





Hipertansiyon Tahmini İçin Temel Bileşen Analizinin Kullanımı

Principal Component Analysis Using For Estimating Hypertension

Halil Murat Ünver¹  Yunus Kökver²  Aydın Çiftçi³ 

¹Kırıkkale Üniversitesi, Mühendislik Fakültesi, Bilgisayar Müh. Bl. Kırıkkale

²Ankara Üniversitesi Elmadag Meslek Yüksekokulu Elmadag Ankara

³Kırıkkale Üniversitesi, Tıp Fakültesi İç Hastalıkları ABD Kırıkkale

Başvuru/Received: 06/02/2020

Kabul / Accepted: 02/12/2020

Çevrimiçi Basım / Published Online: 31/12/2020

Son Versiyon/Final Version: 31/12/2020

Öz

Amaç: Otuz yaş ve üzerindeki 150 hastadan, hipertansiyona etki etmesi muhtemel bilgilerden; cinsiyet, yaş, lipid profili, trigliserid, vücut kütle indeksi, ürik asit ve sigara kullanımı verileri toplanmış ve bir hipertansiyon veritabanı oluşturulmuştur. Bu kişilerden 65'i sağlıklı, geriye kalan 85 kişi ise hipertansiyon hastasıdır. Bu veritabanından hipertansiyon hastalığının Temel Bileşen Analizi kullanılarak tahmin edilmesi amaçlanmıştır.

Gereç ve Yöntem: Naive Bayes, Çok Katmanlı Algılayıcı Ağ (ÇKA), Karar Tablosu ve C4.5 sınıflandırma algoritmaları uygulanmış, ardından Temel Bileşenler Analizi uygulanarak hipertansiyon veritabanının boyutu indirgenmiş ve aynı sınıflandırma algoritmaları tekrar uygulanmış ve sonuçlar karşılaştırılmıştır.

Bulgular: Aynı şartlarda işleme sokulan algoritmalarından en başarılı sonucu %88 doğruluk oranıyla Naive Bayes sınıflandırıcısı vermiştir. Naive Bayes sınıflandırıcısını sırasıyla %85,33 başarı oranıyla Karar Tablosu algoritması, %82,67 başarı oranıyla ÇKA algoritmaları takip etmiştir. Hipertansiyon veritabanına TBA analizi uygulanıp, aynı şartlarda aynı algoritmalar tekrar işleme sokulup, TBA uygulanmayan sonuçlarla kıyaslandığında ise, C4.5 algoritması normalden %4 daha başarılı sonuç vererek en başarılı algoritma olmuştur. C4.5 algoritmasını sırasıyla %2,67 daha başarılı sonuç veren Karar Tablosu algoritması ve %1,33 daha başarılı sonuç veren ÇKA izlemiştir.

Sonuç: Naive Bayes sınıflandırıcı haricindeki tüm algoritmalarda Temel Bileşenler Analizi'nin sınıflandırma başarısını artırdığı görülmüştür.

Anahtar Kelimeler

"Hipertansiyon, temel bileşen analizi, veri madenciliği, karar ağacı"

Abstract

Aim: 150 patients which aged 30 years and over were exposed to possible hypertension; age, gender, lipid profile, body mass index, triglyceride, cigarette use and uric acid data are collected and hypertension database are created. 65 people is healthy, and the remaining one is suffering from hypertension. It is aimed to estimate the hypertension disease from this database using the Principal Component Analysis.

Material and Method: Decision Table, Naive Bayes, C4.5 and Multilayer Perceptron Network(MLP) classification algorithms are applied to this database, then the size of the hypertension database is reduced by applying Principal Component Analysis and the same methods are applied again and the results are compared.

Results: The most successful result of the algorithms that were processed under the same conditions gave Naive Bayes classifier with 88% accuracy. Naive Bayes classifier was followed by the Decision Table algorithm with success rate of 85.33%, and ÇKA algorithms with success rate of 82.67%. If the TBA analysis is applied to the hypertension database and the same algorithms are re-processed under the same conditions and the TBA is compared to the untreated results, the C4.5 algorithm is normally the most successful algorithm with 4% more successful results. The Decision Table algorithm, which yielded C4.5 algorithm with 2.67% more success rate respectively, and ÇKA which has a more successful result than 1.33%.

Conclusion: Algorithms except the Naive Bayes algorithm, improved their classification accuracy rate.

Key Words

"Hypertension, principal component analysis, data mining, decision tree logic,"

1. Giriş

Hipertansiyon rahatsızlığı, günümüzde birçok kişide gözlenen önemli bir sağlık problemidir. Hipertansiyon sinsi bir şekilde herhangi ciddi bir belirti oluşturmadan kişinin başta böbrekleri olmak üzere, kalp ve diğer organlarına ciddi zararlar verebilir. Her yıl hipertansiyon komplikasyonlarına bağlı olarak 9,4 milyon insan ölmektedir. Kalp rahatsızlıklarına bağlı ölümlerin yaklaşık %45'inden, inmeye bağlı ölümlerin ise yaklaşık %51'inden hipertansiyon doğrudan sorumludur. Hipertansiyona maruz kalma sıklığı yaşlanmayla beraber artmaktadır. İlerleyen zamanlarda birçok kişide rutin yapılan ölçümlerde rastlanabilir. Hipertansiyonun organlarda meydana getirdiği hasar ve ortaya çıkan komplikasyonların tedavi maliyetlerinin yüksek olması, hipertansiyonun erken teşhisinin önemini ortaya koymaktadır (*A global brief on hypertension / A global brief on Hyper tension*, 2013) .

Hipertansiyonu olan insanların yaklaşık olarak %95'inde bir neden olmaz, sadece risk faktörlerinden bahsedilebilir. Bu gruba primer hipertansiyon denilmektedir; %5'lik grupta kalan insanlarda ise hipertansiyonu başlatan bir neden vardır. Sekonder hipertansiyon olarak adlandırılan bu grubun büyük kısmını böbrek ve böbrek damar hastalıkları oluşturmaktadır (Carretero & Oparil, 2000).

Hipertansiyon ile ilgili literatürde birçok çalışma mevcuttur. Sınıflandırıcı algoritmalarla yapılan çalışmalar özetlenmiştir. Yapılan çalışmada karar ağacı algoritmalarından CHIAD, C4.5, C5.0 kullanılmış, 127886 kayıt kullanılarak hipertansiyon tahmininde en etkili değişkenlerin vücut kütle indeksi, idrar proteini, kan glikozu, kolesterol değerleri ve yaş olduğu saptanmıştır (Chae, Ho, Cho, Lee, & Ji, 2001).

Diğer bir çalışmada, 15-64 yaş aralığındaki erkeklerin, belirlenen 5 tedavi çeşidinin hangi yaş aralığında daha etkin sonuç verdiği için veri madenciliği algoritmaları kullanılmıştır (Almazyad, Ahamad, Siddiqui, & Almazyad, 2010).

Hipertansiyon hastalığı riskini tahmin etmek amacıyla yapılan çalışmada (Ture, Kurt, Turhan Kurum, & Ozdamar, 2005), sınıflandırma teknikleri performansı karşılaştırılmıştır. Retrospektif olarak yapılan çalışmada 3 karar ağacı, 4 istatistiksel yöntem, 2 yapay sinir ağı kullanılmış ve performansları karşılaştırılmıştır.

Siddiqui ve ark. (Siddiqui & Aljumah, 2014), Suudi Arabistan'da hipertansiyon hastaları üzerinde yapmış oldukları çalışmada veri madenciliği sınıflandırma tabanlı teknikleri kullanmışlar ve hipertansiyonla mücadelede en iyi yöntemin sigarayı bırakmak olduğunu görmüşlerdir. Sigaradan sonra sırasıyla en etkili mücadele yöntemlerinin egzersiz, diyet, kilo düzeni ve ilaç kullanımı olduğu belirlenmiştir.

Türk ve ark.'nın (Türk, Barişçi, Çiftçi, & Ekmekçi, 2015) yapmış oldukları çalışmada, 30 yaş ve üzeri 150 kişiden aldıkları verilerden, hipertansiyon sınıflandırması yapmışlar ve Multi Layer Perceptron ve Jordan Elman Neural Networks algoritmalarının başarısını karşılaştırmışlardır. Hipertansiyon tahmininde Jordan Elman Neural Networks algoritmasının daha başarılı sonuçlar verdiği görülmüştür.

Kökver ve ark.'nın (KÖKVER, BARIŞÇI, ÇİFTÇİ, & EKMEKÇİ, 2014) yaptıkları çalışmada hipertansiyon sınıflandırması için C4.5, Naive Bayes ve Çok Katmanlı Algılayıcı Ağlar kullanılmış ve C4.5 algoritmasının sınıflandırma başarısının daha iyi olduğu görülmüştür.

Çalışmanın kalan kısımları; materyal ve metot, bulgular, tartışma ve sonuç bölümlerinden oluşmaktadır.

2. Gereç ve Yöntem

Sistolik(yüksek) kan basıncının 140 mmHg ve üstü, diastolik (düşük) kan basıncımsa 90 mmHg ve üstü olması durumu hipertansiyon olarak tanımlanır (Alderman, 2004; Chalmers et al., 1999). Hipertansiyon hastalığına doğrudan veya dolaylı olarak birçok etken sebep olabilir. Bu çalışmada hipertansiyona en çok etki edebilecek parametreler 2 uzman hekim tarafından belirlenmiştir. Bu parametreler aşağıda açıklamalarıyla birlikte verilmiştir:

Yaş ve Cinsiyet: Çalışmada 30 yaş ve üzeri Kadın ve Erkeklerden veri alınmıştır.

Body Mass Index (BMI): Bir kişinin boyunun, kilosuna göre normal olup olmadığını gösteren bir kriterdir. Kişinin kilosunun, boyunun karesine bölünmesiyle elde edilir. Bu durum, Eşitlik 1'de gösterilmiştir.

$$BMI = \frac{\text{ağırlık(kg)}}{\text{uzunluk(m)}^2} \quad (1)$$

Sigara Kullanımı: Günlük tükettiği paket sayısı ile, sigara içtiği yıl sayısının çarpımıyla elde edilmiştir.

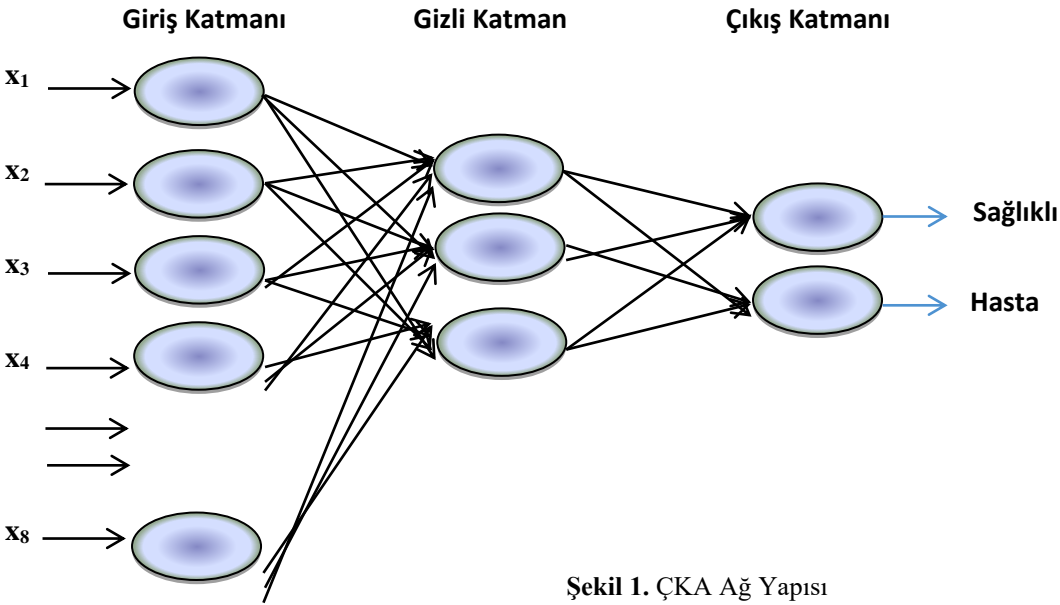
Ürik Asit: Ürik asit yüksekliği, hipertansiyona ve böbrekle bağlantılı damarlarda hasara neden olur.

Lipid Profili: Lipid profile olarak trigliserid, HDL ve LDL değerleri alınmıştır.

2.1. Çok Katmanlı Algılayıcı Ağı (ÇKA)

ÇKA, doğrusal olmayan olayları öğrenme yeteneğine sahip bir ağıdır. Bu ağ modeli özellikle mühendislik problemlerini olmak üzere birçok uygulamada sıkça kullanılmaktadır (Demuth, Beale, Jess, & Hagan, 2014; Sancho-Gómez, García-Laencina, &

Figueiras-Vidal, 2009). Ağın yapısı bir adet giriş katmanı, bir veya daha fazla sayıda ara katman ve bir adet çıkış katmanından oluşmaktadır (Haykin, 2007). Ağın yapısı Şekil 1'de görülmektedir.



Şekil 1. ÇKA Ağ Yapısı

Giriş Katmanı: Bu katmandaki nöronlar, dış dünyadan gelen bilgileri alır. Bu katmandaki her bir nöron, bir sonraki katmandaki tüm nöronlara bağlıdır. Bu nöronlardaki bilgiler, işleme tabi tutulmadan ilk olarak gizli katmana gönderilir (Bishop, 1995).

Gizli Katman: Gizli katmandaki her nöron önceki katman nöronlarının çıktılarını alır ve ağırlıklı toplamını, diğer bir deyişle net bilgi sinyalini oluşturur. Bu sinyal daha sonra bir sonraki katmandaki nöronlar tarafından kullanılmak üzere, bu iletim fonksiyonuna göre o nöronun çıkış sinyaline dönüştürülür.

Çıkış Katmanı: Çıkış katmanındaki her bir nöron, son gizli katmandan gelen bilgi sinyallerinin ağırlıklı toplamı olan bir net bilgi sinyali oluşturur. Bu net bilgi daha sonra bir etkinleştirme işlevinden geçirilir ve ağırlıklı bir çıktı sinyali elde edilir. Dış dünya bilgisi genellikle analog veridir, bu bilginin sayısallaştırılması ve bu sayısal bilginin $[0, 1]$ ölçeğine dönüştürülmesi gerekir.

Gizli katmanlarda ve çıkış katmanındaki nöronların çıktısını hesaplamak için, girdi net fonksiyonundan hesaplanmalıdır, bunun için ağırlık ve bilgi çarpımı kullanılır.

Bu durum Eşitlik 2'de verilmiştir.

$$Net_j = \sum_{i=1}^n W_{ij} F_i \quad (2)$$

Burada Net_j bu katmanın j-inci işleminin net giriş sinyali bilgisini ifade eder. F_i birinci katmandan i-inci nöronun çıkış bilgi sinyalidir ve W_{ij} önceki katmandaki i-inci nöron ile o katmanın j-inci nöronunun ağırlığıdır.

2.2. C4.5 Algoritması

C4.5 algoritması, bilgi entropisi kavramını kullanarak bir eğitim veri setinden karar ağacı üretir. Küçük veya basit karar ağaçlarını keşfetmeye çalışır (Azuaje, 2006). Temel durumlara göre algoritmanın eylemleri aşağıdaki gibidir:

- Listenin tüm örnekleri aynı sınıfta olduğunda, karar ağacı için o sınıfı seçmemizi söyleyen sadece bir yaprak düğüm oluşturur.
- Hiçbir özellik bir bilgi kazancı sağlamazsa, sınıfın beklenen değerini kullanarak ağaçtan daha yukarıya bir karar düğümü oluşturur.
- Daha önce görülmemeyen bir sınıfla karşılaşıldığında, beklenen değeri kullanarak ağaçtan daha yukarıya bir karar düğümü oluşturur (Sehn Körting & Sehn Korting).

2.3. Naive Bayes Algoritması

Veritabanı içerisindeki verilerin, önceden belirlenmiş sınıflara ait olma ihtimallerini tahmin eden bir sınıflandırıcıdır. İstatistikteki Bayes teoremine dayanmaktadır. Bu teorem; belirsizliğe sahip bir durum için model oluşturularak, bu durum ile ilgili gerçekçi gözlemler neticesinde belirli sonuçlar elde edilmesine imkan sağlar. Belirsizlik durumlarında karar verme bakımından çok başarılıdır (J. Wang, 2010).

Bayes Kuralı;

A ve B rastgele sayılar;

$$P(A | B) = P(B | A)P(A) / P(B)$$

P(A) : A olayının müstakil olasılığı

P(B) : B olayının müstakil olasılığı

(3)

$P(A | B)$: B olayının olduğu bilindiği zaman A olayının olma olasılığı
 $P(B | A)$: A olayının olduğu bilindiği zaman B olayının olma olasılığı

2.4. Karar Tablosu Algoritması

Karar tablosu algoritması sınıflandırma yapmak sebebiyle bir karar tablosu oluşturmaktadır. Eğitim setinin özneliklerine göre belirlenen karar tablosunu kullanarak kural tabanlı sınıflandırma yapar (G. Wang, Yu, COMPUTERS, & 2002, n.d.). Karar Tablosu; şartlar setini ve mantık tablo formatında kolayca ifade edilebilir olduğunda onların iş sonuçlarını göstermek için kullanılır. Karar Tablosu aynı zamanda ve farklı şartlar altında farklı işler içeren işlemlerin tutarlılığını ve bütünlüğünü doğrulamak için de kullanılır. Karar Tablosu, çok kompleks ve geniş şart setleri için Karar Ağaçları'ndan daha iyi çalışır.

2.5. Temel Bileşenler Analizi

Temel Bileşenler Analizi (TBA), sınıflandırma, imaj sıkıştırma vb. alanlarda kullanılan, bir değişkenler setinin varyans-kovaryans yapısını, lineer birleşimleri vasıtasıyla açıklayarak, boyutunun indirgenmesi ve yorumlanmasını gerçekleştiren, çok değişkenli istatistiksel bir yöntemdir. TBA, veri seti içerisindeki önemli bilgileri ortaya çıkarmada etkili bir yöntemdir. Verilerdeki genel özellikleri bulup, boyut sayısının azaltılmasını ve bu şekilde verinin sıkıştırılmasını sağlamaktadır (Abdi & Williams, 2010).

TBA'nın çalışma yöntemi 4 adımda belirlenmektedir.
 X vektörünün m ortalama vektörünün hesaplanması:

$$i = 1, 2, \dots, N \text{ olmak üzere, } m = \frac{1}{N} \sum_{i=1}^N X_i \quad (4)$$

Kovaryans matrisinin hesaplanması:

$$E = \frac{1}{N} \sum_{i=1}^N (X_i - m)(X_i - m)^T \quad (5)$$

Burada X, 1'den N'e kadar bir matristir. E beklenti operatör işleci ve T ise transpozunu temsil etmektedir.

Kovaryans matrisinin öz değerlerinin hesaplanması ve genel form biçimli sıralanışı:

$$\Sigma - \lambda I = 0 \quad (6)$$

Burada I özdeşlik matrisini temsil eder ve NxN ölçülerindedir.

Son olarak ise TBA'nın her temel bileşeninin öz vektörlerinin oluşturulması ile yöntem tamamlanmaktadır.

3. Bulgular

Hipertansiyon hastalığına etki eden faktörler, 2 tane alanında uzman hekim ile yapılan ortak bir çalışma ile belirlenmiş ve bu faktörlerin; cinsiyet, yaş, vücut kütle indeksi, lipit profili, trigliserit, ürik asit ve sigara kullanımı olması gerektiği vurgulanmıştır. Veriler alınırken, kişilerin 30 yaş ve üzerinde olmasına, hamile olmamasına ve ilaç tedavisi başlamamış olmasına özen gösterilmiş, 150 kişiden veriler alınmış ve veritabanına sağlıklı ve hasta olarak 2 sınıfta kaydedilmiştir.

Toplanan veriler, Weka veri madenciliği programı kullanılarak sınıflandırılmıştır. Sınıflandırma için kullanılacak verilerin dağılımı Tablo 1'de gösterilmiştir.

Tablo 1. Sınıflandırma için kullanılacak verilerin dağılımı

Sınıf	Sayı
Sağlıklı	65
Hasta	85
Toplam	150

150 adet veriden 85 adedi "Hasta" olarak, kalan 65 veri ise "Sağlıklı" olarak sınıflandırmaya girmiştir. Test seçeneği olarak yüzde ayırma yöntemi uygulanmıştır. Verilerin yarısı eğitim için, yarısı da test verisi olarak kullanılmıştır.

Tablo 2, Tablo 3, Tablo 4 ve Tablo 5'de sırasıyla ÇKA, C4.5, Naive Bayes ve Karar Tablosu sınıflandırıcısının hipertansiyon tahmini için Karmaşıklık Matrisi verilmiştir.

Tablo 2. ÇKA için karmaşıklık matrisi

a	b	Sınıf
38	10	a=Hasta
3	24	b=Sağlıklı
39	9	a= Hasta – TBA ile
3	24	b= Sağlıklı - TBA ile

Tablo 3. C4.5 için karmaşıklık matrisi

a	b	Sınıf
34	14	a= Hasta
4	23	b= Sağlıklı
36	12	a= Hasta – TBA ile
3	24	b= Sağlıklı - TBA ile

Tablo 4. Naive Bayes için karmaşıklık matrisi

a	b	Class
41	7	a= Hasta
2	25	b= Sağlıklı
42	6	a= Hasta – TBA ile
5	22	b= Sağlıklı - TBA ile

Tablo 5. Karar tablosu için karmaşıklık matrisi

a	b	Class
41	7	a= Hasta
4	23	b= Sağlıklı
45	3	a= Hasta – TBA ile
6	21	b= Sağlıklı - TBA ile

Tablo 6, Tablo 7, Tablo 8 ve Tablo 9’da sırasıyla ÇKA, C4.5, Naive Bayes ve Karar Tablosu sınıflandırıcısının detaylı doğruluk (Detailed Accuracy) değerleri verilmiştir.

Tablo 6. ÇKA için Detaylı Doğruluk Tablosu

Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı	Sınıf
%82,67	0,792	0,111	0,927	0,854	0,941	a=Hasta
	0,889	0,208	0,706	0,787	0,941	b=Sağlıklı
	0,827	0,146	0,847	0,83	0,941	Ağırlıklı Ortalama
%84	0,813	0,111	0,929	0,867	0,942	a=Hasta-TBA ile
	0,889	0,188	0,727	0,8	0,942	b=Sağlıklı-TBA ile
	0,84	0,139	0,856	0,843	0,942	Ağırlıklı Ortalama - TBA ile

Tablo 7. C4.5 için detaylı doğruluk tablosu

Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı	Sınıf
%76	0,708	0,148	0,895	0,791	0,78	a=Hasta
	0,852	0,292	0,622	0,719	0,78	b=Sağlıklı
	0,76	0,2	0,796	0,765	0,78	Ağırlıklı Ortalama
%80	0,75	0,111	0,923	0,828	0,819	a=Hasta-TBA ile
	0,889	0,25	0,667	0,762	0,819	b=Sağlıklı-TBA ile
	0,8	0,161	0,831	0,804	0,819	Ağırlıklı Ortalama - TBA ile

Tablo 8 (devam). Naive Bayes için detaylı doğruluk tablosu

Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı	Sınıf
%88	0,854	0,074	0,953	0,901	0,945	a=Hasta
	0,926	0,146	0,781	0,847	0,945	b=Sağlıklı
	0,88	0,1	0,891	0,882	0,945	Ağırlıklı Ortalama

Tablo 8. Naive Bayes için detaylı doğruluk tablosu

Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı	Sınıf
	0,875	0,185	0,894	0,884	0,922	a=Hasta-TBA ile
%85,33	0,815	0,125	0,786	0,8	0,922	b=Sağlıklı-TBA ile
	0,853	0,164	0,855	0,854	0,922	Ağırlıklı Ortalama - TBA ile

Tablo 9. Karar tablosu için detaylı doğruluk tablosu

Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı	Sınıf
	0,854	0,148	0,911	0,882	0,9	a=Hasta
%85,33	0,852	0,146	0,767	0,807	0,9	b=Sağlıklı
	0,853	0,147	0,859	0,855	0,9	Ağırlıklı Ortalama
	0,938	0,222	0,882	0,909	0,9	a=Hasta-TBA ile
%88	0,778	0,063	0,875	0,824	0,9	b=Sağlıklı-TBA ile
	0,88	0,165	0,88	0,878	0,9	Ağırlıklı Ortalama - TBA ile

Sınıflandırma algoritmalarının TBA'lı ve TBA'sız karşılaştırmalı detaylı analizi Tablo 10'da verilmiştir.

Tablo 10. Algoritmaların Detaylı Analizi

Algoritma	Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı
ÇKA	%82,67	0,827	0,146	0,847	0,83	0,941
ÇKA-TBA ile	84 %	0,84	0,139	0,856	0,843	0,942
C4.5	%76	0,76	0,2	0,796	0,765	0,78
C4.5- TBA ile	%80	0,8	0,161	0,831	0,804	0,819
Naive Bayes	%88	0,88	0,1	0,891	0,882	0,945
Naive Bayes-TBA ile	%85,33	0,853	0,164	0,855	0,854	0,922

Tablo 10 (devam). Algoritmaların Detaylı Analizi

Algoritma	Doğruluk	TP Oranı	FP Oranı	Hassasiyet	F-Ölçütü	Roc Alanı
Karar Tablosu	%85,33	0,853	0,147	0,859	0,855	0,9
Karar Tablosu-TBA ile	%88	0,88	0,165	0,88	0,878	0,9

Tablo 10'a göre, aynı şartlarda işleme sokulan algoritmalarından en başarılı sonucu %88 doğruluk oranıyla Naive Bayes sınıflandırıcısı vermiştir. Naive Bayes sınıflandırıcısını sırasıyla %85,33 başarı oranıyla Karar Tablosu algoritması, %82,67 başarı oranıyla ÇKA algoritmaları takip etmiştir. Hipertansiyon sınıflandırmada en başarısız algoritma ise %76 başarı oranıyla C4.5 algoritması olmuştur. Hipertansiyon veritabanına TBA analizi uygulanıp, aynı şartlarda aynı algoritmalar tekrar işleme sokulup, TBA uygulanmayan sonuçlarla kıyaslandığında ise, C4.5 algoritması normalden %4 daha başarılı sonuç vererek en başarılı algoritma olmuştur. C4.5 algoritmasını sırasıyla %2,67 daha başarılı sonuç veren Karar Tablosu algoritması ve %1,33 daha başarılı sonuç veren ÇKA izlemiştir. En başarısız sınıflandırmayı ise normalden %2,67 daha düşük başarı oranıyla Naive Bayes algoritması vermiştir.

4. Tartışma

Kişilerdeki risk faktörleri ve laboratuvar parametrelerini içeren algoritmalar oluşturarak giderek sıklığı artan hipertansiyonun erken tanı ve tedavisi için bir sistem oluşturma amacı ile yapılan çalışmamızda daha önce yapılan çalışmalar ile paralel sonuçlar elde edilmiştir. Tüm dünyada sıklığı giderek artan, önemli oranda morbidite ve mortaliteye neden olan diyabet, kanser, kalp hastalıkları vs. gibi pek çok hastalığın erken tanılarına yönelik çok sayıda çalışma yapılmaktadır. Bu çalışmalardan amaç hastalık ile ilgili daha erken ipuçları bularak erken tanı konulması ve erken tedaviye başlanmasıdır. Hatta en ideal amaç tanı koymadan önceki aşamada bu hastalığın gelişme riskinin hesaplanması ve riski fazla olan bireylerde hastalık gelişimini önlemeye yönelik önlemlerin alınmasıdır (Abd El Baky Mahmoud, Shaaban, & Ali Ramzy, 2018; Li et al., 2018; Tang & Zhou, 2019).

Sınıflandırma bazlı çalışmalar az sayıdadır. Aynı veritabanı kullanılarak yapılan çalışmalardan, Türk ve arkadaşları'nın yapmış olduğu (Türk et al., 2015), 30 yaş üzeri hiç ilaç kullanmayan ve sistemik bilinen hastalığı olmayan kişiler ile sağlıklı kontrollerin karşılaştırıldığı çalışmalarında, katılımcıların cinsiyet, yaş, boy, kiloları, HDL, LDL, trigliserit, ürik asit düzeyleri ölçüldü, sigara kullanımı sorgulandı. Çalışmanın amacı sağlam ve hasta kişilerin verilerini bilgisayar ortamında karşılaştırarak, sağlıklı kişilerde hastalığa yol açabilecek parametreler konusunda risk analizi ve sınıflandırması yapmaktı. Bunların 65'i normal, 85'i hastalardan oluşmaktaydı. Bu veriler, kantitatif analiz yöntemlerini işleyerek sınıflandırıcılara aktarıldı. Her bir hastadan elde edilen verilere Yapay Sinir Ağı (YSA) modelleri uygulandı. Elde edilen sonuçlar normal ya da hasta olarak sınıflandırdı. Çok Katmanlı Algılayıcı (MLP) nöral ağı kullanılarak, hastaların %80,4'ü ve normal bireylerin %81,8'i doğru olarak sınıflandırılmıştır. Jordan Elman nöral ağını kullanarak, hastaların %85,3'ü ve normal bireylerin %87,8'i doğru sınıflandırılmıştır. Hasta ve sağlıklı sınıflandırması ayrı ayrı hesaplanmış ve hasta sınıflandırmasındaki en yüksek doğruluk oranı %85,3 olarak hesaplanırken, sağlıklı sınıflandırmasındaki en yüksek doğruluk oranı %87,8 olarak hesaplanmıştır. Çalışmada veri tabanının hangi oranda eğitim setine ayrıldığından bahsedilmemiştir.

Yine aynı veri tabanını kullanan çalışmalardan, Kökver ve arkadaşları'nın yapmış olduğu çalışmada (KÖKVER et al., 2014), 150 adet veri, veri madenciliği sınıflandırıcılarıyla incelenmiştir. Veriler yine normal veya hasta olacak şekilde iki sınıfta incelenmiştir. Ayrıca sınıflandırma sonuçlarına göre bir karar ağacı oluşturularak, hipertansiyona doğrudan ve dolaylı olarak etki eden faktörler belirlenmiştir. Sınıflandırma algoritmalarından Naive Bayes, C4.5 ve Çok Katmanlı Algılayıcının (MLP) kullanıldığı çalışmada, C4.5 algoritmasının daha başarılı sonuçlar verdiği gözlenmiştir. Hipertansiyon sınıflandırmasındaki en yüksek doğruluk oranı %91,67 olarak hesaplanmış fakat, söz konusu çalışmada veri setinin %76'sı eğitim için, geri kalan %24'ü test verisi olarak kullanılmıştır. Yani 150 adet veriden 36 tanesi ile sistem test edilmiştir. Bizim çalışmamızda ise 75 tane veri ile sistem test edilmiş ve daha fazla veri ile optimum bir sonuca ulaşılmıştır. Yapılan bu çalışmada en yüksek doğruluk oranı %88 olarak hesaplanmıştır. Bu çalışmanın bir diğer avantajı ise, hesaplanan doğruluk oranlarının TBA yöntemi kullanılarak daha da iyileştirilebilmiş olmasıdır.

5. Sonuç

Bu çalışmada, 30 yaş ve üzerindeki 150 kişiden, hipertansiyona etki etmesi muhtemel bilgilerden cinsiyet, yaş, trigliserid, lipid profili, ürik asit, vücut kütle indeksi ve sigara kullanımı verileri toplanmış ve bir hipertansiyon veritabanı oluşturulmuştur. Uygulanan algoritmalarından en başarılı sonucu veren sınıflandırıcı %88 ile Naive Bayes olmuştur. Ardından Temel Bileşenler Analizi (TBA) uygulanarak hipertansiyon veritabanının boyutu indirgenmiş ve aynı sınıflandırma algoritmaları tekrar uygulanmış ve sonuçlar karşılaştırılmıştır. Bu durumda C4.5 algoritması TBA uygulanmadan önceki durumuna göre %4 daha başarılı sonuç vererek en başarılı algoritma olmuştur. Naive Bayes sınıflandırıcı haricindeki tüm algoritmalarda Temel Bileşenler Analizi'nin sınıflandırma başarısını artırdığı görülmüştür.

İleriki yıllarda bu tür çalışmaların artacağını, kişilerin daha hastalanmadan erken tanılarının konulacağını, belki de hastalık riski olan kişilerin önceden risk analizlerinin yapılarak gerekli önlemler ile hiç hastalık gelişmeyebileceği veya bu ihtimalin çok az olacağını öngörmekteyiz.

Maddi Destek ve Çıkar İlişkisi

Bu çalışma 2011/71 numaralı Kırıkkale Üniversitesi Bilimsel Araştırma Projesi olarak desteklenmiştir. Yazarların çıkara dayalı bir ilişkisi yoktur.

Referanslar

- A global brief on hypertension / A global brief on Hyper tension.* (2013). Retrieved from http://apps.who.int/iris/bitstream/10665/79059/1/WHO_DCO_WHD_2013.2_eng.pdf?ua=1
- Abd El Baky Mahmoud, M., Shaaban, M. A. A., & Ali Ramzy, A. (2018). Clinical role of serum Copeptin in acute coronary syndrome. *Egyptian Heart Journal*, 70(3), 155–159. <https://doi.org/10.1016/j.ehj.2018.04.008>
- Abdi, H., & Williams, L. J. (2010, July 1). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/wics.101>
- Alderman, M. H. (2004). JNC 7: Brief summary and critique. In *Clinical and Experimental Hypertension*. <https://doi.org/10.1081/CEH-200032158>
- Almazayad, A. S., Ahamad, M. G., Siddiqui, M. K., & Almazayad, A. S. (2010). Effective hypertensive treatment using data mining in Saudi Arabia. *Journal of Clinical Monitoring and Computing*, 24(6), 391–401. <https://doi.org/10.1007/s10877-010-9260-2>
- Azuaje, F. (2006). Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2nd edition: San Francisco: Morgan Kaufmann Publishers; 2005:560. ISBN 0-12-088407-0, £34.99. *BioMedical Engineering OnLine*. <https://doi.org/10.1186/1475-925X-5-51>
- Bishop, C. (1995). Neural networks for pattern recognition. Retrieved from <https://books.google.com/books?hl=tr&lr=&id=T0S0BgAAQBAJ&oi=fnd&pg=PP1&dq=Bishop+C.+Neural+Networks+for+Pattern+Recognition+1995%3B+Oxford+Univ.+Press,+N+Y.&ots=jN20qL8Dod&sig=2-F-ubwZLqvUINsroDpVxQ1Dt6w>
- Carretero, O. A., & Oparil, S. (2000). Essential Hypertension. *Circulation*, 101(3), 329–335. <https://doi.org/10.1161/01.CIR.101.3.329>
- Chae, Y. M., Ho, S. H., Cho, K. W., Lee, D. H., & Ji, S. H. (2001). Data mining approach to policy analysis in a health insurance domain. *International Journal of Medical Informatics*, 62(2–3), 103–111. [https://doi.org/10.1016/S1386-5056\(01\)00154-X](https://doi.org/10.1016/S1386-5056(01)00154-X)
- Chalmers, J., MacMahon, S., Mancia, G., Whitworth, J., Beilin, L., Hansson, L., ... Clark, T. (1999). 1999 World Health Organization-International Society of Hypertension Guidelines for the management of hypertension. Guidelines sub-committee of the World Health Organization. *Clinical and Experimental Hypertension (New York, N.Y. : 1993)*, 21(5–6), 1009–1060. <https://doi.org/10.3109/10641969909061028>
- Demuth, H., Beale, M., Jess, O. De, & Hagan, M. (2014). Neural network design. Retrieved from <https://dl.acm.org/citation.cfm?id=2721661>
- Haykin, S. (2007). Neural networks: a comprehensive foundation. Retrieved from <https://dl.acm.org/citation.cfm?id=1213811>
- KÖKVER, Y., BARIŞÇI, N., ÇİFTÇİ, A., & EKMEKÇİ, Y. (2014). DETERMINING AFFECTING FACTORS OF HYPERTENSION WITH DATA MINING TECHNIQUES. *Qualitative Studies*, 9(2), 15–25. <https://doi.org/10.12739/NWSA.2014.9.2.1A0352>
- Li, J., Wang, L., Wang, Q., Xin, Z., Liu, Y., & Zhao, Q. (2018). Diagnostic value of carotid artery ultrasound and hypersensitive C-reactive protein in Type 2 diabetes mellitus patients with acute myocardial infarction in Chinese population. *Medicine (United States)*, 97(41). <https://doi.org/10.1097/MD.0000000000012334>
- Sancho-Gómez, J. L., García-Laencina, P. J., & Figueiras-Vidal, A. R. (2009). Combining missing data imputation and pattern classification in a multi-layer perceptron. *Intelligent Automation and Soft Computing*, 15(4), 539–553. <https://doi.org/10.1080/10798587.2009.10643048>
- Sehn Körting, T., & Sehn Korting, T. (n.d.). *C4.5 algorithm and Multivariate Decision Trees GeoDMA-Geographic Data Mining Analyst View project Brazil Data Cube View project C4.5 algorithm and Multivariate Decision Trees*. [researchgate.net](https://www.researchgate.net/publication/267945462). Retrieved from <https://www.researchgate.net/publication/267945462>
- Siddiqui, M. K., & Aljumah, A. A. (2014). Hypertension Interventions using Classification Based Data Mining. *Research Journal of Applied Sciences, Engineering and Technology*, 7(17), 3593–3602. Retrieved from <http://www.mayoclinic.com/health/high->
- Tang, L., & Zhou, X. J. (2019). Diffusion MRI of cancer: From low to high b-values. *Journal of Magnetic Resonance Imaging*,

49(1), 23–40. <https://doi.org/10.1002/jmri.26293>

Ture, M., Kurt, I., Turhan Kurum, A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, 29(3), 583–588. <https://doi.org/10.1016/j.eswa.2005.04.014>

Türk, F., Barişçi, N., Çiftçi, A., & Ekmekçi, Y. (2015). Comparison of Multi Layer Perceptron and Jordan Elman Neural Networks for Diagnosis of Hypertension. *Intelligent Automation and Soft Computing*, 21(1), 123–134. <https://doi.org/10.1080/10798587.2014.959312>

Wang, G., Yu, H., COMPUTERS, D. Y.-C. J. O., & 2002, undefined. (n.d.). Decision table reduction based on conditional information entropy. *Cjc.Ict.Ac.Cn*. Retrieved from <http://cjc.ict.ac.cn/eng/qwjse/view.asp?id=1076>

Wang, J. (2010). *Encyclopedia of Data Warehousing and Mining, Second Edition*. *Encyclopedia of Data Warehousing and Mining, Second Edition*. Information Science Reference. <https://doi.org/10.4018/978-1-60566-010-3>