

Çeşitli Çekirdek Fonksiyonları ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bir Klinik Uygulama

Emek GÜLDOĞAN¹, Ahmet Kadir ARSLAN^{1,a}, Jülide YAĞMUR²

¹İnönü Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya, Türkiye.

²İnönü Üniversitesi, Turgut Özal Tıp Merkezi, Kardiyoloji Anabilim Dalı, Malatya, Türkiye.

ÖZET

Amaç: Bu araştırmanın birincil amacı; çeşitli çekirdek fonksiyonları ile oluşturulan destek vektör makinesi modellerinin, Akut Koroner Sendromlu hastalarda diabetes mellitusu sınıflandırma performanslarının incelenmesi ve karşılaştırılmasıdır. Bu araştırmanın ikincil amacı ise, destek vektör makinesi modeli oluşturulurken kullanılan çeşitli çekirdek fonksiyonlarının parametrelerinin optimize edilerek en iyi sınıflandırma performansını elde etmeve çalışmaktır.

Gereç ve Yöntem: Bu çalışmada incelenen veriler, İnönü Üniversitesi Turgut Özal Tıp Merkezi Kardiyoloji Anabilim Dalı için geliştirilen veritabanından geriye yönelik (retrospektif) olarak seçilmiştir. Çalışmadaki söz konusu veriler Akut Koroner Sendromlu hastalarda tip 2 diabetes mellitus ile değişik demografik ve klinik değişkenleri içermektedir. Akut Koroner Sendromlu hastalarda tip 2 diabetes mellitus'un sınıflandırılması için Destek Vektör Makinesi modelleri kullanılmıştır. İlgili modeller, ANOVA radyal tabanlı fonksiyon, bessel, doğrusal, Gaussian radyal tabanlı fonksiyon, laplace, polinomiyal ve sigmoid çekirdekleri ile oluşturulmuştur.

Bulgular: Laplace çekirdek fonksiyonu ile oluşturulan en iyi sınıflama performansına sahip destek vektör makinesi modeline ilişkin doğruluk, ROC eğrisi altında kalan alan, duyarlılık ve özgüllük [seçicilik] ölçütleri ile % 95 güven aralığı değerleri sırasıyla; 0.9804 (0.9716 - 0.987), 0.9332 (0.9096 - 0.9567), 0.9999 (0.9791 - 1.000) ve 0.9776 (0.9675 - 0.9852) olarak elde edilmiştir.

Sonuç: İncelenen değişik çekirdek fonksiyonları ile oluşturulan modeller arasında söz konusu performans ölçütleri dikkate alındığında, en iyi sınıflama performansı laplace Destek Vektör Makinesi modelinden elde edilmiştir. İlerleyen çalışmalarda, farklı klinik verilerde değişik çekirdek fonksiyonlu Destek Vektör Makinesi modelleri ile diğer makine öğrenmesi ya da veri madenciliği algoritmalarının kullanılması hastalıkların sınıflandırma başarısını artırabilecektir.

Anahtar Sözcükler: Destek Vektör Makinesi, Çekirdek Fonksiyonları, Akut Koroner Sendromu, Tip II Diabetes Mellitus.

ABSTRACT

The Performance Exploration of Support Vector Machines Models Constructed with Various Kernel Functions: A Clinical Application Objective: The primary aim of this study is to examine and compare the classification performance of support vector machine models generated by various core functions used to classify diabetes mellitus in acute coronary syndrome patients. The secondary aim is to optimize the parameters of the various kernel functions which are used for constructing the support vector machine model and to achieve the best classification performance.

Material and Method: The data examined in this study were selected retrospectively from the database developed for İnönü University Turgut Özal Medical Center Cardiology Department. The study included type 2 diabetes mellitus and various demographic and clinical variables in acute coronary syndrome patients. The Support Vector Machine model was used to classify type 2 diabetes mellitus in acute coronary syndrome patients. The related models are constructed by ANOVA radial basis function, bessel, linear, Gaussian radial basis function, laplace, polynomial and sigmoid kernel functions.

Results: The best classification performance was obtained by Support Vector Machine model constructed by laplace kernel function based on the results of performance metrics. The accuracy, area under ROC curve, sensitivity and specificity metrics with 95% CI were calculated as; 0.9804 (0.9716 - 0.987), 0.9332 (0.9096 - 0.9567), 0.9999 (0.9791 - 1.000) and 0.9776 (0.9675 - 0.9852), respectively.

Conclusion: When the performance metrics were taken into account, the best classification performance was achieved from the Laplace Support Vector Machine model. In subsequent studies, the use of Support Vector Machine models with different kernel functions and other machine learning or data mining algorithms in different clinical trials may improve the classification success of the diseases.

Keywords: Support Vector Machine, Kernel Functions, Acute Coronary Syndrome, Type II Diabetes Mellitus.

Akut Koroner Sendrom (AKS), acil servise başvurulmanın ve hastanelerin koroner yoğun bakım birimleri başta olmak üzere hastaların yatırılma sebeplerinin başında yer almaktadır. Akut Koroner Sendromun neden olduğu iş gücü kaybı, morbidite ve mortaliteye ek olarak yüksek hastane maliyetlerine sebep olması açısından günümüzde en önemli sağlık problemlerinden biri olarak kabul edilmektedir (1). Akut koroner Sendromu, kalp kasının bir kısmının düzgün çalışmaması veya işlevinin kaybolması gibi koroner atardamarlarda

azalmış kan akışına bağlı olarak ortaya çıkan bir sendromdur. En sık rastlanan belirtilerinden bazıları çoğunlukla sağ omuza yayılan göğüs ağrısı ile bulantı ve terlemedir. Akut Koroner Sendromu, genellikle ST yükselmeli miyokard enfarktüsü (STEMI, % 30), ST yükselmeli olmayan miyokard enfarktüsü (NSTEMI, % 25) veya kararsız angina (% 38) olmak üzere üç nedenden ortaya çıkmaktadır. Bu tipler, EKG / EKG görünümüne göre ST segment yükselmeli olmayan miyokard enfarktüsü ve ST segment yükselmeli miyokard enfarktüsü olarak adlandırılır (2, 3).

^aYazışma Adresi: Ahmet Kadir ARSLAN, İnönü Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya, Türkiye
Tel: 0422 341 0660
Geliş Tarihi/Received: 28.05.2017
e-mail: ahmetkadirarslan@gmail.com
Kabul Tarihi/Accepted: 03.08.2017

Diabetes mellitus (DM), kandaki glikoz düzeyinin yükselmesiyle neticelenen, çoğunlukla kalıtsal ve çevre ile ilgili faktörlerin bir araya gelmesi ile meydana gelen bir hastalıktır. Dünya Sağlık Örgütü (WHO) verilerine göre, diyabetli insanların sayısı 1980'de 108 milyon iken, 2014'te bu sayı 422 milyona; hastalığın 18 yaş üstü erişkinlerde genel prevalansı ise 1980'de % 4.7 iken, 2014'de % 8.5'e yükselmiştir. Diyabetin 2030'da tüm dünyadaki ölüm nedenleri arasında 7. sırada olacağı öngörülmektedir. Orta ve düşük gelirli ülkelerde, diyabet prevalansı hızla artmaktadır. Ayrıca, diyabet; körlük, böbrek yetmezliği, kalp krizi, inme ve alt ekstremitte amputasyonunun ana nedenlerindedir (4).

Destek vektör makinesi modelleri son zamanlarda popüler bir algoritma haline gelmektedir. Destek vektör makineleri, sınıflandırma veya regresyon problemleri için kullanılabilen denetimli/danışmanlı bir makine öğrenme/veri madenciliği algoritmasıdır. Bu süreçte verileri dönüştürmek için çekirdek çözümü (kernel trick) adı verilen bir teknik kullanır. Bu çekirdek çözümü, veri dönüşümlerine dayanarak olası çıktılar arasında en uygun sınırı bulur. Başka bir ifadeyle, çekirdek çözümü son derece karmaşık veri dönüşümleri yapar ve daha sonra bu verilerin, tanımlanan etiketler veya çıktılara dayalı olarak nasıl ayrıldığı belirlenir. Destek vektör makineleri, optimal bir hiperdüzlem oluşturmak için hata fonksiyonunu en aza indirmek amacıyla kullanılan yinelemeli bir eğitim algoritması kullanır. Bu sayılan esnek ve önemli özellikleri nedeniyle, bu çalışmada ele alınan sınıflama probleminin çözümünde değişik çekirdek fonksiyonlarının tahmin performanslarını incelemek için destek vektör makineleri seçilmiştir.

Bu araştırmanın birincil amacı; çeşitli çekirdek fonksiyonları ile oluşturulan destek vektör makinesi modellerinin, Akut Koroner Sendromlu hastalarda diabetes mellitus'u sınıflandırma performanslarının incelenmesi ve karşılaştırılmasıdır. Bu araştırmanın ikincil amacı ise, destek vektör makinesi modeli oluşturulurken kullanılan çeşitli çekirdek fonksiyonlarının parametrelerini optimize ederek en iyi sınıflandırma performansını elde etmeye çalışmaktır.

GEREÇ VE YÖNTEM

Bu çalışma, Malatya Klinik Araştırmalar Etik Kurulu'nun 2016/161 protokol numaralı izni ile onaylanmıştır. Bu çalışmada, anabilim dalımız tarafından Kardiyoloji Anabilim Dalı için geliştirilen PHP tabanlı veri giriş, sorgulama, silme, güncelleme, vb. işlemleri yapan yazılım kullanıldı. Bu kapsamda, Akut Koroner Sendromlu hastalarda DM'nin varlığı ve yokluğu, aşağıda belirtilen değişkenlere dayalı olarak sınıflandı. Bu değişkenlere ait tanımlayıcı bilgiler Tablo 1'de verilmiştir.

İki grup arası tahmini glikoz düzeyi farkı 20, varsayılan ortak standart sapması 55, tip I hata (alfa) 0.05 ve tip II hata (beta) 0.10 olduğunda, her grupta en az 160 olmak üzere toplamda en az 320 birey gerektiği güç analizi ile hesaplandı (5). Bu planlanan araştırma kapsamına 1378 bireye ilişkin veriler alınmıştır.

Bu çalışmada uygulanan veritabanlarında bilgi keşfi (VTBK) süreci aşamaları aşağıda verilmiştir:

- Veri seçimi:** Veritabanlarından elde edilen ve bu çalışmada bağımlı/hedef değişken olarak DM, bağımsız/tahminleyici değişkenler olarak ise Tablo 1'de detaylıca açıklanan faktörler veri seçimi kapsamında incelenmiştir.
- Veri önileme:** Veri setindeki kayıp değer bulunan sınıflar verisetinden çıkarılmıştır. Aşırı/aykırı değer tespiti lokal aykırı faktörü (LOF) analizi ile yapılmış ve tespit edilen aşırı/aykırı değerler verisetinden çıkarılmıştır. Aşırı/aykırı değer tespitinde, yerel aykırı faktör (LOF) algoritması (6) kullanıldı. LOF, yakın zamanda geliştirilmiş olan yoğunluğa dayalı aşırı/aykırı gözlem tespitinde kullanılan yöntemlerden biridir. LOF, diğer aşırı/aykırı gözlem saptama algoritmalarıyla karşılaştırıldığında birçok avantaja sahiptir. LOF, bir gözlemin diğer komşu nesnelere olan uzaklığını bir sayısal ölçüğe dönüştürerek, yerel aşırı/aykırı gözlemlerin tespit edilmesini sağlar (7).
- Veri dönüşümü ve veri indirgeme:** Veri setindeki sayısal değişkenler için standardizasyon (Z-dönüşümü) uygulanmıştır. Standardizasyona ilişkin denklem aşağıda verilmiştir:

$$x_{\text{yeni}} = \frac{x - \bar{x}}{s}$$

Burada \bar{x} : örneklem aritmetik ortalamasını ve s : örneklem standart sapmasını ifade etmektedir (8).

- Veri madenciliği:** Çeşitli çekirdek fonksiyonları kullanılarak oluşturulan DVM modelleri kullanılarak veri setinden ilişki, örüntüler çıkarma ve tahminler yapılmıştır. Bu çalışmada çeşitli DVM modelleri, Tablo 2'de ayrıntılı şekilde tanımlanan değişik çekirdek fonksiyonları ile oluşturulmuş ve incelenen veri setine uygulanmıştır. Hiperparametre optimizasyonu için en yaygın ve en çok bilinen ızgara araması (grid search) yönteminde, hiperparametreler tespit edilen sabit bir adım büyüklüğü ile kullanılabilir olan en büyük aralıkta uygulanır ve her bir sıralanış biçimi (kombinasyon) için performans ölçütlerine göre değerlendirilir (9). Tablo 3'de DVM modelinin oluşturulmasında kullanılan çekirdek fonksiyonlarının optimizasyon parametreleri, aralıkları ve kombinasyon sayıları verilmiştir.

Tablo 1: Çalışmada kullanılan değişkenlerin tanımları

	Değişkenler	Değişken tipi	Değişken Açıklaması	Değişken Rolü
Demografik	Diabetes mellitus	Kategorik	Var/Yok	Bağımlı/Hedef
	Akut koroner sendrom tipi	Kategorik	Unstable Angina/STEMI/NSTEMI	Bağımsız/T ahminleyici
	Yaş	Nümerik	Doğal sayı	Bağımsız/T ahminleyici
	Cinsiyet	Kategorik	Kadın/Erkek	Bağımsız/T ahminleyici
	Vücut kitle indeksi	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
Hastalık öyküsü	Hipertansiyon	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Sigara içme durumu	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Renal yetmezlik öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Miyokardiyal enfarktüs öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Akut koroner sendrom aile öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Malignite öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Hiperlipidemi öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Periferik arter öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Koroner arter by-pass öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	İnme hastalığı öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici
Kalp yetmezliği öyküsü	Kategorik	Var/Yok	Bağımsız/T ahminleyici	
İlaç öyküsü	Aspirin	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Clopidogrel	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Beta bloker	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Kalsiyum kanal bloker	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Statin	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Digoksin	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Anjiyotensin dönüştürücü enzim inhibitör	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Anjiyotensin reseptör bloker diğer	Kategorik	Var/Yok	Bağımsız/T ahminleyici
Laboratuvar	Kreatinin	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Kan üre azotu	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Kolesterol	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Trigliserid	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Düşük yoğunluklu lipoprotein	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Yüksek yoğunluklu lipoprotein	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Sistolik kan basıncı	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Diastolik kan basıncı	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici
	Diüretik	Kategorik	Var/Yok	Bağımsız/T ahminleyici
	Glikoz	Nümerik	Pozitif reel sayı	Bağımsız/T ahminleyici

Tablo 2: Çekirdek tipleri ve fonksiyonları

Çekirdek tipi	Fonksiyon
Doğrusal	$k(x, y) = x^T y + c$
Polinomiyal	$k(x, y) = (\alpha x^T y + c)^d$
Hiperbolik Tanjant (Sigmoid)	$k(x, y) = \tanh(\alpha x^T y + c)$
Bessel	$k(x, x') = -Bessel_{(nu+1)}^n(\sigma x - x' ^2)$
ANOVA RBF	$k(x, y) = \sum_{k=1}^n \exp(-\sigma(x^k - y^k)^2)^d$
Gaussian RBF	$k(x, y) = \exp(-\gamma\ x - y\ ^2)$
Laplace	$k(x, y) = \exp\left(-\frac{\ x - y\ }{\sigma}\right)$

C (cost, maliyet) parametresi, ayırıcı hiperdüzlemin düzgünlüğü ile eğitim verilerinin yanlış sınıflandırılması arasındaki dengeyi kontrol eder (10). Maliyet parametresinin aksine σ (sigma), scale (ölçek) ve degree (derece) parametreleri çekirdek fonksiyonu parametreleridir. Sigma; Gaussian RBF, Laplace, ANOVA

RBF ve Bessel çekirdek fonksiyonlarının parametresi, ölçek; Hiperbolik Tanjant (Sigmoid) ve Polinomiyal çekirdek fonksiyonlarının parametresi, derece; Polinomiyal, ANOVA RBF ve Bessel çekirdek fonksiyonlarının parametresidir.

Tablo 3: Çekirdek fonksiyonlarının optimizasyon parametrelerine ilişkin tanımlayıcı bilgiler

Çekirdek Fonksiyonu	Optimizasyon Parametreleri	Parametre Aralığı	Kombinasyon Sayısı
Gaussian RBF	C, sigma	(2^{-2} - 2^5), (0.02-0.20)	24
Hiperbolik Tanjant (Sigmoid)	C, ölçek	(2^{-2} - 2^5), (10^{-3} - 10^4)	64
Polinomiyal	C, derece, ölçek	(2^{-2} - 2^5), (1-3), (10^{-3} - 10^4)	192
Doğrusal	C	(2^{-2} - 2^5)	8
Laplace	C, sigma	(2^{-2} - 2^5), (0.02-0.20)	24
ANOVA RBF	C, sigma, derece	(2^{-2} - 2^5), (0.02-0.20), (1-3)	72
Bessel	C, sigma, derece	(2^{-2} - 2^5), (0.02-0.20), (1-3)	72

- e. **Değerlendirme ve yorumlama:** Çeşitli çekirdek fonksiyonları kullanılarak oluşturulan DVM modellerinin tahmin performansları; 10-katlı çapraz geçerlilik tekniği ile değerlendirilmiştir. Ayrıca, performans metriklerinden doğruluk, duyarlılık, özgüllük (seçicilik) ve ROC eğrisi altında kalan kullanılarak ilgili model çıktıları yorumlanmıştır. Bu metriklerle ilişkin ayrıntılı formül tanımlamaları aşağıda verilmiştir:

$$\begin{aligned} \text{Doğruluk} &= (\text{DP}+\text{DN})/(\text{DP}+\text{DN}+\text{YP}+\text{YN}) \\ \text{Duyarlılık} &= \text{DP}/(\text{DP}+\text{YP}) \\ \text{Özgüllük (Seçicilik)} &= \text{DN}/(\text{DN}+\text{YN}) \\ \text{Pozitif tahmin değeri} &= \text{DP}/(\text{DP}+\text{YN}) \\ \text{Negatif tahmin değeri} &= \text{DN}/(\text{DN}+\text{YP}) \end{aligned}$$

Burada; DP: doğru pozitif sayısı, DN: doğru negatif sayısı, YP: yanlış pozitif sayısı, YN: yanlış negatif sayısı olarak tanımlanır.

VTBK sürecindeki analizlerde RStudio Version 1.0.143 yazılımı (11) ve yazılım içindeki ilgili paketler kullanılmıştır. Kullanılan paketler; LOF analizi için DMwR (12), veri madenciliği aşaması için caret (13) ve kernlab (14), ROC eğrisi çizimi için ggplot2 (15) ve ROCR (16) paketleridir.

Biyoistatistiksel Analiz

Nicel veriler ortalama ve standart sapma, nitel veriler ise sayı ve yüzde özetlendi. Verilerin normal dağılıma uygunluğu Kolmogorov Smirnov testi ve varyansların homojenlik kontrolü ise Levene testi ile değerlendirildi. Verilerin analizinde bağımsız örneklerde t-testi, Pearson Ki-kare testi, Yates Düzeltilmiş Ki-kare testi ve Fisher Kesin Ki-kare testi uygun olan yerlerde kesin (exact) yaklaşım dayalı olarak kullanıldı. Analizlerde IBM SPSS Statistics version 24.0 for Windows paket

programı (17) kullanıldı. $p<0.05$ değeri istatistiksel olarak önemli kabul edildi.

BULGULAR

Aşırı/aykırı değer analizi sonucunda 2 gözlem DM olmayan bireylerin oluşturduğu veri setinden çıkarılmıştır. Bu durumda dağılım Tablo 4'de gösterilmiştir.

Tablo 4: DM değişkenin dağılımı

DM	Sayı	Yüzde (%)
Yok	1176	85.3
Var	202	14.7
Toplam	1378	100.0

Nitel değişkenlerin DM'nin varlığı ve yokluğuna göre dağılımı Tablo 5'de verilmiştir. Bu bulgulara göre, cinsiyet, anjiyotensin dönüştürücü enzim inhibitör, anjiyotensin reseptör bloker diğer, aspirin, beta bloker, diüretik, statin, hipertansiyon, kalp yetmezliği öyküsü, miyokardiyal enfarktüs öyküsü, renal yetmezlik öyküsü ve sigara içme durumu değişkenleri açısından DM varlığı ve yokluğu arasında istatistiksel olarak fark vardır ($p<0.05$). Buna karşın, kalsiyum kanal bloker, clopidogrel, Akut Koroner Sendrom aile öyküsü, koroner arter by-pass öyküsü, hiperlipidemi, kalp yetmezliği öyküsü, malignite öyküsü, periferik arter öyküsü ve inme hastalığı öyküsü değişkenleri açısından DM varlığı ve yokluğu arasında istatistiksel olarak fark yoktur ($p>0.05$).

Nicel değişkenlerin DM'nin varlığına ve yokluğuna göre tanımlayıcı istatistikleri Tablo 6'da verilmiştir. Tablo 6 incelendiğinde, kan üre azotu, glikoz, yüksek yoğunluklu lipoprotein, vücut kitle indeksi değişkenleri açısından DM varlığı ve yokluğu arasında istatistiksel olarak fark varken ($p<0.05$), Düşük yoğunluklu lipop-

rotein, trigliserid, nabız, diastolik kan basıncı, sistolik kan basıncı, kolesterol değişkenleri açısından DM varlığı ve yokluğu arasında istatistiksel olarak farklılık yoktur ($p>0.05$).

DVM'nin tüm çekirdek fonksiyonlarının çeşitli performans metriklerine göre sınıflandırma performansları (% 95 güven aralığı ile) Tablo 7'de verilmiştir.

Tablo 5: Nitel değişkenlerin DM'nin varlığı ve yokluğuna göre dağılımı

Değişken adı/Kategorileri		Diabetes Mellitus				p
		Yok (n=1176)		Var (n=202)		
		Sayı	% *	Sayı	% *	
Cinsiyet	Kadın	224 ^a	19.0	69 ^b	34.2	<0.001
	Erkek	952 ^a	81.0	133 ^b	65.8	
Anjiyotensin dönüştürücü enzim inhibitör	Yok	1116 ^a	94.9	172 ^b	85.1	<0.001
	Var	60 ^a	5.1	30 ^b	14.9	
Anjiyotensin reseptör bloker diğer	Yok	1130 ^a	96.1	183 ^b	90.6	0.001
	Var	46 ^a	3.9	19 ^b	9.4	
Aspirin	Yok	1065 ^a	90.6	144 ^b	71.3	<0.001
	Var	111 ^a	9.4	58 ^b	28.7	
Beta bloker	Yok	1084 ^a	92.2	173 ^b	85.6	0.004
	Var	92 ^a	7.8	29 ^b	14.4	
Diüretik	Yok	1120 ^a	95.2	168 ^b	83.2	<0.001
	Var	56 ^a	4.8	34 ^b	16.8	
Statin	Yok	1146 ^a	97.4	187 ^b	92.6	0.001
	Var	30 ^a	2.6	15 ^b	7.4	
Hipertansiyon	Yok	655 ^a	55.7	77 ^b	38.1	<0.001
	Var	521 ^a	44.3	125 ^b	61.9	
Kalp yetmezliği öyküsü	Yok	1173 ^a	99.7	199 ^b	98.5	0.044
	Var	3 ^a	0.3	3 ^b	1.5	
Miyokardiyal enfarktüs öyküsü	Yok	1133 ^a	96.3	186 ^b	92.1	0.01
	Var	43 ^a	3.7	16 ^b	7.9	
Renal yetmezlik öyküsü	Yok	1168 ^a	99.3	196 ^b	97.0	0.01
	Var	8 ^a	0.7	6 ^b	3.0	
Sigara içme durumu	Yok	463 ^a	39.4	125 ^b	61.9	<0.001
	Var	713 ^a	60.6	77 ^b	38.1	
Kalsiyum kanal bloker	Yok	1128	95.9	195	96.5	0.827
	Var	48	4.1	7	3.5	
Clopidogrel	Yok	1164	99.0	201	99.5	0.705
	Var	12	1.0	1	0.5	
Aile öyküsü	Yok	952	81.0	172	85.1	0.155
	Var	224	19.0	30	14.9	
Koronar arter by-pass öyküsü	Yok	1156	98.3	200	99.0	0.760
	Var	20	1.7	2	1.0	
Hiperlipidemi	Yok	1143	97.2	192	95.0	0.161
	Var	33	2.8	10	5.0	
Kalp yetmezliği öyküsü	Yok	1117	95.0	187	92.6	0.217
	Var	59	5.0	15	7.4	
Malignite öyküsü	Yok	1170	99.5	201	99.5	0.999
	Var	6	0.5	1	0.5	
Periferik arter öyküsü	Yok	1172	99.7	200	99.0	0.215
	Var	4	0.3	2	1.0	
İnme	Yok	1166	99.1	202	100.0	0.374
	Var	10	0.9	0	0.0	

*: Sütun; APA stiline göre her satırda farklı harfleri taşıyan değerler, istatistiksel olarak birbirinden farklıdır ($p<0.05$).

Tablo 6: Nicel değişkenlerin DM'nin varlığına ve yokluğuna göre tanımlayıcı istatistikleri

Değişkenler	Diabetes Mellitus				p
	Yok (n=1176)		Var (n=202)		
	Ortalama	Standart Sapma	Ortalama	Standart Sapma	
Kan üre azotu	17.91 ^a	11.04	21.52 ^b	12.62	<0.001
Glikoz	129 ^a	57	210 ^b	99	<0.001
Yüksek yoğunluklu lipoprotein	36 ^a	9	35 ^b	8	0.036
Kolesterol	180 ^a	36	182 ^a	43	0.525
Kreatinin	1.05 ^a	2.06	1.58 ^b	4.38	0.094
Düşük yoğunluklu lipoprotein	119.8 ^a	30.3	115.5 ^a	34.2	0.10
Trigliserid	148 ^a	61	155 ^a	64	0.145
Nabız	73 ^a	11	72 ^a	12	0.272
Diastolik kan basıncı	80 ^a	12	80 ^a	13	0.097
Sistolik kan basıncı	132 ^a	19	132 ^a	20	0.902
Vücut kitle indeksi	25.21 ^a	2.22	25.72 ^b	2.97	0.021

APA stiline göre her satırda farklı harfleri taşıyan değerler, istatistiksel olarak birbirinden farklıdır (p<0.05).

Tablo 7: DVM'nin tüm çekirdek fonksiyonlarının çeşitli performans metriklerine göre sınıflandırma performansları (%95 güven aralığı ile)

Çekirdek Fonksiyonları	Doğruluk	ROC eğrisi altında kalan alan (AUC)	Duyarlılık	Özgüllük (Seçicilik)
Gaussian RBF	0.9448 (0.9315 - 0.9563)	0.8139 (0.7805-0,8473)	0.9922 (0.9572 - 0,9998)	0.94 (0.9254 - 0,9525)
Hyperbolic tangent (Sigmoid)	0.8716 (0.8527 - 0,8888)	0.5906 (0.5631 - 0.618)	0.7359 (0.5967 - 0,8474)	0.877 (0.8581 - 0,8942)
Polinomiyal	0.8846 (0.8666 - 0.901)	0.6495 (0.6171 - 0,6819)	0.7529 (0.6475 - 0,8401)	0.8933 (0.8752 - 0,9096)
Bessel	0.8832 (0.865 - 0,8997)	0.6261 (0.5955 - 0,6566)	0.8154 (0.6997 - 0,9008)	0.8865 (0.8681 - 0,9032)
Doğrusal	0.8788 (0.8604 - 0,8956)	0.6256 (0.5948-0,6564)	0.73973 (0.6238 - 0,8355)	0.88569 (0.8681 - 0,9033)
Laplace	0.9804 (0.9716 - 0.987)	0.9332 (0.9096 - 0,9567)	0.9999 (0.9791 - 1.000)	0.9776 (0.9675 - 0,9852)
ANOVA RBF	0.8803 (0.8619 - 0,8969)	0.6182 (0.5883 - 0,6482)	0.7937 (0.6730 - 0,8853)	0.8844 (0.8659 - 0,9012)

TARTIŞMA

Klinik açıdan koroner arter hastalık (KAH) ve trombotik komplikasyonları sanayileşmiş ülkelerde morbidite ve mortalitenin önde gelen nedenlerindedir. KAH oranının önümüzdeki on yılda hızlı bir şekilde artacağı ve yaşlanmaya sebep olacağı tahmin edilmektedir. Ayrıca, önümüzdeki yıllarda obezite, tip 2 DM ve metabolik sendromun prevalansında dünya çapında bir artış beklenmekle beraber, genç insanlar arasında kardiyovasküler risk faktörlerinin de önemli derecede artış göstermesi öngörülmektedir. Özellikle tip 2 DM olan bireylerde kardiyovasküler hastalık riski, DM olmayan bireylere kıyasla 2-4 kat daha fazladır. Tip 2 DM hastalarının %75'inde mortalite sebebi KAH'tır (18). Akut Koroner Sendromu, KAH'nın ilerlemiş bir klinik tablosu biçiminde meydana gelmektedir. Bu nedenle Akut Koroner Sendromu ile ilişkili risk faktör-

lerinin belirlenmesi ve kontrol altına alınması, kardiyovasküler hastalıklarının önlenmesi (birincil koruma) ve tanısı konulmuş hastalıkların tekrarlama durumlarının önlenmesi (ikincil koruma) için oldukça önemlidir.

Bu çalışmada, Tip 2 DM olan Akut Koroner Sendromlu bireylerden oluşan bir hasta örneğinde, KAH gelişmesini etkilediği kabul edilen DM'nin sınıflandırılması için değişik çekirdek fonksiyonları ile geliştirilen DVM modellerinin tahmin performanslarının değerlendirilmesi amaçlanmıştır. Yapılan yayın taraması bulgularına göre, koroner arter hastalığının özel bir türü olan Akut Koroner Sendromlu hastalardan oluşan bireylerde tip 2 diyabet mellitusun varlığı ya da yokluğunu tıbbi bilgi keşfi süreci ile inceleyen bir çalışmaya rastlanmamıştır. Bu bağlamda, ANOVA RBF, bessel, doğrusal, Gaussian RBF, laplace, polinomiyal ve sig-

moid çekirdek fonksiyonları ile geliştirilen DVM modelleri elde edilmiştir. Bu DVM modelleri arasında, DM'yi sınıflamada en yüksek doğruluk değeri Laplace çekirdek fonksiyonu (0.9804) ile hesaplandı. Doğruluk değeri dikkate alındığında, DM'yi sınıflamada en yüksek değerler sırasıyla; Gaussian RBF (0.9448), Polinomial (0.8846), Bessel (0.8832), ANOVA RBF (0.8803), Hyperbolic tangent (Sigmoid) (0.8716) ve doğrusal (0.8788) olarak elde edilmiştir. ROC eğrisi altında kalan alan açısından tahminlenen değişik DVM modelleri değerlendirildiğinde, çekirdek fonksiyonlarına ilişkin en yüksek değerler sırasıyla; Laplace (0.9804), Gaussian RBF (0.8139), Polinomial (0.6495), Bessel (0.6261), doğrusal (0.6256), ANOVA RBF (0.6182) ve Sigmoid (0.5906) olarak bulunmuştur. Performans ölçütleri arasından duyarlılık değerleri büyükten küçüğe sıralandığında; Laplace (0.9999), Gaussian RBF (0.9922), Bessel (0.8154), ANOVA RBF (0.7937), Polinomial (0.7529), doğrusal (0.7397)

ve Sigmoid (0.7359) olarak elde edilmiştir. Ayrıca, oluşturulan modeller özgüllük [seçicilik] değerleri açısından incelendiğinde, en yüksek özgüllük [seçicilik] değerleri sırasıyla; Laplace (0.9776), Gaussian RBF (0.94), Polinomial (0.8933), Bessel (0.8865), ANOVA RBF (0.8844), doğrusal (0.8857) ve Sigmoid (0.877) olarak bulunmuştur.

Sonuç olarak; incelenen değişik çekirdek fonksiyonları ile oluşturulan modeller arasında söz konusu performans ölçütlerinden; doğruluk, duyarlılık, özgüllük (seçicilik) ve ROC eğrisi altında kalan alan değerleri dikkate alındığında, en iyi sınıflama performansı Laplace Destek Vektör Makinesi modelinden elde edilmiştir. İlerleyen çalışmalarda, farklı klinik verilerde değişik çekirdek fonksiyonlu Destek Vektör Makinesi modelleri ile diğer makine öğrenmesi ya da veri madenciliği algoritmalarının kullanılması hastalıkların sınıflandırma başarısını artırabilecektir.

KAYNAKLAR

1. Sezgin AT, Yıldırım A, Müderrisoğlu H. Akut Koroner Sendromlar III: Kararsız angina pectoris/ST yükselmesiz miyokard infarktüsü. *İç Hastalıkları Dergisi* 2003; 10: 121-30.
2. Amsterdam EA, Wenger NK, Brindis RG, et al. 2014 AHA/ACC Guideline for the management of patients with non-ST-elevation acute coronary syndromes: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2014; 130: 2354.
3. Grech ED, Ramsdale DR. Acute coronary syndrome: unstable angina and non-ST segment elevation myocardial infarction. *BMJ* 2003; 326: 1259-61.
4. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract* 2014; 103: 137-49.
5. Minitab I. MINITAB statistical software. Minitab Release 2015; 16.
6. Breunig MM, Kriegel H-P, Ng RT, Sander J, editors. LOF: identifying density-based local outliers. In: *ACM Sigmod Record* 2000; 29: 93-104.
7. Lee J, Kang B, Kang S-H. Integrating independent component analysis and local outlier factor for plant-wide process monitoring. *J Process Control* 2011; 21: 1011-21.
8. Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPN consortium. *J Am Med Inform Assoc* 2013; 20: e341-e8.
9. Eminoğlu U. Matlab GUI kullanılarak rüzgâr türbini sistemleri için tasarım optimizasyonu paketinin geliştirilmesi ELECO 2012 Elektrik-Elektronik ve Bilgisayar Mühendisliği Sempozyumu; 29 Kasım - 01 Aralık 2012; Bursa.
10. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995; 20: 273-97.
11. Team R. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. 2016.
12. Torgo L, Torgo ML. Package "DMwR". Comprehensive R Archive Network. 2013.
13. Kuhn M. Caret package. *J Stat Softw* 2008; 28: 1-26.
14. Zeileis A, Hornik K, Smola A, Karatzoglou A. Kernlab-an S4 package for kernel methods in R. *J Stat Softw* 2004; 11: 1-20.
15. Wickham H. ggplot2: elegant graphics for data analysis. *J Stat Softw* 2010; 35: 65-88.
16. Sing T, Sander O, Beerwinkler N, Lengauer T. ROCr: visualizing classifier performance in R. *Bioinformatics* 2005; 21: 3940-1.
17. Corp I. IBM SPSS Statistics for Windows, version 24.0. IBM Corp Armonk, NY; 2017.
18. Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of hyperglycaemia in type 2 diabetes: a patient-centered approach. Position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia* 2012; 55: 1577-96.