T.C.
ERCİYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

TÜİK YAŞAM MEMNUNİYETİ ANKETLERİ ÜZERİNE VERİ MADENCİLİĞİ UYGULAMASI

Hazırlayan
Serkan DEMİRCAN

Danışman
Doç. Dr. Sinem KULLUK

Yüksek Lisans Tezi

Haziran 2015
KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Adı-Soyadı : Serkan DEMİRCAN

İmza : 

YÖNERGEYE UYGUNLUK

"TÜİK YAŞAM MEMNUNİYETİ ANKETLERİ ÜZERİNE VERİ MADENCİLİĞİ UYGULAMASI" adlı Yüksek Lisans / Doktora tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi'ne uygun olarak hazırlanmıştır.

Tezi Hazırlayan

Serkan DEMİRCAN

Danışman

Doç. Dr. Sinem KULLUK

Endüstri Mühendisliği ABD Başkanı

Doç. Dr. İbrahim DOĞAN

Doç. Dr. Sinem KULLUK danışmanlığında **Serkan DEMİRCAN** tarafından hazırlanan **"TÜİK Yaşam Memnuniyeti Anketleri Üzerine Veri Madenciliği Uygulaması "** adlı bu çalışma, jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalında **Yüksek Lisans** tezi olarak kabul edilmiştir.

17 /06 / 2015

JÜRİ:

Danışman : Doç. Dr. Sinem KULLUK

Üye : Doç. Dr. Lale ÖZBAKIR

Üye : Yrd. Doç. Dr. Kumru UYAR

ONAY:
Bu tezin kabulü Enstitü Yönetim Kurulunun 23/06/2015 tarih ve 2015/25-07 sayılı kararı ile onaylanmıştır.

23 / 06 / 2015

Prof. Dr. Kâzım KEŞLİOĞLU

Enstitü Müdürü

TÜİK YAŞAM MEMNUNİYETİ ANKETLERİ ÜZERİNE VERİ MADENCİLİĞİ UYGULAMASI

Serkan DEMİRCAN
Erciyes Üniversitesi, Fen Bilimleri Enstitüsü
Yüksek Lisans Tezi, Haziran 2015
Tez Danışmanı: Doç. Dr. Sinem KULLUK

ÖZET

Mutluluk ve yaşam memnuniyeti kavramları, dünya genelinde felsefe kapsamında çok eskiden bu yana, çeşitli boyutlarıyla ele alınan bir konudur. Mutluluk kavramı, yıllar içinde bireylerin maddi veya manevi değerleriyle kendi içinde duyduğu öznel beğeniden, toplumun takdirini kazanacak değerlere sahip olmasına, bunların çeşitli oranlardaki bileşimine, sürdüğü yaşam biçiminden hoşnutluğa kadar çeşitli görünümleriyle ortaya konulmuştur. Oysa günümüzde algılanan mutluluk, bireyin kendi içindeki, kendine özgü değerlendirmenin bir çıktısı olarak kendisini mutlu olarak hissedip hissetmemesidir.

Veri toplama yöntemiyle mutluluk düzeyi ölçümü, dünyada 1940'ların ikinci yarısında başlamıştır. Bu alandaki teorik ve uygulamalı çalışmalar "World Data Base of Happiness" veritabanında yer almaktadır. Ayrıca, yaşam memnuniyeti ile ilgili "Journal of Happiness Studies" adında bir süreli yayın da mevcuttur. Resmi istatistik olarak Türkiye'de yapılan, mutluluk konusunu ele alan ilk araştırma, Türkiye İstatistik Kurumu (TÜİK) tarafından "Yaşam Memnuniyeti Araştırması" (YMA) adıyla 2003 yılında gerçekleştirilmiştir. YMA, TÜİK' in toplumsal içerikli ve aynı zamanda öznel öğeler içeren ilk araştırması olma özelliğini taşımaktadır. Araştırmada bireylerin mutluluk, umut, temel yaşam alanlarındaki genel memnuniyeti ve bu alanlardaki kamu hizmetlerinden memnuniyeti ölçülmektedir.

Bu tez kapsamında, TÜİK tarafından 2012 yılında fert ve hane bazında yapılan yaşam memnuniyeti anketi ham verileri üzerinde sınıflandırma uygulaması Weka yazılımı kullanılarak gerçekleştirilmiştir. Öncelikle, anketler üzerinde hedef değişken(ler) ve bu değişkenleri etkileyebilecek girdi değişkenleri belirlenmiştir. Daha sonra Veri

Madenciliği (VM) çalışmalarının zamansal olarak en büyük kısmını kapsayan, veri bütünleştirme, kayıp değerlerin doldurulması, tutarsız ve aşırı uç değerlerin temizlenmesi gibi veri önişleme işlemleri gerçekleştirilmiştir. Elde edilen veri kümeleri ile sınıflandırma algoritmaları çalıştırılarak veri kümelerini ifade eden sınıflandırma kuralları belirlenmiştir. Ayrıca, elde edilen sınıflandırma doğruluklarını artırmak ve hedef değişkenini etkilemeyen girdi değişkenlerini elimine etmek için nitelik indirgeme işlemi gerçekleştirilerek, oluşturulan yeni veri kümeleri ile tekrar sınıflandırma yapılmıştır. Sınıflandırma çalışmaları kapsamında 5 sınıflandırma algoritması ele alınmış, uygulama sonucunda "EĞER … O HALDE …" şeklinde kural kümeleri oluşturulmuş ve sonrasında ise algoritmaların doğruluk düzeyleri ve performansı kıyaslanmıştır.

Sınıflandırma işlemleri sonucunda ortalama % 70 doğruluk değerlerine ulaşılmıştır. Elde edilen sonuçlar düşük gibi görünse de gerçek hayat uygulamaları için kabul edilebilir değerlerdir ve elde edilen kurallar yeni örneklerin sınıflandırılması amacıyla kullanılabilir.

Anahtar kelimeler: Veri Madenciliği, Sınıflandırma, TÜİK, Yaşam Memnuniyeti

THE APPLICATION OF DATA MINING UPON THE QUESTIONAIRES OF TÜİK ON LIFE SATISFACTION

Serkan DEMİRCAN
University of Erciyes, The Institute of Science
Master's Thesis, June 2015
Thesis Supervisor: Assoc. Prof. Dr. Sinem KULLUK

ABSTRACT

Since the very ancient times, the concept of happiness and life satisfaction have been issues which have been taken up in utter detail. Throughout the years, the concept of happiness has been argued in various forms with the economic and spiritual values of the individuals, from the subjective credit of the individuals' economic and spiritual values in himself to the possession of the values appreciated by the society,the combination of those in various proportion and the pleasure of the lifestyle he experiences. However, happiness regarded today in individual himself is whether to feel happy or not.

The level of happiness estimate with the data collection method started in the second half of the 1940s in the world. The theorotical and practical studies in this field takes place 'World Data Base'. Furthermore, a periodical journal called 'Journal of Happiness Studies' related to life pleasure is also available. As the official studies conducted in Turkey, the first research handling the subject of happiness, was carried out by Turkish Statistic Institution (TÜİK) named 'Life Satisfaction Research' (YMA) in the year of 2003. The first researh concerning the subject of happiness,YMA, is the first research of TÜİK regarding the social and also subjective and objective content. In the research, the individuals' happiness, expectations, the overall pleasure in the fundemental public utilities in such fields are evaluated.

Within the framework of this thesis , the life pleasure questionaire conducted by TÜİK in 2012, the classification application on raw data was fulfilled by Weka Software. Primarily, the objective and the input variables that can influence such variables were

estimated. Later on, data unification that chronologically include a great deal of Data Mining (VM), the refilling of lost values, and the purifying of such extreme values were realized.

Having studied the algorithm classification with the gathered data sets, the classification rules representing the data sets were determined. In addition, in order to increase the accuracy of the collected data sets and to eliminate the input variables that didn't impact the objective variable by the realization of the quality degradation process, reclasification was made with the newly formed data sets. With regard to the classification studies, five classification algorithms were handled, and as a consequence of that, condition sets; 'if…, in that case….' were formed and following that, the accuracy levels and the performance of algorithms were compared.

As a result of the classification process, roughly 70 percent of accuracy was reached. Even though the obtained results seem to be low, they are reasonable values regarding the real life applications and also with the aim of new samples' classification, such rules can be used.

Key words: Data Mining, Classification, TÜİK, Life Satisfaction

TEŞEKKÜR

"TÜİK 2012 Yaşam Memnuniyeti Anketi Üzeri Veri Madenciliği Çalışması" adlı tezin düşünce aşamasından sunumun gerçekleşmesi anına kadar her türlü konuda yardımlarını benden esirgemeyen ve faydalı bir çalışma oluşması amacıyla her aşamada titizlilikle kontroller yapan değerli hocam Doç. Dr. Sinem KULLUK'a teşekkürlerimi iletiyorum.

Ayrıca uzman görüşlerinden ve anket üzerinde anlaşılmayan noktalara getirdiği açıklamalarından dolayı TÜİK Gaziantep Bölge Müdürlüğü Sosyal Sektör İstatistikleri Takım Sorumlusu Sayın Vesile Gülsüm Bozkır'a, yabancı dil konusundaki yardımlarından dolayı Gaziantep Üniversitesi Yabancı Diller Yüksek Okulu Temel İngilizce Hazırlık bölümünde görevli Sayın Zeynep Deniz Durdu'ya ve tüm tez dönemi boyunca gerek motivasyon gerekse de diğer konularda benden desteğini esirgemeyen çok değerli aileme teşekkürü borç bilirim.

Serkan
DEMİRCAN

İÇİNDEKİLER

TÜİK 2012 YILI YAŞAM MEMNUNİYETİ ANKETLERİ ÜZERİNE VERİ MADENCİLİĞİ UYGULAMASI

2. BÖLÜM

SINIFLANDIRMA

3. BÖLÜM
GEREÇ ve YÖNTEM

4. BÖLÜM

5. BÖLÜM

KISALTMALAR

vd.	ve diğerleri
VM	Veri Madenciliği
TÜİK	Türkiye İstatistik Kurumu
YMA	Yaşam Memnuniyeti Anketi
VTBK	Veri Tabanlarından Bilgi Keşfi
KDD	Bilgi Keşfi ve Veri Madenciliği (Knowledge Discovery and Data mining)
İMKB	İstanbul Menkul Kıymetler Borsası
A.Ş.	Anonim Şirket
CRM	Müşteri İlişkileri Yönetimi (Customer Relations Management)
C&RT	Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees)
LSD	Küçük Kareli Sapma (Little Squared Daviation)
RIPPER	Hata Azaltma Amaçlı Tekrarlanan Artımlı Budama (Repeated Incremental Pruning to Produce Error Reduction)
SIGKDD	Bilgi Keşfi ve Veri Madenciliği Özel İlgi Grubu (Special Interest Group on Knowledge Discovery and Data Mining)

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Teknolojinin gelişmesiyle birlikte, insan ihtiyaçları da artarak kendini sürekli yenilemiş; ekonomik, sosyal, kültürel, bilimsel alanlar ve teknoloji alanlarında geniş bir yelpazede rekabet piyasası oluşturmuştur. Bu piyasanın ortaya çıkması hem devletlerarası, hem de devlet içi kalkınmanın giderek büyümesine neden olmuştur. Ancak bu kalkınmayı sağlarken bazı bilimsel yöntem ve bilgi teknolojilerine ilişkin gelişmeleri ve uluslararası göstergeleri takip etmek, ulusal ve uluslararası bilgi ağı ve bilgi akış sisteminin oluşturulmasını koordine etmek ve diğer ülkeler ve uluslararası kuruluşlarla işbirliğini sağlamak gerekmektedir. Bu durum ise ülkelerdeki ekonomi, sosyal, demografi, kültür, çevre, bilim ve teknoloji alanları ile gerekli görülen diğer alanlardaki geçmişe dayalı verileri toplama ve depolama ihtiyacını beraberinde getirmiştir. 1926 yılında bu amaçla, ilk adı Merkezi İstatistik Dairesi olan günümüzdeki adı ile Türkiye İstatistik Kurumu (TÜİK) kurulmuştur.

TÜİK bütün bu gelişmeler neticesinde yaşamını sürdüren insanoğlunun toplumsal yapısında meydana gelen gelişmeleri takip ederek geleceğe ışık tutmak amacıyla Yaşam Memnuniyeti Araştırması (YMA) yapmaya karar vermiştir. Yaşam Memnuniyeti Araştırması'nın ilki, 2003 yılından itibaren Hanehalkı Bütçe Anketi'ne ek bir modül olarak gerçekleştirilmiştir. Araştırma 2003 yılından itibaren her yıl düzenli olarak yapılmaktadır.

TÜİK 2012 yılında Türkiye'de yaklaşık 4000 hane ve 8000 fert üzerinde yaşam memnuniyeti anketini uygulamıştır. Bu anketlerde gizli olan bilgileri kullanıcıların anlayabileceği dilsel kurallar şeklinde ifade edebilmek ve gelecekte bu bilgileri kullanarak çıkarım yapabilmek amacıyla tez kapsamında veri madenciliğinin en çok kullanılan görevlerinden biri olan sınıflandırma çalışması gerçekleştirilmiştir. Böylece Türkiye'de yaşayan insanların toplumsal yapısıyla ilgili genel kanaatler elde edilmiştir.

1. BÖLÜM
GENEL BİLGİLER VE LİTERATÜR ARAŞTIRMASI

1.1. Veri Madenciliğine Genel Bakış

Veri Madenciliği genel bir bakış açısı ile değerlendirilecek olursa, öncelikle VM kavramının neden ortaya çıktığının ifade edilmesi gerekir. Bu tanımlama, VM nedir? VM hangi fonksiyonları içerir? VM işletmelere ne ölçüde katma değer sağlayabilir? gibi soruların cevabını verecektir.

Bilgisayarın insan hayatına girmesinden itibaren, bilgisayarlar her geçen gün biraz daha gelişmiş ve değişmiştir. Kullanıcı gereksinimleri doğrultusunda değerli verilerin depolanması isteği ile veri tabanları ortaya çıkmıştır. Günümüzde hemen hemen bütün organizasyonlar bilgisayar teknolojisi ve internetin de etkisi ile kayıtlarını elektronik veri tabanlarına taşımışlardır. Bu veri tabanlarının boyutları 'terabayt'larla ifade edilmektedir. Zamanla veri tabanlarında gizli, stratejik ve politik değere sahip olabilecek, işe yarar örüntüler ve şablonlar elde etmenin mümkün olacağı düşüncesi yayılmıştır [16].

Bilgisayarların bilgi saklama kapasitelerinin artmasıyla birlikte bilgi kaydı yapılan alanların sayısı da artmaktadır. Bu sebepten dolayı, karar vericiler için eldeki verilerin analizi, tahminlenmesi ve bilgi çıkarımının önemi gün geçtikçe artmaktadır. Bilgisayar sistemleri ile üretilen veriler tek başlarına değersizdir, çünkü çıplak gözle bakıldığında bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman bir anlam ifade etmeye başlar. Bu yüzden günümüzde büyük miktardaki verileri işleyebilen teknikleri kullanabilmek büyük önem kazanmıştır. Bu ham veriyi bilgiye veya anlamlı hale dönüştürme işlemleri veri madenciliği ile yapılabilmektedir [3].

Veri tabanı teknolojilerinde sağlanan inanılmaz gelişmeler, işletmelerin veri toplamalarını ve sahip oldukları tüm süreçlerin her bir fonksiyonuna ait verileri kayıt

altına almalarını daha kolay hale getirmiştir. Bu şekilde büyük boyutlu verilerin toplanması, çok çeşitli bilgi depolama ortamlarını gerektirmiş ve işletmelerde veri ambarlarının ortaya çıkmasına yol açmıştır. Ancak bu veriler her geçen gün ulaşılması daha da zor hal alarak, bilgi sistemlerinin içinde gömülüp gitmesine yol açmıştır.

Bilgisayar ve iletişim teknolojilerindeki gelişmelere paralel olarak donanımın ucuzlaması, verilerin uzun süre depolanmasına dolayısıyla büyük kapasiteli veri tabanların oluşmasına neden olmuştur. Bu sebeple büyük boyutlu veri tabanlarında istenilen anlamlı, kullanılabilir ve ilginç bilgiye erişmek yeni bir disiplinin doğmasına neden olmuştur. Verilerin çeşitli istatiksel metotlarla analiz edilmesi, kurumların karar verme sürecinin etkinliğini ve yeni stratejiler geliştirmesine katkı sağlamıştır [40].

İşletmeler, çok yönlü bilgisayarlara ve iletişim sistemlerine sahip olmalarına rağmen, karar mekanizması konumundaki yöneticiler, uzmanlar ve danışmanlar, kritik bilgilere ulaşamama sıkıntısı yaşamaktadırlar. Bu anlamda veri yönünden zengin, fakat bilgi bakımından fakir bir ortamda tek çıkış yolu olarak VM gözükmektedir.

VM, özellikle kar ve pazar payı elde edebilmek için yoğun rekabetin yaşandığı pazarlama alanında ön plana çıkmaktadır. Hangi müşteri, hangi ürünü, ne zaman satın alabilir, kimler tedarikçilerinden vazgeçmekte ve bu tür müşterileri vazgeçirmek/geri kazanmak için neler yapılabilir, ürünün değerini yitirmesine hangi değişkenler neden olmaktadır, vb. soruların cevapları veri yığınlarının altında gizlidir ve cevapları bulabilmek için veri madenciliği çözümleri gereklidir [23].

VM ile büyük veri yığınlarından oluşan veri tabanı içerisinden, gerekli öz bilgilerin çekilip çıkarılması amaçlanmaktadır. VM, büyük miktarda verinin istatistik, matematik disiplinleri, modelleme teknikleri, veri tabanı teknolojisi ve çeşitli bilgisayar programları vasıtasıyla analiz edilerek anlamlı bilgiye dönüştürülmesini sağlayan bir sürecin adıdır. VM ile ilgili bazı diğer tanımlamalar aşağıda verilmektedir.

Karar destek sistemi için faydalı olabilecek, büyük hacimli veri içerisinden anlamlı, gizli kalmış bilgilerin çıkarıldığı ve arka planında istatistik, yapay zekâ ve veri tabanları bulunan veri analiz tekniğine VM adı verilmektedir.

VM, veri tabanları, istatistik ve yapay öğrenme konularının tekniklerini kullanarak; büyük miktardaki veri yığını içinden, gelecekle ilgili tahmin yapılmasını sağlayacak bağıntı ve kuralların aranmasıdır.

Verilerin bilgiye dönüştürülmediği sürece hiçbir anlam ifade etmediği bilinen bir gerçektir. Ham veriden bilgiye ulaşma süreci olarak tanımlanabilecek VM, çok miktarda verileri işleyerek hedef bilgilere ulaşmayı sağlar. Verilerin anlamlı olarak analiz edilmesi ve içerisinde gizlediği bilgilerin ortaya çıkarılması doğru bilgi yönetimi ile ve VM ile mümkündür. VM, verilerden pratik bilgi elde etmede ve buna bağlı olarak eylem planları oluşturmada temel rolü üstlenir.

1.2. Veri ambarı

1991 yılında ilk kez William H. Inmon tarafından ortaya atılan veri ambarı kavramı, yönetimin kararlarını desteklemek amacı ile çeşitli kaynaklardan elde ettikleri bilgileri zaman değişkeni kullanarak veri toplama olarak tanımlanmaktadır. Kısaca veri ambarları, birçok veri tabanından alınarak birleştirilen verilerin toplandığı depolardır. Veri ambarlarının en önemli özelliği, kullanıcılara veri hakkında farklı detay düzeyleri sağlayabilmesidir. Detayın en alt düzeyi arşivlenen kayıtların kendisi ile ilgili iken, daha üst düzeyler zaman gibi daha fazla bilginin toplanması ile ilgilidir. Veri ambarları ciddi yatırımlar gerektirmekte ve uygulanması bir yıl veya daha uzun zaman almaktadır [17].

Veri ambarı, yerleşik sistemlerde ve diğer dışsal sistemlerde var olan verilerin ayıklanması ve temizlenmesi; karar verme mekanizmalarına hizmet edecek şekilde hazırlanması, doğru şekilde saklanması, çeşitli son kullanıcı programları aracılığıyla veriye erişilmesi ve belirleyici veri ilişkilerinin aranıp bulunması işlemlerinin tümünü içeren bir aktiviteler zinciridir. Veri ambarının kullanımı çoğunlukla karar destek teknolojilerinin bir toplamını gerekli kılar [89].

Bir Veri ambarının yapısı organizasyon içindeki bütün son kullanıcılara, verileri ve işlem sonuçlarını sunan, en gelişmiş iletişimi sağlayan bir dizi birbiriyle bütünleşik alt bileşenlerden oluşur. Bunlar; [24]

- Operasyonel Veri Tabanı / Harici Veri Tabanı Katmanı,
- Enformasyon Ulaşım Katmanı,

- Veri Ulaşım Katmanı,
- Metadata,
- İşlem Yönetim Katmanı,
- Uygulama Haberleşmesi Katmanı,
- Veri Ambarı Katmanı,
- Veri Sunum Katmanıdır.

Veri ambarlarının, verilerde değişiklik yapmak amacıyla değil, sadece verileri okumaya yönelik olarak oluşturulması nedeniyle; veri ambarında veriler, analiz yapmayı kolaylaştıran bir formatta tutulmaktadır. Analiz sırasında; sorgular, raporlar, karar destek sistemleri veya istatistiksel hesaplardan faydalanılır. Veri ambarları, birbiriyle bütünleşik olmayan uygulamaların bütünleştirilmesine olanak sağlar.

Veri ambarlarının genel özellikleri, günlük hayatta kullanılan işlemsel veritabanlarıyla karşılaştırmalı olarak şöyledir [19];

- İşlemsel veritabanlarında yer alan veri, bir süzme işlemi sonucunda veri ambarına aktarılır.
- Zaman yelpazesi her iki sistemde farklılık gösterir. İşlemsel ortamdaki veri çok taze, veri ambarındaki ise eskidir.
- Veri ambarı özet bilgileri içerebilir. İşlemsel veri ise içermez.
- Bütünleştirmeyi sağlamak için verinin önemli bir kısmı belirli bir dönüşümden sonra veri ambarına aktarılır.

Veri ambarları, sağlık sektöründen bilişim sistemlerine, işletmelerin pazarlama bölümünden üretime, geleceğe dönük tahminlerde bulunmada, sonuçlar çıkarmada ve işletmelerin yönetim stratejilerini şekillendirmede kullanılan bir sistemdir. Maliyet yönünden pahalı bir yatırım olsa bile sonuçtaki getirisi bu maliyetten kat kat fazladır.

Veri ambarları günümüzde çeşitli amaçlarla kullanılmaktadır. Bu amaçlardan bazıları aşağıda belirtilmiştir [18];

- Müşterilerin fark edilemeyen satın alma eğilimlerini tespit etmek
- Satış analizi ve eğilimler üzerinde yoğunlaşmak,

- Finansal analiz (Maliyetlerin azaltılması dolayısıyla rekabet avantajının sağlanması)
- Stratejik Analiz (Bir Karar Destek Sistemi olmasından dolayı)
- İş ilişkilerin belirlenmesi
- Müşteri ihtiyaçlarına güvenilir biçimde ve anında cevap vermek.

Veriyi yönetmek için "veri ambarı", verileri çözümleyip bilgiye ulaşılabilmesi için "veri madenciliği" yöntemleri ortaya çıkmıştır [19]. Veri ambarı aşaması veri madenciliği sürecinde önemli bir aşamadır. Bu süreç toplam maliyet ve zamanın önemli bir kısmını almaktadır. Madenciliği yapılacak veri tek bir yapı içerisinde bulunmayabilir. Bu nedenle bilginin tek bir çatı altında toplanması gerekir. Fakat, veri ambarı oluşturma aşaması, sadece verinin tek bir çatı altında toplanması değildir. Aynı zamanda toplanan veriler içerisinde var olan hataların ve belirsizliklerinde temizlenmesi aşamasıdır. Bu aşamada veri bazı alt işlemlere tabi tutulmaktadır. Bu işlemler Veri Toplama, Uyumlandırma, Birleştirme ve Temizlenme, Seçme ve Dönüştürmedir [91].

Veri ambarlarında gizli olan bilginin ortaya çıkarılması için veri madenciliği yöntemlerinden faydalanılmaktadır.

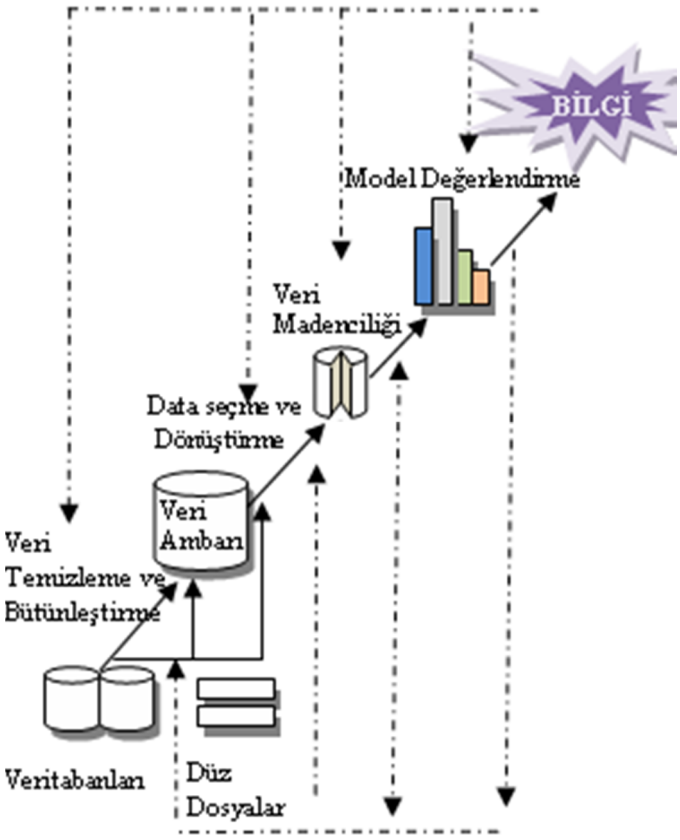
1.3. Veri Tabanlarından Bilgi Keşfi

Veri madenciliği terimi, çoğu zaman veri tabanlarından bilgi keşfi (VTBK) ile aynı anlamda kullanılır. Ancak VTBK, veri madenciliğinden daha fazla anlamlar ifade eder. VM, VTBK süreci içinde merkezi bir adımdır.

Büyük veri tabanlarında saklı olan, ilginç ve değerli olan bilgiyi algılamak ve erişmek oldukça zordur. VTBK süreci, bu değerli, önceden bilinmeyen, kullanılabilir bilgiye belirli metotlar uygulayarak veride gizli olan bilgiyi tanımlamada çok büyük rol oynamaktadır [20]. VTBK süreci Şekil 1.1'de gösterilmiştir.

Uygulama alanının incelenmesi: Konuyla ilgili bilgi ve uygulama amaçların belirlenmesidir.

Amaca uygun veri kümesinin oluşturulması: Analiz edilecek verinin hangi veri tabanında yapılacağını belirterek, veri seçimi ya da keşif edilecek alt veri örnekleri oluşturulmasıdır.



Şekil 1.1. Veri Tabanlarından Bilgi Keşfi Süreci [21]

Veri temizleme ve önişlem: Gürültülü ve tutarsız verilerin veri kümesinden silinmesidir.

Veri azaltma ve veri dönüşümü: Analizde gerekli özellikleri (boyutları) seçme, özellikler arasındaki ilişkiyi belirleme, veri dönüşümü ya da veri birleşimi yaparak boyut azaltma işlemidir.

Veri madenciliği görevi seçme: Sınıflandırma, birliktelik kuralları, kümeleme gibi veri madenciliği görevlerinden amaca uygun olanının seçilmesi işlemidir.

Veri madenciliği algoritması seçme: Seçilen veri madenciliği görevini yerine getirebilmek için görevle ilgili algoritmalar seçerek, algoritmalar üzerinde sonuçlar elde etme aşamasıdır.

Model değerlendirme ve bilgi sunma: Algoritmalardan elde edilen sonuçlardan en iyisini veren algoritmayı seçerek elde edilen modelin değerlendirilerek kullanıcıya sunulmasıdır.

Bulunan bilginin yorumlanması: Elde edilen sonuçların yorumlanarak dış dünyaya aktarılmasıdır [24-26].

1.4. Veri Madenciliği

Veri Madenciliği, 1960'lı yıllarda veri analizi sorunlarının bilgisayar ile çözülmeye başlanması ile ortaya çıkmış olup, 'Veri Madenciliği' ismi 1990'lı yıllarda bilgisayar mühendisleri tarafından ortaya atılmıştır [7].

Veri Madenciliğini veri dağları altındaki elmasları, altınları, külçeleri özel yazılımlar ile keşfetmek olarak tanımlamak mümkündür [8].

Literatürde yer alan bazı veri madenciliği tanımları ise aşağıda verilmektedir;

- Veriden örüntü elde etmek için belirli algoritmaların uygulamasıdır [9].
- İstatistiksel modelleri, matematiksel algoritmaları ve makine öğrenmesi modellerini de içeren veri analizi araçlarının kullanımını içermektedir [10].
- Daha önceden bilinmeyen örüntüleri ve anlaşılabilir bilgileri geniş veri tabanlarından seçmeyi, keşfetmeyi ve bu bilgileri modellemeyi içerir [11].
- Organizasyonların veri ambarlarındaki çok önemli bilgilere yoğunlaşmasına yardımcı olan, çok büyük veri tabanlarında saklı, akıllı bilgiyi ortaya çıkaran yeni bir teknolojidir [12].
- Veri tabanı, yapay zekâ ve istatistik dünyasını birbirine bağlayan son zamanlarda çıkmış bir alandır [13].
- Verinin, karar almak için kullanılan bilgiye dönüşümü- Bilgi Keşfi'nin tetikleyicisidir [14].
- Veri kümeleri içinden akıllı örüntüler keşfetmek için algoritmalar kullanan bir süreçtir [15].

Bu tanımlarda da görüldüğü gibi VM çok disiplinli bir çalışma alanıdır. Veri tabanı teknolojisi, istatistik, yapay zekâ, matematik, yönetim bilişim sistemleri veri madenciliğinde birlikte çalışmaktadır [16].

Veri madenciliğinin amaçları öngörü, tanıma, sınıflandırma ve en iyileme olarak dört başlık altında toplanabilir [4].

Öngörü; hangi ürünlerin hangi dönemlerde, hangi koşullarda, hangi miktarda satılacağına ilişkin ya da eğitimde hangi davranışlar sonucunda hangi olayların gerçekleşebileceği gibi kestirimlerde bulunmak şeklinde tanımlanabilir.

Tanıma; bir müşterinin, aldığı ürünlerden veya kullanıldığı programlar ve yaptığı işlemlerden tanınması olarak ifade edilebilir.

Sınıflandırma; birçok parametrenin birleştirilerek, kategorilere ayrılması olarak tanımlanabilir.

En iyileme; belirli kısıtlamalar çerçevesinde zaman, yer, para ya da hammadde gibi sınırlı kaynakların kullanımını en iyileme ve üretim miktarı, satış miktarı ya da kazanç gibi değerleri en büyükleme olarak tanımlanabilir [5].

VM; yöneticilere karar verme konusunda önemli ölçüde yardımcı olurken, analiz edilmesi, raporlanması gereken verilere erişim kolaylığı da büyük önem taşımaktadır. Bunu, veri ambarları olanaklı kılmıştır [16].

1.5. Veri Madenciliği Ne Değildir?

İdeal durumda tüm kurumlar faaliyetleri sonucunda elde ettikleri verileri değerlendirerek, kullanılabilir sonuçlar elde etmeyi hedeflemelidirler. Ancak uygulamalara bakıldığında, kurumların önemli bir kısmının verileri toplamanın ötesine geçmedikleri gözlenmektedir. Gelişim çizgisine bakıldığında verilerin doğru bir şekilde toplanması başlangıç noktasıdır. Elde edilen verilerden yapılan sorgulamalar ve detaylı analizler ile elde edilen sonuçları, veri madenciliği olarak değerlendirmemek gereklidir. Bir ölçüde bunlar da veri madenciliğidir ancak daha doğru tanımı veri düzenleme olarak adlandırılabilir. VM; veri toplamak, mevcut verilerden sorgulamalar yapmak veya gelişmiş analiz teknikleri kullanmanın ötesinde bir noktadır [23].

Bir restoran zincirinde; şubelerin yaptığı ciro miktarlarını belirleme, ürünlerin daha fazla satıldığı noktaların ve saatlerin belirlenmesine yönelik analizler veya bir satış şirketinde; müşterilerin hangi bölgelerde, ne kadar devamlılık gösterdiğine dair performansını belirlemek veri madenciliği değildir. Gelir ve yaş ilişkisine dayalı bir değişken, bir sonuç ve yeterli sayıda veriden oluşan bir modeli tanımlayarak, yaşa göre gelir tahmini yapmak da veri madenciliği değildir. Yüz değişkenin olduğu, değişkenler

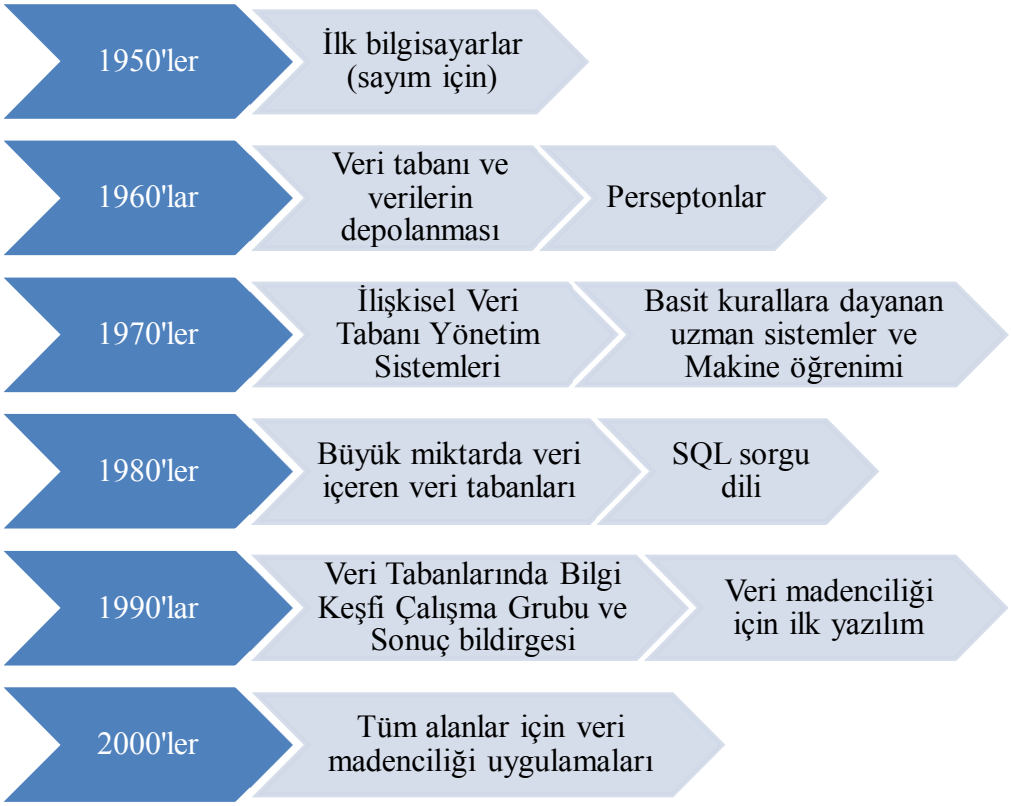
arasında sadece rakamsal değerlerin değil, sıralı (örnek: yüksek-orta-düşük) veya sırasız (örnek: evli-bekâr-dul) kategorilerin bulunduğu, milyon tane verinin olduğu ancak doğru algoritmalar ve güçlü bir bilgisayar ile sonuca ulaşmanın mümkün olduğu modelleri kurmak veri madenciliğidir. Algoritmalar doğrusal regresyondan daha karmaşık olmakla birlikte, kavram aynıdır, mevcut verileri kullanarak tahmin veya tanımlama yapmak [23].

1.6. Veri Madenciliğinin Tarihi

Günümüzde neredeyse her eve bilgisayar girmiştir ve internet erişimi hemen hemen her yerden sağlanmaktadır. Disk kapasitelerinin artması, her yerden bilgiye ulaşma olasılığı, bilgisayarların çok büyük miktarlarda veri saklamasına ve daha kısa sürede işlem yapmasına olanak sağlamıştır. Geçmişten günümüze verilerin her zaman yorumlanması, bilgi elde edilmesi istenmiştir ve bunun için de donanımlar oluşturulmuştur. Bu sayede bilgi, geçmişten günümüze taşınır hale gelmiştir [3].

1950'li yıllarda ilk bilgisayarlar sayımlar için kullanılmaya başlamıştır. 1960'larda ise veri tabanı ve verilerin depolanması kavramı teknoloji dünyasında yerini almıştır. 1960'ların sonunda bilim adamları basit öğrenmeli bilgisayarlar geliştirebilmişlerdir. Minsky ve Papert, günümüzde sinir ağları olarak bilinen perseptron'ların sadece çok basit olan kuralları öğrenebileceğini göstermişlerdir [2]. 1970'lerde İlişkisel Veri Tabanı Yönetim Sistemleri uygulamaları kullanılmaya başlanmıştır. Bilgisayar uzmanları bununla beraber basit kurallara dayanan uzman sistemler geliştirmişler ve basit anlamda makine öğrenimini sağlamışlardır. 1980'lerde veri tabanı yönetim sistemleri yaygınlaşmış ve bilimsel alanlarda, mühendislikte vb. alanlarda uygulanmaya başlamıştır. Bu yıllarda şirketler, müşterileri, rakipleri ve ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu veri tabanlarının içerisinde çok büyük miktarlarda veri bulunmaktadır ve bunlara SQL veri tabanı sorgulama dili ya da benzeri diller kullanarak ulaşılabilir. 1990'larda artık içindeki veri miktarı katlanarak artan veri tabanlarından, faydalı bilgilerin nasıl bulunabileceği düşünülmeye başlanmıştır. Bunun üzerine çalışmalara ve yayınlara başlanmıştır. 1989, KDD (IJCAI)-89 Veri Tabanlarında Bilgi Keşfi Çalışma Grubu toplantısı ve 1991, KDD (IJCAI)-89'un sonuç bildirgesi sayılabilecek "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop" makalesinin KDD (Knowledge Discovery and Data Mining) ile

ilgili temel tanım ve kavramları ortaya koyması ile süreç daha da hızlanmış ve nihayet 1992 yılında veri madenciliği için ilk yazılım gerçekleştirilmiştir. 2000'li yıllarda VM sürekli gelişmiş ve hemen hemen tüm alanlarda uygulanmaya başlanmıştır. Alınan sonuçların faydaları görüldükçe, bu alana ilgi artmıştır. Veri madenciliğinin tarihsel gelişim süreci, Şekil 1.2'de gösterilmiştir [3].



Şekil 1.2. Veri madenciliğinin tarihsel süreci [3]

1.7. Veri Madenciliği Süreci

Veri madenciliği süreci, verileri veri ambarından alır, bunları derler, düzenler ve yorumlar. Veri madenciliği algoritmalarının uygulanması öncesinde ham veri üzerinde bazı ön işlemlerin yerine getirilmesi söz konusu olabilir. Kurumların oluşturduğu birçok veri tabanında bilgiler eksik, yanlış, tekrarlı ve gereksiz olabilir. Bundan dolayı ham veri, sırasıyla temizleme, bütünleştirme, indirgeme ve dönüştürme gibi işlemlerden geçirildikten sonra, veri madenciliği algoritmaları uygulanarak sonuçlar elde edilir.

Eğer veri madenciliği verisi, bir veri ambarından sağlanıyorsa, bu işlemlere gerek kalmayabilir çünkü bu tür işlemler veri ambarı hazırlanırken yerine getirilir. Ön işleme işlemlerinin uygulanmasının ardından, veri analiz için hazır hale getirilir. VM yöntemlerinin bu işlenmiş veriye uygulanmasıyla sonuçlar elde edilir. Sonuçlar analiz edilerek veri içindeki örüntüler açığa çıkarılır.

Veri hazırlama ile ilgili işlemleri veri madenciliği kavramı içinde düşünürsek, veri madenciliğinin bir süreç olarak değerlendirmesi gerekir. Veri madenciliği sürecinin adımları Şekil 1.3'de gösterilmektedir [39].

Şekil 1.3. Veri Madenciliği Sürecinin Adımları

Verilerin Toplanması: Bu aşama verilerin nasıl toplanacağı ile ilgilidir. Verilerin oluşturulması yani toplanması sürecinde iki farklı yaklaşım vardır. Eğer süreç uzman kontrolünde yapılırsa tasarlanmış deney; uzman kontrolü olmadan yapılırsa gözlemsel yaklaşım olarak adlandırılır [41].

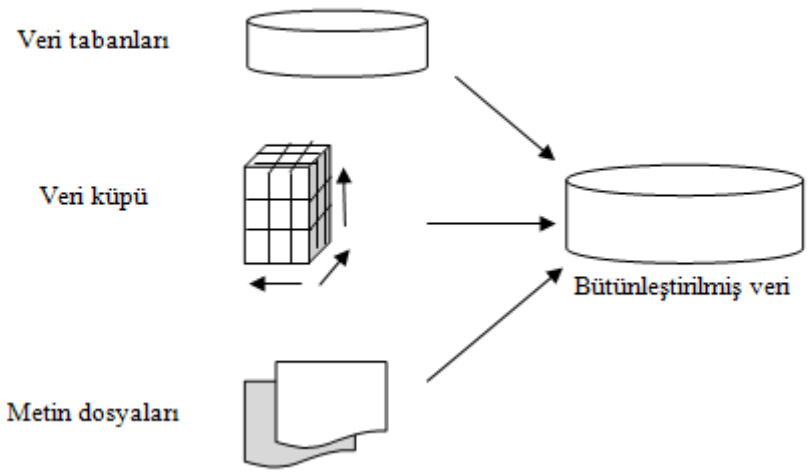
Problemin Belirlenmesi ve Verilerin Anlaşılması: Problemin belirlenmesi ve verilerin anlaşılması kısmı veri madenciliği uygulamasının ilk aşamasını oluşturmaktadır. Problemin belirlenmesi aşamasında bilinmeyen bağımlılıklara göre değişkenler belirlenir ve bir model oluşturmak için veriler arası ilişkilerden hipotez veya hipotezler oluşturulmaya çalışılır [145].

Veri Hazırlama: Veri tabanları içindeki verinin ve bu veriye dayalı olarak elde edilen veri madenciliği sonuçlarının kalitesinin artırılması, veriyi analize hazırlarken dikkat edilmesi gereken en önemli noktadır. Veri madenciliği işlemlerini kolaylaştırmak ve verimliliği artırmak için veri tabanındaki veriler, bir "ön işleme" aşamasından geçirilir [34]. Verinin hazırlanması süreci olarak da kabul edilen bu işlemler özellikle veri tabanındaki yanlış değerleri ve veriler arasındaki tutarsızlıkları kaldırmayı amaçlar.

Veri Temizleme: Bu aşamada, kayıp ya da eksik değerlerin tamamlanması, aykırı değerlerin belirlenerek gürültünün azaltılması ve verilerdeki tutarsızlıkları ortadan kaldırmak için kullanılan birçok teknik gerçekleştirilir. Veri kümesini bu tür eksik ya da kayıp değerlerden arındırmak için kullanılan yöntemlerden en önemlileri ve yaygın biçimde kullanılanları şunlardır [34]:

- Kayıp verilerin bulunduğu sorunlu kayıtların sayısı çok az ise, bu kayıtlar tamamıyla silinir.
- Kayıp değerlerin "BİLİNMİYOR" veya "∞" gibi sabit bir değer tanımlanması mümkündür; ancak veri madenciliği programları bu değeri ortak bir değer gibi algılayabilir. Bu nedenle çok tercih edilmemektedir.
- Sayısal değerlere sahip değişkenler için hesaplanacak ortalama değer kayıp değerler yerine kullanılabilir.
- Kayıp değer, mevcut verilere dayanarak en fazla kullanılan değer (mod) ile tamamlanabilir.
- Regresyon veya karar ağacı gibi teknikler ile en uygun değer belirlenerek bu değer kayıp değer yerine kullanılabilir.

Veri Bütünleştirme: Veri ambarının oluşturulması esnasında değişik kaynaklardan (veri tabanları, veri küpleri, metin dosyaları, vb.) elde edilen veriler arasında uyum sağlamak için verilerin bütünleştirilmesi işlemlerinin yerine getirilmesi gereklidir. Bu işlem sırasında, mevcut bütün veri kaynakları karakteristiklerine, özelliklerine ve toplanma seviyelerine göre eşleştirilerek, tutarlılık sağlanır. Kurum eğer birden fazla bilgi kaynaklarına sahipse kayıtlı olan verilerin bir araya getirilmesi aşamasında, aynı verinin farklı biçimlerde tutulması söz konusu olması durumunda veri bütünleştirmesi işlemine başvurulur. Bir tedarikçi veya müşteriye ait aynı verinin farklı biçimlerde saklanmış olması buna örnek olarak verilebilir. Veri bütünleştirme işlemi Şekil 1.4'de gösterilmiştir.



Şekil 1.4. Veri Bütünleştirme [39]

Veri İndirgeme: Veri madenciliği sürecinin ön işlemleri aşamasında ele alınan veri tabanında çok fazla kayıt ve gerekli olmayan nitelikler (değişkenler) olması durumunda veri indirgeme işlemi yapılır. Bu durumda verinin temel özelliklerini kaybetmeden, bir kısmını atarak azaltma yoluna gidilip gidilmeyeceği sorunu ile karşılaşılır [35]. Bu sorunu azaltmak için veri indirgeme yöntemine başvurulur.

Niteliklerdeki tutarsızlıklar elde edilen veri kümesinde fazlalıklara neden olabilir. Bu fazlalıkları tespit etmek için korelasyon analizinden yararlanılmaktadır. Örneğin korelasyon analizi sonucu, müşteri kimliği ile müşteri numarası isimli nitelikler arasında yüksek bir ilişki bulunursa bunlardan biri veri deposundan çıkarılarak indirgeme yapılır

[36]. Gereksiz boyut sayısı sorununun üstesinden gelmek içinse; nitelik (değişken) seçimi ve/veya kayıt seçimi işlemlerine başvurulabilir.

Veri Dönüştürme: Verilerin farklı ölçek ya da kod ile kaydedilerek; veri bütünlüğünü bozacak şekilde kaydedildiği durumlarda başvurulan bir yöntemdir. Örnek olarak; ağırlık ölçüsünün bir veri tabanında gram ile kaydedilirken diğerinde kilogram ile kaydedilmiş olması verilebilir. Veri dönüştürmede kullanılan yaklaşımlar aşağıda açıklanmıştır.

Ondalık Ölçekleme: Ondalık ölçekleme ile normalleştirmede, ele alınan değişkenin değerlerinin ondalık kısmı hareket ettirilerek normalleştirme gerçekleştirilir. Söz konusu ölçekleme, sayısal değerlerin -1 ile +1 arasında yer almalarını sağlayacak biçimde dönüştürülmesine karşılık gelir. Hareket edecek ondalık nokta sayısı, değişkenin maksimum mutlak değerine bağlıdır. Ondalık ölçeklemede değer, değeri 1'den küçük en büyük sayıya dönüştürecek 10'un derecesine bölünür. Örneğin 900 maksimum değer ise, n=3 olacağından 900 sayısı 0,9 olarak normalleştirilir [37].

Min-Max Normalleştirme: Orijinal veri üzerinde doğrusal bir dönüşüm yapmak için Min-Max normalleştirme kullanılır. Veriler bu yöntem aracılığıyla genellikle 0-1 aralığına dönüştürülür.

Z-Skor Standartlaştırma: Z-Skor standartlaştırma verileri dönüştürmek amacıyla kullanılan bir diğer yöntem olarak bilinir. Ele alınan verinin ortalama ve standart sapma değerlerini kullanan bu yöntem, istatistiksel veri dönüştürme teknikleri arasında yer alır ve yaygın biçimde kullanılır.

Farklı tekniklerle gerçekleştirilebilen normalizasyon işlemi, veri boyutunun küçültülmesi amacıyla kullanılabileceği gibi, verilerle gerçekleştirilecek işlemlerin uygun aralıklara normalize edilmiş değerlerle yapılarak işlemlerin daha hızlı gerçekleştirilip ve daha anlamlı ve kolay yorumlanabilir sonuçlar almak amacı ile de kullanılabilir [38].

Modelin Kurulması: Tanımlanan problem için uygun modelin belirlenmesi aşamasında, mümkün olduğunca çok sayıda model kurularak ve bu kurulan modellerin denenmesi ile gerçekleşir. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele ulaşıncaya kadar yinelenen bir aşamadır [41].

Modelin Yorumlanması: Model elde edilmesi beklenen hedefleri karşılamaya yeterli bulunduğu zaman, süreç bazlı daha geniş bir perspektiften değerlendirme yapılır. Bu değerlendirme sürecinde modelin doğru kurulup kurulmadığı, gelecekte kullanılabilecek farklı verilerin neler olabileceği, modelin genişletilmesi gibi konuları içerir [41].

1.8. Veri Madenciliği Uygulama Alanları

Veri madenciliğinin uygulama alanları ana başlıkları altında şöyle sıralanabilir;

Web Uygulamaları

- Kullanıcı taraflı bilgiler (tarayıcı, dil, vb) ışığında alt yapı düzenlemeleri [25],
- Kullanıcıların profillerini çıkarma ve zaman içindeki değişimleri takip etme, sitedeki beğenilen ya da beğenilmeyen köşeleri tespit etme [26],
- Kullanıcı profillerine göre site perspektifi düzenleme,
- Site haritası, bağlantılar, vb. düzenlemeleri,
- Kullanıcıların gezinti şekli/hızı sitenin içerik, yapılandırma ve alt-yapı açısından performansı hakkında fikir verir [26],
- Kullanıcı profillerine uygun ürünlerin reklam kampanyalarını, kullanıcının en çok ziyaret ettiği sayfalara koyma [27],
- En sık beraber ziyaret edilen çift sayfaları belirleme [27],
- Farklı web şablonları, temaları arasındaki kullanıcı isteklerini değerlendirme,
- Form verilerinin toplanmasındaki zorlukları en aza indirme yöntemleri geliştirme,
- Kötü niyetli kullanıcı isteklerini belirleyip, bunlara karşı alınması gereken önlemleri belirleme [25].

İşletme Alanı

- Müşterilerin satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi,
- Müşteri ilişkileri yönetimi,

- Müşteri değer analizi,
- Satış tahmini [24],
- Kendi müşterisini rakibine kaptıran işletmeler için, müşterilerle ilgili analizler yaparak rakiplerini tercih etme nedenlerine yönelik bilgiler elde etme ve gelecek dönemlerde kaybetme olasılığı olan müşterilerine yönelik tahminlerde bulunma ve onları kaybetmemek için strateji geliştirme,
- Ürün veya hizmette hangi özelliklerin ne derecede müşteri memnuniyetini etkilediği, hangi özelliklerinden dolayı müşterinin bunları tercih ettiğini ortaya çıkarma,
- Müşterilerin kredi risklerinin hesaplanarak hangi müşterilerin kredi riskinin yüksek olduğunu, hangi müşterilerin geri ödemesini zamanında yapamayabileceğini kestirme,
- Kredi kartı ödemelerini aksatan, gecikmeli olarak yapan veya hiç yapmayanların özelliklerinden yola çıkılarak bundan sonra aynı duruma düşebilecek muhtemel kişileri saptama,
- Ürün talebi bazında müşteri görünümlerini belirleyerek, müşteri segmentasyonuna gitmek ve çapraz satış olanakları yaratmada,
- Piyasada oluşabilecek değişikliklere mevcut müşteri portföyünün vereceği tepkinin firma üzerinde yaratabileceği etkinin tespitinde kullanma,
- En kârlı mevcut müşterileri saptayarak, potansiyel müşteriler arasından en kârlı olabilecekleri belirleme. Kârlı müşterileri tespit ederek onlara özel kampanyalar uygulama. En masraflı müşterileri daha masrafsız müşteri haline dönüştürme. Örneğin en çok bankacılık işlemi yapanlar ortaya çıkarılıp bunlar şube bankacılığı yerine daha masrafsız internet bankacılığına yönlendirilebilir.
- Bir ürün veya hizmetle ilgili bir kampanya programı oluşturmak için hedef kitlenin seçiminden başlayarak bunun hedef kitleye hangi kanallardan sunulacağı kararına kadar olan süreçte veri madenciliği kullanma,
- Kurum teknik kaynaklarının en uygun şekilde kullanılmasını sağlamakta kullanma,
- Geçmiş ve mevcut yapı analiz edilerek geleceğe yönelik tahminlerde bulunma Özellikle ciro, kârlılık, pazar payı gibi analizlerde de veri madenciliği kullanılabilir [28].

Endüstri Alanı

- Kalite kontrol analizlerinde,
- Lojistikte,
- Üretim süreçlerinin optimizasyonunda kullanılabilir [6].

Eğitim Alanı

- Kütüphane kullanıcılarının erişim örüntülerinin keşfi [29],
- Öğrenci Seçme Sınavına giren öğrencilerin profillerinin ve tercihlerinin öğrenci başarılarına etkisi [30].

1.9. Veri Madenciliğinde Karşılaşılan Problemler

Veri madenciliğinde karşılaşılan problemler aşağıdaki şekilde özetlenebilir;

Kayıp Veriler: VM algoritmasının uygulanacağı veri kümesinde bazı kayıtlar hiç girilmemiş olabilir. Kayıp verilere sahip bir veri tabanına uygulanabilecek yöntemlerden ilki kayıp verinin bulunduğu kaydı veri tabanından çıkarmaktır. Eğer kayıp verili kayıt sayısı, toplam kayıt sayısına göre oldukça az ise bu kayıtların veri tabanından çıkarılması mümkündür. Diğer taraftan kayıp veri sayısı yüksekse veya bu kayıtlara ait diğer değerler önemliyse kayıp değerlerin yerine genel bir sabit kullanılabilir veya kayıp verilerin yerine tüm verilerin ortalama değeri kullanılabilir [32]. Daha etkin olarak ise regresyon kullanılarak diğer değişkenlerin yardımı ile kayıp verilerin değerleri tahmin edilebilir. Ayrıca zaman serileri analizi, bayesyen sınıflandırma, karar ağaçları, maksimum beklenti gibi teknikler de kayıp verilerin tahmininde kullanılabilir [33].

Yanlış ya da Aşırı Uç Veriler: Kullanılan veri tabanında hatalı ya da tutarsız değerler içeren verilere örnek olarak, yaş bilgisinin 180 girilmesi verilebilir. Doğru olması mümkün olmayan böyle verilere gürültülü veri denir. Bu tür veriler için veri düzgünleştirme tekniği kullanılır. Tutarsız verilerden kaynaklanabilecek problemleri ortadan kaldırmada en etkin yöntem bu tarzdaki uç değerlerin veri tabanından çıkarılmasıdır. Diğer taraftan veri düzgünleştirme, her bir kümede en çok kullanılan veriyle (mod) veya her kümenin aritmetik ortalamasının alınarak küme içindeki verilerin bu aritmetik ortalamayla (medyan) değiştirilmesiyle yapılır. Kullanılabilecek

bir başka yöntem ise kenardaki verilerin birbirlerinden farkının küme elemanı sayısına bölünmesiyle elde edilen değerin o küme elemanına atanmasıdır.

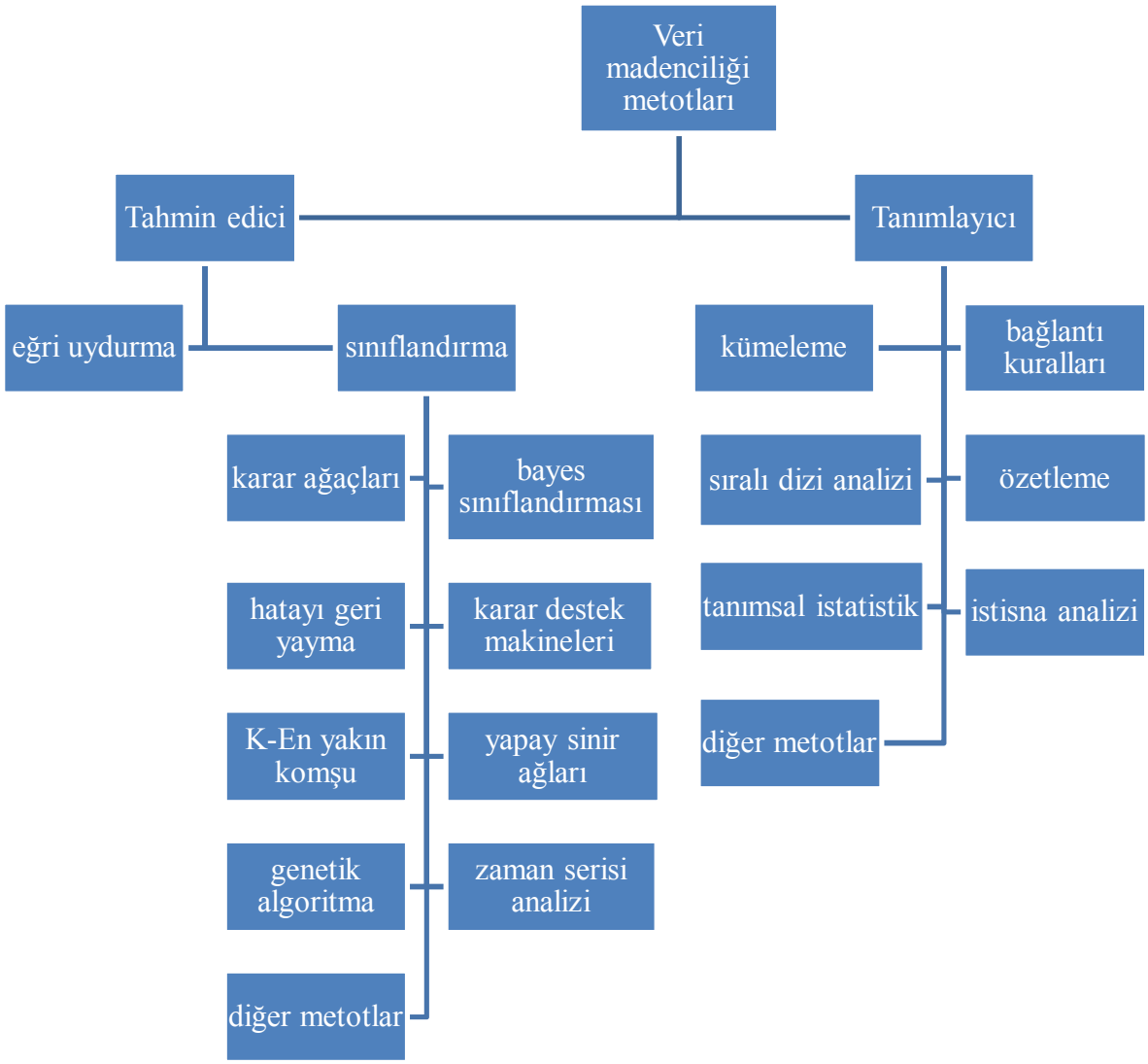
Gereksiz Veriler: Aynı veri tabanı içerisinde hem yaş hem de doğum tarihi bilgisinin verilmesi gibi aynı anlama gelen veya herhangi ilişkisi olmayan, bilgisayar çalışma zamanını artıran ve sonuçların güvenilirliğini ve kalitesini etkileyebilecek veriler gereksiz veri olarak adlandırılır. Böyle durumlarda veri boyutunun indirgenerek mevcut değişkenlerin birleştirilmesi en iyi çözümdür. Bu amaçla en sık kullanılan yöntemler dalga dönüşümü ve temel bileşenler analizidir [32].

Değişken Değerlerinin Birbirinden Ayrık Olduğu Durumlar: Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu takdirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve bu değişkenler diğerlerinin rollerini önemli ölçüde azaltır. Bu nedenle bir dönüşüm yöntemi uygulayarak söz konusu değişkenlerin normalleştirilmesi ve standartlaştırılması uygun bir yol olacaktır [31]. Bu amaçla en sık kullanılan yöntemler verinin alabileceği minimum ve maksimum değeri kullanan min-maks normalizasyonu ve ortalama ve standart sapma değerlerini kullanan sıfir-ortalama normalizasyonudur [33].

Kullanılacak Algoritmaya Bağlı Yapılandırmalar: Veri madenciliğinde kullanılan bazı algoritmalar sadece sayısal değerlerle, bazıları sadece kategorik değerlerle bazıları ise ikili değerlerle işlem yaparlar. Mevcut veri, kullanılacak belirli türdeki verilerle çalışabilen algoritmaya uygun hale getirilmelidir.

1.10. Veri Madenciliği Yöntemleri

Veri madenciliğinde kullanılan yöntemler, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici yöntemlerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçların bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Tanımlayıcı yöntemlerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır. Şekil 1.5'de veri madenciliği metotları gösterilmektedir [40].



Şekil 1.5. Veri madenciliği metotları [22]

Veri madenciliğinin üstlendiği görevler geniş bir bakış açısıyla dört boyutta incelenebilir: Sınıflandırma, tahmin, bölümlendirme/kümeleme, tanımlama/özetleme. Görevin ana hedefine bağlı olarak, veri madenciliği sürecinin çıktısı tahminleyici modeller veya tanımlayıcı bilgiler olacaktır. VM süreci tarafından türetilen tahminleyici modeller sınıflandırma veya tahmin görevleri söz konusu olduğunda kullanışlı iken; tanımlayıcı bilgiler, bölümlendirme/kümeleme ve özetleme türünden problemler için uygun olacaktır. Her bir VM uygulaması farklı hedef ve durumlara sahip olduğundan, farklı VM tekniği kümelerini gerektirmektedir.

VM, veri tabanında beklenmeyen ilişkilerin bulunmasına yardım etmesi açısından geleneksel istatistiksel analizlerden farklılık göstermektedir. Yüksek boyutluluk ve büyük hacimli verilerden ötürü, veri madenciliğinde geleneksel istatistiksel yöntemlerin kullanımı sınırlanmıştır [42].

1.10.1 Sınıflandırma ve Regresyon

Sınıflandırma ve regresyon, önemli veri sınıflarını ortaya koyan veya gelecek veri eğilimlerini tahmin eden modelleri kurabilen iki veri analiz yöntemidir. Sınıflandırma kategorik değerleri tahmin ederken, regresyon süreklilik gösteren değerlerin tahmin edilmesinde kullanılır. Örneğin, bir sınıflandırma modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir. Sınıflandırma ve regresyon modellerinde kullanılan başlıca teknikler şunlardır [45]:

1- Karar Ağaçları
2- Yapay Sinir Ağları
3- Genetik Algoritmalar
4- K-En Yakın Komşu
5- Bellek Temelli Nedenleme
6- Naive-Bayes.

Sınıflandırma, günlük yaşamda çok sıklıkla başvurulan bir işlemdir. Sınıflandırma ile nesneler bölünerek ayrıştırılır, yani karşılıklı olarak özel ya da genel kategorilerden her biri bir sınıf olarak atanabilir. Pek çok pratik karar verme işlemi, bir sınıflandırma problemi olarak formüle edilebilir. Örneğin kişiler ya da nesneler birçok kategoriden biri olabilir [46].

Sınıflandırma, farklı sınıflardaki, değişik öğeleri ayırma sürecidir. Bu sınıflar, iş kuralları, sınıf sınırları veya bazı matematiksel fonksiyonlar olabilir. Sınıflandırma işlemi, sınıflandırılmış olan öğenin, bilinen bir sınıf değeri ile özellikleri arasındaki bir ilişki üzerine inşa edilebilir. Bu sınıflandırma tipi, “denetimli öğrenme” olarak isimlendirilir. Eğer bir sınıfın bilinen örnekleri yoksa bu sınıflandırma denetimsizdir. En yaygın denetimsiz sınıflandırma yaklaşımı, kümelemedir. Kümeleme teknolojisinin en

yaygın uygulamaları, perakende ürünlerde birliktelik analizi (market sepet analizi) ve dolandırıcılık tespitidir [47].

Veri önişleme, parametre seçimi ve test kümesi seçimi veri madenciliği uygulamasında ortaya çıkacak olan modelin başarısını etkiler. Dolayısı ile yapılan karşılaştırma sonuçları büyük ölçüde uygulamacıya bağlıdır [48].

1.10.1.1 Karar Ağaçları ile Sınıflandırma

Temel olarak bir karar ağacının oluşturulup, veri tabanındaki her kaydın bu ağaca uygulanması ve çıkan sonuca göre kaydın sınıflandırılması esasına dayanır. Karar ağaçlarının yapılandırılması ve uygulanması hem görsel, hem de anlaşılabilirlik açısından daha kolaydır. Kullanılan algoritmaya göre karar ağacının şekli değişebilmekte ve farklı sınıflandırma sonuçları verebilmektedir. Karar ağaçlarına bağlı olarak geliştirilen birçok algoritma kök, düğüm ve dallanma kriteri seçimlerinde izledikleri yol açısından farklılık göstermektedirler.

Karar ağaçlarının oluşturulması sırasında dallanmaya hangi nitelikten başlanacağı oldukça önemlidir çünkü olası tüm ağaç yapılarını ortaya çıkararak içlerinden en uygun olanı ile başlamak mümkün değildir. Bu sebeple karar ağacı algoritmalarının çoğu daha başlangıçta birtakım değerleri hesaplayarak ona göre ağaç oluşturma yoluna gitmektedir. Bu hesaplamalardan biri de entropiye dayalıdır. Entropi, bir veri kümesi içindeki belirsizlik ve rastgeleliği ölçmek için kullanılır ve 0 ile 1 arasında değer alır. Bütün olasılıklar eşit olduğunda, entropi maksimum değerini alacaktır [31]. Entropiye dayalı karar ağaçları ile sınıflandırma algoritmalarının en önemlileri aşağıdaki gibidir.

ID3 [49]: ID3, makine öğrenme ve bilişim teorisine bağlı olarak verilen örnekler içinde en ayırıcı değişkeni bulan bir algoritmadır. Temel olarak kategorik nitelikleri sınıflandırır ve veri tabanı dallandırılmadan önce ve sonra doğru sınıflandırma yapmak için gelen bilgiler arasındaki farkı kullanarak, öncelikli düğüme ve dallanmalara karar verir.

C4.5 [50]: ID3 algoritmasından farklı olarak, sayısal değerlere sahip nitelikler için de karar ağaçlarının oluşturulmasını sağlar. Diğer taraftan karar ağacı oluştururken kayıp verileri almaması sebebiyle daha anlamlı kurallar sunan ağaçlar üretebilir. Kayıp veriler ise diğer veri ve değişkenler kullanılarak tahmin edilir.

CART [51]: CART algoritması, her karar düğümünden sonra ağacın iki dala ayrılması ilkesine dayanır. Bu teknikte dallanma kriteri belirlenirken kayıp veriler önemsenmez.

1.10.1.2 Yapay Sinir Ağları ile Sınıflandırma

Çıktı katmanını elde etmek için ağırlıkların hesaplanmasına dayanan yapay sinir ağları ile sınıflandırmada eğitim veri kümesi üzerinde hesaplanan ağırlıklar, test veri kümesi üzerinde kullanılarak öğrenmenin ne kadar gerçekleştiği belirlenir. Elde edilen ağırlıkların etkinliği doğrulanana kadar ağırlıklar üzerinde düzeltme ve yeniden hesaplama işlemleri gerçekleştirilir. Ağırlıklar yardımıyla yeni bir verinin hangi sınıfa ait olduğu öğrenme süreci tamamlanması ile belirlenebilir. Öğrenme süreci uzun süren yapay sinir ağları, oldukça duyarlı sınıflandırmalar yapabilmektedir.

1.10.1.3 İstatistiğe Dayalı Algoritmalar

CHAID, regresyon, lojistik regresyon, Bayesyen sınıflandırma ve zaman serileri analizi yaklaşımı gibi istatistiksel yöntemler veri madenciliğinde sınıflandırma algoritması olarak kullanılmaktadır. Bunlardan en sık kullanılan Bayesyen sınıflandırma algoritması ve regresyon analizi, sınıflanmış verileri kullanarak yeni bir verinin mevcut sınıflara girme olasılığını hesaplar.

1.10.1.4 Mesafeye Dayalı Algoritmalar ile Sınıflandırma

Bu algoritmalar mevcut verilerin birbirlerine olan uzaklıklarını hesaplayarak sınıflandırma yapan k-en yakın komşu algoritması ve en küçük mesafe sınıflandırıcısıdır. k-en yakın komşu algoritması yaygın olarak kullanılır. Mesafeye dayalı algoritmalar sınıfları belli bir örnek kümedeki gözlem değerlerini kullanarak yeni gözlemlerin hangi sınıfa ait olduğunu belirlemek amacı ile kullanılır. Örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip k adet gözlemin seçilmesi esasına dayanır. k-en yakın komşu algoritması bilimsel çalışmalarda geliştirilen birçok algoritmanın temelini oluşturur.

1.10.2 Kümeleme Yöntemi

Kümeleme analizi grupları kesin olarak bilinmeyen birimleri, değişkenleri birbiriyle benzer alt kümelere (grup, sınıf) ayırmaya yardımcı olan çok değişkenli istatistiksel analiz yöntemlerinden biridir. Ayırma, aynı gruptaki gözlemler birbirine benzer iken, farklı gruplardaki gözlemler birbirinden farklı olacak şekilde yapılmaktadır [43]. Kümeleme analizinin temel amacı, birimleri sahip oldukları karakteristik özellikleri temel alarak gruplandırmaktır. Kümeleme analizi son yıllarda gündemde olan analiz yöntemlerinden biridir. Bu yöntem özellikle bilim ve iş alanında birçok durumda uygulanabilen, kolay yorumlanabilen ve etkili olan bir yöntem olma özelliğini taşır. Bu nedenle birçok bilim alanında bu yöntemden yararlanılmaktadır [52].

Gözlemler ya da değişkenler kümelemenin konusu olabilir. Birimlerin kümelenmesinde bir gözlemler setinin kümelere sınıflanması amaçlanmakta, ancak ne grup sayısı ne de grup üyeliği bilinmektedir. Değişkenlerin kümelenmesinde ise, birbiriyle ilişkili değişken gruplarının yani ortak faktör yapılarının ortaya çıkarılması amaçlanmakta, dolayısıyla faktör analizi ile benzerlik görülmektedir [53].

Kümeleme analizi, hemen hemen tüm bilim alanlarında yararlanılan bir yöntemdir. Tıp, biyoloji, psikoloji, sosyoloji, arkeoloji gibi belirsizlik koşullarının ve karmaşık oluşumların bulunduğu bilim alanlarında ise daha yoğun olarak yararlanılan bir yöntemdir. Örneğin, tıp alanında; hastalıkların sınıflandırılması, psikiyatride; paranoya, şizofreni gibi semptomların doğru sınıflandırılması (teşhis profilleri oluşturması), laboratuvar bulguları ile klinik bulguların oluşturduğu veri matrislerinden hastalık alt gruplamalarının ya da yeni semptomların tanımlanması gibi amaçlarla kümeleme analizinden yararlanılmaktadır [54].

Kümeleme analizinin başlıca varsayımları, veri matrislerinin analiz öncesi tahmin ve kriter değişkenleri alt matrislerine bölüştürmemesi ve verilerin kısmen homojen, kısmen heterojen oluşudur. Kümeleme analizinde ilk aşama, bir benzerlik veya uzaklık ölçüsünün seçilmesidir. Sonra kullanılacak kümeleme tekniğine (hiyerarşik veya bölümlemeli gibi) yönelik bir karar verilir. Üçüncü adımda seçilen teknik için kullanılacak olan kümeleme yöntemi türü (hiyerarşik kümeleme tekniğinde centroid yöntemi gibi) seçilir. Son aşamada ise küme sayısı belirlenerek kümeleme sonucu yorumlanır [53] .

Kullanıcının amacına ve kullanım alanına göre kümeleme analizinin amaçları aşağıdaki gibi sıralanabilir [55]:

- Doğru tiplerin belirlenmesi
- Model oluşturmak
- Gruplara dayalı tahmin
- Hipotez testi
- Veri araştırma (inceleme)
- Hipotez oluşturma
- Veri indirgemedir.

Kümeleme analizinde hiyerarşik ve bölümlemeli olmak üzere iki tür kümeleme yöntemi vardır. Hiyerarşik olan kümeleme yönteminde kaç küme olacağı bilinmemektedir ve başlangıçta n birey için n tane küme mevcuttur. En yakın iki küme birleştirilir ve küme sayısı bir indirgenerek yinelenmiş uzaklıklar matrisi bulunur. Bu işlem n-1 kez tekrarlanır. Hiyerarşik olan kümeleme yönteminde, ağaç diyagramları ile gösterilen sonuçlara dendogram denir. Hiyerarşik olan küme yönteminde kullanılan teknikler şöyledir: Tek Bağlantı Tekniği, Tam Bağlantı Tekniği, Ortalama Bağlantı Tekniği, Küme Merkezleri Tekniği, Ward's Tekniği. Tek bağlantı tekniği, en kısa mesafe esasına dayanır. Birbirine en yakın iki gözlem bulunur ve küme çekirdeği ilk aşamaya oturtulur. Tam bağlantı tekniği, tek bağlantı tekniğine benzer. Farkı, en uzak iki gözlemden başlamasıdır. Ortalama bağlantı tekniğinde, bir kümenin ortasına düşen gözlem esas alınırken; Küme Merkezleri Tekniğinde, bir kümeyi oluşturan gözlemlerin ortalamaları esas alınmaktadır. Ward's tekniğinde ise, bir kümenin ortasına düşen gözlemin, aynı kümenin içinde bulunan gözlemlerden ortalama uzaklığı esas alınarak toplam sapma karelerinden yararlanılır [56].

1.10.2.1 Hiyerarşik Algoritmalar

BIRCH, CURE, CHAMELEON, SLINK ve ROCK algoritmaları hiyerarşik kümeleme algoritmalardır. Hiyerarşik algoritmalar, birbirine en fazla benzeyen iki nesneyi bir kümede toplar. Bu, işlem maliyeti oldukça yüksek bir süreçtir. Çünkü her toplamadan önce tüm nesneler karşılaştırılır. Bölücü yaklaşımın da gerektirdiği hesaplama yükü

yığışımsal yaklaşıma benzer. Ayrıca, hiyerarşik kümeleme yöntemlerine örnek olarak Ward'ın minimum varyans yöntemi de verilebilir [72].

SLINK (En yakın komşu algoritması) [57]: Her bir verinin ayrı bir küme olarak ele alındığı ve aşamalı olarak bu kümelerin birleştirildiği bir yapıya sahiptir. Bu algoritmada iki kümenin birbirine olan uzaklığı, o kümelerdeki birbirine en yakın verilerin birbirine olan uzaklığı olarak kabul edilir. Eğer eldeki uzaklık verisi belli bir eşik değerini geçiyorsa kümeler birleştirilir.

CURE (Temsilciler kullanarak kümeleme) [58]: Veri tabanı içinde diğer verilerden uzakta bulunan ve sayıları az olup aslında hiç bir kümeye ait olmaması gereken uç verilerin kümeleme kalitesini etkilememesi amacıyla geliştirilmiş bir algoritmadır. En yakın komşu algoritmasındaki toplaşım ve yakınlık prensibine dayanır.

En uzak komşu algoritması: En yakın komşu algoritmasından farklı olarak iki kümenin birbirine olan uzaklığı, kümelerdeki birbirine en uzak verilerin arasındaki uzaklıkla belirlenir.

CHAMELEON [59]: İki kümenin birbirine olan uzaklığının yanı sıra birbirine olan benzerlikleri bilgisini de kullanır. İki kümenin birleştirilmesi esnasında, kümelerin birbirine olan benzerliği ve yakınlığı ile bu kümelerin kendi iç benzerlikleri ve yakınlıkları karşılaştırılır. Böylece daha kaliteli ve homojen kümeler elde edilir. En yakın komşu algoritmasındaki toplaşım ve yakınlık prensibine dayanır.

BIRCH (Hiyerarşi kullanarak dengelenmiş iteratif azaltma ve kümeleme) [60]: Temel olarak gürültülü verilerin kontrol edilmesi amacıyla büyük boyutlu veri tabanlarının kümelenmesi için geliştirilmiştir. En uzak komşu algoritmasında olduğu gibi bölünür bir yapıya sahip olan algoritma, sadece sayısal verilere uygulanabilmektedir.

1.10.2.2 Bölümlemeli Algoritmalar

Kümeler arasındaki minimum ya da maksimum uzaklığın, kümelerin iç benzerlik kriterlerinin ve küme sayısının kullanıcı tarafından belirlendiği algoritmalardır. Hiyerarşik algoritmalardan daha hızlı çalışan bölümlemeli algoritmalar, bu özelliklerinden dolayı büyük veri tabanlarının kümelenmesi için daha uygundur.

k-ortalamalar algoritması [61]: Verilerin kümelerin ortalamalarına göre önceden belirlenmiş k adet kümeye ayrılması esasına dayanan bir algoritmadır. Toplam ortalama hatanın minimize edilmesini amaçlayan algoritma sadece sayısal verilerde kullanılabilir. Bu alandaki algoritmaların çoğu k-ortalamalar algoritmasının geliştirilmesiyle ortaya çıkmıştır.

k-medoid algoritması [62]: Sayısı önceden belirlenmiş k adet kümenin her biri için k adet medoid belirlenmesi ile başlayan algoritma veri tabanındaki diğer verilerin kendilerine en çok benzeyen medoidlerin etrafına toplanması esasına dayanır. Medoid ise kümenin merkezine yakın uzaklıkta bulunan noktayı temsil etmektedir.

CLARA algoritması (Geniş uygulamaların kümelenmesi) [62]: k-medoid algoritmasından farklı olarak tüm veri tabanını tarayarak medoid noktaları belirlemek yerine veri tabanından rastgele oluşturulan bir örnek küme üzerinde benzer şekilde çalışır. İki algoritma karşılaştırıldığında CLARA algoritmasının büyük boyutlu veri tabanları için daha güvenli olduğu ve daha kısa süre içinde kümeleme yapabildiği belirtilmiştir.

CLARANS algoritması (Rastgele aramaya dayalı geniş uygulamaları kümeleme) [63]: k-medoid ve CLARA algoritmalarının geliştirilmiş bir halini barındıran CLARANS algoritması Şebeke diyagramından yararlanan bir yapıya sahiptir. CLARA algoritmasına benzer olarak bütün veri tabanı taranmazken yapılan örnekleme dinamik bir yapıya sahiptir.

1.10.2.3 Yoğunluğa Dayalı Algoritmalar

Yoğunluk-tabanlı kümeleme yaklaşımı ise diğer bir hiyerarşik olmayan kümeleme yaklaşımıdır. Yoğunluk-tabanlı kümeleme algoritması verinin yoğun olduğu alanı küme olarak aldığı için, rastsal şekillere sahip kümeler bulmada sorun yaşamaz. Tipik algoritma örnekleri arasında; DBSCAN, DENCLUE ve OPTICS algoritmaları sayılabilir [74].

1.10.2.4 Izgara Tabanlı Algoritmalar

Izgara-tabanlı kümeleme yaklaşımı veri noktalarından çok hücreleri göz önüne alan bir yaklaşımdır. Bu özelliğinden dolayı, ızgara-tabanlı kümeleme algoritmaları genel olarak

tüm kümeleme algoritmalarından hesapsal olarak daha etkindir. Bu yaklaşıma örnek olarak; STING, STING+, WaveCluster, CLIQUE ve GDILC verilebilir [73].

1.10.3 BİRLİKTELİK KURALLARI

Birliktelik kuralları, veri kümesi içindeki hareketlerin (işlemlerin) analiz edilerek bu hareketler ya da kayıtlar arasında sıklıkla bir arada görülenlerin tespit edilmesi işlemidir. Birliktelik kuralları ticaret, mühendislik, fen ve sağlık sektörlerinin de içinde bulunduğu diğer birçok alanda uygulanmaktadır [66].

Birliktelik kuralları ile bir ilişkide yer alan niteliklerin değerleri arasındaki bağımlılıklar, anahtarda yer almayan diğer niteliklerin gruplandırılması ile bulunur. Bu kurallar ilk olarak Agrawal tarafından 1994'te geliştirilmiştir [64].

Çok sayıda verinin depolandığı bir veri tabanı içinde çeşitli nitelikler arasında hemen fark edilmeyen bir takım ilişkiler mevcut olabilir. Bu tip ilişkilerin ortaya çıkartılması stratejik kararların alınmasına yardımcı olabilir. Ancak, bu ilişkilerin çok sayıda verinin içinden elde edilmesi basit bir süreç değildir. Bu süreç birliktelik kuralı madenciliği olarak adlandırılmaktadır [65].

Market sepet çözümlerinde satılan ürünler arasındaki ilişkileri ortaya koymak için "destek" ve "güven" gibi iki ölçütten yararlanılır. Bu ölçütlerin hesaplanmasında "destek sayısı" adı verilen bir değer kullanılır. "kural destek ölçütü" bir ilişkinin tüm alışverişler içinde hangi oranda tekrarlandığını belirler. Kural güven ölçütü, A ürün grubunu alan müşterilerin B ürün grubunu da alma olasılığını ortaya koyar [32].

1.11. Veri Madenciliği Literatürü

1.11.1 Türkiye'deki Veri Madenciliği Çalışmaları ve Uygulamaları

Pek çok alanda etkili bir şekilde kullanılmaya başlanan veri madenciliği, günümüzün en çok uygulanan disiplinlerinden birisi olmuştur. Her geçen sene kendisine daha da yaygın bir kullanım alanı bulmakla birlikte, kolay uygulanabilirliği ve etkili sonuçlar ortaya çıkarması sayesinde, kurum ve kuruluş yöneticileri tarafından en çok başvurulan yöntemlerden biri haline gelmiştir. Literatürde yer alan veri madenciliği uygulamaları,

eğitim, ticaret, mühendislik, bankacılık ve borsa, tıp ve telekomünikasyon başlıkları altında sınıflandırılarak aşağıda özetlenmiştir.

Mühendislik Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Veri madenciliği, mühendislik alanında etkin olarak kullanılmaktadır. Bu çalışmalardan, Kıyas Kayaalp [75] 2007 yılında yaptığı yüksek lisans tezinde, veri madenciliği tekniği ile üç fazlı asenkron motordaki sargı spirleri arasında oluşabilecek kısa devre veya yalıtım bozuklukları ve motor milinde oluşabilecek mekanik dengesizlik hatalarının tespiti gerçekleştirilmiştir.

Ali İnan tarafından [76] 2006 yılında yapılan kişilerin konum bilgilerinin toplanması, kullanımı ve dağıtılması ile ilgili bir çalışmada, konum-zaman veri tabanlarında veri madenciliğini mümkün kılmak için özel olarak tasarlanan algoritmalara gerek duyulmuş ve zaman-mekân nitelikleri olan veriler için bir gizliliği koruyan veri madenciliği tekniği ve iki ön-işleme tekniği önerilmiştir. bunlar (1) Dağıtık kümeleme, (2) Merkezi anonimleştirme ve (3) Dağıtık anonimleştirmedir. Önerilen tekniklerin güvenlik ve performans analizleri de yapılmış ve sonuçta mantıklı varsayımlar altında minimum mahrem bilgi kaybıyla veri madenciliğinin mümkün olduğu gözlemlenmiştir.

Gökhan Yavaş [77] tarafından 2003 yılında gerçekleştirilen başka bir çalışmada ise mobil kullanıcıların hareket modellerinin veri madenciliği kullanılarak çıkarılması ve bu modeller kullanılarak mobil kullanıcıların daha sonraki hareketlerinin tahmin edilmesi için yeni bir algoritma geliştirilmiştir. Sunulan algoritmanın performansı simülasyonlar yardımıyla iki farklı tahmin yöntemiyle karşılaştırılmıştır. Performans sonuçları algoritmanın diğer metotlardan daha doğru tahminler yapabildiğini göstermiştir.

Sibel Kırmızıgül Çalışkan ve İbrahim Soğukpınar [78] 2008 yılında, veri madenciliği yöntemlerinden “K-ortalamalar” ve “K en yakın komşu” yöntemlerinin iyileştirilmesi amacıyla; nüfuz tespiti için kümelemeyi ve sınıflandırmayı, denetimli ve denetimsiz öğrenimi, k-ortalamalar ve k en yakın komşu yöntemlerini bir arada kullanan hibrit bir yapı geliştirmiştir. Geliştirilen uygulamada en hızlı sonucu veren k-ortalamalar uygulaması ile test kümesi daha küçük alt kümelere ayrılarak k en yakın komşu yönteminin zaman karmaşası ve bellek gereksinimi azaltılmıştır.

Nevcihan Duru ve Mücella Canbay [79] 2007 yılında veri madenciliği ile deprem verilerinin analizi üzerine bir çalışma gerçekleştirmişlerdir. Bu çalışma deprem verileri kullanılarak seçilen bir bölgeye ait sismik tehlikenin diğer deyişle gerçekleşme olasılığının veri madenciliği yönünden ele alınarak incelenmesini kapsamaktadır. Uygulama, dünya ölçeğindeki her noktanın analizini yapacak şekilde geliştirilmiş olup, ihtiyaç halinde programa eklemeler yapmak suretiyle, başka bu tür çalışmalar yapacak şekilde tasarlanmıştır.

Yaşar Doğan [80] tarafından 2004 yılında Deniz Harp Okulu'nda, su altı taktik duyarga ağlarında veri madenciliği tabanlı hedef sınıflandırması çalışması hazırlanmıştır. Bu çalışmada, açık, sığ ve çok sığ sularda denizaltı, küçük sualtı taşıma araçları, sualtı mayınları ve dalgıçları sınıflandırmada maliyeti çok az olan mikro duyargalar kullanılmıştır. Algoritma, yüzeydeki şamandıralara bağlı ve ayarlanabilir derinliklere indirilebilen duyargalardan oluşan taktik su altı duyarga ağları için tasarlanmıştır. Sınıflandırmada veri madenciliği tekniği olarak karar ağacı algoritmaları kullanılmıştır.

Eyüp Sıramkaya [81]'nın 2005 yılında hazırladığı bir uygulamada internet üzerinden ulaşılabilen basın-yayın kaynaklarında yer alan görsel ve metinsel verilerin hızlı ve etkin bir şekilde erişimi ve bu kaynaklardan anlamlı ve önemli bilgilerin çıkarılması hedeflenmiştir. Bir ara yüz ile kullanıcının bu bilgileri sorgulaması sağlanmıştır. Çalışma, Birliktelik Kural Madenciliği tekniklerinden Apriori Algoritması kullanılarak uygulanmıştır. Ayrıca kişi-kişi ilişkilerini bulmakta, isimlerdeki harflerin konumlarının birbirlerine göre uzaklıklarını temel alarak bulanık mantık kurallarının uygulandığı bir algoritma kullanılmıştır Bu uygulamadaki amaç kullanıcıların arama yapmak istedikleri kişilerin isimlerini yazarken yapabilecekleri yazım hatalarını elemektir.

Yomi Kastro [82] 2006 yılında, bir yazılımın yeni sürümlerindeki hata oranını eski sürümlerine göre olan değişikliklerini temel alarak tahmin eden bir model ortaya koyma amaçlı bir uygulama gerçekleştirmiştir. Bu model, aynı zamanda bir yazılım ürününe katılan yeniliklerin, hata ayıklama değişiklikleri gibi değişiklik türlerinin, hata oluşturma ihtimallerine olan katkısını ayrı ayrı anlamaya yardımcı olmaktadır.

Coşku Erdem [83], 2006 yılında, matematiksel morfoloji kullanarak yoğunluk temelli kümeleme adında bir uygulama gerçekleştirmiştir. Bu uygulamadaki algoritma veri depolarının imgelere benzerliğinden yola çıkarak bir imge işleme tekniği olan gri tonlu

morfolojinin çok boyutlu veri üzerine uygulanması temeline dayanmaktadır. Önerilen bu algoritmanın gerek sentetik gerekse doğal veri üzerindeki başarımı değerlendirilmiş ve uygun parametrelerle çalıştırıldığında başarılı ve yorumlanabilir sonuçlar üretebildiği görülmüştür.

T. Tugay Bilgin [84] 2009 yılında gerçekleştirdiği bir çalışmada, veri akış diyagramları ve veri akışı tabanlı veri madenciliği süreçleri görselleştirilmesini açıklamıştır. Üç farklı tür veri akış tabanlı yazılımı incelemiş ve detaylı özelliklerini karşılaştırmıştır.

2004 yılında Serkan Toprak [85] tarafından, ilişkisel veri tabanları üzerinde çoklu ilişkisel yapıdaki ortak kuralları bulmayı sağlayan bir uygulama geliştirilmiştir. Uygulama altyapısı olarak ilişkisel veri tabanlarındaki desenleri tanımlayabilen, bu desenleri eklerle geliştirebilen ve bu desenlerin çeşitli ölçmeleri için gerekli sayımları veri tabanından temel yetilerle alan bir yapı kullanılmıştır. Bu çalışma, Apriori algoritmasını, arama alanını daha da küçültmek için kullanarak ve altyapı tarafından desteklenmeyen özyinelemeli desenlerin bulunmasını sağlayarak altyapıya yenilikler getirmiştir. Uygulama bir veri madenciliği yarışması olan KDD Cup 2001'den alınan örnek genlerde yer tahmini problemi ile test edilmiş ve ortaya çıkan sonuçlar yarışmayı kazanan yaklaşımın sonuçlarıyla karşılaştırılmıştır.

Ulaş Baran Baloğlu [86] tarafından 2006 yılında gerçekleştirilen uygulamada, DNA veri kümesinde bulunan biyolojik sıralar üzerinde veri madenciliği yapılarak tekrarlı örüntüler ve potansiyel motifler çıkartılmıştır. E. coli bakterilerinden alınmış DNA sıralarında, önerilen yöntem denenerek uygulanabilirliği ve üstün yanları gösterilmiştir.

Barış Yıldız [87] 2010 yılında, sık kümelerin bulunması için gizliliği koruyan bir yaklaşım önermiştir. Ayrıca bu çalışmada Matrix Apriori algoritması üzerinde değişiklikler yapılmış ve sık küme gizleme çerçevesi de geliştirilmiştir.

Yasemin Kılınç [88] 2009 yılında hazırladığı bir çalışmada, birliktelik kuralları için bir yöntem sunmuştur. Apriori algoritmasının ürettiği kurallar elenerek bir elektronik firmasında üretim ve mal giriş kalite verileri üzerinde uygulanmıştır. Ortaya çıkarılan kurallar test verileri ile doğrulanmış ve sonuçlar analiz edilmiştir.

Tıp Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Barış Aksoy [89] tarafından 2009 yılında Dekompresyon Analizinin kümeleme analizi üzerine bir veri madenciliği uygulaması gerçekleştirmiştir. Bu çalışmada, farklı kümeleme algoritmaları (k-ortalama, COBWEB, EM) ile Divers Alert Network (Dalgıçların Acil Durum Ağı)'nın dalış yaralanmaları bildirim formlarından elde edilen belirti ve bulgu listeleri kullanılarak dekompresyon hastalığı sınıflandırılmış ve sonuçlar klasik sınıflandırma yöntemleri, yeni istatistiksel sınıflandırma yöntemleri ve tedavi sonuçları ile karşılaştırılmıştır. Ayrıca teşhiste yardımcı olabilecek birliktelik kuralları elde edilmiştir.

Pınar Yıldırım vd. [90] tarafından 2008 yılında yapılan çalışmada, hastane bilgi sistemlerindeki veri madenciliği uygulamalarına değinilmiştir.

Şengül Doğan ve İbrahim Türkoğlu [91] tarafından 2008 yılında gerçekleştirilen bir çalışmada, kan biyokimya parametreleri ile demir eksikliği anemisi teşhisinde, hekime yardımcı olacak ve kolaylık sağlayabilecek bir karar destek sistemi oluşturulmuştur. Tasarlanan sistemde 96 hasta verisi değerlendirilmiş ve sistemin sonuçları, doktorun verdiği kararlarla tamamen örtüşmüştür.

Mustafa Danacı vd. [92] tarafından 2010 yılında gerçekleştirilen çalışmada kanser çeşitlerinden biri olan ve kadınlar arasında en sık görülen meme kanseri hakkında kısa bilgi verilmiştir. Daha sonra Xcyt örüntü tanıma programı yardımı ile doku hakkında genel veriler elde edilmiş, Weka programı kullanılarak meme kanseri hücrelerinin tahmin ve teşhisi yapılmıştır.

Bankacılık ve Borsa Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Nihal Ata vd. [93] tarafından 2007 yılında gerçekleştirilen çalışmada, yaşam çözümlemesi yöntemleri veri madenciliği konusu çerçevesinde ele alındıktan sonra kredi kartı sahiplerine ait bir veri kümesi için yaşam olasılıkları ve regresyon modelleri incelenmiştir. Buna göre çalışmada yaş, gelir ve medeni durumun, müşterilerin kredi kartı kullanmayı bırakmalarını etkileyen önemli risk faktörleri olduğu görülmüştür.

Ali Sait Albayrak ve Şebnem Koltan Yılmaz [44] tarafından 2009 yılında gerçekleştirilen bir çalışmada, İMKB 100 endeksinde sanayi ve hizmet sektörlerinde faaliyet gösteren 173 işletmenin 2004–2006 yıllarına ait yıllık finansal göstergelerinden

yararlanarak veri madenciliği tekniklerinden birisi olan karar ağaçları tekniği uygulanmıştır. Ayrıca Ali Sait Albayrak tarafından gerçekleştirilen başka bir çalışmada, yerli ve yabancı olarak önceden grup üyeliği belirlenmiş bankaların sınıflandırılmasında veri madenciliği tekniklerinden diskriminant, lojistik regresyon ve karar ağacı modelleri kullanılarak bankalarla ilgili seçilmiş likidite, gelir-gider, karlılık ve faaliyet oranları sonuçları karşılaştırılmıştır. Araştırmanın sonuçları, bankaların sınıflandırmasında karar ağacı modelinin diskriminant ve lojistik regresyon modellerine üstünlük sağlayarak alternatif etkili bir sınıflandırma tekniği olarak kullanılabileceğini göstermiştir.

H. Ali Ata ve İbrahim H. Seyrek [94] tarafından 2009 yılında gerçekleştirilen çalışmada, denetçiler tarafından yaygın olarak bilinmeyen bazı veri madenciliği teknikleri, finansal tablolardaki hileleri tespit etmeye yardımcı olmak üzere kullanılmıştır. Çalışma İMKB'de işlem gören ve imalat sektöründe faaliyet gösteren 100 firmanın bilgilerine dayalı olarak gerçekleştirilmiştir. Araştırma sonucunda kaldıraç oranı ve aktif karlılık oranının finansal tablo hilesini tespit etmede önemli finansal oranlar olduğu belirlenmiştir.

İpek Savaşçı ve Rezan Tatlıdil [95] tarafından 2006 yılında müşteri ilişkileri yönetimi üzerine bir çalışma gerçekleştirilmiştir. Bu çalışmada bireysel bankacılık alanında uygulanan müşteri ilişkileri yönetim süreci incelenmiş ve müşteri sadakatinin oluşturulmasını sağlayan kredi kartlarında uygulanan CRM stratejileri değerlendirilmiştir.

Eğitim Alanında Gerçekleştirilen Veri Madenciliği Uygulamaları

Konya Selçuk Üniversitesi'nde Onur İnan [96] tarafından, hazırlık sınıfı, birinci sınıf ve mezun durumunda olan öğrenciler üzerinde, üniversite veri tabanındaki veriler kullanılarak; öğrencilerin başarılarını etkileyen etmenler, başarı düzeyleri, üniversiteyi kazanan öğrenci portföyleri ve mezun olamayan öğrencilerin okulu bitirmelerini etkileyen etmenler üzerinde çalışmalar gerçekleştirilmiş ve sonuçları yorumlanmıştır.

Y. Ziya Ayık vd. [97] tarafından yapılan çalışmada, Atatürk Üniversitesi öğrencilerinin mezun oldukları lise türleri ve lise mezuniyet dereceleri ile kazandıkları fakülteler arasındaki ilişki, veri madenciliği teknikleri kullanılarak incelenmiştir.

Ahmet Selman Bozkır vd. [98] tarafından gerçekleştirilen bir çalışmada, ÖSYM tarafından 2008 ÖSS adayları için resmi internet sitesi üzerinden yapılan anket verileri üzerinde veri madenciliği yöntemleri kullanılarak, öğrencilerin başarılarını etkileyen faktörler araştırılmıştır. Bu araştırmada, veri madenciliği yöntemlerinden karar ağaçları ve kümeleme kullanılmıştır. Buna benzer bir çalışma Şenol Zafer Erdoğan ve Mehpare Timor [99] tarafından 2005 yılında gerçekleştirilmiş, öğrencilerin üniversiteye giriş sınavı sonuçları ve öğrencilerin başarıları arasındaki ilişki, kümeleme analizi ve k ortalamalar algoritması teknikleri uygulanarak incelenmiştir. Bu çalışmanın KPSS'ye uygulanmış bir modeline benzeyen çalışmayı da Hüseyin Özçınar [100] 2006 yılında gerçekleştirmiştir. Frekans analizi ve regresyon analizi yöntemleri kullanılarak derslere ve yıllara göre verinin özellikleri incelenmiştir. Oluşturulan regresyon modeli ile KPSS sonuçlarının değişimi üzerinde anlamlı katkısı olan değişkenler incelenmiş ve oluşturulan modellerin tahmin doğrulukları, ortalama mutlak hata ve ortalama hata kareler kökü değerleri kullanılarak karşılaştırılmıştır.

Ahmet Selman Bozkır ve Ebru Sezer [101] tarafından 2009 yılında gerçekleştirilen başka bir çalışmada ise Hacettepe Üniversitesi Beytepe Kampüsü'ndeki öğrenci ve çalışanların, gıda tüketim desenleri incelenmiştir. Çalışmada, karar ağaçları ve birliktelik kuralları uygulanmıştır ve çalışma sonunda % 80 başarıyla, gıda tüketim deseninin ortaya çıkarıldığı görülmüştür.

Hidayet Takçı ve İbrahim Soğukpınar [35] tarafından 2002'de gerçekleştirilen bir çalışmada kütüphane sitesi web günlüklerine dayalı olarak kütüphane kullanıcılarının erişim örüntüleri bulunmaya çalışılmıştır. Bu çalışma yapılırken istatistiksel yöntemler kullanılmıştır.

Murat Kayri [102] tarafından 2008 yılında gerçekleştirilen bir çalışmada, öğrencilerin performans göstergelerinin sürekli izlenebilmesi ve ürünler arasındaki örüntünün bilgisayar sistemleri tarafından oldukça kolay yapılabildiği e-portfolyo değerlendirmeleri için veri madenciliğinde kullanılan yöntemler alternatif bir ölçme yaklaşımı olarak önerilmiştir.

Ticari Alanda Gerçekleştirilen Veri Madenciliği Uygulamaları

Anarberk Kalıkov [103] tarafından, bir yayınevi firmasının internet sitesindeki veriler dikkate alınarak, veri madenciliği birliktelik kuralları tekniği ile sepet ve sipariş

tabloları incelenmiştir. Hangi ürünlerin kategorisinin değiştirilmesi gerektiği, kullanıcıların meslek ve ilgi alanı dağılımları, müşteri ilgi alanlarına göre satış grafikleri ve kullanıcıların ödeme seçenekleri ile ilgili bir veri madenciliği uygulaması gerçekleştirilmiştir.

Sinem Akbulut [104] tarafından yapılan çalışmada, bir kozmetik markasının müşteri grupları ve ayrılma eğilimi gösteren müşteri kesimi belirlenerek; bu müşterilere özel pazarlama stratejileri geliştirilmesi hedeflenmiştir. Bölümleme için kümeleme teknikleri, ayrılacak müşteri kesitini belirlemek için sınıflama teknikleri kullanılmıştır.

Feridun Cemal Özçakır ve A. Yılmaz Çamurcu [105] tarafından gerçekleştirilen bir çalışmada, bir firmanın pastane satış verileri üzerinde veri madenciliği uygulaması olarak birliktelik kuralları ile bir yazılım tasarlanmıştır. Genelde aynı ürün grubuna ait ürünlerin, en sık birlikte satın alınan ürünler olduğu görülmüştür.

Feyza Gürbüz vd. [106] tarafından gerçekleştirilen başka bir çalışmada, Türkiye'de bir hava yolu işletmesinin parça söküm raporları üzerinde veri madenciliği çalışması gerçekleştirilmiştir. Çalışmanın amacı, uçaklarda kullanılan parçaların, herhangi bir arıza oluşmadan önce düzeltici ve önleyici işlemlerin yapılması için ikaz seviyelerinin tespit edilmesine yönelik kural geliştirmektir. Sonuç olarak parçaların ikaz seviyelerini temsil edecek anlamlı bir kural elde edilmiş ve bulunan kurallar doğrulukları ve güvenilirlikleri bakımından test edilmiştir.

Mehmet Aydın Ulaş [107] tarafından 2001 yılında yapılan bir yüksek lisans çalışmasında, sepet analizi gerçekleştirilmiştir. Büyük süpermarket zinciri olan Gima Türk A.Ş.'nin verileri üzerine Apriori algoritması uygulanmış ve ortaya çıkan sonuçlar incelenmiştir. Ayrıca mal satışları arasındaki ilişkileri bulmak amacıyla da, bileşen analizi ve k-ortalama öbeklemesi metotları kullanılmıştır.

Çağatan Taşkın ve Gül Gökay Emel [86] tarafından 2010 yılında veri madenciliğinde kümeleme yaklaşımları ve Kohonen ağları ile perakendecilik sektöründe bir uygulama gerçekleştirilmiştir. Bu uygulamada; bir perakende işletmesinin müşterilerinin Kohonen ağları ile kümelenmesi ele alınmıştır. Kümeleme analizinin amacı; ele alınan işletmeye, pazar bölümlendirmesi ve hedef pazar seçimi gibi stratejik pazarlama kararlarında yardımcı olması için önceden bilinmeyen kritik müşteri özellikleri ve önem derecelerini de ortaya çıkararak gerekli öngörüyü sağlamaktır.

Fatma Güntürkün [108] 2007 yılında işletmelerin kalite iyileştirmelerini araştıran bir yüksek lisans çalışması hazırlamıştır. Bu çalışmada, sürücü koltuğu kalitesi için müşteri memnuniyeti verisi analiz edilmiştir. Müşterinin sürücü koltuğundan memnuniyetini etkileyen en önemli değişkenlerin belirlenmesi için karar ağaçları yaklaşımı uygulanmıştır. Bu uygulamadan elde edilen sonuçlar diğer bir çalışmada aynı veri kümesine uygulanmış ve lojistik regresyon analizinden elde edilen sonuçlarla karşılaştırılmıştır.

TÜİK Verileri Üzerine Veri Madenciliği Uygulamaları

Selim Tüzüntürk [109] tarafından 2010 yılında, TÜİK'in 2000 yılında yapmış olduğu Türkiye'deki 81 ilin sosyoekonomik gelişmişliği hakkındaki anket verileri kullanılarak OLAP analizi yapılmıştır. Kullanılan veriler ankette yer alan 59 değişkenden oluşmaktadır ve toplam veri sayısı 4779'tır. Çalışmada SPSS 13.0 Paket programı kullanılmıştır.

Zeynep Behrin Güven ve Turgay Tugay Bilgin [110] tarafından yapılan çalışmada WEKA programı ile Türkiye İstatistik Kurumu'ndan (TUIK) alınan veri seti üzerinde zaman serileri madenciliği algoritmaları uygulanmıştır. 2001-2010 yılları arasındaki nüfus verileri kullanılarak daha sonraki yıllar için nüfus tahmini yapılmıştır.

Semih Erdem [121] tarafından 2011 yılında, İleri Veri Tabanı Sistemleri dersi dönem projesi için veri madenciliğine yönelik seçim konusu ele alınmıştır. Projede 1999, 2002, 2007 Türkiye milletvekilleri genel seçimleri baz alınmıştır. Genel seçimler için Marmara Bölgesi'nde yer alan Çanakkale ve Balıkesir illeri seçilmiştir. Proje kapsamında illerin ekonomik ve sosyal yapıları, göç durumları, yaş grupları, partilere göre oy dağılımları, işsizlik oranı vb. yapıları incelenmiştir. İlçelerde ise kentsel ve kırsal kesimin partilere göre oy dağılımları üzerinde durulmuştur. Partiler ise bu üç genel seçime katılanlar arasından seçilmiştir. İllerde sıralamada ilk 5'te yer alan, ilçelerde ise ilk 3'te yer alan partiler ele alınmıştır.

Tuğba Değirmenci [111] tarafından 2014 yılında TÜİK'in düzenli olarak yürütmekte olduğu Hane Halkı Bütçe Araştırması teknik boyutu ile ele alınarak, 2012 araştırma verileri üzerinde geniş çaplı bir kümeleme ve birliktelik analizi uygulamasına yer verilmiştir.

1.11.2 Dünya'daki Veri Madenciliği Çalışmaları ve Uygulamaları

Nada Lavrac vd. [112] tarafından 2006 yılında Sloven halk sağlığı ve bakımı bölgesel yönetimi ve planlaması ile ilgili Bölgesel Kamu Sağlık Enstitüsü tarafından elde edilen veri tabanları üzerine veri madenciliği çalışmaları yapılmış ve istatistiksel teknikler uygulanmıştır. Nüfusun kamu sağlık hizmetlerinin kullanılabilirliği ve erişilebilirliği açısından tipik olmayan alanlarını belirlemek amacı ile seçilmiş Celje bölgesinde, kamu sağlık kaynaklarının organizasyonel yönleri üzerine çalışılmıştır.

Yas A. Alsultanny [113] tarafından 2013 yılında market piyasası ihtiyaçlarını tahmin etmek için Naive Bayes sınıflandırıcıları, karar ağaçları ve karar kuralları teknikleri kullanılmıştır. Eğitim tabloları oluşturularak Naive Bayes tekniği uygulanmış, bu tabloların setleri çalışanların işlerinde iken sürekliliğini etkileyen 4 faktör kullanılarak üretilmiştir. Sınıflandırmanın bilinmeyen yönlerini test etmek için öncelikli olasılıklar ve koşulların sonuçlarını tablo haline getirmek ve diğer örneklerin sınıflandırılmasını tahmin etmek için eğitim tabloları kullanılmış, elde edilen bilgiler işgücü piyasasındaki istihdamın bilinmeyen örneklerini sınıflandırmıştır.

Francisco Torres-Avilés vd. [114] tarafından 2014 yılında potansiyel yeni direnç genlerini tanımlamak için regresyon, veri madenciliği ve kümelemenin yararlılığı değerlendirilmiştir ve Phytophthora zararlısı ile aşılanmış ve aşılanmamış domatesler ele alınarak bu iki farklı durum ayrı ayrı 3 yöntemle de analiz edilmiştir. Uygulanmış yöntemlerle elde edilen 10 yeni direnç geni tahminleri en güvenilir olarak seçilip potansiyel direnç genleri olarak rapor edilmiştir.

Abdullah A. Aljumah vd. [115] tarafından 2012 yılında regresyon tabanlı veri madenciliği tekniği kullanılarak diyabetik hastalardaki öngörü analizi üzerinde yoğunlaşılmıştır. Suudi Arabistan'daki bulaşıcı olmayan hastalıkların veri setlerinin risk faktörleri Dünya Sağlık Örgütü'nden elde edilerek analiz için kullanılmıştır. Tedavinin tercih edilen yöntemleri araştırılmıştır. Çalışmalar sonucunda, genç yaş grubundaki hastalarda ilaç tedavisinin yan etkileri önlemek için geciktirilebileceği kanaatine varılmıştır.

Tim Van den Bulcke vd. [116] tarafından 2011 yılında Belçika'daki PCMA tarama merkezi tarafından yapılan yeni doğan taramasının sistematik bir parçası olarak türetilememiş tarama yöntemiyle toplanan 44159 kan numuneleri veri setini

kullanılarak; eşik optimizasyon metodu ve bir parametre ile C4.5 karar ağaçları, lojistik regresyon ve sırt lojistik regresyon veri madenciliği metotları uygulanmış ve tanısal destek aracı olarak uygulanabilirliği ölçülmüştür. Tek katmanlı çapraz doğrulama ayarı içerisinde, değişkenleri ve sınıflandırma eşiklerini içeren geniş yelpazedeki model parametrelerinin her bir modeli için ızgara arama uygulanmıştır.

Imran Khan vd. [117] tarafından 2013 yılında saatlik olarak kaydedilen enerji tüketimi ve puant talebin verileri kullanılarak anormal aydınlatma enerji tüketimini belirlemek için, 3 farklı veri madenciliği tekniği uygulanmıştır. İki uç algılama yöntemi aynı veri seti içerisindeki anormal tüketimi belirlemek için her sınıf ve kümeye uygulanmıştır. Anormal tüketimle ilgili her sınıf ve kümede, normalden gelen değişim miktarı modifiye standart puanlar kullanılarak belirlenmiştir. Manuel olarak hataları algılamaya veya yanlış uyarıların teşhisine gerek kalmadan işletme maliyetini ve süresini azaltmak için enerji yönetim sistemlerinin kurulmasının yararlı olacağı kanaatine varılmıştır.

Alireza Pakgohar vd. [118] tarafından 2010 yılında lojistik regresyon, sınıflandırma ve regresyon ağaçları gibi veri madenciliği teknikleri ile İran'daki trafik kazalarına ilişkin trafik verileri analiz edilmiştir. Bu güncel araştırma bulgularının hükümete daha iyi yol tasarımı ve trafik yönetimi için yardımcı olacağı umulmuştur.

Wipada Chanthaweethip ve Sumanta Guha [119], 2012 yılında HIV hastalarını bakan hekimlere yardımcı olmak için tahmin sonuçlarının görsel temsilini sağlayarak tedavi sonucunu tahmin etmek için zamansal veri madenciliğini önermişler, geçici soyutlamayı her biri bir sembol ile tipik olarak temsil edilen ayrık kategoriler halindeki zaman serisi verilerini sınıflandırmak için kullanmışlardır. Yapay sinir ağları, öğrenme işlemi sırasında dengelenmemiş verinin büyüklük sorunu olduğu yerlerde meydana gelen bu çalışmada kullanılmıştır.

Dake Zhang ve Kang Jiang [120], 2012 yılında Çin'deki yangın nedeniyle oluşan kaza kayıplarının çok büyük olması nedeniyle veri analizleri, kaza analiz teorisi ve Microsoft SQL Server iş zekası bileşenlerinin kombinasyonu ile birlikte; kümeleme analizi, birliktelik kuralları, zaman serileri ve karar ağaçları ile yangın kaza vakaları ve verilerinin analizi üzerinde çalışmış ve veri madenciliğini gerçekleştirmişlerdir. Burada veri madenciliğinin, olayların analizinde uygulama yöntemlerine bir referans sağlaması

ve yangın kazalarındaki veri işleme ve bilgi analizinde belli bir akademik değere ve pratik öneme sahip olması amaçlanmıştır.

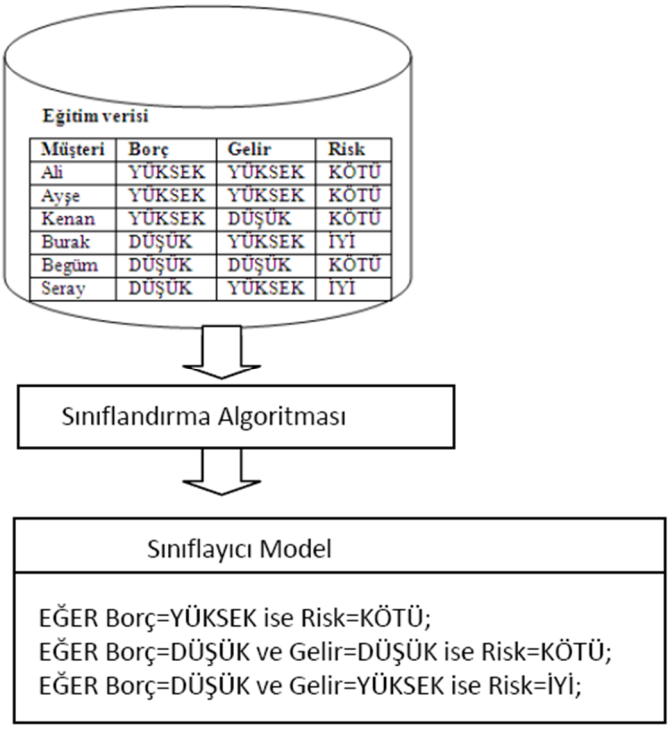
2. BÖLÜM
SINIFLANDIRMA

Verilerin içerdiği ortak özelliklere göre ayrıştırılması sınıflandırma olarak adlandırılmaktadır. Örneğin bir sınıftaki öğrenciler; cinsiyetleri, hangi burca sahip oldukları, yaşadıkları evlerin kira mı yoksa kendilerine mi ait olduğu gibi farklı kriterlere göre sınıflandırılabilir. Sınıflandırma bir öğrenme algoritmasına dayanır. Öğrenmenin amacı, bir sınıflandırma modelinin yaratılmasıdır [32]. Sınıflandırmada modele girdi nitelikleri ile birlikte kesikli formda olan hedef nitelik de, modelin öğrenmesi için sunulur.

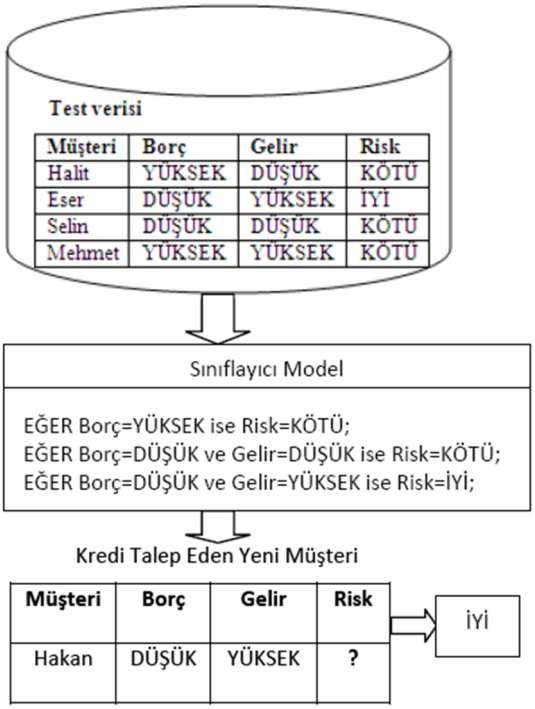
Sınıflandırma süreci iki aşamadan oluşmaktadır. İlk aşama, veri kümesine uygun bir modelin ortaya konulmasıdır. Söz konusu model, veri tabanındaki kayıtların nitelikleri veya bir başka deyişle alan isimleri kullanılarak elde edilir. Sınıflandırma modelinin oluşturulması için veri tabanının bir kısmı eğitim verileri olarak kullanılır. Bu veriler veri tabanından rastgele seçilir. Şekil 2.1'de görüldüğü gibi eğitim verileri üzerinde bir algoritma uygulanarak sınıflandırma modeli elde edilir. İlgili örnekte, müşteri, borç, gelir girdi niteliklerini, risk ise çıktı niteliğini yani sınıfı ifade etmektedir [21].

İkinci aşamada, test verileri üzerinde elde edilen sınıflandırma kuralları test edilir. Örneğin, Şekil 2.2'de gösterilen "Hakan" isimli yeni bir müşterinin kredi talebinde bulunduğu varsayılsın. Bu müşterinin risk durumunu belirlemek için, örnek verilerden elde edilen karar kuralı doğrudan veriye uygulanır. Bu müşteri için Borç=DÜŞÜK, Gelir=YÜKSEK olduğu biliniyorsa Risk=İYİ olduğu anlaşılır.

Yukarıda test sonucunda elde edilen modelin doğru olduğu kabul edilecek olursa, bu model diğer veriler üzerinde uygulanır. Elde edilen sonuç model, mevcut ya da muhtemel müşterilerin gelecekteki kredi risklerini belirlemede kullanılır [32]



Şekil 2.1. Sınıflandırmada model kurma süreci [6]



Şekil 2.2. Sınıflandırmada modelin test süreci [6]

Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler izleyen bölümlerde genel olarak açıklanmıştır [122].

2.1. Karar ağaçları

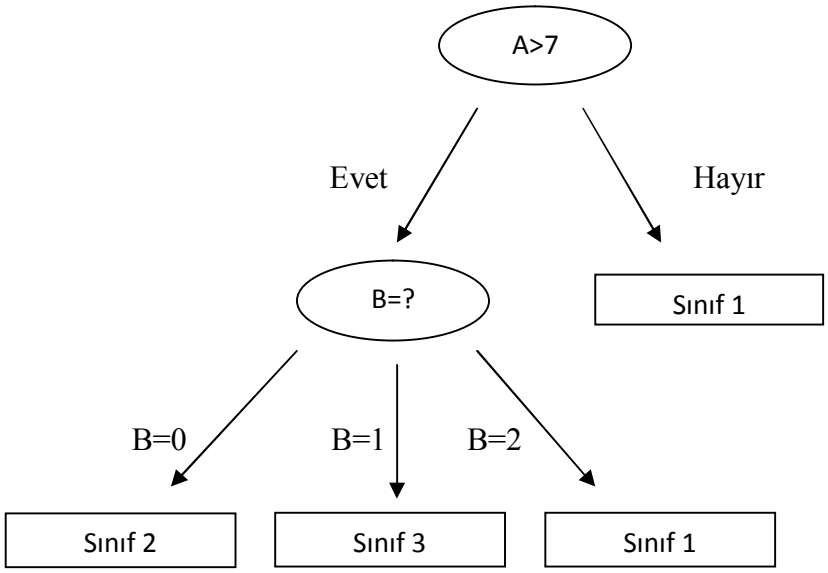
Temelleri AID (Automatic Interaction Detector) yöntemi ile atılan karar ağacı modelleri çeşitli algoritmalar ile sürdürülmüştür. Morgan ve Sonquist adlı araştırmacılar tarafından 1970'li yılların başlarında önerilen ve kullanılan AID algoritması**,** karar ağacı tabanlı ilk algoritma ve yazılımdır. Bilgisayar biliminde veri grubunu bir karar ağacı ile tanımlama işlemi, uygulanan bir yöntem olmasına rağmen, bu yöntem öz bilgiyi elde etmede uzun yıllar tercih edilmemiştir [134].

1984 yılında Berkeley Üniversitesi'nden Leo Breiman ve Charles J. Stone ile Stanford Üniversitesi'nden Jerry Friedman ve R. Olshen tarafından basılan "Classification And Regression Trees" adlı kitapta yeni bir karar ağacı yordamı olan C&RT (Classification and regression trees–sınıflandırma ve regresyon ağaçları) algoritmalarının kullanılmasından bahsedilmektedir. Bu çalışma, yöntemin istatistik biliminde yer edinmesini sağlamıştır. 1986 yılında J.R. Quinlan adlı araştırmacı karar ağaçlarına yeni bir algoritma eklemiştir. Bu karar ağacı algoritması literatüre ID3 algoritması olarak geçmiştir. 1993 yılında ise Quinlan "Programs For Machine Learning" adlı kitabında C4.5 karar ağacı algoritmasını ortaya koymuştur [135].

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 yap "kök" ve bunların arasında kalan yapılar ise "dal" olarak adlandırılır [50].

Karar ağacı yapılarında, her düğüm bir nitelik üzerinde gerçekleştirilen testi, her dal bu testin çıktısını, her yaprak düğüm ise sınıfları temsil eder. En üstteki düğüm kök düğüm olarak adlandırılır. Karar ağaçları, kök düğümden yaprak düğüme doğru çalışır [123]. . Şekil 2.3'de sadece iki niteliğe (A-B) bağlı basit bir karar ağacı görülmektedir.

Şekil 2.3'de görülen basit bir karar ağacı örneğinde elips şeklinde gösterilen A>7 niteliği "kök", B=? gibi elips şeklinde gösterilen B niteliğinin alabileceği 0, 1 ve 2 değerleri "dal" ve dikdörtgenler ile gösterilen Sınıf1, Sınıf2 ve Sınıf3 nitelikleri ise "yaprak" olarak adlandırılır.



Şekil 2.3. A ve B niteliklerine bağlı bir karar ağacı [6]

Sınıflandırmanın yapıldığı karar ağaçlarında önemli konulardan birisi de kökten itibaren dallanmanın (bölümlenmenin) hangi niteliğe göre olacağıdır. Bu amaçla;

- Entropiye dayalı algoritmalar
- Sınıflandırma ve regresyon ağaçları (CART) algoritmaları
- Bellek tabanlı sınıflandırma algoritmaları tercih edilebilir.

Entropiye dayalı bölümlemede "ID3, C4.5 ve C5.0" algoritmaları, Sınıflandırma ve Regresyon ağaçlarında (CART) "Twoing ve Gini" algoritmaları, Bellek tabanlı sınıflandırma yöntemlerine ise "k-en yakın komşu" algoritmaları sayılabilir [32]. Bu algoritmalar birbirlerinden kök, düğüm ve dallanma kriteri seçimlerinde izledikleri yol açısından ayrılırlar [31]. Tablo 2.1, en bilinen karar ağacı algoritmalarının genel özelliklerini göstermektedir.

Tablo 2.1. Karar ağacı algoritmalarının özellikleri [136]

Karar Ağacı Algoritması	**Özellikler**
C&RT	Gini'ye dayalı ikili bölme işlemi mevcuttur. Son veya uç olamayan her bir düğümde iki adet dal bulunmaktadır. Budama işlemi ağacın karmaşıklık ölçüsüne dayanır. Verinin hazırlanmasına gereksinim duyar.

Tablo 2.1. Karar ağacı algoritmalarının özellikleri (devamı) [136]

C4.5 ve C5.0 (ID3 karar ağacı algoritmasının geliştirilmiş versiyonları)	Her düğümden çıkan çoklu dallar ile ağaç oluşturur. Dalların sayısı tahmin edicinin kategori sayısına eşittir. Tek bir sınıflayıcıda birden çok karar ağacını birleştirir. Ayırma işlemi için bilgi kazancı kullanılır. Budama işlemi her yapraktaki hata oranına dayanır**.**
CHAID	Ki-kare testlerini kullanarak bölme işlemini gerçekleştirir. Dalların sayısı iki ile tahmin edicinin kategori sayısı arasında değişir.
SLIQ	Hızlı ölçeklenebilir bir sınıflayıcıdır. Hızlı ağaç budama algoritması mevcuttur.
SPRINT	Büyük veri kümeleri için idealdir. Bölme işlemi tek bir niteliğin değerine dayanır. Tüm bellek sınırlamaları üzerinde nitelik listesi veri yapısı kullanarak işlem yapar.

2.1.1 ID3 algoritması

ID3 algoritması; makine öğrenmesi ve bilişim teorisine dayanarak verilen örnekler içinde, en ayırıcı özelliğe sahip olan değişkeni bulan bir algoritmadır [124]. Bunun için de entropi kavramından yararlanır. Entropi, beklentisizliğin maksimumlaşmasıdır [125]. Başka bir ifadeyle, eldeki bilgilerin sayısallaştırılmasıdır. Dunham [126] entropinin bir veri kümesi içindeki belirsizlik, şaşkınlık ve rastgeleliği ölçmek için kullanıldığını söylemektedir. Eldeki bütün veriler tek bir sınıfa ait olsa, örneğin herkes aynı yaşta olsa, herhangi bir kişiye yaşı sorulduğunda alınacak yanıt şaşırtıcı olmaz; bu durumda entropi sıfır (0) olur. Entropi sayısal olarak [0 –1] aralığında bir değere sahiptir. Tüm olasılıklar (p1,p2,…,pi) eşit olduğunda entropi en yüksek (1) değerine, tüm örnekler aynı sınıfa sahip olduğunda ise en düşük (0) değerine sahip olur.

S isimli bir kaynak veri seti olsun, pi olasılık dağılımına bağlı S kaynağının tamamının entropi hesabı H(P) şu şekilde hesaplanır [127]:

$$H(P) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (2.1)$$

Aynı zamanda S veri kaynağındaki tüm T alt niteliklerinde kendi aralarında entropileri şöyle hesaplanır.

$$H(X, T) = \sum_{i=1}^{n} \frac{|T_i| H(T_i)}{|T|} \quad (2.2)$$

Veritabanı bölünmeden önce doğru sınıflandırma yapmak için gelen bilgiyle H(P), veritabanı bölündükten sonra doğru sınıflandırma için gelen bilgi H(X,T) arasındaki farka kazanç adı verilir. Kazanç (2.3) eşitliği ile hesaplanır.

$$\text{Kazanç}(X, T) = H(P) - H(X, T) \quad (2.3)$$

Burada bölümleme (dallanma) kriter seçimi yapılırken kazanç niteliklerinden en büyüğü öncelikli nitelik olarak seçilir ve seçilen nitelik üzerinden diğer niteliklere ait tekrar H(P) toplam entropi ve H(X,T) alt nitelik entropi hesabı yapılarak yeni kazançlar elde edilir. Bulunan kazançlardan en büyüğü seçilerek yeni alt bölümlemeler (dallanmalar) gerçekleştirilir [6].

Uygulamada Kazanç ölçütü adı verilen yukarıdaki formül yerine daha iyi sonuçlar veren Kazanç Oranı adı verilen formül daha çok kullanılmaktadır [127]. T kümesinde X niteliğinin değerini belirlemekte gereken bilgi miktarını ortaya koymak için bu yol bulunmuştur. Söz konusu bilgi $H(P_{X,T})$ ile ifade edilir. $P_{X,T}$ ifadesi X değerlerinin olasılık dağılımıdır ve şu şekilde hesaplanır [6]:

$$P_{X,T} = \left(\frac{|T_1|}{|T|}, \frac{|T_2|}{|T|}, \ldots, \frac{|T_k|}{|T|}\right) \quad (2.4)$$

Burada H(PX,T) miktarı T kümesindeki X niteliği için bilgi bölünmesidir. Bu değer ise (2.5) eşitliği ile hesaplanır.

$$H(P_{X,T}) = H\left(\frac{|T_1|}{|T|}, \frac{|T_2|}{|T|}, \ldots, \frac{|T_k|}{|T|}\right) \text{ veya } H(P_{X,T}) = -\sum_{i=1}^{k} \frac{|T_i|}{|T|} \log_2\left(\frac{|T_i|}{|T|}\right) \quad (2.5)$$

Yukarıda bulunan $H(P_{X,T})$ değeri ve Kazanç ölçütü yardımıyla da kazanç oranı aşağıdaki gibi hesaplanır:

$$\text{Kazanç Oranı } (X,T) = \frac{Kazanç\ (X,T)}{H(P_{X,T})} \quad (2.6)$$

2.1.2 C4.5 (C5 ve J48) algoritmaları

C4.5 algoritması da ID3 algoritmasında olduğu gibi Quinlann tarafından geliştirilen bir sınıflama algoritmasıdır. ID3 algoritmasından, sayısal değerlere sahip nitelikler ve

bilinmeyen değerlere sahip nitelikler için de geliştirilmiş bir karar ağacı oluşturması bakımından üstünlükleri vardır. Karar ağacı oluştururken kayıp verileri hesaba katmayarak, kazanç oranı hesaplanırken sadece verileri eksik olmayan diğer kayıtlar kullanılır. Böylece daha duyarlı ve daha anlamlı kurallar çıkartabilen bir ağaç üretmiş olur [49].

Kategorik olmayan sayısal değerli niteliklere ilişkin C4.5 karar ağacı modeli oluşturulurken değerleri iki aralığa bölmek için rastgele eşikler bulunmaktadır. En uygun t eşik değerini hesaplamak için birçok yöntem vardır. Eşik değerinin belirlenmesi amacıyla, en büyük bilgi kazancını sağlayacak biçimde bir eşik değer belirlenir. Bu amaçla nitelik değerleri sıralanır ve {v1, v2, …, vn} biçimini alır. Eşik değeri kullanarak nitelik değer iki parçaya ayrılır. Eşik değeri olarak {vi, vi+1] aralığının orta noktası yada aritmetik ortalaması alınabilir. Veri setindeki sayısal bilgileri kategorik biçime dönüştürdükten sonra ID3 algoritmasındaki gibi önce tüm veri setinin entropi hesabı, sonra her nitelik için ayrı entropi hesabı yapılır. Yine her nitelik için ayrı kazanç oranları hesaplandıktan sonra elde edilen kazanç oranlarından en büyük değerli nitelik kök ya da bir sonraki bölümleme niteliği olarak atanır [31].

Veri tabanındaki bilgilerde kayıplar olduğunda C4.5 algoritması bu sorunu gidermek için iki çözüm sunmaktadır. Birincisinde, eğer kayıp veriler veri setindeki verilerin çoğunluğunu kapsamıyorsa kayıp verilerin olduğu kayıtlar veri tabanından çıkarılır ve algoritma geriye kalan veri tabanı üzerine uygulanır. Ancak veri tabanındaki kayıp verilerin sayısı fazla ise o zaman ikinci çözüm kullanılır [6]. İkinci çözüm ile kayıp verilerle de çalışacak bir algoritma uygulanır. Algoritma uygulanmadan önce kayıp verilere sahip örneklerde kazanç ölçütünün hesaplamak için bir F düzeltme faktöründen yararlanılır. Bu amaçla ilk olarak kayıp veriler çıkarılarak H(P) toplam entropi ve H(X,T) sınıflar için entropiler ID3 algoritmasındaki gibi hesaplanır. F faktörü kullanılarak kazanç ölçütü düzeltilir:

$$F=\frac{Veri\ tabanında\ değeri\ bilinen\ niteliğe\ sahip\ örneklerin\ sayısı}{Veri\ tabanındaki\ tüm\ örneklerin\ sayısı} \tag{2.7}$$

Yeni kazanç ölçütü ise (3.8) eşitliği ile belirlenir [21].

$$\text{Kazanç(X)} = F\big(H(T) - H(X,T)\big) \tag{2.8}$$

Karmaşık görünümlü karar ağaçlarında bir alt ağacı atarak yerine bir yaprak (sınıf niteliği) yerleştirmeye karar ağacının budanması adı verilir. Alt ağacın yerine yaprak yerleştirmekle, algoritma öngörülü hata oranını azaltmayı ve sınıflandırma modelinin kalitesini artırmayı amaçlar [128].

C4.5 algoritmasının ID3 algoritmasına göre diğer bir üstünlüğü de doğruluk ölçütü kullanarak karar ağacını budamaktır. Ağaçtaki her düğüm için Ucf üst güven sınırı iki terimli dağılımların istatistiksel tablolarını kullanarak elde edilebilir [128]. Verilen düğümde Ucf parametresi Ti ve E'nin bir fonksiyonudur. C4.5 algoritması %25 güven sınırı kullanır ve verilen her bir düğümdeki Ti, U%25 (|Ti|/E) düğüm yapraklarının güven aralığı ile karşılaştırılır. Her bir yaprakta ağırlıklar durumların toplam sayısıdır. Eğer alt ağaçtaki kök düğümün beklenen hatası, yapraklardaki U%25 toplam ağırlıktan daha küçük ise (alt ağacın beklenen hatası), o zaman alt ağaç yok edilir, yerine kök düğüm konulur. Böylece budanmış ağaçta yeni bir yaprak olarak yer alır [32].

C5 algoritması da C4.5 algoritmasına dayalı geliştirilmiş bir karar ağacı algoritmasıdır. J48 algoritması ise C4.5 karar ağacı algoritmasının WEKA (Waikato Environment for Knowledge Analysis) açık kaynak kodlu paket programına uyarlanmış versiyonudur. J48 algoritması; aynı veri seti üzerinden WEKA paket programında Naive Bayes, Lojistik Regresyon, ID3, JRIP, PART ve Sinir Ağları gibi bilinen sınıflandırma algoritmalarına göre genellikle doğruluğu en yüksek sınıflandırma algoritmasıdır [5].

2.1.3 CART Algoritması

CART algoritması, ağaç yapısına dayalı olarak sınıflandırma ve regresyon modellerinin türetilmesi için yaygın olarak kullanılan bir istatistiksel prosedürdür. CART ağaç modeli, tek değişkenli ikili kararların bir hiyerarşisini içerir. CART verileri iki alt kümeye ayırdığı için her bir alt küme içindeki durumlar, bir önceki alt kümeden daha homojen olacaktır. Bu ardışık süreç, homojenlik kriterine ulaşılıncaya veya diğer bazı durma kriterleri sağlanıncaya değin kendini tekrar eder. Aynı kestirim değişkeni ağaçta farklı düzeylerde pek çok kez kullanılabilir. Ağacın yapısı önceden belirlenmemekte, verilerden türetilmektedir [138]. CART, kök düğümünde, verilerin iki gruba bölünmesi için en iyi değişkenin seçilmesini sağlar ve farklı bölümlendirme kriterleri kullanır. Bu bölümlendirme kriterlerinin tümü, her bir alt kümedeki sınıf etiketlerini mümkün

olduğunca homojen olacak biçimde bölümlendirir [142]. Bölümlendirme prosedürü çocuk düğümlere veya alt düğümlerin her birine ardışık olarak uygulanır [139].

CART ağaçları, kesin bir heterojenlik ölçüsüne bağlı olarak düğümlere ayrılmış iki değerli ağaçlardır ve bu nedenle de sonuçta homojen dallar oluşmaktadır [140]. Ağacın hedefi, benzer veya aynı çıktı değerlerine sahip olma eğiliminde olan alt gruplar yaratmaktır. CART modelleri için bölünmelerin bulunmasında kullanılan dört farklı heterojenlik ölçüsü mevcuttur. Kategorik hedef değişkenler için Gini, Twoing veya (sıralayıcı hedef değişkenleri için) sıralı Twoing, sürekli hedef değişkenler için ise en küçük kareli sapma (LSD) kullanılabilir.

Gini indeksi aşağıdaki şekilde yazılabilir:

$$g(t) = 1 - \sum_j p^2(\frac{j}{t}) \tag{2.9}$$

Her hangi bir düğümde durumlar kategoriler arasında eşit biçimde dağıldığında Gini indeksi 1- (1/k) maksimum değerini alır. Bir düğümdeki durumlar aynı kategoriye ait olduğunda ise Gini indeksi 0' a eşit olacaktır [141].

Twoing indeksi, hedef değişken kategorilerinin iki süper sınıfa bölümlendirilmesine dayalıdır ve ardından bu iki süper sınıfa dayalı olarak kestirim değişkenindeki en iyi bölünmeyi bulur. t düğümünde s bölünmesi için Twoing kriter fonksiyonu şu şekilde tanımlanabilir [138]:

$$\Phi(s,t) = p_L\, p_R * \left[\sum_j \left| p\left(\frac{j}{t_L}\right) - p(\frac{j}{t_R})\right|\right] \text{^2} \tag{2.10}$$

Fonksiyonda yer alan t_L ve t_R, s bölünmesi tarafından yaratılan düğümleri göstermektedir. s bölünmesi, bu kriteri maksimize eden bölünme olarak belirlenir. İki süper sınıf olan Cl ve C2 aşağıdaki biçimde tanımlanabilir:

$$C_1 = \left\{j: p\left(\frac{j}{t_L}\right) \geq p(\frac{j}{t_R})\right\} \qquad C_2 = C - C_1 \tag{2.11}$$

Burada C, hedef değişkenin kategori kümesidir.

Sıralı Twoing indeksi, sıralayıcı hedef değişkenleri için Twoing indeksinin değiştirilmiş şeklidir. Sıralı Twoing kriterindeki farklılık yalnızca bitişik kategorilerin süper sınıflar ile birleştirilmesidir. Örneğin bir değişkenin 4 kategorisi olsun. Twoing kriteri 1 ve 4'ü

bir süper sınıf ve 2 ve 3' ü de diğer bir süper sınıf olarak belirlemiş olsun. Bununla beraber kategoriler sıralı olduğundan 1 ve 4 kategorileri birleştirilemez çünkü bunlar bitişik kategoriler değillerdir. Sıralı Twoing indeksi bu durumu göz önüne aldığından 1 ve 4 gibi kategoriler bitişik olmadığından birleştirilemez [137].

Sürekli hedef değişkenleri için en küçük kareli sapma (LSD) heterojenlik ölçüsü kullanılmaktadır. LSD ölçüsü R(t), t düğümü için basit (ağırlıklandırılmış) düğüm içi varyansıdır ve düğüm için risk tahminine eşittir. R(t)'nın formülü aşağıdaki şekildedir [138]:

$$R(t)=\frac{1}{N_w(t)}\sum_{i\varepsilon t} W_n f_n(y_l - \bar{y}(t))^2 \quad (2.12)$$

Nw(t), t düğümündeki ağırlıklandırılnış durum sayısı, Wn i durumu için mevcut ise ağırlıklandırılmış değişken değeri, f_n mevcut ise frekans değişkeninin değerini, y_l hedef değişkenin değerini ve $\bar{Y}(t)$ ise t düğümü için ağırlıklı ortalamayı göstermektedir.

Sonuçta elde edilen ağacın büyüklüğü, karmaşık budama sürecinin bir sonucudur. Çok büyük bir ağaç, uyumun üzerinde ve çok küçük ağaç, yetersiz tahmin gücüne sahip olacaktır. Ağaç yapısının hiyerarşik formu, CART gibi algoritmaları ağaç yapısına dayanmayan diğer sınıflandırma algoritmalarından açık bir şekilde ayırır [137].

2.1.4 REPTree algoritması

Açılımı azaltılmış hata ayıklama ağacıdır (reduced error pruning tree). REPTree algoritmasının temeli entropi ile bilgi kazancının hesaplanması ve varyanslardan kaynaklanan hataların azaltılması prensibine dayanır [150]. Bu metot ilk defa Quinlan tarafından önerilmiştir [151]. Bu yöntem yardımıyla karar ağacı modelinin karmaşıklığı, "azaltılmış hata ayıklama yöntemi" ve varyanslardan kaynaklanan hatanın azaltılması ile azaltılır [185,186]. Karar ağacı böl ve yönet yaklaşımını doğrulayan veri yapısı ile oluşturulur ve denetimli öğrenme için kullanılır.

2.1.5 NBTree algoritması

NBTree algoritması, Naive Bayes sınıflandırıcı ve C4.5 karar ağacı algoritmasının kombinasyonundan oluşan bir algoritmadır ve öğrenilen bilgi bir ağaç formunda temsil edilir [152]. Bu ağaç yinelemeli inşa edilir. Entropi değerini sınırlamak amacıyla sürekli

nitelikler için bir eşik belirlenir. NBTree algoritması kesikli çıktı değerlerine sahip değişkenlerde daha iyi sonuç vermektedir. Eğer veri boyutu arttırılırsa algoritmanın performansı da artar. Ancak sürekli değerlere sahip değişkenler kullanıldığı zaman, değişken etkileşimlerini dikkate almadığından iyi sonuç vermeyebilir. Veri boyutu çok yüksek olduğunda karar ağaçları iyi performans göstermemesine rağmen, NBTree algoritması bu eksikliklerin üstesinden gelir [151].

2.1.6 PART algortiması

PART algoritması, doğru kuralları elde etmek amacıyla küresel optimizasyon gerçekleştirmeden sınıflandırma yapan basit bir algoritmadır [21]. PART algoritması C4.5 ve RIPPER kural öğreniminin bir kombinasyonudur. Bu temel "ayır ve fethet" prosedürünü kullanarak sınırsız bir karar listesi oluşturur. "Ayır ve yönet" komutu ile bir ağaç oluşturmak için C4.5 prosedürlerini kullanarak her iterasyonda kısmi bir C4.5 karar ağacı oluşturarak en iyi düğümü kurala dönüştürür [153].

2.1.7 JRip algoritması

JRip algoritrması 1995 yılında W. W. Cohen tarafından hayata geçirilmiştir [1]. Algoritmada "hata azaltma amaçlı tekrarlanan artımlı budama" (RIPPER) adında kural öğrenen bir önerme uygulanmıştır. RIPPER kural öğrenimini uygulayan Cohen, bu yolla bireysel kuralları değiştirerek veya revize ederek kuralların doğruluğunu artırmaya çalışmıştır [154].

2.2. K-NN (En Yakın Komşuluk) Algoritması

KNN, eğitimli öğrenme algoritmasıdır ve amacı, yeni bir örnek geldiğinde var olan öğrenme verisi üzerinde sınıflandırma yapmaktır. Algoritma, yeni bir örnek geldiğinde, onun en yakın K komşusuna bakarak örneğin sınıfına karar verir [143].

Bu algoritma ile yeni bir vektörü sınıflandırabilmek için doküman vektörü ve eğitim dokümanları vektörleri kullanılır. Tüm eğitim dökümanları ve kategorisi belirlenecek olan doküman vektörel olarak ifade edildikten sonra bu vektörler K-NN algoritması ile karşılaştırılırlar [145].

Bu yöntem örnek kümedeki gözlemlerin her birinin, sonradan belirlenen bir gözlem değerine olan uzaklıklarının hesaplanması ve en küçük uzaklığa sahip k sayıda gözlemin seçilmesi esasına dayanmaktadır. Uzaklıkların hesaplanmasında, i ve j noktaları için Eşitlik (3.13)'de verilen Öklid uzaklık formülü kullanılabilir [32].

$$D(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2.13)$$

Algoritmada ilk olarak, K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır. Söz konusu nokta ile diğer tüm noktalar arasındaki uzaklıklar tek tek hesaplanır. Yukarıda hesaplanan uzaklıklara göre satırlar sıralanır ve bunlar arasından en küçük olan k tanesi seçilir. Seçilen satırların hangi kategoriye ait oldukları belirlenir ve en çok tekrar eden kategori değeri seçilir. Seçilen kategori, tahmin edilmesi beklenen gözlem değerinin kategorisi olarak kabul edilir [32].

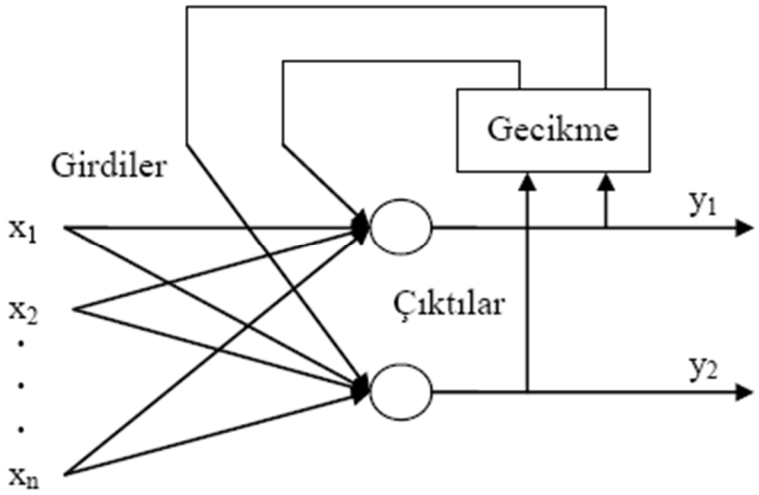
Uygulamada K-NN algoritmasının çok kullanılmasındaki nedenler aşağıdaki gibi sıralanabilir:

- Uygulanabilirliği basit bir algoritma olması.
- Gürültülü eğitim dokümanlarına karşı dirençli olması.
- Eğitim dokümanları sayısı fazla ise etkili olması
- Bu metot ölçeklendirilebilir bir metottur ve çok geniş veri tabanları üzerinde de uygulanabilir [144].

2.3. Yapay Sinir Ağları

Gizli katmana sahip olmayan ve girdi ve çıktı katmanları olmak üzere iki tip katmana sahip yapay sinir ağlarına perceptron adı verilmektedir. Bir yapay sinir ağı ise perceptron modelinden daha karmaşık bir yapıya sahiptir [129] . Yapay sinir ağları, katmanları arasındaki bağlantıların tiplerine göre genel olarak, ileri beslemeli (feedforward) ve yinelemeli (recurrent) olmak üzere iki kategoride sınıflandırılırlar.

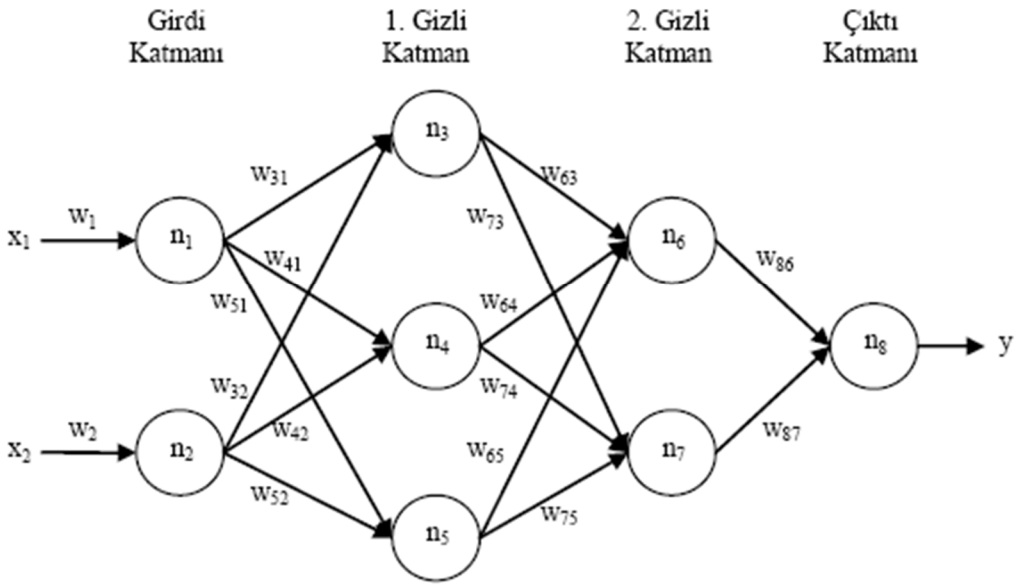
Yapay sinir ağında döngüsel yol oluşturan bir geri besleme (feedback) hattı varsa, bu ağ yinelemeli ağ olarak adlandırılır. Bu yapıda genellikle senkronizasyonu sağlamak amacıyla bir gecikme (delay) bileşeni de kullanılmaktadır. Şekil 3.4'de bir yinelemeli yapay sinir ağı mimarisi görülmektedir [128].



Şekil 2.4. Yinelemeli yapay sinir ağı mimarisi [132]

Yinelemeli yapay sinir ağları dinamik yapılı ağlardır ve mimarileri statik olanlara göre geri besleme içermelerinden dolayı temel olarak farklıdır. Yinelemeli ağlar, ileri beslemeli olanlara göre daha küçük bir mimariye sahip olmaktadırlar [130].

Yapay sinir ağlarında en yaygın kullanılan mimari, yapılarındaki esneklik, iyi temsil yetenekleri ve geniş sayıdaki öğrenme algoritmaları ile ileri beslemeli ağlardır [131] . İleri beslemeli yapay sinir ağları (Feedforward Neural Networks) bir girdi katmanı, bir veya daha fazla gizli katmanı ve bir de çıktı katmanı olan ağlardır. Girdi katmanı, eğitim seti girdilerini alarak ağırlıklandırır ve çıktılarını kendisinden sonra gelen gizli katmana eş zamanlı olarak gönderir. Bir gizli katmanın çıktıları, kullanılan gizli katman sayısına bağlı olarak bir sonraki gizli katmanın girdileri olabilir. Gizli katman sayısı analizcinin takdirine göre seçilebilir, ancak uygulamada genellikle tek gizli katman kullanılmaktadır. Son gizli katmanın ağırlıklandırılmış çıktıları, çıktı katmanının girdileridir ve bu katman çıktı olarak yapay sinir ağının verilen eğitim seti için yaptığı kestirimi verir [21]. Şekil 3.5'de ileri beslemeli 2 gizli katmana sahip bir yapay sinir ağı gösterilmektedir.



Şekil 2.5. İleri beslemeli yapay sinir ağı [132]

2.4. Naive-Bayes Sınıflandırma Algoritması

Naive Bayes sınıflandırma algoritması, adını Matematikçi Thomas Bayes'den alır ve sınıflandırma/ kategorilendirme algoritmasıdır. Naive Bayes sınıflandırması olasılık ilkelerine göre tanımlanmış hesaplamalar ile sisteme sunulan verilerin sınıfını yani kategorisini tespit etmeyi amaçlar.

Naive Bayes sınıflandırmasında sisteme belirli bir oranda öğretilmiş veri sunulur (Örn: 100 adet). Öğretim için sunulan verilerin mutlaka bir sınıfı/kategorisi bulunmalıdır. Öğretilmiş veriler üzerinde yapılan olasılık işlemleri ile sisteme sunulan yeni test verileri, daha önce elde edilmiş olasılık değerlerine göre işletilir ve verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır. Elbette öğretilmiş veri sayısı ne kadar çok ise, test verisinin gerçek kategorisini tespit etmek o kadar kesin olabilmektedir [133].

2.5. WEKA Veri Madenciliği Paket Programı

Weka, makine öğrenimi amacıyla Waikato Üniversitesinde geliştirilmiş ve "Waikato Environment for Knowledge Analysis" kelimelerinin baş harflerinden oluşmuş veri madenciliği yazılımın ismidir [71]. Java dilinde geliştirilmiş olması ve kütüphanelerinin

jar dosyaları halinde geliyor olması sayesinde, JAVA dilinde yazılan projelere kolayca entegre edilebilmesi, kullanımını daha da yaygınlaştırmıştır [67]. Weka, iş zekası alanında en çok kullanılan 10 yazılımdan birisi olup, yine iş zekası konusunda en çok kullanılan özgür yazılımlar sıralamasında ilk 3 sırada yer almaktadır [68].

Weka, tamamen modüler bir tasarıma sahip olup, içerdiği özelliklerle veri kümeleri üzerinde görselleştirme, veri analizi, iş zekası uygulamaları, veri madenciliği gibi işlemler yapabilmektedir. Weka yazılımı, kendisine özgü olarak bir *.arff* uzantısı desteği ile gelmektedir. Ancak Weka yazılımının içerisinde CSV dosyalarını da ARFF formatına çevirmeye yarayan araçlar mevcuttur. Temel olarak aşağıdaki 3 Veri Madenciliği işlemi Weka ile yapılabilir [196]:

- Sınıflandırma (Classification)
- Kümeleme (Clustering)
- İlişkilendirme (Association)

Ayrıca yukarıdaki işlemlere ilave olarak, veri kümeleri üzerinde ön ve son işlemler yapılabilir.

- Veri Ön işleme
- Görselleştirme

Son olarak Weka Kütüphanesinde veri kümelerini içeren dosyalar üzerinde çalışan çok sayıda hazır fonksiyon bulunmaktadır.

Weka paket programının ismi, 200'e yakın IEEE makalesinde ve 5,700 civarında ACM makalesinde ismi doğrudan geçmektedir [69]. 2005 yılında, Weka, dünyanın en prestijli veri madenciliği ödüllerinden olan SIGKDD "Data Mining and Knowledge Discovery Service Award", ödülüne layık görülmüştür [70].

WEKA çalıştırıldıktan sonra Şekil 3.6'da görüldüğü gibi, Application menüsünde çalışılabilecek modlar listelenmektedir. Bunlar komut modunda çalışmayı sağlayan Simple CLI, projeyi adım adım görsel ortamda gerçekleştirmeyi sağlayan Explorer ve projeyi sürükle bırak yöntemiyle gerçekleştirmeyi sağlayan KnowledgeFlow seçenekleridir [47].

Şekil 2.6. WEKA'da applications menüsü [47]

Weka veri madenciliği yazılımında bünyesinde onlarca sınıflandırma algoritması barındırmaktadır. Bu algoritmalardan bazılarının genel özellikleri aşağıda verilmektedir.

ZeroR: 0-R sınıflandırıcısını kullanarak, nümerik sınıflar için ortalamayı, nominal sınıf değeri için ise modu hesaplayıp sınıflandırma işlemi yapar. Numerik, nominal, string, data, ikili ve tekli verilerde çalışır.

PART: Karar listesi üreterek sınıflandırma yapar. "ayır ve yönet" kuralını kullanır. Her iterasyonda kısmi C4.5 karar ağacı oluşturur ve en iyi yaprağı kurala dönüştürür. Numerik, nominal, data, ikili ve tekli verilerde çalışır.

DecisionTable: Basit bir karar tablosu çoğunluk sınıflandırıcı kullanarak sınıflandırma yapar. Numerik, nominal, data, ikili ve tekli verilerde çalışır.

Id3: Budanmamış karar ağacına dayalı ID3 algoritmasını kullanarak sınıflandırma yapar. Sadece nominal niteliklerle çalışır. Eksik verilere izin verilmez. Boş yapraklar sınıflandırılmamış verilere neden olabilir. Nominal, ikili ve tekli verilerde çalışır.

J48: Budanmış ya da budanmamış C4.5 karar ağacı oluşturarak sınıflandırma yapar. Numerik, nominal, data, ikili ve tekli verilerde çalışır.

NBTree: Naive Bayes sınıflandırıcı ile yapraklarda bir karar ağacı üreterek sınıflandırma yapar. Numerik, nominal, data, ikili ve tekli verilerde çalışır.

REPTree: Hızlı bir karar ağacı öğrenicisidir. Bilgi kazancı/varyansı kullanıp bir karar/regresyon ağacı oluşturarak, onu indirgenmiş hata budama yoluyla budar. Yalnızca numerik niteliklerin değerlerini sıralar. Eksik değerler ilgili niteliklerin parçalara bölünmesiyle ele alınır.

JRip: Bu algoritma William W. tarafından önerilen "tekrarlı artırımlı budama ile hata azaltımı üretme" (RİPPER) kural öğrenicisi ile önerme uygular. Numerik, nominal, data, ikili ve tekli verilerde çalışır.

SimpleCart: Minimum maliyet-karmaşıklık budaması uygular. Algoritma eksik değerleri ile uğraşırken yedek bölünme metodu yerine kademeli (kesirli) örnekler metodunu kullanır. Numerik, nominal, ikili ve tekli verilerde çalışır.

3. BÖLÜM

GEREÇ VE YÖNTEM

Bu çalışmanın amacı, insanların genel memnuniyet ve umut düzeyini etkileyen faktörleri belirleyerek ve bu faktörlerden hangilerinin ne derece etkili olduğunu ortaya koyarak sınıflandırma yapmaktır. Bu amaçla, TÜİK'ten 2012 yılına ait Yaşam Memnuniyeti Anketi mikro verileri talep edilmiş ve bu veriler üzerinde çalışmalar gerçekleştirilmiştir.

3.1. Veri Kümesinin Oluşturulması

Yaşam Memnuniyeti Araştırması (YMA), bireylerin genel mutluluk algısını, toplumsal değer yargılarını, temel yaşam alanlarındaki genel memnuniyetini ve kamu hizmetlerinden memnuniyetini ölçmek, memnuniyet düzeylerinin zaman içindeki değişimini takip etmek amacıyla, ilki 2003 yılında Hane halkı Bütçe Anketi'ne ek bir modül olmak üzere 2004 yılından itibaren düzenli olarak TUİK tarafından gerçekleştirilmektedir. Anket kapsamında, örnekleme grubunda yer alan hanelerdeki 18 yaş ve üzerindeki kişiler ile görüşme yapılmaktadır.

TUİK tarafından yapılan YMA'da Hane halkı ve Fert bazında olmak üzere 2 farklı anket uygulanmaktadır. Çalışma kapsamında ele alınan 2012 yılı Hane halkı anketinde 4069 haneye 50 soru, Fert anketinde de bu hanelerde yaşayan 18 yaş ve üzerindeki 7956 kişiye 221 soru sorulmuştur (EK-1). Bu anketlere ilişkin mikro veriler elde edilmiştir.

3.2. Veri Ön İşleme

3.2.1 Veri Analizi ve Veri Temizleme

Çalışma kapsamında ele alınan YMA'da sınıflandırma yapılacak hedef değişkeni/değişkenleri belirleyebilmek için TUİK'de görevli uzman kişilerin görüşlerine başvurulmuş ve sınıflandırma yapılabilecek üç hedef değişken belirlenmiştir. Bu değişkenler B39-Umut düzeyi, B13_1-Sağlık hizmetlerinden memnuniyet ve B13_2-Asayiş hizmetlerinden memnuniyettir.

B39-Umut düzeyi

Uzman bilgisi neticesinde, Fert anketinden umut düzeyi ile alakalı ankette yer alan genel memnuniyet soruları dışındaki diğer soruların genel memnuniyet veya umut düzeyi durumunu öğrenme amaçlı sorulmadığı ve bu hedef değişken için kullanılmasının tutarsız sonuçlar vereceği kanaatine varılmıştır. Ayrıca, Hane halkı Anketi kapsamında sorulan H17-Hanenin aylık toplam geliri ve H18-Bu gelirle hanenin temel ihtiyaçlarını karşılama düzeyi değişkenlerinin de kişilerin umut düzeyini etkilediği belirlenmiş ve bu nedenle ilgili hedef değişken için yapılacak sınıflandırma çalışmasında girdi değişkeni olarak ele alınmıştır.

B39-Umut düzeyi hedef değişkenine ait girdilerden B03-Son bir hafta içinde ücretli ya da ücretsiz olarak bir işte çalışma durumu değişkeni ile ilgili tutarsız 44 veri sınıflandırma işlemi dışında tutulmuştur. Bu tutarsızlık kişinin önceki sorulara "çalıştı" olarak cevap verip, sonraki soruda "çalışmama nedeni" sorusuna cevap verilmesinden kaynaklanmaktadır.

B13_1-Sağlık hizmetlerinden memnuniyet ve B13_2-Asayiş hizmetlerinden memnuniyet

Fert anketinden Sağlık hizmetlerinden memnuniyet ve Asayiş hizmetlerinden memnuniyet hedef değişkenleri için uzman kişi, ankette yer alan soruların girdi değişkeni olarak yeterli olmayacağı konusunda uyarmıştır. Ancak bu hedef değişkeni alanlarının çok geniş bir yelpazede düşünülebileceği ve bu anketleri de uygulayabilmenin mümkün olamayacağı konusunda fikir birliğine varılarak mevcut girdi değişkenleri ile sınıflandırma yapılmasına karar verilmiştir.

B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkeni ile ilgili, B26-Bu sağlık kuruluşunu neden tercih ediyorsunuz?, B27_1-Muayene ve tahlil için randevu almakta sorun yaşıyor musunuz?, B27_2-Temizlik/hijyen konusunda sorun var mı?, B27_3-Yapılan muayeneden memnun musunuz?, B27_4-Doktorların hastalara davranışında sorun var mı?, B27_5-Hemşirelerin/hastabakıcıların hastalara davranışında sorun var mı?, B27_6-Doktor ve sağlık personeli sayısı yeterli mi?, B27_7-Muayene ve tahlil ücretlerini yüksek buluyor musunuz?, B27_8-İlaç fiyatlarında sorun görüyor musunuz?, B27_9-Muayene ve tahlil için sıra beklemede sorun var mı?, B27_10-Muayene ve katkı payı ücreti ödemeyi sorun olarak görüyor musunuz? ve B28-2012 yılı içinde sağlık hizmeti aldınız mı? girdi değişkenindeki 43 boş değer, kişiler şimdiye kadar sağlık kuruluşlarına hiç başvurmadığı için veri kümesinden temizlenmiştir.

B13_1-Sağlık hizmetlerinden memnuniyet ve B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenlerine ait girdilerden B03-Son bir hafta içinde ücretli ya da ücretsiz olarak bir işte çalışma durumu değişkeni ile ilgili tutarsız 44 veri sınıflandırma işlemi dışında tutulmuştur. Bu tutarsızlık kişinin önceki sorulara "çalıştı" olarak cevap verip, sonraki soruda "çalışmama nedeni" sorusuna cevap verilmesinden kaynaklanmaktadır.

3.2.2 Veri Bütünleştirme ve Eksik Veriler

B39-Umut düzeyi

Girdi değişkenlerinden B03-Son bir hafta içinde ücretli veya ücretsiz olarak bir işte çalışma durumu, B04-Çalışmama nedeni ve B05-Çalışılan iş/işyerinin kamu, özel olma durumu değişkenleri ortak bilgi içermektedir. Bu nedenle bu değişkenler, bütünleştirme işlemi yapılarak B03-Son bir hafta içinde ücretli veya ücretsiz olarak bir işte çalışma durumu değişkeni adı altında toplanmıştır. Tablo 3.1 ilgili değişkene uygulanan bütünleştirme işlemini özetlemektedir.

B06-Çalışılan işteki durum, B07-İşyerinin iktisadi faaliyet kodu ve B11_7-İşten elde ettiğiniz kazançtan memnuniyet girdi değişkenlerindeki 4564 boş değere, anket düzenlenen kişinin B03-Son bir hafta içinde ücretli ya da ücretsiz olarak bir işte çalışma durumuna "çalışmadı" cevabını vermelerinden dolayı "çalışmadı" değeri atanmıştır. B11_2-Evlilikten memnuniyet değişkenindeki boş değerlere, evli olmayanların soruya

cevap vermemesinden dolayı "evli değil" değeri atanmıştır. B11_3-Şimdiye kadar alınan eğitimden memnuniyet değişkenindeki boş değerlere, eğitim almayanların soruya cevap vermemesinden dolayı "eğitim almadı" değeri verilmiş, B11_6-İşten memnuniyet ve B12_4-İşyerindeki kişilerle ilişkilerden memnuniyet değişkenlerindeki boş değerlere, çalışmayanların soruya cevap vermemesinden dolayı "çalışmıyor" değeri atanmıştır.

Tablo 3.1. B03-Son bir hafta içinde ücretli veya ücretsiz olarak bir işte çalışma durumu değişkeni için bütünleştirme işlemi

Bütünleştirilen Değişken Kodu	Orijinal Değişken Değerleri	Bütünleştirilen Değişken Değerleri	İlişki Kurulan Değişkenler
B03	Çalıştı	Özel	B03-B05
		Kamu	
	Çalışmadı fakat işi ile ilgisi devam ediyor	Çalışmadı fakat işi ile ilgisi devam ediyor	-
	Çalışmadı	İş bulamama	B03-B04
		Mevsimlik çalışıyor	
		Eğitim/Öğretime devam ediyor	
		Ev işleri ile meşgul	
		Emekli	
		Özürlü veya hasta	
		Yaşlı	
		İrad sahibi	
		Diğer	

B13_1-Sağlık hizmetlerinden memnuniyet

B29-En son sağlık hizmeti alımı sırasında herhangi bir sorun yaşadınız mı? ve B30-2012 yılında en son sağlık hizmeti aldığınız sağlık kuruluşu hangisidir? girdi değişkenlerine ait 1830 boş değer, 2012 yılında sağlık hizmeti alınmadığından dolayı "sağlık hizmeti almadı" olarak doldurulmuştur.

B13_2-Asayiş hizmetlerinden memnuniyet

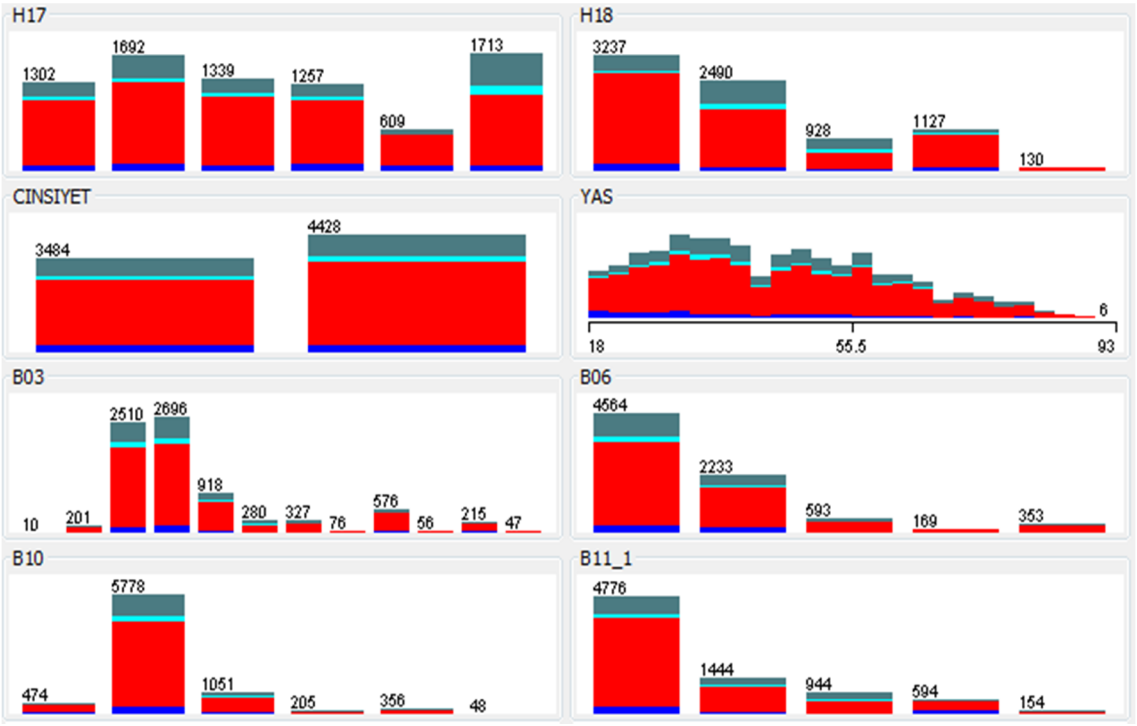
B13_2 hedef değişkenine ait girdi değişkenlerinde ayrıca bir bütünleştirme işlemine ihtiyaç duyulmamıştır.

3.2.3 Aşırı Uç veriler

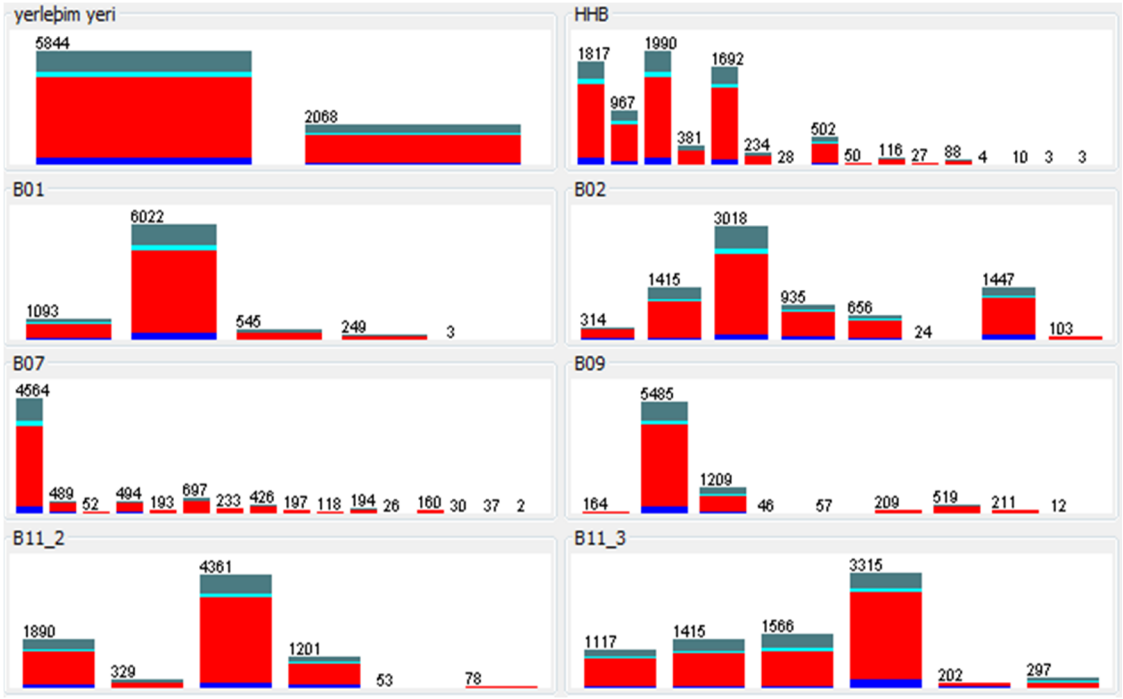
Veri kümelerinde değişkenlerin aşırı uç değer alıp almadıklarını analiz yapabilmek için değişkenler için histogramlar oluşturulmuştur.

B39-Umut düzeyi

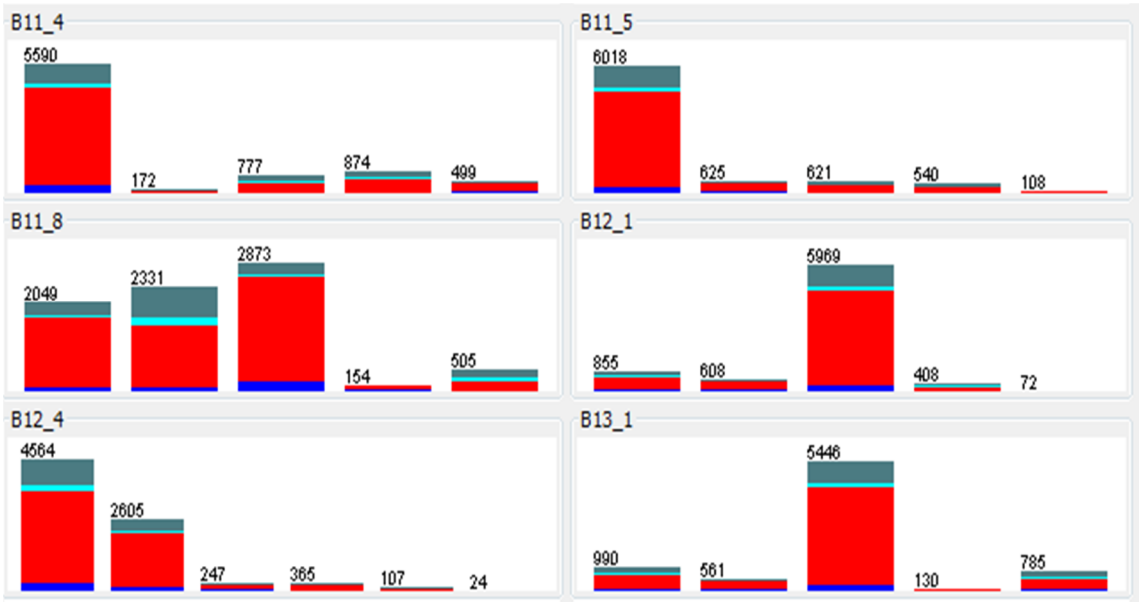
Şekil 3.1-3.6 B39-Umut düzeyi hedef değişkenine ait veri kümesinde yer alan değişkenler için oluşturulan histogramları göstermektedir.



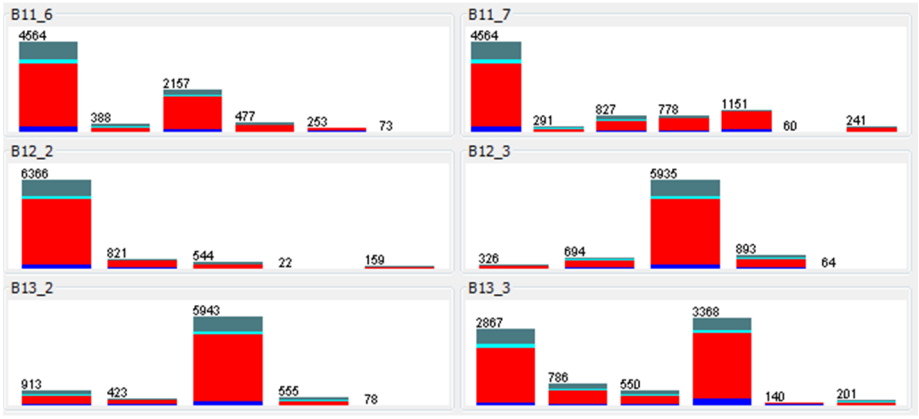
Şekil 3.1. B39-Umut düzeyi hedef değişkenine ait girdi değişken histogramları-1



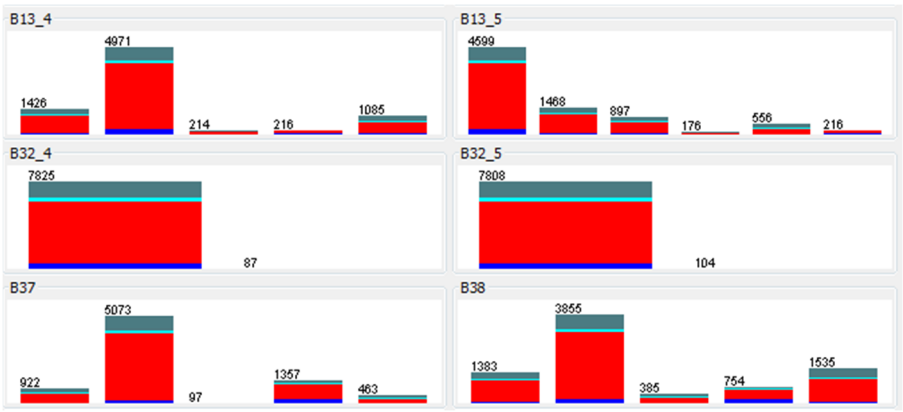
Şekil 3.2. B39-Umut düzeyi hedef değişkenine ait girdi değişken histogramları-2



Şekil 3.3. B39-Umut düzeyi hedef değişkenine ait girdi değişken histogramları-3



Şekil 3.4. B39-Umut düzeyi hedef değişkenine ait girdi değişken histogramları-4



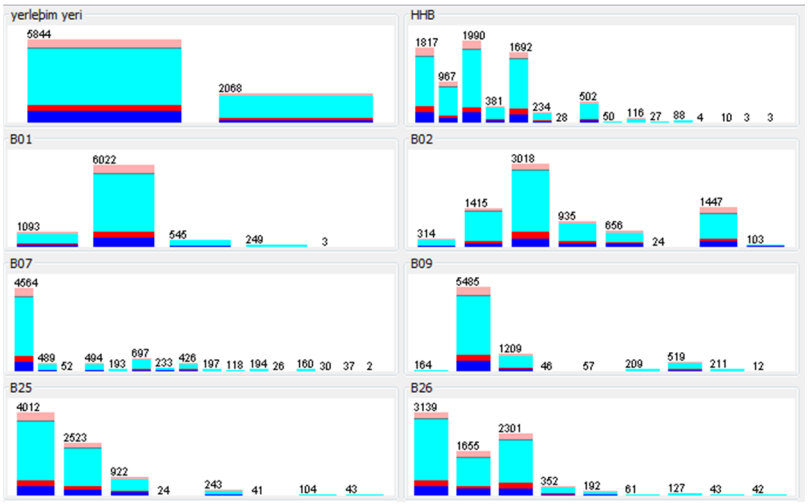
Şekil 3.5. B39-Umut düzeyi hedef değişkenine ait girdi değişken histogramları-5

Şekil 3.6. B39-Umut düzeyi hedef değişkenine ait girdi değişken histogramları-6

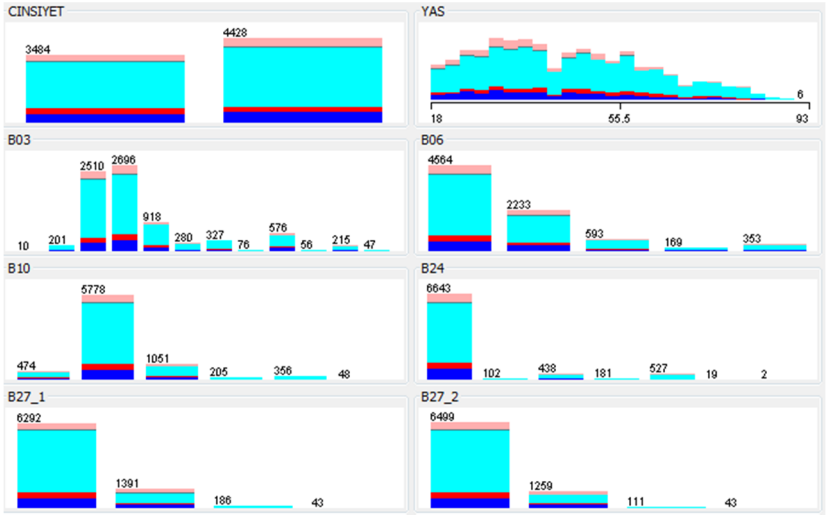
Oluşturulan histogramlar incelendiğinde aşağı yönde veya yukarı yönde hiçbir aşırı uç değere rastlanmamıştır.

B13_1-Sağlık hizmetlerinden memnuniyet

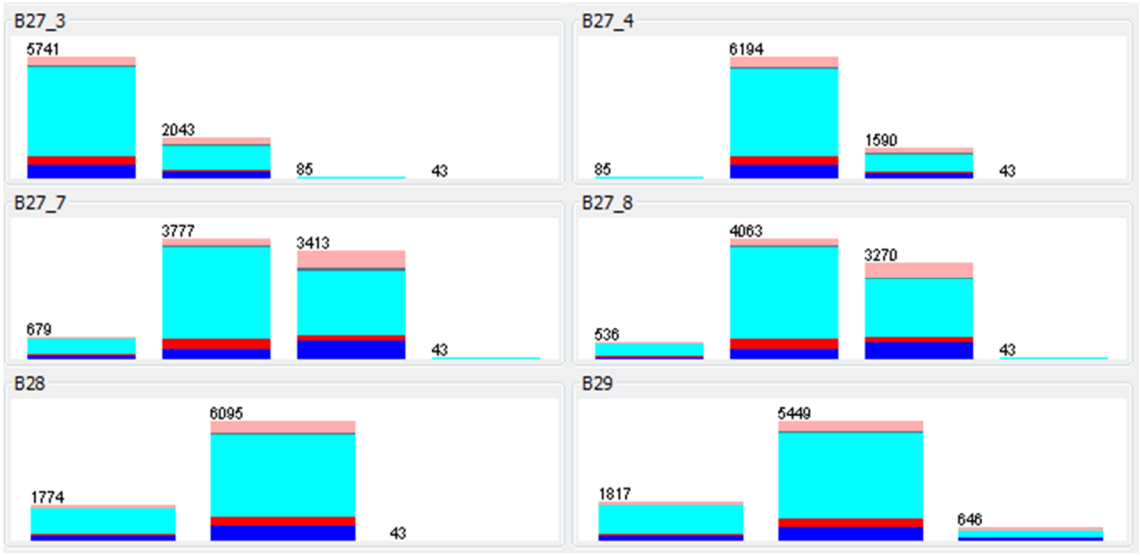
B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait veri kümesinde yer alan değişkenler için oluşturulan histogramlar, Şekil 3.7-3.10 'da gösterilmektedir.



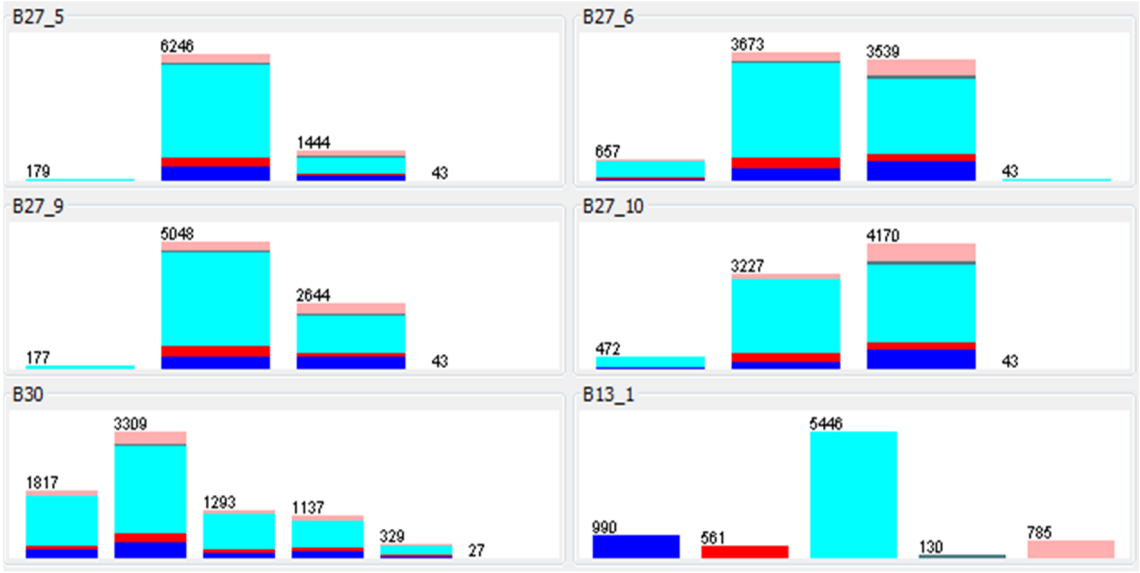
Şekil 3.7. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-1



Şekil 3.8. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-2



Şekil 3.9. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-3

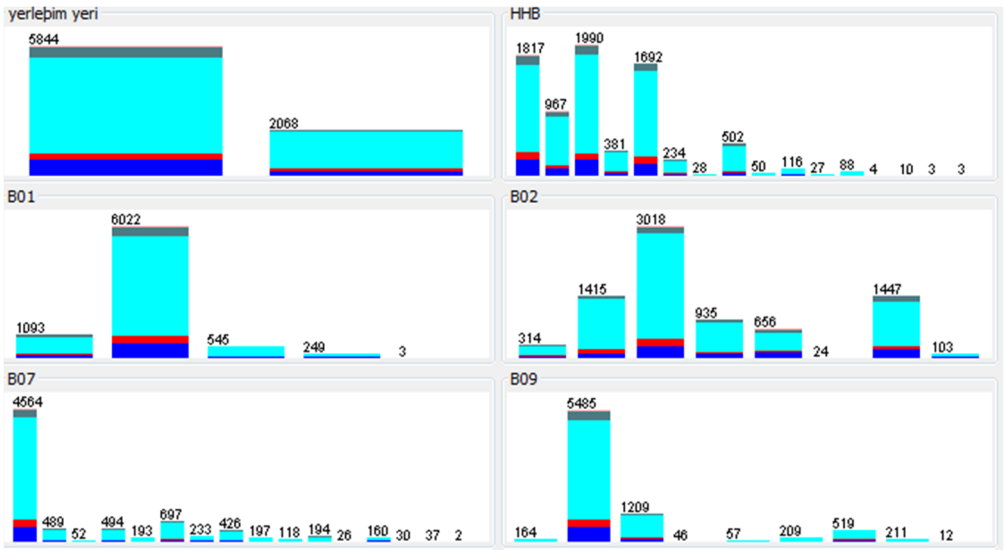


Şekil 3.10. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-4

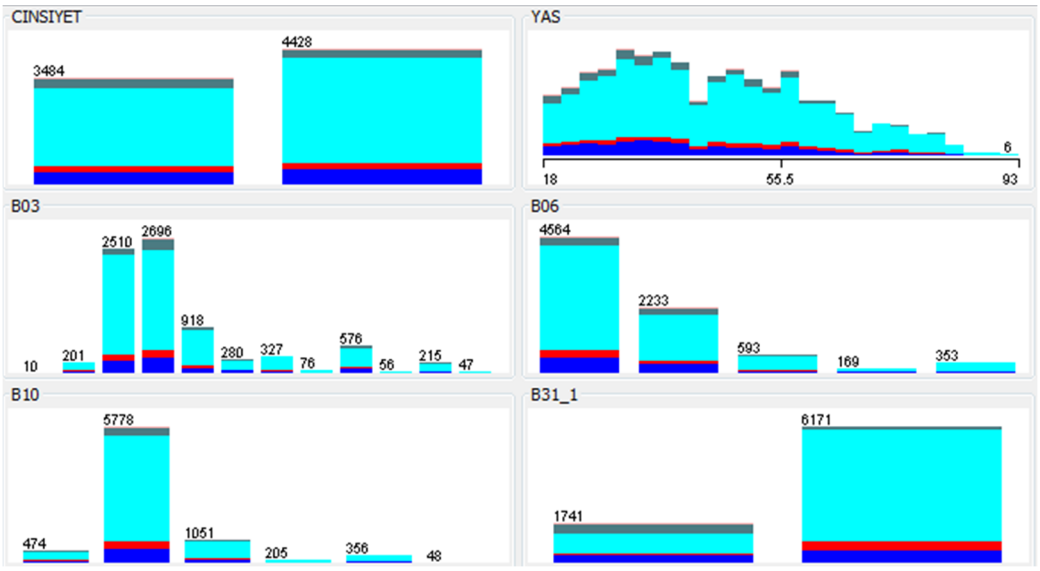
B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait oluşturulan histogramlar incelendiğinde, hiçbir aşırı uç değere rastlanmamıştır.

B13_2-Asayiş hizmetlerinden memnuniyet

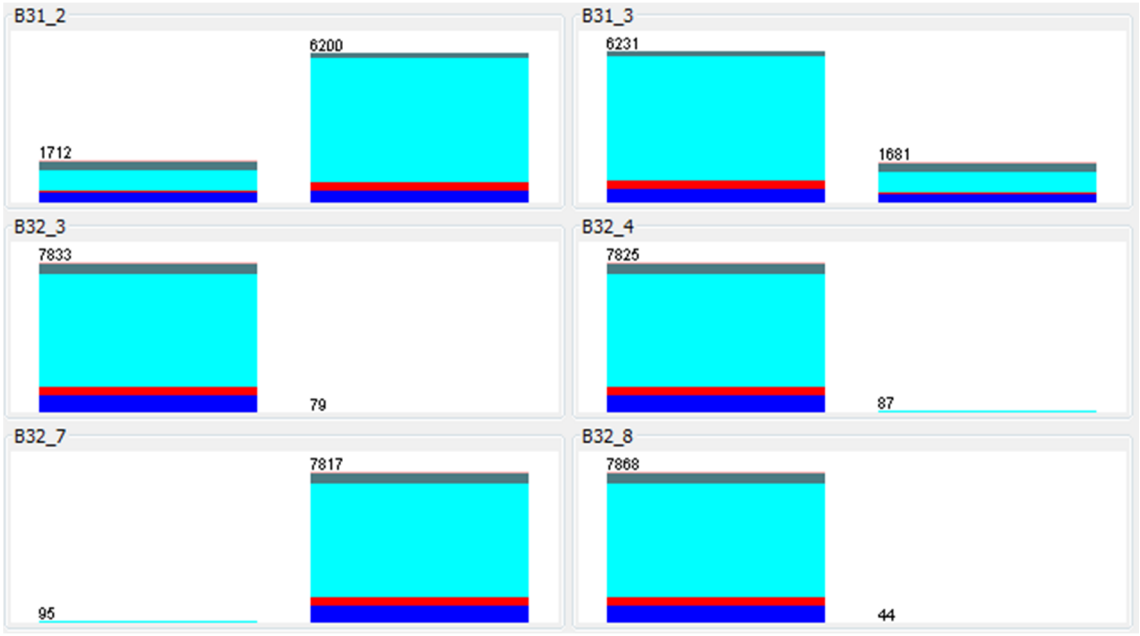
B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenine ait veri kümesinde yer alan değişkenler için oluşturulan histogramlar, Şekil 3.11-3.14 ‘de gösterilmektedir.



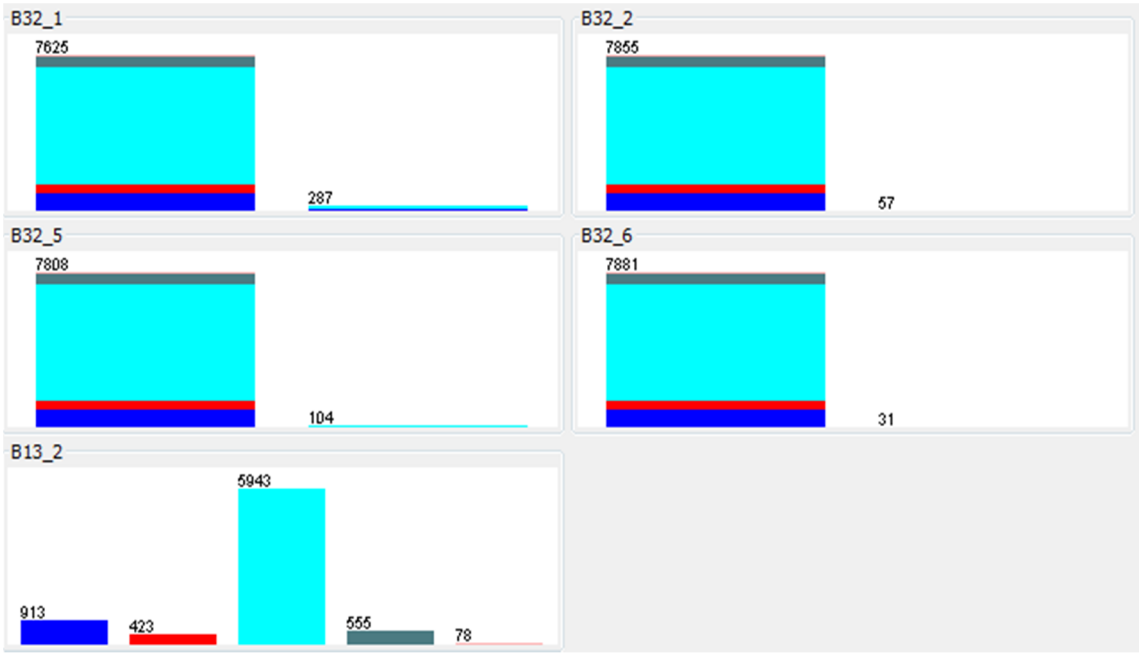
Şekil 3.11. B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-1



Şekil 3.12. B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-2



Şekil 3.13. B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-3



Şekil 3.14. B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenine ait girdi değişken histogramları-4

Yine ele alınan B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkeni ile ilgili girdi değişkenleri histogramları incelendiğinde herhangi bir aşırı uç değere rastlanmadığı görülmektedir.

3.3. Veri Ön İşleme Sonucu Elde Edilen Veri Kümeleri

Tablo 3.2, 3.3 ve 3.4 sırasıyla B39-Umut düzeyi, B13_1-Sağlık hizmetlerinden memnuniyet ve B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenleri için veri ön işleme sonucu elde edilen değişkenleri ve değişken açıklamalarını göstermektedir.

Tablo 3.2. B39-Umut düzeyi hedef değişkeni için değişken tanımlamaları

Değişken Kodu	Değer Aralığı veya Kategorik değer	Değişken tipi	Değişken Açıklaması
B39	[1,2,3,4]	Nominal	Umut düzeyi
Yerleşim yeri	[1,2]	Nominal	Yerleşim yeri
HHB	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]	Nominal	Hane halkı büyüklüğü
CINSIYET	[1,2]	Nominal	Cinsiyet
YAS	$18 \leq x \leq 93$	Nümerik	Yaş
B01	[1,2,3,4,99]	Nominal	Medeni durum
B02	[1,2,3,4,5,6,7,8]	Nominal	Eğitim durumu
B03	[1,2,3,4,5,6,7,8,9,10,11,12]	Nominal	Çalışma durumu
B06	[1,2,3,4,5]	Nominal	Çalışılan işteki durum
B07	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]	Nominal	İşyerinin faaliyet kodu
B09	[1,2,3,4,5,6,7,8,98]	Nominal	Hayatta en çok kimin mutlu ettiği
B10	[1,2,3,4,5,98]	Nominal	Hayatta en çok neyin mutlu ettiği
B11_1	[1,2,3,4,5]	Nominal	Sağlıktan memnuniyet
B11_2	[1,2,3,4,5,6]	Nominal	Evlilikten memnuniyet
B11_3	[1,2,3,4,5,6]	Nominal	Alınan eğitimden memnuniyet
B11_4	[1,2,3,4,5]	Nominal	Oturulan konuttan memnuniyet

Tablo 3.2. B39-Umut düzeyi hedef değişkeni için değişken tanımlamaları (devamı)

B11_5	[1,2,3,4,5]	Nominal	Oturulan semtten memnuniyet
B11_6	[1,2,3,4,5,6]	Nominal	İşten memnuniyet
B11_7	[1,2,3,4,5,6,7]	Nominal	Çalışılan işin kazancından memnuniyet
B11_8	[1,2,3,4,5]	Nominal	Aylık hane halkı gelirinden memnuniyet
B12_1	[1,2,3,4,5]	Nominal	Akrabalarla ilişkiden memnuniyet
B12_2	[1,2,3,4,5]	Nominal	Arkadaşlarla ilişkiden memnuniyet
B12_3	[1,2,3,4,5]	Nominal	Komşularla ilişkiden memnuniyet
B12_4	[1,2,3,4,5,6]	Nominal	İş arkadaşları ile ilişkilerden memnuniyet
B13_1	[1,2,3,4,5]	Nominal	Sağlık hizmetlerinden memnuniyet
B13_2	[1,2,3,4,5]	Nominal	Asayiş hizmetlerinden memnuniyet
B13_3	[1,2,3,4,5,6]	Nominal	Adli hizmetlerden memnuniyet
B13_4	[1,2,3,4,5]	Nominal	Eğitim hizmetlerinden memnuniyet
B13_5	[1,2,3,4,5,6]	Nominal	SGK hizmetlerinden memnuniyet
B13_6	[1,2,3,4,5,6]	Nominal	Ulaştırma hizmetlerinden memnuniyet
B19	[11,12,13,14,2]	Nominal	Faydalanılan Sosyal Güvenlik Kurumu
B32_4	[1,2]	Nominal	Aile fertlerinden kötü muamele gördünüz mü?
B32_5	[1,2]	Nominal	Şantaj ya da tehdit olayı yaşadınız mı?
B32_6	[1,2]	Nominal	Cinsel suçlardan mağduriyet yaşadınız mı?
B32_7	[1,2]	Nominal	Dolandırıcılıktan mağduriyet yaşadınız mı?
B37	[1,2,3,4,5]	Nominal	Evinizde yalnız otururken kendinizi ne kadar güvende hissediyorsunuz?

Tablo 3.2. B39-Umut düzeyi hedef değişkeni için değişken tanımlamaları (devamı)

B38	[1,2,3,4,5]	Nominal	Gece yalnız yürürken kendinizi ne kadar güvende hissediyorsunuz?
H17	[1,2,3,4,5,6]	Nominal	Hanenin aylık toplam net geliri
H18	[1,2,3,4,5]	Nominal	Bu gelirle hanenin ihtiyaçlarının karşılama düzeyi

Tablo 3.3. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkeni için değişken tanımlamaları

Değişken Adı	Değer Aralığı veya Kategorik değer	Değişken tipi	Değişken Açıklaması
B13_1	[1,2,3,4,5]	Nominal	Sağlık hizmetlerinden memnuniyet
Yerleşim yeri	[1,2]	Nominal	Yerleşim yeri
HHB	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]	Nominal	Hane halkı büyüklüğü
CINSIYET	[1,2]	Nominal	Cinsiyet
YAS	18≤x≤93 arası	Numerik	Yaş
B01	[1,2,3,4,99]	Nominal	Medeni durum
B02	[1,2,3,4,5,6,7,8]	Nominal	Eğitim durumu
B03	[1,2,3,4,5,6,7,8,9,10,11,12]	Nominal	Çalışma durumu
B06	[1,2,3,4,5]	Nominal	Çalışılan işteki durum
B07	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]	Nominal	İşyerinin faaliyet kodu
B09	[1,2,3,4,5,6,7,8,98]	Nominal	Hayatta en çok kimin mutlu ettiği
B10	[1,2,3,4,5,98]	Nominal	Hayatta en çok neyin mutlu ettiği
B24	[1,2,3,4,5,90,98]	Nominal	Hastalandığınızda masraflarınızı hangi kanalla karşılıyorsunuz?

Tablo 3.3. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkeni için değişken tanımlamaları (devamı)

B25	[1,2,3,4,5,6,7,8]	Nominal	Hastalandığınızda ilk nereye gidersiniz?
B26	[1,2,3,4,5,6,7,98,99]	Nominal	Bu sağlık kuruluşunu neden seçiyorsunuz?
B27_1	[1,2,3,99]	Nominal	Muayene ve tahlillerde randevu almakta sorun yaşıyor musunuz?
B27_2	[1,2,3,99]	Nominal	Temizlik/hijyen konusunda sorun var mı?
B27_3	[1,2,3,99]	Nominal	Yapılan muayeneden memnun musunuz?
B27_4	[1,2,3,99]	Nominal	Doktorların hastalara davranışında sorun var mı?
B27_5	[1,2,3,99]	Nominal	Hemşire/Hasta bakıcıların hastalara davranışında sorun var mı?
B27_6	[1,2,3,99]	Nominal	Doktor / sağlık personeli sayısı yeterli mi?
B27_7	[1,2,3,99]	Nominal	Muayene ve tahlil ücretlerini yüksek buluyor musunuz?
B27_8	[1,2,3,99]	Nominal	İlaç fiyatlarında sorun görüyor musunuz?
B27_9	[1,2,3,99]	Nominal	Muayene ve tahlil için sıra beklemede sorun var mı?
B27_10	[1,2,3,99]	Nominal	Muayene için katkı payı ödemeyi sorun olarak görüyor musunuz?

Tablo 3.3. B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkeni için değişken tanımlamaları (devamı)

B28	[1,2,99]	Nominal	2012 yılı içinde sağlık hizmeti aldınız mı?
B29	[1,2,99]	Nominal	En son sağlık hizmeti alımı sırasında herhangi bir sorun yaşadınız mı?
B30	[1,2,3,4,5,99]	Nominal	2012 yılında en son sağlık hizmeti aldığınız sağlık kuruluşu hangisidir?

Tablo 3.4. B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkeni için değişken tanımlamaları

Değişken Adı	Değer Aralığı veya Kategorik değer	Değişken tipi	Değişken Açıklaması
13_2	[1,2,3,4,5]	Nominal	Asayiş hizmetlerinden memnuniyet
Yerleşim yeri	[1,2]	Nominal	Yerleşim yeri
HHB	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]	Nominal	Hane halkı büyüklüğü
CINSIYET	[1,2]	Nominal	Cinsiyet
YAS	18≤x≤93 arası	Numerik	Yaş
B01	[1,2,3,4,99]	Nominal	Medeni durum
B02	[1,2,3,4,5,6,7,8]	Nominal	Eğitim durumu
B03	[1,2,3,4,5,6,7,8,9,10,11,12]	Nominal	Çalışma durumu
B06	[1,2,3,4,5]	Nominal	Çalışılan işteki durum
B07	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16]	Nominal	İşyerinin faaliyet kodu
B09	[1,2,3,4,5,6,7,8,98]	Nominal	Hayatta en çok kimin mutlu ettiği
B10	[1,2,3,4,5,98]	Nominal	Hayatta en çok neyin mutlu ettiği

Tablo 3.4. B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkeni için değişken tanımlamaları

B31_1	[1,2]	Nominal	Polis veya jandarma olaylara zamanında müdahale ediyor mu?
B31_2	[1,2]	Nominal	Polis ve jandarmanın vatandaşa davranışından memnun musunuz?
B31_3	[1,2]	Nominal	Polis veya jandarmanın verdiği trafik hizmetinden memnun musunuz?
B32_1	[1,2]	Nominal	Kapkaç, yankesicilik vb. hırsızlık olayı yaşadınız mı?
B32_2	[1,2]	Nominal	Gasp olayı yaşadınız mı?
B32_3	[1,2]	Nominal	Yaralanma, darp olayı yaşadınız mı?
B32_4	[1,2]	Nominal	Aile fertlerinden kötü muamele gördünüz mü?
B32_5	[1,2]	Nominal	Şantaj, tehdit olayı yaşadınız mı?
B32_6	[1,2]	Nominal	Cinsel suçlardan mağduriyet yaşadınız mı?
B32_7	[1,2]	Nominal	Dolandırıcılıktan mağdur oldunuz mu?
B32_8	[1,2]	Nominal	Başka bir suçtan mağduriyet yaşadınız mı?

3.4. Sınıflandırma

3.4.1 Verinin eğitim ve test kümelerine parçalanması

Sınıflandırmada ele alınan veri kümesi eğitim kümesi ve test kümesi olmak üzere iki parçaya ayrılır. Eğitim kümesi sınıflandırıcının öğrenmesi için kullanılırken, test kümesi sınıflandırma sonucu elde edilen modelin değerlendirilmesi için kullanılır. Verinin eğitim ve test kümelerine parçalanması ile ilgili literatürde en çok kullanılan iki yöntem "% ayrım" ve "k-katlı çapraz doğrulama"dır.

% ayrım, tüm veriyi tek bir eğitim ve test kümesine böler. Yöntemde verinin çoğunluğu (genellikle 2/3) eğitim kümesine atanırken, kalan kısmı (1/3) test kümesi olarak belirlenir. Dolayısıyla model hep aynı eğitim kümesi üzerinde öğrenir ve aynı test kümesi üzerinde test edilir.

k-katlı çapraz doğrulama yöntemi ise, tüm veriyi kullanarak sınıflandırma yapmayı sağlayan bir metottur. Yöntemde, seçilen k değerine göre k adet grup oluşturulur ve her seferinde gruplardan bir tanesi test kümesi olarak belirlenerek, oluşturulan modeli doğrulamak için kullanılır; kalan k-1 grup ise eğitim kümesi olarak belirlenerek, model bu küme üzerinde inşa edilir. Süreç her defasında k parçadan farklı birini test, kalanları eğitim kümesine atayarak k defa tekrarlanır ve doğruluk oranı, k adet doğruluk oranının ortalaması olarak sunulur. k-katlı çapraz doğrulama yönteminin avantajı veri üzerindeki varyasyonu azaltması ve veri rastgeleliği sağlamasıdır.

k-katlı çapraz doğrulamanın avantajlarından dolayı tez çalışmasında veri kümesinin eğitim ve test kümelerine parçalanmasında k-katlı çapraz doğrulama kullanılmıştır. Verinin parçalanacağı k değeri olarak ise hem veri kümesinin yüksek boyutlu olması, hem de literatürde en çok tercih edilmesi nedeniyle k=10 değeri tercih edilmiştir.

3.4.2 Performans ölçütleri

Tez çalışması kapsamında sınıflandırma algoritmaları ile elde edilen sonuçların performansını değerlendirmek için Weka 3.6.9 veri madenciliği yazılımı tarafından da performans ölçütü olarak sunulan doğruluk, kappa istatistiği, doğru pozitif (TP) oranı, yanlış pozitif (FP) oranı, kesinlik, duyarlılık, F-ölçütü, ROC-alanı ve karışıklık matrisi kullanılmıştır.

Test Doğruluğu: Sınıflandırıcının hiç görmediği örnekler yani test veri kümesi üzerinde elde ettiği doğruluk yüzdesini vermektedir. Test doğruluğu Eşitlik 3.1 ile hesaplanmaktadır.

$$Test\ Doğruluğu = \frac{Test\ kümesinde\ doğru\ sınıflandırılan\ örnek\ sayısı}{Test\ kümesindeki\ örnek\ sayısı} \quad (3.1)$$

Kappa İstatistik: Kappa istatistik değeri, karşılaştırmalı uyuşmanın güvenilirliğini ölçen bir istatistik yöntemdir ve [0,1] arasında değer alır. 0 ya da 0'a yakın bir değer alması sınıflar arası uyuşmanın olmadığını, 1'e yakın değer alması sınıfların birbirleriyle uyuştuğunu ifade eder. İki sınıf arasındaki karşılaştırmalı uyuşmanın güvenilirliğinde Cohen'in Kappa katsayısı, ikiden fazla sabit sayıda sınıf için uyuşmanın güvenilirliğinde Fleiss'in Kappa katsayısı kullanılır. Fleiss'in Kappa katsayısı eşitlik *3.2'deki* gibi formüle edilir:

$$K = \frac{\bar{P} - \overline{P_e}}{1 - \overline{P_e}} \quad (3.2)$$

(3.2) eşitliğindeki 1-$\overline{P_e}$ faktörü rastgelelik ötesinde ne derece uyuşma olabileceğinin mümkün olduğunu gösterir, $\bar{P}$- $\overline{P_e}$ gözlemlenenin gerçekte ne derecede rastgelelik ötesinde uyuşma derecesinin ortaya çıktığını açıklar.

Karışıklık Matrisi: Her bir satırı gerçek sınıftaki örnekleri temsil ederken, her bir sütunu tahmin sınıfındaki örnekleri temsil eden ve tablo düzeni kurarak sınıf performanslarını görsel olarak sunan matristir. İki sınıflı bir problemde sınıflandırma sonucu oluşturulan karışıklık matrisi Tablo 3.5'te gösterilmektedir. Tabloda tp (doğru pozitif) ve tn (doğru negatif) doğru sınıflandırmaları; fp (yanlış pozitif) ve fn (yanlış negatif) yanlış sınıflandırmaları ifade etmektedir.

Tablo 3.5. İki sınıflı problemler için karışıklık matrisi

		Tahmin edilen	
		negatif	pozitif
gerçek	negatif	a (TN)	b (FP)
	pozitif	c (FN)	d (TP)

Doğru Pozitif (TP) Oranı: Doğru pozitif oranı, koşul ifadesi doğru iken sonuç (sınıf) ifadesinin de doğru olma yüzdesini ifade eder. Duyarlılık değerine eşittir. Aşağıdaki Eşitlik 3.3 gibi formüle edilir:

$$TP\ oranı = \frac{sınıf\ x\ olarak\ sınıflandırılanların\ miktarı}{toplamdaki\ gerçek\ sınıf\ x\ miktarı} = \frac{tp}{tp+fn}$$

(3.3)

Yanlış Pozitif (FP) Oranı: Koşul ifadesi doğru iken sonuç ifadesinin yanlış olma yüzdesini ifade eder. Eşitlik 3.4'teki gibi formüle edilir:

$$FP\ oranı = \frac{sınıf\ x\ olarak\ yanlış\ sınıflandırılanların\ miktarı}{toplamdaki\ gerçek\ sınıf\ x\ miktarı} = \frac{fp}{fp+tn} \quad (3.4)$$

Kesinlik: Hassasiyet veya doğruluk değerini ifade eder ve bu değer ne kadar yüksek ise model o kadar etkindir. Eşitlik 3.5'deki gibi formüle edilir:

$$Kesinlik = \frac{doğru\ olarak\ sınıflandırılan\ x\ örneklerinin\ miktarı}{toplam\ x\ olarak\ sınıflandırılan\ miktar} = \frac{tp}{tp+fn} \quad (3.5)$$

F-ölçütü: Yüksek olması arzu edilen "kesinlik" ve "Duyarlılık" değerlerinin kombinasyonundan oluşan bir değerdir. Eşitlik 3.6'daki gibi formüle edilir:

$$F-\text{ölçütü} = \frac{2*Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \quad (3.6)$$

ROC alanı: ROC alanı değeri modelin ne kadar iyi bir sonuç verdiğini gösteren bir değerdir ve 0,7 ile 1 arasında bir değer alması beklenir. 1- özgüllük (specificity) olarak hesaplanır. Özgüllük eşitlik 3.7'deki gibi formüle edilir:

$$\text{Özgüllük} = \frac{TN}{TN+FP} = \frac{a}{a+b} = \frac{doğru\ negatiflerin\ sayısı}{doğru\ negatiflerin\ sayısı+\ yanlış\ pozitiflerin\ sayısı}$$

(3.7)

3.4.3 Nitelik Seçimi Yapmadan Sınıflandırma

Sınıflandırma yapmak amacıyla Weka 3.6.9 veri madenciliği yazılımı kullanılmıştır. Oluşturulan karar ağaçları ve kurallar tablosunu yorumlayabilmek amacıyla yazılımda mevcut sınıflandırma algoritmalarından PART, JRip, J48, REPTree ve SimpleCart algoritmaları seçilmiştir.

3.4.3.1 B39-Umut Düzeyi Hedef Değişkeni için Sınıflandırma

B39-Umut düzeyi hedef değişkenine ait veri ön işleme sonucu elde edilen veri kümesi 4 sınıflı bir problemdir. 7912 örnek ve 38 adet girdi değişkeni içermektedir. Aşağıda ele alınan sınıflandırma algoritmaları ile tüm veri kümesi üzerinde 10-katlı çapraz doğrulama ile elde edilen sonuçlar verilmektedir.

J48 algoritması ile sınıflandırma

Şekil 3.15 B39-Umut düzeyi veri kümesinde J48 sınıflandırma algoritması ile elde edilen sonuçları göstermektedir.

```
=== Summary ===

Correctly Classified Instances        5425               68.5667 %
Incorrectly Classified Instances      2487               31.4333 %
Kappa statistic                          0.1205
Mean absolute error                      0.2138
Root mean squared error                  0.3546
Relative absolute error                 92.1124 %
Root relative squared error            104.1018 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.066     0.011      0.284     0.066     0.107      0.646    1.0
                 0.928     0.808      0.731     0.928     0.818      0.609    2.0
                 0.093     0.019      0.184     0.093     0.124      0.531    4.0
                 0.131     0.056      0.353     0.131     0.191      0.603    3.0
Weighted Avg.    0.686     0.58       0.607     0.686     0.623      0.607

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   33  439    4   23 |    a = 1.0
   53 5163   77  269 |    b = 2.0
    8  244   33   68 |    c = 4.0
   22 1215   65  196 |    d = 3.0
```

Şekil 3.15. J48 algoritması ile B39-Umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde J48 algoritması ile % 68,57 doğruluk ile sınıflandırma gerçekleştirilmiş ve 778 kural oluşmuştur.

Kappa istatistik değeri, 0'a yakın bir değer olduğundan sınıf değerleri uyuşmamaktadır. Karışıklık matrisi incelendiğinde birinci sınıfa ait 499 verinin sadece 33'ünün, dördüncü sınıfa ait 353 verinin sadece 33'ünün ve üçüncü sınıfa ait 1498 verinin sadece 196'sının doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerler, veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5562 verinin 5163'ü yine ikinci sınıfa ait olarak doğru sınıflandırılmış, 53'ü birinci sınıfa, 77'si dördüncü sınıfa ve 269'u üçüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

JRip algoritması ile elde edilen sonuçlar

Şekil 3.16 B39-Umut düzeyi veri kümesinde JRip sınıflandırma algoritması ile elde edilen sonuçları göstermektedir.

```
=== Summary ===

Correctly Classified Instances        5524               69.818  %
Incorrectly Classified Instances      2388               30.182  %
Kappa statistic                          0.0727
Mean absolute error                      0.2256
Root mean squared error                  0.3408
Relative absolute error                 97.2258 %
Root relative squared error            100.0513 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.08      0.01       0.36      0.08      0.131      0.55     1.0
                 0.966     0.912      0.715     0.966     0.822      0.529    2.0
                 0.04      0.005      0.264     0.04      0.069      0.534    4.0
                 0.065     0.021      0.418     0.065     0.112      0.529    3.0
Weighted Avg.    0.698     0.646      0.616     0.698     0.61       0.531

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   40  453    2    4 |    a = 1.0
   63 5373   21  105 |    b = 2.0
    1  312   14   26 |    c = 4.0
    7 1378   16   97 |    d = 3.0
```

Şekil 3.16. JRip algoritması ile B39-Umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde JRip algoritması ile % 69,82 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 18 kural oluşmuştur.

Kappa istatistik değeri 0'a yakın bir değer olduğundan sınıf değerleri uyuşmamaktadır. Karışıklık matrisi incelendiğinde, birinci sınıfa ait 499 verinin sadece 40'ının, dördüncü sınıfa ait 353 verinin sadece 14'ünün ve üçüncü sınıfa ait 1498 verinin sadece 97'sinin doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5562 verinin 5373'ü yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 63'ü birinci sınıfa, 21'i dördüncü sınıfa ve 105'i üçüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

PART algoritması ile elde edilen sonuçlar

Şekil 3.17 B39-Umut düzeyi veri kümesinde PART algoritması ile elde edilen sonuçları göstermektedir.

```
=== Summary ===

Correctly Classified Instances        5072               64.1052 %
Incorrectly Classified Instances      2840               35.8948 %
Kappa statistic                          0.1595
Mean absolute error                      0.2049
Root mean squared error                  0.3811
Relative absolute error                 88.2961 %
Root relative squared error            111.9022 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.158     0.036      0.226     0.158     0.186      0.651    1.0
                 0.819     0.644      0.75      0.819     0.783      0.622    2.0
                 0.116     0.029      0.159     0.116     0.134      0.567    4.0
                 0.266     0.131      0.322     0.266     0.292      0.597    3.0
Weighted Avg.    0.641     0.481      0.61      0.641     0.623      0.617

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   79  355   10   55 |    a = 1.0
  198 4553  131  680 |    b = 2.0
   15  193   41  104 |    c = 4.0
   57  966   76  399 |    d = 3.0
```

Şekil 3.17. PART algoritması ile B39-Umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde PART algoritması ile % 64,11 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 765 kural oluşmuştur. Kappa istatistik değeri 0'a yakın bir değer olduğundan sınıf değerleri uyuşmamaktadır. Karışıklık matrisi kısmı incelendiği zaman birinci sınıfa ait 499 verinin sadece 79'unun, dördüncü sınıfa ait 353 verinin sadece 41'inin ve üçüncü sınıfa ait 1498 verinin sadece 399'sinin doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. ikinci sınıfa ait 5562 verinin 4553'ü yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 198'i birinci sınıfa, 131'i dördüncü sınıfa ve 680'i üçüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

REPTree algoritması ile elde edilen sonuçlar

Şekil 3.18 B39-Umut düzeyi veri kümesinde REPTree algoritması ile elde edilen sonuçları göstermektedir.

```
=== Summary ===

Correctly Classified Instances        5475               69.1987 %
Incorrectly Classified Instances      2437               30.8013 %
Kappa statistic                          0.0835
Mean absolute error                      0.2105
Root mean squared error                  0.3358
Relative absolute error                 90.6865 %
Root relative squared error             98.587  %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.052     0.008      0.299     0.052     0.089      0.715    1.0
                 0.95      0.878      0.719     0.95      0.819      0.669    2.0
                 0.045     0.006      0.246     0.045     0.077      0.664    4.0
                 0.098     0.041      0.359     0.098     0.154      0.673    3.0
Weighted Avg.    0.692     0.626      0.603     0.692     0.614      0.672

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   26  458    2   13 |    a = 1.0
   51 5286   24  201 |    b = 2.0
    2  286   16   49 |    c = 4.0
    8 1320   23  147 |    d = 3.0
```

Şekil 3.18. REPTree algoritması ile B39-Umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde REPTree algoritması ile % 69,2 doğruluk, 417 kural ile elde edilmiştir.

Kappa istatistik değeri sınıfların uyuşmadığını göstermektedir. Karışıklık matrisi incelendiğinde, birinci sınıfa ait 499 verinin sadece 26'sının, dördüncü sınıfa ait 353 verinin sadece 16'sının ve üçüncü sınıfa ait 1498 verinin sadece 147'sinin doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5562 verinin 5286'sı yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 51'i birinci sınıfa, 24'ü dördüncü sınıfa ve 201'i üçüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

SimpleCart algoritması ile elde edilen sonuçlar

Şekil 3.19 B39-Umut düzeyi veri kümesinde SimpleCart algoritması ile elde edilen sonuçları göstermektedir.

```
=== Summary ===

Correctly Classified Instances        5563               70.3109 %
Incorrectly Classified Instances      2349               29.6891 %
Kappa statistic                          0.0108
Mean absolute error                      0.2306
Root mean squared error                  0.3396
Relative absolute error                 99.3677 %
Root relative squared error             99.7114 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.513    1.0
                 0.996     0.988      0.705     0.996     0.826      0.524    2.0
                 0         0          0         0         0          0.527    4.0
                 0.014     0.004      0.438     0.014     0.027      0.529    3.0
Weighted Avg.    0.703     0.695      0.578     0.703     0.585      0.525

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
    0  498    0    1 |    a = 1.0
    0 5542    0   20 |    b = 2.0
    0  347    0    6 |    c = 4.0
    0 1477    0   21 |    d = 3.0
```

Şekil 3.19. SimpleCart algoritması ile B39-Umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde SimpleCart algoritması ile % 70,31 doğruluk ile sınıflandırılma gerçekleştirilmiş ve algoritma tek kural oluşturarak tüm örnekleri en çok rastlanan ikinci sınıf değerine atamıştır.

Kappa istatistik değeri 0'a yakın bir değer olduğundan sınıf değerleri uyuşmamaktadır. Karışıklık matrisi kısmı incelendiği zaman birinci sınıfa ait 499 verinin ve dördüncü sınıfa ait 353 verinin tamamının yanlış sınıflara atanması bu sınıflara ait TP oranı, Kesinlik, F-ölçütü ve FP oranı değerlerinin 0 olmasına neden olmuştur. Ayrıca üçüncü sınıfa ait 1498 verinin de 21'nin doğru sınıfa atanması ile birlikte ağırlıklı ortalama TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. ikinci sınıfa ait 5562 verinin 5542'si yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 20'si üçüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

Tablo 3.6 B39-Umut düzeyi veri kümesinde ele alınan 5 sınıflandırma algoritması ile elde edilen etkinlik ölçüt değerlerini toplu olarak göstermektedir.

Tablo 3.6. B39 Umut düzeyi hedef değişkeni için elde edilen toplu sonuçlar

Algoritma	TP oranı	FP oranı	Kesinlik	F-ölçütü	ROC alanı	Doğruluk	Kural sayısı
J48	0,686	0,58	0,607	0,623	0,607	% 68,56	778
JRip	0,698	0,646	0,616	0,61	0,531	% 69,81	18
PART	0,641	0,481	0,61	0,623	0,617	% 64,10	765
REPTree	0,692	0,626	0,603	0,614	0,672	% 69,19	417
SimpleCart	0,703	0,695	0,578	0,585	0,525	% 70,31	1

Tablo 3.6'dan da görüldüğü gibi ele alınan veri kümesinde en yüksek % 70 doğruluk değeri elde edilmiştir. Bu durumun en büyük sebebi, en çok kullanılan sınıf değeri dışındaki tüm sınıf değerlerinin çok düşük doğruluk oranlarında sınıflandırılması hatta bazılarının bazı algoritmalarda tamamen yanlış sınıflandırılmasından kaynaklanmaktadır. Algoritmalar sadece en çok veriye sahip sınıf değeri için güzel

sonuçlar vermiş ancak bu sonuçlar ulaşılmak istenen TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerine ulaşmayı sağlayamamıştır. Kendi aralarında kıyaslayacak olursak doğruluk bakımından en iyi sonucu % 70,31 ile SimpleCart algoritması vermiştir. Kural sayıları kıyaslanacak olursa en az kural en iyi sonuç veren SimpleCart algoritmasında elde edilmiş ancak bu algoritma kural oluşturmaya gerek duymayarak tüm sınıf değerlerini verilen cevapların % 70'ini yani büyük çoğunluğunu oluşturan ikinci sınıfa atamıştır. Tablo incelendiğinde genel olarak kural sayılarının azalmasının doğruluk oranlarını arttırdığı gözlenmiştir.

3.4.3.2 B13_1 Sağlık Hizmetlerinden Memnuniyet Hedef Değişkeni için Sınıflandırma

B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkenine ait veri ön işleme sonucu elde edilen veri kümesi 5 sınıflı bir problemdir. 7869 örnek ve 27 adet girdi değişkeni içermektedir. Aşağıda, ele alınan sınıflandırma algoritmaları ile tüm veri kümesi üzerinde 10-katlı çapraz doğrulama ile elde edilen sonuçlar verilmektedir.

J48 algoritması ile elde edilen sonuçlar

Şekil 3.20 B13_1 Sağlık hizmetlerinden memnuniyet veri kümesinde J48 algoritması ile elde edilen sonuçları göstermektedir.

İlgili veri kümesinde J48 algoritması ile % 68,6 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 346 kural oluşmuştur.

Kappa istatistik değeri sınıf değerlerinin uyuşmadığını göstermektedir. Karışıklık matrisi incelendiğinde üçüncü sınıfa ait 982 verinin sadece 56'sının, birinci sınıfa ait 561 verinin sadece 1'inin, beşinci sınıfa ait 130 verinin sadece 1'inin ve dördüncü sınıfa ait 779 verinin sadece 55'inin doğru sınıflandırıldığı görülmektedir. Bu ise, TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5417 verinin 5285'i yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 70'i üçüncü sınıfa, 11'i beşinci sınıfa ve 51'i dördüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

```
=== Summary ===

Correctly Classified Instances        5398               68.5983 %
Incorrectly Classified Instances      2471               31.4017 %
Kappa statistic                          0.0797
Mean absolute error                      0.1898
Root mean squared error                  0.3162
Relative absolute error                 95.7581 %
Root relative squared error            100.4526 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.057     0.022      0.271     0.057     0.094      0.551    3.0
                 0.002     0          0.333     0.002     0.004      0.537    1.0
                 0.976     0.891      0.708     0.976     0.82       0.593    2.0
                 0.008     0.004      0.034     0.008     0.013      0.493    5.0
                 0.071     0.015      0.342     0.071     0.117      0.6      4.0
Weighted Avg.    0.686     0.617      0.579     0.686     0.588      0.583

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   56    2  881    3   40 |    a = 3.0
    3    1  551    2    4 |    b = 1.0
   70    0 5285   11   51 |    c = 2.0
   16    0  102    1   11 |    d = 5.0
   62    0  650   12   55 |    e = 4.0
```

Şekil 3.20. J48 algoritması ile B13_1-Sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

JRip algoritması ile elde edilen sonuçlar

Şekil 3.21 B13_1 Sağlık hizmetlerinden memnuniyet veri kümesinde JRip algoritması ile elde edilen sonuçları göstermektedir.

İlgili veri kümesinde JRip algoritması ile % 68,7 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 8 kural oluşmuştur. Yine kappa istatistik değeri, sınıfların uyuşmadığını göstermektedir. Karışıklık matrisi kısmı incelendiği zaman birinci sınıfa ait 561 verinin tamamının yanlış sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve FP oranı değerinin 0 olmasına neden olmuştur. Ayrıca üçüncü sınıfa ait 982 verinin sadece 8'inin, beşinci sınıfa ait 130 verinin sadece 1'inin, dördüncü sınıfa ait 779 verinin sadece 13'ünün doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5417 verinin 5384'ü yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 17'si

üçüncü sınıfa, 1'i beşinci sınıfa ve 15'i dördüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

```
=== Summary ===

Correctly Classified Instances        5406               68.7    %
Incorrectly Classified Instances      2463               31.3    %
Kappa statistic                          0.0182
Mean absolute error                      0.1973
Root mean squared error                  0.3148
Relative absolute error                 99.5058 %
Root relative squared error            100.0168 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.008     0.005       0.195     0.008     0.016      0.508    3.0
                 0         0           0         0         0          0.505    1.0
                 0.994     0.974       0.693     0.994     0.816      0.513    2.0
                 0.008     0.001       0.167     0.008     0.015      0.5      5.0
                 0.017     0.005       0.26      0.017     0.031      0.509    4.0
Weighted Avg.    0.687     0.672       0.53      0.687     0.567      0.511

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
    8    0  955    1   18 |    a = 3.0
    2    0  557    0    2 |    b = 1.0
   17    0 5384    1   15 |    c = 2.0
    3    0  124    1    2 |    d = 5.0
   11    0  752    3   13 |    e = 4.0
```

Şekil 3.21. JRip algoritması ile B13_1-Sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

PART algoritması ile elde edilen sonuçlar

Şekil 3.22 B13_1 Sağlık hizmetlerinden memnuniyet veri kümesinde PART algoritması ile elde edilen sonuçları göstermektedir.

İlgili veri kümesinde PART algoritması ile % 61,67 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 934 kural oluşmuştur. Karışıklık matrisinde en iyi sınıflandırılan sınıfın en çok örneğe sahip ikinci sınıf olduğu görülmektedir.

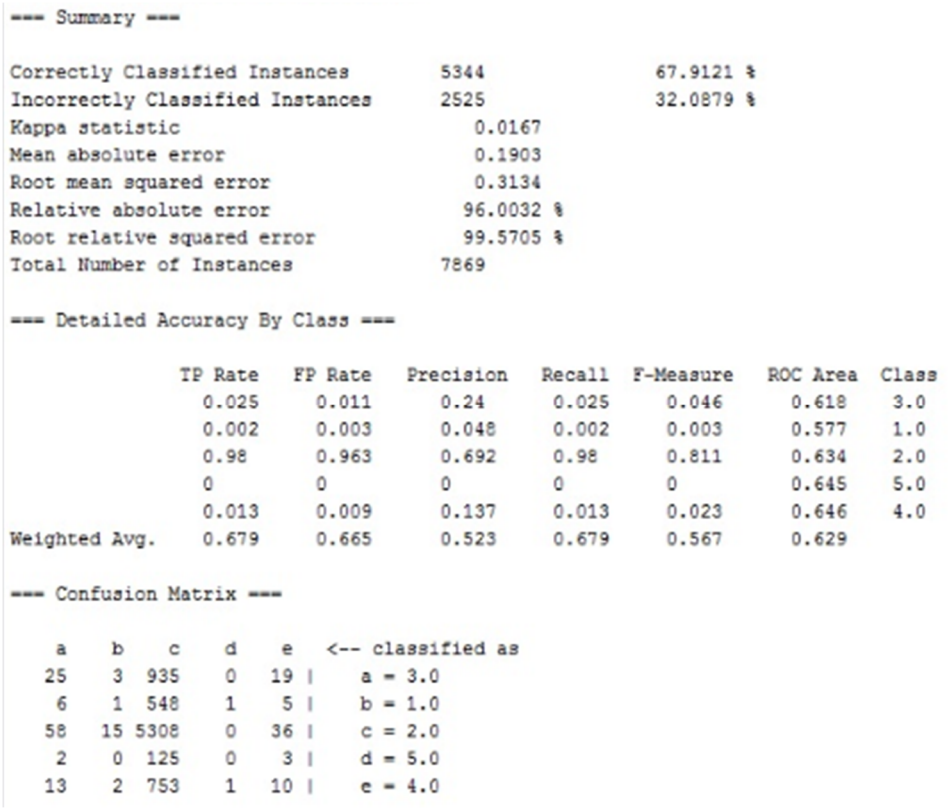
```
=== Summary ===

Correctly Classified Instances        4853               61.6724 %
Incorrectly Classified Instances      3016               38.3276 %
Kappa statistic                          0.1209
Mean absolute error                      0.1812
Root mean squared error                  0.3543
Relative absolute error                 91.3872 %
Root relative squared error            112.5465 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.166     0.085      0.218     0.166     0.188      0.558    3.0
                 0.052     0.034      0.105     0.052     0.069      0.543    1.0
                 0.835     0.682      0.73      0.835     0.779      0.608    2.0
                 0.038     0.007      0.081     0.038     0.052      0.534    5.0
                 0.169     0.064      0.226     0.169     0.194      0.592    4.0
Weighted Avg.    0.617     0.489      0.561     0.617     0.585      0.594

=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
 163   36  656    8  119 |    a = 3.0
  39   29  468    3   22 |    b = 1.0
 403  177 4524   26  287 |    c = 2.0
  28    4   69    5   24 |    d = 5.0
 116   31  480   20  132 |    e = 4.0
```

Şekil 3.22. PART algoritması ile B13_1-Sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

REPTree algoritması ile elde edilen sonuçlar

Şekil 3.23 B13_1 Sağlık hizmetlerinden memnuniyet veri kümesinde REPTree algoritması ile elde edilen sonuçları göstermektedir.

İlgili veri kümesinde REPTree algoritması ile % 67,91 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 405 kural oluşmuştur. Kappa istatistik değeri 0'a yakın bir değer olduğundan sınıf değerleri uyuşmamaktadır. Karışıklık matrisinde üçüncü sınıfa ait 982 verinin sadece 25'inin, birinci sınıfa ait 561 verinin sadece 1'inin ve dördüncü sınıfa ait 779 verinin sadece 10'unun doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. Ayrıca beşinci sınıfa ait 130 verinin tamamının yanlış sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve FP oranı değerinin 0 olmasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5417 verinin 5308'i yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 58'i üçüncü sınıfa, 15'i birinci sınıfa ve 36'sı dördüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.



```
=== Summary ===

Correctly Classified Instances        5344               67.9121 %
Incorrectly Classified Instances      2525               32.0879 %
Kappa statistic                          0.0167
Mean absolute error                      0.1903
Root mean squared error                  0.3134
Relative absolute error                 96.0032 %
Root relative squared error             99.5705 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.025     0.011      0.24      0.025     0.046      0.618    3.0
                 0.002     0.003      0.048     0.002     0.003      0.577    1.0
                 0.98      0.963      0.692     0.98      0.811      0.634    2.0
                 0         0          0         0         0          0.645    5.0
                 0.013     0.009      0.137     0.013     0.023      0.646    4.0
Weighted Avg.    0.679     0.665      0.523     0.679     0.567      0.629

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   25    3  935    0   19 |    a = 3.0
    6    1  548    1    5 |    b = 1.0
   58   15 5308    0   36 |    c = 2.0
    2    0  125    0    3 |    d = 5.0
   13    2  753    1   10 |    e = 4.0
```

Şekil 3.23. REPTree algoritması ile B13_1-Sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

SimpleCart algoritması ile elde edilen sonuçlar

Şekil 3.24 B13_1 Sağlık hizmetlerinden memnuniyet veri kümesinde SimpleCart algoritması ile elde edilen sonuçları göstermektedir.

İlgili veri kümesinde SimpleCart algoritması ile % 68,66 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 8 kural oluşmuştur. Karışıklık matrisine göre en iyi sınıflandırılan sınıf, ikinci sınıftır. Diğer sınıflarda yanlış sınıflandırılan örnek sayısı oldukça fazladır.

Tablo 3.7 B13_1-Sağlık hizmetlerinden memnuniyet veri kümesinde ele alınan 5 sınıflandırma algoritması ile elde edilen etkinlik ölçüt değerlerini toplu olarak göstermektedir.

```
=== Summary ===

Correctly Classified Instances        5403               68.6618 %
Incorrectly Classified Instances      2466               31.3382 %
Kappa statistic                          0.0294
Mean absolute error                      0.1952
Root mean squared error                  0.3128
Relative absolute error                 98.4845 %
Root relative squared error             99.3753 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.022    0.009      0.265     0.022     0.041      0.56     3.0
                 0        0          0         0         0          0.523    1.0
                 0.99     0.959      0.695     0.99      0.817      0.564    2.0
                 0        0          0         0         0          0.592    5.0
                 0.023    0.007      0.254     0.023     0.042      0.583    4.0
Weighted Avg.    0.687    0.662      0.537     0.687     0.572      0.563

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   22    0  943    0   17 |    a = 3.0
    1    0  558    0    2 |    b = 1.0
   27    0 5363    0   27 |    c = 2.0
   10    0  113    0    7 |    d = 5.0
   23    0  738    0   18 |    e = 4.0
```

Şekil 3.24. SimpleCart algoritması ile B13_1-Sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

Tablo 3.7. B13_1 Sağlık hizmetlerinden memnuniyet veri kümesinde elde edilen toplu sonuçlar

Algoritma	TP oranı	FP oranı	Kesinlik	F-ölçütü	ROC alanı	Doğruluk	Kural sayısı
J48	0,686	0,617	0,579	0,588	0,583	% 68,59	346
JRip	0,687	0,672	0,53	0,567	0,511	% 68,7	8
PART	0,617	0,489	0,561	0,585	0,594	% 61,67	934
REPTree	0,679	0,665	0,523	0,567	0,629	% 67,91	405
SimpleCart	0,687	0,662	0,537	0,572	0,563	% 68,66	8

Ele alınan veri kümesinde en iyi doğruluk değerini JRip algoritması 8 kural ile vermiştir. Kural sayıları kıyaslanacak olursa en az kural en iyi sonuç veren JRip

algoritmasında ve ona çok yakın bir sonuç elde eden SimpleCart algoritmasında elde edilmiştir. Tablo incelendiğinde genel olarak kural sayılarının azalmasının doğruluk oranlarını arttırdığı gözlenmiştir.

3.4.3.3 B13_2 Asayiş Hizmetlerinden Memnuniyet Değişkeni için Sınıflandırma

B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenine ait veri ön işleme sonucu elde edilen veri kümesi 5 sınıflı bir problemdir. 7912 örnek ve 22 adet girdi değişkeni içermektedir. Aşağıda, ele alınan sınıflandırma algoritmaları ile tüm veri kümesi üzerinde 10-katlı çapraz doğrulama ile elde edilen sonuçlar verilmektedir.

J48 algoritması ile elde edilen sonuçlar

B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde J48 algoritması ile elde edilen sonuçlar Şekil 3.25'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5922               74.8483 %
Incorrectly Classified Instances      1990               25.1517 %
Kappa statistic                          0.1025
Mean absolute error                      0.1527
Root mean squared error                  0.2835
Relative absolute error                 92.0358 %
Root relative squared error             98.4619 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.059     0.019      0.292     0.059      0.098      0.576    3.0
                 0         0.001      0         0          0          0.552    1.0
                 0.977     0.878      0.771     0.977      0.862      0.642    2.0
                 0.11      0.016      0.345     0.11       0.167      0.698    4.0
                 0         0.001      0         0          0          0.521    5.0
Weighted Avg.    0.748     0.663      0.637     0.748      0.67       0.632

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   54    2  813   43    1 |    a = 3.0
    4    0  417    2    0 |    b = 1.0
   68    2 5807   62    4 |    c = 2.0
   53    1  437   61    3 |    d = 4.0
    6    1   62    9    0 |    e = 5.0
```

Şekil 3.25. J48 algoritması ile B13_2-Asayiş hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde J48 algoritması ile % 74,85 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 244 kural oluşmuştur. Karışıklık matrisinde beşinci sınıfa ait 78 verinin ve birinci sınıfa ait 423 verinin tamamının yanlış sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve FP oranı değerinin 0 olmasına neden olmuştur. Ayrıca üçüncü sınıfa ait 913 verinin sadece 54'ünün, dördüncü sınıfa ait 555 verinin sadece 61'inin doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5943 verinin 5807'si yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 68'i üçüncü sınıfa, 2'si birinci sınıfa, 4'ü beşinci sınıfa ve 62'si dördüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

JRip algoritması ile elde edilen sonuçlar

B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde JRip algoritması ile elde edilen sonuçlar Şekil 3.26'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5932               74.9747 %
Incorrectly Classified Instances      1980               25.0253 %
Kappa statistic                          0.0122
Mean absolute error                      0.1651
Root mean squared error                  0.2883
Relative absolute error                 99.477  %
Root relative squared error            100.1231 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.002     0.002      0.118     0.002     0.004      0.5      3.0
                 0         0          0         0         0          0.499    1.0
                 0.996     0.986      0.753     0.996     0.858      0.506    2.0
                 0.022     0.003      0.353     0.022     0.041      0.511    4.0
                 0         0          0         0         0          0.511    5.0
Weighted Avg.    0.75      0.741      0.604     0.75      0.647      0.505

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
    2    0  905    6    0 |    a = 3.0
    1    0  420    0    2 |    b = 1.0
   11    0 5918   14    0 |    c = 2.0
    2    0  541   12    0 |    d = 4.0
    1    0   75    2    0 |    e = 5.0
```

Şekil 3.26. JRip algoritması ile B13_2-Asayiş hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde JRip algoritması ile % 74,97 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 3 kural oluşmuştur. Kappa istatistik değeri 0'a yakın bir değer olduğundan sınıf değerleri uyuşmamaktadır. Karışıklık matrisinde beşinci sınıfa ait 78 verinin ve birinci sınıfa ait 423 verinin tamamının yanlış sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve FP oranı değerinin 0 olmasına neden olmuştur. Ayrıca üçüncü sınıfa ait 913 verinin sadece 2'sinin, dördüncü sınıfa ait 555 verinin sadece 12'sinin doğru sınıflandırılması TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerinin istenen seviyede olmamasına neden olmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5943 verinin 5918'i yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 11'i üçüncü sınıfa ve 14'ü dördüncü sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

PART algoritması ile elde edilen sonuçlar

B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde PART algoritması ile elde edilen sonuçlar Şekil 3.27'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5612               70.9302 %
Incorrectly Classified Instances      2300               29.0698 %
Kappa statistic                          0.1297
Mean absolute error                      0.1502
Root mean squared error                  0.305
Relative absolute error                 90.5039 %
Root relative squared error            105.9285 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.105     0.06       0.187     0.105     0.135      0.549    3.0
                 0.028     0.011      0.125     0.028     0.046      0.573    1.0
                 0.909     0.756      0.784     0.909     0.842      0.654    2.0
                 0.178     0.039      0.255     0.178     0.21       0.676    4.0
                 0.013     0.003      0.045     0.013     0.02       0.528    5.0
Weighted Avg.    0.709     0.578      0.636     0.709     0.665      0.637

=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
  96   13  701   98    5 |    a = 3.0
  11   12  394    5    1 |    b = 1.0
 290   66 5404  174    9 |    c = 2.0
  95    4  351   99    6 |    d = 4.0
  21    1   43   12    1 |    e = 5.0
```

Şekil 3.27. PART algoritması ile B13_2-Asayiş hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde PART algoritması ile % 70,93 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 574 kural oluşmuştur. En iyi değerlere veri kümesinde en çok örneğe sahip ikinci sınıf değerinde yakalanmıştır. İkinci sınıfa ait 5943 verinin 5404'ü yine ikinci sınıfa atanarak olarak doğru sınıflandırılmış, 290'ı üçüncü sınıfa, 66'sı birinci sınıfa, 174'ü dördüncü sınıfa ve 9'u beşinci sınıfa atanarak yanlış sınıflandırma gerçekleştirilmiştir.

REPTree algoritması ile elde edilen sonuçlar

B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde REPTree algoritması ile elde edilen sonuçlar Şekil 3.28'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5875               74.2543 %
Incorrectly Classified Instances      2037               25.7457 %
Kappa statistic                          0.0506
Mean absolute error                      0.1533
Root mean squared error                  0.2825
Relative absolute error                 92.3818 %
Root relative squared error             98.0949 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.03      0.018      0.174     0.03      0.051      0.619    3.0
                 0.002     0.002      0.067     0.002     0.005      0.564    1.0
                 0.979     0.926      0.761     0.979     0.857      0.671    2.0
                 0.05      0.01       0.286     0.05      0.086      0.768    4.0
                 0         0          0         0         0          0.645    5.0
Weighted Avg.    0.743     0.699      0.616     0.743     0.656      0.666

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   27    2  853   30    1 |    a = 3.0
    3    1  418    1    0 |    b = 1.0
   81   11 5819   32    0 |    c = 2.0
   34    1  492   28    0 |    d = 4.0
   10    0   61    7    0 |    e = 5.0
```

Şekil 3.28. REPTree algoritması ile B13_2-Asayiş hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde REPTree algoritması ile % 74,25 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 324 kural oluşmuştur.

SimpleCart algoritması ile elde edilen sonuçlar

B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde SimpleCart algoritması ile elde edilen sonuçlar Şekil 3.29'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5924               74.8736 %
Incorrectly Classified Instances      1988               25.1264 %
Kappa statistic                          0.0365
Mean absolute error                      0.1607
Root mean squared error                  0.2841
Relative absolute error                 96.8467 %
Root relative squared error             98.6641 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.014     0.006      0.245     0.014     0.027      0.583    3.0
                 0         0          0         0         0          0.537    1.0
                 0.99      0.958      0.757     0.99      0.858      0.594    2.0
                 0.049     0.008      0.303     0.049     0.084      0.642    4.0
                 0         0          0         0         0          0.632    5.0
Weighted Avg.    0.749     0.721      0.618     0.749     0.654      0.593

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   13    0  880   20    0 |    a = 3.0
    1    0  421    1    0 |    b = 1.0
   24    0 5884   35    0 |    c = 2.0
   14    0  514   27    0 |    d = 4.0
    1    0   71    6    0 |    e = 5.0
```

Şekil 3.29. SimpleCart algoritması ile B13_2-Asayiş hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde SimpleCart algoritması ile % 74,87 doğruluk ile sınıflandırılma gerçekleştirilmiş ve algoritma tek kural oluşturarak tüm örnekleri ikinci sınıf değerine atamıştır.

Tablo 3.8 B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde ele alınan beş algoritma ile elde edilen sonuçları toplu olarak göstermektedir.

Tablodan da görüldüğü gibi PART algoritması hariç diğer algoritmalar birbirlerine yakın sonuçlar vermiştir. Ele alınan veri kümesi, % 75 doğruluk ile sınıflandırılmıştır.

Tablo 3.8. B13_2 Asayiş hizmetlerinden memnuniyet veri kümesinde elde edilen toplu sonuçlar

Algoritma	TP oranı	FP oranı	Kesinlik	F-ölçütü	ROC alanı	Doğruluk	Kural sayısı
J48	0,748	0,663	0,637	0,67	0,632	% 74,84	244
JRip	0,75	0,741	0,604	0,647	0,505	% 74,97	3
PART	0,709	0,578	0,636	0,665	0,637	% 70,93	574
REPTree	0,743	0,699	0,616	0,656	0,666	% 74,25	324
SimpleCart	0,749	0,721	0,618	0,654	0,593	% 74,87	1

Kural sayıları kıyaslanacak olursa en az kural SimpleCart algoritmasında çıkmıştır ancak bu algoritma kural oluşturmaya gerek duymayarak tüm sınıf değerlerini verilen cevapların % 75'ini yani büyük çoğunluğunu oluşturan ikinci sınıfa atamıştır. En iyi sonuç veren JRip algoritmasında ise 3 kural elde edilmiştir. Tablo incelendiğinde genel olarak kural sayılarının azalmasının doğruluk oranlarını arttırdığı gözlenmiştir.

3.4.4 Nitelik Seçimi ile Sınıflandırma

Eğer hedef değişken için belirlenmiş olan girdi değişkenleri hedef değişkenini tam olarak etkilemiyorsa, bu sınıflandırma işlemi sonucunun daha düşük doğruluk oranı alınmasına neden olabilir. Bu nedenle de girdi değişkenlerinin etki gücünden tam emin olmak için nitelik seçimi kullanılması gerekir.

Tez çalışması kapsamında, uzman görüşü alınarak belirlenen ve veri ön işlemeden geçen tüm girdi değişkenleriyle sınıflandırma neticesinde elde edilen sonuçları daha da iyileştirmek adına nitelik seçimi gerçeklenmiştir. Nitelik seçimi, girdi değişkenleri ile hedef değişken arasındaki ilişkiyi değerlendirerek girdi değişkeninin hedef değişkeni etkileme gücünü ölçer ve tüm değişkenler arasından hedef değişkenini etkileme gücü yüksek olan değişkenleri ortaya koyar.

Nitelik seçimi, filtreleme (filtering) ve sarmalama (wrapping) olmak üzere 2 farklı şekilde yapılabilir. Filtreleme yönteminde tüm girdi değişkenleri hedef değişkenini etkileme oranına göre en etkili olanından en etkisiz olanına kadar sıralanarak

değişkenleri indirgeme yoluna başvurur. Sarmalama yönteminde ise hedef değişkenini etkileme gücüne bakılarak en etkin değişkenlerden alt küme oluşturularak seçim yapılır.

Tez kapsamında, nitelik seçiminde iyi merit değeri verdiği için sarmalama yöntemlerinden ConsistencySubsetEval yöntemi kullanılmıştır. Merit, girdi değişkenlerinin hedef değişkeni ifade etmedeki önemini belirlemeye yarayan soyut bir ölçüdür. ConsistencySubsetEval yönteminde, öğrenme algoritması olarak çeşitli denemeler sonucunda en yüksek sonuçları veren J48 algoritması tercih edilmiştir.

Nitelik indirgeme (sarmalama) yönteminde kullanılan ConsistencySubsetEval yöntemi, Hi Lu ve arkadaşları [149] tarafından sunulan tutarlılık ölçütünü Eşitlik (3.8)'de gösterildiği gibi kullanır.

$$Tutarlılık_s = 1 - \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N} \tag{3.8}$$

s, bir değişken altkümesidir. J, s için değişken değerlerinin farklı kombinasyonlarının sayısıdır. $|D_i|$, değişken değeri kombinasyonun i. oluşum sayısıdır. $|M_i|$, i. değişken değeri kombinasyonu için çoğunluk sınıfının eleman sayısıdır. N, veri kümesindeki örneklerin toplam sayısıdır [146]. ConsistencySubsetEvaluater'ı kullanmak için veri kümesi, U. M. Fayyad ve arkadaşlarının yöntemi gibi, uygun bir yöntem kullanılarak sayısal değişkenlerin kesiklendirilmesine ihtiyaç duyabilir [147]. Bu araştırma metodu, değişkenlerin listesini üreten bir ileri seçim araştırması olarak kullanılır [148]. Değişkenler, daha sonra değişken kümesinin tutarlılığında genel katkılarına göre sıralanır.

3.4.4.1 B39 Umut düzeyi veri kümesi nitelik seçimi ve sınıflandırma sonuçları

Nitelik seçimi sırasında, ConsistencySubsetEval yöntemi toplamda 626 değişken alt kümesi oluşturulup değerlendirilmiş ve seçilen değişkenler ışığında merit (yararlılık) değeri olarak 1 tam değerini bulmuştur. Tablo 3.9 nitelik seçimi sonrası B39 Umut düzeyi veri kümesinde elde edilen değişkenleri göstermektedir.

Tablo 3.9. B39 umut düzeyi değişkeni için seçilen değişkenler

Seçilen değişken kodu	Değişken tanımı	Değer aralığı veya (kategorik değer sayısı)	Değişken tipi
H17	Hanenin aylık toplam net geliri	[1,2,3,4,5,6]	Nominal
H18	Bu gelirle hanenin temel ihtiyaçlarını karşılama düzeyi	[1,2,3,4,5]	Nominal
Yerleşim yeri	Yerleşim yeri	[1,2]	Nominal
HHB	Hane halkında yaşayan kişi sayısı	[1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16]	Nominal
B02	Eğitim durumu	[1,2,3,4,5,6,7,8]	Nominal
B03	Son bir hafta içinde ücretli ya da ücretsiz olarak bir işte çalışma durumu	[1,2,3,4,5,6,7, 8,9,10,11,12]	Nominal
B06	Çalışılan işteki durum	[1,2,3,4,5]	Nominal
B09	Hayatta en çok kimin mutlu ettiği	[1,2,3,4,5,6,7,8,98]	Nominal
B10	Hayatta en çok neyin mutlu ettiği	[1,2,3,4,5,98]	Nominal
B11_1	Sağlıktan memnuniyet	[1,2,3,4,5]	Nominal
B11_2	Evlilikten memnuniyet	[1,2,3,4,5,6]	Nominal
B11_3	Şimdiye kadar alınan eğitimden memnuniyet	[1,2,3,4,5,6]	Nominal
B11_8	Aylık hane halkı gelirinden memnuniyet	[1,2,3,4,5]	Nominal
B12_3	Komşularla ilişkilerden memnuniyet	[1,2,3,4,5]	Nominal
B13_3	Adli hizmetlerden memnuniyet	[1,2,3,4,5,6]	Nominal
B13_6	Ulaştırma hizmetlerinden memnuniyet	[1,2,3,4,5,6]	Nominal
B19	Hangi sosyal güvenlik kurumundan yararlanıyorsunuz?	[11,12,13,14,2]	Nominal
B38	Yaşadığınız çevrede gece yalnız yürürken kendinizi ne kadar güvende hissediyorsunuz?	[1,2,3,4,5]	Nominal

J48 Algoritması ile nitelik seçimi sonrası elde edilen sınıflandırma sonuçları

J48 algoritması ile B39 umut düzeyi veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.30'da gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5505               69.5779 %
Incorrectly Classified Instances      2407               30.4221 %
Kappa statistic                          0.071
Mean absolute error                      0.2159
Root mean squared error                  0.3395
Relative absolute error                 93.0263 %
Root relative squared error             99.67   %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.052     0.005      0.406     0.052     0.092      0.62     1.0
                 0.965     0.902      0.717     0.965     0.823      0.633    2.0
                 0.068     0.009      0.255     0.068     0.107      0.633    4.0
                 0.059     0.028      0.331     0.059     0.101      0.65     3.0
Weighted Avg.    0.696     0.64       0.604     0.696     0.608      0.635

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   26  460    5    8 |    a = 1.0
   28 5366   34  134 |    b = 2.0
    3  288   24   38 |    c = 4.0
    7 1371   31   89 |    d = 3.0
```

Şekil 3.30. J48 algoritması ile B39 umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde J48 algoritması ile % 69,58 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 228 kural oluşmuştur.

Değişken seçimi öncesine göre birinci, üçüncü ve dördüncü sınıf değerlerinin doğruluk oranı azalmış ancak ikinci sınıf değerinden elde edilen doğruluk oranı artmıştır. Veri kümesini en büyük kısmını da ikinci sınıf değerine ait örnekler teşkil ettiği için, nitelik seçimi olmadan elde edilen % 68,57 doğruluk değeri % 69,58'e yükselmiştir. Kural sayıları açısından ise nitelik seçimi ile 778 kural 228 kurala indirgenmiştir.

JRip Algoritması ile nitelik seçimi sonrası elde edilen sınıflandırma sonuçları

JRip algoritması ile B39 umut düzeyi veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.31'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5540               70.0202 %
Incorrectly Classified Instances      2372               29.9798 %
Kappa statistic                          0.0582
Mean absolute error                      0.2271
Root mean squared error                  0.3403
Relative absolute error                 97.8582 %
Root relative squared error             99.9224 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.1       0.007      0.476     0.1       0.166      0.554    1.0
                 0.975     0.934      0.712     0.975     0.823      0.52     2.0
                 0.028     0.004      0.27      0.028     0.051      0.519    4.0
                 0.037     0.015      0.373     0.037     0.068      0.52     3.0
Weighted Avg.    0.7       0.66       0.613     0.7       0.604      0.522

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   50  446    2    1 |    a = 1.0
   49 5424   13   76 |    b = 2.0
    2  324   10   17 |    c = 4.0
    4 1426   12   56 |    d = 3.0
```

Şekil 3.31. JRip algoritması ile B39 umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde nitelik seçimi sonrası JRip algoritması ile doğruluk oranı % 69,82'den % 70,02'ye çıkmıştır. Nitelik seçimi olmadan JRip algoritması ile 18 kural elde edilirken nitelik seçimi ile bu sayı yarıya indirgenmiştir.

PART algoritması ile elde edilen sonuçlar

PART algoritması ile B39 umut düzeyi veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.32'de gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5195               65.6598 %
Incorrectly Classified Instances      2717               34.3402 %
Kappa statistic                          0.1492
Mean absolute error                      0.207
Root mean squared error                  0.3646
Relative absolute error                 89.1823 %
Root relative squared error            107.0435 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.146     0.028      0.262     0.146     0.188      0.641    1.0
                 0.856     0.695      0.745     0.856     0.797      0.634    2.0
                 0.088     0.023      0.151     0.088     0.111      0.579    4.0
                 0.219     0.11       0.318     0.219     0.259      0.618    3.0
Weighted Avg.    0.657     0.512      0.607     0.657     0.626      0.629

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   73  358   14   54 |    a = 1.0
  161 4763   93  545 |    b = 2.0
   11  206   31  105 |    c = 4.0
   34 1069   67  328 |    d = 3.0
```

Şekil 3.32. PART algoritması ile B39 umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde nitelik seçimi sonrası PART algoritması ile doğruluk oranı % 64,10'dan % 65,66'ya çıkmıştır. Nitelik seçimi olmadan PART algoritması ile 765 kural elde edilirken nitelik seçimi ile bu sayı 626'ya indirgenmiştir.

REPTree algoritması ile elde edilen sonuçlar

REPTree algoritması ile B39 umut düzeyi veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.33'de gösterilmektedir.

İlgili veri kümesinde nitelik seçimi sonrası REPTree algoritması ile doğruluk oranı % 69,18'den % 68,67'ye azalmıştır. Nitelik seçimi olmadan REPTree algoritması ile 417 kural elde edilirken nitelik seçimi ile bu sayı 637'ye yükselmiştir.

```
=== Summary ===

Correctly Classified Instances        5433               68.6678 %
Incorrectly Classified Instances      2479               31.3322 %
Kappa statistic                          0.0516
Mean absolute error                      0.2127
Root mean squared error                  0.3385
Relative absolute error                 91.6597 %
Root relative squared error             99.3757 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.048     0.01       0.24      0.048     0.08       0.689    1.0
                 0.954     0.909      0.713     0.954     0.816      0.658    2.0
                 0.034     0.007      0.188     0.034     0.058      0.677    4.0
                 0.062     0.033      0.303     0.062     0.103      0.661    3.0
Weighted Avg.    0.687     0.647      0.582     0.687     0.601      0.661

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
   24  467    2    6 |    a = 1.0
   67 5304   27  164 |    b = 2.0
    2  295   12   44 |    c = 4.0
    7 1375   23   93 |    d = 3.0
```

Şekil 3.33. REPTree algoritması ile B39 umut düzeyi veri kümesinde sınıflandırma sonuçları

SimpleCart algoritması ile elde edilen sonuçlar

SimpleCart algoritması ile B39 umut düzeyi veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.34'te gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5567               70.3615 %
Incorrectly Classified Instances      2345               29.6385 %
Kappa statistic                          0.012
Mean absolute error                      0.2301
Root mean squared error                  0.339
Relative absolute error                 99.1291 %
Root relative squared error             99.5303 %
Total Number of Instances             7912

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.01      0          0.625     0.01      0.02       0.521    1.0
                 0.997     0.988      0.705     0.997     0.826      0.53     2.0
                 0.006     0          0.667     0.006     0.011      0.535    4.0
                 0.009     0.003      0.406     0.009     0.017      0.536    3.0
Weighted Avg.    0.704     0.695      0.642     0.704     0.586      0.531

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
    5  493    0    1 |    a = 1.0
    3 5547    1   11 |    b = 2.0
    0  344    2    7 |    c = 4.0
    0 1485    0   13 |    d = 3.0
```

Şekil 3.34. SimpleCart algoritması ile B39 umut düzeyi veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde nitelik seçimi sonrası SimpleCart algoritması ile doğruluk oranı % 70,31'den % 70,36'ya çıkmıştır. Nitelik seçimi olmadan SimpleCart algoritması ile kural oluşturulmayıp tüm sınıf değerleri verilen cevapların % 70'ini oluşturan ikinci sınıf değerine atanırken nitelik seçimi ile bu sayı 21'e yükselmiştir.

Tablo 3.10 B39 Umut düzeyi veri kümesinde ele alınan nitelik seçimi sonrası 5 sınıflandırma algoritması ile elde edilen etkinlik ölçüt değerlerini toplu olarak göstermektedir.

Tablo 3.10. B39 Umut düzeyi hedef değişkeni için elde edilen toplu sonuçlar

Algoritma	TP oranı	FP oranı	Kesinlik	F-ölçütü	ROC alanı	Doğruluk	Kural sayısı
J48	0,696	0,64	0,604	0,608	0,635	69,57	228
JRip	0,7	0,66	0,613	0,604	0,522	70,02	9
PART	0,657	0,512	0,607	0,626	0,629	65,66	626
REPTree	0,687	0,647	0,582	0,601	0,661	68,66	637
SimpleCart	0,704	0,695	0,642	0,586	0,531	70,36	21

Yukarıdaki tablodan da anlaşılacağı gibi uygulanan tüm algoritmalarda istenen sonuç tam olarak alınamamıştır. Bu durumun en büyük sebebi en çok kullanılan sınıf değeri dışındaki tüm sınıf değerlerinin çok düşük doğruluk oranlarında sınıflandırılması hatta bazılarının bazı algoritmalarda tamamen yanlış sınıflandırılmasından kaynaklanmaktadır. Algoritmalar sadece en çok veriye sahip sınıf değeri için güzel sonuçlar vermiş ancak bu sonuçlar ulaşılmak istenen TP oranı, Kesinlik, F-ölçütü ve ROC alanı değerlerine ulaşmayı sağlayamamıştır. Kendi aralarında kıyaslanacak olursa doğruluk bakımından en iyi sonucu SimpleCart algoritması vermiştir. Nitelik seçimi öncesine göre kıyaslanacak olursa REPTree algoritması dışında tüm algoritmalarda doğruluk oranlarında çok küçük de olsa bir artış gözlenmiştir. Genel olarak en iyi sonuç nitelik seçimi sonrası SimpleCart algoritmasında % 70,36 doğruluk ile alınmıştır. Kural sayıları kıyaslanacak olursa en az kural JRip algoritmasında çıkmış ancak en iyi sonuç

veren SimpleCart algoritmasında ise 21 kural elde edilmiştir. Tablo incelendiğinde kural sayılarındaki azalmanın doğruluk oranlarını arttırdığı gözlenmemiştir.

3.4.4.2 B13_1 Sağlık hizmetlerinden memnuniyet veri kümesi nitelik seçimi ve sınıflandırma sonuçları

Değişken seçimi işlemi sonrasında, ConsistencySubsetEval yöntemi değişken seçimi işlemi için toplamda 380 veri alt kümesi oluşturulup değerlendirilmiş ve seçilen değişkenler ışığında merit (yararlılık) değeri olarak 0,999 yani 1 tam değerini bulmuştur. Tablo 3.11 nitelik seçimi sonrası B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde elde edilen değişkenleri göstermektedir.

Tablo 3.11. B13_1 sağlık hizmetlerinden memnuniyet değişkeni için seçilen değişkenler

Seçilen değişken kodu	Değişken tanımı	Değer aralığı veya (kategorik değer sayısı)	Değişken tipi
Yerleşim yeri	Yerleşim yeri	[1,2]	Nominal
Cinsiyet	Cinsiyet	[1,2]	Nominal
YAS	18 ve yukarıdaki fert yaşı	18≤x≤93	Numerik
HHB	Hane halkında yaşayan kişi sayısı	[1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16]	Nominal
B01	Medeni durum	[1,2,3,4,99]	Nominal
B02	Eğitim durumu	[1,2,3,4,5,6,7,8]	Nominal
B03	Son bir hafta içinde ücretli ya da ücretsiz olarak bir işte çalışma durumu	[1,2,3,4,5,6,7, 8,9,10,11,12]	Nominal
B07	İşyerinin iktisadi faaliyet kodu	[1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16]	Nominal
B09	Hayatta en çok kimin mutlu ettiği	[1,2,3,4,5,6,7,8,98]	Nominal
B10	Hayatta en çok neyin mutlu ettiği	[1,2,3,4,5,98]	Nominal
B24	Hastalandığınızda masraflarınızı hangi kanalla karşılıyorsunuz?	[1,2,3,4,5,90,98]	Nominal

Tablo 3.11. B13_1 sağlık hizmetlerinden memnuniyet değişkeni için seçilen değişkenler (devamı)

B25	Hastalandığınızda ilk nereye gidersiniz?	[1,2,3,4,5,6,7,8]	Nominal
B26	Bu sağlık kuruluşunu neden seçiyorsunuz?	[1,2,3,4,5,6,7,98,99]	Nominal
B27_2	Temizlik/hijyen konusunda sorun var mı?	[1,2,3,99]	Nominal
B27_3	Yapılan muayeneden memnun musunuz?	[1,2,3,99]	Nominal
B27_5	Hemşire/Hasta bakıcıların hastalara davranışında sorun var mı?	[1,2,3,99]	Nominal
B27_6	Doktor / sağlık personeli sayısı yeterli mi?	[1,2,3,99]	Nominal
B27_7	Muayene ve tahlil ücretlerini yüksek buluyor musunuz?	[1,2,3,99]	Nominal
B27_8	İlaç fiyatlarında sorun görüyor musunuz?	[1,2,3,99]	Nominal
B27_9	Muayene ve tahlil için sıra beklemede sorun var mı?	[1,2,3,99]	Nominal
B27_10	Muayene için katkı payı ödemeyi sorun olarak görüyor musunuz?	[1,2,3,99]	Nominal
B29	En son sağlık hizmeti alımı sırasında herhangi bir sorun yaşadınız mı?	[1,2,99]	Nominal
B30	2012 yılında en son sağlık hizmeti aldığınız sağlık kuruluşu hangisidir?	[1,2,3,4,5,99]	Nominal

Aşağıda B13_1 sağlık hizmetlerinden memnuniyet için nitelik seçimi işlemi sonucunda elde edilen sınıflandırma sonuçları algoritma algoritma verilmiştir.

J48 algoritması ile elde edilen sonuçlar

J48 algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.35'te gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5386               68.4458 %
Incorrectly Classified Instances      2483               31.5542 %
Kappa statistic                          0.0759
Mean absolute error                      0.1902
Root mean squared error                  0.3161
Relative absolute error                 95.9284 %
Root relative squared error            100.4233 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.051     0.019      0.272     0.051     0.086      0.557    3.0
                 0         0.001      0         0         0          0.533    1.0
                 0.974     0.893      0.707     0.974     0.819      0.593    2.0
                 0.015     0.004      0.063     0.015     0.025      0.465    5.0
                 0.073     0.017      0.317     0.073     0.119      0.594    4.0
Weighted Avg.    0.684     0.619      0.553     0.684     0.587      0.582

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   50    2  880    6   44 |    a = 3.0
    2    0  551    0    8 |    b = 1.0
   67    3 5277   11   59 |    c = 2.0
   15    0  101    2   12 |    d = 5.0
   50    1  658   13   57 |    e = 4.0
```

Şekil 3.35. J48 algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde J48 algoritması ile % 68,45 doğruluk ile sınıflandırılma gerçekleştirilmiş ve 312 kural oluşmuştur.

Değişken seçimi öncesine göre beşinci ve dördüncü sınıf değerlerinin doğruluk oranı artmıştır. Ancak üçüncü, birinci ve ikinci sınıf değerinden elde edilen doğruluk oranı azaldığı ve veri kümesini en büyük kısmını da ikinci sınıf değerine ait örnekler teşkil ettiği için, nitelik seçimi olmadan elde edilen 68,57 doğruluk değeri % 68,45'e gerilemiştir. Kural sayıları açısından ise nitelik seçimi ile 346 kural 312 kurala indirgenmiştir.

JRip algoritması ile elde edilen sonuçlar

JRip algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.36'te gösterilmektedir.

```
=== Summary ===

Correctly Classified Instances        5414               68.8016 %
Incorrectly Classified Instances      2455               31.1984 %
Kappa statistic                          0.0223
Mean absolute error                      0.1966
Root mean squared error                  0.3145
Relative absolute error                 99.1856 %
Root relative squared error             99.9252 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.016     0.005       0.327     0.016     0.031      0.505    3.0
                 0         0           0         0         0          0.504    1.0
                 0.994     0.972       0.693     0.994     0.817      0.511    2.0
                 0.008     0.001       0.111     0.008     0.014      0.504    5.0
                 0.017     0.004       0.302     0.017     0.032      0.507    4.0
Weighted Avg.    0.688     0.67        0.55      0.688     0.569      0.509

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   16    0  949    1   16 |    a = 3.0
    1    0  559    0    1 |    b = 1.0
   20    0 5384    1   12 |    c = 2.0
    1    0  127    1    1 |    d = 5.0
   11    0  749    6   13 |    e = 4.0
```

Şekil 3.36. JRip algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

İlgili veri kümesinde nitelik seçimi sonrası JRip algoritması ile doğruluk oranı % 68,7'den % 68,8'e çıkmıştır. Nitelik seçimi olmadan JRip algoritması ile 8 kural elde edilirken nitelik seçimi ile bu sayı yarıya indirgenmiştir.

PART algoritması ile elde edilen sonuçlar

PART algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.37'te gösterilmektedir.

İlgili veri kümesinde nitelik seçimi sonrası PART algoritması ile doğruluk oranı % 61,67'den % 60,77'ye azalmıştır. Nitelik seçimi olmadan PART algoritması ile 934 kural elde edilirken nitelik seçimi ile bu sayı 892'ye indirgenmiştir.

```
=== Summary ===

Correctly Classified Instances        4782               60.7701 %
Incorrectly Classified Instances      3087               39.2299 %
Kappa statistic                          0.0999
Mean absolute error                      0.1828
Root mean squared error                  0.3552
Relative absolute error                 92.2141 %
Root relative squared error            112.8602 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.154     0.091      0.194     0.154     0.172      0.553    3.0
                 0.037     0.034      0.078     0.037     0.051      0.56     1.0
                 0.83      0.694      0.725     0.83      0.774      0.609    2.0
                 0.015     0.007      0.034     0.015     0.021      0.506    5.0
                 0.145     0.064      0.2       0.145     0.168      0.589    4.0
Weighted Avg.    0.608     0.498      0.549     0.608     0.575      0.595

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
  151   38  686    7  100 |    a = 3.0
   41   21  459    4   36 |    b = 1.0
  413  180 4495   31  298 |    c = 2.0
   35    5   69    2   19 |    d = 5.0
  138   26  488   14  113 |    e = 4.0
```

Şekil 3.37. PART algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

REPTree algoritması ile elde edilen sonuçlar

REPTree algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.38'te gösterilmektedir.

İlgili veri kümesinde nitelik seçimi sonrası REPTree algoritması ile doğruluk oranı % 67,91'dan % 67,96'ya çıkmıştır. Nitelik seçimi olmadan REPTree algoritması ile 405 kural elde edilirken nitelik seçimi ile bu sayı 399'ye indirgenmiştir.

```
=== Summary ===

Correctly Classified Instances        5348               67.9629 %
Incorrectly Classified Instances      2521               32.0371 %
Kappa statistic                          0.0162
Mean absolute error                      0.1903
Root mean squared error                  0.3132
Relative absolute error                 95.9974 %
Root relative squared error             99.5061 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.024     0.011      0.242     0.024     0.044      0.618    3.0
                 0         0.003      0         0         0          0.577    1.0
                 0.981     0.964      0.692     0.981     0.812      0.635    2.0
                 0         0          0         0         0          0.639    5.0
                 0.013     0.009      0.139     0.013     0.024      0.645    4.0
Weighted Avg.    0.68      0.666      0.52      0.68      0.567      0.63

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   24    3  936    0   19 |    a = 3.0
    6    0  549    1    5 |    b = 1.0
   54   14 5314    0   35 |    c = 2.0
    2    0  125    0    3 |    d = 5.0
   13    2  754    0   10 |    e = 4.0
```

Şekil 3.38. REPTree algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

SimpleCart algoritması ile elde edilen sonuçlar

SimpleCart algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde nitelik seçimi sonrası elde edilen sınıflandırma sonuçları Şekil 3.39'te gösterilmektedir.

İlgili veri kümesinde nitelik seçimi sonrası SimpleCart algoritması ile doğruluk oranı % 68,66'dan % 68,65'e azalmıştır. Nitelik seçimi olmadan SimpleCart algoritması ile 8 kural elde edilirken nitelik seçimi ile bu sayı 12'ye yükselmiştir.

```
=== Summary ===

Correctly Classified Instances        5402               68.6491 %
Incorrectly Classified Instances      2467               31.3509 %
Kappa statistic                          0.0352
Mean absolute error                      0.1942
Root mean squared error                  0.3121
Relative absolute error                 97.9662 %
Root relative squared error             99.1387 %
Total Number of Instances             7869

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.027     0.009      0.303     0.027     0.05       0.579    3.0
                 0         0          0         0         0          0.523    1.0
                 0.988     0.951      0.697     0.988     0.817      0.581    2.0
                 0         0          0         0         0          0.602    5.0
                 0.027     0.01       0.226     0.027     0.048      0.599    4.0
Weighted Avg.    0.686     0.657      0.54      0.686     0.574      0.579

=== Confusion Matrix ===

    a    b    c    d    e   <-- classified as
   27    0  932    0   23 |    a = 3.0
    2    0  556    0    3 |    b = 1.0
   25    0 5354    0   38 |    c = 2.0
    9    0  113    0    8 |    d = 5.0
   26    0  732    0   21 |    e = 4.0
```

Şekil 3.39. SimpleCart algoritması ile B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde sınıflandırma sonuçları

Tablo 3.12 B13_1 sağlık hizmetlerinden memnuniyet veri kümesinde ele alınan nitelik seçimi sonrası 5 sınıflandırma algoritması ile elde edilen etkinlik ölçüt değerlerini toplu olarak göstermektedir.

Tablo 3.12. B13_1 sağlık hizmetlerinden memnuniyet hedef değişkeni için elde edilen toplu sonuçlar

Algoritma	TP oranı	FP oranı	Kesinlik	F-ölçütü	ROC alanı	Doğruluk	Kural sayısı
J48	0,684	0,619	0,553	0,587	0,582	68,44	312
JRip	0,688	0,67	0,55	0,569	0,509	68,80	4
PART	0,608	0,498	0,549	0,575	0,595	60,77	892

Tablo 3.12. B13_1 sağlık hizmetlerinden memnuniyet hedef değişkeni için elde edilen toplu sonuçlar (devamı)

REPTree	0,68	0,666	0,52	0,567	0,63	67,96	399
SimpleCart	0,686	0,657	0,54	0,574	0,579	68,64	12

Yukarıdaki tablodan da anlaşılacağı gibi doğruluk bakımından en iyi sonucu JRip algoritması vermiştir. Nitelik seçimi öncesine göre kıyaslanacak olursa J48, PART ve SimpleCart algoritmalarının doğruluk oranlarında çok küçük bir azalma, JRip ve REPTree algoritmalarının doğruluk oranlarında çok küçük bir artış gözlenmiştir. Genel olarak en iyi sonuç nitelik seçimi sonrası JRip algoritmasında % 68,80 doğruluk ile alınmıştır. Kural sayıları kıyaslanacak olursa en az kural en iyi sonuç veren JRip algoritmasında elde edilmiştir. Tablo incelendiğinde genel olarak kural sayılarındaki azalmanın doğruluk oranlarını arttırdığı gözlemlenmiştir.

3.4.4.3 B13_2 Asayiş hizmetlerinden memnuniyet veri kümesi nitelik seçimi ve sınıflandırma sonuçları

Değişken seçimi işlemi sonrasında, ConsistencySubsetEval yöntemi değişken seçimi işlemi için toplamda 254 veri alt kümesi oluşturulup değerlendirilmiş ve seçilen değişkenler ışığında merit (yararlılık) değeri olarak 0,944 değerini bulmuştur. Tablo 3.13 nitelik seçimi sonrası B13_2 asayiş hizmetlerinden memnuniyet veri kümesinde elde edilen değişkenleri göstermektedir.

Tablo 3.13. B13_2 asayiş hizmetlerinden memnuniyet değişkeni için seçilen değişkenler

Seçilen değişken kodu	Değişken tanımı	Değer aralığı veya (kategorik değer sayısı)	Değişken tipi
Yerleşim yeri	Yerleşim yeri	[1,2]	Nominal
HHB	Hane halkı büyüklüğü	[1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16]	Nominal
CINSIYET	Cinsiyet	[1,2]	Nominal

Tablo 3.13. B13_2 asayiş hizmetlerinden memnuniyet değişkeni için seçilen değişkenler (devamı)

YAS	Yaş	18≤x≤93 arası	Numerik
B01	Medeni durum	[1,2,3,4,99]	Nominal
B02	Eğitim durumu	[1,2,3,4,5,6,7,8]	Nominal
B03	Çalışma durumu	[1,2,3,4,5,6,7,8,9,10,11,12]	Nominal
B06	Çalışılan işteki durum	[1,2,3,4,5]	Nominal
B07	İşyerinin faaliyet kodu	[1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16]	Nominal
B09	Hayatta en çok kimin mutlu ettiği	[1,2,3,4,5,6,7,8,98]	Nominal
B10	Hayatta en çok neyin mutlu ettiği	[1,2,3,4,5,98]	Nominal
B31_1	Polis veya jandarma olaylara zamanında müdahale ediyor mu?	[1,2]	Nominal
B31_2	Polis ve jandarmanın vatandaşa davranışından memnun musunuz?	[1,2]	Nominal
B31_3	Polis veya jandarmanın verdiği trafik hizmetinden memnun musunuz?	[1,2]	Nominal
B32_1	Kapkaç, yankesicilik vb. hırsızlık olayı yaşadınız mı?	[1,2]	Nominal
B32_2	Gasp olayı yaşadınız mı?	[1,2]	Nominal
B32_3	Yaralanma, darp olayı yaşadınız mı?	[1,2]	Nominal
B32_4	Aile fertlerinden kötü muamele gördünüz mü?	[1,2]	Nominal
B32_5	Şantaj, tehdit olayı yaşadınız mı?	[1,2]	Nominal
B32_6	Cinsel suçlardan mağduriyet yaşadınız mı?	[1,2]	Nominal
B32_7	Dolandırıcılıktan mağdur oldunuz mu?	[1,2]	Nominal
B32_8	Başka bir suçtan mağduriyet yaşadınız mı?	[1,2]	Nominal

B13_2 asayiş hizmetlerinden memnuniyet değişkenini ait girdi değişkenlerinin tamamı değişken indirgeme işlemi sonucunda da 0,944 gibi yüksek bir merit değerinde seçilmiştir. Dolayısıyla daha önce bahsettiğimiz değişken seçme işlemi öncesindeki sonuçlarla aynı sonuçlara ulaşılmıştır. Nitelik seçimi öncesi ile sonrası arasında herhangi bir fark olmadığından en iyi sonuç JRip algoritmasında % 74,97 doğruluk ile alınmıştır.

3.5. Sınıflandırma Sonucu Oluşan Kurallar Listesi

Aşağıda B39-Umut düzeyi, B13_1-Sağlık hizmetlerinden memnuniyet ve B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkenlerine ait en iyi sonuçları veren algoritmalara ait kural listeleri verilmiştir.

B13_2-Asayiş Hizmetlerinden Memnuniyet Hedef Değişkenine ait Kurallar Listesi

B13_2-Asayiş hizmetlerinden memnuniyet hedef değişkeni için en iyi sonuç nitelik indirgeme öncesi ve sonrasında aynı değişkenler kullanıldığı için değişmemiş ve JRip algoritmasında % 74,97 doğrulukla elde edilmiş ve 3 kural oluşmuştur. Kurallar aşağıda ifade edilmiştir:

1. **Eğer** polis veya jandarma olaylara zamanında müdahale etmiyorsa **ve** polis veya jandarmanın verdiği trafik hizmetinden memnun değil ise **ve** hayatta en çok mutlu eden şey sevgi ise **ve** polis ve jandarmanın vatandaşa davranışından memnun değil ise ve 31≤Yaş≤40 ise **o halde** asayiş hizmetlerinden MEMNUN DEĞİLdir. (38/17)

2. **Eğer** polis ve jandarmanın vatandaşa davranışından memnun değil ise **ve** eğitim durumu 4 yıllık yüksekokul veya fakülte ise **ve** son bir hafta içinde ücretli ya da ücretsiz bir işte çalışma durumu özel sektör çalışanı ise **ve** Yaş≤26 ise **o halde** asayiş hizmetlerinden ORTA DÜZEYDE MEMNUNdur. (13/4)

3. **Eğer** 1 ve 2 numaralı kurallar değilse **o halde** asayiş hizmetlerinden MEMNUNdur. (7861/1932)

B13_1-Sağlık Hizmetlerinden Memnuniyet Hedef Değişkenine ait Kurallar Listesi

B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkeni için en iyi sonuç nitelik indirgeme sonrasında JRip algoritmasında % 68.80 doğrulukla elde edilmiş ve 4 kural oluşmuştur. Kurallar aşağıda ifade edilmiştir:

1. **Eğer** muayene ve tahlil için beklemede sorun varsa **ve** temizlik ve hijyen konusunda sorun varsa **ve** yapılan muayeneden memnun değilse **ve** 25≤Yaş≤28 ise **o halde** sağlık hizmetlerinden ORTA DÜZEYDE MEMNUNdur. (33/16)

2. **Eğer** ilaç fiyatlarında sorun varsa **ve** temizlik ve hijyen konusunda sorun varsa **ve** eğitim durumu 4 yıllık yüksekokul veya fakülte ise **ve** 2012 yılında en son sağlık hizmeti aldığı sağlık kuruluşu devlet hastanesi ise **o halde** sağlık hizmetlerinden ORTA DÜZEYDE MEMNUNdur. (20/6)

3. **Eğer** muayene ve tahlil ücretlerinin yüksek buluyorsa **ve** eğitim durumu 4 yıllık yüksekokul veya fakülte ise **ve** son bir hafta içinde ücretli ya da ücretsiz bir işte çalışma durumu özel sektör çalışanı ise **ve** işyerinin iktisadi faaliyet kolu imalat sanayii ise **o halde** sağlık hizmetlerinden ORTA DÜZEYDE MEMNUNdur. (17/6)

4. **Eğer** 1, 2 ve 3 numaralı kurallar değilse **o halde** sağlık hizmetlerinden MEMNUNdur. (7799/2398)

B39-Umut Düzeyi Hedef Değişkenine ait Kurallar Listesi

B39-Umut düzeyi hedef değişkeni için en iyi sonuç nitelik indirgeme sonrasında SimpleCart algoritmasında % 70.36 doğrulukla elde edilmiş ve 21 kural oluşmuştur. Kurallar aşağıda ifade edilmiştir:

1. **Eğer** aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz veya çok güvensiz hissediyorsa **o halde** umut düzeyi UMUTLUdur. (3680/889)

2. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz veya çok güvensiz hissetmiyorsa **ve** adli

hizmetlerden memnuniyet memnun veya fikri yoksa **o halde** umut düzeyi UMUTLUdur. (252/129)

3. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz veya çok güvensiz hissetmiyorsa **ve** adli hizmetlerden memnuniyet memnun veya fikri yok değilse **ve** aylık hane halkı gelirinden memnuniyet çok memnunsa **o halde** umut düzeyi ÇOK UMUTLUdur. (19/6)

4. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz veya çok güvensiz hissetmiyorsa **ve** adli hizmetlerden memnuniyet memnun veya fikri yok değilse **ve** aylık hane halkı gelirinden memnuniyet çok memnun değilse **ve** hane halkı büyüklüğü 2 veya 4 veya 3 veya 12 veya 11 veya 9 veya 14 veya 13 veya 16 veya 15 ise **o halde** umut düzeyi UMUTLUdur. (44/35)

5. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz veya çok güvensiz hissetmiyorsa **ve** adli hizmetlerden memnuniyet memnun veya fikri yok değilse **ve** aylık hane halkı gelirinden memnuniyet çok memnun değilse **ve** hane halkı büyüklüğü 2 veya 4 veya 3 veya 12 veya 11 veya 9 veya 14 veya 13 veya 16 veya 15 değilse **o halde** umut düzeyi ÇOK UMUTLUdur. (13/9)

6. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yararlandığı sosyal güvenlik kurumu Banka Sandığı veya emekli sandığı veya BAĞ-KUR veya SSK ise **o halde** umut düzeyi UMUTLUdur. (1049/633)

7. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yararlandığı sosyal güvenlik kurumu Banka Sandığı veya emekli sandığı veya BAĞ-KUR veya SSK değilse **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz hissediyorsa **ve**

şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **o halde** umut düzeyi UMUTLUdur. (221/185)

8. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yararlandığı sosyal güvenlik kurumu Banka Sandığı veya emekli sandığı veya BAĞ-KUR veya SSK değilse **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz hissediyorsa **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil değilse **o halde** umut düzeyi UMUTLU DEĞİLdir. (11/9)

9. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun ise **ve** yararlandığı sosyal güvenlik kurumu Banka Sandığı veya emekli sandığı veya BAĞ-KUR veya SSK değilse **ve** yaşadığı çevrede gece yalnız yürürken kendini güvenli veya orta düzeyde güvenli veya güvensiz hissetmiyorsa **o halde** umut düzeyi UMUTLU DEĞİLdir. (31/53)

10. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun ise **ve** adli hizmetlerden memnuniyet memnun ise **o halde** umut düzeyi UMUTLUdur. (60/30)

11. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun ise **ve** adli hizmetlerden memnuniyet memnun değil ise ve sosyal güvenlik kurumuna kayıtlı değilse **ve** çalışılan işteki durum ücretsiz aile işçisi ise **o halde** umut düzeyi UMUTLUdur. (8/1)

12. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun ise **ve** adli hizmetlerden memnuniyet memnun değil ise ve sosyal güvenlik kurumuna kayıtlı değilse **ve** çalışılan işteki durum ücretsiz aile işçisi değil ise **o halde** umut düzeyi UMUTLU DEĞİLdir. (34/22)

13. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun ise **ve** adli hizmetlerden memnuniyet memnun değil ise ve sosyal güvenlik kurumuna kayıtlı değil değilse **ve** eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise ise **o halde** umut düzeyi UMUTLU DEĞİLdir. (13/10)

14. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun ise **ve** adli hizmetlerden memnuniyet memnun değil ise ve sosyal güvenlik kurumuna kayıtlı değil değilse **ve** eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise değil ise **o halde** umut düzeyi UMUTLUdur. (70/48)

15. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun değil ise ve eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise veya ilköğretim/ortaokul/mesleki ortaokul veya ilkokul veya doktora ise **ve** şimdiye kadar alınan eğitimden memnuniyet çok memnun veya memnun değil veya eğitim

almadı veya hiç memnun değil ise **ve** komşularla ilişkilerden memnuniyet memnun veya orta düzeyde memnun ise **ve** ulaştırma hizmetlerinden memnuniyet memnun veya çok memnun veya hiç memnun değilse **o halde** umut düzeyi UMUTLUdur. (15/18)

16. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun değil ise ve eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise veya ilköğretim/ortaokul/mesleki ortaokul veya ilkokul veya doktora ise **ve** şimdiye kadar alınan eğitimden memnuniyet çok memnun veya memnun değil veya eğitim almadı veya hiç memnun değil ise **ve** komşularla ilişkilerden memnuniyet memnun veya orta düzeyde memnun ise **ve** ulaştırma hizmetlerinden memnuniyet memnun veya çok memnun veya hiç memnun değil değilse **o halde** umut düzeyi HİÇ UMUTLU DEĞİLdir. (15/13)

17. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun değil ise ve eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise veya ilköğretim/ortaokul/mesleki ortaokul veya ilkokul veya doktora ise **ve** şimdiye kadar alınan eğitimden memnuniyet çok memnun veya memnun değil veya eğitim almadı veya hiç memnun değil ise **ve** komşularla ilişkilerden memnuniyet memnun veya orta düzeyde memnun değil ise **o halde** umut düzeyi UMUTLU DEĞİLdir. (12/9)

18. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun değil ise ve eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise veya ilköğretim/ortaokul/mesleki ortaokul veya ilkokul veya doktora ise **ve** şimdiye

kadar alınan eğitimden memnuniyet çok memnun veya memnun değil veya eğitim almadı veya hiç memnun değil değilse **o halde** umut düzeyi UMUTLUdur. (41/41)

19. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil ise **ve** evlilikten memnuniyet memnun değil ise ve eğitim durumu 4 yıllık yüksekokul/fakülte veya iki/üç yıllık yüksekokul veya yüksek lisans veya lise/mesleki lise veya ilköğretim/ortaokul/mesleki ortaokul veya ilkokul veya doktora değil ise **o halde** umut düzeyi UMUTLU DEĞİLdir. (77/60)

20. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil değilse **ve** aylık hane halkı gelirinden memnuniyet hiç memnun değil veya orta düzeyde memnun veya memnun veya çok memnun ise **o halde** umut düzeyi HİÇ UMUTLU DEĞİLdir. (16/12)

21. **Eğer** Eğer aylık hane halkı gelirinden memnuniyet memnun veya çok memnun veya orta düzeyde memnun değilse **ve** sağlıktan memnuniyet memnun veya çok memnun veya orta düzeyde memnun değil ise **ve** şimdiye kadar alınan eğitimden memnuniyet memnun veya orta düzeyde memnun veya çok memnun veya eğitim almadı veya memnun değil değilse **ve** aylık hane halkı gelirinden memnuniyet hiç memnun değil veya orta düzeyde memnun veya memnun veya çok memnun değilse **o halde** umut düzeyi UMUTLU DEĞİLdir. (10/9)

4. BÖLÜM

BULGULAR

TÜİK 2003 yılından itibaren gerçekleştirdiği YMA anketinin her sene sonuçlarını yayınlarken, her anket sorusunu kendi içinde değerlendirerek bazı istatistikî bilgiler sunmaktadır. Ancak araştırılan tüm anket sorularını birbirleri ile anlamlı gruplar oluşturup bir bütün olarak inceleyerek genel yargılarda bulunma yoluna gidilmek istendiği zaman, TÜİK tarafından sunulan raporlar yetersiz kalacak ve veri madenciliği devreye girecektir.

Bu çalışmada TÜİK tarafından bireylerin genel mutluluk algılamasını, toplumsal değer yargılarını, temel yaşam alanlarındaki genel memnuniyetini ve kamu hizmetlerinden memnuniyetini ölçmek, memnuniyet düzeylerinin zaman içindeki değişimini takip etmek üzere gerçekleştirilen YMA anketi kapsamında sorulan sorular yardımıyla veri madenciliği sınıflandırma çalışması yapılarak, yukarıda bahsedilen konuların birbirleri ile olan ilişkileri incelenip her konu ile ilgili araştırılan sorulara ait elde edilen cevaplardan anlamlı kurallar oluşturulmaya çalışılmıştır.

Sınıflandırma çalışması kapsamında öncelikle anketteki soruları temsil etme gücü yüksek olan B39-Umut düzeyi, B13_1-Sağlık hizmetlerinden memnuniyet ve B13_2-Asayiş hizmetlerinden memnuniyet değişkenleri hedef değişken olarak belirlenmiştir. Kullanılacak algoritmalar olarak, veri ön işleme çalışmalarını müteakip sınıflandırmada en iyi sonuçları veren PART, JRip, J48, REPTree ve SimpleCart algoritmaları seçilmiştir. Veri bütünleştirme, tutarsız ve eksik verilerin silinmesi, boş değerlerin tamamlanması gibi uygulanabilecek tüm veri ön işleme çalışmasından sonra sınıflandırma gerçekleştirilmiş ve sonuçlar yorumlanmıştır. Alınan sonuçların daha da iyileştirilebileceği düşüncesiyle nitelik indirgeme işlemi de yapılmış bu işlemde ise daha iyi merit (seçilen girdi değişkenlerinin hedef değişkenini temsil etmedeki

yararlılığı) değeri verdiği için ConsistencySubsetEval yöntemi, nitelik seçme modu olarak ise direk en etkili girdi değişkenlerini veren "Use trainig mode" kullanılmıştır. Araştırma metodu olarak ise belirli bir kural seçilerek en umut verici düğümün genişletilmesi yardımıyla bir grafik araştırıp en iyi çözümü bulana kadar işlemi tekrarlayan "BestFirst" metodu kullanılmıştır. Nitelik seçimi sonucu elde edilen niteliklerle de sınıflandırma gerçekleştirilip bir önceki duruma göre kıyaslama yapılmıştır. B39-Umut düzeyi hedef değişkeni için % 70, B13_1-Sağlık hizmetlerinden memnuniyet hedef değişkeni için % 69 ve B13_2- Asayiş hizmetlerinden memnuniyet hedef değişkeni için % 75 doğruluğa ulaşılmıştır.

Genel olarak bakıldığında tüm algoritmalar hedef değişkene ait sınıflardan, hangisine ait örnek daha fazla ise o sınıf değerine ait sonuçları en iyileme yoluna gitmiş (default class) diğer sınıflara ait sonuçlar istenilen seviyede olmamıştır. Bu da doğrudan ağırlıklı performans ölçütlerinin düşük olmasına neden olmuştur.

Dikkate değer ikinci bir nokta ise nitelik indirme işlemi sırasında seçilen girdi değişkenlerine ait merit değeri çok iyi bir değer olarak bulunurken bu girdi değişkenleri ile alınan sınıflandırma sonuçları, nitelik indirgeme işlemi öncesinde alınan sınıflandırma sonuçlarına göre çok fazla değişmemiştir. Bu durum ise hedef değişkene ait her bir sınıf değerine verilen cevapların birbiri ile uyuşmamasından yani aynı sınıfa giden bir girdi değişkenine birbirinden farklı birçok cevabın verilmesinden kaynaklanmaktadır.

5. BÖLÜM

SONUÇ, TARTIŞMA ve ÖNERİLER

Çok büyük boyutlu veri yığınlarında gizli olan bilgiyi çıkarmak, veriden anlamlı bilgiler elde etmek için uygulanacak yöntem veri madenciliğidir. Veri madenciliği işlemi, elde bulunan veri kaynaklarına dayanarak daha önce keşfedilmemiş bilgileri ortaya çıkarma, bunlara göre karar verme ve senaryo planları oluşturmaktır. Bu çalışmada veri madenciliği görevlerinden en çok kullanılan biri olan sınıflandırma, TUİK verileri üzerinde gerçekleştirilmiştir.

Kalkınma Bakanlığı'na bağlı TÜİK, istatistik alanında ulusal ve uluslararası öncelikleri dikkate alarak, istatistiksel veri ihtiyacı duyulan alanları ve veri derleme yöntemlerini ilgili kurum ve kuruluşlarla işbirliği ile belirleyerek, ihtiyaç duyulan alanlarda veri üretilmesini ve analizini sağlamaktadır. TÜİK, doğrudan veri madenciliğinin uygulanmadığı ancak veri madenciliği çalışmaları için büyük kapasiteli veri ambarı deposu görevi gören önemli bir kuruluştur. TÜİK sanayi, tarım, ticaret, enflasyon, inşaat ve konut, çevre ve enerji, ulaştırma ve haberleşme, nüfus ve demografi, sağlık ve sosyal koruma vb. birçok alanda istatistiksel araştırmalar ve analizler yapmaktadır. Bu çalışmada kullanılan mikro veri seti nüfus ve demografi alanında toplumsal yapı ve cinsiyet istatistiklerine bağlı olarak araştırılan Yaşam Memnuniyeti Anketi'dir.

TÜİK YMA'da bireylerin kişisel analizinden yaşadığı çevreye, belediye hizmetlerinden gelecekle ilgili beklentilere, toplumsal baskı algısından toplumsal ilgi alanlarına kadar pek çok konuda bilgiyi araştırmıştır. Tez çalışması kapsamında bireylerin umut düzeyi, sağlık hizmetlerinden ve asayiş hizmetlerinden memnuniyet durumları sınıflandırma kuralları ile ortalama % 70 doğruluklarla elde edilmiştir

YMA'da bireylerin genel memnuniyet durumları ve umut düzeylerini etkileyen durumlar ayrıntılı olarak incelenmiş iken sağlık hizmetleri, adli hizmetler ve asayiş

hizmetleri ile ilgili araştırma alanları yetersiz kalmıştır. Eğitim hizmetleri ile ilgili araştırmalar fert bazında değil hane bazında incelenmiş, ulaştırma hizmetleri, elektronik ortamlarda sunulan kamu hizmetleri ile ilgili hemen hemen hiçbir ayrıntılı araştırma yapılmamıştır. Adli hizmetlerden memnuniyet, sağlık hizmetlerinden memnuniyet ve asayiş hizmetlerinden memnuniyet hedef değişkenleri için seçilen girdi değişkenlerinde merit (yararlılık) değeri 1 ya da 1'e çok yakın değerler bulunmasına rağmen bu hedef değişkenler için istenilen düzeyde doğruluklar elde edilememiştir. Bunun en büyük nedeni ise anketlerde hedef değişkeni tam olarak ifade edecek girdi nitelikleriyle ilgili soruların yukarıda da değinildiği gibi sorulmamasıdır.

Elde edilen doğrulukların beklenenden düşük çıkmasındaki başka bir neden ise bireyler arasındaki görüş farklılıklarıdır. Bu konuya verilebilecek en bariz örneklerden birisi, aynı hane halkı büyüklüğüne, aylık aynı hane halkı gelirine ve diğer şartlara sahip iki hane kıyaslandığında bir hane için bu gelirle hanenin temel ihtiyaçlarını karşılama düzeyi kolay iken diğer hane için zor olabilmesidir. Ya da bu hanelere ait aylık hane halkı gelirinden memnuniyet durumunun birisinde memnun şeklinde iken diğerinde memnun değil şeklinde olabilmesidir. Bu durum değişkenler üzerinde yapılan nitelik seçimi ve diğer veri ön işleme işlemlerine rağmen doğruluğun çok fazla değişmemesine neden olmaktadır.

YMA'da bireylerin genel memnuniyet durumları ve umut düzeyi hedef değişkenleri için yapılan sınıflandırmada, örneklerin çoğunluğunun bir sınıfa ait olması ve diğer sınıflara ait az örnek olması (dengesiz sınıflandırma problemi) sebebiyle sınıflandırıcılar yoğun örnek grubuyla ilgili sınıflandırma yapmaya yönelmektedir. Böylesi dengesiz sınıflandırma problemleri için maliyet-duyarlı sınıflandırma çalışmak daha anlamlıdır.

Dolayısıyla gelecek çalışmalar olarak ilk etapta ilgili veri kümesi ile maliyet-duyarlı sınıflandırma hedeflenmektedir. Doğruluk yüzdelerinin daha da istenilen düzeye getirilebilmesi amacıyla çalışma kapsamında ele alınan tek bir yıla ait anketler 3-5 yılı içeren anketler şeklinde genişletilerek veri kümesinin artırılması, bu amaçla TUİK'den yeni mikro veri talebinde bulunulması planlanmaktadır.

KAYNAKLAR

1. Cohen, William W., 1995. "Fast Effective Rule Induction", *Twelfth International Conference on Machine Learning*, 115-123.
2. Adriaans, P., Zantinge, D., 1997. Data Mining, , Boston, MA, USA **Addison Wesley Longman Publishing**.
3. Savaş, S., Topaloğlu, N., Yılmaz, M., 2012. Veri madenciliği ve Türkiye'deki uygulama örnekleri**, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi** Yıl:11 Sayı: 21 Bahar 2012 s. 1-23
4. Yarımağan, Ü., 2000. Veri tabanı sistemleri, **Akademi&Türkiye Bilişim Vakfı**, Ankara, s.7-9.
5. Aydoğan, E.K., 2008. Veri madenciliğinde sınıflandırma problemleri için evrimsel algoritma tabanlı yeni bir yaklaşım: Rough-Mep algoritması, Doktora tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.
6. İçeli, Namık., 2012. The Success Analysis of Basic Computer Course of Divriği Nuri Demirağ Vocational College's Students by Using Data Mining Method, **Mesleki Bilimler Dergisi, ISSN:2146-7420, MBD 2012, 1**(1): 18 – 37
7. Öğüt, S., 2009. Veri madenciliği kavramı ve erişim süreci. **http://www.sertacogut.com/papers/Sertac_Ogut__Veri_Madenciligi_Kavrami_ve_Gelisim_Sureci.pdf**
8. Özdamar, E.Ö., 2002. Veri madenciliğinde kullanılan teknikler ve bir uygulama. Yayımlanmamış Yüksek Lisans Tezi. İstanbul: Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı.
9. Fayyad, U.M., Shapiro, G.P., Smyth, P., 1996. From data mining to knowledge discovery in databases. **AI Magazine 17(3)**, 37-54. http://www.daedalus.es/fileadmin/daedalus/doc/MineriaDeDatos/fayyad96.pdf
10. Surajit, A.N., Surajit, C., Bernhardt, J. ve Fayyad, U., 2000. Integration of data mining and relational databases, *Proceeding of the 26th Conference on Very Large Databases*, http://www.tcs-trddc.com/tecs/integration-of-data-mining.pdf
11. Show, M.J., Subramaniam, C., Tan, G.W. ve Welge, M.E., 2001. Knowledge management and data mining for marketing. **Decision support systems, 31**, 127-137.

12. Thealing, K., 2009. Data mining and analytic technologies. http://www.thearling.com/index.htm
13. Lindell, Y., Israel, R. ve Pinkas, B., 2000. Privacy preserving data mining. Journal of Cryptology. http://www.aladdin.cs.cmu.edu/workshops/privacy/slides/pdf/linell_pinkas.pdf
14. Müller, J.A. ve Lemke, F., 2000. Self-Organising data mining. Libri Books on Demand. http://www.knowledgeminer.com/pdf/sodm.pdf
15. DeRosa, M., 2004. Data mining and data analysis for counterterrorism. Center for Strategic and International Studies (CSIS). http://csis.org/files/media/csis/pubs/040301_data_mining_report.pdf
16. Çankırı, S., Kartal, E., Yıldırım, K., Gülseçen, S. , 2009. *ÜNAK 2009 Bilgi Çağında Varoluş: "Fırsatlar ve Tehditler" Sempozyumu*, Yeditepe Üniversitesi, İstanbul - TÜRKİYE, Bildiriler Kitabı.
17. Murray, J., Mackinnon ve Ned Glick, 1999.'Data Mining and Knowledge Discovery in Databases- An Overview', J.Statists., Vol.41, No.3, s.260.
18. http://web.sakarya.edu.tr/~cagil/e-isletme/VeriAmbarlari.ppt
19. Altan, M., Veri tabanı yönetimi veri ambarı ders notu, Trakya Üniversitesi Bilgisayar Mühendisliği
20. David L., Olson D. D., Advanced Data Mining Techniques p-11- 53
21. Han J., Kamber M., 2006, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher San Francisco
22. Lori, B. A., 2006.Data Mining for Information Professionals.
23. Argüden, Y., Erşahin, B., 2008, Veri Madenciliği Veriden Bilgiye, Masraftan Değere, ARGE Danışmanlık Yayınları No: 10, ISBN: 978-975-93641-9-9 1. Basım, s. 17
24. Baykal, A., 2006, Application Fields of Data Mining, **D.Ü.Ziya Gökalp Eğitim Fakültesi Dergisi, 7**, 95-107
25. Christensen M., Hermiz K., Manganaris S., Zerkle D., 2000, "A data mining analysis of RTID alarms", Computer Networks34, s.571-577.
26. Bing, L., 2007, "Web data mining: Exploring Hyperlinks, Contents and Usage Data", ISBN-10 3-540-37881-2, Springer-Verlag Berlin Heidelberg.
27. Güvenç, E., 2001, "Student performance assesment in higher education using data mining", Master tezi, Boğaziçi Üniversitesi, İstanbul, s.5-14.

28. Hudairy H., 2004, "Data mining and decision making support in the governmental sector", Faculty of Graduate School of The University of Louisville, Kentucky.

29. Soğukpınar İ., Takçı H., 2002, "Kütüphane kullanıcılarının erişim örüntülerinin keşfi", **Bilgi dünyası dergisi, Cilt:3, sayı:1**, s.12-26.

30. Delioğlu S., Dolgun M.Ö., Özdemir T.G., 2007 , "Öğrenci Seçme Sınavında (ÖSS) öğrencilerin tercih profillerinin veri madenciliği yöntemleriyle tespiti", *Bilişim 07 kongresi*, Ankara.

31. Silahtaroğlu, G., (2008). "Veri madenciliği – Sınıflandırma teknikleri ve algoritmalar", Kavram ve algoritmalarıyla temel veri madenciliği, Papatya yayın, İstanbul, s.9-82.

32. Özkan Y., 2008. "Karar ağaçları ile sınıflandırma-Sınıflandırma ve regresyon ağaçları-Kümeleme-Birliktelik kuralları", Veri madenciliği yöntemleri, Editör: R.Çölkesen, Papatya yayın, İstanbul, 51-166.

33. Tapkan, P., Özbakır, L., Baykasoğlu, A., 2011**.** Weka ile Veri Madenciliği Süreci ve Örnek Uygulama, *Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi*, s. 248-249

34. Han, J., Kamber, M., 2011. Data Mining Concepts and Techniques,; Third Edition, Morgan Kaufman Elsevier Inc, ABD

35. Stepaniuk, J. R., 2008. Granular Computing in Knowledge Discovery and Data Mining, Springer.

36. Maimon, O., Rokach, L., 2007. Data Minining And Knowledge Discovery Handbook, Springer Science Business Media Inc., New York.

37. Oğuzlar, A., 2003. "Veri Ön İşleme", **Ege Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Sayı 21**, Temmuz-Aralık 2003, s.73, http://iibf.erciyes.edu.tr/dergi/sayi21/ aoguzlar.pdf

38. Khemka, A., 2003. A Colloborative Predictive Data Mining Model, Yayınlanmamış Yüksek Lisans Tezi, Faculty of University of Missouri-Kansas City, Missouri

39. Dondurmacı, G. A., Çınar, A., 2014. Finans Sektöründe Veri Madenciliği Uygulaması, **Akademik Sosyal Araştırmalar Dergisi**, **Sayı: 2/1**, s.258-271

40. Kaya, H., Köymen, K., 2008.Veri Madenciliği Kavramı ve Uygulama Alanları, Doğu Anadolu Bölgesi Araştırmaları

41. Acar Şaylan, Ç., 2013. "Böbrek Nakli Geçirmiş Hastalarda Akıllı Yöntem Tabanlı Yeni Öznitelik Seçme Algoritması Geliştirilmesi", Kadir Has Üniversitesi Fen Bilimleri Enstitüsü
42. Wang, S., ve Wang H., 2002. "Knowledge Discovery Through Self-Organizing Maps: Data Visualization And Query Processing", Knowledge And Information Systems, 4.
43. Koltan Yılmaz, Ş., Patır, S., 2011. Kümeleme Analizi Ve Pazarlamada Kullanımı, **Akademik Yaklaşımlar Dergisi Cilt:2** Sayı:1
44. Albayrak, A., Yılmaz, Ş.K., 2009. "Veri Madenciliği: Karar Ağaçları ve İMKB Verileri Üzerine Bir Uygulama", **Süleyman Demirel Üniversitesi İktisadi Ve İdari Bilimler Fakültesi Dergisi, 14(1)**, ss. 31–52.
45. Özekes, S., Çamurcu, A.Y., 2002. Türkiye Veri Madenciliğinde Sınıflama ve Kestirim Uygulaması, **Marmara Üniversitesi Fen Bilimleri Dergisi, 18**
46. Alan, M. A., 2012. Veri Madenciliği Ve Lisansüstü Öğrenci Verileri Üzerine Bir Uygulama, **Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, Sayı 33**
47. Nisbet R., Elder J., Miner, G., 2009. Handbook of Statistical Analysis And Data Mining Applications, Elsevier, London
48. Coşkun, C., Baykal, A., 2011. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması, *Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri*, Malatya
49. Quinlan, J.R., 1986. "Induction of Decision Trees ", Journal of Machine Learning, vol. 1, 81-106,
50. Quinlan, J.R., 1993. "C4.5: Program for Machine Learning", Morgan Kaufmann Publishers, San Mateo, CA,
51. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. "Classification and Regression Trees", Wadsworth, Belmont,
52. Ada, M. B., Altunay, F., Civelek, M., Kaplan, S., Koç, P., Kümeleme Analizi
53. Öz, B., Taban, S., Kar, M., 2008. Kümeleme Analizi ile Türkiye ve AB Ülkelerinin Beşeri Sermaye Göstergeleri Açısından Karşılaştırılması, **Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 10(1)**
54. Özdamar, K., 1999. Paket Programlarla İstatistiksel Veri Analizi 2, Kaan Kitabevi, Eskişehir.

55. Çakmak, Z., Uzgören, N., Keçek, G., 2005. **Kümeleme Analizi Teknikleri ile İllerin Kültürel Yapılarına Göre Sınıflandırılması ve Değişimlerinin İncelenmesi, Sayı:12**, 15-36

56. Altun Ada, A., 2011. Kümeleme Analizi ile AB Ülkeleri Ve Türkiye'nin Sürdürülebilir Kalkınma Açısından Değerlendirilmesi **Dumlupınar Üniversitesi, Sosyal Bilimler Enstitüsü Dergisi, Sayı 29**

57. Sibson, R., 1973. "SLINK: An Optimally Efficient Algorithm for the Single Link Cluster Method", **The Computer Journal, vol. 16(1)**, 30-34

58. Guha, S., Rastogi, R., Shim, K., 1998. "CURE: An Efficient Clustering Algorithm for Large Databases", in: *Proceedings of the 1998 ACM SIGMOD Intenational Conference on Management Data*, vol. 27(2), 73-84

59. Karypis, G., Han, E., Kumar, V., 1999. "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", IEEE Computer, vol. 32(8), 68-75

60. Zhang, T., Ramakrishnan, R., Livny, M., 1996. "BIRCH: An Efficient Data Clustering Method for Very Large Databases", in: *Proceedings of the 1996 ACM SIGMOD International Conference on Management Data*, vol. 25(2)

61. Macqueen, J., 1967. "Some Methods for Classification and Analysis of Multivariate Observations", in: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 281-297

62. Kauffman, L., Rousseeuw, P.J., 1990. "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley and Sons

63. Raymond, T.NG, Han, J., 1994. "Efficient and Effective Clustering Methods for Spatial Data Mining", in: *Proceedings of the 20th VLDB Conference*, 144-155

64. Altınışık, U., 2006. "Öğrenci bilgi sisteminde veri madenciliğinin uygulanması", yayınlanmamış master tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli, s.12-17.

65. Ghosh A., Nath B., 2004. "Multi objective rule mining using genetic algorithms", **Information sciences, cilt:163**, sayı1-3, s.123-133.

66. Dolgun, M. O., 2006. "Buyuk Alısveris Merkezinden Yapılan Satıslar İçin Sepet Analizi", Hacettepe Universitesi Fen Bilimleri Enstitusu, Yuksek Lisans Tezi, Ankara

67. Şeker, Ş. E., 2013. İş Zekası ve Veri Madenciliği (Weka ile), ISBN 9786051276717. Cinius

68. http://analyticstraining.com/2011/10-most-popular-analytic-tools-in-business/

69. http://ieeexplore.ieee.org/search/searchresult.jsp?reload=true&queryText=weka

70. Piatetsky-Shapiro, G., 2005. "KDnuggets news on SIGKDD Service Award 2005", http://www.kdnuggets.com/news/2005/n13/2i.html

71. http://tr.wikipedia.org/wiki/Weka

72. Kuo, R.J., Ho, L.M., Hu, C.M., 2002. "Cluster Analysis In Industrial Market Segmentation Through Artificial Neural Network", **Computers&Industrial Engineering, Vol:42**, Issue:2-4, s. 393.

73. HSU, C.-H., 2009. "Data Mining to Improve Industrial Standards and Enhance Production and Marketing: An Empirical Study in Apparel Industry", **Expert Systems with Applications, Vol:36**, Issue:3, s. 4186; MA ve CHOW, s. 504.

74. Taşkın, Ç., Emel, G. G., 2010. Veri Madenciliğinde Kümeleme Yaklaşımları ve Kohonen Ağları ile Perakendecilik Sektöründe bir Uygulama, **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, C.15**, S.3 s.395-409.

75. Kayaalp, K., 2007. Asenkron Motorlarda Veri Madenciliği ile Hata Tespiti, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü.

76. İnan, A., 2006. Privacy Preserving Distributed Spatio-Temporal Data Mining, Yüksek Lisans Tezi, Sabancı University, Computer Science and Engineering,

77. Yavaş, G., 2003. Using A Data Mining Approach For The Prediction of User Movements in Mobile Environments, Yüksek Lisans Tezi, Bilkent University, Institute of Engineering and Science.

78. Çalışkan, S.K., Soğukpınar, İ., 2008. "KxKNN: K-Means ve K En Yakın Komşu Yöntemleri ile Ağlarda Nüfuz Tespiti", *2. Ağ ve Bilgi Güvenliği Sempozyumu*, 16-18 Mayıs, Girne, 120-124.

79. Duru, N., Canbay, M., 2007. "Veri Madenciliği ile Deprem Verilerinin Analizi", *International Earthquake Symposium*, Kocaeli, 556-560.

80. Doğan, Y., 2004. A Data Mining Based Classification Algorithm for Tactical Underwater Sensor Networks, Yüksek Lisans Tezi, Turkish Naval Academy, Computer Engineering.

81. Sıramkaya, E., 2005. Veri Madenciliğinde Bulanık Mantık Uygulaması, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü.

82. Kastro, Y., 2006. A Defect Prediction Method For Software Versioning, Yüksek Lisans Tezi, Boğaziçi University, Computer Engineering.

83. Erdem, C., 2006. Density Based Clustering Using Mathematical Morphology, Yüksek Lisans Tezi, Middle East Technical University, Information Systems.

84. Bilgin, T.T., 2009. "Veri Akışı Diyagramları Tabanlı Veri Madenciliği Araçları ve Yazılım Geliştirme Ortamları", Akademik Bilişim 09, 11-13 Şubat, Harran Üniversitesi, Şanlıurfa, 807-814.

85. Toprak, S., 2004. Data Mining For Rule Discovery in Relational Databases, Middle East Technical University, Computer Engineering.

86. Baloğlu, U.B., 2006. DNA Sıralarındaki Tekrarlı Örüntülerin ve Potansiyel Motiflerin Veri Madenciliği Yöntemiyle Çıkarılması, Fırat Üniversitesi, Fen Bilimleri Enstitüsü.

87. Yıldız, B., 2010. Impacts Of Frequent Itemset Hiding Algorithms On Privacy Preserving Data Mining, İzmir İnstitute of Technology, Computer Engineering.

88. Kılınç, Y., 2009. Mining Association Rules For Quality Related Data In An Electronics Company, Middle East Technical University, Industrial Engineering.

89. Aksoy, B., 2009. Cluster Analysis Of Decompression Illness, Galatasaray University, Institute of Science and Engineering.

90. Yıldırım, P., 2008. Uludağ, M. ve Görür, A., "Hastane Bilgi Sistemlerinde Veri Madenciliği", Akademik Bilişim 2008, 30 Ocak - 01 Şubat, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 429-434.

91. Doğan, Ş., Türkoğlu, İ., 2008. "Iron-Deficiency Anemia Detection From Hematology Parameters By Using Decision Trees", **International Journal of Science & Technology, Cilt 3**, No 1, 85-92.

92. Danacı, M., Çelik, M., Akkaya, A.E., 2010. "Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi", *Akıllı Sistemlerde Yenilikler ve Uygulama Sempozyumu*, 21-24 Haz. 2010, Kayseri, 9-12.

93. Ata, N., Özkök, E., Karabey, U., 2008. "Survival Data Mining: An Application To Credit Card Holders", **Sigma Mühendislik ve Fen Bilimleri Dergisi, Cilt 26**, No 1, 33-42.

94. Ata, A.H., Seyrek, İ.H., 2009. "The Use of Data Mining Techniques in Detecting Fraudulent Financial Statements: An Application on Manufacturing Firms", **Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt 14**, No 2, 157-170.

95. Savaşçı, İ., Tatlıdil, R., 2006. "Bankaların Kredi Kartı Pazarında Uyguladıkları CRM (Müşteri İlişkiler Yönetimi) Stratejisinin Müşteri Sadakatine Etkisi", **Ege Akademik Bakış Dergisi, Cilt 6**, No 1, 62-73.

96. İnan, O., 2003. Veri Madenciliği, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü.

97. Ayık, Y.Z., Özdemir, A. ve Yavuz, U., 2007. "Lise Türü ve Lise Mezuniyet Başarısının Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi", **Sosyal Bilimler Enstitüsü Dergisi, Cilt 10**, No 2.

98. Bozkır, A.S., Sezer, E. ve Gök, B., 2009. "Öğrenci Seçme Sınavında (ÖSS) Öğrenci Başarımını Etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle Tespiti", *5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09)*, 13-15 Mayıs, Karabük Üniversitesi, Karabük, 37-43.

99. Erdoğan, Ş.Z. ve Timor, M., 2005. "A Data Mining Application In A Student Database", **Journal Of Aeronautics and Space Technologies, Cilt 2**, No 2, 53-57.

100. Özçınar, H., 2006. KPSS Sonuçlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi, Yüksek Lisans Tezi, Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü.

101. Bozkır, A.S., Sezer, E., 2009. "Usage of Data Mining Techniques in Discovering The Food Consumption Patterns of Students and Employees of University", *Balkan-Kafkas ve Türk Devletleri Uluslararası Mühendislik Sempozyumu*, 22-24 October, Isparta, 104-109.

102. Kayri, M., 2008. "Elektronik Portfolyo Değerlendirmeleri İçin Veri Madenciliği Yaklaşımı", **Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, Cilt 5**, No 1, 98-110.

103. Kalıkov, A., 2006. Veri Madenciliği ve Bir E-Ticaret Uygulaması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

104. Akbulut, S., 2006. Veri Madenciliği Teknikleri ile Bir Kozmetik Markanın Ayrılan Müşteri Analizi Ve Müşteri Segmentasyonu, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü.

105. Özçakır, F.C., Çamurcu, A.Y., 2007. "Birliktelik Kuralı Yöntemi İçin Bir Veri Madenciliği Yazılımı Tasarımı ve Uygulaması". **İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, No 12**, 21-37.

106. Gürbüz, F., Özbakır, L., Yapıcı, H., 2009. "Türkiye'de Bir Havayolu İşletmesine Ait Parça Söküm Raporlarına İlişkin Veri Madenciliği Uygulaması", **Gazi Üniversitesi Mimarlık Mühendislik Fakültesi Dergisi, Cilt 24**, No 1, 73-78.

107. Ulaş, M.A., 2001. Market Basket Analysis For Data Mining, Yüksek Lisans Tezi, Boğaziçi University, Computer Engineering.

108. Güntürkün, F., 2007. A Comprehensive Review Of Data Mining Applications İn Quality İmprovement And A Case Study, Yüksek Lisans Tezi, Middle East Technical University, Statistics.

109. Tüzüntürk, S., 2010. Veri Madenciliği ve İstatistik, **Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi Cilt XXIX, Sayı 1**, s. 65-90

110. Güven, Z. B., Bilgin, T. T., Zaman Serileri Madenciliği Kullanılarak Nüfus Artışı Tahmin Uygulaması, Maltepe Üniversitesi, Yazılım Mühendisliği Bölümü, İstanbul

111. Değirmenci, T., 2014. Resmi İstatistiklerde Veri Madenciliği Yaklaşımı, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü Yüksek Lisans Tezi

112. Lavrac, N., Bohanec, M., Pur, A., Cestnik, B., Debeljak, M., Kobler, A., July 2006. Data mining and visualization for decision support and modeling of public health-care resources, Jozef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia, **Journal of Biomedical Informatics, 40** 438–447

113. Alsultanny, Y. A., 2013. Labor Market Forecasting by Using Data Mining, *International Conference on Computational Science, Arabian Gulf University*, Manama, Kingdom of Bahrain, Procedia Computer Science 18, 1700 – 1709

114. Torres-Avilês, F., Romeo, J. S. ve Lôpez-Kleine, L., 2014. Data mining and influential analysis of gene expression data for plant resistance gene identification in tomato (Solanum lycopersicum), Electronic Journal of Biotechnology, Departamento de Matemática y Ciencia de la Computación, Universidad de Santiago de Chile, Av. Libertador Bernardo O'Higgins 3363, Santiago, Chile

115. Aljumah, A. A., Ahamad, M. G., Siddiqui, M. K., 2012. Application of data mining: Diabetes health care in young and old patients, Salman bin Abdulaziz University, Saudi Arabia, Journal of King Saud University, **Computer and İnformation Sciences 25**, 127-136

116. Bulcke, T. V., Broucke, P. V., Hoof, V. V., Wouters, K., Broucke, S. V., Smits, G., Smits, E., Proesmans, S., Genechten, T. V., Eyskens, F., 2011. Data mining

methods for classification of Medium-Chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data, i-ICT, University Hospital Antwerp, Wilrijkstraat 10, 2650 Edegem, Belgium, **Journal of Biomedical Informatics, 44** 319-325

117. Khan, I., Capozzoli, A., Corgnati, S. P., Cerquitelli, T., 2013. Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques, The Mediterranean Green Energy Forum 2013, MGEF-13, **Energy Procedia, 42** 557 – 566, TEBE Research Group, Department of Energy (DENERG), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
118. Pakgohar, A.*[a], Tabrizi, R. S.*[b], Khalili, M.*[c] , Esmaeili, A.*[d], 2010. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach, **Procedia Computer Science, 3** 764–769, [a]Faculty of Statistics, Payame Noor University(PNU),Branch of Ardakan,Iran, [b]MBA, Multimedia University, FOM, Cyberjaya, Malaysia, [c]Departement of Statistics , Payame Noor University(PNU), Branch of Meybod, Iran, [d]Faculty of Traffic, Oloome Entezami University(Police), Tehran, Iran.
119. Chanthaweethip, W., Guha, S., 2012. Temporal Data Mining and Visualization for Treatment Outcome Prediction in HIV Patients, *Proceedings of the International Neural Network Society Winter Conference* (INNS-WC 2012), Asian Institute of Technology, Pathumthani 12120, Thailand
120. Zhang, D.,ve Jiang, K., 2012. Application of Data Mining Techniques in the Analysis of Fire Incidents, *International Symposium on Safety Science and Engineering in China*, 2012 (ISSSE-2012), **Procedia Engineering, 43** 250 – 256, School of Safety & Environmental Engineering, Capital University of Economics & Business, Beijing 100070, China
121. Erdem, S., 2011. İleri Veri Tabanları Sistemleri Dersi Genel Seçim Çalışmaları, Çanakkale 18 Mart Üniversitesi, Bilgisayar Mühendisliği Dönem Projesi
122. Akpınar H., 2000. "Veri tabanlarında bilgi keşfi ve veri madenciliği", **İ.Ü. İşletme Fakültesi Dergisi, C:29**, sayı:1/ s.1-22 .
123. Chiu T., Wei C., 2002. "Turning telecommunications call details to churn prediction: a data mining approach", **Expert systems with applications, 23** s.103-102.
124. Mitchell, T., 1997. Machine learning, McGraw-Hill International, London, s.52-81.

125. Fiske, J., 1998. Introduction to communication studies, y.y., Routledge.

126. Dunham, M.H., 2003). Data mining ıntroductory and advanced topics, Pearson education inc., Prentice Hall, New Jersey, s.8.

127. Shannon, C.E., 1948. A Mathematical theory of communication, **The Bell system technical journal, vol.27**, s.379-423, 623-656.

128. Kantardzic, M., 2003. "Chapter 9: Artificial Neural Networks Chapter 1-1.4", Data Mining Concepts, Models, Methods and Algorithms, John Wiley & Sons.

129. Tan, P.-N., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining, Pearson, Addison-Wesley, Boston, MA, USA.

130. Goh, S. L., Mandic, D. P., 2003. "Recurrent Neural Networks with Trainable Amplitude of Activation Functions", **Neural Networks, 16** 1095-1100.

131. Ma, L., Khorasani, K., 2003. "A New Strategy for Adaptively Constructing Multilayer Feedforward Neural Networks", **Neurocomputing, 51** 361-385.

132. Irmak, S., Köksal, C. D., Asilkan Ö., 2012. "Predicting Future Patient Volumes of The Hospitals By Using Data Mining Methods", **International Journal of Alanya Faculty of Business**, **Vol:4**, No:1, s. 101-114

133. **http://kodcu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/**

134. Emel, G. G., Taşkın, Ç., 2005. "Veri Madenciliğinde Karar Ağaçları ve bir Satış Analizi Uygulaması", **Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi**, Cilt: 6 Sayı: 2

135. Lee, S. J., Keng, S., 2001. "A Review Of Data Mining Techniques", **Industrial Management&Data Systems 1**(101), s. 41-46.

136. Bounsaythip, C., Esa, R. R., 2001. "Overview of Data Mining For Customer Behavior Modeling", **VTT Information Technology Research Report**, Version:1, s. 1-53.

137. Oğuzlar, A., 2004. "CART Analizi ile Hanehalkı İşgücü Anketlerinin Özetlenmesi", **Uludag Üniversitesi iktisadi ve İdari Bilimler Dergisi**, Sayı: 3-4

138. Answer Tree 3.0 User's Guide 2001. SPSS Inc., USA.

139. Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining, MIT Press, USA.

140. Ahola, J., Rinta-Runsala E., 2001. "Data Mining Case Studies İn Customer Profiling", **VTT lnfonnation Technology Research Report**, TIEI-2001-29.

141. Apte, C., Weiss, S., 1997. "Data mining with decision trees and decision rules", **Future Generation Computer Systems, 13**, ss.197-210.

142. http://www.ifpri.orglthemes/mpı8/teChguidltg03.pdf

143. http://ab.org.tr/ab08/bildiri/71.pdf

144. http://www.metinmadenciligi.com/kaynaklar/mm1.pdf

145. http://www.mademir.com/2010/12/k-nn-algoritmas.html

146. Kononenko, I., 1994. "Estimating attributes: Analysis and extensions of relief". *Proceedings of the Seventh European Conference on Machine Learning*, pp. 171-182.

147. Fayyad, U. M., Irani, K. B., 1993. "Multi-interval discretisation of continuous-valued attributes". *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pp. 1022-1027,

148. Aizerman, M. A., Braverman, E. M., Rozoner, L.I., 1964. "Theoretical foundations of thepotential function method in pattern recognition learning". **Automation and Remote Control, 25**:826-837.

149. Liu, H., Setiono, R., 1996. "A probabilistic approach to feature selection". *Proceedings of the 13th International Conference on Machine Learning*. pp. 319-327.

150. Witten I.H., Frank E., 2005. Data mining: practical machine learning tools and techniques – 2nd ed.. the United States of America, Morgan Kaufmann series in data management systems.

151. Sandeep V. Sabnani. 2008. Computer Security: A Machine Learning Approach, Technical Report, MSc in Information Security at Royal Holloway, University of London

152. Pumpuang P., Srivihok A. , Praneetpolgrang P., 2008. "Comparisons of Classifier Algorithms: Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students". *IEEE International Conference*.

153. Mohamed, W., Salleh M., Omar, H., 2012. "A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms", *IEEE International Conference on Control System*, Computing and Engineering, Penang, Malaysia

154. Leon, F., Zaharia, M. H., Galea, D., 2004. “Performance Analysis of Categorization Algorithms,” *International Symposium on Automatic Control and Computer Science*.

EKLER

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu

Kolon adı	Değişken Adı
Formno	Anket yapılan hane form numarası
Yerleşim yeri	
Fert sıra no	Hanehalkındaki 18 ve yukarı yaştaki anket yapılan kişilere verilen sıra numarası
Hane halkı büyüklüğü (HHB)	Hanehalkında yaşayan kişi sayısı
Cinsiyet	
YAŞ	18 ve yukarıdaki fert yaşı
B01	Medeni durum
B02	Eğitim durumu
B03	Son bir hafta içinde ücretli ya da ücretsiz olarak bir işte çalışma durumu
B04	Çalışmama nedeni
B05	Çalışılan iş/iş yerinin kamu, özel olma durumu
B06	Çalışılan işteki durum
B07	İşyerinin iktisadi faaliyet kodu
B08	Bir bütün olarak yaşamınızı düşünüdüğünüzde ne kadar mutlusunuz?
B09	Hayatta en çok kimin mutlu ettiği
B10	Hayatta en çok neyin mutlu ettiği
B11_1	Sağlıktan memnuniyet
B11_2	Evlilikten memnuniyet
B11_3	Şimdiye kadar alınan eğitimden memnuniyet
B11_4	Oturulan konuttan memnuniyet
B11_5	Oturulan semtten veya mahalleden memnuniyet
B11_6	(Eğer çalışıyorsa) İşten memnuniyet
B11_7	(Eğer çalışıyorsa) İşten elde ettiğiniz kazançtan memnuniyet
B11_8	Aylık hanehalkı gelirinden memnuniyet
B12_1	Akrabalarla ilişkilerden memnuniyet
B12_2	Arkadaşlarla ilişkilerden memnuniyet
B12_3	Komşularla ilişkilerden memnuniyet
B12_4	(Eğer çalışıyorsa) İşyerindeki kişilerle ilişkileren memnuniyet
B13_1	Sağlık hizmetlerinden memnuniyet
B13_2	Asayiş (güvenlik) hizmetlerinden memnuniyet
B13_3	Adli hizmetlerden memnuniyet
B13_4	Eğitim hizmetlerinden memnuniyet
B13_5	Sosyal Güvenlik Kurumu'nun hizmetlerinden memnuniyet

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu (devamı)

B13_6	Ulaştırma hizmetlerinden memnuniyet
B14	Elektronik ortamda sunulan kamu hizmetlerinden yararlanıyor musunuz?
B14_1	Elektronik ortamda sunulan kamu hizmetlerinden memnuniyet
	2012 yılı içinde, yapılan işe ilişkin sorunlar ÜCRETLİ/ MAAŞLI/YEVMİYELİ ÇALIŞANLARA SORULMUŞTUR
B15_1	İdari konularda
B15_2	Ücretler arası farklılık
B15_3	Ücret miktarında
B15_4	Çalışma koşullarında
B16	Yaşadığınız yerleşim yeri belediye sınırları içinde mi?
	Belediye hizmetlerinden memnuniyet durumu
B17_1	Çöp ve çevresel atık toplama hizmeti
B17_2	Kanalizasyon hizmeti
B17_3	Şebeke suyu hizmetleri
B17_4	Toplu taşıma hizmetleri
B17_5	Zabıta hizmetleri
B17_6	Yol/kaldırım yapımı hizmetleri
B17_7	Yeşil alanların miktarı
B17_8	Hava kirliliği ile mücadele
B17_9	Belediyenin sağlık, spor merkezi olanakları
B17_10	Belediyenin imar/iskan/ruhsat işlemleri
B17_11	Engellilere yönelik düzenlemeler
B17_12	Hasta ve yoksullara yardımları
B17_13	Sergi, festival, kermes,konser faaliyetleri
B17_14	Meslek edindirme/el becerisi geliştirme kursları
B17_15	Işıklandırma ve temizlik hizmetleri
B17_16	İtfaiye/cenaze hizmetleri
B17_17	Sokak levhaları ve dış kapı numaralandırma hizmeti
B17_18	Gıda üreten tesislerin denetimi
	İl özel idare hizmetlerinden memnuniyet durumu
B18_1	Kanalizasyon hizmeti
B18_2	Şebeke suyu hizmetleri
B18_3	Yol yapımı hizmetleri
B18_4	İmar/iskan/ruhsat işlemleri
B18_5	Engellilere yönelik düzenlemeler
B18_6	Hasta ve yoksullara yardımları
B18_7	Sergi, festival, kermes,konser faaliyetleri
B18_8	Meslek edindirme/el becerisi geliştirme kursları
B18_9	Işıklandırma ve temizlik hizmetleri

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu (devamı)

B18_10	Sokak levhaları ve dış kapı numaralandırma hizmeti
B19	Hangi sosyal güvenlik kurumundan yararlanıyorsunuz?
B20	Kimin üzerinden sosyal güvenlik kapsamındasınız?
B21_1	Bağlı bulunduğunuz sosyal güvenlik kuruluşu vasıtasıyla aldığınız sağlık hizmetlerinin kalitesinde sorun yaşıyor musunuz?
B21_2	Sağlık harcamalarının geri ödenmesi ile ilgili işlemlerde sorun yaşıyor musunuz?
B21_3	İlaç alımı ile ilgili işlemlerde sorun var mı?
B21_4	Sosyal güvenlik kuruluşunda çalışan personelin davranışlarında sorun görüyor musunuz?
B21_5	Emeklilik işlemlerinde sorun yaşıyor musunuz? (yalnızca emeklilere sorulmuştur)
B21_6	Emekli maaşlarının miktarı yeterli mi? (yalnızca emeklilere sorulmuştur)
B22	2012 yılı içinde bağlı olduğunuz Sosyal Güvenlik Kurumuna herhangi bir iş için gittiniz mi?
B23	Sosyal Güvenlik Kurumunda sorun yaşadınız mı?
B24	Hastalandığınızda, tedavi, ilaç vb. masraflarınızı genellikle hangi kanalla karşılıyorsunuz?
B25	Hastalandığınızda genellikle ilk nereye başvurursunuz?
B26	Bu sağlık kuruluşunu neden tercih ediyorsunuz?
B27_1	Muayene ve tahlil için randevu almakta sorun yaşıyor musunuz?
B27_2	Temizlik/hijyen konusunda sorun var mı?
B27_3	Yapılan muayeneden memnun musunuz?
B27_4	Doktorların hastalara davranışında sorun var mı?
B27_5	Hemşirelerin/hastabakıcıların hastalara davranışında sorun var mı?
B27_6	Doktor ve sağlık personeli sayısı yeterli mi?
B27_7	Muayene ve tahlil ücretlerini yüksek buluyormusunuz?
B27_8	İlaç fiyatlarında sorun görüyor musunuz?
B27_9	Muayene ve tahlil için sıra beklemede sorun var mı?
B27_10	Muayene için katkı payı ücreti ödemeyi sorun olrak görüyor musunuz?
B28	2012 yılı içinde sağlık hizmeti aldınız mı?
B29	En son sağlık hizmeti alımı sırasında herhangi bir sorun yaşadınız mı?
B30	2012 yılında en son sağlık hizmeti aldığınız sağlık kuruluşu hangisidir?
B31_1	Polis veya jandarma olaylara zamanında müdahale ediyor mu?
B31_2	Polis veya jandarmanın vatandaşa davranışından memnun musunuz?

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu (devamı)

B31_3	Polis veya jandarmanın verdiği trafik hizmetinden memnun musunuz?
B32_1	Kapkaç, yankesicilik vb hırsızlık olayı yaşadınız mı?
B32_11	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_12	Başvurmama nedeniniz nedir?
B32_2	Gasp olayı yaşadınız mı?
B32_21	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_22	Başvurmama nedeniniz nedir?
B32_3	Yaralanma, darp olayı yaşadınız mı?
B32_31	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_32	Başvurmama nedeniniz nedir?
B32_4	Aile fertlerinizin herhangi birinden kötü muamele gördünüz mü?
B32_41	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_42	Başvurmama nedeniniz nedir?
B32_5	Herhangi bir nedenden dolayı şantaj, tehdit olayı yaşadınız mı?
B32_51	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_52	Başvurmama nedeniniz nedir?
B32_6	Cinsel suçlardan dolayı mağduriyet yaşadınız mı?
B32_61	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_62	Başvurmama nedeniniz nedir?
B32_7	Dolandırıcılık sebebiyle herhangi bir mağduriyet yaşadınız mı?
B32_71	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_72	Başvurmama nedeniniz nedir?
B32_8	Bunların dışında herhangi bir suçtan dolayı mağduriyet yaşadınız mı?belirtiniz....
B32_81	Bu olay için Polise/Jandarmaya başvurdunuz mu?
B32_82	Başvurmama nedeniniz nedir?
B33_1	Mahkemelerdeki işlemlerde sorun var mı?
B33_2	Davaların karara bağlanma süresinde sorun var mı?
B33_3	Yasaların herkese adil ve tarafsız uygulanmasında sorun var mı?
B33_4	Avukatlık hizmetlerinin kalitesinde sorun var mı?
B34	2012 yılı içinde mahkemeye başvurunuz oldu mu?
B35	Mahkeme sürecinde herhangi bir sorun yaşadınız mı?
	Ulaştırma hizmetlerinden memnuniyet
B36_1	Karayolu ulaşımı
B36_2	Denizyolu ulaşımı
B36_3	Havayolu ulaşımı
B36_4	Demiryolu ulaşımı
B37	Evinizde yalnız otururken kendinizi ne kadar güvende hissediyorsunuz?

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu (devamı)

B38	Yaşadığınız çevrede, gece yalnız yürürken kendinizi ne kadar güvende hissediyorsunuz?
B39	Umut düzeyi
B40	Türkiye'de yaşayan insanların refah düzeyini "0" basamağı en düşük "10" basamağı en yüksek" düzey olarak düşündüğünüzde kendinizi hangi düzeyde görüyorsunuz?
B41	5 yıl öncesi ile karşılaştırdığınızda bugünkü durumunuzu (maddi veya manevi) nasıl görüyorsunuz?
B42	Gelecek 5 yıllık dönemi düşündüğünüzde, genel olarak durumunuzun nasıl olacağını tahmin ediyorsunuz?
B43_1	Genel olarak hayatınız sizce nasıl olacak?
B43_2	Kişisel iş durumunuz sizce nasıl olacak??
B43_3	Hanenizin mali durumu nasıl olacak?
B43_4	Türkiye'deki iş/çalışma durumu nasıl olacak?
B43_5	Türkiye'nin içinde bulunduğu ekonomik durum nasıl olacak?
B44	Türkiye'nin Avrupa Birliği'ne üye olmasının sizin yaşamınızı ne yönde etkileyeceğini düşünüyorsunuz?
B45	Türkiye'nin Avrupa Birliği'ne üye olması konusunda bir referandum (halk oylaması) yapılırsa, siz ne yönde oy kullanırsınız?
	Ülkemiz aşağıdaki konularda önümüzdeki 5 yılda nasıl değişecek?
B46_1	Ekonomik açıdan
B46_2	Sosyal haklar ve özgürlükler açısından
B46_3	Kamu hizmetlerinin sunumu açısından
B46_4	Devletin şeffaflığı açısından
B46_5	Dünyadaki saygınlığı açısından
	Son 1 yılda kişinin hayatında meydana gelen değişiklikler
B47_1	İşe girdim(iş arayanlar için)
B47_2	Yeni iş yeri açtım
B47_3	İşimi kaybettim
B47_4	İflas ettim/dükkan kapattım
B47_5	Gelirim azaldı
B47_6	Gelirim arttı
B47_7	Tasarruflarımda azalma oldu
B47_8	Tasarruf yapmaya başladım
B47_9	Daha ucuz ürünlerin tüketimine yöneldim
B47_10	Eğlence/tatil harcamalarımı kıstım
B47_11	Göç etmek zorunda kaldım
B47_12	Araba aldım
B47_13	Ev, arsa, yazlık vb. aldım

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu (devamı)

B47_14	Arabamı sattım
B47_15	Ev, arsa, yazlık vb. sattım
B47_16	Borçlandım
B47_17	Borçlarımı ödedim
	Toplumda itibarlı olmak için aşağıdakiler ne derece önemlidir?En önemli olan üçünü sıralayınız.
B49_1	Düzgün bir aile yaşamı
B49_2	Para
B49_3	Sosyal çevre
B49_4	Onurlu/şerefli/ahlaklı bir yaşam
B49_5	Eğitim
B49_6	Meslek/yapılan iş
B49_7	Diğer
	Aşağıdaki konularda çevrenizdeki diğer kişilerin durumları sizin için ne kadar önemlidir?
B50_1	Kılık kıyafetleri
B50_2	Aile yaşam biçimleri
B50_3	Ev ve kişisel eşyaları(cep tel, araba vb.)
B50_4	Arkadaş çevreleri
B50_5	Çocuklarının başarıları
B50_6	İnsanların eve giriş çıkış saatleri
B50_7	Yaptıkları işleri
B50_8	Gelir düzeyleri
B50_9	Dini inançları
B50_10	Siyasi düşünceleri
B50_11	Eğitim düzeyleri
	Aşağıdaki konularda çevrenizdeki kişilerin (akraba, arkadaş, komşu vb.) sizinle ilgili görüşleri sizin için ne kadar önemlidir?
B51_1	Kılık kıyafetiniz hakkındaki düşünceleri
B51_2	Aile yaşam biçiminiz hakkındaki düşünceleri
B51_3	Ev ve kişisel eşyalarınız hakkındaki düşünceleri (cep tel, araba vb.)
B51_4	Arkadaş çevreniz hakkındaki düşünceleri
B51_5	Çocuklarınızın başarısı hakkındaki düşünceleri
B51_6	Eve giriş çıkış saatiniz hakkındaki düşünceleri
B51_7	Yaptığınız işiniz hakkındaki düşünceleri
B51_8	Geliriniz hakkındaki düşünceleri
B51_9	Dini inançlarınız hakkındaki düşünceleri
B51_10	Siyasi düşünceleriniz hakkındaki düşünceleri
B51_11	Eğitim düzeyiniz

EK-1 : TÜİK Yaşam Memnuniyeti Anketi Fert değişkenler tablosu (devamı)

	Toplumsal baskı algısı
B52_1	Cinsiyetinden dolayı
B52_2	Medeni durumundan dolayı
B52_3	Yaşından dolayı
B52_4	Gelenek göreneklerinden dolayı
B52_5	Dini inanç ve davranışlarından dolayı
B52_6	Siyasi görüşden dolayı
B52_7	Memleketinden dolayı
B52_8	İşinden dolayı
B52_9	Kılık kıyafetinden dolayı
B52_10	İşsiz, çalışmıyor olmaktan dolayı
B52_11	Gelir düzeyinden dolayı
	Toplumsal konulara olan ilgi
B53_1	Siyaset
B53_2	Çevre, doğa sorunları(hava, su kirliliği vb.)
B53_3	Bilim ve teknoloji
B53_4	Kültür, sanat, edebiyat
B53_5	Sendika/Dernek faaliyetleri
B53_6	Ekonomi
B53_7	Spor
B53_8	Moda
B53_9	Din
B53_10	Sağlıkla ilgili konular
B53_11	Müzik
B54_1,B54_2,B54_3	Ülkenin en önemli 3 sorunu
B54_D	Diğer seçeneğinin açıklaması

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı, Soyadı:	Serkan DEMİRCAN
Uyruğu:	Türkiye (T.C.)
Doğum Tarihi ve Yeri:	3 Kasım 1990, Gaziantep
Medeni Durumu:	Bekar
Tel:	+90 534 338 39 83
E-mail:	serkan_27gs@hotmail.com

EĞİTİM

Derece	**Kurum**	**Mezuniyet Tarihi**
Lisans	Erciyes Üni., M.F., Endüstri Müh.	2013
Lise	Ayten Kemal Akınal Anadolu Lisesi	2008

SERTİFİKALAR

C sınıfı İş Güvenliği Uzmanlığı

YABANCI DİL

İngilizce