SOSYAL BİLİMLER ENSTİTÜSÜ

T.C.
OSMANİYE KORKUT ATA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI

İNSANİ GELİŞMİŞLİK ENDEKSİNİN SINIFLANDIRMA BAŞARILARININ KARŞILAŞTIRILMASINDA KARAR AĞACI YÖNTEMLERİNİN KULLANILMASI

YÜKSEK LİSANS TEZİ

Ayşe YILDIZ

OSMANİYE – 2015

T.C.
OSMANİYE KORKUT ATA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YÖNETİM BİLİŞİM SİSTEMLERİ ANABİLİM DALI

İNSANİ GELİŞMİŞLİK ENDEKSİNİN SINIFLANDIRMA BAŞARILARININ KARŞILAŞTIRILMASINDA KARAR AĞACI YÖNTEMLERİNİN KULLANILMASI

YÜKSEK LİSANS TEZİ

DANIŞMAN

Prof. Dr. Murat TÜRK

Ayşe YILDIZ

OSMANİYE – 2015

TEZ ONAYI

Osmaniye Korkut Ata Üniversitesi Sosyal Bilimler Enstitüsü Müdürlüğü'ne;
Osmaniye Korkut Ata Üniversitesi Sosyal Bilimler Enstitüsü "Yönetim Bilişim Sistemleri" Ana Bilim Dalı "13YBSYL1104" nolu öğrencisi "Ayşe YILDIZ" tarafından "Prof. Dr. Murat TÜRK" danışmanlığında hazırlanan "İnsani Gelişme Endeksi'nin Sınıflandırma Başarılarının Karşılaştırılmasında Karar Ağacı Yöntemlerinin Kullanılması" başlıklı bu çalışma aşağıda imzaları bulunan jüri üyeleri tarafından oy birliği/çokluğu ile Yüksek Lisans Tezi olarak kabul edilmiştir.

Başkan: Prof. Dr. Murat TÜRK ……..

Üye: Doç. Dr. H. Ali ATA ……..

Üye: Yrd. Doç. Dr. Emre YAKUT ……..

Yukarıdaki Jüri kararı Osmaniye Korkut Ata Üniversitesi Sosyal Bilimler Enstitüsü Yönetim Kurulu'nun 17/06/2015 tarih ve ..... /……..... sayılı kararı ile onaylanmıştır.

Prof. Dr. Mustafa TANÇ
Sosyal Bilimler Enstitüsü Müdürü

Bu tezde kullanılan özgün bilgiler, şekil, çizelge ve fotoğraflardan kaynak göstermeden alıntı yapmak 5846 sayılı Fikir ve Sanat Eserleri Kanunu hükümlerine tabidir.

T.C.

OSMANİYE KORKUT ATA ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

YEMİN METNİ

Yüksek Lisans tezi olarak sunduğum "İnsani Gelişme Endeksi'nin Sınıflandırma Başarılarının Karşılaştırılmasında Karar Ağacı Yöntemlerinin Kullanılması" adlı çalışmanın, tezin proje safhasından sonuçlanmasına kadarki bütün süreçlerde bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurulmaksızın yazıldığını ve yararlandığım eserlerin Bibliyografya'da gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve onurumla beyan ederim.

Ayşe YILDIZ

17.06.2015

ÖZET

İNSANİ GELİŞMİŞLİK ENDEKSİNİN SINIFLANDIRMA BAŞARILARININ KARŞILAŞTIRILMASINDA KARAR AĞACI YÖNTEMLERİNİN KULLANILMASI

Ayşe YILDIZ

Osmaniye Korkut Ata Üniversitesi Sosyal Bilimler Enstitüsü
Yüksek Lisans Tezi, Yönetim Bilişim Sistemleri Anabilim Dalı
Danışman: Prof. Dr. Murat TÜRK
Haziran 2015, 59 sayfa

Günümüzde bilişim alanındaki hızlı gelişmeler sonucunda yapılan her işlem bilgisayarlara kaydedilmektedir. Bu kaydedilen veriler dev veri tabanlarını oluşturmaktadır. Kurumlar için rekabet avantajı sağlayacak olan önemli bilgiler bu veri yığınları içerisinde kaybolmaktadır. Geleneksel istatistiksel yöntemler ile bu büyük boyuttaki verilerin çözümlenmesi mümkün olmadığı için alternatif olarak veri madenciliği ortaya çıkmıştır.

Bu çalışmanın amacı UNDP (United Nations Development Programme)'nin her yıl yayınlamış olduğu İnsani gelişme raporunda bulunan ülkelerin gelişmişlik seviyelerini ortaya koyan İnsani gelişme endeksinin veri madenciliği tekniklerinden C5.0, CHAID ve C&RT algoritmaları kullanılarak sınıflandırılması ve sınıflandırma başarılarının karşılaştırılarak en iyi tekniğin seçilmesidir.

Analiz sonucunda C5.0 algoritmasının diğer algoritmalara göre daha yüksek bir sınıflandırma başarısı sağladığı ortaya konulmuştur.

Anahtar Kelimeler: Veri Madenciliği, Karar Ağaçları, C5.0

ABSTRACT

USING THE METHODS OF DECISION TREE IN COMPARASION OF CLASSIFICATION SUCCESSES OF HUMAN DEVELOPMENT INDEX

Ayşe YILDIZ

Osmaniye Korkut Ata University, Institute of Social Sciences
Department of Management Information Systems
Supervisor: Prof. Dr. Murat TÜRK
June 2015, 59 pages

Today, all transactions as a result of rapid developments in the field of informatics are recorded in computers. These recorded data constitute giant databases. Important information that will allow for competitive advantage for organizations is lost in these data stacks. Because it is not possible for traditional statistical methods to analyze large size data, data mining has emerged as an alternative.

The purpose of this resarch is to classify Human Development Index which reveals development levels of countries in human development report published by UNDP annually using C 5.0, CHAID, C&RT algorithms, three of data mining techniques, and to select the best technique by comparing classification successes.

Key Words: Data Mining, Decision Trees, C5.0

İÇİNDEKİLER

ÜÇÜNCÜ BÖLÜM

UYGULAMA

TABLOLAR

ŞEKİLLER

GRAFİKLER

KISALTMALAR

DVM : Destek Vektör Makineleri

EYS : Expected Years of Schooling

GNI : Gross National Income

GNI HDI : GNI Per Capita Rank Minus HDI Rank

GSMG : Gayri Safi Milli Gelir

HDI : Human Development Index

İGE : İnsani Gelişme Endeksi

LEB : Life Expectancy at Birth

LSD : En Küçük Kareli Sapma Yöntemi

s. : Sayfa

UIS : İstatistik Enstitüsü

UNDP : United Nations Development Programme

UNESCO : United Nations Educational, Scientific and Cultural Organization

vb. : Ve Benzeri

VM : Veri Madenciliği

WB : Dünya Bankası

YSA : Yapay Sinir Ağları

ÖNSÖZ

Çalışmalarım boyunca yardım ve katkılarıyla beni yönlendiren Danışman Hocam Prof. Dr. Murat TÜRK'e ve bu çalışmanın gerçekleştirilmesinde, her konuda yardımlarını benden esirgemeyen, fikirleri ve desteği ile yanımda olan Yrd. Doç. Dr. Emre YAKUT'a, bana katkılarından dolayı Doç. Dr. M. Fedai ÇAVUŞ'a, Doç. Dr. Bülent ÖZ'e, Yrd. Doç. Dr. Esengül İPLİK'e ve Doç. Dr. Alper AYTEKİN'e teşekkür ederim. Ayrıca çalışmam boyunca beni destekleyen Öğr. Gör. Abdurrahman AKMAN'a, Nebi KARAKELLE'ye, Jayant KUMAR'a, Dr. Hakkı Seçkin ÇETİN'e, çevirilerimde bana yardımcı olan Uzm. Burak AYÇİÇEK'e ve tecrübelerinden faydalandığım
Arş. Gör.Volkan Soner ÖZSOY'a teşekkürü bir borç bilirim.

Attığım her adımda manevi destekleriyle, sevgi ve sabırlarıyla beni hiçbir zaman yalnız bırakmayan çok değerli aileme sonsuz teşekkür ederim.

Ayşe YILDIZ

Osmaniye, 2015

Anneme ve Babama…

BİRİNCİ BÖLÜM

GİRİŞ

Teknolojinin gelişmesiyle beraber artan rekabet şartlarında bilginin önemi artmıştır. Bir gün boyunca yaptığımız her türlü hareket birden fazla veri tabanında veri olarak girilmektedir. Telefonla görüşme süremiz, ziyaret ettiğimiz internet siteleri, bu sitelerdeki ziyaret süremiz, satın alma davranışlarımız, e-devlet ve bankacılık işlemlerimiz vb. her türlü çevrimiçi işlemler veya market alışverişlerimiz trafikte ki mobese kayıtlarımız gibi çevrimdışı işlemlerin hepsi veri tabanlarına kaydedilmektedir. Bütün bu kayıtlar milyarlarca veriyi içermektedir ve bu veri yığınları arasında faydalı olabilecek önceden tahmini mümkün olmayan bilgiler kaybolmaktadır. Bu noktada veri madenciliğine ihtiyaç duyulmuştur. VM, dev veri yığınları içerisinde gizli kalmış olan bilgilerin, örüntülerin ortaya çıkarılarak anlamlı hale getirilmesi işlemidir.

VM; dolandırıcılık tespiti, gereksiz cerrahi müdahalelerinin azaltılmasını, mevcut müşterilerin satın ama davranışlarının belirlenerek satış politikalarının düzenlenmesini, hastalıklı genlerin tespit edilerek erken müdahale edilmesini mümkün kılar.

Bu çalışmanın amacı, İnsani Gelişme Endeksi'nin sınıflandırılmasında veri madenciliği algoritmalarından C5.0, C&RT, CHAID kullanılarak sınıflandırma başarısının ortaya konulmasıdır. 3 algoritmanın sonuçları karşılaştırılarak en yüksek sınıflandırma başarısını veren algoritma tespit edilmiştir. UNDP tarafından her yıl düzenli olarak yayınlanan raporlardan 2011-2012-2013 yılları olmak üzere 3 yıla ait veriler kullanılmıştır. Algoritmalar veri setlerine uygulanarak performansları karşılaştırılmıştır.

Çalışma 3 bölümden oluşmaktadır. Birinci bölümde, VM tanımlanmış ve kullanım alanları açıklanmıştır. İkinci bölümde VM'nin yöntemleri ve algoritmaları hakkında bilgi verilmiştir. Üçüncü bölümde, İnsani Gelişme Endeksi'nin genel tanımlamaları yapılarak hesaplama yönteminden bahsedilmiştir. C5.0, CHAID ve C&RT algoritmaları kullanılarak analiz gerçekleştirilmiştir. 3 yöntemin sonucu karşılaştırılmış ve elde edilen sonuçlar tablo ve grafiklerle desteklenmiştir. Tüm bu araştırmaların ve analizin yorumlarına sonuç bölümünde yer verilmiştir.

İKİNCİ BÖLÜM

VERİ MADENCİLİĞİ BİLİMİ

2.1. Veri Madenciliği Nedir?

Veri Madenciliği; Verilerin bilgisayar ortamında saklanmasıyla birlikte sürekli artan veri yığınlarının depolanması için kullanılan veri tabanları da veri yığınlarına bağlı olarak genişlemiştir. Bunun sonucunda önemli olan bilgiler bu yığınların arasında kaybolmuştur. Veri madenciliği veri tabanlarında gizli kalmış olan önemli, kıymetli bilgileri, ilişkileri ve örüntüleri ortaya çıkararak bu bilgilerden kar veya fayda elde etmeyi amaçlar. Literatürde veri madenciliğinin birçok tanımı vardır. Bunlardan bazıları;

- Shalvi ve DeClaris (1998), veri madenciliğini, belirli bir alanda ve belirli bir amaç için toplanan veriler arasındaki gizli kalmış ilişkilerin(desenlerin, modellerin vb.) ortaya konulması olarak tanımlamışlardır (Acar Şaylan, 2013, s.3).
- Hand (1998), veri madenciliği; istatistik, veri tabanı teknolojisi, örüntü tanıma, makine öğrenme ile etkileşimli yeni bir disiplin ve geniş veritabanlarında önceden tahmin edilemeyen ilişkilerin ikincil analizi olarak tanımlanmıştır (Akbulut, 2006, s.3).
- David (1999), veri madenciliğinin büyük hacimli verilerdeki örüntüleri araştıran matematiksel algoritmaları kullandığını belirtmiştir. Ayrıca VM hipotezleri keşfederek, sonuçları birleştirmek için insan yeteneğini kullanır. Veri madenciliğinin bir bilimden fazlası olup, aynı zamanda bir sanat olduğunu söylemiştir (Akbulut, 2006, s.3).
- Veri madenciliği geniş veri tabanlarındaki birliktelikleri ortaya çıkarır (Akbulut, 2006, s.4).
- Veri madenciliği, önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veri tabanlarından elde edilmesi ve bu bilgilerin işletme kararları verilirken kullanılmasıdır (Silahtaroğlu, 2013, ss.7-9).
- Veri madenciliği, kurumlarda biriken veri içerisinde kurum için yararlı olanlarını bulup ortaya çıkarma ve ölçekli veriler arasından "değeri olan" bir bilgiyi elde etme işidir (Özkan, 2013, s.37).
- Veri madenciliği, büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılması tekniğidir (Carus ve Mesut, 2005, s.121).

- Veri madenciliği, depolanmış yüksek miktardaki veriden istatistiksel ve matematiksel teknikler gibi desen tanımlayıcı teknolojiler kullanarak anlamlı ve yeni ilişkiler, desenler ve trendler keşfetme sürecidir (Acar Şaylan, 2013, s.3).

Bu konuda Atılgan (2011, s.11) şu tanımları vermiştir.

- Veri madenciliği, verideki geçerli, alışılmışın dışında, kullanışlı ve anlaşılabilir örüntülerin (pattern) belirlenmesi sürecidir. - Fayyad.
- Veri madenciliği, geniş veri tabanlarından daha önceden bilinmeyen, anlaşılabilir ve kullanılabilir bilgi çıkarsama ve bu bilgiyi kritik karar almak için kullanma sürecidir. - Zekulin.
- Veri madenciliği, bilgi keşfi sürecinde veriden daha önceden bilinmeyen ilişkileri ve yapıları ayırt etmek için kullanılan yöntemlerdir. - Ferruza.
- Veri madenciliği, geniş veri tabanlarında bilinmeyen ve beklenmeyen bilgi örüntülerini araştıran karar destek sürecidir (Atılgan, 2011, s.11).

Yapılmış olan tanımların birbirine çok benzediği görülmektedir. VM konusunda tüm tanımların ortak ifadesi; "anlamlı ve değerli bilginin ortaya çıkarılması", "gizli örüntülerin ve ilişkilerin belirlenmesi", "daha önceden bilinmeyenin keşfedilmesi" sürecidir.

2.2. Veri Madenciliğinin Tarihçesi

1950'li yıllarda matematikçiler veri madenciliği teknikleri üzerine çalışarak mantık ve bilgisayar bilimleri alanlarında yapay zeka ve makine öğrenme alanlarını ortaya çıkarmışlardır. 1960'lı yıllarda istatistikçiler yeni bir algoritma keşfetmişlerdir. Veri madenciliğinin ilk adımlarını oluşturan bu algoritmalar regresyon analizi ve en büyük olabilirlik kestirimidir. Daha sonraki 20 yıllık süreçte önce verilerin sınıflara ayrılması ardından bu sınıflar arasında ilişkisel bağlantıların kurulması ile veri tabanı kavramı ortaya çıkarılmıştır. 1990'lı yıllara gelindiğinde ise veri tabanında bilgi keşfinin ilk adımları oluşturulmuş ve bununla birlikte büyük veri tabanları için veri ambarı geliştirilmiş ve aynı zamanlarda yeni teknolojilerle beraber veri madenciliği yaygın olarak kullanılmaya başlanmıştır. Veri madenciliğinin tarihsel gelişim kronolojisi Tablo 2.1 de verilmiştir (Acar Şaylan, 2013, s.3)

Tablo 2.1: Veri Madenciliğinin Tarihsel Gelişimi

Tarih	**Basamaklar**	**Sorular**	**Kullanılabilir Teknolojiler**	**İlgili Yazılımlar**
1960'lar	Veri toplama, Veritabanı Yönetim Sistemleri	Benim son 5 yıldaki toplam kârım nedir?	Bilgisayar, Disk, Düz dosyalar	Fortran
1980'ler	Veriye ulaşım, Veri sorgulama	Geçen Mart İstanbul'daki birim satış miktarı nedir?	Daha hızlı ve ucuz bilgisayarlar, daha fazla depolama alanı, ilişkisel veritabanları	Oracle, IBM, DB, SQL
1990'lar	Veri ambarları, Karar destek sistemleri	Geçen Mart İstanbul'daki birim satış miktarı nedir? Ankara ile karşılaştırmalı olarak görmek istiyorum.	Daha hızlı ve ucuz bilgisayarlar, Daha fazla depolama alanı, İlişkisel veritabanları, OLAP, Çok boyutlu veritabanları, Veri ambarları	SQL Standart, Veri Ambarları, OLAP, Darwin, IBM Intelligent Miner, SPSS Crisp DM, SAS Miner, Angoss Knowledge
1990'larınsonu 2000'ler	Veri Madenciliği Web Madenciliği	Ankara'da gelecek ayki birim satışlarım ne durumda olacak ? Neden?	Daha hızlı ve ucuz bilgisayarlar, Daha fazla depolama alanı, İlişkisel veritabanları, Gelişmiş bilgisayar algoritmaları	Oracle Data Miner, IBM DB2 UDB Mining, SPSS Clementine, SAS Enterprise Miner

Kaynak: (Acar Şaylan, 2013, s.4).

2.3. Veri Madenciliği Uygulama Alanları

Veri yığınları arasında değerli ve kullanılabilir bilgileri keşfeden veri madenciliği, verinin üretildiği büyük veri ambarlarına sahip her ortamda uygulama alanı bulmuştur.

Bunlar;

Satış ve Pazarlama alanında veri madenciliği uygulamaları:

- Müşteri sınıflandırma, hedef müşteri belirleme,
- Müşterilerin satın alma sıklıkları ve satın alma örüntülerinin belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki ilişkilerin saptanması,
- Mevcut müşterilerin elde tutulması, yeni müşteri edinilmesi,
- Pazar sepeti analizi,
- Müşteri ilişkileri yönetimi,
- Çapraz satış analizleri, müşteri değerleme,
- Satış tahminleri,
- Posta kampanyalarına cevap verme oranının artırılması,

Bankacılık alanında veri madenciliği uygulamaları:

- Finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı harcamalarına göre müşterilerin gruplandırılması,
- Kredi kartı dolandırıcılığının tespiti,
- Hisse senetlerinin değer değişim tahminleri,
- Kredi taleplerinin değerlendirilmesi,
- Risk analizleri, risk yönetimi,
- Sigortacılık,

Biyoloji, tıp ve genetik alanında veri madenciliği uygulamaları:

- Kanser Tespiti,
- Down sendromu tespiti,
- Tıbbi teşhis, hastalık tespiti,
- Test sonuçlarının tespiti,
- Gereksiz biyopsi ve MR çekimlerinin önlenmesi,
- Bitki türlerinin ıslahı,
- DNA sıra analizi,
- Yem ve ilaç türlerinin keşfi,
- Gen haritasının analizi,

Güvenlik alanında veri madenciliği uygulamaları:

- Ses ve yüz tanıma teknolojisi,
- Dolandırıcılık tespiti,

Perakendecilik alanında veri madenciliği uygulamaları:

- Alışveriş sepeti analizleri,
- Tedarik zinciri yönetimi uygulamaları,
- Mağaza optimizasyonu başlıcalarıdır.

2.4. Veri Madenciliği Süreci

Bir veri tabanına veri madenciliği analizinin uygulanması için veriler belirli süreçlerden geçirilerek analize hazırlanmalıdır. Veri tabanı bazı kayıtlar yönünden eksik veriler veya aşırı uç değerler içerebilir. Böyle kayıtlar analizin doğruluğunu tehdit edebilir. Bu gibi veri sorunları için analiz öncesi veriler analize hazırlanmalıdır. Doğru bir analiz için izlenmesi gereken adımlar:

- Problemin tanımlanması,
- Verilerin hazırlanması,
- Modelleme,
- Değerlendirme,
- Sonuç.

2.4.1. Problemin Tanımlanması

Problemin tanımlanması için öncelikle analizin hangi amaçla yapıldığı belirlenmelidir. Problem belirlenirken, analist problemin ilgili olduğu alanda uzman olmayabilir. Bu durumda alanında bir uzmandan destek alınarak problem, uzman dışındakilerin de anlayabileceği bir hale getirilir. Problemin net belirlenmemesi, Kısıtlarının tanımlanmaması durumunda analizin sürekli bu aşamaya geri dönmesi gerekebilir. Bu da maliyet ve zaman kaybına neden olur. Analizin başarılı olması için problem net olarak belirlenmelidir.

2.4.2. Verilerin Hazırlanması

Verilerin hazırlanması, diğer bir deyişle veri önişleme, veri madenciliği süreçlerinin en uzun zaman gerektirenidir. Büyük çaplı veri yığınları, genellikle tutarsızlıklar, problemler ya da ilgili oldukları konuya özel değerler içerdikleri için kullanımı mümkün

olmayan ham verilerde içerebilir. Bu safhada, operasyonel işlemler sonrasında elde edilen veriler, veri madenciliği uygulamalarına uygun hale gelmeleri amacıyla bir takım işlemlere tabi tutulurlar. Bu işlemler verinin kalitesini arttırırken veri madenciliği uygulamasının da veri üzerinde daha rahat işlem yapılabilmesini sağlar (Yakut, 2012, s.8).

Bu işlemler şu başlıklar altında incelenebilir;

- Veri temizleme
- Veri normalleştirme
- Veri indirgeme
- Veri dönüşümü

2.4.2.1. Veri Temizleme

Veri temizleme aşamasında seçilen analiz teknikleri için gerekli verinin kalitesi arttırılmaya çalışılmaktadır (Yakut, 2012, s.9).
Bazı uygulamalarda, üzerinde çözümleme yapılacak olan verilerin istenen özelliklere sahip olmadığı görülebilir. Örneğin eksik verilerle ve uygun olmayan verilerin oluşturduğu tutarsız verilerle karşılaşılabilir. Veri tabanında yer alan tutarsız ve hatalı verilere gürültü denir. Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekebilir (Özkan, 2013).

2.4.2.2. Veri Normalleştirme

Bazı durumlarda veriyi direk analize tabi tutmak uygun olmayabilir. Değişkenlerin sahip olduğu çok büyük ve çok küçük değerler analizin doğruluğunu azaltır. Bu nedenle değişkenler normalleştirilmelidir. Bu amaçla kullanılan teknikler;

2.4.2.2.1. Min-max Normalleştirmesi

Min-maks yönteminde min, en küçük değeri max ise, en büyük değeri tanımlar. Verilerin 0.0 - 1.0 aralığına indirgenilmesi amaçlanır. Bunun için aşağıdaki formül kullanılır (http://ilkucar.com).

$$Xnormal = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

X normal= Dönüştürülmüş değer

x = Gözlem değeri

x min = En küçük gözlem değeri

x max = En büyük gözlem değerini ifade eder.

2.4.2.2.2. Z – Skor Standartlaştırma

Sıkça kullanılan bir dönüşüm biçimidir. Bu yöntem, verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esas alınır ve şu bağıntı kullanılır (https://www.academia.edu).

$$X^* = \frac{x-\bar{x}}{\sigma_x} \qquad (2.2)$$

X^*: dönüştürülmüş değer

x : gözlem değerleri

$\bar{x}$: verilerin aritmetik ortalaması

σ_x: gözlem değerlerinin sapmasını ifade eder.

2.4.2.3. Veri İndirgeme

Veri indirgeme teknikleri, daha küçük hacimli olarak ve veri kümesinin indirgenmiş bir örneğinin elde edilmesi amacıyla kullanılır. Bu sayede elde edilen indirgenmiş veri kümesine veri madenciliği teknikleri uygulanarak daha etkin sonuçlar elde edilebilir. Veri indirgeme yöntemleri aşağıdaki biçimde özetlenebilir (Oğuzlar, 2003):

1. Veri Birleştirme veya Veri Küpü
2. Boyut indirgeme
3. Veri Sıkıştırma
4. Kesikli hale getirme

2.4.2.4. Veri Dönüşümü

Veriler; farklı kaynaklar ve veritabanlarından alınması nedeniyle farklı dosya uzantısı ve yapılarına sahiptirler. Bu yapı ve uzantıları dönüştürmek elzemdir. Verilerin kullanılacak algoritmaya uygunluğu sürecinde bu dönüştürme işleminden istifade edilir.

2.4.3. Modelleme

Bu aşamada daha önce hazırlanmış olan veriler kullanılarak modelleme gerçekleştirilir. Modelleme şu adımları içerir (Hand, Mannila ve Smyth, 2011):

- Veriye uygun modelin seçimi,
- Modellere ilişkin alternatiflerin değerlendirilmesi,
- Modele uygun algoritmaların ve hesaplama metotlarının belirlenmesi.

2.4.4. Değerlendirme

Modelin yayılma aşamasına geçmeden önce analizin amaçlarını tam olarak gerçekleştirdiğinden emin olmak için modelin eksiksiz bir şekilde değerlendirilmesi ve modeli gerçekleştirmek için oluşturulan adımların gözden geçirilmesi önemli bir adımdır. Temel amaç, Problemde yeteri derecede dikkate alınmayan bir sorununun olup olmadığını belirlemektir. Bu evrenin sonunda veri madenciliği sonuçlarının kullanımıyla ilgili bir karara ulaşılır (Küçüksille, 2009, s.34).

2.4.5. Sonuç

Bu aşamada veri madenciliği yöntemleri ile elde edilen sonuçların düzenlenmesi, yorumlanması ve sunuma hazır hale getirilmesi işlemleri gerçekleştirilir. Sunumda anlaşılabilirliği arttırmak amacıyla görselleştirme tekniklerinden faydalanılabilir (Özkan, 2013).

2.5. VERİ MADENCİLİĞİ YÖNTEMLERİ

Veri madenciliği konusunda birçok yöntem ve algoritma geliştirilmiştir. Temel olarak veri madenciliği modelleri üç ayrı modeldir:

- Sınıflandırma,
- Kümeleme,
- Birliktelik kuralı,
- Destek vektör makineleri.

2.5.1. Sınıflandırma

Sınıflandırma veri madenciliğinde en yaygın kullanılan yöntemdir. Veri tabanlarındaki gizli kalmış bilgileri ve önemli örüntüleri farklı algoritmalar kullanarak ortaya çıkarır.

Verilerin sınıflandırılması için belirli bir süreç izlenir. Veritabanının bir kısmı eğitim seti olarak ayrılır ve sınıflandırma kuralları türetilir. Daha sonra bu kurallar yardımıyla yeni veya benzer bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Çıktılar, önceden bilindiği için sınıflama, veri kümesini denetimli (supervised) olarak öğrenir (Özkan, 2013).

Sınıflandırma tekniklerinde yaygın olarak kullanılan modeller:

- Karar Ağaçları (Decision Trees)
- Yapay Sinir Ağları (Artificial Neural Networks)
- Genetik Algoritmalar (Genetic Algorithms)
- K-En Yakın Komşu (K-Nearest Neighbor)
- Naive-Bayes
- Lojistik Regresyon Analizidir (Giudici, 2003, ss.4-6).

2.5.1.1. Karar Ağaçları

Karar ağaçları, veri madenciliğinde kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde yaygın kullanıma sahiptir (Akpınar, 2000, ss.1-22).

Karar ağaçları 3 kısımdan oluşur. Düğüm denilen kök, düğümlere bağlanan dallar ve dallara bağlı olan yapraklar.

Karar düğümü, gerçekleştirilecek testi belirtir. Bu testin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olur. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleştirilir ve bu ayrılma işlemi üst seviyedeki ayrımlara bağımlıdır. Ağacın her bir dalı sınıflama işlemini tamamlamaya adaydır. Eğer bir dalın ucunda sınıflama işlemi gerçekleşmiyorsa, orada bir karar düğümü oluşur. Ancak belirli bir sınıf oluşuyorsa, o dalın sonunda yaprak vardır. Bu yaprak, veri üzerinde belirlenmek istenen sınıflardan biridir. Karar ağacı işlemi kök düğümünden başlar ve yukarıdan aşağıya doğru yaprağa ulaşana dek ardışık düğümleri takip ederek gerçekleşir (Özekes, 2003, ss.65-81).

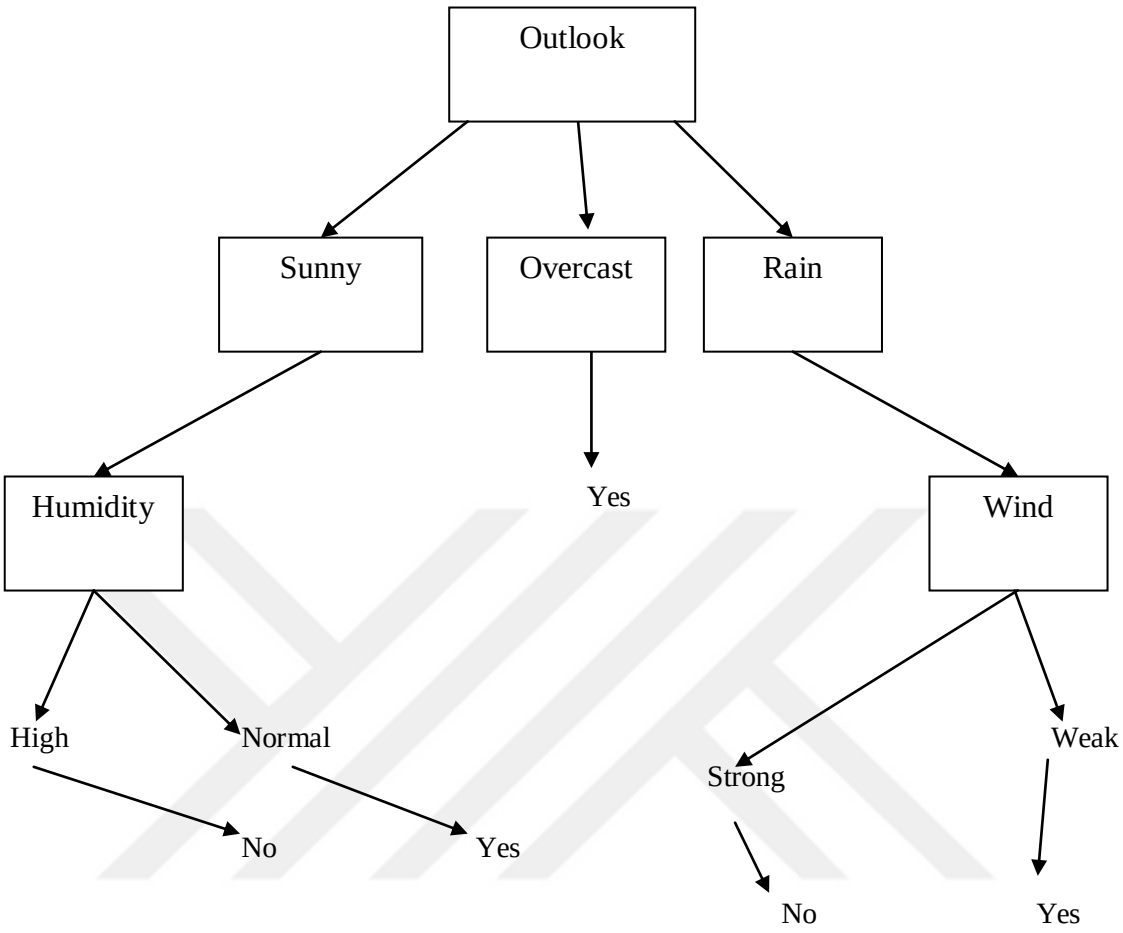
Başlangıçta bütün öğrenme örnekleri kök düğümde yer alır, örnekler seçilmiş özelliklere göre tekrarlamalı olarak bölündükten sonra ağacı temizlemek için (tree pruning) gürültü ve istisna kararları içeren dallar belirlenir ve budama işlemi gerçekleştirilir.

Karar ağacı tekniğini ile verinin sınıflandırılması üç aşamadan oluşur:

- Öğrenme: Önceden sonuçları bilinen verilerden (eğitim verisi) model oluşturulur.
- Sınıflama: Yeni bir veri seti (test verisi) modele uygulanır, bu şekilde karar ağacının doğruluğu belirlenir. Test verisine uygulanan bir modelin doğruluğu, yaptığı doğru sınıflamanın test verisindeki tüm sınıflara oranıdır. Her test örneğinde bilinen sınıf, model tarafından tahmin edilen sınıf ile karşılaştırılır.
- Uygulama: Eğer doğruluk kabul edilebilir oranda ise, karar ağacı yeni verilerin sınıflanması amacıyla kullanılır (Argüden ve Erşahin, 2008).

Karar ağacı algoritmasının yaygın olarak kullanıldığı sahalar (Telcioğlu, 2007),

- Belirli bir sınıfın muhtemel üyesi olacak elemanların belirlenmesi,
- Çeşitli vakaların yüksek, orta, düşük risk grupları gibi çeşitli kategorilere ayrılması,
- Gelecekteki olayların tahmin edilebilmesi için kuralların oluşturulması,
- Parametrik modellerin kurulmasında kullanılmak üzere çok miktarda değişken içeren veri kümesinden faydalı olacakların seçilmesi,
- Sadece belirli alt gruplara özgü olan ilişkilerin tanımlanması,
- Kategorilerin birleştirilmesi ve sürekli değişkenlerin kesikli hale dönüştürülmesidir.



Şekil 2.1: Basit Bir Karar Ağacı Örneği

Kaynak: www.cs.princeton.edu.

Bu örnekte tipik öğrenmeli bir karar ağacı verilmiştir. Ağaç havanın tenis oynamak için uygun olup olmadığının belirlenmesi için kural üretmektedir. Ağaçtan kuralları çıkarmak için en üstteki düğüm olarak adlandırılan kısımdan başlayarak aşağı doğru inilir. "İf-Then" yapısı diye ifade edilen "Eğer" ifadeleri kullanarak kurallar şu şekilde çıkarılır:

Kural 1: Eğer hava bulutlu ise tenis oynanır.

Kural 2: Eğer hava güneşliyse neme bakılır, eğer nem normal ise tenis oynanır.

Kural 3: Eğer hava güneşli nem yüksekse tenis oynanmaz.

Kural 4:Eğer hava yağmurluysa rüzgara bakılır, eğer rüzgar güçlü ise tenis oynanmaz, zayıf ise oynanır.

Kurallardan görüldüğü üzere çıkacak kuralın sayısı kural çıkarıcının yorumlamasına bağlı olarak değişebilmektedir. Karar ağacının yaygın kullanılma sebeplerinden biride ağacın bir uzmanlık gerektirmeden kolaylıkla okunup anlamlandırılabilmesidir.

En yaygın kullanılan karar ağacı algoritmaları (Taşdemir, 2012, s.49);

• ID3 (Quinlan, 1986),

• C4.5 (Quinlan, 1993),

• CHAID (Kass 1980),

• C&RT (Breiman ve Friedman, 1984).

2.5.1.1.1. ID3 Algoritması

Quinlan tarafından karar ağaçlarının sınıflandırma yapması için geliştirilmiş en basit algoritmadır. Sayısal niteliklerle ve eksik verilerle çalışmaz. Entropiye dayalı bir algoritmadır. Karar Ağaçlarının oluşturulması esnasında dallanmaya başka bir ifadeyle bölümlemeye hangi nitelikten başlanacağı önem taşımaktadır. Çünkü sınırlı sayıda kayıttan oluşan bir eğitim kümesinden yararlanarak olası tüm ağaç yapılarını ortaya çıkarmak ve içlerinden en uygun olanı seçerek ondan başlamak kolay değildir (Kantardzic, 2003, s.125).

Karar ağacı algoritmaları başlangıç aşamasında bazı değerleri hesaplayarak bu değerler yardımıyla ağaç oluştururlar. Ağacın dallanması entropinin alacağı değere göre değişiklik gösterir.

Entropi, Bir sistemdeki belirsizliğin ölçüsüne "Entropi" denir (Özkan, 2013). Dunham entropinin bir veri kümesi içindeki belirsizlik, şaşkınlık ve rastgeleliği ölçmek için kullanıldığını söyler.

Eldeki bütün veriler tek bir sınıfa ait olsaydı, örneğin herkes aynı futbol takımını tutsaydı, bir kişiye tuttuğu takımı sorduğumuz zaman alacağımız cevap bizi şaşırtmayacaktı; bu durumda entropi sıfır(0) olacaktı. Entropi 0-1 arasında bir değer alır. Bütün olasılıklar eşit olduğunda entropi maksimum değerine ulaşır (Silahtaroğlu, 2013). Matematiksel olarak formüle edilirse,

$$H= -\sum_{i=1}^{n} P_i \log_2 pi \quad \text{(Okafor, 2005,s.5)} \qquad (2.3)$$

Örnek (Özkan, 2013,s.57):

Risk kümesi on elemanlı bir kümedir,

Risk={var,var,var,yok,var,yok,yok,var,var,yok} Bu küme için entropi hesabı yapmak için; C_1sınıfı "var", C_2 sınıfı "yok" olarak adlandırılır.Bu durumda,

$|C_1| = 6$

$|C_2| = 4$ için olasılıklar $p_1 = \frac{6}{10} = 0.6$ ve $p_2 = \frac{4}{10}$ biçiminde hesaplanır. Olasılık dağılımı da şu şekilde yazılır:

$$p_{Risk} = \frac{6}{10}, \frac{4}{10}$$

$H(Risk) = -\sum_{i=1}^{n} p_i \log_2(pi)$ eşitliğini kullanarak Risk kümesi için entropi şu şekilde hesaplanır:

$$H(Risk) = -\left(\frac{6}{10}\log 2\frac{6}{10} + \frac{4}{10}\log 2\frac{4}{10}\right)$$

$$= 0.97$$

2.5.1.1.2. C4.5 Algoritması

J.Quinlan tarafından 1993'te sayısal niteliklere sahip verilerden de karar ağaçları oluşturabilmek ve ID3 algoritmasını geliştirmek amacıyla oluşturulmuştur. Bir düğümden çıkan çoklu dallarla ağaç üretir. Dalların sayısı ön görülen kategorilerin sayısına eşittir. Tek bir sınıflandırıcı içerisinde birden fazla karar ağacını birleştirir. Dallanma için bilgi kazancı kullanılır. Her yaprağın hata oranı esas alınarak budama işlemi yapılır (Bounsaythip ve Runsala, 2001, s.29).

C4.5, ID3'e göre üstünlükleri (Koyuncugil, 2006, ss.73-75):

• Kayıp veriye sahip tahmin ediciler hala kullanılabilmektedir,

• Sürekli değere sahip tahmin ediciler kullanılabilmektedir,

• Budama yapılabilmektedir.

• Kural çıkarma işlemi yapılabilmektedir.

Budama(prunning): Gürültülü veya aykırı veri içeren kısımları ortadan kaldırmak için yapılır. Budanmış ağaçların daha küçük ve daha az karmaşıktır (http://www.tutorialspoint.com).

İki tür budama işlemi vardır, Ön budama (preprunning) ve son budama (post prunning). Ön budama, genelde ağaç oluşturulurken bazı önemsiz, null ve anlamsız dallar ağaca

hiç eklenmez. Son budama da ise ağacın bütün dalları oluşturulur ve sonra saptanan kurallara göre budama işlemi yapılır.

2.5.1.1.3. Chaid Algoritması

Hedef değişkeni, tahmin edici değişkenler ile ilişki düzeyine göre sınıflandırma amacıyla; diğer karar ağacı algoritmalarından farklı olarak ikiden fazla gruba ayırarak dallanan algoritma, tüm olası alt grupları ağaç biçiminde kolay anlaşılır biçimde göstermektedir (Koyuncugil, 2006, s.75).

Bu algoritmanın en belirgin özelliği hem bir karar ağacı algoritması olması hem de istatistiğe dayalı bir algoritma olmasıdır (Silahtaroğlu, 2013).

Chaid sürekli ve kategorik tüm değişken tipleriyle çalışabilmektedir. Bununla beraber, sürekli tahmin edici değişkenler otomatik olarak analizin amacına uygun olarak kategorize edilmektedir (Koyuncugil, 2006, s.75).

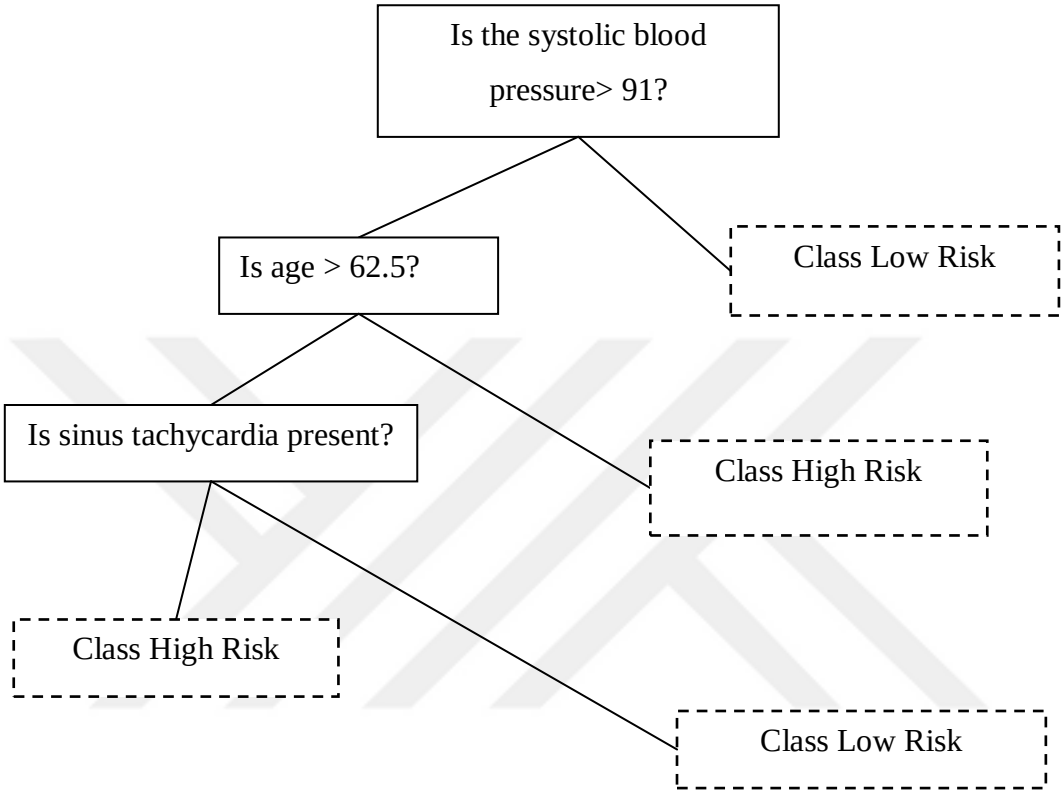
Chaid, bölümlendirme amaçlı kullanılan etkili bir istatistiksel tekniktir. Bir istatistiksel testin anlamlılığını kriter olarak kullanarak, bir potansiyel ön kestirici değişkenin tüm değerlerini değerlendirir. Hedef değişkene veya aynı anlama gelmek üzere bağlı değişkene göre istatistiksel olarak homojen (benzer) olarak değerlendirilen tüm değerleri birleştirir ve diğer tüm değerleri heterojen (benzer olmayan) olarak değerlendirir. Ardından karar ağacındaki ilk dalın formuna göre en iyi ön kestirici değişkenin seçilmesiyle, her bir düğümün seçilen değişkenin homojen değerlerinin bir grubunu oluşturmasını sağlar. Bu süreç ardıl olarak ağaç tamamıyla büyüyene kadar sürer. Kullanılan istatistiksel test, hedef değişkenin ölçüm düzeyine bağlı olarak değişir. Eğer hedef değişken sürekli bir değişken ise, F testi kullanılır. Eğer hedef değişkeni kategorik ise ki-kare testi kullanılır (Oğuzlar, 2004, s.84)

2.5.1.1.4. C&RT Algoritması

1984 yılında Breiman ve arkadaşları tarafından Regresyon ve sınıflandırma merkezli olarak oluşturulmuştur. Kök düğümden başlayarak iki çocuk düğüme bölünerek inşa edilir(http://public.dhe.ibm.com).

C&RT sözde karar ağacı oluşturmak için geçmiş verileri kullanan bir sınıflandırma yöntemidir. Karar ağaçları daha sonra yeni verileri sınıflandırmak için kullanılır. C&RT'nin kullanılabilmesi için sınıfların önsel değerlerinin bilinmesi gerekir. Karar

ağacı oluşturulurken, sözde öğrenme kümesi oluşturmak için, tüm gözlemler için önceden atanmış sınıflar ile geçmiş veri kümesinin değerleri kullanır. Hem sayısal hem de kategorik veriler ile çalışabilir. San Diego Tıp Merkezi tarafından hastalarının sınıflandırılması için kullanılan sınıflandırma ağacı Şekil 2.2'de gösterilmiştir.



Şekil 2.2: San Diego Tıp Merkezi Hasta Sınıflandırma Ağacı

Kaynak: (Timofeev, 2004).

Ağacın hedefi benzer veya aynı çıktı değerlerine sahip olma eğiliminde olan alt gruplar oluşturmaktır. C&RT 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) yöntemi kullanılabilir (Oğuzlar, 2004, s.83)

Gini: Bu algoritma karar ağacı oluşturulmasında kullanılan bir algoritmadır. C&RT ağacının ilk hangi nitelikten bölüneceği ve bölünme değeri Gini indeks değerine bakılarak karar verilir. Gini indeks değeri veri setindeki varlıkların oranı olarak tanımlanabilir. İki varlığın Gini değeri aynı çıkarsa sonuç dağılımları aynı demektir. Eğer veri setindeki bir nitelikte 3 veya daha fazla seçenek bulunuyorsa ve ikiden fazla

bölünmeye izin verilmediği için birbirine yakın seçenekler gruplandırılır (Adak ve Yurtay, 2013).

Gini indeksi Gini sol ve Gini sağ değerinden elde edilir. Hesaplamalar şu şekilde yapılır:

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left[\frac{L_i}{|T_{sol}|}\right]^2 \quad (2.4)$$

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left[\frac{R_i}{|T_{sağ}|}\right]^2 \quad (2.5)$$

k: Sınıf sayısı

T: Bir düğümdeki örnek sayısı

T sol: Sol koldaki örneklerin sayısı

T sağ: Sağ koldaki örneklerin sayısı

Li: Sol kolda i kategorisindeki örneklerin sayısı

Ri: Sağ kolda i kategorisindeki örneklerin sayısı

$$Gini_j = \frac{1}{n}\left(|T_{sol}|Gini_{sol} + |T_{sağ}|Gini_{sağ}\right) \quad (2.6)$$

Hesaplanan değerler arasından en küçük olan seçilerek bölünme başlatılır. Kalan verilere de aynı işlem basamakları uygulanarak diğer bölünmeler hesaplanır.

Twoing: Bu algoritma özniteliklerin içerdiği değerler göz önüne alınarak eğitim kümesi aday bölünme olarak adlandırılan iki ayrı dala ayrılır. Bir t düğümünde sağ $(\boldsymbol{t_{sağ}})$ve sol $(\boldsymbol{t_{sol}})$şeklinde kümelerden oluşan iki dal bulunur. Regresyon ağacı oluşturulmasında kullanılacak her bir veri sağ ve sol dala bölünmeye adaydır. Twoing kuralında öncelikle her bir aday için sağ ve sol taraftaki dalda olma olasılıkları hesaplanır. Her bir aday verinin sol taraftaki dala bölünme olasılığı P_{sol}ve $P(j/t_{sol})$, sağ taraftaki dala bölünmesi olasılığı ise $P_{sağ}$ve $P(j/t_{sağ})$, şeklinde ifade edilir. Olasılıkların hesaplanmasının ardından t düğümündeki s aday bölünmelerinin uygunluk ölçüsü:

$$\theta\left(\frac{s}{t}\right) = 2P_{sol}P_{sağ}\sum_{j=1}^{n}\left|P(\frac{j}{t_{sol}}) - P(j/t_{sağ})\right| \quad (2.7)$$

Şeklin de hesaplanır. Bu eşitlikte j özniteliklere ait sınıf etiket değerini sembolize eder. Dallanmayı oluşturacak satırı belirlemek için, hesaplama sonucu elde edilen değerler içerisinden en büyük olanı seçilir. Böylece, regresyon ağacının ilk ikili dallanması

gerçekleştirilmiş olur. Ağaç dallanması için alt kümelere işlem basamakları tekrar uygulanır (Larose, 2005).

2.5.1.2. Yapay Sinir Ağları

Yapay sinir ağları insan beyninin en temel özelliği olan öğrenme fonksiyonunu taklit ederek öğrenmeyi ve karar vermeyi gerçekleştiren bilgisayar sistemidir. Öğrenme işlemi örnekler yardımı ile yapay sinir ağına öğretilerek gerçekleştirilir. Bu ağlar birbirine bağlı süreç elemanlarından (yapay sinir hücrelerinden) oluşur. Her bağlantının bir ağırlık değeri vardır. Yapay sinir ağının sahip olduğu bilgi bu ağırlık değerlerinde saklanarak ağa iletirilir. Yapay sinir ağları bilinen hesaplama yöntemlerinden farklı bir hesaplama yöntemi kullanmaktadır. Bulundukları ortama kolayca uyum sağlayabilen, adaptif, eksik bilgi ile çalışabilen, belirsizlikler altında karar verebilen, hatalara karşı esnek olan bir yöntemdir. Oluşturulacak olan ağın yapısının belirlenmesinde, ağ parametrelerinin seçiminde, belirli bir standardın olmaması, problemlerin sadece nümerik bilgiler ile gösterilebilmesi, eğitimin nasıl bitirileceğinin bilinmemesi ve ağın davranışlarını açıklayamamasına rağmen kullanımı yaygındır.

Sınıflandırma, ilişki tanımlama, sinyal süzme, veri küçültme ve optimizasyon uygulamalarında yapay sinir ağları büyük başarılar elde etmiştir Veri madenciliği, optik karakter taşıma, optimum rota belirleme, parmak izi tanıma, malzeme analizi, iş çizelgelemesi ve kalite kontrol, tıbbi analizler gibi bir çok alanda uygulanmaktadır (Öztemel, 2003, s.23).

Bir yapay sinir ağı 5 bölümden oluşur:

Girdiler

Proses elamanın dış ortamdan bilgileri (verileri) alan elemanlarıdır. Veriler bu safhada bir işleme tabi tutulmadan direk olarak iletilirler (Siyambaş, 2014).

Ağırlıklar

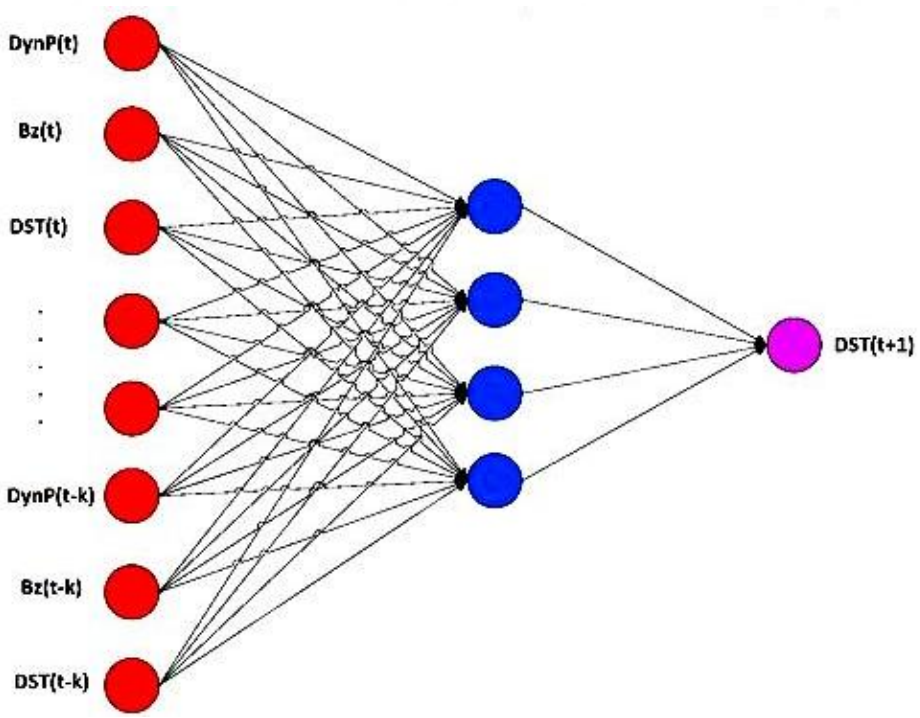
Ağırlıklar, yapay sinir hücresine girdi katmanından gelen bilgilerin hücrenin üzerindeki etkisini gösteren katsayılardır. Bütün girişlerin kendine ait bir ağırlığı vardır ve bu değerler değişken veya sabit olabilmektedir. Eğer ağırlığın değeri küçükse o girişin sinire zayıf bağlanmış olduğunu gösterirken, büyük değere sahip ağırlıklar da sinire

güçlü bağlanmış anlamına gelir (Bulut, 2011). Fakat ağırlığın küçük ve büyük olması önemli olup olmadığı anlamına gelmez.

Toplama İşlemi: Toplama işleminde, her bir giriş değerinin bağlı olduğu ağırlıkla çarpımlarının toplamı ve Θj eşik değerinin toplama işlemi gerçekleştirilir. Elde edilen bu toplam aktivasyon fonksiyonuna gönderilir (Bulut, 2011).

Aktivasyon Fonksiyonu: Aktivasyon fonksiyonu, toplama işleminden gelen sonucu, üzerinde gerekli işlemler gerçekleştirildikten sonra çıkışa iletir. Etkinleştirme fonksiyonu veya transfer fonksiyonu olarak da isimlendirilen aktivasyon fonksiyonun çıkışı hesaplamak için farklı formülleri vardır (Bulut, 2011).

Basit bir yapay sinir ağı görünümü Şekil 2.3'te verilmiştir.



Girdi Katmanı Gizli Katman Çıktı Katmanı

Şekil 2.3: YSA'nın Genel Yapısı

Kaynak:(https://filebox.ece.vt.edu).

Bir yapay sinir ağı 3 katmandan oluşur; girdi katmanı, gizli katman ve çıkış katmanı. Girdi katmanı dışarıdan gelen bilgiyi alarak işlem yapmadan gizli katmana iletir. Bu katmanda bilgi işlem yapılmadan aktarılır. Girdiden gelen bilgi gizli katmanda işlenir. Gizli katmanının ağırlıklarının sayısı değişkendir. Gizli katman ağ ile ilgili

çözümlemeyi yapıp problemi çözdükten sonra çıktı katmanına iletir. Çıktı katmanı da ağın çözümleme sonucunu rapor eder.

2.5.1.3. Genetik Algoritmalar

Çok değişkenli ve doğrusal olmayan optimizasyon problemlerinin çözümünde kullanılan sezgisel bir yöntemdir. Evrimin işleyişini örnek alarak 1975 yılında John Holland tarafından oluşturulmuştur. En iyi olan hayatta kalır ilkesini temel alarak En iyi olan birimlerin genleri sonraki iterasyondaki genlere aktarılır ve kötü olan birimler elenir.

Bir problemi çözebilmek için öncelikle rastgele başlangıç çözümleri belirlenmektedir. Daha sonra bu çözümler birbirleriyle eşleştirilerek performansı yüksek olan çözümler üretilmektedir. Bu şekilde sürekli çözümler birleştirilerek yeni çözümler aranmaktadır. Bu arama en iyi sonuç üretilmeyinceye kadar devam etmektedir. Genetik algoritmalar ile problemlerin çözülmesinde arzu edilen sonucu üretecek özelliklerin kalıtım yoluyla başlangıç çözümlerinden elde edilen yeni çözümlere onlardan da daha sonraki çözümlere geçtiği kabul edilmektedir (Öztemel, 2003).

Genetik algoritmaları belirli özelliklere göre sınıflandırmak zordur. Bunun sebebi, genetik operatörlerin, bütün parametrelerin, amaç fonksiyonunun ve uygunluk fonksiyonunun probleme özel olarak değişebilme durumundan kaynaklanmaktadır. Genetik algoritma yapısı optimum veya optimuma yakın sonuca daha hızlı ulaşmak için her türlü şekilde uyarlanabilmektedir. Hatta genetik operatörlerde bile köklü değişiklikler yapılabilmektedir. Mutasyonun uygun olmayan çözüm yaratma olasılığı bulunan bir problemde kullanılmaması veya problemde geleneksel mutasyon operatörü yerine kromozomda uygun olmayan çözümü engelleyen probleme özel tasarlanmış mutasyon operatörü kullanılması ve kromozomun ikili sayı sisteminde kodlanması yerine onlu sayı sisteminde ya da altmış dörtlük sayı sisteminde kodlanması örnek olarak verilebilir. Genetik algoritmaları 3 türdedir (Telcioğlu, 2007),

- Basit genetik algoritmalar: Alışılagelmiş genetik algoritmalar olarak adlandırılır.
- Paralel genetik algoritmalar: Birden çok işlemci bulunan bir sistemde hesap yükünün işlemcilere uygun şekilde dağıtılması olarak ifade edilir.
- Melez genetik algoritmalar: Genetik algoritma içerisine bir başka tekniğin oturtulmasıyla daha iyi performans sağlayan algoritmalardı.

2.5.1.4. K En Yakın Komşu Algoritması

K En Yakın Komşu yöntemi, sınıflandırma problemini çözen denetimli öğrenme yöntemleri arasında yer alır. Yöntemde; sınıflandırma yapılacak verilerin öğrenme kümesindeki normal davranış verilerine göre benzerlikleri hesaplanarak; en yakın olduğu düşünülen k verinin ortalamasıyla, belirlenen eşik değere göre sınıflara atamaları yapılır. Önemli olan, her bir sınıfın özelliklerinin önceden net bir şekilde belirlenmiş olmasıdır. Öğrenmeli bir yöntemdir. Yöntemin performansını k en yakın komşu sayısı, eşik değer, benzerlik ölçümü ve öğrenme kümesindeki normal davranışların yeterli sayıda olması kriterleri etkilemektedir (Çalışkan ve Soğukpınar, 2015, ss.120-123).
Genellikle büyük veri tabanlarında tercih edilen bir sınıflandırma tekniğidir. Sınıflandırılmak istenen nesnenin ait olduğu sınıfı, en yakınında yer alan K birim nesneden en fazla birime ait olanla aynı kümede sınıflandırması mantığına dayanmaktadır (Koyuncugil ve Özgülbaş, 2009, s.27).
Kolay anlaşılabilir bir algoritmadır ve genellikle sayısal nitelikteki verilerde kullanılır. Gürültülü verilerle çalışabilmesi avantajlarındandır. Öklid, Manhattan ve Minkowski Uzaklık Ölçütleriyle en yakın komşu hesaplanır.

Öklid uzaklığı şu formül ile hesaplanmaktadır : (http://www-users.cs.umn.edu)

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots (x_{ip} - x_{jp})^2} \quad (2.8)$$

Manhattan uzaklığı:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}| \quad (2.9)$$

Minkowski uzaklığı:

$$d(i,j) = \left(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \cdots + |x_{ip} - x_{jp}|^q\right)^{\frac{1}{q}} \quad (2.10)$$

2.5.1.5. Naive Bayes

Naive Bayes yöntemi hedef değişken ile bağımsız değişken arasındaki ilişki durumunu ortaya koyar. Bu yöntemde verilere ait özellikler verinin sınıf bilgilerinden varsayımda bulunduğu sınıfa ait olma olasılığını her sınıf niteliği için hesaplar ve en yüksek

olasılığa sahip sınıfın olasılığını kabul ederek sınıflandırır. Hızlı ve kullanımı kolaydır. Daha çok metin sınıflandırılmasında kullanılır.

2.5.1.6. Lojistik Regresyon Analizi

Lojistik regresyonun iki amacı vardır: Sınıflandırma ve bağımlı- bağımsız değişken arasındaki neden sonuç ilişkilerini incelemektedir. Matematiksel olarak esnek ve yorumlanması kolay bir modeldir. Normal dağılım sürekli dağılım gibi şartlı kısıtlama durumları yoktur. Bağımlı değişkenler kesikli değerler alır. Bağımlı değişkenin alabileceği değerlerin gerçekleşme olasılığı bulunur. Sonuç değişkeni evet-hayır, başarılı-başarısız gibi ikili veya çokludur. Biyoloji, tıp, ekonomi ve taşımacılık alanında yaygın kullanıma sahiptir.

2.5.2. Kümeleme

Kümeleme bir dizi içinde bulunan birlerine göre yüksek benzerliğe sahip nesneleri benzerliklerine göre gruplama yöntemidir. Bir dizideki nesneleri sınıflara atayan bir prosedür ve denetimsiz sınıflandırmaya bir örnektir. Denetimsiz olması nedeni, sınıflandırma sırasında önceden tanımlanmış sınıfların ve eğitim örneklerinin bulunmamasıdır (Jiang, et al., 2004). Genel kabul görmüş iki tip kümeleme algoritması vardır:

- Hiyerarşik kümeleme,
- Hiyerarşik olmayan kümeleme.

2.5.2.1. Hiyerarşik Kümeleme

Hiyerarşik kümeleme en basit yöntemlerdendir. Bu yöntemde kümeleme işlemi toplaşımlı ve bölücü olarak yapılabilir. Toplamalı kümeleme de veri tabanındaki her bir veri bir sınıf olarak görülür ve benzer özelliklere sahip olanlar birleştirilir. Bölücü kümeleme de ise, tüm bireysel nitelikler tek bir kümeymiş gibi görülür ve özelliklerine göre bölünerek kümelenir. Küme sayısını manuel olarak belirlenir. Bu dezavantaj olarak görülebilir.

Kümeleme analizinde şu etkenleri kullanıcı belirler:

- Olgular arasındaki benzerliği veya mesafeyi belirlemek için bir kriter,
- Verileri sınıflandırmak için küme sayısı,
- Kümeleme için hangi yöntemin kullanılacağı (http://www.norusis.com).

Kümeleme analizinde olgular arasındaki mesafe ve uzaklığı belirlemek için Öklid uzaklığı Manhattan uzaklığı ve Minskowski uzaklığı kullanılır.
Hiyerarşik kümeleme algoritmaları: Slink, Chameleon, Birch, Cure, Rock algoritmaları örnek verilebilir.

2.5.2.2. Hiyerarşik Olmayan Kümeleme

Bu teknik büyük veritabanlarına uygulanır. Kullanıcının önceden bir küme sayısı girmesi istenir. Genellikle K- ortalama tekniği kullanılır.

K ortalama
MacQueen " k − ortalama" terimini her bir birimin en yakın merkezli kümeye atanması süreci anlamında kullanmıştır. k − ortalama tekniği, gözlemleri kümelerin önceden belirlenen sayısına göre gruplandırmakla işleme başlar. Böylece her biri tek gözlemden oluşan k tane küme ile işleme başlanır ve her bir yeni gözlem en yakın ortalamalı gruba eklenir. Gruba yeni bir gözlem eklendikten sonra küme ortalaması yeniden hesaplanır. Bu süreç tüm gözlemler gruplara atanıncaya kadar devam eder. Tüm gözlemler gruplara atandıktan sonra atandıkları küme ortalamasından daha yakın küme ortalaması varsa, gözlemlerin yerleri değiştirilmektedir. Amaç diğer kümeleme yöntemlerinde olduğu gibi, gerçekleştirilen kümeleme işlemi sonucunda elde edilen kümelerin, küme içi benzerliklerinin maksimum, kümeler arası benzerliklerinin ise minimum olmasını sağlamaktır. Küme benzerliği, kümenin ağırlık merkezi kabul edilen bir birim ile kümedeki diğer birimler arasındaki uzaklıkların ortalama değer ile ölçmektir (Atbaş, 2008).

2.5.3. Birliktelik Kuralı

Büyüyen veri tabanları ve bu büyük veri yığınları içerisinde gizli kalmış birçok önemli bilgi ve örüntü vardır. Birliktelik kuralları veritabanları içerisinde gizli kalmış birliktelikleri ortaya çıkarmak için oluşturulmuş bir algoritmadır. Bu algoritma sayesinde gizli örüntüler birliktelik analizleri ortaya çıkarılmıştır. Bir müşterinin bir ürünü satın alırken bir başka ürünü de bu ürünle birlikte satın alma olasılığı hesaplanarak raflar ve kampanyalar bu birliktelikler doğrultusunda gerçekleştirilmektedir. Örneğin bir gömlek alan bir kişinin kravat alma olasılığı hesaplanarak gömlek raflarının yanına kravat stantları yerleştirilmiştir. Bu uygulama

sayesinde çapraz satış işlemi hedeflenmiştir. Süper marketlerde de birçok uygulaması vardır. Örneğin puding raflarının yanına süt rafları konumlandırılarak tamamlayıcı ürün politikası uygulanmıştır. Bütün bu politikaların temelinde birliktelik analizi ile saptanan bilgiler vardır. Ürün broşürleri, satış politikaları, müşteri memnuniyeti, sms kampanyaları, kişiye özel indirimler birliktelik kuralları sonucunda belirlenmektedir. Birliktelik analizinde kullanılan algoritmalar: AIS, SETM, Apriori, AprioriTid algoritmasıdır. Yaygın kullanılan Apriori algoritmasıdır. Destek ve güven ölçütü kullanarak bir eşik değer belirlenir ve çıkan sonuç eşik değeriyle karşılaştırarak arada ki ilişkiyi ortaya koyar.

2.5.4. Destek Vektör Makineleri

DVM'ler yapısal olarak, düşük boyutlu bir giriş uzayından alınan vektörleri, yüksek boyutlu bir diğer uzaya doğrusal olmayan bir biçimde taşıyan bir dönüşümdür. Dönüşümü gerçekleyen makine, sistem ya da ağ, dönüşümü belirleyen bir çekirdek ile tanımlanır. Sınıflama problemlerinde, yeteri kadar yüksek boyutlu uzaya taşınan vektörler doğrusal ayrıştırılabilir duruma gelirler. En uygun doğrusal ayrıştırıcı, ayrıştıran düzlemler arasından sınıflara uzaklığı en çok olanıdır. Pay olarak adlandırılan en yakın uzaklık; yüzeye en yakın olan vektörlerin belirlenmesi ile bulunur. DVM'ye de adını veren ve destek vektörler olarak adlandırılan bu vektörler, ayrıştıran düzlemi belirler ve DVM'lerin tasarımı için etkin bir yol sunarlar. Örüntü tanıma alanının geleneksel yöntemleri, eğitim kümesi üzerindeki başarımın en çoklanmasına ve böylece deneysel riskin en azlanmasına dayanırlar. DVM'ler deneysel ve yapısal risklerin ikisini de en azlayacak şekilde eğitilirler. DVM'lerin tasarımında genelleme hatası için verilen bir üst sınır en azlanır. Yaygın kullanım bulan ve etkin bir yöntem sunan DVM'lerin, yapısal olarak geliştirilme süreci devam etmekte ve artan bir eğilimle yeni uygulama alanlarında denenmektedir (Uçar, 2006).

ÜÇÜNCÜ BÖLÜM

UYGULAMA

İnsani Gelişme Endeksi (İGE) nedir?

İnsani Gelişme Endeksi (İGE) bir ülkenin ortalama kazanımlarını insani gelişmenin üç temel alanında ölçen özet bir karma endekstir: sağlık, bilgi ve gelir. İlk kez müteveffa Pakistanlı ekonomist Mahbub ul Haq tarafından, Nobel ödüllü Amartya Sen ve dönemin önde gelen diğer düşünce adamları ile işbirliği içerisinde Birleşmiş Milletler Kalkınma Programı (UNDP) tarafından 1990 yılında yayınlanmıştır. Ulusal kalkınmayı ölçmek amacıyla kullanılan gelir düzeyi ve ekonomik büyüme hızı gibi alışılageldik ölçümlere alternatif olarak ortaya konmuştur. İnsani gelişmenin üç temel boyutundaki uzun vadeli gelişmeyi değerlendirmek için kullanılan özet bir ölçüm yöntemidir. Bu üç temel boyut, uzun ve sağlıklı bir yaşam, bilgiye erişim ve insana yakışır bir yaşam standardıdır. Uzun ve sağlıklı yaşam boyutu, ortalama yaşam beklentisiyle, bilgiye erişim, yetişkin nüfus arasında ortalama okula gitme süresiyle, okula başlama yaşındaki çocuklar için beklenen okula devam süresi, yaşa dayalı, okula kaydolma konusunda hüküm süren oranların okul çağındaki bir çocuğun yaşamı boyunca aynı kalması durumunda, çocuğun toplam öğrenim görme süresi beklentisiyle değerlendirilmektedir. Yaşam standardı da, kişi başına düşen Gayri Safi Milli Gelir (GSMG) ile ölçülmektedir. Ülkelerin birbiriyle mümkün olduğunca iyi bir şekilde kıyaslanabilmesi için İGE, özellikle Birleşmiş Milletler Nüfus Bölümü (UNPD), Birleşmiş Milletler Eğitim, Bilim ve Kültür Teşkilatı UNESCO İstatistik Enstitüsü (UIS) ve Dünya Bankası'ndan (WB) sağlanan uluslararası veriler temel alınarak hazırlanmaktadır (undp.org).

Bir ülkenin kalkınmasının değerlendirilmesinde dikkate alınacak tek ölçütün ekonomik büyüme olmadığını, esas ölçütün kişiler ve onların kapasiteleri olduğunu vurgulamak için her yıl geliştirilen hesaplama teknikleri ile kayda değer ilerlemeleri belgeler. İGE aynı zamanda seçilen ulusal politikaların sorgulanmasında, Gayri Safi Milli Hasıla (GSMH) düzeyi aynı olan iki ülkenin nasıl olup da bu derece farklı insani gelişme sonuçlarına sahip olabilecekleri sorusunun cevabını sorgular. Rapor her yıl farklı temalar ile Küresel sorunları dile getirmekte ve çözümler aramaktadır. Dünya üzerindeki her ülke bu rapora dahil edilmek istenmekte ama güvenilir ve kesin verilere ulaşılamadığı için bazı ülkeler rapora dahil edilmemektedir.

Yıllara göre İnsani Gelişme Raporları temaları

1990 İnsani Gelişme Kavramı ve Ölçülmesi

1991 İnsani Gelişmenin Finansmanı

1992 İnsani Gelişmenin Küresel Boyutları

1993 Halk Katılımı

1994 İnsani Güvenliğin Yeni Boyutları

1995 Toplumsal Cinsiyet ve İnsani Gelişme

1996 Ekonomik Büyüme ve İnsani Gelişme

1997 Yoksulluğun Ortadan Kaldırılmasında İnsani Gelişme

1998 İnsani Gelişme İçin Tüketim

1999 İnsani Yüzü Olan Küreselleşme

2000 İnsan Hakları ve İnsani Gelişme

2001 Yeni Teknolojileri İnsani Gelişme İçin Kullanmak

2002 Parçalanmış Bir Dünyada Demokrasiyi Pekiştirmek

2003 Binyıl Kalkınma Hedefleri: İnsan Yoksulluğunu Ortadan Kaldırmak İçin Uluslararası Söz

2004 Günümüzün Farklı Dünyasında Kültürel Özgürlük

2005 Uluslararası İşbirliği Bir Kavşak Noktasında: Eşit Olmayan Bir Dünyada Yardım, Ticaret ve Güvenlik

2006 Kıtlığın Ötesinde: Güç Dengesizliği, Yoksulluk ve Küresel Su Krizi

2007/2008 İklim Değişikliğiyle Mücadele: Bölünmüş Bir Dünyada İnsan Dayanışması

2009 Engelleri Aşmak: Göç ve İnsani Gelişme

2010 Ulusların Gerçek Zenginliği: İnsani Gelişmenin Yolları

2011 Sürdürülebilirlik ve Eşitlik: Herkes İçin Daha İyi Bir Gelecek

2013 Güneyin Yükselişi: Farklılıklar Dünyasında İnsani gelişme

2014 İnsani İlerlemeyi Sürdürmek: Kırılganlıkları Azaltmak ve Dayanıklılık Oluşturmak (http://hdr.undp.org)

2015 İnsani Gelişme Bağlamında Çalışma Kavramını Yeniden Düşünmek (Rapor henüz yayınlanmamıştır.)

İGE 0 ve 1 arasında değerler almakta ve 1 yaklaşan değerler gelişmişlik düzeyinin yüksek olduğunu göstermektedir. UNDP gelişmişlik boyutunu 3 farklı kategoriye ayırmış ve Tablo 3.1'de gösterilmiştir.

Tablo 3.1: Gelişmişlik Sınıfları

0.550-0.699	Orta
0.700-0.799	Yüksek
>0.800	Çok yüksek

Kaynak:http://www.tr.undp.org

İGE'nin hesaplanması 2010 öncesi dönemlerde aşağıdaki formülle tanımlanmıştır (http://en.wikipedia.org).

$$x - index = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2.11)$$

Burada min (x)ve max (x) : x in alabileceği en yüksek (max) ve en az (min) değerdir.
İnsani Gelişme Endeksi aşağıdaki verileri temsil eder ve ortalamasını alır.

- Yaşam uzunluğu dağılımı = $\frac{HU-25}{85-25}$
- Eğitim dağılımı (ED) = $\frac{2}{3} \times ED + \frac{1}{3} \times GED$
- Yetişkin okur yazar oranı =$\frac{YOYO-0}{100-0}$
- Okullara kayıtlı öğrenci sayısı = $\frac{CGER-0}{100-0}$

HU: Yaşam uzunluğu
YOYO: Yetişkin okur yazar oranı
OKÖS: Okullara kayıtlı öğrenci sayısı yüzdesi
2010 sonrasında bu formüller değiştirilerek:
1. Yaşam Beklentisi Endeksi (LEI) $= \frac{LE-20}{85-20}$
2. Eğitim Endeksi (EL) =$\frac{MYSI+EYSI}{2}$
Okullaşma Endeksi (Mysi) 2.1 ortalama Yıl = $\frac{MYS}{15}$
Okullaşma Endeksi (EYSI) 2.2 Beklenen Yıl =$\frac{EYS}{18}$
3. Gelir Endeksi (II) =$\frac{In(GNIpc)-In(100)}{In(75{,}000)-In(100)}$

İGE (HDI) =$\sqrt[3]{LEI \times EI \times II}$ (2.12)

LE: Doğumda beklenen yaşam süresi

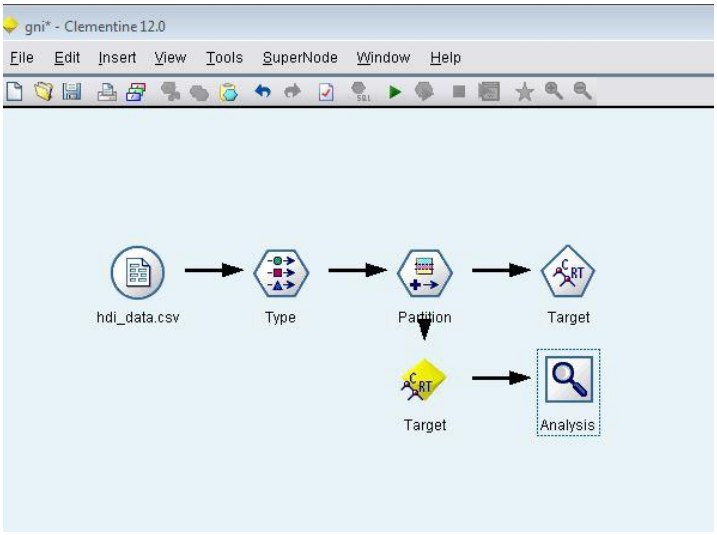
MYS: Ortalama eğitim yılı

EYS: Bütün hayat boyunca eğitime harcanacak olan yıl

GNI: Gayri safi milli gelire göre kişi başına düşen satın alma gücü.

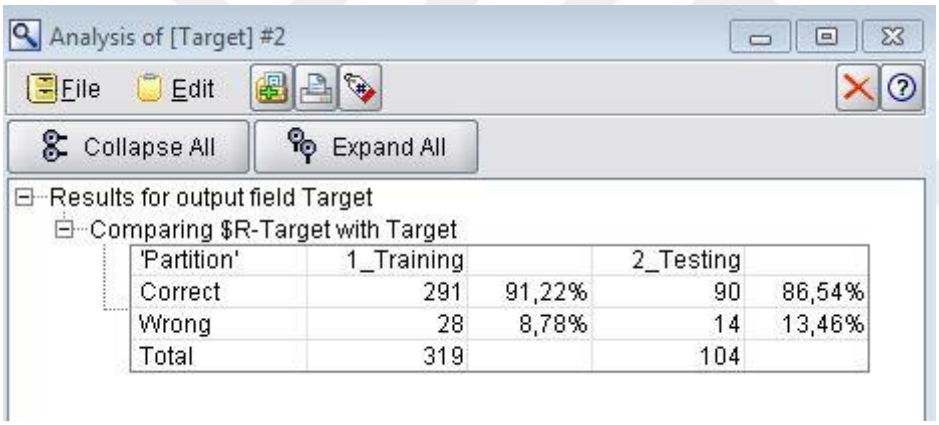
2010 yılından sonra hesaplamalarda aritmetik ortalama yerine geometrik ortalama kullanılmaya başlanmıştır. Bu nedenle bu çalışmada 2011-2012-2013 yıllarına ait 141 ülkenin veriler kullanılmıştır. Bu ülkeler EK 1 de gösterilmiştir. Bu veriler http://hdr.undp.org adresinden temin edilmiştir. 3 yıla ait tablodan Life expectancy at birth, Mean years of schooling, Expected years of schooling, Gross national income, GNI per capita rank minus HDI rank, Nonincome HDI Value sütunlarındaki değerler alınmıştır. Bu çalışmanın amacı İnsani Gelişme Endeksi'nin sınıflandırma başarısını karar ağacı algoritmaları ile tespit etmektir. Verilerin Analizinde yöntem olarak C&RT, C5.0 ve CHAID algoritmaları kullanılmıştır. Ülkeler gelişmişlik sınıflarına göre çok gelişmiş ülkelerden başlanarak 1-3 arasında değerler atanmıştır. Atanan değerler veri setinde bulunan İGE sütununa yerleştirilmiştir. Veriler için gerekli düzenleme işlemleri uygulandıktan sonra veriler Excel dosyası şeklinde kaydedilerek analiz için hazır hale getirilmiştir. Clementine programına Excel verileri tanıtılmıştır. Veriler %80'e %20 oranında eğitim seti ve test seti olarak ayrılmıştır.

GİNİ Algoritması



Şekil 3.1: GİNİ Algoritmasının Genel Görünümü

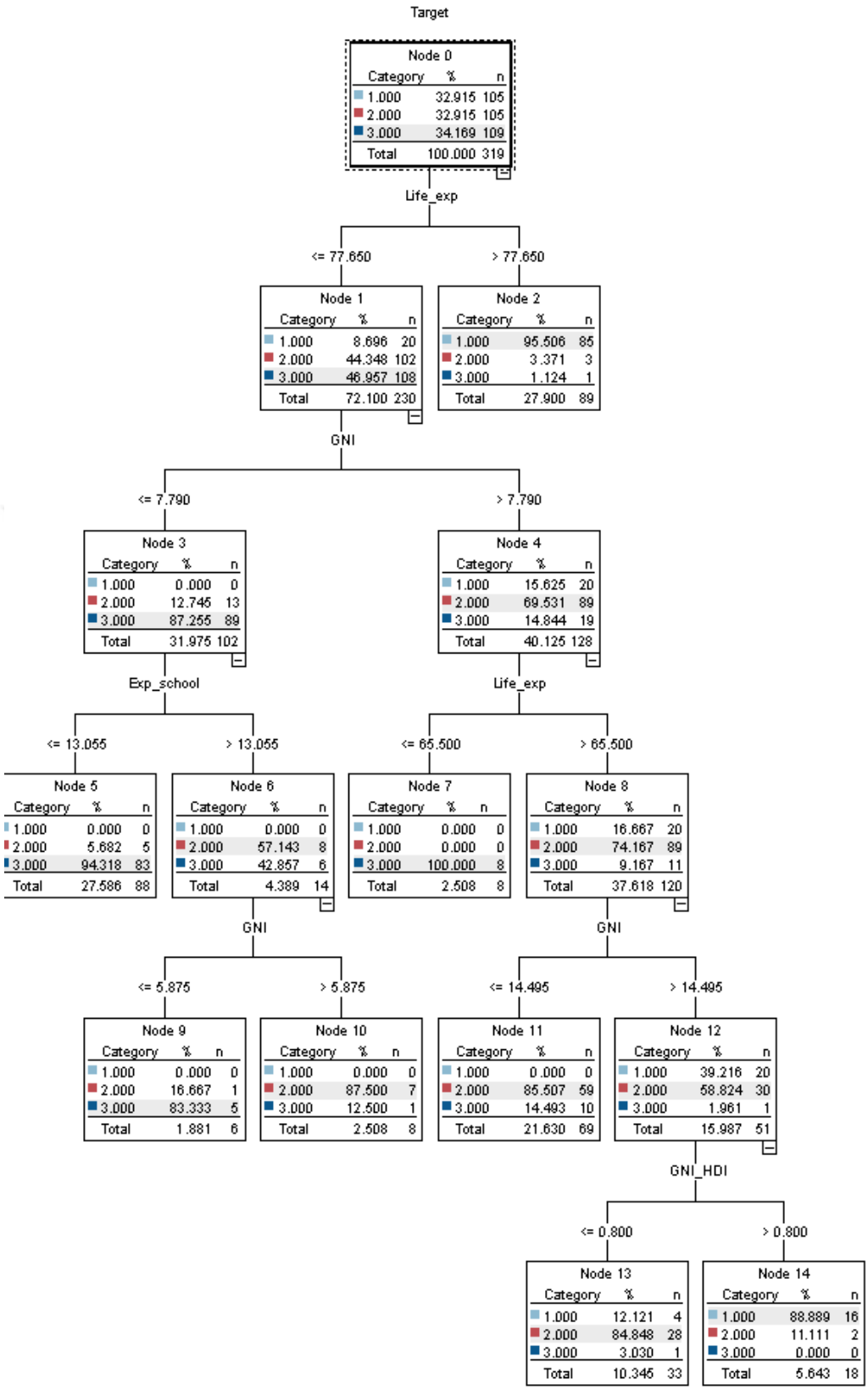
Analizin genel görünümü yukarıda Şekil 3.1 de verilmiştir. Bu analiz sonucunda aşağıdaki rakamlar elde edilmiştir.



'Partition'	1_Training		2_Testing	
Correct	291	91,22%	90	86,54%
Wrong	28	8,78%	14	13,46%
Total	319		104	

Şekil 3.2: GİNİ Algoritmasının Sınıflandırma Başarısı

Şekil 3.2 de gösterildiği gibi veriler %91.22 başarı oranıyla doğru sınıflandırılmıştır.



Şekil 3.3: GİNİ Algoritması ile Oluşturulmuş Karar Ağacı

Karar ağacının genel görünümü Şekil 3.3'te gösterilmiştir.

Karar ağacından şu kurallar elde edilmiştir:

1. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.650'den büyük ise veriler %95.506 doğruluk oranı ile "Çok Yüksek Gelişmiş" sınıfına atanır.

2. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.790'dan küçük veya eşit ve beklenen okullaşma yılı (Expected years of schooling) 13.055'ten küçük veya eşit ise veriler %94.318 başarı ile "Orta Gelişmiş" sınıfına aktarılır.

3. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.790'dan küçük veya eşit ve beklenen okullaşma yılı (Expected years of schooling) 13.055'ten büyük ve Gayri safi milli gelir (Gross national income)değeri 5.875'ten küçük ise veriler%83.33 ile "OrtaGelişmiş"sınıfına aktarılır. Değilse %87.500 sınıflandırma başarısı ile "Yüksek Gelişmiş" sınıfına aktarılır.

4. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.790'dan büyük ve doğumda beklenen yaşam süresi (Life expectancy at birth) 65.500'den küçük veya eşit ise veriler %100 sınıflandırma başarısı ile "Orta Gelişmiş" sınıfa dahil edilir.

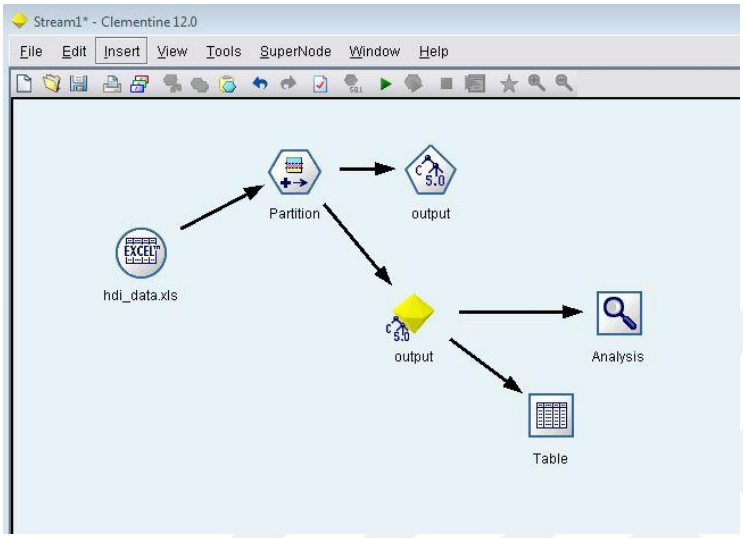
5. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.790'dan büyük ve doğumda beklenen yaşam süresi (Life expectancy at birth) 65.500'den büyük ve Gayri safi milli gelir (Gross national income)değeri 14.495'ten küçük veya eşit ise veriler %85.507 oranı ile "Yüksek Gelişmiş" sınıfa dahil edilir.

6. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.790'dan büyük ve doğumda beklenen yaşam süresi (Life expectancy at birth) 65.500'den büyük ve Gayri safi milli gelir (Gross national income)değeri 14.495'ten büyük ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.800'den büyük ise

veriler %88.889 oranı ile “Çok Yüksek Gelişmiş” sınıfına dahil edilir. Değilse %84.848 ile veriler “Yüksek Gelişmiş” sınıfına aktarılır.

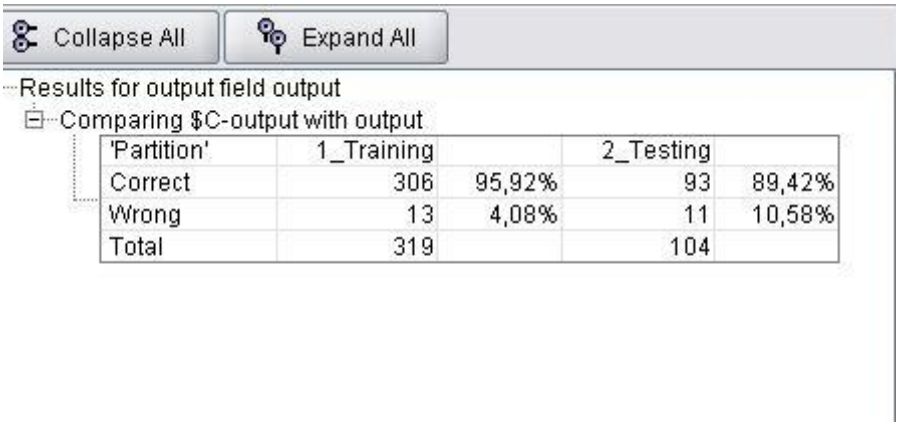
C5.0 Algoritması

Analizin genel görünümü Şekil 3.4’te gösterildiği gibidir.



Şekil 3.4: C5.0 Algoritmasının Genel Görünümü

Bu analiz sonucunda Şekil 3.5’te ki rakamlar elde edilmiştir.

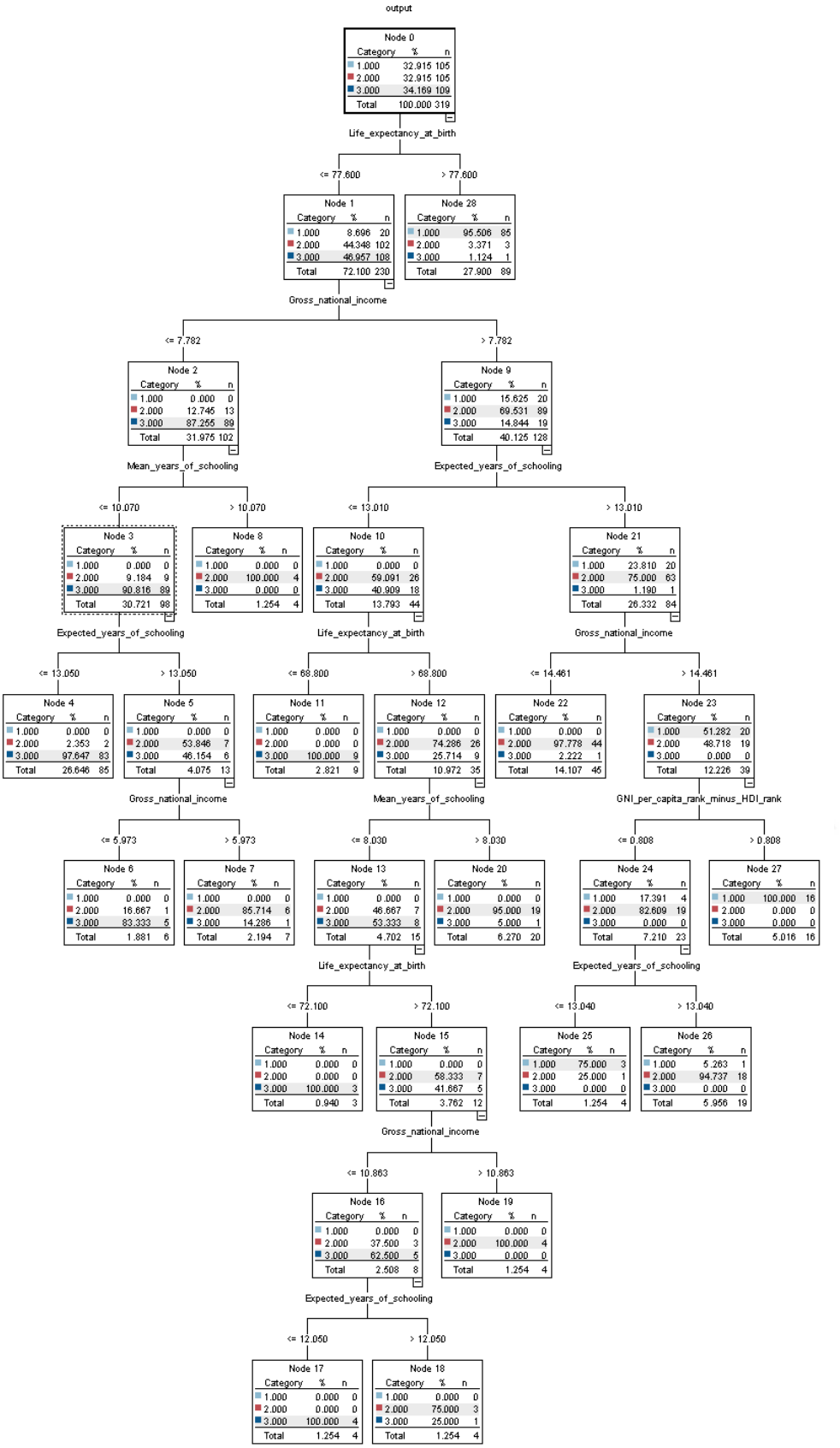


'Partition'	1_Training		2_Testing	
Correct	306	95,92%	93	89,42%
Wrong	13	4,08%	11	10,58%
Total	319		104	

Şekil 3.5: C5.0 Algoritmasının Sınıflandırma Başarısı

Şekil 3.5’te gösterildiği gibi veriler %95.92 başarı oranıyla doğru sınıflandırılmıştır.

Bu analiz sonucunda bulunan karar ağacının genel görünümü Şekil 3.6’da ki gibidir.



Şekil 3.6: C5.0 Algoritması ile Oluşturulmuş Karar Ağacı

Karar ağacından şu kurallar elde edilmiştir:

1. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den küçük veya eşit ve Ortalama okullaşma yılı (Mean years of schooling)10.070'den küçük veya eşit ise ve beklenen okullaşma yılı (Expected years of schooling) 13.050'den küçük veya eşit ise veriler %97.647güven ile "Orta Gelişmiş"sınıfına dahil edilir.

2. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den küçük veya eşit ve Ortalama okullaşma yılı (Mean years of schooling)10.070'den küçük veya eşit ise ve beklenen okullaşma yılı (Expected years of schooling) 13.050'den büyük ise Gross national income değeri 5.973'ten küçük ise veriler %83.33 güven ile "Orta Gelişmiş" sınıfına dahil edilir.

3. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den küçük veya eşit ve Ortalama okullaşma yılı (Mean years of schooling)10.070'den küçük veya eşit ise ve beklenen okullaşma yılı (Expected years of schooling) 13.050'den büyük ise Gross national income değeri 5.973'ten büyük ise %85.714 güven ile "Yüksek Gelişmiş" sınıfına dahil edilir.

4. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den küçük veya eşit ve Ortalama okullaşma yılı (Mean years of schooling)10.070'den büyük is %100 güven ile veriler "Yüksek Gelişmiş" sınıfına dahil edilir.

5. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den büyük ise %95.506 ile veriler"Çok Yüksek Gelişmiş"sınıfına dahil edilir.

6. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den büyük ve beklenen okullaşma yılı (Expected years of schooling) 13.010'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den küçük veya eşit ise veriler % 100 güven ile "Orta Gelişmiş" sınıfına dahil edilir.

7. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve beklenen okullaşma yılı (Expected years of schooling) 13.010'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030'dan büyük ise %95 güven ile "Yüksek Gelişmiş" sınıfına dahil edilir.

8. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den büyük ve beklenen okullaşma yılı (Expected years of schooling) 13.010'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 72.100'den küçük ise %100 güven ile "Orta Gelişmiş" sınıfına dahil edilir.

9. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den büyük ve beklenen okullaşma yılı (Expected years of schooling) 13.010'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 72.100'den büyük ve Gayri safi milli gelir (Gross national income)değeri 10.863'ten büyük ise veriler %100 güven ile "Orta Gelişmiş" sınıfına dahil edilir.

10. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri7.782'den büyük ve beklenen okullaşma yılı (Expected years of schooling) 13.010'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030'dan küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 72.100'den büyük ve Gayri safi milli gelir (Gross national income)değeri 10.863'ten küçük ve beklenen okullaşma yılı (Expected years of schooling) 12.050'dan küçükse %100 güven ile "Orta Gelişmiş" sınıfına değilse %75 güven ile "Yüksek Gelişmiş"sınıfına dahil edilir.

11. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010' küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den küçük veya eşit ise %100 güven ile "Orta Gelişmiş" sınıfına dahil edilir.

12. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010' küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030 'dan büyük ise veriler %95 güven ile "Yüksek Gelişmiş" sınıfına dahil edilir.

13. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010' küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030 'dan küçük eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth)72.100'den büyük ve Gayri safi milli gelir (Gross national income)değeri 10.863'ten büyük ise veriler %100 güven ile "Yüksek Gelişmiş" sınıfına dahil edilir.

14. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010' küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030 'dan küçük eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth)72.100'den büyük ve Gayri safi milli gelir (Gross national income)değeri 10.863'ten küçük ve ya eşit ise beklenen okullaşma yılı (Expected years of schooling) 12.050 den büyük ise %75 güven ile "Yüksek Gelişmiş" sınıfına değilse %100 güven ile "Orta Gelişmiş" sınıfına dahil edilir.

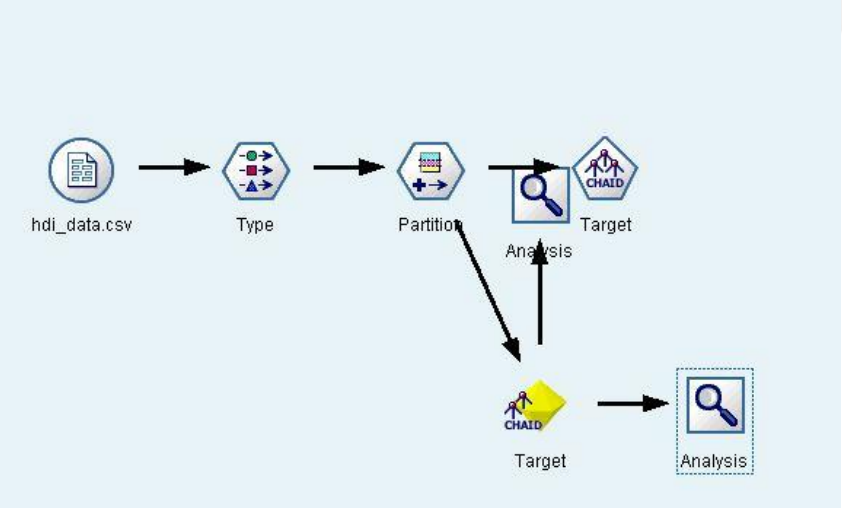
15. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010' küçük veya eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth) 68.800'den büyük ve ortalama okullaşma yılı (Mean years of schooling) 8.030 'dan küçük eşit ve doğumda beklenen yaşam süresi (Life expectancy at birth)72.100'den küçük ve ya eşit ise %100güven ile "Orta Gelişmiş" sınıfına dahil edilir.

16. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010'dan büyük ve Gayri safi milli gelir (Gross national income)değeri 14.461'den küçük veya eşit ise %97.778 ile "Yüksek Gelişmiş" sınıfına dahil edilir.

17. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010'dan büyük ve Gayri safi milli gelir (Gross national income)değeri 14.461'den büyük ve ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank)değeri0.808'den büyük ise %100 güven ile ile "Çok Yüksek Gelişmiş" sınıfına dahil edilir.

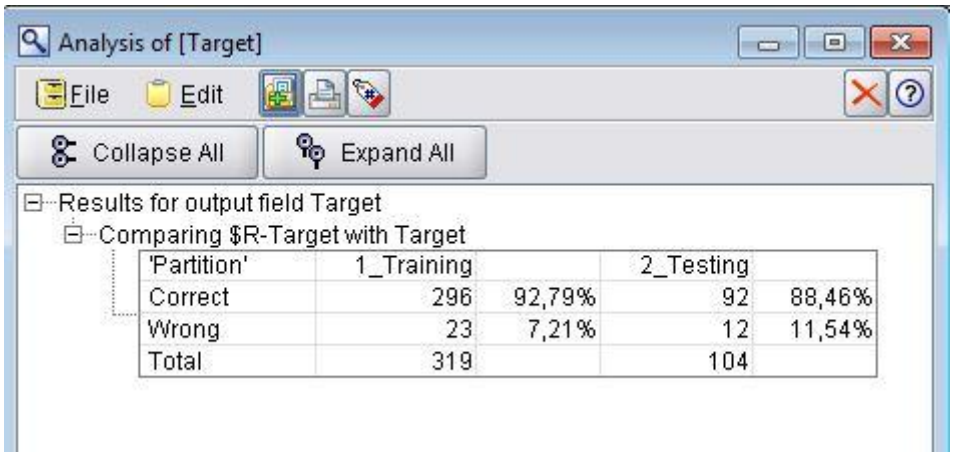
18. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77.600'den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 7.782'den büyük ve 13.010'dan büyük ve Gayri safi milli gelir (Gross national income)değeri 14.461'den büyük ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank)değeri0.808'den küçük veya eşit ve beklenen okullaşma yılı (Expected years of schooling) büyük 13.040'dan büyük ise veriler %94.737güven ile "Yüksek Gelişmiş" sınıfına değilse %75 güven ile "Çok Yüksek Gelişmiş" sınıfına dahil edilir.

CHAID Algoritması



Şekil 3.7: CHAID Algoritmasının Genel Görünümü

Analizin genel görünümü yukarıda Şekil 3.7 de verilmiştir. Bu analiz sonucunda Şekil 3.8'de ki rakamlar elde edilmiştir.

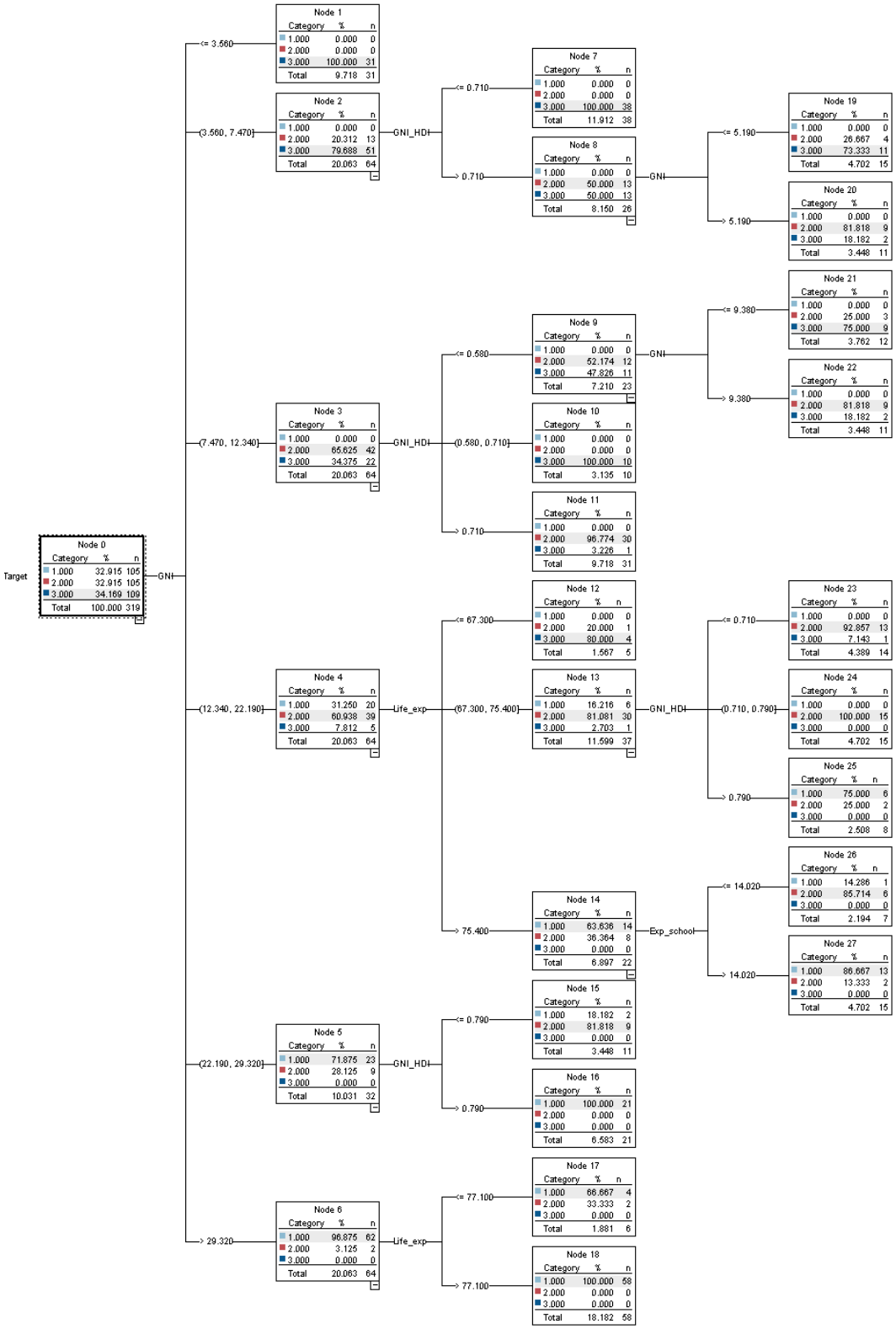


'Partition'	1_Training		2_Testing	
Correct	296	92,79%	92	88,46%
Wrong	23	7,21%	12	11,54%
Total	319		104	

Şekil 3.8: CHAID Algoritmasının Sınıflandırma Başarısı

Şekil 3.8'de gösterildiği gibi veriler %92.79 başarı oranıyla doğru sınıflandırılmıştır.

Karar ağacının genel görünümü Şekil 3.9'da verilmiştir.



Şekil 3.9: CHAID Algoritması ile Oluşturulmuş Karar Ağacı

Karar ağacından şu kurallar elde edilmiştir:

1. Eğer Gayri safi milli gelir (Gross national income)değeri 3.560 ‘dan küçük veya eşit ise veriler %100 doğruluk oranı ile “Orta Gelişmiş” sınıfına atanır.

2. Eğer Gayri safi milli gelir (Gross national income) değeri 3.560 ile 7.470 arasında ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) 0.710’dan küçük ve eşit ise veriler %100 başarı ile “Orta Gelişmiş” sınıfına atanır.

3. Eğer Gayri safi milli gelir (Gross national income) değeri 3.560 ile 7.470 arasında ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) 0.710’dan büyük ise Gayri safi milli gelir (Gross national income) değeri 5.190’dan küçük veya eşitse “Orta Gelişmiş” sınıfına atanır. Değilse “Yüksek Gelişmiş” sınıfına atanır.

4. Eğer Gayri safi milli gelir (Gross national income)değeri 7.470 ile 12.340 arasında ise insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.710’dan büyük ise veriler %96.774 başarı ile “Yüksek Gelişmiş” sınıfına atanır.

5. Eğer Gayri safi milli gelir (Gross national income)değeri 7.470 ile 12.340 arasında ise insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.580 ile 0.710 arasındaysa %100 başarı ile “Orta Gelişmiş” sınıfına atanır.

6. Eğer Gayri safi milli gelir (Gross national income)değeri 7.470 ile 12.340 arasında ise insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.580 ‘den küçük veya eşit ve Gayri safi milli gelir (Gross national income)değeri 9.380’den küçük veya eşit ise %75 başarı ile veriler “Orta Gelişmiş” sınıfına atanır. Değilse “Yüksek Gelişmiş” sınıfına atanır.

7. Eğer Gayri safi milli gelir (Gross national income) değeri 12.340 ile 22.190 arasında ve doğumda beklenen yaşam süresi (Life expectancy at birth) 67.300’den küçük is veriler % 80 değer ile “Orta Gelişmiş” sınıfına atanır.

8. Eğer Gayri safi milli gelir (Gross national income) değeri 12.340 ile 22.190 arasında ve doğumda beklenen yaşam süresi (Life expectancy at birth)

67.300 ile 75.400 arasında ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.710' dan küçük veya eşit ise veriler % 92.857 değer ile "Yüksek Gelişmiş" sınıfına atanır.

9. Eğer Gayri safi milli gelir (Gross national income) değeri 12.340 ile 22.190 arasında ve doğumda beklenen yaşam süresi (Life expectancy at birth) 67.300 ile 75.400 arasında ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.790' dan büyük ise %75 değer ile "Çok Yüksek Gelişmiş" sınıfına atanır.

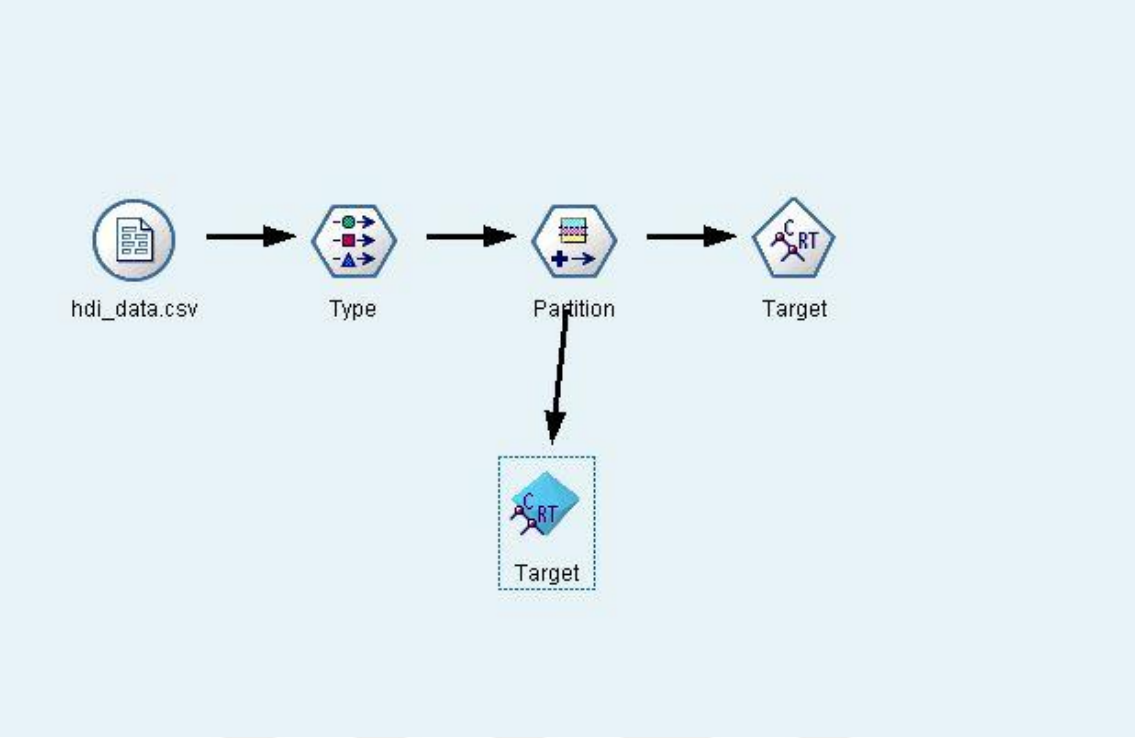
10. Eğer Gayri safi milli gelir (Gross national income) değeri 12.340 ile 22.190 arasında ve doğumda beklenen yaşam süresi (Life expectancy at birth) 67.300 ile 75.400 arasında ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.710 ile 0.790 arasındaysa veriler %100 ile "Yüksek Gelişmiş" sınıfına atanır.

11. Eğer Gayri safi milli gelir (Gross national income) değeri 12.340 ile 22.190 arasında ve doğumda beklenen yaşam süresi (Life expectancy at birth) 75.400'den büyük ise okullaşma yılı beklentisi (Expected years of schooling) değeri 14.020'den küçük veya eşit ise veriler %85.714 değeri ile "Yüksek Gelişmiş" sınıfına atanır. Değilse "Çok Yüksek Gelişmiş" sınıfına atanır.

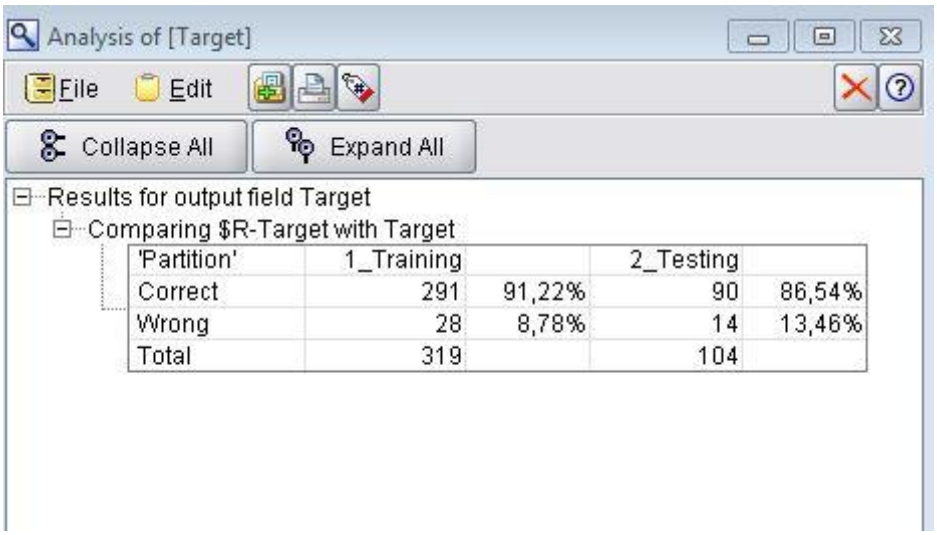
12. Eğer Gayri safi milli gelir (Gross national income)değeri 22.190 ile 29.320 arasındaysa insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0.790'dan büyük ise veriler %100 değeri ile "Çok Yüksek Gelişmiş" sınıfına atanır. Değilse %81.818 "Yüksek Gelişmiş" sınıfına atanır.

13. Eğer Gayri safi milli gelir (Gross national income)değeri 29.320'den büyük ise doğumda beklenen yaşam süresi (Life expectancy at birth) değeri büyük ise 77.100 'den "Çok Yüksek Gelişmiş" sınıfına atanır. Değilse % 66.667 ile "Çok Yüksek Gelişmiş" sınıfına atanır.

C&RT (Simple) Algoritması



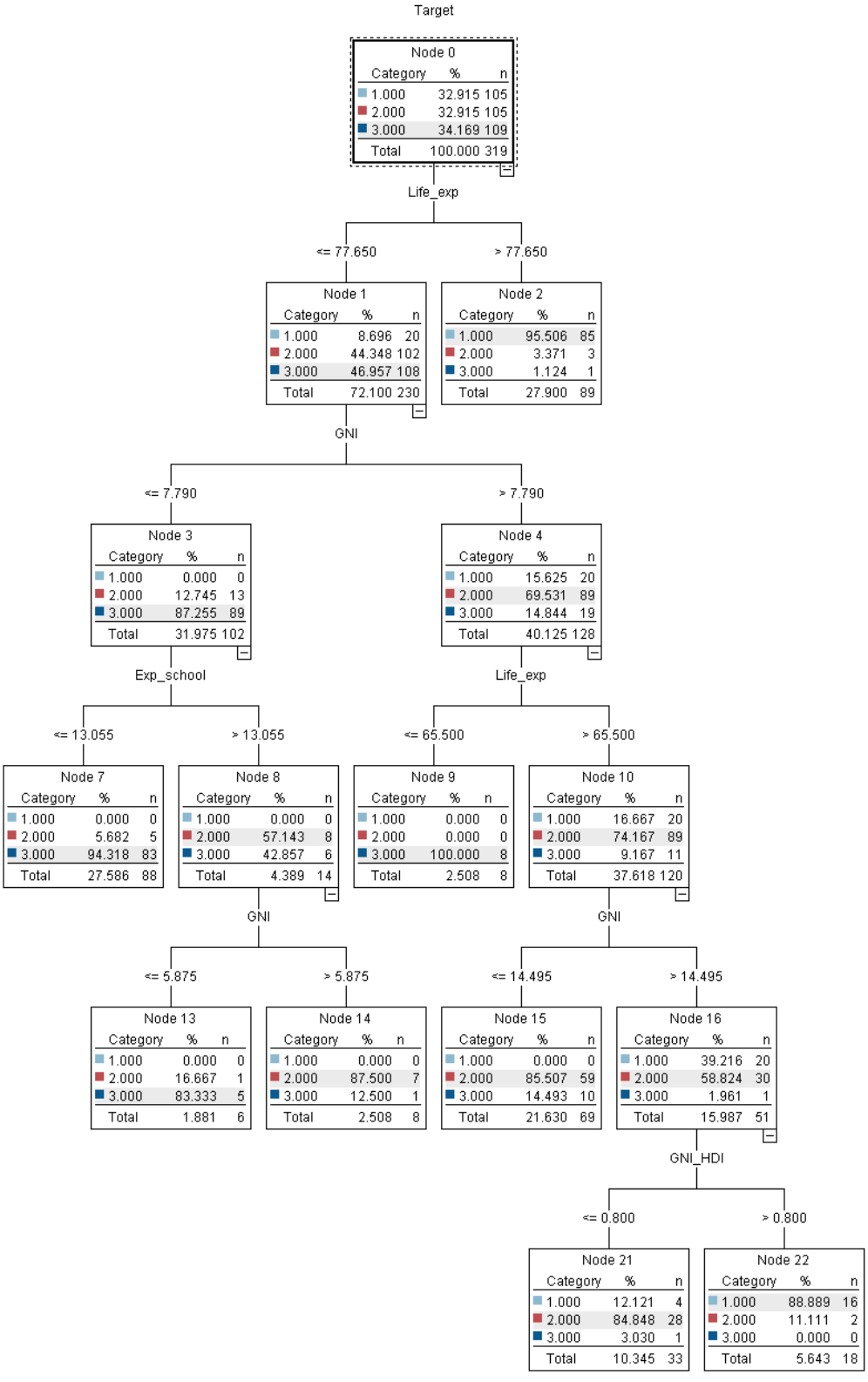
Şekil 3.10: C&RT (Simple) Algoritmasının Genel Görünümü



'Partition'	1_Training		2_Testing	
Correct	291	91,22%	90	86,54%
Wrong	28	8,78%	14	13,46%
Total	319		104	

Şekil 3.11: C&RT (Simple) Algoritmasının Sınıflandırma Başarısı

C&RT (Simple) Algoritmasının Sınıflandırma Başarısı Şekil 3.11'de gösterildiği gibi veriler %91.22 başarı oranıyla doğru sınıflandırılmıştır. Karar ağacının genel görünümü Şekil3.12'de gösterilmiştir.



Şekil 3.12: C&RT (Simple) Algoritması ile Oluşturulmuş Karar Ağacı

Karar ağacından şu kurallar çıkarılmıştır:

1. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77,650' den büyük ise %95.506 oranı ile "Çok Yüksek Gelişmiş" sınıfına aktarılır.

2. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77,650' den küçük veya eşit ve gayri safi milli gelir (Gross national income)değeri 7,790'dan küçük veya eşit ve okullaşma yılı beklentisi (Expected years of schooling) değeri 13,055 'ten küçük veya eşit ise veriler % 94.318 oranı ile "Orta Gelişmiş" sınıfına aktarılır.

3. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77,650' den küçük veya eşit ve gayri safi milli gelir (Gross national income)değeri 7,790'dan küçük veya eşit ve okullaşma yılı beklentisi (Expected years of schooling) değeri 13,055 'ten büyük ve gayri safi milli gelir (Gross national income)değeri 5,875'ten büyük ise %87.500 ile veriler "Yüksek Gelişmiş" sınıfına aktarılır. Değilse veriler %83.33 ile "Orta Gelişmiş" sınıfına aktarılır.

4. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77,650' den küçük veya eşit ve gayri safi milli gelir (Gross national income)değeri 7,790'dan büyük ve doğumda beklenen yaşam süresi (Life expectancy at birth) değeri 65,500'den küçük veya eşit ise veriler %100 ile "Orta Gelişmiş" sınıfına aktarılır.

5. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77,650' den küçük veya eşit ve gayri safi milli gelir (Gross national income)değeri 7,790'dan büyük ve doğumda beklenen yaşam süresi (Life expectancy at birth) değeri 65,500'den büyük ve safi milli gelir (Gross national income)değeri 14.495'ten küçük veya eşit ise veriler %85.507 ile "Yüksek Gelişmiş" sınıfına aktarılır.

6. Eğer doğumda beklenen yaşam süresi (Life expectancy at birth) 77,650' den küçük veya eşit ve gayri safi milli gelir (Gross national income)değeri 7,790'dan büyük ve doğumda beklenen yaşam süresi (Life expectancy at birth) değeri

65,500’den büyük ve safi milli gelir (Gross national income)değeri 14.495’ten büyük ve insani gelişmişlik endeksinde kişi başına düşen milli gelir sıralaması (GNI per capita rank minus HDI rank) değeri 0,800’den büyük ise veriler %88.889 ile “Çok Yüksek Gelişmiş” sınıfına aktarılır. Değilse %84.848 ile “Yüksek Gelişmiş” sınıfına aktarılır.

SONUÇ

Küreselleşen dünyada artan rekabet koşulları bilgiye olan ihtiyacı arttırmıştır. Kurumların yaptığı bütün işlemler bilgisayarlara kaydedilmektedir. İşlemler içlerinde önemli bilgiler barındırmaktadır. Veriler ham halde yarar sağlamayacağı için, dev veri tabanlarında bulunan karmaşık verilerin bilgiye dönüştürülerek anlamlandırılması gerekmektedir. Bu anlamlandırma işlemi çeşitli disiplinlerden yararlanılarak yapılmaktadır. Manuel hesaplamanın mümkün olmadığı, istatistiğin cevap veremediği, kimi zaman yetersiz kaldığı durumlar arttığı için yeni bir teknik, yeni bir bilim olarak veri madenciliği ortaya çıkmıştır.

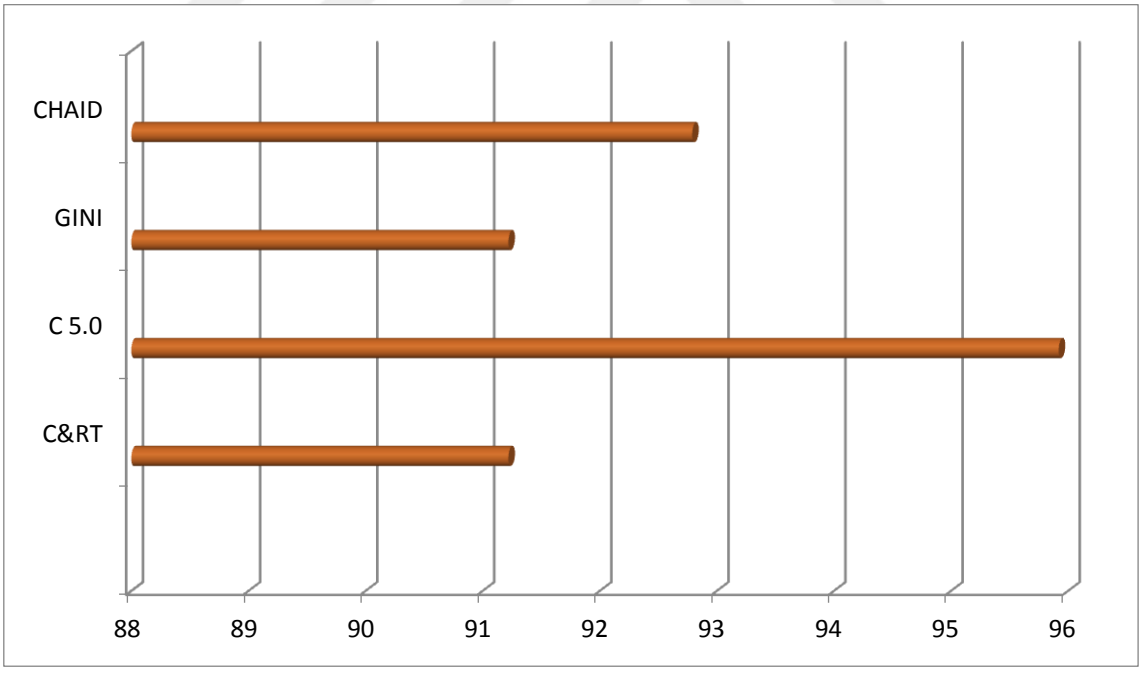
VM, farklı disiplinlerle iş birliği içinde çalıştığı için kapsamı geniştir. Bilim sayılabilecek niteliklere sahip olan bu teknik, dev veri tabanlarında anlamlandırılmayı bekleyen verileri ortaya çıkararak anlamlandırıp bilgiye dönüştürür. Veriler düzenlenip gerekli dönüştürme işlemlerinden geçirilerek problem çözümüne uygun hale getirilir. Belirlenen probleme uygun model seçilerek, veriler analize tabi edilir. Analiz sonucunda modelin uygunluğu test edilerek sonuçlar yorumlanır. Uygulama alanı oldukça geniş bir bilimdir. Bankacılık, Finans, Tıp, Savunma Sanayi, Eğitim, Üretim, Kalite gibi birçok alanda uygulama alanı bulmuştur. Kurumlar için rekabet avantajı sağlayacak stratejilerin geliştirilmesinde, kredi skorlamasın da yaygın olarak kullanılmaktadır. VM, Yapay Sinir Ağları ile uyum içerisinde çalışarak; görüntü işleme, desen tanıma teknikleriyle Askeriye ve Savunma Sanayisine önemli katkılarda bulunmuştur.

VM, tıp alanında devrim sayılabilecek uzman sistemlerin oluşturulmasında büyük katkıları olmuştur. Örneğin, DNA sıra analizlerinin yapılarak hastalıklı hücrelerin tespiti mümkün olabilmektedir. Ayrıca, bu yöntem gebelik sırasında bebek sıvısına yapılan müdahaleleri ortadan kaldırılarak düşük risklerini azaltılmış ve anne karnında operasyon yapılmadan down sendromu teşhisini kolaylaştırılmıştır.

Bunlara ek olarak, günümüzde kadınların en büyük korkusu olan meme kanseri için çektirilen insan sağlığı için zararlı ve kanseri tetikleyici etkiye sahip olan mamografi çekiminin azaltılmasına yönelik çalışmalar devam etmektedir. Üretimin bağımlı olduğu değişkenler belirlenerek stok birikimi azaltılmıştır. Üretimdeki zaman ve hammadde kayıpları minimuma indirilmektedir. Tüm bu işlemler farklı VM teknikleriyle yapılmaktadır. Bu çalışmada sınıflandırma teknikleri kullanılarak İnsani Gelişme Endeksinin sınıflandırma başarısı ölçülmüştür.

İnsani Gelişme Endeksi her yıl UNDP tarafından düzenli olarak yayınlanmaktadır. İnsani gelişmenin gelirden farklı olarak eğitim, sağlık gibi unsurların gelişmeye olan katkısıyla hesaplanmaktadır. Gelirden bağımsız olarak hesaplanan bu endeks milli gelirin yüksek, gelişmenin düşük olduğu ülkelerdeki eşitsizliklere dikkat çekmeyi hedeflemiştir. Her yıl farklı bir tema ile yayınlanan bu rapor toplumların gelişmişlik seviyesini belirlemenin yanı sıra küresel ortak sorunlara da dikkat çekerek çözüm sunan politikalar geliştirmektedir. İnsani Gelişme Endeksi 0 ile 1 arasında değerler alır. Değer 1' e yaklaştıkça gelişmişlik seviyesi artar. Çok yüksek, Yüksek ve Orta olarak sınıflar ayrılmıştır.

Bu çalışmada veri madenciliği algoritmalarıyla İnsani Gelişme Endeksi'nin sınıflandırılması yapılarak en yüksek başarıyı veren algoritma tespit edilmiştir. Gelişmişlik seviyesi yerine 1-3 arası kodlama yapılarak 141ülkeninin verileri analize tabi tutulmuştur. Analizde C5.0, CHAID, C&RT(Gini) ve C&RT(Simple) algoritmaları uygulanmıştır. Veriler eğitim seti ve test seti olarak bölünmüştür. Literatürde olduğu gibi bu oran %80-%20 olarak alınmıştır. Analiz sonucunda algoritmaların sınıflandırma başarıları Grafik 1 de gösterilmiştir:



Grafik 1: Sınıflandırma Başarıları

C 5.0 : %95.92

CHAID : %92.79

C&RT(Gini) : %91.22

C&RT(Simple) : %91.22

En yüksek sınıflandırma başarısını C5.0 algoritması %95.92 oranıyla kazanmıştır. Karar ağacı ile farklı kurallar üretilmiştir. Üretilen karar ağacı sayesinde dallanmayı gerçekleştiren nitelikler belirlenerek yorumlanmıştır. Dallanma kriterleri Ek 2 de belirtilmiştir. İnsani Gelişme Endeksinin sınıflandırılmasında "doğumda beklenen yaşam süresi" önemli bir nitelik olarak bulunmuştur. Verilerin algoritmaya uygunluğunu ve testin doğruluğunu ölçmek için korelasyon analizi yapılmıştır. Bu analiz sonucunda dallanma kriterleriyle değişken arasında 1'e yakın değerli negatif ve pozitif korelasyona sahip ilişkiler tespit edilmiştir. Bu niteliklerin aldığı değerlere bakılarak testin doğruluğu ve modelin uygunluğu onaylanmıştır. İnsani Gelişme Endeks değeri, değişkenlerine uygun olarak hesaplamalar yapılarak sınıflandırılmış ve sınıflandırmanın hangi kriterlere dayandığı göstermek amacıyla da karar ağaçları üretmiştir.

Var olan kaynaklarla yakından alakalı olan bu çalışma, bundan sonra bu alanda yapılacak birçok çalışma için kaynak niteliğindedir. Çalışma sınıflandırma başarısı için örnek olarak işlenebilir ve geliştirmeye açıktır. Bu çalışma sonucunda daha az zamanda daha az işlem yükü ile sınıflandırma yapmanın mümkünlüğünü ortaya koymak için yapılmıştır ve analiz sonucu ile desteklenmiştir. Gelecek çalışmalar için farklı sınıflandırma yöntemleri ile bu başarıyı yakalamak için yöntem belirlemek, farklı istatistiksel analizler yapılarak farklı modellerin kurulması önerilebilir.

KAYNAKÇA

Acar Şaylan, Ç. (2013). *Böbrek Nakli Geçirmiş Hastalarda Akıllı Yöntem Tabanlı Yeni Öznitelik Seçme Algoritması Geliştirilmesi*. (Yüksek Lisans Tezi). Kadir Has Üniversitesi/Fen Bilimleri Enstitüsü, İstanbul.

Adak, M. F. ve Yurtay, N. (2013). Gini Algoritmasını Kullanarak Karar Ağacı Oluşturmayı Sağlayan Bir Yazılımın Geliştirilmesi. Bilişim Teknolojileri Dergisi, 6 (3).

Akbulut, S. (2006). *Veri Madenciliği Teknikleri İle Bir Kozmetik Markanın Ayrılan Müşteri Analizi ve Müşteri Segmentasyonu*. (Yüksek Lisans Tezi). Gazi Üniversitesi/Fen Bilimleri Enstitüsü, Ankara.

Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İşletme Fakültesi Dergisi, 29, 1-22.

Argüden, Y. ve Erşahin, B. (2008). *Veri Madenciliği: Veriden Bilgiye, Masraftan Değere.* İstanbul: ARGE Danışmanlık Yayınları No: 10. http://www.arguden.net/wp-content/uploads/2013/02/veri-madenciligi.pdf [Erişim Tarihi: 14 Mayıs 2015].

Atbaş, A. C. G. (2008). *Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma*. (Yüksek Lisans Tezi). Ankara Üniversitesi/Fen Bilimleri Enstitüsü, Ankara.

Atılgan, E. (2011). *Karayollarında Meydana Gelen Trafik Kazalarının Karar Ağaçları Ve Birliktelik Analizi ile İncelenmesi*. (Yayınlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi/Fen Bilimleri Enstitüsü, Ankara.

Bounsaythip C. ve Runsala, E. R. (2001). *Overview of Data Mining for Customer Behavior Modeling.*?

Carus, A. ve Mesut, A. (2005). *Web Kullanım Veri Madenciliği Uygulaması*. II. Mühendislik Bilimleri Genç Araştırmacılar Kongresi MBGAK, İstanbul. http://altanmesut.trakya.edu.tr/pubs/A1-17.pdf [Erişim Tarihi: 14 Mayıs 2015].

Çalışkan, S. K. ve Soğukpınar, İ. (2008) *K Means ve K En Yakın Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti*. http://www.emo.org.tr/ekler/8c1874c96244659 ek.pdf [Erişim Tarihi: 14 Mayıs 2015].

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.*

Jiang, Daxin, Chun Tang, and Aidong Zhang. (2004). *Cluster analysis for gene expression data: A survey*. *Knowledge and Data Engineering, IEEE Transactions on* 16.11

Koyuncugil, A. S. (2006). *Bulanık Veri Madenciliği ve Sermaye Piyasalarına Uygulanması*. (Doktora Tezi). Ankara Üniversitesi/Fen Bilimleri Enstitüsü, Ankara.

Koyuncugil, A. S. ve Özgülbaş, N. (2009). Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. Bilişim Teknolojileri Dergisi, 2 (2), 21.

Küçüksille, E. (2009). *Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İMKB Hisse Senetleri Piyasasında Bir Uygulama.* (Doktora Tezi). Süleyman Demirel Üniversitesi/Sosyal Bilimler Enstitüsü, Isparta.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introductionto Data Mining. John Wiley&Sons,* Inc., New York.

Nas, B. B. (2011). *YSA ve DVM Yöntemleri İle Taşınmaz Değerlemesi İçin Bir Yaklaşım Geliştirme*. (Yüksek Lisans Tezi). Selçuk Üniversitesi/Fen Bilimleri Enstitüsü, Konya.

Oğuzlar, A. (2003). *Veri Önişleme*. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 21, 67-76. http://iibf.erciyes.edu.tr/dergi/sayi21/aoguzlar.pdf [Erişim Tarihi: 14 Mayıs 2015].

Oğuzlar, A. (2004). CART Analizi ile Hanehalkı İşgücü Anketi Sonuçlarının Özetlenmesi. Atatürk Üniversitesi İİBF Dergisi, 18 (3-4), 79-90.

Okafor, A. (2005). *Entropy based techniques with applications in data mining.* Universty of Florida, s:5

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. İstanbul Ticaret Üniversitesi Dergisi, 65-82.

Özkan, Y. (2013). *Veri Madenciliği Yöntemleri.* İstanbul: Papatya.

Öztemel, E. (2003). *Yapay Sinir Ağları*. İstanbul: Papatya.

Öztemel, E. (2012). *Yapay Sinir Ağları.* İstanbul: Papatya.

Padhraic, S. (2001) . *Breaking Out of the Black-Box: Research Challenges in Data Mining*, California.

Silahtaroğlu, G. (2013). *Veri Madenciliği*. İstanbul: Papatya.

Siyambaş, Y. (2014). *Çeliğin Farklı Kesme Şartlarında Delinmesinde Delik Kalitsinin Araştırılması ve Sonuçlarının YSA'da Modellenmesi*. (Yüksek Lisans Tezi). Gazi Üniversitesi/Fen Bilimleri Enstitüsü, Ankara.

Taşdemir, M. (2012). *Veri Madenciliği*. (Yüksek Lisans Tezi). Dicle Üniversitesi/Sosyal Bilimler Enstitüsü, Diyarbakır.

Telcioğlu, M. B. (2007). *Veri Madenciliğinde Genetik Programlama Temelli Yeni Bir Sınıflandırma Yaklaşımı ve Uygulanması*. (Yüksek Lisans Tezi). Erciyes Üniversitesi/Fen Bilimleri Enstitüsü, Kayseri.

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*. Berlin, December 20.

Uçar, A. (2006). *Destek Vektör Makine Tabanlı Bulanık Sistemler, Yeni Bir Gürbüz Sınıflayıcı ve Regresör Tasarımı*. (Doktora Tezi). Fırat Üniversitesi/Fen Bilimleri Enstitüsü, Elazığ.

Yakut, E. (2012). *Veri Madenciliği Tekniklerinden C5.0 Algoritması ve Destek Vektör Makineleri ile Yapay Sinir Ağlarının Sınıflandırma Başarılarının Karşılaştırılması: İmalat Sektöründe Bir Uygulama*. (Doktora Tezi). Atatürk Üniversitesi/Sosyal Bilimler Enstitüsü, Erzurum.

İNTERNET KAYNAKLARI

http://ilkucar.com/AYT/AYT MI 10 Normallestirme.pdf

http://www.cse.buffalo.edu/DBGROUP/bioinformatics/papers/survey.pdf

http://www.norusis.com/pdf/SPC v13.pdf

http://www.tutorialspoint.com/data mining/dm dti.htm

https://filebox.ece.vt.edu/~s14ece6504/projects/baa18 nsharp3 space weather/index.html

https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v4.2.pdf

https://www.academia.edu/6807989/Veri madencili%C4%9Fi nedir

https://www.cs.princeton.edu/courses/archive/spring07/cos424/papers/mitchell-dectrees.pdf

ftp://public.dhe.ibm.com/software/analytics/spss/support/Stats/Docs/Statistics/Algorithms/13.0/TREE-CART.pdf

http://hdr.undp.org

http://www.tr.undp.org/content/dam/turkey/docs/Publications/hdr/2014%20Human%20Development%20Report%20-%20English.pdf

http://en.wikipedia.org/wiki/Human Development Index

http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap2 data.pdf

EKLER

EK 1: Analizde Kullanılan Ülkeler

Norway
Australia
Switzerland
Netherlands
UnitedStates
Germany
NewZealand
Canada
Singapore
Denmark
Ireland
Sweden
Iceland
UnitedKingdom
HongKong,China(SAR)
Korea(Republicof)
Japan
Liechtenstein
Israel
France
Austria
Belgium
Luxembourg
Finland
Slovenia
Italy
Spain
CzechRepublic
Greece
BruneiDarussalam
Qatar
Cyprus
Estonia
SaudiArabia
Lithuania
Poland
Andorra
Slovakia
Malta
UnitedArabEmirates
Chile
Portugal
Hungary

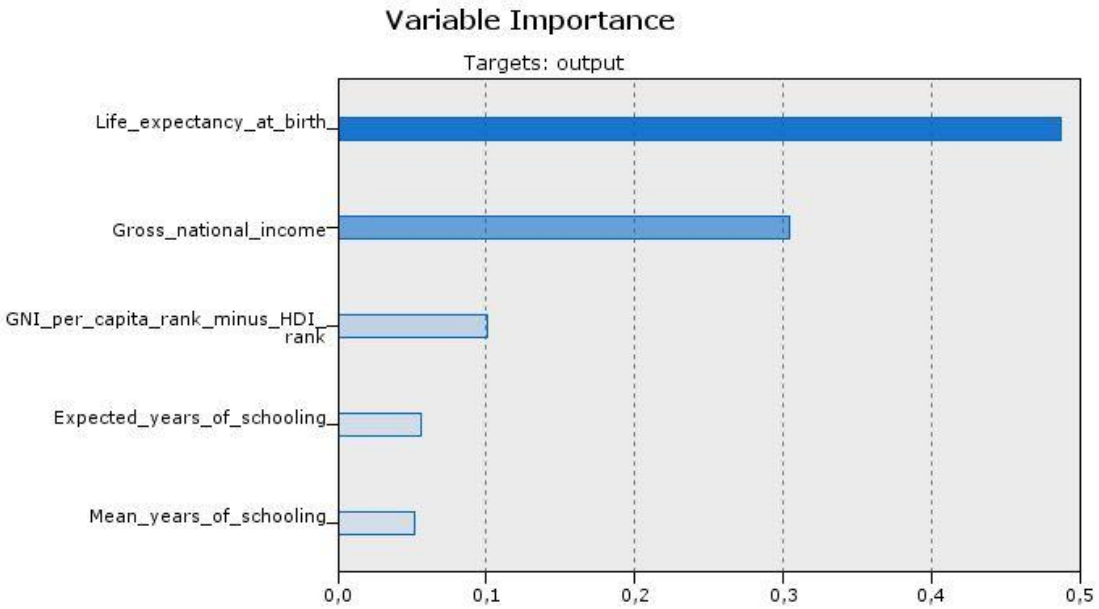
Bahrain
Cuba
Kuwait
Croatia
Latvia
Argentina
Uruguay
Bahamas
Montenegro
Belarus
Romania
Libya
Oman
RussianFederation
Bulgaria
Barbados
Palau
AntiguaandBarbuda
Malaysia
Mauritius
TrinidadandTobago
Lebanon
Panama
Venezuela(BolivarianRepublicof)
CostaRica
Turkey
Kazakhstan
Mexico
Seychelles
SaintKittsandNevis
SriLanka
Iran(IslamicRepublicof)
Azerbaijan
Jordan
Serbia
Brazil
Georgia
Grenada
Peru
Ukraine
Belize
TheformerYugoslavRepublicofMacedonia
BosniaandHerzegovina
Armenia
Fiji
Thailand

Tunisia
China
SaintVincentandtheGrenadines
Algeria
Dominica
Albania
Jamaica
SaintLucia
Colombia
Ecuador
Suriname
Tonga
DominicanRepublic
Maldives
Mongolia
Turkmenistan
Samoa
Palestine,Stateof
Indonesia
Botswana
Egypt
Paraguay
Gabon
Bolivia(PlurinationalStateof)
Moldova(Republicof)
ElSalvador
Uzbekistan
Philippines
SouthAfrica
SyrianArabRepublic
Iraq
Guyana
VietNam
CapeVerde
Micronesia(FederatedStatesof)
Guatemala
Kyrgyzstan
Namibia
TimorLeste
Honduras
Morocco
Vanuatu
Nicaragua
Kiribati
Tajikistan
India

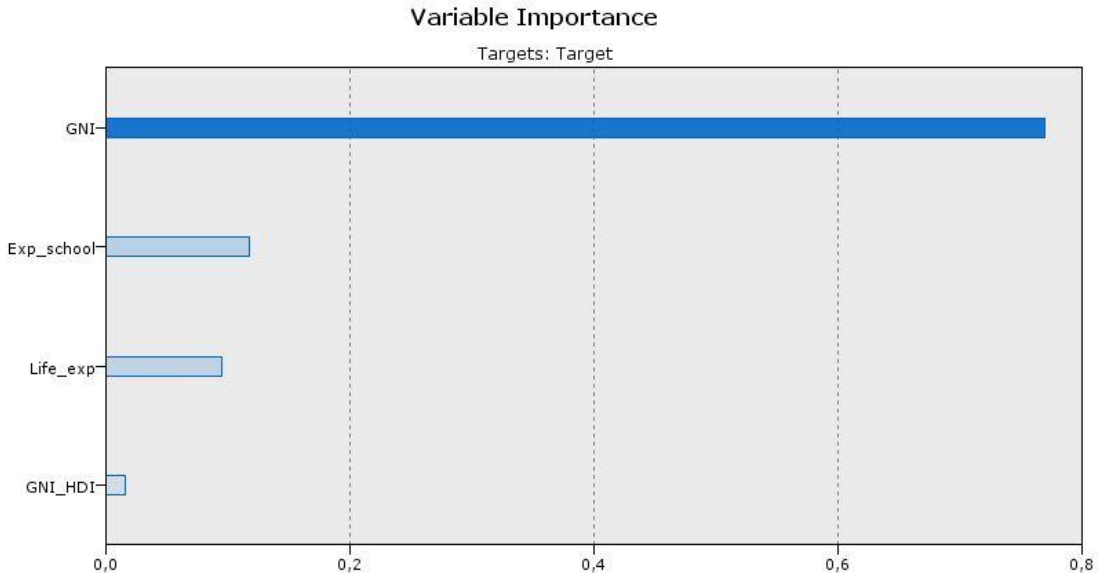
Bhutan
Cambodia
Ghana
LaoPeople'sDemocraticRepublic
Congo
Zambia

EK 2: Önem Derecesine Göre Bölünme Kriterleri

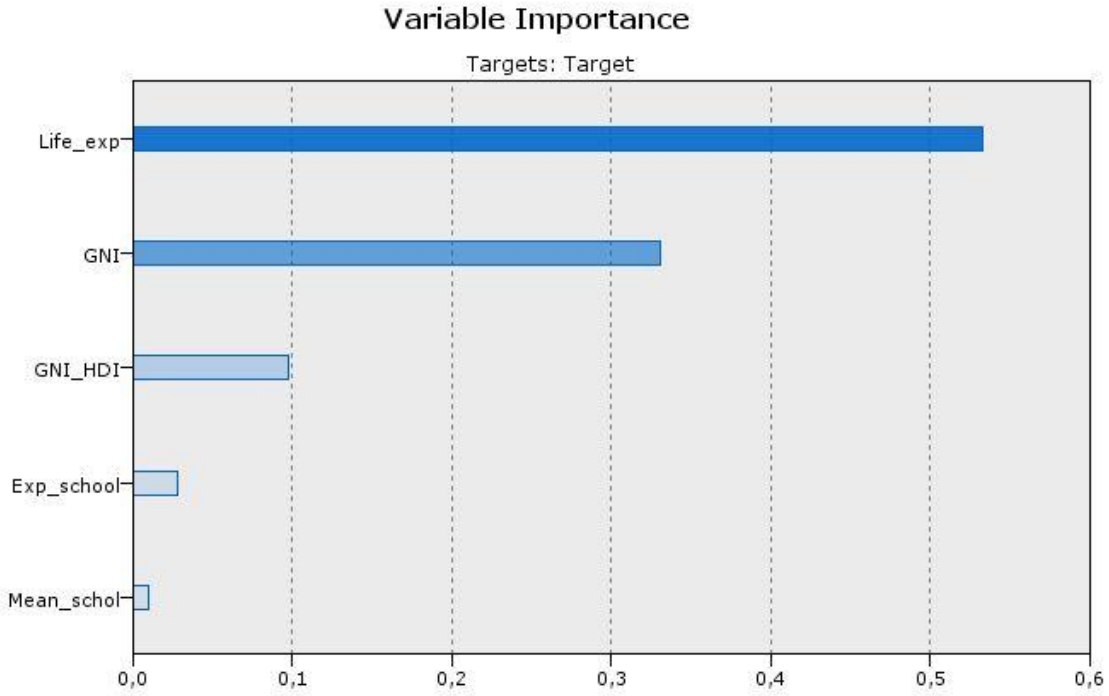
C 5.0 Algoritması



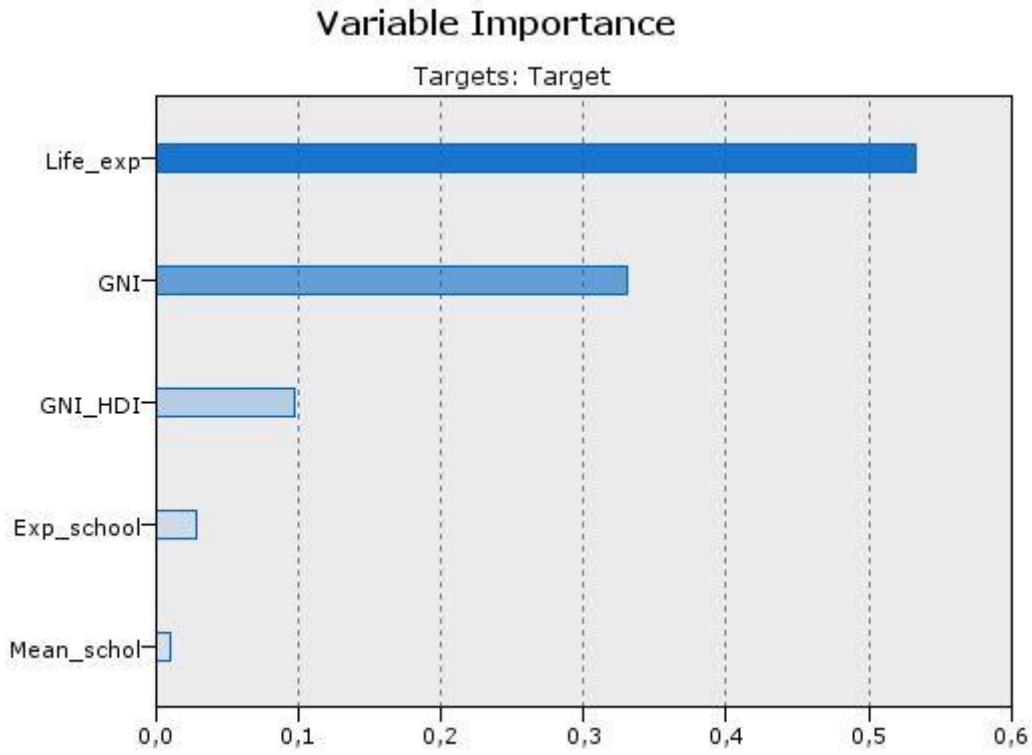
CHAID Algoritması



C&RT Algoritması



GİNİ Algoritması



ÖZGEÇMİŞ

Adı Soyadı : Ayşe YILDIZ

Doğum Tarihi : 1989

E-Posta Adresi : ayseyildizybs@windowslive.com

Öğrenim Durumu:

Derece	Bölüm/Program	Üniversite	Bitirme Yılı
Lise	Eşit Ağırlık	Toroslar Lisesi	2006
Lisans	Yönetim Bilişim Sistemleri	Bartın Üniversitesi	2013
Yüksek Lisans	Yönetim Bilişim Sistemleri	Osmaniye Korkut Ata Üniversitesi	2015