T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANA BİLİM DALI

DİYABET HASTALARINA AİT VERİLERİN İSTATİSTİKSEL YÖNTEMLER VE VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK İNCELENMESİ

Yüksek Lisans Tezi

Merve DÜNDER

Danışman

Doç. Dr. Erol TERZİ

SAMSUN
2021

TEZ KABUL VE ONAYI

Merve DÜNDER tarafından, **Doç. Dr. Erol TERZİ** danışmanlığında hazırlanan **"Diyabet Hastalarına Ait Verilerin İstatistiksel Yöntemler ve Veri Madenciliği Teknikleri Kullanılarak İncelenmesi"** başlıklı bu çalışma, jürimiz tarafından 16.2.2021 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Yüksek Lisans Tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı Üniversitesi Ana Bilim/Ana Sanat Dalı	**İmza**	**Sonuç**
Başkan	Doç. Dr. Tolga ZAMAN Çankırı Karatekin Üniversitesi İstatistik Anabilim Dalı		☒ Kabul ☐ Ret
Üye **(Danışman)**	Doç. Dr. Erol TERZİ Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı		☒ Kabul ☐ Ret
Üye	Doç. Dr. Hasan BULUT Ondokuz Mayıs Üniversitesi İstatistik Anabilim Dalı		☒ Kabul ☐ Ret

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

ONAY

… / … / …

Prof. Dr. Ali BOLAT

Enstitü Müdürü

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmamdaki mevcut tüm bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

İmza
22/03/2021
Öğrenci Adı SOYADI
Merve DÜNDER

TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı : Diyabet hastalarına ait verilerin istatistiksel yöntemler ve veri madenciliği teknikleri kullanılarak incelenmesi

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 14/01/2021 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 21
Tek kaynak oranı : % 3 çıkmıştır.

İmza
14 /01/ 2021
Danışman Adı SOYADI
Doç. Dr. Erol TERZİ

ÖZET

DİYABET HASTALRINA AİT VERİLERİN İSTATİSTİKSEL YÖNTEMLER VE VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK İNCELENMESİ

Merve DÜNDER
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
İSTATİSTİK ANA BİLİM DALI
Yüksek Lisans, Mart/2021
Danışman: Doç. Dr. Erol TERZİ

Günümüz teknolojisi hızla ilerlemekte ve her geçen gün kullanım alanlarına paralel olarak gücü de artmaktadır. Bundan dolayı bilgisayarların veri tabanlarında tutulan verilerin analiz edilmesi sonucu, bu verilerden elde edilen bilgiler, karar vericiler için önem kazanmaktadır. Bilgisayar sistemleri ile elde edilen veriler tek başına yetersizdir. Bu veriler belli amaçlar doğrultusunda işlendiği zaman anlamlı olmaya başlarlar. Bu yüzden büyük miktardaki verileri işleyebilen teknikleri kullanabilmek veri madenciliği için çok önemlidir. Bu çalışmada veri madenciliğinin günümüzde geldiği noktaya değinildi ve veri madenciliği üzerine yapılan çalışmalar incelenerek günümüz sorunlarından biri olan diyabet hastalığına ait veri seti üzerinde bir uygulaması yapıldı. Araştırmada kullanılan diyabet verisine veri madenciliği yöntemlerinden kümeleme analizi, karar ağaçları, lojistik regresyon ve birliktelik kuralları uygulandı. Sonuç olarak diyabet hastaları üzerinde kullanılan insülin, glyburide ve glipzide ilaçlarının daha etkili olduğu saptandı.

Anahtar Sözcükler: Veri Madenciliği, Karar Ağacı Algoritmaları, Kümeleme Analizi

ABSTRACT

ANALYSIS of DATA FOR DIABETS PATIENTS USING STATISTICAL METHODS AND DATA MINING TECNIQUES

Merve DÜNDER
Ondokuz Mayıs University
Institute of Graduate Studies
Department of statistics
Master, March / 2021
Supervisor: Doç. Dr. Erol TERZİ

Today's technology is advancing rapidly, and its power is increasing day by day. As a result of the analysis of the data held in the databases of the computers, the information obtained from these data is important for the decision makers. Data obtained by computer systems are insufficient. When these data are processed for certain purposes, they begin to be meaningful. Therefore, it is important for data mining to use techniques that can process large amounts of data. In this study, the point that data mining has reached today was touched upon and studies on data mining were examined and an application was made on diabetes data, which is one of today's problems. Data mining methods such as cluster analysis, decision trees, logistic regression and association rules were applied to the diabetes data used in the study. As a result, it was determined that insulin, glyburide and glipzide drugs used on diabetic patients were effective.

Keywords: Data Mining, Decision Tree Algorithms, Clustering Analysis.

ÖNSÖZ VE TEŞEKKÜR

Akademik eğitim sürecimin bir üst noktası olan yüksek lisans çalışmamın her aşamasında göstermiş olduğu anlayışı, yapmış olduğu maddi ve manevi tüm destekleri için danışman hocam sayın Doç. Dr. Erol Terzi 'ye saygılarımı ve sonsuz teşekkürlerimi sunarım. Lisans ve yüksek lisans eğitimim boyunca yardım, bilgi ve tecrübeleri ile bana sürekli destek olan bölüm hocalarıma teşekkür ederim. Tez çalışmamın araştırma ve kaynak edinme aşamalarında bana yardımcı olan ve uygulama kısmında yol gösterici olan eşim Dr. Öğr. Üyesi Emre DÜNDER'e ve çalışmalarım boyunca maddi ve manevi destekleriyle beni hiçbir zaman yalnız bırakmayan sevgili anneme sonsuz teşekkür ederim.

Adı SOYADI
Merve DÜNDER

İÇİNDEKİLER

SİMGE ve KISALTMALAR

TTB: Türk Tabipleri Birliği

DM: Diyabetes Mellitus

IDF: Uluslararası Diyabet Federasyonu

VM: Veri Madenciliği

ADA: American Diabetes Association

OECD: Ekonomik İşbirliği ve Kalkınma Teşkilatı

GSYİH: Gayri Safi Yurt İçi Hasıla

DSÖ: Dünya Sağlık Örgütü

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

İçinde yaşadığımız bilişim çağında elektronik ortamda mevcut verinin hızla artması ve birçok verinin birikmesi ile veri tabanlarında bilgi keşfi olarak adlandırılan yeni bir paradigma ortaya çıkmıştır. Daha yaygın bir kullanımla bu alana Veri Madenciliği denilmektedir.

Veri madenciliği analizleri, elde bulunan veri kümelerinin durumuna ve verilerin kullanım amacına göre veri madenciliği yöntemleri ile yapılmaktadır. Veri madenciliği yöntemlerini uygulamadan önce analiz için önemli olacak değişkenlerin belirlenmesi gerekmektedir. Değişkenler belirlendikten sonra veri madenciliği yöntemleri ile analizler gerçekleştirilmektedir. Böylece, veri madenciliği ile doğru bilgilere hızlı ve güvenilir bir şekilde ulaşılmaktadır.

Veri madenciliğinin kullanımı belirli uygulama alanlarıyla sınırlandırılamaz. Veri madenciliğini verinin üretilip kayıt altına alındığı her alanda kullanmak mümkündür. Sağlık, endüstri, mühendislik, pazarlama, bankacılık ve eğitim alanları veri madenciliğinin yoğun olarak kullanıldığı başlıca uygulama alanlarıdır (Han ve Kamber, 2006).

Tıp, veri madenciliğinin yoğun olarak kullanıldığı bir bilimdir. Sağlık sistemi politikalarının ve yönetsel kararlarının temeli veri ve veriden elde edilmiş bilgidir. Bu bilgiler hasta düzeyinde daha iyi sağlık hizmeti sunumu, sağlık kurumlarının daha iyi yönetilmesi, kaynakların etkin kullanımı ve sağlık politikalarının oluşturulması amaçları ile kullanılmaktadır. Sağlık verileri hastaneler, diğer sağlık kurumları, sigorta şirketleri ve ilgili kamu kurumları başta olmak üzere birçok kuruluş tarafından toplanmaktadır. Büyük miktarda verinin ilk çağrıştırdığı kavram veri madenciliğidir. Veri Madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir. (Koyuncugil ve Özgülbaş 2009).

Günümüzde en çok karşılaştığımız ve tam olarak kesin bir tedavisi olmayan hastalıklardan biri diyabet hastalığıdır. Bu çalışmada 10485 diyabetli hastaya ait veriler kullanılarak veri madenciliği tekniklerinden karar ağaçları, lojistik regresyon, birliktelik kuralları ve kümeleme analizi uygulamaları gerçekleştirilmiştir.

2. LİTERATÜR ÇALIŞMALARI

Günümüzün gelişen bilişim sistemleri teknolojileri, sağlık kuruluşlarında hasta takip sistemlerinin gelişmesi ile sağlık sektöründe veri madenciliği uygulamalarının yaygınlaşmasını sağlamıştır. Literatürde birçok hastalık ile ilgili veri madenciliği çalışmaları yapılmakta, özellikle hastalarla ilgili detaylı verilerinin oluşu, hastalığın çeşitli tiplerde bulunması (tip 1, tip 2, hamilelik diyabeti) ve komplikasyonlarının çok önemli ve etkilerinin mortalite ve imputasyonlara neden oluşu sebebiyle diyabet hastalığı sağlıkta veri madenciliği uygulamalarının odak konusu olmaktadır.

Özyazar (2019), sürdürülebilir tıbbi sistemler için diyabet verilerini kullanarak veri madenciliği yöntemlerini uyguladığı bir çalışmada bir hastanenin, tip 2 diyabet hastalığı üzerine çalışmıştır. Araştırmada yeni bir değişken oluşturup kullanılması hedeflenmiştir. Toplamda 13 farklı özellik (macrovascular complications, microvascular complications, co-disease, GFR degree, Hba1cdegree, Albumin degree, BMI degree, lipid profile, glucose level risk degree, creatinine degree, HDL/LDL, TCOLL/LDL and ALB/Cr) listelenmiştir.

Ergin ve diğerleri (2020), diyabet ve sosyo-ekonomik durum arasındaki ilişkiyi ve sosyo-ekonomik durumun diyabete etkisini belirlemek amacıyla uygulanan çalışmada kümeleme analizi ve iz regresyon analizi sonuçlarına göre GSYİH (Gayri Safi Yurt İçi Hasıla)'nın yüksek olduğu OECD(Organisation for Economic Co-operation and Development) ülkelerinde diyabet prevalansının daha düşük olduğu sonucuna ulaşılmıştır. Çalışmada kullanılan veri setinin analizi R project programlama dili ile uygulanmıştır (Ergin ve diğerleri, 2020).

Çiçek (2014), tip 2 diyabetli hastaların klinik kılavuzların gereğini yerine getirmek için veri madenciliği yönetimlerini kullanmıştır. Çalışmada doktorlara ilaç dozu ayarlarken yardımcı olabilecek bir araç oluşturmak hedeflenmiştir. Girdi parametreleri olarak cinsiyet, yaş, alanin aminotransferaz, HDL kolesterol, kreatinin, LDL kolesterol, mikroalbuminüri, trigliserid ve üre olarak belirlemiştir. Sonuç olarak metformin isimli ilaç için doz ayarlamayı planlayabilmiştir.

Uçar (2014), diyabetik hastalarda kronik böbrek yetmezliğinin veri madenciliği tekniklerini kullanarak teşhis edilmesini araştırmıştır. Bu çalışmanın amacı risk grubu içerisinde yer alan hipertansiyon ve diyabet hastalarının böbrek

yetmezliği hastalığı riski taşıyıp taşımadığına karar veren sağlık uzmanının karar verme sürecini destekleyecek ve bunu yaparken de veri madenciliği tekniklerini kullanacak web tabanlı bir uygulama geliştirmektir. Uygulama, Ege Üniversitesi Tıp Fakültesi Hastanesi Biyokimya Laboratuvarında yer alan diyabet ve hipertansiyon hastalarının verileri kullanılarak geliştirilmiştir. Veri madenciliği işlemleri için 74322 adet kayıt üzerinde Naive Bayes algoritması kullanılmıştır ve 7296 adet kayıt üzerinde %78 başarı ile test edilmiştir.

Adalı (2019), diyabet hastalarının verilerini kullanarak veri madenciliği tekniklerini karşılaştırmıştır. Bu çalışmada hem potansiyel hastaların risk oranını belirlemek, hem de diyabetlilerin tedavileri boyunca yol gösterici bir uzman sistem geliştirmek için veri madenciliği tekniklerinin karşılaştırılmasını amaçlamıştır. İstanbul diyabet hastanesi verilerini kullanılmıştır. Diyabet hastalarının sahip olduğu veriler kullanılarak ANFIS, multinominal lojistik regresyon, bayes ağı yardımı ve rough ile kestirimler yapılmıştır. Sonuç olarak, ANFIS' in daha etkili bir öğrenme ve kestirim aracı olduğu görülmüştür.

Şaşar (2017), diyabet hastalarındaki Hba1c parametresine etki eden faktörlerin veri madenciliği yöntemleri ile tahminini araştırmıştır. Bu çalışmadaHba1cparametresini etkileyen faktörler incelenmiştir. Hekimlere teşhis ve tedavi süresinde yardımcı olabilecek sonuçların saptanması hedeflenmiştir. Çalışmada Köyceğiz ve Dalaman devlet hastanelerinden elde edilen veriler kullanılmıştır. Verilerdeki ilgisiz ve artık verilerin çıkartılması, girdi verilerinin azaltılması ve daha kaliteli verilerle çalışmak için öznitelik seçme algoritmaları kullanılmıştır. Başarı oranı yüksek olan algoritmalarla öznitelikler seçilmiştir. Elde edilen veriler veri madenciliği yöntemleri kullanılarak sınıflandırılmış ve başarıları karşılaştırılmıştır. Çalışmada Yapay Sinir Ağları modelinin diğer algoritmalardan daha başarılı sonuçlar ürettiği ve %94'lere varan tahmin başarısı olduğu görülmüştür.

Çerkezi (2013), veri madenciliği yöntemlerini kullanarak diyabetik retinopati hastalığının teşhis edilme sürecinin araştırıldığı bir çalışmada gerçek veriler kullanılarak K-en yakın komşu, ağırlıklı oylama KNN (Weighted K-nearest neighbor) ve Bayes algoritmaları ile sınıflandırma yapılmış ve elde edilen sonuçlar tartışılmıştır. Çalışmada kullanılan veriler, Sakarya Üniversitesi Eğitim

ve Araştırma Hastanesi Göz Polikliniği bölümünden alınmıştır. Sonuç olarak veri madenciliği yöntemlerini kullanarak parametreler üzerinden diyabetik retinopati hastalıkların teşhisi yapılmıştır.

Şahin (2016), veri madenciliği yöntemleri ile diyabet hastalığına sebep olan faktörlerin araştırıldığı bir çalışmada kişilere ait fiziksel özellikler ve kan testi verilerinin veri madenciliği yöntemleri ile diyabet hastası olup olmadıklarını araştırmıştır. Yapılan çalışmada 768 kadına ait veriler Dünya Sağlık Örgütü (WHO) kriterlerine göre düzenlenmiş, bu veriler veri seti haline getirilerek farklı sınıflandırma algoritmaları uygulanmış ve sonuçlar kıyaslanmıştır. Kıyaslama sonucunda%86 başarı elde edilmiştir.

Rahman ve Afroz (2012), diyabet teşhisi için farklı sınıflandırma tekniklerinin karşılaştırılması amacıyla literatürde yer edinmiş bir çalışmada WEKA, TANEGRA ve MATLAB adlı üç veri madenciliği aracını kullanarak bireylere ait parametreleri analiz etmiştir. Toplamda 768 deneğin kullanıldığı bu araştırmada sınıflandırma yöntemleri arasındaki farklar test edilmiştir. Weka'da en güvenilir sonuç J48 algoritması %81.33 ile belirlenmiştir. TANAGRA'da Naive Bayes %100 doğruluk sağlamıştır ve son olarak MATLAB'da ANFIS %78,79 doğruluk sağlamıştır. Sonuç olarak en güvenilir yöntemin TANAGRA olduğu bilgisine ulaşılmıştır.

Aljumah ve diğerleri (2013), genç ve yaşlı diyabet gruplarını veri madenciliği yöntemlerini kullanarak incelendiği bir çalışmada regresyon tabanlı veri madenciliği tekniklerini kullanarak diyabet tedavisini araştırmıştır. Çalışmada kullanılan veri seti farklı yaş grupları ve bunlara uygulanan tedavi yöntemlerinin etkinliği analiz edilmiştir. Sonuç olarak yan etkilerden kaçınmak için genç yaş grubundaki hastalar için ilaç tedavisinin ertelenebileceği sonucuna varılmıştır. Buna karşılık yaşlılık grubundaki hastalara derhal ilaç tedavisi uygulanması gerektiği raporlanmıştır.

Breault ve diğerleri (2002), Amerika Birleşik Devletleri'nde elde edilen bir veri seti ile araştırmada iki değişken ayrıntılı olarak tartışmış; komorbidite indeksi ve sonuçlarla ilgili glisemik kontrolün bir ölçüsü olan Hba1c öngörücülerinin ikili hedef değişkeni olan sınıflandırma ve regresyon ağaçlarında (CART) sınıflandırma ağacı yaklaşımını kullanmıştır. Sonuç olarak Hba1c hastaları için

komorbidite indeksi bulma olasılığı (hastaların ilişkili hastalıklara yakalanma riski) 6,5 yaşın altındaki hastalarda yaşlı olan hastalara göre 3,2 kat daha yüksektir.

Iyer ve diğerleri (2015), diyabet tanısının güvenilir bir şekilde teşhis edilmesi için veri madenciliğinin sınıflandırma metodunu incelemek amacıyla diyabetin ABD (Amerika Birleşik Devletleri) için beşinci sırada ölümcül olduğunu ve bu sebeple teşhis ve tedavide yöntem belirlemek için veri madenciliği yöntemlerini kullanmanın önemini belirtmişlerdir. Araştırmada karar ağaçları ve Bayes algoritmaları kullanılmıştır. Elde edilen sonuçlara göre, her iki yöntemin de hata oranında küçük farklılıklar olduğu gözlenmiştir.

Devi ve Shyla (2016), diyabet mellitusu tahmin etmek için kullanılan çeşitli veri madenciliği teknikleri isimli yayınladıkları makale çalışmasında Naive Bayes, J48, PLS-LDA, SVM, BLR, MLP, K-NN, bayes ağı algoritmalarını incelemişlerdir. Çalışmayı gerçekleştirirken Tanagra, WEKA ve MATLAB kullanılmıştır. Sonuçlara göre J48 algoritması %99.87 oranında, C4.5 algoritması %86 oranında doğrulukla tahmin ettiği saptanmıştır.

3. VERİ MADENCİLİĞİ

3.1. Veri Madenciliğinin Kısa Tarihi

Veri madenciliği kavramı ilk olarak 1960'lı yıllarda teknolojinin ilerlemesi ile ortaya çıkmıştır. Bilim insanları o dönemlerde, bilgisayar yardımıyla, yeterince uzun bir tarama yapıldığında, istenilen verilere ulaşmanın mümkün olacağı gerçeği kabullenildikten sonra çalışmalarda veri taraması adı ile araştırmalarını gerçekleştirdi (Öğüt, 2009).

1970'li yıllarda veri tabanı işletim sistemi uygulamaları kullanılmaya başlanmıştır. Bu sayede bilim insanları basit kurallara dayanan sistemler geliştirerek makine öğrenimini sağlamışlardır. 1980'lerde veri tabanı yönetim sistemleri yaygınlaşmış ve bilimsel alanlarda uygulanmaya başlanmıştır. Bu yıllarda şirketler; müşteriler, rakipler ve ürünler ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Bu veri tabanlarının içerisinde çok büyük miktarlarda veri bulunmaktadır ve bunlara SQL veri tabanı sorgulama dili ya da benzeri diller kullanarak ulaşılabilmektedir (Savaş, 2012).

1990'larda bilgisayar mühendisleri, geleneksel istatistiksel yöntemlerinin yerine algoritmik bilgisayar modülleri kullanmışladır. Bu yıllarda veri tabanlarındaki bilgiler hızla arttığı için büyük miktardaki verilerden faydalı işe yarar bilgilerin nasıl elde edilebileceği üzerine çalışmalar yapılmıştır. 2000'li yıllarda artık veri madenciliği kavramı bilim dünyasında popülerlik kazanmış, geliştirilmiş ve gün geçtikçe önemi artmaktadır. (Erdoğan, 2019)

3.2. Veri Madenciliği Tanımları

- Veri Madenciliği verideki gizli, önceden bilinmeyen ve potansiyel olarak faydalı enformasyonun önemsiz olmayanlarının açığa çıkarılmasıdır (Piatetsky-Shapiro ve Fawley 1991).

-Veri madenciliği büyük verilerde ilginç beklenmedik veya değerli yapıların keşfidir (Hand, 2007).

- Veri madenciliği veri tabanlarında önceden bilinmeyen örüntüleri ve bu bilgileri tahmin modelleri oluşturmak için kullanma süreci olarak tanımlanır (Hearst, 1999).

- Veri madenciliği, istatistik tanıma ve matematik teknikleri gibi desen tanıma teknolojilerini kullanarak depolarda depolanan büyük miktarda veriyi eleyerek anlamlı yeni korelasyon kalıpları ve eğilimleri keşfetme sürecidir (Wu ve arkadaşları, 2013).

-Veri madenciliği, toplanan verilerden anlamlı bilgiler çıkarmak, veri içerisinde gizli olan birtakım örüntüleri ve eğilimleri tespit etmek ve çeşitli değişkenler arasında ilişkiler bulmak ve böylece karar vermeye yardımcı olmak amacıyla uygulanan bir yaklaşımdır. Veri madenciliğinde istatistik, yapay zekâ, makine öğrenmesi gibi farklı alanlarda geliştirilmiş birçok teknik ve yöntem kullanılmaktadır (Seyrek ve Ata, 2010).

-Veri madenciliği; büyük miktardaki verileri, istatistiksel ve matematiksel tekniklerin yanı sıra patern tanıma teknolojileri de kullanılarak, faydalı yeni korelasyonlar, paternler ve trendler keşfetme sürecidir (Larose, 2005).

3.3. Veri Madenciliği Uygulama Alanları

Veri madenciliği, teknolojinin gelişmesi ile artan verileri kolay bir şekilde kullanılabilir duruma getirerek veri madenciliği yöntemleriyle anlamlı bilgilerin ortaya çıkarılmasını sağlamaktadır. Analizlerde sağladığı kolaylıklar nedeniyle çeşitli departmanlarda uygulanabilmektedir (Silahtaroğlu,2018). Veri madenciliğinin kullanıldığı başlıca alanlar şunlardır:

-Tıp

-Pazarlama yönetimi

-Müşterilerin alışveriş alışkanlıklarının tespiti

-Kredi skorlama

-Reklam kampanyalarının etkinliğini artırma

-Müşteri sadakati ve yeni müşteri kazanımı

-Pazar sepeti analizi

-Dolandırıcılık tespiti ve risk yönetimi

-Sinyal işleme (telekomünikasyon)

-Biyoloji (DNA sıra işleme)

-Fiyatlandırma

-Ulaştırma

-Tedarik zinciri analizi

-Sosyal ağ analizi

-Coğrafi bilgi sistemleri

-Silahlı kuvvetler güvenlik uyarıları

-Bilgi güvenliği

-Eğitim - öğretim

3.4. Veri Madenciliğinin Diyabet Tedavisinde Önemi

Diyabet Mellitus (DM) pankreas adlı salgı bezinin yeterli miktarda insülin hormonu üretememesi ya da ürettiği insülin hormonunun etkili bir şekilde kullanılamaması durumunda gerçekleşen komplikasyonlara neden olan bir hastalıktır. Kronik DM kontrol altına alınmadığı zaman özellikle gözlerin, böbreklerin, sinirlerin, kalp ve damarların ciddi anlamda hasar görmesine neden olur. Belirgin hiperglisemi belirtileri arasında poliüri, polidipsi, kilo kaybı, bazen polifaji ve bulanık görme vardır. Büyüme ve bazı enfeksiyonlara duyarlılık da kronik hiperglisemiye eşlik edebilir. Diyabetin uzun vadeli komplikasyonaları arasında retinopati ve nefropati ve periferik nöropati bulunur. DM hastaları iki farklı kategoriye ayrılırlar: Tip1 ve Tip2 diyabet. Tip1 Diyabet insülin sekresyonunun eksikliğinde ortaya çıkar. Tip2 diyabet ise insüline direnç ve yetersiz insülin salınımı ile oluşan DM tipidir.

Günümüzün bilişim sistemleri teknolojilerinin ve sağlık kuruluşlarında hasta takip sistemlerinin gelişime ihtiyaç duyulması veri madenciliği uygulamalarının yaygınlaşmasını sağlamıştır. Literatürde diyabet hastalığı ile ilgili çeşitli veri madenciliği çalışmaları yapılmakta, özellikle hastalarla ilgili detaylı verilerin oluşu, hastalığın çeşitli tiplerde bulunması (tip 1, tip 2, hamilelik diyabeti) ve komplikasyonlarının çok önemli ve etkilerinin mortalite ve imputasyonlara neden oluşu sebebiyle diyabet hastalığı sağlıkta veri madenciliği uygulamalarının odak konusu olmaktadır. Bu aşamada veri madenciliği diyabet

tedavisi sırasında hastaların demografik ve fizyolojik özelliklerine göre kontrol altına alınma sürecinin düzenlenmesini sağlar.

3.5. Veri Madenciliği Süreci

3.5.1. Veri Toplama, Temizleme ve Dönüştürme

Veri madenciliğinin ilk aşaması verilerin toplanmasıdır. Toplanılan veriler birçok farklı ortamda depolanabilmektedir. Yapılması gereken ilk şey veri tabanlarından veya veri ambarlarından uygun verileri çekmektir. Daha sonra veriler iki ayrı grupta toplanır (Test ve analiz). Elimizdeki bu verileri bazen dönüştürmemiz gerekebilir. Bunun sebebi ise kullanılan bazı veri madenciliği algoritmalarının elimizdeki veri tipine göre daha olumlu sonuç vermesidir. Ama daha öncesinde veri temizleme işlemi uygulanmalıdır. Amaç uygun olmayan veya hatalı verileri ayıklamaktır. Bu işlemde eksik veriler bazı tekniklerle doldurulur.

3.5.2. Model kurma ve Değerlendirme

Model kurma veri madenciliğinin en temel unsurudur. Öncelikle modelin doğru bir şekilde kurulabilmesi için yapılacak araştırmanın amacı çok iyi bir şekilde belirlenmelidir. Daha sonra eldeki veriler için veri madenciliği algoritmaları çalıştırılır ve en doğru sonucu veren algoritma uygulanır. En doğru sonucu bulmak için ise çeşitli yöntemler mevcuttur.

3.5.3. Raporlama

Bu aşamada elde ettiğimiz veri madenciliği bulguları gösterilir. Son olarak elde edilen bu bulgular değerlendirilir.

Her bir veri madenciliği modelinin bir yaşam döngüsü vardır. Bazı araştırmalarda bu yaşam döngüsü uygulanır ve bu modelin eğitilmesine gerek yoktur. Fakat çoğu kez yeni veriler geldikçe modelin yeniden eğitilmesi gerekir. Sonuç olarak bir model kurulduktan sonra veri setinde değişiklik yapıldıysa model güncellenmelidir (Tekerek, 2011).

4. VERİ MADENCİLİĞİ MODELLERİ

Veri madenciliğinin uygulama aşaması için çok sayıda algoritma geliştirilmiştir. Uygulanacak algoritmanın belirlenmesi verilerin niteliğine göre değişebilmektedir. Veri madenciliği modelleri kullanıcıların model arayışlarına yardımcı olmak ve istenilen sonuca ulaşmak için ipuçları belirlemelerine de katkıda bulunmaktadır.

Veri madenciliği yöntemleri deneysel çalışmalar olduğu için çok sayıda algoritmanın ayrı ayrı sınanıp, kullanılacak olan algoritmanın seçimine denemeler yapılarak ulaşılması daha güvenilir olacaktır. Birçok algoritmanın denendiği verilerde en başarılı sonuçları veren algoritmalar seçilebilir. Veriler zaman içinde farklılaşma gösterdiği için kullanılan modellerin de zaman içinde yenilenmesi gerekebilir.

Veri madenciliğinde kullanılan modeller, tahmin edici ve tanımlayıcı olmak üzere iki ana başlık altında incelenmektedir. Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır. Bu modeller sınıflandırma algoritmaları olarak mesafeye dayalı algoritmalar, yapay sinir ağaçları ve karar ağaçları modellerinden oluşur (Denizli, 2019).

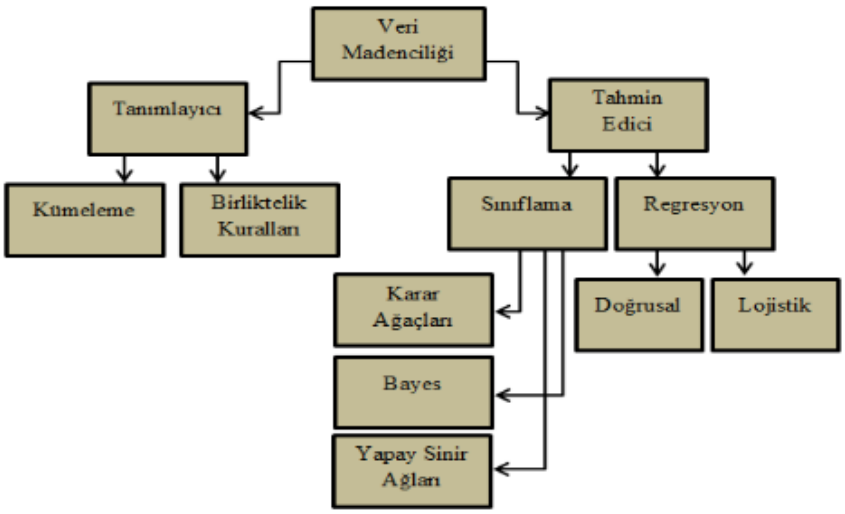
Tanımlayıcı modeller ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlamamızı ve veri tabanındaki verilerin genel özelliklerini karakterize etmemizi sağlamaktadır (Makhabel, 2015).

Veri madenciliği modelleri fonksiyonlarına göre:

1-Sınıflama (classification) ve Regresyon(regression)

2-Kümeleme

3-Birliktelik kuralları şeklinde sıralanırlar.



Şekil 4.1. Veri madenciliği modelleri (Saygılı, 2013)

4.1. Sınıflandırma Teknik ve Algoritmaları

Veri seti içerisindeki verilerin ortak özelliklerine göre ayrıştırılması işlemine sınıflandırma adı verilir. Başka bir deyişle elimizdeki hangi sınıfa ait olduğu bilinmeyen verinin sınıfının belirlenmesi sürecidir.

Sınıflandırma teknikleri en çok kullandığımız veri madenciliği tekniklerinden biridir. Dolandırıcılık tespiti, kalite kontrol çalışmaları, örüntü tanıma ve pazarlama konuları sınıflandırma tekniklerinin en sık kullanıldığı alanlardır. Sınıflandırma teknikleri aslında tahmin edici tekniklerdir.

Sınıflandırma algoritması bir veri kümesi üzerinde tanımlanmış olan sınıflar arasına veri setindeki verileri yerleştirme işlemidir. Sınıflandırma algoritmaları istenilen ve oluşturulmuş kümelerin dağılım şeklini öğrenirler ve daha sonra yeni gelen verileri doğru şekilde sınıflandırmaya çalışırlar. Sınıflandırma modelleri bağımlı değişkenin kategorik olduğu tahminleyici modellerdir. Burada amaç sınıfların belirlenerek bağımsız değişkenin verilen değerlerine göre bağımlı değişkenin değerini tahmin etmektir (Denizli, 2019).

Verilerin sınıflandırma süreci iki aşamadan oluşur (Han ve Kamber, 2006:30). İlk olarak veri tabanına uygun model seçilmelidir. Bu model veri tabanındaki verilerin özelliklerinin kullanılarak oluşturulduğu bir modeldir. Sınıflandırma modelinin kurulması aşamasında veri tabanından bir kısım rasgele olarak ele alınan

veriler kullanılarak bir sınıflandırma modeli oluşturulur. Daha sonra ikinci adımda ise kalan kısım veriler ile sınıflandırma kuralları uygulandıktan sonra aynı kurallar bu verilere de uygulanarak test edilir. Eğer ikinci adımda kullanılan verilerin sınıflandırma kuralları sonucunda elde edilen modelin doğruluğu ispatlanırsa bu model bir sınıflama modeli olarak belirlenebilir.

Sınıflandırma algoritmaları istatistiğe ve mesafeye dayalı algoritmalar, karar ağaçları, yapay sinir ağları gibi yöntemlerle uygulanabilmektedir.

4.1.1. İstatistiğe Dayalı Algoritmalar

Sınıflandırma tekniklerinden biri olan istatistiğe dayalı algoritmalar uygulanırken istatistiksel yöntemler ele alınır. Bağımlı ve bağımsız değişkenler belirlenir ve kullanılacak modeller bu değişkenlere göre oluşturulur. Bu tekniklerden en bilineni ise CHAID algoritmasıdır.

4.1.1.1. CHAID Algoritması

Karar ağaçları (Classification tree; decision tree, CR&T) tekniklerinden CHAID (Chi-squared Automatic Interaction Detection) analizi sınıflama yöntemi ile regresyon analizinin bir arada kullanıldığı bir yöntemdir (Kayri ve Boysan, 2007). CHAID analizi kategorik değişkenlere ilişkin veri kümesini ve bağımlı değişkeni alt gruplara böler. Bu alt kümeler küçük tahmin edici gruplardan oluşur. Ki-kare analizi kullanılarak başlangıç değişkenleri bağımsız olarak yeniden kategorileştirilir. Değişkenlerin bölünmeye uygun olup olmadığına bonferroni düzeltilmiş "p" değeri kullanılarak karar verilir. Bu analiz yöntemi tüm olası alt grupları ağaç biçiminde anlaşılır bir şekilde göstermektedir sonuç olarak bağımlı değişken üzerine etkisi istatistiksel olarak anlamlı olan F değerine sahip olan değişken CHAID şemasında ilk sırada yerini alır (Koyuncugil ve Özgülbaş, 2008).

Her bir alt grup için elde edilmesi gereken ki-kare değerleri

$$Ki - Kare = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(Gij - Bij)^2}{Bij} \qquad (1)$$

4.1.1.2 Basit ve Çoklu Lojistik Regresyon

Değişkenler arasındaki ilişkinin şiddeti hakkındaki bilgiye korelasyon katsayısı ile ulaşılabilir. İki değişken arasındaki ilişkinin matematiksel modeli ise regresyon analizi ile elde edilir. Eğer ikiden fazla değişken ile regresyon analizi uygulanıyorsa buna çoklu regresyon analizi adı verilir. Regresyon analizi iki ya da daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan bir yöntemdir. Eğer yöntemde kullanılacak olan değişken nicel değişkenler ise parametrik testler, nitel değişkenler kullanılacak ise parametrik olmayan testler kullanılmaktadır. Regresyon analizinde değişkenlerden biri bağımlı iken diğeri mutlaka bağımsız olmak zorundadır. Örneğin diyabet olma durumunun(Y_i), yaş(X_{i1}), kilo(X_{i2}), genetik yatkınlık(X_{i3}), ve yeme düzeni(X_{i4}), ile arasındaki ilişkiyi modelleyebilmek için regresyon yöntemlerine ihtiyaç duyarız. Nitel veri analizlerinde ise kullanılabilecek yöntemlerden biri de lojistik regresyon analizidir (Pasin, 2019).

Lojistik regresyon analizi çok değişkenli regresyon analizine her ne kadar benziyor olsa da aralarında önemli farklılıklar vardır. Çoklu regresyon analizinde:

- Bağımlı değişkenin normal dağıldığı

-Bağımsız değişkenler arasında çoklu doğrusal bağlılık olmadığı

-Hata terimlerinin sıfır ortalamalı olduğu

-Varyansın normal dağıldığı

- Gözlemler arasında oto korelasyon bulunmadığı varsayılmaktadır. Bu varsayımlardan herhangi birinin gerçekleşmemesi durumunda çoklu regresyon uygulanamaz. Lojistik regresyonun diğer yöntemlerde olduğu gibi değişkenin normal dağılması ve sürekli olması varsayımları gerektirmeyen durumu sayesinde en çok tercih edilen regresyon analizi yöntemlerinden biridir (Johnson, 1988).

Lojistik regresyon analizi bağımlı değişkenin kategorik olarak gözlendiği durumlarda bağımsız değişkenlerle sebep-sonuç ilişkisini belirlemede kullanılan yöntemdir. Burada dikkat edilmesi gereken bağımlı değişkenin niteliğidir. Bağımlı değişkenin iki kategorili olması durumunda binary, ikiden fazla ve sırasız kategorili

olduğu durumda çok kategorili multinominal, ikiden fazla ve sıralı kategorili olduğu durumda ise ordinal lojistik regresyon modeli kullanılmaktadır. Lojistik regresyonun temel amacı bağımsız değişkenlerin, bağımlı değişkenleri ne derecede tahmin edebildiğini araştırmaktır.

Lojistik regresyon uygulaması yapılırken bağımsız değişkenin dağılımına ilişkin herhangi bir varsayım sağlanması gerekmese de oluşturulan modelin sağlıklı sonuçlar verebilmesi için bazı varsayımların sağlanması gerekmektedir. Bu varsayımlar şunlardır:

- Gelecek tahmini yapılabilmesi için oluşturulan matematiksel modelde bilgi karmaşasına yol açacak değişkenlerin modele dahil edilmesi modelin oluşmasında güçlük yaratabilmektedir. Bu sebeple modelin matematiksel kalıbının doğru bir şekilde oluşturulması gerekmektedir.

- Bağımsız değişkenler arasında çoklu doğrusal bağlantı olmamalıdır. Eğer bağımsız değişkenler arasında çoklu bağlantı var ise bu modelin güvenilirliğini anlamlı ölçüde olumsuz etkiler.

- Lojistik regresyon modelinin hata terimleri arasında oto korelasyon yoktur.

- Gözlem sayısı tahmin edilecek parametre sayısından fazla olmalıdır.

Lojistik regresyon en az değişken ile bağımlı ve bağımsız değişken arasındaki ilişkiyi gösteren bir model kurmaya ve sınıflama yapabilmeye yardımcı olan bir yöntemdir. Lojistik regresyonu doğrusal regresyondan ayıran en temel özelliği lojistik regresyonda bağımlı değişkenin kategorik olmasıdır. Doğrusal regresyonda bağımlı değişkenin değeri tahmin edilir. Fakat lojistik regresyonda bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilir.

Lojistik regresyonun avantajları:

- Regresyon modelindeki bağımlı değişken kesikli iken bağımsız değişkenin hem sürekli hem kesikli olabilmesi

- Modeldeki bağımsız değişkenlerin olasılık fonksiyonlarının dağılımları üzerinde herhangi bir kısıt olmaması

-Lojistik regresyon modeli üzerindeki parametrelerin kolay ve açıklanabilir olması

- Lojistik regresyondaki parametrelerin olasılık değerlerinin 0 ile 1 aralığında olması

- Lojistik regresyonun yaygın kullanılan SPSS ve Minitab gibi programlarda uygulanabilmesi.

Lojistik regresyon modeli doğrusal regresyon modelinin bir alternatifidir.

Basit doğrusal regresyon modeli:

$$Yi = \beta + \beta Xi + \varepsilon \qquad (2)$$

şeklinde ifade edilir.

Y; bağımlı değişken,

β; regresyon katsayıları,

X; bağımsız değişkenler,

Ɛ; hata terim

Doğrusal regresyon modelinde bağımsız değişkenler üzerinde bir kısıt yok iken bağımlı değişkenin sürekli olma koşulu vardır. Y_i bağımlı değişkeni -∞ ve +∞ değerleri arasında yer alır. Bağımlı değişkenin (Y_i) 0-1 değerleri gibi kategorik değerler alması durumunda ise dağılım bozulmaktadır. Böyle durumlarda *i*'nci gözlemin 1 değerini alma olasılığı P(yi = 1) ise beklenen değer :

$$E(yi) = 1 \times P(yi = 1) + 0 \times P(yi = 0) = P(yi = 1) \qquad (3)$$

eşitliği ile elde edilir.

Doğrusal regresyon analizinde kullanılan yöntemler lojistik regresyonda da uygulanabilir. Lojistik regresyon fonksiyonu:

$$\pi(x) = E(Y|x) = \frac{e^{\beta 0+\beta 1x}}{1+e^{\beta 0+\beta 1x}} \qquad 0 \leq E(Y|x) \leq 1 \qquad (4)$$

Şeklinde ifade edilir.

4.1.1.3. Bayesyen Sınıflandırma

Olasılık teorisi içerisinde incelenen bir B olayının bilindiği durumda A olayı için olasılık değeri, A olayının bilindiği durumda B olayı için olasılık değerinden farklıdır. Ancak bu iki olay arasında kuvvetli bir ilişki vardır ve bu ilişkiye bayes teoremi denilmektedir.

Bu teorem bir rassal değişken için olasılık dağılımı içinde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişkiyi gösterir. Bu şekli ile Bayes teoremi bütün istatistikçiler için kabul edilir bir ilişkiyi açıklar. Elimizde olan sınıflanmış verileri kullanarak yeni bir verinin var olan sınıflardan herhangi birine girme olasılığını hesaplar.

Xi sınıf üyeliği bilinmeyen bir veri örneği olsun ve örnek $xi = (x1, x2, \dots, xn)$ nitelik değerlerinden oluşsun. Bu örnek sınıfta m adet sınıf olduğu ve C1, C2, ...,Cm sınıf değerleri olduğu düşünülürse, xi'nin Ci sınıfında olma olasılığı aşağıdaki formül ile değerlendirilebilir.

$$P(Ci/xi) = \frac{p\,(xi/Ci)\;P(Ci)}{P(xi)} \qquad (5)$$

Burada $P(xi)$, xi değerinin $P(Ci)$ sınıfının veri tabanında bulunma olasılığıdır. Hesaplardaki işlem yükünü azaltmak ve yeni veriyi sınıflandırmak için paydada yer alan $P(xi)$ değerleri birbirine eşit olduğundan sadece pay değerlerini karşılaştırmak yeterlidir. Bu değerler içinden en büyük olanı seçilir ve bilinmeyen örneğin bu sınıfa ait olduğu belirlenir.

4.1.2. Mesafeye Dayalı Sınıflandırma Algoritmaları

Sınıflandırma yapılırken verilerin birbirlerine olan uzaklık yakınlık durumundan faydalanarak sınıflandırmanın yapılması mesafeye dayalı sınıflandırma teknikleri olarak adlandırılmaktadır. Bu tekniklerden en çok kullanılanı ise "K-en yakın komşu algoritmasıdır".

4.1.2.1. K en yakın komşu algoritması

K en yakın komşu (k-nn) algoritması yeni bir verinin en yakın(k) komşu verilerine göre etiketlendiği sınıflandırma yöntemi olarak kullanılabilir (Gorunescu, 2011).

Bu algoritmada sınıflandırılmak istenen yeni verinin daha önceki verilerden k tanesine yakınlığına bakılır. Örneğin k=4 için yeni bir eleman sınıflandırılmak istensin. Bu durumda eski sınıflandırılmış elemanlardan en yakın 4 tanesi alınır. k değeri çok küçük olursa model çok etkilenir. Çok büyük olursa da tek bir sınıf gibi olur. Yani k sayısının sınıflandırmaya etkisi vardır. Kolay anlaşılabilir bir algoritma olduğu için uygulaması basit olması ve gürültüye sahip veriler için de olumlu sonuçlar ortaya koyması avantajlarındandır. Fakat uzaklığa bağlı bir sınıflandırma yöntemi olduğu için uzaklık ölçütünün kullanılmasına dair bir kesinlik olmaması ve tahminin bölgesel bilgilere dayalı olması ve bu nedenle aykırı değerlerden etkilenmesi olasılığı dezavantajıdır.

Bu algoritmanın adımları:

- Elde edilen yeni bilgi sınıfa eklenir

- K komşusuna bakılır

- Uzaklık fonksiyonları kullanılarak(çoğu kez öklid) uzaklık hesaplanır.

- En yakın neresi ise bu veri oraya atanır.

Bazı uzaklık fonksiyonları şu şekildedir:

- Manhattan uzaklık fonksiyonu:

$$d(i,j):|x_{i1}- x_{j1}|+|x_{i2}-x_{j2}|+\ldots. \qquad (6)$$

- Minkowski uzaklık fonksiyonu:

$$d(i,j):(|x_{i1}-x_{j1}|^{q}+|x_{i2}-x_{j2}|^{q}+\ldots)^{1/q} \qquad (7)$$

- Öklid uzaklık fonksiyonu:

$$d(i,j):(|x_{i1}-x_{j1}|^{2}+|x_{i2}-x_{j2}|^{2}+\ldots)^{1/2} \qquad (8)$$

4.1.2.2. En küçük mesafe sınıflandırıcısı

Bu algoritma da k en yakın komşu algoritmasındaki gibi mesafeye dayalı sınıflandırma tahmininde bulunur. En küçük mesafe sınıflandırıcısı öklid fonksiyonunu kullanarak herhangi bir değerin hangi sınıfa ait olduğunu belirleyen bir algoritmadır.

Öklid formülü: $dis(X,Y)=\sqrt{\sum_{n=1}^{n}(xi-yi)^2}$ (9)

Algoritmanın aşaması:

$X_0=\frac{\sum_{i=1}^{nN} x \min}{N}$

$Dz\ (x) = X^T X_0 - ½\ (X_0^T\ X_0)$

X verisini en büyük sınıfa ata

X_0: Her bir sınıfın orta noktası

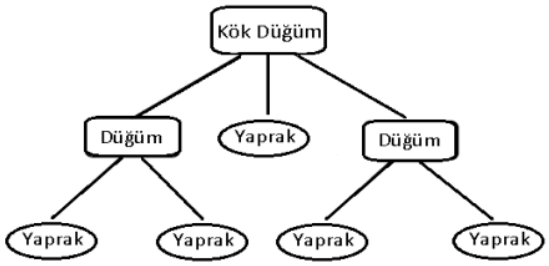
Sınıflandırılması istenen x değeri Dz değeri büyük olan sınıfa atanır.

4.1.3 Karar Ağaçları ve Algoritmaları

Karar ağaçları sınıflandırma yöntemlerinden en çok kullanılan yöntemlerdendir. Diğer yöntemlere göre karar ağaçlarının anlaşılması ve yorumlanması daha kolaydır. Karar ağaçları başarılı modeller üretebilmesi ile en öne çıkan sınıflandırma yöntemlerinden biridir. Karar ağaçları yönteminde sınıflandırma yapmak için elimizdeki verilerden ağaç oluşturulur ve bu veriler bu ağaca dağıtılır, çıkan sonuca göre de sınıflandırma işlemi gerçekleşmiş olur. Yani veri tabından elde edilen karar ağaçlarına göre hangi sınıfa ait olduğunu bilmediğimiz bir veri elimize geldiğinde oluşturulan kural dizisine göre hangi sınıfta olması gerektiği tahmin edilir (Silahtaroğlu,2013).

Karar ağaçları adından da anlaşılabileceği üzere ağaç şeklinde gösterilebilmektedir. Düğüm noktaları bulunan bu ağaçlar dal ve yapraklardan oluşmaktadır. Sınıflandırma işlemi en tepede bulunan kök düğüm ile başlar. Sınıflandırma süreci yaprak elde edilinceye kadar niteliklerin alt dallara ayrılmasının

devam etmesiyle oluşur. Burada düğümler test işlemine dahil olan verileri gösterir. Yapraklar ise veri setindeki bir karar sınıfını temsil etmektedir. Dalın sonunda sınıflandırma tamamlanmadıysa tekrar düğüm oluşturulur.



Şekil 4.2. Karar ağacı örneği (Silahtaroğlu,2013)

Karar ağaçlarının avantajları şu şekilde sıralanabilir:

- Kolay anlaşılabilir ve yorumlanabilir olmasından dolayı daha kullanışlıdır.

- Çok büyük boyutlardaki veri setlerine rahatlıkla uygulanabilir.

- Karar ağaçları non-parametriktir ve diğer tekniklerde olduğu gibi istatistiksel varsayımlarla uğraşmaz

- Sürekli ve nitel değişkenlerle kullanılabilir

- En çok kullanılan veri madenciliği yöntemi olduğu için ulaşılabilir vaziyette olan birçok kaynaktan yararlanılabilir.

Karar ağaçlarının bazı dezavantajları ise şu şekilde sıralanabilir:

- Özellikle büyük veri setlerinde en uygun ağacı bulmak bazen zor olabilir. Yararlı karar ağacı oluşturulmaya çalışılırken karmaşık ve büyük dallı ağaçlar oluşturulabilir. Ayrıca bu büyük veri setlerinin sınıflandırılması daha maliyetli olabilir.

- Girdi sayısının az olduğu durumda ağaç yeteri kadar dallanamayabilir.

- Bazı algoritmalar sadece kategorik değişkenlerle çalışır.

Karar ağaçları oluşturulurken hangi algoritmanın kullanılacağına karar vermek çok önemlidir. Kullanılan algoritmaya göre ağaç yapısı farklılık gösterebilir. Bu nedenle farklı ağaç yapıları farklı kurallar oluşturabilmektedir. Karar ağaçlarına dayalı olarak geliştirilen algoritmalar kök düğümün ve diğer düğümlerin seçimlerinde izledikleri yol açısından birbirlerinden ayrılırlar, ancak bazı noktalarda ise birbirlerine benzemektedirler ya da benzerlik gösterirler (Silhataroğlu, 2013).

Karar ağaçlarında kullanılan temel algoritma adımları birbirlerine benzemektedirler. İlk olarak kök düğüm oluşturulur ve eğer tüm bilgiler aynı sınıfa ait ise düğüm yaprağa dönüşür. Eğer aynı sınıfa ait değil ise bölünme gerçekleşir ve dallar oluşturulur. Genel olarak kategorik değişkenler kullanılır. Bir düğümde bulunan bütün örnekler aynı sınıfa ait ve örnekleri bölecek nitelik kalmadığında(başka örneklem kalmadığında) bölünme işlemi sonlandırılır. Karar ağaçları için geliştirilen algoritmalardan bazıları şunlardır:

-ID3 algoritması (Iterative Dichotomiser 3)

-C4.5 ve C5 algoritmaları

- CART algoritması (Classification and Regression Trees)

- SPRINT algoritması (Scalable PaRallezlizable INduction of Decision Trees)

4.1.3.1. ID3 Algoritması

İlk olarak J. Ross Quinlan tarafından Sydney Üniversitesinde geliştirilmiştir. Bu algoritma diğer değişkenler arasından sınıflamada kullanılacak en ayırt edici değişkeni bulurken entropi kavramından yararlanır. Entropi; eldeki verilerin sayısallaştırılması, beklentisizliğin maksimumlaştırılması ve belirsizliğin ölçülmesi için kullanılır. Entropi değeri 0-1 arasında değer alır. Örneğin herhangi bir mağazada son kalan 10 adet ürünün aynı olması durumunda mağazaya giren bir kişiye hangi ürünü alacağının sorulması halinde alınacak cevap bellidir ve entropi değeri "0" değerini alır. Bütün olasılıkların eşit olduğu durumda ise entropi maksimum değerini alır. Aşağıdaki formül ile ifade edilir:

$$H(T) = -\sum pi\ log2(pi) \qquad (10)$$

Karar ağacının oluşturulacağı veri setinin tamamının entropisi hesaplanır. Fakat dallanma gerçekleştiği zaman oluşturulan her bir bölüm için entropinin yeniden hesaplanması gerekir. Bu algoritmada veri tabanında doğru dallandırma yapmak için bilgi kazancı (information gain) kullanır. Dallandırma sayısı arttıkça doğru sınıflandırmalar için gereken bilgi azalmaktadır. Kazanımı hesaplamak için: verilerin ham halinin entropisi ile her bir nitelik için hesaplanan entropilerin ağırlık toplamı arsandaki fark alınır. Bu fark hangi alt bölüm için büyük ise o alt bölüme doğru dallandırma işlemi gerçekleştirilir. Bilgi kazancının formülü aşağıdaki gibi ifade edilir;

Kazanç(X,T) = H(T) – H(X,T) (11)

$$H\ (X,T) = \sum_{i=1}^{n} (\frac{Ti}{T})$$

Buradaki amaç Kazanç (X,T) değerini maksimum yapan değeri bulmaktır (Diler, 2016).

4.1.3.2. C4.5 ve C5 Algoritması

C4.5 algoritması ID3 algoritmasının geliştirilmesiyle elde edilmiştir. ID3 algoritmasıyla oluşturulan karar ağacı modellerinde eksik veriler yok sayılır. Fakat C4.5 algoritması veri setinde bulunan eksik gözlemleri diğer verilerin özellik yapılarına göre tahmin ederek bir karar ağacı oluşturabilir. Burada da temel amaç verileri sınıflara bölmektir. C5 algoritması ise, C4.5 algoritmasının ticari bir sürümüdür. C4.5'in geliştirilmiş halidir. C4.5'e kıyasla daha hızlı ve güvenilir kurallar oluşturulduğu gözlemlenmiştir.

C4.5 algoritmasında ID3 algoritmasından farklı olarak aşırı öğrenmeyi engellemek için kazanç yerine kazanç oranı kullanılmaktadır.

Kazanç Oranı (D,S)= Kazanç(D;S)/ Bölünme bilgisi(D,S)

$$Bölünme\ bilgisi\ (D,S) = H(\frac{D1}{D}, \dots, \frac{Ds}{D}) \quad (12)$$

Burada maksimum kazanç oranına sahip olan nitelik bölünme niteliği olarak seçilir.

4.1.3.3. CART (Classification and Regression trees)

İlk olarak 1984'te önerilmiş ve uygulamaları yapılmıştır. Bu teknik ID3 algoritmasında olduğu gibi dalları ayırmada en uygun kriteri seçmek için entropiden yararlanır. Fakat burada en uygun kriteri belirlemek için ID3 ve C4.5 algoritmalarından farklı bir formülizayson kullanılır (Yaşasınoğlu, 2019).

Cart algoritmasında entropi değeri şu şekilde hesaplanır:

Herhangi bir t düğümündeki s dallara ayrılma kriteri ϕ (s/t) olarak gösterilirse;

$$\phi(s/t)=2P_L P_R \sum |P(J|tL) - P(J|tR))| \quad (13)$$

t:Dallanmanın yapılacağı düğüm

C: Kriter

L: Ağacın sol tarafı

R: Ağacın sağ tarafı

PL ,PR:Öğrenim kümesindeki bir kaydın sağda ve solda olma olasılıkları

$P(J|tL|)$ ve $P(J|tR|)$: CJ sınıfındaki bir kaydın sağda ve solda olma olasılıkları

4.1.3.4. SPRINT Algoritması

Sprint algoritması(Scalable Parallelizable Induction of Decision Tree) ağaç yapısında en uygun dallanmayı sağlayabilmek için her bir değişkene ait veriyi sıraya dizerek ağaç yapısını bu şekilde oluşturmaktadır. Bu algoritmada dallandırma kriteri olarak Gini indeksinden faydalanılır. Büyük veri kümeleri için oldukça uygun bir algoritmadır. Öncelikle tüm değişkenler için ayrı bir değişken listesi hazırlanmaktadır. Değişken sayısı kadar tablo oluşturulmaktadır. Burada her bir tabloda kullanılacak olan değişken, sınıf ve sıra numarası yer almaktadır.

4.1.4. Birliktelik Kuralları ve Market Sepet Analizi

Günümüzde birçok alandaki veriler bilgisayarların veri tabanlarına işlenmektedir. Bu verilerden kullanışlı ve işe yarayan bilgileri elde edebilmek için kullanılan bir diğer veri madenciliği yöntemi ise birliktelik kurallarıdır. Birçok kullanım alanına sahip olan birliktelik kuralları özellikle niteliklerin veya nesnelerin bir arada olma durumlarını araştırmak için kullanılır (Yüksel, 2018).

Birliktelik kurallarının araştırılmasında çok fazla nitelik üzerinde durulduğundan oldukça maliyetlidir. Bu nedenle önceki çalışmalarda belirlenen birliktelik kurallarının korunması büyük önem taşımaktadır (Ateş ve Karabatak, 2017).

Avantajları:

- Büyük miktardaki veri setleri üzerinde uygulanabilmesi

- Açık ve anlaşılır sonuçlar elde etmesi

- Anlaşılabilir hesaplamalar olması

Dezavantajları:

- Seyrek görülen özelliklerin göz ardı edilmesi

- Büyük veri setlerinden anlaşılır veri bulabilmek için çok fazla zaman alması

Birliktelik kuralı problemi ilk defa 1993 yılında market sepet verisi üzerinde uygulanmıştır. Birliktelik kuralları tanımlayıcı veri madenciliği yöntemlerinden biridir ve büyük miktardaki veri yığınları içerisinden daha önce oluşturulmayan örüntüleri oluşturmak için kullanılır. Son yıllarda marketlerde uygulanan barkod tanıma sistemlerinin firmaların satış durumlarını uygun bir veri tabanına işleyebilmelerini sağlamıştır.

Market sepet verileri olarak adlandırılan bu veriler daha çok süpermarketlerden elde edilebilmektedir. Bu verileri kullanarak bünyesine fayda sağlayan birçok kuruluş vardır. Temel amaç ürün satışlarının birbirleri arasındaki ilişkiyi belirleyerek şirketin karını artırmaktır (Ateş ve Karabatak, 2017).

Süpermarketler gün sonunda yaptıkları satışlardan elde ettikleri verilerin içerisinden anlamlı ilişkiler elde etmek için market sepet analizini kullanırlar. “Eğer müşteriler A ürününü alıyorlarsa %x oranında B ürününü de alabilirler” şeklinde bir sonuç elde edebilmektedirler ve bu sayede bünyelerine fayda sağlayabilmektedirler. Katalog düzenlemesinde, çapraz satış ve mağaza raflarının düzenlenmesinde market sepet analizlerinden elde edilen bilgilerden yararlanılır.

Büyük veri tabanlarında birliktelik kuralları 2 adımda uygulanır:

1- Sık tekrarlanan gözlemler bulunur.
2- Sık tekrarlanan öğelerden anlamlı birliktelik kuralları oluşturulur.

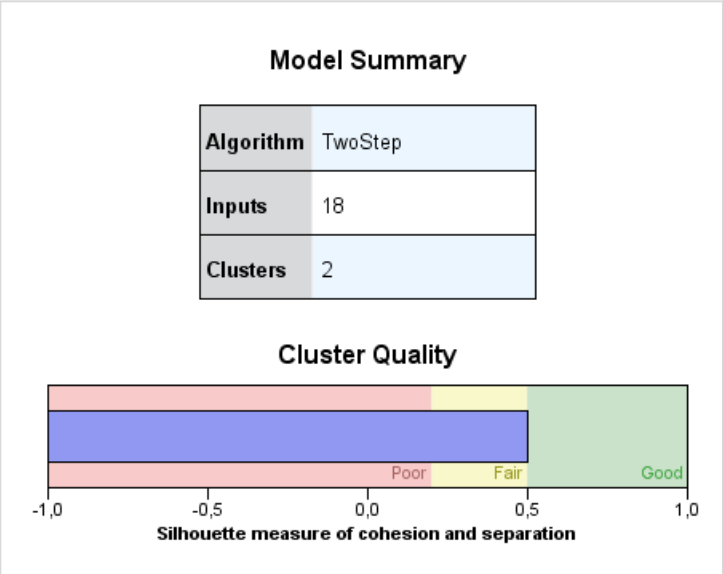
Birliktelik kurallarında en sık kullanılan algoritma apriori algoritmasıdır.

4.1.4.1. Apriori Algoritması

Apriori algoritması 1994 yılında Agraval ve Srikant tarafından oluşturulmuştur. En bilinen ve yaygın olarak kullanılan birliktelik kuralları algoritmasıdır. Bu algoritma teknik olarak sık tekrarlanan bir gözlem kümesinin alt kümesinin de sık tekrarlanan bir gözlem kümesinin olacağını esas alır. Apriori algoritması veri tabanından birliktelik kuralları oluşturabilmek için çok sayıda tarama uygular ve bu taramaların ilkinde elde edilen sık tekrarlanan kümelerini, ikinci taramada ise aday gözlem kümelerini oluşturmak amacıyla kullanır. Üçüncü taramada ise, ikinci tarama sırasında elde edilen tekrarlanan gözlem kümelerini bu tarama sırasında aday gözlem kümeleri olarak kullanır ve bu iterasyon hiç tekrarlanan gözlem kümesi kalmayıncaya kadar devam eder.

5.UYGULAMA ve ANALİZ

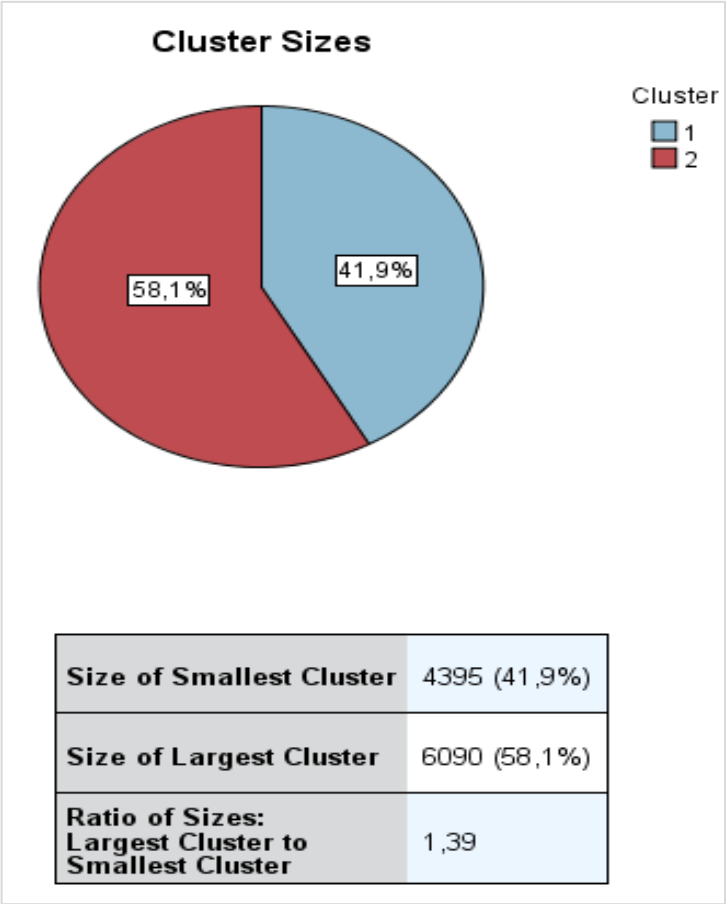
Bu çalışmada 10485 diyabet mellitus (DM) hastasının kullandıkları ilaçların Hba1c değerleri üzerindeki etkisi veri madenciliği teknikleri ile araştırılmıştır. Araştırmaya katılma şartı bireyin diyabet hastası olması ve en az bir ilaç kullanmasıdır. Çalışmada kullanılan veri seti "**Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records**" adlı çalışmadan elde edilmiştir (Strack ve diğerleri, 2014). Çalışmanın istatistiksel analiz aşamasında öncelikle verileri sınıflandırmak amacıyla kümeleme analizi uygulanmıştır. Daha sonra üç farklı karar ağacı tekniği (CHAID, CART ve QUEST) kullanılmıştır. Bağımlı değişkenin Hba1c olduğu durum için kullanılan ilaçların, cinsiyetin ve yaşın bağımsız olarak belirlenme durumunda lojistik regresyon analizi uygulanmış ve hangi ilacınHba1cüzerinde ne derece etkili olduğu bilgisine ulaşılmıştır. Son olarak birliktelik kuralları algoritması ile hangi ilaçların birlikte kullanıldığı araştırılmıştır. Araştırmaların tamamı SPSS ve R project paket programları ile değerlendirilmiştir.



Şekil 4.3. Silhouette indeksine göre iki aşamalı küme analizi çıktısı

Şekil 3'te araştırmaya dahil olan 18 değişkene ait kümeleme dağılımının sonuçları verilmiştir. Sonuçlara göre araştırmaya dahil edilen bu 18 değişken 2 farklı

kümede toplanmıştır. Ayrıca silhouette indeksi 0.50 olarak hesaplanmış ve kümeleme kalitesi orta düzeyli olarak hesaplanmıştır.



Size of Smallest Cluster	4395 (41,9%)
Size of Largest Cluster	6090 (58,1%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1,39

Şekil 4.4 Kümeleme şeması

Şekil 4.4'te kümeleme sonuçlarına göre elde edilen grafik sonuçları verilmiştir. Araştırmaya katılan bireylerin %41.9'u birinci kümeyi, %58.1'i ise ikinci kümeyi oluşturmaktadır.

Tablo 4.1 Araştırmaya dahil olan bireylerin ilaç kullanma durumlarına göre kümeleme sonuçları

Değişken	Küme		p
	1. Küme (n=4395)	2. Küme (n=6090)	
Metformin			
-	4395 (100%)	3110 (51.1%)	**<0.001**
+	0 (0.00%)	2980 (48.9%)	
Repaglinide			
-	4395 (100%)	5805 (95.3%)	**<0.001**
+	0 (0.00%)	285 (4.68%)	
Nateglinide			
-	4395 (100%)	6009 (98.7%)	**<0.001**
+	0 (0.00%)	81 (1.33%)	
Chlorpropamide			
-	4395 (100%)	6083 (99.9%)	**0.047**
+	0 (0.00%)	7 (0.11%)	
Glimepiride			
-	4395 (100%)	5304 (87.1%)	**<0.001**
+	0 (0.00%)	786 (12.9%)	
Glipizide			
-	4395 (100%)	4382 (72.0%)	**<0.001**
+	0 (0.00%)	1708 (28.0%)	
Glyburide			
-	4395 (100%)	4695 (77.1%)	**<0.001**
+	0 (0.00%)	1395 (22.9%)	
Tolbutamide			
-	4394 (100.0%)	6089 (100.0%)	1.000
+	1 (0.02%)	1 (0.02%)	
Pioglitazone			
-	4395 (100%)	5205 (85.5%)	**<0.001**
+	0 (0.00%)	885 (14.5%)	
Rosiglitazone			
-	4395 (100%)	5227 (85.8%)	**<0.001**
+	0 (0.00%)	863 (14.2%)	
Acarbose			
-	4395 (100%)	6044 (99.2%)	**<0.001**
+	0 (0.00%)	46 (0.76%)	
Miglitol			
-	4395 (100%)	6083 (99.9%)	**0.047**
+	0 (0.00%)	7 (0.11%)	
Tolazamide			
-	4395 (100%)	6085 (99.9%)	0.079
+	0 (0.00%)	5 (0.08%)	
İnsulin			
-	0 (0.00%)	2334 (38.3%)	**<0.001**
+	4395 (100%)	3756 (61.7%)	

glyburide_metformin			
-	4395 (100%)	6016 (98.8%)	**<0.001**
+	0 (0.00%)	74 (1.22%)	
glipizide_metformin			
-	4395 (100%)	6088 (100.0%)	0.513
+	0 (0.00%)	2 (0.03%)	
metformin_rosiglitazone			
-	4394 (100.0%)	6090 (100%)	0.419
+	1 (0.02%)	0 (0.00%)	
Hba1c			
8'den düşük	1085 (24.7%)	1987 (32.6%)	**<0.001**
8 veya 8'den yüksek	3310 (75.3%)	4103 (67.4%)	

(+): ilacı kullanıyor, (-): ilacı kullanmıyor

Tablo 4.1'de araştırmaya dahil olan bireylerin ilaç kullanma durumları ile küme grupları arasındaki ilişki testi sonuçları verilmiştir. Sonuçlara bakıldığında birinci kümede bulunan bireylerin hepsi sadece insülin tedavisi görmektedir. İkinci kümede ise insülin+OAD (oral antidiabetik ilaçlar) kullanan bireyler bulunmaktadır. Tablodaki metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, glyburide, pioglitazone, rosiglitazone, acarbose, miglitol, insulin ve glyburide+metformin ilaçlarının kullanım durumları oluşturulan kümeler üzerinde anlamlı bir etkiye sahiptir. Glipizide+metformin, tolbutamide, tolazamide ve metformin_rosiglitazone ilaçlarının kullanma durumları ise oluşturulan kümeler üzerinde anlamlı bir etkiye sahip değildir. Bunun nedeni ise bu ilaçları araştırmaya dahil olan bireylerin neredeyse hepsinin kullanmamasıdır.

Tabloya dahil edilen Hba1c değerleri 2 farklı grupta ele alınmıştır. Bireylerin Hba1c değeri 8'den düşük ve 8 veya 8'den yüksek olarak değerlendirilmiştir (DSÖ). Sonuçlara bakıldığında ikinci kümede bulunan ve Hba1c değeri 8'den düşük olan bireyler, birinci kümede bulunan veHba1c değeri 8'den düşük olan bireylere göre daha fazladır.

Tablo 4.2 Hba1c değerini modellemek için uygulanan lojistik regresyon modellerinin performans metrikleri

Model	-2 Log Olabilirlik	Cox-Snell R^2	Nagelkerke R^2	DSO
Full model	11718.375	0.088	0.125	70.611
Backward model	11722.920	0.087	0.125	70.770
Final model	11730.475	0.087	0.124	70.892

Tablo 4.2'de Hba1c değerini modellemek için uygulanan lojistik regresyon modellerinin performans metrikleri gösterilmektedir. Performans metriklerine göre üç modelin R^2 ölçüleri ve doğru sınıflama oranları (DSO) birbirine oldukça yakındır. Son modelin sonuçlarına göre modelin doğru sınıflama oranı %70.892 düzeyindedir ve R^2 ölçüleri de sıfırdan farklıdır.

Tablo 4.3. Hba1c değerini modellemek için uygulanan lojistik regresyon modellerinin tahmin sonuçları

Variable	Full model			Backward-model			Final model		
	OR	Wald	p	OR	Wald	p	OR	Wald	p
Cinsiyet (1)	0.966	0.574	0.449	-	-	-	-	-	-
Yaş	-	629.024	**<0.001**	-	641.877	**<0.001**	-	640.027	**<0.001**
Yaş (1)	33.154	22.723	**<0.001**	33.069	22.700	**<0.001**	33.115	22.718	**<0.001**
Yaş (2)	21.217	77.148	**<0.001**	21.186	77.222	**<0.001**	21.217	77.293	**<0.001**
Yaş (3)	11.065	83.368	**<0.001**	11.055	83.584	**<0.001**	11.068	83.664	**<0.001**
Yaş (4)	6.393	84.034	**<0.001**	6.368	84.385	**<0.001**	6.352	84.205	**<0.001**
Yaş (5)	3.949	61.495	**<0.001**	3.931	61.946	**<0.001**	3.934	62.017	**<0.001**
Yaş (6)	3.122	45.363	**<0.001**	3.115	45.769	**<0.001**	3.105	45.520	**<0.001**
Yaş (7)	1.972	16.549	**<0.001**	1.969	16.648	**<0.001**	1.972	16.719	**<0.001**
Yaş (8)	1.298	2.456	0.117	1.298	2.468	0.116	1.300	2.501	0.114
Yaş (9)	0.989	0.004	0.947	0.992	0.002	0.961	0.994	0.001	0.972
Metformin(1)	1.004	0.007	0.932	-	-	-	-	-	-
Repaglinide (1)	1.015	0.014	0.906	-	-	-	-	-	-
Nateglinide (1)	1.104	0.165	0.685	-	-	-	-	-	-
Chlorpropamide (1)	1.267	0.088	0.767	-	-	-	-	-	-
Glimepiride (1)	0.839	4.210	**0.040**	0.837	4.350	**0.037**	0.832	4.622	**0.032**
Glipizide (1)	0.801	12.179	**<0.001**	0.801	12.327	**<0.001**	0.798	12.821	**<0.001**
Glyburide (1)	0.728	20.584	**<0.001**	0.728	21.035	**<0.001**	0.724	21.699	**<0.001**
Tolbutamide (1)	0.000	0.000	0.999	-	-	-	-	-	-
Pioglitazone (1)	0.954	0.359	0.549	-	-	-	-	-	-
Rosiglitazone (1)	0.976	0.090	0.764	-	-	-	-	-	-
Acarbose (1)	0.988	0.001	0.972	-	-	-	-	-	-
Miglitol (1)	0.213	1.999	0.157	0.215	1.984	0.159	-	-	-
Tolazamide (1)	3.401	1.673	0.196	3.373	1.653	0.199	-	-	-
İnsulin (1)	0.554	110.351	**<0.001**	0.555	116.285	**<0.001**	118.508	0.000	0.553
Glyburide metformin (1)	1.026	0.010	0.918	-	-	-	-	-	-
Glipizide metformin (1)	1.95E+09	0.000	0.999	1.99E+09	0.000	0.999	-	-	-
Metformin_ osiglitazone (1)	0.000	0.000	1.000	-	-	-	-	-	-
Constant	1.68E+09	0.000	1.000	0.000	0.000	0.999	18.351	0.000	2.387

Tablo 4.3'te Hba1c değerini modellemek için uygulanan üç farklı lojistik regresyon modelinin tahmin sonuçları verilmektedir. Modellerde her bağımsız değişken için odds oranları (OR), Wald istatistikleri ve anlamlılık değerleri verilmiştir. Tahmin edilen modellerde yer almayan değişkenler (-) yer almaktadır. Son modelin lojistik regresyon sonuçlarına göre yaş grupları, glimepiride, glipizide ve glyburide ilaç kullanım durumlarınınHba1c değerinin modeli üzerinde istatistiksel olarak anlamlı etkiye sahiptir ($p<0.05$).

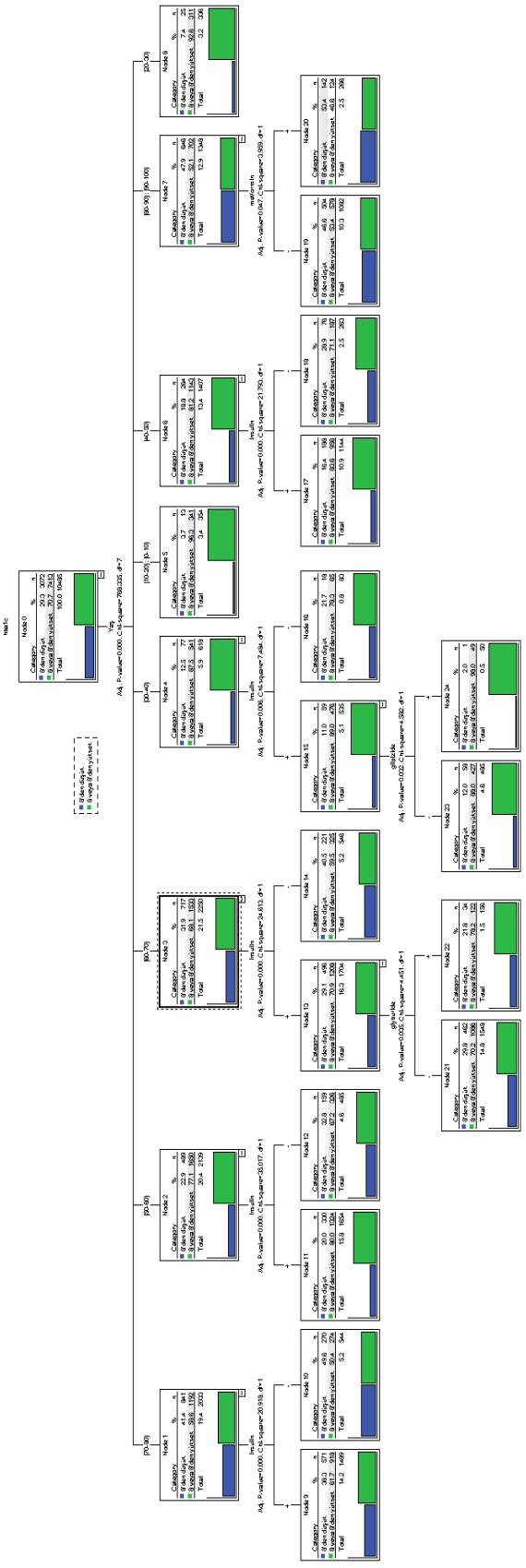
Araştırmada bağımlı değişken olarak alınan Hba1c değerini modellemek için ikili (binary) lojistik regresyon analizi uygulanmıştır. Lojistik regresyon analizi için üç aşamalı bir yol izlenmiştir. Öncelikle hipotez testi aşamasında Hba1c ile istatistiksel olarak anlamlı ilişkiye sahip olan demografik faktörler üzerinden lojistik regresyon modeli (Full-model) tahmin edilmiştir. Daha sonra full-model içerisinde istatistiksel olarak anlamlı bulunmayan değişkenleri elemek için backward-lojistik regresyon analizi tekniği ile değişken seçimi uygulanmıştır. Backward aşamasında modele değişken eklerken $\alpha = 0.05$, modelden değişken çıkarırken de $\alpha = 0.10$ hata payı baz alınmıştır. Son olarak, backward aşamasında istatistiksel olarak anlamlı bulunan bağımsız değişkenler üzerinden lojistik regresyon modeli kurularak nihai çıkarımlar bu modele (Final-model) göre yapılmıştır.

Tablo 4.4 Birliktelik kuralları sonuçları

Birliktelik kuralları		Destek	Güven	Lift
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, acarbose, tolazamide=-, glyburide=-_metformin=-, glipizide=-_metformin=-}	=> **{insulin=+}**	0.501	0.935	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, tolbutamide=-, acarbose, tolazamide=-, glyburide=-_metformin=-, glipizide=-_metformin=-}	=> **{insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, acarbose, tolazamide=-, glyburide=-_metformin=-, glipizide=-_metformin=-, metformin=-_rosiglitazone=-}	=> **{insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, acarbose, miglitol, tolazamide=-, glyburide=-_metformin=-, glipizide=-_metformin=-}	=> **{insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, acarbose, tolazamide=-, glyburide=-_metformin=-}	=> **{insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-,	=> **{insulin=+}**	0.501	0.934	1.202

glyburide=-, tolbutamide=-, acarbose, tolazamide=-, glyburide=-_metformin=-}				
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, acarbose, tolazamide=-, glyburide=-_metformin=-, metformin=-_rosiglitazone=-}	**=> {insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, acarbose, miglitol, tolazamide=-, glyburide=-_metformin=-}	**=> {insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, tolbutamide=-, acarbose, tolazamide=-, glyburide=-_metformin=-, metformin=-_rosiglitazone=-}	**=> {insulin=+}**	0.501	0.934	1.202
{metformin=-, chlorpropamide=-, glipizide=-, glyburide=-, tolbutamide=-, acarbose, miglitol, tolazamide=-, glyburide=-_metformin=-}	**=> {insulin=+}**	0.501	0.934	1.202

Tablo 4.4'te birliktelik kuralları sonuçları verilmiştir. Sonuçlara bakıldığında oluşan kombinasyonların kaldıraç (lift) değerlerinin aynı olduğu görülmektedir. Buna göre oluşan kombinasyonların özgünlükleri de birbirine oldukça yakındır. Birliktelik kurallarına bakıldığında bireyler metformin, chlorpropamide, glipizide, glyburide, tolbutamide, acarbose, miglitol ve tolazamide ilaçlarının hiçbirini kullanmıyorsa %93.4 insülin kullanmaktadır.



Şekil 4.5. CHAID algoritmasına göre karar ağaçları sonuçları

Şekil 5'te CHAID ve Kapsamlı CHAID algoritması ile elde edilen karar ağacı gösterilmektedir. Bu sonuca göre sekiz açıklayıcı değişken modelde yer almaktadır. Modelin doğru sınıflama oranı %70.9'tür

CHAID algoritmasına göre yaş değişkeninin Hba1c değeri üzerinde anlamlı bir etkiye sahiptir ($p<0.05$). İlk olarak araştırmaya katılan bireylerin %29.3'ü 8'den düşük, %70.7'si ise 8 ve 8'den büyük değere sahiptir. Oluşan ağacın dallarına baktığımızda, yaşı 70-80 arasında olan bireylerin %41.4'ününHba1c değeri 8'den düşük, %58.6'sının iseHba1c değeri 8 veya 8'den yüksektir. Bu sınıfın insülin kullananlarının %38.3'ününHba1c değeri 8'in altında ve %61.7'sinin ise 8 veya 8'in üstündedir. Ayrıca bu sınıfta bulunan ve insülin kullanmayan bireylerin %49.6'sınınHba1c değeri 8'in altında ve %50.4'ünün ise 8 veya 8'in üstündedir. Sonuç olarak insülin kullanımı 70-80 arasındaki bireylerde anlamlı etkiye sahiptir.

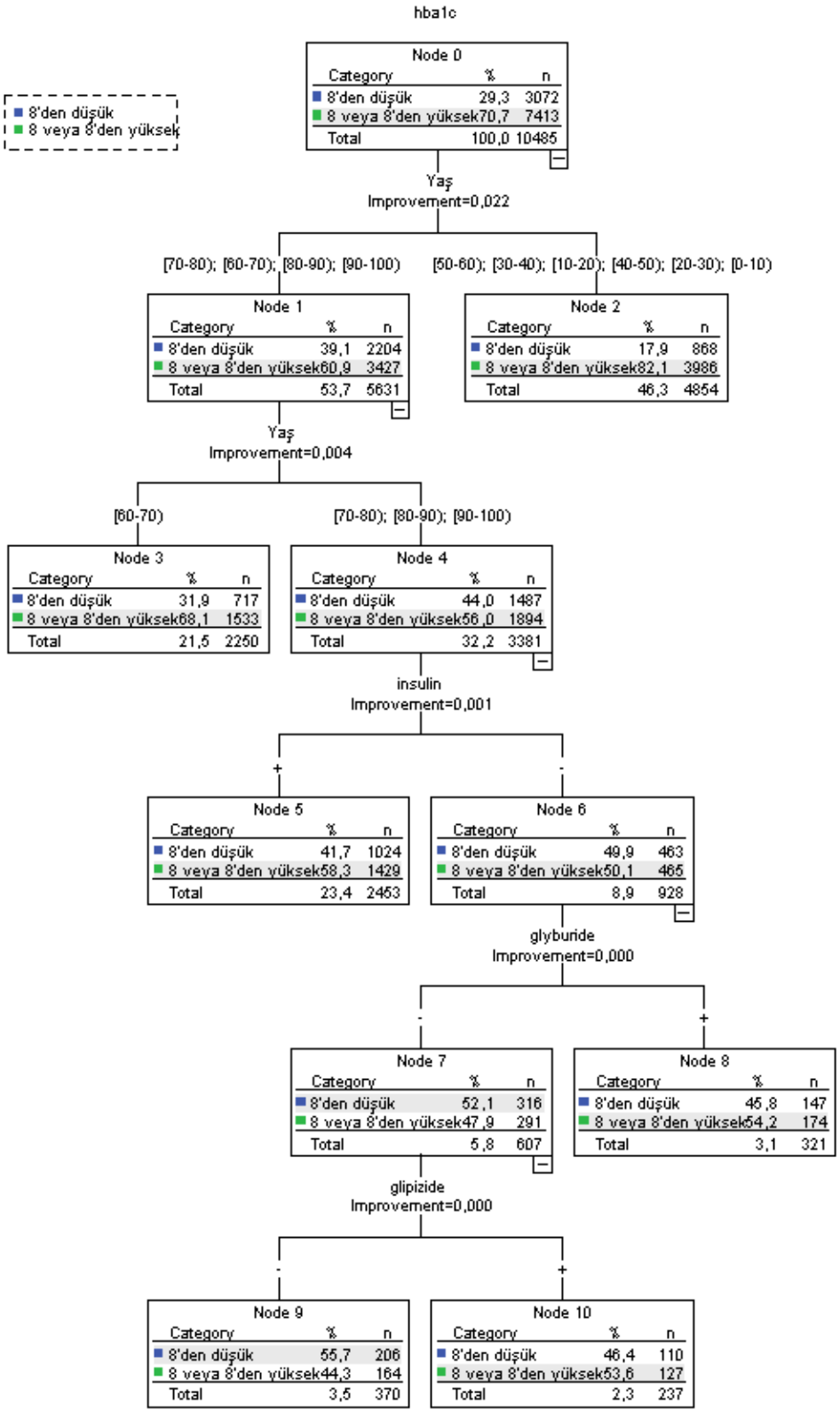
Bir diğer oluşan dala baktığımızda 50-60 yaş arasında bulunan bireylerin %22.9'unun Hba1c değeri 8'den düşük, %77.1'inin iseHba1c değeri 8 veya 8'den yüksektir. Bu sınıfın insülin kullananlarının %20'sininHba1c değeri 8'in altında ve %80'inin ise 8 veya 8'in üstündedir. Ayrıca bu sınıfta bulunan ve insülin kullanmayan bireylerin %32.8'nin Hba1c değeri 8'in altında ve %67.2'sinin ise 8 veya 8'in üstündedir. Sonuç olarak insülin kullanımı 50-60 yaş arasındaki bireylerde anlamlı etkiye sahiptir ($p<0.05$).

Yaşı 60-70 arasında olan bireylerin %31.9'uunHba1c değeri 8'den düşük, %69.1'inin ise 8 veya 8'den fazladır. İnsülin kullanımı bu yaş grubu üzerinde anlamlı bir etkiye sahiptir ($p<0.05$). Ayrıca insülin kullanan bireylerin glyburide kullanma durumları da yaşı 60-70 arasında bulunan bireyler için anlamlı bir etkiye sahiptir. Hem insülin hem glyburide kullanan bireylerinHba1c değerinin 8'den küçük olma oranı, sadece insülin kullananlara göre daha azdır.

Yaşı 30-40 arasında olan bireylerin %12.5'ininHba1c değeri 8'den düşük, %87.5'inin ise 8 veya 8'den fazladır. İnsülin kullanımı bu yaş grubu üzerinde anlamlı bir etkiye sahiptir($p<0.05$). Ayrıca insülin kullanan bireylerin glpizide kullanma durumları da yaşı 30-40 arasında bulunan bireyler için anlamlı bir etkiye sahiptir. Hem insülin hem glpizide kullanan bireylerinHba1c değerinin 8'den küçük olma oranı, sadece insülin kullananlara göre daha azdır.

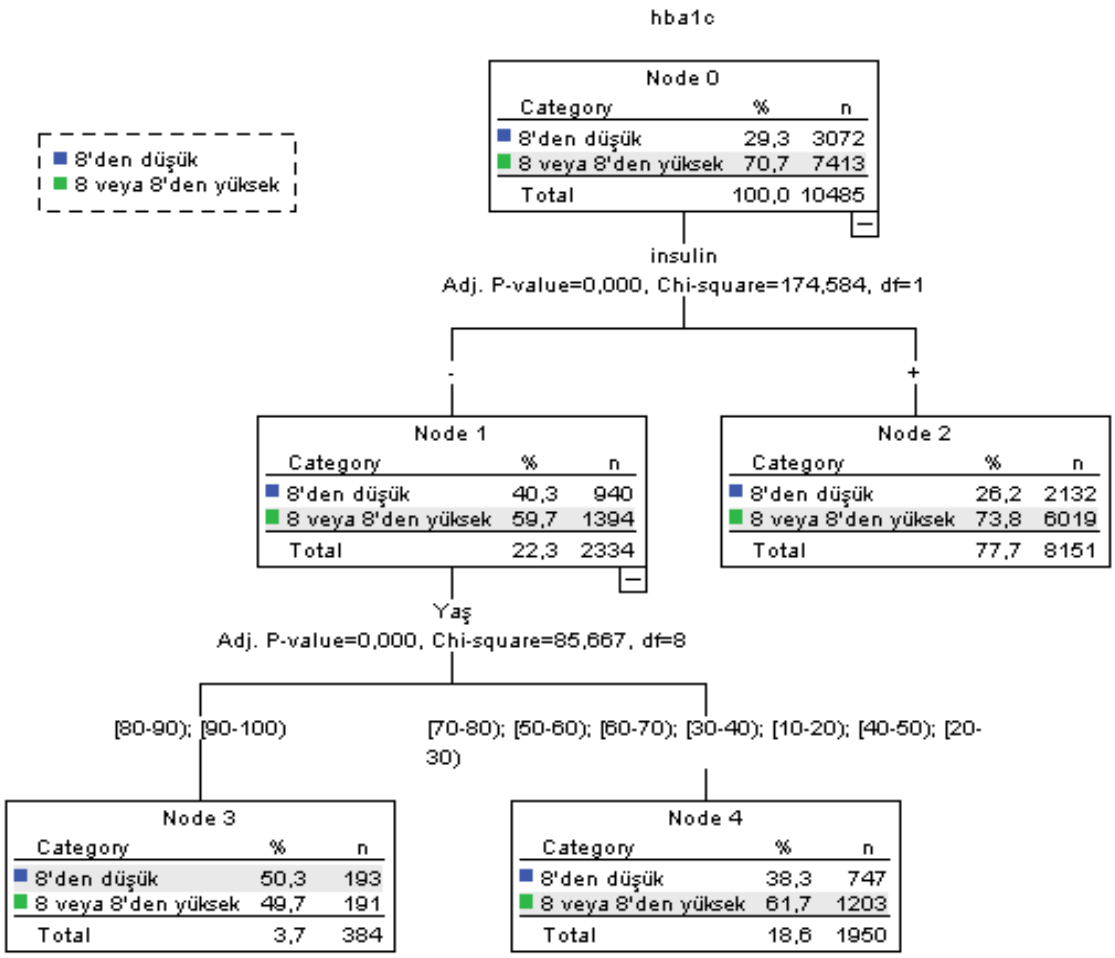
Yaşı 40-50 arasında olan bireylerin %18.8'ininHba1c değeri 8'den düşük, %81.2'sinin iseHba1c değeri 8 veya 8'den yüksektir. Bu sınıfın insülin kullananlarının %29.9'ununHba1c değeri 8'in altında ve %71.1'inin ise 8 veya 8'in üstündedir. Ayrıca bu sınıfta bulunan ve insülin kullanmayan bireylerin %16.4'ününHba1c değeri 8'in altında ve %80.6'sının ise 8 veya 8'in üstündedir. Sonuç olarak insülin kullanımı 40-50 arasındaki bireylerde anlamlı etkiye sahiptir ($p<0.05$).

Yaşı 80-100 arasında olan bireylerin %47.9'ununHba1c değeri 8'den düşük, %52.1'nin iseHba1c değeri 8 veya 8'den yüksektir. Bu sınıfın metformin kullananlarının %53.4'ünün Hba1c değeri 8'in altında ve %46.6'ının ise 8 veya 8'in üstündedir. Ayrıca bu sınıfta bulunan ve metformin kullanmayan bireylerin %46.6'sınınHba1c değeri 8'in altında ve %53.4'ünün ise 8 veya 8'in üstündedir. Sonuç olarak metformin kullanımı 80-100 arasındaki bireylerde anlamlı etkiye sahiptir ($p<0.05$).



Şekil 4.6. CART algoritması için karar ağacı

Şekil 6'da CART algoritması ile elde edilen karar ağacı gösterilmektedir. Bu sonuca göre beş açıklayıcı değişken modelde yer almaktadır. Modelin doğru sınıflama oranı ise %71.1'dir.CART algoritmasının sonuçlarına göre yaş değişkeni Hba1c üzerinde anlamlı bir etkiye sahiptir ($p<0.05$). Bu algoritma yaşı 0-60 ve 70-100 olarak iki dala ayırmıştır. Dallara bakıldığında yaşı 70-100 arasında olan ve insülin kullananların Hba1c değerinin 8'den düşük olma oranı, yaşı 70-100 arasında olan ve glyburide kullananların oranından daha düşüktür. Dallar incelendiğinde yaşı 70-100 arasında olan bireylerin sadece insülin, sadece glyburide, sadece glipzide kullandığı görülmektedir.



Şekil 4. 7. QUEST algoritması için karar ağacı

Şekil 7'de QUEST algoritması ile elde edilen karar ağacı gösterilmektedir. Bu sonuca göre yalnızca iki açıklayıcı değişken modelde yer almaktadır. Modelin doğru sınıflama oranı %97.4'tür. Bu algoritmaya göre insülin kullanımı Hba1c üzerinde anlamlı bir etkiye sahiptir ($p<0.05$). Sonuçlara göre insülin kullanan ve Hba1c değeri 8'in altında olan bireylerin oranı insülin kullanmayan bireylere göre daha düşüktür. İnsülin kullanmayan bireylerde ise yaş değişkeni anlamlıdır. İnsülin kullanmayıp yaşı 80-100 arasında olan veHba1c değeri 8'in altında olan bireylerin oranı, yaşı 200-70 arasında olan bireylere göre daha yüksektir.

6. SONUÇ ve DEĞERLENDİRME

Bu çalışmada 10485 DM hastasının kullandığı ilaçların bu hastalara ait Hba1c değerlerini nasıl etkilediği, hangi ilaçların diyabet tedavisinde etkili olduğu araştırılmıştır. Araştırma aşamasında her bir bireyin en az bir ilaç kullanma şartı ile oluşturulan veri setinin uygulama aşamasında veri madenciliği yöntemleri kullanılmıştır. 0-100 yaş arasında kadın ve erkek hastaların oluşturduğu veri seti içerisinde 17 farklı ilacın diyabet üzerindeki etkisi irdelenmiştir. Çalışmayı özgün kılan en büyük özelliklerinden biri yaş ve cinsiyet sınırlaması olmaması ve bu büyük veri setinden veri madenciliği teknikleri ile elde edilen bilgilerin R programlama dili ve SPSS paket programı kulanılarak literatüre katkıda bulunabilmesidir (Strack ve diğerleri, 2014)

Araştırmanın uygulama aşamasında Hba1c değerini modellemek için uygulanan üç farklı lojistik regresyon modeline göre her bağımsız değişken için odds oranları (OR), Wald istatistikleri ve anlamlılık değerleri verilmiştir. Son modelin lojistik regresyon sonuçlarına göre yaş grupları, glimepiride, glipizide ve glyburide ilaç kullanım durumlarınınHba1c değerinin modeli üzerinde istatistiksel olarak anlamlı etkiye sahiptir ($p<0.05$).

Araştırmada bağımlı değişken olarak alınan Hba1c değerini modellemek için uygulanan lojistik regresyon analizi için üç aşamalı bir yol izlenmiştir. Öncelikle hipotez testi aşamasında Hba1c ile istatistiksel olarak anlamlı ilişkiye sahip olan demografik faktörler üzerinden lojistik regresyon modeli (Full-model) tahmin edilmiştir. Daha sonra full-model içerisinde istatistiksel olarak anlamlı bulunmayan değişkenleri elemek için backward-lojistik regresyon analizi tekniği ile değişken seçimi uygulanmıştır. Backward aşamasında modele değişken eklerken $\alpha = 0.05$, modelden değişken çıkarırken de $\alpha = 0.10$ hata payı baz alınmıştır. Son olarak, backward aşamasında istatistiksel olarak anlamlı bulunan bağımsız değişkenler üzerinden lojistik regresyon modeli kurularak nihai çıkarımlar bu modele (Final-model) göre yapılmıştır (tablo 3).

Çalışmanın karar ağacı uygulamalarına göre yaşı 70-100 arasında olan bireylerde sadece metaformin ve sadece insülin kullanımının etkili olduğu ve metaformin kullanımının bu yaş grubu üzerinde daha olumlu etkisinin olduğu saptanmıştır. Yaşı 40-60 arasında olan bireylerde sadece insülin kullanımının

hastaların Hba1c değeri üzerinde anlamlı ve olumlu bir yönde etkisinin olduğu bilgisine rastlanmıştır. Sonuç olarak diyabet hastalığı tedavisinde uygulanan veri madenciliği tekniklerine göre en etkili 3 ilaç sırasıyla insülin, glyburide ve glipizide olduğu görülmektedir (tablo 3).

Diyabet günümüzün en büyük problemlerinden biri olduğu için bu konuda literatüre yerleşmiş birçok çalışma bulunmaktadır. Bu çalışmada R project ile veri madenciliği teknikleri kullanılarak diyabet tedavisinde kullanılan ilaçların değerlendirilmesi uygulanmıştır.

Bu çalışmada bireylerin sadece cinsiyetleri yaşı ve kullandıkları ilaçların bilgisi kullanılarak R project ve SPSS paket programı ile veri madenciliği modelleri kullanılmıştır. Veri madenciliği, veri setinin aşırı derin ve büyük olduğu durumlarda sahip olduğu modellerle en uygun algoritmaları kullanarak bir çözüm elde edilmesini sağlar. Diyabet hastalığının da toplum üzerindeki yer edinimine ve önemine bakılırsa bu konuda daha detaylı araştırmalar yapılabilir. Bireylerin günlük yaşantıları, egzersiz yapma durumları, beslenme şekli, diyabet hastalığı hakkındaki görüşleri, daha önce diyabet hastalığı hakkında eğitim alıp almadığı, hangi besini ne kadar sıklıkla tükettiği gibi özellikleri ile kan değerleri detaylı olarak araştırılabilir.

7. KAYNAKLAR

Adalı (2009). *Benchmarking Data Mining Techniques For Segmenting Diabetes Patients/* Bahçeşehir Üniversitesi / Fen Bilimleri Enstitüsü / Bilgisayar Mühendisliği Bölümü / Bilgisayar Mühendisliği Anabilim Dalı

Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), 127-136.

American Diabetes Association. (2014). Diagnosis and classification of diabetes mellitus. *Diabetes care*, *37*(Supplement 1), S81-S90.

Ateş, Y. ve Karabatak, M. (2017). Nicel Birliktelik Kuralları ile Çoklu Minimum Destek Değeri. Fırat Üniversitesi Mühendislik Bilimleri Dergisi, 29(2), Sf. 57-65.

Breault, J. L., Goodall, C. R., & Foes, P. J. (2002). Data mining a diabetic data warehouse. Artificial intelligence in medicine, 26(1-2), 37-54.

Çerkezi (2013). Veri madenciliği yöntemlerini kullanarak diyabetik retinopati hastalığının teşhisi/ Sakarya Üniversitesi / Fen Bilimleri Enstitüsü / Bilgisayar ve Bilişim Mühendisliği Anabilim Dalı

Çiçek (2014). Applying data mining techniques to implement the clinical guidelines for the management of the patients with type 2 diabetes: medication dose adjustments. Fatih Üniversitesi / Fen Bilimleri Enstitüsü / Bilgisayar Mühendisliği Anabilim Dalı

Devi, M. R., & Shyla, J. M. (2016). Analysis of various data mining techniques to predict diabetes mellitus. International journal of applied engineering research, 11(1), 727-730.

Diler, S (2016). Veri madenciliği süreçleri ve karar ağaçları algoritmaları ile bir uygulama. Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı Van.

Erdoğan (2019). Türkiye'de Lisans Düzeyindeki İstatistik Ders İçeriklerinin Veri Madenciliği Yöntemleri ile Analizi/ Çukurova Üniversitesi / Fen Bilimleri Enstitüsü / İstatistik Anabilim Dalı

Ergin, G. O., Dünder, E., Ökçün, S., & Koçkaya, G. (2020). Sosyoekonomik Durumun Diyabet ile İlişkisi ve Diyabete Etkisinin İncelenmesi. Türkiye Diyabet ve Obezite Dergisi, 4(2), 71-78.

Hearst, M. A. (1999, June). Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 3-10). Association for Computational Linguistics.

G. Piatetsky-Shapiro, W. J. Fawley (1991) . Knowledge Discovery in Databases. AAAI/MIT Pres

Han, J., Kamber, M., 2006. Data mining concepts and techniques. Morgan Kaufmann, 770s.

Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622

Hearst, M. A. (1999, June). Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 3-10). Association for Computational Linguistics.

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

Johnson, D. E. (1988). Applied Multivariate Methods for Data Analysts/ California: Duxbury Press.

Kayri, M., & Boysan, M. (2007). Araştırmalarda CHAID analizinin kullanımı ve baş etme stratejileri ile ilgili bir uygulama/ Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 40(2), 133-149.

Koyuncugil, A., & Özgülbaş, N. (2008). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. Bilişim Teknolojileri Dergisi, 2(2s).

Larose, D., T., 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, 217s.

Makhabel, B. (2015). Learning data mining with R. Packet Publishing Ltd.

Öğüt, S., 2009. Veri Madenciliği Kavramı ve Gelişim Süreci / Yeditepe Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, 12s.

Özyazar, Ö. (2019). Data mining applications for sustainable medical systems: A study on diabetes. Marmara Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı İstanbul.

Pasin, B. (2019). Türkiye halkı verilerine dayalı sigara kullanımını etkileyen faktörlerin belirlenmesinde çok değişkenli lojistik regresyon analiz teknikleri kullanılarak yapılan bir çalışma. Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı Ekonometri Bilim Dalı İzmir.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahman, R. M., & Afroz, F. (2013). Comparison of various classification techniques using different data mining tools for diabetes diagnosis. Journal of Software Engineering and Applications, 6(3), 85-97.

Savaş, S., Topaloğlu, N., Yılmaz, M., 2012. Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(21): 1-23.

Saygılı, A., Veri Madenciliği ile Mühendislik Fakültesi Öğrencilerinin Okul Başarılarının Analizi. Yıldız Teknik üniversitesi. Bilgisayar Mühendisliği Anabilim Dalı İstanbul

Seyrek, İ. H., & Ata, H. A. (2010). Veri Zarflama Analizi ve Veri Madenciliği ile Mevduat Bankalarında Etkinlik Ölçümü. Journal of BRSA Banking & Financial Markets, 4(2).

Silahtaroğlu, G., 2013. Veri Madenciliği Kavram ve Algoritmaları, Papatya Yayınevi

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

Strack, B., Deshazo, P., Gunning's, C., Lomo, L., Ventura, S., Cios, K., Clore, N., 2014 Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. Department of Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA.

Şahin, G. B., Gökhan, T., & Çetin, A. Veri Madenciliği Yöntemleri ile Diyabet Hastalığına Sebep Olan Faktörlerin Tespiti.

Şaşar (2017). Diyabet HastalarındakiHba1cParametresine Etki Eden Faktörlerin Veri Madenciliği Yöntemleri ile Tahmin Edilmesi Muğla Sıtkı Koçman Üniversitesi Fen Bilimleri Enstitüsü Bilişim Sistemleri Mühendisliği Anabilim Dalı

Yaşasınoğlu, N. (2019). Türkiye'de Anlaşmalı Boşanma ile Sosyal Güvenlik Kurumundan Aylık Alanlar Üzerine Veri Madenciliği Cart Algoritması Uygulaması/ Gazi Üniversitesi / Fen Bilimleri Enstitüsü / İstatistik Anabilim Dalı

Yüksel, T. (2018). Dağıtık sistemlerde birliktelik kuralları ile sepet analizi. İstanbul Aydın Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Bilgisayar Mühendisliği Bilim Dalı İstanbul

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and engineering, 26(1), 97-107.

ÖZ GEÇMİŞ

Merve DÜNDER, 16.09.1992 tarihinde DİYARBAKIR' da doğdu. Samsun Cumhuriyet Anadolu Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi Fen/Edebiyat Fakültesi'nden 2014 yılında mezun oldu.

İletişim Bilgileri

E-mail : dundermerve@gmail.com

Telefon : 17211061

ORCID ID: https://orcid.org/0000-0003-1738-4373