Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama

Betül Yazıcı[1], Fethiye Yaslı[1], Hande Yıldız Gürleyik[2], Umut Orçun Turgut[2]
Mehmet S. Aktas[1],Oya Kalıpsız[1]

[1]Bilgisayar Mühendisliği Bölümü, Elektrik-Elektronik Fakültesi
Yıldız Teknik Üniversitesi, İstanbul

[2] Ar-Ge Merkezi, Cybersoft, İstanbul

E-posta: {betulyazicii@gmail.com,fethiyeyasli@gmail.com, hande.gurleyik@cs.com.tr, umut.turgut@cybersoft.com.tr, mehmet@ce.yildiz.edu.tr, oya@ce.yildiz.edu.tr}

Özet. Günümüzde pek çok kurum mevcut verilerini ilişkisel veri tabanlarında saklamakta ve modellemelerini bu verileri kullanarak gerçekleştirmektedir. Kurumsal veri modellerinin karmaşıklığı, veriye ait özelliklerin çokluğu ve veri miktarının fazlalığı, veri üzerinde her türlü analizin (kümeleme, regresyon, vb.) yapılmasını zorlaştırmaktadır. Bu nedenle veri kümeleri üzerinde tahmin gücü yüksek özelliklerin belirlenebilmesi için kolay kullanılabilir, yaygın kullanıma sahip mevcut araçlarla (R, Weka) entegre olabilecek ve karşılaştırmalı olarak en iyi tahmini üretebilecek yazılımlara ihtiyaç bulunmaktadır. Literatürde, özellikleri inceleyen üç temel yaklaşım vardır. Bunlar, entropi yöntemiyle belirsizliği ölçen Bilgi Teorisi Ki-kare (X^2) istatistiğini kullanarak özelliklerin taşıdığı bilginin birbirinden farklılığını ölçen Geleneksel İstatistik ve negatif entropiyi kullanarak bilgi değerini ölçen Öngörüsel Analiz yaklaşımlarıdır. Bu araştırma kapsamında bahsedilen ihtiyaçlara yanıt verebilmek amacıyla Öngörüsel Analiz yaklaşımını kullanan ve tahmin gücü yüksek özellikleri belirleyen bir yazılım üretilmiştir. Bu bildiriyle yazılımın geliştirilmesi sürecinde kullanılan yöntemler, teknikler, algoritmalar ve geliştirilen yazılım detaylı olarak anlatılmıştır. Geliştirilen yöntemler aynı bankacılık veri kümesinde uygulanmış ve sonuçları karşılaştırmalı olarak analiz edilerek yorumlanmıştır.

Anahtar Kelimeler: Veri Madenciliği, Öngörüsel Yaklaşım, Özellik Seçimi, Bilgi Kazancı, Bilgi Değeri, Kazanım Oranı

1. Giriş

Özellikler, gerek gözetimli gerekse de gözetimsiz yöntemler tercih edilerek bağımlı değişkeni açıklamada kullanılan etkenlerdir. Özellik Seçimi işlemi, bağımlı değişkenle ilgisi olmayan, tahminleyici bilgisi az veya hiç olmayan özellikleri eleyerek bağımlı değişkeni açıklama kabiliyeti yüksek olan özelliklerin tespitini sağlamaktadır [1]. Bu çalışmada gözetimli veri kümeleri üzerinde uygulanan özellik seçim yöntemleri kullanılmıştır.

Son on yılda sınıflandırma algoritmalarının üstünde uygulandığı veri kümelerindeki özellik sayıları binleri hatta on binleri bulmaktadır. Bu nedenle araştırmacılar özellik seçme yöntemlerine her zamankinden daha fazla ihtiyaç duymaktadırlar [2]. Seçilen özelliklerle yapılan sınıflandırmada, işlem sayısı azalmakta, gürültülü ve ilgisiz özellikler özgün veriden çıkarılarak sınıflama başarısı arttırılmakta ve özellikler üzerinden yapılabilen sınıflama yorumları artmakta veya kolaylaşmaktadır. Bunlara ek olarak model eğitim zamanı kısalmakta, daha az ölçüm yapılmakta ve daha az bellek kullanılmaktadır. Bu yararlar, modeli tanımanın anlamlı ve daha kolay olmasını sağlamaktadır.

Korelasyonları yüksek birçok özelliğin bulunduğu ve örnek sayısının az olduğu veri kümelerinde özellik seçme algoritmalarının önemi bir kat daha artmaktadır. Bu durumlarda özellik seçme algoritmaları hem veri kümesi içindeki gürültülü, sapkın ve gereksiz özellikleri eleyerek verilerin daha sağlıklı ifade edilmesini sağlamakta hem de örneklemdeki kayıt sayısının az olduğu hallerde sınıflandırıcı algoritmanın başarı oranını artırmaktadır [3].

Bu bildiride, bankacılık sektörüne ait örnek veri kümesi kullanılarak tahminleyici özelliklerin belirlenmesi üzerine farklı algoritmalar (Bilgi Kazancı, Kazanım Oranı, Bilgi Değeri) uygulama kapsamında geliştirilmiş, sonuçları Weka ve R kullanılarak karşılaştırmalı biçimde irdelenmiştir.

2. Özellik Seçimi

Özellik seçimi, kümeleme veya regresyon işlemleri için kullanılacak özelliklerin belirlenmesi aşamasında, tüm özellik kümesi sütunlarından bağımlı değişkenle olan ilişkiyi açıklamada, ilgisiz sütunların elenmesi ve açıklayıcı gücü yüksek sütun alt kümelerinin belirlenmesi işlemidir. Özellik seçimi genel olarak doğruluk ve ölçeklenebilirlik için kullanılmaktadır. İlk bakışta, veri kümesindeki tüm özelliklerin analize konu edilmesiyle, sınıflandırma veya bağımlı değişkeni açıklayan regresyon algoritmalarının başarılı sonuçlar vereceği akla gelmektedir. Oysa bu düşünce pek çok özellik içeren veri kümelerinde her zaman doğru olmayabilir. Veri kümesindeki her özellik bağımlı değişken hakkında açıklayıcı ya da tahminleyici bilgi taşımayabilir. Dolayısıyla özelliklerin tahminleyici bilgi taşıma durumuna göre ayırt edilip analize konu edilmesi gerekir. Genel bir ifadeyle aksi durum, regresyon modelinde bağımsız değişken enflasyonu yaratırken modelin katsayıları açısından istatistiksel olarak anlamlılığını azaltıcı etki yaratır [10]. Başka bir ifadeyle, veri kümesi içindeki bazı

özellikler işlem performansını olumsuz etkileyecek gürültüye sahip olduğundan bu özelliklerin veri kümesi içinden silinmesi, işlem sonucunun doğruluğunun artmasında etkili olabilmektedir. Diğer taraftan algoritmalarda kullanılacak veri boyutunun azaltılması da işlem gücü, hafıza ihtiyacı ve depolama alanı gibi işlem süreci üzerinde etkili konularda zaman tasarrufu sağlar.

2.1. Özellik Seçimi Yöntemleri

Bu bölümde özellik seçmede kullanılan yöntemler kısaca ele alınmaktadır. Bu yöntemler sınıf etiketi olan gözetimli veri kümesi üzerinde uygulanan yöntemlerdir. Literatürde kullanılan özellik seçme yöntemleri bunlarla sınırlı olmamakla birlikte veri kümesinin pek çok özellik içerdiği durumlarda gözetimli analize konu olacak özelliklerin belirlenmesinde sıklıkla kullanılan yöntemler bu başlık altında incelenmektedir.

2.1.1 Bilgi Kazancı Yöntemi

Bilgi Kazancı entropiye dayalı özellik seçim yöntemidir. Entropi, bir sistemdeki düzensizliğin ya da belirsizliğin ölçüsüdür ve (1) numaralı formüldeki gibi ifade edilmektedir. Entropi 0 ve 1 aralığında değerler alır ve 1 değerine yaklaştıkça belirsizlik artar. Yüksek entropiye sahip veri daha çok bilgi içerir.

$$E(D) = -\sum_{k=1}^{m} p_i log_i\ (p_i) \quad (1)$$

p_i, *D* veri kümesindeki "*i*" sınıfının olasılığıdır ve "*i*" sınıfına düşen örnek sayısının tüm veri kümesindeki toplam örnek sayısına bölünmesiyle elde edilir.

Bilgi Kazancı yöntemi, en ayırt edici özelliği belirlemek için kullanılır ve veri kümesindeki her bir özellik için ölçülür. *D* veri kümesi, *n* tane alt bölüme *X* özelliğinden bölünecekse *X*'e ait bilgi kazancı hesaplanması (2) numaralı formülle gerçekleştirilir.

$$Bilgi\ Kazancı\ (D,X) = E(D) - \sum_{k=1}^{n} p(D_i)E(D_i) \quad (2)$$

E(D); veri kümesinin *X* üzerinden bölünmeden önceki entropisini, $E(D_i)$*; i* alt bölümünün *X* üzerinden bölünme olduktan sonraki entropisini ve *p(Di)* ise *i* alt bölümünün *X* üzerinden bölünme olduktan sonraki olasılığını göstermektedir[4]. Veri kümesinin bölünmeden önceki belirsizliğinin yüksek olması, verinin, bilgi verici niteliğinin olduğunu göstermektedir. Bölünmeden sonraki belirsizliğinin düşük çıkmasıysa bu yöntemin veriyi dallara ayırma işlemini düzgün yaptığını göstermektedir. (2) numaralı formüle göre *E(D)*'nin yüksek çıkarken $p(Di)E(D_i)$ çarpımları toplamının düşük çıkması bilgi kazancını artırmaktadır.

2.1.2 Kazanım Oranı Yöntemi

Bilgi Kazancı yöntemi çok çeşitli değerlere sahip özellikleri seçme eğiliminde olduğundan sonuçları sapmalı bir yöntemdir[11]. Bu sapmanın azaltılması amacıyla Kazanım Oranı yöntemi oluşturulmuştur. Sapmayı azaltmak için bölünme bilgisi (Split Information) kullanılmaktadır. Bölünme Bilgisi (3) numaralı formülde gösterilmektedir.

$$Bölünme\ Bilgisi\ \ (S) = -\sum_{i=1}^{v}\left(\frac{|S_i|}{|S|}\right) log_2\ (\frac{|S_i|}{|S|}) \quad (3)$$

Kazanım Oranı, bilgi kazancı değerlerini, bölünme bilgisine oranlayarak bir çeşit normalizasyona tabi tutar. Bu terim nitelik değerinin veriyi nasıl böldüğü konusunda hassastır[5].

$$Kazanım\ Oranı\ (A) = Bilgi\ Kazancı\ (A)/Bölünme\ Bilgisi\ (S) \quad (4)$$

(3) ve (4) numaralı formüller kullanılarak en yüksek kazanım oranına sahip özellikler belirlenmiş olur.

2.1.3 Bilgi Değeri Yöntemi

Bilgi değeri, veri kümesindeki özelliklerin tahminleyici gücünü hesaplayan istatistiksel bir yöntemdir. Özelliklerin taşıdığı bilgi değerine göre tahminleyici güçleri arasında karşılaştırma yapmak mümkün olmaktadır. Bilgi değerinin ölçülmesinde bir hipotezi destekleyen kanıtları birleştirmek için kullanılan ve niceliksel bir yöntem olan Kanıtsal Ağırlık'a yer verilmektedir. Kanıtsal Ağırlık, özelliklerin tahmin gücünü hedeflenen sınıfa göre analiz eder ve konuyu olumlu ve olumsuz olmak üzere iki taraflı olarak ele alır. Burada bahsedilen iki taraflı durum, bireyin bir ürünü satın alma veya almama ihtimali olabileceği gibi bir kredi müşterisinin kredi borcunu ödeyip ödeyememesi durumu gibi kesikli, ayrık bir durum da olabilir. Kanıtsal Ağırlık tanımıyla özellik bazında bu durumların birbirinden ne kadar ayrışık olduğu belirlenebilir[12, 13].

(5) numaralı denklemde pay ve paydada sırasıyla, kredi kartı alanların ve almayanların olasılık dağılımı ifade edilmektedir. Olasılık dağılımlarının birbirine oranının doğal logaritması bize Kanıtsal Ağırlık değerini vermektedir ve bu değer (6) numaralı denklemde gösterildiği gibi Bilgi Değeri hesaplanırken kullanılmaktadır. Ürünü satın alanların dağılımı satın almayanların dağılımına eşitse olasılık dağılımlarının oranı 1'e eşit olacak ve bunun doğal logaritmadaki karşılığı sıfır olacaktır. Satın alan ve almayanların dağılımının birbirinden ne kadar ayrışık olduğunu anlayabilmek için olasılık dağılımlarının birbirinden o kadar farklı olması beklenmektedir. Böylece iki kümenin birbirinden farklı bilgi taşıdığı ve ayrışık olduğu kanaatine varılabilir. Olasılık dağılımlarının birbirine eşit olması, maksimum belirsizliğe işaret eder, Kanıtsal Ağırlığı 0'a yakınsatır, Bilgi Değerini azaltır [11,13].

$$\boldsymbol{Kanıtsal\ Ağırlık} = \boldsymbol{ln}\left(\frac{(\boldsymbol{Kredi\ Kartı\ Alanların\ Dağılımı})_i}{(\boldsymbol{Kredi\ Kartı\ Almayanların\ Dağılımı})_i}\right) \quad (5)$$

$$Bilgi\ Değ. = \sum((K.KartıAlanların\ Dağ.)_i - (K.Kartı\ Almayanların\ Dağ.)_i) * Kanıtsal\ Ağırlık \quad (6)$$

Veri kümesinde bilgi değeri yüksek çıkan özelliklerin tahminleyici gücü yüksektir. Bilgi değeri yöntemi, (6) numaralı denklemden çıkan sonuçları (BD<0,02) tahminleyici gücü yok, (0,02<BD<0,1) zayıf, (0,1<BD<0,3) orta, (0,3<BD<0,5) kuvvetli ve (0,5<BD) şüpheli derecede kuvvetli olacak şekilde etiketlendirmektedir[12]. Uygulamalı literatür [11, 12, 13], tahminleyici gücü orta ve kuvvetli olarak etiketlenen özelliklerin regresyon ve sınıflama modelleri için tercih edildiğini belirtmektedir.

2.1.4 Ki-kare Özellik Seçimi Yöntemi

Ki-kare testi (X^2) iki değişken arasındaki ilişkinin bağımlı veya bağımsız olduğunu belirlemeye yarayan ayrık veriler için kullanılan bir hipotez test yöntemidir. Ki-kare istatistiğine dayalı özellik seçimi metodu iki adımı içermektedir. Yöntemin ilk kısmında özelliklerin sınıflara göre ki-kare istatistikleri hesaplanır. İkinci kısımdaysa serbestlik derecesi ve belirlenen önemlilik seviyesine göre ki-kaynaşımı prensibiyle ki-kare değerlerine bakılarak veri seti içerisindeki tutarsız özellikler bulunana kadar art arda özelliklerin ayrıştırılması gerçekleştirilir[6]. Veri kümesi içinde yer alan bir özellik için hesaplanan ki-kare değeri, o özelliğin sınıf içerisindeki bağımlılığını ölçmektedir. Sıfır değerine sahip bir özellik o küme içinde bağımsız olduğunu gösterir. Yüksek bir ki-kare değerine sahip olan özellik, veri kümesi için daha tanımlayıcıdır. Ki-kare değerinin hesaplanmasında kullanılan genel eşitlikler aşağıda verilmektedir[6].

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (7)$$

$$E_{ij} = \frac{(R_i * C_j)}{N} \quad (8)$$

(7) numaralı denklemdeki eşiklikte *k* veri kümesindeki sınıf sayısı, A_{ij} gözlenen frekans değeri (*i* satır, *j* sütün) ve E_{ij} ise *Aij'*nin beklenen (teorik) frekans değeridir. (8) numaralı denklemde R_i *i*'inci aralıktaki veri sayısı, C_j *j*'inci sınıftaki gözlemlerin sayısı, *N* sınıflardaki gözlemlerin toplamını ifade etmektedir. Bu yöntem nümerik değerler için aralık belirleme kapsamında kullanılmaktadır.

3.Veri Kümesi Tanıtımı

Bu uygulamada Şekerbank'a ait örnek veri kümesi kullanılmıştır. Veri kümesi 6 özellik, 1658 veri ve 1 sınıf etiketinden oluşmaktadır. Çalışmada kullanılan veri kümesi, Şekerkart Visa kart ürüne ait ürün sahipliğinin ve bu ürüne başvuruda bulunan müşterilerin riskliliğini ölçen banka skoru, Kredi Kayıt Bürosu (KKB) skoru,

müşterilerin aylık geliri, eğitim düzeyi, faaliyet gösterdikleri sektör ve cinsiyet özelliklerini kapsamaktadır.

Sınıf etiketi 2 farklı kategorik değerle ifade edilmektedir. Örnek veride sınıf etiketi ürün sahipliğiyle temsil edilmektedir. Bunun için 2014 yılı içinde Şekerkart Visa Kart kredi kartı ürünü için yapılmış tüm başvurular dikkate alınmıştır. Başvurusu kabul edilen müşteri artık ürüne sahip olarak değerlendirilmiş ve veri kümesinde "1" ile işaretlenmiş, başvurusu kabul edilmeyen müşterininse ürün sahibi olmadığı kabul edilmiş ve bu durum veri kümesinde "0" ile işaretlenmiştir.

Özellikler şu şekildedir:

- Banka Skoru: Müşterilerin risklilik seviyelerini belirlemede, bankanın içsel derecelendirmeye dayalı yaklaşımlar kullanarak hesapladığı kredi notu olup başvuru yapan kişinin veya kuruluşun kredi alma yeterliliğini ölçer. Nümerik bir değerdir. Test veri kümesinde [60-943] arasında değerler almaktadır.
- KKB Skoru: Türkiye'deki tüm kredi kurumlarındaki kişisel pozisyonu, güvenilirliği ve müşteri olma potansiyelini gösteren nümerik bir değerdir. Test veri kümesinde [276-1.576] arasında değerler alır.
- Aylık Gelir: Müşterilerin aylık gelirlerinin bulunduğu özelliktir. Nümerik bir değerdir. Test veri kümesinde [846,00TL-200.000,00TL] arasında değerler almaktadır.
- Cinsiyet: Müşterilerin cinsiyet bilgisinin bulunduğu özelliktir. Kategorik bir özelliktir. Erkek veya kadın olabilir.
- Eğitim Durumu: Müşterilerin eğitim durumunun tutulduğu kategorik bir özelliktir. Test veri kümesinde 8 farklı eğitim derecesine sahip müşteri bulunmaktadır. 8 en yüksek, 1 en düşük eğitim seviyesini göstermek üzere kategoriler sıralanarak ayrıntıları Tablo 3'te sunulmuştur. Eğitim kategorilerinde belirtilmesi gereken husus, kavramsal olarak ilköğretim kategorisindeki müşterilerin ortaokul kategorisine taşınarak ilköğretim kategorisinin çalışma dışında bırakılabileceğidir. Ancak Şekerbank müşteri veri tabanında her iki eğitim kategorisi de ayrı ayrı yer aldığından çalışma kapsamında veri kümesinin orijinaline sadık kalınmasına karar verilmiştir.
- Sektör Adı: Müşterinin çalıştığı sektörün kod bilgisinin tutulduğu kategorik bir özelliktir. Test veri kümesinde 40 farklı sektör bulunmaktadır.

Veri kümemizde nümerik değer alan özelliklerimiz Banka Skoru, KKB Skoru ve Aylık Gelir'dir. Bu özelliklerin veri kümesi içinde aldıkları minimum, maksimum değerlerle ortalaması ve standart sapması Tablo 1'de gösterilmektedir. Tablo 1'e göre Aylık Gelir standart sapması çok yüksek dengesiz dağılım gösteren bir özelliktir ve aykırı değer barındırmaktadır. Bu aykırı değerler nümerik özelliklerin aralıklara bölünmesi için uygulanan ki-kaynaşımı algoritmasıyla dengeli hale getirilmiştir.

Tablo 1: Nümerik özelliklerin istatistiksel bilgileri

	Banka Skoru	**KKB Skoru**	**Aylık Gelir (TL)**
Minimum Değer	60	276	846
Maksimum Değer	943	1.576	200.000
Ortalama	673,41	1.165,96	6.925,13
Standart Sapma	94,99	222,15	12.921,81

Tablo 2'de veri kümesindeki erkek ve kadın müşteri dağılımları gösterilmektedir.

Tablo 2: Cinsiyet özelliği için verinin dağılımı

Kategorik Sıralama	**1**	**0**	
Cinsiyet	Erkek	Kadın	Toplam
Müşteri Sayısı	1.248	409	1.657
% Dağılım	0,75	0,25	100

Tablo 3'te Eğitim Durumu özelliğinin 8 farklı dalı gösterilmektedir. Bu dalların veri kümesindeki dağılımları yüzde olarak verilmiştir. Veri kümemizin %50'sinden fazlasını lise mezunu müşteriler oluşturmaktadır. Buna göre veri kümemiz dengeli bir dağılım göstermemektedir. Benzer durum sektör özelliğinde de söz konusudur; ancak banka uygulamasında müşterilerin faaliyet gösterdiği 40 farklı sektör olduğu için burada tablo özeti kullanılmamıştır.

Tablo 3: Eğitim Durumu özelliği için verinin dağılımı

Kategorik sıralama	8	7	6	5	4	3	2	1	
Eğitim Durumu	Doktora	Lisansüstü	Lisans	Lise	İlköğretim	Ortaokul	İlkokul	Okul Bitirmemiş	Toplam
Müşteri Sayısı	6	51	459	916	140	55	29	1	1.657
%Dağılım	0,4	3,1	27,7	55,2	8,4	3,3	1,8	0,10	100

4. Gerçekleştirilen Testler ve Niceliksel Sonuçları

Bu çalışma için geliştirilmiş olan FetBet[1] yazılımının yapacağı özellik seçimini doğrulamak amacıyla, aynı veri kümesine hem FetBet yazılımıyla hem de Weka ve R

[1] FetBet yazılımının geliştirilmesindeki temel amaç teorik olarak bilinen veri madenciliği özellik seçim yöntemlerinden birkaçını uygulamaya geçirmektir. FetBet yazılımı geliştirilirken entropi tabanlı özellik seçim yöntemleri tercih edilmiş ve nümerik verilerin aralıklandırılması konusunda karşılaşılan zorluk Ki-kaynaşımı yönteminin FetBet yazılımında kullanılması suretiyle aşılmıştır.

programlarıyla Bilgi Kazancı, Kazanım Oranı ve Bilgi Değeri yöntemleri uygulanmış ve elde edilen sonuçlar karşılaştırılmıştır. Aşağıda tüm yöntemlerin üç farklı uygulamadaki çıktıları ve elde edilen sonuçlara dair yorumlar paylaşılmaktadır.

Test veri kümesi için Bilgi Kazancı yöntemi bahsi geçen uygulamalarda icra edilerek çıktılar karşılaştırmalı olarak Tablo 4'te sunulmuştur.

Tablo 4: Bilgi Kazancı Yöntemine Göre Weka, R ve FetBet Yazılımlarının Karşılaştırılması

	Weka			**FetBet**			**R**		
	Bilgi Kazancı Ağırlıkları		Kümülatif Ağırlık	Bilgi Kazancı Ağırlıkları		Kümülatif Ağırlık	Bilgi Kazancı Ağırlıkları		Kümülatif Ağırlık
BİLGİ KAZANCI	Banka Skoru	0,163	**0,56**	Banka Skoru	0,183	**0,47**	Banka Skoru	0,113	**0,56**
	KKB Skoru	0,081	**0,84**	KKB Skoru	0,129	**0,81**	KKB Skoru	0,056	**0,84**
	Sektör	0,024	**0,92**	Aylık Gelir	0,037	**0,90**	Sektör	0,016	**0,92**
	Eğitim Durumu	0,012	**0,96**	Sektör	0,024	**0,96**	Eğitim Durumu	0,008	**0,97**
	Aylık Gelir	0,010	**0,99**	Eğitim Durumu	0,013	**0,99**	Aylık Gelir	0,007	**1,00**
	Cinsiyet	0,001	**1,00**	Cinsiyet	0,002	**1,00**	Cinsiyet	0,001	**1,00**
	Toplam	**0,291**		**Toplam**	**0,388**		**Toplam**	**0,201**	

Veri kümesini iyi tahminleyen özellik sıralaması incelendiğinde Weka ve FetBet uygulamalarının Aylık Gelir özelliği dışında benzer sıralama yaptığı görülmektedir. Aylık Gelir özelliğinin uygulamalarda farklı sıralanmasının sebebi, nümerik özelliklerin aralıklara ayrılması için kullanılan yöntemlerin uygulama bazlı farklılık göstermesidir. FetBet yazılımı Ki-kaynaşımı yöntemini kullanırken, Weka uygulaması bir aralıktaki bölümlemeyi belirlemede Minimum Betimleme Uzunluk Esası'nı dikkate alan deneyimsel bir yöntemi izlemektedir [7].

Kümülatif ağırlık %95 seçildiğinde Banka Skoru, KKB Skoru özellikleri FetBet, Weka ve R sonuçlarına göre ilk iki sırada, Aylık Gelir ise diğerlerinden farklı olarak FetBet uygulama sonucunda üçüncü sırada yer almaktadır. Kategorik verilerin bilgi kazancı değerlerinin üç yazılımda da birbirine çok yakın çıktığı gözlenmektedir. Bunun sebebi kategorik veriler için aralıklandırmaya ihtiyaç duyulmamasıdır.

Bilgi Kazancı yöntemi için kullanılan test veri kümesi, Kazanım Oranı yöntemi için de kullanılmış; Weka, FetBet ve R uygulamalarından elde edilen sonuçlar karşılaştırmalı olarak Tablo 5'te sunulmuştur.

Tablo 5: Kazanım Oranı Yöntemine Göre Weka, R ve FetBet Yazılımlarının Karşılaştırılması

KAZANIM ORANI	Weka			FetBet			R		
	Kazanım Oranı Ağırlıkları		Kümülatif Ağırlık	Kazanım Oranı Ağırlıkları		Kümülatif Ağırlık	Kazanım Oranı Ağırlıkları		Kümülatif Ağırlık
	Banka Skoru	0,125	**0,56**	Banka Skoru	0,044	**0,49**	Banka Skoru	0,125	**0,56**
	KKB Skoru	0,06	**0,83**	KKB Skoru	0,023	**0,74**	KKB Skoru	0,06	**0,83**
	Aylık Gelir	0,023	**0,94**	Aylık Gelir	0,008	**0,83**	Aylık Gelir	0,023	**0,94**
	Eğitim Durumu	0,007	**0,97**	Eğitim Durumu	0,007	**0,91**	Eğitim Durumu	0,007	**0,97**
	Sektör	0,006	**0,99**	Sektör	0,006	**0,97**	Sektör	0,006	**0,99**
	Cinsiyet	0,002	**1,00**	Cinsiyet	0,002	**1,00**	Cinsiyet	0,002	**1,00**
	Toplam	**0,223**		**Toplam**	**0,09**		**Toplam**	**0,223**	

Kazanım oranı yönteminin sonucuna göre en iyi tahminleyici ilk üç özelliğe bakıldığında tüm yöntemlerde sıralamanın Banka Skoru, KKB Skoru ve Aylık Gelir olduğu görülmektedir[2]. Alan bilgisinden kaynaklanan beklentimiz de bir müşterinin, kredi kartı alıp almama durumunu etkileyecek en önemli özelliklerin Banka Skoru, KKB Skoru ve borç ödeme gücünü gösteren Aylık Gelir olacağı yönündedir[3]. Yöntemden elde edilen sonuç beklentimizle örtüşmektedir.

Bilgi değeri yöntemi temel alınıp bu yönteme Ki-kaynaşımı algoritmasının uygulanmasıyla elde edilen FetBet sonuçları Tablo 6'te sunulmuştur.

[2] Kazanım Oranı ve Bilgi Kazancı yöntemlerinin sonuçları kümülatif ağırlıklar açısından ele alındığında FetBet yazılımının Weka ve R uygulamasından farklı sonuçlar verdiği Tablo 5'te görülmektedir. Bu farklılığın FetBet yazılımına Ki-kaynaşımı yönteminin uygulanmasından ve Kazanım Oranı yönteminin aralıklandırmadaki değişikliğe olan hassasiyetinden kaynaklandığı düşünülmektedir. Ki-kaynaşımının Banka Skoru, KKB skoru ve Aylık Gelir özelliklerine uygulanmasıyla kazanım oranı ağırlıklarının Weka ve R'a göre azaldığı ancak kümülatif ağırlıklar açısından sıralamanın değişmediği Tablo 5'te gözükmektedir. Eğitim durumu, Sektör ve Cinsiyet gibi kategorik özelikler için aralıklandırma gerekmediğinden Kazanım Oranı ağırlığı Weka, FetBet ve R'da aynı sonucu vermiştir.

[3] Bu çalışmada özellik seçim yöntemleri uygulamasına konu olan değişkenler yurt içi piyasada perakende bankacılıkta faaliyet gösteren Şekerbank A.Ş.'ne aittir ve rastgele seçilmiş örnek bir ürüne ait müşteri bazında ulaşılabilen özelliklerle sınırlıdır. Her ürün için o ürünü satın alma ve almama davranışını belirleyen özellikler belli ölçüde hem müşteri hem de ürünün özelliklerine göre farklılık gösterebilir. Analizimiz banka tarafından kullanımına izin verilen özellikler üzerinden gerçekleştirilmiştir.

Tablo 6: Bilgi Değeri Yöntemine Göre Özelliklerin Tahminleyici Gücü

BİLGİ DEĞERİ	FetBet					
	Bilgi Değeri Ağırlıkları		Tahmin Gücü Etiketi	Bilgi Değeri Ağırlıkları		Tahmin Gücü Etiketi
	Banka Skoru	**0,913**	Şüpheli derecede kuvvetli[4]	Sektör	**0,123**	Orta
	KKB Skoru	**0,65**	Şüpheli derecede kuvvetli[5]	Eğitim Durumu	**0,066**	Zayıf
	Aylık Gelir	**0,141**	Orta	Cinsiyet	**0,007**	Zayıf

Bilgi Değeri, özelliklerin tahmin gücünü bulurken, özellik bazında ikili her durum için Kanıtsal Ağırlık hesaplamaktadır. Kanıtsal Ağırlık hesaplamasıyla veri kümesinin dengesi de ölçülmektedir. Bilgi Kazancı ve Kazanım Oranı yöntemlerinin sonuçlarına bakıldığında Banka Skoru ve KKB Skoru en önemli özellikler olarak tespit edilmiş; Bilgi Değeri yöntemiyle de bunu destekler bir sonuca ulaşılmıştır. Bilgi Değeri yöntemi aylık gelir ve sektör özelliklerini orta kuvvette tahminleyici olarak bulurken eğitim durumu ve cinsiyet özelliklerini zayıf tahminleyici olarak bulmuştur.

Özellik seçiminde bu çalışmaya konu olan üç yöntem, Weka, R ve geliştirilen FetBet uygulamalarında aynı bankacılık veri kümesi kullanılarak araştırılmıştır. Özelliklerin tahmin gücü, araştırmaya konu olan yöntemlerin farklı uygulamalar kullanıldığında aynı sıralamayı verip vermedikleriyle test edilmiştir. Sonuçlar bu açıdan ele alındığında, Bilgi Kazancı yöntemi Weka ve R uygulamalarında farklı sıralama gösterirken Ki-kaynaşımının kullanıldığı FetBet algoritmasında alan bilgisine dayalı beklentimizi karşılayan sonucu verdiği gözlenmiştir. Bilgi Kazancı yönteminin sapma yarattığı bilgisiyle Kazanım Oranı yönteminin bu sapmayı düzelten bir yapısı olduğu dikkate alındığında, FetBet uygulamasıyla ulaşılan sonucun Weka ve R'dan elde edilen Kazanım Oranı sonucuyla aynı sıralamaya ulaşmış olması, geliştirilen uygulamanın doğru çalıştığını göstermektedir. Bilgi Kazancı ve Kazanım Oranı yöntemleri karşılaştırıldığında Kazanım Oranı yönteminden gelen sıralamanın beklentimizle örtüşmesi nedeniyle sezgisel olarak da Kazanım Oranı yönteminin daha iyi sonuçlar ürettiğini düşünülmektedir.

---

[4] "Şüpheli derecede kuvvetli" etiketi, Bilgi Değeri yönteminin uygulamalı literatürde karşımıza çıkan bir sınıflama etiketidir[8]. Özellikle bankacılık sektörüne ait ikili durumu açıklayan özelliklerin belirlenmesinde kullanılmaktadır. Örneğin banka kredisi kullanan müşterilerin kredi borcunu zamanında ve tam olarak ödeyip ödeyememe durumuna göre sınıflandırıldığı iyi – kötü ayırımında kullanılan özelliklerin tespitinde dikkate alınır. Bilgi değeri 0,5'in üzerinde hesaplanan özellikler için bu etiket tahmin gücü "kuvvetli"den sonra gelmektedir. "Şüpheli derecede kuvvetli" olan özelliklere, sınıflama veya regresyon modeline dâhil edilme aşamasında özel dikkat gerektiği literatürde not düşülmektedir.

[5] FetBet yazılımıyla uygulanan Bilgi Değeri yönteminin sağlaması Excel'de yapılmıştır. Weka'da, Bilgi Değerini hesaplayan bir yazılıma ulaşılamadığından ve R'daki yazılımın da sonuçları istediğimiz sınıflamaya oturtmadığı gözlendiğinden karşılaştırma analizi Excel'de yapılmıştır. Buna göre Banka Skoru FetBet yazılımına paralel olarak "Şüpheli derecede kuvvetli" bulunmuştur. KKB skoruysa "Kuvvetli" olarak etiketlenmiştir. Bu farkın KKB skorunun herhangi bir müdahaleye açık olmayan dışsal bir veri olmasından kaynaklandığı düşünülmektedir. Banka Skoru ise içsel hesaplanmakta ve banka tarafından manipülasyona konu olabilmektedir.

Bilgi Kazancı ve Kazanım Oranı yöntemlerine girdi olarak verilen veri kümesi, FetBet uygulamasına dahil edilen Bilgi Değeri yöntemine de uygulandığında; Banka Skoru, KKB skoru ve Aylık Gelir özellikleri tahmin gücü şüpheli derecede kuvvetli ve orta olarak etiketlenmiştir. Sektör, eğitim ve cinsiyet gibi kategorik nitelik gösteren özelliklerse orta ve zayıf olarak etiketlenmiştir. Bu durum özelliklerin taşıdığı Kanıtsal Ağırlığın zayıflığına, yani bu özelliklerin taşıdığı belirsizliğin yüksek olmasına bağlanabilir. FetBet uygulamasından elde edilen Bilgi Değeri sonuçları sıralama açısından ele alındığında ilk üç sıradaki özelliğin değişmediği ve genel olarak FetBet uygulamasının Kazanım Oranı yöntemiyle benzer sonuçlar verdiği gözlenmektedir.

Uygulanan Yöntemlerin sonuçları birbirlerini destekler niteliktedir ancak bu çalışma özellikle aralarında nedensel bir ilişki olmadan ya da oransallık içermeyen aralıklandırılmış kategorik değişkenler için geliştirilebilir. Örneğin, sektör isimleri rastgele numaralandırılmıştır ve rakamlar arasında artan ya da azalan bir ilişkinin sektörler arası geçişle anlamlı bir ilişkisi yoktur. Bu durumun Kanıtsal Ağırlığı azaltan ve belirsizliği artıran etki gösterdiği düşünülmektedir. Bilgi değeri yöntemine göre Sektör orta kuvvette bir özellik olarak sonuç verse de sektör özelinde bu sonucun güvenilirliği tartışılmalıdır. Benzer bir şekilde eğitim durumu 0 ile 8 arasında derecelendirilmiştir; ancak bu kategorik değişkende her 1 puanlık artış eğitim düzeyindeki artışla doğrusal orantılı değildir. Özellik seçimi yöntemiyle tahmin gücünü tespit ettiğimiz kategorik değişkenlerin aralıklandırılması için geliştirilebilecek bir yazılım, regresyon analizi aşamasında katsayı tahmininde kolaylık sağlayacaktır.

Kaynaklar

1.Genç, H., Cataltepe, Z., and Pearson, T., “A New PCA/ICA Based Feature Selection Method”, In: IEEE 15th In Signal Processing and Communications Applications, pp 1-4 (2007).

2.Gülgezen, G., Cataltepe, Z., and Yu, L., “Stable Feature Selection Using MRMR Algorithm / MRMR Algoritması Kullanılarak Kararlı Özellik Seçimi”, 17th IEEE Conference on Signal Processing and Communication Applications (SIU 2009), Antalya, Turkey.

3.Çetişli, B., “Using Neuro-Fuzzy Classifier With Lingustic Hedges For Feature Selection”, Eskişehir Osmangazi Üniversitesi Müh.Mim.Fak.Dergisi (2006).

4.Akman, M., "Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi Sağlık Alanında Bir Uygulama "Yüksek Lisans Tezi, Ankara Üniversitesi, (2010).

5.Şaylan, Ç., “Böbrek Nakli Geçirmiş Hastalarda Akıllı Yöntem Tabanlı Yeni Öznitelik Seçme Algoritması Geliştirilmesi”, Yüksek Lisans Tezi, Kadir Has Üniversitesi Yönetim Bilişim Sistemleri Yüksek Lisans Programı, İstanbul (2013).

6.Kavzoğlu, T., “Heyelan Duyarlılık Analizinde Ki-Kare Testine Dayalı Faktör Seçimi”, V. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu, 14-17 Ekim 2014, İstanbul.

7.Fayyad, U., Irani, K., “Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning”, Proceedings of the 13thInternational Joint Conference on Artificial Intelligence, Chambery, France, 1993, 1022-1027.

8.Upadhyay, R., “Information Value and Weight of Evidence – A Case Study from Banking”, http://ucanalytics.com/,10 Mayıs 2015.

9.Ramanathan, R,. “Introduction to Economics with Applications”, Dryden Press (1992).

10.Han, J. And Kamber, M., Data Mining, Second Edition Concepts and Techniques 2nd Edition ISBN: 978-1-55860-901-3 The Morgan Kaufmann Series, (2006).

11.Lin, A. Z. “Variable Reduction in SAS by Using Weight of Evidence and Information Value”, SAS Global Forum 2013 Proceedings, Paper 095-213.

12.Siddiqi, N. (2006). Credit Risk Scorecards, Hoboken, NJ: John Wiley & Sons, Inc.

13.Lund B. and Brotherton D. (2013). “Information Value Statistic”, MWSUG 2013, Proceedings, Midwest SAS Users Group, Inc., paper AA-14.