



**AĞAÇ TABANLI MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE
GENOMİK VERİLERİN SINIFLANDIRILMASINA YÖNELİK
KLİNİK KARAR DESTEK SİSTEMİNİN GELİŞTİRİLMESİ**

İpek BALIKÇI ÇİÇEK

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

**Tez Danışmanı
Prof. Dr. Saim YOLOĞLU**

Doktora Tezi- 2023

T.C.
İNÖNÜ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

AĞAÇ TABANLI MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE GENOMİK
VERİLERİN SINIFLANDIRILMASINA YÖNELİK KLİNİK KARAR DESTEK
SİSTEMİNİN GELİŞTİRİLMESİ

İpek BALIKÇI ÇİÇEK

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı
Doktora Tezi

Tez Danışmanı
Prof. Dr. Saim YOLOĞLU

MALATYA

2023

T.C. İNÖNÜ ÜNİVERSİTESİ
Sağlık Bilimleri Enstitüsü Müdürlüğüne

ETİK BEYANI

İnönü Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak “Prof. Dr. Saim YOLOĞLU” danışmanlığında hazırlayıp sunduğum “Ağaç Tabanlı Makine Öğrenmesi Yöntemleri İle Genomik Verilerin Sınıflandırılmasına Yönelik Klinik Karar Destek Sisteminin Geliştirilmesi” başlıklı Doktora tezim içinde elde ettiğim verileri, bilgileri, belgeleri akademik ve etik kurallar çerçevesinde elde ettiğimi, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tezimde yararlandığım eserlere bilimsel kurallara uygun atıfta bulunarak kaynak gösterdiğimi, tezimin özgün olduğunu, tezimin çalışma ve yazımında patent ve telif haklarını ihlal edici bir davranışımın olmadığını, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim./...../20...

İpek BALIKÇI ÇİÇEK

İÇİNDEKİLER

ÖZET.....	vii
ABSTRACT	viii
SİMGELER VE KISALTMALAR DİZİNİ.....	ix
ŞEKİLLER DİZİNİ.....	x
TABLolar DİZİNİ	xi
1. GİRİŞ	1
2. GENEL BİLGİLER	4
2.1. Böbrek	4
2.2. Toksisite	5
2.2.1. Nefrotoksisite	5
2.3. Metotreksat (MTX)	5
2.3.1. Metotreksatın Böbrek Toksisitesi	6
2.4. Genomik.....	7
2.4.1. Yapısal genomik.....	9
2.4.2. Fonksiyonel genomik	10
2.4.3. Polimeraz zincir reaksiyonu (PCR).....	10
2.4.4. Denatürleyici Kademeli Jel Elektroforezi (DGGE)	11
2.4.5. Southern Blotlama.....	12
2.5. DNA Dizileme Analizi.....	13
2.5.1. Yeni Nesil Dizileme Teknolojisi (YND)	14
2.5.2. Mikroçip (Microarray)	15
2.5.3. RNA Dizileme Analizi	16
2.5.4. Transkriptomik.....	17
2.5.5. Kodlamayan RNA'lar	19
2.5.6. Uzun Kodlanamayan RNA'lar (LncRNA).....	20
2.5.7. LncRNA'ların Sınıflandırılması.....	21

2.5.8. LncRNA’ların İşlevleri	22
2.5.9. LncRNA’ların Hastalıklarla İlişkisi	23
2.6. Biyoinformatik	24
2.7. Diferansiyel Eksprese Edilmiş olan Genlerin Belirlenmesi	24
2.8. Yapay Zekâ, Makine Öğrenmesi ve Ağaç Tabanlı Yöntemler	25
3. MATERYAL VE METOT	27
3.1. Veri Seti	27
3.2. Veri Tabanlarında Bilgi Keşfi Süreci (VTBK) ve Tıbbi Bilgi Keşfi Süreci (TBKS)	27
3.3. İş süreçlerinin anlaşılması	29
3.4. Verinin anlaşılması.....	30
3.5. Veri önışleme aşaması.....	30
3.5.1. Değişken seçimi	30
3.5.2. Lasso	32
3.5.3. Elastik net.....	32
3.5.4. Boruta.....	33
3.6. Model aşaması.....	34
3.6.1. Sınıflandırma yöntemleri	34
3.6.2. Ağaç Tabanlı Topluluk Öğrenme Yöntemleri	35
3.6.3. Karar Ağaçları	35
3.6.4. Random Forest (RF).....	36
3.6.5. Arttırma (boosting) algoritması.....	37
3.6.6. AdaBoost algoritması.....	38
3.6.7. Gradient Boosting algoritması	399
3.6.8. XGBoost (Extreme Gradient Boosting) algoritması	39
3.6.9. Torbalı Sınıflandırma ve Regresyon Ağaçları (Bagged CART).....	40
3.6.10. Açıklanabilir Yapay Zeka Yöntemleri	41

3.6.11. Yerel Yorumlanabilir Model-Agnostik Açıklama	41
3.6.12. Shapley Katkı Açıklamaları	42
3.7. Veri analizi	43
3.7.1. Biyoinformatik analizler	43
3.7.2. Biyoistatistiksel analizler	43
3.8. Değerlendirme aşaması	44
3.9. Ürün aşaması	44
3.9.1. Geliştirilen Web Tabanlı Yazılım	44
3.9.2. “Giriş” Sekmesi.....	45
3.9.3. “Dosya Yükleme” Sekmesi.....	44
3.9.4. “Grafik” Sekmesi	44
3.9.5. “Açıklama” ve “Alıntılama” Sekmesi.....	44
4. BULGULAR	49
4.1. Biyoinformatik Analiz Sonuçları	49
4.2. Biyoistatistik Analiz ve Modelleme Sonuçları	52
5. TARTIŞMA	63
6. SONUÇ VE ÖNERİLER	67
KAYNAKLAR	68
EKLER.....	68
EK-1. Özgeçmiş	78
EK-2. Etik Kurul Onayı	79

TEŞEKKÜR

Bu çalışmanın planlanması, yürütülmesi sürecinde benden desteğini esirgemeyen, lisansüstü eğitimim boyunca, bilgi ve hoşgörülerinden yararlandığım bölüm kurucumuz aynı zamanda danışman hocam Sayın Prof. Dr. Saim YOLOĞLU'na sonsuz teşekkürlerimi sunarım. Lisansüstü eğitimim boyunca desteğini hiç esirgemeyen ve her zaman deneyimlerini paylaşarak bana ışık tutan değerli hocam Sayın Prof. Dr. Cemil ÇOLAK' a ve bu tez çalışmasının izleme raporlarında değerli görüşleriyle çalışmama katkı sağlayan Sayın Prof. Dr. Hakan PARLAKPINAR'a teşekkürlerimi sunarım. Ayrıca anabilim dalımızdaki diğer çok kıymetli hocalarım Doç. Dr. Harika G. G. BAĞ, Doç. Dr. Emek GÜLDOĞAN ve Dr. Öğr. Üyesi Ahmet Kadir ARSLAN'a çok teşekkür ederim. Ayrıca 17 yıldır her anımda yanımda olup destekleyen canım arkadaşım, kardeşim ve aynı zamanda aynı bölümde görev yaptığım kıymetli asistan arkadaşım Arş. Gör. Zeynep KÜÇÜKAKÇALI'ya, aynı anabilim dalında görev yaptığım bir diğer asistan arkadaşım ama sadece mesai ile kalmayıp her anımda yanımda olan, kızarken bile beni sevebilen, kızına da benim ismini koyan can dostum, arkadaşım Arş. Gör. Şeyma YAŞAR'a ve son olarak destekleri için asistan arkadaşım Fatma Hilal YAĞIN'a teşekkür ederim.

Bu tez çalışmamda yardımını hiç esirgemeyen, destekleriyle hayatımın her alanında olduğu gibi bu süreçte de beni hiçbir zaman yalnız bırakmayan sevgili eşim Can ÇİÇEK ile kızım Asya Umay ÇİÇEK'e sonsuz teşekkür ederim. Son olarak her zaman yanımda olan sevgili kardeşlerime ve beni yetiştirip bugünlere gelmemi sağlayan sevgili anne ve babama sonsuz teşekkürlerimi sunarım.

Arş. Gör. İpek BALIKÇI ÇİÇEK

ÖZET

Ağaç Tabanlı Makine Öğrenmesi Yöntemleri İle Genomik Verilerin Sınıflandırılmasına Yönelik Klinik Karar Destek Sisteminin Geliştirilmesi

Amaç: Bu çalışmanın amacı, MTX ile nefrotoksitesi olan ve patolojisi olmayan sıçanlardan alınan böbrek doku örneklerinin genomik analizleri sonucu elde edilen büyük verilerin biyoinformatik analizlerinin yapılması ve sonuçların görselleştirilmesi amacıyla bir yazılım geliştirilmesidir. Ayrıca verilerin makine öğrenmesi yöntemlerinden ağaç tabanlı öğrenme yöntemleri ile modellenmesi ve açıklanabilir yapay zeka yöntemleri ile modelin yorumlanabilirliğini sağlayarak nefrotoksite tanısına yönelik olası biyobelirteçlerin belirlenmesi amaçlanmıştır.

Materyal ve Metot: Bu çalışmada 20 adet dişi Wistar Albino cinsi sıçan alınarak oluşturulmuş bir deney düzeneğinden elde edilen genomik veriler kullanılmıştır. Biyoinformatik analiz yapmak için R programlama dili kullanıldı. Modellerde ağaç tabanlı makine öğrenmesi yöntemlerinden karar ağaçları, Random Forest, AdaBoost, Baged CART ve XGBoost modelleri kullanılmıştır. XGBoost modelinin yorumlanabilirliğini artırmak için Local Interpretable Model-Agnostic Explanations ve SHapley Additive Explanations yöntemleri kullanılmıştır. Modellerin ve açıklanabilir yapay zeka yöntemlerinin analizinde Python programlama dili kullanılmıştır.

Bulgular: Çalışmada kullanılan genomik veri seti 16.386 lncRNA ifadesini içermektedir. Biyoinformatik analiz sonuçlarına göre nefrotoksite ve kontrol grupları için lncRNA ların 35 tanesi yukarı regülasyon göstermiş iken 17 tanesi aşağı regülasyon göstermiştir. Boruta değişken seçimi ile seçilmiş olan lncRNA lar ile yapılan modellemeler sonucunda performans ölçütlerine göre XGBoost yöntemi en başarılı makine öğrenimi yöntemi olmuştur. SHAP'ın bir sonucu olarak, Nefrotoksite için öngörücü biyobelirteç adaylarından en önemli ilk üçü ma-XR_591534.3 (LOC103691816), ma-XR_351582.4 (LOC102555118), ma-XR_005499541.1 (LOC120099962) idi.

Sonuç: Bu çalışma yapılan biyoinformatik analiz, modellemeler ve modelleme yorumlanabilirliği sonucunda nefrotoksitesi olan sıçanlar ile kontrol grubunda yer alan sıçanların lncRNA ekspresyon verileri kullanılarak nefrotoksite için olası genomik biyobelirteçleri belirlenmiştir.

Anahtar Kelimeler: Açıklanabilir Yapay Zeka, Ağaç Tabanlı Öğrenme, Karar Destek Sistemi, Genomik, Makine Öğrenmesi, Sınıflandırma, Yapay Zekâ

ABSTRACT

Development of a Clinical Decision Support System for Classification of Genomic Data with Tree-Based Machine Learning Methods

Aim: The aim of this study is to develop a software for bioinformatic analysis of large data obtained as a result of genomic analysis of kidney tissue samples taken from rats with nephrotoxicity and without pathology with MTX and to visualize the results. In addition, it was aimed to model the data with tree-based learning methods, one of the machine learning methods, and to determine possible biomarkers for the diagnosis of nephrotoxicity by providing the interpretability of the model with explainable artificial intelligence methods.

Material and Method: In this study, genomic data obtained from an experimental setup created by taking 20 female Wistar Albino rats were used. R programming language was used to perform bioinformatic analysis. Decision trees, Random Forest, AdaBoost, Bagged CART and XGBoost models from tree-based machine learning methods were used in the models. Local Interpretable Model-Agnostic Explanations and SHapley Additive Explanations methods were used to improve the interpretability of the XGBoost model. Python programming language was used in the analysis of models and explainable artificial intelligence methods.

Results: The genomic dataset used in the study includes 16,386 lncRNA expressions. According to the results of bioinformatics analysis, 35 of the lncRNAs for nephrotoxicity and control groups showed up-expression, while 17 of them showed down-expression. As a result of the models made with lncRNAs selected by Boruta variable selection, the XGBoost method has been the most successful machine learning method according to performance criteria. As a result of SHAP, the top three most important candidates for predictive biomarkers for Nephrotoxicity were RNA-XR_591534.3 (LOC103691816), RNA-XR_351582.4 (LOC102555118), RNA-XR_005499541.1 (LOC120099962).

Conclusion: As a result of the bioinformatic analysis, models and modeling interpretability performed in this study, possible genomic biomarkers for nephrotoxicity were determined by using lncRNA expression data of rats with nephrotoxicity and rats in the control group.

Keywords: Explainable Artificial Intelligence, Tree-Based Learning, Decision Support System, Genomics, Machine Learning, Classification, Artificial Intelligence

SİMGELER VE KISALTMALAR DİZİNİ

ABY	:Akut Böbrek Yetmezliği
cDNA	:Hedef DNA veya complementary DNA
DGGE	: Denatüran Gradyent Jel Elektroforezi
LIME	:Yerel Yorumlanabilir Model-Agnostik Açıklama
lncRNA	:Long non-coding RNA
log2FC	:log2-kat değişim
MTX	:Metotreksat
ncRNA	:non-coding RNA
PCR	:Polimeraz zincir reaksiyonunun
RF	:Random Forest
SHAP	:Shapley Katkı Açıklamaları
VTBK	:Veri Tabanlarında Bilgi Keşfi Süreci
XGBoost	:Extreme Gradient Boosting
YND	:Yeni nesil DNA dizileme yöntemleri

ŞEKİLLER DİZİNİ

Şekil No	Sayfa No
Şekil 2.1. Omik teknolojileri ve aralarındaki ilişki	8
Şekil 2.2. Polimeraz zincir reaksiyonun basamakları.....	11
Şekil 2.3. Denatürleyici kademeli jel elektroforezinin aşamaları.....	12
Şekil 2.4. YND'nin çalışma prensibi.....	14
Şekil 2.5. cDNA için Mikroçip yönteminin çalışma şekli.....	16
Şekil 2.6. Kodlanamayan RNA'ların sınıflandırılması	20
Şekil 2.7. LncRNA'ların genomdaki sentez edildiği bölgelere göre sınıflandırılması ..	22
Şekil 2.8. LncRNA'lara ilişkin hücresel işlevler.....	23
Şekil 2.9. Yapay zekâ, makine öğrenmesi ve ağaç tabanlı yöntemlere ilişkin hiyerarşi	26
Şekil 3.1. VTBK süreci.....	28
Şekil 3.2. Özellik seçimi yardımıyla bir özellik alt kümesinin seçimi	31
Şekil 3.3. Boosting Akış Diyagramı	38
Şekil 3.4. Gradient Boosting Adımları	39
Şekil 3.5. XGBoost'un avantajları.....	40
Şekil 3.6. “Giriş” Sekmesi.....	40
Şekil 3.7. “Dosya Yükleme” Sekmesi	40
Şekil 3.8. “Grafik” Sekmesi	40
Şekil 4.1. Nefrotoksisitesinde olan sıçanlar ile kontrol grubu sıçanların arasında tanımlanan farklı şekilde ekspresyon gösteren lncRNA' lar için volkan grafiği	51
Şekil 4.2. En Fazla Değişiklik Gösteren lncRNA ifadelerinden 50 tanesi ısı haritası ...	52
Şekil 4.3. XGBoost modeline ilişkin performans metriklerinin grafiği	56
Şekil 4.4. Nefrotoksisite grubundaki ilk sıçan için LIME sonuçları	57
Şekil 4.5. Nefrotoksisite grubundaki ikinci sıçan için LIME sonuçları	58
Şekil 4.6. Nefrotoksisite grubundaki üçüncü sıçan için LIME sonuçları	59
Şekil 4.7. Normalize edilmiş SHAP değerlerine göre önemlilik sıralamaları.....	60
Şekil 4.8. SHAP grafiği	61

TABLÖLAR DİZİNİ

Tablo No	Sayfa No
Tablo 4.1. Biyoinformatik Analiz Sonuçları.....	499
Tablo 4.2. İstatistik Analiz Sonuçları.....	53
Tablo 4.3. Çalışmada kullanılan makine öğrenimi modellerinin performans ölçütleri sonuçları (%95 güven aralığı (GA))	55



1. GİRİŞ

Günümüzde kanser vakalarının artmasından dolayı kemoterapötik ajanların kullanımı fazla oranda artmıştır. Folat antagonisti olduğu bilinen bir antimetabolik ajan olan metotreksat (MTX) hücrenin çoğalma fazında hücreye etki edip hücre bölünmesini engellemektedir. MTX lenfoma, lösemi, osteosarkom, akciğer kanseri, boyun ve baş tümörleri, kanser türlerinde de kemoterapötik ajan olarak kullanılmaktadır. Ayrıca çok uzun zamandır romatoid artrit, psöriatik artrit, neoplazmalar, lösemiler, otoimmün hastalıkların tedavi edilmesinde yer almıştır (1). Günümüzde ise vaskülitlerde, inflamatuvar bağırsak hastalıklarında, bağ dokusu hastalıklarında, tıbbi düşüklerin tedavisinde de tek başına veya kombine bir şekilde verilebilen antimetabolit grubunda yer alan bir ilaçtır (2). MTX'in tedavi amacıyla kullanımı hayati organların üzerindeki toksik etkileri de beraberinde getirmektedir. Psöriazis ya da akut lösemi gibi hastalıkların tedavisi için kullanılmakta olan yüksek dozdaki MTX'in karaciğerde siroza sebep olabilecek kadar hepatik toksisite ile ilişkili olabildiği bildirilmiştir (3). Ayrıca MTX'in toksisitelerinin başlıca görüldüğü yerler bağırsak lümeni, karaciğer, hücre yenilenmesinin çok olduğu kemik iliği ve böbrek dokusu olduğu klinik çalışmalarda ve deneysel çalışmalarda belirtilmektedir (4). MTX'in metabolize olduğu yer karaciğer ve metabolitlerinin ise %90'ı böbrek aracılığıyla atıldığından karaciğerde ve böbrekte ciddi hasarlara sebep olabilmektedir (5).

Böbrekler vücudumuzda birçok görevi yerine getirdiği için hayati öneme sahiptirler. Elektrolit ve su dengesini sağlanmasında, asit ile baz dengesinin düzenlenmesinde, kan basıncının düzenlenmesinde, vücudun eritrosit yapımında, D3 vitamininin aktifleşmesinde, ilaçların ve toksinlerin vücuttan atılabilmesinde böbreklerin çok önemli görevleri vardır (6). Böbrekler vücuttan zararlı maddelerin uzaklaştırılması ve aktif taşıma görevinden dolayı birçok ilacın ve toksik maddenin hedef organı durumundadır (7). Bundan dolayı, ilaçla tedavi edilen hastaların % 20'den fazlasında görülen ilaç toksisitesi sonucu akut böbrek hasarı oluşmaktadır. İlaça bağlı olarak ortaya çıkan nefrotoksisite böbrek yetmezliği sebebiyle ölüme yol açabilmektedir. Genellikle MTX'in uzun süreli ve yüksek dozda kullanımı hayati organlar üzerindeki toksik etkilerinin yanında kişiyi nefrotoksiste açısından da olumsuz etkilemektedir. Yapılmış

olan alıřmalarda ilaca baėlı gerekleřen nefrotoksisite durumunda hcre lmnden dolayı ciddi bbrek hasarı durumunun ortaya ıktıėı bildirilmiřtir (8).

Son zamanlarda hastalıkların belirlenmesinde genomik veriler olduka fazla kullanılmakta olup bu veriler kullanılarak hastalıkların gen boyutunda incelenip kiřisel profillerin oluřturulması, kiřisel olarak tedavi ve yan tesirlerin dzenlenmesinde de olduka nemli bir etkiye sahip olmuřlardır. İlaça baėlı olan gen deėiřikliklerinin belirlenebilmesi ilaca baėlı olarak meydana gelen bbrek hasarının saptanabilmesi iin kritik bir neme sahiptir. İlaça baėlı ortaya ıkan nefrotoksisite ile iliřkili olan genleri belirlemek iin birok alıřma yapılmıřtır. Ancak ilaca baėlı olarak nefrotoksisite iin genomik alıřmaları olmakla beraber long non-coding RNA'lar (lncRNA) ile iliřkili olan alıřmalara pek rastlanmamaktadır. Bu yzden bu alandaki eksiklikleri giderebilmek adına literatrde bu tr alıřmalara ihtiya doėmuřtur (9).

Saėlık sektr ierisinde tanı ve teřhisi yalnızca insan gc ile ynetilmek mmkn olmamaktadır. Hastalıkların nlenebilmesi, bilgi eksikliėi nedeniyle kaynaklanan personellerin hatalarını azaltabilmenin yanı sıra, hastalıkların doėru řekilde ve zamanında teřhis edilip tedavi edilmesi iin yapay zek teknolojileri kullanmaktadır (10). Yapay zekanın bir alt dalı olan makine ėrenmesinde veriye dayalı olarak ėrenmeyi gerekleřtirerek yeni verilerle karřılařtıėında yeni veriye iliřkin tahmin yapabilmeyi amalamaktadır. Makine ėrenimi sistemleri, karar verebilme becerisinde insan kararına baėlı olan ihtiyacı ortadan kaldırmayı veya insan-makine arasındaki iřbirliėine dayalı olarak elde edebilmeyi amalar (11). Saėlık alanına makine ėrenmesinin girmesiyle tıbbi kayıtların saklanması, tıbbi grntleme, tıbbi aıdan karar verme, ila geliřtirebilme, erken tanı ve tedavi ve daha pek ok saėlık konularında nemli geliřmelere katkıda bulunduėu bilinmektedir (12). Makine ėreniminin bir alt alanı olan aėa tabanlı makine ėrenme yntemlerinden Extreme Gradient Boosting (XGBoost) daha gl bir sınıflandırıcı meydana getirebilmek iin bir dizi zayıf sınıflandırıcıyı birleřtirerek glendirmeyi kullanan bir topluluk yaklařımıdır (13). Temel bir ėrenciden bařlayarak, gl ėrenci yinelemeli olarak eėitilir. Diėer makine ėrenimi algoritmalarına gre yaklařık 10 kat daha hızlı iřlem yapma kapasitesine sahip olup, hız ve performans aısından olduka avantajlı bir konuma sahiptir. Ek olarak, genel performansı artıran ve fazla uydurmayı veya fazla ėrenmeyi azaltan bir dizi dzenlemeye sahiptir (14).

Makine öğrenimi yöntemlerinin modelleme sonucunda elde edilen sonuçların daha yorumlanabilir ve açıklanabilir olması amacıyla bazı yöntemlere ihtiyaç duyulmuştur. Bu gereksinimlere dayalı olarak açıklanabilir yapay zeka kavramı ortaya atılmıştır. Açıklanabilir yapay zeka modelleme sonucu elde edilen sonuçların daha anlaşılabilir ve açıklanabilir olması amacıyla her bir gözlem için çıktı ve girdi değişkenleri arasındaki ilişkileri ortaya koyarak modelin daha anlaşılabilir olması için geliştirilen yöntemler bütünüdür (15). Sağlık alanında hastalığı teşhis etmek için sınıflandırma modellerinin kullanımı, büyük ölçüde oluşturulan modellerin araştırıcı tarafından yorumlanabilmesine ve açıklanabilmesine bağlıdır. Sağlık alanında oluşturulan yapay zeka modellerinin açıklanabilirliğini artırmanın birçok farklı yolu vardır ve değişken önemliliği bunlardan biridir. Bu amaçla kullanılan açıklanabilir yapay zeka yöntemleri, belirli bir sınıflandırma için hastaya özel bir açıklama sağlar, böylece herhangi bir karmaşık sınıflandırıcının klinik ortamda daha basit bir şekilde açıklanmasına olanak tanır (16).

Bu çalışma ile MTX ile nefrotoksisite oluşturulmuş sıçanlar ile patoloji oluşturulmamış sıçanlardan alınmış olan biyolojik örneklerin genomik analizleri sonucu elde edilen büyük verilerin biyoinformatik analizlerinin yapılması, geliştirilecek bir yazılım ile sonuçların görselleştirilmesi ve açıklanabilir yapay zekâ ve makine öğrenmesi yaklaşımlarından olan ağaç tabanlı yöntemler ile modellenmesi sonucunda ilaca bağlı nefrotoksisite tanısında olası biyobelirteçlerin belirlenmesi amaçlanmıştır.

2. GENEL BİLGİLER

2.1. Böbrek

Böbrekler, kırmızımsı renkte olup karın bölgesinin arka duvarına yaslanmış olan, göğüs kafesinin hemen altında, geniş, fasulye biçimli bir çift organdır. Böbrekler dolaşım desteğini ana atardamar olan aorttan doğrudan alır ve yine ilgili damarları aracılığıyla alt ana toplardamarına kirli kanı gönderir. Böbrekler idrar yolu ile vücutta yer alan zararlı maddelerin vücuttan uzaklaştırıldığı için boşaltım sisteminin merkezi konumunda yer almaktadır. Bu bağlamda idrar yolunun başlangıcı olarak kabul edilen pelvis adı verilen idrar toplama bölgesinden idrar yolunun kısımları arasında yer alan diğer kısımlarla yani üreter, mesane ve üretra bölgeleri ile bağlantı kurmaktadır. Vücut için gereken bazı hormonların üretildiği bölge olmasının yanında, her iki böbreğin üzerinde yer alan böbrek üstü salgı bezleriyle de komşuluk kurmaktadır. Her iki böbrek de hafif oblik pozisyonudadır. İnsanın fizyolojik yaşantısını normal bir şekilde devam ettirebilmesi için çok önemli fonksiyonlara sahip olup, elektrolit, mineral, sıvı ve asit baz dengesinin sağlanabilmesinde çok önemli rol oynayan organlardır.

Karaciğerin bulunduğu yerden dolayı sol böbrek, sağ böbrekten 1-2 cm kadar daha yukarıdadır. Sol böbreğin boyu sağ böbreğe kıyasla daha uzundur. Beslenme alışkanlıklarına bağlı olarak değişiklik gösterebilmekle beraber böbreklerin uzun eksenlerinin uzunluğu 12- 14 cm, ön-arka genişliği 3 cm, ortalama ağırlıkları ise 140-170 g kadardır (17). Böbreğin görevi arteriyel kan basıncının dengesi, hormonların bazılarının salgılanması, atık ürünlerin vücuttan uzaklaştırılması, idrar oluşumu, vücut sıvılarının dengesinin sağlanması, asit baz dengesi sağlanması ve elektrolit yoğunluğunun düzenlenmesi olaylarıdır (18) .

Vücut dengesinin düzenlenmesinde çok önemli görevleri olan böbrekler, ekspresyon fonksiyonları, yüksek düzeydeki kan perfüzyonu ve metabolik yollar içerisindeki etkinlikleri ile toksinlerin etkilerine, ilaçların yan etkilerine oldukça fazla maruz kalırlar ve toksik maddelerin de hedef organlarıdır (19) .

2.2. Toksisite

Canlı organizmalara zarar verebilecek bitkisel, hayvansal, mineral ya da sentetik maddeler tarafından organizmanın geçici olarak ya da sürekli bir şekilde bozulması olarak tanımlanmaktadır. Zehirlenme veya intoksikasyon olarak da bilinmektedir.

2.2.1. Nefrotoksisite

Böbreklerin yüksek düzeyde perfüzyon hızı olması, tübül lümeninde pek çok maddeyi konsantre edebilmeleri ve metabolizma hızlarının da yüksek olması nedeniyle zararlı maddelerin etkilerine karşı aşırı hassastır. Nefrotoksisite böbreğin yapısının kimyasallar sebebiyle hasar görmesi ve böbrek fonksiyonunda olumsuz bir şekilde etkileyerek azalma meydana gelmesi durumudur (20). Bu durum sonucunda normal şartlarda böbreklerin çıkarmış olduğu hem azotlu hem de azotsuz atıkları olan kreatinin ve üre gibi bileşenlerin vücutta birikmesine neden olur. Bundan dolayı sıvı dengesinin bozulmasına, bazı metabolik sorunlara ve vücutta yer alan diğer organların da fonksiyonlarının bozulmalarına neden olur (21).

Renal perfüzyonunda azalmaya neden olan nefrotoksik metabolitler, hücrelere oksijenin taşınmasını engelleyip, Adenozin trifosfat (ATP) kullanımını artırarak mitokondride meydana gelen enerji üretimini bozarlar (22). Lipid peroksidasyonu, oksidatif stresin ve glutatyon tüketiminden dolayı kimyasallara bağlı olan böbrek toksikasyonuna sebebiyet verdiği düşünülmektedir. Böbreklerden atılmış olan kimyasallar öncelikle toksik hücresel hasar meydana getirir sonrasında böbrek fonksiyonlarını bozup etkilerini gösterirler (23). Nefrotoksisiteden sorumlu olan toksinler; eksojen, ilaçlar, endojen (miyogloblin, hemogloblin) maddeler ve hayvansal yapıdaki toksinlerdir (24). Nefrotoksisite çok uzun bir süredir biliniyor olmasına rağmen, nefrotoksinlere karşı duyulan ilgi son yirmi yıl içerisinde dikkat çekici boyutta artmıştır. Nefrotoksisite için yapılan araştırmaların büyük bir çoğunluğu iyatrojenik zarar ve farmasötik ajanlar üzerine odaklanılmıştır. Çünkü doğal kaynaklı maddelerle nefrotoksisite meydana gelse de, sorunun büyük bölümü adı geçen iki ana başlıktan ortaya çıkmaktadır (25).

2.3. Metotreksat (MTX)

MTX, antineoplastik grubu içerisinde sentetik folik asit antimetaboliti olacak şekilde etki gösteren bir ilaçtır. Daha önceleri ametopterin olarak da bilinmekte olan

MTX, temel hücre yapılarının sentezi sırasında çeşitli basamaklarda orijinal bir şekilde sentezlenen ve hücre proliferasyonunda pirimidin ve pürin sentezi için gereken bir enzim olan dihidrofolat redüktazı (DHFR) inhibe edebilmek için yapısal olacak şekilde tasarlanmış birkaç folik asit 5 antagonistinden biridir. Koenzim ya da substrat üzerinde yer alan özgül bağlanma noktalarına bağlanıp enzimin çalışma sistemini inhibe eden ilaçtır (5, 26). İlk olarak 1940'lı yıllarda MTX, çocukluk çağı lösemisi üzerine yapılan klinik çalışmalar için kullanılmaya başlamıştır. Bu yıllarda, Yella Pragada Subbarow karaciğerden folik asit ekstraksiyonu ile ilgili çalışmalar yapmıştır. Diğer bazı araştırmacılar da aynı zaman içerisinde deney hayvanlarına intravenöz yoluyla folik asit uygulamışlar ve bu uygulama sonucunda kanser hücrelerine ilişkin proliferasyonunun durduğunu bildirmişlerdir (27). Lösemide Subbarow ve arkadaşları'nın 1950'lerde sentezlediği 4-aminopteroylglutamik asit (aminopterin) ve ametopterin moleküllerinin klinik kullanım açısından uygun olduğu onaylanmış ve yaygın bir şekilde kullanılmaya başlamıştır (28). Daha sonra da malign ve malignite haricinde MTX'in birçok hastalık için tedavi amacıyla kullanımı ile Dünya Sağlık Örgütü yüksek başarı derecesine sahip olan temel ilaçların arasına almıştır. Düşük dozları otoimmün hastalıkların tedavisi için etkin terapötik bir ajan olarak kullanılan MTX, yüksek dozları birçok malignite tedavisi için yaygın bir şekilde kullanılmaktadır (2).

2.3.1. Metotreksatın Böbrek Toksisitesi

MTX atılımının %90'ının yapıldığı yerin böbrek olmasından dolayı toksisitesinin fazla olduğu dokuların arasında böbrek de bulunmaktadır (5). Glomerülerden eliminasyonu filtre işlemi gerçekleştikten sonra aktif tübüler sekresyonla birlikte tekrardan emilir (2). Böbrek ilacın farmakokinetiği üzerine çok önemli bir role sahip olduğu için tedavi edilen hastalara uygulanırken doz ayarı yapılmalı ve tolerasyon durumu göz önünde bulundurularak doz azaltılmalı veya tedavinin sonlandırılması gerekebilir. MTX ve metabolitlerinin çoğu atılımı böbrek yoluyla gerçekleşir ve böbrek tübüllerinde birikmesi nedeniyle potansiyel bir toksisitedir. MTX'in böbrek toksisitesinin sistemi; kimyasalın doğrudan toksisitesi, DNA sentezinin bozulmasıyla birlikte getirdiği nötrofil infiltrasyonu, enzim inhibisyonları ve oksidatif stresidir (29).

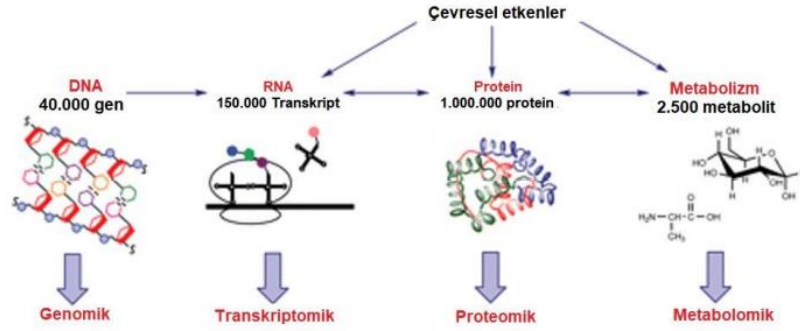
Nötrofillerin kümeleşmesinin ve aynı zamanda aktif hale gelmesi miyeloperoksidaz (MPO) ve serbest radikaller sentezini oluşturmaktadır. Ratlarla yapılmış olan çalışmalarda intraperitoneal MTX uygulamasının sonucunda MPO,

Malondialdehid (MDA) seviyesindeki artış ve GSH düzeyinde azalma ile oksidatif stres açısından anlamlı sonuçlar elde edilmiştir (30).

Akut lösemi ve buna benzer kanser tedavisinde yüksek doz MTX kullanımının ardından serum kreatinin düzeyindeki artış, akut böbrek yetmezliği (ABY) oluşabileceği yapılan çalışmalarda belirtilmiştir (31). Yüksek doz MTX uygulaması sonrasında gelişebilecek ABY için üriner alkalizasyon ve yeterli miktarda idrar çıkışının sağlanması gerekmektedir. Öktem ve ark. çalışmasında MTX uygulayarak renal hasar oluşturulmuş olan lipid peroksidasyonu için süperoksit dismutaz (SOD), katalaz (CAT), glutatyon peroksidaz (GSH-Px) değişkenlerinin düzeylerinde azalma gözlemlenirken, MDA düzeyinde ise artış gözlemlenmiştir. Bir başka çalışma olan Öktem ve ark. ise MTX'in yol açtığı nefrotoksistide dokularda ksantin oksidaz, serum adenozin deaminaz (ADA) ve nitrik oksit düzeylerindeki artış oldukça belirgin bir şekilde gözlenmiştir (32).

2.4. Genomik

2001 yılında insan genom projesinin (Human Genome Project, HGP) tamamlanması ile moleküler biyoloji alanında, izole edilmiş olan biyolojik moleküllerin incelenmesinden daha büyük biyolojik moleküllerin analizinin genişlemesi gerektiği fikri vurgulanmıştır. Omik veriler biyoloji, genetik, ilaçbilim beslenme, toksikoloji gibi pek çok alanda kullanılmakta olan ve başarıyla uygulanan en yeni teknolojiler arasında yer almaktadır. Moleküler bir terime "Omik" son ekinin eklenmesi, incelenmekte olan biyolojik sisteme ilişkin bir bölümünü kapsayan bir dizi molekülün (proteinler veya genler gibi) daha kapsamlı olacak şekilde değerlendirilmesini ifade eder (33). Omik teknolojileri, büyük biyolojik moleküllerini incelemek amacıyla uygulanan biyobelirteçleri keşfetmeye yarayan araçlardır. Omik tekniklerinin çeşidi sürekli olarak artış göstermiş olsa da, en gelişmiş olan beş omik teknolojisi transkriptomik, epigenomik, genomik, metabolomik ve proteomik olduğu bilinmektedir (34). Omik teknolojileri ve aralarındaki ilişki Şekil 2.1 ile verilmiştir.



Şekil 2.1. Omik teknolojileri ve aralarındaki ilişki (35)

Genomik, proteomik tanımlarında geçmekte olan ~omik eki, gen ve protein sözcüklerinin sonuna eklenip bütünlük, tümlük anlamı kazandırmaya yaramaktadır. Dolayısıyla “genom” sözcüğü ile bütün genleri, “proteom” sözcüğü ile de belirli bir yer ve zaman içerisindeki bütün proteinlerin ifade edildiği söylenebilir (36).

Genom ve genlerle ilişkili fonksiyonel-organizasyonel bilgiyi açıklamaya yönelik olarak yapılan çalışmaların bütünü olacak şekilde özetlenebilecek olan genomik;

- Bir canlıda bulunan tüm genlerin tek tek tanımlanmasına,
- Genlerin hem birbiriyle hem de çevrenin genlerle olan etkileşimlerinin araştırılmasına,
- Genlerin yer-zaman-miktar olarak üretiminin ve aktivasyonlarının incelenmesine aracılık etmektedir (37).

Genomik, bazı virüslerin ve canlıların biyolojik gelişimlerinde gerekten olan genetik bilginin aktarılmasını sağlayan DNA'nın yapısını ve işlevlerini detaylı bir şekilde ele alan bilim dalı olarak da tanımlanabilmektedir. Bir organizmaya ait genomdaki bütün genetik yapıların belirlenebilmesi, dizi analizinin yapılp haritasının çıkarılabilmesi gibi DNA yapısının ve işlevinin kapsamlı olarak araştırılmasına genomik teknolojisi denilmektedir (38). Bir genom, bir organizmaya ait olan bütün genler de dâhil olacak şekilde eksiksiz bir DNA setidir. Ayrıca, genomik tüm genomların işlevlerinin ve yapısının bir araya getirilebilmesi ve analiz edilebilmesi için yüksek verimli DNA dizilimini ve biyoinformatik yöntemlerin kullanarak genomların dizilenmesini ve analiz edilmesini içermektedir. Genomikteki ilerlemeler sayesinde araştırma ve sistem biyolojisinde biyolojik sistemlerin anlaşılabilirliğini kolaylaştırabilecek keşif yapamaya dayanan bir devrim tetiklenmiştir. Genomik teknolojinin tıp alanındaki ana hedefi

hastalıkları önleyebilmek, bireye özgü olacak şekilde hastalıkların tanı ve tedavisini düzenleyebilmektir (39).

Genomik çalışmalar hem hastalıkla ilgili genlerin hem de hastalığın fizyolojik süreci ile ilgili genlerin tanımlanmasına, genler arasında yer alan yapısal ve fonksiyonel etkileşimlerin ortaya konulmasına, genlere ait ekspresyon profillerinin ortaya çıkartılmasına ve genetik zeminde farklı organizmaların da karşılaştırılmasına olanak tanır (40).

Genomik çıktıların bilişim teknolojileri yardımıyla işleyerek anlamlandırıp saklayan, biyoinformatik ve otomasyon alanlarındaki ilerlemeler ile gelişen bir bilim dalıdır (41). Ayrıca genomik hemen hemen her konuda araştırma yaparak durum değerlendirmesi yapabilen uygun disiplindir. Onkoloji, İmmünoloji, Farmakoloji, Mikrobiyoloji, Biyokimya, vb. tıbbın hemen her alanında araştırmalar doğru seçilmiş bir karşılaştırma tekniği ile yapılabilmektedir. Karşılaştırmalı çalışmalar sayesinde, ilaca verilen yanıtın tahmini ve böylelikle kişiselleştirilmiş ilaç gelişimi, kanserleşme ve prognoz tahmini, immün yanıtın yapısı ile transplantasyon sonucunu erken tahmin etme, mikroorganizmalar ve alt gruplarının genomu gibi araştırmalar yapılabilir.

Genomik; yapısal genomik ve fonksiyonel genomik olmak üzere iki alt gruba ayrılmaktadır.

2.4.1. Yapısal genomik

Yapısal genomik; genlerin kromozomların üzerinde bulundukları yerlerin gösterilmesine imkan tanıyan genom haritalaması ve DNA nükleotid dizisini saptanmaya yarayan DNA dizi analizi yöntemleri yardımıyla organizmaya ilişkin genetik bilgilerin açığa çıkmasına yardımcı olur. Yapısal genomik ile organizmaya ait genetik bilgilerin ortaya çıkartılmasına imkan tanısa da, ortaya çıkan bu bilgi yalnızca fotoğrafik olup genlerin fonksiyonlarına ilişkin herhangi bir bilgi vermez. Bu dizilerle ilgili fonksiyonu bilmek ise en önemli olan noktadır. Bu noktada karşılaşılan eksikliği, geçmişte genlerden fonksiyonu bilinmeyen genlerle farklı olup ama akraba bir türe ait homolog ve fonksiyonunu bilinen genlerin karşılaştırılması ile kapatmaya çalışmışlardır. Bugün yöntem sınırlılıkları sebebiyle kullanılmamaktadır ve bu yüzden yapısal genomik yerini fonksiyonel genomik üzerine çalışmalara bırakmıştır (37).

2.4.2. Fonksiyonel genomik

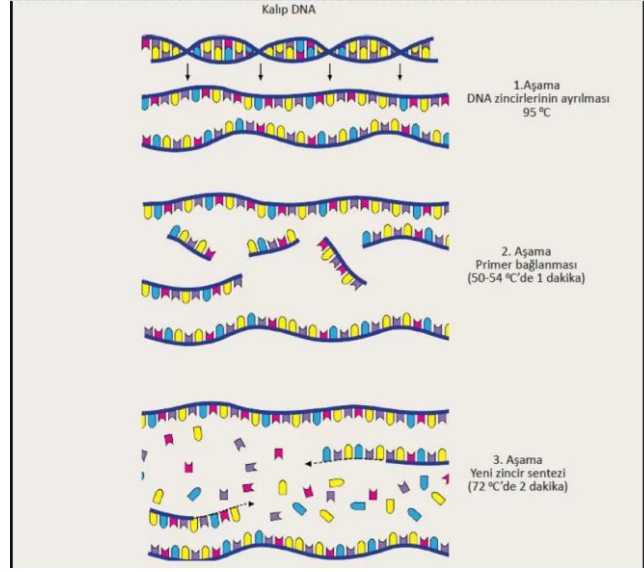
Fonksiyonel genomik; biyolojik açıdan genomik bilgiye anlam kazandırabilmek amacıyla, genlerin ekspresyon seviyelerine, genlerin fonksiyonlarına ve genlerin birbirleriyle olan ilişkilerine açıklık getirilmesine olanak tanıyan genom ölçekli sistematik yaklaşımları içine alan bir disiplindir. Genlere ait fonksiyonların öğrenilmesine ilaveten organizma açısından önemlerinin anlaşılması konusunda da yardımcı olur. Fonksiyonel genomik çalışmaları mRNA'lar üzerinden yürütülür. Bunun sebebi, DNA'ya ilişkin fonksiyonun mRNA'lar yardımıyla gösterilmesidir (40).

Genomik çalışmalar, geleneksel yöntemlere ek olarak yüksek verimli yöntemlerin yardımıyla yapılmaktadır.

Geleneksel yöntemler için polimeraz zincir reaksiyonu temelli, elektroforez temelli, blotlama temelli ve DNA dizi analizleri yöntemleri ele alınırken yüksek verimli yöntemler de mikroyarray yöntemi ele alınmaktadır.

2.4.3. Polimeraz zincir reaksiyonu (PCR)

Bir Polimeraz zincir reaksiyonunun (Polymerase Chain Reaction, PCR) keşfedilmesi biyoloji bilimi için önemli bir buluştur. Bu yöntem adını, DNA'nın bir parçasını DNA Polimeraz enzimini kullanarak in vitro yöntem yardımıyla çoğaltılmasından almaktadır. PZR, DNA havuzundan bir parça DNA'nın çoğaltılmasına olanak tanıyan basit ve enzimatik bir işlemdir (42). PCR yönteminde saç, deri, periferik kan, ve tükürük ve mikroorganizma gibi farklı türde bitkisel ve hayvansal kökenli materyal DNA kaynağıdır. PCR için gereken DNA miktarı laboratuvar yöntemleri ile elde edilip bu yöntemlerle yeterli miktarda DNA kopyası analiz edilir. PCR yönteminde termal döngü cihazı kullanılır. Termal döngü ise belirli bir dizisi olan PCR örneği ısıtılıp daha sonra soğutulmasıdır. PCR termal döngüsü üç basamaktan oluşmaktadır. DNA'nın 92-95 °C de birkaç dakika tutulması ile DNA zincirinin ayrılması (denatürasyon), reaksiyon sıcaklığı 37-65 °C'ye düşürülüp açılan primerlerin (oligonükleotid primerlerin) açılan hedef DNA zincirlerine bağlanması (yapışma) ve DNA zincirlerinin üzerine yapışmış olan primerlerin DNA polimeraz enzimi yardımıyla uzatılmasıdır (43). Tekrarlanan termal döngüler sonucunda DNA'nın milyarlarca kopyası elde edilebilir. Şekil 2.2'de PCR basamakları gösterilmektedir.



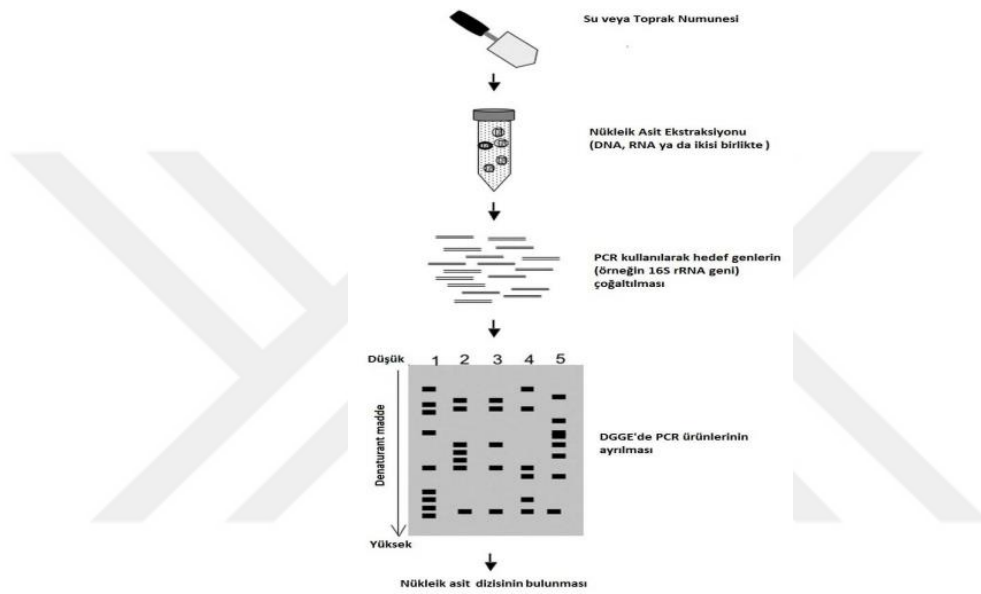
Şekil 2.2. Polimeraz zincir reaksiyonun basamakları (43)

2.4.4. Denatürleyici Kademeli Jel Elektroforezi (DGGE)

DGGE yönteminde, jel üzerinde farklı nükleik asit dizilerinden aynı boyutta olanların hareketine bağlı olarak gerçekleşmektedir. Yani, elektroferez üzerinde farklı rRNA dizilerinin her biri kendilerine özgü olan farklı denatüre bölgeleri içerisinde erimeye başlar. Böylelikle numunedeki genetik biyoçeşitliliği yansıtacak olan band desenleri üretilmektedir. Numunedeki baskın olan türlerin sayısı bantların sayısı ile ifade edilmektedir (44).

DNA parçalarından Bbz dizilimi farklı olanların kimyasal denatüran gradyanı artan jel içinde ilerleyişi esnasında göstermiş oldukları erime (DNA'nın çift sarmalının birbirinden ayrılma şekli) özellikleri de birbirinden farklılık gösterir. Çift sarmallı olan DNA parçasına ilişkin erime sıcaklığı, eş bazlar arasında yer alan hidrojen bağlarının komşu bağlar ile arasındaki etkileşim aracılığıyla belirlenir. DGGE yönteminde denatüre gradyanı artmış olan jel üzerindeki DNA parçasına ait çift sarmal birbirinden yavaşça ayrılır. Sarmal açılınca jel üzerindeki DNA parçasının hareketliliği azalır ve DNA hareketsiz kalıncaya kadar bu durum devam eder. Genellikle formamit ya da üre kimyasal denatüre gradyan olarak kullanılır. Jelin uygun bir yöntem ile boyanması ile ayrılmış olan DNA parçaları görüntülenir (45). Birden çok numunenin aynı jel üzerinde yürütülebilir olması, DGGE'nin en önemli avantajıdır. Dolayısıyla değişen çevresel koşullarının ve

zamanın etkisiyle meydana gelen popülasyon değişimlerini izlemeye imkan tanır. Dahası, klonlama ve dizi analizi yöntemleri gibi pahalı, zahmetli ve numune sayısı arttıkça vakit alıcı bir teknik değildir. Genomikte DGGE, klonlanmış olan ve PCR ile çoğaltılan DNA’da oluşan tek baz değişimlerinin ve polimorfizmin belirlenmesi için etkili bir analiz yöntemidir. Şekil 2.3’de Denatürleyici kademeli jel elektroforezinin aşamaları gösterilmiştir.



Şekil 2.3. Denatürleyici kademeli jel elektroforezinin aşamaları (46)

2.4.5. Southern Blotlama

1973'te Edwin Southern tarafından geliştirilmiş olan ve hala kullanılmakta olan bu biyoloji tekniği, DNA'ya ilişkin yapının incelenmesi için kullanılmaktadır. Birçok diğer DNA moleküllerinin arasından spesifik olan DNA molekülünü tespit ederek ve bu spesifik sekansı tanımlayabilmek amacıyla bir organizmaya ait toplam DNA'sını (genomunu) analiz etmektedir.

Southern Blotlamanın aşamaları şu şekildedir.

1. İlk aşamada bir restriksiyon endonükleazıyla genomik DNA küçük parçalara ayrılarak hazırlanır.
2. Jel elektroforeziyle DNA fragmentleri boyutları dikkate alınarak ayrılır.
3. Jel içinde çift sarmallı DNA parçaları tek sarmallar şeklinde denatüre edilir.

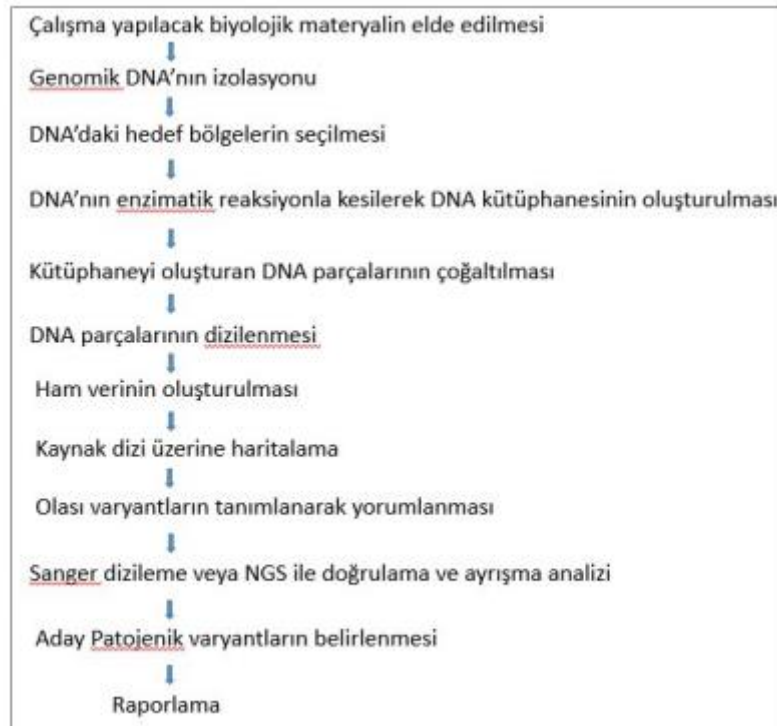
4. DNA jel üzerinden bir blotlama membranına (naylon filtre ya da nitroselüloz) aktarılır.
5. Aktarım bittikten sonra, membran tüm bantları orijinal olacak şekilde jel üzerinde taşır.
6. Daha sonra membran numunede belli bir DNA sekansının tamamlayıcı bir sekansı olacak şekilde tasarlanmış bir prob olarak isimlendirilen küçük bir RNA veya DNA parçasıyla muamele edilir. Bu, probun membran üzerinde yer alan spesifik bir DNA fragmanı ile hibridizasyonu sağlanmaktadır.
7. Son aşamada, hibridizasyondan sonra bir floresan veya bir radyoaktif atom boya aracılığıyla işaretlenir. Böylece zar üzerindeki ilgili DNA fragmanının diğer birçok farklı DNA fragmanının içinden tespit edilmesine imkan tanır (47).

2.5. DNA Dizileme Analizi

Son Dizileme, genom çalışmaları için önemli bir yere sahip olup canlılara ait genetik bilgilerinin tutulduğu DNA üzerindeki nükleotitlerin sıralamasının bulunması sürecidir. Pek çok farklı hastalık ve genom üzerinde yapılmış olan araştırmalar ve araştırmaların sonucunda ortaya çıkan keşifler dizilemenin ve dizileme teknolojisinin gelişmeleri sonucudur. 1975 yılında Frederick Sanger tarafından DNA dizileme için, zincir sonlandırma yönteminden yola çıkarak geliştirilen yöntem, 1977’te Sanger ve arkadaşları tarafından yayımlanıp literatüre sunulmuştur (48). Aynı yıl içerisinde, kimyasal parçalama yöntemine dayanan DNA dizileme yöntemi başka bir çalışma ile Maxam ve Gilbert tarafından yayımlanmıştır. Sanger’in geliştirildiği yöntem, yüksek doğruluğa sahip olduğu için DNA dizileme de altın standart olarak bilinmektedir. Sanger dizileme yaklaşımı, DNA dizileme alanında “Birinci Nesil Dizileme” yöntemi olarak kabul görüp, İnsan Genom Projesi de Sanger dizileme ile yapılmıştır. İnsan Genom Projesi, tarihin on üç yıllık tamamlanma süresiyle maliyetli projeler arasındadır. Bu sürecin başarılı bir şekilde tamamlanması Sanger dizileme ile olsa da, maliyet olarak daha az ve Sanger dizileme kadar güvenilir olan yeni bir teknolojinin varlığına ihtiyaç duyulmuştur (49). 2004 yılında bu ihtiyaç doğrultusunda, maliyet ve zaman problemini ortadan kaldıracak yeni bir proje Ulusal İnsan Genomu Araştırma Enstitüsü ile başlatılmıştır. Bu proje, DNA’nın paralel olacak şekilde dizilenmesi temeline dayalı olan yaklaşımların üzerine kurulan yeni bir teknolojinin ortaya çıkmasına sebep olmuştur. Bu yönde geliştirilmiş olan yaklaşımların tümü “Yeni Nesil Dizileme (YND)” olarak isimlendirilmiştir.

2.5.1. Yeni Nesil Dizileme Teknolojisi (YND)

YND yöntemi DNA'nın enzimatik reaksiyonlar yardımıyla kesilerek oldukça fazla sayıda DNA parçası ile bir kütüphanenin meydana getirilmesi ve kütüphaneyi meydana getiren DNA parçalarının çoğaltılması temeline dayanmaktadır. Paralel sekanslama ile milyonlarca sayıda küçük DNA parçalarının eş zamanlı olacak şekilde dizilenmesi gerçekleştirilir; böylece genomda yer alan bazların her birinin birden fazla okunabilir ve varyasyonların tespiti daha doğru bir şekilde elde edilir (50). Sistemin çalışma prensibi aşağıdaki basamaklarından oluşmaktadır; çalışma yapılacak olan biyolojik materyal elde edilir, elde edilmiş olan biyolojik materyallerden genomik DNA'nın izolasyonu yapılır, İzole edilmiş olan DNA'daki hedef bölgeler seçilir, enzimatik reaksiyon yardımıyla DNA kesilerek DNA kütüphanesi oluşturulur, Kütüphaneyi oluşturmuş olan DNA parçaları çoğaltılır, DNA parçalarının dizilenmesi gerçekleştirilir, Dizileme sonrasında ham veri oluşturulur, Kaynak dizinin üzerinde haritalama gerçekleştirilir. Olası değişimler tanımlanarak yorumlanır, Sanger dizileme yöntemi veya YND yöntemi ile doğrulama ve ayrışma analizi yapılır, aday patojenik değişimler belirlenir ve son olarak da elde edilmiş olan bu veriler raporlanır (Şekil 2.4) (51).



Şekil 2.4. YND'nin çalışma prensibi (51)

Günümüzde dizileme yönteminin kullanımı açısından farklılık gösteren YND platformları bulunmaktadır. Tüm dünyada bugün YND sistemlerinden SOLID, Roche 454, Illumina Miseq, Ion Torrent, Pacific Biosciences yaygın bir şekilde kullanılmaktadır. Bu platformların hepsinde ortak bulunan çalışma aşamaları DNA kütüphanesinin oluşturulması, dizileme, görüntüleme ve elde edilmiş olan verilerin analizinin yapılması basamaklarını içermektedir (50).

2.5.2. Mikroçip (Microarray)

Mikroçip yöntemi 90'lı yılların ortasına doğru gelişmeye başlayan, günümüzde ise hücrelerdeki ve dokulardaki çok fazla sayıda gen ekspresyon düzeyini toplu bir şekilde incelemeye imkan veren güçlü ve yeni bir teknolojidir ve aynı anda binlerce DNA'nın analiz edilmesine olanak sağlamaktadır (52).

Bir mikroorganizmaya ait olan tüm genler mikroskop lamı ölçeğinde bir yüzeye spotlanarak ve genin ekspresyon seviyelerinin binlercesi sadece tek bir deney ile eş zamanlı bir şekilde çalışılabilir (53). Bu kadar fazla sayıdaki genin eş zamanlı olarak çalışılmasından dolayı ve örnek hacminin çok küçük bir kısmının yetecek olabilmesi sebebiyle yüksek verimli bir yöntem olarak kabul edilmektedir. DNA Mikroçip teknolojisi, ekspresyon analizleri, mutasyon analizleri, genler ve genlere ait polimorfizmlerinin araştırılması, evrimsel çalışmalar, potansiyel olan terapötik ajanların geliştirilmesi, sekans analizleri, genlerin sınıflandırılmasını kapsayan birçok alanda kullanılabilmektedir (54). Mikroçip teknolojisi beş temel basamaktan oluşmaktadır (37).

1. Hedef DNA, complementary DNA (cDNA) dizilerine ait tamamlayıcı dizileri kapsayan problemlerin immobilize edildiği mikroarray platformunun hazırlanması veya temin edilerek alınması.

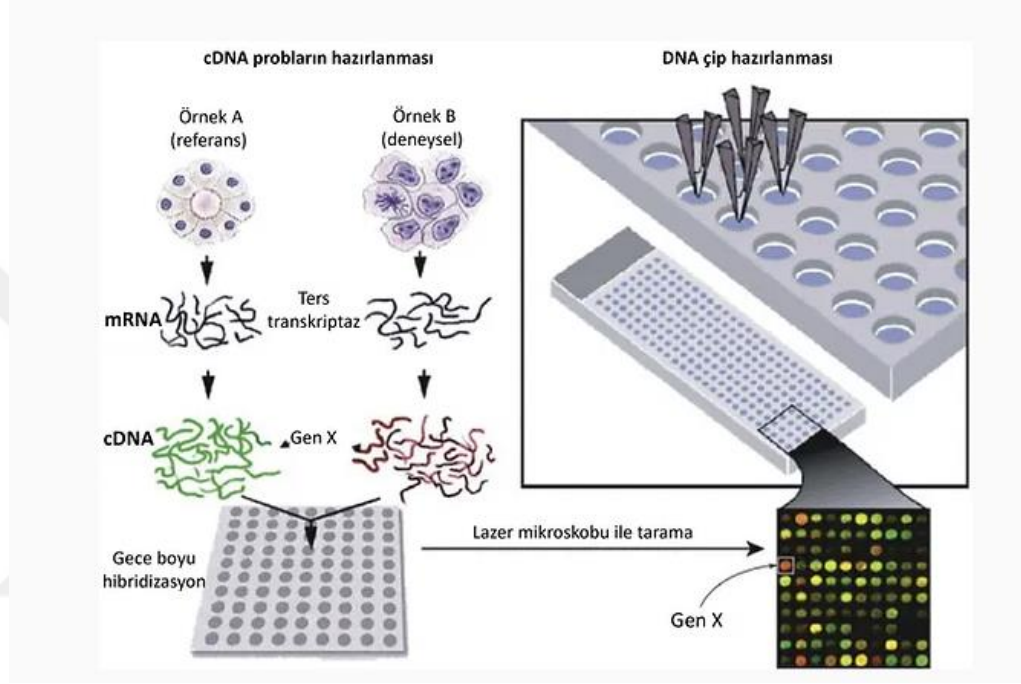
2. İncelenecek olan örnekten floresan kullanarak işaretlenmiş DNA, cDNA, complementary RNA (cRNA)'nın hazır hale getirilmesi.

3. İşaretlenen DNA, cDNA, cRNA ile mikroçip platformunda yer alan hibridizasyon solüsyonunun içerisinde karşılaştırılması.

4. Yıkama işleminden sonra platformun yüzeyindeki hibridizasyonun olup olmadığı okuyucu veya tarayıcı yardımıyla analiz edilerek görüntünün bilgisayarda depo edilmesi.

5. Depolanmış olan görüntü bir yazılım ile değerlendirilerek mikroçip platformunun üzerindeki noktalardan hangilerinde hibridizasyon olup olmadığı ve niceliksel açıdan seviyesinin (koyu mavi-yeşil-sarı kırmızı renk yelpazesi ile) değerlendirilerek yorumlanması.

Mikroçip yöntemine ait gösterim Şekil 2.5’de verilmiştir.



Şekil 2.5. cDNA için Mikroçip yönteminin çalışma şekli (37)

2.5.3. RNA Dizileme Analizi

RNA dizileme (RNA-Seq), bir hücreye ait transkriptoma ilişkin görüş sağlayabilmek amacıyla yüksek verimli sekanslama yöntemleri kullanılmaktadır. Sanger teknolojisi ve mikrodizi tabanlı olan yöntemler ile karşılaştırıldığı zaman RNA-Seq, transkriptoma ilişkin dinamik doğanın çok yüksek bir kapsamını ve daha fazla yüksek çözünürlüğünü sağlamaktadır. YND yöntemlerinde ortak özellik, kılcal elektroforez tabanlı Sanger dizileme cihazlarındaki kapasiteyi aşacak dizileme reaksiyonları bir tek deneyde gerçekleştirebilmesidir. Gen ekspresyon düzeyini ölçebilmenin ötesinde RNA-Seq’in ürettiği veriler yeni transkriptlerin keşfedilmesini, alele özgü ekspresyon seviyesinin belirlenmesini, alternatif olarak eklenmiş olan genlerin tanımlanmasını kolaylaştırır (55).

Son yıllarda RNA-seq yöntemi hızla gelişmekte olan YND teknolojilerinden biridir. RNA-seq ile beraber yaygın bir şekilde kullanılan YND yöntemleri içerisinde genlerin kodlayan sekansları olarak bilinen ekzonların tümünün sekanslanması ile uygulanmakta olan Tüm Ekzom Dizileme (Whole Exome Sequencing, WES) yöntemidir. Yaygın olarak kullanılan bir başka YND yöntemi ise Tüm Genom Dizileme (Whole Genome Sequencing, WGS) olarak bilinen bütün genomik DNA'nın dizilenmesi temeline dayanan yöntemdir. Tüm YND yöntemlerinin ortak bir çalışma prensibi vardır. Bu çalışma prensibi ise çok fazla miktarda transkriptomik (RNA) veya genetik (DNA) sekans bilgisinin üretilmesidir. RNA-seq metodunda biyolojik materyalin ifade ettiği gen bölgelerinin ifadelerinin dizileri elde edilir.

RNA-seq, nadir rastlanan hastalıkların, kanser ve kansere benzeyen karmaşık hastalıkların mekanizmalarında sorumlu olan genleri düşük maliyetler ile hızlı bir şekilde tespit edebilmektedir. YND analizi, daha önceden farklı bir fenotipe bağlı olduğu bilinen bir gende bulunan patojenik bir mutasyonu tanımlayabilmek amacıyla ayrışma analizi gibi bir yöntem ile beraber kullanılabilir (56).

2.5.4. Transkriptomik

Transkriptom, bir hastalığın hücresel gelişimini ve hastalığa ilişkin mekanizmasının anlaşılması için hayati derecede önemli olan, anlık bir zaman diliminde, moleküler süreç veya gelişim için oluşan bütün transkriptlerin toplamı olarak ifade edilmektedir. Transkriptomun karmaşık verileri, YND ile RNA dizileme, mikrodizileme gibi yöntemleri kullanır; fakat mikrodizileme yönteminde, RNA moleküllerini ölçebilmek ve sınıflandırabilmek için kapasitesi sınırlıdır. Yüksek çözünürlüğe sahip olan RNA-seq yöntemleri hızlı, tekrarlanabilir, pratik ve geniş çaplı araştırılmış olan transkriptomu kullanabilmektedir (57).

Transkriptom, yüksek derecede karmaşıktır ve pek çok sayıda kodlayıcı ve kodlanamayan RNA türünü kapsamaktadır. Tarihsel olarak, RNA molekülleri, moleküler biyolojinin merkezinde yer aldığı gibi, genlerin ve proteinlerin arasında bir ara ürün olarak yer almaktadır. Bu sebeple, haberci RNA (mRNA), proteinleri genetik kod yardımıyla kodladıklarından dolayı en fazla incelenen RNA türleri olmuşlardır. Protein kodlamakta olan mRNA'ya ek olarak, fonksiyonel açıdan çeşitli kodlanamayan RNA (ncRNA) molekülleri grubu da bulunmaktadır. Bilinen ncRNA'ların çoğu, mRNA translasyonunda yer almakta olan ribozomal RNA'lar (rRNA) ve transfer RNA'lar

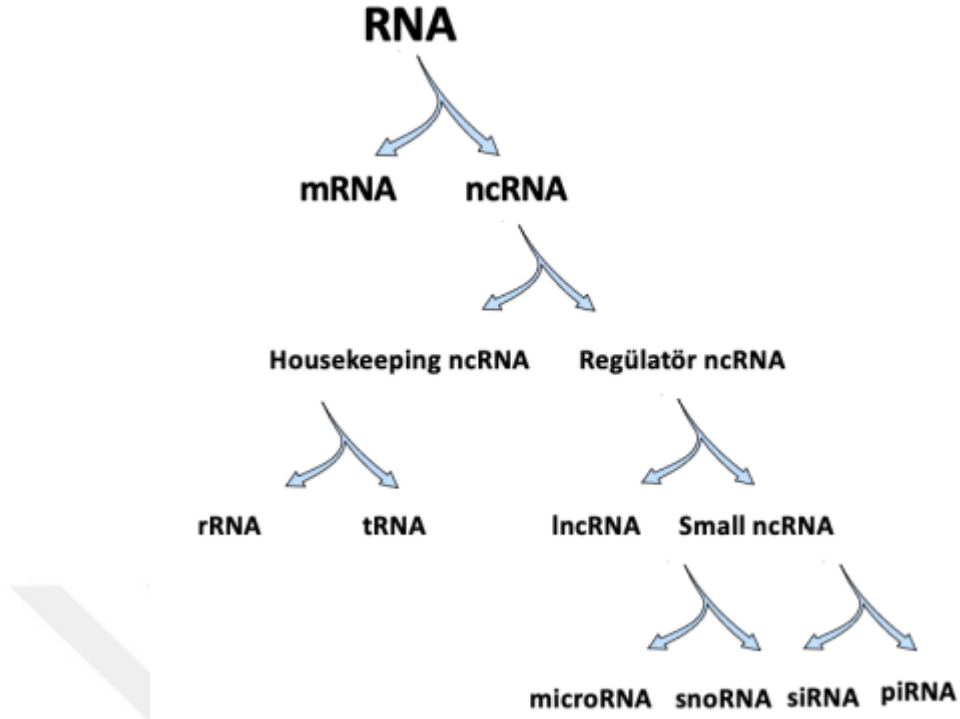
(tRNA), eklemede bulunan küçük nükleer RNA (snRNA'lar) ve rRNA'ların modifikasyonunda bulunan küçük nükleolar RNA'lar (snoRNA'lar) gibi temel olan hücresel görevleri yerine getirmektedir. Yakın zamanlarda, yeni ncRNA sınıfları keşfedilmiştir. İkisi de transkripsiyondan sonraki seviyede gen ekspresyonunu düzenlemekte olan mikroRNA (miRNA) ve piwi-etkileşimli RNA'yı (piRNA) kapsayan küçük kodlanamayan RNA'lar bu tür yeni bir ncRNA sınıfına örnek verilebilir. Uzun kodlamayan RNA'lar (lncRNA'lar) bir diğer kayda değer ncRNA sınıfıdır. Fonksiyonel bir sınıf olan, lncRNA'lar ilk defa farelerde cDNA kitaplıklarında büyük ölçekli dizilemesinde tanımlanmıştır. LncRNA'lar için, transkripsiyonel kontrol, kromatin yeniden modelleme ve transkripsiyon sonrasında işleme dahil edilen sayısız moleküler fonksiyon keşfedilmiştir, fakat LncRNA'ların büyük çoğunluğu daha tam olarak karakterize edilememiştir.

RNA-seq, RNA ve transkriptom miktarı hakkında YND yöntemini kullanıp çok yönlü bir araştırma sağlamaktadır. Dizileme yöntemleri yardımıyla, miRNA, lncRNA, rRNA, dairesel RNA'lar (circRNA), tRNA, küçük inhibitör RNA'lar (siRNA) ve diğer RNA sınıfları (örn., piRNA ve snRNA) ile ilişkili olan pek çok değişiklik gen seviyesinde incelenebilmektedir.

Karmaşık transkriptom verileri, fizyolojik ve gelişimsel süreç sırasında sürekli bir şekilde değişiklik göstermektedir. Genlerin moleküler aktivitesi sırasında mutasyonlar, polimorfizmler nedeniyle ortaya çıkan farklılıklarla beraber kararsız olan bir genetik mekanizma meydana gelir (58). Yazılım teknolojilerinin ve hesaplamalı bilimlerin gelişmesiyle beraber epigenom, genom ve transkriptom gibi multidisipliner alanlardaki çalışmalara odaklanmış, böylece karmaşık olan hastalıkları karakterize etmek, kararsız genetik mekanizmasını anlayabilmek için yeni nesil yöntemlerin bu konularda uygulanması daha da kritik bir hal almıştır. Bu sırada, kodlayan RNA'lar ve kodlanamayan RNA'lar, bir hücre veya hücre popülasyonunda yer alan bütün RNA transkriptlerine atıfta bulunup, transkriptom kavramına dahil edilmiştir. İlgili olan dokulardan ve hücre gruplarından elde edilmiş olan tam transkriptom analizi, mutasyonları tespit edebilmeye, genlerin görevlerini anlamaya, konakçı-patojen ilişkisinde yer alan hedefleri belirlemeye, hastalıklı dokulara ait farklı olan gen ekspresyon seviyelerini ortaya koymaya ve biyobelirteçleri tanımlamayabilmeye konularında yardımcı olabilmektedir (59).

2.5.5. Kodlamayan RNA'lar

İnsan genomunda dizilenmekte olan DNA'daki bilginin %90'ından fazlasının RNA'ya transkribe edilir daha sonra RNA'nın da görevi bu şifreyi proteine dönüştürmektir. Protein ürünlerine dönüşebilen yalnızca (%2)'lik az bir kısımdır. DNA'dan transkribe edilen fakat protein olarak ürüne çevrilmeyen transkriptlere kodlanamayan RNA'lar (ncRNA'lar) denilmektedir (60). Literatüre bakıldığında, biyolojik sistemlerdeki bazı fonksiyonlarda kodlanamayan RNA oranının, kodlayan RNA oranından daha yüksek olduğu öne sürülmüştür. Protein kodlamakta olan genlerin yaklaşık olarak %60'ı ncRNA'lar tarafından hedeflenir. Bu nedenle, RNA transkriptlerinde çoğu ncRNA'lar primer rol oynar. ncRNA'lar uzunluklarına göre kısa kodlanamayan RNA'lar ve uzun kodlanamayan RNA'lar olmak üzere iki ana kategoriye ayrılmaktadır. Kısa kodlanamayan RNA'lar 200 nükleotidden daha kısa RNA'lardır. Büyük oranda miRNA, küçük kodlanamayan RNA (sncRNA) ve küçük snRNA kısa kodlanamayan RNA'lardır. 200 nükleotidden daha fazla nükleotit içeren RNA'lara ise uzun kodlanamayan RNA'lar (long non-coding RNA, lncRNA) dır (61). Kodlanamayan RNA'lar işlevleri bakımından housekeeping ve regülatör (düzenleyici) RNA'lar olmak üzere sınıflandırılabilirler. Housekeeping ncRNA'lar, tRNA ve rRNA'dan oluşur. Diğer taraftan, düzenleyici RNA'lar sncRNA'lar ve lncRNA'lardan oluşur. Yaklaşık olarak 15-40 baz aralığında değişmekte olan küçük ncRNA'lar; miRNA, snoRNA'lar, siRNA'lar ve piRNA'lar ile etkileşimde olan RNA'lardır (Şekil 2.6) (61).



Şekil 2.6. Kodlanamayan RNA'ların sınıflandırılması (61)

ncRNA'lar; gelişimsel süreçlerde, hücrel savunmada, hücre farklılaşmasında, DNA replikasyonunda, transkripsiyon sürecinde ve post-transkripsiyonel sürecinde görev almaktadırlar. ncRNA'larda ortaya çıkan bozukluklar birçok hastalığa neden olmaktadır. ncRNA'lar immün yetmezlik hastalıklarda, kanser de dahil olmak üzere nörodejeneratif hastalıklarda ve kardiyovasküler hastalıklar gibi pek çok hastalığa ilişkin fizyolojik ve patolojik süreçte görev almaktadırlar. Hastalıkların gelişimine sebep olduğu düşünülen ncRNA'lar, pek çok yeni tedavi yaklaşımlarının biyobelirteç arayışlarında sadece hedef olarak değil aynı zamanda tedavi aracı olarak da görülmektedirler (62).

2.5.6. Uzun Kodlanamayan RNA'lar (LncRNA)

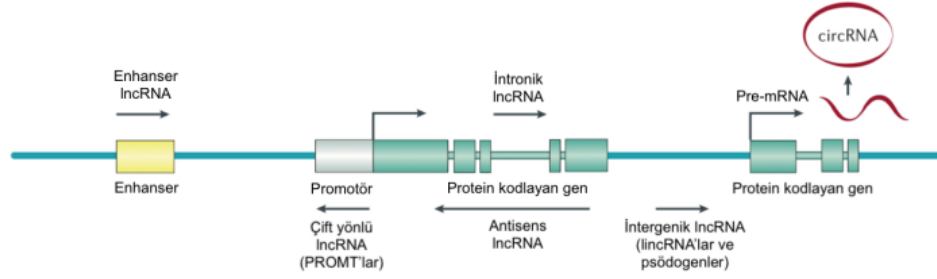
Uzun kodlanamayan RNA'lar (lncRNA'lar), nükleotid uzunluğu 200'den fazla olan ve genellikle protein kodlayamayan bir RNA sınıfıdır. 1989'da ilk memeli lncRNA'sı keşfedilmiş olup H19, daha sonra memelilerde X kromozom inaktivasyonu için önemli rol oynayan lncRNA Xist tespit edilmiştir. LncRNA'ların büyük çoğunluğu protein kodlamayan transkriptler sınıfında yer alır. Bugüne kadar tanımlanmış lncRNA'ların kodlandığı gen sayısının yaklaşık olarak 55 000 olduğu bilinmektedir. İnsan genomunun büyük çoğunluğunu bu rakam temsil etmektedir (63). Bu genlerden hücrel

süreçleri düzenlemedeki işlevi ve etki mekanizması bilinenlerin sayısı hala çok azdır. Kısa kodlanamayan RNA'lara göre lncRNA'ların yapısı ve işlevi oldukça çeşitlidir. lncRNA'lara ilişkin bazı özellikler mRNA'ya benzeyen bir yapıdadır. Örnek verilmek istenirse, lncRNA'lar, 5' başlık, poly-A kuyruğa sahip olup kırılma mekanizması aracılığıyla işlevsel özellik açısından olgun transkript haline gelmektedir. Dahası, mRNA'ların polimeraz II tarafından sentezlendiği gibi lncRNA'lar da polimeraz II tarafından sentezlenir (64). Genellikle protein kodlayan genlerle karşılaştırıldığı zaman lncRNA'lar, düşük ifade düzeyine sahiptir. Buna karşın, lncRNA'ların birçoğu dokuda ve gelişimsel aşamalarda düzenleyici rollere sahiptir. Daha önceleri lncRNA'ların, transkripsiyon işleminin yan ürünleri olduğu düşünülmekteydi. Ancak, lncRNA'ların, transkripsiyonel, translasyonel ve translasyon sonrasındaki süreçlerde hücrelerdeki gen ifadesini ve işlevini düzenleyip çeşitli biyolojik süreçlerde yer aldığı bilinmektedir. lncRNA'lar yaygın olarak farklı kanser çeşitlerinde anormal ifade düzeyine sahiptir. Bu sebeple lncRNA'ların, insan karsinogenezinde çok önemli bir rolü olduğu düşünülmektedir. Kanser başlangıcında ve kanserin ilerleme aşamalarında lncRNA'lar rol oynamaktadır (65).

2.5.7. lncRNA'ların Sınıflandırılması

lncRNA'ların uzunluğu, hücre içi konumu ve moleküler işlevleri açısından kısa ncRNA'lardan farklı olduğu için lncRNA'ları sınıflandırmak daha güçtür (66) .

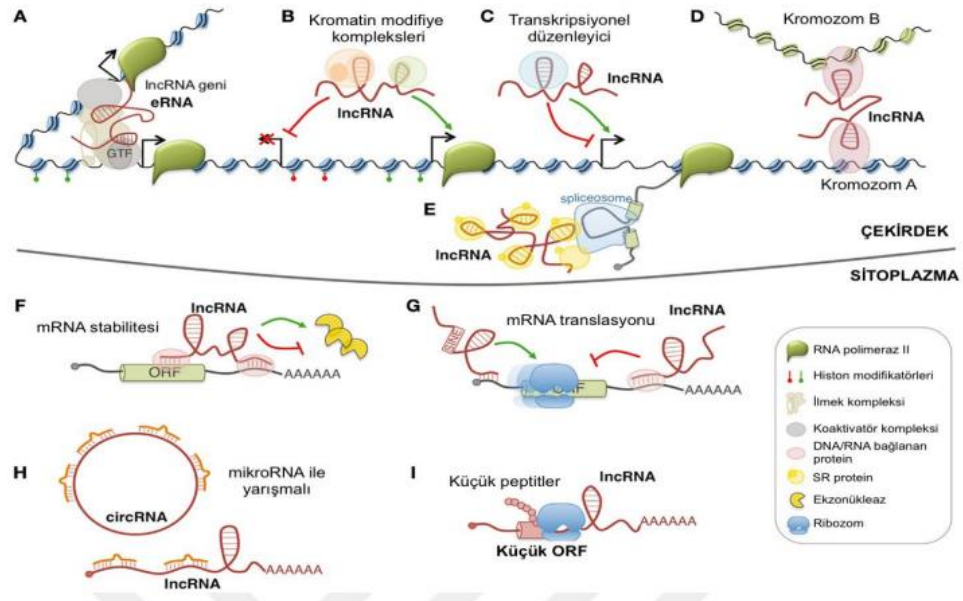
Genomik organizasyonlarına göre, lncRNA'lar beş ana kategoride sınıflandırmak mümkün olabilir. Bunlar intergenik lncRNA, antisense lncRNA, intronik lncRNA, enhancer lncRNA ve çift yönlü lncRNA'dır. İntergenik lncRNA protein kodlamakta olan iki gen arasındaki bölgeden transkribe olur. Antisense lncRNA, protein kodlamakta olan genlerin antisens yönünden transkribe olur. İntronik lncRNA, protein kodlamakta olan genlerin intronik bölgesinden transkribe olur. Enhancer lncRNA, genoma ait olan enhancer bölgelerinden transkribe olur. Çift yönlü lncRNA ise protein kodlamakta olan genlere ilişkin promotör bölgelerinden çift yönlü olacak şekilde transkribe olurlar. Olgunlaşmamış mRNA'ların bazılarının da ters kırılma ile lncRNA olarak kabul görülen circRNA'ları oluşturur (Şekil 2.7) (67).



Şekil 2.7. LncRNA'ların genomdaki sentez edildiği bölgelere göre sınıflandırılması (68)

2.5.8. LncRNA'ların İşlevleri

LncRNA'lar, gen ifadesinin her bir basamağında rol oynamaktadırlar. Çekirdekte genoma ilişkin yapının ve transkripsiyonun kontrolünü sağlarlar. Sitoplazmada ise mRNA'nın kararlılığının, translasyon ve translasyon sonrasında modifikasyonların kontrolünü sağlarlar. Bu bağlamda LncRNA'lar, önemli düzenleyiciler olarak kabul edilmektedirler (69). LncRNA'lar bulundukları konum değiştikçe mekanizmalardaki rolleri değişmektedir. LncRNA'lar ın çekirdekte ve sitoplazmada birçok düzenleme mekanizmalarında etkin olduğu bilinmektedir. Özellikle, kromatin ve DNA modifikasyonlarında, kromozomların konformasyonunda, RNA transkripsiyonunda, mRNA'nın olgunlaşma sürecinde, mRNA'nın bozulmasında, mRNA translasyonunda ve translasyon sonrasında modifikasyonların düzenlenmesinde ve bunun dışında pek çok hücrel süreci de etkilerler (Şekil 2.8) (68).



Şekil 2.8. LncRNA'lara ilişkin hücresel işlevler (68)

Çekirdek lncRNA'ları; enhancer RNA olduğunda transkripsiyonu, transkripsiyon faktörlerine ilişkin aktiviteleri, transkripsiyon düzenleyici olan komplekslerin gene bağlanmasını, olgunlaşmamış mRNA'da kesip-çıkarma işlem mekanizmasını ve kromozomların konformasyonuna etki edip gen ifadesini düzenleyebilmektedir. Sitoplazmadaki lncRNA'lar ise mRNA kararlılığını düzenleyip mRNA'nın ifadesini, mRNA'nın translasyonunu düzenleyebilmektedir. Ayrıca miRNA'ların üzerine toplanmasını sağlayarak mRNA'ya bağlanabilmesini engelleyebilmektedir ve biyolojik aktiviteye sahip olan kısa peptit dizilerinin bazıları için açık okuma çerçevesi (ORF) içerebilmektedir (68).

2.5.9. LncRNA'ların Hastalıklarla İlişkisi

LncRNA'ların transkripsiyonda, transkripsiyon sonrasında ve protein kodlamakta olan genlere ilişkin bazal düzenlenmede rol aldığı bilinmektedir. LncRNA'ların aşırı düzeyde ekspresyonunda, eksikliğinde veya mutasyonunda, pek çok hastalığa neden olmaktadır. Kolon kanseri, lösemi, prostat kanseri, meme kanseri, sedef hastalığı, hepatoselüler karsinoma ve Alzheimer gibi daha pek çok hastalık nedeni olarak lncRNA'ların düzensiz ekspresyon seviyesi ile ilişkilendirilmektedir (70). LncRNA'ların onkogenler, tümör baskılayıcılar veya terapötik potansiyelleri gibi davranabilme özellikleri vardır. Bu farklılıkların temel sebebi çok bileşenli olan komplekslere dahil

olabilme özelliği ve karmaşık yapıya sahip olmaları olarak düşünülmektedir. IncRNA'lar bir dokuya veya hücre tipine özgü olan gen ekspresyonunda değişiklik göstermesiyle, birçok kanser türünde tümör baskılayıcı ve onkojenik roller üstlendiği görülmüştür (71).

2.6. Biyoinformatik

Biyoinformatik kelimesi, ilk olarak 1970'li yıllarda Ben Hesper ve Paulien Hogeweg tarafından 'biyotik sistemlerde bilişim süreçlerinin incelenmesi' tanımıyla kullanılmıştır (72). Bilgisayar teknolojisinde yer alan gelişmeler ile beraber biyolojik verilerin üretilebilmesindeki kapasite artmıştır. Bu artışla üretilmiş olan büyük miktardaki veriyi verimli ve etkin bir biçimde işlenmesine olan ihtiyaç ile biyoinformatik bilimi gelişmiştir. Biyoinformatik bilgisayar bilimleri, biyoloji ve istatistik dallarından meydana gelen disiplinlerarası bir bilim dalı olarak bilinir. Biyoinformatik biliminin temelde üç amacı bulunmaktadır (73). Bu amaçlar, farklı araştırma gruplarının ürettiği verilerin depolanıp diğer çalışmalarda da kullanılması, bu verilerin etkili bir şekilde analiz edilmesini sağlayacak kaynakların ve araçların geliştirilmesi son olarak ise geliştirilmiş olan bu araçlar yardımıyla verilerin analiz edilerek biyolojik yönden anlamlı sonuçların elde edilmesini sağlamaktır. Bu doğrultuda, biyoinformatiğe ilişkin uygulama alanlarından bazıları kanser benzeri genetik hastalıkların sebeplerinin araştırılması, bu hastalıklar için etken olan genlerin tespit edilebilmesi ve ilaç keşifleridir. Biyoinformatiğin temel amaçlarından biri olan biyolojik verilerin depo edilmesi ve analizi yüksek başarımlı olan hesaplama gücüne ihtiyaç duymaktadır. Dahası, veri analizinde gereken araçların geliştirilmesinde yazılım geliştirme süreçleri kullanılmaktadır. Bunlar dışında, makine öğrenme, sinir ağları, veri madenciliği, veritabanının tasarımı ve geliştirilmesi, dinamik programlama, genetik algoritmalar, doğal dil işleme, hesaplamalı geometri, grafik ve grafik arayüzleri gibi konular bilgisayar bilimiyle ilgilenmekte olan insanların biyoinformatik bilimindeki çalışma alanlarıdır (74).

2.7. Diferansiyel Eksprese Edilmiş olan Genlerin Belirlenmesi

Biyolojik farklı iki durum arasında bulunan moleküler düzeydeki farkın anlaşılması için, diferansiyel eksprese edilmiş olan genlerin (kontrole göre farklı olarak ifade edilmiş olan genler); log2 kat değişiminin (log2FC) ve istatistiksel anlamlılık düzeyini gösteren p değerinin ayarlanıp, sayısal veriler kullanılarak tespit edilmesi gerekmektedir. İstatistiksel analizlerin sonucunda, hastalıklı olan bir grubun kontrolle kıyaslandığında

genlerin hangisinin hangi düzeyde ifade edildiğinin, hangi genlerin yukarı regülasyon veya hangi genlerin aşağı regülasyon gösterdiğini anlayabiliriz (75).

2.8. Yapay Zekâ, Makine Öğrenmesi ve Ağaç Tabanlı Yöntemler

Yapay zeka, insan gibi davranabilen, insandaki davranışları taklit eden, mantıklı karar verebilme yeteneği olan ve anlamlı bir şekilde cevap verebilen bir makine öğretilisidir. Bunları yaparken ise temel gereksinimi öğrenimdir. Pek çok bilim adamı yıllar boyunca bilgisayar zekasının öğrenme olmadığı sürece bir işe yaramayacağını düşünmüşlerdir. Fakat bilim adamlarının bu düşüncesi makine öğrenmesinin ve daha sonra derin öğrenmenin ortaya çıkması ile değişmiş ve hızlı bir şekilde gelişmiştir (76).

Makine öğrenmesi, yapay zekanın alt dallarından biridir ve büyük veri setlerinden kullanışlı, anlamlı bilgiler elde edebilmek için en önemli alanlardan biri olmaya başlamıştır. Makine öğrenme alanları günümüz teknolojileri gelişmesi ile birlikte giderek artmaktadır. Bunun nedeni ise verilerin boyutunun ve karmaşıklığının günden güne artmasıdır. Büyük boyutlu ve karmaşık verilerin analiz edilmesi daha da zorlaşmıştır. Bu durumda büyük boyutlu ve karmaşık verilerin analizinin yapılabilmesi için makine öğrenmesinin kullanımı zorunlu bir hal almıştır (76).

Ağaç tabanlı yöntemleri, problemi oluşturacak olan veri setlerinin yapısına bağlı olarak bir ağaç yapısı şekline sahip olan sınıflandırma modelleri ve regresyon modelleri oluşturmaktadırlar. Sözü edilen ağaç yapılarının oluşturulması için kullanılmakta olan karar kuralları anlaşılabilir olduğundan bu yöntemlerin kullanımı yaygın hale gelmiştir. Sınıflama ve regresyon modellerinde kullanılmakta olan ağaç tabanlı yöntemlerin ilk uygulandığı alan sosyal bilimler alanı olup sonrasında ise istatistik alanında yapılan çalışmalar sonucunda ağaç tabanlı yöntemler uygulamalı ve teorik yönlerinin ele alınması ile ağaç tabanlı yöntemler için yeni algoritmalar önerilmiştir (77). Yapay zekâ, makine öğrenmesi ve ağaç tabanlı yöntemlere ilişkin hiyerarşi Şekil 2.9’da verilmiştir.



Şekil 2.9. Yapay zekâ, makine öğrenmesi ve ağaç tabanlı yöntemlere ilişkin hiyerarşi

3. MATERYAL VE METOT

3.1. Veri Seti

İlacı bağı nefrotoksisiteye ilişkin olası biyobelirteçlerin saptanabilmesi, klinik düzeyde nefrotoksisite durumunun sınıflandırılmasının yapılabilmesi ve yapılan genomik analizlerin görselleştirilmesi amacıyla bir yazılım geliştirilmesi için tez kapsamında İnönü Üniversitesi Bilimsel Araştırma Projeleri (BAP) Birimi tarafından desteklenen TOA-2021-2593 kodlu projeden elde edilen veri seti proje yürütücüsünün izni alınarak kullanılmıştır. Tez çalışması kapsamında kullanılan veri seti lncRNA lardan oluşan bir veri seti olup bu veriler aşağıdaki koşullarda gerçekleşen bir sıçan deneyinden elde edilmiştir.

20 adet dişi Wistar Albino sıçanın (yaş: 3 aylık; ağırlık: 250±20 g) alınmasıyla deney düzeneği oluşturulmuştur. Deney düzeneğinde

- Kontrol grubu: Deneyin ilk günü taşıt çözücü olarak kullanılan serum fizyolojik intraperitoneal (i.p.) olarak uygulanmış olan grup.
- MTX grubu: Deneyin ilk günü 20mg/kg MTX tek doz i.p. olarak uygulanmış olan grup.

şeklindedir.

Deneyin 4. günü yüksek doz anestezi ile ketamin (225 mg/kg i.p.) ve ksilazin (24 mg/kg i.p.) sıçanlardan böbrek doku örnekleri alınarak genomik analizler yapılarak veriler elde edilmiştir. Tez çalışması için İnönü Üniversitesi Sağlık Bilimleri Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu Başkanlığı'ndan 09.05.20231 tarih ve 2023/4637 sayılı etik kurul onayı alındı.

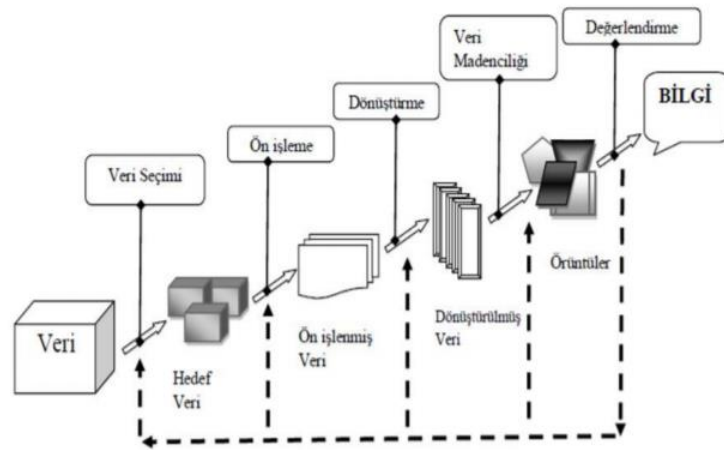
3.2. Veri Tabanlarında Bilgi Keşfi Süreci (VTBK) ve Tıbbi Bilgi Keşfi Süreci (TBKS)

Veri tek başına bir değer ifade etmez, bir amaç doğrultusunda işlendiği zaman bilgi meydana gelmektedir. Verinin bilgiye dönüştüğü bu sürece veri analizi denilmektedir. Günümüzdeki verilerden elde edilmiş olan bu bilgiler yakın gelecek içerisinde de kullanılabilecek ve gelecekte doğru tahminler yapabilmeyi sağlayacaktır. Elde edilmiş olan verilerden anlamlı bilgilerin üretildiği bu sürecin adı Veri tabanlarında Bilgi Keşfi (VTBK) süreci olarak bilinmektedir. Veri tabanlarında bilgi keşfi, depo

edilmiş olan veri içerisinde yer alan faydalı, yeni, geçerli ve sonuçta anlaşılabilir örüntülerin ortaya çıkarılmasını içeren bir süreçtir. Uygulama alanının öğrenilmesi ile bu sürecin ilk adımı başlamaktadır. Veri tabanlarında bilgi keşfinin son adımında ise, elde edilmiş olan bilgiyi görüntüleme yöntemleri ve bilgi gösterimi yöntemlerinin kullanılması ile araştırmacıya sunmaktır (78).

Veri madenciliği; elde edilmiş verilerden önceden bilinmeyen, net olmayan fakat kullanışlı hale gelebilecek potansiyeli olan bilginin ortaya çıkarılmasıdır. Diğer bir ifade ile veri madenciliği, büyük ve karmaşık veri kümelerinden anlamlı bilgiler elde edebilmek için, bilgisayar destekli olan bilgi çözümleme işlemi olarak da ifade edilebilir (79).

VTBK, üzerinde çalışılacak olan veri setinde gizli halde bulunan bilginin, örüntünün ve özelliklerinin ortaya çıkarılması için uygulanan ve 5 adımdan meydana gelen bir süreçtir. VTBK süreci, veri tabanlarından veri setinin elde edilmesi, veri önışleminin yapılması, veri dönüşümünün gerçekleşmesi ve indirgemesi, veri madenciliği ve veri madenciliğinin çıktılarının değerlendirilmesi ve yorumlanması adımlarından meydana gelir. VTBK süreci adımları tıp alanında kullanıldığında ise Tıbbi Bilgi Keşfi Süreci (TBKS) olarak ifade edilmektedir (80). VTBK süreci Şekil 3.1’de gösterilmiştir.



Şekil 3.1. VTBK süreci (81)

VTBK sürecinin adımları aşağıda ifade edildiği şekildedir:

- Veri Seçimi (Data Selection): Bu adımda birden çok veri kümesinden, üzerinde sorgulama yapılacak uygun olan örnek bir veri kümesinin oluşturulması aşamasıdır. Veri

toplama (data collection) ve çeşitli veri kümeleri içerisindeki verilerin birleştirilmesi işleminin yapılması işlemi de bu adımda yer almaktadır. Toplama, tanımlanmış olan problemde faydalı olabileceği düşünülen verilerin ve bu verilere ulaşılacak veri kaynaklarının karar verilmesi adımıdır.

- Veri Ön işleme (Data Pre-processing): Veri setinde yer alan hatalı değerler bulunur, veri setinde bulunan kayıp değerlerin değer ataması çeşitli yöntemler kullanılarak yapılır, aşırı/aykırı değerler çeşitli makine öğrenmesi yöntemleri ile tespit edilir ve veri setinden çıkarılır (82).

- Veri Dönüşümü (Data Transformation): Veri setinde yer alan farklı biçimdeki veri tiplerinin ve değerlerinin normalizasyon, Box-Cox, standardizasyon (z-dönüşüm), vb. dönüşümlerinin uygulanması ile veri setini meydana getiren gözlemler dönüşüm yöntemleri kullanılarak belirli bir standarda uygun hale getirilir.

- Veri İndirgeme (Data Feature selection/ Data Reduction): Seçilmiş olan veri setindeki değişkenlerin sayısı uygulanacak olan sınıflandırma yöntemlerine veya regresyon yöntemlerine bağlı olarak (wrapper) veya bu yöntemlerde bağımsız olacak şekilde (filter) yaklaşımları kullanılarak indirgenir.

- Veri Madenciliği (Data Mining): Bu adımda veri madenciliğinin yöntemleri ve algoritmaları uygulanır. Çeşitli kümeleme, sınıflandırma ve regresyon analizlerinin kullanılması ile veri setindeki bilgi ve örüntü çıkarılması işlemidir.

- Değerlendirme (Evaluation): Bu adımdan önceki adımlar sonucunda elde edilmiş olan veri madenciliği çıktılarının yorumlanması gerçekleştirilir ve modelin performansı değerlendirilir (83).

3.3. İş süreçlerinin anlaşılması

Bu adımda metotreksat ile nefrotoksisite oluşturulmuş sıçanlar ile patoloji oluşturulmamış sağlıklı sıçanlardan alınan böbrek doku örneklerinin genomik analizlerinin yapılması ile elde edilen büyük verilerin değişkenleri tahminleyici/açıklayıcı/girdi değişkeni, nefrotoksisite'nin varlığı/yokluğu ise çıktı/hedef/sonuç değişkenleri olarak alınacaktır.

3.4. Verinin anlaşılması

Böbrek doku örneklerinin genomik analizlerinin sonucunda elde edilen büyük veriler üzerinde analizlere başlamadan önce veride kayıp/eksik, aşırı/aykırı, hatalı veya sapan gözlemlerin olup olmadığının tespit edileceği adımdır. Çalışmadaki verideki eksiklikler, hatalar tanımlanıp giderilmesi için yapılacak işlemlere karar verilecektir.

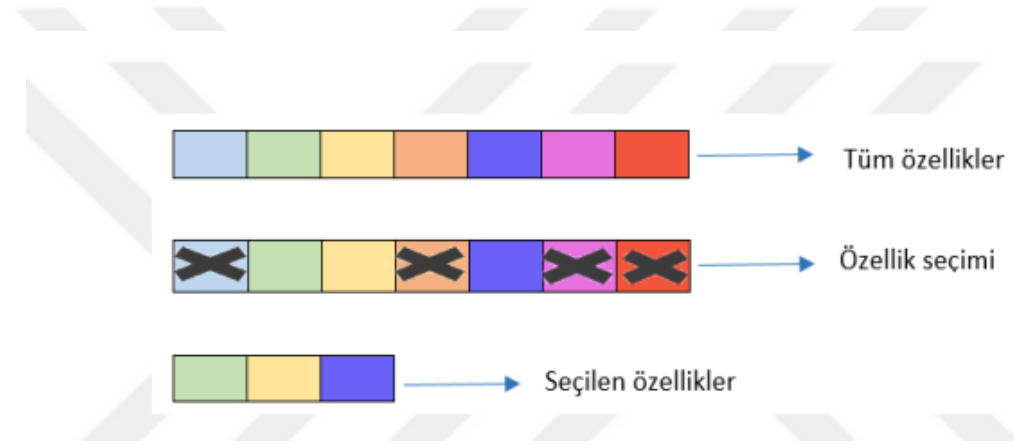
3.5. Veri ön işleme aşaması

Bu adımda ise kullanılacak olan büyük veride kayıp değer bulunuyorsa, kayıp/eksik değer sayılarının değişken bazındaki durumu göz önünde bulundurularak, kayıp değerlerin veriden çıkartılması veya kayıp değerler yerine veriye yeni gözlemlerin eklenmesi ile sorun çözülebilir. Veriden eksik olan değerlerin çıkartılması gözlem sayısının ciddi bir şekilde azalmasına ve yeterli olan örneklem sayısının yetersiz hale dönüşmesine neden olabilir. Bu durumda ise istatistiksel analizlerin gücünde azalmaya sebebiyet verecektir. Ayrıca kayıp değerlerin veriden çıkartılması için kayıp değerlerin rasgele dağılmış olması gereklidir. Aksi takdirde yanlılık problemi de oluşabilir (84). Bu problemlerden dolayı kayıp gözlemler olduğunda ortalama atama, yakın noktalar medyan ataması gibi basit düzeyde atama yöntemi kullanılabilir. Ancak bu çalışmada ilgili değişken(ler)in yapısına dikkat ederek lojistik regresyon, random forest, vb. modeller kullanılarak kayıp değer ataması yapılarak tamamlanacaktır. Ayrıca bu aşamada gerekli görülürse verilerden nicel olanlar için veri dönüşümü yöntemlerinden standardizasyon, normalizasyon gibi yöntemler ile uygun formlara dönüştürülecektir. Kullanılacak modellerin eğitilmesi sırasında oluşabilecek aşırı/eksik uyum problemleri ve modellerin performans başarısını arttırabilmek için değişken seçim yöntemlerinden Elastik net, Lasso veya Boruta yöntemlerinden uygun olan(lar)ı kullanılacaktır. Böylece verinin güvenilirliği arttırılmış olacaktır. Bu aşamanın sonunda veride gerekli ön işlemler yapılmış olup amaca uygun bir şekilde zaman kaybına sebep olmadan veri modelde işlenmeye hazır olacaktır.

3.5.1. Değişken seçimi

Değişken günümüzde yaşanan bilgi teknolojilerindeki meydana gelen gelişmeler ve yeniliklerle birlikte veri toplamak ve depolamak oldukça kolay hale gelmiştir. Böylece büyük veri tabanlarına ulaşılabilirlik daha da artmış ve pek çok alanda büyük veri tabanları oluşturulmuştur. Bunun sonucunda ise bu veri tabanlarında depolanmakta olan veri miktarı da katlanarak artmaya başlamıştır. Teknolojideki bu gelişmelere paralel

olarak veri toplama ve depolama maliyetlerinde de düşüş olmuştur. Fakat elektronik ortamda depo edilen bu verileri insanların kolayca anlaşılabilir olacak şekilde bilgi haline getirebilmesi bilgi teknolojisindeki gelişme ile aynı paralellikte gelişmemiştir. Söz konusu olan veri yığınlarını analiz edilebilmesinde geleneksel yöntemler yetersiz kalmıştır. Ayrıca verilerin boyutunun önemli ölçüde artış göstermesiyle makine öğrenmesi ve veri madenciliği alanlarında bulunan mevcut öğrenme algoritmalarının performanslarında ciddi sorunlara yol açmaktadır (85). Özellik seçimi adıyla da bilinen değişken seçimi, başlangıçtaki özelliklerin sayısının azaltıldığı ve modellemede daha iyi performans elde edebilmek için yeterli bilgiyi sahip olan ham veri setini temsil edebilen en iyi alt kümenin seçildiği süreçtir. Şekil 3.2’de ham veriyi temsil edebilen bir özellik alt kümesi seçimine örnek verilmiştir.



Şekil 3.2. Özellik seçimi yardımıyla bir özellik alt kümesinin seçimi

Özellik seçimi konusunda yapılan çalışmalara 1960’larda başlanmıştır. Büyük veri kümelerinin artışı ile pek çok alanda özellik seçimine gösterilen ilgi artmıştır. Mikrodizi teknolojisi, kanser hastalıkları başta olmak üzere tıp biliminde birçok alanda araştırmacılara yarar sağlayabilecek binlerce gen açıklama veri setine eş zamanlı ulaşabilme olanağı sağlamıştır. Araştırmacılar, mikrodizi teknolojisi ile birlikte gelişen ve çok büyük boyutlu olan gen açıklama verilerinde binlerce gen tarafından temsil edilen az sayıda bulunan örneklem sorunu ile karşı karşıya gelmişlerdir (86). Buradaki gen sayısının çok büyük olması sınıflandırmada bazı dezavantajlara neden olması durumu söz konusu olabilir. Özellik seçimi yardımıyla bu dezavantajların yok edilmesi ve birçok avantaja dönüşmesi sağlanacaktır. Özellik seçimi ile sınıflandırma performansını arttırmayan, bilgilendirici olmayan, alakasız ve gereksiz özellikler kaldırılır. Uygun özellikler seçilerek modeldeki değişkenler azaltılarak karmaşık olmayan bir model elde

etmek, aşırı öğrenmeyi önlenmesini sağlamak ve genelleme performansını yükseltmek, değişken sayısı azaltılarak model eğitiminde gerekli olan süreyi azaltmak, veri toplamada ve depolamada maliyeti azaltmak mümkün hale getirilebilecektir. Araştırmacılar bu faydaları dikkate aldığına, binlerce gen arasından hastalıkları belirleyebilmede önemli olan genlerin belirlendiği yani bir başka ifade ile özellik seçimi yöntemlerinin kullanıldığı ve özellik seçimi sonrasında seçilmiş olan bu genler ile sınıflandırmanın gerçekleştiği çalışmalara yönelmişlerdir (87).

3.5.2. Lasso

En Küçük Mutlak Büzölme ve Seçim Operatörü (LASSO) özellik seçimi 1996 yılında ilk kez Tibshirani tarafından En Küçük Kareler tahmin doğruluğunu artırabilmek için önerilmiş ve zaman içerisinde de kullanım alanları fazlaca genişletilmiştir. Günümüzde ise başta sağlık alanı olmak üzere pek çok farklı alanlarda kullanılmaktadır (88). Kullanıldığı alanlar göz önüne alındığında, özellikle çok büyük boyutlu veri setlerinde gözlem sayısının değişken sayısından az olduğu veri setlerine odaklanıldığı çalışmalar olduğu görülmektedir. LASSO tahmincisi, En Küçük Kareler tahmin denklemine belirli bir ceza alarak tahmin doğruluğunu artırmayı amaçlar. LASSO tahmin edicisinde, ceza kısıtlamaları nedeniyle tahmin denkleminde bir veya daha fazla değişken sıfıra daraltılır. Böylece özellikle büyük boyutlu veri setlerinde yorumlanması kolay modellerin elde edilmesini sağlar. LASSO, özellik seçimi ve değişken kestirimini aynı anda gerçekleştirerek işlem kolaylığı sağlaması nedeniyle yaygın olarak tercih edilen bir kestirim aracıdır (89).

3.5.3. Elastik net

LASSO değişken seçimi yöntemindeki bazı eksiklikleri giderebilmek için Zou ve Hastie tarafından 2005 yılında Ridge ve LASSO değişken yöntemlerinin bir arada kullanılması ile Elastik Net değişken seçimi yöntemi önerilmiştir. Kısaca bir şekilde bu eksiklikten bahsedilecek olursa; gözlem sayısının değişken sayısından az olduğu durumlarda ($n < p$) LASSO değişken seçimi yöntemi en çok n değişken seçebilir ve bu durumda bu değişken seçimi yöntemi kısıtlayıcı olabilmektedir. Dahası veri setinde aralarında yüksek korelasyon bulunan değişken grupları olduğunda LASSO yöntemi değişken grubunun içinden yalnızca bir tanesini modele dahil etmektedir ve diğer değişkenleri ise modele dahil etmeyip dışarı atmaktadır. Genomik veri setlerinde yani gen ekspresyonu ile elde edilen veri setlerinde benzer işlevli genler birbiri ile yüksek

korelasyonlu olduğundan modelleme sürecinde grup halinde modele eklenmesi gerekmektedir. Bu durumda Elastik Net değişken seçimi yöntemi önerilmektedir. Elastik Net yöntemi değişken seçimi yaparken ve katsayılarda daraltma yaparken ilişkili olan değişken gruplarını da seçebilme yeteneğine sahiptir (90).

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^m \beta_j X_{ij})^2 + (1 - \lambda) \sum_{j=1}^m \beta_j^2 + \lambda \sum_{j=1}^m |\beta_j| \quad (1)$$

Yukarıdaki eşitlik (1)'de λ "0" alındığında denklem Ridge değişken seçimi yöntemine; "1" alındığında ise LASSO değişken seçimi yöntemine dönüşmektedir.

3.5.4. Boruta

Boruta değişken seçim yöntemi ilk kez 2010 yılında Kursa ve Rudnicki tarafından önerilmiştir. Boruta değişken seçim yöntemi Rastgele Orman sınıflandırma algoritması etrafında bir sarmalayıcı olarak yani Random Forest sınıflandırıcısının temelini oluşturan fikre dayanarak tasarlanmıştır. Bu yöntemde sisteme rasgele değişkenler ekleyip ve rasgele örneklemeler topluluğundan sonuçlar toplayıp rasgele değişimlerin ve korelasyonların yanıltıcı etkisini azaltabilir. Burada, bu fazladan rasgelelik eklenmesiyle hangi değişkenlerin gerçekten önemli olduğuna dair daha net bir görüş sağlanması amaçlanmıştır.

Gölge unsurları adı verilen tüm özelliklerin karışık kopyalarını oluşturarak verilen veri setine rasgelelik ekler. Ardından, bu genişletilmiş veri kümesinde (orijinal değişkenler artı gölge değişkenleri) rasgele bir orman sınıflandırıcısı eğitir ve Ortalama Azalma Doğruluğu gibi bir özellik önem ölçüsü uygular ve her özelliğin önemini değerlendirir. Her yinelemede, Boruta Algoritması gerçek bir özelliğin daha yüksek öneme sahip olup olmadığını kontrol eder. Gölge özelliklerinin en iyisi ve çok önemsiz sayılan özellikleri sürekli olarak kaldırır. Son olarak, Boruta Algoritması, tüm özellikler onaylandığında veya reddedildiğinde veya belirli bir rasgele orman sınırına ulaştığında durur. Öte yandan Boruta, yanıt değişkeniyle güçlü veya zayıf bir şekilde ilgili olan tüm özellikleri bulur. Bu, hangi insan genlerinin (öngörücüler) belirli bir tıbbi duruma (yanıt değişkeni) bir şekilde bağlı olduğunu belirlemekle ilgilenilebilecek biyomedikal uygulamalar için çok uygundur. Boruta yönteminin çalışma prensibi aşağıda verilmiştir (91).

1. Tüm değişkenlerin gölge değişkenler denilen kopyaları eklenip veri seti birleştirilir.
2. Korelasyonları yok edebilmek için eklenmiş olan değişkenler kendi aralarında karıştırılır.
3. Birleştirilmiş veri setine Random Forest sınıflandırma yöntemi uygulanır ve Z skorları elde edilir.
4. Gölge değişkenler içinde en yüksek olan Z skoru belirlenir ve bu gölge değişkenden daha yüksek skorlu orjinal değişkenler belirlenir.
5. Her bir değişken için, en yüksek skora sahip olan gölge değişken ile istatistiksel analizler uygulanır.
6. En yüksek Z skora sahip olan gölge değişkenden istatistiksel açıdan önemli ölçüde daha yüksek öneme sahip olan değişkenler “önemli” kabul edilirken daha düşük öneme sahip olan değişkenler “önemsiz” kabul edilir.
7. Bütün gölge değişkenler veri setinden çıkarılır.
8. Tüm değişkenlere bir önemlilik ölçüsü atanıncaya ya da algoritmanın önceden belirlenmiş olan Random Forest sınırlarına ulaştığında işlemler sonlandırılır.

3.6. Model aşaması

Tanımlanan problem için hazırlanmış olan veri setine en uygun modelin bulunabilmesi, çok sayıda modelin kurularak denenmesi ile mümkündür. Bu aşamada makine öğrenimi modelinin kurulması, en iyi olduğu düşünülen modele varılıncaya kadar tekrarlanır. Modelin iyileştirilmesi için uygun olan optimizasyon yöntemleri kullanılabilir (92).

Bu çerçevede, çalışmada VTBK sürecinde model aşamasındaki işlemler için kullanılması planlanan modellere ve algoritmalara yönelik açıklamalar aşağıda sunulmuştur.

3.6.1. Sınıflandırma yöntemleri

Sınıflandırma, bir veri kümesinde önceden belirlenmiş olan çıktılarına uygun şekilde tanımlı olan farklı sınıflara veriyi dağıtmaktır. Sınıflandırma yöntemleri, verilen eğitim kümesi yardımıyla bu dağılımın şeklini öğrenerek sınıfının belirli olmayan test verileri geldiği zaman doğru bir şekilde sınıflandırmaya çalışırlar (93). Veri kümesinde bulunan bu sınıfları belirtmeye yarayan değerlere etiket adı verilir ve hem eğitim hem de test

sırasında verinin hangi sınıfta olduğunu belirlenmesi amacıyla kullanılırlar. Sınıflandırma işlemini gerçekleştirebilmek için istatistik ve makine öğrenimine dayanan pek çok yöntem geliştirilmiştir (94). Bunlardan doğrusal regresyon analizi, lojistik regresyon analizi, ayırma analizi istatistik temeline dayanan sınıflandırma yöntemlerinden bazılarıdır. Karar ağaçları, en yakın komşu yöntemi, yapay sinir ağları, destek vektör makineleri gibi yöntemler ise başlıca makine öğrenimi temelli sınıflandırma yöntemleri olarak bilinir. Genetik Algoritmalar, karar Ağaçları, K-En Yakın Komşu, Naive-Bayes, Yapay Sinir Ağları gibi bazı yöntemler de makine öğrenimi temeline dayana sınıflandırma yöntemlerinden bazılarıdır.

3.6.2. Ağaç Tabanlı Topluluk Öğrenme Yöntemleri

Aynı problemlere ilişkin ortak çözümleri içerisinde bulunduran, güvenilirlik açısından ise en yüksek tahminleri üretebilen modeli üretmeyi amaçlayan, bir diğer ifade ile nihai kararların alınması aşamasında birden çok görüşü dikkate alarak onlar arasında en yüksek güvenilirliği olan sonuçları araştırmacıya sunan yöntemler olarak tanımlanmaktadır (11).

Ağaç tabanlı topluluk öğrenme yöntemlerinin ilk yıllarında makine öğrenimi yöntemlerinde genel olarak bilinen bir sorun yüksek varyans problemini azaltarak ve böylece makine öğrenimi yöntemlerinin doğruluğunu arttırmıştır. Ayrıca kısa zaman içerisinde makine öğrenimi yöntemleri arasında en sık karşılaşılan kayıp özellikleri bulma, öz değer seçimi ve veri ön işleme aşamasında başarılı çözümler sunmuştur (95).

3.6.3. Karar Ağaçları

Karar ağaçları 1973 yılında Bierman ve Friedman tarafından önerilmiştir. Değişkenleri parçalayıp bir ağaç oluşturulması temeline dayanmaktadır. Karar ağaçları, veri madenciliğinin ve bilgi keşfinin hem popüler hem de güçlü yöntemlerinden biri olarak bilinmektedir. Karar ağaçları, verinin içerisinde yer alan kuralların hiyerarşik ve sıralı olacak şekilde gösterilmesini sağlayan yöntemdir. Karar ağaçları, araştırmacının karşılaştığı probleme yönelik var olan bilgiyi daha anlaşılır halde gösterimini sağlayan, karar seçenekleri ile olasılıklara bağlı olan durumların belli bir düzen çerçevesinde sıralama yaparak ortaya koymaya yarayan görsel bir modelleme tekniğidir. Böylece, karar ağaçları yöntemi hem kararları hem de sonuçlarını gösterebilen hiyerarşik bir modeli ifade etmektedir. Herkesin kolayca anlayabileceği grafiksel yapısı ve kuralları olduğu için pek çok alanda yaygın bir şekilde kullanılmaktadır (96). Veri madenciliğinde,

sınıflandırma modellerinden biri olan karar ağaçları modeli, öngörü değerine sahip olan bir modeldir. Karar ağaçlarının ilk aşamasından başlanarak son karar seçeneğine kadar sorular sorulur ve bu sorulara verilen cevaplar doğrultusunda karar ağacının yapısı meydana getirilir, bu karar ağacının yapısına bakılarak kurallar (eğer-ise kuralı) yazılabilmektedir (97).

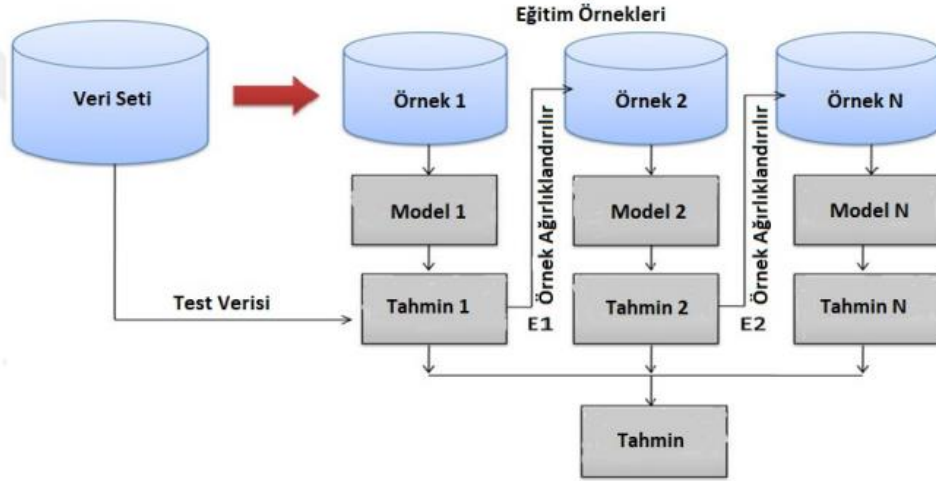
3.6.4. Random Forest (RF)

Breiman tarafından 2001 yılında önerilen rastgele orman yaklaşımı, Bagging ve Random Subspaces yöntemlerinin kombinasyonu olan birçok karar ağacına sahip bir makine öğrenme algoritmasıdır. RF algoritmasında, çok sayıda karar ağacının hesaplamalarının nihai bir sonuç elde etmek üzere birleştirildiği denetimli bir makine öğrenimi algoritmasıdır. Basit ama etkili olduğu için popülerdir. Kullanım kolaylığı ve esnekliği sayesinde, hem sınıflandırma hem de regresyon problemlerini ele aldığı için benimsenmesini hızlandırdı (98). 1996'da Leo Breiman torbalama yöntemini tanıttı; bu yöntemde, bir eğitim setindeki rasgele bir veri değiştirilerek seçilir; bu, bireysel veri noktalarının birden fazla seçilebileceği anlamına gelir. Birkaç veri oluşturulduktan sonra, bu modeller daha sonra bağımsız olarak eğitilir ve regresyon veya sınıflandırma türüne bağlı olarak, bu tahminlerin ortalaması veya çoğunluğu daha doğru bir tahmin verir. Rastgele orman algoritması, ilişkisiz bir karar ağaçları ormanı oluşturmak için hem torbalamayı hem de özellik rasgeleliğini kullandığından, torbalama yönteminin bir uzantısıdır. Özellik toplama veya " rasgele altuzay yöntemi " olarak da bilinen özellik rasgeleliği, karar ağaçları arasında düşük korelasyon sağlayan, özelliklerin rasgele bir alt kümesini oluşturur. Bu, karar ağaçları ile rastgele ormanlar arasındaki temel farktır. Karar ağaçları tüm olası özellik ayrımlarını dikkate alırken, rastgele ormanlar bu özelliklerin yalnızca bir alt kümesini seçer. RF algoritmasında veri seti önce rastgele iki kısma ayrılır: öğrenme için eğitim verileri ve RF algoritmasında öğrenme düzeyini test etmek için doğrulama verileri. Bunu takiben, veri setinden "bootstrap yöntemi" kullanılarak rastgele birçok karar ağacı oluşturulur. Düğüm konumlarında rastgele seçilen belirleyiciler, her ağacın dallanmasını belirler. RF sonuç tahmini, ağacın tüm sonuçlarının ortalamasıdır. Sonuç olarak, belirli ağırlıklar için, her bir ağacın RF tahmini üzerinde etkisi vardır. Bu yöntem bir "kara kutu" özelliği gösterdiğinden, her ağaç ayrı ayrı değerlendirilmez. Alt kümelerden eğitim verilerini rastgele kabul etme ve rastgele yöntemlerle ağaçlar oluşturma yeteneği nedeniyle, RF algoritması diğer makine öğrenimi algoritmalarından daha iyi performans gösterir. Ayrıca, bootstrap örnekleme kullanılarak rastgele seçilen

farklı alt veri kümeleri üzerinde eğitim yapıldığından, rastgele orman algoritması aşırı uydurma seviyesini korur. (99, 100).

3.6.5. Arttırma (boosting) algoritması

1989 yılında Schapire önderliğinde geliştirilmiştir. Schapire, Freud ve Friedman'ın yaptığı çalışmalar ile daha da geliştirilmiştir. Arttırma, tahmine dayalı veri analizindeki hataları azaltmak için makine öğreniminde kullanılan bir yöntemdir. Arttırma algoritmasının prensibi, hatadan öğrenmeyi amaçlayan, yavaş öğrenmeye dayanan sıralı bir yöntemdir. Arttırma algoritmaları, yüksek performanslı modelleri oluşturabilmek için düşük performanslı birkaç modeli birleştirmektedir. Böylece birden çok zayıf öğreniciyi tek bir güçlü öğrenme modeline dönüştürerek makine modellerinin tahmin doğruluğunu ve performansını iyileştirir (101). Arttırma algoritması sürecinde; İlk olarak, eğitim verilerinden bir model oluşturulur. Daha sonra birinci modelde var olan hataları düzeltmek için ikinci model kurulur. Bu prosedüre devam edilir ve tam eğitim veri seti doğru bir şekilde tahmin edilene veya maksimum model sayısı eklenene kadar modeller eklenir. Nihai tahminler, tüm bireysel modellerin tahminlerinin birleştirilmesiyle yapılır. Arttırma, regresyon ve sınıflandırma problemlerinde kullanılabilir ve doğrusal olmayan verilerle uğraşmak için güçlü bir araçtır. Ayrıca aşırı öğrenmeye karşı güçlüdür, yani genelleştirilebilirlikten ödün vermeden genellikle yüksek düzeyde doğruluk elde edebilir. Tahmine dayalı modellerin doğruluğunu arttırdığı da iyi bilinmektedir. Bu nedenle arttırma, birçok pratik makine öğrenimi uygulaması için popüler bir seçimdir (102). Arttırma algoritmaları kapsamında, Ağaç tabanlı yöntemlerden olan Gradient Boosting, AdaBoost, CatBoost, XGBoost, LightGBM, Skolastik GradyanBoost yöntemleri, vb. bulunmaktadır. Boosting yöntemine ilişkin akış diyagramı Şekil 3.3'de verilmiştir.



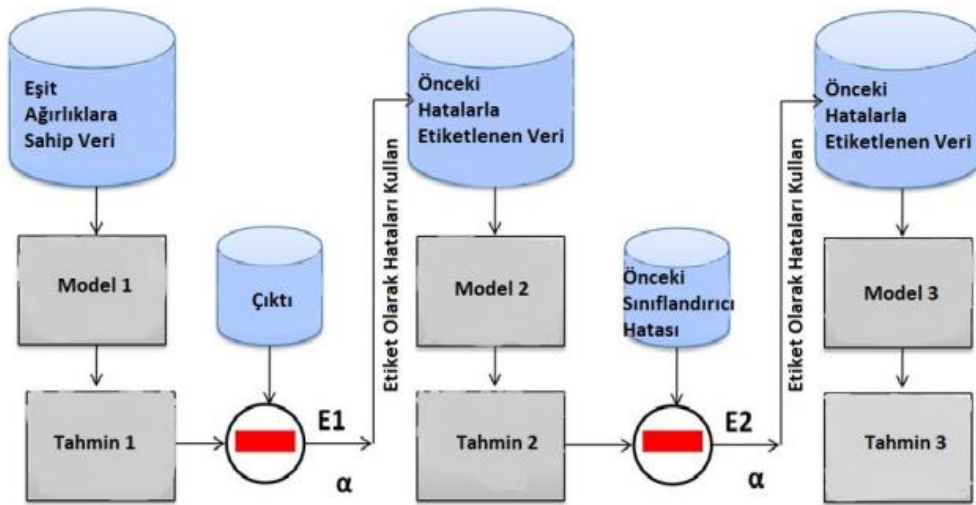
Şekil 3.3. Boosting Akış Diyagramı (102)

3.6.6. AdaBoost algoritması

Adaptive Boosting, Robert Schapire ve Yoav Freund tarafından 1996 yılında önerilmiş olan makine öğreniminde topluluk hızlandırıcı sınıflandırıcı olarak kullanılan bir Boosting tekniğidir. Her bir örneğe ağırlıklar yeniden atandığı için ve yanlış sınıflandırılmış olan örnekler ise daha yüksek ağırlıklar atanmasından dolayı, Adaptive Boosting olarak adlandırılır. Güçlendirme, denetimli öğrenme için yanlılığı ve varyansı azaltmak için kullanılır. Öğrenenler ilkesi sırayla büyüye göre çalışmaktadır. İlki haricinde, sonraki her bir öğrenen, önceden yetiştirilmiş olan öğrenenlerden yetiştirilir. Basit bir ifade ile, zayıf öğrenciler güçlü öğrencilere dönüştürülür (103). Adaboost yöntemindeki temel kavram, sınıflandırıcılara ait olan ağırlıkları ayarlamak ve veri örneğini olağandışı gözlemlerin doğru tahminlerini sağlayacak şekilde her yinelemede eğitmektir. Adaboost algoritmasının çalışma prensibinin adımları şu şekildedir. İlk olarak, Adaboost rasgele olacak şekilde bir eğitim alt kümesini seçer. Son eğitimin doğru tahminine dayanarak eğitim setini seçip AdaBoost makine öğrenimi modelini yinelemeli olacak şekilde eğitir. Yanlış sınıflandırılmış olan gözlemlere daha yüksek ağırlık atanır ve böylece bir sonraki iterasyonda bu gözlemler sınıflandırmada yüksek olasılık elde eder. Ayrıca, sınıflandırıcının doğruluğuna bağlı olarak her adımda eğitimli sınıflandırıcıya ağırlık atar. Daha doğru olan sınıflandırıcı yüksek ağırlık alacaktır. Bu süreç, tüm eğitim verileri hatasız olarak uyana veya belirtilmiş olan maksimum tahmin edici sayısına ulaşana kadar yinelenir (103).

3.6.7. Gradient Boosting algoritması

Gradient Boosting, 2001 yılında Friedman öncülüğü ile tanıtılmış olan güçlü bir makine öğrenimi yöntemidir. Gradient Boosting, sınıflandırma ve regresyon ve problemlerinde kullanılabilen bir tahmin modeli olup genellikle karar ağaçlarına benzer şekilde zayıf tahmin modellerinin topluluk formunu üretir. Gradient Boosting algoritmasının temeli artırma yöntemlerine dayanmaktadır. Arttırma yöntemine dayandığı için sırasıyla pek çok sayıda zayıf öğreniciyi oluşturmayı ve zayıf öğrenicileri karmaşık bir modelle birleştirmeyi amaçlamaktadır. Gradient Boosting yönteminin çalışma prensibi şu şekildedir. İlk olarak ilk veri seti kullanılarak bir karar ağacı meydana getirilir. Böylece tahmin edilen ile çıktının arasındaki hata elde edilir. Veri kümesindeki örnekler için bu hata değeri yeni çıktı değerleri olacak şekilde kullanılır. Veri setindeki hatalar kullanılarak yeni bir karar ağacı meydana getirilir. Önceki ağacın ürettiği hatayı tekrar meydana getirmek için ağaç eğitilir. Önceki çıktı ile tahmin arasında elde edilen hata istenilen düzeye ulaşana kadar ağaç ekleme işlemine devam edilir (104). Gradient Boosting adımları Şekil 3.4’de verilmiştir.



Şekil 3.4. Gradient Boosting Adımları (104)

3.6.8. XGBoost (Extreme Gradient Boosting) algoritması

Extreme Gradient Boosting kısaltması olan XGBoost, temel olarak gradient boosting ve karar ağacı yöntemlerine dayanan bir makine öğrenimi algoritmasıdır. XGBoost yöntemi 2002 yılında Friedman önderliğinde geliştirilmiştir. 2016 yılındaki konferansta Washington Üniversitesi'nden iki araştırmacı Carlos Guestrin ve Tianqi

Chen tarafından makale olarak sunulmuşlardır ve böylece makine öğreniminde oldukça popüler hale gelmiştir (105). XGBoost, finans, enerji, sağlık vb. birçok alanda kullanılmıştır. Performans ve hız açısından diğer algoritmalar ile kıyaslandığında oldukça avantajlıdır. Dahası XGBoost, yüksek tahmin performans gücü sayesinde diğer yöntemlerden 10 kat daha hızlı olduğu bilinmektedir. Bunlara ek olarak XGBoost yöntemi, aşırı öğrenmeyi azaltabilen ve genel performansı iyileştirmeyi sağlayan bir dizi regularizationu içermektedir. XGBoost yöntemi de Gradient boosting yöntemi gibi, bir dizi zayıf öğreniciyi arttırma ile birleştirerek güçlü bir öğrenici oluşturan topluluk yöntemidir. Bu iki yöntem arasındaki temel farklılıklar uygulama detaylarında gizlidir. Farklı regularization teknikleri kullanan XGBoost, ağaçların karmaşıklığını kontrol edebildiği için daha iyi performans gösterebilmektedir (106). XGBoost'un sahip olduğu avantajları Şekil 3.5'de gösterilmiştir.



Şekil 3.5. XGBoost'un avantajları (102)

3.6.9. Torbalı Sınıflandırma ve Regresyon Ağaçları (Bagged CART)

Sınıflandırma ve Regresyon Ağaçları (CART), bir sınıflandırıcı olarak yaygın olarak kullanılan parametrik olmayan bir karar ağacı yöntemidir. Bu ağaç ve diğerleri, Rastgele Orman gibi daha karmaşık algoritmaların temelini oluşturur. Algoritması ikili ağaçlara dayanmaktadır. CART karar ağacı yöntemi, karar ağacını oluşturmak için verileri ikili bölümlere ayırır. CART ağacı, hangi değişkenlerin veri sınıflandırması için daha fazla bilgi aldığını belirlemek için Gini indeksini kullanır. Gini indeksi düşük olan

değişkenlere kategorize edilirken daha fazla ağırlık verilir. CART, en düşük Gini indeksine sahip her boyutta veya değişkende ayırıcı nokta için en iyi değeri belirlemek üzere test ve hata yöntemini kullanır. Verilerdeki küçük bir değişiklik, CART gibi (stabil olmayan algoritmalar olarak da adlandırılır) karar ağaçlarında farklı sınıflara neden olabilir. Bu sorunun üstesinden gelmek için birçok sınıflandırıcıyı bünyesinde barındıran topluluk sınıflandırıcıları kullanılmaktadır. Torbalama, topluluk sınıflandırma yöntemlerinden biridir. Torbalama, karar ağacı tahmin doğruluğunu artırmak için en sık kullanılan topluluk yaklaşımlarından biridir. Bu stratejide, her sınıflandırıcı, verinin bir bölümünü sınıflandırarak modelini oluşturur ve kaydeder. Son olarak bu sınıflamalar arasından oylama yöntemi kullanılarak en yüksek oyu alan sınıf son sınıflandırıcı olarak seçilir (107, 108).

3.6.10. Açıklanabilir Yapay Zeka Yöntemleri

Makine öğrenme ve yapay zeka modelleri, tahmin doğruluğu, süreç verimliliği ve araştırma üretkenliği açısından iyidir. Ancak mevcut makine öğrenme/yapay zeka modelleri, yorumlama sürecini ve tahmin sonuçlarını açıklamakta genellikle zayıftır. Bu durum, oluşturulan tahmin modellerinin anlaşılmasının önünde bir engel haline gelmektedir. Sonuç olarak klinisyenler veya sağlık çalışanları, bir modelin sonuçlarının veya tahminlerinin yeterli düzeyde açıklayıcı olmadığını düşünebilirler (15, 16). Bundan dolayı oluşturulan yapay zeka tabanlı modellerin açıklanabilirliğini sağlamak için Yerel Yorumlanabilir Model-Agnostik Açıklama (LIME) yöntemi, Shapley Katkı Açıklamaları (Shapley Additive Explanations), Kısmi Bağımlılık Grafikleri ve Bireysel Koşullu Beklenti (Individual conditional expectation) grafiği, vb. sıklıkla kullanılan yöntemlerdir. Bu yöntemleri kullanmaktaki amaç, oluşturulan modellerin yorumlanabilirliğini ve açıklanabilirliğini artırmaktır.

3.6.11. Yerel Yorumlanabilir Model-Agnostik Açıklama

Makine öğrenimi, bilim ve teknolojideki son gelişmelerin çoğunun merkezinde yer almaktadır. Makine öğrenimi modellerinin çoğunun, iç işleyişi hakkında bir fikir edinmek neredeyse imkansız olduğundan, işlevsel olarak kara kutulardır. Makine öğrenimi kara kutu modellerini yorumlama yöntemleri, sonuçların şeffaflığını artırır ve karşılığında algoritmaların güvenilirliği ve adaletine ilişkin içgörü oluşturur. Bununla birlikte, yorumların kendileri, sonuçlara olan güveni sarsan ve modelin güvenilirliği konusunda endişelere yol açan önemli belirsizlikler içerebilir. Bundan dolayı oluşturulan

yapay zeka tabanlı modellerin açıklanabilirliğini sağlamak için Yerel Yorumlanabilir Model-Agnostik Açıklama (LIME) yöntemi sıklıkla kullanılmaktadır. LIME'ı kullanmaktaki amaç, oluşturulan modellerin yorumlanabilirliğini ve açıklanabilirliğini artırmaktır.

Sağlık alanında hastalığı teşhis etmek için sınıflandırma modellerinin kullanımı, büyük ölçüde oluşturulan modellerin araştırmacı tarafından yorumlanabilmesine ve açıklanabilmesine bağlıdır. Sağlık alanında oluşturulan yapay zeka modellerinin açıklanabilirliğini artırmanın birçok farklı yolu vardır ve değişken önemliliği bunlardan biridir. Bu amaçla kullanılan LIME, belirli bir sınıflandırma için hastaya özel bir açıklama sağlar, böylece herhangi bir karmaşık sınıflandırıcının klinik ortamda daha basit bir şekilde açıklanmasına olanak tanır. LIME, verideki her bir değişkenin modeldeki her bir tahmine (hastaya özel) ne kadar katkıda bulunduğunu belirleyebilir (109). LIME yöntemini kullanarak, hangi değişkenlerin modeldeki her bir tahmini hangi derecede ve hangi yönde etkilediğini veya hangi değişkenin modeldeki her bir tahminin sonuçları üzerinde diğer değişkenlere kıyasla daha fazla etkiye sahip olduğu belirlenebilmektedir. Sonuç olarak LIME yöntemi herhangi bir sınıflandırma modelini kullanarak her bir tahmin için açıklanabilirlik sağlamaktadır (110).

3.6.12. Shapley Katkı Açıklamaları

Makine öğrenimi modellerini yorumlamadaki zorluklar ve bunların tahminleri, ML'nin pratik uygulanabilirliğini ve güvenini sınırlar. Model yorumlanabilirliği genellikle bireysel özelliklerin (bağımsız değişkenler) modelin sonuçlarına katkısının tahmin edilmesine bağlıdır. Makine öğrenimi modellerinin yorumlanmasına yardımcı olacak açıklanabilir yaklaşımlara ihtiyaç vardır. Bu amaca yönelik olarak, SHapley Additive Explanations (SHAP) metodolojisi yakın zamanda tanıtılmıştır (111).

SHAP, makine öğrenmesinde modelin bireysel tahminlerini açıklamak için kullanılan bir yöntemdir. SHAP, teorik olarak optimal Shapley değerlerine dayanmaktadır. Shapley değerinin teknik tanımı, "bir değişken değerinin tüm olası koalisyonlar üzerindeki ortalama marjinal katkısı" dır. Başka bir deyişle, Shapley değerleri, tüm olası değişken kombinasyonlarını kullanarak bir gözlem (örnek) için tüm olası tahminleri dikkate alır. Bu nedenle SHAP, küresel ve yerel tutarlılık ve yorumlanabilirlik sağlayan birleşik bir yaklaşımdır. Bu kapsamda SHAP'ın amacının, her değişkenin tahmine katkısını hesaplayarak herhangi bir gözlemin tahminini açıklamak olduğu şeklinde ifade edilebilir (112).

Sonuç olarak; SHAP metodolojisi, makine öğrenimi modellerinin ve tahminlerinin yorumlanmasını sağlar ve herhangi bir makine öğrenimi modelinden gelen bireysel tahminler için değişken önemliliği değerlerini verir. Bu nedenle, SHAP analizi, birçok makine öğrenimi yaklaşımının kara kutu doğasına ışık tutmaktadır. Tahminlere değişken katkılarının büyüklüğünü ve yönünü gösteren sayısal değerler belirlendikten sonra, değişken katkılarının görselleştirmelerini sağlamaktadır (113).

3.7. Veri analizi

Yapılacak biyoinformatik ve biyoistatistik analizler aşağıdadır.

3.7.1. Biyoinformatik analizler

lncRNA ekspresyon profillerinin incelendiği nefrotoksisite oluşturulan sıçanlar ile kontrol grubu sıçanlarının örnekleri için, R programlama dili kullanılarak diferansiyel ekspresyon analizleri yapılmıştır. Diferansiyel ifade analizi ile belirli bir gen için okuma sayılarında gözlenmiş olan bir farkın önemli olup olmadığının, yani sadece doğal rastgele varyasyon sebebiyle beklenenden değerden daha büyük olup olmadığına karar verebilmek için kullanılmakta olan istatistiksel analizdir. Biyoinformatik analizler R programlama diliyle oluşturulmuş bir kod dizini ile yapılmıştır. İstatistiksel analiz sonucu önem sırasına göre verilen bir gen tablosu ve genlere ilişkin bir grafik ile sunulmaktadır. Ayrıca ayarlanmış P ve 2 tabanında logaritmik kat değişiklikleri ($\log_2FC$) sonuç tablosunda yer almaktadır. Yukarı regüle genleri $\log_2FC > 1$ ile tanımlanırken, aşağı regüle genleri $\log_2FC < -1$ ile tanımlandı. Ekspresyona uğrayan genler ise volkan grafiği ile verilmiştir (114). Grafikte lncRNA lardan yukarı ekspresyon gösterenler kırmızı, aşağı ekspresyon gösteren lncRNA lar ise mavi ile gösterilmektedir. Ayrıca iki grup arasındaki ifadelerinde değişiklik olmayan lncRNA lar siyah renk ile verilmiştir.

3.7.2. Biyoistatistiksel analizler ve Modelleme

Değişkenlerin normal dağılım gösterip göstermediği Shapiro Wilk testi ile incelendi. Veriler ortanca (minimum-maksimum) veya ortalama $\pm$ standart sapma olarak sunuldu. Normal dağılmayan verileri karşılaştırmak için Mann-Whitney U testi ve normal dağılım gösteren verileri karşılaştırmak için bağımsız örneklem t testi kullanıldı. Her bir genin olasılık oranını (etki büyüklüğünün bir ölçüsü) tahmin etmek için lojistik regresyon analizi yapıldı. Lojistik regresyon için Hosmer ve Lemeshow'un uyum iyiliği testi ve

omnibus model katsayıları testi hesaplanmıştır. P değeri <0.05 istatistiksel olarak anlamlı kabul edildi. Analizde IBM SPSS Statistics 26.0 programı kullanılmıştır.

Çalışma kapsamında kullanılması planlanan ağaç tabanlı modellerinin uygulanmasında ve sonrasında açıklanabilir yapay zeka modellemesi için Python programlama dili kullanılacaktır (115). Model geçerliliği için 10 katlı cross validaton yöntemi kullanıldı. n-katlı çapraz doğrulama yönteminde veri önce n parçaya bölünür ve kullanılan model n parçaya uygulanır. n parçadan biri test için kullanılırken, diğer n-1 parça modeli eğitmek için kullanılır. Elde edilen değerlerin ortalaması, çapraz doğrulama yöntemi için değerlendirilir.

3.8. Değerlendirme aşaması

Bu projede, doğruluk, duyarlılık, seçicilik, kesinlik, pozitif tahmin değeri, negatif tahmin değeri, eğri altında kalan alan (AUC) ve F1-skor, Matthews Korelasyon Katsayısı (MCC), pozitif olabilirlik oranı, negatif olabilirlik oranı gibi metrikler %95 güven aralığı ile analizlerin performans değerlendirmesinde kullanılacaktır.

3.9. Ürün aşaması

Bu tez kapsamında biyoinformatik analizleri gerçekleştirmek için R programlama dilinde yazılan ardışık kod dizilimi ile diferansiyel ekspresyon analizleri yapılmıştır (116). Oluşturulan kod dizininde limma, ggplot2, ggrepel paketleri kullanılmıştır. Limma paketi gen ekspresyonu mikrodizi verilerinin analizi için kullanılan bir Bioconductor paketidir. ggplot2, ggrepel paketleri ise veri görselleştirilmesinde kullanılan R paketleridir. Biyoinformatik analizlerden sonra modelleme aşamasında Python programlama dilindeki paketler kullanılarak LIME ve SHAP açıklanabilir yapay zeka yöntemleri sayesinde sonuçları bireysel ve küresel olarak görsel bir şekilde araştırmacılara sunulmuştur. Python programlama dili ile web tabanlı yazılım yapmaya olanak sağlayan Streamlit kütüphanesi kullanılarak 2 adet yazılım yapılmıştır (117). Yazılımlardan ilki heatmap, volkan grafiği olmak üzere 2 adet grafikten oluşan ve genomik verileri görselleştirebilen “Grafiksel Genomik Yazılımı”dır. Diğer yazılım ise “Boyut İndirgeme Analiz Yazılımı” adı ile değişken sayısı oldukça çok olan genomik verilerde boyut indirgemek ve değişken sayısını azaltmak amacıyla yapılmıştır.

3.9.1. Geliştirilen Web Tabanlı Yazılımlar

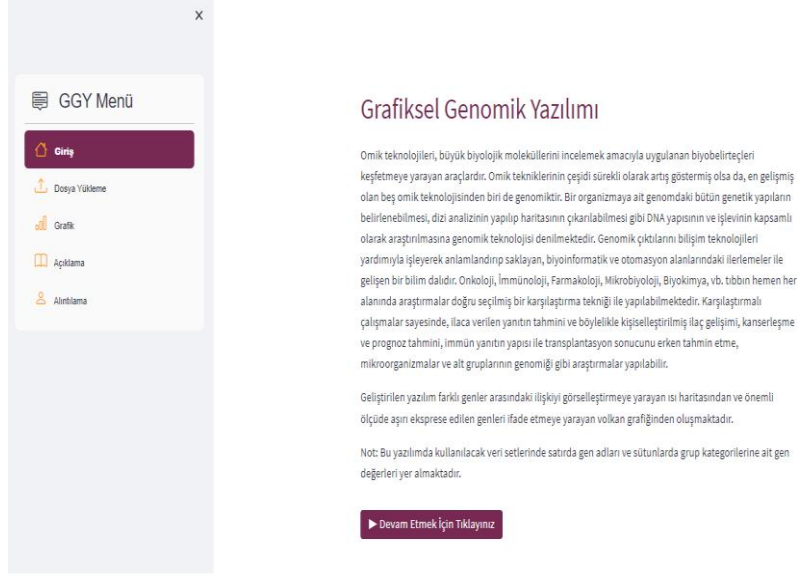
Geliştirilen web tabanlı yazılımlardan ilki yapılan genomik analizleri görselleştirmeye yarayan 2 çeşit grafikten oluşmaktadır. Bu grafiklerden heatmap grafiği gen ifadesi verilerini görselleştirmenin yaygın bir yöntemi olup verileri bir ısı haritası olarak görüntülemektir. Isı haritaları, her ekseninde çizilen iki değişken arasındaki ilişkileri göstermek için kullanılır. Isı haritalarında veriler, her satırın bir geni ve her sütunun örnekleri temsil ettiği bir ızgarada görüntülenir. Kutuların rengi ve yoğunluğu, gen ifadesindeki değişiklikleri temsil etmek için kullanılır. Bir diğer grafik olan volkan grafiği ise değişimin büyüklüğüne (kat değişimi) karşı istatistiksel önemi (p değeri) gösteren bir dağılım grafiği türüdür. İstatistiksel olarak da önemli olan büyük kat değişiklikleri ile genlerin hızlı görsel olarak tanımlanmasını sağlar. Bunlar biyolojik olarak en önemli genler olabilir. Bir volkan grafiğinde, en yukarı regüle edilen genler sağa doğru, en aşağı regüle edilen genler sola doğru ve istatistiksel olarak en önemli genler yukarıya doğru gösterilmektedir.

Boyut indirgemek için yapılan yazılımda ise temel bileşenler analizi temelli boyut indirgeme yapılmakta olup elde edilen grafik ile yüksek verimli dizileme deneylerinin (örn. RNA-seq, GWAS) genellikle yüksek boyutlu veri setlerinin (birkaç yüz ila binlerce örnek) boyutlarının azaltılması amaçlanmıştır. Örneğin, RNA-seq deneylerinde PCA, yüksek boyutlu bir RNA-seq veri kümelerindeki gen ekspresyon modellerini ve biyolojik varyasyonu anlamaya yardımcı olur.

Geliştirilen ilk yazılım “Giriş”, “Dosya Yükleme”, “Grafik”, “Açıklama” ve “Alıntılama” olmak üzere 5 ana bölümden oluşmaktadır. Diğer yazılım da benzer sekmelerden oluşmakta olup bu benzerlikten dolayı sadece ilk yazılıma ait sekmeler tanıtılacaktır.

3.9.2. “Giriş” Sekmesi

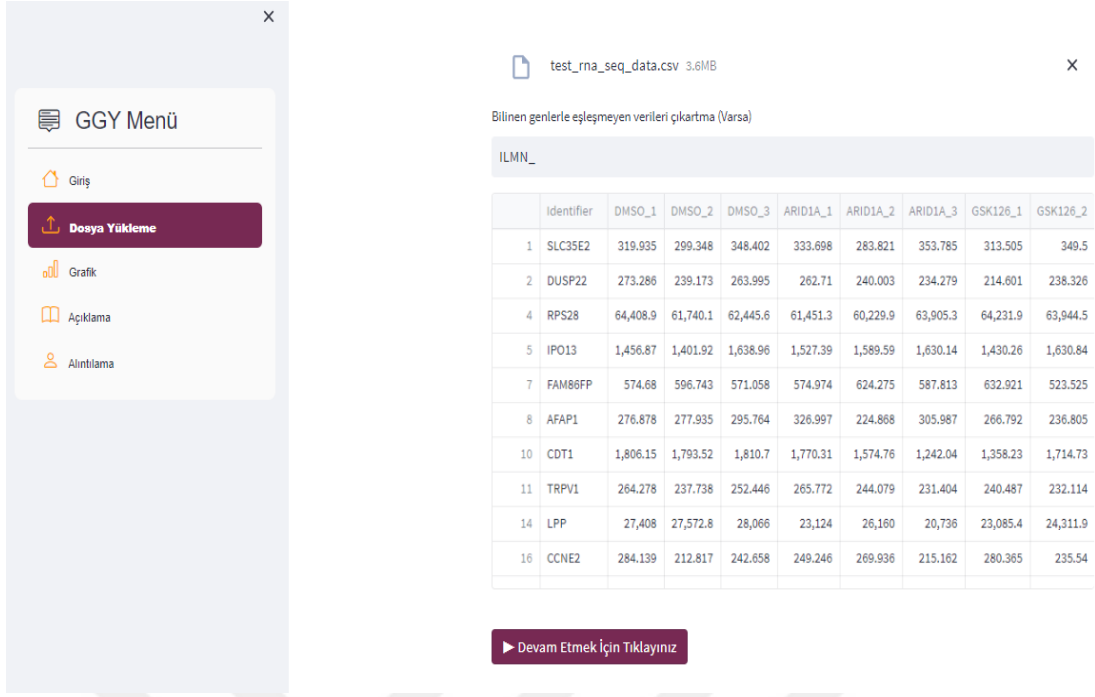
Bu sekmede yazılım ile ilgili genel bilgilerin yer aldığı ve geliştirilen yazılımın anlatıldığı bir bölüm ve yazılımda kullanılacak veri setini anlatan bilgilerin yer aldığı bir bilgilendirme bölümü yer almaktadır. Giriş sekmesine ait görüntü Şekil 3.6’da verilmiştir.



Şekil 3.6. “Giriş” Sekmesi

3.9.3. “Dosya Yükleme” Sekmesi

“Dosya Yükleme” sekmesinde yazılıma uygun veri dosyası yüklenir. Bu bölümde yazılım hakkında bilgi vermek amacıyla örnek bir veri seti ile kullanıcılar karşılanmaktadır. Böylece kullanıcılar, yazılımın hangi koşullar altında kullanılabileceği hakkında fikir sahibi olurlar. Bu aşamadan sonra kullanıcılar kendi veri dosyalarını yazılıma yükleyerek analize devam edebilirler. Bu sekmeyle ait görüntü Şekil 3.7’de verilmiştir.



Şekil 3.7. “Dosya Yükleme” Sekmesi

3.9.4. “Grafik” Sekmesi

Grafik aşamasında kullanıcılar çizmek istedikleri grafik türünü ve grupları seçerek çalışmaların uygun grafikleri çizdirebilirler. Ve sonuçlara ulaşırlar. Bu sekmeye ait görüntü Şekil 3.8’de verilmiştir.



Şekil 3.8. “Grafik” Sekmesi

3.9.5. “Açıklama” ve “Alıntılama” Sekmesi

“Açıklama” sekmesinde geliştirilen yazılım kapsamında yer verilen grafiklere ilişkin açıklamaların yer verildiği bölüm yer almaktadır. “Alıntılama” sekmesinde ise yazılımla ilgili alıntılamanın nasıl yapılacağına ait bilgi bulunmaktadır.



4. BULGULAR

4.1. Biyoinformatik Analiz Sonuçları

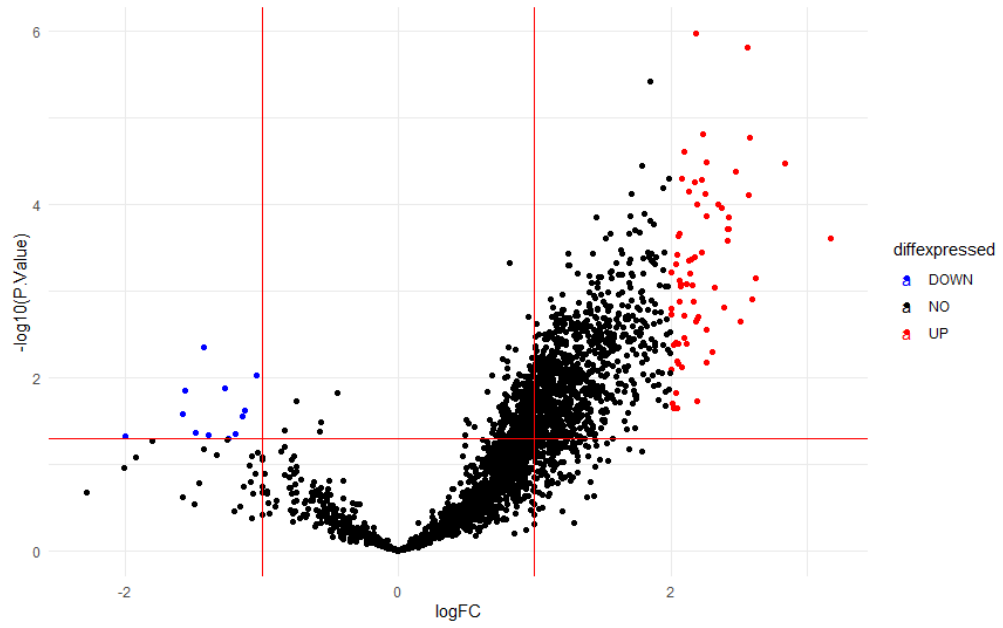
Yapılan genomik analizler sonucunda böbrektoksitesisi olan ve olmayan sıçanlardan 16,386 adet lncRNA incelenmiş ve bunların ekspresyon seviyeleri verilmiştir. Bu lncRNA ların 35 tanesi yukarı regülasyon ($\log_2FC > 1$) göstermiş iken 17 tanesi aşağı regülasyon ($\log_2FC < -1$) göstermiştir. Bu farklı şekilde ifade edilen 52 lncRNA nın hastalık durumu ile ilişkili olabilecek aday lncRNA olabileceği söylenir. Biyoinformatik analizine göre minimum ayarlanmış-p değerleri ile ilişkili ilk on sonuç Tablo 4'de özetlenmiştir.

Tablo 4.1. Biyoinformatik Analiz Sonuçları

ID	Gen Adı	Log2F C	t	P değeri	Adj P Value	B	diffexpres sed
rna- XR_005503408. 1	LOC12010 2212	2.1877 87	6.9831 22	1.05E- 06	0.00205 8	5.6674 49	UP
rna- XR_351582.4	LOC10255 5118	2.5654 95	6.7939 07	1.55E- 06	0.00205 8	5.3162 69	UP
rna- XR_001840278. 2	LOC10835 2129	2.5831 14	5.7615 63	1.69E- 05	0.00898	3.1254 3	UP
rna- XR_001839007. 2	LOC10835 1528	2.1015 16	5.4960 13	2.47E- 05	0.01040 5	2.7620 08	UP
rna- XR_590665.2	LOC10369 1349	2.2599 17	5.3659 34	3.29E- 05	0.01040 5	2.4933 18	UP
rna- XR_005497370. 1	LOC12009 8816	2.8339 66	5.3538 55	3.38E- 05	0.01040 5	2.4682 67	UP
rna- XR_005499594. 1	LOC10369 1306	2.4797 34	5.7867 8	4.21E- 05	0.01046	2.3087 85	UP
rna- XR_591534.3	LOC10369 1816	2.0790 6	5.1794 67	5E-05	0.01046	2.1047 76	UP
rna- XR_005496784. 1	LOC12009 8521	2.2246 8	5.1647 58	5.16E- 05	0.01046	2.0739 67	UP
rna- XR_005497310. 1	LOC12009 8801	2.1765 17	5.1356 78	5.51E- 05	0.01046	2.0129 9	UP

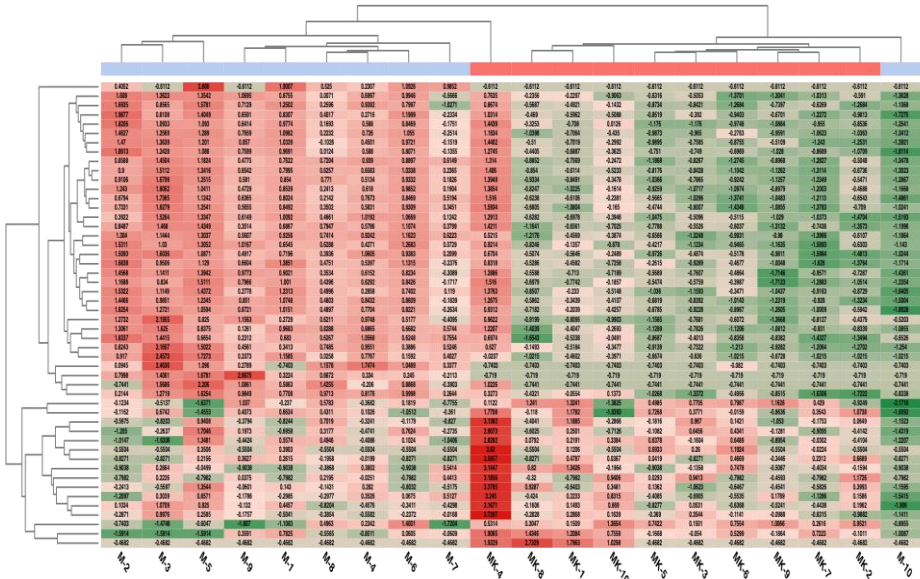
Tablo 4.1’de gösterilen lncRNA ifadelerine dikkat edildiğinde hepsinin Log2FC değerlerinin birden büyük olduğu gözlemlenmektedir. Bu nedenle lncRNA ifadelerinin Log2FC değerleri hepsi yukarı ekspresyon göstermiştir. Bu yukarı ekspresyon gösteren lncRNA lar 2.833965561, 2.58311362, 2.565494997, 2.479734103, 2.259916722, 2.224679868, 2.187786899, 2.176516723, 2.101515792, 2.079059666 Log2FC değerleri ile rna-XR_005497370.1 (LOC120098816), rna-XR_001840278.2 (LOC108352129), rna-XR_351582.4 (LOC102555118), rna-XR_005499594.1 (LOC103691306), rna-XR_590665.2 (LOC103691349), rna-XR_005496784.1 (LOC103691349), rna-XR_005503408.1 (LOC120102212), rna-XR_005497310.1 (LOC120098801), rna-XR_001839007.2 (LOC108351528), rna-XR_591534.3 (LOC103691816) id numaralı lncRNA lar olarak elde edilmiştir.

Diferansiyel olarak ifade edilmiş olan genlerin görselleştirilmesinde kullanılan Volkan grafiği (yanardağ grafiği) Şekil 4’de gösterilmektedir. Bir volkan grafiği, değişimin büyüklüğüne (kat değişimi) karşı istatistiksel önemi (P değeri) gösteren bir dağılım grafiği türüdür. Bir volkan grafiğinde, p değerinin negatif logaritmasının y eksenine (genellikle 10 tabanı) çizilmesiyle oluşturulur. X eksen, iki koşul arasındaki kat değişiminin logaritmasıdır. İstatistiksel olarak da önemli olan büyük kat değişiklikleri ile genlerin hızlı görsel olarak tanımlanmasını sağlar. Bunlar biyolojik olarak en önemli genler olabilir. Bir volkan grafiğinde, en yukarı regüle edilen genler sağa doğru, en aşağı regüle edilen genler sola doğru ve istatistiksel olarak en önemli genler yukarıya doğru gösterilmektedir.



Şekil 4.1. Nefrotoksisitesinde olan sıçanlar ile kontrol grubu sıçanların arasında tanımlanan farklı şekilde ekspresyon gösteren lncRNA' lar için volkan grafiği

İfade seviyeleri karşılaştırıldığında en fazla değişiklik gösteren lncRNA ifadelerinden 50 tanesi için heatmap (ısı haritası) gösterimi şekil 2 gibi tespit edilmiştir. Isı haritaları, her ekseninde çizilen iki değişken arasındaki ilişkileri göstermek için kullanılır. Hücre renklerinin her ekseninde nasıl değiştiğini gözlemleyerek, bir veya her iki değişken için değerlerde herhangi bir model olup olmadığını gözlemleyebilme imkanı sunar.



Şekil 4.2. En Fazla Değişiklik Gösteren lncRNA ifadelerinden 50 tanesi ısı haritası

M vs MK kıyaslamasında en fazla değişiklik gösteren 50 lncRNA için fazla ekspres olanlar kırmızı ve baskılanan ifade seviyesi için ise yeşil renkli gösterimde uygulama numunelerinin kontrole göre farklı ifade profili sergilediği görülmektedir. Ancak bazı numunelerin (M-10, MK-4 ve MK-2 gibi) uygulama ve kontrol gruplarından sapma gösterdiği belirlenmiştir.

4.2. Biyoistatistik Analiz ve Modelleme Sonuçları

Çalışmada hastalık durumu ile ilişkili olabilecek lncRNA ların ortaya konulabilmesi adına farklı regülasyonlar gösteren (aşağı-yukarı) lncRNA lardan değişken seçimi yöntemlerinden Boruta değişken seçimi yöntemi kullanılarak hastalık durumu ile ilişkili olabilecek 31 lncRNA seçilmiştir. Tablo 4 seçilen ifadeler ile veri setinin açıklamaları, incelenen hedef değişken için seçilen ifadelerle ait tanımlayıcıları, istatistik anlamlılığı ve hedef değişken için gen başına olasılık oranını (OR) içermektedir.

Tablo 4.2. İstatistik Analiz Sonuçları

Gen Adı	ID	M		MK		OR	p
		Ortalama±Standart Sapma	Medyan (Min-Maks)	Ortalama±Standart Sapma	Medyan (Min-Maks)		
LOC102555118	rna-XR_351582.4	226.4±116.41	248(42-447)	35.6±16.85	34(17-66)	1.077	0.001*
LOC106736471	rna-NR_133655.1	102.4±76.73	88(5-257)	11.9±9.48	10.5(1-34)	1.098	0.005*
LOC103691349	rna-XR_590665.2	281.2±123.78	294(49-470)	68.4±78.82	46(26-290)	1.017	0.001**
LOC108351528	rna-XR_001839007.2	454.2±191.95	486.5(96-661)	117.1±118.09	80.5(55-449)	1.011	0.001**
LOC120098801	rna-XR_005497310.1	166±92.76	164.5(29-370)	38.6±36.34	26.5(13-139)	1.032	0.001**
LOC120094778	rna-XR_005489439.1	140±90.62	125(28-296)	28.9±14.03	30(9-51)	1.099	0.004*
LOC120099280	rna-XR_005498350.1	109.6±68.4	96(13-206)	21.6±21.84	15.5(7-82)	1.053	0.002**
LOC120096007	rna-XR_005492056.1	134.4±91.26	123.5(17-332)	32.6±28.96	25.5(6-111)	1.040	0.004**
LOC120098788	rna-XR_005497230.1	27.3±13.61	29.5(3-52)	4.6±4.62	2.5(0-15)	1.291	<0.001**
LOC120098190	rna-XR_005496257.1	85.5±54.7	70(9-172)	19.5±18.58	16(4-70)	1.064	0.004**
LOC108348888	rna-XR_005496888.1	71.2±32.64	75.5(12-112)	17.1±20.59	11.5(3-74)	1.060	0.002**
LOC103691816	rna-XR_591534.3	210.4±116.14	230.5(54-421)	49.2±36.54	40.5(19-147)	1.030	0.001**
LOC120098816	rna-XR_005497370.1	220.6±173.89	186(48-552)	31.3±22.35	30(6-73)	1.080	0.007*
LOC120096731	rna-XR_005493563.1	6.6±6.64	3.5(0-18)	13.2±14.34	8(3-51)	-	0.093**
LOC120098521	rna-XR_005496784.1	362.1±181.28	349.5(74-587)	88.8±100.42	58(33-369)	1.012	0.001**
LOC120102202	rna-XR_005503371.1	84.1±63.57	73(13-208)	15.6±11.47	13.5(3-37)	1.115	0.008*
LOC102549457	rna-XR_358189.4	77.7±42.9	75.5(8-154)	21.1±26.93	12.5(4-96)	1.048	0.007**
LOC120102261	rna-XR_005503535.1	215.2±138.84	176.5(16-442)	47±27.2	38.5(26-116)	1.035	0.003**
LOC120100781	rna-XR_005500805.1	51.1±23.38	49(11-82)	14.4±20.5	8(2-71)	1.071	0.002**

LOC1083 48808	rna- XR_00549 6283.1	42.2±26.1	37.5(5-84)	9.1±5.61	9(2-19)	1.1 76	0.003 *
LOC1036 91306	rna- XR_00549 9594.1	6.2±4.47	5.5(0-12)	0.6±0.52	1(0-1)	6.6 82	0.001 **
LOC1025 52040	rna- XR_00183 9839.2	3.8±4.49	3(0-15)	0.1±0.32	0(0-1)	9.7 92	0.002 **
LOC1200 99889	rna- XR_00549 9330.1	282.2±232.78	197(41-831)	68.2±86.7	35(24-308)	1.0 14	0.002 **
LOC1200 99800	rna- XR_00549 9033.1	53.7±33.95	45(5-102)	14.2±19.85	9.5(1-69)	1.0 60	0.004 **
LOC1200 97836	rna- XR_00549 5645.1	32.8±16.73	28.5(13-62)	7.7±4.32	8.5(1-14)	2.1 56	0.001 *
LOC1201 02212	rna- XR_00550 3408.1	18.5±10.54	14.5(8-42)	4±2.31	3.5(2-10)	2.3 18	<0.00 1**
LOC1025 55751	rna- XR_00549 7840.1	54.9±45.9	41(1-162)	12.1±12.54	8.5(3-47)	1.0 90	0.008 **
LOC1201 02327	rna- XR_00550 3688.1	50.7±46.08	41.5(1-165)	9.8±8.04	7.5(3-30)	1.1 24	0.005 **
LOC1200 99962	rna- XR_00549 9541.1	1±0.94	1(0-3)	2.3±0.82	2(1-4)	1.1 62	0.005 **
LOC1083 52129	rna- XR_00184 0278.2	26±18.34	21(0-59)	5.8±6.94	3(2-25)	1.1 63	0.008 **
LOC1025 54372	rna- XR_35343 8.4	48.4±27.61	49.5(3-84)	12.1±6.05	11.5(4-21)	1.1 32	0.002 *

*: Bağımsız örneklerde t testi; **: Mann Whitney U testi; OR: Odds ratio; SD: Standard sapma

Tablo 4.2'deki istatistik analiz sonuçlarına göre rna-XR_005493563.1 (LOC120096731) lncRNA dışında tüm lncRNA ifadeleri için nefrotoksitesinde olan sıçan grubu ile kontrol grubu arasında istatistiksel olarak önemli farklılıklar tespit edilmiştir ($p<0.05$). Tüm ifadeler için yapılan lojistik regresyon modelleri anlamlı olup OR oranları belirlenmiştir.

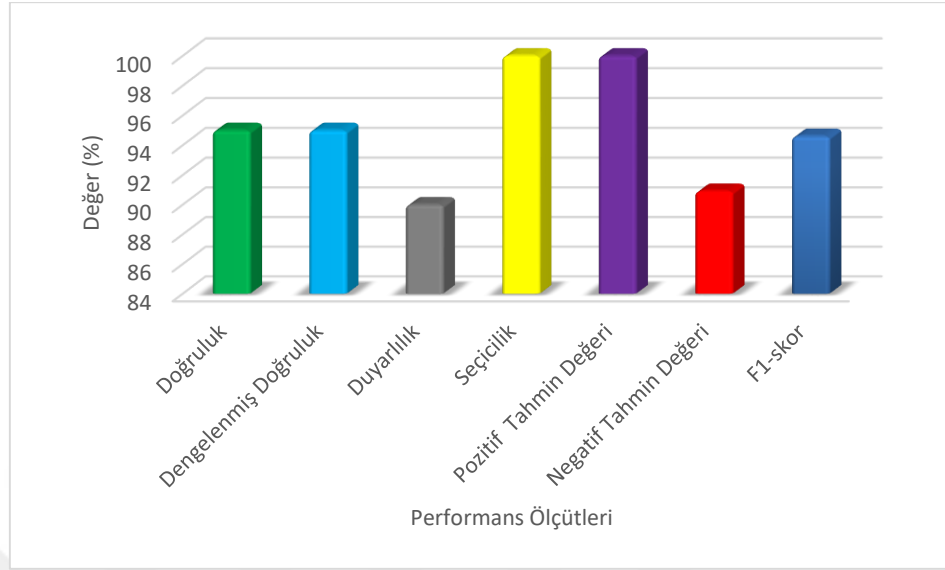
Çalışmada kullanılan makine öğrenimi modellerinin performans ölçütlerine ilişkin sonuçlar Tablo 4.3'de verilmiştir.

Tablo 4.3. Çalışmada kullanılan makine öğrenimi modellerinin performans ölçütleri sonuçları (%95 güven aralığı (GA))

Değer (%) (95% GA) / Model	XGBoost	RF	Karar Ağacı	AdaBoost	Bagged CART
Doğruluk	95 (85.4-100)	83.3 (53.5-100)	85 (69.4-100)	85 (69.4-100)	83.3 (53.5-100)
Dengelenmiş Doğruluk	95 (85.4-100)	83.3 (53.5-100)	85 (69.4-100)	85 (69.4-100)	83.3 (53.5-100)
Duyarlılık	90 (55.5-99.75)	100 (29.2-100)	80 (44.4-97.5)	90 (55.5-99.75)	83.3 (53.5-100)
Seçicilik	100 (69.15-100)	66.7 (9.4-99.2)	90 (55.5-99.7)	80 (44.4-97.5)	66.7 (9.4-99.2)
Pozitif Tahmin Değeri	100 (66.4-100)	75 (19.4-99.4)	88.9 (51.8-99.7)	81.8 (48.2-97.7)	75 (19.4-99.4)
Negatif Tahmin Değeri	90.91 (58.7-99.8)	100 (15.8-100)	81.8 (48.2-97.7)	88.9 (51.8-99.7)	100 (15.8-100)
F1-skor	94.6 (85-100)	85.7 (48-100)	84.2 (68.2-100)	85.7 (70.4-100)	85.7 (57.7-100)

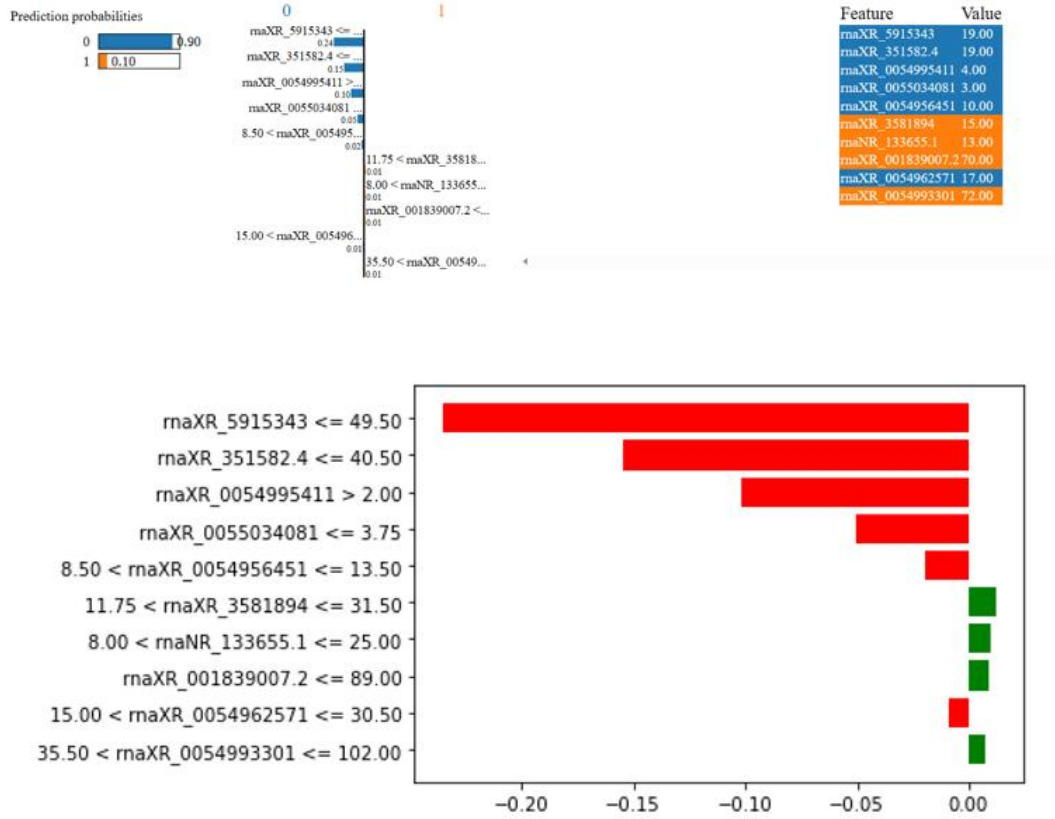
RF, Karar Ağacı, AdaBoost, Bagged CART ve XGBoost modellerinin sınıflandırma performansına göre RF modelinden doğruluk %83.3, dengelenmiş doğruluk %83.3, duyarlılık %100, seçicilik %66.7, pozitif tahmin değeri %75, negatif tahmin değeri %100 ve F1-skor %85.7 olarak elde edilmiştir. Karar Ağacı modelinden doğruluk %85, dengelenmiş doğruluk %85, duyarlılık %80, seçicilik %90, pozitif tahmin değeri %88.9, negatif tahmin değeri %81.8 ve F1-skor %84.2 olarak elde edilmiştir. AdaBoost modelinden doğruluk %85, dengelenmiş doğruluk %85, duyarlılık %90, seçicilik %80, pozitif tahmin değeri %81.8, negatif tahmin değeri %88.9 ve F1-skor %85.7 olarak hesaplanmıştır. Bagged CART modelinden doğruluk %83.3, dengelenmiş doğruluk %83.3, duyarlılık %66.7, seçicilik %75, pozitif tahmin değeri %100, negatif tahmin değeri %100 ve F1-skor %85.7 olarak hesaplanmıştır. XGBoost modelinden doğruluk %95, dengelenmiş doğruluk %95, duyarlılık %90, seçicilik %100, pozitif tahmin değeri %100, negatif tahmin değeri %90.91 ve F1-skor %94.6 olarak hesaplanmıştır.

Kullanılan ağaç tabanlı modellerden en iyi sonucu veren XGboost modelinin performans metrikleri Şekil 4.3. 'de verilmektedir.



Şekil 4.3. XGBoost modeline ilişkin performans metriklerinin grafiği

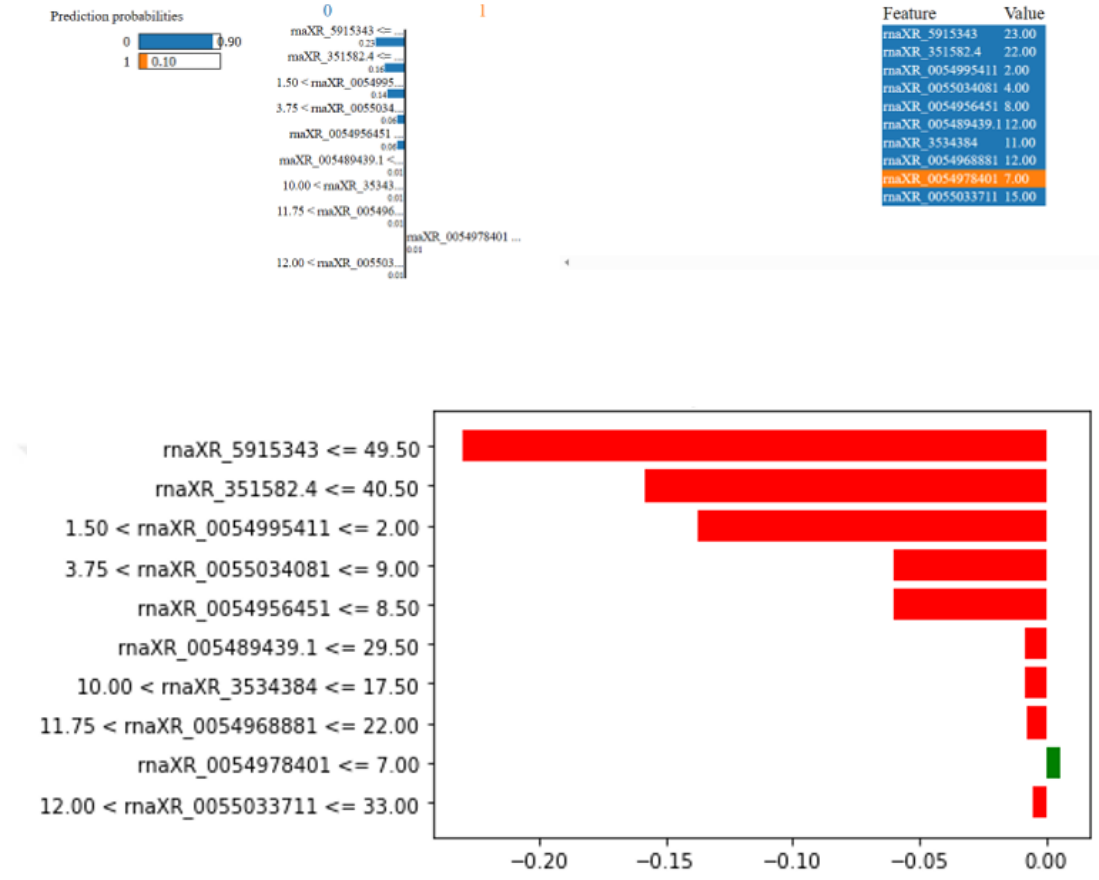
Kullanılan ağaç tabanlı modellerin arasında en iyi sonucu veren XGBoost modeline açıklanabilir yapay zeka yöntemlerinden yerel yöntemlerden biri olan LIME ve küresel yöntemlerden biri olan SHAP uygulanmıştır. Şekil 4.4, Şekil 4.5 ve Şekil 4.6'da yerel açıklama örnekleri için LIME yöntemi sonucunda elde edilen sonuçlardan ilk üçü aşağıda verilmiştir. Şekil 4.4, Şekil 4.5 ve Şekil 4.6'da, yeşil çubuklar hedef değişken ile pozitif korelasyona sahip özellikleri gösterirken, kırmızı çubuklar hedef değişken ile negatif korelasyona sahip özellikleri göstermektedir.



Şekil 4.4. Nefrotoksisite grubundaki ilk sıçan için LIME sonuçları

Buna göre; Şekil 4.4'deki sıçanın %90 olasılık ile Nefrotoksisite olmadığı tahmin edilmiştir. Bu sıçanın yanlış negatif olarak tahmin edilmesinin sebebi rnaXR_591534.3 (LOC103691816), rnaXR_351582.4 (LOC102555118), rnaXR_005499541.1 (LOC120099962), rnaXR_005503408.1 (LOC120102212), rnaXR_005495645.1 (LOC120097836), rnaXR_358189.4 (LOC102549457), rnaNR_133655.1 (LOC106736471), rnaXR_001839007.2 (LOC108351528), rnaXR_005496257.1 (LOC120098190), rnaXR_005499330.1 (LOC120099889) lncRNA ekspresyon verilerinin değerleridir. Bu sıçanın rnaXR_591534.3 <= 49.50 den düşük değerleri, rnaXR_351582.4 <= 40.50'den düşük değerleri, rnaXR_005499541.1 > 2.00'den yüksek değerleri, rnaXR_005503408.1 <= 3.75' den düşük değerleri, 8.50 < rnaXR_005495645.1 <= 13.50 arasındaki ve 15.00 < rnaXR_005496257.1 <= 30.50 arasındaki değerleri Nefrotoksisite olma olasılığını azaltmıştır. Buna karşılık 11.75 < rnaXR_358189.4 <= 31.50 arasındaki değerleri, 8.00 < rnaNR_133655.1 <= 25.00 arasındaki değerleri, rnaXR_001839007.2 <= 89.00' den düşük değerleri, ve 35.50 <

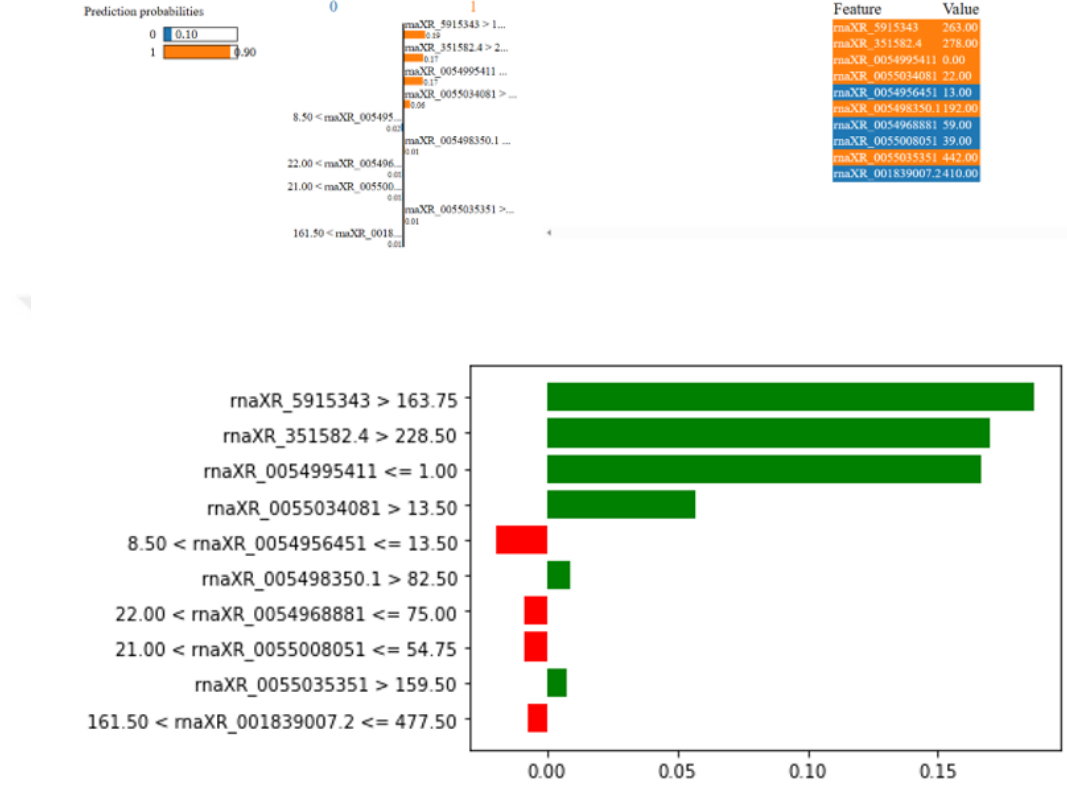
rnaXR_005499330.1 $\leq$ 102.00 arasındaki değerleri Nefrotoksisite olma olasılığını arttırmıştır.



Şekil 4.5. Nefrotoksisite grubundaki ikinci sıçan için LIME sonuçları

Buna göre; Şekil 4.5’deki sıçanın %90 olasılık ile Nefrotoksisite olmadığı tahmin edilmiştir. Sıçanın yanlış negatif olarak tahmin edilmesinin sebebi rnaXR_591534.3 (LOC103691816), rnaXR_351582.4 (LOC102555118), rnaXR_005499541.1 (LOC120099962), rnaXR_005503408.1 (LOC120102212), rnaXR_005495645.1 (LOC120097836), rnaXR_005489439.1 (LOC120094778), rnaXR_353438.4 (LOC102554372), rnaXR_005496888.1 (LOC108348888), rnaXR_005497840.1 (LOC102555751) ve rnaXR_005503371.1 (LOC120102202) lncRNA ekspresyon verilerinin değerleridir. Bu sıçanın rnaXR_591534.3 $\leq$ 49.50’ den düşük değerleri, rnaXR_351582.4 $\leq$ 40.50’ den düşük değerleri, 1.50 < rnaXR_005499541.1 $\leq$ 2.00 arasındaki değerleri, 3.75 < rnaXR_005503408.1 $\leq$ 9.00 arasındaki değerleri, rnaXR_005495645.1 $\leq$ 8.50’ den düşük değerleri, rnaXR_005489439.1 $\leq$ 29.50’ den

düşük değerleri, $10.00 < \text{rnaXR_3534384} \leq 17.50$ arasındaki değerleri, $11.75 < \text{rnaXR_005496888.1} \leq 22.00$ arasındaki değerleri ve $12.00 < \text{rnaXR_005503371.1} \leq 33.00$ arasındaki değerleri Nefrotoksisite olma olasılığını azaltmıştır. Buna karşılık $\text{rnaXR_005497840.1} \leq 7.00$ arasındaki değerleri Nefrotoksisite olma olasılığını arttırmıştır.

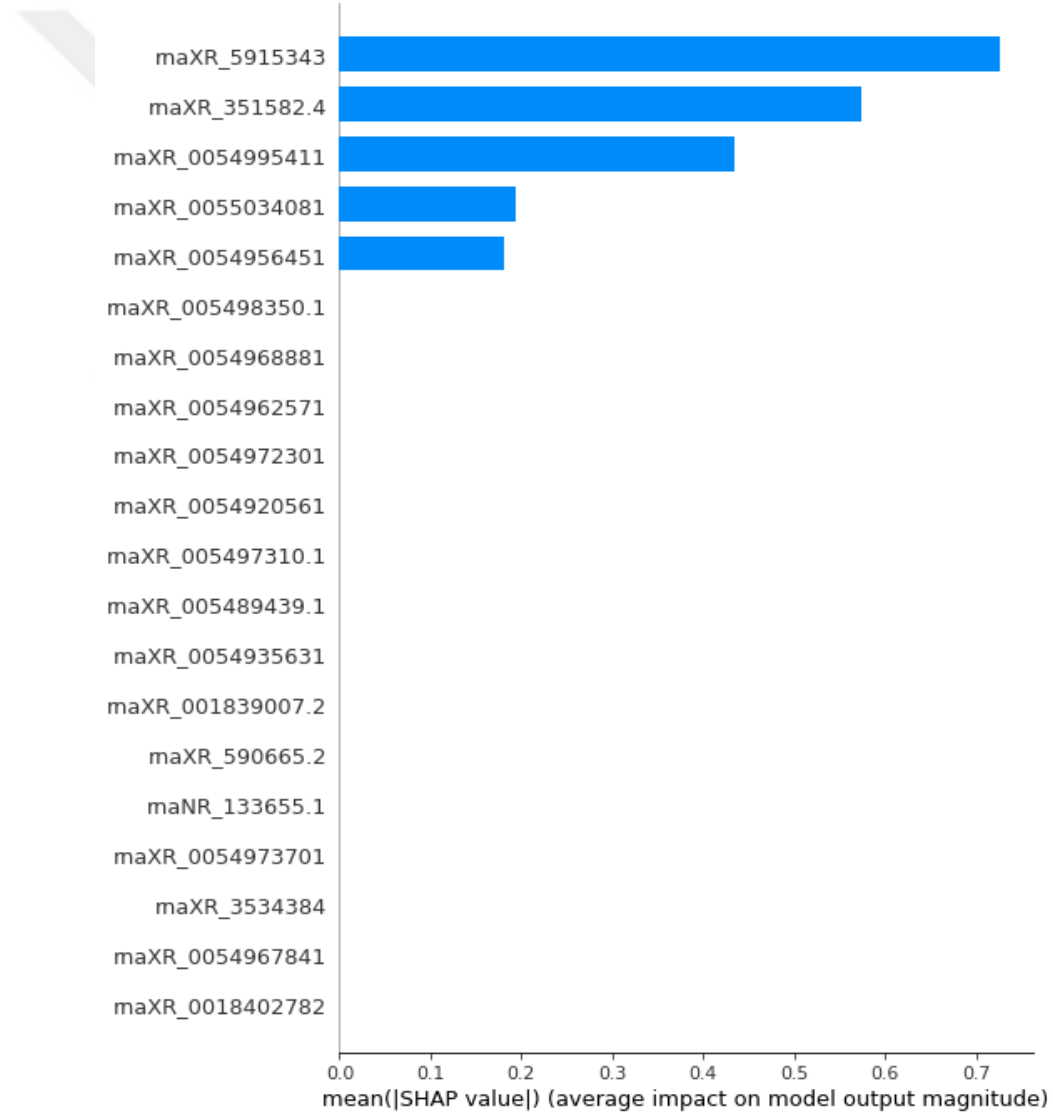


Şekil 4.6. Nefrotoksisite grubundaki üçüncü sıçan için LIME sonuçları

Buna göre; Şekil 4.6'daki sıçanın %90 olasılık ile Nefrotoksisite olduğu tahmin edilmiştir. Bu sıçanın doğru olarak tahmin edilmesinin sebebi rnaXR_591534.3 (LOC103691816), rnaXR_351582.4 (LOC102555118), rnaXR_005499541.1 (LOC120099962), rnaXR_005503408.1 (LOC120102212), rnaXR_005495645.1 (LOC120097836), rnaXR_005498350.1 (LOC120099280), rnaXR_005496888.1 (LOC108348888), rnaXR_005500805.1 (LOC120100781), rnaXR_005503535.1 (LOC120102261) ve rnaXR_001839007.2 (LOC108351528) lncRNA ekspresyon verilerinin değerleridir. Bu sıçanın $8.50 < \text{rnaXR_005495645.1} \leq 13.50$ arasındaki değerleri, $22.00 < \text{rnaXR_005496888.1} \leq 75.00$ arasındaki değerleri, $21.00 < \text{rnaXR_005500805.1} \leq 54.75$ ve $161.50 < \text{rnaXR_001839007.2} \leq 477.50$ arasındaki

değerleri Nefrotoksisite olma olasılığını azaltmıştır. Buna karşılık rnaXR_591534.3 > 163.7500'den yüksek değerleri, rnaXR_351582.4 > 228.50' den yüksek değerleri, rnaXR_005499541.1 <= 1.00' den düşük değerleri, rnaXR_005503408.1 > 13.50'den yüksek değerleri, rnaXR_005498350.1 > 82.50'den yüksek değerleri ve rnaXR_005503535.1 > 159.50' den yüksek değerleri Nefrotoksisite olma olasılığını arttırmıştır.

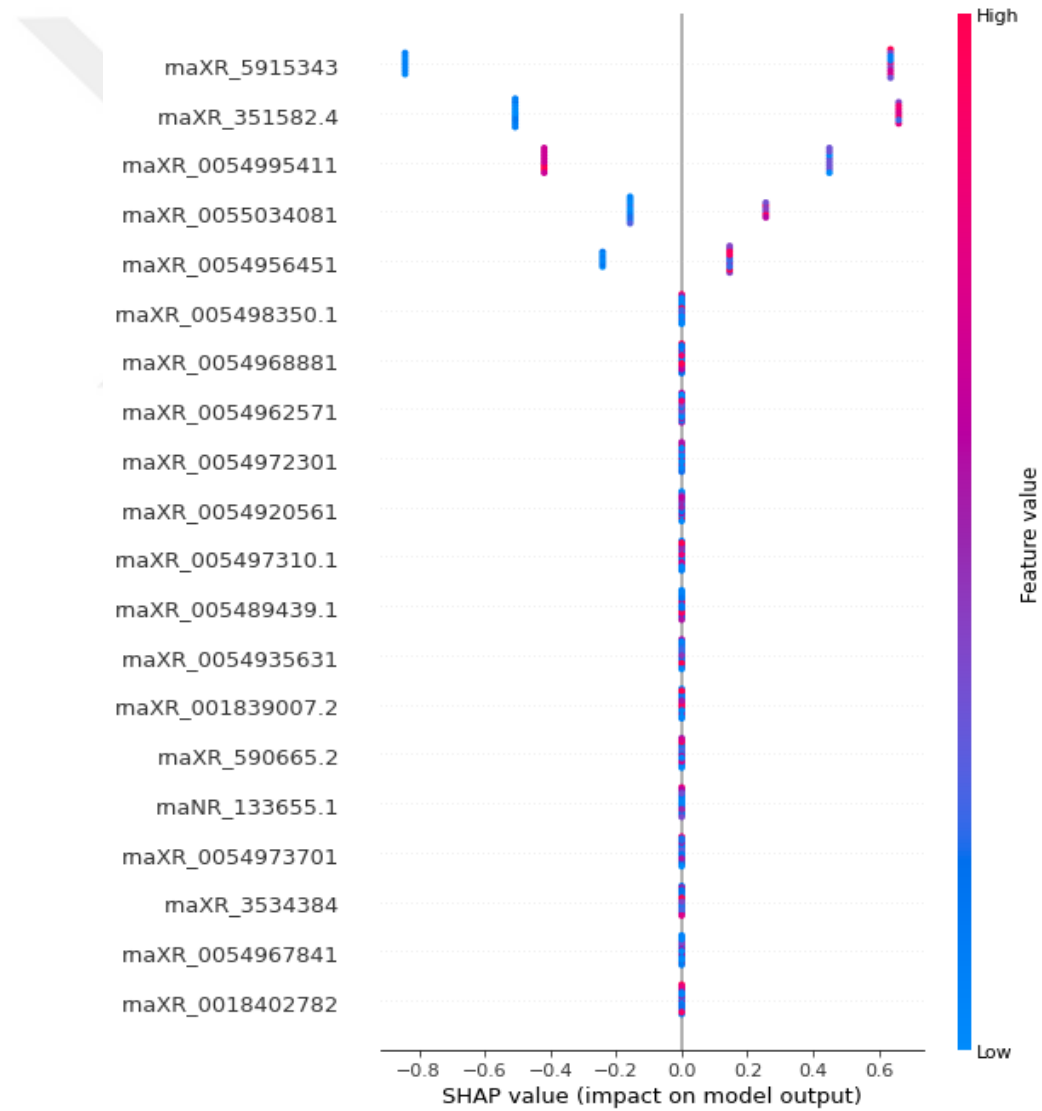
Modeldeki değişkenlerin Nefrotoksisite olma riskini nasıl etkilediğini göstermek amacıyla görsel olarak açıklamak için SHAP yöntemi kullanılmıştır. Şekil 4.7, normalize edilmiş SHAP değeri tarafından değerlendirilen risk faktörlerini ve bunların önemlilik sıralamalarını göstermektedir. Bu değişken önemlilikleri azalan sırada verilmektedir.



Şekil 4.7. Normalize edilmiş SHAP değerlerine göre önemlilik sıralamaları

Yukarıda Şekil 4.7 göz önünde bulundurulduğunda; rna-XR_591534.3 (LOC103691816), rna-XR_351582.4 (LOC102555118), rna-XR_005499541.1 (LOC120099962), rna-XR_005503408.1 (LOC120102212) ve rna-XR_005495645.1 (LOC120097836) lncRNA ekspresyon verilerinin değerlerinin Nefrotoksisite için belirleyici risk faktörlerinden en önemli beş risk faktörü olduğu söylenebilir.

Şekil 4.8, pozitif ve negatif SHAP değerleri dikkate alınarak oluşturulmuştur. Pozitif bir SHAP değeri, hedef değişkene (Nefrotoksisite) yapılan katkının pozitif olduğunu, negatif bir SHAP değeri ise katkının negatif olduğunu göstermektedir. Buna ek olarak grafikteki noktalar maviye yaklaştıkça değişkenin değeri azalır, pembeye yaklaştıkça artar.



Şekil 4.8. SHAP grafiği

Şekil 4.8'e göre daha yüksek rna-XR_591534.3 (LOC103691816), rna-XR_351582.4 (LOC102555118), rna-XR_005503408.1 (LOC120102212), rna-XR_005495645.1 (LOC120097836) daha düşük rna-XR_005499541.1 (LOC120099962) lncRNA ekspresyon verilerinin değerlerinin Nefrotoksisite riskini arttırdığı gözlemlenmiştir.



5. TARTIŞMA

Yaşayan organizmaların ilaçlara ve kimyasallara maruz kalması sıklıkla toksisite ile sonuçlanır. Her bir bileşiğin vücudun tümüne toksik etkisinin olduğu söylene de; birçok bileşiğin toksik etkisinin bazı organlara olan toksik etkisi daha belirgindir, bunlardan biri de böbreklerdir. Böbrekler, zararlı maddelerin etkilerine karşı çok hassastır. Bu durumun sebepleri arasında böbreklerin yüksek perfüzyon hızına sahip olması, tübül lümeninde birçok maddeyi konsantre etmeleri ve metabolizma hızlarının yüksek olması sayılabilir. Birçok materyal böbrekler için nefrotoksiktir. Bu materyaller arasında ilaçlar, ağır metaller, organik kimyasallar, hemoglobin, miyoglobin ve yılan zehiri olabilmekte, nefrotoksisite olguları rapor edilmektedir (20). Türkiye’de ve Dünya’da böbrek hastalığı olan kişi sayısı her geçen gün artmaktadır ve nefrotoksik ilaç kullanımı, böbrek hastalığı için en yüksek risk faktörleri arasındadır (118). Önemli miktarda ilaç ve metabolitleri idrarla atılır. İlaçlar, toplumdan ve hastaneden edinilen akut böbrek yetmezliği olaylarının yaklaşık yüzde 20'sine neden olur. Yaşlı yetişkinler arasında ilaca bağlı nefrotoksisite insidansı yüzde 66'ya kadar çıkabilir. 30 yıl öncesine karşılaştırıldığında, günümüzün hastaları daha yaşlı, diyabet ve kardiyovasküler hastalık insidansı daha yüksek, birden fazla ilaç alıyor ve böbrek fonksiyonlarına zarar verme potansiyeline sahiptir. Bu yüzden böbrekte elde edilen yüksek konsantrasyon göz önüne alındığında, nefrotoksisite ve ilaç etkileşimlerini içeren yan etkiler oldukça yaygındır. İlaçlar akut böbrek hasarının (ABY) yaklaşık %20-40'ına, belki de yaşlı popülasyonda ABH'nin %60'ına kadar yüksek oranda neden olur (119). Nefrotoksisite önemli morbidite ve mortaliteye yol açabilen bir risk olmaya devam etmektedir. Meydana gelen böbrek toksisitelerinin azaltmak için nefrotoksik ilaç kombinasyonlarından kaçınılmalı, mümkün olduğunda nefrotoksik olmayan alternatif ilaçlar kullanılmalı, tedavinin başlangıcından önce temel böbrek fonksiyonu değerlendirilmeli, tedavi sırasında böbrek fonksiyonunun ve hayati belirtilerin izlenmeli yani kısaca etki mekanizmaları açıklanmalıdır. Böylece buna bağlı olarak tedavi yöntemlerinin geliştirilmesi hastaların mortalite riski açısından oldukça önemlidir (120). Çalışmamızda metotreksat ile oluşturulmuş nefrotoksitesisi olan sıçanlar ile kontrol grubundan oluşan sıçanlardan alınan örnekler alınarak genomik, histopatolojik ve biyokimyasal analizler yapılmıştır.

Yapılan çalışmada nefrotoksitesisi olan sıçanlar ile kontrol grubu sıçanların böbrek dokularından elde edilen genomik veriler kullanılmıştır. Genomik analizler sonucunda böbrek toksitesisi olan ve olmayan sıçanlardan 16.386 adet lncRNA

incelenmiş ve bunların ekspresyon seviyeleri verilmiştir. Bu lncRNA ların 35 tanesi yukarı regülasyon göstermiş iken 17 tanesi aşağı regülasyon göstermiştir. Yapılan ilk on sonucun verildiği biyoinformatik analizlerin bulguları göz önünde bulundurulduğunda (Tablo 4.1.) iki grup arasındaki ekspresyon kat değişikliklerini belirlemek için kullanılan Log2FC değerlerine göre, rna-XR_005497370.1 (LOC120098816) id li lncRNA, nefrotoksisitesi olan grupta kontrol grubuna göre 7.13 kat daha yüksek gen ekspresyonuna sahiptir. Benzer şekilde Log2FC değerleri en yüksekten en düşüğe göre rna-XR_001840278.2 (LOC108352129), rna-XR_351582.4 (LOC102555118), rna-XR_005499594.1 (LOC103691306), rna-XR_590665.2 (LOC103691349), rna-XR_005496784.1 (LOC103691349), rna-XR_005503408.1 (LOC120102212), rna-XR_005497310.1 (LOC120098801), rna-XR_001839007.2 (LOC108351528), rna-XR_591534.3 (LOC103691816) id numaralı lncRNA lar sırası ile 5.99 kat, 5.92 kat, 5.58 kat, 4.79 kat, 4.67 kat, 4.56 kat, 4.52 kat, 4.29 kat ve 4.23 kat daha yüksek gen ekspresyonuna sahip olduğu elde edilmiştir.

Çalışmada kullanılan Boruta değişken seçimi yöntemi ile seçilen 31 lncRNA ile hedef (Nefrotoksisite) değişkeni alınarak yapılan ağaç tabanlı makine öğrenimi modellerinden RF, Karar Ağacı, AdaBoost, Bagged CART ve XGBoost modelleri ile yapılan modellemeler sonucunda elde edilen performans ölçüleri şu şekildedir: RF modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skor sırasıyla %83.3, %83.3, %100, %66.7, %75, %100, ve %85.7 olarak elde edilmiştir. Karar Ağacı modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skor sırasıyla %85, %85, %80, %90, %88.9, %81.8, ve %84.2 olarak elde edilmiştir. AdaBoost modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skor sırasıyla %85, %85, %90, %80, %81.8, %88.9, ve %85.7 olarak elde edilmiştir. Bagged CART modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skor sırasıyla %83.3, %83.3, %83.3, %66.7, %75, %100, ve %85.7 olarak elde edilmiştir. XGBoost modelinden elde edilen doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri, negatif tahmin değeri ve F1-skor sırasıyla %95, %95, %90, %100, %100, %90.91, ve %94.6 olarak elde edilmiştir. Performans ölçütlerinin değerleri göz önünde bulundurulduğunda XGBoost yöntemi en başarılı makine öğrenimi yöntemi olmuştur. Buna göre XGBoost yönteminin

yüksek performans metrikleri sonucunda iki grubu doğru bir şekilde sınıflandırabileceğini göstermektedir. Yüksek performans ölçütlerinin değerleri göz önünde bulundurulduğunda XGBoost yönteminin iki grubu doğru bir şekilde sınıflandırabileceğini göstermektedir.

Son on yılda, büyük veri kümelerinin ve daha yüksek bilgi işlem gücünün kullanılabilirliğiyle, makine öğrenimi yöntemleri çok çeşitli durumlarda yüksek performans elde etmesine rağmen kullanılan makine öğrenimi modelinin tahminlerini ve sonuçlarını neden veya nasıl elde edebildiğine ilişkin doğrudan açıklamalar sağlamak zordur. Bu problemler ışığında açıklanabilir yapay zeka son zamanlarda daha fazla dikkat çekmeye başlamıştır. Kısaca açıklanabilir yapay zeka; yapay zeka uygulamalarını kullanıcılar tarafından anlaşılır hale getirmeyi amaçlayan yöntem veya tekniklerin bütünüdür. Açıklanabilir yapay zekada amaç, genel olarak kavranması zor bir süreçte sahip olan yapay zekanın kararlarının ardındaki hesaplamalı çıkarımları genel kullanıcılar ve araştırmacılar tarafından anlaşılır kılmaktır. Bu nedenle açıklanabilir yapay zeka yöntemleri geliştirilmiş ve farklı modellere uygulanmıştır (121).

Bu çalışmada kullanılan LIME yöntemlerinin sonucunda modeldeki değişkenlerden rnaXR_591534.3 (LOC103691816), rnaXR_351582.4 (LOC102555118), rnaXR_005499541.1 (LOC120099962), rnaXR_005503408.1 (LOC120102212), rnaXR_005495645.1 (LOC120097836), rnaXR_358189.4 (LOC102549457), rnaXR_133655.1 (LOC106736471), rnaXR_001839007.2 (LOC108351528), rnaXR_005496257.1 (LOC120098190), rnaXR_005499330.1 (LOC120099889), rnaXR_005489439.1 (LOC120094778), rnaXR_005496888.1 (LOC108348888), rnaXR_353438.4 (LOC102554372), rnaXR_005497840.1 (LOC102555751), rnaXR_005503371.1 (LOC120102202), rnaXR_005498350.1 (LOC120099280), rnaXR_005500805.1 (LOC120100781) ve rnaXR_005503535.1 (LOC120102261) id li lncRNA'lar bireysel tahminlerde önemli rol oynamaktadır.

SHAP yöntemi ile edilen değişken önemliliklerine göre, rna-XR_591534.3 (LOC103691816), rna-XR_351582.4 (LOC102555118), rna-XR_005499541.1 (LOC120099962), rna-XR_005503408.1 (LOC120102212) ve rna-XR_005495645.1 (LOC120097836) id li lncRNA'lar en yüksek değişken önemliliklerine sahip olup Nefrotoksisite için öngörücü biyobelirteç adayları olarak kullanılabilir.

Yapılan istatistik analiz sonuçlarına göre, deęişken seçimi ile elde edilen 31 lncRNA dan rna-XR_005493563.1 (LOC120096731) lncRNA dıřındaki tüm genler, iki grup için istatistiksel olarak anlamlı farklılıklar göstermiştir.

Odds deęerleri hesaplanan lncRNA lardan, rna-XR_001839839.2 ve rna-XR_005499594.1 lncRNA'lar daha büyük odds oranları ile nefrotoksisitesi olan grupta önemli ölçüde daha yüksek katlarda ekspresyon göstermiştir. Seçilmiş olan tüm lncRNA ların OR oranları 1 den büyük olup bu lncRNA ların nefrotoksisite grubunda yüksek katlarda yukarı ekspresyona sahip olduęu gözlemlenmiştir. Çalışmada hesaplanmış olan OR deęerleri ile Log2FC deęerlerinin birbirlerini destekledięi görülmüştür.



6. SONUÇ VE ÖNERİLER

Sonuç olarak, bu çalışma ile nefrotoksisitesi olan sıçanlar ile kontrol grubunda yer alan sıçanların lncRNA ekspresyon verileri kullanılarak nefrotoksisite için potansiyel genomik biyobelirteçler tanımlanmış oldu. Gelecekteki çalışmalar ile keşfedilen lncRNA'ların güvenilirliği daha kapsamlı analizler kullanılarak değerlendirilebilir ve klinik olarak doğrulandıktan sonra bu lncRNA'lara dayalı olarak terapötik prosedürler oluşturulabilir ve klinik kullanımları detaylandırılabilir.



KAYNAKLAR

1. Cetinkaya A, Bulbuloglu E, Kurutas EB, Kantarceken B. N-acetylcysteine ameliorates methotrexate-induced oxidative liver damage in rats. *Med Sci Monit Basic Res* 2006, 12(8): 274-8.
2. Bedoui Y, Guillot X, Selambarom J, Guiraud P, Giry C, Jaffar Bandjee MC, Stephane Ralandison S, Gasque P. Methotrexate an old drug with new tricks. *Int J Mol Sci* 2019, 20(20): 5023.
3. Hytioglou P, Tobias H, Saxena R, Abramidou M, Papadimitriou CS, Theise ND. The canals of hering might represent a target of methotrexate hepatic toxicity. *Am J Clin Pathol* 2004, 121(3): 324-9.
4. Sener G, Eksioglu Demiralp E, Cetiner M, Ercan F, Sirvancı S, Gedik N, Yegen MO. L-Carnitine ameliorates methotrexate-induced oxidative organ injury and inhibits leukocyte death. *Cell Biol Toxicol* 2006, 22(1): 47-60.
5. Armağan İ. Metotreksat'ın Karaciğer ve böbrekte neden olduğu hasarda oksidatif stresin rolü. *SDÜ Tıp Fak Derg* 2015, 22(4): 151-5.
6. Guyton A, Hall J. *Tıbbi Fizyoloji*, 11. Baskı. İstanbul, Nobel Tıp Kitabevi, 2007: 303-4.
7. Aşçı H, Özer MK. Metotreksat kaynaklı karaciğer ve böbrek hasarında misoprostolün koruyucu etkisi: *SDÜ Tıp Fak Derg* 2011, 2(2): 125-6.
8. Asvadi I, Hajipour B, Asvadi A, Asl N, Roshangar L, Khodadadi A. Protective effect of pentoxifylline in renal toxicity after methotrexate administration. *Eur Rev Med Pharmacol* 2011, 15(9): 1003-9.
9. Wang EJ, Snyder RD, Fielden MR, Smith RJ, Gu YZ. Validation of putative genomic biomarkers of nephrotoxicity in rats. *Toxicol* 2008, 246(2-3): 91-100.
10. Islam MT, Raihan M, Farzana F, Raju MGM, Hossain MB. An empirical study on diabetes mellitus prediction for typical and non-typical cases using machine learning approaches. *10th IEEE/ICCCNT* 2019, 45670: 1-7.
11. Polikar R. Ensemble learning. In: Zhang C, Ma Y (eds). *Ensemble machine learning*, 1sted. Newyork, Springer, 2012: 1-85.
12. Uzun T. Yapay zeka ve sağlık uygulamaları. *İzmir Katip Çelebi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi* 2020, 3(1): 80-92.
13. Ogunleye A, Wang QG. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans Comput Biol Bioinform* 2019, 17(6): 2131-40.

14. Davagdorj K, Pham VH, Theera Umpon N, Ryu KH. XGBoost-based framework for smoking-induced noncommunicable disease prediction. *Int J Environ Res Public Health* 2020, 17(18): 6513.
15. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Comput Methods Programs Biomed* 2022, 226: 107161.
16. Gimeno M, Villar S, Agirre X, Prosper F, Rubio A, Carazo F. Explainable artificial intelligence for precision medicine in acute myeloid leukemia. *Front Immunol* 2022, 13: 1-13.
17. Eralp MO. Deneysel Olarak Oluşturulan Böbrek Iskemi Reperfüzyon Sonrası Oluşan Böbrek Hasarına Karşı Carnosol'un Koruyucu Etkisinin İncelenmesi. Trakya Üniversitesi Tıp Fakültesi Acil Tıp Anabilim Dalı. Uzmanlık tezi, Edirne: Trakya Üniversitesi, 2013.
18. Seçgin Ş. Sıçanlarda Metotreksat Kaynaklı Böbrek Toksisitesi Üzerine Morinin Koruyucu Etkisi. Lisansüstü Eğitim Enstitüsü, Tıbbi Biyokimya Anabilim Dalı. Yüksek Lisans tezi, Karabük: Karabük Üniversitesi, 2021.
19. Maden M, Aslan V. Deneysel gentamisin nefrotoksisitesinde üriner enzim aktivitelerinin önemi. *Tr J of Veterinary and Animal Sciences* 1999, 23: 29-42.
20. James CW, Steinhaus MC, Szabo S, Dressler RM. Tenofovir-related nephrotoxicity: case report and review of the literature. *Pharmacotherapy* 2004, 24(3): 415-8.
21. Schrier RW, Wang W, Poole B, Mitra A. Acute renal failure: definitions, diagnosis, pathogenesis, and therapy. *J Clin Investig* 2004, 114(1): 5-14.
22. Özbek Bilgin A. Farelerde Parasetamolle İndüklenen Akut Böbrek Toksisitesi Üzerine 5-Ht7 Agonist Ve Antagonistlerinin Etkilerinin Araştırılması. Sağlık Bilimleri Enstitüsü, Tıbbi Farmakoloji Anabilim Dalı. Doktora tezi, Erzurum: Atatürk Üniversitesi, 2015.
23. Somani SM, Husain K, Whitworth C, Trammell GL, Malafa M, Rybak LP. Dose-dependent protection by lipoic acid against cisplatin-induced nephrotoxicity in rats: antioxidant defense system. *J Pharmacol Toxicol* 2000, 86(5): 234-41.
24. Patzer L. Nephrotoxicity as a cause of acute kidney injury in children. *Pediatr Nephrol* 2008, 23(12): 2159-73.
25. Uber AM, Sutherland SM. Nephrotoxins and nephrotoxic acute kidney injury. *Pediatr Nephrol* 2020, 35(10): 1825-33.

26. Zakeri Milani P, Shirani A, Nokhodchi A, Mussa Farkhani S, Mohammadi S, Shahbazi Mojarrad J, Mahmudyan M, Gholikhani T, Farshbaf M, Velizade H. Self-assembled peptide nanoparticles for efficient delivery of methotrexate into cancer cells. *Drug Dev Ind Pharm* 2020, 46(4): 521-30.
27. Farber S, Diamond LK, Mercer RD, Sylvester Jr RF, Wolff JA. Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid (aminopterin). *N Engl J Med* 1948, 238(23): 787-93.
28. Malaviya AN. Landmark papers on the discovery of methotrexate for the treatment of rheumatoid arthritis and other systemic inflammatory rheumatic diseases: a fascinating story. *Int J Rheum Dis* 2016, 19(9): 844-51.
29. Kolli VK, Abraham P, Isaac B, Selvakumar D. Neutrophil infiltration and oxidative stress may play a critical role in methotrexate-induced renal damage. *Chemotherapy* 2009, 55(2): 83-90.
30. Jahovic N, Cevik H, Sehirli AO, Yegen BC, Sener G. Melatonin prevents methotrexate-induced hepatorenal oxidative injury in rats. *J Pineal Res* 2003, 34(4): 282-7.
31. Izzedine H, Launay Vacher V, Karie S, Caramella C, De Person F, Deray G. Is low-dose methotrexate nephrotoxic? Case report and review of the literature. *Clin Nephrol* 2005, 64(4): 315-9.
32. Oktem F, Yilmaz HR, Ozguner F, Olgar S, Ayata A, Uzar E, Uz E. Methotrexate-induced renal oxidative stress in rats: the role of a novel antioxidant caffeic acid phenethyl ester. *Toxicol Ind Health* 2006, 22(6): 241-7.
33. Hasin Y, Seldin M, Lusi A. Multi-omics approaches to disease. *Genome Biol* 2017, 18(1): 1-15.
34. Vlaanderen J, Moore LE, Smith MT, Lan Q, Zhang L, Skibola CF, Rothman N, Vermeulen R. Application of OMICS technologies in occupational and environmental health research; current status and projections. *Occup Environ Med* 2010, 67(2): 136-43.
35. Horgan RP, Kenny LC. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *Obstet Gynecol* 2011, 13(3): 189-95.
36. Bal SH, Budak F. Genomik, proteomik kavramlarına genel bakış ve uygulama alanları. *Uludağ Üniversitesi Tıp Fakültesi Dergisi* 2013, 39(1): 65-9.

37. Bařaran E, Smer A, Cansaran Duman D. General outlook and applications of genomics, proteomics and metabolomics. *Turk Hij Den Biyol Derg* 2010, 67(2): 85-96.
38. Ordovas JM, Corella D. Nutritional genomics. *Annu Rev Genomics Hum Genet* 2004, 5: 71-118.
39. Kadakkuzha BM, Puthanveetil SV. Genomics and proteomics in solving brain complexity. *Mol Biosyst* 2013, 9(7): 1807-21.
40. Przybyla L, Gilbert LA. A new era in functional genomics screens. *Nat Rev Genet* 2022, 23(2): 89-103.
41. Martin DB, Nelson PS. *From genomics to proteomics: techniques and applications in cancer research*. *Trends Cell Biol* 2001, 11: 60-5.
42. arhan A, Ercan E, Yalinkaya T. Dijital PZR ve kullanım alanları. *Turk Hij Den Biyol Derg* 2016, 73(2): 183-98.
43. Garibyan L, Avashia N. Research techniques made simple: polymerase chain reaction (PCR). *J Invest Dermatol* 2013, 133(3): 6-14.
44. Nicomrat D, Dick WA, Tuovinen OH. Assessment of the microbial community in a constructed wetland that receives acid coal mine drainage. *Microb Ecol* 2006, 51(1): 83-9.
45. Dorigo U, Volatier L, Humbert JF. Molecular approaches to the assessment of biodiversity in aquatic microbial communities. *Water Res* 2005, 39(11): 2207-18.
46. Nakatsu C. Soil microbial community analysis using denaturing gradient gel electrophoresis. *Soil Sci Soc Am J* 2007, 71(2): 562-71.
47. Green MR, Sambrook J. Analysis of DNA by Southern blotting. *Cold Spring Harb Protoc* 2021, 2021(7): 251-6.
48. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 1977, 74(12): 5463-7.
49. Rizzo JM, Buck MJ. Key Principles and Clinical Applications of “Next-Generation” DNA Sequencing. *Cancer Prev Res* 2012, 5(7): 887-900.
50. Buermans H, Den Dunnen J. Next generation sequencing technology: advances and applications. *Biochim Biophys Acta Mol Basis Dis* 2014, 1842(10): 1932-41.
51. Kızmaz MZ, Paylan İC, Erkan S. DNA Dizilemenin Tarihsel Geliřimi. *Gaziosmanpařa Bilimsel Arařtırma Dergisi* 2017, 6(2): 47-53.

52. Iida K, Nishimura I. Gene expression profiling by DNA microarray technology. *Crit Rev Oral Biol Med* 2002, 13(1):35-50.
53. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995, 270(5235): 467-70.
54. Yoltaş A, Karaboz İ. DNA mikroarray teknolojisi ve uygulama alanları. *Elektronik Mikrobiyoloji Dergisi TR* 2010, 8: 1-19.
55. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harb Protoc* 2015, 2015(11): 951-69
56. Fernandez Marmiesse A, Gouveia S, Couce ML. NGS technologies as a turning point in rare disease research, diagnosis and treatment. *Curr Med Chem* 2018, 25(3): 404-32.
57. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome profiling in human diseases: new advances and perspectives. *Int J Mol Sci* 2017, 18(8): 1652.
58. Behura SK, Kelleher AM, Spencer TE. Evidence for functional interactions between the placenta and brain in pregnant mice. *FASEB J* 2019, 33(3): 4261-72.
59. Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson MV, Zhang Z, Harland MR. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell Mol Life Sci* 2015, 72: 3425-39.
60. Ali SA, Peffers MJ, Ormseth MJ, Jurisica I, Kapoor M. The non-coding RNA interactome in joint health and disease. *Nat Rev Rheumatol* 2021, 17(11): 692-705.
61. Wei MM, Zhou GB. Long non-coding RNAs and their roles in non-small-cell lung cancer. *Genom Proteom Bioinform* 2016, 14(5): 280-8.
62. Ping G, Xiong W, Zhang L, Li Y, Zhang Y, Zhao Y. Silencing long noncoding RNA PVT1 inhibits tumorigenesis and cisplatin resistance of colorectal cancer. *Am J Transl Res* 2018, 10(1): 138-49.
63. Carbonell Sala S, Uszczyńska Ratajczak B, Lagarde J, Johnson R, Guigo R. Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing (CLS). Functional Analysis of Long Non-Coding RNAs: *Springer* 2021, 2254: 133-59.
64. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* 2016, 17(1): 47-62.

65. Bach DH, Lee SK. Long noncoding RNAs in cancer cells. *Cancer Lett* 2018, 419: 152-66.
66. Ma L, Bajic VB, Zhang Z. On the classification of long non-coding RNAs. *RNA Biol* 2013, 10(6): 924-33.
67. Wong CM, Tsang FC, Ng IOL. Non-coding RNAs in hepatocellular carcinoma: molecular functions and pathological implications. *Nat Rev Gastroenterol Hepatol* 2018, 15(3): 137-51.
68. Morlando M, Ballarino M, Fatica A. Long non-coding RNAs: new players in hematopoiesis and leukemia. *Front Med* 2015, 2: 1-5.
69. McDonel P, Guttman M. Approaches for understanding the mechanisms of long noncoding RNA regulation of gene expression. *Cold Spring Harb Perspect Biol* 2019, 11(12): 1-17.
70. Liu Q, Sun S, Yu W, Jiang J, Zhuo F, Qiu G, Xu S, Jiang X. Altered expression of long non-coding RNAs during genotoxic stress-induced cell death in human glioma cells. *J Neurooncol* 2015, 122(2): 283-92.
71. Inamura K. Major tumor suppressor and oncogenic non-coding RNAs: clinical relevance in lung cancer. *Cells* 2017, 6(2): 12.
72. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol* 2011, 7(3): 1-5.
73. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? An introduction and overview. *Yearb Med Inform* 2001, 10(01): 83-100.
74. Cohen J. Bioinformatics—an introduction for computer scientists. *ACM Comput Surv* 2004, 36(2): 122-58.
75. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinformatics* 2019, 20(6): 2044-54.
76. Zencirli K. Bipolar Parsiyel Protez Uygulanmış Kalça Kırıklı Hastalarda Makine Öğrenme Yöntemleri ile Perioperatif Prognoz ve Maliyet Analizi. Ortopedi ve Travmatoloji Anabilim Dalı. Tıpta Uzmanlık tezi, Erzurum: Atatürk Üniversitesi, 2020.
77. Eser G. Tahmin Problemleri İçin Regresyon Ağacı ve Komşuluk Tabanlı Yöntemler Geliştirilmesi: Kalıpcılık Sektöründe Bir Uygulama. Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalı. Yüksek Lisans tezi, Bursa: Uludag Üniversitesi, 2021.

78. Şen F. Veri Madenciliği ile Birliktelik Kurallarının Bulunması. Fen Bilimleri Enstitüsü, Bilgisayar ve Bilişim Mühendisliği Anabilim Dalı. Yüksek Lisans tezi, Sakarya: Sakarya Üniversitesi, 2008.
79. Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, Lyu J. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil Med Res* 2021, 8: 1-12.
80. Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. *Comput Methods Programs Biomed* 2016, 130: 87-92.
81. Sağaltıcı D, Alay FD, Cihan E, İlhan N. Veri Madenciliği Yöntemleri İle Meteorolojik Verilerden Kayıp Güneş Işınım Değerlerinin Tahmini. *Harran Üniv Müh Derg* 2018, 3(2): 49-53.
82. Koyuncugil A, Özgülbaş N. Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi*. 2009, 2(2): 21-32.
83. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. *Comput Methods Programs Biomed* 2015, 119(3): 181-5.
84. Brown ML, Kros JF. Data mining and the impact of missing data. *Ind Manag Data Syst* 2003, 103(8): 611-21.
85. Jain D, Singh V. Feature selection and classification systems for chronic disease prediction: A review *Egypt Inform J* 2018, 19(3): 179-89.
86. Liu S, Xu C, Zhang Y, Liu J, Yu B, Liu X, Dehmer M. Feature selection of gene expression data for cancer classification using double RBF-kernels. *BMC Bioinform* 2018, 19(1): 1-14.
87. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng* 2014, 40(1): 16-28.
88. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006, 101(476): 1418-29.
89. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 2007, 94(3): 691-703.
90. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Methodol* 2005, 67(2): 301-20.
91. Kursu MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010; 36(11): 1-13.

92. Saltz JS, Krasteva I. Current approaches for executing big data science projects-a systematic literature review. *PeerJ Comput Sci* 2022; 8: 862.
93. Betancur RR, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, Lecointre G, Orti G. Phylogenetic classification of bony fishes. *BMC Evol Biol* 2017, 17: 1-40.
94. Akpınar H. *Data: Veri Madenciliği Veri Analizi*, 1. Baskı. İstanbul, Papatya Yayıncılık, 2014: 10-24.
95. Hu L, Li L. Using Tree-Based Machine Learning for Health Studies: Literature Review and Case Series. *Int J Environ Res Public Health* 2022, 19(23): 16080.
96. Mao Y, He Y, Liu L, Chen X. Disease classification based on eye movement features with decision tree and random forest. *Front Neurosci* 2020, 14: 798.
97. Hajjej F, Alohalı MA, Badr M, Rahman MA. A comparison of decision tree algorithms in the assessment of biomedical data. *Biomed Res Int* 2022, 9449497: 1-9.
98. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst Appl* 2019, 134: 93-101.
99. Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 2006, 9(2): 181-99.
100. Asadi S, Roshan S, Kattan MW. Random forest swarm optimization-based for heart diseases diagnosis. *J Biomed Inform* 2021, 115: 1-13.
101. Telli S. Topluluk Öğrenme Yöntemini Kullanarak Twitter Verisi Üzerinde Duygu Algılama ve Tanıma. Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı. Yüksek Lisans tezi, İzmir: Ege Üniversitesi, 2019.
102. Yangın G. XGboost ve Karar Ağacı Tabanlı Algoritmaların Diyabet Veri Setleri Üzerine Uygulaması. Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı. Yüksek Lisans tezi, İstanbul: Mimar Sinan Güzel Sanatlar Üniversitesi, 2019.
103. Sevinç E. An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. *Comput Ind Eng* 2022, 165: 107912.
104. Dixit A. *Ensemble Machine Learning: A Beginner's Guide that Combines Powerful Machine Learning Algorithms to Build Optimized Models*, 1st ed. England, Packt Publishing Ltd, 2017: 103-48.

105. Ramraj S, Uzir N, Sunil R, Banerjee S. Experimenting XGBoost algorithm for prediction and classification of different datasets. *Int J Control Theory Appl* 2016, 9(40): 651-62.
106. Moore A, Bell M. XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. *Clin Med Insights Cardiol* 2022, 16: 1-6.
107. Stafford C, Marrero WJ, Naumann RB, Lich KH, Wakeman S, Jalali MS. Identifying key risk factors for premature discontinuation of opioid use disorder treatment in the United States: A predictive modeling study. *Drug Alcohol Depend* 2022, 237: 109507.
108. Choubin B, Abdolshahnejad M, Moradi E, Querol X, Mosavi A, Shamshirband S, Shamshirband S, Ghamisi P. Spatial hazard assessment of the PM10 using machine learning models in Barcelona. *Spain Sci Total Environ* 2020; 701: 134474.
109. Kumarakulasinghe NB, Blomberg T, Liu J, Leao AS, Papapetrou P. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. *IEEE Computer-Based Medical Systems* 2020, 33: 7-12.
110. Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Comput Methods Programs Biomed* 2020, 126: 104041.
111. Rodriguez Perez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *J Comput Aided Mol Des* 2020, 34(10): 1013-26.
112. Anjum M, Khan K, Ahmad W, Ahmad A, Amin MN, Nafees A. New SHapley Additive ExPlanations (SHAP) Approach to Evaluate the Raw Materials Interactions of Steel-Fiber-Reinforced Concrete. *Materials* 2022, 15(18): 6261.
113. Alkadhim HA, Amin MN, Ahmad W, Khan K, Nazar S, Faraz MI, Imran M. Evaluating the Strength and Impact of Raw Ingredients of Cement Mortar Incorporating Waste Glass Powder Using Machine Learning and SHapley Additive ExPlanations (SHAP) Methods. *Materials* 2022, 15(20): 7344.
114. Yan H, Zheng G, Qu J, Liu Y, Huang X, Zhang E, Cai Z. Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. *J Cell Physiol* 2019, 234(12): 23785-97.

115. Hao J, Ho TK. Machine learning made easy: a review of scikit-learn package in python programming language. *J Educ Behav Stat* 2019, 44(3): 348-61.
116. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. Shiny: web application framework for R. *R package version* 2017, 1(5): 2017.
124. Wettenhall JM, Smyth GK. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* 2004, 20(18): 3705-6.
117. Nápoles-Duarte J, Biswas A, Parker MI, Palomares-Baez J, Chávez-Rojó M, Rodríguez-Valdez L. Stmol: A component for building interactive molecular visualizations within streamlit web-applications. *Front Mol Biosci* 2022, 9: 990846.
118. Liyanage T, Ninomiya T, Jha V, Neal B, Patrice HM, Okpechi I, Zhao M, Lv J, Garg AX, Knight J, Rodgers A, Gallagher M, Kotwal S, Cass A, Perkovic V. Worldwide access to treatment for end-stage kidney disease: a systematic review. *Lancet* 2015, 385(9981): 1975-82.
119. Kim SY, Moon A. Drug-induced nephrotoxicity and its biomarkers. *Biomol Ther* 2012, 20(3): 268.
120. Bazmandegan G, Amirteimoury M, Kaeidi A, Shamsizadeh A, Khademalhosseini M, Nematollahi MH, Hassanipour M, Fatemi I. Sumatriptan ameliorates renal injury induced by cisplatin in mice. *Iran J Basic Med Sci* 2019, 22(5): 563.
121. Dosilovic FK, Brcic M, Hlupic N. Explainable artificial intelligence: A survey. *IEEE MIPRO* 2018, 41: 210-5.

EKLER

EK-1. Özgeçmiş



EK-2. Etik Kurul Onayı

