T.C.
BURSA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ÖZNİTELİK SEÇİM YÖNTEMLERİNDEKİ YANLILIK ETKİSİNİN SINIFLANDIRMA BAŞARISI AÇISINDAN DEĞERLENDİRİLMESİ

YÜKSEK LİSANS TEZİ

Semih Can BOZOK

Bilgisayar Mühendisliği Anabilim Dalı

NİSAN 2023

T.C.
BURSA TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

ÖZNİTELİK SEÇİM YÖNTEMLERİNDEKİ YANLILIK ETKİSİNİN SINIFLANDIRMA BAŞARISI AÇISINDAN DEĞERLENDİRİLMESİ

YÜKSEK LİSANS TEZİ

Semih Can BOZOK
(19376482007)

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Ergün GÜMÜŞ

NİSAN 2023

BTÜ, Lisansüstü Eğitim Enstitüsü'nün 19376482007 numaralı Yüksek Lisans Öğrencisi Semih Can BOZOK, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "ÖZNİTELİK SEÇİM YÖNTEMLERİNDEKİ YANLILIK ETKİSİNİN SINIFLANDIRMA BAŞARISI AÇISINDAN DEĞERLENDİRİLMESİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Dr. Öğr. Üyesi Ergün GÜMÜŞ** .....
Bursa Teknik Üniversitesi

Jüri Üyeleri : **Doç. Dr. Can EYÜPOĞLU** .....
Milli Savunma Üniversitesi

Dr. Öğr. Üyesi Seçkin YILMAZ .....
Bursa Teknik Üniversitesi

Teslim Tarihi : Mayıs 2023
Savunma Tarihi : 26 Nisan 2023

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, Bursa Teknik Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

İNTİHAL BEYANI

Bu tezde görsel, işitsel ve yazılı biçimde sunulan tüm bilgi ve sonuçların akademik ve etik kurallara uyularak tarafımdan elde edildiğini, tez içinde yer alan ancak bu çalışmaya özgü olmayan tüm sonuç ve bilgileri tezde kaynak göstererek belgelediğimi, aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

Öğrencinin Adı Soyadı: Semih Can BOZOK

İmzası:

Aileme,

ÖNSÖZ

Tez çalışmasında sonsuz bir şekilde bana yardımcı olan, desteğini hiçbir zaman esirgemeyen, çalışmada bana yol gösterici olan değerli tez danışmanım Dr. Öğr. Üyesi Ergün GÜMÜŞ Hocama teşekkürlerimi sunmaktayım. Bursa Teknik Üniversitesinin yüksek lisans öğrencileri için verdiği desteklere ve söz konusu destekleri öğrencilere ulaştırmak için çalışan tüm birimlerine de teşekkürlerimi sunmaktayım.

Nisan 2023

Semih Can BOZOK
(Araştırma Görevlisi)

İÇİNDEKİLER

Sayfa

KISALTMALAR

ANOVA : Analysis of Variance
ASM : Adjusted Stability Measure
BMI : Body Mass Index
CV : Cross Validation
ÇG : Çapraz Geçerleme
DVM : Destek Vektör Makineleri
LRC : Logistic Regression Classifier
LRS : Lojistik Regresyon Sınıflandırıcısı
mRMR : Minimum Redundancy Maximum Relevance
RFC : Random Forest Classifier
RFE : Recursive Feature Elimination
ROC : Receiver Operating Characteristic
ROS : Rastgele Orman Sınıflandırıcısı
SVM : Support Vector Machines
TmRMR : Temporal Minimum Redundancy Maximum Relevance
USM : Unadjusted Stability Measure

SEMBOLLER

| |	**:** Kümedeki eleman sayısını veren operatör
∩	**:** Küme kesişimi operatörü
∪	**:** Küme birleşimi operatörü
c	**:** Elde edilen öznitelik alt kümelerinin sayısı
C	**:** Hedef sınıf
D	**:** İlgililik
e	**:** Üstel notasyon
$\mathbf{F^{ki}}$, $\mathbf{F^{kj}}$	**:** Öznitelik alt kümeleri
I	**:** Karşılıklı bilgi
i, j	**:** Sayaç değişkenleri
$\mathbf{k_i}$, $\mathbf{k_j}$	**:** Öznitelik alt küme çiftlerinin eleman sayısı
max	**:** Maksimize işlemi
min	**:** Minimize işlemi
n	**:** Toplam öznitelik sayısı
r	**:** İki öznitelik alt kümesi içinde aynı olan özniteliklerin sayısı
R	**:** Tekrarlama
S	**:** Öznitelik seti
$\mathbf{S_A}$, $\mathbf{S_S}$	**:** Öznitelik kümelerindeki benzerlik
SE	**:** Standart hata
$\mathbf{s_i}$, $\mathbf{s_j}$	**:** Öznitelik alt kümeleri
$\mathbf{x_i}$, $\mathbf{x_j}$	**:** Öznitelik setindeki öznitelik çiftleri
∈	**:** Kümenin elemanıdır operatörü
ϕ	**:** Optimize edip birleştirme operatörü

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

ÖZNİTELİK SEÇİM YÖNTEMLERİNDEKİ YANLILIK ETKİSİNİN SINIFLANDIRMA BAŞARISI AÇISINDAN DEĞERLENDİRİLMESİ

ÖZET

Günümüz dünyasında veri her yerde, bol bir şekilde, rahatlıkla ulaşılabilir bir haldedir. Veri bol, elde etmesi kolay ama sürekli olarak artan bir yapıda olması nedeniyle işlenmesi, anlamlı hale getirilmesi giderek zorlaşmaktadır. Özellikle büyük veri çalışmaları, görüntü tabanlı çalışmalar, veri akışı tabanlı çalışmalarda özniteliklerin anlamlı alt kümeler şeklinde azaltılması önem kazanmaktadır. Öznitelik seçimi yapılmazsa, işlemci gücü yoğun bir şekilde kullanılmakta, sınıflandırıcıların eğitim süresi uzamakta ve bu durum da bazı verileri işlenemez hale getirmektedir. Makine öğrenmesinde öznitelik seçimi günümüzde çok ilgi gören bir çalışma alanıdır. Öznitelik seçimi verideki özniteliklerin sayısını azaltarak boyutsallık lanetinden (curse of dimensionality) kaçınmayı amaçlamaktadır. Bu amaç için veriyi çeşitli yaklaşımlarla incelemeye alır, çeşitli karar verme mekanizmaları kullanarak en anlamlı olan öznitelikleri seçer.

Öznitelik seçimi yapılırken verinin doğasında bulunan etkiler nedeniyle yanlılık etkisi oluşabilmektedir. Yanlılık etkisi öznitelik seçimini olumsuz yönde etkilemektedir. Öznitelik seçiminde önemli konulardan biri de kullandığımız eğitim, geçerleme (validation) ve test kümesinin iterasyonlar bazlı değişiminin yanlılık etkisi ortaya çıkarmasıdır. Örnek sayısı ile seçilen özniteliklerin değişimi arasındaki ilişki de önemli bir konudur. Örnek sayısının fazla olduğu durumlarda öznitelik seçimi yaptığımızda her seferinde benzer öznitelik alt kümesinin seçilmesi beklenmektedir. Yanlılık etkisinden kaçınmak için çeşitli çapraz geçerleme yöntemleri kullanmak etkiyi azaltma yönünde olumlu bir durum oluşturmaktadır. Veriyi farklı çapraz geçerleme yöntemleri kullanarak öznitelik seçimine sokmamız farklı benzerlik metriği oranı vererek yanlılık etkisinin hangi yöntemde daha az olduğu hakkında bize bilgi vermektedir. Bu konuyla ilgili araştırmalar yoğun bir ilgiyle sürmektedir.

Tez çalışmasında üç farklı veri seti ve üç farklı öznitelik seçim yöntemi kullanılarak öznitelik seçimi yapılmıştır. Söz konusu öznitelik seçim yöntemleri de üç farklı çapraz geçerleme yöntemi ve üç farklı sınıflandırıcı ile çalıştırılmıştır. Bu sayede seksen bir farklı çalıştırma yapılmıştır. Yapılan çalışmalar için iki farklı benzerlik metriği kullanılarak yanlılık etkisi gözlemlenmiştir. Elde edilen sonuçlara göre veri setinden ve öznitelik seçim yönteminden bağımsız olarak yanlılık etkisinin en az olduğu çapraz geçerleme yöntemi tespit edilmiştir.

Anahtar kelimeler: Öznitelik Seçimi, Yanlılık Etkisi, Çapraz Geçerleme.

EVALUATION OF THE BIAS EFFECT IN FEATURE SELECTION METHODS IN TERMS OF CLASSIFICATION ACCURACY

SUMMARY

In today's world, data is everywhere, abundant and easily accessible. Data is abundant, easy to obtain, but due to its continuously increasing structure, it is becoming increasingly difficult to process and make it meaningful. Especially in big data studies, image-based studies, data stream-based studies, it is important to reduce attributes into meaningful subsets. Without feature selection, processing power is used intensively, the training time of classifiers is prolonged and this makes some data unprocessable. Feature selection in machine learning is a field of study that has received much attention. Feature selection aims to avoid the curse of dimensionality by reducing the number of features in the data. For this purpose, it examines the data with various approaches and selects the most meaningful attributes using various decision-making mechanisms.

When selecting attributes, a bias effect may occur due to the inherent effects of the data. The bias effect negatively affects attribute selection. One of the important issues in feature selection is that the iterative change of the training, validation and test set we use can introduce bias effects. The relationship between the number of samples and the variation of the selected attributes is also an important issue. When we select attributes when the number of instances is large, we expect a similar subset of attributes to be selected each time. In order to avoid the bias effect, using various cross validation methods is a positive way to reduce the effect. Using different cross validation methods for feature selection gives us different similarity metric ratios and gives us information about which method has less bias effect. Research on this topic continues with intense interest.

In this thesis, three different datasets and three different feature selection methods were used for feature selection. These feature selection methods were also run with three different cross validation methods and three different classifiers. In this way, eighty-one different runs were performed. Bias effect was observed by using two different similarity metrics for the studies. According to the results obtained, the cross validation method with the least bias effect was determined independently of the dataset and feature selection method.

Keywords: Feature Selection, Bias Effect, Cross Validation.

1. GİRİŞ

Öznitelik seçimi, makine öğrenmesi için kullanılan verilerdeki en anlamlı öznitelikleri bulmak için kullanılmaktadır. Öznitelik, analiz etmeye çalıştığımız veride bulunan ölçümlenebilir olan en küçük birimdir. Söz konusu en küçük birim literatürde bulunan veri setlerinde sütunlara karşılık gelmektedir. Veride bulunan öznitelikler veriyi elde ederken dikkat ettiğimiz noktalardır ancak, veriyi elde ederken bu özniteliklerin hangilerinin en anlamlı olduklarını bilemeyebiliriz. Özellikle büyük verilerde, görüntü tabanlı verilerde, gen ifadelerinin bulunduğu verilerde öznitelik sayısı çok fazladır.

Sınıflandırma problemlerinde çok fazla sayıda olan özniteliklerin, anlamlı alt kümeler şeklinde elde edilmesi önemli bir çalışma alanı olmuştur. Daha anlamlı alt kümeler elde etmek veriyi daha iyi anlamlandırmaya, öğrenim süresini kısaltmaya, boyutsallık lanetinden (curse of dimensionality) kaçınmaya yardımcı olmaktadır.

Öznitelik seçimi için çeşitli algoritmalar bulunmaktadır. Söz konusu algoritmalar, istatistiki yöntemleri baz alarak çalışanlar, tahminsel yaklaşım uygulayarak yinelemeli şekilde çalışanlar ve bu iki yaklaşımı baz alarak melez olarak türetilmiş şekilde çalışanlar olarak üçe ayrılmaktadır.

Öznitelik seçimi için söz konusu yöntemlerden birini kullanırken yapılan seçimlerde "yanlılık etkisi" (bias effect) denen bir fenomen oluşmaktadır. Yanlılık etkisi, bazı özniteliklerin sınıfla yapay olarak ilişkilerinin yüksek olduğuna bu durumun da verinin doğasına göre o özniteliklerin seçilmesine neden olabileceği şeklinde tanımlanmaktadır [1].

Yanlılık etkisinden kaçınmak için öznitelik seçimi yapmadan önce veriyi rastgele hale getirmemize yardımcı olan farklı çapraz geçerleme yöntemlerinin denenmesi önerilmiştir [2].

1.1 Tezin Amacı

Tezin amacı literatürde sıklıkla karşımıza çıkan çeşitli veri setleri üzerinde öznitelik seçimi yapmak ve seçim yaparken oluşan yanlılık etkisini azaltmak için önerilen çeşitli

çapraz geçerleme yöntemlerini kullanmaktır. Amacın can alıcı noktası, kullanılan çeşitli çapraz geçerleme yöntemlerini de kendi aralarında yanlılık etkisi bakımından incelemek ve hangisinin daha kararlı bir yapıda olduğunu bulmaktır. Farklı veri seti kullanılmasındaki amaç verinin doğasına göre oluşan yanlılık etkisinin her bir veri seti üzerinde farklı oluştuğunu görebilmemiz için yapılmıştır.

1.2 Literatür Araştırması

Öznitelik seçimi ve yanlılık etkisi ile ilgili yapılan literatür araştırmasında çok sayıda araştırmacı tarafından yapılmış çalışmalar bulunmaktadır. Bu bölümde söz konusu çalışmaların detayları açıklanmıştır.

Singhi ve Liu sınıflandırma problemleri üzerine öznitelik seçimi kullanmışlardır. Sınıflandırma başarımında yanlılık etkisini çeşitli istatistiksel yöntemlere göre analiz etmişler ve etki eden faktörlerin neler olduklarını bulmuşlardır [1].

Krawczuk ve Łukaszuk gen verisi üzerinde sınıflandırma başarımını yükseltmek için öznitelik seçimi kullanmışlardır. Dört farklı öznitelik seçim yöntemini yüksek boyutlu yedi ayrı veri seti üzerinde denemişler ve yirmi sekiz ayrı senaryoda yanlılık etkisinin %2,6 ile %41,67 arasında değiştiğini gözlemlemişlerdir. Yanlılık etkisini azaltmak için farklı çapraz geçerleme yöntemlerinin kullanılmasını önermişlerdir [2].

Maggipinto ve arkadaşları, Alzaymır hastalığı üzerine elde ettikleri bir veride sınıflandırma başarımını yükseltmek için öznitelik seçimi kullanmışlardır. Sınıflandırma başarımında yanlılık etkisinin ROC eğrisi altında kalan alan bazlı olarak %10 ile %30 göreceli oranda değiştiğini bulmuşlardır [3].

Markowetz ve Spang, gen ifadeleri üzerine sınıflandırma yaparken öznitelik seçimi kullanmışlardır. Sınıflandırma başarımında yanlılık etkisini, tasarlamış oldukları iç döngülü çapraz geçerleme ile azaltmışlardır. Tasarladıkları yöntemin diğer yöntemlerle olan farklılıklarını ortaya koymuşlardır [4].

Park ve Kwon, metin sınıflandırma problemleri üzerinde yanlılık etkisini düzeltmek için Gini Index algoritmasını geliştirip yeni bir algoritma tasarlamışlardır. Tasarladıkları algoritmanın sınıflandırma başarımında ortalama olarak %20 iyileştirme yaptığını bulmuşlardır [5].

Demircioğlu, medikal görüntüler üzerinde en anlamlı öznitelikleri bulmak için öznitelik seçim yöntemi kullanmıştır. Öznitelik seçimi yapmadan önce çapraz geçerleme yapılmazsa veri sızıntısı olduğunu bunun da yanlılık etkisi olarak ortaya çıkacağını gözlemlemiştir. Literatürde bulunan on farklı medikal görüntü veri seti üzerinde iki farklı deney gerçekleştirmiştir. İlk deneyde çapraz geçerlemeden önce öznitelik seçimi yanlış uygulanmış, sonrasında her bir kat için çapraz geçerleme içinde öznitelik seçimi doğru bir şekilde uygulanarak yapılmıştır. İki deney arasında yanlılık etkisi, kesinlik (accuracy) metriği olarak %17 düzeylerine kadar çıkmıştır [6].

Li ve arkadaşları, öznitelikler ve sınıf için bir Bayes modeli kullanıldığında yanlılık etkisinin nasıl önlenebileceğini araştırmışlardır. Seçilmemiş özniteliklerin tam değerleri unutulsa bile sınıfla olan korelasyonlarının seçilmeleri için çok küçük olduğu bilgisini korumak gerektiği ve bu durumu koşullandırmanın iyi olacağını bulmuşlardır. Çalışmalarında bu fikrin ikili sınıflı verilerde Naive Bayes yöntemi için nasıl uygulanabileceğini göstermişlerdir. Yöntemi, gen ifadesini kolon kanseriyle ilişkilendiren veri alt kümelerine uygulamışlar ve öznitelik seçiminden kaynaklanan yanlılığı düzeltmenin tahmin performansını artırdığını görmüşlerdir [7].

Munson ve Caruana, on dokuz farklı veri seti üzerinde öznitelik seçimi yaparken tek ve torbalanmış karar ağacı (bagged decision trees) sınıflandırıcısı ikilisi kullanarak yanlılık analizi yapmışlardır. Torbalama (bagging) kullanıldığında yanlılık etkisinin azaldığını saptamışlardır [8].

Tran ve arkadaşları, öznitelik seçimi yanlılığının olup olmadığıyla ilgili olarak on farklı veri setinde çapraz geçerleme yapılmadan ve çapraz geçerleme yapılarak Parçacık Sürü Optimizasyonu (Particle Swarm Optimisation) ile öznitelik seçimi yapmışlardır. Parçacık Sürü Optimizasyonunda yerel optimuma takılma sorununu da algoritmayı değiştirip yeni bir optimizasyon algoritması önererek çözmüşlerdir. Kullandıkları 10 Kat Çapraz Geçerleme yöntemi sayesinde yanlılık etkisinin aşırı uyum (overfitting) sorununa neden olmasını engellemişlerdir [9].

Ni ve arkadaşları, öznitelik seçiminde yanlılığı gidermek için önerilen çapraz geçerleme, çift döngülü çapraz geçerleme yöntemlerinin küçük örneklemli verilerde kararlılığı sağlamadığını gözlemlemişlerdir. Çapraz geçerleme kullanmadan, modelden bağımsız olarak Minimum Beklenen Yanlış Sınıflandırma Maliyeti (Minimum Expected Cost of Misclassification) isimli bir yaklaşım önermişlerdir.

Yaptıkları çalışmada 10 katlı çift döngülü çapraz doğrulama tabanlı sarmal (wrapper) öznitelik seçim metodunda gürültüsüz hata oranından yaklaşık %33 daha yüksek hata oranına sahip olduğunu ve 10 katın tamamında ayırt edici özellikleri tutarlı bir şekilde tanımlayamadığını gösterdiğini bulmuşlardır [10].

Ambroise ve McLachlan, kanser hastalığı için binlerce genin ifadesini içeren bir veri seti üzerinde öznitelik seçimi yapmışlardır. Yaptıkları çalışmada 10 Kat Çapraz Geçerleme kullanıldığında Birini Dışarıda Bırak (Leave-One-Out) Çapraz Geçerlemeye göre daha az yanlılık etkisi gözlemlemişlerdir. Daha önce yapılan başka bir çalışmada da çapraz geçerleme kullanılmadığı için hatalı seçim yapıldığını da açıklamışlardır [11].

Wood ve arkadaşları, gen ifadeleri üzerine öznitelik seçimi yapmışlardır. Normal çapraz geçerleme ve iki seviyeli çapraz geçerleme kullanarak ikisi arasında yanlılık etkisinin değişimini gözlemlemişlerdir. Kullandıkları veri setinde %3 ile %5 arasında değişim olduğunu raporlamışlardır [12].

1.3 Hipotez

Öznitelik seçimi yaparken verinin doğasında bulunan nedenlerden dolayı yanlılık etkisinin oluşması beklenmektedir. Tez çalışmasında yanlılık etkisini azaltmak için literatürde önerilen çeşitli çapraz geçerleme yöntemleri kullanılacaktır. Söz konusu çapraz geçerleme yöntemlerinin kendi aralarında yanlılık etkisi açısından farklılıklar oluşturması beklenmektedir. Bu yöntemlerden hangisinin öznitelik seçimi bakımından daha kararlı bir yapıda olduğu ve yöntemin, kullanılan veri setiyle ilişkisi incelenecektir.

2. MAKİNE ÖĞRENMESİ

Makine öğrenmesi, deneyim yoluyla kendini geliştiren makinelerin (bilgisayarların) nasıl oluşturulacağı kısmında çalışmalar yapan bir disiplindir [13]. Makine öğrenmesi, makineyi verimli haliyle kullanmak için çeşitli yöntemlerle onu optimize etmeye çalışmak olarak da tanımlanabilir.

Makine öğrenmesi yapılırken ya geçmiş deneyimlerle makineyi besleriz ya da bir miktar veriyi makineye veririz. Makine bunlara bakarak çeşitli istatistiksel, cebirsel hesaplamalar yapar ve bir model elde eder. Bu modeli de kullanarak benzer özellikte olan bir miktar veriyi tekrar girdi olarak verdiğimizde model, belirli bir oranda tahminde bulunur, verilen yeni girdiyi anlamlandırmaya çalışır.

Makineye verilen veriler sonucunda makine bir çıkarımda bulunur. Bu çıkarıma "model" ismi verilmektedir. Modeli elde etmek de makine öğrenmesinin "öğrenme" kısmı olmaktadır. Bilgisayar biliminde kullanılan klasik yöntemlerde, "girdi" verilir, "model" kullanıcı tarafından oluşturulur ve sonuç olarak "çıktı" beklenir. Makine öğrenmesinde hem "girdi" verilir hem "çıktı" verilir, sonuç olarak makinenin bunlara bakıp bir model oluşturması beklenir. Model elde edildikten sonra bu modeli kullanarak yeni gelen veriler üzerinden tahminler yapılır.

Makine öğrenmesi sınıflandırma, öbekleme, bağlanım, boyut azaltma, öznitelik seçimi, korelasyon analizi konularını kapsar. Tez çalışmasında öznitelik seçimi, öznitelik seçiminde oluşan yanlılık etkisi ve yanlılık etkisini gidermek için önerilen çapraz geçerleme yöntemlerinin detayları üzerinde çalışma yapılmıştır.

2.1 Sınıflandırma

Sınıflandırma, modele verilen etiketli eğitim verisine karşılık olarak kendisine verilen test verisini tahminlemesidir. Model girdilere bakarak öğrenir, girdideki öznitelikler arasındaki ilişkiyi anlamlandırır ve test etmek için verilen yeni girdileri tahminlemek amacıyla elde ettiği çıkarımlara bakar. Veriyi kategorize ederek girdi haline getirmek gerekmektedir. Aynı şekilde de tahminlemeyi de kategorik olarak yapar.

Sınıflandırma genel olarak ikiye ayrılır; iki sınıflı sınıflandırıcı, ikiden çok sınıflı sınıflandırıcı. İkili sınıflandırmada verideki sınıf etiketi iki kategoridedir. Tez çalışmasında kullanılan veriler bu sınıflandırma problemine girmektedir. Çünkü kullanılan üç veri setinde de sınıf etiketi iki ayrı kategoriden oluşmaktadır. Çok sınıflı sınıflandırıcı, bir sınıflandırma probleminde verideki sınıf etiketi iki kategoriden fazla olursa bu durum çok sınıflı sınıflandırma problemidir şeklinde tanımlanabilir.

Literatürde çok sayıda sınıflandırıcı bulunmaktadır. Sınıflandırıcılar verinin doğasına göre seçilerek kullanılmaktadır. Sınıflandırıcılar çeşitli matematiksel yaklaşımlar baz alınarak geliştirilmiştir. Sigmoid fonksiyonunu baz alarak çalışan, Bayes teoremini baz alarak çalışan, uzayda verinin konumlarına bakarak en yakın komşulara göre karar verme bazlı çalışan, hiyerarşik yaklaşım kullanarak şartlar şeklinde bir nevi ağaç dalları oluşturma bazlı olarak çalışan yöntemler ve daha birçok yöntem mevcuttur.

Tez çalışmasında literatürde çok kez karşımıza çıkan üç sınıflandırıcı kullanılmıştır. Söz konusu sınıflandırıcılar, öznitelik seçiminde elde edilen öznitelik alt kümesinin sınıflandırma başarımını elde etme bakımından kullanılmıştır. Ayrıca söz konusu sınıflandırıcılar tez çalışmasında kullanılan öznitelik seçim yöntemleriyle uyumlu parametrelerle çalıştıkları için de aşağıda detayları anlatılan üç sınıflandırıcı seçilmiştir.

2.1.1 Destek vektör makineleri

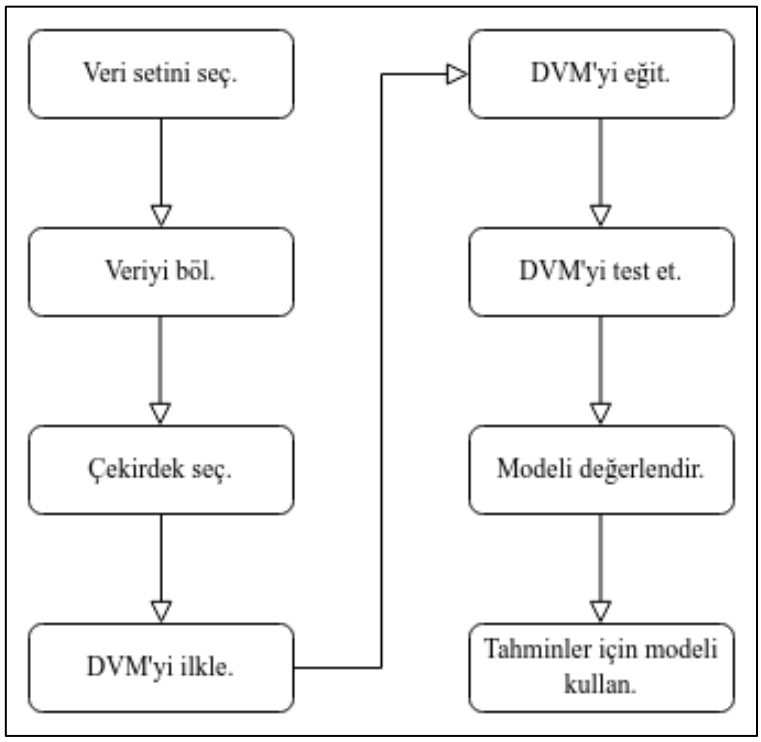
Destek Vektör Makineleri literatürde “Support Vector Machines, SVM” şeklinde karşımıza çıkmaktadır ve iki gruplu sınıflandırma problemleri için 1995 yılında önerilen bir algoritmadır [14]. Algoritma, genel olarak iki sınıf, uzaya yerleştirildiğinde sınıftaki örnek kümelerinin arasındaki mesafeyi maksimize etme mantığıyla çalışmaktadır [15].

Destek Vektör Makineleri (DVM) literatürde görüntü tanıma, metin sınıflandırma ve biyoinformatik dahil olmak üzere çeşitli alanlarda yaygın olarak incelenmiş ve kullanılmıştır.

Tez çalışmasında üç farklı veri seti üzerinde ve üç farklı öznitelik seçiminde DVM kullanılmıştır. DVM kullanılırken parametre olarak çekirdek (kernel) parametresi, düzenleme (regularization) parametresi, durdurma toleransı parametresi ve maksimum iterasyon parametresi veriye uygun şekilde ayarlanarak kullanılmıştır. Ayarlanan söz

konusu dört parametre hem üç farklı öznitelik seçiminde hem de üç farklı veri setinde de uygunluğu test edilerek belirlenmiştir.

DVM'nin çalışma mantığını daha iyi anlamak için ilgili akış diyagramı aşağıda verilmiştir (Şekil 2.1).



Şekil 2.1 : Destek Vektör Makineleri akış diyagramı.

2.1.2 Rastgele orman sınıflandırıcısı

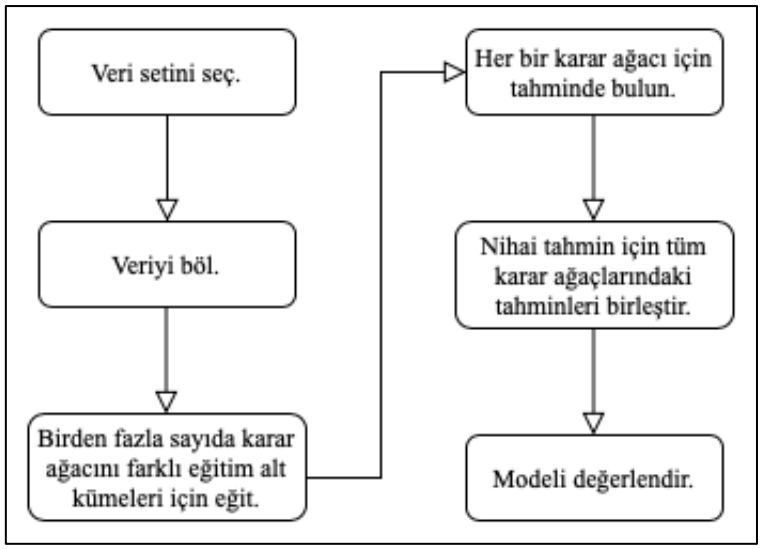
Rastgele Orman Sınıflandırıcısı literatürde "Random Forest Classifier, RFC" şeklinde karşımıza çıkmaktadır. Rastgele Orman Sınıflandırıcısı (ROS), sınıflandırma için kullanılan popüler bir makine öğrenimi algoritmasıdır. Algoritma tahminde bulunmak amacıyla çok fazla sayıda karar ağacı oluşturarak çalışmaktadır [16].

ROS, karar ağacı sınıflandırıcısının geliştirilmiş halidir. ROS, çok sayıda karar ağacı oluşturmaktadır. Oluşturulan çok sayıda karar ağacı, verideki özniteliklerin rastgele alt kümeleri şeklinde düğümlerden oluşmaktadır. ROS'da verideki örnekler yerine koyma (replacement) yapılarak kullanılmaktadır. Düğümler rastgele olarak bölümlere ayrılmaz, içlerinde en anlamlı olan düğüm ayrımları seçilerek ilerleme yapılmaktadır.

Tez çalışmasında üç farklı veri seti üzerinde ve üç farklı öznitelik seçiminde ROS kullanılmıştır. ROS kullanılırken parametre olarak aldığı oluşturacağı karar ağacı sayısı, oluşturacağı karar ağacında kaç öznitelik olacağı, yapraklardaki minimum

örnek sayısı, yeni düğümler için gerekli olan örnek sayısı gibi parametreler veriye uygun şekilde ayarlanarak kullanılmıştır. Ayarlanan söz konusu parametreler hem üç farklı öznitelik seçiminde hem de üç farklı veri setinde de uygunluğu test edilerek belirlenmiştir.

ROS'un çalışma mantığını daha iyi anlamak için ilgili akış diyagramı aşağıda verilmiştir (Şekil 2.2).



Şekil 2.2 : Rastgele Orman Sınıflandırıcısı akış diyagramı.

2.1.3 Lojistik regresyon sınıflandırıcısı

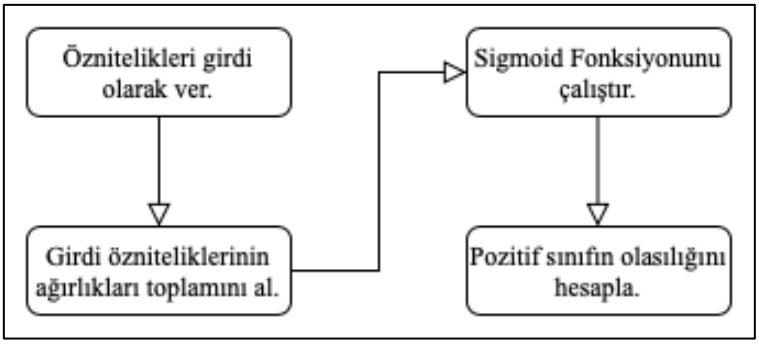
Lojistik Regresyon Sınıflandırıcısı literatürde "Logistic Regression Classifier, LRC" şeklinde karşımıza çıkmaktadır. Lojistik Regresyon Sınıflandırıcısı (LRS), sınıflandırma yapmak için iki olası sonuçtan birini tahminleyerek çalışmaktadır. LRS, pozitif sonucun (örneğin doğru veya evet) olasılığına bakarak girdideki özniteliklerin fonksiyonunu oluşturmayla modelleme yapar [17].

Lojistik regresyon modelinin çıktısı, bir karar eşiğine (örneğin 0,5) bağlı olarak yeni örnekleri sınıflandırmak için kullanılabilen 0 ile 1 arasında değişen bir olasılık puanıdır. Model, gerçek sınıf etiketlerinin bilindiği bir dizi etiketli örnek kullanılarak eğitilir ve buradaki amaç, gözlemlenen verilerin olasılığını en üst düzeye çıkaran model parametreleri kümesini bulmaktır [18].

Tez çalışmasında üç farklı veri seti üzerinde ve üç farklı öznitelik seçiminde LRS kullanılmıştır. LRS kullanılırken parametre olarak aldığı maksimum iterasyon sayısı, eğitim kısmında modeli optimize etmek için kullanacağı çözücü, durdurma kriteri toleransı, aşırı uyumdan kaçınmak için alacağı düzenleme (regularization) parametresi

gibi parametreler veriye uygun şekilde ayarlanarak kullanılmıştır. Ayarlanan söz konusu parametreler hem üç farklı öznitelik seçiminde hem de üç farklı veri setinde de uygunluğu test edilerek belirlenmiştir.

LRS'nin çalışma mantığını daha iyi anlamak için ilgili akış diyagramı aşağıda verilmiştir (Şekil 2.3).



Şekil 2.3 : Lojistik Regresyon Sınıflandırıcısı akış diyagramı.

3. ÖZNİTELİK SEÇİMİ

Veri madenciliği ve makine öğrenmesinde yüksek boyutlu verinin analizi zor tamamlanan bir alan olmuştur. Öznitelik seçimi verideki yararı az olan, bağı az olan öznitelikler ile en anlamlı olan öznitelikleri ayrıştırır, en anlamlı olan öznitelikleri kullanarak söz konusu problemi çözmeyi kolaylaştırıp etkili bir yol sunar. Bu yolu sunarken hesaplama zamanını azaltır, öğrenme oranını artırır, öğrenme modelinin veya öğrenme verisinin daha anlaşılır olmasını sağlar [19].

Öznitelik seçiminin kullanım alanı çok geniş kapsamdadır. Bunlar; görüntü tanıma, metin madenciliği, ihlal tespiti, biyoinformatik tabanlı veri analizi, hata tespiti vb. şeklinde örneklendirilebilir [19].

Öğrenme metoduyla olan ilişkisine göre öznitelik seçim metotları filtre yöntemler, sarmal yöntemler ve hibrit yöntemler şeklinde sınıflandırılmaktadır.

3.1 Filtre Yöntemler

Filtre yöntemlerde (Filter methods) özniteliklerin anlamlı alt kümeleri oluşturulurken sınıfla olan ilişkileri dikkate alınır. Burada örnek olarak her bir özniteliğin sınıfla olan korelasyonunun hesaplanması şeklinde de düşünebiliriz. Sarmal yöntemlere kıyasla filtre yöntemler daha hızlı çalışır ve filtre yöntemlerin hesaplama maliyeti daha düşüktür. Filtre yöntemlerde ekseriyetle istatistiksel dağılımlar kullanılmaktadır. Filtre yöntemlerin temellerine örnek olarak bilgi kazancı (information gain), ki-kare testi (chi-square test), Fisher skoru verilmektedir.

3.1.1 ANOVA filtre öznitelik seçim yöntemi

ANOVA Filtre Öznitelik Seçim Yöntemi, sınıflandırma problemlerinde ilgili özniteliklerin seçilmesi için kullanılan popüler yöntemlerden biridir. İstatistiksel temeli olan bir öznitelik seçim yöntemidir. Her sınıf için öznitelik değerlerinin ortalamaları hesaplanmakta, hesaplanan ortalamaların farkı arasında bir ölçüte (F-İstatistik, F-Statistic) göre karar vermektedir. Yüksek F-İstatistiğe sahip özniteliklerin sınıflandırma için daha önemli olduğu kabul edilmektedir [20].

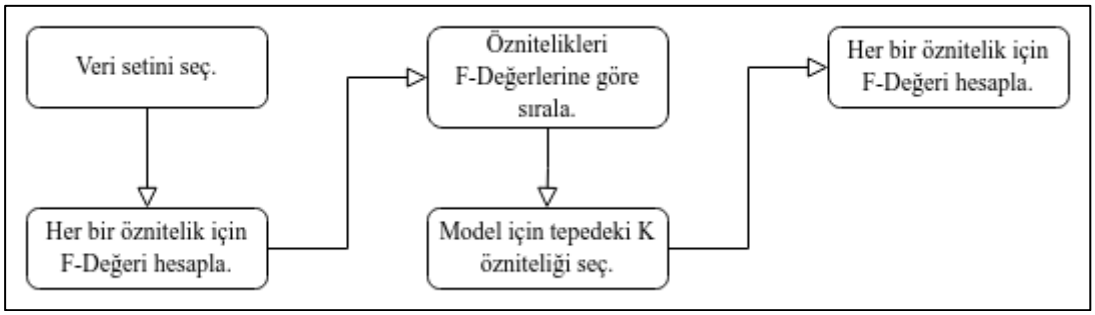
Ding ve arkadaşları bakterilerin özelleşmiş bir proteini üzerinde özniteliklerin belirlenmesi için ANOVA'yı özel bir öznitelik seçim yöntemiyle birlikte kullanmışlardır. ANOVA'nın ilgili öznitelikleri seçtiğini ve sınıflandırma başarımını artırdığını görmüşlerdir [20].

Johnson ve Synovec jet yakıtı karışımlarının sınıflandırılması üzerinde özniteliklerin belirlenmesi için ANOVA Filtre Öznitelik Seçim Yöntemini kullanmışlardır. İki farklı jet yakıtı arasındaki %1'lik farkı ifade edebilecek olan öznitelikleri seçtiğini görmüşlerdir [21].

Sheikhan ve arkadaşları, ses tanıma sistemlerinde duygu durumlarını Modüler Yapay Destek Vektör Makineleri (Modular Neural Support Vector Machines) adını verdikleri sınıflandırıcıda kullanmak amacıyla ANOVA kullanıp öznitelik seçimi yaptırmışlardır. Çalışma sonucunda özniteliklerin %22'sinin atılmasıyla bile sınıflandırma başarımında %2,2 oranında iyileştirme görmüşlerdir [22].

Elssied ve arkadaşları, spam e-postaları sınıflandırmak için kullandıkları Destek Vektör Makineleri sınıflandırıcısına öznitelikleri vermeden önce ANOVA kullanarak öznitelik seçimi yaptırmışlardır. Çalışma sonucunda sınıflandırma başarımında %1'lik iyileştirme görmüşlerdir [23].

ANOVA Filtre Öznitelik Seçim Yönteminin çalışma mantığını daha iyi anlamak için ilgili akış diyagramı aşağıda verilmiştir (Şekil 3.1).



Şekil 3.1 : ANOVA Filtre öznitelik seçim yönteminin akış diyagramı.

3.1.2 Minimum tekrarlama maksimum ilgililik

Türkçe olarak Minimum Tekrarlama Maksimum İlgililik şeklinde isimlendirebileceğimiz öznitelik seçim yöntemi, literatürde "minimum Redundancy Maximum Relevance, mRMR" şeklinde karşımıza çıkmaktadır.

mRMR, öznitelik seçerken özniteliğin sınıfla olan karşılıklı bilgisinin (mutual information) yüksek olmasını aynı zamanda da özniteliklerin kendi aralarında karşılıklı bilgisinin düşük olmasını ister [24].

$$max\ D(S,C), \qquad D = \frac{1}{|S|}\sum_{x_i \epsilon S} I(x_i;C) \tag{3.1}$$

Denklem 3.1'de *S* öznitelik seti, *C* hedef sınıf, *D* ilgililik, *I* karşılıklı bilgi, x_i öznitelik setindeki her bir öznitelik olmak üzere her bir özniteliğin sınıfla olan ortalama karşılıklı bilgisine bakılır. Böylece "Maximum Relevance" için hedef sınıfımızla en yüksek ilgililik (relevance) içeren öznitelik seçilmiş olur. En yüksek ilgililik derken ilgililik yerine "korelasyon" (correlation) ya da "karşılıklı bilgi" de diyebiliriz [24].

$$min\ R(S), \qquad R = \frac{1}{|S|^2}\sum_{x_i,x_j \epsilon S} I(x_i,x_j) \tag{3.2}$$

Denklem 3.2'de *S* öznitelik seti, *R* tekrarlama (redundancy), *I* karşılıklı bilgi, x_i, x_j öznitelik setindeki öznitelik çiftleri olmak üzere her bir öznitelik çifti için çiftler arasındaki ortalama karşılıklı bilgi hesaplanır. Böylece "minimum Redundancy" değeri elde edilir. Elde edilen değer de minimize edilmek istenir. Bu durumda, iki öznitelik birbiriyle çok benzeşiyorsa bunlardan birinin ayırt ediciliği kalmayacağı için sınıf - öznitelik bazlı değerlendirmeye de katkısının olmayacağı söylenebilir [24].

$$max\ \phi\ (D,R), \phi = D\ -\ R \tag{3.3}$$

Denklem 3.3'te elde edilen *D* ve *R* değerleri birbirlerinden çıkarılarak sınıfla bağı yüksek ama diğer özniteliklerle bağı düşük olan öznitelik seçilmiş olur [24].

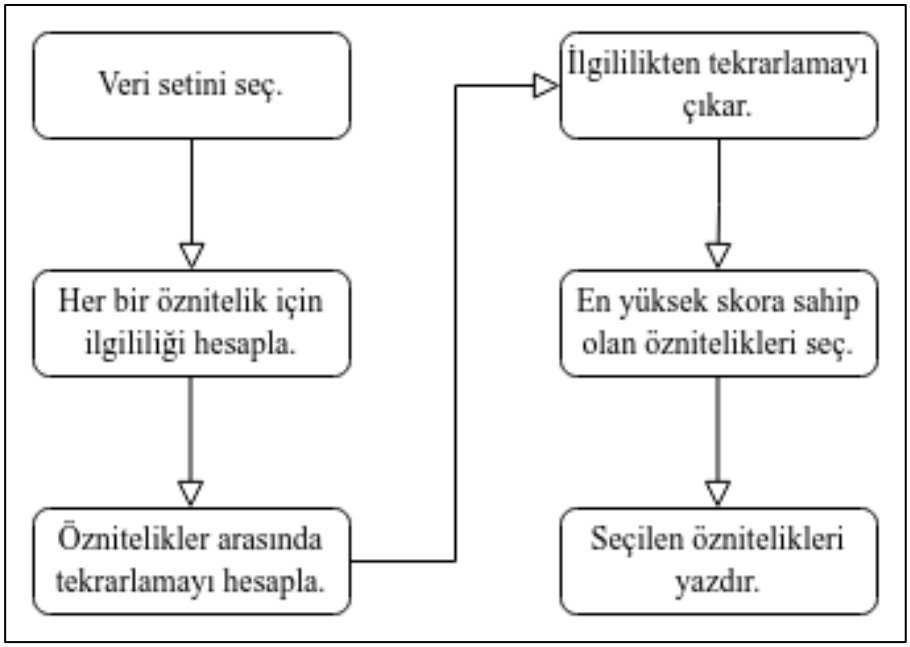
mRMR, Peng ve arkadaşları tarafından 2005 yılında öne sürülmüştür. Öne sürülmesi sonrasında Ding ve Peng, yine aynı yılda yaptıkları bir çalışmada 6 farklı gen ifadesi veri seti üzerinde 4 farklı sınıflandırıcıda tutarlı bir şekilde sınıflandırma başarımında iyileştirme gördüklerini raporlamışlardır [25].

Radovic ve arkadaşları tarafından yapılan bir çalışmada gen ifadesi veri setlerinde kullandıkları öznitelik seçim yöntemlerinin çoğunda veri düzleştirmesi yapmadan çalışma yapılmadığını görmüşlerdir. Gen ifadesi veri setlerinde çok sınıflı zamansal veriler bulunduğu için mevcut mRMR yönteminde bilgi kaybı yaşadıklarını

raporlamışlardır. Bilgi kaybını önlemek için önerdikleri yeni yönteme Temporal minimum Redundancy Maximum Relevance (TmRMR) ismini vermişlerdir. TmRMR her bir zaman adımındaki F-İstatistiği değerlerinin ortalamasını alarak "ilgi düzeyini" hesaplamakta ve dinamik bir zaman eğriltme yaklaşımı kullanarak "fazlalığı" hesaplamaktadır. TmRMR yönteminin üç farklı gen ifadesi veri seti üzerinde sınıflandırma başarımı bakımından daha iyi performans gösterdiği ifade edilmiştir [26].

Ramírez-Gallego ve arkadaşları, mRMR'ı, öznitelik seçimi yapmak için büyük verilerde kullandıklarında yaptığı hesaplamanın uzun sürdüğünü ve öznitelik sayısından dramatik bir şekilde etkilendiğini gözlemlemişlerdir. Bu amaçla hesaplama maliyetinin azaltmak için çeşitli çalışmalar yapıp hızlı mRMR (Fast mRMR) ismini verdikleri yöntemi önermişlerdir. Önerilen yöntemin eskisine göre çok daha hızlı çalışmakta olduğunu çeşitli veri setleri üzerinde deneyerek göstermişlerdir [27].

mRMR Öznitelik Seçim Yönteminin çalışma mantığını daha iyi anlamak için ilgili akış diyagramı aşağıda verilmiştir (Şekil 3.2).



Şekil 3.2 : mRMR öznitelik seçim yönteminin akış diyagramı.

3.2 Sarmal Yöntemler

Sarmal yöntemler (Wrapper methods) yineleme mantığı kullanarak çalışmaktadır. Sarmal yöntemlerde veri, eğitim ve test şeklinde ayrılır, bir sınıflandırıcıyla birlikte öznitelikler değerlendirme kriterine göre öznitelik alt kümesine eklenir, çıkarılır.

Açgözlü yaklaşım kullanılarak en iyi sınıflandırma başarımını veren alt küme oluşturulur. En yüksek başarıyı veren öznitelik alt kümesi, öznitelik seçiminin çıktısı olarak verilmektedir.

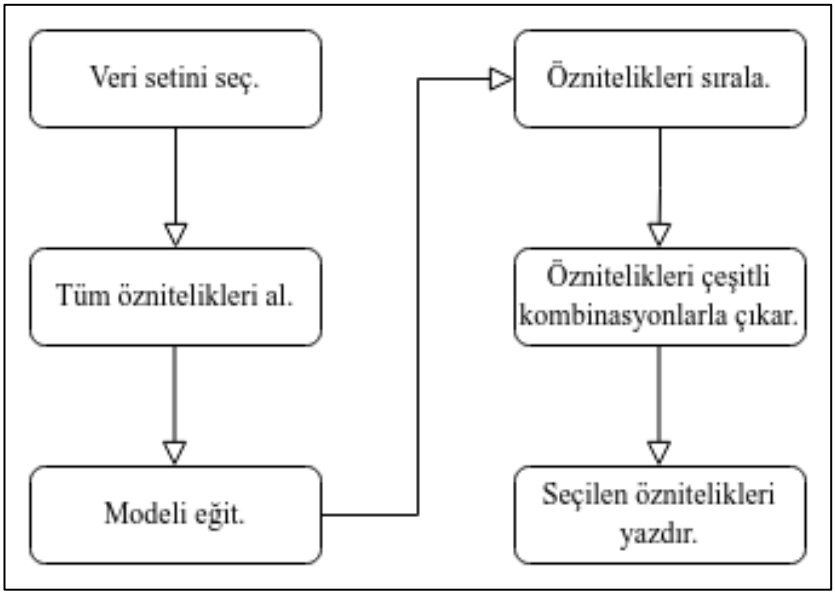
3.2.1 Özyinelemeli öznitelik eleme

Türkçeye Özyinelemeli Öznitelik Eleme olarak çevirebileceğimiz "Recursive Feature Elimination, RFE" öznitelik seçim yöntemi, sarmal tabanlı bir öznitelik seçim yöntemidir. RFE, yinelemeli bir şekilde ayrı ayrı modeller eğitir ve değerlendirme kriterine göre en anlamsız öznitelikleri, öznitelik kümesinden çıkarır. Bu işlem yapılırken her adımda değişme olasılığı olan öznitelikler kümesiyle sınıflandırıcıyı besleyerek sınıflandırma başarımında nasıl bir etki bıraktığını da görme olanağımız mevcuttur.

Guyon ve arkadaşları, kolon kanseri sınıflandırması üzerine elde edilmiş olan bir veri seti üzerinde RFE'yi DVM ile birlikte kullanmayı önermişlerdir. Seçilen özniteliklerle sınıflandırma yaptıklarında daha iyi sınıflandırma performansı sağladıklarını ve seçilen özniteliklerin kanserle biyolojik olarak ilgili olduğunu deneysel olarak göstermişlerdir. Kullandıkları veri setinde, RFE ile elde ettikleri 4 özniteliği kullandıklarında başarımın %98 çıktığını, öznitelik seçimi yapmadan sınıflandırma yaptıklarında %86 başarım elde ettiklerini raporlamışlardır [28].

Chen ve Jeong, örnek sayısının az olduğu ama öznitelik sayısının fazla olduğu veri setlerinde RFE kullanıldığında belirli bir özniteliğin zayıf olarak addedilmesiyle öznitelik alt kümesinden çıkarılması sonrasında problemler yaşattığını fark etmişlerdir. RFE'de bulunan öznitelik çıkarma mantığına yeni bir katkıda bulunarak "Enhanced RFE"yi öne sürmüşlerdir. Kullandıkları 7 farklı veri setinde RFE ve "Enhanced RFE"yi kıyaslamışlardır. "Enhanced RFE"nin sınıflandırma başarımında %12'ye yakın iyileştirme yaptığını raporlamışlardır [29].

RFE Öznitelik Seçim Yönteminin çalışma mantığını daha iyi anlamak için ilgili akış diyagramı aşağıda verilmiştir (Şekil 3.3).



Şekil 3.3 : RFE öznitelik seçim yönteminin akış diyagramı.

3.3 Hibrit Yöntemler

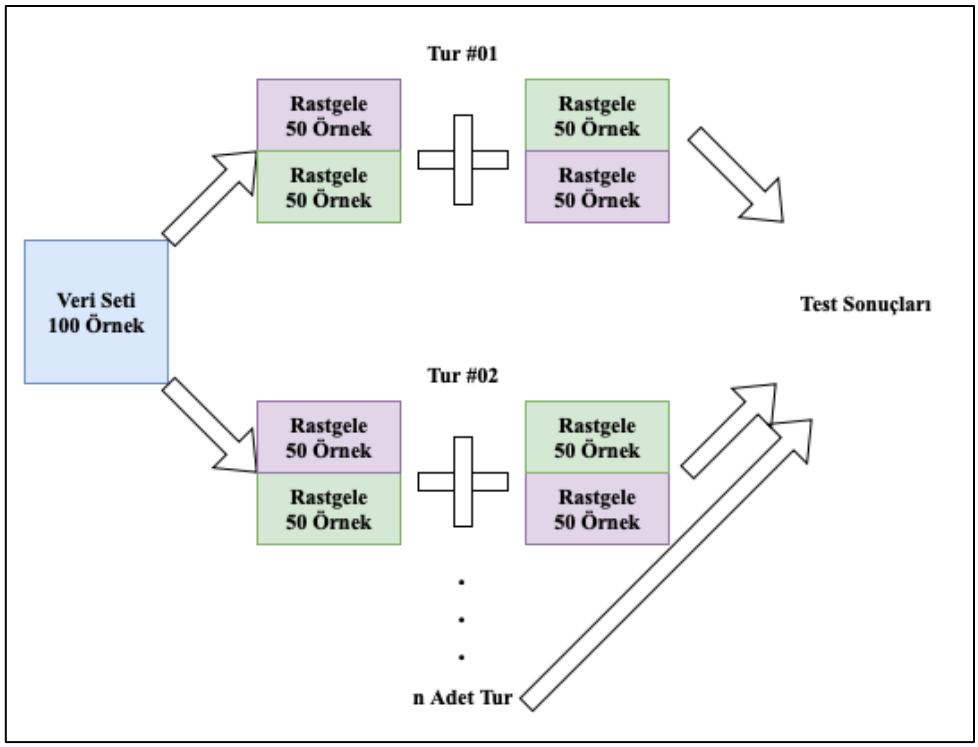
Hibrit yöntemler (Hybrid methods), filtre yöntemler ile sarmal yöntemlerin iyi yönlerinin birleştirilmesiyle oluşturulmuş yöntemlerdir. Modelde her bir iterasyonda en anlamlı özniteliklerin seçilmesi bazlı olarak ilerlediği için iteratif olarak çalışmaktadırlar.

4. ÇAPRAZ GEÇERLEME

Bu bölümde tez çalışmasında kullanılan çapraz geçerleme (cross validation, CV) yöntemleri hakkında bazı tanımlar sunulacaktır. Çapraz geçerleme (ÇD) yöntemlerinin detaylarından bahsedilecektir.

4.1 5x2 Kat Çapraz Geçerleme

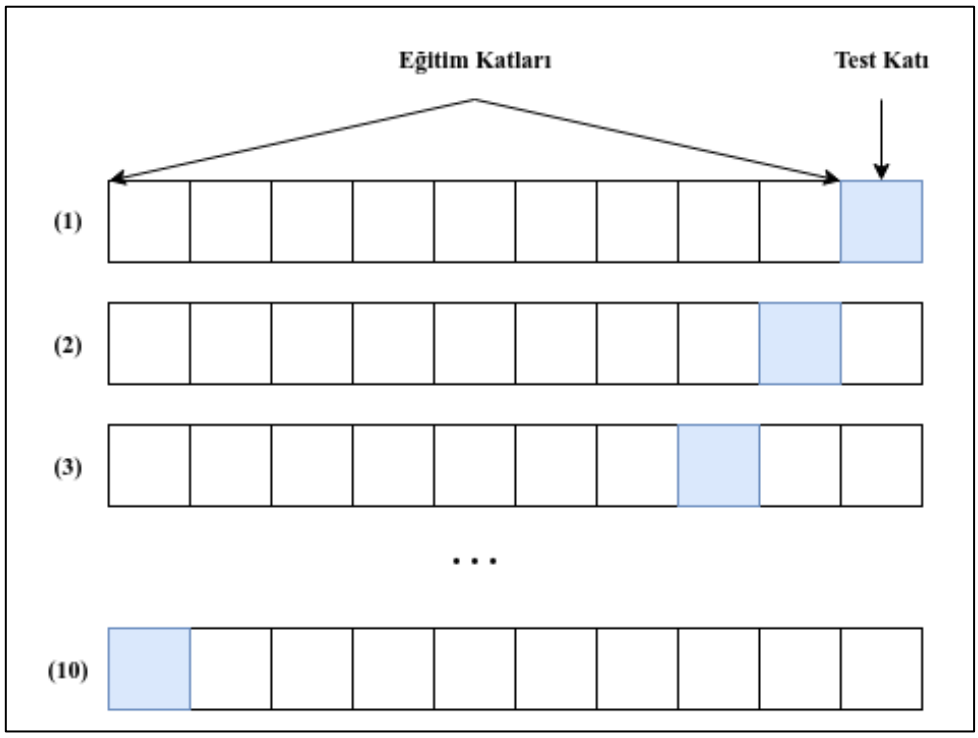
5x2 Kat Çapraz Geçerleme yöntemi iç içe çalışma mantığı gütmektedir. İç çapraz geçerlemede, veri setini eğitim ve test katlarına ayırmak için kullanılan bir dış K kat çapraz geçerleme döngüsü vardır. Dış döngüye ek olarak, eğitim ve geçerleme katını kullanarak en uygun modeli seçmek için kullanılan bir iç K kat çapraz geçerleme döngüsü vardır. Bu durumu açıklayan şekil aşağıda verilmiştir (Şekil 4.1).



Şekil 4.1 : 5x2 Kat Çapraz Geçerleme [30].

4.2 10 Kat Çapraz Geçerleme

10 Kat Çapraz Geçerlemede öncelikle verideki örneklerin yerleri değiştirilir ve 10 eşit kata bölünecek şekilde katlara ayrılır. İlk iterasyonda, 9 kat eğitim için kullanılır kalan 1 kat test için kullanılır. Model, eğitim katları ile eğitilir sonra test katı ile değerlendirilir. Sonraki iterasyonda daha önce test için kullanılmayan 1 kat test için seçilip kalanları da eğitim katları olarak kullanılır. Bu haliyle model tekrar eğitilmeli ve yeni test katı ile test edilmelidir. Bu işlem 10 kez olacak şekilde tekrarlandığında 10 Kat Çapraz Geçerleme yapılmış olmaktadır. 10 Kat Çapraz Geçerlemeyi açıklayan şekil aşağıda verilmiştir (Şekil 4.2).

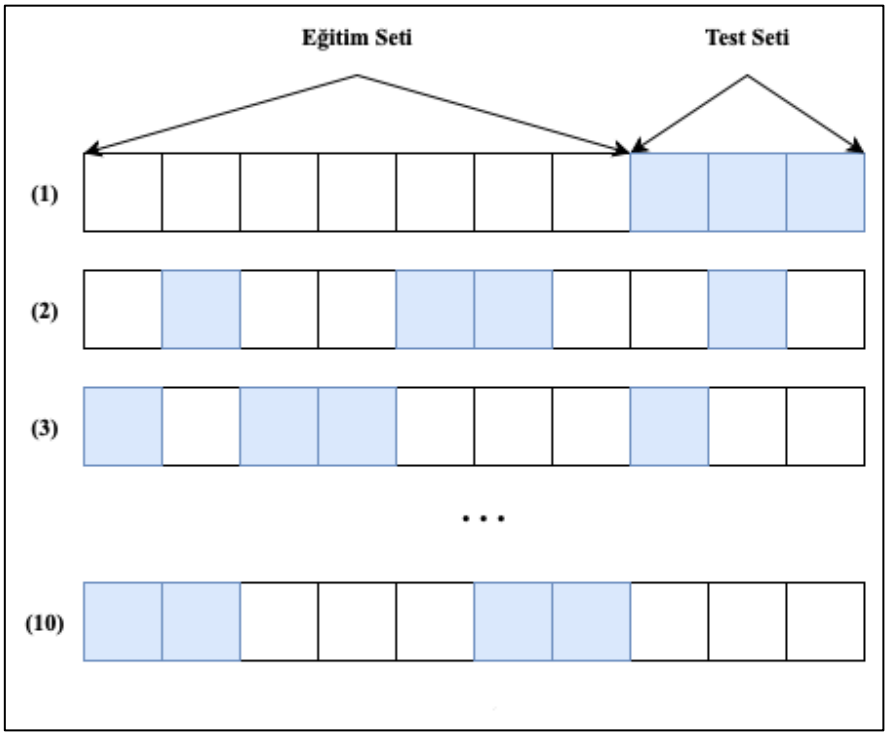


Şekil 4.2 : 10 Kat Çapraz Geçerleme [31].

4.3 Monte Carlo Çapraz Geçerleme

Monte Carlo Çapraz Geçerlemede eğitim ve test seti oranları rastgele olarak elde edilmektedir. Her iterasyonda bu oranlar değişmektedir. Örneğin, ilk iterasyonda %70 - %30 oranlarıyla eğitim - test seti yapılmışsa sonraki iterasyonda rastgele olarak %62,5 - %37,5 alınmış olabilir. İlk iterasyonda eğitim setimizle model eğitilmeli ve test setiyle test edilmelidir. Sonraki iterasyonda farklı bir oranda olan eğitim - test seti alınmalı, eğitim setiyle model eğitilmeli ve test setiyle sınanmalıdır. Bu şekilde ilerleyerek iterasyon sayısını artırdığımızda yapılan işlem Monte Carlo Çapraz

Geçerleme ismiyle anılır. Burada yerine koyma yapılarak (with replacement) örnekleri seçme söz konusu olduğu için aynı test örneklerinin birden fazla kez farklı iterasyonlarda seçilme olasılığı mevcuttur. Yerine koyma ve genel olarak Monte Carlo Çapraz Geçerlemenin mantığı aşağıdaki şekilde verilmiştir (Şekil 4.3).



Şekil 4.3 : Monte Carlo Çapraz Geçerleme [32].

5. ÖZNİTELİK SEÇİM YÖNTEMLERİNDEKİ YANLILIK ETKİSİ

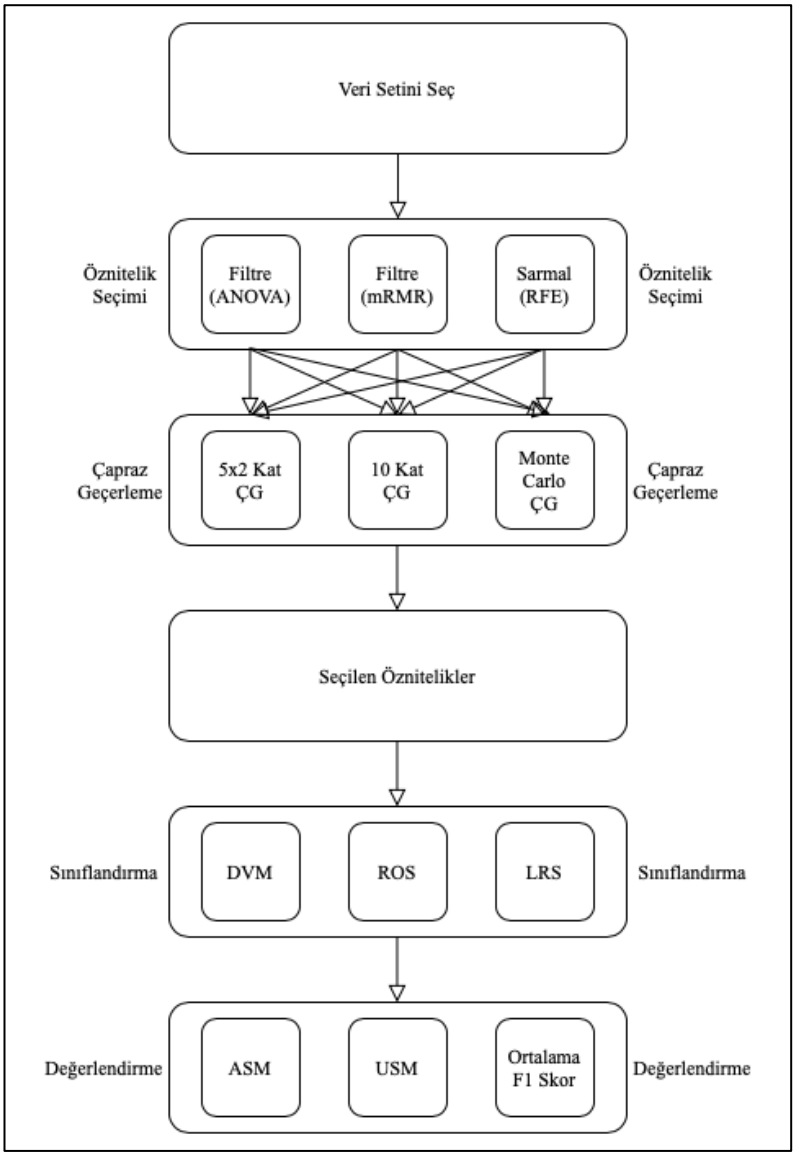
Öznitelik seçimi yapılırken verinin doğasından kaynaklanan nedenlerle yanlış seçimler yapılması olasıdır. Yapılan yanlış seçimler yanlılık etkisine (bias effect) neden olabilmektedir. Yanlılık etkisi gerçekte ilgili olmayan özniteliklerin seçimine ya da önemli özniteliklerin öznitelik alt kümesine seçilmemesine neden olabilmektedir.

Öznitelik seçiminde yanlılık etkisini azaltmak için seçim yöntemindeki kriterleri iyi değerlendirmek gerektiği, kriterin uygun olduğuna karar verdikten sonra çeşitli veri setleri üzerinde deneme yapmak gerektiği ve çapraz geçerleme yöntemleri kullanmak gerektiği literatürde sıklıkla önerilmektedir.

5.1 Öznitelik Seçimi İçin Kullanılan Yaklaşım

Yapılan tez çalışmasında öncelikle veri setleri literatürde taranarak elde edilmiş ve veri setlerinde veri ön işleme adımları yapılmıştır. Veri ön işleme adımından sonra her bir veri setinde ANOVA Filtre, mRMR ve RFE isimli üç farklı öznitelik seçim yöntemiyle öznitelik seçimi yapılmıştır. Her bir öznitelik seçim yöntemi kullanılırken üç farklı sınıflandırıcı (DVM, ROS ve LRS) ve üç farklı çapraz geçerleme yöntemi (5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme) sırayla kullanılarak bir öznitelik seçim yöntemi için üç sınıflandırıcı, üç çapraz geçerleme olmak üzere 9 çalıştırma gerçekleştirilmiştir. Yapılan 9 çalıştırmada da 29 deneme yapılmıştır. Bu sayede 29 deneme içinde "Seçilen öznitelik alt kümesinde değişiklik var mı, yok mu?" ve "Seçilen bu öznitelik alt kümesiyle sınıflandırma yapılsa ortalama F1 Skoru ne olurdu?" sorularının cevabını elde etme şansımız olmuştur. Bu sonuçların elde edilmiş olmasıyla öznitelik seçim yöntemlerindeki yanlılık etkisini ölçmek için literatürde bulunan Ayarlanmış Kararlılık Metriği (Adjusted Stability Measure, ASM) ve Ayarlanmamış Kararlılık Metriği (Unadjusted Stability Measure, USM) sayesinde incelemeler yapılmıştır. Yanlılık etkisinin çözümüne yönelik olarak önerilen çapraz geçerleme yöntemleri kendi aralarında kıyaslanmış ve veri setinin, öznitelik seçim yöntemlerinin, yanlılık etkisine nasıl bir etki uyguladığı gözlemlenmiştir.

Tez çalışmasında uygulanan adımlar aşağıdaki akış diyagramında verilmiştir (Şekil 5.1).



Şekil 5.1 : Tez çalışmasının akış diyagramı.

5.2 Öznitelik Seçimindeki Yanlılık Etkisi İçin Kullanılan Metrikler

Tez çalışmasında kullanılan ASM ve USM metrikleri genel olarak seçilen iki öznitelik alt kümesi arasındaki benzerliği ölçmektedir. Çalışmada yanlılık etkisini gözlemlemek için 29 deneme yapılmış olduğu için 29 denemede 29 öznitelik alt kümesi elde edilmiştir. Elde edilen öznitelik alt kümeleri için ASM ve USM değerleri hesaplanmış ve öznitelik seçim yöntemi, çapraz geçerleme yöntemi bazında karşılaştırma yapılarak yanlılık etkisi gözlemlenebilmiştir.

5.2.1 Ayarlanmış kararlılık metriği

Ayarlanmış kararlılık metriği literatürde "Adjusted Stability Measure, ASM" şeklinde karşımıza çıkmaktadır. Ayarlanmış kararlılık metriği (ASM) seçilen iki öznitelik alt kümesindeki benzerliği ölçmektedir. Benzerlik için de iki öznitelik alt kümesi arasında kıyaslama yapmak amacıyla öznitelik alt kümelerinin eleman sayıları, ortak elemanlarının sayısı vb. bilgilerle aşağıdaki denklemler üzerinden hesaplama yapmaktadır [2].

$$S_A(F^{K_i}, F^{K_j}) = \frac{r \; - \; (k_i k_j \; / \; n)}{min \; (k_i \; , \; k_j) \; - \; max \; (0, \; k_i \; + \; k_j \; - \; n)} \tag{5.1}$$

Denklem 5.1'de S_A iki öznitelik alt kümesi (F^{ki}, F^{kj}) arasındaki benzerlik, r iki öznitelik alt kümesi içinde aynı olan özniteliklerin sayısı, k_i öznitelik alt kümelerinden birincisinin eleman sayısı, k_j öznitelik alt kümelerinden ikincisinin eleman sayısı, n toplam öznitelik sayısı olmak üzere her bir öznitelik alt kümesi için bu değer hesaplanmaktadır. S_A, öznitelik alt kümeleri arasında aynı elemanlar olmadığı zaman 0 değerini almakta ve küçük alt kümedeki tüm öznitelikler büyük alt kümede de olduğunda yani $r = min(k_i, k_j)$ olduğunda S_A değeri mümkün olan en yüksek değer olan 1 değerini almaktadır [2].

$$ASM = \frac{2}{c(c \; - \; 1)} \sum_{i \; = \; 1}^{c \; - \; 1} \; \sum_{j \; = \; i \; + \; 1}^{c} S_A(F^{K_i}, F^{K_j}) \tag{5.2}$$

Denklem 5.2'de S_A iki öznitelik alt kümesi (F^{ki}, F^{kj}) arasındaki benzerlik, c elde edilen öznitelik alt kümelerinin sayısı olmak üzere her bir öznitelik alt kümesi için hesaplanan S_A değerlerinin ortalaması alınarak ASM değeri hesaplanmış oluyor [2].

5.2.2 Ayarlanmamış kararlılık metriği

Ayarlanmamış kararlılık metriği literatürde "Unadjusted Stability Measure, USM" şeklinde karşımıza çıkmaktadır. Ayarlanmamış kararlılık metriği (USM), seçilen iki öznitelik alt kümesi arasındaki benzerliği ölçmektedir. Benzerlik için de iki öznitelik alt kümesi arasında kıyaslama yapmak amacıyla öznitelik alt kümelerinin eleman sayıları, kümedeki tümleyen eleman sayısı, birleşim kümesinin eleman sayısı vb. bilgilerle aşağıdaki denklemler üzerinden hesaplama yapmaktadır [33].

$$S_S(s_i, s_j) = 1 - \frac{|s_i| + |s_j| - 2|s_i \cap s_j|}{|s_i| + |s_j| - |s_i \cap s_j|} = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \quad (5.3)$$

Denklem 5.3'te S_S iki öznitelik alt kümesi (s_i, s_j) arasındaki benzerliği, s_i öznitelik alt kümelerinden birincisini, s_j öznitelik alt kümelerinden ikincisini, "| |" operatörü de ilgili öznitelik alt kümesinin eleman sayısını göstermektedir. Kısaca, öznitelik alt kümelerinde, kesişim kümesindeki eleman sayısı, birleşim kümesindeki eleman sayısına bölündüğünde S_S değeri elde edilir [33].

$$USM = \frac{2}{c(c-1)} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} S_S(s_i, s_j) \quad (5.4)$$

Denklem 5.4'te S_S iki öznitelik alt kümesi (s_i, s_j) arasındaki benzerliği, c elde edilen öznitelik alt kümelerinin sayısını göstermek üzere her bir öznitelik alt kümesi için hesaplanan S_S değerlerinin ortalaması ile USM değeri hesaplanmış olur [33].

6. BULGULAR VE YORUMLAR

Bu bölümde tez çalışmasında yapılan işlemlerin genel çerçevesi, çalışmanın sonuçları ve parametreler bakımından karşılaştırma yapılarak analizler sunulacaktır.

6.1 Veri Setleri

Tez çalışmasında üç farklı veri seti kullanılmıştır. İlk veri seti "Breast Cancer" ismiyle literatürde bulunan bir veri setidir. Söz konusu veri seti "University of Wisconsin Clinical Sciences Center" tarafından elde edilmiş olan bir veri setidir [34]. Bu veri setinde 32 adet öznitelik bulunmaktadır. Bu öznitelikler içinde bir öznitelik de sınıf etiketini temsil eden "Diagnosis" özniteliğidir. "Diagnosis" özniteliği her bir örnek için "M" ya da "B" karakterleriyle verilmiştir. Ayrıca veri setinde 569 adet örnek bulunmaktadır. Veri setindeki öznitelikler meme dokusunda şüphe uyandıran dokunun özelleşmiş bir aspirasyon cihazıyla ölçümlediği doku çapı, doku yoğunluğu, dokunun çevresi vb. bilgilerdir. "Breast Cancer" veri setindeki öznitelikleri açıklayan çizelge aşağıda verilmiştir (Çizelge 6.1).

Çizelge 6.1 : "Breast Cancer" veri setindeki öznitelikler ve açıklamaları.

Öznitelik	Özniteliğin Açıklaması
Radius Mean	Dokunun merkezinden çevresine olan uzaklıkların ortalamasını veren virgüllü sayı
Texture Mean	Dokunun gri ölçekli değerlerinin ortalamasını veren virgüllü sayı
Perimeter Mean	Dokunun çevresini ortalama olarak veren virgüllü sayı
Area Mean	Dokunun alanını ortalama olarak veren virgüllü sayı
Smoothness Mean	Dokudaki yarıçap uzunluklarının yerel varyasyonunu ortalama olarak veren virgüllü sayı
Compactness Mean	$((\text{Perimeter})^2 / (\text{Area} - 1))$'i ortalama olarak veren virgüllü sayı
Concavity Mean	Doku konturünün iç bükey kısımlarının şiddetini ortalama olarak veren virgüllü sayı
Concave Points Mean	Doku konturünün iç bükey kısımlarının sayısını ortalama olarak veren virgüllü sayı
Symmetry Mean	Dokunun simetrisini ortalama olarak veren virgüllü sayı
Fractal Dimension Mean	Dokunun ("kıyı şeridi yaklaşımını (coastline approximation)" - 1) ortalama olarak veren virgüllü sayı

Çizelge 6.1 (devam) : "Breast Cancer" veri setindeki öznitelikler ve açıklamaları.

Öznitelik	Özniteliğin Açıklaması
Radius SE	Dokunun merkezinden çevresine olan uzaklıkların standart sapmasını veren virgüllü sayı
Texture SE	Dokunun gri ölçekli değerlerinin standart sapmasını veren virgüllü sayı
Perimeter SE	Doku çevresinin standart sapmasını veren virgüllü sayı
Area SE	Doku alanının standart sapmasını veren virgüllü sayı
Smoothness SE	Dokudaki yarıçap uzunluklarının yerel varyasyonunun standart sapmasını veren virgüllü sayı
Compactness SE	((Perimeter)^2 / (Area - 1))'in standart sapmasını veren virgüllü sayı
Concavity SE	Doku konturünün iç bükey kısımlarının şiddetinin standart sapmasını veren virgüllü sayı
Concave Points SE	Doku konturünün iç bükey kısımlarının sayısının standart sapmasını veren virgüllü sayı
Symmetry SE	Doku simetrisinin standart sapmasını veren virgüllü sayı
Fractal Dimension SE	Dokunun ("kıyı şeridi yaklaşımının (coastline approximation)" - 1) standart sapmasını veren virgüllü sayı
Radius Largest	Dokunun merkezinden çevresine olan en büyük 3 uzaklığının ortalamasını veren virgüllü sayı
Texture Largest	Dokunun gri ölçekli en büyük 3 değerlerinin ortalamasını veren virgüllü sayı
Perimeter Largest	Dokudaki en büyük 3 çevrenin ortalamasını veren virgüllü sayı
Area Largest	Dokudaki en büyük 3 alanın ortalamasını veren virgüllü sayı
Smoothness Largest	Dokudaki en büyük 3 yarıçap uzunluğunun yerel varyasyonunu ortalama olarak veren virgüllü sayı
Compactness Largest	((Perimeter)^2 / (Area - 1))'daki en büyük 3 değerin ortalamasını veren virgüllü sayı
Concavity Largest	Doku konturünün iç bükey kısımlarının şiddetindeki en büyük 3 değeri ortalama olarak veren virgüllü sayı
Concave Points Largest	Doku konturünün iç bükey kısımlarındaki en büyük 3 değeri ortalama olarak veren virgüllü sayı
Symmetry Largest	Dokunun simetrisindeki en büyük 3 değeri ortalama olarak veren virgüllü sayı
Fractal Dimension Largest	Dokunun ("kıyı şeridi yaklaşımındaki (coastline approximation)" - 1) en büyük 3 değeri ortalama olarak veren virgüllü sayı
Diagnosis	Dokunun iyi huylu ya da kötü huylu olduğu veren karakter

İkinci veri seti "Diabetes" ismiyle literatürde verilen "National Institute of Diabetes and Digestive and Kidney Diseases" tarafından elde edilmiş olan bir veri setidir [35]. Bu veri setinde 8 adet öznitelik bulunmaktadır. Bunlar "Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction",

“Age”, “Outcome” isimleriyle veri setinde bulunmaktadır. 8 özniteliğin içinde bir öznitelik de sınıf etiketini temsil eden “Outcome” özniteliğidir. “Outcome” özniteliği her bir örnek için “0” ya da “1” değeriyle verilmiştir. Ayrıca veri setinde 768 adet örnek bulunmaktadır. “Diabetes” veri setindeki öznitelikleri açıklayan çizelge aşağıda verilmiştir (Çizelge 6.2).

Çizelge 6.2 : “Diabetes” veri setindeki öznitelikler ve açıklamaları.

Öznitelik	Özniteliğin Açıklaması
Pregnancies	Kaç kez hamile kaldığını veren tam sayı
Glucose	Oral glukoz tolerans testinde iki saatte plazma glukoz konsantrasyonu veren tam sayı
BloodPressure	Diyastolik kan basıncını veren tam sayı
SkinThickness	Triseps deri kıvrım kalınlığını veren tam sayı
Insulin	İki saatlik serum insülini veren tam sayı
BMI	Vücut kitle indeksini veren ondalıklı sayı
DiabetesPedigreeFunction	Diyabet soyağacı fonksiyonunu veren ondalıklı sayı
Age	Yıl bazında yaş bilgisini veren tam sayı
Outcome	Diyabet hastası olup olmadığını veren tam sayı

Üçüncü veri seti “Ionosphere” ismiyle çeşitli veri tabanlarında paylaşılmış olan bir veri setidir. Bu veri setinde 35 adet öznitelik bulunmaktadır ama sınıf etiketini veren öznitelik dışında özniteliklerin detayları verilmemiştir. Veri seti “Space Physics Group of the Johns Hopkins University Applied Physics Laboratory” tarafından yürütülen bir çalışmada 16 yüksek frekans anteninin iyonosfer tabakasında bulunan serbest elektronları gözlemlemesi ile elde edilmiştir [36]. Bu öznitelikler içinde bir öznitelik de sınıf etiketini temsil eden “Outcome” özniteliğidir. “Outcome” özniteliği her bir örnek için “B” ya da “G” karakterleriyle verilmiştir. Ayrıca veri setinde 351 adet örnek bulunmaktadır.

6.2 “Breast Cancer” Veri Setinde Öznitelik Seçiminin Uygulanması

Bu bölümde “Breast Cancer” veri seti üzerinde ANOVA Filtre, mRMR ve RFE öznitelik seçim yöntemleri kullanılarak elde edilen bulgular yorumlarıyla birlikte verilmiştir. Öznitelik seçimi yapılırken yanlılık etkisini analiz etmek için farklı çapraz geçerleme yöntemleri uygulanmıştır.

6.2.1 ANOVA filtre öznitelik seçim yöntemi

"Breast Cancer" isimli veri setimizde 32 adet öznitelik bulunmaktadır. Bu öznitelikler, ANOVA Filtre öznitelik seçim yönteminde kullanılacak şekilde girdi olarak verilmiştir. Öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle ANOVA Filtre ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde ANOVA Filtre ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da ANOVA Filtre ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.3).

Çizelge 6.3 : "Breast Cancer" veri setinde ANOVA filtre öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,815	0,926	0,858	0,833	1	0,842	0,833	1	0,842
ROS	0,815	0,926	0,909	0,833	1	0,912	0,833	1	0,910
LRS	0,812	0,919	0,901	0,833	1	0,895	0,833	1	0,895

ANOVA Filtre öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda 10 Kat Çapraz Geçerleme ve Monte Carlo Çapraz Geçerleme kullanıldığında USM'nin 1 çıktığını görmekteyiz. Bu da seçilen öznitelik alt kümesinin hiçbir zaman değişmediğini, aynı özniteliklerden oluştuğunu göstermektedir. Benzer şekilde en yüksek ASM değeri de 10 Kat Çapraz Geçerleme ve Monte Carlo Çapraz Geçerleme yöntemlerinde elde edilmiştir. ANOVA Filtre öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları %84 ile %91 arasında değişmektedir.

6.2.2 mRMR öznitelik seçim yöntemi

Bu kısımdaki denemeler mRMR öznitelik seçim yöntemi kullanılarak elde edilmiştir. mRMR ile öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle mRMR ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde mRMR ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da mRMR ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.4).

Çizelge 6.4 : "Breast Cancer" veri setinde mRMR öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,543	0,859	0,891	0,550	0,882	0,904	0,532	0,852	0,912
ROS	0,530	0,859	0,929	0,540	0,872	0,941	0,506	0,839	0,934
LRS	0,518	0,844	0,908	0,540	0,871	0,915	0,507	0,823	0,913

mRMR öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda tüm çapraz geçerleme yöntemlerinde USM'nin değişim aralığının benzer çıktığını görmekteyiz. Bu da seçilen öznitelik alt kümesinin mRMR yönteminde üç farklı çapraz geçerleme için benzer yanlılık etkisi oluşturduğunu göstermektedir. mRMR öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları %89 ile %91 arasında değişmektedir.

6.2.3 RFE öznitelik seçim yöntemi

Farklı öznitelik seçim yöntemi kullanılarak yapılan denemelerin içinde son olarak RFE öznitelik seçim yöntemi kullanılmıştır. RFE ile öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle RFE ve DVM sabit tutularak üç

farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde RFE ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da RFE ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.5).

Çizelge 6.5 : "Breast Cancer" veri setinde RFE öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,556	0,738	0,926	0,624	0,718	0,935	0,697	0,795	0,936
ROS	0,614	0,744	0,922	0,809	0,914	0,932	0,769	0,863	0,930
LRS	0,528	0,753	0,924	0,606	0,840	0,929	0,621	0,762	0,930

RFE öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda 10 Kat Çapraz Geçerleme ve Monte Carlo Çapraz Geçerleme yöntemleri kullanıldığında USM'nin değer aralığı geniş çıkmıştır. Bu da yanlılık etkisinin bir göstergesidir. RFE öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları %92 ile %93 arasında değişmektedir.

6.3 "Diabetes" Veri Setinde Öznitelik Seçiminin Uygulanması

Bu bölümde "Diabetes" veri seti üzerinde ANOVA Filtre, mRMR ve RFE öznitelik seçim yöntemleri kullanılarak elde edilen bulgular yorumlarıyla birlikte verilmiştir. Öznitelik seçimi yapılırken yanlılık etkisini analiz etmek için farklı çapraz geçerleme yöntemleri uygulanmıştır.

6.3.1 ANOVA filtre öznitelik seçim yöntemi

"Diabetes" isimli veri setimizde 8 adet öznitelik bulunmaktadır. Bu öznitelikler, ANOVA Filtre öznitelik seçim yöntemi kullanılacak şekilde girdi olarak verilmiştir. Öznitelik seçtirme yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle ANOVA Filtre ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde ANOVA Filtre ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da ANOVA Filtre ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.6).

Çizelge 6.6 : "Diabetes" veri setinde ANOVA filtre öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,710	0,915	0,596	0,625	0,978	0,568	0,627	0,959	0,592
ROS	0,727	0,941	0,624	0,625	1	0,631	0,630	0,943	0,625
LRS	0,705	0,942	0,625	0,625	0,988	0,620	0,634	0,931	0,623

ANOVA Filtre öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda 10 Kat Çapraz Geçerleme kullanıldığında USM'nin 1 ve 1'e yakın değerler verdiğini görmekteyiz. Bu da seçilen öznitelik alt kümesinin neredeyse hiçbir zaman değişmediğini, neredeyse aynı özniteliklerden oluştuğunu göstermiştir. ANOVA Filtre öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları %56 ile %63 arasında değişmektedir.

6.3.2 mRMR öznitelik seçim yöntemi

Bu kısımdaki denemeler mRMR öznitelik seçim yöntemi kullanılarak elde edilmiştir. mRMR ile öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız

için öncelikle mRMR ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde mRMR ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da mRMR ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.7).

Çizelge 6.7 : "Diabetes" veri setinde mRMR öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,617	0,895	0,561	0,5	1	0,564	0,502	0,951	0,587
ROS	0,591	0,863	0,619	0,5	1	0,616	0,503	0,911	0,618
LRS	0,622	0,899	0,622	0,5	0,986	0,613	0,504	0,905	0,621

mRMR öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda 10 Kat Çapraz Geçerleme kullanıldığında USM'nin 1 ve 1'e yakın değerler verdiğini görmekteyiz. Bu da seçilen öznitelik alt kümesinin neredeyse hiçbir zaman değişmediğini, neredeyse aynı özniteliklerden oluştuğunu göstermiştir. mRMR öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları ANOVA Filtre yöntemine benzer olarak %56 ile %62 arasında değişmektedir.

6.3.3 RFE öznitelik seçim yöntemi

Farklı öznitelik seçim yöntemi kullanılarak yapılan denemelerin içinde son olarak RFE öznitelik seçim yöntemi kullanılmıştır. RFE ile öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle RFE ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde RFE ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da RFE ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç

farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgeye baktığımızda yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.8).

Çizelge 6.8 : "Diabetes" veri setinde RFE öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,725	0,926	0,623	0,761	0,937	0,623	0,713	0,923	0,618
ROS	0,416	0,915	0,627	0,717	0,949	0,636	0,591	0,859	0,629
LRS	0,692	0,909	0,627	0,625	1	0,633	0,625	0,988	0,638

RFE öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda 10 Kat Çapraz Geçerleme ve Monte Carlo Çapraz Geçerleme yöntemleri kullanıldığında USM'nin değer aralığı geniş çıkmıştır. Bu da yanlılık etkisinin bir göstergesidir. RFE öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları %61 ile %63 arasında değişmektedir.

6.4 "Ionosphere" Veri Setinde Öznitelik Seçiminin Uygulanması

Bu bölümde "Ionosphere" veri seti üzerinde ANOVA Filtre, mRMR ve RFE öznitelik seçim yöntemleri kullanılarak elde edilen bulgular yorumlarıyla birlikte verilmiştir. Öznitelik seçimi yapılırken yanlılık etkisini analiz etmek için farklı çapraz geçerleme yöntemleri uygulanmıştır.

6.4.1 ANOVA filtre öznitelik seçim yöntemi

"Ionosphere" isimli veri setimizde 35 adet öznitelik bulunmaktadır. Bu öznitelikler, ANOVA Filtre öznitelik seçim yönteminde kullanılacak şekilde girdi olarak verilmiştir. Öznitelik seçtirme yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle ANOVA Filtre ve DVM sabit tutularak üç farklı çapraz

geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde ANOVA Filtre ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da ANOVA Filtre ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.9).

Çizelge 6.9 : "Ionosphere" veri setinde ANOVA filtre öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,649	0,709	0,885	0,816	0,965	0,889	0,778	0,882	0,889
ROS	0,618	0,699	0,919	0,821	0,990	0,930	0,790	0,907	0,929
LRS	0,638	0,727	0,896	0,821	0,990	0,898	0,783	0,871	0,900

ANOVA Filtre öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda 10 Kat Çapraz Geçerleme kullanıldığında USM'nin 1'e yakın değerler verdiğini görmekteyiz. Bu da seçilen öznitelik alt kümesinin neredeyse hiçbir zaman değişmediğini, neredeyse aynı özniteliklerden oluştuğunu göstermiştir. Ayrıca 5x2 Kat Çapraz Geçerleme yöntemindeki USM değerleri diğer iki çapraz geçerleme yöntemine göre daha düşük çıkmıştır. ANOVA Filtre öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları %88 ile %93 arasında değişmektedir.

6.4.2 mRMR öznitelik seçim yöntemi

Bu kısımdaki denemeler mRMR öznitelik seçim yöntemi kullanılarak elde edilmiştir. mRMR ile öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle mRMR ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde mRMR ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da mRMR ve LRS sabit tutularak üç farklı çapraz geçerleme

yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.10).

Çizelge 6.10 : "Ionosphere" veri setinde mRMR öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,566	0,881	0,895	0,593	0,908	0,895	0,554	0,855	0,892
ROS	0,615	0,891	0,936	0,595	0,911	0,947	0,593	0,889	0,944
LRS	0,638	0,885	0,896	0,598	0,910	0,894	0,578	0,887	0,896

mRMR öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda tüm çapraz geçerleme yöntemlerinde USM'nin değişim aralığının benzer çıktığını görmekteyiz. Bu da seçilen öznitelik alt kümesinin mRMR yönteminde üç farklı çapraz geçerleme için benzer yanlılık etkisi oluşturduğunu söylemiştir. mRMR öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları ANOVA Filtre yöntemine benzer olarak %89 ile %94 arasında değişmektedir.

6.4.3 RFE öznitelik seçim yöntemi

Farklı öznitelik seçim yöntemi kullanılarak yapılan denemelerin içinde son olarak RFE öznitelik seçim yöntemi kullanılmıştır. RFE ile öznitelik seçimi yaptığımızda yanlılık etkisinin nasıl oluştuğunu anlamamız için öncelikle RFE ve DVM sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Sonrasında aynı şekilde RFE ve ROS sabit tutularak üç farklı çapraz geçerleme, devamında da RFE ve LRS sabit tutularak üç farklı çapraz geçerleme yöntemi denenmiştir. Böylece her sınıflandırıcı için üç farklı çapraz geçerleme kullanarak 9 farklı çalıştırma yapılmıştır. Üç farklı çapraz geçerleme yöntemi için yanlılık etkisi iki farklı metrik ile hesaplanmış ve sınıflandırıcılar için ortalama F1 Skorunu metrik olarak verebilmemize imkân

sağlayan 29 deneme yapılmıştır. Aşağıdaki çizelgede yanlılık metriği olarak ASM, USM hesaplanmış ve son olarak da 29 denemenin ortalama F1 Skoru verilmiştir (Çizelge 6.11).

Çizelge 6.11 : "Ionosphere" veri setinde RFE öznitelik seçim yöntemindeki yanlılık metrikleri ve ortalama sınıflandırma başarımları.

	5x2 Kat Çapraz Geçerleme			10 Kat Çapraz Geçerleme			Monte Carlo Çapraz Geçerleme		
	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor	ASM	USM	Ortalama F1 Skor
DVM	0,424	0,581	0,892	0,678	0,751	0,896	0,645	0,668	0,899
ROS	0,562	0,708	0,923	0,748	0,911	0,934	0,690	0,790	0,928
LRS	0,486	0,625	0,896	0,763	0,930	0,893	0,688	0,768	0,894

RFE öznitelik seçim yöntemi; DVM, ROS, LRS yöntemleri ve 5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme ile 29 kez deneme sonuçlarına baktığımızda USM'nin değişiminin fazla olduğunu görmekteyiz. Bu da RFE'nin her çalışmasında çok farklı öznitelik alt kümeleri seçtiğini göstermiştir. Bu nedenden dolayı yanlılık etkisi bu veri seti için diğer öznitelik seçim yöntemlerine göre daha yüksek çıkmıştır. RFE öznitelik seçim yönteminde kullanılan üç farklı sınıflandırıcı ve çapraz geçerleme için ortalama F1 Skorları diğer üç öznitelik seçim yöntemine benzer olarak %89 ile %93 arasında değişmektedir.

6.5 Anlamlılık Testinin Uygulanması

"Breast Cancer", "Diabetes" ve "Ionosphere" isimli veri setleri için üç farklı öznitelik seçim yöntemi, üç farklı sınıflandırıcı ve üç farklı çapraz geçerleme yöntemiyle 81 farklı çalıştırma yapılmıştır. Yapılan her bir çalıştırma için de 29 deneme gerçekleştirilmiştir. 29 deneme için öncelikle her bir çapraz geçerleme yöntemi kullanılarak veri seti karıştırılmış, karışan veri setinde öznitelik seçimi yapılmıştır. Öznitelik seçimi sonucunda her bir 29 deneme için öznitelik alt kümesi ortaya çıkmıştır. Ortaya çıkan öznitelik alt kümesiyle sınıflandırma yaptığımızda da 29 deneme için F1 Skorları elde edilmiştir. Elde edilen F1 Skorlarında anlamlı bir farklılık var mı diye bakmak amacıyla T-Testi uygulanmıştır.

T-Testi iki grubu kıyaslamak adına yapılan bir istatistiksel testtir. Söz konusu istatistiksel testte iki grubun ortalamaları arasında anlamlı bir fark olup olmadığına bakılmaktadır [36].

T-Testini gerçekleştirmek için iki grubun ortalamaları arasındaki fark elde edildikten sonra farkın standart hatasına bölünmesiyle elde edilen T-Değeri (T-Statistic) hesaplanmaktadır [37].

P-Değeri (P-Value), bir hipotez testinde boş hipoteze karşı kanıtın gücünün bir ölçüsü olarak verilmiştir. T-Testi bakımından konuya bakılırsa sıfır hipotezi, karşılaştırılan iki grubun ortalamaları arasında anlamlı bir fark olmadığıdır. P-Değeri, sıfır hipotezinin doğru olduğunu varsayarak gözlemlenenden daha uç bir T-Değeri elde etme olasılığımızı söylemektedir [38]. Tez çalışmasında P-Değeri, öznitelik seçim yöntemi ve çapraz geçerleme yöntemine göre farklı sınıflandırıcıların F1 Skorları arasındaki benzerliği ifade etmek amacıyla kullanılmıştır.

6.5.1 "Breast cancer" veri setinde t-testinin uygulanması

"Breast Cancer" veri seti için üç farklı öznitelik seçim yöntemi, üç farklı sınıflandırıcı ve üç farklı çapraz geçerleme yöntemiyle 27 farklı çalıştırma yapılmıştır. Yapılan çalıştırmalar sonucunda elde edilen F1 Skorlarında anlamlı bir farklılık var mı diye bakmak amacıyla T-Testi uygulanmıştır. Aşağıdaki çizelgede söz konusu veri seti için yapılan çalıştırmalara göre T-Değeri, P-Değeri hesaplanmış ve öznitelik seçim metodu, çapraz doğrulama ikilisi bazında verilmiştir (Çizelge 6.12).

Çizelge 6.12 : "Breast Cancer" veri setine uygulanan T-Testinin sonuçları.

	DVM – ROS		DVM – LRS		ROS – LRS	
	T-Değeri	P-Değeri	T-Değeri	P-Değeri	T-Değeri	P-Değeri
ANOVA Filtre – 5x2 Kat ÇG	-6,991	3,5835e-09	-5,876	2,4151e-07	7,717	2,2608e-10
ANOVA Filtre – 10 Kat ÇG	-9,303	5,8517e-13	-7,047	2,8953e-09	10,257	1,8065e-14
ANOVA Filtre – Monte Carlo ÇG	-4,711	1,6706e-05	-3,675	0,0005	4,176	0,0001

Çizelge 6.12 (devam) : "Breast Cancer" veri setine uygulanan T-Testinin sonuçları.

	DVM – ROS		DVM – LRS		ROS – LRS	
	T-Değeri	P-Değeri	T-Değeri	P-Değeri	T-Değeri	P-Değeri
mRMR – 5x2 Kat ÇG	-9,136	1,0885e-12	-3,899	0,0002	14,608	9,3823e-21
mRMR – 10 Kat ÇG	-9,958	5,3225e-14	-2,802	0,0069	17,399	3,0034e-24
mRMR – Monte Carlo ÇG	-6,697	1,0949e-08	-0,212	0,8324	6,282	5,2638e-08
RFE – 5x2 Kat ÇG	3,063	0,0033	0,735	0,4653	-0,742	0,4611
RFE – 10 Kat ÇG	1,757	0,0843	2,575	0,0127	1,350	0,1825
RFE – Monte Carlo ÇG	2,055	0,0446	2,419	0,0188	0,045	0,9644

Çizelgeye baktığımız zaman RFE ve Monte Carlo Çapraz Geçerleme kullanıldığında ROS ile LRS arasında ortalama F1 Skorları bakımından benzer değerler olduğunu görmekteyiz. Çünkü RFE ve Monte Carlo Çapraz Geçerleme kullanıldığında ROS ile LRS arasında elde edilen P-Değeri diğer P-Değerlerinden yüksek çıkmıştır.

6.5.2 "Diabetes" veri setinde t-testinin uygulanması

"Diabetes" veri seti için üç farklı öznitelik seçim yöntemi, üç farklı sınıflandırıcı ve üç farklı çapraz geçerleme yöntemiyle 27 farklı çalıştırma yapılmıştır. Yapılan çalıştırmalar sonucunda elde edilen F1 Skorlarında anlamlı bir farklılık var mı diye bakmak amacıyla T-Testi uygulanmıştır. Aşağıdaki çizelgede söz konusu veri seti için yapılan çalıştırmalara göre T-Değeri, P-Değeri hesaplanmış ve öznitelik seçim metodu, çapraz doğrulama ikilisi bazında verilmiştir (Çizelge 6.13).

Çizelge 6.13 : "Diabetes" veri setine uygulanan T-Testinin sonuçları.

	DVM – ROS		DVM – LRS		ROS – LRS	
	T-Değeri	P-Değeri	T-Değeri	P-Değeri	T-Değeri	P-Değeri
ANOVA Filtre – 5x2 Kat ÇG	-4,293	7,0599e-05	-4,644	2,1086e-05	-0,779	0,4393
ANOVA Filtre – 10 Kat ÇG	-7,769	1,8559e-10	-6,602	1,5685e-08	4,455	4,0678e-05
ANOVA Filtre – Monte Carlo ÇG	-3,135	0,0027	-2,884	0,0055	0,434	0,6658

Çizelge 6.13 (devam) : "Diabetes" veri setine uygulanan T-Testinin sonuçları.

	DVM – ROS		DVM – LRS		ROS – LRS	
	T-Değeri	P-Değeri	T-Değeri	P-Değeri	T-Değeri	P-Değeri
mRMR – 5x2 Kat ÇG	-4,907	8,3404e-06	-5,239	2,5237e-06	-2,288	0,0259
mRMR – 10 Kat ÇG	-5,999	1,5279e-07	-5,729	4,1698e-07	1,278	0,2065
mRMR – Monte Carlo ÇG	-2,465	0,0168	-2,675	0,0097	-0,432	0,6674
RFE – 5x2 Kat ÇG	-2,583	0,0124	-2,732	0,0084	0,324	0,7468
RFE – 10 Kat ÇG	-5,317	1,9024e-06	-4,826	1,1129e-05	1,250	0,2165
RFE – Monte Carlo ÇG	-2,131	0,0374	-3,585	0,0007	-1,581	0,1195

Çizelgeye baktığımız zaman RFE ve 5x2 Kat Çapraz Geçerleme kullanıldığında ROS ile LRS arasında ortalama F1 Skorları bakımından benzer değerler olduğunu görmekteyiz. Çünkü RFE ve 5x2 Kat Çapraz Geçerleme kullanıldığında ROS ile LRS arasında elde edilen P-Değeri diğer P-Değerlerinden yüksek çıkmıştır.

6.5.3 "Ionosphere" veri setinde t-testinin uygulanması

"Ionosphere" veri seti için üç farklı öznitelik seçim yöntemi, üç farklı sınıflandırıcı ve üç farklı çapraz geçerleme yöntemiyle 27 farklı çalıştırma yapılmıştır. Yapılan çalıştırmalar sonucunda elde edilen F1 Skorlarında anlamlı bir farklılık var mı diye bakmak amacıyla T-Testi uygulanmıştır. Aşağıdaki çizelgede söz konusu veri seti için yapılan çalıştırmalara göre T-Değeri, P-Değeri hesaplanmış ve öznitelik seçim metodu, çapraz doğrulama ikilisi bazında verilmiştir (Çizelge 6.14).

Çizelge 6.14 : "Ionosphere" veri setine uygulanan T-Testinin sonuçları.

	DVM – ROS		DVM – LRS		ROS – LRS	
	T-Değeri	P-Değeri	T-Değeri	P-Değeri	T-Değeri	P-Değeri
ANOVA Filtre – 5x2 Kat ÇG	-32,098	9.5427e-38	-18,060	5,0739e-25	20,197	2.1962e-27
ANOVA Filtre – 10 Kat ÇG	-34,986	9.5968e-40	-9,226	7,7848e-13	26,414	2.6882e-33
ANOVA Filtre – Monte Carlo ÇG	-13,687	1.6416e-19	-3,843	0,0003	9,567	2.2171e-13

Çizelge 6.14 (devam) : "Ionosphere" veri setine uygulanan T-Testinin sonuçları.

	DVM – ROS		DVM – LRS		ROS – LRS	
	T-Değeri	P-Değeri	T-Değeri	P-Değeri	T-Değeri	P-Değeri
mRMR – 5x2 Kat ÇG	-52,799	1.8644e-49	-1,918	0,0602	48,809	1.3856e-47
mRMR – 10 Kat ÇG	-47,371	7.1118e-47	1,601	0,1150	52,324	3.0625e-49
mRMR – Monte Carlo ÇG	-17,622	1.6394e-24	-1,492	0,1412	17,206	5.0897e-24
RFE – 5x2 Kat ÇG	-28,574	4.4416e-35	-4,969	6,6850e-06	24,997	4.6510e-32
RFE – 10 Kat ÇG	-28,795	2.9671e-35	2,101	0,0401	23,931	4.3441e-31
RFE – Monte Carlo ÇG	-11,355	3.7437e-16	1,735	0,0882	12,971	1.6402e-18

Çizelgeye baktığımız zaman mRMR ve Monte Carlo Çapraz Geçerleme kullanıldığında DVM ile LRS arasında ortalama F1 Skorları bakımından diğerlerine göre benzer değerler olduğunu görmekteyiz. Çünkü mRMR ve Monte Carlo Çapraz Geçerleme kullanıldığında DVM ile LRS arasında elde edilen P-Değeri diğer P-Değerlerinden yüksek çıkmıştır.

7. SONUÇ VE ÖNERİLER

Bu tez çalışmasında çeşitli veri setleri üzerinde öznitelik seçimi yapılmış ve bir nevi boyut azaltma işlemi bu şekilde gerçekleştirilmiştir. Öznitelik seçim yöntemlerinde, seçim yaparken yanlılık etkisi nedeniyle seçilen öznitelik kümelerinde farklılıklar gözlemlenebilmektedir. Bu farklılıkların etkisini azaltmak için literatürde çeşitli çapraz geçerleme yöntemlerinin kullanılması önerilmektedir. Yapılan çalışmada öncelikle veri setleri üzerinde veri ön işleme adımları yapılmıştır. Veri ön işleme adımından sonra her bir veri setinde ANOVA Filtre, mRMR ve RFE isimli üç farklı öznitelik seçim yöntemiyle öznitelik seçimi yapılmıştır. Her bir öznitelik seçim yöntemi kullanılırken üç farklı sınıflandırıcı (DVM, ROS ve LRS) ve üç farklı çapraz geçerleme yöntemi (5x2 Kat Çapraz Geçerleme, 10 Kat Çapraz Geçerleme, Monte Carlo Çapraz Geçerleme) sırayla kullanılarak bir öznitelik seçim yöntemi için üç sınıflandırıcı, üç çapraz geçerleme olmak üzere 9 çalıştırma gerçekleştirilmiştir. Yapılan 9 çalıştırmada da 29 deneme yapılmıştır. Bu sayede 29 deneme içinde "Seçilen öznitelik alt kümesinde değişiklik var mı, yok mu?" ve "Seçilen bu öznitelik alt kümesiyle sınıflandırma yapılırsa ortalama F1 Skoru ne olurdu?" sorularının cevabını elde etme şansımız olmuştur. Bu sonuçların elde edilmiş olmasıyla öznitelik seçim yöntemlerindeki yanlılık etkisi ASM ve USM metrikleri üzerinden incelenmiştir. Yanlılık etkisinin öznitelik seçimine etkisini azaltmak için literatürde önerilen çapraz geçerleme yöntemleri kendi aralarında kıyaslanmış ve veri setinin, öznitelik seçim yöntemlerinin, öznitelik seçimine nasıl bir etki uyguladığı gözlemlenmiştir.

"Breast Cancer" veri setinde ANOVA Filtre öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi, 10 Kat Çapraz Geçerleme ve Monte Carlo Çapraz Geçerleme olarak çıkmıştır. Aynı veri setinde mRMR öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme olarak çıkmıştır. Aynı veri setinde RFE öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme ve ve Monte Carlo Çapraz Geçerleme olarak çıkmıştır. Buradan çıkarım yaparsak "Breast Cancer" veri setinde yanlılık etkisinin öznitelik kümesindeki

çeşitliliği en az etkilediği çapraz geçerleme yönteminin 10 Kat Çapraz Geçerleme yöntemi olduğu görülmüştür.

“Diabetes” veri setinde ANOVA Filtre öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme yöntemi çıkmıştır. Aynı veri setinde mRMR öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme yöntemi çıkmıştır. Aynı veri setinde RFE öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme yöntemi çıkmıştır. Buradan çıkarım yaparsak “Diabetes” veri setinde yanlılık etkisinin öznitelik kümesindeki çeşitliliği en az etkilediği çapraz geçerleme yönteminin 10 Kat Çapraz Geçerleme yöntemi olduğu görülmüştür.

“Ionosphere” veri setinde ANOVA Filtre öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme çıkmıştır. Aynı veri setinde mRMR öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme yöntemi çıkmıştır. Aynı veri setinde RFE öznitelik seçim metodu için en kararlı çapraz geçerleme yöntemi 10 Kat Çapraz Geçerleme yöntemi çıkmıştır. Buradan çıkarım yaparsak “Ionosphere” veri setinde yanlılık etkisinin öznitelik kümesindeki çeşitliliği en az etkilediği çapraz geçerleme yönteminin 10 Kat Çapraz Geçerleme yöntemi olduğu görülmüştür.

Veri setimizdeki özniteliklerin doğasına göre yanlılık etkisinin sonucu farklılıklar göstermektedir. Kullanılan üç veri seti için bu farklılıkları azaltmak amacıyla kullanılan çapraz geçerleme yöntemleri içinde 10 Kat Çapraz Geçerleme yöntemi veri setinden bağımsız olarak en kararlı olan yöntemdir.

Anlamlılık testlerine baktığımızda “Breast Cancer”, “Diabetes” “Ionosphere” veri setlerinde kullanılan farklı çapraz geçerleme yöntemlerinde birbirine yakın ortalama F1 Skorları veren çapraz geçerleme yöntemleri farklı çıkmıştır. Buradan hareketle durumun veri setine de bağlı olduğu sonucu çıkmıştır.

Tez çalışması sonucunda öneri olarak vermek istenenleri kısaca özetlemek gerekirse, ileriki çalışmalarda öznitelik seçim yöntemlerini ve çapraz geçerleme yöntemlerini artırarak daha fazla kombinasyonla daha anlamlı sonuçlar elde edinilebilir.

KAYNAKLAR

[1] **Singhi, S. K., & Liu, H**. (2006). Feature subset selection bias for classification learning. In *Proceedings of the 23rd international conference on Machine learning* (pp. 849-856).

[2] **Krawczuk, J., & Łukaszuk, T.** (2016). The feature selection bias problem in relation to high-dimensional gene data. *Artificial intelligence in medicine, 66*, 63-71.

[3] **Maggipinto, T., Bellotti, R., Amoroso, N., Diacono, D., Donvito, G., Lella, E., ... & Alzheimer's Disease Neuroimaging Initiative.** (2017). DTI measurements for Alzheimer's classification. *Physics in Medicine & Biology, 62*(6), 2361.

[4] **Markowetz, F., & Spang, R.** (2005). Molecular diagnosis. *Methods of information in medicine, 44*(03), 438-443.

[5] **Park, H., & Kwon, H. C.** (2011). Improved Gini-index algorithm to correct feature-selection bias in text classification. *IEICE transactions on information and systems, 94*(4), 855-865.

[6] **Demircioğlu, A.** (2021). Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. *Insights into Imaging, 12*(1), 1-10.

[7] **Li, L., Neal, R. M., & Zhang, J.** (2008). A method for avoiding bias from feature selection with application to naive bayes classification models. *Bayesian Analysis, 3*(1), 171-196.

[8] **Munson, M. A., & Caruana, R.** (2009). On feature selection, bias-variance, and bagging. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20* (pp. 144-159). Springer Berlin Heidelberg.

[9] **Tran, B., Xue, B., Zhang, M., & Nguyen, S.** (2016). Investigation on particle swarm optimisation for feature selection on high-dimensional data: Local search and selection bias. *Connection Science, 28*(3), 270-294.

[10] **Ni, W., Xu, N., Dai, H., & Huang, S. H.** (2020). Reducing feature selection bias using a model independent performance measure. *International Journal of Data Science, 5*(3), 229-246.

[11] **Ambroise, C., & McLachlan, G. J.** (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences, 99*(10), 6562-6566.

[12] **Wood, I. A., Visscher, P. M., & Mengersen, K. L.** (2007). Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics, 23*(11), 1363-1370.

[13] **Jordan, M. I., & Mitchell, T. M.** (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260.

[14] **Cortes, C., & Vapnik, V.** (1995). Support-vector networks. *Machine learning, 20*, 273-297.

[15] **Huang, X., Shi, L., & Suykens, J. A.** (2013). Support vector machine classifier with pinball loss. *IEEE transactions on pattern analysis and machine intelligence, 36*(5), 984-997.

[16] **Belgiu, M., & Drăguţ, L.** (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31.

[17] **Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X.** (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[18] **Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H.** (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[19] **Cai, J., Luo, J., Wang, S., & Yang, S.** (2018). Feature selection in machine learning: A new perspective. *Neurocomputing, 300*, 70-79.

[20] **Ding, H., Feng, P. M., Chen, W., & Lin, H.** (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Molecular BioSystems, 10*(8), 2229-2235.

[21] **Johnson, K. J., & Synovec, R. E.** (2002). Pattern recognition of jet fuels: comprehensive GC× GC with ANOVA-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 60*(1-2), 225-237.

[22] **Sheikhan, M., Bejani, M., & Gharavian, D.** (2013). Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. *Neural Computing and Applications, 23*, 215-227.

[23] **Elssied, N. O. F., Ibrahim, O., & Osman, A. H.** (2014). A novel feature selection based on one-way anova f-test for e-mail spam classification. *Research Journal of Applied Sciences, Engineering and Technology, 7*(3), 625-638.

[24] **Peng, H., Long, F., & Ding, C.** (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence, 27*(8), 1226-1238.

[25] **Ding, C., & Peng, H.** (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology, 3*(02), 185-205.

[26] **Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z.** (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC bioinformatics, 18*(1), 1-14.

[27] **Ramírez-Gallego, S., Lastra, I., Martínez-Rego, D., Bolón-Canedo, V., Benítez, J. M., Herrera, F., & Alonso-Betanzos, A.** (2017). Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for

high-dimensional big data. *International Journal of Intelligent Systems, 32*(2), 134-152.

[28] **Guyon, I., Weston, J., Barnhill, S., & Vapnik, V.** (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*, 389-422.

[29] **Chen, X. W., & Jeong, J. C.** (2007). Enhanced recursive feature elimination. *Sixth international conference on machine learning and applications (ICMLA 2007)* (pp. 429-435). IEEE.

[30] **Url-1** <*https://is.gd/RZdZF1*>, erişim tarihi 01.04.2023.

[31] **Url-2** <*https://is.gd/QnwYFv*>, erişim tarihi 01.04.2023.

[32] **Url-3** <*https://is.gd/EUWXVH*>, erişim tarihi 01.04.2023.

[33] **Lustgarten, J. L., Gopalakrishnan, V., & Visweswaran, S.** (2009). Measuring stability of feature selection in biomedical datasets. *AMIA annual symposium proceedings* (Vol. 2009, p. 406). American Medical Informatics Association.

[34] **Street, W. N., Wolberg, W. H., & Mangasarian, O. L.** (1993). Nuclear feature extraction for breast tumor diagnosis. *Biomedical image processing and biomedical visualization* (Vol. 1905, pp. 861-870). SPIE.

[35] **Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S.** (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the annual symposium on computer application inmedical care* (p. 261). American Medical Informatics Association.

[36] **Sigillito, V. G., Wing, S. P., Hutton, L. V., & Baker, K. B.** (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest, 10*(3), 262-266.

[36] **Kim, T. K.** (2015). T test as a parametric statistic. *Korean journal of anesthesiology, 68*(6), 540-546.

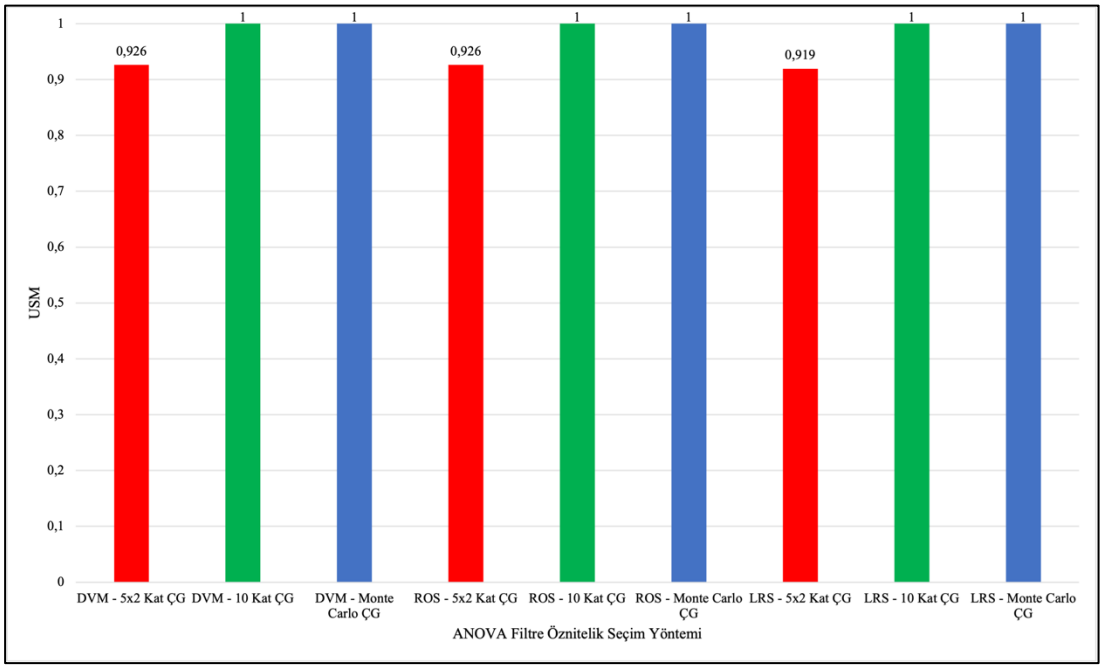
[37] **Gerald, B.** (2018). A brief review of independent, dependent and one sample t-test. *International journal of applied mathematics and theoretical physics, 4*(2), 50-54.

[38] **Thiese, M. S., Ronna, B., & Ott, U.** (2016). P value interpretations and considerations. *Journal of thoracic disease, 8*(9), E928.

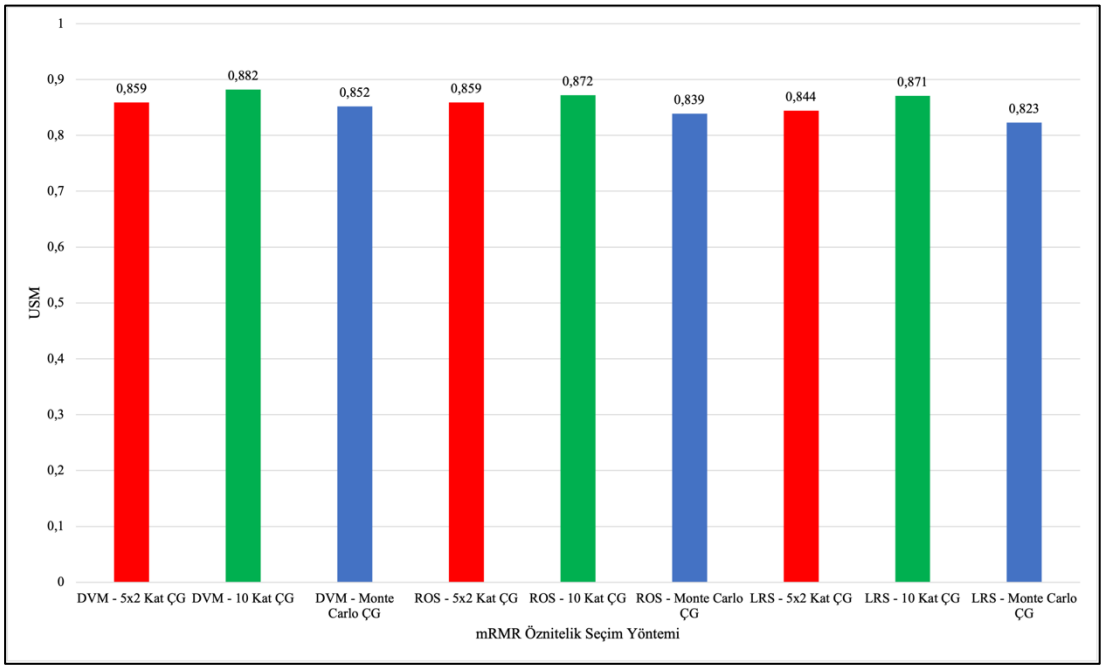
EKLER

EK A: Veri Seti ve Öznitelik Seçim Yöntemi Bazlı USM Bar Grafikleri

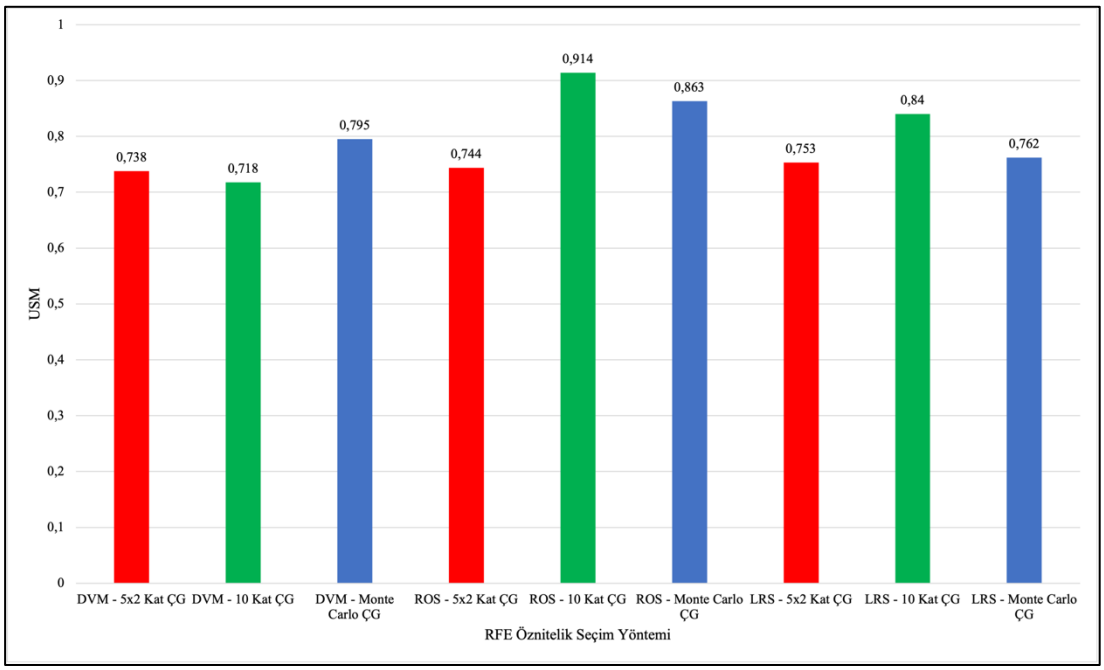
EK A



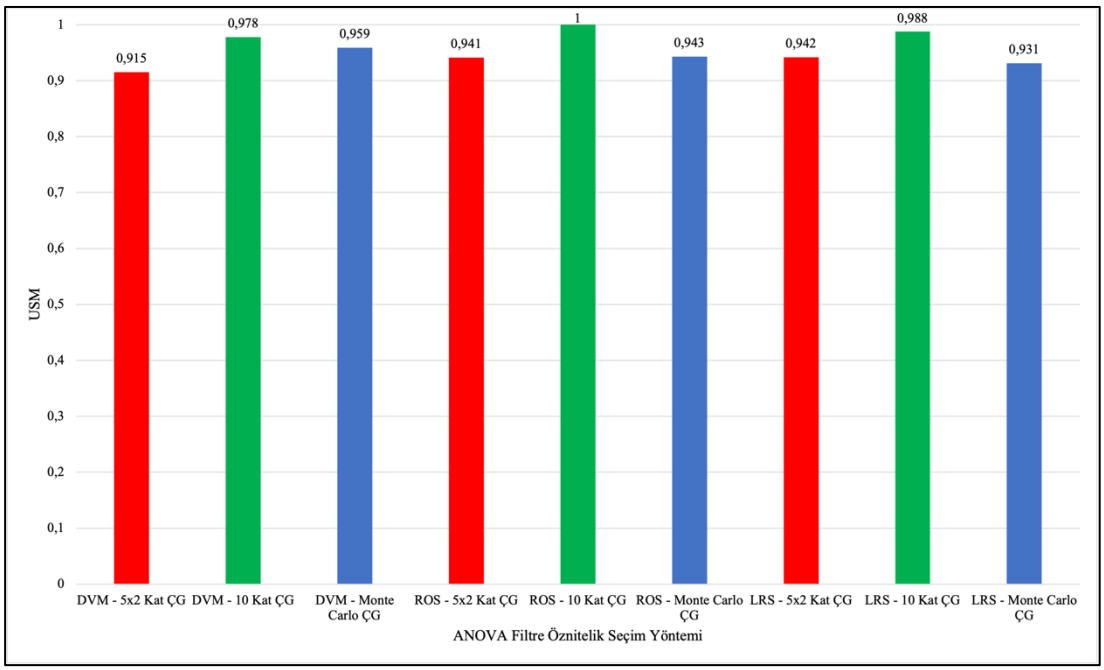
Şekil A.1 : "Breast Cancer" veri seti ANOVA Filtre öznitelik seçim yöntemi USM bar grafiği.



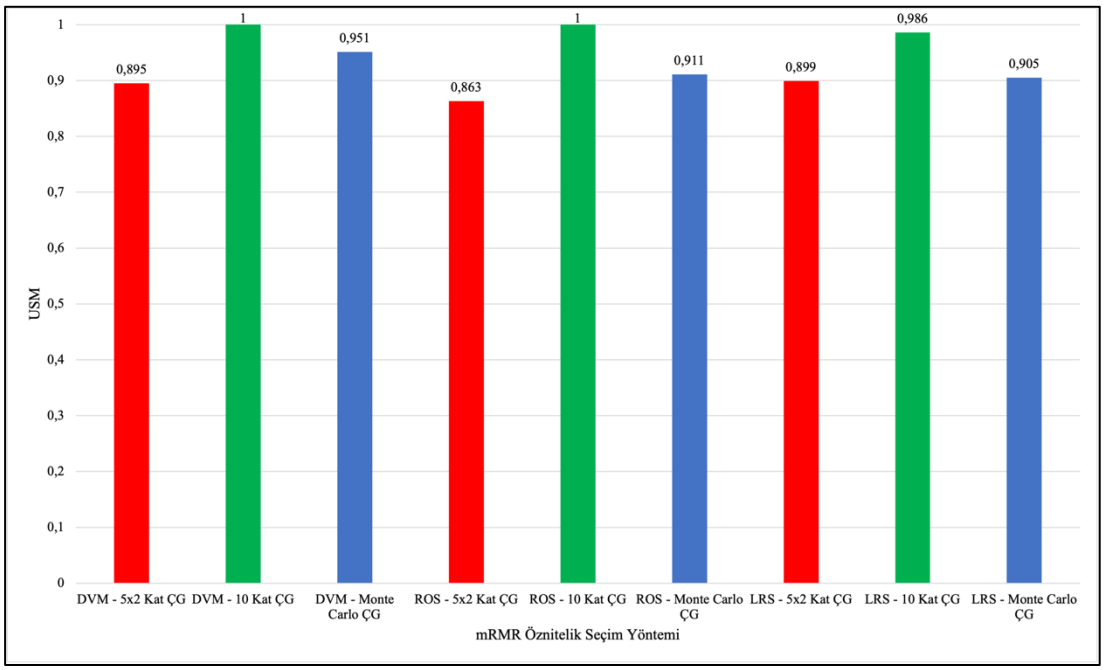
Şekil A.2 : "Breast Cancer" veri seti mRMR öznitelik seçim yöntemi USM bar grafiği.



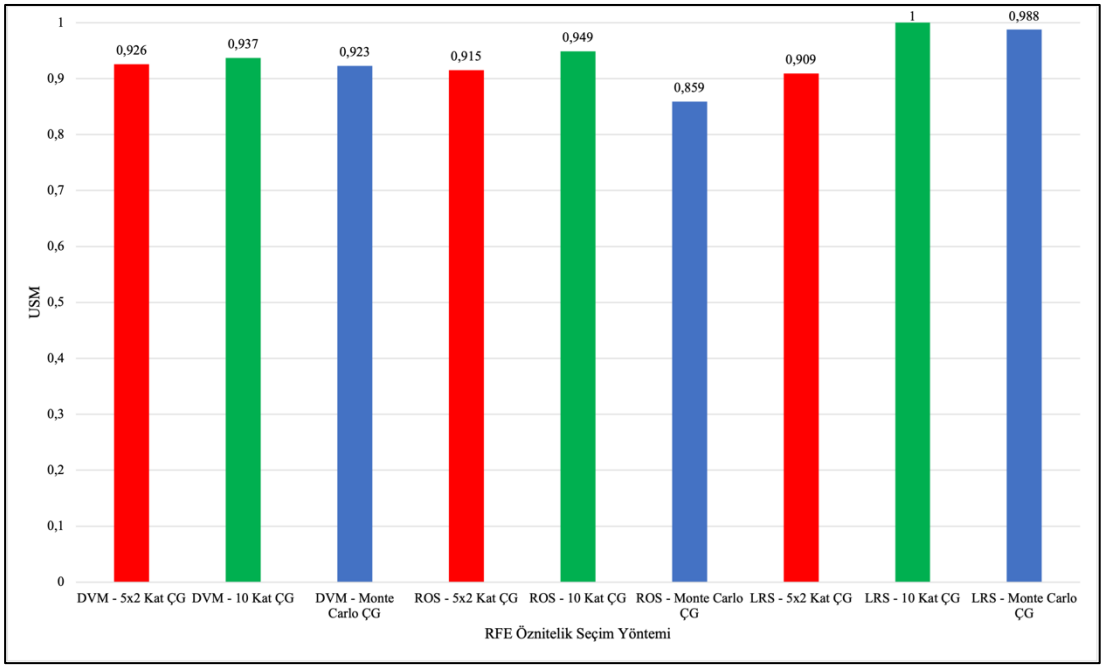
Şekil A.3 : "Breast Cancer" veri seti RFE öznitelik seçim yöntemi USM bar grafiği.



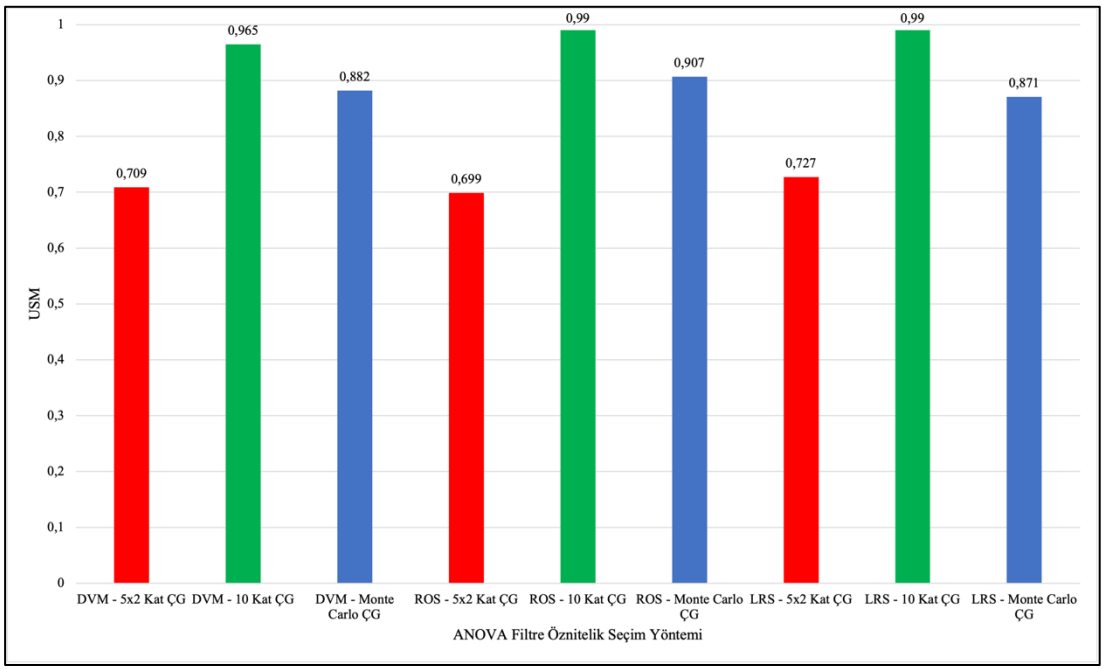
Şekil A.4 : "Diabetes" veri seti ANOVA Filtre öznitelik seçim yöntemi USM bar grafiği.



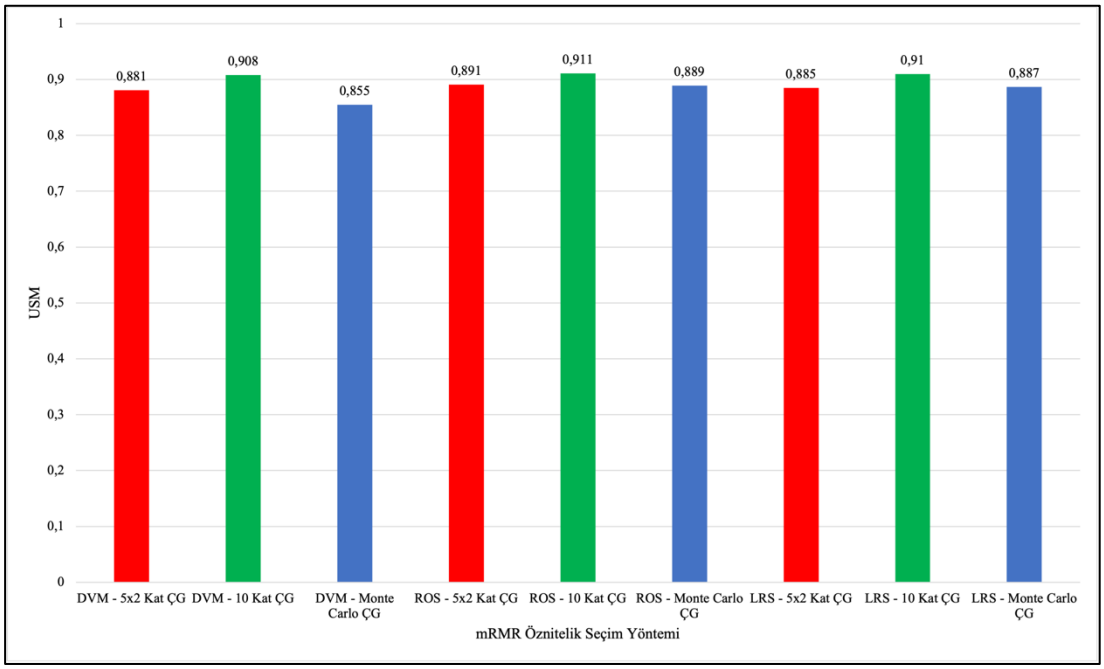
Şekil A.5 : "Diabetes" veri seti mRMR öznitelik seçim yöntemi USM bar grafiği.



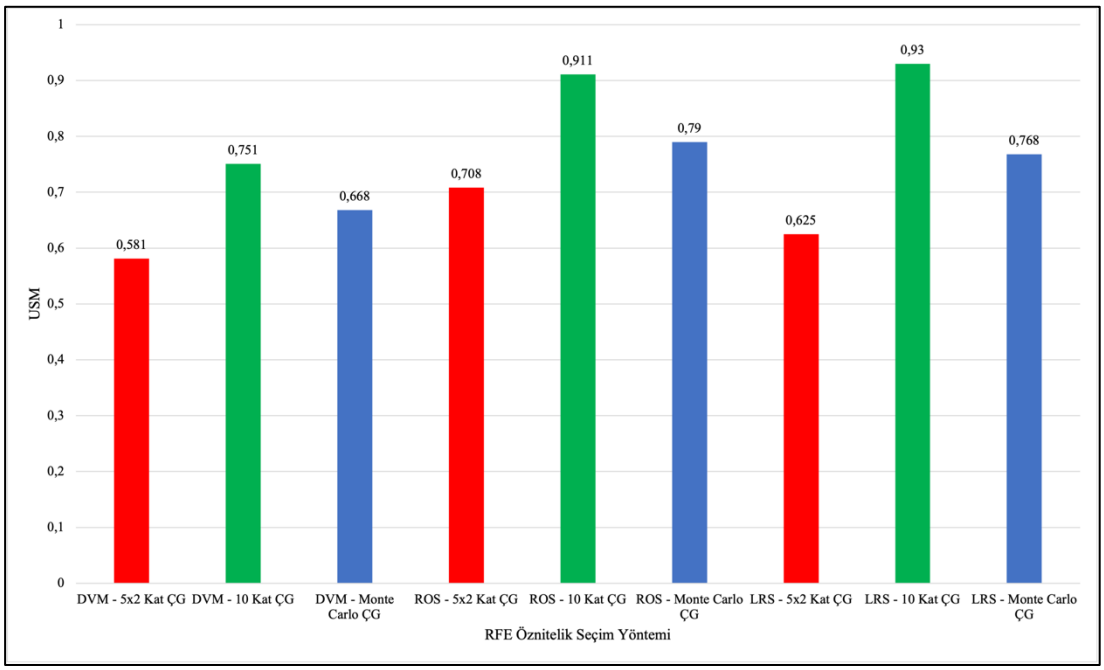
Şekil A.6 : "Diabetes" veri seti RFE öznitelik seçim yöntemi USM bar grafiği.



Şekil A.7 : "Ionosphere" veri seti ANOVA Filtre öznitelik seçim yöntemi USM bar grafiği.



Şekil A.8 : "Ionosphere" veri seti mRMR öznitelik seçim yöntemi USM bar grafiği.



Şekil A.9 : "Ionosphere" veri seti RFE öznitelik seçim yöntemi USM bar grafiği.

ÖZGEÇMİŞ

Ad-Soyad **:** Semih Can BOZOK

Doğum Tarihi ve Yeri **:**

E-posta **:**

ÖĞRENİM DURUMU:

- **Lisans** **:** 2018, Erciyes Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü

MESLEKİ DENEYİM VE ÖDÜLLER:

- Yozgat Bozok Üniversitesi (Araştırma Görevlisi), Kasım 2021 - Halen çalışıyor.
- ……..

TEZDEN TÜRETİLEN ESERLER, SUNUMLAR VE PATENTLER:

- …….
- ……..
- ……

DİĞER ESERLER, SUNUMLAR VE PATENTLER:

- …….
- ……..