YÖNETİM BİLİŞİM SİSTEMLERİ DERGİSİ
http://dergipark.ulakbim.gov.tr/ybs/

Yayın Geliş Tarihi: 30.05.2016
Yayına Kabul Tarihi: 15.08.2016
Online Yayın Tarihi: 05.10.2016

Cilt:1, Sayı:3, Yıl:2016, Sayfa 223-233
ISSN: 2148-3752

TÜRKİYE'DEKİ SEÇMEN EĞİLİMLERİNİN C4.5 KARAR AĞACI ALGORİTMASI İLE BELİRLENMESİ

Ali BAYIR

Şebnem ÖZDEMİR

Sevinç GÜLSEÇEN

Özet

Seçimler siyasi eğilimlerin ortaya konulduğu, seçmenin nihai kararının sonuçlandığı ve yönetenlerin oylanarak belirlendiği sistemdir. Bu sistemler bireylerin algıları ve tercihleri doğrultusunda çıktılar üretmektedir. Bu çıktılar veri madenciliği yöntem ve teknikleri sayesinde gerçek sonuca belirli değerlerle yakınsayacak şekilde önceden tahmin edilebilmektedirler. Bu çalışmanın amacı; verilerin fazla olduğu alanlardan biri olan siyasi seçimlerdeki seçmen eğilimlerinin veri madenciliği yöntem ve teknikleri kullanılarak tespit edilmesi ve bilgi keşfi sürecini açığa çıkarmaktır. Çalışmanın uygulama kısmında 2011 yılı Türkiye milletvekili genel seçimlerinden önce KONDA Araştırma ve Danışmanlık Firması tarafından hazırlanan seçim sonuçlarının tahminine yönelik anket ve kamuoyu araştırma sonuçları veri seti olarak kullanılmıştır. Sınıflandırmaya dayalı veri madenciliği teknikleri ile bu veri seti analiz edilmiş ve böylece seçmenlerin oy verecekleri partileri seçmelerini etkileyen kuralların neler olduğu ortaya çıkarılmıştır. Veri seti niteliklerinin kategorik yapıda olması nedeniyle karar ağacı algoritmalarından, C4.5 sınıflandırma algoritması kullanılarak model oluşturulmuştur. Algoritma uygulamalarında yazılım platformlarından R programlama dili ve Rstudio ortamından faydalanılmıştır. Kurgulanan modelin performans değerlendirme karşılaştırmaları yapılmıştır. Sonuç olarak karar ağacı algoritmasının siyasi eğilimleri belirlemede önemli bir yardımcı araç olduğu, partilerin bu aracı kullanarak seçim propagandalarını yurtdışı örneklerinde olduğu gibi yönlendirebilecekleri görüşüne varılmıştır.

Anahtar Kelimeler: Veri Madenciliği, C4.5. Karar Ağacı Algoritmaları, Siyasi Seçim, Seçmen Eğilimi.

DETERMINATION OF VOTING TENDENCIES IN TURKEY THROUGH C4.5 DESICION TREE ALGORITHM

Abstract

Political elections are systems containing subjects such as political tendencies put forward, final decision of voters concluded and political leaders appointed. This system produces output according to the perceptions and preferences of individuals. These outputs can be predicted by data mining methods and techniques to be converged to the actual results with specific values. The main objectives of this study are defining voting tendencies in politics which is an example of subjects containing huge data through data mining methods and revealing the process of information exploration. On the application part of this disquisition, survey results which was prepared before 2011 elections of Turkey by KONDA Research and Consultancy company are presented as data set and this data set is analyzed via various data mining methods and the rules of behaviors that define voting tendencies are revealed. Because the data set is categorical, model is produced using data mining method such as C4.5 decision tree algorithms. In application, R programming language was chosen for versatility and package support about statistical programming and produced models were compared with R in terms of performance evaluation.

Keywords: Data Mining, C4.5 Decision Tree Algorithms, Political Elections, Voting Tendencies.

GİRİŞ

Dünyada kişi başına düşen bilgileri hesaplamak için yapılan işlemler sonucunda her 14 ayda bir bilgi kapasitesinin iki katına çıkmakta, verilerin genel olarak üretildiği bilgisayarların sayısı ise 18 ayda bir iki katına yükselmektedir. 1986 yılından 2007 yılına kadar geçen süreçte dünyadaki veri boyutu her 40 ayda bir tahminen iki katına çıkmıştır (Hilbert, 2011). Örneğin tek bir jet uçuşunda 30 dakikada bir 10 Terabyte (1TB = 1012bayt) veri üretilebilmekte ve günde 25000'den fazla havayolu uçuşunun evrene kattığı veri hacminin değeri Petabyte (1 PB = 1024 TB)'lara ulaşmaktadır (Dijcks, 2013). Dijital evrenin 2010 yılının başından 2020 yılının sonuna kadar 50 kat büyüme gerçekleşeceği öngörülmekte, 2020 yılı sonuna kadar oluşacak veri oranı 40000 Eksabayt (1EB = 1024 PB) üzerinde olması beklenmektedir (Gantz ve Reinsel, 2012). Verideki bu hızlı artış; aşırı bilgi artışı, büyük veri gibi pek çok yeni kavramın doğmasına neden olmuştur. Bu yeni kavramların yanı sıra işlenmiş verinin karara/kurallara dönüştürülmesi, veriden bilgi eldesi pek çok alanda önemli bir rekabet avantajı sağlamaktadır (Hegland, 2003). Veri madenciliği yöntem ve teknikleri bu avantajın elde edilmesinde önemli ve popüleritesi giderek artan araçlar olarak kabul edilmektedirler. Öyle ki finans, sigorta, pazarlama, biyomedikal tedavi, internet müşteri analizi, tıp, müşteri ilişkileri, biyoloji ve astronomi gibi birçok alanda veri madenciliğinden faydalanılmaktadır (Özekes, 2003; Shi vd., 2015).

Veri hacminin giderek arttığı, değerli bilgi barındıran bir diğer araştırma alanı siyasettir. Tüm dünyada giderek daha fazla siyasetçi kendi siyasi kampanya kararlarını kolaylaştırmak için büyük veri yığınlarının analizleri ile elde ettikleri bilgilerin zenginliğinden yararlanmaktadırlar. Bu sayede siyasi hedeflerinde başarı elde edebilmeleri, nasıl bir yol izleyecekleri konusunda daha belirgin adımlar atabilmeleri mümkün olabilmektedir. Özellikle seçim kampanyalarında seçmenlere yönelik bilgilerin özetlenmesi, suç unsurlarının azaltılması amacıyla devlet kanunlarının düzenlenmesine yönelik kampanyalar yürütülmesi, siyasal kampanya düzenleyenlerin seçmenlerin eğilimlerine göre politika oluşturmaları ve strateji belirlemeleri, siyasal partilerin seçmenlerle ve onların oy verme eğilimleri ile ilgili bilgi toplamaları, başkanlık sistemi kampanyalarında seçmenlere hangi tür mesaj veya hareketlerin ikna edici olabileceğine yönelik çeşitli algoritmalar geliştirilmesinde veri madenciliği yöntem ve tekniklerinden faydalanılmaktadır. Örneğin 2012 Amerika Başkanlık seçimlerinde bölge bölge oy sonuçlarını tahmin etmeye yönelik özel bir veri madenciliği modeli kullanılmış, elde edilen model ile gerçek seçim sonuçlarına 0. 5'den daha haz hata payı ile yakınsayan tahminler yapılmıştır (Larose ve Larose, 2015).

Seçmen eğilimlerinin ortaya çıkarılması hedeflenen bu çalışmada "veri madenciliği teknikleri" kullanılarak Türkiye'de milletvekili genel seçimlerinden önce yapılan seçim sonuçlarının tahminine yönelik anket ve kamuoyu araştırmasına dair elde edilen veri kümesinin analiz edilmesi ve böylece seçmenlerin oy verecekleri partileri seçmelerini etkileyen kuralların neler olduğunun ortaya konulması amaçlanmaktadır.

YÖNTEM

Araştırmada KONDA Araştırma ve Danışmanlık şirketi tarafından 2011 yılı Türkiye Milletvekili Genel Seçimlerinden önce 3584 kişinin katılımıyla gerçekleştirdikleri kamuoyu yoklaması ile elde edilen veriler kullanmıştır. Verilerin analizi için veri madenciliği yöntem ve tekniklerinden C4.5 karar ağacı algoritması seçilmiştir. Algoritmasının uygulanmasında analiz ortamı olarak R dili ve RStudio ortamından faydalanılmıştır. Verilerden bilgi keşfinin belli bir standart çerçevesinde gerçekleştirilebilmesi için Kantardzic (2011) tarafından hazırlanmış olan veri madenciliği süreç adımları izlenmiştir:

- Problemin belirlenmesi ve hipotezin formüle edilmesi
- Ham verilerin toplanması
- Verilerin ön işlemden geçmesi
- Modelin tahmin edilmesi

- Modelin yorumlanması ve sonuç çıktısı oluşturulması

Problemin Tespiti ve Hipotezin Kurulması

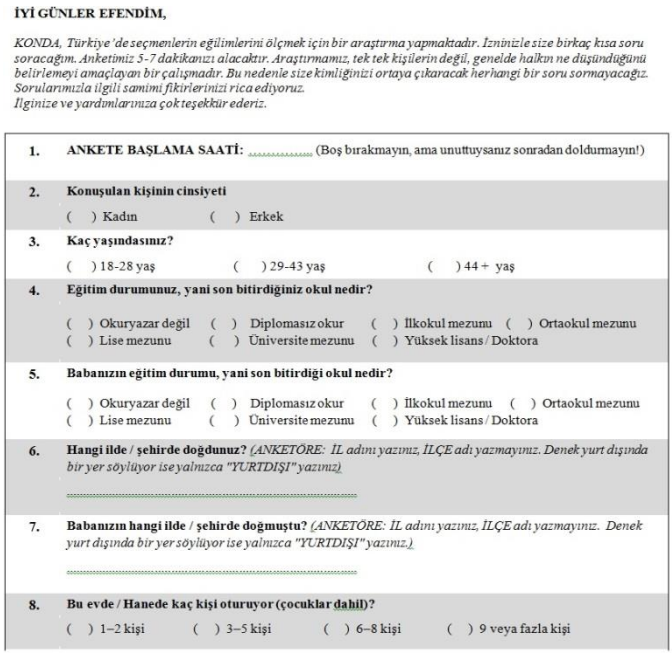
Seçmenlerin seçimlerden önce hangi siyasal partiye oy vermeye eğilimli oldukları ve oy verme davranışını hangi yönde gerçekleştirdikleri seçmen eğilimi olarak tanımlanabilmektedir. Seçmenler; karar verme sürecinde cinsiyet, yaş, meslek, eğitim düzeyi, inanç, ailevi bağ, ekonomik yararlanma beklentisi, ideoloji ile parti ve lidere bağlılık, kişileri takip etme veya izleme eğilimindedirler (Doğan ve Göker, 2010).

Türkiye'de seçimlerden önce kamuoyu araştırmaları yapan şirketler tarafından düzenlenen ve uygulanan seçim anket çalışmalarının içeriği, seçmen eğilimlerinin belirlemeye yöneliktir. Seçmen anketlerinin doğru ve etkin olarak kullanılması ile ortaya çıkacak verilerin daha doğru bir şekilde değerlendirilmesi ve bu bağlamda seçmen eğilimlerinin daha net ortaya çıkması sağlanmış olacaktır. Bu çalışmanın önemi, söz konusu anketlerin seçmen tercihlerinin sayısal olarak ortaya konulması, bu tercihleri etkileyen niteliklerin analiz edilmesi ve karar kuralları biçiminde ortaya konulması ve böylece seçmen tercihlerinin daha derinlemesine çözümlenmesidir.

Verileri Tanımlama ve Anlama

KONDA Araştırma ve Danışmanlık şirketi tarafından 2011 yılı Türkiye Milletvekili Genel Seçimlerinden önce kamuoyu yoklamasında, seçmenlerin eğilimlerini ölçmek için 29 soruluk anket çalışması gerçekleştirilmiştir. Kamuoyu yoklamasında kullanılan anket formunun bir kısmının ekran görüntüsü Şekil 1'de verilmektedir.

Şekil 1. Sorulan sorulara ait anket formunun bir kısmının ekran görüntüsü



İYİ GÜNLER EFENDİM,

KONDA, Türkiye'de seçmenlerin eğilimlerini ölçmek için bir araştırma yapmaktadır. İzninizle size birkaç kısa soru soracağım. Anketimiz 5-7 dakikanızı alacaktır. Araştırmamız, tek tek kişilerin değil, genelde halkın ne düşündüğünü belirlemeyi amaçlayan bir çalışmadır. Bu nedenle size kimliğinizi ortaya çıkaracak herhangi bir soru sormayacağız. Sorularımızla ilgili samimi fikirlerinizi rica ediyoruz.
İlginize ve yardımlarınıza çok teşekkür ederiz.

1. **ANKETE BAŞLAMA SAATİ:** ..... (Boş bırakmayın, ama unuttuysanız sonradan doldurmayın!)

2. **Konuşulan kişinin cinsiyeti**
( ) Kadın ( ) Erkek

3. **Kaç yaşındasınız?**
( ) 18-28 yaş ( ) 29-43 yaş ( ) 44 + yaş

4. **Eğitim durumunuz, yani son bitirdiğiniz okul nedir?**
( ) Okuryazar değil ( ) Diplomasız okur ( ) İlkokul mezunu ( ) Ortaokul mezunu
( ) Lise mezunu ( ) Üniversite mezunu ( ) Yüksek lisans / Doktora

5. **Babanızın eğitim durumu, yani son bitirdiği okul nedir?**
( ) Okuryazar değil ( ) Diplomasız okur ( ) İlkokul mezunu ( ) Ortaokul mezunu
( ) Lise mezunu ( ) Üniversite mezunu ( ) Yüksek lisans / Doktora

6. **Hangi ilde / şehirde doğdunuz?** *(ANKETÖRE: İL adını yazınız, İLÇE adı yazmayınız. Denek yurt dışında bir yer söylüyor ise yalnızca "YURTDIŞI" yazınız)*
.....

7. **Babanızın hangi ilde / şehirde doğmuştu?** *(ANKETÖRE: İL adını yazınız, İLÇE adı yazmayınız. Denek yurt dışında bir yer söylüyor ise yalnızca "YURTDIŞI" yazınız.)*
.....

8. **Bu evde / Hanede kaç kişi oturuyor (çocuklar dahil)?**
( ) 1–2 kişi ( ) 3–5 kişi ( ) 6–8 kişi ( ) 9 veya fazla kişi

Anket çalışmasına farklı illerden toplam 3584 denek katılmıştır. Uygulanan anket sonuçları Excel ortamında düzenlenerek birleştirilmiş ve işleme hazır hale getirilmiştir. Sınıflandırılmış verilere ait 36 nitelik bulunmaktadır. Genel olarak anket sorularından elde edilen bu niteliklerin açıklamaları Tablo1'de verilmektedir.

Tablo 1. Anket sorularından elde edilen bu niteliklerin açıklamaları

Sütun/ değişken	Soru metni
id	Formu numarası
DB	Anketin yapıldığı mahalle/köyün veritabanı kodu
soru_02	Cinsiyet
soru_03	Yaş
soru_04	Eğitim durumu
soru_05	Baba eğitim durumu
soru_06.1	Doğum yeri
soru_07.1	Baba doğum yeri
soru_08	Bu evde / hanede kaç kişi oturuyor?
soru_09	Çalışma durumu
soru_10	Asla oy vermem dediği parti
soru_11	Türkiye'nin sorunlarını hangi parti çözebilir?
soru_12	Bugünlerde seçim yapılsa sizce kim kazanır?
soru_13	Türkiye'yi kim yönetsin?
soru_13.2	Kim yönetsin (gruplanmış)
soru_14	Genel olarak oy tercihinizi sayacağım sebeplerden hangisi etkiliyor?
soru_15	22 Temmuz 2007 seçimlerinde kime, hangi partiye oy verdiniz?
soru_16	Seçim yapmak durumunda olsaydınız, aşağıdakilerden hangisinin en önemlisi olduğunu söylerdiniz?
soru_17	Böyle değişiklikler çoğu sinir bozucu olacaktır.
soru_18	Toplumsal barış ve huzur için böylesi değişiklikleri desteklerim.
soru_19	Böylesi değişikliklerin hayatıma olumlu katkısı olacağına inanıyorum.
soru_20	Bugün seçim olsa hangi partiye oy verirsiniz?
soru_20.1	Bugün seçim olsa (gruplanmış)
soru_21	Kendinizi siyasi görüş olarak nerede tanımlarsınız?
soru_22	Örtünme durumu
soru_23	Etnik köken
soru_24	En çok izlediğiniz televizyon kanalı
soru_25	Din / Mezhep
soru_26	Dindarlık
soru_27	Son bir ay içinde geçinebilmek için bu kurumların herhangi birinden bir yardım aldınız mı?
soru_28	Aylık hane geliri
soru_29	Oturulan evin tipi
bölge	Anketin yapıldığı bölge
kırkent	Yerleşim kodu
soru_06	06 Doğum yeri (il)
soru_07	07 Baba doğum yeri (il)

Verilerin Ön İşlemden Geçirilmesi

Her veri seti analizinde ilk görev, verilerin toplanması, verilerin tanımlanması, temizlenmesidir. Bu işlemden sonra veriler; analiz edilebilir ve sonuç elde edilebilir hale gelmektedir (Dasu ve Johnson, 2003). Veri madenciliğinde güvenilirliğin artırılması için veri ön işleme yapılmak zorundadır (Oğuzlar, 2003).

Veri setindeki verilerin fazla ve detaylı olması sebebiyle önişlemden geçirilmiştir. Ön işleme tabi tutulan verilerin veri madenciliği teknikleri ile analiz edilmesi amaçlanmış, bu kapsamda veri setinde bulunan niteliklere ait değerler oluşturulmuştur. Veri seti değişkenlerin incelemesinde, 36 nitelikten, "id", "DB", "bölge", "kırkent" nitelikleri veri analizinde kullanılmayacağı için ilk etapta çıkartılmıştır. Ankette bulunan altıncı ve yedinci sorulara verilen cevaplar "soru_06" ve "soru_07" niteliği şeklinde anket hazırlayıcısı tarafından eklenmiş, daha sonra verilen cevapları bölgelere göre düzenlemek için "soru_06.1" ve "soru_07.1" niteliğini eklemiştir. Veri setindeki "soru_13" ve "soru_20" nitelikleri anketin hazırlayıcıları tarafından gruplanmış olarak "soru_13.2" ve "soru_20.1" sütunları eklenmiştir. Uygulamada aynı değişkenlerin iki nitelikte de kullanılmaması için "soru_13" ve "soru_20" sütunlarda çıkartılmış, veri seti 28 niteliğe indirgenmiştir.

Hazırlanan veri setinde değişkenlerden bazıları "Cevap Yok" olarak belirtilmiştir. Anket çalışmasında sorulan sorulara bireylerin görüş bildirmemesi veya tercihini gizlemesi sebebiyle o hücre boş olarak geçmektedir. Genel olarak bir veri setinde belirtilmeyen boş hücreler, "kayıp veri" (missing data) olarak tanımlanmaktadır. Bu doğrultuda genellikle bir analizde yer alan değişkenlerin herhangi birinde eksik bir değer var ise bu uygun olmayan analizlere neden olmaktadır (Little ve Rubin 1987). "Cevap Yok" hanelerinin veri setinden çıkartılması karar yapılarınızı etkileyecektir. Bu nedenle değişkenin, kodlamalarda yapılan nümerik atamalarının kullanılması ve veri analiz teknikleri sonuçlarına göre yorumlamaların bu faktör göz önünde bulundurularak yapılmasına karar verilmiştir.

Veri seti analizinin özünde seçmen eğilimlerinin sonucunu gösteren parti seçimine dair nitelik bağlayıcılık taşımaktadır. Bu nedenle veri setinde hedef nitelik olarak "Bugün seçim olsa hangi partiye oy verirsiniz" sorusuna verilen cevapların gruplandırıldığı "soru_20.1" niteliği seçilmiştir. Veri setinde yapılan incelemelerde bazı niteliklerin anlamları ve kodlamaları arasında benzerlikler olduğu ve denekler tarafından verilen cevap karşılaştırmalarında benzerlik olabileceği değerlendirilmiş ve değişkenler arasındaki korelasyon değerleri analiz edilmiştir. Bu analiz için IBM SPSS Statistics 23.0 programı kullanılmıştır. Hedef nitelik ile bağımsız nitelikler arasında alınan korelasyon oranları Tablo 2'de verilmektedir.

Tablo 2. Hedef nitelik ve bağımsız nitelikler arasındaki korelasyon

Değişkenler Arasındaki Korelasyonlar			
	soru_20.1		**soru_20.1**
soru_02	,026	**soru_16**	,077
soru_03	-,003	**soru_17**	,378
soru_04	,198	**soru_18**	,336
soru_05	,167	**soru_19**	,398
soru_06.1	,044	**soru_20.1**	1,000
soru_07.1	,054	**soru_21**	,506
soru_08	,188	**soru_22**	,415
soru_09	,099	**soru_23**	,243
soru_10	,520	**soru_24**	,348
soru_11	,789	**soru_25**	,296
soru_12	,549	**soru_26**	,326
soru_13.2	,891	**soru_27**	,080
soru_14	,236	**soru_28**	,128
soru_15	,749	**soru_29**	,070

BULGULAR

C4.5 algoritması uygulanmasında, veri setinin eğitim olarak ayrılan kısmı tüm veri setinin %60-%70-%80-%90 olarak ayrılması ile elde edilen sonuçlar ortaya konmuştur. Ortaya çıkan tahminler test verisi olarak ayrılan %40-%30-%20-%10 bölümler üzerinde uygulanmış ve sonuçlar elde edilmiştir. Veri setinin belirlenen oranlarda uygulanması ile elde edilen doğruluk ve yanlış sınıflandırma değerlerine ait karşılaştırma tablosu Tablo 3'de verilmiştir.

Tablo 3. C4.5 Doğruluk ve Yanlış Sınıflandırma Değerleri

	Genel Doğruluk Değerleri	Yanlış Sınıflandırma Değerleri
%60-%40	0.694619147	0.305380853
%70-%30	0.691519105	0.308480895
%80-%20	0.714285714	0.285714286
%90-%10	0.72752809	0.27247191

Veri setinde en uygun yanlış sınıflandırma oranı %27,2 ile % 90 Eğitim seti - %10 Test seti olarak ayrılan bölümden elde edildiği görülmektedir. Bu ayrışma ile oluşan karar ağacı yapısının "M" değerine ve budanma olması/olmaması durumuna göre hata oran değişimi Tablo 4'de görülmektedir.

Tablo 4. C4.5 Algoritmasının Hata Oranı Değişimi

M (minimum yaprak sayısı)	U (budanmamış)	Yaprak Sayısı	Ağaç Boyutu	Hata
3	EVET	1178	1403	32.00%
3	HAYIR	314	375	27.80%
5	EVET	662	787	32.00%
5	HAYIR	189	225	28.90%
8	EVET	443	527	31.40%
8	HAYIR	114	136	28.30%
10	EVET	363	432	31.40%
10	HAYIR	91	108	28.60%
20	EVET	180	214	29.40%
20	HAYIR	65	77	27.20%
30	EVET	121	144	28.00%
30	HAYIR	49	58	27.20%
40	EVET	92	109	27.80%
40	HAYIR	45	53	27.50%
50	EVET	74	87	28.30%
50	HAYIR	35	41	28.60%
100	EVET	37	44	32.80%
100	HAYIR	19	22	32.00%

Eğitim örneklerinden oluşan bir dizinin saflığını ölçmek amacıyla verilen bir entropi, eğitim verisini sınıflandırmadaki bir niteliğin etkinliğinin ölçümü olarak tanımlanabilmektedir. Bilgi kazancı olarak adlandırılan bu ölçüm basitçe belirlenen niteliğe göre örneklerin bölümlenmesiyle oluşan entropide ortaya çıkması beklenen azalmadır (Mitchell, 1997). Veri madenciliğinde ağaç yapısı niteliklerin önem derecesini belirten bilgi kazancı ile oluşturulur. Bilgi kazancı yüksek olan yani önem derecesi yüksek olan nitelikler ağacın kök veya köke yakın düğümlerini oluşturmaktadır. Veri setinin tüm niteliklerine ait bilgi kazanç değerleri Tablo 5'de verilmektedir.

Tablo 5. Niteliklerin Bilgi Kazanç Değerleri

Nitelik	**Bilgi Kazancı (Gain)**	**Nitelik**	**Bilgi Kazancı (Gain)**
Kim kazanır	0.2947	Ev nüfusu	0.0458
Siyasi görüş	0.2676	Eğitim durumu	0.0384
Oy vermeyeceği parti	0.2489	Babanın eğitim durumu	0.0339
Etnik köken	0.1641	İş durumu	0.0246
Oy tercihini etkileyen sebepler	0.1095	Gelir	0.0236
Tercih edilen TV kanalı	0.1061	Ülke düzeni-söz hakkı-insancıl toplum tercihi	0.0233
Anayasa değişimini olumlu bulmak	0.1057	Oturduğu ev tipi	0.0130
Örtünme durumu	0.1011	Geçim yardımı alma durumu	0.0099
Anayasa değişimini olumsuz bulmak	0.0959	Yaş	0.0080
Anayasa değişimine destek durumu	0.0779	Babanın doğum yeri	0.0072
Dindarlık durumu	0.0664	Doğum yeri	0.0064
Dini mezhep	0.0545	Cinsiyet	0.0062

Verilen bir eğitim veri setiyle en iyi çalışmayı yapacak kararı üretmek, sıklıkla örnek gürültüsü çok hassas, büyük bir ağaç oluşturur. Bu şekildeki karar ağaçları, test edilmemiş örnekler ile iyi bir performans göstermez. Tahminin hata oranını azaltmak için bu ağaçlara budama yöntemi uygulanması gerekmektedir. Budama, sınıflandırma için ağaçtaki çok az katkı sağlayan kısımların çıkarılmasıdır. C4.5, bir alt ağacın kısımlarının tahmini hata oranlarını

toplar ve bu alt ağacın bir yaprakla değiştirilmesi durumundaki hata oranıyla bunu karşılaştırır. Eğer tek yaprak olması durumunda alt ağaçtan daha yüksek hata oranına sahip değilse bu alt ağaç budanır (Hssina vd., 2014). C4.5 algoritmasından elde edilen bazı karar kuralları aşağıda sıralanmıştır (Tablo 6).

Kural 1

EĞER "Bugünlerde seçim olsa kim kazanır" **AKP İSE** ve
"Etnik Köken" **TÜRK İSE** ve
"Siyasi görüş" **ORTANIN SOLU İSE**
Oy Vereceği Parti = **CHP**

Kural 2

EĞER "Bugünlerde seçim olsa kim kazanır" **AKP İSE** ve
"Etnik Köken" **TÜRK İSE** ve
"Siyasi görüş" **SOL İSE**
Oy Vereceği Parti = **BDP**

Kural 3

EĞER "Bugünlerde seçim olsa kim kazanır" **KOALİSYON OLUR İSE**
"Etnik Köken" **TÜRK İSE** ve
"Siyasi görüş" **ORTANIN SAĞI İSE**
Oy Vereceği Parti = **MHP**

Kural 4

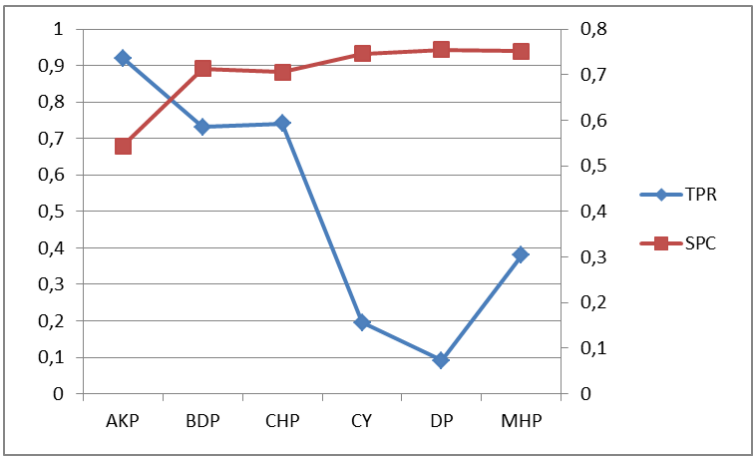
EĞER "Bugünlerde seçim olsa kim kazanır" **KOALİSYON OLUR İSE** ve
Etnik Köken **TÜRK İSE** ve
"Siyasi görüş" **BU TANIMLAR BENCE GEÇERSİZ İSE** ve
"Anayasa değişikliği sinir bozucu olacaktır" **DOĞRU İSE**
Oy Vereceği Parti = **CHP**

Kural 5

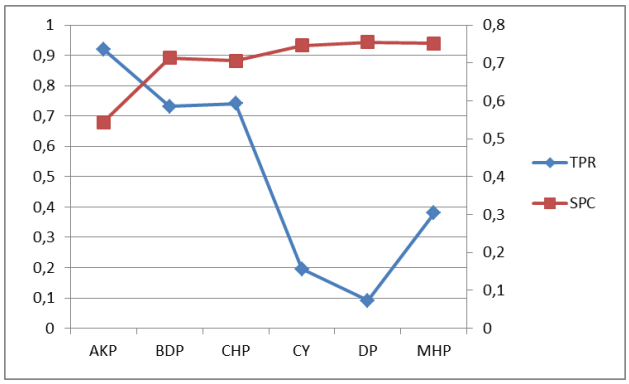
EĞER "Bugünlerde seçim olsa kim kazanır" **AKP İSE** ve
"Etnik köken" **KÜRT-ZAZA İSE** ve
"Siyasi görüş" **BU TANIMLAR BENCE GEÇERSİZ İSE**
"Oy tercihini etkileyen sebep" **SON DAKİKACI İSE**
Oy Vereceği Parti = **AKP**

Algoritma uygulamasında ortaya çıkan en iyi doğruluk değerine sahip eğitim seti oranına göre hedef niteliğinin tüm değerleri analiz edilmiştir. C4.5 algoritmasına ait %60 eğitim seti ayrışımında Duyarlılık (TPR), Belirleyicilik (SPC) karşılaştırmaları grafiği Şekil 2'de, %70 eğitim seti ayrışımında TPR, SPC karşılaştırmaları grafiği Şekil 3'de %80 eğitim seti ayrışımında TPR, SPC karşılaştırmaları grafiği Şekil 4'de %90 eğitim seti ayrışımında TPR, SPC karşılaştırmaları grafiği Şekil 5'de verilmiştir.

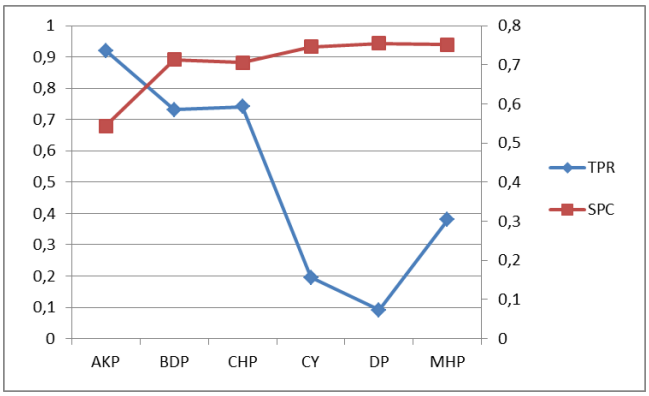
Şekil 2. %60 Eğitim seti ayrışımında C4.5 algoritması hedef nitelik değerlerine göre TPR-SPC Grafiği



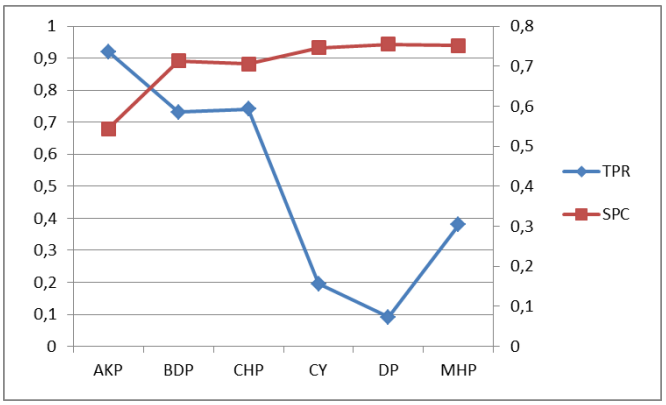
Şekil 3. %70 Eğitim seti ayrışımında C4.5 algoritması hedef nitelik değerlerine göre TPR-SPC Grafiği



Şekil 4. %80 Eğitim seti ayrışımında C4.5 algoritması hedef nitelik değerlerine göre TPR-SPC Grafiği



Şekil 5. %90 Eğitim seti ayrışımında C4.5 algoritması hedef nitelik değerlerine göre TPR-SPC Grafiği



Yapılan araştırmada C4.5 karar ağaç algoritmasının genel doğruluk ve hata oranları hesaplamalarında veri setinin eğitim için ayrılan %60, test için %40 ayrılmış ve aynı sırada %70-%30, %80-%20, %90-%10'luk diğer bölümlemeler veri setinin gözlemlerinden rastgele olarak seçilmiştir. C4.5 algoritmasının genel doğruluk değeri %69,1 - %72,7 arasında değiştiği görülmüş ve %90'lık bölünen eğitim setinde en yüksek başarı oranına sahip model elde edilmiştir. C4.5 karar ağaç algoritmasının en yüksek genel doğruluk oranını veren test kümesi üzerinde yapılan incelemelerde AKP, CHP, MHP, BDP, DP ve CY parametrelerinin duyarlılık oranı karşılaştırmalarında %93,8 ile AKP, %75,2 ile CHP, %82,1 ile BDP, %28,5 ile MHP, %20 ile CY ve %13,6 ile DP, belirleyicilik oranı karşılaştırmalarında %55,1 ile AKP, %71,9 ile CHP, %71,9 ile BDP, %77,5 ile MHP, %75,8 ile CY ve %76,6 ile DP olduğu elde edilmiştir.

SONUÇ

İnsanın etkileşim içerisinde bulunduğu çevre, aile, sosyal koşullar, inanç, etnik köken, parti, lider, kişinin eğitim düzeyi, gelir seviyesi, meslek, sosyo-ekonomik statü, bölgesel ikametgâh, işitsel ve görsel medya, ideolojik yapı gibi birçok etmen insanın siyasal eğilimini etkileyebilmektedir. Bu etkileşim ve etmenler çerçevesinde; seçmenlerin siyasal davranışlarının geçişli bir yapı sergileyebileceği açıktır. Buna bağlı olarak partilerin güç kaybetmesi, siyasi konjonktürün değişmesi ve beklenmedik parti seçimi ve/veya çıkışları gözlenebilmektedir.

Veri madenciliği yöntem ve tekniklerinin uygulandığı diğer sektörlerdeki başarılı örnekleri/sonuçlarına bakılarak, siyaset-seçmen davranışı gibi kritik bir alanda da değişim takibi ve sonuç tahmini açısından önemli bir avantaj sağlayacağı düşünülmektedir. Bu düşünce doğrultusunda yapılan çalışmada; KONDA Araştırma ve Danışmanlık şirketi tarafından 2011 yılı Türkiye Milletvekili Genel Seçimlerinden önce 3584 kişiden derlenen verilerin sınıflandırma tekniklerinden C4.5 karar ağacı kullanılarak analizi gerçekleştirilmiştir. Kurulan ağaç ile bir siyasi partinin seçiminde literatürde etkisi belirtilen faktörler dışında partiye özgü karar yolları da elde edilmiştir. Karar ağacından elde edilen kurallar; bir x partisinin seçmen profilinde ve oy alma/almama nedenlerindeki değişimi açıklayabilecek niteliktedir. Kurulan modelin başarı ve sınıf bazlı hassasiyet-belirleyicilik değerlerindeki yüksek yüzdeler ile C4.5 algoritmasının seçim öncesindeki sürecin doğru yönetilmesinde destek olabileceğini söylemek mümkündür.

Veri madenciliği yöntem ve tekniklerinin siyasi seçimlerde kullanılması, mikro seviyede seçmenlerin genel eğilimlerine göre hareket tarzının belirlenmesi ve oy verme algısındaki yeni-eski parametrelerin dinamik biçimde değerlendirilmesini sağlayacaktır. Makro düzeyde ise toplumun genel isteklerinin daha objektif kriterlerle belirlenerek, siyasi haritanın şekillendirilebilecektir.

İnsan davranışının göreceli tahmin edilemezliğine (Cziko, 1989) rağmen belirli koşullar sağlandığında sosyal bilimlerde matematik bilimi sayesinde yüksek bir seviyede tahmin edilebilir bir sonuç alınması araştırmayı bu yönde değerli kılmaktadır. Veri madenciliği ile seçmen eğilimlerinin analizi türünden bir çalışmanın Türkiye'de henüz yapılmamış olması, analiz için popüleritesi giderek artan ve veri madenciliğinde güçlü bir araç kabul edilen R dilinin kullanılmış olması araştırmanın güçlü yönleridir.

KAYNAKÇA

Cziko, G.A. (1989) Unpredictability and Indeterminism in Human Behavior: Arguments and Implications for Educational Research, *Educational researcher*,*18*(3), 17-25.

Dasu, T., Johnson, T.(2003) *Exploratory Data Mining and Data Cleaning*, John Wiley &Sons Publication, New Jersey, Inc., ISBN:0-471-26851-8.

Dijcks, J.P. (2013) *Oracle:Big Data for the Enterprise*, http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf, [Ziyaret Tarihi: 29 Mayıs 2015].

Doğan, A. ve Göker, G.(2010) Yerel Seçimlerde Seçmen Tercihi: 29 Mart Yerel Seçimleri Elazığ Seçmeni Örneği, *Eskişehir Osman Gazi Üniversitesi İİBF Dergisi*, 2, 159-187.

Gantz, J. ve Reinsel, D. (2012) *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, Biggest Growth in the Far East*, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf,[Ziyaret Tarihi: 5 Haziran 2015].

Hegland, M.(2003) *Data Mining–Challenges, Models, Methods And Algorithms*, http://datamining.anu.edu.au/publications/2003/dm_script_hegland.pdf, [Ziyaret Tarihi: 10 Haziran 2015].

Hilbert, M., ve Lopez, P. (2011) The World's Technological Capacity to Store, Communicate, and Compute Information, *Science*, 332, 60-65.

Hssina, B., Merbouha, A., Ezzikouri, H. ve Erritali, M. (2014) A comparative study of decision tree ID3 and C4.5, *International Journal of Advanced Computer Science and Applications*, 4(2), 13–19.

Kantardzic, M. (2011) *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons Publication, New Jersey, ISBN:978-1-118-02912-1.

Larose, D.T. ve Larose, C.D. (2015) *Data Mining and Predictive Analytics*, John Wiley & Sons Publication, New Jersey, ISBN: 978-1-118-11619-7.

Little, R. J.A. ve Rubin, D.B. (1987) *Statistical Analysis with Missing Data, Second ed.*, John Wiley &Sons Publication, New York, Inc.,ISBN:0-471-18386-5.

Mitchell, T. (1997) *Machine Learning*, McGraw-Hill, Maidenhead, U.K., ISBN: 0070428077.

Oğuzlar, A. (2003) Veri Ön İşleme, *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 21, 67–76.

Özekes, S.(2003) Veri Madenciliği Modelleri ve Uygulama Alanları, *İstanbul Ticaret Üniversitesi Dergisi*, 3, 64-82.

Shi, Y., Zhang, L., Tian, Y. ve Li, X. (2015) *Intelligent Knowledge: A Study beyond Data Mining*, Springer, London, ISBN: 978-3-662-46193-8.