Kokain Kullanımının Sınıflayıcı Makine Öğrenmesi Algoritmaları ile Tahmini: UCI Drug Consumption Veri Kümesi Üzerine Bir Analiz

Abdülkadir Ipek
Mühendislik Fakültesi / Ticari Bilimler
Fakültesi Başkent Üniversitesi
Ankara/Türkiye
22396338@mail.bask
ent.edu.tr

***Abstract*—**Bu çalışmada, bireylerin kokain kullanım durumunun tahmin edilmesi amacıyla UCI Machine Learning Repository'de yer alan Drug Consumption adli veri kümesi kullanılmıştır. Veri kümesi; demografik bilgiler, kişilik envanteri skorları ve çeşitli yasal ve yasa dışı maddelerin kullanım düzeylerini içeren çok boyutlu davranışsal özelliklerden oluşmaktadır. Çalışma kapsamında kokain kullanım düzeyi bir sınıflama problemi olarak ele alınmış ve Lojistik Regresyon, Destek Vektör Makinesi ve Rassal Orman olmak üzere üç farklı denetimli makine öğrenmesi algoritması kullanılarak modelleme gerçekleştirilmiştir. Modellerin Performansı; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru metrikleri üzerinden değerlendirilmiştir. Ayrıca, özellik seçimi sürecinin model performansı üzerindeki etkisini incelemek amacıyla Bilgi Kazancı, ReliefF ve Ki-Kare Testi olmak üzere üç farklı özellik seçimi algoritması uygulanmış; modeller hem tüm özelliklerin bulunduğu orijinal veri kümesi ile hem de azaltılmış özelliklerden oluşan veri kümesi ile yeniden eğitilmiştir. Elde edilen bulgular; kokain kullanımının, veri kümesinde bulunan diğer uyuşturucu maddeler ve şahısların demografik özellikleri ile güçlü bir ilişki içerisinde olduğunu göstermiştir. Ayrıca; makine öğrenmesi algoritmaları arasında Rassal Orman algoritmasının, hem özellik seçimi öncesinde hem de sonrasında en yüksek ve istikrarlı performansı sergilediği görülmüştür. Özellik seçimi algoritmalarının uygulanması sonrasında, modellerin daha az sayıda değişkenle benzer performans seviyelerine ulaşabildiği görülmüş, bu durumun özellik özellik seçiminin maliyet-performans dengesi açısından etkin bir yaklaşım sunduğunu ortaya koymuştur.

Keywords—Veri Madenciliği, Makine Öğrenmesi, Özellik Seçimi, Sınıflama Algoritmaları, Kokain Kullanımı

I. GİRİŞ

A. Veri Kümesi Hakkında

Uyuşturucu kullanımının bireysel ve toplumsal etkileri, modern veri analitiği ve davranışsal bilimlerin ortaklaşa incelediği bir alan haline gelmiştir. Özellikle kokain gibi aşırı bağımlılık üreten maddelerin kullanımını etkileyen çeşitli demografik, psikolojik ve davranışsal etkenlerin anlaşılması hem fen bilimlerinde hem de uygulamalı bilimlerde önemli bir araştırma konusu haline gelmiş bulunmaktadır. Bu çalışma kapsamında kullanılan UCI Drug Consumption veri kümesi, içerisinde bulundurduğu örneklerin (şahısların) yaş, cinsiyet, etnik köken gibi sosyo-demografik özellikleriyle beraber kişilik envanteri skorları ve çeşitli diğer uyuşturucu maddelerle kokainin eş zamanlı kullanım düzeylerini içeren kapsamlı bir davranışsal yapı sunmaktadır.
Veri kümesinde bulunan tüm özellikler aşağıda listelenmiştir:
ID: Örneğin numarası,
Age: Örneğin yaşı,
Gender: Örneğin cinsiyeti,
Education: Örneğin eğitim düzeyi,
Country: Örneğin yaşadığı ülke,
Ethnicity: Örneğin etnik kökeni,
Nscore: Örneğe ait Nevrotiklik skoru,
Escore: Örneğe ait Dışadönüklük skoru,
Oscore: Örneğe ait Deneyime Açıklık skoru,
Ascore: Örneğe ait Uyumluluk skoru,
Cscore: Örneğe ait Sorumluluk skoru,
Impulsiveness: Örneğe ait Dürtüsellik skoru,
SS: Örneğe ait Duyum Arayışı skoru,
Alcohol: Örneğin alkol kullanım düzeyi,
Amphetamines: Örneğin amfetamin kullanım düzeyi,
Amyl_nitrite: Örneğin amil nitrat kullanım düzeyi,
Benzodiazepines: Örneğin benzodiazepin kullanım düzeyi,
Cannabis: Örneğin esrar kullanım düzeyi,
Chocolate: Örneğin çikolata kullanım düzeyi,
Caffeine: Örneğin kafein kullanım düzeyi,
Crack: Örneğin crack kullanım düzeyi,
Ecstasy: Örneğin ekstazi kullanım düzeyi,
Heroin: Örneğin eroin kullanım düzeyi,
Ketamine: Örneğin ketamin kullanım düzeyi,
Legal_highs: Örneğin yasal uyarıcı madde kullanım düzeyi,
LSD: Örneğin LSD kullanım düzeyi,
Methadone: Örneğin metadon kullanım düzeyi,
Mushrooms: Örneğin psikoaktif etkiler yaratan mantar kullanım düzeyi,
Nicotine: Örneğin nikotin kullanım düzeyi,
VSA: Örneğin VSA kullanım düzeyi,
Semeron: Örneğin semeron kullanım düzeyi.
Veri kümesinde bulunan tüm uyuşturucu maddeler için kullanım düzeyleri 0-6 arasında kategorik bir ölçekle kodlanmıştır. Bu ölçek, bireyin ilgili maddeyi hiç kullanmamış olmasından son 24 saat içerisinde kullanmış olmasına kadar uzanan 7 seviyeli bir sıklık derecesini temsil etmektedir. Kısaca tanımlamak gerekirse:
CL0 (veya 0): Hiç Kullanmamış,
CL1 (veya 1): 10+ Yıl Önce Kullanmış,
CL2 (veya 2): Son 10 Yıl İçerisinde Kullanmış,
CL3 (veya 3): Son 1 Yıl İçerisinde Kullanmış,
CL4 (veya 4): Son 1 Ay İçerisinde Kullanmış,
CL5 (veya 5): Son 1 Hafta İçerisinde Kullanmış,
CL6 (veya 6): Son 24 Saat İçerisinde Kullanmış.
Bu kodlama sayesinde söz konusu maddelerin hem geçmiş kullanım sıklıkları hem de güncel kullanım davranışları nicel bir biçimde analiz edilebilmektedir.

Benzer şekilde, veri kümesinde bulunan özelliklerden olan Beş Faktör Kişilik skorları (yukarıda belirtilen Nscore, Escore, Oscore, Ascore, Cscore), Duyum Arayışı (Sensation

Seeking veya kümedeki adıyla SS) ve Dürtüsellik (Impulsiveness) skorları da uyuşturucu kullanım düzeylerinde kullanılan numerikleştirme yöntemine benzer olarak, bir puan skalası yardımı ile nicel hale getirilmiştir. Şahısların:
Beş Faktör Kişilik özelliklerinin skorları -3,5 ile 3,5 arası değişen ondalıklı değerler ile,
Duyum Arayışı skorları -2 ile 2 arasında değişen ondalıklı değerler ile,
Dürtüsellik skorları ise -2,6 ile 2,91 arasında değişen ondalıklı değerler ile temsil edilmiştir.
Veri setinin çok boyutlu yapısı ve şahıs davranışlarını yukarıda özetlenen metrikler ile temsil etmesi, onu makine öğrenmesi modelleri için son derece elverişli bir kaynak haline getirmektedir.

B. Temel Amaçlar

Bu analizin temel amacı, kokain kullanımını etkileyen ve/veya tetikleyen etkenlerin veri madenciliği prensipleri ile incelenmesi, bu prensipler ışığında sınıflama algoritmalarının performanslarının karşılaştırılması ve özellik seçimi yöntemlerinin modellerin doğruluğu üzerindeki etkilerinin değerlendirilmesi olarak özetlenebilir. Bu doğrultuda Lojistik Regresyon (Logistic Regression veya LR), Destek Vektör Makinesi (Support Vector Machine veya SVM) ve Rassal Orman (Random Forest veya RF) gibi yaygın ve etkili sınıflama algoritmaları kullanılmış; Bilgi Kazancı (Information Gain), ReliefF ve Ki-Kare olmak üzere üç farklı özellik seçimi algoritması kullanılarak modeller hem saf veri kümesi ile hem de azaltılmış özelliklere sahip veri kümesi ile eğitilmiştir.

C. Kokain Kullanımının Bir Sınıf Değişkeni Olarak Yapılandırılması

Bu çalışmanın temel konusu olan kişilerin kokain kullanımı, veri kümesinde "Cocaine" değişkeni ile 0-6 arasında değerler alabilen bir kategorik değişken olarak temsil edilmektedir. Bu değişken bireylerin kokain kullanım sıklığını gösterirken, diğer maddelerin kullanım düzeyleri, kişilik skorları ve demografik bilgilerin bulunduğu özellikler bağımsız değişkenler olarak modellemeye dahil edilmiştir. İkili bir sınıf mantığı oluşturmak ve analiz esnasında karmaşıklığı gidermek adına, kokain kullanım düzeyi "CL0-CL1 = Kullanmıyor, CL2-CL6 = Kullanıyor" şeklinde bir dönüşümle yeniden tasarlanmıştır. Bu sayede kokain, çalışmaya söz konusu olan tahmin edilmesi beklenen uyuşturucu madde kullanımı olarak seçilmiştir.

II. LİTERATÜR TARAMALARI

A. Çoklu Madde Kullanımı Davranışları

Uyuşturucu madde kullanımının kimi durumda tekil bir davranış olmaktan ziyade birbirini tetikleyen ve aynı bireyde birlikte ortaya çıkan bir örüntüye sahip olduğu birçok çalışmada vurgulanmıştır. Çoklu madde kullanımı, madde kullanan veya bağımlılık sorunu olanlarda sıkça karşılaşılan bir durumdur. Aidiyet ve sevgi arayışı, eğlenmek, sosyalleşmek, bireyleşmek/bağımsızlaşmak, rahatlamak, kaygıyı gidermek, kontrol elde etmek/güçlü olduğunu hissetmek gibi nedenlerle [1] başlayan madde kullanımı, erken dönemlerde terk edilmezse zamanla hem kullanım miktarı artmakta, hem de kullanılan maddeler çeşitlenmektedir [2]. Bu davranış örüntüsüne sahip olan kimseler yalnızca tek bir madde kullanmakla kalmamakta; alkol, nikotin, kafein gibi diğer uyarıcı/psikoaktif maddeleri de eş zamanlı ya da birbirini takip eden dönemlerde kullanabilmektedir. Örneğin, yukarıda bahsedilen yasal ve yasa dışı uyuşturuculardan birisi olan nikotin ele alınacak olursa; tütün ya da nikotin bağımlılığı, bağımlılık yapan ve yapmakta olan diğer maddelere ulaşmanın ilk basamağı olarak görülebilir [3]. Benzer şekilde alkol, kafein gibi bağımlılık yapıcı maddelerin de kokain gibi uyuşturucu maddelerin kullanımına ön ayak olduğu ve bunu takiben bu maddelerin eş zamanlı kullanımına sebep olacağı da öngörülebilir.

B. Kişilik Özellikleri ve Demografik Etkenlerin Kokain Kullanımındaki Yeri

Kişilik özelliklerinin madde kullanımı üzerindeki etkisi, davranışsal bilimler alanının sıklıkla ilgilendiği bir konudur. Özellikle Beş Faktör Kişilik Modeli gibi şahısların kişilik özelliklerini açıklamada bilimsel metodolojiyi sıklıkla kullanan modeller, kişilerin risk alma eğilimleri ve bağımlılık davranışlarını açıklamada oldukça sık kullanılmaktadır. Bu modelde nevrotiklik düzeyi yüksek bireylerin stresle baş etme mekanizması olarak uyuşturucu maddelere yönelme ihtimalinin daha yüksek olduğu; dışadönüklük ve deneyime açıklık gibi özelliklerin ise uyarıcı madde arayışı ve yeni deneyimlere açıklık ile ilişkilendirildiği çeşitli çalışmalarda rapor edilmiştir. (…) Dürtüsel, uyum sorunu olan, nevrotik özellikler taşıyan ve dışadönük kişilik özellikleri madde kullanım bozuklukları gelişiminde önemli birer bir risk faktörü olarak görülmektedir [4].
Demografik değişkenler de madde kullanım davranışlarının anlaşılmasında önemli bir rol oynamaktadır. Yaş, bireylerin risk alma davranışlarının zaman içerisindeki değişimini gösterirken, cinsiyet, eğitim düzeyi, bireyin sosyal çevresi gibi etkenler de maddeye erişim, sosyal kabul ve kullanım sıklığı üzerinde belirleyici olabilmektedir. Örneğin Zeki Karataş'ın 2019 yılında gerçekleştirdiği bir araştırmada, katılımcıların %47,2'sinin uyuşturucu maddeyi ilk kez 18 yaşından önce, %34,7'sinin 19-23 yaş aralığında, %18,1'inin ise 24 yaş ve daha büyükken kullandığı görülmüştür. İlk seferinde esrar kullananların oranı %92,5'tir. Katılımcıların % 58,7'si merak, %57,8'i arkadaşların etkisi, %34,3'ü özenti, %28,1'i sorunlu sosyal çevre, %19,6'sı stresli yaşam olayları, %18'i alkol kullanımı, %17,7'si düzensiz aile yaşamı, %13,1'i işsizlik ve ekonomik sorunlar, %13,1'i eğlence merkezine gitme, %12,2'si öfkesini kontrol edememe, %11'i tanı konulmamış psikolojik sorunları olma, %7,6'sı iletişim problemleri olma, %3,7'si okulla ilgili sorunları olma, %3,7'si küçük yaşta şiddet görme nedeniyle madde kullanmaya başladığını belirtmiştir [5].

İncelenen bu çalışmalar genel olarak madde kullanımının çok boyutlu bir yapıya sahip olduğunu, kişilik özellikleri ve demografik etkenlerin madde kullanım davranışları üzerinde belirleyici bir rol oynadığını ortaya koymaktadır. Yukarıda belirtilen çalışmalara ayrıca katkı sağlama amacı güden bu analiz; davranışsal, demografik ve çoklu madde kullanımına ilişkin değişkenleri bir arada değerlendirerek farklı sınıflama algoritmalarının ve özellik seçimi yöntemlerinin kokain kullanımını tahmin etmedeki performanslarını karşılaştırmalı biçimde incelemeyi amaçlamakta ve bu yönüyle literatüre metodolojik bir katkıda bulunmayı hedeflemektedir.

III. BULGULAR

A. Makine Öğrenmesi Algoritmaları, Özellik Seçimi Algoritmaları, Kullanılan Metrikler ve Oluşturulan Modeller

Analiz kapsamında kokain kullanımının tahmin edilmesi amacıyla üç farklı sınıflama algoritması kullanılmıştır. Bunlar; Lojistik Regresyon (Logistic Regression veya LR), Destek Vektör Makinesi (Support Vector Machine veya SVM) ve Rassal Orman (Random Forest veya RF) algoritmalarıdır. LR, ikili sınıflama problemlerinde yaygın olarak kullanılan doğrusal bir model olup, bağımsız değişkenler ile bağımlı değişken arasındaki olasılık ilişkisini modellemektedir. SVM, yüksek boyutlu veri setlerinde sınıflar arasındaki en uygun ayırıcı hiper-düzlemi bulmayı amaçlayan güçlü bir sınıflama algoritmasıdır. RF ise birden fazla karar ağacının birleşiminden oluşan, doğrusal olmayan ilişkileri yakalamada başarılı bir topluluk öğrenmesi algoritmasıdır. Bu algoritmaların sınıflama metotları olarak seçilmesinin sebebi hem doğrusal hem de doğrusal olmayan sınıflama yaklaşımlarının aynı veri kümesi üzerinde karşılaştırmalı olarak değerlendirilmesine imkan sağlamasıdır.

Oluşturulan modellerin performansı yalnızca kullanılan algoritmaya değil, aynı zamanda modele dahil edilen özelliklerin niteliğiyle ve sayısıyla da doğrudan ilişkilidir. Özellikle yüksek boyutlu veri kümelerinde, nesnelerin sınıfını ayırmada düşük katkıda bulunan özelliklerin modele dahil edilmesi, hem hesaplama maliyetini (zaman ve enerji bakımından) arttırmakta, hem de modelin genelleme yeteneğini olumsuz etkileyebilmektedir. Bu nedenle, çalışmada sınıflama aşamasından önce ve sonra karşılaştırmalı analiz yapılabilmesi amacıyla çeşitli özellik seçimi algoritmaları kullanılmış; bu algoritmaların ortak olarak belirlediği en zayıf özellikler veri kümesinden çıkarılarak modeller tekrar çalıştırılmıştır.

Bu çalışma kapsamında Bilgi Kazancı (Information Gain), ReliefF ve Ki-Kare Testi'nden oluşmak üzere üç farklı özellik seçimi algoritması kullanılmıştır. Bilgi Kazancı, bir özelliğin sınıf değişkeni hakkında sağladığı bilgi miktarını ölçmektedir. Bu yöntem, bir özelliğin sınıf belirsizliğini ne ölçüde azalttığını nicel bir biçimde ifade ederken özellikle sınıf ile güçlü bağımlılık gösteren değişkenlerin belirlenmesinde etkilidir. ReliefF, örnekler arasındaki benzerlik ve farklılık ölçütlerini dikkate alarak bir özelliğin aynı sınıfa ait örnekleri birbirine yaklaştırma ve farklı sınıfa ait örnekleri ayırt etme yeteneğini ölçmektedir. Bu yönüyle ReliefF, doğrusal olmayan ilişkileri de yakalayabilen, örnek tabanlı bir özellik seçimi algoritmasıdır. Ki-Kare Testi ise kategorik veya ayrık değişkenler için sıklıkla kullanılan istatistiksel bir algoritma olup, bir özelliğin sınıf değişkeninden bağımsız olup olmadığını sınamaktadır. Düşük Ki-Kare Testi skoruna sahip özellikler, sınıf değişkeni ile anlamsız bir ilişkiye sahip olarak değerlendirilir.

Son olarak, oluşturulan sınıflama modellerinin başarımı; doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve F1 skoru metrikleri kullanılarak değerlendirilmiştir. Doğruluk metriği, modelin tüm örnekler üzerindeki genel doğru tahmin oranını ifade ederken; kesinlik, pozitif olarak sınıflandırılan örneklerin ne kadarının gerçekten pozitif olduğunu göstermektedir. Duyarlılık metriği, gerçek pozitif örneklerin ne ölçüde doğru biçimde tespit edilebildiğini ölçerken; F1 skoru ise kesinlik ve duyarlılık metriklerinin harmonik ortalamasını alarak bu iki ölçüt arasındaki dengeyi yansıtır. Ayrıca, sınıflama hatalarının daha ayrıntılı biçimde analiz edilebilmesi amacıyla karmaşıklık matrisleri de incelenmiştir.

B. Özellik Seçimi Algoritmalarının Çalıştırılması, En Değerli ve En Değersiz Özellikler

Bu çalışmada kullanılan özellik seçimi algoritmaları, her bir özelliğin sınıf değişkeni olan kokain kullanımı üzerindeki ayırt edicilik gücünü farklı bakış açılarıyla değerlendirmiştir. Aşağıda her bir özellik seçimi algoritmasının en değerli bulduğu 20 özellik listelenmiştir:

Görsel III.B.1. Ki-Kare Testi Algoritmasının Belirlediği En Değerli 20 Özellik

Ki-Kare Testi algoritmasının en değerli bulduğu 20 özellik:

	feature	chi2_score
28	VSA	5.416264
20	Ecstasy	1.627103
19	Crack	1.497346
3	Country	1.352482
17	Chocolate	1.066078
21	Heroin	0.811662
22	Ketamine	0.678674
24	LSD	0.675168
12	Alcohol	0.517096
16	Cannabis	0.506823
23	Legal_highs	0.469846
0	Age	0.468094
14	Amyl_nitrite	0.311435
26	Mushrooms	0.299053
29	Semeron	0.261184
27	Nicotine	0.218980
15	Benzodiazepines	0.197157
2	Education	0.184572
25	Methadone	0.150993
7	Oscore	0.133712

Ki-Kare Testi algoritması, kokain kullanan bireyleri belirlemede en çok fayda sağlayan özellikler olarak VSA, Ekstazi, Crack, alkol, esrar ve nikotin kullanım düzeylerini ve yaş, eğitim düzeyi gibi şahısların demografik özelliklerini de sınıflama esnasında kullanılmaya değer özellikler olarak belirlemiştir. Bir diğer özellik seçimi algoritması olan Bilgi Kazancı ise:

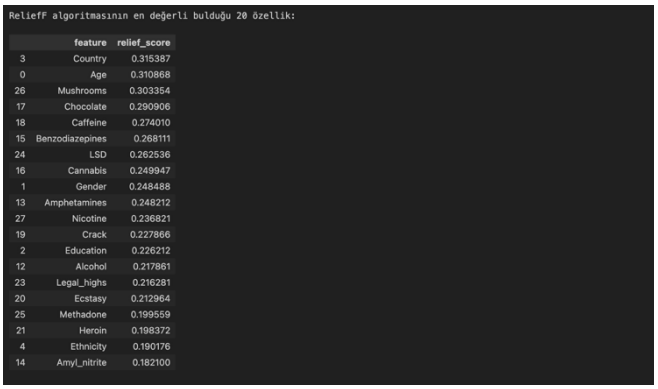
Görsel III.B.2. Bilgi Kazancı Algoritmasının Belirlediği En Değerli 20 Özellik

Information Gain algoritmasının en değerli bulduğu 20 özellik:

	feature	mi_score
4	Ethnicity	0.005860
7	Oscore	0.005546
3	Country	0.005150
1	Gender	0.004134
15	Benzodiazepines	0.002768
9	Cscore	0.002151
27	Nicotine	0.002102
24	LSD	0.001977
10	Impulsiveness	0.001962
8	Ascore	0.001922
18	Caffeine	0.001902
12	Alcohol	0.001619
11	SS	0.001562
16	Cannabis	0.001175
25	Methadone	0.001149
17	Chocolate	0.000719
5	Nscore	0.000693
23	Legal_highs	0.000461
26	Mushrooms	0.000453
13	Amphetamines	0.000217

şahısların etnik kökeni, Deneyime Açıklık skorları, Nevrotiklik skorları, ülkeleri, cinsiyetleri gibi demografik ve davranışsal özelliklerinin ve nikotin, benzodiazepin, kafein ve alkol gibi maddelerin kullanım düzeylerinin, kokain kullanım düzeyini sınıflama esnasında diğer özelliklerden daha fazla katkı sağladığını ortaya çıkarmıştır. Son olarak:

Görsel III.B.3. ReliefF Algoritmasının Belirlediği En Değerli 20 Özellik



ReliefF algoritmasının en değerli bulduğu 20 özellik:

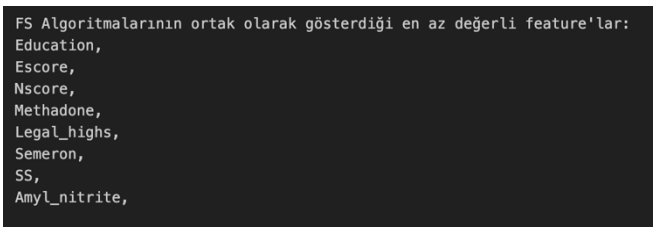
	feature	relief_score
3	Country	0.315387
0	Age	0.310868
26	Mushrooms	0.303354
17	Chocolate	0.290906
18	Caffeine	0.274010
15	Benzodiazepines	0.268111
24	LSD	0.262536
16	Cannabis	0.249947
1	Gender	0.248488
13	Amphetamines	0.248212
27	Nicotine	0.236821
19	Crack	0.227866
2	Education	0.226212
12	Alcohol	0.217861
23	Legal_highs	0.216281
20	Ecstasy	0.212964
25	Methadone	0.199559
21	Heroin	0.198372
4	Ethnicity	0.190176
14	Amyl_nitrite	0.182100

ReliefF algoritması; şahısların ülkesi, yaşı, cinsiyeti, eğitim düzeyi ve etnik kökeni gibi demografik özelliklerini ve psikoaktif etkiler yaratan mantar, çikolata, kafein, esrar, nikotin gibi uyuşturucu maddelerin kullanımını da kokain kullanımının sınıflandırılmasında kullanılabilecek değerli özellikler olarak belirlemiştir.

Her bir özellik seçimi algoritması, *III.A.* bölümünde belirtildiği üzere farklı yöntem ve metodolojiler kullanarak özelliklere belirli skorlar vermektedir. Bu sonuçlar, bir veri kümesindeki özelliklerin sınıf ayırmadaki kıymetini belirlemek için tek bir yöntemin kullanılamayacağını, kesin bir sonuca ulaşılabilmesi için birden fazla metodun denenmesi ve en optimal sonucun elde edilmesi gerektiğini de kanıtlar niteliktedir.

Şimdi ise her bir algoritma, kendi ölçütlerine göre belirlediği düşük katkı sağlayan özellikleri elde etmiş; bu özellikler, sınıflama performansına sınırlı düzeyde katkı sunan özellikler olarak yorumlanmıştır. Özellik seçimi sürecinde, yalnızca tek bir algoritmanın düşük önemde bulduğu değişkenleri elemek yerine, üç farklı özellik seçimi algoritmasının ortak olarak düşük katkılı olarak değerlendirdiği özellikler dikkate alınmıştır. Aşağıdaki görsel, üç farklı özellik seçimi algoritmasının kendi içerisinde değersiz bulduğu *ilk 20 özellikten ortak olan 8 adedini* göstermektedir:

Görsel III.B.4. Özellik Seçimi Algoritmalarının Ortak Olarak Belirlediği En Az Değerli Özellikler



```
FS Algoritmalarının ortak olarak gösterdiği en az değerli feature'lar:
Education,
Escore,
Nscore,
Methadone,
Legal_highs,
Semeron,
SS,
Amyl_nitrite,
```

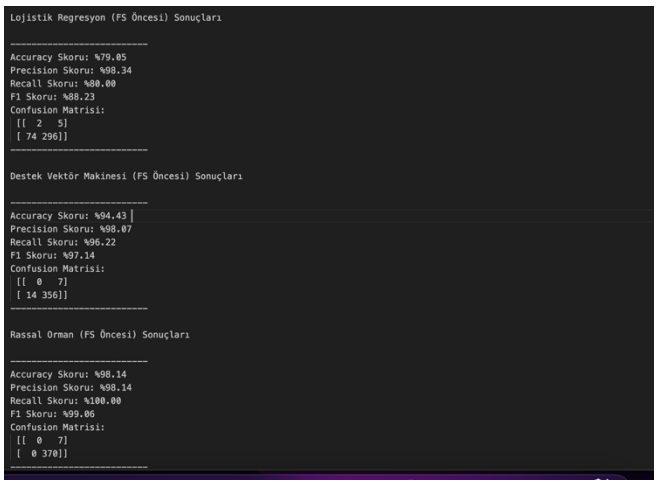
Görsel III.B.4.'de de görüldüğü üzere, kokain kullanıcılarını sınıflamada en az katkıyı sağlayan özellikler sırasıyla eğitim düzeyi, Dışadönüklük skoru, Nevrotiklik skoru, metadon kullanım düzeyi, yasal uyuşturucu kullanım düzeyi, semeron kullanım düzeyi, Duyum Arayışı skoru ve amil nitrat kullanım düzeyi olarak açıklanabilir.

Analizin bir sonraki aşamasında oluşturulan modeller veri kümesinin hem orijinal haliyle hem de yukarıdaki özelliklerin veri kümesinden çıkarılmış haliyle iki kez eğitilmiş ve herhangi bir performans düşüşü veya artışı yaşanıp yaşanmadığı gözlemlenmiştir.

C. Özellik Seçimi Algoritmaları Çalıştırılmadan Önce Model Performansları

Bu bölümde herhangi bir özellik seçimi algoritması kullanılmadan veri kümesi üzerindeki tüm bağımsız değişkenler kullanılarak eğitilen sınıflama modellerinin performansları incelenmiştir. Bir sonraki görselde; modellerin, yukarıda bahsedilen metrikler bazında elde ettiği skorlar görülmektedir:

Görsel III.C.1. Özellik Seçimi Algoritmaları Çalıştırılmadan Önce Model Performansları



```
Lojistik Regresyon (FS Öncesi) Sonuçları

--------------------------
Accuracy Skoru: %79.05
Precision Skoru: %98.34
Recall Skoru: %80.00
F1 Skoru: %88.23
Confusion Matrisi:
 [[  2   5]
 [ 74 296]]
--------------------------

Destek Vektör Makinesi (FS Öncesi) Sonuçları

--------------------------
Accuracy Skoru: %94.43
Precision Skoru: %98.07
Recall Skoru: %96.22
F1 Skoru: %97.14
Confusion Matrisi:
 [[  0   7]
 [ 14 356]]
--------------------------

Rassal Orman (FS Öncesi) Sonuçları

--------------------------
Accuracy Skoru: %98.14
Precision Skoru: %98.14
Recall Skoru: %100.00
F1 Skoru: %99.06
Confusion Matrisi:
 [[  0   7]
 [  0 370]]
--------------------------
```

Görsel III.C.1.'de görüldüğü üzere, özellik seçimi algoritmaları uygulanmadan önce LR algoritması %79,05 oranında doğruluk skoruna, %98,34 oranında kesinlik skoruna, %80 oranında duyarlılık skoruna ve %88,23 oranında bir F1 skoru değerine sahiptir. SVM algoritması %94,43 oranında doğruluk skoruna, %98,07 oranında kesinlik skoruna, %96,22 oranında duyarlılık skoruna ve %97,14 oranında bir F1 skoru değerine sahiptir. Son olarak RF algoritması ise %98,14 oranında doğruluk skoruna, %98,14 oranında kesinlik skoruna, %100 oranına duyarlılık skoruna ve %99,06 oranında bir F1 skoru değerine sahiptir.

Bu sonuçlar birlikte değerlendirildiğinde, doğrusal bir sınıflama algoritması olan Lojistik Regresyon algoritmasının, veri kümesindeki karmaşık ve doğrusal olmayan ilişkileri yakalamakta sınırlı kaldığı görülmektedir. Her ne kadar LR algoritması yüksek bir kesinlik skoruna sahip olsa da, kullanılan diğer algoritmalara kıyasla aldığı düşük duyarlılık skoru, özellikle aktif kokain kullanıcılarının bir kısmını doğru bir biçimde tespit edemediğine işaret etmektedir. Bu durum, modelin pozitif sınıfı ayırt etme konusunda temkinli davrandığını ve bazı gerçek pozitif örnekleri negatif olarak sınıflandırdığını göstermektedir. Destek Vektör Makinesi algoritması ise Lojistik Regresyon'a kıyasla daha dengeli bir performans sergilemiş, hem yüksek doğruluk hem de yüksek duyarlılık değerleri ile aktif kokain kullanıcılarını daha başarılı biçimde ayırt edebilmiştir. Bu durum, SVM'nin yüksek boyutlu veri kümelerinde sınıflar arasındaki sınırları daha etkin bir şekilde belirleyebilme yeteneğiyle ilişkilendirilebilir. Özellikle kişilik skorları ve davranışsal değişkenler gibi doğrusal olmayan örüntüler içeren veri kümelerinde, SVM'nin güçlü bir sınıflayıcı olduğu görülmektedir.

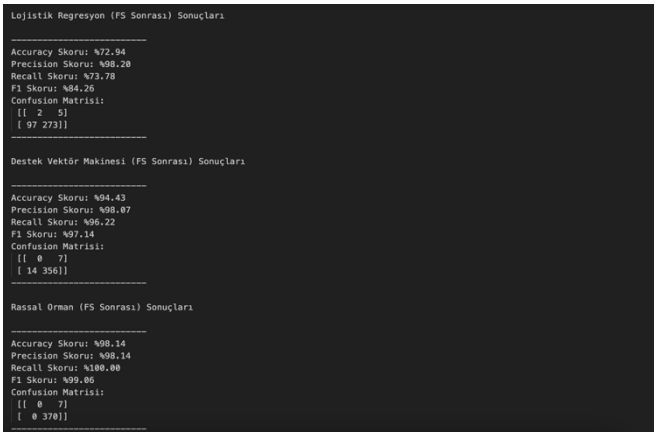
Rassal Orman algoritması ise tüm metrikler bazında en yüksek performansı sergileyerek çalışmada kullanılan modeller arasında öne çıkmıştır. RF'nin %100 duyarlılık

değerine ulaşması, test kümesindeki tüm aktif kokain kullanıcılarının doğru biçimde sınıfladığını göstermektedir. Bu üstün performans, Rassal Orman algoritmasının birden fazla karar ağacını bir araya getirerek karmaşık ilişkileri başarılı bir biçimde modelleyebilme yeteneğinden kaynaklanmaktadır. Ayrıca, RF algoritmasının özellikler arası etkileşimleri doğal biçimde yakalayabilmesi, davranışsal ve demografik değişkenlerin birlikte değerlendirilmesi gereken bu tür veri kümeleri için önemli bir avantaj sağlamaktadır.

D. Özellik Seçimi Algoritmaları Çalıştırıldıktan Sonra Model Performansları

Bu bölümde ise, özellik seçimi algoritmaları uygulandıktan ve belirlenen değersiz özellikler veri kümesinden çıkarıldıktan sonraki model performansları incelenmiş ve yorumlanmıştır. Bir sonraki görselde; modellerin, daha önce bahsedilen metrikler bazında elde ettiği skorlar görülmektedir:

Görsel III.D.1. Özellik Seçimi Sonrası Modellerin Elde Ettiği Skorlar



Görsel III.D.1.'de görüldüğü üzere, özellik seçimi algoritmaları uygulandıktan sonra LR algoritması %72,94 oranında doğruluk skoruna, %98,20 oranında kesinlik skoruna, %73,78 oranında duyarlılık skoruna ve %84,26 oranında bir F1 skoru değerine sahiptir. SVM algoritması %94,43 oranında doğruluk skoruna, %98,07 oranında kesinlik skoruna, %96,22 oranına duyarlılık skoruna ve %97,14 oranında bir F1 skoru değerine sahiptir. RF algoritması ise özellik seçimi algoritmaları uygulanmadan önceki metrik skorlarını koruyarak %98,14 oranında doğruluk skoruna, %98,14 oranında kesinlik skoruna, %100 oranına duyarlılık skoruna ve %99,06 oranında bir F1 skoru değerine sahiptir.

LR modelinde özellikle doğruluk ve duyarlılık metriklerinde özellik seçimi öncesine kıyasla sınırlı da olsa bir performans düşüşü yaşandığı gözlemlenmektedir. SVM algoritması, mevcut değerleriyle özellik seçimi öncesindeki değerlerini neredeyse tamamen korumuştur. Bu durum, SVM algoritmasının daha az sayıda ancak daha ayırt edici özelliklerle çalışmaya elverişli yapısını ve özellik seçimi sürecine görece daha dayanıklı olduğunu göstermektedir. RF algoritması, aldığı skorlar ile tüm metrikler açısından en yüksek performansı sergilemeye devam etmiştir. Bu sonuç, Rassal Orman algoritmasının çoklu karar ağaçlarına dayalı yapısı sayesinde düşük katkı sağlayan özelliklerin model performansı üzerindeki olumsuz etkilerini doğal olarak dengeleyebildiğini ortaya koymaktadır.

Sonuçlandırmak gerekirse; özellik seçimi algoritmalarının amacı her zaman model metriklerinin değerlerini artırmak değil, benzer doğruluk değerlerine daha düşük hesaplama maliyeti, daha sade ve daha düşük boyutlu modeller ile ulaşmaktır. Dolayısıyla bu çalışmada elde edilen sonuçlar, özellik seçimi sonrasında özellikle Rassal Orman algoritmasının performansını koruduğunu, dolayısıyla modelin daha az sayıda özellik ile benzer tahmin gücüne ulaşabildiğini göstermektedir. Bu durum, özellik seçiminin maliyet- performans dengesi açısından başarılı bir biçimde uygulandığının açıkça gözlemlenebilmesine olanak sağlamıştır.

E. Bulguların Literatür ile Karşılaştırılması

Elde edilen bulgular, literatür taraması bölümünde ele alınan çoklu madde kullanımı ve davranışsal ve demografik risk faktörlerine ilişkin teorik çerçeve ile büyük ölçüde örtüşmektedir. Özellikle çalışmada birer özellik kullanılan uyuşturucu kullanım düzeyleri ile kokain kullanım düzeyi arasındaki eş zamanlılık, başka bir deyişle bu değişkenlerin özellik seçimi algoritmaları tarafından değerli özellikler olarak nitelendirilmesi, literatür bölümünde vurgulanan çoklu madde kullanım örüntülerini destekler niteliktedir. Bu durum, madde kullanımının tekil bir davranış olmaktan ziyade birbirini tetikleyen ve birlikte ortaya çıkan bir yapı sergilediği yönündeki bulgularla uyumludur.

Yaş, cinsiyet, eğitim düzeyi, ülke, etnik köken gibi demografik özellikler veya Duyum Arayışı skoru ve Dürtüsellik gibi davranışsal özellikler de uyuşturucu madde kullanımlarına benzer olarak özellik seçimi algoritmaları tarafından değerli özellikler olarak belirlenmiştir. Örneğin Ki-Kare Testi algoritması, yaş ve eğitim düzeyi özelliklerini en değerli 20 özellik arasında, Bilgi Kazancı algoritması etnik köken, ülke, cinsiyet, Dürtüsellik skoru gibi özellikleri en değerli 20 özellik arasında ve ReliefF algoritması ise ülke, yaş, cinsiyet, eğitim düzeyi ve etnik köken özelliklerini en değerli 20 özellik arasında barındıracak şekilde puanlamıştır.

Buna karşılık, literatürde madde kullanımına yatkınlık açısından önemli olduğu belirtilen Beş Faktör Kişilik Modeline ait özelliklerin, sınıflama odaklı bu çalışmada özellik seçimi algoritmaları tarafından düşük ayırt ediciliğe sahip olarak değerlendirilmesi dikkat çekicidir. Örneğin Ki-Kare Testi algoritması, Beş Faktör Kişilik Modelinden yalnızca Deneyime Açıklık özelliğini en değerli 20 özellik içerisinde belirtmiş, modelde bulunan diğer özellikleri sınıf ayırmada değerli olan özellikler olarak belirlememiştir. Bilgi Kazancı algoritması da benzer olarak bu modelden Deneyime Açıklık ve Sorumluluk özelliklerini sınıf ayırmak için değerli özellikler olarak belirlemiş, modelin diğer özellikleri bu listede yer almamıştır. ReliefF algoritması ise modelden hiçbir özelliği listeye dahil etmemiştir. Bu durum, kişilik değişkenlerinin davranışın ortaya çıkışını açıklamada önemli olabileceğini ancak mevcut kullanım durumunu tahmin etmede diğer madde kullanım göstergelerine kıyasla daha dolaylı bir rol üstlendiği yorumunun yapılmasına olanak sağlayabilir.

IV. SONUÇ VE DEĞERLENDİRME

Bu çalışma kapsamında, UCI Drug Consumption veri kümesi kullanılarak bireylerin kokain kullanım durumunun tahmin edilmesi amaçlanmış; farklı sınıflama algoritmaları ve özellik

seçimi yöntemleri karşılaştırmalı olarak incelenmiştir. Çalışmanın temel motivasyonu, davranışsal, demografik ve çoklu madde kullanımına ilişkin değişkenlerin birlikte ele alındığı bir veri kümesi üzerinde, veri madenciliği tekniklerinin etkinliğini değerlendirmek ve bu tekniklerin kokain kullanım düzeyini tahmin etme başarısını ortaya koymaktır.

Elde edilen bulgular, kokain kullanımının diğer maddelerin kullanımı ile güçlü bir eş zamanlılık sergilediğini açıkça göstermektedir. Bu sonuçlar, literatürde sıklıkla vurgulanan çoklu madde kullanımı örüntüsü ile büyük ölçüde örtüşmekte ve madde kullanım davranışlarının tekil değil, birbirini tetikleyen çok boyutlu bir yapı sergilediğini desteklemektedir. Makine öğrenmesi algoritmaları açısından değerlendirildiğinde, Rassal Orman algoritmasının hem özellik seçimi öncesinde hem de sonrasında en yüksek performansı sergilediği görülmüştür. RF algoritması; doğruluk, kesinlik, duyarlılık ve F1 skoru gibi tüm metriklerde istikrarlı ve yüksek sonuçlar üretmiş; özellik seçimi sonrasında da performansını koruyarak daha az sayıda değişkenle benzer tahmin gücüne ulaşabilmiştir. Destek Vektör Makinesi algoritması, özellik seçimi sonrasında performansını büyük ölçüde korumuş; Lojistik Regresyon algoritması ise özellik seçimi sonrası az da olsa bir performans düşüşü yaşamıştır. Bu durum, kullanılan algoritmaların veri kümesinin yapısına ve özellik sayısına karşı farklı derecelerde hassasiyet gösterdiğini ortaya koymaktadır. Özellik seçimi algoritmalarının sonuçları incelendiğinde; Bilgi Kazancı, ReliefF ve Ki-Kare yöntemlerinin ortak olarak bazı kişilik özelliklerini düşük ayırt edici özellik olarak değerlendirdiği görülmüştür. Literatürde madde kullanımına yatkınlık açısından önemli olduğu belirtilen bu değişkenlerin, sınıflandırma odaklı bu çalışmada daha zayıf katkı sağlaması dikkat çekicidir. Bu bulgu, kişilik özelliklerinin davranışın ortaya çıkışını açıklamada önemli olabileceğini, ancak mevcut kullanım durumunu tahmin etmede doğrudan ve güçlü göstergeler olan diğer madde kullanım değişkenlerine kıyasla daha dolaylı bir rol üstlendiğini göstermektedir.

Özellik seçimi algoritmalarının genel performansına göz atılacak olursa; özellik seçimi algoritmalarının amacı her zaman model performansını artırmak değil; benzer doğruluk değerlerine daha sade, daha düşük boyutlu ve daha düşük hesaplama maliyetine sahip modeller ile ulaşmaktır. Bu çalışma kapsamında elde edilen sonuçlar, özellikle Rassal Orman algoritmasının özellik seçimi sonrasında performansını koruyabildiğini ve bu yönüyle maliyet-performans dengesi açısından başarılı bir model sunduğunu göstermektedir. Çalışma, davranışsal ve çok boyutlu veri kümelerinde farklı sınıflama ve özellik seçimi yaklaşımlarının etkisini ortaya koyarak, veri madenciliği tekniklerinin madde kullanım davranışlarının analizinde etkin biçimde kullanılabileceğini göstermektedir.

V. KAYNAKÇA

[1]: Doç. Dr. Nesrin Dilbaz & Uzm. Dr. Aslı Enez Darçın (2011) Şizofreni ve Madde Kullanım Bozukluğu Eş Tanılı Hastalarda Tedavi, Klinik Psikofarmakoloji Bülteni-Bulletin of Clinical Psychopharmacology, 21:1, 80-90, DOI: 10.5350/KPB-BCP201121114

[2]: Çevik, M., & Kızmaz, Z. (2021). UYUŞTURUCU BAĞIMLILARININ DEMOGRAFİK ÖZELLİKLERİ VE MADDE KULLANIM ALIŞKANLIKLARI. Adıyaman Üniversitesi Sosyal Bilimler Enstitüsü Dergisi(37), 470-506. https://doi.org/10.14520/adyusbd.823646

[3]: Sağlam, L. (2017). Nikotin Bağımlılığının Klinik Değerlendirilmesi. Güncel Göğüs Hastalıkları Serisi, 4(1), 78- 89.

[4]: Yüncü Z, Kesebir S, Özbaran B, Çelık Y, Aydın C. 2009. Madde Kullanım Bozukluğu Olan Ergenlerin Ebeveynlerinde Psikopatoloji ve Mizaç: Kontrollü Bir Çalışma. Turk Psikiyatri Dergisi, 20

[5]:Karataş, Z. (2020). Madde Kullanım Bozukluğu Olan Yetişkinlerin Sorunlarının Açıklanmasında Aile İşlevleri ve Çeşitli Demografik Değişkenlerin Rolü. Toplum ve Sosyal Hizmet, 31(1), 70-105.

VI. EKLER

Analize ait kaynak kodlarının bulunduğu GitHub Repository: https://github.com/kdairutt/UCI-Drug-Consumption-Analysis

UCI Drug Consumption Data Set: https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified