ETSO KAYITLI İŞLETMELERİN İŞÇİ İSTİHDAMLARININ ELAZIĞ EKONOMİSİNE ETKİLERİNİN CHAID ANALİZİ İLE İNCELENMESİ

Selman AKTAŞ

İstatistik Anabilim Dalı

Yrd. Doç. Dr. Nurhan HALİSDEMİR

OCAK-2017

T.C

FIRAT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ETSO KAYITLI İŞLETMELERİN İŞÇİ İSTİHDAMLARININ ELAZIĞ EKONOMİSİNE ETKİLERİNİN CHAID ANALİZİ İLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

Selman AKTAŞ

122133104

Tezin Enstitüye Veriliş Tarihi:

Tezin Savunulduğu Tarih: 30.01.2017

Tez Danışmanı : **Yrd. Doç. Dr. Nurhan HALİSDEMİR (F.Ü)**

Diğer Jüri Üyeleri : **Doç. Dr. Mehmet GÜRCAN (F.Ü)**
Doç. Dr. Kamil ALAKUŞ (OMÜ)

OCAK-2017

ÖNSÖZ

Çalışmamda bana en büyük katkıyı sağlayan ve sürekli olarak akademik anlamda desteğini hiç eksik etmeyen danışman hocam Sayın Yrd. Doç. Dr. Nurhan HALİSDEMİR' e ve Doç. Dr. Mehmet GÜRCAN' a ve Doç. Dr. Sinan ÇALIK' a teşekkür ederim.

Selman AKTAŞ

Elazığ-2017

İÇİNDEKİLER

ÖZET

Bu çalışmada veri madenciliği modellerinden sınıflama modelleri içerisinde ki - kare ağaçlarından biri olan CHAID analizi kullanılmıştır. Çalışmada 2014-2015 yılları arasında Elazığ Ticaret ve Sanayi Odasına kayıtlı olan işletmelerin işçi istihdamlarının Elazığ ekonomisine olan etkisi CHAID analizi ile inceleme sonucunda işletmelerin temizlik giderleri masraflarının etkisi görülmüştür

Anahtar kelimeler: CHAİD analizi

SUMMARY

INVESTIGATION OF THE EFFECTS OF EMPLOYED EMPLOYMENT REGULATIONS OF WORKERS ON ELAZIĞ ECONOMY BY CHAID ANALYSIS

In this study, CHAID analysis which is one of chi – square trees in classification models which is one of data mining models is used. The effect of the employment of the workers registered in the Elazığ Chamber of Commerce and Industry between years: 2014 and 2015 on the Elazığ economy during the period of 2014-2015 was analysed by the CHAID analysis and the effect of cleaning costs of the enterprises

Keywords: CHAİD analysis

ŞEKİLLER LİSTESİ

TABLO LİSTESİ

1. GİRİŞ

Bilgisayar teknolojisindeki yenilikler ile birlikte donanımların çoğalması sonucu fiyatlarının ucuzlaması verilerin uzun süre depolanmasına büyük kapasiteli veri tabanların oluşmasını sağlamıştır. Bu nedenle büyük veri tabanları istenilen bilgiye ulaşılmasını detaylı incelemeler ve araştırılmalar yapılmasını sağlayarak farklı metotlar getirmiştir. Veriler çeşitli istatistiksel metotlarla analizler yaparak kuruluşların, kurumların ve fertlerin yeni ve farklı stratejiler geliştirmesine karar aşamasında etkili sonuçlar üretmesine destek olmuş ve hatta karar aşamasında otorite haline gelmiştir.

19. yy 'da Osmanlı devleti ilk olarak ticaret, sanayi ve ziraat odaları Mehmet Said Paşa tarafından başlatıldı ve II. Abdülhamit döneminde resmen kuruldu Cumhuriyet döneminde Odalar ve Borsalar 5590 sayılı Kanun 8 Mart 1950 tarihinde hazırlanmış, 15 Mart 1950 tarihinde 7457 sayılı Resmi Gazete'de yayınlanarak yürürlüğe girmiştir. Ticaret ve sanayi odalarının görevleri çıkardıkları yasayla şu şekildedir;

Türkiye Odalar ve Borsalar Birliği ile Odalar ve Borsalar Kanunu (5174 Sayılı Yasa) l. Maddesi "Bu Kanunun amacı; ticaret ve sanayi odaları, ticaret odaları, sanayi odaları, deniz ticaret odaları, ticaret borsaları ile Türkiye Odalar ve Borsalar Birliğinin kuruluş ve işleyişine ilişkin esasları düzenlemektir." demektedir. 5174 Sayılı Yasanın 12. Maddesine Göre; Odaların görevleri şunlardır:

a) Meslek ahlâkını, disiplini ve dayanışmayı korumak ve geliştirmek, ticaret ve sanayinin kamu yararına uygun olarak gelişmesine çalışmak.

b) Ticaret ve sanayiyi ilgilendiren bilgi ve haberleri derleyerek ilgililere ulaştırmak, ilgili kanunlar çerçevesinde resmî makamlarca istenecek bilgileri vermek ve özellikle üyelerinin mesleklerini icrada ihtiyaç duyabilecekleri her çeşit bilgiyi, başvuruları durumunda kendilerine vermek veya bunların elde edilmesini kolaylaştırmak, elektronik ticaret ve internet ağları konusunda üyelerine yol gösterecek girişimlerde bulunmak, bu konularda gerekli alt yapıyı kurmak ve işletmek.

c) Ticaret ve sanayiye ait her türlü incelemeleri yapmak, bölgeleri içindeki iktisadî, ticarî ve sınaî faaliyetlere ait endeks ve istatistikleri tutmak, başlıca maddelerin piyasa fiyatlarını takip ve kaydetmek ve bunları uygun vasıtalarla yaymak.

d) 26 ıncı maddedeki belgeleri düzenlemek ve onaylamak.

e) Meslek faaliyetlerine ait konularda resmî makamlara teklif, dilek ve başvurularda bulunmak; üyelerinin tamamının veya bir kesiminin meslekî menfaati olduğu takdirde meclis kararı ile bu üyeleri adına veya kendi adına dava açmak.

f) Çalışma alanları içindeki ticarî ve sınaî örf, adet ve teamülleri tespit etmek, Bakanlığın onayına sunmak ve ilân etmek.

g) Üyeleri tarafından uyulması zorunlu meslekî karar almak.

h) Yurt içi ve yurt dışı fuar ve sergilere katılmak.

ı) Gerektiğinde 507 sayılı Esnaf ve Küçük Sanatkârlar Kanununun 125 inci maddesinde sayılan mal ve hizmetlerin azamî fiyat tarifelerini, kendi üyeleri için, Bakanlıkça çıkarılacak yönetmeliğe uygun olarak tespit etmek ve onaylamak.

j) Deniz ticaretinin kamu yararına, millî ulaştırma ve deniz ticareti politikasına uygun şekilde gelişmesine çalışmak.

k) Millî ve milletlerarası deniz ticaretine ait incelemeler yapmak ve bu konudaki bilgileri sağlamak, Türkiye limanları arası ve yurt dışı navlun, acente komisyonu ve ücretleri ile liman masrafları gibi bilgileri toplamak ve bunları mümkün olan vasıtalarla en seri şekilde yaymak, dünya deniz ticaretindeki en son gelişmeleri izlemek, istatistikler tutmak ve bunları ilgililere duyurmak.

l) Deniz ticaretine ait ticarî örf, teamül ve uygulamaları tespit ve ilân etmek, navlun anlaşmaları, konşimento ve benzeri evraka ait tip formları hazırlamak.

m) Yabancı gemi sahip ve donatanları ile denizcilikle ilgili müesseselere Türkiye limanlarının imkânları, çalışma şekilleri, tarifeleri ve liman masrafları hakkında bilgi vermek ve onlardan benzeri bilgileri sağlamak.

n) Deniz ticareti ile ilgili milletlerarası kuruluşlara üye olmak ve delege bulundurmak.

o) İlgililerin talebi üzerine deniz ticareti ile ilgili ihtilaflarda hakemlik yapmak.

p) Deniz acenteliği hizmet ücret tarifelerini hazırlamak ve Bakanlığın onayına sunmak.

r) Sair mevzuatın verdiği görevlerle, ilgili kanunlar çerçevesinde Birlik ve Bakanlıkça verilecek görevleri yapmak.

s) Birliğin belirlediği standartlara göre üye kayıtlarını tutmak ve üyelik aidatlarına ilişkin belgeleri saklamak ve bunları Birliğe talep halinde bildirmek.

t) Mevzuatla bakanlıklara veya diğer kamu kurum ve kuruluşlarına verilen işlerin, bu Kanunda belirtilen kuruluş amaçları ve görev alanı çerçevesinde odalara tevdii halinde bu işleri yürütmek.

u) Üyelerinin ihtiyacı olan belgeleri vermek ve bunlara ilişkin gerekli hizmetleri yapmak.

v) Yurt içi fuarlar konusunda yapılacak müracaatları değerlendirip Birliğe teklifte bulunmak.

y) Üyeleri hakkındaki tüketici şikâyetlerini incelemek ve kuruluş amaçları doğrultusunda diğer faaliyetlerde bulunmak.

z) Ticaret ve sanayi odalarınca, odalar ayrı olan illerde ise sanayi odalarınca sanayiciler için kapasite raporları düzenlemek. Odalar, bunlardan başka mevzuat hükümleri çerçevesinde;

i. Ticaret mallarının niteliklerinin belirlenmesine yönelik laboratuvarlar kurmak veya bunlara iştirak etmek, uluslararası kalibrasyon, test ölçme laboratuvarı kurmak veya iştirak etmek, belgelendirme hizmetleri sunmak,

ii. Milli Eğitim Bakanlığının izin ve denetiminde ticaret, denizcilik ve sanayi ile ilgili kursları açmak, açılan kurslara yardımda bulunmak, yurt içinde ve dışında ihtiyaç duyulan alanlar için öğrenci okutmak ve stajyer bulundurmak; meslekî ve teknik eğitim ve öğretimi geliştirme ve yönlendirme çalışmaları yapmak, kendi üyelerinin işyerleriyle sınırlı olmak üzere, 3308 sayılı Meslekî Eğitim Kanunu kapsamına alınmamış meslek dallarında bu uygulamaya ilişkin belgeleri düzenlemek,

iii. İlgililerin talebi halinde, ticarî ve sınaî ihtilaflarda hakem olmak, tahkim kurulları oluşturmak,

iv. Açılmış veya açılacak olan sergiler, panayırlar, umumi mağazalar, depolar, müzeler ve

Belirtilen maddeler ile kurumun bulunduğu ili daha sonra bölgesini ve daha sonra ülke genelinde işletmelere öncülük edip ekonomiyi hareketlendirmek, istihdam sağlamak, işletmelere rehberlik ederek ticari faaliyetlerini güçlendirmek amaçlarındadır, kuruma bağlı işletmeler deki ekonomik hareketler işçi istihdamı, ham madde temini v.b girdiler ekonomik döngüyü oluşturmaktadır. Bir ildeki ekonomik yapıyı incelemek, araştırma yapmak, üretkenlik adına stratejiler geliştirebilmek için gerekli bilginin en sağlıklı ve en güvenilir şekilde temin edilebileceği yerler Ticaret ve Sanayi Odalarıdır.

Çalışmamızda Elazığ ili Ticaret ve Sanayi Odasının il deki ekonomik gidişata nasıl bir katkı sağladığını inceledik.

1.1. Veri Madenciliği

1.1.1.Veri Madenciliğinin Tarihçesi

Veri Madenciliği, uzun bir geçmişi olan teknoloji alanının evrim geçirmiş şekli olarak tanımlanabilir. İlk bilgisayar ABD'li John Mauchly ve J. Presper Eckert tarafından 1946 yılında ABD ordusuna hizmet için II. Dünya savaşında üretilmiştir. Ve ENIAC (Electrical Numerical Integrator And Calculator) adı verilmiştir. Bu ilk sayısal bilgisayar olan ENIAC kişisel bilgisayarların doğuşu olarak da adlandırılabilinir ve 30 tonluk ağırlığında 170 m²lik alanı kaplamaktaydı. Bilgisayarlar 70 yıllık bir sürede şuan kullandığımız masaüstü boyutlarına gelmiştir. Sayısal bilgisayar olan ENIAC' ın ortaya çıkışı ile birlikte, bilim insanları;

1950'li yıllarda mantık ve bilgisayar bilimleri alanlarında çalışarak, yapay zekâ (Artificial intelligence) ve makine öğrenme (Machine Learning) kavramlarının ortaya çıkmasına neden olmuştur. 1960lı yıllarda istatistikçiler yeni bir algoritma geliştirmişlerdir. Örneğin; regresyon analizi (regression analysis), en çok olabilirlik kestirim (maximum like lihoodestimates), sinir ağları (neural Networks) vb. gibi metotlar Veri Madenciliği'nin ilk adımlarını oluşturmuştur. Ayrıca veri tabanı sistemlerinin gelişmesi ile birlikte çok büyük ebatlarda metin evraklarının, korunmasını ve bilgilerin tekrar kullanabilmesini sağlamıştır.

1970, 1980, 1990lı yıllarda gelişen bilgisayarlar yeni programlama dillerinin oluşmasını sağlamıştır bu gelişmeler ile "genetic algorithms", "EM algorithms", "K-meansclustering" ve "decision tree algorithms" gibi algoritmaların da oluşmasını sağlanmıştır.

Veri ambarı, büyük veri tabanlarının oluşmasıyla ve bilgi keşfinin ilk adımlarının oluşmasıyla meydana gelmiştir. Bununla birlikte gelişen teknoloji ve eklenen yenilikler Veri Madenciliğini değiştirerek bilgisayar teknolojisinde yaygın kullanılan temel bir iş haline getirmiştir.

Veri madenciliği ile ilgili yapılan ilk uygulamanın ise Pazar Sepet Analizi olduğu bilinmektedir. Bazı kaynaklar Veri Madenciliği'ni ilk olarak 1980 yılında Londra'da John Graunt adında bir kişinin Doğal ve Siyasi söylemler Mortalite Bonosu çalışması üzerine

uyguladığını söylemektedir. John Graunt'uno yıllarda ölüm verileri ile ilgili ayrıntılı bir analiz yaparak, model oluşturup bu konuda çalıştığını ve daha sonra da bu modele şehirlerdeki veba hastalığından ölenlerin verilerini kullanarak tahminlerde bulunduğu söylenmektedir.

1.1.2. İstatistik ve Veri Madenciliği Arasındaki İlişki

İstatistiksel metotların verilerin indirgenmesi ve modellenmesi gibi temel veri ön işleme aşamalarında ve çıktıların değerlendirilmesi veya yorumlanmasında yararları çok fazladır.

Veri madenciliği istatistik ile birçok yönden çok yakın ilişki içindedir (Zhao ve Luan 2006). Veri madenciliği ile istatistiğin ortak özelliği "veriden öğrenilmesi"(Ganesh, 2002) veya "verinin bilgiye dönüştürülmesidir (Kuonen, 2004). Bu iki düşünce verilerin anlamını ne olduğunu bilmek içindir. Belirsizlikleri ortadan kaldırmak ve konular hakkında bilgiler ve öneriler vermek adına oluşturulmuştur. İstatistik ve Veri madenciliği bir olaya etki eden elemanları tespit etmeyi konu edinmiştir bunun yanında oluşan sonuçlar ile gelecekte meydana gelebilecek olayları anlayıp önlem alıp yada daha iyi sonuçlara ulaştırmayı konu edinmiştir.

Veri madenciliğinin istatistik ile benzer taraflarının olmasına rağmen farklı yönleri de azımsanmayacak kadar fazladır. Zhao ve Luan dört temel farklılıktan bahsetmektedir.

Tablo 1.1. İstatistiksel analiz ve Veri Madenciliğinin farklılıkları

İstatistiksel Analiz	***Veri Madenciliği***
İstatistikçiler genellikle bir hipotez ile başlarlar.	Veri madenciliği hipoteze gerek duymaz.
İstatistikçiler hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorundadırlar.	Veri madenciliği algoritmaları eşitlikleri otomatik olarak geliştirir.
İstatistiksel analizler sadece sayısal verileri kullanır.	Veri madenciliği farklı tiplerde data kullanır (örneğin metin, ses) sadece sayısal veriyi değil.
İstatistikçiler kirli veriyi analizleri sırasında bulur ve filtre ederler.	Veri madenciliği temiz veriye dayanır.

Bunlar; Tablo 1.1'de verildiği gibi teorinin rolü, genellenebilirlik, hipotez testi ve güven düzeyidir. İstatistik teori ve hipotezler olmadan başlayamaz. Teorinin kılavuzluğu olmadan tüm olayların ve gözlemlerin sonucu yanlış olabilir. İstatistiksel analizler önceden verilen bilgiye bağlı teori ile başlar ve teorinin kabul edilmesi veya ret edilmesi durumlarının neden olduğunu sorgular. İstatistik yapısal olarak bilgiyi doğrulamayla ilgilenir. Veri madenciliği teorinin doğrulamayı amaçlamaz. Bu bilgisayarın otomatik olarak örüntüleri bulacağı veya öngörü yapacağı anlamına gelmez. Aksine, veri madenciliğinet ve belirgin komutlar ister. Veri madenciliği İstatistik ile karşılaştırıldığında değişkenler arası ilişkileri en az varsayımlar ile inceler.

İstatistik tüme varımı konu alırken, Veri Madenciliği tümden gelim'i kullanır. İstatistiksel araştırmada örneklemin kendisiyle işlem yapılması durumu yok denecek kadar azdır. Analistler seçilen örneklem kümesi ana kütle ile ilgili sonuçlar çıkarmayı hedefler. İstatistik bireyselliği ortadan kaldırmayı hedeflemez, amaç olarak konu olan şeylerin benzerliklerin bulunmasına çalışır. Veri madenciliği ise kapsamlı bilgilerin, yerel bilgilerin ya da kişileştirilmiş bilginin peşindedir.

Hipotezin veri madenciliği için mana ifade etmez; çünkü teori veya hipotez genelleştirilmiş bilgilerdir Veri Madenciliği ise ana kütlenin özelleştirilmiş hali ile ilgilenir. Veri madenciliğinde çok büyük veri setleri kullanılır bu sebeple İstatistikte ki anlamlılık düzeyi de çıkarım yapmaktaki önemini kaybeder.

Diğer bazı önemli farklılıklar şöyledir:

- İstatistiksel araştırmalarda veriler akıldaki belirli sorular için toplanır ve bu sorulara yanıt bulmak için analiz edilir. İstatistiksel deney tasarımı ve alan araştırması gibi alt disiplinler veri toplamak için en iyi yollarla ilgili ipuçlarını sağlarlar. Veri madenciliğinde ise veriler, veri madenciliği uygulamak için değil diğer bazı amaçlar için kullanılır (Oğuzlar, 2003).
- Klasik istatistiksel uygulamalar ve veri madenciliği arasındaki en temel farklılık, veri kümesinin büyüklüğüdür (Oğuzlar, 2004).İstatistikte kullanılan veri kümesindeki veri sayısı binler de iken Veri Madenciliğinde milyarlarca veri kullanılır.
- Veri madenciliği uygulamalarında veri kümesinin önceden işlemler geçirmesi veri kalitesi için oldukça gereklidir. Önceden işlenmiş veri nitelikli yani kaliteli

veri olur buda doğru sonuçların elde edilmesini sağlayacaktır. Fakat, veri madenciliği yapılan analizlerinde verinin bir ön işlemden geçmesi büyük zaman sorunudur çünkü veri miktarı milyon veya milyarlarca adettir. İstatistiki çalışmalarda veri ön işlemesi yapılmadan direkt analize geçilir ve veri adetinin Veri Madenciliğine oranla düşük olmasına rağmen buda ikisi arsında ki farklardan biridir.

1.1.3.Veri Madenciliğinin Tanımı

Veri madenciliğinin bir çok farklı tanımı vardır. Bunlar;

- Veri madenciliği büyük veri tabanlarındaki gizli bilgi ve yapıyı açıklamak için, çok sayıda veri analizi aracını kullanan bir süreçtir (Oğuzlar, 2004).
- Kuonen (2004) veri madenciliğini iş kararlarının alınabileceği doğru, alışılmamış, faydalı ve anlaşılabilir örüntüler veya modeller olarak tanımlamaktadır.
- Veri madenciliği büyük hacimli verilerden öz bilgi'nin çıkarılması sürecidir (Ganesh 2002). Bir başka ifade ile veri madenciliği büyük ve karmaşık verilerde beklenmeyen patikaların, değerli yapıların ve ilginç ilişkilerin keşfedilmesi bilimidir.
- Veri madenciliği genel anlamda, büyük miktarda veri içerisinden, gizli kalmış, değerli, kullanılabilir bilgilerin açığa çıkarılmasıdır (Koyuncugil, 2007).
- Veri madenciliği ve öz bilgi keşfi, verilerde daha önceden bilinmeyen, anlamlı ve değerli bilgiler elde etme işlemidir (Yıldırım, Uludağ ve Görür, 2007).
- Bilgisayar teknolojilerinin sağlamış olduğu çok hızlı veri işleme ve yüksek hacimde veri depolama imkânları yardımıyla ve farklı disiplinlerin katkısıyla sağlanan araçlarla, sahip olunan çok büyük hacimlerdeki veriden, karar vericinin etkin ve daha fazla bilgiye dayalı karar vermesinde kullanabilmesi amacıyla önceden bilinmeyen, gizli, örtük, klasik metotlarla ortaya çıkarılması güç, faydalı, ilginç, anlaşılabilir; ilişki, örüntü, bağıntı veya trendlerin otomatik veya yarı otomatik bir şekilde ortaya çıkarılması olarak tanımlanır (Şentürk, 2006).

Genel anlamda veri madenciliği verilerin farklı bir gözle inceleyip analiz yaparak kullanışlı bilgiyi özet haline getirerek inceleyen süreçtir. Veri madenciliği, birbiriyle ilişkili çok büyük veri tabanlarının incelenip bunlar arasındaki düzeni ve ilişki bulma işidir. Veriler üzerinde çözümlemeler yapmak ve veriyi çözüp bilgiyi elde etmek için meydana

getirilmiş bir yöntemdir. Veri madenciliği İstatistiki programlarıyla meydana gelmiş yada İstatistikte ki gibi sorgulama ve doğrulama süreci değildir. Veri madenciliğinin veri hacmi çok büyüktür aynı şekilde değişken sayısı da bir hayli fazladır**.** Dünya da her geçen gün teknolojik açıdan gelişmeler ve ilerlemeler görülüyor, bu gelişmeler doğru sonuçlara ulaşabilmek çıkarımlar yapabilmek için elektronik ortamlarda saklanıyor ve korunuyor. Bu işlemin çokluğu bilgilerin kayıtlar altına alınmasını, korunmasını sağlıyor ve bir rutin haline geldiği için önceki dönemlere oranla daha uygun maliyetler sağlıyor. Bu nedenlerle veri tabanlarında oluşan zenginlik sayesinde kayıtlı olaylar ışında analizler yapılarak geleceğe dair çıkarımlar ve yapılan işlerde kolaylıklar sağlamak adına tahminlerde bulunulabiliyor.

1.1.4.Veri Madenciliğinin Amacı

Günlük yaşamda bilgisayarların hayatımıza girmesi ile birlikte yapılan her işlem sayısal ortamda kayıt altına alınmaktadır. Örneğin market alışverişlerinde alınan veya iade edilen her bir ürünün manyetik ortamda yapılan giriş çıkış işlemi, benzer hastanelerdeki hasta kayıtları, sinema ve devlet dairelerindeki kayıtlar, yollarda bulunan kameraların kayıtları, yapılan telefon görüşmeleri gibi kısacası her yerde, her yapılan işlem bir veri topluluğu meydana getirmekte ve verinin olduğu yerde de veri tabanı meydana gelmektedir.

Veri madenciliği sayesinde veri tabanlarının içindeki bilgiler kolaylıkla çekilip kullanılabiliyor. Bilgiyi çekip kullanma işlemi matematik modeller, istatistik ve birçok bilgisayar programı ile yapılabiliyor. Çok büyük miktarlarda verilerin incelenmesi veri madenciliğinin konusu olduğundan veri madenciliği veri tabanları ile ilişki içindedir. Birçok veri tabanının birleştirilip bundan kolayca faydalanılacak bir özet oluşturulması veri ambarı mantığını oluşturmuş, bu işleyiş günümüz de yaygın olarak kullanılıp gelişmeler göstermektedir. Böylelikle bilgiye kolay ve hızlı bir şekilde ulaşılmasını sağlamıştır veri ambarları.

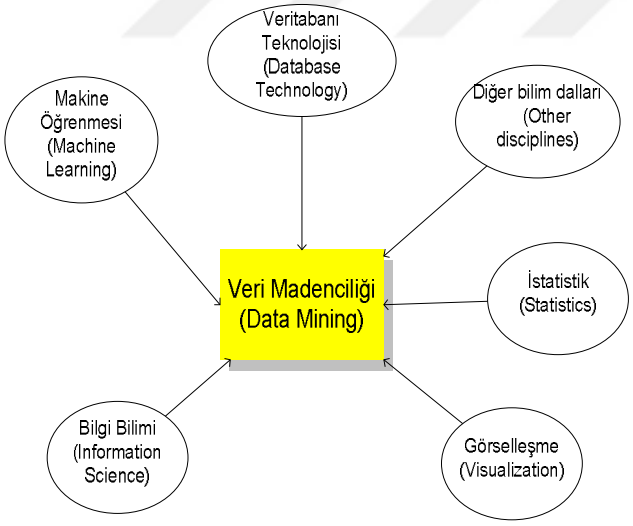
Gelişen teknoloji ile birlikte, veriye erişim kolay olmasına karşın milyonlar hatta milyarlar seviyesinde veri kaydının olduğu bir veri tabanında bu verilerden birtakım çıkarsamalar yaparak bilgiye ulaşmak zorlamış ve klasik istatistiksel yöntemler bu verilerden bilgi çıkarmak için yeterli gelmemeye başladığından, verileri analiz edebilmek için yeni tekniklere ihtiyaç duyulmuştur. 2001 yılında en büyük işleme sahip veri tabanları

318 terabyte iken, bu rakam 2003 yılında1029 terabytelara çıkarak, iki yıl içerisinde müthiş bir veri artışı olmuş ve halen olmaya da devam etmektedir. Bütün bu veri yığınları içerisinde altın değerinde keşfedilmeyi bekleyen bilgiler bulunmaktadır.

Kurumların kadrolarında görev yapan üst yöneticilerin, bu dev boyuttaki veri yığınları içerisinden bilgiye ulaşarak ileriye dönük, kurumun iyileştirilmesine yönelik kararlar almaları bilgiye erişemedikleri sürece imkânsız hale gelmektedir. Bu verilerin üst yönetime bilgi olarak dönmesi için, bu verilerin uygun yazılımlar aracılığı ile bir takım işlemlerden geçirilerek bilgiye dönüştürülmesi ve üst yönetime sunulması gerekir. Bu bilgilere klasik istatistiksel yöntemlerle erişmek de artık çok zorlaştığından, yeni bir takım tekniklere ihtiyaç duyulmuş ve Veri Madenciliği kavramı ortaya çıkmıştır.

1.1.5.Veri Madenciliği Literatür Araştırması

Veri madenciliği aşağıdaki şekil 1.1'de şemalandırıldığı gibi birçok disiplin içeren bir alandır. Bunlar; veritabanı teknolojileri, istatistik, makine öğrenme, görselleşme ve diğerleridir.



Şekil1.1. Veri Madenciliğinin birleştiği dallar

Veri madenciliği bankacılık, pazarlama, sigortacılık, sağlık gibi birçok alanlarda kullanılmaktadır. Sektörlerdeki farklılık gözetilmeden her alanda veri madenciliği kullanılır. Bu ise çok büyük veri ambarlarının oluşmasına fayda ve kaynak sağlamaktadır. Bu dağılım 2003 yılı verilerine göre sektörler itibarıyla tablo 1.2'de özetlenmiştir.

Tablo 1.2. Veri madenciliğinin uygulama alanlarının ana başlıkları

Bankacılık(51)	%12
Biyoteknoloji/Genomik(11)	%3
Kredi Puanlama(35)	%8
CRM(52)	%12
Doğrudan Pazarlama/Fundraising(34)	%8
e-Ticaret(11)	%3
Eğlence/Müzik(4)	%1
Sahtecilik Algılama(31)	%7
Kumar(2)	%0
Hükümet uygulamaları(12)	%3
Sigorta(24)	%6
Yatırım/Hisse(5)	%1
Önemsize-posta/Anti-spam(5)	%1
Sağlık/HR(15)	%4
imalat (19)	%5
Medikal/ilaç (12)	%3
Perakende(25)	%6
Bilim(17)	%4
Güvenlik/Anti-terör(5)	%1
Telekom(23)	%5
Seyahat/Hospitality(8)	%2
Web (9)	%2
Diğer(11)	%3

1.1.6. Veri madenciliğinin Kullanıldığı Alanlar

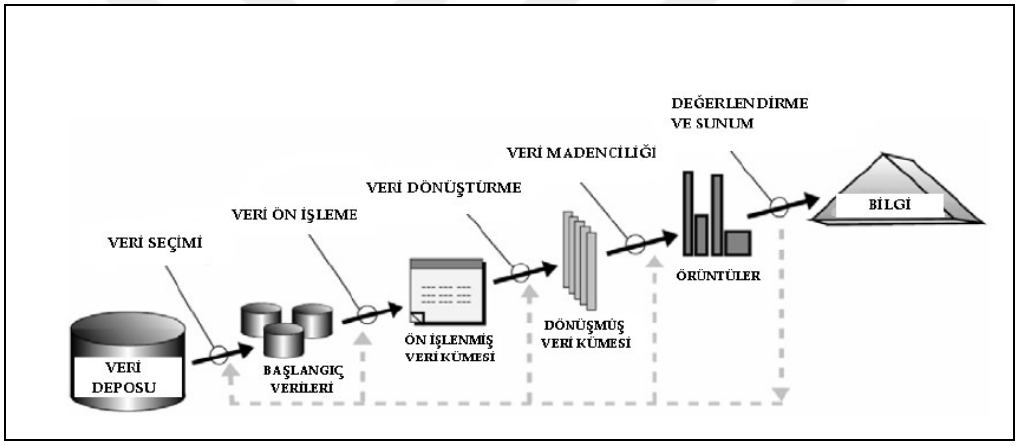
- **Pazarlama Yönetimi;**
  - Pazar sepeti analizi
  - Elde olan müşterilerin korunması, yeni müşterilerin kazanılması
  - Müşterilerin satın alma örüntülerinin belirlenmesi
  - Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması
  - Posta kampanyalarında cevap verme oranının artırılması
  - Risk yönetimi ve dolandırıcılık

- **Diğer Uygulamalar;**
- İşaret işleme
- Biyoloji
- Tıp

1.1.7. Veri Madenciliğinde Bilgi Keşfi Süreç Ve Aşamaları

Veri tabanlarında bilgi süreci:

- Veriyi temizleme
- Veri dönüştürme
- Veri küçültme
- Veri madenciliği



Şekil 1.2. Veri keşif sürecinin aşamaları

1.1.7.1.Veri Temizleme

Pratikte, ortamdaki veri eksik, gürültülü, yanlış veya devamlılık göstermiyor olabilir. Veri temizleme rutinleri, eksik değerleri tamamlamak, gürültüyü veya hatalı verileri düzeltmek için kullanılır:

1. Eksik Değerler

Eksik değerler üzerinde uygulanacak metotlar şunlardır:

a. Eksik değerleri analizden çıkartmak: Bu yöntem, eğer çok fazla nitelikte eksik değerler varsa, verimlilik sağlamaz.

b. Eksik değerleri elle doldurmak: Bu yöntem genelde zaman alıcıdır ve büyük veritabanları için yapılabilir değildir.

c. Eksik değerler için global bir değer belirlemek: Eksik nitelikler için global değişken vermek, analiz programının bu değeri sıklıkla kullanıldığı için bir değer olarak kabul edip hesaplamalara dahil etmesini sağlayabileceğinden yanlış analiz sonuçları ortaya çıkartabilir.

d. Eksik değerler içeren niteliğin ortalama değeri ile veriyi tamamlamak: Genelde, nitelik bir gelir bilgisiyse, ortalama kullanmak doğru sonuçlar verebilir.

e. En olası değer ile eksik veriyi tamamlamak: Bumetot için regresyon, Bayesian-yaklaşımı ya da karar ağaçları kullanılabilir.

2. Gürültülü Değerler

Gürültü, bir değişkendeki rastlantısal hata oranıdır. Verideki gürültüyü yok etmek için aşağıdaki yöntemler kullanılır:

a. Kutulama: Bu yöntemde, veriler, komşu değerlerine göre sıralanırlar. Sıralanmış bu veriler, belirli sayıda "kutulara" konur. "Ortalama değere göre kutulamada her değer, o kutuya dâhil edilen tüm değerlerin ortalama değeri ile yer değiştirir. Aynı şekilde "Medyan değerine göre kutulamada değerler medyanla, "Limitlere göre kutulama'da ise, değerler en yüksek ya da en düşük değere yakınlıklarına göre bu iki değerle yer değiştirirler.

b. Demetleme: Birbirine benzer değerler gruplara yada demetlere bölünerek her bir demetin sınır çizgileri belirlenir.

c. Regresyon: Veri, bir fonksiyona sokularak, o fonksiyon üzerine yerleşmesi sağlanır ve böylece gürültülü değerler otomatik olarak elenmiş olurlar.

1.1.7.2. Veri Dönüştürme

Veri dönüştürmede amaç, veriyi analiz edebilmek için uygun hale getirmektir. Bunun için aşağıdaki yöntemler kullanılabilir:

a. Düzleştirme: Verideki gürültüyü azaltmak için kullanılır. Kutulama, demetleme ve regresyon bu gruba girer.

b. Birleştirme: Veriyi özetleme ve toparlama için kullanılan yöntemdir. Örneğin günlük satış verileri, aylık veya yıllık olarak tutulabilir.

c. Genelleştirme: Düşük seviyeli veriler, daha yüksek seviyelerdekilere dönüştürülerek hiyerarşik hale getirilebilir. Örneğin, sokak adı bilgisi yerine şehir, hatta ülke adı kullanılması, ya da rakamsal yaş değerleri yerine "genç", "orta yaşlı" veya "yaşlı" değerlerinin kullanılması gibi.

d. Normalleştirme: Değişken veriyi aralıklar şeklinde tutmak için kullanılır.

e. Yeni nitelik oluşturma: Veriyi daha iyi analiz edebilmek için, ona yeni nitelikler ve değerler eklenmesi işlemidir.

1.1.7.3. Veri Küçültme

Veriyi analiz ederken bazen çok büyük veriler analizi olumsuz yönde etkileyebilir. Bu sebeple verinin boyutunun küçültülmesi denenebilir. Bunun için en fazla kullanılan yöntemlerden biri Histogramlardır. Histogramlar, veri dağılımlarını yaparken kutulama yöntemini kullanırlar ve veri miktarını verimli olabilecek hale getirerek azaltırlar. Veriyi bölümlerken Histogramlarda çeşitli kurallar kullanılabilir.

a. Eşit genişlikli: Verinin konduğu her kutunun genişliği veya aralık değeri aynıdır.

b. Eşit derinlikli (Eşit boylu) : Her kutudaki frekans dağılımı yaklaşık aynıdır. Yani eşit sayıda veri içerir.

Veri küçültmede kullanılan başka bir yöntem demetlemedir. Demetleme, veriyi gruplara ayırır. Bir grubun içindeki değerler birbirine yakınken, diğer gruplardakilere o kadar uzak olmalıdırlar. Bir demetin kalitesi, onun çapı ile ölçümlenebilir.

Bu çap, bir demetin içindeki iki objenin birbirine olan maksimum uzaklığına eşittir. Veriyi küçültürken kullanılan en son yöntem örneklemedir. Örnekleme, büyük miktardaki verileri çok daha küçük boyutlarda sunmaya olanak sağlar. Çeşitleri şunlardır:

a. Yerine iadesiz rastgele örnekleme: N adet değerden n tanesinin rastgele seçilmesi ile oluşturulur, bu durumda her değerin seçilme olasılığı 1/N'dır ve birbirine eşittir.

b. Yerine iadeli rastgele örnekleme: Bu örnekleme çeşidi, yerine iadesiz örneklemeye benzer ancak bu kez, seçilen her değer, yeniden seçilebilme üzere genel verinin içine yeniden iade edilebilen seçim yapılır.

4. Veri Madenciliği

Bir veri madenciliği operasyonunda farklı veri madenciliği teknikleri kullanılabilir. Her tekniğin kendine göre avantaj ve dezavantajları vardır. Bunların en fazla kullanılanları, karar ağaçları, sinir ağları, demetleme algoritmaları, k-en yakın algoritmaları, genetik algoritmalar, bulanık mantık ve bağ analizi olarak sayılabilir

a. Sinir Ağları

Sinir ağları, çok gelişmiş bir modelleme yöntemidir insan beyninden esinlenerek oluşturulmuştur ve çalışma yapısı buna benzemektedir, analizden kazanım yapılan bilgiyle gözlemlerden çıkacak sonuçlar arasında ilişki kurar. Giriş değerlerinden çeşitli kuralları öğrenirler ve örüntüleri ortaya çıkartarak parametreleri yeni veri üzerinde uygularlar. Sinir ağları, tahminlemelerde, kredi puanlamada ve risk analizinde oldukça faydalıdır.

Bir sinir ağının yapısında girişi, çıkışı ve işlem bölümleri bulunan düğümler bulunur. Bir değeri tahminlemede kullanılmak üzere bir model oluşturabilmek için bu düğümlerin giriş değerleri çeşitli şekillerde kombine edilir. Her bir düğümün değeri, onu besleyen diğer düğümlerin toplam ağırlıklarından yola çıkılarak hesaplanır. Bir modeli oluşturmada önemli olan, doğru sonuçları ortaya koyabilecek uygun bağlantı ağırlıklarını bulabilmektir.

Sinir ağları, bir verideki ilişkileri ve örüntüleri ortaya çıkartmak için kullanılırlar. Bu veri, bir pazar araştırmasının sonuçları ya da farklı koşullarda bulunan bir üretim işleminin ortaya çıkan sonuçları olabilir. Kullanıldığı alan ne olursa olsun, sinir ağları geleneksel yöntemlerin dışında işlemler yaparlar. Geleneksel yöntemde, analizi yapan kişi, bilgisayara durumu ve kuralları tek tek tanıtarak öğretir. Sinir ağlarının bu kodlamaya

ihtiyaçları yoktur. Sadece ham veriyle, ve iyi bir test sürecinden sonra, ortaya tahminleme yapabilecek durumda olan bir analiz programı çıkartırlar.

b. Karar Ağaçları

Karar ağacı, veriyi sınıflandırma ve tahminleme yapmada kullanılan popüler bir veri madenciliği tekniğidir. Karar ağaçlarının sinir ağlarından daha çekici olan tarafı, onlardan farklı olarak ortaya kurallar çıkartabilmeleridir. Bu kurallar, kullanıcıların kolayca anlayabileceği şekilde ifade edilirler, bu da analizi kolay bir hale getirir.

Karar ağaçları, ağaç şeklinde sınıflandırıcı bir yapıdır ve ağaçtaki her düğüm ya biryaprak düğümü ya da karar düğümünü simgeler.

c. K-en yakın komşuluğualgoritmaları

Bu algoritmalar, aslında demetlemenin bir çeşididir. K değeri, komşu olan kayıtların sayısını simgeler. Verilen N adet ilk örnek örüntüye ve bunların doğru sınıflandırılmasına göre, algoritma sınıflandırılmamış olan bir örüntüyü en yakın komşu gruba bağlar. Sınıflandırmanın doğruluğu, k değerinin artmasıyla artış gösterir. Ancak bu yöntem, eğer veri gürültülüyse hatalı sonuçlar ortaya koyar.

Ayrıca, geçmişe yönelik veriye ihtiyaç duyulur.

d. Genetik algoritmalar

Genetik algoritmalar, evrimsel gelişimi taklit ederek çalışırlar. Optimum çözüme mutasyon ve seçme yöntemiyle ulaşılır. Yüksek uygunluğu olan çözümler seçilir ve mutasyona uğratılarak daha yüksek uygunluklu çözümler üretmek üzere yeniden kullanılır. Genelde genetik algoritmalar, iş tarifelerinde veya motor dizaynlarında kullanılırlar.

e. Bulanık mantık (FuzzyLogic)

Genelde kontrol sistemlerinde kullanılırlar. Sabit, değişmeyen veya kesin değişkenler yerine olabilme ihtimali olan değişkenler kullanılan yöntemdir. 0 ve 1 arasında değişkenlik gösteren sonuçlar verir ve bu sonuçlar ile niteliğin netliğe ne kadar yakın yada ne kadar uzak olduğu belirlenir.

f. Bağ analizi (Link Analysis)

Bir verinin içindeki bağları ortaya çıkartan yöntemdir. Genelde ürün ve müşterinin arasındaki bağların ortaya çıkartıldığı market sepeti analizinde, hedefe yönelik pazarlama alanında veya stok fiyatlarının değişimlerinde kullanılır.

g. Diğer Teknikler (OLAP)

OLAP, bilgiyi çoklu boyutta ve hiyerarşide sunabilmek için kullanılır. Örneğin çok büyük boyutlardaki satış istatistik verileri arasında, kullanıcılara hangi ürünlerin satış oranlarının daha fazla değişkenlik gösterdiğini sunabilir. Kullanıcıların, daha önemli olan değişkenlere ve değerlere odaklanmasını sağlar.

OLAP ve veri madenciliği arasındaki en büyük fark, OLAP'ın kullanıcı odaklı olması yani analiz yapan kişinin bir hipotez öne sürerek OLAP programını kullanıp bunu test etmesi; ancak veri madenciliğinde bunun tam tersi olarak, veri madenciliği programının kendisinin bir hipotez oluşturmasıdır. Böylece bir veri madenciliği aracı, kullanıcının büyük bir veri kaynağında görmesinin çok zor olabileceği bir örüntüyü ortaya çıkartabilir. OLAP ise daha çok bütünleştirmeler, hesaplamalar ve sonuç karşılaştırmalarını grafiksel ortamda incelemeye olanak tanır.

1.8. Veri Madenciliği Modelleri

Veri madenciliği kullanıldığı amaç ve alanlara göre değişik modellere ayrılırlar. Bu modeller değer tahmin modeli, kümeleme analiz, bağlantı analizi ve dolandırıcılık tespitidir. Bu modeller literatürde operasyon veya yöntem isimleriyle de kullanılmaktadır. Bu yöntemler uygulamalarda değişik amaçlar için kullanılırken birçok teknik ve algoritmalardan faydalanılır. Bu teknikler tahminleyici, tanımlayıcı ve her iki yaklaşımı da içerenlerdir. Daha modern bir yapılandırma olarak veri madenciliği üç ayrı modelden oluşmaktadır. Bunlar sınıflandırma, kümeleme ve bağlantı analizidir. Bu modellerin her biri için geliştirilmiş teknik ve algoritmalar vardır. Bu yöntemler şu şekildedir;

1.Sınıflandırma Modeli

a. İstatistiğe Dayalı Algoritmalar:

i. Bayes sınıflandırma

ii. Regresyon

iii. CHAID

b. Mesafeye Dayalı Algoritmalar:

*i.*En yakın komşu

*ii.*En küçük mesafe sınıflandırıcısı

c. Karar Ağaçları

*i.*CART

*ii.*ID3

iii. C4.5 ve C5

*iv.*Sprint

d. Genetik Algoritmalar

e. Yapay Sinir Ağları

2.Kümeleme Modeli

a. Hiyerarşik Yöntemler:

i. SLIQ Algoritması

ii. Cure Algoritması

iii. CHAMELEON Algoritması

iv. BIRCH Algoritması

v. CLUDCH Algoritması

b. Bölümlemeli Yöntemler:

i. K-Ortalama Algoritması

ii, PAM Algoritması

iii. CLARA Algoritması

iv. CLARANS Algoritması

c. Yoğunluğa Dayalı Algoritmalar

i. DBSCAN Algoritması

ii. OPTICS Algoritması

iii. DENCLUDE Algoritması

d. Grid Temelli Algoritmalar:

i. STING Algoritması

ii .Dalga Kümeleme

iii. CLIQUE Algoritması

e. Genetik Algoritmalar

f. Yapay Sinir Ağları

3.Bağlantı Analiz Modeli

a.Apriot Algoritması

b.SETM Algoritması

c.Apriori TID Algoritması

d. GRI Algoritması

1.9. Karar Ağaçları

Karar ağaçları sınıflandırma sorularında sürekli ve yaygın kullanılan algoritmalardan biridir. Diğer yöntemlerle kıyaslandığında karar ağaçlarının yapılandırılması ve anlaşılması daha kolay denilebilir (Agrawal, 1993). Sınıflandırma yapılabilmesi için bu teknikte bir ağaç meydana getirilir ve sonra, veritabanındaki her kayıt bu ağaca uygulanır daha sonra kayıt sonuçları incelenir ve bir sınıflandırma yapılabilinir. Ana iki adımdan oluşur: Birincisi ağacın kurulması, ikincisi de verilerin teker teker ağaca uygulanarak sınıflandırmanın gerçekleştirilmesi şeklindedir.

Karar ağaçlarının daha net anlaşılması açısından matematiksel olarak şöyle tanımlanabilir:

$D=(t_1 \ldots t_2)$ bir veritabanı olsun.

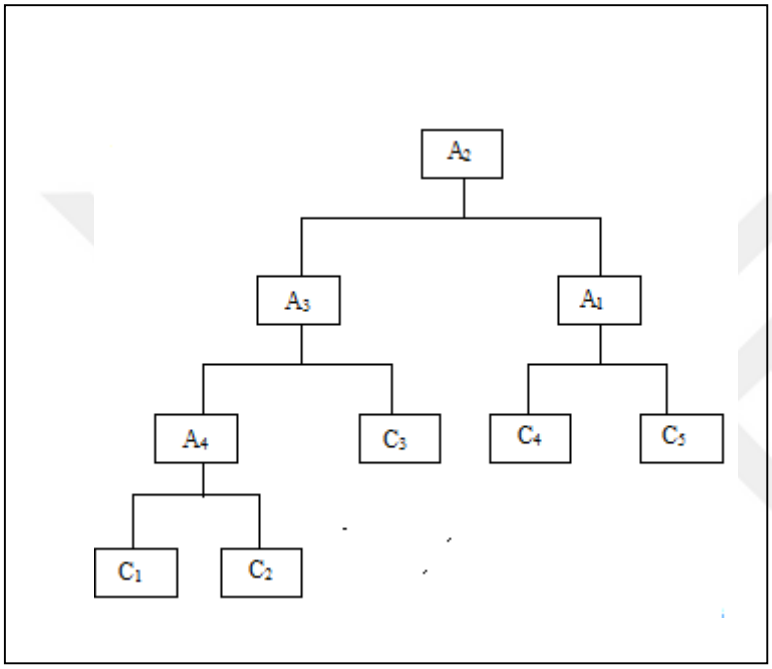
Buradaki her t_i, $t_i=<t_i \ldots t_i>2$ den oluşmaktadır ve bu veritabanı $\{A_1, A_2 \ldots A_n\}$ alanlarından oluşmaktadır.

Bunun dışında, C={C_1 …C_n} kadar da sınıf verilmiş olsun. Bu durumda bir karar ağacı aşağıdaki gibi tanımlanabilir; bir sonraki sayfada da şekilsel olarak gösterilmiştir:

-Her bir düğümü Ai alanıyla isimlendirilmiş

-Her düğümden ayrılan kollar bu alanla ilgili bir soruya cevap veren

-Her yaprağın bir sınıf olduğu bir ağaçtır (Dunham, 2003).



Şekil 1.3.Örnek bir Karar Ağacı

Şekil 1.3 de örnek bir karar ağacı görülmektedir. Ağaçtaki A_1, A_2,…,A_n'dan her biri bir düğümü oluşturmaktadır. Her düğüm kendinken sonra iki dala ayrılmaktadır. Bu ayrılma işlemi sürecinde A_i düğümü hakkında cevabı veritabanında bulunacak bir soru sorulmakta ve verilen yanıta göre bir dal izlenmektedir. Ağaçtaki C_1 …C_n'lerin her biri birer yapraktır ve aynı zamanda bir sınıfı temsil etmektedirler.

Karar Ağaçları kural üretmekle birlikte bazı parametre değerleri de sunarlar. Bunlardan en önemlilerinden bir tanesi kaldıraç (lift) kavramıdır. Kaldıraç, her hangi bir düğümdeki kayıtların, tüm ağaca kıyasla ne kadarının hedef sınıfa ait olduğunu gösteren değerdir. Örneğin; % 255 kaldıraç (lift) değeri o düğümdeki kayıtların 2.55 kat daha fazla oranla belirlenen sınıfa ait olduğunu belirtir.

Kaldıraç hesabını bir örnek üzerinden inceleyelim. Aşağıdaki bir spor salonunu müşteri/üye süreleri ve müşterilerin sistemi terk edip etmedikleri verilmiştir.

Tablo1.3. Üyelik süreleri ve devam tablosu

Üyelik süresi	Terk
1-5	0
6-15	1
6-15	1
1-5	0
6-15	1
6-15	0
1-5	1
1-5	0
1-5	0

Veriler dikkatlice incelendiğinde aşağıdaki iki kural bulunacaktır.

Kural 1: IF üyelik =1-5 THEN *Terk* = 0

Kural 2: If üyelik =6-15 THEN *Terk* = 1

Kural 1 ve veriler birlikte incelendiğinde toplam 5 adet 1-5 kaydın dördünde 0 sınıfı yer almıştır. Tüm veri kümesindeki 0 sınıfı sayısı da 5 olduğu için kural 1 düğümündeki kaldıraç değeri şu şekilde hesaplanır.

$$(4/5)/(4/9)= 1.80$$

Kural 2 için ise, toplam 4 adet 6-15 kaydın üçünde 1 sınıfı vardır. O halde kural 2 düğümünün kaldıraç değeri:

$$(3/4)/(4/9)= 1.69$$

Olur.1.69 kaldıra değerinin yorumu şu şekilde yapılabilir:

"Müşteri terkini önlemek için yapılacak bir promosyon çalışmasında, mevcut müşterilerden hiç bir kurala dayanmadan rastgele seçim yaparak bir promosyon çalışması

yapmak yerine, kural ikiye göre bir çalışma yapılacak olursa, işletmeyi terk etmesi muhtemel müşteriye ulaşma şansımız 1.69 kat artacaktır."

Karar ağaçları oluşturulurken hangi uygulamanın seçileceğine dikkat edilmelidir. Farklı ağaç şekilleri de farklı sınıflandırma sonuçları doğuracaktır. Kök denilen ilk düğümü oluşturan Ainin değişik olması, en sondaki yaprağa ulaşılırken kullanılacak yöntemi ve aka bininde sınıflandırmayı da farklılaştıracaktır. Hem kök düğümün hem de bundan sonraki her bir düğümün seçilmesinde en büyük neden o noktadan dallara ayrıldığında veritabanının geri kalan kısmının belli eşit kısımlara ayrılmış olmasıdır. Örneğin veri tabanındaki mevcut cevap evet/hayıt gibi ise iki eşit parçaya, evet/hayır/belki gibi üç değişkenliyse imkânlar dâhilinde üç eşit parçaya bölünmesi gerekmektedir. Bunun nedeni, en basit şekilde istenen cevabın ya da sınıfın elde edilmesidir.

Bu işlem şu oyuna benzer: Bir kişi aklından bir şey tutar; tutacağı şey evrendeki herhangi bir şey olabilir. Örneğin elma, araba gibi nesneler; yürümek, gülmek gibi eylemler veya sıcak, soğuk, kırmızı gibi sıfatlarda olabilir. Diğer oyuncu ise yanıtı *evet* ya da *hayır* olacak şekilde en fazla soruda tutulan bu şeyi bulmaya çalışır. Bu oyunu sürekli oynayanlar ilk sorunun ne olacağını bilirler. Öyle bir soru sorulmalıdır ki, yanıtı ister evet ister hayır olsun evrendeki her şeyi ikiye bölmeli ve yarısı elenmelidir. *"Tuttuğumuz şey, canlı mı?'* sorusu böyle bir sorudur. Bundan sonra ise gelecek cevap durumlarına göre sorular değişecektir. İşte yapılmak istenen ağaç böylesi bir modeldir.

Tablo 1.4. Cinsiyet, boy, kilo ve beden tablosu

Cinsiyet	Kilo	Boy	Beden
K	48	170	Orta
K	49	151	Küçük
K	52	158	Orta
K	56	165	Orta
E	59	160	Küçük
K	61	159	Orta
E	62	162	Küçük
E	63	174	Orta
K	68	168	Orta
K	69	177	Büyük
E	72	170	Orta
E	74	165	Küçük
E	85	175	Orta
E	85	190	Büyük
E	98	190	Büyük

Ağaç oluşturulduktan sonra, her kayıt bu ağaç üzerine uygulanır ve her kayıt ulaşılan yaprağa göre sınıflandırılmış olur. Oluşturulan ağaç aslında çok fazla 'ifthen' den meydana gelir.

Şimdi örnek bir veritabanı üzerinde karar ağaçları algoritmalarının genel çalışma ilkelerini inceleyelim.

Karar ağacı kurulurken boy, kilo, cinsiyet değişkenleri birer düğümü temsil edecek ve her bir düğümden sonra dallanma yaratılarak diğer düğümlere geçiş sağlanacaktır. Bu işlem yapılırken en önemli nokta hangi değişkenin ilk düğüm, yani 'kök düğüm'olacağının hesaplanmasıdır. Burada öyle bir değişken seçilmeli ki verilen yanıt ne olursa olsun, diğer değişkenlerle kıyaslandığında elimizdeki veritabanı kabaca iki eşit parçaya bölüne bilmelidir. Kök düğümden sonra ağacın alt dallarını oluşturacak düğümleri belirlemek için de kök düğümün belirlenmesindeki yol uygulanır. Daha önce de belirtildiği gibi farklı algoritmalar farklı türde ağaç kurabilirler. Ancak her farklı ağaç, kendisi gibi farklı sınıflandırmalara yol açacaktır. Bu nedenledir ki, düğümlerin ve düğümlerden dallara ayrılmanın tespiti oldukça önemlidir.

Verilen tablo temel alınarak aşağıdaki türlerde ağaç kurmak olasıdır. Ancak, bunlardan sadece bir tanesi en doğru sınıflandırmayı gerçekleştirecektir. Uygulamada, farklı iki ağaç hemen hemen aynı ya da birbirine çok benzer sınıflandırma sonuçları verebilir. Bu da değişken sayısının fazla olmasına bağlıdır. Fazla miktarda değişken olması aynı ya da benzer sonuca farklı yollardan ulaşmayı mümkün kılar.

Dallara ayrılırken kullanılacak kriterin hesaplanması da önemlidir. Örneğin cinsiyet sadece iki değer aldığı için bu düğümden ayrılma iki yöne –kadın/erkek- olacaktır; ancak, boy düğümünden ayrılırken kullanılacak dal sayısı kaç olacaktır? Veri tabanındaki tüm boy değerleri için ayrı bir dal mı seçilmeli? Yoksa bu boy seçenekleri 160<boy<165, 166<boy<175, 176<boy<182, 183<boy<183+,…şeklinde aralıklara mı bölünmeli? Aralıklara bölünecekse bu aralık değerleri ne olmalıdır? Kolayca öngörüleceği gibi bu aralık değerleri kadın için farklı, erkek için farklı olacaktır. Bütün bu soruların yanıtları veri tabanında önceden bildirilmiş (C) sınıflara bakılarak verilecektir.

Karar ağacını kaynak alarak meydana getirilen algoritmalar şu kodlama ile çalışır;

```
D:Öğrenme veritabanı
T:Kurulacak ağaç
T=0        // başlangıçta ağaç boş küme
Dallara ayırma kriterlerini seç
T=Kök düğümü belirle
T=Dallara ayrılma kurallarına göre kök düğümü dallara ayır;
her bir dal için
do
bu düğüme gelecek değişkeni belirle
if( durma koşuluna ulaşıldı)
yaprak ekle ve dur
else
loop
```

Verilen kaba kodda sözü edilen durma/sonlandırma koşulunu açıklamakta yarar vardır. Karar ağacı kurulurken eldeki veri tabanının bir kısmı öğrenme işlemi için kullanılarak ağaç oluşturulacaktır; bu arada veri tabanının bir kısmı da oluşturulan ağacı test etmek için kullanılır. Ağaç oluşturulurken sistemin çalışıp çalışmadığı belirlenir. Eğer ağaç belirlenen düzeyde çalışıyorsa dallanma durdurulur ve sınıflandırma tamamlanır. Programdaki durdurma kriteri ağacın hassasiyetini de meydana getirir. Geç durdurulan bir ağaç daha fazla dallanacak ve ağaç daha geniş olacak, çalışma süresi uzayacaktır. Bunun karşılığında ise daha duyarlı sonuç verecektir. Erken durdurulan ağaç ise her ne kadar daha hızlı çalışsa da tam öğrenmenin gerçekleşmeme olasılığını her zaman taşıyacaktır (Dunham, 2003).

Ağaç oluşturmada kullanılan yöntemlerden biride budama yöntemidir. Budama ağaçta meydana gelen sonuca etkisi olmayan ve sınıflandırmaya da hiçbir etkisi olmayan dalların ağaçtan çıkarılmasıdır. Başka bir deyişle gereksiz detayların sonuçtan çıkartılıp

ağaçtan atılması durumudur. Ağacın hassasiyetinin azalmasına neden alt dallara dallara ulaşan veri sayısının azalmasıdır buda birçok düğümün ve dalın oluşmasıyla olur (Cabena, 1998).Budamanın gerçekleştirilmesi için kullanılan algoritmanın işleyiş biçimi budamanın hızı açısından önemlidir; ancak daha önemli bir unsur ise budamanın hangi ölçüte göre yapılacağının belirlenmesidir.

Kullanılan algoritmaların çoğunda varsayılan (default) değer olarak %5-%30 arası değerlerden düşük anlamlılık gösteren değerler budanırken, bu anlamlılığın belirlenmesi kullanıcıya bırakılmaktadır. Budama, gerek ağacın kurulumu esnasında gerekse de kurulduktan sonra yapılabilir. Ağacın kurulumundan önce istenen saflık değeri de ön budama için kullanılan bir değerdir yani ağaç kurulurken verilen saflık değeri dikkate alınarak ağacın yaprakları belirli bir yüzde değerdeki örneğin %70, %95 gibi saflık değerine ulaşınca ağaca yaprak atanarak, diğer dallar için işleme geçilir.

Karar ağaçları baz alınarak oluşturulan karar ağacı sayısı çok fazladır. Bu algoritmalar kök, düğüm ve dallanma kriteri seçimi aşamasında seçilen yöntemlerle farklılık gösterirler.

1.9.1. ID3Algoritması

ID31 algoritması entropiyi bütün değişkenlerin içinden en ayırt edici özelliğe sahip değişkeni bulmak için kullanır. Entropi kavramı, mevcut verinin sayısallaştırılmasıdır (Dunham, 2003).Entropi hiç beklentinin olmamasıdır(Fiske, 1998). Dunham'a göre entropi veri kümesinin sahip olduğu belirsizliği, şaşkınlığı ve rastgeleliği hesaplamayı sapladığını ifade eder. Mevcut bilgilerin tamamının tek bir sınıfın verilerinin olmasını örneklenirse; örneğin herkes aynı filmi izleseydi, herhangi bir kişiye izlediği film sorulduğunda verilecek cevap aynı olacaktı ve sonuç olarak da entropide sıfır (0) olacaktı. Entropi 0-1 arasında bir değer alır. Tüm ihtimallerin eşit olduğu durumlarda entropi en yüksek değere sahip olacaktır.

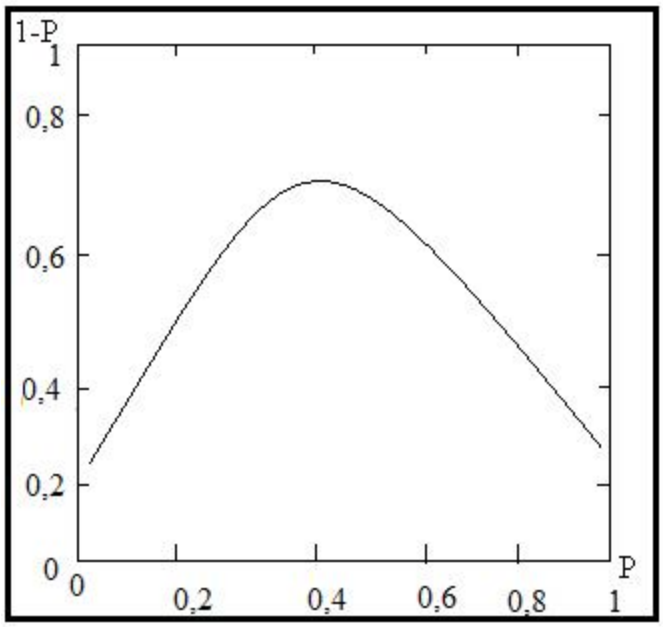
Entropi matematiksel olarak şöyle tanımlanabilir:

$<p_1, p_2 \ldots, p_n>$ olasılıkları gösterir ise bu olasılıkların hepsinin toplamı1 (bir) olmalıdır.

$\sum pi = 1$, böylelikle entropibu şekilde olacaktır:

$H(p_1, p_2,\ldots,p_n)=\sum(p_i\log(1/p_i))$

Bir veritabanının bütününün entropisine bakılır; fakat bu veritabanı farklı alt dallanmalar yaparsa her bir alt dallanmanın da entropisine bakılır. Şekil 1.4H(p, 1-p) veritabanının herhangi bir durumundaki halini göstermektedir.



Şekil 1.4. Entropi – H (p, 1-p)

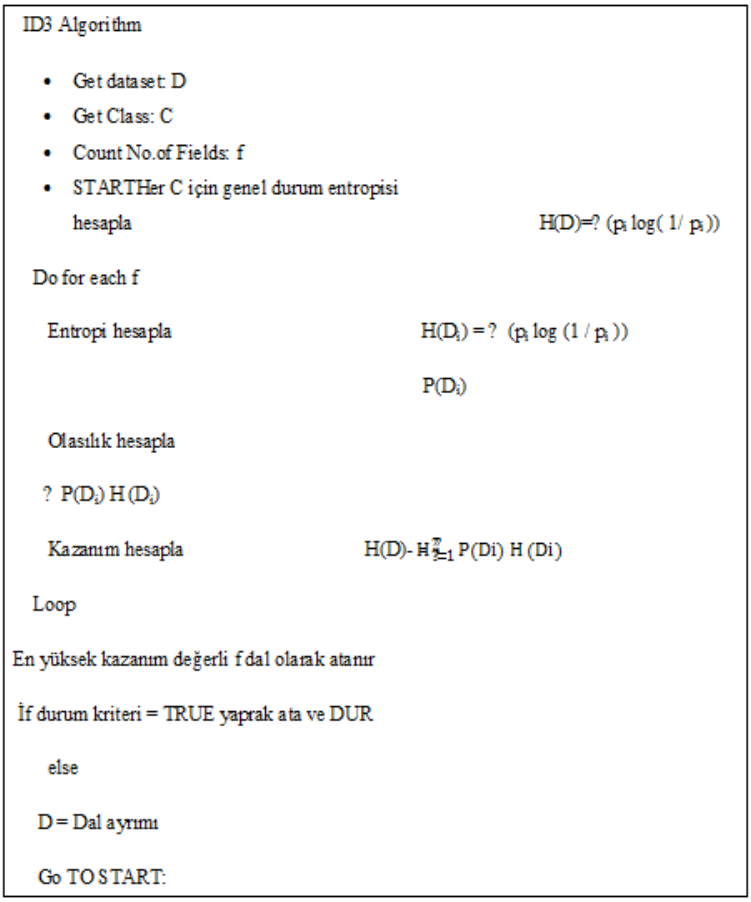
ID3 algoritması veritabanında dallanma yapmadan önce sınıflandırmanın doğru olması için; eldeki veriyle, veritabanı dallandıktan sonra doğru sınıflandırma yapmak amacıyla elde edinilen veriyle aralarındaki fark kullanılır ve öncelikli düğümü ve dallanmaları seçer. Oluşan fark ise kazanım olarak isimlendirilir. Gerçekten de veritabanı bölününce, yani dallanmalar oluştukça doğru sınıflandırma için gerekli bilgi sayısı da azalacaktır.

ID3 algoritmasında kazanım hesabı bu şekilde yapılır:

Ham şekildeki verilerin yani düzeltmeler yapılmadan ilk halindeki şekilleri ile alt dallanmalarının ağırlıklı ortalamaları arasındaki farka bakılır. Yapılan bu hesaplamayla farkın en fazla olduğu noktadan başlanarak alt dallanmaların doğru dallanma olması sağlanır.

Kazanım(D;S) = H(D) - ∑P(Di)H(Di)

Algoritmanın adımları kabaca aşağıdaki gibidir:



Şekil1.5. ID3 Algoritması adımları

1.9.2C 4.5 ve C 5 Algoritmaları

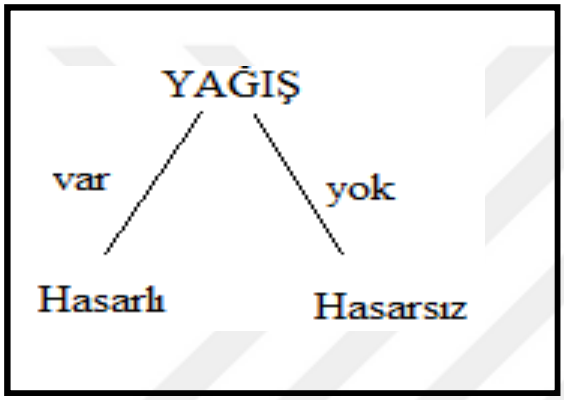
C4.5 algoritması ID3 algoritmasına şu konular açısından üstünlük sağlamaktadır. Karar ağacı oluşturulurken kayıp veriler hesaba katılmaz. Yani, kazanım oranı belirlenirken kullanılan kayıtlar verileri tam olan kayıtlardır eksik veriler kullanılmaz.C4.5 algoritması, kayıp verileri diğer veri ve değişkenler yardımıyla öngörerek kazanım oranının hesaplanmasında kullanılır (Dunham, 2003).Bu sayede anlamlı kurallar türetebilen ve hassasiyeti fazla ağaçlar oluşturulabilinir.

C4.5 ağacın budanması işlemi için iki ayrı yöntem kullanılır:

Alt ağaç yerleştirmesi (Quinlan, 1986) ağacın sahip olduğu alt-ağaçlar yapraklar olarak değiştirilir. Ancak, bu dönüşümün olması için yerleştirilecek yaprağın hata oranının, alt-ağacın hata oranından az olması lazımdır. Örneğin, Şekil 1.6daki ağaçta (Alt-

ağaç)yapılan test işlemi sonucunda yağış var diye gelen 3 kayıttan birinin hasarlı ikisinin hasarsız olduğunu düşünelim. Burada oluşan hata sayısı 2 olacaktır. Yağış yok diye gelen 3 kayıttan 1'i hasarlı, 2 tanesi hasarsız olsun. Bu durumda, gelen toplam 6 kayıttan (2+1) 3 tanesinin hatalı sonuç vereceği açıktır. Bu alt-ağacı hatasız yaprağına dönüştürecek olursak hata sayısı ikiye düşecektir. Bu durumda budama gerçekleştirilebilir.

Diğer bir yöntem ise şudur; herhangi bir alt ağacın, kullanılan ağacı sürekli kullanılan ağacın yerine geçmesi yöntemidir. Aynı şekilde bu yöntemde de değişikliklerin yapılmasıyla hata oranında düşüşler görülmesi gerekir.



Şekil 1.6. Ağaç halinde sorgulama

ID3 algoritması, değişkenleri birçok alt bölüme ayırır ve bu ayırma aşırı öğrenmeye neden olabilir. Bunu önlemek için Quinlan kazanım yerine aşağıdaki kazanım oranını kullanmıştır.

Kazanım Oranı(D,S) = Kazanım(D,S)/Ayırma Bilgisi(D,S)

Ayırma Bilgisi(D;S) = H($|D1|/|D|,\ldots,|Ds|/|D|$)

- Getdataset: D
- Get Class: C
- CountNo.of Fields: f
- Start: C sınıfı için genel durum entropisini $H(D) = \sum(p_i \log (1/p_i))$ hesapla

Do foreach f

Ayrılma bilgisini hesapla $H (|D_1|/|D|,\ldots\ldots,|D_s|/|D|)$

Kazanım oranını hesapla

Kazanım Oranı (D,S) = Kazanım (D,S) / Ayırma Bilgisi (D,S)

Loop

An küçük değerli kazanım oranına sahip değişkeni düğüm olarak ata.

İf durma kriteri = TRUE yaprak ata ve DUR

Else

D= Dal ayrımı

Go TO START:

Şekil1.7. C 4.5 VE C 5 Algoritmaları

1.9.3. CART Algoritması

CART tekniği tıpkı ID3 algoritmasındaki gibi en iyi dallanma kriterini bulmak için entropiden yardım alır (Dunham, 2003). CART en iyi dallara ayırma kriterini seçmek için ise ID3, C4.5 ve C5 te olmayan değişik bir formüle sahiptir. CART algoritmasının en iyi dallanma kriteri şu şekilde hesaplanır (Yohannes ve Webb, 1999). CART ya da C&RT olarak literatürde yer alan bu algoritma ikili ağaçlar üretir. Dolayısıyla, hangi düğümün kök ya da düğüm olacağına karar vermesinin yanı sıra, belirlenen düğümün hangi noktasından düğümün ikiye ayrılması gerektiğini de hesaplamak durumundadır. Başka bir deyişle, algoritma dallanma yapmadan önce uygun değişkeni seçer, bu sırada seçilen değişken de eğer farklı değerler varsa bunlarında nasıl gruplanacağının belirlenmesini de yapar. Bu işlemleri aşağıdaki parametreleri kullanarak gerçekleştirir.

Herhangi bir t düğümündeki s dallara ayrılma kriteri $\Psi(s/t)$ olarak gösterilirse:

$\Psi(s/t) = 2PLPR\sum \left| P(Cj \mid tl) - P(Cj \mid tR) \right|$

t: Dallanma meydana gelecek düğüm

c:Kriter

L: Ağacın sağ tarafı

R: Ağacın sol tarafı

PL, PR: Öğrenim kümesindeki bir kaydın sağda veya solda olma olasılığı.

$P(Cj \mid tl)$ ve $P(Cj \mid tR)$:Cj sınıfındaki bir kaydın sağda veya solda olma olasılığı.

CART dallara bölünme kriterini hesaplarken kayıp verileri dikkate almaz. Belirtildiği algoritmanın ürettiği ağaç yapısı da diğer algoritmalardan farklı olarak ikili ağaçtır. Hesaplanan $\Psi(s/t)$ değerleri içerisinde en büyük değeri taşıyan nokta düğüm olarak seçilir ve aynı işlemler diğer karar ağacı algoritmalarında olduğu gibi tüm yapraklara ulaşıncaya kadar devam ettirilir.

1.9.4. SPRINT Algoritması

Daha öncede de belirtildiği gibi ID3, CART, C4.5 ve C5 gibi algoritmalar ilk olarak derinlik kuralına göre çalışırlar ve veriler sürekli olarak sıralanır bunun nedeni ise en iyi dallara ayırma kriterine ulaşmayı sağlamaktır. Bu durum SLIQ algoritmasında daha basittir ve her değişken için bir liste yapılıp sıraya dizilme işlemi sadece bir kez yapılır. SPRINT algoritması bu yönüyle SLIQ algoritmasına benzer ve sözü edilen diğer algoritmalardan ayrılır.

SPRINT çalışmaya her değişkeni tek tek listeleyerek çalışmaya başlar. Bu listelerde kullanılacak olan değişkenler, sınıflar ve sıra numaraları yazar. Böylece veri tabanındaki değişken sayısıyla aynı sayıda tablo meydana gelir. Sürekli değerleri taşıyan tablolar sürekli değer değişkenine göre sıraya dizilirken, kategorik veri taşıyan diğer tablolar sıra numarasına göre sıralı kalacaktır. Eğitim kümelerinin sayesinde oluşturulan ilk tablolar sınıflandırma ağacının köküyle bağlantılı hale getirilir. Ağaç dallanıp düğümler yeni dallara ayrıldıkça düğümlere ait değişken listeleri de ayrılarak yeni oluşan dallarla alakalandırılır. Liste bölündüğünde içindeki kayıtların sıralaması değiştirilmez; böylece oluşturulan yeni listelerin bir daha kendi içlerinde sıraya dizilmesine gerek kalmaz.

Ayrılma safhasına gelen düğümler için $C_{üst}$ veC_{alt} diyea dlandırılan ve düğümdeki sınıf dağılımlarına ulaşma sebebiyle kullanılan histogramlar belirlenir. Alt dallara bölünme şartı için isetıpkı SLIQ algoritmasında ki gibi GINI indeksi kullanılır. Herhangi bir S kümesinin (gini S) indeksi aşağıdaki gibi hesaplanır.[Brieman, 1984]

$gini(S)=1 -\sum p^2_j$

Burada p_j, S kümesi içinde j sınıfının tekrar değeridir. Eğer S kümesi S_1 ve S_2 gibi alt kümlere ayrılırsa K kümesinin $gini_{ayrılmış}$ (S) değeri;

$gini_{ayrılmış}(S) = \sum_{i=1}^{t} \frac{ni}{n} gini(Si)$

şeklinde hesaplanır. Aşağıda algoritmanın adımları verilmiştir.

Veri Seti: D
Sınıflar: C
Değişkenler: f
Durma Kriteri: dk
START
Do foreach f
Calculate Splitting value $gini_{SPLIT}$ (K) = $\sum_{i=1}^{t} \frac{ni}{n} gini(Ki)$
Loop
En küçük $GINI_{SPLIT}$ değerli değişkeni düğüm olarak ata
İf durma kriteri = TRUE then yaprak ekle VE DUR
Else
D = NextBranch
Go TO START:

Şekil 1.8. SPRINT algoritması adımları

2. MATERYAL VE METOD

2.1.CHAİD Algoritması

CHAID algoritması ilk olarak ‘’An Exploratory Technique for Invetigating Large Quantities of Categorical Data’’ ismi altında sunulmuştur. İstatistikte bilinen Ki-kare test yöntemi kullanılarak bir karar ağacı oluşturur. Dolayısıyla hem, bir karar ağacı hem de istatistiğe dayalı bir algoritmadır. CHAID, CART algoritması gibi ikili ağaç oluşturur. Dolayısıyla her bir değişkendeki farklı değer sayısını ikiye indirerek ağacı oluşturmaya devam eder.

Algoritma adımları aşağıdaki gibi özetlene bilir;

Girdiler: Veri kümesi

Değişken sayısı

Her bir değişkendeki farklı değer sayısı

Adımlar:

1.DO FOR *her bir tahminleyici değişken (X)*

2.IF (X >*iki farklı değer*)

DO UNTIL (X = *iki farklı değer*)

Tüm değerleri diğer değerlerle ikili *Ki-Kare* testinde sok ve en büyük *p* değeri veren ikilileri tek değer olarak etiketle.

3.Tüm iki değerli değişkenler için sınıf değişkenine göre *ki-kare* değerini hesapla.

4.En küçük *p* değerine sahip değişkenden dal yarat

LOOP

5.DUR

Yukarıdaki algoritma ağaç tamamlana kadar dönecektir. Algoritmanın özünü oluşturan Ki-Kare test istatistiği ise şu şekilde hesaplanır.

Öncelikle veriler aşağıdaki gibi (2Xd)'lik bir tablo haline dönüştürülür. Buradaki d değeri sınıf değişkenindeki değer sayısı (sınıf sayısı) 2 ise ilgili değişkendeki değer sayısını ifade etmektedir.

Tablo 2.1. Ki-kare hesaplama tablosu

	Sınıf 1	**Sınıf 2**	**Sınıf 3**	**Toplamlar**
Değer 1	a	b	c	a+b+c
Değer 2	d	e	f	d+e+f
Toplam	a+d	b+e	c+f	a+b+c+d+e+f=N

Algoritma incelendiğinde tüm değişkenlerin iki farklı değere farklı olması gerektiği ve eğer bir değişken ikiden farklı değere sahipse o değişken kendi içinde iki ayrı değere düşene kadar yine ki-kare testi uygulandığı görülmektedir. Dolayısıyla bu algoritma için 2Xd lik tablo yeterlidir. İstatistikte kullanılan ki-kare testinde ise daha büyük boyutlu tablolar kullanıla bilmektedir. Daha sonra;

$$X^2 = \sum \frac{(\text{gözlenen} - \text{beklenen})2}{beklenen}$$

buradaki gözlenen ve beklenen değer ise şu şekilde hesaplanır.

Gözlenen değer a, b, c, d, e, f değerlerinden her biridir. Herbir gözlenen değerin karşılığı beklenen değerse o değerin sütün ve satır toplamlarının genel toplama bölümüdür. Örneğin a değerinin beklenen değeri aşağıdaki gibi hesaplanır.

a(beklenen)=[(a+d) (a+b+c)] / N

Daha sonra ki-kare tablosu yardımıyla, bu değerin karşılığı olan p olasılık değeri belirlenir. Ki-kare tablosunu kullanmak için iki ayrı parametreye ihtiyaç vardır; bunlardan biri serbestlik derecesi, diğeri ise ki-kare değeridir.

CHAID algoritmasının 2Xd' lik tablo yapısı gereği serbestlik derecesi her zaman (2-1)X(d-1)=sd olur. Bir başka değişle serbestlik derecesi sınıf sayısı-1 olarak hesaplanır.

Tablo da serbestlik derecesi=2 için ki-kare değerlerinin karşılığı p değeri bulunmaktadır. Daha farklı serbestlik dereceleri için EXCEL programında CHDIST (ki-kare değeri; anlamlılık düzeyi) formülü kullanılabilir. Bunun yanı sıra p değeri yerine doğrudan ki-kare değeri de kullanıla bilir. Böyle durumlarda, algoritmanın ikinci adımındaki en büyük p değeri yerine en küçük ki-kare değeri, algoritmanın üçüncü adımındaki en küçük p değeri yerine ise en büyük ki-kare değeri seçilebilir.

3. BULGULAR

3.1.Uygulama

Bu çalışmada Elazığ ili Ticaret ve Sanayi Odasına kayıtlı işletmelerden elde edilen veriler ışığında işletmelerin Elazığ ekonomisine yaptıkları katkılar incelenmiştir. Yapılan araştırmalar sonucunda işletmelerin sağladığı istihdam sayıları belirlenmiştir. Bu bilgiler doğrultusunda işletmelerin ekonomiye katkısı belirlenmiştir. Ekonomik getiri açısında bakıldığında bunun sadece işçi istihdamı ile kalmayıp üretimde kullanılan ham maddeden işletmede kullanılan envanterler Elazığ ilinde faaliyet gösteren işletmelerden temin edildiğinden dolayı ekonomik açıdan getiri sağlamaktadır. İşletmelerin muhasebe gider kayıtlarına bakıldığında resmi olarak 47 adet gider kolu görünmektedir. Bu gider kolları her işletmede gider olarak vergilendirilmektedir, bu açıdan bakıldığında yapılan her harcama başka bir işletmeye gelir sağlayacaktır ve il içinde ekonomik bir döngü sağlayıp ilde ki ticaret faaliyetlerini oluşturacaktır. Bu gider kalemleri aşağıdaki tabloda belirtilmiştir. İşletmeler çeşitli şekillerde sınıflandırılabilirler; İş Yaptığı Alanlara göre, Tüketicilerin türüne göre, Üretilen ürün ve hizmet türüne göre, Üretim elemanlarının mülkiyetine sahipliğine göre, Büyüklüklerine göre, Yasal yapılarına göre gibi sınıflandırmalar yapılabilinir.

Bu çalışmada büyüklüklerine göre sınıflandırılan işletmeler ile inceleme yapılmıştır. Avrupa Birliği ne uyum sağlama sürecinde 2005 yılında yapılan yönetmelikte bu gruplama şöyledir; mikro işletme, küçük işletme, orta işletme ve büyük işletmeler olmak üzere işletmeleri dört ana gruba bölmektedir.

Tablo 3.1. Vergilendirilebilen gider kalemleri

1	Kırtasiye	25	Araç Sigortalama
2	Yiyecek İçecek	26	Küçük Demirbaşlar
3	Telefon Faks	27	Büro Bakım Onarım
4	İski Su	28	Araç Yakıt
5	Elektrik	29	Kurye
6	Reklam Tanıtım	30	Gazete Ve Dergi
7	Kargo Kurye	31	Araç Otopark
8	Temizlik	32	Araç Bakım Ve Onarım
9	Teknik Servis	33	Matbaa
10	SmmYmm Avukat	34	Maaş
11	Burs Gideri	35	Banka Komisyon
12	İnternet Sarf Malz.	36	Kredi Kartı Komisyonları
13	Noter	37	Bağkur
14	Bilg. Sarf Malzemeleri	38	Prim
15	Vergi Resim Ve Harçlar	39	Yurt İçi Ulaşım
16	Diğer Çeşitli	40	Araç Vergi
17	Personel Sağlık	41	Ssk
18	Şehir İçi Ulaşım	42	Ogs
19	Temsil Ve Ağırlama	43	Kanunen Kabul Edilmeyen
20	Apartman Aidatları	44	Fiyat Farkları
21	İşyeri Sigortalama	45	Yurtdışı Ulaşım &Konaklma
22	Kiralar	46	Araç Kiralama
23	İgdaş Doğalgaz	47	Kıdem İhbar
24	Diğer Haberleşme Giderleri		

Mikro İşletme: Yıllık satışı 1 milyon TL nin altında olan ve 10 kişinin altında işçi çalıştıran işletmelerdir.

Küçük İşletme: Yıllık satışı 1 milyonun TL üzerinde olan ve 10 ile 50 kişi arasında işçisi olan işletmelerdir.

Orta Büyüklükteki İşletme: Yıllık satışı 25 milyon TL ye kadar olan ve 50 ile 250 arasında işçi çalıştıran işletmelerdir.

Büyük İşletmeler: Yıllık satışı 25 milyon TL üzeri olan ve 250 kişi ve daha fazlasını istihdam eden işletmelerdir.

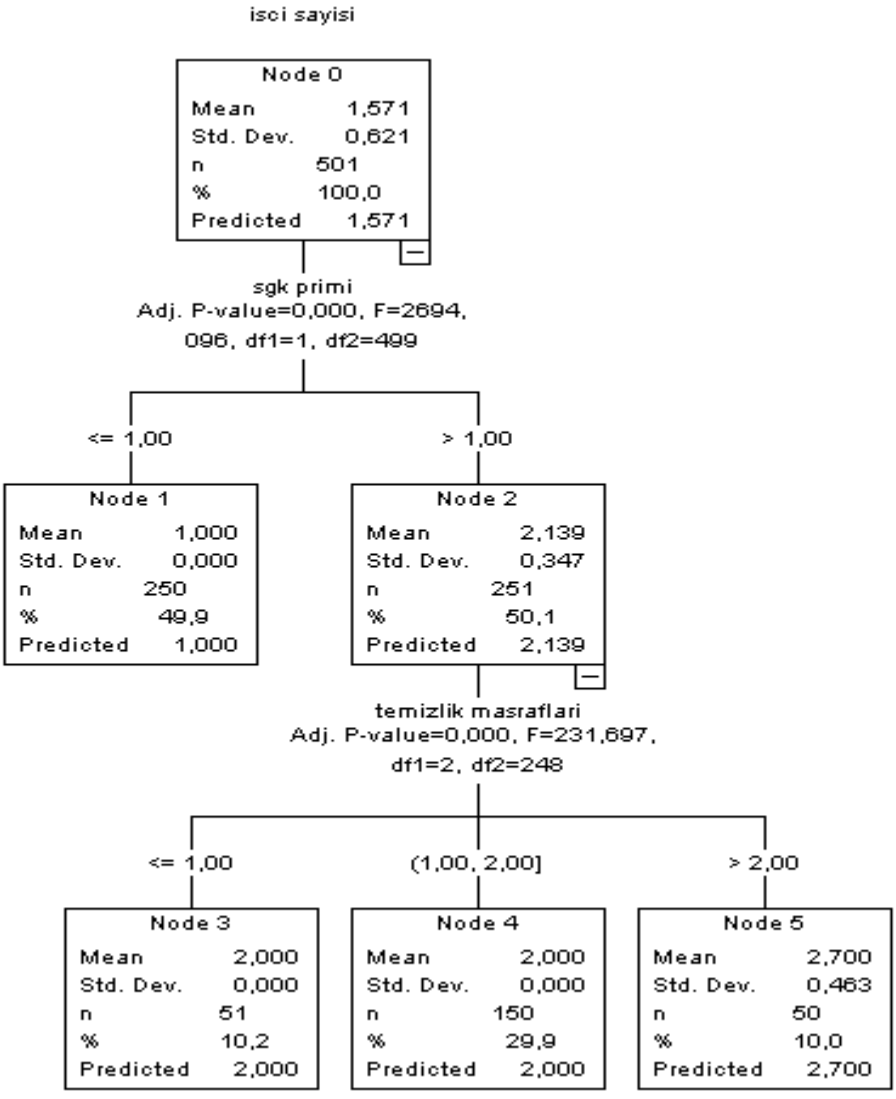
Elazığ ilinde net olmamak ile birlikte yaklaşık 2692 limited şirket, 2163 şahıs şirketi, 276 anonim şirket,286 kooperatif, 12 adet kolektif şirket, 41 banka merkez ve şubesi ve 1 adet holding bulunmaktadır .Bu çalışmada Elazığ ilin deki işletmelerden 501 adedi incelenmiştir, inceleme rassal seçilmiş küçük,orta ve büyük işletmelerde yapılmıştır. Chaid analizi bir karar ağacı ve algoritma da olduğundan bir bağımlı değişken diğer bir adıyla hedef değişken ve bu bağımlı değişkeni açıklayacak, destekleyecek veya etkilerini gözlemleyebilmemizi sağlayan değişkenlere ihtiyaç duyar bu değişkene de bağımsız değişken denilmektedir ve bu iki değişken ile analiz yapılmaktadır İşletmeler istihdam sağladıkları işçi sayıları bağımlı değişkenine elektrik gideri, doğalgaz gideri, su gideri, işletme gıda masrafları, kırtasiye masrafları, araç yakıt giderleri, temizlik masrafları ve sgk primleri çalışmada bağımsız değişkenler olarak seçilmiştir.

Tablo 3.2. Değişkenlerin listesi

	DEĞİŞKENLER
Y	İşçi istihdamı: İşletmede çalıştırılan kişi sayısı
X_1	Elektrik kullanımı: İşletme elektrik faturası gideri
X_2	Su kullanımı: İşletme de kullanılan su ve su faturası gideri
X_3	Doğalgaz kullanımı: İşletmenin doğalgaz faturası gideri
X_4	İşletme gıda masrafları: İşletme yiyecek giderleri
X_5	Kırtasiye masrafları: İşletmede kullanılan kırtasiye malzemesi giderleri
X_6	Araç yakıt giderleri: İşletme bünyesindeki taşıtların yakıt giderleri
X_7	Sgk primi: İşletme de çalışan kişilerin sgk prim gideri
X_8	Temizlik masrafları: İşletme temizliğinde kullanılan temizlik malzeme gideri

Örneklem kümemiz Elazığ ili Ticaret ve Sanayi Odasına kayıtlı 5488 işletmeden seçilmiştir, kayıtlı olan kuruluşlardan yaklaşık olarak 1200'ü aktif değildir. Mevcut işletmelerden rassal olarak 338 küçük ve 162 büyük ve orta büyüklükteki işletme seçilmiştir. Bu işletmelerden sgk primi ödemesi 371 tanesi 110,000 TL ye kadar ödeme yaparken 130'u 110,000 TL üzeri ödeme yapmıştır, elektrik gideri olarak gösterilen fatura bedeli 232'si 5000 TL ye kadar 269'u 5000 TL üzeri fatura ödemişlerdir, doğalgaz faturaları 8000 TL ye kadar 337 işletme 8000 TL üzeri ödeyen 164 işletme, su faturası 750 TL ye kadar ödeyen 349 işletme 750 TL üzeri ödeyen 152 işletme, iş yeri gıda masrafı 10,000 TL' ye kadar 205 işletme 10,000 TL üzeri ödeyen 296 işletme, kırtasiye masrafı

1000 TL ye kadar olan 298 işletme 1000 TL üzeri 203 işletme, işyeri temizlik masrafı 2.000 TL ye kadar olan 301 ve 2,000 TL üzeri olan 200 işletme görülmüştür masraflar aylık olarak belirtilmiştir.



Şekil 3.1. Yapılan analizin karar ağacı

4. SONUÇ VE TARTIŞMA

Elazığ ilinde Avrupa Birliği 2005 yılı uyum yasasına göre işletmeler büyüklükleri bakımından 4 gruba ayrılır. 2005 yılından sonra ilgili yasa çerçevesinde hükümetler bu tür işletmelere bazı kalemlerde kolaylık sağlamaktadır; bundan ötürü elektrik, su, doğal gaz, kırtasiye, yakıt ve gıda masrafları, SGK ve temizlik kalemlerine göre şirketler lehine muafiyet getirmektedir. Hatta istihdam ettiği işçi sayısı arttıkça ödediği vergi ve diğer kalemlerde de azalmalar gözlemlenmiştir.

Elazığ ili Ticaret ve Sanayi Odasına bağlı 5488 işletmenin şehir ekonomisine katkısı SGK primi ve temizlik malzemeleri yönünden gerçekleşmiştir tabloda da görüldüğü gibi 8 kalemin ortalaması SGK primine göre 1 in altında 250 şirket ve 1 in üzerinde 251 işletme de gerçekleşmiş 1 in altında kalanların düğümü sonlandırılmış 1 in üzerinde olanlar ise 2. Düğüme geçmiş ve burada da temizlik malzemeleri giderleri tüm gider ortalamasına göre 1 in altında olan 51 işletme, 1 ile 2 arasında 150 ve 2 nin üzerinde 50 olmak üzere toplam 251 işletme şehir ekonomisine katkı sunmaktadır.

Elazığ da faaliyet gösteren ve Ticaret ve Sanayi Odası'na kayıtlı işletmelerin elektrik, su, doğal gaz, ve araç yakıt giderlerinin şehir ekonomisine yapacakları katkı işçi istihdamı dışında getiri sağlamamaktadır. Dolaylı olarak SGK priminde ve çalıştırılan işçi sayısında kazanım sağlamaktadır, bunun dışında kırtasiye, temizlik ve gıda masrafları doğrudan şehir ekonomisine katkı sağlaması gereken kalemlerdir. Ancak bu kalemlerden sadece temizlik ürünleri yapılan analiz sonucunda etkili çıkmıştır. Bunun sebepleri ayrıca araştırılması gereken konudur

İlgili kurum ve kuruluşlar şehir ekonomisine katkı sağlayacak kalemlerde artışa giderken; bunun yanı sıra işletmelerin harcama kalemlerinin abartılı ve yanlış gösterilenlerini detaylı ve kapsamlı incelemeli ve yerel işletmelerden doğru bilgilerin temini sağlanılmalıdır.

KAYNAKLAR

Agarwal, S. (1993). "Influence of Formalization on Role Stress, Organizational Commitment, and Work Alienation of Salespersons: A Cross-National Comparative Study", Journal of International Business Studies, Vol. 24, No. 4, 4th Qtr.,pp. 715-739

Breiman, L.(1984), Classification and regression trees, Wads worthInc, Monterey, California.

Cabena, P.(1998). Discovering Data Mining from Concept to Implemantation, Prentice Hall, New Jersey

Dunham, M.H.(2003). Data Mining: Introductory and Advanced Topics, Prentice Hall Publication.

Fiske, S. T., (1998) "Prejudice, stereo typing, and discrimination", D. T. Gilbert, S. T. Fiske, ve G. Lindzey (der.), The handbook of social psychology, 4. baskı, McGraw-Hill, New York, s. 357-411.

Ganesh, S.(2002). Data Mining: Should it be Included in th'Statistics' Curriculum? TheSixth International Conference on Teaching Statistics, Cape Town, South Africa, 7–12 July.

Kalaycı, Ş.(2010), "SPSS uygulamalı Çok Değişkenli İstatistik Teknikleri." Asil Yayın Dağıtım, Ankara

Koyuncugil. A. S. (2007), A.S."Veri Madenciliği ve Sermaye Piyasalarına Uygulanması." Sermaye Piyasası Kurulu Araştırma Raporu: 1-17.

Kuonen, D. (2004). Data mining and statistics: what is the connection? The Data Administration Newsletter, 30.0.

Oğuzlar, A.(2003), "Veri Ön İşleme", Erciyes Üniversitesi İİBF Dergisi, 21, 67-76.

Oğuzlar, A. (2004), "Çok Değişkenli Kontrol Grafikleri ve Bir Uygulama", Uludağ Üniversitesi İİBF Dergisi, 23 (2), 65-74.

Quinlan, J.(1986), Ross. "Induction of decisiontrees." Machine learning 1.1: 81-106.

Silahtaroğlu, G. (2008), "Veri madenciliği." Papatya Yayınları, İstanbul

Spiegel, M. ve Stephens, L. (1999). "İstatistik. "Nobel Yayın Dağıtım, Ankara

Şentürk, A.(2006), "Veri Madenciliği Kavram ve Teknikler." Ekin Kitabevi, Bursa, 114s .

Yıldırım, P., Uludağ, M.ve Görür, A. (2007). "Hastane Bilgi Sistemlerinde Veri Madenciliği." Çanakkale On Sekiz Mart Üniversitesi Akademik Bilişim.

Yohannes, Y. ve Webb, P. (1999). Classification and regression trees, cart: A user manual for identifying indicators of vulner ability to famine and chronic foodin security. Microcomputers in Policy Research No. 3. Wahington

Zhao, C.M. ve Luan, J. (2006). Data Mining: Going Beyond Traditional Statistics. New Directions for Institutional Research, No. 131, pp. 7–16

ÖZGEÇMİŞ

1990 yılında Elazığ ilinde doğan Selman AKTAŞ, orta öğrenimini Elazığ Dumlupınar İlköğretim Okulunda 2003 yılında tamamlamış, lise öğrenimini Elazığ Mehmet Akif Ersoy Lisesi'nde 2007 yılında tamamlamıştır, Yüksek Öğrenimini ise Fırat Üniversitesi İstatistik Bölümü'nde 2008-2012 yılları arasında tamamlamıştır. 2013 yılında Fırat Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Olasılık Süreçleri ve Olasılık Teorisi Bilim Dal'ında yüksek lisans eğitimine başlamış ve halen eğitimini sürdürmektedir.