



**KARACİĞER HASTALIKLARI İLE İLİŞKİLİ BELİRLİ GAZLARIN
ELEKTRONİK BURUN TEKNOLOJİSİ İLE TESPİTİ**

Mehmet Kerem ÖZEZEN

**YÜKSEK LİSANS TEZİ
İLERİ TEKNOLOJİLER ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FENBİLİMLERİ ENSTİTÜSÜ**

OCAK 2022

ETİK BEYAN

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

- 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 çalışmada 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.

Mehmet Kerem ÖZEZEN

13/01/2022

KARACİĞER HASTALIKLARI İLE İLİŞKİLİ GAZLARIN ELEKTRONİK BURUN TEKNOLOJİSİ İLE TESPİTİ

(Yüksek Lisans Tezi)

Mehmet Kerem ÖZEZEN

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Ocak 2022

ÖZET

Sağlık alanında yüzyıllardır hekimlerin başvurdukları koku ile teşhis yöntemi, son yıllarda mühendislik alanında oldukça ilgi odağı hâline gelmiş ve biyolojik olarak hayvanların koku alma sistemlerinden ilham alan elektronik burunlar geliştirilmiştir. Literatürde, başlıca sağlık, gıda sektörü, sanayi ve askeri çalışmalar başta olmak üzere çok çeşitli alanlar için elektronik burunlar önerilmiştir. Bununla birlikte, insan nefesindeki organik uçucuların analizi öncü çalışmalar olmaya devam etmektedir. Bu çalışmada, insan nefesindeki konsantrasyon miktarları, alkole bağlı olmayan yağlı karaciğer hastalığı ve karaciğer sirozu ile ilişkilendirilmiş, C_3H_6O (aseton), C_4H_8O (2-bütanon) ve C_2H_6S (dimetil sülfid) analitleri, kuvarz kristal mikrobalsans sensörleri ile tespit edilmeye çalışılmıştır. Sensörlerin zamana bağlı frekans değişimlerini matematiksel olarak modelleyen makine öğrenmesi algoritmaları için veri setleri oluşturulmuş ve çeşitli algoritmaların performansları test edilmiştir. Sonuçlarına göre en yüksek doğruluk oranına sahip Lojistik Regresyon, XGBoost, Destek Vektör Makineleri ve Rastgele Orman algoritmaları ile bir topluluk modeli (ensemble model) kurulmuş ve bu model ile, C_3H_6O için AUC:0.97, F1 Skoru: 0.87, C_4H_8O için AUC:0.97, F1 Skoru: 0.89, C_2H_6S için ise AUC:0.99, F1 Skoru:0.91 başarısı elde edilmiştir. Bu çalışmada, analitlerin ppm değerleri, sınıflandırma sonuçlarına göre sırasıyla RMSE: 2.396874, 2.853136 ve 2.775164 hata oranlarına sahip bir regresyon işlemi ile hesaplanmıştır.

Bilim Kodu : 92432
Anahtar Kelimeler : E-Burun, QCM Sensörü, Makine öğrenmesi, NAFLD, Siroz
Sayfa Adedi : 99
Danışman : Prof. Dr. Selim ACAR

DETECTION OF CERTAIN GASES ASSOCIATED WITH LIVER DISEASES WITH
ELECTRONIC NOSE TECHNOLOGY

(M.Sc. Thesis)

Mehmet Kerem ÖZEZEN

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

January 2022

ABSTRACT

The olfactory diagnosis method, which has been used by mediciners for centuries in the field of health, has become the focus of attention for engineering in recent years, and electronic noses that are biologically inspired by the olfactory systems of animals have been developed. In the literature, electronic noses have been proposed for a wide variety of fields, mainly in health, food sector, industry and military sevicees. However, the analysis of organic volatiles in human breath remains pioneering work. In this study, concentrations of C_3H_6O (acetone), C_4H_8O (2-butanone) and C_2H_6S (dimethyl sulfide) analytes in human breath were associated with non-alcoholic fatty liver disease and liver cirrhosis, attempted to be detected with quartz crystal microbalance sensors. Data sets were created for machine learning algorithms that mathematically model the time-dependent frequency changes of sensors and the performances of various algorithms were tested. According to the results, an ensemble model was established with Logistic Regression, XGBoost, Support Vector Machines and Random Forest algorithms with the highest score, and with this model the success rate, AUC: 0.97, F1 Score: 0.87 for C_3H_6O , AUC: 0.97, F1 Score: 0.89 for C_4H_8O , and AUC: 0.99, F1 Score: 0.91 for C_2H_6S was captured. In this study, the ppm values of the analytes were calculated by a regression process with RMSE: 2.396874, 2.853136 and 2.775164 error rates, respectively, according to the classification results.

Science Code : 92432

Key Words : E-Nose, QCM Sensors, Machine learning, NAFLD, Cirrhosis

Page Number : 99

Supervisor : Prof. Dr. Selim ACAR

TEŞEKKÜR

Yüksek lisans eğitimim süresince, bilgi ve tecrübesi ile desteğini esirgemeyen, birlikte çalışma imkanı sunan ve kariyer hedeflerimin tek bir alana kanalize edilmesine yardımcı olan saygı değer hocam Prof. Dr. Selim ACAR’a teşekkürlerimi sunarım. Bu çalışmada kullanılan sensörlerin geliştirilmesi rolünü üstlenen Sayın Zafer ŞEN’e, çalışmalarım süresince değerli fikirleriyle bakış açımı genişleten Murat ARTUÇ’a, verdikleri destek için değerli arkadaşlarım; Ali Orkun ÇAĞIRTEKİN, Ahmad AJJAQ ve Özlem BARİN ‘e teşekkür ederim. Ayrıca bu çalışmayı bitirebilmem için maddi, manevi her türlü desteğini esirgemeyen, bu günlere gelmeme vesile olan annem Muna YÜCEOL ÖZEZEN, babam Halil ÖZEZEN, anneannem Meliha YÜCEOL, dedem Adil YÜCEOL ve sevgili eşim Gülçin TÜRKKAN’a sonsuz teşekkür ederim. Bu tez çalışmasına, 18/2020-02 No’lu proje ile Gazi Üniversitesi, Bilimsel Araştırma Projesi Destek Birimi tarafından kaynak oluşturulmuştur. Destekleri için teşekkür ederim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	xi
RESİMLERİN LİSTESİ	xiv
SİMGELER VE KISALTMALAR.....	xv
1. GİRİŞ.....	1
2. SENSÖRLER	7
2.1. Elektriksel Değişim Tabanlı Gaz Sensörler (Kimyasal Direnç Sensörleri)	7
2.1.1. Metal oksit yarı iletken sensörü	8
2.1.2. İletken polimer yapılı sensörler.....	9
2.1.3. Nano-malzeme yapılı sensörler.....	9
2.2. Optik ve Kalorimetrik Değişimlere Dayalı Sensörler	10
2.2.1. Optik sensör.....	10
2.2.2. Kolorimetrik sensör.....	13
2.3. Akustik Varyasyonlara Dayalı Sensörler	14
2.3.1. Yüzey akustik dalga sensörü	14
2.3.2. Kuvars kristal mikrobalsans sensörü	15
3. MAKİNE ÖĞRENMESİ.....	21
3.1. Giriş.....	21
3.1.1. Sınıflandırma ve regresyon	21
3.1.2. Karışıklık matrisi.....	22

	Sayfa
3.1.3. Sensitivity ve specificity kavramları	23
3.1.4. ROC ve AUC grafikleri.....	24
3.1.5. Bias ve Variance.....	25
3.1.6. Odds/log-odds	26
3.1.7. Keşifçi veri analizi (EDA).....	27
3.1.8. Validasyon (doğrulama) stratejileri.....	27
3.1.9. Veri sızıntıları.....	27
3.1.10. Veri kodlaması (encoding)	28
3.1.11. Dengesiz veri setleri	30
3.1.12. Ensembling (topluluk öğrenmesi)	30
3.2. Denetimli (Supervised) Öğrenme	31
3.2.1. Karar ağacı tabanlı olmayan algoritmalar	32
3.2.2. Karar ağacı tabanlı algoritmalar	42
4. MATERYAL VE YÖNTEM	65
4.1. Kuvarz Kristal Mikrobalans Sensörleri.....	65
4.2. Veri Toplama İşlemi	66
5. ARAŞTIRMA BULGULARI	71
5.1. Sensör Tepkileri	71
5.2. Makine Öğrenmesi Algoritmalarının Sonuçları.....	76
6. SONUÇ	85
KAYNAKLAR	89
ÖZGEÇMİŞ	99

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 1.1. Bazı hastalıklar ve metabolik rahatsızlıkların kimyasal gazlar ve aromalar ile ilişkileri.....	3
Çizelge 1.2. Tespit edilen bileşiklerin, ppbv cinsinden tanımlayıcı istatistikleri	4
Çizelge 3.1. Karışıklık matrisi	22
Çizelge 3.2. Karışıklık matrisi örnek gösterimi	23
Çizelge 3.3. Etiket kodlaması tekniğine bir örnek.....	29
Çizelge 3.4. One-Hot kodlama tekniğine bir örnek	29
Çizelge 3.5. İkili kodlama tekniğine bir örnek	30
Çizelge 3.6. Örnek bir veri seti	33
Çizelge 3.7. Örnek bir karar ağacı parametreleri.....	42
Çizelge 3.8. Örnek veriler.....	46
Çizelge 3.9. Sıralandırılmış örnek verileri.....	46
Çizelge 3.10. Ortalamaları alınmış örnek veriler.....	46
Çizelge 3.11. Bootstrapped veri seti oluşturmak için kullanılacak orjinal veri seti	54
Çizelge 3.12. Bootstrapped veri seti	54
Çizelge 3.13. Skor tablosu.....	54
Çizelge 3.14. Out of Bag olarak tanımlanan verileri içeren tablo	55
Çizelge 3.15. Out of Bag skor tablosu.....	55
Çizelge 3.16. $\{(x_i, y_i)\}_{i=1}^n$ olarak ifade edilen basit bir veri setine ait tablo	56
Çizelge 3.17. Veri seti üzerinde hesaplanan ilk artakalan değerlerini gösteren tablo	57
Çizelge 3.18. Veri seti üzerinde hesaplanan $m=2$ için artakalan değerlerini gösteren tablo	59
Çizelge 3.19. Sınıflandırma için örnek veri setini gösteren tablo.....	60
Çizelge 3.20. Sınıflandırma için artakalanları gösteren tablo.....	62
Çizelge 4.1. QCM sensörlerinde kullanılan polimer yapılı bileşikler	65
Çizelge 4.2. Analitlere ait Antonie katsayıları.....	66

Çizelge	Sayfa
Çizelge 4.3. MFC'ye ait pin bilgileri.....	67
Çizelge 5.1. Algoritmaların doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru değerleri	79

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. a) n-tipi MOS ve b) p-tipi MOS'un gaz reaktivitesi sonucu oluşan direnç değişimini gösteren şema	8
Şekil 2.2. Fiber optik sensörlerin şematik görünümü	11
Şekil 2.3. Bragg yasasının şematik gösterimi	13
Şekil 2.4. Bir kalorimetrik sensörün hedef gaz ile teması sonrası oluşan renk değişimi	14
Şekil 2.5. a) Tek portlu SAW rezonatörünün; b) bir kimyasal tanıma katmanına sahip iki bağlantı noktalı bir SAW rezonatörünün tasarımı	15
Şekil 2.6. a) Doğal formunda bir kuvarz kristali, b) Kalınlık shear modu	16
Şekil 2.7. Kristal plaka kesimleri	17
Şekil 2.8. a) Moleküllerin kristal tarafından tutulmasıyla oluşan kütle değişimi bilgisinin frekans olarak hesaplanmasının temsili gösterimi ve b) altın elektrodlu bir kuvarz kristali	18
Şekil 2.9. a) Osilasyonun temel ve b) üçüncü modunun temsili gösterimi	18
Şekil 2.10. Butterworth - van Dyke (BvD) eşdeğer devresi	20
Şekil 3.1. a) Sınıflandırma problemi, b) Regresyon problemi	21
Şekil 3.2. Makine öğrenmesi algoritmalarının ROC grafiği üzerindeki temsili gösterimi	24
Şekil 3.3. AUC grafiği temsili gösterimi	25
Şekil 3.4. Bias ve variance kavramlarının gösterimi	26
Şekil 3.5. Denetimli makine öğrenmesi modeli için oluşturulan eğitim seti	32
Şekil 3.6. İki değişkenli doğrusal regresyon modeli	34
Şekil 3.7. Sinir ağı modeli	39
Şekil 3.8. Aktivasyon fonksiyonları	41
Şekil 3.9. Karar Ağacı yapısının temsili gösterimi	42
Şekil 3.10. Programalama dillerine göre veri bilimci olan ya da olmayan kişi sayıları ifade edilmiştir	43
Şekil 3.11. Karar ağacı gösterimi	44

Şekil	Sayfa
Şekil 3.12. Örnek karar ağacı oluşum süreci	44
Şekil 3.13. Örnek karar ağacı oluşum süreci	45
Şekil 3.14. Örnek karar ağacının son hâli.....	45
Şekil 3.15. Örnek-2 karar ağacı	47
Şekil 3.16. Örnek veri setinin artakalan değerleri	49
Şekil 3.17. Örnek olarak kullanılacak farazi bir deneyin sıcaklığa göre başarısını gösteren grafik.....	50
Şekil 3.18. Veri setleri verilen deneye ait karar ağacı	50
Şekil 3.19. Sınırlandırmalar sonucu oluşan karar ağaçları	51
Şekil 3.20. Veri Setleri verilen deneye ait karar ağacı ve artakalan değerleri.....	52
Şekil 3.21. Veri seti ile kurulan artakalanları gösteren basit bir regresyon ağacı	57
Şekil 3.22. Veri Seti ile kurulan artakalanları gösteren basit bir regresyon ağacının son hâli	58
Şekil 3.23. Veri seti ile kurulan $m=2$ için artakalanları gösteren basit bir regresyon ağacı.....	59
Şekil 3.24. Veri seti ile kurulan $m=2$ için artakalanları gösteren basit bir regresyon ağacının son hâli	59
Şekil 3.25. Artakalanlar ile oluşturulan bir regresyon ağacı	62
Şekil 4.1. Gaz ölçüm düzeneği	67
Şekil 4.2. MFC kontrol düzeneği.....	68
Şekil 4.3. Sıvı seviyesi kontrol sistemi ve solenoid valf devresi.....	68
Şekil 5.1. Sensör 1'in analitlere karşı verdiği tepki.....	72
Şekil 5.2. Sensör 2'nin analitlere karşı verdiği tepki.....	73
Şekil 5.3. Sensör 3'ün analitlere karşı verdiği tepki.....	73
Şekil 5.4. Sensör 4'ün analitlere karşı verdiği tepki.....	74
Şekil 5.5. Sensör 5'in analitlere karşı verdiği tepki.....	74
Şekil 5.6. Sensör 6'nın analitlere karşı verdiği tepki.....	75
Şekil 5.7. Sensör 7'nin analitlere karşı verdiği tepki.....	75

Şekil	Sayfa
Şekil 5.8. Sensör 8'in analitlere karşı verdiği tepki.....	76
Şekil 5.9. Karışıklık matrisleri.....	78
Şekil 5.10. Aseton analitine ait ROC (reciever operator characteristics curve) grafiği..	80
Şekil 5.11. 2-bütanon analitine ait ROC (reciever operator characteristics curve) grafiği	80
Şekil 5.12. Dimetil sülfid analitine ait ROC (reciever operator characteristics curve) grafiği	81
Şekil 5.13. Veri analizi ve makine öğrenmesi süresince atılan adımları ifade eden akış şeması	81
Şekil 5.14. Aseton anlitine ait regresyon işlemi	82
Şekil 5.15. 2-bütanon anlitine ait regresyon işlemi	83
Şekil 5.16. Dimetil sülfid anlitine ait regresyon işlemi	83
Şekil 5.17. Regresyon işlemlerine tabi tutulan analitlerine ait MSE hata oranlarının kutu grafiği ile gösterimi	84

RESİMLERİN LİSTESİ

Resim	Sayfa
Resim 4.1. QCM sensör birimi	66

SİMGELER VE KISALTMALAR

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

Simgeler	Açıklamalar
$\dot{u}_x$	Hız
η_q	Efektif Viskozite
C_2H_4	Etilen
C_2H_6S	Dimetil Sülfid
C_3H_6O	Aseton
C_4H_8O	2-Bütanon
R_{jm}	Terminal Bölgesi
V_q	Kesme Dalga Hızı
Z_L	Yük Empedans
cm^2	Santimetre Kare
f_0	Temel Frekans
h_q	Kristalin Kalınlığı
$h_0(x)$	Bağımlı Değişken
ϵ_{22}	Dielektrik Geçirgenliği
θ_m	Girdi Parametresi
μ_q	Kesme Sertliği
ρ_q	Kuvarz Yoğunluğu
Δ	Değişim
$^{\circ}C$	Celcius
A	Amper, Piezoelektrik Aktif Alanı
A, B, C	Antonie Katsayıları
b	Bias
cm	Santimetre
d	Kristal Katmanların Aralığı, Veri Girdi Sayısı
dyne	Kuvvet Birimi
f(x)	Fonksiyon

Simgeler**Açıklamalar**

$J(\theta)$	Maliyet Fonksiyonu
K	Kuvarz Elektromekanik Bağlantı Katsayısı, Kelvin
$L(y, f(x))$	Kayıp Fonksiyonu
M	Karar Ağacı Sayısı
m	Kütle
mA	Mili Amper
N	Rezonans Modu
P	Olasılık
ppb	Milyarda Parça
ppm	Milyonda Parça
R	Direnç
V	Voltaj
W	Ağırlık Matrisi
X	Girdi Verisi
Y	Çıktı Verisi
λ	X Işınının Dalga Boyu, Akustik Dalga Boyu, Düzenleştirme Katsayısı
CH_4	Metan
CO_2	Karbon Di-Oksit
CO	Karbon Mono-Oksit
NO_2	Azot Di-Oksit
Q	Işın Açısı
Γ	Kristalin Yarı Bant Genişliği
γ	Olasılıksal Değer
σ	Stres

Kısaltmalar**Açıklamalar**

AC	Alternatif Akım
ADC	Alternatif-Direkt Akım Çevirici
AE	Auto-Encoder (Oto-Kodlayıcı)
AI	Artificial Intelligence (Yapay Zeka)

Kısaltmalar**Açıklamalar****ANN, YSA, SNN**

Yapay Sinir Ağları (Algoritma)

AT, BT, SC

Kristal Plaka Kesimi

AUC

Area Under Curve (Eğri Altında Kalan Alan)

CNT

Karbon Nano Tüp

DAC

Direkt-Alternatif Akım Çevirici

DC

Direkt Akım

DT

Decision Tree (Karar Ağacı)

EDA

Exploratory Data Analysis (Keşifçi Veri Analizi)

GBM

Gradient Boosting Machine (Algoritma)

GNB

Gaussian Naive Bayes (Algoritma)

GO

Grafen Oksit

Hz

Hertz

IDT

Interdigital Transdüser

KHz

Kilo Hertz

KNN

K-En Yakın Komşu (Algoritma)

LC

Liver Cirrhosis (Karaciğer Sirozu)

LDA

Doğrusal Diskriminant Analizi (Algoritma)

LLE

Locally Linear Embedded (Algoritma)

MFC

Kütle Akış Kontrolcüsü

MHz

Mega Hertz

MOS

Metal Oksit Yarı İletken

NAFLD

Alkole Bağlı Olmayan Yağlı Karaciğer Hastalığı

Pani

Polianilin

PCA

Temel Bileşenler Analizi (Algoritma)

PhC

Fotonik Kristal

PID

Oransal İntegral Türevsel Denetleyici

PPy

Polipinol

PTh

Politiyofen

PVC

Polivinil Klorür

PVP

Polivinil Prolidon

PWM

Sinyal Genişlik Modülasyonu

QCM

Kuarz Kristal Mikrobalans

Kısaltmalar**Açıklamalar****RF**

Radyo Frekans, Rastgele Orman (Algoritma)

rGO

İndirgenmiş Grafen Oksit

ROC

Receiver Operating Characteristic

SAW

Yüzey Akustik Dalga

SVM

Destek Vektör Makineleri (Algoritma)

t-SNE

t-Dağıtılmış Stokastik Komşu Gömme (Algoritma)

1. GİRİŞ

Elektronik burun (e-burun), canlıların koku alma sistemlerinden esinlenerek, farklı gaz konsantrasyonlarını algılamak ve ayırt etmek için tasarlanmış bir cihazdır. E-burun, gazları algılayan ve gaz etkileşimi nedeniyle oluşan direnç, kütle, vb. değişimlerini elektrik sinyallerine dönüştüren aktif gaz sensörlerinden ve her bir gazın belirleyici "parmak izi" sinyallerini anlamlı bir bilgiye dönüştüren örüntü tanıma sistemlerinden oluşur [1-4]. Bu örüntü tanıma sistemleri, Karar Ağaçları (DT), Gaussian Naive Bayes (GNB), Yapay sinir ağları (ANN) gibi "Makine Öğrenmesi" algoritmalarını içerir. E-burun teknolojisi bir çok araştırmacı için ilgi odağı olmuş ve bir çok çalışmada bu sistemler kullanılmıştır [5-10]. E-burun cihazları tarım, tıbbi teşhis, gıda kalite kontrolü, askeriye, kozmetik, vb. alanlarında yaygın olarak kullanılmaktadır [11-18]. E-burun üzerine yapılan çalışmalar, yeni sensörler geliştirme [19], donanım tasarlama [4,20,21], veri ön işleme ile örüntü tanıma algoritmaları uygulama ve geliştirme [22-24] olarak ayrılır.

Şen, Z.'nin Petrol Kaynaklı Hidrokarbonlardan BTEX Kompleks Gaz Karışımlarının Sınıflandırılması İçin QCM Gaz Sensör Dizisi Geliştirilmesi, başlıklı yüksek lisans tezinde organik gazların tespiti için kullanılan QCM sensörleri üzerine çalışılmıştır. Jetsprey metodu ile oksim, oksim kompleksleri, ftalosiyanın ve polimer ara yüzey malzemeleri kuvarz kristaller üzerine kaplanarak elde edilen sensörler üzerine BTEX (Benzen, Toluen, Etilbenzen ve Ksilen) ve interfere olarak farklı kimyasal gruplarından seçilen kloroform, etilasetat, trietilamin ve propanol gazları uygulanarak 4'er farklı konsantrasyonda ölçülmüştür. Ayrıca sensörlerin farklı oranlardaki nem tepkileri incelenmiştir. En iyi tepkiyi veren 8 adet sensör ile bir dizi oluşturulmuş ve uygulanan gazlara karşı oluşan frekans değişimlerinin PCA analizleri yapılmıştır [25].

Senyücel, N.'nin, Zararlı Kokuları Algılayıp İnternet Üzerinden İleten Elektronik Burun Tasarımı, başlıklı doktora tezinde önerilen elektronik burun, Kontrol Ünitesi, Hava Emme Ünitesi, Hava Tahliye Ünitesi ve Algılama Ünitesi bölümlerinden oluşur. Algılama Ünitesinde hava sıkıştırılarak sensörlerin tepki süresinin kısaltılması amaçlanmıştır. Metal oksit gaz sensörleri ile yanıcı ve patlayıcı gazların yoğunluk sıralaması yapılamadığı için uzman bir sistem geliştirilmiş ve analizin yapılabilmesi sağlanmıştır. Örüntü tanıma sistemi olarak yapay sinir ağları, gazların alt patlama sınırının tespiti amacıyla kullanılmıştır.

İnternet üzerinden cihazın durumu ve elde edilen sonuçlar yayınlanabilmiş ve cihaz kontrolleri gerçek zamanlı olarak yapılabilmiştir [26].

Zhang, L. ve arkadaşları, sistematik olarak elektronik burun teknolojisini tanımlamakta ve mevcut sorunlarının etkili ve verimli bir şekilde nasıl ele alınacağını gösteren belirli yöntemler, modeller ve teknikler sunmaktadır. Araştırmacıların gelecekte başlatabileceği ve başarabileceği en yeni algoritmik zorlukları inceleyen ve ele alan ilk kitaptır. Bu kitap özellikle dört algoritmik soruna odaklanmaktadır. Bölüm II'de, genel elektronik burun için, koku tanıma algoritmaları ve konsantrasyon tahmin algoritmaları dahil olmak üzere, kalıp tanıma ve makine öğrenimindeki algoritmik zorluklar açıklanmaktadır. Bölüm III'te, elektronik burun teknolojisindeki sensör sürüklenme sorunu, algoritmaların spesifik çözme şemaları ile tanıtılmakta ve açıklanmaktadır. Bölüm IV'te, ENT'nin genel kullanımını bozan parazit sorunları belirli yaklaşımlar ve tekniklerle tanıtılmakta ve tanımlanmaktadır. Son olarak, Bölüm V'de, ENT cihazlarının büyük ölçekli endüstriyel uygulamasını önleyen ayrıklık sorunu etkili modeller ve algoritmalar ile tanıtılmakta ve açıklanmaktadır [27].

Li, H. ve arkadaşları tarafından yapılan başka bir çalışmada, endüstriyel gazları tanımlamak için çekirdek ayırıcı analiz (KDA) algoritmasına dayanan bir yöntem sunmaktadır. Dört tipik endüstriyel gazın koku baskıları elektronik bir burunla toplanmıştır. Toplanan gazların ekstrakte edilen özellikleri, ana bileşen analizi (PCA), doğrusal diskriminant analizi (LDA), PCA + LDA ve KDA dahil olmak üzere farklı sınıflandırma algoritmaları kullanılarak gaz tanımlaması için kullanılmıştır. Daha iyi sınıflandırma sonuçları elde etmek için orijinal yüksek boyutlu verilerin boyutları düşürülmüş ve iyi bir sınıflandırıcı seçilmiştir. KDA algoritması, $c = 10$ çekirdek fonksiyonunun ofsetini ve $d = 5$ serbestlik derecesini seçerek % 100'lük yüksek bir sınıflandırma doğruluğu sağlamıştır. Bu doğruluğun PCA kullanılarak elde edilenlerden % 4.17 daha yüksek olduğu bulunmuştur. Standart sapma durumunda, KDA algoritması en yüksek tanıma oranına ve en az zaman tüketimine sahip olduğu belirlenmiştir [28].

Liu, H. ve arkadaşları farklı şarapların kokularını tespit etmek için metal oksit yarı iletken (MOS) sensörleri kullanılarak taşınabilir bir elektronik burun prototipi geliştirmiştir. Koku tespiti, üretim alanları, fermantasyon süreçleri dahil olmak üzere farklı özelliklere sahip şarapların ayrılmasını kolaylaştırır. Dört popüler makine öğrenme algoritması; XGBoost, RF, SVM ve BPNN farklı sınıflandırma görevleri için tanımlama modelleri oluşturmak için

kullanıldı. Deneyisel sonuçlar, BPNN'nin üretim alanlarını ve çeşitlerini belirlemede sırasıyla %94 ve %92,5 doğruluklarla en iyi performansı elde ettiğini göstermektedir; SVM, sırasıyla %67.3 ve %60.5 doğruluklarla, fermantasyon süreçlerini belirlemede en iyi performansı elde etti. Sonuçlar, optimal bir algoritmanın seçilmesinden sonra farklı şarapları özelliklerine göre ayırt etmek için kullanılabilen geliştirilmiş E-burnunun etkinliğini göstermektedir [1].

Tarihte hastalıklarının koku duyusuna bağlı olarak teşhis edilmesi, en güvenilir tedavi yöntemlerinden biri olarak kabul edilmektedir. Antik Yunan doktoru Hipokrat, MÖ 400 civarında hastalıkları teşhis etmek için 'koku sanatını' kullanmıştır. Bazı hastalıkların, fizyolojik veya metabolik bozuklukların, pozitif göstergeleri olarak belirtilen kimyasal gazlar ve aromalar listesi Çizelge 1.1'de verilmiştir.

Çizelge 1.1. Bazı hastalıklar ve metabolik rahatsızlıkların kimyasal gazlar ve aromalar ile ilişkileri

Hastalık	Vücut Kaynağı	Tanımlayıcı Gazlar ve Aromalar	Referanslar
Prerenal Azotemi	İdrar	Konsantre idrar kokusu ve kamembert krem peyniri	[29]
İdrar Yolları Enfeksiyonu	İdrar	Amonyak	[30]
Diabetik Ketoasidoz	Nefes	Çürümüş elma ve aseton	[31]
Diyabet	Nefes	Aseton	[30]
Fetor Hepaticus	Nefes	Yeni biçilmiş yonca	[32]
Kanserli Tümör	Nefes	Formaldehit	[33]
Akciğer Kanseri	Nefes	Hekzan, Metilpentan	[34]
Pseudomonas Enfeksiyonu	Deri ve Ter	Üzüm	[30]
Sıraca Hastalığı	Deri	Bayatlamış Bira	[35]
Üremi	Nefes	Monometilamin, Dimetilamin, Trimetilamin, Amonyak	[32,36-38]

Erken evre karaciğer hastalığı olan alkolsüz yağlı karaciğer hastalığının (NAFLD) ya da siroz, karaciğer fibrozu, karaciğer kanseri, hepatik ensefalopati gibi daha ciddi karaciğer hastalıklarının erken teşhisinin kolaylaşması, karaciğerin geri dönüşümsüz hasara uğramasından önce koruyucu önlem alınmasına izin vererek hayat kurtarabilir. Karaciğer hastalıklarının teşhisinde genellikle biyopsi gibi yöntemler kullanılır. Biyopsiler invaziv ve

pahalıdır, hastalar üzerinde stres oluşturur ve belirli riskler teşkil eder. Bu nedenle, erken evre hastalık tespitinde kullanım için pratik ve uygulanabilir olmadığı anlamına gelir. Karaciğer vücuttaki birçok önemli metabolik olayın merkezinde yer alır ve bu nedenle metabolik atıklar, karaciğerin durumunu takip etmek ve hastalığın erken belirtilerini tespit etmek için ideal bir hedef sunar. Çeşitli çalışmalarda karaciğerin sağlığının tespiti için kan veya idrar örnekleri de kullanmıştır, ancak nefesdeki uçucu gazların tespiti, gerekli adımların atılabilmesi için daha hızlı ve uygun bir çalışma mekanizması sunar.

Van den Velde, S. ve arkadaşları tarafından yapılan bir çalışmada, karaciğer sirozunun nefesteki spesifik koku bileşiklerini incelemek için gaz kromatografisi-kütle spektrometrisi kullanılmıştır. Çeşitli derecelerde karaciğer sirozu olan 50 hasta (19 kadın) ve yaşları uyumlu 50 sağlıklı gönüllü (29 kadın) çalışmaya dahil edilmiş ve sonuç olarak Çizelge 1.2’de gösterilen gazların sağlıklı ve hasta bireyler arasında değişiklik gösterdiği gözlenmiştir [39].

Çizelge 1.2. Tespit edilen bileşiklerin, ppbv cinsinden tanımlayıcı istatistikleri [39]

	Hastalıklı Ortanca	Alt Çeyrek	Üst Çeyrek	Sağlıklı Ortanca	Alt Çeyrek	Üst Çeyrek	Belirleyici İşaret
Aseton	765.13	465.81	1739.16	212.25	144.56	325.92	+
2-pentanon	1.56	0.98	2.69	0.38	0.25	0.61	+
2-bütanon	3.11	1.61	6.86	0.38	0.06	1.42	+
Dimetil sülfid	29.02	10.44	61.52	13.79	7.51	23.24	+
Dimetil disülfid	0.00	0.00	0.03	0.00	0.00	0.06	-
Dimetil trisülfid	0.00	0.00	0.00	0.00	0.00	0.00	-
Karbon disülfid	0.01	-0.60	0.13	-0.03	-0.04	0.03	-
Alil metil sülfid	0.09	0.00	0.26	0.08	0.04	0.19	-
İndol	0.00	0.00	0.05	0.18	0.06	0.32	+
Skatol	0.00	0.00	0.00	0.00	0.00	0.00	-
Dimetil selenit	0.08	0.00	0.15	0.56	0.38	0.70	+
1-propanol	7.80	1.90	9.50	8.10	2.65	10.76	-

Bu tez çalışmasının amacı, karaciğer sirozu ve alkolsüz yağlı karaciğer hastalığı ile doğrudan ilişkili olan dimetil sülfid, 2-bütanon ve aseton gazlarını saptama, sınıflandırma ve bu gazların konsantrasyonlarını belirlemektir. Ayrıca klinik düzeyde kolaylıkla kullanılabilecek bir cihazın geliştirilmesi için öncü bir çalışma olması plânlanmış, örüntü tanıma algoritması olarak kullanılacak çeşitli makine öğrenmesi algoritmalarının performansları test edilmiş ve sensör tepkileri veri bilimi esasları altında işlenmiştir.

Bu tez kapsamında Bölüm 1 ile genel bilgiler, literatür özetleri ve tezin amacı ile ilgili bilgiler sunuldu. Bölüm 2’de, gaz sensör çeşitleri ve çalışma prensipleri anlatıldı. Bölüm 3’de ise veri bilimi konsepti ve makine öğrenmesi algoritmaları tanıtıldı. Bölüm 4’de, gaz sensörlerinin nasıl kullanıldığı ve verilerin nasıl toplandığı ifade edildi. Bölüm 5 ise sensör tepkilerini, örüntü tanıma algoritması olarak kullanılan makine öğrenmesi algoritmalarının sonuçlarını göstermektedir.

2. SENSÖRLER

Son yıllarda, gaz halindeki bileşenleri algılamak için sensörler, çeşitli ilke ve malzemelere dayalı olarak üretilmiştir. Gaz sensörlerinin, uzay araştırmaları, tıp, endüstriyel çalışmalar, gıda sektörü başta olmak üzere bir çok alanda kullanımı artmaktadır [40]. Örneğin, CH₄, LPG ve H₂ gibi havadaki yanıcı gazları algılayan metal oksit yarı iletken gaz sensörleri, şu anda evlerde gaz kaçağı alarmları için büyük ölçekte kullanılmaktadır [41].

Gaz sensörleri, büyük önem taşıyan kimyasal sensörlerdir. Bir kimyasal sensör, yardımcı elektronik bileşenler ile, kimyasal bilgiyi, frekans, akım veya voltaj değişimi olarak elektronik sinyal biçimine dönüştüren yapıdır [42]. Yarı iletken gaz sensörleri, elektrokimyasal gaz sensörleri, optik gaz sensörleri ve akustik gaz sensörleri olmak üzere çeşitli gazların tespiti için farklı gaz sensörleri kullanılmıştır. Gaz sensörleri uçucu gazların türü ve konsantrasyonuna bağlı olarak geliştirilmiştir. Sensörlerin kullanım farklılıkları, performans, hassasiyet, seçicilik, algılama limiti, tepki süresi gibi bazı özelliklere dayanmaktadır. Bu sensörlerden tasarım geometrisinin basit olması imalat kolaylığı sağlaması, farklı algılama malzemeleriyle uyumlu olması, yüksek tersinirlik ve tekrarlanabilirliğe sahip olması, düşük gaz konsantrasyonlarında dahi yüksek hassasiyete sahip olması ve düşük maliyete sahip olması beklenir [40]. Bu bölümde, gaz sensörleri çalışma mekanizmalarına göre gruplandırılarak açıklanmıştır.

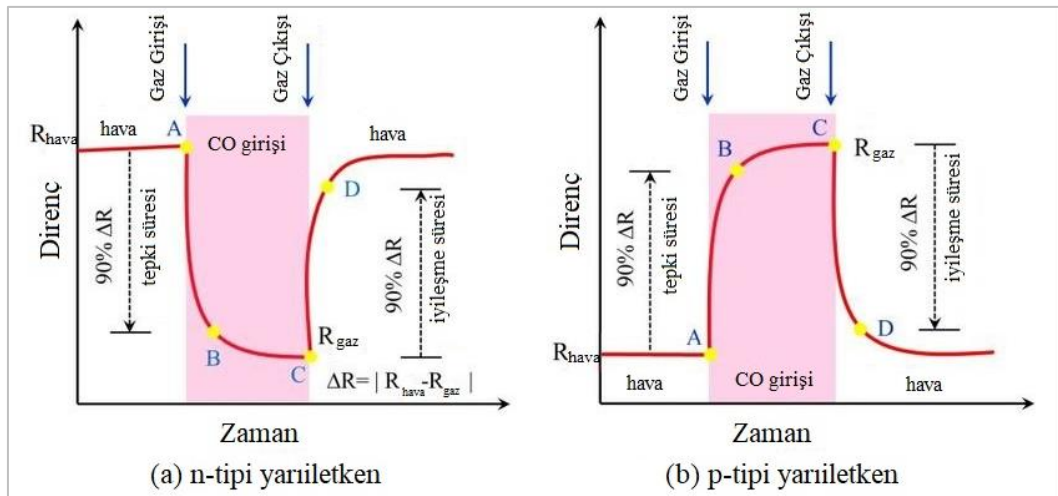
2.1. Elektriksel Değişim Tabanlı Gaz Sensörler (Kimyasal Direnç Sensörleri)

Kimyasal direnç sensörünün çalışma prensibi, hedef gazın, sensör yüzeyindeki numune ile etkileşmesi sonucu numunenin direncini değiştirmesi ilkesine dayanır. Hedef gaz, algılama numunesiyle temas ettiğinde, elektrotları birbirine bağlayan algılama malzemesinin fiziksel ve kimyasal özellikleri değişir ve bunun sonucunda algılama malzemesinin direnci artar veya azalır [43]. Dirençteki bu değişim saptanması için sensör elektrotları, harici bir veri işleme birimine bağlanır. Böylelikle ortamdaki hedef gazın konsantrasyonu ölçülmüş olur ya da varlığı tespit edilir.

Kimyasal direnç sensörleri, silikon substrat üzerine monte edilen iki elektrot arasındaki boşluğu dolduran hassas malzeme ile üretilir. Elektrotlar, yüksek iletkenlik ve eylemsizlik özellikleri nedeniyle genellikle altın veya kromdan yapılır. Kimyasal direnç sensörü, yüksek hassasiyet, üretim kolaylığı, geniş algılama malzemeleri yelpazesi gibi çok çeşitli avantajlara sahip olmasına rağmen seçicilikten yoksundur ve ortamdaki nem ve sıcaklığa karşı oldukça hassastır [44].

2.1.1. Metal oksit yarı iletken sensörü

Metal oksit yarı iletken (MOS) sensörleri, kimyasal direnç sensörlerinin bir versiyonudur. Doğruluk oranları, uzun ömürleri, kısa tepki süreleri ve elektronik devrelere kolay entegrasyonları nedeniyle tercih edilirler. Sensörün çalışma mekanizması, indirgeyici veya yükseltgeyici (oksitleyici) gaz algılama malzemesi ile temas ettiğinde iletkenliğin değişmesine bağlıdır. Bu etkileşim, sensörde bulunan algılama malzemesinin fiziksel ve kimyasal özelliklerini değiştirir [44]. Her metal oksit algılama malzemesi, uygun bir sıcaklıkta belirli bir gaza duyarlıdır. MOS algılama malzemesi n-tipi ve p-tipi malzemeler olarak sınıflandırılır [45]. n-tipi MOS'taki yük taşıyıcıları elektronlardır. Şekil 2.1'de gösterildiği gibi n-tipi MOS'un yükseltgeyici gazla reaktivitesi iletkenliği azaltır, indirgeyici gazla reaktivitesi ise iletkenliği artırır. p-tipi MOS'taki yük taşıyıcıları deliklerdir. p-tipi MOS'un yükseltgeyici gazla reaktivitesi iletkenliği artırır, indirgeyici gazla reaktivitesi ise iletkenliği azaltır. İletkenlikteki bu değişiklikler, sensörün temel çalışma prensibini oluşturan harici bir devre veya sinyal işleme ünitesi ile ölçülür [46].



Şekil 2.1. a) ntipi MOS ve b) p-tipi MOS'un gaz reaktivitesi sonucu oluşan direnç değişimini gösteren şema [47]

2.1.2. İletken polimer yapılı sensörler

İletken polimer sensörlerinde, metal oksit yarı iletken gaz sensörleri gibi direnç değişimi esas alınmıştır. Polipirol (PPy), polianilin (Pani), politiyofen (PTh) ve bunların türevleri gibi iletken polimerler, gaz sensörlerinin aktif katmanları olarak kullanılmaktadır [48]. Polimer sensörü, polimer türevleriyle bağlantılı olan yüksek iletkenliğe sahip olması nedeniyle genellikle altından yapılan iki elektrottan oluşur. Polimer malzeme, aktif algılama katmanını oluşturur. Bu iletken polimerler, ilgili polimerin oksitlenmesinin kimyasal veya elektrokimyasal prosesi yoluyla sentezlenebilir. Bu elektrotlar ve iletken polimerler, silikon substratta üretilir. İletken polimer sensörler, yüksek hassasiyet, basit tasarım, kısa tepki süresi ve oda sıcaklığında stabilite gibi birçok avantaja sahiptir. Ancak, sıcaklık değişimlerine ve neme karşı hassastırlar.

2.1.3. Nano-malzeme yapılı sensörler

Gaz algılama için kullanılan metal oksit yarı iletken gaz sensörleri ve iletken polimer sensörlerinden daha duyarlı sensörlere ihtiyaç duyulması alternatif materyallerin kullanımının önünü açmıştır [49]. Bu alternatiflerden biri nanomalzemelerin kullanımıdır. Yüksek yüzey alanı/hacim oranları, uygulamaya göre işlevselleştirilebilme olasılığı, önemli ölçüde artırılmış tepki hızı ve nano ölçekli özellikler nedeniyle, nanomalzeme yapılı sensörler büyük bir potansiyele sahiptirler[50]. Gaz algılama teknolojisi için en yaygın karbon nanomalzemelerden ikisi karbon nanotüpler (CNT'ler) ve indirgenmiş grafen oksitlerdir (rGO).

Karbon nanotüpler keşfedildiğinde olağanüstü elektriksel, mekanik, optik, termal ve kimyasal özelliklere sahip oldukları görülmüştür. Karbon nanotüpler, tek duvarlı ve çok duvarlı olarak bilinen bir veya daha fazla grafen katmanından oluşan içi boş silindirlerdir [51]. Tek bir grafen tabakasıyla yapısal benzerliğine rağmen, tek duvarlı CNT'ler çaplarına ve kiralitelerine (grafen tabakalarının CNT'leri oluşturmak için nasıl yuvarlandıklarına) bağlı olarak metalik veya yarı iletken olabilir. Bu özellikler, onların p-n yapıları gibi nano ölçekli elektronik cihazlar olarak kullanılmalarına izin verir [52].

CNT ile hedef gaz arasındaki etkileşim, kütle artışı gibi CNT'nin fiziksel özelliklerinde değişikliklere neden olur, böylelikle direnç artar ve iletkenlik azalır. CNT sensörü,

ölçülebilir değişken olarak iletkenliği kullanır. Elektrik akımındaki bu değişiklik, CNT sensörlerinin temel çalışma prensibini oluşturan özelliklerdir.

Tasarım olarak CNT sensörü, MOS gaz sensörüne daha çok benzer. Bu sensörler, aktif algılama katmanı olarak işlev gören karbon nanotüpler tarafından köprülenen elektrotlardan oluşur. Kaynak ve drenaj elektrotları bu CNT ile bağlanırken, silikon substrat üzerine inşa edilmiş silikon dioksit tabakası gibi dielektrik bariyer tabakası tarafından yalıtılır [53].

Gaz algılama uygulamaları için bir başka ilginç malzeme, iki boyutlu (2D) bir karbon olan grafen oksittir (GO) [54]. Ağır oksitlenmiş GO elektriksel olarak iletken değildir. Ancak, GO, çoğu zaman indirgeme yoluyla iletken bir malzemeye dönüştürülebilir [55]. Gaz sensörlerinin üretiminde kullanılan bu malzeme, indirgenmiş grafen oksit (rGO) olarak isimlendirilir [56]. rGO'nun yüksek düzeyde iletken olması, düşük konsantrasyonlardaki hedef gazları tespit etmek için yeterli sinyal üretimini sağlar.

Saf grafen ile karşılaştırıldığında, rGO'nun gaz algılama için bir malzeme olarak birçok avantajı vardır. rGO, basit kimyasal yaklaşımlar kullanılarak pahalı olmayan reaktiflerden büyük miktarlarda sentezlenebilir [57]. İletkenlik, gerilme mukavemeti vb. gibi rGO'nun özellikleri, büyük ölçüde GO indirgeme derecesine bağlıdır, bu da onların indirgeme koşulları tarafından ince bir şekilde ayarlanabilecekleri anlamına gelir [58]. rGO'nun zengin kimyası, potansiyel olarak rGO'yu gaz algılama için çok yönlü bir platform yapan farklı fonksiyonel gruplar ile daha fazla kimyasal modifikasyona izin verir.

Algılama uygulamaları için rGO'nun tüm avantajlarına rağmen, rGO tabanlı gaz sensörleri için birçok iyileştirme yapılması gerekmektedir. Bir rGO pulunun yüzeyinde farklı gaz molekülleri adsorbe edilebilir ve iletkenliğini değiştirebilir, bu da rGO tabanlı sensörlerin zayıf seçiciliğine neden olur [54].

2.2. Optik ve Kalorimetrik Değişimlere Dayalı Sensörler

2.2.1. Optik sensör

Optik sensörler, fotonların tespitinin doğrudan elektronik sinyallerle sonuçlandığı aktif bir alanda, ışığın yayılmasının kontrolü, manipülasyonu veya tespitine dayalı olarak çalışır.

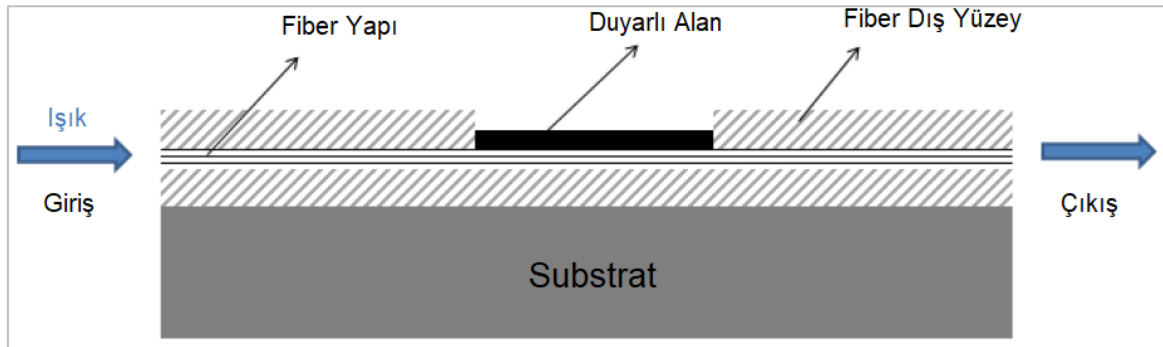
Yaygın olarak kullanılan iki optik sensör, fiber optik gaz sensörleri ve fotonik kristal gaz sensörleridir.

Fiber optik gaz sensörleri

Fiber optik sensörler, Şekil 2.2'de gösterildiği gibi algılama katmanı, optik fiber ve alt katmandan oluşur. Optik fiber, algılama katmanını kısmen ışığa maruz bırakmak için alt katman üzerinde tutulur. Algılama membranı, analit ile algılama katmanı arasındaki etkileşimin meydana geldiği, kırılma indisi gibi fiziksel ve kimyasal değişiklikler üreten fiberin üzerine yerleştirilir. Bu sayede, fiber optik sensörler, algılama katmanında ölçülebilir bir optik veya optoelektrik değişiklikler yaratan analitleri algılayabilir [59].

Fiber optik sensörlerin çalışma mekanizması, PWM sinyalleri ve analitler-algılayıcı katman etkileşimi ile açıklanabilir. PWM tabanlı fiber optik gaz sensörlerinde, istenen analitlere maruz kalan algılama katmanının özelliklerinde oluşan değişiklikler sinyalin darbe genişliğini değiştirir [59].

Fiber optik sensörler için algılama katmanları, polivinilpirolidon (PVP) veya polivinil klorür (PVC) içinde solvatokromik gibi boyalar kullanılarak hazırlanır. Boyanın yük transfer karakteri nedeniyle, ışık analitlere maruz kaldığında, algılama tabakası kırılma indisini değiştirir [60].



Şekil 2.2. Fiber optik sensörlerin şematik görünümü [40]

Fiber optik sensörler, bir optik fiberin alt tabakaya eklenmesi ve ardından bir algılama malzemesinin yerleştirilmesiyle yapılır. Optik fiberi barındırmak için kuvars substrat üzerinde istenen eğrilik yarıçapına sahip kavisli V şeklinde bir oluk oluşturulur [61]. Algılama membranı, belirli bir uygulama için kullanılan algılama malzemelerine bağlı

olarak farklı yöntemler kullanılarak yapının üst kısmında biriktirilir. Polimerler, fiber optik sensörlerde algılama malzemeleri olarak yaygın olarak kullanılmaktadır. Polimerik algılayıcı tabaka durumunda, polimer film bir çözücü içinde çözülür ve optik fiber tutulan substrat üzerinde spin kaplama, daldırma kaplama veya sprey kaplama [59] ile biriktirilir.

Fotonik kristal gaz sensörleri

Fotonik kristal (PhC), algılama için farklı kırılma indisine sahip dielektrik malzemelerin periyodik düzenlemelerini yaygın olarak kullanan kırılma indisine dayalı sensörlerdir [62]. Bu sensörler, nanometre boyutundaki kimyasal bileşikler, sıcaklık, basınç ve nem gibi çevresel parametreleri algılama, sensör tasarım esnekliği kazanma, elektrik sinyallerinden elektromanyetik paraziti önleyerek daha fazla güvenlik sağlama yeteneği ile yüksek hassasiyet elde etmek için potansiyel adaylar olarak önerilmiştir.

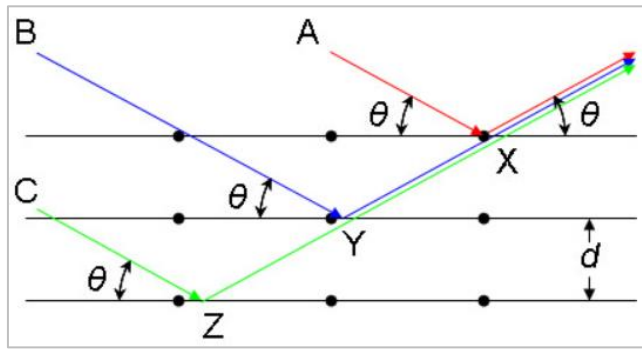
PhC gaz sensörleri, yüksek oranda periyodik mikro veya nano yapılara veya desenlere sahip dielektrik malzemeleri içerir. Kullanılan dielektrik malzemelerin periyodik olarak düzenlenmesi, belirli dalga boylarındaki ışığın PhC'den geçmesine izin veren bir fotonik bant aralığı oluşturur. Bu fotonik özellik, farklı PhC malzemeleri ve desenleme kullanılarak hedef ışığın özelliklerine dayalı olarak tasarlanır [63]. Orta kızılötesi bölgede soğurma çizgileri sergiledikleri için CO₂, CH₄ veya CO gazlarını algılamak için kullanılan bu cihazların yaygın örneği orta kızılötesi PhC'lerdir. Gaz algılama uygulamasında, iki faktörün neden olduğu kırınım dalga boyu değişimini ölçerek analitleri tespit etmek için PhC sensörleri kullanılır. PhC'lerde ışığın kırınımı ve algılama mekanizması Bragg yasası (Şekil 2.3) ile açıklanabilir:

$$n\lambda = 2d \times \sin\theta \quad (2.1)$$

λ , x-ışınının dalga boyu, d , kristal katmanların aralığı (yol farkı), θ gelen açı (gelen ışın ile saçılma düzlemi arasındaki açı) ve n bir tamsayıdır. Gelen ışığın ve PhC tabakasının konumunun sabit olduğunu varsayarsak, efektif kırılma indisi, ışığın izin verilen dalga boyunu değiştiren tek değişken olur.

Analitlerin kırılma indisi, PhC malzemelerinkinden farklıdır. Kırılma indisindeki bu fark, algılama için gerekli bir etkidir. PhC algılama katmanı, ışık kaynağı (kızılötesinden görünür ışığa kadar değişen) ve dedektör arasına yerleştirilir [65]. Analitler periyodik yapı

üzerinde yoğunlaşır veya kalıplardan geçer. Her bir kalıbın boyutu ve mesafesi genellikle mikro metre mertebesinde, ve PhC katman hacminin %30-40'ını kapsar. Ek olarak, desenler istenen kırınım dalga boyuna bağlı olarak çeşitli şekillerde optimize edilebilir ve PhC gaz sensöründeki algılama tabakasının boyutu yaklaşık 1 cm genişliğe kadar düşürülebilir [66]. PhC gaz sensörleri, algılama odasını PhC'lerle değiştirerek geleneksel optik gaz sensörlerinin özelliklerini daha da geliştirmek için potansiyel adaylardır. Geleneksel cihazlar, ~10–50 cm boyutlarında ışık yolunu korumak için aynalar kullanır, bu nedenle minyatürleştirilmeleri zordur. PhC'ler bu optik algılama odasını daha küçük ve daha basit bir konfigürasyonla değiştirebilir. Bununla birlikte, PhC tabanlı gaz algılama cihazlarının minyatürleştirilmesi, harici ışık kaynağının ve foto dedektörün boyutu nedeniyle hala sınırlamalara sahiptir.



Şekil 2.3. Bragg yasasının şematik gösterimi [64]

2.2.2. Kolorimetrik sensör

Kolorimetrik gaz sensörleri, yalnızca sınırlı miktarda enerjinin bulunduğu ve düşük maliyetli sensörlerin gerekli olduğu uygulamalarda kullanılır [67]. Bu tür uygulamalarda kolorimetrik sensörler CO_2 , C_2H_4 , NO_2 gibi gazları güvenilir bir şekilde algılayabilir. Düşük maliyetli sensörler, sensör ağlarına entegre edilebilir ve diğer sensör prensipleriyle birleştirilebilir [68]. Kolorimetrik sensör, hedef gaz sensörün algılama katmanı ile temas ettiğinde renk değiştiren implante edilmiş birkaç kimyasal-soğurucu boyaya sahiptir. Kimyasal-soğurucu boya ile gaz arasındaki bu etkileşim kimyasal işleminden geçer ve boyanın rengi değişir [69]. Renkteki bu değişiklik ayrıca uygun bir görüntü işleme sistemi ile hesaplanır ve hedef gazın varlığı tespit edilir. Bu renk değişimi gazın konsantrasyonu ile ilişkilidir ve diğer gazlarla etkileşimi yoktur. Reaksiyon oda sıcaklığında gerçekleşir, bu nedenle ısıtma gücünün gerekli olmadığı anlamına gelir. Güç tüketimi birkaç miliwatt civarındadır. Kolorimetrik sensörler,

uçucu madde tespiti için en düşük maliyetli gaz sensörüdür. Ancak sensör yanıtının tersinmezliği, uzun yanıt süresi, düşük doğruluk, yüksek görüntü işleme gücü gibi dezavantajları vardır [70].



Şekil 2.4. Bir kalorimetrik sensörün hedef gaz ile teması sonrası oluşan renk değişimi [68]

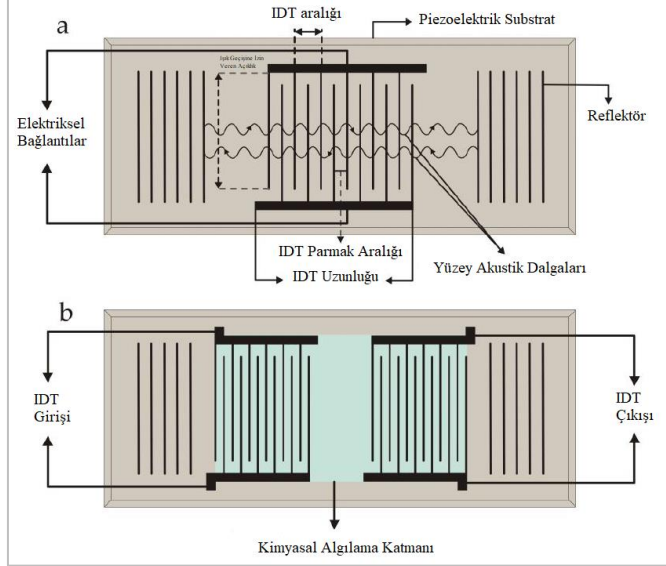
2.3. Akustik Varyasyonlara Dayalı Sensörler

2.3.1. Yüzey akustik dalga sensörü

Piezoelektrik yüzey akustik dalga cihazları ilk olarak 1965 yılında White ve Voltmer tarafından geliştirilmiştir [71]. En temel formlarında, bu cihazlar, Şekil 2.5.b'de gösterildiği gibi, düzlemsel bir piezoelektrik alt tabaka üzerinde üretilmiş ve bir boşlukla ayrılmış iki interdigital transdüserden (IDT) oluşur. Bir interdigital dönüştürücüye uygulanan bir RF voltajı, piezoelektrik etki nedeniyle akustik dalgaları başlatan alternatif bir elektrik alanı ile sonuçlanır. Kimyasal algılama bölgesininin hedef gaz ile teması doğal frekansta değişime neden olur. Alıcı dönüştürücü, akustik sinyali tekrar elektriksel forma dönüştürür. Böylelikle hedef gazın konsantrasyonunun hesaplanabilmesi için gerekli bilgiler elde edilir. Frekansta oluşan değişiklikler empedans analizörü ile doğru bir şekilde ölçülebilir veya kablosuz olarak alınıp işlenebilir. SAW gaz sensöründe meydana gelen aşamalar kısaca şu şekildedir: Akustik dalgalar üretmek için piezoelektrik substratın birbirine bağlı dönüştürücü (IDT) ile uyarılması, hedef gazlarla etkileşimleri sonucu akustik dalgaların modülasyona uğraması ve birbirine bağlı dönüştürücü ile akustik dalgadaki değişimin tespiti.

Yüzey akustik dalga sensörü, yüksek hassasiyete ve kısa tepki süresine sahip olması, kablosuz çalışabilmesi gibi avantajlara sahiptir. Avantajlarına rağmen, üretim sürecinin

karmaşık olması, gaz tespiti için bir gerekli olan bir dizi konfigürasyonda çalıştırılmasının zor olması gibi dezavantajları da bulunmaktadır.



Şekil 2.5. a) Tek portlu SAW rezonatörünün; b) bir kimyasal tanıma katmanına sahip iki bağlantı noktalı bir SAW rezonatörünün tasarımı [72]

2.3.2. Kuvars kristal mikrobalsans sensörü

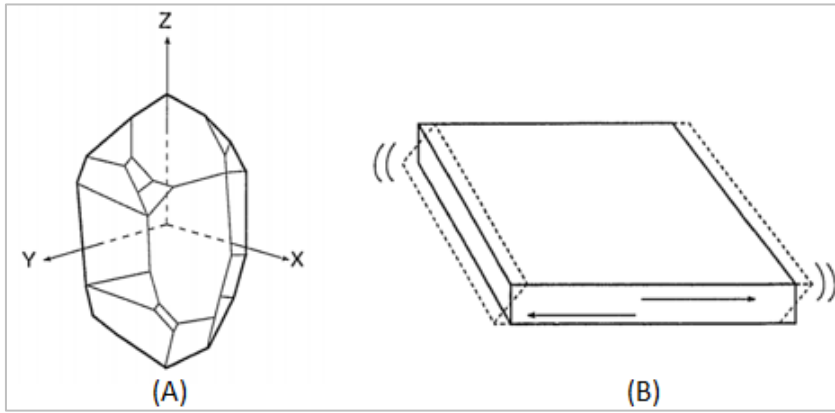
Geçtiğimiz yıllarda, frekans analizi için birincil ölçüm tekniği olarak Kuvarz Kristal Mikrobalsans (QCM) sensörlerini kullanan çok sayıda yayın ortaya çıkmıştır [5, 9, 16-18, 25, 73-77]. Moleküler düzeneklerin, yumuşak katmanların, atomların ve moleküllerin adsorpsiyonunun analizi, hücre büyümesi ve bakteri etkileşimleri çalışmaları, malzemelerin karakterizasyonu ve özellikle biyofizik, biyokimya ve farmakoloji gibi çalışmalarda bu sensörler özellikle tercih edilmiştir [78-82].

QCM sensörlerinin çalışma mekanizması

QCM sensörlerinin çalışma mekanizması piezoelektrik etkiye dayanmaktadır. Piezoelektrik etki, ilk olarak 1880'de Pierre ve Jacques Curie kardeşler tarafından keşfedilmiştir [83]. Benzersiz bir özellik olarak, bu etki, fiziksel stres altında elektrik alan üretilmesi ya da elektrik alan altında mekanik bir deformasyon (büzülme veya genişleme) oluşmasıdır [84]. Piezoelektrik etki, kristaller, belirli seramikler, ve çeşitli proteinler gibi bazı katı maddelerde saptanmıştır [85,86]. Piezoelektrik oluşumunu açıklamak için kristalin her bir molekülüne bakılması gerekir. Bir katı, elektriksel olarak nötr olabilsede, her molekülün dipol olarak da

bilinen bir polarizasyonu vardır. Bu, molekülü oluşturan atomların ve her molekülün bir ucunun daha negatif, diğer ucunun ise pozitif yüklü olduğu moleküllerin pozisyon-alma şeklinin bir sonucudur. Bir kuvvetin veya elektrik alanının uygulanması sonucu yüklü atomlar yer değiştirir ve net bir dipol oluşturur.

Piezoelektrik kuvars kristal rezonatörleri, tek bir doğal veya sentetik kuvars kristalinden hassas şekilde kesilmiş levhalardır. Kusursuz doğal formunda bir kuvars kristali Şekil 2.6.a'da gösterilmektedir.

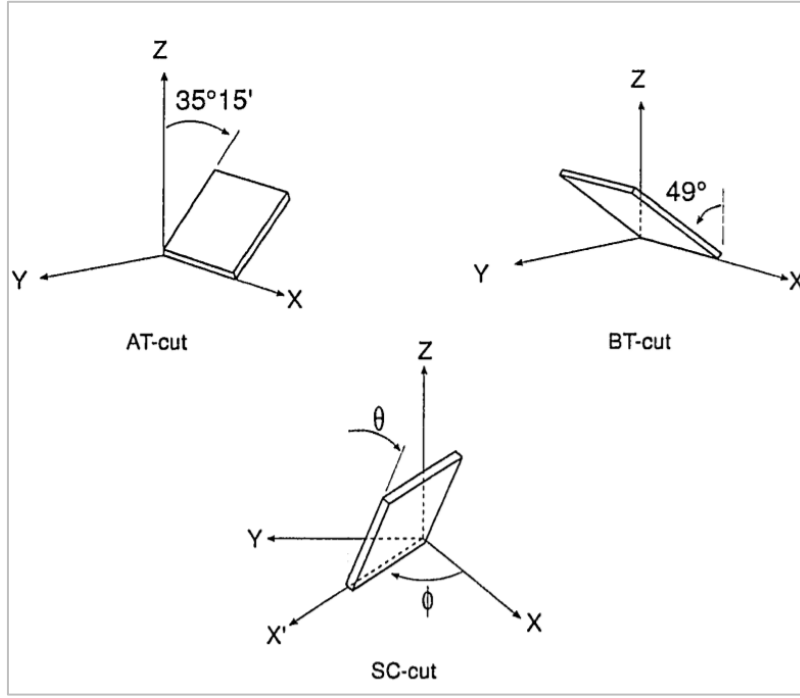


Şekil 2.6. a) Doğal formunda bir kuvarz kristali, b) Kalınlık shear modu [87]

Piezoelektrik kuvars kristalleri birçok yönde rezonans modu sergiler. Üç farklı türde titreşim sergileyen bir dikdörtgen katı yapı örnek olarak gösterilebilir. Rezonanslar bu örnekte uzunlamasına, yanıl ve burulma olarak gerçekleşir. Genel olarak, kristallerin tek bir titreşim modunda olması arzu edilir (Şekil 2.6.b). İstenmeyen modların bastırılması, belirli bir kristalografik oryantasyonda bir kuvars kristal levhanın kesilmesiyle elde edilebilir. Rezonans modunu etkileyebilecek diğer faktörler, kristal üzerindeki elektrotların konfigürasyonunu, destekleyici yapıyı ve osilatör devresini içerir [25]. Yüksek frekanslı kalınlık kesme (shear) modu, bir kuvars kristal rezonatör için kütle değişimine olan yüksek duyarlılığı nedeniyle oldukça tercih edilen bir titreşim modudur. Bu modda rezonatörün plaka hareketleri esasen tek boyutludur [25].

Bir kuvars kristal rezonatör plakasının saf kalınlık kesme modunda salınım yapması için plakaların kristal eksenlerine göre belirli bir oryantasyona, yani döndürülmüş Y-kesim (Y-cut) ailesine göre kesilmesi gerekir. Şekil 2.7'de gösterilen AT- ve BT-kesim kristal plakaları, tek başına döndürülmüş Y-kesimlerini temsil etmektedir. AT- ve BT- kesim

kuvars kristali rezonans frekansları, genellikle oda sıcaklığına yakın sıcaklıklarda, sıcaklık değişimlerine karşı duyarsız, kütle değişimlerine karşı da oldukça duyarlıdır [25]. Bu nedenle QCM sensör çalışmalarında tercih edilmektedirler. Standart AT-kesim dönüş açısını değiştirerek çalışma sıcaklığı aralığını değiştirmek de ayrıca mümkündür [88].



Şekil 2.7. Kristal plaka kesimleri [87]

1959'da Sauerbrey [89] ilk olarak bir kalınlık kesme modu rezonatörünün rezonans frekansındaki kaymanın kristalin yüzeyinde biriken kütle ile orantılı olduğunu göstermiştir. Kalınlık kesme modu rezonatörü, kristalin, bir elektrik osilatör devresinde frekans belirleyici eleman olarak çalıştığı, genellikle kuvars kristal mikrobals olarak adlandırılan bir gravimetrik kütle sensörü olarak yapılandırılır. QCM, Şekil 2.8'de gösterildiği gibi her iki tarafta elektrotlarla kaplanmış AT veya BT kesimli kuvars kristalinden oluşan ince bir diskten oluşur. QCM sensörü elektrotlarına AC gerilim uygulanması ile oluşan rezonans salınımları, kalınlık shear modunda Şekil 2.9'da gösterildiği gibi farklılık göstermektedir. Bu durumda;

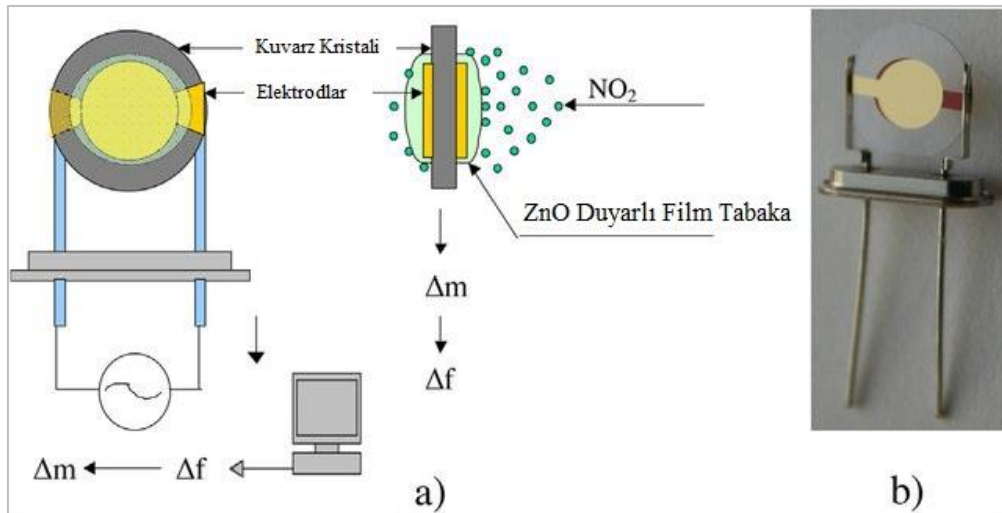
$$h_q = N\left(\frac{\lambda}{2}\right) \quad (2.2)$$

h_q , kristalin kalınlığı, λ , akustik dalga boyu ve N , rezonans modunu temsil eden rezonatör harmonik sayısıdır.

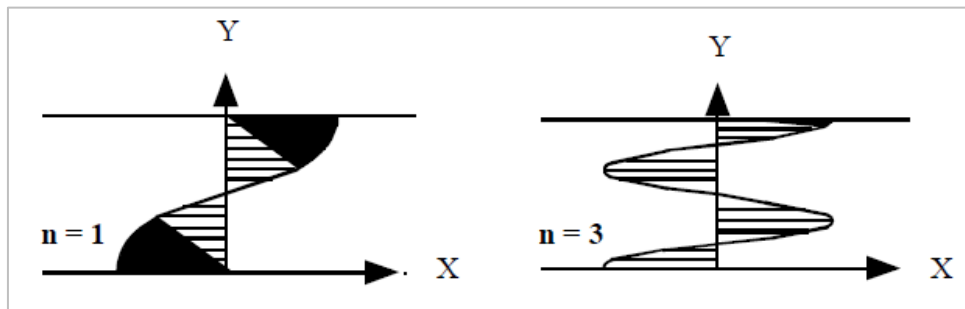
$$v_q = \sqrt{\left(\frac{\mu_q}{\rho_q}\right)} \quad (2.3)$$

v_q , kesme dalgası hızı, ρ_q , kuvarz yoğunluğu (2.65 g/cm^3) ve μ_q , kesme sertliğidir ($2.95 \times 10^{11} \text{ dyne/cm}^2$). Buradan f_o , temel rezonans frekansı;

$$f_o = N \left(\frac{v_q}{2h_q} \right) \quad (2.4)$$



Şekil 2.8. a) Moleküllerin kristal tarafından tutulmasıyla oluşan kütle değişimi bilgisinin frekans olarak hesaplanmasının temsili gösterimi ve b) altın elektrodlu bir kuvarz kristali [90]



Şekil 2.9. a) Osilasyonun temel ve b) üçüncü modunun temsili gösterimi [91]

Ticari olarak temin edilebilen QCM rezonans frekansı aralıkları 5-20 MHz'dir. Temel olarak, QCM'nin frekans aralığı, kesme kalınlığına veya QCM kafesinin kalınlığına göre değişir. Başka bir deyişle, kesme kalınlığı azaldıkça QCM'nin rezonans frekansı artacaktır; aksine kalınlık arttıkça frekans azalacaktır [88]. Uygulamada, rezonans frekansı yalnızca QCM'nin kesme kalınlığına veya rezonans moduna değil, aynı zamanda QCM yüzeyinde biriken kütleye bağlı olarak değişir. Rezonans frekansını değiştiren harici kütle, QCM üzerine kaplanmış seçici ince filmler veya filme emilen veya film üzerine adsorbe edilen gaz molekülleridir. Harici kütlenin frekansı değiştirmesinin nedeni kalınlıktaki artıştır. QCM'ye eklenen kütle ile rezonans frekansının değişimi arasındaki matematiksel ilişki, Sauerbrey tarafından ince, katı ve tek biçimli filmlerin durumuna dayalı olarak türetilmiş ve 1959'da yayınlanmıştır [89].

$$\frac{\Delta f}{f_0} = - \frac{\Delta h_q}{h_q} = - \frac{\Delta \rho_q h_q}{\rho_q h_q} \quad (2.5)$$

Denklem daha sonra aşağıdaki şekilde geliştirilmiştir;

$$\Delta f = -2f_0^2 \frac{\Delta m}{A\sqrt{\rho\mu}} \quad (2.6)$$

Δf (Hz) rezonans frekansındaki değişiklik ve f_0 (MHz) kristalin temel rezonans frekansı ise, Δm (g) QCM yüzeyine kaplanan ve tutunan malzemenin kütlesidir. A (cm^2) iki altın elektrot tarafından tanımlanan piezoelektrik olarak aktif alanıdır.

QCM sensörlerinin elektriksel eşdeğer devresi

QCM sensörlerinin elektriksel özelliklerini, rezonansa yakın bir frekans aralığında Butterworth - van Dyke eşdeğer devresi ile tanımlamak mümkündür. Statik kol olarak da adlandırılan elektriksel kol yalnızca elektrotlar arasında kalan kuvarsın kapasitansını ve harici bağlantı kapasitanslarını temsil eder. Hareketli kol olarak adlandırılan akustik kol ise kuvars üzerindeki akustik rezonansları temsil eder.

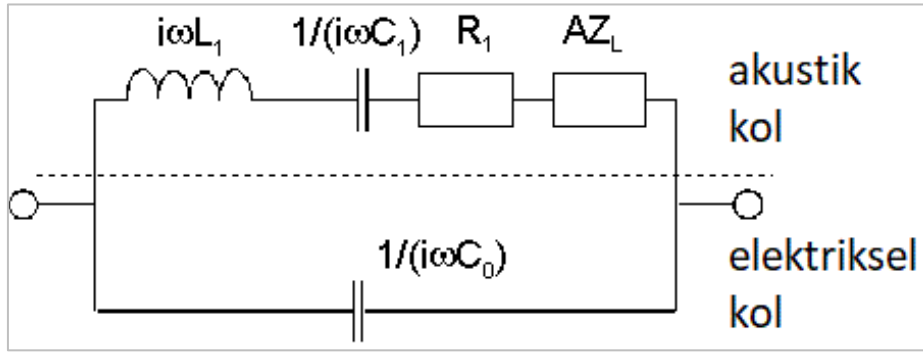
$$C_0 = \frac{\varepsilon_{22}A}{h_q} \quad (2.7)$$

$$C_1 = \frac{8K^2C_0}{(N\pi)^2} \quad (2.8)$$

$$L_1 = \frac{1}{\omega_s^2 C_1} \quad (2.9)$$

$$R_1 = \frac{\eta_q}{\mu_q C_1} \quad (2.10)$$

Burada N , tonlama sayısıdır; A ve h_q , kristal plakanın alanı ve kalınlığıdır; $\omega_s = 2\pi f_s$ ise f_s pertürbe edilmemiş kristalin seri rezonans frekansıdır; K^2 , ε_{22} , μ_q ve η_q sırasıyla kuvars elektromekanik bağlantı katsayısının karesi, dielektrik geçirgenliği, kesme sertliği ve efektif viskozitedir.



Şekil 2.10. Butterworth - van Dyke (BvD) eşdeğer devresi [92]

Yük empedansı Z_L , kristal yüzeyindeki stres, σ ve hız, $\dot{u}_x$ oranı olarak verilir. Z_L kompleks bir sayı olarak ifade edilir: $\text{Re}(Z_L)$ gerçek kısmı, hız ile fazdaki gerilim bileşenini ifade eder ve arayüzeydeki enerji kaybını temsil eder; $\text{Im}(Z_L)$ sanal kısmı, hız ile faz dışı gerilim bileşenini ifade eder ve arayüzeyde depolanan enerjiyi temsil eder. Pertürbe edilmemiş kristalin rezonans frekansına kıyasla, frekans kaymasının küçük olduğudurumlarda, frekans kayması ile yük empedansı arasında orantılı bir ilişki vardır:

$$\frac{\Delta f}{f_0} = \frac{-\text{Im}(Z_L)}{\pi Z_q} \quad (2.11)$$

$$\frac{\Delta \Gamma}{f_0} = \frac{\text{Re}(Z_L)}{\pi Z_q} \quad (2.12)$$

Burada Δf ve $\Delta \Gamma$, sırasıyla rezonans frekansının kayması ve kristalin yarı bant genişliğidir; $Z_q = (\rho_q \mu_q)^{1/2}$ akustik empedansı belirtir, burada ρ_q ve μ_q , AT-cut kuvarsın yoğunluğu ve kesme sertliğidir.

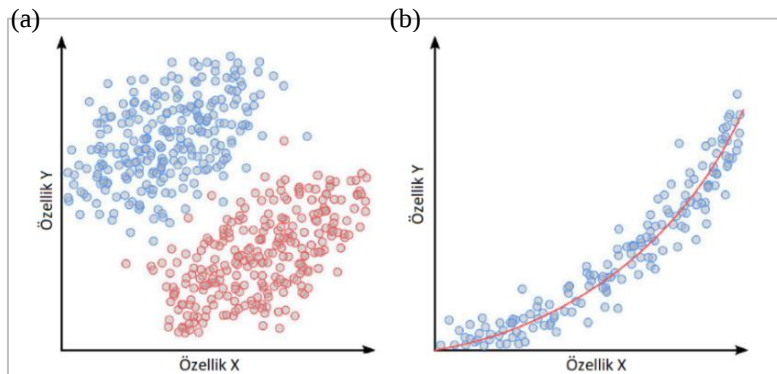
3. MAKİNE ÖĞRENMESİ

En basit haliyle yapay zeka, problem çözmeyi sağlamak için bilgisayar ve veri bilimini, istatistik ve matematiği birleştirerek kullanan bir dizi algoritmanın genel bir tanımlamasıdır. Sıkça bahsedilen makine öğrenmesi ve derin öğrenme ise yapay zekanın alt alanlarını oluşturur. Makine öğrenimi, bir bilgisayar programının insan müdahalesi olmadan öğrenebileceği ve yeni verilere uyum sağlayabileceği kavramdır, girdi verilerine dayalı tahminler veya sınıflandırmalar yapan sistemleri oluşturturur. Arthur Samuel (1959) makine öğrenimine, “Bilgisayara, açıkça programlanmadan öğrenme yeteneği veren çalışma alanıdır.” şeklinde bir tanımlama yapmıştır. Derin öğrenme ise, insan beyninin veri işleme ve karar verme sürecinde kullandığı kalıpları taklit edebilen bir yapay zeka (AI) işlevidir. Derin sinirsel öğrenme veya derin sinir ağı olarak da bilinir.

Makine öğrenmesinde, ilk olarak veriler elde edilir ve algoritmalarının en iyi çalışabileceği durumlara göre belirli ön-işlemden geçirilir. Öğrenme süreci ise ana hatlarıyla denetimli, denetimsiz ve takviyeli öğrenme olarak çeşitli kategoriler altında değerlendirilebilir. Bu süreçte verilerden kurallar ve kalıplar elde edilir ve uygun bir model oluşturulur. Öğrenme ile oluşturulan modelin çıktı sonucu hata analizi yapılarak güncellenir ve model geliştirilir. Elde edilen eğitim modeli yeni veriler üzerinde test edilir ve tahmin edebilme yeteneği tespit edilir.

3.1. Giriş

3.1.1. Sınıflandırma ve regresyon



Şekil 3.1. a) Sınıflandırma problemi, b) Regresyon problemi [93]

Tahmin edilen hedef değişken, az sayıda ayırık değere sahip ise bu öğrenme problemine sınıflandırma problemi adı verilir [94]. Kişinin cinsiyeti, evi olup olmadığı, belirli bir rahatsızlığı olup olmadığı, arabasının markası gibi problemler, sınıflandırma problemlerine örnek olarak gösterilebilir (Şekil 3.1.a). Hedef değişken sürekli olduğunda ise bu duruma regresyon problemi adı verilir. Gelir/gider hesabı, hava sıcaklığı, basınç, frekans, yaş, boy/kilo, kalp ritmi gibi sayısal değerlerin tahmini regresyon problemidir (Şekil 3.1.b).

3.1.2. Karışıklık matrisi

Sınıflandırma yaparken kullanılacak bir çok makine öğrenmesi algoritması bulunmaktadır. Bu algoritmalarından hangisinin belirli bir problem (veri) üzerinde daha iyi tahmin yapabilirdiği ve bu algoritmaların hangi noktada en iyi performansı verdiği belirlenmelidir. Bu süreçte her bir metod için Çizelge 3.1’ de gösterilen karışıklık matrisi oluşturulur.

Çizelge 3.1. Karışıklık matrisi

		Gerçek Sonuç	
		Hedef: 1 (Evet)	Hedef: 0 (Hayır)
Tahmin Edilen Sonuç	Hedef: 1 (Evet)	Gerçek Pozitif	Hatalı Pozitif
	Hedef: 0 (Hayır)	Hatalı Negatif	Gerçek Negatif

Çizelge 3.1’e göre gerçek pozitif değerler, sonucun 1 olduğunu ve bu sonucun doğru bir şekilde makine öğrenmesi algoritması tarafından tahmin edilebildiğini, benzer şekilde gerçek negatif değerler ise gerçek sonuç ile tahmin sonucunun örtüştüğünü ve bu sonucun 0 olduğunu ifade eder. Hatalı negatif değerler, sonucun 1 olduğunu ancak tahminin 0 olduğunu, hatalı pozitif değerler ise sonucun 0 olduğunu ancak tahminin 1 olduğunu ifade eder.

Veri üzerinde, algoritmanın kaç kez doğru ya da hatalı tahminde bulunduğu karışıklık matrisi ile belirlenir. Her bir algoritma için karışıklık matrisleri kurulur ve birbirleriyle karşılaştırılır, böylelikle daha yüksek tahmin yeteneğine sahip algoritma tercih edilir.

Karışıklık matrisi boyutları sınıflandırılacak hedef verinin çeşitliliği ile doğru orantılıdır. Çizelge 3.2, üç farklı sınıflandırma verisine ait bir karışıklık matrisinin temsili gösterimidir. Doğru tahmin edilen veri sayısı yeşil bölgeler içerisinde gösterilmiştir.

Çizelge 3.2. Karışıklık matrisi örnek gösterimi

		Gerçek Sonuç		
		Nesne Arabadır	Nesne Motosiklettir	Nesne Bisiklettir
Tahmin Edilen Sonuç	Nesne Arabadır	75	9	11
	Nesne Motosiklettir	5	41	51
	Nesne Bisiklettir	3	28	53

Uyarı: Her bir algoritma için oluşturulan karışıklık matrisi değerleri karar verilemeyecek kadar birbirine yakın olabilir. Bu durumda Sensitivity, Specificity ya da ROC, AUC gibi kavramlar, tercih sürecinde kullanılır.

3.1.3. Sensitivity ve specificity kavramları

“Sensitivity”, hedefin 1 olma durumunun, “Specificity” ise hedefin 0 olma durumunun hangi yüzde ile doğru tanımlandığını ifade eder. Böylece;

$$\text{Sensitivity} = \frac{\text{Gerçek Pozitif Sayısı}}{\text{Gerçek Pozitif Sayısı} + \text{Hatalı Negatif Sayısı}} \quad (3.1)$$

$$\text{Specificity} = \frac{\text{Gerçek Negatif Sayısı}}{\text{Gerçek Negatif Sayısı} + \text{Hatalı Pozitif Sayısı}} \quad (3.2)$$

Tablo 3.1.2.2 ile verilen örnek üzerinden açıklanacak olursa;

$$\text{Sensitivity}_{Araba} = \frac{74}{74+5+3} = 0,90$$

$$\text{Sensitivity}_{Motosiklet} = \frac{41}{41+28+9} = 0,52$$

$$\text{Sensitivity}_{Bisiklet} = \frac{53}{53+51+11} = 0,46$$

$$\text{Specificity}_{Araba} = \frac{41+51+28+53}{(41+51+28+53)+(9+11)} = 0.89$$

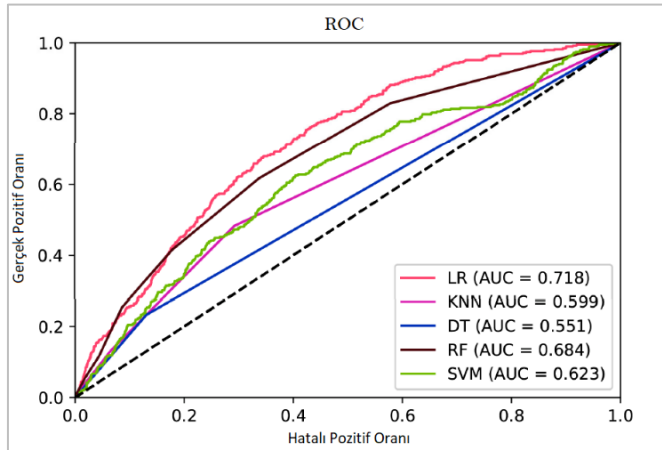
$$\text{Specificity}_{Motosiklet} = \frac{74+11+3+53}{(74+11+3+53)+(5+51)} = 0.71$$

$$\text{Specificity}_{Bisiklet} = \frac{74+9+5+41}{(74+9+5+41)+(11+51)} = 0.67$$

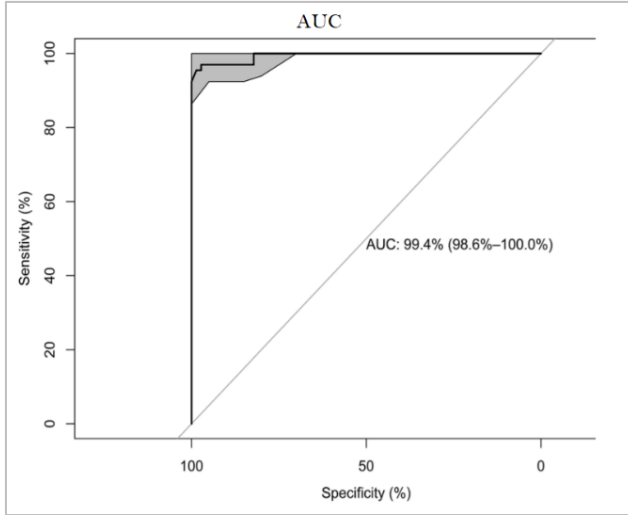
Tek bir makine öğrenmesi için oluşturulan karışıklık matrisine ait Sensitivity ve Specificity değerleri örnekte verildiği gibi hesaplanmıştır. Bu hesaplamalara göre saptanan nesnenin araba olduğu tahmini 90%, araba olmadığı tahmini ise 89% ile doğrudur. Farklı algoritmalara göre karışıklık matrisleri kurulmalı ve sensitivity/specificity değerleri hesaplanmalıdır. Farklı algoritmaların tercihi, nesnelerin tahmin edilebilirliğinin önemine göre değişkenlik gösterir.

3.1.4. ROC ve AUC grafikleri

Makine öğrenmesi algoritmaları yapılarına veya kullandıklarını verinin özelliklerine göre çok fazla sayıda karışıklık matrisine ihtiyaç duyabilir. Bu aşamada karışıklık matrisi ile elde edilecek bilgilerin özetlenebilmesi için Receiver Operator Characteristic (ROC) grafikleri (Şekil 3.2) tercih edilir. Bu grafikte y eksenini gerçek pozitif değerler oranını (sensitivity), x eksenini hatalı pozitif değerler oranını (1-specificity) verir. ROC grafiklerinde gerçek pozitif oranının yüksek, hatalı pozitif oranının düşük olduğu bölgeler, makine öğrenmesi algoritmaları için optimum performans verir.



Şekil 3.2. Makine öğrenmesi algoritmalarının ROC grafiği üzerindeki temsili gösterimi [95]



Şekil 3.3. AUC grafiği temsili gösterimi [96]

Area Under the Curve (AUC), Şekil 3.3'te gösterildiği gibi ROC eğrisi altında kalan bölgenin alanını temsil eder. Bu alanın büyüklüğü, kullanılan makine öğrenmesi algoritmasının performansı ile doğru orantılıdır.

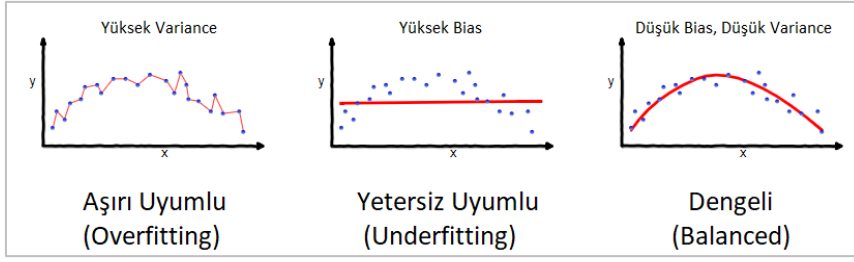
Uyarı: Veri setine göre Hedef: 0 olduğu durumların fazla olmasıyla, x ekseninde hatalı pozitif oranlar yerine precision (doğruluk) oranı kullanılabilir.

$$\text{Precision} = \frac{\text{Gerçek Pozitif Sayısı}}{\text{Gerçek Pozitif Sayısı} + \text{Hatalı Pozitif Sayısı}} \quad (3.3)$$

Sonuç olarak ROC grafikleri makine öğrenmesi için uygun eşik değerlerinin tespiti için kullanılırken, AUC grafikleri uygun algoritmanın tercihi için kullanılır.

3.1.5. Bias ve Variance

Makine öğrenmesi algoritmaları, doğaları gereği bazı hatalara sahiptirler. Uygun eğitim verilerinin yanı sıra uygun karmaşıklık ve esnekliğe sahip modellerin seçilmesi hata miktarının azalmasına neden olur. Bu noktada bias ve variance tanımları bilinmelidir.



Şekil 3.4. Bias ve variance kavramlarının gösterimi [98]

Bias, makine öğrenmesi algoritmasına ait fonksiyonun öğrenilmesini kolaylaştırmak için bir model tarafından yapılan basitleştirici varsayımlardır. Bu varsayımlar, modelin ortalama tahmini ile tahmin etmeye çalıştığı doğru değer arasındaki farka eşittir. Yüksek bias'a sahip model, eğitim verilerine çok az dikkat eder ve modeli aşırı basitleştirir (Şekil 3.4). Eğitim ve test verilerinde her zaman yüksek hataya yol açar. Genel olarak, doğrusal algoritmalar yüksek bias özelliğine sahiptir, bu şekilde öğrenme hızı artırılmış olur, ancak genellikle daha az esnektir ve algoritmaların basitleştirici varsayımlarını karşılamayan karmaşık problemler üzerinde daha düşük tahmin performansına sahiptirler [97].

Variance, farklı eğitim verilerinin kullanılması durumunda hedef fonksiyonun tahmininin değişeceği miktardır. Hedef fonksiyon, bir makine öğrenme algoritması tarafından eğitim verilerinden tahmin edilir, bu nedenle algoritmanın bir miktar varyansa sahip olması beklenir [98]. Yüksek varyansa sahip makine öğrenimi algoritmaları, eğitim verilerinin özelliklerinden güçlü bir şekilde etkilenir bu nedenle test aşamasında kullanılan veriler üzerinde, tahmin yeteneği düşüktür (Şekil 3.4).

3.1.6. Odds/log-odds

Odds ve log- odds kavramları olasılıksal yaklaşımlardır. Lojistik regresyon gibi bazı makine öğrenimi algoritmalarında kullanılmaktadırlar. Odds, başarı olasılığı p 'nin, başarısızlık olasılığı $(1-p)$ 'ye oranı olarak tanımlanır. Odds, kazanma ve kaybetme oranları arasında asimetrik bir yapı oluşturur. Bu nedenle log-odds (logit) kavramı tercih edilmektedir.

$$\text{Odds} = \frac{p}{1-p}, \quad p \in (0,1) \quad (3.4)$$

$$\text{Log-Odds} = \ln\left(\frac{p}{1-p}\right) \quad (3.5)$$

3.1.7. Keşifçi veri analizi (EDA)

Keşifçi veri analizi (EDA), veri bilimciler tarafından veri kümelerini analiz etmek ve araştırmak ve genellikle veri görselleştirme yöntemlerini kullanarak temel özelliklerini özetlemek için kullanılır [99]. Bu yöntem ile verilerin düzenlenmesi, anormalliklerinin giderilmesi, veriler üzerinde uygun istatistiksel yöntemlerin uygulanması ve verilerin sahip olduğu, değişkenler arası ilişkileri doğru anlayabilmek için gereklidir.

EDA araçlarını kullanarak, verileri görselleştirme, veriler ile hedef veriler arasındaki bağlantıları gözlemleme, denetimsiz öğrenme teknikleri ile verileri gruplama, eksik verileri tanımlama gibi bir çok görev yerine getirilebilir.

3.1.8. Validasyon (doğrulama) stratejileri

Validasyon, eğitilmiş bir modelin genelleme yeteneğinin test verileriyle değerlendirilmesidir [100]. Veri özelliklerine ve farklı senaryolara göre çeşitli validasyon stratejileri [101] bulunmaktadır. Bunlardan bazıları:

- Eğitim/Test Setlerinin Bölünmesi; en temel yöntemlerden biridir. Bu stratejide veri, rastgele olacak şekilde (genellikle eğitim için %70-80, test için %20-30) bölünür. Bu şekilde model performansı farklı veriler altında test edilir.
- “K-Fold” Çapraz Doğrulama; veriyi bir yerine birden çok kez bölerek kombinasyonlarını doğrular. Bu strateji ile modelin düşük bias’a sahip olması hedeflenir.
- İç İç Çapraz Doğrulama; eğitim veri kümesinde “overfitting” sorununun üstesinden gelmeye çalışan, hiperparametre optimizasyonu ve model seçimine yönelik bir yaklaşımdır.

3.1.9. Veri sızıntıları

Yüksek performansa sahip modeller, gerçek dünya verileri ile test edildiklerinde eski performanslarına ulaşamadıkları gözlemlenebilmektedir. Bu tutarsızlık durumuna veri sızıntısı adı verilir.

Tipik olarak, bir veri kümesini, eğitim ve test kümelerine bölerken amaç, ikisi arasında hiçbir verinin paylaşılmasını sağlamaktır. Veri sızıntısı genellikle test setinde gerçekçi olmayan yüksek performans seviyelerine neden olur, çünkü model, eğitim seti ile işlenmiş veriler üzerinde çalıştırılır. Model, eğitim seti verilerini etkin bir şekilde ezberler ve bu test veri seti örnekleri için değerleri doğru şekilde çıkarabilir. Böyle bir model gerçek dünyadan alınan yeni veriler üzerinde kullanıldığında, performans beklenenden çok daha düşük olacaktır. Verilerin, çapraz doğrulama öncesi doğru bir şekilde normalize edilmesi, eğitim/test/doğrulama olarak üç gruba ayrılması, veri seti içindeki birbiriyle eş değerlerin elimine edilmesi, veri sızıntılarını azaltmak için gerekli operasyonlardan bazılarıdır.

3.1.10. Veri kodlaması (encoding)

Makine öğrenmesi işlemleri sırasında verinin doğru şekilde biçimlendirilmiş olması performansı etkileyen önemli faktörlerden biridir. Ham veriler, işlenmesi gereken çeşitli yapılandırılmış veya yapılandırılmamış veri türlerini içermektedir. Yapılandırılmış veriler, sayısal ve kategorik veriler olarak sınıflandırılır. Yaygın olarak yapılandırılmış veri türlerinden bazıları, sürekli, ayırık, nominal, sıralı ve iki bileşenli verilerdir.

Sürekli veriler, sayısal veriler olarak sınıflandırılır ve belirli bir aralıkta herhangi bir değer alabilir. Ortalama ve standart sapma bulma, sürekli veriler üzerinde yapılabilecek aritmetik işlemlerdir. Ayırık veriler, sadece tam sayılar ve katlarını değer olarak tutabilen verilerdir. Otoparktaki arabaların sayısı, sınıftaki öğrencilerin sayısı örnek olarak gösterilebilir. Nominal veriler, kendisiyle ilişkilendirilmiş açık bir sıralaması olmayan kategorik verilerdir. Nominal veriler üzerinde ortalama, medyan veya standart sapma hesaplaması gibi herhangi bir istatistiksel işlem yapılamaz. Kişilerin sahip olduğu vatandaşlık numaraları buna örnek olarak verilebilir. Sıralı veriler, belirli bir sıraya sahip kategorik verilerdir. Sınavlardan alınan puanlar, bir otelin sahip olduğu yıldız sayısı bu tür verilere örnek olarak verilebilir. İki bileşenli veriler, herhangi bir sıra almayan ve girdi olarak sadece iki değer alabilen nominal verilerin özel bir türüdür. Bu veriler üzerinde gerçekleştirilebilecek işlemler, nominal veriler ile benzerdir.

Veri setinde bulunan özelliklerin ve veri tiplerinin belirlenmesinden sonraki adım, verilerin makine öğrenmesi modellerine uygun şekilde işlenmesidir. Makine öğrenmesi algoritmaları kategorik değişkenler üzerine doğrudan uygulanamaz bu nedenle değişkenlerin

sayısallaştırılması gerekmektedir. Bu işlem ise veri kodlaması olarak tanımlanır. Veri biliminde üç çeşit kodlama tekniği mevcuttur:

Etiket Kodlaması (Label Encoding); kategorik verilere 1'den N'ye kadar bir değer atanmasıdır (N, verilerde bulunan farklı kategorilerin sayısıdır). Bu tür bir kodlama tekniği, sıralı verilere uygulanır. Değerin atanması artan veya azalan bir sırada gerçekleşir. Etiket kodlaması tekniğine örnek olarak:

Çizelge 3.3. Etiket kodlaması tekniğine bir örnek

Film Türü	Kod
Aksiyon	0
Belgesel	1
Dram	2
Gerilim	3
Korku	4

“One-Hot” Kodlaması; özellikte bulunan farklı kategoriler için, özelliğin varlığına veya yokluğuna bağlı olarak 1 veya 0'dan oluşan bir vektör oluşturma işlemidir. Vektörlerin sayısı, veri kategorilerinin sayısına bağlıdır. Belirli bir özellik için mevcut kategorilerin sayısı çok fazlaysa, bu teknik veri kümesinde bulunan sütunların sayısını artırır ve algoritmanın öğrenme oranını önemli ölçüde azaltır. Bu teknik genellikle veri kümesinde bulunan nominal verilere uygulanır. One-hot kodlama tekniğine örnek olarak:

Çizelge 3.4. One-Hot kodlama tekniğine bir örnek

Araba Türü	Kod_Seda	Kod_Coupe	Kod_Sport	Kod_SW	Kod_Hb	Kod_Con	Kod_SUV	Kod_Mv	Kod_PT
Sedan	1	0	0	0	0	0	0	0	0
Coupe	0	1	0	0	0	0	0	0	0
Sport	0	0	1	0	0	0	0	0	0
Station Wagon	0	0	0	1	0	0	0	0	0
Hatchback	0	0	0	0	1	0	0	0	0
Convertible	0	0	0	0	0	1	0	0	0
SUV	0	0	0	0	0	0	1	0	0
Minivan	0	0	0	0	0	0	0	1	0
Pickup Truck	0	0	0	0	0	0	0	0	1

İkili kodlama,; sütunların yalnızca iki kategoriye sahip olduğu One-Hot kodlamanın özel bir durumudur. Söz konusu sütundaki veriler 0 ya da 1 değerleridir ve hiyerarşi belirtmezler. İkili kodlama tekniğine örnek olarak:

Çizelge 3.5. İkili kodlama tekniğine bir örnek

Cinsiyet	Kod
Erkek	0
Kadın	1

3.1.11. Dengesiz veri setleri

Dengesiz veri kümeleri, çoğunluk (negatif) ve azınlık (pozitif) sınıflarının bulunduğu, sınıf dağılımının sınıflar arasında tekdüze olmadığı, sınıflandırma problemi için geçerli bir durumdur. Dengesiz veri kümeleri makine öğrenmesi algoritmalarının bir sınıfa ait değerlere daha çok odaklanmasına ve düşük skorlu tahminlerde bulunmasına neden olur. “Oversampling” ve “Undersampling” yöntemleri veri kümelerinin dengeli bir şekilde düzenlenmesi için kullanılır.

Undersampling, tüm sınıfları azınlık sınıfının miktarına indirgeyerek seviyeleri eşitlemek anlamına gelir. Bunu bir örnekle açıklamak gerekirse: Sınıf sütunu olarak da bilinen dört bağımsız sütun ve bir bağımlı özelliğe sahip 50 satırlık bir veri kümemiz olsun. Böylelikle sınıf sütununda dört etiket bulunacaktır: 1, 2, 3 ve 4. Etiket 1'in 39 örneği, etiket 2'nin 32 örneği, etiket 3'ün 29 örneği ve Etiket 4'ün 31 örneği vardır. Yukarıda bahsedilen veri kümesine yetersiz örnekleme uygulamak için, Etiket 1, Etiket 2, ve Etiket 4'ü Etiket 3 ile aynı miktarda örneğe indirgememiz gerekir. Böylece, bu özel durumda her etiketin her biri 29 örneğe sahip olacaktır. Oversampling ise farklı olarak, sınıfların satırlarını çoğunluk sınıfına eşit olacak şekilde çoğaltma işlemidir.

3.1.12. Ensembling (topluluk öğrenmesi)

Tek bir algoritmaya dayalı çalışan ve geniş veri setlerine sahip projelerin, yüksek doğruluk oranı yakalayabilmesi oldukça zordur. Algoritmaların kombinasyonları şeklinde oluşturulan birleşik modeller, topluluk öğrenmesi ile veri üzerinde genelleştirmeler yapabilmek ve daha düşük hata oranları elde etmek mümkündür. Topluluk öğrenmesi bazı teknikler kullanılarak elde edilebilir.

Bagging

Bagging, paralel olarak birbirinden bağımsız çalışan modellerin, rastgele ayrılan veri seti grupları ile çalıştığı ve nihai kararın oylama ile belirlendiği bir işlemdir. Bu teknik için kullanılan algoritmalar başlıcası “Random Forest” algoritmasıdır ve bootstrapping (Karar Ağacı Tabanlı Algoritmalar/Random Forest başlığı altında detaylıca açıklanmıştır) yöntemini kullanarak çalışır.

Boosting

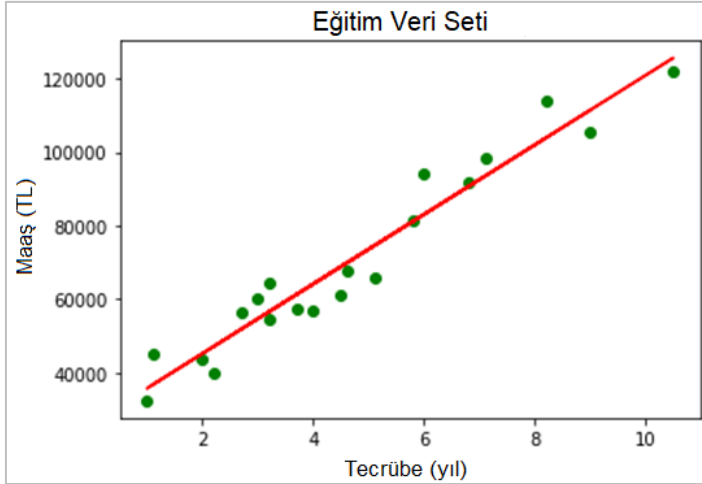
Boosting, genel olarak bias’ı azaltan ve güçlü tahmine dayalı modeller oluşturan sıralı bir topluluk yöntemidir. Zayıf öğrenenleri (rastgele tahminlerden biraz daha iyi sonuç veren modeller) bir araya getirerek güçlü bir öğrenen oluşturmaya hedefler. En sık kullanılan algoritmalar adaboost ve gradient boosting machine algoritmalarıdır. Boosting, sonraki eğitim adımlarını etkilemek için önceki eğitim adımlarının yinelenmelerinden alınan yanlış sınıflandırılmış verileri kullanarak, önceki hataların yeniden öğrenilmesine ve onların iyileştirilmesine olanak tanır.

Stacking

Stacking, çeşitli makine öğrenmesi algoritmaları kullanılarak oluşturulan tahminlerin ikinci katman öğrenme algoritmasında girdi olarak kullanıldığı verimli bir topluluk yöntemidir. Bu ikinci katman algoritması, yeni bir dizi tahmin oluşturmak için model tahminlerini oylama yöntemi ile birleştirir. Böylelikle, bir sınıflandırma veya regresyon görevinde bir dizi iyi performans gösteren modelin yeteneklerinden yararlanılabilmekte ve topluluktaki herhangi bir modelden daha iyi performansa sahip tahminler yapılabilmektedir.

3.2. Denetimli (Supervised) Öğrenme

Denetimli makine öğrenimi, hedef görevin sonuçlarını tahmin etmek için harici olarak sağlanan örnekleri kullanarak, genel modeller ve hipotezler üretebilen algoritmaların oluşturulmasıdır [102]. Denetimli makine öğrenmesi bir diğer deyişle, elde edilmiş verileri en iyi şekilde tanımlayacak $y = f(x)$ fonksiyonlarını üretir; x girdileri, y ise modelin sonuçlarını simgeler.



Şekil 3.5. Denetimli makine öğrenmesi modeli için oluşturulan eğitim seti [103]

Basitçe, edinilen tecrübenin kişinin maaşını ne kadar arttırdığına yönelik örnek bir grafik Şekil 3.5’de ki gibi verilmiştir. Bu örnekte yeşil noktalar kişileri temsil etmektedir. Kırmızı hat kişileri en iyi şekilde temsil edecek matematiksel bir yaklaşımdır. Bu yaklaşıma göre, işe alınacak yeni bir çalışanın maaşı diğer parametreler göz ardı edilerek kolayca hesaplanabilir. Örneğin 7.5 yıllık tecrübeye sahip bir kişinin maaşı yaklaşık 9800 TL olacaktır. Bu örnekte girdi sayısı sadece birdir (tecrübe). Ancak girdi sayısı çoğaltılarak yeni yaklaşımlar elde edilebilir. İki girdili bir problemde grafik üç boyutlu olacaktır. Daha fazla girdili problemlerde ise model görsel olarak temsil edilemez ancak matematiksel olarak formüle edilebilir.

Denetimi makine öğrenmesinin bir çok farklı yaklaşımı (algoritması) mevcuttur ve yapılan çalışmalarla bu yaklaşımlar geliştirilmeye devam edilmektedir [104-106]. Başlıca algoritmalar detaylı bir şekilde örnekler verilerek açıklanmıştır.

3.2.1. Karar ağacı tabanlı olmayan algoritmalar

Doğrusal regresyon

Doğrusal regresyon, gözlemlenen veriler ile doğrusal bir denklemi eşleştirerek iki değişken arasındaki ilişkiyi modellemeye çalışır [107]. Birbirleriyle ilişkilendirilebilir olması gereken bu değişkenlerden birisi bağımsız iken, diğeri bağımlı değişken olarak kabul edilir.

Doğrusal bir regresyon çizgisi, x 'in bağımsız, $h_0(x)$ 'nin bağımlı değişken olduğu $h_0(x) = \theta_0 + \theta_1 x$ biçiminde bir denkleme sahiptir. Doğrunun eğimi θ_1 iken, θ_0 kesme noktasıdır ve bu parametreler ağırlık (weights) olarak adlandırılırlar. Bir örnek üzerinden açıklanacak olursa;

Çizelge 3.6. Örnek bir veri seti

Arabanın Ağırlığı (kg)	Beygir Gücü (hp)	Yakıt Tüketimi (MPG)
999	110	12
998	67	9.5
1969	235	17

$$h_0(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad (3.6)$$

Burada x 'ler R^2 'deki iki boyutlu vektörlerdir (Şekil 3.6). Örneğin, $x_1^{(i)}$, eğitim setindeki i . arabanın ağırlığı ve $x_2^{(i)}$, beygir gücü, $h_0(x)$ ise hedef parametre olan arabanın harcadığı yakıt miktarıdır. Bu denklem genelleştirilerek $x_0 = 1$ için;

$$h(x) = \sum_{i=0}^d \theta_i x_i = \theta^T x \quad (3.7)$$

olarak yazılabilir. Burada x ve θ vektör, d ise girdi parametrelerinin sayısıdır. Tahmin sonucu elde edilen değerler, gerçek değerlerle örtüşmelidir. Bu nedenle θ parametresinin alacağı değerlerin en iyiye yakın olması gerekir. Bu noktada rastgele seçilen θ değerlerinin denklemde yerine yazılmasıyla bulunan $h(x)$ ile gerçek değerler olan (örnekte aracın tükettiği yakıt miktarı) y arasındaki fark hesaplanmalı ve θ değerleri devamlı güncellenerek bu fark en aza indirgenmelidir. Bu fark cost function (maliyet fonksiyonu) kullanılarak bulunur, burada n verinin boyutudur;

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad (3.8)$$

θ değerleri “Gradient Descent Algorithm” yöntemi ile güncellenir. Bu yöntem;

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (3.9)$$

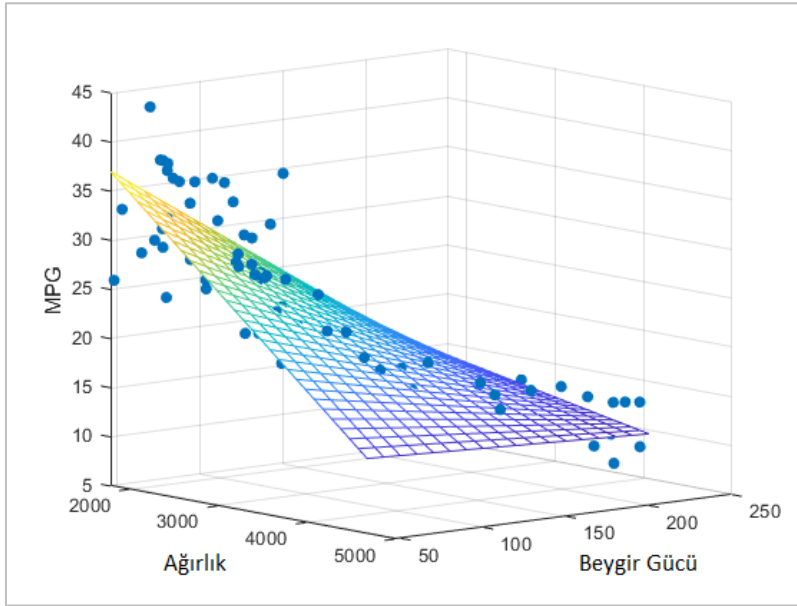
$j=0,1,2,3,\dots,d$ aynı an da olacak şekilde uygulanır. α azalma miktarını değiştiren sabit bir değerdir. Genellikle 0,1 ya da 0,01 olabilmektedir. Denlemde $J(\theta)$ yerine yazıldığında ise denlemin açık hâli;

$$\theta_j := \theta_j - \alpha \left(y^{(i)} - h_{\theta}(x^{(i)}) \right) x_j^{(i)} \quad (3.10)$$

Denklemin son hâli ise;

$$\theta := \theta - \alpha \sum_{i=1}^n \left(y^{(i)} - h_{\theta}(x^{(i)}) \right) x^{(i)} \quad (3.11)$$

olacaktır. Optimum θ değerleri ile oluşturulan matematiksel fonksiyon, eğitim verileri ile yüksek performanslı çalışacaktır. Ancak bu uyum modelin overfit olmasına neden olabilmektedir. Bu nedenle düzenlileştirme yöntemleri kullanılmaktadır.



Şekil 3.6. İki değişkenli doğrusal regresyon modeli [108]

Regularization (düzenlileştirme)

Düzenlileştirme, işlevi verilen eğitim setine uygun şekilde hataları azaltmak ve overfit olma durumunu önlemek için kullanılan bir tekniktir. Yaygın olarak kullanılan düzenlileştirme teknikleri şunlardır:

Losso (L1) Regresyonu:

$$J(\theta) = \sum_{i=1}^n (y_i - \sum_{j=0}^d (\theta_j x_{ij}))^2 + \lambda \sum_{j=0}^d (|\theta_j|) \quad (3.12)$$

Ridge (L2) Regresyonu:

$$J(\theta) = \sum_{i=1}^n (y_i - \sum_{j=0}^d (\theta_j x_{ij}))^2 + \lambda \sum_{j=0}^d (\theta_j^2) \quad (3.13)$$

Elastic-Net Regresyonu:

$$J(\theta) = \frac{\sum_{i=1}^n (y_i - \sum_{j=0}^d (\theta_j x_{ij}))^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=0}^d (\theta_j^2) + \alpha \sum_{j=0}^d (|\theta_j|) \right) \quad (3.14)$$

Lojistik regresyon

Lojistik regresyon bir sınıflandırma algoritmasıdır. Tüm regresyon analizleri gibi, lojistik regresyon da tahmine dayalı bir analizdir. Lojistik regresyon, verileri tanımlamak ve bir bağımlı değişken ile bir veya daha fazla sıra veya oran düzeyinde bağımsız değişken arasındaki ilişkiyi açıklamak için kullanılır. Bir başka deyişle lojistik regresyon, özellikle kategorik değişkenlerin sınıflandırılmasında kullanılan ve özellikler ile belirli sonuçlar arasında olasılıksal bir ilişki bulmaya odaklanan bir makine öğrenmesi algoritmasıdır. Lojistik terimi, modellenen log olasılık olasılıklarını ifade eder. Odds terimi, bir olayın meydana gelme olasılığının gerçekleşmeme olasılığına oranı olarak tanımlanır.

$$\text{odds} = \frac{P(\text{olay})}{1-P(\text{olay})}, \quad (3.15)$$

$$P(y = 1|X) = p(x), \quad X \in \mathbb{R}, \quad p(x) \in [0,1] \text{ için}$$

Lojistik regresyon modeli, lineer denklemin çıktısını 0 ile 1 arasında sıkıştırmak için sigmoid fonksiyonunu kullanır. 'z', logit'e bölünen log oranlarına karşılık geldiğinde, lojistik fonksiyon şu şekilde tanımlanır:

$$f(z) = 1/(1 + e^{-z}) \quad (3.16)$$

$$\log \left(\frac{p(x)}{1-p(x)} \right) = z = \beta X \quad (3.17)$$

Lojistik regresyon, parametre tahmini için Maksimum Likelihood kullanır:

$$l(\beta) = \sum_{i=1}^n y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i)) = \sum_{i=1}^n y_i \beta x_i - \log(1 + e^{\beta x_i}) \quad (3.18)$$

$$\text{ve } \beta = \underset{\beta}{\operatorname{argmax}} l(\beta) \quad (3.19)$$

Bu son olabilirlik denklemi, logaritmalar ve üsler gibi cebirsel olmayan terimlerden oluşur. Bu tür denklemlere “transcendental” denklemler de denir ve tam olarak hesaplanamazlar. Ancak sayısal yöntemler yaklaşım için kullanılabilir. Newton Raphson Metodu dikkate alınırsa:

$$\beta^{t+1} = \beta^t - \frac{\nabla_{\beta} l(\beta^t)}{\nabla_{\beta\beta} l(\beta^t)} = \beta^t + (X^T W^t X)^{-1} X(Y - \hat{Y}^t), \quad (3.20)$$

burada beta katsayısına göre $\nabla_{\beta} l(\beta^t)$ birinci, $\nabla_{\beta\beta} l(\beta^t)$ ise ikinci türevi ifade etmektedir.

K - En yakın komşu

K-en yakın komşular (KNN), test verileri ile tüm eğitim noktaları arasındaki mesafeyi hesaplayarak test verileri için doğru sınıfı tahmin etmeye çalışan bir tür denetimli öğrenme algoritmasıdır. KNN algoritması, test verilerinin, 'K' eğitim verilerinin sınıflarına ait olma olasılığını hesaplar ve en yüksek olasılığa sahip sınıf seçilir. Regresyon durumunda, değer seçilen 'K' eğitim noktalarının ortalamasıdır. K-NN çalışması aşağıdaki algoritma temelinde açıklanabilir:

1. Adım: Komşu verilerin sayısını ifade eden K katsayısı belirlenmelidir.

Uygun K değerini bulmak için tanımlanmış istatistiksel yöntemler bulunmamaktadır. Bu nedenle rastgele belirlenen bir K değeri işleme alınır ve yeni K değerleriyle karşılaştırılarak optimum değeri bulunur.

2. Adım: K sayıda komşunun uzaklıkları hesaplanmalıdır:

Test noktası ile her eğitim noktası arasındaki mesafe çeşitli yöntemler ile hesaplanabilmektedir. Bunlardan en yaygın olarak bilinenleri; Öklid, sürekli veriler için Manhattan ve kategorik veriler için Hamming yöntemleridir.

$$\text{Öklid} \quad : \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (3.21)$$

$$\text{Manhattan} \quad : \sum_{i=1}^k |x_i - y_i| \quad (3.22)$$

$$\text{Hamming} \quad : \sum_{i=1}^k |x_i - y_i| = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad (3.23)$$

3. Adım: Hesaplanan mesafelere göre en yakın K sayıdaki komşular belirlenmelidir.

4. Adım: Yeni veri, sınıflandırma işlemi için en yüksek komşu sayısına ait kategoriye atanmalı, regreyon işlemi için ise en yakın komşuların ortalamasına eşitlenmelidir.

Gaussian naive bayes

Gaussian Naive Bayes tahminçiler arasında bağımsızlığı varsayan oldukça pratik bir Bayes öğrenme tekniğidir. Bir başka deyişle, GNB, bir sınıftaki bir özneliliğin varlığının başka herhangi bir özneliliğin varlığı ile ilgili olmadığını varsayar. Matematiksel olarak ifade edilen Bayes kuralı koşullu olasılığı:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, X: \{X_1, \dots, X_m\} \text{ için; } [109] \quad (3.24)$$

$P(Y|X)$ = X koşulu sağlandığında Y koşulunun meydana gelme olasılığıdır.

$P(X|Y)$ = Y koşulu sağlandığında X koşulunun meydana gelme olasılığıdır.

$P(Y)$ = Y koşulunun meydana gelme olasılığıdır.

$P(X)$ = X koşulunun meydana gelme olasılığıdır.

GNB, koşullu olasılığı tahmin ederek çalışır ve y sınıf etiketi verildiğinde niteliklerin koşullu olarak bağımsız olduğunu varsayar. Koşullu bağımsızlık varsayımının matematiksel ifadesi şu şekilde verilir:

$$P(y = X|Y) = \prod_{i=1}^m P(y = X_i|Y) \quad [110] \quad (3.25)$$

Her bir öznitelik seti $X: \{X_1, \dots, X_m\}$, m öznitelik özelliklerinden oluşur. Bir test veri setini sınıflandırmak için GNB, Denklem 3.25'i kullanarak her Y sınıfının sonsal olasılığını hesaplar:

$$P(Y|X) = \frac{(\prod_{i=1}^m P(X_i|Y)) P(Y)}{P(X)} \quad [111] \quad (3.26)$$

$P(X)$ her Y için statiktir, bu nedenle $\prod_{i=1}^m P(X_i|Y)$ ifadesini maksimize eden sınıf seçilir. NBC, X_i 'nin sınıf koşullu olasılığını hesaplamak yerine, her bir X_i , verilen Y'nin koşullu olasılığını hesaplamak için koşullu bağımsızlık varsayımını kullanır.

Sinir ağları (neural networks)

Yapay sinir ağları (YSA) veya simüle edilmiş sinir ağları (SNN'ler) olarak da bilinen sinir ağları, makine öğreniminin bir alt kümesidir ve derin öğrenme algoritmalarının kalbinde yer alır. Adları ve yapıları, biyolojik nöronların birbirine sinyal gönderme şeklini taklit ederek insan beyninden esinlenmiştir.

Yapay sinir ağları (YSA), bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanı içeren bir düğüm katmanından oluşur (Şekil 3.7). Her düğüm veya yapay nöron diğerine bağlanarak ilişkili bir ağırlık ve eşikçe sahip olur. Herhangi bir düğümün çıktısı belirtilen eşik değerinin üzerindeyse, o düğüm etkinleştirilir ve ağırlık bir sonraki katmanına veri gönderilir. Aksi takdirde, ağırlık bir sonraki katmanına hiçbir veri iletilmez. Bu eşik değeri “aktivasyon fonksiyonları” ile belirlenir.

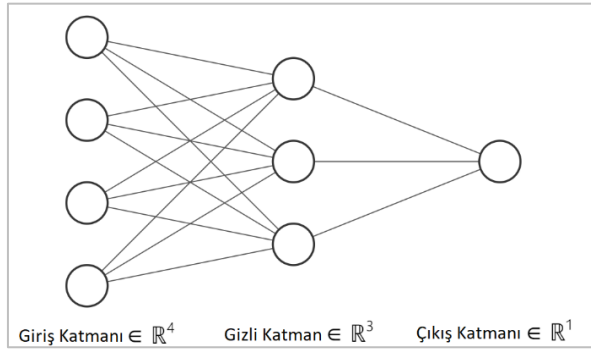
Sinir ağları, w , matris çarpımları ile x , giriş verilerinin kombinasyonlarını içeren, geniş türde doğrusal olmayan $h_\theta(x)$ modelleri/parametreleri anlamına gelir.

$$h_\theta(x) = \max(wx + b, 0) \quad (3.27)$$

b terimi genellikle "bias" olarak w vektörü ise "weight (ağırlık)" vektörü olarak adlandırılır. Burada $h_{\theta}(x)$ tek bir değer döndürür: pozitif değer için $wx + b$ veya sıfır. Sinir ağı bağlamında, $\max\{t, 0\}$ işlevine "Rectified Linear Activation Unit, ReLU" veya doğrultulmuş doğrusal birim denir ve genellikle $\text{ReLU}(t) \triangleq \max\{t, 0\}$ ile gösterilir. ReLU bir aktivasyon fonksiyonudur. Aktivasyon fonksiyonu, nöron için çıktı değerini hesaplar ve bu çıktı değeri farklı bir sinaps yoluyla sinir ağının bir sonraki katmanına iletilir. Genel olarak kullanılan aktivasyon fonksiyonları, Şekil 3.8'de ifade edilmiştir.

$x \in \mathbb{R}^d$ şeklinde, giriş verisi birden fazla boyuta sahip olduğunda, tek nöronlu bir sinir ağı şu şekilde yazılır:

$$h_{\theta}(x) = \text{ReLU}(w^T x + b), w \in \mathbb{R}^d, b \in \mathbb{R} \text{ ve } \theta = (w, b) \quad (3.28)$$



Şekil 3.7. Sinir ağı modeli

Denetimli öğrenme probleminde, maliyet (kayıp) fonksiyonu kullanarak tahmin edilen $h_{\theta}(x)$ değeri ile gerçek y değeri arasındaki hata miktarı hesaplanmalı ve modelin doğruluğu değerlendirilmelidir. Doğrusal regresyonda olduğu gibi maliyet fonksiyonu, ortalama kareler hatası (MSE) denklemi kullanılarak hesaplanır;

$$J(\theta) = \text{MSE} = \frac{1}{2m} \sum_{i=1}^m (y - h_{\theta}(x))^2 \quad (3.29)$$

Bu süreçte amaç, herhangi bir gözlem için uygunluğun doğruluğunu sağlayarak maliyet fonksiyonunu en aza indirmektir. Optimizasyon yöntemleri, model ağırlıklarını ve bias'ı ayarlarken, minimum değere ulaşmak için maliyet fonksiyonunu kullanır. Bazı optimizasyon yöntemleri ve ilgili denklemleri;

Gradient Descent: (θ , ağırlık/parametre olacak şekilde)

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (3.30)$$

Stochastic Gradient Descent: Gradient Descent'den farklı olarak tüm veri seti üzerinde çalışmaz.

$$\theta := \theta - \alpha \frac{\partial}{\partial \theta} J(\theta) \quad (3.31)$$

Stochastic Gradient Descent + Momentum: Gradient'lerin gelişmesini engelleyen yerel minimum değerleri için momentum değerleri eklenmiştir. ($v_0 = 0$, ilk hız ve $\rho=0,9$ olacak şekilde)

$$v_{t+1} = \rho v_t - \alpha \frac{\partial}{\partial \theta} J(\theta) \quad \theta := \theta + v_{t+1} \quad (3.32)$$

Nesterov Momentum:

$$v_{t+1} = \rho v_t - \alpha \frac{\partial}{\partial \theta} J(\theta + \rho v_t) \quad \theta := \theta + v_{t+1} \quad (3.33)$$

AdaGrad: (İlk G değeri 0 olacak şekilde)

$$G := G + \left(\frac{\partial}{\partial \theta} J(\theta)\right)^2 \quad \theta := \theta - \frac{\alpha \frac{\partial}{\partial \theta} J(\theta)}{\sqrt{G+10^{-7}}} \quad (3.34)$$

Adam: (İlk first_moment ve second_moment değeri 0, beta1=0.9, beta2=0.999 ve t , döngüdeki belirli bir an olacak şekilde)

$$\begin{aligned} \text{first_moment} &:= \text{beta1} \cdot \text{first_moment} + (1-\text{beta1}) \cdot \frac{\partial}{\partial \theta} J(\theta) \\ \text{second_moment} &:= \text{beta2} \cdot \text{second_moment} + (1-\text{beta2}) \cdot \left(\frac{\partial}{\partial \theta} J(\theta)\right)^2 \\ \text{first_unbias} &:= \frac{\text{first_moment}}{1-(\text{beta1})^t} \quad \text{second_unbias} := \frac{\text{second_moment}}{1-(\text{beta2})^t} \\ \theta &:= \theta - \frac{\alpha \cdot \text{first_unbias}}{\sqrt{\text{second_unbias} + 10^{-7}}} \end{aligned} \quad (3.35)$$

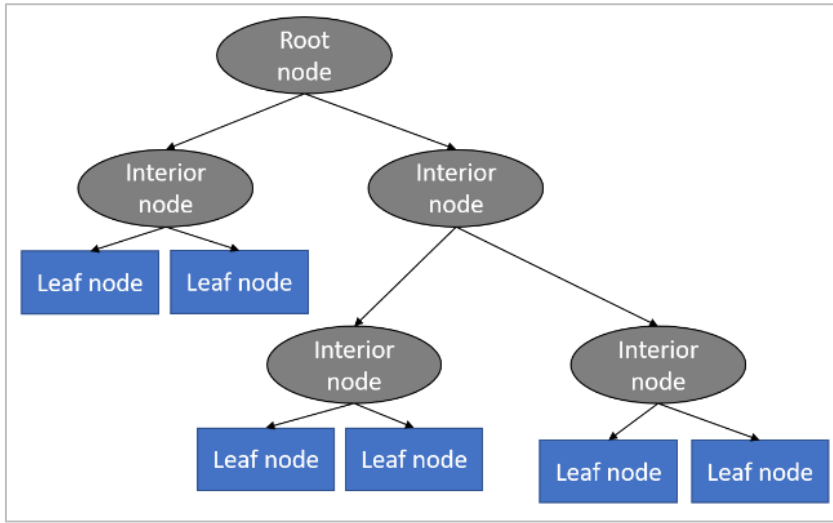
Aktivasyon Fonksiyonu	Fonksiyon Grafiği	Fonksiyon Denklemi	Denklemin x'e Göre Türevi	Değer Aralığı
Birim Fonksiyon		$f(x) = x$	$f'(x) = 1$	$(-\infty, \infty)$
İkili Sıvı Fonksiyonu (Binary Step)		$f(x) = \begin{cases} 0 & : x < 0 \\ 1 & : x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & : x \neq 0 \\ ? & : x = 0 \end{cases}$	$\{0, 1\}$
Sigmoid		$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$ [4]	$f'(x) = f(x)(1 - f(x))$	$(0, 1)$
Tanh		$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$	$f'(x) = 1 - f(x)^2$	$(-1, 1)$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2 + 1}$	$(-\frac{\pi}{2}, \frac{\pi}{2})$
Softsign		$f(x) = \frac{x}{1 + x }$	$f'(x) = \frac{1}{(1 + x)^2}$	$(-1, 1)$
Inverse Square Root Unit (ISRU)		$f(x) = \frac{x}{\sqrt{1 + \alpha x^2}}$	$f'(x) = \left(\frac{1}{\sqrt{1 + \alpha x^2}} \right)^3$	$(-\frac{1}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}})$
Inverse Square Root Linear Unit (ISRLU)		$f(x) = \begin{cases} \frac{x}{\sqrt{1 + \alpha x^2}} & : x < 0 \\ x & : x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \left(\frac{1}{\sqrt{1 + \alpha x^2}} \right)^3 & : x < 0 \\ 1 & : x \geq 0 \end{cases}$	$(-\frac{1}{\sqrt{\alpha}}, \infty)$
Square Nonlinearity (SQNL)		$f(x) = \begin{cases} 1 & : x > 2.0 \\ x - \frac{x^2}{4} & : 0 \leq x \leq 2.0 \\ x + \frac{x^2}{4} & : -2.0 \leq x < 0 \\ -1 & : x < -2.0 \end{cases}$	$f'(x) = 1 \mp \frac{x}{2}$	$(-1, 1)$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & : x < 0 \\ x & : x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & : x < 0 \\ 1 & : x \geq 0 \end{cases}$	$[0, \infty)$
Bipolar Rectified Linear Unit (BReLU)		$f(x_i) = \begin{cases} ReLU(x_i) & : i \bmod 2 = 0 \\ -ReLU(-x_i) & : i \bmod 2 \neq 0 \end{cases}$	$f'(x_i) = \begin{cases} ReLU'(x_i) & : i \bmod 2 = 0 \\ -ReLU'(-x_i) & : i \bmod 2 \neq 0 \end{cases}$	$(-\infty, \infty)$
Leaky Rectified Linear Unit (Leaky ReLU)		$f(x) = \begin{cases} 0.01x & : x < 0 \\ x & : x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0.01 & : x < 0 \\ 1 & : x \geq 0 \end{cases}$	$(-\infty, \infty)$
Parametric Rectified Linear Unit (PReLU)		$f(\alpha, x) = \begin{cases} \alpha x & : x < 0 \\ x & : x \geq 0 \end{cases}$	$f'(\alpha, x) = \begin{cases} \alpha & : x < 0 \\ 1 & : x \geq 0 \end{cases}$	$(-\infty, \infty)$ [2]
Randomized Leaky Rectified Linear Unit (RReLU)		$f(\alpha, x) = \begin{cases} \alpha x & : x < 0 \\ x & : x \geq 0 \end{cases}$ [3]	$f'(\alpha, x) = \begin{cases} \alpha & : x < 0 \\ 1 & : x \geq 0 \end{cases}$	$(-\infty, \infty)$
Exponential Linear Unit (ELU)		$f(\alpha, x) = \begin{cases} \alpha(e^x - 1) & : x \leq 0 \\ x & : x > 0 \end{cases}$	$f'(\alpha, x) = \begin{cases} f(\alpha, x) + \alpha & : x \leq 0 \\ 1 & : x > 0 \end{cases}$	$(-\alpha, \infty)$
SoftPlus		$f(x) = \ln(1 + e^x)$	$f'(x) = \frac{1}{1 + e^{-x}}$	$(0, \infty)$
Bent Identity		$f(x) = \frac{\sqrt{x^2 + 1} - 1}{2} + x$	$f'(x) = \frac{x}{2\sqrt{x^2 + 1}} + 1$	$(-\infty, \infty)$
Sigmoid Linear Unit (SiLU)		$f(x) = x \cdot \sigma(x)$ [5]	$f'(x) = f(x) + \sigma(x)(1 - f(x))$ [6]	$[\approx -0.28, \infty)$
Soft Exponential		$f(\alpha, x) = \begin{cases} -\frac{\ln(1 - \alpha(\pi + \alpha))}{\alpha} & : \alpha < 0 \\ x & : \alpha = 0 \\ \frac{e^{\alpha x} - 1}{\alpha} + \alpha & : \alpha > 0 \end{cases}$	$f'(\alpha, x) = \begin{cases} \frac{1}{1 - \alpha(\pi + \alpha)} & : \alpha < 0 \\ e^{\alpha x} & : \alpha \geq 0 \end{cases}$	$(-\infty, \infty)$
Soft Clipping		$f(\alpha, x) = \frac{1}{\alpha} \log \frac{1 + e^{\alpha x}}{1 + e^{\alpha(x-1)}}$	$f'(\alpha, x) = \frac{1}{2} \sinh\left(\frac{p}{2}\right) \operatorname{sech}\left(\frac{px}{2}\right) \operatorname{sech}\left(\frac{p}{2}(1-x)\right)$	$(0, 1)$
Sinus		$f(x) = \sin(x)$	$f'(x) = \cos(x)$	$[-1, 1]$
Sinc		$f(x) = \begin{cases} 1 & : x = 0 \\ \frac{\sin(x)}{x} & : x \neq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & : x = 0 \\ \frac{\cos(x)}{x} - \frac{\sin(x)}{x^2} & : x \neq 0 \end{cases}$	$[\approx -0.217234, 1]$
Gaussian		$f(x) = e^{-x^2}$	$f'(x) = -2xe^{-x^2}$	$(0, 1]$

Şekil 3.8. Aktivasyon fonksiyonları [112]

3.2.2. Karar ağacı tabanlı algoritmalar

Karar ağaçları (decision trees)

Karar Ağaçları, denetimli bir sınıflandırma yaklaşımıdır [113]. Bu yaklaşım root (kök) ve nodes (düğümler) gibi yapılardan oluşmaktadır (Şekil 3.9). Kavram olarak gerçek bir ağacın dalları ve yaprakları ilham olarak alınmıştır. [114].



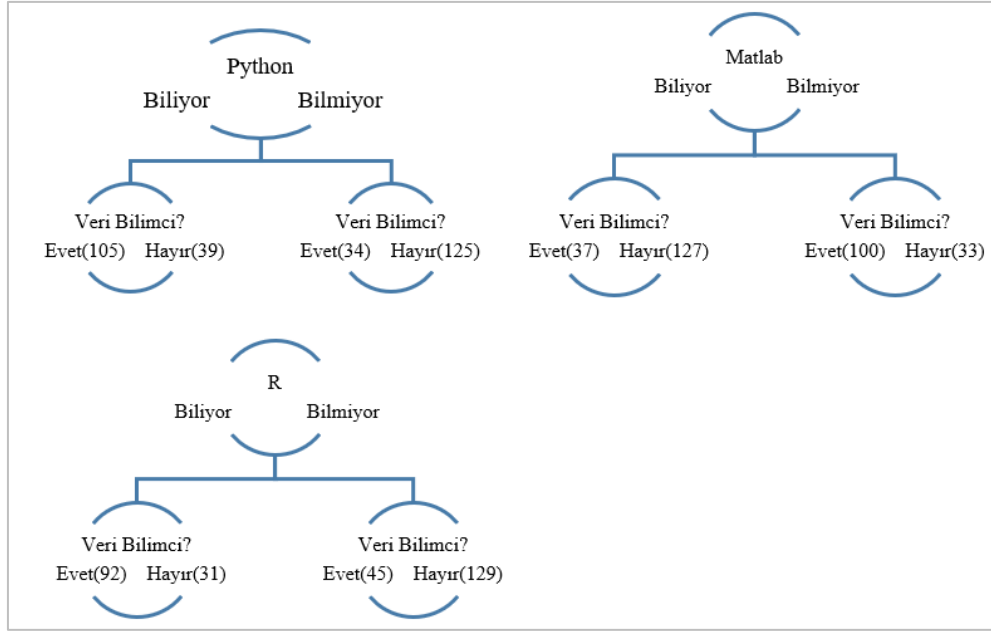
Şekil 3.9. Karar Ağacı yapısının temsili gösterimi [115] (İfadeler bilimsel terim olarak kullanılacağı için İngilizce olarak kullanımı tercih edilmiştir)

Karar ağacı algoritmalarının çalışma süreci açıklanırken bazı tanımlamalar bilinmelidir: Root Node, tüm ağacın başlangıç noktası olarak kabul edilir. Interior Node, karar verme sürecinde henüz sonuçlanmamış düğümleri ifade eder. Leaf Node ise, ağacın son düğümlerini ifade eder. Karar verme süreci burada son bulmuştur.

Karar ağaçları, temelde evet/hayır sorularının cevabı olarak kurulur. Çizelge 3.7’de verilen örnekte Python, Matlab ve R programlama dillerine hakim olan ya da olamayan insanların veri bilimci olup olmadıklarını tahmin eden bir karar ağacı oluşturulacaktır.

Çizelge 3.7. Örnek bir karar ağacı parametreleri

Python	Matlab	R	Veri Bilimci
Evet/Hayır	Evet/Hayır	Evet/Hayır	Evet/Hayır



Şekil 3.10. Programalama dillerine göre veri bilimci olan ya da olmayan kişi sayıları ifade edilmiştir

Verilen örnekte hiçbir leaf node 100% evet ya da hayır olmadığı için bu durum “impure” olarak ifade edilir. Root node tanımlanırken hangi yapının daha iyi olduğunu belirlemek gerekir. Bunun için ise “gini impurity” hesaplamaları yapılır ve sonuçlar karşılaştırılır.

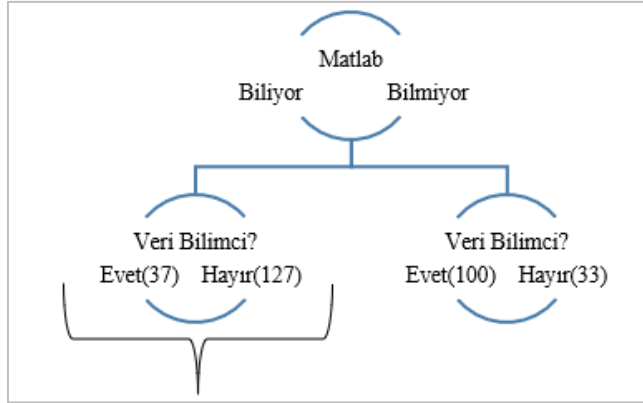
Her bir leaf için $\text{gini impurity} = 1 - (\text{evet olma olasılığı})^2 - (\text{hayır olma olasılığı})^2$ (3.36)

Python için gini impurity; $1 - \left(\frac{105}{105+39}\right)^2 - \left(\frac{39}{105+39}\right)^2 = 0.395$ (144 kişiyi temsil eder)

$$1 - \left(\frac{34}{34+125}\right)^2 - \left(\frac{125}{34+125}\right)^2 = 0.336 \text{ (159 kişiyi temsil eder)}$$

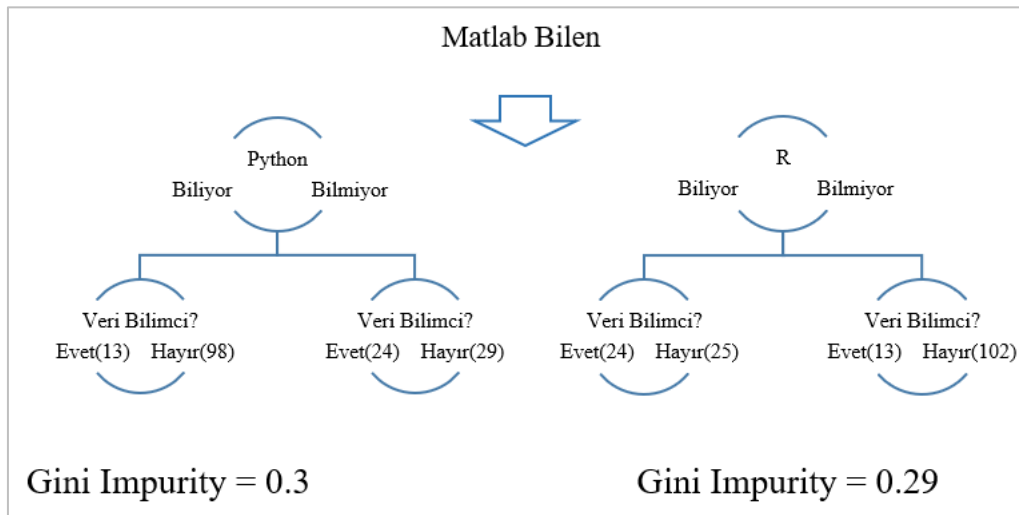
$$\frac{144}{144+159} \times 0.395 + \frac{159}{144+159} \times 0.336 = 0.364$$

Aynı strateji kullanılarak, Matlab için gini impurity = 0.36; R için gini impurity = 0.381 olarak bulunur. Bu hesaplamaların sonucunda Matlab için bulunan gini impurity değerinin daha düşük olması nedeniyle en iyi olarak tercih edilecektir.

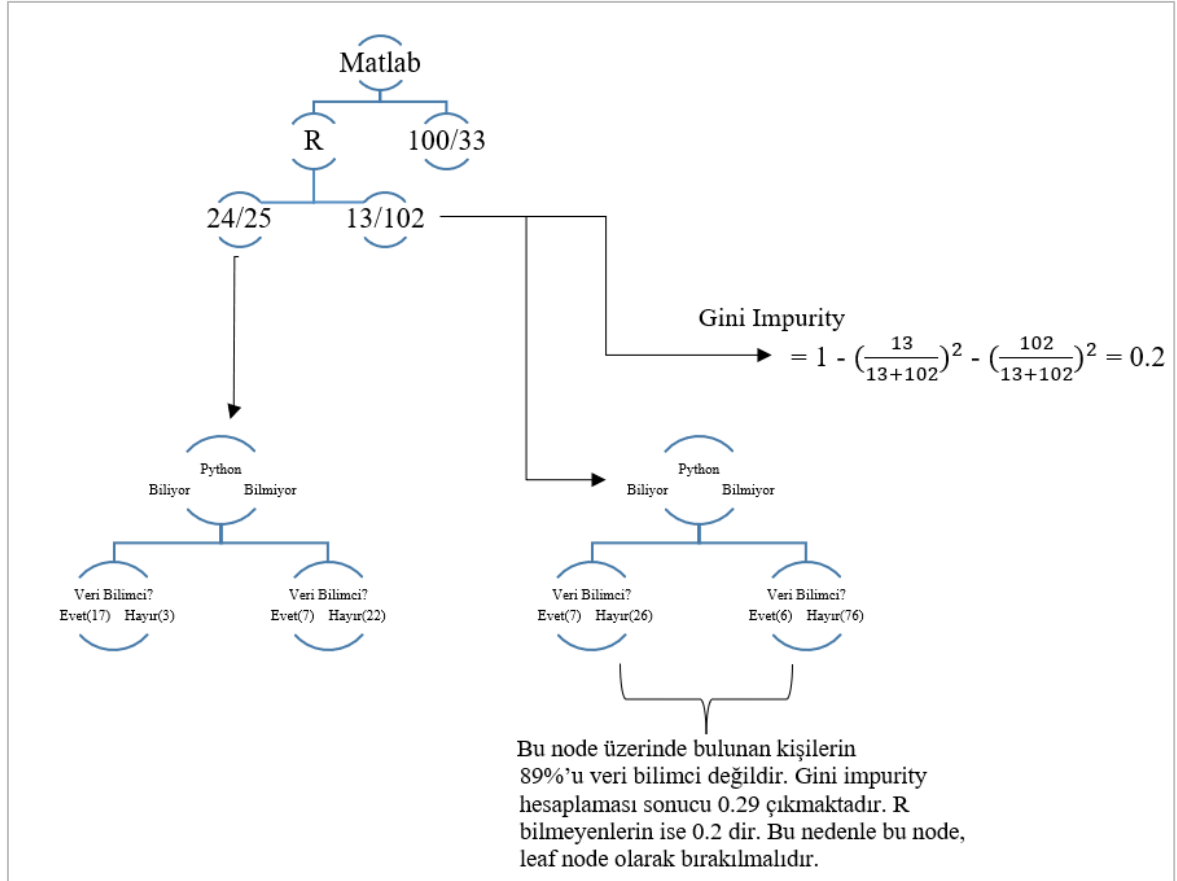


Şekil 3.11. Karar ağacı gösterimi

Bundan sonraki aşamada Matlab bilen kişilerin Python ve R proramlama dillerine göre durumları değerlendirilmelidir.

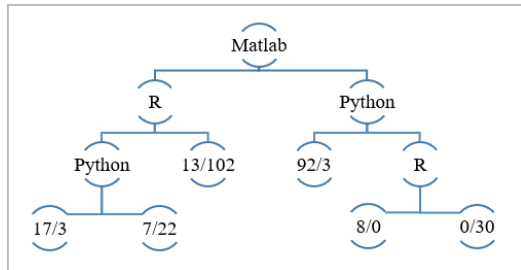


Şekil 3.12. Örnek karar ağacı oluşum süreci



Şekil 3.13. Örnek karar ağacı oluşum süreci

Kısaca tekrarlamak gerekirse karar ağacı oluştururken her bir parametre için gini impurity skorları hesaplanmalı, skor geliştiği sürece yeni node'lar oluşturup daha düşük skora sahip parametrelerle yola devam edilmeli ve en düşük skora sahip node, leaf node olarak bırakılmalıdır.



Şekil 3.14. Örnek karar ağacının son hâli

Oluşturulan karar ağacına göre Matlab, R ve Python deneyimi olan kişilerin 17 si veri bilimci olarak tanımlanırken, programlama dillerine hakim olmayan kişiler veri bilimci olarak tanımlanmamıştır.

Karar ağaçları oluşturulurken kullanılan parametreler, evet/hayır sorularına cevap verebildiği gibi bu parametreler numerik ifadeler içeriyor olabilir. Bu durumda izlememiz gereken adımlar değişiklik gösterecektir. Verdiğimiz örnekten devam edecek olursak kişinin veri bilimci olarak tanımlanması, kendi alanında ne kadar süreyle çalıştığıyla da ilişkilendirilebilir:

Çizelge 3.8. Örnek veriler

Süre (gün)	Veri Bilimci?
220	Evet
180	Evet
225	Evet
190	Hayır
155	Hayır

⇒ İlk olarak numerik veriler küçükten büyüğe doğru sıralanmalıdır.

Çizelge 3.9.Sıralandırılmış örnek veriler

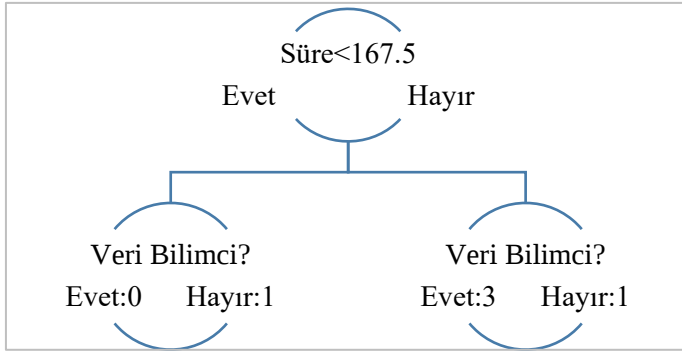
Süre (gün)
155
180
190
220
225

⇒ Sıralama tamamlandıktan sonra verilerin ikili gruplar hâlinde ortalamaları alınmalıdır.

Çizelge 3.10. Ortalamaları alınmış örnek veriler

Süre (gün)
155
180
190
220
225
167.5
185
205
222.5

⇒ Son olarak da her bir ortalama değeri için impurity skoru hesaplanmalıdır.



Şekil 3.15. Örnek-2 karar ağacı

Sol node için gini impurity: $1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2 = 0$

Sağ node için gini impurity: $1 - \left(\frac{3}{3+1}\right)^2 - \left(\frac{1}{3+1}\right)^2 = 0.375$

Genel: $\frac{1}{1+4} \times 0 + \frac{4}{1+4} \times 0.375 = 0.30$

Diğer ortalama değerler için elde edilen genel gini impurity skorları sırasıyla <185 için 0.47, <205 için 0.27, <222.5 için 0.4 tür. Bu durumda en küçük gini impurity değerini süre<205 vermiştir. Bu değer Python ya da R gibi programlama dillerinin bilinmesi durumuyla karşılaştırılacağı zaman bir alt node olarak kullanılabilir ya da farklı numerik verilerle karşılaştırılabilir.

Özellik seçimi ve eksik verilerin tamamlanması

Karar ağacı oluşturulurken gini impurity skorları esas alınır. Ağacın geliştirilmesi ise bu skorun azalma eğiliminde olmasına bağlıdır. Eğer parametrelerden birinin (örneğin Python) skorunda iyileştirmeye gidilemiyorsa kişilerin veri bilimci olup olmadığına dair ayırım, bu parametreler kullanılarak yapılamaz. Bu nedenle karar ağacı oluştururken bu parametrelerin ağaca dahil edilmemesi gerekir. Bu duruma ise özellik seçimi adı verilir. Karar ağacına sınırlama getirilmesi farklı yaklaşımlarla da sağlanabilir. Hesaplanan gini impurity skorları arasında büyük farklılıklar mevcut ise bir eşik değeri belirlenebilir ve karar ağacının genişleyerek overfit (makine öğrenmesi modelinin eğitim setlerine fazlasıyla adapte olması nedeniyle test setlerinde başarılı olamaması durumu, ezberleme (tezin ilerleyen bölümlerinde detaylıca açıklanacaktır)) olması engellenebilir.

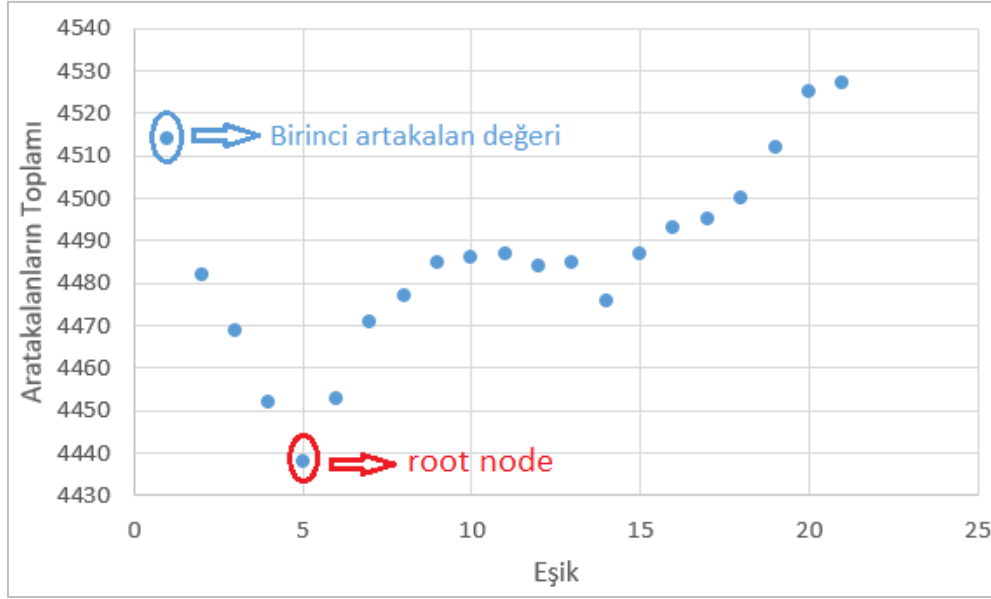
Parametreler oluşturulurken kullanılan yöntemler ve sistemler farklılık gösterir. Sensör verileri, istatistiksel veriler, araştırma verileri, vb. parametrelerin içeriğini oluşturur ve bu parametreler oluşturulurken veri kaybının olması çoğu zaman kaçınılmazdır. Eksik verilerin tamamlanması için bazı yöntemler uygulanabilir. Eğer parametrenin içeriği hayır'a göre daha fazla evet barındırıyorsa eksik olan kısım evet olarak güncellenebilir. Alternatif olarak farklı bir parametreyle bir kolerasyon yakalanabilir ve bu farklı parametrenin rehberliğinde işlem yapılabilir. Numerik verilerde ise eksik veriler, genelin ortalaması alınarak ya da verilerin ortancası ile tamamlanabilir. Bunun dışında kolerasyon yakalanan farklı bir parametreyle lineer regresyon modeli oluşturularak eksik veriler tahmin edilebilir.

Regresyon ağaçları

Regresyon ağaçları her bir leaf node'un numerik değerleri temsil etmesi için kurulur. Birden çok numerik parametreye sahip veri setleri için özellikle lineer regresyon algoritmasına göre daha efektif sonuçlar verir. Regresyon ağaçları oluşturulurken, eğitim verileri rastgele çizilen hayali bir çizgiyle iki gruba ayrılır. Her bir grubun ortalaması bulunur ve artakalan (residuals) hesabı yapılır:

$$\text{Artakalan} = (\text{İlk gruba ait değerler} - \text{ilk grubun ortalama değeri})^2 + (\text{ikinci gruba ait değerler} - \text{ikinci grubun ortalama değeri})^2 \quad (3.37)$$

Bu hesaplamalar hangi değerlerin root node, interior node ya da leaf node olacağını belirler, en küçük artakalan değeri root node olarak yerini alır.



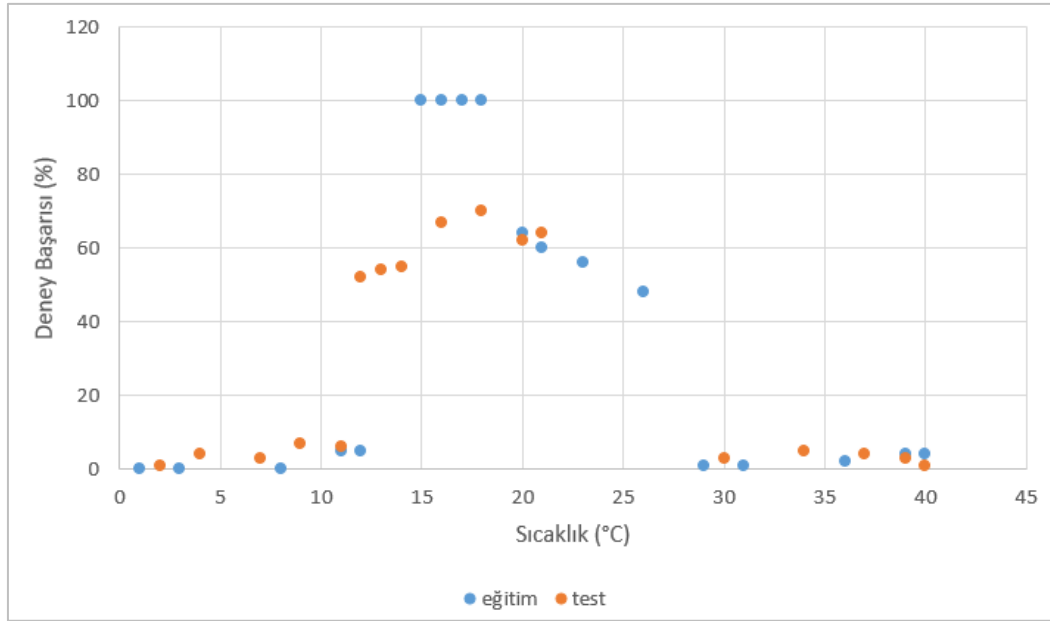
Şekil 3.16. Örnek veri setinin artakalan değerleri

Root node'un belirlenmesinden sonra interior node'lar bulunur. Çizilen hayali çizgi ile sağ ve sol gruplara ayrılan veriler kendi aralarında ikili gruplar oluşturur. Her bir grup için artakalan hesabı yapılır. Elde edilen değerler kendi aralarında karşılaştırılır ve en küçük değerler interior node'ları oluşturur. Bu süreç leaf node'lar elde edilene kadar devam eder.

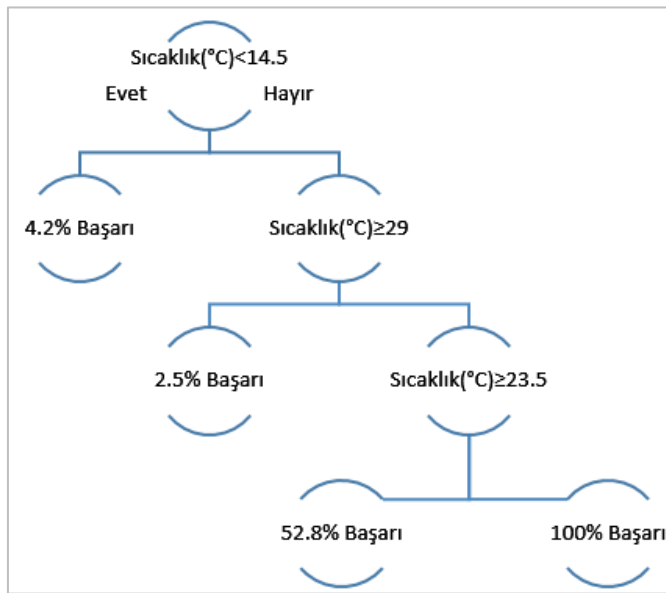
Uyarı: Kurulan bir regresyon modeli, eğitim setlerine mükemmel bir şekilde entegre olmuşsa muhtemelen overfit olmuştur ve yeni veriler için tahmin yeteneği zayıftır. Bir diğer deyişle model düşük Bias'a, yüksek Variance'a sahiptir. Modelin overfit olmaması için bazı çözüm yolları mevcuttur. Bunlardan en basit olanı, regresyon ağacı kurulurken yapılan gruplaşmalara bir sınırlama getirmektir (aynı grup içinde bulunan değerler arasındaki standart sapma çok küçük olmamalıdır).

Regresyon ağaçları için sınırlandırma (pruning)

Regresyon ağaçlarına sınırlandırma getirebilmek için bazı yöntemler kullanılmalıdır. Bu yöntemlerden biri "Cost Complexity Pruning (Weakest Link Pruning)" olarak adlandırılır.



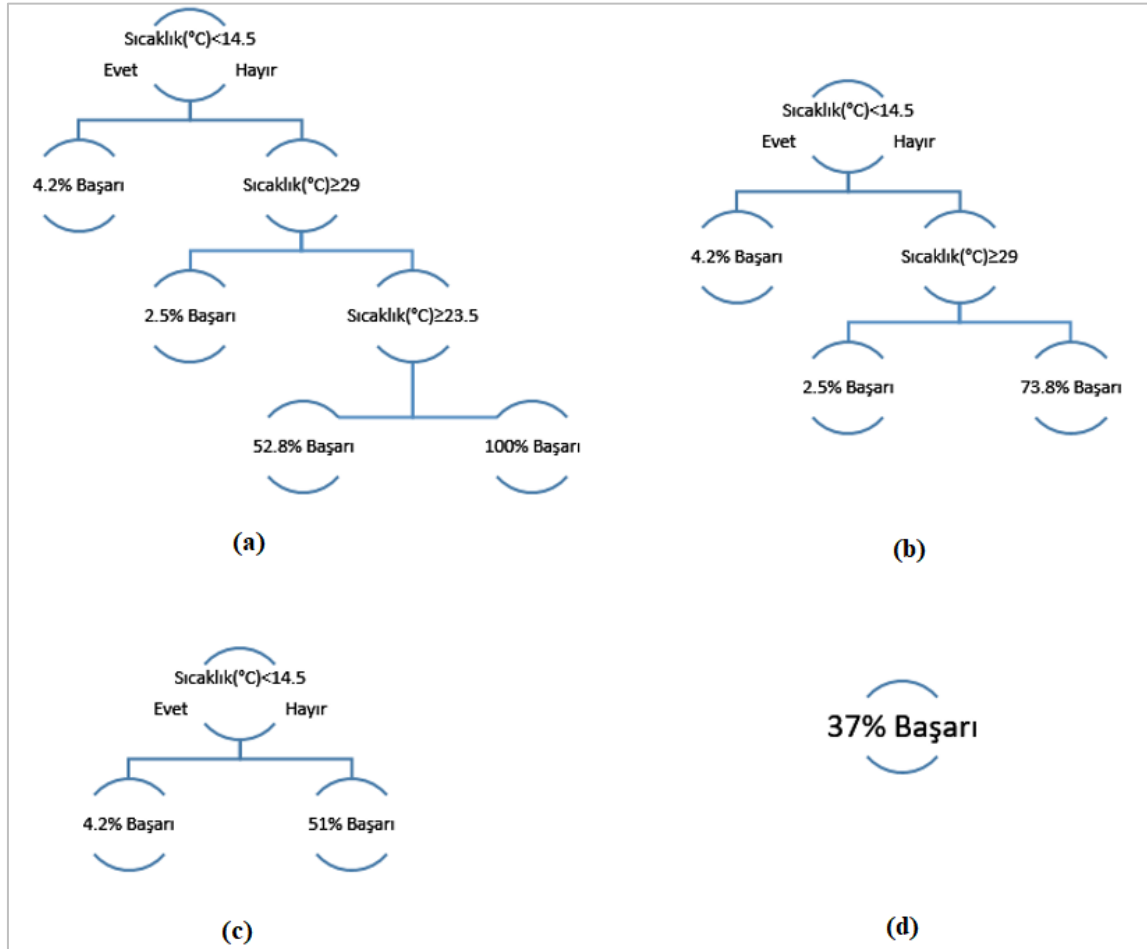
Şekil 3.17. Örnek olarak kullanılacak farazi bir deneyin sıcaklığa göre başarısını gösteren grafik



Şekil 3.18. Veri setleri verilen deneye ait karar ağacı

Örnekte eğitim setlerini başarılı bir şekilde yansıtabilen bir karar ağacı verilmiştir. Veri setlerinin bulunduğu grafik değerlendirildiğinde, başlangıçtaki eğitim ve test verilerinin birbirine yakın olduğu görülmektedir. Bu nedenle hesaplanacak olan artakalanlar birbirinden çok farklı çıkmayacaktır. Ancak sıcaklığın 20 ile 30 °C olduğu kısımda yer alan veriler arasında oldukça büyük farklılıklar vardır. Bu nedenle modelin overfit olduğu söylenebilir. Bazı leaf node'ların saf dışı bırakılması overfit probleminin giderilmesine yardımcı olabilir.

Bu örnek baz alındığında interior leaf olan $Sıcaklık(^{\circ}C) \geq 23.5$, 73.8% başarı olarak $((52.8 + 100)/2)$ değiştirilerek leaf node haline getirilebilir. Karar ağacında yapılan bu değişiklik eğitim setini önceki kadar iyi temsil edememektedir, ancak modelin test seti için tahmin edebilme yeteneği oldukça artırılmış olacaktır. Böylelikle modelin overfit olmasının önüne geçilecek ve sınırlandırma başarılı olacaktır. Interior node'ların leaf node'lar haline getirilmesiyle sınırlandırma devam edebilir:



Şekil 3.19. Sınırlandırmalar sonucu oluşan karar ağaçları

Bu süreçte önemli olan sınırlandırmanın nerede son bulması gerektiğidir. İlk olarak artakalanlar hesaplanmalıdır. Bu hesaplama doğrudan ilk olarak oluşturulan karar ağacı bütünü üzerinde uygulanarak gösterilecektir.

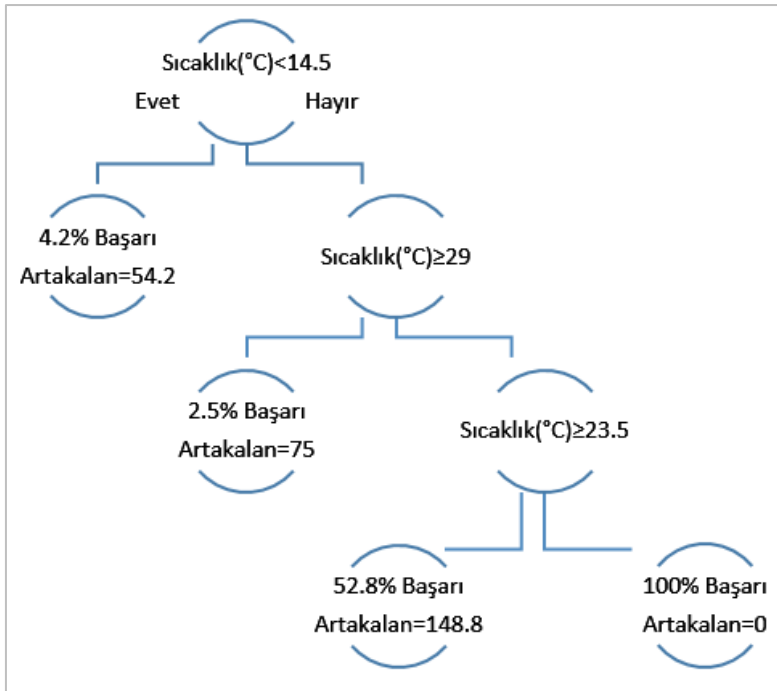
- $Sıcaklık(^{\circ}C) < 14.5$ olduğu durumda artakalan:

$$(0 - 4.2)^2 + (0 - 4.2)^2 + (0 - 4.2)^2 + (5 - 4.2)^2 + (5 - 4.2)^2 + (5 - 4.2)^2 = 54.2$$

Not: Sayılar grafikte bulunan Sıcaklık($^{\circ}\text{C}$)<14.5 koşulunu sağlayan değerler ve bu değerlerin başarı yüzdesidir.

- Sıcaklık($^{\circ}\text{C}$) ≥ 29 olduğu durumda artakalan: 75
- Sıcaklık($^{\circ}\text{C}$) ≥ 23.5 olduğu durumda artakalan: 148.8 ve 0

$$\text{Toplam} = 54.2 + 75 + 148.8 + 0 = 278$$



Şekil 3.20. Veri Setleri verilen deneye ait karar ağacı ve artakalan değerleri

Bu hesaplamalar sınırlandırılmış diğer karar ağaçlarına uygulandığında sırasıyla artakalanları 5494.8, 19243.7 ve 28897.2 çıkmaktadır. Görüldüğü üzere artakalanlar ağacın sınırlandırılmasıyla orantılı şekilde artmaktadır. Bunun nedeni sınırlandırılan ağacın git gide eğitim modeliyle yakaladığı senkronizasyonun azalmasıdır. Bu durumda Cost Complexity Pruning, ağaç skoru hesaplanarak bulunmalıdır:

$$\text{Ağaç Skoru} = \text{Artakalan} + \alpha T \quad (3.38)$$

Burada α sabit bir parametre (çapraz validasyon kullanılarak belirlenir, bu örnekte 10000 sayısı kullanılacaktır) ve T ağaçta bulunan leaf node sayısıdır. Bu durumda ağaç skorları:

$$278 + 10000 \times 4 = 40278$$

$$5494.8 + 10000 \times 3 = 35494.8$$

$$19243.7 + 10000 \times 2 = 39243.7$$

$$28897.2 + 10000 \times 1 = 38897.2$$

Bu durumda en küçük ağaç skoru bir kez sınırlandırılan karar ağacına aittir, eğitim ve test setleri bir arada düşünüldüğünde tercih edilmesi gereken modeldir.

Uyarı: α sabitinin 22000 olarak belirlendiği durumda, en düşük ağaç skoru, en fazla sınırlandırılan karar ağacına ait olacaktı. Benzer şekilde α sabitinin 15000 olarak belirlendiği durumda, en düşük ağaç skoru, iki kez sınırlandırılan karar ağacına ait olacaktı. Anlaşıldığı üzere α için tercih edilecek değerler karar aşamasında farklılık yaratmaktadır. En doğru α sabitinin belirlenebilmesi için yapılması gerekenler basittir. Eğitim ve test setleri için kullanılan verilerin tamamıyla yeniden bir karar ağacı oluşturulmalıdır (karar ağacı farklı olarak sadece eğitim setleriyle değil tüm veri seti kullanılarak oluşturulmalıdır). Artakalan değerleri yeniden hesaplanmalı ve her bir sınırlandırılmış alt-ağaç için en düşük ağaç skorları bulunmalıdır. Bu süreçte α sabiti sıfırdan başlatılarak arttırılır ve sınırlandırılmış alt-ağaçlara en düşük ağaç skorunu atayacak α sabitleri belirlenir. Daha sonra veri seti tekrardan eğitim ve test setleri oluşturacak şekilde bölünür, eğitim setleri kullanılarak yeni bir karar ağacı ve sınırlandırılmış alt-ağaçları oluşturulur. Sadece test seti verileri kullanılarak artakalanlar hesaplanır ve belleğe alınır. Bütün bir süreç tekrardan başlar (Veri seti yeni bir eğitim-test seti oluşturacak şekilde bölünür. Eğitim setlerine göre karar ağaçları oluşturulur. Test seti kullanılarak artakalanlar hesaplanır) ve 10-K Fold Cross Validation tamamlanana kadar devam eder. Belleğe alınan karar ağaçlarının artakalan değerinin ortalaması alınır ve en düşük ortalama sahip karar ağacının α değeri ideal olarak tanımlanmış olur. Son olarak tüm veri setinden oluşturulmuş orjinal karar ağaçlarına geri dönülür ve belirlenmiş α değerlerinden ideal olana sahip karar ağacı, aynı şekilde ideal karar ağacı olarak belirlenir.

Random forest

Random Forest, bir çok karar ağacından oluşan sınıflandırma, regresyon ve diğer görevler için kullanılan bir denetimli makine öğrenmesi algoritmasıdır. Karar ağaçları gibi basit yapılı olan ancak yeni veriler ile daha esnek çalışabilen bu algoritma, test sürecinde karar

ağaçlarına göre daha yüksek bir kesinlik sunar. Random Forest algoritması oluşturulurken belirli adımlar takip edilmelidir:

- i) Bootstrapped (önyüklenmiş) veri seti oluşturulur.

Çizelge 3.11. Bootstrapped veri seti oluşturmak için kullanılacak orjinal veri seti

Göğüs Ağrısı	Nefes Darlığı	Çarpıntı	Kişinin Kilosu (kg)	Kalpde Yağlanma
Hayır	Hayır	Hayır	57	Hayır
Evet	Evet	Evet	82	Evet
Evet	Evet	Hayır	95	Hayır
Evet	Hayır	Evet	76	Evet

Bootstrapped veri seti, orjinal veri setinden rastgele değerler alarak oluşturulur ve her bir değer birden fazla kullanılabilir.

Çizelge 3.12. Bootstrapped veri seti

Göğüs Ağrısı	Nefes Darlığı	Çarpıntı	Kişinin Kilosu (kg)	Kalpde Yağlanma
Evet	Evet	Evet	82	Evet
Hayır	Hayır	Hayır	57	Hayır
Evet	Hayır	Evet	76	Evet
Evet	Hayır	Evet	76	Evet

- ii) Bootstrapped veri seti kullanılarak, parametreler tamamen rastgele olacak şekilde bir karar ağacı oluşturulur.
- iii) Orjinal veri setinden yeni bir bootstrapped veri seti, ve bu veri setinden de yeni bir karar ağacı oluşturulur. Bu işlem, karar ağacı havuzu oluşturulana kadar devam eder (100 karar ağacı ideal olarak belirlenebilir).

Elde edilen çeşitlilik, tek bir karar ağacı yapısına göre random forest algoritmasının daha etkili çalışmasını sağlar. Bu aşamadan sonra algoritma, yeni bir veri ile test edilir. Yeni verideki parametreler, oluşturulan tüm karar ağaçları üzerinde test edilir, ve bir skor tablosu oluşturulur.

Çizelge 3.13. Skor tablosu

Kalpde Yağlanma	
Evet 79	Hayır 21

Skor tablosuna göre test setinin sonucu kişinin karaciğer yağlanmasıya sahip olduğunu gösterir. Bootstrapped veriler kullanılarak skor tablosu oluşturmak ve bu tablo üzerinden tahminde bulunmak “bagging” olarak ifade edilir. Random Forest algoritması bir ensemble modeldir (ilerleyen bölümlerde detaylıca açıklanacaktır).

Bootstrapped veri seti oluşturulurken tamamen rastgele seçimler yapılması, bazı verilerin kullanılmamasıyla sonuçlanabilir. Bu durum “Out of Bag Dataset” olarak isimlendirilir.

Çizelge 3.14. Out of Bag olarak tanımlanan verileri içeren tablo

Göğüs Ağrısı	Nefes Darlığı	Çarpıntı	Kişinin Kilosu (kg)	Kalpde Yağlanma
Evet	Evet	Hayır	5	Hayır

Karar ağaçları oluşturulurken kullanılmayan bu veriler, her bir ağaç üzerinde teste tabi tutulur ve kendi skor tablosu oluşturulur. Bu örnekte;

Çizelge 3.15. Out of Bag skor tablosu

Out of Bag Verisi Tahminleri	
Evet 37	Hayır 63

Bu durumda random forest algoritmasına ait 63 karar ağacı doğru tahmin yaparak veri sonucunun “hayır” olduğunu tespit etmiştir. Tüm Out of Bag verileri için bu işlem tekrarlanmalıdır. Doğru bir şekilde tahmin edilebilen Out of Bag verilerinin oranı bize kurulan Random Forest algoritmasının ne kadar kesin sonuçlar verebildiğini gösterir. Yanlış tahminde bulunulan veriler “Out of Bag Error” olarak ifade edilir.

Gradient boosted machines

Gradient Boosted Machines (GBM'ler), birçok alanda başarılı olduğu kanıtlanmış son derece popüler bir makine öğrenme algoritmasıdır ve Kaggle gibi veri bilimi yarışmalarını kazanmanın önde gelen yöntemlerinden biridir. GBM algoritmasını anlamak için temel bilgilerin edinilmesi gerekir [116].

Chain rule (zincir kuralı)

Zincir kuralı, bileşik fonksiyonlarla çalışmamızı sağlayan önemli bir türev kuralıdır [117]. Bileşik fonksiyon, iki veya daha fazla fonksiyondan oluşan bir yapıdır. f ve g iki fonksiyon olsun, o zaman zincir kuralı, $f \circ g$ veya $g \circ f$ gibi bileşik fonksiyonların türevini bulmamıza yardımcı olur. $f \circ g$ bileşik fonksiyonu göz önüne alındığında, zincir kuralı türevi şöyle görünür:

$$(f \circ g)' = (f' \circ g) \cdot g' \quad (3.39)$$

x, y, z değişkenlerine sahip üç değişkenli bir fonksiyonda, $z = f(y)$ ve $y = g(x)$ olduğunda zincir kuralı türevi şuna benzer:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx} \Rightarrow f'(y) g'(x) = f'(g(x)) g'(x) \quad (3.40)$$

Regresyon işlemi, GBM'ler için doğrusal regresyon işlemlerinden farklıdır. Bu problemi sırasıyla çözebilmek için;

Giriş Verisi: $\{(x_i, y_i)\}_{i=1}^n$, türevlenebilir Kayıp Fonksiyonu $L(x_i, F(x))$ olarak belirlenir.

Giriş verisi olarak verilen tanımlama, eğitim verisini ve tahmin edilmiş olan değerlerin gerçek değerlerle ne kadar örtüştüğünü ifade eder.

Çizelge 3.16. $\{(x_i, y_i)\}_{i=1}^n$ olarak ifade edilen basit bir veri setine ait tablo (n , veri setinin büyüklüğü, x girdi verileri, y ise çıktı verisidir)

Kişinin Boyu(metre)	Kişinin Sık Tükettiği Yiyecek	Kişinin Cinsiyeti	Kişinin Ağırlığı
1.6	Pizza	Erkek	88
1.6	Hamburger	Kadın	76
1.5	Makarna	Kadın	56

Kayıp Fonksiyonu $L(x_i, F(x))$ ise;

$$\frac{d}{dTahmin} \left(\frac{1}{2} (Gözlem - Tahmin)^2 \right) = - (Gözlem - Tahmin) \text{ olarak ifade edilir.} \quad (3.41)$$

$$1. \text{ Adım: Sabit bir değerle model kurulur: } F_0(x) = \operatorname{argmin} \sum_{i=1}^n L(y_i, \gamma) \quad (3.42)$$

$$F_o(x) = (Tahmin - 88) + (Tahmin - 76) + (Tahmin - 56) = 0; Tahmin = 73.3$$

Tahmin edilen 73.3 değeri bir leaf node'tur, ve $F_o(x)$ fonksiyonunun başlangıç değeridir. (Daha büyük problemlerin çözümünde optimum $F_o(x)$ değeri Gradient Descent ile hesaplanır.)

2. Adım: $m = 1, \dots, M$ için:

$$(a) r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad i = 1, \dots, n \text{ için hesaplanır.} \quad (3.43)$$

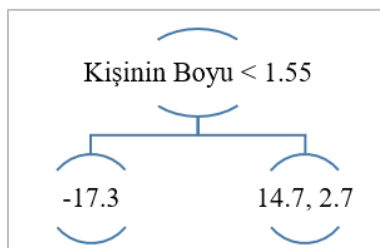
$m = 1, \dots, M$ bir döngüyü (loop) ifade eder ve tahmin sürecinde M adet karar ağacının kurulacağını söyler (pratikte $M=100$ olarak belirlenir), m ise her bir karar ağacını temsil eder.

r_{im} ifadesi ise artakalandır. $-\frac{d}{dTahmin} \left(\frac{1}{2} (Gözlem - Tahmin)^2 \right) = (Gözlem - Tahmin)$ 'e eşittir ve $F_{m-1}(x)$, $m = 1$ olduğunda $F_o(x) = 73.3$ 'tür. Bu şartlar altında hesaplanan ilk artakalanlar Çizelge 3.18'de gösterildiği gibi olacaktır.

Çizelge 3.17. Veri seti üzerinde hesaplanan ilk artakalan değerlerini gösteren tablo

Kişinin Boyu (metre)	Kişinin Sık Olarak Tükettiği Yiyecek	Kişinin Cinsiyeti	Kişinin Ağırlığı	Artakalanlar (r_{i1})
1.6	Pizza	Erkek	88	14.7
1.6	Hamburger	Kadın	76	2.7
1.5	Makarna	Kadın	56	-17.3

(b) r_{im} değerleri ile bir regresyon ağacı kurulur ve $j = 1, \dots, J_m$ olacak şekilde R_{jm} terminal bölgesi oluşturulur.



Şekil 3.21. Veri seti ile kurulan artakalanları gösteren basit bir regresyon ağacı

Leaf node'lar, R_{jm} (m, karar ağacı indeksi, j ise leaf node'ların indeksidir) terminal bölgesidir. Bu durumda $R_{1,1} = -17.3$, $R_{2,1} = 14.7$, 2.7 'dir.

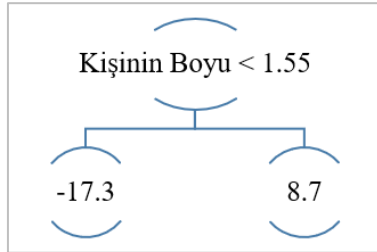
Uyarı: Birden fazla parametreye sahip GBM ağaçları her zaman daha geniş kurulur. Şekil 3.21 sadece bir gösterimdir.

$$(c) j = 1, \dots, J_m \text{ için } \gamma_{jm} = \operatorname{argmin}_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma) \text{ hesaplanır.} \quad (3.44)$$

Şekil 3.21'de görüldüğü üzere bazı leaf node'lar birden fazla değere sahip olabilir. Bu durum çıktı değerinin kesinleştirilmesini gerektirir. Değerler denklemde yerine yazılacak olursa;

$$\begin{aligned} \gamma_{2,1} &= \operatorname{argmin} \left[\frac{1}{2} (88 - (73.3 + \gamma))^2 + \frac{1}{2} (76 - (73.3 + \gamma))^2 \right] \\ &= \frac{d}{d\gamma} \left[\frac{1}{2} (14.7 - \gamma)^2 + \frac{1}{2} (2.7 - \gamma)^2 \right] = 0 \end{aligned}$$

γ buradan 8.7 olarak elde edilerek Şekil 3.22'deki gibi yazılır.



Şekil 3.22. Veri Seti ile kurulan ortakalanları gösteren basit bir regresyon ağacının son hâli

$$(d) F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{ij}) \text{ güncellenir.} \quad (3.45)$$

Bu bölümde bütün veriler için yeniden tahminde bulunulur. $F_m(x)$ tahmin edilirken, bir önceki tahmin olan $F_{m-1}(x)$ kullanılır. v ise 0 ile 1 arasında modelin overfit olmasını engelleyen bir değerdir. Verideki her bir örnek için;

$$F_1(x) = 73.3 + 0.1 \times 8.7 = 74.2 \text{ (Kişinin boyu 1.55 metreden büyük olduğu için)}$$

$$F_1(x) = 73.3 + 0.1 \times 8.7 = 74.2$$

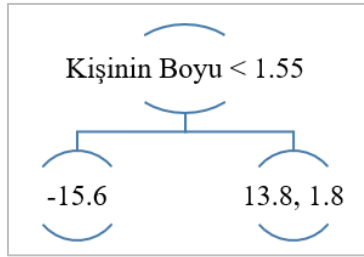
$$F_1(x) = 73.3 + 0.1 \times (-17.3) = 71.6$$

Yeni tahmin değerleri ilk tahmin olan 73.3'e göre gelişme kaydederek, gerçek değere yaklaşmıştır. Bu aşamalar M değeri kadar tekrar eder.

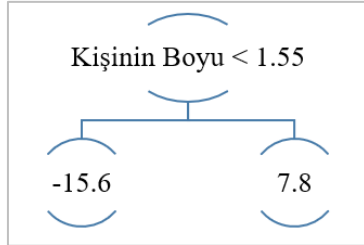
$m = 2$ için;

Çizelge 3.18. Veri seti üzerinde hesaplanan $m=2$ için artakalan değerlerini gösteren tablo

Kişinin Ağırlığı	Kişinin Tahmini Ağırlığı	Artakalanlar (r_{i2})
88	74.2	13.8
76	74.2	1.8
56	71.7	-15.6



Şekil 3.23. Veri seti ile kurulan $m=2$ için artakalanları gösteren basit bir regresyon ağacı



Şekil 3.24. Veri seti ile kurulan $m=2$ için artakalanları gösteren basit bir regresyon ağacının son hâli

$$F_2(x) = 73.3 + 0.1 \times 8.7 + 0.1 \times 7.8 = 74.98$$

$$F_2(x) = 73.3 + 0.1 \times 8.7 + 0.1 \times 7.8 = 74.98$$

$$F_2(x) = 73.3 + 0.1 \times (-17.3) + 0.1 \times (-15.6) = 70.14$$

3. Adım: $m=M$ için, $F_m(x)$ çıktısı güncel olarak elde edilir ve tahmin yeteneği yeni veriler ile test edilir.

Sınıflandırma işlemi, GBM’ler için lojistik regresyon işlemleriyle benzerlik gösterir. Örnek üzerinden gidilecek olursa:

Çizelge 3.19. Sınıflandırma için örnek veri setini gösteren tablo

Kişi Drama Seviyor mu?	Kişinin Yaşı	Kişinin Sevdiği Renk	Kişi “TC” Serisini Seviyor mu?
Evet	12	Mavi	Evet
Hayır	87	Beyaz	Evet
Hayır	44	Mavi	Hayır

Verilen veri setinde, kişinin “Trois Couleurs” serisini sevip sevmediği tahmin edilecektir.

Giriş Verisi: $\{(x_i, y_i)\}_{i=1}^n$, türevlenebilir kayıp fonksiyonu $L(x_i, F(x))$ olarak belirlenir.

Lojistik regresyonda olduğu gibi $\log(\text{likelihood})$, kayıp fonksiyonu olarak kullanılır. Örnekte olduğu gibi küçük veri setleri tercih edildiğinde, modelin daha uygun oluşturulabilmesi için denklem “—” değerini alır. y_i ise eğitim setinde gözlemlenen hedef değerlerdir ve sırasıyla 1,1,0 olduğu bilinmektedir;

$$\text{Olasılık, } p = \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}} \quad (3.46)$$

$$\log(\text{likelihood}) = \text{Kayıp Fonksiyonu} = \sum [y_i \times \log(p) + (1 - y_i) \times \log(1 - p)] \quad (3.47)$$

$$\Rightarrow -[y \times \log(p) + (1 - y) \times \log(1 - p)]$$

$$\Rightarrow -y \times \log(p) - (1 - y) \times \log(1 - p)$$

$$\Rightarrow -y \times \log(p) - \log(1 - p) + y \times \log(1 - p)$$

$$\Rightarrow -y \times [\log(p) - \log(1 - p)] - \log(1 - p)$$

$$\Rightarrow -y \times \left[\frac{\log(p)}{\log(1-p)} \right] - \log(1 - p)$$

$$\Rightarrow -y \times \left[\log \frac{p}{(1-p)} \right] - \log(1 - p)$$

$$\Rightarrow -y \times \left[\log \frac{p}{(1-p)} \right] - \log(1 - p)$$

$$\Rightarrow -y \times \log(\text{odds}) - \log(1 - p)$$

$$\Rightarrow -y \times \log(\text{odds}) - \log \left(1 - \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}} \right)$$

$$\Rightarrow -y \times \log(\text{odds}) - \log \left(\frac{1}{1 + e^{\log(\text{odds})}} \right)$$

$$\begin{aligned}
&\Rightarrow -y \times \log(odds) - [\log(1) - \log(1 + e^{\log(odds)})] \\
&\Rightarrow -y \times \log(odds) - [-\log(1 + e^{\log(odds)})] \\
&\Rightarrow -y \times \log(odds) + \log(1 + e^{\log(odds)})
\end{aligned}$$

Türevlenebilir kayıp fonksiyonu buradan;

$$\frac{d}{d \log(odds)} - y \times \log(odds) + \log(1 + e^{\log(odds)})$$

$$\Rightarrow -y + \frac{1}{1 + e^{\log(odds)}} \times e^{\log(odds)} \quad (3.48)$$

$$\Rightarrow -y + p \text{ olarak bulunur.} \quad (3.49)$$

$$1. \text{ Adım: Sabit bir değerle model kurulur: } F_o(x) = \operatorname{argmin} \sum_{i=1}^n L(y_i, \gamma) \quad (3.50)$$

$L(y_i, \gamma)$ kayıp fonksiyonunda y_i gözlemlenen değerler, γ ise $\log(odds)$ karşılığıdır. Buradan;

$$(-1+p) + (-1+p) + (0+p) = 0, \quad p = 2/3$$

$$F_o(x) \Rightarrow \gamma = \log(odds) = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{2/3}{1-2/3}\right) = \log(2) = 0.69$$

Bu değer GBM karar ağacı oluşturulurken kullanılan başlangıç leaf node'sidir.

2. Adım: $m = 1, \dots, M$ için:

$$(a) \quad r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad i = 1, \dots, n \text{ için hesaplanır.} \quad (3.51)$$

$$- \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \text{ ifadesi,}$$

$$- \frac{d}{d \log(odds)} [-y \times \log(odds) + \log(1 + e^{\log(odds)})] \text{ türev denkleminde eşittir.}$$

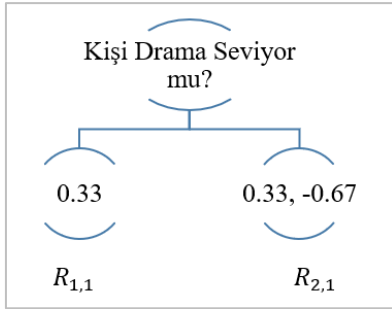
$$\text{Hesaplandığı üzere } r_{im}, \text{ artakalanlar } - \left[-y + \frac{e^{\log(odds)}}{1 + e^{\log(odds)}} \right] \text{ olacaktır.}$$

$$y - \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}} = y - \frac{e^{0.69}}{1 + e^{0.69}} = y - 0.67$$

Çizelge 3.20. Sınıflandırma için artakalanları gösteren tablo

Kişi “Trois Couleurs” Serisini Seviyor mu?	Artakalanlar
Evet	0.33
Evet	0.33
Hayır	-0.67

(b) r_{im} değerleri ile bir regresyon ağacı kurulur ve $j = 1, \dots, J_m$ olacak şekilde R_{jm} terminal bölgesi oluşturulur.



Şekil 3.25. Artakalanlar ile oluşturulan bir regresyon ağacı

$$(c) j = 1, \dots, J_m \text{ için } \gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma) \text{ hesaplanır.} \quad (3.52)$$

$L(y_i, F_{m-1}(x_i) + \gamma)$ ifadesi kayıp fonksiyonudur. Bu nedenle,

$-y \times \log(\text{odds}) + \log(1 + e^{\log(\text{odds})})$ şeklinde yazılabilir. Böylelikle kayıp fonksiyonu;

$-y \times F_{m-1}(x_i) + \gamma + \log(1 + e^{F_{m-1}(x_i) + \gamma})$ olur. Oluşturulan ilk ağacın ilk leaf node'si buradan;

$$\gamma_{1,1} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{1j}} -y \times F_{m-1}(x_i) + \gamma + \log(1 + e^{F_{m-1}(x_i) + \gamma}) \text{ olur.}$$

Regresyon işleminde olduğu gibi bu aşamadan sonra $\gamma_{1,1}$ 'e ait denklemin γ 'ya göre türevi alınmalı ve daha sonra γ için çözülmelidir. Ancak işlem kalabalığı nedeniyle alternatif

çözüm yollarına gidilmiştir. Burada 2. Dereceden Taylor Serisi ile kayıp fonksiyonuna yaklaşım gerçekleştirilebilir;

$$L(y_i, F_{m-1}(x_i) + \gamma) \approx L(y_1, F_{m-1}(x_1)) + \frac{d}{dF_0}(y_1, F_{m-1}(x_1))\gamma + \frac{1}{2} \frac{d^2}{dF_0^2}(y_1, F_{m-1}(x_1))\gamma^2$$

$$\Rightarrow \frac{d}{d\gamma}(y_1, F_{m-1}(x_1) + \gamma) \approx \frac{d}{dF_0}(y_1, F_{m-1}(x_1)) + \frac{d^2}{dF_0^2}(y_1, F_{m-1}(x_1))\gamma = 0$$

Buradan;

$$\gamma = \frac{-\frac{d}{dF_0}(y_1, F_{m-1}(x_1))}{\frac{d^2}{dF_0^2}(y_1, F_{m-1}(x_1))} = \frac{-\text{Kayıp Fonsiyonunun 1. Türevi}}{\text{Kayıp Fonsiyonunun 2. Türevi}}$$

$$= \frac{y - \frac{e^{\log(\text{odds})}}{1 + e^{\log(\text{odds})}}}{\frac{e^{\log(\text{odds})}}{(1 + e^{\log(\text{odds})})^2}} = \frac{\text{Artakalanlar}}{p \times (1-p)}$$

$$\gamma_{1,1} = \frac{0.33}{0.67 \times (1-0.67)} = 1.5$$

$$\gamma_{2,1} = [L(y_2, F_{m-1}(x_2) + \gamma) + L(y_3, F_{m-1}(x_3) + \gamma)]$$

$$\gamma_{2,1} \approx L(y_2, F_{m-1}(x_2)) + L(y_3, F_{m-1}(x_3))$$

$$+ \left[\frac{d}{dF_0}(y_2, F_{m-1}(x_2)) + \frac{d}{dF_0}(y_3, F_{m-1}(x_3)) \right] \gamma$$

$$+ \frac{1}{2} \left[\frac{d^2}{dF_0^2}(y_2, F_{m-1}(x_2)) + \frac{d^2}{dF_0^2}(y_3, F_{m-1}(x_3)) \right] \gamma^2$$

$$\frac{d}{d\gamma}\gamma_{2,1} \approx \left[\frac{d}{dF_0}(y_2, F_{m-1}(x_2)) + \frac{d}{dF_0}(y_3, F_{m-1}(x_3)) \right] + \left[\frac{d^2}{dF_0^2}(y_2, F_{m-1}(x_2)) + \frac{d^2}{dF_0^2}(y_3, F_{m-1}(x_3)) \right] \gamma = 0$$

$$\gamma = \frac{-\left[\frac{d}{dF_0}(y_2, F_{m-1}(x_2)) + \frac{d}{dF_0}(y_3, F_{m-1}(x_3)) \right]}{\left[\frac{d^2}{dF_0^2}(y_2, F_{m-1}(x_2)) + \frac{d^2}{dF_0^2}(y_3, F_{m-1}(x_3)) \right]} = \frac{y_2 + y_3 - p_2 - p_3}{p_2 \times (1-p_2) + p_3 \times (1-p_3)} = \frac{\text{Artakalan}_2 + \text{Artakalan}_3}{p_2 \times (1-p_2) + p_3 \times (1-p_3)}$$

İlk karar ağacı olduğu için $p_1 = p_2 = p_3 = 0.67$ olur ve buradan $\gamma_{2,1} = -0.77$ gelir. Böylelikle ilk ağacın çıktı değerleri hesaplanmış olur.

$$(d) F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{ij}) \text{ güncellenir.} \quad (3.53)$$

$$F_1(x) = 0.69 + 0.8 \times 1.5 = 1.89$$

$$F_1(x) = 0.69 + 0.8 \times (-0.77) = 0.07$$

$$F_1(x) = 0.69 + 0.8 \times (-0.77) = 0.07$$

3. Adım: $m=M$ için, $F_m(x)$ çıktısı güncel olarak elde edilir ve tahmin yeteneği yeni veriler ile test edilir.

4. MATERYAL VE YÖNTEM

4.1. Kuarz Kristal Mikrobalans Sensörleri

Bu çalışmada tasarlanmış olan elektronik burun sisteminde, kuvars kristal mikrobals (QCM) sensörleri başarıyla kullanılmıştır. Bir QCM sensörü, analite duyarlı bir numune (polimer yapılar) ile kaplanmış bir kuvars kristalinden oluşur. Analit, numune ile temas ettiğinde, numunenin yüzey yapısı nedeniyle adsorbe edilir. Bu durum QCM sensörünün kütlelerinin artmasına ve dolayısıyla rezonans frekansında bir değişikliğe neden olur. Belirli bir polimerle kaplanmış bir QCM sensörü, birden fazla analite yanıt verebilmektedir, dolayısıyla QCM tabanlı elektronik burun sistemleri, farklı numunelerin sağlayabileceği bilgilerden yararlanabilmek için bir dizi sensör kullanır. Tasarlanan elektronik burun sisteminde, QCM sensörlerine ait sekiz farklı numune kullanılmış ve çeşitli analitlere karşı verdikleri tepkiler elde edilmiştir. QCM yüzeylerine kaplanan numunelerin isimleri Çizelge 4.1’de verilmiştir.

Çizelge 4.1. QCM sensörlerinde kullanılan polimer yapıları bileşikler

	NUMUNENİN ADI
Sensör 1	Alkoks Fenil (3 Karbonlu) Oksim Bileşiğinin Nikel Kompleksi
Sensör 2	Alkil Zincirli Dinitril Bileşiği
Sensör 3	Uzun Zincirli Dinitril Bileşiği
Sensör 4	Polivinli Prolidon İsimli Polimer
Sensör 5	Kükürtlü Polioksolu Tetra Nikel Ftalosiyanın
Sensör 6	OH’lı Polioksolu Tetra Nikel Ftalosiyanın
Sensör 7	Alkoks Fenil (6 Karbonlu) Oksim Bileşiğinin Nikel Kompleksi
Sensör 8	Alkil Zincirli Aminli Oksim Bileşiğinin Nikel Kompleksi

Analitlere ait konsantrasyon değerleri Antonie Eşitliği ile hesaplanan buhar basınçlarından elde edilmiştir. Bu eşitlik Denklem 4.1’de gösterildiği gibidir. Burada sırasıyla P, analitin buhar basıncı, T, sıcaklığı ve A,B,C, Antonie katsayılarıdır. Sıcaklığa bağlı olarak Antonie Eşitliğine göre analitlerin sahip olduğu katsayılar ve konsantrasyon değerleri Çizelge 4.2’ verilmiştir.

$$\log (P) = A - \frac{B}{(C+T)} \quad (4.1)$$

Çizelge 4.2. Analitlere ait Antonie katsayıları

	A	B	C	T (°C)	Konsantrasyon Aralığı
Aseton	4.42448	1312.253	-32.445	16	6722-24007 ppm
2-Bütanon	3.9894	1150.207	-63.904	16	2499-8927 ppm
Dimetil Sülfid	4.28713	1201.134	-29.906	4	9170-32751 ppm

QCM sensörleri, nanogram düzeninde küçük kütle değişikliklerini ölçmek için kullanılır. Frekans kararlılığı ve kütle çözünürlüğü saptanmış bir sensör, osilatör devresi aracılığı ile titreştirilir ve oluşan doğal frekans TTL sinyali olarak çıktı verir. Bu sinyaller frekans sayıcı devresi ile seri port üzerinden bilgisayara aktarılmaktadır. Bu devreler Şekil 4.1’de verildiği gibi sensör birimini oluşturur.

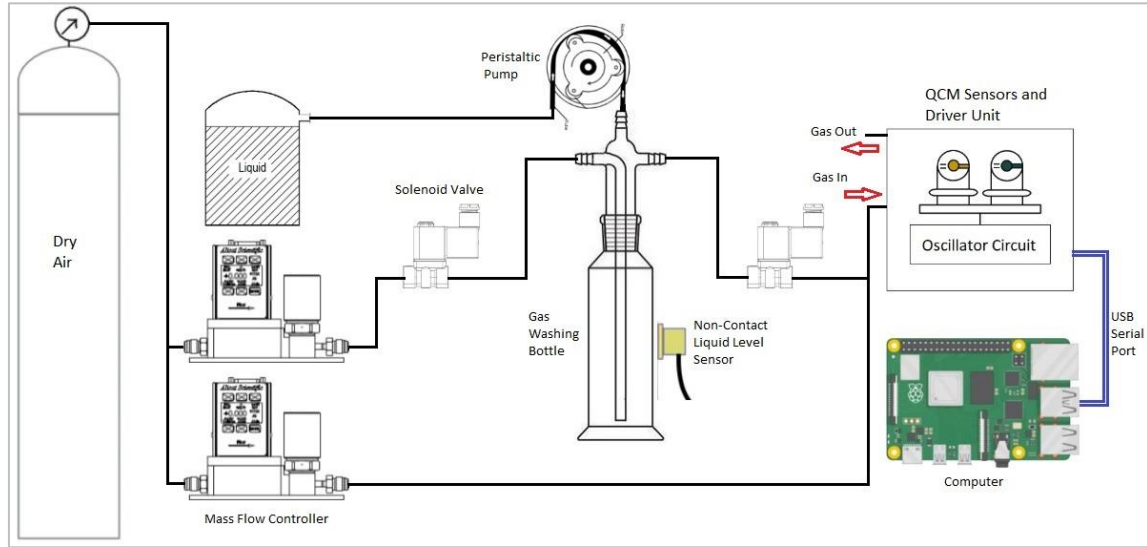


Resim 4.1. QCM sensör birimi

4.2. Veri Toplama İşlemi

Karaciğer hastalıklarının teşhisinde, elektronik burun sistemleri için ayırt ediciliğe sahip aseton, 2-bütanon ve dimetil sülfid kimyasalları oda sıcaklığında sıvı halde bulunmaktadır. Bu kimyasalların gaz halinde doğru miktarlarda sensör birimine iletilmeleri bir dizi konfigürasyonun gerçekleştirilmesini zorunlu kılmıştır. Şekil 4.1’de gösterilen bu konfigürasyona göre kuru hava, gaz akış kontrolcülerini aracılığı ile doğrudan sensör ünitesine ve belirli aralıklarla gaz yıkama şişesi üzerinden sensör ünitesine aktarılmaktadır. Böylelikle sensör içerisinde biriken olası kimyasallar temizlenecek, sensörlerin doğal frekanslarında

titreşimi sağlanacak ve belli aralıklarla istenilen kimyasalın sensörlere gönderilmesi ile frekans değişimi gözlemlenecektir.

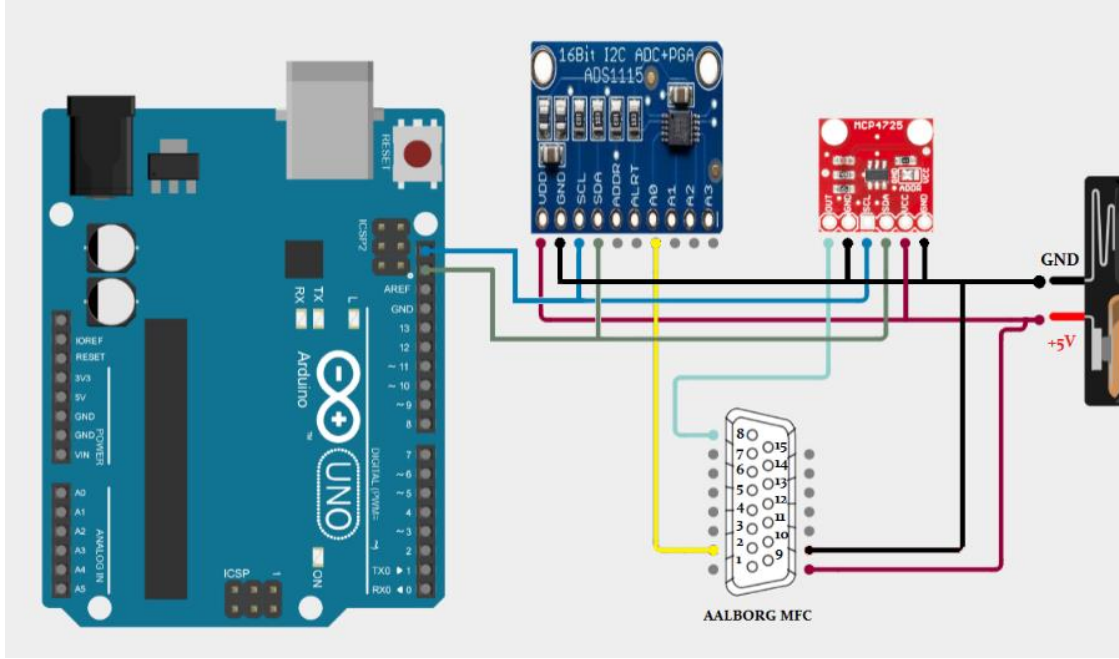


Şekil 4.1. Gaz ölçüm düzeneği

Sensör ünitesine gönderilen kimyasalın konsantrasyonu, gaz yıkama şişesinden geçen havanın akış miktarı, sccm (dakikada standart santimetre küp) ile doğru orantılıdır. Bu akış PID kontrolöre sahip kütle akış kontrolcülerini (mfc) ile sağlanmaktadır. Bu çalışmada bütçe kısıtlamaları nedeniyle iki farklı mfc kullanılmıştır. Şekil 4.1’de gösterilen gaz yıkama şişesi ile bağlantılı mfc1, yüksek hassasiyete ve kısa tepki süresine sahiptir ve kendi terminal programı ile kontrol edilebilmektedir. Öte yandan doğrudan sensör ünitesine bağlı olan mfc2 ise manuel voltaj uygulanarak kontrol edilebilmektedir. Bu nedenle voltaj değerini girdi olarak alabilen ve ADC/DAC kartları ile mfc2’ye bu voltaj değerini atayabilen bir arduino programı geliştirilmiştir (Şekil 4.2). Mfc2’ye ait 8 numaralı pin ayar noktasıdır (Çizelge 4.3) ve buraya verilecek 5 Vdc, 500 ml/dk akış değerine eşittir.

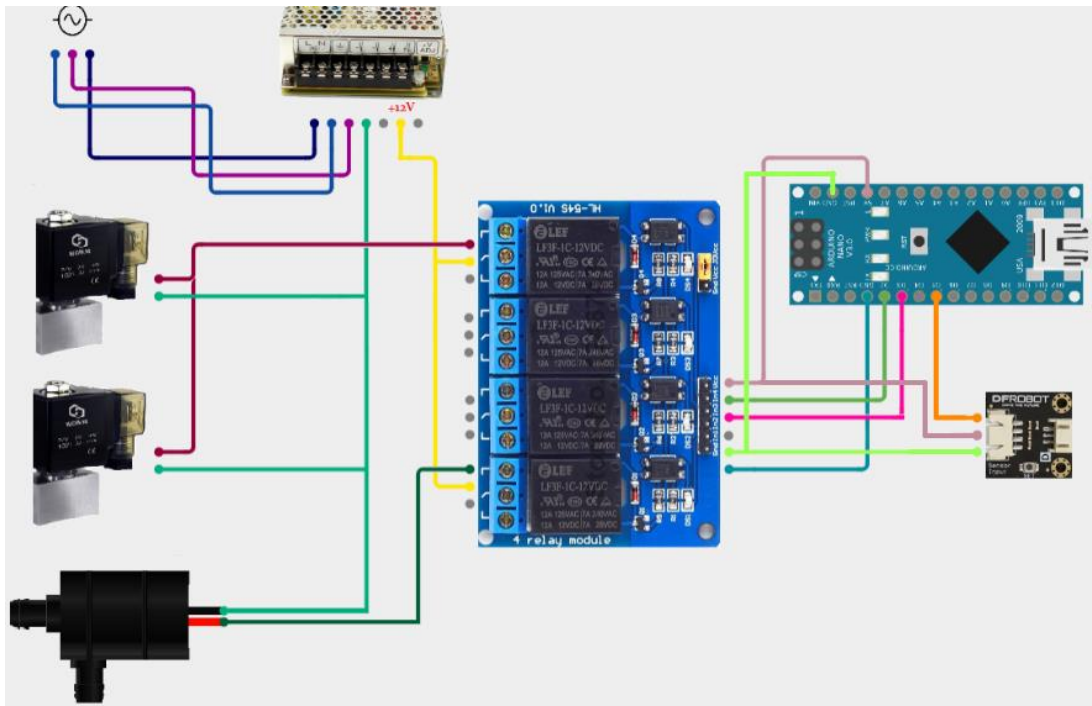
Çizelge 4.3. MFC’ye ait pin bilgileri

1&2	Vdc Çıkış	5&7	+12 Vdc Güç Kaynağı
3&4	Tahliye	8&10	0-5 Vdc Ayar Noktası
3&12	Valf OFF Kontrol	9&14	4-20 mA Çıkış (Pasif Yükler İçin)
5&13	Yardımcı +12 Vdc Güç Çıkışı	10&11	+5 Vdc Kontrol Kaynağı
6	Atanmamış		



Şekil 4.2. MFC kontrol düzeneği

Gaz ölçüm düzeneğinde hava ile gaz yıkama şişesi arasındaki bağlantılar solenoid valfler aracılığı ile gerçekleştirilmektedir. 12 Vdc Güç kaynağına bağlı röle kartıyla beslenen solenoid valfler, arduino nona mikro denetleyici kartı ile 5 dakika kapalı/açık döngüsü altında kontrol edilmektedir.



Şekil 4.3. Sıvı seviyesi kontrol sistemi ve solenoid valf devresi

Gaz yıkama şişesi içerisinde kuru hava ile temas eden analitler, sensör birimine taşınmaktadır. Bu işlem sırasında gaz yıkama şişesi haznesinde bulunan analit miktarı azalır ve kuru hava ile taşınan analit miktarında da benzer şekilde bir azalma gözlenir. Sensör tepkilerini doğrudan etkileyen bu durum karşısında, bir seviye stabilizasyon sistemi geliştirilmiştir. Bu sistemde gaz yıkama şişesine bağlanan bir sıvı seviyesi tespit sensörü, analit miktarının azalmasıyla mikrodnetleyici kartına 0 değerini gönderir ve bu kart peristaltik pompanın devreye girmesi için komut gönderir. Peristaltik pompa, analit seviyesi eski konumuna gelene kadar çalışmaya devam eder. Bu işlemin, sensör ünitesine taşınan analit miktarını etkilememesi için mikrodnetleyici kart, yalnızca solenoid valfler kapalıyken peristaltik pompayı çalıştıracak şekilde programlanmıştır.

5. ARAŞTIRMA BULGULARI

Her bir analit için gaz konsantrasyon değerleri Çizelge 4.2’de verilen ppm değerleri arasında arttırılmış ve sensörlerin tepkileri kaydedilmiştir. Ölçümler esnasında kuru hava akış hızı 500 ml/dk seviyesinde sabit tutulmuştur. Test edilecek analit gaz, 5 dakikakılık periyotlar ile gaz sistemine, kütle akış kontrolcülerini yardımı ile istenen gaz konsantrasyonlarında bilgisayar kontrollü olarak verilmiştir. Hücre içerisinde farklı sensör materyaline sahip 8 adet QCM sensör kullanılmıştır. Aseton ve 2-bütanon analitlerinin zararsız olması, kolay ulaşılabilir olması ve rahatlıkla ölçülebilir olması nedeniyle ölçümler tekrarlanmış ve sonuçların tutarlı olduğu gözlemlenmiştir. Ancak dimetil sülfid analitinin, düşük buharlaşma sıcaklığı, teneffüsü ile insan metabolizmasında yaratacağı olası problemler gibi zorlukları nedeniyle tek seferde ölçülmüş ve sonuçları işleme alınmıştır.

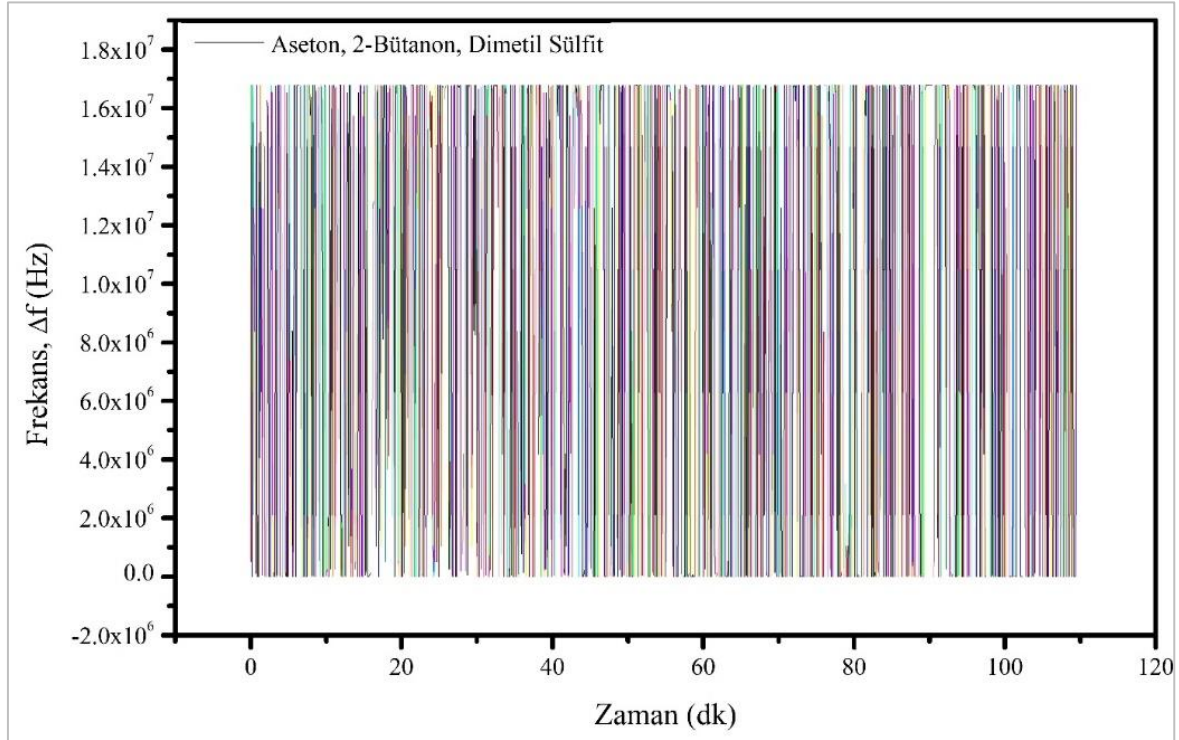
Ölçümler, sensör tepkilerinin makine öğrenmesi algoritmaları tarafından matematiksel olarak modellenebilmesi için ve bu modellerin performanslarının test edilebilmesi için çeşitli konsantrasyonlarda yapılmıştır. Makine öğrenmesi algoritmalarının başarıları eğitim için kullanılacak verilerin özellikleri ile doğrudan ilişkilidir. Bu nedenle eğitim verileri, düzenli olarak alınmış, çeşitli konsantrasyonları bir arada bulunduran ölçümlerden oluşmaktadır. Ardından bu modellerin performanslarının test edilebilmesi için her bir analit ile belirli gaz konsantrasyonunda ölçümler yapılmıştır. Eğitim seti için alınan ölçümler sırası ile Şekil 5.1’den 5.8’e kadar paylaşılmıştır.

5.1. Sensör Tepkileri

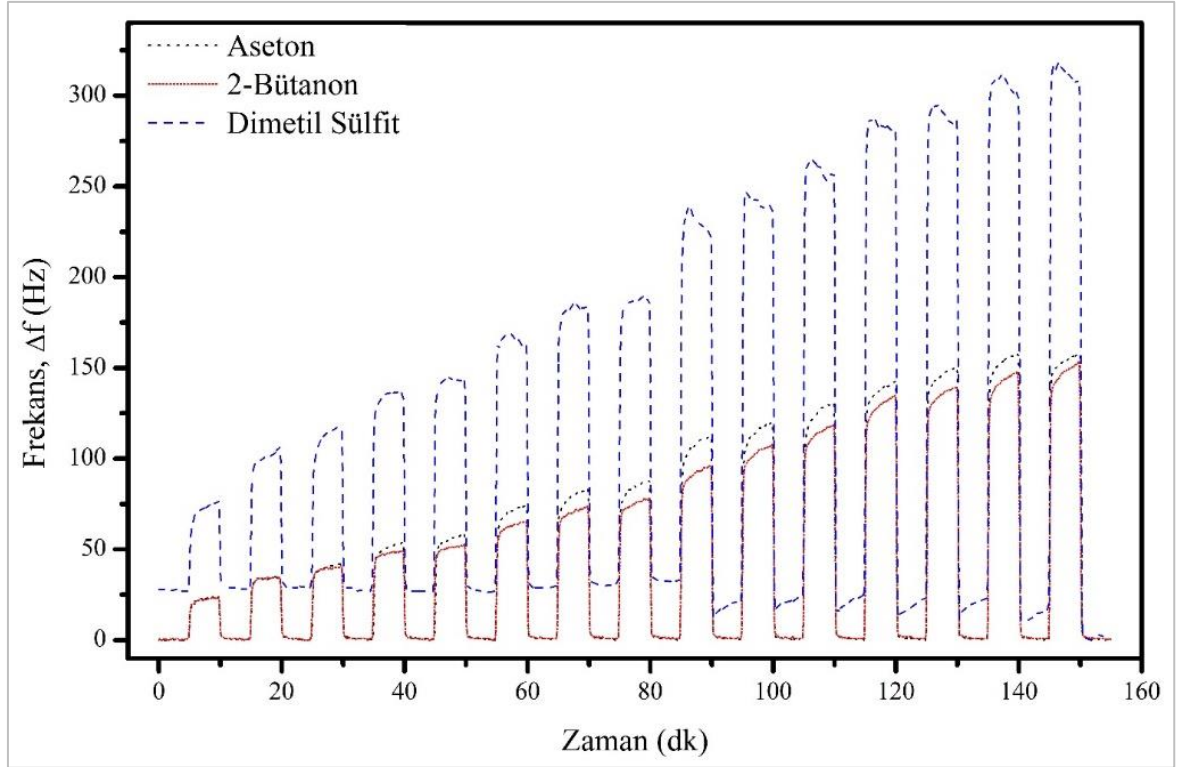
Ölçümler sonrasında Sensor 1’in (Şekil 5.1), analitlerin ve konsantrasyon değerlerinin tamamı için gürültü olarak isimlendirilen anlamsız tepkiler verdiği anlaşılmıştır. Sensör 2 (Şekil 5.2), aseton, 2-bütanon ve dimetil sülfid analitleri için kararlı ve cevap süresinin kısa olduğu gözlemlenmektedir. Ancak aseton ve 2- bütanon değerleri birbirine oldukça yakındır ve görece düşüktür. Bu nedenle, bu sensörün dimetil sülfid için daha yüksek bir ayırt ediciliğe sahip olduğu söylenebilir. Bununla beraber dimetil sülfid analitinin temel seviyenin uzağında kaldığı gözlemlenmektedir. Sensör yüzeyinin, analitten arındırılma süresinin yetersiz olabileceği bu durumun nedeni olarak düşünülmektedir. Sensör 3 (Şekil 5.3), tüm analitler için anlamlı tepki vermiştir, kararlıdır ve cevap süresi kısadır. Ancak sonuçlarının birbirine oldukça yakın olması bu sensörün ayırt edici bir özelliğe sahip bir sensör olarak kullanımını

zorlaştırmıştır. Sensör 4'ün verdiği tepki (Şekil 5.4), test ölçümleri göz önüne alındığında, sensör ünitesinden kaynaklı olarak farklılık göstermiştir. Bu nedenle dimetil sülfid analiti için ideal bir ölçüm alınamamıştır. Ancak aseton ve 2-bütanon analitleri anlamlı tepkiler vermiştir. Sensör 5'te (Şekil 5.5), aseton ve dimetil sülfid analitlerinin benzer tepkiler verdiği, ancak üç analit için de anlamlı tepkilerin elde edilebildiği gözlemlenmiştir. Sensör 6 (Şekil 5.6), dimetil sülfid analitinin yüksek konsantrasyonları için kararlılığını yitirmeye başlamıştır, aseton ve 2-bütanon analitleri için ayırt ediciliği bulunmaktadır. Sensör 7 (Şekil 5.7), tüm analitler için anlamlı tepkiler vermiştir. Ancak dimetil sülfid analiti için Sensör 2'de olduğu gibi temel seviyenin uzağında kalmıştır. Sensör 8 (Şekil 5.8) ise, aseton ve dimetil sülfid analitleri için temel seviyenin uzağında kalmış ve yüksek konsantrasyonlarında kararlılığını yitirmiştir. Yalnızca 2-bütanon analiti için anlamlı tepki vermiştir.

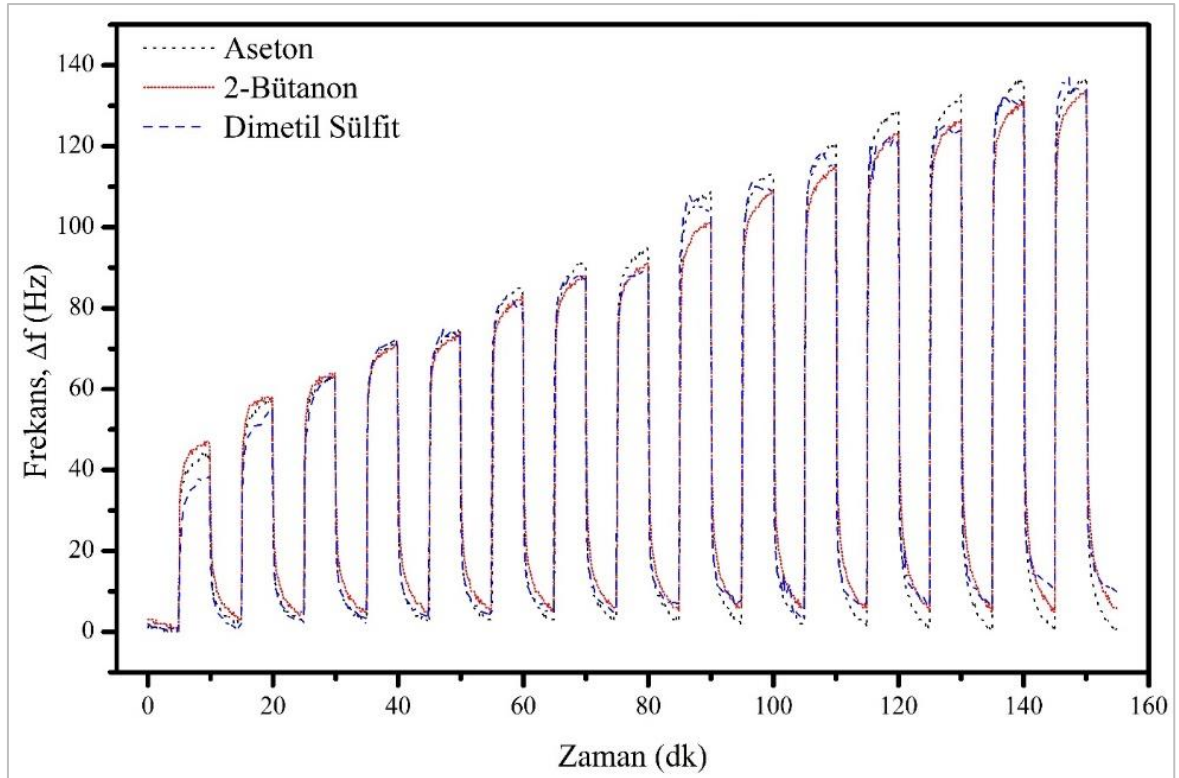
Analitlerin yüksek konsantrasyonlarında, Sensör 1 hariç tüm sensörler anlamlı tepkiler vermiştir. Oksijen, karbon di-oksit, karbon mono-oksit, azot, argon gibi inorganik uçuculara tepki vermemeleri bu sensörlerin en büyük avantajlarıdır. Sensör yüzeyinde farklı numunelerin tercihi ve yüksek hassasiyetli osilatörlerin kullanılması, ppb düzeyinde tespit yapılabilmesi için gereklidir.



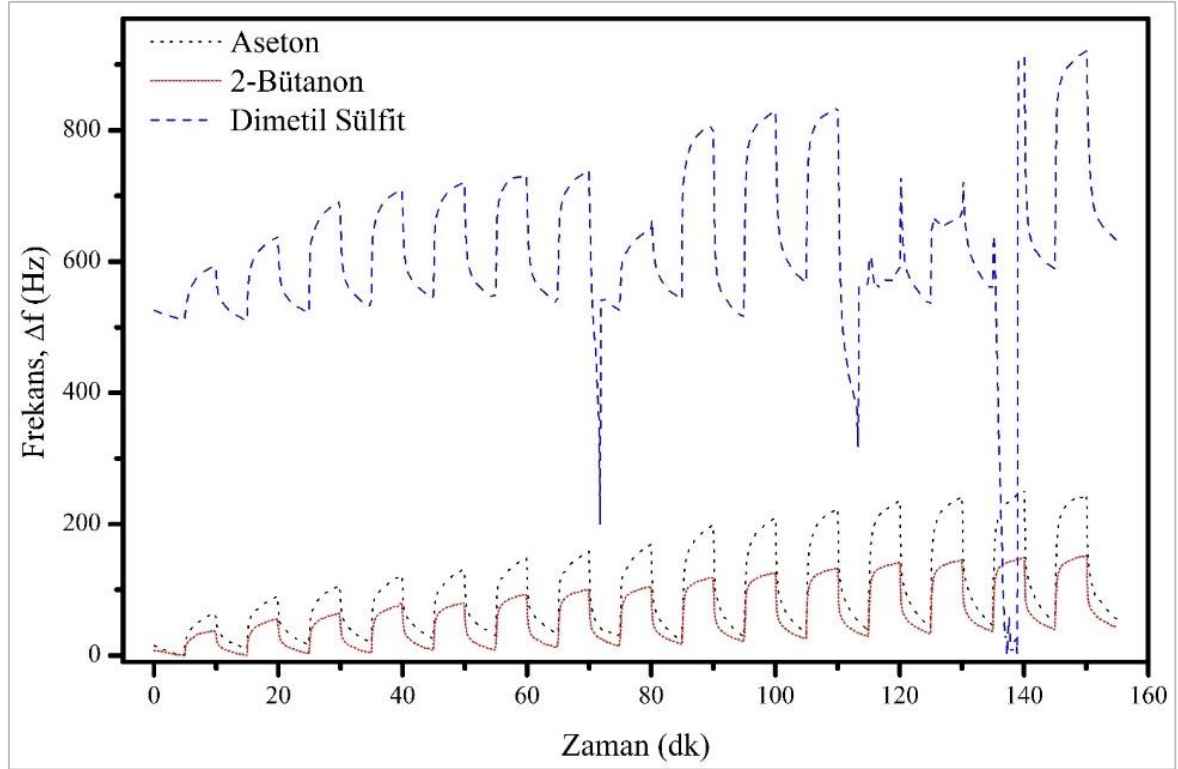
Şekil 5.1. Sensör 1'in analitlere karşı verdiği tepki



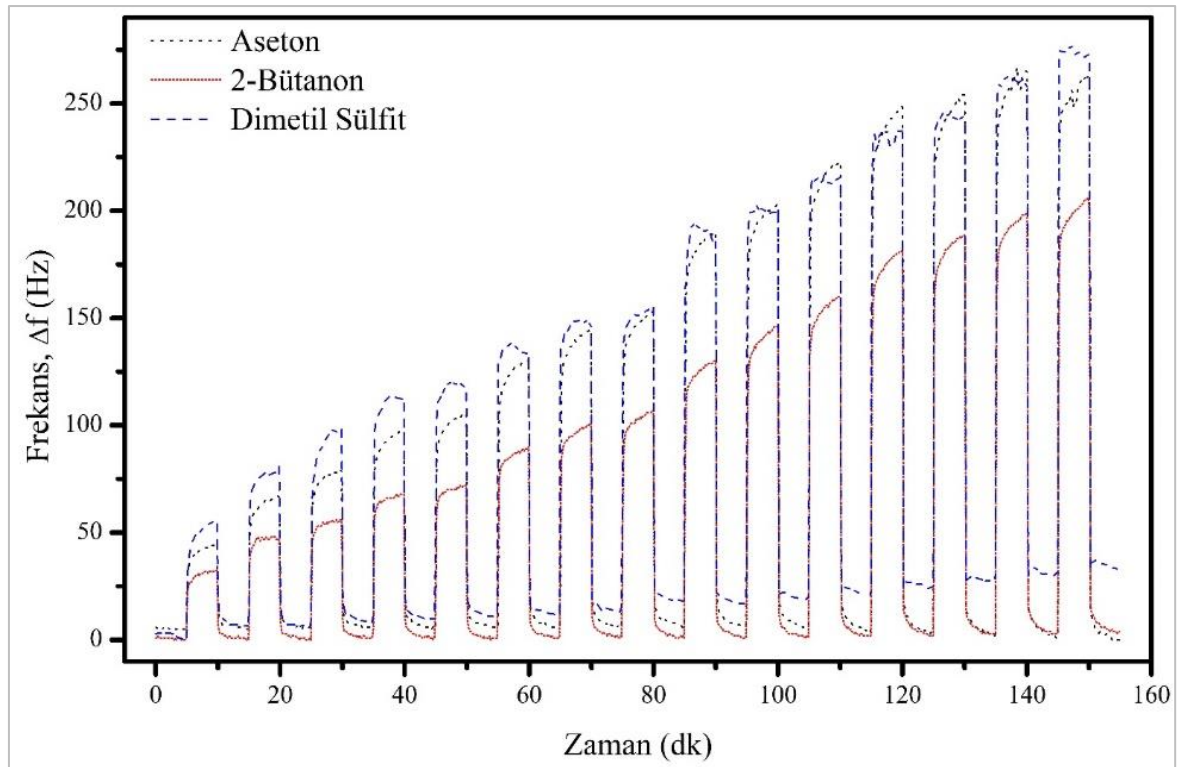
Şekil 5.2. Sensör 2'nin analitlere karşı verdiği tepki



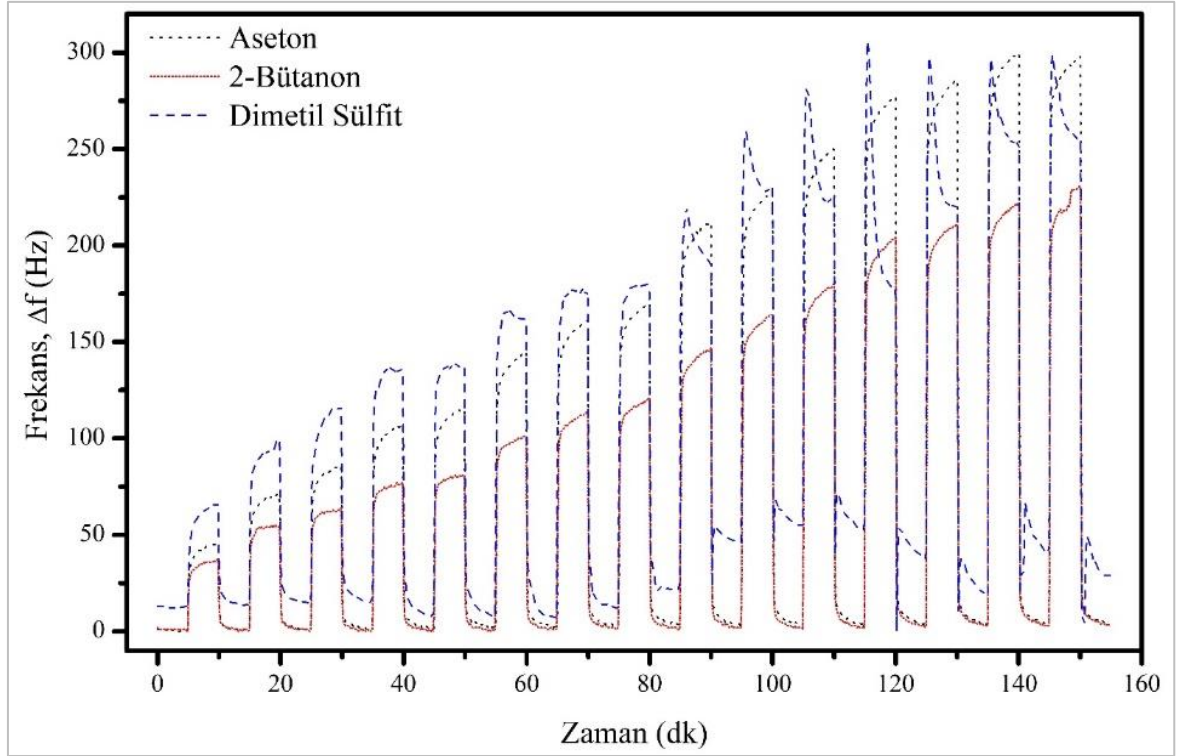
Şekil 5.3. Sensör 3'ün analitlere karşı verdiği tepki



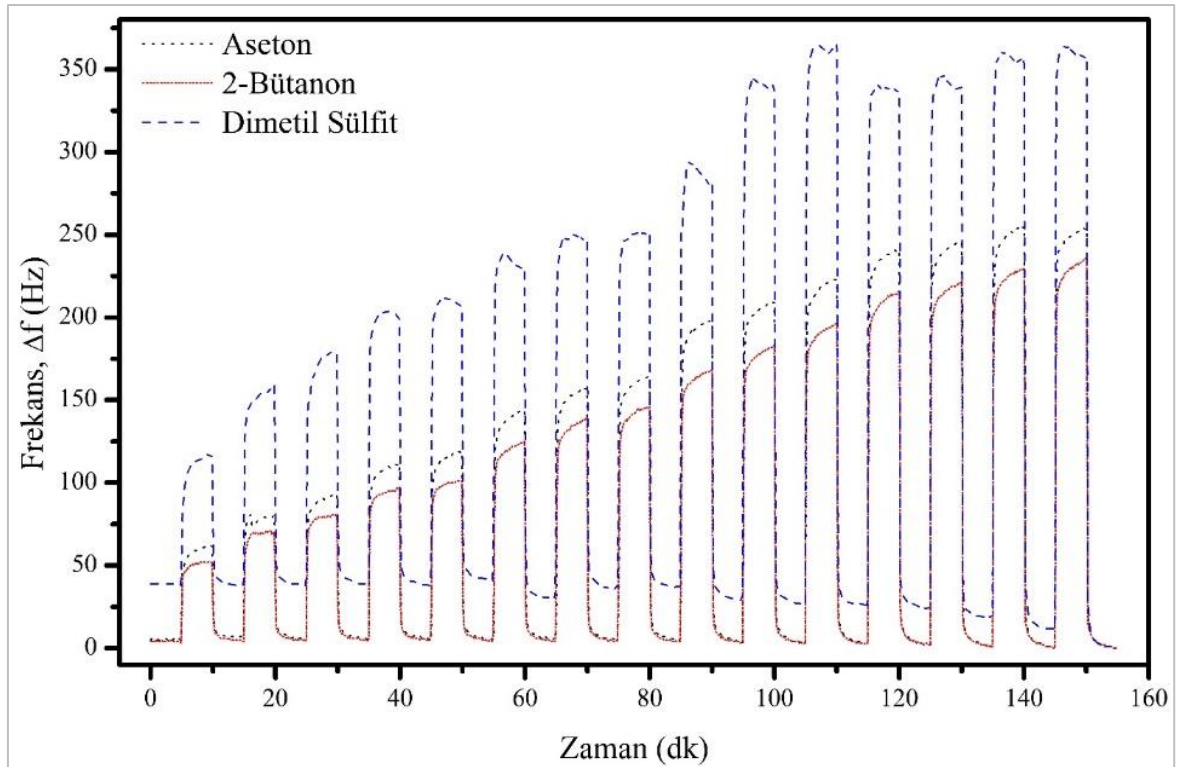
Şekil 5.4. Sensör 4'ün analitlere karşı verdiği tepki



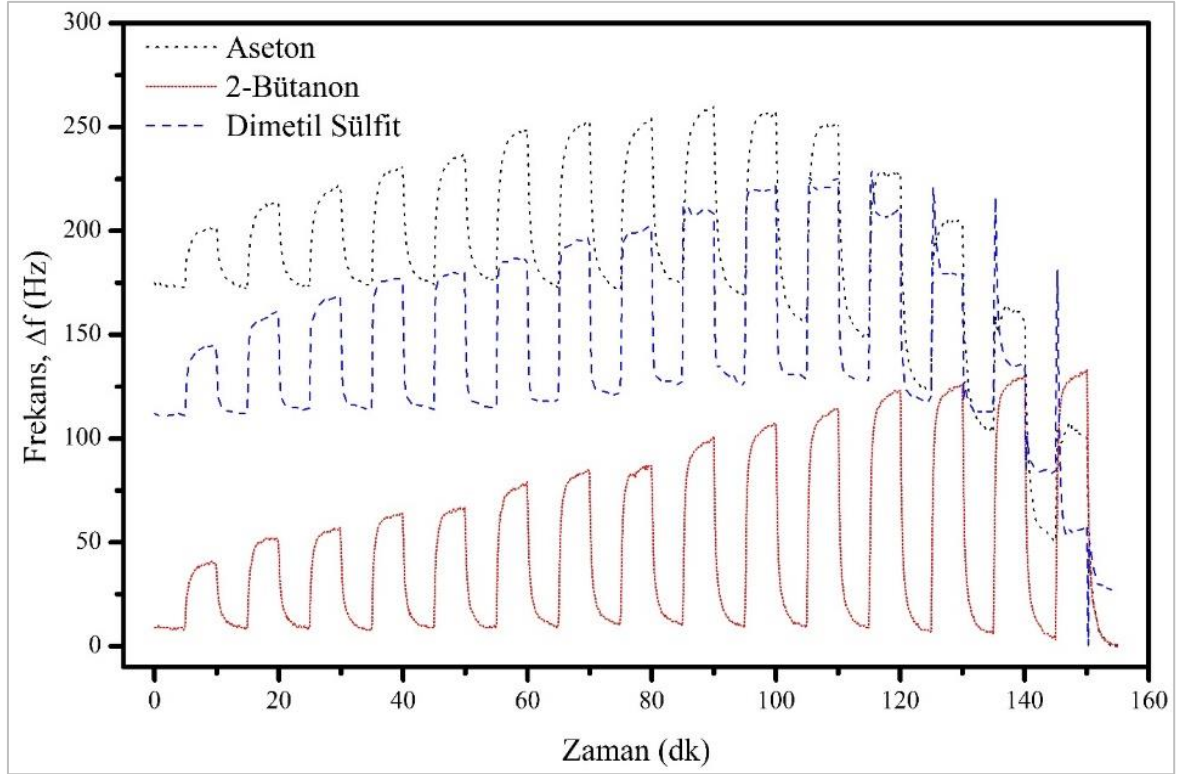
Şekil 5.5. Sensör 5'in analitlere karşı verdiği tepki



Şekil 5.6. Sensör 6'nın analitlere karşı verdiği tepki



Şekil 5.7. Sensör 7'nin analitlere karşı verdiği tepki



Şekil 5.8. Sensör 8’in analitlere karşı verdiği tepki

5.2. Makine Öğrenmesi Algoritmalarının Sonuçları

Kuvarz kristal mikrobalsans sensörleri ile her bir analit için elde edilen frekans değerleri, uygun örüntü tanıma algoritmaları ile sınıflandırma ve regresyon işlemleri yapılmasına olanak sağlar. Bu çalışmada geleneksel sınıflandırma yöntemlerinden K- En Yakın Komşu, Lojistik Regresyon, Destek Vektör Makineleri, Gaussian Naive Bayes, Sinir Ağları, Karar Ağaçları, Rastgele Orman, Gradient Boosting Machine, Extreme Gradient Boosting ve genellikle özellik çıkarımı (feature extraction) ve boyut küçültme gibi operasyonlar için tercih edilen, Doğrusal Diskriminant Analizi algoritmaları test edilmiş ve yüksek performans gösteren algoritmalar ile bir topluluk modeli (ensemble model) geliştirilmiştir.

Veri seti olarak kullanılacak bazı sensör tepkileri üzerinde tespit edilen anlık anormallikler ve aykırı değerler temizlenmiş, eksik veriler takip edilen verilerin bir ortalaması alınarak tamamlanmış ve temel yapı ortaya çıkarılmıştır. Bununla beraber veri seti üzerinde bulunan minimal dengesizlik, “Undersampling” yöntemlerinden biri olarak kullanılan “NearMiss” yaklaşımı ile çözümlenmiştir. Analitlerin bilgisini içeren hedef vektör matrisi, “Label Encoding” yöntemi ile kodlanmış ve makine dili için uygun hâle getirilmiştir.

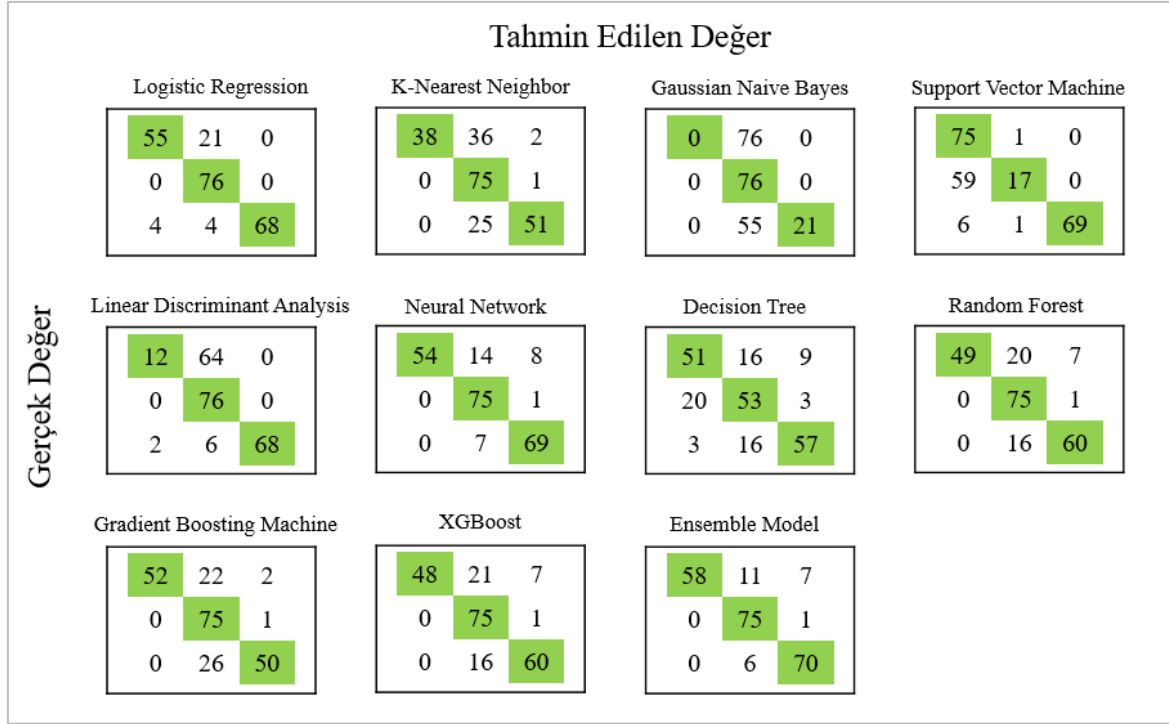
Her bir analit için 15 konsantrasyon değeri kullanılarak oluşturulan eğitim seti ile tek bir konsantrasyon değerinin bilgisini içeren test setleri, veri sızıntılarını önleyip gerçekçi bir sınıflandırma sonucu elde etme amacıyla bağımsız zamanlarda ölçülmüştür.

Literatürde, Temel Bileşenler Analizi (Principal Component Analysis, PCA), Lokal Doğrusal Gömüleme (Locally Linear Embedded), t-Dağıtılmış Stokastik Komşu Gömme (t-Distributed Stochastic Neighbor Embedding, t-SNE) ve Otokodlayıcı (Auto-Encoder) gibi boyut küçültme, veriyi farklı yaklaşımlarla temsil etme ve yüksek varyans ile tutma gibi görevleri üstlenerek özellik çıkarımını sağlayan ve denetimli sınıflandırıcıların daha yüksek doğruluk oranını yakalamasına katkıda bulunan denetimsiz makine öğrenmesi algoritmaları bu çalışmada test edilmiş, ancak veri setinin yapısı nedeniyle kullanımı uygun bulunmamıştır.

Sensör tepkileri incelendiğinde fark edilen, sensörlerin farklı analitlere karşı benzer tepkiler göstermesi denetimsiz makine öğrenmesi algoritmaları için problem teşkil etmiştir. Örneğin, Rastgele Orman Algoritması ile birlikte kullanılan PCA, 0.27 doğruluk oranı yakalarken, XGBoost ile kullanılan t-SNE, 0.36 doğruluk oranı yakalamıştır. Bu oranlar sınıflandırma görevi için oldukça düşüktür. Bu nedenle veri seti doğrudan denetimli makine öğrenmesi algoritmalarıyla modellenmiş ve geliştirilmiştir.

Rastgele alınan üç farklı konsantrasyon değeri ile test için bir girdi matrisi oluşturulmuş, bu matris her analit için 76×5 boyutlarına sabitlenmiştir. Satır sayısı, sensörlere ait frekans değerlerini, sütun sayısı ise işleme dahil edilen sensör sayısını ifade etmektedir. Girdi matrisi ile yapılan tahmin, 76×1 boyutunda, analitlere ait bilgileri barındıran bir çıktı matrisi verecektir.

Doğru ve yanlış tahminlerin sayısını, sayı değerleri ile özetleyerek her sınıf için ayıran karışıklık matrisleri, her bir algoritma için Şekil 5.9'da olduğu gibi verilmiştir. Karışık Matrislerinde satırlar gerçek değerleri, sütunlar ise tahmin edilen değerleri ifade etmektedir. Yeşil olarak ifade edilmiş sayı değerleri ise doğru tahmin edilen değerlerin sayısıdır. Rastgele Orman (Random Forest), Destek Vektör Makineleri (Support Vector Machine), Lojistik Regresyon ve XGBoost ile kurulan Topluluk Modeli'ne (Ensemble Model) ait karışıklık matrisinde, 58, aseton, 75, 2-bütanon ve 70, dimetil sülfid değeri doğru tahmin edilmiştir.



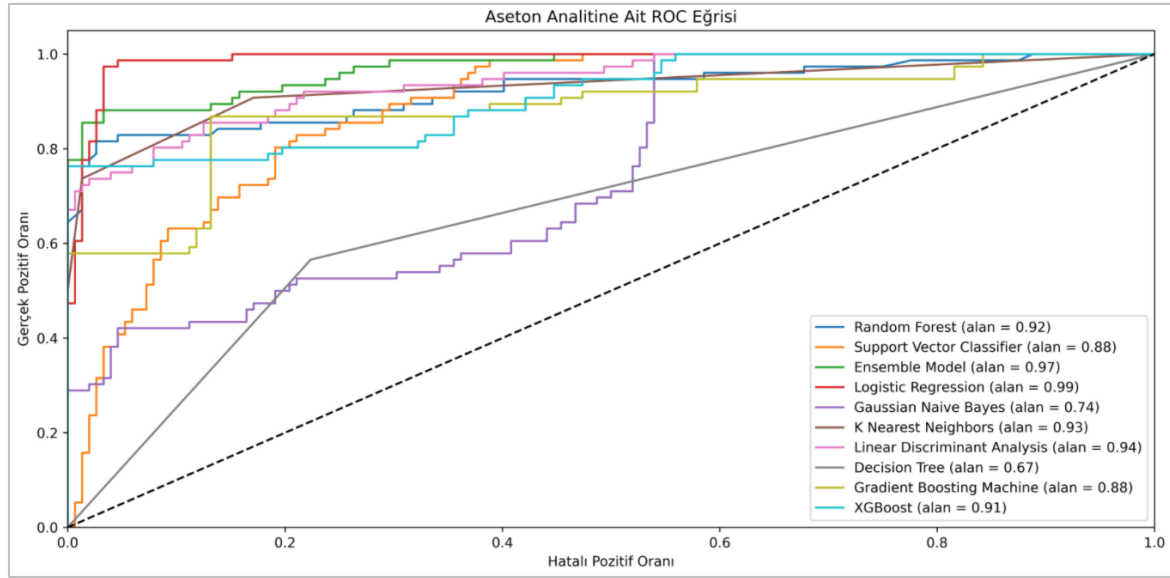
Şekil 5.9. Karışıklık matrisleri

Doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru metrikleri model performansları ile ilgili bilgi vermektedir. Doğru tahmin edilen değerlerin toplamının, toplam veri kümesine oranı şeklinde verilen doğruluk (accuracy), kurulan topluluk modeli için 0.89 olarak hesaplanmıştır. Ancak bu metrik veri bilimi projeleri için tek başına yeterli bir bilgi verememektedir. Bu nedenle diğer metrikler göz önünde bulundurulmalıdır. Bu noktada kesinlik, bize doğru olarak tahmin edilen değerlerin gerçekte ne kadarının doğru olduğunu ifade etmektedir. Bu metrik her bir analit için değişkenlik göstermektedir ve topluluk modeli için sırasıyla 1.00 (aseton), 0.82 (2-bütanon), 0.90 (dimetil sülfid) oranları olarak hesaplanmıştır. Bir diğer metrik olan duyarlılık ise doğru olarak tahmin etmemiz gereken işlemlerin ne kadarını doğru olarak tahmin ettiğimizi göstermektedir ve bu değerler topluluk modeli için sırasıyla 0.76 (aseton), 0.99 (2-bütanon), 0.92 (dimetil sülfid) oranları olarak hesaplanmıştır. İdeal olarak bir modelden yüksek kesinlik ve duyarlılığa sahip olması beklenmektedir. Bu metrikleri tek bir sayı değeri olarak ifade eden, kesinlik ve duyarlılık değerlerinin harmonik ortalaması şeklinde olan F1 skoruna göre topluluk modelinin performansı her bir analit için sırasıyla 0.87, 0.89, 0.90 olarak hesaplanmıştır.

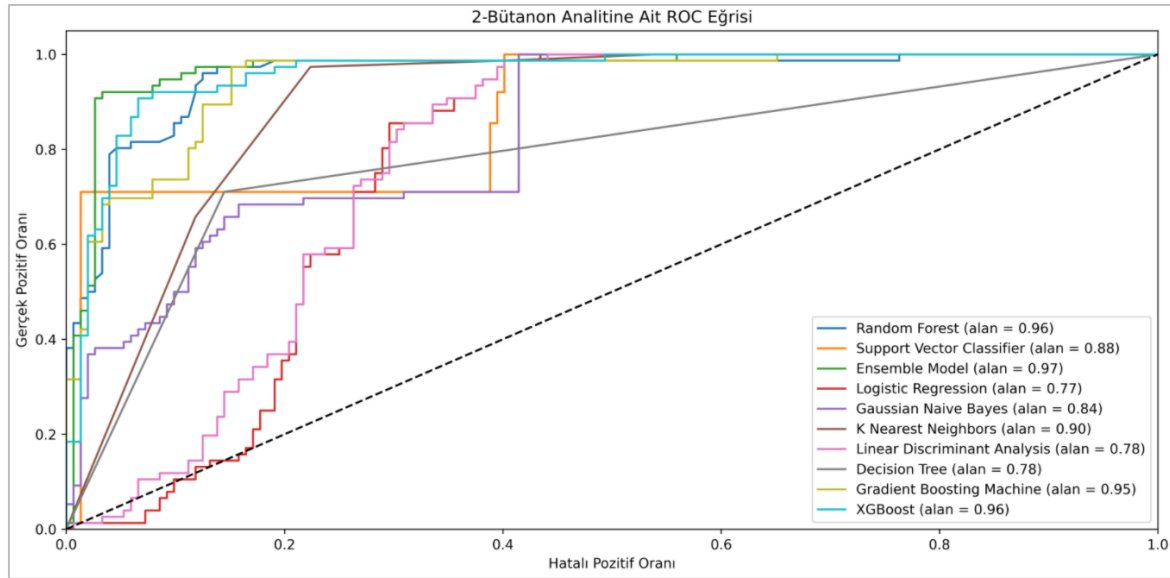
Çizelge 5.1. Algoritmaların doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru değerleri

Model	Analit	Accuracy	Precision	Recall	F1 Skoru
Logistic Regression	Aseton	0.87	0.93	0.72	0.81
	2-Bütanon		0.75	1.00	0.86
	Dimetil Sülfid		1.00	0.89	0.94
K-Nearest Neighbor	Aseton	0.72	1.00	0.50	0.67
	2-Bütanon		0.55	0.99	0.71
	Dimetil Sülfid		0.94	0.67	0.78
Gaussian Naive Bayes	Aseton	0.43	0.00	0.00	0.00
	2-Bütanon		0.37	1.00	0.54
	Dimetil Sülfid		1.00	0.28	0.43
Support Vector Machine	Aseton	0.71	0.54	0.99	0.69
	2-Bütanon		0.89	0.22	0.36
	Dimetil Sülfid		1.00	0.91	0.95
Linear Discriminant Analysis	Aseton	0.68	0.86	0.16	0.27
	2-Bütanon		0.52	1.00	0.68
	Dimetil Sülfid		1.00	0.89	0.94
Neural Network	Aseton	0.33	0.00	0.00	0.00
	2-Bütanon		0.33	1.00	0.50
	Dimetil Sülfid		0.00	0.00	0.00
Decision Tree	Aseton	0.71	0.69	0.67	0.68
	2-Bütanon		0.62	0.70	0.66
	Dimetil Sülfid		0.83	0.75	0.79
Random Forest	Aseton	0.81	1.00	0.64	0.78
	2-Bütanon		0.68	0.99	0.80
	Dimetil Sülfid		0.88	0.79	0.83
Gradient Boosting Machine	Aseton	0.78	1.00	0.68	0.81
	2-Bütanon		0.61	0.99	0.75
	Dimetil Sülfid		0.94	0.66	0.78
XGBoost	Aseton	0.80	1.00	0.63	0.77
	2-Bütanon		0.67	0.99	0.80
	Dimetil Sülfid		0.88	0.79	0.83
Ensemble Model	Aseton	0.89	1.00	0.76	0.87
	2-Bütanon		0.82	0.99	0.89
	Dimetil Sülfid		0.90	0.92	0.91

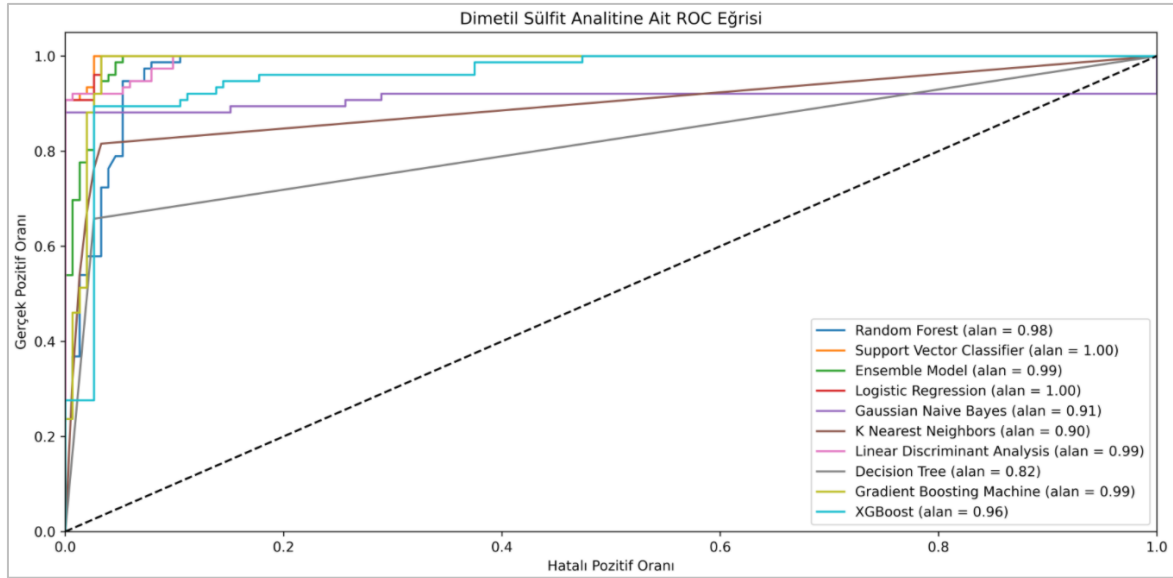
ROC eğrisi, her noktada belirli bir karar eşiğine karşılık gelen bir duyarlılık/özgüllük çiftini temsil eden bir başka metriktir. ROC eğrisi altında kalan alan (AUC) modelin performansı ile ilgili bilgi vermektedir. ROC, tahmin edilen sınıfları kullanan F1 skorundan farklı olarak, tahmin edilen skorları girdi olarak almaktadır. Genellikle ağırlıklı olarak dengesiz olmayan veri setleri için ROC/AUC kullanılmaktadır. AUC sonuçlarına göre topluluk modeli, her bir analit için 0.97, 0.97 ve 0.99 değerlerini vermiştir.



Şekil 5.10. Aseton analitine ait ROC (reciever operator characteristics curve) grafiği

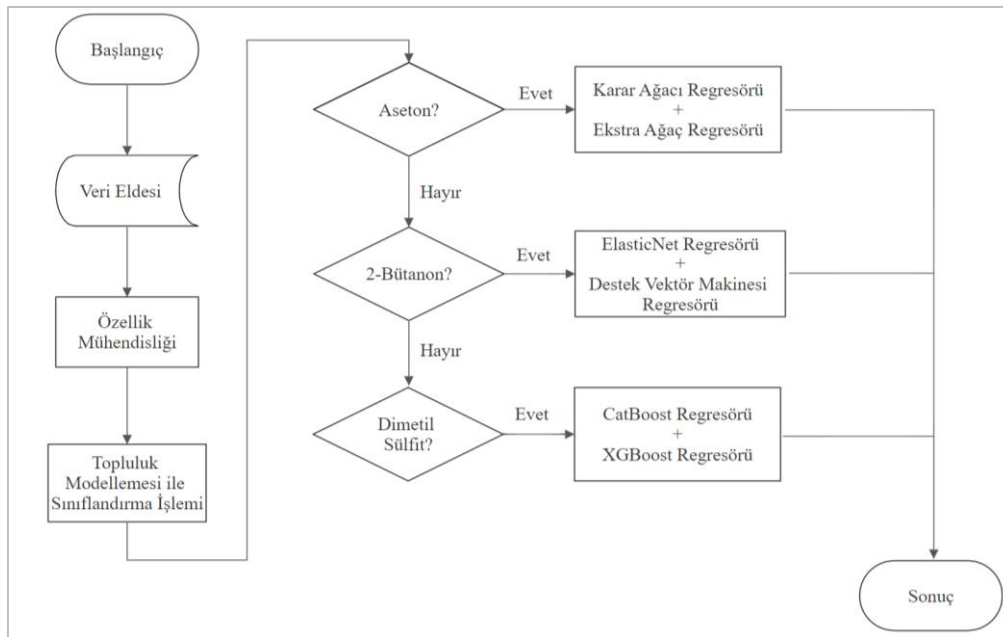


Şekil 5.11. 2-bütanon analitine ait ROC (reciever operator characteristics curve) grafiği



Şekil 5.12. Dimetil sülfid analiti için ROC (receiver operator characteristics curve) grafiği

Sınıflandırmayı takip eden regresyon işlemi, scikit-learn istatistiksel paketlerini kullanan XGBoost, CatBoost, Çoklu Doğrusal Regresyon, Karar Ağacı, Ekstra Ağaç ve Destek Vektör Makineleri, algoritmaları ile gerçekleştirilmiştir. GridSearchCV ile optimize edilen ve bir topluluk modeli altında birleştirilen bu algoritmalar her bir analitin sensör tepkileri için farklı performanslar sergilemiştir. Bu nedenle sınıflandırma ile elde edilen sonuca göre regresyon işlemleri Şekil 5.13’de gösterilen akış şemasına göre gerçekleştirilmiştir.

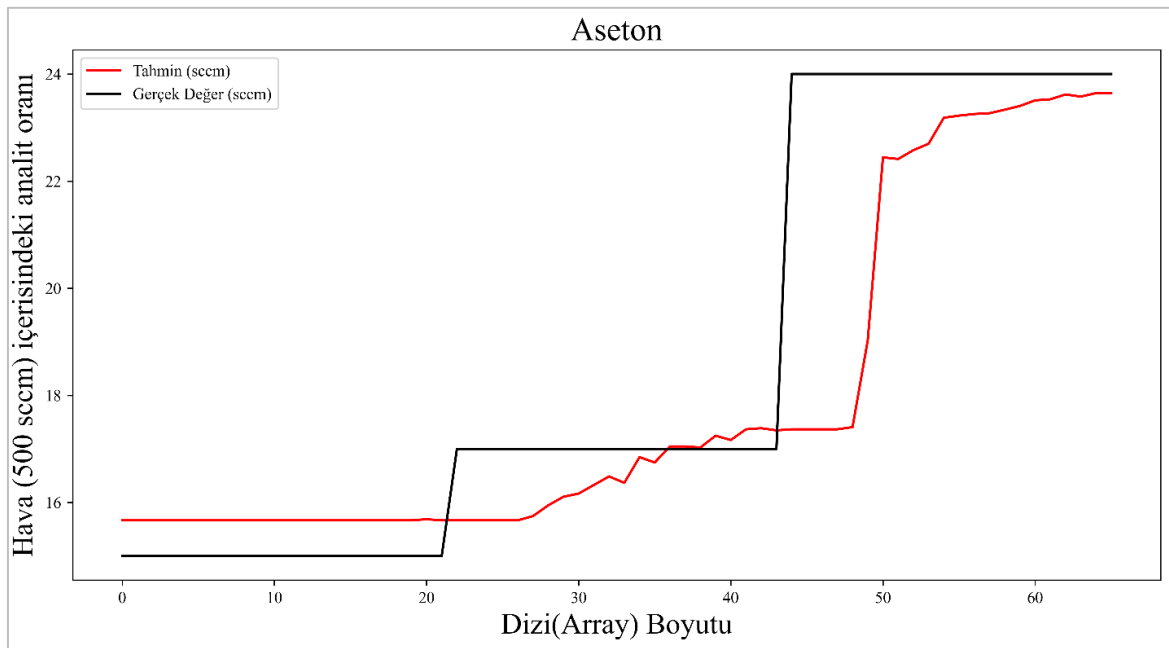


Şekil 5.13. Veri analizi ve makine öğrenmesi süresince atılan adımları ifade eden akış şeması

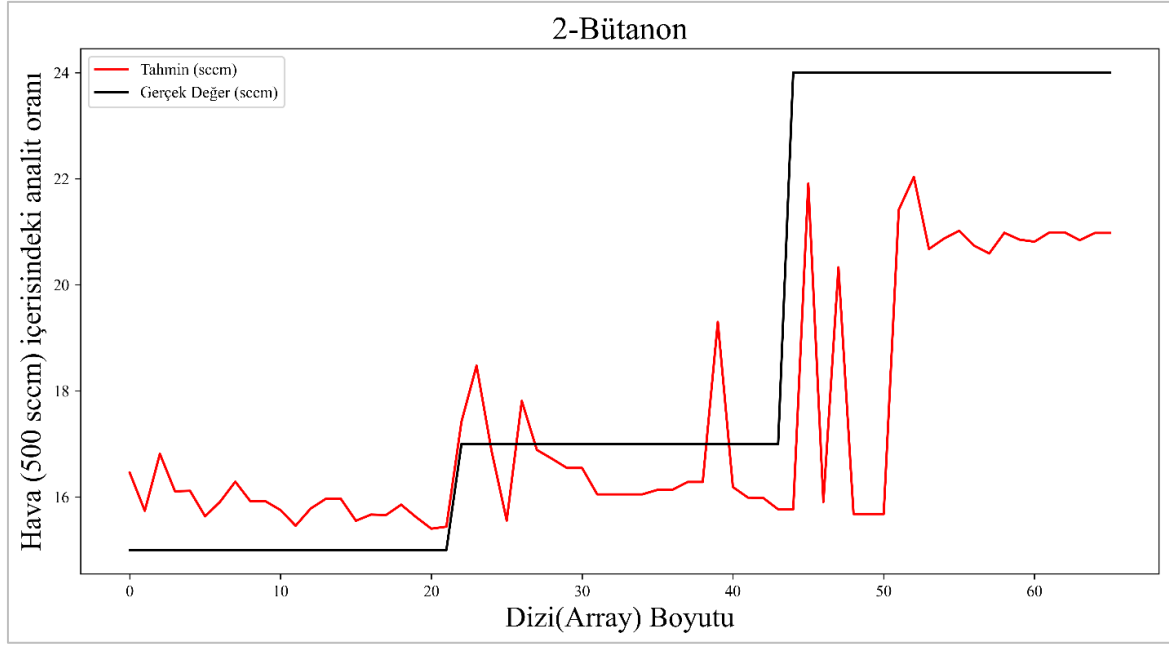
Her bir analitin rastgele seçilen üç farklı konsantrasyon değeri oluşturulan topluluk modelleri ile tahmin edilmeye çalışılmış ve elde edilen tahmin değerleri ile gerçek değerler Şekil 5.14-5.16’da gösterilmiştir.

$$\text{Ortalama karekök sapmaları} = \sqrt{\sum_{i=1}^N \frac{(\hat{y}_i - y_i)^2}{N}} \quad (\text{burada } N \text{ örnek sayısıdır}); \quad (5.1)$$

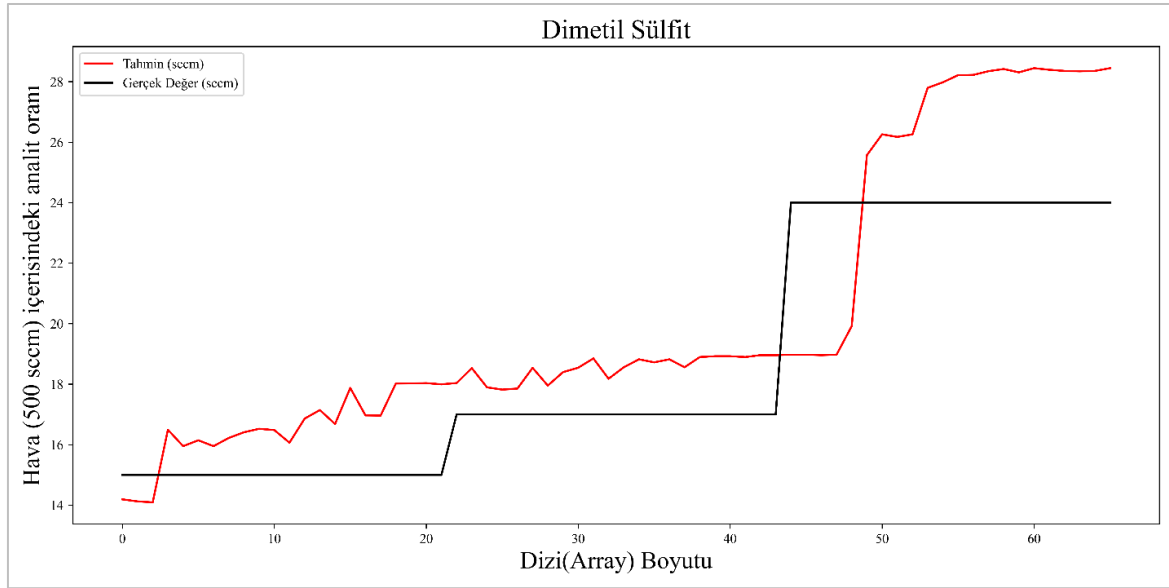
sırasıyla 2.396874, 2.853136 ve 2.775164 olarak bulunmuştur.



Şekil 5.14. Aseton anlitine ait regresyon işlemi

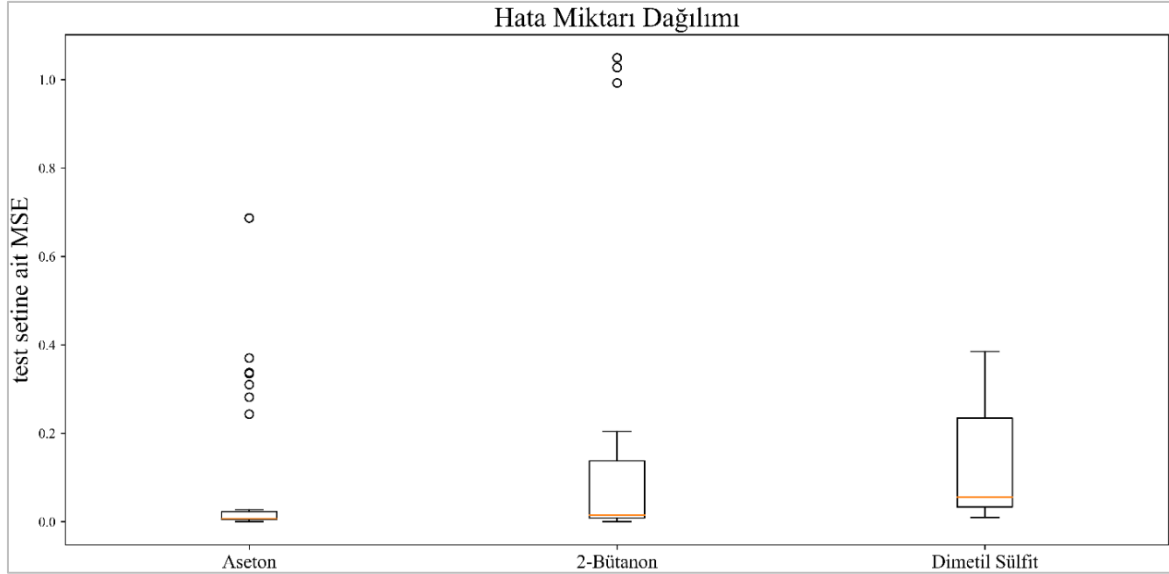


Şekil 5.15. 2-bütanon anlitine ait regresyon işlemi



Şekil 5.16. Dimetil sülfid anlitine ait regresyon işlemi

Şekil 5.17, test setindeki tüm hedef analitin, kendi topluluk modelleri tarafından elde edilen tahmin değerlerinin ortalama hatalar karesi dağılımının kutu grafiğidir. Hata dağılımının dimetil sülfid > 2-bütanon > aseton olduğu gözlemlenmektedir. Ortalama karekök sapmaları ile karşılaştırıldığında, analitlere ait hata dağılımının farklı çıkması, aseton ve 2-bütanon analitlerinin aykırı değerlere sahip olmasıyla açıklanmaktadır.



Şekil 5.17. Regresyon işlemlerine tabi tutulan analitlerine ait MSE hata oranlarının kutu grafiği ile gösterimi

Bu çalışmada veri bilimi esaslarının ve makine öğrenmesi algoritmalarının kullanımı öncelikli olarak hedeflenmiş, tercih edilecek sensörlerin hazır olarak alınması kararlaştırılmıştır. Farklı numuneler kaplanarak üretilen bu sensörler, hedeflenen hassasiyet değerlerinin uzağında kalmış, analitlerin ppb düzeyindeki konsantrasyonlarının eldesi mümkün olmamıştır. Böylelikle bu çalışma yüksek konsantrasyonlarda alınan ölçümler üzerinden devam etmiştir. Tercih edilen algoritmalar ile, her bir analit için yüksek hassasiyete ve ayırt ediciliğe sahip sensörlerin geliştirilmesi, karaciğer hastalıklarıyla ilişkili gazların sınıflandırılması ve regresyonu işlemlerini farklı bir noktaya taşıyacak ve hastalıkların teşhisini mümkün kılacaktır.

6. SONUÇ

Bu tezin temel amacı, elektronik bir burnun, uçucu organik bileşikleri (VOC) kullanarak alkole bağlı olmayan yağlı karaciğer hastalığının ve dünya çapında önde gelen mortalite ve morbidite nedenlerinden biri olan karaciğer sirozunun erken evre teşhisi için bir tanı aracı olarak geliştirilmesidir. Elektronik burun teknolojisi, metabolik atıklarından kaynaklanan gaz fazı biyobelirteçlerini analiz etmek için kullanılabilir, bu da onu, test başına düşük maliyetli ve invazif olmayan bir tanı aracı hâline getirir.

İnsan vücudundaki VOC kaynaklarından birisi insan nefesidir. Nefesteki kimyasallar vücudun işlevi hakkında değerli bilgiler verir. Bu nedenle, bu çalışma, tercih edilen örnek olarak insan nefesine odaklanmıştır.

Aseton, 2-bütanon ve dimetil sülfid analitlerinin nefeste bulunan konsantrasyonlarındaki artışı, karaciğer hastalıklarının varlığına işaret etmektedir. Bu nedenle nefeste bulunan azot, oksijen, karbon di-oksit, karbon mono-oksit, argon gibi inorganik gazlara duyarlılığı düşük bir sensör olan kuvarz kristal mikrobals (QCM) sensörleri tercih edilmiştir.

QCM sensörleri piezoelektrik bir malzemeden oluşmaktadır. Bu sebeple sensör elektrotlarına bir AC gerilim uygulanması, kristal yüzeyinde doğal bir titreşim frekansı oluşmasına neden olmaktadır. Bu frekans değeri, sensör kristali üzerine etki eden kütlenin miktarıyla değişkenlik göstermektedir. QCM sensörleri, analitler ile etkileşimin sağlanabilmesi için kristal yüzeyi üzerinde büyütülmüş numunelere ihtiyaç duyar ve numuneler ile etkileşime giren analitler kristal yüzeyinde bir kütle artışına neden olur, bu artış doğal frekans değerindeki azalma ile sonuçlanır. Bu durum Sauerbrey denklemi ile matematiksel olarak ifade edilmiştir.

Bu çalışmada, farklı numuneler ile donatılmış toplamda 16 adet, 8 çift QCM sensörünün, kullanılacak analitlere olan tepkileri ölçülmüş ve bu sensörlerin, kararlılık, seçicilik, duyarlılık, tekrarlanabilirlik gibi özellikleri incelenmiştir. Tüm analitler için standart koşullar sağlanmış ve konsantrasyonlara bağlı 15 farklı π değerinin elde edilmesi amaçlanmıştır. Bu sonuçlara göre Sensör 1'in hiçbir analite karşı anlamlı bir tepki verememesi, Sensör 4 ve Sensör 8'in karalı olamaması gibi nedenlerle bu sensörler sınıflandırma işlemleri için tercih edilmemiştir.

Oda şartlarında sıvı hâlde bulunan analitlerin, gaz fazında sensör birimine ulaştırılması amacıyla, kütle akış kontrolcülerini ile gaz yıkama şişeleri kullanılan bir deney düzeneği tasarlanmıştır. Bu sistemde analitleri ppm seviyesinde taşıma görevini üstlenen kuru hava, zaman ayarlı solenoid valfler ile yönlendirilmektedir. Bu deney düzeneğine, ölçüm süresince azalan gaz yıkama şişelerindeki analit miktarının, konsantrasyon değerlerini etkilemesi nedeniyle, valfler ile senkronize çalışan bir seviye kontrolcü eklenmiştir. Bu kontrolcü, bir sıvı seviyesi tespit sensörü, peristaltik pompa ve mikro denetleyici karttan oluşmaktadır.

Sınıflandırma ve regresyon işlemi, analitlere ait sensör tepkilerinin makine öğrenmesi algoritmaları ile modellenmesi ile gerçekleştirilmektedir. Eğitim ve test veri setleri oluşturulacak şekilde ölçülen sensör tepkileri, keşifçi veri analizi ile incelenmiş ve gerekli veri ön işleme operasyonları ile düzenlenmiştir.

Sınıflandırma, 10 farklı makine öğrenmesi algoritması ile gerçekleştirilmiş, yüksek performans gösteren 4 algoritma ile de topluluk modeli oluşturularak nihai sonuç elde edilmiştir. Her bir algoritma ve oluşturulan topluluk modeli için karışıklık matrisleri elde edilmiş, bu matrisler ile de doğruluk, kesinlik, duyarlılık ve F1 skoru gibi metrikler hesaplanmıştır. Modelin tahmin edebilme yeteneğini tek bir sayı değeri ile ifade edebilen F1 skoruna göre topluluk modeli, her bir analit için sırasıyla 0.87, 0.89, 0.91 değerlerini vermiştir. Model performanslarını farklı noktalardan ele alan ve bir diğer başarı göstergesi olarak kabul edilen ROC grafiklerine göre ise, topluluk modeli 0.97, 0.97, 0.99 alan değerlerini vermiştir. Böylelikle aseton, 2-bütanon ve dimetil sülfid analitlerinin frekans değerlerine göre başarılı bir şekilde sınıflandırılabilirdiği görülmüştür.

Sınıflandırma sonrası, sınıflandırma için kullanılan test seti regresyon işlemine tabi tutulmuştur. Her bir analit için kendi regresyon topluluk modeli oluşturulmuş ve sonuçları değerlendirilmiştir. Buna göre en düşük hata payıyla tahmin edilen konsantrasyon değerinin aseton analitine ait olduğu görülmüştür.

Bu çalışmanın pratikte uygulanabilirliği, tercih edilen sensörler ve oluşturulan eğitim setleri ile doğrudan ilişkilidir ve birden çok anabilim dalının senkronize çalışmasını gerektirir. Farklı yaş gruplarından sağlıklı olan ve karaciğer hastalıklarının erken evresinde bulunan insanlardan gerekli nefes örnekleri toplanmalıdır. Alınan nefes örnekleri gaz kromatografisi ile incelenmeli, karaciğer sirozu hastalığına sahip insanların ve sağlıklı insanların

nefeslerinde bulunan gazların konsantrasyonları daha geniş bir çalışma ile ayırt edilmelidir. Bu aşamadan sonra ayırt ediciliğe sahip her bir gaz için yüksek duyarlılığa, diğer gazlar için ise düşük duyarlılığa sahip sensörler geliştirilmelidir. Bu sensörlerin, ppb seviyesindeki konsantrasyonlar için çalışabilmesi esastır. İnsan nefesinden elde edilen en düşük ve en yüksek konsantrasyon değerleri arasında, gazların tepkileri nem faktörü de dahil edilerek toplanmalı ve örüntü tanıma algoritmaları için çeşitli olasılıklar altında yüksek boyutlu eğitim setleri oluşturulmalıdır. Bu süreç, gerçek bir hastaya ait nefes örneğinin sınıflandırma ve regresyon işlemlerine tabi tutulabilmesi için elzemdir.

KAYNAKLAR

1. Liu, H., Li, Q., Yan, B., Zhang, L., and Gu, Y. (2019). Bionic electronic nose based on MOS sensors array and machine learning algorithms used for wine properties detection. *Sensors*, 19(1), 45.
2. Righettoni, M., Tricoli, A., Gass, S., Schmid, A., Amann, A., and Pratsinis, S. E. (2012). Breath acetone monitoring by portable Si: WO₃ gas sensors. *Analytica chimica acta*, 738, 69-75.
3. Kablan, Y. (2018). *Elektronik burun ile farklı kalitedeki Doğu Karadeniz çaylarının sınıflandırılması*. Yayınlanmamış Yüksek Lisans Tezi, Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü, Trabzon
4. Benrekia, F., Attari, M., and Bouhedda, M. (2013). Gas sensors characterization and multilayer perceptron (MLP) hardware implementation for gas identification using a field programmable gate array (FPGA). *Sensors*, 13(3), 2967-2985.
5. Tekçe, F. (2005). *Gaz karışımlarını QCM sensör dizisi ve yapay sinir ağı kullanarak bulunması*. Yayınlanmamış Yüksek Lisans Tezi, Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü, Kütahya
6. Si, P., Mortensen, J., Komolov, A., Denborg, J., and Møller, P. J. (2007). Polymer coated quartz crystal microbalance sensors for detection of volatile organic compounds in gas mixtures. *Analytica chimica acta*, 597(2), 223-230.
7. Srivastava, S., Mishra, G., and Mishra, H. N. (2019). Probabilistic artificial neural network and E-nose based classification of *Rhyzopertha dominica* infestation in stored rice grains. *Chemometrics and Intelligent Laboratory Systems*, 186, 12-22.
8. Özen, P., Ebeoğlu, M. A., (2012). *QCM Tabanlı sıvı ortam algılama sistemi tasarımı*, ELECO Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu, 484-488.
9. Xu, X., Cang, H., Li, C., Zhao, Z. K., & Li, H. (2009). Quartz crystal microbalance sensor array for the detection of volatile organic compounds. *Talanta*, 78(3), 711-716.
10. Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., and Jung, J. (2020). Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. *Neural Networks*, 121, 396-408.
11. Wilson, A. D., & Baietto, M. (2009). Applications and advances in electronic-nose technologies. *Sensors*, 9(7), 5099-5148.
12. Peris, M., & Escuder-Gilabert, L. (2009). A 21st century technique for food control: Electronic noses. *Analytica Chimica Acta*, 638(1), 1-15.
13. Gobbi, E., Falasconi, M., Zambotti, G., Sberveglieri, V., Pulvirenti, A., and Sberveglieri, G. (2015). Rapid diagnosis of Enterobacteriaceae in vegetable soups by a metal oxide sensor based electronic nose. *Sensors and Actuators B: Chemical*, 207, 1104-1113.

14. Caron, A., Redon, N., Thevenet, F., Hanoune, B., and Coddeville, P. (2016). Performances and limitations of electronic gas sensors to investigate an indoor air quality event. *Building and Environment*, 107, 19-28.
15. Herrero, J. L., Lozano, J., Santos, J. P., and Suárez, J. I. (2016). On-line classification of pollutants in water using wireless portable electronic noses. *Chemosphere*, 152, 107-116.
16. Fidan, E. (2019). *QCM sensörlerinin zehirli gazlara karşı algılama özelliklerinin incelenmesi*. Yayımlanmamış Yüksek Lisans Tezi, Ömer Halisdemir Üniversitesi Fen Bilimleri Enstitüsü, Niğde.
17. Koçan, M. (2008). *QCM tabanlı elektronik burnun diyabet hasta nefesine verdiği tepkilerin incelenmesi*. Yayımlanmamış Yüksek Lisans Tezi, Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü, Kütahya.
18. Aydemir, F. (2017). *QCM tabanlı sensör dizisiyle sıvı algılama sistem tasarımının gömülü sistemlerle gerçekleştirilmesi*. Yayımlanmamış Doktora Tezi, Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü, Kütahya.
19. Çağırtekin, A., O. (2019). *Zn_{1-x}Sn_xO MOS sensörlerinde UV ışığı altında NO gaz algılama parametrelerinin incelenmesi*. Yayımlanmamış Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
20. Benrekia, F., Attari, M., Bermak, A., and Belhout, K. (2009, March). FPGA implementation of a neural network classifier for gas sensor array applications. In *2009 6th International Multi-Conference on Systems, Signals and Devices*, Austria 1-6.
21. García, G. J., Jara, C. A., Pomares, J., Alabdo, A., Poggi, L. M., and Torres, F. (2014). A survey on FPGA-based sensor systems: towards intelligent and reconfigurable low-power sensors for computer vision, control and signal processing. *Sensors*, 14(4), 6247-6278.
22. Sun, Y., Zhang, H., Zhao, T., Zou, Z., Shen, B., and Yang, L. (2020). A new convolutional neural network with random forest method for hydrogen sensor fault diagnosis. *Institute of Electrical and Electronics Engineers Access*, 8, 85421-85430.
23. Kabir, M. M., Shahjahan, M., and Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. *Expert Systems with Applications*, 39(3), 3747-3763.
24. Al-Tashi, Q., Rais, H. M., Abdulkadir, S. J., Mirjalili, S., and Alhussian, H. (2020). A review of grey wolf optimizer-based feature selection methods for classification. *Evolutionary machine learning techniques*, 273-286.
25. Şen, Z. (2010). *Petrol kaynaklı hidrokarbonlardan BTEX kompleks gaz karışımlarının sınıflandırılması için QCM gaz sensör dizisi geliştirilmesi*. Yayımlanmamış Yüksek Lisans Tezi, Gebze Teknik Üniversitesi Mühendislik ve Fen Bilimleri Enstitüsü, Kocaeli.

26. Senyücel, N. (2013). *Zararlı kokuları algılayıp internet üzerinden ileten elektronik burun tasarımı*. Yayınlanmamış Doktora Tezi, Trakya Üniversitesi Fen Bilimleri Enstitüsü, Edirne.
27. Zhang, L., Tian, F., Zhang, D. (2018). *Electronic nose: algorithmic challenges*. Singapore: Springer, 3-339.
28. Li, H., Luo, D., Sun, Y., and GholamHosseini, H. (2019). Classification and Identification of Industrial Gases Based on Electronic Nose Technology. *Sensors*, 19(22), 5033.
29. Najarian, J. S. (1980). The diagnostic importance of the odor of urine. *The New England journal of medicine*, 303(19), 1128.
30. Pavlou, A. K., and Turner, A. P. F. (2000). Sniffing out the truth: clinical diagnosis using the electronic nose. *Clinical Chemistry and Laboratory Medicine*, 38(2), 99-112.
31. Phillips, M. (2002). Detection of volatile organic compounds in breath. In *Disease Markers in Exhaled Breath*. N. Markczin, S. A. Kharitonov, M. H. Yacoub, P. J. Barnes, (Eds.); Marcel Dekker: New York, NY, USA, 219-231.
32. Hayden, G. F. (1980). Olfactory diagnosis in medicine. *Postgraduate medicine*, 67(4), 110-116.
33. Lee, S., Hensley, J. M., and Allen, M. G. (2003, June). High sensitivity detection of broadband gas using time domain multiplexed QC lasers. In *Conference on Lasers and Electro-Optics*, United States of America.
34. Phillips, M. (1999). *U.S. Patent No. 5,996,586*. Washington, DC: U.S. Patent and Trademark Office.
35. Liddell, K. (1976). Smell as a diagnostic marker. *Postgraduate Medical Journal*, 52(605), 136-138.
36. Phillips, M. (2001). Breath test for dedection of various diseases, United States Patent, 6,221,026.
37. Di Natale, C., Macagnano, A., Martinelli, E., Paolesse, R., D'Arcangelo, G., Roscioni, C., and D'Amico, A. (2003). Lung cancer identification by the analysis of breath by means of an array of non-selective gas sensors. *Biosensors and Bioelectronics*, 18(10), 1209-1218.
38. Fend, R., Bessant, C., Williams, A. J., and Woodman, A. C. (2004). Monitoring haemodialysis using electronic nose and chemometrics. *Biosensors and Bioelectronics*, 19(12), 1581-1590.
39. Van den Velde, S., Nevens, F., van Steenberghe, D., and Quirynen, M. (2008). GC–MS analysis of breath odor compounds in liver patients. *Journal of Chromatography B*, 875(2), 344-348.

40. Joseph, A. (2019). *Design and analysis of high-frequency quartz crystal microbalance sensor array with concentric electrodes and dual inverted mesa structure for multiple gas detection*. Yayınlanmamış Doktora Tezi, University of Windsor, Kanada.
41. Sberveglieri, G. (2012). *Gas sensors: principles, operation and developments*. Springer Science & Business Media.
42. Awang, Z. (2014). Gas sensors: A review. *Sens. Transducers*, 168(4), 61-75.
43. Gerasimov, G. (2016). Graphene-based gas sensors. In *Advanced Environmental Analysis*, 133-153.
44. Mirzaei, A., Kim, S. S., and Kim, H. W. (2018). Resistance-based H₂S gas sensors using metal oxide nanostructures: A review of recent advances. *Journal of hazardous materials*, 357, 314-331.
45. Fine, G. F., Cavanagh, L. M., Afonja, A., and Binions, R. (2010). Metal oxide semiconductor gas sensors in environmental monitoring. *Sensors*, 10(6), 5469-5502.
46. Zappa, D., Galstyan, V., Kaur, N., Arachchige, H. M. M., Sisman, O., and Comini, E. (2018). "Metal oxide-based heterostructures for gas sensors"-A review. *Analytica Chimica Acta*, 1039, 1-23.
47. Santhaveesuk, T. (2015). Metal oxide semiconducting gas sensor technology. *Nakhon Phanom University Journal*, 1, 2.
48. Bai, H., and Shi, G. (2007). Gas sensors based on conducting polymers. *Sensors*, 7(3), 267-307.
49. Kumar, N., Sahatiya, P., and Dubey, P. (2014). Fabrication of CNT based gas sensor using interdigitated gold electrodes. *Procedia Materials Science*, 6, 1976-1980.
50. Wang, L. C., Tang, K. T., Teng, I. J., Kuo, C. T., Ho, C. L., Kuo, H. W., and Chang, C. P. (2011). A single-walled carbon nanotube network gas sensing device. *Sensors*, 11(8), 7763-7772.
51. Ruiz, L. A. P. (2015). *Highly efficient carbon-nanostructures-based gas sensors for biomedical and other applications*. Yayınlanmamış Yüksek Lisans Tezi, Technische Universität Dresden, Almanya.
52. Baughman, R. H., Zakhidov, A. A., and De Heer, W. A. (2002). Carbon nanotubes--the route toward applications. *Science*, 297(5582), 787-792.
53. Xiao, Z., Kong, L. B., Ruan, S., Li, X., Yu, S., Li, X., and Li, S. (2018). Recent development in nanocarbon materials for gas sensor applications. *Sensors and Actuators B: Chemical*, 274, 235-267.
54. Lipatov, A., Varezchnikov, A., Wilson, P., Sysoev, V., Kolmakov, A., and Sinitskii, A. (2013). Highly selective gas sensor arrays based on thermally reduced graphene oxide. *Nanoscale*, 5(12), 5426-5434.

55. Gómez-Navarro, C., Weitz, R. T., Bittner, A. M., Scolari, M., Mews, A., Burghard, M., and Kern, K. (2007). Electronic transport properties of individual chemically reduced graphene oxide sheets. *Nano Letters*, 7(11), 3499-3503.
56. Xu, S., Sun, F., Pan, Z., Huang, C., Yang, S., Long, J., and Chen, Y. (2016). Reduced graphene oxide-based ordered macroporous films on a curved surface: general fabrication and application in gas sensors. *ACS Applied Materials & Interfaces*, 8(5), 3428-3437.
57. Segal, M. (2009). Selling graphene by the ton. *Nature Nanotechnology*, 4(10), 612-614.
58. Huang, X. M., Liu, L. Z., Zhou, S., and Zhao, J. J. (2020). Physical properties and device applications of graphene oxide. *Frontiers of Physics*, 15(3), 1-70.
59. Khan, M., Rahaman, R., and Kang, S. W. (2014). A high sensitivity and wide dynamic range fiber-optic sensor for low-concentration VOC gas detection. *Sensors*, 14(12), 23321-23336.
60. Pal, B. P. (2014). Optical fiber sensors: A versatile technology platform for sensing. *Journal of the Indian Institute of Science*, 94(3), 283-310.
61. Khan, M. R. R., Kang, B. H., Lee, S. W., Kim, S. H., Yeom, S. H., Lee, S. H., and Kang, S. W. (2013). Fiber-optic multi-sensor array for detection of low concentration volatile organic compounds. *Optics Express*, 21(17), 20119-20130.
62. Troia, B., Paolicelli, A., De Leonardis, F., and Passaro, V. M. (2013). Photonic crystals for optical sensing: A review. *Advances in Photonic Crystals*, 241-295.
63. Zhang, Y. N., Zhao, Y., and Lv, R. Q. (2015). A review for optical sensors based on photonic crystal cavities. *Sensors and Actuators A: Physical*, 233, 374-389.
64. İnternet: Bragg's Law. Web: [https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_\(Analytical_Chemistry\)/Instrumental_Analysis/Diffraction_Scattering_Techniques/Bragg's_Law#:~:text=from%20crystal%20surface,-,Introduction,same%20angle%20of%20scattering%2C%20%CE%B8](https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Instrumental_Analysis/Diffraction_Scattering_Techniques/Bragg's_Law#:~:text=from%20crystal%20surface,-,Introduction,same%20angle%20of%20scattering%2C%20%CE%B8) Erişim Tarihi: 19.12.2021.
65. Zhao, Y., Zhang, Y. N., and Wang, Q. (2011). Research advances of photonic crystal gas and liquid sensors. *Sensors and Actuators B: Chemical*, 160(1), 1288-1297.
66. Xu, H., Wu, P., Zhu, C., Elbaz, A., and Gu, Z. Z. (2013). Photonic crystal for gas sensing. *Journal of Materials Chemistry C*, 1(38), 6087-6098.
67. Cho, S. H., Suh, J. M., Eom, T. H., Kim, T., and Jang, H. W. (2021). Colorimetric sensors for toxic and hazardous gas detection: A review. *Electronic Materials Letters*, 17(1), 1-17.
68. Suslick, K. S., Bailey, D. P., Ingison, C. K., Janzen, M., Kosal, M. E., McNamara III, W. B., and Nakagaki, S. (2007). Seeing smells: development of an optoelectronic nose. *Quimica Nova*, 30, 677-681.

69. Owyung, R. E., Panzer, M. J., and Sonkusale, S. R. (2019). Colorimetric gas sensing washable threads for smart textiles. *Scientific Reports*, 9(1), 1-8.
70. Xiao-wei, H., Xiao-bo, Z., Ji-yong, S., Zhi-hua, L., and Jie-wen, Z. (2018). Colorimetric sensor arrays based on chemo-responsive dyes for food odor visualization. *Trends in Food Science & Technology*, 81, 90-107.
71. White, R. M., and Voltmer, F. W. (1965). Direct piezoelectric coupling to surface elastic waves. *Applied Physics Letters*, 7(12), 314-316.
72. Mujahid, A., and Dickert, F. L. (2017). Surface acoustic wave (SAW) for chemical sensing applications of recognition layers. *Sensors*, 17(12), 2716.
73. Doğan, E. (2010). *QCM tabanlı taşınabilir gaz ölçüm cihazı geliştirme*. Yayımlanmamış Yüksek Lisans Tezi, Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü, Kütahya.
74. Suzuki, R., Emura, T., Tokura, Y., Kawamura, N., and Hori, Y. (2020). A quartz crystal microbalance based portable gas sensing platform for on-demand human breath monitoring. *Institute of Electrical and Electronics Engineers Access*, 8, 146166-146171.
75. Chang, B., and Zhao, T. (2008). Quartz crystal microbalance study of DNA immobilization and hybridization for DNA sensor development. *Department of Civil and Environmental Engineering, Michigan State University*.
76. Grammoustianou, A., Papadakis, G., and Gizeli, E. (2017). Solid-phase isothermal DNA amplification and detection on quartz crystal microbalance using liposomes and dissipation monitoring. *Institute of Electrical and Electronics Engineers Sensors Letters*, 1(5), 1-4.
77. Joseph, A., and Emadi, A. (2020). A high frequency dual inverted mesa QCM sensor array with concentric electrodes. *Institute of Electrical and Electronics Engineers Access*, 8, 92669-92676.
78. Speight, R. E., and Cooper, M. A. (2012). A survey of the 2010 quartz crystal microbalance literature. *Journal of Molecular Recognition*, 25(9), 451-473.
79. Yang, M., and He, J. (2016). Graphene oxide as quartz crystal microbalance sensing layers for detection of formaldehyde. *Sensors and Actuators B: Chemical*, 228, 486-490.
80. Johannsmann, D. (2015). The quartz crystal microbalance in soft matter research, *Soft and Biological Matter*, 191-204.
81. Kananizadeh, N., Rice, C., Lee, J., Rodenhausen, K. B., Sekora, D., Schubert, M., and Li, Y. (2017). Combined quartz crystal microbalance with dissipation (QCM-D) and generalized ellipsometry (GE) to characterize the deposition of titanium dioxide nanoparticles on model rough surfaces. *Journal of Hazardous Materials*, 322, 118-128.

82. Joshi, T., Voo, Z. X., Graham, B., Spiccia, L., and Martin, L. L. (2015). Real-time examination of aminoglycoside activity towards bacterial mimetic membranes using quartz crystal microbalance with dissipation monitoring (QCM-D). *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1848(2), 385-391.
83. Manbachi, A., and Cobbold, R. S. (2011). Development and application of piezoelectric materials for ultrasound generation and detection. *Ultrasound*, 19(4), 187-196.
84. Duan, X. (2014). *Lipid based sensing of organic vapours: A study combining AFM and QCM*, Yayınlanmamış Doktora Tezi, Durham Üniversitesi, İngiltere.
85. Lemanov, V. V. (2000). Piezoelectric and pyroelectric properties of protein amino acids as basic materials of soft state physics. *Ferroelectrics*, 238(1), 211-218.
86. Berlincourt, D. (1971). Piezoelectric crystals and ceramics. In *Ultrasonic transducer materials*, Boston, MA: Springer, 63-124.
87. Lu, C., and Czanderna, A. W. (1984). Methods and phenomena 7: Applications of Piezoelectric Quartz Crystal Microbalances.
88. Levenson, L. L. (1967, January). Condensation coefficients of argon krypton and xenon measured with a quartz crystal microbalance at 4.2 Degrees K. In *Nuovo Cimento*, 5(2), 321-+.
89. Sauerbrey, G. (1959). Verwendung von schwingquarzen zur wägung dünner schichten und zur mikrowägung. *Zeitschrift Für Physik*, 155(2), 206-222.
90. Boyadjiev, S. I., Georgieva, V., Yordanov, R., Raicheva, Z., and Szilágyi, I. M. (2016). Preparation and characterization of ALD deposited ZnO thin films studied for gas sensors. *Applied Surface Science*, 387, 1230-1235.
91. Bizet, K., Gabrielli, C., and Perrot, H. (1999). Biosensors based on piezoelectric transducers. *Analisis*, 27(7), 609-616.
92. İnternet: Kuvarz Kristal Mikrobalans Sensör. Web: https://stringfixer.com/tr/Quartz_crystal_microbalance Erişim Tarihi: 19.12.2021.
93. Oehlmann, P., Osswald, P., Blanco, J. C., Friedrich, M., Rietzel, D., and Witt, G. (2021). Modeling fused filament fabrication using Artificial neural networks. *Production Engineering*, 15(3), 467-478.
94. İnternet: Ng, A., Ma, T., (2020). Stanford University: CS229 Lecture Notes. Web: <http://cs229.stanford.edu/syllabus.html> Erişim Tarihi: 19.12.2021.
95. Park, H., and Kim, K. (2019). Comparisons among machine learning models for the prediction of hypercholesterolemia associated with exposure to lead, mercury, and cadmium. *International Journal of Environmental Research and Public Health*, 16(15), 2666.

96. Yang, H. Y., Wang, Y. C., Peng, H. Y., and Huang, C. H. (2021). Breath biopsy of breast cancer using sensor array signals and machine learning analysis. *Scientific Reports*, 11(1), 1-9.
97. İnternet: Machine Learning. Web: <https://machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/> Erişim Tarihi: 19.12.2021.
98. İnternet: Machine Learning. Web: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229> Erişim Tarihi: 11.12.2021.
99. İnternet: Data Analysis. Web: <https://www.ibm.com/cloud/learn/exploratory-data-analysis> Erişim Tarihi: 11.12.2021.
100. Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press.
101. İnternet: Data Analysis. Web: <https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7> Erişim Tarihi: 11.12.2021.
102. İnternet: Machine Learning. Web: <https://www.ibm.com/topics/supervised-learning> Erişim Tarihi: 15.10.2021.
103. İnternet: Machine Learning. Web: <https://aihubprojects.com/linear-regression-from-scratch/> Erişim Tarihi: 15.10.2021.
104. Assiri, A. S., Hussien, A. G., and Amin, M. (2020). Ant Lion Optimization: variants, hybrids, and applications. *Institute of Electrical and Electronics Engineers Access*, 8, 77746-77764.
105. Shi, B., Heidari, A. A., Chen, C., Wang, M., Huang, C., Chen, H., and Zhu, J. (2020). Predicting Di-2-ethylhexyl phthalate toxicity: Hybrid integrated harris hawks optimization with support vector machines. *Institute of Electrical and Electronics Engineers Access*, 8, 161188-161202.
106. Kostopoulos, G., Karlos, S., Kotsiantis, S., and Ragos, O. (2018). Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, 35(2), 1483-1500.
107. İnternet: Machine Learning. Web: <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm> Erişim Tarihi: 15.10.2021.
108. İnternet: Machine Learning. Web: <https://www.mathworks.com/help/stats/regress.html> Son Erişim Tarihi: 16.10.2021.
109. Ghosh, S., Mondal, S., & Ghosh, B. (2014, February). A comparative study of breast cancer detection based on SVM and MLP BPN classifier. In *2014 First International Conference on Automation, Control, Energy and Systems (ACES)*, India, 1-4.
110. Osareh, A., and Shadgar, B. (2010, April). Machine learning techniques to diagnose breast cancer. In *5th international symposium on health informatics and bioinformatics*, Ankara, Türkiye, 114-120.

111. Dana Bazazeh, D. and Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *5th Int. Conf. on Electronic Devices, Systems and Applications (ICEDSA)*, Ras Al Khaimah, UAE.
112. İnternet: Neural Networks Web: <https://tr.pinterest.com/pin/697565429768377179/> Erişim Tarihi: 03.12.2021.
113. Zhao, Y., and Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research*, 41(12), 1955-1959.
114. Ali, J., Khan, R., Ahmad, N., and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.
115. İnternet: Decision Tree. Web: <https://medium.com/analytics-vidhya/a-deep-dive-to-decision-trees-6575e016d656> Erişim Tarihi: 04.12.2021.
116. İnternet: Gradient Boosting Machine. Web: http://uc-r.github.io/gbm_regression Erişim Tarihi: 09.12.2021.
117. İnternet: Chain Rule. Web: <https://machinelearningmastery.com/the-chain-rule-of-calculus-even-more-functions/> Erişim Tarihi: 09.12.2021.



GAZİ GELECEKTİR..