
GAZİ ÜNİVERSİTESİ
Gazi
1926

GEN İFADESİ VERİLERİNE ÇOK KRİTERLİ KARAR VERME YÖNTEMLERİNİN UYGULANMASI

Meryem Gülşah PAMUK

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HAZİRAN 2019

Meryem Gülşah PAMUK tarafından hazırlanan "GEN İFADESİ VERİLERİNE ÇOK KRİTERLİ KARAR VERME YÖNTEMLERİNİN UYGULANMASI" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Bilgisayar Mühendisliği Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Dr. Öğr. Üyesi İsmail ATACAK

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ...……

Başkan: Doç. Dr. Necaattin BARIŞÇI

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Üye: Dr. Öğr. Üyesi Metin YILDIZ

Biyomedikal Mühendisliği Ana Bilim Dalı, Başkent Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 14/06/2019

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

……

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Meryem Gülşah PAMUK

14/06/2019

GEN İFADESİ VERİLERİNE ÇOK KRİTERLİ KARAR VERME YÖNTEMLERİNİN UYGULANMASI

(Yüksek Lisans Tezi)

Meryem Gülşah PAMUK

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2019

ÖZET

Gen ifadesi verileri analiz edilerek başta kanser olmak üzere genetik faktörlerin etkili olduğu hastalıkların teşhisinden tedavisine değin uzanan geniş bir yelpazede önemli bilgilere ve öngörülere ulaşılabilir. Gen ifadesi verilerinden hastalıklar hakkında anlamlı çıkarımlarda bulunmak için makine öğrenmesi temelli tahmin modellerinin uygulanması yaygın bir yaklaşımdır. Tahmin amaçlı yararlanılan bu tür modellerin yapısında uygun bir sınıflandırıcıya ek olarak uygun bir öznitelik seçim yönteminin yer alması başarı oranını artırır. Bu tezde kolon tümörü ve lenfoma gen ifadesi verileri üzerinde iki aşamalı bir çalışma gerçekleştirilmiştir. İlk aşamada, 2 çeşit sınıflandırıcının ve bilgi kuramı tabanlı 10 adet öznitelik seçim yönteminin dâhil edildiği bir değerlendirme süreci tesis edilmiştir. Sınıflandırıcılar her bir öznitelik seçim yöntemi ile ayrı ayrı kombine edilerek birbirinden farklı 20 tahmin modeli oluşturulmuştur. Tahmin modellerinin performansları Çok Kriterli Karar Verme (ÇKKV) disiplini çerçevesinde 5 farklı kritere göre değerlendirilmiştir. Bu amaçla, Analitik Hiyerarşi Süreci (AHS) ve Çok Kriterli Optimizasyon ve Uzlaşma Çözümü (Vise Kriterijumska Optimizacija I Kompromisno Resenje - VIKOR) yöntemlerini birleştirerek uygulayan bütünleşik AHS-VIKOR yöntemi kullanılmıştır. Değerlendirmeler sonucunda her bir veri kümesi için tahmin modellerinin uzlaşık bir sıralaması elde edilmiş ve sınıflandırma performansını optimize eden modeller belirlenmiştir. İkinci aşamada ise öznitelikleri değerlendirme görevinin VIKOR yöntemi ile ele alındığı hibrit bir öznitelik seçim mekanizması önerilmiştir. Önerilen yöntemde bilgi kuramı tabanlı öznitelik seçim yöntemlerinin değerlendirme fonksiyonları birer karar kriteri olarak kullanılmış olup, ortaya çıkan çok kriterli öznitelik seçim problemi VIKOR yöntemiyle çözüme kavuşturulmuştur. VIKOR tabanlı öznitelik seçiminde her sınıflandırıcı için en etkili sonuca ulaşılması amaçlanmış ve bunun için ilk aşama sonucunda elde edilen uzlaşık sıralamalardan yararlanılmıştır. Böylece tez çalışmasının iki ana aşaması birbiriyle ilişkilendirilmiştir. Kolon tümörü ve lenfoma verileri üzerinde 2 çeşit sınıflandırıcı temel alınarak gerçekleştirilen deneylerde, önerilen yöntem hâlihazırda kullanılan diğer öznitelik seçim yöntemleri ile mukayese edilmiştir. Karşılaştırmalarda sınıflandırma performansı ölçüt olarak alınmıştır. Sonuçlar önerilen yöntemin öznitelik seçimi ve sınıflandırma performansında kayda değer bir gelişme sağladığını göstermiştir.

Bilim Kodu	:	92431
Anahtar Kelimeler	:	Bilgi kuramsal öznitelik seçimi, bütünleşik AHS-VIKOR yöntemi, ÇKKV, makine öğrenmesi
Sayfa Adedi	:	152
Danışman	:	Dr. Öğr. Üyesi İsmail ATACAK

APPLICATION OF MULTIPLE CRITERIA DECISION MAKING METHODS TO GENE EXPRESSION DATA

(M. Sc. Thesis)

Meryem Gülşah PAMUK

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2019

ABSTRACT

Analysis of gene expression data can provide important information and predictions in a wide range from diagnosis to treatment for diseases such as cancer where genetic factors are effective. It is a common approach to apply machine learning based prediction models to make meaningful inferences about diseases from gene expression data. The presence of an appropriate feature selection method in addition to an appropriate classifier in the structure of such models used for prediction increases the success rate. In this thesis, a two-stage study is conducted on colon tumor and lymphoma gene expression data. In the first stage, an evaluation process including 2 types of classifiers and 10 information theory based feature selection methods is established. The classifiers are combined with each of the feature selection methods separately and 20 prediction models which are different from each other are constituted. The performances of the prediction models are evaluated according to 5 different criteria within the framework of Multiple Criteria Decision Making (MCDM) discipline. For this purpose, the integrated AHP-VIKOR method, which applies the combination of Analytic Hierarchy Process (AHP) and Multi-Criteria Optimization and Compromise Solution (Vise Kriterijumska Optimizacija I Kompromisno Resenje - VIKOR) methods, is used. As a result of the evaluations, a compromise ranking of the prediction models for each dataset is obtained and the models that optimize the classification performance are determined. In the second stage, a hybrid feature selection mechanism is proposed in which the task of evaluating features is addressed by VIKOR method. In the proposed method, the evaluation functions of information theory based feature selection methods are used as decision criteria and the resulting multi-criteria feature selection problem is solved by VIKOR method. In VIKOR based feature selection, it is aimed to reach the most effective result for each classifier, and for this purpose, the compromise rankings obtained as a result of the first stage are utilized. Thus, the two main stages of the thesis study have been interrelated with each other. In the experiments carried out by using 2 types of classifiers on colon tumor and lymphoma datasets, the proposed method is compared with other feature selection methods already used. In comparisons, classification performance is taken as the criterion. The results demonstrate that the proposed method provides a remarkable improvement in feature selection and prediction performance.

Science Code : 92431

Key Words : Information theoretical feature selection, integrated AHP-VIKOR method, MCDM, machine learning

Page Number : 152

Supervisor : Assist. Prof. Dr. İsmail ATACAK

TEŞEKKÜR

Çalışmalarım süresince kıymetli öneri ve yardımlarıyla tezimin şekillenmesine büyük katkı sağlayan danışman hocam, Sayın Dr. Öğr. Üyesi İsmail ATACAK'a en içten duygularımla teşekkür ederim. Akademik bilgi ve deneyimlerinden faydalandığım Gazi Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine saygı ve teşekkürlerimi sunarım. Bu süreçte beni yalnız bırakmayan ve daima destek olan babama, manevî desteği için kardeşim Emre PAMUK'a, dünyalar tatlısı yeğenlerim Zehra ile Berra'ya, ayrıca ilgi ve anlayışları için tüm aileme yürek dolusu teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklamalar
$\mathbf{P_e}$	Sınıflandırmada hata olasılığı
$\boldsymbol{\mu}$	Aritmetik ortalama
$\boldsymbol{\sigma}$	Standart sapma
$\mathbf{p(x)}$	Olasılık kütle fonksiyonu
$\mathbf{p(x\|y)}$	Koşullu olasılık kütle fonksiyonu
$\mathbf{p(x,y)}$	Bileşke olasılık dağılımı
$\mathbf{H(X)}$	Entropi
$\mathbf{H(X,Y)}$	Bileşik entropi
$\mathbf{H(X\|Y)}$	Koşullu entropi
$\mathbf{I(X;Y)}$	Karşılıklı bilgi
$\mathbf{I(X;Y\|Z)}$	Koşullu karşılıklı bilgi
$\mathbf{F}$	Aday öznitelik kümesi
$\mathbf{f}$	Aday öznitelik
$\mathbf{C}$	Hedef sınıf
$\mathbf{S}$	Seçili öznitelik alt kümesi
$\mathbf{\|S\|}$	Seçili alt kümenin eleman sayısı
$\mathbf{s}$	Seçili öznitelik
$\mathbf{s_1}$	Seçilen ilk öznitelik
$\mathbf{s_n}$	Seçilen son öznitelik
$\mathbf{J}$	Amaç fonksiyonu
$\boldsymbol{\beta}$	Fazlalık teriminin göreli önem ağırlığı
$\boldsymbol{\delta}$	Fazlalık kontrolünde kullanılan eşik değeri
$\mathbf{I(C;\{f,S\})}$	Birleşik karşılıklı bilgi
$\mathbf{U(f,C)}$	Simetrik belirsizlik
$\mathbf{CR(f,s)}$	Korelasyon oranı
$\mathbf{w(f)}$	Aday özniteliğin ağırlığı
$\mathbf{mF_S}$	Seçili özniteliklerin toplam fazlalık miktarı

Simgeler	Açıklamalar
MB_S	Seçili özniteliklerin hedef sınıfla toplam ilgisi
k	En yakın komşu sayısı
$d(x_i,y_i)$	Eğitim ve test örneği arası öznitelik mesafesi
$P(C_i)$	Belirli bir sınıfa ait önsel olasılık
$P(X)$	Test örneğinin önsel olasılığı
$P(X\|C_i)$	Sınıf koşullu olasılık
F_1	F-ölçütü
P	Pozitif sınıfın örnek sayısı
N	Negatif sınıfın örnek sayısı
$A[a_{ij}]_{n\times n}$	İkili karşılaştırma matrisi
a_norm_{ij}	Normalleştirilmiş ikili karşılaştırma değeri
$W[w_i]_{n\times 1}$	Kriter ağırlıklarını tutan vektör
λ_{max}	İkili karşılaştırma matrisine ait en büyük özdeğer
A_j	Belirli bir alternatif
f_{ij}	Karar matrisi elemanı
f_i^*	Belirli bir kritere ilişkin en iyi değer
f_i^-	Belirli bir kritere ilişkin en kötü değer
S^*	Çoğunluk kuralına göre karar verme stratejisi
R^*	Karşıt görüşlerin asgari pişmanlığı ile ilgili strateji
v	Çoğunluk kuralı stratejisinin ağırlığı
Q	Uzlaşık sıralama indeksi
DQ	Kabul edilebilir avantaj eşiği
K	Kriter seti
P_l	Artırımlı arama döngüsünde kullanılan karar matrisi
p_{ij}	Aday öznitelikle ilgili değerlendirme puanı

Kısaltmalar	Açıklamalar
A	Adenin
AHP	Analytic hierarchy process
AHS	Analitik hiyerarşi süreci

Kısaltmalar	Açıklamalar
AKKBÖS	Ağırlıklandırılmış koşullu karşılıklı bilgi öznitelik seçimi
AUC	Area under the curve
AUPRC	Area under the precision-recall curve
BDE	Binary differential evolution
BDF	Binary dragonfly
BKBM	Birleşik karşılıklı bilgi maksimizasyonu
C	Sitozin
cDNA	Komplementer (complementary) deoksiribonükleik asit
CI	Consistency index
CMI	Conditional mutual information
CMIFS	Conditional mutual information-based feature selection
CMIM	Conditional mutual information maximization
CR	Consistency ratio
ÇAKV	Çok amaçlı karar verme
ÇKKV	Çok kriterli karar verme
DAÖS	Dinamik ağırlıklandırma tabanlı öznitelik seçimi
DEA	Data envelopment analysis
DKBÖS	Düzgün bilgi dağılımı altında karşılıklı bilgi ile öznitelik seçimi
DN	Doğru negatif
DNA	Deoksiribonükleik asit
DP	Doğru pozitif
DRF-CFS	Distributed ranking filter approach employing correlation-based feature selection
DVM	Destek vektör makineleri
DWFS	Dynamic weighting-based feature selection
ELECTRE	Elimination and choice expressing reality
FICA	Fuzzy imperialist competitive algorithm
G	Guanin
GRA	Gray relational analysis
GSFSJNE	Gene selection algorithm with the fisher score based on joint neighborhood entropy
IWSSr	Incremental wrapper subset selection with replacement

Kısaltmalar	Açıklamalar
JMIM	Joint mutual information maximization
KB	Karşılıklı bilgi
KBF	Karşılıklı bilgi fark kriteri
KBO	Karşılıklı bilgi oran kriteri
KBÖS	Karşılıklı bilgi temelli öznitelik seçimi
K/B/Y	Kazanma/beraberlik/yenilgi
KKB	Koşullu karşılıklı bilgi
KKBM	Koşullu karşılıklı bilgi maksimizasyonu
KKBÖS	Koşullu karşılıklı bilgi tabanlı öznitelik seçimi
kNN	k-En yakın komşu (k-Nearest neighbor)
LOOCV	Leave-one-out cross-validation
MCC	Matthews correlation coefficient
MCDM	Multiple criteria decision making
MDL	Minimum description length
mFMİ	Minimum fazlalık maksimum ilgi
MI	Mutual information
MID	Mutual information difference criterion
MIFS	Mutual information based feature selection
MIFS-U	Mutual information feature selector under uniform information distribution
MIQ	Mutual information quotient criterion
MKK	Matthews korelasyon katsayısı
mRMR	Minimum redundancy maximum relevance
mRNA	Mesajcı ribonükleik asit
NB	Naive Bayes
NKBÖS	Normalleştirilmiş karşılıklı bilgi öznitelik seçimi
NMIFS	Normalized mutual information feature selection
PRC	Precision-recall curve
RCI	Random consistency index
RelaxMRMR	Relax minimum redundancy maximum relevance
RNA	Ribonükleik asit
ROC	Receiver operating characteristic

Kısaltmalar	Açıklamalar
SAW	Simple additive weighting
SVM-RFE	Support vector machines recursive feature elimination
T	Timin
TOPSIS	Technique for order preference by similarity to ideal solution
U	Urasil
VÇKÖS	VIKOR ile çok kriterli öznitelik seçimi
VIKOR	Vise kriterijumska optimizacija i kompromisno resenje
VZA	Veri zarflama analizi
WCMIFS	Weighted conditional mutual information feature selection
WEKA	Waikato environment for knowledge analysis
WRFS	Feature selection based on weighted relevancy
YN	Yanlış negatif
YP	Yanlış pozitif

1. GİRİŞ

Modern dünyada, bilimsel araştırma alanlarının yanı sıra günlük yaşamın da her detayında teknolojik gelişmeler belirleyici rol üstlenmektedir. Teknolojinin nüfuz ettiği her türlü faaliyet alanında üretilen veri miktarları büyük bir artış göstermiştir. Gelişen teknolojiyle birlikte bir yandan sürekli veri artışı olurken diğer yandan bu verilerin depolanmasına özen gösterilmiştir. Çünkü belirli bir uğraşı sonucu üretilmiş olan verilerin saklanması, sahip olunan bilgi ve birikimin artırılması için uyulması gereken ilk adımdır. Bu sayede daha sonraki adımlarda, mevcut verilerin anlamlandırılmasına yönelik çeşitli yöntemler uygulanabilmektedir.

Devasa büyüklükteki veriler içinden elde edilen anlamlı bilgiler, çözülmesi gereken sorunlara zamanında müdahale etmeyi ve fırsatlara kolaylıkla ulaşmayı sağlar. Bu bağlamda; akıllı cihaz teknolojileri, sosyal ağlar, finans, pazarlama, tıbbi tanılama, genetik bilgi analizi gibi pek çok alanda yaşanan gelişmeler depolanan verilerin etkin bir şekilde işlenmesi konusunu güncel hâle getirmiştir. Gelecekle veya bilinmeyenle ilgilenen tahminci modellerin geliştirilmesine kapı aralayacak büyüklükteki veriler; makine öğrenmesi, karar analizi, veri madenciliği, biyoenformatik gibi veri odaklı pek çok disiplince araştırılmaktadır.

Gen ifadesi verilerinin elde edilmesini kolaylaştıran mikrodizi teknolojisi ve bu alanda ortaya çıkan çeşitli yenilikler, insan sağlığının merkezindeki yerini koruyan genetik bilgiden hastalıklar hakkında tahminlerde bulunmayı mümkün kılmıştır. Aynı hastalığa sahip bireylerden ya da farklı kategorilerdeki hastalardan elde edilerek veri kümeleri şeklinde depolanan gen ifadeleri, karşılaştırmalı analizlere dayanan araştırmalara kaynaklık etmektedir. Gelişen teknolojiye rağmen, gen ifadesi verilerinin sistemli bir şekilde hastalardan toplanması uzun ve zahmetli bir uğraş gerektirir. Dolayısıyla veri kümeleri genellikle az sayıda örnekten oluşur. Bununla birlikte, gen ifadesi verileri öznitelik sayısı bakımından büyüklük göstermektedir.

Yüksek boyutlu verilerin iyi bilinen örneklerinden olan gen ifadesi verileri, hastalık teşhisinden tedavisine değin uzanan pek çok süreçte öngörülerde bulunmayı sağlar. Gen ifade verisi analizleri; hastalıkların teşhisinde, seyir sürecini tahmin etmede, hastalıkların

altında yatan genetik ve epigenetik mekanizmaların araştırılmasında, ilaç etkileşimlerinin tespitinde ve tedavi süreçlerinde önemli rol oynar.

Gen ifadesi verilerinden hastalıklarla ilgili tahminlerde bulunmaya yönelik araştırma süreçleri genellikle bir sınıflandırma problemi şeklinde ortaya koyulur. Bu tür problemlerin çözümü için; veri ön işleme, öznitelik seçimi ve sınıflandırma gibi makine öğrenmesinin konusu içinde olan birtakım yöntemlerden yararlanılır. Şunu not etmek gerekir ki öznitelik seçimi ve sınıflandırma yöntemleri; veri madenciliği, örüntü tanıma, görüntü işleme, biyoenformatik gibi birçok araştırma alanının alt başlıklarında yer almakla birlikte, bu tez çalışmasında makine öğrenmesi başlığı altında ele alınmaktadır.

Makine öğrenmesi sürecinde kullanılan algoritmaların başarımları her problem verisi için aynı değildir. Sınıflandırma problemlerinin çözümünde kullanılmış olan onlarca öznitelik seçici ve sınıflandırıcı yöntem bulunmaktadır. Ancak bu yöntemlerin arasında, üzerinde uzlaşılmış bir en iyi yoktur. Eldeki problem verisine hangi yöntemlerin en uygun olduğunun belirlenmesi için deneysel araştırmalara başvurulması gerekir. Bu amaç doğrultusunda, öznitelik seçimi ve sınıflandırma yöntemleri içinden belirli kombinasyonlarla oluşturulmuş tahmin modelleri gen ifadesi verilerine uygulanarak performans değerleri karşılaştırılabilir. Tez çalışması içinde sıkça tekrarlanan *tahmin modeli* kavramı, burada bahsedildiği gibi, sınıflandırma probleminin çözümüne yönelik olarak belirli bir öznitelik seçim yöntemi ile bir sınıflandırma yönteminin birlikte kullanımını ifade eden makine öğrenmesi temeline dayalı modellere karşılık gelmektedir.

Gen ifadesi verilerinden hastalıkları teşhis etmek veya hastayı ilgilendiren gelişmeleri öngörmek için makine öğrenmesi yaklaşımıyla tasarlanan modellerde, örnek verilerle eğitilen sınıflandırma algoritmaları test verileri üzerinde tahminci olarak kullanılır. Sınıflandırıcı algoritmaların eğitimi öncesinde veriler birtakım algoritmik hazırlık işlemlerine tâbi tutulur. Verilerden gürültünün etkilerini gidermek ve veri kalitesini artırmak için uygulanan hazırlık aşamasına veri ön işleme denmektedir. Ön işlemler sınıflandırma başarısını artırmada etkili olsa da yeterli değildir. Gen ifadesi verileri, çok fazla sayıdaki gene özgü ekspresyon değerlerinden oluşur. Dolayısıyla gen ifadesi verilerinin sınıflandırılmasında zorluk oluşturan unsurların başında boyutluluk gelir. Yüksek boyutlu verilerde öznitelik sayısının çokluğu, öğrenme sürecini karmaşıklaştırarak

sınıflandırma performansını düşürmektedir. Bu performans düşüşünden kaçınmak için öznitelik seçimi yapılır.

Bir mikrodizi deneyine dâhil edilmiş olan her bir gen, verilerin işlenmesi aşamasında ayrı bir öznitelik olarak işlem görür. Gen ifadesi verileri için gen seçme de denen öznitelik seçim işlemi sonucunda, veri kümesini en küçük boyutta ve en anlamlı şekilde ifade eden gen alt kümesinin bulunması amaçlanır. Bu amaca yönelik önerilen çok sayıda öznitelik seçim yöntemi vardır. Bu konuyla ilgili literatürde yer alan çalışmalar gözden geçirildiğinde, var olan yöntemlerin birbirlerine göre üstünlüklerini karşılaştıran birçok araştırmaya rastlanması mümkündür. Birkaç örnek vermek gerekirse: Kanserle alâkalı çeşitli gen ifadesi verileri üzerinde yapılan bir araştırmada, 8 farklı öznitelik seçim yönteminin; k-En yakın komşu (k-Nearest Neighbor - kNN), C4.5 karar ağacı, Naive Bayes (NB) ve Destek Vektör Makineleri (DVM) sınıflandırıcıları ile elde ettiği performans değerleri karşılaştırılmıştır [1]. İyi bilinen 7 mikrodizi veri kümesi üzerinde 3 gen seçim tekniği ve 21 farklı sınıflandırma yönteminin uygulanması suretiyle gerçekleştirilen başka bir araştırmada, sınıflandırıcılar 4 ayrı kategoride gruplandırılarak değerlendirilmiştir [2]. Bu amaçla her veri kümesinin, 3 farklı öznitelik seçim işlemi sonunda ayrı yöntem kategorilerindeki sınıflandırıcılarla ulaştığı hata oranları karşılaştırılmıştır. Bir diğer çalışmada glioma gen ifadesi verileri üzerinde; DVM, kNN ve rastgele orman sınıflayıcıları ile 8 öznitelik seçim yönteminin etkinliği değerlendirilmiştir [3]. Literatürde rastlanan benzer birçok çalışmada mikrodizi gen ifadesi verileri sınıflandırılmış ve öznitelik seçiminin, tahmin doğruluğunu artırmadaki önemi vurgulanmıştır. Dahası bu tür araştırmaların sonuçlarına bakıldığında, aynı veriler ve sınıflayıcılar kullanılsa dahi farklı öznitelik seçicilerin seçtiği gen setlerinin farklılık göstermesine bağlı olarak tahmin doğruluğunun değiştiği rahatlıkla gözlemlenebilir [1, 4]. Kanıksanmış olan bu durum, öznitelik seçiminin sınıf tahminindeki belirleyiciliğini ortaya koymaktadır.

Gen ifadesi verilerinin sınıflandırılmasına ilişkin problem alanında kanser verileri çok sık işlenmiştir. Bilindiği üzere kanser, günümüzün en ölümcül hastalıkları arasındaki yerini korumaktadır. Küresel Kanser Gözlemevi'nin istatistiklerine göre 2018 yılında, başta; akciğer, göğüs, kolorektum, prostat ve mide kanserleri olmak üzere 18 078 957 yeni kanser vakası ortaya çıkmıştır [5]. Kansere bağlı tahmini ölüm sayısı 9 555 027 olarak bildirilmiştir [5]. Kanserli dokularda genetik işleyişin bozulması nedeniyle gen

ifadelerinde ortaya çıkan değişimler; hastalığın tanısı, seyri ve alt tiplerinin belirlenmesi hakkında bilgi veren önemli işaretçilerdir. Bu nedenle, gen ifadesi verileri üstünde gerçekleştirilen analizler kanserle ilgili araştırmalarda önemli yer tutmaktadır.

Belirli bir problem verisi için makine öğrenmesi alanındaki çeşitli yöntemler arasında kıyas yapmak veya bunlar arasından en uygun olanları belirlemek maksadıyla Çok Kriterli Karar Verme (ÇKKV) yaklaşımından yararlanılabilir. ÇKKV birden çok kriterin etkisi altında şekil alan karar problemlerini incelemektedir. ÇKKV problemlerinin çözümünde uygulanan etkili yöntemlerden biri; Analitik Hiyerarşi Süreci (AHS) ve Çok Kriterli Optimizasyon ve Uzlaşma Çözümü (Vise Kriterijumska Optimizacija I Kompromisno Resenje - VIKOR) yöntemlerini bir araya getiren bütünleşik AHS-VIKOR yöntemidir. Bu yöntemde AHS ile karar kriterlerinin ağırlıkları belirlenirken, VIKOR yöntemi ile alternatif çözümler çoklu kriterlere göre karşılaştırılarak uzlaşık sıralama ve uzlaşık çözümler elde edilir. Bütünleşik AHS-VIKOR yaklaşımı ve bulanık uzantıları; yenilenebilir enerji plânlaması, yer seçimi, proje seçimi, tedarikçi seçimi gibi çeşitli alanlarda ÇKKV problemlerinin çözümü için uygulanmıştır [6-9]. Bu uygulamalarda bütünleşik AHS-VIKOR yönteminin belirli bir ÇKKV problemi çerçevesinde modifiye edilerek kullanımı öne çıkmıştır.

Bu tez çalışmasında makine öğrenmesi ve ÇKKV disiplinlerinin bir araya getirilişi üzerinde durulmuş ve bu yöndeki çalışmalar örnek alınarak gen ifadesi verileri üzerinde ayrıntılı bir araştırma sistemi ortaya konmuştur. Kanser hastalığı ile ilgili olan 2 farklı gen ifadesi verisi makine öğrenmesi ve ÇKKV disiplinleri çerçevesinde analiz edilmiştir. İki genel aşaması olan bu çalışmanın ilk aşamasında, kanser verilerinin sınıflandırılması probleminin çözümünde kullanılabilecek olan makine öğrenmesi temelli çeşitli modeller performans ölçümlerine göre değerlendirilmiştir. Bu doğrultuda, bütünleşik AHS-VIKOR yöntemi kullanılarak değerlendirme sürecinin birden çok ölçütün katılımına uygun şekilde plânlanması sağlanmıştır. İkinci aşamada ise VIKOR yönteminden yararlanılarak yeni bir öznitelik seçim stratejisi geliştirilmiştir. Deneysel uygulamada; kolon tümörü ve lenfoma gen ifadesi veri kümeleri kullanılmıştır.

İlk aşama kapsamında bütünleşik AHS-VIKOR yöntemi, literatürdeki kullanımına benzer bir şekilde, gen ifadesi verilerine uygulanan tahmin modellerinin değerlendirilmesi amacıyla modifiye edilmektedir. Bu doğrultuda öncelikle ÇKKV ortamını oluşturan karar

bileşenleri belirlenir. Alternatifler olarak belirlenen makine öğrenmesi temelli 20 farklı tahmin modeli; 10 ayrı bilgi kuramsal öznitelik seçim yöntemi ve birbirinden farklı algoritmik işleyişe sahip popüler 2 sınıflandırma yönteminin ikili eşleşmeleri şeklinde ifade edilmektedir. Sınıflandırmada; kNN ve NB yöntemleri esas alınmıştır. İlgili veri kümeleri için sınıflandırma çözümünde en iyi sonuçlara ulaşan alternatiflerin tespiti, karar kriterlerini teşkil eden 5 farklı performans metriğine göre gerçekleştirilir. Bu süreçte, karar kriterlerinin ağırlıkları AHS yöntemi ile hesaplandıktan sonra aynı ağırlıklar tüm veri kümeleri için VIKOR yöntemine girdi değerleri olarak gönderilmektedir. AHS'den elde edilen ağırlıklardan yararlanılarak VIKOR yöntemi ile oluşturulan sıralama ve elde edilen uzlaşık çözümler ilk aşamanın çıktılarıdır. Tasarlanan bu değerlendirme süreci, gen ifadesi verilerinin her biri için ayrı ayrı uygulanmıştır.

Bu tez çalışmasının ikinci aşamasında, bilgi kuramı tabanlı kriterleri kullanan çok kriterli bir öznitelik seçim yaklaşımı önerilmektedir. Bu aşama, VIKOR yöntemi ekseninde şekillendirilmiştir. Önerilen yöntem öznitelikleri değerlendirmek için; bilgi kuramsal öznitelik seçim yöntemlerinin amaç fonksiyonlarını karar kriterleri olarak kullanmakta ve bu kriterleri VIKOR yöntemi ile bütünleştirmektedir. Önerilen yöntem, VIKOR ile Çok Kriterli Öznitelik Seçimi (VÇKÖS) olarak adlandırılmıştır. Bu yöntemle öznitelik seçimi gerçekleştirilirken birden fazla bilgi kuramsal kritere bağlı kalınır. Burada dikkat çekilmesi gereken nokta, bir sınıflandırma problemine uygulanan çözüm yaklaşımına göre iyi performans gösteren kriterlerin genellikle farklılık göstermesidir. Dolayısıyla, 2 veri kümesi üzerinde 2 ayrı sınıflandırma yöntemiyle gerçekleştirilmesi amaçlanan 4 farklı sınıflandırma görevi söz konusu olup; VÇKÖS yönteminin tasarımında veri kümesi ve sınıflandırıcı yöntem farkının dikkate alınması gerekir. VÇKÖS yöntemi geliştirilirken bu husus dikkate alınmış ve 4 farklı sınıflandırma çözümünün her biri için en uygun olan kriter seti ayrı ayrı belirlenmiştir. Tüm sınıflandırma görevlerinin ifası için toplamda 4 farklı kriter seti elde edilmiştir. Bu süreçte çoklu kriterler ilk aşamada gerçekleştirilen deneylerin sonuçlarına bakılarak seçilmiştir. Kısaca bahsetmek gerekirse kriter belirleme işi, 2 ayrı veri kümesinin her biri için mevcut olan uzlaşık sıralamalar göz önüne alınarak yapılır. Sıralama sonuçları sınıflandırıcı yöntemler bazında ayrıştırılıp, kNN ve NB yöntemleri ile ilintili tahmin modellerini içeren 2 ayrı sıralı liste elde edilir. Bu listeler ilgili veri kümesi üzerinde kNN ve NB yöntemleri ile gerçekleştirilen sınıflandırma çözümlerinde daha etkili sonuçlara ulaşmayı sağlayan öznitelik seçim kriterlerinin hangileri olduğu hakkında bilgi verir. Bu bilgiler ışığında sınıflandırma çözümlerinin her

biri için en uygun olan öznitelik seçim kriterleri belirlenir. VIKOR ise bu kriterlerle gerçekleştirilen bireysel değerlendirmeler arasında uzlaşı tesis edip, öznitelikleri seçmek için kullanılmıştır. Özet olarak ikinci aşamada; kolon tümörü ve lenfoma verileri üzerinde kNN ve NB yöntemleri ile isabetli sınıf tahmini yapabilmek için en uygun öznitelik seçim kriterleri belirlenmekte ve bilgi kuramsal kriterlerin ortaklaşa kararını VIKOR yöntemine göre bulan çok kriterli bir öznitelik seçim yöntemi önerilmektedir. Bu yöntemin 2 ayrı sınıflandırıcı ile elde ettiği performans her bir veri kümesi üzerinde test edilmiştir. Önerilen yöntemle ulaşılan sınıflandırma performansının çoğu durumda diğer bilgi kuramsal öznitelik seçicilerin elde ettiği performans değerlerinden daha yüksek olduğu deneylerle gösterilmiştir.

Bu tez çalışmasının benzer çalışmalardan ayrılmasını sağlayan en belirgin yönü; gen ifadesi verileri üzerinde tahmin performansını artırmak için bilgi kuramsal öznitelik seçim kriterleri ve VIKOR yöntemini temel alan hibrit bir öznitelik seçim yaklaşımı geliştirilmesidir. Bu sayede öznitelik seçimiyle ilgili araştırmalara katkıda bulunulmuştur. Her iki disiplini kapsamı içine alan bu yeni yöntem ile başarılı sonuçlar elde edilmiştir.

Tez çalışmasının ilerleyen bölümleri aşağıda belirtildiği şekilde organize edilmiştir. Bölüm 2'de; gen ifadesi ve mikrodiziler hakkında bilgi verilmiş ve gen ifadesi verilerinin makine öğrenmesi yöntemleriyle işlenişi üzerinde durulmuştur. Öznitelik seçimine ilişkin tafsilâtlı bilgi verilmiştir. Bölüm 3'te; karar analizi kapsamında yer alan başlıca kavramlar açıklanmış, karar verme teorik anlamda genel hatlarıyla ele alınmış ve ÇKKV süreci anlatılmıştır. Ayrıca makine öğrenmesi ve ÇKKV disiplinlerinin birlikte kullanımına ilişkin literatüre değinilmiştir. Bölüm 4'te tez çalışması kapsamında kullanılan materyaller ve metotlar açıklanmıştır. Bu bölümde; veri kaynakları, araştırmada kullanılan yöntemler, araştırma prosedürünün çerçevesi ve önerilen öznitelik seçim yaklaşımının yapısı hakkında detaylı bilgiler verilmiştir. Bölüm 5'te deneysel çalışmalar ve çalışmalardan elde edilen sonuçlar ayrıntılı şekilde açıklanmıştır. Ayrıca önerilen yöntemin başarısı çalışma kapsamında kullanılan bilgi kuramsal öznitelik seçicilerle karşılaştırılmıştır. Karşılaştırma sonuçları çizelgeler ve grafikler ile görsel değerlendirmeye sunulmuştur. Bölüm 6'da sonuçlara yer verilmiştir.

2. GEN İFADESİ VERİLERİ VE MAKİNE ÖĞRENMESİ

2.1. Gen İfadesi ve Mikrodizi Teknolojisi

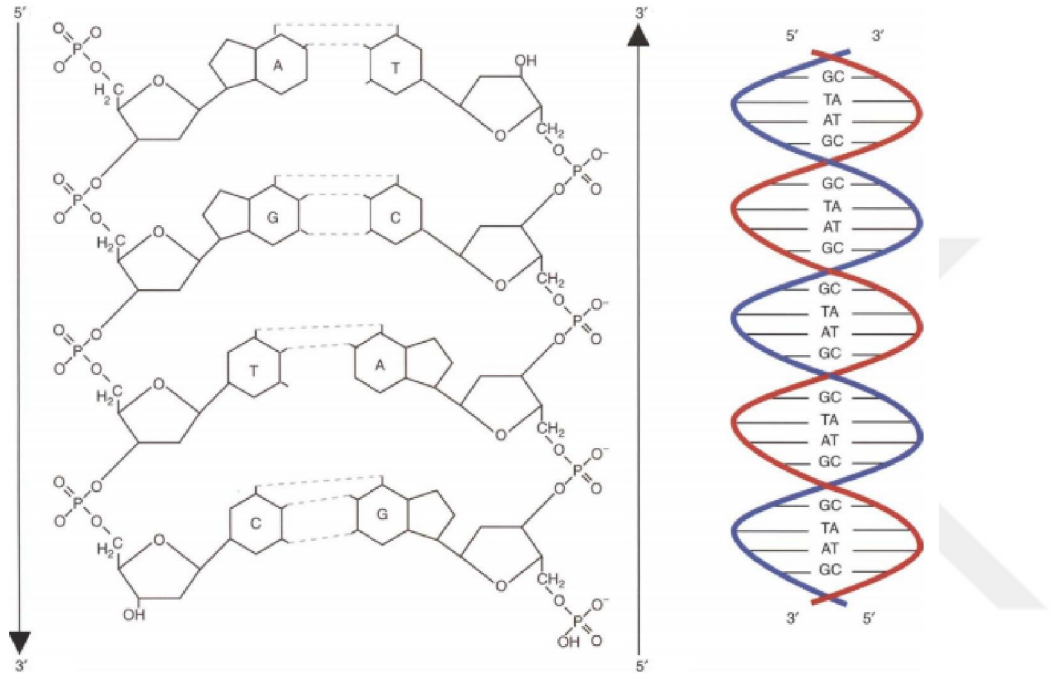
Hücre yapısı içinde Deoksiribonükleik Asit (DNA) ve Ribonükleik Asit (RNA) olmak üzere iki çeşit nükleik asit vardır ve her bir nükleik asit, baz olarak bilinen küçük yan grupların bağlı olduğu şeker ve fosfat molekülü zincirinden oluşur [10]. Bu uzun polimerik zincirin moleküler yapı birimi; bir fosfat, bir şeker ve bir azotlu organik baz bileşimini içeren nükleotidlerdir. DNA ve RNA dizilimindeki nükleotidler, yan gruplar olarak eklenen dört farklı azotlu bazdan birinde farklılık gösterir [10]. Her ikisi de Adenin (A), Guanin (G) ve Sitozin (C) içerir, ancak DNA'da dördüncü baz olarak Timin (T) bulunurken RNA'nın yapısında T yerine Urasil (U) bazı yer alır [10].

Nükleik asitler çift sarmallı bir yapıya sahiptir. Çift sarmallı yapı; biri diğerine tamamlayıcı, başka bir deyişle komplementer olan iki polinükleotid iplikçik içerir. Nükleik asidin bir iplikçiğinde bulunan nükleotid yapı taşlarının sırası biliniyorsa diğer iplikçikteki nükleotid yapı taşlarının sırası da bilinir; çünkü hidrojen bağları aracılığıyla A daima karşı iplikçikteki T bazı ile eşleşirken, C ise her zaman G bazına bağlanır [11]. RNA'da ise A her zaman U bazı ile eşleşir. Yani bir sarmalın zincirleri, Şekil 2.1'de gösterildiği gibi, birbirini bütünleyici anti-paralel dizgeye sahiptir.

Bir nükleik asit parçasının spesifikliği, yapısında bulunan bazların dizilimi tarafından ifade edilir ve bu baz dizilimi belirli bir proteinin aminoasit dizisini belirten basit bir koddur [12]. DNA genetik materyalin en değişmez parçasıdır ve protein sentezi genler tarafından kontrol edilir [12]. Hücre ve dokularda canlılığı sağlayan biyolojik süreçlerin sürdürülmesi ve düzenlenmesi amacıyla çeşitli işlevlere sahip olan genler, organizmaların kalıtsal bilgisini taşıyan DNA molekülü üzerinde belirli uzunluktaki nükleotid dizileri halinde bulunur. Kısa ve öz bir tanımlama yapmak gerekirse genler, RNA ve protein sentezinde kullanılacak olan kodların tutulduğu işlevsel DNA parçalarıdır.

Her bir gen, kendine has biyolojik anlamı olan bir DNA sekansıdır. Bir genin baz sekansı, gen ürünlerinin sentezlenmesi için gerekli olan bir tür başlangıç verisidir. Moleküler biyolojinin genel kabulü olan santral dogmaya [13] göre, DNA'dan RNA'ya ve RNA'dan

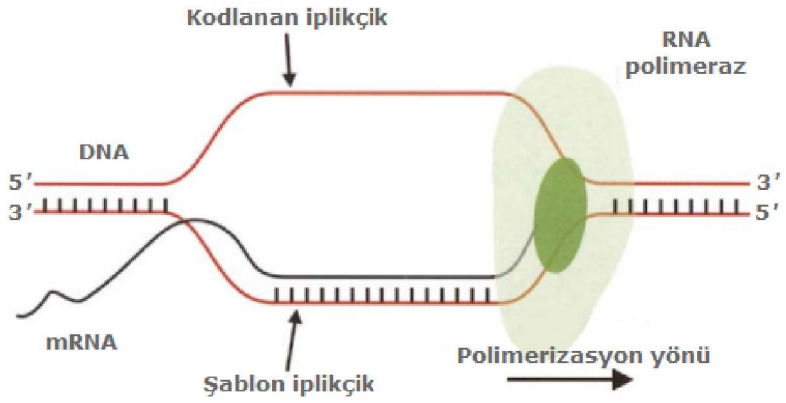
proteine doğru biyolojik bilgi akışı gerçekleşir. Bu bilgi akışının önemi; gen ürünlerinin elde edilmesinde rol oynamasıdır. Genin gerektiğinde aktifleşmesi ve sakladığı şifreden gen ürünlerinin sentezlenmesi o genin ifade edilmesidir. Gen ekspresyonu, gen anlatımı ya da daha yaygın kullanılan şekliyle gen ifadesi sırasında gerçekleşen biyolojik süreç transkripsiyon ve translasyon adımları ile tanımlanır.



Şekil 2.1. Çift iplikli bir helis oluşturmak üzere birbirine sarılmış iki polinükleotid zinciri içeren insan genomik DNA'sı [14]

Transkripsiyon, helisel yapıdaki eşleşmiş iki sarmaldan oluşan DNA molekülünün ilgili gen biriminde sarmalların açılması ve tek iplikten genetik kodun mesajcı Ribonükleik Asit (mRNA) sentezi yoluyla kopyalanmasıdır. Birbiri ile eşleşmiş olan iplikçikler komplementer yapıda olduğu için, DNA'dan mRNA'ya kodlanacak dizilim bilgisi bulunduğu iplikten değil kodlanmayan karşı iplikten üretilir. Bu iplikçik kalıp veya şablon olarak anılır. RNA polimeraz II enzimi yardımıyla gerçekleşen transkripsiyon işlemi Şekil 2.2'de basitçe gösterilmiştir. Transkripsiyon ürünü olan mRNA, sentezlenecek proteinin yapı taşları olan aminoasitlerin dizilimini bildiren kodu içermektedir. Bu kodu oluşturan üçlü baz gruplarına kodon denir. Protein sentezi için gerekli bilgi, mRNA ile hücre çekirdeğinden sitoplazmaya taşınır. Translasyon; protein sentezinden sorumlu olan ribozom organelinde, kodon bilgilerine karşılık gelen aminoasitlerin birleştirilerek aminoasit zincirinin oluşturulmasıdır. Özetle; transkripsiyonla DNA'dan elde edilen

taşınabilir genetik bilgi translasyon adımında, protein üretimi için gerekli bilgiye çevrilerek kullanılır.



Şekil 2.2. Bir DNA sekansının, bir komplementer RNA zinciri üretmek üzere transkripsiyon işlemiyle kopyalanması [14]

Bir organizmada bulunan genlerin bütününe genom denir. Canlının farklı yaşamsal faaliyetlerinin yönetildiği farklı doku hücrelerinde, genomun etkinlik haritası değişmektedir. Yani, belirli görevleri yerine getirmek için özelleşmiş organ, doku ve hücre tiplerinde işlevsellik gösteren genler farklıdır. Çünkü bir genin ifade edilmesi belli bir biyolojik faaliyetin gerçekleştirilmesine yöneliktir. Organizmayı etkileyen faktörlerin çoğu, genlerin faaliyetlerini de etkiler. Farklılaşan hücresel, çevresel veya metabolik koşullara göre ihtiyaç olan gen ürünlerini elde etmeye veya gereksiz olan gen ürünlerinin çoğalmasını önlemeye yönelik kontrol ve düzenlemeler gen ifadesinin regülasyonu olarak adlandırılır. Gen regülasyonu ile hücrelerdeki genlerin ifade düzeylerinin kontrolü; transkripsiyonel, transkripsiyon sonrası, translasyonel veya translasyon sonrası seviyelerde çeşitli moleküller yardımıyla sağlanabilir [15].

Gen aktivitesinde genetik faktörlere bağlı değişimler olduğu gibi, çevresel şartlara bağlı transkript farklılıkları da görülebilir. Mutasyon denen genetik hatalar ve genetik çeşitliliği ifade eden polimorfizmler dışında, DNA'nın nükleotid sekansına bağlı olmayan ancak kalıtımla aktarılan gen ifadesi değişiklikleri, genetik üstü anlamına gelen epigenetiğin konusudur [16]. Epigenom; insan genomuyla ilişkili olan, DNA'ya bağlanabilen, genleri açma veya kapama gibi faaliyetleri yönlendiren ve hücrelerdeki gen ekspresyonunu düzenleyen kimyasal bileşik ve proteinlerdir [17]. Genetik yapı ve çevresel etmenlerin katkısıyla ortaya çıkan fenotip, canlının dışsal biyolojik özellikleridir. İnsan vücudundaki

400 doku ve hücre tipinin her birinin fenotipi, sahip olduğu transkribe genlerin kendine özgü ekspresyon deseni ile tanımlanır [18]. Maruziyetler, yaşam stili, beslenme, diyet, spor gibi çevresel unsurların fenotipik etkilerle gen ifadelerinde yol açtığı kalıcı değişiklikler, epigenetik modifikasyonlara örnek verilebilir. Bu modifikasyonlar genetik değişikliklerle tetiklenenlere eşdeğer işlevsel sonuçlar doğurur [19]. O hâlde epigenetik modifikasyonların bazı hastalıkların gelişimiyle olan ilişkisi şaşırtıcı değildir. Kanser değişmiş epigenetik desenle bağlantılı başlıca hastalıklar arasındadır [18].

Kanser; genetik ve epigenetik hatalar nedeniyle normal bir hücrenin zararlı, yayılmacı ve metastatik bir tümör hücresine dönüşmesidir [16]. Kompleks bir hastalık olan kanserin altında yatan sebepler henüz tam olarak gün yüzüne çıkarılamamış olsa da; bazı genlerin nükleotid sekansındaki polimorfizmler, genlerin işlevini yitirmesi, mutasyonlar ve epigenetik yatkınlık gibi faktörlerin kanser gelişiminde etkili olduğu bilinmektedir. Bu faktörlerin zararları, kontrolsüz hücre bölünmesi ve kanserle sonuçlanabilecek düzeye ulaşabilir. Hücre bölünmesinin denetlenememesi sonucu anormal büyümenin olduğu bölgelerde bezecik veya yumru şeklinde tümör denen kitleler oluşur. Tümör hücrelerinin kan ve lenf yoluyla vücuda yayıldığı metastatik olgularda kanser başka dokuları ele geçirme potansiyeline sahiptir.

Gen ifadesinin regülasyonu, yaşamsal ihtiyaçlara ve koşullara göre organizmanın genomik faaliyetlerini düzene koymak ve optimize etmek için kritik önemdedir. Gen ifadesi değerlerinin alışılmışa aykırı olan değişimleri kanser hastalığı ile yakından ilgilidir. Gen regülasyonunda meydana gelen genetik veya epigenetik kaynaklı bozulmalar; genlerin ihtiyaç duyulmayan hücrelerde, normalden farklı düzeylerde ya da anormal zamanlarda ifade edilmesine neden olur. Fenotipi değişmiş ya da kanserleşmiş hücrelerdeki ekspresyon desenleri bozulan genler kanser patogenezi ve prognozu hakkında bilgi verir.

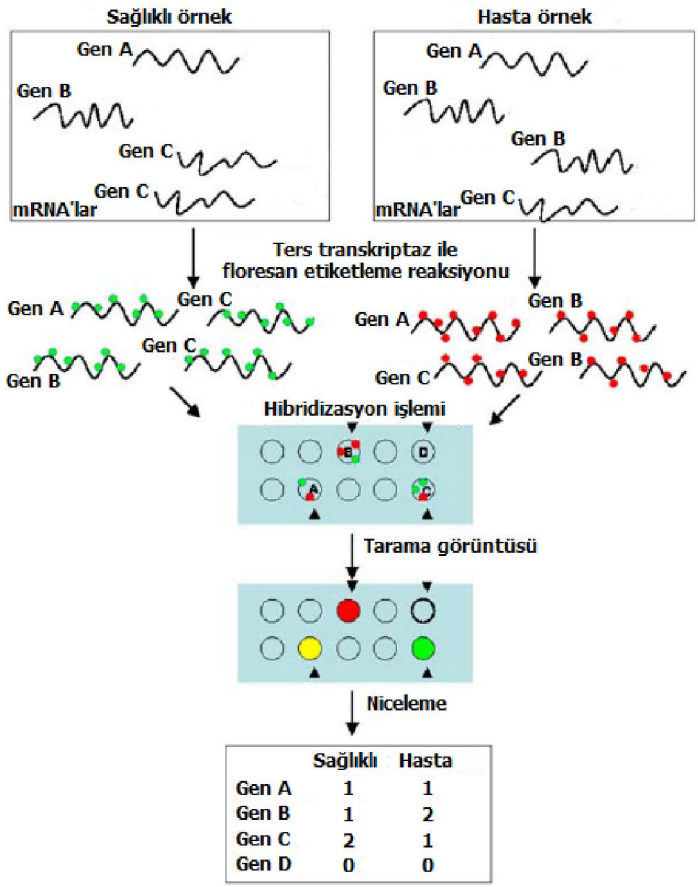
GENCODE verilerine göre insan genomunda yaklaşık 58 000 gen ve 20 000 civarında protein kodlayan gen bulunmaktadır [20]. Organizmanın yaşamsal döngülerinde her bir genin nasıl bir biyolojik işleve sahip olduğu, metabolik ve çevresel değişkenlerden ne yönde etkilendiği gibi hususların aydınlatılması için gen ifadesi değerleri analiz edilir. Genlerin tek başlarına ele alınmaları suretiyle kendilerine özgü görevleri ve işleyişleri hakkında birtakım bilgiler elde edilebilir. Bu bilgiler önemli olmasına rağmen, biyolojik süreçleri anlamlandırmada yetersiz kalır. Gen faaliyetlerinin yeterli derecede anlaşılması

için daha geniş kapsamlı araştırmaların yapılması gerekir. Genleri, moleküler ağlardaki etkileşimlerini ve diğer genlerle olan ilişkilerini göz ardı ederek izole bir biçimde ele almak yerine hücresel metabolizmada birlik içinde analiz etmeye imkân veren çeşitli teknolojiler geliştirilmiştir. Mikrodiziler, moleküllerin sistem çapındaki değişikliklerini ölçmek için geliştirilen OMICs teknolojilerinden biridir [15]. Biyoçip veya gen çipi de denen DNA mikrodizileri ile gerçekleştirilen deneylerde on binlerce genin ifade düzeyi tek bir seferde ölçülebilir. Bir mikrodizi deneyinde genlerin ifade düzeyleri ışık sinyalleri hâlinde temsil edilir. Bu sinyaller yüksek yoğunluklu bir çip üzerinde aynı anda görüntülendiği için, araştırma kapsamında muazzam sayıda gen bir arada ele alınabilir. Böylelikle probleme bütünsel bir açıdan bakılması mümkün hâle gelir. Gen çipinden yayılan ışık sinyalleri görüntü formunda kayda geçirilir. Görüntü verilerinin bir dizi teknikle işlenmesi sonucu elde edilen sayısal değerli veriler deney sonunda veri kümeleri şeklinde kaydedilir. Bu veri kümelerinin yüksek seviyeli analizleri neticesinde keşfedilen örüntüler ve ortaya çıkarılan anlamlı bulgular, geniş çaplı multigenik araştırmalara ivme kazandırmaktadır.

Mikrodizi deneyleriyle elde edilen veriler kanserin hücresel fenotipi hakkında bilgi edinmeyi sağlar ve kanserin genom çapında izlenebilirliği açısından büyük öneme sahiptir. Kanser genetiği ile ilgili mikrodizi deneyleri; moleküler düzeyde tanı, prognoz belirleme, kanser alt tiplerini sınıflandırma, ilaç geliştirme gibi çok çeşitli amaçlarla gerçekleştirilir. Deneylerde; kanserli ve sağlıklı hücreler, kimyasallara maruz kalmış ya da ilaçla reaksiyona girmiş hücreler, farklı tipte kanser hastalarından alınan hücreler, bir örnekten zaman veya koşul değişimi gözetilerek alınan hücreler karşılaştırılır. Böylece kanser üzerinde etkili olan genler veya koşullar ortaya çıkarılabilir. Mikrodizi deneylerinde süreç, gen çipi ve örneklerin hazırlanması ile başlatılır; peşi sıra uygulanan hibridizasyon, tarama ve görüntü işleme basamakları ile sürdürülür. Hasta ve sağlıklı olan iki örnekle gerçekleştirilmiş bir DNA mikrodizi deneyine ilişkin süreç, Şekil 2.3'te ana hatlarıyla özetlenmektedir.

DNA mikrodizileri, mikroskop lamı gibi cam ya da katı bir yüzey üzerinde açılmış spot denen binlerce kuyucuktan oluşur. Basit bir tabirle mikrodiziler; şekilsel olarak gözenekli bir matris formunda tasarlanan çiplerdir. Gen ekspresyonunu ölçmede, DNA sekansının kendi komplementeri ile hibridize olma özelliği kullanılır. Çip hazırlanırken bu husus dikkate alınır ve biyolojik araştırma kapsamında ifade düzeyi ölçülmek istenen genlere

karşılık komplementer (complementary) Deoksiribonükleik Asit (cDNA) klonları oluşturulur. Klonlanmış cDNA molekülleri, mikrodizi üzerine gerekli miktarlarda basılmak üzere hazırlanır. Bir deneysel çalışma kapsamındaki genlere karşılık gelen izole edilmiş bu moleküller, çip üzerinde belirli pozisyonlarla tanımlanan kuyucukların içine robotik teknolojiyle sağlam bir şekilde tutturulur. Böylece her bir genin DNA sekansı, konumu bilinen bir spotla eşleştirilmiş olur.



Şekil 2.3. Sağlıklı ve hasta örnekler ile gerçekleştirilen mikrodizi deneyinin şeması [21]

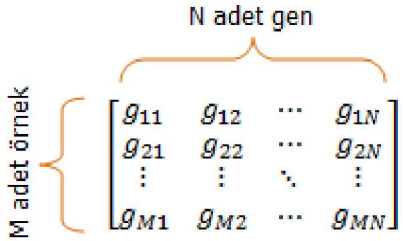
Mikrodizi deneyi ile test edilmesi plânlanan ve bu çerçevede dizilimleri ve çip üzerindeki yerleri bilinen cDNA moleküllerine prob denir. Mikrodiziler probların oluşturulma şekline göre cDNA mikrodizisi ya da oligonükleotid mikrodizisi diye ayrı ayrı isimlendirilirler. cDNA mikrodizileri hazırlanırken, yukarıda da bahsedildiği gibi, başka bir ortamda oluşturulan cDNA klonları çip üzerine basılır. Bundan ayrı olarak, DNA probları doğrudan spotlar üzerinde sentezlenebilir. Bu teknikle üretilen ve genellikle daha kısa sentetik DNA sekansları içeren biyoçipler oligonükleotid mikrodizileri diye bilinmektedir.

Bir cDNA mikrodizi deneyinin gerçekleştirilebilmesi için gerekli ilk şey kullanıma hazır bir çiptir. Gen çipi temin edildikten sonra, deneyde kullanılacak örneklerin hazırlığına geçilir. İlk önce, belirli biyolojik örneklerde ifade edilen genleri temsil eden mRNA transkriptleri toplanır. Bu aşamada, mRNA popülasyonunun hücresel içerik bütününden ayrılmasını sağlamak için moleküler biyolojinin alanına giren çeşitli işlemler uygulanır. mRNA popülasyonları elde edildikten sonra hibridizasyon basamağında kullanılacak olan komplementer zincirlerin hazırlanması gerekir. Bu amaçla, biyolojik örnekteki mRNA moleküllerine karşılık gelen daha kararlı bir cDNA örneklemi üretilir [22]. Enzimler yardımıyla gerçekleştirilen bu işleme ters transkripsiyon denir. Aynı sırada cDNA molekülleri, hibridizasyon sinyali olarak işlev gören floresan nükleotidlerle etiketlenmektedir. Floresanla işaretlenmiş tek zincirli cDNA moleküllerinden oluşan örnekleme hedef denir. Hedefler oluşturulurken; referans örnekten üretilen cDNA örneklemi yeşil boya, test örneğinden hazırlanan cDNA örneklemi ise kırmızı renkte boya ile işaretlenerek aralarında ayrım yapılabilmesi sağlanır. Daha sonra, referans ve test örneklerinden hazırlanmış olan ve dizilimleri bilinmeyen bu işaretli hedefler karıştırılır. Elde edilen karışım tek bir çip üzerine uygulanır. Hedeflerin birleştirilip eş zamanlı uygulanması; hedef popülasyonlarının hibridizasyon için rekabet etmelerine imkân vermek içindir.

Hibridizasyon, her iki örneğin transkriptlerini temsil eden işaretli hedefler ve gen çipi üzerindeki cDNA sekansları arasında gerçekleşen komplementer tabanlı eşleşme reaksiyonunu tabir eder. Baz eşleşmesi gerçekleşen zincirlerin birbirlerine bağlanması için deney çipinin uygun sıcaklıkta belirli bir süre bekletilmesi gerekir. Yeterli süre sonunda hibridizasyon tamamlanır. Hibridizasyon basamağının esas amacı; ölçümü yapılacak ışık sinyallerini ilgili spotlarda üretmektir. Bu nedenle anlamsız sinyallerin yok edilmesi, yani mikrodizinin gürültü üreten hibridize olmamış moleküllerden temizlenmesi gerekir. Hibridizasyon sonrasında deneysel kalıntılardan arındırılan mikrodizi ölçümler için hazır hâle gelir. Her ışıklı spot biyolojik örneklerde ifade edilen bir gene karşılık gelir ve yaydığı ışık bu genin transkripsiyon ürünü olan mRNA moleküllerinin yoğunluğunu gösterir. O hâlde hibridizasyon sonlandığında açığa çıkan ışık sinyalleri örneklerdeki gen aktivitesi hakkında bilgi edinmeyi sağlar. Şöyle ki; çok aktif olan genlerin mRNA transkripti fazla olacağından ilgili spotlar daha parlaktır [22]. Işık olmayan ve siyah renkte görüntülenen spotların ilişkilendirildiği genler, örneklerin hiçbirinde aktif değildir ve bu ışıksız spotlarda hibridizasyon gerçekleşmemiştir [23]. Bir spotta kırmızı renk baskın gelmişse ilgili gen

test örneğinde (hasta ya da tümörlü örnek) normalden fazla ifade edilmektedir; spot yeşil ise gen referans örnekte (normal örnek) daha fazla ifade edilmektedir [22]. Eğer bir gen test ve referans örneklerinde eşit olarak ifade ediliyorsa spotlandığı bölge sarı renkte görüntülenir [23].

Bir mikrodizi deneyinin ana adımlarından son ikisi; tarama ve görüntü işlemedir. Mikrodizi deneyi ile teste tâbi tutulan genlerin iki farklı deney koşulu altındaki ifade düzeyleri kırmızı ve yeşil floresan yoğunluklarına göre ölçülür. Bu ölçüm için, floresan etiketleri kırmızı ve yeşil floresana uygun lazerler ile uyarılarak her bir spot taranır. Tarama adımının sonunda, spotların yaydığı ışık değerlerinin piksel piksel belirtildiği iki sayısal görüntü oluşturulur. İki ayrı renk kanalına göre elde edilen bu görüntüler, gen ifade oranı değerlerinin hesaplandığı ham verilerdir [24]. Görüntü işleme yapılarak piksel değerlerinden spotların sinyal yoğunluk değerleri elde edilir. Elde edilen nicel değerler; logaritmik dönüşüm, normalizasyon gibi bir takım işlemlerden geçirilerek veri matrisi şeklinde kaydedilir. Tekrar etmek gerekirse hibridizasyon deneyi ile üretilen her veri noktası, test örneğinin gen ifade düzeyi ile referans koşulun gen ifade düzeyi arasındaki oranı temsil eder [25, 26]. Mikrodizi deneyleri sonucunda elde edilen gen ifadesi değerlerinin saklandığı bir veri kümesinde; her bir satır bir örnek veya zaman noktasını, her bir sütun özel bir geni, her bir matris elemanı belirli bir genin ilgili örnekte veya zaman noktasında ölçülen ifade düzeyini gösterir [27]. Yani M adet örneğin her biri için N adet genin ifade değeri ölçümleri (g_{11}, ..., g_{1N}; ...; g_{M1}, ..., g_{MN}) elde edilip, Şekil 2.4'te gösterildiği gibi matris formunda bir veri kümesi oluşturulur.



Şekil 2.4. Bir gen ifadesi veri kümesinin matris formatında gösterilişi

Mikrodizi verileri, sistem çapında tahmin modellerinin geliştirilmesi ve bunun yüksek doğruluklu bir şekilde yapılabilmesi için güçlü bir alt yapı sunar. Bununla birlikte, mikrodizi verileri ile yapılan çalışmalarda çeşitli sorunlar ve kısıtlılıklarla karşılaşılmaktadır. Bunlardan bazıları şu şekilde sıralanabilir;

- Piyasaya sürülen mikrodizilerin üretiminde belli bir standart yoktur. Dahası, esas alınan deney protokolü ve uygulanan yöntemler her çalışmada birebir aynı değildir. Çip farkı, ilgili proba özgü olmayan hibridizasyon sinyalleri veya tarama aşamalarındaki farklar nedeniyle deneylerde optik arka plan gürültüsü değişebilir.
- Farklı deneysel platformlardan veya farklı laboratuvarlardan elde edilen veriler arasında geçerli karşılaştırmalar yapılmasını sağlamak üzere mikrodizi deneylerinin uyumluluğu araştırılmalı, gen ifadesi ölçümlerinin hassasiyetini etkileyen faktörler ortaya çıkarılmalı ve verilerin depolanmasına ilişkin standartlar belirlenmelidir [28-30]. Bu gibi hususların bilimsel çerçevede ayrıntılı olarak ele alınması için kapsamlı çalışmalara ve özel projelere gereksinim duyulur.
- Fabrikasyon hataları, spotlama hataları, yanlış tarama ya da düşük görüntü çözünürlüğü sonucunda gen ifadesi verilerinde rastlanan çok sayıda eksik değer, araştırma süreçlerini olumsuz etkiler [31].
- Veri kümelerinin oluşturulması için gerekli numunelere ulaşmak her zaman kolay değildir.
- Mikrodizi teknolojisi yardımıyla biyolojik örneklerdeki mRNA transkriptlerinin türleri belirlenip miktarları ölçülebilir; ancak bu ölçümler gen regülasyonunu tümüyle aydınlatmaz, sadece gen ifadesi sürecinin belirli bir kısmı hakkında bilgi verir.
- Mikrodizi deneyleri sonucunda elde edilen yüksek boyutlu veriler oldukça karmaşıktır. Bu verilerden anlamlı bilgilere ulaşmak isteyen araştırmacılar etkili analiz yöntemlerine ihtiyaç duyar. Bunun için verilerin; biyoenformatik, veri madenciliği, makine öğrenmesi gibi araştırma disiplinleriyle sistematik bir biçimde ele alınması gerekir [32, 33].
- Mikrodizi teknolojisi ile elde edilen bilgilerin bir kısmı klinik uygulamalara aktarılabilse de mevcut teknolojilerin klinik uygulama bakımından hâlen yetersiz kaldığı söylenebilir [34].

2.2. Gen İfadesi Verilerinin Makine Öğrenmesi Tabanlı Analizi

Mikrodizi teknolojisinin ortaya çıkışı ile gündeme gelen önemli meselelerden birisi, gen ifadesi verilerinin işlenmesinde karşılaşılan zorluklardır. Makine öğrenmesi yaklaşımları, aşılması gereken bu zorluklara karşı etkili çözümler sunduğu için gen ifadesi verilerinin analizinde sıkça kullanılmaktadır.

2.2.1. Makine öğrenmesi ve sınıflandırma

Makine öğrenmesi; sistemli olarak toplanan iyi tanımlı verileri etkili bir şekilde kullanmayı ve daha önce görülmemiş veri örneklerinin bilinmeyenleri hakkında belirli bir analiz hedefine göre çıkarım yapmayı sağlayan yöntemler bütünüdür. Bu yöntemler, çözüm yolu kesin kurallarla belirlenemeyen problemlerin analizinde önemli yer tutar. Makine öğrenmesi yöntemlerinin asıl ilgi çeken yönü; belirli değerlerle nitelenmiş olan veri örneklerini kullanması ve bu değerlere göre şekillenen bir öğrenme faaliyeti içinde bulunmasıdır. O hâlde, bir uğraşı veya araştırma kapsamında elde edilen verilerin, nitelikleri bilinen ve bu nitelikler altında hangi değerleri aldığı belirli olan bir örnek kümesi formunda kayda geçirilmesi önemlidir. Çünkü bu sayede, elde bulunan veriler yeni örneklerle ilgili muğlâk noktaları açıklığa kavuşturmak için makine öğrenmesi disiplini çerçevesinde pratik ve etkili bir şekilde kullanılabilir.

Denetimli ve denetimsiz öğrenme makine öğrenmesinin iki alt başlığıdır. Herhangi bir bağlamda etiket bilgisi eklenmemiş veriler üzerinde çıkarımlarda bulunmayı sağlayan kümeleme algoritmaları denetimsiz öğrenmenin konusudur. Kümeleme problemlerinde, daha önceden tasnif edilmemiş veriler ele alınır. Bu verileri benzerliklerine göre en uygun şekilde gruplamak için makine öğrenmesi temelli çözüm yolları öneren çeşitli algoritmalar kullanılır. Denetimli öğrenme ise sınıflandırma algoritmalarını konu alır. Birçok gerçek yaşam probleminde sınıflandırma kuralı bilinmemekle birlikte mevcut olan tek bilgi etiketli örneklerden oluşan bir veri kümesidir [35]. Veri kümesindeki örnekler, tahminci özniteliklerin alabileceği olası değerler skalasında nitelendirilirler. Öznitelikler aracılığıyla veri örnekleri ile sınıflandırma algoritması arasında bağlantı kurulur. Buradan da anlaşılacağı üzere, sınıflandırma problemlerinin çözümü için, özniteliklerin sınıf değişkeniyle olan ilişkilerinin analiz edilmesi gerekir. Başka bir şekilde ifade etmek gerekirse; test örneklerini birbirinden ayırt etmek için öznitelik uzayını başarılı bir şekilde ayıran bir öngörü modeli oluşturmak gerekir. Öngörü modeli oluşturmak; geçmişte tecrübe edilmiş durumları göz önüne alarak, bir uygunluk kriterini sağlayan muhtemel çözümlerin bulunduğu bir uzayda verilere en uygun modeli araştırmaktır [35].

En basit tanımlama ile sınıflandırma örneklerden öğrenmedir [32]. Sınıflandırmada amaç; bilinen bir etiket değişkeni ile ilişkilendirilmiş ve belirli nitelik değerlerine sahip veri örneklerini analiz ederek, yeni ve bilinmeyen örnekleri etiketlemek için uygun olan çözüm

yolunu öğrenmektir. Bu amaçla ilk önce, aynı öznitelik kümesi ile tanımlanmış eğitim örnekleri kullanılarak öznitelikler açısından sınıf etiketlerinin dağılımını belirleyen genelleyici bir öngörü modeli oluşturulur [36]. Oluşturulan model daha sonra, öznitelik değerleri bilinen ancak henüz sınıflandırılmamış test verilerine, mevcut sınıf etiketlerini atamak için kullanılır [36]. Sınıflandırma problemlerinin çözümü için farklı kavramsal esaslara dayanan çeşitli yaklaşımlar benimsenmiştir. Örnek tabanlı öğrenme [36, 37], istatistiksel öğrenme [36, 38], çok katmanlı algılayıcılar [39], DVM [40] ve karar ağaçları [36, 41] sınıflandırma problemlerinin çözümü için kullanılan temel yaklaşımlar arasında yer almaktadır.

Kanser genetiği açısından bakıldığında, veri örneklerinin kanser araştırması dâhilinde belirlenen spesifik kategoriler içinden hangisi ile ilişkili olduğunu tespit etmek için sınıflandırma işlemine başvurulur. Etiket değişkenleri genellikle kanser fenotipi hakkında ayırıcı bir bilgi verir. Çoğu çalışmada; kanserli ve sağlıklı örneklerin tahmini, kanser tipleri veya kanser alt tiplerinin ayırt edilmesi, hastalığın seyri ile ilgili biyolojik olguların tahmini gibi sınıflandırma problemleri üzerinde durulmuştur [42-46]. Tahminlerden kesin sonuç alınması; erken teşhis, doğru ve zamanında tıbbî müdahale, etkili tedavi yöntemlerinin uygulanması ve hastanın tedaviye cevap vermesini sağlama açısından çok önemlidir. Gen ifade verisi örnekleri arasında ayrım yapabilecek yüksek doğruluklu tahmin modelleri, kanserle ilgili diyagnostik ve prognostik süreçlerde; biyopsi, dokularda morfolojik değişimlerin incelenmesi, kan testleri, radyografi gibi konvansiyonel yöntemlerden yararlanan hekimlere ciddi bir karar desteği sağlar.

Gen ifadesi verilerinin etkin bir biçimde işlenmesi için birbirine alternatif olabilecek birçok makine öğrenmesi temelli model önerilmiştir. Verilerden modelleme yaparken aslında, sınıflandırma probleminin en iyi çözümünü bulmak için bir optimizasyon işlemi gerçekleştirilir [35]. Nitelik sayısı çok ancak örnek sayısı az olan gen ifadesi verileri yüksek boyutlu verilerin tipik bir örneğidir. Yüksek boyutlu verilerin modellenmesi sürecinde ortaya çıkabilecek tüm pürüzlerin sınıflandırma algoritmasıyla giderilmesi imkân dışıdır. Tahmin performansında düşüşe yol açan etmenleri olabildiğince azaltmak için bazı yardımcı bileşenlere ihtiyaç duyulur. Modelin verimliliği ve optimizasyonu açısından önemli yer tutan bu bileşenler; veri ön işleme ve öznitelik seçimi başlıkları altında incelenebilir. Bunlar haricinde, modelleme sürecinde performans doğrulaması için başvurulan bazı tekniklerden ayrı bir başlık altında söz edilecektir.

2.2.2. Veri ön işleme

Mikrodizi gen ifadesi verilerinde görülen; gürültü, eksik veri gibi problemli durumlar ve eğitim verisi azlığı gibi sınırlılıklar tahmin doğruluğunu kötü yönde etkilemektedir. Bu nedenle gen ifadesi verilerinin analizi genellikle ön işlem süreci ile başlatılır. Ön işlem adımında, verilerin daha kolay ve etkili işlenmesi için algoritmik destek sağlayan bir dizi teknik uygulanır. Verilerin uygun ön işlemlere tâbi tutulması; öznitelik seçimi, sınıflandırma gibi daha üst seviyeli analizlerin performansını olumlu yönde etkiler. Mevcut çalışmalarda özellikle; veri impütasyonu, standardizasyon ve ayrıklaştırma işlemleri arasından gerekli görülenlerin gen ifadesi verilerine uygulandığı görülmektedir [4, 47, 48].

Veri impütasyonu

Mikrodizi teknolojisi ile elde edilen veri kümelerinde, üretim hataları veya deneysel hatalar nedeniyle eksik veri noktalarına rastlanması alışılagelmiş bir problemdir. Veri örneklerindeki kayıp ve eksik öznitelik değerleri sebebiyle ortaya çıkan gürültünün giderilmesi için çeşitli yöntemler önerilmiştir. En basit yöntem; eksik gözlemler içeren özniteliklerin veri kümesinden çıkarılmasıdır; fakat böyle bir yönteme başvurmak özniteliğin sağladığı yararlı bilgilerin kaybolmasına neden olur [31]. Benzer şekilde, eksik değer içeren örneklerin silinmesi tercih edilebilir ancak veri azlığı sebebiyle riskli bir yaklaşımdır. Öte yandan eksik verilerin analizi için kullanılan impütasyon tekniklerinde, verileri silmek yerine eksik gözlem noktalarını uygun değerlerle tamamlamayı öne çıkaran çözüm yollarına başvurulur. Sayısal öznitelikler için eksik değerlerin; sıfırla, sabit bir değerle ya da ilgili özniteliğin eğitim verilerindeki mevcut değerlerinin ortalaması ile değiştirilmesi daha kabul edilebilir çözümler arasında sayılabilir [49, 50]. Bunun yanı sıra, veri kümesinin eksik olmayan kısmından istifade edilerek eksik değerlerin tahmin edilmesi için; karmaşık istatistiksel modeller, regresyon ve optimizasyon algoritmaları, harici alan bilgisi kaynaklarıyla desteklenen çeşitli yaklaşımlar önerilmiştir [49-51].

Standardizasyon

Veri kümelerinin çoğunda, tüm öznitelik seti için tanımlanmış olan tek tip ve durağan bir değer aralığı yoktur. Özniteliklerin ölçeklendiği değer aralıklarının farklı olması, öznitelik seçimi ve sınıflandırma gibi müteakip analiz aşamalarındaki göreli önem seviyelerini

gerçeğe aykırı biçimde etkileyebilmektedir [52]. Birbirinden farklı dinamik aralıklarda değer alan öznitelikler, standardizasyon işlemiyle daha dar veya kısıtlı bir aralığa ölçeklenerek hesapsal önem açısından aralarında denklik sağlanabilir. Böylece, özniteliklerin başa baş değerlendirilebileceği bir zeminde daha dengeli ve tarafsız analizler gerçekleştirilir. Bu hedef doğrultusunda sık kullanılan yöntemlerden biri; özniteliklerin sıfır ortalama ve birim varyansa sahip olacak şekilde standartlaştırılmasıdır [47, 53-54].

Ayrıklaştırma

Tahmin modellerinde yer verilen bazı öznitelik seçimi ve sınıflandırma algoritmaları yalnızca ayrık verileri kullanarak işlem yapabilir. Ayrık veri girişi gerektiren analizler gerçekleştirilecekse verilerin sürekli öznitelik değerleri, ön işlem sürecinde bir ayrıklaştırma işlemiyle kullanıma uygun hâle getirilir. Ayrıklaştırmada, sayısal bir özniteliğin değer aldığı sürekli aralık, sıralı ve birbiriyle çakışmayan alt aralıklar şeklinde bölünür. Ayrıklaştırılmış yeni değer aralığı, üretilmiş olan sonlu sayıdaki ayrık bölümlerden her birine karşılık hususî bir ayrık durum değeri ihtiva eder. Ayrıklaştırma neticesinde her bir sürekli öznitelik değeri, bulunduğu aralığı temsil eden ayrık durum değeriyle ifade edilir. Pratikte, gen ifadesi verilerinin çok geniş bir sayısal spektrumda aldığı sürekli değerleri, büyük ölçüde azaltılmış ayrık durum değerlerinden oluşan bir alt kümeye eşlediği için ayrıklaştırma bir veri azaltma tekniği olarak görülebilir [55]. Gen ifadesi verilerinin analizi için ayrıklaştırma işleminin sağladığı avantajlar şöyle sıralanabilir;

- Ayrık durumların kullanılması sayesinde ham verilerin barındırdığı biyolojik ve teknik gürültünün büyük kısmı soğurulur [55, 56].
- Ayrık değerler bilgi düzeyindeki temsillere daha yakın olduğu için kullanımı ve kavranması sürekli değerlere göre daha kolaydır [57, 58].
- Ayrıklaştırma veri miktarını azaltırken aynı zamanda tahmin doğruluğunu koruma hatta geliştirme potansiyeline sahiptir [57].

Ayrıklaştırma işlemi denetimli ve denetimsiz yöntemler kullanılarak gerçekleştirilebilir. Denetimsiz yöntemler, alt aralık sayısını sınıf etiketiyle ilgisi olmayan bir kurala göre belirleyerek ayrıklaştırma yapar. Eşit genişlikli ayrıklaştırma [59] gen ifadesi verilerinin analizinde popüler olan en basit denetimsiz yöntemlerden biridir [60, 61]. Bu yöntem, özniteliğin değer aralığını, verilen bir aralık sayısına göre eşit genişlikteki alt bölümlere

ayırmaya dayanır. Denetimsiz ayrıklaştırmanın gen ifadesi verilerine uygulanışında sıkça karşılaşılan bir diğer yaklaşım, alt aralıkları belirleyen kesme noktalarını hesaplamak için özniteliğin ortalama ve standart sapma değerlerine göre hazırlanmış formülasyonlar kullanmaktır [47, 53-55, 62]. Örneğin; yalnızca ortalama değeri kullanılarak ikili ayrıklaştırma yapılabilir yahut ortalamayla beraber standart sapma değeri kullanılarak öznitelikler 3-durumlu veya 9-durumlu olacak şekilde ayrıklaştırılabilir [62, 63]. Denetimli yöntemlerde ise sürekli bir özniteliğin değer aralığını, sınıf bilgisine göre en anlamlı alt aralıklara bölme amacı ön plândadır [57]. Minimum Tanım Uzunluğu (Minimum Description Length - MDL) yöntemi [64] gen ifadesi verilerini ayrıklaştırmada yaygın olarak kullanılan denetimli yöntemler arasında gösterilebilir [65, 66].

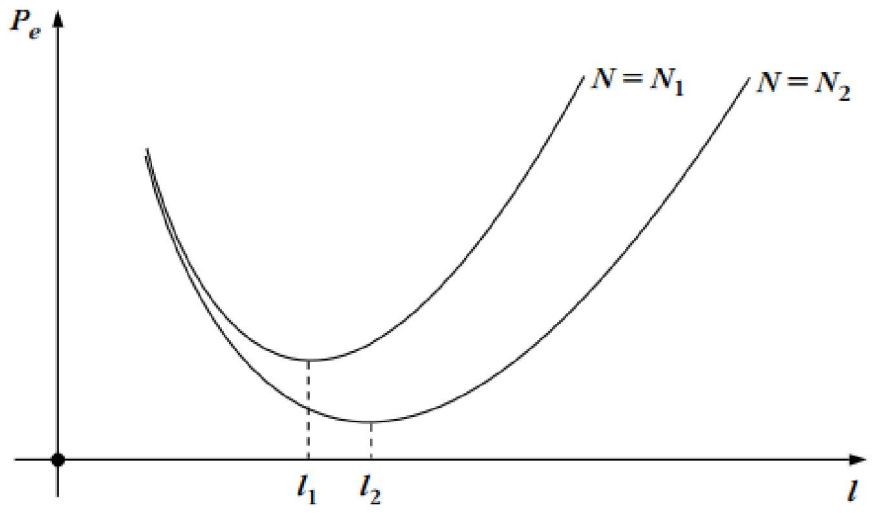
2.2.3. Öznitelik seçimi

Çoğu gen ifadesi veri kümesi, modeli eğitmek için mevcut olan 100'den daha az sayıdaki veri örneğine kıyasla on binlere varan gen sayısı ile karakterizedir [23]. Öznitelik sayısı ve örnek sayısı arasındaki bu ilişki tahmin doğruluğunu sınırlandırır. Sınıflandırmada yüksek genelleme performansı elde edebilmek için, eğitim örneklerinin sayısı özniteliklerin sayısına göre yeterince büyük olmalıdır [52]. Bunun nedeni sınıflandırılacak veri örneklerinin dağılımının teorik olarak bilinmemesidir. Eğitim verisi örnekleri azken öznitelik uzayı büyükse dağılımın parametreleri hakkında edinilen bilgi daha azdır. Yani, öznitelikler için yeterli değişkenlikte veri noktası yoksa doğru sınıflandırmayı sağlayan parametre değerlerini kestirmek zorlaşır. Mikrodizi gen ifadesi verileri, yararlı özniteliklerin yanı sıra çok sayıda gereksiz ve tekrarlı öznitelikten oluşur ve özellikle sınıflandırma görevi için tüm boyutlar gerekli değildir [67]. Bu nedenle veri ön işleme aşaması sonrasında mikrodizi verileri boyutluluk açısından analiz edilir.

Yüksek boyutluluk ve örnek azlığı sebebiyle gelişmesi muhtemel olan sakıncalı bir durum; uygulanan sınıflandırıcının eldeki veriye aşırı uyum sağlamasıdır. Çok sayıda değişkene göre az sayıda örnek kullanılarak eğitilen sınıflandırıcı ezberci bir tahmin modeline sahip olacağı için, eğitim verilerini iyi sınıflandırsa bile yeni örnekleri sınıflandırmaya çalıştığında kötü performans gösterecektir [42].

Sınıflandırıcının aşırı uyum problemi kendini zirve fenomeninde gösterir [68]. Pratikte, N tane örnekten oluşan sonlu bir küme için; özniteliklerin sayısının arttırılmasıyla

performansta başlangıçta bir iyileşme elde edilir, ancak kritik bir değerden sonra öznitelik sayısının artması sınıflandırmada P_e hata olasılığının artmasına yol açar [52]. Bu durum *zirve fenomeni* olarak adlandırılır ve sınırlı sayıda eğitim verisi altında, öznitelik sayısının nispeten düşük bir sayıda tutulması gerektiği anlamına gelir [52]. N ve l sırasıyla eğitim kümesi ve öznitelik uzayının büyüklüğü olmak üzere, N'nin büyüklüğüne göre genel hata eğilimi Şekil 2.5'te gösterilmektedir. Burada, $N_2 >> N_1$ durumunda $l_2 > l_1$ olduğu görülmektedir. Yani, sınıflandırma hatasını en aza indirmek için kullanılması gereken öznitelik sayısında, eğitim kümesinin büyüklüğüne bağlı olarak bir miktar değişiklik gözlemlenir. Zirve fenomenine ilişkin son bir not düşmek gerekirse; bir a değişkeni 2 ile 10 aralığında değerler almak üzere, genellikle en düşük hata olasılığına $l = N/a$ sayıda öznitelikle ulaşılır [52]. Buna göre, 100 civarında örnek bulunduran çoğu veri kümesi için, yaklaşık 10 ile 50 arasında seçili öznitelikten oluşan bir alt küme ile en iyi sınıflandırmayı yapmak mümkündür.



Şekil 2.5. Farklı büyüklükteki eğitim kümeleri için sınıflandırma hatası olasılığının öznitelik sayısına göre değişimi [52]

Boyutluluğun istenmeyen etkileri tahmin performansını düşürmesiyle sınırlı değildir. Yüksek boyutluluk, sınıflandırma yöntemlerinin eğitilmesi ve test edilmesi süreçlerinde hesaplamaların artmasına yol açarak analizi zorlaştırır. Genellikle boyutsallık arttıkça hesaplama maliyeti de katlanarak artar [69]. Yüksek boyutluluğun ortaya çıkardığı negatif etkiler göz önünde bulundurulduğunda, sınıflandırma amaçları için gen ifadesi verilerinde boyutluluğun azaltılması gerekliliği ön plâna çıkmaktadır.

Boyutluluğun azaltılmasında; özellik çıkarma ve öznitelik seçme olmak üzere iki esas tercih söz konusudur [70]. Özellik çıkarma yöntemleri, orijinal özniteliklere uyguladığı dönüşümlerle yeni bir indirgenmiş öznitelik kümesi türetir. Dönüşümle türetilen yeni özniteliklerde fiziksel anlam korunmaz [70]. Dolayısıyla özellik çıkarma stratejisi çerçevesinde uygulanan analizlerin yorumlanması zordur ve analizlerden elde edilen sonuçların okunabilirliği düşüktür [70]. Öznitelik seçim tekniklerinde ise özniteliklerin fiziksel anlamlarının korunması amaçlanır ve herhangi bir dönüşüm uygulanmadan sonuca ulaşılır [70, 71]. Yani, özniteliklerin bir alt kümesi seçilerek veriler en küçük boyutla ifade edilmeye çalışılır. Orijinal öznitelik uzayını değiştiren özellik çıkarım yöntemlerine kıyasla seçim yoluyla yapılan indirgeme alan uzmanları tarafından daha kolay yorumlanabildiği için gen ifadesi verilerinde öznitelik seçimi ana tercih hâline gelmiştir [70, 72, 73].

Öznitelik seçim yöntemleri kullanılarak, ayrı sınıfların örneklerini ayırt etme görevinde en yüksek başarıyı sağlayan gen alt kümesinin bulunması amaçlanır. Muazzam sayıda gen arasından sadece küçük bir bölümün, kanser başlangıcı veya ilerlemesi ile ilgili olması muhtemeldir [53, 74]. Biyolojik olarak bilgilendirici olan bu genler; işaretleyici gen ya da biyobelirteç olarak anılır [46, 53, 75]. Biyobelirteç tespiti ve sınıflandırma performansını geliştirme; gen ifadesi verilerinin makine öğrenmesi tabanlı analizinde birbiriyle iç içe ele alınan iki ana hedeftir [54, 76]. Bu hedeflere ulaşmak için genellikle; sınıflandırıcı yöntem ve öznitelik seçim yöntemi kombinasyonu şeklinde tasarlanan tahmin modelleri uygulanır. Gen ifadesi verileri üzerinde gerçekleştirilen araştırmalarda, veri kümelerinin genel karakteristiği sebebiyle öznitelik seçimi belirgin bir önem kazanmıştır. Özellikle yüksek boyutlu veriler için tahmin modeli tasarımının kaçınılmaz bir parçası olan öznitelik seçimi, gen ifadesi verilerinin analizinde en çok üzerinde durulan konular arasında yer alır [68].

Stokastik bir optimizasyon problemi olan öznitelik seçimi [77], sınıflandırma görevi doğrultusunda en anlamlı özniteliklerin keşfine odaklıdır. Bu optimizasyon probleminin çözümü için iki temel prosedüre ihtiyaç duyulur. Bunlardan ilki, orijinal öznitelik kümesi içinden aday öznitelikleri belirlemek için gerçekleştirilen arama prosedürüdür. İkincisi ise arama prosedürüyle belirlenen adayların performanslarını ölçmeyi sağlayan değerlendirme prosedürüdür. Bir öznitelik seçim yöntemi, arama ve değerlendirme prosedürlerini birlikte çalıştırarak adayların performanslarını kıyaslar ve genellikle yineli adımlarla seçim işlemini sürdürür. Öznitelik seçim süreci, tanımlanmış bir durdurma şartı sağlanıncaya

değin devam eder. Öznitelik seçimini sonlandırmak için genellikle kullanılan durdurma kriterleri şunlardır [78];

- Ön tanımlı öznitelik sayısına veya yineleme sayısına ulaşılması.
- İki ardışık yineleme arasında belirli bir iyileştirme yüzdesine ulaşılması.
- Belirli bir değerlendirme fonksiyonuna göre en uygun öznitelik alt kümesinin elde edilmesi.

Öznitelik seçme yoluyla ayırıcı özniteliklerin en küçük ve etkili alt kümesini türetmek için geliştirilmiş olan çok sayıda yöntem vardır. Öznitelik seçim probleminin stokastik karakteri nedeniyle, prensipte iki seçim yöntemi benzer esaslara dayansa dahi birinin seçtiği ayırıcı alt küme öteki yöntem tarafından önerilen alt küme ile aynı değildir. Öznitelik seçim yöntemlerini birden çok kıstasa göre kategorilere ayırmak mümkündür. Bu kıstaslar değerlendirme veya arama prosedürlerinde uygulanan yaklaşımların temel farklarını öne çıkarır. Öznitelik seçim yöntemleri, literatürde çokça değinilen taksonomiler göz önüne alınarak, takip eden alt başlıklarda ele alınmaktadır.

Öznitelik seçiminde sınıflandırıcı bağımlılığı

Değerlendirme prosedürünün öğrenme algoritmasıyla ilişkisini dikkate alan taksonomiye göre öznitelik seçim yöntemleri; filtre [79, 80], sarmal [81, 82], gömülü [83, 84] ve hibrit [85-87] yöntemler olarak dört başlık altında gruplandırılır [88]. Burada temel ayrım yönü; özniteliklerin değerlendirilmesine yönelik sürecin, öğrenme algoritması ile bir bağının olup olmaması veya aralarında bir bağımlılık var ise bunun ne şekilde gerçekleştiğidir.

Filtreleme yöntemlerinde değerlendirmeler verilerden elde edilen genel istatistiksel ölçümlere dayanır. Öznitelik uzayındaki adaylar için gerçekleştirilen arama sınıflandırıcının hipotez alanındaki aramadan ayrılmıştır; dolayısıyla filtre temelli seçim sürecinde, sınıflandırıcı yöntem ile bağımlılıklar görmezden gelinmektedir [73]. Filtreleme yaklaşımı çerçevesinde özniteliklerin birbirleri arasındaki bağlantıları ve sınıf değişkeniyle aralarındaki anlamlı ilişkileri tespit etmek için; olasılıksal uzaklık, tutarlılık, sınıflar arası mesafe, bilgi-teorik veya olasılıksal bağımlılık benzeri ölçümler kullanılır [77, 78, 89].

Sarmal öznitelik seçimi, belirli bir sınıflandırıcı güdümünde gerçekleştirilir. Sarmal yöntemlerde, en iyi tahmin performansını temin eden alt kümenin keşfi hedeflenir. Buna

binaen, aday öznitelik alt kümelerinin önemi sınıflandırma performans ölçütlerine göre değerlendirilir. Sarmal yöntemler; öğrenme algoritmasının tümevarımsal eğilimlerine daha uygun olan gen kümelerini bularak üstün performans avantajı sağlayabilir [87].

Gömülü yöntemler, öznitelik seçimi ve sınıflandırıcı tasarımını bütünleştirerek daha sıkı işlenmiş öğrenme modellerinin ortaya çıkmasını sağlar. Gömülü öznitelik seçiminde optimal alt küme araması; öznitelik alt kümeleri ve ilgili sınıflandırma hipotezinin birleşik uzayında gerçekleştirilen yerleşik bir arama olarak görülebilir [73, 84]. Hibrit yaklaşımlar farklı türden öznitelik seçim yöntemlerini daha avantajlı şekillerde bir araya getirebilir. Hibrit yöntemlerin genellikle rastlanan formu filtre-sarmal öznitelik seçimidir.

Farklı kategorilerdeki yöntemlerin birbirlerine göre üstün ve zayıf yönleri mevcuttur. Tahmin modelinde kullanılan sınıflandırıcıya göre ayırıcı özniteliklerin en iyi alt kümesi farklılık göstereceği için, öznitelik seçiminde sınıflandırıcı faktörünü tamamen yok saymak her zaman iyi bir fikir değildir [90]. Bu bağlamda, sınıflandırıcının kendine özgü sezgilerini ve öznitelik seçimindeki yönlendirici etkilerini ihmal ettiği bilinen filtre temelli yöntemler, tahmin performansını optimize etme açısından zayıf olabilir [69]. Öğrenme sürecinin öznitelik seçiminden ayrılmasının avantajlı yönü ise hesap yükünün azalmasıdır. Bu sayede filtreleme yöntemleri, gen ifadesi verileri gibi büyük boyutlu veriler için kolayca ölçeklenebilir ve sayısal olarak basit ve hızlı çözüm sağlar [72, 91]. Filtre temelli yöntemler sınıflandırıcıdan bağımsız olduğu için, öznitelik seçimi bir ön işlem adımının uygulanışına benzer şekilde, yani öğrenme aşaması öncesinde sadece bir kez uygulanır. Filtreleme sonrasında ise eldeki seçili öznitelikler herhangi bir sınıflandırıcı ile değerlendirilebilir [92]. Dolayısıyla filtreleme yöntemleri diğer yöntemlere nispeten daha genel çözümler sunar [92].

Sarmal öznitelik seçiminde değerlendirme kriteri olarak sınıflandırıcı performansı kullanıldığı için filtreleme yöntemlerinden daha yüksek doğruluk elde edilebilir [84]. Buna karşın, aranan her alt kümenin değerlendirilmesi için sınıflandırıcının tekrar tekrar eğitimi ve test edilmesi yoğun bir hesap yükünü beraberinde getirir. Sınıflandırıcı bağımlılığıyla ortaya çıkan yüksek yoğunluklu hesaplamalar sarmal yöntemlerin yavaş çalışmasına ve aşırı uyum riskinin artmasına yol açar.

Gömülü yöntemler tahmin doğruluğu açısından sarmal yöntemlerle karşılaştırılabilir sonuçlara ulaşmakta, bunun yanı sıra daha az hesaplama gerektirmektedir [70]. Gömülü yöntemler, sınıflandırıcıyı iterasyonlar içinde çağırarak sonucu denetlemek yerine sınıflandırıcının tasarım kurallarını öznitelik seçimiyle birleştiren bir yol izler. Bu nedenle hesaplama açısından sarmal yöntemlerden daha verimlidir, ancak öznitelik seçim süreci tamamen tasarımda vurgulanan sınıflandırma algoritmasının yapısına özgü şekillendiği için başka sınıflandırıcılarla kullanılamaz. Sarmal yöntemlerin de sınıflandırıcı bağımlılığı yüksektir, ancak böyle olsa bile öznitelik seçim sürecinin farklı sınıflandırma yöntemlerine göre uyarlanma şansı vardır.

Tek değişkenli ve çok değişkenli öznitelik seçimi

Öznitelik seçim yöntemleri, nitelikler arası ilişkilerin hesaba katılıp katılmamasına göre tek değişkenli ve çok değişkenli yöntemler olarak iki gruba ayrılmıştır. Bu ayrılma daha çok, filtre temelli öznitelik seçiminin alt kategorilerini temsil eder. Tek değişkenli yöntemlerde özniteliklerin değerleri sınıf ayrımına katkılarına göre birer birer hesaplanır. Her bir öznitelik için hesaplanan bireysel değerler bir önem sıralaması belirlemek için kullanılır. Üst sıralara sınıf etiketi ile ilgili en çok bilgi taşıyan öznitelikler yerleşirken sıralamanın sonlarına gidildikçe sınıflandırma göreviyle ilgisi olmayan gereksiz öznitelikler yer alır. Sınırlı sayıdaki özniteliği kapsayan ayırıcı bir alt küme oluşturmak için bir durdurma kriteri kullanılması gerekir. Yani, bir durdurma kriteri kullanılarak en iyi sıralama değerlerine sahip olan özniteliklerden seçili bir alt küme oluşturulur. Gen seçiminde uygulanan tek değişkenli filtreleme yöntemlerine; bilgi kazancı [3, 93], *t*-testi [68], gini indeksi [3, 68], *t*-skor [82] ve fisher korelasyon skorlama [74, 94] yöntemleri örnek olarak gösterilebilir.

Tek değişkenli yöntemlerde özniteliklerin ayrımcı yetenekleri yalnızca sınıf değişkeni ile ilgisine bakılarak analiz edildiği için sadece gereksiz öznitelikler açığa çıkarılabilir. Öte yandan, tahmin performansını bozan yedekli öznitelikler ve tahmin performansını iyileştiren etkileşimli öznitelikler tek değişkenli öznitelik seçimiyle keşfedilemez. Dolayısıyla bu yöntemlerin seçtiği öznitelikler sınıflandırma probleminin çözümünde genellikle yetersiz kalır. Öznitelikleri daha ayrıntılı şekilde değerlendirmeye olanak veren bazı yararlı bilgiler öznitelikler arasındaki ilişkilerin irdelenmesi suretiyle açığa çıkarılabilir. Çok değişkenli yöntemler niteliklerin istatistikî ilişkilerinin ölçülmesine ve bu

bilgilerin değerlendirme prosedüründe işlerlik kazanmasına imkân verir. Böylece ayırıcı özniteliklerin seçimine ciddi bir destek sağlar. Genel anlamda; çok değişkenli bir yöntemle elde edilen öznitelik alt kümesi, tek değişkenli bir yöntemle elde edilen alt kümeye göre daha sade ve küçük, ama daha etkili ve yüksek ayrımcı güce sahiptir. Kısacası, çok değişkenli yöntemlerle daha özlü alt kümelerin elde edilmesi mümkündür. Gen ifadesi verileri açısından çok değişkenli yöntemler, gen gruplarının farklı sınıf türlerindeki diferansiyel davranışını ele alır [95].

Sarmal yöntemlerde takip edilen analiz süreci alt küme araştırmasına dayandığı için temelde çok değişkenli öznitelik seçiminin doğasıyla uyuşur. Sarmal yöntemlerde öznitelik ilişkilerinin ölçümü sınıflandırıcının gerçekleştirdiği kara kutu değerlendirmesinin içine dâhildir. Aday bir alt kümenin elemanları arasındaki ilişkiler çözümlenerek ölçülmez, ancak sınıflandırma performansı bu ilişkilerden doğan nicel sonuçları yansıtır. Bundan farklı olarak; öznitelikler arasındaki istatistiksel ilişkilerin bir dereceye kadar ölçülebilmesi amacıyla önerilen çok değişkenli filtre yaklaşımlarında [73], öznitelikler önceden tanımlanmış bağıntılar kullanılarak *açık döngü* [78] içinde değerlendirmeye tâbi tutulur. Bu özel bağıntılar hem özniteliklerin ilgi düzeylerini (relevance) ölçmeye hem de özniteliklerin kendi aralarındaki ilişki ve etkileşimleri analiz etmeye yönelik işlem terimleri içerir. Korelasyon tabanlı öznitelik seçimi [69, 93] ve bilgi teorisi tabanlı yöntemler [61, 80] gen seçimi için uygulanan çok değişkenli filtreleme kategorisindeki yöntemlere örnek olarak gösterilebilir.

Yeri gelmişken öznitelikler arasında ne tür sayısal ilişkiler olabileceği hakkında bilgi vermek ve bu konuya ilişkin belli başlı kavramları açıklamak faydalı olacaktır. Öznitelikler sınıf etiketinin tahmini hususunda birbirleriyle zayıflatıcı veya dayanışmacı ilişki içinde bulunabilir. Bu ilişki tiplerini tanımlamak amacıyla kullanılan fazlalık (redundancy) ve etkileşim (interaction) kavramlarının literatürde sıkça irdelendiği görülür. Seçili öznitelik kümesinin sınıflandırma görevine yönelik sunduğu bilgi birikimini doğru bir stratejiyle artırmak için aday öznitelikler ve seçili öznitelikler arasında fazlalık ölçümü yapılır. Birbirleri ile aynı temel bilgilere sahip olan yüksek korelasyonlu öznitelikler tahmin sürecini karmaşıklaştırarak tahmin doğruluğunu düşürür [96]. Aday özniteliklerin ilgi düzeylerinin ölçülmesi neticesinde, bu adaylar arasından pek çoğunun sınıflandırma çözümü için önem arz ettiğine inanılabilir. Yani, sınıf etiketi ile ilgisinin yüksek olduğu belirlenen öznitelikler çok sayıda olabilir. Gerçekte ise ilgili özniteliklerin yalnızca bir

kısmı sınıflandırma çözümü için zarurîdir. Fazlalık analizi, sınıflandırma performansının optimizasyonu için en gerekli olan öznitelikleri ortaya çıkarmaya yardımcı olur. Fazlalık oluşturan öznitelikler; seçili öznitelik kümesi tarafından temin edilen mevcut bilginin ihtiyaç dışında tekrarlandığını gösteren özniteliklerdir. Diğer ifade şekilleriyle artıklı ya da yedekli denen bu öznitelikler aralarında istatistiksel anlamda çok fazla bilgi paylaştığı için, içlerinden biri veya birkaçı seçilince geriye kalanlar seçkinliğini kaybeder. Öznitelik seçiminde genellikle, sınıflandırma için yararlı olan asgarî yedeksiz öznitelikler kümesi tanımlanmaya çalışılır ve bu yüzden anlamlı düzeyde ek bilgi sağlamayan yedekli özniteliklerin kaldırılmasına odaklı ölçüm ve analizler yapılır [96]. Sonuç olarak fazlalık ölçümü daha bilgilendirici adayların seçimine katkı sağlamaktadır.

Sınıf etiketiyle ilgisi olmadığı için seçilmeyen veya seçili özniteliklerle fazlalık oluşturduğu gerekçesiyle elenen bazı özniteliklerin aslında gözden kaçırılmış bir öneminin olması muhtemeldir. Çünkü bazı öznitelikler sadece, spesifik bir öznitelik grubunun üyesi iken sınıflandırma çözümüne fark edilir bir katkı sağlayabilir. Bu tür öznitelikler arasında etkileşim ilişkisi vardır. Birden çok özniteliğin tamamlayıcı bilgilerinin bir araya gelmesi sonucunda sınıf etiketini belirleyici kritik bilgilerin ortaya çıkması etkileşim terimi ile ifade edilmektedir. Öznitelik kombinasyonlarından oluşturulan kümeler, tahmin performansı bağlamında, içerdikleri özniteliklerin bireysel performanslarından izlenemeyen birleşik bir etkiye sahip olabilir [76]. Etkileşim gösteren grubun herhangi bir özniteliği yoksa geriye kalan öznitelikler eldeki problemle alâkasız hale gelir [97]. XOR problemi etkileşim konusunu basitçe açıklayan tipik bir örnek olarak gösterilir [98]. Açıktır ki XOR probleminin değişkenleri bireysel olarak sonuç etiketi hakkında hiçbir anlam taşımamaktadır, ancak iki girdi değişkeni bir araya getirildiğinde sonucu tamamen belirler [98]. Bu; girdi değişkenlerinin sonuç etiketini belirleme hususunda görevdeş olduğu, yani etkileşim ilişkisi gösterdiği anlamına gelir.

Çok değişkenli çözümlemelerin artı ve eksi yönlerini birkaç cümleyle özetlemek faydalı olacaktır. Çok değişkenli filtreleme yöntemlerinin, ilgi analizine ilâveten sağladığı fazlalık ve etkileşim analizleri öznitelik seçiminde anahtar roller üstlenir. İlgisiz ve yedekli öznitelikleri elerken aynı zamanda etkileşim içindeki içsel öznitelik kümelerini korumak önemlidir [98]. Tek değişkenli filtreleme yöntemleri çok hızlı çalışmasına rağmen yedekli öznitelikleri elimine edemez ve etkileşimli öznitelik gruplarını korumakta zorlanır. Bunun anlamı şudur; çok değişkenli yöntemlerle yapılan öznitelik seçimi daha etraflı analizlere

dayanır. Dolayısıyla bu yöntemler kullanılarak daha bilgilendirici öznitelik alt kümelerinin keşfedilmesi mümkündür. Bununla beraber çok değişkenli analizler sırasında daha fazla hesaplama süresi kullanılır.

Öznitelik seçiminde arama stratejileri

Öznitelik seçim yöntemlerini ayıran bir diğer önemli taksonomide ise arama prosedürü kıstas alınır. Arama prosedürü bilhassa, alt küme araması yapan sarmal yöntemler için ayırıcı bir kıstastır. Öznitelik seçim teknikleri, sınıflandırma probleminin çözümü doğrultusunda mümkün olan en ayrımcı alt kümenin bulunmasını sağlamakla görevlidir. Dolayısıyla her öznitelik seçim yöntemi izlediği analitik yapıya uygun bir arama stratejisi kullanır. Arama stratejileri, kapsamlı arama ve sezgisel arama şeklinde kategorize edilebilir [92, 99]. Tüm olası alt kümeleri değerlendirerek en uygun sonucu bulmayı garanti eden kapsamlı aramanın yüksek boyutlu veri kümelerinde kullanılması oldukça maliyetlidir [78]. Öznitelik sayısı n arttıkça, 2^n-1 ile belirlenen alt küme sayısının kombinatoryal patlaması, en uygun alt kümeyi bulma görevini hesaplanabilir olmaktan hızla uzaklaştırır [71]. Yani öznitelik sayısı arttıkça arama katmanında üstel bir şekilde büyüyen karmaşıklık, hesaplama yükünü karşılayacak geniş kaynaklar gerektirir. Uygulamada, pek çok problem için optimal çözümlere takılmak yerine makul bir sürede iyi çözümler elde etmek daha çok ilgi çekmiştir [100]. Bu nedenle, alt optimal öznitelik seçimi yapan ve daha pratik olan sezgisel arama yaklaşımları benimsenmiştir. Sezgisel yaklaşımda, aranan adayların hedefe olan katkıları, seçim sürecinin önceki adımından elde edilen sonuçları gözeten bir değerlendirme prosedürü rehberliğinde hesaplanır. Sezgisel aramada temel beklenti; deneyimlerle veya hassasiyetli seçimlerle yönlendirilen arama sürecinde, iyi alt optimum ve hatta global optimum çözümlerin diğer verimsiz alt kümelerden önce aranmasıdır [90]. Bu aynı zamanda, arama maliyetinin azaltılması maksadıyla ihtiyaç duyulan algoritmik alt yapının sağlanması anlamına gelir.

Sezgisel arama, deterministik ve rastgele arama olarak ikiye ayrılabilir [35, 73]. Bir deterministik arama yöntemi, aynı veri üzerinde çalıştırıldığı takdirde her zaman aynı sonucu üretir [99]. Deterministik arama örneği olarak sayılabilecek belli başlı yöntemler; ardışık arama, *l* ekle *r* çıkar yöntemi ve kayan aramadır. Ardışık arama birbirini takip eden sıralı adımlardan oluşur. Ardışık aramayla öznitelik seçimi, tayin edilen yöne göre, sıralı ileriye doğru seçim [101] veya sıralı geriye doğru seçim [102] yöntemleri ile yapılabilir.

Yön tayini öznitelik seçiminde izlenecek olan gidişatı belirler. İleri ve geri yönler sırasıyla ekleme ve eleme yoluyla seçim yapılması anlamına gelir. Ardışık yöntemlerde, aç gözlü tepe tırmanma yaklaşımı benimsenmiştir [78]. Sıralı ileriye doğru seçim yöntemi boş bir gen alt kümesi ile başlar ve mevcut alt kümeye tüm olası tek öznitelik genişletmeleri geçici şekilde eklenerek değerlendirilir [69]. Arama döngüsünün her tekrarında, en iyi sonuca götüren bir öznitelik kalıcı olarak seçili alt kümeye eklenir [69]. Sıralı geriye doğru seçim yöntemi ise tam öznitelik kümesi ile başlatılır ve en düşük önemdeki öznitelikler kademeli olarak kaldırılır [90, 91]. Sıralı geriye doğru aramada elenen özniteliklerin, sıralı ileriye doğru aramada ise eklenen özniteliklerin bir daha aday statüsünde değerlendirilmesine izin verilmez, dolayısıyla öznitelik seçimi geri dönüşsüzdür. Arama kuralının katılığı, geçmişte alınan kararların yeniden gözden geçirilmesini engeller. Bu durum alt küme çeşitliliğinin kısıtlanmasına yol açarak en iyi çözüme ulaşma ihtimalini azaltır. Geriye yönelik kontrolleri mümkün kılmak için *l* ekle *r* çıkar [103] veya kayan arama [81] gibi daha esnek stratejilere başvurulabilir. *l* ekle *r* çıkar yönteminde, döngüsel olarak, *l* adım sıralı ileri yönlü seçim ve ardından *r* adım sıralı geri yönlü seçim gerçekleştirilir [91]. Böylece daha önceki adımlarda seçilen öznitelikler sonradan çıkarılabilir, ancak yeterince iyi bir çözüm elde etmek için en uygun sabit *l* ve *r* değerlerini hesaplamak kolay değildir [100]. Şurası da var ki; ilerleme ve geri izleme adımlarından oluşan çevrimi sabit değerlere göre tanzim etmek öznitelik seçiminde istenen başarıyı temin etmeyebilir. Bu dezavantajlardan kurtulmak için sıralı ileriye doğru kayan seçim ve sıralı geriye doğru kayan seçim yaklaşımları önerilmiştir. Kayan arama temelli yaklaşımlarda ileri ve geri seçim adımlarının sayısı önceden sabitlenmek yerine prosedürün farklı aşamalarında dinamik olarak kontrol edilir [91, 100]. İleri yönlü kayan aramada, seçili kümeye eklenen her öznitelikten sonra geriye doğru kontrol yapılarak, önceki adıma göre performansı düşürdüğü belirlenen öznitelik veya öznitelikler silinir. Eğer performans düşürücü öznitelik zaten yoksa veya kalmadıysa ekleme işlemi kaldığı yerden devam eder. Geriye doğru kayan aramada ise benzer süreç tersten işletilir. Bu şekilde öznitelikler yeniden değerlendirilebilir ve alt kümeler birden çok kez taranabilir [100].

Gen seçimini konu alan mevcut çalışmaların bir kısmında deterministik arama yöntemleri uygulanmıştır [93, 99, 104]. Bunun yanı sıra birçok çalışmada rastgele arama tabanlı yaklaşımların tercih edildiği görülür [46, 105, 106]. Rastgele arama, olası tüm çözümleri araştırmaya gerek kalmadan yeterince iyi öznitelik alt kümeleri seçmek için sıkça tercih edilen bir sezgisel arama stratejisidir [100]. Bir rastgele arama yönteminin aynı veri

üzerinde her zaman aynı çözümü üretmesi beklenmez [99]. Arama prosedüründe kullanılan rastgelelik bölgesel zirve değerlerden kaçmaya fırsat verir. Bu nedenle, rastgele arama yöntemlerinde yerel optimuma sıkışma riski deterministik arama yöntemlerine göre daha azdır [71]. Rastgele arama prosedürü uygulayan sarmal yöntemlerin çoğunda, benzetilmiş tavlama veya popülasyona dayalı bir yöntem olan genetik algoritma kullanılmaktadır [69, 71].

Yüksek boyutlu uzayda öznitelik seçim probleminin çözümü için sarmal ve gömülü yöntemler yerine filtre tabanlı yöntemlerin kullanılması daha uygun görülmüştür [82, 92]. Bu tez çalışmasında gen seçimi için, çok değişkenli filtreleme yöntemleri kategorisinde yer alan bilgi kuramsal öznitelik seçim yöntemlerinden yararlanılmaktadır. Bilgi kuramsal yöntemlerin arama yapısından bahsetmek suretiyle bu bölümü toparlamak faydalı olacaktır. Bilgi kuramı tabanlı öznitelik seçiminde arama stratejisi olarak genellikle aç gözlü arama kullanılmıştır [107]. Bilgi kuramsal yöntemlerde belirli bir değerlendirme kriterine göre bir seferde bir öznitelik seçen artırımlı arama yapısı esas alınır ve her yinelemede değerlendirme kriteri tek bir öznitelik açısından maksimize edilmeye çalışılır [97]. Ancak daha önce ifade edildiği üzere; çok değişkenli değerlendirme kriteri, aday öznitelikler ile seçili öznitelik grubu arasındaki istatistiksel ilişkileri bir dereceye kadar hesaba katmak için formüle edilmiş kullanışlı terimler içerir. Dolayısıyla bu yöntemlerde, arama prosedürünün yinelenmeleri süresince bir yandan artırımsal adımlarla seçili küme belirlenirken diğer yandan bu kümenin elemanları birer girdi değişkeni olarak değerlendirme prosedüründe rol üstlenebilir.

2.2.4. Performans doğrulama

Öznitelik seçimi, sınıflandırma performansını destekleyici etkisi açısından tahmin modeli tasarımının ana öğelerinden birisi konumundadır. Öznitelik seçimi tamamlandığında elde edilen sonucun doğrulanması gerekir. En bilinen doğrulama yaklaşımı, seçili özniteliklerden oluşan alt kümenin sınıflandırma performansının ölçülmesidir. Verilerden modelleme yapılırken hem öznitelik seçimi hem de sınıflandırmanın birincil hedefi tahmin performansının iyileştirilmesidir. Bu bağlamda, sınıflandırma performansını ölçmek için kullanılan metriklerin tahminci modelleri değerlendirmek için kullanılması da mümkündür. Daha açık ifade etmek gerekirse; hem öznitelik seçim yöntemini hem sınıflandırma

yöntemini hem de bu yöntemleri bünyesine katmış olan bir tahmin modelini değerlendirmek için çoğu zaman aynı ölçütlerden yararlanılır.

Model performansını ölçmek amacıyla veri örneklerinin bir kısmı test örneği olarak ayrılır, geriye kalan kısmı ise eğitim örneği olarak kullanılır. Daha kapsamlı ve genellenebilir bir sonuca ulaşmak için, eğitim ve test kümelerinin belirlenmesinde dönüşümlü bir yol izleyen çapraz doğrulama teknikleri kullanılmaktadır. Bu tekniklerin en bilineni *k*-kat çapraz doğrulamadır. *k*-kat çapraz doğrulamada veri kümesi eşit boyutlu *k* adet parçaya ayrılır. Sonra, *k* kez tekrarlanan bir döngü içinde bu parçalardan her biri birer kez test kümesi olarak seçilir. Her yinelemede test kümesi görevini devralan parça hariç tutulur ve öteki örneklerin tümü eğitim için kullanılır. Döngü tamamlandığında, eldeki *k* adet performans ölçümünün aritmetik ortalaması alınarak sonuç değerine ulaşılır. Genelleme hatasını tahmin etmede *k* için 10 değerinin kullanımı yaygındır [72].

k-kat çapraz doğrulamada *k* değerinin örnek sayısına eşit alınması LOOCV (Leave-One-Out Cross-Validation) yöntemi olarak adlandırılır. Bu yöntemde veri kümesindeki her bir örnek bir kez test kümesi olacak şekilde model kurulumunun dışında bırakılır. Bütün veri örnekleri bir defa dolaşılarak ölçümler tamamlanır. Modelin, örnek sayısınca eğitilmesi ve test edilmesi sonucunda elde edilen tüm değerlerin aritmetik ortalaması tahmin performansını verir. LOOCV yöntemi, örnek sayısı az olan veri kümelerinde uygulanabilir bir doğrulama yöntemidir [99].

3. KARAR VERME

Yaşamın her kesitinde ve her döneminde, farklı orijinlerde gelişen pek çok karar problemiyle karşılaşılmaktadır. İnsanlar, sorumlu veya yetkili oldukları alanlarda ortaya çıkan problemlere çözüm getirmek için ya da bir takım hedeflere ulaşmak niyetiyle; kişisel, kurumsal, örgütsel, toplumsal yahut bilimsel düzeyde kararlar almaya ihtiyaç duyarlar. Her kararın, belirsizliklere son vermek açısından bağlayıcı bir unsur olduğunu hatırlamak ve gelişigüzel karar vermekten kaçınmak gerekir. Bir karar probleminin yüzeysel yaklaşımlarla ele alınması çözüm arayışının başarısızlıkla sonuçlanmasına yol açabilir. Bu nedenle, mevcut bilgi ve birikim güçlü muhakeme yetisi çerçevesinde dikkatlice değerlendirilmelidir. Karar problemine ilişkin elde yeterli verinin bulunması ve karar verici pozisyonunda ehliyetli kimselerin görevlendirilmesi kararsızlıkların aşılmasında önem taşıyan hususlardır. Bunların yanı sıra, karar süreçleri sonunda azamî fayda sağlayabilmek için öne sürülmüş olan çeşitli sayısal formülasyonlar bir hayli ilgi çekmiştir. Zaman içerisinde, karar problemlerinin çözümünde metodolojik yollara başvurmanın önemi giderek artmış; karar vermede bilimsel yöntemlerin kullanılması yaygınlaşmıştır. Belirli kriterlere uygun şekilde karar vermek maksadıyla ileri sürülen yöntemler ve bu yöntemlerin karar problemlerine tatbik edilişi akademik literatürde oldukça sık çalışılmış konular arasında yer almaktadır.

3.1. Karar Vermeye İlişkin Temel Kavramlar

Kararsızlık, kaygı ve tereddütleri ihtiva eden bir soruna işaret eder ve aslında yaşam boyu en sık karşılaşılan problemlerden bir tanesidir. Kararsızlığın sebebiyet verdiği duraksama hâlinden çıkmak için mevcut alternatifler arasından en uygun olanı seçilmeli ve bu seçim doğrultusunda harekete geçilmelidir. Seçim yapma işlemi söz konusu olduğuna göre, bir karar probleminde zihni kurcalayan hususların açıklığa kavuşturulması için tercihe dayalı çözüm yollarına başvurulması gerekir. O hâlde karar verme; bir tercih belirtmeyi ve bu tercihle belirtilen çözüm yolunu uygulamaya koymayı ifade eder. Karar verme sayesinde mantıksal temelde tereddütlerin nihayet bulduğu bir sonuca varılır. Bu katî sonuca karar denir ve uygulamaya esas teşkil eden tercihi temsil eder. Daha yalın bir tanımlama yapmak gerekirse karar; yanıt bekleyen bir belirsizlik hâlini sonlandırmaya yönelik verilen bilinçli bir karşılıktır.

Karar probleminin farkına varılması, karar verme sürecinin başlangıç noktasıdır. Öncelikle, fark edilen problemi tanımak ve tanımlamak gerekir. Bir karar probleminin yapısı birtakım bileşenler aracılığıyla daha anlaşılır biçimde açıklanabilir. Karar probleminin tanımlanmasında ve karar verme sürecinin tesisinde rol alan elemanlar şunlardır [108, 109]:

- Karar verici; karar sürecini kontrol altında tutan ve karar verme mesuliyetini üstlenen kişi veya topluluktur. Karar verici mevcut bilgileri ve geçmişten gelen tecrübelerini kullanarak problem alanına müdahale eder; karar sürecini yapılandırmak için gerekli çabayı gösterir. Bu bakımdan, karar vericinin yeterli ölçüde serbestliğe sahip olması önem taşır.
- Amaç; karar vericinin karar sürecindeki faaliyetleri ile ulaşmayı hedeflediği durumdur.
- Karar kriteri; karar vericinin alternatif çözüm yolları arasından seçim yapabilmek için kullandığı değer yargısıdır. Kriter kelimesi yerine anlamdaşı olan ölçüt kelimesi de yaygın olarak kullanılmaktadır. Karar kriteri ya da diğer kullanım şekliyle karar ölçütü, amaca ulaşmada katkı veya maliyet belirten bir değişkeni ifade eder.
- Alternatif (seçenek); belirlenen amaçlar doğrultusunda değerlendirmeye tâbi tutulan mümkün çözümlerden her birisidir. Karar verici kendi kontrolünde olan bir seçenek kümesi içinden tercih yapar. Eğer bir tercih yapmaktan söz ediliyorsa bu, en az iki ayrı seçeneğin analiz edilmesi anlamına gelir. Alternatiflerin önem dereceleri, kriterler altında elde edilen değerlendirme puanlarına göre hesaplanır.
- Karar ortamı (olaylar); karar vericinin kontrolü altında olmayan ancak seçim sürecini etkileyen çevre şartlarını ifade eder.
- Sonuç; her bir alternatif ve olay birlikteliği neticesinde ortaya çıkan değerdir.

Karar vermede kalitatif ve kantitatif olmak üzere iki genel yaklaşımdan bahsedilebilir [110]. Karar verici, daha önce bir benzeriyle karşılaştığı yahut kolayca çözülebileceğini düşündüğü problemleri ele alırken, temel bilgi ve deneyimlerini kullanarak karar verme sürecinde sezilerini ön plana çıkarabilir [110]. Karar verme, geçmişten gelen bilgi birikimi doğrultusunda karar vericinin tasarrufunda gerçekleştiriliyorsa böyle bir yaklaşıma kalitatif karar verme denir. Ancak çoğu zaman sezgiler karar problemi ile başa çıkmak için yeterli değildir. Pek çok karar probleminde, karar vericinin işe yarar ve tutarlı bir sonuca bir çırpıda ulaşması zordur. Karar verici, karar sürecinin yapılandırılmasına müdahil olabilse dahi yanılgıya düşmemek için salt öznel yargılardan mümkün mertebe uzak durmalıdır.

Karar vericinin bocalayıp çelişkiye düşmesine neden olan çapraşık problemlerin analizi için kalitatif yöntemler genellikle elverişsizdir. İsabetli karar vermenin zor olduğu zamanlarda, yanılma riskinin görmezden gelinmesi doğru olmaz. Dolayısıyla daha objektif çözüm önerilerine ihtiyaç duyulur. Böyle girift görünen bir durum varsa karar süreci matematiksel yöntemler kullanılarak daha disiplinli bir şekilde yönetilebilir. Bunun için kantitatif yöntemlere başvurulur. Kantitatif karar verme yöntemleri; mevcut verileri analiz ederek karar sürecinde ortaya çıkan tüm soru işaretlerini mantıksal bir çerçevede çözüme kavuşturmayı amaç edinen yöntemlerdir. Bu yöntemlerde, karar verme sürecinin denetimli ilerleyişi için sayısal hesaplamalara dayalı analizler esas alınır. Sayısal yaklaşımlar profesyonel karar vermeye geçişte etkin rol oynamaktadır.

Karar mekanizması en az bir ölçüte göre dizayn edilir. Bunun anlamı şu ki; tek bir kritere odaklanmak mümkün olduğu gibi, karar mekanizmasına birden çok kriterin katılımı da sağlanabilir. Bu bakımdan, karar vermede kullanılan yöntemler tek kriterli ve çok kriterli olmak üzere iki alt sınıfa ayrılmıştır. Gerçekte karşılaşılan problemlerin pek çoğunda, uygulamaya geçilecek olan çözümün birden fazla karar kriterini aynı anda sağlaması istenir. Dolayısıyla çoklu kriterler varlığında karar verilmesi gerekir. Çok kriterli karar verme, bu tez çalışmasının ana ekseninde yer alan konulardan olup; çalışmanın izleyen bölümünde genel hatlarıyla ele alınmaktadır.

3.2. Çok Kriterli Karar Verme (ÇKKV)

Karar verme sürecinde alternatifler ve amaçlar arasındaki ilgi kriterler aracılığıyla kurulmaktadır. Alternatiflerin, belirli bir amaç doğrultusunda bir veya daha fazla kriteri karşılaması beklenir. Tek kriterli karar vermede dikkatler yalnızca bir kriter üzerinde toplanmakta ve analizler sırf bu kriterle bağlantılı olarak gerçekleştirilmektedir. Ancak tek bir kritere göre optimize edilen karar, başlangıçta belirlenen amaç ve istekleri gereken ölçüde karşılamıyor olabilir. Bu durum verilen kararın yeterince ikna edici olmadığı düşüncesini akla getirir. Öyleyse belirli bir amaca ne derecede ulaşıldığını değişik yönlerden ölçmek için karar sürecinde bir dizi farklı kriterden yararlanmak makul bir yaklaşımdır. Karar verici, önemli gördüğü kriterleri sürece dâhil etmeli ve potansiyel çözümler arasından en doğru olanın izini sürerken yol göstermesi için bu kriterleri kullanmalıdır. Bu sayede, hem karar probleminin daha gerçekçi bir şekilde

modellenmesine imkân tanınmış olur hem de verilen kararın değişik yönlerden onanmış olması sağlanır.

Karar vermede başlıca ekollerden biri olan ÇKKV; karmaşık bir karar problemini birim kriterler altında ayırıp her kriter için öncelik sıralaması elde edilmesini, ardından bunu gitgide geliştirerek bütünsel tercih sırasının elde edilmesini sağlamaktadır [111, 112]. Çoklu kriterler söz konusu olduğunda bu kriterlerin tümüne göre optimal olan bir alternatif bulabilmek oldukça güçtür. Çünkü çoğu durumda, mevcut alternatifler arasında tam anlamıyla ideal bir alternatif yoktur. Öyleyse, mümkün olan en iyi çözümü arayıp durmak yerine var olan seçenekler arasındaki en iyi çözümü bulmaya odaklanmak daha doğru olur. Bu bağlamda ÇKKV, akla uygun matematiksel bir bakış açısı sağlamaktadır. ÇKKV yaklaşımları uygulanarak, optimal olmasa bile uzlaşmaya dayalı memnun edici bir sonuç bulunabilir; üstelik birbirleri ile çelişir durumda olan birçok kriter bir arada ele alınabilir [112]. ÇKKV disiplini çerçevesinde önerilen uzlaşmacı çözüm; kriterler arası denge ve tolerans gözetilerek elde edilen ve aynı zamanda bütünsel çerçevede en uygulanabilir olan çözümü simgeler. Açıkça anlaşıldığı üzere bir ÇKKV probleminin çözümü; kesikli bir uzaydaki, zaten bilinen, net olarak belirtilmiş, sayılı seçenekten birini veya birkaçını kapsar [113, 114]. Eğer karar problemi; sürekli bir uzaydaki sonsuz seçenek varlığında belirli amaç fonksiyonlarının optimizasyonuyla ilgiliyse ve en uygun alternatifin tasarlama yapılmak suretiyle elde edilmesini gerektiriyorsa bu problem Çok Amaçlı Karar Verme'nin (ÇAKV) konusudur [113, 114].

Bir ÇKKV probleminin detaylarıyla ortaya koyulabilmesi için; amaçların, tercih kümesinin ve çoklu kriterlerin açık ve anlaşılır olmasına özen gösterilmesi gerekir. Çoklu kriterler belirlenirken dikkate alınması gereken hususlardan bazılarının altını çizmekte fayda vardır. Karşılıklı eşsizlik bu hususlardan biri olup; her bir kriterin bir diğeri tarafından ölçülemeyen ayrı bir özelliği ölçmesi gerektiğini ifade eder [115, 116] Önemli hususlardan biri de kriterler arası artıksızlık sağlanmasıdır, yani kriterler kümesi olabildiğince küçük tutulmalıdır [115]. Artıksız kriterler problemin daha iyi tanımlanmasına yardımcı olup, alternatifler arası kıyaslamalarda ve karar sonucu üzerinde etki gösterir; gereksiz yere kullanılan kriterler ise alternatiflerin mukayese edilmesine yönelik bir etki göstermez [116]. Karar analizi açısından anlamsız olan bu tür kriterler problemi karmaşıklaştırıp, hedefe odaklanma konusunda engel teşkil edebilir.

ÇKKV problemleri; seçim, sıralama ve sınıflama başlıkları altında üç kola ayrılır [114, 115]. Seçim problemlerinde muhtemel çözüm seçenekleri içerisinden en uygun olan bir veya birkaç tanesinin belirlenmesi istenir; sıralama problemlerinde alternatifler tercih edilme önceliklerine göre en cazip olandan en avantajsız olana doğru sıraya koyulur; sınıflama problemlerinde ise alternatiflerin alt sınıflara ayrışması, böylelikle benzer alternatiflerin aynı sınıf altında toplanması istenir [114, 115]. Sıralama ve sınıflama problemleri genellikle daha karmaşık bir karar verme sürecinin bir alt bileşeni olarak ortaya çıkmaktadır [115, 116].

ÇKKV problemlerinin genel özellikleri bilinmeli ve problem çözümünde kullanılan yöntemlerin hangi kavramlar etrafında formüle edildiği iyi anlaşılmalıdır. Bu kapsamda öncelikle nitelikler arası zıtlıktan, arkasından kriter ağırlıklandırma ve karar matrisi gibi bazı önemli kavramlardan söz edilecektir. Nitelikler arası zıtlık ÇKKV problemlerinin tipik özelliklerinden biridir [117, 118]. Bir dizi kritere göre gerçekleştirilen değerlendirmelerin uyuşmazlık gösterip birbirine ters düşmesi nitelikler arası zıtlık olarak tabir edilir. Kriterler altındaki değerlendirme puanlarına göre alternatiflerin tercih sırasını gösteren listeler arasında birebir uyuşma oldukça seyrek görülmekte olup, çoğu durumda aralarında çelişki açığa çıkmaktadır. Yani bir alternatiften diğerine geçildiğinde kriterlerden kimilerinin değeri iyileşirken, kimilerinin değerinde kötüleşme meydana gelir [119]. Bu; bazı kriterler açısından daha avantajlı olan bir alternatifin bazı kriterler açısından zayıf olması sonucu ister istemez ortaya çıkan bir durumdur [119]. Aslına bakılırsa nitelikler arası zıtlık, ÇKKV probleminin içinden çıkılması zor bir hâle gelmesine zemin hazırlayan en önemli etmendir [119]. Farklı kriterler altındaki değerlendirme sonuçlarının mantıken birbirini tutmaması, ÇKKV probleminin çözümünde kriterler arası tavizci bir yapının tesisini gerektirir. Bir çözüm kümesi üzerinde anlaşma sağlanabilmesi için karar kriterleri arasında toleransa dayalı bir denge kurulması gerekir, aksi takdirde anlaşma sağlanması neredeyse imkânsızdır. Bu yüzden, ÇKKV problemlerine çözüm aranırken esasında kriterler arasında ödünleşime müsaade edilerek sistemli bir şekilde uzlaşıya varılmak istenir [8].

Yeri gelmişken kriter ağırlıklandırma işlemine açıklama getirmek uygun olacaktır. Öznel veya nesnel kriter ağırlıklarının ortaya çıkarılması ÇKKV sürecinin önemli bir adımıdır [118, 120]. Birden fazla kriter kullanılması durumunda kriterler arasında hassas bir denge kurulması istendiği için hesaplamaların titizlikle gerçekleştirilmesi gerekir. Eğer

hesaplamalara etkisi açısından kriterler arasında anlamlı bir fark olduğu düşünülmüyorsa çoklu kriterlerin hepsinin eşdeğer ağırlıkta olduğu varsayılabilir. Ancak eşit ağırlık atama gibi sıradan bir yaklaşım her problem için başarılı bir çözüm stratejisi değildir. Bilindiği üzere kriterler bir ÇKKV problemindeki farklı boyutları temsil eder. Doğrusu yalnızca bu düstur göz önüne alındığında bile, çoklu kriterlerden her birinin farklı önem ağırlığına sahip olduğu izlenimi güç kazanmaktadır. Karar sonucuna etkisi olan bu önem ağırlıklarını belirlemek ve analiz sürecinde kullanmak yerinde bir tutum olacaktır. Çünkü kriter ağırlıklandırma, çoklu kriterler arasındaki ödünleşme stratejisini daha yararlı hâle getirmek için yapılmaktadır. Ağırlıklar nesnel yöntemlerle hesaplanabildiği gibi karar vericinin tercihlerini dikkate alan öznel yöntemlerle de hesaplanabilir; hatta ağırlıkları doğrudan karar verici atayabilir [118]. Nesnel ağırlıklar, problem çözümüne esas teşkil eden veriler üzerindeki sayısal hesap ve ölçümlere dayanır; fakat öznel ağırlıkların belirlenmesi sürecinde karar vericinin inisiyatif ve yargılarına başvurulması gerekir [118]. Çözüme yansıtılmak istenen kriter ağırlıkları uygun bir yöntem yardımıyla belirlendikten sonra genellikle bir vektör formunda hesaplamalara dâhil edilir.

ÇKKV yöntemlerinin hemen hemen hepsinde kıyasçı yapının temelini oluşturan bir karar matrisi kullanılır. Karar matrisi; tüm alternatif ve kriterlerin sırasıyla satır ve sütunlarda temsil edildiği veri yapısını belirtmekte olup, bu matrisin satır ile sütunlarının kesişim noktalarında alternatiflerin kriterler altındaki değerlendirme sonuçları tutulmaktadır [121].

Son olarak, bu konunun iyice pekişmesi amacıyla ÇKKV yöntemleri hakkında genel bilgi ve izahata yer verilecektir. Karar probleminin çözüme kavuşturulmasında birden çok kriterin etkisi önemseniyorsa ÇKKV yöntemleri uygulanır. Çok kriterli bir karar ortamının modellenmesi amacıyla kullanılan matematiksel yöntemler karar sürecinin eksiksiz bir şekilde yapılandırılması açısından son derece önemlidir. Çoklu kriterlerden oluşan karar problemlerine çözüm sunulurken bu kriterler arasında uzlaşı sağlamaya yönelik yoğun bir düşünsel çaba sarf edilmesi gerekir. Karar verici bu düşünsel uğraşıya doğrudan doğruya, yani bilimsel bir yöntemden yardım almadan yön vermekte zorlanabilir. Uygulamada, bu konudaki iş yükünün büyük kısmını ÇKKV yöntemleri üstlenir. Yani bir ÇKKV yöntemi kapsamında; karar sürecinin mantıksal çerçevesi ve hesaplamalarda kullanılacak olan formüller belirgin bir şekilde ortaya koyulmuştur. Dolayısıyla karar vericinin, ilgili ve yetenekli olduğu alan dışında karar bilimine yönelik herhangi bir konuda çok fazla uzmanlaşmasına gerek yoktur. Kısacası; uzlaşı esasına göre hareket edilerek en avantajlı

ödünleşmenin gerçekleştirilmesi için ÇKKV yöntemleri analitik bir alt yapı sağlamaktadır. Bu yöntemler kullanılarak hedefe yönelik işlemlerin yerine getirilmesi ve karar analizlerinin kolaylıkla yürütülmesi sağlanır.

Ana hatlarıyla tarif etmek gerekirse bir karar verme yaklaşımı; genel mantık ve anlayışa uygun olmalı, rahatlıkla açıklanabilen sade bir yapıya sahip olmalı, görüş birliğini teşvik etmeli ve aşırı ihtisaslaşmayı gerektirmeden pratik bir şekilde kullanılabilmelidir [122]. ÇKKV yöntemleri geliştirilirken genel anlamda bu hususlar dikkate alınmış olup, bu sayede basit ama kullanışlı yöntemler önerilmiştir. ELECTRE (Elimination and Choice Expressing Reality), Gri İlişkisel Analiz (Gray Relational Analysis - GRA), Veri Zarflama Analizi (VZA) (Data Envelopment Analysis - DEA), Basit Toplamlı Ağırlıklandırma (Simple Additive Weighting - SAW), İdeal Çözüme Benzerliğe göre Tercih Sıralama Tekniği (Technique for Order Preference by Similarity to Ideal Solution - TOPSIS), AHS ve VIKOR yöntemleri ÇKKV disiplini uyarınca önerilmiş olan popüler yöntemlere örnek olarak gösterilebilir [123, 124].

ÇKKV yöntemleri pek çok problem alanına uyarlanabildiği için bu yöntemlerin geniş bir uygulama alanı bulduğu ve bu kapsamda geniş bir literatür oluştuğu görülmektedir. Literatür araştırmalarına bakılırsa ÇKKV yöntemlerinin; tedarik zinciri yönetimi, yer seçimi, materyal seçimi, insan kaynakları yönetimi, finansal yönetim, pazarlama, su kaynakları planlaması, sürdürülebilir enerji planlaması, performans değerlendirme gibi problem alanlarında yaygın şekilde kullanıldığı görülür [125, 126]. Makine öğrenmesi ve ÇKKV disiplinlerinin beraber kullanımını konu alan çalışmalar görece daha azdır; bu disiplinlerin problem çözümünde sentezci bir yaklaşımla uygulandığı çalışmalara da oldukça az rastlanmaktadır. Tez çalışması kapsamında önemli olan bu konuya açıklık kazandırmak için bir sonraki bölümde bu hususta bir literatür araştırmasına yer verilmesi uygun görülmüştür.

3.3. Literatürde Makine Öğrenmesi Alanında Uygulanan ÇKKV Yaklaşımları

Karar verici için; amaçlar, alternatifler ve çelişir durumdaki çoklu kriterler arasındaki bağlantıları kurup, bu bileşenler etrafında karar ortamını modellemek akıl karıştırıcı bir süreçtir. Karar verici bu süreci en uygun şekilde sonuca ulaştırmanın çabası içindedir. ÇKKV metotları sayesinde, karar problemi ile ilgili veriler kolayca faydalı hâle

getirilmektedir. Bu nedenle ÇKKV yöntemleri, çeşitli araştırma ve uygulama alanlarında ortaya çıkan karar problemlerinin çözümü için karar vericiler tarafından sıkça tercih edilen yöntemlerdir. Literatürde ÇKKV yöntemleri ile makine öğrenmesi yöntemleri bazı araştırmalar kapsamında bir arada kullanılmıştır. Bu çalışmaların büyük bir bölümü, makine öğrenmesi tabanlı algoritmaların değerlendirilmesi amacıyla ÇKKV yöntemlerinden yararlanılmasını konu alır. Bazı çalışmalarda ise araştırılan problemin çözümü için her iki disiplinden belirli yöntemler seçilmekte olup, bu yöntemlerin bileşimi ile elde edilen yeni bir yaklaşım önerilmektedir.

Makine öğrenmesi kapsamında ele alınan özellik çıkarma, öznitelik seçimi, kümeleme ve sınıflandırma problemlerinin çözümüne yönelik önerilmiş çok sayıda algoritma mevcuttur. Bir makine öğrenmesi yönteminden elde edilen performans, ele alınan problem verisine göre değişir. Belirli bir problem üzerinde iyi performans gösteren bir algoritma, başka problemlerin çözümünde yeterince etkili olmayabilir. Örneğin, sınıflandırma problemlerinin çözümü için her zaman en iyi sonucu verdiği kabul edilen bir sınıflandırıcı yoktur. Öte yandan, eldeki veriye en uygun olan algoritmanın kullanılması problem çözümünde iyi bir performans yakalamaya yardımcı olur. Makine öğrenmesi disiplini kapsamındaki çeşitli algoritmalar arasından hangisinin belirli bir problemin çözümü için en uygun olduğuna deneysel yollar kullanılarak karar verilir. Şu da var ki; algoritmaların farklı yönlerden değerlendirilmesini sağlamak amacıyla kullanımda olan çok sayıda metrik vardır. Algoritma değerlendirme göreviyle ilgili özgün bir öneme sahip olan her bir metrik belirttiği performans değerleri ile çözüme katkı sağlayabilir. Performans değerlendirmesi yapılırken çoğu zaman, birden çok metriğin aynı anda kullanılabilmesi için uygun alt yapı sağlayan yöntemlere ihtiyaç duyulur. Çünkü birden fazla metriğin dikkate alındığı değerlendirme süreçlerinde birbirinden farklı metriklere göre ölçülen performans sonuçlarının çelişir durumda olması karar vericilerin işini zorlaştırır. Bu durumda algoritma seçim problemi; bir dizi kriter ve bu kriterlere göre hesaplanan deneysel performans değerleri esas alınarak bir ÇKKV problemi olarak modellenebilir [127, 128].

Literatürde, belirli görevler için makine öğrenmesi algoritmalarının seçimine yönelik önerilen ÇKKV temelli çeşitli araştırma çerçevelerine rastlanması mümkündür. Bu araştırmalarda çoğunlukla, algoritmaları belirli veriler üzerinde gösterdiği performans değerlerine göre ÇKKV yaklaşımı temelinde değerlendirme çabası ön plândadır. Bunun için çeşitli algoritmalar karar sürecinde birer alternatif olarak temsil edilir. Ayrıca,

algoritmaların performansını göstermesi için kullanılan metrikler içinden bir kriter kümesi belirlenir. Bir kriter kümesinin oluşturulması için genellikle performans ölçümünde kullanılan metriklerden yararlanılır, ancak bunun yerine algoritma seçiminde etkili olan başka nitelikler de kullanılabilir. Alternatiflerin performans değerleri, belirlenen kriter kümesi uyarınca belirli ÇKKV yöntemleri ile ele alınır. Aşağıda, ÇKKV yöntemlerinin algoritma seçim problemlerine çözüm getirilmesi amacıyla kullanılmasını konu alan bazı çalışmalar hakkında açıklayıcı bilgiler verilmiştir.

Kou ve arkadaşları tarafından finansal verilerin analiz edildiği bir çalışmada, veriler üzerinde 6 adet kümeleme algoritmasının sıralaması VIKOR, TOPSIS ve VZA olmak üzere 3 ÇKKV yöntemi ile ayrı ayrı gerçekleştirilmiştir [128]. Bu amaçla kredi riski ve iflas riski ile ilgili 3 ayrı veri kümesi kullanılmıştır. ÇKKV yöntemleri tarafından kullanılacak olan kriter setinin oluşturulması içinse kümeleme algoritmalarını değerlendirmede kullanılan 11 ayrı küme geçerlilik indeksi belirlenmiştir. Değerlendirmelerden elde edilen sonuçlara göre; ÇKKV yöntemleriyle oluşturulan sıralamaların genel olarak en iyi kümeleme algoritmaları üzerinde hemfikir olduğu belirtilmiştir. Ayrıca, herhangi bir veri kümesi için tüm ölçümlerde en iyi performansı elde edebilen bir algoritmanın olmadığı, bu sebeple kümeleme algoritmalarını çoklu kriterlere göre değerlendirmenin gerekli olduğu ifade edilmiştir.

Finansal veriler üzerinde gerçekleştirilen başka bir çalışmada, temerrüde düşen banka kredilerini tahmin etmek için bir dizi tahmin modeli kullanılmış ve bu modeller TOPSIS yöntemiyle değerlendirilmiştir [129]. İlk başta, banka kredisi verileri üzerinde temel bileşenler analizi ve bağımsız bileşen analizi olmak üzere 2 ayrı özellik çıkarım yöntemi uygulanmıştır. Böylelikle boyutları indirgenmiş olan 2 yeni veri kümesi elde edilmiştir. Daha sonra, bu veri kümelerinde görülen sınıf dengesizliği problemi örnekleme yapılmak suretiyle çözülmüştür. Tahmin modelleri oluşturulurken; NB, lineer lojistik regresyon, sınıflandırma ve regresyon ağacı, kNN ve C4.5 olmak üzere toplamda 5 ayrı sınıflandırıcı dikkate alınmıştır. Bu yöntemler, özellik çıkarma yoluyla boyutları indirgenen veri kümeleri üzerinde kullanılarak sınıflandırma yapılmış ve tahmin performansı 4 farklı metrikle ölçülmüştür. Tahmin modelleri, elde edilen performans değerleri temelinde TOPSIS yöntemiyle değerlendirilmiştir. Değerlendirme süreci veri kümelerine ayrı ayrı uygulanmış olup; bu süreç sonunda temerrüt tahmininde her iki veri kümesi için de kNN algoritmasının en iyi potansiyele sahip olduğu sonucuna varılmıştır.

Peng ve arkadaşları, sınıflandırma performansını artırmak ve veri kümeleri için uygun sınıflandırıcıları belirlemek amacıyla entegre bir şema önermiştir [130]. Bu entegre şemada öznitelik seçim tekniklerinin ve sınıflandırma algoritmalarının değerlendirilmesi olmak üzere iki tür algoritma değerlendirme görevi ÇKKV problemi olarak formüle edilmiştir. Bu kapsamda; 11 öznitelik seçim yöntemi, 9 sınıflandırıcı ve 5 ayrı ÇKKV yöntemi kullanılmıştır. Veri kümeleri iki farklı aşamada ele alınmıştır. İlk aşamada geleneksel öznitelik seçimi yapılmış olup, sınıflandırma algoritmaları veri kümelerine uygulanarak tahmin sonuçları elde edilmiştir. İkinci aşamada; geleneksel öznitelik seçim teknikleri ve ÇKKV yöntemlerini bütünleştiren bir yaklaşım önerilmiştir. Burada ilk olarak, geleneksel tekniklerle öznitelik seçimi yapılmış; her öznitelik için 11 ayrı öznitelik seçim tekniğinden elde edilen sayısal değerler kullanılarak o özniteliğin önemini belirten genel bir ağırlık değeri hesaplanmıştır. Bu adımdan sonra, önerilen öznitelik seçim yaklaşımının daha da geliştirilmesi amacıyla ÇKKV yöntemleri sürece dâhil edilmiştir. Daha iyi performans gösteren öznitelik seçim teknikleri ÇKKV yöntemleri kullanılarak belirlenmiş ve belirlenen tekniklere göre özniteliklerin ağırlıkları yeniden hesaplanmıştır. Elde edilen ağırlık değerlerine göre öznitelik seçimi yapılmıştır. Seçilen öznitelikler sınıflandırmada kullanılmıştır. Önerilen yaklaşımın, tahmin doğruluğunu ne şekilde etkilediğini incelemek için birinci ve ikinci aşamanın sınıflandırma sonuçları karşılaştırılmıştır. Son olarak, 5 farklı performans ölçütüne göre sınıflandırıcıları karşılaştırmak için ÇKKV yöntemleri yine uygulanmış ve sınıflandırma probleminin çözümüne yönelik öneride bulunulmuştur.

Yukarıdakine benzer içerikli bir çalışmada, yazılım kusuru tahmininde doğruluk ve güvenilirliğin artırılması için önemli bir adım olan öznitelik seçimi ele alınmış olup, geleneksel öznitelik seçim yöntemleri ile ÇKKV yöntemleri bütünleştirilmiştir [131]. Yazılım hatası ile ilgili 4 ayrı veri kümesi üzerinde 11 öznitelik seçim tekniğini 5 ayrı ÇKKV yöntemi kullanarak değerlendiren araştırmacılar, yüksek sıralanmış 5 tekniğe göre öznitelikleri ağırlıklandırıp öznitelik seçimi yapmışlardır. Yani, tahminci öznitelikler belirlenirken ÇKKV yöntemlerinin algoritma seçimine yönelik önerileri hesaba katılmıştır. Önerilen yaklaşımın doğrulanması için 9 sınıflandırma yöntemi kullanılarak deneysel bir çalışma gerçekleştirilmiştir. Ayrıca öznitelik seçiminden sonra ÇKKV yöntemleri sınıflandırma algoritmalarının değerlendirilmesi maksadıyla tekrar kullanılmıştır.

Yazılım kusuru konusunun ele alındığı bir başka çalışmada ise kullanıcı tercihleri algoritma seçim sürecine yansıtılmış ve yazılım kusuru tahmini için uygun bir sınıflandırıcı

seçme görevi ÇKKV problemi olarak modellenmiştir [132]. Yazılım kusuruyla ilgili 10 veri kümesi üzerinde 38 farklı sınıflandırma algoritmasının performansının karşılaştırılması amacıyla 4 ayrı ÇKKV yöntemi kullanılmıştır. Bu süreçte 13 değerlendirme kriterinden yararlanılmış olup, karar vericinin bu kriterlere ağırlık atamasına izin verilmiştir. Böylece kendi tercihlerini sıralama prosedürüne dâhil edebilmesi için karar vericiye imkân tanınmıştır. Sınıflandırma algoritmaları; kriter ağırlıkları ve veri kümelerinden elde edilen deneysel sonuçlar dikkate alınarak ÇKKV yöntemleriyle değerlendirilmiştir.

Hsu ve Pai önerdikleri yaklaşımda, finansal krizi analiz etmek için farklı öznitelik seçim yapılarına sahip DVM modellerini kullanmış ve aralarından en uygun olanı seçmek için VIKOR yöntemini uygulamıştır [133]. Bu çalışmada öncelikle, topluluk öğrenmesine dayanan kombinasyon prensipleri kullanılarak çoklu öznitelik seçim şemaları oluşturulmuştur. DVM ile birlikte kullanılması hâlinde bu şemalardan hangisinin finansal krizin tahmini için daha iyi sonuç verdiği araştırılmıştır. Bu çerçevede, yapı seçimi problemi 3 ayrı sınıflandırma ölçütü kullanılarak bir ÇKKV görevine dönüştürülmüştür. 11 farklı öznitelik seçim yapısı ile kombine edilen DVM sınıflandırıcıları VIKOR yöntemiyle değerlendirilmiştir.

Kuru göz teşhisinde önemli yeri olan gözyaşı filmi sınıflandırma problemi Barral ve arkadaşları tarafından makine öğrenmesi, ÇKKV yöntemleri ve çatışma yönetimi çerçevesinde ele alınmıştır [134]. Öncelikle, gözyaşı filmi lipit tabakası sınıflandırma probleminin çözülmesi için sınıf binarizasyonu, öznitelik seçimi ve sınıflandırma yöntemlerinin katılımıyla oluşturulan çeşitli çözüm seçenekleri aday gösterilmiştir. Bu çözümler oluşturulurken 4 binarizasyon tekniği, 3 filtre temelli öznitelik seçici ve 5 sınıflandırıcıdan yararlanılmıştır. Önerilen tüm çözüm seçenekleri karar kriterleri olarak belirlenen 7 performans ölçütüne göre TOPSIS, VIKOR ve GRA yöntemleri ile ayrı ayrı değerlendirilmiştir. Böylece sınıflandırma performansını optimize etme başarılarına göre çözüm seçenekleri sıralanmıştır. Değerlendirme sürecinde uygulanan ÇKKV yöntemlerinin çelişen sıralamalar üretmesi nedeniyle çatışma yönetimine ihtiyaç duyulmuştur. Çatışma yönetimi için sıralamalar arasındaki istatistiksel ilişki ölçüsüne dayanan bir teknik uygulanmıştır. Sıralamadaki çatışmaların çözümüne yönelik uygulanan işlemler neticesinde ÇKKV yöntemlerinin çelişkili sonuçları tek bir sıralamada birleştirilmiştir.

Çatışma yönetiminin ele alındığı bir başka çalışmada ise çok sınıflı sınıflandırma görevinde etkili algoritmaların tespiti için ÇKKV yöntemleri kullanılmış olup, 4 farklı ÇKKV yönteminin kullanılması sonucu sıralamada ortaya çıkan çatışmaların çözülmesi için bir füzyon yaklaşımı önerilmiştir [135]. Önerilen füzyon yaklaşımında, uyumsuz ÇKKV sıralamaları üç adımda ele alınmıştır. İlk adımda sınıflandırma algoritmalarının sıralanması için 4 çeşit ÇKKV yöntemi uygulanmıştır. Sıralama sonuçları arasında güçlü uyumsuzluklar varsa ikinci adıma geçilmiş ve her ÇKKV yönteminin optimal ağırlığı belirlenmiştir. Optimal ağırlıklar, ideal sıralama ile tüm ÇKKV yöntemlerinin sıralaması arasındaki toplam farkı minimumda tutmayı sağlayan ağırlıklardır. Üçüncü adımda 4 ÇKKV yöntemi yeniden uygulanmıştır. Ancak bu kez her ÇKKV yöntemi, ilk adımda elde ettiği sıralama puanları ve kendi ağırlık değeri kullanılarak uygulanmıştır. Böylece sınıflandırma algoritmalarının ikincil sıralamaları elde edilmiştir. Füzyon yaklaşımı kamusal erişime açık 4 veri kümesi üzerinde test edilmiştir. Veriler üzerinde 7 farklı sınıflandırma algoritması esas alınarak gerçekleştirilen deneylerde ilk başta oldukça farklı sıralamalar elde edildiği, ancak önerilen yaklaşımın uygulanması sonucu ikincil sıralamalarda güçlü bir anlaşma sağlandığı ifade edilmiştir.

Chattopadhyay ve Mitra, mikro finans kuruluşlarının performansını analiz etmek için en uygun olan sınıflandırıcıyı TOPSIS yöntemini kullanarak belirlemeyi önermiştir [136]. İlk adımda, mikro finans kuruluşlarının büyüme ve sosyal yardım ikiz hedeflerini yerine getirme yeteneğini analiz etmeye yarayan ızgara temelli yeni bir yaklaşım öne sürülmüştür. Bu yaklaşıma göre mikro finans kuruluşlarının ikiz hedeflere ulaşma bakımından performansları 2×2'lik bir sınıflandırma matrisi üzerinde bulunan 4 kısımdan birine denk düştüğü için performans analizi bir sınıflandırma problemi şeklinde ele alınabilir hâle gelmiştir. Yani mikro finans kuruluşları iki yöndeki performanslarına bağlı olarak 4 bölüm hâlinde sınıflandırılmıştır. İkiz hedefleri karşıladığı için başarılı olduğu düşünülen kuruluşlar birinci bölüme yerleştirilmiştir. Birinci bölüm ile ikiz hedefleri aynı anda karşılayamayan diğer üç bölüm arasında ayrım yapılması amaçlandığı için mevcut durum ikili bir sınıflandırma problemi biçiminde yalınlaştırılmıştır. Izgara temelli performans değerlendirme yaklaşımına göre tanımlanan bu ikili sınıflandırma problemi; mikro finans kuruluşlarının operasyonel ve finansal alanlardaki tahminci özellikleri göz önüne alınarak 6 farklı sınıflandırıcı ile ele alınmıştır. Sınıflandırma sonuçları 9 ayrı performans kriterine göre elde edilmiş olup, bu sonuçlar TOPSIS yöntemi ile değerlendirilerek sınıflandırma yöntemlerinin bir sıralaması elde edilmiştir. Kısacası, mikro finans kuruluşları arasında

performans ayrımı yapmaya en uygun olan sınıflandırıcılar TOPSIS yöntemi kullanılarak belirlenmiştir.

Literatürde ÇKKV ile makine öğrenmesi disiplinlerini bir araya getiren diğer çalışmalara bakıldığında bunların büyük kısmının hibrit yöntem geliştirilmesi yönünde yapılan çalışmalar olduğu görülmektedir. Hibrit yöntemlerde; makine öğrenmesi ve ÇKKV alanlarında önceden beri kullanılan bazı formülasyonlar ele alınır ve geçerli düşünsel dayanakları olan yeni bir öneriye uygun şekilde bunlar birleştirilir. Böylece belirli araştırma hedeflerine yönelik daha verimli yaklaşımların ortaya çıkması sağlanır. Makine öğrenmesi ve ÇKKV temelli hibrit yaklaşımların önerildiği bazı çalışmalarla ilgili aşağıda bilgi verilmiştir.

Kartal ve arkadaşları, makine öğrenmesi algoritmalarını ÇKKV yöntemleriyle bütünleştirerek çok nitelikli envanter analizini etkili bir şekilde yürütmeyi sağlayan hibrit bir metodoloji önermiştir [137]. Envanter kontrolü konusunda temelleşen yaklaşımlardan biri olan ABC analizi, envanter kalemlerinin önemlerine göre A, B ve C etiketleri altında farklı sınıflara ayrılmasına dayanır. Hibrit metodoloji kapsamında; ABC analizi çok kriterli bir sınıflama problemi olarak ele alınmış olup, analiz sonucunda elde edilen sınıf bilgilerinin makine öğrenmesi temelli süreçlerde kullanılması suretiyle envanter sınıflandırma görevine yönelik verimli modeller üretilmesi amaçlanmıştır. Önerilen metodoloji büyük ölçekli bir otomotiv şirketinin 715 farklı stok kalemi içeren endüstriyel envanterinin analizi için kullanılmıştır. İlk olarak stok kalemlerinin ham özellikleri belirlenip, bu ham özelliklerden karar kriterleri türetilmiştir. Bunun ardından, çelişir durumdaki çoklu niteliklere dayanan envanter sınıflama problemi; SAW, AHS ve VIKOR olmak üzere 3 farklı ÇKKV yöntemi ile çözülerek stok kalemlerinin her birinin sınıfı belirlenmiştir. Bu noktadan sonra mevcut durumun, sınıf etiketli veriler üzerine kurulu tahmin modelleri üretmeye uygun olduğu düşüncesi etkili olmuştur. Bu doğrultuda çok nitelikli envanter sınıflandırması bir tahmin problemi olarak ele alınmıştır. Stok kalemlerinin sınıflarını tahmin etmek için mevcut veriler eğitim ve test kümeleri olarak ayrılıp, 5 ayrı makine öğrenmesi algoritması uygulanmıştır. Tüm ÇKKV yöntemleri için her bir algoritmanın tahmin performansı değerlendirilmiştir. Sonuçlar arasında kıyaslamalar yapılarak, ABC envanter analizini en iyi yürüten modellerin hangileri olduğu belirlenmiştir. Değerlendirmeler kapsamında önerilen yaklaşıma vurgu yapılmış ve bu

yaklaşım çerçevesinde makine öğrenmesi algoritmalarının envanter sınıflandırma problemlerine uygulanabilir olduğunun altı çizilmiştir.

Öznitelik seçim probleminin çözümü için kullanılan yöntemlerden bir kısmı sonuç kümesi olarak özniteliklerin iyiden kötüye doğru sıralamasını veren bir liste döndürür. Bu yöntemler değerlendirme prosedürü içinde birbirlerinden farklı değerlendirme kriterleri kullandıkları için her bir yöntemin ürettiği sıralama listesi diğerlerinden farklıdır. Bu durum sıralamada tutarsızlık görülmesine yol açar. Jaya ve Tamilselvi, sözü edilen tutarsızlık problemini ele almak için çoklu değerlendirme kriterlerine dayalı yeni bir öznitelik seçim yöntemi önermiştir [138]. Önerilen yöntem uyarınca; mesafe, bağımlılık ve bilgi kriterlerine göre öznitelikler sıralanmış ve bu 3 farklı değerlendirme kriterine göre elde edilen sıralama sonuçlarının tutarlı bir sıralama altında birleştirilmesi için bulanık AHS yöntemi kullanılmıştır. Öncelikle her değerlendirme kriteri için bireysel öznitelik sıralaması elde edilmiş olup, bu sıralamaya göre ikili karşılaştırma matrisi oluşturulmuştur. Saaty tarafından önerilen önem ölçeği [139] temelinde oluşturulan karşılaştırma matrisleri bulanıklaştırılmış, ardından bulanık AHS ile işlenerek özniteliklerin genel tercih sıralamaları elde edilmiştir. Buna ek olarak sınıflandırma sürecinde kullanılacak olan optimal öznitelikler belirlenmiştir. Yazarlar önerdikleri yöntemin, kredi riski sınıflandırma görevi için en uygun öznitelikleri seçme görevindeki etkinliğini iki ayrı veri kümesi üzerinde test etmiştir. Test sonuçları değerlendirilmiş ve sıralamada tutarsızlığı ortadan kaldırmaya yönelik önerilen yöntemin, kredi riski tahmininde ulaşılan performans değerlerini iyileştirdiği ifade edilmiştir.

Saghapour ve arkadaşları, proteomik veriler üzerinden kanser evrelerinin tahmininde başarıyı artırmak için hibrit bir öznitelik sıralama yöntemi önermiştir [140]. Bu yöntem uyarınca, protein ekspresyon verileri üzerinde uygulanan filtre temelli öznitelik seçim tekniklerinin sonuçları TOPSIS yöntemiyle birleştirilerek en ayırt edici proteinler belirlenmiştir. TOPSIS yöntemi ile gerçekleştirilen öznitelik sıralama prosedüründe iş akışı şu şekildedir: Öncelikle 10 ayrı filtreleme yönteminin her biri için sıralama sonuçları elde edilmiştir. Öznitelikler elde ettikleri sıra değeriyle ters orantılı şekilde puanlandıktan sonra TOPSIS yöntemi uygulanmıştır. TOPSIS yönteminde öznitelik seçim yöntemleri ve protein ifadeleri, sırasıyla kriterler ve alternatifler olarak kabul edilmiştir. Bu çerçevede, sıralama değerine göre proteinlere verilen puanlar, alternatiflerin kriterler altındaki derecelerine karşılık gelmiştir. TOPSIS ile elde edilen nihaî sıralamaya göre öznitelik

seçim işlemi gerçekleştirilmiştir. Böylelikle kanser aşamalarının öngörülmesi için gerekli olan en iyi protein biyobelirteçleri belirlenmiştir. Belirlenen bu proteinler kanser evrelemesi hakkında doğru öngörüde bulunmayı sağlayan tahminci öznitelikleri temsil etmekte olup, sınıflandırma algoritmalarının girdi kümesini oluşturmak üzere kullanılmıştır. Önerilen hibrit yöntem, 7 kanser veri kümesi üzerinde 7 farklı sınıflandırıcı model ile test edilmiştir. Elde edilen sonuçlar rapor edilmiş, böylece önerilen yöntemin diğer yöntemlere kıyasla hem daha yüksek stabilite ve üstün sınıflandırma performansı avantajına sahip olduğu hem de uygulanan sınıflandırıcıya karşı daha az hassas olduğu kanıtlanmıştır.

Nguyen ve Nahavandi, tek değişkenli bir puanlama kriterine göre işlem yapan bireysel gen seçim yöntemlerinin sunduğu sıralama sonuçlarını bütünleştirmek için AHS'nin bir modifikasyonuna dayanan yeni bir gen seçim yöntemi önermiştir [141]. Önerilen modifiye AHS yöntemi; 5 farklı gen sıralama yöntemiyle belirlenen en ayırt edici genleri sistematik bir hiyerarşi ile bir araya getirerek sınıflandırma için en bilgilendirici ve kararlı gen alt kümesini araştıran hibrit bir yöntemdir. Başlangıçta 5 ayrı kritere göre genlerin nicel sıralama puanları elde edilmiştir. Bu nicel değerlere göre genler arasında ikili karşılaştırmalar yapılarak genlerin göreceli önemleri belirlenmiştir. Bu süreçte genler arası mesafe temel alınmıştır. Belirli bir kritere göre iki gen arasındaki mesafeyi hesaplamak için o iki genin verilen kriter altındaki nicel değerlendirme puanlarının mutlak farkı alınmıştır. Hesaplanan mesafe değerleri 1 ila 10 değer aralığı içine ölçeklenmiştir. Buraya kadar yapılan işlemler neticesinde, kriterlerin her biri için gerçek sayılar üzerine kurulu bir ikili karşılaştırma matrisi elde edilmiştir. Bu matrislere AHS yönteminin işlem adımları uygulanarak genlerin sıralama puanlarını temsil eden özvektörler hesaplanmış ve elde edilen 5 özvektör bir araya getirilip performans matrisi oluşturulmuştur. Performans matrisi üzerinden genlerin sıralama sonucu belirlenmiş ve en iyi sıralanan genler sınıflandırma için seçilmiştir. Bir dizi deneysel karşılaştırma yapılarak, genleri değerlendirmede 5 bireysel yöntemin avantajlarını derleyen modifiye AHS yönteminin diğer yöntemlere göre büyük bir performans üstünlüğü ve sağlamlık sergilediği gösterilmiştir [141].

Gen seçimi probleminin ele alındığı bir başka çalışmada ise genler arasından daha iyi sınıflandırma doğruluğu sağlayan en bilgilendirici alt kümenin seçilmesi amacıyla TOPSIS ve F-skor yöntemini bir araya getiren yeni bir yöntem önerilmiştir [142]. Önerilen yöntem

kolon kanseri mikrodizi veri kümesi üzerinde uygulanıp, performansı değerlendirilmiştir. Birleşik gen seçim yönteminde, önce TOPSIS metoduna göre genler sıralanarak üst sıradaki 250 gen seçilmiş, peşinden F-skor yöntemi uygulanmıştır. Toplam gen sayısı 100, 50, 20 ve en son 10 gene kadar küçültülerek veri kümesinin boyutları indirgenmiştir. Elde edilen çıktı kümeleri; NB, karar ağacı, kNN ve DVM olmak üzere dört farklı sınıflandırıcıyla test edilmiş ve en iyi performansın 10 gen içeren gen alt kümesi ile elde edildiği sonucuna ulaşılmıştır.

Birçok sezgisel öznitelik seçim yönteminde; sınıf ilgisi, artıklık, koşullu bağımsızlık gibi öznitelik özellikleri temel alınmış olup, birden fazla öznitelik özelliğinden oluşturulan kombinasyonlar değerlendirme kriteri olarak kullanılmıştır [143]. Bu kriterlerin çoğu, öznitelik özellikleri arasında bir denge kurulması gerektiği için ön argümanlar veya sabit katsayılardan yararlanılarak formüle edilmiştir [143]. Ön argümanlar veya katsayılar öznitelik seçiminde belirli bir değerlendirme sistemi kurmak için kullanılan çoklu öznitelik özelliklerinin göreceli önem ağırlıklarını tam anlamıyla yansıtmadığından dolayı öznitelik seçiminde parametre ayarı yapmayı gerektirmeyen yöntemlere ihtiyaç duyulmuştur [143]. Zhang ve arkadaşları bu konu üzerine yaptıkları bir çalışmada öznitelik seçimini çoklu değerlendirme endekslerine sahip bir etkinlik değerlendirme süreci olarak ele alıp, verimli öznitelikleri etkin bir şekilde aramak için VZA temelli parametrik olmayan bir öznitelik seçim çerçevesi önermiştir [143]. Bu çerçeveye göre özniteliklerin verimlilik puanları elde edilerek sıralamaları belirlenmiştir. Önerilen yöntemin NB, kNN, C4.5 karar ağacı ve DVM sınıflandırıcıları ile gösterdiği performans 12 veri kümesi üzerinde deneysel olarak değerlendirilmiş ve çeşitli öznitelik seçim yöntemleri ile karşılaştırılmıştır. Önerilen süper verimli VZA yönteminin çoğu durumda diğer öznitelik seçim yöntemlerinden daha üstün performans sergilediği deneysel sonuçlar ile ortaya konulmuştur.

Zheng ve Padmanabhan, sınıflandırma görevlerinin yerine getirilmesinde bireysel modellerden daha iyi performans gösteren topluluk modelleri oluşturmak amacıyla VZA temelli yeni bir yaklaşım önermiştir [144]. Bu yaklaşım uyarınca, *k*-sınıflı sınıflandırma görevleri için model birleşimi problemini bir VZA formülasyonuna dönüştüren iki yöntem geliştirilmiştir. Önerilen iki yöntem ve model birleştirmede yaygın olarak kullanılan iki teknik arasında karşılaştırma yapılması amacıyla 20 veri setinde üstünde kapsamlı deneyler gerçekleştirilmiştir. Böylece VZA temelli model birleşiminin çeşitli sınıflandırma

problemleri için diğer yaklaşımlardan daha iyi performans elde ettiği deneysel olarak gösterilmiştir.

Sınıflandırma ve öznitelik seçme makine öğrenmesinin temel konuları arasında yer almakla birlikte bu konuların ele alındığı ÇKKV araştırmaları görece az sayıdadır. Daha önce de ifade edildiği üzere; bu iki disiplini bir araya getiren mevcut çalışmalar genel anlamda algoritma değerlendirme ve hibrit yöntem geliştirme olmak üzere iki ana eksende şekillenmiştir. İlkinde, belirli veri kümeleri üzerinde algoritma değerlendirmede farklı hedef ve kısıtlara yönelik kullanılan çeşitli karar kriterlerinin uyumsuz sonuçlarını tek ve kapsamlı bir sonuçta birleştirmek için ÇKKV yöntemlerinden yararlanılmıştır. İkincisinde ise makine öğrenmesi ve ÇKKV yöntemlerini harmanlayan hibrit yöntemler geliştirilerek çeşitli problemlere yeni çözüm önerileri getirilmiştir. Bu tez çalışması, burada sözü edilen iki araştırma çerçevesiyle bağlantılı iki ana aşamadan oluşmaktadır. Tez çalışması kapsamında yürütülen araştırmanın detayları bir sonraki bölümde ele alınmaktadır.

4. MATERYAL VE METOTLAR

İki aşamalı olarak gerçekleştirilen bu tez çalışmasında gen ifadesi verilerinin analizi için makine öğrenmesi ve ÇKKV disiplinlerinin kolektif kullanımı ele alınmıştır. İlk aşamada, gen ifadesi verilerinin sınıflandırılması için çeşitli tahmin modelleri aday olarak belirlenmiş olup, bu modellerin sınıflandırma görevindeki başarısı araştırılmıştır. Araştırma kapsamında bütünleşik AHS-VIKOR yöntemi kullanılarak tahmin modeli seçim problemi bir ÇKKV problemi biçiminde ele alınmıştır. Böylelikle verilen bir sınıflandırma probleminin çözümünde en etkili olan bir veya birkaç tahmin modeli belirlenmiştir. İkinci aşamada ise hibrit bir öznitelik seçim yaklaşımı önerilmiştir. Önerilen yaklaşımda, öznitelik seçim probleminin çözümü birden çok öznitelik seçim kriteri kullanılmak suretiyle bir VIKOR formülasyonuna dönüştürülmüştür. Araştırma çerçevesinin belirgin olarak ortaya koyulması için bu bölümde sırasıyla; konuya ilişkin veri kaynakları tanıtılmış, bu tez çalışması kapsamında makine öğrenmesi ve ÇKKV disiplinleri altında yer alan yöntemlerden hangilerinin kullanıldığı açıklanmış, tahmin modellerini değerlendirmek için uygulanan araştırma prosedürü tanıtılmış ve son olarak önerilen öznitelik seçim yapısının ayrıntılarına inilmiştir.

4.1. Veri Kümeleri ve Veri Hazırlama Süreci

Bu tez çalışması kapsamında kullanılmak üzere; kolon tümörü ve lenfoma isimleriyle bilinen 2 ayrı gen ifadesi verisi belirlenmiştir. Bunlardan ilki, yani kolon tümörü verisi oligonükleotid dizisi kullanılarak elde edilmiştir. Lenfoma verisi ise cDNA mikrodizi deneyi ile elde edilmiş gen ifadesi değerlerini içerir.

Kolon tümörü verisi [145], normal ve tümörlü doku örnekleri arasındaki ayrımı incelemek üzere kullanılan diyagnostik temelli bir veri kümesidir. Veri kümesi 2000 gen içermekte olup, 62 doku örneğinden elde edilen gen ifadesi değerlerini gösterir [146]. Örneklerin 22 tanesi normal doku, 40 tanesi ise tümörlü doku sınıfına aittir.

Lenfoma verisi [147]; tümörlerin gen ifadesi temelinde moleküler olarak sınıflandırılması göreviyle ilgili olup, lenf kanserine ilişkin 9 alt türün örneklerini içerir. Toplamda 96 örnekten oluşan veri kümesinde 4026 genin ifade değeri ölçümleri mevcuttur. Örneklerin

sınıflara dağılımı 2 ila 46 örnek arasında değişmektedir. Lenfoma veri kümesi lenfomanın 9 alt türüne göre; 46, 2, 2, 10, 6, 6, 9, 4 ve 11 örnekten oluşan alt sınıflara ayrılmıştır. Kolon tümörü ve lenfoma veri kümeleri ile ilgili bilgilerin özeti Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Veri kümeleri

Veri Kümesi	Örnek Sayısı	Gen Sayısı	Sınıf Sayısı
Kolon tümörü	62	2000	2
Lenfoma	96	4026	9

Veri hazırlama sürecinde ihtiyaca göre; eksik verilerin doldurulması, standardizasyon ve ayrıklaştırma işlemlerine başvurulmuştur. Bu süreçte kullanılmak üzere seçilen tüm işlemler basit eğiticisiz veri ön işleme tekniklerdir. Tez çalışması kapsamında gen ifadesi verilerine aşağıda açıklamaları verilen ön işlemler uygulanmıştır.

Ön işlem sürecinin ilk adımında, veri kümesindeki her eksik veri noktası ilgili olduğu özniteliğin diğer veri örneklerindeki mevcut değerlerinin aritmetik ortalaması ile doldurulmuştur. Bunun ardından, standardizasyon adımına geçilmiştir. Standardizasyon işleminde; bir X özniteliği için ortalama ve standart sapma değerleri sırasıyla μ ve σ olmak üzere, n adet veri örneğinin X için aldığı değerler (X_k; k = 1, 2, ..., n), Eş. 4.1'de verilen formüle göre yeni değerlerine (Z_k; k = 1, 2, ..., n) dönüştürülmüştür. Böylece öznitelikler sıfır ortalama ve birim varyans değerlerine ayarlanarak standartlaştırılmıştır.

$$Z_k = \frac{X_k - \mu}{\sigma}, \qquad k = 1, 2, \dots, n \tag{4.1}$$

Son ön işlem adımında sürekli sayısal öznitelik değerleri ayrıklaştırılmıştır. Ayrıklaştırma tahmin sonucunu önemli derecede etkileyeceği için gen ifadesi verilerine uygulanan ayrıklaştırma yönteminde biyolojik verinin içsel doğası dikkate alınmalıdır [55]. Bu nedenle her bir gen için; o genin düşük, normal ve aşırı ifadesine karşılık gelen 3-durumlu bir ayrıklaştırma yöntemi kullanılmıştır. Bu yöntem, bir genin örnek verilerden elde edilen ortalama ve standart sapma değerleri μ ve σ olmak üzere; $\mu-\sigma/2$ değerinden küçük veri noktalarını birinci ayrık durum değerine, $\mu-\sigma/2$ ile $\mu+\sigma/2$ arasında değer alan veri

noktalarını ikinci ayrık durum değerine, $\mu+\sigma/2$ değerinden büyük veri noktalarını ise üçüncü ayrık durum değerine dönüştürmektedir [53].

4.2. Bilgi Kuramı Tabanlı Öznitelik Seçimi

Bilgi kuramı [148], rastlantı değişkenleri arasındaki ilişkileri ölçmeyi ve rastlantı değişkenlerinden bilgi edinme amaçlı yararlanmayı sağlar. Dolayısıyla, öznitelik seçim görevi için teorik bir alt yapı sunmaktadır. Bilgi kuramı tabanlı öznitelik seçiminde, bilgi entropisi kavramı temel alınarak özniteliklerin birbirleriyle ve sınıf etiketiyle aralarındaki ilişkiler ölçülmektedir.

4.2.1. Bilgi kuramı ile ilgili temel kavramlar

Entropi şans değişkenlerinin veya diğer ifade şekliyle rassal değişkenlerin belirsizliğinin bir ölçüsüdür [149]. Entropi hesabında, rassal değişkenin aldığı değerlerin olasılıkları dikkate alınır [4]. Ayrık bir X rassal değişkeninin değer aldığı potansiyel x durumlarındaki olasılık değerleri $p(x)$ olasılık kütle fonksiyonu ile gösterilir. Buna göre $x \in X$ olmak üzere X'in entropisi $H(X)$ Eş. 4.2 ile tanımlanmaktadır.

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{4.2}$$

Entropi tanımında logaritmanın tabanı 2 olarak alınmaktadır. Burada x ayrık kategorilerinin marjinal olasılıkları logaritma işlemiyle ele alınarak X değişkeninin tüm olası durumlarını ifade etmek için gereken bit sayıları bulunur. Bulunan her bit sayısı değeri için ilgili marjinal olasılık değerinin ağırlık olarak kullanılması suretiyle elde edilen ağırlıklı toplam entropiyi ifade eder. Buna göre entropi esasen, bir rastlantı değişkenini tanımlamak için gerekli olan ortalama bilgiyi ölçmektedir [148]. Entropi negatif değer almaz ve sayısal tanım aralığı $0 \leq H(X) \leq 1$ biçiminde verilir. Her değişken durumunun eşit muhtemel hâlde oluşu bilgi entropisinin en yüksek değerinde olması anlamına gelir [149]. Dolayısıyla, bir X değişkenin belirli bir x durumu için olasılığı arttıkça entropisi düşer ve tahmin edilmesi kolaylaşır. Entropinin tanımı (X, Y) şeklinde gösterilen bir ayrık rassal değişken çifti için genişletilerek bileşik entropi tanımlanır [148]. X ve Y değişkenleri

için x ve y durumlarının birlikte oluşma olasılığını ifade eden bileşke dağılım $p(x,y)$ olmak üzere bu iki rassal değişkenin bileşik entropisi $H(X,Y)$ Eş. 4.3 ile formüle edilir.

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(x,y) \tag{4.3}$$

Bir ayrık rassal değişken çifti (X,Y) bileşik şekilde ele alınırken Y için durum değeri bilindiğinde X'in kalan belirsizliğine koşullu entropi denir [150]. Koşullu olasılık kütle fonksiyonu $p(x|y)$ ile gösterilmek üzere koşullu entropi $H(X|Y)$ Eş. 4.4 ile ifade edilir.

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, p(x|y) \tag{4.4}$$

Bilgi entropisini kullanarak iki rastlantı değişkeni X ve Y arasında paylaşılan bilgiyi niceleyen Karşılıklı Bilgi (KB) (Mutual Information - MI), değişkenler arasındaki hem doğrusal hem de doğrusal olmayan bağımlılıkları yakalamaya imkân veren bir ilişki ölçüsüdür [148, 151]. KB aynı zamanda; $I(X;Y) = I(Y;X)$ özdeşliğini sağlayan simetrik bir değer olup, Y'nin X hakkında verdiği bilgiyi ve X'in Y hakkında verdiği bilgiyi gösterir [148]. KB değeri Eş. 4.5'te verilen formüle göre hesaplanır. Bununla birlikte KB ölçüsü, bilgi entropisiyle olan ilgisi açısından Eş. 4.6 ile ifade edilir.

$$I(X;Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\, log\, \frac{p(x,y)}{p(x)p(y)} \tag{4.5}$$

$$I(X;Y) = H(X) - H(X|Y) \tag{4.6}$$

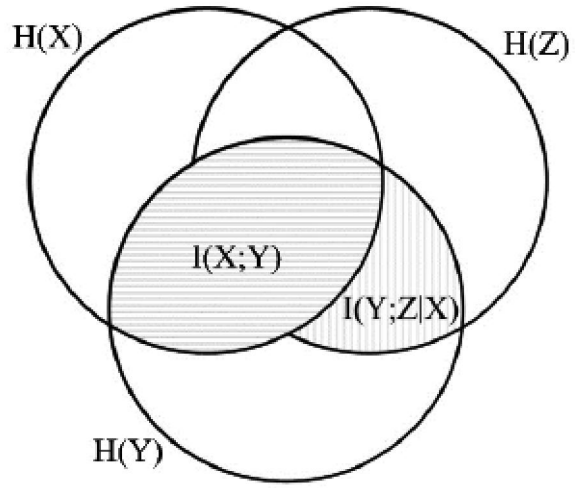
O hâlde KB değeri, Y'nin bilgisine bağlı olarak X'in belirsizliğindeki eksilme olarak ifade edilebilir [148]. İki rastlantı değişkeni arasındaki KB değeri 0 ise bu değişkenler istatistiksel olarak bağımsızdır, KB değerinin büyük olması ise değişkenlerin yakından ilgili olduğunu gösterir [150].

Koşullu Karşılıklı Bilgi (KKB) (Conditional Mutual Information - CMI), koşullu entropi temelli bir bilgi kuramı ölçüsü olup, Z verildiğinde X ve Y değişkenleri tarafından

paylaşılan bilgi miktarı olarak tanımlanır [4]. $I(X;Y|Z)$ şeklinde sembolize edilen KKB ölçüsü Eş. 4.7'de verilen matematiksel ilişki ile ifade edilir.

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{4.7}$$

Bilgi paylaşım ilişkilerinin şematik olarak tarif edilmesi bu ilişkilerin daha kolay anlaşılmasını sağlayabilir. Bilgi kuramı ölçülerine göre 3 rastlantı değişkeni arasındaki bilgi paylaşım ilişkisi Şekil 4.1'de gösterilmiştir.



Şekil 4.1. Rastlantı değişkenleri arasındaki bilgi paylaşım ilişkisi [107]

4.2.2. Karşılıklı bilgi (KB) tabanlı öznitelik seçim yöntemleri

Öznitelik seçimi, sınıf değişkeni ile ilgili en çok bilgiyi temin eden niteliklerin belirlenmesini gerektiren bir optimizasyon problemidir. Bu problemin çözümü için sınıf değişkeni ve öznitelikler arasında KB tabanlı hesaplamalar yapmaya dayanan çeşitli yöntemler önerilmiştir. Bu yöntemlerde, sınıf etiketi ve öznitelikler KB tabanlı formülasyonların değişkenleri olarak işleme alınmaktadır. n öznitelikli bir F öznitelik seti içerisinden C sınıf etiketi hakkında azami bilgi veren $l < n$ öznitelikli alt kümeyi ($S \subset F$) bulma problemi, KB tabanlı öznitelik seçiminde $I(C;S)$ değerini optimize eden S alt kümesini bulma görevine dönüşür [152]. Bu süreçte iki önemli problem ortaya çıkmaktadır. Birincisi; en bilgilendirici alt kümenin keşfi için tüm olası alt kümeleri aramaya çalışmanın hesapsal yükü oldukça fazladır [150, 152]. Bu nedenle arama prosedüründe, aç gözlü arama stratejisi genel tercih hâline gelmiştir. KB tabanlı öznitelik

seçim yöntemlerinde ideal aç gözlü seçim yapısını ifade eden algoritmik bir şablon esas alınır. İdeal aç gözlü seçim, aşağıda verilen temel adımları içerir [150].

- Başlangıçta, n tane aday öznitelik içeren orijinal öznitelik kümesi F ile gösterilir ve henüz boş bir küme olan seçili alt küme S ile ifade edilir.
- Aday özniteliklerin KB değerleri, yani $\forall f \in F$ için $I(C;f)$ değeri hesaplanır.
- Sınıfla ilgisi en yüksek olan, yani $I(C;f)$ değeri maksimum olan f özniteliği orijinal kümeden seçili kümeye aktarılır $F \leftarrow F \setminus \{f\}, S \leftarrow \{f\}$.
- Aç gözlü seçim stratejisinin kullanıldığı bir döngü içerisinde $l < n$ öznitelik elde edilene dek en uygun özniteliklerin seçimine devam edilir. Bu süreçte; seçili öznitelikler göz önünde bulundurularak aday özniteliklerin, $\forall f \in F$ için $I(C;\{f,S\})$ şeklinde ifade edilen birleşik KB değerleri hesaplanır. Her adımda bu ölçüm değerini maksimize eden f özniteliği aday kümeden seçili kümeye aktarılır $F \leftarrow F \setminus \{f\}, S \leftarrow \{f\}$.
- Seçili öznitelik kümesi S sonuç olarak döndürülür.

KB tabanlı öznitelik seçiminde karşılaşılan ikinci önemli problem; $I(C;\{f,S\})$ bilgisini maksimize etmek için gerekli hesaplamaların çokluğu nedeniyle değerlendirme prosedüründe ortaya çıkan zorluktur [150]. Öznitelik sayısı arttıkça öznitelikler arasında paylaşılan bilgi miktarının ölçümü de zorlaşmaktadır. Bu sorunsalın ele alınması için kesin olmayan çeşitli sezgisel formüller türetilir. Bir öznitelik vektörünün bileşen olduğu karşılıklı bilgiyi ölçmek için gerekli hesaplamalar, bireysel öznitelikler arasındaki hesaplamalarla sezgisel olarak ifade edilmeye çalışılır [152]. Yani, çoklu değişkenler arasındaki karşılıklı ilişkileri anlamak için değişkenlerin ikili veya üçlü ilişkilerine bakılır. Böylelikle karmaşık hesaplamaların basit hesaplamalarla ikame edilmesi amaçlanır.

KB tabanlı öznitelik seçiminde, hesaplanabilirlik ile ilgili hususlar dikkate alınarak ideal aç gözlü seçim yapısına göre uyarlanan çeşitli yöntemler geliştirilmiştir [53, 61, 107, 152]. Bu yöntemlerle öznitelikler arasındaki karşılıklı bilgilerin kolaylıkla hesaplanması ve gerçeğine yakın değerlerin elde edilebilmesi hedeflenmiştir. Her yöntem, değerlendirme prosedüründe diğer yöntemlerden farklı bir amaç fonksiyonunun maksimizasyonu ile ilgilenir. Amaç fonksiyonu değişkenler arasındaki ilişkilerin ölçümü için tanımlanmış bir değerlendirme kriteri olup, öznitelik seçimi terminolojisinde sıkça kullanılan ilgi, fazlalık ve etkileşim analizlerini gerçekleştirmeye yönelik terimlerin birleşiminden meydana gelir.

Bu tez çalışmasında, öznitelik seçim probleminin çözümünde kullanılagelen bilgi kuramı kriterleri arasından 10 tanesi ele alınmıştır. Ele alınan tüm kriterler ideal aç gözlü seçim stratejisi ile tanıtılan işlem adımlarını takip etmektedir, bununla birlikte ilk özniteliğin seçimi sonrasında kullandıkları amaç fonksiyonları bakımından farklılık gösterirler. Ayrıca bu 10 yöntem arasından 9 tanesi ilk özniteliğin seçiminde KB ölçüsünü yalın olarak kullanır. Sadece biri KB ölçüsünü kullanmayıp, onun yerine simetrik belirsizlik ölçüsünü kullanmaktadır. Bilgi kuramı kriterleri ile ilgili daha ayrıntılı açıklamalar aşağıda verilmiştir.

Karşılıklı Bilgi Temelli Öznitelik Seçimi (KBÖS) (Mutual Information Based Feature Selection - MIFS) algoritmasında, aday öznitelikler arasından sınıf değişkeni ile ilgisi en yüksek olan aynı zamanda seçili alt küme elemanları ile fazlalık derecesi en düşük olan özniteliğin tespiti hedeflenir [152]. KBÖS yöntemi ile önerilen öznitelik değerlendirme kriterine göre; C sınıf etiketi, f aday öznitelik, s seçili öznitelik kümesi S'in bir elemanı olmak üzere her artırımsal adımda Eş. 4.8'de gösterilen J amaç fonksiyonunu maksimize eden bir öznitelik seçili kümeye eklenir.

$$J = \underbrace{I(C;f)}_{\text{ilgi}} - \beta \underbrace{\sum_{s \in S} I(f;s)}_{\text{fazlalık}} \tag{4.8}$$

Görüldüğü üzere KBÖS yönteminde; $I(C;\{f,S\})$ bilgisi ölçülmek yerine yalnızca $I(C;f)$ ve $I(f;s)$ ölçümleri yapılmıştır [150]. Burada β parametresi fazlalık teriminin göreli önemini belirten ağırlıktır ve değeri genellikle 0,5 ve 1 arasında değişir [152]. Fazlalıklar, sınıf etiketiyle ilgili olabilen ancak seçili küme varlığında dikkate değer ek bilgi sağlamayan özniteliklerdir. Bunun için fazlalıklar öznitelik kümesi içindeki istenmeyen unsurlar olup, değerlendirme sürecinde bir ceza puanıyla seçkinlik derecelerinin düşürülmesi gerekir. Fazlalıkların cezalandırılması için KB tabanlı yöntemlerde, ilgi terimi ile fazlalık terimi arasında genellikle fark alma işlemi gerçekleştirilmektedir.

Sınıf hakkında seçili kümede mevcut olmayan ek bilgiler sağlayan, ancak seçili özniteliklerle ilişkisi yüksek olan bazı aday öznitelikler KBÖS yöntemi tarafından iyi analiz edilemez. Dolayısıyla bu öznitelikler KBÖS algoritmasının performansında düşüşe yol açabilir [150]. Şöyle ki; bir s özniteliği verildiğinde aday f niteliğinin sınıf değişkeni C

ile arasındaki karşılıklı bilginin geriye kalan kısmı, yani $I(C;f|s)$ değeri yüksek olabilir. Aynı zamanda $I(s;f|C)$ değeri de yüksek olabilir. $I(s;f|C)$ bilgisi öznitelik seçimi açısından önem ifade etmemesine rağmen KBÖS yönteminde bu bilgi fazlalık analizine dâhil edilmektedir. Düzgün Bilgi Dağılımı Altında Karşılıklı Bilgi ile Öznitelik Seçimi (D-KBÖS) (Mutual Information Feature Selector under Uniform Information Distribution - MIFS-U) [150]; burada sözü edilen problemi ele alma amacıyla geliştirilen metotlardan biridir. D-KBÖS yönteminde, $I(C;f|s)$ değerini maksimize eden özniteliklerin bulunabilmesi için revize bir formül önerilmiştir. Önerilen formül, seçili özniteliğin entropisini ifade eden bölgede düzgün bilgi dağılımının olduğu varsayımına dayanmaktadır. Yani, $H(s)$ bölgesi içinde aday öznitelik ve sınıf etiketi varlığındaki bilgi paylaşımını temsilen ortaya çıkan alt bölgeler düzgün bilgi dağılımına sahiptir. Bu varsayım doğrultusunda KBÖS ile önerilen amaç fonksiyonu gözden geçirilip, Eş. 4.9'da ifade edilen D-KBÖS yöntemine özgü amaç fonksiyonu ortaya konmuştur.

$$J = I(C;f) - \beta \sum_{s \in S} \frac{I(C;s)}{H(s)} I(f;s) \tag{4.9}$$

KBÖS yönteminde olduğu gibi D-KBÖS kriterinde de ilgi terimi ile fazlalık terimi arasındaki göreceli önemi belirten ön tanımlı bir β değeri kullanılır. Bununla birlikte, formülden de anlaşıldığı üzere fazlalık teriminde her seçili s özniteliği için o özniteliğin sınıfla paylaştığı bilgiye bakılarak $I(C;s)/H(s)$ oranı ile tanımlanan bir anlamlılık değeri belirlenmiş olup, bu değer $I(f;s)$ bilgisini ağırlıklandıran bir faktör olarak kullanılmıştır.

KBÖS ve D-KBÖS ile önerilen amaç fonksiyonlarında ilgi ve fazlalık analizi için kullanılan terimlerin birbirleriyle karşılaştırılamaz oluşu göze çarpmaktadır [97]. Çünkü öznitelik seçim sürecinde artırımlı arama stratejisi izlenmektedir ve giderek büyüyen bir seçili öznitelik kümesi söz konusudur. Bu durumda, fazlalık analizindeki kümülatif toplam işlemi sebebiyle ilgi terimi gitgide önemsizleşir ve seçili alt kümeyle artıksız olan özniteliklerin seçilmesi daha önemli bir amaç hâline gelir [97]. Böyle bir nedenden dolayı ilgisiz özniteliklerin seçilmesine geçit vermemek için Minimum Fazlalık Maksimum İlgi (mFMİ) (Minimum Redundancy Maximum Relevance - mRMR) veya Normalleştirilmiş Karşılıklı Bilgi Öznitelik Seçimi (NKBÖS) (Normalized Mutual Information Feature Selection - NMIFS) yöntemlerinden yararlanılabilir. Bu yöntemlerde fazlalık terimine

kullanıcı tanımlı bir önem ağırlığı atanmaz, bunun yerine toplam fazlalık seçili kümenin o adımdaki eleman sayısına bölünür.

mFMİ yöntemi [80] KBÖS yöntemine benzemekle birlikte aralarındaki ayrım fazlalık parametresinden dolayı ortaya çıkmıştır. mFMİ yönteminde formülasyon, ilgi ve fazlalık terimleri arasında kullanılan operatörün türüne göre ikiye ayrılır. Buna göre formülasyon; fark operatörü kullanılıyorsa Karşılıklı Bilgi Fark kriteri (KBF) (Mutual Information Difference criterion - MID), bölme operatörü kullanılıyorsa Karşılıklı Bilgi Oran kriteri (KBO) (Mutual Information Quotient criterion - MIQ) adını almaktadır [53]. Seçili alt kümenin eleman sayısı $|S|$ ile belirtilmek üzere, KBF kriterinin temsil ettiği amaç fonksiyonu Eş. 4.10'da verilmiştir.

$$J = I(C;f) - \frac{1}{|S|}\sum_{s\in S} I(f;s) \tag{4.10}$$

KBO kriterinde ilgi ve fazlalık terimleri arasındaki oran esas alınmış olup, Eş. 4.11'de verilen amaç fonksiyonunun maksimizasyonu önerilmiştir.

$$J = I(C;f)\Big/\frac{1}{|S|}\sum_{s\in S} I(f;s) \tag{4.11}$$

NKBÖS yönteminde, fazlalık analizi için KB ölçümünün normalleştirilmesine dayanan bir formülasyon kullanılır [97]. Bu yöntemde; aday öznitelik *f* ile seçili öznitelik *s* arasındaki KB ölçümünün Eş. 4.12'de gösterilen sınırlar arasında değer aldığı ve bir özniteliğin entropisi büyük ölçüde değişebildiğinden dolayı KB ölçüsünün global bir öznitelikler kümesine uygulanmadan önce normalizasyona tâbi tutulması gerektiği vurgulanmıştır [97].

$$0 \le I(f,s) \le \min\{H(f), H(s)\} \tag{4.12}$$

Seçili alt küme *S* ile aday nitelik *f* arasındaki fazlalık ölçüsünü temsilen ortalama normalleştirilmiş KB değerinin kullanıldığı NKBÖS yönteminde Eş. 4.13 ile verilen amaç fonksiyonu esas alınmıştır.

$$J = I(C;f) - \frac{1}{|S|}\sum_{s\in S}\frac{I(f;s)}{\min\{H(f),H(s)\}} \tag{4.13}$$

KBÖS, D-KBÖS, mFMİ ve NKBÖS yöntemlerinde dikkat çeken bir husus, hem ilgi hem de fazlalık analizinde KB ölçüsünün esas alınmış olmasıdır. Bu yöntemlerde görülen ortak bir problem; fazlalık oluşturan özniteliklerin cezalandırılması için aday öznitelik ile seçili alt kümedeki öznitelikler arasındaki KB değerine göre, sınıf etiketi dikkate alınmadan hesaplama yapılmasıdır [61]. Sözü edilen yöntemler sınıf değişkeni ile ilgisi olan ve sınıf değişkeni ile ilgisi olmayan fazlalık ölçüleri arasındaki farkı iyi analiz edememektedir [153]. Hâlbuki sınıf bilgisi göz ardı edildiğinde özniteliklerin birbirleri arasındaki bilgi paylaşım miktarının büyük olması bu özniteliklerin fazlalık ilişkisi içinde olduklarına hükmetmek için yeterli değildir. D-KBÖS yönteminde hedefle ilişkili fazlalıkların hesaplanabilmesine yönelik varsayımsal bir çözüm önerisi getirilmiştir. Bunun dışında, KKB ölçüsü kullanılarak formüle edilen değerlendirme kriterleri fazlalık analizinin hedeften bağımsızlaşmasını önleme hususunda yararlı olabilmektedir.

Öznitelik seçimine konu olan değişkenler arasındaki birleşik KB değerini temsil eden $I(C;\{f,S\})$ ifadesi Eş. 4.14'te gösterildiği gibi ayrıştırılabilir. Bu durumda, verilen bir S alt kümesi için $I(C;S)$ ölçümü sabit kalacağından dolayı $I(C;f|S)$ değerinin maksimize edilmesi gerektiği açıkça anlaşılmaktadır [153]. Buna istinaden, KKB ölçüsü öznitelik seçiminde yaygın olarak kullanılmıştır.

$$I(C;\{f,S\}) = I(C;S) + I(C;f|S) \tag{4.14}$$

Koşullu Karşılıklı Bilgi Maksimizasyonu (KKBM) (Conditional Mutual Information Maximization - CMIM) yöntemi; seçili kümedeki bir s özniteliği verildiğinde f ve C değişkenleri arasındaki bilgi miktarının, yani KKB değerinin ölçülmesine dayanır [154]. Eğer f ile C arasında s özniteliğinin içermediği farklı bilgiler paylaşılıyorsa bu durum, aday öznitelik f tarafından hedef sınıfın belirsizliğini azaltacak yeni bilgiler sağlandığı anlamına gelir. KKB değerinin artışı yararlı olup, bu ölçümün maksimize edilmesi istenir. KKBM yönteminde tüm özniteliklerin sınıfla olan KB değerleri hesaplanarak analiz süreci başlatılır. İlk öznitelik seçildikten sonraki adımlarda KKB ölçümleri minimumların maksimumu stratejisi ile hesaba katılır. KKBM yönteminde maksimize edilmesi amaçlanan fonksiyon Eş. 4.15'te formüle edilmiştir.

$$J = \min_{s \in S} I(C; f|s) \tag{4.15}$$

KKBM yönteminin her adımında, aday özniteliklerin sınıfla olan KKB değerleri seçili alt kümenin her bir elemanına göre tek tek hesaplanır. Bunun neticesinde *f* özniteliklerinin her biri için eldeki seçili alt kümenin eleman sayısı adedince KKB değeri elde edilir. Adayın seçili kümeye ne kadar katkı sağladığını ölçmek için bu değerler arasından minimum olanı baz alınır. Yani bir aday öznitelik için hesaplanan KKB değerlerinin en düşüğü, o özniteliğin sağladığı faydalı bilgi miktarını temsil etme amaçlı kullanılır. Özetlemek gerekirse KKBM yönteminde; her bir aday için KKB ölçümleri sonucu elde edilen değerler içinden en düşük olanı dikkate alınarak aday öznitelikler kıyaslanır ve en yüksek değerli öznitelik seçilir.

KKB ölçüsünü etkin bir şekilde kullanan yöntemlerden bir diğeri Koşullu Karşılıklı Bilgi Tabanlı Öznitelik Seçimi (KKBÖS) (Conditional Mutual Information-Based Feature Selection - CMIFS) yöntemidir [153]. KKBÖS yöntemi; özniteliklerin hem sinerji hem de fazlalık ilişkilerinin dikkate alınmasını, bunun yanı sıra fazlalık analizinin hedef sınıfla olan ilgiye göre yapılmasını sağlar.

KKBÖS yönteminde özniteliklerin sinerji ve fazlalık ilişkileri etkileşim bilgisine dayandırılmış olup; formülasyonda etkileşim bilgisi ile KKB arasındaki bağlantıdan yararlanılmıştır. Etkileşim bilgisi [155] bir dizi özniteliğin tümü tarafından paylaşılan ancak bu özniteliklerin herhangi bir alt kümesi üzerinden elde edilmesi mümkün olmayan bilgi miktarı olarak tanımlanmaktadır [61, 153]. Bu bilginin negatif değer alması öznitelikler arasında fazlalık ilişkisinin mevcut olduğunu, pozitif değer alması ise özniteliklerin işbirliği içinde olduğunu gösterir [153]. Öznitelik seçimiyle ilgili değişkenler açısından etkileşim bilgisi $I(C; f; S)$ şeklinde belirtilmek üzere bu bilginin KKB ölçüsüyle arasındaki ilişki Eş. 4.16'da verilen bağıntı ile ifade edilmektedir.

$$I(C; f|S) = I(C; f) + I(C; f; S) \tag{4.16}$$

Burada verilen bağıntı analiz edilerek, etkileşim (fazlalık veya sinerji) ilişkisi içindeki özniteliklerin tespitine yardımcı olan KKBÖS kriteri önerilmiştir. KKBÖS yönteminin ilk adımında hedef sınıfla en ilgili olan öznitelik seçilerek s_1 şeklinde sembolize edilir. İkinci

adımda $I(C;f|s_1)$ ölçüsünü maksimize eden öznitelik seçilir. Daha sonraki adımlarda ise Eş. 4.17 ile verilen amaç fonksiyonu temel alınır ve fonksiyonu maksimize eden f özniteliği seçilir. Bu fonksiyonda s_n her zaman son seçilen özniteliği temsil etmektedir.

$$J = I(C;f|s_1) - I(f;s_n|s_1) + I(f;s_n|C) \tag{4.17}$$

KKBÖS yönteminde ayrıca, değeri 0 ve 1 arasında olan bir eşik değeri (δ) göz önüne alınarak yedekli özniteliklerin tespiti yapılır. Bu doğrultuda Eş. 4.18'de tanımlanan koşullar temel alınmakta ve bu koşulları sağlayan öznitelikler araştırılmaktadır.

$$I(C;f) > 0 \;\; ve \;\; \frac{I(C;f|s_n)}{I(C;f)} \leq \delta \tag{4.18}$$

Bu iki koşulu sağlayan öznitelikler hedef sınıfın belirsizliğinde sadece ihmal edilebilecek bir düzeyde eksilme meydana getiren yedekli özniteliklerdir. Çünkü belirtilen koşullar; sınıf değişkeni ile ilgili olmasına rağmen s_n varlığında ilgililik bakımından önemini yitiren özniteliklerin tespitini sağlamaktadır. Dolayısıyla, ilk adım haricindeki tüm adımlarda öznitelik seçiminden sonra bu iki koşul kontrol edilerek fazlalık oluşturan öznitelikler belirlenir ve bu öznitelikler aday kümesinden elenir.

Aday özniteliklerin değerlendirilmesinde KB ve KKB ölçülerini esas alan yöntemlerden biri de Dinamik Ağırlıklandırma Tabanlı Öznitelik Seçimi (DAÖS) (Dynamic Weighting-Based Feature Selection - DWFS) yöntemidir [4]. DAÖS yönteminde aday özniteliklerin ilgi ve fazlalık düzeyleri analiz edilmekle kalınmamış, aynı zamanda etkileşimli özniteliklerin keşfine yönelik bir öznitelik seçim çerçevesi sunulmuştur. Önerilen çerçevede, aday öznitelik f ile hedef sınıf C arasındaki ilgi düzeyinin hesaplanması için $U(f,C)$ ile sembolize edilen simetrik belirsizlik ölçüsü kullanılmıştır. Simetrik belirsizlik, KB ölçüsünü daima 0 ve 1 arasında normalleştirilmiş bir değere ölçekler ve Eş. 4.19'da verilen formül ile ifade edilir.

$$U(f,C) = 2 \times \frac{I(C;f)}{H(f) + H(C)} \tag{4.19}$$

DAÖS yönteminde aday ile seçili öznitelikler arasındaki fazlalık ve karşılıklı bağımlılıkların hesaplanması için korelasyon oranı adı verilen bir formül altında KB ve KKB ölçüleri bir araya getirilmiştir. Öznitelik değerlendirme süreçlerinde özniteliklerin bir kısmının sınıf etiketiyle ilgisiz olduğuna karar verilir, diğer bir kısmı ise aralarında oldukça yüksek istatistiksel ilişki bulunduğu gerekçesiyle elenir. Karşılıklı bağımlılık; ilgisiz olduğu veya fazlalık oluşturduğu düşünülen ancak aslında gereksiz olmayan, üstelik grup hâlindeyken yüksek ayrımcı güce sahip olan özniteliklerin, yani yararlı içsel öznitelik gruplarının araştırılmasına yönelik kullanılan bir ilişki ölçüsüdür. Korelasyon oranı, karşılıklı bağımlılık ilişkisi içindeki özniteliklerin tespitine ve gerçekten fazlalık oluşturan özniteliklerin ortaya çıkarılmasına yardımcı olur. $CR(f,s)$ şeklinde sembolize edilen korelasyon oranının değeri Eş. 4.20'de verilen formül ile belirlenir ve bu değer −1 ile 1 arasında değişebilir.

$$CR(f,s) = 2 \times \frac{I(C;f|s) - I(C;f)}{H(f) + H(C)} \tag{4.20}$$

Korelasyon oranının negatif olması seçili öznitelik s nedeniyle f ile C arasındaki ilginin düşüş gösterdiğini ifade eder [4]. Pozitif değerler ise seçili özniteliğin etkisiyle f ile C arasındaki ilişkinin artış oranını gösterir [4]. DAÖS yönteminde adayların dinamik olarak ağırlıklandırılmasına dayanan bir şema kullanılmıştır. Bu şemada her adayın ağırlığı son seçilen öznitelikle etkileşimine göre belirlenmektedir. Ağırlıklı özniteliklerin ele alındığı değerlendirme sisteminde Eş. 4.21 ile verilen amaç fonksiyonunun maksimizasyonu istenir. Bu fonksiyonda $w(f)$, aday özniteliğin seçili özniteliklere göre fazlalık ve karşılıklı bağımlılığını özetleyen ağırlıktır, bununla birlikte ilgi düzeyinin göreli önemini belirtir.

$$J = U(f,C) \times w(f) \tag{4.21}$$

Başlangıçta tüm adayların ağırlıkları eşit olup, hepsine 1 değeri atanır. Aday özniteliklerin ağırlıkları, iteratif öznitelik seçim adımları boyunca korelasyon oranına göre sürekli güncellenmektedir. Her adayın güncel ağırlığı, seçili kümeye son eklenen s özniteliği dikkate alınarak Eş. 4.22'de ifade edilen ağırlık hesabıyla elde edilir. Bu durumda f son seçilen öznitelikle karşılıklı bağımlı ise ağırlığı artar artıklı ise ağırlığı düşer.

$$w(f) = w(f) \times \big(1 + CR(f,s)\big) \tag{4.22}$$

Öznitelikler arası bilgilerin kümülatif toplamına dayanan yöntemler aday özniteliklerin önemi hakkında yanılgıya sebebiyet verebilir. Bunun nedeni seçili alt küme elemanlarının yalnızca birkaç tanesiyle yüksek düzeyde ilişkili olan ancak bunların büyük çoğunluğundan bağımsız olan adayların, kümülatif toplam almaya dayanan analizlerde düşük fazlalık miktarına sahipmiş gibi algılanmasıdır [61]. Böyle bir durumdaki aday özniteliğin oluşturduğu fazlalık silik kalacağı için bu öznitelik amaç fonksiyonunda yüksek değere sahip olabilir. Bu sorunsalın ele alınması maksadıyla Birleşik Karşılıklı Bilgi Maksimizasyonu (BKBM) (Joint Mutual Information Maximization - JMIM) yöntemi önerilmiştir [61]. BKBM yönteminin değerlendirme prosedüründe minimumların maksimumu stratejisi kullanılmış olup, $I(C;\{f,s\})$ şeklinde ifade edilen birleşik KB ölçüsünün maksimizasyonu hedeflenmiştir. Bu suretle kümülatif toplama stratejisinin dezavantajı ortadan kaldırılmıştır. BKBM yönteminde önerilen amaç fonksiyonu Eş. 4.23 ile ifade edilmektedir.

$$J = \min_{s \in S}\big(I(C;s) + I(C;f|s)\big) \tag{4.23}$$

KKBM ve BKBM yöntemleri, minimumların maksimumu stratejisine göre düzenlenen amaç fonksiyonları bakımından benzerlik gösterirler [61]. İki yöntem arasındaki fark ise KKBM yönteminde aday özniteliklerin değerlendirilmesi sırasında seçili küme ile hedef sınıf arasındaki sabit ilişkinin göz önüne alınmıyor oluşudur. Oysaki BKBM ile tanıtılan değerlendirme kriteri, $I(C;s)$ ile temsil edilen sabit ilişki miktarını da hesaba katarak işlem yapmayı sağlar. Bu sayede, aday öznitelik ile önceden seçilmiş özniteliklerden en az biri arasındaki birleşik KB miktarının yüksek olması sağlanacağından dolayı seçilen alt kümenin ayrımcılık gücü artırabilmektedir [61].

Bu tez çalışmasında kullanılmakta olan son yöntem Ağırlıklandırılmış Koşullu Karşılıklı Bilgi Öznitelik Seçimi (AKKBÖS) (Weighted Conditional Mutual Information Feature Selection - WCMIFS) yöntemidir [107]. WCMIFS yönteminde aday özniteliklerin sınıf etiketi ile ilgi düzeyi hesaplanırken hem KB hem de KKB ölçümleri hesaba katılmıştır. Buna ek olarak, ilgi ve fazlalık terimleri arasındaki göreli önemi düzenlemek için problem bağımlı bir K katsayısının kullanılması önerilmiştir. Böylece, değişen problemler karşısında dinamik bir değerlendirme sistemi oluşturulmuştur [107]. Sözü edilen K ağırlık parametresi Eş. 4.24 ile tanımlanmaktadır.

$$K = \text{logsig}\left(\frac{\sum_{s\in S} I(f;s) + mF_S}{I(C;f) + MB_S}\right) \tag{4.24}$$

Burada mF_S ve MB_S sırasıyla seçili özniteliklerin fazlalık miktarlarının toplamını ve bu özniteliklerin hedef sınıfla olan toplam ilgisini temsil eder. Aday *f* özniteliğinin fazlalık ve ilgi düzeylerinin de oranlamaya katılımı sağlanıp, elde edilen oran değeri logaritmik sigmoid (s-biçimli) fonksiyon kullanılarak 0 ve 1 arasında normalleştirilir [107]. Böylelikle ilgi ve fazlalık arasında eldeki probleme özgü bir bağlantı kurmayı sağlayan *K* değeri hesaplanmış olur. Amaç fonksiyonunu maksimize eden öznitelikler artırımsal şekilde seçilirken, problemi yorumlayan dinamizmin tesisi için mF_S ve MB_S değerlerinin her seçilen öznitelikle birlikte güncellenmesi gerekir. AKKBÖS yönteminde *K* katsayısı kullanılarak formüle edilen amaç fonksiyonu Eş. 4.25'te ifade edilmektedir.

$$J = I(C;f) + \sum_{s\in S} I(C;f|s) - 2K\sum_{s\in S} I(f;s) \tag{4.25}$$

AKKBÖS yönteminde önerilen amaç fonksiyonunun açıklanması gereken bir yönü fazlalığın iki kez hesaba katılmasıdır. Bunun nedeni ilgi analizinin iki farklı terimin toplamına göre gerçekleştirilmesi olup, buna karşılık ilgi ve fazlalık terimleri arası dengeyi koruma maksadıyla fazlalık terimi 2 ile çarpılmıştır. İlgi analizi amaçlı gerçekleştirilen KKB ölçümünde aday özniteliğin sınıfla paylaşılan bilgisi tüm seçilmiş öznitelik alt kümesine göre alındığı için, burada seçilmiş öznitelik sayısı arttıkça KKB ölçüsünün kümülatif değeri artar, buna bağlı olarak maksimum ilgililiğin temini hususundaki etkisi de büyür [107]. Buna karşın öznitelik seçimi prosedürünün ilk adımlarında KB değerleri öznitelik seçiminde daha etkin olmaktadır [107].

Genel bir değerlendirme yapmak gerekirse; KB tabanlı yöntemler kullanılarak aday özniteliklerin hem hedef sınıfla olan ilgisi hem de seçili özniteliklerle olan fazlalık ve görevdeşlik ilişkileri ölçülebilmektedir. Bu yöntemlerde bir özniteliğin sınıf etiketi ile ilgili oluşu genellikle sınıf değişkeniyle arasındaki KB ölçümüne göre belirlenir ve basittir. Asıl zorluk özniteliklerin fazlalık ve etkileşim ilişkilerinin ölçümünde ortaya çıkmaktadır. Çünkü fazlalık ve etkileşim miktarlarının seçili alt kümedeki özniteliklere göre çok değişkenli bir şekilde yansıtılması gerekir. Sonuç olarak aday özniteliklerin önem derecelerinin doğru bir şekilde hesaplanması için ilgi ve fazlalık analizleri arasında

sezgisel geçiş sağlayan çeşitli amaç fonksiyonlarına ihtiyaç duyulmuştur. Bu fonksiyonlar, en ayrımcı alt kümeyi bulmakla görevli birer kriteri temsil etmektedir. Bu bölümde; KBÖS, D-KBÖS, KBF, KBO, NKBÖS, KKBM, KKBÖS, DAÖS, BKBM ve AKKBÖS olmak üzere KB tabanlı 10 ayrı öznitelik seçim kriteri hakkında bilgi verilmiştir. Bu kriterlerin birbirlerine göre teorik üstünlükleri ve zayıflıklarına ilişkin önemli noktalar vurgulanmıştır. Ancak şunu not etmek gerekir ki; öznitelik seçim kriterlerinin sınıflandırma görevindeki deneysel performansları eldeki veri kümesinin tesiri altında değişim gösterebilir.

4.3. Sınıflandırma Yöntemleri

Sınıflandırma yöntemleri; belirli bir öznitelik kümesi ve sınıfsal çıkış ile tanımlanan veri örneklerinin analizi suretiyle yeni karşılaşılan örneklerin ilgili olduğu sınıfı belirleme işini öğrenen algoritmalardır. Girdi verileri ve sınıf etiketlerinin ilişkilerini modelleyerek yeni sorgu örneklerinin sınıfı hakkında tahminde bulunmayı sağlayan çok sayıda sınıflandırma yöntemine rastlanmaktadır [37-41]. kNN ve NB sınıflandırıcıları bu alanda başı çeken istatistiksel yöntemler arasında yer almakta olup, bu tez çalışmasında kullanılmak üzere seçilmiştir. Bu yöntemler ilerleyen alt başlıklarda ele alınmıştır.

4.3.1. k-En yakın komşu (kNN) algoritması

Sınıflandırma problemlerini ele almak için kullanılan çözüm stratejilerinden biri örnek tabanlı öğrenmedir. Örnek tabanlı öğrenme, yalnızca belirli örneklerin sınıf bilgisini kullanarak sınıfı bilinmeyen örnekler hakkında tahmin üretmeyi sağlayan çözüm yöntemlerinin genel adıdır [156]. k-En yakın komşu veya kısaca kNN algoritması [37] örnek tabanlı öğrenme yöntemlerinden biri olup, sınıflandırılmak istenen örnek ve etiketli veri örnekleri arasındaki mesafe ölçümlerini temel alır.

kNN yönteminde *n* öznitelikli eğitim örnekleri, *n* boyutlu bir örnek uzayında yer alan birer veri noktası olarak temsil edilir ve sınıfı bilinmeyen yeni bir örneğin bu veri noktalarına olan uzaklığı *n* boyutun her biri dikkate alınarak hesaplanır [36]. Burada amaç, test örneğine en yakın olan eğitim örneklerini ortaya çıkarmaktır. Çünkü kNN yönteminde, test örneğinin sınıfının belirlenmesine yönelik olarak yalnızca ona en çok benzeyen birkaç örneğin bilgilendirici olabileceği görüşü önemli yer tutar. En yakın komşuluktaki veri

örneklerinin tespiti için belirli bir mesafe ölçüsüne ihtiyaç duyulur ve genellikle Öklid mesafesi kullanılır. $X = \{x_1, x_2, x_3, ..., x_n\}$ bir eğitim verisinin $Y = \{y_1, y_2, y_3, ..., y_n\}$ ise bir test verisinin öznitelik kümesini belirtsin, eğitim ve test örneklerinin öznitelik değerleri arasındaki $d(x_i, y_i)$, $i = \{1, 2, 3, ..., n\}$ uzaklıklarının toplamına dayanan bir formülasyon kullanılarak Eş. 4.26'da gösterildiği gibi D mesafe değeri hesaplanır.

$$D = \sqrt{\sum_{i=1}^{n} d(x_i, y_i)^2} \tag{4.26}$$

Test örneği ile tüm eğitim örnekleri arasındaki mesafeler araştırılarak ön tanımlı bir k değeri adedince komşu eğitim örneği seçilir. Sonrasında ise bu k adet örneğe göre bir genelleme yapılır. Bu genellemede; tahminci özniteliklerinin aldığı değerler bakımından birbirine benzer olan örneklerin sınıf bakımından da birbirine benzer olacağı düşüncesi esas alınmaktadır. Dolayısıyla belirlenen örneklerin sınıfına bakılarak k en yakın örnek kümesinde en çok örneğine rastlanan sınıfın, test örneğinin sınıfı olduğuna karar verilir.

kNN yönteminin avantajı sadeliği ve kolay anlaşılır oluşudur [3]. Örnek tabanlı öğrenme yöntemlerinde sınıflandırıcı tasarımı yapılmaz, yani genelleme yapmayı sağlamak üzere eğitilmiş herhangi bir modelin varlığı söz konusu değildir. Bir sorgu örneği ortaya çıkana dek eğitim verilerinin atıl durumda kalması, kNN yönteminde tüm işlem yükünün test aşamasında birikmesine yol açar. Sıfır eğitim süresine karşın test örneklerinin sınıfı sorgulanırken gerçekleştirilen hesaplamaların çokluğu nedeniyle süre ve bellek gibi kaynakların kullanımında artış meydana gelir [36, 157]. Eğitim aşamasında bellek gereksiniminin azaltılamaması nedeniyle test aşamasının maliyetli oluşu kNN algoritmasının negatif yönlerinden biridir [157]. Buna ilâve olarak, kNN yönteminin ilgisiz ve gürültülü özniteliklere karşı toleransı düşüktür [156, 158]. kNN yönteminin performansı, örneklerin sınıflara dağılımının dengesiz oluşundan etkilenebilir. Mesafe ölçüsü tercihine ve seçilen k komşu sayısına bağlı olarak kNN performansı değişmektedir [36, 157]. Uygulamada k için 1 değeri yaygın olarak kullanılmıştır [86, 99]. 1-En yakın komşu yönteminde test örneğinin sınıfı ona en yakın olan tek bir eğitim örneğinden öğrenilmektedir. Eğer $k > 1$ ise çoğunluk kararına uyularak sınıflandırma yapılır. Bu durumda, k değeri sınıf sayısının katı olmayan sayılar arasından seçilerek çoğunluk kararında muhtemel bir berabere kalma durumunun önüne geçilebilir.

kNN istatistiksel öğrenmeyi esas alan yöntemlerden biri olmasına karşın sınıflandırma sürecinde analiz edilen veri noktalarının temelindeki olasılık yapısını dikkate almaz [37]. Başka bir şekilde ifade etmek gerekirse kNN yöntemi dağılımdan bağımsızdır ve parametrik olmayan istatistik alanında ele alınır [37]. Çok genel istatistiksel varsayımlar altında, en yakın komşuluğa göre yapılan sınıflandırmanın hatalı sonuç verme olasılığı, optimal Bayes karar politikasının en fazla iki katı hatalı sınıflandırma oranı ile sınırlıdır [37, 156].

4.3.2. Naive Bayes (NB) sınıflandırıcısı

kNN ile NB istatistiksel dayanakları farklı olan sınıflandırma yöntemleridir. NB yönteminde kNN yönteminin aksine öznitelikler ve sınıf değişkeni arasındaki ilişkiler eğitim verileri yardımıyla modellenir ve olasılıksal bir sınıflandırıcı tasarımı gerçekleştirilir. NB sınıflandırıcısı, yeni karşılaşılan örneklerin sınıfını Bayes olasılık teoremi ilkelerine göre tespit eder [38]. Şunu not etmek gerekir ki; Bayes sınıflandırma kuralı ile NB sınıflandırıcısı birbirinin aynısı değildir. NB sınıflandırıcısı, Bayes karar teorisinden türetilen ve uygulanışı daha basit olan bir yöntemi ifade eder [35].

Sınıfı bilinmeyen n öznitelikli bir test örneği, n boyutlu bir $X = (X_1, X_2, X_3, ..., X_n)$ vektörü ile temsil edilmek üzere, Bayes sınıflandırıcıları bu örneğin hangi sınıfa ait olduğu problemini Bayes teoremindeki koşullu olasılık kuralına göre çözmektedir. X ile belirtilen öznitelik deseninin, mevcut C_i, $i = 1, 2, 3, ..., m$ sınıfları arasından belirli birine ait olma olasılığı $P(C_i|X)$ şeklinde sembolize edilir ve Eş. 4.27'de verilen koşullu olasılık formülü kullanılarak hesaplanması gerekir.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{4.27}$$

Bu formülde $P(C_i)$ i. sınıfın önsel olasılığını temsil etmekte olup, C_i sınıfına ait örnek sayısının tüm örnek sayısına oranıdır [52]. $P(X|C_i)$; C_i sınıfında X'in bulunma olasılığını ifade etmektedir. $P(X)$ sınıfı araştırılan test örneğinin önsel olasılığıdır. Her C_i sınıfı için hesaplanan değerler mukayese edilerek olası m tane sınıf arasından hangisinin olasılık payının daha yüksek olduğuna bakılır. $P(X)$ olasılığının, sınıf indeksini gösteren i değişkeninden bağımsız olduğu koşullu olasılık formülünden anlaşılmaktadır. Bu

durumda; kıyaslama sonucunu etkilemeyeceği için $P(X)$, yani payda göz ardı edilir ve X verisine en uygun C_i sınıfını arama problemi Eş. 4.28'de gösterildiği gibi bir maksimum bulma problemine dönüştürülerek ifade edilir.

$$C_{NB} = \arg\max_{C_i}\big(P(X|C_i)P(C_i)\big) \tag{4.28}$$

Burada C_{NB}, olasılık değeri maksimize edilmek istenen hedef sınıfı temsil etmektedir. Formüle göre tüm sınıflar içinde en yüksek olasılık değerine sahip olan sınıfın, test örneğinin sınıfı olarak tahmin edilmesi gerekir. Ancak X, n öznitelikli bir vektör olduğu için öznitelik bağımlılıklarının modellenmesi ve çoklu değişkenlerin sınıf-koşullu olasılıklarının elde edilmesi gerekliliği ortaya çıkmaktadır. Bu gereklilik karmaşık bir süreci beraberinde getirir. Öznitelik bağımlılıkları sınıflandırma kararını zorlaştıran önemli bir etkendir. Boyutluluk arttıkça, $P(X|C_i)$ ile belirtilen sınıf-koşullu olasılıkların en doğru şekilde hesaplanması için çok daha fazla sayıda veri noktasına ihtiyaç duyulur [52].

Koşullu olasılıkların hesaplanmasında ortaya çıkan karmaşıklık tahminci öznitelikler arasında bir bağımlılık olmadığı varsayımı ile azaltılabilir. NB yöntemi pratikte tercih edilen popüler sınıflandırıcılardan biri olup, öznitelikler arasındaki istatistiksel bağımlılıkları yok sayar. Böylece koşullu olasılıkların tahmininde doğruluktan ödün vermek pahasına karmaşık hesaplamaları basitleştirir ve eğitim verisi talebini kontrol altında tutmanın bir yolunu sunar [52]. Bu varsayıma göre NB ile tanımlanan sınıflandırma kuralı, X için en olası sınıfı öngörmeye yönelik maksimum bulma formülünü Eş. 4.29'daki gibi değiştirir.

$$C_{NB} = \arg\max_{C_i}\left(P(C_i)\prod_{j=1}^{n}P\big(X_j|C_i\big)\right) \tag{4.29}$$

Formülde ifade edildiği üzere özniteliklerin sınıflandırma sonucuna olan etkileri sınıf-koşullu olasılıklarının çarpımı yoluyla hesaba katılır. Böylece belirli bir test örneği basit bir hesaplama ile en olası sınıfa atanmış olur.

Sınıflandırma problemlerinde özniteliklerin bağımsız olduğu kabulü genellikle gerçeği yansıtmaz. NB sınıflandırıcısı öznitelik bağımlılıklarını modelleyemiyor olmasına rağmen

pek çok sınıflandırma görevinde sergilediği başarılı performanstan dolayı çokça tercih edilen bir yöntem olagelmiştir [35, 52, 159]. Ayrıca hızlı çalışması, düşük bellek gereksinimi ve kolay uygulanışı nedeniyle ilgi çeken sınıflandırma yöntemleri arasında yer almaktadır [159].

4.4. Tahmin Modellerini Değerlendirmede Kullanılan Metrikler

Sınıflandırma performansı, öznitelik seçim algoritmalarını doğrulamak için en etkili ve tabiî yollardan biridir [4]. Dolaylı olarak şöyle söylenebilir; ölçülen sınıflandırma performansı, tahmin modelinde kullanılan sınıflandırıcının ve öznitelik seçici yöntemin birlikte gösterdiği performansı yansıtır. Buradan hareketle, tez çalışması kapsamında belirli gen ifadesi verileri için birden çok modelin kıyaslandığı değerlendirme süreçlerinde ölçüt olarak sınıflandırma performansı kullanılmıştır.

Bir tahmin modelinin performansı değerlendirilirken, modelin etiketlendirme işlemini doğru yapıp yapmadığına bakılır. Buna göre bir tahminci modelin belirli sayıda sorgu örneğinden oluşan bir test kümesiyle sınanması ve elde edilen sonuçların belli başlı performans metriklerine göre anlamlandırılması gerekir. Tahmin performansının ölçümü için tanımlanmış olan çok sayıda metrik vardır. Her bir metrik model performansının başka bir yönden analizini sağlar ve görece zayıf ve üstün özelliklere sahip olabilir.

Tahmin sürecinde test sonuçları genellikle bir hata matrisi ile ifade edilir ve performans metrikleri temelde bu matrise dayanır. Hata matrisi ikili bir sınıflandırma probleminde sınıflandırıcının elde ettiği tahminlerin, mevcut sınıf etiketleri ile karşılaştırılması sonucu elde edilir ve tahmin sonuçlarının nicel değerlerle ifade edilebilmesi maksadıyla kullanılır. Sınıflandırıcıdan çıkan sonuçların doğru veya hatalı oluşu göz önüne alınarak hata matrisinde 4 ayrı bölge tanımlanmıştır. İkili bir sınıflandırma problemi için P pozitif sınıftaki örnek sayısı ve N negatif sınıftaki örnek sayısı olsun. D ve Y ise test örneğinin asıl etiketi karşısında sınıflandırıcının atadığı tahmini etiketin doğru veya yanlış olma durumunu ifade etmek için kullanılsın. Buna göre test sonuçları; doğru pozitif (DP), doğru negatif (DN), yanlış pozitif (YP) ve yanlış negatif (YN) olmak üzere 4 alt bölüme ayrıştırılabilir [160]. DP ve DN sırasıyla doğru tahmin edilen pozitif ve negatif örneklerin sayısını temsil etmektedir [161]. YP pozitif olarak tahmin edilen negatiflerin sayısıdır. YN ise negatif olarak tahmin edilen pozitif test örneklerinin sayısını temsil eder [161]. Yeni bir

örneğin sınıflandırma sonucu hata matrisi aracılığıyla tanımlanan bu bölgelerden (DP, DN, YP ve YN) birine denk düşer [162]. İkili bir sınıflandırma problemi için hata matrisinin genel gösterimi Çizelge 4.2'de verilmiştir.

Çizelge 4.2. Hata matrisi

		Öngörülen Sınıf		
		Pozitif	Negatif	Toplam
Gerçek Sınıf	Pozitif	DP	YN	P
	Negatif	YP	DN	N

Performans ölçümünde genel anlamda, tahminci model tarafından test örneklerine atanan sınıf etiketlerinin gerçeği ne kadarıyla yansıttığına bakılır. Bu tez çalışmasında, tahmin modellerinin değerlendirilmesinde 5 farklı metrikten yararlanılmıştır. Bu 5 metriğe ilişkin gerekli açıklamalar takip eden alt başlıklarda verilmiştir.

4.4.1. Doğruluk

Doğruluk (accuracy), performans ölçümünde yaygın şekilde kullanılan metriklerden biri olup, test örneklerinin doğru sınıflandırılabilme oranını gösterir [163]. Doğruluk ölçüsü Eş. 4.30'da ifade edilen şekilde formüle edilir.

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \tag{4.30}$$

Sınıflar arasında örnek dağılımının dengesiz olduğu veri kümelerinde örnek çoğunluğuna sahip olan sınıflar yanlı sınıflandırma sonuçları elde edilmesine yol açar [164]. Sınıf dengesizliği ile karşı karşıya kalındığında doğruluk ölçüsü ile yapılan değerlendirmelerin yetersiz kalması önemli bir dezavantajdır [164]. Doğruluk değeri yalnızca sınıflandırıcının genel performansı hakkında bilgi verir, pozitif ve negatif etiketli sınıflar hakkında detaylı bir hata analizi yapmayı sağlamaz [162]. Hataların başka metrikler yardımıyla daha ayrıntılı incelenmesi mümkündür.

4.4.2. F-ölçütü

F-ölçütü iki farklı metriğin ağırlıklı harmonik ortalamasına dayanır [160]. Bunların ilki anma (recall) veya bilinen diğer adıyla duyarlılık (sensitivity) ölçüsüdür. Anma, yalnızca pozitif sınıf için hesaplanan doğruluk değeridir. Gerçekte pozitif etiketli olan örnekler içinde sınıflandırıcı öngörüsünün ne ölçüde başarıya ulaştığını tespit etmek için kullanılır. Buna göre anma değeri, pozitif olduğu tahmin edilebilen pozitif örnek sayısının test kümesindeki mevcut pozitif örnek sayısına bölümü ile elde edilir [161]. Sınıf bazında yapılan bir ölçümdür ve değerlendirmelerde sınıfsal dengesizlikten doğan yanılgıların etkisini azaltmaya yardımcı olur. Anma Eş. 4.31'de verilen formül ile ifade edilmektedir.

$$Anma = \frac{DP}{DP + YN} \tag{4.31}$$

F-ölçütü bünyesinde birleştirilen metriklerin ikincisi ise kesinliktir. Kesinlik (precision) değeri; pozitif etikete sahip olup sınıflandırıcı tarafından da pozitif olarak etiketlenen örnek sayısının, pozitif olarak etiketlendirilen örneklerin toplam sayısına oranlanması ile ele edilir [161]. Yani, tahmin modeli tarafından pozitif sınıfa atanan örnekler içinde gerçekten pozitif sınıfa mensup olanların oranını ölçer. Kesinlik değeri Eş. 4.32'de verilen formül kullanılarak hesaplanır.

$$Kesinlik = \frac{DP}{DP + YP} \tag{4.32}$$

F-ölçütü (F-measure), kesinlik ve anma ölçümlerini bir araya getirip, bütünsel bir ölçüm değeri elde eden metriktir. Böylece iki ayrı ölçüm sonucunun değerlendirme sürecinde eş zamanlı olarak ele alınmasını sağlar. F-ölçütü değeri (F_1) Eş. 4.33'te verilen formüle göre elde edilmektedir.

$$F_1 = \frac{2 \times Kesinlik \times Anma}{Kesinlik + Anma} \tag{4.33}$$

4.4.3. Matthews korelasyon katsayısı (MKK)

Hata matrisi ile tanımlanan DP, DN, YP ve YN değerlerinin hepsi Matthews Korelasyon Katsayısı (MKK) (Matthews Correlation Coefficient - MCC) [165] bünyesinde bir araya getirilip tek bir performans değeri elde edilir [164]. MKK değeri Eş. 4.34'te ifade edildiği şekilde hesaplanmaktadır.

$$MKK = \frac{(DP \times DN) - (YP \times YN)}{\sqrt{(DP + YN)(DP + YP)(DN + YP)(DN + YN)}} \tag{4.34}$$

MKK değeri −1 ila 1 arasında değişmekle birlikte sadece 0'dan büyük değerler anlamlı tahmin sonuçları elde edildiğini ifade eder, zira 0 rastlantıya bağlı tahmin doğruluğuna karşılık gelmektedir [160]. Doğruluk ve F-ölçütü göreli sınıf dağılımına duyarlı metrikler olup, veri dengesizliği durumunda genellikle güvenilir olmayan ölçüm sonuçlarına yol açarlar [164]. MKK ise veri dengesizliğine karşı dayanıklı olan metriklerden biridir [164].

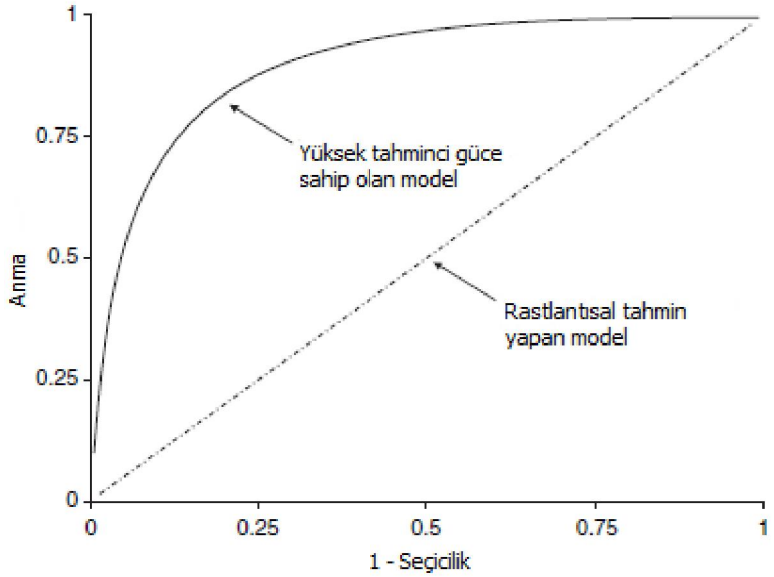
4.4.4. ROC eğrisi altında kalan alan (AUC)

Alıcı İşletim Karakteristiği (Receiver Operating Characteristic - ROC) eğrisi ve Eğri Altında Kalan Alan (Area Under the Curve - AUC) hakkında bilgi verilmeye başlanmadan önce seçiciliğin ne olduğuna ilişkin açıklama yapılması yararlı olacaktır. Seçicilik veya özgüllük (specificity); negatif örneklerin isabetli tahmin edilme oranıdır ve sınıflandırıcının negatif sınıftaki genel etkinliğini ölçmek için kullanılır. Seçicilik değeri, doğru negatif tahminlerin gerçekte negatif olan örneklerin tümüne oranlanması ile elde edilir. Seçicilik Eş. 4.35'te verilen formül ile ifade edilmektedir.

$$Seçicilik = \frac{DN}{DN + YP} \tag{4.35}$$

ROC eğrisi anma ve yanlış pozitif oranı arasındaki dengeyi karakterize etmektedir [166]. Burada yanlış pozitif oranı olarak adlandırılan ölçü, seçiciliği 1'e tamamlayan değeri (1−seçicilik) ifade eder. İkili bir sınıflandırma probleminde performans sonucu, verileri tahmini sınıflara ayırmayı sağlayan karar eşiği seçiminden etkilenir. ROC eğrisi potansiyel eşik değerlerine göre anmayı yanlış pozitif oranın bir fonksiyonu olarak ifade eden

performans eğrisidir [167]. Buna göre bir ROC grafiğinde yatay koordinatlar yanlış pozitif oran değerlerini düşey eksen ise karşılık gelen anma değerlerini ifade etmek için kullanılır. Grafik çizimi sırasında çeşitli karar eşikleri denenerek iki boyutlu bir koordinat düzleminde yanlış pozitif oran ve anma değer çiftlerini temsil eden noktalar işaretlenir ve işaretli noktaların birleşimi ile ROC eğrisi elde edilir [167]. ROC eğrileri grafiksel olarak Şekil 4.2'deki gibi gösterilir.



Şekil 4.2. Varsayımsal ROC eğrileri [168]

ROC grafiklerinde model performansı, eğri altındaki alanın büyüklüğüne bakılarak ifade edilir. AUC bu alanı temsil eden metriktir ve değeri 0 ila 1 arasında değişebilir. Bir ROC grafiğinde rastlantısal tahmin performansına karşılık gelen alt sınır köşegen çizgisi ile temsil edilir [169]. Bu çizginin AUC değeri 0,5'tir. Tahmin gücü bakımından sadece 0,5 üzerinde AUC değerine sahip eğrilerin belirli bir anlamı olabilir. Modelin öngörü gücü arttıkça anma değeri artıp yanlış pozitif oranı azalacağı için ROC eğrisi rastgelelikten uzaklaşarak sol üst köşeye yaklaşır [160, 169]. Dolayısıyla, AUC değeri 1'e daha yakın olan eğriler daha başarılı tahmin sonuçlarını ifade eder.

AUC ölçüsünün en önemli avantajı muhtemel sınıfsal dengesizlik düzeylerini hesaba katmayı sağlamasıdır. AUC bir tahmin modelinin pozitif ve negatif sınıflardan rastgele seçilmiş iki örneği doğru tanıyabilme olasılığı, yani modelin hatalı tahminden kaçınma yeteneği olarak ifade edilebilir [161, 170]. ROC analizlerinde, değerlendirmelerin sınıf

bazında gözlem yapmayı sağlayan iki ayrı ölçümün optimizasyonuna odaklı olması ve belirli bir eşiğe bağlı kalınmaması sebebiyle AUC değeri sınıf dengesizliğine karşı hassas değildir [167]. Dolayısıyla AUC; doğruluk ve f-ölçütünden daha güvenilir değerlendirme sonuçları sunmaktadır [164].

4.4.5. Kesinlik-anma eğrisi altında kalan alan (AUPRC)

Kesinlik-Anma Eğrisi (Precision-Recall Curve - PRC) tahmin modelleri için performans göstergesi olarak kullanılan grafiklerden biridir. PRC grafikleri, değişen karar eşiklerine göre elde edilen kesinlik-anma değer çiftleri kullanılarak çizilir ve kesinlik ile anma arasındaki ödünleşim hakkında bilgi edinmeyi sağlar [166]. Kesinlik-Anma Eğrisi Altında Kalan Alan (Area Under the Precision-Recall Curve - AUPRC) hesaplanarak tahmin performansını gösteren genel bir değer elde edilebilir. PRC grafiklerinde düşey eksendeki $P/(P+N)$ noktasından geçen yatay çizgi taban çizgisi olarak adlandırılır ve rastlantısal tahmin performansına sahip bir sınıflandırıcıyı temsil eder [160]. AUPRC değeri, verinin sınıfsal dengesine göre hareketli olan bu taban çizgisine göre anlam kazanır. Bu nedenle örneklerin sınıflara dağılımı, tahmin sonucunun ne denli başarılı olduğunu belirlemede önemlidir [160]. Eğrinin taban çizgisinden uzaklaşarak yükselmesi AUPRC değerini artırır. PRC eğrisinin sağ üst köşeye yakınlığı, kusursuz bir tahmin modeline yaklaşıldığını ifade eder [171]. Etkinliği yüksek olan tahmin modellerinin AUPRC değeri 1'e daha yakındır. Şunu not etmek gerekir ki; AUC değeri optimum seviyeye yakın bir sınıflandırıcı, kesinlik-anma uzayında optimal olmayabilir [166, 171]. Bir başka önemli nokta da şu ki; AUPRC sınıf dengesizliğine karşı duyarlılığı düşük olan ölçüm metriklerindendir. Üstelik bazı durumlarda çok dengesiz veriler için PRC grafiklerinin ROC eğrilerinden daha uygun olduğu öne sürülmüştür [167, 171].

4.5. Bütünleşik AHS-VIKOR Yöntemi

Birden çok alternatifin çok sayıda karar kriteri dikkate alınarak kıyaslandığı karar süreçlerinde hassas bir uzlaşıya ulaşma gereği, karar vermeyi karmaşık ilişkilerle örülü bir açmaz haline getirmektedir. Bu durumda ÇKKV yöntemleri kullanılarak çoklu karar kriterleri için optimize edilmiş uzlaşmacı çözümlerin elde edilmesi amaçlanır. Bütünleşik AHS-VIKOR; AHS [172] ve VIKOR [173] yöntemlerini birleştirerek uygulayan etkili bir ÇKKV yaklaşımıdır. Bu yaklaşım; kriter ağırlıklarını belirlemek için AHS yönteminin,

uzlaşma sonucunu belirlemek içinse VIKOR yönteminin kullanıldığı bütünleşik bir süreci ifade eder. Bütünleşik AHS-VIKOR yöntemi dört ana aşamayla tasvir edilebilir. Bunlar sırasıyla; alternatiflerin, kriterlerin, kriter ağırlıklarının ve uzlaşık kararın belirlenmesi aşamalarıdır. Buna göre bütünleşik AHS-VIKOR yöntemi ile uygulanan işlem adımları aşağıdaki gibi özetlenebilir:

- Karar sürecinin ilk aşamasında problemin çözümüne yönelik çeşitli seçeneklerin tespiti yapılır ve bunlar içinden karar alternatifleri olarak kullanılmak istenen m tane alternatif (A_1, A_2, A_3, ..., A_m) belirlenir.
- Karar vermeye ilişkin değerlendirme sisteminin hangi kriterler üzerine kurulu olacağı araştırılarak n adet karar kriteri (K_1, K_2, K_3, ..., K_n) belirlenir.
- Her bir kriterin hedefe olan katkısı değerlendirilir ve AHS yöntemi kullanılarak kriter ağırlıkları hesaplanır. Hesap edilen değerler bir w ağırlık vektörü ile ifade edilir.
- Karar sürecinin son aşamasında ise, bir önceki adımda elde edilen karar ağırlıklarından yararlanılmak suretiyle VIKOR yöntemi uygulanır. Böylelikle alternatifler değerlendirilerek uzlaşık sonuçlar elde edilir.

ÇKKV problemlerinde belirli bir çözüm üzerinde uzlaşı sağlanması için kriterler arasında bir denge kurulması gerekir. Çözüm arayışında her kriterin ifade ettiği öneme göre ağırlıklandırılması, karar vermede birden çok kriter arasında kurulması istenen dengenin en faydalı uzlaşma politikası ile elde edilmesine katkı sağlar. Karar verici kendine yöneltilen problemi ele alırken kullanmak istediği kriterlere ağırlık atama konusunda zorlanıyorsa ağırlıkları doğrudan belirlemek yerine inisiyatiflerini karar sürecine yansıtmasına imkân tanıyan sübjektif yöntemlerden birini kullanabilir. AHS yalın hâliyle karar verme süreçlerinde kullanılabilen başlıca yöntemlerden biri olmasına karşın yalnızca kriter ağırlıklandırma amacıyla kullanımı da yaygındır. Bu yöntemde kriter ağırlıkları karar vericinin basit mantıksal vargılarına dayanılarak hesaplanır. AHS ile kriter ağırlıkları kararlaştırılırken uygulanan işlemler aşağıda ayrıntılı şekilde açıklanmaktadır [7, 8, 172].

Öncelikle bir kriterin diğerinden kaç kat önemli olduğunu ifade etmeyi sağlayan sayısal bir ölçek yardımıyla kriterler arasında basit ikili karşılaştırmalar yapılır. Bu suretle kriterlerin birbirlerine olan üstünlükleri sayısal değerlerle ifade edilmiş olur. Karşılaştırmalarda kullanılması önerilen temel ölçek Çizelge 4.3'te gösterilen değerlerden oluşmaktadır. Kriter sayısı n olmak üzere, kriterlerin ikişerli şekilde karşılaştırılması sonucu elde edilen tüm değerler $n \times n$ boyutlu bir kare matrise kaydedilir. Bu matrise ikili karşılaştırma matrisi

denmekte olup, $A[a_{ij}]_{n\times n}$ ile sembolize edilir. İkili karşılaştırma matrisinin hem satır hem de sütunları aynı kriterler kümesini temsil etmektedir. Burada *i*. satır ile *j*. sütundaki kriterlerin ikili karşılaştırma değerini gösteren a_{ij}, *j*. kritere göre *i*. kriterin önemi belirten nispî değerdir. $a_{ij} > 1$ olması *i*. kriterin *j*. kriterden daha önemli olduğunu gösterir. *A* matrisinde, a_{ij}'nin asal köşegene göre simetriği olan eleman a_{ji}, *j*. kriterin *i*. kritere göre ikili karşılaştırma değerini gösterir. Bu değer $a_{ji} = 1/a_{ij}$ bağıntısı kullanılarak elde edilir. Ayrıca *A* matrisinde asal köşegen elemanlarının değeri 1'e eşittir ve $a_{ii} = 1$ şeklinde ifade edilir. Çünkü asal köşegen üzerinde her kriter kendisiyle karşılaştırılma hâlinde bulunur.

Çizelge 4.3. İkili karşılaştırmalarda kullanılan temel ölçek

Göreli Önem Derecesi	Tanım
1	Eşit önem
3	Orta önem
5	Güçlü önem
7	Çok güçlü önem
9	Mutlak önem
2, 4, 6, 8	Ara değerler

Kriter ağırlıkları karşılaştırma matrisi üzerinde gerçekleştirilen matematiksel işlemlerle belirlenmektedir. Öncelikle $A[a_{ij}]$ karşılaştırma matrisinin normalize edilmesi istenir. Bunun için her bir matris elemanı Eş. 4.36'da ifade edildiği gibi ait olduğu sütundaki tüm elemanların değerleri toplamına bölünür. Böylece her elemanın a_norm_{ij}, $i = 1, 2, 3, ..., n$; $j = 1, 2, 3, ..., n$ ile ifade edilen normalleştirilmiş karşılığı hesaplanır.

$$a\_norm_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}} \tag{4.36}$$

İkili karşılaştırma matrisinin normalleştirilmesiyle elde edilen matrisin satır ortalamaları alınarak kriter ağırlıklarını gösteren w_i, $i = 1, 2, 3, ..., n$ değerleri bulunur. Ağırlıkların hesaplanmasında Eş. 4.37'de verilen formül kullanılmaktadır.

$$w_i = \sum_{j=1}^{n} \frac{a\_norm_{ij}}{n} \tag{4.37}$$

Hesaplamalar sonucunda tüm kriterlere karşılık gelen bir $W[w_i]_{n\times 1}$ ağırlık vektörü elde edilmiş olur. Son adımda ise hesaplanan ağırlıklara göre karşılaştırma matrisinin tutarlı olup olmadığına bakılır. Bunun için ilk önce karşılaştırma matrisine karşılık gelen skaler özdeğerin hesaplanması gerekir. Bu değer λ_{max} ile belirtilmektedir. Tutarlılığın doğrulanmasına yönelik ilk olarak, A karşılaştırma matrisi ile W ağırlık vektörü arasında matris çarpımı yapılmak suretiyle $D[d_i]_{n\times 1}$ sütun vektörü elde edilir. Bu çarpım işlemi Eş. 4.38 ile ifade edilmektedir.

$$D[d_i]_{n\times 1} = A\left[a_{ij}\right]_{n\times n} \times W[w_i]_{n\times 1} \tag{4.38}$$

Ardından, D vektörünün elemanları W ağırlık vektörünün eş indisli elemanlarına bölünür ve elde edilen bölüm değerlerinin aritmetik ortalaması alınarak λ_{max} belirlenir. Bu işlemlerin uygulanması sırasında Eş. 4.39'da verilen formül kullanılmaktadır.

$$\lambda_{max} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i}{w_i} \tag{4.39}$$

Özdeğer hesabından sonra, Eş. 4.40'ta verilen formüle göre tutarlılık indeksi (Consistency Index - CI) hesaplanır. Elde edilen CI değeri Eş. 4.41'de gösterildiği gibi rassal indeks (Random Consistency Index - RCI) değerine bölünür. Böylece tutarlılık oranı (Consistency Ratio - CR) hesaplanmış olur. Burada RCI, ön tanımlı bir değerler dizisini belirtir. Kriter sayısına göre tanımlı standart RCI değerleri Çizelge 4.4'te gösterilmektedir.

$$CI = (\lambda_{max} - n)/(n-1) \tag{4.40}$$

$$CR = CI/RCI \tag{4.41}$$

Tutarlılığın doğrulanması için CR değerinin 0,1 olarak belirlenen eşikten küçük olması gereklidir. CR değeri eşikten küçükse tutarsızlık tolere edilebilir düzeydedir. CR değeri eşikten büyükse ikili karşılaştırmalarda belirgin düzeyde tutarsızlık olduğu anlaşılır. Bu

durumda ikili karşılaştırma matrisinde gereken düzenlemeler yapılmalı ve ağırlıklar yeniden hesaplanmalıdır.

Çizelge 4.4. Rassal indeks değerleri

Kriter sayısı	2	3	4	5	6	7	8	9
RCI değeri	0	0,58	0,9	1,12	1,24	1,32	1,41	1,45

AHS yöntemiyle kriter ağırlıkları hesaplandıktan sonra bütünleşik süreç VIKOR yöntemi ile devam ettirilir. ÇKKV problemlerinin çözümünde yaygın şekilde kullanılan VIKOR yöntemi alternatifleri ideal çözüme yakınlığına göre sıralamayı ve en uygun alternatifleri seçerek uzlaşık çözüm önerileri üretmeyi sağlar. VIKOR ile önerilen uzlaşık çözüm maksimum grup faydasını ve karşıt görüştekilerin minimum bireysel pişmanlığını ifade eder [174]. VIKOR ile uzlaşık çözümlerin elde edilme süreci ve bu süreçte izlenen işlem prosedürü aşağıda ayrıntılı şekilde açıklanmaktadır [114, 116, 174].

VIKOR yönteminde ilk olarak, kriterlere göre alternatiflerin durumu değerlendirilir ve tüm alternatiflerin her bir kriter karşısında belirlenen değerleri bir karar matrisine yerleştirilir. Bir A_j alternatifi için K_i kriteri altında elde edilen değer f_{ij} ile ifade edilmek üzere karar matrisi tüm f_{ij}, $i = 1, 2, 3, ..., n$; $j = 1, 2, 3, ..., m$ değerlerini belirten matristir. Karar matrisi üzerinde kriterler bazında araştırı yapılarak her bir kriter altındaki mevcut değerlerin en iyi ve en kötü olanları tespit edilir. Bu işlem adımı Eş. 4.42'de formüle edilmektedir.

$$f_i^* = \max_j f_{ij}, \quad f_i^- = \min_j f_{ij} \tag{4.42}$$

Burada i. kritere ilişkin en iyi ve en kötü değerler sırasıyla f_i^* ve f_i^-, $i = 1, 2, 3, ..., n$ sembolleri ile temsil edilmektedir. Bununla birlikte kriter fonksiyonu fayda değil de maliyet belirtiyorsa hesaplamalar Eş. 4.43'te ifade edildiği şekilde gerçekleştirilir.

$$f_i^* = \min_j f_{ij}, \quad f_i^- = \max_j f_{ij} \tag{4.43}$$

Karar alternatifleri ile kriter fonksiyonlarının en iyi değerleri arasındaki uzaklıklar S_j ve R_j, $j = 1, 2, 3, ..., m$ sembolleri ile belirtilen iki ayrı ölçüye göre hesaplanır. S_j değeri j.

alternatif için; tüm kriterler ile ilintili oldukları en iyi değerler arası uzaklıkların ağırlıklı toplamını göstermekte olup, Eş. 4.44'te verilen formüle göre belirlenir.

$$S_j = \sum_{i=1}^{n} w_i\left(f_i^* - f_{ij}\right)/\left(f_i^* - f_i^-\right) \tag{4.44}$$

Burada w_i, i. kriterin önem ağırlığını belirtir. Bütünleşik AHS-VIKOR çerçevesinde w_i, AHS ile elde edilen kriter ağırlıklarını temsil etmektedir. R_j değeri j. alternatif için bireysel mesafelerin üst sınırını, yani kriterler ile ilişkili oldukları ideal değerler arasındaki mesafelerden maksimum olanı temsil eder. R_j, j = 1, 2, 3, ..., m puanının hesaplanmasında Eş. 4.45 ile ifade edilen formül kullanılır.

$$R_j = \max_i\left[w_i\left(f_i^* - f_{ij}\right)/\left(f_i^* - f_i^-\right)\right] \tag{4.45}$$

Tüm alternatifler için S ve R puanları elde edildikten sonra bunlar içindeki en büyük ve en küçük değerler belirlenir. Buna göre Eş. 4.46'da verilen hesaplamalar yardımıyla S^*, S^-, R^*, R^- değerleri elde edildikten sonra Eş. 4.47'de ifade edilen formül kullanılarak her bir alternatifin Q_j, j = 1, 2, 3, ..., m puanı hesaplanır.

$$S^* = \min_j S_j\,,\quad S^- = \max_j S_j,\quad R^* = \min_j R_j\,,\quad R^- = \max_j R_j \tag{4.46}$$

$$Q_j = \frac{v\left(S_j - S^*\right)}{\left(S^- - S^*\right)} + \frac{(1-v)\left(R_j - R^*\right)}{\left(R^- - R^*\right)} \tag{4.47}$$

Formülde de görüldüğü üzere Q ölçüsü S^* ve R^* stratejileri ile elde edilen iki ölçüye ayrıştırılabilir. Burada R^* ile temsil edilen strateji bireysel kriter uzaklıklarının en kötü üst sınırıyla ilgili iken S^* ile temsil edilen strateji her bir karar kriterine göre hesaplanan uzaklıkların kümülatif toplamına dayanır. R^* stratejisi ile belirlenen değerin düşük olması karşıt görüşlerin minimum bireysel pişmanlığını ifade eder. Bununla birlikte, tüm kriterlerin toplam etkisine göre çözüme gidilmesi kriterlerin çoğunluk kararını yansıtır. Dolayısıyla S^* stratejisi maksimum fayda kuralını ifade eder. Bu iki strateji Q indeksi içinde bir araya getirilmiştir. Q indeksi, karar vermede her iki stratejinin etkisini dikkate almayı sağlar. Burada v maksimum grup faydasını sağlayan stratejinin ağırlığı olup, değeri

0 ila 1 arasında değişebilir. $1-v$ ise bireysel pişmanlığın önem ağırlığını belirtir. $v > 0{,}5$ ise karar çoğunluk kuralına doğru eğilim gösterir. $v = 0{,}5$ olduğunda karar maksimum grup faydası ve rakibin asgari bireysel pişmanlığı arasındaki görüş birliğini gösterir. Bu ağırlık için genellikle iki stratejiyi dengeleyen $v = 0{,}5$ değeri kullanılır.

Uzlaşık çözüm arayışında, her alternatifin ideal alternatife yakınlığı değerlendirilir. Buna göre asıl istenen; *j.* alternatifin tercih puanını gösteren Q_j değerinin küçük olmasıdır. S_j ve R_j daha büyük değerler aldıkça *j.* alternatifin ideal çözüme uzaklığı artar, tercih sırası ise düşer. En iyisi her iki ölçümün de minimum olmasıdır. Karşılaştırmalarda Q puanı öne çıkarılmakla birlikte S, R ve Q indekslerinin hepsinden yararlanılmaktadır. Bu nedenle S, R ve Q indekslerine göre sıralama yapılıp, iyiden kötüye doğru sıralı durumdaki alternatifleri tutan üç ayrı liste elde edilir. Q listesi uzlaşık sıralamayı göstermekte olup, Q(minimum) ile en iyi sıralanan alternatif a' aşağıda açıklanan iki koşul karşılandığı takdirde uzlaşık çözüm olarak önerilir.

İlk koşul *kabul edilebilir avantaj* koşulu olarak adlandırılır ve Q değerine göre en iyi sıralanan iki alternatif arasında önemli düzeyde fark olup olmadığını kontrol etmeyi sağlar. En yüksek sıralı alternatif a' ile ikinci en yüksek sıralı alternatif a'' arasındaki Q değeri farkı DQ eşiğine göre denetlenir. Kabul edilebilir avantaj koşulu Eş. 4.48 ile ifade edilmektedir.

$$Q(a'') - Q(a') \geq DQ \tag{4.48}$$

Burada m alternatif sayısı olmak üzere, DQ ile belirtilen eşik değerinin matematiksel ifadesi Eş. 4.49'da verilmektedir.

$$DQ = \begin{cases} 0{,}25, & m \leq 4 \\ 1/(m-1), & m > 4 \end{cases} \tag{4.49}$$

İkinci koşul *karar vermede kabul edilebilir istikrar* koşulu olup, Q listesinin ilk sırasındaki a' alternatifinin, S ve R değerlerine göre elde edilen diğer iki listenin en az birinde daha en iyi alternatif olarak sıralanmış olmasını gerektirir. Bu iki koşuldan biri sağlanmıyorsa çözüm önerisinde birden fazla alternatife yer verilir. Sadece ikinci koşulun sağlanmaması durumunda, karar vermede istikrar yoktur. Bu durumda a' ve a'' alternatifleri uzlaşık

çözümler olarak önerilir. İlk koşulun kabul edilmediği durumlarda ise uzlaşık çözümler; a', a'', ..., $a^{(M)}$ olmak üzere M tane alternatiften oluşur. Birden çok alternatifin uzlaşık çözüm önerisine dâhil edilmesinin gerekçesi; Q listesine göre en iyi sıralı ilk M alternatif arasında belirgin düzeyde üstünlük farkı oluşmamasıdır. Burada $a^{(M)}$, maksimum M için Eş. 4.50'de ifade edilen matematiksel ilişki ile belirlenmektedir.

$$Q\left(a^{(M)}\right) - Q(a') < DQ \tag{4.50}$$

Uzlaşık çözümün belirlenmesiyle birlikte bütünleşik AHS-VIKOR ile karar verme süreci tamamlanmış olur. Son olarak şunu belirtmek gerekir; VIKOR yöntemi ile bütünleşik AHS-VIKOR arasındaki tek ayrım kriter ağırlıklarının belirlenmesinde ortaya çıkmaktadır. VIKOR yöntemiyle karar verme sürecinde kriter ağırlıklarına herhangi bir şekilde değer ataması yapılmamışsa, bu durumda toplamları 1 olacak şekilde her kriter için eşit ağırlık kullanılabilir. Yani kriter sayısı n göz önüne alınmak suretiyle tüm kriterlerin ağırlığı $w_i = 1/n$, $i = 1, 2, 3, ..., n$ olarak belirlenir.

4.6. Araştırma Metodolojisi

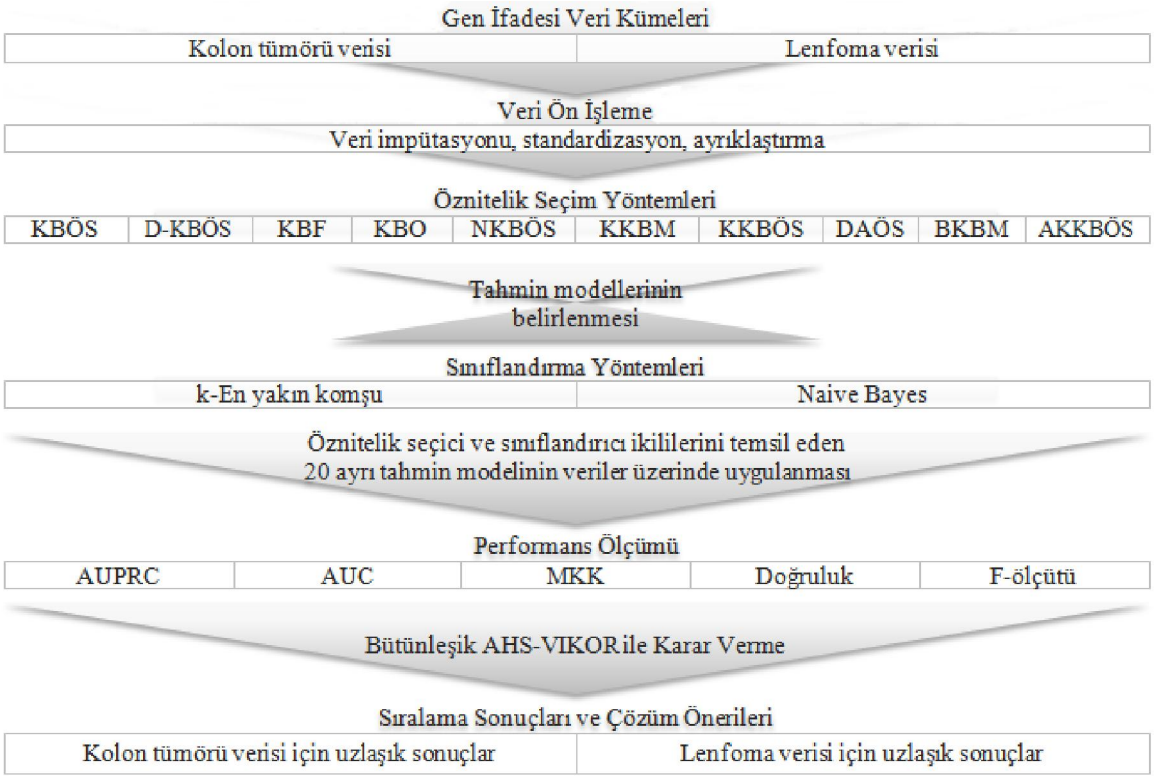
Bu tez çalışmasının temel analiz hedefleri genelden özele doğru şu şekilde sıralanabilir:

- Belirli öznitelik seçim yöntemleri ve sınıflandırıcıların ikili kombinasyonlarını temsil eden çeşitli tahmin modellerini kullanarak gen ifadesi verileri üzerinde sınıflandırma görevlerini gerçekleştirmek,
- Sınıflandırma problemlerinin çözümünde başarıyı artıran tahmin modellerini çoklu kriterlerin göz önüne alındığı bir değerlendirme süreci ile belirlemek,
- Gen ifadesi verilerinin sınıflandırılmasında daha iyi performans sonuçları elde etme amacıyla klasik öznitelik seçim yöntemleri ile ÇKKV yaklaşımını bütünleştiren hibrit bir öznitelik seçim yapısı geliştirmek.

Bu tezde yukarıda sözü edilen hedefler çerçevesinde iki aşamalı bir çalışma yürütülmüştür. Takip eden alt başlıklarda bu aşamaların uygulanma biçimi hakkında ayrıntılı bilgiye yer verilmektedir.

4.6.1. Tahmin modellerinin bütünleşik AHS-VIKOR yöntemi ile değerlendirilmesi

Analiz hedefleri doğrultusunda iki ana aşamaya ayrılan tez çalışmasının birinci aşamasında gen ifadesi verilerinin sınıflandırılması için belirli tahmin modelleri kullanılmış ve ÇKKV görevine dönüştürülen bir değerlendirme süreci uygulanarak bu modellerin başarısı karşılaştırılmıştır. İlk aşama kapsamında takip edilen araştırma prosedürü Şekil 4.3'te ifade edilmektedir.



Şekil 4.3. Tahmin modellerini değerlendirmeye yönelik uygulanan araştırma prosedürü

Araştırma plânı uyarınca; kolon ve lenfoma verilerini analiz etmeye yönelik oluşturulan çeşitli tahmin modelleri adım adım ilerleyen sistemli bir değerlendirme süreci içinde ele alınmıştır. Başlangıçta verilerin bazı ön işlemlerden geçirilmesi gerekebilir. Ön işlem adımında ihtiyaca göre; eksik verilerin doldurulması, standardizasyon ve ayrıklaştırma işlemlerine başvurulmuştur. Ardından, ÇKKV yaklaşımının gerektirdiği şekilde problemin modelini kurmakla işe başlanmıştır. ÇKKV problemlerinde modeller, istenilen hedeflere bağlı kalınarak, bu hedeflere ulaşma ölçütü olan çoklu kriterler altında alternatiflerin değerlendirildiği bir süreci ifade etmektedir [114]. Belirli bir veri kümesi için tahmin modellerinin başarı sırasının belirlenmesi ve en başarılı olanların tespit edilmesi amacıyla

gerçekleştirilen modelleme sürecinde alternatifleri temsil etmek üzere 20 tahmin modeli kullanılmıştır. Bu süreçte tahminci modelleri değerlendirmeye yönelik kullanılan 5 çeşit metrikten her biri birer karar kriteri rolü üstlenmiştir. Böylece amaç, çözüm alternatifleri ve kriterler belirlenip, değerlendirme süreci bir ÇKKV problemi olarak modellenmiştir. Tahmin modellerinin değerlendirilmesine yönelik ortaya koyulan bu ÇKKV probleminin çözümü içinse bütünleşik AHS-VIKOR yöntemi uygulanmıştır. Model kurulum süreci ve karar elemanlarının belirlenmesine ilişkin adımlar aşağıda daha ayrıntılı şekilde açıklanmaktadır.

Gen ifadesi verilerinin analizi sürecinde, makine öğrenmesi temelli tahmin modelleri yaygın olarak kullanılmaktadır. Gen ifadesi verileri boyutsallıktan muzdarip olduğundan dolayı tahmin modellerinin yapısında, öznitelik seçim prosedürünün yer alması bilhassa önemlidir. Önerilen araştırma plânına göre tahmin modellerinin kapsam ve çerçevesi, öznitelik seçimi ve sınıflandırma adımları ile belirlenmiştir. Bu iki önemli analiz adımında kullanılmak üzere seçilen yöntemler tahmin performansını doğrudan etkilemektedir. Bu tez çalışmasında gen ifadesi verilerinin analizi için 2 sınıflandırma yöntemi (kNN ile NB) ve 10 öznitelik seçim yöntemi (KBÖS, D-KBÖS, KBF, KBO, NKBÖS, KKBM, KKBÖS, DAÖS, BKBM ve AKKBÖS) esas alınmıştır. Sınıflandırıcılar ile öznitelik seçim yapıları ikili kombinasyonlar hâlinde bir araya getirilip, birbirinden farklı 20 tahmin modeli oluşturulmuştur. Yani her bir model, gen ifadesi verilerini sınıflandırmak amacıyla bir arada kullanılan öznitelik seçim yöntemi ve sınıflandırıcı ikililerinden farklı birini temsil etmektedir. ÇKKV sürecinde ise bu modeller karar alternatifleri olarak kullanılmıştır.

Her bir gen ifadesi verisi için, mevcut tüm modellerin başarı düzeyleri ölçülerek değerlendirme süreci başlatılmaktadır. Bu süreçte modellerin hata ve başarı oranlarının farklı açılardan değerlendirilmesi için; AUPRC, AUC, MKK, doğruluk ve f-ölçütü olmak üzere 5 farklı metrikten yararlanılmıştır. Bu performans metrikleri ÇKKV sürecinde karar kriterleri olarak temsil edilmiştir. Tahminci modellerin performansına etki eden önemli bir unsur öznitelik seçimi sırasında kullanılan durma kriteridir. N veri kümesinin örnek sayısı, a ise 2 ila 10 arasında değer alan bir değişken olmak üzere öznitelik seçiminde genellikle N/a sayıda öznitelik kullanılarak optimum tahmin performansı elde edilebilmektedir [52]. Bu bilgi ışığında; a değeri 2 alındığında, 62 örnekli kolon tümörü verisi için yaklaşık 31 özniteliğin ($62/2 = 31$), 96 örneğe sahip lenfoma verisi içinse yaklaşık 48 özniteliğin ($96/2 = 48$) tahmin performansını maksimize etmede yeterli olduğu söylenebilir. Ancak

deneysel süreçte istisnaî durumların da ortaya çıkabileceği düşünülmüş olup, daha kesin sonuçlara ulaşılabilme maksadıyla bu öznitelik sayıları birkaç kat daha büyük alınmıştır. Kolon tümörü verisiyle ilgili deneysel süreçte tahmin modellerinin tümünde öznitelik seçim prosedürü için seçili öznitelik sayısının 100'e ulaşması durma kriteri olarak belirlenmiştir. Lenfoma verisi içinse öznitelik seçiminde durma kriteri 200 öznitelik olarak tanımlanmıştır. Ayrıca her bir modelin performansı en iyi AUC değerine ulaştığı duruma endeksli şekilde belirlenmiştir. Yani performans değerleri elde edilirken, AUC kriterini maksimize eden seçili öznitelik sayısı baz alınmıştır. AUC ölçüsü veri dengesizliğine karşı dayanıklı olmakla beraber performans ölçümünde de sık kullanılır. Burada AUC kriterinin seçilme nedeni, özellikle lenfoma veri kümesinde veri dağılımının dengesiz olmasıdır. Tüm ölçümler 10-kat çapraz doğrulama tekniğinden yararlanılmak suretiyle gerçekleştirilmiş ve tahmin modellerin performans değerleri 5 ayrı metriğe göre elde edilmiştir.

Mevcut tahmin modelleri için farklı metrikler altında elde edilen sıralama sonuçlarının birbirine aykırı olması değerlendirme sürecinin bir ÇKKV problemi olarak modellenmesinin temel gerekçesidir. Modeller arasında gerçekleştirilen sıralama ve seçim süreçlerinde farklı karar kriterlerinin ortaya koyduğu uyumsuz ve çelişkili kararların uzlaştırılması maksadıyla bütünleşik AHS-VIKOR yöntemine başvurulmuştur.

Bütünleşik AHS-VIKOR, iki ayrı ÇKKV yöntemi olan AHS ve VIKOR yöntemlerinin birleştirilerek uygulanmasına dayanır. Karar sürecinde avantaj ve dezavantajlarına göre her bir kritere ağırlık atamak ve bu ağırlıklar eşliğinde çözüme gitmek tercih edilen bir yaklaşımdır. AHS kendi iç yapısında sunduğu analiz adımları sayesinde kriter ağırlıklarını hesaplamaya yardımcı olmaktadır. Bütünleşik AHS-VIKOR yaklaşımında AHS, karar vericiden alınan değerlendirmeler doğrultusunda kriterlerin nitel önem derecelerinin karşılaştırılmasına izin vererek öznel bir ağırlıklandırma alt yapısı sağlar. VIKOR ise her bir alternatifin ağırlıklı kriterler altındaki değerlendirme sonuçlarını birleştirip, bütünsel bir tercih puanı elde etmeyi sağlamaktadır. Bu çerçevede; karar vericinin 5 performans kriteri için belirlediği göreli önem dereceleri ve bu 5 kriter altında 20 tahmin modelinin değerlendirilmesi sonucu elde edilen performans bilgileri bütünleşik AHS-VIKOR yöntemi uyarınca işleme alınarak, tahmin modellerinin gen ifadesi verileri üzerindeki başarımları mukayese edilmiştir.

Birinci aşama sonunda, bütünleşik AHS-VIKOR yöntemiyle gerçekleştirilen analizler tamamlanarak her bir veri kümesi için tahmin modellerinin uzlaşık sıralaması elde edilmiş ve en iyi performans gösteren tahmin modelleri uzlaşık çözümler olarak belirlenmiştir.

4.6.2. Bilgi kuramı kriterlerine dayanan VIKOR ile hibrit çok kriterli öznitelik seçimi

Stolovitzky, gen ifadesi verileri ve gen seçim yöntemleri arasındaki ilişkiyi kör adamların fili tanımlamaya çalışmasına benzetmiş ve bununla ilgili kısa öyküyü hatırlatmıştır [95]. Öykü şöyledir: İlk kez fille karşılaşan körler ona dokunarak tanımlamak isterler. Ancak kör adamların her biri filin başka bir uzvuna temas ettiği için kendince başka bir görüş bildirir. Örneğin filin kulağını tutan adam filin bir yelpazeye benzediğini, filin hortumuna dokunan adam ise onun bir yılan gibi olduğunu ifade eder. Bu görüşler her ne kadar tamamen yanlış olmasa da, körler filin ne olduğu konusunda yanılgıya düşmektedir. Gen ifadesi verilerinin analiz süreci de bu metaforla bağdaştırılabilir. Zira her bir gen seçim yöntemi kendine özgü analiz adımları kullanarak hikâyenin yalnızca belirli bir dilimini mercek altına alır [95]. Özetlemek gerekirse; gen seçimine yönelik analizlerin, gen ifade verileri içinde saklı olan istatistiksel desenlere parça parça ışık tutması bütünsel bakışı gözden kaçırmaya sebebiyet vermektedir [95]. Bu soruna çözüm bulma yönündeki çabalar öznitelik seçiminde topluluk yöntemlerinin, hibrit yaklaşımların ve eklektik analizlerin gelişimine zemin hazırlamıştır.

Öznitelik seçiminde çoklu yöntemlerin birleşimi

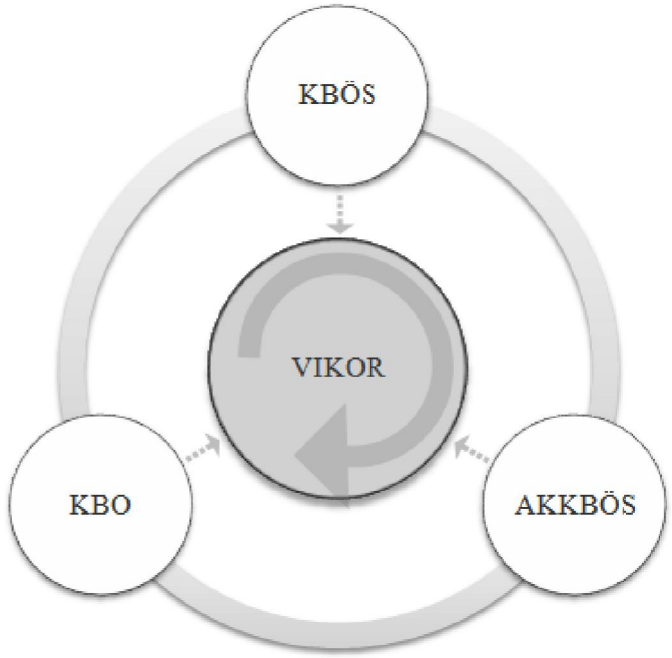
Denetimli öğrenmede, rastgele tahminden daha iyi performans gösteren ve hataları en azından bir miktar ilişkisiz olan çeşitli yöntemlerin birleşimi ile bireysel bir yöntem kullanmaktan daha iyi sonuçlar elde edilebileceği keşfedilmiştir [175, 176]. Yeni örnekleri sınıflandırmak için tekli yöntemlerin bireysel kararlarını bir şekilde birleştirmeyi öneren topluluk yöntemleri tahmin performansının geliştirilmesinde yararlı olmuştur [176]. Model birleşiminin sınıflandırma doğruluğu üzerindeki etkileri göz önüne alınarak benzer tekniklerin öznitelik seçim görevi üzerinde etkili olup olmadığı da araştırılmıştır [177]. Birden çok yöntemin birleşimi esas alınarak, özellikle mikrodizi verileri gibi büyük boyutlu ve küçük örnekleme sahip veriler üzerinde öznitelik seçiminde daha iyi performans gösteren yöntemler geliştirilebileceği gösterilmiştir [66, 177].

Öznitelik seçiminde, sınıflandırma performansının mümkün olduğunca geliştirilmesi amacıyla en anlamlı öznitelikleri içeren alt kümeyi elde etme çabası hâkimdir. Öznitelik seçimi için kullanılan farklı değerlendirme kriterleri farklı öznitelik alt kümeleri üretir [141]. Seçili alt kümenin değişimi sınıflandırma performansında da değişime yol açar. Bu nedenle öznitelik seçiminde daha kapsamlı ve daha tutarlı sonuçlara ulaşabilmenin yolları araştırılmıştır. Bu bağlamda, ayrı yöntemlerin tek bir model altında birleştirilmesi düşüncesi dikkat çekmiştir. Örneğin; bireysel yöntemlerin çıktılarını birleştirmek üzere uygulanan sınıflandırıcı topluluklarının yapımına benzer şekilde öznitelik seçimi toplulukları oluşturulmuştur [177]. Yine bu çerçevede, en ayırt edici özniteliklerin tespiti için, özniteliklerin çeşitli bireysel yöntemlerce seçilme sıklığını kıstas olarak kullanmak gibi daha basit yollar da denenmiştir [3].

Öznitelik seçiminde çoklu yöntemleri bütünleştiren modeller oluşturmanın bir yolu da ÇKKV yöntemlerini kullanmaktır. Birden çok öznitelik seçim kriterini tek bir değerlendirme süreci içinde bir araya getirmek için ÇKKV prensipleri üzerine kurulu çeşitli yaklaşımlar önerilmiştir [140, 141]. ÇKKV yöntemlerinin çelişkili durumları ele almaya uygun oluşu ve bütünsel tercihleri ortaya çıkarmadaki etkinliği öznitelik seçiminde ÇKKV yöntemlerine yönelimi motive etmektedir. Bu tez çalışmasının ikinci aşamasında birden fazla öznitelik seçim kriterini bir ÇKKV formülasyonu içinde bir araya getiren yeni bir öznitelik seçim yöntemi önerilmiştir.

Önerilen öznitelik seçim yöntemi

Birden fazla yöntemin bir araya getirilmesi için izlenen bir prosedürün yapısında genel olarak; tekli yöntemlerden hangilerinin birleştirileceğinin belirlenmesi, bu birleşimin üyesi olan yöntemlerin bireysel çıktılarının elde edilmesi ve uygun bir model oluşturularak bu çıktıların birleştirilmesi adımları yer alır [144, 177]. Tez çalışmasının ikinci aşamasında bu adımlar dikkate alınarak, çok kriterli öznitelik seçimi yapılması amacıyla ÇKKV tabanlı hibrit bir yöntem önerilmiştir. Kolon tümörü ve lenfoma gen ifadesi verilerinin sınıflandırılmasında en etkin genlerin tespiti için bilgi kuramsal öznitelik seçim yöntemlerinin değerlendirme kriterleri VIKOR yöntemi çerçevesinde bir araya getirilmiştir. Önerilen öznitelik seçim şemasının örnek bir gösterimi Şekil 4.4'te verilmiştir. Bu şematik gösterim, 3 çeşit kriter esas alınarak oluşturulan VIKOR tabanlı öznitelik seçim yapısını temsil etmektedir.

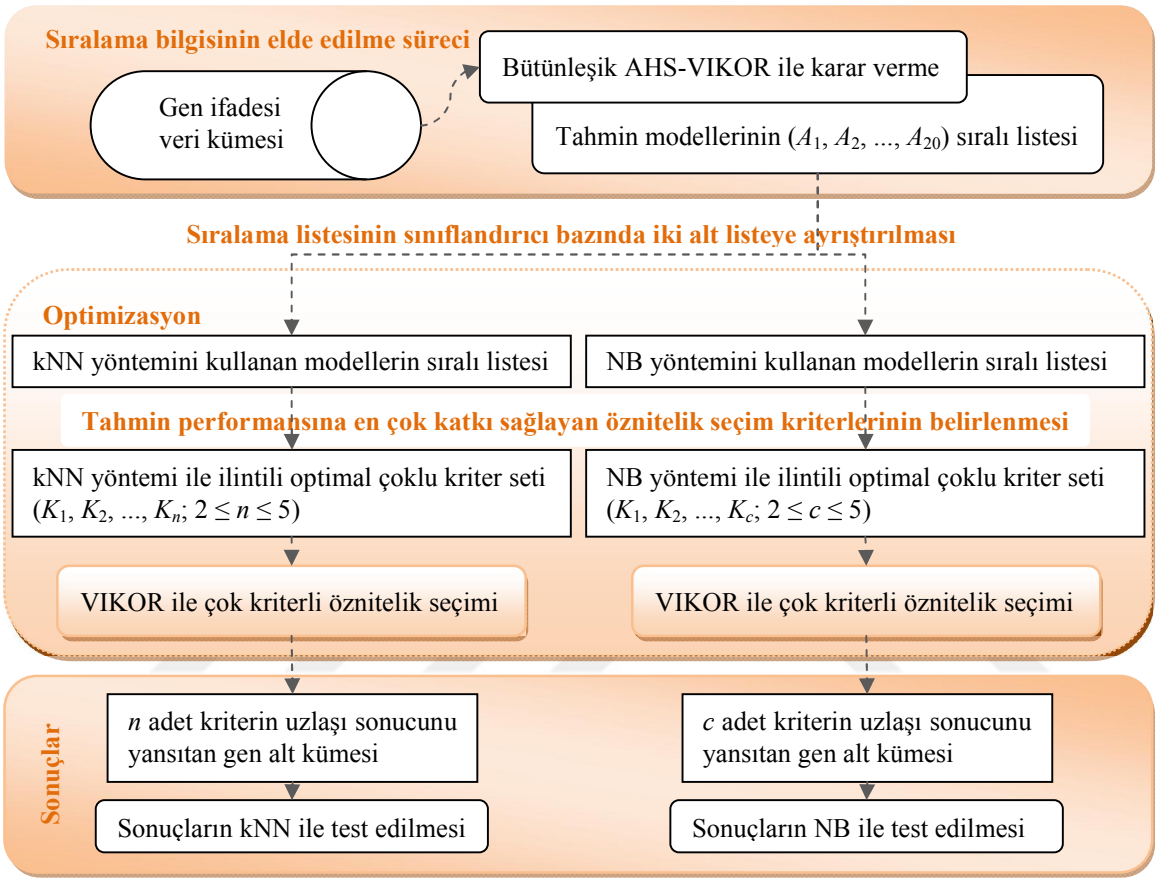


Şekil 4.4. VIKOR yöntemi ile çok kriterli öznitelik seçiminin şematik gösterimi

Önerilen yöntemde; KBÖS, KBO, AKKBÖS gibi bilgi kuramı tabanlı çeşitli öznitelik seçim yöntemlerinin amaç fonksiyonları birer karar kriteri olarak kullanılıp, VIKOR ile öznitelik seçimi gerçekleştirilmektedir. Böylece tekli kriterler bir ÇKKV mekanizması altında birleştirilerek daha güçlü ayrımcı yeteneği olan öznitelik alt kümelerinin bulunması amaçlanmıştır. Bu çerçevede karar alternatifleri olarak ele alınan öznitelikler, belirli öznitelik seçim kriterlerine göre elde edilen değerlendirme puanları esas alınarak VIKOR yönteminin işlem adımları ile analiz edilmiş ve en bilgilendirici özniteliklerin hangileri olduğuna karar verilmiştir. Önerilen yöntem VIKOR yönteminin her öznitelik seçim adımında tekrar kullanımına dayanır. Dolayısıyla VIKOR artırımsal öznitelik seçim adımları boyunca yineli olarak uygulanmaktadır. ÇKKV temelli bu yeni yaklaşıma VIKOR ile çok kriterli öznitelik seçimi, kısaca VÇKÖS yöntemi adı verilmiştir. Bilgi kuramsal öznitelik seçim kriterlerini ÇKKV yaklaşımı çerçevesinde birleştiren VÇKÖS yöntemi hibrit bir yöntem olmanın yanı sıra öznitelik seçiminde model birleşimini uygulamaktadır. VÇKÖS yöntemi, gen seçim işlemine göre modifiye edilmiş olan bir tür VIKOR yöntemi olarak tarif edilebilir.

Karar analizlerinde karar sonucunu belirlemek için ihtiyaç duyulan bilgilerin elde edilme süreci dikkate değer boyutta süre ve emek sarfına neden olabilmektedir. Önerilen öznitelik seçiminde bu sarfı azaltmak için tez çalışmasının birinci ve ikinci aşamaları arasında bir

geçiş sağlanmıştır. Birinci aşamadan başlanarak VIKOR ile çok kriterli öznitelik seçiminin test edilmesine kadarki sürecin hangi adımlardan oluştuğunu göstermek için Şekil 4.5'te bu süreci açıklayan bir akış şeması verilmiştir. Burada açıklanan süreç kolon tümörü ve lenfoma gen ifadesi verilerinin her biri için ayrı ayrı uygulanmaktadır.



Şekil 4.5. Tezin iki ana aşaması arasındaki geçiş işlemi ve VÇKÖS sürecine ilişkin iş akışı

Tez aşamaları arasında kurulan ilişkinin mantıksal temelinde, öznitelik seçiminin hem problem verisi hem de tercih edilen sınıflandırıcı yöntem göz önünde bulundurularak optimize edilmesi düşüncesi vardır. Belirli bir veri kümesi ele alınırken en bilgilendirici özniteliklerin her sınıflandırıcı için aynı olmaması göz önünde bulundurulmalıdır. Öznitelik seçiminde sınıflandırıcı yöntem tercihine bağlı olarak optimizasyon yapılması daha faydalıdır. Dolayısıyla her bir veri kümesi için, kNN ve NB sınıflandırıcılarıyla beraber kullanımı hâlinde tahmin performansına en çok katkı sağlayan bilgi kuramsal kriterlerin belirlenmesi yararlı olacaktır. Bu konuya ilişkin bilgi ihtiyacının karşılanması için ilk aşamadan gelen uzlaşık sonuçlar üzerinde basit analizler yapılmıştır.

Hatırlarsak tez çalışmasının ilk aşamasında kNN ve NB sınıflandırıcıları; KBÖS, D-KBÖS, KBF, KBO, NKBÖS, KKBM, KKBÖS, DAÖS, BKBM ve AKKBÖS olmak üzere bilgi kuramı tabanlı 10 ayrı öznitelik seçim kriteri ile kombine edilerek tahmin modelleri oluşturulmuştu. VIKOR yöntemiyle çok kriterli öznitelik seçimi yapılırken bu 10 kriterin hepsini kullanmak karar sürecini karmaşıklaştırarak sonuca negatif etki edebilir. Öznitelik seçim probleminin çözümü için ileri sürülen plân gereğince ilk iş olarak, hangi bireysel yöntemlerin bir arada kullanılacağı belirlenmelidir. Yani VÇKÖS yöntemi ile çözüme gidilirken çoklu kriter setini oluşturmak için bilgi kuramı kriterlerinden hangilerinin kullanılacağının kararlaştırılması gerekir. Bu doğrultuda tez çalışmasının ilk aşama çıktıları analiz edilerek kolon ve lenfoma verileri için kNN ve NB sınıflandırıcılarının hangi öznitelik seçim yapılarıyla daha iyi sonuç verdiği araştırılmıştır.

Kriter belirleme sürecinde, veri kümeleriyle ilgili uzlaşık sıralamalar üzerinde basit bir ayrıştırma yapılmıştır. Şöyle ki; bir gen ifadesi veri kümesi için birinci aşama sonunda edinilen uzlaşık sıralama bilgisi 20 elemanı olan sıralı bir listedir. Bu liste 2 farklı sınıflandırıcının her birinin 10 ayrı öznitelik seçim yapısıyla elde ettiği sonuçların başarı sırasını tutmaktadır. Bu sıralı liste şu şekilde analiz edilmiştir: Uzlaşık sıralama listesindeki elemanlar hangi sınıflandırma yöntemiyle bağlantılı olduğuna bakılarak 2 kısma ayrıştırılmıştır. Yani sınıflandırıcı bakımından ortak olan modeller sıralamadaki üstünlükleri korunarak daha küçük listeler içinde toplanmıştır. Böylece bir uzlaşık sıralama listesinden, 10'ar elemanlı 2 ayrı sıralı liste elde edilmiştir. Her liste, bilgi kuramı kriterlerinin ilgili sınıflandırma çözümündeki rolü hakkında bilgi taşımaktadır. Böylece her bir veri kümesi için kNN ve NB yöntemleriyle kombine edilen 10 farklı öznitelik seçim yapısının başarı sırası belirlenmiş olup; bunlar arasından en iyi sıralanan en az 2 en fazla 5 tanesinin amaç fonksiyonları karar kriterleri olarak seçilmeye aday gösterilmiştir. Burada kriter sayısı için üst sınırın 5 olarak belirlenmesinin nedeni, değerlendirmelerde toplamda 10 çeşit öznitelik seçim kriterinin kullanılmış olmasıdır. Tahminen bunlardan yarısının model birleşiminde iyi bir performans yakalamak için yeterli olacağı kanaatiyle hareket edilmiştir. Bunu doğrulama amacıyla bir dizi deneme yapılması ve her bir sınıflandırma çözümü için uygun kriter setinin karşılaştırmalı analizler yoluyla belirlenmesi gerekir. Bu noktaya kadar gerçekleştirilen işlemleri şöyle özetlemek mümkündür: VIKOR yapısı içinde bir araya getirilmek istenen çoklu kriterlerin belirlenmesine yönelik olarak; kolon verisinin kNN ile, kolon verisinin NB ile, lenfoma verisinin kNN ile ve lenfoma verisinin NB ile sınıflandırılması olmak üzere 4 ayrı sınıflandırma çözümünün birbirinden ayrı

şekilde ele alınması amaçlanmıştır. Bu amaçla tez çalışmasının birinci aşamasından elde edilen uzlaşık sıralamalardan yararlanılarak her bir sınıflandırma çözümü için en avantajlı olan 5 kriter belirlenmiştir. Böylece VIKOR yapısı içinde hangi tekli kriterlerin birleşimiyle daha iyi sonuç elde edileceği sorunsalının çözüm alanı daraltılmış ve bu problemin daha kolay ele alınması sağlanmıştır.

Öznitelik seçim kriterlerinin 4 ayrı sınıflandırma çözümündeki önemi değerlendirildikten sonra VÇKÖS yönteminin algoritma tasarımı yapılmıştır. Önerilen yöntemle gerçekleştirilen ileri yönlü aç gözlü öznitelik seçim işleminde durma kriteri olarak ön tanımlı öznitelik sayısı (*dk*) kullanılmıştır. Kolon tümörü verisi için VÇKÖS ile toplamda 100 öznitelik seçildiğinde seçim işlemi sonlandırılmaktadır. Lenfoma verisi içinse bu değer *dk* = 200 olarak belirlenmiştir. VÇKÖS algoritmasında; değerlendirmeye tâbi tutulan öznitelik (gen) sayısı *m*, bilgi kuramsal kriter sayısı *n* ve iterasyon indisi *l*, $l = 1, 2, ..., dk$ sembolleri ile ifade edilmiştir. VÇKÖS algoritması kapsamında $i = 1, 2, ..., m$; $j = 1, 2, ..., n$ olmak üzere *i*. özniteliğin *j*. kritere göre elde edilen değerlendirme puanı p_{ij} ile belirtilmiş olup, iteratif öznitelik seçimi süresince izlenen algoritmik prosedür aşağıda adım adım açıklanmıştır:

1. Değişkenlerin başlangıç değerlerinin atanması: Ön işlemlerden geçirilmiş gen ifade verisine ait aday öznitelikler $F = \{f_i, i = 1, 2, ..., m\}$ kümesi ile, bilgi kuramsal öznitelik seçim kriterleri $K = \{K_j, j = 1, 2, ..., n;\ 2 \le n \le 5\}$ kümesi ile, seçili öznitelikler $S = \{\}$ boş kümesi ile belirtilir. Bunların yanı sıra $w_j = 1/n; j = 1, 2, ..., n$ bağıntısı kullanılarak kriter ağırlıkları belirlenir. Maksimum grup faydası stratejisinin ağırlığını ifade eden *v* değişkenine 0,5 ila 0,7 aralığında uygun bir değer ataması yapılır ve iterasyon indisine ilk değeri ($l = 0$) atanır.
2. Artırımsal öznitelik seçim döngüsünün başlangıcı (döngü etiketi): $m = m - l$; $l = l + 1$.
3. Öznitelik (alternatif) değerlendirmeleri: *l*. döngü adımı kapsamında, $\forall K_j \in K$ kriter fonksiyonuna göre $\forall f_i \in F$ aday özniteliği için $p_{ij} = K_j(f_i)$ değerlendirme puanı hesaplanır. Eğer $l > 1$ ise p_{ij} puanı hesaplamaları, ilgili K_j kriteri tarafından öznitelikler arası fazlalık veya etkileşim bilgilerinin hesaplanmasında kullanılan formül uyarınca seçili alt küme elemanları göz önünde bulundurularak gerçekleştirilir.
4. Karar matrisinin oluşturulması: Değerlendirme puanları *l*. döngü adımına özgü hazırlanan $P_l[p_{ij}]_{m \times n}$ karar matrisine Eş. 4.51'de gösterildiği gibi yerleştirilir.

$$P_l = \begin{array}{c|c} & \begin{matrix} K_1 & \cdots & K_n \end{matrix} \\ \hline \begin{matrix} f_1 \\ \vdots \\ f_m \end{matrix} & \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{m1} & \cdots & p_{mn} \end{bmatrix} \end{array} \tag{4.51}$$

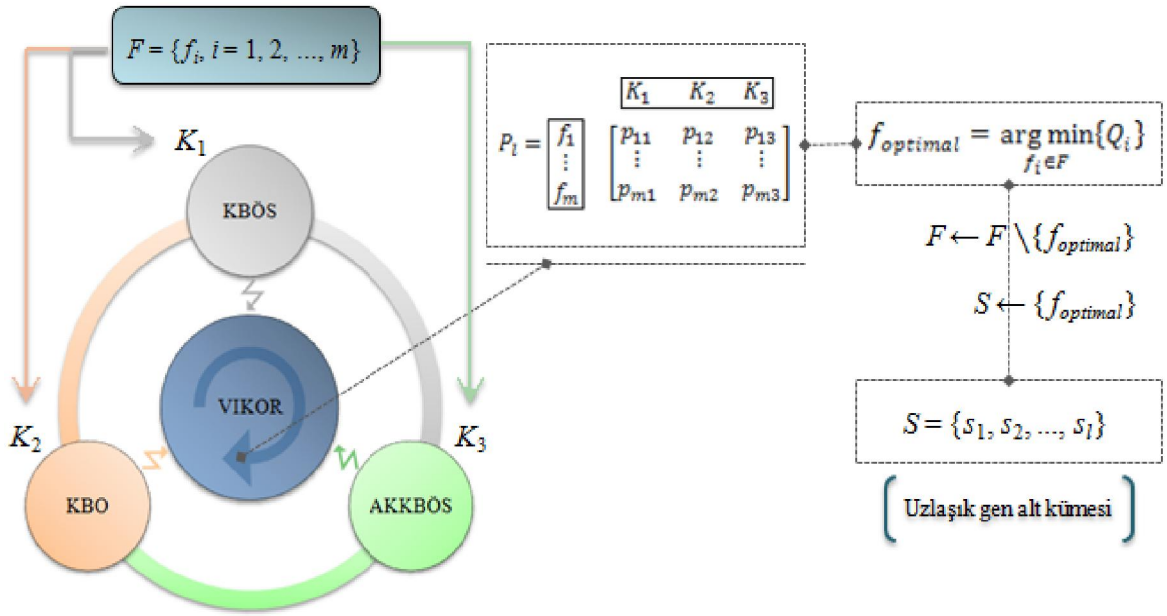
5. Aday özniteliklerin Q indeksi puanlarının hesaplanması: P_l matrisi üzerinden VIKOR hesaplamaları ile $\forall f_i \in F$ için uzlaşık değerlendirme puanı Q_i hesaplanır.
6. Öznitelik tespiti: l. adımdaki optimal özniteliğin hangisi olduğu Eş. 4.52'de verilen formül kullanılarak belirlenir.

$$f_{optimal} = \arg\min_{f_i \in F} \{Q_i\} \tag{4.52}$$

7. Öznitelik seçimi: En uygun olduğu tespit edilen öznitelik, aday öznitelik kümesinden seçili alt kümeye aktarılır; $F \leftarrow F \setminus \{f_{optimal}\}$, $S \leftarrow \{f_{optimal}\}$.
8. Değişken değerlerinin güncellenmesi: Öznitelik seçimini takiben kriterlerle ilişkili değişkenlerin güncel değerleri belirlenir. Örneğin DAÖS yöntemi kriter seti içinde yer alan yöntemlerden biri ise aday özniteliklerin DAÖS kriterine göre belirlenen ağırlıkları seçili özniteliğe göre yeniden hesaplanır. AKKBÖS yöntemi kriter olarak kullanılmakta ise AKKBÖS kriteriyle bağlantılı MB_S ve mF_S değerleri son seçilen niteliğin bilgi paylaşım ilişkileri dikkate alınarak güncellenir. Yani belirli bir kritere özgü değişken değerleri güncellenirken hesaplamalar o kriterin bilinen tanımı çerçevesinde gerçekleştirilir, burada değişen tek şey hesaplamalarda yalnızca o kritere göre değil de n adet kriterin uzlaşık kararına göre belirlenmiş seçili özniteliklerin kullanılmasıdır.
9. Durma şartının kontrol edilmesi: Eğer $l < dk$ ise döngüye devam et (adım 2'ye dön) komutu, değilse döngüden çık komutu uygulanır.
10. Sonuç: Seçili öznitelik at kümesi $S = \{s_1, s_2, ..., s_{dk}\}$ sonuç olarak döndürülür.

Yukarıdaki işlem adımlarına göre VÇKÖS yöntemi ile öznitelik seçim sürecinin uygulanışı şu şekildedir: Öncelikle n ayrı kriterden oluşan kriter setinin tüm elemanları eşit önem ağırlığı ile ağırlıklandırılır. Uygulamada ilgili veri kümesindeki her bir öznitelik farklı bir alternatif olarak temsil edilir ve tüm kriterlere göre aday özniteliklerin değerlendirme puanları elde edilerek karar matrisine yerleştirilir. Karar matrisi, VIKOR analizleriyle ele alınıp tüm özniteliklerin uzlaşık sıralaması elde edildikten sonra sıralamanın tepesindeki bir öznitelik seçili alt kümeye aktarılır. Böylelikle VÇKÖS yönteminin temsil ettiği

bütünsel tercih prosedürüne göre en seçkin olan ilk öznitelik s_1 belirlenmiş olmaktadır. Durma şartı olarak tanımlanan *dk* adet öznitelik seçilinceye değin döngüsel adımlarla bu işleme devam edilmektedir. Bilgi kuramı kriterleri tarafından esas alınan ideal açgözlü seçim prosedürünün genel yapısı gereğince, ilk özniteliğin seçilmesinden sonraki adımlarda tüm bireysel kriter değerlendirmeleri seçili özniteliklerin de hesaba katılması suretiyle elde edilir. Her adımda aday öznitelikler için *n* ayrı karar kriterine göre elde edilen bireysel çıktılar VIKOR analizleriyle değerlendirilir. Bunun neticesinde elde edilen *Q* indeksi puanları genel tercihi belirlemek için kullanılıp, minimum *Q* değerine sahip olan öznitelik optimum öznitelik olarak seçilir. Burada VIKOR analizleri, KB tabanlı yöntemlerin artırımsal öznitelik seçim yapısı içine gömülü durumdadır. Dolayısıyla her yinelemede bir alternatif eksilmekle birlikte geriye kalan aday özniteliklerin hedef sınıf ve seçili küme elemanları ile ilişkileri hesaba katılarak yeni bir karar matrisi oluşturulur. VÇKÖS yönteminde yalnızca alternatiflerin *Q* indeksine göre sıralanmasına kadarki VIKOR işlem adımlarından yararlanılmıştır. Bundan sonraki diğer analiz adımlarına ihtiyaç yoktur. Yani orijinal VIKOR yöntemindeki gibi uzlaşık çözüm araştırması yapılmaz. VIKOR ile çok kriterli öznitelik seçimi sona erdiğinde *n* ayrı kriterin konsensüs kararını yansıtan uzlaşık bir öznitelik alt kümesi elde edilmektedir. VÇKÖS algoritmasının *l*. döngü adımı içinde gerçekleştirilen işlemler zihinde daha kalıcı olması için Şekil 4.6'da gösterildiği gibi şematize edilebilir.



Şekil 4.6. VÇKÖS algoritmasının artırımsal adımları içinde gerçekleştirilen işlemler

VÇKÖS yönteminin algoritma tasarımı çerçevesinde son olarak, kriter setine ait elemanların ve çoğunluk kuralına göre karar verme stratejisinin ağırlığı olan v değerinin uygun şekilde belirlenmesi amacıyla performans karşılaştırmasına dayanan bir araştırı yapılmıştır. Hatırlarsak kolon ve lenfoma verileri üzerinde kNN ve NB yöntemleriyle gerçekleştirilen sınıflandırmalarda en etkili çözümlerin hangi öznitelik seçim kriterleriyle elde edildiği daha önceden belirlenmişti. Bu çerçevede her bir sınıflandırmada en iyi sıralı ilk 2 kriterden başlanarak sırasıyla en iyi ilk 3, en iyi ilk 4 ve en iyi ilk 5 kriterden oluşan kümelerin kriter seti olarak denenmesi uygun bir araştırma çerçevesi olabilir. Ancak gözlemler doğrultusunda, oldukça iyi tahmin performansına ulaşılması veya kabul edilebilir bir hesaplama süresinin aşılması durumları da dikkate alınarak yalnızca ilk 2 ve ilk 3 kriter üzerinde deneme yapılması daha uygun görülmüştür. Bu doğrultuda VÇKÖS yönteminde kriter seti tercihinin tahmin performansı üzerindeki etkisi araştırılırken çoğu kez 2 ve 3 elemanlı kriter setleri kullanılmış olup, sadece hatasız sınıflandırmaya ulaşılamadığı durumlarda 4 ve 5 elemanlı kriter setlerinin de araştırmaya dâhil edilmesi yoluna başvurulmuştur. Bu süreçte ayrıca, VIKOR yöntemine ilişkin v değişken değeri için 0,5; 0,6 ve 0,7 sayıları denenerek bunlardan en uygun olanın seçilmesi amaçlanmıştır. Bu değer çoğunluk kuralına göre karar verme stratejisinin ağırlığını ifade eder. Burada v ağırlığına atanacak olan uygun değerin araştırılması sürecinde niçin 0,5; 0,6 ve 0,7 sayılarının denendiğine açıklık getirmek faydalı olacaktır. Şöyle ki; VIKOR yöntemi çerçevesinde v ağırlığına 0,5 değerinin atanması, çoğunluk kuralına göre karar verme stratejisi ile karşıt görüşlerin asgarî bireysel pişmanlığını ifade eden strateji arasındaki uzlaşıyı temsil eder. Bunun yanı sıra; karar kriterlerinin sayısı n olmak üzere, v değeri $v = (n+1)/(2\times n)$ formülüne göre belirlenebilir [178]. O hâlde $2 \leq n \leq 5$ olmak üzere; kriter sayısı 2 ise v değeri 0,75 $(3/4)$; kriter sayısı 5 ise v değeri 0,6 $(6/10)$ çıkmaktadır. Bu nedenle v için değer araştırması yapılırken; 0,5 ile 0,75 arasında yer alan 3 değerden, yani 0,5; 0,6 ve 0,7 değerlerinden yararlanılması uygun görülmüştür.

Değişen kriter setleri ve v değerlerine göre bir dizi deneme yapılmış ve buradan elde edilen performans değerleri değerlendirilerek kriter belirleme süreci tamamlanmıştır. Aynı zamanda v ağırlığı için uygun değerler belirlenmiştir. Kolon ve lenfoma verilerinin her biri için 2'şer tane olmak üzere toplamda 4 ayrı kriter seti elde edilmiştir. Buna göre VÇKÖS yöntemi 4 farklı kriter seti kullanılarak her sınıflandırma çözümü için ayrı ayrı uygulanmıştır.

5. DENEYSEL UYGULAMA VE DEĞERLENDİRMELER

Bu bölümde kolon tümörü ve lenfoma gen ifadesi veri kümeleri üzerinde gerçekleştirilen deneysel çalışma hakkında bilgi verilmiş ve deney sonuçları tablo ve grafiklerle özetlenmiştir. Öncelikle tahmin modellerinin bütünleşik AHS-VIKOR ile değerlendirilmesine ilişkin deneysel çalışma, ardından VIKOR ile çok kriterli öznitelik seçimine ilişkin deneysel uygulama detaylı şekilde açıklanmıştır.

Veri kümeleri üzerinde uygulanan makine öğrenmesi temelli tüm işlem adımlarında ve bütünleşik AHS-VIKOR ile gerçekleştirilen değerlendirme sürecinde, ayrıca önerilen çok kriterli öznitelik seçim yönteminin uygulanmasında WEKA (Waikato Environment for Knowledge Analysis) [179] kaynak kodlarından yararlanılmış olup, işlemler NetBeans IDE 8.2 editörü üzerinde gerçekleştirilmiştir.

5.1. Tahmin Modellerinin Değerlendirilmesine Yönelik Deneysel Uygulama

Deneysel uygulama sürecinde öncelikle veri kümeleri ön işlemden geçirilmiştir. Lenfoma veri kümesindeki eksik veri noktaları eğitim verisi ortalamaları ile doldurulmuştur. Kolon tümörü veri kümesi eksik değer içermediği için veri impütasyonu işlemine tâbi tutulmamıştır. Standartlaştırma ve 3-durumlu ayrıklaştırma işlemleri ise tüm veri kümelerine uygulanmıştır.

Veri ön işleme aşamasından sonraki süreç bütünleşik AHS-VIKOR modelinin kurulum süreci olarak ifade edilebilir. Bu süreçte ilk önce gen ifadesi verilerinin sınıflandırılmasında kullanılmak üzere; kNN ve NB yöntemleri ile bilgi kuramı tabanlı öznitelik seçim kriterlerini ikili kombinasyonlarla bir araya getiren 20 farklı tahmin modeli oluşturulmuş ve Çizelge 5.1'de gösterildiği gibi bu modeller karar alternatifleri olarak ifade edilmiştir. kNN yönteminde en yakın komşu sayısını ifade eden k değişkenine 1 değeri atanmış ve bütün deneysel uygulama süreci boyunca bu değer esas alınarak işlem yapılmıştır.

Tahmin modellerinin performans ölçümünde 5 ayrı metrikten yararlanılmış olup, bu metrikler Çizelge 5.2'de gösterildiği gibi karar kriterleri olarak ifade edilmiştir. Bu süreçte

gereksiz yere kriter kullanımından kaçınmaya özen gösterilerek model başarımı hakkında kapsamlı bilgi sağlayan metrikler tercih edilmiştir.

Çizelge 5.1. Karar alternatiflerini temsil eden tahmin modelleri

Alternatifler	Tahmin Modelleri		Alternatifler	Tahmin Modelleri	
A_1	KBÖS	kNN	A_{11}	KBÖS	NB
A_2	D-KBÖS	kNN	A_{12}	D-KBÖS	NB
A_3	KBF	kNN	A_{13}	KBF	NB
A_4	KBO	kNN	A_{14}	KBO	NB
A_5	NKBÖS	kNN	A_{15}	NKBÖS	NB
A_6	KKBM	kNN	A_{16}	KKBM	NB
A_7	KKBÖS	kNN	A_{17}	KKBÖS	NB
A_8	DAÖS	kNN	A_{18}	DAÖS	NB
A_9	BKBM	kNN	A_{19}	BKBM	NB
A_{10}	AKKBÖS	kNN	A_{20}	AKKBÖS	NB

Çizelge 5.2. Karar kriterlerini temsil eden performans metrikleri

Kriterler	Metrikler
K_1	AUPRC
K_2	AUC
K_3	MKK
K_4	Doğruluk
K_5	F-ölçütü

Performans ölçümlerine ilişkin şu nokta önemlidir: İkili sınıflandırma problemlerinde çeşitli metriklere göre kolaylıkla ölçüm yapılabilmektedir. İkiden çok sınıf söz konusu ise problemin bir dizi ikili sınıflandırma problemi şeklinde ele alınması faydalı olabilir.

WEKA platformunda çok sınıflı bir sınıflandırma problemi ile karşılaşıldığında, her bir sınıfın birer kez pozitif sınıf geriye kalanlarınsa negatif sınıf olarak kabul edilmesi suretiyle elde edilen bireysel değerler pozitif sınıfın büyüklüğüne göre ağırlıklı ortalama alınarak birleştirilmektedir. Deneylerde hem kolon hem de lenfoma verilerinde ölçüm sonucu olarak, WEKA tarafından hesaplanan bu ağırlıklı ortalama değerleri kullanılmıştır.

5.1.1. AHS ile kriter ağırlıklarının hesaplanması

Performans metriklerine ilişkin avantaj ve dezavantajlar göz önüne alınarak kriterler nispî önemlerine göre derecelendirilmiştir. Kolon tümörü ve lenfoma veri kümelerinde örneklerin sınıflara dağılımı dengeli değildir. Bu durum göz önünde bulundurularak ikili karşılaştırmalarda, veri dengesizliğinden daha az etkilenen metrikler daha büyük önem derecesi ile ölçeklenmiştir. Doğruluk ve f-ölçütü sınıf dengesizliğine karşı dayanıksız metrikler arasında yer alır. Dolayısıyla bu ikisinin AUPRC, AUC ve MKK metriklerinden daha önemsiz olduğu kabul edilmiştir. Doğruluk ve f-ölçütü kendi aralarında eşit önem derecesiyle ölçeklenmiştir. Benzer şekilde AUC ve MKK metriklerine eşit önem derecesi atanmıştır. PRC grafiklerinin çok dengesiz veri kümeleri için bazen ROC grafiklerinden daha bilgilendirici olduğu ileri sürülmüştür [171]. Bundan yola çıkılarak AUPRC ölçütünün; AUC ve MKK metriklerine göre bir derece daha önemli olduğuna karar verilmiştir. Bu değerlendirmeler AHS yöntemi ile ele alınmıştır. AHS ile kriter ağırlıklandırma prosedürü çerçevesinde temel ölçek (Bkz. Çizelge 4.3) kullanılarak Çizelge 5.3'te gösterilen ikili karşılaştırma matrisi oluşturulmuştur.

Çizelge 5.3. Kriterlerin göreli önemlerini yansıtan ikili karşılaştırma matrisi

Kriterler	K_1	K_2	K_3	K_4	K_5
K_1	1	2	2	3	3
K_2	0,5	1	1	2	2
K_3	0,5	1	1	2	2
K_4	0,33	0,5	0,5	1	1
K_5	0,33	0,5	0,5	1	1

Karşılaştırma matrisi üzerinde AHS yönteminin işlem adımları uygulanarak Çizelge 5.4'te gösterilen kriter ağırlıkları elde edilmiştir. Karar ağırlıkları elde edildikten sonra ikili karşılaştırmaların tutarlılığını kontrol etmeye yönelik işlem adımları uygulanmıştır. İşlemler sonucunda λ_{max} = 5,0133 olarak hesaplanmış ve tutarlılık oranını ifade eden CR değeri 0,003 çıkmıştır. Buna göre CR < 0,1 şartı sağlanmıştır, yani karşılaştırma matrisi tutarlıdır. AHS hesaplamaları sonucunda belirlenen karar ağırlıklarında herhangi bir değişiklik yapılmaksızın aynı ağırlıklar her iki veri kümesi için kullanılmıştır.

Çizelge 5.4. Karar kriterlerinin AHS yöntemi ile hesaplanan ağırlıkları

Karar Kriteri	K_1	K_2	K_3	K_4	K_5
Kriter Ağırlığı	0,3683	0,2064	0,2064	0,1094	0,1094

5.1.2. Ağırlıklı kriterler kullanılarak VIKOR yönteminin uygulanması

Bütünleşik AHS-VIKOR yaklaşımı çerçevesinde AHS'den sonra VIKOR yönteminin uygulanmasına geçilmiştir. VIKOR yönteminin her veri kümesi için ayrı ayrı uygulanması gerekir. Bu doğrultuda öncelikle kolon tümörü ve lenfoma veri kümeleri için alternatif tahmin modellerinin kriterler altındaki değerlendirmelerine ilişkin bilgiler 10-kat çapraz doğrulama stratejisi ile elde edilmektedir. Ölçümler esnasında parametre ayarları şu şekilde yapılmıştır: KBÖS ve D-KBÖS yöntemlerinde fazlalık teriminin göreli önemini ifade eden β değeri 0,5 olarak belirlenmiştir. KKBÖS yöntemi diğer yöntemler gibi yalnızca öznitelikleri sıralamak için kullanılmıştır. Yani değerlendirme sürecinin genel yapısı gereği KKBÖS yönteminde δ eşik değerine −1 atanarak yedekli özniteliklerin silinmesi önlenmiştir. Bunun sonucu olarak KKBÖS yönteminde fazlalık analizi yapılamamış ve fazlalıklar göz ardı edilmiştir. kNN yönteminde k için 1 değeri kullanılmıştır. Ayrıca daha önce de belirtildiği gibi doğruluk dışındaki tüm kriterlerde ölçüm sonucu, ağırlıklı ortalama hesaplaması ile belirlenmiştir. Öznitelik seçiminde örnek sayısı ve deneysel bulgulardan yararlanılarak bir durdurma kriteri belirlenmesi gerekmektedir. Kolon ve lenfoma veri kümeleri için seçili alt küme boyutunun sırasıyla 100 ve 200 özniteliğe ulaşması durma kriteri olarak belirlenmiştir. Bunlar dışında ise WEKA varsayılan değerleri değiştirilmeden kullanılmıştır.

Parametre ayarlamaları yapıldıktan sonra tahmin modelleri kolon tümörü verileri üzerinde uygulanıp, performans değerleri elde edilmiştir. Ölçümler neticesinde elde edilen performans bilgileri kullanılarak kolon tümörü veri kümesi için EK-1'de gösterilen karar matrisi oluşturulmuştur. Kolon verisinden sonra lenfoma veri kümesi için performans ölçümleri gerçekleştirilerek karar matrisinin eleman değerleri belirlenmiştir. Bu değerlere göre EK-2'de gösterilen karar matrisi oluşturulmuştur. Karar matrislerinde A_1, A_2, ..., A_{20} tahmin modellerini (Bkz. Çizelge 5.1); K_1, K_2, ..., K_5 performans metriklerini (Bkz. Çizelge 5.2) ifade etmektedir. Ayrıca karar matrislerinde her bir alternatifin bitişiğindeki $|S|$ değeri ilgili modelin yapısına katılan tahminci öznitelik sayısına karşılık gelmektedir. Diğer bir ifadeyle, AUC ölçüsüne göre en yüksek tahmin performansının elde edildiği öznitelik sayısını belirtmektedir. Kolon ve lenfoma veri kümeleri için tahmin modellerinin 5 ayrı kriter altındaki performanslarına göre elde edilen çelişkili sıralama sonuçları EK-3'te gösterilmiştir. Her bir kriter altında elde edilen sıralama listesinde alternatifler en iyiden en kötüye doğru sıralanmış olup, sıra değeri eşit olan alternatifler sağ üst köşelerine iliştirilen asteriks $^{(*)}$ işareti belli edilmiştir. Ayrıca iki ayrı performans değerine endeksli eşitlik durumlarının ardışık sırada olması hâlinde birbiriyle karışmaması içinse ikinci duruma ait eşit sıralı alternatifler çift asteriks $^{(**)}$ ile işaretlenmiştir.

Alternatiflerin kolon ve lenfoma gen ifadesi verileri üzerinde gösterdiği performansa ilişkin birer karar matrisi oluşturulmasının ardından bu matrisler bütünleşik AHS-VIKOR yöntemi çerçevesinde sırasıyla analiz edilmiştir. Bu süreçte mevcut karar ağırlıkları dikkate alınarak VIKOR yönteminin işlem adımları uygulanmış olup, strateji ağırlığını ifade eden v değeri tüm veri kümeleri için 0,5 olarak belirlenmiştir. Değerlendirmelerde 5 adet kriterin tümü fayda kriteri olarak işleme alınmaktadır. Çünkü bu kriterler altında elde edilen değerlerin sayısal olarak artışı amaca ulaşma yönündeki ilerlemeyi gösterir.

Kolon ve lenfoma verileri üzerinde VIKOR yöntemi, AHS'den gelen ağırlıklı kriterlere göre (Bkz. Çizelge 5.4) uygulanmıştır. Bu çerçevede her bir alternatifle ilgili S, R ve Q değerleri elde edilmiştir. Elde edilen değerler EK-4'te gösterilmiştir. Ardından S, R ve Q indekslerinde tutulan değerlere göre alternatiflerin her bir indeks altındaki sırası belirlenmiştir. Bunlar içinden Q indeksi altında elde edilen sıralama, ilgili veri kümesi için tahmin modellerinin uzlaşık sırasını ifade etmektedir. Her bir veri kümesi için ayrı ayrı elde edilen sıralama listeleri tek bir çizelge hâline getirilerek Çizelge 5.5'te gösterilmiştir. Burada A_1, A_2, ..., A_{20} tahmin modellerini temsil eden alternatifleri (Bkz. Çizelge 5.1) ifade

etmektedir. *Q* indeksi altındaki uzlaşık sıralamada, sıralı alternatiflerin karşılığı olan modeller parantez içlerinde verilmiştir. Bazı alternatifler arasında puan farkı bulunmaması sebebiyle bu alternatiflerin sıralamaları arasında eşitlik durumu ortaya çıkmıştır. Atanan sıra numaraları ise yalnızca indis sırasını yansıtmaktadır. Bir sıralama listesinde çizelge sıra düzeni nedeniyle ardışık sıralarda gösterilen, ancak gerçekte aynı sıra değerini paylaşan alternatifler sağ üst köşelerine asteriks $^{(*)}$ işareti ilâve edilerek belirtilmiştir.

Çizelge 5.5. Alternatiflerin *S*, *R* ve *Q* indekslerine göre sıralanışını gösteren listeler

Sıra Numarası	Veri Kümeleri					
	Kolon			Lenfoma		
	S	R	Q	S	R	Q
1	A_8	A_8	A_8 (DAÖS & kNN)	A_2^*	A_2^*	A_2^* (D-KBÖS & kNN)
2	A_{20}	A_{20}	A_{20} (AKKBÖS & NB)	A_4^*	A_4^*	A_4^* (KBO & kNN)
3	A_{14}	A_{10}	A_{14} (KBO & NB)	A_{11}^*	A_{11}^*	A_{11}^* (KBÖS & NB)
4	A_{11}	A_1	A_{10} (AKKBÖS & kNN)	A_{12}^*	A_{12}^*	A_{12}^* (D-KBÖS & NB)
5	A_2	A_{14}	A_{11} (KBÖS & NB)	A_{14}	A_{14}	A_{14} (KBO & NB)
6	A_{10}	A_6	A_1 (KBÖS & kNN)	A_{20}	A_{20}	A_{20} (AKKBÖS & NB)
7	A_6	A_{11}	A_6 (KKBM & kNN)	A_{19}	A_{19}	A_{19} (BKBM & NB)
8	A_1	A_{12}	A_2 (D-KBÖS & kNN)	A_{13}	A_{13}	A_{13} (KBF & NB)
9	A_{12}	A_2	A_{12} (D-KBÖS & NB)	A_{16}	A_{16}	A_{16} (KKBM & NB)
10	A_{16}	A_4	A_4 (KBO & kNN)	A_{10}	A_9	A_{10} (AKKBÖS & kNN)
11	A_4	A_{18}	A_{16} (KKBM & NB)	A_6	A_{10}	A_9 (BKBM & kNN)
12	A_{13}	A_{13}^*	A_{18} (DAÖS & NB)	A_9	A_6	A_6 (KKBM & kNN)
13	A_{18}	A_{15}^*	A_{13} (KBF & NB)	A_1	A_3	A_1 (KBÖS & kNN)
14	A_{15}	A_{16}^*	A_{15} (NKBÖS & NB)	A_3	A_1	A_3 (KBF & kNN)
15	A_{19}	A_{19}	A_{19} (BKBM & NB)	A_{15}	A_{15}	A_{15} (NKBÖS & NB)
16	A_7	A_{17}	A_7 (KKBÖS & kNN)	A_8	A_8	A_8 (DAÖS & kNN)
17	A_9	A_7	A_{17} (KKBÖS & NB)	A_{17}	A_{17}	A_{17} (KKBÖS & NB)

Çizelge 5.5. (devam) Alternatiflerin *S*, *R* ve *Q* indekslerine göre sıralanışını gösteren listeler

18	A_3	A_9	A_9 (BKBM & kNN)	A_5	A_{18}	A_5 (NKBÖS & kNN)
19	A_5	A_3	A_3 (KBF & kNN)	A_{18}	A_5	A_{18} (DAÖS & NB)
20	A_{17}	A_5	A_5 (NKBÖS & kNN)	A_7	A_7	A_7 (KKBÖS & kNN)

Bütünleşik AHS-VIKOR ile karar verme sürecinde gelinen son aşama uzlaşık çözümün bulunması aşamasıdır. Bu aşamada kabul edilebilir avantaj ve karar vermede kabul edilebilir istikrar koşulları kontrol edilmiştir. Kolon ve lenfoma veri kümeleri için uzlaşık çözümün bulunmasına yönelik işlemlerin sonuçları aşağıda sırayla açıklanmış ve bütünleşik AHS-VIKOR yöntemi ile elde edilen uzlaşık sonuçlar değerlendirilmiştir.

Kolon ve lenfoma veri kümeleri üzerinde değerlendirmeye tâbi tutulan alternatiflerin sayısı $m = 20$ olduğu için kabul edilebilir avantaj koşulunun kontrolünde kullanılan DQ eşik değerleri aynıdır. $DQ = 1/(m-1)$ şeklinde ifade edilen formüle göre $DQ = 0{,}0526$ olarak hesaplanmıştır. Kolon tümörü veri kümesinde Q listesinin başında yer alan ilk iki alternatif arasındaki değer farkı $Q(A_{20}) - Q(A_8) = 0{,}1383$ olarak hesaplanmıştır. Bu fark değeri DQ eşik değerinden daha büyüktür. Yani; $0{,}1383 > 0{,}0526$ olması sebebiyle kabul edilebilir avantaj koşulu karşılanmış durumdadır. Ayrıca A_8; *S*, *R* ve *Q* indekslerinin hepsinde en düşük değeri alan alternatiftir. Bu durumda her iki koşul da sağlanmış olup, Q(minimum) değerine sahip olan A_8 alternatifi uzlaşık çözüm olarak belirlenmiştir. A_8 alternatifi, DAÖS ve kNN yöntemlerinin bir arada kullanımını temsil eden tahmin modelidir. Uzlaşık çözümün bulunmasıyla birlikte kolon tümörü verisi için karar süreci sona ermiştir.

Lenfoma veri kümesinde *S*, *R* ve *Q* listelerinin hepsinin en üst sırasında A_2 alternatifi yer almaktadır. Bu, karar vermede kabul edilebilir istikrarın sağlandığını gösterir. *Q* listesinin başındaki ilk iki alternatifin puan farkı $Q(A_4) - Q(A_2) = 0$ olarak hesaplanmıştır. Bu fark değeri, $DQ = 0{,}0526$ olarak belirlenen eşik değerinden küçüktür. Bu durumda kabul edilebilir avantaj koşulu sağlanmaz ve çoklu alternatiflerden oluşan bir uzlaşık çözüm kümesi önerilmesi gerekir. Lenfoma verisi için A_2, A_4, A_{11} ve A_{12} alternatiflerinin *S*, *R* ve *Q* değerleri 0 olarak hesaplanmış ve birbirine eşit çıkmıştır. Bu nedenle *S*, *R* ve *Q* indekslerine göre elde edilen sıralamalarda ilk 4 alternatifin sıralaması birbiriyle aynı olup,

bu alternatifler birinci sırayı paylaşmaktadır. Bu çerçevede kabul edilebilir avantaj şartına göre; en düşük Q değerine sahip olan ilk 4 alternatif arasında anlamlı bir fark olmadığı tespit edilmiştir. Bununla birlikte, Q listesinin ilk sırasındaki A_2 alternatifi ile 5. sırasındaki A_{14} alternatifinin puan farkı, $Q(A_{14}) - Q(A_2) = 0{,}0638$ çıkmıştır. Bu fark DQ değerini aştığı için A_{14} alternatifine gelindiğinde kabul edilebilir düzeyde avantaj farkının oluştuğu belirlenmiştir. Dolayısıyla; Q listesinde en üstte sıralanan A_2, A_4, A_{11} ve A_{12} alternatifleri uzlaşık çözümler olarak tavsiye edilmiştir. Bu alternatifler tarafından temsil edilen tahmin modelleri sırasıyla; D-KBÖS ile kNN, KBO ile kNN, KBÖS ile NB ve D-KBÖS ile NB yöntemlerinin birlikte kullanımına karşılık gelmektedir.

5.1.3. Uzlaşık sonuçların irdelenmesi

Bütünleşik AHS-VIKOR ile elde edilen uzlaşık sonuçlar, sınıflandırıcı yöntemler ve öznitelik seçiciler bakımından ayrı ayrı değerlendirilmiştir. Kolon tümörü verisi için elde edilen uzlaşık sıralamaya göre kNN yöntemi sıralamada ilk 10'a giren alternatifler içinde daha çok yer tutar, bununla birlikte uzlaşık çözümde de temsil edilmiştir. Görünen o ki; kNN, NB sınıflandırıcısına bakışla daha avantajlıdır. Öznitelik seçimi perspektifinden bakıldığında ise uzlaşık çözümde yer alan DAÖS yönteminin performansı dikkat çekmektedir. Öte yandan DAÖS yöntemi kolon verisi üzerinde kNN yöntemiyle birlikte kullanıldığında çok iyi sonuç verirken, NB yöntemiyle ancak vasat seviyede sayılabilecek bir sıralama değeri elde etmiştir. AKKBÖS yöntemi öznitelik seçiciler arasında fark edilir üstünlük sergileyen yöntemlerden bir diğeridir. Uzlaşık sıralamada 2. ve 4. sıradaki A_{20} ve A_{10} alternatiflerinin öznitelik seçim yapısını oluşturan AKKBÖS yöntemi en iyi ilk 5 model içinde 2 kez temsil edilmiştir. Dikkate değer sonuçlara ulaşan öteki yöntemler; KBO, KBÖS, D-KBÖS ve KKBM yöntemleridir. Göze çarpan sonuçlardan biri de NKBÖS, BKBM, KKBÖS ve KBF yöntemleriyle oluşturulan tahmin modellerinin kötü sıralanmış olmasıdır.

Lenfoma verisi için elde edilen uzlaşık sıralama sonucuna göre NB sınıflandırıcısı ile oluşturulmuş tahmin modellerinin sayısı ilk 10 sıra içinde daha çoktur. En iyi ilk 10 sırada NB sınıflandırıcısını içeren tahmin modellerinin sayısı 7 iken kNN sınıflandırıcısını içeren tahmin modellerinin sayısı 3'tür. Bunun yanı sıra uzlaşık çözüm kümesi içinde NB ve kNN sınıflandırıcılarını içeren 2'şer model bulunmaktadır. En başarılı tahmin sonucuna ulaştığı belirlenen A_2 (D-KBÖS ile kNN) ve A_4 (KBO ile kNN) modellerinin yapısında kNN

sınıflandırıcısı yer alırken A_{11} (KBÖS ile NB) ve A_{12} (D-KBÖS ile NB) modellerinin yapısında NB sınıflandırıcısı bulunmaktadır. D-KBÖS yöntemi hem A_2 hem de A_{12} alternatifinin yapısında yer almasıyla uzlaşık çözüm kümesinde 2 defa temsil edilmiştir. Bu nedenle D-KBÖS lenfoma verisi üzerinde öznitelik seçim görevinin ifası için en etkili yöntemlerden biridir. Benzer şekilde en iyi sıra değerine sahip olan ilk 5 modelden 2'sinin (A_4 ve A_{14} alternatifleri) yapısında KBO yönteminin yer alması bu yöntemin lenfoma veri kümesi için öznitelik seçiminde baskın bir önemi olduğunu göstermektedir. KBÖS ve KBF yöntemlerinin NB sınıflandırıcısı ile birlikte kullanıldığı modellerden iyi sıralama sonuçları elde edilmiş olsa da kNN sınıflandırıcısıyla aynı başarıya ulaşılamamıştır. AKKBÖS, BKBM ve KKBM yöntemleri ile oluşturulan modeller orta düzeylerde başarı göstermiştir. NKBÖS, DAÖS ve KKBÖS yöntemleri ise lenfoma verisi üzerinde diğer öznitelik seçim yöntemlerine nazaran daha düşük tahmin performansına yol açmıştır. Burada şunu not etmek gerekir; KKBÖS yönteminde negatif eşik seçimi orijinal öznitelik uzayının korunmasını sağlar, öte yandan negatif eşik fazlalıkların cezalandırılamamasına yol açar. Bu ise performans düşüşüne neden olmaktadır. Hem kolon hem de lenfoma verileri üzerinde bütünleşik AHS-VIKOR ile elde edilen değerlendirme sonuçları burada sözü edilen durumu teyit etmiştir.

5.2. VIKOR ile Çok Kriterli Öznitelik Seçimine Yönelik Deneysel Uygulama

Kolon ve lenfoma veri kümeleri için tahmin modellerinin uzlaşık sırasının ve en iyi modelleri gösteren uzlaşık çözüm kümesinin elde edilmesiyle birlikte tezin ilk aşamasıyla alâkalı deneyler tamamlanmıştır. Bundan sonraki deneysel süreç ikinci aşamayla ilgili olup, VIKOR yöntemi ile hibrit çok kriterli öznitelik seçiminin uygulanma süreci ve test edilme işlemlerinden oluşmaktadır.

5.2.1. Sınıflandırıcı yöntemler bazında en avantajlı olan öznitelik seçim kriterlerinin belirlenmesi

İkinci aşama çerçevesinde öncelikle kolon ve lenfoma gen ifadesi verileri için ilk aşamanın çıktıları, yani tahmin modellerinin uzlaşık sıralamaları analiz edilerek her bir sınıflandırma görevi için hangi öznitelik seçim kriterlerinin daha etkili olduğuna karar verilmiştir. Burada amaç her bir sınıflandırma görevi için VÇKÖS yöntemine girdi olarak gönderilecek kriterlerin belirlenmesine yardımcı olmaktır. Kriter tespitinin uzlaşık

sıralamalara dayandırılması, tez çalışmasının iki ana aşaması arasında kurulan faydacı bir bağlantıya işaret etmektedir.

Öznitelik seçiminin optimize edilmesi amaçlandığı için sınıflandırıcı bakımından ortak paydada buluşan tahmin modelleri dikkate alınarak her bir sınıflandırıcı altında kriterlerin başarı sırası ayrı ayrı belirlenebilir. Bu bağlamda kolon ve lenfoma verilerinin her biri için 20 adet alternatifin bütünleşik AHS-VIKOR ile değerlendirilmesi neticesinde elde edilen uzlaşık sıralama sonuçları (Bkz. Çizelge 5.5) ayrıştırılarak her bir sıralama listesi kNN ve NB sınıflandırıcılarına göre iki ayrı sıralı liste içinde toplanmıştır. Ayrıştırma işlemiyle kolon ve lenfoma verileri için elde edilen ikişer liste Çizelge 5.6'da gösterilmiştir. Bu çizelgede verilen sıralama listeleri temel alınarak, kolon ve lenfoma veri kümeleri üzerinde daha iyi bir öznitelik alt kümesi elde edebilmek için kNN ve NB sınıflayıcıları bazında en uygun olan 5'er kriter belirlenmiştir. Yani en iyi sıralı ilk 5 alternatif dikkate alınmış ve bunların yapısında yer alan öznitelik seçiciler belirlenip, alternatif sıralamasıyla uyumlu bir şekilde ele alınmıştır.

Çizelge 5.6. Uzlaşık sıralama sonuçlarının kNN ve NB sınıflandırıcılarına göre ayrılmış durumu

	kNN		NB	
Sıralama	Kolon	Lenfoma	Kolon	Lenfoma
1	A_8	A_2	A_{20}	A_{11}
2	A_{10}	A_4	A_{14}	A_{12}
3	A_1	A_{10}	A_{11}	A_{14}
4	A_6	A_9	A_{12}	A_{20}
5	A_2	A_6	A_{16}	A_{19}
6	A_4	A_1	A_{18}	A_{13}
7	A_7	A_3	A_{13}	A_{16}
8	A_9	A_8	A_{15}	A_{15}
9	A_3	A_5	A_{19}	A_{17}
10	A_5	A_7	A_{17}	A_{18}

Kolon problem verisi üzerinde kNN sınıflandırıcısı bazında en avantajlı kriterler; A_8, A_{10}, A_1, A_6 ve A_2 ile temsil edilen tahmin modellerinin yapısında yer alan DAÖS, AKKBÖS, KBÖS, KKBM ve D-KBÖS kriterleridir. NB sınıflayıcısı içinse A_{20}, A_{14}, $A_{11,}$ A_{12} ve A_{16} modellerinin yapısındaki AKKBÖS, KBO, KBÖS, D-KBÖS ve KKBM kriterleri avantajlı kriterler olarak belirlenmiştir. Lenfoma verisinin kNN sınıflandırıcısına göre analizinde öznitelik seçimi için sıralamanın en üstündeki 5 kriter seçilmiştir. Bunlar A_2, A_4, $A_{10,}$ A_9 ve A_6 modellerinin yapısında yer alan D-KBÖS, KBO, AKKBÖS, BKBM ve KKBM kriterleridir. Lenfoma verisinin NB sınıflandırıcısına göre analizinde ise KBÖS, D-KBÖS, KBO, AKKBÖS ve BKBM en avantajlı kriterler olarak belirlenmiştir. Bunlar NB sınıflandırıcısı bazında en iyi sıralı A_{11}, A_{12}, A_{14}, A_{20} ve A_{19} modellerinin bünyesinde temsil edilen öznitelik seçim kriterleridir. Belirlenen kriterler Çizelge 5.7'de sıralı olarak ifade edilmiştir. Bu bireysel kriterler çok kriterli öznitelik seçim yapısıyla betimlenen birleşimin üyesi olmaya aday gösterilmiştir.

Çizelge 5.7. Her bir veri kümesi için sınıflandırıcı bazında en başarılı öznitelik seçiciler

	kNN		NB	
Sıralama	Kolon	Lenfoma	Kolon	Lenfoma
1	DAÖS	D-KBÖS	AKKBÖS	KBÖS
2	AKKBÖS	KBO	KBO	D-KBÖS
3	KBÖS	AKKBÖS	KBÖS	KBO
4	KKBM	BKBM	D-KBÖS	AKKBÖS
5	D-KBÖS	KKBM	KKBM	BKBM

5.2.2. VÇKÖS yapısı içinde bir araya getirilmek üzere uygun kriterlerin belirlenmesi ve çoğunluk kuralı stratejisine uygun ağırlık değerinin atanması

Belirli bir veri kümesi üzerinde kNN ve NB sınıflandırıcıları bazında en başarılı performansı elde eden kriterlere bakılarak (Bkz. Çizelge 5.7), tahmin probleminin sınıflandırıcı bazında birbirinden ayrı olan çözümlerine yönelik uygun kriter setlerinin belirlenmesi mümkündür. Bu amaçla deneysel bir araştırma sürecine ihtiyaç duyulmuştur. Buna ek olarak, VÇKÖS tasarımında çoğunluk kuralına göre karar verme stratejisinin ağırlığını ifade eden v değerinin ayarlanması gereklidir. Bu doğrultuda farklı sayıda

kriterler ve çeşitli v değerlerinin denendiği karşılaştırmalı bir araştırma yapılmıştır. VÇKÖS yönteminde bilgi kuramı tabanlı birden çok yöntemin birleştirilmesi ve her öznitelik seçim adımında yeni bir karar matrisi üzerinde işlem yapılması sebebiyle kriter sayısı arttıkça hesaplama süresi artış göstermektedir. Bu nedenle deneylerde sınıflandırıcı bazında en başarılı öznitelik seçicilerden (Bkz. Çizelge 5.7) yalnızca ilk 2'si ve ilk 3'ü kriter seti oluşturmak üzere kullanılmıştır. Diğerlerinden farklı olarak, kolon verileri üzerinde NB sınıflandırıcısı ile gerçekleştirilen deneylerde 2 veya 3 kriterle hatasız sınıflandırma sonucuna ulaşılamadığı gerekçesiyle 4 ve 5 elemanlı kriter setleriyle de denemeler yapılmıştır.

VÇKÖS yöntemiyle gerçekleştirilen öznitelik seçim işleminde kriter setini oluşturan yöntemler için tezin birinci aşamasında belirlenen parametre ayarları değiştirilmeden kullanılmıştır. Dolayısıyla KBÖS ve D-KBÖS yöntemlerinde fazlalık teriminin göreceli önemini belirten β parametresinin değeri yine 0,5 olarak belirlenmiştir. kNN yönteminde k için 1 değeri kullanılmıştır. Öznitelik seçiminde kolon tümörü ve lenfoma verileri için sırasıyla 100 ve 200 seçili özniteliğe ulaşılması durma kriteri olarak belirlenmiştir.

Deneylerde öncelikle tez çalışmasının ilk aşamasındaki ön işlem adımlarının aynısı uygulanarak veriler analizlere hazır hâle getirilmiştir. Ön işlemlerden sonra kolon tümörü ve lenfoma verileri üzerinde tanımlı olan tahmin problemlerinin çözümü için VÇKÖS yöntemi kNN ve NB sınıflandırıcısı ile birlikte kullanılarak test edilmiştir. Buna göre her bir veri kümesi üzerinde kNN ve NB sınıflandırıcıları için öznitelik seçiminde VÇKÖS yönteminin 2 ve 3 elemanlı kriter setleriyle kullanımına dayanan çeşitli tahmin modelleri oluşturulmuştur. Öte yandan, kolon ve lenfoma verileri üzerinde sınıf tahmini yapmaya yönelik tanımlanan modeller kapsamında VÇKÖS yönteminin performansı test edilirken v ağırlığının çeşitli değerlerinin karşılaştırılması amacıyla deneylerin kapsamı daha da genişletilmiştir. VÇKÖS yönteminde v ağırlığına sırasıyla 0,5; 0,6 ve 0,7 değerleri atanarak performans sınaması yapılmıştır. Bu yolla elde edilen tahmin performansı 10-kat çapraz doğrulama kapsamında AUC ve doğruluk kriterlerine göre ölçülerek kolon ve lenfoma verileri üzerinde en iyi performansın hangi kriter seti ve v değeri ile elde edildiği tespit edilmiştir. Kolon verisi için elde edilen en iyi tahmin değerleri Çizelge 5.8'de gösterilmiş olup, birbirinden farklı kriter setleri ve çeşitli v değerlerine göre en başarılı sonuca ulaşılan durumlar ise koyu renkle vurgulanmıştır. Burada ayrıca her bir tahmin

sonucunun bitişiğinde # işareti ile belirtilen tamsayı değeri, o tahmin performansının elde edildiği seçili öznitelik sayısını ifade etmektedir.

Çizelge 5.8. Kolon tümörü verisi için VÇKÖS yönteminde birbirinden farklı kriter setleri ve çeşitli *v* değerleri kullanılarak elde edilen en yüksek tahmin performansları

kNN sınıflandırıcısı ile oluşturulan tahmin modellerinin performans sonuçları				
	VÇKÖS (DAÖS, AKKBÖS) & kNN		VÇKÖS (DAÖS, AKKBÖS, KBÖS) & kNN	
v	AUC	Doğruluk	AUC	Doğruluk
0,5	0,989 #28	0,984 #28	0,977 #39	0,984 #31
0,6	1,00 #30	0,984 #26	0,972 #37	0,968 #16
0,7	1,00 #32	1,00 #32	0,966 #15	0,952 #15
NB sınıflandırıcısı ile oluşturulan tahmin modellerinin performans sonuçları				
	VÇKÖS (AKKBÖS, KBO) & NB		VÇKÖS (AKKBÖS, KBO, KBÖS) & NB	
v	AUC	Doğruluk	AUC	Doğruluk
0,5	0,983 #24	0,935 #7	0,990 #36	0,952 #6
0,6	0,989 #10	0,935 #11	0,989 #19	0,952 #6
0,7	0,989 #10	0,935 #11	0,989 #19	0,952 #6
	VÇKÖS (AKKBÖS, KBO, KBÖS, D-KBÖS) & NB		VÇKÖS (AKKBÖS, KBO, KBÖS, D-KBÖS, KKBM) & NB	
v	AUC	Doğruluk	AUC	Doğruluk
0,5	0,990 #24	0,952 #9	0,988 #22	0,952 #13
0,6	0,984 #25	0,935 #8	0,988 #29	0,935 #6
0,7	0,989 #17	0,935 #8	0,985 #47	0,952 #7

VÇKÖS yönteminin kolon ve lenfoma verileri üzerinde KNN ve NB sınıflandırıcıları ile test edildiği deneylerde, çeşitli kriter setleri ve değişen *v* değerlerine göre en başarılı sonuçların belirlenmesi için öncelikle AUC ve doğruluk performanslarının yüksekliği dikkate alınmıştır. Eğer her iki ölçüm bakımından denklik gösteren birden çok en iyi

durum söz konusu ise o zaman 2 veya 3 kriterin, 4 veya 5 kritere göre hep daha öncelikli olduğu varsayılmıştır. Ayrıca, 2 veya 3 kriterli setler kendi içinde değerlendirilirken daha küçük sayıda seçili öznitelikle elde edilen sonuçların daha avantajlı olduğuna karar verilmiştir. Eğer her iki ölçüm bakımından denklik gösteren birden çok en iyi durum aynı sayıda öznitelikle elde edilmiş ise o zaman göz önünde bulundurulması gereken yine kriter setinin eleman sayısıdır. Çünkü daha az sayıda kriter daha düşük hesap yükü anlamına gelmektedir. Kıyaslamalar kolon verileri üzerinde VÇKÖS yönteminin kNN sınıflandırıcısı ile birlikte en iyi sonucu 2 elemanlı kriter seti için ve $v = 0,7$ durumunda elde ettiğini, NB sınıflandırıcısı ile birlikte en iyi sonucu 3 elemanlı kriter seti için ve $v = 0,5$ durumunda elde ettiğini göstermiştir. Aslında kolon verileri üzerinde VÇKÖS yönteminin 4 elemanlı kriter seti ve $v = 0,5$ değeri kullanılarak NB sınıflandırıcısı ile birlikte uygulandığı durum da performans ölçümleri ve seçili öznitelik sayıları bakımından uygundur, ancak hesaplama verimliliği için 4 yerine 3 elemanlı kriter seti tercih edilmiştir. Kolon tümörü verisi için bundan sonraki deneysel süreçte VÇKÖS yöntemi burada belirlenen kriterler ve v değerleri kullanılarak uygulanmıştır. kNN yöntemiyle gerçekleştirilen deneyler için kriterler olarak DAÖS ve AKKBÖS; NB yöntemiyle gerçekleştirilen deneyler için kriterler olarak AKKBÖS, KBO ve KBÖS esas alınmıştır. Ayrıca kNN yöntemiyle gerçekleştirilen deneyler için v değeri 0,7; NB yöntemi içinse v değeri 0,5 olarak alınmıştır.

Kolon tümörü verisi için izlenen deneysel prosedürün aynısı lenfoma verileri için uygulanarak çeşitli kriterler ve v ağırlık değerlerine göre elde edilen en iyi tahmin sonuçları belirlenmiştir. Bu sonuçlar elde edildikleri seçili öznitelik sayıları ile birlikte Çizelge 5.9'da gösterilmiştir. En iyi ilk 2 ve en iyi ilk 3 kriterden oluşan kriter setleri ile birlikte v değişkeni için 3 ayrı durum değeri denenerek elde edilen sonuçlar arasından en başarılı olanlar koyu renk tonuyla belirtilmiştir.

Lenfoma verisi üzerinde VÇKÖS yöntemi 3 elemanlı kriter setiyle kullanıldığı zaman v değişkenine 0,5; 0,6 veya 0,7 değerleri atandığı durumların tümünde kNN sınıflandırıcısı altında hatasız tahmin performansına ulaşılmıştır. Ayrıca VÇKÖS yönteminde 2 elemanlı kriter seti kullanıldığında v değerinin 0,5 veya 0,7 seçilmesi de kNN sınıflandırıcısı bazında hatasız tahmin performansına ulaşılmasını sağlamıştır. Ancak 3 elemanlı kriter seti ve $v = 0,7$ değeri kullanıldığında aynı sonuç daha az sayıda öznitelikle elde edilmiştir. Benzer şekilde VÇKÖS yöntemi çerçevesinde 2 ve 3 elemanlı kriter setleri ve v

değişkeninin hemen hemen tüm değerlerinde NB sınıflandırıcısı için AUC ve doğruluk metrikleri altında en iyi tahmin performansına ulaşılmıştır. Ancak 3 elemanlı kriter seti ve v = 0,7 değerinde bu sonuca daha az sayıda öznitelikle ulaşmanın mümkün olduğu gözlenmiştir. Sonuç olarak lenfoma verisi üzerinde VÇKÖS yöntemiyle kNN ve NB sınıflandırıcıları için en iyi tahmin performansı, 3 elemanlı kriter setleriyle ve ağırlık değerinin v = 0,7 olduğu durumda elde edilmiştir. Bu nedenle lenfoma verisi üzerinde kNN yöntemiyle gerçekleştirilen deneyler için kriterler olarak D-KBÖS, KBO ve AKKBÖS; NB yöntemiyle gerçekleştirilen deneyler için kriterler olarak KBÖS, D-KBÖS ve KBO esas alınmıştır. Ayrıca lenfoma verisi için hem kNN hem de NB yöntemiyle gerçekleştirilen deneylerde v değeri 0,7 olarak alınmıştır.

Çizelge 5.9. Lenfoma verisi için VÇKÖS yönteminde birbirinden farklı kriter setleri ve çeşitli v değerleri kullanılarak elde edilen en yüksek tahmin performansları

kNN sınıflandırıcısı ile oluşturulan tahmin modellerinin performans sonuçları				
	VÇKÖS (D-KBÖS, KBO) & kNN		VÇKÖS (D-KBÖS, KBO, AKKBÖS) & kNN	
v	AUC	Doğruluk	AUC	Doğruluk
0,5	1,00 #47	1,00 #47	1,00 #101	1,00 #101
0,6	0,997 #160	0,990 #36	1,00 #36	1,00 #36
0,7	1,00 #141	1,00 #141	1,00 #34	1,00 #34
NB sınıflandırıcısı ile oluşturulan tahmin modellerinin performans sonuçları				
	VÇKÖS (KBÖS, D-KBÖS) & NB		VÇKÖS (KBÖS, D-KBÖS, KBO) & NB	
v	AUC	Doğruluk	AUC	Doğruluk
0,5	1,00 #41	1,00 #41	1,00 #93	1,00 #93
0,6	1,00 #31	1,00 #31	1,00 #34	0,990 #20
0,7	1,00 #56	1,00 #56	1,00 #27	1,00 #27

Deneysel bulgulara göre, hatasız sınıflandırma sonucu elde etmek için genellikle en iyi ilk 2 veya 3 kriterden yararlanmak yeterli olmuştur. 2 veya 3 kriterle hem tahmin performansı hem de seçili öznitelik sayısı bakımından etkili sonuçlara ulaşılmıştır. Deneysel sürecin

buraya kadar olan kısmı neticesinde; kolon verisi üzerinde kNN yöntemiyle, kolon verisi üzerinde NB yöntemiyle, lenfoma verisi üzerinde kNN yöntemiyle ve lenfoma verisi üzerinde NB yöntemiyle gerçekleştirilecek olan 4 farklı sınıflandırma çözümüne yönelik kullanılmak istenen VÇKÖS yöntemleri tüm yönleriyle tarif edilmiştir. Tahmin süreçlerinde VÇKÖS yönteminin tasarımında kullanılması gereken kriter setleri ve v ağırlık değerleri artık belirli durumdadır. Bu çerçevede her bir gen ifadesi verisi için öznitelik seçiminde VÇKÖS yöntemi temel alınarak oluşturulan tahmin modelleri Çizelge 5.10'da gösterilmiş olup, her bir model için VÇKÖS yapısı içinde kullanılan çoklu kriterler ile v değişkenine atanan değerler parantez içlerinde belirtilmiştir. Ayrıca her bir modelle ilişkili olan VÇKÖS yöntemi, kriter seti ve v değeri bakımından ötekilerden farklı bir öznitelik seçim yapısını temsil ettiği için bu yapılar 1'den 4'e kadar numaralandırılmış olup, böylece aralarında ayrım yapılabilmesi amaçlanmıştır.

Çizelge 5.10. VÇKÖS yöntemi kullanılarak oluşturulan tahmin modelleri

	Tahmin Modeli
Veri Kümesi	VÇKÖS (Çoklu kriterler; v değeri) & Sınıflandırıcı
Kolon	VÇKÖS1 (DAÖS, AKKBÖS; v = 0,7) & kNN
Kolon	VÇKÖS2 (AKKBÖS, KBO, KBÖS; v = 0,5) & NB
Lenfoma	VÇKÖS3 (D-KBÖS, KBO, AKKBÖS; v = 0,7) & kNN
Lenfoma	VÇKÖS4 (KBÖS, D-KBÖS, KBO; v = 0,7) & NB

5.2.3. VÇKÖS yönteminin kıyaslamalı olarak değerlendirilmesi

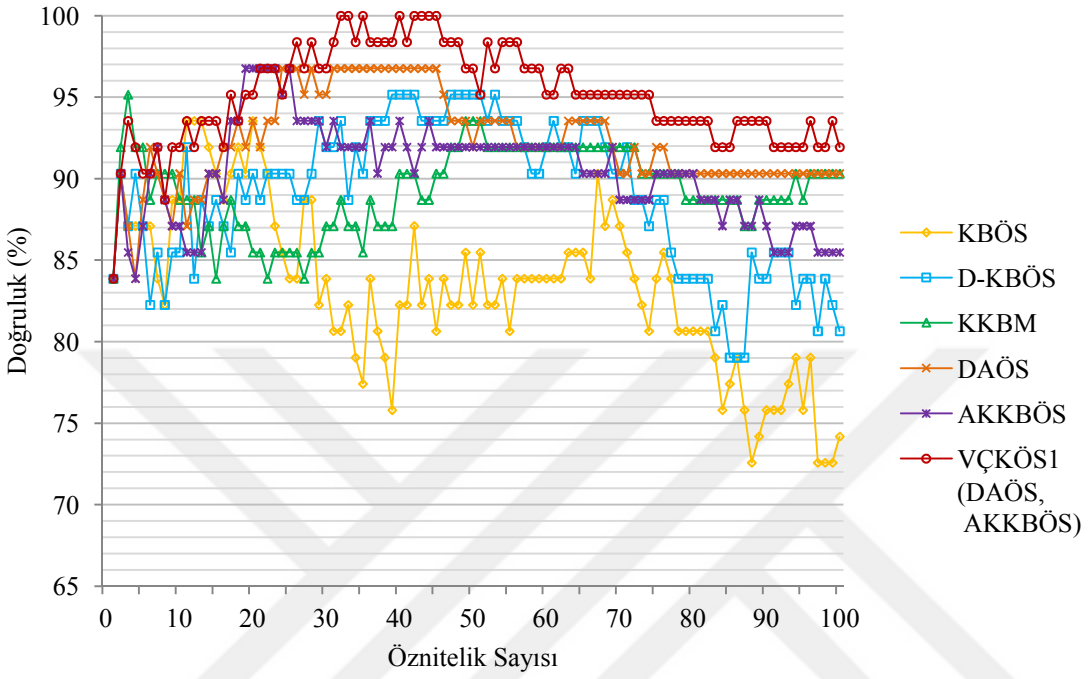
Tahmin performansını geliştirmede VÇKÖS yönteminin ne ölçüde etkili olduğunun görülebilmesi için kıyaslamalı bir deneysel çalışma gerçekleştirilmiştir. VÇKÖS yöntemi kNN ve NB sınıflandırıcıları temelinde KBÖS, D-KBÖS, KBF, KBO, NKBÖS, KKBM, KKBÖS, DAÖS, BKBM ve AKKBÖS olmak üzere bilgi kuramı tabanlı diğer 10 öznitelik seçim yöntemi ile mukayese edilmiştir. Kıyaslamalar hem ortalama performans değerlerine hem de en iyi performans değerlerine göre yapılmıştır. Tüm deneysel karşılaştırma süreçlerinde seçili öznitelik sayısı kolon verisi için en fazla 100, lenfoma verisi için en fazla 200 olarak belirlenmiştir. Daha önceki deneylerde olduğu gibi, ön işlemden

geçirilmiş veriler kullanılmıştır. kNN yönteminde en yakın komşu sayısını ifade eden k değeri 1 seçilmiştir. KBÖS ve D-KBÖS yöntemlerinde β parametresine 0,5 değeri, KKBÖS yönteminde δ parametresine -1 değeri atanmıştır. Tahmin sonuçlarının elde edilmesinde 10-kat çapraz doğrulama kullanılmıştır.

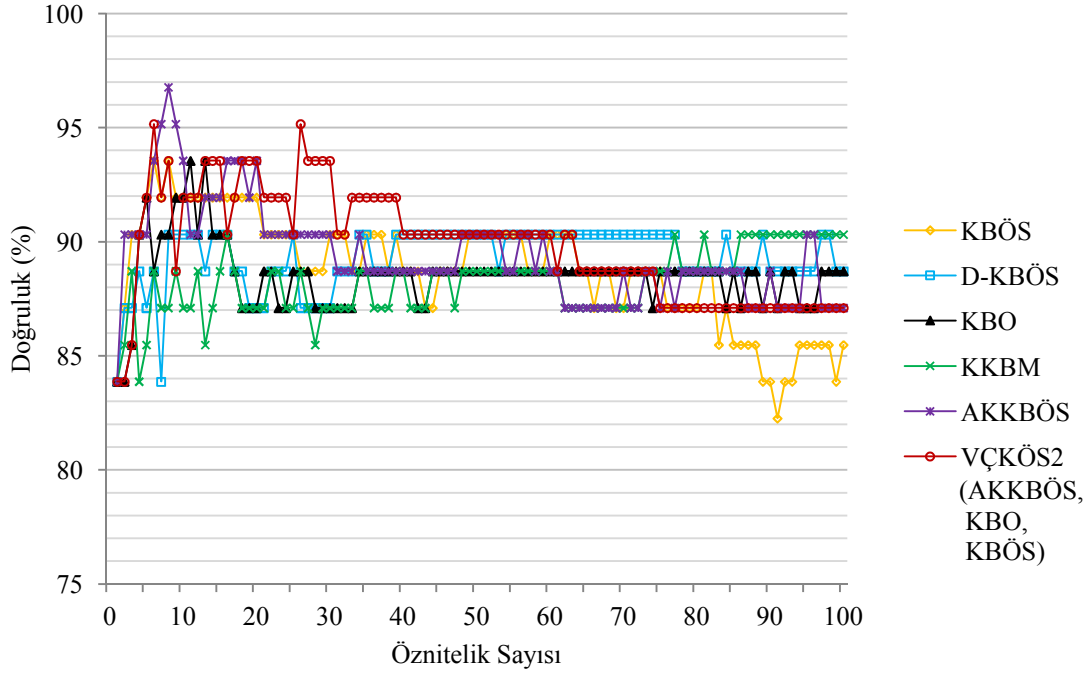
Deneyler, 2 veri kümesi üzerinde 2 sınıflandırıcı yöntemle uygulanmak istenen 4 farklı sınıflandırma çözümünün her biri için ötekilerden farklı bir kriter seti kullanılarak tasarımlanan VÇKÖS yapıları ile gerçekleştirilmiştir. VÇKÖS yönteminin 4 farklı sınıflandırma çözümünün her birindeki başarısı, ilgili sınıflandırmada en iyi performansları sergileyen diğer öznitelik seçim yöntemleri ile (Bkz. Çizelge 5.7) grafiksel olarak karşılaştırılmıştır. Grafiklerde her bir yöntemin 1 ila durma kriteri olarak belirlenen sınır değer arasındaki tüm öznitelik sayılarında elde ettiği performans değerleri gösterilmiştir. Performans sonuçları tüm grafiklerde yüzde değerleri olarak ifade edilmiştir. VÇKÖS yöntemi 4 farklı sınıflandırma çözümünün her biri için diğer öznitelik seçim yöntemleriyle ortalama ve en iyi performans değerleri bakımından karşılaştırılmıştır. Kolon tümörü verisi ele alınırken seçili öznitelikler içinde sırasıyla ilk 30, 50 ve 100 öznitelik için elde edilen ortalama performans değerleri karşılaştırılmıştır. Lenfoma verisi ele alınırken seçili öznitelikler içinde sırasıyla ilk 30, 50, 100, 150 ve 200 öznitelik için elde edilen ortalama performans değerleri karşılaştırılmıştır. Bu süreçte performans ölçümünde oldukça sık kullanılan AUC ve doğruluk metriklerinden yararlanılmıştır. Benzer şekilde; VÇKÖS yönteminin kNN ve NB yöntemleriyle birlikte elde ettiği en iyi performans değerleri ile diğer öznitelik seçim yöntemlerinin en iyi performans sonuçları AUC ve doğruluk olmak üzere 2 ayrı metrik altında mukayese edilmiştir. Bu çerçevede ortalama ve en iyi performans değerleri çizelge formatında ifade edilmiştir. Kıyaslamalara ilişkin ölçüm sonuçlarının gösterildiği tüm çizelgelerde en alt satır VÇKÖS yönteminin diğer yöntemlere karşı elde ettiği Kazanma/Beraberlik/Yenilgi (K/B/Y) sayılarını ifade etmektedir.

İlk olarak kolon problem verisi üzerinde deneysel karşılaştırmalar gerçekleştirilmiştir. Kolon verileri üzerinde kNN ve NB sınıflandırıcıları ile birlikte en iyi performans sonuçlarına ulaştığı tespit edilen 5'er tane öznitelik seçim yöntemi (Bkz. Çizelge 5.7) ile VÇKÖS yöntemi elde ettikleri doğruluk değerleri bakımından grafik üstünde karşılaştırılmıştır. Grafiksel gösterimlerde 4 farklı VÇKÖS yapısı arasında ayrım sağlanabilmesi için bunlar daha önce kendilerine atanan numaraları ile (Bkz. Çizelge 5.10) ifade edilmiştir. Ayrıca, ilgili VÇKÖS yapısında kriter setini oluşturan çoklu kriterler

parantez içinde de belirtilmiştir. Kolon tümörü verisi için her bir öznitelik seçim yöntemi ile seçilen 1 ila 100 arası özniteliğin kNN ve NB yöntemleriyle birlikte sergilediği doğruluk performansları sırasıyla Şekil 5.1'de ve Şekil 5.2'de gösterilmiştir.



Şekil 5.1. Kolon tümörü verisi için öznitelik seçicilerin kNN yöntemi ile birlikte elde ettiği doğruluk değerleri



Şekil 5.2. Kolon tümörü verisi için öznitelik seçicilerin NB yöntemi ile birlikte elde ettiği doğruluk değerleri

Kolon tümörü verisi üzerinden kNN ve NB yöntemlerine göre çizilen performans grafiklerinin her ikisi de VÇKÖS yönteminin çoğu durumda karşılaştırılabilir derecede iyi veya en iyi doğruluk değerlerine ulaştığını göstermektedir. VÇKÖS yöntemi özellikle kNN yöntemi ile birlikte kullanıldığı durumda oldukça iyi performans sergilemiştir. Grafiksel olarak verilen bu sonuçların dikkat çeken bir diğer yanı; öznitelik seçicilerin NB sınıflandırıcısı ile birlikte kullanıldıklarında genel anlamda %95 sınırının altında kalan doğruluk değerleri elde etmesidir (Bkz. Şekil 5.2). Buna karşın kNN sınıflandırıcısı ile çoğu kez bu sınırı aşan performans değerlerine ulaşıldığı gözlenmiştir (Bkz. Şekil 5.1).

Seçili öznitelik kümesinin eleman sayısı için üst sınır *N* sırasıyla 30, 50 ve 100 seçilmek üzere, kolon verisi üzerinde öznitelik seçicilerin 1 ila #*N* adet arası elemana sahip seçili alt kümelerden kNN ile elde ettiği performansların ortalamaları Çizelge 5.11'de gösterilmiştir.

Çizelge 5.11. Kolon tümörü verisi üzerinde öznitelik seçicilerin kNN sınıflandırıcısı ile birlikte sergilediği ortalama performans değerleri

Öznitelik Seçici	Ortalama AUC			Ortalama Doğruluk		
	#30	#50	#100	#30	#50	#100
KBÖS	0,9084	0,8766	0,8484	0,8833	0,8568	0,8332
D-KBÖS	0,8873	0,9137	0,9114	0,8817	0,9035	0,8894
KBF	0,8634	0,8785	0,8892	0,8742	0,8803	0,8744
KBO	0,8875	0,8989	0,8991	0,8731	0,8823	0,8748
NKBÖS	0,8560	0,8696	0,8836	0,8516	0,8603	0,8615
KKBM	0,8805	0,8980	0,9160	0,8753	0,8823	0,8935
KKBÖS	0,8783	0,8907	0,8861	0,8694	0,8742	0,8765
DAÖS	0,9304	0,9365	0,9356	0,9140	0,9323	0,9231
BKBM	0,8604	0,8566	0,8817	0,8656	0,8568	0,8682
AKKBÖS	0,9066	0,9166	0,9106	0,9102	0,9145	0,9039
VÇKÖS1 (DAÖS, AKKBÖS)	0,9436*	0,9630*	0,9593*	0,9371*	0,9577*	0,9510*
K/B/Y	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0

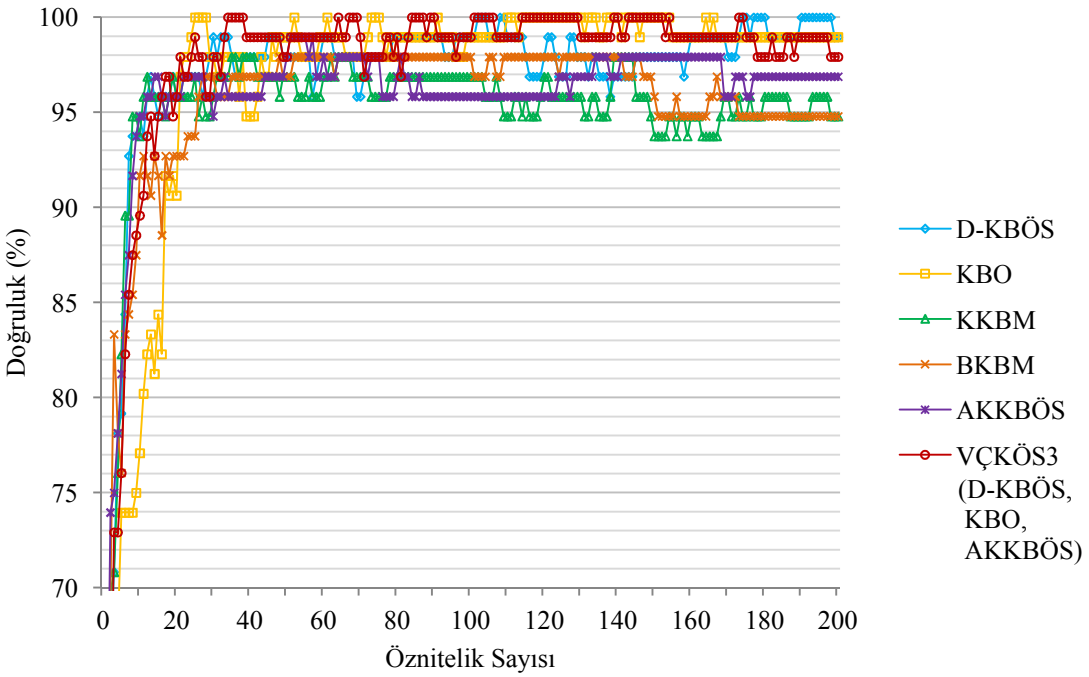
Burada sınıflandırma kolon verisi üzerinde kNN ile yapıldığı için VÇKÖS yönteminin ölçüm sonuçları VÇKÖS1 yapısı esas alınarak elde edilmiştir. Ortalama performans mukayesesi kapsamında yukarıda verilen çizelgede, tüm yöntemler arasındaki en iyi değerler sağ üst köşelerine iliştirilen asteriks $^{(*)}$ işareti ile belirtilmiştir. Bunun yanı sıra VÇKÖS yönteminin diğer yöntemlere kıyasla elde ettiği kazanma, beraberlik ve yenilgi durumlarına ilişkin sayısal bilgiler çizelgenin en altındaki K/B/Y satırında verilmiştir. Kolon tümörü verisi üzerinde kNN sınıflandırıcısının VÇKÖS yöntemi ile birlikte kullanılması sonucu elde edilen ortalama performans değerlerinin tamamı, 11 öznitelik seçim yöntemi ile elde edilen sonuçlar arasındaki en yüksek değerleri temsil etmektedir. Kolon verileri üzerinde öznitelik seçicilerin 30, 50 ve 100 öznitelik için NB sınıflandırıcısı ile birlikte elde ettiği ortalama performans değerleri Çizelge 5.12'de gösterilmiş olup, tüm yöntemler arasındaki en başarılı sonuçlar asteriks $^{(*)}$ işareti kullanılarak vurgulanmıştır.

Çizelge 5.12. Kolon tümörü verisi üzerinde öznitelik seçicilerin NB sınıflandırıcısı ile birlikte sergilediği ortalama performans değerleri

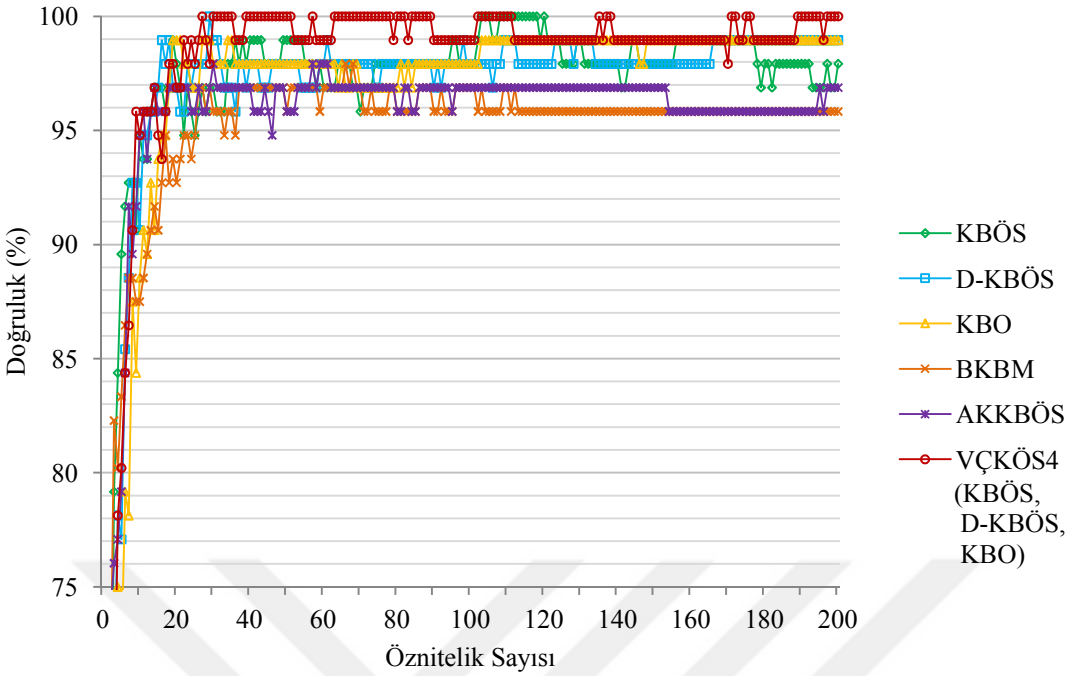
Öznitelik Seçici	Ortalama AUC			Ortalama Doğruluk		
	#30	#50	#100	#30	#50	#100
KBÖS	0,9583	0,9475	0,9335	0,9075	0,9013	0,8877
D-KBÖS	0,9609	0,9678	0,9725	0,8828	0,8890	0,8929
KBF	0,9613	0,9671	0,9706	0,8849	0,8855	0,8823
KBO	0,9630	0,9687	0,9714	0,8887	0,8865	0,8855
NKBÖS	0,9570	0,9621	0,9649	0,8785	0,8806	0,8835
KKBM	0,9491	0,9619	0,9712	0,8726	0,8752	0,8821
KKBÖS	0,8803	0,8922	0,9102	0,8634	0,8687	0,8711
DAÖS	0,9515	0,9583	0,9595	0,9113	0,9097	0,8973
BKBM	0,9206	0,9382	0,9518	0,8806	0,8777	0,8781
AKKBÖS	0,9662*	0,9719	0,9751	0,9151	0,9052	0,8937
VÇKÖS2 (KBÖS, AKKBÖS, KBO)	0,9652	0,9726*	0,9761*	0,9172*	0,9139*	0,8982*
K/B/Y	9/0/1	10/0/0	10/0/0	10/0/0	10/0/0	10/0/0

Burada sınıflandırma çözümü kolon verisi ve NB yöntemi ile ilintili olduğu için VÇKÖS yöntemi VÇKÖS2 yapısı esas alınarak uygulanmıştır. VÇKÖS yönteminin diğer yöntemler karşısında elde ettiği K/B/Y durumlarının sayısı çizelgenin en alt satırında belirtilmiştir. K/B/Y değerlerinden de anlaşıldığı üzere; kolon verisi üzerinde NB sınıflandırıcısına göre elde edilen ortalama AUC ve ortalama doğruluk değerleri içindeki en yüksek 6 sonuçtan 5 tanesi VÇKÖS yönteminin sonucudur.

Lenfoma çok dengesiz bir veri kümesi olduğundan dolayı seçili alt kümelerin performansları doğruluk ölçümünün yanı sıra bir de AUC ölçümüne göre ele alınmıştır. Lenfoma verisi üzerinde kNN ve NB sınıflandırıcıları ile birlikte en iyi performans sonuçlarına ulaştığı tespit edilen öznitelik seçim yöntemlerinden 5'er tanesi (Bkz. Çizelge 5.7) ile VÇKÖS yöntemi doğruluk ve AUC ölçüleri bakımından grafik üstünde karşılaştırılmıştır. Öznitelik seçim yöntemleri ile üretilen ve eleman sayısı 1'den 200'e kadar olan seçili öznitelik alt kümelerinin kNN ve NB sınıflandırıcılarıyla birlikte kullanılması sonucu elde edilen doğruluk değerleri sırasıyla Şekil 5.3'te ve Şekil 5.4'te gösterilmiştir. Performans değerlerinin elde edilmesiyle ilgili süreçte VÇKÖS yöntemi lenfoma verisi üzerinde kNN sınıflayıcısı ile birlikte kullanılacağında VÇKÖS3 yapısı, NB sınıflayıcısı ile birlikte kullanılacağında ise VÇKÖS4 yapısı esas alınmıştır.

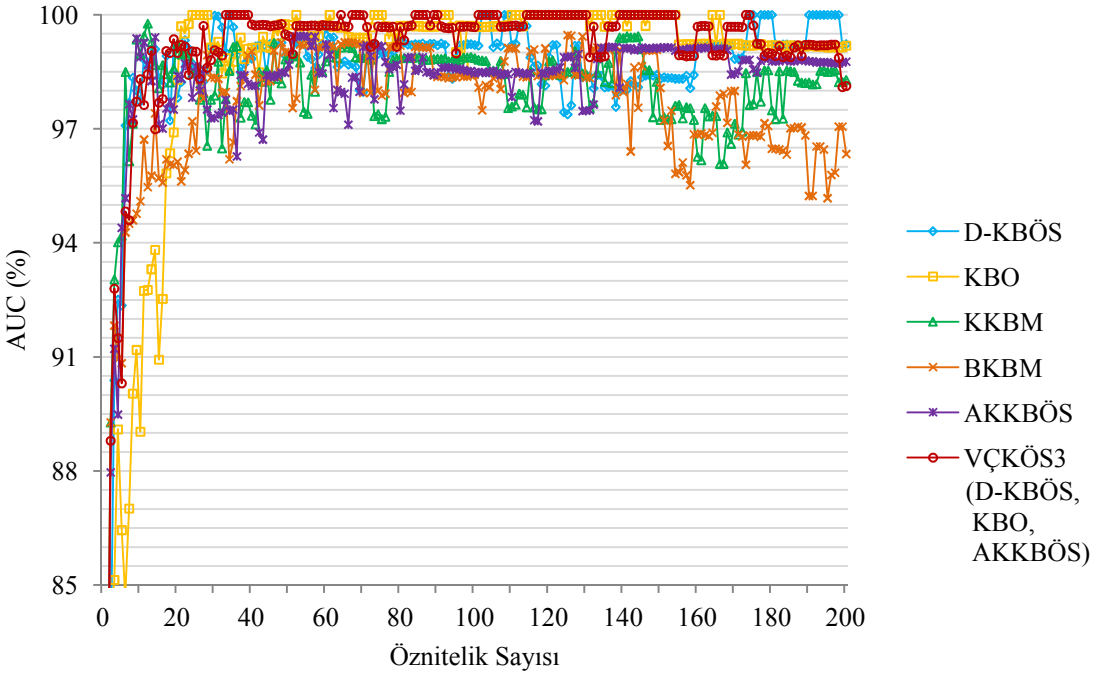


Şekil 5.3. Lenfoma verisi için öznitelik seçicilerin kNN yöntemi ile birlikte elde ettiği doğruluk değerleri

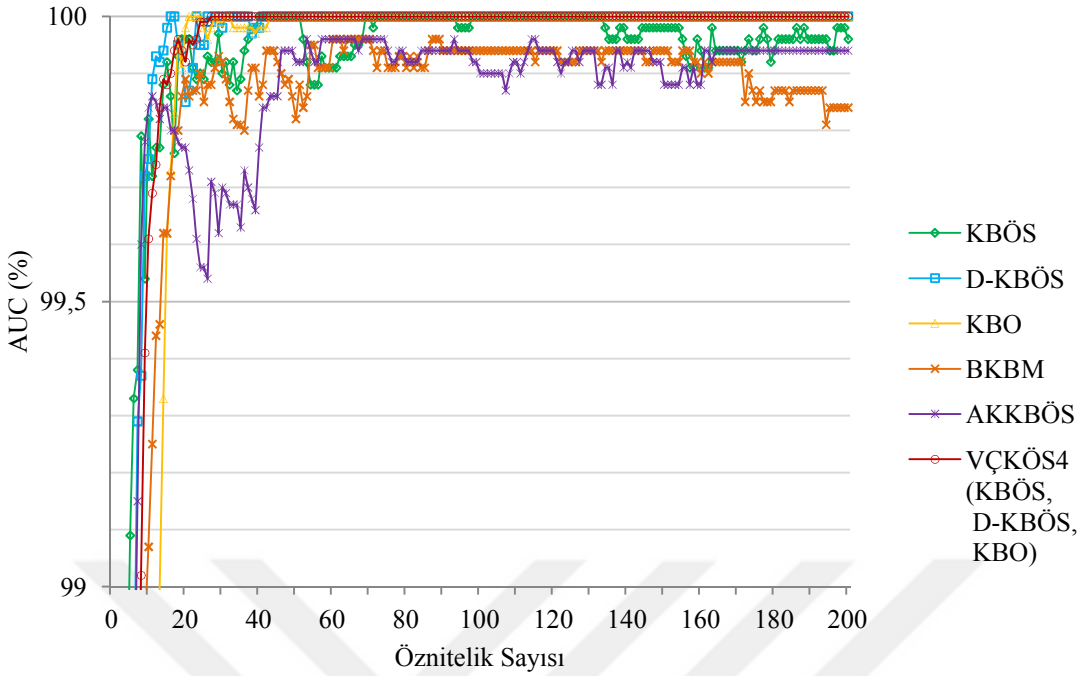


Şekil 5.4. Lenfoma verisi için öznitelik seçicilerin NB yöntemi ile birlikte elde ettiği doğruluk değerleri

Lenfoma verisi üzerinde öznitelik seçicilerin 1 ila 200 arası öznitelik seçerek kNN ve NB yöntemleriyle elde ettiği AUC değerleri sırasıyla Şekil 5.5'te ve Şekil 5.6'da gösterilmiştir.



Şekil 5.5. Lenfoma verisi için öznitelik seçicilerin kNN yöntemi ile birlikte elde ettiği AUC değerleri



Şekil 5.6. Lenfoma verisi için öznitelik seçicilerin NB yöntemi ile birlikte elde ettiği AUC değerleri

Doğruluk ve AUC ölçümlerine ilişkin grafiksel sonuçlar, lenfoma veri kümesi üzerinde VÇKÖS yöntemiyle çoğu durumda oldukça iyi tahmin performansları elde edildiğini göstermiştir. Özellikle NB yöntemiyle gerçekleştirilen sınıflandırmada VÇKÖS yöntemi ile seçilen 27. öznitelikten itibaren artırımsal olarak elde edilen tüm seçili alt kümelerde AUC performansının %100'ü bulması dikkat çekmektedir (Bkz. Şekil 5.6).

Lenfoma verisi üzerinde; bilgi kuramı tabanlı öznitelik seçim yöntemlerince elde edilen 1 ile *#N* adet arası seçili özniteliğe sahip alt kümeler kNN sınıflandırıcısı ile birlikte kullanılıp performanslarının ortalama değeri belirlenmiş ve bu değerler Çizelge 5.13'te gösterilmiştir. Burada ortalama performans ölçümü *N* için sırasıyla 30, 50, 100, 150 ve 200 değerleri denenerek yapılmıştır. Ölçümler lenfoma verisi üzerinde kNN yöntemine göre gerçekleştirildiği için VÇKÖS formülasyonu VÇKÖS3 yapısı esas alınarak uygulanmıştır. Dolayısıyla D-KBÖS, KBO ve AKKBÖS kriterlerini kapsayan kriter seti kullanılmıştır. Kıyaslama durumlarının her biriyle ilgili sonuçlar incelenip, bunlar içindeki en yüksek değerler asteriks $^{(*)}$ ile işaretlenmiştir. Lenfoma verilerinin sınıfının tahmin edilmesinde, VÇKÖS yönteminin kNN sınıflayıcısı ile birlikte elde ettiği ortalama AUC ve ortalama doğruluk değerlerinde seçili özniteliklerin sayısının artmasına bağlı olarak bir artış gözlemlenmiştir. Bu durum K/B/Y kıyaslamalarının sonuçlarını olumlu yönde etkilemiştir.

Çizelge 5.13. Lenfoma verisi üzerinde öznitelik seçicilerin kNN sınıflandırıcısı ile birlikte sergilediği ortalama performans değerleri

Öznitelik Seçici	Ortalama AUC					Ortalama Doğruluk				
	#30	#50	#100	#150	#200	#30	#50	#100	#150	#200
KBÖS	0,9690	0,9770	0,9800	0,9772	0,9741	0,9101	0,9325	0,9360	0,9327	0,9299
D-KBÖS	0,9656	0,9759	0,9831	0,9843	0,9864	0,9132	0,9406*	0,9614	0,9676	0,9732
KBF	0,9575	0,9684	0,9750	0,9780	0,9790	0,9101	0,9304	0,9472	0,9536	0,9566
KBO	0,9346	0,9574	0,9769	0,9843	0,9864	0,8465	0,8969	0,9422	0,9604	0,9680
NKBÖS	0,9454	0,9577	0,9652	0,9676	0,9713	0,8625	0,8994	0,9268	0,9335	0,9409
KKBM	0,9691*	0,9742	0,9804	0,9817	0,9805	0,9118	0,9356	0,9516	0,9540	0,9528
KKBÖS	0,9210	0,9289	0,9434	0,9492	0,9536	0,8417	0,8681	0,8988	0,9099	0,9158
DAÖS	0,9570	0,9611	0,9619	0,9511	0,9351	0,9146*	0,9308	0,9330	0,9224	0,9058
BKBM	0,9489	0,9623	0,9746	0,9784	0,9755	0,8885	0,9200	0,9494	0,9583	0,9562
AKKBÖS	0,9632	0,9695	0,9779	0,9805	0,9826	0,9139	0,9335	0,9518	0,9567	0,9604
VÇKÖS3 (D-KBÖS, KBO, AKKBÖS)	0,9653	0,9781*	0,9875*	0,9912*	0,9915*	0,8983	0,9348	0,9620*	0,9734*	0,9771*
K/B/Y	7/0/3	10/0/0	10/0/0	10/0/0	10/0/0	4/0/6	8/0/2	10/0/0	10/0/0	10/0/0

Lenfoma verisi üzerinde; bilgi kuramı tabanlı öznitelik seçim yöntemlerinin NB sınıflandırıcısı ile bir arada uygulanması sonucu elde edilen ortalama performans değerleri Çizelge 5.14'te gösterilmiştir. Burada performans ölçümleri lenfoma verisi üzerinde ve NB yöntemine göre gerçekleştirildiği için VÇKÖS yöntemi VÇKÖS4 yapısı esas alınarak uygulanmıştır. Kıyaslamalarda performans ortalamaları alınan seçili öznitelik kümelerinin eleman sayısı için alt sınır her zaman 1, üst sınır *N* sırasıyla 30, 50, 100, 150 ve 200 öznitelik olarak alınmıştır. Kıyaslama durumlarının her biri için (#*N*), verilen sonuçlar içindeki en yüksek değer belirlenip asteriks (*) ile işaretlenmiştir. Çizelgenin en altında VÇKÖS ile diğer yöntemlerin kıyaslama sonuçlarını yansıtan K/B/Y skorları verilmiştir.

Lenfoma verisi için, kNN'de olduğu gibi NB sınıflandırıcısı bazında elde edilen ortalama performans sonuçlarına bakıldığında da, seçili öznitelik sayısının artmasına bağlı olarak VÇKÖS yönteminin ortalama başarısının yükseldiği fark edilmektedir. Aynı zamanda

öznitelik sayısı arttıkça VÇKÖS yönteminin diğer yöntemlere kıyasla ortalama performans değerleri bakımından yenilgiye düştüğü durum sayısı giderek azalmıştır.

Çizelge 5.14. Lenfoma verisi üzerinde öznitelik seçicilerin NB sınıflandırıcısı ile birlikte sergilediği ortalama performans değerleri

Öznitelik Seçici	Ortalama AUC					Ortalama Doğruluk				
	#30	#50	#100	#150	#200	#30	#50	#100	#150	#200
KBÖS	0,9856*	0,9913*	0,9955*	0,9970*	0,9976*	0,9205*	0,9444	0,9611	0,9699	0,9730
D-KBÖS	0,9805	0,9883	0,9941	0,9961	0,9971	0,9146	0,9387	0,9582	0,9661	0,9712
KBF	0,9811	0,9882	0,9938	0,9956	0,9963	0,9010	0,9219	0,9371	0,9430	0,9432
KBO	0,9706	0,9823	0,9911	0,9941	0,9956	0,8920	0,9273	0,9516	0,9640	0,9704
NKBÖS	0,9794	0,9869	0,9928	0,9947	0,9955	0,8503	0,8735	0,8968	0,9013	0,9060
KKBM	0,9834	0,9893	0,9939	0,9956	0,9966	0,9059	0,9235	0,9461	0,9537	0,9567
KKBÖS	0,9659	0,9732	0,9832	0,9862	0,9882	0,8285	0,8531	0,8865	0,8985	0,9047
DAÖS	0,9770	0,9767	0,9663	0,9540	0,9438	0,8750	0,8767	0,8730	0,8474	0,8231
BKBM	0,9800	0,9875	0,9934	0,9954	0,9963	0,8948	0,9225	0,9440	0,9491	0,9514
AKKBÖS	0,9822	0,9885	0,9940	0,9957	0,9966	0,9142	0,9346	0,9514	0,9572	0,9579
VÇKÖS4 (KBO, KBÖS, D-KBÖS)	0,9810	0,9886	0,9943	0,9962	0,9971	0,9170	0,9496*	0,9724*	0,9790*	0,9824*
K/B/Y	6/0/4	8/0/2	9/0/1	9/0/1	8/1/1	9/0/1	10/0/0	10/0/0	10/0/0	10/0/0

Deneysel çalışma kapsamında son olarak öznitelik seçim yöntemlerinin en iyi performans değerlerinin karşılaştırması yapılmıştır. Bu kapsamda her bir öznitelik seçim yönteminin kolon ve lenfoma verileri üzerinde kNN sınıflandırıcısı ile birlikte elde ettiği en iyi performans değeri belirlenip Çizelge 5.15'te gösterilmiştir. Çizelgede verilen her bir değerin bitişiğinde o değerin elde edildiği seçili öznitelik sayısı # ile belirtilmiş ve 11 ayrı yönteme göre elde edilen sonuçlar içinden en iyi olanları göstermek için koyu renkle vurgulama yapılmıştır. Kolon verisi üzerinde uygulanan VÇKÖS yönteminde DAÖS ile AKKBÖS'ün kriterler olarak kullanıldığı VÇKÖS1 yapısı, lenfoma verisi üzerinde uygulanan VÇKÖS yönteminde ise D-KBÖS, KBO ve AKKBÖS'ün kriterler olarak kullanıldığı VÇKÖS3 yapısı esas alınmıştır (Bkz. Çizelge 5.10).

Çizelge 5.15. Kolon tümörü ve lenfoma verileri üzerinde öznitelik seçicilerin kNN sınıflandırıcısı ile birlikte sergilediği en iyi performans değerleri

Öznitelik Seçici	Kolon		Öznitelik Seçici	Lenfoma	
	AUC	Doğruluk		AUC	Doğruluk
KBÖS	0,9739 #12	0,9355 #11	KBÖS	0,9990 #7	0,9896 #44
D-KBÖS	0,9739 #48	0,9516 #39	D-KBÖS	1,00 #102	1,00 #102
KBF	0,9364 #49	0,9194 #4	KBF	0,9915 #109	0,9792 #25
KBO	0,9585 #11	0,9355 #10	KBO	1,00 #25	1,00 #25
NKBÖS	0,9375 #59	0,9032 #2	NKBÖS	0,9887 #186	0,9688 #79
KKBM	0,9784 #50	0,9516 #3	KKBM	0,9977 #12	0,9792 #35
KKBÖS	0,9540 #52	0,9194 #2	KKBÖS	0,9813 #168	0,9479 #70
DAÖS	0,9886 #25	0,9677 #24	DAÖS	0,9910 #22	0,9792 #22
BKBM	0,9420 #79	0,9194 #2	BKBM	0,9947 #125	0,9792 #48
AKKBÖS	0,9773 #18	0,9677 #19	AKKBÖS	0,9945 #53	0,9896 #51
VÇKÖS1	1,00 #32	1,00 #32	VÇKÖS3	1,00 #34	1,00 #34
K/B/Y	10/0/0	10/0/0	K/B/Y	8/2/0	8/2/0

Her bir öznitelik seçim yönteminin kolon ve lenfoma verileri üzerinde NB sınıflandırıcısı ile birlikte elde ettiği en yüksek AUC ve doğruluk değeri Çizelge 5.16'da gösterilmiştir. Çizelgede verilen her bir değerin bitişiğinde o değerin elde edildiği seçili öznitelik sayısı # ile belirtilmiş ve 11 ayrı yönteme göre elde edilen sonuçlar içinden en iyi olanları göstermek için koyu renkle vurgulama yapılmıştır. Kolon verisi üzerinde uygulanan VÇKÖS yönteminde AKKBÖS, KBO ve KBÖS'ün kriterler olarak kullanıldığı VÇKÖS2 yapısı, lenfoma verisi üzerinde uygulanan VÇKÖS yönteminde ise KBÖS, D-KBÖS ve KBO'nun kriterler olarak kullanıldığı VÇKÖS4 yapısı esas alınmıştır (Bkz. Çizelge 5.10).

Seçili öznitelik alt kümelerinden elde edilen optimum tahmin performansları baz alınarak VÇKÖS yöntemi ile diğer 10 yöntem arasında kıyaslamalar yapılmış ve bu kıyaslamalardan elde edilen K/B/Y skorları çizelgelerin en alt satırında gösterilmiştir. K/B/Y skorları VÇKÖS yönteminin diğer yöntemlere karşı çoğunlukla üstün geldiğini göstermiştir. kNN sınıflandırıcısı bazında gerçekleştirilen kıyaslamalarda VÇKÖS yönteminin öteki yöntemler karşısında yenilgi aldığı durum hiç yokken yalnızca birkaç kez

berabere kalma durumu ortaya çıkmıştır (Bkz. Çizelge 5.15). NB sınıflandırıcısı bazında gerçekleştirilen kıyaslamalarda ise VÇKÖS yöntemi yalnızca bir kez kolon verisi üzerinde AKKBÖS yöntemine karşı yenilmiştir. Buna ilâveten lenfoma verisi üzerinde en iyi AUC ve en iyi doğruluk performansı açısından diğer yöntemlere karşı birkaç kez beraberlik durumunda kalmıştır (Bkz. Çizelge 5.16).

Çizelge 5.16. Kolon tümörü ve lenfoma verileri üzerinde öznitelik seçicilerin NB sınıflandırıcısı ile birlikte sergilediği en iyi performans değerleri

Öznitelik Seçici	Kolon		Öznitelik Seçici	Lenfoma	
	AUC	Doğruluk		AUC	Doğruluk
KBÖS	0,9864 #5	0,9355 #6	KBÖS	1,00 #103	1,00 #103
D-KBÖS	0,9818 #25	0,9032 #8	D-KBÖS	1,00 #29	1,00 #29
KBF	0,9795 #18	0,9194 #4	KBF	0,9998 #72	0,9792 #45
KBO	0,9886 #10	0,9355 #11	KBO	1,00 #21	0,9896 #19
NKBÖS	0,9750 #14	0,9032 #5	NKBÖS	0,9996 # 92	0,9583 #87
KKBM	0,9841 #50	0,9032 #16	KKBM	0,9998 #151	0,9792 #66
KKBÖS	0,9364 #87	0,8871 #4	KKBÖS	0,9955 #198	0,9375 #67
DAÖS	0,9761 #51	0,9355 #14	DAÖS	0,9952 #11	0,9688 #10
BKBM	0,9716 #52	0,9194 #4	BKBM	0,9996 #66	0,9792 #66
AKKBÖS	0,9886 #17	0,9677 #8	AKKBÖS	0,9996 #57	0,9792 #30
VÇKÖS2	0,9898 #36	0,9516 #6	VÇKÖS4	1,00 #27	1,00 #27
K/B/Y	10/0/0	9/0/1	K/B/Y	7/3/0	8/2/0

Ortalama sınıflandırma performansı, optimum sınıflandırma performansı gibi çeşitli ölçümler temel alınarak gerçekleştirilen deneysel karşılaştırmalar şunu göstermiştir; kolon tümörü ve lenfoma veri kümelerinde öznitelik seçimi için VÇKÖS yönteminin kullanılması, sınıflandırmada daha iyi performans değerlerine ulaşılmasını sağlamıştır. Bilgi kuramı tabanlı diğer yöntemlere kıyasla VÇKÖS yöntemi K/B/Y karşılaştırmalarında daha üstün sonuçlar elde etmiştir. Buna göre VÇKÖS yöntemi kolon tümörü ve lenfoma

veri kümeleri üzerinde daha iyi tahminci öznitelik alt kümelerinin elde edilmesine ve tahmin performansının geliştirilmesine katkı sağlamıştır.

Bu bölümde son olarak, öznitelik seçim problemlerine yönelik yeni çözüm yollarının tanıtıldığı çalışmalar ele alınmıştır. Literatürdeki son yıllarda geliştirilmiş bazı öznitelik seçim yöntemlerinin [85, 180-185] kısa açıklamaları ve bu yöntemlerle kolon veri kümesi üzerinde elde edilen performans sonuçları Çizelge 5.17'de gösterilmiştir.

Çizelge 5.17. Kolon tümörü verisi üzerinde uygulanan çeşitli öznitelik seçim yöntemleri aracılığıyla sınıflandırmada ulaşılan performans değerleri

Literatür	Yöntem	Tahmin Performansı		
[180]	Simetrik belirsizlik, Fuzzy Imperialist Competitive Algorithm (FICA) ve Incremental Wrapper Subset Selection with replacement (IWSSr) yöntemlerini bir arada uygulayan hibrit yöntem	Bayesian sınıflandırıcısı altında 10-kat çapraz doğrulama ile elde edilen doğruluk: 0,936		
[85]	Binary Differential Evolution (BDE) algoritmasına dayanan bir sarmal yöntem ile sıralama temelli bir filtreleme yöntemini bir araya getiren BDE-X_{rank} yöntemi	BDE-kNN_{rank}	BDE-NB_{rank}	BDE-DVM_{rank}
		AUC: 0,755	AUC: 0,855	En iyi doğruluk: 0,75
[181]	Korelasyon Tabanlı Öznitelik Seçimini Kullanan Dağıtık Sıralama Filtresi Yaklaşımı (Distributed Ranking Filter Approach Employing Correlation-based Feature Selection - DRF-CFS)	kNN	NB	DVM
		Sınıflandırma doğruluğu: 0,8	Sınıflandırma doğruluğu: 0,85	Sınıflandırma doğruluğu: 0,8
[182]	Bilgi Kazancı/Standart Genetik Algoritma yöntemi	Genetik programlama ile elde edilen sınıflandırma doğruluğu: 0,855		
[183]	Destek Vektör Makineleri-Özyinelemeli Öznitelik Eleme (Support Vector Machines Recursive Feature Elimination - SVM-RFE) yöntemini uygulayıp, İkili Yusufçuk (Binary Dragonfly - BDF) arama stratejisi yardımıyla DVM sınıflandırma performansını optimize eden yöntem	Üç farklı DVM çeşidine göre elde edilen ortalama sınıflandırma doğruluğu: 0,975		
[184]	Birleşik Komşuluk Entropisine Dayanan Fisher Skoru ile Gen Seçimi Algoritması (Gene Selection Algorithm with the Fisher Score Based on Joint Neighborhood Entropy - GSFSJNE)	kNN	C4.5	DVM
		Sınıflandırma doğruluğu: 0,841	Sınıflandırma doğruluğu: 0,816	Sınıflandırma doğruluğu: 0,843
[185]	Relax Minimum Redundancy Maximum Relevance (RelaxMRMR) yaklaşımı	kNN	NB	DVM
		Sınıflandırma doğruluğu: 0,853	Sınıflandırma doğruluğu: 0,892	Sınıflandırma doğruluğu: 0,874

Literatüre bakıldığında öznitelik seçimini konu alan çalışmaların bir kısmında lenfoma verisinin örnek sayısının daha az olduğu ve ikili sınıflandırma problemi biçiminde ele alındığı görülmektedir [85, 181]. Son yıllarda yapılmış öznitelik seçimi odaklı bazı çalışmalar hakkında özet bilgiler ve bu çalışmalarda lenfoma veri kümesi üzerinde elde edilen tahmin performansı değerleri Çizelge 5.18'de verilmiştir.

Çizelge 5.18. Lenfoma verisi üzerinde uygulanan çeşitli öznitelik seçim yöntemleri aracılığıyla sınıflandırmada ulaşılan performans değerleri

<table>
<tr><th>Literatür</th><th>Yöntem</th><th colspan="3">Tahmin Performansı</th></tr>
<tr><td>[180]</td><td>Simetrik belirsizlik, Fuzzy Imperialist Competitive Algorithm (FICA) ve Incremental Wrapper Subset Selection with replacement (IWSSr) yöntemlerini bir arada uygulayan hibrit yöntem</td><td colspan="3">Bayesian sınıflandırıcısı altında 10-kat çapraz doğrulamadan elde edilen doğruluk: 0,941</td></tr>
<tr><td rowspan="2">[85]</td><td rowspan="2">Binary Differential Evolution (BDE) algoritmasına dayanan bir sarmal yöntem ile sıralama-temelli bir filtreleme yöntemini bir araya getiren BDE-X_{rank} yöntemi</td><td>BDE-kNN_{rank}</td><td>BDE-NB_{rank}</td><td>BDE-DVM_{rank}</td></tr>
<tr><td>AUC: 0,827</td><td>AUC: 0,837</td><td>En iyi doğruluk: 0,929</td></tr>
<tr><td rowspan="2">[181]</td><td rowspan="2">Korelasyon Tabanlı Öznitelik Seçimini Kullanan Dağıtık Sıralama Filtresi Yaklaşımı (Distributed Ranking Filter Approach Employing Correlation-based Feature Selection - DRF-CFS)</td><td>kNN</td><td>NB</td><td>DVM</td></tr>
<tr><td>Sınıflandırma doğruluğu: 0,933</td><td>Sınıflandırma doğruluğu: 0,933</td><td>Sınıflandırma doğruluğu: 0,867</td></tr>
<tr><td rowspan="2">[185]</td><td rowspan="2">Relax Minimum Redundancy Maximum Relevance (RelaxMRMR) yaklaşımı</td><td>kNN</td><td>NB</td><td>DVM</td></tr>
<tr><td>Sınıflandırma doğruluğu: 0,876</td><td>Sınıflandırma doğruluğu: 0,885</td><td>Sınıflandırma doğruluğu: 0,91</td></tr>
<tr><td>[186]</td><td>InteractedLasso regresyon modeli</td><td colspan="3">Ortalama DVM doğruluğu: 0,887</td></tr>
<tr><td rowspan="2">[187]</td><td rowspan="2">Ağırlıklı İlgi Tabanlı Öznitelik Seçimi (Feature Selection based on Weighted Relevancy - WRFS)</td><td colspan="2">kNN</td><td>DVM</td></tr>
<tr><td colspan="2">Ortalama sınıflandırma doğruluğu: 0,807</td><td>Ortalama sınıflandırma doğruluğu:0,731</td></tr>
</table>

Kolon ve lenfoma verilerinin VÇKÖS yöntemiyle ele alınması sürecinde takip edilen analiz basamakları ve uygulanan ön işlem ve doğrulama teknikleri literatürdeki diğer yöntemlerle birebir aynı olmasa da, bu veriler üzerinde daha önceden ulaşılmış sonuçlara bakıldığında VÇKÖS yönteminin öznitelik seçimi ve tahmin performansının iyileştirilmesi hususunda ümit vaat eden bir yöntem olduğu anlaşılmaktadır. VÇKÖS yöntemi ile

öznitelik seçimi yapılarak kolon tümörü verilerinin sınıflandırılmasında yakalanan en yüksek başarı oranı (doğruluk değeri) kNN ve NB sınıflayıcıları için sırasıyla 1,00 ve 0,952 olarak hesaplanmıştır. Lenfoma verilerinin sınıflandırılmasında yakalanan en yüksek başarı oranı ise her iki sınıflayıcı için 1,00 olarak hesaplanmıştır. Bu değerler literatürde kolon ve lenfoma verileri üzerinde öznitelik seçimi yapıldıktan sonra elde edilen sınıflandırma performansı değerleriyle karşılaştırılabilir derecede iyi veya bunlardan daha üstündür.

6. SONUÇ

Mikrodizi gen ifadesi verileri özellikle kanser gibi genetik etmenlerle ilgisi olan hastalıklar üzerinde etraflı ve derinlikli araştırmalar yapılmasına imkân sağlamaktadır. Gen ifadesi verileri üzerinde gerçekleştirilen analizler genel anlamda; veri ön işleme, öznitelik seçimi, sınıflandırma gibi birkaç aşamada ele alınır. Öznitelik sayısının çokluğundan oluşan boyutluluk sebebiyle, gen ifadesi verilerinin işlenmesinde çeşitli zorluklarla karşılaşılmaktadır. Bu verileri etkin bir şekilde kullanarak hastalıklar hakkında doğru öngörülere ulaşmak, yüksek boyuta sahip veri kümesi içindeki faydasız ve anlamsız bilgileri ayıklayabilmekle mümkündür. Verilerden modelleme yapılırken tahminci öznitelik olarak nelerin kullanılacağına iyi karar verilmesi gerekir. Bu nedenle gen ifadesi verilerine uygulanan bir tahmin modelinin başarısında, sınıflandırıcı yöntem kadar önemli rolü olan bir başka unsur öznitelik seçiminde kullanılan yöntemdir. Tahmin görevinde başarısızlığın önlenmesi için uygun bir öznitelik seçim yönteminin model tasarımına dâhil edilmesi gerekir.

Makine öğrenmesi yöntemlerinden faydalanılarak tahmin problemlerinin çözümüne yönelik çok sayıda güçlü tahmin modeli tanımlanabilir. Bu tahmin modelleri, gen ifadesi verileri üzerinde yapılan araştırmaların kolay ve hızlı ilerlemesine katkı sağlamaktadır.

Bu tezde gen ifadesi verileri üzerinde iki ana aşaması olan bir çalışma gerçekleştirilmiştir. İlk aşamada kolon tümörü ve lenfoma gen ifadesi verileri üzerinde çeşitli tahmin modellerinin performanslarının değerlendirilmesi amacıyla bütünleşik AHS-VIKOR yöntemi uygulanmıştır. Uygulama sürecinde öncelikle, bilgi kuramı temelli öznitelik seçim yöntemlerinden olan; KBÖS, D-KBÖS, KBF, KBO, NKBÖS, KKBM, KKBÖS, DAÖS, BKBM ve AKKBÖS yöntemleri ile kNN ve NB sınıflandırıcıları arasında ikişerli şekilde kombinasyon yapılarak 20 farklı model oluşturulmuş ve bu modeller tahmin probleminin çözümüne aday gösterilmiştir. Ardından AUPRC, AUC, MKK, doğruluk ve f-ölçütü kriterlerine göre adayların performans değerleri elde edilmiştir. Bu kriterler AHS yöntemi kullanılarak öznel ağırlıklandırmaya tâbi tutulmuştur. AHS yöntemiyle elde edilen ağırlıklar ve tahmin modellerinden elde edilen performans sonuçları VIKOR yöntemi ile ele alınarak uzlaşık sonuçlar bulunmuştur. Böylece gen ifadesi verilerinde tahmin probleminin çözümüne yönelik model seçme işi bir ÇKKV problemi olarak formüle edilip,

tahmin modelleri karar analizleri ile değerlendirilmiştir. Tahmin performansı açısından bu modellerin birbirlerine göre üstünlüğünü gösteren puanlama değerleri ve sıralama çıktıları elde edilmiştir. Bütünleşik AHS-VIKOR yöntemiyle gerçekleştirilen bu değerlendirme süreci sayesinde, kolon tümörü ve lenfoma gen ifadesi verilerinin sınıflandırılması için aday olarak belirlenen 20 model arasından hangilerinin bu problemi en uygun şekilde çözdüğü tespit edilmiştir.

Kolon verileri üzerinde elde edilen uzlaşık sonuçlara göre tahmin modellerinin (öznitelik seçici & sınıflandırıcı) iyiden kötüye doğru sıralanışı şu şekildedir: DAÖS & kNN, AKKBÖS & NB, KBO & NB, AKKBÖS & kNN, KBÖS & NB, KBÖS & kNN, KKBM & kNN, D-KBÖS & kNN, D-KBÖS & NB, KBO & kNN, KKBM & NB, DAÖS & NB, KBF & NB, NKBÖS & NB, BKBM & NB, KKBÖS & kNN, KKBÖS & NB, BKBM & kNN, KBF & kNN, NKBÖS & kNN. Lenfoma verileri üzerinde tahmin modellerinin iyiden kötüye doğru başarı sırası ise şu şekildedir; D-KBÖS & kNN, KBO & kNN, KBÖS & NB, D-KBÖS & NB, KBO & NB, AKKBÖS & NB, BKBM & NB, KBF & NB, KKBM & NB, AKKBÖS & kNN, BKBM & kNN, KKBM & kNN, KBÖS & kNN, KBF & kNN, NKBÖS & NB, DAÖS & kNN, KKBÖS & NB, NKBÖS & kNN, DAÖS & NB, KKBÖS & kNN.

Uzlaşık sıralama sonuçları sınıflandırıcı yöntemler bazında değerlendirildiğinde kolon tümörü verileri üzerinde kNN yöntemiyle, lenfoma verileri üzerinde ise NB yöntemiyle daha başarılı tahmin sonuçlarına ulaşıldığı görülmüştür. Kolon tümörü verileri üzerinde öznitelik seçiminde DAÖS ve AKKBÖS yöntemlerinin daha avantajlı olduğu; KBO, KBÖS, D-KBÖS ve KKBM yöntemleri ile kayda değer sonuçların elde edildiği, bununla birlikte NKBÖS, BKBM, KKBÖS ve KBF yöntemlerinin daha düşük performanslı olduğu sonucuna varılmıştır. Lenfoma verileri üzerinde öznitelik seçim görevinin ifasında D-KBÖS ve KBO yöntemlerinin daha etkili olduğu, KBÖS ve KBF yöntemlerininse sadece NB sınıflandırıcısıyla iyi performans gösterdiği tespit edilmiştir. Bunun yanı sıra AKKBÖS, BKBM ve KKBM yöntemlerinden orta düzeyde performans elde edilirken NKBÖS, DAÖS ve KKBÖS yöntemlerinden düşük performans alındığı sonucuna varılmıştır.

Tez çalışmasının ilk aşaması sonucunda, kolon tümörü verisi için DAÖS ile kNN yöntemlerinin birleşiminden oluşan tahmin modeli uzlaşık çözüm olarak önerilmiştir.

Lenfoma verisi içinse 4 ayrı tahmin modelini içine alan bir çözüm kümesi elde edilmiştir. Bu kümenin elemanları; D-KBÖS ile kNN, KBO ile kNN, KBÖS ile NB ve D-KBÖS ile NB yöntemlerinin kombine edilmesiyle oluşturulan tahmin modellerini temsil etmektedir.

Tez çalışmasının ikinci aşamasında, öznitelik seçiminde birden çok kritere uygun olarak karar vermeyi sağlayan hibrit bir yöntem önerilmiştir. Önerilen yöntemde; tahminci özniteliklerin çoklu kriterlere göre elde edilen değerlendirme sonuçları VIKOR yönteminin işlem adımlarından yararlanılarak bütünleştirilip, uzlaşık bir seçili alt küme oluşturulmuştur. Öznitelik seçiminde VIKOR yöntemi yardımıyla bilgi kuramı tabanlı bireysel kriterleri bir araya getiren bu hibrit yaklaşıma VÇKÖS yöntemi adı verilmiştir.

Makine öğrenmesinde tahmin problemlerinin çözümüne yönelik önerilen yöntemlerin başarısı problem verisine göre şekillenmektedir. Bunun yanı sıra, farklı türde sınıflandırma yöntemlerine göre en uygun özniteliklerin belirlenmesi önemlidir. Bu çerçevede VÇKÖS yöntemi kolon ve lenfoma gen ifadesi verileri üzerinde kNN ve NB yöntemleri ile gerçekleştirilmesi planlanan her bir sınıflandırma çözümü için özel bir kriter seti kullanılarak uygulanmıştır. Bu kriter setleri belirlenirken tez çalışmasının ilk aşamasında uygulanan değerlendirme sürecinin sonuçlarından yararlanılmıştır. Kolon ve lenfoma verileri için tahmin modellerinin tercih edilme sırasını gösteren uzlaşık sıralama listeleri, her bir sınıflandırma çözümü için önemli olan öznitelik seçim kriterlerinin belirlenmesi amacıyla kullanılmıştır. Bu amaçla uzlaşık sıralamalar KNN ve NB sınıflandırıcılarına göre ayrılarak her bir sınıflandırıcı bazında öznitelik seçim kriterleri içerisinden en uygun olanlar tespit edilmiştir. Her bir sınıflandırma çözümü için belirlenen en başarılı 5 kriter arasından ilk 2'si veya ilk 3'ü VÇKÖS yöntemi çerçevesinde birleştirilmek üzere seçilmiştir. Buna göre kolon tümörü verileri KNN yöntemi ile sınıflandırılırken öznitelik seçim kriteri olarak DAÖS ve AKKBÖS yöntemleri kullanıldığında en iyi sonuçlar elde edilmiştir. Kolon tümörü verilerinin sınıflandırılmasında NB ile birlikte en iyi sonuç veren kriterler AKKBÖS, KBO ve KBÖS kriterleri olarak belirlenmiştir. Lenfoma verilerinin kNN ile sınıflandırılmasında D-KBÖS, KBO ve AKKBÖS kriterleri daha başarılı bir çözüm elde etmeyi sağlarken, NB ile sınıflandırılmasında çözümü optimize etmek için KBÖS, D-KBÖS ve KBO kriterlerinin kullanışlı olduğu belirlenmiştir. Belirlenen her bir kriter topluluğu ilgili olduğu sınıflandırma çözümünde tahmin performansının geliştirilmesi amacıyla VÇKÖS yönteminin yapısında kriter seti olarak kullanılmıştır.

VÇKÖS yöntemi ile en ayrımcı alt kümenin elde edilmesine yönelik süreç, kriterler arasında uzlaşı arayışında VIKOR yönteminin analiz adımlarının kullanımına dayanır. Bu süreçte; öznitelik seçim prosedürünün yinelenen her seçim adımında VIKOR analizleri tekrar uygulanmış ve aday öznitelikler kararlaştırılan kriter seti uyarınca değerlendirilip, VÇKÖS yöntemiyle elde edilen genel tercih puanlarına göre yeniden sıralanmıştır. Her adımda uzlaşık sıralamanın en başındaki bir öznitelik seçilerek artırımsal stratejiyle seçim işlemine devam edilmiştir. Böylece kolon ve lenfoma verilerine ait öznitelikler birden çok öznitelik seçim kriterine bağlı kalınarak VÇKÖS yöntemi ile değerlendirmeye tâbi tutulmuş ve çoklu kriterlerin uzlaşma kararını yansıtan seçili öznitelik alt kümeleri elde edilmiştir.

VÇKÖS yönteminin etkinliğinin gösterilmesi amacıyla deneysel uygulama sürecinde bireysel, ortalama ve en iyi performans değerleri bağlamında değerlendirmeler yapılmıştır. Deneylerde VÇKÖS yönteminin tasarımında temel alınan 2 çeşit gen ifadesi verisi kullanılmıştır.

Kolon ve lenfoma verileri üzerinde uygulanan her bir sınıflandırma çözümü için VÇKÖS yöntemi ile diğer öznitelik seçim yöntemleri artırımsal seçim adımları boyunca elde ettikleri öznitelik alt kümelerinin bireysel performansları bakımından karşılaştırılmıştır. Grafikler üstünde verilen performans değerleri, VÇKÖS yöntemi ile elde edilen tahminci özniteliklerin kullanıldığı sınıflandırma çözümlerinde optimum veya karşılaştırılabilir derecede iyi tahmin sonuçlarına ulaşıldığını göstermiştir.

Kolon tümörü verisi üzerinde ortalama performans değerleri bakımından VÇKÖS yönteminin 30, 50 ve 100 öznitelik için kNN sınıflayıcısıyla hesaplanan AUC ve doğruluk değerleri sırasıyla 0,944; 0,963; 0,959 ve 0,937; 0,958; 0,951 çıkmıştır. Bu değerler bilgi kuramsal diğer 10 öznitelik seçim yönteminin sonuçlarıyla mukayese edildiğinde en yüksek ortalama performans değerleri olma başarısını göstermiştir. Kolon verisi için VÇKÖS yönteminin NB sınıflayıcısıyla hesaplanan ortalama AUC ve ortalama doğruluk değerleri 30, 50 ve 100 öznitelik için sırasıyla 0,965; 0,973; 0,976 ve 0,917; 0,914; 0,898 çıkmıştır. Bunlar içinde yalnızca 30 öznitelik için 0,965 olarak hesaplanan ortalama AUC değeri AKKBÖS yönteminin elde ettiği 0,966 değerinden daha düşüktür. Geriye kalan tüm değerler öteki yöntemlere kıyasla en üstün sonuçları ifade etmektedir.

Lenfoma verisi üzerinde VÇKÖS yönteminin kNN sınıflayıcısıyla birlikte 30, 50, 100, 150 ve 200 öznitelik için elde ettiği ortalama AUC ve ortalama doğruluk değerleri sırasıyla 0,965; 0,978; 0,988; 0,991; 0,992 ve 0,898; 0,935; 0,962; 0,973; 0,977 çıkmıştır. Bu değerler bilgi kuramsal diğer 10 öznitelik seçim yönteminin test sonuçlarıyla karşılaştırıldığında; 30 öznitelikte ortalama AUC değeri KBÖS, D-KBÖS ve KKBM yöntemlerinin elde ettiği AUC sonucuna nispeten daha düşük çıkmıştır, ancak diğer 7 yöntem tarafından elde edilen sonuçlardan daha üstündür. Buna karşın VÇKÖS yöntemi 50, 100, 150 ve 200 öznitelikte elde ettiği ortalama AUC değerleri bakımından öteki yöntemlerin tümüne üstün gelmiştir. VÇKÖS yönteminin 30 öznitelikle elde ettiği ortalama doğruluk diğer 4 yöntem tarafından elde edilen sonuçlara üstün gelmiş ve 6 yöntem karşısında düşük çıkmıştır. VÇKÖS yönteminin 50 öznitelikle elde ettiği ortalama doğruluk 8 yönteme üstün gelmiş, D-KBÖS ve KKBM yöntemleri karşısında düşük çıkmıştır; 100, 150 ve 200 öznitelikle elde ettiği ortalama doğruluk öteki yöntemlerin tamamına karşı üstün gelmiştir.

Lenfoma verisi üzerinde VÇKÖS yönteminin 30, 50, 100, 150 ve 200 öznitelik için NB sınıflayıcısıyla hesaplanan ortalama AUC değerleri sırasıyla 0,981; 0,989; 0,994; 0,996 ve 0,997 çıkmıştır. Ortalama doğruluk değerleri ise 30, 50, 100, 150 ve 200 öznitelik için sırasıyla 0,917; 0,950; 0,972; 0,979 ve 0,982 çıkmıştır. VÇKÖS yönteminin NB sınıflayıcısı ile birlikte elde ettiği ortalama AUC değerleri diğer 10 yönteme karşı 30, 50, 100, 150 ve 200 öznitelik için sırasıyla 6 kazanma, 4 yenilgi; 8 kazanma, 2 yenilgi; 9 kazanma, 1 yenilgi; 9 kazanma, 1 yenilgi ve 8 kazanma, 1 beraberlik, 1 yenilgi durumu elde etmiştir. VÇKÖS yönteminin NB yöntemi ile elde edilen ortalama doğruluk değerleri diğer 10 yöntem karşısında yalnızca bir kez, 30 öznitelik için KBÖS yöntemi karşısında düşük çıkmıştır. Bunun haricinde 30 öznitelik için diğer 9 yönteme karşı, 50, 100, 150 ve 200 öznitelik içinse 10 öznitelik seçicinin hepsine karşı üstünlük elde edilmiştir.

Bireysel ve ortalama performans karşılaştırmalarının yanı sıra bir de optimum performans değerlerinin karşılaştırması yapılmıştır. VÇKÖS yöntemi; kolon tümörü verisi üzerinde kNN sınıflandırıcısı ile birlikte kullanıldığında seçili 32 öznitelik için AUC ve doğruluk değerleri 1,00 olarak ölçülmüştür. NB sınıflandırıcısı ile birlikte kullanıldığında seçili 36 öznitelik için AUC değeri 0,990 ve seçili 6 öznitelik için doğruluk değeri 0,952 olarak ölçülmüştür. VÇKÖS yöntemi lenfoma verisi üzerinde kNN sınıflandırıcısı ile birlikte kullanıldığında seçili 34 öznitelik için AUC ve doğruluk değerleri 1,00 olarak ölçülmüştür.

NB sınıflandırıcısı ile birlikte kullanıldığında ise seçili 27 öznitelik için AUC ve doğruluk değerleri 1,00 olarak ölçülmüştür. VÇKÖS yöntemi ile en iyi tahmin performansı bağlamında elde edilen bu sonuçlar diğer yöntemlerin performans sonuçları ile karşılaştırılmıştır. Deneysel karşılaştırma durumlarının her biri için VÇKÖS yönteminin diğer 10 yönteme kıyasla elde ettiği kazanma, beraberlik ve yenilgi sayısı K/B/Y biçiminde verilen özet gösterimle ifade edilmiştir. Bu kıyaslamalara göre VÇKÖS yöntemi optimum performans değerleri bakımından tüm karşılaştırmalarda yalnızca bir kez, kolon verisi üzerinde NB sınıflandırıcısı ile elde ettiği doğruluk değeri bakımından AKKBÖS yöntemi karşısında yenik düşmüştür. Diğer karşılaştırmalarda çoğunlukla kazanmış, bazı durumlarda ise beraberlik sonucu elde etmiştir. Beraberlik sonuçları az sayıda olmakla birlikte lenfoma verisi üzerinde gerçekleştirilen performans ölçümlerinde ortaya çıkmıştır.

VÇKÖS yöntemiyle seçilen özniteliklerin sınıflandırmada elde ettiği performans değerleri, son yıllarda yapılan öznitelik seçimi odaklı çalışmalarda kolon ve lenfoma verileri üzerinde ulaşılan tahmin performansı değerlerinin çoğundan daha iyi düzeydedir [85, 180-187].

Özet olarak bu tez çalışmasında gen ifadesi verileri makine öğrenmesinin konusu olan bir dizi yöntem ve ÇKKV yöntemleri kullanılarak analiz edilmiştir. ÇKKV yöntemleri değerlendirme ve seçim problemlerinin çözümü için matematiksel hesaplamalara dayalı ilkeli bir araştırma zemini sunmaktadır. Bu noktadan hareketle ilk önce, kolon tümörü ve lenfoma verilerinin sınıflandırılmasına yönelik en etkili tahmin modellerinin tespiti için bütünleşik AHS-VIKOR yöntemi uygulanmıştır. Uygulama neticesinde, çoklu kriterlere göre model değerlendirme sonuçlarını ifade eden uzlaşık sıralamalar ve uzlaşık çözümler elde edilmiştir. Bu sayede birbirinden farklı model değerlendirme kriterlerine göre elde edilen mukayese sonuçları arasındaki çelişkiler yok edilmiş ve tahmin modeli seçim problemine belirli bir kriter seti üzerine kurulu sistemli bir çözüm getirilmiştir. Sonuçlar model seçiminde ÇKKV yaklaşımını kullanmanın faydalı ve etkili bir çözüm yolu olduğunu göstermiştir. Bu süreç tamamlandıktan sonra, VIKOR yöntemi esas alınarak çok kriterli yeni bir öznitelik seçim yapısı geliştirilmiştir. Kısaca VÇKÖS yöntemi olarak adlandırılan bu yapı ile tekli öznitelik seçim kriterlerinin avantajları bir araya getirilmiştir. VÇKÖS yöntemi kolon ve lenfoma verileri üzerinde kNN ve NB sınıflayıcılarının her biri için ayrı bir optimal kriter seti ile uygulanmıştır. Farklı sınıflandırma çözümleri için en avantajlı kriterler belirlenirken tahmin modellerinin ilk aşamada elde edilen sıralama

sonuçları kullanılmıştır. Kriter tespitinin uzlaşık sıralamalara dayandırılması aracılığıyla tez çalışmasının iki ana aşaması arasında faydacı bir bağlantı kurulmuştur. Tespit edilen kriter setleri kullanılarak VÇKÖS yöntemi gen ifadesi verileri üzerinde uygulanmıştır. VÇKÖS yöntemi ile elde edilen AUC ve doğruluk değerleri bilgi kuramı tabanlı 10 ayrı yöntemin test sonuçlarıyla karşılaştırılmıştır. Böylece VIKOR tabanlı öznitelik seçim yapısının gen ifadesi verilerinde tahmin performansının iyileştirilmesine katkı sağladığı deneysel olarak ortaya konmuştur.

Bu tez çalışmasının ilk aşaması kapsamında tahmin modellerinin değerlendirilmesi için bütünleşik AHS-VIKOR yöntemi ile uygulanan analiz sürecinin bir benzeri; 6 öznitelik seçim yöntemi ve 3 sınıflandırıcı kullanılarak oluşturulan 18 alternatif ile 8 ayrı performans kriterine göre göğüs kanseri verileri üzerinde uygulanmış olup, ilgili çalışma Uluslararası Yenilikçi Mühendislik Uygulamaları Konferansı'nda [188] yayımlanmıştır.

KAYNAKLAR

1. Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15), 2429-2437.

2. Lee, J. W., Lee, J. B., Park, M., and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4), 869-885.

3. Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Computer Science*, 23, 5-14.

4. Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., and Wang, K. (2013). Feature selection using dynamic weights for classification. *Knowledge-Based Systems*, 37, 541-549.

5. İnternet: IARC. All cancers. Globocan 2018. URL: http://www.webcitation.org/query?url=https%3A%2F%2Fgco.iarc.fr%2Ftoday%2Fdata%2Ffactsheets%2Fcancers%2F39-All-cancers-fact-sheet.pdf&date=2019-05-10, Son Erişim Tarihi: 10.05.2019.

6. Kaya, T., and Kahraman, C. (2010). Multicriteria renewable energy planning using an integrated fuzzy VIKOR & AHP methodology: the case of Istanbul. *Energy*, 35(6), 2517-2527.

7. Sennaroglu, B., and Celebi, G. V. (2018). A military airport location selection by AHP integrated PROMETHEE and VIKOR methods. *Transportation Research Part D: Transport and Environment*, 59, 160-173.

8. Karaman, B. (2014). *0-1 Hedef Programlama Destekli Bütünleşik AHP-VIKOR Yöntemi: Hastane Yatırımı Projeleri Seçimi,* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 1-3.

9. Luthra, S., Govindan, K., Kannan, D., Mangla, S. K., and Garg, C. P. (2017). An integrated framework for sustainable supplier selection and evaluation in supply chains. *Journal of Cleaner Production*, 140, 1686-1698.

10. Crick, F. H. C. (1957). Nucleic acids. *Scientific American*, 197(3), 188-203.

11. National Institute of General Medical Sciences. (2010). *The new genetics* [Brochure]. Bethesda, MD: NIH.

12. Crick, F. H. C. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12, 138-163.

13. Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561-563.

14. Zhang, S., Davidson, D. D., Zhang, D. Y., Parks, J. A., and Cheng, L. (2008). Principles of clinical molecular biology. In L. Cheng and D. Y. Zhang (Eds.), *Molecular genetic pathology*. Totowa, NJ: Humana, 2-32.

15. Qin, J., Yan, B., Hu, Y., Wang, P., and Wang, J. (2016). Applications of integrative OMICs approaches to gene regulation studies. *Quantitative Biology*, 4(4), 283-301.

16. Rodenhiser, D., and Mann, M. (2006). Epigenetics and human disease: translating basic biology into clinical applications. *Canadian Medical Association Journal*, 174(3), 341-348.

17. İnternet: U.S. National Library of Medicine. What is epigenetics?. Genetics Home Reference. URL: http://www.webcitation.org/query?url=https%3A%2F%2Fghr.nlm.nih.gov%2Fprimer%2Fhowgeneswork%2Fepigenome&date=2019-05-10, Son Erişim Tarihi: 10.05.2019.

18. Carlberg, C. and Molnár, F. (2016). *Mechanisms of gene regulation* (Second Edition). Dordrecht, Netherlands: Springer, 3, 168.

19. Sawan, C., Vaissière, T., Murr, R., and Herceg, Z. (2008). Epigenetic drivers and genetic passengers on the road to cancer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 642(1-2), 1-13.

20. Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonel Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1), D766-D773.

21. Wong, G. (2005). Introduction. In J. Tuimala and M. M. Laine (Eds.), *DNA microarray data analysis*. Second Edition. Helsinki, Finland: CSC–Scientific Computing Limited, 19.

22. Singh, R. K., and Sivabalakrishnan, M. (2015). Feature selection of gene expression data for cancer classification: a review. *Procedia Computer Science*, 50, 52-57.

23. George, G. V. S., and Raj, V. C. (2011). Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. *International Journal of Computer Science and Engineering Survey*, 2(3), 16-27.

24. Khan, J., Saal, L. H., Bittner, M. L., Chen, Y., Trent, J. M., and Meltzer, P. S. (1999). Expression profiling in cancer using cDNA microarrays. *Electrophoresis*, 20(2), 223-229.

25. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338), 680-686.

26. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares Jr, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1), 262-267.

27. Lee, C. P., and Leu, Y. (2011). A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, 11(1), 208-213.

28. Arrais, J. P., Carreto, L., Santos, M. A., and Oliveira, J. L. (2008). *A microarray information database*. International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies, IEEE, 170-175.

29. Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y. Q., Ye, S. Q., and Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2(5), 345-349.

30. Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., and Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, 10(1), 46.

31. Chiu, C. C., Chan, S. Y., Wang, C. C., and Wu, W. S. (2013). Missing value imputation for microarray data: a comprehensive comparison study and a web tool. *BMC Systems Biology*, 7(6), S12.

32. Selvaraj, S., and Natarajan, J. (2011). Microarray data analysis and mining tools. *Bioinformation*, 6(3), 95-99.

33. Pirooznia, M., Yang, J. Y., Yang, M. Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(1), S13.

34. Guarnaccia, M., Gentile, G., Alessi, E., Schneider, C., Petralia, S., and Cavallaro, S. (2014). Is this the real time for genomics?. *Genomics*, 103(2-3), 177-182.

35. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armananzas, R., Santafe, G., Perez, A., and Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1), 86-112.

36. Kotsiantis, S. B. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, 31(3), 249-268.

37. Cover, T. M., and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

38. John, G. H., and Langley, P. (1995). *Estimating continuous distributions in Bayesian classifiers*. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, 338-345.

39. Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. New York: Spartan Books.

40. Vapnik, V. N. (1998). *Statistical learning theory*. New York: John Wiley.

41. Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronics age*. Edinburgh: Edinburgh University Press.

42. Cruz, J. A., and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2, 59-77.

43. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8-17.

44. Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., and Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5), 631-643.

45. Fu, L. M., and Fu-Liu, C. S. (2004). Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. *FEBS Letters*, 561(1-3), 186-190.

46. Liu, J. J., Cutler, G., Li, W., Pan, Z., Peng, S., Hoey, T., Chen, L., and Ling, X. B. (2005). Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, 21(11), 2691-2697.

47. Zhao, G., and Liu, S. (2016). Estimation of discriminative feature subset using community modularity. *Scientific Reports*, 6, 25040.

48. Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(1), 9.

49. Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 11(2), 253-264.

50. Moorthy, K., Mohamad, M. S., and Deris, S. (2014). A review on missing value imputation algorithms for microarray gene expression data. *Current Bioinformatics*, 9(1), 18-22.

51. Liew, A. W. C., Law, N. F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, 12(5), 498-513.

52. Theodoridis, S. and Koutroumbas, K. (2009). *Pattern recognition* (Fourth Edition). Burlington, MA: Academic Press, 13-14, 59, 263, 265-267.

53. Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(02), 185-205.

54. Li, Z., Xie, W., and Liu, T. (2018). Efficient feature selection and classification for microarray data. *PloS One*, 13(8), e0202167.

55. Gallo, C. A., Cecchini, R. L., Carballido, J. A., Micheletto, S., and Ponzoni, I. (2016). Discretization of gene expression data revised. *Briefings in Bioinformatics*, 17(5), 758-770.

56. Gallo, C. A., Carballido, J. A., and Ponzoni, I. (2011). Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinformatics*, 12(1), 123.

57. Liu, H., Hussain, F., Tan, C. L., and Dash, M. (2002). Discretization: an enabling technique. *Data Mining and Knowledge Discovery*, 6(4), 393-423.

58. Dash, R., Paramguru, R. L., and Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, 2(3), 29-37.

59. Dougherty, J., Kohavi, R., and Sahami, M. (1995). *Supervised and unsupervised discretization of continuous features*. Proceedings of the Twelfth International Conference on Machine Learning, Morgan Kaufmann, 194-202.

60. Chen, Y., Zhang, Z., Zheng, J., Ma, Y., and Xue, Y. (2017). Gene selection for tumor classification using neighborhood rough sets and entropy measures. *Journal of Biomedical Informatics*, 67, 59-68.

61. Bennasar, M., Hicks, Y., and Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42(22), 8520-8532.

62. Sun, J., Passi, K., and Jain, C. K. (2016). *Improved microarray data analysis using feature selection methods with machine learning methods*. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 1527-1534.

63. Kursun, O., Sakar, C. O., Favorov, O., Aydin, N., and Gurgen, F. (2010). Using covariates for improving the minimum redundancy maximum relevance feature selection method. *Turkish Journal of Electrical Engineering and Computer Sciences*, 18(6), 975-989.

64. Fayyad, U. M., and Irani, K. B. (1993). *Multi-interval discretization of continuous-valued attributes for classification learning*. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1022-1027.

65. Zeng, X. Q., and Li, G. Z. (2014). Supervised redundant feature detection for tumor classification. *BMC Medical Genomics*, 7(2), S5.

66. Cai, J., Luo, J., Liang, C., and Yang, S. (2017). A novel information theory-based ensemble feature selection framework for high-dimensional microarray data. *International Journal of Performability Engineering*, 13(5), 742-753.

67. Hengpraprohm, S., and Chongstitvatana, P. (2009). Feature selection by weighted-SNR for cancer microarray data classification. *International Journal of Innovative Computing, Information and Control*, 5(12), 4627-4636.

68. Hua, J., Tembe, W. D., and Dougherty, E. R. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3), 409-424.

69. Hira, Z. M., and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015, 1-13.

70. Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: a review. In C. C. Aggarwal (Ed.), *Data classification: algorithms and applications*. CRC Press, 37-64.

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

72. Ahmad, F. K., Norwawi, N. M., Deris, S., and Othman, N. H. (2008). *A review of feature selection techniques via gene expression profiles*. Proceedings of the International Symposium on Information Technology, IEEE, 2, 1-7.

73. Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.

74. Goodison, S., Sun, Y., and Urquidi, V. (2010). Derivation of cancer diagnostic and prognostic signatures from gene expression data. *Bioanalysis*, 2(5), 855-862.

75. Alkuhlani, A., Nassef, M., and Farag, I. (2017). Multistage feature selection approach for high-dimensional cancer data. *Soft Computing*, 21(22), 6895-6906.

76. Liu, Q., Sung, A. H., Chen, Z., Liu, J., Chen, L., Qiao, M., Wang, Z., Huang, X., and Deng, Y. (2011). Gene selection and classification for cancer microarray data based on machine learning and similarity measures. *BMC Genomics*, 12(5), S1.

77. Meyer, P. E., Schretter, C., and Bontempi, G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2(3), 261-274.

78. Ang, J. C., Mirzal, A., Haron, H., and Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5), 971-989.

79. Liu, H., and Setiono, R. (1995). *Chi2: feature selection and discretization of numeric attributes*. Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence, IEEE, 388-391.

80. Peng, H., Long, F., and 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.

81. Pudil, P., Novovičová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11), 1119-1125.

82. Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, 31(2), 91-103.

83. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3), 389-422.

84. Maldonado, S., Weber, R., and Famili, F. (2014). Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information Sciences*, 286, 228-246.

85. Apolloni, J., Leguizamón, G., and Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. *Applied Soft Computing*, 38, 922-932.

86. Yang, C. H., Chuang, L. Y., and Yang, C. H. (2010). IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. *Journal of Medical and Biological Engineering*, 30(1), 23-28.

87. Ruiz, R., Riquelme, J. C., and Aguilar-Ruiz, J. S. (2006). Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 39(12), 2383-2392.

88. He, S., Chen, H., Zhu, Z., Ward, D. G., Cooper, H. J., Viant, M. R., Heath, J. K., and Yao, X. (2015). Robust twin boosting for feature selection from high-dimensional omics data with label noise. *Information Sciences*, 291, 1-18.

89. Dash, M., and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1-4), 131-156.

90. Wang, L., Wang, Y., and Chang, Q. (2016). Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods*, 111, 21-31.

91. Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., and Fong, S. (2018). Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, 26(1), 329-339.

92. Gnana, D. A. A., Balamurugan, S. A. A., and Leavline, E. J. (2016). Literature review on feature selection methods for high-dimensional data. *International Journal of Computer Applications*, 136(1), 9-17.

93. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., and Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282, 111-135.

94. Loo, L. H., Roberts, S., Hrebien, L., and Kam, M. (2007). New criteria for selecting differentially expressed genes. *IEEE Engineering in Medicine and Biology Magazine*, 26(2), 17-26.

95. Stolovitzky, G. (2003). Gene selection in microarray data: the elephant, the blind men and our algorithms. *Current Opinion in Structural Biology*, 13(3), 370-376.

96. Raychaudhuri, S., Sutphin, P. D., Chang, J. T., and Altman, R. B. (2001). Basic microarray analysis: grouping and feature reduction. *TRENDS in Biotechnology*, 19(5), 189-193.

97. Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2), 189-201.

98. Zeng, Z., Zhang, H., Zhang, R., and Yin, C. (2015). A novel feature selection method considering feature interaction. *Pattern Recognition*, 48(8), 2656-2666.

99. Inza, I., Sierra, B., Blanco, R., and Larrañaga, P. (2002). Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent & Fuzzy Systems*, 12(1), 25-33.

100. Yusta, S. C. (2009). Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters*, 30(5), 525-534.

101. Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers*, C-20(9), 1100-1103.

102. Marill, T., and Green, D. M. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9(1), 11-17.

103. Strearns, S. D. (1976). *On selecting features for pattern classifiers*. Proceedings of the 3rd International Conference on Pattern Recognition, 71-75.

104. Wang, A., An, N., Chen, G., Li, L., and Alterovitz, G. (2015). Accelerating wrapper-based feature selection with k-nearest-neighbor. *Knowledge-Based Systems*, 83, 81-91.

105. Jirapech-Umpai, T., and Aitken, S. (2005). Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6(1), 148.

106. Kar, S., Sharma, K. D., and Maitra, M. (2015). Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive k-nearest neighborhood technique. *Expert Systems with Applications*, 42(1), 612-627.

107. Celik, C., and Bilge, H. S. (2015). Feature selection with weighted conditional mutual information. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 30(4), 585-596.

108. Doğan, M. (1985). *İşletmelerde karar verme teknikleri*. İzmir: Bilgehan Basımevi.

109. Dağdeviren, M., and Eren, T. (2001). Analytical hierarchy process and use of 0-1 goal programming methods in selecting supplier firm. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 16(2), 41-52.

110. Karakaya, K. (2003). *İstanbul Boğazı'ndan geçen gemilerin emniyetli geçişinin analitik hiyerarşi prosesi kullanarak analizi*. Yayımlanmamış Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli.

111. Stewart, T. J. (2003). Decision-making approaches. In H. Bidgoli (Ed.), *Encyclopedia of information systems*. New York: Academic Press, 1, 535-549.

112. Çıtak, Ş. (2013). *Bir Elektronik Firmasında Çok Ölçütlü Stok Sınıflandırma,* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 16.

113. Hwang, C. L., and Yoon, K. (1981). Introduction. In M. Beckmann and H. P. Kunzi (Eds.), *Lecture Notes in Economics and Mathematical Systems: Vol. 186*. *Multiple attribute decision making: methods and applications: a state-of-the-art survey*. Berlin, Heidelberg: Springer-Verlag, 1-15.

114. Koçak, D. (2016). *Klasik ve Bulanık Çok Kriterli Karar Verme Yöntemleri ve Uygulama,* Yüksek Lisans Tezi, Gazi Üniversitesi Sosyal Bilimler Enstitüsü, Ankara, 10-12.

115. Habenicht, W., Scheubrein, B., and Scheubrein, R. (2002). Multiple-criteria decision making. In U. Derigs (Ed.), *Theme 6.5 "Optimization and Operations Research" of the Encyclopedia of Life Support Systems (EOLSS), Developed under the auspices of the UNESCO*. Oxford, UK: EOLSS Publishers.

116. Arıcan, İ. E. (2014). *Olimpiyat Oyunlarına Türkiye'den Ev Sahipliği Yapacak Aday Şehrin Seçim Sürecinin Çok Ölçütlü Karar Verme Yöntemleri ile Değerlendirilmesi,* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 23-24.

117. Chen, S. J., and Hwang, C. L. (1992). Multiple attribute decision making – an overview. In M. Beckmann and W. Krelle (Eds.), *Lecture Notes in Economics and Mathematical Systems: Vol. 375*. *Fuzzy multiple attribute decision making: methods and applications*. Berlin, Heidelberg: Springer-Verlag, 16-41.

118. Öztel, A. (2016). *Çok Kriterli Karar Verme Yöntemi Seçiminde Yeni Bir Yaklaşım,* Doktora Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 3-13.

119. Can, Ş. (2012). *Bir Savunma Sanayi Firmasında Çok Kriterli Alt Yüklenici Seçim Problemi ve Çözümü,* Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 13.

120. Yoe, C. (2002). Trade-off analysis planning and procedures guidebook. *US Army Corps of Engineers, 310, Alexandria*.

121. Bernroider, E. W. N., and Mitlohner, J. (2005). Characteristics of the multiple attribute decision making methodology in enterprise resource planning software decisions. *Communications of the IIMA*, 5(1), 6.

122. Saaty, T. L. (1994). How to make a decision: the analytic hierarchy process. *Interfaces*, 24(6), 19-43.

123. Tzeng, G. H. and Huang, J. J. (2011). *Multiple attribute decision making: methods and applications*. New York: CRC Press.

124. Velasquez, M., and Hester, P. T. (2013). An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, 10(2), 56-66.

125. Mardani, A., Jusoh, A., Nor, K., Khalifah, Z., Zakwan, N., and Valipour, A. (2015). Multiple criteria decision-making techniques and their applications–a review of the literature from 2000 to 2014. *Economic Research-Ekonomska Istraživanja*, 28(1), 516-571.

126. Mardani, A., Zavadskas, E. K., Govindan, K., Senin, A. A., and Jusoh, A. (2016). VIKOR technique: a systematic review of the state of the art literature on methodologies and applications. *Sustainability*, 8(1), 37.

127. Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2), 1-39.

128. Kou, G., Peng, Y., and Wang, G. (2014). Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences*, 275, 1-12.

129. Kou, G., Peng, Y., and Lu, C. (2014). MCDM approach to evaluating bank loan default models. *Technological and Economic Development of Economy*, 20(2), 292-311.

130. Peng, Y., Kou, G., Ergu, D., Wu, W., and Shi, Y. (2012). An integrated feature selection and classification scheme. *Studies in Informatics and Control*, 21(3), 241-248.

131. Kou, G., Peng, Y., Shi, Y., and Wu, W. (2012). Classifier evaluation for software defect prediction. *Studies in Informatics and Control*, 21(2), 117-126.

132. Peng, Y., Wang, G., and Wang, H. (2012). User preferences based software defect detection algorithms selection using MCDM. *Information Sciences*, 191, 3-13.

133. Hsu, M. F., and Pai, P. F. (2013). Incorporating support vector machines with multiple criteria decision making for financial crisis analysis. *Quality & Quantity*, 47(6), 3481-3492.

134. Peteiro-Barral, D., Remeseiro, B., Méndez, R., and Penedo, M. G. (2017). Evaluation of an automatic dry eye test using MCDM methods and rank correlation. *Medical & Biological Engineering & Computing*, 55(4), 527-536.

135. Peng, Y., Kou, G., Wang, G., and Shi, Y. (2011). FAMCDM: A fusion approach of MCDM methods to rank multiclass classification algorithms. *Omega*, 39(6), 677-689.

136. Chattopadhyay, M., and Mitra, S. K. (2017). Applicability and effectiveness of classifications models for achieving the twin objectives of growth and outreach of microfinance institutions. *Computational and Mathematical Organization Theory*, 23(4), 451-474.

137. Kartal, H., Oztekin, A., Gunasekaran, A., and Cebi, F. (2016). An integrated decision analytic framework of machine learning with multi-criteria decision making for multi-attribute inventory classification. *Computers & Industrial Engineering*, 101, 599-613.

138. Jaya, Y. B. J., and Tamilselvi, J. J. (2015). Fuzzified MCDM consistent ranking feature selection with hybrid algorithm for credit risk assessment. *Research Journal of Applied Sciences, Engineering and Technology*, 11(12), 1397-1403.

139. Saaty, T. L. (1980). *The analytic hierarchy process: planning, priority setting, resource allocation*. New York: McGraw Hill.

140. Saghapour, E., Kermani, S., and Sehhati, M. (2017). A novel feature ranking method for prediction of cancer stages using proteomics data. *PloS One*, 12(9), e0184203.

141. Nguyen, T., and Nahavandi, S. (2016). Modified AHP for gene selection and cancer classification using type-2 fuzzy logic. *IEEE Transactions on Fuzzy Systems*, 24(2), 273-287.

142. Abd-el Fattah, I. M., Khedr, W. I., and Sallam, K. M. (2013). A TOPSIS based method for gene selection for cancer classification. *International Journal of Computer Applications*, 67(17), 39-44.

143. Zhang, Y., Yang, A., Xiong, C., Wang, T., and Zhang, Z. (2014). Feature selection using data envelopment analysis. *Knowledge-Based Systems*, 64, 70-80.

144. Zheng, Z., and Padmanabhan, B. (2007). Constructing ensembles from data envelopment analysis. *INFORMS Journal on Computing*, 19(4), 486-496.

145. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12), 6745-6750.

146. Chandra, B., and Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, 44(4), 529-535.

147. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769), 503-511.

148. Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory* (Second Edition). Hoboken, NJ: John Wiley & Sons, 5-25.

149. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423.

150. Kwak, N., and Choi, C. H. (2002). Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 13(1), 143-159.

151. Pascoal, C., Oliveira, M. R., Pacheco, A., and Valadas, R. (2017). Theoretical evaluation of feature selection methods based on mutual information. *Neurocomputing*, 226, 168-181.

152. Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4), 537-550.

153. Cheng, H., Qin, Z., Feng, C., Wang, Y., and Li, F. (2011). Conditional mutual information-based feature selection analyzing for synergy and redundancy. *Etri Journal*, 33(2), 210-218.

154. Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5, 1531-1555.

155. McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19(2), 97-116.

156. Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1), 37-66.

157. Guo, G., Wang, H., Bell, D., Bi, Y., and Greer, K. (2003). *KNN model-based approach in classification*. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE", Springer, 986-996.

158. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

159. Domingos, P., and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3), 103-130.

160. Tharwat, A. (In press). Classification assessment methods. *Applied Computing and Informatics*.

161. Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4), 427-437.

162. Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation. In A. Sattar and B. Kang (Eds.), *Lecture Notes in Computer Science: Vol. 4304*. *AI 2006: Advances in artificial intelligence.* Berlin, Heidelberg: Springer, 1015-1021.

163. Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27-38.

164. Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS One*, 12(6), e0177678.

165. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2), 442-451.

166. Wang, H., Khoshgoftaar, T. M., and Napolitano, A. (2014). *Choosing the best classification performance metric for wrapper-based software metric selection for defect prediction*. Proceedings of the 26th International Conference on Software Engineering and Knowledge Engineering, KSI Research, 540-545.

167. Japkowicz, N. (2013). Assessment metrics for imbalanced learning. In H. He and Y. Ma (Eds.), *Imbalanced learning: foundations, algorithms, and applications*. Hoboken, NJ: John Wiley & Sons, 187-206.

168. Linden, A. (2006). Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice*, 12(2), 132-139.

169. Flach, P. A., and Kull, M. (2015). *Precision-recall-gain curves: PR analysis done right*. Proceedings of the 29th Annual Conference on Neural Information Processing Systems, Neural Information Processing Systems, 838-846.

170. Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. (2010). *A comparison of AUC estimators in small-sample studies*. Proceedings of the Third International Workshop on Machine Learning in Systems Biology, Microtome Publishing, 3-13.

171. Davis, J., and Goadrich, M. (2006). *The relationship between Precision-Recall and ROC curves*. Proceedings of the 23rd International Conference on Machine Learning, ACM, 233-240.

172. Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European Journal of Operational Research*, 48(1), 9-26.

173. Opricovic, S. (1998). *Multicriteria Optimization of Civil Engineering Systems,* PhD Thesis, Faculty of Civil Engineering, Belgrade.

174. Opricovic, S., and Tzeng, G. H. (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research*, 156(2), 445-455.

175. Hansen, L. K., and Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), 993-1001.

176. Dietterich, T. G. (1997). Machine learning research: four current directions. *Artificial Intelligence Magazine*, 18(4), 97-136.

177. Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). *Robust feature selection using ensemble feature selection techniques*. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer-Verlag, 313-325.

178. Opricovic, S. (2009). A compromise solution in water resources planning. *Water Resources Management*, 23(8), 1549-1561.

179. Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, 20(15), 2479-2481.

180. Moradkhani, M., Amiri, A., Javaherian, M., and Safari, H. (2015). A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm. *Applied Soft Computing*, 35, 123-135.

181. Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2015). Distributed feature selection: an application to microarray data classification. *Applied Soft Computing*, 30, 136-150.

182. Salem, H., Attiya, G., and El-Fishawy, N. (2017). Classification of human cancer diseases by gene expression profiles. *Applied Soft Computing*, 50, 124-134.

183. Medjahed, S. A., Saadi, T. A., Benyettou, A., and Ouali, M. (2017). Kernel-based learning and feature selection analysis for cancer diagnosis. *Applied Soft Computing*, 51, 39-48.

184. Sun, L., Zhang, X. Y., Qian, Y. H., Xu, J. C., Zhang, S. G., and Tian, Y. (2019). Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. *Applied Intelligence*, 49(4) 1245-1259.

185. Vinh, N. X., Zhou, S., Chan, J., and Bailey, J. (2016). Can high-order dependencies improve mutual information based feature selection?. *Pattern Recognition*, 53, 46-58.

186. Zhang, Z., Tian, Y., Bai, L., Xiahou, J., and Hancock, E. (2017). High-order covariate interacted Lasso for feature selection. *Pattern Recognition Letters*, 87, 139-146.

187. Zhang, P., Gao, W., and Liu, G. (2018). Feature selection considering weighted relevancy. *Applied Intelligence*, 48(12), 4615-4625.

188. Pamuk, M. G., and Atacak, I. (2018). *Evaluation of machine learning-based schemes by using integrated AHP-VIKOR method: a case study in determination of breast cancer prognosis*. Proceedings of the International Conference on Innovative Engineering Applications, CIEA, 667-676.

EKLER

EK-1. Kolon tümörü verisine ait karar matrisi

Çizelge 1.1. Kolon tümörü veri kümesi için tahmin modellerinin kriterler altında ölçülen performans değerleri

		Kriterler				
Alternatifler	$\|S\|$	K_1	K_2	K_3	K_4	K_5
A_1	12	0,9700	0,9739	0,8591	0,9355	0,9355
A_2	48	0,9620	0,9739	0,8959	0,9516	0,9518
A_3	49	0,9159	0,9364	0,8043	0,9032	0,9047
A_4	11	0,9607	0,9585	0,8584	0,9355	0,9347
A_5	59	0,9112	0,9375	0,8172	0,9032	0,9051
A_6	50	0,9673	0,9784	0,8640	0,9355	0,9361
A_7	52	0,9411	0,9540	0,8261	0,9194	0,9197
A_8	25	0,9839	0,9886	0,9295	0,9677	0,9677
A_9	79	0,9172	0,9420	0,8043	0,9032	0,9047
A_{10}	18	0,9721	0,9773	0,8591	0,9355	0,9355
A_{11}	5	0,9872	0,9864	0,8232	0,9194	0,9179
A_{12}	25	0,9839	0,9818	0,8043	0,9032	0,9047
A_{13}	18	0,9821	0,9795	0,7564	0,8871	0,8876
A_{14}	10	0,9896	0,9886	0,8335	0,9194	0,9204
A_{15}	14	0,9782	0,9750	0,7564	0,8871	0,8876
A_{16}	50	0,9857	0,9841	0,7564	0,8871	0,8876
A_{17}	87	0,9471	0,9364	0,7256	0,8710	0,8721
A_{18}	51	0,9795	0,9761	0,7648	0,8871	0,8885
A_{19}	52	0,9760	0,9716	0,7360	0,8710	0,8729
A_{20}	17	0,9896	0,9886	0,8591	0,9355	0,9355

EK-2. Lenfoma verisine ait karar matrisi

Çizelge 2.1. Lenfoma veri kümesi için tahmin modellerinin kriterler altında ölçülen performans değerleri

		Kriterler				
Alternatifler	$\|S\|$	K_1	K_2	K_3	K_4	K_5
A_1	7	0,9979	0,9990	0,9306	0,9479	0,9453
A_2	102	1,00	1,00	1,00	1,00	1,00
A_3	109	0,9772	0,9915	0,9689	0,9792	0,9778
A_4	25	1,00	1,00	1,00	1,00	1,00
A_5	186	0,9651	0,9887	0,9496	0,9688	0,9593
A_6	12	0,9900	0,9977	0,9543	0,9688	0,9678
A_7	168	0,9422	0,9813	0,9189	0,9375	0,9381
A_8	22	0,9713	0,9910	0,9596	0,9792	0,9690
A_9	125	0,9846	0,9947	0,9689	0,9792	0,9778
A_{10}	53	0,9843	0,9945	0,9839	0,9896	0,9879
A_{11}	103	1,00	1,00	1,00	1,00	1,00
A_{12}	29	1,00	1,00	1,00	1,00	1,00
A_{13}	72	0,9998	0,9998	0,9628	0,9688	0,9715
A_{14}	21	1,00	1,00	0,9887	0,9896	0,9891
A_{15}	92	0,9951	0,9996	0,9236	0,9375	0,9370
A_{16}	151	0,9998	0,9998	0,9566	0,9688	0,9699
A_{17}	198	0,9872	0,9955	0,9065	0,9271	0,9346
A_{18}	11	0,9755	0,9952	0,9059	0,9271	0,9264
A_{19}	66	0,9996	0,9996	0,9714	0,9792	0,9804
A_{20}	57	0,9996	0,9996	0,9715	0,9792	0,9804

EK-3. Kolon ve lenfoma verileri üzerinde kriterlere göre alternatiflerin sıralanışı

Çizelge 3.1. Kolon tümörü ve lenfoma veri kümeleri için tahmin modellerinin 5 ayrı kriter altında elde edilen sıralama sonuçları

Kolon tümörü verisi					Lenfoma verisi				
Kriterler					Kriterler				
K_1	K_2	K_3	K_4	K_5	K_1	K_2	K_3	K_4	K_5
A_{14}^{*}	A_8^{*}	A_8	A_8	A_8	A_2^{*}	A_2^{*}	A_2^{*}	A_2^{*}	A_2^{*}
A_{20}^{*}	A_{14}^{*}	A_2	A_2	A_2	A_4^{*}	A_4^{*}	A_4^{*}	A_4^{*}	A_4^{*}
A_{11}	A_{20}^{*}	A_6	A_1^{*}	A_6	A_{11}^{*}	A_{11}^{*}	A_{11}^{*}	A_{11}^{*}	A_{11}^{*}
A_{16}	A_{11}	A_1^{*}	A_4^{*}	A_1^{*}	A_{12}^{*}	A_{12}^{*}	A_{12}^{*}	A_{12}^{*}	A_{12}^{*}
A_8^{*}	A_{16}	A_{10}^{*}	A_6^{*}	A_{10}^{*}	A_{14}^{*}	A_{14}^{*}	A_{14}	A_{10}^{**}	A_{14}
A_{12}^{*}	A_{12}	A_{20}^{*}	A_{10}^{*}	A_{20}^{*}	A_{13}^{**}	A_{13}^{**}	A_{10}	A_{14}^{**}	A_{10}
A_{13}	A_{13}	A_4	A_{20}^{*}	A_4	A_{16}^{**}	A_{16}^{**}	A_{20}	A_3^{*}	A_{19}^{*}
A_{18}	A_6	A_{14}	A_7^{**}	A_{14}	A_{19}^{*}	A_{15}^{*}	A_{19}	A_8^{*}	A_{20}^{*}
A_{15}	A_{10}	A_7	A_{11}^{**}	A_7	A_{20}^{*}	A_{19}^{*}	A_3^{*}	A_9^{*}	A_3^{**}
A_{19}	A_{18}	A_{11}	A_{14}^{**}	A_{11}	A_1	A_{20}^{*}	A_9^{*}	A_{19}^{*}	A_9^{**}
A_{10}	A_{15}	A_5	A_3^{*}	A_5	A_{15}	A_1	A_{13}	A_{20}^{*}	A_{13}
A_1	A_1^{*}	A_3^{*}	A_5^{*}	A_3^{*}	A_6	A_6	A_8	A_5^{**}	A_{16}
A_6	A_2^{*}	A_9^{*}	A_9^{*}	A_9^{*}	A_{17}	A_{17}	A_{16}	A_6^{**}	A_8
A_2	A_{19}	A_{12}^{*}	A_{12}^{*}	A_{12}^{*}	A_9	A_{18}	A_6	A_{13}^{**}	A_6
A_4	A_4	A_{18}	A_{13}^{**}	A_{18}	A_{10}	A_9	A_5	A_{16}^{**}	A_5
A_{17}	A_7	A_{13}^{*}	A_{15}^{**}	A_{13}^{*}	A_3	A_{10}	A_1	A_1	A_1
A_7	A_9	A_{15}^{*}	A_{16}^{**}	A_{15}^{*}	A_{18}	A_3	A_{15}	A_7^{*}	A_7
A_9	A_5	A_{16}^{*}	A_{18}^{**}	A_{16}^{*}	A_8	A_8	A_7	A_{15}^{*}	A_{15}
A_3	A_3^{*}	A_{19}	A_{17}^{*}	A_{19}	A_5	A_5	A_{17}	A_{17}^{**}	A_{17}
A_5	A_{17}^{*}	A_{17}	A_{19}^{*}	A_{17}	A_7	A_7	A_{18}	A_{18}^{**}	A_{18}

EK-4. Bütünleşik AHS-VIKOR yöntemi ile hesaplanan indeks puanları

Çizelge 4.1. Kolon tümörü ve lenfoma verileri üzerinde uygulanan her bir tahmin modeli için karar süreci sonunda elde edilen *S*, *R* ve *Q* değerleri

	Veri Kümeleri					
	Kolon			Lenfoma		
Alternatifler	S	R	Q	S	R	Q
A_1	0,2947	0,0921	0,2621	0,3361	0,1522	0,3858
A_2	0,2582	0,1297	0,2945	0,00	0,00	0,00
A_3	0,8244	0,3462	0,9634	0,3715	0,1453	0,3952
A_4	0,4009	0,1358	0,3921	0,00	0,00	0,00
A_5	0,8286	0,3683	0,9984	0,5650	0,2224	0,6029
A_6	0,2840	0,1048	0,2740	0,2840	0,1002	0,2874
A_7	0,5789	0,2278	0,6375	0,9384	0,3683	1,0000
A_8	0,0268	0,0268	0,00	0,4481	0,1829	0,4870
A_9	0,7962	0,3401	0,9369	0,2891	0,0981	0,2872
A_{10}	0,2714	0,0822	0,2332	0,2297	0,1000	0,2582
A_{11}	0,2392	0,1076	0,2504	0,00	0,00	0,00
A_{12}	0,3255	0,1267	0,3320	0,00	0,00	0,00
A_{13}	0,4293	0,1752	0,4675	0,1743	0,0816	0,2036
A_{14}	0,2059	0,0972	0,2144	0,0566	0,0248	0,0638
A_{15}	0,4654	0,1752	0,4899	0,3907	0,1676	0,4357
A_{16}	0,3942	0,1752	0,4457	0,1902	0,0952	0,2306
A_{17}	0,8313	0,2064	0,7630	0,5429	0,2051	0,5677
A_{18}	0,4454	0,1667	0,4651	0,6343	0,2064	0,6182
A_{19}	0,5449	0,1959	0,5696	0,1300	0,0627	0,1545
A_{20}	0,1445	0,0713	0,1383	0,1298	0,0625	0,1540

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı	: PAMUK, Meryem Gülşah
Uyruğu	: T.C.
Doğum tarihi ve yeri	: 10.08.1991, Mersin
Medeni hali	: Bekâr
Telefon	: 0 (506) 827 59 92
e-mail	: gulsahmeryempamuk.24@gmail.com

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi / Bilgisayar Müh.	Devam ediyor
Lisans	Selçuk Üniversitesi / Bilgisayar Müh.	2014
Lise	İçel Anadolu Lisesi	2009

İş Deneyimi

Yıl	Yer	Görev
-	-	-

Yabancı Dil

İngilizce

Yayınlar

1. Pamuk, M. G., and Atacak, I. (2018). *Evaluation of machine learning-based schemes by using integrated AHP-VIKOR method: a case study in determination of breast cancer prognosis*. Proceedings of the International Conference on Innovative Engineering Applications, CIEA, 667-676.

Hobiler

Tarihi ve kurgu romanlar, edebiyat

GAZİ GELECEKTİR..