International Journal of Scientific and Technological Research
ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/5-12-12
Vol.5, No.12, 2019

Classification of Variables Affecting Birth Weight by Decision Trees and K-Nearest Neighbor Methods

Sadi Elasan (Corresponding author)
Department of Biostatistics, Faculty of Medicine,
Van Yuzuncu Yil University, Van, Turkey
E-mail: sadielasan@gmail.com
ORCID: 0000-0002-3149-6462

Siddik Keskin
Department of Biostatistics, Faculty of Medicine,
Van Yuzuncu Yil University, Van, Turkey

Abstract
Objective: The aim of this study was to determine the factors affecting the birth weight of infants by using some Decision Trees and K-Nearest Neighbor methods with high accuracy and to evaluate the performance of the algorithms in the classification of low birth weight. **Material and Methods:** The algorithms used for classification can generally be examined under two headings as "unsupervised" and "supervised". "Decision trees" and "k-nearest neighbor" algorithms in supervised data mining; nonparametric methods and has predictive feature. With these algorithms applied for classification purposes, explanatory variables which are most effective on the birth weight of babies have been determined. From decision trees; "CART, CHAID, exhaustive CHAID, QUEST, Random Forest and C4.5" algorithms have been used. In k-nearest neighbor algorithm; "Euclidean" and "Manhattan" distance measurements have been applied. **Results:** The highest estimation rate in terms of sensitivity has been observed in the "CART" algorithm with 88.4%. The highest estimation rate in terms of specificity criterion has been seen 98.2% in the "Random Forest" algorithm. The highest estimation rate in terms of accuracy criterion has been seen 94.5% in the "C4.5" algorithm. The lowest rate in terms of the risk estimate has been observed in the "C4.5" of 5.6%. **Conclusion:** When the results are examined; it can be said that all algorithms work with "good classification, high estimation and low error rate". This study may contribute to early investigations of the birth weight of newborn babies, whether it is low birth weight or not, and thus taking preventive measures.

Keywords: Cross Validation; Supervised Methods; Euclidean Distance; Risk Estimation; Classification

DOI: 10.7176/JSTR/5-12-12

Bu çalışma, Sadi Elasan tarafından Nisan 2019'da sunulmuş doktora tezinden üretilmiştir.

Doğum Ağırlığını Etkileyen Değişkenlerin Karar Ağaçları ve K-En Yakın Komşu Yöntemleriyle Sınıflandırılması

Özet
Amaç: Bu çalışmada, bebeklerin doğum ağırlığına etki eden faktörlerin bazı Karar Ağaçları ve K-En Yakın Komşu yöntemleri kullanılarak yüksek doğrulukla erken belirlenmesi ve düşük doğum ağırlığını sınıflandırmada algoritmaların performanslarının değerlendirilmesi amaçlanmıştır. **Gereç ve Yöntemler:** Veri madenciliğinde, sınıflandırma amacıyla kullanılan algoritmalar genel olarak; "denetimsiz (unsupervised)" ve "denetimli (supervised)" olmak üzere iki başlık altında incelenebilir.

Denetimli veri madenciliğinde "karar ağaçları (decision trees)" ve "k-en yakın komşu (k-nearest neighbor)" algoritmaları; parametrik olmayan yöntemler arasında olup, tahmin edici özelliğe sahiptir. Sınıflandırma amacıyla uygulanan bu algoritmalarla, çalışmadaki bebeklerin doğum ağırlığı üzerine etkili olan açıklayıcı değişkenler belirlenmiştir. Karar ağaçlarından; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5" algoritmaları kullanılmıştır. K-en yakın komşu algoritmasında; "Öklid" ve "Manhattan (City block)" uzaklık ölçüleri kullanılarak uygulama yapılmıştır. **Bulgular:** Sınıflandırma performansları göz önüne alınarak, en iyi tahmin değerini veren algoritmalar belirlenmeye çalışılmıştır. Bu sonuçlara göre; Duyarlık (Sensitivity) ölçütü bakımından en yüksek tahmin oranı %88.4 ile "CART" algoritmasında gözlenmiştir. Özgüllük (Specificity) ölçütü bakımından en yüksek tahmin oranı %98.2 ile "Rastgele Orman" algoritmasında görülmüştür. Genel doğruluk ölçütü bakımından ise en yüksek tahmin oranı %94.5 ile "C4.5" algoritmasında gözlenmiştir. Risk (hata) tahmin ölçütü bakımından en düşük algoritma, %5.6 ile "C4.5" algoritması olmuştur. **Sonuç:** Genel olarak sonuçlar incelendiğinde; tüm algoritmaların "iyi sınıflandırma, yüksek tahmin ve düşük hata oranı" ile çalıştığı söylenebilir. Bu çalışma, yeni doğacak bebeklerin doğum ağırlığının, düşük doğum ağırlığında olup olmayacağına erken karar verme ve böylece koruyucu tedbirlerin alınması açısından araştırmacılara katkı sağlayabilir.

Anahtar kelimeler: Çapraz Geçerlik; Denetimli Yöntemler; Öklid Uzaklığı; Risk Tahmini; Sınıflama

GİRİŞ

Veri madenciliğinde, sınıflandırma amacıyla kullanılan algoritmalar genel olarak; "denetimsiz (unsupervised)" ve "denetimli (supervised)" olmak üzere iki başlık altında incelenebilir. Denetimli veri madenciliğinde "karar ağaçları (decision trees)" ve "k-en yakın komşu (k-nearest neighbor)" algoritmaları; parametrik olmayan yöntemler arasında olup, tahmin edici özelliğe sahiptir. Karar ağaçları ve K-en yakın komşu, büyük veri tabanlarıyla kolayca entegrasyonu, güvenilirliğinin yüksek, maliyetinin düşük, sonuçlarının görsel ve kolay yorumlanabilir olması gibi nedenlerle sınıflama yöntemleri içerisinde sıkça kullanılan algoritmalardandır. Veri madenciliğinde sınıflandırma amacıyla kullanılan bu algoritmalar; veri temizleme, veri dönüştürme ve indirgeme işlemlerini yönetebilmekle birlikte; bölme, durdurma, birleştirme ve budama gibi işlemleri yapabilmektedir.

Bu çalışmada, veri madenciliğinde sınıflandırma amacıyla kullanılan karar ağacı yöntemlerinden; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman, C4.5" ve "K-En Yakın Komşu" yöntemlerinin Kadın Hastalıkları ve Doğum verisi üzerindeki performanslarının incelenmiştir. Bu çalışma ile bebeklerin doğum ağırlığına etki eden faktörlerin yüksek doğrulukla erken belirlenmesi ve düşük doğum ağırlığını sınıflandırmada algoritmaların performanslarının değerlendirilmesi amaçlanmıştır.

Karar Ağaçları Yöntemi: İlk olarak Breiman tarafından önerilen karar ağaçları, parametrik olmayan tahmin edici özelliğe sahiptir. Karar ağaçları, genel bir ifadeyle, kural çıkarma algoritmalarıdır.[1] Karar ağaçlarını oluşturacak değişkenler kategorik veya sürekli özellikte olabilir. Karar ağaçları; cevap değişkeninin sürekli olması durumunda regresyon ağacı (regression tree), kategorik olması durumunda ise sınıflama ağacı (classification tree) olarak adlandırılmaktadır. Bu farklılığa rağmen karar ağaçları, iki durum için de benzer şekilde oluşturulmaktadır. Klasik istatistik yöntemlerde, veriden bir fonksiyon elde edildikten sonra bu fonksiyonun anlaşılabilir bir kural olarak yorumlanması zor iken karar ağaçları oluşturulduktan sonra kök düğümden yaprak düğümlere doğru inilerek, her dal bir kural oluşturacak şekilde fonksiyon yazılabilir.

Karar ağaçlarının oluşturulmasındaki en önemli adım, veri setindeki değişkenlerin sınıflamasını sağlayacak dallanmanın hangi kritere veya hangi değişkene göre yapılacağının belirlenmesidir. Bu aşamada, belirsizliği en yüksek olan değişken belirlenerek ağacın kök düğümünde test için kullanılır. Bunu belirlemeye yönelik geliştirilmiş literatürde farklı yaklaşımlar vardır. Bunlardan en önemli olan yaklaşımlar; Entropiye dayalı olan, "bilgi kazancı (information gain) ve bilgi kazanç oranı"[4,5,6] "Twoing kuralı"[1], "Gini kriteri"[1] ve "Ki-kare olasılık"[7] tablo istatistiğidir.

K-En Yakın Komşu Yöntemi: K-en yakın komşu yöntemi ilk olarak Cover ve Hart tarafından önerilmiş olup, belirlenen veri noktasının yer aldığı sınıfın veya en yakın komşunun, *k*-değerine göre belirlendiği bir sınıflandırma metodudur.[2] Veri madenciliği sınıflandırma yöntemlerinden olan k-en yakın komşu yöntemi, örüntü (model) tanımada kullanılmak amacıyla, parametrik olmayan bir yöntem olarak geliştirilmiştir. K-en yakın komşu yöntemi, gözlemlerin yer alacağı sınıfı ve en yakın komşuyu, *k*-değerine göre belirleyen bir sınıflama yöntemidir. Gözlemler veya nesneler arası uzaklığa dayalı sınıflandırma yapan denetimli veri madenciliği algoritmalarındandır. Örüntü tanıma, yapay zeka, veri madenciliği, istatistik, bilişsel psikoloji, tıp ve biyoinformatik gibi birçok alanda kullanılmaktadır.

K-en yakın komşu algoritması, uzaklık veya yakınlık hesaplaması yardımıyla sınıflandırma yapar. Bu sınıflama algoritmasının temelinde, "örnek uzayında birbirine yakın olan nesneler muhtemelen aynı kategoriye aittir" düşüncesi yer alır. Algoritmanın amacı, bireyleri ya da nesneleri, bu nesnelere ait özelliklerden yararlanarak, önceden belirlenene sınıflara veya gruplara en doğru şekilde atamaktır. Yöntem ayrıca yeni bir gözlemin de sınıflamasını sağlar. Sınıflandırılmak istenen gözlem, öğrenme veri seti yardımıyla, en yakınında bulunan *k* tane gözlemden en fazla benzer olanlarla aynı veri setinde sınıflandırılması yapılır.

GEREÇ VE YÖNTEM

Çalışmada, "doğum ağırlığını etkileyen değişkenlerin karar ağaçları ve k-en yakın komşu yöntemleriyle sınıflandırılması" amacıyla, tanımlayıcı istatistikleri Tablo 1'de verilen 910 kadına ait (Süleymaniye eğitim ve araştırma hastanesi etik inceleme kurulu tarafından onaylanmış)[3] veri seti kullanılmıştır. Bu veri setinden; 4'ü sürekli yapıda olmak üzere toplam 34 adet değişken seçilmiştir. Veri seti, yaşlarına göre 3 gruptan (Kontrol n=301 | x̄=27, Adolesan n=306 | x̄=16, İleri yaş n=303 | x̄=41) oluşmaktadır. Bebeklerin %31.2'sinin doğum ağırlığı "düşük (≤2500g)" kategoride yer alırken, %68.8'inin doğum ağırlığı "normal (>2500g)" kategoride yer almıştır. Çalışmada kullanılan sürekli ve kategorik değişkenler ile bunlara ait tanımlayıcı istatistikler (ortalama, standart sapma, minimum, maksimum, sayı ve yüzde) Tablo 1'de özetlenmiştir.

Tablo 1. Çalışmada ele alınan değişkenler ve tanımlayıcı istatistikler

		n	Ortalama	Std. Sapma	Minimum	Maksimum
Anne Yaşı	Kontrol	301	26.60	4.20	20.0	39.0
	Adolesan	306	15.78	0.89	13.0	17.0
	İleri Yaş	303	40.93	0.95	40.0	44.0
	Genel	910	27.73	10.6	13.0	44.0
Doğum Haftası		910	35.62	3.58	26.0	40.0
Gravida		910	2.69	1.60	1.00	7.00
Parite		910	1.18	1.23	0.00	6.00

		n	%			n	%			n	%
Bebek Doğ. Ağır.	Düşük	284	31.2	Polihidromnios	Yok	873	95.9	SGA	Yok	739	81.2
	Normal	626	68.8		Var	37	4.1		Var	171	18.8
Doğum Şekli	Normal	700	76.9	Oligohidramnio	Yok	824	90.5	Gestasyonel Diyabet	Yok	812	89.2
	Sezaryen	210	23.1		Var	86	9.5		Var	98	10.8
Preeklampsi	Yok	861	94.6	Fetal Ölüm	Yok	868	95.4	PROM	Yok	861	94.6
	Var	49	5.4		Var	42	4.6		Var	49	5.4
Eklampsi	Yok	897	98.6	Erken Doğum	Yok	877	96.4	Koryoamniyonit	Yok	888	97.6
	Var	13	1.4		Var	33	3.6		Var	22	2.4
Anne karında oksijensizlik	Yok	863	94.8	Amniyosentez	Yok	885	97.3	Urinenfe	Yok	801	88.0
	Var	47	5.2		Var	25	2.7		Var	109	12.0
Acil Sezaryen	Yok	851	93.5	ART	Yok	885	97.3	Fetolum	Yok	900	98.9
	Var	59	6.5		Var	25	2.7		Var	10	1.1
Oksitosin Takılma	Yok	898	98.7	Sigara	Yok	816	89.7	Antehemo	Yok	810	89.0
	Var	12	1.3		Var	94	10.3		Var	100	11.0
Prespont	Yok	809	88.9	Anemi	Yok	729	80.1	Posthemo	Yok	835	91.8
	Var	101	11.1		Var	181	19.9		Var	75	8.2
Previa	Yok	889	97.7	Çoklu Gebelik	Yok	866	95.2	Nonvertp	Yok	861	94.6
	Var	21	2.3		Var	44	4.8		Var	49	5.4
Nullipar (ilk doğumu)	Yok	564	62.0	Multiparite (çoklu doğum)	Yok	346	38.0	Gestasyonel Hipertansiyon	Yok	837	92.0
	Var	346	38.0		Var	564	62.0		Var	73	8.0

ART: Yardımcı üreme teknikleri, SGA: Bebeğin anne karnında gelişimi, Fetal Distress: Koryoamniyonit:Uterus enfeksiyonu. Anne karnında oksijensizlik, Prom: Erken su gelme, Nonvertp: Geliş anomalisi, Prespont: Kend. erken doğum, Previa: Plasentanın rahim ağzına yapışması, Eklampsi: Gebede nöbet geçirme, Polihidromnios: Su fazlalığı, Oligohidromnio: Su azlığı, Posthemo: Kanama, Multiple:Çoklu gebelik, Amniyosentez: Sıvı alınması

Çalışmada "Bebek Doğum Ağırlığı (normal>2500g | düşük≤2500g)" cevap değişkeni olarak alınmış ve bu değişken ile açıklayıcı değişkenler arası ilişkileri belirlemek üzere, "karar ağaçları" yöntemlerinden; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5" algoritmaları kullanılmıştır. Yine bu verilere "K-En Yakın Komşu" algoritması da uygulanarak, bu yöntemlerin performansları incelenmiştir. Elde edilen sonuçlara göre "düşük doğum ağırlığını etkileyen değişkenler" belirlenmiştir.

Verilerin analizi için "Python (ver.3.5)", "Weka (ver.3.9) ve "SPSS (ver.25)" istatistik paket programları kullanılmıştır.
Karar ağaçları ve k-en yakın komşu algoritmalarının performansını belirlemede, gerçek ve tahmin değerlerine ait; "Risk Katsayısı, Duyarlık (sensitivity), Özgüllük (specificity), Genel Doğruluk Oranı (accuracy)" ve "MAPE katsayısı" ölçütleri Tablo 2'deki gibi hesaplanmıştır.

Tablo 2. Performans ölçülerinin (tanı testlerinin) hesaplanması

Performans Ölçütü	Açıklama	Hesaplama
Duyarlık (Sensitivity)	Gerçekte "Pozitif" olanlar içinden, "Pozitif" olarak tahmin edilenlerin oranı	GP/(GP+YN)
Özgüllük (Specificity)	Gerçekte "Negatif" olanlar içinden, "Negatif" olarak tahmin edilenlerin oranı	GN/(GN+YP)
G. Doğr. Oranı (Accuracy)	Gerçekte "Pozitif" ve "Negatif" olanların toplam içindeki oranı	(GP+GN)/(GP+YP+YN+GN)
MAPE	Ortalama mutlak yüzde hata (mean absolute percentage error)	| Gerçek-Tahmin | / Gerçek * 100
GP: Gerçek Pozitif, GN: Gerçek Negatif, YP: Yanlış Pozitif, YN: Yanlış Negatif		

Düşük doğum ağırlığı, doğum ağırlığının 2500 gramdan az (≤2500g) olması durumudur. Bu çalışmada, "düşük doğum ağırlığındaki bebekler", tanı testi bakımından pozitifliği göstermektedir. Buna göre; "gerçekte pozitif (düşük doğum ağırlıklı) olanlar içinden, "Pozitif" olarak tahmin edilenlerin oranı "Duyarlığı (sensitivity)" vermektedir. Dolayısıyla elde edilen uygulama sonuçlarına göre "Duyarlığı yüksek" bulunan algoritmaların performansının, klinik bakımdan önemli olacağı düşünülmektedir.

BULGULAR

Uygulamanın ilk aşamasında, "karar ağacı" yöntemlerinden CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5 algoritmalarına ait ön deneme sonuçlarına göre en iyi performansı veren işlem seçenekleri elde edilmiştir (Tablo 3).

Tablo 3. Karar ağacı algoritmalarında işlem seçenekleri

Seçenekler	CART	CHAID	Ayrıntılı CHAID	QUEST	Rastgele Orman	C4.5
Çapraz Geçerlik	10-kat	10-kat	10-kat	10-kat	10-kat	10-kat
Maks. Ağaç Derinliği (Oto. Seç.)	5	3	3	5	5	5
Min. Dal Düğüm Sayısı	25	25	25	25	25	25
Min. Yaprak Düğüm Sayısı	5	5	5	5	10	10
Cevap değişkeni	Bebek Doğum Ağırlığı (Normal/Düşük)					

Karar ağacı analizlerinde minimum hata oranlarına ulaşabilmek amacıyla yapılan ön deneme sonuçlarına göre 10-kat çapraz geçerlik testi tercih edilmiştir. Benzer şekilde, elde edilen ön deneme sonuçlarına göre; CART, CHAID, Ayrıntılı CHAID ve QUEST için "minimum dal düğüm sayısı" 25 ve "minimum yaprak düğüm sayısı 5, olarak belirlenmiştir. Rastgele Orman ve C4.5 için "minimum dal düğüm sayısı" 25 ve "minimum yaprak düğüm sayısı 10, olarak belirlenmiştir. CART, QUEST, Rastgele Orman ve C4.5 algoritmalarında "maksimum ağaç derinliği" (otomatik seçimle) 5 olarak alınmıştır. CHAID ve Ayrıntılı CHAID algoritmalarında ise ağaç derinliği 3 olarak belirlenmiştir (Tablo 3). Karar Ağaçları analizi için "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman ve C4.5"e ait ön deneme ile sınıflandırma performansı sonuçları Tablo 5'te verilmiştir.
"K-en yakın komşu" yönteminin doğruluğunun test edilmesinde minimum hata oranına sahip *k* değerine ulaşılmaya çalışılmıştır. K-en yakın komşu algoritmasına ait ön deneme sonuçlarına göre en iyi performansı veren işlem seçenekleri elde edilmiştir (Tablo 4).

Tablo 4. K-en yakın komşu algoritmasına ait ön deneme (performans) sonuçları

K-kat Çapraz Geçerlik	Komşu Sayısı (KNN)	K seçiminde Yanlış Sınıflandırma (Hata/Risk %'si)	Bölümlere Atama Yüzdesi (%)	Duyarlık (Sensitivity) (%)	Özgüllük (Specificity) (%)	Gen. Doğr. Oranı (Accuracy) (%)
10	1	10.6	32.1	78.6	93.8	89.4
10	2	12.6	28.8	65.5	97.8	87.4
10	3	10.8	27.6	75.6	96.4	89.2
10	4	9.7	28.4	72.8	98.3	90.3
10	**5**	**7.9**	**29.1**	**78.4**	**97.4**	**92.2**
10	6	8.2	29.2	77.4	97.4	90.9
10	7	9.4	28.7	77.1	96.7	90.6
10	8	10.4	29.2	75.6	96.2	89.6
10	9	11.6	28.4	69.0	98.7	88.4
10	10	12.1	28.1	67.8	97.2	87.9

K-en yakın komşu algoritmasının performansını belirlemek üzere yapılan ön deneme, işlem seçenekleri olarak; "10-kat çapraz geçerlilik" ve "1 ile 10 arası k-en yakın komşu sayısı" kullanılarak alternatif sonuçlar elde edilmiştir. Buna göre, bu algoritmanın performansını gösteren ölçütlerden; "*k* seçiminde yanlış sınıflandırma (hata/risk) oranı, duyarlık (sensitivity), özgüllük (specificity) ve genel doğruluk oranı (accuracy)" ölçütleri verilmiştir (Tablo 4). "Tahmin değerlerine ait performans sonuçları bakımından; en düşük "risk/hata oranı" ve optimum "duyarlık, özgüllük, genel doğruluk oranı" göz önüne alındığında, "*k* komşu sayısının 5 olması" en iyi sonucu vermektedir.
K-en yakın komşu yöntemi uygulamasında, gözlemler arası uzaklıkları hesaplamada, veri tipi (kategorik, sürekli ve sıralı) göz önüne alınarak iki farklı uzaklık ölçüsü (Öklid ve Manhattan) kullanılmıştır. Elde edilen bu sonuçlara göre; "Öklid" ve "Manhattan (City Block)" uzaklık ölçüleri ile diğer işlem seçenekleri (Tablo 4) kullanılarak KNN algoritması uygulanmış ve Tablo 5'teki analiz sonuçları elde edilmiştir. Çıkan sonuçlara göre açıklayıcı değişkenlerin cevap değişkeni (bebek doğum ağırlığı) üzerindeki önem sıralaması da verilmiştir.

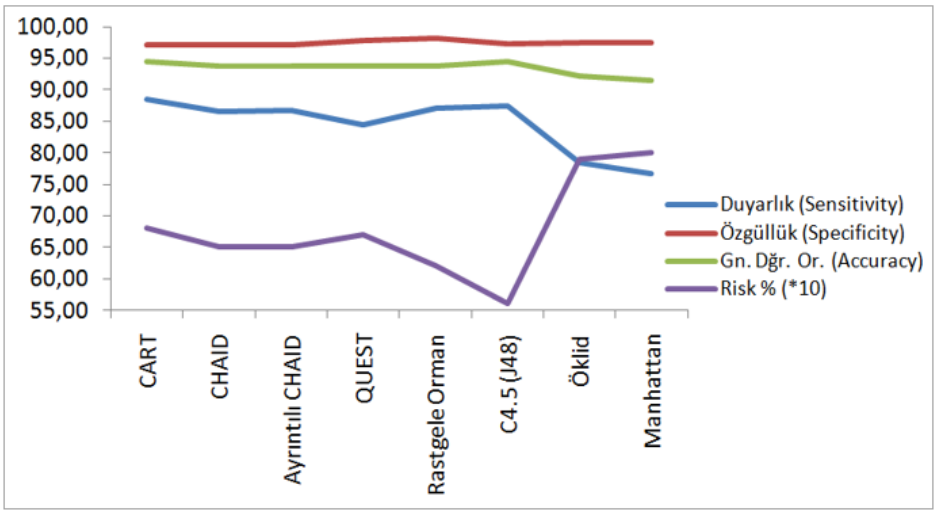
Algoritmaların sınıflandırma performanslarının incelenmesi
Çalışmada kullanılan algoritmaların performansını belirlemede; "Risk Katsayısı, Duyarlık (sensitivity), Özgüllük (specificity), Genel Doğruluk Oranı (accuracy) ve Ortalama Mutlak Hata (MAPE)" ölçütleri kullanılmıştır. Analiz sonucunda, Bebek Doğum Ağırlığı üzerinde etkili olabilecek değişkenler elde edilmiştir (Tablo 5).

Tablo 5. Kullanılan algoritmaların sınıflama performanslarının incelenmesi

		Karar Ağaçları						K-En Yakın Komşu	
Performans Ölçütü		**CART**	**CHAID**	**Ayrıntılı CHAID**	**QUEST**	**Rastgele Orman**	**C4.5**	**Öklid**	**Manhattan**
Duyarlık %		88.4	86.4	86.6	84.5	87.1	87.5	78.4	76.6
Özgüllük %		97.1	97.1	97.0	97.8	98.2	97.3	97.4	97.4
Gen. Doğr. %		94.4	93.6	93.7	93.8	93.8	94.5	92.1	91.5
Risk/Hata (%)		6.8	6.6	6.7	6.7	6.2	5.6	7.9	8.0
MAPE		11.6	13.6	13.4	15.5	16.5	10.9	15.7	16.2
Etkili Değişken	1	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası	Multiparite	Doğ. Haftası	Doğ. Haftası	Doğ. Haftası
	2	SGA	SGA	SGA	SGA	Oligohidro amniyoz	SGA	SGA	SGA
	3	Anne Yaşı	Nullipar	Anne Yaşı	Anne Yaşı	Anne Yaşı	Anne Yaşı	Prespont	Prespont
	4		Anne Yaşı	Nullipar	Nullipar	Pre eklampsi	Prespont	Eklampsi	Gesth
	5		Anemi	Anemi		SGA	Gestasyonel Diyabet	Smoker	Oligohidro amniyoz

MAPE: Mean absolute percentage error (Ortalama mutlak yüzde hata), SGA: bebeğin anne karnında gelişimi, Prespont: kendiliğinden erken doğum, Nullipar: hast. ilk doğ., Multiparite: önceden doğum yapmış, Oligohidroamniyoz: anne karnındaki amniyotik sıvının eksik olması, Preeklampsi: yüksek tansiyon ve organ hasarı

ISSN 2422-8702 (Online), **DOI:** 10.7176/JSTR/5-12-12
Vol.5, No.12, 2019



Şekil 1. Kullanılan algoritmaların sınıflandırma performanslarının karşılaştırılması

Çalışmada ele alınan algoritmaların sınıflandırma performansları Tablo 5'te ve Şekil 1'de gösterilmiştir. Buna göre, kullanılan algoritmaların sınıflandırma performansları genel olarak birbirine yakın bulunmuştur. Ancak bu performans ölçülerine göre en iyi tahmin değerini veren yöntemler aşağıdaki gibi özetlenmiştir.

Duyarlık (Sensitivity): Bu tahmin ölçütü bakımından en yüksek tahmin oranı %88.4 ile CART algoritmasında gözlenmiştir.

Özgüllük (Specificity): Bu tahmin ölçütü bakımından en yüksek tahmin oranı %98.2 ile Rastgele Orman algoritmasında gözlenmiştir.

Genel Doğruluk (Accuracy): Bu tahmin ölçütü bakımından en yüksek tahmin oranının %94.5 ile "C4.5" algoritmasında olduğu tespit edilmiştir.

Risk (hata) tahmini: Bu tahmin ölçütü bakımından en düşük oran %5.6 ile "C4.5" algoritmasında gözlenmiştir.

MAPE: (Ortalama mutlak yüzde hata): Bu tahmin ölçütü bakımından en düşük oran %10.9 ile "C4.5" algoritmasında gözlenmiştir.

Modele giren ve "bebek doğum ağırlığı" cevap değişkeni üzerinde etkili olan açıklayıcı değişkenler incelendiğinde; Rastgele Orman algoritması haricinde (Multiparite), tüm algoritmalarda "doğum haftasının" en etkili değişken olduğu gözlenmiştir. Benzer şekilde, ikinci sırada en etkili açıklayıcı değişkenler algoritmalarda genel olarak "SGA" olurken, Rastgele Orman algoritmasında "Oligohidroamniyoz (anne karnındaki amniyotik sıvının eksik olması)" yer almıştır. Genel anlamda algoritmaların performansları bakımından az değişken ile yüksek performans gösteren algoritmaların tercih edildiği göz önüne alındığında; CART'ın 3 ve CHAID'in 4 adet açıklayıcı değişken ile model oluşturması, diğer değişkenlere (5 adet değişken) göre daha başarılı olduğu söylenebilir.

TARTIŞMA

Tıpta daha çok teşhise karar verme amacıyla kullanılan yöntemlerden olan veri madenciliğinin sağlık sektöründe kullanımı, sağlık hizmetlerinin daha etkin sunumu ve kaynakların daha verimli kullanılması açısından önemlidir. Bu çalışmada; "CART, CHAID, Ayrıntılı CHAID, QUEST, Rastgele Orman, C4.5 karar ağaçları" ve "K-En Yakın Komşu" yöntemlerinin "bebek doğum ağırlığını" etkileyen faktörleri sınıflama ve belirleme amacıyla kullanılmıştır. Bu veri setinden elde edilen uygulama sonuçlarına göre söz konusu yöntemlerin performansları incelenmiştir. Buna göre, her algoritmada bir miktar değişiklik

olmakla birlikte, benzer değişkenler modele girmiştir. Kullanılan algoritmalarda genel olarak ilk sırada "Doğum haftası" açıklayıcı değişkeni yer alırken, Rastgele Orman algoritmasında "Multiparite" değişkeni yer almıştır. Genel olarak, bu çalışmada kullanılan algoritmalara ait analiz sonuçları incelendiğinde; tüm algoritmaların "yüksek tahmin ve düşük hata oranı" ile çalıştığı söylenebilir. Ancak, "C4.5" algoritmasının diğer algoritmalara göre bir miktar daha iyi performansla sınıflandırma sağladığı gözlenmiştir (Tablo 5). Literatürde, birçok alanda veri madenciliği yöntemleri uygulanmış çalışmalar bulunmakta ve kullanılan yöntemlerin sınıflandırma başarıları karşılaştırılmaktadır. Bu yöntemlerin karşılaştırılması akademik çalışmalara ve güncel uygulamalara fayda sağlamaktadır. Tıp alanında, bu algoritmaların sınıflandırmadaki tahmin başarıları ile ilgili çalışmalar sıkça yapılmaktadır. Yapılmış bu çalışmaların genel olarak yüksek performansla sınıflandırma yaptığı, örüntü desenini açıkladığı ve karar verme konusunda yardımcı olduğu görülmektedir.

SONUÇ

Dünya genelinde sağlık hizmetleri, teknolojik değişimden önemli derecede etkilenmektedir. Bu nedenle karar verme süreçlerine yardımcı olabilecek algoritmalar yardımıyla geliştirilecek yapay zeka modelleri geleceğin vazgeçilmez gelişmeleri arasında yer alacaktır. Yapay zeka ile insan sağlığı açısından hızlı teşhis, tedavi planlaması, sonuçların doğruluğunun artması, tıbbi müdahalenin azalması ve kişiye özel tedavi yöntemlerinin belirlenmesi açısından önemlidir. Dünyada bebeklerin %16'sı düşük doğum ağırlığı (≤2500g) ile doğmaktadır. Türkiye'de ise bu oran ortalama %10-12 arasındandır. Giderek artan bu oranlar toplum sağlığı açısından olumsuz sonuçlara yol açmakta ve bebeklerde gelişimsel geriliğe yol açan biyolojik etkenlerden biri olarak görülmektedir.[8] Bebek doğum ağırlığı birçok faktörden etkilenmektedir. Dolayısıyla doğum ağırlığını olumlu ya da olumsuz yönde etkileyen faktörlerin karar vermedeki başarılarının belirlenmesiyle; yeni doğacak bebeklerin doğum ağırlığını belirlenmesi, düşük doğum ağırlığında olup olmayacağına erken karar verilmesi, koruyucu tedbirlerin alınması ve toplum sağlığı açısından önemli olacaktır.

Finansal Kaynak

Bu çalışma sırasında, yapılan araştırma konusu ile ilgili doğrudan bağlantısı bulunan herhangi bir ilaç firmasından, tıbbi alet, gereç ve malzeme sağlayan ve/veya üreten bir firma veya herhangi bir ticari firmadan, çalışmanın değerlendirme sürecinde, çalışma ile ilgili verilecek kararı olumsuz etkileyebilecek maddi ve/veya manevi herhangi bir destek alınmamıştır.

Çıkar Çatışması

Bu çalışma ile ilgili olarak yazarların ve/veya aile bireylerinin çıkar çatışması potansiyeli olabilecek bilimsel ve tıbbi komite üyeliği veya üyeleri ile ilişkisi, danışmanlık, bilirkişilik, herhangi bir firmada çalışma durumu, hissedarlık ve benzer durumları yoktur.

Yazar Katkıları

Fikir/Kavram: *Sadi Elasan;* ***Tasarım:*** *Sadi Elasan;* ***Denetleme/Danışmanlık:*** *Sıddık Keskin;* ***Veri Toplama Ve/Veya İşleme:*** *Orkun Çetin;* ***Analiz Ve/Veya Yorum:*** *Sadi Elasan;* ***Kaynak Taraması:*** *Sadi Elasan;* ***Makalenin Yazımı:*** *Sadi Elasan;* ***Eleştirel İnceleme:*** *Sadi Elasan, Sıddık Keskin;* ***Kaynaklar ve Fon Sağlama:*** *Sadi Elasan;* ***Malzemeler:*** *Sadi Elasan.*

KAYNAKLAR

1. Breiman L, Friedman J, Stone CJ. Olshen RA. Classification and Regression Trees. Taylor and Francis, Chapman&Hall/CRC. 1984.

2. Cover TM ve Hart PE. [Nearest neighbor pattern classification]. IEEE Trans Inf Theory 1967;13(1):21–7.

3. Çetin O, Verit FF, Zebitay AG, Aydın Z, Kurdoğlu Z, Yücel O. [Neither early nor late for becoming pregnant: Comparison of the perinatal outcomes of adolescent. reproductive age and advanced maternal age pregnancies]. Clinical Investigation. Turk J Obstet Gynecol 2015;12(3):151-7.

4. Quinlan JR. [Simplifying decision trees]. Int J Man-Mach Stud 1987;27(3):221-34.

5. Quinlan JR. [Decision trees and decision-making]. IEEE Trans Syst Man Cybern 1990;20(2):339-46.

6. Quinlan JR. C4.5: programs for machine learning. Elsevier; 2014.

7. Mingers J. [An empirical comparison of selection measures for decision-tree induction]. J Mach Learn 1989;3(4):227-43.

8. Sağlık Bakanlığı. Halk Sağlığı Genel Müdürlüğü [internet] 2017. [ET: 27.02.2019] https://hsgm.saglik.gov.tr/tr/ beslenme/gebelik-doneminde-beslenme.html.