



**MYELOPROLİFERATİF HASTALIK ÖN TANILI HASTALARDA
MUTASYON TESTİ SEÇİMİNE İLİŞKİN İSTATİSTİKSEL
SINIFLANDIRMA TABANLI BİR ÇALIŞMA**

Gözde KUBAT

**DOKTORA TEZİ
İSTATİSTİK ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

TEMMUZ 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.

Gözde KUBAT

07/07/2022

MYELOPROLİFERATİF HASTALIK ÖN TANILI HASTALARDA MUTASYON
TESTİ SEÇİMİNE İLİŞKİN İSTATİSTİKSEL SINIFLANDIRMA TABANLI BİR

ÇALIŞMA

(Doktora Tezi)

Gözde KUBAT

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Temmuz 2022

ÖZET

Myeloproliferatif hastalıkların tanısının konulabilmesi için Dünya Sağlık Örgütü'nün belirlediği hastalık tanı şeması doğrultusunda klinisyen tarafından genetik mutasyonların varlığı incelenmektedir. Bu çalışmada; kliniğe başvuru yapmış hastaların tam kan sayımı bulgularından yola çıkılarak oluşturulabilecek model ile uygun tanı testinin kestirilmesi hedeflenmiştir. Bu doğrultuda, kliniğe başvuran hastaların zaman ve maddi açıdan tasarruf etmesi amaçlanmaktadır. Başkent Üniversitesi Ankara Hastanesi Tıbbi Genetik Ana Bilim Dalı Genetik Hastalıklar Değerlendirme Merkezi'ne yönlendirilmiş hastaların bulguları ele alınarak makine öğrenmesi yöntemi yardımıyla sınıflandırma algoritmaları kullanılarak kestirim yapılmıştır. Değişkenlere ait tanımlayıcı istatistikler, kategorik değişkenler için frekans ve yüzde ile ifade edilirken, sürekli değişkenler için ortalama ve standart sapma ile ifade edilmiştir. Kestirimde Naive Bayes, k-en yakın komşu, doğrusal diskriminant analizi, destek vektör makinesi ve karar ağacı gibi sınıflandırma algoritmaları kullanılmıştır. Algoritmalar ile doğruluk, özgüllük, duyarlılık gibi belirleyici değerler elde edilmiştir. Yapılan kestirimlerde elde edilen doğruluk değerlerine göre incelenmiş ve en iyi model seçilmeye çalışılmıştır. İncelenen tam kan sayım değerleri ile mutasyon varlığı arasında istatistiksel olarak anlamlı farklılıklar tespit edilmiştir. Ele alınan Naive Bayes, k-en yakın komşu, doğrusal diskriminant analizi, destek vektör makinesi ve karar ağacı algoritmaları ile oluşturulan modellerde doğruluk değerleri %70 olarak saptanmıştır. Kullanılan algoritmalarda elde edilen doğruluk oranı orta seviyede olmasına rağmen benzer çalışmaların literatürde yer almamış olması sebebiyle çalışma sonuçlarının alana önemli katkısının olacağı sonucuna varılmıştır.

Bilim Kodu : 91440

Anahtar Kelimeler : Myeloproliferatif hastalık, makine öğrenmesi, karar ağacı, k-en yakın komşu, doğrusal diskriminant analizi

Sayfa Adedi : 74

Danışman : Prof. Dr. Bülent ÇELİK

İkinci Danışman : Prof. Dr. Feride İ. ŞAHİN

A STATISTICAL CLASSIFICATION BASED STUDY ON SELECTION OF
MUTATION TEST IN PATIENTS WITH PREDIAGNOSIS OF
MYELOPROLIFERATIVE DISEASE

(Ph. D. Thesis)

Gözde KUBAT

GAZİ UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

July 2022

ABSTRACT

In order to diagnose myeloproliferative diseases, the presence of genetic mutations is examined by the clinician in line with the disease diagnosis scheme determined by the World Health Organization. In this study, it is aimed to predict the appropriate diagnostic screening test with the model created based on complete blood count findings of patients who applied to the clinic. Accordingly, it is aimed to save time and financial for patients who apply to the clinic. Prediction was made using classification algorithms with the help of machine learning method by considering the findings of patients referred to the Baskent University Ankara Hospital, Department of Medical Genetics Genetic Diseases Diagnosis Center. Descriptive statistics of the study data were given as frequency and percentage for categorical variables, mean and standard deviation for continuous variables. Classification algorithms such as Naive Bayes, k-nearest neighbor, linear discriminant analysis, support vector machines and decision tree were used in the estimation. With the algorithms, determinative values such as accuracy, specificity and sensitivity were obtained. The estimations made were examined according to the accuracy rates obtained and the best model was tried to be selected. Statistically significant differences were found between the examined complete blood count values and the presence of mutation. The accuracy rates were found to be around 70% in the models created with Naive Bayes, k-nearest neighbor, linear discriminant analysis, support vector machines and decision tree algorithms. Although the accuracy rate obtained in the algorithms used is at a moderate level, it was concluded that the results of the study would make a significant contribution to the field, since similar studies have not been included in the literature.

Science Code : 91440

Key Words : Myeloproliferative disease, machine learning, decision tree, k-nearest neighbor, linear discriminant analysis

Page Number : 74

Supervisor : Prof. Dr. Bülent ÇELİK

Co-Supervisor : Prof. Dr. Feride İ. ŞAHİN

TEŞEKKÜR

Bu tezin kapağında sadece benim ismim yer alsa da pek çok kişinin katkısı olmuştur. Bu tezi mümkün kılan herkese şükran borçluyum.

Öncelikle çalışmam sırasında bana yol gösteren, beni teşvik eden ve sabırla beni dinleyerek anlayan danışmanlarım Sayın Prof. Dr. Feride İ. ŞAHİN ve Sayın Prof. Dr. Bülent ÇELİK'e, tezimle ilgili yapıcı eleştiri ve önerileriyle beni yönlendiren jüri üyelerim Sayın Prof. Dr. İsmayil Safa GÜRCAN, Sayın Prof. Dr. Hasan BAL, Sayın Prof. Dr. Mehmet Akif BAKIR, Sayın Prof. Dr. Hacı Hasan ÖRKÜ ve Sayın Doç. Dr. Doğukan ÖZEN'e teşekkür etmek isterim. Benim üzerimdeki emeklerinin meyvesini bu çalışmayla sunabilmek bana güç ve cesaret vermektedir.

Doktora sürecime başlamamda bana yol gösteren Sayın Prof. Dr. Hüseyin TATLIDİL'e, eğitim dönemimde bana yeni ufuklar açan ve derslerini aldığım çok kıymetli hocalarıma, tezimin verileri için birlikte mesai harcadığım, bana destek olan ve Başkent Üniversitesi Tıbbi Genetik Ana Bilim Dalı'nda çalışmakta olan Sayın Prof. Dr. Feride İ. ŞAHİN, Sayın Prof. Dr. Zerrin YILMAZ ÇELİK, Sayın Prof. Dr. Yunus Kasım TERZİ, Uzm. Kim. Esra BAŞYİĞİT ve tüm personele, tez çalışmamda manevi desteklerini esirgemeyen çok değerli Doç. Dr. Erkan YILDIZ, Öğr. Gör. Dr. Fereshteh REZAEİ, Öğr. Gör. Özkan KUTLUCAN, Öğr. Gör. İlayda MÜFTÜOĞLU, Gülşah KOÇ ve değerli arkadaşlarıma çok teşekkür ederim.

Tüm hayatım boyunca benden desteklerini hiç esirgemeyen ve yaptıklarımla onları gururlandırdığımı bana hep hissettiren canım annem Lale ÖZER ve canım babam Öztürk ÖZER'e, her zaman yanımda olan, beni destekleyen ve bana yardımcı olan canım eşim Gökhan KUBAT'a, çalışmalarım sırasında bana yol gösteren, hep yanımda hissettiğim ve yaş olarak en küçük olsam da benim biricik arkadaşlarım olan çok sevdiğim canım ablalarım Özge ÖZER KAYA ve Özden ÖZER ÖZKAN OĞLU'na, çalışmalarım sürecinde bana vakit ayıran ve beni motive eden melek annem Nurten KUBAT, çok değerli babam Özkan KUBAT ve biricik kardeşim Gönenç KUBAT'a, beni her zaman destekleyen ve sabırla dinleyen canım Gözde ŞENTÜRK ve Ömer Lütfi ŞENTÜRK'e, bunaldığım anlarda beni motive eden canım yeğenlerim Sarp, Erin ve Nora'ya çok teşekkür ederim. Son olarak; hayatta iyiliğin, çalışkanlığın ve erdemliliğin daima yol göstereceğine inanarak büyüttüğüm, beni bir an bile yalnız bırakmayan, sevildiklerini bilen ve bana sevildiğimi hissettiren, benim bu hayattaki en büyük şanslarım olan canım kızım Defnem ve canım oğlum Demirim'e sonsuz teşekkür ederim.

İÇİNDEKİLER

	Sayfa
ÖZET	v
ABSTRACT.....	vi
TEŞEKKÜR.....	vii
İÇİNDEKİLER	viii
ÇİZELGELERİN LİSTESİ.....	x
ŞEKİLLERİN LİSTESİ.....	xi
1. GİRİŞ.....	1
2. GENEL BİLGİLER	9
2.1. Myeloproliferatif Hastalık	9
2.2. Makine Öğrenmesi Yöntemi.....	12
2.2.1. Tarihçesi.....	12
2.2.2. Uygulama adımları.....	14
2.2.3. Kullanılan algoritmalar	15
2.2.4. Algoritmaların karşılaştırılması	23
3. MATERYAL VE METOT	27
3.1. Çalışmanın Örneklemi	27
3.2. Çalışmanın Yöntemi	27
3.3. Çalışmada Kullanılan Değişkenlere İlişkin Açıklamalar.....	27
4. BULGULAR	29
4.1. Çalışma Verilerine İlişkin R Programlama Dilinde Yapılan İşlemler	29
4.2. Tanımlayıcı İstatistikler	35
4.3. Kullanılan Algoritmalara İlişkin İstatistikler	37
4.3.1. Naive Bayes sınıflandırma algoritması.....	37

	Sayfa
4.3.2. k-en yakın komşu sınıflandırma algoritması.....	39
4.3.3. Doğrusal diskriminant analizi sınıflandırma algoritması.....	41
4.3.4. Destek vektör makinesi sınıflandırma algoritması	43
4.3.5. Gini sınıflandırma algoritması	45
4.3.6. C5.0 sınıflandırma algoritması.....	47
4.4. Algoritma Bulgularının Karşılaştırılması	49
4.5. Değişkenlere İlişkin Fark Karşılaştırması.....	51
4. SONUÇ VE ÖNERİLER.....	55
KAYNAKLAR	59
EKLER.....	69
EK-1. Klinik bulgulara ilişkin açıklamalar	70
ÖZGEÇMİŞ	72

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 3. 1. Çalışmada kullanılan değişkenler	29
Çizelge 4.1. Çalışmada kullanılan değişkenlerde bulunan eksik veri durumunun dağılımı	31
Çizelge 4.2. Çalışmada kullanılan değişkenlerin modeldeki önem durumunun kontrolüne ilişkin dağılım	34
Çizelge 4.3. Demografik özelliklerin dağılımı	35
Çizelge 4.4. Çalışma değişkenlerinin mutasyona göre dağılımı.....	36
Çizelge 4.5. Makine öğrenmesi algoritmalarında kullanılan veri sayılarının dağılımı...	37
Çizelge 4.6. Naive Bayes sınıflandırma algoritması sonuçları	39
Çizelge 4.7. K-en yakın komşu sınıflandırma algoritması sonuçları.....	41
Çizelge 4.8. Doğrusal diskriminant analizi sınıflandırma algoritması sonuçları.....	43
Çizelge 4.9. Destek vektör makinesi sınıflandırma algoritması sonuçları.....	45
Çizelge 4.10. Gini sınıflandırma algoritması sonuçları	47
Çizelge 4.11. C5.0 sınıflandırma algoritması sonuçları.....	49
Çizelge 4.12. En iyi doğruluk değerinin tespit edildiği eğitim verisi oranlarına göre elde edilen analiz sonuçları	50
Çizelge 4.13. Çalışmada ele alınan hasta bulgularının yapılan tarama testine göre dağılımı.....	52

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Myeloproliferatif hastalık ön tanılı hastalara ilişkin tanı testlerinin uygulama yöntemi	10
Şekil 2.2. k-en yakın komşu algoritması ile sınıflandırma şeması	16
Şekil 2.3. Destek vektör makinesi yönteminde doğrusal ayırıcı ile sınıflandırma şeması	19
Şekil 2.4. Destek vektör makinesi yönteminde doğrusal olmayan ayırıcı ile sınıflandırma şeması.....	19
Şekil 2.5. Karar ağacı algoritması ile sınıflandırma şeması.....	21
Şekil 4.1. Çalışmada kullanılan değişkenlerin önem değerlerinin dağılımı	33

1. GİRİŞ

Myeloproliferatif hastalıklar, kırmızı kan hücrelerinin, trombositlerin veya beyaz kan hücrelerinin kontrolsüz çoğalması sonucu oluşmaktadır [1]. Hastalığın görülme sıklığı 100 000 kişide 1 olarak literatürde yer almaktadır [2]. Polisitemia vera, esansiyel trombositoz ve primer myelofibrozu en sık görülen myeloproliferatif hastalıklar olarak sıralanmaktadır [3].

Klinisyene başvuran hastalara Dünya Sağlık Örgütü'nün belirlediği tanı kriterlerine (lökosit, trombosit, dalak büyüklüğü gibi parametreler) göre ön tanı konulmaktadır [4]. Genetik bozukluk ile ilgili hastalıkların kesin tanısının konulabilmesi için mutasyon taramasının yapılması gerekmektedir. Ön tanısı konulan hastalara klinisyen tarafından genetik mutasyon taraması önerilerek klinik çalışmalar başlatılmaktadır. Mutasyon taramalarına ilişkin birden fazla seçenek olmasına karşın [5], Dünya Sağlık Örgütü'nün tanı şemasında yer alan *JAK2* V617F, *JAK2* Ekzon 12, *MPL* ve *CALR* mutasyon taramaları sıklıkla tercih edilmektedir [6].

Dünya Sağlık Örgütü tanı şemasına göre öncelikle hastalarda mutasyon taraması yapılarak *JAK2* V617F mutasyonunun varlığı saptanmaya çalışılmaktadır. *JAK2* V617F mutasyonunun saptandığı durumda hastalığa tanı konulabilmekteyken, mutasyonun saptanamaması durumunda *JAK2* Ekzon 12, *MPL* veya *CALR* mutasyonları araştırılmaktadır.

Günümüzde yapılmakta olan bu klinik araştırmaların süre ve maliyet yönünden hasta kazanımlarının düşük seviyede olduğu değerlendirilmektedir. Kliniğe başvuran hasta için her mutasyon taraması uzun bir süre ve ek bir maliyet gerektirmektedir.

Sağlık sektöründe hastalıkların tanılarının kesinleştirilebilme sürecinde zaman ve maliyet kısıtları ortaya çıkmaktadır [7]. Myeloproliferatif hastalık için kısıtlar değerlendirildiğinde, Dünya Sağlık Örgütü'nün belirlemiş olduğu tanı şeması doğrultusunda, hastalık ön tanısı almış kişiye uygulanacak klinik çalışmaların uzun sürede sonuçlandığı ve yüksek maliyetler olduğu sonucuna varılmaktadır. Bilgi teknolojilerinin gelişmesi, yapay zeka uygulamalarının karar süreçlerine dahil olması ile belirtilen kısıtlar azaltılabilmektedir [8]. Günümüzde birçok alanda kullanılan makine öğrenmesi yöntemi, sağlık alanında da

kullanılmakta ve bu yöntem ile kliniğe başvuran bireyler için fayda sağlayacak modeller oluşturulmaktadır.

Bilişim teknolojilerinin yaygınlaşması ve büyük veri ile yapılan araştırmalarda tercih ediliyor olması sağlık sektöründe elde edilen verilere uygunluğunun yüksek olduğu değerlendirilmektedir. Karakoyun ve Hacıbeyoğlu (2014) tarafından yapılan çalışmada makine öğrenmesi yöntemi kullanılarak biyomedikal verilerine ilişkin altı tane algoritma ile sınıflandırma yapılmıştır. Analizler sonucunda en iyi başarımlar ve en hızlı çalışma zamanını belirleyen algoritma tespit edilerek yorumlanmıştır. Çalışmada kullanılan algoritmalar yapay sinir ağları algoritması sınıflandırma başarısı yönünden en yüksek değere sahip olurken, k en yakın komşu algoritması ise çalışma zamanı açısından en hızlı algoritma olarak tespit edilmiştir [9].

Sevli (2019) çalışmasında makine öğrenmesi yöntemi kullanarak kestirim yaparak karar verme sürecine destek vermeyi hedeflemiştir. Bu doğrultuda, dünya çapında yaygın olarak görülen göğüs kanseri teşhisinde performans karşılaştırması yapmıştır. Çalışmada Wisconsin Üniversitesi veri setini ele alarak 30 değişken üzerinden destek vektör makinesi, Naïve Bayes, rastgele orman, k en yakın komşu ve lojistik regresyon algoritmaları kullanarak kestirim yapmıştır. Yapılan analizler sonucunda verilere ilişkin en iyi sınıflama kestirimi yapan algoritmanın lojistik regresyon algoritması olduğu sonucuna varılmıştır [10].

Koronavirüsün ilk olarak Çin'in Wuhan kentinde ortaya çıkması sonrası dünya çapında yayılmış ve Pandemi olarak ilan edilmiştir. Pandeminin sınırlanabilmesi ve kontrol altına alınabilmesi için dünya çapında önlemler alınmıştır. Özen, Saraç ve Koyuncu (2021) tarafından yapılan çalışmada Amerika Birleşik Devletleri'ndeki COVID-19 hasta sayıları ele alınarak, geleceğe yönelik hasta sayısına ilişkin kestirim yapılmaya çalışılmıştır. Bu kapsamda prophet, polinom regresyon, ARIMA, doğrusal regresyon ve random forest olmak üzere beş tane makine öğrenmesi yöntemi algoritması kullanılmıştır. Algoritmaların performans karşılaştırmalarında ortalama mutlak yüzde hatası (MAPE), ortalama karekök sapması (RMSE) ve ortalama mutlak hata (MAE) değerleri kullanılmıştır. En iyi sonuç ortalama mutlak yüzde hatası (MAPE) değeri ile polinom regresyon algoritması kullanıldığı analizde elde edilmiştir [11].

Çağlıyan ve Köse (2021) çalışmalarında epilepsi hastalarının temel tanı yöntemi olarak değerlendirilen EEG sinyallerine ilişkin verileri ele almışlardır. Çalışmada makine öğrenmesi yöntemi algoritmalarından yapay sinir ağları, doğrusal diskriminant analizi ve k-en yakın komşu algoritmalarını kullanmışlardır. Çalışma sonucunda epileptik nöbet tanımlama sınıflamasını ilişkin algoritmaları test ederek, algoritma başarılarını yorumlamışlardır [12].

Bilgin (2021) çalışmasında makine öğrenmesi yöntemi kullanarak diyabete yakalanma olasılığını tahmin etmeye çalışmıştır. Çalışma kapsamında altı tane makine öğrenmesi algoritması kullanarak en iyi sonucu k-en yakın komşu algoritması ile elde etmiştir. Gerçekleştirilen bu çalışmada oluşturulan model bilgisayar kullanıcı ara yüzü ile birleştirilerek diyabet erken tanı kiti geliştirilmiştir [13].

Kanser ve benzeri hastalıklar uzun süreli tedavi gerektirmekte ve oldukça yüksek maliyetli olmaktadır. Kliniğe başvuran hastalar açısından belirtilen bu harcamalar ve zaman nedeniyle hastalığın tedavisinde erken tanı her geçen gün önem kazanmaktadır. Makine öğrenmesi yöntemi hastalığın erken tanısında sıklıkla kullanılmaya başlanmıştır. Kaya ve diğerleri (2019) tarafından yapılan çalışmada kanser ve benzeri hastalık tanısı olan bireylerin klinik verileri ile oluşturulan sınıflandırma modeli ile önceden teşhis yapılabilmesi hedeflenmiştir. Bu doğrultuda ölüm riskinin azaltılması ve hastanın yaşam süresinin uzatılması için makine öğrenmesi algoritmaları ile en iyi kestirim yapılmıştır [14].

Veri madenciliğinin sıklıkla kullanılmaya başlamasıyla yapılan çalışmalarda kestirimde kullanılan özniteliklerin en aza indirgenebilmesi için hedeflenmektedir. Böylelikle yapılan kestirimler için daha spesifik öznitelikler kullanılarak yorumlamalar yapılabilir. Demirarslan ve Suner (2021) tarafından yapılan çalışmada öznitelik seçim yöntemleri kullanılarak makine öğrenmesi yöntemlerinin hesaplama süresini ve maliyetini düşürerek sınıflama performansının iyileştirilmesi amaçlamıştır. Öznitelik seçim yöntemleri sarmal yöntemler, filtreleme yöntemleri ve gömülü yöntemler olmak üzere üç grupta incelenmiştir. Sağlık verileri ele alınarak yapılan bu çalışmada, makine öğrenmesi yöntemi sınıflandırma algoritmalarından Naive Bayes, destek vektör makinesi ve karar ağacı yöntemleri kullanılmıştır. Çalışmada kullanılan veriler UCI ve Kaggle veri tabanlarından elde edilmiştir. Algoritmaların sınıflandırma performanslarını karşılaştırırken doğruluk değerleri hesaplanmış ve yorumlanmıştır. Yapılan karşılaştırma ile en hızlı kestirim yapan

algoritmanın tespit edilmesi hedeflenmiştir. Bulgulara göre, çok sayıda özneliğe sahip verilerin sınıflandırma performansları, öznelik seçimi yapılmış verilere göre daha düşük bulunmuştur. Sonuç olarak ise, algoritmalar arasında zaman bakımından farklılık oluşmadığı ve örneklem büyüklüğüne göre analiz sürecinin değişiklik gösterdiği tespit edilmiştir [15].

Kardiyovasküler hastalıklar, her yıl tahminen 20 milyona yakın can kaybına neden olmaktadır. Bu sebeple kalp hastalıkları için erken tanı ve tedavi önem arz etmektedir. Son yıllarda yapılan çalışmalarda bilgisayar bilimlerinin hızlı yükseliş göstermesiyle makine öğrenmesi ve yapay zekâ gibi yeni teknikler sağlık alanında kullanılmaya başlanmıştır. Coşar ve Deniz (2021) çalışmalarında kalp hastalığı olan bireylerin ölüm oranının yüksek olması nedeniyle, makine öğrenmesi yöntemi kullanarak kalp hastalığına yakalanmadan önceden kestirim yapabilecek bir model oluşturmayı hedeflemiştir. Ele aldıkları değişkenlerle üç tane makine öğrenmesi algoritması kullanarak ileriye dönük kestirim yapmışlardır. Çalışma sonucunda doğruluk oranlarını değerlendirerek kestirim yapılan modele ilişkin kullanılan parametrelerin yeterli olduğu kanısına varmışlardır [16].

Otistik spektrum bozukluklar sosyal ilişkilerde ve bilişsel gelişimde gecikmelere ya da farklılaşmalara yol açmaktadır. Hastalığın teşhisi bireylerin gelişimleri için oldukça önem arz etmektedir. Son yıllarda yapılan araştırmalara göre Dünya genelinde ve Türkiye’de otistik spektrum bozukluklarda ciddi bir artış gözlenmektedir. Özellikle 12-36 ay arasındaki çocuklarda teşhisin konulabilmesi için yardımcı sistemlerinin geliştirilmesi gerektiği düşüncesi ortaya çıkmıştır. Metlek ve Kayaalp (2020) tarafından yapılan çalışmada otistik spektrum bozukluğunu önceden tahmin edebilecek bir yazılım geliştirmiştir. Bu yazılımda makine öğrenmesi yöntemi kullanarak denetimli ve denetimsiz öğrenme algoritmaları ile kestirimler elde etmişlerdir. Çalışma sonucunda iki tür algoritma sonuçlarını karşılaştırarak denetimli öğrenme algoritmalarında daha iyi kestirimler elde edildiğini tespit etmişlerdir [17].

Meme kanseri, meme hücrelerinde başlayan ve sıklıkla kadınlar arasında görülen bir kanser türüdür. Erken teşhis ve doğru tedavi yöntemi ile meme kanseri hastalarının hayatta kalma oranının arttırılabildiği değerlendirilmektedir. Bektaş ve Babur (2022) tarafından yapılan çalışmada makine öğrenmesi yöntemi kullanılarak meme kanseri teşhisine yönelik kestirim yapılmış ve öznelik seçme yöntemleri kullanılarak da meme kanserinde etkin genler belirlenmeye çalışılmıştır. Kullanılan algoritmalarda elde edilen doğruluk değerlerinin

değerlendirilmesi sonucunda meme kanseri teşhisinde kullanılabilecek parametrelerin tespiti yapılarak oluşturulan model yorumlanmıştır [18].

Üretim sektöründe düşük maliyet ile ürün elde edebilme ve kısa sürede ürün geliştirebilme döngüsünün önemi göz ardı edilememektedir. Wei ve diğerleri (2019) yürüttükleri çalışma kapsamında makine öğrenmesi yöntemi ile üretim sektörüne yönelik bir iyileştirme yapmayı hedeflemişlerdir. Çalışmalarında makine öğrenmesi yöntemi ile malzeme özelliklerinin analiz edilebildiği temel uygulama prosedürlerini tartışarak, sektöre uygun iyileştirmeler önermişlerdir [19].

Makine öğrenmesi yöntemi, epidemiyolojik bilimleri dönüştürme potansiyeline sahip bir bilgisayar bilimi dalı olarak tanımlanmakta ve avantajları nedeniyle de çalışmalarda tercih edilmeye başlanılmıştır. Bi, Goodman, Kaminsky ve Lessler (2019) tarafından yaygın olarak kullanılan beş tane makine öğrenmesi yöntemi algoritması kullanılarak çalışma yorumlanmıştır [20].

Sevinç (2021) tarafından yapılan çalışmada periodontoloji alanında geleneksel tedavi yöntemlerinin yerine, kliniğe başvuran her hasta için ayrı hastalık riskini belirleyerek tedavi edilebilmesini hedeflemiştir. Çalışmaya hastaların demografik bulgularına ek olarak klinik muayene sonuçlarını da dahil edilmiştir. Toplam 21 kategorik bağımsız değişken kullanılarak makine öğrenmesi yöntemi kullanarak risk sınıflandırması yapılmış ve yorumlanmıştır. Çalışma sonucunda periodontal hastalık kestirim aracı geliştirilerek bir bilgisayar yazılım programı oluşturulmuştur. Böylelikle kliniğe başvuran hastalara ilişkin bir risk sınıflandırması yapılmakta ve klinisyene tedavi planlamasında yardım edilmektedir [21].

Kaynar, Arslan, Görmez ve Işık (2018) çalışmalarında bilgisayar kullanımında karşılaşılan tehditleri ele almışlardır. Günümüzde internet kullanımında sıklıkla karşılaşılan virüs tehdidini önleyebilmeyi hedefleyerek makine öğrenmesi yöntemiyle öznitelik seçimi yapmışlardır. Çalışmalarında 9 tane öznitelik seçim algoritması ve 3 tane sınıflama algoritması kullanarak en iyi yöntemi belirlemeye çalışmışlardır. Çalışma sonucunda kullanılan öznitelik seçim yöntemlerinin her üç sınıflama yönteminin de başarı oranını arttırdığı ve modellerin daha hızlı çalışmasını sağladığı yorumunu yapmışlardır [22].

Er ve Çiğ (2020) yapmış oldukları çalışmada müziğin dinleyen kişi üzerinde çeşitli duygular uyandırdığı bilgisiyle yola çıkarak beyin sinyallerini incelemişlerdir. Çalışmalarında müzik parçalarını dinlerken bireyin duygularının tanınması hedeflemişlerdir. Çalışma verileri Elektroansefalografi sinyalleri üzerinden elde edilen özneliliklerden oluşmaktadır. Çalışma kapsamında destek vektör makineleri, k-en yakın komşu ve yapay sinir ağlarını makine öğrenmesi algoritmaları kullanılarak sınıflama yapılmıştır. Çalışma sonucunda en iyi sonuç yapay sinir ağları algoritması ile elde edilmiştir [23].

Gültepe (2019) çalışmasında bazı meteorolojik değişkenler kullanarak hava tahmini yapmayı hedeflemiştir. Çalışmasında hava kirliliğinin nedenleri olarak nüfus yoğunluğunun artması, kentleşme ve endüstriyel gelişmelerini belirterek makine öğrenmesi yöntemi ile kestirim yapmıştır. Yapay sinir ağları, rastgele orman, k-en yakın komşu, lojistik regresyon, karar ağacı, lineer regresyon ve basit Bayes algoritmalarının kullanıldığı bu çalışmada en iyi kestirim yapay sinir ağları ile elde edilirken en kötü kestirim ise lojistik regresyon ile elde edilmiştir [24].

Yiğiter ve diğerleri (2018) tarafından yapılan çalışmada makine öğrenmesi yöntemi kullanılarak ülkemizde sukuk ihraç eden Vakıf Portföy şirketi kira sertifikası ile elde edilen fiyatlar üzerinden yapılmış ve günlük fiyat verileri kullanılarak bir modelleme yapılmıştır. Sukuk alternatif bir finansman aracı olup, İslami kriterlere uygun olan bir menkul kıymet olarak tanımlanmaktadır. Bu çalışmada, verilerin 143 tanesi eğitim verisi olarak kullanılmış, 95 tanesi ise test verisi olarak kullanılmıştır. Çalışma sonucunda 1, 3 ve 5 gün ilerisi için yapılan fiyat tahminleri oldukça başarılı sonuçlanmıştır [25].

Yapılan literatür taraması ile yapay zekanın sağlık bilimleri alanında oldukça etkin olduğu değerlendirilmiştir. Çalışma kapsamında oluşturulan istatistik bir model yardımıyla kliniğe başvuran hastaların süre ve maliyet kazanımlarının arttırılabileceği düşüncesiyle yola çıkılmıştır. Dünya Sağlık Örgütü'nün tanımlamış olduğu tanı kriterlerine göre hastaya tarama testi önerilmekte ve mutasyonun varlığı incelenmektedir. Yapılan bu çalışma ile oluşturulacak istatistiksel bir model klinisyene öncelikle taraması gereken mutasyon ile ilgili bilgi verecektir.

Çalışma kapsamında klinisyene yön verebilmesi için oluşturulan model makine öğrenmesi yöntemi kullanılarak elde edilmiştir. Bu yöntem, günümüzde istatistiksel analizlerde yaygın

olarak kullanılmaya başlanmıştır ve araştırmacılara mevcut verilerden tutarlı ve doğru kestirimler yapabilme imkanı sunmaktadır [26].

Makine öğrenmesi yönteminde eğitim ve test verisi olmak üzere iki farklı veri seti oluşturulmaktadır. Eğitim verisi gerçekte karşılaşılmış durumları gösteren verilerin belirli bir oranı seçilerek algoritmalarda kullanılması olarak tanımlanabilirken, test verisine ilişkin gerçek veriler algoritmalarda kullanılmadan makine öğrenmesi yöntemi ile kestirim yapılmaya çalışılmaktadır [27]. Model ile kestirilmeye çalışılan değişkene ilişkin test verisi değerleri ile kestirim değerleri karşılaştırılarak güven düzeyleri saptanmaktadır [28]. Çalışma doğrultusunda, saptanan bu güven düzeyleri ile klinisyene tanı testi önerilebilmektedir.

Yapılan bu çalışmanın amacı, klinik bulgularına göre myeloproliferatif hastalık ön tanısı konulmuş hastaların demografik özelliklerine, kan sonuçlarına ilişkin verileri ele alarak makine öğrenmesi yardımıyla hangi mutasyonun varlığının araştırılması gerektiğini öneren bir modelin oluşturulması olarak belirlenmiştir.

2. GENEL BİLGİLER

Bu bölümde çalışma verilerinin alana yönelik açıklamasına ve çalışmada kullanılan analiz yöntemine ilişkin detaylı bilgi verilmeye çalışılmıştır. Bölümün ilk kısmında myeloproliferatif hastalığın türlerinden ve hastalığa ilişkin araştırılan mutasyonlardan bahsedilmiştir. Bölümün diğer kısmında ise makine öğrenmesi yönteminin tarihçesi, yöntemde kullanılan algoritmalar, yöntemin uygulanışı ve literatürde yer alan kullanım alanlarından bahsedilmiştir.

2.1. Myeloproliferatif Hastalık

Myeloproliferatif hastalıklar, kemik iliğindeki miyeloid öncül hücrelerin aşırı üretimi ile karakterizedir [29]. Kronik myelositer lösemi, polisitemia vera, esansiyel trombositoz ve primer myelofibrozis olmak üzere dört grupta değerlendirilmektedir [4].

Kronik myeloid lösemi, myeloid hücrelerin çoğalması ile oluşan bir kök hücre hastalığıdır [29]. Kemik iliğinde yüksek oranda myeloid hiperplazi ve kanda yüksek oranda lökosit bulunmasına ek olarak dalak büyümesi ile karakterizedir [30]. Kronik faz, akselere faz ve blastik faz olmak üzere üç evre olarak değerlendirilmektedir [31].

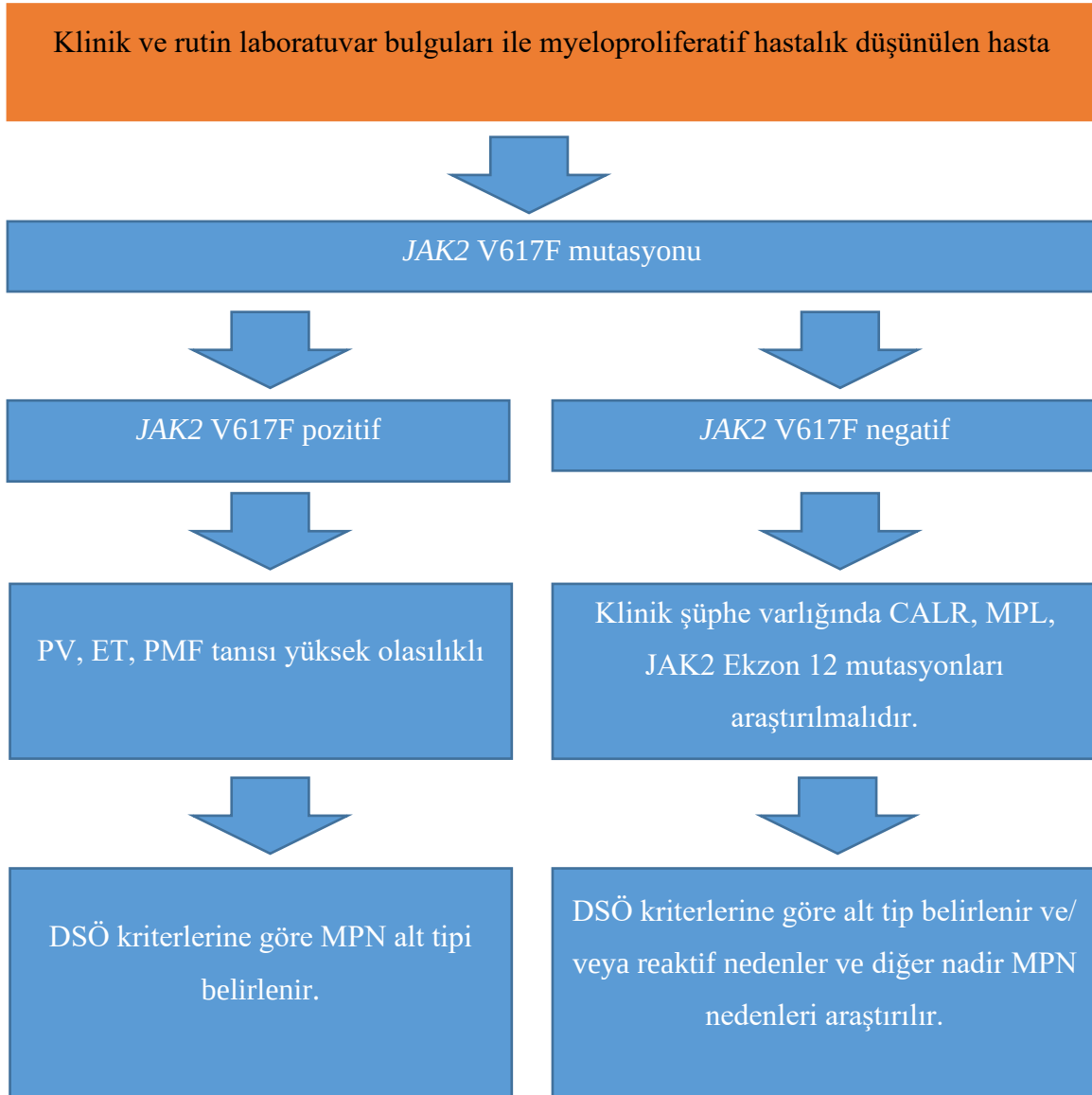
Polisitemia vera, herhangi bir uyaran olmadan kemik iliğinde eritrosit oranının kontrolsüz çoğalması ile karakterize bir hastalıktır [32]. Hastalığın seyrinde kanama eğilimi gözlenmektedir [33]. Hastalığın görülme sıklığı 100 000’de 0,5-2 arasındadır [34].

Esansiyel trombositoz, trombosit sayısının arttığı klonal bir kök hücre hastalığıdır [35]. Görülme sıklığı yıl bazında 100 000 kişide 1-2,5 kadardır [36]. Hastalarda trombosit sayı artışı ile oluşan bozukluklar tromboz veya kanamaya neden olmaktadır [37]. Esansiyel trombositoz ön tanılı hastalarda *JAK2* V617F mutasyonunun varlığının gösterilmesi önerilmektedir. Bu mutasyonun varlığı tanıyı desteklemekte, mutasyon varlığının olmadığı durumlarda ise *MPL* mutasyonunun varlığının araştırılması önerilmektedir [38].

Bir kök hücre hastalığı olan primer myelofibroz, kemik iliğinde fibrozis, splenomegali, sitokin salınımı ile karakterizedir [39]. Nadir bir hastalık olan primer myelofibrozun görülme

sıklığı yılda 100 000’de 0,5-1,5 olarak değerlendirilmektedir [40]. Primer myelofibroz ön tanılı hastalar genellikle halsizlik, kilo kaybı, çarpıntı, karında şişkinlik gibi şikayetler ile kliniğe başvurmaktadır [41].

Myeloproliferatif hastalıkların tanısının kesinleştirilebilmesi için Dünya Sağlık Örgütü tarafından bir tanı şeması oluşturulmuştur. Şekil 2.1’de *JAK2* V617F, *JAK2* Ekzon 12, *MPL* ve *CALR* mutasyon taramalarına ilişkin uygulama şeması yer almaktadır.



DSÖ: Dünya Sağlık Örgütü, PV: Polisitemia vera, ET: Esansiyel trombositemi, PMF: Primer

Şekil 2.1. Myeloproliferatif hastalık ön tanılı hastalara ilişkin tanı testlerinin uygulama yöntemi (Türkiye Hematoloji Derneği, 2016)

Şekil 2.1’de görüldüğü gibi kliniğe başvuran hastaların klinik ve laboratuvar bulguları ışığında klinisyen tarafından hastalara yönelik ön tanı verilmektedir. Bu aşamada klinisyen, Dünya Sağlık Örgütü tanı kriterleri doğrultusunda ön tanılı hastalarda *JAK2* V617F, *JAK2* Ekzon 12, *MPL* ve *CALR* olmak üzere dört farklı hedefte mutasyon varlığını incelemektedir.

JAK2 mutasyonlarının polisitemia vera hastalarının hemen tümünde varlığının gösterilmesi, yine esansiyel trombositoz ve primer myelofibroz tanılarında yaklaşık hastaların yarısında bu mutasyonların gösterilmesi myeloproliferatif hastalık tanı kriterlerinin yenilenmesini sağlamıştır [42]. JAK/STAT yolağı olarak adlandırılan moleküler mekanizma olgun myeloid hücrelerin ve bunların öncüllerinin çoğalması ve farklılaşması sırasında sinyal iletiminde önemli bir rol oynamaktadır. Normal *JAK2* proteininin varlığının sağlıklı hematopoez için gerekli olduğu değerlendirilmektedir [43]. *JAK2* genine ait fonksiyon kazandıran mutasyonlar ilk kez 2005 yılında BCR-ABL negatif myeloproliferatif hastalıkta gösterilmiştir [44]. *JAK2* V617F mutasyonu, genin 14. ekzonunda, yani kromozom 9p24 üzerindeki *JAK2*’nin psödokinaz alanında bulunmakta ve tek bir nükleotit değişimiyle (G1849T) oluşmaktadır [45].

Daha az sıklıkla karşımıza çıkan *JAK2* geni Ekzon 12 mutasyonları tek nükleotit değişimleri, delesyon ya da insersiyonlar şeklinde karşımıza çıkabilmektedir. Polisitemia veralı hastaların %2-3’ünde *JAK2* Ekzon 12 mutasyonlarına rastlanmaktadır [46]. Ekzon 12 ve 14 mutasyonlarında benzer klinik özellikler beklenmektedir [47].

MPL geninde ekzon 10’da yer alan kodon 515 bölgesine ait tek nükleotit değişiklikleri (W515K veya W515L) en yaygın görülen mutasyonlardır. Bunlar *JAK2* sinyal artışına neden olan bir sinyal iletim artışı sağlamaktadır [48].

CALR genine ait mutasyonlar 9. ekzonda yer alır ve iki tipi vardır. Tip 1 mutasyonlarda 52 baz çifti büyüklüğünde delesyon meydana gelirken, Tip 2 mutasyonlarda 5 baz çifti uzunluğunda bir eklenme söz konusudur. Her iki tip mutasyon da çerçeve kaymasına neden olarak sinyal iletimini etkilemektedir [13].

2.2. Makine Öğrenmesi Yöntemi

Makine öğrenmesi yöntemi, çalışma kapsamındaki birimlerin ait oldukları sınıfları kestirebilmek amacıyla günümüzde yaygın olarak kullanılmaya başlanan bir analiz yöntemidir [49]. Makine öğrenmesi yöntemi, yapay zekâ algoritmalarının bir alt boyutu olarak tanımlanmaktadır [50].

Makine öğrenmesi yöntemi teknikleri denetimli öğrenme ve denetimsiz öğrenme olarak iki grupta toplanmaktadır [51]. Denetimli öğrenme teknikleri, önceden gözlemlenmiş ve sonuçları bilinen veriler kullanılarak yeni bir model oluşturmayı, denetimsiz öğrenme teknikleri ise sonuçları bilinmeyen verilere ilişkin bilinmeyen bir yapı bulmayı hedeflemektedir [52].

Bu yöntem ile uygulanan çalışmaların temelinde; araştırmacının elindeki veri setine ilişkin bulgulardan elde edilecek model ile yeni karşılaşılabilecek durumlara ilişkin kestirim yapabilmek bulunmaktadır. Bu süreçte yapay bir öğrenme yöntemi ile makine üzerinden farklı algoritmalar kullanılarak analizler gerçekleştirilmektedir [53].

2.2.1. Tarihçesi

Makine öğrenmesi yöntemi ilk olarak 1950 yıllarında kullanılmaya başlanmıştır. 1950 öncesi yıllarda bu kavram henüz gelişmemekle birlikte, bu yıllara makine öğrenmesi yönteminin temelini oluşturan istatistiksel modellerin ortaya çıktığı ve geliştirildiği dönem olarak adlandırılmaktadır [54].

1950 yılında; İngiliz matematikçi Alan Turing “Computing Machinery and Intelligence” isimli makalesinde makine düşünmesinin gerçekten yapılabileceğini açıklamaya çalışarak, geliştirmiş olduğu Turning testi ile bu konuda bir adım atmıştır. Turning yapmış olduğu bu çalışma ile de ilk defa makine öğrenmesini deneyimleyen araştırmacı olarak bilinmektedir [55].

1950 ve 1960 yılları arasında; makine öğrenmesi ulaşılabilir bir konu olmaya başlayarak dama oyunu ile güncel yaşamımıza girmeye başlamıştır. Bu dönemde, bilgisayar üzerinden ilk uygulamalar başlamış, derin sinir ağları kavramının temelleri atılmıştır [56].

1960 ve 1970 yılları arasında; 1965 yılında Alexey Ivakhnenko ve arkadaşları tarafından grup veri işleme yöntemi kullanılarak, ilk derin öğrenme ağları oluşturulmuştur [57]. 1967 yılında ise, Fix ve Hodges (1989) tarafından k-en yakın komşu algoritması oluşturulmuş [56] ve sonraki yıllarda haritalama alanında kullanılmıştır [58].

1970 ve 1980 yılları arasında; elektrik mühendisi Kohonen ve nöropsikolojist Andersen tarafından çağrışımlı bellek konusunda çalışmalar yapılmış ve denetimsiz öğrenme kuralının temeli atılmıştır [59-60].

1980 ve 1990 yılları arasında; 1982 yılında Kohonen kendi kendine öğrenme sistemi ile haritalama uygulamalarını başlatmış [61] ve Hopfield algoritmalarının genelleştirilebileceğini göstererek, karmaşık problemler için çözümler üretilebileceğini göstermiştir [62]. 1987 yılında Grossberg ve Carpenter tarafından adaptif rezonans teorisi geliştirilerek, denetimsiz öğrenme ile ilgili en karmaşık yapay sinir ağı uygulaması duyurulmuştur [63]. Fukushima tarafından 1986 yılında şekil ve örüntü tanıma amaçlı geliştirilen denetimli öğrenme temelli Neocognitron tanıtılmıştır [64]. 1988 yılında Specht tarafından probabilistik ağlar geliştirilmiştir [65].

1990-2000 yılları arasında; 1991 yılında Specht genel regresyon ağlarını geliştirmiştir [66]. 1993 yılında Alman bilim insanı Schmidhuber derin öğrenme alanında yapmış olduğu ilk çalışmada büyük verilerle analiz yaparak çok katmanlı yapı mantığını geliştirmiştir [67]. 1995 yılında Vapnik tarafından destek vektör makineleri geliştirilmiştir [68]. 1997 yılında IBM tarafından Deep Blue isimli bilgisayar geliştirmiş ve bu bilgisayar satranç oyununda dünya şampiyonunu yenmiştir [69].

2000-2010 yılları arasında; 2005 yılında, metinleri, resim ve videolardaki nesneleri algılayıp ayırt edebilen derin öğrenme metodlarının kullanıldığı Kinect adlı hareket algılama cihazı PrimeSense tarafından geliştirilerek, Microsoft firması tarafından lisanslanmıştır [70].

2010 yılından sonra, nesneleri sınıflama ve tanımla işlemi yapabilen Google Brain yazılımı [71], Facebook üzerinden fotoğraftaki bireyleri tanımlayabilen ve doğrulayabilen Deep Face yazılımı [72], satranç benzeri bir Çin oyununun oynanabilmesi için Deep Mind yazılımı ve yine oyunlarda kazanç sağlayabilmek için Alpha Go yazılımı [73] geliştirilmiştir.

2.2.2. Uygulama adımları

Makine öğrenmesi yönteminde hipotez oluşturulduktan sonra uygulanması gereken süreçler bulunmaktadır [74]. Bu süreçler genel olarak; verilerin toplanması, verilerin düzenlenmesi, algoritmaların denenmesi ve algoritmaların karşılaştırılarak yorumlanması olarak sıralanabilmektedir.

Verinin toplanması

Bir çalışmaya ilişkin planlama sürecinde kurulan hipotezler doğrultusunda öncelikle kullanılacak değişkenlerin belirlenmesi gerekmektedir. Bu aşama oldukça önemli olup, belirlenen değişkenler çalışma kapsamında belirlenen hedeflere ulaşabilir nitelikte olduğu durumlarda araştırma başarı ile tamamlanabilmektedir. Belirlenen değişkenlere ilişkin verilerin toplanması aşamasında gözlem, odak grup veya anket yöntemleri kullanılmaktadır [75].

Veri setinin hazırlanması

Çalışma kapsamında toplanan verilerin istatistiksel analizinin yapılacağı programa göre hazırlanması gerekmektedir. Bu kapsamda, IBM Statistics SPSS, R programlama dili, Python, AMOS, Lisrel veya MetaWin gibi paket programlar kullanılmaktadır. Çalışmada ele alınan değişkenler, kullanılan program gereklilikleri doğrultusunda nitel-nicel, kesikli-sürekli veya bağımlı-bağımsız gibi veri özelliklerini belirtecek şekilde hazırlanmalıdır.

Veri temizleme ve birleştirme

Programa girişi yapılan veri setindeki her bir değişken kendi özelliğine uygun bir şekilde hazırlanmalıdır. Yanlış veri girişlerinin düzeltilmesi bu aşamada gerçekleşmektedir. Veri toplama aşamasında eksik verilerin tamamlanması veya çalışmadan çıkarılması, giriş

hatasından kaynaklı veri düzenlemeleri, yapılacak analizlere uygun olmayan değişkenlerin tanımlanması, çalışma kapsamında oluşan boyutlara göre verilerin birleştirilmesi gibi birçok düzenleme işlemi bu aşamada yapılmaktadır.

Modelin eğitilmesi

Makine öğrenmesi yönteminin temel çalışma prensibinin uygulandığı aşama modelin eğitilmesi adımıdır. Makine öğrenmesinde yapılacak sınıflama için algoritmalar kullanılmaktadır. Bu aşamada modele özgü tanımlanan bağımlı değişkene ilişkin verilerin bir bölümü modele tanıtılarak eğitim verisi olarak değerlendirilir, diğer kısmı ise test verisi olarak algoritmalarla oluşturulan model tarafından kestirilmeye çalışılmaktadır [76].

Modelin değerlendirilmesi

Modelin eğitilmesi adımı uygulanırken algoritmalarla test verisi ile elde edilen sonuçlar, gerçek durumları ile karşılaştırılmaktadır. Bu karşılaştırma yapılırken karışıklık matrisinden yararlanılmaktadır. Karşılaştırma sonucunda doğruluk, duyarlılık, güvenilirlik, F-ölçütü gibi değerler hesaplanarak algoritmaların performansı karşılaştırılmaktadır [77].

2.2.3. Kullanılan algoritmalar

Makine öğrenmesi yönteminde kullanılan algoritmalar denetimli ve denetimsiz öğrenme olarak ikiye ayrılmaktadır. Literatürde en çok kullanılan denetimli makine öğrenmesi algoritması olarak doğrusal regresyon, lojistik regresyon, karar ağaçları ve Naive Bayes algoritmaları başta gelmektedir. Denetimsiz makine öğrenmesi algoritmaları arasında ise k en yakın komşu, temel bileşenler analizi ve dağılmış stokastik komşu gibi algoritmalar sıklıkla kullanılmaktadır. Çalışma kapsamında makine öğrenmesi yönteminde kullanılan sınıflandırma algoritmaları ile değerlendirmeler yapılmıştır. Bu bölümde çalışma kapsamında kullanılan sınıflandırma algoritmaları detaylı bir şekilde açıklanmaya çalışılmıştır.

Naive Bayes sınıflandırma algoritması

Naive Bayes sınıflandırma algoritması, en iyi bilinen ve en çok kullanılan sınıflandırma algoritmasıdır [78] ve basit olasılıksal sınıflandırıcı olarak istatistiksel sınıflandırma teknikleri arasında yer almaktadır. Bu sınıflandırıcının en büyük avantajı, eğitim için az sayıda niteliğe ihtiyaç duymasındır [79]. Bu algoritmanın çok terimli Naive Bayes, Bernoulli Bayes ve Gaussian Naive Bayes olmak üzere üç türü bulunmaktadır [80].

Koşullu bağımsızlık varsayımı ile kurgulanan bu algoritmaya göre [81] bir sonucun çıkma olasılığı, o sonucu etkileyen tüm değerlerin sonucu sağlama olasılıklarının çarpımı ile elde edilmektedir. Elde edilen bu çarpım sonucu üzerinden sınıflamaya ilişkin yorumlama yapılmaktadır [82].

Naive Bayes sınıflandırma algoritmasında sınıf değişkeni (C_i)'ye ilişkin bir koşul oluşturulmaktadır ve Eş. 2.1'de gösterilmiş olan formül yardımı ile olasılıklar hesaplanmaktadır.

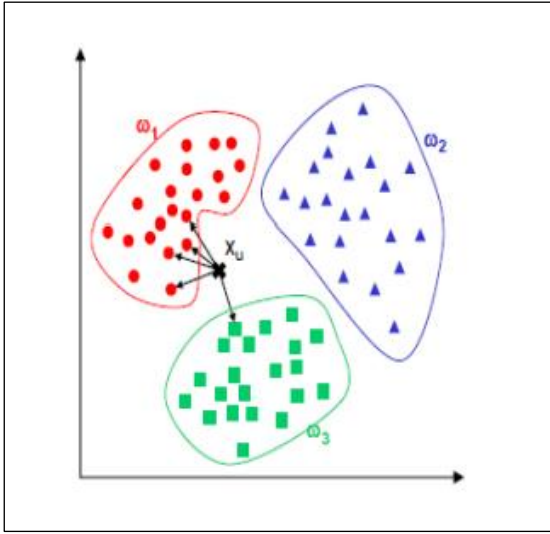
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.1)$$

C_i : Sınıf değişkeni

Olasılıklar hesaplandıktan sonra, veriye ilişkin sınıflandırma için, $P(X|C_i)P(C_i)$ çarpım değerinin en yüksek olduğu sınıf belirlenerek o sınıfa atama yapılır [83].

k-en yakın komşu sınıflandırma algoritması

k-en yakın komşu sınıflandırma algoritması, 1968 yılında Cover ve Hart tarafından geliştirilmiş [84] denetimli bir sınıflandırma algoritmasıdır [85]. Bu algoritmanın, diğer algoritmalara göre uygulanabilirliğinin daha kolay olmasının yanı sıra, diğer algoritmalarından farklı olan özelliği ise eğitim verisine ihtiyaç duymayabilmesidir. Dolayısıyla, işleyiş aşamasında birbirine yakın olan nesneleri aynı sınıf altında toplayarak sınıflandırma yapmaktadır [86]. Algoritmanın çalışma prensibini gösteren grafik Şekil 2.2'de verilmiştir.



Şekil 2.2. k-en yakın komşu algoritması ile sınıflandırma şeması (Türk, 2020)

k-en yakın komşu algoritmasında sınıfı bilinen örnekler üzerinden sınıflama yapıldığı bilgisi ışığında, kestirilen özellikler tek tek işleme alınır. Bu aşamada sınıflandırma yapılabilmesi için eğitim kümesinde yer alan ve o örneğe en yakın k adet örnek seçilir. Daha sonra, seçilen k adet örnekten oluşan küme içerisinde en fazla hangi sınıfa ait örnek varsa kestirim yapılacak örnek bu sınıfa atanmaktadır [87].

Bu algorithmada uygulanan sınıflama işlemlerine ilişkin uzaklık hesaplamasında Eş. 2.2’de verilen Öklid (Euclidean) uzaklığı yöntemi kullanılmaktadır [88].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (dizi_r(x_i) - dizi_r(x_j))^2} \quad i \neq j \quad (2.2)$$

k-en yakın komşu algoritmasının kapsamında belirtilen özellikler göz önüne alındığında iki tane olumsuz özellik ortaya çıkmaktadır. Birincisi, algoritma mantığı test verisinin bilgilerinden çok kendisinin yaptığı hesaplamalara göre çalıştığı için büyük örneklerde çok fazla zaman gerektirmesidir. İkincisi ise, bir örneğe ait mesafe o örneğin tüm özellikleri göz önüne alınarak hesaplandığı için özellik sayısı arttıkça durumlarda mesafe etkilenecek ve sınıflamanın doğruluk değerinin azalacak olmasıdır [89].

Doğrusal diskriminant analizi sınıflandırma algoritması

Doğrusal diskriminant analizi 1930’lu yıllarda Fisher tarafından geliştirilen çok değişkenli istatistiksel yöntemlerden biridir [90]. Algoritma karmaşık yapıdaki veri kümesini yine o kümenin özelliklerini taşıyan daha küçük boyutlu kümelerin yer aldığı bir uzaya yansıtarak sınıflandırma yapmaktadır [91]. Bu doğrultuda algoritmanın temel amacı, verilerin sınıflandırılmasında gerekli olan en ayırt edici özellikleri belirleyip, ayırt edici olmayan özellikleri elemesi olarak değerlendirilebilmektedir. Böylelikle algoritma kullanılarak sınıflandırma işlemi daha kolay yapılabilir.

Doğrusal diskriminant analizinin kullanım amaçları;

- Bağımlı verileri en iyi sınıflandıran tahminlemeyi sınıflar arası farkı en büyük olacak şekilde yapmak,
- Bağımlı ve bağımsız değişkenler arasındaki farkı test etmek,
- Gruplar arası farkı en iyi belirleyen bağımsız değişkeni bulmak,
- Sınıflandırmanın doğruluğunu test etmek,
- Bağımlı değişkenin ne kadarının bağımsız değişkenler tarafından açıklandığını belirlemek olarak sıralanabilmektedir [92].

Doğrusal diskriminant analizinin ölçüsü, belirlenen sınıf merkezleri arası uzaklığı en büyük yaparken, sınıf içi yayılımı en az yapmaktır [93]. Bu yöntem, doğrusal bir sınıflandırıcı olarak analizlerde kullanılmasının yanı sıra, daha karmaşık ve doğrusal olmayan sınıflandırmalarda bir ön işlemci olarak da kullanılabilmektedir.

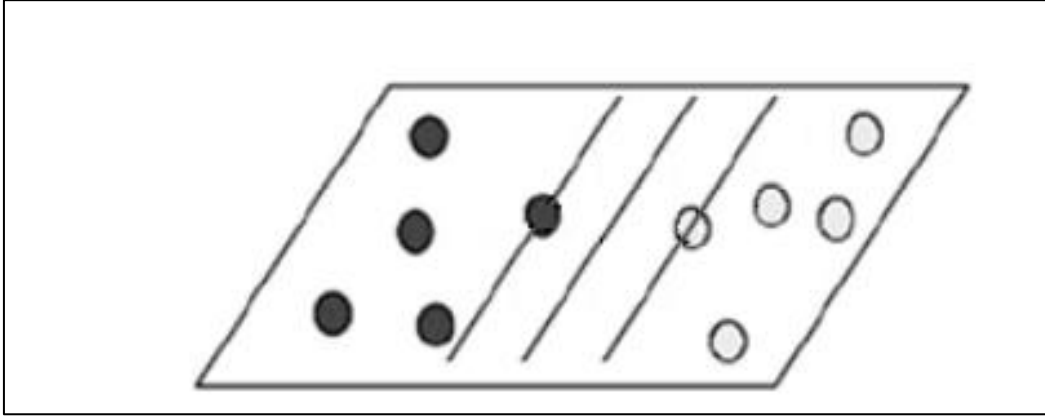
Destek vektör makineleri sınıflandırma algoritması

Destek vektör makineleri 1995 yılında Vapnik tarafından geliştirilen ve güçlü istatistik teorilerine dayanan bir yöntemdir [94]. Algoritmanın temel varsayımı; çalışmaya alınan verinin bağımsız ve aynı dağılıma sahip olması olarak tanımlanmaktadır [95].

Bu yöntemin işleyişinde, bir optimizasyon problemi çözülerek hem doğrusal hem de doğrusal olmayan bir yapıyla sınıflandırma yapabilme yeteneği bulunmaktadır [96]. Veri seti üzerinde yapılan gerekli hesaplamalarla birlikte bir ayırıcı oluşturulmaktadır. Bu kapsamda,

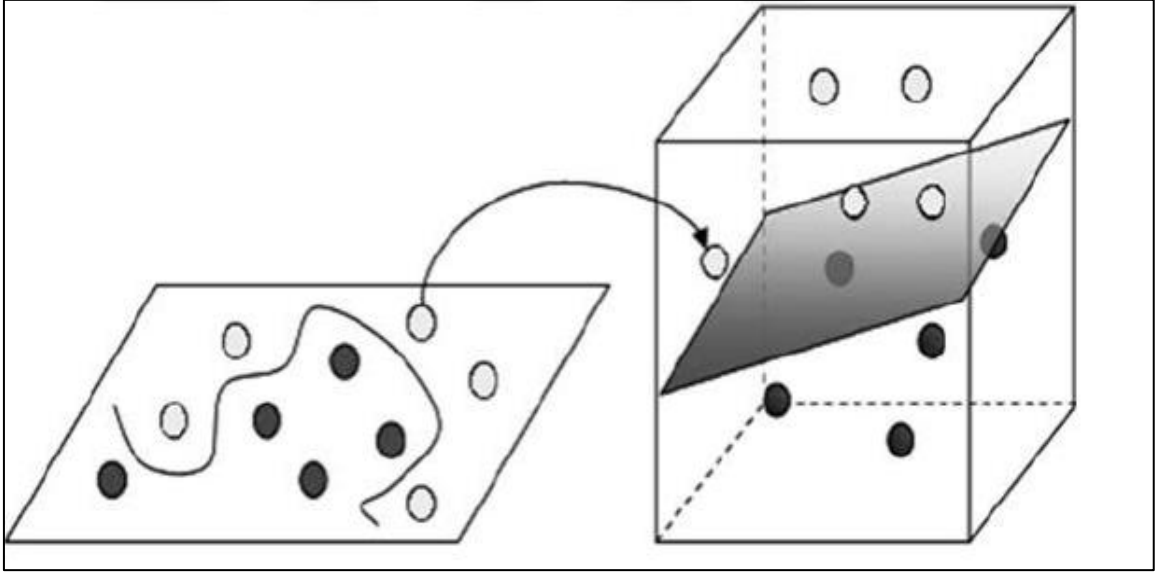
destek vektör makineleri algoritması ile seçilmiş bir veri kümesi için sınıfları ayırabilen sonsuz sayıda doğru oluşturulabileceği bilinmektedir. Sınıflar oluşturulurken en iyi ayırıcı çizginin seçimi, destek vektörleri olarak tanımlanan eğitim numunelerinin alt kümesi kullanılarak gerçekleştirilmektedir. Destek vektörleri arasındaki en büyük sınır uzaklığı, en iyi karar sınırı olarak tanımlanmaktadır. Yöntemin en önemli özelliği, oluşturulan bu ayraç ve sınıflar arası mesafenin eşit uzaklıklarda olmasıdır [97].

Test verisi ile kestirim yapılan veriye ilişkin sınıf atamaları, veriyi birbirinden ayırabilen bir sınıf ayırıcısı ile başlamaktadır. Doğrusal ayrılabilme durumunda Şekil 2.3'te gösterildiği gibi bir yapıyla hareket edilmektedir [98].



Şekil 2.3. Destek vektör makinesi yönteminde doğrusal ayırıcı ile sınıflandırma şeması (Gillmore ve diğerleri, 1979)

Sınıflandırma ayırıcısının doğrusal olmadığı durumlarda Kernel işlevinden yola çıkarak veri setinde örtülü bir dönüşüm kullanılmaktadır. Bu durumda, doğrusal bir düzlem kullanılarak, doğrusal olmayan ayrılabilir bir destek vektörü yapısı ile sınıflandırma yapılabilmektedir [99]. Şekil 2.4'te doğrusal olmayan bir ayırıcı ile yapılan sınıflandırma işlemine ilişkin bir örnek gösterilmektedir.



Şekil 2.4. Destek vektör makinesi yönteminde doğrusal olmayan ayrıcı ile sınıflandırma şeması (Gillmore ve diğerleri, 1979)

Yapılan çalışmaların birçoğunda sınıflandırma performansının en iyi olabilmesi için probleme uygun bir Kernel genişliğinin seçilmesi ve algoritmanın bu doğrultuda yürütülmesi gerekmektedir [97]. Bu algoritma için kullanılan dönüşüm fonksiyonları bulunmaktadır. Uygulamada en sık kullanılan dönüşüm fonksiyonları ve bu fonksiyonların matematiksel denklemleri Eş. 2.3, Eş. 2.4 ve Eş. 2.5'te verilmiştir.

Doğrusal (Linear) fonksiyon

$$K(x_i, x_j) = x_i^T x_j \quad (2.3)$$

Polinomiyal (Polynomial) fonksiyon

$$K(x_i, x_j) = (1 + x_i^T x_j)^p \quad (2.4)$$

Radyal tabanlı (Gaussian/radial-basis) fonksiyon

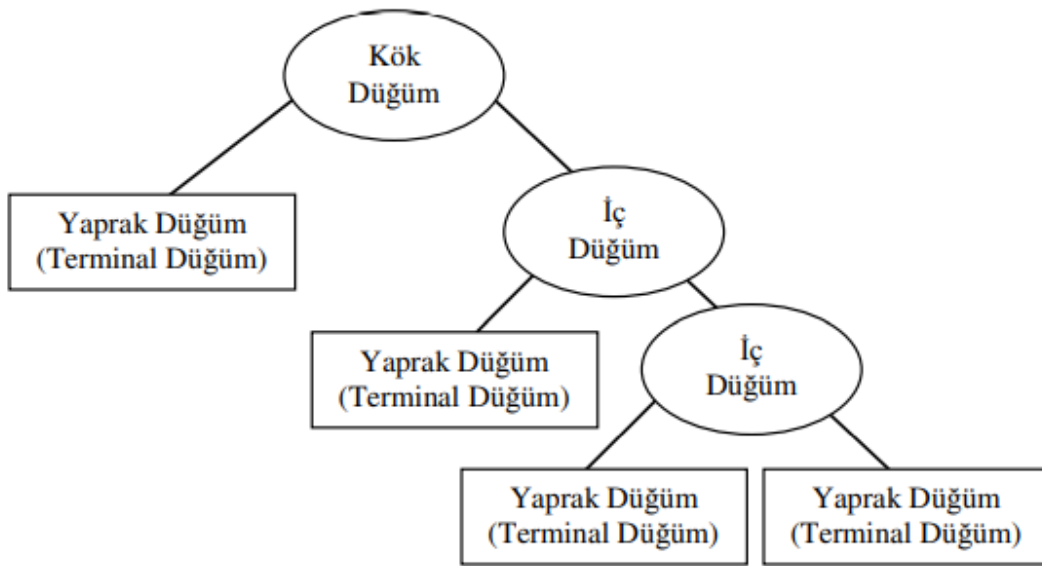
$$K(x_i, x_j) = \exp \left(-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \quad (2.5)$$

Karar ağacı sınıflandırma algoritması

Karar ağacı algoritmasının ilk oluşumu Von Neumann ve Morgenstern tarafından oyunların genişletilmesine yönelik yaptıkları çalışma ile başlamıştır [100]. Daha sonra ise, 1984 yılında Breiman ve diğerleri tarafından karar ağacı algoritması ilk defa istatistiksel bir analiz olarak

kullanılmıştır [101]. Makine öğrenmesi alanında bu algoritmanın kullanıldığı ilk çalışma ise Quainlan tarafından 1986 yılında yapılmış olan çalışmadır [102].

Temel olarak tanımlanan karar ağacı algoritmasından yola çıkılarak, “Random Forest” ve “Gradient Boosted Tree” gibi farklı algoritmalar türetilmiştir [103]. Algoritma, anlaşılabilir ve yorumlanabilir özellikleri ile hem sınıflandırılmış hem de sayısal verilerin işlenmesinde kullanılabilmektedir.



Şekil 2.5. Karar ağacı algoritması ile sınıflandırma şeması (Türk, 2020)

Bu algorithmada Şekil 2.5’teki gibi bir ağaç yapısı oluşturulmaktadır. Ağacın yaprakları seviyesinde oluşturulan sınıf etiketleri ile başlangıçtan çıkan ve bu yapraklara giden özellikler üzerindeki işlemler ifade edilmektedir [104].

Karar ağacı algoritmasının avantajları; kolay modellenenebilir olması, veri işleme kolaylığının olması ve algoritmalar ile elde edilen modelin yorumlama rahatlığının olması olarak sıralanabilirken, dezavantajları yüksek sayıda veriye ihtiyaç duyması ve en iyi ağaç yapısının bulunamama olasılığının olması olarak görülmektedir [105].

Gini algoritması

1984 yılında Breiman tarafından geliştirilen gini algoritmasında gini indeksine göre sınıflandırma yapılmaktadır [106]. Algoritma kapsamında hesaplanan en küçük gini değerine karşılık gelen sınıf test verisine atanarak karar ağacı oluşturulmaktadır [107].

Gini algoritmasının ilk adımı sağ ve sol taraftaki gini değerlerinin Eş. 2.6 ve Eş. 2.7’de yer aldığı gibi hesaplanmasıyla başlamaktadır.

$$Gini_{sol} = 1 - \sum_{i=1}^k \left[\frac{L_i}{|T_l|} \right]^2 \quad (2.6)$$

$$Gini_{sağ} = 1 - \sum_{i=1}^k \left[\frac{L_i}{|T_r|} \right]^2 \quad (2.7)$$

k: Sınıf sayısı

T: Bir düğümdeki örnek sayısı

T_{sol} : Sol sütundaki örneklerin sayısı

$T_{sağ}$: Sağ sütundaki örneklerin sayısı

L_i : Sol sütundaki i kategorisindeki örneklerin sayısı

R_i : Sağ sütundaki i kategorisindeki örneklerin sayısı

Algoritmanın sonraki adımında gini indeksi Eş. 2.8 yardımıyla hesaplanır.

$$Gini_j = \frac{1}{n} (|T_{sol}| Gini_{sol} + |T_{sağ}| Gini_{sağ}) \quad (2.8)$$

n: örnek sayısı

T_{sol} : Sol sütundaki örneklerin sayısı

$T_{sağ}$: Sağ sütundaki örneklerin sayısı

Algoritmanın son adımında j tane nitelik için hesaplanan gini indeksi değerleri arasından en küçük olan değer seçilerek, bölünme işlemi yapılmaktadır. Yapılan bu işlemin sonrasında bir sonraki bölünmeler için ilk adım tekrarlanarak işlem basamakları tamamlanmaktadır [107].

C5.0 algoritması

C5.0 sınıflandırma algoritması karar ağacı yöntemi algoritmalarından birisidir. Quinlan tarafından 1993 yılında geliştirilen bu algoritma, ID3 ve C4.5 algoritmalarının geliştirilmesi sonucu ortaya çıkmıştır [105].

Karar ağacı yöntemindeki ters ağaç yapısı bu algoritmanın da temel çalışma prensibi olmasının yanı sıra, bu algoritmanın bellek kullanımı ve çalışma süresi yönünden C4.5'e göre daha iyi olduğu değerlendirilmiştir. C5.0 algoritmasında karar kurallarını üretmek ve karar ağaçlarını en aza indirmek daha mümkün olmaktadır [108].

C5.0 algoritması ile ilgili ilk kestirimler kategorik veriler üzerinden yapılmaktayken [104], sonraki yıllarda sürekli veriler üzerine de çalışmalar yapılmıştır [105].

Algoritmanın çalışma adımlarında öncelikle veri kümesinin beklenen değeri hesaplanır ve bu hesaplamalar için Eş. 2.9'da yer alan eşitlik kullanılır.

$$Entropi = H(T) = -\sum_{i=1}^k p_i \log_2(p_i) \quad (2.9)$$

k Sol Örnek sayısı

p_i : i . Değişkene ait olasılık

Sonraki adıma geçildiğinde oluşturulan bölümlemelere yönelik bir beklenen değer hesaplanır ve bu hesaplamalar için Eş. 2.10'da yer alan eşitlik kullanılır.

$$H_S(T) = -\sum_{i=1}^k p_i H_S(T_i) \quad (2.11)$$

T_i : i . hedef değişken

Algoritmanın son adımında Eş. 2.12'de belirtildiği gibi, elde edilen beklenen değerler arasındaki fark hesaplanarak, en yüksek kazanımı belirlenmeye çalışılır [105].

$$Bilgi Kazancı(S) = H(T) - H_S(T) \quad (2.12)$$

2.2.4. Algoritmaların karşılaştırılması

Makine öğrenmesi yönteminin son adımında tahminlemeye yönelik oluşturulan modellerin birbiriyle karşılaştırılması ve en iyi modelin seçilmesi gerekmektedir. Yöntemde kullanılan algoritmalar sonucunda elde edilen kestirim değerlerinin birbirleriyle karşılaştırılabilmesi için karışıklık matrisinden yararlanılmaktadır. Bu matriste değişkenlerin gerçek değerleri ile yapılan analizler sonucunda elde edilen kestirim değerleri karşılaştırılmaktadır. Yapılan karşılaştırma ile kestirime ait “Doğru Pozitif”, “Doğru Negatif”, “Yanlış Negatif” ve “Yanlış Pozitif” değerleri elde edilmektedir. Elde edilen değerler Eş. 2.13, Eş. 2.14, Eş. 2.15,

Eş. 2.16 ve Eş. 2.17’deki hesaplamalar yardımıyla sırasıyla “Doğruluk”, “Duyarlılık”, “Özgüllük”, “Kesinlik” ve “F-ölçütü” olarak yorumlanmaktadır [109].

Doğruluk (Accuracy)

Doğru kestirimlerin yüzdesi olarak tanımlanan doğruluk değeri [111], makine öğrenmesi yönteminde kullanılan eğitim verisi yardımıyla kestirim yapılan test verisinin gerçek sonuç verme oranıdır [110] ve Eş.2.13 yardımıyla hesaplanmaktadır.

$$\text{Doğruluk} = \frac{\text{Doğru Pozitif} + \text{Doğru Negatif}}{\text{Doğru Pozitif} + \text{Doğru Negatif} + \text{Yanlış Pozitif} + \text{Yanlış Negatif}} \quad (2.13)$$

Duyarlılık (Sensitivity)

Doğru pozitif oranı olarak da tanımlanan duyarlılık değeri, yapılan kestirim sonucu gerçekte pozitif sınıfa ait olan verilerin doğru olarak tahmin edilme oranını vermektedir [110] ve Eş. 2.14 yardımıyla hesaplanmaktadır.

$$\text{Duyarlılık} = \frac{\text{Doğru Pozitif}}{\text{Doğru Pozitif} + \text{Yanlış Negatif}} \quad (2.14)$$

Özgüllük (Spesitivity)

Negatif belirleyicilik değeri olarak da tanımlanan özgüllük değeri, kestirim yapılan test verisindeki doğru kestirim yapılan negatif sınıf verilerinin oranını vermektedir [111] ve Eş. 2.15 yardımıyla hesaplanmaktadır.

$$\text{Özgüllük} = \frac{\text{Doğru Negatif}}{\text{Yanlış Pozitif} + \text{Doğru Negatif}} \quad (2.15)$$

Kesinlik (Precision)

Bu değer, çalışmada test verisinden ve eğitim verisinden elde edilen pozitif değerlerinin ne kadarının doğru olarak kestirim yapıldığını belirtmektedir [112] ve Eş. 2.16 ile hesaplanmaktadır.

$$Kesinlik = \frac{Doğru Pozitif}{Doğru Pozitif + Yanlış Pozitif} \quad (2.16)$$

F ölçütü

F ölçütü, elde edilen kestirim için daha kesin bir ifade kullanılabilmek amacıyla hesaplanan bir değerdir [113] ve Eş. 2.17’de verildiği gibi kesinlik ve duyarlılığın harmonik ortalaması alınarak elde edilmektedir.

$$F - ölçütü = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \quad (2.17)$$

3. MATERİYAL VE METOT

Çalışma kapsamında ele alınan hasta verilerine ilişkin bilgiler, açıklamalar, elde edilme sürecine ilişkin detaylara bu bölümde yer verilmiştir.

3.1. Çalışmanın Örneklemi

Yapılan bu çalışma kapsamında hasta verilerinin kullanılabilmesine ilişkin öncelikle 24.07.2019 tarihli 19/80 sayılı proje ile Başkent Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurul onayı alınmıştır. Çalışma Helsinki Deklarasyon Prensipleri göz önüne alınarak yürütülmüştür [115].

03.01.2012 ile 11.03.2019 tarihleri arasında Başkent Üniversitesi Ankara Hastanesi Tıbbi Genetik Ana Bilim Dalı Genetik Hastalıklar Merkezi'ne yönlendirilmiş myeloproliferatif hastalık ön tanılı hastalar örnekleme oluşturulmuştur. Çalışma kapsamında toplam 1284 hastanın klinik sonuçları üzerinden analizler gerçekleştirilmiştir.

Çalışma kapsamında tek merkez üzerinden veri toplanması ve incelenen mutasyonların hastanede ilk defa araştırılmaya başlama zamanlarının farklı olması nedeniyle çalışma kapsamın kısıtlar oluşmuştur.

3.2. Çalışmanın Yöntemi

Değişkenlere ait tanımlayıcı istatistikler kategorik değişkenler için frekans ve yüzde ile ifade edilirken, sürekli değişkenler için ortalama ve standart sapma ile ifade edilmiştir. Çalışma kapsamında kullanılan algoritmalara ilişkin kestirimler ve ortalamalar arası farkların araştırıldığı istatistiksel analizler R programlama dili kullanılarak yapılmıştır.

3.3. Çalışmada Kullanılan Değişkenlere İlişkin Açıklamalar

Kliniğe başvuran hastaların öncelikle klinik bulgularının değerlendirilmesi ile tanı sürecine başlanmaktadır. Klinisyen hasta bulguları ile ele alınan hastalık türlerine ilişkin Dünya Sağlık Örgütü'nün belirlemiş olduğu kriterleri göz önüne alarak hastalık ön tanısını

belirleyebilmektedir. Çizelge 3.1’de yer alan klinik bulgular hastane sisteminden elde edilerek çalışmaya başlanmıştır. Klinik bulgularla ilgili değişkenlerin açıklamaları Ek.1’de verilmiştir.

Çizelge 3.1. Çalışmada kullanılan değişkenler

Cinsiyet	AFP	PSA Oranı	MCH	Nötrofil
Yaş	CEA	Hemoglobin	MCV	Lenfosit
Eritropoietin	CA15	Hematokrit	MCHC	Monosit
Ferritin	CA19	Eritrosit	RDW	Eozinofil
Homosistein	PSA	Lökosit	PDW	Bazofil
Paratiroid	Serbest PSA	Trombosit	MPV	Tetkik

4. BULGULAR

Çalışma kapsamında ele alınan hasta verilerinin modellemesine ilişkin analiz sonuçlarına bu bölümde yer verilmiştir. Çalışma kapsamında elde edilen tanımlayıcı istatistikler ve makine öğrenmesi yöntemi kapsamında kullanılan algoritmalara ilişkin analiz sonuçlarına, R programlama dilinde uygulanan kodlara ve makine öğrenmesi yöntemi algoritmalarında elde edilen tüm sonuçlar çizelgelerle özetlenmiş ve yorumlanmıştır.

4.1. Çalışma Verilerine İlişkin R Programlama Dilinde Yapılan İşlemler

Makine öğrenmesi yöntemi aşamaları doğrultusunda veri setindeki değişkenlerin analizlere uygun hale getirilmesi için yapılan işlemler ve kullanılan program kodlamaları aşağıdaki gibi verilmiştir.

Adım 1. Veri seti programa tanıtılır,

```
R> install.packages("readxl")
R> library(readxl)
R> gozr <- read_excel("C:\\Users\\Lenovo\\Desktop\\tez\\yanaliz\\yver.xlsx")
R> ion <- gozr
R> ion=as.data.frame(ion)
R> str(ion)
```

Adım 2. Bağımlı değişken olan tetkik değişkeni için a: JAK2 V617F, b: JAK2 Ekzon 12, c: CALR ve d: MPL olmak üzere sınıf atamaları yapılır,

```
R> for (i in 1:1284) {
R> if (ion[i,30]==1){
R> ion[i,30]= "a"
R> }else if(ion[i,30]==2){
R> ion[i,30]= "b"
R> }else if(ion[i,30]==3){
R> ion[i,30]= "c"
```

```
R> }else if (ion[i,30]==4){
R> ion[i,30]= "d"
R> }
R> }
R> str(ion)
```

Adım 3. Bağımlı değişkenin faktör olarak tanımlama işlemi yapılır,

```
R> ion$tetkik<-as.factor(ion$tetkik)
R> str(ion)
```

Adım 4. Bağımsız değişkenlerden sayısal olarak tanımlanmayanlara ilişkin tanımlama işlemi yapılır,

```
R> ion$Eritropoietin <- as.numeric(as.character(ion$Eritropoietin))
R> ion$Ferritin <- as.numeric(as.character(ion$Ferritin))
R> ion$Paratiriot <- as.numeric(as.character(ion$Paratiriot))
R> ion$AFP <- as.numeric(as.character(ion$AFP))
R> ion$RDW <- as.numeric(as.character(ion$RDW))
R> ion$PDW <- as.numeric(as.character(ion$PDW))
R> ion$monosit <- as.numeric(as.character(ion$monosit))
R> ion$eoziñofil <- as.numeric(as.character(ion$eoziñofil))
R> ion$bazofil <- as.numeric(as.character(ion$bazofil))
R> str(ion)
```

Adım 5. Çalışma kapsamında 1284 hastadan toplanan verilere ilişkin bağımsız değişkenlerdeki eksik verilerin tespiti yapılır, çalışma kapsamında ele alınan değişkenlere ilişkin bulgular Çizelge 4.1’de verilmiştir.

```
R> colSums(is.na(ion))
R> str(ion)
```


Çizelge 4.1. Çalışmada kullanılan değişkenlerde bulunan eksik veri durumunun dağılımı

Değişken	Eksik Veri Sayısı	Değişken	Eksik Veri Sayısı	Değişken	Eksik Veri Sayısı	Değişken	Eksik Veri Sayısı
Cinsiyet	0	CA15	1279	Lökosit	50	Nötrofil	50
Yaş	721	CA19	1277	Trombosit	51	Lenfosit	50
Eritropoietin	837	PSA	1264	MCH	50	Monosit	56
Ferritin	568	Serbest PSA	1273	MCV	50	Eozinofil	56
Homosistein	1219	PSA Oranı	1273	MCHC	50	Bazofil	56
Paratiroid	1269	Hemoglobin	57	RDW	51	Tetkik	0
AFP	1253	Hematokrit	50	PDW	451		
CEA	1269	Eritrosit	51	MPV	90		

Çizelge 4.1’de yer aldığı gibi homosistein, paratiroid, AFP, CEA, CA15, CA19, PSA, serbest PSA ve PSA oranı değişkenlerinde eksik veri oranının oldukça yüksek olduğu tespit edilmiştir. Bu durumun sebebi kliniğe başvuran hastaların genelinden eksik veri oranı az olan diğer klinik bulguların mutlaka istenmesi ve eksik verilerin çok olduğu klinik bulguların ise her hastadan istenmemesinden kaynaklanmaktadır.

Adım 6. Eksik verilere sahip sürekli değişkenlerde değişken ortalaması değişmeden atama işlemi yapılır,

```
R> install.packages("mice")
R> library(mice)
R> tempData <- mice(ion, m=5,maxit=50,meth='pmm',seed=500)
R> summary(tempData)
R> completedData <- complete(tempData,1)
R> completedData$Paratiroid[is.na(completedData$Paratiroid)] <- mean(completedData$
Paratiroid, na.rm = TRUE)
R> completedData$AFP[is.na(completedData$AFP)] <- mean(completedData$AFP,
na.rm = TRUE)
R> completedData$CEA[is.na(completedData$CEA)] <- mean(completedData$CEA,
na.rm = TRUE)
R> completedData$CA15[is.na(completedData$CA15)] <- mean(completedData$CA15,
na.rm = TRUE)
```

```

R> completedData$CA19[is.na(completedData$CA19)] <- mean(completedData$CA19,
na.rm = TRUE)
R> completedData$SerbestPSA[is.na(completedData$SerbestPSA)] <-
mean(completedData$SerbestPSA, na.rm = TRUE)
R> completedData$PSAOranı[is.na(completedData$PSAOranı)] <-
mean(completedData$PSAOranı, na.rm = TRUE)
R> ion.complete <- completedData

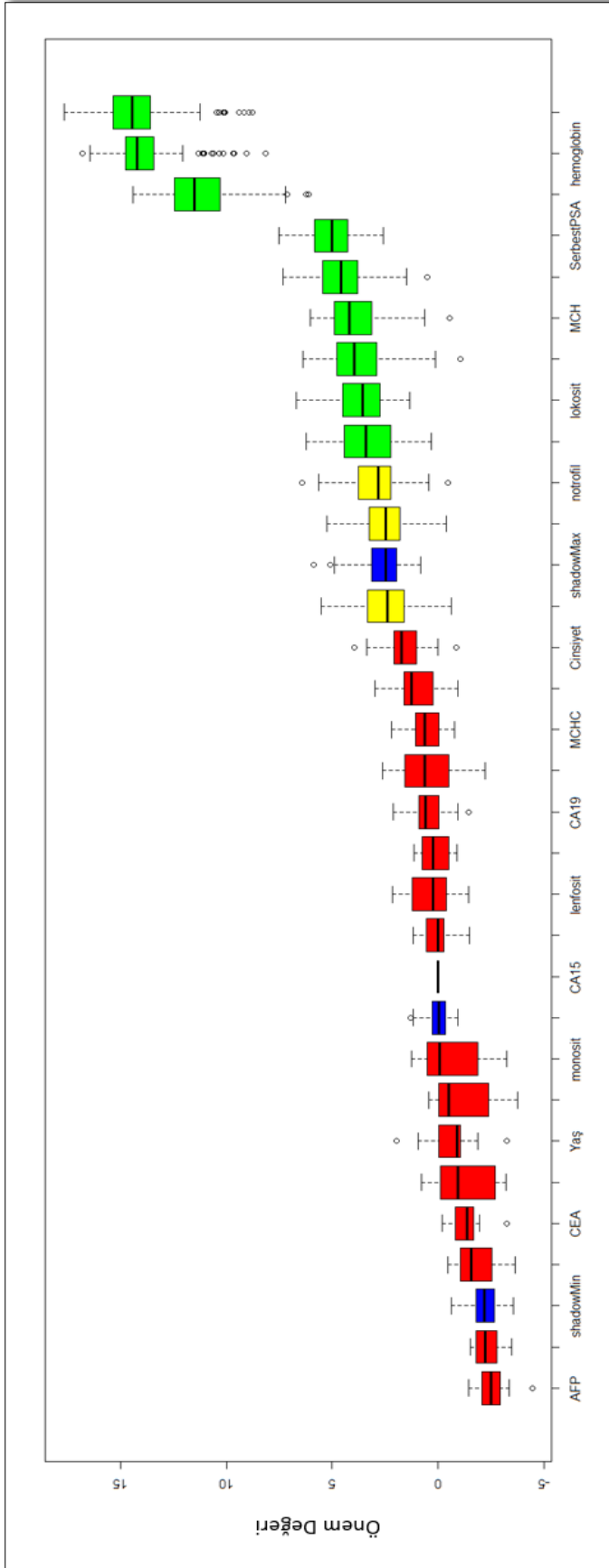
```

Adım 7. Makine öğrenmesi çalışmalarının önemli bir adımı olan değişken seçim aşaması bu adımda gerçekleştirilir. Filtreleme metodu, bütünleşik metotlar ve düzenlenmiş karar ağaçları gibi farklı değişken seçim algoritmaları bulunmaktadır. Bu çalışmada Boruta algoritması kullanılarak değişkenlerin yapılacak kestirimdeki önem düzeyleri belirlenmiştir. Şekil 4.1’de kullanılan değişkenlerin önem dağılımı verilmiştir.

```

R> colSums(is.na(ion))
R> install.packages("Boruta")
R> library("Boruta")
R> imputer = Imputer(missing_values='NaN', strategy='mean', axis=0)
R> Boruta.ion <- Boruta(tetkik~, data=ion.complete, doTrace=2, ntree=500)
R> Boruta.ion
R> plot(Boruta.ion)
R> print(Boruta.ion)

```



Şekil 4.1. Çalışmada kullanılan değişkenlerin önem değerlerinin dağılımı

Elde edilen dağılım grafiğinin yorumlanabilmesi için kullanılan algoritma doğrultusunda belirlenen önem değerlerine ilişkin dağılım test edilir. Bu aşamada her değişken için bir gölge değişken oluşturulur ve hesaplanan Z skoru ile önem değerleri hesaplanır. Yapılan test sonucuna göre elde edilen sonuçların dağılımı Çizelge 4.2’de verilmiştir.

R> attStats(Boruta.ion)

Çizelge 4.2. Çalışmada kullanılan değişkenlerin modeldeki önem durumunun kontrolüne ilişkin dağılım

	Önem Değeri Ortalaması	Önem Değeri Medyanı	Minimum Önem Değeri	Maksimum Önem Değeri	p	Karar
Eritropoietin	2,44	2,38	-0,64	5,51	0,46	Önemli
Ferritin	-1,28	0,94	-3,21	0,77	0,00*	Önemsiz
Homosistein	4,57	4,60	0,51	7,31	0,91	Önemli
Paratiroid	-1,16	-0,49	-3,74	0,43	0,00*	Önemsiz
AFP	-2,58	-2,49	-4,46	-1,45	0,00*	Önemsiz
CEA	-1,35	-1,38	-3,27	-0,19	0,00*	Önemsiz
CA15	0,00	0,00	0,00	0,00	0,00*	Önemsiz
CA19	0,42	0,61	-1,43	2,12	0,00*	Önemsiz
PSA	3,42	3,43	0,31	6,24	0,75	Önemli
Serbest PSA	5,01	5,01	2,57	7,53	0,94	Önemli
PSA Oranı	0,44	0,63	-2,24	2,64	0,03*	Önemsiz
Hemoglobin	1,38	1,42	8,17	1,68	1,00	Önemli
Hematokrit	1,42	1,45	8,80	1,77	1,00	Önemli
Eritrosit	1,12	1,15	6,13	1,44	1,00	Önemli
Lökosit	3,69	3,57	1,35	6,71	0,80	Önemli
Trombosit	2,44	2,48	-0,40	5,27	0,51	Belirsiz
MCH	3,95	4,21	-0,54	6,03	0,79	Önemli
MCV	3,78	3,94	-1,05	6,39	0,80	Önemli
MCHC	0,61	0,64	-0,78	2,18	0,00*	Önemsiz
RDW	0,96	1,24	-0,95	2,99	0,03*	Önemsiz
PDW	0,02	0,02	-1,48	1,16	0,00*	Önemsiz
MPV	1,78	-1,58	-3,65	-0,46	0,00*	Önemsiz
Nötrofil	2,95	2,81	-0,47	6,41	0,63	Belirsiz
Lenfosit	0,39	0,23	-1,44	2,17	0,00*	Önemsiz
Monosit	-0,61	-0,08	-3,26	1,25	0,00*	Önemsiz
Eozinofil	-2,31	-2,22	-3,47	-1,51	0,00*	Önemsiz
Bazofil	0,13	0,24	-0,90	1,14	0,00*	Önemsiz

*p<0,05

Adım 8. Çalışmada yapılan değişken azaltma adımından elde edilen sonuçlar değerlendirildiğinde, eksik veri oranının yüksek olduğu ve önemsiz çıkan değişkenler çalışmadan çıkarılır,

```
R>ion.drop = subset(ion.complete, select = -c(Homosistein, Paratiriot, AFP, CEA, CA15, CA19, PSA, SerbestPSA, PSAOranı))
```

4.2. Tanımlayıcı İstatistikler

Çalışmada ele alınan hastaların cinsiyet ve tarama yapılan mutasyonlara ilişkin dağılım Çizelge 4.3'te verilmiştir. Çalışmaya katılan hastaların %65,0'ı erkek, %35,0'ı kadındır. Kliniğe başvuran hastaların %72,7'sinde JAK2 V617F mutasyonunun varlığı incelenmiştir.

Çizelge 4.3. Demografik özelliklerin dağılımı

Değişkenler		Frekans	%
Cinsiyet	Kadın	449	35,0
	Erkek	835	65,0
Tetkik	JAK2 V617F	933	72,7
	JAK2 Ekzon 12	220	17,1
	CALR	46	3,6
	MPL	85	6,6

Tarama yapılan mutasyonlara ilişkin çalışma değişkenlerinin dağılımı Çizelge 4.4'te verilmiştir. Bu kapsamda,

- JAK2 V617F mutasyon varlığının incelendiği hastalarda MCH, PDW ve eozinofil değerleri,
- JAK2 Ekzon 12 mutasyon varlığının incelendiği hastalarda hematokrit, eritrosit, MCH, PDW ve eozinofil değerleri,
- CALR ve MPL mutasyon varlıklarının incelendiği hastalarda MCH ve PDW değerleri, referans aralık değerlerinin dışında çıkmıştır.

Çizelge 4.4. Çalışma değişkenlerinin mutasyona göre dağılımı

	Referans Aralık	Mutasyon							
		<i>JAK2</i> V617F (n=933)		<i>JAK2</i> Ekzon 12 (n=220)		<i>CALR</i> (n=46)		<i>MPL</i> (n=85)	
		Ort.	S.S.	Ort.	S.S.	Ort.	S.S.	Ort.	S.S.
Eritropoietin	1-25	10,6	13,2	9,4	13,6	21,5	17,9	22,2	32,0
Ferritin	20-200*	105,4	270,4	156,7	564,8	74,2	32,7	90,8	80,2
	20-500**	185,4	327,5	135,7	181,7	172,6	386,1	343,1	558,5
Hemoglobin	12-16	14,6	3,0	16,0	2,4	14,1	3,2	13,1	3,2
Hematokrit	35-46	44,4	8,7	48,3	6,7	42,5	9,5	39,8	9,4
Eritrosit	4-5,2	5,1	1,0	5,6	0,8	4,9	1,2	4,7	1,2
Lökosit	4,5-11	9,6	10,9	8,6	3,8	8,9	4,7	9,5	9,3
Trombosit	150-400	298,4	196,7	278,5	161,1	351,2	213,3	331,5	229,7
MCH	26-32	15,9	11,3	16,0	11,0	17,5	11,6	17,2	11,6
MCV	80-96	86,7	7,4	86,7	6,3	86,9	7,1	86,4	7,9
MCHC	30-36	32,9	1,4	33,1	1,3	33,1	1,2	32,9	1,4
RDW	8-18	15,2	3,2	14,8	3,0	16,0	3,1	16,2	3,3
PDW	9-14	19,3	1,9	19,4	1,8	20,1	1,8	20,0	1,8
MPV	7-12	8,7	1,8	8,6	1,7	8,2	2,1	8,3	1,8
Nötrofil	2-7,8	5,9	6,7	5,1	2,2	5,9	3,9	6,1	6,8
Lenfosit	1-4	2,6	6,7	2,5	1,6	2,2	0,8	2,3	1,6
Monosit	0-1	0,7	1,7	0,7	1,2	0,6	0,2	0,7	1,8
Eozinofil	0-1	1,3	7,8	1,8	9,2	0,2	0,2	0,2	0,2
Bazofil	0-0,2	0,2	2,1	0,2	0,5	0,1	0,2	0,1	0,3

*Kadın hastalar için referans aralık

**Erkek hastalar için referans aralık

***S.S: Standart sapma

Çalışma kapsamında kullanılan makine öğrenmesi yöntemine göre verinin bir bölümü eğitim verisi olarak kullanılırken, diğer kısmı da test verisi olarak kullanılmaktadır. Bu kapsamda algoritmalarla yapılan kestirimlerde %70, %80 ve %90 oranlarında eğitim verisi kullanılmıştır. Bu bilgi ışığında, sırasıyla da %30, %20 ve %10 oranlarında test verisiyle

kestirim yapılmıştır. Belirtilen oranlara ilişkin çalışmada kullanılan eğitim ve test verilerinin sayıları Çizelge 4.5'te gösterilmiştir.

Çizelge 4.5. Makine öğrenmesi algoritmalarında kullanılan veri sayılarının dağılımı

Eğitim Verisi Oranı (%)	Eğitim Verisi Sayısı	Test Verisi Sayısı	Toplam Veri Sayısı
70	901	383	1284
80	1028	256	1284
90	1157	127	1284

4.3. Kullanılan Algoritmalara İlişkin İstatistikler

Makine öğrenmesi yöntemi sınıflandırma algoritmalarından Naive Bayes, k-en yakın komşu, doğrusal diskriminant analizi, destek vektör makinesi, gini ve C5.0 sınıflandırma algoritmaları kullanılarak kestirime yönelik analiz gerçekleştirilmiştir. Algoritmalara ait R programlama dilinde kullanılan kodlamalara ve analiz sonucu elde edilen kestirim değerlerine ilişkin sonuçlara bu bölümde yer verilmiştir.

4.3.1. Naive Bayes sınıflandırma algoritması

Bu bölümde Naive Bayes sınıflandırma algoritmasına ilişkin yapılan analiz sonuçlarına ve bu sonuçları makine öğrenmesi yönteminde elde edebilmek için kullanılan kodlamalara yer verilmiştir.

Algoritmanın programa tanıtım işleminin tamamlanabilmesi için aşağıda verilen kodların öncelikle çalıştırılması gerekmektedir.

```
R> install.packages("klaR")
R> library(caret)
R> library(MASS)
R> library(klaR)
R> library(psych)
R> set.seed(80)
```

Daha sonraki aşamada değişkenler tanımlanır, eğitim ve test verisi seçim işlemleri aşağıdaki kodlamalar kullanılarak gerçekleştirilmektedir.

```
R> pairs.panels(ion.drop[1:11], gap = 0, bg = c("red", "green","blue",
"yellow")[ion.drop$ tetkik], pch = 19)
R> responseY <- as.matrix(ion.drop[,dim(ion.drop)[2]])
R> predictorX <- as.matrix(ion.drop[,1:(dim(ion.drop)[2]-1)])
R> trainIndex <- createDataPartition(responseY, times=1, p = 0.8, list = F)
R> training <- ion.drop[trainIndex,]
R> testing <- ion.drop[-trainIndex,]
R> dim(training); dim(testing);
R> trctrl <- trainControl(method = "cv", number = 5)
```

Değişkenlere ilişkin işlemler tamamlandıktan sonra aşağıdaki kodlamalar yardımıyla analiz gerçekleştirilmektedir.

```
R> model.nb <- train(tetkik ~., data=training, method="nb", trControl= trctrl)
R> predictions <- predict(model.nb, training)
R> confusionMatrix(predictions, training$tetkik)
R> predictions <- predict(model.nb, testing)
R> confusionMatrix(predictions, testing$tetkik)
```

Naive Bayes sınıflandırmasına göre yapılan analiz sonuçları Çizelge 4.6’da verilmiştir. Sonuçlar değerlendirildiğinde %64,8 ile en yüksek doğruluk değeri %80’lik eğitim verisinin kullanıldığı analizde elde edilmiştir. *JAK2 V617F* mutasyonuna ilişkin yapılan kestirimde %82,8 oranında doğru mutasyon tahmin edilmiştir. Bu mutasyonun yanlış tahmin edilme oranı %20 ile diğerlerine oranla daha düşük saptanmıştır. Yapılan kestirime ilişkin doğruluk değerinin mutasyon bazlı değerlendirilmesi sonucunda F ölçütü değeri %77,8 ile *JAK2 V617F* mutasyonu için en yüksek oranda saptanmıştır.

Çizelge 4.6. Naive Bayes sınıflandırma algoritması sonuçları

Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
0,7	0,632	<i>JAK2 V617F</i>	0,824	0,154	0,723	0,771
		<i>JAK2 Ekzon 12</i>	0,152	0,921	0,286	0,198
		<i>MPL</i>	0,000	0,984	0,000	-
		<i>CALR</i>	0,080	0,939	0,083	0,082
0,8	0,648	<i>JAK2 V617F</i>	0,828	0,200	0,733	0,778
		<i>JAK2 Ekzon 12</i>	0,205	0,920	0,346	0,257
		<i>MPL</i>	0,000	0,992	0,000	-
		<i>CALR</i>	0,176	0,937	0,167	0,171
0,9	0,598	<i>JAK2 V617F</i>	0,807	0,118	0,714	0,758
		<i>JAK2 Ekzon 12</i>	0,045	0,924	0,111	0,065
		<i>MPL</i>	0,000	0,976	0,000	-
		<i>CALR</i>	0,000	0,916	0,000	-

4.3.2. k-en yakın komşu sınıflandırma algoritması

Bu bölümde k-en yakın komşu sınıflandırma algoritmasına ilişkin yapılan analiz sonuçlarına ve bu sonuçları makine öğrenmesi yönteminde elde edebilmek için kullanılan kodlamalara yer verilmiştir.

Öncelikle aşağıda verilen kodlamalar ile algoritmanın tanıtılma işlemi gerçekleştirilmektedir.

```
R> library(class)
```

```
R> library(caret)
```

```
R> library(psych)
```

```
R> set.seed(80)
```

Algoritmanın programa tanıtılmasına ilişkin yükleme işlemi tamamlandıktan sonra, değişkenlere ilişkin tanımlamalar yapılarak, eğitim ve test verisine ilişkin bölümlendirme işlemi aşağıdaki kodlamalar ile gerçekleştirilmektedir.

```
R> pairs.panels(ion.drop[1:11], gap = 0, bg = c("red", "green", "blue", "yellow")[ion.drop$etetik], pch = 19)
```

```

R> responseY <- as.matrix(ion.drop[,dim(ion.drop)[2]])
R> predictorX <- as.matrix(ion.drop[,1:(dim(ion.drop)[2]-1)])
R> trainIndex <- createDataPartition(responseY, times=1, p = 0.8, list = F)
R> training <- ion.drop[trainIndex,]
R> testing <- ion.drop[-trainIndex,]
R> dim(training); dim(testing);
R> trctrl <- trainControl(method = "cv", number = 5)

```

Veriye ilişkin tüm kodlamalar tamamlandıktan sonra modele ilişkin aşağıda verilen kodlamalar ile algoritma çalıştırılmaktadır.

```

R> model.knn <- train(tetkik ~., data = training, method = "knn", trControl=trctrl,
tuneLength = 10)
R> model.knn
R> plot(model.knn)
R> train_pred <- predict(model.knn, newdata = training)
R> confusionMatrix(train_pred, training$tetkik )
R> test_pred <- predict(model.knn, newdata = testing)
R> confusionMatrix(test_pred, testing$tetkik ) #check accuracy

```

k-en yakın komşu sınıflandırma algoritması ile yapılan kestirim modeli sonuçlarına göre en yüksek doğruluk değeri 0,732 ile verinin %90'nın eğitim verisi olarak değerlendirildiği analizde saptanmıştır. Algoritma ile elde edilen tüm analiz sonuçları Çizelge 4.7'de verilmektedir.

Çizelge 4.7. k-en yakın komşu sınıflandırma algoritması sonuçları

Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
0,7	0,726	<i>JAK2 V617F</i>	0,996	0,000	0,728	0,841
		<i>JAK2 Ekzon 12</i>	0,000	0,997	0,000	-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	1,000		-
0,8	0,727	<i>JAK2 V617F</i>	1,000	0,000	0,727	0,842
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	1,000		-
0,9	0,732	<i>JAK2 V617F</i>	1,000	0,000	0,732	0,845
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	1,000		-

k-en yakın komşu sınıflandırma algoritma sonuçlarına göre, *JAK2 V6127F* mutasyonu için duyarlılık değeri %100,0 ve diğer üç mutasyonlar için %0 olarak elde edilmiştir.

4.3.3. Doğrusal diskriminant analizi sınıflandırma algoritması

Bu bölümde doğrusal diskriminant analizi sınıflandırma algoritmasına ilişkin yapılan analiz sonuçlarına ve bu sonuçları makine öğrenmesi yönteminde elde edebilmek için kullanılan kodlamalara yer verilmiştir.

Bu aşamada öncelikle algoritmanın programa tanıtılması gerekmektedir ve bu kapsamda algoritmaya ilişkin yüklemeler verilen kodlamalar ile yapılmaktadır.

```

R> install.packages("psych")
R> install.packages("caret")
R> install.packages("e1071")
R> library(psych)
R> library(caret)
R> library(MASS)
R> set.seed(80)

```

Algoritmanın programa tanıtılması ve yükleme işlemi tamamlandıktan sonra değişkenlere ilişkin tanımlamalar yapılarak, eğitim ve test verisine ilişkin bölümlleme işlemi aşağıdaki kodlamalar ile gerçekleştirilmektedir.

```
R> pairs.panels(ion.drop[1:11], gap = 0, bg = c("red", "green","blue",
"yellow")[ion.drop$tetikik], pch = 19)
R> responseY <- as.matrix(ion.drop[,dim(ion.drop)[2]])
R> predictorX <- as.matrix(ion.drop[,1:(dim(ion.drop)[2]-1)])
R> trainIndex <- createDataPartition(responseY, times=1, p = 0.8, list = F)
R> training <- ion.drop[trainIndex,]
R> testing <- ion.drop[-trainIndex,]
R> dim(training); dim(testing);
R> trctrl <- trainControl(method = "cv", number = 5)
```

Veriye ilişkin tüm kodlamalar tamamlandıktan sonra modele ilişkin aşağıda verilen kodlamalar yazılarak algoritmaya ilişkin analiz gerçekleştirilmektedir.

```
R> model.lda <- train(tetikik ~., data=training, method="lda", trControl= trctrl)
R> predictions <- predict(model.lda, training)
R> confusionMatrix(predictions, training$tetikik)
R> predictions <- predict(model.lda, testing)
R> confusionMatrix(predictions, testing$tetikik)
```

Doğrusal diskriminant analizi sınıflandırma algoritmasına göre yapılan analizler sonucunda çalışmada ele alınan mutasyon taramalarına ilişkin elde edilen sonuçlar Çizelge 4.8’da verilmiştir.

Çizelge 4.8. Doğrusal diskriminant analizi sınıflandırma algoritması sonuçları

Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
0,7	0,710	<i>JAK2 V617F</i>	0,961	0,077	0,736	0,834
		<i>JAK2 Ekzon 12</i>	0,000	0,997	0,000	-
		<i>MPL</i>	0,077	0,981	0,125	0,095
		<i>CALR</i>	0,000	0,980	0,300	0,000
0,8	0,715	<i>JAK2 V617F</i>	0,957	0,114	0,742	0,836
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,222	0,984	0,333	0,267
		<i>CALR</i>	0,176	0,971	0,300	0,222
0,9	0,693	<i>JAK2 V617F</i>	0,946	0,000	0,721	0,819
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,000	0,976	0,000	-
		<i>CALR</i>	0,000	0,983	0,000	-

Elde edilen bu sonuçlar doğrultusunda, makine öğrenmesi yönteminde modelin kestirim yapabilmesi için verilerin %80'sinin eğitim verisi olarak kullanıldığı analizlerde en iyi sonuç elde edilmiştir. Yapılan bu kestirimde doğruluk değeri %71,5 olarak hesaplanmıştır. Tespit edilen bu değere ilişkin kestirim yapılmaya çalışılan tarama testlerinin duyarlılığı *JAK2 V6127F* için diğer üç mutasyon taramasından oldukça yüksek ve %95,7 olarak elde edilmiştir.

4.3.4. Destek vektör makinesi sınıflandırma algoritması

Bu bölümde destek vektör makinesi sınıflandırma algoritmasına ilişkin yapılan analiz sonuçlarına ve bu sonuçları makine öğrenmesi yönteminde elde edebilmek için kullanılan kodlamalara yer verilmiştir.

Öncelikle algoritmanın programa tanıtılabilmesi için aşağıda verilen kodlamalar yazılmaktadır.

```
R> library(e1071)
```

```
R> library(caret)
```

```
R> library(psych)
```

```
R> library(MASS)
```

```
R> set.seed(80)
```

Algoritmanın programa tanıtılmasına ilişkin yükleme işlemi tamamlandıktan sonra, değişkenlere ilişkin tanımlamalar yapılarak, eğitim ve test verisine ilişkin bölümlendirme işlemi aşağıdaki kodlamalar ile gerçekleştirilmektedir.

```
R> pairs.panels(ion.drop[1:11], gap = 0, bg = c("red", "green","blue",
"yellow")[ion.drop$tetkik], pch = 19)
R> responseY <- as.matrix(ion.drop[,dim(ion.drop)[2]])
R> predictorX <- as.matrix(ion.drop[,1:(dim(ion.drop)[2]-1)])
R> trainIndex <- createDataPartition(responseY, times=1, p = 0.8, list = F)
R> training <- ion.drop[trainIndex,]
R> testing <- ion.drop[-trainIndex,]
R> dim(training); dim(testing);
R> trctrl <- trainControl(method = "cv", number = 5)
```

Veriye ilişkin tüm kodlamalar tamamlandıktan sonra modele ilişkin aşağıda verilen kodlamalarla algoritma çalıştırılmaktadır.

```
R> model.svm = svm(tetkik ~ ., data = training, kernel = "polynomial", trControl= trctrl)
R> train_pred <- predict(model.svm, training)
R> confusionMatrix(train_pred, training$tetkik)
R> test_pred <- predict(model.svm, testing)
R> confusionMatrix(test_pred, testing$tetkik)
```

Destek vektör makinesi sınıflandırma algoritması kullanılarak yapılan analizler kapsamında doğrusal fonksiyon, polinomiyal fonksiyon ve radyal tabanlı fonksiyon kullanılarak dönüşüm işlemleri gerçekleştirilmiştir. Her üç fonksiyonun kullanıldığı sınıflandırma analizinde sonuçlar birbirinin aynısı olarak elde edilmiş olup, ilgili sonuçlar Çizelge 4.9’da verilmiştir.

Çizelge 4.9. Destek vektör makinesi sınıflandırma algoritması sonuçları

Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
0,7	0,723	<i>JAK2 V617F</i>	0,986	0,029	0,731	0,840
		<i>JAK2 Ekzon 12</i>	0,000	0,994	0,000	-
		<i>MPL</i>	0,077	0,997	0,500	0,133
		<i>CALR</i>	0,040	0,994	0,333	0,071
0,8	0,715	<i>JAK2 V617F</i>	0,984	0,029	0,729	0,838
		<i>JAK2 Ekzon 12</i>	0,000	0,995	0,000	-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	0,983	0,000	-
0,9	0,717	<i>JAK2 V617F</i>	0,979	0,000	0,726	0,833
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	0,983	0,000	-

Çizelge 3.7’de verilen sonuçlara göre, yapılan kestirimlerde her bir test verisi kullanma durumuna ilişkin anlamlı sonuçlar elde edilmiştir. Algoritmaya ilişkin %72,3 doğruluk değeri ile yapılan kestirimler %70’lik test verisinin kullanıldığı analizlerde saptanmıştır. Bu analizde *JAK2 V617F* mutasyona ilişkin kestirim değerleri %98,6’larda hesaplanmıştır. Yapılan kestirimlerde *JAK2 V617F* mutasyonu %98,6 oranında doğru olarak tahmin edilmiştir. Bu doğruluk ve duyarlılık değerleri hesaplanan %84,0’lık F ölçütü değeri ile de desteklenmiştir.

4.3.5. Gini sınıflandırma algoritması

Bu bölümde karar ağacı sınıflandırma algoritmalarından gini sınıflandırma algoritmasına ilişkin yapılan analiz sonuçlarına ve bu sonuçları makine öğrenmesi yönteminde elde edebilmek için kullanılan kodlamalara yer verilmiştir.

Öncelikle algoritmanın makineye tanıtılabilmesi için aşağıda verilen kodlamalar programa yazılmaktadır.

```
R> install.packages("rpart")
```

```
R> library(caret)
```

```
R> library(rpart)
```

```
R> set.seed(80)
```

Algoritmanın programa tanıtılmasına ilişkin yükleme işlemi tamamlandıktan sonra, değişkenlere ilişkin tanımlamalar yapılarak, eğitim ve test verisine ilişkin bölümlendirme işlemi aşağıdaki kodlamalar ile gerçekleştirilmektedir.

```
R> ayrim=createDataPartition(y = ion.drop$tetkik, p = .8, list = FALSE)
R> egitim=ion.drop[ayrim,]
R> test=ion.drop[-ayrim,]
R> egitim_hedef<-ion.drop[[19]]
R> test_hedef<-ion.drop[[19]]
R> egitim_nitelikler<-ion.drop[, -19]
R> test_nitelikler<-ion.drop[, -19]
```

Veriye ilişkin tüm kodlamalar tamamlandıktan sonra modele ilişkin aşağıda verilen kodlamalar yazılarak analiz gerçekleştirilmektedir.

```
R> gini<-rpart(tetkik ~., data=egitim,method="class",minsplit=4)
show(gini)
R> tahmin_gini = predict(gini,test,type="class")
R> cm_gini = confusionMatrix(data = tahmin_gini, reference = test[,19])
R> cm_gini
```

Gini sınıflandırma algoritması ile ele edilen analiz sonuçları Çizelge 4.10'da verilmiştir. Sonuçlara göre, en yüksek doğruluk değeri veri setinin %10'unun test verisi olarak kullanıldığı durumda saptanmıştır. Bu doğruluk değerinde JAK2 V617F mutasyon tarama testinin duyarlılık değeri %100 çıkarken, diğer taramalarda bu değer %0 olarak tespit edilmiştir.

Çizelge 4.10. Gini sınıflandırma algoritması sonuçları

Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
0,7	0,729	<i>JAK2 V617F</i>	1,000	0,000	0,729	0,843
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	1,000		-
0,8	0,719	<i>JAK2 V617F</i>	0,978	0,029	0,728	0,835
		<i>JAK2 Ekzon 12</i>	0,045	0,981	0,333	0,080
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	1,000		-
0,9	0,732	<i>JAK2 V617F</i>	1,000	0,000	0,732	0,845
		<i>JAK2 Ekzon 12</i>	0,000	1,000		-
		<i>MPL</i>	0,000	1,000		-
		<i>CALR</i>	0,000	1,000		-

4.3.6. C5.0 sınıflandırma algoritması

Bu bölümde karar ağacı sınıflandırma algoritmalarından C5.0 sınıflandırma algoritmasına ilişkin yapılan analiz sonuçlarına ve bu sonuçları makine öğrenmesi yönteminde elde edebilmek için kullanılan kodlamalara yer verilmiştir.

Bu aşamada öncelikle algoritmanın makineye tanıtılması gerekmektedir ve bu kapsamda algoritmaya ilişkin yüklemeler aşağıda verilen kodlamalar ile yapılmaktadır.

```
R> install.packages("modeldata")
```

```
R> library(modeldata)
```

```
R> library(caret)
```

```
R> library(C50)
```

```
R> set.seed(80)
```

Algoritmanın programa tanıtılmasına ilişkin yükleme işlemi tamamlandıktan sonra, değişkenlere ilişkin tanımlamalar yapılarak, eğitim ve test verisine ilişkin bölümlendirme işlemi aşağıdaki kodlamalar ile gerçekleştirilmektedir.

```

R> ayrim=createDataPartition(y = ion.drop$tetikik, p = .8, list = FALSE)
R> egitim=ion.drop[ayrim,]
R> test=ion.drop[-ayrim,]
R> egitim_hedef<-ion.drop[[19]]
R> test_hedef<-ion.drop[[19]]
R> egitim_nitelikler<-ion.drop[, -19]
R> test_nitelikler<-ion.drop[, -19]

```

Veriye ilişkin tüm kodlamalar tamamlandıktan sonra modele ilişkin aşağıda verilen kodlamalarla algoritma çalıştırılmaktadır.

```

R> treeModel=C50::C5.0(ion.drop2[1:1284, -19], ion.drop2$tetikik[1:1284])
R> treeModel
R> summary(treeModel)
R> ruleModel=C50::C5.0(tetikik~, data=ion.drop2[1:1284, ], rules=TRUE)
R> ruleModel
R> summary(ruleModel)
R> tahmin_c50 = predict(ruleModel, test, type="class")
R> cm_c50 = confusionMatrix(data = tahmin_c50, reference = test[, 19])
R> cm_c50

```

Çizelge 4.11’de C5.0 sınıflandırma algoritması ile ele edilen analiz sonuçları yer almaktadır. Sonuçlara göre, en yüksek doğruluk değeri veri setinin %10’unun test edildiği durumda ortaya çıkmıştır. Bu doğruluk değerinde *JAK2* V617F mutasyon taramasının duyarlılık değeri %94,6, *JAK2* Ekzon 12 mutasyon taraması için %31,8 ve *MPL* mutasyon taraması için %50,0 çıkarken, diğer mutasyon taraması için %0 olarak hesaplanmıştır.

Çizelge 4.11. C5.0 sınıflandırma algoritması sonuçları

Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
0,7	0,742	<i>JAK2 V617F</i>	0,939	0,231	0,766	0,844
		<i>JAK2 Ekzon 12</i>	0,182	0,946	0,414	0,253
		<i>MPL</i>	0,385	0,997	0,833	0,526
		<i>CALR</i>	0,200	0,997	0,833	0,323
0,8	0,742	<i>JAK2 V617F</i>	0,946	0,200	0,759	0,842
		<i>JAK2 Ekzon 12</i>	0,159	0,958	0,438	0,233
		<i>MPL</i>	0,333	1,000	1,000	0,500
		<i>CALR</i>	0,235	0,996	0,800	0,364
0,9	0,764	<i>JAK2 V617F</i>	0,946	0,265	0,779	0,854
		<i>JAK2 Ekzon 12</i>	0,318	0,952	0,583	0,412
		<i>MPL</i>	0,500	1,000	1,000	0,667
		<i>CALR</i>	0,000	1,000		-

4.4. Algoritma Bulgularının Karşılaştırılması

Bu bölümde, çalışma kapsamında Naive Bayes, k-en yakın komşu, doğrusal diskriminant analizi, destek vektör makinesi, gini ve C5.0 sınıflandırma algoritmaları kullanılarak elde edilen analiz sonuçlarına ilişkin karşılaştırmalara yer verilmiştir. Çalışmada %70, %80 ve %90 olmak üzere üç farklı oranda eğitim verisi kullanılmıştır. Çizelge 4.12’de ele alınan makine öğrenmesi algoritmalarından elde edilen en iyi kestirimler özetlenmiştir.

Çizelge 4.12. En iyi doğruluk değerinin tespit edildiği eğitim verisi oranlarına göre elde edilen analiz sonuçları

	Doğru Sonucun Verildiği Oran	Doğruluk	Tarama Testi	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Naive Bayes	0,8	0,648	JAK 2 V617F	0,828	0,200	0,733	0,778
			JAK 2 Ekzon 12	0,205	0,920	0,346	0,257
			MPL	0,000	0,992	0,000	-
			CALR	0,176	0,937	0,167	0,171
K En Yakın Komşu	0,9	0,732	JAK 2 V617F	1,000	0,000	0,732	0,845
			JAK 2 Ekzon 12	0,000	1,000		-
			MPL	0,000	1,000		-
			CALR	0,000	1,000		-
Doğrusal Diskriminant Analizi	0,8	0,715	JAK 2 V617F	0,957	0,114	0,742	0,836
			JAK 2 Ekzon 12	0,000	1,000		-
			MPL	0,222	0,984	0,333	0,267
			CALR	0,176	0,971	0,300	0,222
Destek Vektör Makineleri	0,7	0,723	JAK 2 V617F	0,986	0,029	0,731	0,840
			JAK 2 Ekzon 12	0,000	0,994	0,000	-
			MPL	0,077	0,997	0,500	0,133
			CALR	0,040	0,994	0,333	0,071
Gini	0,9	0,732	JAK 2 V617F	1,000	0,000	0,732	0,845
			JAK 2 Ekzon 12	0,000	1,000		-
			MPL	0,000	1,000		-
			CALR	0,000	1,000		-
C5.0	0,9	0,764	JAK 2 V617F	0,946	0,265	0,779	0,854
			JAK 2 Ekzon 12	0,318	0,952	0,583	0,412
			MPL	0,500	1,000	1,000	0,667
			CALR	0,000	1,000		-

Çalışma kapsamında kullanılan algoritmalarda yapılan kestirimler birbirlerine göre değerlendirildiğinde,

- Naive Bayes sınıflandırma algoritmasında, çalışmada ele alınan veri setinin %20'sinin test verisi olarak kestirim yapıldığı durumda %64,8 ile en yüksek doğruluk değerine ulaşılmıştır. Mutasyonların duyarlılık değerleri; *JAK2 V6127F* için %82,8, *JAK2 Ekzon 12* için %20,5, *CALR* için %17,6 ve diğer mutasyon taraması için %0 olarak elde edilmiştir.
- K-en yakın komşu sınıflandırma algoritmasında %90 eğitim verisi kullanılarak yapılan analizlerde en yüksek doğruluk oranı elde edilmiştir. Yapılan bu kestirimde *JAK2 V617F* mutasyonuna ilişkin duyarlılık değeri %100,0 ile hiç yanlış kestirim yapılmadan

gerçekleştirilmiştir. Bu sonuçlar doğru bir kestirim yapıldığının göstergesi olarak değerlendirilmiştir. Bu değerlendirme algoritma için hesaplanan ve %84,5 olarak elde edilen F ölçütü değeriyle de desteklenmektedir.

- Doğrusal diskriminant analizi algoritması ile yapılan kestirimde en yüksek doğruluk değeri %80 eğitim verisinin kullanıldığı analizde elde edilmiştir. Bu analizde mutasyonların duyarlılık değerleri; *JAK2* V6127F için %95,7, *MPL* için %22,2, *CALR* için %4,0 ve diğer mutasyon taraması için %0 olarak elde edilmiştir.
- Destek vektör makineleri algoritması ile en yüksek doğruluk değeri veri setinin %30'sinin kestirim yapıldığı modellemede ortaya çıkmıştır. Yapılan kestirimde *JAK2* V617F mutasyonu %98,6 oranında doğru olarak tahmin edilmiştir. Bu doğruluk ve duyarlılık değerleri hesaplanan %84,0'lık F ölçütü değeri ile de desteklenmiştir.
- Gini sınıflandırma algoritmasında %90 eğitim verisi kullanılarak yapılan analizlerde en yüksek doğruluk oranı %73,2 ile elde edilmiştir. Yapılan bu kestirimde *JAK2* V617F mutasyonuna ilişkin duyarlılık değeri %100,0 ile hiç yanlış kestirim yapılmadan gerçekleştirilmiştir. Bu sonuçlar doğru bir kestirim yapıldığının göstergesi olarak değerlendirilmiştir. Bu değerlendirme algoritma için hesaplanan ve %84,5 olarak elde edilen F ölçütü değeriyle de desteklenmektedir.
- C5.0 sınıflandırma algoritmasında %90 eğitim verisi kullanılarak yapılan analizlerde en yüksek doğruluk oranı %76,4 ile elde edilmiştir. Mutasyonların duyarlılık değerleri; *JAK2* V6127F için %94,6, *JAK2* Ekzon 12 için %31,8, *MPL* için %50,0 ve diğer mutasyon taraması için %0 olarak elde edilmiştir.

4.5. Değişkenlere İlişkin Fark Karşılaştırması

Çalışma kapsamında oluşturulan sınıflandırma modellerine ilişkin kullanılan değişkenler ile kestirim yapılan mutasyonlar arasındaki farklılıklar bu bölümde incelenmiş ve yorumlanmıştır.

Yapılan karşılaştırmalarda kullanılan değişkenler eritropoietin, ferritin ve tam kan sayım sonuçlarında elde edilen değerler mutasyon türlerine göre farklılık oluşturma durumları incelenmiştir. Belirtilen analizlere ilişkin sonuçlar Çizelge 4.13'te verilmiştir.

Çizelge 4.13. Çalışmada ele alınan hasta bulgularının yapılan tarama testine göre dağılımı

	Mutasyon										p değeri	Farklılık Gösteren Mutasyonlar
	JAK2 V612F		JAK2 EKZON12		CALR		MPL					
	Ort.	S.S.	Ort.	S.S.	Ort.	S.S.	Ort.	S.S.				
Eritropoietin	10,62	13,25	9,39	13,65	21,51	17,95	22,16	32,00	0,059			
Ferritin	105,39	270,35	156,72	564,76	74,25	32,67	90,83	80,16	0,203			
Hemoglobin	14,63	3,03	16,00	2,42	14,07	3,20	13,15	3,17	0,000*	JAK2		
Hematokrit	44,43	8,69	48,34	6,74	42,46	9,55	39,84	9,38	0,000*	JAK2		
Eritrosit	5,14	1,00	5,59	0,77	4,92	1,18	4,65	1,18	0,000*	JAK2		
Lökosit	9,62	10,92	8,60	3,83	8,91	4,74	9,49	9,34	0,58			
Trombosit	298,43	196,75	278,47	161,10	351,15	213,34	331,53	229,70	0,047*	JAK2		
MCH	15,9	11,3	16,0	11,0	17,5	11,6	17,2	11,6	0,648			
MCV	86,70	7,37	86,72	6,32	86,91	7,10	86,42	7,89	0,983			
MCHC	32,91	1,38	33,10	1,28	33,08	1,20	32,91	1,40	0,25			
RDW	15,23	3,20	14,79	3,02	15,96	3,07	16,16	3,32	0,003*	CALR ve MPL		
PDW	19,31	1,87	19,42	1,78	20,08	1,85	19,97	1,81	0,158			
MPV	8,66	1,84	8,65	1,68	8,24	2,11	8,26	1,76	0,131			
Nötrofil	5,89	6,68	5,12	2,17	5,86	3,89	6,14	6,77	0,383			
Lenfosit	2,55	6,72	2,48	1,56	2,15	0,83	2,29	1,60	0,040*	JAK2		
Monosit	0,71	1,67	0,68	1,24	0,57	0,23	0,72	1,80	0,938			
Eozinofil	1,31	7,83	1,78	9,21	0,20	0,22	0,20	0,24	0,009*	JAK2		
Bazofil	0,22	2,11	0,17	0,48	0,13	0,25	0,15	0,32	0,952			

*p<0,05 **S.S.: Standart Sapma

Yapılan istatistiksel karşılaştırmalar sonucunda hemoglobin, hematokrit, eritrosit, RDW ve eozinofil değerlerine göre yapılan mutasyon taramaları arasında istatistiksel olarak anlamlı fark bulunurken ($p<0,05$), diğer klinik değerlere göre istatistiksel olarak anlamlı farklılık saptanmamıştır ($p>0,05$). Belirtilen istatistiksel anlamlı bulgular değerlendirildiğinde; *JAK2* mutasyonları hemoglobin, hematokrit, eritrosit, trombosit, lenfosit ve eozinofil değerlerinde istatistiksel olarak anlamlı farklılık gösterirken, *CALR* ve *MPL* mutasyonları RDW değerinde istatistiksel olarak anlamlı farklılık göstermiştir.

4. SONUÇ VE ÖNERİLER

Myeloproliferatif hastalıklara yönelik yapılan literatür incelemesinde çalışmalarda ele alınan dört mutasyonun bir arada incelendiği çalışmaya nadir rastlanmıştır. Yapılan incelemelere ait özet bilgilere bu bölümde yer verilmiştir.

- Bahsi ve Yiğenoğlu (2019) yapmış oldukları araştırmada, *JAK2 V617F*, *CALR* ve *MPL* mutasyonlarının hastadaki varlığını incelerken, mevcut hasta gruplarındaki durum analizini yorumlamışlardır [1].
- Mağın ve diğerleri (2010) tarafından yapılan çalışmada polisitemia vera hastalığı tanısında *JAK2* mutasyonunun önemi incelenerek, bir olgu üzerinden yorumlama yapılmıştır [20].
- Bayram (2019) çalışmasında hastaların klinik bulguları ile sadece *JAK2* mutasyonunun varlığını incelemiş olup, bu mutasyonun klinik bulgular ile anlamlı bir ilişkisi olduğu sonucuna varmıştır [22].
- Öte (2015) çalışmasında, *JAK2 V617F*, *CALR* ve *MPL* mutasyonlarının varlığını inceleyerek *JAK2 V617F* ve *MPL* mutasyon varlığının negatif çıktığı hastalarda *CALR* mutasyon varlığının incelenmesinin tanının belirlenme sürecine zaman kazandırdığı sonucuna varmıştır [29].
- Edib ve diğerleri (2015) çalışmalarında, myeloproliferatif hastalıklara ilişkin *JAK2 V617F* mutasyonuna sahip hastalara yönelik bir değerlendirme yapmışlardır. Çalışmalarında bu mutasyonun hastalığın tanısında önemli rol üstlendiği sonucuna varmışlardır [31].
- Scott ve diğerleri (2007) tarafından yapılan çalışmada *JAK2* Ekzon 12 mutasyonuna ilişkin inceleme yapılarak, bu mutasyonun hastalık tanısında önemli rol oynadığı sonucuna varılmıştır [34].
- Uysal ve diğerleri (2020) tarafından yapılan çalışmada, myeloproliferatif hastalıklara ilişkin tanı saptama aşamasında kullanılan mutasyon taramaları ile konulan tanı üzerine bir karşılaştırma yapılmıştır. Bu doğrultuda *JAK2 V617F*, *CALR* ve *MPL* olmak üzere üç mutasyon incelenmiş ve myeloproliferatif hastalıklarda bu üç mutasyonun en sık tekrarlanan mutasyon olduğunu sonucuna varılmıştır [116].
- Zulkeflee ve diğerleri (2021) tarafından yapılmış olan çalışmada, yine aynı üç mutasyon varlığı incelenmiş olup, mutasyon varlığında konulan tanılar değerlendirilerek yorumlama yapılmıştır. Bahsedilen bu çalışmada, bir önceki çalışmadan farklı olarak

sadece *JAK2* V617F ve *CALR* mutasyonlarının hastalık tanısında önemli role sahip olduğu saptaması yapılmıştır [117].

- Li ve diğerleri (2017) tarafından yapılan çalışma kapsamında, *JAK2* V617F, *JAK2* Ekzon 12, *CALR* ve *MPL* mutasyonları ele alınarak incelenmiştir. Ele alınan bu çalışmada diğer çalışmalara benzer olarak dört mutasyonun tanı ile karşılaştırması yapılarak sonuçlandırılmıştır [118].
- Lang ve diğerleri (2018), Çin’de yaşayan Uygur ve Han uyruklu hastaların klinik özellikleri ile *JAK2* V617F, *CALR* ve *MPL* mutasyonları arasında ilişkilendirme yapmışlardır. Yaptıkları bu çalışmada iki uyruk arasında saptanan farkları değerlendirerek yorumlamışlardır [119].
- Passamonti ve diğerleri (2011) yaptıkları çalışmada *JAK2* V617F ve *JAK2* Ekzon 12 mutasyon varlığının polisitemia vera hastalığı tanısı ile ilişkilendirmesini incelemişlerdir. Çalışma sonucunda iki mutasyon arasında benzerlik saptamışlardır [120].
- Lin ve diğerleri (2015) tarafından yapılan çalışmada, *JAK2* V617F, *JAK2* Ekzon 12, *MPL* ve *CALR* mutasyonlarının varlığının tespitinin tanı aşamasına katkısı araştırılmış ve tanı kriteri olarak kullanılabileceği sonucuna varılmıştır [121].
- Soliman ve diğerleri (2020) tarafından Mısırlı myeloproliferatif hastalara yönelik bir çalışma yapılmıştır. Bu çalışmada *JAK2* V617F, *MPL* ve *CALR* mutasyonları değerlendirilmiş ve yorumlanmıştır [122].

Yapılan literatür incelemesinde daha önceki çalışmalarda, hastalık ve mutasyon varlığının karşılaştırıldığı, klinik bulgulara göre değerlendirildiği ve mutasyon varlığının tanı tespitindeki öneminin araştırıldığı saptanmıştır. Yapılan bu çalışmalara göre klinik bulgular ile mutasyonlar arasında bir model oluşturulmadığı tespit edilmiştir.

Ele alınan bu çalışmada, myeloproliferatif ön tanılı hastaların klinik bilgileri ile incelenmesi gereken mutasyonu tespit edebilecek bir model oluşturulmaya çalışılmıştır. İncelenen modelde, *JAK2* V617F, *JAK2* Ekzon 12, *CALR* ve *MPL* mutasyonlarının varlığının arandığı hastaların klinik bulguları kullanılmıştır. Çalışma kapsamında oluşturulan model sonucunda, klinisyene doğruluk değerleri ile değerlendirilen *JAK2* V617F, *JAK2* Ekzon 12, *CALR* ve *MPL* mutasyonlarından birisi önerilmeye çalışılmıştır.

Myeloproliferatif hastalıklar, bireyin kemik iliğinde oluşan bozukluklar olarak tanımlanmaktadır. Bu bozukluklar hücre öncüllerinin üretimi ile olgun hücre sayısında artışa neden olmakla birlikte, diğer kan hücrelerinin miktarında da artış veya azalışlara yol açmaktadır.

Myeloproliferatif hastalık ön tanısı ile kliniğe başvuru yapmış hastalar Dünya Sağlık Örgütü'nün belirlemiş olduğu kriterler doğrultusunda, klinisyen tarafından genetik tarama testlerine yönlendirilmekte ve tarama sonucuna göre hastalığın tanısı konulmaktadır. *JAK2* V617F mutasyonu ile hastalık şiddeti arasında ilişki olduğu ve bu mutasyonun pıhtılaşma oluşumunda etken olduğu birçok çalışmada değerlendirilmiştir [123]. Bu kapsamda, kliniğe başvuran hastalarda *JAK2* V617F mutasyonunun varlığı araştırılarak tanı konulmaya çalışılmaktadır. Mutasyonun durumu tespit edildikten sonra sonuç pozitif ise tanı kesinleştirilmekteyken, sonucun negatif çıkması durumunda diğer mutasyonların varlığı araştırılmaktadır.

Dünya Sağlık Örgütü tarafından önerilen ve uygulanmakta olan prosedür, başvuru yapan hastalar için hem maliyetli hem de uzun zaman gerektiren bir süreç olarak ortaya çıkmaktadır. Bu doğrultuda çalışma ile bir model kurgulanmıştır.

Çalışma kapsamında, Myeloproliferatif hastalıklarda gözlemlenen *JAK2* V617F, *JAK2* Ekzon 12, *CALR* ve *MPL* olmak üzere dört mutasyon ele alınmıştır. Hastaların tam kan sayım sonuçlarından oluşturulan model ile hangi mutasyonun varlığının araştırılması gerektiği belirlenmeye çalışılmıştır.

Analiz sonucunda elde edilen istatistiksel olarak anlamlı farklılıklar değerlendirildiğinde,

- Hemoglobin, hematokrit, eritrosit, lenfosit ve eozinofil değerleri *JAK2* mutasyonlarında yüksek seviyede saptandığı için bu hastalarda belirtilen mutasyonların incelenmesi gerektiği,
- Eritrosit dağılım genişliğinde saptanan istatistiksel anlamlılık doğrultusunda hastaların *CALR* ve *MPL* mutasyonlarına ilişkin taramaya yönlendirilmesi gerektiği

sonucuna varılmıştır. Böylece, Dünya Sağlık Örgütü'nün hastalık tanı şemasını destekleyen sonuçların saptandığı değerlendirilmiştir (Bkz. Şekil 1.1).

Sonuç olarak, çalışmada bireyin on sekiz klinik bulgusu ile bir model oluşturulmuş olup, kurgulanan model ile yapılan kestirimin klinisyene hızlı bir yönlendirme sağlayacağı değerlendirilmektedir. Bu model ile oluşturulabilecek bir yazılım programının klinisyen tarafından kullanılması durumunda kliniğe başvurmuş bireyin model kapsamında ele alınan klinik bulgularının programa girilmesi ile yönlendirilmesi sağlanacaktır. Bu yönlendirmede Dünya Sağlık Örgütü'nün tanı şemasında yer alan mutasyonlardan hangisinin varlığının araştırılması gerektiği tahmin edilebilecektir. Bu tahmin kliniğe başvurmuş bireye zaman ve maddi yükümlülüklerde kolaylık sağlayacağı ön görülmektedir.

Makine öğrenmesi yöntemi kullanılarak yapılan analizler sonucunda %76,42'lük doğruluk oranlı ile en doğru sonucu veren algoritma C5.0 olarak belirlenmiştir. Bu algorithmada yapılan kestirimlerde *JAK2* V617F mutasyonu için %94,6, *JAK2* Ekzon 12 için %31,8 ve *MPL* için %50'lik doğru kestirim elde edilerek belirtilen algoritmaların iyi bir sonuç verdiği desteklenmektedir. Bu doğrultuda kliniğe başvuran hastalara klinisyen tarafından %76'luk doğruluk oranı ile üç mutasyon taramasının duyarlılık sonuçları verilebilecek ve klinisyen tarafından öncelikle *JAK2* mutasyonunun araştırılması ile başlanarak tanı konulabileceği söylenebilecektir.

Elde edilen analiz sonuçlarına ek olarak, literatür taramasında incelenen çalışmalarda kullanılan veri setleri ve analiz sonuçları yapılan bu çalışmaya benzerlik göstermediği değerlendirmesi yapılabilmektedir. Bu sonuç yapılan bu çalışmanın özgünlüğünün yüksek seviyede olduğunun bir göstergesi olarak değerlendirilmiştir.

Çalışma kapsamında önerilen modelle elde edilen doğruluk değerinin %70 civarında olması göz önüne alındığında, çalışmanın örneklem büyüklüğünün genişletilerek kestirim yapılan mutasyon taramalarındaki örneklem sayılarındaki dengesizlik ortadan kaldırılarak tekrarlanabileceği düşünülmektedir. Böylelikle daha sonra yapılacak çalışmalarda kestirime ilişkin hesaplamalarda doğruluk değerinin yükseltilebileceği ön görülmektedir.

KAYNAKLAR

1. Bahsi, T. ve Yiğenoğlu, T. N. (2019). Myeloproliferatif neoplazilerde CALR, JAK2 ve MPL gen mutasyonlarının sıklığının ve birlikteliğinin değerlendirilmesi; Tek merkez deneyimi. *Acta Oncologica Turcica*, 52(3), 388-392.
2. Akyoli P., Yıldız A., Afacan Öztürk, H.B., Maral, S., Aras, M. R., Yılmaz, F., Sağlam, B., Tiglioğlu, M. and Albayrak, M. (2021) Disease and clinical characteristics of patients with chronic myeloproliferative neoplasms: 11-year single center experience. *ACTA Oncologica Turcica*, 54(2), 206-216.
3. Aydeniz, E.G. (2020). *Hematolojik malignetelerde obstetrik komplikasyonların yönetimi*. (Birinci Baskı). Ankara: Akademisyen Yayınevi, 213-222.
4. Türkiye Hematoloji Derneği Miyeloproliferatif Hastalıklarda Tanı ve Tedavi Kılavuzu. (2016). web: <https://thd.org.tr/BAK/15/kronik-myelositer-losemi-ve-kronik-myeloproliferatif-neoplaziler-bilimsel-a-k/calismalar/>. Son Erişim Tarihi: 04.04.2022.
5. Saeidi, K. (2016). Myeloproliferative neoplasms: Current molecular biology and genetics. *Critical Reviews in Oncology/Hematology*, 98, 375–389.
6. Michiels, J. J., Tevet, M., Trifa, A., Niculescu-Mizil, E., Lupu, A., Vladareanu, A. M., Bumbea, H., Ilea, A., Dobrea, C., Georgescu, D., Patrinoiu, O., Popescu, M., Murat, M., Dragan, C., Mihai, F., Zurac, S., Angelescu, S., Iova, A., Popa, A., Gogulescu, R. and Popov, V. (2016). WHO clinical molecular and pathological criteria for classification and staging of myeloproliferative neoplasms (MPN) caused by MPN driver mutations in the JAK2, MPL and CALR genes in the context of new 2016 WHO classification: Prognostic and therape. *MAEDICA – A Journal of Clinical Medicine*, 11(1), 5-25.
7. Duru Çetinkaya, P., Aktoğu Özkan, S., Yalnız E. ve Erer, O. F. (2009). Pulmoner venöz tromboembolili hastalarda tanı süresi ve maliyet analizi. *İzmir Göğüs Hastanesi Dergisi*, 23(1), 17-23.
8. Akalın, B. ve Veranyurt, Ü. (2020). Sağlıkta dijitalleşme ve yapay zeka. *Sağlık Yönetimi Dergisi*, 2(2), 131-141.
9. Karakoyun, M. ve Hacıbeyoğlu, M. (2014). Biyomedikal veri kümeleri ile makine öğrenmesi sınıflandırma algoritmalarının istatistiksel olarak karşılaştırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Mühendislik Bilimleri Dergisi*, 16(48), 30-41.
10. Seveli, O. (2019). Performance comparison of different machine learning techniques in diagnosis of breast cancer. *European Journal of Science and Technology*, 16, 176-185.
11. Özen, N. S., Saraç, S. and Koyuncu, M. (2021). Prediction of COVID-19 cases in the United States of America with machine learning algorithms. *European Journal of Science and Technology*, 22, 134-139.
12. Çağlıyan, B. and Köse, U. (2021). Epilepsy EEG data classification via machine learning techniques. *European Journal of Science and Technology*, 23, 163-172.

13. Bilgin, G. (2021). Makine öğrenmesi algoritmaları kullanılarak erken dönemde diyabet hastalığı riskinin araştırılması. *Zeki Sistemler Teori ve Uygulamaları Dergisi*, 4(2), 55-64.
14. Kaya, U., Yılmaz, A. and Dikmen, Y. (2019). Deep learning methods used in the field of health. *European Journal of Science and Technology*, 16, 792-808.
15. Demirarslan, M. and Suner, A. (2021). The effect of feature selection methods to classification performance in health datasets. *Journal of Artificial Intelligence in Health Sciences*, 1(1), 6-11.
16. Coşar, M. ve Deniz, E. (2021). Detection of heart diseases using machine learning algorithms. *European Journal of Science and Technology*, 28, 1112-1116.
17. Metlek, S. ve Kayaalp, K. (2020). Otistik spektrum bozukluğunun makine öğrenme algoritmaları ile tespiti. *Zeki Sistemler Teori ve Uygulamaları Dergisi*, 3(2), 60-68.
18. Bektaş, B. ve Babur, S. (2016, Kasım). *Machine learning based performance development for diagnosis of breast cancer*. TIPTEKNO: 2016 Medical Technologies National Congress; Antalya, 1-4.
19. Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., Wei, Z. ve Lei, M. (2019). Machine learning in materials science. *InfoMat*, 1(3), 338-358.
20. Bi, Q., Goodman, K. E., Kaminsky, J. ve Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American Journal of Epidemiology*, 188(12), 2222-2239.
21. Sevinç, A. S. (2021). *Makine öğrenmesi ile periodontal hastalık risk tahmini*, Yayınlanmamış Uzmanlık Tezi, Bolu Abant İzzet Baysal Üniversitesi Diş Hekimliği Fakültesi Periodontoloji Ana Bilim Dalı, Bolu.
22. Kaynar, O., Arslan H., Görmez, Y. ve Işık, Y. E. (2018). Makine öğrenmesi ve öznelilik seçim yöntemleriyle saldırı tespiti. *Bilişim Teknolojileri Dergisi*, 11(2), 175-185.
23. Er, M. B. ve Çiğ, H. (2020). Recognition of human emotions by machine learning method using Turkish music stimuli. *Gazi University Journal of Science Part C: Design Design and Technology*, 8(2), 458-474.
24. Gültepe, Y. (2019). Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme. *Avrupa Bilim ve Teknoloji Dergisi*, 16, 8-15.
25. Yiğiter, Ş. Y., Sarı, S., Karabulut, T. ve Başakın, E. E. (2018). Kira sertifikası fiyat değerlerinin makine öğrenmesi metodu ile tahmini. *Uluslar arası İslam Ekonomisi ve Finansı Araştırmalar Dergisi*, 4(3)0 74-82.
26. Elf, S., Abdelfattah, N. S., Chen, E., Perales-Patón, J., Rosen, E. A., Ko, A., Peisker, F., Florescu, N., Giannini, S., Wolach, O., Morgan, E. A., Tothova, Z., Losman, J. A., Schneider, R. K., Al-Shahrour, F. and Mullally, A. (2016). Mutant calreticulin requires both its mutant c-terminus and the thrombopoietin receptor for oncogenic transformation. *Cancer Discovery*, 6(4), 368–381.
27. Yang, Z. R. (2010). Machine learning approaches to bioinformatics. *World Scientific Publishing*. 1-15.

28. Malley, J. D., Malley, K. G. and Pajevic, S.(2011). *Statistical leearning for biomedical data*. (First Edition). Cambridge University Press, 1-13.
29. Rumi, E. and Cazzola, M. (2017). Diagnosis, risk stratification, and response evaluation in classical myeloproliferative neoplasms. *Blood*, 129(6), 680-691.
30. Haznedaroğlu, İ. C. (2006). Kronik miyeloid lösemi KML; Türk Hematoloji Derneği. *Moleküler Hematoloji Kursu*, 34-35.
31. Oğuz, E. (2014). *Aktivasyonla indüklenen sitidin deaminazın gen ekspresyon düzeyi ile kronik miyeloid löseminin progresyonu arasındaki ilişkinin araştırılması*. Tıpta Uzmanlık Tezi, İstanbul Üniversitesi İstanbul Tıp Fakültesi İç Hastalıkları Ana Bilim Dalı, Ankara, 3-7.
32. Mağın, H., Akdoğan, M. F., Uyanık, Ç., Gücün, M., Palit, F. ve Duranay, M. (2010). Polisitemia vera tanısında JAK-2 gen mutasyonunun önemi: Olgu sunumu ve literatürün gözden geçirilmesi. *Bakırköy Tıp Dergisi*, 6(1), 35-38.
33. Yönel, İ. ve Sargın, F. D. (2015). Polisitemia vera: Patogenez, teşhis ve tedavide güncel bilgiler. *İstanbul Tıp Fakültesi Dergisi*, 78(3), 83-91.
34. Bayram, M. (2011). *Polisitemia vera tanısıyla izlenen hastalarda JAK2 gen mutasyonunun klinik ve komplikasyonlarla ilişkisi*. Tıpta Uzmanlık Tezi, Uludağ Üniversitesi Tıp Fakültesi İç Hastalıkları Ana Bilim Dalı, Bursa, 2.
35. Bayram, M., Özkocaman, V., Özkalemkaş, F., Ali, R., Karkucak, M., Özçelik, T., Irmak, G., Yakut, T., Ocakoğlu, G., Aydın, T. ve Tunalı, A. (2011). Esansiyel trombositoz tanısıyla izlenen olgularda JAK-2 gen mutasyonu ve komplikasyonlarla ilişkisi. *Uludağ Üniversitesi Tıp Fakültesi Dergisi*, 37(1), 13-16.
36. Çetintepe, T., Çetintepe, L., Mutlu, C. A., Solmaz, Ş., Acar, C., Namdaroğlu, S. ve Gediz, F. (2018). Esansiyel trombositoz tanılı hastalarımızın retrospektif değerlendirilmesi. *İzmir Eğitim ve Araştırma Hastanesi Tıp Dergisi*, 22(3), 101-104.
37. Ünal, S. (2019). *Esansiyel Trombositoz ve Primer Myelofibrozisli Hastalarda Kalretikülin ve JAK-2 Mutasyonu Araştırılması*. Tıpta Uzmanlık Tezi, Gazi Üniversitesi Tıp Fakültesi İç Hastalıkları Ana Bilim Dalı, Ankara, 19-24.
38. Irmak, G. (2013). *Esansiyel trombositemi tanısıyla izlenen olgularda JAK2 gen mutasyonu, komplikasyonlar ve tedavi ilişkisinin değerlendirimi*. Tıpta Uzmanlık Tezi, Uludağ Üniversitesi Tıp Fakültesi İç Hastalıkları Ana Bilim Dalı, Bursa, 2-7.
39. Koçak, M. Z. (2015). *BCR ABL (-) Miyeloproliferatif Neoplazilerde BETA-2 Glikoprotein I Antikörünün Klinik Önemi*, Tıpta Uzmanlık Tezi, Selçuk Üniversitesi Tıp Fakültesi İç Hastalıkları Ana Bilim Dalı, Konya, 19-24.
40. Çağırğan, S. (2014, 6-8 Mart). *Primer myelofibroзда allojeneik hematopoetik kök hücre nakli*. 8. Ulusal Kemik İliği Transplantasyonu ve Kök Hücre Tedavileri Kongresi, Antalya, 60-64.
41. Öte, E. O. (2011). *JAK2 ve MPL mutasyonu negatif esansiyel trombositoz ve primer myelofibroza hastalarında kalretikülin (CARL) mutasyon durumunun belirlenmesi*. Tıpta Uzmanlık Tezi, Başkent Üniversitesi Tıp Fakültesi Tıbbi Genetik Ana Bilim Dalı, Ankara, 10-14.

42. Arber, D. D., Orazi, Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., Bloomfield, C. D., Cazzola, M. and Vardiman, J. W. (2016). The 2016 revision to the world health organization classification of myeloid neoplasms and acute leukemia. *Blood*, 127(20), 2391-2405.
43. Ebid, G. T., Ghareeb, M., Salaheldin, O. and Kamel, M. M. (2015). Prevalence of the frequency of JAK2 (V617F) mutation in different myeloproliferative disorders in Egyptian patients. *International Journal of Clinical Experimental Pathology*, 8(9), 11555-11559.
44. Gonzalez, M. S., De Brasi, C. D., Bianchini, M., Gargallo, P., Stanganelli, C., Zalcberg, I. and Larripa, I. B. (2014). Improved diagnosis of the transition to JAK2 V617F homozygosity: The key feature for predicting the evolution of myeloproliferative neoplasms. *PLoS One*, 9(1), e86401.
45. Bandaranayake, R. M., Ungureanu, D., Shan, Y., Shaw, D. E., Silvennoinen, O. and Hubbard, S. R. (2012). Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *National Structural & Molecular Biology*, 19(8), 754–759.
46. Scott, L. M., Tong, W., Levine, R. L., Scott, M. A., Beer, P. A., Stratton, M. R., Futreal, P. A., Erber, W. N., McMullin, M. F., Harrison, C. N., Warren, A. J., Gilliland, G., Lodish, H. F. and Green, A. R. (2007). JAK2 Exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. *New England Journal of Medicine*, 356(5), 459–468.
47. Passamonti, F., Elena, C. E., Schnittger, S., Skoda, R. C., Green, A. R., Girodon, F., Kiladjian, J. J., McMullin, M. F., Ruggeri, M., Besses, C., Vannucchi, A. M., Lippert, E., Gisslinger, H., Rumi, E., Lehmann, T., Ortmann, C. A., Pietra, D., Pascutto, C., Haferlach, T. and Cazzola, M. (2011). Molecular and Clinical Features of the Myeloproliferative Neoplasm Associated with JAK2 Exon 12 Mutations. *Blood* 117(10), 2813–2816.
48. Pikman, Y., Lee, B. H., Mercher, T., McDowell, E., Ebert, B. L., Gozo, M., Cuker, A., Wernig, G., Moore, S., Galinsky, I., DeAngelo, D. J., Clark, J.J., Lee, S. J., Golub, T. R., Wadleigh, M., Gilliland, D. G. and Levine, R. L. (2006) MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. *PLOS Medicine*, 3, 1140–1151.
49. Vapnik, V. (1999). *The nature of statistical learning theory*. (Second Edition). Springer Science & Business Media, 1-21.
50. Alpaydın, E. (2020). *Introduction to machine learning*. (Fourth Edition). Cambridge, Massachusetts: MIT PRESS, 1-12.
51. Gong, Y. and Xu, W. (2007). *Machine learning for multimedia content analysis*. (First Edition). Springer, New York, NY.
52. Murphy. K.P. (2012). *Machine Learning: A Probabilistic Perspective*. (First Edition). Cambridge, Massachusetts: MIT PRESS, 1-14.
53. Yu, J. and Tao, D. (2019). *Modern machine learning techniques and their applications in cartoon animation research*. (First Edition). John Wiley & Sons, 2-9.
54. Acungil, M. (2020). *Nörobilimde makine öğrenmesi ve geleceği*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul, 5-6.

55. Özdemir, R. (2021). *Makine Öğrenmesindeki sınıflandırma yöntemlerinin karşılaştırılması ve e-ticaret üzerinde bir uygulama*, Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 25-27.
56. Ertürk, Z. K. ve Ertürk, B. (2021). Acil serviste yapay zeka kullanımı. *Aksaray Üniversitesi Tıp Bilimleri Dergisi*, 2(2), 40-41.
57. Arslan, K. (2020). Eğitimde yapay zeka ve uygulamaları. *Batı Anadolu Eğitim Bilimleri Dergisi*, 11(1), 71-88.
58. Tan, F. G., Yüksel, A. S., Aydemir, E. ve Ersoy, M. (2021). Derin öğrenme teknikleri ile nesne tespiti ve takibi üzerine bir inceleme. *Avrupa Bilim ve Teknoloji Dergisi*, 25, 159-171.
59. Fix, E. and Hodges, J. L. (1989). Discriminatory analysis, nonparametric discrimination: consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3), 238-247.
60. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.
61. Anderson, J. A. (1972). A simple neural network generating on interactive memory. *Mathematical Biosciences*, 14, 197-220.
62. Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.
63. Hopfield, J. J. (1988). Artificial neural networks. *IEEE Circuits and Devices Magazine*, 4, 2.
64. Carpenter, G. A. and Grossberg, S. A. (1987). ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, 26(3), 4919-4930.
65. Fukushima, K. (1987). Neural network model for selective attention in visual pattern recognition and associative recall. *Applied Optics*, 26(23), 4985-4992.
66. Specht, D.F., (1988, 24-27 Temmuz). *Probabilistic neural networks for classification, mapping, or associative memory*. IEEE Conference on Neural Networks, San Diego, CA.
67. Specht, D.F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6), 568-576.
68. Schmidhuber, J. (1993, 2 Mart-1 Nisan). *A neural network that embeds its own meta-levels*. IEEE Conference on Neural Networks, San Diego, CA.
69. Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9, 293-300.
70. Seirawan, Y. (1997). The implications of kasparov vs. deep blue. *Communications of ACM*, 40(8), 21-25.
71. Özbay, E. (2013). *Nesnelerin üç boyutlu modellenmesi için kinect tabanlı bir uygulama*. Yüksek Lisans Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 15-16.

72. Şeker, E. (2020). Yapay zeka tekniklerinin/ uygulamalarının siber savunmada kullanımı. *Uluslararası Bilgi Güvenliği Mühendisliği Dergisi*, 6(2), 108-115.
73. Tan, F. G., Yüksel, A. S., Aydemir, E. ve Ersoy, M. (2021). Derin öğrenme teknikleri ile nesne tespiti ve takibi üzerine bir inceleme. *European Journal of Science and Technology*, 25, 159-171.
74. Anadolu, B. (2019) Dijital hikâye anlatıcılığı bağlamında yapay zekanın sinemaya etkisi: Sunspring ve it's no game filmlerinin analizi. *Erciyes İletişim Dergisi*, 1, 39-56.
75. Demirci, M. (2022). *Makine öğrenmesi yaklaşımı kullanılarak siber e-dolandırıcılık saldırılarının tespiti*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 24-25.
76. Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2008). *Bilimsel araştırma yöntemleri*. (Yirmi Beşinci Baskı). Ankara: Pegem A Yayıncılık, 129-171.
77. Batista, G. E., Prati, R. C. and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29.
78. Güldal, H. ve Çakıcı, Y. (2017). Ders yönetim sistemi yazılımı kullanıcı etkileşimlerinin sınıflandırma algoritmaları ile analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 21(4), 1355-1367.
79. Chen, S., Webb, G. I., Liu, L. and Ma, X. (2019). A novel selective naive bayes algorithm. *Knowledge Based Systems*, 192(2020), 1-12.
80. Zhang, Z. (2016). Naïve bayes classification in R. *Annals of Translational Medicine*, 4(12), 241.
81. Aizubaidi, F. Y. H. (2021). *Detect malware url using naive bayes algorithm*. Yüksek Lisans Tezi, Altınbaş Üniversitesi Lisansüstü Çalışmalar Enstitüsü, İstanbul, 39-42.
82. Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015). Heart diseases detection using naive bayes algorithm. *International Journal of Innovative Science, Engineering and Technology*, 2(9), 441-444.
83. Şaylan, Ç.A. (2013). *Böbrek Nakli Geçirmiş Hastalarda Akıllı Yöntem Tabanlı Yeni Öznitelik Seçme Algoritması Geliştirilmesi*, Yüksek Lisans Tezi, Kadir Has Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 23.
84. Güldal, H. ve Çakıcı Y. (2017). Ders yönetim sistemi yazılımı kullanıcı etkileşimlerinin sınıflandırma algoritmaları ile analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 21(4), 1355-1367.
85. Mulak, P. ve Talhar, N. (2015). Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset. *International Journal of Science and Research*, 4(7), 2101-2104.
86. Konieczny, J. and Stojek, J. (2021) Use of the k-nearest neighbour classifier in wear condition classification of a positive displacement pump. *Sensors (Basel)*, 21(18), 6247.

87. Türk, Ö. (2020). Determination of emotional status from EEG time series by using EMD based local binary pattern method. *European Journal of Technique*, 10(2), 313-321.
88. Köktürk, F. (2012). *k-en yakın komşuluk, yapay sinir ağları ve karar ağaçları yöntemlerinin sınıflandırma başarılarının karşılaştırılması*. Bülent Ecevit Üniversitesi Sağlık Bilimleri Enstitüsü, Zonguldak, 15-20.
89. Karakoyun, M. ve Hacıbeyoğlu, M. (2014). Biyomedikal veri kümeleri ile makine öğrenmesi sınıflandırma algoritmalarının istatistiksel olarak karşılaştırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 16(48), 30-42.
90. Hu, L. Y., Huang, M. W., Ke, S. W. and Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *Springerplus*, 5(1), 1304.
91. Bo, S., Junping, D. and Tian, G. (2009, Şubat 16-18). *Study on the improvement of k-nearest -neighbor algorithm*. 2009 International Conference on Artificial Intelligence and Computational Intelligence, Avusturya.
92. Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *In Annals of Eugenics*, 7(2), 179-188.
93. Torkkola, K. (2001, Kasım). *Linear discriminant analysis in document classification*. The 2001 IEEE International Conference on Data Mining, California, USA.
94. Filiz, Z. ve Yaprak, B. (2009). A study on classifying parenting styles through discriminant analysis. *Eğitimde Kuram ve Uygulama Dergisi*, 5(2), 195-209.
95. Ye, Q., Fu, L., Zhang, Z., Zhao, H and Naiem, M. (2018) Lp- and Ls-norm distance based robust linear discriminant analysis. *Neural Networks : The Official Journal of The International Neural Network Society*, 105, 393-404.
96. Göldoğan, E., Yağmur, J., Yoloğlu, S., Asyalı, M. H. ve Çolak, C. (2015). Myocardial infarction classification with support vector machine. *Journal of Turgut Ozal Medical Center*, 22(4), 221-224.
97. Yakut, E., Elmas, B. ve Yavuz, S. (2014). Yapay sinir ağları ve destek vektör makineleri yöntemleriyle borsa endeksi tahmini. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 19(1), 139-157.
98. Hsu, C., Chen, M. C. and Chen, L. S. (2010). Intelligent ica-svm fault detector for non-gaussian multivariate process monitoring. *Expert Systems with Applications*, 37(4), 3264-3273.
99. Tolun, S. (2008). *Destek vektör makineleri: Banka başarısızlığının tahmini üzerine bir uygulama*. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul, 72-84.
100. Gilmore, G. G., Hersh, H., Caramazza, A. and Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception and Psychophysics*, 25(5), 425-431.
101. Rozenberg, G., Back, T. and Kok, J. N. (2012). *Handbook of Natural Computing*. (First Edition), Berlin: Springer, 479-506.

102. Demirel, Ş. ve Yakut, S. G. (2019). Karar ağacı algoritmaları ve çocuk işçiliği üzerine bir uygulama. *Social Sciences Research Journal*, 8(4), 52-65.
103. Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40, 229-242.
104. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81-106.
105. Alan, M. A. (2014). Karar ağaçlarıyla öğrenci verilerinin sınıflandırılması. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 28(4), 101-112.
106. Chern, C. C., Chen, Y. J. and Hsiao, B. (2019) Decision tree-based classifier in providing telehealth service. *BMC Medical Informatics and Decision Making*, 19(1), 104.
107. Başbağ, M., Çağan, E., Sayar, M. ve Fırat, M. (2020). *Veri madenciliği yöntemleri tarım alanında uygulamaları*. (Birinci Baskı), Çanakkale: Rating Academy Yayınları, 115-119.
108. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*. (Birinci Baskı), Routledge: New York, 38-40.
109. Yakut, E. ve, Korkmaz A. (2020). İnsani gelişmişlik endeksinin karar ağacı algoritmaları ile modellenmesi: BM’de bir uygulama 2010-2017 dönemi. *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, 20(2), 65-84.
110. Wu, M.T. (2022) Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom. *Scientific Reports*, 12(1), 3095.
111. Dağ, O. (2018). *Binary classification via GMDH-type neural network algorithm*. Doktora Tezi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara, 14-18.
112. Alan, A. ve Karabatak M. (2020). Veri seti - sınıflandırma ilişkisinde performansa etki eden faktörlerin değerlendirilmesi. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 32(2): 531-540.
113. Cesur, S., Kınıklı, S., Hatipoğlu, Ç. A., Arslan, K., Karakök, T. ve Demircan, Ş. A. (2019). Vankomisine dirençli enterokokların saptanmasında iki farklı kromojenik besiyerinin karşılaştırılması. *Turkish Journal of Clinics and Laboratory*, 10(3): 319-323.
114. Erdaş, Ç. B. ve Sümer, E. (2020, Kasım 19-20). *Nörodejeneratif hastalıkların tespiti için derin öğrenme tabanlı bir yaklaşım*. Tıprtekno’20: Tıp Teknolojileri Kongresi; Online, 320-323.
115. İnternet: Declaration of Helsinki. (2001) *Bulletin of the World Health Organization*, 79 (4). Web: <https://apps.who.int/iris/bitstream/handle/10665/268312/PMC2566407.pdf?sequence=1&isAllowed=y> adresinden 17 Haziran 2022’de alınmıştır.
116. Uysal, A., Altın, Ş., Uysal, S. ve Çebi, A. H. (2020). Genetic analysis of BCR-ABL negative chronic myeloproliferative diseases at initial diagnosis and their clinical effects. *Çukurova Medical Journal*, 45(3), 933-936.

117. Zulkeflee, R. H., Zulkafli, Z., Johan, M. F., Husin, A., İslam, A. and Hassan, R. (2021). Clinical and laboratory features of JAK2 V617F, CALR, and MPL mutations in Malaysian patients with classical myeloproliferative neoplasm (MPN). *International Journal of Environmental Research and Public Health*, 7582(18), 1-14.
118. Li, M. Y., Chao, H. Y., Sun, A. N., Qiu, H. Y., Jin, Z. M., Tang, X. W., Han, Y., Fu, C. C., Chen S. N. and Wu, D. P. (2017). Clinical significance of JAK2, CALR and MPL gene mutations in 1 648 Philadelphia chromosome negative myeloproliferative neoplasms patients from a single center. *Zhonghua xue ye xue za zhi = Zhonghua Xueyexue Zazhi*, 38(4), 295-300.
119. Lang, T., Nie, Y., Wang, Z., Huang, Q., An, L. and Wang, Y. (2018). Correlation analysis between JAK2, MPL, and CALR mutations in patients with myeloproliferative neoplasms of Chinese Uygur and Han nationality and their clinical characteristics. *Journal of International Medical Research*, 46(11), 4650-4659.
120. Passamonti, F., Elena, C., Schnittger, S., Skoda, R. C., Green, A. R., Girodon, F., Kiladjian, J. J., McMullin, M. F., Ruggeri, M., Besses, C., Vannucchi, A. M., Lippert, E., Gisslinger, H., Rumi, E., Lehmann, T., Ortmann, C. A., Pietra, D., Pascutto, C., Haferlach, T. and Cazzola, M. (2011). Molecular and clinical features of the myeloproliferative neoplasm associated with JAK2 exon 12 mutations. *Blood*, 117(10), 2813-2816.
121. Lin, Y., Liu, E., M, J., Li, Q., Cao, Z., Wang, J., Jia, Y., Zhang, H., Song, Z., Ai, X., Shi, L., Feng, X., Li, C., Wang, J. and Ru, K. (2015). The Prevalence of JAK2, MPL, and CALR mutations in Chinese patients with BCR-ABL1-negative myeloproliferative neoplasms. *American Journal of Clinical Pathology*, 144(1), 165-171.
122. Soliman, E. A., El-Ghlban, S., El-Aziz, S. A., Abdelaleem, A., Shamaa, S. and Abdel-Ghaffar, H. (2020). JAK2, CALR, and MPL mutations in Egyptian patients with classic Philadelphia-negative myeloproliferative neoplasms. *Clinical Lymphoma Myeloma and Leukemia*, 20(10), 645-651.
123. Akgün Çağlayan, G. (2015). *Myeloproliferatif hastalıklarda JAK2V617F mutasyonu ile trombosit fonksiyonları arasındaki ilişkinin araştırılması*. Tıpta Uzmanlık Tezi, Pamukkale Üniversitesi Tıp Fakültesi İç Hastalıkları, Ankara, 10-13.

EKLER

Çizelge 1.1. Çalışmada kullanılan klinik bulgulara ilişkin açıklamalar

EK-1. Klinik bulgulara ilişkin açıklamalar

Klinik Bulgu	Referans Aralık	Birim	Açıklama
Eritropoietin	1-25	mU/ml	Eritropoietin (EPO) hipoksiye (oksijen yetersizliğine) yanıt olarak büyük oranda böbreklerde üretilen ve kırmızı kan hücresi üretimini uyaran hormondur. Böbrek dışında karaciğer ve doku makrofaj hücrelerinde üretilabilmektedir.
Ferritin	20-200*	ml/ng	Ferritin ise vücutta bulunan demirden farklı olarak, var olan demiri depolayan ve salınımını kontrol eden bir tür protein kompleksidir.
	20-500**		
Homosistein	5-17	Mmol/L	Homosistein bedeninizde methiyonin aminoasidinden sentezlenen bir madde. Genelde et ürünleri, süt ürünleri ve yumurtada oldukça fazla bulunan metiyonin aminoasidini bitkisel kaynaklar da az miktarda içerebilir.
Paratiroid	<6	pikogram	Paratiroid, boyunda tiroid bezinin hemen bitişiğinde bulunan, parathormon isimli hormonu üreten bir bezdir. Paratiroid bezleri, kanda ve kemiklerdeki kalsiyum ile fosfor miktarını kontrol altında tutmaktan sorumludur.
AFP	10-150	ng/mL	Alfa fetoprotein (AFP), fetusun karaciğerinde ve mide-barsak yolunda ve erişkinlerde bazı anormal (ör, tümörlü dokularda) üretilen bir proteindir.
CEA	0 – 2,5	ng/mL	Bazı hücrelerin yüzeyinde bulunan bir maddedir. Embriyonik gelişim sırasında gastrointestinal sistemin hücreleri tarafından üretilen bir glikoproteindir.
CA15	<30	U/ mL	Normal göğüs hücreleri tarafından üretilen bir proteini ve tümör belirteçidir.
CA19	<37	U/ mL	Başta pankreas olmak üzere çeşitli kanserlerde yükselen bir tümör belirteçidir.
PSA	4	ng/ml	PSA (prostat spesifik antijen), prostat bezinden salgılanan ve meninin kıvamını sağlayan bir enzimdir.
Serbest PSA	0-6,5	ng/ml	Kandaki serbest prostat spesifik antijen seviyesidir.
PSA Oranı	<25	%	Serbest PSA/ PSA

EK-1 (devam). Klinik bulgulara ilişkin açıklamalar

Klinik Bulgu	Referans Aralık	Birim	Açıklama
Hemoglobin	12-16	g/ dl	Kırmızı kan hücrelerinde bulunan ve oksijenin vücuda taşınmasını sağlayan çok önemli bir proteindir.
Hematokrit	35-46	%	Kırmızı kan hücrelerinin dolaşımdaki kan miktarına göre hacminin oranını ifade eder.
Eritrosit	4-5,2	milyon/mm ³	Kemik iliğinde üretilen ve görevi vücudun her noktasına kan taşımak olan kırmızı kan hücreleridir.
Lökosit	4,5-11	mcl	Yabancı maddelerle savaşan, akyuvar olarak bilinen beyaz kan hücreleridir.
Trombosit	150-400	hücre/mL	Kan pıhtılaşmasını hızlandıran hücre parçalarıdır.
MCH	26-32	pg	Tek bir kırmızı kan hücresinde bulunan ortalama hemoglobin miktarına karşılık gelir.
MCV	80-96	fl	Kırmızı kan hücrelerinizin ortalama büyüklüğünün bir ölçüsüdür.
MCHC	30-36	g/dl	Tek bir alyuvarda birim hacim başına hemoglobin miktarının bir hesaplamasıdır.
RDW	8-18	fl	Kırmızı kan hücrelerinin büyüklüğündeki ve hacmindeki değişiklikleri gösterir.
PDW	9-14	fl	Kandaki trombositlerin büyüklüğünün dağılımına dair bir ölçüdür.
MPV	7-12	fl	Kanın pıhtılaşmasına yarayan elemanların yaklaşık boyutlarını vermektedir.
Nötrofil	2-7,8	mcL	Bağışıklık sisteminin önemli parçası olan beyaz kan hücrelerinin bir türüdür.
Lenfosit	1-4	mcL	Vücudun ana bağışıklık hücre türlerinden bir tanesidir.
Monosit	0-1	mcL	Bakteri, virüs ve mantarlarla savaşan bir tür beyaz kan hücresidir.
Eozinofil	0-1	mcL	Bağışıklık sistemimizde görev alan bir hücre çeşididir.
Bazofil	0-0,2	mcL	Akyuvar hücresidir.



GAZİ GELECEKTİR..