ÜLKELERDEN GİDEN MÜLTECİ SAYILARI ÜZERİNE VERİ MADENCİLİĞİ UYGULAMASI

Ali Mertcan KÖSE*

Ayça Çakmak PEHLİVANLI**

Özet

Mülteci; ırkı, dini, milliyeti, belli bir sosyal gruba mensubiyeti veya siyasi düşünceleri nedeniyle zulüm göreceği konusunda haklı bir korku taşıyan ve bu yüzden ülkesinden ayrılan ve korkusu nedeniyle geri dönemeyen veya dönmek istemeyen kişi olarak nitelendirilmiştir. Bu nitelik temel alınarak yapılan araştırmalarda mülteci sayıları elde edilmiştir. Elde edilen bu sayılarla beraber, kişilerin mülteci olma sebebi üzerine etki ettiğini düşündüğümüz sosyo-ekonomik değişkenler veri setimize eklenmiştir. Bu çalışmada kullanılan veri seti 2008-2013 yılları arasında 215 ülkeden giden mültecilerin sosyal-ekonomik özellikleri ve kategorize edilmiş mülteci sayılarından oluşmaktadır. K-Yakın Komşu, Naive Bayes ve Karar Ağacı gibi yaygın kullanılan veri madenciliği teknikleri ile elde edilen sınıflama oranları karşılaştırmalı olarak çalışmamızda yer almıştır. Yapılan çalışmalarda k kat çapraz geçerliliği yapılmış, K- Yakın Komşu algoritmasının en iyi sınıflama oranını verdiği görülmüştür. Diğer yöntemlerin etkinliğini artırmak için Temel Bileşenler Analizi ile boyut indirgemesi yapılmış, sözü geçen yöntemler indirgenmiş veri seti üzerine uygulanarak sınıflama başarısının arttığı gözlenmiştir.

Anahtar Kelimeler: Mülteci, Veri madenciliği teknikleri, Temel Bileşenler Analizi

IMPLEMENTION OF DATA MİNİNG TECHNİCS CONSİDERİNG NUMBER OF REFUGEES

Abstract

Refugees are people fleeing conflict or persecution. because of their race, religion, nationality, membership in a particular social group, or political opinion. Refugees are defined and protected in international law, and must not be expelled or returned to situations where their life and freedom are at risk. The final refugees numbers used in this study were collected based on the given definition. In addition to obtained numbers, we also added socio-economic variables which affect '' being a refugee''. The data set used in this study includes both socio-economic attributes of refugees from 215 countries beetween 2008 and 2013 categorized refugee numbers. We used common data mining techniques such as Naive Bayes Decision Trees and K- Nearest Neighborhood classification algorithm in order to compare classification ratios. According to our study we obtained best classification ratio with K-Neighborhood algorithm with k-fold cross validation. In order to increase the classification success rate, it is also applied Principle Component Analysis. As a result of dimension reduction, better classification results were observed.

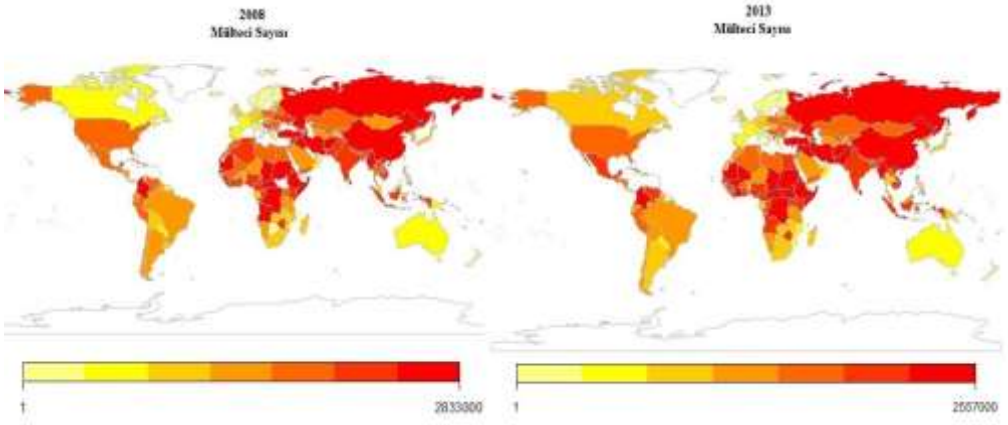
Key Words: Refugee, Data mining techniques, Principal Compenent Analysis

* Mimar Sinan GSÜ., İstatistik Bölümü, alimertcankose@gmail.com

** Yrd. Doç. Dr., Mimar Sinan GSÜ., İstatistik Bölümü, ayca.pehlivanli@msgsu.edu.tr

GİRİŞ

Mülteci Birleşmiş Milletlerin tanımına göre; ırkı, dini, milliyeti, belli bir sosyal gruba mensubiyeti veya siyasi düşünceleri nedeniyle zulüm göreceği konusunda haklı bir korku taşıyan ve bu yüzden ülkesinden ayrılan ve korkusu nedeniyle geri dönemeyen veya dönmek istemeyen kişi olarak nitelendirilmiştir. Bu nitelik temel alınarak yapılan araştırmalarda mülteci sayıları elde edilmiştir. Elde edilen bu sayılarla beraber, ülkelerden mülteci eden kişilerin mülteci olma sebebi üzerine etki ettiğini düşündüğümüz sosyal ve ekonomik değişkenler veri setimizin oluşmasında etkin bir rol oynamıştır. Bu oluşumda veri madenciliği yöntemleri kullanılmıştır. Veri madenciliğine basit bir tanım yapmak gerekirse, büyük ölçekli veriler arasından bilgiye ulaşma, bilgiyi bir madenden değerli bir cevheri çıkarıyormuşçasına gün yüzüne çıkarma işidir. Veri madenciliği, eldeki verilerden üstü kapalı, çok net olmayan, önceden bilinmeyen ancak potansiyel olarak kullanışlı bilginin çıkarılmasıdır.(Kalikov,2006) Bu bilgilerin çıkarılması için düzenlenen veri setimiz 2008-2013 yıllarındaki mülteci sayıları ve diğer sosyo-ekonomik değişkenler üzerinden belli veri madenciliği tekniklerinin kullanılmasıyla sınıflama olasılıklarının bulunması ve karşılaştırılmalarının yapılmasıdır. Ayrıca Sınıflama oranlarının yükseltilmesi için bazı uygulamalar yapılmıştır. Mülteci Sayısı ile ilgili değişimi göstermek amacıyla Şekil 1 de Dünya haritasında 2008 ve 2013 yıllarındaki mülteci sayıları görselleştirilmiştir. Görsele göre bazı ülkelerin renklerinin değişimi net olarak görülmüştür.



Şekil 1. 2008-2013 Mülteci sayısı değişimi

İLGİLİ ÇALIŞMA

Mülteci sayıları ve bununla birlikte etki ettiğini düşündüğümüz değişkenler üzerine yapacağımız bu çalışma, 1070 gözlemden oluşan 2008-2013 yılları arasındaki 215 ülkeden giden mülteci sayıları ve buna eşdeğer olarak diğer sosyo- ekonomik değişkenlerimizden oluşmaktadır. Veri setimizdeki değişkenler Dünya Bankasının veri tabanından elde edilmiştir ve Tablo 1 de gösterilmiştir.

Tablo1. Veri setinde bulunan değişken sayısı

Değişken Kodu	Değişken Açıklaması
Country Code	Ülke Kodu
FoodProdIndx	Gıda Üretim İndeksi
InflGDPdflt	Gsyih deflatörüne göre Enflasyon Oranı
GDP_growth	Gsyih göre büyüme oranı
MortltyrteUndr5	5 yaşından küçükler için ölüm oranı
GDPperCapita	Gsyih göre Kişi başına gelir
Agrcltrlnd	Tarımsal alan oranı
RailLines	Tren yolları uzunluğu
AirTrnsprt	Hava taşımacılığı, Dünya çapında tescilli olan taşıyıcıların kalkışları
Lifeexpect	Beklenen yaşam süresi
Urbnpop	Kentsel Nüfus Oranı
Rrlpop	Kırsal Nüfus Oranı
Unemplyment	İşsizlik Oranı
Lbrforcepart	İş gücüne katılım oranı
Lbrforce	İş gücü
Militaryexp	Askeri Harcama oranı
Refugewasyl	Ülke veya sığınma ülkesi tarafında olan Mülteci Nüfus
Refugeeorigin	Ülkelerden giden Mülteci Nüfus

METOTLAR

Veri madenciliğinde sıklıkla kullanılan algoritmaların başında sınıflama teknikleri gelir. Karar ağaçları algoritması, istatistiğe dayalı algoritmalar, uzaklık matrisine dayalı algoritmalar ve yapay sinir ağları sınıflama teknikleri olarak dört ana başlıktan oluşmaktadır. Bu çalışmada karar ağacı, istatistiğe dayalı ve uzaklık matrisine dayalı algoritmalar kullanılarak uygulama yapılmış ve uygulama yapılan algoritmalar hakkında kısa bilgiler verilmiştir. Algoritmaların etkinliğini arttırmak için Temel Bileşenler Analizi uygulanarak karşılaştırmalar yapılmıştır. Böylelikle tahmini değerlerin elde edilmesi için kullanılacak en iyi analiz yolu belirlenmiştir.

I-Naive Bayes Algoritması:

Naive Bayes sınıflandırma tekniği, elde var olan, hali hazırda sınıflanmış verileri kullanarak yeni bir gözlemin mevcut sınıflarda herhangi birine girme olasılığını hesaplayan bir yöntemdir. Bayes kuralına dayalı geliştirilmiş bu algoritma Naive Bayes sınıflandırma tekniği olarak adlandırılır. Bayes teorimi şu şekilde ifade edilir.

$$P(C_1|x_i) = \frac{P(C_1|x_i)\,P(C_1)}{P(x_i|C_1)\,P(h_1)+P(x_i|C_2)\,P(C_2)}$$

Burada C_1 ve C_2 olarak gösterilen iki ayrı hipotezin, başka bir deyişle iki ayrı sınıfın olduğu kabul edilmiştir. $P(C_1|x_i)$ x_i'nin C_1 sınıfında olma olasılığını ifade etmektedir. $P(x_i)$, x_i değerinin verisetindeki bulunma sıklığı/ sayısıdır.

Eğer m adet hipotez- sınıf- olduğu düşünülürse bu durumda kural;

$$P(x_i)=\sum_{j=1}^{m}(x_i|C_j)\,P(C_j)$$

Şeklinde olacak ve bu durumda x_i nin C_1 sınıfında olma olasılığı aşağıdaki bağıntı ile hesaplanır.

$$P(C_1|x_i) = \frac{P(x_i|C_1)\,P(C_1)}{P(x_i)}$$

Bayes algoritması, öncelikle kendisine verilen öğrenme kümesinde $P(C_j)$ değerini her sınıfın verilen öğrenme kümesi içinde bulunma sıklığını hesaplar. Daha sonra, x_i'ler sayılarak $P(x_i)$ değeri bulunur. Benzer şekilde herbir sınıfta, herbir x_i değerinin bulunma sıklığı $P(x_i|C_1)$, C_j'ler içinde x_i'lerin sayılmasıyla elde edilir.

II-K –En Yakın Komşu Algoritması:

En sık tercih edilen algoritmalardan birisidir. İngilizce kaynaklarda K- Nearest Neigbour ya da KNN şeklinde ifade edilir. Bu algoritmada sınıflandırma yapılırken veri setindeki her bir gözlemin diğer gözlemlerle olan uzaklığı hesaplanır. Ancak, bir gözlem için diğer gözlemden sadece k adedi göz önüne alınır. Algoritmanın isminden anlaşılacağı gibi bu k adet gözlem, uzaklığı hesaplanan noktaya diğer gözlemlere kıyasla en yakın olan gözlemdir.

Algoritmada k değeri önceden seçilir; değerinin yüksek olması birbirine benzemeyen noktaların bir araya toplanmasına, çok küçük seçilmesiyle birbirine benzediği, yani aynı sınıfın noktaları oldukları halde, bazı noktaların ayrı sınıflara konmasına ya da o tür noktalar için ayrı sınıfların açılmasına neden olur.(Silahtaroğlu,2008:65)

III-C4.5 (J48) Karar Ağacı Algoritması:

Verilerin sınıflandırma yöntemlerinden biri karar ağaçları ile sınıflandırma adını taşımaktadır. Uygulama istatistikte makine öğrenmesi başlığı altında bir çok karar ağacı algoritması geliştirilmiştir. Örneklerden oluşan bir küme kullanılarak karar ağacının oluşmasını sağlayan bir yöntemdir. Karar ağaçları akış şemalarına benzeyen yapılardır. Her bir nitelik bir düğüm tarafından temsil edilir. dallar ve yapraklar ağaç yapısının elemanlarıdır. En son yapı yaprak, en üst yapı kök ve bunların arasında kalan yapılar ise dal olarak isimlendirilir.(Quinlan, 1993). Diğer yöntemlerin daha çok kategorize edilmiş verilerde kullanılmasından dolayı bir eksiklik oluşmuştur. Bu sebeple sayısal verilerde daha iyi bir sonuç elde etmek amacıyla C4.5 algoritması geliştirilmiştir.(Özkan,2008:53)

IV-Temel Bileşenler Analizi:

Bir diğer adı Karhunen- Loeve metodudur. Türkçesi temel bileşenler analizi olan PCA, tanıma, sınıflandırma, görüntü sıkıştırma alanlarında kullanılan, bir değişkenler setinin varyans - kovaryans yapısını, bu değişkenlerin doğrusal birleşimleri vasıtasıyla açıklayarak boyut indirgenmesi ve yorumlanmasını sağlayan, çok değişkenli bir istatistik yöntemidir. Bu yöntemde karşılıklı bağımlılık yapısı gösteren, ölçüm sayısı (n) olan (p) adet değişken; doğrusal, dikey (ortogonal) ve birbirinden bağımsız olma özelliklerini taşıyan (k) tane yeni değişkene dönüştürülmektedir. PCA, verideki gerekli bilgileri ortaya çıkarmada oldukça etkili bir yöntemdir. Yüksek boyutlu verilerdeki genel özellikleri bularak boyut sayısının azaltılmasını, verinin sıkıştırılmasını sağlar. Boyut azalmasıyla bazı özellikleri kaybedileceği kesindir; fakat amaçlanan, bu kaybolan özelliklerin veri seti hakkında çok az bilgi içeriyor olmasıdır.(Yılmaz,Çamurcu ve Doğan,2010:252)

PCA ayrıca veri içindeki örüntüyü bulmaya çalışır. Bu yüzden örüntü bulma tekniği olarak da kullanılabilir. Çoğunlukla verinin sahip olduğu çeşitlilik, tüm, boyut takımından

seçilen küçük bir boyut setiyle yakalanabilir. Böylece PCA kullanılarak yapılan boyut küçültme işlemleri, daha küçük boyutlu veri setlerinin ortaya çıkmasını sağlar ve böylece yüksek boyutlu verilere uygun olmayan teknikler bu veri seti üzerinde rahatça çalışabilir. Verideki gürültüler, örüntülerden daha güçsüz olduklarından, boyut küçültme sonucunda bu gürültüler temizlenebilir. Bu özellik hem veri madenciliğinde hem de diğer veri analizi algoritmalarında özellikle kullanışlıdır. (Yılmaz,Çamurcu ve Doğan,2010:252)

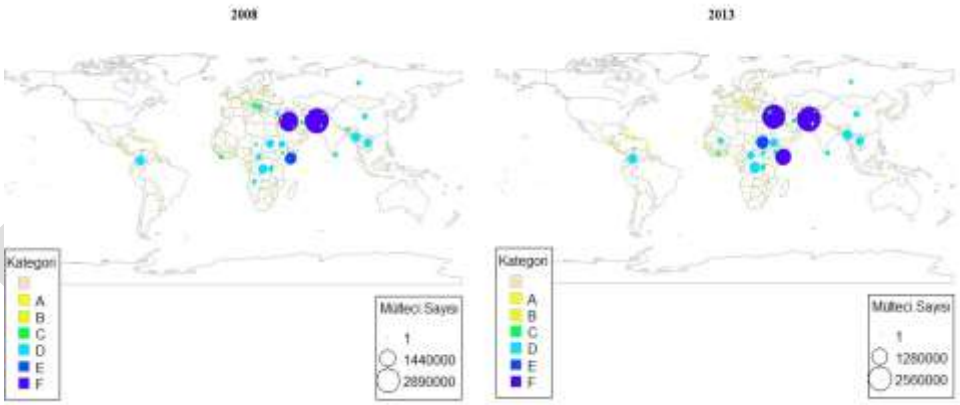
UYGULAMA

2008- 2013 yılları arasındaki ülkelerden giden mülteci sayısı ve diğer değişkenlerimizin oluşturduğu veri setimiz üzerindeki uygulama WEKA 3.6 programı kullanılarak yapılmıştır.(Hall ve ark,2009) Algoritmanın çalışabilmesi için mülteci sayısı kategorize edilmiştir. Bu kategorize Tablo 2 de gösterilmiştir. Kategorize işleminden sonra Weka 3.6 Programına verilerimiz girilmiştir.

Tablo 2. Mülteci sayılarına ilişkin kategoriler

Mülteci sayısı	Kategori Adı
1-10000	A
10000-50000	B
50000-100000	C
100000-500000	D
500000-1000000	E
1000000-	F

Mülteci sayılarına ilişkin 2008 ve 2013 yıllarına ait Kategorize edilmiş verilerin Görseli şekil 2 de gösterilmiştir.



Şekil 2. Mülteci sayılarının 2008 ve 2013 yıllarına ait kategorize edilmiş gösterimi

Sınıflama Olasılıkları Tablo 3 te gösterilmiştir.

Tablo 3. Sınıflama yöntemlerine ilişkin sonuçlar

Metot	Doğru Sınıf Sayısı	Doğru Sınıflama Oranı
Naive Bayes	642	%67.0146
C4.5 Karar Ağacı	708	%73.904
K-yakın komşu	825	%86.1169

Naive Bayes Sınıflandırma tekniğine göre kategorize edilmiş mülteci sayılarının olasılıkları Tablo 4 te gösterilmiştir.

Tablo4. Naive bayes ile kategorilere göre mülteci sayıları olasılıkları

Kategori	F	B	A	C	D	E
Oran	0.01	0.13	0.74	0.04	0.07	0.01

Metotlara göre Doğru sınıflama oranları, Roc değerleri ve bazı değerlendirme istatistikleri olasılıklar Tablo 5 de gösterilmiştir.

Tablo5. Metotlara ilişkin değerlendirme istatistikleri

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Naive Bayes	0.75	0.024	0.281	0.75	0.409	0.933	F
	0.699	0.159	0.393	0.699	0.503	0.86	B
	0.71	0.1	0.953	0.71	0.814	0.753	A
	0.488	0.067	0.247	0.488	0.328	0.778	C
	0.318	0.037	0.389	0.318	0.35	0.798	D
	0.375	0.043	0.068	0.375	0.115	0.698	E
Weighted Avg.	0.67	0.1	0.796	0.67	0.71	0.773	
	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
Karar ağacı	0	0	0	0	0	0.433	F
	0	0	0	0	0	0.49	B
	1	1	0.739	1	0.85	0.497	A
	0	0	0	0	0	0.489	C
	0	0	0	0	0	0.481	D
	0	0	0	0	0	0.399	E
Weighted Avg.	0.739	0.739	0.546	0.739	0.628	0.493	
	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
K-yakın komşu	0.583	0.002	0.778	0.583	0.667	0.749	F
	0.886	0.055	0.703	0.886	0.784	0.887	B
	0.874	0.04	0.984	0.874	0.926	0.895	A
	0.805	0.023	0.611	0.805	0.695	0.877	C
	0.818	0.057	0.514	0.818	0.632	0.881	D
	0.375	0.003	0.5	0.375	0.429	0.686	E
Weighted Avg.	0.861	0.042	0.893	0.861	0.87	0.889	

Temel Bileşenler yapıldıktan sonraki doğru sınıflama oranları Tablo 6 da ve Tablo 7 de gösterilmiştir.

Tablo6. İndirgenmiş veri ile elde edilen sonuçlar

Metot	Doğru Sınıf Sayısı	Doğru Sınıflama Oranı
Naive Bayes	759	%79.2276
C4.5 Karar Ağacı	888	%92.6931
K-yakın komşu	900	%93.9457

Tablo7. İndirgenmiş veri ile elde edilen sonuçlar

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Naive Bayes	0.667	0.003	0.727	0.667	0.696	0.935	F
	0.683	0.072	0.583	0.683	0.629	0.887	B
	0.825	0.136	0.945	0.825	0.881	0.83	A
	0.805	0.053	0.402	0.805	0.537	0.896	C
	0.697	0.043	0.548	0.697	0.613	0.884	D
	0.5	0.016	0.211	0.5	0.296	0.678	E
Weighted Avg.	0.792	0.115	0.839	0.792	0.808	0.844	
	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
Karar ağacı	0.75	0.006	0.6	0.75	0.667	0.872	F
	0.846	0.025	0.832	0.846	0.839	0.926	B
	0.984	0.064	0.978	0.984	0.981	0.878	A
	0.61	0.009	0.758	0.61	0.676	0.833	C
	0.788	0.019	0.754	0.788	0.77	0.883	D
	0.125	0.002	0.333	0.125	0.182	0.742	E
Weighted Avg.	0.927	0.052	0.924	0.927	0.925	0.881	
	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
K-yakın komşu	0.833	0.003	0.769	0.833	0.8	0.89	F
	0.902	0.022	0.86	0.902	0.881	0.941	B
	0.989	0.02	0.993	0.989	0.991	0.835	A
	0.585	0.011	0.706	0.585	0.64	0.781	C
	0.788	0.018	0.765	0.788	0.776	0.898	D
	0.375	0.006	0.333	0.375	0.353	0.685	E
Weighted Avg.	0.939	0.019	0.94	0.939	0.939	0.85	

Tablolara göre tekniklerin doğru sınıflama oranları ve bazı istatistikler gösterilmiştir. Tablolar ayrıntılı incelendiğinde görülmüştür ki indirgenmiş veri ile elde edilen sonuçlar orijinal veri ile elde edilen sonuçlara göre oldukça yüksektir. Özellikle tablo 5 ve tablo 7 karşılaştırıldığında görülecektir ki, ROC değerleri indirgenmiş veri ile elde edilen sınıflama

sonuçlarında 1'e oldukça yakın bulunmuştur. Bu da sınıflama başarısının önemli bir göstergesidir.

SONUÇ

Tablo 8. İndirgenmiş ve İndirgenmemiş değerlerin Karşılaştırma Tablosu

Metot	Doğru Sınıf Sayısı	Doğru Sınıflama Oranı	Metot	Doğru Sınıf Sayısı	Doğru Sınıflama Oranı
Naive Bayes	642	%67.0146	**Naive Bayes**	759	%79.2276
C4.5 Karar Ağacı	708	%73.904	**C4.5 Karar Ağacı**	888	%92.6931
K-yakın komşu	825	%86.1169	**K-yakın komşu**	900	%93.9457

Mülteci sayıları üzerine yapılan çalışmada veri madenciliği teknikleri kullanılarak değişkenlerimiz üzerindeki sınıflama oranları bulunmuştur. Yapılan ilk çalışmada kullanılan teknikler üzerinden en iyi sınıflamayı K yakın komşu yöntemi %86.1, C4.5 %74.0, Naive Bayes ise %67.0 olarak yapmıştır. Yöntemlerin etkinliğini artırmak için veri setimiz üzerinde Temel Bileşenler Analizi yapıldıktan sonra bir değerlendirme yapılmıştır. Bu değerlendirme sonucunda %67.0 olan Naive Bayes sınıflandırma oranı Temel Bileşenlerden sonra %79.2 oranına, %74.0 olan C4.5 algoritmasının %92.7 ve %86.1 olan K yakın komşu yöntemi %93.9' a yükseldiği gözlemlenmektedir. Bu değerlere göre, tahmini değerlerin daha iyi bir sonuç vermesi açısından Temel Bileşenlerin veri setimiz üzerindeki etkinliği arttırdığını söyleyebiliriz. Ayrıca veri setimizi oluşturan sosyo-ekonomik değişkenlerin sınıflama tekniklerine ilişkin olasılık değerlerine göre mülteci sayılarına ilişkin tahmini değerleri yapılabilir. Ve sınıflama oranı yüksek olan algoritmanın ülkelerden giden mülteci sayılarına ilişkin tahmin modelinde etkin olduğunu söyleyebiliriz.

KAYNAKÇA

Şeker, Ş., E., (2013), İş Zekası ve Veri Madenciliği Weka İle, Cinius Yayınları, İstanbul.

Pektaş, A., O.,(2013), SPSS ile Veri Madenciliği, Dikey Eksen Yayın, İstanbul.

Çokluk, Ö., Şekercioğlu, G.,(2012), Büyüköztürk, Ş., Sosyal Bilimler için Çok Değişkenli İstatistik SPSS ve LISREL uygulamaları, Pegem Akademi, Ankara.

Silahtaroğlu, G., (2008), Kavram ve Algoritmalarıyla Temel Veri Madenciliği, Papatya Yayıncılık, İstanbul.

Özkan, Y., (2008), Veri Madenciliği Yöntemleri, Papatya Yayıncılık, İstanbul.

Yıldız, K., Çamurcu Y., Doğan, B., (2010), Veri Madenciliğinde Temel Bileşen Analizi ve Negatifsiz Matris Çarpanlarına Ayırma Tekniklerinin Karşılaştırmalı Analizi, Muğla Üniversitesi 10. Akademik Bilişim Konferansı Bildirileri, s. 252.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H,(2009), The WEKA Data Mining Software: An Update, SIGKDD Explorations, Volume 11, Issue 1.

Miljkovic, D., Gajic, L., Kovacevic, A., Konjovic, Z., (2010), The Use of Data Mining for Basketball Matches Outcomes Prediction, IEEE 8th International Symposium on Intelligent Systems and Informatics, p.309-312

Abedin, J., Mittal,H.,V., (2014), R Graphs Cookbook, Packt Publishing Ltd., Birmingham.

Birleşmiş Milletler Mülteciler yüksek komiserliği, http://www.unhcr.org/turkey/home.php?page=29 [Erişim tarihi 17.1.2016]

Dünya Bankası Veri sitesi, http://databank.worldbank.org/data/home.aspx[Erişim tarihi 17.1.2016]