T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÖĞRENCİLERDE ALKOLLÜ İÇECEK KULLANIMININ VERİ MADENCİLİĞİ YÖNTEMLERİ İLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

Nihal Zuhal KAYALI

Enstitü Anabilim Dalı	**:**	**BİLGİSAYAR VE BİLİŞİM MÜHENDİSLİĞİ**
Tez Danışmanı	**:**	**Doç. Dr. Nilüfer YURTAY**

Haziran 2019

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ÖĞRENCİLERDE ALKOLLÜ İÇECEK KULLANIMININ VERİ MADENCİLİĞİ YÖNTEMLERİ İLE İNCELENMESİ

YÜKSEK LİSANS TEZİ

Nihal Zuhal KAYALI

Enstitü Anabilim Dalı	**BİLGİSAYAR VE BİLİŞİM MÜHENDİSLİĞİ**

Bu tez 10/06/2019 tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

Doç. Dr.	**Dr. Öğr. Üyesi**	**Dr. Öğr. Üyesi**
Nilüfer YURTAY	**Veysel Harun ŞAHİN**	**Bahadır HİÇDURMAZ**
Jüri Başkanı	**Üye**	**Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde herhangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir üniversitede herhangi bir tez çalışmasında kullanılmadığını beyan ederim.

Nihal Zuhal KAYALI

10.06.2019

TEŞEKKÜR

Yüksek lisans ve lisans eğitimim boyunca değerli bilgi ve deneyimlerinden yararlandığım, her konuda bilgi ve desteğini almaktan çekinmediğim, araştırmanın planlanmasında yardımlarını esirgemeyen, teşvik eden, aynı titizlikte beni yönlendiren, akademik hayatımın mimarı olarak nitelendirebileceğim değerli danışman hocam Doç. Dr. Nilüfer Yurtay'a ve Öğr. Gör. Yüksel Yurtay'a teşekkürlerimi sunarım.

Beni dürüst ve ahlaklı bir Cumhuriyet kadını olarak yetiştiren babama ve anneme hiçbir özveriden kaçınmadan beni bu günlere getirdikleri için, fedakârlıklarını asla ödeyemeyeceğimi bilerek en derin sevgi ve saygılarımı sunmak istiyorum. Bu çalışmanın her aşamasında ve ne zaman ihtiyaç duysam yanımda hissettiğim dert ortağım biricik eşim Ömer Kayalı ve beni sürekli motive edip bana ışık tutan dostum Özge Gökçe iyi ki varsınız.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

AI	: Yapay Zeka
CRISP-DM	: Veri Madenciliği için Sektörler Arası Süreç
CRM	: Müşteri İlişkileri Yönetimi
DF	: Serbestlik değeri
EVM	: Eğitimsel Veri Madenciliği
KDD	: Veritabanlarında Bilgi Keşfi
KMO	: Kaiser-Meyer-Olkin
SEMMA	: Örnek, Keşfet, Değiştir, Modelle ve Değerlendir
TÜİK	: Türkiye İstatistik Kurumu

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Veri Madenciliği, Sınıflandırma, Karar Ağacı, C 5.0 Algoritması, Alkollü İçki Tüketimi

Günümüzde alkol ve benzeri bağımlılık yapıcı maddelere bağımlı olma riski özellikle gençler için her geçen gün artan şekilde tehlike yaratmaktadır. Bu tehlikeli maddeleri kullanmaya başlama yaşı ülkeden ülkeye değişmektedir. Alınan tüm yasal düzenlemelere rağmen bağımlılıkların 10'lu yaşlara kadar düşmesinin önüne geçilememiştir. Gençler yetişkinlere kıyasla hem risk almaya daha fazla açıktır hem de daha az deneyime sahiptir. Tehlikeli deneyimler gençler için cezbedicidir ve yetişkinlerin dünyasına bir tür ait olma çabası olarak görülürler. Gençlerde alkole bağımlılık çok daha hızlı gerçekleşebildiği gibi gençlik dönemindeki sakıncalı alkollü içecek tüketimi, ilerleyen yıllarda benzer tehlikeli maddelere karşı oluşabilecek bağımlılığı riskini yükseltmektedir. Bu bağımlılıkların başlamadan önlenmesi için bağımlılık yapıcı sebeplerin irdelenmesi ve çözülmesi gerekir.

Bu çalışmada veri madenciliği süreçlerinden bahsedilerek popüler veri madenciliği yöntemlerinden lojistik regresyon analizi ve C 5.0 algoritması ile gençlerde alkol tüketiminin belirli faktörler üzerindeki ilişkisini inceleyen bir çalışma yapılmıştır.

INVESTIGATION OF THE USE OF ALCOHOLIC BEVERAGES WITH DATA MINING METHODS IN STUDENTS

SUMMARY

Keywords: Data Mining, Classification, Decision Tree, C 5.0 Algorithm, Consumption of Alcoholic Beverages

Nowadays, the risk of being dependent on alcohol and other dangerous substances is increasing danger especially for young people. The age of commencement of using these dangerous substances varies from country to country, but it has fallen to the age of 10 despite all legal regulations. Young people are more vulnerable to taking risks than adults and have less experience. Dangerous experiences are attractive for young people and are an effort to belong to the world of adults. Alcohol addiction is much faster in young people, and alcohol consumption during the youth period increases the risk of alcohol dependence in the following years.

In this study, data mining processes are discussed and regression analysis and C 5.0 algorithm of popular data mining methods and the relationship between alcohol consumption and specific factors in young people is investigated.

BÖLÜM 1. GİRİŞ

Yeni düşünce, yeni teknoloji yetenekleri ve dijital işlere doğru ilerlemenin yol açtığı veri kullanımı radikal olarak değişirken, günümüzde tüm alanlarda veriler dramatik bir hızla toplanmakta ve biriktirilmektedir. İnsanların hızla büyüyen sayısal veri hacimlerinden faydalı bilgi elde etmelerine yardımcı olan yeni nesil hesaplama teorilerine ve araçlarına acil ihtiyaç vardır. Bu teoriler ve araçlar, veritabanlarında ortaya çıkan bilgi keşfinin (KDD) konusudur. KDD teorik anlamda, anlamlı verileri ortaya çıkarabilmek için yöntem ve tekniklerin geliştirilmesi ile ilgilidir. KDD sürecinin ele aldığı temel problem, ilk bakışta kolayca anlamak ve sindirmek için çok hacimli olan düşük seviyeli verileri; kısa bir rapor gibi daha kompakt, verileri oluşturan sürecin açıklayıcı bir yaklaşımı veya modeli ile ifade edilebilecek şekilde daha soyut ve gelecekteki vakaların tahmini için daha yararlı olabilecek diğer formlara eriştirmektir. Sürecin özünde desen keşfi ve çıkarımı için özel veri madenciliği yöntemlerinin kullanılması vardır.

1.1. Veri Madenciliğinin Tanımı ve Tarihi

İnsana dair hemen her alanda veritabanlarındaki verilerin muazzam şekilde çoğalması bu karmaşık verileri anlamlı ifadelere dönüştürmek için kullanılacak güçlü ve yeni araçların yaratılması konusunda büyük bir talep yaratmıştır. Araştırmacılar bu ihtiyacı karşılama çabasıyla makine öğrenmesi, örüntü tanıma, istatistiksel veri analizi, veri görselleştirme, sinir ağları, vb. yöntemlerin keşfedilmesi ve geliştirilmesini araştırmıştır. Bu çabalar "Veri Madenciliği ve Bilgi Keşfi" olarak adlandırılan yeni bir araştırma alanının ortaya çıkmasına neden olmuştur. Mevcut Bilgi Çağı, olağanüstü bir veri artışı ile karakterize edilir ve her türlü insani çaba hakkında bilgi üretilmekte ve depolanmaktadır. Bu veriler sürekli artan oranlarda veritabanlarında kaydedilir, böylece bilgisayar teknolojileri ile kolayca erişilebilinir. Bu verilerin çok büyük

hacimli oluşu içlerinden faydalı ve görev temelli bilginin nasıl çıkarılacağı konusunda sorun yaratmıştır. Yapılacak çıkarma işlemleri için veri analizi tekniklerinden geleneksel olarak regresyon analizi, küme analizi, sayısal taksonomi, çok boyutlu analiz, diğer çok değişkenli istatistiksel yöntemler, stokastik modeller, zaman serisi analizi, doğrusal olmayan tahmin teknikleri ve diğer yöntemler kullanılmıştır. Bu teknikler birçok pratik problemi çözmek için yaygın olarak kullanılmaktadır. Bununla birlikte, bunlar öncelikle niceliksel ve istatistiksel veri özelliklerinin çıkarılmasına yöneliktir ve bu nedenle kendi içsel sınırlamaları vardır.

Örneğin, istatistiksel bir analiz verilerdeki değişkenler arasında kovaryans ve korelasyonlar belirleyebilir. Ancak bu bağımlılıkları karakterize edip kavramsal seviye ve prosedürde var olma nedenlerinin basit bir açıklamasını yapamaz. Bu ilişkilerin daha üst düzey mantık tarzı tanımlamalar ve yasalar biçiminde bir gerekçesini de geliştiremez. Bir istatistiksel veri analizi, verilen faktörlerin merkezi eğilimini ve varyansını belirleyebilir ve bir regresyon analizi, bir dizi veri noktasına bir eğriye sığabilir. Bununla birlikte, bu teknikler düzenlemelerin niteliksel bir tanımını yapamaz ve verilerin bağımlılıklarını belirleyemezler. Ayrıca keşfedilen düzenlilik ile başka bir alandaki düzenlilik arasında bir analoji çizemezler. Sayısal taksonomi tekniği, varlıkların sınıflandırılması olabilir ve varlıklar arasında sayısal bir benzerlik belirtebilir. Ancak, aynı kategorideki varlıklar için yaratılan ve varsayılan sınıfların nitel tanımını oluşturamaz.

Benzerlik ölçümlerinin yanı sıra benzerliği tanımlayan öznitelikler daha önceden bir veri analisti tarafından tanımlanmalıdır. Yani bu teknikler ilgili özellikleri otomatik olarak üretemezler ve farklı veri analiz problemleriyle değişen ilişkilerini belirlemek için arkaplan alan bilgisinden kendi kendilerine yararlanamazlar. Bahsi geçen görevlerin yerine getirilmesi için bu bilgi ve verileri içeren sembolik akıl yürütme görevlerinin bir veri analizi sistemi ile donatılmış olması gerekir. Özetle, geleneksel veri analizi teknikleri, faydalı veri yorumlamalarını kolaylaştırır ve verilerin arkasındaki süreçle ilgili önemli bilgiler oluşturmaya yardımcı olur. Bu yorum ve görüşler veritabanlarını inşa edenler tarafından aranan nihai bilgidir. Ancak bu tür bilgiler bu araçlarla yaratılmaz, bunun yerine veri analizi ile elde edilmek zorundadır.

Bu kısıtlamaları aşacak olan yeni veri analizi araçlarına artan ihtiyacı karşılamak için araştırmacılar, makine öğrenmede geliştirilen fikirlere ve yöntemlere yöneldiler. Makine öğrenimi alanı, bu amaç için doğal bir fikir kaynağıdır, çünkü bu alanda araştırmanın özü gerçeklerden ve arka plan bilgisinden bilgi edinmek için hesaplamalı modeller geliştirmektir. Bu ve ilgili çabalar, sıklıkla veri madenciliği ve bilgi keşfi olarak adlandırılan yeni bir araştırma alanının ortaya çıkmasına neden olmuştur. "Veri madenciliği" ve "KDD" terimlerinin tam anlamıyla ilgili kafa karışıklığı oluşmuştur. KDD, 1995 yılında verilerin tüm verilerden çıkarılması sürecini tanımlamak için önerilmiştir. Bu bağlamda bilgi, veri elemanları arasındaki ilişkiler ve kalıplar anlamına gelir. Veri Madenciliği özel olarak KDD sürecinin keşif aşaması için kullanılmalıdır.

Veri ya da bilgi keşfi olarak da adlandırılan Veri Madenciliği kavramının şimdiye kadar pek çok tanımı yapılmıştır. Basit bir ifadeyle, kullanılabilir verileri daha büyük bir ham veri kümesinden elde etmek için kullanılan bir işlem olarak tanımlanır. Makine öğrenimi, istatistik ve veri tabanı sistemlerinin kesişiminde kullanılan yöntemleri içeren büyük veri kümelerindeki kalıpları keşfetme sürecidir [1]. Veri madenciliği, çeşitli kaynaklardan gelen mevcut veri hacminin ham verilerinden veri analizi ile önceden bilinmeyen ilginç örüntüleri ayıklar. Bu örüntüler giriş verilerinin bir özeti olarak görülebilir ve daha ileri analizler ile eyleme dönüştürülebilir iç görüler gibi faydalı bilgiler çıkarılır. Veri madenciliği bilgisayar bilimlerinin disiplinler arası bir alt alanıdır ve genel olarak akıllı yöntemler kullanılarak bir veri kümesinden bilgi ayıklamak ve bilgiyi daha etkin kullanım için anlaşılabilir bir yapıya dönüştürmek olan bir hedeftir [2, 3, 4]. Beklenmeyen ilişkileri bulmak için veri setlerini yeni yöntemlerle analiz etme sürecidir. Veri madenciliği yoluyla elde edilen ilişkilere ve özetlere genellikle verilerden örtük, bilinmeyen ve potansiyel faydalı bilgiler çıkaran modeller veya modeller olarak atıfta bulunulur. İstatistik, veri bilimi, veritabanı teorisi ve makine öğrenimi gibi birçok tekniği harmanlayan bir bilgisayar bilimi alt alanıdır.

Bilgisayar teknolojisinin çoğalması, yaygınlığı ve artan gücü, veri toplama, depolama ve etkileşimini arttırmıştır. Veri setlerinin büyüklüğü ve karmaşıklığı arttıkça,

doğrudan uygulamalı veri analizi, dolaylı, otomatik veri işleme ile giderek daha da artmaktadır.

Veri madenciliğinin kökleri klasik istatistik bilimi, yapay zeka (AI) ve makine öğrenmesi olmak üzere üç ana başlıkta toplanabilir. İstatistik Bilimi, veri madenciliği ile ilerleyen birçok teknolojinin temelidir. Regresyon analizi, standart dağılım, standart sapma, standart varyans, ayırt edici analiz, küme analizi ve güven aralığı bu temelleri oluşturur. Bunların hepsi veri ve veri ilişkilerini incelemek için kullanılır. Sezgisel yaklaşımın üzerine kurulu AI, insan düşüncesi benzerini işlemeyi istatistiksel yaklaşımların üzerine uygular. Makine öğrenmesi ise, istatistik ve AI'nın birleşimidir. AI'nın bir evrimi olarak kabul edilebilir, çünkü AI sezgiselliğini gelişmiş istatistiksel analizle birleştirir. Makine öğrenimi, bilgisayar programlarının çalıştıkları veriler hakkında bilgi edinmeye çalışır, böylece programlar çalışılan verinin nitelikleri temelinde farklı kararlar verir, temel kavramlar için istatistikleri kullanır ve hedeflerine ulaşmak için daha gelişmiş AI sezgileri ve algoritmaları ekler.

"Veri Madenciliği" terimsel kullanımının 1990'lı yıllarda ortaya çıkmış olması ve tarihinin adının yeni teknoloji haberleriyle sık sık anılması sebebiyle kısa bir süre önce başladığı düşünülür. 1763 yılında Bayes Teoremi ve 1805 yılında Regresyon Analizi gibi veri içindeki desen ve örüntüleri tanımlayan istatistiki yöntemlerle başlayan aslında uzun bir geçmişe sahip olan bir disiplindir.

1763 yılında Thomas Bayes'in ölümünden sonra bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösteren Bayes teoremini içeren makalesi yayınlamıştır.

1805 yılında Adrien-Marie Legendre ve Carl Friedrich Gauss, cisimlerin Güneş etrafındaki yörüngelerini (kuyruklu yıldızlar ve gezegenler) belirlemek için regresyonu uygulamıştır. Regresyon analizinin amacı değişkenler arasındaki ilişkileri tahmin etmektir ve bu durumda kullandıkları özel yöntem en küçük kareler yöntemidir. Regresyon, veri madenciliğindeki kilit araçlardan biridir.

1936 yılında büyük miktarda verinin toplanmasını ve işlenmesini mümkün kılan bilgisayar çağının başlangıcı olmuştur. 1936 tarihli bir makalede, Hesaplanabilir Sayılar Üzerine Alan Turing, günümüz bilgisayarları gibi hesaplamalar yapabilen bir Evrensel Makine fikrini sunmuştur. Günümüz bilgisayarı, Turing'in öncülük ettiği konseptler üzerine inşa edilmiştir.

1943 yılında Warren McCulloch ve Walter Pitts, bir sinir ağı için kavramsal bir model oluşturan ilk insanlar olmuştur. Sinirsel aktiviteye ilişkin fikirlerin mantıksal bir hesabı olan bir makalede, bir ağdaki bir nöron fikrini açıklanmıştır. Bu nöronların her biri girdileri almak, girdileri işlemek ve çıktı üretmek işlemlerini yapabilir.

1965 yılında Lawrence J. Fogel, evrimsel programlama uygulamaları için Decision Science, Inc. adında yeni bir şirket kurmuştur. Gerçek dünyadaki sorunları çözmek için özellikle evrimsel hesaplamayı kullanan ilk şirket olmuştur.

1970'li yıllarda gelişmiş veritabanı yönetim sistemleri ile terabayt ve petabayt verilerini depolamak ve sorgulamak mümkün olmuştur. Ek olarak, veri ambarları, kullanıcıların işlem odaklı bir düşünce biçiminden verileri daha analitik bir şekilde izlemelerine olanak sağlamıştır. Bununla birlikte, çok boyutlu modellerin bu veri ambarlarından sofistike içgörüleri çıkarmak henüz çok sınırlıdır.

1975 yılında John Henry Holland, Genetik algoritmaların çığır açan kitabı olan Doğal ve Yapay Sistemlerde Adaptasyon yazdı. Bu çalışma alanını başlatan, teorik temelleri sunan ve uygulamaları inceleyen kitaptır [5].

1980'lerde HNC, "Veritabanı Madenciliği" ifadesini markalaştırılmıştır. Ticari adıyla, DataBase Mining Workstation olarak tescillemiştir. Yapay sinir ağı modelleri oluşturmak için kullanılan bu araç artık mevcut değildir. Aynı zamanda, bu süreçte karmaşık algoritmalar konu uzmanlarının ilişkilerin ne anlama geldiğine dair bir neden belirlemesini sağladığı verilerden ilişkileri öğrenebilir.

1989 “Veritabanlarında Bilgi Keşfi” (KDD) terimi Gregory Piatetsky-Shapiro tarafından yazılmıştır. Aynı zamanda KDD adlı ilk atölyeyi kurulmuştur.

1990’lar Veri topluluğunda “Veri Madenciliği” terimi ortaya çıktı. Perakende şirketleri ve finans topluluğu, verileri analiz etmek ve müşteri tabanını artırma eğilimlerini tanımak, faiz oranlarındaki dalgalanmaları, hisse senedi fiyatlarını, müşteri talebini tahmin etmek için veri madenciliğini kullanılmaktadır.

1.2. Kullanıldığı Alanlar

Veri Madenciliği sektörel bazda Müşteri İlişkileri Yönetimi (CRM)/Müşteri Analitiği, bankacılık, doğrudan pazarlama, kredi puanlama, telekominikasyon, dolandırıcılık tespiti, satış, sağlık, finans, bilim, reklamcılık, e-ticaret, sigortacılık, web madenciliği, sosyal ağlar, ilaç sanayi, biyoteknoloji gibi çok önemli alanlarda kullanılmaktadır.

Veri Madenciliği, günümüzde ağırlıklı olarak tüketici odaklı, perakende, finansal, iletişim ve pazarlama organizasyonları olan şirketler tarafından, kurumsal karı arttırmak, işlem verilerini ayrıntılandırmak ve fiyatlandırmayı, müşteri tercihlerini ve ürün konumlandırmasını belirlemek, satışları etkilemek, müşteri memnuniyeti için kullanılmaktadır. Veri madenciliğinde, bir satıcı belirli müşteri segmentlerine hitap etmek için ürünler ve promosyonlar geliştirmek için müşteri alım satım kayıtlarını kullanabilir.

Veri madenciliği sağlık sistemlerini iyileştirmek için büyük potansiyele sahiptir. Bakımı iyileştiren ve maliyetleri düşüren en iyi uygulamaları belirlemek için veri ve analitik kullanılır. Araştırmacılar, çok boyutlu veritabanları, makine öğrenmesi, soft-computing, veri görselleştirme ve istatistik gibi veri madenciliği yaklaşımlarını kullanmaktadır. Her kategorideki hastaların hacmini tahmin etmek için veri madenciliği kullanılabilir ve bu sayede hastaların doğru yerde ve doğru zamanda uygun bakım almasını sağlayan süreçler geliştirilebilir. Veri madenciliği ayrıca sağlık sigortacılarının sahtekarlığı ve kötüye kullanımı tespit etmelerine yardımcı olabilir.

Pazar sepeti analizi, belirli bir ürün grubu satın alırsanız başka bir ürün grubu satın alma olasılığınızın yüksek olduğu teorisine dayanan bir modelleme tekniğidir. Bu teknik, perakendecinin bir alıcının satın alma davranışını anlamasını sağlayabilir. Bu bilgiler, satıcının alıcının ihtiyaçlarını bilmesine ve mağazanın düzenini buna göre değiştirmesine yardımcı olabilir. Farklı demografik gruplardan elde edilen sonuçlar farklı mağazalar arasında, farklı demografik gruplardaki müşteriler arasında karşılaştırmalı analiz yapılabilir.

Eğitimsel Veri Madenciliği (EVM) ise eğitim ortamlarından gelen verilerden bilgi keşfedecek yöntemler geliştirmekle ilgilenmektedir. EVM'nin amaçları, öğrencilerin gelecekteki öğrenme davranışını öngörmek, eğitim desteğinin etkilerini incelemek ve öğrenme hakkında bilimsel bilgiyi ilerletmek olarak tanımlanmaktadır. Veri madenciliği bir kurum tarafından doğru kararlar almak ve ayrıca öğrencinin sonuçlarını tahmin etmek için kullanılabilir. Sonuçlarla kurum neyin öğretileceğine ve nasıl öğretileceğine odaklanabilir. Öğrencilerin öğrenme örüntüleri yakalanabilir ve onlara öğretmek için teknikler geliştirmek için kullanılabilir.

Bilgi, bir üretim işletmesinin sahip olabileceği en iyi varlıktır. Veri madenciliği araçları, karmaşık üretim sürecindeki kalıpları keşfetmek için çok yararlı olabilir. Üretim Mühendisliğinde Veri Madenciliği, sistem düzeyinde tasarımda, ürün mimarisi, ürün portföyü ve müşteri ihtiyaçları verileri arasındaki ilişkileri çıkarmak için kullanılabilir. Ayrıca, ürün geliştirme süresini, maliyetini ve diğer görevler arasındaki bağımlılıkları tahmin etmek için de kullanılabilir.

CRM'de müşteri kazanma ve elde tutma, ayrıca müşteri sadakatini arttırma ve müşteri odaklı stratejileri uygulama ile ilgilidir. Bir müşteriyle uygun bir ilişkiyi sürdürmek için bir işletmenin veri toplaması ve bilgileri analiz etmesi gerekir. Veri madenciliğinin rol oynadığı yer burasıdır. Veri madenciliği teknolojileri ile toplanan veriler analiz için kullanılabilir. Müşteriyi elde tutmaya odaklanılacak yerin kafası karışmak yerine, çözüm arayanlar filtrelenmiş sonuçlar alır.

Dolandırıcılık eylemiyle milyarlarca dolar kaybediliği bilinen bir gerçektir. Geleneksel sahtekarlık tespit yöntemleri zaman alıcı ve karmaşıktır. Veri madenciliği, anlamlı modeller sağlamaya ve verileri bilgiye dönüştürmeye yardımcı olur. Geçerli ve yararlı olan herhangi bir bilgi bilgidir. Mükemmel bir sahtekarlık algılama sistemi tüm kullanıcıların bilgilerini korumalıdır. Denetlenen bir yöntem, örnek kayıtların toplanmasını içerir. Bu kayıtlar sahte veya sahte olmayan olarak sınıflandırılmıştır. Bu veriler kullanılarak bir model oluşturulmuştur ve kaydın sahte olup olmadığını belirlemek için algoritma yapılmıştır.

Bir kaynağın bütünlüğünü ve gizliliğini tehlikeye atacak herhangi bir işlem izinsizdir. İzinsiz girişi önlemek için yapılan savunma önlemleri, kullanıcı kimlik doğrulaması, programlama hatalarından kaçınma ve bilgi korumasını içerir. Veri madenciliği, anomali (aykırılık) tespitine bir odak seviyesi ekleyerek izinsiz giriş tespitini geliştirmeye yardımcı olabilir. Bir analistin, bir etkinliği ortak günlük ağ faaliyetlerinden ayırt etmesine yardımcı olur. Veri madenciliği ayrıca, problemle daha alakalı olan verilerin elde edilmesine yardımcı olur.

Bir suçluyu yakalamak kolaydır, ondan gerçeği çıkarmak zordur. Kolluk, suçları araştırmak, şüpheli teröristlerin iletişimini izlemek için madencilik tekniklerini kullanabilir. Bu dosya metin madenciliği de içeriyor. Bu işlem, genellikle yapılandırılmamış metin olan verilerde anlamlı kalıplar bulmaya çalışır. Önceki araştırmalardan toplanan veri örnekleri karşılaştırılmış ve yalan tespiti için bir model oluşturulmuştur. Bu model ile ihtiyaçlara göre süreçler oluşturulabilir.

Geleneksel pazar araştırması, müşterileri segmentlere ayırmamıza yardımcı olabilir, ancak veri madenciliği derinlere iniyor ve pazar etkinliğini artırıyor. Veri madenciliği, müşterileri farklı bir segmente hizalamaya yardımcı olur ve ihtiyaçları müşterilere göre uyarlayabilir. Pazar her zaman müşterileri elde tutmakla ilgilidir. Veri madenciliği, kırılganlığa dayalı bir müşteri segmenti bulmaya izin verir ve şirket onlara özel teklifler sunabilir ve memnuniyeti artırabilir.

Her yerde bilgisayarlı bankacılık ile yeni işlemlerle büyük miktarda veri üretilmesi bekleniyor. Veri madenciliği, işletme bilgileri ve piyasa fiyatlarında yöneticiler tarafından açıkça görülmeyen kalıplar, nedensellikler ve korelasyonları bularak işletme ve işletme problemlerinin çözülmesine katkıda bulunabilir çünkü hacim verileri çok büyüktür veya uzmanlar tarafından çok hızlı bir şekilde taranır. Yöneticiler bu bilgiyi daha iyi bölümlere ayırma, hedefleme, elde etme, elde tutma ve karlı bir müşteriyi sürdürme konusunda bulabilirler.

Kurumsal gözetim, bir şahsın veya grubun davranışlarının bir kurum tarafından izlenmesidir. Toplanan veriler çoğunlukla pazarlama amacıyla kullanılır veya başka şirketlere satılır, ancak devlet kurumlarıyla düzenli olarak paylaşılır. İşletmeler tarafından müşterileri tarafından arzu edilen ürünlerini uyarlamak için kullanılabilir. Veriler, arama geçmişini ve e-postalarını analiz ederek reklamların arama motoru kullanıcısına hedeflendiği Google ve Yahoo'daki hedefli reklamlar gibi doğrudan pazarlama amaçları için kullanılabilir.

Tarih, araştırmalarda devrimci değişimlere tanık olduğumuzu gösteriyor. Veri madenciliği, veri temizliği, veri ön işleme ve veri tabanlarının entegrasyonuna yardımcı olmaktadır. Araştırmacılar, veritabanında araştırmada herhangi bir değişiklik getirebilecek benzer verileri bulabilirler. Birlikte ortaya çıkan herhangi bir dizilimin tanımlanması ve herhangi bir aktivite arasındaki korelasyon bilinir. Veri görselleştirme ve görsel veri madenciliği bize verileri net bir şekilde sunar.

Kriminoloji, suçun özelliklerini tanımlamayı amaçlayan bir süreçtir. Aslında suç analizi, suçları ve suçlularla ilişkilerini araştırmayı ve tespit etmeyi içerir. Suç veri kümelerinin yüksek hacmi ve bu tür veriler arasındaki ilişkilerin karmaşıklığı kriminolojiyi veri madenciliği tekniklerini uygulamak için uygun bir alan haline getirmiştir. Metin tabanlı suç raporları kelime işlem dosyalarına dönüştürülebilir. Bu bilgiler suç eşleştirme işlemini gerçekleştirmek için kullanılabilir.

Veri Madenciliği yaklaşımları, veri bakımından zengin olduğu için Biyoinformatik için ideal görünmektedir. Madencilik biyolojik verileri, biyolojide ve tıp ve sinirbilim

gibi diğer ilgili yaşam bilimleri alanlarında toplanan büyük veri kümelerinden faydalı bilgilerin çıkarılmasına yardımcı olur. Biyoinformatiğe veri madenciliği uygulamaları gen bulma, protein fonksiyon çıkarımı, hastalık teşhisi, hastalık prognozu, hastalık tedavi optimizasyonu, protein ve gen etkileşimi ağ rekonstrüksiyonu, veri temizliği ve protein alt hücresel yerleşim tahminini içerir.

1.3. Etkileyen Etmenler ve Karşılaşılan Problemler

Veri madenciliği çok güçlü olmasına rağmen, uygulanması sırasında birçok zorlukla karşılaşmaktadır. Zorluklar performans, veriler, kullanılan yöntem ve tekniklerle vb. İlişkili olabilir. Veri madenciliği süreci, zorluklar veya sorunlar doğru bir şekilde tanımlanıp doğru bir şekilde belirlendiğinde başarılı olur.

1.3.1. Gürültülü ve eksik veri

Gerçek dünyadaki veriler heterojen, eksik ve gürültülüdür. Büyük miktarlardaki veriler normalde yanlış veya güvenilmez olacaktır. Bu problemler, verileri ölçen aygıtların hataları veya insan hataları nedeniyle olabilir. Bir perakende zincirinin 200 dolardan fazla harcayan müşterilerin e-posta kimliğini topladığını ve faturalandırma personelinin ayrıntılarını sisteme girdiğini varsayıldığında kişi, yanlış verilerle sonuçlanan e-posta kimliğini girerken yazım hatası yapabilir. Bazı müşteriler bile, eksik verilerle sonuçlanan e-posta kimliklerini ifşa etmeye hazır olmayabilir. Hatta sistem veya insan hatası nedeniyle veriler değişebilir. Tüm bunlar, madenciliği gerçekten zorlaştıran, gürültülü ve eksik verilerle sonuçlanmaktadır.

1.3.2. Dağıtılmış veri

Gerçek dünya verileri, genellikle dağınık bilgisayar ortamlarında farklı platformlarda depolanır. Veri tabanlarında, bireysel sistemlerde ve hatta internette olabilir. Temel olarak kurumsal ve teknik nedenlerden dolayı tüm verileri merkezi bir veri havuzuna getirmek pratik olarak çok zordur. Örneğin, farklı bölge ofisleri, verilerini depolamak için kendi sunucularına sahip olabilir, ancak tüm ofisleri (milyon terabayt) merkezi bir

sunucudaki tüm ofislerden depolamak mümkün olmayabilir. Bu nedenle, veri madenciliği, dağıtılmış verilerin madenciliğini sağlayan araçların ve algoritmaların geliştirilmesini gerektirir.

1.3.3. Karmaşık veri

Gerçek dünya verileri gerçekten heterojendir ve görüntüler, ses ve video, karmaşık veriler, zamansal veriler, mekânsal veriler, zaman serileri, doğal dil metni vb. İçeren multimedya verileri olabilir. Bu farklı türdeki verileri ele almak ve gerekli bilgileri elde etmek gerçekten zordur. Çoğu zaman, ilgili bilgileri elde etmek için yeni araçlar ve metodolojiler geliştirilmelidir.

1.3.4. Performans

Veri madenciliği sisteminin performansı esas olarak kullanılan algoritmaların ve tekniklerin verimliliğine bağlıdır. Tasarlanan algoritmalar ve teknikler işarete uymuyorsa, veri madenciliği sürecinin performansını olumsuz yönde etkileyecektir.

1.3.5. Arka plan bilgisinin birleşmesi

Arka plan bilgisi dahil edilebiliyorsa, daha güvenilir ve doğru veri madenciliği çözümleri bulunabilir. Tanımlayıcı görevler daha yararlı bulgular ortaya çıkarabilir ve öngörücü görevler daha doğru tahminler yapabilir. Ancak, arka plan bilgisini toplamak ve birleştirmek karmaşık bir süreçtir.

1.3.6. Veri görüntüleme

Veri madenciliği veri madenciliğinde çok önemli bir süreçtir, çünkü çıktıyı kullanıcıya öngörülebilir bir şekilde gösteren ana işlemdir. Elde edilen bilgiler, aslında neyi iletmek istediğinin tam anlamını iletmelidir. Ancak çoğu zaman, bilgiyi son kullanıcıya doğru ve anlaşılması kolay bir şekilde sunmak gerçekten zordur. Giriş

verileri ve çıktı bilgilerinin gerçekten karmaşık olması, çok etkili ve başarılı olması için veri görselleştirme tekniklerinin başarılı olması için uygulanması gerekir.

1.3.7. Veri gizliliği ve güvenliği

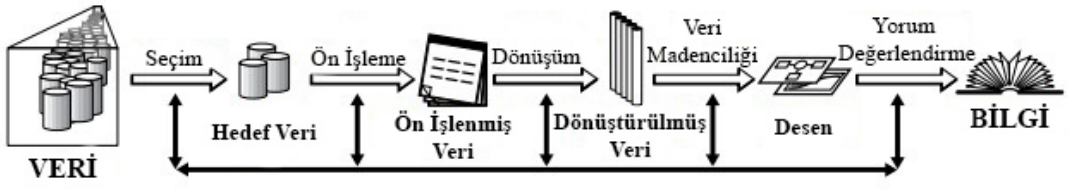
Veri madenciliği normalde veri güvenliği, gizlilik ve yönetişim açısından ciddi sorunlara yol açmaktadır. Örneğin, bir satıcı, satın alma ayrıntılarını analiz ettiğinde, satın alma alışkanlıkları ve müşterilerin tercihleri hakkında izinleri olmadan bilgi verir.

1.4. Veri Madenciliği Süreci ve Metotları

1.4.1. Veri madenciliği süreci

Bu bölümde KDD, Veri Madenciliği için Sektörler Arası Süreç (CRISP-DM) ve SEMMA metodolojilerine değinilecektir. KDD verilerdeki geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabilir kalıpları belirlemeye yönelik önemli süreçtir.

KDD işlemi gerekli herhangi bir ön işleme, alt örnekleme ve dönüşüm ile birlikte bir veri tabanı kullanarak, önlemlerin ve eşiklerin spesifikasyonuna göre bilgiyi elde etmek için veri madenciliği yöntemlerini kullanma sürecidir [6]. Bu süreç şekil 1.1'de gösterildiği gibi beş aşamada ele alınmıştır.



Şekil 1.1. KDD süreci [6]

Seçim aşamasında, bir hedef veri seti oluşturmaya veya keşfin gerçekleştirileceği bir değişkenler veya veri örnekleri alt grubuna odaklanmaya dayanır. Ön işleme aşamasında, tutarlı veri elde etmek için hedef veri temizliği ve ön işlemden yapılmasından oluşur. Dönüşüm aşamasında boyutsallığı azaltma veya dönüşüm

yöntemleri kullanarak verilerin dönüşümünden oluşur. Veri madenciliği aşamasında, Veri Madencidiliği hedefine (genellikle tahmin) bağlı olarak, belirli bir temsili biçimde ilgi kalıplarının aranmasından oluşur. Yorum ve Değerlendirme aşaması ise çıkarılan kalıpların yorumlanması ve değerlendirilmesinden oluşur.

CRISP-DM; veri madenciliği için sektörler arası süreç anlamına gelen bir veri madenciliği metodolojisidir. CRISP-DM DaimlerChryrler, SPSS ve NCR şirketlerinin birlikteliğinden oluşan bir konsorsiyumun çabaları sonucunda geliştirilmiştir. Bu metodoloji, bir veri madenciliği projesinin planlanmasında yapısal bir yaklaşım sunmaktadır. Bu sağlam ve kanıtlanmış bir metodolojidir. CRISP-DM süreci altı aşamadan oluşur.

SEMMA, örnekleme, keşfetme, değiştirme, modelleme, değerlendirme anlamına gelen bir kısaltmadır. En büyük istatistik ve iş zekâsı yazılımı üreticilerinden biri olan SAS Enstitüsü tarafından geliştirilen sıralı adımların bir listesidir. Genişleyen veri madenciliği alanında, endüstriden bağımsız olarak kullanıcıların veri madenciliği projelerine uygulayabilecekleri çeşitlendirilmiş ve yinelemeli veri madenciliği süreci için standart bir metodolojidir.

KDD, CRISP-DM ve SEMMA veri madenciliği süreçleri arasında yapılan karşılaştırma Tablo 1.1.'deki gibidir. Araştırmacıların ve veri madenciliği uzmanlarının çoğunun bildiği gibi, KDD süreç modeli eksiksiz ve doğru olarak görüldüğü için daha çok tercih edilmektedir. Buna karşılık CRISP-DM ve SEMMA süreçleri ise çoğunlukla şirket odaklı işletmelerde sıklıkla kullanılmaktadır. CRISP-DM ve SEMMA karşılaştırıldığında ise süreçler bakımından CRISP-DM SEMMA'ya kıyasla daha eksiksiz bulunmaktadır. Tüm bu süreç modelleri, veri madenciliğinin pratik senaryolara nasıl uygulayabileceklerini bilmeleri için insanlara ve uzmanlara rehberlik eder ve yardım eder [7].

Tablo 1.1. Veri Madenciliği süreçleri [6]

KDD	CRISP-DM	SEMMA
Uygulamanın Geliştirilmesi ve Anlaşılması	Problemi, İşi Kavrama	-
Hedef Veri Kümesi Oluşturma	Veriyi Anlama	Örnek
Veri Temizleme ve Ön İşleme		Keşif
Veri Dönüşümü	Veriyi Hazırlama	Modifiye etme
Veri Madenciliği	Modelleme	Modelleme
Yorum ve Değerlendirme	Değerlendirme	Değerlendirme
Keşfedilen Bilgiyi Kullanma	Yayınlama	-

1.4.2. Veri madenciliği metotları

Veri Madenciliğinde istatistiksel analizlere dayalı pek çok yöntem bulunmakla birlikte temel olarak üç gruba ayrılır. Bunlar Sınıflandırma, Kümeleme ve Birliktelik Kuralları olarak adlandırılır.

1.4.2.1. Sınıflandırma

Sınıflandırma, koleksiyondaki öğeleri hedef kategorilere veya sınıflara atayan bir veri madenciliği işlevidir. Sınıflandırmanın amacı, verilerdeki her durum için hedef sınıfı doğru bir şekilde tahmin etmektir. Bir öğrenme aşaması ve bir sınıflandırma aşamasından oluşan iki aşamalı bir süreçtir. Öğrenme adımında, bir sınıflandırma modeli oluşturulur ve sınıflandırma adımı, oluşturulan veriler için sınıf etiketlerini önceden yapılandırmak için kullanılır. Sınıflandırma, verilen eğitim verilerinin yardımıyla verileri sınıflandırır.

Bir karar ağacı, her sınıfın yorumlanmasının veya sınıflandırma kurallarının grafiksel bir tasviridir. Regresyon, bir değişkenin değeri, bir ilişkiden belirli bir sınıfa giden bir veri dizisini eşlemek yerine, dizgeye dayalı olarak tahmin edildiğinde yararlıdır. Bazı yaygın sınıflandırma algoritmaları ise karar ağacı, sinir ağları, lojistik regresyondur.

1.4.2.2. Kümeleme

Kümeleme, bir kümenin içinde bulunan nesnelerin yüksek benzerliğe sahip olacağı ve iki kümenin nesnelerinin birbirine benzemeyeceği sınıflar ve kümeler halinde bir veri grubunu organize etme tekniğidir. Kümeleme yöntemi de tıpkı sınıflandırma yöntemi gibi nesneleri bir veya daha fazla özellik ile gruplar halinde karakterize eden bir tür öğrenme yöntemidir. Bu süreçler benzer görünmesine rağmen veri madenciliği bağlamında aralarında temel bir fark bulunmaktadır. Sınıflandırma, verilen eğitim verilerinin yardımıyla verileri sınıflandırır. Öte yandan, kümeleme verileri sınıflandırmak için farklı benzerlik ölçütleri kullanır.

1.4.2.3. Birliktelik kuralları

Sık kullanılan örüntüleri, ilişkileri, korelasyonları veya nedenselliği bulmaya yarayan bir veri madenciliği metodudur. Market-Sepet Analizi olarak sıklıkla anılan Birliktelik Kuralları, örnek olarak müşterilerin “alışveriş sepetlerine” yerleştirdikleri farklı ürünler arasındaki ilişkileri ve ilişkileri bularak müşteri alışkanlıklarını anlamaya yarar. Bu yarar ise ilgili yönetimin bu analizleri doğru kullanması ile satış karını arttırabileceğini gösterir.

BÖLÜM 2. KAYNAK ARAŞTIRMASI

2.1. Alkollü İçecek Tüketimi Üzerine Yapılmış Sosyal Çalışmalar

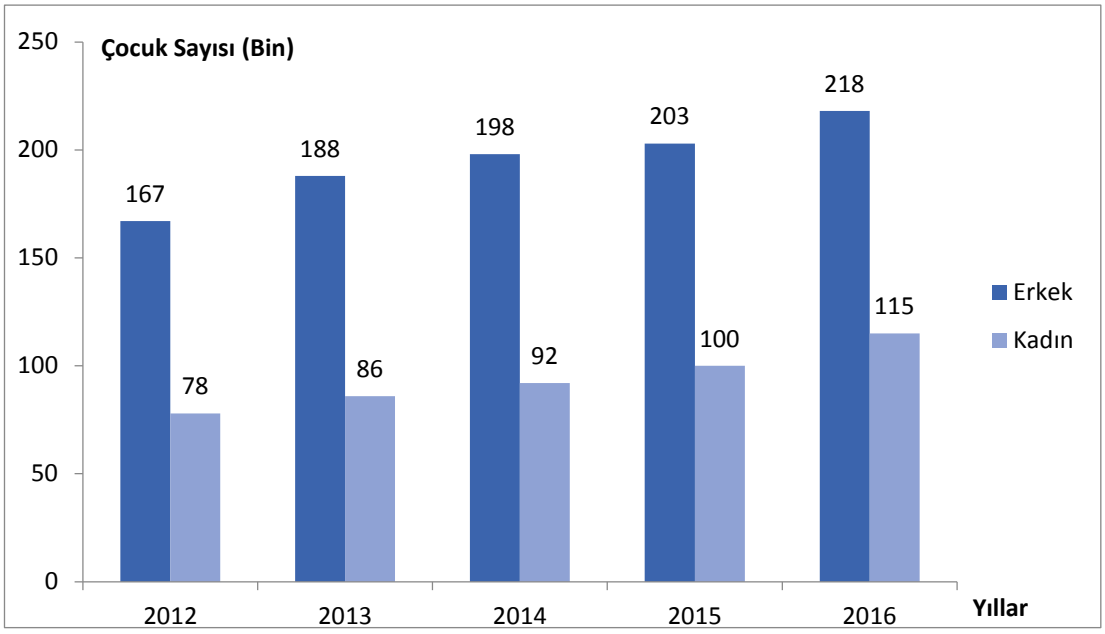
Alkol tüketimi yaş ortalamasının her geçen gün düştüğünü göz önüne aldığımızda gençler arasındaki yaygınlığı tüm dünyada bir sorun haline gelirken, gençlerin alkol tüketimine yönelmesinde aile, arkadaş ortamı, çevre gibi sosyal etkenlerin rol oynadığı da bilinmektedir. Alkollü içki tüketimi, genç nüfus ağırlıklı ve gelişmekte olan ülkemiz içinde çok önemli toplumsal bir sorundur. Bağımlılık yapıcı madde tüketim yaşı ise günden güne düşmektedir. Çoğu genç başta tecrübe etme, heves gibi amaçlarla alkol kullanmakta daha sonra bıraksa bile başta masum görünebilecek bu denemeler ileride alkol veya zararlı madde kullanımına hazırlık anlamına da gelmektedir [8]. Hiyerarşik sırada ilerleyen sigara, alkol ve madde bağımlılığı zararlı alışkanlıklarını yapılan çalışmalarda göstermektedir [9,10]. Bu nedenle zararlı alışkanlıklara eğilim göstermiş gençlerin demografik özelliklerinden faydalanarak gelecekte eğilim göstermesi ihtimali bulunan gençlerin tespit edilmesi, yönlendirilmesi ve kazanılması pek çok açıdan önem arz etmektedir.

Dünya Sağlık Örgütü iş birliğiyle 2004 yılında Amerika, Kanada ve Avrupa ülkelerinde yapılmış bir çalışma sonucunda; 11 yaşındaki öğrencilerin %15'inin, 13 yaşındaki öğrencilerin %40'ının ve 15 yaşındaki öğrencilerin %62'sinin hayatları boyunca en az bir kez sigara içtikleri sonucuna erişilmiştir [11].

2016 yılında yapılan bir çalışmaya göre ülkemizde 2004 ve 2015 yılları arasında kişi başına düşen en düşük tüketimin 1,31 litre saf alkol ile 2005 yılında; en yüksek tüketimin ise 1,55 litre saf alkol ile 2012 yılında olduğu belirlenmiştir [12]. 2015 yılındaki kişi başına tüketim ise 1,39 litre saf alkoldür. Dünya Sağlık Örgütü'nün 2018 yılında yayınladığı Küresel Alkol ve Sağlık Raporuna göre Türkiye'deki alkollü

içecek tüketiminin yaygınlığı batı toplumlara göre daha düşüktür [13].

Ülkemizde Türkiye İstatistik Kurumu'nun (TÜİK) 1 Ağustos 2017 yılında açıkladığı bir çalışmada suça sürüklenmiş çocukların %33'ünün bağımlılık yapan madde kullandığı ortaya çıkmıştır. TÜİK tarafından yapılan bu çalışmada uygulanan anket formu, 1997-2006 yılları arasında 27 ilde Jandarma Genel Komutanlığı'na ve Emniyet Genel Müdürlüğü'ne bağlı güvenlik birimlerinde uygulanmış olup, 2007 yılından itibaren 81 ilde uygulanmaya başlamıştır. "Güvenlik birimlerine suça sürüklenme nedeni ile getirilen 108 bin 675 çocuğun 36 bin 87'sinin bağımlılık yapan madde kullandığı görüldü. Bağımlılık yapan madde kullanan çocukların %84,5'ini 15-17 yaş grubu, %15'ini ise 12-14 yaş grubundaki çocuklar oluşturdu. Çocukların %72,9'unun sigara, %8,6'sının sigara ve alkol, %4'ünün sigara ve esrar, %2,9'unun esrar, %2'sinin ise sigara, alkol ve esrar kullandığı görüldü" [14].



Şekil 2.1. Güvenlik birimine getirilen çocuklar 2012-2016 [15]

2017 yılında Düzce'de üç farklı tipteki okuldan toplamda 2340 öğrenci üzerinde yapılan bir anket sonucunda sigara ve alkollü içecek tüketiminin frekansını ve bu kullanımları etkileyen faktörleri belirmek amacıyla bir çalışma yapılmıştır. Yapılan bu çalışma sonucunda; hayatları süresince en az bir kez sigara içen öğrencilerin oranı %35, çoğunlukla her gün bu tüketimde bulunan öğrencilerin oranı ise %20 olarak

bulunmuştur. Alkol tüketimi için ise bu oranlar %19 ve %2 olarak bulunmuştur [15].

15 - 22 yaşları arasındaki belirli sayıda öğrenciden elde edilmiş gerek kişisel gerek okul gerekse aile durum bilgilerini içeren veriler ışığında öğrenciyi alkol tüketimine itebilecek olası faktörler analiz edilecektir.

2.2. Literatür Taraması

Literatürde gençlerin alkol kullanımını etkileyen farklı faktörlerin etkisi çeşitli çalışmalarda incelenmiş ve sonuçlar elde edilmiştir. 1994 yılında yapılmış bir çalışmanın sonucunda ailenin çocuklarına normalde ayırdığından daha fazla zaman ayırmasının çocukların alkol ve hatta sigara kullanımını azalttığını belirtilmiştir [16].

2000 yılında Konya il merkezinde farklı düzeyde öğrenim görülen toplam 14 eğitim merkezinde 12 ve 21 yaşları arasındaki gençlerin dahil edildiği bir çalışma yapılmıştır. Yapılan bu çalışmada öğrencilere sosyolojik ve demografik özellikleri ile alkollü içecek kullanımı ilgili bilgiler içeren bir anket formu uygulanıp sonuçları SPSS programı ile lojistik regresyon analizi yöntemi ile değerlendirilmiştir. Bu çalışma gurubundaki gençlerin alkollü içecek kullanımları ile ailenin yüksek gelir seviyesine sahip olması durumuna, annenin çalışıyor olmasına ve annenin eğitim düzeyinin yüksek olmasına ilişkin anlamlı bir ilişki bulunmuştur [17].

2004 yılında Ankara'nın Çankaya ilçesinde yapılmış olan bir çalışmada; farklı sosyal ve ekonomik özelliklere sahip iki lisede okuyan %66'sı 15-16 yaşlarında olan toplam 380 öğrenci üzerinde zararlı madde kullanımını etkileyen faktörler analiz edilmiştir. Bilgisayar ortamında SPSS programı üzerinde Ki-Kare yöntemi kullanılmıştır. Çalışma sonucunda özellikle yüksek oranda devamsızlık gösteren, aile ilişkileri iyi olmayan ve ailesinde alkollü içecek ve sigara tüketimi olan öğrencilerin daha yüksek oranda bağımlılık yapıcı madde kullandığı tespit edilmiştir [18].

2008 yılında Portekiz'de aynı veri seti ile yapılan çalışmada ise eğitimde iş zekasını incelemek amacıyla karar ağaçları, rastgele orman, yapay sinir ağları ve destek vektör makineleri test edilmiştir. Çalışmada yapılan tüm deneyler, veri madenciliği tekniklerinin kullanımını kolaylaştıran R ortamı için açık kaynak kodlu bir kütüphane olan RMiner kullanılarak yapılmıştır. Çalışma sonucunda öğrencinin daha önceki dönemlerdeki başarısının gelecekte başarılarını etkilediği bulunmuştur. Ayrıca öğrencilerin alkol tüketimi, arkadaşlarıyla dışarıda vakit geçirmesi gibi sosyal değişkenlerinde bu sonucu pekiştirdiği belirtilmiştir [19].

2014 yılında Bartın ilinde ortaöğretim düzeyindeki öğrencilerde madde bağımlılığı durumunun aldığı halin ve görülen yaygınlık düzeyinin saptanması amacıyla bir çalışma yapılmıştır. Aynı zamanda bu çalışma madde bağımlılığı ile öğrencilerin farklı özelliklerinin birbirini etkileyip etkilemediği konusu da gözlemlenmiştir. Toplamda dokuz farklı ortaöğretim kurumundan rastgele seçilen 545 öğrencinin anket sorularına verdikleri cevaplardan bir veri seti elde edilmiştir. Değişkenler arasındaki bağımlılığın anlamlılık açısından test edilmesi için Pearson Ki-Kare tesi kullanılmıştır. Çalışma sonucunda genel manada sırasıyla; öğrencilerin sigara içmeyi merak etme dürtüleri, arkadaş etkisinde kalmaları ve yalnızlık sebepleri ise sigara içmelerinin en ciddi sebepleri olarak belirlenmiştir [20].

BÖLÜM 3. MATERYAL VE YÖNTEM

3.1. Materyal

Bu çalışmada kullanılmış olan, internet erişimine açık "UCI Machine Learning Repository" sitesinden alınan Student Alcohol Consumption veri seti içeriği Portekiz'de 2005-2006 yılında eğitim gören devlet okulu öğrencilerine aittir [21]. Verilerin toplanmasında öğrencilerin okul raporları ve raporlardaki eksikliklerin tamamlanmasında, uzmanlarca gözden geçirilen 37 soruluk anket formu uygulanmıştır. Çalışmada, matematik sınıfı öğrencilerine ait 1044 veri kullanılmıştır. Verilere ilişkin detaylı bilgi aşağıdaki Tablo 3.1.'de gösterilmiştir.

Tablo 3.1. Veri set içeriği

Sütun Adı	Açıklama	Veri Tipi	İçerik
okul	Öğrencinin okulu	binary	"GP" - Gabriel Pereira veya "MS" - Mousinho da Silveira
cinsiyet	Öğrencinin cinsiyeti	binary	"F" - kadın veya "M" - erkek
yas	Öğrencinin yaşı	numeric	numeric: 15 ila 22 arası
adres	Öğrencinin ev adresi türü	binary	"U" - kentsel veya "R" - kırsal
aile_uyesayisi	Ailedeki üyelerin sayısı	binary	"LE3" - 3 veya daha düşük veya "GT3" e eşit - 3'den büyük
ebeveyn_birliktelik	Ebeveynin birlikte yaşama durumu	binary	"T" birlikte yaşamak veya "A" - ayrı
A_egitim	Annenin eğitimi	numeric	0 - yok, 1 - ilköğretim (4. sınıf), 2-5 ila 9. sınıf, 3 - orta öğretim veya 4 - yükseköğretim
B_egitim	Babanın eğitimi	numeric	0 - yok, 1 - ilköğretim (4. sınıf), 2-5 ila 9. sınıf, 3- orta öğretim veya 4 - yükseköğretim

Tablo 3.1. (Devamı)

A_is	Annenin işi	nominal	“öğretmen”, “sağlık” bakımı ile ilgili, sivil “hizmetler” (örneğin idari veya polis), “at_home” veya “diğer”
B_is	Babanın işi	nominal	“öğretmen”, “sağlık” bakımı ile ilgili, sivil “hizmetler” (örneğin idari veya polis), “at_home” veya “diğer”
okul_tercihsebep	Bu okulun seçilme nedeni	nominal	“ev” e yakın, okul “itibar”, “ders” tercihi veya “diğer”
veli	Öğrencinin velisi	nominal	“anne”, “baba” veya “diğer”
yol_sure	Evden okula gidiş dönüş süresi	numeric	1 - <15 dakika, 2 - 15 - 30 dakika, 3 - 30 dakika - 1 saat veya 4 -> 1 saat
calisma_sure	Haftalık çalışma süresi	numeric	1 - <2 saat, 2 - 2 - 5 saat, 3-5 - 10 saat veya 4 -> 10 saat
zayif	Geçmiş sınıf başarısızlıklarının	numeric	1 <= n <3 ise n, başka 4
burs	Ekstra eğitim desteği	binary	Evet veya hayır
aile_destegi	Aile içi eğitim desteği	binary	Evet veya hayır
ders_ucretiodemesi	Adet ders ücreti ödemesi	binary	Evet veya hayır
aktivite	Müfredat dışı etkinlikler	binary	Evet veya hayır
anaokulu	Anaokuluna devam etti	binary	Evet veya hayır
master_istegi	Yükseköğrenim görmek istiyor	binary	Evet veya hayır
internet_erisimi	Evde internet erişimi	binary	Evet veya hayır
romantic	Romantik bir ilişkisi var mı?	binary	Evet veya hayır
aile_iliskikalitesi	Aile ilişkileri kalitesi	numeric	1’den - çok kötü - 5’e - mükemmel
bos_zaman	Okul sonrası serbest zaman- boş	numeric	1’den - çok düşükten 5’e - çok yüksek
ark_disaridavakit	Arkadaşlarıyla dışarı çıkıyor	numeric	1’den - çok düşükten 5’e - çok yüksek
Dalc	Günlük alkol tüketimi	numeric	1’den - çok düşükten 5’e - çok yüksek

Tablo 3.1. (Devamı)

Walc	Hafta sonu alkol tüketimi	numeric	1’den - çok düşükten 5’e - çok yüksek
saglik	Güncel sağlık durumu	numeric	1’den - çok düşükten 5’e - çok yüksek
devamsizlik	Okul devamsızlıkları	numeric	0 ila 93 arası
G1	İlk dönem notu	numeric	0 - 20 arası
G2	İkinci dönem notu	numeric	0 - 20 arası
G3	Final notu	numeric	0 - 20 arası

3.2. Yöntem

3.2.1. Veri seti işlemleri

Veri madenciliği araçları için üzerinde işlem yapılacak olan veri setini belirli formatlara sahip olmalıdır. Veri setinin çalışmaya uygun hale getirilmesi amacıyla veri seti bir önişleme tutulur. Bu işlemler sırasıyla eksik, gürültülü ve tutarsız olan verileri iyileştirme amacıyla yapılan veri temizleme işlemi, veri birleştirme işlemi, veri dönüştürme işlemidir. Uygulanacak modele katkısı olmayacağı düşünülen normal dağılım göstermeyen ve modeli gereksiz yere yorup modelin yönetimini zorlaştırabilecek alanların tespiti ve eliminasyonunun yapılmış ve faktör analizi ile de kontrolü sağlanmıştır.

Faktör analizi yapılmadan önce veri setinin korelasyon matrisinde bazı korelasyonların veri setinden çıkarılması gerekir. Bu değişkenlerin korelasyonları %30′dan küçüktür. Bu çıkarma işlemi ile veri seti faktör analizine daha uygun bir hale gelir. Bundan sonraki adımda ise kısmi korelasyon katsayılarına bakılıp bu katsayılar yüksek ise, veri seti iyi temsil edilemeyeceğinden faktör analizinin uygulanmaması gerektiği anlaşılacaktır. Faktör analizinin uygunluğunun araştırılması için gerekli test yaklaşımı literatürde mevcuttur. Bu testte korelasyon matrisinin birim matrise eşit olup olmadığı sınanır. Bartlett küresellik testi olarak adlandırılan bu testte, verilerin çok değişkenli

normal dağılan ana kütleden geldiği örneklerde geçerli olup test sonucunda anlamlılık %5′ten büyük çıkarsa faktör analizi uygulanmamalıdır.

Kaiser-Meyer-Olkin (KMO) Testi, verilerinizin Faktör Analizi için ne kadar uygun olduğuna dair bir ölçektir. Modeldeki her değişken için örnekleme yeterliliğini ve modelin tamamını ölçen bu istatistik, ortak varyans olabilecek değişkenler arasındaki varyans oranının bir ölçütüdür. Oran ne kadar düşük olursa, veriler o kadar uygun olur. KMO 0 ile 1 arasında bir değer döndürür. KMO değerleri 0.6'dan küçük olduğunda örneklemin yeterli olmadığını ve düzeltici önlemlerin alınması gerektiğini göstermektedir. KMO testi formülü aşağıdaki denklem kullanılarak (Denklem 3.1) hesaplanır.

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} u_{ij}^2} \quad (3.1)$$

Burada R = $[r_{ij}^2]$ korelasyon matrisini, U = $[u_{ij}^2]$ ise kısmi kovaryans matrisini ifade eder.

Bartlett testi ise aynı KMO gibi değişkenler arası ilişki gücünü ölçümler. Sağlıklı bir faktör analizinin yapılabilmesi için KMO>0.6 ve Bartlett sig<0.05 şeklinde değerler elde etmek yeterlidir.

3.2.2. Kullanılan algoritmalar

Çalışmada, SPSS ve Clementine programları aracılığı ile Veri Madenciliği tekniklerinden "Lojistik Regresyon" ve en yaygın kullanılan karar ağaçlarından C 5.0 algoritması kullanılmıştır.

Kavramsal açıdan bakıldığında regresyon analizi değişkenlerin arasında fonksiyonel ilişkilerin araştırılması için geniş bir yelpazede kendine yer edinmiş istatistiksel bir yöntemdir [22]. Çalışma kapsamında alkollü içecek tüketiminin özellikle erken yaştaki kullanımının önlenmesi için bu tüketimin çeşitli değişkenler ve özelliklerle

ilişkilendirilmesi sağlanmıştır. Birden fazla sayıdaki değişkenin ilişkisini incelenmesi ve bunun bir model ile ifade edilmesini sağlayan bu istatistiksel analiz sayesinde bağımsız değişkenlerin aldığı değerlere göre bağımlı değişkenin alacağı değerlerin tahmini elde edilir. Regresyon analizinde bağımlı değişken ile arasında ilişki bulunmayan bağımsız değişken veya değişkenler modele dahil edilmez. Bağımsız değişkenlerin p-değerleri hesaplanır. P-değeri bağımsız değişkenin etkisinin olmaması durumu için sıfır hipotezin test edilmesini sağlar. Sıfır hipotezin reddedileceğini göstereceği için P değerinin 0,05'den düşük olması tercih edilir. P değeri bu değerden büyük bir değere sahipse bağımsız değişkenin model içinde istatistiki olarak bir önemi olmayacağı anlaşılır.

Lojistik regresyon, belli bağımsız değişkenlerle kategorik bağımlı değişkendeki değişimi tahmin edip öngörebilme işlemidir. Lojistik Regresyon metodunda Maksimum Olabilirlik Yöntemi (Method of Maximum Likelihood) kullanılır. Bu yöntemde uygun sınıf üyelikleri tahmini ve bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkilerde araştırılmış olur.

Temel anlamıyla karar ağacı, bir kök düğüm, dallar ve yaprak düğümleri içeren bir yapıdır. Her dahili düğüm bir öznitelik üzerinde bir testi temsil eder. Her yaprak düğümü bir sınıfı temsil eder. Bir karar ağacı herhangi bir alan bilgisi gerektirmediği için yapısı itibarı ile anlaşılması kolaydır. Öğrenme ve sınıflandırma adımları basit ve hızlıdır. Veri kümesinin en iyi özelliğini ağacın köküne yerleştirilir, eğitim seti alt kümelere ayırılır. Alt kümeler, her alt kümenin bir öznitelik için aynı değerde veri içerecek şekilde yapılmalıdır. Ağacın tüm dallarında yaprak düğümleri bulana kadar her alt kümede 1. ve 2. adımları tekrarlanır.

Entropiye dayalı ID3 algoritmasının eksiklerini gidermek için geliştirilen C 4.5 algoritmasının ileri bir versiyonu olan C 5.0 algoritması doğruluğu arttırılmış bir tekniktir. Daha düşük hata oranlarına sahipir. C 5.0 algoritması bellek verimliliği sağlamaya, basit ağaç oluşumuna, kural tabanlı modellemelere ve yardımcı olmayan özelliklerin kaldırılmasına otomatik olarak izin verir [23].

BÖLÜM 4. ARAŞTIRMA BULGULARI

Çalışma grubunda ankete katılan 15-22 yaş arasındaki 1044 öğrenciden 591'i (%56,6) kız ve yaş ortalamaları 16.7 iken, 453'ü (%43,4) erkek ve yaş ortalamaları 16.6'dır. Kız öğrencilerden %21'inin, erkek öğrencilerden %44'ünün riskli alkol kullanıcısı kategorisine alınabileceği tespit edilmiştir.

Veri setinde devamsızlığı 5'in altında olan öğrenciler "1" ile daha fazla devamsızlığa sahip olan öğrenciler ise "2" ile kodlanmıştır. Dalc ve Walc sütunlarından haftalık alkol tüketimi hesaplanarak yedi günlük alkol kullanımı üzerinden düzenli olarak haftada en az iki ve üzeri oranında alkol kullanımı gerçekleştiren öğrenciler riskli grup olarak adlandırıldı ve RiskALC kategorisi oluşturulmuştur. Veri seti üzerinde "string" şeklinde yer alan veriler ilgili metotların üzerinde çalışabilmesi amacıyla "numeric" ifadelere dönüştürülmüştür.

Tablo 4.1.' e göre faktörlerin ortak varyanslılık değerleri incelendiğinde hiçbir faktörün faktör yükü 0.30'un altında olmadığı için, faktörlerden herhangi biri faktör analizinden çıkarılarak, faktör analizi yenilenmemiştir. Tablo 4.2.'de ise KMO>0.6 ve Bartlett sig<0.05 çıktığı görülmektedir.

Tablo 4.1. Faktör yüklerinin gösterimi

	Initial	Extraction
school2	1,000	,591
cinsiyet	1,000	,622
yas	1,000	,574
adres	1,000	,642
aile_uyesayisi	1,000	,552

Tablo 4.1. (Devamı)

ebeveyn_birliktelik	1,000	,649
A_egitim	1,000	,723
B_egitim	1,000	,652
A_is	1,000	,509
B_is	1,000	,539
okul_tercihsebep	1,000	,564
veli	1,000	,588
yol_sure	1,000	,587
calisma_sure	1,000	,442
zayif	1,000	,564
burs	1,000	,655
aile_destegi	1,000	,508
ders_ucretiodemesi	1,000	,580
aktivite	1,000	,597
anaokulu	1,000	,474
master_istegi	1,000	,409
internet_erisimi	1,000	,414
romantic	1,000	,575
aile_iliskikalitesi	1,000	,544
bos_zaman	1,000	,599
ark_disaridavakit	1,000	,708
saglik	1,000	,557
devamsizlik	1,000	,386

Tablo 4.2. KMO ve Bartlett testi sonucu

KMO Örnekleme Yeterliliğinin Ölçümü		,646
Bartlett Küresellik Testi	**Ki-Kare**	2441,077
	df	276
	Sig.	,000

Temel Bileşen Analizine göre oluşturulan faktörlerin yük gösterimi (Tablo 4.3.) ve KMO&Bartlett test sonuçları, Maksimum Olabilirlik Yöntemine göre tekrar incelendiğinde Tablo 4.4'te görüldüğü üzere Ki-Kare sonucu 1275 ve df (serbestlik derecesi) değeri ise 378 olarak bulunmuştur. Bu değerler birbirlerine oranlandığında sonuç 3.37 çıkmaktadır. Bu sonucun 5'ten küçük olması uyum indeksleri açısından yüksek oranda kabul edilebilir olduğunu göstermiştir.

Tablo 4.3. Maksimum Olabilirlik Yöntemine göre faktör yüklerinin gösterimi

	Initial	Extraction
school2	1,000	,529
cinsiyet	1,000	,602
yas	1,000	,613
adres	1,000	,654
aile_uyesayisi	1,000	,652
ebeveyn_birliktelik	1,000	,609
A_egitim	1,000	,733
B_egitim	1,000	,674
A_is	1,000	,592
B_is	1,000	,550
okul_tercihsebep	1,000	,640
veli	1,000	,564
yol_sure	1,000	,615
calisma_sure	1,000	,575
zayif	1,000	,545
burs	1,000	,717
aile_destegi	1,000	,508
ders_ucretiodemesi	1,000	,456
aktivite	1,000	,583
anaokulu	1,000	,661
master_istegi	1,000	,551
internet_erisimi	1,000	,492
romantic	1,000	,730
aile_iliskikalitesi	1,000	,550
bos_zaman	1,000	,576
ark_disaridavakit	1,000	,712
saglik	1,000	,662
devamsizlik	1,000	,642

Tablo 4.4. Maksimum Olabilirlik Yöntemine göre KMO ve Bartlett testi sonucu

KMO Örnekleme Yeterliliğinin Ölçümü		,610
Bartlett Küresellik Testi	**Ki-Kare**	1275,673
	df	378
	Sig.	,000

Bu çalışmada bağımlı değişken olan öğrencilerin alkol tüketimini sonuç ölçütü olarak alıp, bağımsız değişken olarak Faktör Yüklerini gösteren tabloda adı geçen faktörler seçilmiştir.

“Cinsiyet”, “ebeveyn_birliktelik”, “A_egitim”, “B_egitim”, “A_is”, “B_is”, “burs”, “aile_destegi”, “anaokulu”, “master_istegi”, “romantic”, “aile_iliskikalitesi”, “ark_disaridavakit”, “devamsizlik” kategorileri giriş “RiskALC” kategorisi hedef olarak belirlenerek çalışılan C 5.0 algoritması sonucunda toplam 11 tanesi ‘riskli’ 4 tanesi ‘risksiz’ sonuç veren kurala ulaşılmıştır. Bu kurallar tablo 4.5’te gösterildiği gibidir.

Tablo 4.5. C 5.0 Karar ağacı kural sonuçları

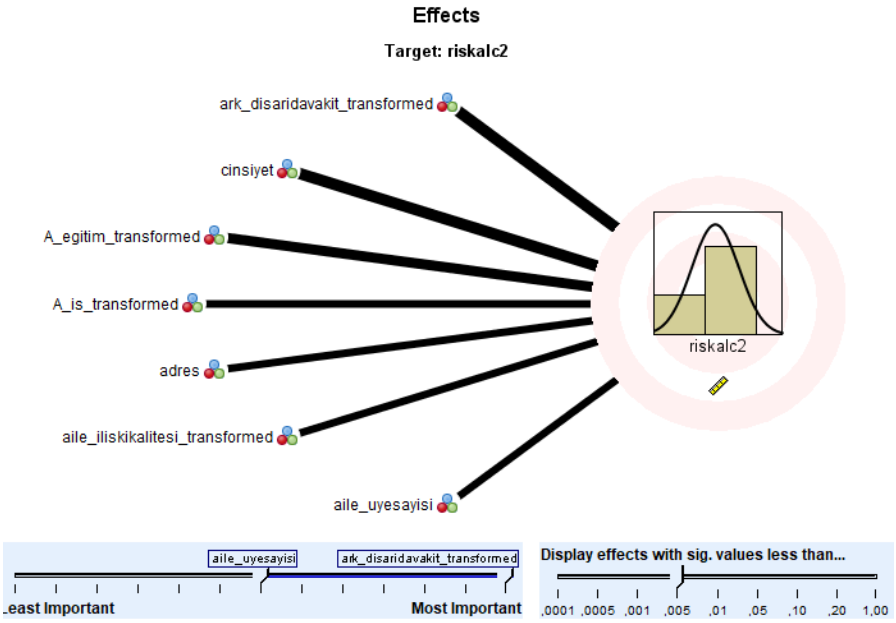
Riskli	
Rule 1 for Riskli (11; 0,923)	if cinsiyet = 1,000 and A_is = 1,000 and burs = 2,000 and ark_disaridavakit = 4,000 then Riskli
Rule 2 for Riskli (10; 0,917)	if cinsiyet = 1,000 and A_is = 5,000 and aile_destegi = 1,000 and aile_iliskikalitesi = 4,000 and ark_disaridavakit = 3,000 then Riskli
Rule 3 for Riskli (7; 0,889)	if cinsiyet = 1,000 and aile_destegi = 1,000 and master_istegi = 2,000 and ark_disaridavakit = 3,000 then Riskli
Rule 4 for Riskli (31; 0,848)	if cinsiyet = 1,000 and aile_destegi = 2,000 and master_istegi = 2,000 then Riskli
Rule 5 for Riskli (4; 0,833)	if cinsiyet = 1,000 and A_is = 5,000 and aile_destegi = 1,000 and aile_iliskikalitesi = 3,000 and ark_disaridavakit = 3,000 then Riskli
Rule 6 for Riskli (4; 0,833)	if cinsiyet = 1,000 and A_egitim = 1,000 and B_egitim = 1,000 and A_is = 3,000 and ark_disaridavakit = 4,000 then Riskli

Tablo 4.5. (Devamı)

Rule 7 for Riskli (53; 0,818)	if cinsiyet = 1,000 and romantic = 2,000 and ark_disaridavakit = 5,000 then Riskli
Rule 8 for Riskli (65; 0,791)	if cinsiyet = 1,000 and anaokulu = 1,000 and ark_disaridavakit = 5,000 then Riskli
Rule 9 for Riskli (19; 0,714)	if cinsiyet = 1,000 and B_egitim = 3,000 and burs = 2,000 and ark_disaridavakit = 4,000 then Riskli
Rule 10 for Riskli (24; 0,654)	if cinsiyet = 1,000 and B_egitim = 2,000 and burs = 2,000 and master_istegi = 1,000 and ark_disaridavakit = 4,000 then Riskli
Rule 11 for Riskli (93; 0,526)	if cinsiyet = 1,000 and ark_disaridavakit = 4,000 then Riskli

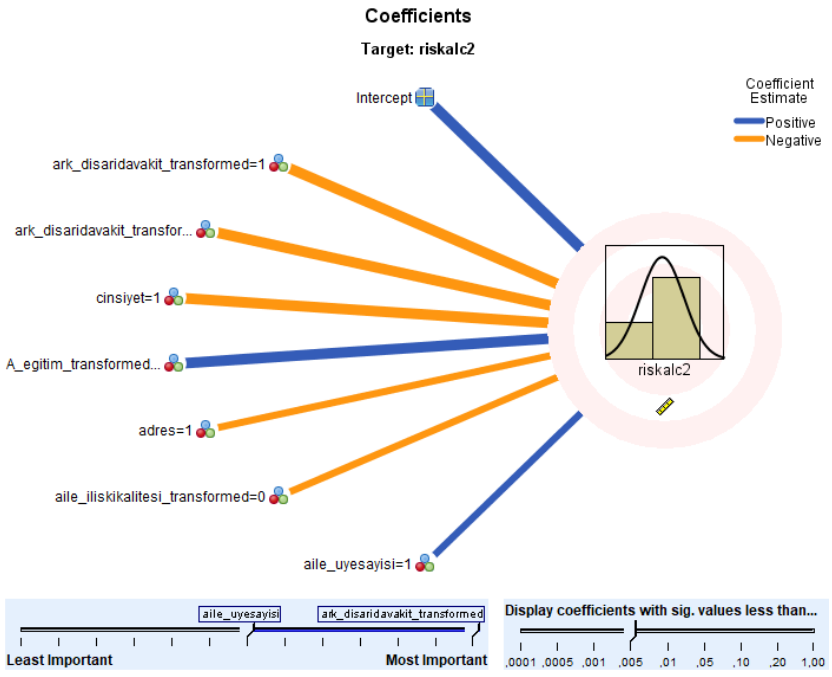
Bu sonuçlara göre Riskli kategoride oluşan kuralları yorumladığımızda kız öğrencilerden, annesi ev hanımı olanların, okul için burs desteği almayan ve arkadaşları ile ortalamanın üzerinde vakit geçirenlerin “Riskli” grupta çıkması oranı %92,3 bulunmuştur. Kız öğrencilerden master yapmak istemeyen ve aile desteği almayan ların “Riskli” grupta çıkması oranı ise %84,8 çıkmıştır. Bir diğer “Riskli” bölge kuralı ise kız öğrencilerden, anne ve babası orta öğretim mezunu, annesi “diğer” kategorisinde çalışan ve arkadaşları ile dışarıda normal süreler içinde vakit geçirenlerin “Riskli” grupta çıkmasının doğruluk olasılığı %83,33 olarak bulunmuştur.

Lojistik Regresyon sonuçları incelendiğinde ise elde edilen sonuçlar Şekil 4.1. ve Şekil 4.2.’deki gibidir.



Şekil 4.1. Dış faktörlerin riskALC'ye etkileri

Şekil 4.1.'de görüldüğü üzere veri setinde bahsi geçen gençlerin alkollü içecek tüketimin sırasıyla "arkadaşları ile dışarıda geçirdikleri vakit", "cinsiyet", "annenin eğitimi", "annenin işi", "adres" yani ikamet edilen bölge bakımından kentsel ya da kırsal alanda yaşama durumu, "aile arasındaki ilişki kaliteleri" ve "ailenin üye sayısı" etkilemektedir.



Şekil 4.2. Dış faktörlerin riskALC'ye pozitif veya negatif etkileri

Şekil 4.1.'deki sonuçlara ek olarak Şekil 4.2.'de görüldüğü üzere öğrencinin alkollü içki tüketimine negatif etkisi olan toplamda dört faktör gözlemlenmiştir. "Arkadaşları ile dışarıda geçirdikleri vakit" iki farklı değerde etkilemek üzere, "cinsiyet", "adres" ve "aile arasındaki ilişki kalitesi" dış faktörlerinin gencin alkol tüketimine negatif etkisi olduğu gözlenmiştir. Arkadaşları ile geçirdiği vakit düşük olan öğrencilerin alkollü içecek tüketme ihtimali artmaktadır. Pozitif etki gösteren dış faktör sayısı ise ikidir. Bunlar ise "annenin eğitimi" ve "aile üye sayısı" olmuştur. "Arkadaşları ile dışarıda geçirdikleri vakit" en yüksek değerleri aldığında, "aile arasındaki ilişki kalitesi" en düşük değeri aldığında, "adresi" kentsel bir alanda olduğunda ve öğrencinin "cinsiyeti" ise erkek olduğunda bu durumların öğrencinin alkol tüketimini arttıracağı ortaya çıkmıştır.

"Annenin eğitimi" düşük olmadığında ve "aile üye sayısı" kategorisinde ise kalabalık aile olmaması risk grubunu pozitif yönde etkilemektedir. Bir diğer ifadeyle bu durumda öğrencinin alkollü içecek tüketme ihtimali düşüktür. "cinsiyet", "annenin eğitimi", "annenin işi", "kentsel ya da kırsal alanda yaşama durumu", "aile arasındaki ilişki kalitesi" ve "ailenin üye sayısı" etkilemektedir.

Şekil 4.3'te gösterilen şekilde uygulanan C5.0 sınıflandırma algoritması ile oluşan karar ağacı yapısı da faktör analizi ve lojistik regresyon işlemlerinin bulgularıyla örtüşen sonuçlar vermiştir.

Şekil 4.3. Veri setine C5.0 algoritmasının uygulanması

Bu sonuçlara bakıldığında algoritma erkek öğrenciler için ortalamanın üzerinde arkadaşlarıyla dışarıda vakit geçirenleri ve erkek öğrencilerden ortalama veya altında arkadaşlarıyla dışarıda vakit geçiren, ailesi ile arasındaki ilişki ve iletişim kalitesi düşük olanları riskli olarak sınıflandırılmıştır. Yine erkek öğrenciler için dışarıda

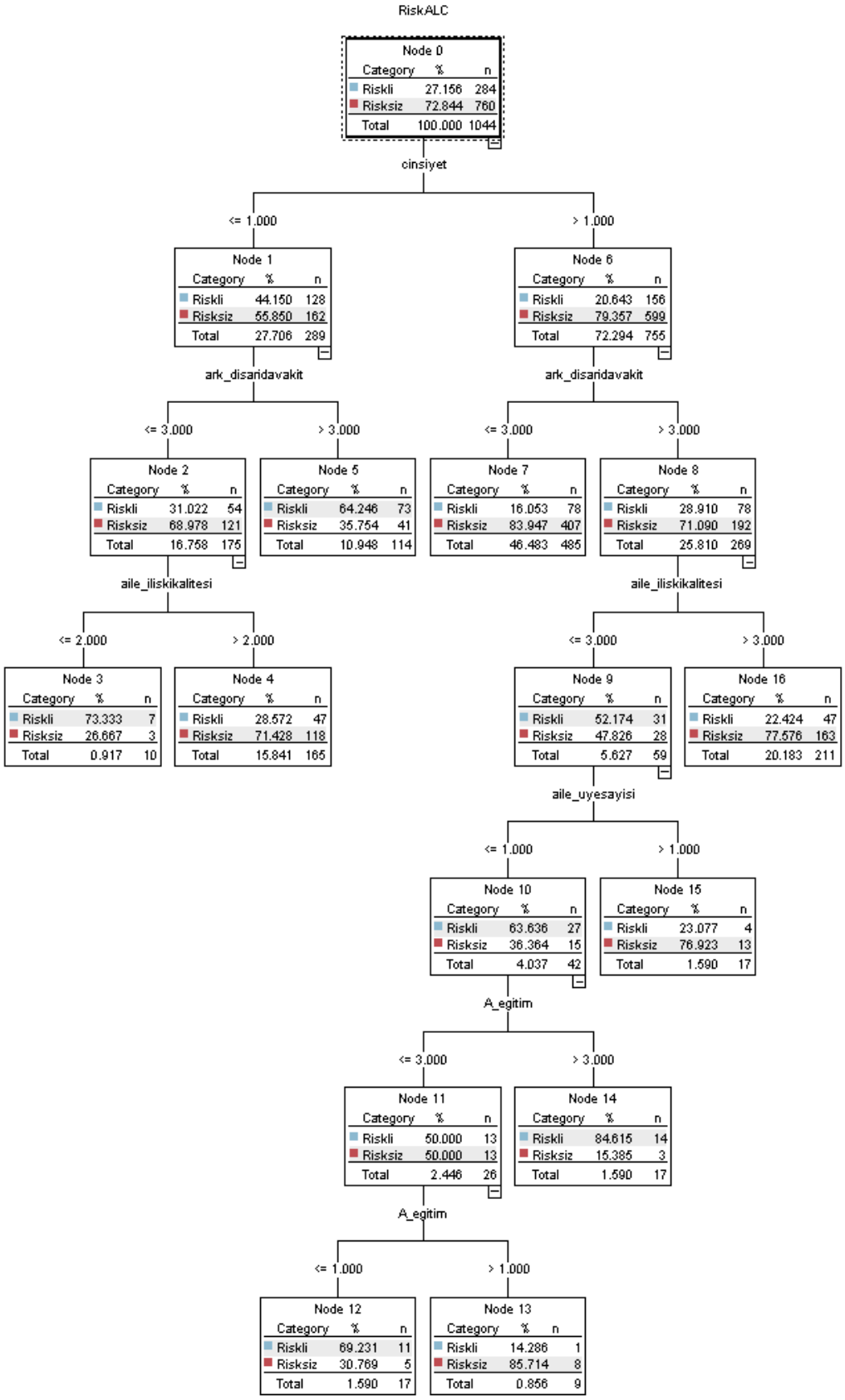
arkadaşlarıyla ortalamanın altında vakit geçirenlerden ailesi ile arasındaki ilişki kalitesi yüksek olanlar risksiz olarak sınıflandırılmıştır.

Kız öğrenciler için ise dışarıda arkadaşlarıyla ortalamanın üzerinde vakit geçirenlerden aile üyeleri arasındaki iletişim kalitesi ortalamanın altında kalanlar ve tek kardeş olanlar riskli grupta yer almıştır. Bu daha önce bahsi geçen çalışmadaki [17] yalnızlık hissinin madde kullanımına olumsuz etkisini desteklemektedir.

Algoritmanın sonucu model olarak aşağıdaki gibi özetlenmiştir. Şekil 4.4'te ise grafiksel anlatımla ağaç yapısı olarak gösterilmiştir. Buradaki n örneklem sayısını ifade etmektedir.

- cinsiyet = 1 [Mode: Risksiz]
    - ark_disaridavakit <= 3 [Mode: Risksiz]
        - aile_iliskikalitesi <= 2 [Mode: Riskli] => Riskli
        - aile_iliskikalitesi > 2 [Mode: Risksiz] => Risksiz
    - ark_disaridavakit > 3 [Mode: Riskli] => Riskli

- cinsiyet > 1 [Mode: Risksiz]
    - ark_disaridavakit <= 3 [Mode: Risksiz] => Risksiz
    - ark_disaridavakit > 3 [Mode: Risksiz]
        - aile_iliskikalitesi <= 3 [Mode: Riskli]
            - aile_uyesayisi <= 1 [Mode: Riskli]
                - A_egitim <= 3 [Mode: Risksiz]
                    - A_egitim <= 1 [Mode: Riskli] => Riskli
                    - A_egitim > 1 [Mode: Risksiz] => Risksiz
                - A_egitim > 3 [Mode: Riskli] => Riskli
            - aile_uyesayisi > 1 [Mode: Risksiz] => Risksiz
        - aile_iliskikalitesi > 3 [Mode: Risksiz] => Risksiz



Şekil 4.4. C5.0 Karar ağacı yapısı

BÖLÜM 5. TARTIŞMA VE SONUÇ

Ergenlik dönemini de kapsayan gençlik dönemi, gencin kişilik özelliklerini farkına varmadığı bir süreçtir. Bu kişilik arayışı esnasında risk almaya yönelik eylemlerin sıklıkla görüldüğü, inişli çıkışlı duygularında etkisi ile ilgili olarak sigara, alkol gibi bağımlılık yaratan zararlı tüketimlerin kolaylıkla ortaya çıkabildiği bir dönemdir. Bu dönemi başarı ile atlatan gencin özgüveni ve kendinde gördüğü değer artar. Aynı zamanda huzurlu ve sıkıntısız bir aile ortamı gencin ruhsal gelişimini sağlıklı yönde etkilerken, sorunlu bir aile ortamında bulunan gençleri olumsuz etkileyerek alkollü içecekler ve sigara gibi bağımlılık yaratan zararlı tüketimlerini arttırıcı bir risk oluşturacaktır.

Gençlerin sigara, alkol ve uyuşturucu madde kullanımına zemin hazırlayan ailesel risk etkenlerine bakıldığında genetik ve aile ilişkileri olarak ikiye ayrıldığı görülmektedir. Özellikle ikizlerle yapılan çalışmalar sonucunda alkol tüketiminin genetik geçiş ile sürdürüldüğü uzun süredir bilinmektedir. Bununla birlikte ebeveynleri alkollü içecek bağımlısı olan erkek çocukların bağımlı olma riskinin, ebeveynleri alkollü içeceklere bağımlı olmayan çocuklara kıyasla daha yüksek olduğu bulunmuştur [24,25].

Lise öğrenimi sonrasında üniversiteye başlayan öğrencilerin aile ve çevre kontrolünden uzaklaşmalarının neticesinde daha özgür bir ortamda bulunuyor olmanın etkisiyle bağımlılık yapıcı maddelerden sigara içme alışkanlığınında arttığı gözlemlenmiştir. Elde edilen bulgular ise bize sigara ve alkollü içecek kullanımının önlenmesi çalışmalarının sadece ortaokul ve lise seviyesinde değil, üniversitelere yönelik olarakta sürdürülmesi gerektiğini göstermiştir [26]. Gençlerin zararlı alışkanlık ve bağımlılıklardan uzak durmasını sağlayacak koruyucu faktörler arasında olumlu özelliklere sahip yaşıtlarının olduğu bir çevrede yer alıyor olmaları önemli bir yere sahiptir [27].

Yapılan bu çalışmada, Portekiz'de 2005-2006 yılında eğitim gören devlet okulu öğrencilerine ait veri seti üzerinde veri madenciliği teknikleri uygulanmıştır. Günümüz teknolojisi sayesinde gençleri alkollü içecek tüketimine iten sebepler ve bu durumun yarattığı sonuçlar günün sonunda birer veridir. Bu veriler işlenip bilgi haline getirildiğinde ise sebeplerin ortadan nasıl kaldırılacağı, sonuçların nasıl hafifletileceği öngörülebilir. Öğrencilerin alkol tüketimlerinin faktörler analizi ve veri madenciliği tekniklerinden karar ağacı yapısı C 5.0 algoritmasnın kullanılması ve lojistik regresyon yöntemleri ile elde edilen sonuçlar ışığında öğrencilerin demografik özelliklerinin alkollü içecek tüketim eğilimlerinin üzerinde etkisi olduğunu söyleyebiliriz. Öğrencileri bu eğilimlere iten faktörler ortadan kaldırıldığında ya da azaltıldığında başta kişinin kendisi olmak üzere toplumada olumlu yönde etki edeceği düşünülebilir.

Çalışma sonucunda aile arasındaki iletişimin kaliteli olduğunu belirten öğrencilerin alkollü içecek tüketimine yönelecek riskli sınıfta bulunmamaları daha önce bahsi geçen [16,17] çalışmalarda elde edilen sonuçla örtüşmektedir. Ayrıca aile üye sayısı az olan veya tek çocuk olduğunu belirten öğrencilerin yalnızlık hissiyle bu maddelere yöneldiğini bildiren daha önce yapılmış araştırmaların [17,19] sonuçlarını desteklemiştir. Bu çalışmanın sonuçları ve literatüre konu olmuş diğer çalışmaların sonuçları birleştirildiğinde; günümüzde zararlı maddelerin tüketimin azaltılmasını sağlayacak önlemler ile genetik risk etkenlerinin oranlarınında düşeceği ve bu sayede gelecek nesillere aktarılan gen faktörlerinden doğacak bağımlılık ihtimallerinin daralacağı düşünülmektedir. Gençlerin bireysel, ailesel ve okul özellikleri dikkate alınarak sosyal yaşam içerisinde uygulanabilecek önlemlerin planlanması etkili bir yol olacaktır. Öğrencilerin zararlı alışkanlıklara sebep olacak madde kullanmalarının nedenlerinin daha kapsamlı inceleneceği çalışmalar yapılması gerektiğinin, risk faktörlerinin belirlemesi ile muhtemel kullanımı engelleyici önlemlerin düşünülmesi ve ilgili kurumların katkılarıyla öğrencilerin aile, okul ve çevre özelliklerinin de dikkate alınacağı önleyici ve koruyucu projeler yürütülmesi sağlıklı olacaktır.

KAYNAKLAR

[1] Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., ... & Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 140.

[2] Clifton, C. (2010). Encyclopædia britannica: definition of data mining. Retrieved on, 9(12), 2010.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics.

[4] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[5] Holland, J. H. (1992). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press.

[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[7] Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM.

[8] Herken, H., ÖZKAN, A. S. Ç., & BODUR, S. (2000). Ögrencilerde Alkol Kullanım Sıklığı ve Sosyal Öğrenme ile ilişkisi.

[9] Mercer GW, Kohn PM(1980): Child-rearing factors, outhoritarianizm drug use attitudes andadolescent drug use a model. J Genet Psychol 163:159-71.

[10] Sieber, M. F., & Angst, J. (1990). Alcohol, tobacco and cannabis: 12-year longitudinal associations with antecedent social context and personality. Drug and alcohol dependence, 25(3), 281-292.

[11] Currie, C., Roberts, C., Settertobulte, W., Morgan, A., Smith, R., Samdal, O., ... & World Health Organization. (2004). Young people's health in context: Health Behaviour in School-aged Children (HBSC) study: international report from the 2001/2002 survey (No. EUR/04/5048327). Copenhagen: WHO Regional Office for Europe.

[12] World Health Organization. Management of Substance Abuse Unit. (2014). Global status report on alcohol and health, 2014. World Health Organization.

[13] World Health Organization. Management of Substance Abuse Unit. (2018). Global status report on alcohol and health, 2018. World Health Organization.

[14] http://www.tuik.gov.tr/PreHaberBultenleri.do?id=24680, Erişim Tarihi: 10 Haziran 2018

[15] Akkuş, D., Karaca, A., Şener, D. K., & Ankaralı, H. (2017). Lise Öğrencileri Arasında Sigara, Alkol Kullanım Sıklığı Ve Etkileyen Faktörler. Anatolian Clinic the Journal of Medical Sciences, 22(1), 36-45.

[16] Cohen, D. A., Richardson, J., & LaBree, L. (1994). Parenting behaviors and the onset of smoking and alcohol use: a longitudinal study. Pediatrics, 94(3), 368-375.

[17] Herken, H., ÖZKAN, A. S. Ç., & BODUR, S. (2000). Ögrencilerde Alkol Kullanım Sıklığı ve Sosyal Öğrenme ile ilişkisi.

[18] Karatay, G., & Kubilay, G. (2004). Sosyoekonomik düzeyi farklı iki lisede madde kullanma durumu ve etkileyen faktörlerin belirlenmesi. Hemşirelikte Araştırma Geliştirme Dergisi, 1(2), 57-70.]

[19] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance

[20] Kurupınar, A., & Erdamar, G. (2014). Ortaöğretim öğrencilerinde görülen madde bağımlılığı alışkanlığı ve yaygınlığı: Bartın ili örneği. Sosyal Bilimler Dergisi, 16(1), 65-84.

[21] UCI Machine Learning Repository. "Student alcohol consumption data set"http://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION , (04.04.2016).

[22] Draper, N. R., & Smith, H. (2014). Applied regression analysis (Vol. 326). John Wiley & Sons.

[23] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

[24] Ünal, M. (1991). Madde bağımlılığı ve alkolizmde aile. Aile ve Toplum Dergisi, 1(2), 80-85.

[25] Alikasifoglu, M. (2002). Ercan O. Ergenlerde madde kullanimi. Turk Pediatri Arsivi, 37, 66-73.

[26] Selma, Ç. İ. V. İ., & ŞAHİN, T. K. Selçuk Üniversitesi Tıp Fakültesi ve Sağlık Hizmetleri Meslek Yüksekokulu Öğrencilerinin Sigara Konusundaki Bilgi-Tutum ve Davranışları. Sosyal Politika Çalışmaları Dergisi, 1(1).

[27] Alikaşifoğlu, M., & Bilimdalı, A. (2008). ERGENLERDE DAVRANIŞSAL SORUNLAR

ÖZGEÇMİŞ

Nihal Zuhal Kayalı, 15.02.1993'te İstanbul'da doğdu. İlk ve ortaokul eğitimini İstanbul'da tamamladı. 2006 yılında Sakarya'da Şehit Üsteğmen Selçuk Esedoğlu Anadolu Lisesi'ni kazanıp 2010 yılında mezun oldu. 2010 yılında başladığı Sakarya Üniversitesi Bilgisayar Mühendisliği Bölümü'nü 2014 yılında bitirdi. 2014 yılında Sakarya Üniversitesi Bilgisayar ve Bilişim Mühendisliği Bölümü'nde yüksek lisans eğitimine başladı. Lisans ve Yüksek Lisans eğitimi sürecinde Kısmi Zamanlı Asistan Öğrenci olarak çalıştı. 2015 yılında Sakarya Teknokent bünyesinde faaliyet gösteren Erkay Teknoloji Geliştirme Hizmetleri şirketinde Bilgisayar Mühendisi olarak çalıştı. 2017 yılının Mart ayında Türk Alman Üniversitesi'nde araştırma görevlisi olarak çalışmaya başladı ve halen Türk Alman Üniversitesi Bilgisayar Mühendisliği Bölümü Yazılım Anabilim Dalı'nda Araştırma Görevlisi olarak görev yapmaktadır