VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Mikrobiyota Verileri İçin Boyut İndirgemede Yeni Bir Yaklaşım

Handan ANKARALI[1]*, Süleyman YILDIRIM[2], Nurgül BULUT[1]

[1]*İstanbul Medeniyet Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, İstanbul*
[2]*İstanbul Medipol Üniversitesi, Tıp Fakültesi, Tıbbi Mikrobiyoloji Anabilim Dalı, İstanbul*

Makale Bilgisi

Başvuru:
13/07/2020
Kabul:
02/01/2021

Özet

İnsan derisi, nazofaringeal ve ağız boşlukları, vajinal sistem ve gastrointestinal sistem ile ilişkili mikroorganizmalar insan mikrobiyotasını oluşturur. Fizyolojik, metabolik ve immun sistem üzerinde oldukça etkilidir ve birçok hastalık ile ilişkisi gösterilmiştir. DNA dizileme teknolojisindeki son gelişmeler, bakteriler için 16S rRNA, 18s rRNA veya ITS gibi marker genlerinin amplikonlarının yüksek verim dizilimi yoluyla, mikrobiyal toplulukların profillenmesi kolaylaşmıştır. Elde edilen veriler, çok büyük sayılarda mikrobiyota türlerine ait frekans değerlerinden oluşur ve bol miktarda sıfır değeri içerir. Mikrobiyota verileri gibi büyük boyutlu verilerin çeşitli istatistik modellerle analiz edilebilmesi için ön işleme aşamasında, sonuca anlamlı katkısı bulunmayan türlerin veri analizinden çıkarılması gerekmektedir. İstatistik literatüründe bu işlem, *boyut indirgeme* veya *değişken eleme* olarak adlandırılmaktadır.

Bu çalışmada, çok sayıda sıfır değeri içeren frekans tipi büyük boyutlu veri setlerinde, boyut indirgeme amacıyla kullanılabilecek yeni bir yaklaşım önerildi. Bu amaçla, tek değişkenli testler, sıfır etkili negatif binomiyal model, sınıflama ve regresyon ağaçları ve değişken seçimi algoritması kullanıldı.

Önerilen yaklaşım, Parkinson hastaları, erken demans ve kontrol bireylerinden elde edilen mikrobiyota cinsleri üzerinde denendi. Değişken seçimi sonucunda 199 bakteri cinsi içinden seçilen 19 adet aday cinsin, klinik açıdan da birçok çalışmada vurgulanan bakteri cinsleri olduğu görüldü. Aday olarak seçilen cinslerin hastalık tanısındaki başarısını değerlendirmek için kurulan multiple logistic regresyon modelinde yeniden stepwise değişken eleme yöntemi kullanıldı ve bu model sonucunda birkaç bakteri cinsi ile başarılı bir şekilde hasta ve kontrol gruplarının ayrımı yapıldı.

Bu çalışma ile önerilen yeni hibrit yaklaşım, birden çok yöntemin ortak kararı neticesinde belirlenen değişkenleri veri analizine alma imkanı sunmaktadır. Benzeri yaklaşımlar farklı yöntemlerle denenerek farklı veri tipleri üzerinde kullanılabilir.

Anahtar kelimeler: *Sıfır etkili modeller; Frekans verisi; Sınıflama ve Regresyon ağaçları; Değişken tarama algoritmaları; Mikrobiyota; Parkinson*

* İletişim e-posta: handanankarali@gmail.com

** Bu çalışmanın bir kısmı III. International Conference on Data Science and Applications 2020'de sözlü olarak sunulmuştur.

A New Approach to Dimension Reduction for Microbiota Data

Abstract

Microorganisms associated with human skin, nasopharyngeal and oral cavities, vaginal tract, and gastrointestinal system make up the human microbiota. It is highly effective on the physiological, metabolic and immune system and has been shown to be associated with many diseases. Recent advances in DNA sequencing technology have facilitated profiling of these microbial communities through high throughput sequencing of amplicons of the marker genes such as 16S rRNA for bacteria, 18S rRNA or ITS. Data generated from such sequencing efforts are preprocessed into composition or relative abundance that are often presented in species abundance (OTU/ASV) tables. The data obtained consists of the frequency of microbiota species in very large numbers and it contains a large amount of zero values. Nonetheless, the high dimensional data in such tables must be treated with dimension reduction techniques to draw sensible conclusions from the data. In the statistical literature, this process is called dimension reduction or variable selection.

The aim in this study is to propose a novel approach to reduce dimensions in high dimensional and inherently zero inflated and frequency character microbiota data. For this purpose, univariate tests, a zero-inflated negative binomial model, classification and regression trees, and a feature selection and variable screening algorithm were used. Using these four methods enabled us to select most important features of the microbiota dataset for the subsequent downstream analyses.

We tested the above approach on our recent microbiota dataset we generated from stool samples of Parkinson's disease patients cohort. Of 199 bacteria genera our approach enabled us to select 19 candidate biomarker genera, which are often implicated in serving critical metabolic activities in human body such as production of short-chain fatty acids. To assess the potential of these candidate biomarkers in differentiating disease and healthy states we developed a multiple logistic regression model and further selected their biomarker potential in a stepwise variable screening.

Big data analysis necessarily entails use of increasingly more sophisticated and combinatorial modalities. Here we successfully demonstrated that hitherto untested combinatorial use of feature selection methods enables more useful predictive models. Similar approaches can be tried with different methods and used on different data types.

***Keywords**: Zero-inflated models; Frequency data; Classification and Regression tree; Variable Screening algorithm; Microbiota; Parkinson's disease*

1 Giriş

Canlı vücudunda yaşayan, sayıları trilyonlarla ifade edilen ve konakçının fizyolojisine doğrudan ya da dolaylı katkısı olan mikroskobik canlı grubu *Mikrobiyota* olarak tanımlanmaktadır [1]. Mikrobiyota verilerinden anlamlı bilgiler üretme sürecinde, dizilerin kalite kontrol ve kümelenmesinden sonra en önemli aşamalardan birisi ön işleme aşamasıdır. Ön işlemede veriler, çeşitli yöntem ve yaklaşımlarla özetlenir, aykırı ve sapan değerler tespit edilir, kayıp veriler çıkarılır veya tahmin edilir ve çalışılan konu ile ilişkisi en yüksek olan taksonomik sınıflar seçilir. OTU, cins veya tür düzeyinde kümeleme neticesinde örneklerdeki göreceli bolluğu hesaplanır. Mikrobiyota çalışmalarında verilerin elde edilme sürecinde yapılan işlemlerin maliyetinin yüksek

olması üzerinde çalışılan denek ve bunlardan alınan örnek sayılarını kısıtlayıcı faktördür. Bu durum istatistik modellerde çok ciddi problemlerin ortaya çıkmasına neden olur ve yüksek seviyeli veri analizlerinin yapılmasını güçleştirir.

Tanımlanan çok sayıda mikrobiyota tür veya cinsi arasından, çeşitli grupları karşılaştırmada aday olanların belirlenmesine Değişken Seçimi adı verilir[2,3]. Aynı model için tam özellik setinden daha iyi sonuçlar veren özellik alt kümeleri mevcut olabilir. Tam veri kullanıldığı durumlarda, veri içinde kaos, aşırı doymuş ve gerçek problemlerle uyumsuz olan sonuçlar veya hesaplama güçlüğü nedenleriyle sorunlar ortaya çıkabilir. Böylece daha az sayıda değişkenle daha kısa sürede başarılı sonuçlara ulaşılabilir [2]. Klasik istatistik modellerde sonuç değişkeni ile ilişkisi anlamlı olmayan ve/veya birbirleri ile yüksek korelasyonlu olan özellikler, adımsal, geriye dönük, ileriye dönük gibi değişken seçim yöntemleri ile ayıklanabilmektedir. Ancak verinin boyutu büyüdükçe uygun değişkenlerin seçiminde bu yöntemler yetersiz kalmış ve farklı yöntem ve yaklaşımlar önerilmeye başlanmıştır. Ayrıca ön işleme aşamasında, değişken seçimi amacıyla sadece bir yöntem kullanılması kararlı ve doğru sonuçlar vermeyebilir.

Değişken seçiminde yaygın kullanılan yöntemlerin hipotez testlerinde kullanılan istatistik yöntemlerle karıştırılmaması gerekir. Çünkü adı geçen testlerin, hem değişken seçiminde hem de hipotez testinde kullanılabileceği vurgulanmaktadır [3]. Literatürde en yaygın kullanılan yöntemlerden birisi *filtre yöntemleridir*. Filtre yöntemleri, bağımlı ve bağımsız değişkenler arasındaki ilişkileri inceleyen çeşitli tek değişkenli istatistik testler,ya α– ve β–diversity ölçüleri ve ısı haritalarından oluşur [2,4,5].

Değişken seçiminde çeşitli makine öğrenmesi algoritmaları kullanan bir diğer yöntem *sarıcı (Wrapper) yöntemdir.* Bu yöntemde, eğitim algoritması kullanılarak bağımsız değişkenlerin farklı alt setleri, "çalışılan değişkenlerin bilgi kazancına" bakılarak belirlenir. Bu süreç en iyi alt seti buluncaya kadar devam eder. Sarıcı yöntemler, ileriye doğru seçim, geriye dönük seçim ve özyinelemeli seçim gibi farklı şekillerde uygulanabilir. En yaygın kullanılan Random Forest algoritması, bağımsız değişkenler arasında var olabilecek çok yüksek ilişkilere, kayıp ve sapan değerlere karşı dayanıklı ve bağımsız değişkenler arası var olabilecek etkileşimleri göstermektedir

Son yıllarda kullanılmaya başlanan *gömülü yöntemler*, filtre ve sarıcı yöntemlerin avantajlarını kombine etmiştir. Değişken seçiminde Gini Importance veya Permutation Importance gibi önem düzeyleri dikkate alınmakta ve seçim, algoritmanın eğitim aşamasında yapılmaktadır. Bu amaçla en yaygın LASSO ve RIDGE regresyon modelleri kullanılmaktadır [4].

Bu çalışmada, mikrobiyota verilerinden uygun taksonomik sınıfların seçimi sürecinde, biri filtre diğeri sarmalayıcı gruba giren iki yöntemin yanı sıra frekans verilerin analizinde kullanılan iki regresyon modelinin birlikte kullanımını göstermek amaçlandı. Bu yaklaşımın temel hedefi, kullanılan dört yöntemin hepsinde kararlılığını koruyan değişkenlerin seçilmesidir. Ayrıca model oluşturma sürecinde, değişken eleme yöntemlerine göre seçilen taksonomik sınıflarla birlikte, klinik ve demografik özellikler yeniden değişken eleme sürecinden geçirilerek tahmin modelinin elde edilmesi planlandı.

2 Gereç ve yöntem

2.1 Örnek veri seti

Uygulama verisi içinde, Kontrol (n=21), erken demans (MCI, n=14) ve Parkinson hastaları (PD, n=14) olmak üzere 3 grup ve toplam 199 mikrobiyota cinsi mevcuttur. Verilerin elde edildiği araştırmanın amacı, Parkinson ve erken demans için ayırıcı tanıda kullanılacak biyobelirteçlerin tespit edilmesi ve modellenmesidir. Bu amaçla çalışmaya dahil olan kişilerin yaş, cinsiyet ve eğitim süresi demografik bilgi olarak kaydedilmiş ve demans ve kognitif bozukluklarda kullanılan 4 ölçek puanı hesaplanmıştır. Ayrıca gaita örneklerinden analiz edilen toplam 199 bakteri cinsinin göreceli bolluğu kaydedilmiş ve dinlenme pozisyonundaki fonksiyonel MR (fMR) verilerine ait 3 farklı bilişsel boyut skorları elde tespit edilmiştir. Yapılan ölçümler aşağıdaki gibi gruplandırılmıştır.

1. **Demografik özellikler ve Klinik değerlendirme skorları:** Yaş, Cinsiyet, Eğitim süresi (yıl), CDR, Stroop, GDS, MMSE klinik skorları
2. **Bilişsel komponent skorları:** Default Mode Network, Right Lateral Executive, Left Lateral Executive

2.1.1 *Değişken seçiminde kullanılan yeni yaklaşım*

Bu çalışmada, değişken seçimi amacıyla; **Negatif Binomiyal Regresyon Modeli** (*Negative Binomial*

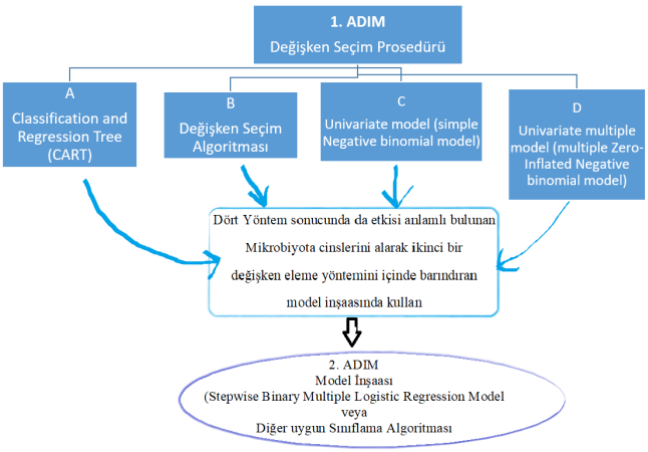
Model, NB), **Sıfır Değer Ağırlıklı Negatif Binomiyal model** (*Zero-inflated Negative Binomial Model, ZINB*), **Sınıflama ve Regresyon ağaçları (*CART*) ve Özellik Seçimi ve Değişken Tarama Algoritması** *(Feature Selection and Variable Screening*) olmak üzere 4 farklı yöntem, hibrit bir yaklaşımla kullanıldı. Bu yöntemler aslında, değişken seçimi amacıyla geliştirilmiş yöntemler değildir. Ancak bilim dünyası büyük veri üzerinde çalışmaya başladığı andan itibaren hipotez testlerinde kullanılan algoritma veya modellerin, ön işleme aşamasında değişken seçimi amacıyla da kullanıldığı sıkça görülmektedir[3]. Seçilen 4 farklı yöntem içinde NB ve ZINB modellerinin değişken seçiminde kullanımına henüz rastlanmamıştır.

NB modelinde bağımlı değişken mikrobiyota cinsleri, bağımsız değişken ise grup (PD, MCI ve kontrol) değişkeni idi. ZINB modelinde; demografik özellikler, bilişsel komponentler ve grup bağımsız değişken, mikrobiyota cinsleri ise bağımlı değişken olarak tanımlandı. Uygun taksonomik sınıfların seçiminde iki modelin de Akaike Bilgi Ölçütü (AIC) ve gruplar arası farka ait *P* değerleri kullanıldı. CART algoritması, gömülü yöntemler grubundan olup veri madenciliğinde sınıflama amacıyla kullanılır. Bu algoritmada grup bağımlı değişken, mikrobiyota cinsleri bağımsız değişken olarak kabul edildi [2]. NB, ZINB ve CART yöntemleri, frekans tipi verilerin analizinde yaygın kullanım alanı bulmuştur. *Özellik Seçimi ve Değişken Tarama Algoritması (Feature Selection and Variable Screening*) 1) değişken seçimi 2) değişken tarama olmak üzere iki aşamadan oluşur. Birinci aşamada, kategorik ve/veya sürekli yapıdaki bağımlı ve bağımsız değişkenler arasındaki ilişkinin şekli dikkate alınmaksızın büyük veri setlerinden aday değişkenler seçilir. Bağımlı ve bağımsız değişkenler kategorik yapıda ise ki-kare ve ***P*** değerleri hesaplanır. Ancak her ikisi veya biri sürekli yapıda ve aykırı gözlemler içeriyorsa ilgili değişken ***k*** tane aralığa bölünür (genel kabul görmüş ***k***=10 dur) ve ki-kare değerleri hesaplanır. Bağımlı ve bağımsız değişkenlerin sürekli yapıda olması durumunda önem düzeyleri korelasyon analizi ile belirlenir. Ayrıca normal dağılım gösteren sürekli yapıda bir bağımlı değişken ve kategorik yapıda bağımsız değişken(ler) varsa önem düzeyleri ***F***-istatistiği ve ***P*** yanılma olasılığı ile ölçülür. Bağımsız değişken sayısı birden fazla ise etkileşimler de incelenebilir. Ancak bu testler sadece doğrusal ilişkileri ortaya koyduğu için gerçek problemlerde karşılaşılan doğrusal olmayan ilişkiler göz ardı edilmektedir. Bu durum değişken seçiminde yanlı kararlara neden olabilir [6].

Kullanılan 4 yöntem içinde sadece CART algoritmasında, incelenen bakteri cinsleri birlikte modele alındı ancak diğer 3 yöntemde her bir cins için ayrı model kuruldu. Bakteri cinslerinin birlikte ele alınması cinsler arası ilişkileri de dikkate almak anlamına gelir [7].

Grupları ayırmadaki etkisi anlamlı olan bakteri cinslerinin, değişken seçim yöntemleri ile belirlenmesi sırasında, yukarıda tanımlanan 4 yönteme göre de ayırt edicilik gücü istatistik olarak anlamlı bulunanlar, *aday (candidate) bakteri cinsi* olarak seçildi. İkinci aşamada, aday bakteri cinsleri, demografik özellikler, klinik skorlar ve bilişsel komponent skorları olmak üzere toplam 29 değişken tanı amaçlı geliştirilecek modelde birlikte dikkate alındı. Bu amaçla, stepwise lojistik regresyon modeli kuruldu. Bu aşamada, yeniden değişken seçim yöntemi (stepwise) kullanılmasının nedeni ise modele alınan değişken sayısının toplam denek sayısına oranla hala büyük olması ve birlikte modele alındıklarında çoklu bağlantı ve/veya anlamsız değişkenlerin ortaya çıkabilme ihtimalinin bulunmasındandır.

Özetle, değişken eleme sürecinde Şekil 1' de verilen adımlar izlendi.



Şekil 1. Değişken eleme süreci

Değişken seçim aşamasında istatistik anlamlılık düzeyi P<0.10 kabul edildi. Hesaplamalarda Statistica (Trial, ver. 18), STATA (ver. 14), SPSS (ver. 22) ve MedCalc (Trial, ver. 16) programları kullanıldı.

3 Bulgular

Çalışmaya 21 kontrol bireyi, 14 erken demans hastası (MCI) ve 14 parkinson hastası (PD) olmak üzere toplam 49 kişi (27 kadın/22 erkek) kişi dahil edildi. Ayrıca katılımcıların yaş ve eğitim süreleri, klinik tanı skorları ve bilişsel komponent skorlarına ait tanımlayıcı istatistikler Tablo 1' de sunuldu.

Değişken seçim veya boyut indirgeme sonuçları

Kullanılan modellerin hepsinde en az iki grubu ayırma başarısı istatistik olarak anlamlı bulunan bakteri cinsleri, ikinci aşamada ayırıcı tanı amacıyla Çalışılan toplam 199 adet bakteri cinsi içinden 19 tanesi, dört yöntemde de en az iki grubu anlamlı düzeyde ayırt ettiği görüldü (Tablo 2).

Tablo 1. Katılımcıların demografik, klinik ve bilişsel skorlarına ait tanımlayıcı istatistikler

	Kontrol			MCI			PD		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Yaş (yıl)	21	58.48	8.42	14	66.21	10.42	14	73.29	9.21
Eğitim süresi (yıl)	21	10.81	5.33	14	4.79	2.42	14	3.79	3.93
CDR	21	0	0	14	0.50	0	14	1.07	0.27
Stroop	21	86.10	27.04	14	128.29	27.20	14	140.14	27.81
GDS	21	4.43	5.29	14	8.07	6.44	14	10.86	7.65
MMSE	21	27.90	1.70	14	24.00	2.25	14	19.86	3.84
Default Mode Network	21	7.10	3.34	14	4.58	2.27	14	-0.01	2.59
Right Lateral Executive	21	23.25	5.78	14	14.31	2.38	14	14.65	3.65
Left Lateral Executive	21	7.62	3.83	14	-1.77	6.18	14	1.43	6.31

Tablo 2. Dört farklı yönteme göre de seçilen bakteri cinsleri

Cins sayısı = 19	
Roseburia	*Eisenbergiella*
Prevotella_9	*Succinivibrio*
Escherichia_Shigella	*Bifidobacterium*
Coprococcus_1	*Eubacterium_hallii_group*
Blautia	*Lachnospiraceae_NC2004_group*
Faecalibacterium	*Anaerostipes*
Eubacterium_rectale_group	*uncultured_Christensenellaceae*
Coprococcus_3	*Lachnospiraceae_ND3007_group*
Coprobacillus	*Unclassified_Lachnospiraceae*
Lactobacillus	

Ayrıca, tüm bakteri cinsleri içinde sadece **NB modeline** göre seçilen ancak diğer 3 yöntemin seçmediği cins sayısı **1**, sadece **özellik seçim ve değişken tarama algoritmasına** göre seçilen cins sayısı **43** ve sadece **CART algoritmasına** göre seçilen cins sayısı **4** olarak belirlendi. Bu durumda toplam **48 bakteri cinsi**, sadece **1** yöntemde anlamlı sonuç verdi. Bu bakteri cinsleri, ikinci aşamada ayırıcı tanı amacıyla geliştirilecek modele dahil edilmedi.

Bunun yanı sıra,

- Sadece **NB model** ve **özellik seçim ve değişken tarama algoritması**nda anlamlı bulunan cins sayısı **13**,
- **CART algoritması** ve **özellik seçim ve değişken tarama algoritmasında** anlamlı bulunan cins sayısı **18**,
- **Özellik seçim ve değişken tarama algoritması** ile birlikte **ZINB modelinde** anlamlı bulunan cins sayısı **16**,
- **ZINB modeli** dışında diğer 3 yöntemde de anlamlı bulunan cins sayısı **4**,
- **NB modeli** dışında diğer 3 yöntemde de anlamlı bulunan cins sayısı **15** ve
- **CART algoritması** dışında 3 yöntemde de anlamlı bulunan cins sayısı **8** olarak belirlendi.

Sonuç olarak, iki veya üç yöntemde en az iki grubu anlamlı düzeyde ayırt ettiği tespit edilen toplam bakteri cinsinin **74** olduğu söylenir.

- Dört yönteme göre de hiçbir grubu başarılı bir şekilde ayırt edemeyen bakteri cinslerinin sayısı **58** olarak belirlendi.

Ayırıcı tanıda kullanılacak model performansı

Değişken seçim aşamasından sonra ayırıcı tanıda kullanılacak olan model inşası aşamasına geçildi. Bu amaçla Kontrol, MCI ve PD gruplarını ikili karşılaştıran çoklu lojistik regresyon modeli kullanıldı. Kullanılan 4 değişken eleme yöntemine

göre en az iki grubu ayırt ediciliği anlamlı bulunan 19 bakteri cinsinin yanı sıra demografik özellikler, klinik skorlar ve bilişsel komponent skoru olmak üzere toplam 29 değişken aynı anda lojistik regresyon modeline alındı. Ancak modele alınan değişken sayısının (k=29) gruplardaki toplam kişi sayısına oranla hala fazla olması, model katsayılarının hatalı tahmin edilmesine yol açacağı bilinmektedir. Kurulan model sonuçları da bunu göstermiştir. Ayrıca kurulacak lojistik regresyon modelinde söz konusu değişkenlerden bazılarının birlikte etkileri anlamsız hale dönüşebilir. Bu nedenle model inşası aşamasında yeniden değişken eleme algoritmalarından birisi olan stepwise yöntemi kullanılmış ve aşağıda özellikleri tanımlanan final modeller elde edilmiştir.

Model-1: *MCI ile Control gruplarını ayırt edici biyobelirteçler*

MCI ve Kontrol gruplarını anlamlı düzeyde ayırt edici özellikler içeren final model, "Model 1" olarak adlandırıldı.

Grup (MCI vs Control)=20.69 - 1.09 (MMSE) + 0.051 (Stroop) + 8.70 (Eubacterium hallii group) ***(Model-1)***

Model-1 incelendiğinde, MMSE skoru arttıkça MCI riskinin azaldığı, Stroop puanı arttıkça MCI riskinin de arttığı ve Eubacterium hallii group abundance arttıkça MCC riskinin anlamlı düzeyde yüksekdiği görüldü.

Bu modelin sınıflama başarısı değerlendirildiğinde, Control grubunda yer alan toplam 21 kişinin 1' i ve MCI grubunda yer alan 14 kişinin 2' si hatalı sınıflandırılmış olup Model-1' in Control bireylerini ayırma başarısı (Specifity) %95.2 ve MCI grubundaki bireyleri ayırma başarısının ise (Sensitivite) %85.7 olduğu görüldü. Genel doğru sınıflama başarısı (Accuracy) ise %91.4' tür.

Model-2: *PDD ile Control gruplarını ayırt edici biyobelirteçler*

PDD ve Control gruplarını ayırt etmedeki etkisi anlamlı bulunan değişkenler Model-2' de sunuldu

Grup (PDD vs Control) = -6.31+0.085 (Stroop) - 4.270 (Roseburia) ***(Model-2)***

Model-2 incelendiğinde, Stroop puanı açısından Control ve PDD grupları arasında anlamlı fark olduğu ve PDD grubunda Stroop puanının anlamlı düzeyde daha yüksek olduğu gözlendi. Ayrıca Roseburia abundance arttıkça PDD riskinin anlamlı düzeyde düştüğü görüldü.

Bu modelin sınıflama başarısı değerlendirildiğinde, Control grubunda yer alan toplam 20 kişinin 1' i ve PDD grubunda yer alan 14 kişinin 1' si hatalı sınıflandırılmış olup Model-2' nin Control bireylerini ayırma başarısı (Specifity) %95.2 ve PDD grubundaki bireyleri ayırma başarısının ise (Sensitivite) %92.9 olduğu görüldü. A Genel doğru sınıflama başarısı (Accuracy) ise %94.3' tür.

Model-3: *PDD ve MCI gruplarını ayırt edici biyobelirteçler*

PDD ve MCI gruplarını anlamlı düzeyde ayırt edici özellikler içeren final model, "Model 3" olarak adlandırıldı.

Grup (PDD vs MCI) = 24.45-1.02 (MMSE)+0.21 (Left Lateral Executive)–5.81 (Eubacterium hallii group) ***(Model-3)***

Model-3 incelendiğinde, MMSE skorunun PDD grubunda MCI grubuna göre anlamlı düzeyde daha düşük olduğu, Left Lateral Executive skorunun PDD grubunda daha yüksek çıktığı ve Eubacterium hallii group abundance arttıkça PDD riskinin anlamlı düzeyde düştüğü görüldü.

Bu modelin sınıflama başarısı değerlendirildiğinde, MCI grubunda yer alan toplam 14 kişinin 2' si ve PDD grubunda yer alan 14 kişinin 3' ü hatalı sınıflandırılmış olup Model-3' ün MCI bireylerini ayırma başarısı (Specifity) %85.7 ve PDD grubundaki bireyleri ayırma başarısının ise (Sensitivite) %78.6 olduğu bulundu. Genel doğru sınıflama başarısı (Accuracy) ise %82.1' dir.

Neticede bu veri seti kullanılarak tanı başarısı oldukça yüksek olan basit modeller geliştirildi.

Bu süreçler, en az sayıda biyobelirteç (*bakteri cinsi, demografik bilgi, klinik skor, bilişsel komponent skoru*) ile en başarılı tanı elde etmeyi hedeflemektedir.

4 Tartışma

Verimli bir modelinin tasarımında, sonuç ile bağlantılı ve bilgilendirici özelliklerin seçilmesi en önemli aşamalardan biridir. Yüksek boyutlu veriler yalnızca daha uzun ve karmaşık hesaplama süreleri gerektirmekle kalmaz, aynı zamanda analizin doğruluğunu da etkileyebilir [3,8]. Bu nedenle, orijinal verilerin özelliklerini olabildiğince koruyan bir alt küme veya gizli faktörleri belirlemek önemli bir konudur. Literatürde var olan ve genellikle üç ana başlık altında toplanan yöntemlerin çeşitli olumlu ve

olumsuz özellikleri vardır [9]. Filtreleme yöntemleri, her bir özelliğin kullanışlılığını tahmin etmek için tasarlanmış bir puan hesaplayarak değişken seçimi yapar. Bu yöntemler, herhangi bir öğrenme modelinden bağımsız olduğu için model ile ilişkili önyargıya sahip değildir. Ayrıca uygulama ve anlama açısından basitlik içerirler ve değişkenlerin anlamlılığı istatistik testlerle belirlenir. Birçok gerçek dünya probleminde, filtre metotları en iyi alt seti bulmakta başarısız olurken, sarmalayıcı ve gömülü yöntemler daha başarılı ve tutarlı değişken seçimi yapmakta ve ilgili özellikleri seçmek için önceden belirlenmiş bir öğrenme algoritması kullanmaktadır. Ancak hesaplama karmaşıklığı ve işlem yoğunluğu fazladır ve seçilen alt setlerde modele aşırı uyum sorunu yaşanmaktadır [4,10]. Gömülü yöntemler kapsamında yer alan LASSO modeli, bağımsız değişkenler arasında yüksek korelasyonlar bulunması durumunda stabil olmayan sonuçlar verebilir bu nedenle L1 ve L2 olarak adlandırılan iki ceza birlikte dikkate alınmalıdır. Bu yöntem Elastic Net olarak adlandırılır. LASSO sonuçlarının stabil olmasını sağlamak için literatürde önerilen, BoLASSO, adaptive Lasso, relaxed Lasso, thresholding, stability selection gibi farklı yaklaşımlar da mevcuttur Ayrıca yüksek korelasyon durumunda RIDGE regresyon yöntemi de kullanılabilir.

Mikrobiyota üzerinde yürütülen araştırmaların veri analizinde, karmaşık model veya algoritmaların başarılı bir şekilde kullanıldığı, buna karşın değişken seçimi konusunda bir mutabakata varılmadığı görülmektedir. Ancak çeşitli yöntemlerle değişken seçiminde önemli bir yol kat edildiği de göz ardı edilmemelidir [11]. Mikrobiyota verilerinin en önemli farklılıkları, frekans tipinde olmaları ve bol miktarda sıfır değeri içermeleridir. Ayrıca boyut sayısının (mesela tanımlanan bakteri türleri) çalışılan denek sayısına oranla çok fazla olması göze batan bir diğer farklılıktır. Bu farklılıklar nedeniyle birçok araştırmadaki veri tipine uyan ve genellenebilen sonuçlar üreten değişken seçim yöntemleri mikrobiyota verileri için başarılı olmayabilir. Literatürde değişken seçimi amacıyla hibrit veya hiyerarşik yaklaşımlara eğilim artmaktadır. Bu çalışmada da 4 farklı yöntem sonuçları bir araya getirilerek değişken seçiminde hibrit bir yaklaşım kullanıldı.

Bazı Mikrobiyota araştırmalarında, veri analizine başlamadan önce %5' ten daha az oranda görülen türler veri setinden çıkarılmış ve filtreleme yöntemlerinden birisi olan Kruskal-Wallis testi ile geriye kalan türler ile gruplar arası ilişkiler incelenmiştir. Bu süreçte anlamlı etkisi bulunan türler, bir makine öğrenmesi algoritmasının öğrenme sürecinde tekrar incelenmiş ve anlamlı bulunan türler, algoritmanın veya modelin test aşamasında kullanılmıştır [2].

Bu çalışmada önerilen hibrit yaklaşımda, 4 farklı yöntemde kararlı bir şekilde önemli bulunan bakteri cinsleri dikkate aldığı için seçimde daha güvenilir sonuçlar verdiği düşünülmektedir. Ayrıca bu çalışmada kullanılan dört yönteme göre de en az iki grubu ayırmada başarılı sonuç veren 19 bakteri cinsi (Tablo 2), birçok araştırmada PD ile sağlıklı bireyleri ayırmada anlamlı etki gösterdiği belirlenmiştir [12]. Buna ilaveten seçilen bakteri cins sayısı, denek sayısı ile orantılandırıldığında makul bir düzeye indiği için tahmin modeli geliştirilirken klasik istatistik modellerle birlikte kullanılan değişken sayısı indirgeme metotları daha güvenilir bir şekilde uygulanabilmektedir.

Sonuç olarak; çalışmada önerilen hibrit yöntem sonuçları ile literatürde var olan yöntem sonuçlarının çeşitli koşullardaki performansı incelenerek başarılarının tartışılması gerekmektedir. Bu amaçla çeşitli fenotiplerle ilişkileri tanımlanmış bakteri türleri üzerinde karşılaştırmalar yapılmasına ihtiyaç bulunmaktadır.

Kaynaklar

[1] Altuntaş Y, Batman A. "Mikrobiyota ve metabolik sendrom". *Turk Kardiyol Dern Ars* , 45(3), 286–296, 2017.

[2] Chen WP, Chang SH, Tang CY, Liou ML, Tsai SJ, Lin YL. "Composition analysis and feature selection of the oral microbiota associated with periodontal disease". *Biomed Res Int*, 2018, 1-14, 2018.

[3] Saeys Y, Inza I, Larra˜naga P. "A review of feature selection techniques in bioinformatics". *Bioinformatics*, 23(19),2507–2517, 2007.

[4] Knights D, Costello EK, Knight R. "Supervised classification of human microbiota". *FEMS Microbiol Rev*, 35(2), 343–359, 2011.

[5] Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. "Metagenomic biomarker discovery and explanation". *Genome Biol.*, 12(6), 1-18, 2011.

[6] Ditzler G, Morrison JC, Lan Y, Rosen GL. "Fizzy: feature subset selection for metagenomics". *BMC Bioinformatics*, 16(358), 1-8, 2015.

[7] Torbati ME, Mitreva M, Gopalakrishnan V. "Application of taxonomic modeling to microbiota data mining for detection of Helminth infection in global populations". *Data (Basel)*, 1(3), 1-23, 2016.

[8] Zhang B, Cao P. "Classification of high dimensional biomedical data based on feature selection using redundant removal". *PLoS ONE*, 14(4), 1-19, 2019.

[9] Mahadeo U, Dhanalakshmi KR. "Stability of feature selection algorithm: A review". *Journal of King Saud University –Computer and Information Sciences*, Article in Press, 1-14, 2019. https://doi.org/10.1016/j.jksuci.2019.06.012

[10] Somol P, Baesens B, Pudil P Vanthienen J, Leuven KU. "Filter- versus Wrapper-based feature selection for credit scoring". *Int J Intell Syst*, 20(10), 985-99, 2005.

[11] Oudah M, Henschel A. "Taxonomy-aware feature engineering for microbiome classification". *Bioinformatics*, 19(227), 1-13, 2018.

[12] Haikal C, Chen QQ, Li JY. "Microbiome changes: an indicator of Parkinson's disease?". *Transl Neurodegener*, 8(38),1-9, 2019.