T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VÜCUT YAĞ YÜZDESİ TAHMİNİ İÇİN ÖZELLİK SEÇİM YÖNTEMLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

Asude ALTIPARMAK BİLGİN

Enstitü Anabilim Dalı	**:**	**ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ**
Enstitü Bilim Dalı	**:**	**ELEKTRİK**
Tez Danışmanı	**:**	**Dr. Öğr. Üyesi Burhan BARAKLI**

Haziran 2022

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VÜCUT YAĞ YÜZDESİ TAHMİNİ İÇİN ÖZELLİK SEÇİM YÖNTEMLERİNİN KARŞILAŞTIRILMASI

YÜKSEK LİSANS TEZİ

Asude ALTIPARMAK BİLGİN

Enstitü Anabilim Dalı	**:**	**ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ**
Enstitü Bilim Dalı	**:**	**ELEKTRİK**

Bu tez 16.06.2022 tarihinde aşağıdaki jüri tarafından oybirliği ile kabul edilmiştir.

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Asude ALTIPARMAK BİLGİN

16.06.2022

TEŞEKKÜR

Lisans ve yüksek lisans eğitimimde tez konumun belirlenmesi, yürütülmesi ve kabulü süresince severek çalıştığım ve desteklerini esirgemeyen değerli danışmanım Dr. Öğr. Üyesi Burhan BARAKLI'ya teşekkürlerimi sunarım.

Eğitimim süresince maddi ve manevi olarak sürekli destek olan ve kendilerini daima yanımda hissettiren sevgili aileme sonsuz teşekkür ederim.

İÇİNDEKİLER

TEŞEKKÜR ……..…… i
İÇİNDEKİLER ……..…… ii
SİMGELER VE KISALTMALAR LİSTESİ …… iv
ŞEKİLLER LİSTESİ ……. vi
TABLOLAR LİSTESİ …….. viii
ÖZET ……. ix
SUMMARY …….. x

BÖLÜM 1.
GİRİŞ …….……..……..…….. 1
1.1. Vücut Yağ Yüzdesine Genel Bakış …….…… 1
1.2. Literatür Taraması …….…….…….... 6
1.3. Tezin Bölümleri …….…….…….….. 9

BÖLÜM 2.
MAKİNE ÖĞRENMESİ …….…….……..... 11
2.1. Regresyon Modelleri …….…….…… 12
2.1.1. Rastgele orman ağaçları …….…… 14
2.1.2. Destek vektör makineleri …….…….…….. 17
2.1.3. Gradyan arttırma regresyon ağacı …….…….…….. 23
2.1.4. Aşırı gradyan arttırma regresyonu …….…….……. 25

BÖLÜM 3.
ÖZELLİK SEÇİMİ …….…….…….…….…….. 29
3.1. Filtre Özellik Seçim Yöntemi …….…….…….……. 33
3.1.1. Karşılıklı bilgi ile özellik seçimi …….…….…… 35

3.1.2. Tek değişkenli istatistiksel test ile özellik seçimi ……..…….. 37
3.2. Sarmal Özellik Seçim Yöntemi …….…….…….…….. 40
3.2.1. Sıralı ileriye doğru özellik seçimi …….…….…….. 42
3.2.2. Sıralı geriye doğru özellik seçimi …….…….…….. 43
3.3. Gömülü Özellik Seçim Yöntemi …….…….…….……. 44
3.3.1. Rastgele orman ağaçları makine öğrenim algoritması kullanılarak özelliklerin önemine dayalı özellik seçimi …. 46
3.3.2. Rastgele orman ağaçları makine öğrenim algoritması kullanılarak özelliklerin önemine dayalı yinelemeli özellik seçimi …….…….…….…….……. 48
3.4. Hibrit Özellik Seçim Yöntemi …….…….…….…… 49
3.4.1. Rastgele karıştırma ile hibrit özellik seçimi …….…… 50
3.4.2. Yinelemeli özellik eleme ile hibrit özellik seçimi …… 52

BÖLÜM 4.
DENEYSEL ÇALIŞMALAR VE SONUÇLAR …….…….…… 54
4.1. Veri Setine Genel Bakış …….…….…….…….. 54
4.2. Regresyon Parametreleri ve Karşılaştırma Metrikleri …….…….. 58
4.2.1. Parametre ayarları …….…….…….……. 58
4.2.2. Regresyon başarımı karşılaştırma metrikleri …….…….. 59
4.3. Özellik Seçiminde Kullanılan Regresyon Yöntemleri …….…….. 60
4.4. Deneysel Sonuçlar …….…….…….…….……. 61
4.5. Sonuçlar …….…….…….…….…… 74

KAYNAKÇA …….…….…….…….…… 75
ÖZGEÇMİŞ …….…….…….…….. 80

SİMGELER VE KISALTMALAR LİSTESİ

$f(x,\alpha)$	: $\hat{y}$ için tahmin fonksiyonu
$L(y,y)$	: Kayıp fonksiyonu
$R(\alpha)$	: Risk fonksiyonu
X	: Özellik kümesi
$\hat{y}$	: Bağımlı değişken y 'nin tahmin değeri
$\overline{y}$	: Bağımlı değişkenlerin ortalaması
y_j	: j örneği için bağımlı değişkenin gerçek değeri
$\hat{y}_j$	: j örneği için bağımlı değişkenin tahmin değeri
GBR	: Gradyan arttırma regresyon
MAE	: Medyan mutlak hata
MAPE	: Ortalama mutlak yüzde hatası
MI	: Karşılıklı bilgi
ML	: Makine öğrenimi
MSE	: Ortalama karesel hata
R^2	: Belirleme katsayısı
RF	: Rastgele orman ağaçları
RFE	: Yinelemeli özellik eleme
RFI	: Rastgele orman ağaçları ile özellik önemine dayalı özellik seçimi
RRFI	: Rastgele orman ağaçları ile özellik önemine dayalı yinelemeli özellik seçimi
RS	: Rastgele karıştırma yöntemi
SBS	: Sıralı geriye doğru seçim
SFS	: Sıralı ileriye doğru seçim

SVM	: Destek vektör makineleri
SVR	: Destek vektör regresyonu
UST	: Tek değişkenli istatistiksel test
XGBR	: Aşırı gradyan arttırma regresyonu

ŞEKİLLER LİSTESİ

Şekil 2.1. Karar ağaçlarının genel yapısı ……..…….. 14
Şekil 2.2. Rastgele orman ağaçlarının genel yapısı …….…….. 16
Şekil 2.3. ε-duyarsız yaklaşımı ile doğrusal destek vektör makineleri şematik gösterimi …….……..…….…….. 19
Şekil 2.4. ε-duyarsız yaklaşımı ile doğrusal olmayan destek vektör makineleri şematik gösterimi …….…….……. 21
Şekil 2.5. Hata ve iterasyona bağlı olarak gradyan artırma regresyonunun şematik gösterimi …….…….…….…….. 23
Şekil 2.6. Karar ağaçlarına dayalı çözümler …….……. 28
Şekil 3.1. Özellik seçim işlemleri için temel adımlar …….…… 31
Şekil 3.2. Filtre FS modeli genel yapısı…….…….. 34
Şekil 3.3. Karşılıklı bilgi ile filte özellik seçim yöntemi akış diyagramı……. 37
Şekil 3.4. Tek değişkenli istatistiksel test ile özellik seçim yöntemi akış diyagramı …..…….…….…… 39
Şekil 3.5. Sarmal FS modeli genel yapısı …….…….. 40
Şekil 3.6. Sıralı ileriye doğru özellik seçim yöntemi akış diyagramı …….. 42
Şekil 3.7. Sıralı geriye doğru özellik seçim yöntemi akış diyagramı ……..… 44
Şekil 3.8. Gömülü FS modeli genel yapısı …….…….…….……... 45
Şekil 3.9. Rastgele orman ağaçları tarafından belirlenen özelliklerin önemine dayalı özellik seçim yöntemi akış diyagramı …….……...….. 47
Şekil 3.10. Rastgele orman ağaçları tarafından belirlenen özelliklerin önemine dayalı yinelemeli özellik seçim yöntemi akış diyagramı …….…… 49
Şekil 3.11. Hibrit FS modeli genel yapısı …….…….……...….. 50
Şekil 3.12. Rastgele karıştırma ile hibrit özellik seçim yöntemi akış diyagramı .. 51
Şekil 3.13. Yinelemeli özellik eleme ile hibrit özellik seçim yöntemi akış diyagramı …….…….…….…….……...… 53

Şekil 4.1. Özellik seçim işlemleri için temel adımlar ……..…….…….. 54
Şekil 4.2. Özellikler arasındaki korelasyon ilişkisini gösteren renkli ısı haritası ... 56
Şekil 4.3. VS1'deki özelliklerin ikili ilişkileri …….…….…….…. 57
Şekil 4.4. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MSE değerlerinin karşılaştırılması …….…….……... 70
Şekil 4.5. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre R^2 değerlerinin karşılaştırılması …….…….……... 70
Şekil 4.6. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAE değerlerinin karşılaştırılması …….…….……... 71
Şekil 4.7. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAPE değerlerinin karşılaştırılması …….…….…… 71
Şekil 4.8. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MSE değerlerinin karşılaştırılması …….…….……... 72
Şekil 4.9. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre R^2 değerlerinin karşılaştırılması …….…….……... 72
Şekil 4.10. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAE değerlerinin karşılaştırılması …….…… 73
Şekil 4.11. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAPE değerlerinin karşılaştırılması …….…….. 73

TABLOLAR LİSTESİ

Tablo 4.1. VS2 veri setindeki 25 özelliğin BFP ile ilişki katsayıları …….. 55
Tablo 4.2. VS1 için özelliklerin gözlem değerleri hakkında bilgi …..…….. 56
Tablo 4.3. VS1 için seçilen özellik sayısına göre model eğitim süresi (sn) ……. 62
Tablo 4.4. VS2 için seçilen özellik sayısına göre model eğitim süresi (sn) ……. 62
Tablo 4.5. VS1 ve VS2 için özellik seçimi yapılmadan önce modellerin performansı ……. 63
Tablo 4.6. VS1'e uygulanan farklı FS yöntemlerinin RF modeline göre performansları ……. 64
Tablo 4.7. VS1'e uygulanan farklı FS yöntemlerinin SVR modeline göre performansları ……. 64
Tablo 4.8. VS1'e uygulanan farklı FS yöntemlerinin GBR modeline göre performansları ……. 65
Tablo 4.9. VS1'e uygulanan farklı FS yöntemlerinin XGBR modeline göre performansları ……. 65
Tablo 4.10. VS2'ye uygulanan farklı FS yöntemlerinin RF modeline göre performansları ……. 66
Tablo 4.11. VS2'ye uygulanan farklı FS yöntemlerinin SVR modeline göre performansları ……. 66
Tablo 4.12. VS2'ye uygulanan farklı FS yöntemlerinin GBR modeline göre performansları ……. 67
Tablo 4.13. VS2'ye uygulanan farklı FS yöntemlerinin XGBR modeline göre performansları ……. 67
Tablo 4.14. FS metotları kullanılarak ML modeline göre performansı gelişen durumların sayısı ……. 69

ÖZET

Anahtar kelimeler: Özellik seçimi, makine öğrenmesi, vücut yağ yüzdesi

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 ölçülmesi için çeşitli cihaz ve ekipman mevcuttur. Ancak bunlar ile ölçümün gerçekleştirilmesi yüksek maliyetlidir ve kullanımları sınırlıdır. Bu değerin 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. Bir regresyon problemi olarak ele alınan vücut yağ yüzdesi, literatürdeki birçok çalışma tarafından başarılı şekilde tahmin edilmiştir. Ancak veri setindeki yüksek boyutlu, alakasız ve gereksiz verilerin bulunması 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. Özellik seçimi sayesinde daha az veri ile çalışmak işlem yükünün azalmasının yanı sıra boyut probleminin etkisini azaltarak modelin tahmin performansını geliştirebilir, öğrenme süreci hızlandırır ve problemin makine öğrenmesi modelleri tarafından daha iyi anlaşılmasını sağlar.

Bu çalışmada vücut yağ yüzdesi tahmini için sekiz farklı özellik seçim algoritması 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.

COMPARISON OF FEATURE SELECTION METHODS FOR ESTIMATION OF BODY FAT PERCENTAGE

SUMMARY

Keywords: Feature selection, machine learning, body fat percentage

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.

Various devices and equipment are available for measuring body fat percentage. However, these are costly to measure and their use is limited. Determining this value quickly, easily, inexpensively and with high accuracy is as important as diagnosing obesity. It is possible to reliably calculate the body fat percentage value, which can be calculated from anthropometric data with machine learning algorithms. Body fat percentage, which is considered as a regression problem, has been successfully estimated by many studies in the literature. However, the presence of high-dimensional, irrelevant and redundant data in the data set distort the accuracy of machine learning algorithms and increases the training time of the model. There are feature selection algorithms that provide higher accuracy by using machine learning algorithms with fewer features. Thanks to feature selection, working with less data reduces the computational cost, as well as reducing the effect of the size problem, improving the prediction performance of the model, accelerating the learning process and providing a better understanding of the problem by machine learning models.

In this study, eight 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.

BÖLÜM 1. GİRİŞ

1.1. Vücut Yağ Yüzdesine Genel Bakış

Obezite son yıllarda önemli bir sağlık problemi haline gelmiştir. Bireyin sosyal yaşamına olan olumsuz etkilerinin yanında birçok sağlık problemine de sebep olmaktadır. Obezite, aşırı beslenme durumunda yağ hücrelerinin depolanmasından kaynaklanır. Yağ kütlesinin, yağsız kütleye oranındaki dengesizlik olarak tanımlanmaktadır (Edelman ve ark., 2014, McLellan, 2002).

Gelişmiş ülkelerde genellikle beslenme alışkanlıklarından dolayı obez birey sayısının fazla olması beklenilen bir durumdur. Ancak gelişmekte olan ülkeler arasında da hızla yayılması bu durumun artık küresel bir problem olarak ele alınmasına sebep olmuştur.

Yüksek obezite oranı, hızlı yiyecek zincirlerinin ve bu yiyecekleri tüketenlerin sayısının fazla olduğu ABD gibi ülkeler ile bağdaştırılsa da aslında sadece bu gibi ülkeler ile sınırlı değildir. Örneğin günlük olarak hızlı tüketim oranının az olduğu Çin'de bile son yıllara göre aşırı kiloluluk oranı erkeklerde 3 katına çıktığı, kadınlarda da hipertansiyon oranının ABD'ye benzer olarak 2 katına çıktığı görülmektedir. Ayrıca, dünya çapında her gün teşhis edilen diyabet vakalarının yarısından fazlası Hindistan ve Çin'e aittir. Mısır'daki kadınların da neredeyse yarısı aşırı kiloludur ve Mısır ABD ile aynı diyabet oranına sahiptir. Tanzanya gibi en fakir ülkelerde bile obezite ve diyabet vakaları artış göstermektedir (McLellan, 2002). Dolayısıyla hızlı tüketim, obezite probleminin artmasında en büyük neden değildir. Bu durumdan sorumlu birçok etken bulunmaktadır. Örneğin, ihracat maliyeti düşük ve tadı lezzetli olan şekerin birçok yiyeceğin içerisine eklenmesi ekstra kalori artışına neden olmaktadır. Ayrıca teknolojik gelişmeler ile pekte sağlıklı olmayan mısır, soya, pamuk gibi çeşitli tohumlardan elde edilen yenilebilir yağların da şeker gibi sıkça yiyeceklere

eklenmesi obezite oranının artmasından sorumlu hale gelmiştir. Bu gibi sebepler ile günlük alınan kalori miktarı artarken teknolojik gelişmelerin bir etkisi olarak harcanan kalori miktarının da sınırlı olması aradaki dengeleri bozmaktadır.

Obezite, kalp-damar hastalıkları, kanser, hipertansiyon, diyabet gibi birçok farklı hastalık türüne de neden olmaktadır (McLellan, 2002). Kalbin yapısına ve işlevine olumsuz etkisinden dolayı kan basıncının artması, gluko/metabolic sendrom, hipertansiyon, koroner kalp hastalığı, kalp yetmezliği, atriyal fibrilasyon gibi farklı kalp damar hastalıklarının oluşmasına sebep olmaktadır (Lavie ve ark., 2016).

Obezite sadece yetişkinler için bir tehlike değildir. Okul çağındaki çocuklar arasında da hızla artmaktadır. Çocuklar için akademik başarıların daha önemli görülmesinden dolayı zamanlarının çoğunu okul işleri için harcamaktadır. Bu durum aktivite sürelerini azalmasına ve dolayısıyla obezitenin çocuklar arasında artmasına sebep olmaktadır. Ayrıca çocukların oyun alanlarının binalar ile işgal edilerek fiziksel aktivitelerinin kısıtlanması da obezitenin artmasında etkilidir. Hem çocuklarda hem de yetişkinlerde giderek artan bu tehlikeli durum, bireyleri ve devletleri ekonomik açıdan da etkilemektedir (McLellan, 2002, Edelman ve ark., 2014).

Obezite, hemen müdahale edilmesi gereken bir problemdir. Tedavi edilmediği müddetçe başlıca kalp damar hastalıkları gibi vücutta birçok tahribata sebep olacak bir faktördür. Obezitenin tedavisi, tüm yaş grupları için mümkün olduğu kadar erken başlanmalıdır (Ortega ve ark., 2016). Obezitenin, tedavisi kadar erken teşhisi de önemlidir. Erken tedavi edilmeye başlanıldığında obezitenin vücutta yol açabilceği zararlar da en aza indirilebilmeye çalışılır (Csige ve ark., 2018).

Dünya Sağlık Örgütü tarafından ağırlık ve boy bilgileri ile hesaplanan vücut kitle indeksinin (BMI) obezite teşhisinde kullanılabileceği onaylanmıştır. BMI genel beslenme durumunun bir göstergesi olarak vücut ağırlığının vücut boyunun karesine oranı olarak ifade edilir. Değer, 25kg/m^2 değerinden büyük olduğu durumlar aşırı kiloluğa karşılık gelirken, 30 kg/m^2 olduğu durumlar obeziteye karşılık gelmektedir. BMI, sadece kas ve kemik kütlesi gibi yağsız kütleyi ölçtüğü ve bireyleri sadece kilo

durumlarına göre gruplandırdığı için vücut hakkında yeterli bir bilgi sağlamamaktadır. Aynı BMI değerine sahip birçok farklı çeşitte vücut yapısı olabilir. Vücut yağı yüzdesi (BFP) vücut bilgilerini yaşa cinsiyete vs. göre gruplar. Obeziteyi değerlendirmek için vücudun gerçek durumunu gösteren BFP değerini ölçmek daha doğru bir yoldur (Srdić ve ark., 2012, Kupusinac ve ark., 2017, Stokić ve ark., 2014)

Aşırı kiloluk, genellikle BMI değerinin 25kg/m^2 den büyük olduğu durumlardır. BMI 22.5-25 kg/m^2 olan bireyler en uygun yaşam kalitesine sahiptir. BMI 30-35 kg/m^2 olan orta derecede obeziteye sahip bireylerin 3 yıl, BMI 40-50 kg/m^2 olan aşırı obeziteye sahip bireylerin 10 yıl kadar yaşam süreleri kısalmaktadır. Aşırı obezitenin yaşam süresine etkisi, ömür boyu sigara içmenin etkisi aynıdır. Genel olarak BMI vücut ölçüsünü ölçmek için kolay bir yol olarak görülse de sadece iki parametreye göre hesaplandığından oluşturabileceği sağlık sorunlarını tahmin etmekte yetersiz performans göstermektedir. Bu ölçümlere ek olarak bel çevresi, bel:kalça oranı ve bel:boy oranı gibi alternatif ölçümler kalp damar hastalıklarını ve ölüm riskini öngörmek için daha doğru bilgi sağlar. Bir birey normal BMI değerine sahip olmasına rağmen orantısız olarak yüksek bel çevresi ölçüsüne sahip olabilir. Bu yüzden kilo yönetiminin etkisi ve kilo ile ilişkili hastalıkların tespiti önemli bir hal almıştır. BMI değerinin dikkate alınmasından ziyade BFP dağılımı son yıllarda daha önemli hale gelmiştir (Huxley ve ark., 2010, Hu ve ark., 2017). Bel:boy oranının hipertansiyon, diyabet, kalp damar hastalığı gibi riskli durumların göstergesi olarak en ilişkili olduğu, BMI'nın ise en ilişkisiz gösterge olduğu belirlenmiştir (Leea ve ark., 2008).

Ağırlık, vücut yağı hakkında bilgi sağlayan ölçülmesi kolay bir nicelik olsa da genel vücut yapısı hakkında bilgi olmadığında tek başına anlamlı bir ifade değildir (Ferenci ve Kovács, 2018). Literatürde, klinik tartışmalarda ya da hayat sigortası sektöründen gelen raporlarda ağırlık, obezitenin veya şişmanlığın bir ölçütü olarak alınır. Herhangi bir ağırlık indeksinin şişmanlık ya da vücut yağı bilgisi için kabul edilebilir bir gösterge olması mümkün değildir. En iyi gösterge, vücut boyu ile en düşük ve vücut yağı ile en yüksek korelasyona sahip olan bilgidir (Keys ve ark., 1972).

Çalışmalar gösteriyor ki 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. BFP değeri kalp damar hastalıkları, diyabet, metabolik hastalıklar ile ilgili oluşabilecek riskleri öngörmek için daha önemli bir göstergedir. Ayrıca yaşlanma sürecinde boy uzunluğu azalırken, yağ dokusu genellikle viseral yağ olarak toplanır. Yani vücut ağırlığı ve BMI değeri sabit kalsa bile BFP değeri artmaktadır. Bununla ilgili olarak olarak aynı BMI değerine sahip kadın ve erkek arasından göre BFP değerleri erkeklere göre daha yüksek olabilir (Kupusinac ve ark., 2017, Stokić ve ark., 2014).

BFP, toplam vücut yağı ölçümüdür ve obezitenin doğru teşhisi için gereklidir. Vücut yağ yüzdesinin tahmini için antropometri (vücut kütlesi, vücuttaki belirli 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. Antropometrik ölçümler vücut yağ oranı tahmini için yaygın olarak kullanılır ve düşük maliyetlidir. BMI, vücut yağ kütlesi hakkında yeterli bilgi sağlamamaktadır. BMI, vücudun beslenme durumunun bir göstergesidir, yağ kütlesi ölçümü değildir. BFP tahmini için BMI, yaş ve cinsiyet gibi veriler kullanılarak birçok formül ile hesaplanabilir. Ancak BMI ve BFP arasındaki ilişkinin doğrusal olup olmadığı konusunda belirsizlikler ve tartışmalar vardır. DEXA ve UWW teknikleri antropometrik ölçümlerden vücut yağı için daha doğru tahminler gerçekleştirir ancak bu ölçümleri gerçekleştirmek yüksek maliyetlidir ve kullanımları sınırlıdır. Vücut yağının çeşitli formüller ile yaş, cinsiyet ve BMI değerine bağlı olarak tahmin etmek hem düşük maliyetli olduğundan hem de tıbbi bir cihaz gerektirmediğinden geniş çapta kullanımı uygun görülmüştür (Stokić ve ark., 2014). Bu yüzden BFP değerinin temel sosyodemografik veriler, temel antropometrik veriler ve rutin kan alımından elde edilen temel laboratuvar parametreleri gibi daha düşük maliyet ile ölçülebilir parametrelerden tahmin edilebilmesi önemlidir (Fernández-Sánchez ve ark., 2011, Ferenci, 2013, Ferenci ve Kovács, 2018).

Araştırma alanlarında makine öğrenmesi uygulamaları giderek artmaktadır. BFP tahmini, makine öğrenimi (ML) algoritmaları ile gerçekleştirilebilmekte ve bir regresyon problemi olarak ele alınmaktadır (Baraklı ve Küçüker, 2018). Regresyon problemleri ile genellikle bağımsız değişkenler ile bağımlı değişkenin sürekli değerinin tahmini gerçekleştirilir. Bir regresyon modelinin probleme uygun olması doğru sonuçların alınması açısından önemlidir. Yanlış model seçimi modelin yanlış tahminler yapmasına neden olabilir. Kullanılacak veriye en uygun model araştırılmalıdır (Maulud ve Abdulazeez, 2020). Literatürde BFP tahmini için gerçekleştirilen çalışmalar, genellikle yaş, boy, ağırlık, çeşitli vücut ölçümleri gibi ölçülmesi kolay antropometrik verileri kullanarak ML metotları ile tahmin gerçekleştirmeyi amaçlamıştır (Uçar ve ark., 2021). Bu çalışmalar genellikle herhangi bir tıbbi cihaz kullanmadan düşük maliyet ile yüksek doğrulukta tahminler gerçekleştirmeyi sağlamıştır.

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 modelin tahmin doğruluğu, hesaplama yükü, genelleştirebilme yeteneği gibi birçok açıdan önemlidir. Gereksiz, alakasız ve 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 (Chandrashekar ve Sahin, 2014). Literatürde FS yöntemleri kullanarak ML algoritmaları ile BFP tahmini gerçekleştiren çeşitli çalışmalar bulunmaktadır.

Bu çalışmada literatürdeki benzer çalışmalardan da yola çıkarak BFP tahmini için FS yöntemlerinin modelin performansına etkisi karşılaştırılmıştır. Çalışma 2 ayrı veri seti üzerinde gerçekleşmiştir. Veri setlerine 4 farklı tipte 8 FS yöntemi kullanarak özellik seçimi yapılması sağlanmıştır. FS ile daha az sayıda özellik uzayına sahip olan yeni veri setlerine 4 farklı regresyon modeli uygulanmıştır. Performansların değerlendirilmesinde literatürde sıkça kullanılan 4 performans metriği kullanılmıştır. Genel olarak özellik seçiminin ML algoritmalarının eğitim süreleri üzerindeki etkisi gösterilmiştir.

1.2. Literatür Taraması

BFP tahmini ML algoritmaları ile gerçekleştirilebilmekte ve bir regresyon problemi olarak ele alınmaktadır (Baraklı ve Küçüker, 2018). Literatürdeki BFP tahmini gerçekleştiren çalışmalar incelenmiştir. Birçok çalışma ML metotları ile antropometrik ölçümler kullanılarak problemin çözülmesi amaçlamıştır (Uçar ve ark., 2021). Literatürdeki BFP tahmini çalışmaları genellikle herhangi bir tıbbi cihaz kullanılmadan düşük maliyet ile yüksek doğrulukta tahmin gerçekleştirmeyi amaçlamıştır. Bazı çalışmalar özellik seçim yöntemlerini ekleyerek BFP tahmini gerçekleştirmişlerdir.

Kupusinac A. vd. çalışmalarında daha yüksek doğruluğa sahip BFP tahmini amacıyla yapay sinir ağları (ANN) tekniğini kullanarak yeni bir yaklaşım göstermişlerdir. Yaş, cinsiyet, BMI değerlerini kullanarak BFP tahmini yapmışlardır. Çalışmalarında MATLAB Neural Network araç kutusunu kullanmışlardır. Ortalama kare hata, ortalama mutlak yüzde hatası metriklerini kullanarak çalışmalarını değerlendirmişlerdir. Sonuçlarını literatürde bulunan BFP hesaplama formülleri ile karşılaştırdığında Kupusinac ve arkadaşlarının önerdiği metot literatürdeki çalışmalara benzer maliyet ve komplekslik göstermiş olsa da daha yüksek doğrulukta BFP tahmini gerçekleştirmiştir (Stokić ve ark. , 2014).

Shao Y.E. çalışmasında vücut yağ yüzdesi 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 vücut yağ yüzdesinin tahmini için kullanmıştır. Önerilen bu hibrit model diğer tek aşamalı modellere göre daha iyi tahmin yaptığını sonuçlarda göstermiştir (Shao, 2014).

Ferenci T. vd. çalışmalarında yetişkin erkeklere ait yaş cinsiyet, ağırlık, boy, bel çevresi gibi kolay ölçülebilen laboratuvar sonuçlarını kullanarak vücut yağ yüzdesi değerinin tahminini iyileştirmeyi amaçlamışlardır. Linear regression, ileri beslemeli sinir ağı ve destek vektör makineleri tekniklerini kullanmışlardır. Önyükleme doğrulaması ile en uygun parametreleri seçmişlerdir. Sonuçlarda az bir fark ile destek vektör makineleri tekniğinin diğer iki tekniğe göre daha doğru tahmin yaptığını göstermişlerdir. Hatayı kök ortalama kare hata (RMSE) ve belirleme katsayısı (R^2) metriklerine göre değerlendirmişlerdir (Ferenci ve Kovács, 2018).

Baraklı B. ve Küçüker A. çalışmalarında sosyodemografik ve temel antropometrik bilgilerden oluşan orijinal veri seti ile birlikte bu veri setinden yüksek ilintili belirli sayıda özellikler seçilerek ve temel bileşenler belirlenerek 15 adet veri seti daha üretmişlerdir. Ayrıca özellik seçimi ve özellik azaltımının regresyon yöntemlerinin başarısına olan katkısını incelemişlerdir. Özellik seçimi 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 modellerini kullanarak BFP değerini tahmin etmişlerdir. Performanları değerlendirmek için medyan mutlak hata (MAE), RMSE, R^2, korelasyon katsayısı (R), ortalama mutlak yüzde hatası (MAPE) metriklerini kullanmışlardır. Deneysel çalışmalar sonucunda, SVM modeli ile yeni türetilen veri setleri kullanılarak yapılan çalışmalar orijinal veri seti kullanılarak yapılan çalışmalara göre daha yüksek doğruluğa sahip tahminler gerçekleştirmişlerdir. RF modeli ile SVM modelinin tersi olarak, orijinal veri seti kullanılarak yapılan çalışmalar yeni türetilen setler kullanılarak yapılan çalışmalara göre daha yüksek doğruluğa sahip tahminler gerçekleştirmişlerdir. Ayrıca RF yönteminin SVR yönteminden karşılaştırma metrikleri bakımından daha üstün olduğunu tespit etmişlerdir (Baraklı ve Küçüker, 2018).

Chiago R. vd. çalışmalarında biri antropometrik ölçümleri içeren diğeri fiziksel inceleme ve laboratuvar ölçümleri içeren 2 ayrı veri seti kullanarak BFP tahmini için gelişmiş göreceli destek vektör makineleri (IRE-SVM) yaklaşımı geliştirmiştir. Önerilen metot unbiased bir tahmin modeli elde etmek için amaç fonksiyonuna yanlı

hata kontrolü terimi eklenmesini kapsamaktadır. Ayrıca anlamlı bilgiyi kaybetmeden gereksiz ve ilgisiz özelliklerin kaldırılmasını içeren özellik seçme tekniği uygulamışlardır. SVM, göreceli destek vektör makineleri (RE-SVM), ANN'nin bir tipi olan çok katmanlı algılayıcı, RF, aşırı gradyan arttırma regresyon olmak üzere 5 farklı tahmin modelini çalışmada önerilen model ile karşılaştırmak için kullanmışlardır. Modellerin performansını 4 farklı performans metriklerine göre değerlendirmişlerdir. Wilcoxon testi kullanarak önerilen model için istatistiksel analiz gerçekleştirmişlerdir. Uygulama sonuçlarında her iki veri setinde de kendi önerdikleri yaklaşımlarının diğer yaklaşımlardan daha doğru tahmin gerçekleştirdiğini gözlemlemişlerdir (Chiong ve ark., 2020).

Keivanian F. vd çalışmalarında, çok katmanlı algılayıcı yapay sinir ağı tahmin modeli kullanarak özellik seçimi için yeni bir bulanık uyarlamalı ikili küresel öğrenme kolonizasyon yöntemi önermişlerdir. BFP için antropometrik veriler kullanmışlardır. Önerilen metot ile karşılaştırma yapılması amacıyla çok katmanlı algılıyıcı ile hibritleyerek bazı iyi bilinen meta-sezgisel yöntemin ikili versiyonları çalışmaya dahil etmişlerdir. Sonuçlarını RMSE, , hata kareleri ortalaması (MSE) performans kriterlerine göre değerlendirmişlerdir. Ayrıca metotlar çalışma sürelerine, seçilen özellik sayılarına göre de değerlendirilmiştir. Sonuç olarak önerilen tahmin modeli çok katmanlı algılayıcı tabanlı 4 tane ikili özellik seçme yöntemlerinden ve çoklu regresyon ve çok değişkenli uyarlanabilir regresyon hibrit modelinden üstün başarılar elde edilmiştir (Keivanian ve ark. , 2019).

Uçar M. K. vd. çalışmalarında en az veri kullanarak yüksek doğruluğa sahip vücut yağ yüzdesi tahminini hibrit makine öğrenmesi modelleri kullanarak gerçekleştirmeyi amaçlamışlar. Antropometrik veriler kullanmışlardır. Uygulamada kullanılan PCA analizi temelli özellik seçme algoritması ile farklı özellik alt grupları oluşturarak önerilen hibrit modeller ile durumları test etmişlerdir. SVM, çok katmalı ileri beslemeli sinir ağı (MLFFNN), DT makine öğrenme modellerini kullanmanın yanı sıra bu modellerden elde edilen 4 farklı hibrit model ile birlikte toplam 7 makine öğrenmesi modeli kullanarak deneysel sonuç elde etmişlerdir. Sonuçlarını RMSE, MAPE, ortalama mutlak fark (MAD), standart hata (SH), R, R^2 veMSE performans kriterlerine

göre değerlendirmişlerdir. Sonuç olarak önerilen hibrit modeller ile daha az parametre kullanılması zaman ve ölçüm maliyeti açısından avantaj sağlarken, ekonomik açıdan da fayda sağlamıştır (Uçar ve ark., 2021).

Hussain A. S vd çalışmalarında, destek vektör makineleri ve duygusal yapay sinir ağlarına dayalı hibrit bir model önermişlerdir. Bu hibrit model, destek vektör makineleri kullanılarak özellik seçimi ve duygusal sinir ağları ile vücut yağ oranı tahmini gerçekleştirmektedir. Yüksek tahmin oranları elde etmek ve hedef değişkeni etkileyen önemli faktörleri belirlemek için sinir ağının temel yapısını çeşitli duygusal işlevlerle birleştirmişlerdir. Vücut yağ oranı tahminş için 8 özellik bulunan bir veri seti kullanmışlardır. Sonuçlarını RMSE, bağıl kök ortalama kare hata (RRMSE), R^2 metriklerine göre değerlendirmişlerdir. Çalışmada önerilen metot, birçok makine öğrenmesi algoritması ile karşılaştırıldı. Tüm metrik değerleine göre önerilen metot, diğer makine öğrenmesi algoritmalarından daha üstün sonuçlar elde etmiştir (Hussain ve ark., 2021).

1.3. Tezin Bölümleri

Tezin birinci bölümünde obezite teriminden bahsedilmiştir. Obezitenin teşhisinde en ilgili gösterge BFP değerinin makine öğrenimi yöntemleri ile hesaplanmasında FS yöntemlerinin kullanılmasının gerekliliği ve çalışmanın amaca belirtilmiştir. Literatürdeki benzer çalışmalar incelenmiş kullandıkları tekniklerden, performans metriklerinden bahsedilmiştir.

İkinci bölümde ML algoritmalarına değinilmiştir. ML'nin avantajlarından ve yüksek veriler ile kullanılmasındaki olumsuzluklardan bahsedilmiştir. Bu tezde kullanılan 4 ML algoritmasının genel yapısı hakkında şekiller ve matematiksel ifadeler ile bilgi verilmiştir.

Üçüncü bölümde FS yöntemleri ve genel yapıları anlatılmaktadır. FS yöntemleri kullandıkları prensiplere göre kendi içinde 4 gruba ayrılmaktadır. Bu tezde her gruptan

farklı FS yöntemleri olacak şekilde şekiller ve matematiksel ifadeler kullanılarak 8 adet yöntemden bahsedilmiştir.

Son olarak dördüncü bölümde, ilk üç bölümdeki teorik bilgiler doğrultusunda tezin amacına yönelik olarak gerçekleştirilen uygulamaların sonuçları tablolar halinde çıkış değerleri verilmiştir. Sonuç değerleri karşılaştırılmıştır ve çıktı değerleri grafikler kullanılarak görselleştirilmiştir.

BÖLÜM 2. MAKİNE ÖĞRENMESİ

ML, giriş verilerine karşılık istenilen çıkış değerine en çok yaklaşmayı amaçlayan tekrarlı bir hesaplama sürecidir. Bu teknik girdi verileri ile tekrarlı olarak çıkış verileri üretir. Her adımda ürettiği çıkışın istenilen çıkışa en yakın olması için model yapılarını deneme yanılma yolu ile değiştirir. Buna öğrenme süreci denir ve aslında insanların hatalardan doğruları öğrenme şeklinden ilham alınmıştır. Öğrenme sürecinde verilen verilere en yakın değeri üretmek için modele uygun olmaya çalışırken görmediği verilerde de başarılı sonuçlar verebilmeyi amaçlar. Bu öğrenme sürecinde modelin en iyi sonuçları oluşturması için çeşitli teknik vardır. Bazı algoritmalar giriş değişkenlerinden seçimler yaparken bazıları verilere ağırlık verebilir. Kullanılan ML teknikleri tekrarlı olarak en uygun parametreler ile optimize olabilir. En uygun sonuçlar için giriş verilerinden olasılık dağılımları belirleyebilir. ML teknikleri farklı yapılara sahip olsa da tüm teknikler iki önemli avantaj sağlar. Birincisi, bir problem için tüm ihtimalleri düşünerek tek tek talimatların belirtilmesinden ziyade öğrenme süreci ile daha önce görmediği veriler üzerinde de bir çıktı oluşturabilirler, bu durum insan çabasının azalmasını sağlar. İkincisi ise bir insanın yapabileceğinden daha karmaşık ve detaylı öğrenebilme potansiyeline sahip oldukları için hataları en aza indirirler (Naqa ve Murphy, 2015).

Günümüzde, obezite araştırmacıları ve sağlık uzmanları çok sayıda veriye ulaşabilmektedir. Sensörler, akıllı telefonlar, elektronik medikal kayıtlar, sigorta veri tabanları ve kamuya açık ulusal sağlık verileri bu alanda gelişmiş matematiksel analizlere ihtiyaç duymaktadır. ML, veriyi analiz etme, uyarlama, öğrenme, tahmin etme konusunda güçlü bir algoritma dizisi sunduklarından araştırma alanlarında ML uygulamaları artmaktadır. ML, genel olarak sınıflandırma ve regresyon problemlerinde kullanılır. Doğrusal/lojistik regresyon, karar ağaçları, yapay sinir ağları, derin öğrenme gibi çeşitli teknikler kullanılabilir. Bunlardan bazıları sadece

sınıflandırma için kullanılabilirken bazıları hem sınıflandırma hem regresyon problemlerinde kullanılabilir. Ayrıca ML, veriyi karakterize etmek veya tanımlamak için de kullanılabilir. Örneğin bir veri setinde bulunan birkaç bağımsız değişken birbiri ile ilişkili olabilir. Aralarındaki bu ilişkiyi tespit etmek hedef değişkenin belirlenmesinde önemli bir etken olmaktadır (DeGregory ve ark., 2018).

2.1. Regresyon Modelleri

İstatistikte, girdi verilerinden sürekli bir çıktı değeri elde etme regresyon işlemi olarak tanımlanır. ML'de regresyon işlemi bir denetimli öğrenme tipidir. Denetimli öğrenmede, her girdi değeri için istenen bir çıktı değeri bilinir. Denetimli öğrenme, bu girdi değeri ile karşılık istenen çıktı değerini modele vererek aralarındaki ilişkiyi kuran bir fonksiyon ile öğrenmeyi gerçekleştirir (Alpaydın, 2010).

ML süreci oldukça geniş kapsamlıdır. Temel olarak örüntü tanıma, regresyon ve sınıflama problemleri gösterilebilir. Bu çalışmada BFP hesabı regresyon tahmini olarak değerlendirilmektedir. Araştırmalarda ML algoritmalarından regresyon modeli sık kullanılan bir istatistiksel metottur. Regresyon uygulamaları genellikle iki amaç için kullanılır. Regresyon analizlerinin daha sık kullanılan amaçlarından ilki bağımsız değişkenler ile bağımlı değişkenin tahminini gerçekleştirmektir. İkinci amacı ise bazı durumlarda bağımsız ve bağımlı değişkenler arasındaki ilişkileri belirlemek için regresyon tekniği kullanmaktır. Kullanılacak olan regresyon modelinin probleme uygun olması modelin performansı açısından önemli bir noktadır. Yanlış ya da probleme uygun olmayan bir regresyon modelinin seçilmesi yanlış değerlendirme sonuçlarının çıkmasına neden olabilir (Maulud ve Abdulazeez, 2020).

Regresyon problemlerinde model, gerçek değere en yakın tahmin değerini gerçekleştirmeyi amaçlar. Koşullu dağılım fonksiyonu $F(x|y)$'e göre her rassal giriş değişkeni x'e karşılık gerçek y değeri vardır. ML algoritması giriş değişkenleri dizisi için tahmin değeri oluşturan $f(x,\alpha), \alpha \in \wedge$ fonksiyonunu kullanır. Burada $\wedge$, α özelliğinin elemanı olduğu özellik dizisidir ve verilen giriş değeri x için ML modeli

en iyi yaklaşımı gerçekleştiren $f(x,\alpha)$ fonksiyonu ile eşitlik (2.1)'de verilen $\hat{y}$ tahmin değeri oluşturulur.

$$\hat{y} = f(x,\alpha) \tag{2.1}$$

Regresyon problemlerinde amaç, eşitlik (2.2)'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 hata, eşitlik (2.3)'te verilen kayıp fonksiyonu ile ifade edilir. Risk fonksiyonu minimum olması için kayıp fonksiyonunu minimum yapan eşitlik (2.4)'te verilen $f(x,\alpha_0)$ değeri bulunmalıdır.

$$R(\alpha) = \int L(y, f(x,\alpha)) dF(x,y) \tag{2.2}$$

$$L(y, f(x,\alpha)) = (y - f(x,\alpha))^2 \tag{2.3}$$

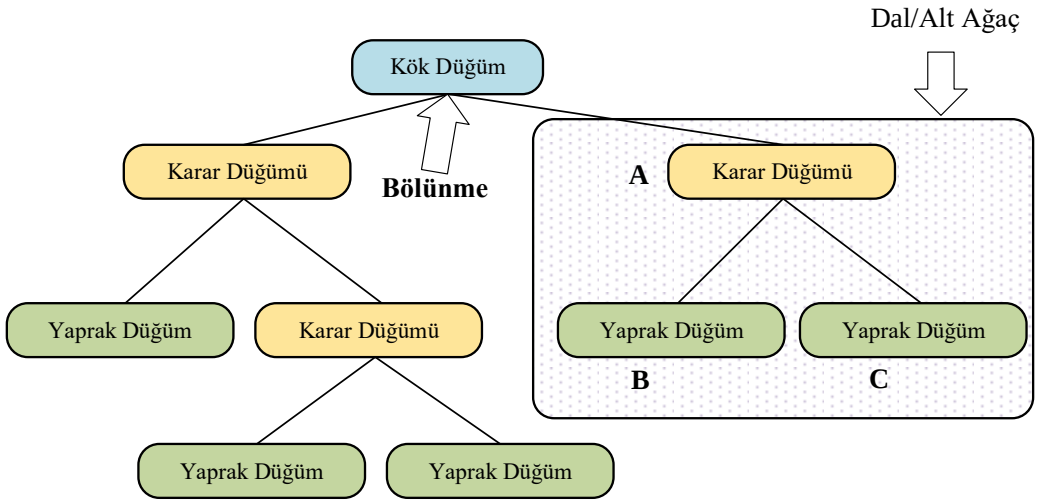
$$f(x,\alpha_0) = \int y \; dF(y \mid x) \tag{2.4}$$

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, gradyan arttırma regresyon (GBR), aşırı gradyan arttırma regresyon (XGBR) yöntemleri kullanılmıştır (Vapnik, 1995).

2.1.1. Rastgele orman ağaçları

RF, bireysel karar ağaçlarından oluşan ağaç topluluğudur. Bu yüzden DT model yapısına benzemektedir. DT modeli Şekil 2.1.'de gösterildiği gibi düğümlerden ve o düğümdeki dallardan oluşan bir yapıya sahiptir. En üstteki düğüm kök düğümdür ve buradan sonra sorular ile bölünerek dallanmalara başlanır (Pekel, 2020). 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 (Fan ve ark., 2006).



Not: B ve C düğümleri A düğümünün alt neslidir.

Şekil 2.1. Karar ağaçlarının genel yapısı

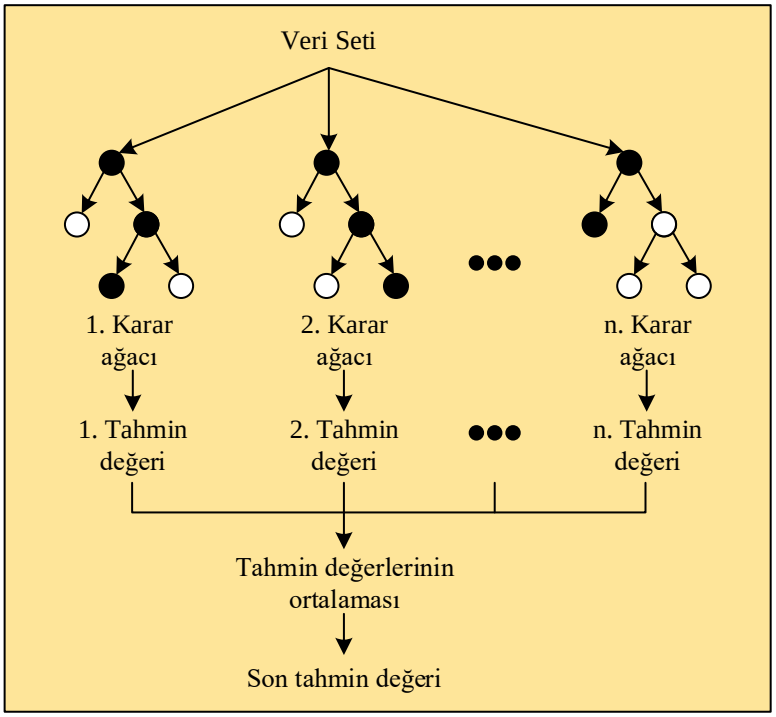
DT, tekrarlı olarak veriyi daha küçük parçalara bölen 'hızlı böl ve yönet' yöntemini benimseyen bir algoritmadır. Karar ağaçları birçok avantaja sahiptir. DT'ler hem sınıflandırma problemleri için kategorik veriler ile hemde regresyon problemleri için sayısal veriler ile çalışma esnekliği sağlar. İstatiksel testler ile modeli doğrulamaya izin verir ve bu da DT'yi güvenilir bir model yapar. Bağımlı ve bağımsız değişkenler boş veri içeriyorsa bunların üstesinden gelebilir. Sınıflandırma problemlerinde çok çıkışlı veriler üzerinde de başarılı şekilde çalışabilir. DT'nin birçok avantajının yanında dezavantajları da mevcuttur. DT'de veride oluşabilecek küçük değişiklikler

ağaç yapısının değişmesine dolayısıyla çıkış değerinin değişmesine neden olacağı için kararsız yapıdadırlar. Kararsız yapıları kötü tahminler yapmalarına neden olabilir. En uygun doğruluk için veri setinde oluşabilecek küçük değişikliklerden kaçınılmalıdır. Modelde aşırı öğrenme oluşmasına karşı parametre ayarları yapılmalıdır.

DT'de tahmin performansını etkileyen birkaç parametre vardır. Bu parametrelerden bazıları; maksimum derinlik, minimum bölme, minimum yapraktaki veri sayısıdır. Maksimum derinlik, bir ağacın kök düğümden yaprak düğüme kadar en fazla büyüyebileceği derinliği belirler. Daha derin ağaçlar veri hakkında daha fazla bilgi elde edilmiş olur. Minimum bölme iç düğümleri bölmek için gereken minimum örnek sayısı anlamına gelmektedir. Örneğin minimum bölme değeri 5 ise düğümde 6 örnek olduğunda bölmeye izin verilir. Minimum yapraktaki veri sayısı, yaprak düğümünde olması gereken minimum örnek sayısı anlamına gelir. Örneğin yapraklardan birinde 3 örnek varsa ve yaprak örnek sayısı 4 ise bu yaprakta bölünme olmayacaktır (Pekel, 2020).

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 içerisinden kayıp fonksiyonunu en minimize eden bölücü seçilir. Ağacın erişebileceği maksimum derinliğe kadar tüm düğümlerin bölünmesi devam eder (Decision Trees, 2022).

RF yapılarının DT'den farkı, ağaç topluluklarından oluşmasıdır. Şekil 2.2.'de verildiği gibi RF yapısında birçok DT ağaçları bulunmaktadır. Bir RF regresyonunun nihai tahmin değeri, her bireysel ağaçların tahmin değerinden elde edilir. Her bireysel ağaç, orijinal veri setindeki gözlem değeri ile aynı olacak şekilde rastgele örnekler yani veri setindeki satırlar seçilir. Bu durumda bir satırdan birden çok bulunabilir. Ayrıca her bireysel ağaç aynı özellikler ile çalışıp benzer sonuçların alınmaması için özelliklerden seçimler yapılır. Her ağaç farklı özellik ve örnek birleşimi içermiş olur. Eşitlik (2.5)'te verildiği gibi RF regresyon tahmini, ilgili girdi değeri için her ağaçtan hesaplanan çıktı değerlerinin ortalaması alınarak nihai tahmin değeri oluşturulur.



Şekil 2.2. Rastgele orman ağaçlarının genel yapısı

$$\bar{h}(x) = (1/K)\sum_{k=1}^{K} h(x;\theta_k) \quad , \quad k = 1,2,...,K \tag{2.5}$$

Burada K ağaç sayısıdır ve $h(x;\theta_k)$, x girişi için ağaçların tahmin değerlerinin elde edildiği tahmin fonksiyonudur. RF tahmin değeri $\bar{h}(x)$, ağaçların tahmin değerlerinin ortalaması ile hesaplanır. θ_k bağımsız ve özdeş dağıtılmış rastgele vektördür (Segal, 2004). 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 (Svetnik ve ark., 2003).

RF, bireysel olarak oluşturulan her ağaç yapısında rastgele örnek verileri seçilir ve rastgele özellikler seçilerek farklı özellik alt kümelerinin oluşturulur. Bu durum her ağacın farklı tahmin değerini oluşturmasını sağlar. Rastgelelik, ağaçlar arasında korelasyonun az olması için yapılır. Böylece tahmin hatalarının azalması sağlanır.

Ağaçların bireysel hatalarını azaltmak için korelasyonun düşük tutulması önemlidir. Bireysel olarak ağaçların boyutu da önemlidir. Ancak daha fazla hafıza alanı kullanma ve düşük hızlarda çalışma gibi olumsuz açıdan etkileri olabilir (Segal, 2004).

RF'de rastgele örnek seçimi ve rastgele seçilen özellik alt kümesi arasından şeçilen en iyi ayırıcı ile bir ağaç maksimum boyuta kadar büyür ve budanmaz. RF, ağaçlar topluluğu olarak bakıldığında hesaplama karmaşıklığının oluşabileceği düşünülebilir. Ancak RF algoritması özellik sayısı çok olduğu durumlarda bile oldukça etkilidir. RF ve DT karşılaştırılınca iki fark gözlemlenebilir. Birincisi DT genellikle ağaç büyütmek için her düğümdeki aday bölünme performansı için tüm özellikleri test ederken RF özelliklerin tümü yerine rastgele belirli bir kısmını seçerek test eder. İkincisi, tek karar ağaçlarında en uygun tahmin için genellikle biraz budama işlemi gerekir. RF'de budama işlemi yapılmaz. Özellik sayısının çok fazla olduğu durumlarda RF tek ağaca göre daha kısa sürede eğitim gerçekleştirdiği ilgili çalışmalarda görülmüştür (Svetnik ve ark., 2003).

Özellik seçimi ağaç büyütme sürecinin temelinde olduğu için karar ağaçlarının performansı genellikle alakasız özelliklerin varlığından etkilenmez. RF alakasız özelliklerin bulunmasına karşı daha dayanıklı olmalıdır. Bu yüzden eğer özellik seçimi ile ilgisiz özellikler çıkarma işlemi uygulanacaksa RF'den daha doğru tahminler yapması beklenmemelidir. Ağaç sayısı RF'de önemli bir parametredir. Ağaçların sayısı parametresi için diğer durdurma kriteri gerektiren algoritmaların aksine, hesap yükünün fazlalığı dışında, çok fazla ağacın bulunmasının olumsuz bir etkisi yoktur. Diğer parametreleri ise DT için kullanılanlar ile aynıdır. (Svetnik ve ark., 2003).

2.1.2. Destek vektör makineleri

Destek vektör makineleri, ilk kez (Vapnik, 1995) tarafından önerilmiş olup istatistiksel öğrenme temeline dayanan bir teoridir. SVM doğrusal/doğrusal olmayan problemler için sınıflama, regresyon ve aykırı değer tespiti yapabilen çok yönlü ve güçlü denetimli bir ML algoritmasıdır. SVM'leri 1990'lı yılların başlarında SVM'lerin az sayıda veri ile öğrenebilmesi, modelin hatalarına karşı dayanıklılığı ve hesaplama verimliliği gibi

yeteneklerinden dolayı güvenilebilir ve başarılı bir şekilde kullanılabilirler (Géron, 2019, Gholami ve Fakhari, 2017).

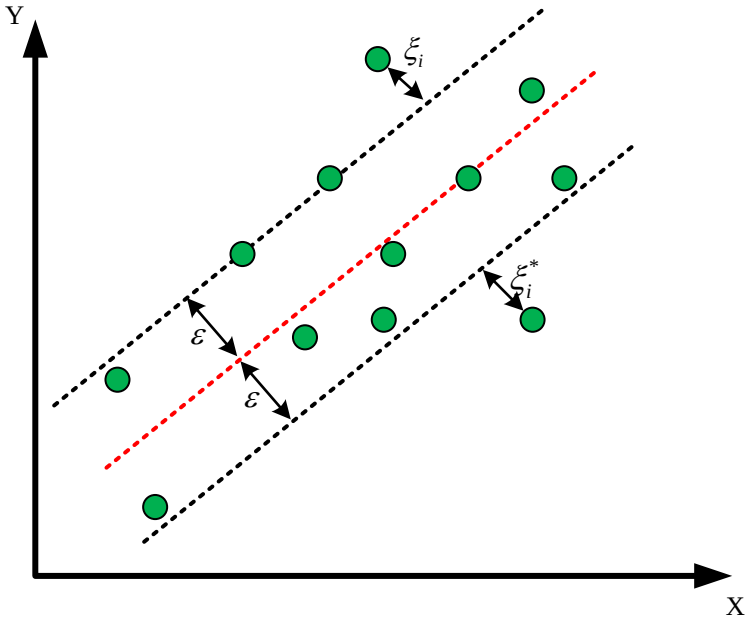
Sınıflandırma problemlerinde, sınıfları ayırmak için birçok düzlem ile her sınıfın noktalarına maksimum uzaklıkta olacak şekilde bir doğrular çizer. Bu doğrular arasındaki mesafeye marj denir (Géron, 2019). Hiperdüzlem, tüpün sınırlarının dışında kalan eğitim verileri olan destek vektörleri cinsinden temsil edilir (Awad ve Khanna, 2015). Marjın daha doğru sınıflandırma yapılabilmesi için mümkün olduğu kadar büyük olması gerekir. Sınıfları birbirinden ayıran çizgilere en yakın noktadaki verilere destek vektörleri denir. Regresyon problemlerin de ise amaç sınıflandırmanın aksine marjın sınırları arasına mümkün olduğu kadar fazla örneği sığdırmaktır. Marjın genişliği parametreler ile kontrol edilir (Géron, 2019).

SVM, maksimum marjı sağlayan en uygun hiperdüzlemi destek vektörleri ile temsil eder. SVM'nin iyi genelleştirme yeteneği ve seyrek çözümü regresyon problemlerine uyarlanmasını sağlar. SVR büyük ve küçük yanlış tahminleri eşit şekilde cezalandıran simetrik kayıp fonksiyonunu kullanarak eğitim sağlar. Vapnik'in ε-duyarsız yaklaşımı, tahmin fonksiyonu etrafında simetrik olarak minimum yarıçaplı esnek ε-tüp ile bir ε-duyarsız bölge oluşturur. Bu tüp, model kompleksliği ile tahmin hatasını dengelerken sürekli değer fonksiyonuna en yaklaşan tüpü bulmak için optimizasyon problemini yeniden formüle eder. Klasik regresyon modelinde, Şekil 2.3.'te verildiği gibi doğrusal veri analizlerinde $\pm\varepsilon$ marjın dışındaki veriler tahmin değeri hatalarını esnek değişkenler ξ_i, ξ_i^* ile tolere edilebilir. Gerçek değer ve tahmin değeri arasındaki hatayı minimize etmek için SVR, ε değerinden daha büyük olan hataları cezalandıran ε-duyarsız kayıp fonksiyonu kullanır. Böylece belirlenen eşiğinin altındaki hataların mutlak değeri tahminin hem üstünde hem de altında yok sayılır. Tüpün dışında kalan noktalar cezalandırılırken tüpün içindekiler ve fonksiyonun hem altında hem üstünde kalan noktalar cezalandırılmaz. ε değeri tüpün genişliğini belirler, küçük değerlerde hataya daha az tolerans gösterir ve destek vektörlerinin sayısını dolayısıyla çözümün seyrekliğini etkiler. Eğer ε değeri azalırsa, tüpün sınırları daralacaktır. Böylece sınırların etrafında daha çok veri noktası bulunması daha çok destek vektörü olmasına

neden olacaktır. Bu durum istenmeyen aşırı öğrenmeye sebebiyet verebilir. Aynı şekilde ε değeri arttıkça tüpün sınırları genişleyecek ve daha az nokta sınırların etrafında olacaktır (Awad ve Khanna, 2015).

Gerçek değer ile tahmin değeri arasındaki tutarsızlığı ölçmek için tanımlanan ε parametresi, en iyi fonksiyonu bulmayı amaçlar. Klasik istatistiksel yaklaşımlar iyi bir tahmin doğruluğu sağlamak amacıyla analizlerine veri değeri ekler. Ancak giriş uzayına eklenecek birkaç aykırı değer modelin esnekliğinin bozulmasına sebep olabilir (Gholami ve Fakhari, 2017).

SVR'nin en temel avantajı hesaplama karmaşıklığının giriş veri uzayının boyutuna bağlı olmamasıdır. Ayrıca mükemmel genelleştirme kapasitesine ve yüksek tahmin doğruluğuna sahiptir (Awad ve Khanna, 2015).



Şekil 2.3. ε -duyarsız yaklaşımı ile doğrusal destek vektör makineleri şematik gösterimi

SVR'de bir doğrusal karar fonksiyonu eşitlik (2.6)'da verildiği gibi ifade edilir.

$$f(x) = \langle w, x \rangle + b \tag{2.6}$$

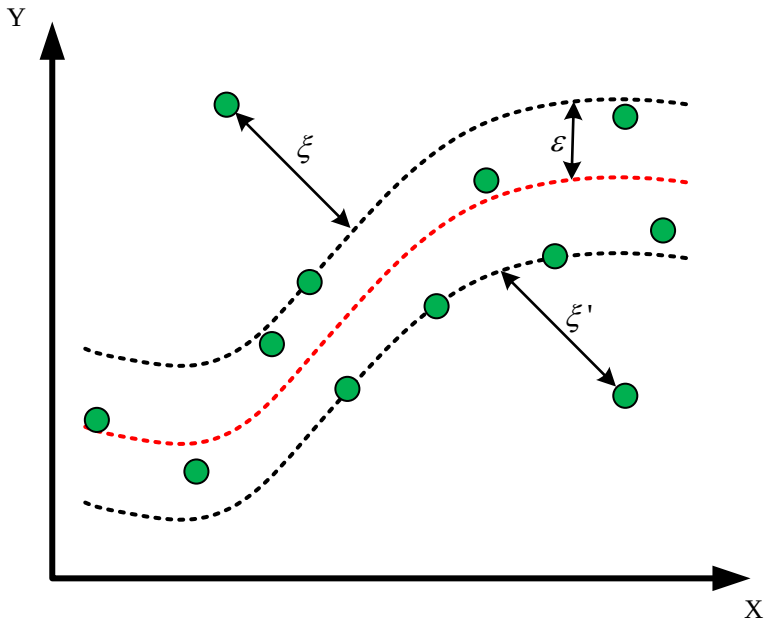
Doğrusal SVR'de, eşitlik (2.7)'de verilen bir kısıt altındaki kayıp fonksiyonunu, ε esnek kenar payı değerlerini kullanarak ve model karmaşıklığını da azaltarak minimize etmeyi amaçlar (Awad ve Khanna, 2015).Burada m veri sayısıdır. SVM'de hatalar, esnek değişkenler ξ_i, ξ_i^* ile tolere edilebilir. C parametresi, optimizasyon problemlerinde model karmaşıklığı ile hataların tolere edilme derecesi arasındaki dengeyi belirleyen düzenleme parametresidir. C parametresi ile modele ceza uygulanarak tahmin fonksiyonun veriye uyması sağlanır (Cherkassky ve Ma, 2004, Ito ve Nakano, 2003, Smola ve Schölkopf, 2004).

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

$$Kısıt: \begin{cases} y_i - \langle w, x_i \rangle + b \le \varepsilon + \xi_i & i = 1, 2, ..., m \\ \langle w, x_i \rangle - b - y_i \le \varepsilon + \xi_i^* & \\ \xi_i, \xi_i^* \ge 0 & \end{cases} \tag{2.7}$$

SVR, doğrusal ve doğrusal olmayan problem için kullanılabilirler (Géron, 2019) ve (Gholami ve Fakhari, 2017). Şekil 2.4.'te verilen doğrusal olmayan SVR analizi, doğrusal durumlar için geçerli olan analize benzerdir. Ancak verilerin yüksek boyutlu Hilbert uzayına (Vapnik, 1995) haritalandırma problemi ile çözülür. Doğrusal olmayan problemlerde ikili çözüm yöntemi önerilir.

Eşitlik (2.8)'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. Eşitlik (2.8)'de verilen L_d fonksiyonundaki w, b, ξ_i, ξ_i^* değerleri için kısmi türev alınarak eşitlik (2.9) elde edilir.



Şekil 2.4. ε -duyarsız yaklaşımı ile doğrusal olmayan destek vektör makineleri şematik gösterimi

$$L_d = \begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\left(\xi_i + \xi_i^*\right) - \sum_{i=1}^{m}\left(\eta_i\xi_i + \eta_i^*\xi_i^*\right) \\ -\sum_{i=1}^{m}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) \end{cases} \tag{2.8}$$

$$\max L_d = -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i - x_j\rangle - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*)$$
$$\text{Kısıt}: \sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*) = 0 \quad \text{ve} \quad \alpha_i, \alpha_i^* \in [0, C] \tag{2.9}$$

Eşitlik (2.9) doğrusal problemler için uygundur. Doğrusal olmayan problemlerde yüksek uzaya haritalama $\Phi: R^n \to H$ ve kernel hilesi $K(x,y) = \langle \Phi(x), \Phi(y)\rangle$ kullanılarak eşitlik (2.10) yazılır. Böylece doğrusal olmayan sistemler için verilen optimizasyon problemi kullanılır (Stoean ve ark., 2006).

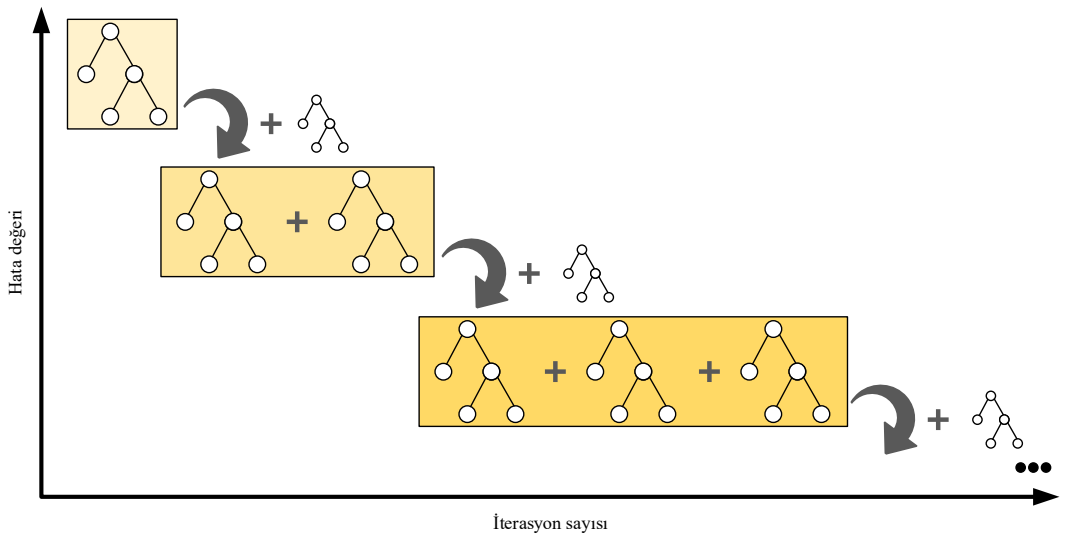
$$\max L_d = -\frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i - x_j) - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*)$$
$$\text{Kısıt}: \sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*) = 0 \quad \text{ve} \quad \alpha_i, \alpha_i^* \in [0, C] \tag{2.10}$$

SVR uygulanırken iyi bir genelleştirme yapabilmesi kullanılan C, ε, kernel gibi parametrelere bağlıdır. Bu parametreler kullanılacak olan domaine göre değişir. Kernel fonksiyon tipi ve bunun dâhil olduğu parametrelerin, C parametresinin ve ε-duyarsız parametresinin uygun değerleri seçilmelidir. Ancak seçilecek bu parametrelerin matematiksel eşitlikler ile ifade edilebilen bir başlangıç değeri olmadığı için en uygun değerlerinin belirlenmesi basit bir işlem değildir. Parametreler için farklı değerlerin kombinasyonu kullanılarak elde edilen çıktıların karşılaştırılıp en başarılı sonuçları veren değerler seçilmelidir (Gholami ve Fakhari, 2017).

C parametresi optimizasyon problemlerinde model karmaşıklığını ve ε değerinden büyük olan sapmaların tolere edilme derecesi arasındaki dengeyi belirler. Seçilecek olan değer çok küçük C değeri (örneğin C=0.01) için gözardı edilebilir bir ceza uygulanmış olur ve SVR fonksiyonun kompleksliği azalır. Çok büyük C değeri (örneğin C=0.01) için ise önemli bir ceza uygulanmış olur ve SVR fonksiyonu veriye uymaya çalışır ve bu durum aşırı öğrenmeye sebep olabilir. ε parametresi, verileri eğitmek için kullanılan ε-duyarsız bölgenin genişliğini belirler. ε değeri destek vektörlerinin sayısını belirleyebilir. Çok dar bir ε-tüp oluşacak şekilde seçilirse (örneğin ε=0.0) veri noktalarını içerisinde alabilecek yeterli marj oluşmayacak ve SVR fonksiyonu veriye uymaya çalışacaktır. Kalın bir ε-tüp oluşacak şekilde ε değeri seçilirse (örneğin ε=0.25) yeterli marja sahip olacağından SVR fonksiyonu düzleşme eğiliminde olacaktır. Büyük ε değeri daha az destek vektörlerinin seçilmesine dolayısıyla modelin daha az karmaşık olmasına etki eder (Cherkassky ve Ma, 2004, Ito ve Nakano, 2003).

2.1.3. Gradyan arttırma regresyon ağacı

Arttırma metotları, sıra ile birden fazla temel modeller geliştirerek tahmin doğruluğunun arttırılmasını amaçlar. Şekil 2.5.'teki gibi her iterasyonda yapısına yeni temel model eklemeleri yaparak 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 (Zhang ve Haghani, 2015).



Şekil 2.5. Hata ve iterasyona bağlı olarak gradyan artırma regresyonunun şematik gösterimi

Arttırma metotları kullanışlı bilgiler elde etmek için eğitim verisini her ardışık modelde yeniden örnekler. Eğitim esnasında her adımda bir önceki modelin hatası dikkate alınır. Bu metotta veri setindeki her verinin eğitim verisi olarak seçilme ihtimali eşit değildir. Kötü tahmin yapan verilere yüksek ağırlıklar verilerek daha çok seçilmeleri sağlanır. Bu yüzden her yeni oluşturulan modelde bir önceki modelde yanlış tahminler yapan veriler vurgulanır. Bu özellik ile GBR'yi eşit olasılıkla seçilen ve rastgele yer değiştiren örneklerden oluşan modeli eğiten RF yönteminden ayırır. Arttırma algoritmaları, önceki modelin katsayısını ve parametrelerini değiştirmeden

yeni bir model ekleyerek kayıp fonksiyonunu minimize etmeye çalışır. Regresyon problemlerinde, arttırma metodu bir fonksiyonel gradyan azaltma şeklindedir. Yani kayıp fonksiyonunu her adımda en iyi şekilde azaltan bir temel model eklenerek optimize edilir. GBR, daha doğru tahminler gerçekleştirebilmek için ardışık olarak kayıp fonksiyonun negatif gradyanı ile yüksek derecede ilişkili olan, genel topluluğa uygun yeni temel öğrenme modelleri ekler (Zhang ve Haghani, 2015, Natekin ve Knoll, 2013).

Arttırma metotlarının geleneksel makine öğrenmesi algoritmalarından farkı, optimizasyon fonksiyon uzayında tutulmaktadır. Yani eşitlik (2.11)'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 (Zhang ve Haghani, 2015, Natekin ve Knoll, 2013).

$$\hat{f}(x) = \hat{f}^{M}(x) = \sum_{i=0}^{M} \hat{f}_i(x) \tag{2.11}$$

Temel öğrenici tahmin fonksiyonu $h(x,\theta)$ ve temel öğreniciden elde edilen tahmin değeri $\hat{\theta}$ değeridir. $t.$ iterasyondaki topluluk tahmin fonksiyonu eşitlik (2.12)'de ve optimizasyon kuralı ise eşitlik (2.13)'te verilmiştir (Natekin ve Knoll, 2013) ve (Friedman, 2001). N veri sayısı olmak üzere fonksiyon tahmini için her iterasyonda optimizasyon kuralını minimum yapan adım sayısı ρ ve θ değerleri seçilmelidir.

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x,\theta_t) \tag{2.12}$$

$$(\rho_t,\theta_t) = \arg\min \sum_{i=1}^{N} L\left(y_i, \hat{f}_{t-1}\right) + \rho h(x_i,\theta) \tag{2.13}$$

RF gibi yaygın kullanılan topluluk tekniklerinde olduğu gibi GBR'de de nihai tahmin ağaç topluluğundaki modellerin tahminlerinin ortalaması olarak ifade edilebilir. Kurulan modele ağaç eklenerek eğitim hatasının daha düşük olması sağlanabilir. Ancak model eğitim verisine çok yakın olursa kötü genelleme yapmasına neden olacaktır. Ağaç sayısının yani iterasyon sayısının artması ile model daha kompleks yapıya dönüşür ve aşırı öğrenme problemi ortaya çıkar. Bunun için en uygun ağaç sayısı belirlenmelidir. Aşırı öğrenme problemi gradyan arttırma iterasyon sayısı ile kontrol edilebilir. İterasyon sayısı ve öğrenme oranı arasında da bir denge vardır. Aynı iterasyon sayısına sahip durumlarda öğrenme oranı azaldıkça eğitim riski artma eğiliminde olacaktır. Performansı etkileyen diğer parametre ise ağaçların derinliğidir. Ağaç derinliği bir ağaçtaki düğümlerin sayısı olarak ifade edilir. Çok derin ağaçlar kötü model performansına ve hesaplama karmaşıklığının fazla olmasına neden olabilr. Bu yüzden giriş değişkenlerinin etkileşimine bağlı olan bu değer tüm ağaçlar için en uygun değerde sınırlandırılmalıdır.

GBR, kompleks doğrusal olmayan problemlerinde de etkili şekilde kullanılabilir. Farklı problemlere uyum sağlayabilecek esnekliğe sahip olması gibi avantajlarının yanı sıra dezavantajları da mevcuttur. Dezavantajlarından biri hafıza kullanımıdır. Öğrenme süreci için bir tahmin modelinin depolanması arttırma iterasyon sayısına bağlıdır. Kullanılacak uygulamaya bağlı olarak bazı durumlarda binler mertebesinde iterasyon gerekebilir. Bu durumda her adımdaki parametrelerin depolanması gereklidir. Bellek tüketimi ile ilgili bu tür sorunlar seyrek temel öğreniciler ile veya topluluk basitleştirme yöntemi ile aşılabilir. Ancak genel topluluk algoritmalarının bellek tüketimi ile ilgili bu tür sorunlar ortaktır (Natekin ve Knoll, 2013, Zhang ve Haghani, 2015).

2.1.4. Aşırı gradyan arttırma regresyonu

XGBR, ölçeklenebilir ağaç arttırma makine öğrenmesi algoritmasıdır (Chen ve Guestrin, 2016). Ölçeklenebilir olmasını önemli sistem ve algoritmik optimizasyon sağlar. XGBR, boş veriler ile çalışabilmeye duyarlıdır ve örnek ağırlıklarını ele almayı sağlayan ağırlıklı nicel çizim ile ağaç öğrenme algoritmasına yenilikler katmıştır.

Paralel ve dağıtılmış hesaplamalar ile modelin çalışmasını hızlandırır. Çekirdek dışı hesaplamalar kullanan XGBR, büyük verileri işlerken etkili önbellek optimizasyonu ile hesaplamaların hızlı gerçekleştirilmesini sağlar.

GBR metoduna benzer yapıda çalışmaktadır. XGBR algoritmasının farkı, modelin karmaşıklığını cezalandıran bir düzenleme parametresinin eklenmesidir. Eşitlik (2.14)'te XGBR yönteminin kullandığı amaç fonksiyonu verilmiştir. GBR'den farklı olarak düzenleme parametresi Ω vardır. Eğer Ω sıfıra ayarlanırsa model GBR metodu gibi çalışır. Düzenleme parametresi Ω, birçok değişken için ağırlıkları sıfıra çekmeye çalışır ve böylece yüksek boyutlu problemlerde önemli bir rol oynayan özellik seçimini gerçekleştirir (Carmona ve ark., 2019).

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad \text{burada } \Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{2.14}$$

Burada ω, yapraklardaki skor vektörü, λ düzenleme parametresi ve γ yaprak düğümü bölmek için gereken minimum kayıptır. Burada l, tahmin değeri ile gerçek değer arasındaki farkı ölçen değişebilir konveks kayıp fonksiyonudur. Düzenleme parametresi Ω sayesinde aşırı öğrenmeden kaçınmak için son öğretilen ağırlıklar düzeltilir. Düzenlenmiş amaç fonksiyonu tahmin yeteneği daha yüksek modelleri seçme eğiliminde olacaktır.

XGBR, ağaç karmaşıklığı, öğrenme oranı, düzenleme parametresi gibi birçok 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 (Sagi ve Rokach, 2021).

XGBR, eşitlik (2.15)'te verilen genel eşitlikte daha hızlı optimizasyon işlemi gerçekleştirmek için amaç fonksiyonun türevi alınır. Formülü daha da basitleştirmek için sabit terim değerini kaldırırsak;

$$\tilde{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right)$$
$$\text{burada } g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \text{ ve } h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \tag{2.15}$$

j yaprağının örnek dizisi $I_j = \left\{ i \,|\, q\left(x_i\right) = j \right\}$ ise kayıp fonksiyonu eşitlik (2.16) verildiği gibi yeniden yazılır.

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{n}\left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\sum_{i\in I_j}\left(g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T\end{aligned} \tag{2.16}$$

j yaprağının en uygun ağırlığı eşitlik (2.17)'den hesaplanan amaç fonksiyonu için en uygun değer eşitlik (2.18)'den elde edilir.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{2.17}$$

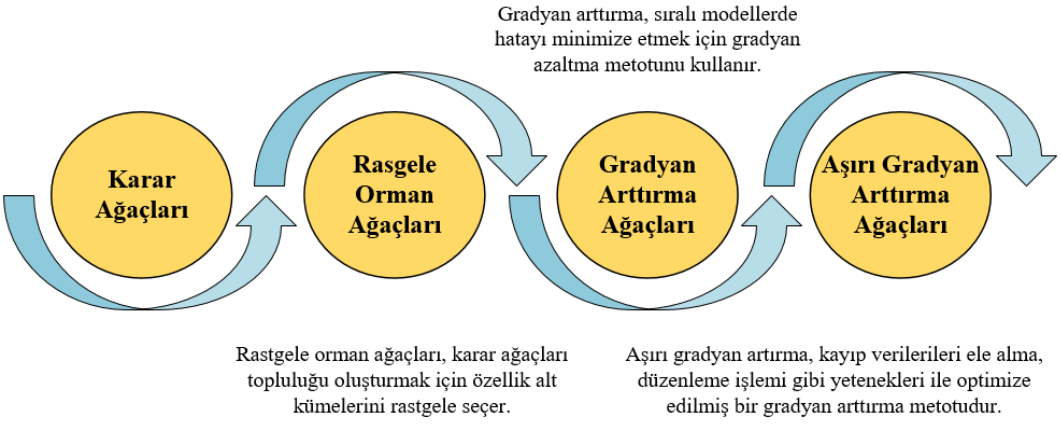
$$\tilde{L}^{(t)}\left(q\right) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i \right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{2.18}$$

Eşitlik (2.18), ağaç yapısı q için bir tür puanlama fonksiyonu olarak kullanılabilir. I_L ve I_R bir düğümün sağ ve sol örnekleri olarak ifade edilirse bölünme işleminden sonra genellikle bölme adaylarını değerlendirmek için kullanılabilecek kayıp fonksiyonu eşitlik (2.19) ile ifade edilir.

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{2.19}$$

Yüksek boyutlu veri setlerinde aday bölme noktalarını bulmak için her örneğe eşit ağırlık vermek başarısız alt kümelerin seçimine sebep olabilir. XGBR, bu problemi çözmek için dağıtılmış ağırlıklı nicel çizim sunmaktadır.

Şekil 2.6.'da karar ağaçlarına dayalı bazı çözümlerin birbirine evrilmesi gösterilmiştir. Bir problem için en uygun çözümün elde edilmesini sağlayan karar ağacına dayalı metotlardaki dezavantajları düzenlemeye çalışmak amacıyla diğer metotlar geliştirilmişlerdir.



Şekil 2.6. Karar ağaçlarına dayalı çözümler

BÖLÜM 3. ÖZELLİK SEÇİMİ

Yüksek boyutlu veri setlerinin makine öğrenmesi modelleri tarafından kullanılması büyük problemdir. Özellik seçimi, yüksek boyutlu giriş setleri ile başa çıkabilmek için modelin öğrenme aşamasından önce gerçekleşen önemli bir veri ön işlem basamağıdır (Bolón-Canedo ve ark., 2013). Uygun verilerin seçilmesi öğrenme modelinin performansına, modelin genelleştirebilme yeteneğine katkı sağlar. Ayrıca daha az veri ile çalışmak hem işlem yükünü azaltır hem de problemin makine öğrenmesi modelleri tarafından daha iyi anlaşılmasını sağlar.

ML tekniklerinde eğitim esnasında genellikle hata fonksiyonunu azaltmak için veri sayısının arttırılması yapılabilir. Ancak bu durum aşırı öğrenme problemine yol açma riski taşımaktadır. Hatanın azalması için çok sayıda veri ile eğitim yapmak, çok karmaşık model seçmek aşırı öğrenmeye sebep olur. Sonuç olarak eğitim aşamasında çok verimli sonuçlar elde edilirken tahmin aşamasında çok kötü tahmin sonuçları oluşur. Aşırı öğrenme problemini çözmek için en basit yaklaşımlardan biri modelin karmaşıklığını azaltmaktır. Bu işlem özellik seçim yöntemi ile özellik uzayının boyutunu azaltarak sağlanabilir. (Awad ve Khanna, 2015)

Giriş boyutunu azaltma amacıyla özellik seçim ve boyut azaltma olmak üzere iki temel teknik mevcuttur. Bunlar benzer uygulamalar olsa da aralarında fark vardır. Özellik seçimi, sadece ilgili özellikleri tanımlar, ancak değişkenleri orijinal formlarında tutar. Boyut azaltma yöntemi ise değişkenleri birleştirerek verileri dönüştürür (Otchere ve ark., 2022).

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

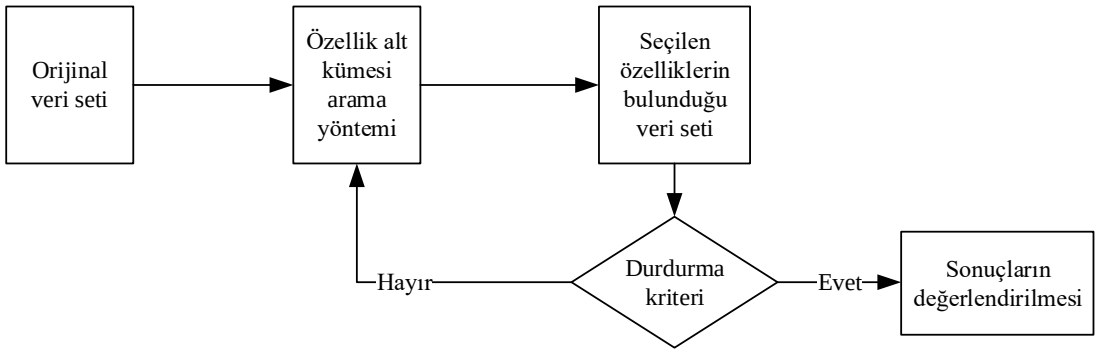
ağırlıklara göre özellikler önem sırasına göre sıralanır. Alt küme değerlendirmesi ise araştırma stratejilerine göre aday özellik alt kümeleri oluşturulur ve değerlendirilir. Genellikle FS sürecinde bir alt küme oluşturulup bu küme değerlendirildikten sonra durdurma kriteri sağlanıyorsa model performansın değerlendirilmesi yapılır (Kumar ve Minz, 2014).

Son yıllarda yüksek boyutlu verilerin varlığı makine öğrenmesi algoritmalarının kullanımını zorlaştırmaktadır. ML algoritmalarını daha etkili şekilde kullanabilmek için veri ön işleme esastır. Alakasız, gereksiz, gürültülü verilerin kaldırılmasını içeren özellik seçimi, önemli bir veri ön işleme tekniğidir (Kumar ve Minz, 2014). Değerlendirme metriğini optimize eden, değerlendirme metriğinin belirli bir kısıtını karşılayan daha küçük boyutlu, boyut ve değerlendirme metriği arasındaki en iyi dengeyi sağlayan özellik alt kümeleri genellikle bu yaklaşımlar dikkate alınarak seçilir.

Hedef değişkenin varyansı, bağımsız değişkenlerin sayısı ile artar veya parametrelerin en küçük kareler kullanılarak uygun hale getirilmesi ile doğrusal model için artar. Bu şekilde en iyi tahminler modelde hiç değişken olmadığında elde ediliyormuş gibi görülebilir. Örneğin bağımsız değişkenlerinin değerine bakılmaksızın, her zaman bağımlı değişken için aynı sabit değer tahmin edilirse tahminler sıfır varyansa sahip olacaktır ama muhtemelen büyük yanlılık içerecektir. Bir modele değişken eklendikçe artan varyansa karşılık yanlılığın azalması dengeyi sağlar (Miller, 2002). Bir değişken tahmin değerine sahip değilse değişken eklemek sadece varyansı arttıracaktır. Eğer bir değişkenin eklenmesi yanlılıkta az bir fark oluşturuyorsa tahmin varyansındaki artış yanlılığı düşürmekten daha faydalı olabilir. Model verilerden bağımsız seçilmediğinde önemli yanlılık problemleri oluşur. Giriş değişkenleri vektörüne ekstra değişken eklemek yanlılığı azaltmayacaktır. Ayrıca tahmin için en iyi özellik alt kümesi, tahmin yapmak istediğimiz giriş vektörlerinin aralığının bir fonksiyonudur. Eğer gözlem örneklerinin sayısı arttırılabilirse varyans azalacaktır. En iyi tahmini gerçekleştiren özellik alt kümesindeki değişken sayısı, modeli kalibre etmek için kullanılan örneklerin boyutuyla birlikte artacaktır. Tahmin denklemlerinde ne kadar değişkenin olacağı yanlılığın anlaşılması ile ilgilidir. Bir problemdeki örnek sayısı özelliklerin

sayısından çok az olursa, araştırma uzayı seyrek popülasyon olacağı için ML algoritmaları öğrenmede zorlanacaktır. Model, gürültülü veri ile ilgili, alakalı veri arasındaki farkı ayırt edemeyecektir.

Özellik seçim yöntemlerinin temel adımları Şekil 3.1.'de verilmiştir. Genellikle FS süreci alt küme oluşturma, alt kümeyi değerlendirme, durdurma kriteri sağlayıp sağlamadığına bakma ve sonucu doğrulama sırasıyla gerçekleşir. Alt küme oluşturmada 2 temel nokta vardır. Biri, her durumda başlangıç noktasına karar vermektir. Aramanın başlayacağı noktaya karar vermek için ileri, geri, bileşik, ağırlıklı ve rastgele metotlar düşünülebilir. Diğeri ise sıralı arama, üstel arama veya rastgele arama gibi özel bir strateji ile özellik seçimi ile ilişkili arama organizasyonudur (Kumar ve Minz, 2014). En yeni üretilen aday alt küme var olan birçok değerlendirme kriterine göre değerlendirilir.



Şekil 3.1. Özellik seçim işlemleri için temel adımlar

Özellik seçimi ilişkili özelliklerin veya aday özellik alt kümelerinin seçilme sürecidir. Değerlendirme kriteri, en uygun özellik alt kümesini oluşturmak için kullanılır. Özellik sayısı N olan bir veri seti için 2^N aday özellik alt kümesi vardır.

Orijinal özellik seti A ve L optime (maksimize) edilmeyi değerlendirme kriteridir ve $L: A' \subseteq A \to R$ olarak tanımlanır. Aday özelliklerin alt kümesi aşağıdaki hususlara göre değerlendirilebilir:

a. $|A| = m \,\&\, |A'| = n$, $L(A')$ maksimize edilir, burada $m > n$ ve $A' \subset A$.

b. $m > n$ durumunda en az sayıda özellik bulunan alt kümeyi bulmak için $L(A') > 0$ olacak şekilde bir θ eşik değeri belirlenir.

c. En uygun özellik alt kümesi için $|A'|$ ve optimizasyon fonksiyonu $L(A')$ bulunur.

Teorik olarak alaka düzeyini korumak için $a_k \in A$ olan her bir özellik a_k için ağırlık w_k değeri atanır. En iyi özellik alt kümesi L optimizasyon kriterinin maksimum olduğu durumdur.

Alt küme oluşturma belirli arama stratejileri kullanarak yapılan bir arama sürecidir. Belirlenen değerlendirme kriterine göre üretilen özellik alt kümesi bir önceki en iyi alt küme ile değerlendirilir. Eğer yeni alt küme öncekinden daha iyi ise güncelleme yapılır ve oluşturulan alt küme artık yeni en iyi alt kümedir. Bu işlem durdurma kriteri karşılana dek devam eder. Durdurma kriterine varıldıktan sonra oluşturulmuş olan en uygun alt kümenin doğrulanması gerekir. Doğrulama işlemi yapay ya da gerçek veri setleri kullanılarak gerçekleştirilebilir.

Özellikler genellikle alakalı, alakasız ve gereksiz olmak üzere 3 şekilde nitelendirilebilir (Karagiannopoulos ve ark., 2007). Alakalı özellikler, çıkış değerini etkileyen ve diğer özelliklerin onun rolünü üstlenemediği özelliklerdir. Alakasız özellikler, çıkış değeri üzerinde bir etkisi olmayan ve her bir örnek için rastgele değer üretilen özelliklerdir. Gereksiz özellikler ise, bir başka özellik tarafından rolleri üstlenilebilen özelliklerdir.

Özellik seçim algoritmalarında bir durdurma kriteri olmadığı durumlarda özellik seçim süreci ya tüm ihtimalleri değerlendiren kapsamlı olarak adlandırılan şekilde ya da sonsuza kadar çalışır. Durdurma kriteri, daha iyi bir alt kümenin oluşturulamayana kadar özelliklerin eklenmesi veya çıkarılması ya da bazı değerlendirme kriterlerine göre en uygun özellik alt kümesi oluşması olabilir.

FS yönteminin amacı orijinal özellikleri içeren veri setinden önemli birkaç özelliği seçmektir (Frénay ve ark., 2013). Özellik seçimi, boyut probleminin etkisini azaltarak modelin tahmin performansını geliştirebilir ve veri setinin boyutunun ciddi şekilde azalmasını sağlar. Öğrenme süreci hızlanır ve problemin daha iyi anlaşılmasını sağlar.

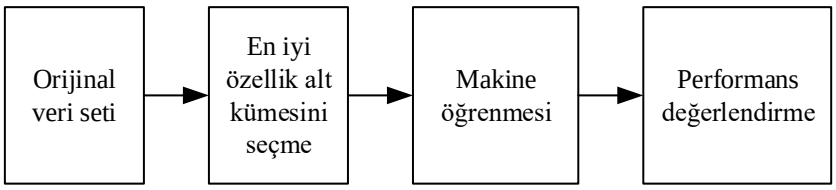
Genel olarak, iyi bir modelin nispeten az tahmin edicisi vardır. Bu nedenle yorumlanması kolaydır ve bu tür bir modele tutumlu denir. Ancak önemli özellikler atlanırsa, regresyon katsayısı tahminleri (ve dolayısıyla uygun değerler) yanlı olabilir (Eberly, 2007).

Özellik seçimi için 4 tip yaklaşım mevcuttur. Birincisi, Filtre FS yöntemi, eğitim verisinin genel karakteristiğini kullanır. İkincisi, Sarmal FS yöntemi, en uygun özelik 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 içerisinden en uygun özellik alt kümesini bulmaya çalışır. Hibrit yöntemler, filtre yöntemlerin yüksek etkililiğinden, sarmal yöntemlerin ise yüksek doğruluğundan faydalanır (Kumar ve Minz, 2014, Jović ve ark., 2015). Bu veri ön işlem sürecinde aynı eğitim verisi kullanılsa bile farklı FS algoritmaları ile farklı özellik alt kümeleri oluşturulup farklı performanslar görülebilir.

3.1. 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 aşamalı strateji uygulanır. 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 3.2.'de filtre özellik seçim yönteminin genel yapısı gösterilmiştir (Miao ve Niu, 2016). Filtre FS metot, bir öğrenme algoritması

kullanılmadan özelliklerin alt kümesini değerlendiren bağımsız bir ölçü içerir. Bu yüzden algortitma hesaplama sürecinde etkili ve hızlı olma gibi avantajının yanı sıra tek başına kullanışlı olmayan ancak diğer özellikler ile birleştiğinde kullanışlı olabilecek özellikleri seçmediğinde kaçırabilir.



Şekil 3.2. Filtre FS modeli genel yapısı

Filtre metot, önemli olan özelliklerin belirlenmesi için öncelikle gözetimli analiz oluşturur ve modele uygular. Bir özelliğin veri setinden çıkarılacağına veya veri setinde kalacağına istatistiksel öneme dayalı bir analiz ile karar verir. Analiz sadece bir kez uygulanır.

Filtre metot 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 (Otchere ve ark., 2022).

Filtre metot için farklı değerlendirme kriterleri mevcuttur. Karşılıklı bilgi ve onun varyansları, korelasyon faktörü, gürültü varyansı, determinasyon katsayıları vs. gibi kriterler örnek verilebilir. Bir özellik tek başına olduğunda hedef için herhangi bir bilgi içermiyor olabilir ancak başka bir özellik ile birleşimi kullanışlı bilgi içerebileceğinden çok değişkenli ilişki tespiti yapmak önemlidir. Çoğu regresyon problemi veri setleri değişkenler arasında doğrusal olmayan bir ilişki vardır. Bu nedenle, regresyon problemlerinde FS için kullanılan bir uygunluk kriteri doğrusal olmayan ilişkileri de tespit edebilmelidir. Tahmin edici örneklerin dayanıklılığı, ilk olarak eğer erişilebilen veri sınırlı ise tahmincinin mümkün olduğu kadar yansız ve düşük varyansa sahip olması anlamına gelmektedir. Veri setindeki küçük değişiklikleri FS'de değişikliklere neden olabilir. Buna tutarlılık denilmektedir (Degeest ve ark.,

2021). Gerçek veri setleri farklı tipte ve etkide gürültüler içerebilir. Tahmin fonksiyonu bu gürültüyü belirli bir noktaya kadar tolere edebilmelidir.

Bu çalışmada karşılıklı bilgi ve tek değişkenli istatistiksel test kullanarak filtre özellik seçim yöntemi gerçekleştirilmiştir.

3.1.1. Karşılıklı bilgi ile özellik seçimi

Karşılıklı bilgi (MI), iki rastgele değişken arasındaki bağlılığı 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 bağlılıklar ile sınırlıdır (Frénay ve ark., 2013).

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 MI değerini tahmin etmektir (Kraskov ve ark., 2004).

MI, varolan iki rastgele değişken arasındaki bağlılığı ölçen bir niceliktir. X ve Y rastgele değişkenlerin olasılık yoğunluk fonksiyonu f_X ve f_Y ve domainleri x ve y ’dir. Ortak olasılık yoğunluk fonksiyonu $f_{X,Y}$ olduğuna göre X ve Y arasındaki MI değeri eşitlik (3.1)’deki gibi ifade edilebilir.

$$I(X;Y) = -\int_x \int_y f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)\, f_Y(y)} dxdy \tag{3.1}$$

Entropi ve şartlı entropi terimlerinin ifadesi ise sırasıyla eşitlik (3.2) ve (3.3)’de verilmiştir. Bu eşitliklerden MI değeri eşitlik (3.4)‘teki gibi elde edilir.

$$H(X) = -\int_x f_X(x) \log f_X(x) dx \tag{3.2}$$

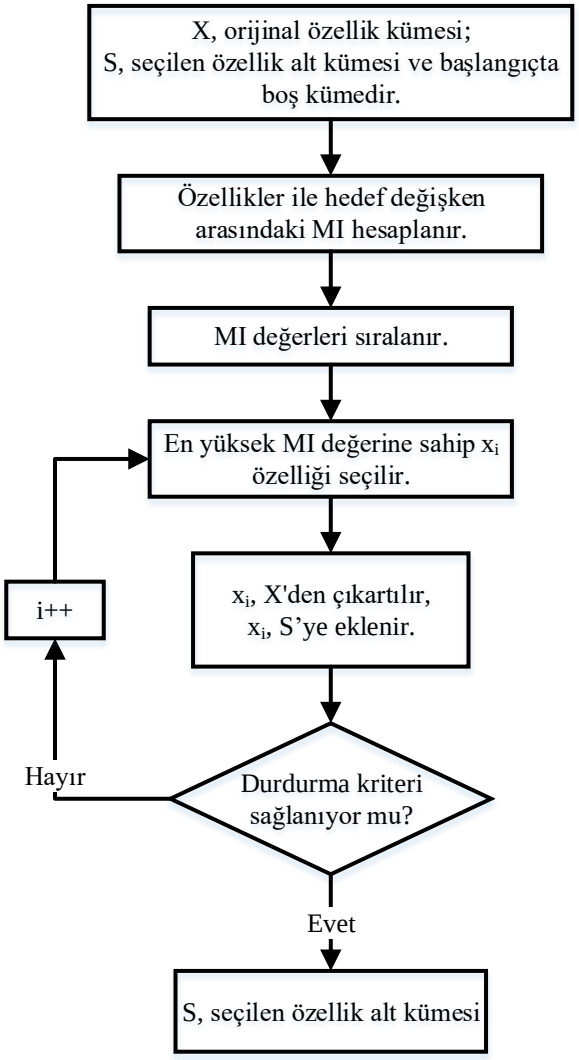
$$H(Y \mid X) = -\int_x \int_y f_{X,Y}(x,y) \log \frac{f_X(x)}{f_{X,Y}(x,y)} dxdy \tag{3.3}$$

$$I(X;Y) = H(Y) - H(Y \mid X) \tag{3.4}$$

Giriş X ve çıkış Y arasındaki regresyon problemleri için Y 'nin entropisi $H(Y)$, sabittir ve özelliklerin seçimine bağlı değildir. Bu yüzden, $I(X;Y)$ maksimize eden özelliklerin seçilmesi X için Y 'nin şartlı entropisi $H(Y \mid X)$ minimize eden özelliklerin seçilmesi ile elde edilir.

MI, simetriktir $I(X;Y) = I(Y;X)$ ve pozitif değer alır. Ortak olasılık yoğunluğu $f_{X,Y}(x,y)$ ve marjinal yoğunlukların $f_X(x) f_Y(y)$ sonucu arasındaki bir uzaklık ölçümü olarak yorumlanabilir. Böylelikle eğer X ve Y istatistiksel olarak birbirinden bağımsız ise $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ MI değeri sıfırdır. Bunun aksine eğer X ve Y daha bağımlı ise MI değeri daha yüksek çıkar.

Şekil 3.3.'de MI kullanılarak gerçekleştirilen filtre özellik seçimine ait akış diyagramı verilmiştir. Her bağımsız değişken için bağımlı değişken ile arasındaki MI hesaplanır ve MI değerleri sıralanr. En yüksek MI değerine sahip özellikten başlanarak durdurma kriterini sağlanana kadar seçilen özellikler ile yeni özellik alt kümesi oluşturulur.



Şekil 3.3. Karşılıklı bilgi ile filte özellik seçim yöntemi akış diyagramı

3.1.2. Tek değişkenli istatistiksel test ile özellik seçimi

Tek değişkenli istatistiksel test (UST) en iyi özellik alt kümelerini bulmaya çalışan bir istatistiksel testtir. Özelliklerin hedef değişken ile bir ilişkisinin olup olmadığı test edilir. Her özellik için bir test skoru oluşturulur. Yüksek skora sahip özellikler özellik alt kümesine seçilir.

Tek değişkenli test denilmesinin sebebi, bu teknik toplu olarak tüm özellikler ile ilgilenmez. Özelliklerin birbiri arasındaki etkileşimi araştırmak yerine her özelliği ayrı

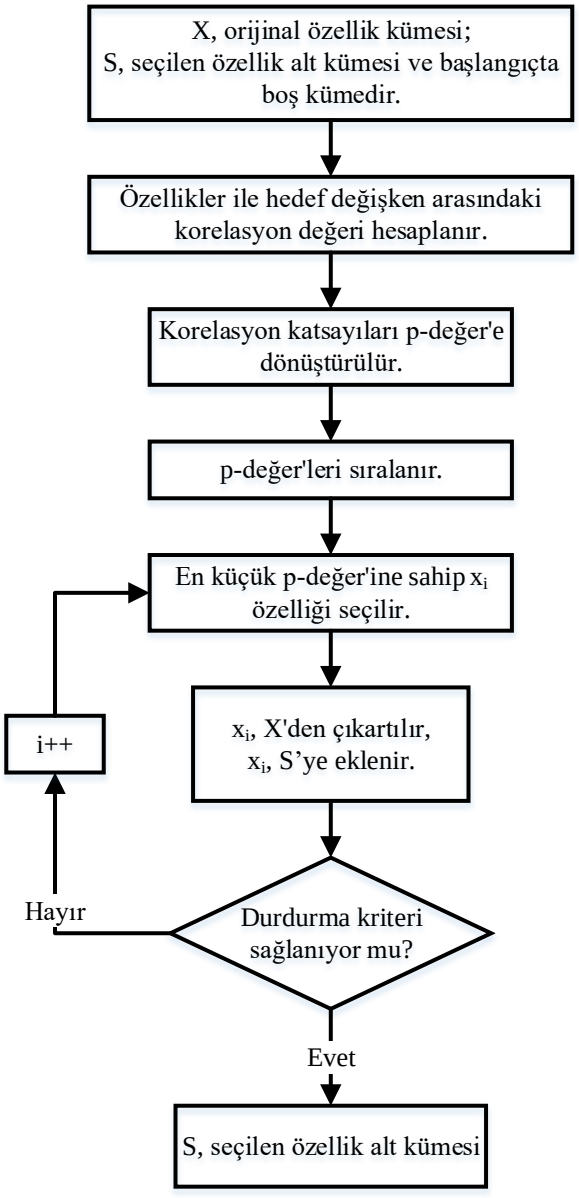
olarak değerlendirir ve özellik ile hedef arasında önemli ilişki olup olmadığını ele alır (Jović ve ark., 2015).

Doğrusal regresyon F-test, regresyon modelinin veriye hiç özellik bulunmayan modelden daha iyi uyup uymadığını test etmek için kullanılır. Yani tek bağımsız değişken bulunan veri setinde, özellik ve hedef arasında önemli ilişki olup olmadığını test eden basit doğrusal regresyona eşittir. Bu teknikte, veri seti birçok basit doğrusal regresyon modellerine bölünür ve her modelden elde edilen F-skoru, aslında özelliğin F-skor değerini döndürür. Bu test sadece her özelliğin hedef ile arasında doğrusal ilişki olup olmadığını doğrular. Bu test kullanılmak istenildiğinde eğer doğrusal olmayan bir ilişki var ise özelliklere özel dönüşüm metotları uygulanmalıdır (Jović ve ark., 2015).

Tek değişkenli FS olan UST en iyi özellikleri seçerek çalışır. Bu metot, iki değişken arasındaki doğrusal ilişkiyi ölçer. Hedef değişken ve özellik değişkenin gauss dağılım izlediğini varsayar. Eğer böyle değilse bu test sonucu kullanışlı olmayacaktır.

İki değişken arasındaki karşılıklı korelasyon değeri Pearson's R korelasyonuna eşitlik (3.5)'e göre hesaplanır. Bu eşitlikten elde edilen değer önce F-skor değerine sonra p-değer'ine dönüştürülür. Burada p_i korelasyon katsayılarını ifade eder, F_i ile F-skor değeri hesapır ve n hedef değişkenin uzunluğudur. Her özellik ve hedef değişken arasındaki ilişki p-değer'e göre sıralanır. Şekil 3.4.'te bu metota ait akış diyagramı verilmiştir. Bir özelliğin p-değer'i ne kadar küçük ise o özellik o kadar önemlidir (Emura ve ark., 2019). Buna bağlı olarak sıralanan p-değer'lerden en küçük değerlerden başlanarak özellikler seçilir.

$$\rho_i = \frac{\left(X[:,i] - mean\left(X[:,i]\right)\right) * \left(y - mean\left(y\right)\right)}{std\left(X[:,i]\right) * std\left(y\right)}$$
$$F_i = \frac{\rho_i^2}{1-\rho_i^2} * \left(n-2\right) \quad (3.5)$$



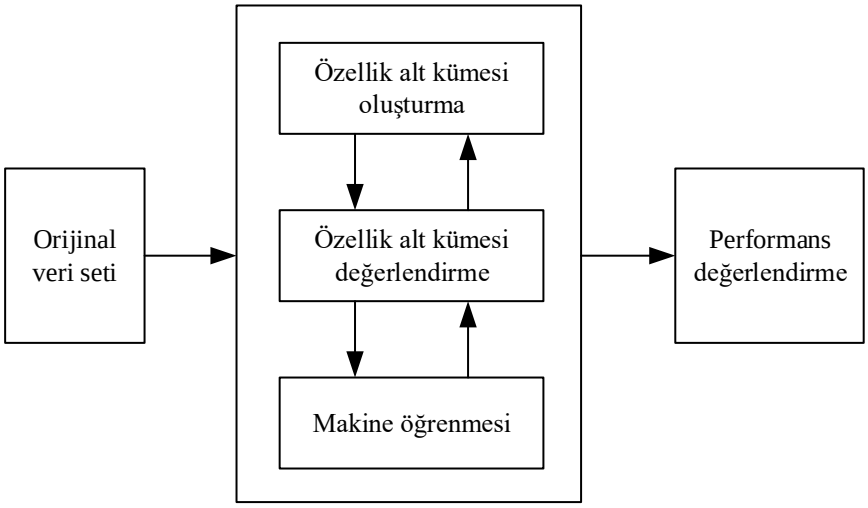
Şekil 3.4. Tek değişkenli istatistiksel test ile özellik seçim yöntemi akış diyagramı

UST metotu, özellikleri istatistiksel puanlama fonksiyonlarına bağlı olarak bir özellik sıralaması oluşturur. Özellikler öğrenme modeline verilmeden önce yapılan bir ön

işlemdir. Tek değişkenli bu metot, önemli olan özellikleri belirler (Subho ve ark., 2019).

3.2. 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 yönteme göre daha iyi sonuçlar gösterir (Kumar ve Minz, 2014). Şekil 3.5.'te sarmal özellik seçim metodunun genel yapısı verilmiştir.



Şekil 3.5. Sarmal FS modeli genel yapısı

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 (Raschka, 2022).

Sarmal özellik seçim yöntemi, ML algoritması ile farklı özellik alt kümelerini değerlendirmeye çalışır ve en iyi performans gösteren alt kümeyi seçer. En temel metot

sıralı ileriye doğru seçme (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. Geriye doğru seçme (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 (Karagiannopoulos ve ark., 2007).

Sarmal metodun başlıca iki problemi vardır. Birincisi, farkı özellik alt kümeleri ile birçok tahmin modeli oluşturma zorunda olduğu için hesap yükünün fazla olmasıdır. Diğeri ise, belirli regresyon problemleri için sınırlı kullanımları olduğundan genelleme yetenekleri zayıftır. Her iki probleminde üstesinden gelmek için filtre metotu kullanılabilir. Filtre FS metotları sarmal FS metotlarından daha hızlıdır ve herhangi bir ML algoritması ile kullanılabilir (Frénay ve ark., 2013).

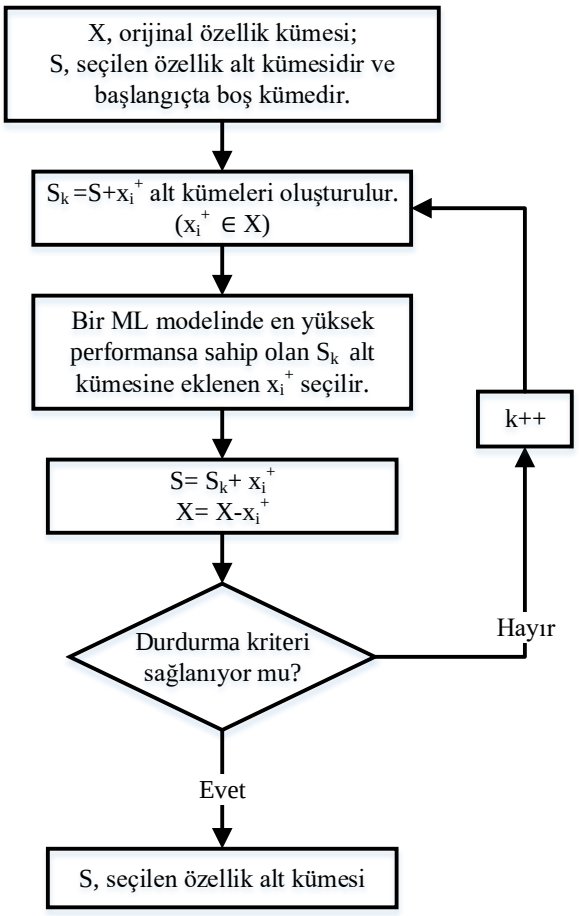
SFS, özellikleri tek tek eklendiğinden birbiri ile bağımlı olan özellikleri seçmede başarısız olabilir. Ancak küçük boyutlu verilerde küçük etkili alt kümeleri hızlı bulabilir. SBS'de de özelliklerin arasındaki bağlılık dikkate alındığından değerlendirmeler daha maliyetlidir.

Sarmal yöntemler, seçilen özellikleri değerlendirmek için yapılarına dâhil olan ML algoritmaları bulunur. Bu yüzden yapıları gereği aynı eğitim verisi ile kullanılsalar bile farklı öğrenme algoritmaları ile değerlendirildiklerinde en iyi performans gösterdikleri özellik alt kümeleri değişebilir. Sarmal FS metotu regresyon modelinin performansını geliştirebilir ancak daha fazla hesap karmaşıklığı içerebilir. Filtre metotlar hesap karmaşıklığı açısından daha uygundur ama özelliklerin alaka düzeyini belirlemede daha başarısızlardır. Ayrıca filtre metot ile seçilen özellikler öğrenme algoritmasının korelasyon katsayısını azaltabilir (Karagiannopoulos ve ark., 2007).

Bu çalışmada sarmal tip özellik seçiminden sıralı ileriye doğru ve sıralı geriye doğru özellik seçim yöntemleri kullanılmıştır.

3.2.1. Sıralı ileriye doğru özellik seçimi

SFS algoritması, d boyutlu X özellik kümesinden $x_1, x_2, ...x_k$ özelliklerini belirli 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^+, eşitlik (3.6)'da gösterildiği gibi problemde kullanılan J kriter fonksiyonunu maksimum yapan özelliktir ve en iyi performansı göstermeyi amaçlar. Eşitlik (3.7)'de gösterildiği gibi kriteri sağlayan x^+ özelliği, özellik alt kümesine eklenir ve k değeri 1 arttırılarak eşitlik (3.6) ve (3.7) tekrarlanır (Ferri ve ark., 1994). Bu işlemler belirlenen durdurma kriteri karşılanana kadar devam eder. Şekil 3.6.'da akış diyagramı gösterilmiştir.



Şekil 3.6. Sıralı ileriye doğru özellik seçim yöntemi akış diyagramı

SFS, ilk adımda hedef ile en alakalı özelliği seçerek işleme başlar. İkinci adımda, ilk adımda seçilen özellik ile beraber hangi özellik eklenirse hedef ile en alakalı alt kümeye olacağını belirler ve durdurma kriteri sağlanana kadar bu şekilde devam eder.

$$x^{+} = \arg\max J\left(X_k + x\right), \text{ burada } x \in X - X_k \tag{3.6}$$

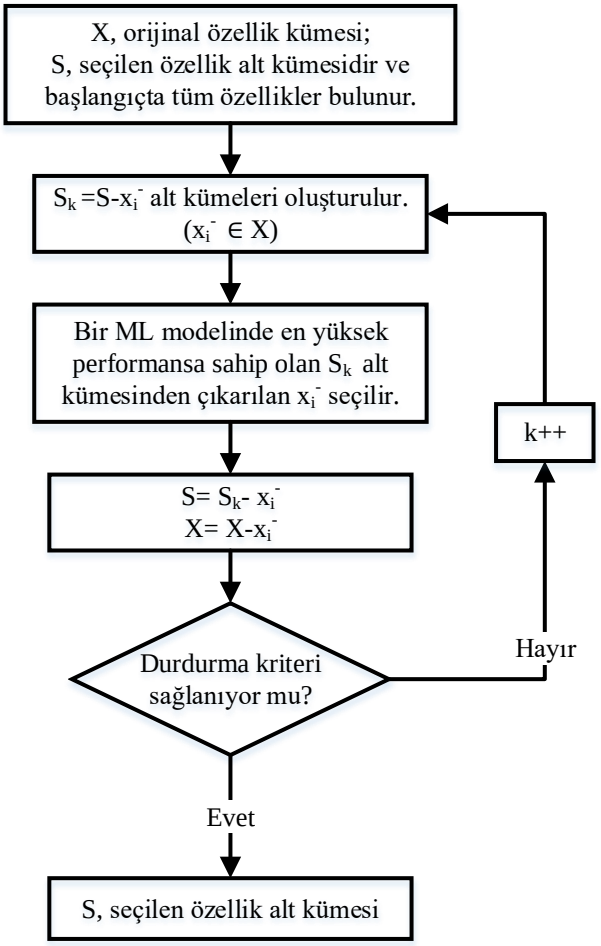
$$X_{k+1} = X_k + x^{+} \tag{3.7}$$

3.2.2. Sıralı geriye doğru özellik seçimi

SBS algoritması, d boyutlu X özellik kümesinden $x_1, x_2, ...x_k$ özelliklerini belirli kritere göre seçerek X_k özellik alt kümesini oluşturur $\left(k < d\right)$. 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 $\left(k < d\right)$. Verilen özellik küme seti boyutu $\left(k = d\right)$ ile algoritma başlar. Sonraki adımda X_k özellik kümesinden x^{-} özelliği çıkarılır. x^{-}, eşitlik (3.8)'de gösterildiği gibi problemde kullanılan J kriter fonksiyonunu maksimum yapan özelliktir ve en iyi performansı göstermeyi amaçlar. Eşitlik (3.9)'da gösterildiği gibi kriteri sağlayan x^{-} özelliği, özellik alt kümesinden çıkarılır ve k değeri 1 azaltılarak eşitlik (3.8) ve (3.9) tekrarlanır. Bu işlemler belirlenen durdurma kriteri karşılanana kadar devam eder (Ferri ve ark., 1994). Şekil 3.7.'de akış diyagramı gösterilmiştir.

$$x^{-} = \arg\max J\left(X_k - x\right), \text{ burada } x \in X_k \tag{3.8}$$

$$X_{k-1} = X_k - x^{-} \tag{3.9}$$



Şekil 3.7. Sıralı geriye doğru özellik seçim yöntemi akış diyagramı

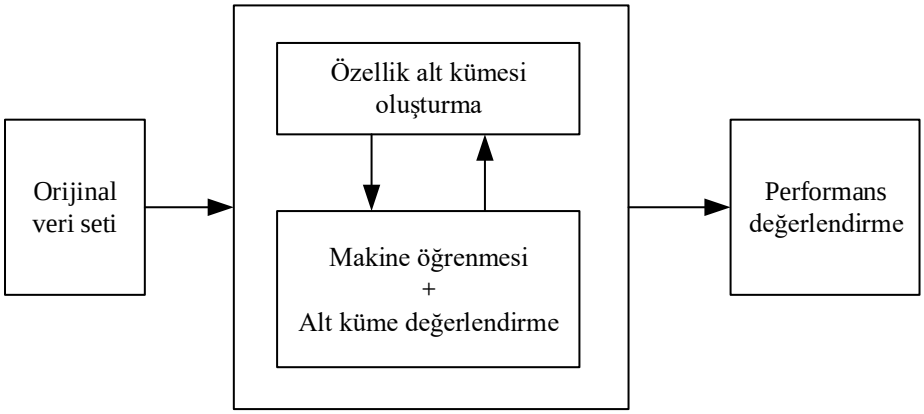
3.3. Gömülü Özellik Seçim Yöntemi

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 (Saeys ve ark., 2008). Gömülü FS yönteminin genel yapısı Şekil 3.8.'de verilmiştir.

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 (Kuhn ve Johnson, 2019).

Bir filtre yaklaşım, sarmal yaklaşıma göre daha az hesap karmaşıklığına sahiptir. Çünkü alt küme değerlendirmek için bağımsız alt küme değerlendirme kriteri kullanır (Kumar ve Minz, 2014). Gömülü yaklaşım, filtre yaklaşıma zaman karmaşıklığı açısından benzerdir. Düşük zaman karmaşıklığı ve yüksek ölçeklenebilir özellik seçim algoritmaları en uygun olandır.



Şekil 3.8. Gömülü FS modeli genel yapısı

Gömülü yaklaşım, en iyi özellikleri 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. En çok kullanılan gömülü metotlar, genelde katsayı değerlerinin bir eşik değerine bağlı olarak özelliklerin katsayılarını ya azaltan ya da cezalandıran teknikler kullanılır. Bu teknikler, tahmine dayalı algoritmaları optimize etmek için cezalandırmalar getirerek daha az özellik ile çalışılmasını sağlar. Örneğin, Lasso algoritması (Fonti ve Belitser, 2017) katsayıları mutlak sıfıra düşüren bir cezalandırma tipi kullanır. Katsayısı sıfır olan bir özellik,

modelden çıkarılmış olur. Ağaç toplulukları genelde benzer yapıdadırlar ancak RF gibi bazı algoritmalar bir ağaç oluşurken alakasız özellikler üzerinde bölmeleri zorlar.

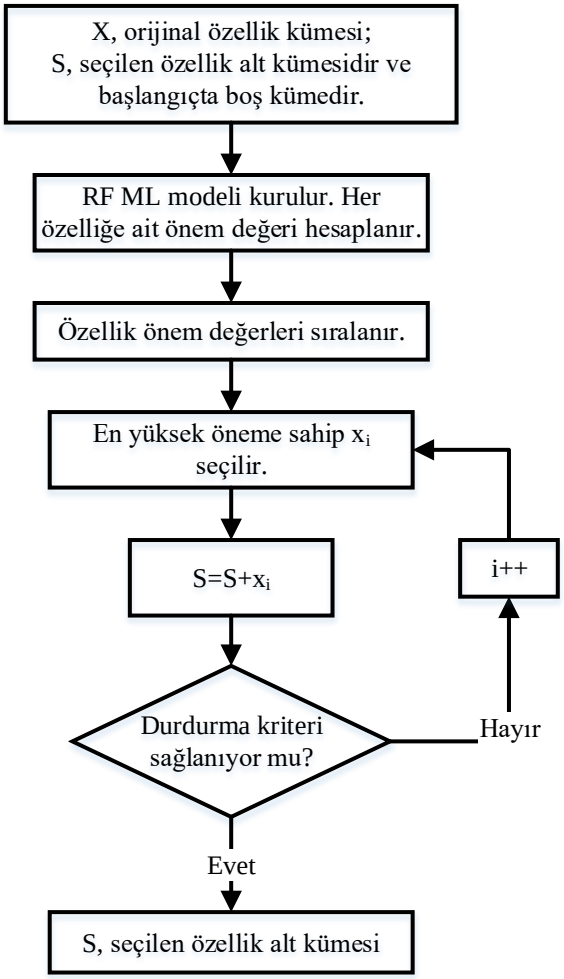
Gömülü yaklaşımlar, özellik seçme süreci modele uydurma süreci ile birlikte oluştuğundan nispeten hızlıdır ve ek başka herhangi bir özellik seçme tekniği gerektirmez. Ayrıca, modeli optimize etmeye çalışan amaç fonksiyonu ile seçilen özellikler arasında doğrudan bir ilişki kurulmasını sağlar (Kuhn ve Johnson, 2019).

Bu çalışmada gömülü tip özellik seçiminden RF tarafından belirlenen özelliklerin önemine bağlı olarak özellik seçimi ve yinelemeli olarak RF tarafından belirlenen özelliklerin önemine bağlı özellik seçim yöntemleri kullanılmıştır.

3.3.1. 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 düğümden 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. Safsızlık kriteri için regresyon problemlerinde hem MSE hem de MAE ile varyans azaltma hesaplanabilir Bölme değişkenlerinin seçimi için safsızlığın yanlılığından dolayı sonuç özellik önemi metriği de yanlıdır (Grömping, 2009).



Şekil 3.9. Rastgele orman ağaçları tarafından belirlenen özelliklerin önemine dayalı özellik seçim yöntemi akış diyagramı

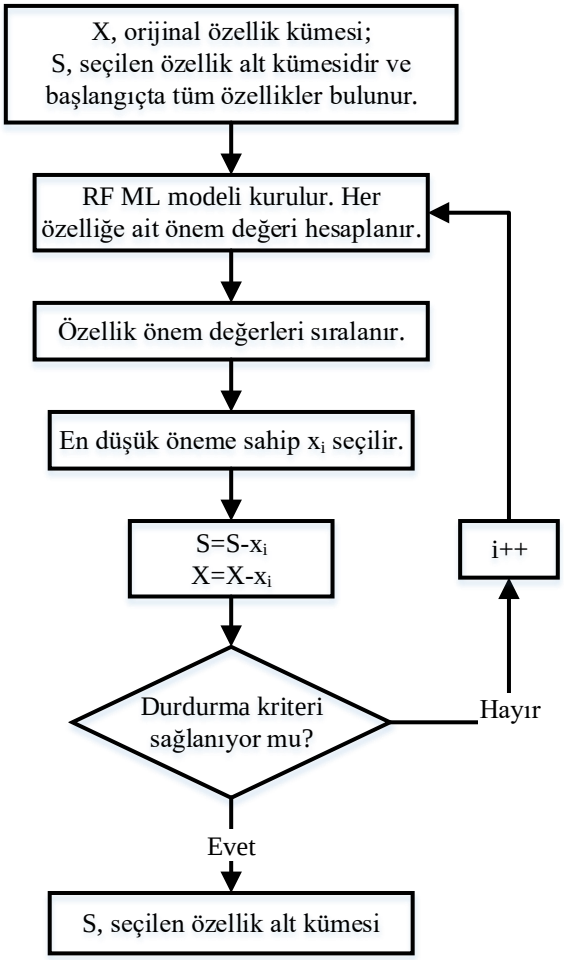
RF ile özellik önemine dayalı özellik seçim yöntemi (RFI), ağaçlar topluluğunda hesaplanan ve sadece ilişkili değişkenlere bağlı olan önemlilik değerine göre bir seçim gerçekleştirir. Şekil 3.9.'da RFI metoduna ait akış diyagramı verilmiştir. Regresyon problemlerinde bir ağaç eğitildiği zaman, her özelliğin 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. RF'de değişkenin nihai önemini belirlemek için her bir özelliğin sağladığı safsızlık azalmasının ağaçların ortalaması alınarak bulunur. (Strobl ve ark., 2007). 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 (Louppe, 2014). Değişken önemi için

eğitim esnasında, önyükleme örneğinde seçilmeyen verilerden oluşan ilgili torba dışı veriler seçilir ve bu verilerin tahmin hatası hesaplanır ve orman üzerinden ortalaması alınır. Eğitimden sonra belirli bir özelliğin önemi ölçülmek istendiğinde, bu özelliğin değerleri rastgele karıştırılırve tekrar tahmin hatası hesaplanır. O değişkenin önem değeri karıştırmadan önceki ve sonraki hata farkının ortalaması alınarak hesaplanır (Random forest, 2022).

3.3.2. Rastgele orman ağaçları makine öğrenim algoritması kullanılarak özelliklerin önemine dayalı yinelemeli özellik seçimi

RF kullanılarak yinelemeli özellik seçim yöntemi (RRFI), 3.3.1.'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. Şekil 3.10.'da RRFI metoduna ait akış diyagramı verilmiştir. RF ilk koşturulmasında 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 koşturulup 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, koşular arasında değil sadece bir koşu için karşılaştırılmış olur (Darst ve ark., 2018).

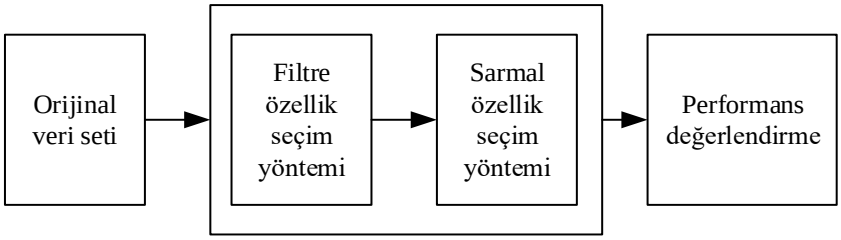
Yinelemeli bir FS ile sıralı bir FS arasında fark vardır. Yinelemeli metotlar, özellik elimine işlemi için özelliklerin ağırlık katsayılarını ya da özellik önemlerini kullanır ve her iterasyonda mevcut durum için yeniden hesaplamalar gerçekleştirir. Sıralı metotlar ise özellikleri elimine etmek ya da eklemek için kullanıcı tarafından belirlenen performans metriğine dayalı işlem yapmaktadır. Sıralı seçimlerde örneğin 2 özellik bir arada oldukları için düşük öneme sahip olabilirler. Fakat bunlardan sadece biri veri setinde bulunduğunda daha yüksek önem verebilme ihtimali vardır. Bu yüzden tek başına olduğunda önemli olacak özelliklerin kaçırılmasını önlemek için yinelemeli seçimleri kullanmak avantajlı olabilir.



Şekil 3.10. Rastgele orman ağaçları tarafından belirlenen özelliklerin önemine dayalı yinelemeli özellik seçim yöntemi akış diyagramı

3.4. 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 ile de en iyi özellik alt kümesi bulunmaya çalışılır.Şekil 3.11.'de hibrit özellik seçim yönteminin genel yapısı verilmiştir. Hibrit metotlar, filtre metotların yüksek etkililiğinden, sarmal metotların ise yüksek doğruluğundan faydalanır (Jović ve ark., 2015).



Şekil 3.11. Hibrit FS modeli genel yapısı

Bu çalışmada gömülü tip özellik seçiminden RF tarafından belirlenen özelliklerin önemine bağlı olarak özellik seçimi ve yinelemeli olarak RF tarafından belirlenen özelliklerin önemine bağlı özellik seçim yöntemleri kullanılmıştır.

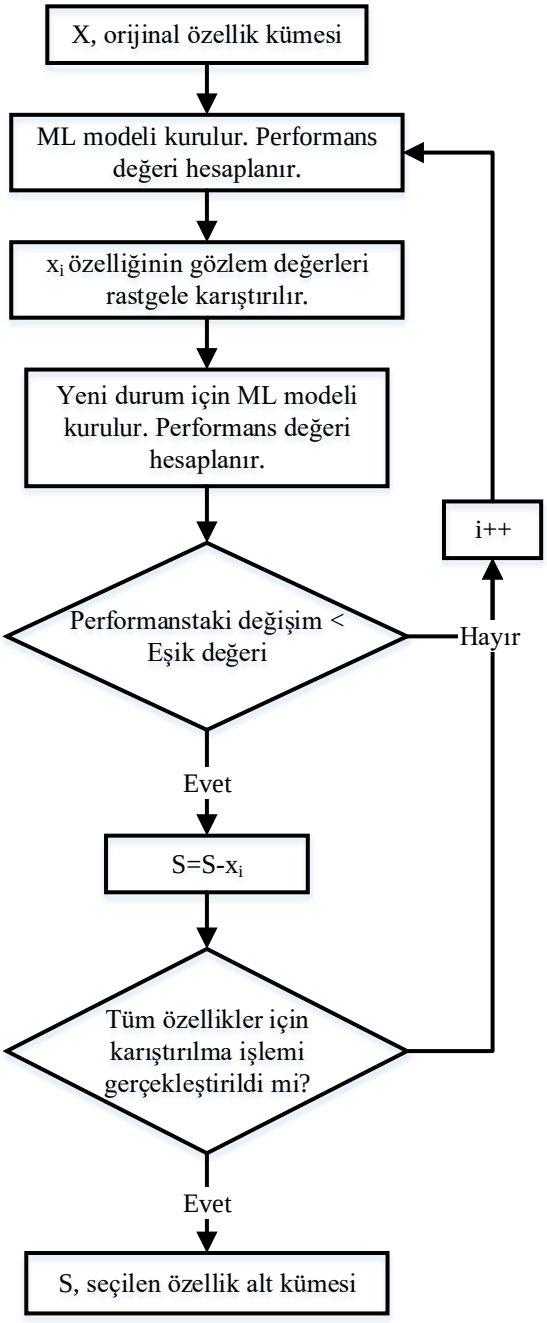
3.4.1. Rastgele karıştırma ile hibrit özellik seçimi

Rastgele karıştırma yöntemi (RS) bir özelliğin gözlem değerlerini rastgele karıştırılmasına dayalı özellik seçimi yapmaktadır. Hibrit özellik seçim yöntemi, filtre metotlar gibi belirli kriterlere göre özellikleri tek tek değerlendirir ve sarmal yöntemler gibi aday özellik alt kümelerini belirli bir öğrenme algoritması ile test ederek en uygun alt kümesini seçer.

Şekil 3.12.'de RS'ye ait akış diyagramı verilmiştir. 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 (Kursa ve Rudnicki, 2010). Performanstaki düşüşü değerlendirmek için bir eşik değeri belirlenmelidir.

Bu çalışmada, SVR ML metodu ile model inşa edilerek eğitim gerçekleştirilmiştir. Performans değerlendirmeleri için MSE performans metriği kullanılmıştır. Veri setindeki bir özelliğin gözlem değerleri rastgele karıştıtılarak yeni durumdaki

performans değeri hesaplanmıştır. Her özellik için tek tek karıştırma işlemi yapılıp performansta oluşan değişimler kaydedilmiştir. Belirlenen eşik değerine göre performanstaki değişimler değerlendirilmiştir.



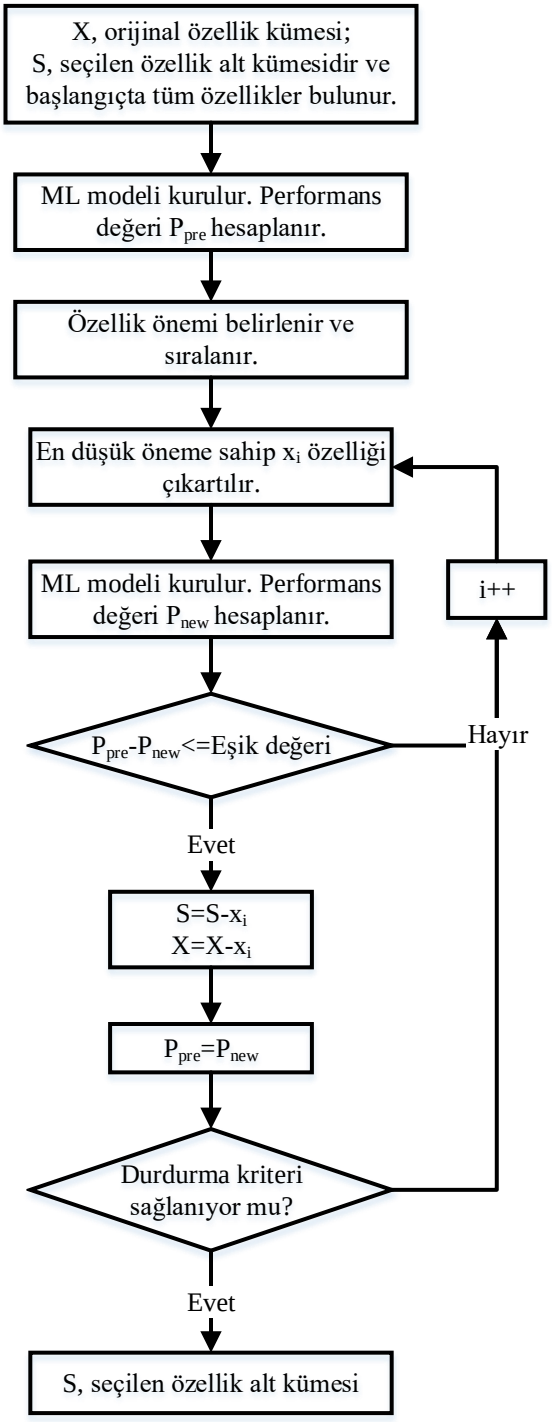
Şekil 3.12. Rastgele karıştırma ile hibrit özellik seçim yöntemi akış diyagramı

3.4.2. Yinelemeli özellik eleme ile hibrit özellik seçimi

Yinelemeli özellik eleme ile özellik seçim yöntemi (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 yararsız 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 (Chen ve Jeong, 2007).

Bu FS yöntemi Şekil 3.13.'de akışn diyagramı verildiği gibi önce tüm özelliklerin bulunduğu veri seti ile işleme başlar. Bir ML metodu ile performans değerlendirilir ve özellikler önem değerlerine göre sıralanır. Sonraki adımda, en az öneme sahip özellik veri setinden çıkarılması durumunda yeniden eğitilen modelin performansı değerlendirilir. Özellik çıkarıldıktan sonra performansta belirli eşik değerine göre bir iyileşme söz konusu ise bu özellik veri setinden kalıcı olarak çıkartılır. Bir sonraki adımda karşılaştırma yapılacak yeni performans değeri bu değer ile güncellenir. Durdurma kriteri karşılanan kadar işlemler bir sonraki adımda en az öneme sahip özelliğin veri setinden çıkartılıp modelin performansının değerlendirilmesi ile yinelemeli olarak devam eder.

Bu çalışmada, GBR ML metodu ile model inşa edilerek eğitim gerçekleştirilmiştir. Veri setinden her çıkarılan özellik için performans R^2 performans metriğine göre değerlendirilmiştir. Performanstaki düşüş değerleri seçilmek istenen özellik sayısına göre bir eşik değeri ile karşılaştırılmıştır. Eğer bir özellik çıkarıldıktan sonra R^2 performans değeri artmış ise o özellik veri setinden çıkarılmıştır. Aksi durumda özellik veri setinde kalmıştır.



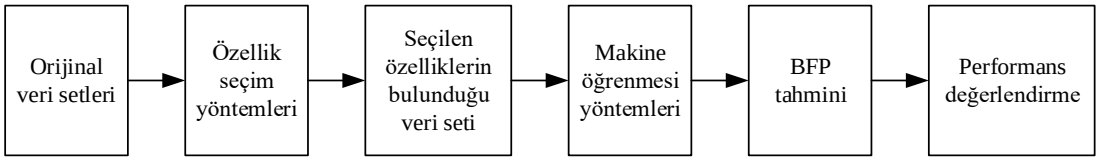
Şekil 3.13. Yinelemeli özellik eleme ile hibrit özellik seçim yöntemi akış diyagramı

BÖLÜM 4. DENEYSEL ÇALIŞMALAR VE SONUÇLAR

Bu çalışmada BFP regresyon tahmini için 2 adet veri seti kullanılmıştır. Veri setlerine 4 tip özellik seçim yöntemlerinden 8 FS algoritması kullanılarak veri setinden özelliklerin seçilmesi sağlanmıştır. Yeni oluşturulan özellik alt kümelerine 4 ML algoritması uygulanarak BFP tahmini gerçekleştirilmiştir. 13 özellik bulunan VS1 setinden 5 (%40), 8 (%60), ve 11 (%85) özellik seçilmiştir. 38 özellik bulunan VS2 setinden 8 (%20), 20 (%40), ve 32 (%85) özellik seçilmiştir. Seçilen özellikler ile kurulan ML modellerinin BFP tahmin değerleri 4 karşılaştırma metriği kullanılarak değerlendirilmiştir. Ayrıca ML modellerinin giriş özellik sayılarına göre eğitim süreleri gösterilmiştir.

4.1. Veri Setine Genel Bakış

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 4.1. ile verilmiştir. Regresyon metotlarının 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 4.1. Özellik seçim işlemleri için temel adımlar

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

Tablo 4.1. VS2 veri setindeki 25 özelliğin BFP ile ilişki katsayıları

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

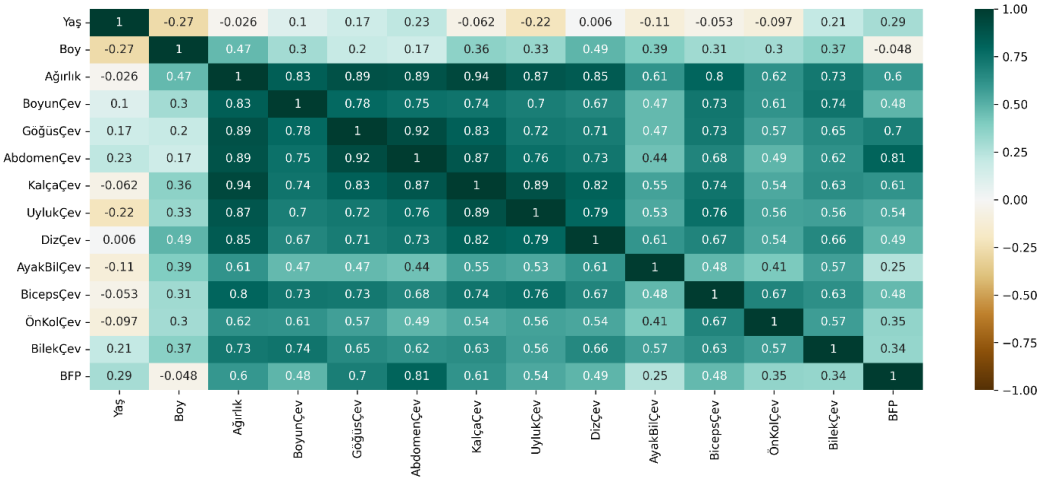
VS1 veri setindeki özelliklerin ikili korelasyonunu gösteren renkli ısı haritası Şekil 4.2.'de verilmiştir. Birbiri arasında yüksek ilişkiye sahip özellikler mevcuttur. Örneğin bağımsız özelliklerden göğüs çevresi ile abdomen çevresi, ağırlık ile kalça çevresi

arasında diğerlerine göre daha yüksek ilişki bulunmaktadır. Bağımlı değişken BFP ile abdomen çevresi arasında yüksek ilişki vardır.

Tablo 4.2. VS1 için özelliklerin gözlem değerleri hakkında bilgi

Özellik Adı	Örnek Sayısı	Ortalama	Standart Sapma	Minimum	Maksimum
Yaş	248	44,92	12,67	22	81
Boy (cm)	248	178,59	6,51	162,56	196,85
Ağırlık (kg)	248	81,24	13,15	56,70	164,72
Boyun Çevresi (cm)	248	38,02	2,41	31,10	51,20
Göğüs Çevresi (cm)	248	100,88	8,32	83,40	136,20
Abdomen Çev. (cm)	248	92,65	10,68	70,40	148,10
Kalça Çevresi (cm)	248	99,91	7,04	85,30	147,70
Uyluk Çevresi (cm)	248	59,44	5,15	49,30	87,30
Diz Çevresi (cm)	248	38,60	2,37	33,00	49,10
Ayak Bileği Çev. (cm)	248	23,12	1,70	19,10	33,90
Biceps Çevresi (cm)	248	32,31	2,99	25,30	45,00
Ön Kol Çevresi (cm)	248	28,68	2,01	21,00	34,90
Bilek Çevresi (cm)	248	18,24	0,92	15,80	21,40
BFP (%)	248	19,25	8,22	3,00	47,50

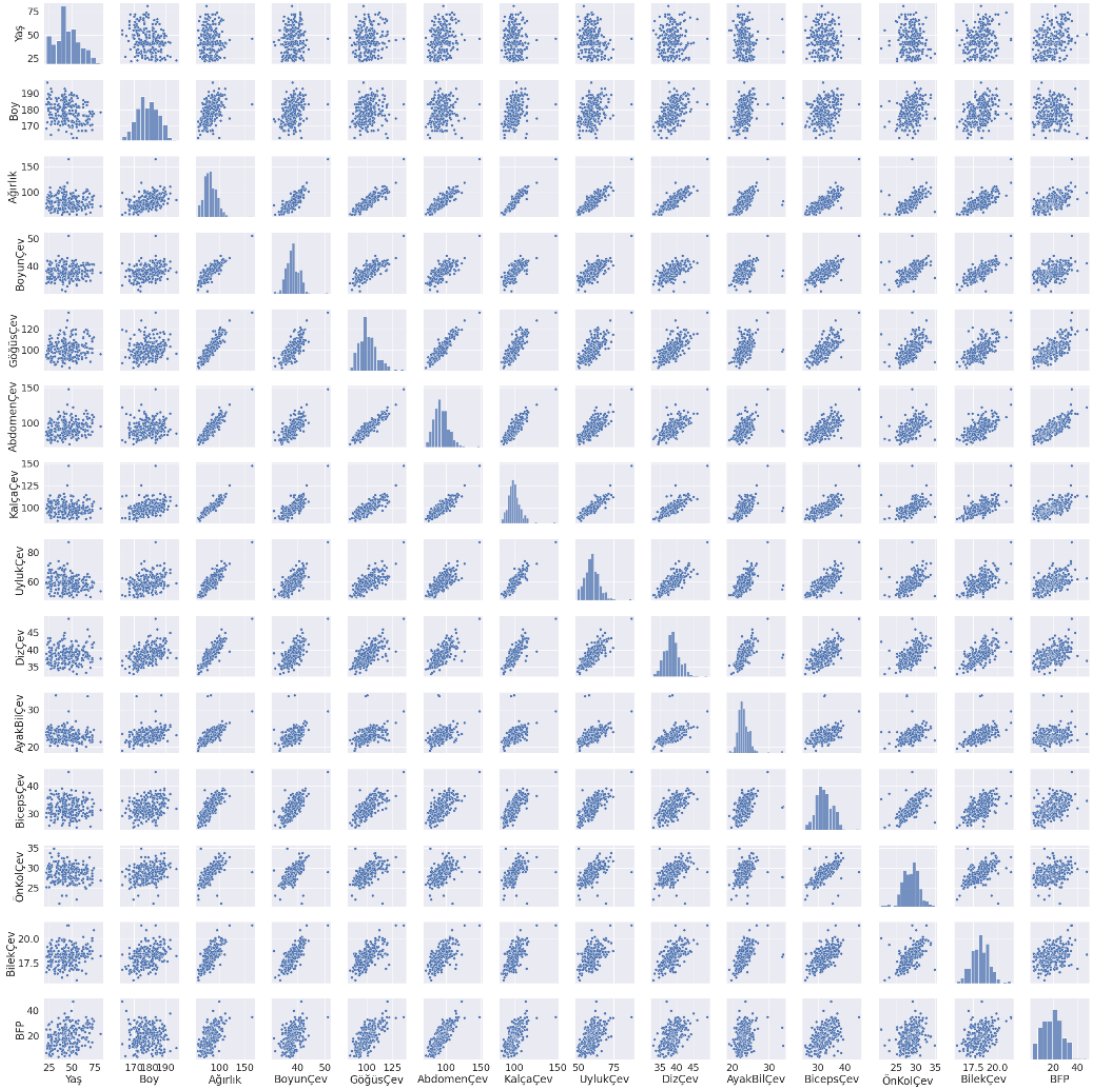
Şekil 4.3.'te de VS1 veri setindeki özellikler arasındaki tüm ikili ilişkiler farklı bir açıdan verilmiştir. Bu gösterimden, özellikler arasında doğrusal ve doğrusal olmayan ilişkilerin bulunduğu görülmektedir. Örneğin abdomen çevresi ile kalça çevresi arasında doğrusal bir ilişki var iken boy, yaş gibi 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.



Şekil 4.2. Özellikler arasındaki korelasyon ilişkisini gösteren renkli ısı haritası

Tablo 4.1.'de VS2 veri setindeki VS1 veri setinden istatistiksel yöntemlerle elde edilen 25 özelliğin BFP hedef özelliği ile olan ilişkisi gösterilmiştir. Verilen değerler korelasyon katsayılarıdır.

Tablo 4.2.'de VS1 veri setinde bulunan özelliklere ait gözlem değerleri hakkında genel bilgi mevcuttur. 22 ve 81 yaş aralığında %3 ile %47,5 BFP değerine sahip 248 bireye ait çeşitli vücut ölçümleri toplanmıştır. Tabloya göre her verinin ortalama değerleri etrafında nasıl değişkenlik gösterdiğini standart sapma bilgisi ile ulaşılmaktadır.



Şekil 4.3. VS1'deki özelliklerin ikili ilişkileri

4.2. Regresyon Parametreleri ve Karşılaştırma Metrikleri

4.2.1. Parametre ayarları

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. 5 katmanlı çapraz doğrulamada N adet gözlem değeri bulunan veri setinde, her setin N/5 örneği test için ve geri kalan kısım eğitim ve doğrulama işlemi için kullnaılmaktadır. Ayrıca veri setinin %80'i eğitim için, %20'si test ve değerlendirmeler için kullanılmıştır. Bu çalışmada RF, SVR, GBR ve XGBR için kullanılan parametreler:

a. 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.
b. 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.
c. 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.

d. 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.

4.2.2. 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 ortalama karesel hata (MSE), belirleme katsayısı (R^2), ortalama mutlak yüzde hatası (MAPE), medyan mutlak hata (MAE) metrikleri kullanıldı. Eşitlik (4.1), (4.2), (4.3) ve (4.4)'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. Eşitlik (4.1)'de verilen MSE, hatanın karesini alarak küçük hataları dahi cezalandırır. Sıfıra ne kadar yakın olursa o kadar iyi performans gösterdiği anlamına gelir.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2 \tag{4.1}$$

R^2, hataların karelerinin toplamı ile hesaplanır. Bağımlı değişkendeki toplam varyasyonun yüzde kaçının bağımsız değişkendeki varyasyon tarafından açıklandığını belirler. Eşitlik (4.2)'de $\bar{y}$ bağımlı değişkenin gerçek değerlerin ortalamasıdır.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{i=1}^{n}\left(y_j - \bar{y}\right)^2} \tag{4.2}$$

MAPE, bağımlı değişkenin gerçek değeri ile tahmin değeri arasındaki farkın yüzdelik cinsten ifadesi eşitlik (4.3)'teki gibidir.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_j - \hat{y}_j\right|}{y_j} \times 100 \tag{4.3}$$

MAE, hedef değişkenin gerçek değeri ile tahmin edilen değeri arasındaki farkın mutlak değerinin ortalamasıdır. MAE, eşitlik (4.4)'te verildiği gibi ifade edilir.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{4.4}$$

4.3. Özellik Seçiminde Kullanılan Regresyon Yöntemleri

Bu çalışmada filtre, sarmal, gömülü ve hibrit özellik seçim yöntemleri kapsamında 8 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.

a. MI ve UST FS yöntemleri, her özellikliğin hedef özellik ile arasındaki ilişki ile ilgilenir. Durdurma kriterini sağlayana dek hedef özellik ile en çok ilişkili olan bağımsız özellikler seçilmiştir. FS yöntemleri özellikleri seçerken bir ML tekniği kullanmaz. Sadece değerlendirme aşamasında ML kullanır.
b. 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.

c. 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.
d. 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.

4.4. Deneysel Sonuçlar

VS1 ve VS2 deney setleri kullanılarak çeşitli FS metotlarının farklı ML modelleri üzerindeki etkileri incelenmiştir. Tablo 4.3. ve Tablo 4.4.'te program 150 kez çalıştırıldıktan sonra ML modellerinin eğitim sürelerinin ortalamaları seçilen özellik sayısına göre verilmiştir. 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.

VS1 ve VS2 veri setlerinde özellik seçimi yapan 8 FS algoritmasının bağımlı değişken için tahminleri 4 farklı ML algoritması ile değerlendirilmiştir. Tablo 4.5. ile Tablo 4.13. arasında verildiği gibi sonuçlar 4 farklı performans metriği kullanılarak karşılaştırılmıştır. VS1 veri seti için özellik sayısı, orijinal veri setinin yaklaşık %40 (5 özellik)'ı, %60 (8 özellik)'ı ve %85 (11 özellik)'i olacak şekilde FS yöntemlerinin durdurma kriterleri ayarlanmıştır. VS2 veri seti için özellik sayısı ise yaklaşık %20 (8 özellik)'si, %40 (20 özellik)'ı ve %85 (32 özellik)'i olacak şekilde FS yöntemlerinin

durdurma kriterleri ayarlanmıştır. Tablo 4.5.'te FS algoritmaları uygulanmadan önceki VS1 ve VS2 veri setleri için ML modellerinin performansları verilmiştir.

Tablo 4.3. 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 4.4. 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	0,05775

Tablo 4.6. ile Tablo 4.9. arasında VS1 veri setine farklı FS algoritmaları uygulanarak sırasıyla 5, 8, 11 özellik seçilmiştir ve farklı ML modellerinin başarısı çeşitli performans metriklerine göre verilmiştir. Tablo 4.6. ile Tablo 4.9. arasında verilen ve VS1 veri setinden 5 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar hibrit FS metotlarından elde edilmiştir. Buradaki 4 FS tipine göre 5 özelliğin seçildiği ve 4 ML modeli ile tahmin değerlerinin hesaplandığı 16 performans sonucuna göre hibrit metot 14 durumda en iyi performansı göstermiştir. Bu performanslardan 9 sonuç RS hibrit FS metoduna aittir. Ayrıca verilen 2 farklı sarmal ve gömülü metot türleri kendi içlerinde aynı sonuçları vermiştir. Yani bu yöntemler aynı özellikleri seçmişlerdir. Bu yöntemler orijinal giriş uzayı boyutu az olan veri seti için farklı çözümler göstermemişlerdir. Benzer durum diğer sarmal ve gömülü FS yöntemlerin genelinde de görülebilmektedir. Tablo 4.6. ile Tablo 4.9. arasında verilen ve VS1 veri setinden 8 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar 5 özelliğin seçildiği duruma benzer olarak hibrit FS metotlarından elde edilmiştir. Buradaki 16 performans sonucuna göre hibrit metot 12 durumda en iyi performansı göstermiştir. Bu performanslardan 8 sonuç RS hibrit FS metoduna aittir. Tablo 4.6. ile Tablo 4.9. arasında verilen VS1 veri setinden 11 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar sarmal FS metotlarından elde edilmiştir. Buradaki 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 4.5. 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 4.10. ile Tablo 4.13. arasında VS2 veri setine farklı FS algoritmaları uygulanarak sırasıyla 8, 20, 32 özellik seçilmiştir ve farklı ML modellerinin başarısı çeşitli performans metriklerine göre verilmiştir. Tablo 4.10. ile Tablo 4.13. arasında verilen VS2 veri setinden 8 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar sarmal FS metotlarından elde edilmiştir. Toplam 16 performans sonucuna sarmal FS metot 14 durumda en iyi performansı göstermiştir. Bu performanslardan 10 durumda SBS metodu, 4 durumda da SFS metodu en iyi performansı göstermiştir. Tablo 4.10. ile Tablo 4.13. arasında 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 4.10. ile Tablo 4.13. arasında VS2 veri setinden 32 özellik seçen FS algoritmalarından genellikle en başarılı sonuçlar sarmal FS metotlarından elde edilmiştir. Toplam 16 performans sonucuna göre sarmal FS metotlarının başarılı olduğu 10 durumdan 6'sında SFS metodu en başarılı sonuçları vermiştir.

Tablo 4.6. ile Tablo 4.9. arasında filtre FS metotlarının en başarılı sonucu verdiği yalnızca 3 durum vardır. Bu performansların 2 durumu UST FS yöntemine aittir ve 1 durumda ise iki filtre FS yöntemi aynı sonucu vermiştir. Tablo 4.10. ile Tablo 4.13. arasında ise filtre FS metotlarından UST FS yöntemi sadece 1 durumda en başarılı sonucu vermiştir. İki veri seti için genelde yüksek oranda özellik seçildiğinde filtre seçim yöntemleri nadiren de olsa başarılı sonuçlar vermiştir.

Tablo 4.6. VS1'e uygulanan farklı FS yöntemlerinin RF modeline göre performansları

Özellik Sayısı		**5**				**8**				**11**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	13,733	0,794	2,843	0,161	13,178	0,798	2,984	0,169	12,637	**0,809**	2,917	0,167
	UST	13,733	0,794	2,843	0,161	13,178	0,798	2,984	0,169	12,637	**0,809**	2,917	0,167
	SFS	12,694	0,802	2,753	**0,151**	12,642	0,810	2,850	**0,152**	**12,458**	0,808	**2,850**	**0,154**
	SBS	12,694	0,802	2,753	**0,151**	12,642	0,810	2,850	**0,152**	**12,458**	0,808	**2,850**	**0,154**
	RFI	12,205	0,809	2,816	0,153	12,754	0,808	2,802	0,153	12,854	0,802	2,887	0,158
	RRFI	12,205	0,809	2,816	0,153	12,944	0,798	2,860	0,157	12,854	0,802	2,887	0,158
	RFE	12,304	0,807	2,748	0,162	12,975	0,807	2,871	0,158	12,660	0,805	2,873	0,157
	RS	**11,286**	**0,830**	**2,626**	0,162	**12,308**	**0,815**	**2,730**	0,163	12,730	0,803	2,873	0,160

Tablo 4.7. VS1'e uygulanan farklı FS yöntemlerinin SVR modeline göre performansları

Özellik Sayısı		**5**				**8**				**11**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	12,021	0,809	2,769	0,162	11,855	0,809	2,729	0,159	10,528	0,828	2,618	0,154
	UST	12,021	0,809	2,769	0,162	11,953	0,818	2,772	0,161	11,318	0,830	2,670	**0,151**
	SFS	10,694	0,825	2,605	0,159	10,833	0,823	2,621	0,160	11,196	**0,838**	2,676	0,161
	SBS	10,694	0,825	2,605	0,159	10,833	0,823	2,621	0,160	11,196	**0,838**	2,676	0,161
	RFI	10,477	0,829	2,639	0,158	10,934	0,821	2,620	0,158	10,990	0,826	2,644	0,158
	RRFI	10,477	0,829	2,639	0,158	10,981	0,822	2,632	0,157	10,990	0,826	2,644	0,158
	RFE	**10,310**	**0,831**	**2,587**	**0,156**	11,762	**0,832**	2,784	**0,155**	10,771	0,827	2,631	0,159
	RS	11,440	0,812	2,694	0,161	**10,287**	0,832	**2,580**	0,157	**10,030**	0,836	**2,545**	0,154

Tablo 4.8. VS1'e uygulanan farklı FS yöntemlerinin GBR modeline göre performansları

Özellik Sayısı		**5**				**8**				**11**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	13,987	0,778	3,084	0,180	16,176	0,777	3,280	0,183	13,875	0,796	3,099	0,174
	UST	16,394	0,777	3,230	0,176	15,189	0,774	3,104	0,179	13,875	0,796	3,099	0,174
	SFS	14,968	0,787	3,090	0,183	15,036	0,786	2,975	0,179	13,521	**0,816**	2,882	0,164
	SBS	14,968	0,787	3,090	0,183	15,036	0,786	2,975	0,179	13,521	**0,816**	2,882	0,164
	RFI	14,473	0,788	3,050	0,170	14,383	0,787	2,932	0,168	12,955	0,812	2,876	0,163
	RRFI	14,473	0,788	3,050	0,170	13,858	0,787	3,022	0,172	12,955	0,812	2,876	0,163
	RFE	13,126	0,803	**2,856**	0,176	14,552	0,800	3,004	**0,157**	12,732	0,812	2,877	**0,152**
	RS	**12,943**	**0,818**	2,966	**0,170**	**11,152**	**0,809**	**2,686**	0,169	**11,728**	0,800	**2,801**	0,165

Tablo 4.9. VS1'e uygulanan farklı FS yöntemlerinin XGBR modeline göre performansları

Özellik Sayısı		**5**				**8**				**11**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	14,961	0,788	3,112	0,171	17,068	0,759	3,115	0,180	13,517	0,782	2,982	0,172
	UST	15,554	0,772	3,074	0,165	13,975	0,798	**2,721**	0,162	13,517	0,782	2,982	0,172
	SFS	13,588	0,796	2,852	**0,160**	14,297	0,775	2,911	0,179	12,964	**0,805**	**2,760**	**0,150**
	SBS	13,588	0,796	2,852	**0,160**	14,297	0,775	2,911	0,179	12,964	**0,805**	**2,760**	**0,150**
	RFI	15,678	0,791	2,969	0,185	**12,203**	0,817	2,760	**0,155**	**12,640**	0,795	2,953	0,158
	RRFI	15,678	0,791	2,969	0,185	13,846	0,792	2,877	0,160	**12,640**	0,795	2,953	0,158
	RFE	15,078	0,802	3,016	0,161	12,518	**0,833**	2,807	0,158	14,045	0,802	2,980	0,165
	RS	**11,278**	**0,831**	**2,737**	0,165	13,113	0,793	2,894	0,167	13,089	0,796	2,854	0,160

Tablo 4.10. VS2'ye uygulanan farklı FS yöntemlerinin RF modeline göre performansları

Özellik Sayısı		**8**				**20**				**32**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	13,471	0,769	3,009	0,190	14,310	0,774	3,137	0,186	12,899	0,792	2,940	0,175
	UST	13,481	0,769	3,016	0,190	14,441	0,780	3,118	0,183	12,864	**0,795**	2,967	0,181
	SFS	12,490	0,795	2,834	0,175	12,425	0,793	**2,771**	**0,170**	**12,374**	0,788	**2,815**	0,176
	SBS	**12,033**	**0,819**	**2,832**	**0,165**	12,361	**0,795**	2,832	0,174	12,767	0,792	2,909	0,175
	RFI	13,582	0,792	2,971	0,177	13,188	0,792	2,939	0,170	12,791	0,788	2,951	0,178
	RRFI	13,750	0,795	3,019	0,171	12,714	0,794	2,893	0,174	12,891	0,792	2,961	0,177
	RFE	13,079	0,799	2,856	0,168	12,370	0,788	2,879	0,177	12,776	0,785	2,946	0,177
	RS	12,680	0,789	2,887	0,176	**12,327**	0,793	2,851	0,177	13,259	0,788	2,979	**0,175**

Tablo 4.11. VS2'ye uygulanan farklı FS yöntemlerinin SVR modeline göre performansları

Özellik Sayısı		**8**				**20**				**32**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	12,035	0,803	2,777	0,172	12,582	0,807	2,883	0,174	11,479	0,814	2,700	0,156
	UST	11,818	0,807	2,813	0,174	12,919	0,801	2,945	0,173	11,318	0,815	2,694	0,157
	SFS	**11,013**	**0,820**	2,675	0,165	10,976	0,821	2,638	0,158	11,187	0,822	2,677	0,154
	SBS	11,158	0,812	2,706	0,163	11,140	0,822	2,661	0,159	11,088	0,824	2,691	0,155
	RFI	12,175	0,801	2,806	0,168	12,065	0,808	2,779	0,159	10,857	0,823	2,625	0,152
	RRFI	11,234	0,816	**2,618**	0,165	**9,983**	**0,837**	2,534	0,153	11,681	0,809	2,789	0,163
	RFE	12,206	0,800	2,778	**0,162**	10,025	0,836	**2,455**	**0,148**	**9,753**	**0,841**	2,508	0,151
	RS	12,652	0,805	2,862	0,169	11,509	0,818	2,660	0,153	9,958	0,837	**2,492**	**0,145**

Tablo 4.12. VS2'ye uygulanan farklı FS yöntemlerinin GBR modeline göre performansları

Özellik Sayısı		**8**				**20**				**32**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	14,928	0,745	3,244	0,190	15,276	0,752	3,386	0,188	13,542	0,809	2,959	0,159
	UST	16,366	0,748	3,350	0,194	16,818	0,746	3,425	0,188	14,312	0,795	3,057	0,165
	SFS	**12,930**	**0,835**	2,897	0,174	14,913	0,791	3,072	0,184	13,562	**0,815**	3,072	0,172
	SBS	13,874	0,792	**2,860**	**0,161**	14,220	0,777	**2,960**	0,177	**13,075**	0,799	**2,904**	**0,159**
	RFI	13,382	0,763	3,077	0,190	14,858	0,785	3,121	0,171	13,243	0,806	3,022	0,164
	RRFI	15,333	0,774	3,157	0,180	13,808	0,800	3,116	0,175	13,602	0,802	2,992	0,161
	RFE	13,652	0,794	2,944	0,166	**13,496**	**0,816**	3,042	**0,163**	13,412	0,810	3,066	0,171
	RS	13,497	0,780	3,038	0,202	14,995	0,788	3,131	0,188	13,438	0,798	2,930	0,181

Tablo 4.13. VS2'ye uygulanan farklı FS yöntemlerinin XGBR modeline göre performansları

Özellik Sayısı		**8**				**20**				**32**			
Performans Metriği		**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**	**MSE**	**R^2**	**MAE**	**MAPE**
Özellik Seçim Yöntemi	**MI**	16,293	0,739	3,345	0,201	14,778	0,778	3,169	0,184	14,673	0,800	3,109	0,175
	UST	15,876	0,758	3,289	0,195	16,657	0,777	3,356	0,192	14,429	0,788	3,050	0,172
	SFS	14,950	0,806	3,035	0,180	14,329	0,805	3,110	0,174	**11,221**	0,819	**2,794**	**0,155**
	SBS	**10,920**	**0,836**	**2,678**	**0,156**	14,208	0,782	3,021	0,174	12,586	**0,828**	2,957	0,169
	RFI	14,202	0,788	3,155	0,195	14,447	0,803	3,010	0,179	12,819	0,825	2,954	0,162
	RRFI	14,316	0,785	3,075	0,178	14,576	0,799	3,159	0,177	13,238	0,819	2,992	0,159
	RFE	13,891	0,809	3,036	0,177	**12,499**	**0,830**	**2,998**	**0,162**	12,593	0,828	2,946	0,174
	RS	15,767	0,780	3,143	0,195	14,597	0,790	3,016	0,182	14,538	0,808	3,046	0,186

Tablo 4.6. ile Tablo 4.13. arasındaki tüm durumlar ele alındığında toplamda 96 farklı durum için performans değerlendirilmiştir Bunlardan 47 durumda hibrit FS yöntemi, 39 durumda sarmal FS yöntemi, 6 durumda gömülü FS yöntemi ve 4 durumda filtre FS yöntemi performans metriklerine göre en iyi sonucu vermiştir.

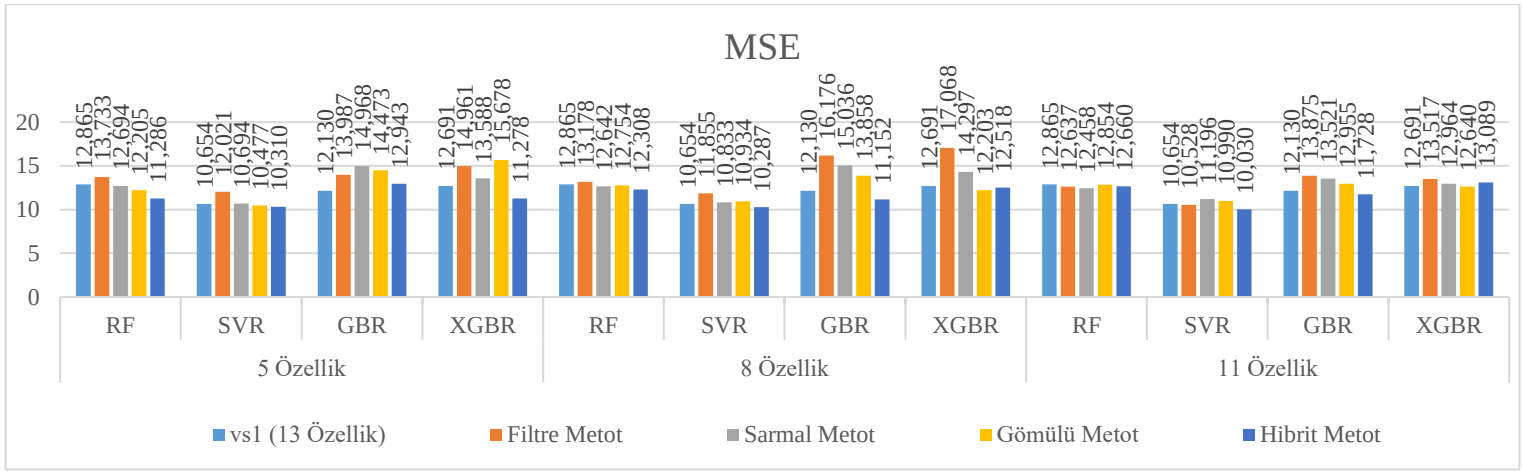
Tablo 4.6. ile Tablo 4.13. arasında verilen ML model performansları her FS tekniği için karşılaştırılabilir. Bu tablolarda her FS yöntemi, 96 durum için değerlendirilebilmektedir. Filtre FS yöntemlerinden MI yöntemi 17 durumda, UST yöntemi 17 durumda orijinal veriler ile kurulan modellerin performansını geliştirmiştir. Sarmal FS yöntemlerinden SFS yöntemi 39 durumda, SBS yöntemi 40 durumda orijinal veriler ile kurulan modellerin performansını geliştirmiştir. Gömülü FS yöntemlerinden RFI yöntemi 29 durumda, RRFI yöntemi 30 durumda orijinal veriler ile kurulan modellerin performansını geliştirmiştir. Hibrit FS yöntemlerinden RFE yöntemi 47 durumda, RS yöntemi 44 durumda orijinal veriler ile kurulan modellerin performansını geliştirmiştir.

Şekil 4.4. ile Şekil 4.11. arasında 4 tip olarak özellik seçim yöntemlerinin performansları ile orijinal veri setinin performansı farklı metrik değerlerine göre karşılaştırılabilmektedir. Burada 4 FS tipi değerlendirilmek için alt FS yöntemlerinden en iyi sonuçları veren değerler alınmıştır. FS yöntemlerinin orijinal veri seti kullanılarak gerçekleştirilen tahminler ile karşılaştırıldığında modelin performansına katkısı olabileceği görülmektedir. Farklı ML modellerinin seçilen özelliklerden oluşturulan özellik alt kümelerini girdi olarak kullandıkları performansları ve orijinal veri setini girdi olarak kullandıkları performansları verilmiştir. Kullanılan 2 veri seti için 4 ML modelinin kurulduğu, tahmin sonuçlarının 4 performans metriğine göre değerlendirildiği ve 3 farklı sayıda özellik alt kümesinin seçildiği düşünülünce her tip FS algoritmasının kullanıldığı 96 farklı durum oluşmaktadır. 96 durum için değerlendirilen FS algoritmalarından hibrit FS metodu 62 durumda, sarmal FS metodu 38 durumda, gömülü FS metodu 42 durumda ve filtre FS metodu 20 durumda kurulan ML modellerinin performansını geliştirmiştir. FS ve ML metotlarının performanslarının geliştiği durumların sayısının dağılımı Tablo 4.14.'te verilmiştir.

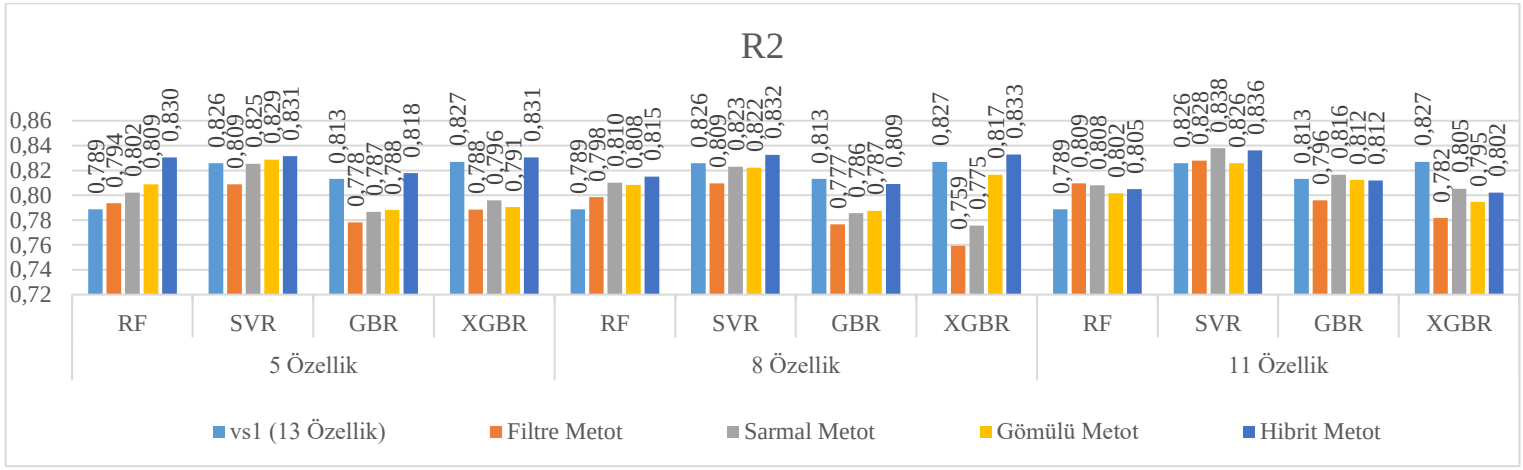
Hibrit metotlar genellikle model performansını geliştirmiştir. RF ML modeli genellikle özellik seçimi uygulandıktan sonra performansı gelişmiştir.

Tablo 4.14. FS metotları kullanılarak ML modeline göre performansı gelişen durumların sayısı

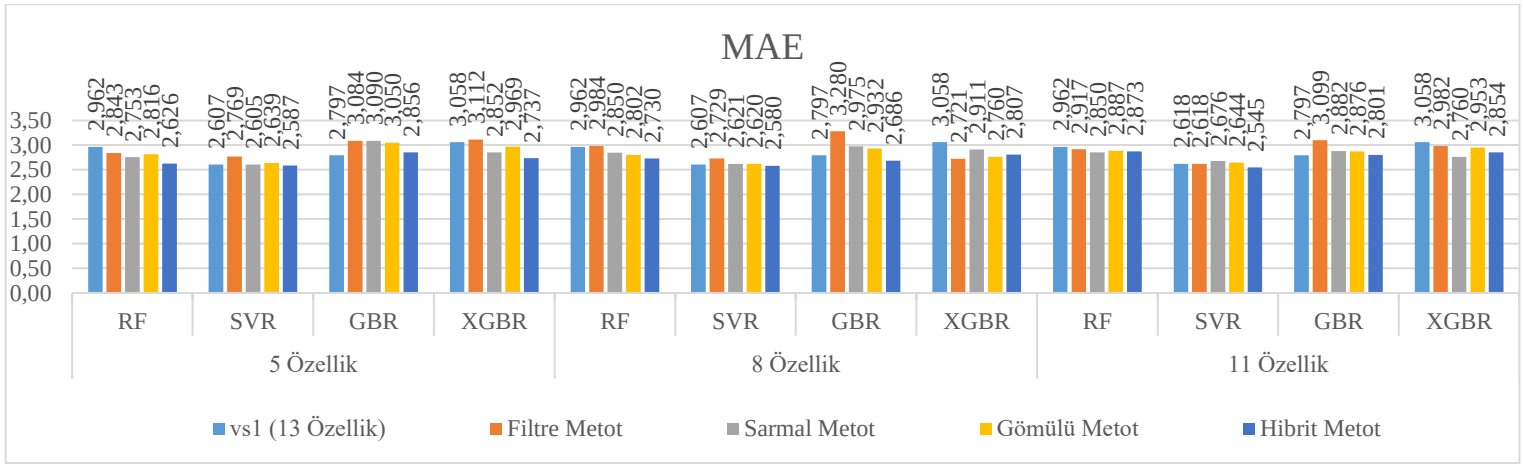
	Hibrit FS	**Gömülü FS**	**Sarmal FS**	**Filtre FS**
RF	18	15	19	9
SVR	17	7	4	5
GBR	12	6	10	4
XGBR	15	10	8	2



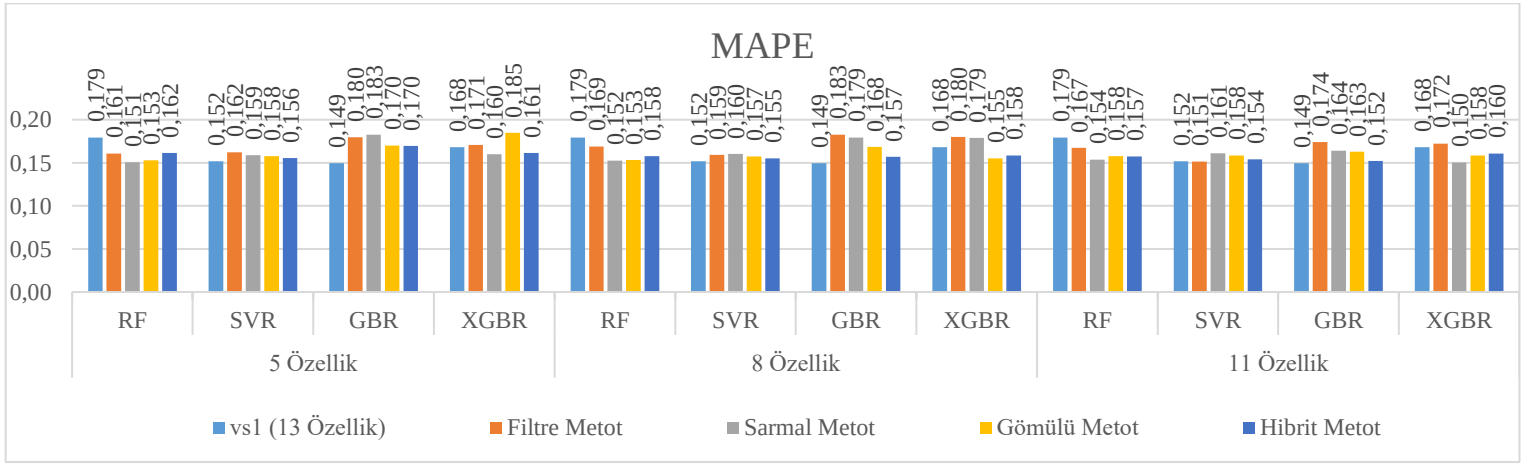
Şekil 4.4. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MSE değerlerinin karşılaştırılması



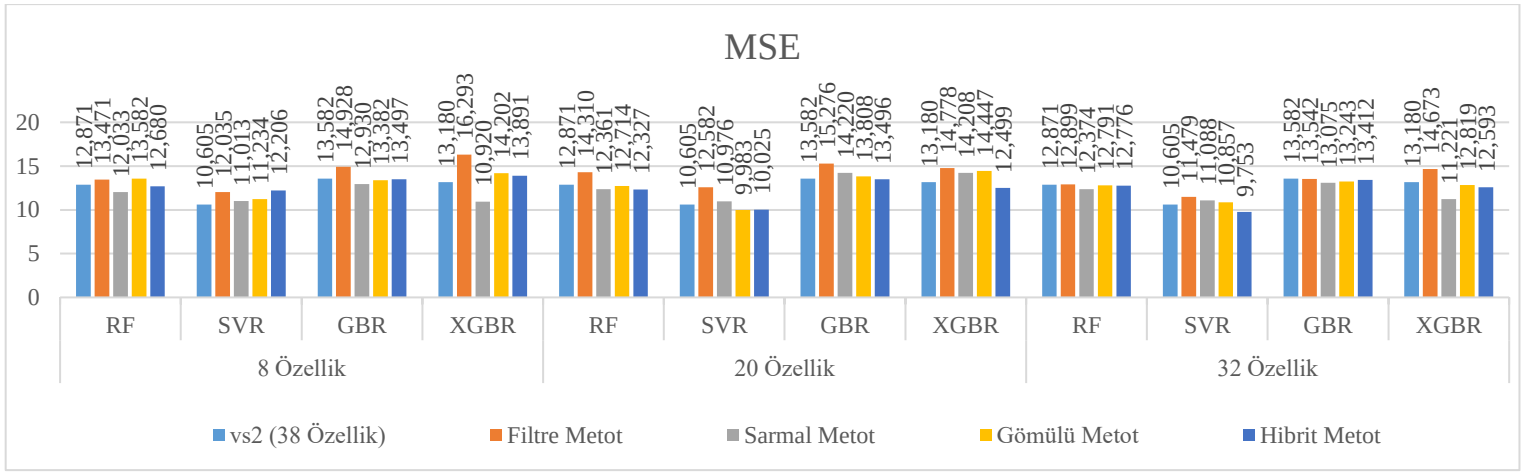
Şekil 4.5. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre R^2 değerlerinin karşılaştırılması



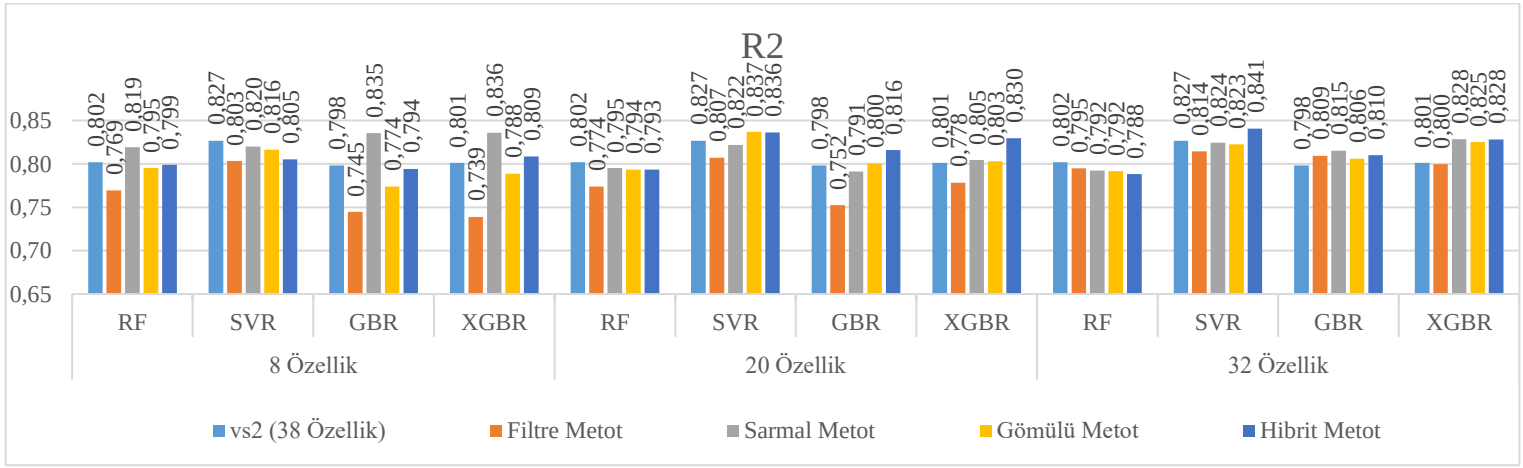
Şekil 4.6. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAE değerlerinin karşılaştırılması



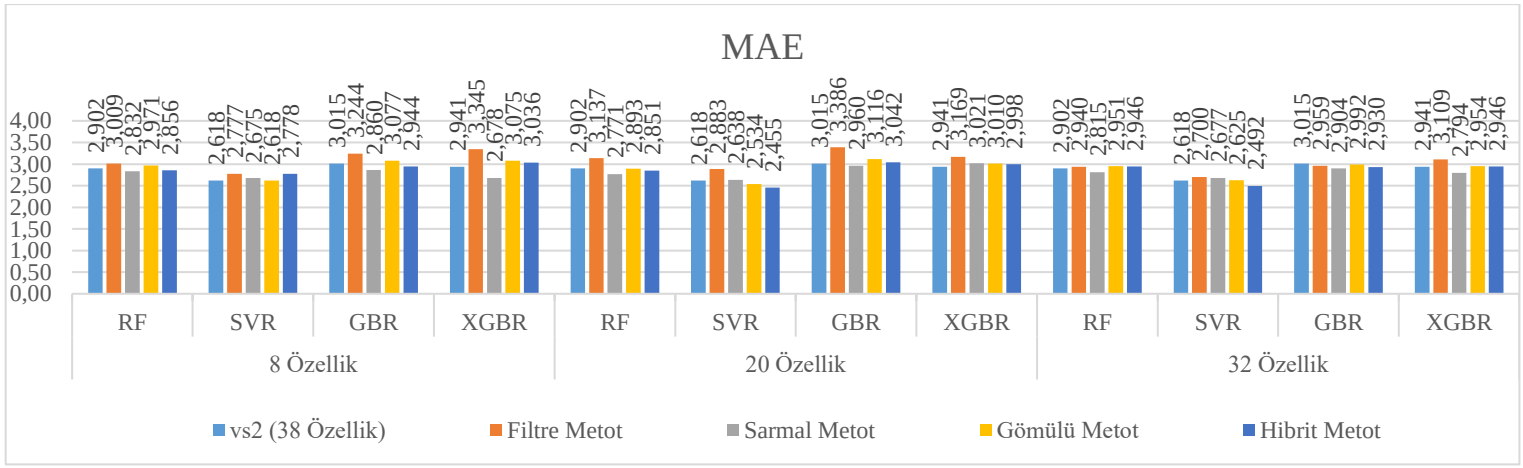
Şekil 4.7. VS1'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAPE değerlerinin karşılaştırılması



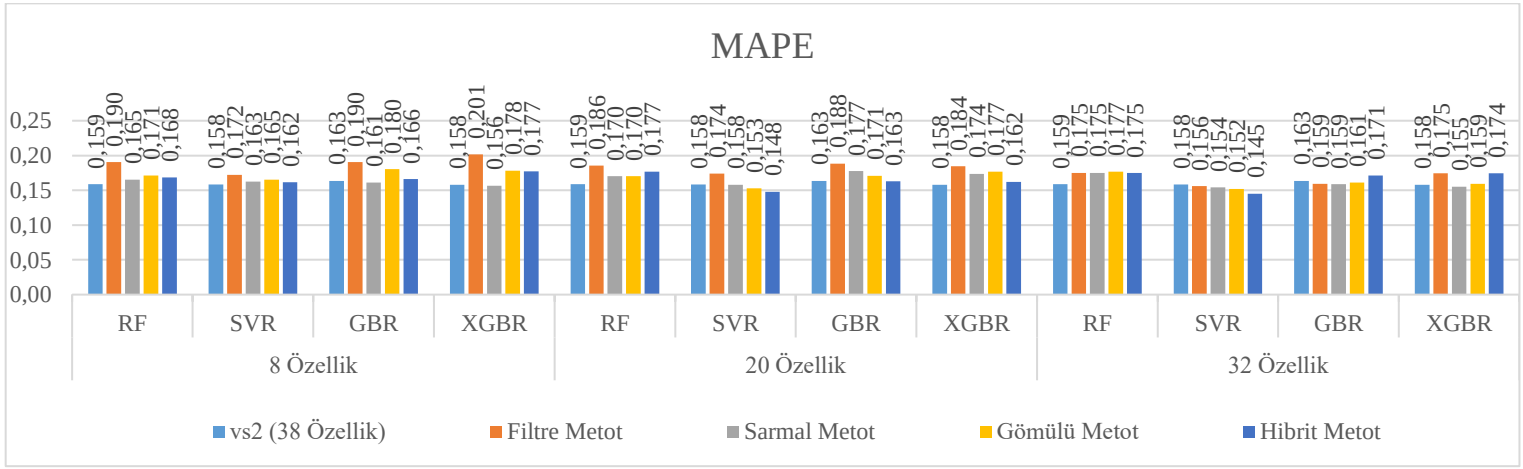
Şekil 4.8. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MSE değerlerinin karşılaştırılması



Şekil 4.9. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre R^2 değerlerinin karşılaştırılması



Şekil 4.10. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAE değerlerinin karşılaştırılması



Şekil 4.11. VS2'den farklı sayılarda özellik seçen FS metotlarının ML modellerine göre MAPE değerlerinin karşılaştırılması

4.5. Sonuçlar

Bu çalışmada, 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 8 FS yönteminin karşılaştırılması amaçlanmıştır. FS algoritmaları ile seçilen özellikler 4 farklı ML algoritması ile eğitilmiştir. Elde edilen sonuçlar 4 performans metriği ile değerlendirilmiştir. BFP için tahmin gerçekleştirmek için 4 ML modelinin kurulduğu, tahmin sonuçlarının 4 performans metriğine göre değerlendirildiği ve her veri setinden 3 farklı sayıda özellik alt kümesinin seçildiği 96 farklı durum oluşmuştur. Bunlardan 47 durumda hibrit FS yöntemi, 39 durumda sarmal FS yöntemi, 6 durumda gömülü FS yöntemi ve 4 durumda filtre FS yöntemi performans metriklerine göre en iyi sonucu vermiştir. FS yöntemlerinin model performansını geliştirebileceği gözlemlenmiştir. Her FS algoritmasının gerçekleştirdiği 96 durumdan 62 durumda hibrit FS yöntemi, 41 durumda sarmal FS yöntemi, 38 durumda gömülü FS yöntemi ve 20 durumda filtre FS yöntemi kullanılan ML modellerine göre performansı geliştirmiştir. Genellikle FS uygulanarak seçilen veriler ile yapılan tahminlerin performansı, FS uygulanmadan VS1 ve VS2 setleri ile yapılan tahminlerin performansına göre daha başarılı olabilir.

Orijinal veri setlerine kıyasla daha az özellik kullanılarak gerçekleştirilen tahminlerde hesaplama yükü azalmıştır. Böylece ML modellerinin eğitim süreleri kısalarak FS yöntemlerinin olumlu etkileri gözlemlenmiştir.

KAYNAKÇA

Alpaydın, E. 2010. Introduction to machine learning. MIT Press, 1-579.

Awad, M., Khanna, R. 2015. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Apress, 67-80.

Baraklı, B., Küçüker, A. 2018. Karar destek makineleri ve rastgele orman ağaçları yöntemleri ile vücut yağ yüzdesinin tahmini. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 6(4): 430-445.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A. 2013. A review of feature selection methods on synthetic data. Knowledge and Information Systems, 34: 483–519.

Carmona, P., Climent, F., Momparler, A. 2019. Predicting failure in the U.S. banking sector: An extreme gradient boosting approach. International Review of Economics & Finance, 61: 304-323.

Chandrashekar, G., Sahin, F. 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16-28.

Chen, T., Guestrin, C. 2016. XGBoost: A scalable tree boosting system. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, California, 785–794.

Chen, X.-w., Jeong, J. C. 2007. Enhanced recursive feature elimination. Sixth International Conference on Machine Learning and Applications, Cincinnati, 429-435.

Cherkassky, V., Ma, Y. 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks, 17(1): 113-126.

Chiong, R., Fan, Z., Hu, Z., Chiong, F. 2020. Using an improved relative error support vector machine for body fat prediction. Computer Methods and Programs in Biomedicine, 198(105749).

Csige, I., Ujvárosy, D., Szabó, Z., Lőrincz, I., Paragh, G., Harangi, M., Somodi, S. 2018. The Impact of Obesity on the Cardiovascular System. Journal of Diabetes Research, 2018(3407306): 1-12.

Darst, B. F., Malecki, K. C., Engelman, C. D. 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genetics, 19(65).

Decision Trees., https://scikit-learn.org/stable/modules/tree.html#regression., Erişim Tarihi: 01.05.2022.

Degeest, A., Frénay, B., Verleysen, M. 2021. Reading grid for feature selection relevance criteria in regression. Pattern Recognition Letters, 148: 92-99.

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., Thomas, D. M. 2018. A review of machine learning in obesity. Obesity Reviews, 19(5): 668-685.

Eberly, L. E. 2007. Multiple linear regression. İçinde: Methods in Molecular Biology, Springer, 165-187).

Edelman, C. L., Kudzma, E. C., Mandle, C. L. 2014. Health Promotion Throughout the Life Span (8th ed.). Mosby.

Emura, T., S. M., Chen, H.-Y. 2019. compound.Cox: Univariate feature selection and compound covariate for predicting survival. Computer Methods and Programs in Biomedicine, 168: 21-37.

Fan, G.-Z., Ong, S. E., Koh, H. C. 2006. Determinants of house price: A decision tree approach. Urban Studies, 43(12): 2301-2315.

Ferenci, T. 2013. Two applications of biostatistics in the analysis of pathophysiological processes. Óbuda Univeristy, Diss. PhD Thesis.

Ferenci, T., Kovács, L. 2018. Predicting body fat percentage from anthropometric and laboratory measurements using artificial neural networks. Applied Soft Computing, 67: 834-839.

Fernández-Sánchez, A., Madrigal-Santillán, E., Bautista, M., Esquivel-Soto, J., Morales-González, A., Esquivel-Chirino, C., Morales-González, J. A. 2011. Inflammation, oxidative stress, and obesity. International Journal of Molecular Sciences, 12(5): 3117-3132.

Ferri, F., Pudil, P., Hatef, M., Kittler, J. 1994. Comparative study of techniques for large-scale feature selection. Machine Intelligence and Pattern Recognition, 16: 403-413.

Fonti, V. F., Belitser, E. N. 2017. Paper in Business Analytics Feature Selection using LASSO. Vrije Universiteit Amsterdam.

Frénay, B., Doquire, G., Verleysen, M. 2013. Is mutual information adequate for feature selection in regression?. Neural Networks, 48: 1-7.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5): 1189-1232.

Géron, A. 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly, 1-1109.

Gholami, R., Fakhari, N. 2017. Chapter 27 - Support vector machine: principles, parameters, and applications. İçinde: Handbook of Neural Computation. Academic Press, 515-535.

Grömping, U. 2009. Variable importance assessment in regression: linear regression versus random forest. The American Statistician, 63(4): 308-319.

Hu, L., Huang, X., You, C., Li, J., Hong, K., Li, P., Cheng, X. 2017. Prevalence of overweight, obesity, abdominal obesity and obesity-related risk factors in southern China. PLoS One, 12(9): 1-14.

Hussain, S. A., Cavuş, N., Şekeroğlu, B. 2021. Hybrid machine learning model for body fat percentage prediction based on support vector regression and emotional artificial neural networks. Applied Sciences, 11(21): 1-16.

Huxley, R., Mendis, S., Zheleznyakov, E., Reddy, S., Chan, J. 2010. Body mass index, waist circumference and waist:hip ratio as predictors of cardiovascular risk—a review of the literature. European Journal of Clinical Nutrition, 64: 16-22.

Ito, K., Nakano, R. 2003. Optimizing support vector regression hyperparameters based on cross-validation. Proceedings of the International Joint Conference on Neural Networks, 2003, Portland, 2077-2082.

Jović, A., Brkić, K., Bogunović, N. 2015. A review of feature selection methods with applications. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 1200-1205.

Karagiannopoulos, M., Anyfantis, D. S., Kotsiantis, S., Pintelas, P. 2007. Feature selection for regression problems. Computer Science.

Keivanian, F., Chiong, R., Hu, Z. 2019. 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, 1-4.

Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N., Taylor, H. L. 1972. Indices of relative weight and obesity. Journal of Chronic Diseases, 25(6-7): 329-343.

Kraskov, A., Stögbauer, H., Grassberger, P. 2004. Estimating mutual information. American Physical Society, 69(6): 066138.

Kuhn, M., Johnson, K. 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models, Chapman and Hall/CRC, 1-310.

Kumar, V., Minz, S. 2014. Feature selection: A literature review. Smart Computing Review, 4(3): 211-229.

Kupusinac, A., Stokić, E., Sukić, E., Rankov, O., Katić, A. 2017. What kind of relationship is between body mass index and body fat percentage. Journal of Medical Systems, 41(5): 1-7.

Kursa, M. B., Rudnicki, W. R. 2010. Feature selection with the boruta package. Journal of Statistical Software, 36(11): 1-11.

Lavie, C. J., Schutter, A. D., Parto, P., Jahangir, E., Kokkinos, P., Ortega, F. B., Milani, R. V. 2016. Obesity and Prevalence of Cardiovascular Diseases and Prognosis—The Obesity Paradox Updated. Progress in Cardiovascular Diseases, 58(5): 537-547.

Leea, C. M., Huxleya, R. R., Wildmanb, R. P., Woodward, M. 2008. Indices of abdominal obesity are better discriminators of cardiovascular risk factors than BMI: a meta-analysis. Journal of Clinical Epidemiology, 61(7): 646-653.

Louppe, G. 2014. Understanding Random Forests From Theory To Practice. Liège: Faculty of Applied Sciences Department of Electrical Engineering & Computer Science.

Maulud, D. H., Abdulazeez, A. M. 2020. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4): 140-147.

McLellan, F. 2002. Obesity rising to alarming levels around the world. The Lancet, 359(9315): 1412.

Miao, J., Niu, L. 2016. A survey on feature selection. Procedia Computer Science, 91: 919-926.

Miller, A. 2002. Subset Selection in Regression, CRC Press, 1-247.

Naqa, I. E., Murphy, M. J. 2015. What is machine learning?. İçinde: Machine Learning in Radiation Oncology, Springer, 3-11.

Natekin, A., Knoll, A. 2013. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7(21): 1-21.

Ortega, F. B., Lavie, C. J., Blair, S. N. 2016. Obesity and Cardiovascular Disease. Circulation Research, 118(11): 1752-1770.

Otchere, D. A., Ganat, T. O., Ojero, J. O., Tackie-Otoo, B. N., Taki, M. Y. 2022. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. Journal of Petroleum Science and Engineering, 208(E): 1-11.

Pekel, E. 2020. Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology, 139: 1111-1119.

Random forest., https://en.wikipedia.org/wiki/Random_forest., Erişim Tarihi: 25.03.2022.

Raschka, S., SequentialFeatureSelector: The popular forward and backward feature selection approaches incl. floating variants. http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/., Erişim Tarihi: 21.03.2022.

Saeys, Y., Abeel, T., Peer, Y. V. 2008. Robust feature selection using ensemble feature selection techniques. İçinde: Lecture Notes in Computer Science, Springer.

Sagi, O., Rokach, L. 2021. Approximating XGBoost with an interpretable decision tree. Information Sciences, 572: 522-542.

Segal, M. R. 2004. Machine learning benchmarks and random forest regression. UCSF: Center for Bioinformatics and Molecular Biostatistics.

Shao, Y. E. 2014. Body fat percentage prediction using intelligent hybrid approaches. The Scientific World Journal, 2014(383910): 1-8.

Smola, A. J., Schölkopf, B. 2004. A tutorial on support vector regression. Statistics and Computing, 14: 199-222.

Srdić, B., Obradović, B., Dimitrić, G., Stokić, E., Babović, S. S. 2012. Relationship between body mass index and body fat in children-Age and gender differences. Obesity Research & Clinical Practice, 6(2): 167-173.

Stoean, R., Dumitrescu, D., Preuss, M., Stoean, C. 2006. Evolutionary support vector regression machines. 2006 Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, 330-335.

Stokić, E., Kupusinac, A., Doroslovački, R. 2014. Predicting body fat percentage based on gender, age and BMI by using artificial neural networks. Computer Methods and Programs in Biomedicine, 113(2): 610-619.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8(25).

Subho, M. R., Chowdhury, M. R., Chaki, D., Islam, S., Rahman, M. M. 2019. A univariate feature selection approach for finding key factors of restaurant business. 2019 IEEE Region 10 Symposium (TENSYMP), Kolkata, 605-610.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P. 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. The Journal for Chemical Information and Computer Scientists, 43: 1947-1958.

Uçar, M. K., Uçar, Z., Köksal, F., Daldal, N. 2021. Estimation of body fat percentage using hybrid machine learning algorithms. Measurement, 167.

Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. New Jersey: Springer.

Zhang, Y., Haghani, A. 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58(B): 308-324.

ÖZGEÇMİŞ

Adı Soyadı **:** Asude ALTIPARMAK BİLGİN

ÖĞRENİM DURUMU

Derece	**Eğitim Birimi**	**Mezuniyet Yılı**
Yüksek Lisans	Sakarya Üniversitesi / Fen Bilimleri Enstitüsü / Elektrik Elektronik Mühendisliği	Devam ediyor
Lisans	Sakarya Üniversitesi / Mühendislik Fakültesi / Elektrik Elektronik Mühendisliği	2017
Lise	Balıkesir Muharrem Hasbi Anadolu Lisesi	2013

İŞ DENEYİMİ

Yıl	**Yer**	**Görev**
2021-Halen	İstanbul Medeniyet Üniversitesi	Araştırma Görevlisi
2018-2019	Royalcert Belgelendirme ve Gözetim Hizmetleri A.Ş.	Elektrik-Elektronik Mühendisi

YABANCI DİL

İngilizce