AÇIKLANABİLİR YAPAY ZEKÂ YÖNTEMLERİNE DAYALI BİLGİSAYAR DESTEKLİ TAHMİN MODELİNİN GELİŞTİRİLMESİ: METABOLOMİK TEKNOLOJİLERİ ÜZERİNE UYGULAMASI

Fatma Hilal YAĞIN

BİYOİSTATİSTİK ve TIP BİLİŞİMİ ANABİLİM DALI

Tez Danışmanı
Doç. Dr. Harika Gözde GÖZÜKARA BAĞ
İkinci Tez danışmanı
Dr. Öğr. Üyesi Dinçer GÖKSÜLÜK

Doktora Tezi- 2024

T.C.
İNÖNÜ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

AÇIKLANABİLİR YAPAY ZEKÂ YÖNTEMLERİNE DAYALI BİLGİSAYAR DESTEKLİ TAHMİN MODELİNİN GELİŞTİRİLMESİ: METABOLOMİK TEKNOLOJİLERİ ÜZERİNE UYGULAMASI

Fatma Hilal YAĞIN

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı
Doktora Tezi

Tez Danışmanı
Doç. Dr. Harika Gözde GÖZÜKARA BAĞ

İkinci Tez danışmanı
Dr. Öğr. Üyesi Dinçer GÖKSÜLÜK

MALATYA
2024

T.C.

İNÖNÜ ÜNİVERSİTESİ

Sağlık Bilimleri Enstitüsü Müdürlüğüne

ETİK BEYANI

İnönü Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak "Doç. Dr. Harika Gözde GÖZÜKARA BAĞ" ve "Dr. Öğr. Üyesi Dinçer GÖKSÜLÜK" danışmanlığında hazırlayıp sunduğum "Açıklanabilir Yapay Zekâ Yöntemlerine Dayalı Bilgisayar Destekli Tahmin Modelinin Geliştirilmesi: Metabolomik Teknolojileri Üzerine Uygulaması" 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. 22/02/2024

Fatma Hilal YAĞIN

İÇİNDEKİLER

ÖZET vii
ABSTRACT viii
SİMGELER VE KISALTMALAR DİZİNİ ix
ŞEKİLLER DİZİNİ x
TABLOLAR DİZİNİ xii
1. GİRİŞ 1
2. GENEL BİLGİLER 5
2.1. Polikistik Over Sendromu (PKOS) 5
2.2. Biyoinformatik ve Omik Teknolojiler 6
2.2.1. Genomik 6
2.2.2. Transkriptomik 7
2.2.3. Proteomik 7
2.2.4. Metabolomik 8
2.3. Yapay Zekâ ve Makine Öğrenimi 13
2.4. Açıklanabilir Yapay Zekâ 14
2.4.1. Yorumlanabilir Yöntemler 15
2.4.2. Model-Agnostik Açıklamalar 16
2.4.3. Örnek Tabanlı Açıklamalar 18
3. MATERYAL VE METOT 21
3.1. Metabolomik Veri Seti 21
3.2. Veri Önişleme 23
3.2.1. Rasgele Orman Algoritması ile Kayıp Değer Atama 24
3.2.2. Borderline-SMOTE 24
3.2.3. Değişken Seçimi 25
3.3. Çalışmada Kullanılan Ağaç Tabanlı Makine Öğrenmesi Algoritmaları 28
3.3.1. Rastgele Orman Algoritması 28
3.3.2. CatBoost Algoritması 29
3.3.3. XGBoost Algoritması 30
3.3.4. AdaBoost Algoritması 32
3.3.5. LightGBM Algoritması 33
3.3.6. NGBoost Algoritması 34

3.4. Hiperparametre Optimizasyonu 34
3.5. Performans Değerlendirme Metrikleri 35
3.6. Çalışmada Kullanılan Açıklanabilir Yapay Zekâ Algoritmaları 36
3.6.1. Yerel Yorumlanabilir Model-Bağımsız Açıklamalar 36
3.6.2. SHapley Katkı Açıklamaları 37
3.7. Veri analizi 37
3.7.1. Biyoistatistiksel Analizler 37
3.7.2. Biyoinformatik Analizler 38
4. BULGULAR 40
4.1. Biyoinformatik Analizlere İlişkin Bulgular 40
4.2. Biyoistatistiksel Analiz ve Ağaç Tabanlı Makine Öğrenmesi Modellerine İlişkin Bulgular 46
4.3. Açıklanabilir Yapay Zekâ Yöntemlerine İlişkin Bulgular 50
5. TARTIŞMA 55
6. SONUÇ VE ÖNERİLER 61
KAYNAKLAR 62
EKLER 72
EK-1. Özgeçmiş 72
EK-2. Etik Kurul Almama Gerekçesi 73

TEŞEKKÜR

Tez çalışmamın planlanması, yürütülmesi sürecinde deneyimleriyle bana yol gösteren, lisansüstü eğitimim boyunca bilgi ve deneyimleriyle benden desteğini esirgemeyen değerli danışman hocam Sayın Doç. Dr. Harika Gözde GÖZÜKARA BAĞ'a ve ikinci danışmanım Sayın Dr. Öğr. Üyesi Dinçer GÖKSÜLÜK'e, lisansüstü eğitimim boyunca desteklerini hiç esirgemeyen ve her zaman deneyimlerini paylaşarak bana ışık tutan değerli hocalarım Sayın Prof. Dr. Cemil ÇOLAK, Sayın Prof. Dr. Saim YOLOĞLU, Sayın Doç. Dr. Emek GÜLDOĞAN ve Sayın Dr. Öğr. Üyesi Ahmet Kadir ARSLAN'a, bu tez çalışmasının izleme raporlarında değerli görüşleriyle çalışmama katkı sağlayan Sayın Prof. Dr. Elif YEŞİLADA'ya sonsuz teşekkür ve saygılarımı sunarım. Anabilim dalında birlikte görev yaptığım değerli asistan arkadaşlarım Şeyma YAŞAR, İpek BALIKÇI ÇİÇEK ve Zeynep KÜÇÜKAKÇALI'ya destekleri için çok teşekkür ederim. Bu süreçte her daim yanımda olan, her türlü desteğini esirgemeyen sevgili annem Birgül UZUNOĞLU, babam İbrahim UZUNOĞLU ve kardeşlerime teşekkürlerimi sunuyorum. Sevgisi, sabrı ve inancı ile bu zorlu ama öğretici yolculuk boyunca beni motive ederek desteğiyle her başarıma ortak olan çok kıymetli eşim ve hayat arkadaşım Burak YAĞIN'a sonsuz teşekkürlerimi sunuyorum.

Arş. Gör. Fatma Hilal YAĞIN

ÖZET

Açıklanabilir Yapay Zekâ Yöntemlerine Dayalı Bilgisayar Destekli Tahmin Modelinin Geliştirilmesi: Metabolomik Teknolojileri Üzerine Uygulaması

Amaç: Bu çalışmanın amacı polikistik over sendromunun (PKOS) tanısında uygun metabolomik biyobelirteçlerin biyoinformatik ve yapay zekâ (AI) yaklaşımları ile saptanması, ağaç tabanlı makine öğrenme (ML) yöntemleri kullanılarak PKOS için prognostik bir modelin geliştirilmesi ve açıklanabilir AI (XAI) kullanılarak optimal modelin klinik olarak yerel ve küresel açıdan yorumlanmasıdır.

Materyal ve Metot: Veri seti 22 kontrol ve 67 PKOS hastasına ilişkin serum örneklerine uygulanan metabolomik analizler sonucu elde edilen 70 metabolitten oluşmaktadır. PKOS'un moleküler alt yapısını inceleyebilmek için benzerlik ağ, kümelemeye dayalı ısı haritası ve yolak zenginleştirme analizleri uygulandı. Biyobelirteç keşfi için kat değişimi, kısmi en küçük kareler diskriminant analizi (PLS-DA) projeksiyon değişken önem (VIP) skoru, Boruta ve minimum artıklık ve maksimum alaka (MRMR) yöntemleri kullanıldı. Bu dört yöntemde ortak olan metabolitler biyobelirteç olarak belirlendi ve altı farklı ağaç tabanlı ML algoritması [rastgele orman, kategorik arttırma (CatBoost), aşırı gradyan arttırma (XGBoost), hafif gradyan artırma makinesi (LightGBM), uyarlanabilir arttırma (AdaBoost) ve doğal gradyan arttırma (NGBoost)] kullanıldı. Optimal modelin klinik açıklamaları yorumlanabilir model-agnostik açıklamalar (LIME) ve Shapley katkı açıklamaları (SHAP yöntemleri) ile incelendi.

Bulgular: Taurolitokolik asit (M-H)-, Hipoksantin (M-H)-, Homovanilik asit (2M-H)-, S-Adenosil-L-Homosistein (M-H)-, Tridekanoik asit (M-H)- ve Süksinik asit (M-H)- biyobelirteç olarak belirlendi. Optimal tahmin modeli XGBoost için doğruluk, AUC ve duyarlılık ölçütleri sırasıyla %92.3, %95.4, %90.9 skoruna sahipti. LIME sonuçlarına göre %100 olasılıkla PKOS olarak tahmin edilen bir hastada özellikle Tridekanoik asit (M-H)- seviyesinin yüksek olması modelin doğru pozitif kararını açıkladı. SHAP açıklamalarında Tridekanoik asit (M-H)-, Süksinik asit (M-H)- ve Hipoksantin (M-H)- seviyelerindeki artışın, Homovanilik asit (2M-H)- ve Taurolitokolik asit (M-H)- seviyelerindeki azalmanın PKOS riskini arttırdığı gözlendi.

Sonuç: Sonuç olarak XGBoost modeli PKOS'u başarılı bir şekilde tespit edebildi. Biyobelirteç metabolitlerin PKOS ile ilişkisinin anlaşılması, hastalığın moleküler düzeyde açıklanmasının yanısıra daha etkili tanı ve tedavi yöntemleri geliştirmek için önemlidir.

Anahtar Kelimeler: Açıklanabilir Yapay Zekâ, Biyoinformatik, Metabolomik, Makine Öğrenmesi, Polikistik Over Sendromu

ABSTRACT

Development of a Computer-Aided Predictive Model Based on Explainable Artificial Intelligence Methods: Application on Metabolomics Technologies

Aim: The aim of this study was to identify metabolomic biomarkers suitable for the diagnosis of polycystic ovary syndrome (PCOS) using bioinformatics and artificial intelligence (AI) approaches, to develop a prognostic model for PCOS using tree-based machine learning (ML) methods, and to interpret the optimal model clinically locally and globally using explainable AI (XAI).

Material and Method: The dataset consists of 70 metabolites obtained from metabolomic analysis of serum samples of 22 controls and 67 PCOS patients. Similarity network, clustering-based heatmap, and pathway enrichment analyses were applied to examine the molecular substructure of PCOS. For biomarker discovery, fold change, partial least squares discriminant analysis (PLS-DA), projection variable importance (VIP) score, Boruta, and minimum redundancy and maximum relevance (MRMR) methods were used. Metabolites common to these four methods were identified as biomarkers and six different tree-based ML algorithms [random forest, categorical boosting (CatBoost), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), adaptive boosting (AdaBoost) and natural gradient boosting (NGBoost)] were used. Clinical explanations of the optimal model were analyzed by interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP).

Results: Taurolithocholic acid (M-H)-, Hypoxanthine (M-H)-, Homovanilic acid (2M-H)-, S-Adenosyl-L-Homocysteine (M-H)-, Tridecanoic acid (M-H)- and Succinic acid (M-H)- were identified as biomarkers. The accuracy, AUC and sensitivity measures for the optimal prediction model XGBoost were 92.3%, 95.4%, and 90.9%, respectively. The high level of Tridecanoic acid (M-H)- especially in a patient predicted as PCOS with 100% probability according to LIME results, explained the model's true positive decision. SHAP annotation showed that increased levels of Tridecanoic acid (M-H)-, Succinic acid (M-H)- and Hypoxanthine (M-H)- and decreased levels of Homovanilic acid (2M-H)- and Taurolithocholic acid (M-H)- increased the risk of PCOS.

Conclusion: In conclusion, the XGBoost model was able to successfully detect PCOS. Understanding the relationship between biomarker metabolites and PCOS is important to explain the disease at the molecular level as well as to develop more effective diagnostic and therapeutic methods.

Keywords: Explainable Artificial Intelligence, Bioinformatics, Metabolomics, Machine learning, Polycystic Ovary Syndrome

SİMGELER VE KISALTMALAR DİZİNİ

AI : Yapay Zekâ

AdaBoost : Adaptif Boosting

ALE : Birikmiş Yerel Etkiler

CatBoost : Kategorik Arttırma

CV : Çapraz Doğrulama

FDR : Yanlış Keşif Oranları

FC : Kat Değişimi

ICE : Bireysel Koşullu Beklenti

LC-MS : Sıvı kramootografisi-Kütle Spektrometrisi

LightGBM : Light Gradient Boosting Machines

LIME : Yerel Yorumlanabilir Model-Bağımsız Açıklamalar

LS-MS : Sıvı Kromatografisi Kütle Spektrometrisi

MRMR : Minimum Artıklık ve Maksimum Alaka

NMR : Nükleer Manyetik Rezonans

NGBoost : Natural Gradient Boosting

PDP : Kısmi Bağımlılık Grafiği

PKOS : Polikistik Over Sendromu

PLS-DA : Kısmi En Küçük Kareler Diskriminant Analizi

ROC : Alıcı İşletim Karakteristiği

SMOTE : Sentetik Azınlık Aşırı Örnekleme Tekniği

SHAP : SHapley Katkı Açıklamaları

VIP : Projeksiyon Değişken Önem

XAI : Açıklanabilir Yapay Zekâ

XGBoost :E Xtreme Gradient Boosting

ŞEKİLLER DİZİNİ

Şekil No **Sayfa No**

Şekil 2.1. Normal ve Polikistik Over Görünümü 5
Şekil 2.2. Omik teknolojileri 6
Şekil 2.3. Genotip-fenotip sürekliliği 9
Şekil 3.1. MRMR değişken seçimi yöntemi 28
Şekil 3.2. Rasgele orman algoritması 29
Şekil 3.3. AdaBoost algoritması gösterimi 33
Şekil 3.4. LIME Adımları 37
Şekil 3.5. Kara Kutu Modeli ve SHAP yöntemi 37
Şekil 4.1. Metabolitlere İlişkin Moleküler Yapı Benzerlik Ağ Analizi Sonuçları 40
Şekil 4.2. Diferansiyel Ekspresyon Gösteren Metabolitlere İlişkin Volkano Grafiği 42
Şekil 4.3. İlk 25 Önemli Metabolitin Ekspresyonuna İlişkin Isı Haritası 43
Şekil 4.4. PLS-DA Modelinde VIP Grafiği 43
Şekil 4.5. Taurolitokolik asit (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği 46
Şekil 4.6. Hipoksantin (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği 47
Şekil 4.7. Homovanillik asit (2M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği 47
Şekil 4.8. S-Adenosil-L-Homosistein (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği 47
Şekil 4.9. Tridekanoik asit (M-H)- için orjinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği 48
Şekil 4.10. Süksinik asit (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği 48
Şekil 4.11. XGBoost Modelinin Performans Ölçütlerine İlişkin Grafiksel Gösterim 50
Şekil 4.12. XGBoost Modelinin Tahmin Sonuçlarına Göre Doğru Negatif (Kontrol) Bir Hasta için LIME Açıklamaları 50
Şekil 4.13. XGBoost Modelinin Tahmin Sonuçlarına Göre Doğru Pozitif (PKOS) Bir Hasta için LIME Açıklamaları 51

Şekil 4.14. Biyobelirteç Adayı Metabolitlerin Ortalamaya Göre Önem Sırası (|SHAP değeri|) 52
Şekil 4.15. XGBoost modelinin SHAP özet grafiği ile yorumlanması 53

TABLOLAR DİZİNİ

Tablo No **Sayfa No**

Tablo 2.1. Yorumlanabilir yöntemlerin özellikleri 16

Tablo 3.1. Metabolitlere ilişkin süper, ana ve alt sınıf bilgileri 21

Tablo 4.1. Biyoinformatik Analiz Bulguları 41

Tablo 4.2. Yolak Zenginleştirme Analizi Bulguları (Kolmogorov-Smirnov testi)..... 44

Tablo 4.3. Biyobelirteç Keşfi Sürecinde Kullanılan Yöntemlerde Belirlenen Önemli Metabolitler..... 46

Tablo 4.4. Ağaç Tabanlı Makine Öğrenmesi Modellerin Performans Ölçütlerine İlişkin Bulgular 49

Tablo 4.5. Normalize edilmiş SHAP değerlerine göre metabolitlerin PKOS tanısındaki önem düzeyleri 54

1. GİRİŞ

Polikistik over sendromu (PKOS), üreme çağındaki kadınlarda en sık görülen endokrinolojik bir hastalık olup, bozulmuş glikoz metabolizması ve insülin direnci de dâhil olmak üzere çoklu metabolik bozukluklarla ilişkilidir. Son çalışmalar, endokrin ve genetik etkenleri de içeren birçok faktörün PKOS patogenezinde rol oynadığını bildirmektedir. PKOS'lu kadınlarda, subfertilite (kısırlık), tip 2 diyabet, hirsutizm (aşırı tüylenme), akne, kardiyovasküler hastalıklar, endometriyal kanser, menstrüel anormallikler ve polikistik yumurtalıkların görülmesi gibi bir dizi komplikasyon riski yüksektir. Komplikasyonların doğru bir şekilde tanımlanması, PKOS'un diğer hormonal bozukluklardan ayrılmasına yardımcı olur (1, 2).

PKOS'da gözlenen komplikasyonlar hem bireysel sağlık hem de sağlık ve sosyal güvenlik sistemleri için önemli mali yükler oluşturabilir. PKOS ile ilişkili komplikasyonların sayısı ve ciddiyeti, tanı, tedavi ve yönetiminin sağlık sistemlerine önemli maliyetler getirmesine neden olmaktadır. Amerika Birleşik Devletleri'ndeki veriler, PKOS'lu hastaların ilk değerlendirme maliyetlerinin 93 milyon dolar olduğunu göstermektedir. Ayrıca menstrüel bozukluklar ve anormal uterin kanama tedavisinde kullanılan hormonal ilaçlar için 1,35 milyar dolar, infertilite tedavileri için 533 milyon dolar, PKOS ile ilişkili diyabet tedavisi için 1,77 milyar dolar ve hirsutizm tedavisi için 622 milyon dolarlık bir mali yüke neden olduğunu göstermektedir. Ülkeden ülkeye değişmekle birlikte, PKOS'un sağlık sistemleri için önemli bir ekonomik yük olduğu bilinmektedir. Bu nedenle, PKOS'un etkilerini ve komplikasyonlarını azaltmak için erken teşhis, etkili tedavi ve yönetim stratejilerinin uygulanması önemlidir (3).

PKOS tanısında farklı yaklaşımlar ve kriterler bulunmaktadır. Amerika Birleşik Devletleri Ulusal Sağlık Enstitüsü (National Institutes of Health, NIH) tarafından 1990 yılında yapılan bir konferansta, klinik veya biyokimyasal hiperandrojenizm ve kronik oligo-anovulasyonun (adet düzensizliği) bir arada bulunduğu durumlar PKOS olarak tanımlanmıştır (4). Ancak, 2003 yılında Hollanda'da gerçekleştirilen bir toplantıda, ultrasonografi (USG) ile polikistik over morfolojisinin (PKOM) varlığı da tanı kriterlerine eklenmiştir. Bu durum, PKOS tanısı alan hasta sayısının artmasına yol açmış, ancak hastalığın daha heterojen bir hal almasına ve hiperandrojenizmi olmayan bazı hastalara da PKOS tanısı konulmasına yol açmıştır (5). Ayrıca, literatürdeki bilimsel kanıtlar, hiperandrojenizmin PKOS'nin patofizyolojisinde ve bu hastalıkla ilişkili

metabolik disfonksiyonda önemli bir rol oynadığını göstermektedir. Bazı araştırmalar, hiperandrojenizmin PKOS'nin temel özelliklerinden biri olduğunu ve tanı kriterlerinin hiperandrojenizmi içermesi gerektiğini öne sürmektedir (6, 7). 2006 yılında Androjen Excess & PKOS Cemiyeti, PKOS tanısı koymak için klinik veya biyokimyasal hiperandrojenizmin olmazsa olmaz şart olduğunu öne sürmüş ve PKOS tanısında yeni bir tanı kriteri önermiştir (8).

Bu farklı tanı yaklaşımları ve kriterler, PKOS tanısının karmaşıklığını artırırken, klinik uygulamada karışıklıklara ve takip süreçlerinde zorluklara neden olmaktadır. Ayrıca, PKOS'un etyopatogenezinin (nedenlerinin) anlaşılması ve kısa/uzun vadeli komplikasyonların takibi gibi konularda da sorunlar yaratabilir (9). Dolayısıyla, PKOS tanısı konulması için kullanılan farklı yaklaşımlar ve tanı kriterleri, PKOS'un anlaşılmasını ve tedavi yönetimini zorlaştırabilir. Bu nedenle, klinik uygulamada ve araştırmalarda daha tutarlı ve evrensel kabul gören tanı kriterleri ve yaklaşımların kullanılması önemlidir.

Kapsamlı çalışmalara rağmen, PKOS patogenezi henüz tam olarak saptanamamışken, PKOS'un oluşum mekanizması da net olarak ortaya konulamamış olup, patofizyolojisi hala belirsizdir (10, 11). Son yıllarda, bu belirsizliğe ışık tutabilmek için "omik" yaklaşımlar kullanılmaktadır. Bu önemli yaklaşımlar arasında yer alan metabolomik; hücreler, biyolojik sıvılar, dokular veya organizmalar içerisinde metabolitler olarak adlandırılan küçük moleküllerin kapsamlı analizlerini içeren çalışmalardır. Metabolomik teknolojisi, sırasıyla hücresel ve genetik ekspresyon düzeylerinde uyarıcılara karşı canlı sistemlerin yanıtlarını ölçen genomik, transkriptomik ve proteomik'den sonra gelen en son omik teknolojidir ve kanıt düzeyi bu omik teknolojilerden daha yüksektir (12).

Metabolomik, metabolik kontrol teorisinden ortaya çıkmıştır ve temelinde bir hücrenin metabolik bileşimi olarak tanımlanan metabolom vardır. Metabonomikte ise enzimatik süreçten ziyade bir kimyasal süreçten meydana gelen moleküller, dış kaynaklı türler, metabolit dalgalanmalar ve endojen metabolitlerin biyolojik sıvı konsantrasyonları değerlendirilmektedir. Bu anlamda metabolomik, metabonomik tarafından kapsanmış sistemlerin bir alt kümesi olarak düşünülebilir. Metabolomik çalışmaların amacı metabolik parmak izi verilerinin birincil kaynakları olarak çeşitli biyolojik sıvılar, doku özütleri ve hücreleri kullanarak bir metabolik fenotiplendirme yapmaktır. Bu konuda serum, seminal plazma, foleküler sıvı ve idrar en çok çalışılan biyolojik sıvılardır (13, 14).

Metabolit biyobelirteçlerin saptanması ve kantifikasyonu analitik, biyokimyasal ve spektral analizin bir kombinasyonu ile gerçekleştirilir. Bu sayede sağlıklı kontrol populasyonu ve spesifik hastalıklara sahip denekler için imza metabolitler oluşturulur. Bu spesifik biyobelirteçlerin konsantrasyonundaki farklılıklar, test deneklerinin kontrollerden farklılaşmasına yardımcı olacak benzersiz metabolik profiller oluşturmak için kullanılır. Metabolitleri ayrıntılarıyla ortaya çıkarmak metabolomiğin kararsız yapısı, karmaşıklığı ve heterojenliği nedeniyle önemli bir analitik zorluk olsa da hızlı teknolojik ilerlemeler geniş bir yelpazede bu moleküllerin tespit edilmesine olanak sağlamaktadır (15).

Metabolomik analizler ile belirlenecek olası biyobelirteçler, PKOS'un erken tanı ve takibi ile tedavi stratejilerinin yönlendirilmesi gibi önemli amaçlarda kullanılabilmektedir. PKOS tanısı ve patogenezinin ortaya konulmasına yönelik serum veya diğer sıvıların deneysel metabolomik analizlerinden elde edilecek verileri ve biyobelirteçleri modern yapay zekâ (AI) ve makine öğrenmesi (ML) yaklaşımları ile tahmin edebilecek araştırmalara ihtiyaç bulunmaktadır (16).

AI ve ML, bilgisayar sistemlerinin tıbbi verileri analiz etmek, öğrenmek, sonuçları tahmin etmek ve kararlar almak için kullanıldığı bir alanı ifade eder. AI/ML karmaşık veri kümelerini analiz etme ve desenleri tanıma yeteneği sayesinde tıbbi teşhis, tedavi planlaması, hasta izlemi ve sağlık yönetimi gibi birçok alanda önemli katkılar sağlamaktadır (17).

AI/ML yöntemleri; hastaların genetik bilgilerini, metabolomik belirteçlerini, hormon seviyelerini, görüntüleme sonuçlarını ve klinik belirtilerini değerlendirerek PKOS teşhisini yapabilir. PKOS'un yönetimi, semptomları hafifletmek, hormonal dengesizlikleri düzeltmek ve uzun vadeli komplikasyonları önlemek için bir tedavi planını içerir. AI/ML yöntemleri, hastaların klinik değişkenlerini ve laboratuvar sonuçlarını analiz ederek, kişiselleştirilmiş tedavi planları geliştirebilir. Bu yöntemler PKOS hastalarında hormonal tedavi, yaşam tarzı değişiklikleri veya ilaçların etkinliği ve yan etkileri hakkında tahminler yapabilir ve en uygun tedavi seçeneklerini belirleyebilir (1).

AI/ML'ye dayalı akıllı sistemler tıp alanında büyük olanaklar sunsa da açıklanabilir yapay zeka'nın (XAI) kullanımı, bu tür karmaşık sistemlerin altında yatan karar verme süreçlerinin, iç çalışma mekanizmalarının ve sonuçlarının insanlar tarafından anlaşılabilir ve izlenebilir olması açısından önemlidir. Açıklanabilirlik, özellikle karmaşık ve büyük veri setleri üzerinde çalışan karmaşık AI sistemlerinde önemlidir (18).

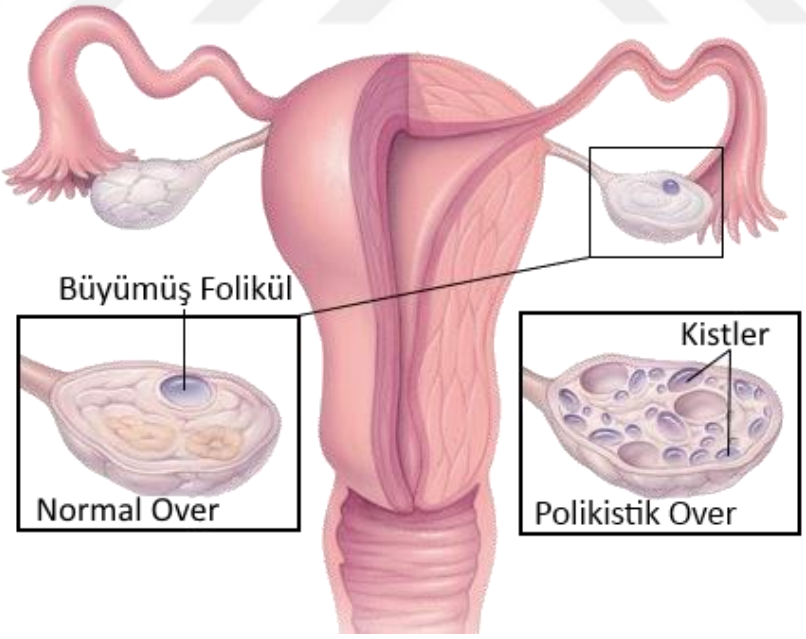
Bu tez çalışmasında, PKOS'un ayırt edilmesi ve biyobelirteçlerinin belirlenebilmesi için deneysel olarak elde edilmiş olan açık kaynaklı metabolomik veriler kullanılarak, ağaç tabanlı ML ve XAI yaklaşımlarına dayalı yorumlanabilir bir tahmin modelinin geliştirilmesi amaçlanmıştır.

2. GENEL BİLGİLER

2.1. Polikistik Over Sendromu (PKOS)

PKOS, kadınların üreme sistemi ve hormonal dengesini etkileyen bir endokrin bozukluktur. Bu sendrom, genellikle üreme çağındaki kadınları etkileyen, belirli semptomlara ve hormonal değişikliklere yol açan çok yaygın bir durumdur. PKOS'lu kadınların ana özelliklerinden biri, overlerindeki foliküllerin (yumurtalık keseleri) olgunlaşmayarak küçük kistler oluşturmasıdır (Şekil 2.1). Bu kistler, düzensiz adet döngüleri, infertilite (kısırlık) ve hormonal değişikliklere neden olabilir (19).

PKOS'un belirtileri arasında adet düzensizlikleri, aşırı tüylenme (hirsutizm), akne, yağlı cilt, kilo artışı, insülin direnci ve bazen tip 2 diyabet riskinin artması yer alır. Ayrıca, PKOS, kadınların sağlığını daha fazla etkileyebilir, özellikle metabolik sendrom ve kardiyovasküler hastalık riskini artırabilir. PKOS'un tanısı, hormonal testler ve ultrasonografi gibi tıbbi incelemelerle konur. Tedavi, semptomlara ve hastanın özel ihtiyaçlarına göre şekillenir ve kilo kontrolü, hormonal düzenleme ve infertilite tedavisi gibi farklı yaklaşımları içerebilir. PKOS, doğru tanı ve tedavi ile yönetilebilen bir durumdur, ancak uzun vadeli sağlık sonuçlarını etkileyebilme potansiyeline sahiptir (19).

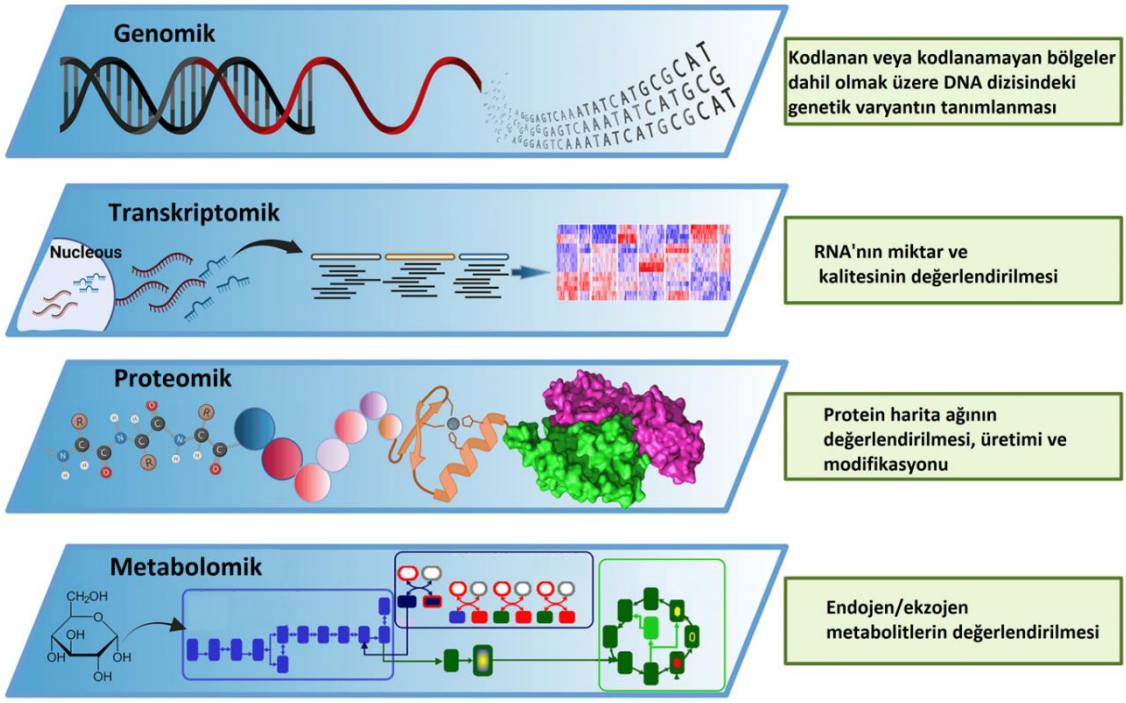


Şekil 2.1. Normal ve Polikistik Over Görünümü

2.2. Biyoinformatik ve Omik Teknolojiler

Biyoinformatik, biyolojik verilerin toplanması, analizi, yorumlanması ve depolanması için bilgisayar bilimleri ve istatistik gibi bilişim araçlarının biyoloji ve tıp alanlarında kullanılmasını sağlayan bir disiplindir. Bu disiplin, genetik bilgi, protein yapısı, evrimsel ilişkiler, hastalıkların nedenleri ve tedavileri gibi biyolojik konuları anlamak ve çözmek amacıyla büyük miktardaki biyolojik verileri işlemek için geliştirilen yöntemlerin tasarlanması ve uygulanmasını içerir. Biyoinformatik, genomik, proteomik, metabolomik ve diğer "omik" verilerin analiziyle ilgilenir ve bu sayede yeni ilaçların keşfi, hastalıkların teşhisi, türlerin evrimsel kökeni gibi birçok biyolojik sorunun çözümünde büyük öneme sahiptir (20).

Biyoinformatik, ML, AI, matematiksel modelleme ve istatistik gibi çok çeşitli disiplinlerle entegre çalışır. Bu sayede biyologlar ve tıp uzmanları, büyük veri setlerini analiz ederek biyolojik sorulara daha hızlı ve etkili bir şekilde yanıt bulabilirler. Bu alandaki gelişmeler, tıp, tarım, çevre bilimleri ve biyoteknoloji gibi birçok sektörde önemli uygulamalar sunmaktadır.



Şekil 2.2. Omik teknolojileri

2.2.1. Genomik

Genomik, organizmaların tüm genetik materyalini, yani Deoksiribonükleik Asit (DNA) veya Ribonükleik Asit'i (RNA) inceleyen ve anlamaya çalışan bir multidisipliner

bilim dalıdır. Genom, bir organizmanın tüm genetik bilgisini içerir ve genlerin sıralanması, organizmanın genetik özelliklerini belirleyen bilgiyi taşır. Genomik, bu genetik bilgiyi analiz etme, yorumlama, karşılaştırma ve organizmanın genetik yapısı ile fenotipik özellikleri arasındaki ilişkiyi anlama sürecini içerir. Genomik araştırmalar, organizmaların evrimini, genetik çeşitliliği, genlerin işlevlerini ve hastalıkların genetik nedenlerini anlamada kritik bir rol oynar (21).

Genomik çalışmaları, yüksek hızlı DNA dizileme teknolojilerinin gelişimi ile büyük bir ivme kazanmıştır. Bu teknolojiler sayesinde birçok organizmanın genomu hızlı ve maliyet etkin bir şekilde sekanslanabilir hale gelmiştir. Genomik araştırmalar, insan genomunun çözülmesi, hastalık genetiklerinin anlaşılması, türlerin evrimsel ilişkilerinin belirlenmesi ve tarım alanında bitki genomları üzerinde çalışmalar yapılmasını sağlayarak birçok farklı alanda önemli bilimsel ve uygulamalı ilerlemeleri tetiklemiştir (21).

2.2.2. Transkriptomik

Transkriptomik, bir organizmanın hücresel seviyede gen ifadesini inceleyen bir araştırma alanıdır. Gen ifadesi, bir organizmanın genetik bilgisinin DNA'dan RNA'ya çevrilmesi ve ardından bu RNA'nın protein üretimi için kullanılmasını içerir. Transkriptomik, bir hücrenin veya dokunun belirli bir zamanda hangi genlerin ifade edildiğini, ne kadar ifade edildiğini ve hangi koşullarda bu ifadenin değişebileceğini anlamaya yönelik çalışmaları kapsar. Bu nedenle transkriptomik, bir organizmanın genetik bilgisini nasıl uyguladığını ve düzenlediğini aydınlatarak, biyolojik süreçlerin ve organizma gelişiminin daha iyi anlaşılmasına katkı sağlar (22).

Transkriptomik çalışmaları genellikle yüksek veri kapasitesine sahip tekniklerle gerçekleştirilir, bu da genomik verilerin analizi için önemlidir. RNA dizileme (RNA-seq) gibi teknikler, bir organizmanın transkriptomunu belirli bir zamanda ayrıntılı bir şekilde inceleme olanağı sunar. Bu tür analizler, gen ifadesi düzenlemesi, hastalık mekanizmaları, çevresel stres yanıtları ve biyolojik gelişim süreçleri gibi bir dizi biyolojik sorunu araştırmak için kullanılır.

Transkriptomik, biyotıp, tıp, bitki bilimi, çevre bilimi ve daha birçok disiplinde uygulama alanlarına sahiptir ve organizmanın hücresel işleyişini anlama ve iyileştirme amacı güden birçok araştırma projesinin temelini oluşturur (22).

2.2.3. Proteomik

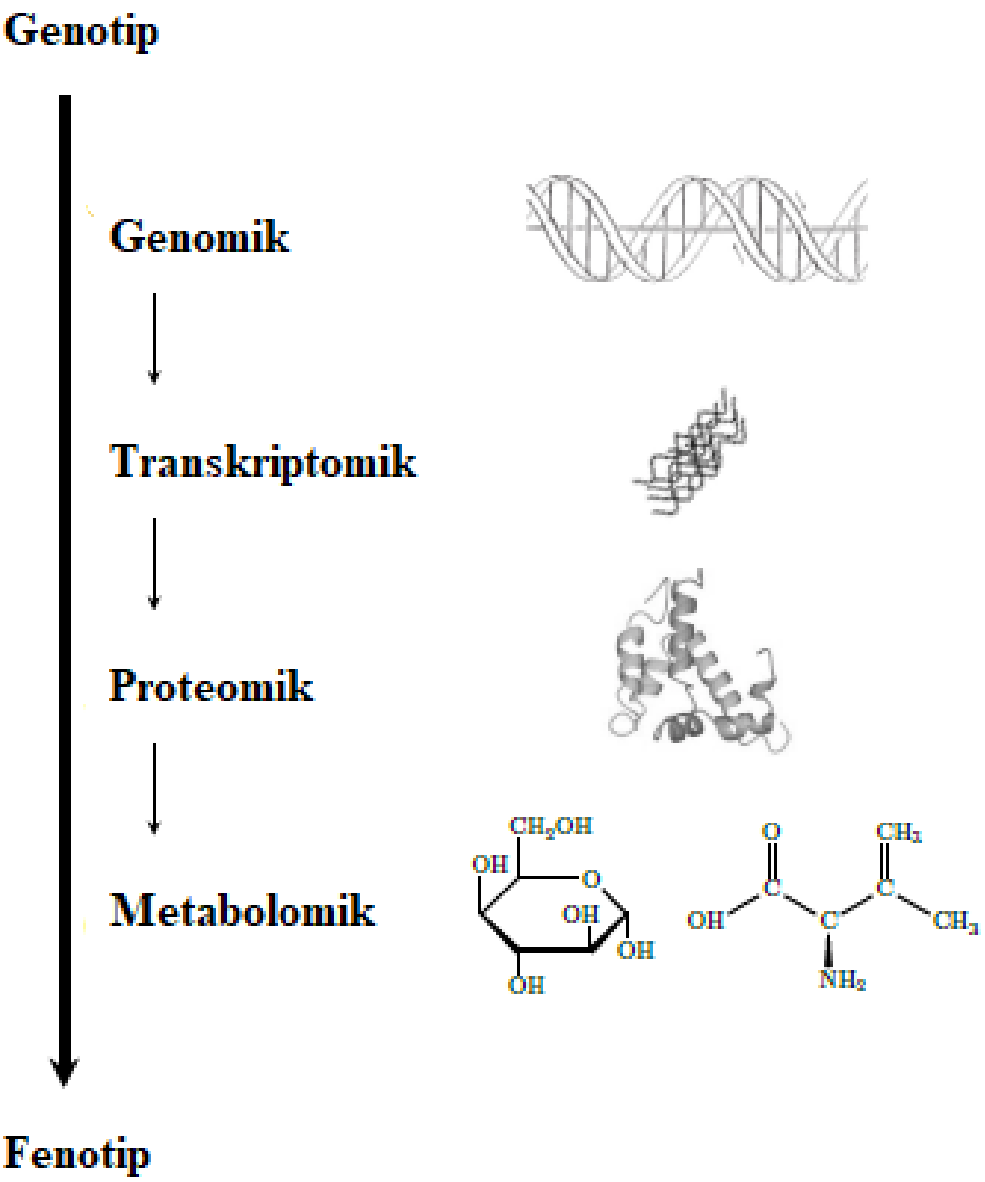
Proteomik, bir organizmanın ya da bir hücrenin belli bir zamandaki tüm proteinlerini anlama ve karakterize etme amacı taşıyan bir multidisipliner bilim dalıdır.

Genler tarafından kodlanan proteinler, bir organizmanın işlevsel yapı taşlarıdır ve biyokimyasal süreçlerde önemli roller üstlenirler. Proteomik, bu proteinlerin türleri, miktarları, işlevleri, etkileşimleri ve düzenlemeleri üzerine odaklanarak, organizma veya hücre seviyesindeki biyolojik süreçlerin daha iyi anlaşılmasına ve hastalıkların mekanizmalarının açıklığa kavuşturulmasına yardımcı olur. Genomikle birlikte organizmaların karmaşık biyolojik yapısının daha kapsamlı bir anlayışını sağlar (23).

Proteomik çalışmaları, yüksek teknolojiye sahip araçlar kullanarak proteinlerin belirlenmesi, miktarlarının tayini, işlevlerinin anlaşılması ve proteinler arası etkileşimlerin araştırılması gibi bir dizi analitik yöntemi içerir. Bu yöntemler, hücresel işleyişin, hastalıkların nedenleri ve tedavileri, ilaç geliştirme, biyolojik süreçlerin anlaşılması gibi birçok alanda uygulanır. Proteomik, biyomedikal araştırmadan bitki bilimine, mikrobiyolojiden çevre bilimine kadar birçok farklı alanda geniş bir uygulama yelpazesi sunar ve biyolojik bilimlerdeki temel araştırma ve uygulamalarda önemli bir rol oynar (23).

2.2.4. Metabolomik

Metabolomik alanı, metabolitler olarak adlandırılan hücresel süreçlerin son ürünlerini incelemektedir. Bu metabolitler, canlı organizmalarda çok sayıda işlevi yerine getiren, kabaca 50 ila 1500 atomik kütle birimi olan küçük moleküllerdir. Metabolitler besinlerin biyotransformasyonu nedeniyle tükürükte bulunurlar, hücrelerde enerji üretimi (ATP) için gereklidirler ve gastrointestinal sistemde dışkı maddesinin oluşumuna ve bozulmasına katkı sağlarlar (24). Metabolomiklerin canlı organizmalardaki yerini daha iyi anlamak için genotip-fenotip sürekliliğini tanıtmak gerekir (Şekil 2.3).



Şekil 2.3. Genotip-fenotip sürekliliği (25)

Genotip, canlı bir organizmada bulunan kalıtsal genetik materyalin koleksiyonudur. Fenotip ise bir organizmanın mevcut fiziksel durumunu temsil eder. Genotipin sabit olduğu varsayılırken fenotip zaman içinde gelişir ve hem dahili (genler, yaş vb.) hem de harici faktörlerden (hastalık, güneş ışığı, stres vb.) etkilenir. Metabolomik en çok fenotiple ilgilidir ve genler tarafından kodlanan proteinlerin aktivitesinin son ürününü temsil eder. Bundan dolayı spesifik fizyolojik tepkilerin ve bunların potansiyel biyobelirteçlerinin anlaşılmasına izin verir (26).

Metabolomik boru hattı veya iş akışı veri işleme ve veri analizi bölümünün takip ettiği deneysel bir bölümden oluşur. Metabolomikte örnek hazırlama, kültürlerin oluşturulması ve kuruluma bağlı olarak belirli bir sırada çok sayıda örneğin çıkarılması

gerektiğinden genellikle günler ila haftalar alır. Bu numune hazırlığının ardından kimyasal analiz yapılır. En yaygın olarak kullanılan teknikler, çoğunlukla sıvı ile kombinasyon halinde nükleer manyetik rezonans (NMR) spektroskopisi ve sıvı kromatografisi kütle spektrometrisidir (LS-MS). Bu analitik çalışmaların herhangi bir aşamasındaki küçük deneysel aksaklıklar, tüm alt analiz üzerinde yıkıcı etkilere sahip olabilir. Veri işleme ve analiz kısmından sonra ilgili metabolitler belirlenir ve sonuçlar yorumlanır. Deneye ve mevcut verilere bağlı olarak metabolit belirleme adımı, kütle spektrumlarının, ürün iyon spektrumlarının, alıkonma sürelerinin, kimyasal kaymaların ve hatta ek analitik deneylerin kullanımını içerir.

Metabolomikte Kullanılan Deneysel Teknikler

Aşağıdaki bölümde, metabolomikte en sık kullanılan analitik teknikler, özellikle LS-MS ve NMR hakkında kısa bir genel bakış sunulmaktadır. Çok sayıda başka olası teknik olmasına rağmen, LC-MS ve NMR en sık kullanılanlardır. Bu tekniklerin işleyişine veya deneylerin nasıl gerçekleştirileceğine dair kapsamlı bir inceleme bu tezin kapsamı dışındadır ve diğer çalışmalarda bulunabilir (27). Bu tez, bu deneysel teknikler tarafından üretilen verilere ve özellikle veri işleme ve analiz yönlerine odaklanmaktadır.

Nükleer Manyetik Rezonans (NMR) Spektroskopisi: NMR spektroskopisi, kütle spektrometrisi ile birlikte metabolomikte en çok kullanılan analitik yaklaşımlardan biridir. NMR, numune içindeki metabolitlerin yapısı hakkında bilgi içeren her numune için bir spektrum üretir. Kısaca, numuneler harici bir manyetik alana yerleştirilir ve daha sonra bir elektromanyetik darbe ile uyarılır. Atomlar temel duruma dönerken elektromanyetik radyasyon yayarlar, bu serbest indüksiyon bozunum sinyalidir. Bu sinyalin Fourier dönüşümü, nihai NMR spektrumuyla sonuçlanır. 1H-NMR spektroskopisinin altında yatan sürecin sezgisel olarak anlaşılması, hidrojen atomlarına ne olduğu dikkate alınarak elde edilebilir. Numune harici bir manyetik alana yerleştirilir. Bu, 1H atomlarının nükleer spin durumlarının harici manyetik alanla hizalanmasına neden olur. Hizalama dış alanla paralel veya anti-paralel olarak iki durumda olabilir. Bu iki durum arasında bir enerji farkı vardır ve çoğu çekirdek daha düşük enerji durumunda olmayı tercih eder. Bununla birlikte, enerji ekleyerek (yani harici elektromanyetik darbe), çekirdekler daha yüksek enerji durumuna uyarılabilir. Daha düşük enerji durumuna geri dönerken bu çekirdekler elektromanyetik radyasyon yayarlar. NMR spektroskopisinin altında yatan temel kavram, bu iki seviye arasındaki enerji farkının manyetik alanın gücü tarafından belirlenmesidir.

Tüm atomlar aynı dış alana yerleştirilse de bir molekül içindeki komşu atomlar dış alandan koruma sağladığı için yerel manyetik alan her atom için farklıdır. Bundan dolayı temel durum ile uyarılmış durum arasındaki enerji farkı atomların çevresine bağlıdır (28). Sonuçta ortaya çıkan spektrumlar numunede bulunan atomları çevreleyen farklı bağlar hakkında bilgi sağlar. Bu bilgi metabolik biyobelirteç keşif deneyleri ile birlikte değerli olan yapısal özellikleri elde etmek için kullanılabilir (29).

Kütle Spektrometrisi (MS): MS, numunedeki moleküllerden oluşan iyonların kütle-yük (m/z) oranını ölçebilen bir yaklaşımdır. Metabolomikte birçok farklı MS türü kullanılır ve m/z tespitinin altında yatan farklı fiziksel prensipler vardır (30). Örneğin, dört kutuplu bir MS, yalnızca belirli m/z değerlerine sahip iyonların cihaz boyunca kararlı bir yörüngeye sahip olmasına izin vermek için dört çubuk kullanır (31). Diğer bir örnek, iyonların cihaz boyunca kabaca aynı yörüngeler sergilediği ancak farklı hızlar elde ettiği uçuş süresi (TOF) kütle spektrometresidir. İyonlar bir potansiyelde hızlandırılır ve daha sonra bir detektör plakasına çarpmadan önce bir serbest uçuş periyodu sergiler. Daha hafif iyonlar daha ağır iyonlara kıyasla dedektörü daha çabuk vurur (32). Orbitrap (33) veya Fourier dönüşümü iyon siklotron rezonans (FTICR) (34) cihazları gibi başka MS teknikleri de mevcuttur. Bu tür spektrometreler iyonların birbirine çarpması prensibine dayalı değildir ve iyonların hareketleriyle indüklenen elektromanyetik sinyali kullanır. Sonuç olarak tüm iyonların aynı anda ölçülmesi ve algılama süresi artırılarak sinyal gürültü oranı arttırılabilmektedir. Son spektrum, tespit edilen sinyalin Fourier dönüşümünden sonra elde edilir. Her yöntemin belirli avantajları, dezavantajları ve performans özellikleri (çözümleme gücü, kayıt hızı, maliyet, vb.) vardır. Genel olarak, Orbitrap ve FTICR teknikleri en yüksek çözümlemeye sahiptir ve FTICR Orbitrap'ten üstün bir tekniktir. Bununla birlikte, TOF'ler ve dört kutuplularla birlikte Orbitrap aletleri, sıvı kromatografi gibi bir önceki ayırma adımına daha kolay bir şekilde bağlanabilir. Aşağıdaki başlıklarda, MS metabolomik deneylerinde rutin olarak kullanılan iki ek teknik sunulmaktadır. İlk olarak, tespitten önce bir ayırma tekniği olarak kullanılan MS ve sonra MS'den gelen numuneyi kütle analizine hazır iyonlara dönüştürmek için kullanılan elektrosprey iyonizasyonu (ESI) teknikleri sunulmaktadır.

Sıvı Kromatografisi (LC)

LC, genellikle doğrudan MS analizinden önce uygulanan bir ayırma adımıdır. Yüzyılı aşkın bir süredir farklı formlarda mevcuttur ve mevcut haliyle bir numune ve çözücü (mobil faz), 100 ila 300 bar basınçlarda bir kolondan pompalanır (35).

Numunedeki bileşikler, hareketli ve durağan faz ile etkileşime bağlı olarak farklı zaman noktalarında ayrışır. Farklı bileşiklerin elüsyon süreleri (RT), hidrofobik (örneğin polar olmayan moleküller) ve hidrojen bağı vb. gibi bir dizi olası etkileşimden etkilenir. Farklı LC teknikleri, belirli etkileşimleri desteklemek için belirli durağan ve hareketli faz kombinasyonları kullanır (36). Örneğin, ters faz LC'de durağan faz hidrofobiktir, dolayısıyla polar bileşikler apolar bileşiklerden daha hızlı ayrışacaktır (37). Normal fazlı LC'de durağan faz hidrofiliktir, dolayısıyla polar bileşiklerin durağan faz için yüksek bir afinitesi (daha yavaş elüsyon) ve apolar bileşiklerin mobil faz için bir afinitesi olacaktır (daha hızlı elüsyon). Bir LC adımının MS'ye bağlanması, verilere fazladan bir boyut eklenmesiyle sonuçlanır. Bileşikler, RT ve kütleye (m/z) göre ayrılır ve bu durum MS'ne aynı anda giren daha az bileşik olduğu için verilere ek çözünürlük sağlar. Sonuç olarak daha temiz ve daha kaliteli spektrumlar elde edilir. Tekniğin dezavantajı, veri boyutu arttıkça tepe algılama yöntemlerinin çok boyutlu verileri işlemesi gerektiği için veri işlemeye ekstra zaman harcanması gerekmesidir (38, 39).

Elektrosprey İyonlaştırma (ESI): ESI, LC cihazlarının MS ile kolayca birleştirilmesine izin vererek metabolomik alanında gerekli bir teknolojik ilerleme sağladı. Örneğin bir LC kolonundan gelen numune, yüklü bir kılcal iğne içinden geçer. İğneden çıkarken sıvı, iğneye uygulanan elektrik potansiyeli ile yüklenir ve ardından küçük damlacıklar halinde dağılır. Bu dağılım, damlacıkların yüzeyinde bulunan Coulomb kuvvetlerinden kaynaklanır. Yüklü damlacıklar daha fazla buharlaşır ve yüklü moleküllerle sonuçlanır. ESI'nın başlıca avantajları, kırılgan moleküllerde parçalanmaya neden olan aşırı enerji eklenmeden çözeltideki bir numunenin gaz fazındaki bozulmamış iyonlara doğrudan dönüştürülmesine izin vermesidir (40, 41).

Metabolomiklerin Sistem Biyolojisindeki Rolü

Sistem biyolojisi, biyolojik bir sistemin tek tek parçalarına odaklanmak yerine bir bütün olarak karakterize etmeyi amaçlayan bir araştırma alanıdır (42). Bu araştırma alanı genler, transkriptler, proteinler, metabolitler ve lipitler gibi moleküler bileşenlerin kapsamlı analizine odaklanmaktadır. Bir organizmanın genomunun şifresini çözmek ve RNA molekülleri aracılığıyla protein seviyelerine kadar takip edebilmek biyolojik işlevi ve düzenlenmesini anlamak için gerekli ilk adımdı (43). Bugüne kadar DNA ve RNA analizlerine yüksek verimli sıralama teknolojileri hakimdi (44). Genler, transkriptler ve proteinler hiyerarşik olarak birbirine bağlıdır. Hepsi bir DNA dizisinden elde edilir ve bu nedenle çeşitlilikleri genom tarafından belirlenir. Ancak bu, yalnızca işlevi belirleyen bilgi olan dizileriyle ilgilidir.

Tüm hücresel bölümlerin asıl yapı taşları, çevreden alınan besinlerdir. Genom kodlu enzimler, değişken besin maddelerinin, tüm hücreyi büyütmek ve çoğaltmak için gerekli olan yüzlerce yapı taşına dönüştürülmesinde temel bir rol oynar. Bir hücrede meydana gelen tüm kimyasal bileşiklerin topluluğu, tüm metabolitlerin kümesi olan metabolomdur. Kavramsal olarak, metabolom, çevrenin etkileşiminden ve genom tarafından proteom aracılığıyla belirlenen hücre içi konfigürasyonun sonucudur. Gen ifadesindeki veya ortamdaki tüm değişiklikler, metabolik seviyede birleşecek ve potansiyel olarak yeni bir duruma neden olacaktır. Değişen koşullar veya stres karşısında zindeliği sağlamak için hücreler, metabolit seviyelerini algılayan ve hücresel veya metabolik süreçleri en iyi şekilde ayarlamak için uyarlanabilir programları tetikleyen düzenleyici mekanizmalar geliştirmiştir. Bu mekanizmalar metabolik değişikliklere yansır ve metabolomun ölçülmesi hücresel durum hakkında benzersiz bilgiler sağlar.

2.3. Yapay Zekâ ve Makine Öğrenimi

AI, makinelerin genellikle insan düzeyinde zekâ gerektiren öğrenme ve problem çözme gibi bilişsel işlevleri taklit etmek için programlanmış yeteneğidir (45). AI önemli bilgileri arama, optimizasyon ve mantıksal akıl yürütme amacıyla kullanılan istatistiksel öğrenme yöntemleri ve davranışa dayalı yaklaşımları içermektedir. Bilgi işlem gücündeki son gelişmeler, araştırma için mevcut verilerin niceliği ve kalitesindeki hızlı artışla birlikte, AI tekniklerinin çok çeşitli alanlarda hızla benimsenmesine yol açmıştır. AI ürün/süreç maliyeti azaltma, tutarlılık, tekrarlanabilirlik, iyileştirilmiş verimlilik, gelişmiş karar alma, yeni ürün ve hizmetlerin geliştirilmesine yardımcı olma gibi çeşitli amaçlarla kullanılabilmektedir. Bunun yanında AI teknolojisi, bilgisayarlarla çalışma şeklimizi dönüştürerek ekonomik büyümeyi hızlandırabilecek yeteneğe sahiptir (45).

ML, bilgisayar sistemlerinin verilerden öğrenme yapabilen algoritmalar kullanarak görevleri gerçekleştirmesini sağlayan bir AI dalıdır. ML, bir programın belirli bir görevi yerine getirmek için programcı tarafından doğrudan talimatlandırılmadan veriye dayalı deneyimlerden öğrenmesine odaklanır. ML, verilerin incelenmesi ve desenlerin tanımlanması için istatistiksel ve matematiksel teknikleri kullanır. Bu desenleri kullanarak, ML algoritmaları tahminler yapabilir, kararlar alabilir, sınıflandırma yapabilir veya veri setlerini analiz edebilir. ML algoritmaları genellikle etiketli (labelled) veya etiketsiz (unlabelled) veri kümelerini kullanır. Etiketli veri kümeleri, örnek verilerin hedef çıktılarıyla birlikte sunulduğu veri setleridir. Örneğin, görüntü tanıma için bir ML algoritması, görüntülerin etiketlendiği bir veri kümesiyle eğitilebilir. Etiketsiz veri kümeleri ise hedef çıktı etiketleri olmadan sadece girdi verilerini içerir.

Bu tür veri kümeleriyle, algoritmalar veri setindeki yapıyı ve desenleri keşfedebilir. ML, farklı yaklaşımlara sahip birçok algoritma ve yöntem içerir. Bunlar arasında doğrusal regresyon, lojistik regresyon, karar ağaçları, destek vektör makineleri, yapay sinir ağları ve derin öğrenme gibi yöntemler bulunur. Bu algoritmalar, veri setlerindeki desenleri öğrenerek, yeni verilere dayalı olarak tahminler yapabilir veya kararlar verebilir. ML, birçok alanda geniş bir uygulama potansiyeline sahiptir. Görüntü ve nesne tanıma, doğal dil işleme, konuşma tanıma, spam filtreleme, pazar segmentasyonu, finansal tahminler ve tıbbi teşhis gibi alanlarda kullanılabilir. ML, büyük miktarda verinin olduğu karmaşık problemleri çözmede ve karar verme süreçlerini otomatikleştirmede büyük bir rol oynamaktadır (45-48).

Son yıllarda derin öğrenme, rastgele ormanlar, destek vektör makineleri (SVM'ler) vb. gibi karmaşık ML algoritmalarının kullanılması, şeffaflık eksikliğine ve dolayısıyla "kara kutu" modellerinin oluşmasına neden olabilir (45, 46). Bu şeffaflık eksikliği, derin öğrenmeye veya karmaşık ML algoritmalarına özgü değildir ve lojistik regresyon, karar ağaçları veya kural tabanlı algoritmalar gibi diğer yorumlanabilir ML algoritmalarının da yüksek boyutlu girdiler için yorumlanması çok zor olabilir (47). ML modellerinin yorumlanmasındaki zorluk, modelin güvenilirliğinin garanti altına alınması gereken mali, tıbbi, adalet vb. uygulamalarda ML sistemlerinin benimsenmesinde büyük bir engel olabilir. Bu engelleri azaltan teknolojilerin geliştirilmesine ihtiyaç olduğu konusunda artan bir fikir birliği vardır (48). ML'de şeffaflığın olmaması, XAI teknikleri geliştirmeye odaklanan yeni bir araştırma alanına yol açmıştır (49). Sonraki bölümde XAI teknikleri açıklanmıştır.

2.4. Açıklanabilir Yapay Zekâ

XAI veya şeffaf/yorumlanabilir AI kullanıcıları tarafından açıkça anlaşılabilen eylemler için açıklamalar veya nedenler üreten bir AI olarak tanımlanabilir. Açıklanabilirlik, AI sistemlerinin belirli bir tahminle ilgili tüm faktörlerin denetlenebilir bir kaydını sağlayarak daha fazla şeffaflık ve yorumlanabilirlik sağlamasına izin vermektedir (50). Açıklanabilirlik ayrıca, işletmelerin uygunluk gereksinimlerini karşılamasını sağlarken algoritmik karar vermenin adil ve etik olmasını sağlamaktadır. Bununla birlikte, modelin karar verme sürecini anlamak çok önemlidir. Tahminlerin güvenilirliğini değerlendirmek için hangi parametrelerin onları etkilediğini bilmek gerekli olabilir. XAI'nin hedeflerinden biri, ML'nin mevcut kara kutu modellerine ilişkin yeni içgörüler sağlamayı vaat eden yenilikçi açıklayıcı algoritmalar geliştirmek ve böylece kullanıcının AI sistemini daha iyi anlamasına ve ona güvenmesine yardımcı

olmaktır (45). XAI, aşağıdaki konulara katkı sağlar ve çeşitli araştırma alanlarında kullanılmaktadır (51).

Şeffaflık: AI modelleri karar vermeyi desteklemek için kullanılır ve bu nedenle, karar verme sürecine dâhil olan tüm tarafların ve bu kararlardan etkilenen kişilerin bunları anlamasını sağlamak hayati önem taşımaktadır.

Adillik: Açıklanabilirlik, sonucu etkileyen ilişkilerin net bir şekilde anlaşılmasını sağlar ve verilerdeki herhangi bir yanlılığı vurgulayarak modelin adil bir şekilde analiz edilmesini sağlar. Bu nedenle, XAI, AI modellerinin haksız veya etik dışı kullanımından kaçınılmasını sağlar.

Etkileşim: Model etkileşimli olmalıdır, yani son kullanıcılara başarıyı garantilemek için modeli geliştirir ve modelle etkileşim kurma olanağı sağlar.

Güven: Bir AI sistemi tarafından sağlanan açıklamalar, AI sisteminin güvenilirliğinin değerlendirilmesine ve böylece sistemin güven kazanmasına izin verir.

Nedensellik: Bir AI modeli verilerden öğrendiğinden, bu veriler arasındaki korelasyonları keşfedebilir. Bu, nedeni ortaya çıkarmak için yeterli olmasa da etki ilişkileri, girdiler ve çıktılar arasındaki nedensel ilişkilerin ilk sezgisini sağlayabilir.

Genel olarak XAI yöntemleri yorumlanabilir yöntemler, model-agnostik açıklamalar ve örnek tabanlı açıklamalar olarak üçe ayrılmaktadır.

2.4.1. Yorumlanabilir Yöntemler

Yapısal Yorumlanabilir yöntemler kural tabanlı, doğrusal regresyon, ağaç tabanlı modeller, lojistik regresyon, Naive Bayes ve hatta K-en yakın komşu (KNN) gibi geleneksel parametrik yaklaşımları içermektedir. Doğrusal regresyonda, tahmin edilen hedef, girdi değişkenlerinin ağırlıklı bir toplamından oluşur. Bu nedenle, doğrusal bir denklemin katsayısı bir tahmini açıklamanın bir yolu olarak kullanılabilir. Lojistik regresyon, sınıflandırma problemini dikkate alan ve sınıflandırma problemleri için olasılıkları modelleyen doğrusal regresyonun bir uzantısıdır. Doğrusal regresyon ile lojistik regresyon arasındaki temel fark, lojistik regresyonun 0 ile 1 arasında bir olasılık sağlamasıdır. Bununla birlikte, regresyon katsayıları etkinin ne kadar güçlü olduğu ve yönü (negatif veya pozitif) hakkında bilgi sağlar. Karar ağacı modelleri, son bir düğüme ulaşılana kadar her düğümde bir kesme eşiği kullanarak verileri birkaç kez böler. Karar ağacı modellerinin temel avantajı, girdiler ve çıktılar arasında doğrusal bir ilişki olmadığında bile çalışmasıdır. Ancak bu yöntem, girdiler ve çıktılar arasında doğrusal bir ilişki ortaya koymaz. Ayrıca, girdi verilerindeki küçük değişiklikler bazen tahmin edilen sonuç üzerinde çok fazla etkiye sahiptir.

Ağaç ne kadar çok düğüme sahip olursa yorumlaması o kadar zordur. Başka bir yaklaşım da karar kurallarıdır (basit IF-THEN-ELSE koşulları). (IF-THEN) kuralları genellikle sınıflandırma problemini çözemez ve doğrusal ilişkileri yorumlarken yanlış sonuçlar verir (52). Örneğin, RuleFit algoritması seyrek doğrusal modelleri dikkate alır. Algoritma karar kuralları biçimindeki etkileşim etkilerini tespit edebilir ve sonuç olarak dâhili yorumlamaya izin verir. Aynı zamanda orijinal değişkenlerin yanı sıra bazı yeni değişkenleri de öğrenir ve bu durum genellikle değişken sayısı arttıkça yorumlanabilirliği engellemektedir (53). Doğrusal modellerin bazı varsayımlarının dikkate alınmasına izin veren genelleştirilmiş toplam modeller (GAM) ve genelleştirilmiş doğrusal modeller (GLM) yorumlanabilir yöntemlerin başka örnekleridir (54, 55).

Bu yöntemler ana yeteneklerine göre şu şekilde sınıflandırılabilir:

- Doğrusallık: Hedef ile değişkenler arasında doğrusal bir ilişki varsa kullanılır.
- Monotonluk: Hedef ile değişken arasındaki ilişkinin, özelliğe bağlı olarak (tüm değer aralığı boyunca) her zaman tek bir tutarlı yöne (artan veya azalan) sahip olduğunu belirler.
- Etkileşim: Doğrusal olmayan değişkenler ve etkileşim, modele her zaman manuel değişken geliştirme yoluyla eklenebilir.

Yukarıda belirtilen hususlar bazı yorumlanabilir yöntemler için Tablo 2.1'de özetlenmiştir (56).

Tablo 2.1. Yorumlanabilir yöntemlerin özellikleri

Algoritma	Doğrusallık	Monotonluk	Etkileşim	Görev
Doğrusal modeller	Evet	Evet	Hayır	Regresyon
Lojistik regresyon	Hayır	Evet	Hayır	Sınıflama
Karar ağaçları	Hayır	Bazen	Evet	Sınıflama
Rulefit	Evet	Hayır	Evet	Sınıflama+Regresyon
Naive Bayes	Evet	Evet	Hayır	Sınıflama
K en yakın komşu	Hayır	Hayır	Hayır	Sınıflama+Regresyon

2.4.2. Model-Agnostik Açıklamalar

Modelden bağımsız yöntemler, açıklama yaklaşımının çeşitli modellerle kullanılabilmesi için açıklamayı ML'den ayırır. Bu ayrımın birçok net faydası vardır:

- Yorumlama yöntemi, çeşitli ML modelleriyle çalışabilir.
- Belirli bir model için farklı yorumlanabilirlik sonuçları sağlanabilir (örneğin, değişkenlerin öneminin görselleştirilmesi) (57).

Kısmi Bağımlılık Grafiği (PDP): ML modellerinin sonucu üzerindeki bir veya iki değişkenin (3-D'de en fazla üç değişken) marjinal etkisi, PDP ile incelenebilmektedir. Modelin genel davranışını ve hedef ile nesne(ler) arasındaki doğrusal veya karmaşık ilişkileri gösterebildiği için küresel bir yöntemdir. PDP yöntemi bireysel işlevlerdeki değişikliklerin bir sonucu olarak tahmindeki değişiklikleri görüntüler ve net bir nedensel yorum sağlar. PDP yöntemi, öğelerin birbiriyle ilişkisiz olduğu varsayımı altında çalışmaktadır. Yöntem tüm nesnelerin tek bir örnek için etkisinden ziyade, nesnenin/nesnelerin tahmin üzerindeki ortalama etkisini göstermektedir (53).

Bireysel Koşullu Beklenti (ICE) Grafiği: ICE grafiği hedef işlev ile ilgilenilen bir değişken arasındaki bağımlılığı göstermek için kullanılmaktadır. Bununla birlikte, ilgilenilen değişkenlerin ortalama etkisini gösteren PDP grafiklerinden farklı olarak, ICE grafikleri, tahminin bir değişkene olan bağımlılığını her örnek için ayrı ayrı, örnek başına bir satırla görselleştirir. ICE grafikleri için yalnızca bir ilgilenilen değişken desteklenir (58).

Birikmiş Yerel Etkiler (ALE) Grafiği: ALE değişkenlerin bir ML modelinin tahminini ortalama olarak nasıl etkilediğini açıklayan bir yöntemdir. ALE grafikleri ilişkili değişkenlerle oldukça iyi çalışır ve PDP'den daha hızlıdır. ALE grafikleri, ilişkili verilere yönelik yansız olsa da, değişkenlerin güçlü bir şekilde ilişkili olduğunda tahmindeki değişiklikleri ayırt etmek zorlaşmaktadır. Bu senaryoda, tahmindeki değişiklikleri anlamak için yalnızca her iki bağlantılı değişkendeki değişiklikleri aynı anda gösteren grafikler incelenmelidir (59).

Değişken Etkileşimi: Değişken etkileşimi, bir tahmin görevi sırasında birbirini etkileyen iki veya daha fazla değişkeni içeren bir yöntem olarak tanımlanabilir. Bu yöntem bir tahmin modelindeki değişkenler arasındaki etkileri, model performansının bir noktası (doğruluk veya hata) olarak ölçebilir. Ayrıca, bir değişkinin rolünü ve bir tahmin modelindeki değişkenler arasındaki etkileşimi anlamak için kullanışlıdır. Model performansını iyileştirmek için de kullanılabilir. Örneğin, bir değişkenin önemi düşük ve negatif etkileşimi yüksekse, onu tahmin modelinden çıkarmak modelin performansını arttırabilir (60).

Küresel Vekil Model: Küresel vekil model, kara kutu modelinin genel davranışını simüle etmeyi amaçlayan bir yaklaşımdır.

Başka bir deyişle, vekil modeller, kara kutu modelinin tahmin işlevini mümkün olduğu kadar yakından kopyalamaya çalışır. Bir üst model, bir yanıt yüzeyi modeli veya bir yaklaşım modeli küresel vekil modeli tanımlamak için kullanılan terimlerdir. Modelin dezavantajı, aynı kara kutu yaklaşımı için farklı mimarilere sahip karar ağaçları gibi birkaç farklı açıklama üretmesidir (61).

Shapley Değerleri: Shapley bu değeri 1953'te tanıtmıştır (62). Yöntem ilgili oyuncular arasında işlevlerin önemini adil bir şekilde dağıtmaya yardımcı olan işbirlikçi oyunlar teorisine dayanır. Her örnek işlev değerinin oyundaki bir oyuncu olduğu varsayılır ve tahmin, toplam ödemeye (yani tahmin) katkılarına bağlı olarak oyuncular (yani işlevler) arasında dağıtılan toplam ödemedir. Çok sayıda nesne olduğunda hesaplamalar zaman alır ve işlemler maliyetli hale gelir. Shapley değeri tüm açıklamaları sağlayan teoriye dayalı bir açıklama tekniğidir ve her nesne için tek bir değer döndürür.

Yöntemin dezavantajı, veri kümesinde ilişkili değişkenler olduğunda sonuçların yorumlanmasının zor olmasıdır (62).

2.4.3. Örnek Tabanlı Açıklamalar

Örnek tabanlı açıklama yaklaşımları, veri kümesindeki bireysel durumları kullanarak model davranışını ve veri dağıtımını agnostik olarak tanımlar. Açıklamalar, "X, Y ile aynıdır ve Y, Z'ye neden olur, bu nedenle tahmin, X'in Z'ye neden olacağını gösterir" şeklindedir. Örnek tabanlı bir açıklamaya örnek olan çekişmeli yöntem, tahminde/çıktıda önemli bir değişiklikle sonuçlanacak girdi değişikliklerini tanımlar. Bireysel tahminler için kullanılan bir diğer yöntem karşı olgusal açıklamalardır. Yöntem, "A olmasaydı, B olmazdı" gibi bir durumun doğrudan ilişkisini detaylandıran bir açıklama vermektedir. Buna rağmen karşı-olgusal açıklamalar her örnek düzeyinde tahmin için çok sayıda doğru açıklamalar (karşı olgular) üretir ve yöntemin zorluğu hangi açıklamanın en iyi olduğuna karar vermektir. Karşı olgusal yaklaşımlar, çalışmak için verilere veya modellere ihtiyaç duymaz ve ML kullanmayan bir sistemle bile çalışabilirler. Bu yaklaşımlar çok sayıda değere sahip kategorik değişkenler için etkisizdir (52).

Yerel Yorumlanabilir Model-Bağımsız Açıklamalar (LIME): LIME, ML modellerinin kararlarını açıklamak için kullanılan bir açıklanabilirlik yöntemidir. LIME, model-agnostik bir yaklaşım sunar, yani herhangi bir makine öğrenimi modeliyle kullanılabilir. Temel olarak, LIME, bir modelin belirli bir örneğe nasıl karar verdiğini anlamak için yerel bir açıklama sağlar. LIME, bir örnek etrafında bir "yakınlık" ölçütü kullanarak örneklem veri noktaları oluşturur.

Bu örneklem veri noktaları, orijinal veri noktasının etrafında yer alır ve modelin bu örnekler üzerinde nasıl davrandığını gösterir. Ardından, bu örneklem veri noktaları, bir anlamlılık ölçütüne dayanarak modelin davranışının nasıl değiştiğini belirlemek için kullanılır. LIME, bu örneklem veri noktalarını daha basit bir açıklanabilir model (genellikle doğrusal regresyon) üzerinde eğiterek sonuçları açıklar. Bu basit model, orijinal modelin yerel davranışını temsil etmeye çalışır. Bu şekilde, LIME, kararların ve sonuçların anlaşılabilir bir şekilde açıklanmasına yardımcı olur. LIME yaklaşımının tıp ve sağlık bilimlerinde klinik karar vermeyi açıklamak için nasıl kullanıldığına dair bazı örnekler ve bazı bilgiler aşağıdaki gibidir:

- LIME rasgele pertürbasyonunu değiştirmek için, Rehman ve arkadaşları (63) bir yığınsal hiyerarşik kümeleme (HC) ve KNN yöntemini kullandılar. Yazarlar, eğitim verilerini kümelemek için HC'yi ve örneği testin en yakın komşusunun altına yerleştirmek için KNN'yi kullanır. Araştırmacılar, bu yöntemin daha istikrarlı açıklama modelleri ürettiğini belirtmektedirler.
- QuadraticLIME (QLIME) olarak adlandırılan ikinci dereceden bir yaklaşım yapısı kullanan Bramhall ve arkadaşları (64), doğrusal LIME ilişkilerini doğrusal olmayan ilişkileri (QLIME) içerecek şekilde değiştirmişlerdir. Karmaşık fonksiyondaki doğrusal yaklaşımları hesaba katarak bu sonuçlara varırlar. Sonuçlar, QLIME kullanıldığında yerel düzeyde LIME lineer bağlantısının ortalama kare hatanın (MSE) arttığını göstermektedir.
- LIME için model tahmin alanı (MPS-LIME), Shea ve arkadaşları (65) tarafından resim verileri için süper piksel bilgilerinin seçilmesine alternatif bir yaklaşım olarak tanıtıldı. Yazarlar, süper pikselleri yönsüz bir grafiğe dönüştürerek olağan grafiği değiştirdiler. MPS-LIME yaklaşımında yanıt işlemi, etkilenen numunelerin sayısını azaltarak çalışma süresini iyileştirir. Yazarlar, tekniklerini Ortalama Mutlak değeri kullanan LIME ile karşılaştırdılar. Hata (MAE) ve R2 belirleme faktörü ve yöntemlerinin netlik, doğruluk ve güncellik açısından daha iyi sonuçlar ürettiğini buldu.

SHapley Katkı Açıklamaları (SHAP): SHAP, model tahminlerinin açıklanması ve değişken önemini belirlemek için kullanılan güçlü bir açıklanabilir yöntemdir. SHAP, Shapley değerleri adı verilen bir kavramı kullanarak, her bir özniteliğin katkısını nicel olarak hesaplar. Shapley değerleri, kooperatif oyun teorisi kavramlarından türetilmiştir. Bir model tahmininde yer alan her bir özniteliğin katkısını hesaplamak için tüm kombinasyonları dikkate alır.

Bu değerler, her değişkenin katkısını hakkaniyetli bir şekilde hesaplar. SHAP, bir modelin her bir tahminindeki her bir değişkenin Shapley değerlerini hesaplayarak açıklama sağlar. Bu değerler, her bir değişkenin tahmin sonucuna olan katkısını gösterir. Değişkenlerin katkıları, pozitif veya negatif olabilir ve sonuç üzerindeki etkilerini nicel olarak gösterir.

SHAP yaklaşımının tıp ve sağlık bilimlerinde klinik karar vermeyi açıklamak için nasıl kullanıldığına dair bazı örnekler ve bilgiler aşağıdaki gibidir:

- Anomalileri saptamak için kullanılan otomatik kodlayıcıları açıklamak için Antvarg ve arkadaşları (66) genişletilmiş bir SHAP tekniği geliştirdiler. Yazarlar, gerçek veri örneklerini yeniden yapılandırılmış çıktıyla karşılaştırarak anormallikleri kategorize etmek için otomatik kodlayıcıyı kullanır. Yazarlar, açıklamaların yeniden yapılanma hatasına dayanması gerektiğini, çünkü nihai ürünün yeniden yapılanma olduğunu düşünüyorlar. En verimli değişkenlere, katkıda bulunan ve telafi eden anormallikler olarak ayrılan SHAP değerleri atanır.
- Sundararajan ve meslektaşları (67), SHAP tekniğinin, belirli işlevlerin gerekli olmadığı durumlarda anlamsız açıklamalar üretme yeteneği de dâhil olmak üzere bir dizi kusuru olduğuna dikkat çekti.
- Aac ve arkadaşları (68), bağımlı değişkenleri barındıracak şekilde KernelSHAP tekniğini genişleterek SHAP değerleri arasındaki ilişkiyi incelediler. Yazarlar ayrıca, bağımlı değişkenler için Shapley küme değerlerini hesaplamak için bir yöntem de gösterdiler. KernelSHAP tekniği, deneysel bir yöntem ve Gaussian kullanarak KernelSHAP yönteminin koşullu dağılımlarını değiştirmek için önerilen dört yaklaşım kullanılarak kapsamlı bir şekilde araştırıldı.
- TreeExplainer adlı bir çerçevede, Lundberg ve arkadaşları (69), küresel modelin yapısının yerel açıklamalarla açıklandığı, ağaçlar için SHAP tekniğinin bir uzantısını geliştirdiler.
- Garcia ve arkadaşları (70) tarafından uzun süreli kısa süreli bellek (LSTM) ağlarını içeren zaman serisi sinyal tahminlerini açıklamak için SHAP tabanlı bir strateji geliştirilmiştir. Bir test setindeki münferit vakalar, eğitim setindeki en temel niteliklere dayalı olarak DeepSHAP algoritması kullanılarak açıklandı. SHAP yöntemi ise değişmeden bırakılmış ve her bir girdi örneği için açıklamalar oluşturulmuştur.

3. MATERYAL VE METOT

3.1. Metabolomik Veri Seti

Bu tez çalışmasında, Metabolomics Workbench veri tabanına ST000726 çalışma kimliği ile kaydedilmiş olan PKOS hastalarına ilişkin negatif iyon modlu metabolomik veri seti kullanılmıştır. İlgili araştırmada PKOS olan ve olmayan hastalarda hedeflenmemiş metabolomikler için 89 örnek analiz edilmiştir. Tüm numuneler, Colorado Çocuk Hastanesinde Etilen Diamin Tetraasetik Asit (EDTA) tüplerine alınmıştır. Toplanan kanlar hemen işlenmiş ve -80 derecede dondurulmuştur. LC-MS analizleri sonucunda elde edilen veriler 89 hastaya ait 70 metabolitin seviyelerini içermektedir (71). Tablo 3.1 ile metabolomik veri setine ilişkin bilgiler sunulmaktadır.

Tablo 3.1. Metabolitlere ilişkin süper, ana ve alt sınıf bilgileri

Metabolitlerin Refmet İsimleri	Süper Sınıf	Ana Sınıf	Alt Sınıf
2-Aminobenzoik asit	Benzenoidler	Benzenler	Aminobenzoik asit
2-Hidroksibutirik asit	Yağ asitleri	Yağ asitleri	Hidroksi YA
2-Hidroksiizobütirik asit	Yağ asitleri	Yağ asitleri	Hidroksi YA
3-Metil-2-oksovalerik asit	Organik asit	Keto asit	Short-chain Keto asit
3-Metilksantin	Nükleik asit	Pürinler	Ksantinler
4-Piridoksik asit	Alkaloidler	Nikotinik asit alkaloidleri	Piridin alkaloidleri
5-Hidroksi-triptofan	Organik asit	Amino asit ve peptitler	Amino asit
Adipik asit	Yağ asitleri	Yağ asitleri	Dikarboksilik asit
Araşidik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Araşidonik asit	Yağ asitleri	Yağ asitleri	Doymamış YA
Behenik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
cis-Erusik asit	Yağ asitleri	Yağ asitleri	Doymamış YA
Dehidroepiandrosteron sülfat	Sterol lipidleri	Steroidler	C19 steroidler
Deoksiadenozin	Nükleik asit	Pürinler	Pürin deoksiribonükleositler
Deoksikolik asit	Sterol lipidleri	Safra asitleri	C24 Safra asit
Deoksiüridin	Nükleik asit	Pirimidinler	Pirimidin deoksiribonükleositler
Hippurik asit	Poliketidler	Fenolik asit	Fenolik asit
Histidin	Organik asit	Amino asit ve peptitler	Amino asit
Homovanilik asit	Poliketidler	Fenolik asit	Fenolik asit

Hidroksifenillaktik asit	Poliketidler	Phenylpropanoids	Cinnamic asit
Hipoksantin	Nükleik asit	Pürinler	HypoKsantinler
İzoSitrik asit	Organik asit	TCA asitleri	TCA asitleri
İzolösin	Organik asit	Amino asit ve peptitler	Amino asit
Ketoleucine	Organik asit	Keto asitleri	Short-chain keto asitleri
Kinurenin	Organik asit	Amino asit ve peptitler	Amino asit
Laurik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
L-Dehidroaskorbik asit	Organoheterosiklik bileşikler	Laktonlar	Gamma butyrolactones
Linoleik asit	Yağ asitleri	Yağ asitleri	Doymamış YA
LPE 18:0	Gliserofosfolipitler	Gliserofosfoetanolaminler	LPE
Lizin	Organik asit	Amino asit ve peptitler	Amino asit
Malik asit	Organik asit	TCA asitleri	TCA asitleri
Margarik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Miristik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Miristoleik asit	Yağ asitleri	Yağ asitleri	Doymamış YA
N-Asetilalanin	Organik asit	Amino asit ve peptitler	Amino asit
N-Asetilmetiyonin	Organik asit	Amino asit ve peptitler	Amino asit
Sinir asiti	Yağ asitleri	Yağ asitleri	Doymamış YA
Oleik asit	Yağ asitleri	Yağ asitleri	Doymamış YA
Oksaloasetik asit	Organik asit	TCA asitleri	TCA asitleri
p-Asetamidofenilglukuronid	Karbonhidratlar	Monosakkaritler	Şeker asitleri
Palmitik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Palmitoleik asit	Yağ asitleri	Yağ asitleri	Doymamış YA
Pantotenik asit	Organik asit	Amino asit ve peptitler	Amino asit
Fenilalanin	Organik asit	Amino asit ve peptitler	Amino asit
Fitanik asit	Prenol lipidleri	Isoprenoids	C20 isoprenoids
Piroglütamik asit	Organik asit	Amino asit ve peptitler	Amino asit
S-Adenosilhomosistein	Nükleik asit	Pürinler	Diğer pürinler
Stearik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Süksinik asit	Organik asit	TCA asit	TCA asit
Taurolitokolik asit	Sterol lipidleri	Safra asitleri	C24 safra asit
Teofilin	Nükleik asit	Pürinler	Ksantinler
trans-Akonitik asit	Organik asit	Karboksilik asitleri	Trikarboksilik asitleri
trans-Zeatin	Nükleik asit	Pürinler	Alkilaminopurinler

Trikosilik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Tridesiklik asit	Yağ asitleri	Yağ asitleri	Doymuş YA
Triptofan	Organik asit	Amino asit ve peptitler	Amino asit
Tirozin	Organik asit	Amino asit ve peptitler	Amino asit
Ürik asit	Nükleik asit	Pürinler	Ksantinler
Üridin	Nükleik asit	Pirimidinler	Pirimidin ribonükleozitler
Valin	Organik asit	Amino asit ve peptitler	Amino asit
Ksantin	Nükleik asit	Pürinler	Ksantinler

3.2. Veri Önişleme

Çalışmada ilk olarak veriler %80 eğitim ve %20 test seti olarak bölündü. Verilerin bölünmesinden sonra eğitim ve test setlerinde kayıp değerlerin atanması ve verilerin normalleştirilmesi ayrı ayrı yapıldı. Veri analitiği süreçlerinde, kayıp verilerle karşılaşıldığında bu kayıpların ele alınması ve veri setinin ileri analizlere uygun hale getirilmesi önemlidir. Metabolomik gibi omik verilerde kayıp veriler, analiz sonuçlarını etkileyebileceği için uygun bir işlem ile bu veriler için atama yapılması gerekmektedir. Çalışmada ilk aşamada kayıp oranı %30'un üzerinde olan değişkenler hariç tutulmuştur ve bu değişkenler ileri analizlere dâhil edilmemiştir. <%5 kayıp değerlere sahip olan değişkenlere ortalama değer ile atamayı ve kayıp değerleri>%5 olan değişkenler için, çoklu atamayı uygulamak amacıyla rasgele orman paketi kullanıldı (72). Eğitim seti, pozitif (PKOS) ve negatif (kontrol) grupları dengelemek için Borderline- Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) aşırı örnekleme algoritması kullanılarak dengelendi. Modellerin stabilitesini geliştirmek için, tüm sürekli değişkenler ortalaması 0 ve standart sapması 1 olan bir dağılım elde edecek şekilde normalleştirildi. Sonrasında biyobelirteç keşfi için değişkenler PKOS ile istatistiksel anlamlılık açısından incelendi ve anlamlı olmayan değişkenler elendi. Bu aşamadan sonra elde edilen değişkenler kümesi içerisinden biyobelirteç adayı metabolitleri belirleyebilmek için dört farklı yaklaşımda ortak olan metabolitler sonraki analizler için seçildi. Dikkate alınan dört farklı yaklaşım aşağıda açıklanmaktadır:

- Kat değişimi (fold change; FC) analizi ile yukarı veya aşağı regülasyonlar incelendi ve anlamlı değişkenler belirlendi.
- Kısmi en küçük kareler diskriminant analizi (PLS-DA) modelindeki projeksiyon değişken önem (VIP) puanı birden büyük olan değişkenler belirlendi.

- ML modellerinin eğitimi sırasında aşırı uyum (overfitting) veya eksik uyum (underfitting) problemleri ile karşılaşmamak, PKOS'a yönelik anahtar metabolitleri seçmek ve hasta değişkenlerinin optimal alt kümesini filtrelemek için Boruta ve Minimum Artıklık ve Maksimum Alaka (MRMR) değişken seçimi yöntemleri uygulandı ve önemli değişkenler belirlendi.

3.2.1. Rasgele Orman Algoritması ile Kayıp Değer Atama

Rasgele orman tabanlı kayıp değer atama, kayıp verilerin tahmin edilmesi için kullanılan bir yöntemdir. Rasgele orman, çoklu karar ağaçlarından oluşan bir topluluk öğrenme algoritmasıdır ve kayıp değerlerin tahmin edilmesi için oldukça etkili bir yöntemdir (73). Rasgele orman tabanlı atama yönteminin aşamaları aşağıdaki gibidir:

- **Veri Hazırlığı:** Kayıp verilerin bulunduğu değişkenler belirlenir ve kayıp verilerin oranı belirlenir.
- **Veri Bölünmesi:** Veri seti, kayıp değer içermeyen bir veri seti (tam veri) ve kayıp değerlere sahip bir veri seti olarak iki alt kümeye bölünür.
- **Rasgele Orman Modeli Oluşturma:** Tam veri seti üzerinde bir Rasgele orman modeli oluşturulur. Bu model, kayıp değerlerin tahmininde temel olarak kullanılacaktır.
- **Kayıp Değerlerin Tahmin Edilmesi:** Kayıp değerlere sahip veri seti, oluşturulan Rasgele orman modeli kullanılarak tahmin edilir. Model, kayıp değerlerin tahmini için diğer değişkenlerin kullanımını dikkate alır ve tahminlerde bulunur.
- **Tahminlerin Uygulanması:** Tahmin edilen kayıp değerler, kayıp değerlerin olduğu veri setine uygulanır. Bu sayede kayıp verilerin yerine geçici olarak tahmin edilen değerler eklenir.
- **Kayıp Değer Atama Sürecinin İterasyonu:** Veri setinde birden fazla değişkende kayıp değer varsa veya kayıp değer miktarı yüksekse, kayıp değer atama süreci birkaç kez tekrarlanabilir. Her iterasyonda, kayıp değerlerin daha iyi tahmin edilmesi için daha fazla veri kullanılabilir.

3.2.2. Borderline-SMOTE

Sınıf dengesizliği, bir veri kümesindeki farklı sınıflar arasındaki gözlemler arasındaki belirgin sayısal farklılıkları ifade eder. Borderline-SMOTE özellikle sınıf dengesizliği sorunuyla başa çıkmak için kullanılan bir örnekleme yöntemidir. Bu yöntem, özellikle sınıf dengesizliği içeren ML problemlerinde yansız/sağlam tahmin modelleri elde etmek ve performansı artırmak için geliştirilmiştir (74).

ML görevlerinde az sayıda örneğe sahip olan azınlık sınıflarında modelin eğitilmesi zorlaşabilir. Borderline-SMOTE, bu sorunu çözmek için SMOTE yöntemini genişleterek özellikle azınlık sınıflardaki sınırlı örnekleri odak alır. SMOTE, azınlık sınıfındaki örnekleri çoğaltarak sentetik örnekler oluşturur. Ancak, sınıf sınırlarında veya karar sınırlarında bulunan örnekler genellikle modelin performansını daha fazla artırabilir. Borderline-SMOTE, bu sınırlarda bulunan örnekleri vurgular ve sentetik örnekleri oluştururken sadece bu örnekleri dikkate alır. Bu sayede, modelin azınlık sınıfındaki sınırlı veriye daha iyi uyum sağlaması ve genelleme yapması amaçlanır (74).

- **Sınırlı Örnek Odaklıdır:** Borderline-SMOTE, azınlık sınıflardaki sınırlı örnekleri hedef alır. Bu, sentetik örneklerin oluşturulurken azınlık sınıfının karar sınırlarında yoğunlaşmasını sağlar. Diğer SMOTE yöntemleri sınıf içinde genel bir yaklaşım benimserken, Borderline-SMOTE, zorlu bölgelere daha fazla odaklanır.
- **Daha İyi Genelleme:** Sınırlı örnekler üzerinde yoğunlaşmak, modelin daha iyi genelleme yapmasına yardımcı olabilir. Model, zorlu karar sınırlarında daha fazla çeşitliliği öğrenir, bu da modelin yeni verilerle daha iyi başa çıkabilme yeteneğini artırabilir.
- **Daha İyi Azınlık Sınıfı Temsili:** Borderline-SMOTE, azınlık sınıfındaki sınırlı örnekleri çoğaltarak sentetik örnekler oluşturduğu için azınlık sınıfının temsilini artırır. Bu, modelin azınlık sınıfını daha dengeli bir şekilde öğrenmesine yardımcı olabilir.

3.2.3. Değişken Seçimi

Değişken seçimi, belirli bir modele dâhil edilecek potansiyel çok sayıda değişken arasından seçim yapma sürecidir; bu, ilgisiz veya gereksiz olanları elemek ve uygun değişkenleri içeren eksiksiz bir değişken listesi oluşturmak anlamına gelir (56). Bu seçimin temel amacı, modele en iyi uyumu sağlayacak bir değişken kümesini belirleyerek doğru tahminler yapmaktır.

Değişken seçiminin, tek başına istatistiksel seçim yöntemlerinden daha çok klinik bilgiye ve önceki literatüre odaklanması gerektiği sıklıkla tavsiye edilir. Veri setleri genellikle nihai modele katkıda bulunmayan bir dizi ek değişken içerir. Gürültülü değişkenleri nihai modele dâhil etmekten kaçınmak için, uygun değişkenlerin seçimi gerçekleştirilmelidir. Binlerce değişken içeren büyük veri setleri (metabolomik, proteomik, genomik, vb.), geleneksel yaklaşımların etkili bir şekilde ele alınmasını ve yönetilmesini zorlaştırır.

Bu gibi veriler ile çalışırken değişken seçimi, veri analizi ve model oluşturma aşamalarında kullanılan değişken sayısını azaltarak işlemci gücünden ve kaynaklardan tasarruf sağlar, modelin daha anlaşılır ve yorumlanabilir olmasını destekler ve aşırı uyum (overfitting) sorununu azaltabilir. Ayrıca, gereksiz veya anlamsız değişkenlerin çıkarılması, modelin genellikle daha iyi performans göstermesini sağlar ve tahmin hatalarını azaltabilir (75).

Kısmi En Küçük Kareler Diskriminant Analizi (PLS-DA) Projeksiyon Değişken Önem (VIP) Skoru: PLS-DA çok değişkenli veri analizi için kullanılan istatistiksel bir yöntemdir ve özellikle yüksek boyutlu veri setlerinde gruplar arasındaki karmaşıklıkların çözümlenmesinde etkili bir araçtır. VIP skoru bir değişkenin PLS-DA modelindeki öneminin bir ölçüsüdür. Bir değişkenin modele yaptığı katkıyı özetler ve PLS-DA bileşenleri ile orijinal değişken arasındaki korelasyonların karelerinin ağırlıklı toplamı olarak hesaplanır. Ağırlıklar, modeldeki PLS-DA bileşeni tarafından açıklanan yüzdesel değişime karşılık gelir. VIP skoru yüksek olan değişkenler, modelin sınıfları daha iyi ayırt ettiği ve sınıflar arasındaki farkları daha iyi açıkladığı düşünülen değişkenlerdir (76).

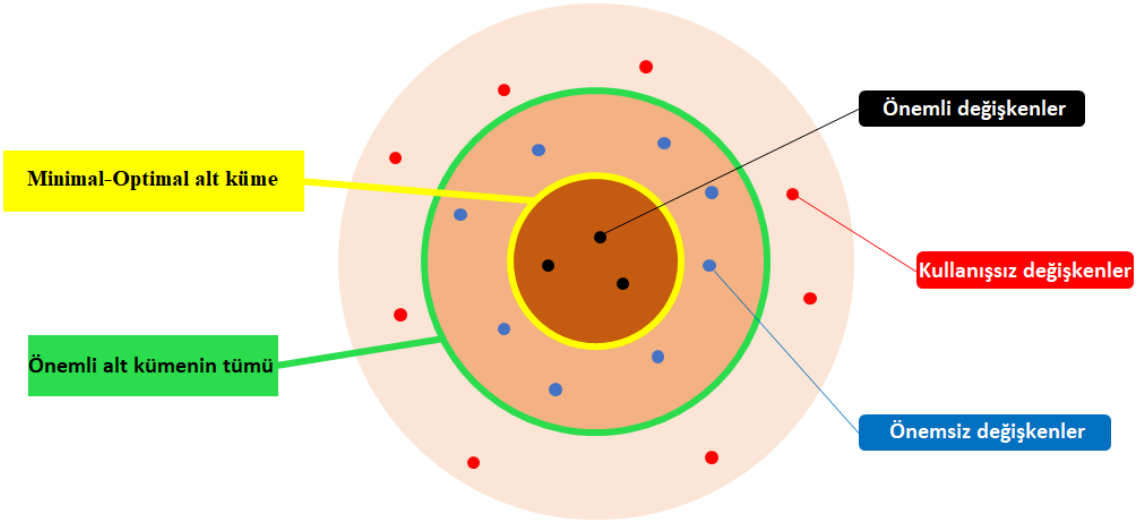
Boruta Değişken Seçimi Yöntemi: Boruta algoritması, R paketi randomForest'te uygulanan rastgele orman sınıflandırma algoritması etrafında oluşturulmuş bir sarmalayıcıdır. Rastgele orman sınıflandırma algoritması nispeten hızlıdır, genellikle parametreler ayarlanmadan çalıştırılabilir ve değişkenin önemine ilişkin sayısal bir tahmin verir. Birden fazla tarafsız zayıf sınıflandırıcının (karar ağaçları) oylanmasıyla sınıflandırmanın yapıldığı bir topluluk yöntemidir. Bir değişkenin önem ölçüsü, nesneler arasındaki değişken değerlerinin rastgele permütasyonundan kaynaklanan sınıflandırma doğruluğu kaybı olarak elde edilir. Sınıflandırma için belirli bir değişkeni kullanan ormandaki tüm ağaçlar için ayrı ayrı hesaplanır. Sonra doğruluk kaybının ortalaması ve standart sapması hesaplanır. Alternatif olarak, ortalama kaybın standart sapmasına bölünmesiyle hesaplanan Z puanı önem ölçüsü olarak kullanılabilir. Ancak Z puanı, dağılımı N (0, 1) olmadığı için rastgele orman algoritması tarafından döndürülen değişken öneminin istatistiksel anlamlılığıyla doğrudan ilişkili değildir. Bununla birlikte Boruta'da ormandaki ağaçlar arasındaki ortalama doğruluk kaybındaki dalgalanmaları hesaba kattığı için önem ölçüsü olarak Z skoru kullanılmaktadır. Z puanı doğrudan önemi ölçmek için kullanılmadığı için, herhangi bir değişkenin öneminin anlamlı olup olmadığına, yani rastgele dalgalanmalardan kaynaklanabilecek önemden ayırt edilip edilemeyeceğine

karar vermek için bazı dış referanslara ihtiyaç vardır. Bu amaçla bilgi sistemi tasarım gereği rastgele olan değişkenlerle genişletilir. Her bir değişken için, değerleri orijinal değişken değerlerinin nesneler arasında karıştırılmasıyla elde edilen karşılık gelen bir 'gölge' değişken yaratılır. Daha sonra bu değişkenleri kullanarak bir sınıflandırma modeli oluşturulur ve gölge değişkenlerin önem kümesi, hangi değişkenlerin gerçekten önemli olduğuna karar vermek için bir referans olarak kullanılır (77).

Minimum Artıklık ve Maksimum Alaka (MRMR) Değişken Seçimi Yöntemi: MRMR, bir veri setinden en önemli ve bilgi taşıyan değişkenleri seçmek için kullanılan bir yöntemdir. MRMR, değişkenlerin veri seti içindeki ilişkilerini değerlendirerek hem özgün hem de anlamlı değişkenlerin belirlenmesine yardımcı olur. İşte MRMR yönteminin temel prensipleri:

- **Minimum Artıklık:** MRMR, seçilen değişkenler arasındaki gereksiz ilişkileri azaltmayı hedefler. Bundan dolayı, seçilen değişkenlerin birbirleriyle mümkün olduğunca az ilişkili olmasını amaçlar. Yani, seçilen değişkenler birbirlerinin yerine geçmeyecek ve aynı bilgiyi taşımayacak şekilde seçilir. Bu sayede, modelin daha az karmaşık ve daha özgün olmasını sağlar.
- **Maksimum Alaka:** MRMR, veri setiyle maksimum ilişkisi olan değişkenleri seçmeyi amaçlar. Seçilen değişkenler veri setinin hedef değişkeniyle (veya sınıf etiketiyle) en fazla ilişkili olacaktır. Bu, modelin doğru tahminler yapabilmesi için gerekli olan bilgiyi içeren değişkenlerin seçilmesini sağlar.

MRMR, değişken seçimi sürecini matematiksel bir optimizasyon problemi olarak ele alır ve her bir değişkenin hedef değişkenle olan ilişkisi hesaplanır ve ardından değişkenler arasındaki gereksiz ilişkileri azaltmak için değişkenler arasındaki ilişkiyi hesaplar (78).



Şekil 3.1. MRMR değişken seçimi yöntemi

3.3. Çalışmada Kullanılan Ağaç Tabanlı Makine Öğrenmesi Algoritmaları

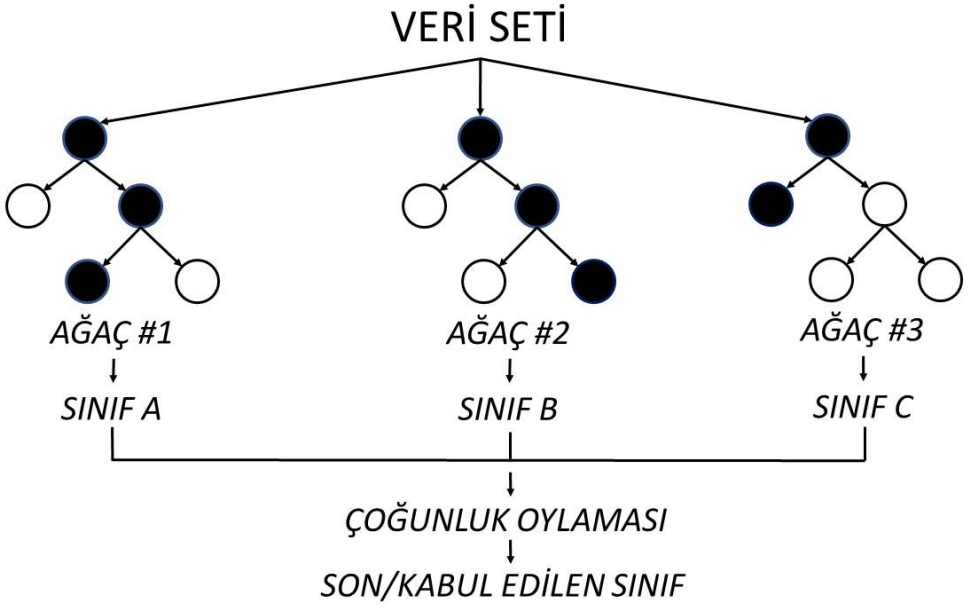
3.3.1. Rastgele Orman Algoritması

Rastgele orman algoritması, sınıflandırma ve regresyon görevi 2001 yılında L. Breiman tarafından önerilmiştir. Rastgele orman, parametrik olmaması, yorumlanabilir olması, verimli olması, yüksek tahmin doğruluğu ve düşük işlem hızı gibi özellikleri nedeniyle büyük ve çok boyutlu veri setleri üzerinde sınıflandırma görevlerini gerçekleştirmek için popüler bir algoritmadır (79).

Sınıflandırıcı olarak birçok karar ağacı üreten ve bunların sonuçlarını nihai tahmin vermek üzere birleştiren bir topluluk öğrenme yöntemidir. Topluluk öğrenme modellerini oluşturmak için uygulanan iki yöntem vardır ve bunlar sıralı topluluk yöntemleri ve paralel topluluk yöntemleridir. Sıralı topluluk yöntemlerinde, temel öğrenenin sonucu, aynı verileri kullanarak bir sonraki adımda öğreneni oluşturmak için geri bildirim olarak kullanılır. Paralel topluluk yöntemlerinde, öğrenme modelleri paralel olarak üretilir, burada temel öğrenici birbirinden bağımsız olarak oluşturulur.

Rastgele ormanda birbirinden bağımsız karar ağaçlarını oluşturmak için paralel topluluk öğrenme yöntemi kullanılır. Nihai öğrenme modelinin karar ağacı olan küçük bir alt kümesini oluşturma süreci aynı zamanda torbalama olarak da bilinir. Bagging, bootstrap ve aggregation olmak üzere iki farklı teknik kullanılarak uygulanan bir kavramdır. Bootstrap, orijinal veri kümesindeki veri satırlarını rastgele seçerek mevcut veri kümesinden ayrı bir veri kümesi oluşturma işlemidir. Bu işlem sırasında aynı satırın birden çok kez seçilip yeni veri setine yerleştirilmesine değiştirmeli örnekleme adı verilmektedir.

Aynı zamanda bazı satırlar hiç seçilmeyebilir veya orijinal veri setinde dokunulmadan bırakılabilir. Artık, öğrenenler olarak birden fazla ağaç oluşturma süreci için, önyükleme veri kümesinden farklı bir özellik ve örnek alt kümesi rastgele seçilir. Örneğin, çok boyutlu bootstrap veri setinden bazı ağaçlar yalnızca 2 değişken kullanılarak oluşturulabilirken, bazı ağaçlar 2'den fazla değişken kullanılarak oluşturulabilir. Bootstrap veri setinden rastgele seçilen veri alt kümesi arasından en iyi bölünme niteliğini seçmek için, sınıflandırma ve regresyon ağaçlarında kullanılan bir yöntem olan gini safsızlığı hesaplanır. Rastgele seçilmiş bir veri seti alt kümesiyle ağacın öğrenicileri veya sınıflandırıcıları oluşturulduktan sonra, bu ağaçlar, kullanıma hazır veriler kullanılarak verimlilikleri açısından test edilir. Torba dışı veriler, bootstrap veri kümesine dahil edilmeden orijinal veri kümesinde bırakılan verilerdir. Temel öğrenicinin verimliliği, bu torba dışı verilerin temel öğreniciye beslenmesiyle tahmin edilir. Artık tüm öğreniciler veya sınıflandırıcılar, toplama işlemi yoluyla hazır veri seti ile test edildiğinde, bu sınıf, modeldeki temel öğrenicilerin veya sınıflandırıcıların çoğunluğu tarafından tahmin edilen nihai tahmin edilen sınıf olarak seçilir. Bu noktada oluşturulan ağaçlar herhangi bir sınıfa ait örnekleri belirlemek için oylarını kullanırlar (79).



Şekil 3.2. Rasgele orman algoritması

3.3.2. Catboost Algoritması

CatBoost, sınıflandırma, regresyon ve sıralama görevi için gradyan artırma ilkesiyle uygulanan bir makine öğrenmesi algoritmasıdır. CatBoost hem sayısal hem de kategorik değişkenlere sahip veri setleri ile kullanılabilmektedir.

CatBoost kategorik değişkenleri ele almak için aşırı uyumu azaltmaya yardımcı olan bir strateji kullanır. Bu stratejide, değişkenler rastgele olarak değiştirilir veya karıştırılır. Her gözlem için, gözlemin aynı kategori değerine sahip, karıştırılmış listede verilenin önüne yerleştirilen ortalama bir etiket değeri hesaplanır. Modelin eğitimi sırasında her gözlem ortalama etiket değeriyle değiştirilir. Topluluk yaklaşımını takip ettiği için yeni bir karar ağacı oluşturulurken veri seti üzerinde yeniden permütasyon yapılır ve kategorik gözlemin sayısal değerinin hesaplanması işleminin tamamı tekrarlanır. CatBoost'ta kategorik değişkenleri sayısal değişkenlere dönüştürmek için uygulanan bir diğer güçlü ve etkili yöntem, veri setinde bulunan gözlemlerin sıklığının hesaplanması ve eğitim sürecinde bunun sayılan değerle değiştirilmesidir. Bu yöntem aynı zamanda algoritmanın performansını da arttırmaktadır. CatBoost'un bir diğer önemli özelliği değişken kombinasyonudur. Değişken kombinasyonunda CatBoost, veri setinde mevcut olan değişkenleri birleştirerek daha güçlü değişkenler üretir. Mevcut tüm değişkenler birbiriyle birleştirilemediği için hesaplama yükü oluşur ve CatBoost bu durumun üstesinden gelebilir.

Bir sonraki bölme için CatBoost, mevcut ağaçta mevcut olan tüm kombinasyonları ve kategorik değişkenleri, veri setindeki tüm kategorik değişkenlerle birleştirir. CatBoost'u XGBoost ve LightGBM gibi diğer gradyan artırıcı karar ağaçlarından farklılaştıran bir özellik ise "Gradyan Yanlılık ile Mücadele"dir. CatBoost'ta, XGBoost ve LightGBM'den farklı olarak ağaç yapısını oluşturmak için degrade tabanlı karar ağaçlarının değiştirilmiş versiyonu ve oluşturulan ağacın yaprak değerlerini ayarlamak için geleneksel degrade artırıcı karar ağaçları kullanılmıştır (80).

3.3.3. XGBoost Algoritması

XGBoost, sınıflandırma, regresyon ve sıralama görevleri için gradient artırma ilkesini uygulayan bir diğer algoritmadır. Chen ve Gustrin tarafından 2016 yılında önerilmiştir. XGBoost, gradyan artırma algoritmasının genişletilmiş bir versiyonu olan aşırı gradyan artırma algoritması anlamına gelir ve temel öğrenenler oluşturmak ve geliştirmek için gradyan artırma algoritmasıyla aynı prensibe dayanmaktadır. XGBoost, artırma tekniği olarak da adlandırılan bir ek sınıflandırıcı oluşturmak için sıralı bir topluluk yaklaşımı kullanır ve denetimli ML sorunları için kullanılır (81).

XGBoost'un öğrenme işlevi, düzenlileştirilmiş amaç fonksiyonu olarak temsil edilir ve amaç fonksiyonu, modelin tahmin görevinde ne kadar iyi performans gösterdiğini ölçmek için tanımlanır. Artıkları ölçmek için XGBoost'ta kullanılabilecek menteşe kaybı, lojistik kayıp, çapraz entropi kaybı ve üstel kayıp gibi birçok kayıp

fonksiyonu olmasına rağmen, genellikle ortalama karesel hata ve lojistik kayıp fonksiyonları kullanılır. Düzenlileştirme terimi eğitim veri setinde aşırı uyumu önlemeye yardımcı olur. Öğrenme modelinin amacı kaybı en aza indirmek ve dolayısıyla modelin verimliliğini arttırmaktır (81).

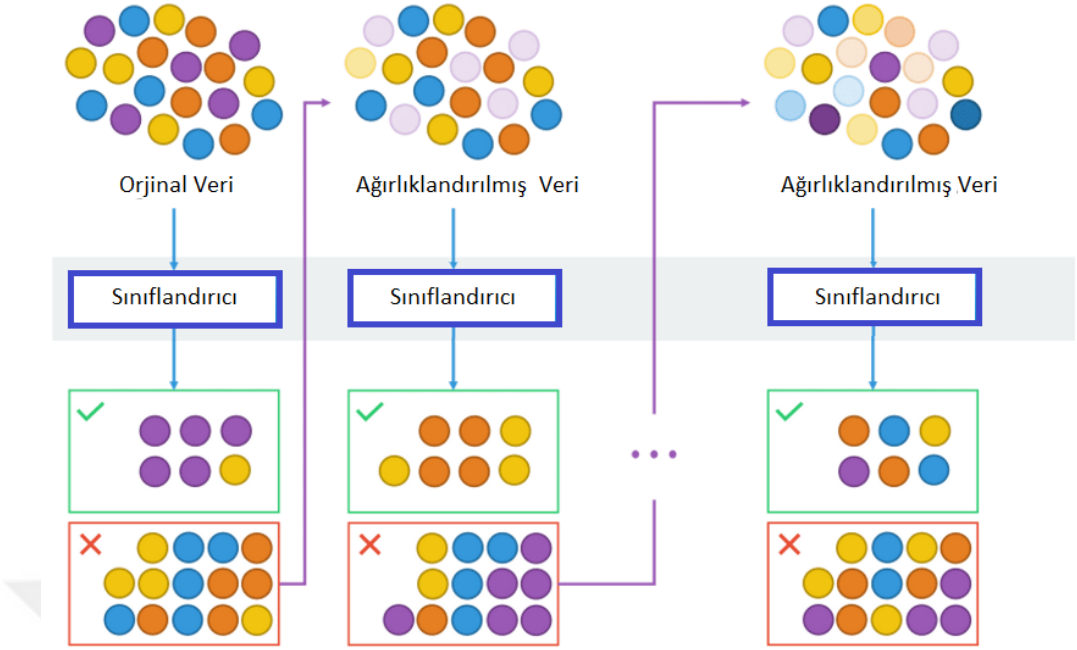
XGBoost'un işlem hızı ve tahmin doğruluğu açısından gradyan artırma makinelerinden çok daha verimli olduğu düşünülmektedir. Değişken ölçeklenebilirliği nedeniyle gradyan artırmadan on kat daha hızlıdır. XGBoost'ta yüksek ölçeklenebilirlik desteği sağlayan başka birçok özellik vardır. XGBoost, aşırı uyumu önlemeye yardımcı olan küçültme ve sütun alt örnekleme özelliklerine sahiptir. Küçültme, ağaç güçlendirmenin her adımından sonra yaprakta yeni eklenen ağırlıkları bazı faktörlerle ölçeklendirir. Her bir ağacın tahmin edilen ağırlığı olan yaprağın nihai ağırlığı azaltıldığında, kalan miktar artar ve bu da yeni ağaçların gelecek tahminini iyileştirmesine yol açar.

Sütun alt örneklemesinde, değişkenlerin alt kümeleri rastgele seçilir, böylece algoritmaların, modeli eğitirken veri setinde mevcut olan tüm değişkenleri dikkate almasına gerek yoktur. Sütun alt örneklemesi ayrıca gerekli eğitim süresinin azaltılmasına da katkıda bulunur. Bir veri setindeki en iyi bölünmüş adayı bulmak için XGBoost tam açgözlü algoritma ve yaklaşık algoritma uygular. Tam açgözlü algoritma, bir veri setinin bir bütün olarak tek bir konuma yerleştirilmesi durumunda kullanılır. Yaklaşık algoritma, dağıtılmış konumlarda gerçekleştirmek için en iyi bölme özelliğine yönelik hesaplamaya ihtiyaç duyulduğunda kullanılır. Ayrıca XGBoost'un bir diğer önemli özelliği de seyrek veri tipinin farkında olmasıdır. XGBoost, çok boyutlu değişken vektörüne sahip yüksek hacimli karmaşık veriler için en uygun çözümdür. Ve gerçek hayatta bu tür veri setleri daha seyrek verilere sahiptir. XGBoost, birden fazla eksik özelliğe sahip veri setleri, sık sıfır değerleri gibi seyrek verilerle beslendiğinde, bu örnekleri ilgili yöne iletmek için ağacın kök düğümünden ayrı bir yol bildirilir. Son olarak XGBoost'un performans hızının da yüksek olmasını sağlayan en önemli özelliği çekirdek dışı hesaplamaları desteklemesidir. Bu özellikte, bilgisayarın birincil belleğine aynı anda sığdırılamayan veriler bloklara bölünerek ikincil bellek konumlarının farklı yerlerinde depolanır. Yürütme işlemi sırasında ihtiyaç duyulduğunda ikincil bellek konumlarından yüklenirler. XGBoost, minimum miktarda kaynak kullanarak son teknoloji ürünü bir çözüm sağladığından, sınıflandırma ve regresyon görevleri için ML ve AI uygulamalarında oldukça popülerdir. Özellikle, büyük veri setlerinde hızlı eğitim süresi, yüksek ölçeklenebilirlik ve yüksek tahmin doğruluğu sağlar (81, 82).

3.3.4. AdaBoost Algoritması

AdaBoost, zayıf sınıflandırıcıları bir araya getirerek güçlü bir sınıflandırıcı oluşturmayı amaçlayan bir topluluk öğrenme algoritmasıdır. AdaBoost, her bir sınıflandırıcıyı eğitirken hatalı tahminlerin ağırlığını artırarak sınıflandırma doğruluğunu geliştirir. Bu sayede, zayıf sınıflandırıcıların birleşimiyle daha yüksek bir performans elde edilir. Bir dizi eğitim örneği verildiğinde, AdaBoost, bu örnekler üzerinde W ağırlık dağılımını korur ve bu dağılım başlangıçta tekdüze olarak ayarlanmıştır. Daha sonra AdaBoost, bileşen öğrenme (Component Learn) algoritmasını bir dizi döngü halinde tekrar tekrar çağırır. t döngüsünde AdaBoost, bileşen öğrenmeye Wt dağılımına sahip eğitim örnekleri sağlar. Buna yanıt olarak bileşen öğrenme, ht sınıflandırıcısını eğitir. Wt dağılımı, eğitim örnekleri üzerindeki tahmin sonuçlarına göre her döngüden sonra güncellenir.

Doğru şekilde sınıflandırılan "kolay" örnekler daha düşük ağırlıklar alırken, yanlış sınıflandırılan "zor" örnekler daha yüksek ağırlıklar alır. Böylece AdaBoost, bileşen öğrenme için daha zor görünen, daha yüksek ağırlıktaki örneklere odaklanır. Bu süreç T döngüleri boyunca devam eder ve son olarak AdaBoost tüm bileşen sınıflandırıcılarını tek bir son hipotez f'de doğrusal olarak birleştirir. Daha düşük eğitim hatalarına sahip bileşen sınıflandırıcılara daha fazla ağırlık verilir. AdaBoost'un önemli teorik özelliği, eğer bileşen sınıflandırıcıları sürekli olarak yarıdan biraz daha iyi bir doğruluğa sahipse, son hipotezin eğitim hatasının üstel olarak hızlı bir şekilde sıfıra düşmesidir. Bu, bileşen sınıflandırıcılarının rastgeleden yalnızca biraz daha iyi olması gerektiği anlamına gelir (83).



Şekil 3.3. AdaBoost algoritması gösterimi

3.3.5. LightGBM Algoritması

LightGBM, sınıflandırma, regresyon ve sıralama görevleri için gradyan artırma algoritmasını uygulamaya yönelik bir makine öğrenimi algoritmasıdır. Guolin Ke ve diğerleri tarafından önerilen LightGBM XGBoost'un çalıştığı degrade artırma prensibiyle çalışır. Hız ve doğruluk açısından XGBoost'un geliştirilmiş versiyonudur.

Birden fazla değişkenin işlenmesi için büyük miktarda veriye ihtiyaç duyulduğunda, XGBoost'ta tüm değişkenler arasında en iyi bölünme noktasını belirlemek için belirli değişkenlerde bulunan her verinin taranması gerekir ve bu da çok zaman alıcıdır. Bu sorunu ortadan kaldırmak için Guolin Ke ve ark. LightGBM için Gradyan tabanlı tek taraflı örnekleme (GOSS) ve Özel Değişken Paketleme (EFB) olmak üzere iki yeni yöntem önerdi. GOSS'ta en iyi bölünme noktasını bulmak amacıyla bilgi kazanımını hesaplamak için yalnızca gradyanları daha yüksek olan veri noktaları dikkate alınır. Daha düşük gradyanlara sahip diğer veriler rastgele örneklenir ve sabit bir değerle çarpılır, böylece daha düşük gradyanlara sahip veriler en iyi bölünme noktasının bulunmasında etkili olabilir. Bu yöntem, en iyi bölünme noktasını bulma sürecinde daha yüksek hız sağlar (84).

Yüksek boyutlu veri setlerinde çok büyük seyreklik olabilir. Yüksek seyrekliğe sahip değişkenler yalnızca eğitim sırasında kaynak gereksinimlerini artırabilir.

Bu sorunu ortadan kaldırmak için bu tür değişkenlerin veri setinden çıkarılması en iyi bölünme noktasının hesaplanmasında da etkili olabilir. LightGBM'de bu sorun, algoritmanın hızına ve verimliliğine katkıda bulunan değişkenlerin sayısını azaltmak için birbirini dışlayan değişkenlerin bir araya getirildiği EFB'nin uygulanmasıyla çözülmektedir. LightGBM bu yönüyle eğitim hızı açısından diğer geleneksel gradyan artırıcı karar ağaçlarından 20 kat daha hızlıdır (84).

3.3.6. NGBoost Algoritması

NGBoost, geleneksel gradient boosting çerçevesini genişleterek doğal gradyan optimizasyon tekniklerini içeren bir ML algoritmasıdır. Gradient boosting, her aşamada önceki aşamanın hatalarını düzeltir şekilde tahmin modeli oluşturan bir topluluk öğrenme tekniğidir. Özellikle regresyon ve sınıflandırma gibi çeşitli tahmin modelleme görevleri için etkilidir (85).

NGBoost, parametre uzayının geometrisini dikkate alan doğal gradyan optimizasyonunu tanıtarak geleneksel gradyan artışını genişletir. Geleneksel gradyan inişi, birinci dereceden gradyan bilgilerini kullanırken, doğal gradyan inişi kayıp fonksiyonunun eğriliğini dikkate alır, bu da daha hızlı ve daha stabil bir yakınsama sağlayabilir. NGBoost, özellikle olasılıklı tahminler için uygundur, çünkü nokta tahminlerin yanı sıra tahmin edilen olasılık dağılımları şeklinde belirsizlik tahminleri sağlar (85).

3.4. Hiperparametre Optimizasyonu

Tahmin modelleri oluştururken algoritmaların sabit olmayan değişkenleri olan hiperparametrelerin belirlenmesi ve ayarlanması gerekir. En iyi tahmin performansını elde etmek için son tahmin modeli oluşturulmadan önce ML modelleri ayarlanmalıdır. Hiperparametreler, sınıflandırıcıların veri setindeki ilişkileri kontrol etmesine olanak tanır. Sonuç olarak, yeni bir veri seti üzerinde herhangi bir tür eğitim yürütülmeden önce hiperparametreler optimize edilmelidir. Bu çalışmada hiperparametrelerin ayarlanması için ızgara arama stratejisi ve 10 kat çapraz doğrulama (GridSearchCV) yöntemi kullanılmıştır. GridSearchCV çok sayıda hiperparametre kombinasyonu üretir, her kombinasyon için modeli hesaplar ve en doğru olanı döndürür (86). GridSearchCV süreci için performans metriği olarak Alıcı işletim karakteristiği (ROC) eğrisi altında kalan alan (AUC) kullanıldı ve sonrasında en yüksek ortalama performansa sahip model seçildi.

3.5. Performans Değerlendirme Metrikleri

Bu tez çalışmasında, doğruluk, duyarlılık, seçicilik, F1 skoru, AUC ve Brier skoru gibi metrikler %95 güven aralığı ile ML modellerinin performansını değerlendirmek için hesaplanmıştır (87).

- **Doğruluk:** Doğru tahmin edilen örneklerin toplam örnek sayısına oranıdır. Sınıflandırma problemleri için sıkça kullanılan bir ölçüttür. Ancak, dengesiz sınıflara sahip veri setlerinde doğruluk tek başına yeterli bir ölçüt olmayabilir (87).
- **Seçicilik:** Seçicilik, bir testin ya da modelin belirli bir durumu, örneğin bir hastalığı olmayan durumu doğru bir şekilde tespit etme yeteneğini ölçen bir diğer önemli istatistiksel ölçüdür. Seçicilik, gerçek negatif sonuçları doğru bir şekilde tanımlama oranını temsil eder. Bu, hastalığı gerçekte taşımayan kişilerin ne kadarının test tarafından doğru bir şekilde tanımlandığını ifade eder (87).
- **Duyarlılık:** Duyarlılık, bir testin ya da modelin belirli bir pozitif durumu doğru bir şekilde tespit etme yeteneğini ölçen önemli bir istatistiksel ölçüdür. Gerçek pozitif sonuçları doğru bir şekilde tanımlama oranını temsil eder. Bu, hastalığı gerçekte taşıyan kişilerin ne kadarının test tarafından doğru bir şekilde tanımlandığını ifade eder (87).
- **F1 Skoru:** F1 skoru, kesinlik ve duyarlılık metriklerinin harmonik ortalamasını temel alarak, modelin pozitif tahminlerinin gerçek pozitiflere karşı dengeli bir performans sergileyip sergilemediğini değerlendirir. Dengesiz sınıflara sahip veri setlerinde kullanılan yaygın bir ölçüttür (87).
- **AUC:** AUC, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir ölçüttür. AUC, ROC eğrisinin altında kalan alanı temsil eder. ROC eğrisi, sınırlayıcı bir değeri (threshold) değiştirerek modelin duyarlılık ve seçicilik değerlerini farklı kombinasyonlarda gösterir. Bu sınırlayıcı değer, modelin çıktı tahminlerini sınıflara ayırmak için kullanılan bir eşik değerdir. Eşik değeri düşürüldüğünde, modelin pozitif tahmin yapma eğilimi artar, bu da duyarlılığın artmasına ve seçiciliğin azalmasına yol açar. Tersine, eşik değeri yükseltildiğinde, modelin negatif tahmin yapma eğilimi artar, bu da seçiciliğin artmasına ve duyarlılığın azalmasına yol açar. ROC eğrisi, duyarlılık ve seçicilik arasındaki trade-off (değiş-tokuş) ilişkisini görselleştirir.

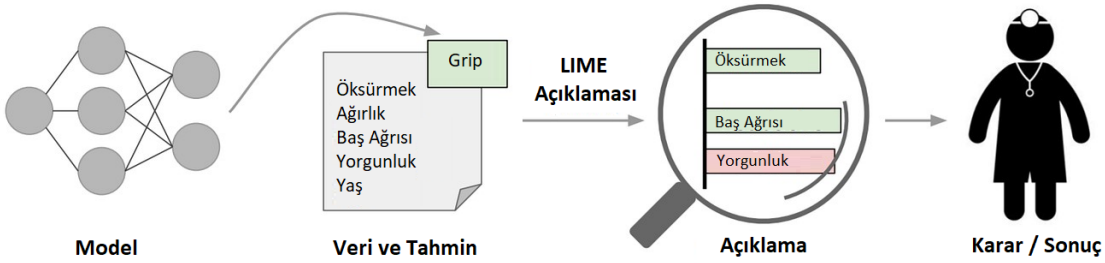
Bir modelin daha iyi performans sergilediği durumlar, ROC eğrisi üzerinde daha sol üst köşeye yakın noktalardır. Eşik değeri ne olursa olsun yüksek duyarlılık ve yüksek seçicilik sağlayan bir model, daha büyük bir AUC değerine sahip olur. AUC, ROC eğrisi ile birlikte kullanılan bir performans ölçütüdür ve modelin sınıflandırma performansını özetler. AUC değeri, 0 ile 1 arasında değişir. AUC değeri 1'e ne kadar yakınsa, modelin daha iyi bir performans sergilediği anlamına gelir. AUC değeri 0.5'e yakınsa, modelin performansı rastgele bir seçimle aynı düzeydedir (87).

- **Model Kalibrasyonu ve Brier Skor:** ML modellerinin kalibrasyonu, modelin tahminlerinin gerçek olasılıklara ne kadar iyi uyduğunun değerlendirilmesini içerir. Tahminlerin doğru olasılıklarla uyumlu olması, modellerin güvenilirliğini artırır ve daha iyi kararlar alınmasına yardımcı olur. Kalibrasyonu değerlendiren ölçütler ML modelinin tahminlerinin gerçeklikle ne kadar iyi uyumlu olduğunu değerlendirmeye yardımcı olur. İyi kalibre edilmiş bir model, doğru olasılıklarla tutarlı tahminler yapar ve bu da modelin güvenilirliğini artırır. Bu tez çalışmasında ML modellerinin kalibrasyonunu değerlendirmek için Brier skoru hesaplanmıştır. Brier Skoru, gerçek sınıf etiketiyle tahmin arasındaki ortalama karesel hata olarak hesaplanır. Daha düşük bir Brier skoru, daha iyi kalibre edilmiş ve daha doğru tahminler yapan bir modeli gösterir. Skor 0 ile 1 arasında değer alır, 0 en iyi, 1 ise en kötü skordur (87).

3.6. Çalışmada Kullanılan Açıklanabilir Yapay Zekâ Algoritmaları

3.6.1. Yerel Yorumlanabilir Model-Bağımsız Açıklamalar

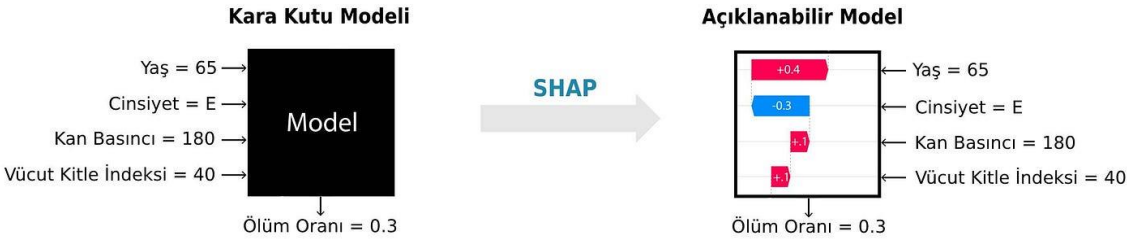
LIME, model bağımsız bir açıklama yöntemidir ve özellikle karmaşık ve kara kutu modellerin nasıl tahminlerde bulunduğunu daha anlaşılır bir şekilde açıklamak için kullanılır. LIME, belirli bir örnek veya tahmin üzerinde modelin davranışını açıklamak için lokal bir açıklama modeli oluşturur. Bu, belirli bir girdinin veya tahminin neden ve nasıl gerçekleştiğini anlamamıza yardımcı olur. LIME, rastgele veri noktalarını oluşturarak ve bu noktaların tahminlerini inceleyerek lokal bir açıklama modeli oluşturur. Bu açıklama modeli, daha basit ve anlaşılır bir yapısı olan, örneğin lineer regresyon gibi bir model olabilir. LIME, açıklanabilirlik gereksinimlerini karşılayan bir yöntemdir (88).



Şekil 3.4. LIME Adımları

3.6.2. SHapley Katkı Açıklamaları

SHAP, tahminlerin açıklanabilirliğini artırmak ve değişkenlerin tahmindeki katkılarını ölçmek için kullanılan bir açıklama yöntemidir. SHAP, koalisyon oyunları teorisinden esinlenen bir yaklaşım kullanır. Her bir değişkenin tahmindeki katkısını hesaplamak için SHAP, birçok farklı değişken kombinasyonunu dikkate alır ve her değişkenin tahmine ne kadar katkıda bulunduğunu hesaplar. SHAP, bir değişkenin katkısını belirlerken tüm diğer değişkenlerin durumunu göz önünde bulundurur. Bu, her bir değişkenin tahmin üzerindeki etkisini anlamamıza yardımcı olur ve karmaşık modellerin iç işleyişini daha anlaşılır kılar. SHAP, açıklanabilirlik ve model yorumlanabilirliği konusundaki ihtiyaçları karşılamak için yaygın olarak kullanılır (88).



Şekil 3.5. Kara Kutu Modeli ve SHAP yöntemi

3.7. Veri analizi

3.7.1. Biyoistatistiksel Analizler

Bu tez çalışması kapsamında, metabolomik analizler sonrasında medyan normalizasyonu, log transformasyonu ve otomatik standartlaştırma uygulanan veriler ortalama ve standart sapma değerlerini ifade eden kutu grafiği ile özetlenmiştir. Veri analizi gerçekleştirilmeden önce veri setinde var olan eksik/kayıp veya hatalı/yanlış değerler, aşırı uç değerlerin belirlenmesine yönelik analizler yapılmıştır.

Verilerin analizinde kullanılacak testin belirlenmesinin varsayımlarından biri olan normal dağılıma uygunluk Shapiro-Wilk testi ile değerlendirilirken, varyansların homojenliğinin kontrolü ise Levene testi ile değerlendirilmiştir. Veriler normal dağıldığı için, bağımsız örneklerde t-testi ile analiz edilmiştir. $p \leq 0.05$ değeri istatistiksel olarak önemli kabul edilmiştir.

ML sürecinde sağlam bir model elde edebilmek için mümkün olduğu kadar çok veri kullanmak gerekir. Veri boyutu yeterince büyük değilse ki bu metabolomik verilerde yaygın bir durumdur, kurulan modeller aşırı uyumlu olma eğilimindedir ve sonuçlar yanıltıcı olabilir. Potansiyel biyobelirteçlerin çıkarımını anlamlı kılmak için çapraz doğrulama, tekrarlı holdout, randomizasyon testi ve önyükleme gibi yeterli model doğrulaması yapılmalıdır. Bu tür yeniden örnekleme tekniklerini uygularken önemli aşamalar, yeterli yeniden örnekleme birimini ve uygun istatistiklerin seçimini içerir. Bu tez çalışmasında oluşturulacak tahmin modellerini doğrulamak için 100 tekrarlı holdout yaklaşımı kullanıldı. 89 hastayı 4:1 oranında bir eğitim seti ve bir test setine bölmek için tabakalı rastgele örnekleme yöntemi kullanıldı. Tekrarlanan 10 kat çapraz doğrulamaya sahip ızgara arama yöntemi, ML modellerinin hiperparametrelerini optimize etmek için kullanıldı. Son olarak her modelin performansı değerlendirildi ve test seti üzerinde karşılaştırıldı.

Daha sağlam bir tahmin performans elde etmek, taraflı sonuçları raporlamaktan kaçınmak ve aşırı uyumu sınırlamak için doğrulama yöntemi farklı rastgele tohumlarla 100 kez tekrarlandı ve bu 100 tekrar üzerinden ortalama performans hesaplandı.

Tez çalışması kapsamında yapılan istatistiksel ve biyoinformatik analizlerin gerçekleştirilmesinde R programındaki MetaboAnalyst paketi ve modelleme aşamasında kullanılan yöntemlerin uygulanmasında Python Versiyon 3.11.5 yazılımı kullanıldı.

3.7.2. Biyoinformatik Analizler

Kimyasal benzerliklerine veya diğer ilgili özelliklerine dayalı olarak metabolitler arasındaki ilişkileri araştırmak ve analiz etmek için ilk olarak benzerlik ağ analizi yapıldı. Metabolomik ifade seviyeleri incelenen PKOS'lu hastalar ile kontrol grubu hastaların örnekleri için, R programlama dili ile geliştirilen MetaboAnalyst paketi kullanılarak diferansiyel ifade analizleri yapıldı. Diferansiyel ifade analizi, iki farklı koşul, tedavi veya grup arasındaki ifade aktivitelerinde nicel farklılıkları bulmak için kullanılan istatistiksel bir yaklaşımdır. Veriler, çok değişkenli analiz için medyan ve pareto ölçeği kullanılarak normalleştirildi.

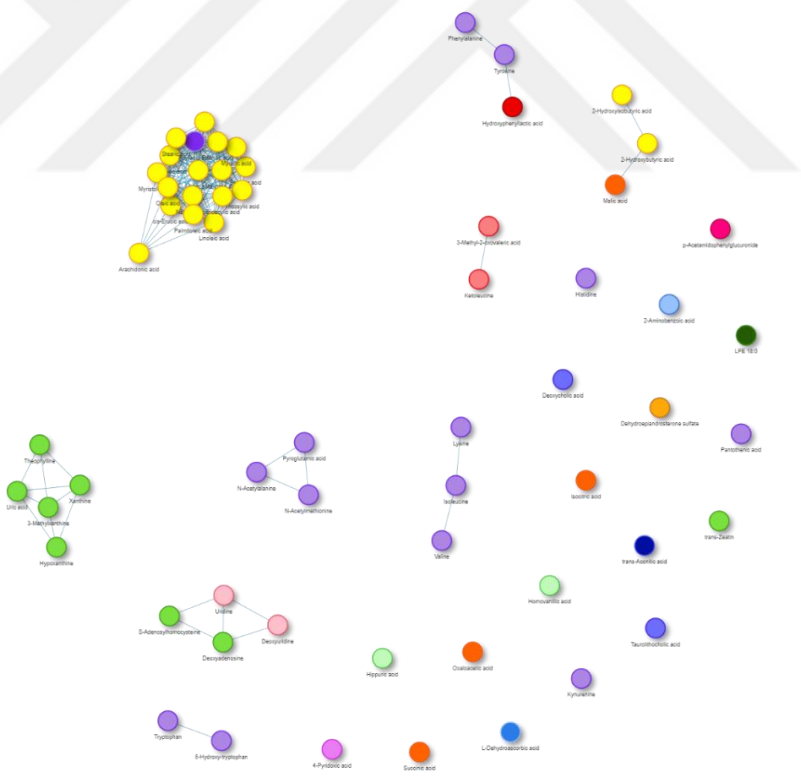
Metabolit seviyelerindeki önemli değişiklikler t testi kullanılarak test edildi ve yanlış pozitifleri en aza indirmek için yanlış keşif oranları (FDR) Benjamini-Hochberg prosedürüne göre belirlendi. PKOS hastalarından ve sağlıklı bireylerden alınan metabolitler arasındaki farkları incelemek için kat değişiklikleri (FC'ler) hesaplandı. FDR'ye göre ayarlanmış p değerleri ≤0.05 ve FC ≥1.2 (yukarı regüle edilmiş) veya FC ≤1.2 (aşağı regüle edilmiş) olan metabolitler PKOS'da biyobelirteç adayı kabul edildi. Keşifsel biyobelirteç analizi için bir Volcano grafiği ile sağlıklı kontrollere kıyasla PKOS hastalarında yukarı veya aşağı regüle edilen metabolitler görselleştirildi. Grup ayrımcılığına katkıda bulunan metabolik imzayı belirlemek ve potansiyel biyobelirteçlerin PKOS'u ayırt etmedeki öngörü performansını değerlendirmek için PLS-DA analizinden elde edilen en önemli ilk 15 metabolit, VIP> 1 temel alınarak seçildi. Verilerin hiyerarşik kümelenmesi için ilk 25 önemli metabolitin ısı haritası çizildi. Isı haritası, haritadaki her renkli hücrenin verilerdeki metabolitlerin logaritmik olarak normalleştirilmiş konsantrasyonuna karşılık geldiği, verilerin sezgisel bir şekilde görselleştirilmesini sağlar. Isı haritasındaki satırlar bireysel metabolitleri, sütunlar ise biyolojik kopyaları temsil eder. Mesafe, Öklid Korelasyonları ve Ward kümeleme algoritması kullanılarak ölçüldü. Son olarak PKOS'da önemli moleküler yolakları bulmak için yolak zenginleştirme analizi uygulandı.

4. BULGULAR

4.1. Biyoinformatik Analizlere İlişkin Bulgular

Metabolitlere ilişkin moleküler yapı benzerliklerinin görselleştirilmesi için Şekil 4.4'de ağ analizi sonuçları sunulmaktadır. Bu yapı benzerliği ağ grafiği, benzerlik katsayısı hesaplamak için MACCSkeys parmak izi türü ve Tanimoto benzerlik yöntemi kullanılarak metabolit listesinden türetilmiştir. Bu görselleştirme yöntemi, metabolitler arasındaki moleküler benzerliklerin ve grup içindeki düzenlemelerin anlaşılmasına olanak tanır.

Ağ grafiği üzerinde, her bir düğüm (node), bir metaboliti temsil eder ve bu düğümler arasındaki bağlantılar, metabolitler arasındaki moleküler benzerlikleri gösterir. Benzer moleküler yapıya sahip metabolitler arasındaki bağlar daha güçlü ve belirgin olarak görselleştirilir. Bu yöntem, metabolitler arasındaki benzerlik ve farklılıkların hızlı bir şekilde değerlendirilmesine olanak sağlar.



Şekil 4.1. Metabolitlere İlişkin Moleküler Yapı Benzerlik Ağ Analizi Sonuçları

Tablo 4.1'de metabolomik veri setine uygulanan kat değişimi analizi sonuçları sunulmaktadır.

Bu analiz, FDR değeri 0.05'ten küçük ve FC değeri 1.2 olarak alınarak PKOS ve kontrol grupları arasında farklılık gösteren metabolitlerin belirlenmesine, gruplar arasındaki moleküler farklılıkların anlaşılmasına ve farklılık gösteren metabolitlerin ekspresyon yönünün belirlenmesine olanak sağlar.

Tablo 4.1. Biyoinformatik Analiz Bulguları

Metabolit İsmi	FC	log_2FC	Düzeltilmiş p Değeri	log_{10p}	Diferansiyel Ekspresyon
Taurolitokolik asit (M-H)-	0.63	-0.67	1.10x10-4	3.96	Aşağı
Valin (M-H)-	0.59	-0.76	1.92x10-4	3.72	Aşağı
Hipoksantin (M-H)-	0.61	-0.71	1.92x10-4	3.72	Aşağı
Homovanilik asit (2M-H)-	0.73	-0.44	1.92x10-4	3.72	Aşağı
S-Adenosil-L-Homosistein (M-H)-	0.52	-0.94	5.44x10-4	3.26	Aşağı
Tridekanoik asit (M-H)-	1.72	0.78	7.27x10-4	3.14	Yukarı
Deoksiüridin (M-H)-	0.63	-0.66	0.017	1.76	Aşağı
Süksinik asit (M-H)-	1.31	0.39	0.024	1.61	Yukarı
18:0 Lizo-Pe (M-H)-	0.78	-0.35	0.024	1.61	Aşağı

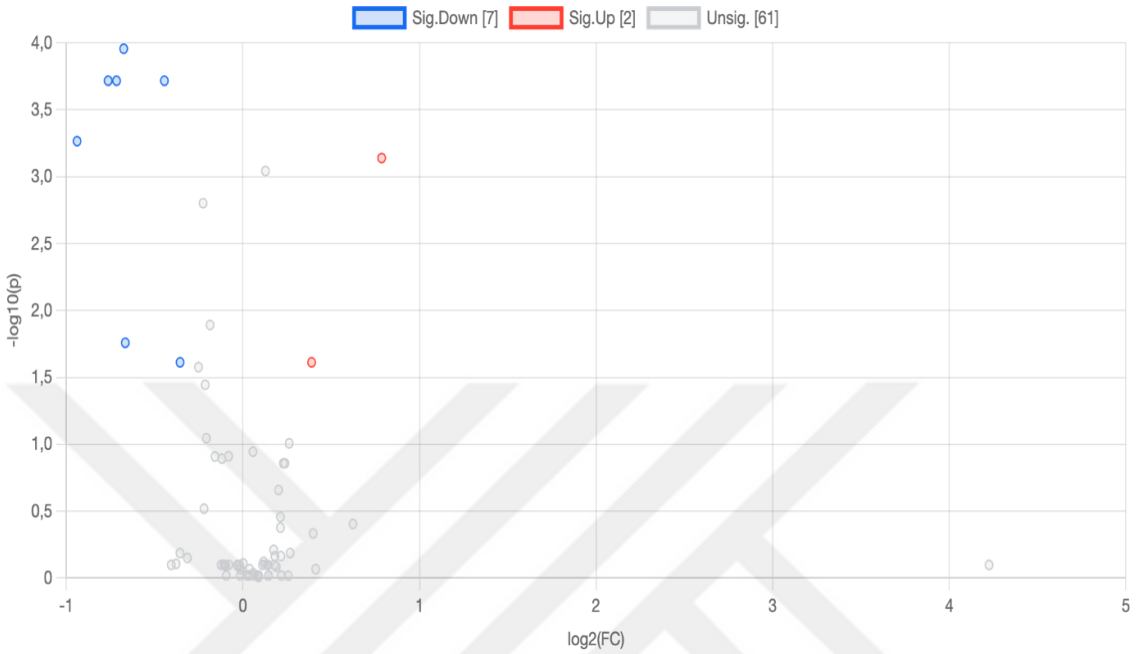
Tablo 4.1'de metabolomik veri setine uygulanan FC analizi sonrasında FDR değeri 0.05'den küçük ve FC değeri 1.2 olarak alındığında 9 metabolitin PKOS ve kontrol gruplarında farklı ekspresyonlara sahip olduğu bulunmuştur. Sonuçlar incelendiğinde, bu 9 metabolitin 2 tanesinin yukarı yönlü ekspresyon gösterdiği görülmektedir (FC > 1.2). Bu metabolitler, PKOS hastalarında kontrol grubuna göre anlamlı bir şekilde artan ekspresyon seviyelerine sahiptir. Diğer yandan, 7 metabolitin aşağı yönlü ekspresyon gösterdiği belirlenmiştir (FC < 1.2). Bu metabolitler ise PKOS grubunda kontrol grubuna göre azalan ekspresyon seviyelerini yansıtmaktadır.

Gruplar arasında diferansiyel ekspresyon gösteren metabolitleri görselleştirmek için kullanılan volkano grafiği Şekil 4.2'de verilmiştir.

Volkano grafikleri, metabolomik verilerde diferansiyel olarak ekspresyon gösteren metabolitleri görselleştirmek için log2 bazındaki kat değişimi önemini çizer.

Grafikte her bir nokta, bir metabolitin bir koşuldan diğerine olan ekspresyon farklılığını ve istatistiksel anlamlılığını temsil eder.

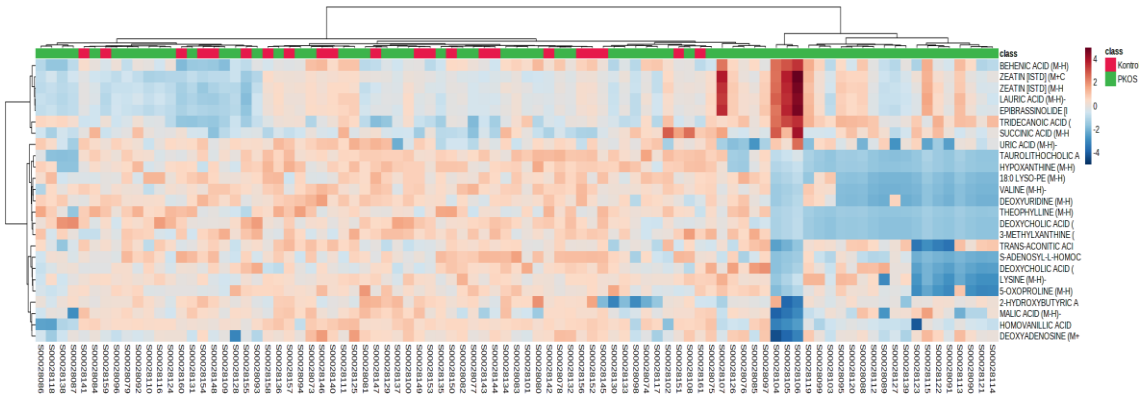
Metabolitlerin ekspresyon farklılıkları büyük ve istatistiksel olarak anlamlı ise, nokta genellikle yüksek bir konumda ve belirgin bir şekilde yer alır. Bu grafik, metabolitlerin önemli biyolojik veya fizyolojik değişiklikleri temsil edip etmediğini değerlendirmek için kullanılır.



Şekil 4.2. Diferansiyel Ekspresyon Gösteren Metabolitlere İlişkin Volkano Grafiği

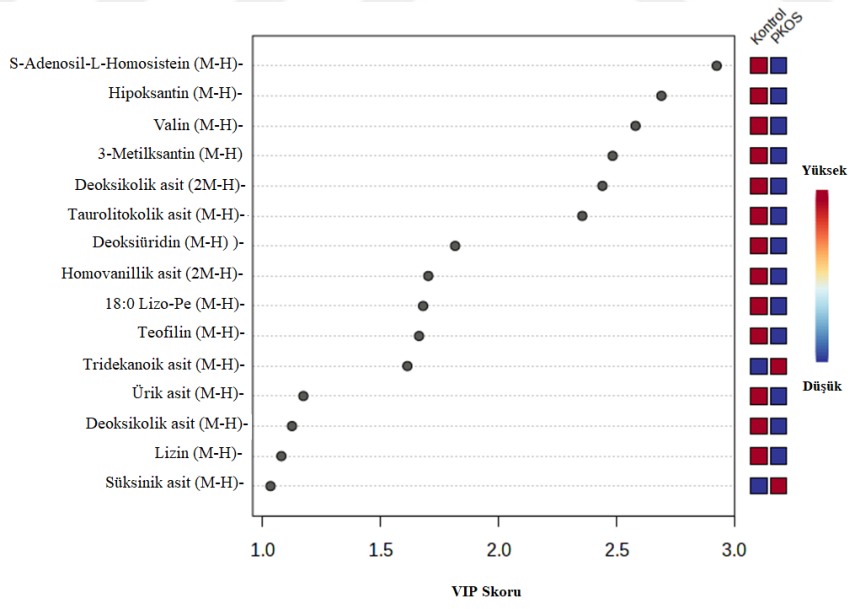
Şekil 4.3'te, gruplar arasında en belirgin farklılık gösteren 25 metabolitin ısı haritası sunulmuştur. Bu ısı haritası, PKOS ve kontrol grupları arasında en yüksek değişikliklerin gözlendiği metabolitlerin belirlenmesine olanak sağlar. Grafikte, PKOS ve kontrol grupları arasında en fazla değişiklik gözlenen 25 metabolitten fazla ekspresyon gösterenler kırmızı renk ile baskılanan ekspresyon gösterenler ise mavi renk ile görselleştirilmiştir. Bu renk kodlaması, her bir metabolitin gruplar arasındaki ekspresyon seviyelerindeki değişiklikleri kolayca ayırt etmeye olanak tanır. Kırmızı renkteki metabolitler, PKOS grubunda kontrol grubuna göre belirgin bir şekilde yüksek ekspresyon gösteren metabolitleri temsil eder. Bu durum, bu metabolitlerin PKOS patofizyolojisi üzerinde potansiyel bir etkiye sahip olduğunu gösterebilir.

Öte yandan, mavi renkteki metabolitler, PKOS grubunda kontrol grubuna göre düşük ekspresyon gösteren metabolitleri ifade eder ve bu metabolitlerin PKOS gelişimindeki olası baskılayıcı rolünü vurgular.



Şekil 4.3. İlk 25 Önemli Metabolitin Ekspresyonuna İlişkin Isı Haritası

Gruplar arasında metabolik profillerdeki farklılıkları daha kapsamlı inceleyebilmek ve daha değerli diferansiyel metabolitleri belirleyebilmek için örnek gruplamasına dayalı denetimli bir analiz protokolü yürüten PLS-DA modeline ilişkin VIP grafiği Şekil 4.4'de sunulmaktadır. VIP, PLS-DA modelindeki metabolitlerin ağırlık değerini ölçer ve her bir metabolitin ekspresyon seviyesinin etki yoğunluğunu ve her gruptaki örnek sınıflandırma/ayrımcılık konusundaki açıklayıcı yeteneğini değerlendirir. Şekil 4.4'de, her bir diferansiyel metabolitin modele genel katkısı, istatistiksel olarak anlamlı kabul edilen VIP skoru> 1 olan metabolitler ile açıklanmaktadır.



Şekil 4.4. PLS-DA Modelinde VIP Grafiği

Şekil 4.4 incelendiğinde S-Adenosil-L-Homosistein (M-H)-, Hipoksantin (M-H)- ve Valin (M-H)- en yüksek VIP skorlarına sahipti ve bu metabolitler PKOS hastaları için PLS-DA VIP yönteminde en önemli ilk üç metabolit olarak belirlendi.

Metabolitlere yapılan yolak zenginleştirme analizi, Tablo 4.2'de sunulmuştur. Bu analiz, metabolitlerin içerdikleri biyolojik bilgilerin daha derinlemesine anlaşılmasına olanak tanıyarak, çeşitli önemli biyolojik yollar belirlenmesine olanak sağlar.

Tablo 4.2. Yolak Zenginleştirme Analizi Bulguları (Kolmogorov-Smirnov testi)

Yolak (skor >3)	Metabolit İsmi	Zenginleştirme skoru (-log10(FDR-düzeltilmiş p-değeri))
Tirozin Metabolizması	Homovanilik asit (2M-H)-	7.035
	Oksaloasetik asit (M-H)-[-H2O]	
	S-Adenosil-L-Homosistein (M-H)-	
	Tirozin (M-H)-	
Sitrik Asit Döngüsü	İzositrik asit (M-H)-	3.402
	Malik asit (M-H)-	
	Oksaloasetik asit (M-H)-[-H2O]	
	Süksinik asit (M-H)-	
Arginin ve Prolin Metabolizması	Oksaloasetik asit (M-H)-[-H2O]	3.367
	S-Adenosil-L-Homosistein (M-H)-	
	Süksinik asit (M-H)-	
Betain Metabolizması	S-Adenosil-L-Homosistein (M-H)-	3.367
Karnitin Sentezi	Lizin (M-H)-	3.367
	S-Adenosil-L-Homosistein (M-H)-	
	Süksinik asit (M-H)-	
Katekolamin Biyosentezi	S-Adenosil-L-Homosistein (M-H)-	3.367
	Tirozin (M-H)-	
Estron Metabolizması	S-Adenosil-L-Homosistein (M-H)-	3.367
Glisin ve Serin Metabolizması	S-Adenosil-L-Homosistein (M-H)-	3.367
Histidin Metabolizması	L-Histidin (M-H)-	3.367
	S-Adenosil-L-Homosistein (M-H)-	
Metiyonin Metabolizması	S-Adenosil-L-Homosistein (M-H)-	3.367
Metilhistidin Metabolizması	L-Histidin (M-H)-	3.367
	S-Adenosil-L-Homosistein (M-H)-	
Nikotinat ve Nikotinamid Metabolizması	S-Adenosil-L-Homosistein (M-H)-	3.367
Fosfatidilkolin Biyosentezi	S-Adenosil-L-Homosistein (M-H)-	3.367
Ubikinon Biyosentezi	S-Adenosil-L-Homosistein (M-H)-	3.367

Tablo 4.2 incelendiğinde tirozin metabolizması, sitrik asit döngüsü, arginin ve prolin metabolizması ile betain metabolizması, PKOS ile bağlantılı en önemli metabolik yollar olarak belirlenmiştir. Tirozin metabolizması, organizmanın enerji üretimi ve hormon düzenlemesinde önemli bir rol oynayan tirozin gibi kritik bileşenler içerir. Sitrik asit döngüsü, hücresel enerji üretiminin merkezi bir aşamasını oluştururken, arginin ve prolin metabolizması, nitrik oksit sentezi ve protein sentezi gibi temel biyolojik süreçlere katkıda bulunan arginin ve prolinin metabolizmasını içerir. Betain metabolizması ise metionin döngüsü içinde yer alarak, metil grubu transferi ve metilasyon reaksiyonlarına katılarak gen ekspresyonunu düzenler. Bu belirlenen önemli metabolik yollar, PKOS'un moleküler düzeydeki temellerini anlamak ve hastalığın gelişim mekanizmalarını keşfetmek için değerli ipuçları sağlamaktadır. Bu yolaklar, potansiyel biyobelirteçlerin ve tedavi hedeflerinin tanımlanmasında önemli bir rol oynayabilir.

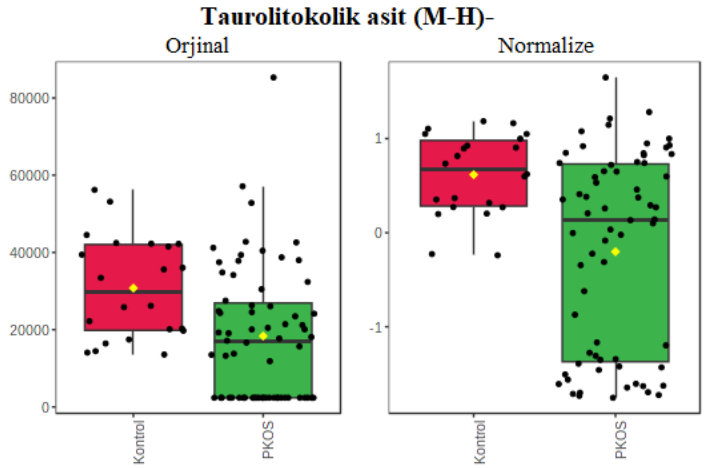
Tablo 4.3'de PKOS hastalarında biyobelirteç keşfi sürecinde kullanılan FC, PLS-DA VIP, Boruta ve MRMR yaklaşımlarının herbirinde belirlenen metabolitler sunulmaktadır. Metabolomik analizlerde biyobelirteç keşfi önemli bir adımdır. Bu nedenle bu tez çalışmasında yukarıda belirtilen dört farklı yaklaşımda ortak olan metabolitler PKOS için biyobelirteç adayı olarak seçilmiştir. Sonuçlar incelendiğinde Taurolitokolik asit (M-H)-, Hipoksantin (M-H)-, Homovanilik asit (2M-H)-, S-Adenosil-L-Homosistein (M-H)-, Tridekanoik asit (M-H)- ve Süksinik asit (M-H)- PKOS'da yenilikçi biyobelirteç adayları olarak belirlenmiştir. Sonraki tahmin modelleri ve XAI analizlerinde bu biyobelirteçler kullanılmıştır.

Tablo 4.3. Biyobelirteç Keşfi Sürecinde Kullanılan Yöntemlerde Belirlenen Önemli Metabolitler

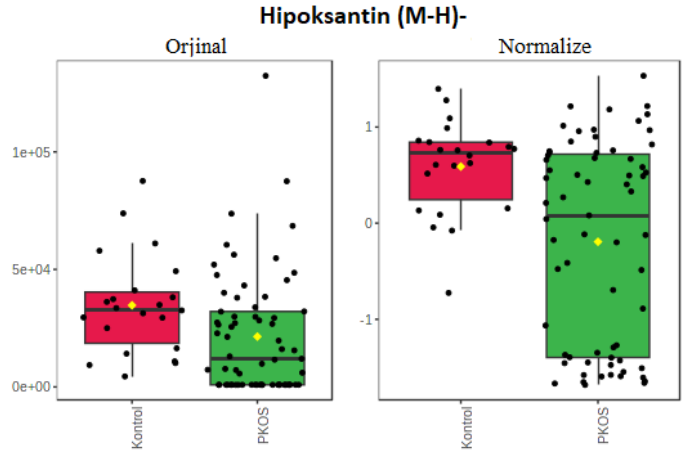
Yöntem	Önemli Metabolitler
FC	Taurolitokolik asit (M-H)-, Valin (M-H)-, Hipoksantin (M-H)-, Homovanillik asit (2M-H)-, S-Adenosil-L-Homosistein (M-H)-, Tridekanoik asit (M-H)-, Deoksiüridin (M-H))-, Süksinik asit (M-H)-, 18:0 LİSO-PE (M-H)-
PLS-DA VIP	S-Adenosil-L-Homosistein (M-H)-, Hipoksantin (M-H)-, Valin (M-H)-, 3-Metilksantin (M-H), Deoksikolik asit (2M-H)-, Taurolitokolik asit (M-H)-, Deoksiüridin (M-H))-, Homovanillik asit (2M-H)-, 18:0 Lizo-Pe (M-H)-, Teofilin (M-H)-, Tridekanoik asit (M-H)-, Ürik asit (M-H)-, Deoksikolik asit (M-H)-, Lizin (M-H)-, Süksinik asit (M-H)-
Boruta	Hipoksantin (M-H)-, Deoksiüridin(M-H)-, Homovanillik asit (2M-H), İzositikasit(M-H)-, Lauricasit(M-H)-, Fenilalanin(M-H)-, Süksinik asit (M-H)-, Teofilin(M-H)- , Tridekanoik asit (M-H)-, S-Adenosil-L-Homosistein (M-H)-, Taurolitokolik asit (M-H)-
MRMR	Homovanillik asit (2M-H)-, İzositrik asit (M-H)-, Tridekanoik asit (M-H)-, Taurolitokolik asit (M-H)-, 3-(4-Hidroksifenil) Laktik asit (M-H)-, S-Adenosil-L-Homosistein (M-H)-, Ksantin (M-H)-, Üridin (M-H)-, Deoksiadenosin (M+CL)-, Laurik asit (M-H)-, Triptofan (M-H)-, Zeatin [ISTD] (M-H)-, Tirozin (M-H)-, Ürik asit (M-H)-, Lizin (M-H)-, İzolösin (M-H)-, Süksinik asit (M-H)-, Fenilalanin (M-H)-, Deoksiüridin (M-H)-, 18:0 Lizo-Pe (M-H))-, Teofilin (M-H)-, P-Asetamidofenil Beta-D-Glukuronid (M-H)-, Fitanik asit (M-H)-, Hipoksantin (M-H)-

4.2. Biyoistatistiksel Analiz ve Ağaç Tabanlı Makine Öğrenmesi Modellerine İlişkin Bulgular

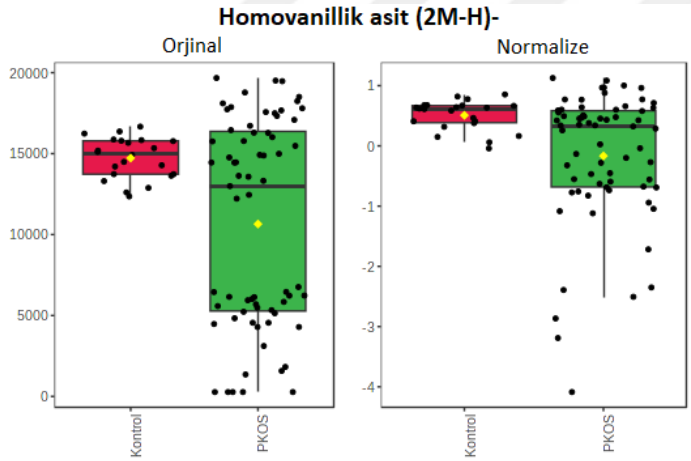
Çalışmada PKOS durumu ile ilişkili olabilecek 6 metabolit belirlenmiştir. Şekil 4.5, 4.6, 4.7, 4.8, 4.9, 4.10'da bu metabolitlerim orjinal ve normalize edilmiş konsantrasyonlarına ilişkin kutu grafikleri verilmiştir.



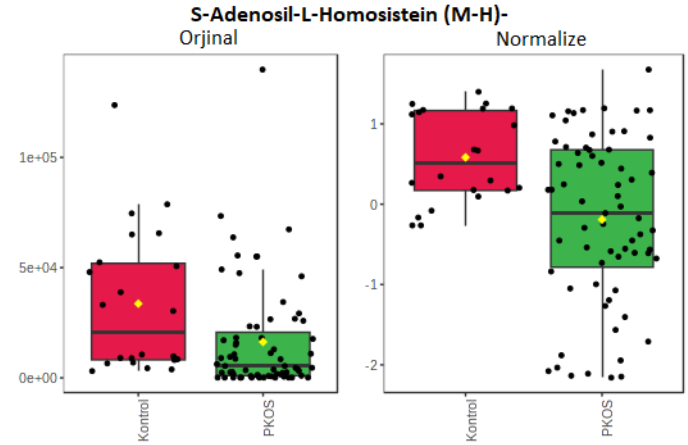
Şekil 4.5. Taurolitokolik asit (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği



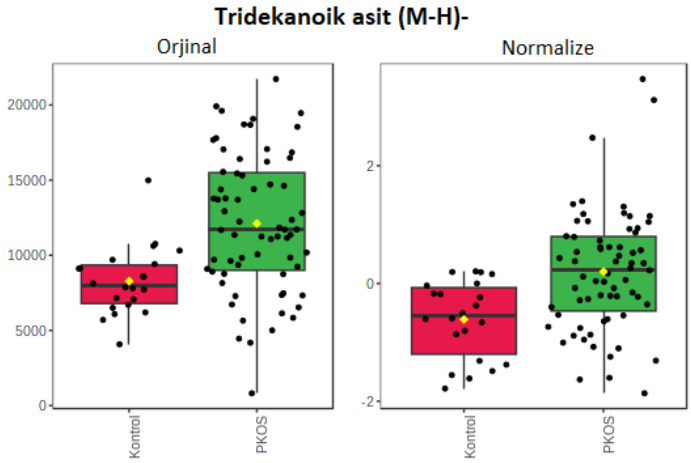
Şekil 4.6. Hipoksantin (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği



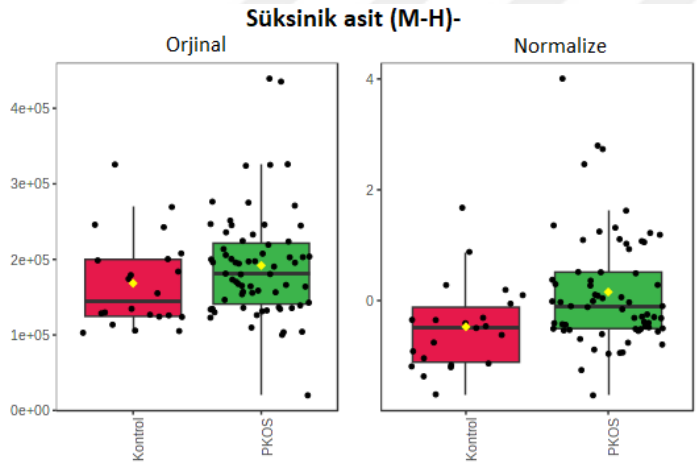
Şekil 4.7. Homovanillik asit (2M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği



Şekil 4.8. S-Adenosil-L-Homosistein (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği



Şekil 4.9. Tridekanoik asit (M-H)- için orjinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği



Şekil 4.10. Süksinik asit (M-H)- için orijinal ve normalize konsantrasyon seviyelerine ilişkin kutu grafiği

İstatistiksel analizlerin sonucunda Taurolitokolik asit (M-H)- (t=3.543; p<0.001), Hipoksantin (M-H)- (t=3.365; p<0.001), Homovanilik asit (2M-H)- (t=2.853; p<0.001), S-Adenosil-L-Homosistein (M-H)- (t=3.327; p=0.001), Tridekanoik asit (M-H)- (t=-3.502; p<0.001) ve Süksinik asit (M-H)- (t=-2.639; p<0.001) metabolitlerinin PKOS ve kontrol grupları arasında farklı konsantrasyon gösterdiği belirlendi.

Bu tez çalışmasında yüksek genelleme yetenekleri ve güçlü performanslarından dolayı 6 farklı ağaç tabanlı ML algoritması PKOS tahmininde kullanıldı. Tablo 4.4'de her bir tahmin modeline ilişkin doğruluk, duyarlılık, seçicilik, F1 skor, AUC ve Brier skoru sunulmaktadır.

Tablodan, XGBoost modelinin PKOS'u tahmin etmek için en iyi performansı elde ettiği görülmektedir. XGBoost modelinin doğruluk, duyarlılık, seçicilik, F1 skoru ve AUC değerleri, diğer modellerden belirgin şekilde daha yüksektir.

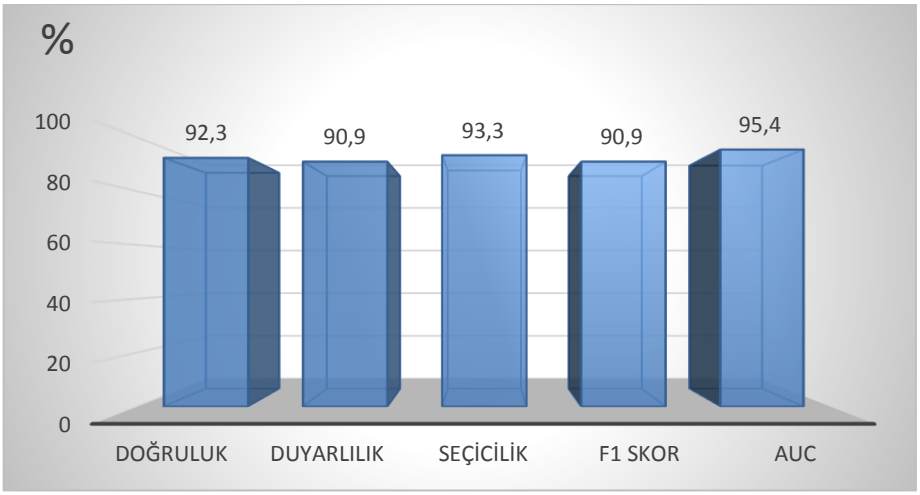
XGBoost modeli [%92.3 (%82.1-%100)] doğruluk ve [%95.4 (%92.1-%98.8)] AUC değerleri elde etmiş olup aynı zamanda model en yüksek duyarlılık [%90.9 (%58.7-%99.8)] skoruna sahipti.

Ayrıca XGBoost modeli en düşük Brier skorunu elde etmiştir, bundan dolayı kalibrasyonu en iyi olan modelin XGBoost olduğu söylenebilir. NGBoost ve LightGBM modelleri de PKOS tahmininde iyi performans göstermiştir ancak bu modellerin doğruluk, duyarlılık ve F1 skoru değerleri XGBoost modelinden biraz daha düşüktür. AdaBoost, CatBoost ve Rastgele orman modelleri ise PKOS hastalığını tahmin etmek için diğer üç modelden daha düşük performans göstermiştir.

Tablo 4.4. Ağaç Tabanlı Makine Öğrenmesi Modellerin Performans Ölçütlerine İlişkin Bulgular

Yöntem	Doğruluk	Duyarlılık	Seçicilik	F1 skor	AUC	Brier Skor
NGBoost	%88.5 (%72.4-%99.8)	%90.0 (%55.5-%99.7)	%87.5 (%61.7-%98.4)	%85.7 (%72.3-%99.2)	%93.1 (%88.7-%97.5)	0.036
XGBoost	**%92.3 (%82.1-%100)**	**%90.9 (%58.7-%99.8)**	**%93.3 (%68.1-%99.8)**	**%90.9 (%79.9-%98.2)**	**%95.4 (%92.1-%98.8)**	**0.023**
LightGBM	%88.9 (%76.2-%100)	%83.3 (%51.6-%97.9)	%92.9 (%66.1-%99.8)	%87.0 (%74.0-%99.9)	%92.8 (%88.1-%97.6)	0.028
AdaBoost	%84.6 (%70.7-%98.5)	%76.9 (%46.2-%95.0)	%92.3 (%64.0-%99.8)	%83.3 (%69.0-%97.7)	%90.5 (%84.6-%94.4)	0.231
CatBoost	%84.6 (%62.3-%98.2)	%81.8 (%48.2-%97.7)	%86.7 (%59.5-%98.3)	%81.8 (%67.0-%96.6)	%89.7 (%84.2-%95.1)	0.242
Rasgele Orman	%88.5 (%76.2-%100)	%83.3 (%51.6-%97.9)	%92.9 (%66.1-%99.8)	%87.0 (%74.0-%99.9)	%92.8 (%88.1-%97.6)	0.041

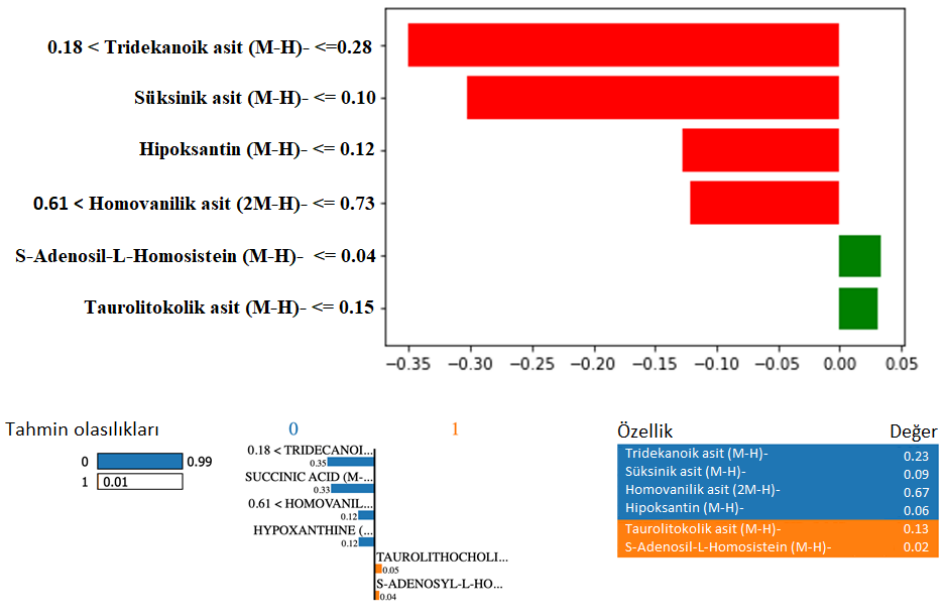
XGBoost modeline ilişkin performans ölçütlerinin grafiksel gösterimi Şekil 4.11'de sunulmaktadır.



Şekil 4.11. XGBoost Modelinin Performans Ölçütlerine İlişkin Grafiksel Gösterim

4.3. Açıklanabilir Yapay Zekâ Yöntemlerine İlişkin Bulgular

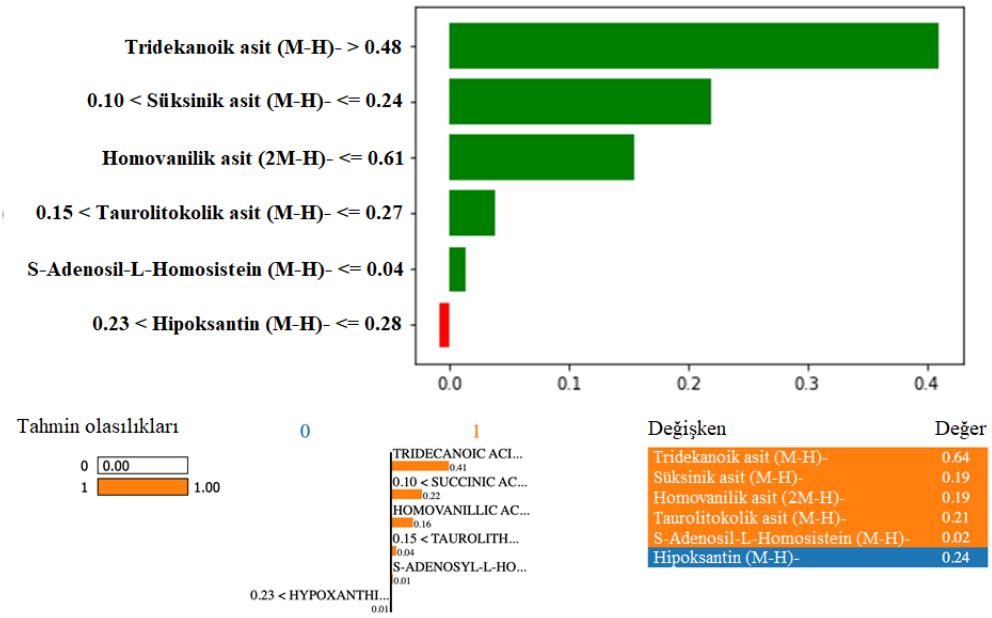
Çalışmada ağaç tabanlı modeller arasında en iyi performansı elde eden XGBoost'u açıklayabilmek için yerel LIME ve küresel SHAP grafikleri uygulanmıştır. Şekil 4.12 ve Şekil 4.13'da sırasıyla XGBoost modelinin doğru negatif ve doğru pozitif olarak tahmin ettiği hastalara ilişkin LIME yerel açıklama sonuçları verilmiştir. Şekil 4.12 ve Şekil 4.13'da yeşil çubuklar PKOS ile pozitif korelasyona sahip metabolitleri gösterirken, kırmızı çubuklar PKOS ile negatif korelasyona sahip metabolitleri göstermektedir.



Şekil 4.12. XGBoost Modelinin Tahmin Sonuçlarına Göre Doğru Negatif (Kontrol) Bir Hasta için LIME Açıklamaları

Şekil 4.12'de hastanın XGBoost modeli %99 olasılık ile PKOS olmadığı tahmin edilmiştir.

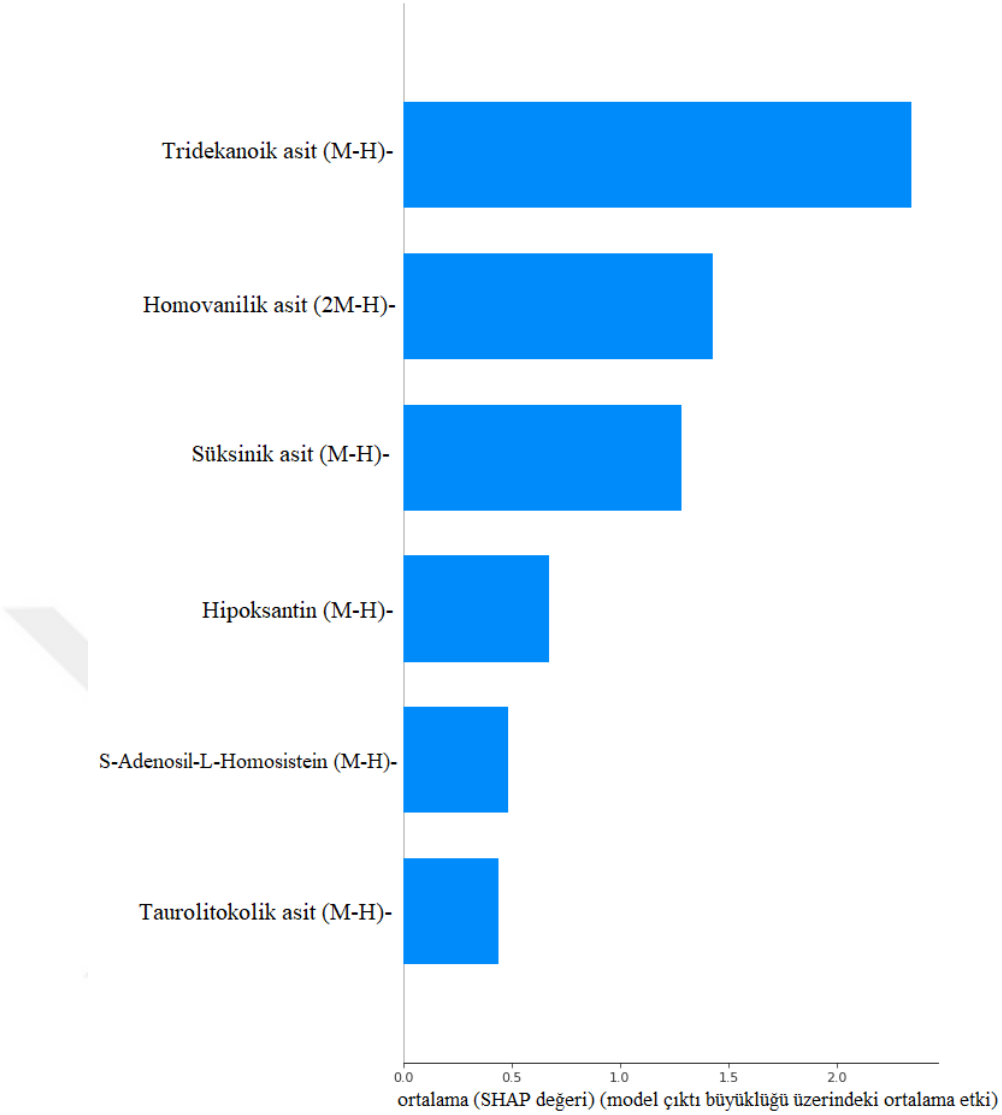
Bu hastanın doğru negatif olarak tahmin edilmesinin sebebi Tridekanoik asit (M-H)-, Süksinik asit (M-H)-, Homovanilik asit (2M-H)- ve Hipoksantin (M-H)- metabolit düzeyleridir. Buna karşılık hastanın S-Adenosil-L-Homosistein (M-H)- ve Taurolitokolik asit (M-H)- metabolit düzeyleri PKOS olasılığını arttırmıştır.



Şekil 4.13. XGBoost Modelinin Tahmin Sonuçlarına Göre Doğru Pozitif (PKOS) Bir Hasta için LIME Açıklamaları

Şekil 4.13'te hastanın XGBoost modeli %100 olasılık ile PKOS olduğu tahmin edilmiştir. Bu hastanın doğru pozitif olarak tahmin edilmesinin sebebi Tridekanoik asit (M-H)-, Süksinik asit (M-H)-, Homovanilik asit (2M-H)-, Taurolitokolik asit (M-H)- ve S-Adenosil-L-Homosistein (M-H)- metabolit düzeyleridir.

XGBoost modelinde metabolitlerin küresel olarak PKOS riskini nasıl etkilediğini görsel olarak açıklamak için SHAP yöntemi kullanılmıştır. Şekil 4.14, normalize edilmiş SHAP değeri tarafından değerlendirilen PKOS'a ilişkin metabolit risk faktörlerini ve bunların önemlilik sıralamalarını göstermektedir. Grafikte metabolit önemlilikleri azalan sırada verilmektedir ve bir metabolitin SHAP değeri ne kadar yüksek ise hastanın PKOS olma olasılığı da o kadar yüksektir.

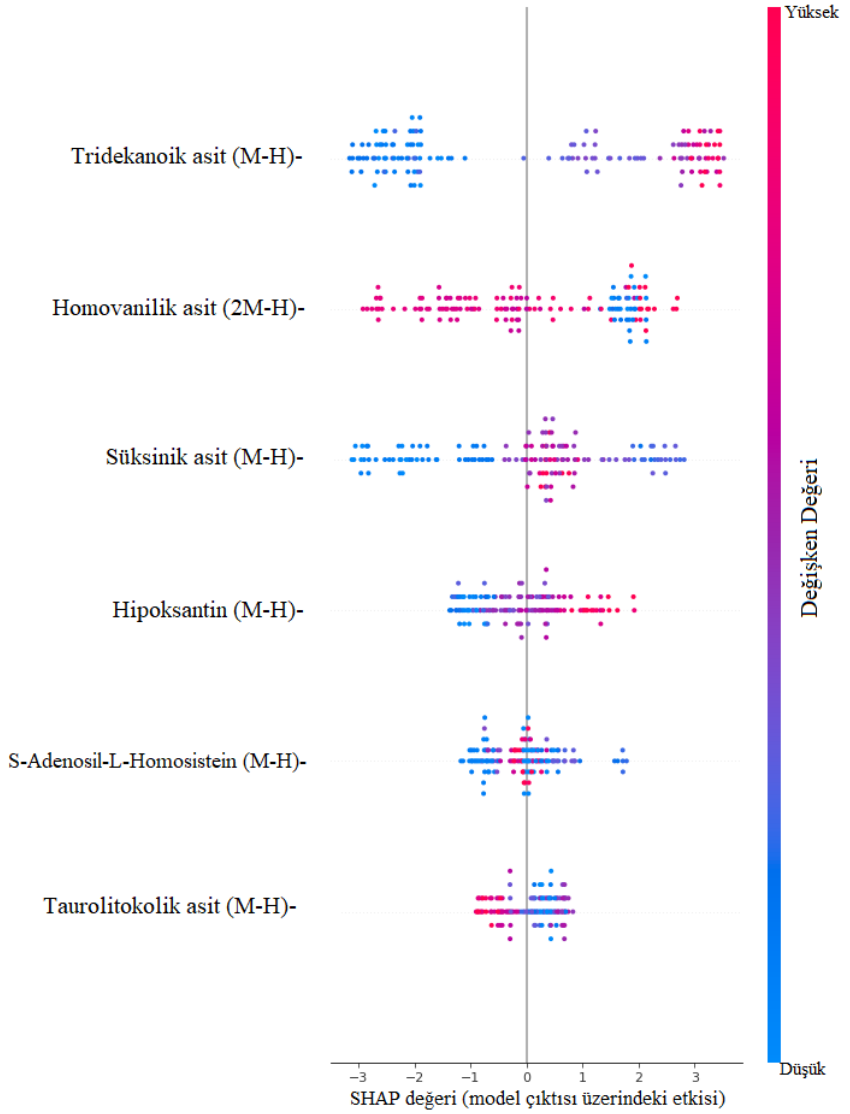


Şekil 4.14. Biyobelirteç Adayı Metabolitlerin Ortalamaya Göre Önem Sırası (|SHAP değeri|)

Şekil 4.14'ün detaylı bir analizi, PKOS ile ilişkilendirilen potansiyel biyobelirteçler arasında öne çıkan üç önemli belirleyici risk faktörünü ortaya koymaktadır. Bu faktörler, sırasıyla, Tridekanoik asit (M-H)-, Homovanilik asit (2M-H)- ve Süksinik asit (M-H)- metabolit değerleridir. Bu metabolitlerin PKOS ile ilişkisinin anlaşılması, hastalığı moleküler düzeyde anlamak ve daha etkili tanı yöntemleri geliştirmek için önemlidir.

Şekil 4.15, biyobelirteç adayı metabolitlerin optimal modelin tahmin kararına olumlu veya olumsuz katkılarını görselleştirmek amacıyla küresel SHAP değerlerini kullanmaktadır. Bu değerler, incelenen metabolitlerin, PKOS hedef değişkenine olan katkılarını net bir şekilde ortaya koymaktadır.

Pozitif bir SHAP değeri, belirli bir metabolitin PKOS'un oluşumu üzerinde olumlu bir etkisi olduğunu gösterirken, negatif bir SHAP değeri bu etkinin olumsuz olduğunu ifade eder. Grafikteki noktaların renkleri, metabolitlerin seviyelerini temsil eder. Maviye yaklaştıkça, metabolitin değeri azalırken, pembe tonlarına yaklaştıkça artar. Bu renk skalası, metabolitlerin konsantrasyonlarındaki değişimleri açıkça göstererek, PKOS riskini belirlemede önemli olan metabolit seviyelerindeki varyasyonları vurgular.



Şekil 4.15. XGBoost modelinin SHAP özet grafiği ile yorumlanması

Şekil 4.15 incelendiğinde, Tridekanoik asit (M-H)-, Süksinik asit (M-H)- ve Hipoksantin (M-H)- seviyelerindeki artışın, aynı zamanda Homovanilik asit (2M-H)- ve Taurolitokolik asit (M-H)- seviyelerindeki azalmanın, PKOS riskini arttırdığı gözlemlenebilir.

Tridekanoik asit (M-H)-, PKOS'lu bireylerde belirgin bir şekilde artmıştır ve bu durum, lipid metabolizması üzerindeki olası etkilerini vurgular.

Homovanilik asit (2M-H)-, sinir sistemi ve dopamin düzenlemesiyle ilişkilendirilmiş olup, PKOS gelişimindeki nöroendokrin faktörlerin rolünü gösterebilir. Süksinik asit (M-H)- ise enerji metabolizması ve mitokondrial aktivite ile bağlantılıdır ve PKOS'un altında yatan metabolik değişimlere işaret edebilir. Bu bulgular, belirli biyolojik bileşenlerin varyasyonlarının PKOS gelişimi üzerindeki potansiyel etkilerini vurgulamaktadır. Bu nedenle, moleküler düzeydeki bu değişikliklerin, PKOS'un anlaşılmasında önemli bir rol oynayabileceği ve tedavi stratejilerinin geliştirilmesinde yol gösterici olabileceği düşünülmektedir.

Tablo 4.5'de, metabolitlerin normalize edilmiş SHAP değerlerine göre önem düzeyleri sunulmaktadır.

Tablo 4.5. Normalize edilmiş SHAP değerlerine göre metabolitlerin PKOS tanısındaki önem düzeyleri

Metabolit	**Metabolit Önemi**
Tridekanoik asit (M-H)-	0.353
Homovanilik asit (2M-H)	0.214
Süksinik asit (M-H)-	0.193
Hipoksantin (M-H)-	0.101
S-Adenosil-L-Homosistein (M-H)-	0.073
Taurolitokolik asit (M-H)-	0.066

Tablo 4.5'e göre, Tridekanoik asit (M-H)-, PKOS tanısında en önemli metabolittir ve önemlilik oranı 0.353 olarak bulunmuştur. Diğer önemli metabolitler'den Homovanilik asit (2M-H), Süksinik asit (M-H)- ve Hipoksantin (M-H)- için önemlilik oranları sırasıyla 0.214, 0.193 ve 0.101 idi.

5. TARTIŞMA

PKOS'un günümüzdeki tanı kriterleri hala tartışılmakta ve önerilen tanı kriterleri bu sendroma sahip bireylerde oldukça heterojen dağılıma neden olmaktadır. Bu nedenle, bu hastalığın tanısı için güvenilir yeni biyokimyasal veya moleküler belirteçlere büyük ölçüde ihtiyaç vardır. Ayrıca, PKOS'u tanımlamak için kullanılabilecek moleküler biyobelirteç(ler) bulunabilirse, PKOS daha fazla bir sendrom olarak adlandırılmaktan ziyade ovaryan ve metabolik bir hastalık olarak tanımlanabilir.

Günümüzde PKOS'da metabolomik çalışmalar oldukça kısıtlıdır. Literatürdeki çalışmalarda PKOS'da en çok karbonhidrat, yağ ve protein metabolizması ile ilgili metabolitlerde bozukluklar saptanmıştır (89-91). Çoğu PKOS hastasında karbonhidrat metabolizması bozulmuştur (92). Obez olmayan PKOS hastalarında, laktat ve glukoneojenik aminoasitlerde belirgin artış ve plazma glukozunda azalma, çizgili kasta artmış glikolizis ve karaciğerde azalmış glukoneojenez ile açıklanabilir. Lipit metabolizması açısından da lipoprotein seviyeleri, düşük molekül ağırlıklı lipoprotein kolesterol ve çok düşük molekül ağırlıklı lipoprotein kolesterol seviyeleri kontrol grubuna göre PKOS hastalarında yüksek olduğu bildirilmiştir. Bu duruma kolesterol ve yüksek molekül ağırlıklı kolesterol seviyelerinde düşüklük eşlik etmektedir (89, 91). Serbest yağ asitlerinin (palmitik, stearik ve linoleik asit gibi) PKOS hastalarında kontrollere göre yükseldiği bildirilmiştir. Bu durum yağ dokusundaki insülin direnci ile bağlantılı bir şekilde lipolisizin artmış olduğuna işaret eder (93, 94). Ayrıca linoleik asit güçlü proinflamatuvar aktiviteye sahiptir ve PKOS'da mevcut olan kronik subakut inflamasyonun sebebi olabilir (95).

Trikarboksilik asit döngüsü (TCA), aerobik organizmaların enerji üretimini sağlayan, asetil koenzim A'nın oksidizasyonu ile seyreden bir dizi kimyasal reaksiyondur. Bununla beraber, TCA döngüsüne katılan, arjinin, sitrullin, glutamin, histidin, metiyonin ve prolin gibi aminoasitler de PKOS'da azalır. Fenilalanin, treonin, tirozin ve valin ise PKOS hastalarında yüksektir (89-91). Sonuç olarak PKOS hastalarında TCA döngüsü bozulmuştur ve bu durum insülin direncinin bir kanıtıdır (96). Hiperandrojenizmin sitrat seviyelerini azaltarak TCA döngüsünü bozduğu belirtilmiştir ve sitrat seviyeleri PKOS hastalarında azalmaktadır (97).

PKOS hastalarında yüksek saptanan valin, tirozin, triptofan, fenilalanin gibi aromatik ve dallı zincirli aminoasitlerin insülin direnci ve diyabet ile ilişkisi gösterilmiştir

(98, 99). İnsülin direnci olmayan PKOS hastalarında da valin seviyeleri artmış ve glisin seviyeleri azalmıştır. Dolayısıyla valin ve glisinin insülin direncinden bağımsız metabolik bozukluklarda rolü olabileceği söylenebilir (89).

Sonuç olarak, PKOS patogenezi hala net değildir. Bunun temel nedeni altta yatan moleküler mekanizmaların tam olarak anlaşılamamasıdır. Metabolizma alanına yaklaştıkça sistemik dengesizlikler, hastalığın progresyonunu ve komplikasyonları öngörecek biyobelirteçler tanımlanabilecektir. PKOS'da protein, lipit ve karbonhidrat metabolizmasında ve TCA döngüsünde bozukluklar olduğu artık bilinmektedir. İnsülin duyarlaştırıcılar ve antiandrojenler ile PKOS hastalarındaki metabolik anormalliklerde düzelmeler saptanmıştır. Bu sonuçlar yeni biyobelirteçlerin tespitine, tanı testlerinde gelişmelere, metabolik taramalarda yeni yöntembilimlere ve tedavinin bireyselleştirilmesine temel oluşturacaktır.

Bu önemli bilgilere ek olarak, bu tez çalışmasında, PKOS'a ilişkin metabolomik seviyedeki değişiklikler incelenmiştir, daha sonra olası biyobelirteçler biyoinformatik ve AI yaklaşımlarına dayalı olarak saptanmış ve ağaç tabanlı ML yaklaşımları ve XAI'yı entegre eden yorumlanabilir bir tahmin modeli geliştirilmiştir. FC analizleri sonrasında PKOS hastalarında kontrole kıyasla Tridekanoik asit (M-H)- ve Süksinik asit (M-H)-) sırasıyla 1.72 ve 1.31 kat fazla konsantrasyon göstermiştir. Taurolitokolik asit (M-H)-, Valin (M-H)-, Hipoksantin (M-H)-, Homovanilik asit (2M-H)-, S-Adenosil-L-Homosistein (M-H)-, Deoksiüridin (M-H)- ve 18:0 Lizo-Pe (M-H)- metabolitlerinin PKOS hastalarında kontrole kıyasla sırasıyla 0.63, 0.59, 0.61, 0.73, 0.52, 0.63 ve 0.78 kat az konsantrasyon gösterdiği belirlenmiştir. Çalışmada yolak analizi sonucunda Tirozin metabolizması, Sitrik Asit döngüsü, Arginin ve Prolin metabolizması ile Betain metabolizmasının PKOS ile bağlantılı en önemli metabolik yollar olduğu bulunmuştur. Moleküler risk faktörlerini belirleyebilmek için FC, PLS-DA VIP, Boruta ve MRMR yaklaşımları ayrı ayrı metabolomik verilere uygulandı ve bu dört yaklaşımda ortak olan metabolitler PKOS için biyobelirteç olarak belirlendi. Sonuç olarak bu dört yaklaşımda ortak olan Taurolitokolik asit (M-H)-, Hipoksantin (M-H)-, Homovanilik asit (2M-H)-, S-Adenosil-L-Homosistein (M-H)-, Tridekanoik asit (M-H)- ve Süksinik asit (M-H)- PKOS'da yenilikçi biyobelirteç adayları olarak belirlenmiştir.

Sonrasında bu biyobelirteçleri kullanarak PKOS'u ayırt edebilmek için altı (XGBoost, NGBoost, LightGBM, CatBoost, AdaBoost, Rasgele orman) farklı ağaç tabanlı ML algoritması kullanıldı ve sonuçlar kapsamlı performans ölçütleri incelenerek karşılaştırıldı.

XGBoost modelinin doğruluk, duyarlılık, seçicilik, F1 skoru ve AUC değerleri, diğer modellerden belirgin şekilde daha yüksekti. XGBoost modeli %92.3 (%82.1-%100) doğruluk ve %95.4 (%92.1-%98.8) AUC değerleri elde etmiş olup aynı zamanda model en yüksek duyarlılık [%90.9 (%58.7-%99.8)] skoruna sahipti. Ayrıca XGBoost modeli en düşük Brier skorunu elde etti, dolayısıyla kalibrasyonu en iyi olan model de XGBoost idi. NGBoost ve LightGBM modelleri de PKOS tahmininde iyi performans gösterdi ancak bu modellerin doğruluk, duyarlılık ve F1 skoru değerleri XGBoost modelinden nispeten daha düşüktü. AdaBoost, CatBoost ve Rastgele orman modelleri ise PKOS hastalığını tahmin etmek için diğer üç modelden daha düşük performans gösterdi. Tüm bu bilgilerin sonucunda XGBoost PKOS tahmini için optimal model olarak belirlendi. Modelin yerel (hasta bazlı) tahminlerini açıklayabilmek için LIME yöntemi kullanıldı. %100 olasılıkla PKOS olarak tahmin edilen bir hastanın (doğru pozitif) LIME sonuçları, hastada Tridekanoik asit (M-H)- konsantrasyon seviyesinin yüksek, Süksinik asit (M-H)- konsantrasyon seviyesinin orta düzeyde ve Homovanilik asit (2M-H)-, Taurolitokolik asit (M-H)- ve S-Adenosil-L-Homosistein (M-H)- metabolitlerinin konsantrasyon seviyelerinin düşük olmasından dolayı modelde doğru pozitif olarak tahmin edildiğini açıkladı. SHAP grafikleri belirlenen biyobelirteçlere dayalı olarak küresel düzeyde modelin klinik yorumlarını elde etmek için oluşturuldu ve Tridekanoik asit (M-H)-, Homovanilik asit (2M-H)- ve Süksinik asit (M-H)- PKOS olasılığını arttıran en önemli üç metabolit olarak belirlendi. Bu metabolitlerin PKOS ile ilişkisinin anlaşılması, hastalığı moleküler düzeyde anlamak ve daha etkili tanı yöntemleri geliştirmek için önemlidir. Tridekanoik asit (M-H)-, Süksinik asit (M-H)- ve Hipoksantin (M-H)- seviyelerindeki artışın, aynı zamanda Homovanilik asit (2M-H)- ve Taurolitokolik asit (M-H)- konsantrasyon seviyelerindeki azalmanın, PKOS riskini arttırdığı gözlendi. Tridekanoik asit (M-H)-, PKOS'lu bireylerde belirgin bir şekilde arttı ve bu durum, lipid metabolizması üzerindeki olası etkilerini vurgulamaktadır. Homovanilik asit (2M-H)-, sinir sistemi ve dopamin düzenlemesiyle ilişkilendirilmiş olup, PKOS gelişimindeki nöroendokrin faktörlerin rolünü gösterebilir. Süksinik asit (M-H)- ise enerji metabolizması ve mitokondrial aktivite ile bağlantılıdır ve PKOS'un altında yatan metabolik değişimlere işaret edebilir.

Güncel bir çalışmada (10) PKOS patofizyolojisinin hala belirsizliğini koruduğu bildirilmiş olup, bu kapsamda PKOS patogenezini daha fazla araştırmak için, PKOS ve kontrol örneklerine ilişkin tüm over DNA mikrodizi gen ifade veri setlerinden hesaplanan gen set düzenliliği (GSR) indeksi kullanılarak PKOS'un sınıflandırılması amaçlanmıştır.

Söz konusu çalışmada; iki, üç ve beş kat CV kullanılarak PKOS tahmini için destek vektör makinesi modeli oluşturulmuş, ilgili modelin performansı çeşitli metrikler ile değerlendirilmiştir. Elde edilen bulgulara göre, tüm çapraz geçerlilik analizlerinde PKOS tahmini için yüksek sınıflandırma performans değerleri elde edilmiştir. Sonuç olarak, PKOS'un over fonksiyonlarının kontrollerden önemli ölçüde farklı olduğu ve PKOS'un ML ile başarılı bir şekilde sınıflandırılabileceği gösterilmiştir. Fakat ilgili çalışmada sadece PKOS ve kontrol örneklerine ilişkin gen ifade veri setleri ele alınmış ve klasik ML yöntemlerinden destek vektör makinesi kullanılarak PKOS sınıflandırması yapılmıştır.

Soni ve diğerleri tarafından yapılan bir çalışmada (100), bir derin öğrenme yaklaşımı olan evrişimsel sinir ağları modeli kullanılarak PKOS'lu olan ve olmayan bireylere ilişkin ultrason görüntülerinin sınıflandırılması yapılmıştır. Önerilen yöntemin, PKOS'lu olan ve olmayan bireylerin ultrason görüntülerinin sınıflandırmasını yüksek doğrulukla gerçekleştirdiği raporlanmıştır. Başka bir çalışmada ise (101), PKOS'lu olan ve olmayan bireylere ilişkin ultrason görüntüleri önişleme aşamasından geçirildikten sonra Gabor Dalgacık algoritması kullanılarak değişken çıkarma (feature extraction) işlemi uygulanmış, sonrasında ise Elman sinir ağları kullanılarak sınıflandırılmıştır. İlgili sistemde, PKOS'lu olan ve olmayan bireyler en fazla %78.1 doğrulukla sınıflandırmıştır. Vikas ve ark. çalışmasında (102), PKOS'lu olan ve olmayan bireylerin 17 adet değişkene (risk faktörü) dayalı olarak çeşitli ML yöntemleriyle sınıflandırılması yapılmıştır. Çalışmanın bulgularına göre, Naive Bayes algoritmasının PKOS'lu olan ve olmayan bireyleri %97.65 doğrulukla sınıflandırmıştır. PKOS'a yönelik bir derleme çalışmasında, yumurtalıkların ultrason görüntülerinden foliküllerin tespiti ve PKOS teşhisi için çeşitli bilgisayar destekli teknikler incelenmiştir. Bu çalışmanın araştırmacıları, bilgisayar destekli folikül tespit yöntemlerinin, ultrason görüntülerini kısa sürede analiz etmesi, tanı hatalarının azaltması gibi avantajlarının olduğunu bildirmiştir (103). Görüntü sınıflandırmayı inceleyen diğer bir çalışmada, PKOS'un tanısına yönelik geleneksel yöntemlerin dezavantajları dikkate alınarak, derinlemesine eğitilmiş bir görüntü kalitesi çekirdeği kullanılarak PKOS tanısı için olasılıklı bir model tasarlanmış ve elde edilen deneysel bulguların, önerilen yöntemin PKOS'un tanı ve sınıflandırılmasında etkin olduğu bildirilmiştir (10).

Yukarıda belirtilen çalışmalarda çeşitli istatistiksel öğrenme ve derin öğrenme yöntemleri kullanılarak PKOS'un farklı genomik değişkenler veya ultrason görüntüleri kullanılarak sınıflandırılması yapılmıştır. Fakat bu çalışmalarda, kanıt düzeyi daha yüksek olan metabolomik analizler ve bu analizlerden elde edilen biyobelirteçler

kullanılmamıştır. Mevcut çalışmamızda metabolomik biyobelirteçleri biyoinformatik ve AI yöntemlerine dayalı olarak belirlenmiştir ve literatürdeki çalışmalarda bulunmayan ileri ML yaklaşımları ve XAI'nın entegre kullanımı ile modeli klinik açıdan yerel ve küresel olarak açıklanmıştır.

Başka bir çalışmada (104), PKOS tanısı alan kadınların serumundaki ayırıcı metabolitleri tanımlamak ve inositollerle tedavinin PKOS'lu kadınların metabolomik yollarını ve metabolik profillerini iyileştirip iyileştiremeyeceğini tanımlamak amaçlanmıştır. Bu pilot çalışmada, PKOS ve inositol + glukomannan'ın birlikte tedavisi ile ilişkili olan karmaşık bir serum molekülleri ağı tanımlanmış ve çalışma bulgularını daha da doğrulamak için klinik ve biyokimyasal kriterler ile belirlenen PKOS hastalarını içeren daha büyük örneklemler üzerinde yapılacak klinik çalışmalara ihtiyaç olduğu belirtilmiştir.

Biyobelirteçleri ele alan bir araştırmada, PKOS'un patogenezinin belirsizliğini koruduğu bildirilmiş, PKOS'un farklı fenotipleri ve erken komplikasyonlarının prognozu için potansiyel metabolik profilleri belirlemek hedeflenmiştir. Bu çalışmada, PKOS ve uzun vadeli komplikasyonlarını öngören yeni metabolik biyobelirteçler tanımlanmıştır. Ayrıca, söz konusu çalışmanın, PKOS için yeni tedavi hedeflerinin geliştirilmesine temel oluşturabilecek olan metabolik yolakların daha iyi anlaşılmasında oldukça önemli sonuçlara sahip olduğu bildirilmiştir (89).

Lim ve ark. tarafından yapılan bir araştırmada, kısa uyku süresi, yetersiz egzersiz ve vardiyalı çalışma gibi yaşam tarzı faktörlerinin tek başına veya kombinasyon halinde, sağlıklı bir popülasyonda tipik olarak PKOS'a ait üreme ve metabolik anormallikler ile ilişkili olup olmadığını incelemiştir. Vardiyalı çalışma, egzersiz veya normal uyku süresi ile PKOS'u tanımlayan androjenik ve over ölçümleri arasında anlamlı bir ilişki gözlenmemiş olup, uyku süresi az olan kadınların menstrüel bozukluk ve insülin direnci açısından artmış risk altında olduğu bildirilmiştir (105).

Diğer bir çalışmada, PKOS hastalarından toplanan genom ölçekli DNA metilasyon verileri kullanılarak PKOS'un klinik fenotiplerinin genomik düzeyde DNA metilasyon deseni ile diferansiyel metillenmiş bölgeler ilişkilendirilerek tespit edilmiştir. İlgili çalışmada, PKOS hastaları için genomik bölge temelli DNA metilasyon profili ortaya konulmuş ve PKOS için endokrin ile ilişkili önemli diferansiyel metillenmiş bölgeler belirlenmiştir (106).

PKOS fenotiplerini ele alan bir çalışmada, PKOS fenotiplerinin sirkadiyen biyobelirteçleri olan kortizol ve melatonin ile olası ilişkileri araştırılmıştır.

Sirkadiyen ritmin iki biyolojik işareti olan serum kortizol ve melatonin ile PKOS fenotipleri arasında anlamlı bir pozitif doğrusal eğilim olduğu ve bu anlamlı ilişkilerin PKOS fenotiplerinin patogenezinde sirkadiyen biyoritimlerin rolünü araştırmak için daha kapsamlı çalışmalara ışık tutabileceği bildirilmiştir (107).

Yukarıda özetlenen araştırmalarda, PKOS ya da PKOS fenotiplerinin tanısına yönelik değişik biyobelirteçler ve etki eden faktörler, genomik veya deneysel olarak elde edilen metabolomik veriler kullanılarak incelenmiştir. Bu tez çalışmasında, metabolomik veriler ağaç tabanlı ML algoritmaları ile birlikte analiz edilerek hangi metabolitlerin hangi seviyelerinin PKOS riskini nasıl etkilediğini incelemek için XAI yöntemleri kullanılmıştır. Literatürde metabolomik biyobelirteç keşfini biyoinformatik ve AI'yı birleştiren ve kapsamlı yöntemlerle doğrulayarak XAI açısından inceleyen bir çalışma yoktur. PKOS'a yönelik sağlam bir tahmin modelinin geliştirildiği bu tez çalışmasında yerel bazda LIME ve küresel bazda SHAP açıklamalarının sonuçları klinisyenlere PKOS ön tanısında yardımcı olabilir. PKOS'a yönelik teröpatik hedeflerde ve tedavi planlamasında belirlenen biyobelirteç metabolitler faydalı olabilir.

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında, çeşitli ağaç tabanlı ML algoritmasını ve XAI yöntemlerini entegre eden bir metadoloji ile PKOS ve sağlıklı kontrollere ilişkin metabolomik biyobelirteçler kullanılarak yorumlanabilir bir sınıflandırma ve tahmin modeli geliştirilmiştir. PKOS'da olası biyobelirteçler biyoinformatik ve AI yaklaşımlarına dayalı olarak belirlenmiştir ve PKOS ile ilişkilendirilen metabolomiklerin moleküler düzeydeki değişiklikleri incelenmiştir. FC analizleri sonucunda, PKOS hastalarında kontrol grubuna kıyasla belirli metabolitlerde önemli konsantrasyon değişiklikleri tespit edilmiştir. Tridekanoik asit ve Süksinik asit, PKOS hastalarında kontrol grubuna göre belirgin şekilde yüksek konsantrasyon göstermiştir. Yolak analizi, Tirozin metabolizması, Sitrik Asit döngüsü, Arginin ve Prolin metabolizması ile Betain metabolizmasının PKOS ile bağlantılı en önemli metabolik yollar olduğunu göstermiştir. FC, PLS-DA VIP, Boruta ve MRMR yaklaşımları kullanılarak belirlenen ortak metabolitler, PKOS için potansiyel biyobelirteçler olarak önerilmiştir. Bu biyobelirteçler arasında Taurolitokolik asit, Hipoksantin, Homovanilik asit, S-Adenosil-L-Homosistein, Tridekanoik asit ve Süksinik asit bulunmaktadır. Altı farklı ağaç tabanlı ML algoritması kullanılarak PKOS'u tahmin etmek için modeller oluşturulmuştur. XGBoost modeli, diğer beş modele kıyasla daha yüksek doğruluk, duyarlılık, seçicilik, F1 skoru ve AUC değerleri elde etmiştir. Bu nedenle, XGBoost modeli PKOS tahmini için optimal model olarak belirlenmiştir. Modelin yerel (hasta bazlı) tahminlerini açıklamak belirlenen biyobelirteçlere dayalı olarak PKOS'un klinik belirtilerini yorumlayabilmek için LIME yöntemi kullanılmıştır. SHAP grafikleri, modelin küresel düzeyde klinik yorumlarını sağlamıştır. Özellikle, Tridekanoik asit, Homovanilik asit ve Süksinik asit, PKOS olasılığını arttıran en önemli üç metabolit olarak belirlenmiştir. Bu çalışmanın bulguları, PKOS'un moleküler düzeyde anlaşılmasına ve etkili tanı yöntemlerinin geliştirilmesine katkıda bulunabilir. Belirlenen biyobelirteçler, gelecekteki klinik uygulamalarda kullanılarak PKOS tanısını destekleyebilir. Sonuç olarak, bu çalışma, PKOS ile ilişkilendirilen metabolomik değişiklikleri anlamak ve tanısal bir model geliştirmek adına değerli bilgiler sunmaktadır.

KAYNAKLAR

1. Silva I, Ferreira C, Costa L, Sóter M, Carvalho L, de C. Albuquerque J, Sales MF, Candido AL, Reis FM, Veloso AA, Gomes KB. Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models. *J Endocrinol Invest* 2022, 45: 497-505.
2. Harwood K, Vuguin P, DiMartino-Nardi J. Current approaches to the diagnosis and treatment of polycystic ovarian syndrome in youth. *Horm Res Paediatr* 2007, 68 (5):209-17.
3. Azziz R, Marin C, Hoq L, Badamgarav E, Song P. Health care-related economic burden of the polycystic ovary syndrome during the reproductive life span. *J Clin Endocrinol Metab* 2005, 90(8):4650-8.
4. Zawadri, J. Diagnostic criteria for polycystic ovary syndrome: towards a rational approach. Polycystic ovary syndrome. *Curr Issues Endocrinol Metab* 1992:377-84.
5. Eshre TR, ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil Steril* 2004, 81(1):19-25.
6. Carmina E, Chu M, Longo R, Rini G, Lobo R. Phenotypic variation in hyperandrogenic women influences the findings of abnormal metabolic and cardiovascular risk parameters. *J Clin Endocrinol Metab* 2005, 90(5):2545-9.
7. Georgopoulos NA, Papadakis E, Armeni AK, Katsikis I, Roupas ND, Panidis D. Elevated serum androstenedione is associated with a more severe phenotype in women with polycystic ovary syndrome (PCOS). *Hormones* 2014, 13(2):213-21.
8. Azziz R, Carmina E, Dewailly D, Diamanti-Kandarakis E, Escobar-Morreale HF, Futterweit W, Janssen OE, Legro RS, Norman RJ, Taylor AE, Witchel SF. Criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: an androgen excess society guideline. *J Clin Endocrinol Metab* 2006, 91(11):4237-45.

9. Corrie L, Gulati M, Singh SK, Kapoor B, Khursheed R, Awasthi A, Vishwas S, Chellappan DK, Gupta G, Jha NK, Anand K, Dua K. Recent updates on animal models for understanding the etiopathogenesis of polycystic ovarian syndrome. *Life Sci* 2021, 280: 119753.

10. Ho C-H, Chang C-M, Li H-Y, Shen H-Y, Lieu F-K, Wang PS-G. Dysregulated immunological and metabolic functions discovered by a polygenic integrative analysis for PCOS. *Reprod Biomed Online* 2020, 40(1):160-7.

11. Sun Z, Chang H-M, Wang A, Song J, Zhang X, Guo J, Peter CK, Lian F. Identification of potential metabolic biomarkers of polycystic ovary syndrome in follicular fluid by SWATH mass spectrometry. *Reprod Biol Endocrinol* 2019, 17(1):1-10.

12. Schneider, M.V., Orchard, S. Omics technologies, data and bioinformatics principles. In: Mayer, B. (eds) *Bioinformatics for Omics Data: Methods and Protocols*. 1nd ed. Totowa, Humana Press, 2011:3-30.

13. Lindon JC, Holmes E, Nicholson JK. Peer reviewed: so what's the deal with metabonomics? *Anal Chem* 2003, 75(17):384-91.

14. Dona AC, Jiménez B, Schäfer H, Humpfer E, Spraul M, Lewis MR, Pearce TM, Holmes E, Lindon JC, Nicholson JK. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem* 2014, 86(19):9887-94.

15. Jayaraman V, Ghosh S, Sengupta A, Srivastava S, Sonawat H, Narayan PK. Identification of biochemical differences between different forms of male infertility by nuclear magnetic resonance (NMR) spectroscopy. *J Assist Reprod Genet* 2014, 31:1195-204.

16. Arslan AK. Derin öğrenme ve topluluk öğrenme yöntemlerine dayalı bilgisayar destekli tanı sisteminin geliştirilmesi: Omik teknolojileri üzerine uygulaması. Sağlık Bilimleri Enstitüsü, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı. Doktora Tezi, Malatya: İnönü Üniversitesi, 2021.

17. Stewart J, Sprivulis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *EMA* 2018, 30(6):870-4.

18. Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. *Wiley Interdiscip Rev Data Min Knowl Discov* 2021, 11(5):1424.

19. Dewailly D. Diagnostic criteria for PCOS: is there a need for a rethink? *Best Pract Res Clin Obstet Gynaecol* 2016, 37:5-11.

20. Fenstermacher D. Introduction to bioinformatics. *JASIST* 2005, 56(5):440-6.

21. Burley SK. An overview of structural genomics. *Nat Struct Mol Biol* 2000, 7(11):932-4.

22. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS Comput Biol* 2017, 13(5):1005457.

23. Cho WC. Proteomics technologies and challenges. *GPB* 2007, 5(2):77-8.

24. Karu N, Deng L, Slae M, Guo AC, Sajed T, Huynh H, Wine E, Wishart DS. A review on human fecal metabolomics: Methods, applications, and the human fecal metabolome database. *Anal Chim Acta* 2018, 1030:1-24.

25. Krumsiek J, Bartel J, Theis FJ. Computational approaches for systems metabolomics. *Curr Opin Biotechnol* 2016, 39:198-8.

26. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol* 2012, 13(4):263-9.

27. Dunn WB, Ellis DI. Metabolomics: current analytical platforms and methodologies. *TrAC Trends Anal Chem* 2005, 24(4):285-94.

28. Zubarev RA, Makarov A. Orbitrap mass spectrometry. *Anal. Chem* 2013, 85(11): 5288–8

29. Smolinska A, Blanchet L, Buydens LM, Wijmenga SS. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Anal Chim Acta* 2012, 750:82-15.

30. Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, Pujos-Guillot E, Verheij E, Wishart D, Wopereis. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*. 2009, 5(4):435-58.

31. Douglas DJ. Applications of collision dynamics in quadrupole mass spectrometry. *J Am Soc Mass Spectrom* 1998, 9(2): 101-113.

32. Cotter RJ. Time-of-flight mass spectrometry. *J Am Chem Soc* 1993, 65(8): 1043-1047

33. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R. The Orbitrap: a new mass spectrometer. *J. Mass Spectrom* 2005,40(4):430-43.

34. Marshall AG, Hendrickson CL, Jackson GS. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* 1998,17(1):1-35.

35. Snyder LR, Kirkland JJ, Dolan JW. *Introduction to modern liquid chromatography*, John W, Sons, 91st ed. New York, 2011, 960.

36. Mikšik I, Tagliaro F, Tesarová E, Deyl Z. *Advanced chromatographic and electromigration methods in biosciences*. Elsevier, 1st ed. Tokyo; 1998, 60.

37. Howard G, Martin AJ. The separation of the C12-C18 fatty acids by reversed-phase partition chromatography. *Biochem J* 1950,46(5):532.

38. Snyder L, Dolan J, Gant JR. Gradient elution in high-performance liquid chromatography: I. Theoretical basis for reversed-phase systems. *J Chromatogr A* 1979,165(1):3-30.

39. MacNair JE, Lewis KC, Jorgenson JW. Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. *Anal Chem* 1997,69(6):983-9.

40. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989,246(4926):64-7.

41. Yang J, Schmelzer K, Georgi K, Hammock BD. Quantitative profiling method for oxylipin metabolome by liquid chromatography electrospray ionization tandem mass spectrometry. *Anal. Chem* 2009,81(19):8085-93.

42. Kesić S. Systems biology, emergence and antireductionism. *Saudi J Biol Sci* 2016,23(5):584-7.

43. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022,40(10):1458-8.

44. May M. Big data, big picture: Metabolomics meets systems biology. *Science* 2017,356(6338):646-8.

45. Cambria E, Malandri L, Mercorio F, Mezzanzanica M, Nobani N. A survey on XAI and natural language explanations. *Inf Process Manag* 2023, 60(1): 103111.

46. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nat Mach Intell* 2020, 2(10): 573-11.

47. Lipton ZC. The mythos of model interpretability, In machine learning, the concept of interpretability is both important and slippery. *Queue* 2018,16(3):31-26.

48. Barbado A, Corcho Ó, Benjamins R. Rule extraction in unsupervised anomaly detection for model explainability: Application to OneClass SVM. Expert Syst Appl 2022, 189: 116100.

49. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H, editors. Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI conference on human factors in computing systems* 2021,1-52

50. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Mag* 2017,38(3):50-7.

51. Arrieta AB, Díaz-Rodríguez N, Del-Ser J, Bennetot A, Tabik S, Barbado A, Herrera F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion* 2020,58:82-33.

52. Mi JX, Li AD, Zhou LF. Review study of interpretation methods for future interpretable machine learning. *IEEE Access* 2020, 8: 191969-6.

53. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat 2*014, *24*(1): 44-21.

54. Danesh T, Ouaret R, Floquet P, Negny S. Interpretability of neural networks predictions using Accumulated Local Effects as a model-agnostic method. *Comput Aided Chem Eng* 2022, 51: 1501-5.

55. Fish J, Belsky V. Generalized aggregation multilevel solver. *Int J Numer Methods Eng* 1997,40(23):4341-61.

56. Alangari N, El Bachir Menai M, Mathkour H, Almosallam I. Exploring evaluation methods for interpretable machine learning: A survey. *Information*, 2023, *14*(8): 469.

57. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* 2016,1135-9.

58. Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat* 2015,24(1):44-21.

59. Huang F, Shangguan W, Li Q, Li L, Zhang Y. Beyond prediction: An integrated post-hoc approach to interpret complex model in hydrometeorology. *Environ Model Softw* 2023, 167: 105762.

60. Oh S. Feature interaction in terms of prediction performance. *App Sci* 2019, 9(23):5191.

61. Zhou Z, Ong YS, Nair PB, Keane AJ, Lum KY. Combining global and local surrogate models to accelerate evolutionary optimization. *IEEE Trans Sys Man Cybern: Syst* 2006, 37(1):66-10.

62. Shapley, LS. *A value for n-person games*, 2th ed. California, RAND Corporation, 1953: 307-10.

63. Zafar, MR, Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Mach learn knowl* 2021, 3(3): 525-16.

64. Bramhall S, Horn H, Tieu M, Lohia N. Qlime-a quadratic local interpretable model-agnostic explanation approach. *SMU Data Scie Rev* 2020, 3(1):4.

65. Shi S, Zhang X, Fan W. A modified perturbed sampling method for local interpretable model-agnostic explanation. *arXiv preprint* 2020, 200207434.

66. Antwarg, L, Miller, RM, Shapira, B, Rokach, L. Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Syst Appl* 2021, 186: 115736.

67. Chen, H, Lundberg, SM, Lee, SI. Explaining a series of models by propagating Shapley values. *Nat Commun* 2022, 13(1): 4512.

68. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif Intell* 2021, 298:103502.

69. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020, 2(1):56-11.

70. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019, 19(1):1-32.

71. Metabolomics Work Bench. https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000726&StudyType=MS&ResultType=5 Son Erişim Tarihi 01 Ocak 2023.

72. Tang H, Jin Z, Deng J, She Y, Zhong Y, Sun W, Ren Y, Cao N, Chen C. Development and validation of a deep learning model to predict the survival of patients in ICU. *J Am Med Inform* 2022, 29(9):1567-76.

73. Yadav ML, Roychoudhury B. Handling missing values: A study of popular imputation packages in R. *Knowl Based Syst* 2018, 160:104-18.

74. Smiti S, Soui M. Bankruptcy prediction using deep learning approach based on borderline SMOTE. *Inf Syst Front* 2020, 22:1067-83.

75. Dash M, Liu H. Feature selection for classification. *Intell Data Anal* 1997, 1(1-4):131-56.

76. Farrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *J Chemom* 2015, 29(10):528-36.

77. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010, 36:1-13.

78. Jo I, Lee S, Oh S. Improved measures of redundancy and relevance for mRMR feature selection. *Comput* 2019, 8(2):42.

79. Breiman L. Random forests. *Mach Learn* 2001, 45:5-32.

80. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data* 2020, 7(1):1-45.

81. Chen, T, Guestrin, C. *Xgboost: A scalable tree boosting system*. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016, 785-9.

82. Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognit Lett* 2020, 136:190-7.

83. Schapire, RE, Singer, Y. *Improved boosting algorithms using confidence-rated predictions*. In Proceedings of the eleventh annual conference on Computational learning theory 1998, 80-11.

84. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017, 30.

85. Mitrentsis G, Lens H. An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. *Appl. Energy* 2022, 309:118473.

86. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011, 12:2825-30.

87. Yagin FH, Alkhateeb A, Raza A, Samee NA, Mahmoud NF, Colak C, Yagin B. An Explainable Artificial Intelligence Model Proposed for the Prediction of Myalgic Encephalomyelitis/Chronic Fatigue Syndrome and the Identification of Distinctive Metabolites. *Diagnostics* 2023, 13(23):3495.

88. Yagin FH, Cicek İB, Alkhateeb A, Yagin B, Colak C, Azzeh M, Akbulut S. Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. *Comput Biol Med* 2023, 154:106619.

89. Zhao Y, Fu L, Li R, Wang LN, Yang Y, Liu NN, Zhang CM, Wang Y, Liu P, Tu BB, Zhang X, Qiao J. Metabolic profiles characterizing different phenotypes of polycystic ovary syndrome: plasma metabolomics analysis. *BMC Med* 2012, 10(1):1-12.

90. Atiomo W, Daykin C. Metabolomic biomarkers in women with polycystic ovary syndrome: a pilot study. *MHR: Basic Sci Reprod Med* 2012, 18(11):546-53.

91. Sun L, Hu W, Liu Q, Hao Q, Sun B, Zhang Q, Mao S, Qiao J, Yan X. Metabonomics reveals plasma metabolic changes and inflammatory marker in polycystic ovary syndrome patients. *J Proteome Res* 2012, 11(5):2937-46.

92. Azziz R, Carmina E, Dewailly D, Diamanti-Kandarakis E, Escobar-Morreale HF, Futterweit W, Janssen OE, Legro RS, Norman RJ, Taylor AE, Witchel SF. The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report. *Fertil Steril* 2009, 91(2):456-88.

93. Basu A, Basu R, Shah P, Vella A, Rizza RA, Jensen MD. Systemic and regional free fatty acid metabolism in type 2 diabetes. *Am J Physiol Endocrinol Metab* 2001, 280(6): 1000-6.

94. Landin K, Lönnroth P, Krotkiewski M, Holm G, Smith U. Increased insulin resistance and fat cell lipolysis in obese but not lean women with a high waist/hip ratio. *Eur J Clin Invest* 1990, 20:530-5.

95. Ojeda-Ojeda M, Murri M, Insenser M, F Escobar-Morreale H. Mediators of low-grade chronic inflammation in polycystic ovary syndrome (PCOS). *Curr Pharm Des* 2013, 19(32):5775-91.

96. Gaster M, Nehlin JO, Minet AD. Impaired TCA cycle flux in mitochondria in skeletal muscle from type 2 diabetic subjects: marker or maker of the diabetic phenotype? *Arch Physiol Biochem* 2012, 118(3):156-89.

97. Jimenez PT, Frolova AI, Chi MM, Grindler NM, Willcockson AR, Reynolds KA, Zhao Q, Moley KH. DHEA-mediated inhibition of the pentose phosphate pathway alters oocyte lipid metabolism in mice. *Endocrinol* 2013, 154(12):4835-44.

98. Newgard CB, An J, Bain JR, Muehlbauer MJ, Stevens RD, Lien LF, Haqq AM, Shah SH, Arlotto M, Slentz CA, Rochon J, Gallup D, Ilkayeva O, Wenner BR, Yancy WS, Eisenson H, Musante G, Surwit RS, Millington DS, Butler MD, Svetkey LP. A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metab* 2009, 9(4):311-26.

99. Wang TJ, Larson MG, Vasan RS, Cheng S, Rhee EP, McCabe E, Lewis GD, Fox CS, Jacques PF, Fernandez C, O'Donnell CJ, Carr SA, Mootha VK, Florez JC, Souza A, Melander O, Clish CB, Gerszten RE. Metabolite profiles and the risk of developing diabetes. *Nature Med* 2011, 17(4):448-53.

100. Soni P, Vashisht S. Image segmentation for detecting polycystic ovarian disease using deep neural networks. *Int J Comput Sci Eng* 2019, 7(3):534-7.

101. Thufailah I, Wisesty U. An implementation of Elman neural network for polycystic ovary classification based on ultrasound images. *Journal of Physics: Conference Series IOP Publishing* 2018.

102. Vikas B, Anuhya B, Chilla M, Sarangi S. A critical study of polycystic ovarian syndrome (PCOS) classification techniques. *Int J Comput Eng Manag* 2018, 4:1-7.

103. Isah OR, Usman A, Tekanyi A. A review on computer assisted follicle detection techniques and polycystic ovarian syndrome (PCOS) diagnostic systems. IJCTT 2015, 28:41-4.

104. Troisi J, Cinque C, Giugliano L, Symes S, Richards S, Adair D, Cavallo P, Sarno L, Scala G, Caiazza M, Guida M. Metabolomic change due to combined treatment with myo-inositol, D-chiro-inositol, and glucomannan in polycystic ovarian syndrome patients: a pilot study. J Ovarian Res 2019, 12:1-11.

105. Lim AJ, Huang Z, Chua SE, Kramer MS, Yong E-L. Sleep duration, exercise, shift work and polycystic ovarian syndrome-related outcomes in a healthy population: a cross-sectional study. *PloS one* 2016, 11(11): 0167048.

106. Jacobsen VM, Li S, Wang A, Zhu D, Liu M, Thomassen M, Kruse T, Tan Q. Epigenetic association analysis of clinical sub-phenotypes in patients with polycystic ovary syndrome (PCOS). *Gynecol. Endocrinol* 2019, 35(8):691-4.

107. Lim AJ, Indran IR, Kramer MS, Yong EL. Phenotypic spectrum of polycystic ovary syndrome and their relationship to the circadian biomarkers, melatonin and cortisol. *Endocrinol Diabetes Metab J* 2019, 2(3): 00047.

EKLER

EK-1. Özgeçmiş

EK-2. Etik Kurul Almama Gerekçesi