullanılmıştır.

C. DENEYSEL SONUÇLAR

VS1 ve VS2 deney setleri kullanılarak çeşitli FS metotlarının farklı ML modelleri üzerindeki etkileri incelenmiştir. Tablo 2 ve Tablo 3'te ML modellerinin eğitim süreleri, seçilen özellik sayısına göre karşılaştırılmıştır. Özellik seçim işlemi yapılmadan tam veri seti ile kurulan modellerin eğitim süresi, özellik seçimi yapıldıktan sonra kurulan modelin eğitim süresinden tüm durumlar için daha uzundur. Her ML algoritmasında eğitim için harcanan süre değişmektedir. Çalışmada VS1 ve VS2 veri setleri kullanıldığında, ML modellerinden eğitim için geçen sürelerden 0,49580 sn ile en uzun RF modeli iken 0,00641 sn ile en kısa SVR modelidir. Her iki veri seti için en uzun ve en kısa olan modeller değişmemektedir, ancak veri setindeki özellik sayısı arttıkça işlem yükü artacağından eğitim süreleri de artmaktadır.

Tablo 2. VS1 için seçilen özellik sayısına göre model eğitim süresi (sn)

Regresyon Yöntemi		RF	SVR	GBR	XGBR
Seçilen Özellik Sayısı	5	0,42679	0,00445	0,04964	0,02053
	8	0,40214	0,00410	0,06329	0,01943
	11	0,44542	0,00392	0,07127	0,04127
	13	0,49580	0,00641	0,08630	0,02365

Tablo 3. VS2 için seçilen özellik sayısına göre model eğitim süresi (sn)

Regresyon Yöntemi		RF	SVR	GBR	XGBR
Seçilen Özellik Sayısı	8	0,25575	0,00492	0,07007	0,02672
	20	0,33445	0,00520	0,11674	0,03991
	32	0,41710	0,00602	0,16547	0,05443
	38	0,42393	0,01176	0,18552	0,18552

VS1 ve VS2 veri setlerinde özellik seçimi yapan 7 FS algoritmasının bağımlı değişken için tahminleri 4 farklı ML algoritması ile değerlendirilmiştir. Tablo 4 - Tablo 10 arasında verildiği gibi sonuçlar 4 farklı performans metriği kullanılarak karşılaştırılmıştır. VS1 veri seti için yaklaşık %40 (5 özellik),

%60 (8 özellik) ve %85 (11 özellik) olacak şekilde seçilen özellik sayısı ayarlanmıştır. VS2 veri seti için ise yaklaşık %20 (8 özellik), %40 (20 özellik) ve %85 (32 özellik) olacak şekilde seçilen özellik sayısı ayarlanmıştır. Tablo 4'te VS1 ve VS2 veri setlerine FS algoritmaları uygulanmadan önceki sırasıyla 13 ve 38 bağımsız değişken bulunan durumlar için ML model performansları verilmiştir.

Tablo 5 - Tablo 7 arasında VS1 veri setine farklı FS algoritmaları uygulanarak sırasıyla 5, 8, 11 özellik seçilmiştir ve farklı performans metriklerine göre ML modellerinin başarıları verilmiştir. Tablo 5'te verilen VS1 veri setinden 5 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar hibrit FS metotlarından elde edilmiştir. Bu tablodaki 16 performans sonucuna göre hibrit metot 14 durumda en iyi performansı göstermiştir. Bu performanslardan 9 sonuç RS hibrit FS metotuna aittir. Ayrıca verilen 2 farklı sarmal ve gömülü metot türleri kendi içlerinde aynı sonuçları vermiştir. Yani bu yöntemler aynı özellikleri seçmişlerdir. Benzer durum

Tablo 6 ile Tablo 7'deki sarmal ve gömülü FS yöntemlerinde de görülebilmektedir.

Tablo 6'da verilen VS1 veri setinden 8 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar Tablo 5'e benzer olarak hibrit FS metotlarından elde edilmiştir. Bu tablodaki 16 performans sonucuna göre hibrit metot 12 durumda en iyi performansı göstermiştir. Bu performanslardan 8 sonuç RS hibrit FS metotuna aittir. Tablo 7'de verilen VS1 veri setinden 11 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar sarmal FS metotlarından elde edilmiştir. Bu tablodaki 16 performans sonucuna göre sarmal metot 8 durumda en iyi performansı göstermiştir. Verilen 2 farklı sarmal metot türünde aynı özellikler seçildiğinden aralarında performans farklılığı olmamıştır.

Tablo 4. VS1 ve VS2 için özellik seçimi yapılmadan önce modellerin performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Orijinal Veri Setleri	VS1	12,865	0,789	2,962	0,179	10,654	0,826	2,607	0,152	12,130	0,813	2,797	0,149	12,691	0,827	3,058	0,168
	VS2	12,871	0,802	2,902	0,159	10,605	0,827	2,618	0,158	13,582	0,798	3,015	0,163	13,180	0,801	2,941	0,158

Tablo 8 - Tablo 10 arasında VS2 veri setine farklı FS algoritmaları uygulanarak sırasıyla 8, 20, 32 özellik seçilmiştir ve farklı performans metriklerine göre ML modellerinin başarıları verilmiştir. Tablo 8'de verilen VS2 veri setinden 8 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar sarmal FS metotlarından elde edilmiştir. Toplam 16 performans sonucuna göre sarmal FS metotlarından 10 durumda SBS metodu, 4 durumda da SFS metodu en iyi performansı göstermiştir. Tablo 9'da verilen VS2 veri setinden 20 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar hibrit FS metotlarından elde edilmiştir. Toplam 16 performans sonucuna göre hibrit FS metotlarının başarılı olduğu 10 durumdan 9'unda RFE metodu en başarılı sonuçları vermiştir. Tablo 10'da VS2 veri setinden 32 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar sarmal FS metotlarından elde edilmiştir. Tablo 10'daki toplam 16 performans sonucuna göre sarmal FS metotlarının başarılı olduğu 10 durumdan 6'sında SFS metodu en başarılı sonuçları vermiştir.

MI'nın başarılı olduğu yalnızca 2 durum vardır. İki veri seti için yüksek oranda özellik seçildiğinde MI, R2 metriğine göre başarılı sonuçlar vermiştir.

Tablo 5. VS1 için farklı FS yöntemleri ile seçilen 5 özelliğin model performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Özellik Seçme Yöntemi	MI	13,733	0,794	2,843	0,161	12,021	0,809	2,769	0,162	13,987	0,778	3,084	0,180	14,961	0,788	3,112	0,171
	SFS	12,694	0,802	2,753	0,151	10,694	0,825	2,605	0,159	14,968	0,787	3,090	0,183	13,588	0,796	2,852	0,160
	SBS	12,694	0,802	2,753	0,151	10,694	0,825	2,605	0,159	14,968	0,787	3,090	0,183	13,588	0,796	2,852	0,160
	RFI	12,205	0,809	2,816	0,153	10,477	0,829	2,639	0,158	14,473	0,788	3,050	0,170	15,678	0,791	2,969	0,185
	RRFI	12,205	0,809	2,816	0,153	10,477	0,829	2,639	0,158	14,473	0,788	3,050	0,170	15,678	0,791	2,969	0,185
	RFE	12,304	0,807	2,748	0,162	10,310	0,831	2,587	0,156	13,126	0,803	2,856	0,176	15,078	0,802	3,016	0,161
	RS	11,286	0,830	2,626	0,162	11,440	0,812	2,694	0,161	12,943	0,818	2,966	0,170	11,278	0,831	2,737	0,165

Tablo 6. VS1 için farklı FS yöntemleri ile seçilen 8 özelliğin model performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Özellik Seçme Yöntemi	MI	13,178	0,798	2,984	0,169	11,855	0,809	2,729	0,159	16,176	0,777	3,280	0,183	17,068	0,759	3,115	0,180
	SFS	12,642	0,810	2,850	0,152	10,833	0,823	2,621	0,160	15,036	0,786	2,975	0,179	14,297	0,775	2,911	0,179
	SBS	12,642	0,810	2,850	0,152	10,833	0,823	2,621	0,160	15,036	0,786	2,975	0,179	14,297	0,775	2,911	0,179
	RFI	12,754	0,808	2,802	0,153	10,934	0,821	2,620	0,158	14,383	0,787	2,932	0,168	12,203	0,817	2,760	0,155
	RRFI	12,944	0,798	2,860	0,157	10,981	0,822	2,632	0,157	13,858	0,787	3,022	0,172	13,846	0,792	2,877	0,160
	RFE	12,975	0,807	2,871	0,158	11,762	0,832	2,784	0,155	14,552	0,800	3,004	0,157	12,518	0,833	2,807	0,158
	RS	12,308	0,815	2,730	0,163	10,287	0,832	2,580	0,157	11,152	0,809	2,686	0,169	13,113	0,793	2,894	0,167

Tablo 7. VS1 için farklı FS yöntemleri ile seçilen 11 özelliğin model performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Özellik Seçme Yöntemi	MI	12,637	0,809	2,917	0,167	10,528	0,828	2,618	0,154	13,875	0,796	3,099	0,174	13,517	0,782	2,982	0,172
	SFS	12,458	0,808	2,850	0,154	11,196	0,838	2,676	0,161	13,521	0,816	2,882	0,164	12,964	0,805	2,760	0,150
	SBS	12,458	0,808	2,850	0,154	11,196	0,838	2,676	0,161	13,521	0,816	2,882	0,164	12,964	0,805	2,760	0,150
	RFI	12,854	0,802	2,887	0,158	10,990	0,826	2,644	0,158	12,955	0,812	2,876	0,163	12,640	0,795	2,953	0,158
	RRFI	12,854	0,802	2,887	0,158	10,990	0,826	2,644	0,158	12,955	0,812	2,876	0,163	12,640	0,795	2,953	0,158
	RFE	12,660	0,805	2,873	0,157	10,771	0,827	2,631	0,159	12,732	0,812	2,877	0,152	14,045	0,802	2,980	0,165
	RS	12,730	0,803	2,873	0,160	10,030	0,836	2,545	0,154	11,728	0,800	2,801	0,165	13,089	0,796	2,854	0,160

Tablo 8. VS2 için farklı FS yöntemleri ile seçilen 8 özelliğin model performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Özellik Seçme Yöntemi	MI	13,471	0,769	3,009	0,190	12,035	0,803	2,777	0,172	14,928	0,745	3,244	0,190	16,293	0,739	3,345	0,201
	SFS	12,490	0,795	2,834	0,175	11,013	0,820	2,675	0,165	12,930	0,835	2,897	0,174	14,950	0,806	3,035	0,180
	SBS	12,033	0,819	2,832	0,165	11,158	0,812	2,706	0,163	13,874	0,792	2,860	0,161	10,920	0,836	2,678	0,156
	RFI	13,582	0,792	2,971	0,177	12,175	0,801	2,806	0,168	13,382	0,763	3,077	0,190	14,202	0,788	3,155	0,195
	RRFI	13,750	0,795	3,019	0,171	11,234	0,816	2,618	0,165	15,333	0,774	3,157	0,180	14,316	0,785	3,075	0,178
	RFE	13,079	0,799	2,856	0,168	12,206	0,800	2,778	0,162	13,652	0,794	2,944	0,166	13,891	0,809	3,036	0,177
	RS	12,680	0,789	2,887	0,176	12,652	0,805	2,862	0,169	13,497	0,780	3,038	0,202	15,767	0,780	3,143	0,195

Tablo 9. VS2 için farklı FS yöntemleri ile seçilen 20 özelliğin model performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Özellik Seçme Yöntemi	MI	14,310	0,774	3,137	0,186	12,582	0,807	2,883	0,174	15,276	0,752	3,386	0,188	14,778	0,778	3,169	0,184
	SFS	12,425	0,793	2,771	0,170	10,976	0,821	2,638	0,158	14,913	0,791	3,072	0,184	14,329	0,805	3,110	0,174
	SBS	12,361	0,795	2,832	0,174	11,140	0,822	2,661	0,159	14,220	0,777	2,960	0,177	14,208	0,782	3,021	0,174
	RFI	13,188	0,792	2,939	0,170	12,065	0,808	2,779	0,159	14,858	0,785	3,121	0,171	14,447	0,803	3,010	0,179
	RRFI	12,714	0,794	2,893	0,174	9,983	0,837	2,534	0,153	13,808	0,800	3,116	0,175	14,576	0,799	3,159	0,177
	RFE	12,370	0,788	2,879	0,177	10,025	0,836	2,455	0,148	13,496	0,816	3,042	0,163	12,499	0,830	2,998	0,162
	RS	12,327	0,793	2,851	0,177	11,509	0,818	2,660	0,153	14,995	0,788	3,131	0,188	14,597	0,790	3,016	0,182

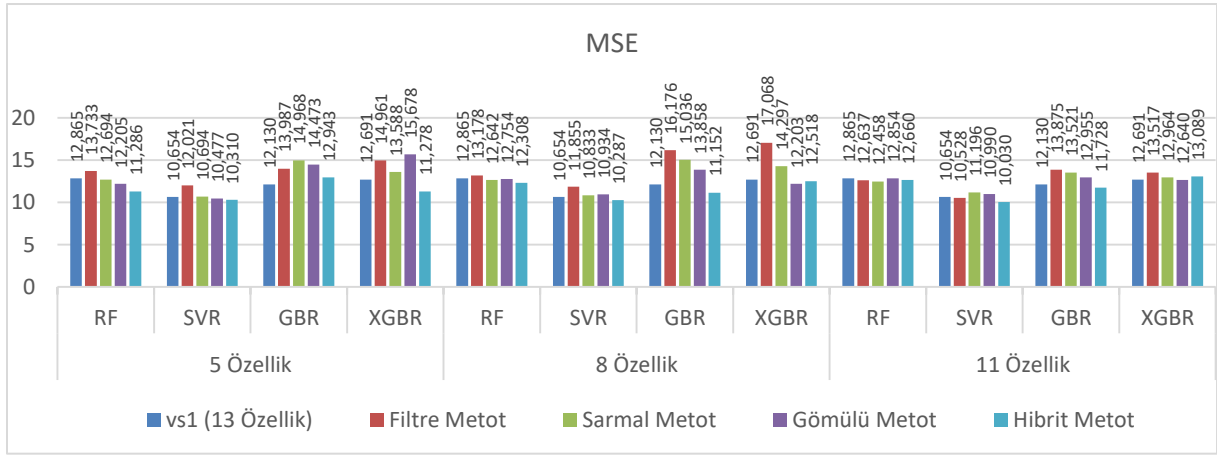
Tablo 10. VS2 için farklı FS yöntemleri ile seçilen 32 özelliğin model performansı

Regresyon Yöntemi		RF				SVR				GBR				XGBR			
Performans Metriği		MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE	MSE	R2	MAE	MAPE
Özellik Seçme Yöntemi	MI	12,899	0,792	2,940	0,175	11,479	0,814	2,700	0,156	13,542	0,809	2,959	0,159	14,673	0,800	3,109	0,175
	SFS	12,374	0,788	2,815	0,176	11,187	0,822	2,677	0,154	13,562	0,815	3,072	0,172	11,221	0,819	2,794	0,155
	SBS	12,767	0,792	2,909	0,175	11,088	0,824	2,691	0,155	13,075	0,799	2,904	0,159	12,586	0,828	2,957	0,169
	RFI	12,791	0,788	2,951	0,178	10,857	0,823	2,625	0,152	13,243	0,806	3,022	0,164	12,819	0,825	2,954	0,162
	RRFI	12,891	0,792	2,961	0,177	11,681	0,809	2,789	0,163	13,602	0,802	2,992	0,161	13,238	0,819	2,992	0,159
	RFE	12,776	0,785	2,946	0,177	9,753	0,841	2,508	0,151	13,412	0,810	3,066	0,171	12,593	0,828	2,946	0,174
	RS	13,259	0,788	2,979	0,175	9,958	0,837	2,492	0,145	13,438	0,798	2,930	0,181	14,538	0,808	3,046	0,186

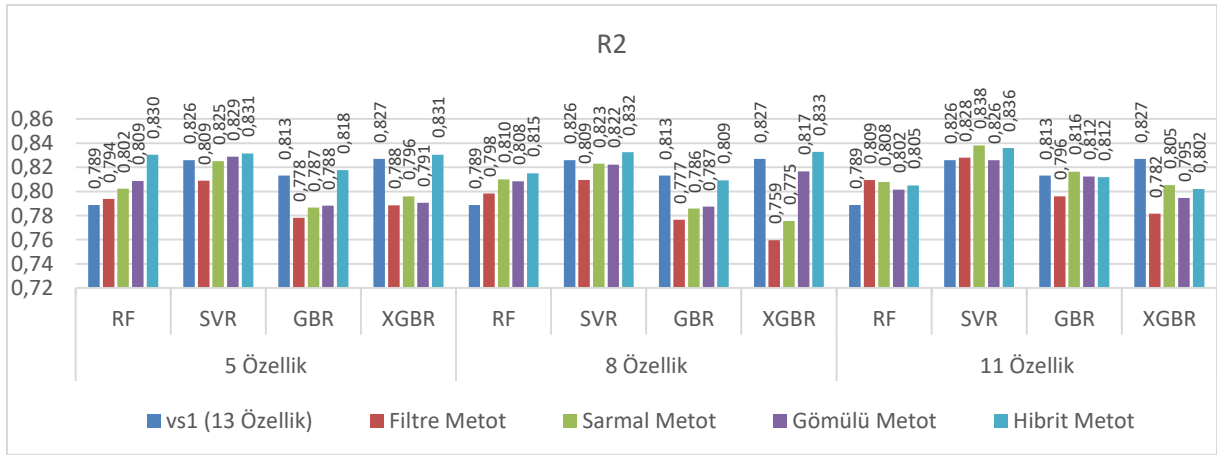
IV. SONUÇ

Giriş bölümünde bahsedilen diğer çalışmalardan farklı olarak bu çalışma, veri setlerinin farklı özellik seçim yöntemleri kullanılarak bunların karşılaştırılmasını sağlamıştır. Çalışmamızda 248 kişiden alınan 13 farklı antropometrik verinin bulunduğu VS1 seti ve bu veri setinden türetilen 25 verinin daha eklenmesi ile oluşan 38 özelliğin bulunduğu VS2 seti kullanılarak BFP tahmini gerçekleştirilmiştir. Bu veri setleri ile 7 FS yönteminin karşılaştırılması amaçlanmıştır. FS algoritmaları ile seçilen özellikler 4 farklı ML algoritması ile eğitilmiştir. Elde edilen sonuçlar 4 performans metriği ile değerlendirilmiştir. Toplamda 96 farklı durum oluşmuştur. Bunlardan 48 durumda hibrit FS yöntemi, 39 durumda sarmal FS yöntemi, 7 durumda gömülü FS yöntemi ve 2 durumda filtre FS yöntemi performans metriklerine göre en iyi sonucu vermiştir. Genellikle FS uygulanarak seçilen veriler ile yapılan tahminlerin performansı, FS uygulanmadan VS1 ve VS2 setleri ile yapılan tahminlerin performansına göre daha başarılıdır. Orijinal veri setlerine kıyasla daha az özellik kullanılarak gerçekleştirilen tahminlerde hesaplama yükü azalmıştır. Böylece ML modellerinin eğitim sürelerinin kısalarak FS yöntemlerinin olumlu etkileri gözlemlenmiştir. Şekil 9 - Şekil 16 arasında FS yöntemlerinin seçtiği özellikler ile yapılan tahminlerin performansı verilmiştir.

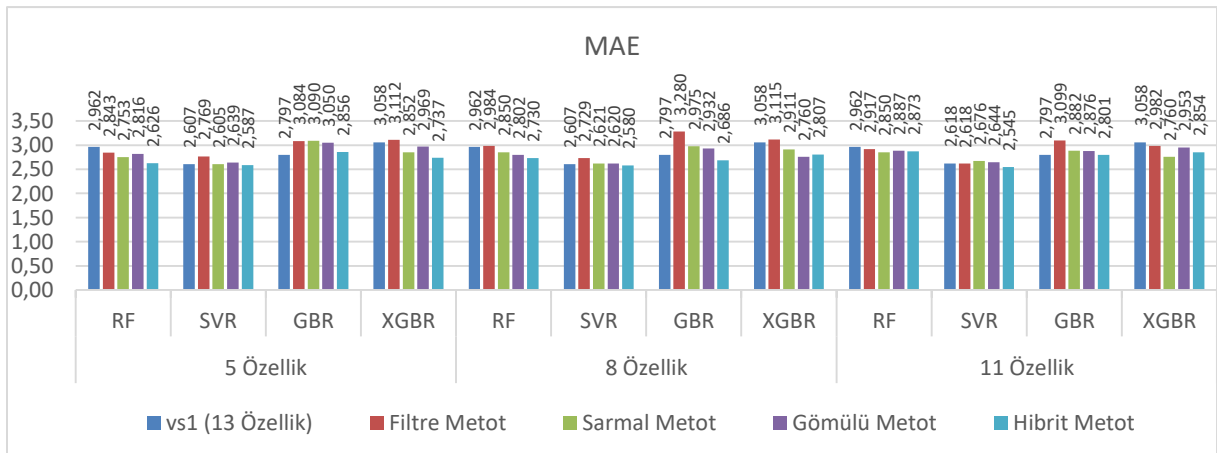
Sonraki çalışmalarda, bu çalışmada kullanılan farklı yöntemler ile hibrit yöntemler geliştirilebilir ve bunlar uygulanarak yeni özellik seçim yöntemleri elde edilebilir. Bu şekilde en iyi FS yöntemleri kombinasyonu oluşturarak sonuçların iyileştirilmesi amaçlanabilir.



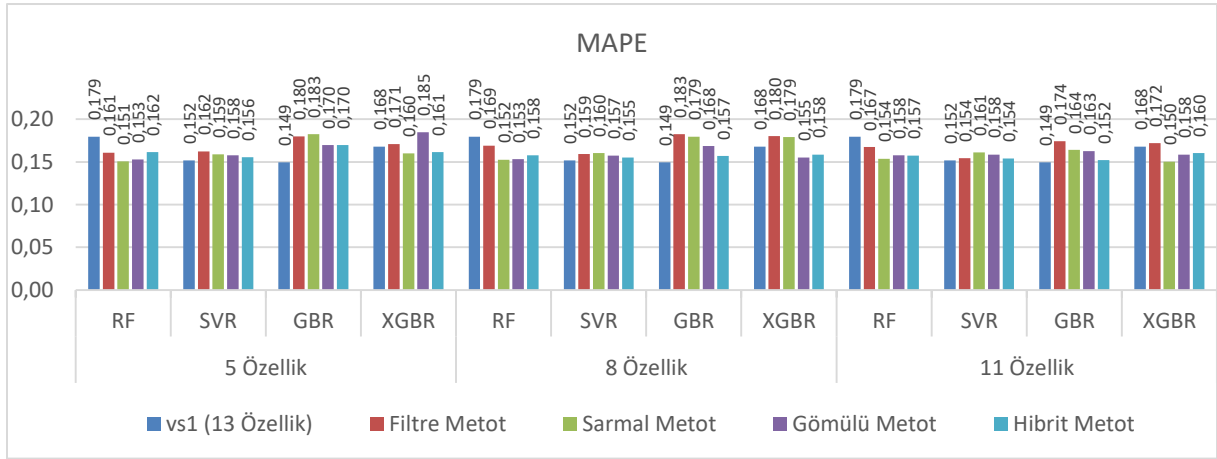
Şekil 9. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MSE değerlerinin karşılaştırılması



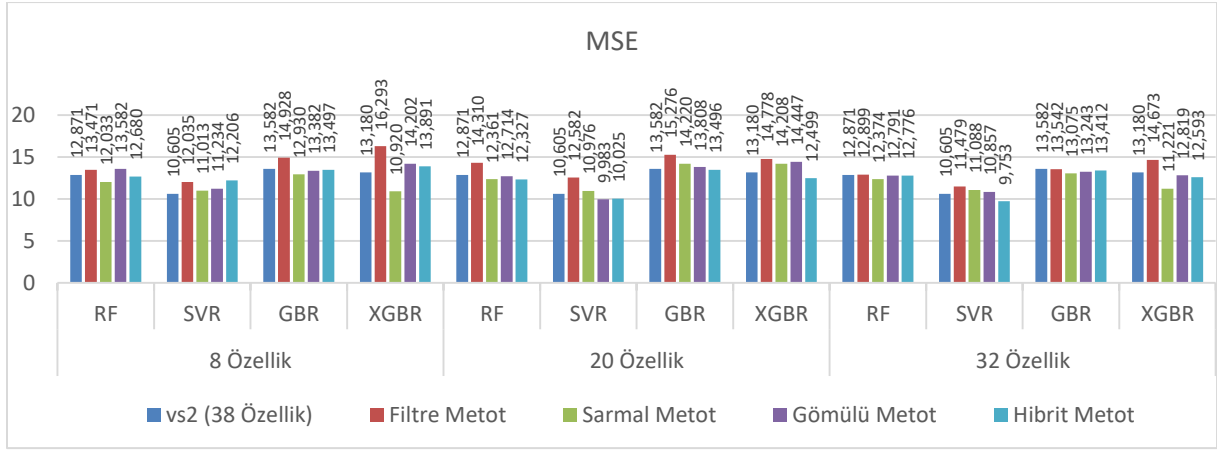
Şekil 10. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre R^2 değerlerinin karşılaştırılması



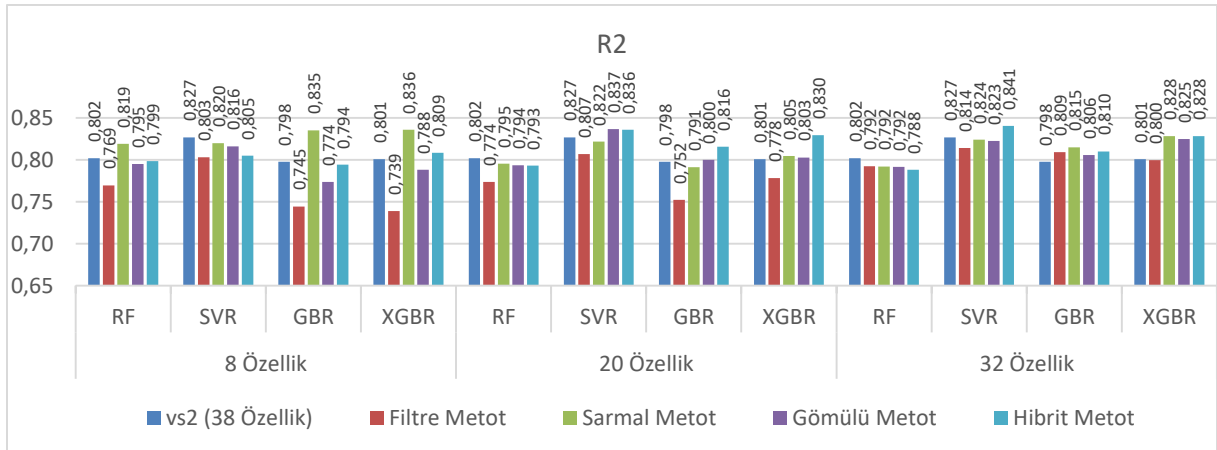
Şekil 11. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAE değerlerinin karşılaştırılması



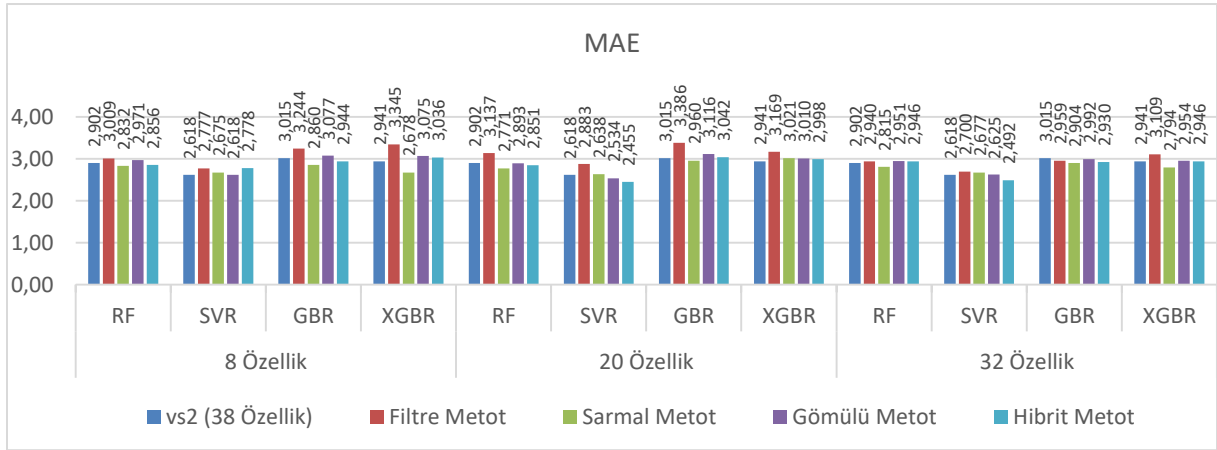
Şekil 12. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAPE değerlerinin karşılaştırılması



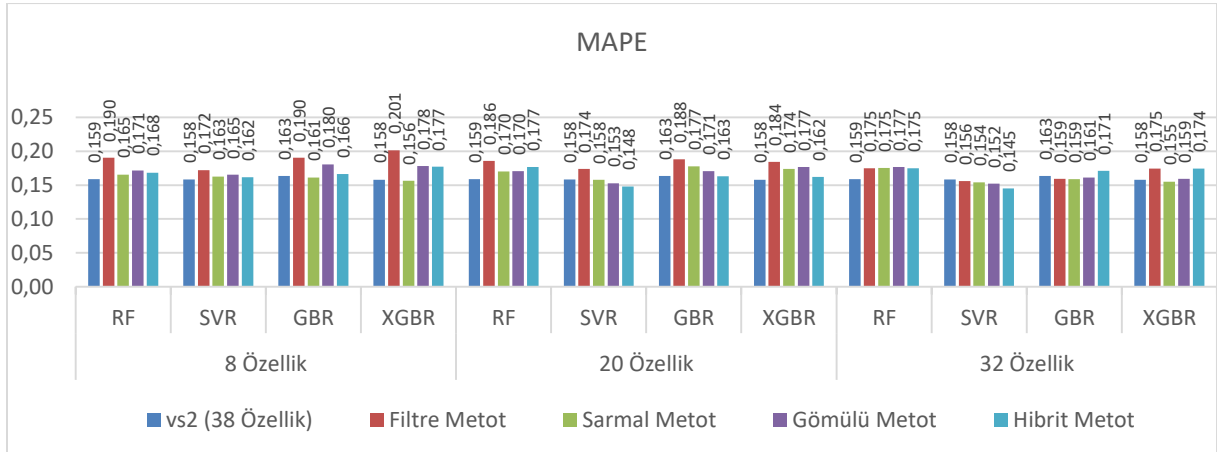
Şekil 13. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MSE değerlerinin karşılaştırılması



Şekil 14. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre R² değerlerinin karşılaştırılması



Şekil 15. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAE değerlerinin Karşılaştırılması



Şekil 16. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAPE değerlerinin karşılaştırılması

V. KAYNAKLAR

- [1] C. J. Laviea, A. D. Schuttera, P. Partoa, E. Jahangira, P. Kokkinosb, F. B. Ortégac, R. Arenad and R. V. Milania, "Obesity and Prevalence of Cardiovascular Diseases and Prognosis- The Obesity Paradox Updated," *Progress in Cardiovascular Diseases*, vol. 58, no. 5, pp. 537-547, 2016.
- [2] F. McLellan, "Obesity rising to alarming levels around the world," *The Lancet*, vol. 359, no. 9315, pp. 1412, 2002.
- [3] C. L. Edelman, C. L. Mandle and E. C. Kudzma, "Health promotion throughout the life span-e-book," *Elsevier Health Sciences*, 2017.
- [4] E. Sukic', A. Katic', E. Stokic', A. Kupusinac and O. Rankov, "What kind of relationship is between body mass index and body fat percentage?," *J. Med. Syst.*, vol. 41, pp. 1-5, 2016.
- [5] A. Kupusinac, E. Stokić and R. Doroslovački, "Predicting body fat percentage based on gender, age and BMI by using artificial neural networks," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 610-619, 2014.

- [6] P. Raj, W. Leslie, L. Lix, and S. Majumdar, "Relationship Among Body Fat Percentage, Body Mass Index, and All-Cause Mortality," *Annals of Internal Medicine*, vol. 164. No. 8, ss. 532-543, 2016.
- [7] B. Baraklı ve A. Küçüker, "Karar Destek Makineleri ve Rastgele Orman Ağaçları Yöntemleri ile Vücut Yağ Yüzdesinin Tahmini," *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, c. 6, s. 4, ss. 430-445, 2018.
- [8] M. K. Uçar, Z. Uçar, F. Köksal, and N. Daldal, "Estimation of body fat percentage using hybrid machine learning algorithms," *Measurement*, vol. 167, 2021.
- [9] G. Chandrashekar, F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, 2014, pp.16-28.
- [10] A. Kupusinac, E. Stokić and R. Doroslovački, "Predicting body fat percentage based on gender, age and BMI by using artificial neural networks," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 610-619, 2014.
- [11] Y. E. Shao, "Body Fat Percentage Prediction Using Intelligent Hybrid Approaches," *The Scientific World Journal*, vol. 2014, 2014.
- [12] T. Ferenci, and L. Kovács, "Predicting body fat percentage from anthropometric and laboratory measurements using artificial neural networks," *Applied Soft Computing*, vol. 67, pp. 834-839, 2018.
- [13] F. Keivanian, R. Chiong and Z. Hu, "A Fuzzy Adaptive Binary Global Learning Colonization-MLP model for Body Fat Prediction," *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*, Paris, France, 2019, pp. 1-4.
- [14] R. Chiong, Z. Fan, Z. Hu and F. Chiong, "Using an improved relative error support vector machine for body fat prediction," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105749, 2020.
- [15] S.A. Hussain, N. Cavus, and B. Sekeroglu, "Hybrid Machine Learning Model for Body Fat Percentage Prediction Based on Support Vector Regression and Emotional Artificial Neural Networks," *Appl. Sci.* vol. 11, s. 9797, 2021.
- [16] K. W. DeGregory, P. Kuiper, T. DeSilvio, J. D. Pleuss, R. Miller, J. W. Roginski, C. B. Fisher, D. Harness, S. Viswanath, S. B. Heymsfield, I. Dungan and D. M. Thomas, "A review of machine learning in obesity," *Obesity Reviews*, vol. 19, no. 5, pp. 668-685, 2018.
- [17] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1st ed., New York, USA, Springer, 1995, pp. 15-32.
- [18] E. Pekel, "Estimation of soil moisture using decision tree regression," *Theoretical and Applied Climatology*, vol. 139, pp. 1111-1119, 2020.
- [19] G. Fan, S. E. Ong and H. C. Koh, "Determinants of House Price: A Decision Tree Approach," *Urban Studies*, vol. 43, no. 12, pp. 2301–2315, 2006.
- [20] Scikit-learn: Machine Learning in Python. (2022, May 01). *Decision Trees*. [Online]. Available: <https://scikit-learn.org/stable/modules/tree.html#regression>

- [21] M. R. Segal, "Machine Learning Benchmarks and Random Forest Regression," *UCSF: Center for Bioinformatics and Molecular Biostatistics*, 2004.
- [22] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 1947-1958, 2003.
- [23] M. Awad, and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, Apress, pp. 67-80, 2015.
- [24] V. Cherkassky, and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Networks*, vol. 17, no. 1, 2004, pp. 113-126.
- [25] K. Ito and R. Nakano, "Optimizing Support Vector regression hyperparameters based on cross-validation," *Proceedings of the International Joint Conference on Neural Networks*, Oregon, USA, 2003, pp. 2077-2082.
- [26] A. J. Smola, and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, 2004, pp. 199-222.
- [27] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., California, USA, O'Reilly, 2019.
- [28] R. Gholami, and N. Fakhari, "Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications," in *Handbook of Neural Computation*, Academic Press, 2017, ch. 27, pp. 515-535.
- [29] R. Stoean, D. Dumitrescu, M. Preuss and C. Stoean, "Evolutionary Support Vector Regression Machines," *2006 Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, Timisoara, Romania, 2006, pp. 330-335.
- [30] Y. Zhang, and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C*, vol. 58, 2015, pp. 308-324.
- [31] A. Natekin, and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013.
- [32] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, 2001, pp. 1189-1232.
- [33] T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, California, USA, 2016, pp. 785-794.
- [34] P. Carmona, F. Climent, and A. Momparler, "Predicting failure in the U.S. banking sector: An extreme gradient boosting approach," *International Review of Economics & Finance*, vol. 61, 2019, pp. 304-323.
- [35] O. Sagi, and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Information Sciences*, vol. 572, 2021, pp. 522-542.
- [36] V. Kumar and S. Minz, "Feature Selection: A literature Review," *Smart Computing Review*, vol. 4, no. 3, pp. 211-229, 2014.

- [37] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2015, pp. 1200-1205.
- [38] J. Miao, and L. Niu, "A Survey on Feature Selection," *Procedia Computer Science*, vol. 91, pp. 919-926, 2016.
- [39] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," *Journal of Petroleum Science and Engineering*, vol. 208, no. E, 2022.
- [40] B. Frénay, G. Doquire, and M. Verleysen, "Is mutual information adequate for feature selection in regression?," *Neural Networks*, vol. 48, 2013, pp. 1-7.
- [41] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *American Physical Society*, vol.69, no.6, 2004.
- [42] S. Raschka. (2022, Mar 21). *SequentialFeatureSelector: The popular forward and backward feature selection approaches incl. floating variants*. 2021. [Online]. Available: http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/.
- [43] M. Karagiannopoulos, D. Anyfantis, S. B. Kotsiantis, and P. E. Pintelas, "Feature Selection for Regression Problems," *8th HellenicEuropean Research on Computer Mathematics and Its Applications*, Athens, Greece, 2007.
- [44] F.J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative Study of Techniques for Large-Scale Feature Selection," *Machine Intelligence and Pattern Recognition*, vol. 16, pp. 403-413, 1994.
- [45] Y. Saeys, T. Abeel, and Y. V. Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," in *Lecture Notes in Computer Science*, Berlin, Heidelberg, Springer, 2008.
- [46] M. Kuhn, and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed., Northwest, USA, CRC Press, 2020, pp. 227-240.
- [47] U. Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," *The American Statistician*, vol. 63, no. 4, pp. 308-319, 2009.
- [48] C. Strobl, A. Boulesteix, A. Zeileis and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 25, 2007.
- [49] G. Louppe, "Understanding Random Forests From Theory To Practice," Ph.D. dissertation, Faculty of Applied Sciences Department of Electrical Engineering & Computer Science, University of Liège, Liège, Belgium, 2014.
- [50] Wikipedia. (2022, Mar 25). *Random forest*. [Online]. Available: https://en.wikipedia.org/wiki/Random_forest.

- [51] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genetics*, vol. 19, no. 65, 2018.
- [52] M. B. Kursa, W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software*, vol. 36, no. 11, 2010.
- [53] X. Chen, and J. C. Jeong, "Enhanced Recursive Feature Elimination," *Sixth International Conference on Machine Learning and Applications*, Cincinnati, USA, 2007, pp. 429-435.