T.C. İSTANBUL ÜNİVERSİTESİ

SOSYAL BİLİMLER ENSTİTÜSÜ

İŞLETME ANABİLİM DALI

SAYISAL YÖNTEMLER BİLİM DALI

DOKTORA TEZİ

SAĞLIK HİZMETLERİNDE VERİ ANALİTİĞİ

NUR KUBAN TORUN

2502140267

TEZ DANIŞMANI

PROF. DR. UMMAN TUĞBA ŞİMŞEK GÜRSOY

İSTANBUL, 2018

T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

DOKTORA
TEZ ONAYI

ÖĞRENCİNİN;

Adı ve Soyadı	:	NUR KUBAN TORUN	**Numarası**	:	2502140267
Anabilim Dalı / Anasanat Dalı / Programı	:	SAYISAL YÖNTEMLER	**Danışmanı**	:	PROF.DR.UMMAN TUĞBA GÜRSOY
Tez Savunma Tarihi	:	11.10.2018	**Saati**	:	11:00
Tez Başlığı	:	SAĞLIK HİZMETLERİNDE VERİ ANALİTİĞİ.			

TEZ SAVUNMA SINAVI, İÜ Lisansüstü Eğitim-Öğretim Yönetmeliği'nin **50. Maddesi** uyarınca yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin **KABULÜNE** OYBİRLİĞİ / ~~OYÇOKLUĞUYLA~~ karar verilmiştir.

JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
PROF.DR.MEHPARE TİMOR		Kabul
PROF.DR.EMEL ŞIKLAR		Kabul
PROF.DR.AHMET METE ÇİLİNGİRTÜRK		Kabul
PROF.DR.UMMAN TUĞBA GÜRSOY	Tuğba	Kabul
DOÇ.DR.NİHAT TAŞ		Kabul

YEDEK JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
PROF.DR.ERGÜN EROĞLU	—	—
DOÇ.DR.ERDAL DİNÇER	—	—

ÖZ

SAĞLIK HİZMETLERİNDE VERİ ANALİTİĞİ

NUR KUBAN TORUN

Bu tezin amacı diyabet şüphesi ile kliniğe muayeneye gelen hastaların oluşturduğu veri seti üzerinden, veri madenciliği yöntemleri kullanarak, bir kişinin diyabetik polinöropati olup olmadığını öngörebilmektir. Çalışmada kullanılan veri seti Bilecik Devlet Hastanesi'nden temin edilmiştir. Veri setindeki değişkenler hastaların elektronik tıbbi kayıtları, şikayet, tanı ve anemnez (hastalık hikayeleri) incelenerek belirlenmiştir. Veri seti sınıflandırma algoritmaları ile analiz edilmiştir. Bunun için veri seti rastgele eğitim ve test veri seti olarak ikiye ayrılmıştır. Eğitim veri setleri ile modeller kurulmuş, ve test veri setleri ile test edilmiştir. K-en yakın komşu algoritması, Naive Bayes sınıflayıcı, Lojistik regresyon , C4.5 karar ağacı algoritması ve birliktelik kuralları uygulanmış ve farklı modeller oluşturulmuştur. Modellerin performansları kıyaslanmıştır. Veri analizleri R programla dili ile RStudio'da yapılmıştır.

Anahtar Kelimeler: Veri Madenciliği, R programlama, Sağlık Hizmetleri, Diyabet

ABSTRACT

DATA ANALYTICS IN HEALTCARE

NUR KUBAN TORUN

The aim of this research is to predict the diabetic polyneuropathy, using data mining methods on the data set of patients who visit clinics due to diabetes related complaint. The data set used in the research was obtained from Bilecik State Hospital. Variables in the data set were determined by examining the patients' electronic medical records, complaints, diagnosis, and anemnia (disease history). Data set was analyzed by classification algorithms. In this manner, the data set was divided into two part: random training and test varible set. Model was built by the training data sets and tested by the test data sets. Different models was created and applied by k-Nearest Neighbors algorithm, Naive Bayes classifier, Logistic regression, C4.5 decision tree algorithm and Association rules. R programming in RStudio was used for data analyzes.

Keywords: Data mining, R programming, Healthcare Service, Diabetes Mellitus

ÖNSÖZ

Diyabet günümüzde yaygın olarak görülen ve geleceği tehdit eden hastalıklardan biridir. Önemi sebebiyle bu tez çalışmasında konu olarak seçilmiştir. Bir devlet hastanesinden alınan veriler düzenlenerek, R pogramlama dili ile analiz edilmiştir.

Modelleme aşamasında K-en yakın komşu algoritması, Naive-Bayes sınıflandırma, Lojistik regresyon analizi, Karar ağaçları ve Birliktelik kuralları teknikleri kullanılmıştır. Elde edilen sonuçlar karşılaştırılmış ve yorumlanmıştır.

Bu tez çalışması sürecinde desteklerini esirgemeyen eşim Tolga Torun'a, eğitim hayatıma katkıları için aileme ve pozitif ve yapıcı tavırlarıyla desteklerini benden hiç esirgemeyen vizyoner ve zarif danışman hocam Prof.Dr. Umman Tuğba Gürsoy'a teşekkür ederim.

Tatlı oğlum Ali Çelebim bana anlayış gösterdiğin için, benim stresimi sende benimle taşıdığın için sana da teşekkür ederim.

NUR KUBAN TORUN

İstanbul, 2018

İÇİNDEKİLER

İKİNCİ BÖLÜM

SAĞLIK HİZMETLERİNDE VERİ MADENCİLİĞİ VE UYGULAMALARI

ÜÇÜNCÜ BÖLÜM

DİYABET HASTALIĞI

DÖRDÜNCÜ BÖLÜM

UYGULAMA: BİLECİK DEVLET HASTANESİ DİYABETİK POLİNÖROPATİ HASTALARINA İLİŞKİN VERİ ANALİTİĞİ

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

KISALTMALAR LİSTESİ

ADA	: American Diabetes Association
ANN	: Artificial Neural Network (Yapay Sinir Ağları)
CART	: Classification and Regression Trees
CHAID	: Chi-Square Automatic Interaction Detector
EWS	: Early Warning Score (Erken Uyarı Skorları)
GP	: Gaussian Process (Gaussian Süreci)
HbA1c	: Glycated Haemoglobin
HDL	: High-Density Lipoprotein
ID3	: Iterative Dichotomiser 3
Kh	: Karbonhidrat
kNN	: K-nearest neighbour (K-en yakın komşu)
LDA	: Liner Diskriminlant Analiz
LDL	: Low-Density Lipoprotein
Max.	: Maximum
Min.	: Minimum
MLP	: Multi-Layer Perceptrons
MOA	: Massive On-line Analysis
NB	: Naive Bayes
QDA	: Quadratik Diskriminlant Analiz
SAS	: Statistical Analysis System
SAP	: Systems Analysis and Program
SQL	: Structured Query Language

SVM : Support Vector Machine

VKİ : Vücut kitle indeksi

WHO : World Health Organization

GİRİŞ

Bilgi çağı veri miktarında artışa yol açmıştır. Özellikle elektronik kayıt sistemlerinin artması ve bireylerin elektronik ortamlarda sıklıkla yer almasıyla birlikte yığın veri ön plana çıkmaktadır. Ancak bu yığın içerisinden faydalı ve doğru veriye ulaşılması ve bu verilerin işlenerek bilgi haline getirilmesi başka bir sorunu beraberinde getirmiştir. Bu soruna çözüm olarak uygulayıcıların ve araştırmacıların karşısına veri madenciliği kavramı çıkmıştır. Veri madenciliği ile birlikte yığın halde bulunan veriler içerisinden belirli teknikler yardımıyla anlamlı bilgiler elde etme yoluna gidilmektedir. Veri madenciliği özellikle hizmet pazarlaması alanında müşteri profilleri oluşturmada ve hedef kitlenin isteklerinin belirlenmesi gibi konularda popülerliğini arttırmıştır. Ancak veri madenciliğinin önemi sadece pazarlama alanında değil; sağlık hizmetleri gibi toplumu ilgilendiren ve anında müdahalenin önemli olduğu stratejik alanlarda da yerini almaya başlamıştır. Elektronik kayıtların artması, hastaların elektronik olarak takibe alınabilmesi, giyilebilir teknolojinin gelişmesine paralel olarak sağlık alanında yığın veriler oluşmaya başlamış ve sağlık hizmetlerinde veri analitiğinin gerekliliği ortaya çıkmıştır. Sağlık hizmetlerinde veri madenciliği ile birlikte tanılar geliştirilmekte, toplumda hastalıkların yaygınlıkları tahmin edilebilmekte ve gerekli durumlarda tahminlemeler yardımıyla uzun vadeli sağlık planları yapılabilmektedir. Sağlık hizmetlerini diğer veri madenciliği tekniklerinin kullanıldığı alanlardan farklı kılan unsur, elde edilen verilerin kişisel olması ve kesin olmamasıdır. Sağlık hizmetlerinde veri madenciliği yöntemleri ile birlikte gizli örüntüler ortaya çıkarılmakta ve bu kesin olmayan veri üzerinden sınıflandırma yapılarak tahmin edilir sonuçlar kuralmaktadır. Sağlık hizmetlerinde veri madenciliği uygulamaları ile oldukça değerli olan bilgiye ulaşılabilmektedir.

Sağlık alanında gün geçtikçe kronik hastalıklarda artış meydana gelmektedir. Dünya genelinde bu kronik hastalıklarla mücadele konusunda ciddi araştırmalar yapılmakta ve mümkün olan her bilgiye başvurulmaktadır. Günümüzün belki de en tehdit içeren kronik hastalığı yaygınlığı bakımından diyabet ya da diğer ismi ile şeker hastalığıdır.

Dünya üzerinde her 100 kişiden 9'u diyabet olmasına rağmen, diyabet tanısı koyulan ve bu hastalığı taşıdığını bilen kişi sayısı azdır. Kalıtsallığın dışında, diyabetin bu denli yaygın olmasının nedenleri arasında sağlıksız beslenme, obezite ve hareket eksikliği sayılabilir. Diyabetin yaygınlığının ve diyabete bağlı hastalıklar arasındaki örüntülerin belirlenmesi açısından veri madenciliği önemli bir yer oluşturmaktadır. Bu amaçla diyabet verilerinden yola çıkılarak tahminler ve öngörüler yapılması, gelecek eylem planları oluşturma, müdahale yöntemleri geliştirme ve bilinçlendirici programları hayata geçirme, diyabet ile mücadele açısından anahtar faktör olarak sağlık kuruluşlarının ve uygulayıcıların karşısına çıkmaktadır.

BİRİNCİ BÖLÜM

VERİ MADENCİLİĞİ

1.1. Veri Madenciliği Nedir?

Veri madenciliği, verinin ayıklaması ya da yığın bir veri içerisinden madencilik yapılarak anlamlı ve işe yarar bilginin çekilmesi aşamalarına verilen isimdir. Veri madenciliği aslında eski zamanlarda toprak içerisinden altın arayan kişilere yapılan bir göndermedir. İnsanlar ellerinde bulunan elekler ile büyük parça toprağı alarak, içerisinden altın aramaktadırlar. Veri madenciliği de aslında tam olarak budur. Anlamlı bir bilgi elde edebilmek amacıyla bir çok veri taranmakta ve analiz edilmektedir. Veri madenciliğinin bu açıdan bakıldığında ismi bilgi arayışında veri madenciliği ya da bilgi madenciliği olarak da kullanılabilmektedir. Ancak bir çok araştırmacı tarafından kullanılan ismi veri madenciliğidir (Jiawei ve Kamber, 2006).

Cabena (vd., 1998) veri madenciliğini, büyük veri kaynaklarından anlamlı bilgi elde etmek için algoritmaları, istatistiği ve görselliği kullanan disiplinlerarası alan olarak tanımlamıştır. Hand (vd., 2001) ise veri madenciliğini gözlenebilir veri setlerinin, anlaşılır ve yararlı bilgiye ihtiyaç duyanlar için analiz edilmesi ve elde edilen bilgilerin raporlanması olarak tanımlamaktadır. Wang ve Weigeng'e (2004) göre veri madenciliği yığın bir veri seti içerisinden gizli ve karmaşık ilişkilerin modern istatistik, akıllı bilgi sistemleri, makine öğrenmesi, örüntü tanıma, karar teorileri, veri mühendisliği ve veri bankası yönetimini birleştirerek çıkarılması olarak tanımlanmaktadır. Ayrıca veri madenciliği, otomatik veya yarı-otomatik biçimlerde verinin analiz edilerek gizli örüntülerin bulunması olarak tanımlamaktadır (Tang ve MacLennan, 2005; Witten ve Frank 2005). Keşfedilen örüntüler anlamlı olmalı ve ekonomik olarak araştırmacıya bir yararı dokunmalıdır. Tan, (vd., 2006), veri madenciliğini, büyük veri kaynaklarında yer alan yararlı bilgilerin otomatik olarak keşfedilmesi süreci olarak tanımlamaktadır. Gartner Group (2007), veri madenciliğini, veri ambarlarında depolanan büyük miktarlardaki verinin istatistiksel ve matematiksel tekniklerle birlikte örüntü tanıma teknolojilerinin de kullanılarak

incelenmesi yoluyla anlamlı, yeni ilişkiler, örüntüler ve eğilimler bulunması süreci olarak tanımlamaktadır.

Veri madenciliği günümüzde araştırmacılarca oldukça fazla kullanılan bir teknik olarak karşımıza çıkmaktadır. Özellikle toplumsal konularda ve endüstriye yönelik konularda veri madenciliğine başvuranların sayısı gitgide artmaktadır. Veri madenciliğine bu denli ilginin artmasının altında yatan ana sebep ise teknolojik gelişmelerdir. Teknolojik gelişmeler ile birlikte verilerin toplanması, depolanması ve çağrılması hususlarında kolaylık yaşanmaya başlanmıştır. Şekil 1'de görüldüğü üzere veri madenciliği belli aşamalardan geçerek günümüze gelmiştir (Jiawei ve Kamber, 2006).

Şekil 1: Geçmişten günümüze veri madenciliğinin gelişimi

Kaynak: Jiawei, H., & Kamber, M. (2006). Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann. https://doi.org/10.1007/978-3-642-19721-5

Veri madenciliğinin kullanılma nedenleri arasında birçok sebep sayılabilir. Veri madenciliğinin sağladığı faydalar aşağıdaki gibi sıralanabilir (Larose, 2005:26):

- Tanımlama
- Tahminleme
- Öngörüde bulunma
- Sınıflama
- Kümeleme
- Birliktelik

1.1.1. Tanımlama

Bazı durumlarda araştırmacılar ve analistler verilerin altında yatan örüntü ve eğilimleri tanımlamanın yollarını aramaktadır. Örüntü ve eğilimlerin iyi bir şekilde tanımlanması, bu örüntülerin ve eğilimlerin pozitif açıklamalarına ulaşmada önemli bir etkendir. Veri madenciliği modelleri mümkün olduğu kadar şeffaf olmalıdır. Modeller temiz örüntüleri tanımlamalı ve kurumsal yorumlama ve açıklamalar için düzeltilebilir olmalıdır. Bazı veri madenciliği modelleri diğerlerine göre açıklamada daha şeffaf olabilmektedir. Örneğin karar ağaçları daha sezgisel ve kullanıcı dostu sonuçlar sağlarken; sinir ağları uzman olmayanlar için daha zor yorumlanabilmektedir.

1.1.2. Tahminleme

Tahminleme de tanımlama gibi kategorik verilerdense sayısal verilere dayanmaktadır. Tahminleme içerisindeki modeller, tahminde bulunacak kişiye hedef değişken ile ilgili değer veren karmaşık kayıtlara dayanmaktadır. Örneğin bir araştırmacının hastanedeki hastaların yaş, cinsiyet, vücut kütle endeksi ve kandaki sodyum seviyesine bağlı olarak hastaların sistolik kan basıncını tahmin etmeye çalıştığını düşünelim; sistolik kan basıncı ile yordayıcı değişkenler arasındaki ilişki

araştırmacıya tahmin modelini verecektir. İstatistik alanında bir çok tahmin edici model bulunmaktadır. Bunlar, nokta tahmin, güven aralık tahminleri, basit doğrusal regresyon ve korelasyon ile çoklu regresyon gibi modellerdir.

1.1.3. Öngörüde Bulunma

Öngörüde bulunma, sınıflandırma ve tahminlemeye oldukça benzemektedir. Ancak öngörüde bulunmayı ayıran faktör, sonuçların gelecek için yapılmasıdır. Örneğin hisse senetlerinin 3 ay sonraki fiyatlarını belirlemek, gelecek sene yaşanabilecek ölümlü trafik kazası sayısını ortaya koymak gibi unsurlar hep gelecek için yapılabilecek öngörülerdir. Belirli şartlar altında, sınıflama ve tahminlemede kullanılan hemen hemen tüm modeller, öngörüde de kullanılmaktadır. Bunlar, nokta tahmin, güven aralık tahminleri, basit doğrusal regresyon ve korelasyon ile çoklu regresyon gibi modellerin yanı sıra, karar ağaçları, yapay sinir ağları ve en yakın komşu metotlarıdır.

1.1.4. Sınıflandırma

Sınıflandırmada genç, orta yaş ve yaşlı insanlar gibi belirli kategorik bir hedef değişken bulunmaktadır. Veri madenciliği modelleri büyük sayıda kayda geçmiş veri setlerini analiz etmektedir. Bu kayıtlı veri setlerinin içerisinde tahmin ediciye veya girdiye yönelik bir çok bilgi bulunmaktadır. Araştırmacı bu durumda, sınıflandırmaya tâbi tuttuğu gelir gruplarını veri setlerinden aldığı bilgi doğrultusunda, meslek, yaş veya cinsiyet ile ilişkilendirerek sınıflandırabilir. Sınıflandırma genel olarak kredi kartı dolandırıcılığı, öğrenciler için özel gereksinimler planlama, kredi başvurularında kişilere kredi verme, hastalıkların teşhisi, vasiyetlerin gerçek kişilerce yazılıp yazılmadığı ya da bir kişinin finansal hareketliliğine dayanarak terörist eylem planlayıp planlamadığı gibi birçok çeşitli alanda kullanılabilmektedir. Sınıflama için en çok kullanılan veri madenciliği metotları yapay sinir ağları, k-en yakın komşu ve karar ağaçları teknikleridir.

1.1.5. Kümeleme

Kümeleme, kayıtların benzerliklerine göre bir arada tutulması farklılık gösterenlerin birbirinden ayrıştırılmasıdır. Kümelemenin sınıflandırmadan farkı, kümelemede

hedef bir değişken bulunmamasıdır. Kümeleme, hedef bir değişkene bağlı olarak sınıflama, tahmin ya da öngörüde bulunma eylemlerini gerçekleştirmemektedir. Kümelemede amaç göreceli olarak homojen alt gruplar veya kümeleri benzerlikleri maksimize ederek veriyi bölümlendirmektir. Kümeleme genelde veri madenciliğinin birinci adımını oluşturmaktadır.

1.1.6. Birliktelik Kuralları

Birliktelik analizinde yapılmak istenen, ele alınan nitelikle birlikte işleyen başka bir niteliği ortaya çıkarmaktır. İşletme dünyasında bu analiz şekli genelde yakın ilgi analizi ya da sepet analizi olarak bilinmektedir. Birliktelik analizinde amaç, iki ya da daha fazla nitelik arasında sayısal bir ilişkiye dayanan gizli kurallar bulmaktır. Birliktelik kuralları genel olarak "if...., then..." algoritması üzerine kuruludur. Örneğin salı akşamı yapılan alışverişlerde, 1000 adet müşterinin toplamda 200 adet bebek bezi aldığı ve bu 200 bebek bezini alan müşterinin de toplamda 50 adet bira da aldığı görülürse, birliktelik kuralına göre "bebek bezi alanlar bira da alır" kuralı çıkacaktır. Bu kuralın desteği 200/1000=%20 ve güvenilirliği 50/200=%25 olarak bulunacaktır.

1.2. Verilerin Madenciliği Adımları

Veri madenciliği uygulaması bir süreçtir. Amaç veri içindeki gizli örüntüyü bulmaktır. Veri madenciliği için gerekli olan veriler işletmelerin SQL kaynaklarından, bilgisayar ortamında yazılmış verilerden, veri bankalarında yer alan verilerden alınabilmektedir. Veri madenciliği süreci problemi tanıma, veriyi toplama, veriyi anlama, veriyi hazırlama, yönteme karar verme ve uygulama aşamalarından oluşmaktadır.

1.2.1. Verinin Hazırlanması

Veri bankalarından elde edilen verilerin, analizlerden önce veri madenciliğine hazırlanması gerekmektedir. Elde edilen veriler, eksik, tamamlanmamış ya da kirli olabilirler. Örneğin veriler eski olabilir, kayıp değerlere sahip olabilir, uç değerler olabilir, veriler kullanılacak veri madenciliği tekniğine uygun olmayabilir ya da

değerlerin her hangi bir dayanağı olmayabilir (Larose, 2005:27-28). Bunun gibi durumlar için aşağıdaki adımlar uygulanmaktadır (Bramer, 2007; Larose, 2005):

- Verinin temizlenmesi
- Kayıp değerler
- Nitelik sayısını azaltma

1.2.1.1. Verilerin Temizlenmesi

Belirli bir kaynaktan elde edilen verilerin hatalardan uzak olduğu düşünülemez. Gerçek dünyanın veri setleri öznel yargılardan kaynaklı ve veri kaydetme aracının arızalarından dolayı hatalarla dolu olabilir. Örneğin 69.72 sayısı kazayla 6.972 olarak ya da kahverengi gözü gösteren etiket mavi göz olarak işaretlenmiş olabilir. Bazı durumlarda, büyük veri setleri içerisinde, en bariz hataları görmek mümkün olmayabilir. Bu durumda araştırmacının yardımına yazılım programları yetişmektedir.

1.2.1.2. Kayıp Değerler

Veri setleri içerisinde yer alan verilerin tamamının tam olarak doldurulmuş olmasını beklemek büyük bir hatadır. Veri setleri içerisinde yer alan eksik verilerin tespiti, araştırmanın anlamlı olması açısından elzem bir durumdur. Örneğin cinsiyete yönelik yapılan tıbbi bir araştırmada, cinsiyetin girilmediği örneklemler araştırma için anlam taşımayacaktır. Kayıp değer karşısında yapılabilecek şeylerin başında kayıp değere sahip örneklemi, araştırma kapsamı dışına almaktadır. Ancak araştırma dışına alınan örneklem, araştırmanın güvenilirliğini etkileyebilir. Ayrıca örneklem çıkarma işleminin kayıp değerlerin oranının örnekleme göre küçük olduğu araştırmalarda yapılması önerilmektedir. Eğer kayıp değer oranı yüksek ise örneklemin çıkarılması tavsiye edilmez. Bunun yerine kayıp değerin veri seti içerisinden bir tahminleme yapılarak frekansa göre ya da ortalamaya dayanarak bir değer atanması gerekmektedir. Kategorik bir nitelik için kayıp değer ataması en sağlıklı şekilde frekansa göre olmaktadır. Ancak sayısal veriler için ortalama belirlemek daha doğru sonuç verecektir.

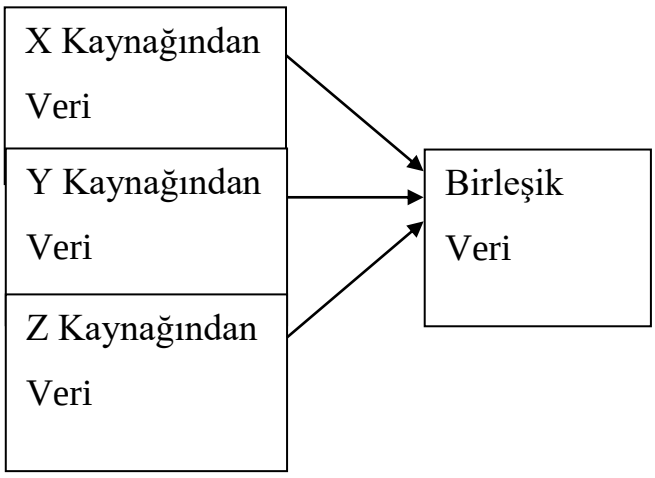
1.2.1.3. Nitelik Sayısını Azaltma

Bazı durumlarda belirli bir niteliği ölçmeye yarayan bir çok nitelik araştırma kapsamına alınmış olabilir. Bu durumlarda araştırma amacından sapabilmektedir. Yazılan uzun algoritmalar hem uzun sürede sonuç verecektir hem de hatalı sonuçlar verebilecektir. Araştırmacı, araştırmada gerçekten neyi bulmayı hedeflediğini belirleyerek, araştırma kapsamına alınan bileşenlerin sayısını düşürebilir, bu sayede araştırmanın boyutunu kısaltmış olur.

1.2.2. Verinin Birleştirilmesi ve Verinin Dönüştürülmesi

Veri bütünlüğü aşaması, ayrı kaynaklardan elde edilmiş verilerin anlamlı ve değerli bilgiye dönüştürülmesi amacıyla birleştirilmesi süreci olarak ifade edilmektedir. Bu aşama verilerin analize hazırlanmasının birincil aşamasını oluşturmaktadır. Verilerin birleştirilmesi aşamasında amaç, gereğinden fazla veri ile uğraşılmasının önüne geçmektedir. Aşağıdaki şekilde X, Y ve Z verileri için birleştirme süreci gösterilmektedir.

Şekil 2: Verilerin birleştirilmesi



Diğer yandan verilerin analize uygun veri setleri haline getirilecek şekilde dönüştürülmesi gerekmektedir. Bazı durumlarda, elde edilen verilerin değerleri ilgili analize uygun olmayabilir. Bu aşamada verilerin uygun hale getirilerek dönüştürülmesi, veri madenciliğinin başarısını arttıracak bir unsur olacaktır.

Verilerin dönüştürülmesinde normalizasyon ve stardardizasyon yöntemleri kullanılmaktadır (Tan, Steinbach ve Kumar, 2006).

1.2.2.1. Normalizasyon

Veri hazırlama sürecinin son adımıdır. Normalizasyon, veri setindeki tüm değerlerin, belirli bir standarda getirilmesi için, aynı değer aralığına karşılık gelecek şekilde dönüştürülmesi işlemidir. Bu işlemin yapılma amacı sayısal bir veri setinde, büyük değer içeren verinin küçük değer üzerinde baskı kurmasını engellemektir. Normalizasyon işlemi ile aynı ölçek düzeyine gelen değişkenler birbiri ile karşılaştırma yapılabilir hale gelecektir. Bu işlem için en çok bilinen yöntemler, minimum-maksimum(min-maks) ve z-skoru yöntemidir (Larose, 2005).

a. Minimum-Maksimum Normalizasyonu

Bu normalizasyon veriyi 0 ile 1 arasındaki değerlere dönüştürür. Bu yöntem de veri içerisindeki en büyük ve en küçük değer belirlenir. Diğer değişkenler buna uygun hale dönüştürülür. En büyük değer 1 değerini, en küçük değer 0 değerini alır. Formülü şu şekilde uygulanır:

$$x_{normalize} = \frac{x - x_{min}}{x_{maks} - x_{min}}$$

Tablo 1: Min-Maks normalizasyon işlemi

Eski değer	Normalleştirme işlemi	Normalize Değer
25	=(25-5)/(25-5)	1
15	=(15-5)/(25-5)	0,6
5	=(5-5)/(25-5)	0
10	=(10-5)/(25-5)	0,4

b. Z-Skoru Standardizasyonu

Z-Skor standardizasyon yönteminde veri ortalaması ve standart sapma değerleri kullanılır. Dönüşüm şu şekilde gerçekleşir:

$$x_{normalize} = \frac{X - \bar{X}}{\sigma_X}$$

X= gözlemin gerçek değeri

$\bar{X}$ =verinin aritmetik ortalamasını

σ_X = Standart sapmayı ifade etmektedir.

Tablo 2: Z-Skor standardizasyon işlemi

Eski değer	Standardizasyon işlemi	Standartize Değer
25	=(25-13,75)/7,395	1,521
15	=(15-13,75)/7,395	0,169
5	=(5-13,75)/7,395	-1,18
10	=(10-13,75)/7,395	-0,507

(Tablodaki değerlere göre aritmetik ortalama $\bar{X}$=13,75 , standart sapma σ_X=7,395 dir.)

1.2.3. Veri Madenciliği Aşaması

Veri madenciliği aşaması, belirli metotlar kullanarak veriye ait örüntülerin ayıklanması aşamasını teşkil etmektedir. Veri madenciliği aşaması karakterleştirme, birliktelik ve korelasyon analizi, sınıflama, kümeleme, öngörü ve uçdeğer analizi gibi bir dizi fonksiyondan oluşmaktadır. Veri madenciliği, büyük sayıda veri setlerinin, algoritmalarca analiz edilmesi aşaması olarak tanımlanmaktadır. Veri madenciliği bünyesinde istatistik, makine öğrenmesi, algoritma, görselleştirme ve veritabanı sistemleri gibi bir çok disiplini barındırmaktadır. Veri madenciliği süreci problemin tanımlanması, verinin toplanması ve hazırlanması, modelin kurulması ve örüntülerin değerlendirilmesi ile bilginin sunulması aşamalarını içermektedir.

1.2.3.1. Problemin Tanımlanması

Problemin tanımlanma aşamasında cevaplanması gereken sorular vardır. Bunlar; belirlenen hedef nedir, tahmin edilmek istenen nedir, tahmin için oluşturulan model ne fayda sağlayacaktır sorularıdır. Bu sorular mutlaka cevaplanarak problem tanımlanmalıdır. Bu aşamada soruların cevaplanması ilerleyen aşamalarda ortaya çıkabilecek sorunları engellemektedir. Eğer problem tanıma süreci iyi yönetilemezse ilerleyen adımlarda sorunlar ortaya çıkabilir.

1.2.3.2. Veri Toplama ve Hazırlama

Problem tanımlandıktan sonraki aşama veriye ulaşma, örnekleme ve dönüştürme aşamalarıdır. Değişik kaynaklardan elde edilen veriler bir tek kaynak altında toplanmaktadır. Analiz edilecek olan veriler birleştirilmiş verilerdir. Veri madenciliğinde toplanacak verinin niteliği kullanılacak olan veri madenciliği metodu düşünülerek iyi bir şekilde tasarlanmalıdır. Araştırmacının kullanabileceği veri türleri şunlardır (Bramer, 2016):

a. Nominal (Kategorik) Veri: Nesneleri kategorilere koymak için kullanılan değişkendir. Matematiksel bir yorumu yoktur. Örneğin, nesnelerin rengi kırmızı, mavi, pembe olabilir.

b. İkili Veri: Nominal değişkenin özel bir halidir. Yalnızca iki olası değer alır: doğru-yanlış, kadın-erkek gibi.

c. Sıralı (Ordinal) Veri: Sıralı veriler nominal verilere benzerler, yalnız ordinal verilerin arasında sıralı bir ilişki vardır. Örneğin küçük, orta ve büyük gibi.

d. Tamsayılı Veri: Sayısal değer alıp dört işlem yapılabilen verilerdir. Örneğin, çocuk sayısı.

e. Aralık- Ölçeği Verileri: Sıfır noktası gerçek yokluk durumunu ifade etmez. Merkezden ya da bir başlangıç noktasından eşit uzaklıkla ölçülmüş nümerik değerleri ifade eder. Fahrenhayt ve Selsiyus sıcaklık dereceleri bu veri tipine girmektedir. Selsiyus sıcaklık derecesiyle ölçülen 10 değeri ve 20 değeri

10 >20 den büyüktür ya da 10 <20 den küçüktür şeklinde yorumlanabilir. Ama biri diğerinin iki katıdır şeklinde yorumlanamaz. Çünkü aynı sıcaklık dereceleri Fahrenayt sıcaklık birimine çevrildiklerinde, aralarındaki kat ilişkisi farklı olacaktır.

f.Oran-Ölçeği Verileri: Oran ölçeği verileri aralık ölçeği verilerine benzemektedir. Buradaki fark geçek sıfır noktasının olmasıdır. Sıfır bir değer ifade eder, gerçek yokluğu gösterir. Matematiksel işlemlerde kullandığımız değerdir. Gerçek sıfır başlangıç noktasına sahip veriler oranlanır ve birbirlerine göre üstünlükleri belirlenir. Ağırlık, boy örnek verilebilir.

1.2.3.3. Model Kurma ve Örüntü Değerlendirilmesi

Elde edilen verilerden yararlanabilmek amacıyla model kurulması gerekmektedir. Bu aşamada kurulan modelin, veriler için uygun olup olmadığın analiz edilmesi ve örüntülerden elde edilen bilginin tanımlanması gerekmektedir. Bulguların doğru ve anlaşılır yorumlanması, araştırmanın ilgi çekiciliğini arttırmaktadır.

1.2.3.4. Bilginin Sunulması

Model içerisinde bir çok girdi, belirli bir amaç doğrultusunda bilgi elde etmek amacıyla kullanılmaktadır. Analiz sonucunda yığınlar içerisinden elde edilen bilginin raporlanması gerekmektedir.

1.3. Veri Madenciliğinin Kullanıldığı Alanlar ve Veri Madenciliği Araçları

Veri madenciliği günümüzde bir çok alanda kullanım açısından önemli bir konuma sahiptir. Veri madenciliğinin kullanım alanları pazarlama, finans, tedarik zinciri/lojistik ve hizmet sektörü başlıkları altında toplanabilmektedir. Veri madenciliğinin yoğun olarak kullanıldığı alanlar aşağıda verilmiştir (Marketsandmarkets.com)

- Tedarik Zinciri Yönetimi
- Bankacılık Sektörü, Finansal Hizmetler, Sigorta Şirketleri
- Tıbbi Hizmetler ve Hayat Bilimi

- Telekom ve IT Hizmetleri
- Hükümet ve Savunma Sanayi
- Enerji ve Kamu Hizmetleri
- Üretici Firmalar
- Diğer (Eğitim, Medya, Eğlence Sektörü)

Veri madenciliği pazarı büyümeye devam eden bir eğilim göstermektedir. 2018 verilerine göre bu pazarın değeri 591.2 milyon dolar civarındadır. 2023 yılında bu miktarın 1 milyar dolar civarına ulaşacağı tahmin edilmektedir. Veri madenciliği araçlarının pazarındaki bu artış şüphesiz günümüz işletmelerinin veri elde etme ve işlemenin önemini kavramasıyla doğru orantılıdır. Özellikle bankalar, finansal hizmet sektöründe yer alan firmalar ve sigorta şirketleri yoğun olarak veri madenciliği kullanmaya başlamıştır.

Veri madenciliği araçlarını üreten işletmeler IBM, SAS Institute, Oracle, Microsoft, MathWorks, Business Insight, Intel, Alteryx, SAP, Rapidminer, Biomax Informatics, Knime, FICO, Salford Systems, Angoss Software, H2O.ai, Reltio, Megaputer Intelligence, Frontline Systems, BlueGranite, Teradata, Suntec India, Dataiku, Wolfram Research ve SenticNet gibi firmalardır (PRnewswire, 2018).

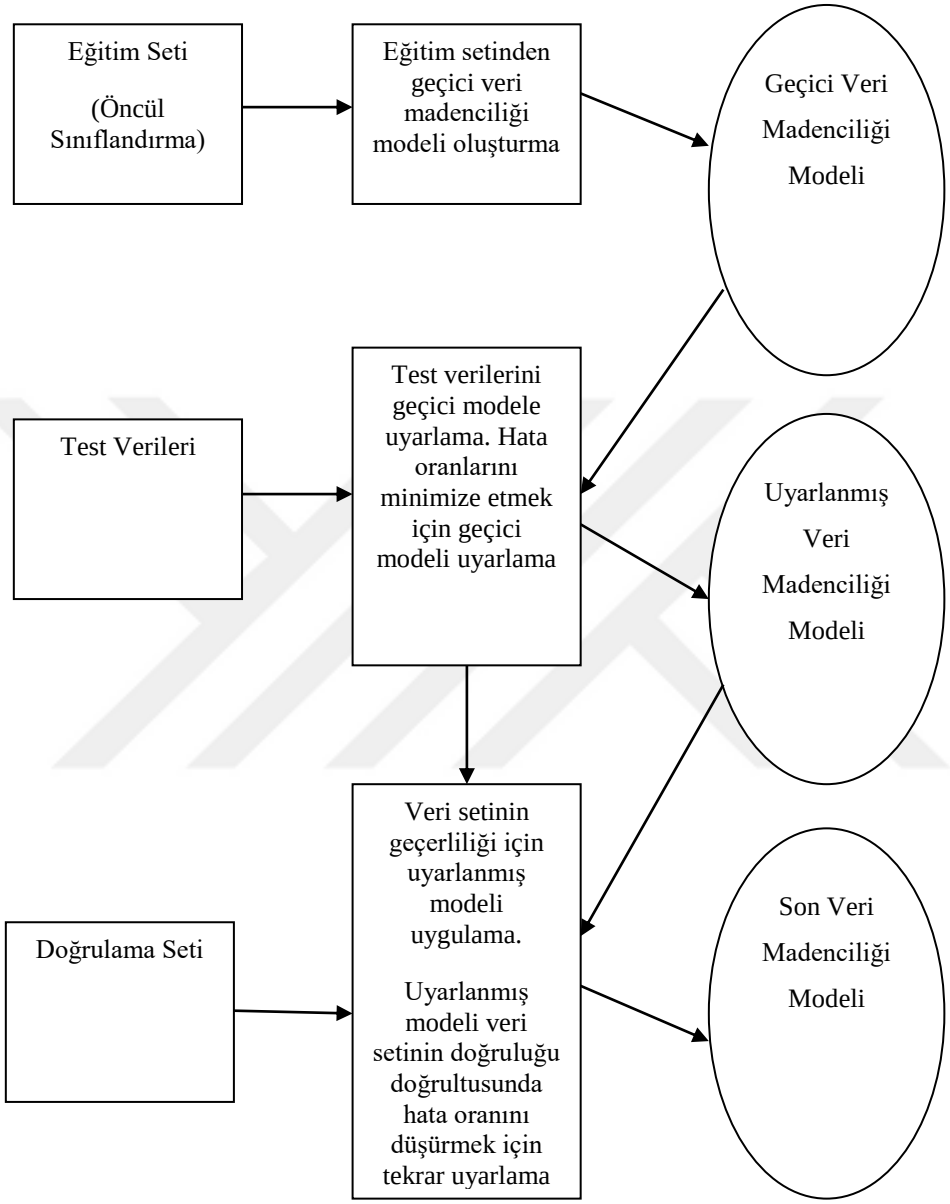
Rapidminer, Weka, R-programlama, Orange, DataMelt, Apache Mahout, ELKI, MOA, KEEL, Rattle, Python ve Knime gibi veri madenciliği programları açık kaynak olarak hizmet vermektedir. Ayrıca IBM SPSS, IBM Cognos, SAS gibi ücretli olarak kullanıcılara sunulan ve kullanımı yaygın olan veri madenciliği araçları da mevcuttur.

1.4. Veri Madenciliği Teknikleri

Veri madenciliği teknikleri temel olarak danışmanlı öğrenme ve danışmansız öğrenme teknikleri olarak iki kategoriye ayrılabilmektedir. Danışmansız öğrenme tekniklerinde hedef nitelik belirlenmez. Veri madenciliği algoritması tüm veri seti içerisinde örüntüler ve yapılar aramaktadır. Bu tekniğin kullanılma nedeni genellikle gizli kalmış etkilerin bulunmasıdır. Ancak birçok veri madenciliği tekniği danışmanlı öğrenme şeklindedir. Danışmanlı öğrenme tekniklerinde hedef bir nitelik belirlenir.

Veri setleri içerisinde yer alan niteliklerden bu hedef nitelik ile yakın ilişki gösteren değerler tahmin edilmeye veya belirlenmeye çalışılır. Danışmanlı öğrenme tekniklerinde veri seti ikiye ayrılmaktadır. Birinci veri seti eğitim veri setidir. Bu eğitim seti içerisinde, tahmin edicilere yönelik öncül bir sınıflandırma yapılmış veriler bulunmaktadır. Ancak bu veri setleri oluşturulurken, algoritmanın ezberci bir yapıya sahip olmaması gerekmektedir. Bu sayede genelleştirme hatalarından kaçınılabilir. Diğer bir unsur ise danışmanlı öğrenme tekniklerinde test verilerinin oluşturulmasıdır. Test verisi içerisinde, hedef nitelik saklı tutularak, eldeki veriler ile hedef nitelik tahmin edilmeye çalışılmaktadır. Danışmanlı öğrenme tekniklerinin başarısını, hedef niteliği en yüksek tahmin etme değeri belirlemektedir. Aşağıdaki şekilde danışmanlı öğrenme veri madenciliği tekniğinin aşamaları gösterilmektedir (Larose, 2006).

Şekil 3: Danışmanlı öğrenme veri madenciliği süreci



Kaynak: Larose, D. T. (2006). Data Mining Methods and Models. JohnWiley & Sons.

1.4.1. K-Means Kümeleme Algoritmaları

K-means algoritmaları, en basit şekliyle kullanılan ve danışmansız öğrenme yöntemi olan kümeleme algoritmaları içerisinde yer almaktadır. Bu teknik 1967 yılında Macqueen tarafından ortaya atılmıştır. Günümüze kadar birçok araştırmacı tarafından kullanılmıştır (Han ve Kamber, 2001). Kümeleme aşaması belirli sayıda küme

içerisinden ele alınan veriyi sınıflamaya yarayan basit bir algoritmadan oluşmaktadır. Algoritma ilk olarak kümelerin merkezini hesaplamaktadır. Merkeze göre de verilerin uzaklığını hesaplamaktadır.Bir gözlem ile her bir kümenin merkezi arasındaki uzaklığa ait olduğu veriye ve en yakın olduğu kümeye göre puan verilerek hesaplama yapılmaktadır. Bu süreç, amaç fonksiyonun hatasının karesini minimize edene kadar devam etmektedir (Jiawei ve Kamber, 2006:402).

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$$

p= kümedeki i'inci nokta,

m_i = i'inci küme

$||x_{ij} - v_j||$ = p ve m_i öklid uzaklığı

C_i = i'inci kümenin veri noktası sayısı

k = küme sayısı

1.4.2. Sınıflandırma Algoritmaları

Sınıflandırma teknikleri çıktıları sınıflandıran, danışmanlı öğrenme süreçleri olarak tanımlanmaktadır. Sınıflandırma teknikleri içerisinde karar ağacı, K-en yakın komşu algoritması, yapay sinir ağları, karar destek makine sistemleri, Naive Bayes ve lojistik regresyon en çok kullanılan tekniklerdendir.

1.4.2.1. Karar Ağaçları Algoritmaları

Karar ağaçları ya da diğer ismi ile sınıflandırma ve regresyon ağaçları geleneksel veri madenciliği ve klasik makine öğrenme algoritmalarındandır. 1980'den bu yana veri madenciliği içerisinde en çok kullanılan tekniklerden birisidir. Karar ağaçlarının bu denli çok kullanılmasının altında yatan temel sebep ise sonuç modelinin kolaylığıdır. Karar ağacı modeli gösterimi kolay olan, anlaşılır ve en önemlisi açıklayıcı bir modeldir (Maindonald, 2012). Karar ağaçlarının en önemli avantajı ise genelleştirme hatasını minimize ederek optimal karar ağacını oluşturmasıdır (Maimon ve Rokach, 2010). İlk karar ağacı ağaçları Quinlan (1986, 1993) tarafından

sunulan ID3 ve C4.5 algoritmalarıdır. Bu algoritmaların pozitif ve etkili sonuç verdiği ispatlanmıştır. Karar ağaçlarında kullanılan algoritmalar şunlardır (Gürsoy, 2009):

- ID3 (Iterative Dichotomiser 3)
- C4.5 (ID3'ün iyileştirilmiş versiyonu)
- C5.0 (C4.5'un geliştirilmiş versiyonu)
- CART (Classification and Regression Trees)
- CHAID (Chi-Square Automatic Interaction Detector)
- Quest

Karar ağacı birçok alanda yaygın bir şekilde kullanılmaktadır. Özellikle tıp, problem çözme ve yönetim bilimi alanlarında karar ağacı tekniğine başvurulmaktadır. Karar ağacının geleneksel bir yapısı bulunmaktadır. Karar ağacı tek bir kök ile başlar, birçok dala ayrılır. Bu dalların en uç noktaları ise yaprakları oluşturur (Maindonald, 2012).

1.4.2.2. K-En Yakın Komşu Algoritması

K-en yakın komşu algoritması (KNN), en çok kullanılan sınıflandırma tekniklerinden birisidir. Bu algoritma ile birlikte bir tahmin ve öngörü yapılması da mümkündür. K en yakın komşu algoritması örnekleme bağlı bir öğrenme tekniğidir. Bu teknik içerisinde yer alan eğitim veri setleri depolanarak, sınıflandırma yapılacağı zaman, bu kaydedilmiş verilere olan benzerlikler ele alınmaktadır (Larose, 2006). KNN'de ayrıca nesnenin en yakın komşu nesneye olan ağırlıklı etkisinin uzaklığı da göz önünde bulundurulmaktadır. Bu bakımdan birbirine en yakın nesneler uzak olanlara göre daha fazla aynı özelliği taşımaktadır. Bu teknik daha şok metin veri sınıflandırılmasında kullanılmaktadır (Taniar, 2008). KNN formülasyonu aşağıdaki gibidir:

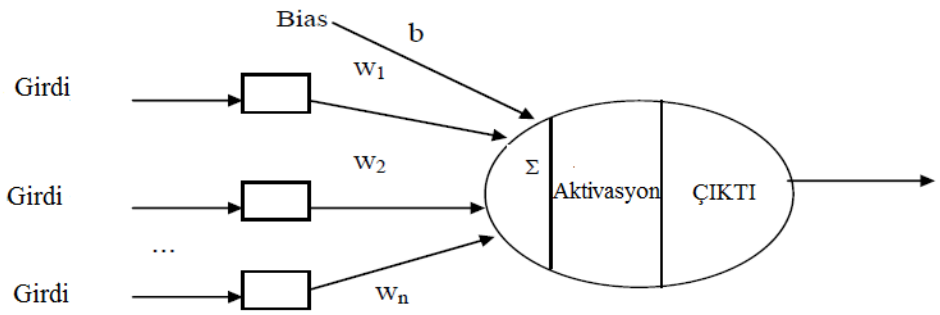
$$\textbf{Benzerlik}\ (x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)}$$

Örneklem içerisinde yer alan n kadar nitelik bulunmaktadır. Sayısal nitelikler için $f(X_i, Y_i) = (X_i - Y_i)^2$ olarak hesaplanır.

1.4.2.3. Yapay Sinir Ağları (YSA)

Bu matematiksel model insan beyninin çalışma düzenini gösteren biyolojik sinir ağlarından ilham alınarak ortaya koyulan bilgi süreç paradigmasıdır. Yapay sinir ağları doğrusal olmayan modellerdir. Doğrusal modellerin en büyük sıkıntısı biaslardır. Ancak bilindiği üzere gerçek hayatta yer alan veriler doğrusal olmayan bir yapı göstermektedir. Bu yüzden yapay sinir ağları doğrusal olmayan modellemeler üzerine kuruludur. Ayrıca yapay sinir ağları, eksik olan ya da gürültü düzeyi yüksek olan ve kesin olmayan veri ve örüntülerden problem çözme özelliğine sahiptir. Yapay sinir ağları, nesneleri sınıflandırmaya yaramakta ve çalışma prensibi tekli akış grafiğine göre Şekil 4'te gösterilmektedir.

Şekil 4: Tekli akış grafiği



Kaynak: Rahman,M. M. (2014). Machine Learning Based Data Pre-processing for the Purpose of Medical Data Mining and Decision Support. Yayınlanmamış Doktora Tezi, University of Hull.

Yapay sinir ağları modeli kurarken karar verilmesi gereken en önemli noktalar, verinin hazırlanması, girdilerin seçimi, network tipinin belirlenmesi ve bu network tipinin tasarlanmasıdır. Yapay sinir ağlarında yer alacak girdilerin iyi bir şekilde belirlenmesi gerekmektedir (Brockmann, Hufnagel, ve Geisel, 2006). Yapay sinir ağlarında girdiye konu olan her veri kategorik dahi olsa standardize edilme anlamında 0 ve 1 arasında değer almak zorundadır. Yapay sinir ağlarında her bir bilgi birbirine bağlı birer işlemci olarak ele alınmaktadır. Her bir işlemcinin birçok bitişik alan ile değişik ağırlığa sahip ilişkisi bulunmaktadır. Yapay sinir ağları girdilerin birbiri ile olan ilişki ağırlıklarını kullanarak ortaya çıktı değerlerini koymaktadır. Bir sinir ağır girdi tabakası, gizli tabaka ve çıktı tabakasından

oluşmaktadır. Buradaki mantık, her bir tabakanın birbiri ile bağıntılı olduğu varsayımıdır. Bu açıdan bazı durumlarda birden fazla gizli tabaka olduğu da görülmektedir (Larose, 2005).

Sinir ağlarında en çok kullanılan model feedforward sinir ağlarıdır. Aynı zamanda bu model multi-layer perceptrons (MLP) olarak da adlandırılmaktadır. Feedforward sinir ağları bir ya da birçok tepki yaratan ya da çıktıya etki eden, tahmin edici veya girdi verilerinin arasındaki ilişkiyi modellemek için uygun bir tekniktir. Bu tekniğin kullanılma nedeni, kaç tane girdi verisinin çıktı verisini etkilediğinin haritalanmasını sağlamaktır. Bir arada kullanılan çoklu tahmin ve haritalama taslakları problem yaratırken, feedforward sinir ağları veri madenciliği için daha uygun olmaktadır. Feedforward sinir ağları aralarında birçok ilişki bulunan nöronlar, nodelar ve hücrelerden oluşmakta ve bunlar katmanlarda yer almaktadır. Her nöron bir bilgi şeklinde çıktıları oluşturmaktadır. Her bir nöronun birbiri ile olan değişik ağırlıkta ilişkisi sonucunda bilgi yaratılmaktadır. Feedforward sinir ağında, modeli oluşturan kişi tabakaların sayısına, bu tabakalardaki her bir node sayısına ve birbiri ile olan ilişkilerine karar vermek zorundadır. Ancak sinir ağlarında, garanti bir şekilde sonuç verecek yazılı kesin bir algoritma bulunmamaktadır (Brockmann, Hufnagel, ve Geisel, 2006).

1.4.2.4. Karar Destek Makine Sistemleri

Karar destek makine ilk olarak 1995 yılında girdi bileşenli vektör olarak sınıflandırmada kullanılmak amacıyla ortaya çıkmıştır. Karar destek makine nesneleri, nesnelerin doğrusal ayrımlarına bakarak iki sınıf altında toplamaktadır. Bu doğrusal ayrıma karar vermede nesneler uzayında yer alan nesnelerin hiperdüzlemleri etki etmektedir. Karar destek makinede temel mantık, maksimum genelleme seviyesini sağlayan hiperdüzlemi bulmak ve aşırı uyumdan kaçınmaktır. Eğitim setlerinde minimal uzaklığa sahip maksimum marjinli hiperdüzleme destek vektörü denilmektedir. Maksimum marjinli hiperdüzlemin pozisyonunu bulmak ve destek vektörü saptamak amacıyla dual optimizasyon problemi kurgulanmaktadır.

Hiperdüzlemin formülasyonu aşağıdaki gibidir:

$$\sum_{j=1}^{m} w_j \varphi_j(x) + b = 0$$

x = Girdi uzayındaki vektör

$\{\varphi_j(x)\}_{j=1}^{m}$ = Nesneler uzayında doğrusal olmayan değişmeye uğramış vektörlerdir.

w_j =vektör ağırlıklarıdır.

b = Bias.

1.4.2.5. Naive Bayes Sınıflandırıcı

Naive Bayes algoritması ele alınan örneğin sınıfa olan üyelik olasılığını tahminlemede kullanılmaktadır. Naive Bayes tekniği, gerçekleşmesi en muhtemelen olasılıkları tespit eden olasılık teoremi üzerine kuruludur. Bu teknikte her bir girdinin aynı önem derecesine sahip olduğu düşünülmektedir (Bramer, 2007). Naive Bayes algoritmasında, bağımsız olmayan bir X nedenler seti içerisinden, bunun sonucu olabilecek olayın olasılığı tespit edilmektedir. Temel çıkış noktası Bayes Teoremi'dir (Merih, 2017).

k veri seti içerisinde karşılıklı özel ve ayrıntılı sınıflandırmalar c_1, c_2, ..., c_k ve öncelikli olasılıklar $P(c_1)$, $P(c_2)$,...., $P(c_k)$ olduğunda, *n* nitelikleri a_1, a_2, ..., a_n ve örneklemdeki değerleri v_1, v_2, ...,v_n olduğunda, c_i sınıfının görünen olasılık oranı aşağıdaki gibi gösterilmektedir (Bramer, 2007):

$$P(c_i) \times P(a_1 = v_1 \text{ and } a_2 = v_2 \ldots \text{ and } a_n = v_n \mid c_i)$$

Tüm niteliklerin varsayımları bağımsız olduğundan formül aşağıdaki forma dönüşmektedir:

$$P(c_i) \times P(a_1 = v_1 \mid c_i) \times P(a_2 = v_2 \mid c_i) \times \ldots \times P(a_n = v_n \mid c_i)$$

Tek bir nitelik için kullanılan formül ise aşağıdadır:

$$P(c_i) \times \prod_{j=1}^{n} P(a_j = v_j \mid class = c_i).$$

1.4.2.6. Lojistik Regresyon

Doğrusal regresyonda değişkenler arasındaki ilişkiler incelenirken, bu değişkenlerin kategorik olmasından çok, sürekli olması tercih edilmektedir. Kategorik verilerin olduğu durumlarda, doğrusal regresyona benzeyen bir yöntem olan lojistik regresyon ortaya çıkmaktadır. Lojistik regresyonda, kategorik veriler arasındaki ilişkiler ortaya koyulmaktadır (Larose, 2006). Lojistik regresyon yöntemi günümüzde özellikle tıp alanında yaygın olarak kullanılmaya başlanmıştır. Lojistik regresyon, diskriminant analizi ve crosstabs yöntemine alternatif olarak uygulanmaktadır. Doğrusal regresyonda değişkenler arasındaki ilişki incelendiği gibi (Şıklar, 2000:5), lojistik regresyonda da ilişkiler incelenmektedir. Ancak lojistik regresyonu doğrusal regresyondan ayıran en önemli fark, bağımlı değişkenin kategorik olmasıdır. Lojistik regresyondaki bu fark hipotezlerde kendisini göstermektedir (Bircan, 2004).

1.4.3. Birliktelik Kuralları

Birliktelik kuralları iş dünyasında ilişki analizleri ya da sepet analizleri olarak adlandırılmaktadır. Birliktelik kurallarında amaç iki ya da daha fazla nitelik arasında birliktelik bulmaya çalışmaktır. Birliktelik kuralının olası algoritması nitelik sayısı ile ilişkilidir. k değerine nitelik sayısı dersek; "$k.2^{k-1}$" olası birliktelik kuralı olacaktır.

Tablo 3.'de örnek olarak bir yol kenarına kurulan tezgah düzeneği şeklinde alışveriş mekanı verilmiştir. Bu alışveriş tezgahında 100 ürün olduğunu varsayalım. Alışveriş olsun ya da olmasın birliktelik kural değeri 100.2^{99}'dur (Larose, 2005).

Tablo 3: Yol kenarı sebze standından yapılan alışveriş

Alışveriş	Satılan Sebzeler
1	Brokoli, yeşil biber, mısır
2	Kuşkonmaz, kabak, mısır
3	Mısır domates, bezelye, kabak

4	Yeşil biber, mısır, domates, bezelye
5	Bezelye, kuşkonmaz, brokoli
6	Kabak, kuşkonmaz, bezelye, domates
7	Domates, mısır
8	Brokoli, domates, yeşil biber
9	Kabak, kuşkonmaz, bezelye
10	Bezelye, mısır
11	Yeşil biber, brokoli, bezelye, kabak
12	Kuşkonmaz, bezelye, kabak
13	Kabak, mısır, kuşkonmaz, bezelye
14	Mısır, yeşil biber, domates, bezelye, brokoli

Sepet analizi yapılırken alışveriş veri formatı ya da tabular veri formatı kullanılmaktadır. Alışveriş veri formatında ID bölmesi ve içerik bölmesi şeklinde sadece iki kısım bulunur. Her bir kayıt tek ifadeyi gösterir. Yukarıdaki tablo alışveriş veri formatına göre hazırlanmıştır. Tabular veri formatında ise her bir alışveriş birbirinden ayrılmaktadır. Alışveriş veri formatında gösterilen 1. satır, tabularda ayrı satırlarda 1...brokoli, 1... yeşil biber, 1... mısır olarak gösterilecektir.

Tablo 4: Birliktelik kuralı tablo veri format örneği

Alışveriş	Kuşkonmaz	Bezelye	Brokoli	Mısır	Yeşil Biber	Kabak	Domates
1	0	0	1	1	1	0	0
2	1	0	0	1	0	1	0
3	0	1	0	1	0	1	1
4	0	1	0	1	1	0	1

5	1	1	1	0	0	0	0
6	1	1	0	0	0	1	1
7	0	0	0	1	0	0	1
8	0	0	1	0	1	0	1
9	1	1	0	0	0	1	0
10	0	1	0	1	0	0	0
11	0	1	1	0	1	1	0
12	1	1	0	0	0	1	0
13	1	1	0	1	0	1	0
14	0	1	1	1	1	0	1

Birliktelik kuralında destek değeri A $\Rightarrow$ B , D içerisindeki A ve B'yi içeren alışverişin oranıdır.

Destek= P(A∩B)= $\frac{A \text{ ve } B' \text{yi içeren tüm alışverişler}}{\text{Toplam Alışveriş Sayısı}}$

Güvenilirlik= P(B/A) = $\frac{A \text{ ve } B' \text{yi içeren alışverişlerin sayısı}}{A' \text{yı içeren alışveriş sayısı}}$

Formülasyona göre kuşkonmaz alıp aynı zamanda bezelye alanların destek oranı 5/14=%35.7'dir. Güvenilirlik ise 5/6=83.3'tür.

1.4.4. Genetik Algoritma

Genetik algoritma popülasyon temelli bir araştırma ve optimizasyon tekniğidir. Bu teknik doğal genetikler ve Darwin'in doğal seleksiyon prensibine bağlı olarak işlemektedir (Goldberg, 1989). Bu teknik ilk olarak 1965 yılında Prof. John Holland tarafından ortaya atılmıştır. Genetik algoritmanın üstlü kodlu GA, gerçek kodlu GA, dağınık GA gibi bir çok versiyonu mevcuttur (Taniar, 2008:115).

1.4.5. Güncel Teknikler

Veri madenciliği özellikle internet ve bilgisayar teknolojilerinin yaygınlaşmasıyla birlikte yeni alanlarda da kullanılmaya başlanmıştır. Bu alanlar zaman serileri veri madenciliği, web madenciliği ve metin madenciliği olarak araştırmalarda yerini almıştır.

1.4.5.1. Zaman Serileri Veri Madenciliği

Zaman serileri teknikleri ile veri madenciliği alanının birleşmesinden zaman serileri veri madenciliği oluşmuştur. Bu yöntemde, zaman serileri analiz etmede veri madenciliği yöntemi kullanıldığından, zaman serileri analizindeki durgunluk ve lineer gereksinimleri ortadan kaldırmaktadır (Aydin, Karaköse ve Akin, 2008).

1.4.5.2. Web Madenciliği

Günümüzde veri madenciliği araştırmalarına konu olan veri setleri internet üzerinden depolanan verilerden oluşmaktadır. Özellikle sosyal medya mecralarında yer alan tweetler, paylaşılan fotoğraflar, yorumlar ve e-postalar göz önünde bulundurulduğunda, ortaya oldukça büyük bir veri çıkmaktadır. Büyük şirketler, bu verilerin muazzamlığının farkına vararak veri madenciliği çalışmalarını web üzerinden verilerle yürütmeye başlamışlardır. Web, örüntüler, müşteri davranışları ve trendler açısından büyük bir kaynak oluşturmaktadır. Web üzerinden faydalı örüntüler ve bilgi elde etmeye yarayan veri madenciliği türüne web madenciliği ismi verilmektedir (Gürsoy, 2017).

1.4.5.3. Metin Madenciliği

Gazetelerde, dergilerde, bilimsel dergilerde ve hatta özet metinlerde bile birçok bilgiye dönüşmeyi bekleyen veri bulunmaktadır. Ancak bu basılı yayınların sayısına bakıldığında ortaya büyük bir veri yığını çıkmaktadır. Bu açıdan, kütüphanecilik bilgisi ile veri bilimi harmanlanarak ortaya metin madenciliği çıkmıştır. Metin madenciliği içerisinde Naive Bayes, En yakın komşu ve karar ağaçları gibi metotlar rahatlıkla uygulanabilmektedir. Ancak metin madenciliğinin kendine has bir veri elde etme yöntemi bulunmaktadır. Metin madenciliği öncesi aranacak olan niteliğe

uygun kelime çantaları hazırlanmaktadır. Bu kelimelerin ikili ya da üçlü kombinasyonlara göre önem dereceleri belirlenmekte ve bu kelimelerin geçtiği her cümle, paragraf ve doküman taranmakta ve elde edilen kelimeler sınıflandırılmaktadır. Cümleler arasında belirli durak kelimelerin belirlenmesi gerekmektedir. İngilizcede bu durak kelimeleri "I, an, a, the, is, you, and, of" gibi 319 adet sözcük olarak belirlenmiştir. Ancak durak kelimeleri her dilde farklılık göstermekte ve metin madenciliği uygulanacak dile göre planlanmalıdır (Bramer, 2007).

Metin madenciliği beş adımda toplanabilir (Gürsoy, 2017):

1. Metin koleksiyonu oluşturma:

   Yayınlardan veri setlerini teşkil edecek şekilde kütüphane oluşturma adımıdır.

2. Metin önişleme:

   İşaretleme, gövdeleme, sözlük oluşturma ve gereksiz kelimeleri ayıklama aşamasıdır. Ayrıca bu aşamada normalleştirme, Türkçeleştirme, dizgi parçalama, büyük/küçük harf dönüşümü, durak kelimelerinin çıkarılması, kelime köklerine ayırma ve terim doküman matrisi oluşturma işlemleri de yapılmaktadır.

3. Veri madenciliği:

   Yapılandırılmış metinlerden anlamlı bilgiler çıkarma aşamasıdır. Kullanılan teknikler arasında konu modelleme, duygu analizi, sosyal ağ analizi, eğilim analizi ve müşteri bağlılık analizi bulunmaktadır.

4. Değerlendirme:

   Elde edilen sonuçların değerlendirilmesi aşamasıdır.

5. Yorumlama

   Değerlendirmeler sonucu elde edilen bilgilerin alana uygun bir şekilde yorumlanması aşamasıdır.

İKİNCİ BÖLÜM

SAĞLIK HİZMETLERİNDE VERİ MADENCİLİĞİ VE UYGULAMALARI

Teknolojideki depolama imkanlarının artması, bilgisayar gibi aygıtların hesaplama güçlerindeki gelişmeler ve özellikle her alanda yaşanan dijitalleşme, sağlık hizmetlerinde veri analizine ve modellemelerine imkan tanımış özellikle sağlık alanında yeni bir dönem başlamıştır (Earley, 2015). Dijitalleşme ile birlikte elde edilen veriler sayesinde, 10 yıl içerisinde veri bilimcilerin ve yazılımcıların tüm biyolojik bilimlerin toplamından daha fazla sağlık alanına katkı sağlayacağı ön görülmektedir (Khosla, 2012). Ancak sağlık alanında verilerin artması, araştırmacıların karşısına yığınla veri içerisinden araştırma ve analiz yapmaya yarayacak, fayda yaratacak ve verileri kullanıma hazır hale getirecek akıllı teknolojilerin de tasarlanma sorununu beraberinde getirmiştir. Bahsi geçen akıllı teknolojiler, bilişim ve bilgisayar biliminin kesişmesi ile ortaya çıkan sağlık bilişimi ismini oluşturmaktadır. Sağlık bilişimi, biyo-bilişim, klinik bilişim, halk sağlığı bilişimi, nöro bilişim gibi alanların bir birleşimidir. Sağlık bilişim bu alanlardan bilgi edinmekte, geri bildirim almakta, depolamakta ve veri madenciliği teknikleri ile analizler yapmaktadır (Herland, Khoshgoftaar ve Wald, 2014:1).

2.1. Sağlık Araştırmalarında Veri Analitikleri

Sağlık hizmetleri üzerinde yapılan araştırmalarda uzamsal veriler kullanılarak gözlemlenen çıktıların tekrarlanma durumları incelenmektedir (Ganguli, 2011:299). Bu incelemeler sırasında istatistiksel analizler kullanılmakta ve kurulan modelin zamanla tekrarlanma olasılıkları, birbiri ile olan ilişkileri ve korelasyonları incelenmektedir. Sağlık hizmetlerinde veri analitikleri, muayene verileri ve klinik karar destek için geliştirilen sağlık puanları ya da indeksleri şeklinde incelenmektedir.

2.1.1. Sağlık Muayenelerinden Veri Analitiği

Tıbbi araştırmalar genel olarak sağlık muayenelerinden elde edilen verilerin, risk faktörlerini ve bu faktörlerin yaygınlığını saptamak amacıyla analiz edilmesi üzerine kuruludur (Ressing, Blettner ve Klug, 2010:187). Muayenelerden elde edilen veriler ile oluşturulan varsayımlar, regresyon analizinin değişik versiyonları sayesinde test edilerek sonuçlar kaydedilmektedir.

2.1.2. Sağlık Puanlama (İndeksleme) Sistemleri

Sağlık muayenelerinden veri analitiği içerisinde sağlık puanlama sistemleri önemli bir yer tutmaktadır. Sağlık taramalarından geçen kişilerin verileri, çoğu zaman sayısal olarak kayıt altına alınmaktadır. Sağlık puanlama sistemi ise bu sayısal verileri ele alarak hesaplamalar yapmakta ve özellikleri belirlemektedir. Elde edilen puanlamaya göre ise sağlığa yönelik bir değerlendirme yapılmaktadır. Karar vericiler, uygulayıcılar ve araştırmacılar sayısal verilere ihtiyaç duymaktadırlar. Örneğin bir hastalığın ülke çapında yaygınlığına yönelik sayısı sağlık puanlama sistemi sayesinde ortaya çıkmaktadır (Kaplan, vd. 1976:478).

2.2. Sağlık Hizmetleri Verilerinden Veri Madenciliği

Veri madenciliği yığın bir veriden anlamlı ve yararlı bilgiler edinmeyi hedeflemektedir. Bu hedefi doğrultusunda araştırmacılar, sınıflandırmalar, regresyon, kümeleme, birliktelik kuralları ve geliştirilen diğer veri madenciliği teknikleri ile bilgi edinmeye çalışmaktadırlar. Özellikle verilerin dijital bir şekilde depolanması ile birlikte birçok araştırmacı sağlık hizmetlerinin kalitesini arttırmak amacıyla veri madenciliği yapma yoluna gitmektedir (Jiawei ve Kamber, 2006:5).

Sağlık hizmetlerinde hazır verilerden veri madenciliği, genellikle sınıflandırıcı şekilde yapılmaktadır. Sınıflandırma, daha önceden belirlenen birçok kategori ile birlikte model kurularak, ilerisi için tahminler yaratmaktır (Jiawei ve Kamber, 2006:18). Yığın halde bulunan verilerin içerisinde, birçok hastalıkla ilgili bilgiler yatmaktadır. Sınıflandırma bu bilgileri açığa çıkarmayı ve ilerisi için önleyici şekilde tahmin edici modeller kurmayı amaçlamaktadır. Sınıflandırma danışmanlı bir öğrenme tekniğidir. Kanser, kalp krizi ve diyabet gibi ölümcül hastalıklarda önleyici

hizmetlerin etkinliği arttırmak amacıyla sınıflayıcı veri madenciliği teknikleri kullanılmaktadır (Sharma, Singh, ve Khatri, 2016).

2.3. Sağlık Araştırmalarından Metin Madenciliği

Kliniklerde bulunan elektronik kayıtlar, sağlık sisteminin bütün yönlerini içerisinde barındırmaktadır. Bu elektronik kayıtlar içerisinde çok büyük miktarda sayısal ve yazılı bilgi bulunmaktadır. Yazılı verilerin analizleri metin madenciliği sayesinde olmaktadır. Metin madenciliği ile birlikte istatistiksel olarak kullanıma elverişli olmayan kalitatif veriler, sayısal değerlere çevrilerek kullanıma kolay hale getirilmektedir. Metin madenciliğinin ilk adımı, belgelerin toplanma aşamasıdır. Sosyoloji ve iletişim alanlarında sıkça kullanılan metin madenciliği elektronik kayıtlarla birlikte sağlık alanında da yer edinmeye başlamıştır. Toplanan tıbbi belgeler ilk olarak analiz edilmekte ve duraklama noktaları belirlenmektedir. Bu aşamadan sonra başlangıç listeleri oluşturma aşaması gelmektedir. Oluşturulan listeler içerisinde kümeler yaratılmakta ve neticesinde bu kümeler yorumlanmaktadır. Yorumlanan kümeler ise sayısal veriler atanarak tahmin edici bir model şeklinde veri madenciliğine hazır hale getirilmektedir. Örneğin sigara ve kanser ile ilişkili kayıtlar inceleme altına alınarak, benzerliklerden yola çıkılarak bu iki alan arasında kümeler oluşturulabilmektedir. Sonrasında bu kümeleri nitelendirecek şekilde sayısal değerler atanarak analize hazır hale getirilmektedir. Metin madenciliğinin sağlıkta yapılacak ölümcül hataları azaltıcı bir etkiye sahip olduğu düşünülmektedir. Özellikle yanlış ilaç kullanımı, tedavi edici uygulamaların kişiler arasındaki farklılıklarının saptanması gibi konularda metin madenciliğinin önemi artmaktadır (Raja, vd., 2008).

2.4. Sağlık Hizmetlerinde Veri Ambarları

Veri ambarları bir organizasyonun geniş anlamda depolanmış kayıtlarından oluşmaktadır. Veri ambarları sağlık hizmetleri açısından oldukça yeni bir alan olarak nitelendirilmektedir. Ancak veri ambarlarının, diğer sektörler açısından kullanımı oldukça eskiye dayanmaktadır. Sağlık sektöründe dijitalleşme ile birlikte kayıtların artması ile birlikte, hastaların bilgilerine ait veri ambarları oluşturulmaya başlanmıştır. Sağlık hizmetleri açısından veri ambarları aşağıdaki yararları taşımaktadır (Berndt, vd., 2001:1; Nimmagadda, Nimmagadda ve Dreher, 2011):

- Sağlık hizmetlerine yönelik stratejik planları desteklemekte ve yönetim kalitesini arttırmaktadır,
- Hastalar, kişiler ve servis sağlayıcılar için iyileştirilmiş sonuçlar sağlanmaktadır,
- Ulusal ve uluslararası çapta kamu sağlığına yönelik girişimlere olanak vermektedir,
- Sağlık hizmetleri sağlayıcılarının sağlık politikalarını ve bütçelerini etkileme gücünü arttırmaktadır.

Sağlık hizmeti sağlayıcılarının veri ambarlarına yönelik veri madenciliği sayesinde, hastaların tedavi sonrası bakımlarına yönelik iyileştirmeler sağlanabilmektedir. Veri ambarları içerisinde hasta ile ilgili birçok bilginin yanı sıra, benzer hastalığa sahip kişilerin bilgileri, tedavi şekilleri ve tedavi sonrası bilgileri bulunmaktadır. Bu verilerin analizleri neticesinde daha doğru sonuçlar alınabilmektedir. Diğer yandan veri ambarları sadece birey açısından değil; tüm nüfus açısından da önemli sonuçları içerebilmektedir. Tüm nüfusu etkileyen kronik hastalıkların yönetilmesi ve mücadelesine yönelik bilgiler, veri ambarlarına yönelik veri madenciliği neticesinde elde edilebilmektedir. Ayrıca sağlık ile ilgili kuruluşlara yürütme ve eylem için gerekli bilginin sağlanması, her bir kurumun kendisini diğer kuruluşlarla kıyaslaması, insan kaynakları, ihtiyaçlar ve arz açısından kurumların planlama yapması, finansal modeller oluşturması ve bütçe konusunda otoritelerle görüşmeler esnasında performans bilgilerinin ortaya koyulması açısından veri ambarları önemlidir (Berndt, vd., 2001:3-4).

2.5. Elektronik Hasta Dosyalarından Veri Madenciliği

Sağlık hizmetlerinde veri madenciliği açısından her bir hastanın kaydının oluşturulması, kullanılması, depolanması ve bu kayıtların gerektiğinde çağrılması önem arz etmektedir. Elektronik hasta dosyaları aynı zamanda yönetimsel ve kliniğe ait büyük veriyi de içermelidir. Bu veriler içerisinde sadece hasta kayıtları değil; aynı zamanda doktorların listesi, montaj sırası, izin belgeleri, hemşirelerin izleyecekleri süreçler, geçmişe yönelik uygulamalar ve gelecek ile ilgili beklentiler de

bulunmalıdır. Bu kayıtların veri madenciliği teknikleri ile analizi neticesinde, geçmişe dayalı bir karar verme aşaması yaratılabilir (Gordon, vd., 1998).

2.6. Kronik Hastalıklar İçin Erken Uyarı Sistemleri ve Veri Madenciliği

Erken uyarı skorları (Early Warning Score-EWS) hastaların kötü durumlarla sonuçlanabilecek sağlık bozukluklarını keşfetmeye yönelik sistemlerden oluşmaktadır. Hastaların yıkımla sonuçlanabilecek durumları genel olarak kayda alınmış kötü durumlarından veya fiziksel parametrelerinden belirlenmektedir. Bu kayıtlar ve fiziksel parametreler önleyici hizmetlerin oluşturulması ve doğru hastaların saptanması açısından önem arz etmektedir. Sağlık hizmetleri sağlayan kuruluşlarda kaynaklardaki yetersizliklerden dolayı ilgilenebilecek hasta sayıları sınırlı sayıdadır. Bu amaçla risk unsuru taşıyan hastaların erkenden belirlenmesi, hastaların ölümcül durumlardan bertaraf edilmesi açısından önemlidir. Erken uyarı sistemleri sistolik kan basıncı, nabız sayısı, solunum oranı, vücut sıcaklığı ve bilinç durumu skorlarından oluşmaktadır (Subbe, vd., 2001:521-522).

2.7. Hastane Enfeksiyonu Kontrolünde Veri Madenciliği Uygulamaları

Sağlık hizmetlerinin başarısı açısından hizmetin sağlandığı kuruluşun steril olması gerekmektedir. Ancak her sene birçok hastanede bulunan hasta, enfeksiyon yüzünden muzdarip olmaktadır. Amerika'da enfeksiyon kapan kişi sayısının yılda 2 milyon civarında olduğu tahmin edilmektedir. Veri madenciliği sayesinde, bilgisayar destekli gözetim sistemleri kurulmaktadır. Bu gözetim sistemleri, yüksek riskli hastaları belirlemekte, olası enfeksiyon durumlarını saptamakta ve geçmişte yaşanan benzer olayların ortaya çıkarılmasına odaklanmaktadır. Hastanelerde kurulan gözetim sistemleri, enfeksiyon ile savaşta ve enfeksiyonun kontrolünde yeni ve ilginç bulguları veri madenciliği teknikleri ile tanımlamaktadır. Bu sistemler genelde birliktelik kuralı kullanmaktadır. Veri madenciliği sayesinde enfeksiyonu kontrol etmekte hassas bir gözetim sistemi kurmak mümkündür (Obenshain, 2004:694).

2.8. Biyoterörizm

Biyolojik ajanlar, canlılarda hastalık oluşturan veya ölümlere yol açan bakteri ve virüs gibi mikroorganizmalara denilmektedir. Biyolojik ajanlar, laboratuvar ortamlarında yapılan değişiklikler ile biyolojik silah olarak kullanılabilmektedir. Biyolojik silahlar üretmekteki amaç düşman ülkenin askeri gücünü azaltmak veya fizyolojik ihtiyaçlarını gidermenin önüne geçmektir. Biyoterörizm hava ya da su gibi yollar ile belli bir alanı etkileyecek şekilde biyolojik hastalık yapıcı etkenlerin bilerek bir ülkeye salınmasıdır. Biyoterörizm içinde bulunduğumuz çağın en büyük tehditlerinden birisidir (Yüksel, 2016). Bu tehdide karşı ülkeler sağlık kayıtları ile koordineli şekilde veri madenciliği yaparak biyoterörizmin önüne geçmeye çalışmaktadır. Örneğin hastanelere gelen çeşitli şikayetler, veri madenciliği esasında incelenerek bu şikayetlerin daha önceden belirtilen semptomlar ile uyuşması halinde bölgenin karantina altına alınmasına kadarki sürecin etkili bir şekilde planlanmasına yardımcı olmaktadır (Talbot, 2001).

2.9. Giyilebilir Teknolojiler

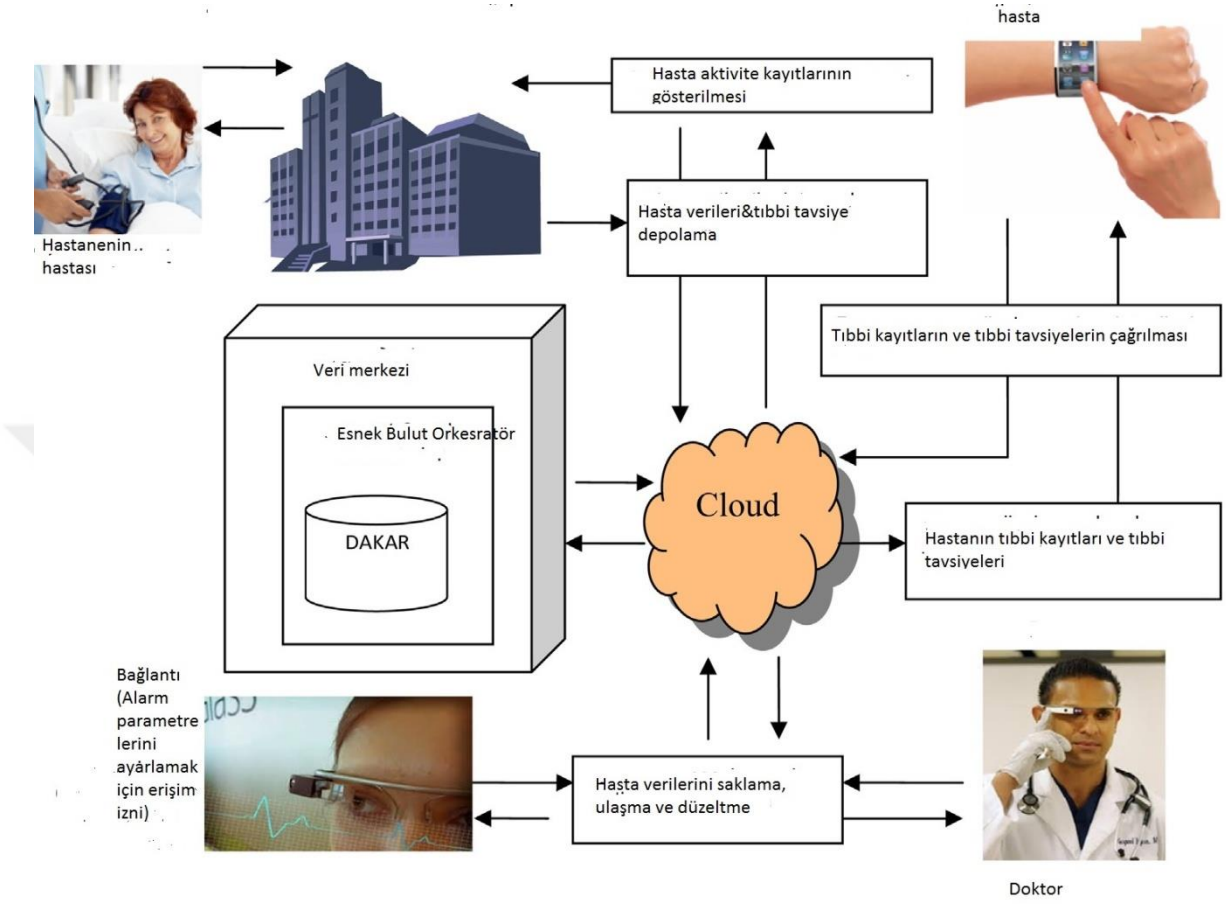
İnsanların vücudunun herhangi bir noktasında direkt giyebildikleri elektronik teknolojilerin hepsine giyilebilir teknoloji ismi verilmektedir. Bu teknolojik aletler, laptopların yaptığı her işlemi yapabilmelerine karşın; laptoplara göre daha elle taşınabilir formda yer almaktadırlar. Giyilebilir teknolojiler gerçek zamanlı olarak, bu teknolojik aletleri giyen kişinin verilerini gerekli yerlere iletmektedir (Tehrani ve Michael, 2014). Tıbbi alanda giyilebilir teknoloji hem sağlık hizmetleri açısından hem de ilaç alanında kullanılmaktadır. Giyilebilir teknolojiler sayesinde, kişilerden 24 saat boyunca veri toplanabilmektedir. Ayrıca giyilebilir teknolojilerin görünmeyen bir sensör özelliği kişilerin sağlık hizmetlerinden aldıkları kaliteyi arttırmakta ve doktorlar ile sürekli iletişim halinde olmayı sağlamaktadır. Sağlık alanında giyilebilir teknolojinin iki kullanım şekli bulunmaktadır. İlk olarak birçok kişinin fitness sırasında edinmek istedikleri atılan adım, geçilen mesafe, yakılan kalori ve diyet gibi unsurların takibini sağlayan bir alet olarak kullanılmaktadır. Diğer kullanım alanı ise tıbbidir. Kanser ya da diyabet gibi hastalıklara sahip kişiler için dizayn edilmiş özel giyilebilir teknoloji bulunmaktadır. Google, Apple ve

Samsung gibi markalar kendi giyilebilir teknolojilerini üretmektedir. Örneğin Apple geliştirdiği giyilebilir teknoloji ile kişinin göz yaşından, kandaki glikoz miktarını saptayabilmektedir (Banaee, vd., 2013).

Giyilebilir teknolojiler ile birlikte kişiler artık kendi sağlıkları ile ilgili sorumlulukları kendileri yüklenebilmektedir. Ayrıca nesnelerin interneti konseptinin gelişmesi ile birlikte sağlık alanında önleyici hizmetler bakımından kişilerin daha aktif olması sağlanmaktadır. Nesnelerin interneti (IoT), araçların sahip oldukları verileri herhangi bir arayüze sahip olmadan kolay bir şekilde paylaşmasına olanak sağlamaktadır. Daha önce müşteri ilişkileri, sadakat yaratma ve müşterileri satın almaya teşvik konularında kullanılan IoT, artık kişilerin sağlığı ile ilgili verileri paylaşmaları noktasında da kullanılmaya başlanmıştır (Canhoto ve Arp, 2017:32). Maliyet avantajından dolayı giyilebilir teknolojiler, ülkelerdeki hükümet düzeyinde planlayıcıların da dikkatini çekmeye başlamıştır (Manyika, vd., 2015).

Giyilebilir teknolojilerin ticari olarak piyasada yer edinmesinin üzerinden çok fazla bir zaman geçmemiştir. İlk giyilebilir teknoloji örneği olarak 1980'lerde Pioneering giyilebilir bilgisayarları piyasaya sunulmuştur. 1981 yılında ise sırt çantası şeklinde dizayn edilen giyilebilir bilgisayarların denemesi olmuştur. Daha sonrasında bu ürün kafaya yerleştirilen bir hal alarak, ismi özel göz olmuştur. 1990'lı yıllarda da giyilebilir teknoloji örnekleri görülmüştür. Park Girişimcilik, kayıt alan "özel göz" denilen ürünü piyasaya sürmüştür. Aynı senelerde, BBN Teknolojileri Pathfinder sistemlerini tasarlamış ve radyasyon tespiti yapan ve GPS özelliği taşıyan bir giyilebilir bilgisayarı piyasaya sürmüştür. Aynı yıllarda insan ve bilgisayar etkileşimi sağlayan "Forget-Me-Not" geliştirilmiştir. Takip eden yıllarda, belirli firmaların yine giyilebilir teknoloji girişimleri olduysa da fiyatların 10.000 doların üzerinde olması sorun teşkil etmiştir. Ticari geçmişi çok da uzun olmayan giyilebilir teknolojilerin kişiler açısından adaptasyon ve kullanımı yaygınlaştırma sorunu bulunmaktadır (Sultan, 2015). Giyilebilir teknolojinin sağlık alanında kullanılma süreci Şekil 5'te verilmiştir.

Şekil 5: Giyilebilir teknolojilerin sağlık alanında işleyişi



Kaynak: Sultan, N. (2015). Reflective thoughts on the potential and challenges of wearable technology for healthcare provision and medical education. International Journal of Information Management, 35(5), 521–526.

Giyilebilir teknolojilerin önündeki en büyük engel, bu teknolojilerin kullanımının güç bulunması, istilacı olarak nitelendirilmesi ve biyolojik algılama imkanlarının şuan için düşük seviyede olmasıdır. Bu açıdan giyilebilir teknoloji için kalıcı ve güçlü sinyaller kullanımı arttırılabilir. Yığın veriler içerisinden giyilebilir teknolojiyi kullananlara fayda yaratmaya yönelik büyük bir heyecan vardır. Günümüz endüstrisi şu an için kalp atışı ve kandaki oksijen miktarı gibi biyolojik sinyalleri güçlü bir şekilde işlemeye yarayan teknoloji için çalışmaktadır. Ancak her bir veri setinin kişiden kişiye göre değişmesi, giyilebilir teknolojinin etkinliğini güçleştirmektedir. Çünkü kişilerin beden tipi, saç rengi, cilt rengi ve diğer bir çok unsuru değişkenlik

gösteren karmaşık yapıdadır. Ayrıca vücuda yerleştirilen giyilebilir teknolojilerin, hareket ile birlikte doğru sonuç verememesi gibi unsurlar da üstesinden gelinmesi gereken konulardandır. Diğer önemli bir engel de batarya sorunudur. Giyilebilir teknolojilerin batarya kapasiteleri ne yazık ki sınırlı kalmaktadır. Bu durum giyilebilir teknolojilerin kullanım alanlarını sınırlamaktadır (Ledger, 2014).

Giyilebilir teknolojiler fonksiyonu, görünümü, insan vücuduna yakınlığı ve diğer parametrelere göre belirli bir sınıflandırmaya tâbi tutulabilir. Tablo 5'te bu yapılan sınıflandırmalar görülmektedir.

Tablo 5: Giyilebilir teknolojilerin sınıflandırmaları

Tip	**Özellik**	**Yapılabilirlikleri**	**Uygulama Alanı**
Akıllı Saat	Düşük İşlem gücü, Kullanıcı dostu, Dokunmatik, Sese duyarlı.	Spesifik bilgileri gösterebilmekte, Ödeme yapabilmekte, Fitness/hareket takibi yapabilmekte, İletişime yaramakta, Navigasyon özelliği var.	İşletmeler, Pazarlama alanı, Sigorta sektörü, Profesyonel sporlar, Eğitim, Bilgi-eğlence.
Akıllı Gözlük	Ekrana dokunmaya ve kafa hareketine duyarlı, Sesle komut, El sallama ile komut, Düşük işletim gücü, Doğrudan kulağa ses sağlama.	Görsellik sağlama, Dil arayüzü, İletişim, Görevlerin koordinasyonu.	Ameliyat, Havacılık, Savunma sektörü, Lojistik, Eğitim, Bilgi-eğlence.
Fitness Takip	Yüksek doğruluk, Su geçirmezlik, Hafiflik, Kablosuz bağlantı imkanı.	Fiziksel iyileşme, Navigasyon, Fitness/hareket takibi, Kalp atışını izleme.	Fitness, Sağlık Hizmetleri, Profesyonel Sporlar, Kapalı ve açık alan sporları.

Akıllı Giysi	Kullanıcıya ekran ya da görüntü imkanı sağlamamakta, Veriler vücut sensörleri ve çalıştırıcılardan sağlanmakta,	Kalp atış takibi, Günlük aktivitelerin takibi, Vücut sıcaklığı takibi, Vücudu otomatik olarak ısıtma ya da soğutma.	Profesyonel spor ve fitness, Tıp alanı, Askeri alan, Lojistik.
Giyilebilir Kamera	Kıyafete ya da vücuda takılabilir, Küçük boyutlara sahip olması, Gece görüş imkanı olması.	Gerçek zamanlı insanların fotoğrafını ya da vidyosunu çekebilir, Canlı yayın yapabilir, Fitness/aktivite takibi yapabilir.	Savunma Sanayi, Fitness, Sanayi, Eğitim.
Giyilebilir Tıbbi Cihaz	Acı yönetimi, Psikolojik takip, Glikoz izleme, Uyku takibi, Beyin aktivitelerinin izlenmesi.	Kardiyovasküler hastalıklar, Psikolojik hastalıklar, Kronik hastalıklar, Diyabet, Ameliyatlar, Sinir bilimi, Dermatoloji, Rehabilitasyon.	Fitness, Kardiyovasküler tıp, Psikiyatri, Ameliyat, Onkoloji, Dermatoloji, Solunum Hastalıkları.

Kaynak: Mardonova, M., & Choi, Y. (2018). Review of Wearable Device Technology and Its Applications to the Mining Industry. Energies, 11(3), 10.

Giyilebilir teknolojilerin sunduğu avantajlar özellikle sağlık alanında oldukça caziptir. Giyilebilir teknolojilerin sağladığı veriler doğrultusunda toplum sağlığı ve bireylerin sağlık durumları birebir ve hastaneye gelmeye gerek kalmaksızın izlenebilmektedir. Kişilerin mevcut durumunda bir olumsuzluk algılandığında, ani müdahale şansı yakalama adına doktorlarla anında iletişime geçilerek, olumsuz durumu önlemeye yönelik tavsiyeler alınabilecektir. Örneğin kalp krizi geçirmek

üzere olan bir hastanın verdiği sinyaller ile erken tanı sistemi oluşturulmuş olacaktır. Diğer yandan giyilebilir teknolojiler, hastaların kullandığı ilaçlara karşı vücut tepkimelerinin de incelenmesine olanak sağlamaktadır. Bu sayede hastaların, alerjik olduğu ilaçlar ya da yan etkiler saptanabilecektir. Ayrıca biyosensörler sayesinde şu kişilerin açlık durumları, yorgunlukları, almaları gereken vitamin ve mineral dengeleri, laktik asit seviyeleri ve vücut ısısı da yakından takip edilebilecektir. Olumsuz koşullar ile karşılaşması muhtemel kişilerin bulunduğu yer GPS yardımıyla belirlenebileceğinden, kişilere ulaşma kolaylaşacak ve anında müdahale edilebilecektir. Her ne kadar giyilebilir teknolojilerin güvenilirliği ve özel hayatı ihlal etme olasılıkları, kullanıcılar tarafından endişe duyulan konular da olsa, gelişen teknoloji ile birlikte, giyilebilir teknolojilerin kullanım yaygınlıkları artacak ve veri madenciliğine yeni bir kaynak oluşturacaktır. Elde edilen veriler ışığında, birçok hastalığa çözüm getirebilecek çıkarım ve tahminler oluşturulabilecektir (Casselman, Onopa, ve Khansa, 2017:1016; Haskins, 2017:69).

2.10. Sağlık ile İlişkili Diğer Konular ve Veri Madenciliği

Veri madenciliğinin sağlık hizmetleri açısından kullanım alanları gün geçtikçe çeşitlenmektedir. Veri madenciliği sayesinde gerçekçi bilgiye ulaşıldıkça, sağlık hizmetlerinde yapılan hatalar özellikle finans boyutu ile de yöneticilerin karşısına çıkmaktadır. Bu açıdan bakıldığında, veri madenciliği sağlık hizmetlerinde, suiistimallerin önüne geçilmesi ve fatura yolsuzluklarının engellenmesi amacıyla büyük yararlar sağlamaktadır. Diğer yandan sağlık hizmetlerinde de günümüzde müşteri memnuniyeti kavramı ortaya çıkmıştır. Özellikle özel sağlık hizmetleri sağlayan kuruluşlar için müşterilerin beklediği kalite anlayışının tespiti veri madenciliği sayesinde analiz edilerek yapılabilmektedir. Diğer bir nokta ise laboratuvar sonuçlarındaki hata ve suiistimallerin minimize edilmesidir. Veri madenciliği sayesinde, hatalar ve suiistimaller belirlenerek hasta güvenliği maksimum seviyeye çıkarılabilmektedir. Sağlık hizmetlerinde maliyet oldukça önemli bir unsurdur. Veri madenciliği sayesinde maliyetlere yönelik haritalama yapılarak, kaynakların etkin ve doğru kullanımı sağlanabilmektedir. Bu sayede sağlık hizmetlerinde finansal riskler belirlenebilir ve finansal kaynakların doğru kullanımı sağlanabilir. Son olarak veri madenciliği yönetsel açıdan da önemli bilgiyi

yöneticilere sağlayan bir yapıyı bünyesinde barındırmaktadır. Yöneticilerin doğru kararlar vermesi ve iyi bir şekilde yönetmesi açısından elde edilecek her bilgiye ihtiyaç vardır (Koyuncugil ve Özgülbaş, 2009).

ÜÇÜNCÜ BÖLÜM

DİYABET HASTALIĞI

Diyabet, insülin hormonuna bağlı olarak canlıların karbonhidrat, yağ ve proteinlerden yeterince yararlanamadığı kronik bir metabolizma hastalığıdır. Bu yüzden diyabet hastaları sürekli tıbbi bakım ihtiyacı duymaktadır (Satman ve ark., 2009). Diyabette ya da tıptaki ismi ile diabetes mellitusta; kanda glikoz seviyesinin artması söz konusudur. Bu artış ile birlikte protein ve karbonhidrat metabolizması bozuklukları, hiperglisemi ve damar sertleşmeleri oluşmaktadır (Satman ve ark,2010). Diyabete bağlı artan ölümlerle birlikte, bu hastalık en önemli sağlık sorunları arasında yerini almış bulunmaktadır (Shaw, vd. 2010). Dünya üzerinde her 100 yetişkinden 9'u diyabet hastalığına sahip olmasına rağmen; bu kişilerin yarısı bu hastalığın farkında değildir. Dünya ilaç sektörüne bakıldığında ise, sağlık harcamalarının %12'sinin diyabet üzerine olduğu görülmektedir. 2015 verilerine bakıldığında diyabetli sayısının 415 milyon kişi olduğu görülmekte ve her sene bu sayının arttığı bilinmektedir. 2017 yılı itibari ile dünyada diyabet riski taşıyan 20 yaş üstü kişi sayısının 352 milyon olduğu düşünülmektedir. Diyabete yapılan harcamalara bakıldığında en büyük harcamaların 1. sırada Amerika Birleşik Milletleri'nde (ABD), 2. sırada ise Çin'de olduğu görülmektedir. Ayrıca diyabetten ölenlerin sayısının 4 milyonun üzerinde olduğu tahmin edilmektedir. 2045 yılında diyabetli sayısının %48 artarak, 629 milyon kişiye ulaşacağı düşünülmektedir (IDF-a, 2017).

Ülkemizde 2017 yılı itibari ile diyabet tanısı koyulan kişi sayısı yaklaşık 7 milyondur. Ancak yaklaşık 2.5 milyon kişiye de tanı koyulmadığı; ancak bu kişilerin diyabet olduğu düşünülmektedir. Türkiye'de diyabetin yaygınlığı %12.8 civarıdır. Türkiye'de diyabete yapılan sağlık harcamalarının miktarı 20-79 yaş aralığında 5.5 milyar dolar, 20 yaş altında ise 25 milyon dolar civarındadır (IDF-b, 2017). Bu sayılar değerlendirildiği, Türkiye'de diyabetli sayısında ve diyabete yönelik sağlık harcamalarında büyük bir artış olduğu görülmektedir (Sosyal Güvenlik Kurumu, 2013). Bu artışın tetikleyicileri arasında sağlıksız beslemenin sonucu olarak obezite ve fiziksel aktivitelerin eksikliği sayılabilir (Türkiye Halk Sağlığı Kurumu, 2014).

Tablo 6: Türkiye diyabetli hasta sayısı

Türkiye'de Diyabet Sayıları	**2017**	**2045**
Tahmini Diyabetli sayısı (20-79 yaş)	12.8	16.5
Ülke payı (%)	12.1	12.1
Diyabetli kişi sayısı (x1000)	6,694.5	11,223.3
Teşhis konulmamış diyabetli sayısı (x1000)	2,558.8	4,289.9
Diyabete bağlı ölümler (x1000)	46.3	
Diyabete Bağlı Sağlık Harcamaları (20-79 yaş)		
Toplam harcamalar (milyon dolar)	5,445.6	7,339.4
Diyabetli kişi başına düşen harcama miktarı (USD)	813.6	653.9
Tip 1 diyabet (0-19 yaş)		
Çocuk ve genç sayısı	25,669	
Hamilelikte Hiperglisemi		
Doğuma bağlı hiperglisemi (x1000)	161.853	

Kaynak: IDF-Diabetes-Atlas (2017). 8th Edition 2017 Country Reports, Turkey.

Tablo 6'da görüldüğü üzere ülkemizde diyabetli sayısında bir artış beklenmekte ve diyabete yönelik sağlık harcamalarında da aynı paralelde bir artış olacağı düşünülmektedir. Diyabet hastası olmasına rağmen henüz teşhis koyulmamış kişilerin sayısı da azımsanamayacak kadar çoktur. Bireylerin diyabete bağlı olarak oluşan belirtilerinin tahlil ederek değerlendirmesi ve sağlık kuruluşlarına başvurması önem arz etmektedir. Buna bağlı olarak diyabetin belirtileri, ağız kuruluğu, aşırı idrara çıkma, susama hissi, kilo kaybı, görme bozukluğu, ayaklarda uyuşma, karıncalanma ve yanma, idrar yolu enfeksiyonları, vajinal iltihap, mantar enfeksiyonları, kaşıntı, ciltte kuruma ve yorgunluk hissi olarak sıralanmaktadır (Turkdiab, 2017). Sağlık kuruluşlarınca diyabet tanısında kullanılan kriterler ise Tablo 7'de gösterilmiştir.

Tablo 7: Diyabetin tanısında kullanılan kriterler

Açlık Plazma Glukozu (APG)	≥ 126 mg/dl
Rastlantısal Plazma Glukozu + diyabet semptomları	≥ 200 mg/dl
Oral Glukoz Tolerans Testi'nde (OGTT) 2.st plazma glukozu	≥ 200 mg/dl
HbA1c	≥ %6.5

Kaynak:http://www.turkdiab.org/admin/PICS/webfiles/Diyabet_tani_ve_tedavi__kitabi.pdf.

Diyabet ile birlikte kişilerde başka hastalıklar da görülmeye başlamıştır. Bu hastalıklar kardiyovasküler hastalıklar, hipertansiyon, retinopati, nefropati, periferik ve otonom nöropati, serebrovasküler hastalıklar ve periferik damar hastalıklar şeklinde kendini göstermektedir. Metabolik kontrol sağlayıcı önlemler ile bu hastalıkların önüne geçilebilmektedir (Turkdiab, 2017; Türkiye Halk Sağlığı Kurumu, 2014:14).

ADA'nın 2015 yılında asemptomatik erişkinlerde diyabet veya prediyabet için belirlendiği risk faktörleri ise şunlardır (ADA, 2015):

- Vücut kitle indeksi (VKİ) ölçümü tüm yetişkinlerde ve kilolularda 25 kg/ m^2'den büyükse veya Asya Amerikalılar da 23 kg / m^2 üstünde ve ek risk faktörleri varsa
- Fiziksel hareketsizlik
- Diyabetli birinci dereceden akraba
- Yüksek riskli ırk / etnisite (örneğin, Latin Afro-Amerikan, Amerikan Yerli, Asya Amerikan, Pasifik Adalı)
- 4 kg ağırlığından büyük bir bebek taşıyan veya gestasyonel diabetes mellitus tanısı almış kadınlar
- Hipertansiyon (≥140 / 90 mmHg veya hipertansiyon tedavisi alan)
- HDL kolesterol seviyesi < 35 mg / dl (0.90 mmol / L) ve / veya trigliserid düzeyi 250 mg / dl (2.82 mmol / L)
- Kadınlarda polikistik over sendromu
- Hemeglobin A1C ≥ 5,7 veya bozulmuş glikoz toleransı, bozulmuş açlık glikozu

- Başka klinik durum; insülin direnci ile ilişkili (örneğin, aşırı şişmanlık)
- Kardiovasküler hastalık öyküsü olma

3.1. Diyabet Tipleri

Diyabet; tip 1, tip 2 ve gebelik diyabeti (gestasyonel diabetes mellitus, GDM) olmak üzere başlıca üç gruba ayrılır. Diyabetli bireylerin çoğunluğunu tip 1 ve tip 2 diyabetli bireyler oluşturmaktadır.

3.1.1. Gebelik Diyabeti (Gestasyonel Diyabet)

Gebelik diyabeti, ilk kez gebelik sırasında ortaya çıkan bir çeşit diyabet şekli olarak belirtilmektedir. Özellik olarak tip II diyabete benzemekte ve gebelerin %2-4'ünde görülmektedir (Who, 1985).

3.1.2. Tip I Diyabet

Tip I diyabet hastalığının temel nedeni, pankreasın ürettiği insülin hormonunun yok olması ve insülin yetersizliği baş göstermesidir (Rother, 2007:1499; ADA, 2010:81). Beta hücrelerinin %80'i yok olduğunda, tip 1 diyabet ortaya çıkmaktadır. Vücutta bulunan beta hücreleri, şeker toleransını düşürmek için yeterli düzeyde değildir (Çınkır, 2011). Tip 1 diyabet genellikle 40 yaş altında insanlarda görülmektedir. Ayrıca bu diyabet türü, diyabetli hastaların %10'unu kapsamaktadır (Diabetes UK, 2009). Tip I diyabetin ortaya çıkmasında genel olarak genetik faktörlerin ve viral enfeksiyonların etkisi bulunmaktadır (Sperling, 2000).

3.1.3. Tip II Diyabet

Tip II diyabet, tip 1 diyabete göre daha yaygın olarak görülmektedir. Her diyabet hastasının %90-95'i tip II diyabet hastasıdır. Bu hastalık kendisini 30 yaş üzerinde göstermektedir. Sıklıkla da 50 ve 60'lı yaşlarda yavaş yavaş ortaya çıkmaktadır. İşte bu yüzden tip II diyabet hastalığına aynı zamanda yetişkin tipi diyabet de denilmektedir (Guyton ve Hall, 2001:951).

Tip II diyabet hastalığında, insülin salgılanmasında artış gözlenmektedir. Bu artış ilerleyen zamanlarda pankreası yormakta ve pankreas kan içerisinde yer alan şekeri düzenlemeye yarayan insülini salgılayamamaya başlamaktadır (Ward, vd. 1984:494; Guyton ve Hall, 2001:951). Tip II diyabet hastalığının oluşmasında rol oynayan

faktörler insülin direnci, hipertrigliseridemi, hipertansiyon, obezite ve kalıtsal faktörlerdir. İnsülin direncini hiperinsülinemi, obezite ve hipertansiyon takip etmektedir **(**Beck-Nielsen, vd. 1995: 23). İnsülin direncinin artmasıyla birlikte hepatik glikoz üretiminde artma, perifer dokular tarafından glikoz alınımında azalma meydana gelmekte ve bunun sonucunda kanda salgılanmış insülin miktarında artma meydana gelmektedir. Bu duruma hiperinsülinemi denilmektedir. Bu hiperinsülinemi durumu daha sonraki evrelerde pankreas beta hücrelerinin yıpranmasına yol açmakta ve beta hücreleri zarar görmektedir (Ward, vd. 1984:493; Groop, vd. 1989; Polonsky, vd. 1996).

40 yaş üzerinde olup aşağıdaki risk faktörlerinden bir ya da birkaçı bulunan kişiler diyabet açısından risklidir (Türkiye Halk Sağlığı Kurumu, 2014:15).

- Ailede şeker hastası olanlar
- Yüksek risk içeren etnik gruba ait olan
- Prediyabet
- Hipertansiyon
- HDL kolesterol <35 mg/dL ve trigliserid >250 mg/dL
- Kardiyovasküler hastalık
- Fazla kilolu veya obez
- Polikistik over sendromu (PCOS)
- Gestasyonel diyabet hikayesi
- 4 kilonun üzerinde bebek doğurma öyküsü
- İnsülin direnci ile ilişkili durumlar (akantozis nigrikans, non-alkolik steatohepatit)
- Şizofreni
- Bazı atipik antipsikotik ve antidepresan ilaçların kullanımı
- Fiziksel inaktivite
- Solid organ (özellikle böbrek) transplantasyonu yapılmış olan kişiler.

3.2. Veri Madenciliğinde Diyabet

Literatürde diyabet hastalığı ve veri madenciliği arasında yoğun bir ilişki olduğu görülmektedir. Diyabet hastalığı ile bağıntılı konularda veri madenciliği teknikleri kullanılmaktadır. Diyabette veri madenciliğinin kullanım amaçlarından bir tanesi

tahmin modeli oluşturarak, ileriye yönelik kestirim yapılmasıdır. Bu kestirimin doğruluğu ve tahmin yeteneği oldukça önemlidir. Bu amaçla araştırmalarda genel olarak veri madenciliği teknikleri arasında bir karşılaştırma yapılmaktadır. Diyabet için önemli bir veri madenciliği tekniği de veri ambarlarıdır. Veri ambarları sayesinde ve metin madenciliği sayesinde, diyabetle alakalı konular ortaya çıkarılarak haritalamalar yapılabilmektedir.

Zhaoli (vd., 2014) diyabet için uyarlanmış geleneksel Çin ilaçlarının saptanması amacıyla veri madenciliği tekniği uyguladığı araştırmasında, en sık kullanılan bitki ve yiyecekleri incelemiş ve diyet, tedavi gibi unsurların temellerini araştırmıştır. Veriler internet yardımıyla SinoMed'den elde edilen 227576 kaydı içermektedir. Bu verilere metin madenciliği uygulanarak sonuçlar ortaya koyulmuştur.

Nimmagadda (vd., 2011) sağlığı korumaya yönelik programlar dizayn etmek ve önermek amacıyla haritalama yapmak ve birçok veriyi gözden geçirmek amacıyla veri ambarı ve veri madenciliği tekniğini kullanarak araştırma yapmıştır. Bu kapsamda diyabet hastalarının verileri incelenmiş ve diyabet hastalarına yönelik sağlıklı beslenme çıkarımları yapılmıştır. Bu çalışmada esas olarak çok boyutlu veri ambarı yaklaşımı kullanılmıştır. Bu kapsamda literatür incelenmiş, 177 boyut belirlenmiş, 53 özellik saptanmış ve yiyecekle alakalı 632 veri incelenmiştir.

Sacchi (vd., 2015) veri madenciliği sayesinde tip II diyabet hastalarının durumunu kötüleştiren faktörleri bulmaya ve bu sayede hastalıkla mücadele etmeye yönelik araştırması kapsamında 953 diyabet hastasını ele almıştır. Klinik verileri elektronik sağlık kayıtlarından ve periyodik olarak kliniklere gelen hastaların laboratuvar sonuçlarından, hayat tarzlarına yönelik davranışlarından ve şikayet tanılarından oluşmaktadır. Çıkan sonuçlara göre hayat tarzını zorlaştıran psikolojik faktörler, fiziksel faktörler (alkol tüketimi veya sigara içimi gibi) belirlenerek yönetsel bir önlem alınması yönünde önerilerde bulunulmuştur.

Zhang (vd., 2015) ilaçları konumlandırmak amacıyla OMICS veri madenciliği yaparak haritalama yapmıştır. Bu kapsamda 16 GWAS makalesini, 17 proteomik çalışmayı ve 18 metabolik makaleyi inceleyerek veri madenciliği yapmış, neticesinde diyabetle ilişkili 115 gen, 56 protein ve 227 metabolik unsuru keşfetmiştir.

Yıldırım (vd., 2011) araştırmasında diyabet hastalarına yönelik dozaj belirlenmesinde veri madenciliği metotlarının kullanımını araştırmıştır. 89 tekil diyabet hastasının tıbbi kayıtları araştırma kapsamında ele alınarak, veri madenciliği tekniklerinden olan ANFIS ve Rough Set metodu kullanılmıştır. Sonuçlara göre bu iki tekniğin karşılaştırılmasında, ANFIS daha doğru sonucu vermiştir.

Dagliati (vd., 2018) araştırmasında tip II diyabet hastalarının tedavisinde öne çıkan faktörleri bulmak amacıyla veri madenciliği içerisinde careflow algoritma tekniğini kullanmıştır. 424 adet İtalyan diyabet hastasının verileri bu araştırma kapsamına dahil edilmiştir. Sonuç olarak hastaların benzerliklerini analiz etmede careflow'un etkinliği ortaya çıkmıştır.

Diyabet ile ilgili bir diğer çalışma da Suudi Arabistan'da yapılmıştır. Al-Nozha (vd., 2004) Suudi Arabistan'da diyabetin yaygınlığını ölçmek amacıyla veri madenciliği tekniğine başvurmuştur. Veriler rastgele seçilmiş 17232 adet ev halkına 5 yıl boyunca diyabetle bağıntılı anket soruları yöneltilerek oluşturulmuştur. Araştırma sonucunda Suudi Arabistan'ın diyabet profili yaş ve cinsiyetlere göre çıkartılmış ve gelecek için sayısal tahminlerde bulunulmuştur. Benzer bir çalışma olarak Aljumah (vd., 2013) araştırmasında genç ve yaşlı diyabet hastalarının sağlığının korunması doğrultusunda regresyona dayalı veri madenciliği tekniğini kullanarak modelleme yapmış ve modeli Oracle Data Miner (ODM) yazılımı ile çözdürmüştür. Analizde araştırmanın yapıldığı Suudi Arabistan için Dünya Sağlık Örgütü'nün (WHO) oluşturduğu risk faktörleri ele alınmıştır. Bu risk faktörlerinin farklı yaş gruplarına göre etkinliğini sıralamıştır.

Breault (vd., 2002) 30 binin üzerinde diyabet hastasının veri ambarlarını inceleyerek sorunlara yönelik çıkarımda bulunmuştur. Sınıflandırma ağacı metodu ile uygulama yapılmış ve hedef değer olarak HgbA1c> 9.5 ve 10 değişken: cinsiyet, yaş, acil bölümüne ziyaret sayısı, diyabet bölümünü ziyaret sayısı, diyabete ek hastalıklar indeksi, dislipidemia, hipertansiyon, kardiyovasküler hastalık, retinopati ve böbrek hastalıkları ele alınmıştır. Elde edilen sonuçlarda yaşlara göre sınıflandırmalar yapılarak, birliktelikler ortaya koyulmuştur.

Sigurdardottir (vd. 2007) çalışmasında tip II diyabetin önleyici eğitiminde glisemik kontrolün etkinliğini arttırmak amacıyla faktörlerin önem sırasını belirlemiştir. Bu faktörleri elde etmek amacıyla "önleyici eğitim" anahtar sözcüğü ile arama yapılarak, bu konu dahilinde yapılan ve 2001-2005 yılları arasında yayınlanan çalışmalar incelenerek, WEKA yardımıyla veri madenciliği yapılmıştır. Bu bağlamda, 464 başlık taratılmış ve 21 makale içerisinde yer alan 18 araştırma, incelemeye alınmıştır. Bu inceleme sonucunda HbA1c seviyesi önleyici eğitim açısından tek başına en önemli faktör olarak bulunmuştur.

Charpentier (vd., 2003) Fransa'da diyabet ve Tip II diyabete bağlı kardiyovasküler risk faktörlerinin kontrol edilmesi amacıyla 575 diyabet uzmanıyla etkileşim içerisine girerek, 2001 yılında 4930 adet hastanın bilgilerine yönelik veri madenciliği yapmıştır. Elde edilen sonuçlara göre kardiyovasküler riskin en üst seviyede olduğu bulunmuştur. Ayrıca glisemiya, kan basıncı ve LDL kolestrol uygun bir faktör olarak görülmemiştir.

Craig (vd., 2007) Asya'nın bir çok yerinde tip I diyabetin yaygınlaşmasına bağlı olarak, glisemik kontrol, diyabet bakımı ve gençlere komplikasyonu amacıyla tip I diyabet bağlamında gilisemik kontrol ve hipergilisemi ile bağıntılı faktörleri ele almıştır. 2312 adet genç ve yetişkinden oluşan bir çalışma grubu oluşturularak parmak ucundan alınan kana dayalı klinik çalışmaları yapılmıştır. Elde edilen sonuçlara göre HbA1c değerlerinin, yaş, cinsiyet, kilograma düşen insülin seviyesi, insülin rejimi, ülke menşei, diyabetin süresi ile yakından ilişkili olduğu istatistiki olarak elde edilmiştir. Diğer yandan kan glikoz değeri ölçümler çoklu regresyona tâbi tutulmuştur. Sonuç olarak glisemik kontrolü başaramayanların, mikrovasküler şikayetler noktasında yüksek risk gruplarını oluşturduğu bulunmuştur.

İspanya'da diyabet üzerine bir çalışma Soriguer-Escofet (vd., 2002) tarafından diyabetin ve otoantikorların yaygınlığını saptamak amacıyla yapılmıştır. Kesitsel bir çalışma olan bu araştırmada 1226 adet hasta araştırmaya dahil edilmiş ve hastaların kan testleri analize tâbi tutulmuştur. Elde edilen sonuçlara göre yetişkinlerde tip II diyabet ve yetişkin gizli otoimmünün güney İspanya'da yaygın olduğu veri madenciliği ile bulunmuştur.

Richards (vd., 2001) diyabet hastalarının klinik kayıtlarına ve ölüm sertifikalarına yönelik veri madenciliği çalışması yapmıştır. Araştırmanın amacı, hastaların ilk kontrole geldikleri zamanda elde edilen gözlemler ile erken ölümler arasında kurallar koymaya ve birliktelikler oluşturmaya yöneliktir. Buna göre ilk ziyarette elde edilen gözlemler ile erken ölüm arasında anlamlı bir birliktelik olduğu bulunmuştur. Ortaya koyulan 32 adet kuraldan 22 tanesi reddedilmiş, 10 tanesi kabul edilmiştir. Son olarak 10 adet kural sayısı 6'ya indirgenmiştir. Bunlar; "sol ayağın normal olmaması", "sol ayağın normal olmaması ve hafif dokunuş sonrası parlaklığın kaybolması"; "sol ayakta bozukluk", "sağ ayak bileği refleksinin olmaması", "sağ ayak bileği refleksinin olmaması ve sağ ayağın hafif dokunuş sonrası normal olmaması", "sağ ayak bileği ve sol ayak bileği refleksinin birlikte olmaması"dır.

Bhramaramba (vd., 2011) diyabetle ilişkili proteinlerin saptanması açısından genomik verilerden veri madenciliği yaparak diskriminant ve principal component analiz tekniklerini kullanmıştır. Bu araştırma fareler üzerinde uygulanmış ve veriye konu olan gen setleri farelerden elde edilmiştir. Araştırmanın sonucu olarak, insanların diyabet özelliklerinin bu iki tekniğe göre diğer canlılara göre farklılık gösterdiği bulunmuştur.

Pradhan ve Sahu (2011) veri madenciliği tekniklerinden olan Yapay Sinir Ağları ve Genetik Algoritma tekniklerini karşılaştırmışlardır. Araştırmada kullanılan veri setini PIMA Hindistan diyabet verileri oluşturmaktadır. Buna göre hamilelik sayısı, plazma glükoz yoğunluğu, kan basıncı, derinin kalınlığı, 2 saatlik serum insülin, vücut kütle endeksi, diyabet kalıtsallığı, yaş ve sınıflandırma verisi (o veya 1) oluşturmaktadır. Elde edilen sonuçlara göre en iyi sonucu Yapay sinir ağları vermiştir.

Bagdi ve Patil (2012) çalışmalarında veri madenciliği tekniklerinden olan ID3, C4.5 karar ağacı tekniklerini kıyaslamışlardır. Çıkan sonuçlara göre C4.5 karar ağacı tekniği en iyi sonucu vermiştir.

Rajesh ve Sangeetha (2012) diyabet tanısında veri madenciliğine yönelik araştırmalarında veri madenciliği tekniklerinden C4.5 sınıflandırma algoritmasını test etmişlerdir. PIMA Hindistan diyabet hastalarına ait verileri kullanarak 768 adet hasta bilgisini ve bunların her biri için 8 değişkeni araştırmasında kullanmıştır. Bu 8

değişken hamilelik sayısı, plazma glukoz yoğunluğu, diastolik kan basıncı, triseps deri katmanının kalınlığı, 2 saatlik serum insülin, vücut kütle endeksi, diyabet kalıtsallığı, yaş ve sınıflandırma verileridir (0 ve 1). Çıkan sonuçlara göre C4.5 algoritmasının doğruluk oranı %91 olarak belirlenmiştir.

Singh ve Kaur (2013) veri madenciliği tekniklerinden olan Sinir Ağları, Yapay Sinir Bulanık Karışım Sistemi, K-en yakın komşu (KNN), Genetik Algoritma, Arka Yayılma Algoritması gibi teknikleri kıyaslamışlardır. Veri setini PIMA Hindistan diyabet veri setinden oluşturmuştur. Elde edilen sonuçlara göre Sinir Ağları'nın doğruluğu en yüksek çıkmıştır.

Lakshmi ve Kumar (2013) veri madenciliği tekniklerini karşılaştırdıkları çalışmalarında, SVM, k-NN, C4.5, BLR, PNN, PLS-DA, MLR, k-means&Apriori ve PLS-LDA, tekniklerini ele almıştır. Veri setleri UCI ve Hindistan Sağlık Araştırma Merkezi'nden elde edilmiştir. Veri seti içerisinde vücut kitle endeksi, kilo, açlık şekeri ve kolestrol gibi unsurlar yer almaktadır. Çıkan sonuçlara göre PLS-LDA en iyi sonucu vermiştir.

Köklü ve Ünal (2013) veri madenciliği tekniklerini karşılaştırdığı çalışmasında Multilayer Perception, J48 ve Navie Bayes sınıflayıcılarını ele almıştır. Veri setlerini PIMA Hindistan diyabet verileri oluşturmaktadır. Çıkan sonuçlara göre Navie Bayes sınıflayıcısı en iyi sonucu vermiştir.

Motka (vd., 2013) araştırmasında diyabet tahminlemesinde farklı veri madenciliği tekniklerini kullanarak karşılaştırma yapmıştır. Bu teknikler Sinir ağları, Sinir ağlarıyla temel bileşen analizi (PCA), yapay sinir bulanık arayüz sistemleri (ANFIS) ve ANFIS+PCA'dır. Veri setleri PIMA Hindistan diyabet verilerinden elde edilmiştir. Elde edilen sonuçlara göre doğrulukta PCA'nın etkisinin büyük olduğu bulunmuştur. PCA ile birlikte kullanılan ANFIS en doğru sonucu vermiştir. Ancak en hızlı sonuç ise PCA ve sinir ağları ile alınmıştır.

Akay, (vd., 2013) araştırmasında sosyal medyadan yararlanarak, veri madenciliği ile diyabet ilaçlarının ve tedavilerin sonuçlarını izleme ve yanıtlama yaklaşımını incelemiştir. Forumlardan elde edilen sözcükler olumlu ve olumsuz sözcükler şeklinde sınıflandırılmıştır. Sözcük veri madenciliği şeklinde teknik kullanılan

araştırmada çıkan sonuçlar doğrultusunda iyileştirmeler ve kamu sağlığı için önerilerde bulunulmuştur.

Sankaranarayanan ve Perumal (2014) araştırmalarında veri madenciliği tekniklerinden olan Sınıflandırma Kuralı ve Karar Ağaçları tekniklerini ele almışlardır. Veri seti içerisinde 8 sürekli değişken ve 768 adet kişi bulunmaktadır. Hamilelik sayısı, plazma glukoz yoğunluğu, ağızdan şeker tolerans testi (GTT), triseps deri kalınlığı, 2 saatlik serum insülin, vücut kütle endeksi, diyabetin kalıtsallığı ve yaş sürekli değişkenleri oluşturmaktadır. Sonuçlara göre karar ağaçlarının sınıflandırma için kullanımının uygun olduğu bulunmuştur.

Kaur ve Chhabra (2014) çalışmalarında diyabete yönelik veri madenciliği tekniklerinden olan MLP, Naive Bayes, REP tree, Random Tree, RandomForest, J48 ve modifiye J48 ve RAD'ı ele alarak değerlendirmiştir. Sonuçlara göre modifiye J48 en iyi sonucu vermiştir. Çalışmada PIMA Hindistan veri seti kullanılmıştır.

Radha ve Srinivasan (2014) veri madenciliği teknikleri içerisinden C 4.5, SVM, k-NN, PNN ve BLR'yi ele alarak karşılaştırma yapmışlardır. Veri setleri olarak PIMA Hindistan diyabet veri setleri ve Hindistan Sağlık Araştırma Merkezi'nin veri setleri kullanılmıştır. Araştırma sonucunda en iyi sonucu veren teknik C4.5 bulunmuştur.

Veena ve Ravikumar (2014) veri madenciliği tekniklerinden olan EM, KNN, K-means, Amalgam KNN ve ANFIS algoritmalarını karşılaştırmışlardır. Veri setlerini PIMA Hindistan diyabet veri setleri oluşturmaktadır. Elde edilen sonuçlara göre Amalgam KNN ve ANFIS en iyi sonucu vermiştir.

Stranieri (vd., 2015) tip II diyabetin belirlenmesinde HbA1c'nin etkinliğini ölçmüştür. Veri seti Avusturya Üniversitesi'nin teşebbüsü ile Diyabet İzleme Komplikasyon Araştırma Girişimi'nden alınan 847 adet diyabet hastasına ait verilerden oluşmaktadır. Veriler içerisinde PG, HbA1c, kolestrol, düşük yoğunlukta lipoprotein, toplam kolestrol, yüksek yoğunlukta lipoprotein ve trigliserit değerleri bulunmaktadır. Araştırmanın sonucu olarak HbA1c'nin diyabeti en iyi tahmin eden değer olduğu bulunmuştur.

Mohammadi (vd., 2015) veri madenciliği teknikleri içerisinden Bayesian Network ve Karar Ağacı tekniklerini kullanmıştır. Veri setlerini diyabet hastalığına sahip

kadınlar oluşturmaktadır. Kriter olarak hamilelik, plazma glükoz yoğunluğu (GTT), kan basıncı, derinin yoğunluğu, insülin miktarı, vücut kitle endeksi, kalıtsallık, yaş ve diyabet sonucunun pozitif veya negatif olması oluşturmaktadır. Elde edilen sonuçlara göre Bayesian Network en doğru sonucu vermektedir.

Saravana Kumar (vd., 2015) çalışmasında diyabetin yaygınlığını, diyabete bağlı olarak oluşan hastalıkları ve bu hastalığa bağlı tedavi yöntemlerini tahmin etmek amacıyla Hadoop/Map reduce algoritması kullanmıştır. Çıkan sonuçlar doğrultusunda, diyabet hastalarına en iyi tedavi önerilerinde bulunulmuştur. Veriler, elektronik ortamlar, klinik verilerinden ve dışsal (hükümet verileri, laboratuvarlar, sigorta şirketlerinin verileri ve eczaneler gibi) kaynaklardan elde edilmiştir. Diğer yandan diyabete bağlı olarak kalp krizi, göz kusurları, felç ve karaciğer hastalıkları gibi bağıntılı hastalıklar da belirlenmiştir.

Vyas (vd., 2016) tarafından farklı canlı tiplerindeki diyabete bağlı protein tipleri arasındaki farklılığı ölçmek amacıyla veri madenciliği tekniklerini kullanarak yapılan araştırmada destek vektör makina, biyomedikal metin madenciliği ve ağ analizleri ele alınmıştır. Araştırmada 1296 farklı protein çeşidi incelenmiştir. Araştırmanın amacı incelenen proteinlerle, gelecekte oluşacak bağıntılı proteinler arasında ilişki kurarak tahminde bulunmaktır. Kurulan modelin %78.2 doğruluğa sahip olduğu görülmüş ve gelecek için diyabete yönelik kullanışlı bir model olduğu ortaya koyulmuştur.

Devi ve Shyla (2016) Hindistan PIMA'dan elde ettikleri 768 adet örnek üzerine veri madenciliği çalışması yapmışlar ve diyabete yönelik tahmin oluşturmuşlardır. Elde edilen veriler MLP, Navie Bayes, PLS-LDA, C4.5, Amalgam KNN, Bayesian Network, ANFIS, modifiye edilmiş J48, Homojenlik testi ve ANN veri madenciliği teknikleri kullanılarak analiz edilmiş ve bu tekniklerin doğrulukları araştırılmıştır. Veri setlerini ise plazma, kan basıncı, derinin kalınlığı, insülin, hamilelik, vücut indeksi, diyabet kalıtsallığı, yaş ve sınıflandırmalar (ölçüt 0,1) oluşturmaktadır. Çıkan sonuçlara göre J48 sınıflayıcı teknik en üst düzey doğruluk oranına sahiptir.

Perveen (vd., 2016) diyabeti öngörmede veri madenciliği sınıflandırma tekniklerinin performansını karşılaştırmıştır. Bilindiği üzere diyabeti belirlemek ve engellemek sağlık ekonomisi açısından önem arz etmektedir. Bu açıdan diyabeti tahmin

edebilmenin önemi ortaya çıkmaktadır. Bu araştırmada, diyabetin risk faktörleri karar ağaçları yardımıyla sınıflandırmaya tâbi tutulmuştur. Karşılaştırma sonucunda adaboost ensemble metodunun diğer metotlara göre daha iyi sonuçlar verdiği görülmüştür.

Harleen ve Bhambri (2016) veri madenciliği teknikleriyle diyabet tahmininde model etkinliği ölçmek amacıyla Navie Bayes ve J48 tekniklerini karşılaştırmışlardır. Veri setini PIMA Hindistan diyabet verileri oluşturmaktadır. Araştırma sonucuna göre Naive Bayes'in doğruluk düzeyi daha yüksek çıkmıştır.

Li (vd., 2018) Çin'de bulunan diyabet hastaları için en iyi tahminlemeyi yapabilecek veri madenciliği tekniğini saptamak adına araştırma yapmıştır. Veri madenciliği tekniklerimden olan destek vektör makine, karar ağacı ve bütünleşik öğrenme modeli (Adaboost ve Bagging) tekniklerini araştırma kapsamına almıştır. Veri seti 25 karakteristik değişkenden oluşmakta ve 2832 adet hastanın kayıtlarını içermektedir. Bu değişkenler, yaş, cinsiyet, etnik köken, eğitim, medeni durum, teşhis metodu, diyabete bağlı şikayetler, sigara kullanımı, alkol tüketimi, diyet programı, fiziksel aktiviteler, hipoglisemik ajanlar, insülin, sigarayı bırakma, limitli alkol tüketimi, takip metodu (telefon, evde ziyaret, klinik), psikolojik durum, takip eden sağlık uygulamaları, tamamlayıcı tedavi, kan basıncı, vücut kitle endeksi, açlık kan şekeridir. Sonuçlara göre en iyi doğruluk oranını Adaboost vermiştir.

Messan ve Zhai (2017) yüksek kan basıncı riskini veri madenciliği teknikleri ile tahmin etmeye yönelik bir çalışma gerçekleştirmişlerdir. Bu amaçla veri madenciliği tekniklerinin doğruluk oranları arasında karşılaştırma yapılmıştır. Veri madenciliği teknikleri GMM, SVM, Lojistik Regresyon, ELM ve yapay sinir ağlarıdır (ANN). Araştırma sonuçlarına göre yapay sinir ağları en yüksek doğruluk oranına sahiptir.

Maniruzzaman (vd., 2017) araştırmasında veri madenciliği tekniklerinin diyabet verilerinin sınıflandırılmasındaki performanslarını ölçmüştür. Veri madenciliği teknikleri arasından en sık kullanılan liner diskriminant analiz (LDA), quadratik diskriminant analiz (QDA) ve Naive Bayes (NB) teknikleri, Gaussian süreç (GP) temelli sınıflandırma tekniği ile karşılaştırılmıştır. PIMA Hindistan diyabet veri

setlerinden elde edilen 768 hastaya ait bilgiye uygulanan analizler sonucunda GP tekniği en iyi sonucu vermiştir.

Alic (vd., 2017) diyabet ve kardiyovasküler bozuklukların sınıflandırılmasında kullanılan makine öğrenmesi teknikleri arasında değerlendirme yapmaya yönelik araştırma kapsamında Yapay Sinir Ağları (YSA) ve Bayesian Ağları (BN) karşılaştırmıştır. Veriler diyabete yönelik araştırmaları kapsayan 20 adet seçilmiş makaleyi içermektedir. Bu 20 adet makale içerisinden 10 tanesi, istenilen hesaplamaları içerdiği için inceleme altına alınmıştır. Çıkan sonuçlara göre Bayesian Ağların diyabet için yüksek doğruluk değerine sahip olduğu ortaya koyulmuştur.

Vijayalakshmi ve Jenifer (2017) veri madenciliği teknikleri ile diyabet için risk faktörlerini analiz ettikleri çalışmalarında, birliktelik kuralı madenciliği ve karar ağacı şeklinde sınıflandırma ve kümeleme teknikleri kullanmışlardır. Araştırma kapsamında bakım evinden alınan 337 hastaya ait veriler analize tâbi tutulmuş ve önem arz eden faktör sıralamalarını içeren sonuçlar paylaşılmıştır. Diğer yandan veri madenciliği tekniklerinin bir karşılaştırması olarak en yüksek doğruluk oranını J48 vermiştir.

Kavakiotis (vd., 2017) araştırmasında, diyabetle ilgili önceki çalışmalarda kullanılan genetik veriler, klinik verileri ve elektronik olarak kaydedilmiş veriler arasında bir ilişki olup olmadığını incelemiştir. Bu açıdan geçmiş çalışmalarda kullanılan veri setleri arasında bir karşılaştırma yapmıştır. Elde ettiği sonuçlara göre, diyabetle ilgili yapılan çalışmalarda kullanılan veriler arasında anlamı bir benzerlik bulunduğu ortaya çıkmıştır.

Turnea ve Ilea (2018) tip II diyabeti tahmin etmek amacıyla veri madenciliği tekniklerini büyük veriye uyarlamışlardır. Çalışmada veri madenciliği tekniklerinden olan karar ağaçları, destek vektör makine (SVM), tümevarımsal öğrenme ve kümeleme teknikleri kullanılmıştır. Veriler klinik verilerinden, antropolojik ölçümlerden, kişisel ve aile hikayelerinden oluşmaktadır. Elde edilen sonuçlara göre sınıflandırma ağaçlarının etkinliği en yüksek çıkmıştır.

3.3. Diyabet Tanıları

Yapılan araştırmalar neticesinde yaş, cinsiyet, hipertansiyon, hiperlipidemi, menopoz, HbA1c, kreatin, toplam kolestrol, LDL, HDL, kırgınlık ve yorgunluk, metformin, gastro özofajial reflü hastalığı, eklem ağrısı, demir eksikliği anemisi, vitamin B12 eksikliği anemisi, aterosklerotik kardiyovasküler hastalık, serebrovasküler hastalıklar, osteoporoz, diyabetik polinöropati, insülin bağımlı olan (tip 1) diyabet ve insülin bağımlı olmayan (tip 2) diyabet arasında ilişki olduğu bilinmektedir.

3.3.1. Yaş

Kişilerde yaşlanma ile birlikte biyolojik ve fizyolojik değişiklikler yaşanmaktadır. Bu yaşanan değişikliklere bağlı olarak belirli sınıflandırmalar yapılmaktadır. Bilimsel olarak belli başlı sınıflandırma çeşitleri araştırmalara konu olmuştur. En kabul gören sınıflandırma Kumar'ın (vd. 2013) yüz görüntüsüne bağlı olarak ortaya koyduğu tahminleme modelinden yola çıkarak oluşturulan yaş aralığıdır. Bu sınıflandırma tahmin modeline göre çocukluk 0-15 yaş, genç yetişkinlik 16-31 yaş, orta düzey yetişkinlik 31-60 yaş ve yaşlı yetişkinlik 60 üstü yaş olarak belirlenmiştir.

3.3.2. Cinsiyet

Cinsiyet ve cinsiyet farklılıklarının birçok hastalıkta, epidemiyoloji, patofizyoloji, tedavi ve sonuçlarda önemli olduğuna dair kanıtlar artmaktadır. Özellikle diyabet gibi bulaşıcı olmayan hastalıklar için cinsiyet farklılıklarının araştırılmasına yönelik çalışmaların önemi artmaktadır. Cinsiyet farklılıkları, cinsiyet kromozomlarındaki farklılıklar, otozomların cinsiyete özgü gen ekspresyonu, cinsiyet hormonları ve bu hormonların organ sistemleri üzerindeki etkileri ile ortaya çıkan farklar kadınlar ve erkekler arasındaki biyolojiye bağlı farklılıklar olarak tanımlanmaktadır. Hem biyolojik hem de psikososyal faktörlere bağlı olarak cinsiyet ve cinsiyet farklılıkları diyabet riskinde ve diyabete bağlı sonuçlarda etkilidirler (Kautzky-Willer, 2016:279).

3.3.3. Hipertansiyon

Hipertansiyon, yüksek tansiyon için başka bir ismidir. Hipertansiyon, şiddetli komplikasyonlara yol açabilmekte ve kalp hastalığı, felç ve ölüm riskini artırmaktadır. Normal kan basıncı değerleri diastolik 80 mmHg ve sistolik 120 mmHg iken; hipertansiyonda diastolik 80 mmHg ve sistolik 130 mmHg'nın üzerindedir (MacGill, 2017).

Sıklıkla diabetes mellitus ile ilişkili hipertansiyon, diyabetik komplikasyonların şiddetinde ve ilerlemesinde önemli bir rol oynamaktadır. Hipertansiyonun diyabetik nefropatinin seyri üzerindeki sıklığı ve etkisi ile persistan mikroalbüminürinin ya gelecekteki açık diyabetik nefropatinin bir göstergesi olarak ya da yeni başlayan hipertansiyonun bir belirteci olarak rolü gözden geçirilmektedir (Rosenstock ve Raskin, 1988).

3.3.4. Hiperlipidemi

Lipidler organizmada önemli fonksiyonlarda görev aldıkları ve hücre membranlarının yapı taşı olduklarından, metbolizmalarının doğru işleyişi oldukça önem kazanmaktadır. Hiperlipidemi lipid metabolizmasının primer bozukluğu şeklinde veya sekonder bozukluklara bağlı olarak görülebilmektedir. Hiperlipideminin birincil nedenleri tek başına hiperkolesterolemi ve hipertrigliseridemi veya hiperkolesterolemi+hipertrigliseridemi kombinasyonu ve HDL kolesterol düşüklüğü şeklinde olabilmektedir. İkincil nedenleri arasında ise diyabet, böbrek sendromu, hipotroidizm, alkolizm, kronik karaciğer hastalığı (obstruktif), protein yapı bozuklukları ve bazı ilaçlarla uzun süren ilaç tedavileri (oral kontraseptifler, tiazid diüretikler ve glukokortikoidler) sayılabilmektedir (Rağbetli, 2009:43).

3.3.5. Menopoz

Menopoz, menstrüasyon dönemi durduktan ve östrojen seviyesi düştükten sonraki yaşam aşaması olarak adlandırılmaktadır. Bazı kadınlarda, menopoz, yumurtalıkların diğer tıbbi nedenlerden dolayı çıkarıldığında, ameliyat sonucu olarak ortaya çıkabilmektedir. Tip 2 diyabet de diğer metabolizma bozuklukları gibi kendisini orta

yaşlarda göstermektedir. Özellikle bu dönem, kadınların menopoz evresinde görülmektedir. Menopozal geçiş, endojen steroid hormonları, vücut kompozisyonu ve vücut yağ dağılımı ve lipit ve metabolik profiller dahil olmak üzere fizyolojik özelliklerde hızlı bir değişim yaşandığı bir evredir. Bu temel değişiklikler, menopoz ve diyabet arasında mekanik bir ilişki yaratmaktadır (Karvonen-Gutierrez, vd. 2016: 1).

3.3.6. Glycated Haemoglobin (HbA1c)

HbA1c, teşhis için uygulanabilirliği konusunda bazı tartışmalar devam etmesine rağmen, halen birçok ülkede resmi olarak (tip 2) diyabet için bir tanı testi olarak kullanılmaktadır. HbA1c'nin sınır değeri ≥48 mmol/mol (≥% 6,5) olarak belirlemiştir (ADA, 2010). HbA1c, sıkı kalite güvence testlerinin yapılmasını ve analizlerin uluslararası referans değerlerine göre hizalanmış kriterlere göre standartlaştırılmasını sağlayan diyabet için bir diyagnostik test olarak kullanılabilmekte ve doğru ölçümünü engelleyen herhangi bir koşul bulunmamaktadır. Diyabeti teşhis etmek için kesme noktası olarak % 6.5'lik bir HbA1c önerilmektedir. % 6,5'ten küçük bir değer, glikoz testleri kullanılarak teşhis edilen diyabeti devre dışı bırakmamaktadır (WHO, 2011:3).

3.3.7. Kreatinin

Kas metabolizmasından kimyasal atık molekülleri üretilmektedir. Keratinin, kaslarda enerji üretimi için büyük öneme sahip bir molekül olan kreatinden üretilmektedir. Yaklaşık olarak vücuttaki kreatinin %2'si bir gün içerisinde kreatinine dönüşmektedir.

Kreatinin kan dolaşımı yoluyla böbreklere taşınmaktadır. Böbrekler kreatininin çoğunu filtrelemekte ve idrardan atmaktadır. Bir atık olmasına rağmen, kreatinin önemli bir teşhis işlevi görür. Kreatinin, böbrek fonksiyonunun oldukça güvenilir bir göstergesi olduğu bulunmuştur. Böbrekler bozulmuş hale geldikçe, kreatinin yükselecektir. Anormal derecede yüksek kreatinin seviyeleri böylelikle, bir hasta herhangi bir semptom bildirmeden önce bile, olası bir arıza veya böbrek yetmezliği konusunda uyarmaktadır. Bu nedenle, standart kan ve idrar testleri, kandaki kreatinin miktarını düzenli olarak kontrol etmektedir. Kandaki normal kreatinin seviyeleri

yetişkin erkeklerde desilitre başına 0.6 ila 1.2 miligram (mg) ve erişkin kadınlarda desilitre başına 0.5 ila 1.1 miligramdır (Metrik sistemde, bir miligram bir gramın binde birine eşit bir ağırlık birimidir ve bir desilitre litrenin onda birine eşit bir hacim birimidir). Genel popülasyona göre, kaslı gençlerde veya orta yaşlı yetişkinlerde kandaki kreatinin daha fazla olabilmektedir. Öte yandan yaşlı insanlar, kanlarında belirlenen normdan daha az kreatinine sahip olabilmektedir. Bebekler, kas gelişimlerine bağlı olarak kreatinin miktarı yaklaşık 0.2 mg veya daha fazla kreatinine sahiptirler (MedicineNet, 2016).

3.3.8. Toplam Kolestrol

Toplam kolestrol kanda desilitre başına miligram olarak ölçülerek ifade edilmektedir. Bir miligram, bir gramın binde birine eşittir. Bir desilitre, litrenin onda birine eşittir. Toplam kolestrolde istenen seviyeler 200 mg/dL'nin altındadır. Sınır çizgisi yüksek seviyeleri 200-239 mg/dL'dir. Yüksek seviyeler ise 240 mg/dL ve yukarısıdır (Ehrlich ve Schroeder, 2009).

3.3.9. High-Density Lipoprotein (HDL)

(HDL) iyi kolesterol olarak adlandırılmaktadır. Çünkü HDL, işlenmemiş kolesterolü işlemek için karaciğere geri taşımaktadır. HDL plak oluşumuna katkıda bulunmamaktadır. HDL için kötü seviyeler 40 mg/dL'nin altındadır. Daha iyi seviyeler 40 ila 59 mg/dL arasındadır. En iyi seviyeler 60 mg / dL ve üstündedir (Ehrlich ve Schroeder, 2009).

3.3.10. Low-Density Lipoprotein (LDL)

LDL, kötü kolesterol olarak adlandırılmaktadır. Çünkü fazla miktarda LDL, arterlerde plak birikmesine sebep vermektedir. LDL için optimal seviyeler 100 mg/dL'nin altındadır. Optimum seviye 100 ila 129 mg/dL arasındadır. Yüksek sınır çizgisi 130 ila 159 mg/dL arasındadır. Yüksek ise 160 ila 189 mg/dL arasındadır. Çok yüksek 190 mg/dL ve üstü seviyelerdedir (Ehrlich ve Schroeder, 2009).

3.3.11. Kırgınlık ve Yorgunluk

Yorgunluk, günlük konuşmalarda yaygın olarak kullanılan ve öznel anlamları olan bir kelimedir. Yorgunluk, uyku hali, yorgunluk, enerji eksikliği ve bitkinlik gibi terimler birbirinin yerine kullanılabilmektedir. Bilimsel literatürde, yorgunluk tanımları oldukça fazladır. Yorgunluk tanımlarında, yorgunluk belirtileri ve etkileri arasındaki ilişkiler ele alınmaktadır. Ancak yorgunluk için temel olarak kullanılan tanım, beyni ve nörolojik sistemleri etkileyen fizyolojik kombinasyonlar şeklindedir. Yorgunluk diyabet hastalarının en büyük şikayetleri arasında yer almaktadır. Kandaki glikoz miktarında yaşanan değişimler gibi diyabete bağlı semptomlar, yorgunluğu tetikleyen unsurlar arasında yer almaktadır (Fritschi ve Quinn, 2010).

3.3.12. Metformin

Tip 2 diyabetin yönetiminde glisemik ve kardiyovasküler risk faktörü hedeflerine ulaşmak için agresif tedaviler gerekmektedir. Bu bağlamda, eski ve geniş çapta kabul görmüş bir madde olan metformin, sadece antihiperglisemik özellikleriyle değil, aynı zamanda glisemik kontrolün ötesinde endotel disfonksiyonu, hemostaz ve oksidatif stres, insülin direnci, lipit profilleri ve yağların yeniden dağılımı gibi etkilerinden dolayı da önemli bir maddedir. Bu özellikler bakımından, metformin sadece antihiperglisemik etkiler için değil; ayrıca kardiyovasküler etkilerin azaltılmasında da katkıda bulunabilir. Metforminin keşfi, yüzyıllardır diyabet tedavisi için Avrupa'da geleneksel olarak ilaç niteliğinde kullanılan bir bitki olan Gallega officinalis'den elde edilen galejine benzer bileşiklerin sentezi ile başlamıştır. Metformin kardiyovasküler ve metobolik etkileri bakımından etkin ve güvenilir bir glikoz düşürücü ilaç olarak tercih edilmektedir (Rojas ve Gomes, 2012).

3.3.13. İnsülin Bağımlı Diabetes Mellitüs (Tip 1)

İnsülin bağımlı diyabet ya da bilinen ismi ile tip 1 diyabet, pankreasta ilerleyen beta-hücre yıkımına yol açan bir dizi olay sonucu insüline bağımlı diyabetin ortaya çıktığı diyabet tipidir. Genellikle otoimmun kaynaklı olarak gelişen hastalığın, çoğunlukla çocukluk çağı ve genç erişkin yaşlarda ortaya çıktığı bilinmektedir (Durna, 2002:11).

3.3.14. Gastro Özofajial Reflü Hastalığı

Gastro özofajial reflü, alınan gıdaların ve/veya mide asidinin yemek borusuna geri kaçmasıdır. Bu kaçışın özellikle yatar pozisyonda ve yemeklerden hemen sonra mide içi basıncın artmasıyla günde 10 defaya kadar olması ve dört dakikadan kısa sürmesi normal kabul edilebilir. Fakat günde 10 defadan fazla ve dört dakikadan uzun olursa veya bu kaçışa bağlı yemek borusunda hasar meydana gelirse buna gastro özofajial reflü hastalığı (GÖRH) denir. Toplumun yaklaşık %20'sinde görülen bir klinik tablo olup kalp kökenli olmayan göğüs ağrılarının en sık nedenidir (Yaman ve Keskin, 2018). Diyabet neticesinde tükürük salgısının azalması gibi durumlardan dolayı gastro özofajial reflü hastalığı tetiklenebilmektedir. Bu durum nedeniyle diyabet ve gastro özofajial reflü hastalığı arasında ilişki bulunmaktadır (Akyüz ve Soyer, 2017:544)

3.3.15. Eklem Ağrısı

Diyabet eklemlere zarar verebilmektedir. Diyabetik artropati denilen bu durumda, ani travmanın neden olduğu ağrılardan farklı olarak, eklemlerde ağrılar zamanla gerçekleşir. Diğer yandan diyabetin ayaklarda değişim, derinin kalınlaşması, omuz renginde değişme, karpal tunnel sendromu gibi diğer etkileri de vardır. Eklem iki kemiğin birleştiği yer olarak ifade edilmektedir. Diyabetin eklemlere verdiği çeşitli zararlar mevcuttur. Özellikle diyabetin sinirleri etkilemesi ile birlikte eklemlerdeki sinirlerde hassasiyet ve ağrı oluşabilmektedir. Ayrıca aşırı kilo almayla birlikte hastaların eklemlerinde ağrılar oluşabilmektedir. Eklem ağrılarının diğer bir nedeni de diyabet ile birlikte bağışıklık sisteminin çökmesidir (Healthline, 2018).

3.3.16. Demir Eksikliği Anemileri

Demir eksikliği dünyada görülen en yaygın anemi şekillerindendir. Demir eksikliğine bağlı olarak kişilerde hemoglobin sentezlenememekte ve kanda glikozillenmiş hemoglobin (HbA1c) değerleri artmaktadır (Bayrak vd. 2009). Diyabete bağlı olarak böbreklerin işlevi zayıfladıkça eritropoetin üretimi de azalmaya başlamaktadır. Eritropoetinin azalmasına bağlı olarak demir eksiklikleri oluşmakta ve anemiler ortaya çıkmaktadır (Wittwer, 2016).

3.3.17. Vitamin B12 Eksikliği Anemileri

B12 vitamini protein hücrelerinin oluşmasında önemli bir vitamindir. B12 vitamini ışığa bağlı olarak proteinlerde düzenleme sağlamaktadır. Ayrıca transkripsiyon ve sentez gibi süreçlerde de rol oynamaktadır. DNA oluşturma sürecinde etkili olan vitamin B12'nin eksikliği vitamin B12 eksikliği anemisi olarak adlandırılmaktadır (Romine, vd. 2017:E1205). Diyabet hastalarında B12 vitamini eksiklikleri sıklıkla görülebilmektedir. Özellikle metformin tedavisi gören diyabet hastalarında, vitamin B12 seviyeleri düşmektedir (Pflipsen, vd. 2009:528).

3.3.18. Aterosklerotik Kardiyovasküler Hastalık

Aterosklerotik kardiyovasküler hastalıkların, diyabet hastalarında sakatlık riskine neden olan ya da ölümle sonuçlanan hastalık türlerinden olduğu bilinmektedir. Diyabet hastalarında aterosklerotik kardiyovasküler hastalık görülme riski, diyabet hastası olmayan kişilere göre yaklaşık 14 yıl daha erken görülme riski taşımaktadır. Koroner arter hastalığı en yaygın görünen aterosklerotik kardiyovasküler hastalıklardandır. Epidemiyolojik çalışmalar koroner arter ile diyabet arasında sıkı bir ilişki olduğunu ortaya koymuştur (Roşu vd. 2018:182).

3.3.19. İnsülin Bağımlı Olmayan Diabetes Mellitüs (Tip 2)

Tip 2 diyabet, insülin direnci bozukluğudur. Vücutta insülin üretilmekte; ancak etkili kullanılamamaktadır. Bu durumu telafi etmek adına pankreas daha fazla insülin hormonu salgılamaktadır. Bir süre sonra insüline karşı duyarsızlık ve hormonal bozukluklar meydana gelmektedir (Ehrlich ve Schoeder, 2009:389).

3.3.20. Serebrovasküler Hastalıklar

Serebrovasküler hastalık beyindeki kan damarlarını ve de serebral dolaşımı etkileyen bir çeşit tıbbi durum olarak ifade edilmektedir. Diyabet hastalarında, şeker nedeniyle diyabet serebrovasküler hastalıklar oluşabilmektedir. Diyabet hastalarının yaklaşık %20-%40'ı serebrovasküler hastalıktan muzdariptir. Serebrovasküler hastalığın tip 2 diyabet hastalarında esas tetikleyicisi aterosklerotik kardiyovasküler hastalıklardır. Diyabetli kişilerde serebrovasküler hastalıklara bağlı olarak beyin hasarları

oluşabilmektedir. 30-44 yaş aralığı felç olma riskini en fazla yaşayan yaştır. (Zhou, Zhang ve Lu, 2014).

3.3.21. Osteoporoz

Diabetes mellitus (özellikle tip 2) ve osteoporoz birbiri ile bağıntılı iki yaygın rahatsızlıktır. Bu iki hastalığın yaygınlığı ise artmaktadır. Tip 1 diabetes mellituslu ergenler, potansiyel pik kemik kütlesine erişemeyebilir ve bu da onları daha fazla kırık riskine sokabilmektedir. Tip 2 diyabeti olan erişkinlerde ise kırık riski artmaktadır (Seeland, vd. 2013, 411).

3.3.22. Diyabetik Polinöropati

Diyabetik nöropati, her organı etkileme potansiyeline sahip heterojen bir grup patolojidir. Diyabetik polinöropati organ yetersizliği gibi klinik sonuçlara yol açmakta ve bu durum da düşük hayat kalitesine ve artan hastalığa (morbiditeye) yol açmaktadır. Diyabetik polinöropati, pozitif ve negatif semptomlarla periferal sinir disfonksiyonu olarak tanımlanmaktadır. Diyabetik polinöropatinin risk faktörleri arasında yaş, erkek cinsiyet, diyabetin süresi, kontrolsüz glisemi, boy, aşırı kilo/obezite ve insülin tedavisini bulunmaktadır (Román-Pintos, et al. 2016:1).

DÖRDÜNCÜ BÖLÜM

UYGULAMA: BİLECİK DEVLET HASTANESİ DİYABETİK POLİNÖROPATİ HASTALARINA İLİŞKİN VERİ ANALİTİĞİ

4.1. Problemin Tanımlanması

Diyabet şüphesi ile dahiliye polikliniğine gelen hastaların oluşturduğu veri seti üzerinde, veri madenciliği yöntemleri kullanılarak, bir kişinin diyabetik polinöropati hastalığının olup olmadığını öngörebilmek bu çalışmanın ana amacıdır.

4.2. Veri Setini Anlama

Kullanılan veri seti Bilecik Devlet Hastanesinden temin edilmiştir. Bilecik İl Sağlık müdürlüğüne bilimsel araştırma izni başvurusu yapılmış. İl Sağlık Müdürlüğünün belirlediği gerekli koşullar sağlanmış, üniversiteden onay alınmış ve süreç sonunda iznin çıkmasıyla çalışmaya başlanmıştır. Bu veri setinde hastalara ait kimlik bilgileri mevcut değildir. Veri seti değişkenleri uzman doktor görüşü alınarak birlikte belirlenmiştir. Değişkenler belirlenirken hastaların anamnez yani tıbbi hikayeleri tek tek okunmuştur. Hasta kayıtlarında bulunan şikayet, tanı, karar ve kullanılan ilaçlar incelenmiştir. Sonunda hastanın nitelikleri (yaş, cinsiyet), laboratuvar sonuçları ve hastalık tanılarından oluşan 22 değişken belirlenmiştir.

Veri setinde 5 adet nümerik değer ile tanımlanmış laboratuvar sonucu bulunmaktadır. Bunlar HbA1c, Kreatinin, Total kolesterol, HDL kolesterol ve LDL kolesteroldür. Geri kalan değişkenler yaş değişkeni hariç kategorik değişkenlerdir. Cinsiyet değişkeninde erkek 1 ile, kadın 2 ile gösterilmiştir. Diğer değikenlerde ki 0 değeri o hastalığın yokluğunu, 1 değeri o hastalığın varlığını tanımlanmaktadır.

Tablo 8: Veri setinde bulunan niteliklere ait özellikler

	Tahmin İçin Kullanılan Verinin Yapısı		
	Değişken	Veri tipi	Veri setinde gösterimi
1	Yaş	Nümerik	
2	Cinsiyet	Kategorik	1=erkek 2=kadın
3	Hipertansiyon	Kategorik	0=yok 1=var
4	Hiperlipidemi	Kategorik	0=yok 1=var
5	Menopoz	Kategorik	0=yok 1=var
6	HbA1c	Nümerik	
7	Kreatinin	Nümerik	
8	Total Kolesterol	Nümerik	
9	HDL Kolesterol	Nümerik	
10	LDL Kolesterol	Nümerik	
11	Kırgınlık ve Yorgunluk	Kategorik	0=yok 1=var
12	Metformin	Kategorik	0=yok 1=var
13	İnsülin Bağımlı Olmayan Diabetes Mellitüs	Kategorik	0=yok 1=var
14	Gastro Özofajial Reflü Hastalığı	Kategorik	0=yok 1=var
15	Eklem Ağrısı	Kategorik	0=yok

			1=var
16	Demir Eksikliği Anemileri	Kategorik	0=yok 1=var
17	Vitamin B12 Eksikliği Anemisi	Kategorik	0=yok 1=var
18	Aterosklerotik Kardiyovasküler Hastalık	Kategorik	0=yok 1=var
19	İnsülin Bağımlı Diabetes Mellitüs	Kategorik	0=yok 1=var
20	Serebrovasküler Hastalıklar	Kategorik	0=yok 1=var
21	Osteoporoz	Kategorik	0=yok 1=var
	Hedef Nitelik		
22	Diyabetik Polinöropati	Kategorik	0=yok 1=var

4.3. Veriyi Hazırlama

Veri seti içerisinde bulunan "boy", "kilo" ve "üre" değişkenleri, çok fazla eksik veri girilmesinden dolayı çıkartılmıştır. Yine veri setinde nümerik bir değişkenin altında "." ile tanımlanan ya da anlamsız sayısal ifade girilen (örneğin -0.0023456) değerler çıkartılmıştır. Total kolesterol değişkeni altında bazı değerler "trigliserit hesaplanamaz" ya da sadece "hesaplanamaz" şeklinde girilmiştir. Bu değerlerde analizden çıkarılmıştır

Veri setinde bulunan uç noktalar incelenmiş, tekrar eden değerler çıkarılmıştır.

Veri seti temizlendikten sonra analiz için hazırlanmaktadır.

4.4. Analize Hazırlık

Bundan sonraki aşamalar RStudio'da yapılmıştır. Uygulama kodları eklerdedir.

Veri seti 7409 gözlem ve 22 değişkenden oluşmaktadır.

Değişkenler sırasıyla: Yaş, Cinsiyet, Hipertansiyon, Hiperlipidemi, Menopoz, HbA1c, Kreatinin, Total Kolesterol, HDL, LDL, Kırgınlık ve Yorgunluk, Metformin, İnsülin Bağımlı Olmayan Diabetes Mellitüs , Gastro Özofajial Reflü Hastalığı , Eklem Ağrısı, Demir Eksikliği Anemileri, Vitamin B12 Eksikliği Anemisi, Aterosklerotik Kardiyovasküler Hastalık, İnsülin Bağımlı Diyaetes Mellitüs, Serebrovasküler Hastalıklar, Osteoporoz ve hedef nitelik olan Diyabetik Polinöropati'dir.

Öncelikle veri setinin yapısı incelenmiş, nümerik ve faktör şeklinde düzenlenmiştir. Nümerik değişkenler nümerik olarak, kategorik değişkenlerde faktör şeklinde tanımlanmıştır. Düzenlendikten sonra şu hale dönüşmüştür.

'data.frame': 7409 obs. of 22 variables:

$ Yas : num 59 65 61 58 34 73 69 65 72 66 ...
$ Cinsiyet : Factor w/ 2 levels "1","2": 2 1 2 2 1 1 1
$ Hipertansiyon : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ Hiperlipidemi : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1
$ Menapoz : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ HbA1c : num 8 7 6.1 7.5 6.8 8.15 8.5 8.9 7 6.7
$ Kreatinin : num 0.8 1.37 0.74 0.72 0.87 1.02 0.95
$ Total.Kolesterol : num 234 131 203 206 168 168 184
$ HDL : num 50 39 64 56 40 61 44 45 41 56
$ LDL : num 153.8 77.2 121.8 111.4 96.8 ...
$ Kırgınlık.ve.Yorgunluk : Factor w/ 2 levels "0","1": 1 1 1 1 1
$ Metformin : Factor w/ 2 levels "0","1": 1 1 2 1 1
$ Insulin.Bagimli.Olmayan.Diyabetes.Mellitus : Factor w/ 2 levels "0","1": 1 1 1 1 1
$ Gastro.Ozofajial.Reflu.Hastalıgı : Factor w/ 2 levels "0","1": 1 1 1 1 1
$ Eklem.Agrisi : Factor w/ 2 levels "0","1": 1 1 1 1 1 1

```
$ Demir.Eksikligi.Anemileri                    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ Vitamin.B12.Eksikligi.Anemisi                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ Aterosklerotik.Kardiyovaskuler.Hastalık      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1
$ Insulin.Bagimli.Diyabetes.Mellitus           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ Serebrovaskuler.Hastalıklar                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ Osteoporoz                                   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
$ Diyabetik.Polinoropati                       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1
```

Veri setinin özeti Tablo 9'da verilmiştir. Bu tabloda kategorik değişkenler ve nümerik değişkenlerin minimum değerleri, 1. kartil, medyan, ortalama, 3. kartil ve maksimum değerleri görülür.

Tablo 9: Veri seti özeti 1

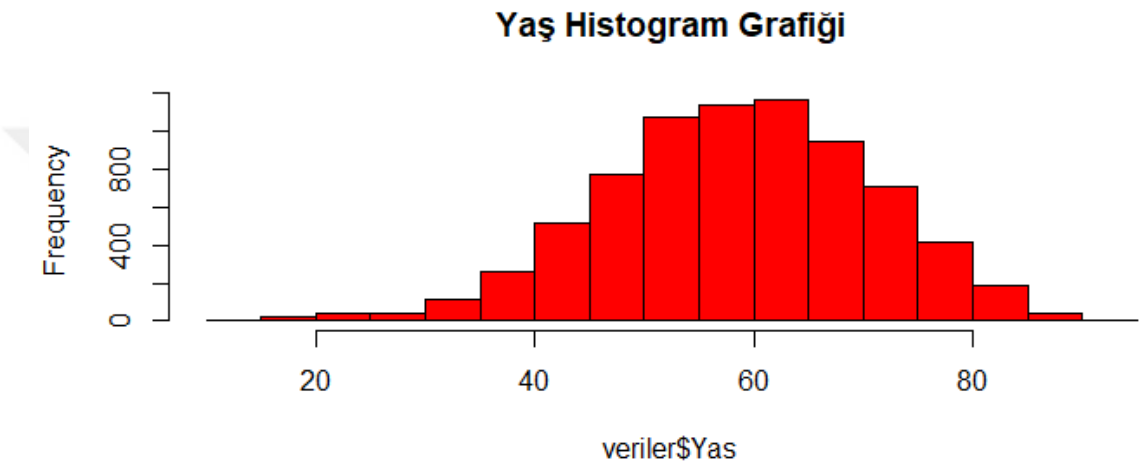
Yaş	Cinsiyet	Hipertansiyon	Hiperlipidemi	Menopoz	HbA1c	Kreatinin	Total Kolesterol	HDL Kolesterol	LDL Kolesterol	Kırgınlık ve Yorgunluk
Min. :14.00000	Erkek :2708	0:5455	0:5644	0:7250	Min. : 4.600000	Min. : 0.320000	Min. : 64.000	Min. : 15.00000	Min. : 15.8000	0:6950
1st Qu.:51.00000	Kadın :4701	1:1954	1:1765	1: 159	1st Qu.: 6.400000	1st Qu.: 0.760000	1st Qu.:170.000	1st Qu.: 40.00000	1st Qu.: 93.4000	1: 459
Median :59.00000					Median : 7.500000	Median : 0.890000	Median :199.000	Median : 47.00000	Median :116.8000	
Mean :58.99137					Mean : 7.929127	Mean : 0.931949	Mean :202.973	Mean : 48.52378	Mean :119.4857	
3rd Qu.:67.00000					3rd Qu.: 8.900000	3rd Qu.: 1.030000	3rd Qu.:232.000	3rd Qu.: 56.00000	3rd Qu.:143.0000	
Max. :93.00000					Max. : 17.500000	Max. : 7.480000	Max. :622.000	Max. :160.00000	Max. :339.0000	

(devamı takip eden sayfada)

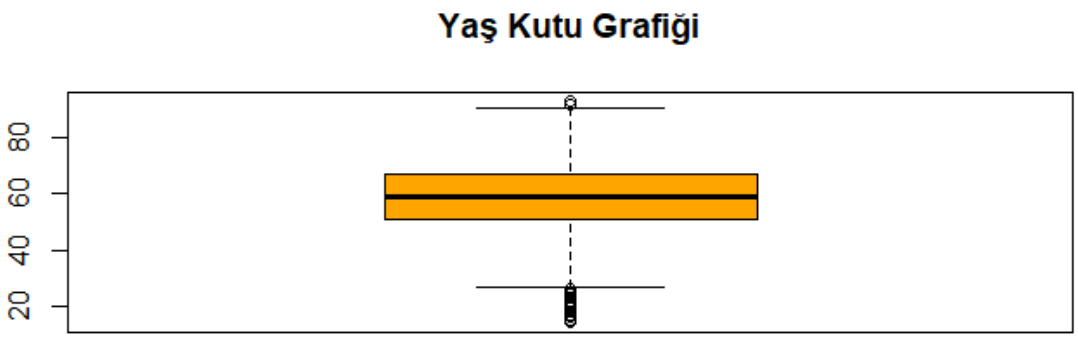
Metformin	İnsülin Bağımlı Olmayan Diabetes Mellitüs	Gastro Özofajial Reflü Hastalığı	Eklem Ağrısı	Demir Eksikliği Anemileri	Vitamin B12 Eksikliği Anemisi	Aterosklerotik Kardiyovasküler Hastalık	İnsülin Bağımlı Diabetes Mellitüs	Serobrovasküler Hastalık	Osteoporoz	Diyabetik Polinöropati
0:4699	0:4506	0:7133	0:7339	0:7362	0:7084	0:7249	0:7358	0:7348	0:7324	var: 285
1:2710	1:2903	1: 276	1: 70	1: 47	1: 325	1: 160	1: 51	1: 61	1: 85	yok:7124

Veri setindeki değişkenler tek tek incelenmiştir. Bunun için her birine uygun grafikler çizilmiştir. Nümerik değişkenler için histogram grafikleri, kategorik değişkenler için çubuk grafikleri çizilmiştir. Ayrıca değişkenler kutu grafikleri ile de gösterilmiştir. Böylece dağılımları hakkında daha kolay bilgi edinilmiştir.

Şekil 6: Yaş değişkeni grafikleri

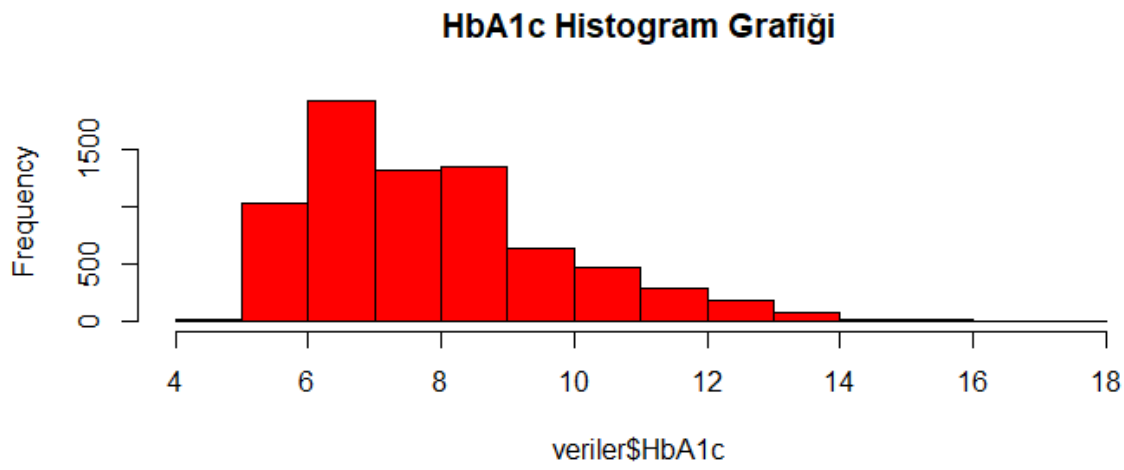


Şekil 7: Yaş kutu grafiği

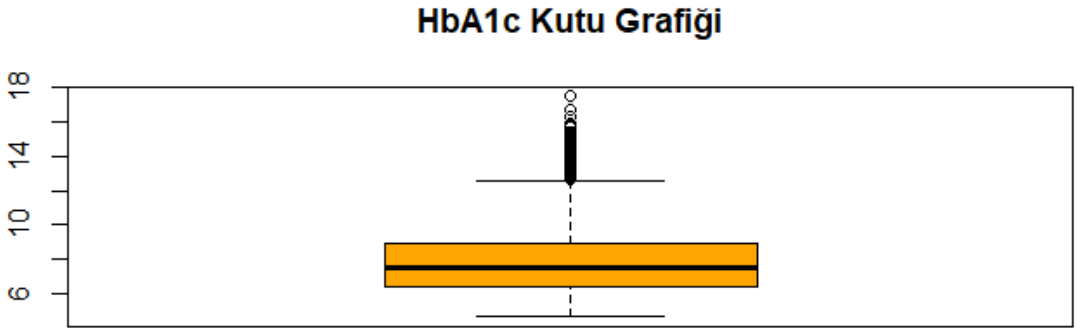


Veri setinde yaş değişim aralığı en küçük 14 yaşındaki hasta ile en büyük 93 yaşındaki hasta arasındadır. Frekanslarına bakıldığında 60 yaş civarı hasta sayısının çok olduğu görülmektedir.

Şekil 8: HbA1c değişkeni grafikleri

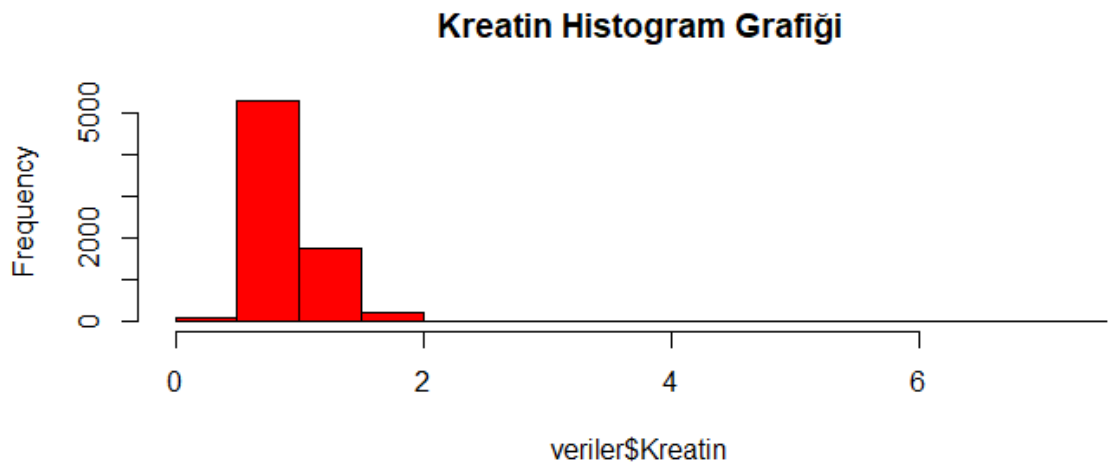


Şekil 9: HbA1c kutu grafiği

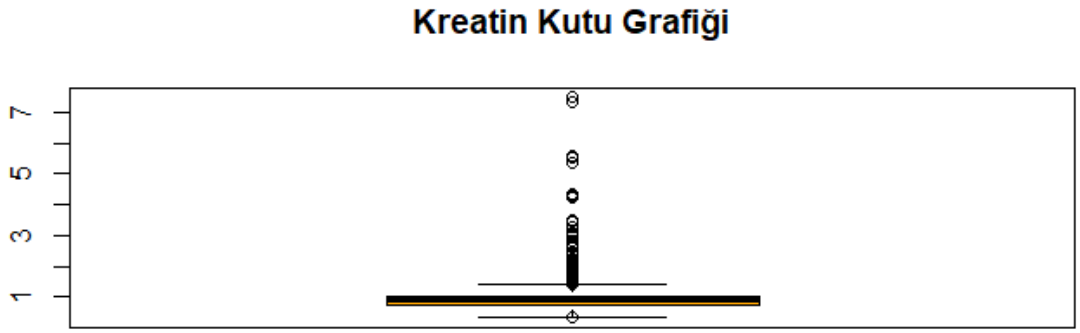


HbA1c değişkeni değerleri 4.6 mmol/L ile 17.5 mmol/L arasında değişmektedir. Ortalaması yaklaşık 8 mmol/L'dir.

Şekil 10: Kreatinin değişkeni grafikleri

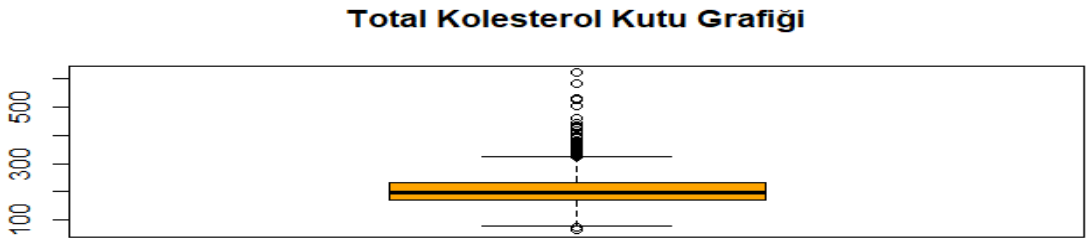


Şekil 11: Kreatinin kutu grafiği

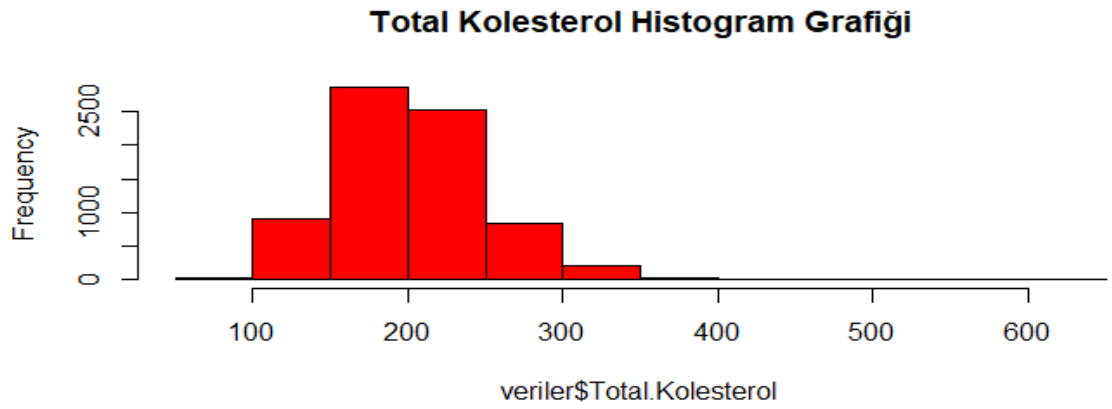


Kreatin değişkeni değerleri 0.32 mg/dL ile 7.48 mg/dL arasında değişmektedir. En çok tekrar eden değerlerin 0.5 mg/dL ile 1 mg/dL değeri arası olduğu görülmektedir. Kutu grafiğinde görüldüğü üzere veri setinde uç değerler vardır. Total kolesterol, HDL ve LDL kolesterolde de aynı şekilde aykırı değerler olduğu görülmüştür. Bu değerlerin analizden çıkartılmamasına karar verilmiştir.

Şekil 12: Total kolesterol değişkeni grafikleri



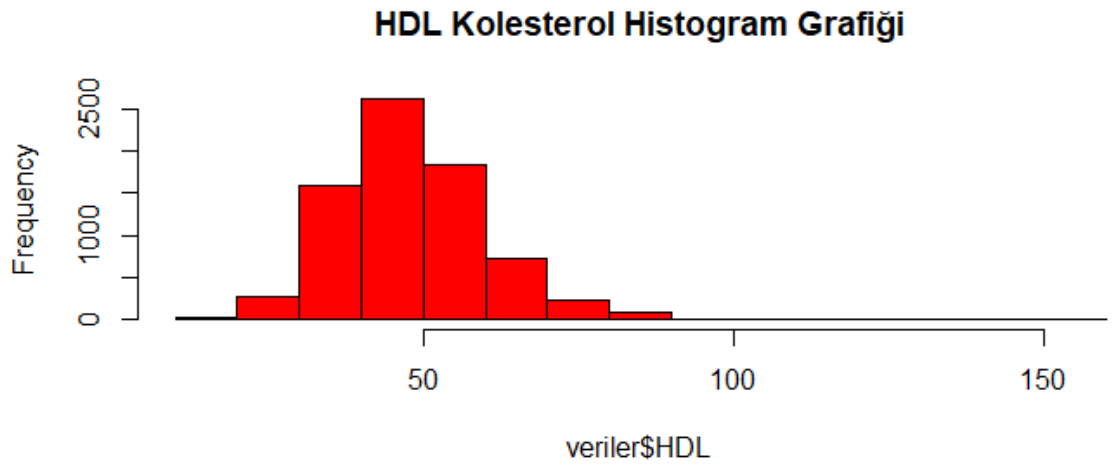
Şekil 13: Total kolesterol histogram grafiği



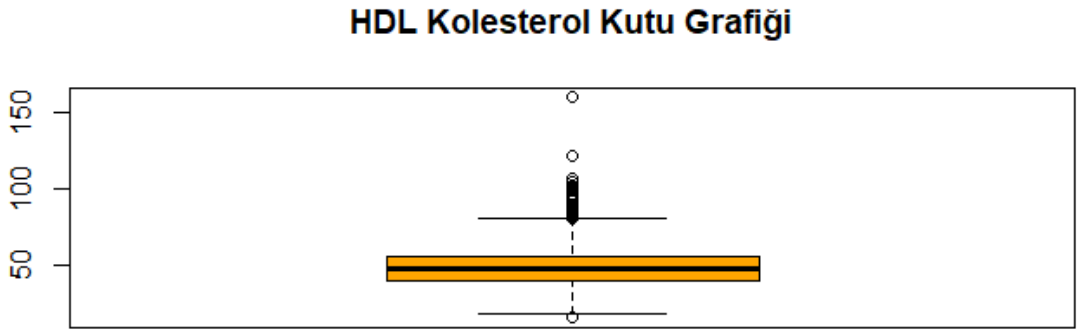
Total kolesterol değişkeni değerleri 64 mg/dL ile 622 mg/dL arasında değişmektedir. Ortalaması 202 mg/dL ‘dir. Sık tekrar edilen değerler 150 mg/dL ile 250 mg/dL arasındadır.

Kutu grafiği incelendiğinde aykırı değerler olduğu görülmektedir. Bu değerler bütünlüğünün bozulmaması adına analizden çıkartılmamıştır.

Şekil 14: HDL kolesterol değişkeni grafikleri

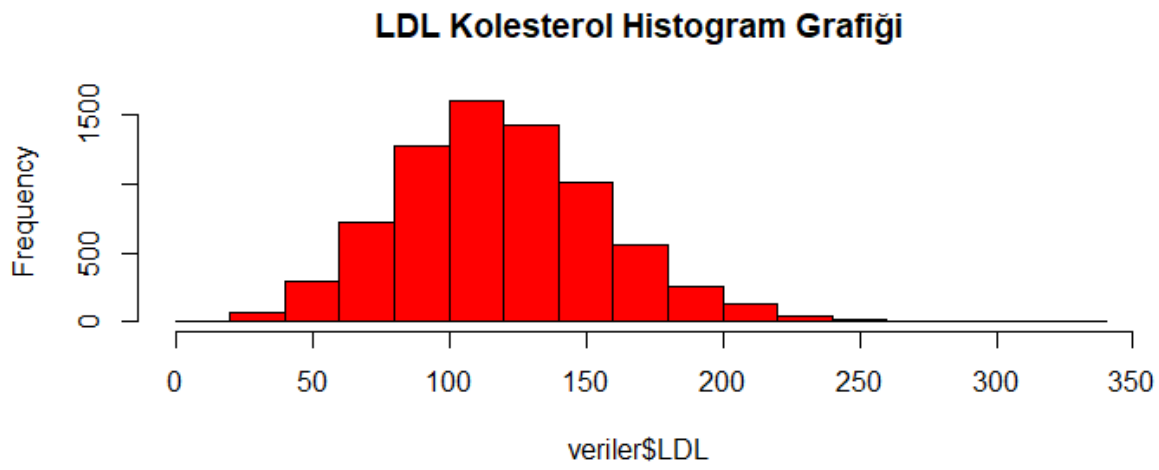


Şekil 15: HDL kolestrol kutu grafiği

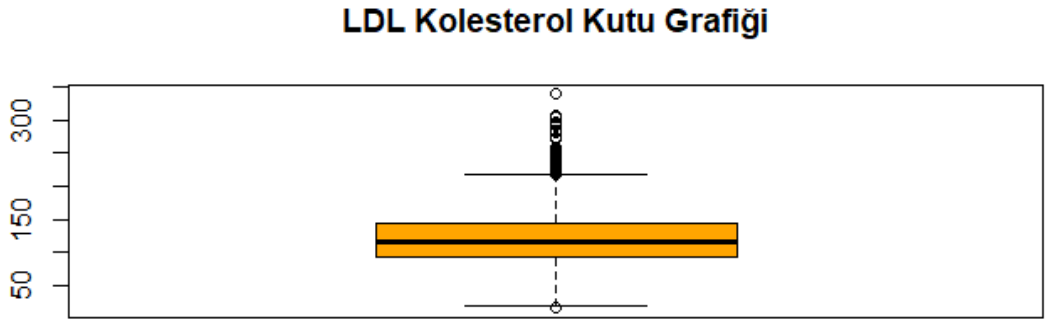


HDL kolesterol değişkeni değerleri 15 mg/dL ile 160 mg/dL arasında değişmektedir. Ortalaması 48.5 mg/dL'dir. Aykırı değerler çıkartılmamıştır.

Şekil 16: LDL kolesterol değişkeni grafikleri

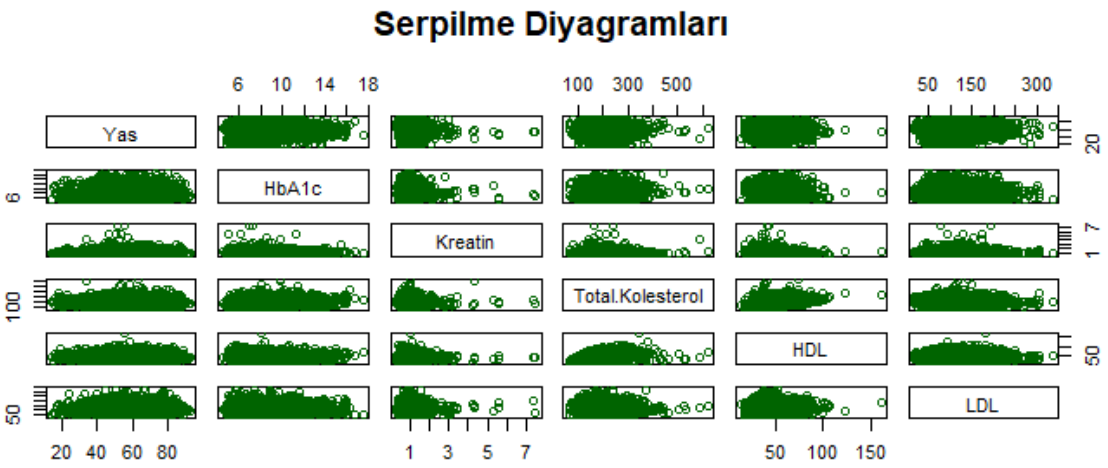


Şekil 17: LDL kolesterol değişkeni kutu grafiği

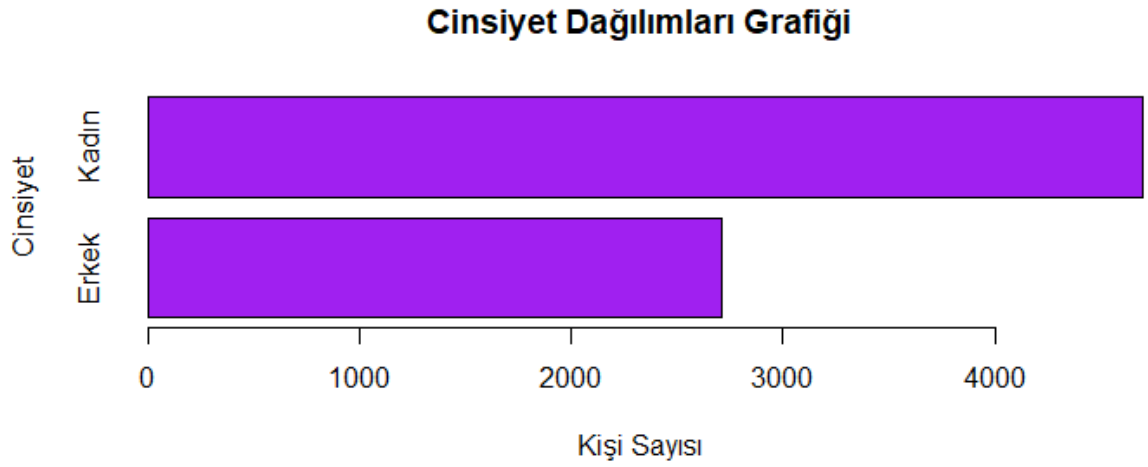


LDL kolesterol değişkeni değerleri 15.8 mg/dL ile 339 mg/dL arasında değişmektedir. Ortalaması 119.5 mg/dL'dir. Aykırı değerler çıkartılmamıştır.

Şekil 18: Serpilme diyagramları

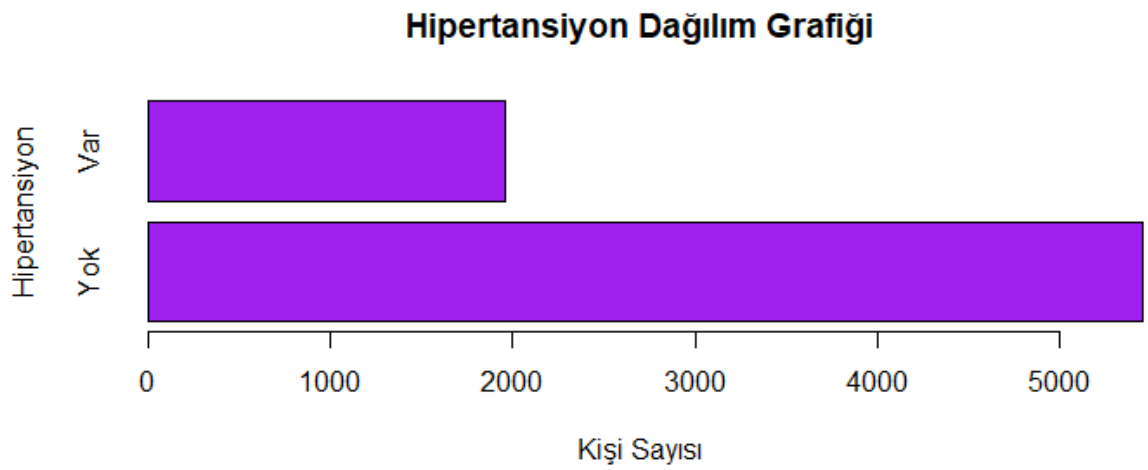


Şekil 19: Cinsiyet dağılımları grafiği



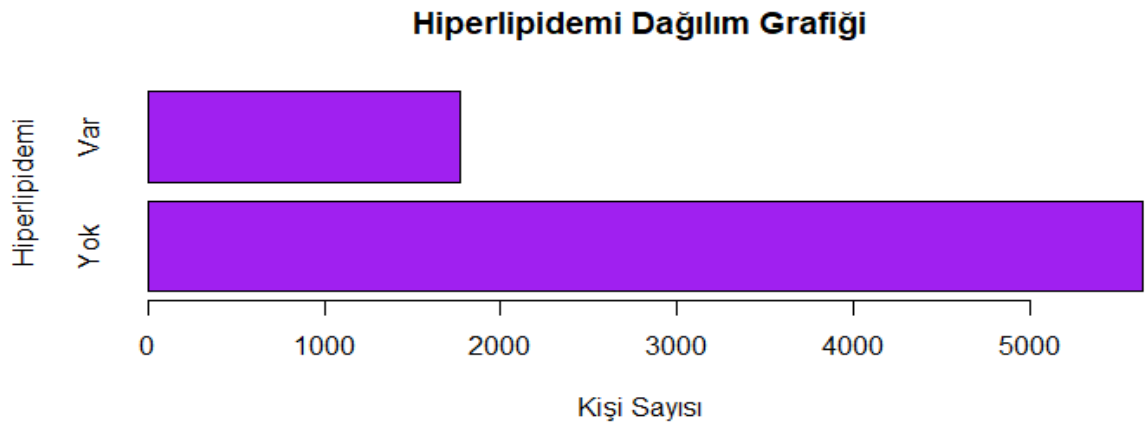
7409 hastanın 2708'i erkek, 4701'i kadındır. Veri setinin % 37'si erkek, %63'ü kadındır.

Şekil 20: Hipertansiyon dağılım grafiği



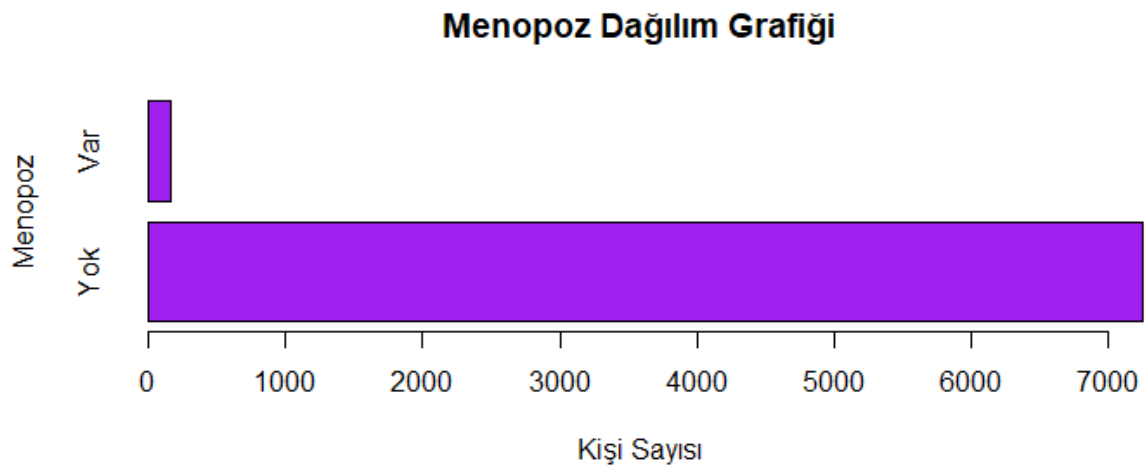
Hipertansiyon hastalığı olan 1954 hasta, olmayan 5455 hasta vardır. Veri setinin %26'sı hipertansiyon tanısı konulan hastalardan, %74'ü hipertansiyon tanısı konulmayan hastalardan oluşmaktadır.

Şekil 21: Hiperlipidemi dağılım grafiği



Hiperlipidemisi olan 1765 hasta, olmayan 5644 hasta vardır.Veri setinin %24'ü hiperlipidemi tanısı konulan, %76'sı hiperlipidemi tanısı konulmayan hastalardan oluşmaktadır.

Şekil 22: Menopoz dağılım grafiği

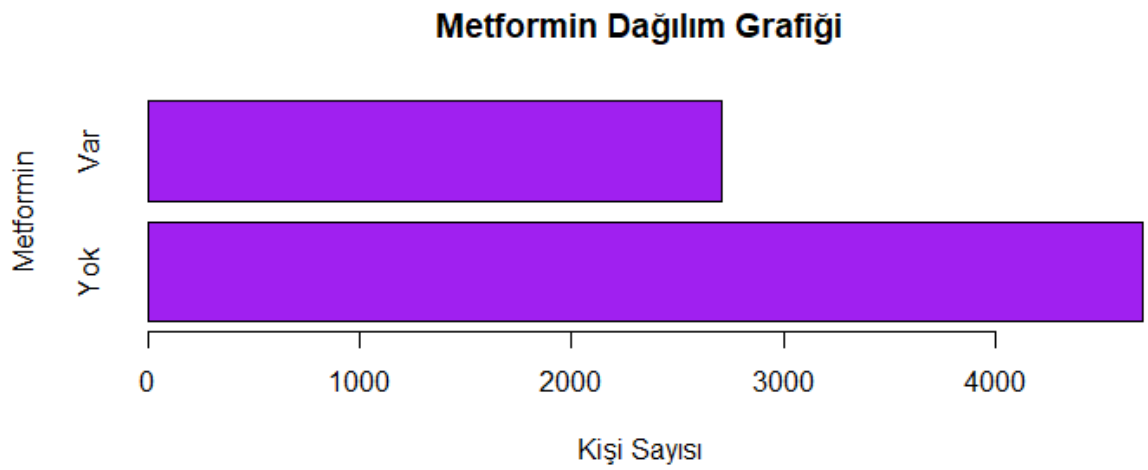


4701 kadın arasından menopoz olan 159 kadın vardır.Kadın hastaların %3.3'ünde menopoz vardır.

Şekil 23: Kırgınlık ve yorgunluk dağılım grafiği

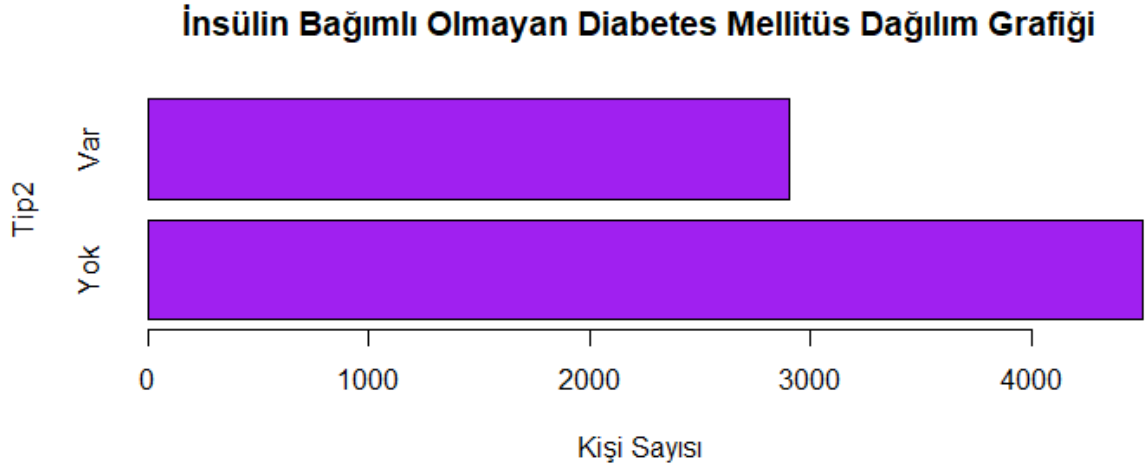
Kırgınlık ve yorgunluk teşhisi konulan 459 hasta, bu teşhis konulmayan 6950 hasta vardır. Veri setinin %6'sı kırgınlık ve yorgunluk tanısı konulan, %94'ü kırgınlık ve yorgunluk tanısı konulmayan hastalardan oluşmaktadır.

Şekil 24: Metformin dağılım grafiği



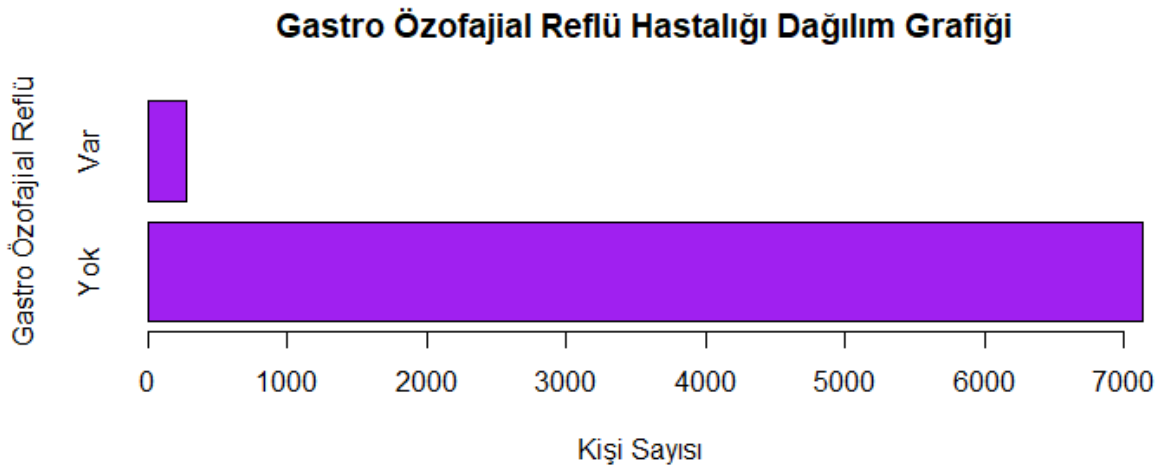
Metformin ilacını kullanan 2710 hasta, kullanmayan 4699 hasta vardır. Veri setinin %37'si metformin ilacını kullanan, %63'ü metformin ilacını kullanmayan hastalardan oluşmaktadır.

Şekil 25: İnsülin bağımlı olmayan diabetes mellitüs dağılım grafiği



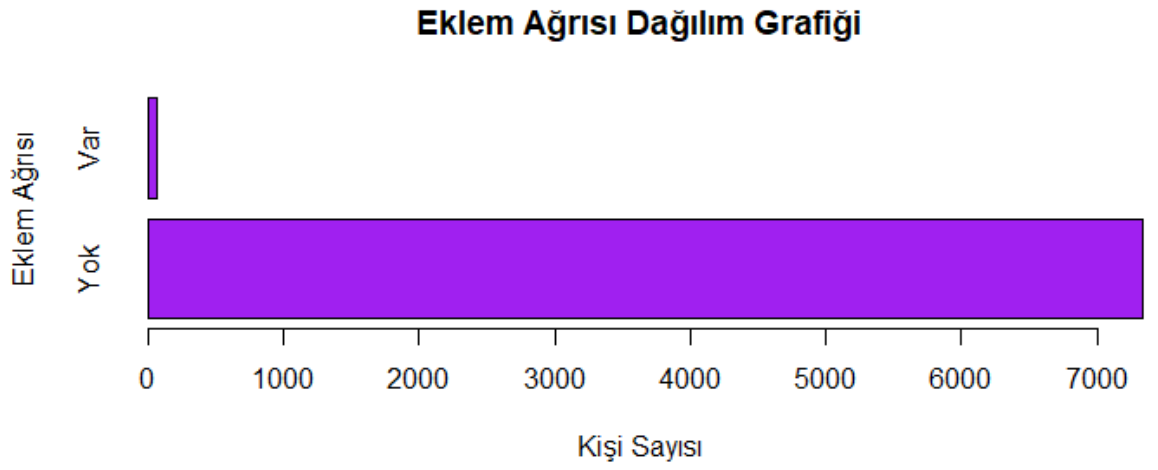
İnsülin bağımlı olmayan diyabet hastası yani tip 2 diyabet hastası 2903 kişi vardır. Tip2 diyabet hastası olmayan 4506 kişi vardır. Veri setinin %39'undan tip2 diyabet hastalığı vardır, %61'inde tip2 diyabet hastalığı yoktur.

Şekil 26: Gastro özofajial reflü hastalığı dağılım grafiği



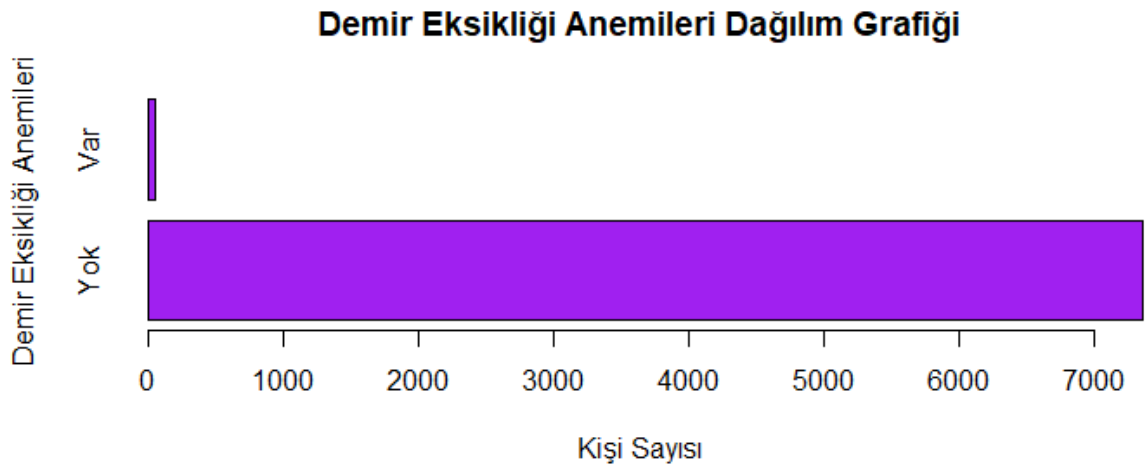
Veri setinde, gastro özofajial reflü hastalığı olan 276 hasta, olmayan 7133 hasta vardır. Veri setinin %4'ü reflü hastalığı olan, %96'sı reflü hastalığı olmayan kişilerden oluşmaktadır

Şekil 27: Eklem ağrısı dağılım grafiği



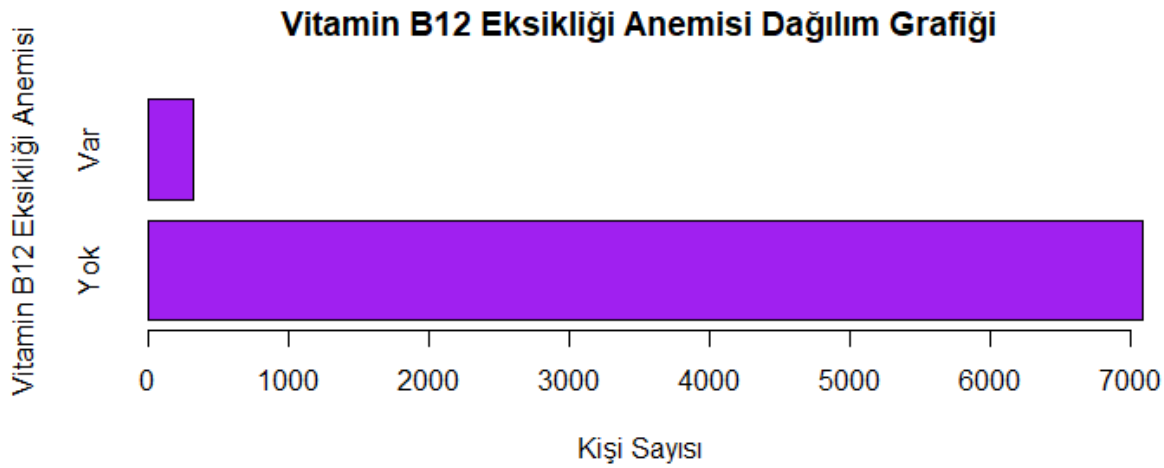
Eklem ağrısı tanısı konulan 70 hasta, bu tanı konulmayan 7339 hasta vardır. %1'lik hasta dilimine eklem ağrsı tanısı konulmuş, %99'luk dilime bu tanı konulmamıştır.

Şekil 28: Demir eksiklikleri dağılım grafiği



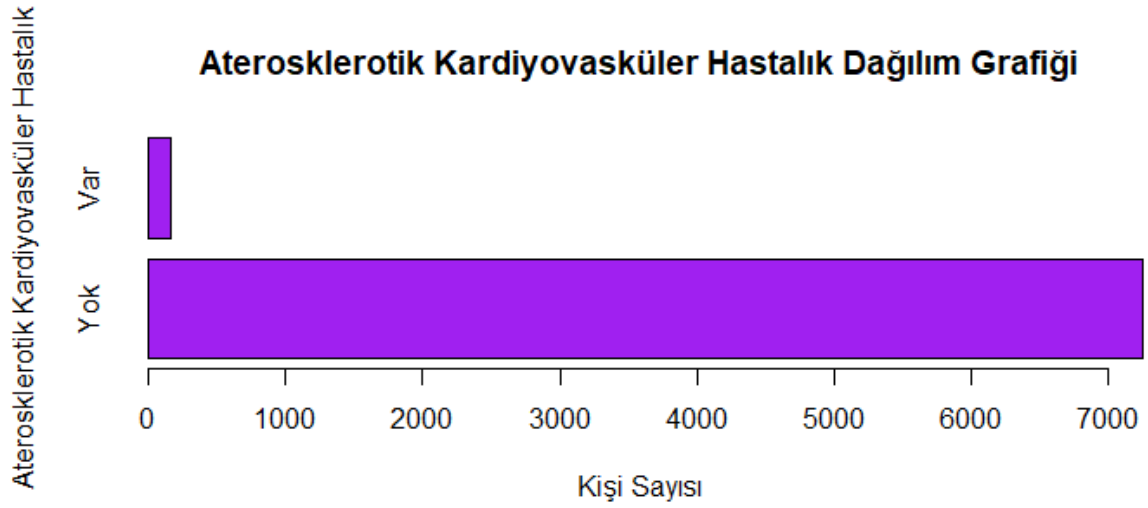
Demir eksikliği anemileri tanısı konulan 47 hasta, bu tanının konulmadığı 7362 hasta vardır. Veri setinin %0.6'sı demir eksikliği anemileri tanısı konulan, %99.4'ü demir eksikliği anemileri tanısı konulmayan hastalardan oluşmaktadır

Şekil 29: Vitamin B12 eksikliği anemisi dağılım grafiği



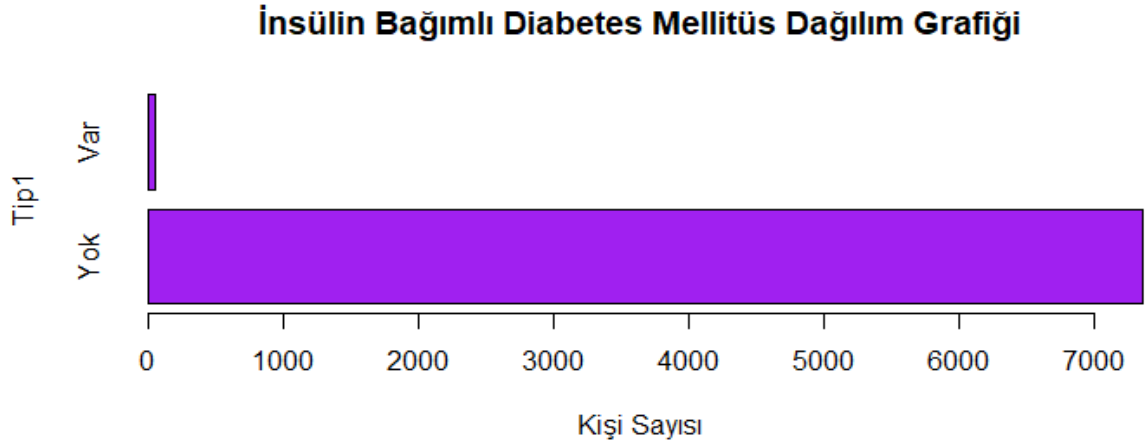
Vitamin B12 eksikliği anemisi tanısı konulan 325 hasta, bu tanının konulmadığı 7084 hasta vardır. Veri setinin %4'ü vitamin B12 eksikliği anemisi tanısı konulan, %96'sı vitamin B12 eksikliği anemisi tanısı konulmayan hastalardan oluşmaktadır.

Şekil 30: Aterosklerotik kardiyovasküler hastalık dağılım grafiği



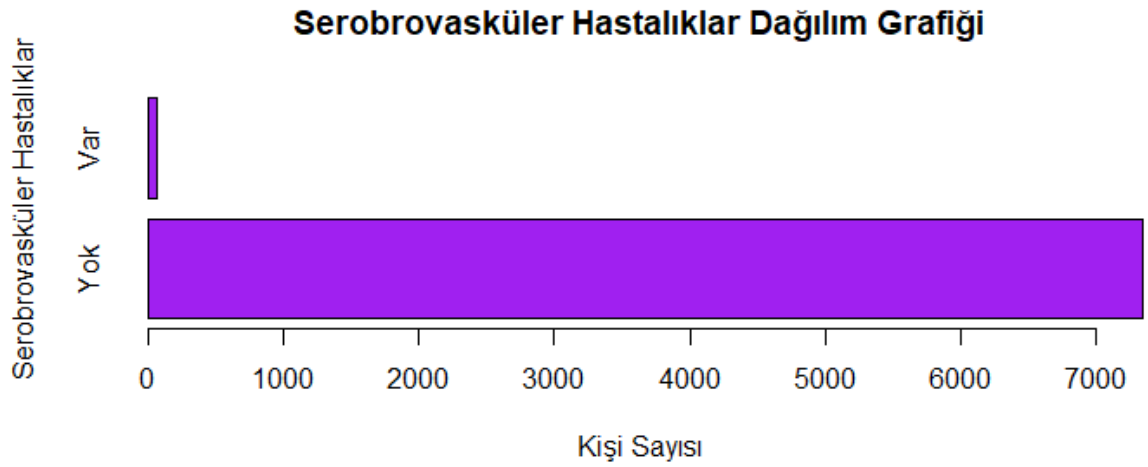
Aterosklerotik kardiyovasküler hastalık tanısı konulan 160 hasta, bu tanının konulmadığı 7249 hasta vardır. Veri setinin %2'si aterosklerotik kardiyovasküler hastalık tanısı konulan, %98'zi aterosklerotik kardiyovasküler hastalık tanısı konulmayan hastalardan oluşmaktadır

Şekil 31: İnsülin bağımlı diabetes mellitüs dağılım grafiği



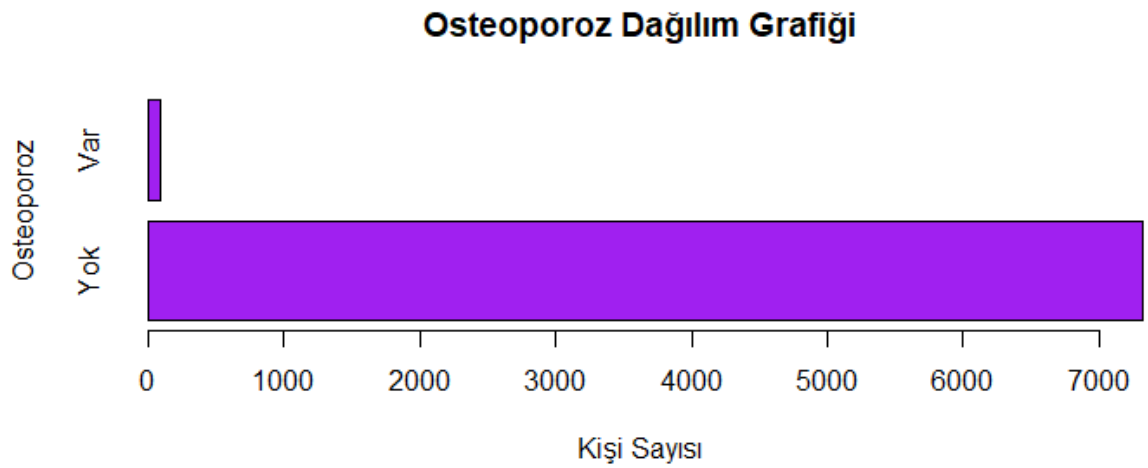
İnsülin bağımlı diyabet hastası yani tip1 diyabet hastası olan 51 hasta, olmayan 7358 hasta vardır. Veri setinin %0.7’ si tip 2 diyabet hastalığı olan, %99.3’ ü tip 2 diyabet hastalığı olmayan kişilerden oluşmaktadır.

Şekil 32: Serobrovasküler hastalıklar dağılım grafiği



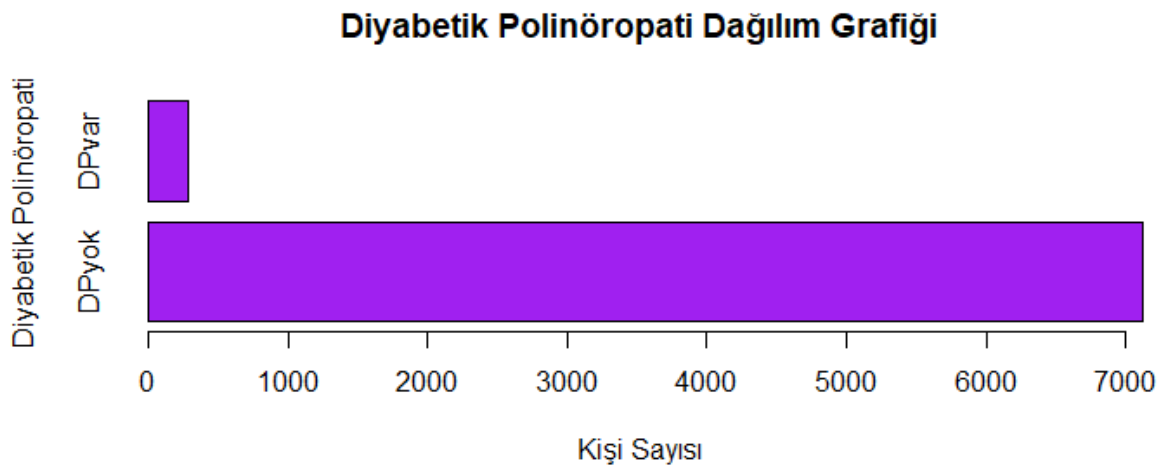
Serobrovasküler hastalık tanısı konulan 61 hasta, bu tanının konmadığı 7348 hasta vardır. Veri setinin %0.8’ i serobrovasküler hastalık tanısı konulan, %99.2’ si serobrovasküler hastalık tanısı konulmayan hastalardan oluşmaktadır

Şekil 33: Osteoporoz dağılım grafiği



Osteoporoz tanısı konulan 85 hasta, konulmayan 7324 hasta vardır. Veri setinin %1.1' i osteoporoz tanısı konulan, %98.9' u osteoporoz tanısı konulmayan hastalardan oluşmaktadır

Şekil 34: Diyabetik polinöropati dağılım grafiği



Diyabetik polinöropati hastası olan 285 hasta, bu tanının konulmadığı 7124 hasta vardır. Veri setinin %4'ü diyabetik polinöropati hastalığı tanısı konulan, %96'sı diyabetik polinöropati hastalığı tanısı konulmayan hastalardan oluşmaktadır

4.5. Veri Dönüştürme

Veri setindeki nümerik değerler normalize edilmiştir. Bu işlem için minimum-maksimum yöntemi kullanılmıştır. En küçük değere 0 , en yüksek değere 1 atanarak, veri seti normalizasyon işlemi sonrası aşağıdaki şekle dönüşmüştür.

Veri seti artık analiz için hazırdır.

Tablo 10: Veri seti özeti 2

Yaş	Cinsiyet	Hipertansiyon	Hiperlipidemi	Menopoz	HbA1c	Kreatinin	Total Kolesterol	HDL Kolesterol	LDL Kolesterol	Kırgınlık ve Yorgunluk
Min. :0000000 0	Erkek:270 8	0:5455	0:5644	0:7250	Min. : 0.000000 0	Min. : 0.0000000	Min. : 0.000000 0	Min. : 0.000000 0	Min. : 0.000000 0	0:6950
1st Qu.: 0.4683544	Kadın:470 1	1:1954	1:1765	1: 159	1st Qu.: 0.139534 9	1st Qu.: 0.0614525 1	1st Qu.: 0.189964 2	1st Qu.: 0.172413 8	1st Qu.: 0.240099 0	1: 459
Median : 0.5696203					Median : 0.224806 2	Median : 0.0796089 4	Median : 0.241935 5	Median : 0.220689 7	Median : 0.312500 0	
Mean : 0.5695160					Mean : 0.258071 8	Mean : 0.0854677 3	Mean : 0.249048 3	Mean : 0.231206 2	Mean : 0.320778 1	
3rd Qu.: 0.6708861					3rd Qu.: 0.333333 3	3rd Qu.: 0.0991620 1	3rd Qu.: 0.301075 3	3rd Qu.: 0.282758 6	3rd Qu.: 0.393564 4	
Max. :1.000000 0					Max. : 1.000000 0	Max. : 1.0000000	Max. : 1.000000 0	Max. : 1.000000 0	Max. : 1.000000 0	

Metformin	İnsülin Bağımlı Olmayan Diabetes Mellitüs	Gastro Özofajial Reflü Hastalığı	Eklem Ağrısı	Demir Eksikliği Anemileri	Vitamin B12 Eksikliği Anemisi	Aterosklerotik Kardiyovasküler Hastalık	İnsülin Bağımlı Diabetes Mellitüs	Serobrovasküler Hastalık	Osteoporoz	Diyabetik Polinöropati
0:4699	0:4506	0:7133	0:7339	0:7362	0:7084	0:7249	0:7358	0:7348	0:7324	var: 285
1:2710	1:2903	1: 276	1: 70	1: 47	1: 325	1: 160	1: 51	1: 61	1: 85	yok:7124

4.6. KNN Algoritması

Sınıflandırma algoritmalarından olan K-En Yakın Komşu (KNN) algoritması danışmanlı öğrenen bir algoritmadır. Yani veri setinden öğrenim yapar. Veri seti, eğitim ve test veri seti olarak ikiye ayırılır. Eğitim veri setine algoritma öğretilir. Test veri setiyle de algoritma modeli test edilir. Gerçek veri ile öngörülen veri kıyaslanır. Kontenjans tablosu (confusion matrix) oluşturulur. Bu matrise göre modelin performans değerlendirme ölçütleri bulunur. Performans değerlendirme ölçütleri kurulan modelin ne kadar performans verdiği ölçer. Bunun için doğruluk oranı, hata oranı gibi ölçütler kullanılır.

KNN algoritması yeni bir veriyi sınıfa atarken, belirlenen bir k değeri kullanır. Bu k değerine göre mevcut örneklem içindeki verilere olan uzaklığı hesaplanır. Bu şekilde en yakın komşusu bulunarak o kümeye atanır.

Bu çalışmada Öklid uzaklığı ile uzaklıklar hesaplanarak K-en yakın komşu algoritması uygulanmıştır. Bunun için, veri setinden yalnızca nümerik değerler taşıyan ve hedef niteliğin de olduğu bir alt küme elde edilmiştir. Bu değişkenler, “Yas”, ”HbA1c”, “Kreatin”, “Total Kolesterol”, “HDL” ve “LDL” değişkenleridir, hedef nitelik de “Diyabetik Polinöropati” değişkenidir.

Formülü uyguladığımızda 7409 veri ve 7 değişkenden oluşan yeni bir data.frame elde edilir.

Veri seti %60’ı eğitim ve %40’ı test veri seti olarak ayrılmıştır. Performans değerlendirme yöntemi olarak “Hold-out” yöntemi kullanılmıştır.

k değeri 1’den 10’ a kadar denenmiştir.

Modelin performansının ölçülmesi için kontenjans tablosu kurulur.

k = 1 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	17	98
DPyok	97	2751

True positives(TP), doğru pozitif değeri 17'dir.

Gerçekte diyabetik polinöropati hastası olan 17 kişi, tahminde de diyabetik polinöropati hastasıdır diye tahmin edilmiştir. Kontenjans tablosu true positives yani doğru pozitif değeri bunu gösterir.

False positives(FP), yanlış pozitif değeri 98'dir

Var olan durum yani gerçekte diyabetik polinöropati hastalığı olmayan 98 kişi, diyabetik polinöropati hastalığı vardır şeklinde tahmin edilmiştir. Yani gerçekte olmayan bir durum tahminde var tahmin edilmişitir. Buna tip 1 hata denir. Kontenjans tablosunda yanlış pozitif sınıfı bu değeri gösterir.

False negatives(FN), yanlış negatif değeri 97'dir.

Gerçekte diyabetik polinöropati hastası olan 97 kişi, tahminde diyabetik polinöropati hastası değildir şeklinde tahmin edilmiştir. Gerçekte var olan(pozitif) bir durumun, tahminde yoktur(negatif) şeklinde bulunmasına tip 2 hata denir. Kontenjans tablosunda yanlış negatif sınıfı bu değeri gösterir.

True negatives(TN), doğru negatif değeri 2751'dır.

Gerçekte diyabetik polinöropati olmayan hastalardan, modelin diyabetik polinöropati değildir biçiminde tahmin ettiği hasta sayısıdır. kontenjans tablosunda doğru negatif sınıfı bu değeri gösterir.

kontenjans tablosunda bulunan değerlere göre modelin doğruluk oranı, hata oranı, duyarlılık oranı, belirleyicilik oranı, kesinlik oranı, negatif öngörü değeri, yanlış pozitif oranı, yanlış negatif oranı ve F-ölçütü hesaplanmıştır.

Tablo 11: k=1 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=1 için
Doğruluk oranı	% 93.4
Hata oranı	%6.5
TPR(Duyarlılık oranı)	%14.9

SPC(Belirleyicilik oranı)	%96.5
PPV(Pozitif öngörü oranı)	%14.7
NPV(Negatif öngörü oranı)	%96.5
FPR(Yanlış pozitif oranı)	%3.3
FNR(Yanlış negatif oranı)	%85
F-ölçütü	%14.8

k=2 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	8	112
DPyok	106	2737

Tablo 12: k=2 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=2 için
Doğruluk oranı	% 92.6
Hata oranı	%7.3
TPR(Duyarlılık oranı)	%7
SPC(Belirleyicilik oranı)	%96
PPV(Pozitif öngörü oranı)	%6.6
NPV(Negatif öngörü oranı)	%96.2
FPR(Yanlış pozitif oranı)	%3.7
FNR(Yanlış negatif oranı)	%92.9
F-ölçütü	%6.8

k=3 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri değerleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	5	26
DPyok	109	2823

Tablo 13: k=3 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=3 için
Doğruluk oranı	% 95.4
Hata oranı	%4.5
TPR(Duyarlılık oranı)	%4.3
SPC(Belirleyicilik oranı)	%99
PPV(Pozitif öngörü oranı)	%16
NPV(Negatif öngörü oranı)	%96.2
FPR(Yanlış pozitif oranı)	%0.8
FNR(Yanlış negatif oranı)	%95.6
F-ölçütü	%6.8

k=4 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	21
DPyok	113	2828

Tablo 14: k=4 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=4 için
Doğruluk oranı	% 95.4
Hata oranı	%4.5
TPR(Duyarlılık oranı)	%0.8
SPC(Belirleyicilik oranı)	%99.2
PPV(Pozitif öngörü oranı)	%4.5
NPV(Negatif öngörü oranı)	%96
FPR(Yanlış pozitif oranı)	%0.7
FNR(Yanlış negatif oranı)	%99.1
F-ölçütü	%14.7

k=5 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	9
DPyok	113	2840

Tablo 15: k=5 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=5 için
Doğruluk oranı	% 95.8
Hata oranı	%4.1
TPR(Duyarlılık oranı)	%0.8
SPC(Belirleyicilik oranı)	%99.6
PPV(Pozitif öngörü oranı)	%10
NPV(Negatif öngörü oranı)	%96.1

FPR(Yanlış pozitif oranı)	%0.3
FNR(Yanlış negatif oranı)	%99.1
F-ölçütü	%16.1

k=6 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	11
DPyok	113	2838

Tablo 16: k=6 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=6 için
Doğruluk oranı	% 95.4
Hata oranı	% 4.5
TPR(Duyarlılık oranı)	% 0.8
SPC(Belirleyicilik oranı)	% 99.2
PPV(Pozitif öngörü oranı)	% 4.5
NPV(Negatif öngörü oranı)	% 96.1
FPR(Yanlış pozitif oranı)	% 0.7
FNR(Yanlış negatif oranı)	% 99.1
F-ölçütü	% 14.7

k=7 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	8
DPyok	113	2841

Tablo 17: k=7 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=7 için
Doğruluk oranı	% 95.9
Hata oranı	%4
TPR(Duyarlılık oranı)	%0.8
SPC(Belirleyicilik oranı)	%99.7
PPV(Pozitif öngörü oranı)	%11
NPV(Negatif öngörü oranı)	%96.1
FPR(Yanlış pozitif oranı)	%0.2
FNR(Yanlış negatif oranı)	%99.1
F-ölçütü	%16.2

k=8 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	4
DPyok	113	2845

Tablo 18: k=8 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=8 için
Doğruluk oranı	% 96
Hata oranı	% 3.9
TPR(Duyarlılık oranı)	% 0.8
SPC(Belirleyicilik oranı)	% 99.8
PPV(Pozitif öngörü oranı)	% 20
NPV(Negatif öngörü oranı)	% 96.1

FPR(Yanlış pozitif oranı)	% 0.1
FNR(Yanlış negatif oranı)	% 99.1
F-ölçütü	% 1.6

k=9 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri;

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	4
DPyok	113	2845

Tablo 19: k=9 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=9 için
Doğruluk oranı	% 96
Hata oranı	% 3.9
TPR(Duyarlılık oranı)	% 0.8
SPC(Belirleyicilik oranı)	% 99.8
PPV(Pozitif öngörü oranı)	% 20
NPV(Negatif öngörü oranı)	% 96.1
FPR(Yanlış pozitif oranı)	% 0.1
FNR(Yanlış negatif oranı)	% 99.1
F-ölçütü	% 1.6

k=10 değeri için kontenjans tablosu ve performans değerlendirme ölçütleri

Kontenjans tablosu

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	1	3
DPyok	113	2846

Tablo 20: k=10 değeri için performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	k=10 için
Doğruluk oranı	% 96
Hata oranı	%3.9
TPR(Duyarlılık oranı)	%0.8
SPC(Belirleyicilik oranı)	%99.8
PPV(Pozitif öngörü oranı)	%25
NPV(Negatif öngörü oranı)	%96.1
FPR(Yanlış pozitif oranı)	%0.1
FNR(Yanlış negatif oranı)	%99.1
F-ölçütü	%1.6

Tablo 21: Tüm k değerleri performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**
Doğruluk oranı	% 93.4	% 92.6	% 95.4	% 95.4	% 95.8	% 95.4	% 95.9	% 96	% 96	% 96
Hata oranı	%6.5	%7.3	%4.5	%4.5	%4.1	% 4.5	%4	% 3.9	% 3.9	%3.9
TPR(Duyarlılık oranı)	%14.9	%7	%4.3	%0.8	%0.8	% 0.8	%0.8	% 0.8	% 0.8	%0.8
SPC(Belirleyicilik oranı)	%96.5	%96	%99	%99.2	%99.6	% 99.2	%99.7	% 99.8	% 99.8	%99.8
PPV(Pozitif öngörü oranı)	%14.7	%6.6	%16	%4.5	%10	% 4.5	%11	% 20	% 20	%25
NPV(Negatif öngörü oranı)	%96.5	%96.2	%96.2	%96	%96.1	% 96.1	%96.1	% 96.1	% 96.1	%96.1
FPR(Yanlış pozitif oranı)	%3.3	%3.7	%0.8	%0.7	%0.3	% 0.7	%0.2	% 0.1	% 0.1	%0.1
FNR(Yanlış negatif oranı)	%85	%92.9	%95.6	%99.1	%99.1	% 99.1	%99.1	% 99.1	% 99.1	%99.1
F-ölçütü	%14.8	%11	%6.8	%14.7	%16.1	% 14.7	%16.2	% 1.6	% 1.6	%1.6

Tablo 22: Tüm k değerleri kontenjans tablosu

k=5	Gerçek	
tahmin	Var	yok
var	1	9
yok	113	2840

k=6	Gerçek	
tahmin	Var	Yok
var	1	11
yok	113	2838

k=7	Gerçek	
tahmin	Var	Yok
var	1	8
yok	113	2841

k=8	Gerçek	
tahmin	Var	Yok
var	1	4
yok	113	2845

k=1	Gerçek	
tahmin	var	Yok
var	17	98
yok	97	2751

k=2	Gerçek	
tahmin	var	yok
var	8	112
yok	106	2737

k=3	Gerçek	
tahmin	var	Yok
var	5	26
yok	109	2823

k=4	Gerçek	
tahmin	var	yok
var	1	21
yok	113	2828

k=9	Gerçek	
tahmin	var	yok
var	1	4
yok	113	2845

k=10	Gerçek	
tahmin	var	yok
var	1	3
yok	113	2846

k-NN algoritması performans değerlendirme ölçütleri oranları incelendiğinde doğruluk oranları için en yüksek değerler k=8, k=9 ve k=10'dur. Yanlış sınıflandırılmış örnek sayısının tüm veriye oranı olan hata oranları için en düşük değerler yine k=8, k=9 ve k=10 değeridir. Duyarlılık oranı yani doğru sınıflandırılmış pozitif değerlerin tüm pozitif değerlere oranı en yüksek k=1 değerinde çıkmıştır. Performans değerlendirme ölçütlerine göre en iyi değerleri veren model k=10 gibi gözükse de kontenjans tablosu incelenmelidir. k=1 kontenjans tablosuna bakıldığında gerçek pozitif değerin 17 çıktığı, yani gerçekte DP olan 17 hastanın tahminde de doğru tahmin edildiği görülmektedir. Oysa diğer matrislerde bu oran sürekli düşmektedir. k=2 matrisinde 8 kişi hem DP olup hemde doğru tahmin edilmiş, k=3 matrisinde 5 kişi hem DP olup hem de doğru tahmin edilmiştir. Bu matrislerden sonraki k değerlerinde, doğru pozitif değerleri 1 çıkmıştır. Bu çalışmada gerçek veriler ile hastalık tahmini yapılmaya çalışıldığından k=10 değeri sonuçları en

yüksek doğruluk oranını ve en düşük hata oranını verse de, karışık matrisinde ortaya çıkan sonuçtan ötürü k=1 değeri tercih edilebilir.

k=1 değeri için kontenjans tablosu incelendiğinde, doğru pozitif(TP) değeri yani gerçekte diyabetik polinöropati olan hastalardan modelin diyabetik polinöropati olarak tahmin ettiği hastaların sayısı 17 dir.

Yanlış pozitif(FP) değeri yani gerçekte diyabetik polinöropati olmayan hastalardan modelin diyabetik polinöropati hastasıdır biçiminde tahmin ettiği kişi sayısı 98'dir. Tip 1 hatayı verir.

Yanlış negatif(FN) değeri yani gerçekte diyabetik polinöropati hastalığı olan kişilerin, modelin diyabetik polinöropati hastası değildir olarak tahmin ettiği kişi sayısı 97'dir. Tip 2 hatayı verir.

Doğru negatif(TN) değeri yani gerçekte diyabetik polinöropati olmayan hastalardan modelin diyabetik polinöropati hastası değildir biçiminde tahmin ettiği kişi sayısı 2751'dir.

KNN algoritması için k=1 ve k= 10 değeri ayrı ayrı incelenmiştir.

4.7. Naive (Basit) Bayes Sınıflandırıcı Algoritması

Naive (Basit) Bayes ile bütün koşullu olasılık değerleri çarpılarak sınıflandırılır. Temeli Bayes teoremine dayanmaktadır. Bayes teoreminde koşullu olasılıklar ile marjinal olasılıklar arasındaki ilişki gösterilmektedir. Naive bayes yöntemi sınıflandırma algoritmaları içerisinde yer almaktadır.

Analiz öncesi değişkenler faktör ve nümerik olarak tanımlanmış, nümerik veriler normalize edilip analize uygun hale getirilmiştir. Diğer snıflandırma algoritmalarında olduğu gibi veri seti eğitim veri seti ve test veri seti olarak ayrılmıştır. Eğitim veri seti %60, test veri seti %40 olarak bölünmüştür. Eğitim ve test veri setine tahmininde kullanılacak nitelik ve hedef nitelik(diyabetik polinöropati) atanmıştır.

Naive Bayes algoritmasının kullanılması için R programına "e1071" paketi yüklenmeli ve kütüphaneden çağrılmalıdır. Bu paketteki naiveBayes() fonksiyonu kullanılmıştır. Model tahmin edilmiş ve aşağıdaki koşullu olasılık değerleri bulunmuştur.

Naive Bayes Classifier for Discrete Predictors

Call: naiveBayes.default(x = egitimNitelikleri, y = egitimHedefNitelik)

A-priori probabilities:

egitimHedefNitelik

DPvar DPyok

0.03846154 0.96153846

Conditional probabilities:

Yas

egitimHedefNitelik [,1] [,2]

DPvar 0.5776149 0.1367536

DPyok 0.5712606 0.1535764

Cinsiyet

egitimHedefNitelik Erkek Kadın

DPvar 0.3918129 0.6081871

DPyok 0.3670175 0.6329825

Hipertansiyon

egitimHedefNitelik Yok Var

DPvar 0.6081871 0.3918129

DPyok 0.7443275 0.2556725

Hiperlipidemi

egitimHedefNitelik Yok Var

DPvar 0.5263158 0.4736842

DPyok 0.7723977 0.2276023

Menopoz

egitimHedefNitelik Yok Var

DPvar 0.97660819 0.02339181

DPyok 0.98035088 0.01964912

HbA1c

egitimHedefNitelik [,1] [,2]

DPvar 0.2433927 0.1368167

DPyok 0.2593545 0.1508886

Kreatinin

egitimHedefNitelik [,1] [,2]

DPvar 0.08237773 0.03948762

DPyok 0.08512562 0.04357342

Total.Kolesterol

egitimHedefNitelik [,1] [,2]

DPvar 0.2420717 0.08373833

DPyok 0.2483020 0.08693878

HDL

egitimHedefNitelik [,1] [,2]

DPvar 0.2311966 0.07822327

DPyok 0.2306601 0.08069054

LDL

egitimHedefNitelik [,1] [,2]

DPvar 0.3475805 0.1285597

DPyok 0.3185527 0.1188165

Kirginlik.ve.Yorgunluk

egitimHedefNitelik Yok Var

DPvar 0.91812865 0.08187135

DPyok 0.93824561 0.06175439

Metformin

egitimHedefNitelik Yok Var

DPvar 0.3742690 0.6257310

DPyok 0.6444444 0.3555556

Insulin.Bagimli.Olmayan.Diyabetes.Mellitus

egitimHedefNitelik Yok Var

DPvar 0.2163743 0.7836257

DPyok 0.6222222 0.3777778

Gastro.Ozofajial.Reflu.Hastaligi

egitimHedefNitelik Yok Var

DPvar 0.90643275 0.09356725

DPyok 0.96584795 0.03415205

Eklem.Agrisi

egitimHedefNitelik Yok Var

DPvar 0.970760234 0.029239766

DPyok 0.990877193 0.009122807

Demir.Eksikligi.Anemileri

egitimHedefNitelik Yok Var

DPvar 0.959064327 0.040935673

DPyok 0.994853801 0.005146199

Vitamin.B12.Eksikligi.Anemisi

egitimHedefNitelik Yok Var

DPvar 0.92982456 0.07017544

DPyok 0.95859649 0.04140351

Aterosklerotik.Kardiyovaskuler.Hastalik

egitimHedefNitelik Yok Var

DPvar 0.94736842 0.05263158

DPyok 0.98105263 0.01894737

Insulin.Bagimli.Diyabetes.Mellitus

egitimHedefNitelik Yok Var

DPvar 0.994152047 0.005847953

DPyok 0.993450292 0.006549708

Serebrovaskuler.Hastaliklar

egitimHedefNitelik Yok Var

DPvar 0.976608187 0.023391813

DPyok 0.991578947 0.008421053

Osteoporoz

egitimHedefNitelik Yok Var

DPvar 0.994152047 0.005847953

DPyok 0.988771930 0.011228070

Model tahmininde bulunan sonuçlar eklerde verilmiştir.

Tahmin edilen değerlerin ve gerçek değerlerin kıyaslanması için kontenjans tablosu elde edilmiştir.

Tablo 23: Naive bayes kontenjans tablosu

Naive Bayes	Gerçek Sınıflar		
Tahmini Sınıflar		DPvar	DPyok
	DPvar	5	41
	DPyok	109	2808

Tablo 24: Naive bayes performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	**Naive Bayes**
Doğruluk oranı	% 94.9
Hata oranı	%5
TPR(Duyarlılık oranı)	%4.3
SPC(Belirleyicilik oranı)	%98.5
PPV(Pozitif öngörü oranı)	%10.8
NPV(Negatif öngörü oranı)	%96.2
FPR(Yanlış pozitif oranı)	%1.3
FNR(Yanlış negatif oranı)	%95.6
F-ölçütü	%6.2

Naive Bayes algoritması kontenjans tablosu sonuçlarına göre gerçekte diyabetik polinöropati hastalığı olan 5 hasta, tahminde de diyabetik polinöropati hastası olarak tahmin edilmiştir. Doğru pozitif değeri 5'tir.

Gerçekte diyabetik polinöropati hastalığı bulunmayan, ama tahminde diyabetik polinöropati hastasıdır çıkan 41 kişi vardır. Yanlış pozitif yani tip 1 hata değeri 41'dir.

Gerçekte diyabetik polinöropati hastası olan, tahminde diyabetik polinöropati hastası değildir çıkan 109 kişi vardır. Yanlış negatif yani tip 2 hata değeri 109'dur.

Gerçekte diyabetik polinöropati hastalığı olmayan, tahminde de diyabetik polinöropati hastalığı yoktur çıkan 2808 kişi vardır. Doğru negatif değeri 2808'dir.

Modelin doğruluk oranı 0.949 ve hata oranı 0.05 çıkmıştır.

4.8. Lojistik Regresyon Algoritması

Bağımsız değişkenlerin iki ya da daha fazla kategori içeren bağımlı değişken üzerindeki etkisini ölçmek amacıyla kullanılan lojistik regresyon, diyabetik polinöropati hastalarının tahminlenmesi için kullanılmıştır.

Lojistik regresyon analizi sınıflandırma yöntemi içerisinde olan bir analiz türüdür. Veri seti eğitim ve test veri seti olarak ayrılmalıdır. Lojistik regresyon analizi yaparken referans değeri önemlidir. Mutlaka istenilen şekilde girilmelidir.

Veri analize sokulmadan veri özetine bakılmıştır. Değişkenler faktör ve nümerik şeklinde tanımlanmıştır. Nümerik nitelikler normalize edildi. Referans kategorisi için istenilen durum için kod yazılmıştır. Referans değeri veriler$Diyabetik.Polinoropati == var olarak belirlenmiştir.

Veri setinin %60 eğitim, %40'ı test veri seti olarak ayrılmıştır.

Logistik regresyon uygulmasında glm() fonksiyonu kullanılmıştır.

```
Call:
glm(formula = Diyabetik.Polinoropati ~ ., family = binomial,
    data = veriler)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0676   0.1508   0.1734   0.2995   1.1340

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             4.532025   0.563276   8.046 8.57e-16
***
Yas                                    -0.318061   0.589457  -0.540 0.589485
CinsiyetKadın                           0.249221   0.170710   1.460 0.144314
HipertansiyonVar                       -0.007493   0.179173  -0.042 0.966643
HiperlipidemiVar                       -0.665641   0.177674  -3.746 0.000179
***
MenopozVar                             -0.278657   0.538517  -0.517 0.604840
```

```
HbA1c                                          0.631205   0.563024   1.121 0.262246
Kreatinin                                      0.071847   1.726072   0.042 0.966798
Total.Kolesterol                               0.085919   0.962923   0.089 0.928901
HDL                                           -0.090639   0.784226  -0.116 0.907987
LDL                                           -0.749019   0.616126  -1.216 0.224103
Kirginlik.ve.YorgunlukVar                      1.281876   0.546786   2.344 0.019059 *
MetforminVar                                  -0.422491   0.181855  -2.323 0.020166 *
Insulin.Bagimli.Olmayan.Diyabetes.MellitusVar -1.378004   0.196923  -6.998 2.60e-12
***
Gastro.Ozofajial.Reflu.HastaligiVar           -0.693411   0.357215  -1.941 0.052239 .
Eklem.AgrisiVar                               -1.166757   0.517676  -2.254 0.024207 *
Demir.Eksikligi.AnemileriVar                  -4.252568   1.294096  -3.286 0.001016
**
Vitamin.B12.Eksikligi.AnemisiVar               0.169234   0.369862   0.458 0.647268
Aterosklerotik.Kardiyovaskuler.HastalikVar    -0.133880   0.540827  -0.248 0.804486
Insulin.Bagimli.Diyabetes.MellitusVar          0.207445   1.050079   0.198 0.843395
Serebrovaskuler.HastaliklarVar                -1.736864   0.513013  -3.386 0.000710
***
OsteoporozVar                                  0.144930   0.639175   0.227 0.820622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1449.6  on 4445  degrees of freedom
Residual deviance: 1263.4  on 4424  degrees of freedom
AIC: 1307.4

Number of Fisher Scoring iterations: 7
```

Oluşan modelde, β katsayıları estimate ile ifade edilmiştir. Katsayılara ait standart hata değerleri Std. Error şeklinde gösterilmiştir. Z-value (Wald istatistiği değeri) değerleri ile p değerleri de verilmiştir. β katsayılarının p değerleri incelendiğinde anlamlı çıkan sonuçların 0.05'den küçük olduğu görülmektedir.

İntercept(sabit değer), HiperlipdemiVar, Kırgınlık ve YorgunlukVar, MetforminVar, İnsülin bağımlı olmayan diyabetes mellitüsVar, Eklem ağrısıVar, Demir eksikliği anemileriVar ve Serobrovasküler hastalıklarVar istatistiksel olarak anlamlı çıkmıştır.

Model kurulurken modele istatistiksel olarak etkisi olmayan değişkenler çıkartılıp model tekrar kurulabilir. Model bu şekilde de denenmiştir. Değişim olmadığı görüldüğü için ilk kurulan model üzerinden analize devam edilmiştir.

β katsayılarına göre lojistik regresyon denklemi şu şekilde oluşur:

$$log\frac{\pi}{(1-\pi)} = 4.53 - 0.31\beta_{Yas} + 0.24\beta_{CinsiyetKadın} - 0.007\beta_{HipertansiyonVar} - 0.66\beta_{HiperlipidemiVar} - 0.27\beta_{MenopozVar} + 0.63\beta_{HbA1c} + 0.07\beta_{Kreatin} + 0.85\beta_{Total.Kolesterol} - 0.09\beta_{HDL} - 0.74\beta_{LDL} + 1.28\beta_{Kırgınlık\ ve\ YorgunlukVar} - 0.422\beta_{MetforminVar} - 1.37\beta_{İnsülin\ Bagımlı\ Diyabetes\ MellitüsVar} - 0.69\beta_{Gastro\ Ozofajial\ Reflu\ HastalıgıVar} - 1.16\beta_{Eklem\ AgrısıVar} - 4.25\beta_{Demir\ Eksikligi\ AnemileriVar} + 0.16\beta_{Vitamin\ B12\ Eksikligi\ AnemisiVar} - 0.13\beta_{Aterosklerotik\ Kardiyovaskuler\ HastalıkVar} + 0.20\beta_{İnsülin\ Bagımlı\ Diyabetes\ MellitüsVar} - 1.73\beta_{Serebrovaskuler\ HastalıklarVar} + 0.14\beta_{OsteoporozVar}$$

β katsayılarında + - işareti, diyabetik polinöropatinin modele pozitif ya da negatif katkısını gösterir. β katsayılarını sıraladığımızda en büyük değere sahip olan Demir eksikliği anemileri değişkeni modele en fazla katkı veren değişkendir.

Tablo 25: Lojistik regresyon güven aralıkları

	2.5 %	97.5 %
(Intercept)	3.42802464	5.636025331
Yas	-1.47337580	0.837253793
CinsiyetKadın	-0.08536364	0.583806327
HipertansiyonVar	-0.35866516	0.343679713
HiperlipidemiVar	-1.01387519	-0.317406167
MenopozVar	-1.33413066	0.776817023
HbA1c	-0.47230087	1.734711096
Kreatinin	-3.31119127	3.454885166
Total.Kolesterol	-1.80137520	1.973213261

HDL	-1.62769404	1.446415906
LDL	-1.95660415	0.458566495
Kirginlik.ve.YorgunlukVar	0.21019554	2.353556676
MetforminVar	-0.77891978	-0.066063039
Insulin.Bagimli.Olmayan.Diyabetes.MellitusVar	-1.76396634	-0.992041171
Gastro.Ozofajial.Reflu.HastaligiVar	-1.39354010	0.006718062
Eklem.AgrisiVar	-2.18138358	-0.152129457
Demir.Eksikligi.AnemileriVar	-6.78895027	-1.716186289
Vitamin.B12.Eksikligi.AnemisiVar	-0.55568168	0.894150182
Aterosklerotik.Kardiyovaskuler.HastalikVar	-1.19388155	0.926122252
Insulin.Bagimli.Diyabetes.MellitusVar	-1.85067123	2.265562205
Serebrovaskuler.HastaliklarVar	-2.74234973	-0.731377524
OsteoporozVar	-1.10783027	1.397690066

Değişkenleri %95 güven aralığındaki değişimi Tablo 25'te gösterilmiştir.

Modelin hedef niteliği ne kadar tahminleyebildiğini görmek için kontenjans tablosu oluşturulmuştur. Bu matris yardımıyla bulunan performans değerlendirme ölçütleri Tablo 26'da verilmiştir.

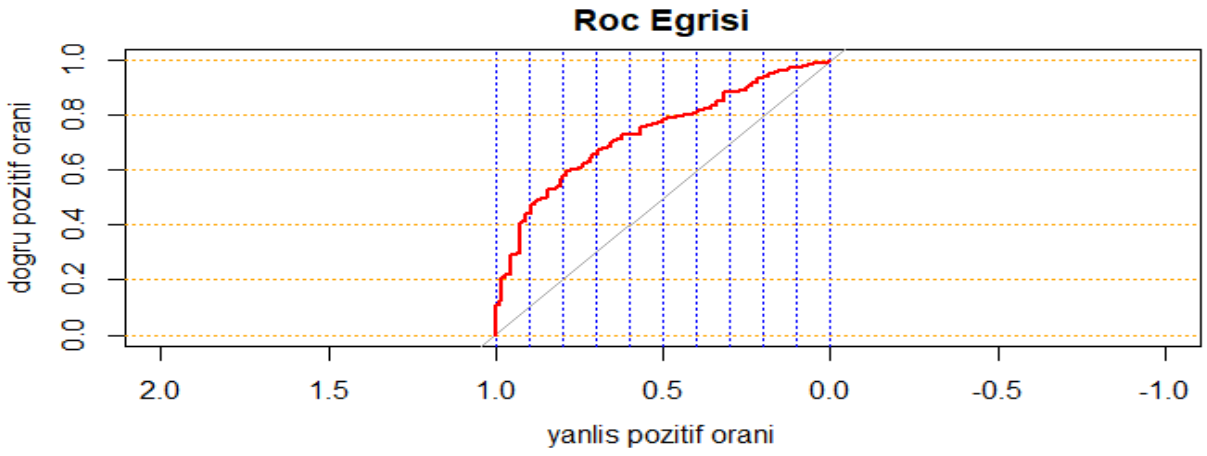
Tablo 26: Lojistik regresyon kontenjans tablosu

Lojistik Regresyon	Gerçek		
Tahmin		DPvar	DPyok
	DPvar	6	33
	DPyok	108	2816

Tablo 27: Lojistik regresyon performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	**Lojistik Regresyon**
Doğruluk oranı	% 95.2
Hata oranı	%4.7
TPR(Duyarlılık oranı)	%5.2
SPC(Belirleyicilik oranı)	%98.8
PPV(Pozitif öngörü oranı)	%15.3
NPV(Negatif öngörü oranı)	%96.3
FPR(Yanlış pozitif oranı)	%1.1
FNR(Yanlış negatif oranı)	%94.7
F-ölçütü	%7.8

Şekil 35: Roc eğrisi



Lojistik regresyon uygulaması için diyabetik polinöropati değişkenin tanımlanmasında Hiperlipidemi, kırgınlık ve yorgunluk, metformin, tip2 diyabet, eklem ağrısı, demir eksikliği anemileri ve serobrovasküler hastalıkların belirleyici değişkenler olduğu görülmektedir.

Kontenjans tablosu incelendiğinde, gerçekte diyabetik polinöropati olan hastaları, modelde de diyabetik polinöropati hastasıdır diye tahmin ettiği 6 kişi vardır. Doğru pozitif değeri 6'dır.

Gerçekte diyabetik polinöropati hastalığı olmayan, tahminde diyabetik polinöropati hastasıdır çıkan 33 kişi vardır. Yanlış pozitif yani tip 1 hata değeri 33'tür.

Gerçekte diyabetik polinöropati hastası olan, tahminde diyabetik polinöropati hastası değildir çıkan 108 kişi vardır. Yanlış negatif yani tip 2 hata değeri 108'dir.

Gerçekte diyabetik polinöropati hastalığı olmayan, tahminde de diyabetik polinöropati hastalığı yoktur çıkan 2816 kişi vardır. Doğru negatif değeri 2816'dır.

Lojistik regresyon algortiması için doğruluk oranı 0.952 ve hata oranı 0.047 çıkmıştır.

Roc eğrisinde "yanlış pozitiflik oranı azalsın, doğru pozitiflik oranı artsın" istenir. Şekil 35 incelendiğinde roc eğrisinin doğru pozitif oranı kısmında artış ve değişim gösterediği görülmektedir.

4.9. C4.5 Algoritması

C4.5 algoritması bir karar ağacı algoritmasıdır. Değişkenleri ağaç şeklinde dallanma yaparak sınıflandırır. C4.5 karar ağacı algoritması uygulanmadan önce veri setinin yapısı incelenmiştir. Değişkenler nümerik ve faktör şeklinde atanmıştır. Sınıflandırma algoritması olduğu için veri seti eğitim ve test veri seti olarak ayrılmıştır. Diğer algoritmalar ile bütünlük oluşturması açısından %60 eğitim veri seti, %40 test veri seti olarak ayrım yapılmıştır.

Uygulamanın yapılabilmesi için R programlamaya RWeka paketi yüklenmiş ve kütüphaneden çağırılmıştır. Paketin içindeki J48() fonksiyonu C4.5 karar ağacı algoritması çözümünde kullanılmıştır.

Eğitim veri setine uygulanan karar ağacı algoritması sonuçları verilmiştir.

```
=== Summary ===

Correctly Classified Instances        4282               96.3113 %
Kappa statistic                          0.0759
Mean absolute error                      0.071
Root mean squared error                  0.1885
Relative absolute error                 95.7912 %
Root relative squared error             97.9996 %
Total Number of Instances             4446

=== Confusion Matrix ===

    a    b   <-- classified as
    7  164 |    a = DPvar
    0 4275 |    b = DPyok
```

Burada correctly classified instances doğru yerleşen tahmin sayısıdır. Bunun toplam 4446 kişi içinden 4282 kişi olduğu gözükmektedir ve %96.3 doğruluk oranına sahiptir. Modelin oluşturduğu ağaç şu şekildedir:

```
J48 pruned tree
------------------

Demir.Eksikligi.Anemileri = Yoksa: DPyok (4417.0/164.0)
Demir.Eksikligi.Anemileri = Varsa
|   Serebrovaskuler.Hastaliklar = Yoksa
|   |   Gastro.Ozofajial.Reflu.Hastaligi = Yoksa
|   |   |   LDL <= 89.4: DPvar (3.0)
|   |   |   LDL > 89.4
|   |   |   |   HDL <= 46
|   |   |   |   |   Kreatinin <= 0.69: DPyok (2.0)
|   |   |   |   |   Kreatinin > 0.69: DPvar (4.0)
|   |   |   |   HDL > 46: DPyok (11.0)
|   |   Gastro.Ozofajial.Reflu.Hastaligi = Varsa: DPyok (4.0)
|   Serebrovaskuler.Hastaliklar = Varsa: DPyok (5.0)

Number of Leaves  :    7

Size of the tree :     13
```

Ağaç yapısı incelenir. Number of leaves yani yaprak sayısı 7 tanedir. Yapraktan sonra parantez içinde verilen değerler o kategoriye ait doğru ve yanlış sınıflandırmayı açıklar.

Örneğin;

Demir.Eksikligi.Anemileri = Yoksa: DPyok (4417.0/164.0)

Burada demir eksikliği olmayan hastalarda Diyabetik Polinöropati yoktur kuralı elde edilmiştir. Parantez ile ifade edilen, bu kategoride 4417 örneğin doğru sınıflandırıldığı, 164 örneğin yanlış sınıflandırıldığıdır.

Karar ağacından elde edilen kurallar şu şekildedir.

KURAL1:

EĞER hastada, demir eksikliği anemileri yoksa Diyabetik Polinöropati yoktur.

KURAL2:

EĞER hastada, demir eksikliği anemileri varsa VE serobrovasküler hastalıklar varsa Diyabetik Polinöropati yoktur.

KURAL3:

EĞER hastada, demir eksikliği anemileri varsa VE serobrovasküler hastalıklar yoksa VE gastro özofajial reflü hastalığı varsa Diyabetik Polinöropati yoktur.

KURAL4:

EĞER hastada, demir eksikliği anemileri varsa VE serobrovasküler hastalıklar yoksa VE gastro özofajial reflü hastalığı yoksa VE LDL <= 89.4 İSE Diyabetik Polinöropati vardır.

KURAL 5:

EĞER hastada, demir eksikliği anemileri varsa VE serobrovasküler hastalıklar yoksa VE gastro özofajial reflü hastalığı yoksa VE LDL>89.4 VE HDL>46 İSE Diyabetik Polinöropati yoktur.

KURAL6:

EĞER hastada, demir eksikliği anemileri varsa VE serobrovasküler hastalıklar yoksa VE gastro özofajial reflü hastalığı yoksa VE LDL>89.4 VE HDL <= 46 VE Kreatinin <=0.69 İSE Diyabetik Polinöropati yoktur.

KURAL7:

EĞER hastada, demir eksikliği anemileri varsa VE serobrovasküler hastalıklar yoksa VE gastro özofajial reflü hastalığı yoksa VE LDL>89.4 VE HDL <= 46 VE Kreatinin>0.69 İSE Diyabetik Polinöropati vardır.

Kontenjans tablosu Tablo 28'de görülmektedir.

Tablo 28: C4.5 kontenjans tablosu

C4.5	Gerçek		
Tahmin		DPvar	DPyok
	DPvar	3	4
	DPyok	111	2845

Tablo 29: C4.5 performans değerlendirme ölçütleri

Performans Değerlendirme Ölçütleri	**C4.5**
Doğruluk oranı	% 96.1
Hata oranı	%3.8
TPR(Duyarlılık oranı)	%2.6
SPC(Belirleyicilik oranı)	%99.8
PPV(Pozitif öngörü oranı)	%42.8
NPV(Negatif öngörü oranı)	%96.2
FPR(Yanlış pozitif oranı)	%0.1
FNR(Yanlış negatif oranı)	%97.3
F-ölçütü	%4.9

C4.5 karar ağacı algoritması kontenjans tablosu sonuçları incelendiğinde, doğru pozitif değerinin 3 çıktığı görülmektedir. Yani gerçekte diyabetik polinöropati olan hastalardan, model tahminde diyabetik polinöropati hastasıdır diye 3 kişiyi doğru tahmin etmiştir.

Gerçekte diyabetik polinöropati hastalığı olmayan kişileri model, diyabetik polinöropati hastasıdır şeklinde tahmin etmiştir. Yanlış pozitif değeri 4'tür.

Gerçekte diyabetik polinöropati hastalığı olan kişileri model, diyabetik polinöropati hastalığı yoktur şeklinde tahmin etmiştir. Yanlış negatif değeri 111'dir.

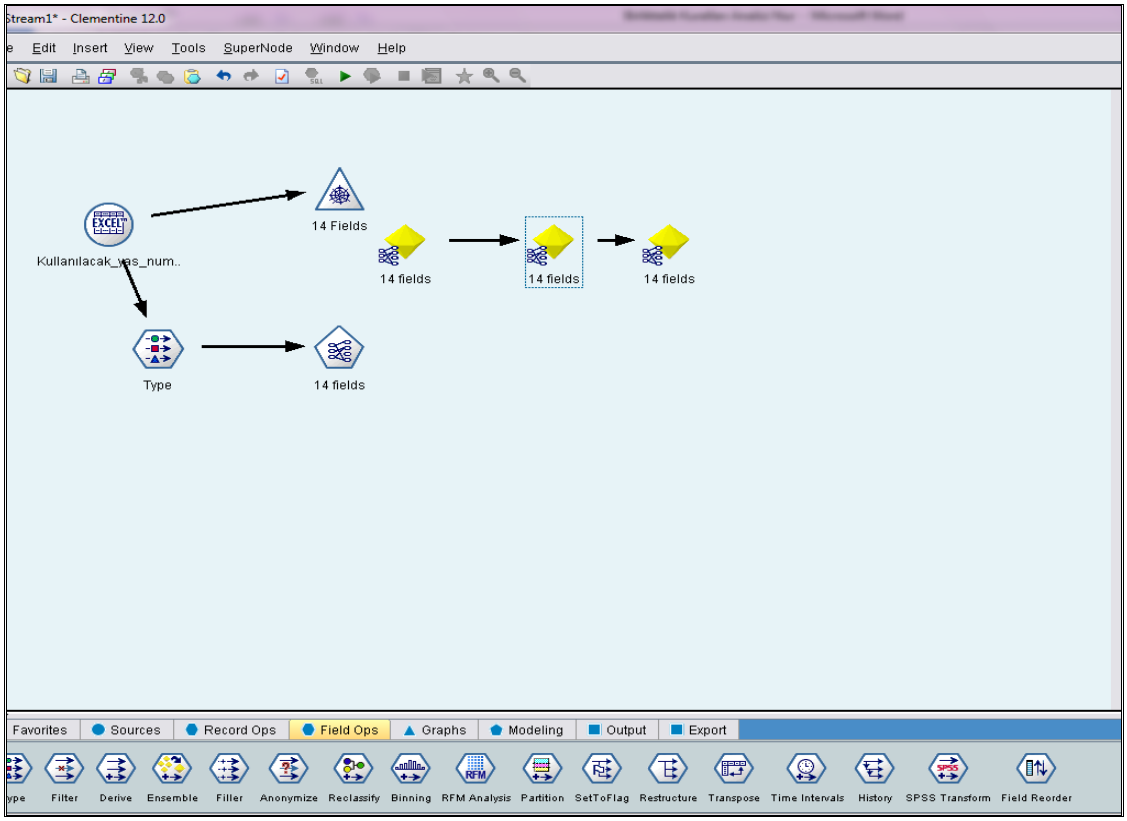
Gerçekte Diyabetik polinöropati olmayan hastaları model diyabetik polinöropati değildir şeklinde, 2845 kişide doğru tahmin etmiştir. Doğru negatif değeri 2845'tir.

Modelin doğruluk oranı 0.961 ve hata oranı 0.038 çıkmıştır.Kurallarda ortaya çıkan belirleyici değişkenler, demir eksikliği anemileri, serobrovasküler hastalıklar, gastro özofajial hastalıklar, HDL, LDL ve Kreatinin değerleridir.

4.10. Birliktelik Kuralları

Minimum destek oranı %15 ve minimum güven oranı %50 olmak üzere uygulanan Apriori algoritması ile kurulan model Şekil 36'da ve elde edilen kural setleri Tablo 30'da görülmektedir.

Şekil 36: Apriori algoritması ile kurulan model



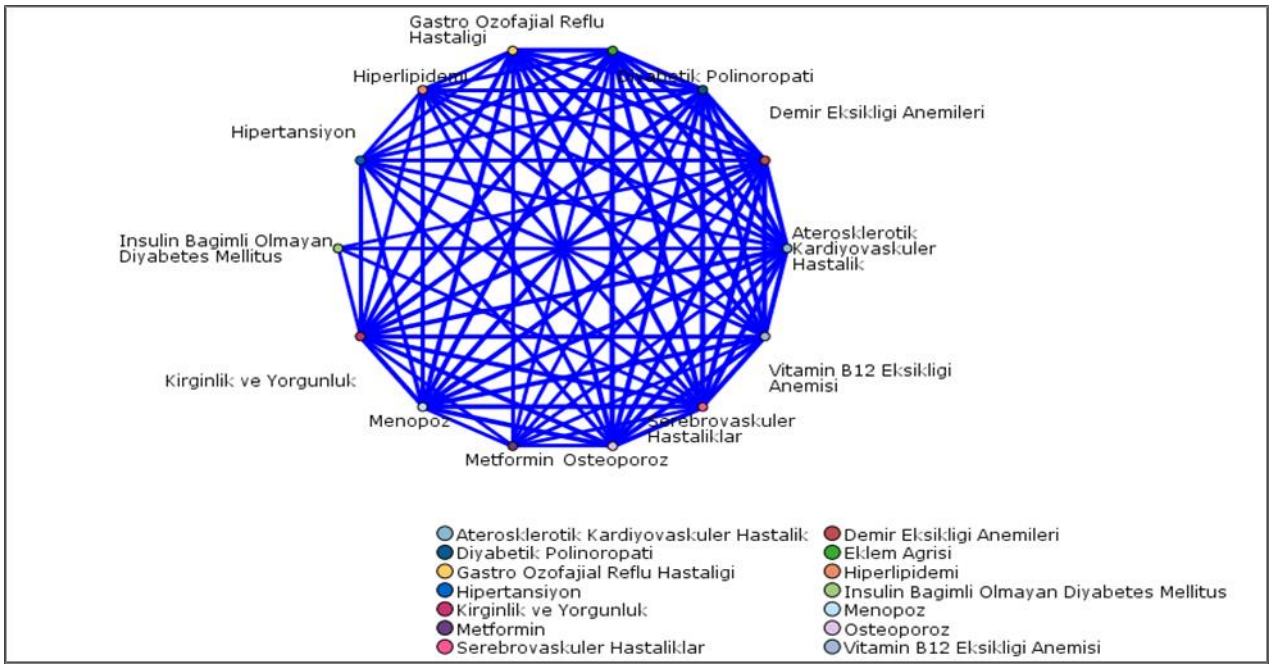
Tablo 30: Elde edilen kural setleri

Consequent	Antecedent	Support %	Confidence %
Insulin Bagimli Olmayan Diyabetes Mellitus	Hipertansiyon Metformin	16,17	69,533
Insulin Bagimli Olmayan Diyabetes Mellitus	Metformin	36,577	67,97
Insulin Bagimli Olmayan Diyabetes Mellitus	Hiperlipidemi Metformin	15,103	67,113
Insulin Bagimli Olmayan Diyabetes Mellitus	Hipertansiyon	26,373	66,888
Metformin	Hiperlipidemi Insulin Bagimli Olmayan Diyabetes Mellitus	15,535	65,248
Insulin Bagimli Olmayan Diyabetes Mellitus	Hiperlipidemi	23,822	65,212
Metformin	Hipertansiyon Insulin Bagimli Olmayan Diyabetes Mellitus	17,641	63,734
Metformin	Insulin Bagimli Olmayan Diyabetes Mellitus	39,182	63,452
Metformin	Hiperlipidemi	23,822	63,399
Metformin	Hipertansiyon	26,373	61,31
Hipertansiyon	Hiperlipidemi Insulin Bagimli Olmayan Diyabetes Mellitus	15,535	50,391

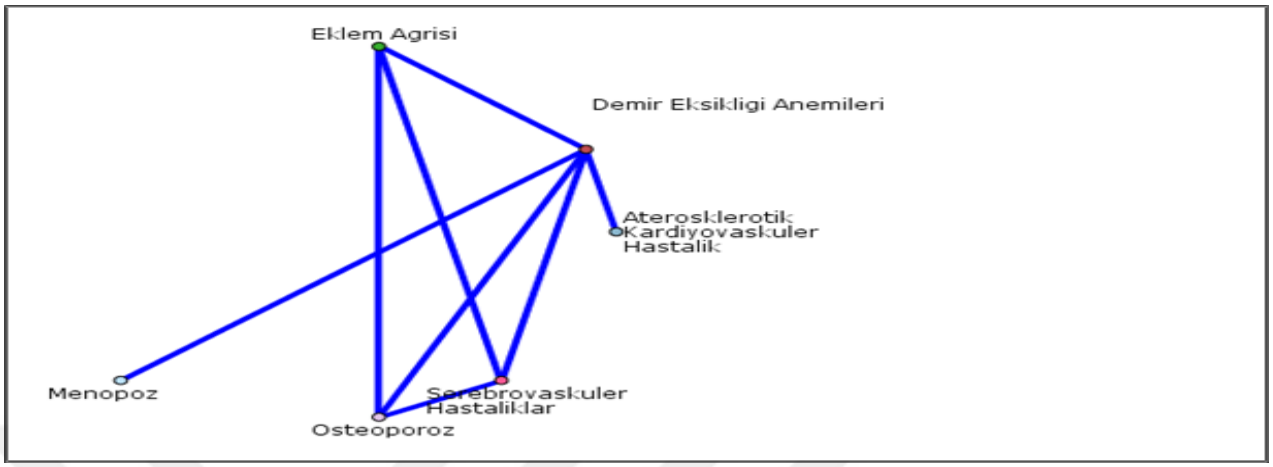
- Hipertansiyonu olan ve metformin ilacı kullanan hastaların %69,533'ü İnsülin bağımlı olmayan diyabetes mellitustur. Destek oranı %16,17'dir.
- Hipirlipidemisi olan hastaların %63,399'u metformin kullanmaktadır. Destek oranı %23,822'dir.

Değişkenler arasındaki ilişkiler şekil 37'deki web grafiğiyle, sıklığı yüksek olan ilişkiler ise Şekil 38 ve 39 ile incelenmiştir.

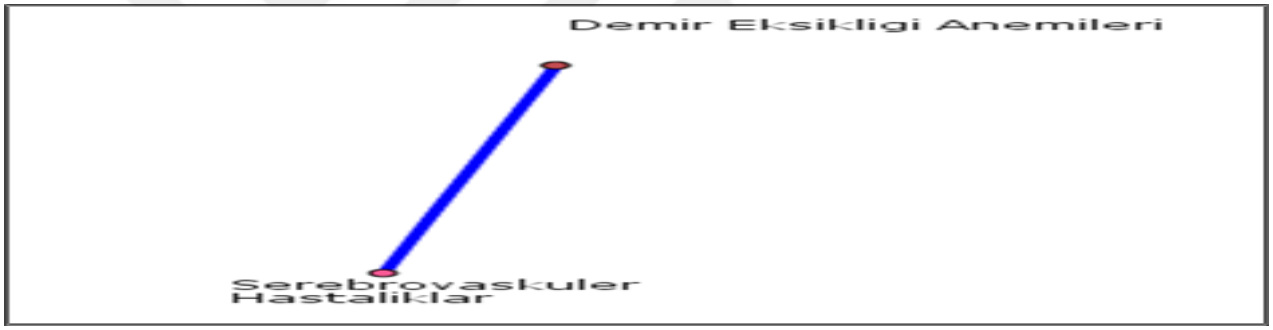
Şekil 37: Değişkenler arası ilişkiler



Şekil 38: Sıklığı yüksek olan ilişkiler 1



Şekil 39: Sıklığı yüksek olan değişkenler 2



Minimum destek oranı % 0 ve minimum güven oranı %50 olmak üzere Birliktelik kuralları analizi sonucu elde edilen kural setleri Tablo 31 ve Tablo 32'de görülmektedir.

Güven oranı %100 olan kural setleri sıralanmak istenmiştir.

- Diyabetik polinöropatisi olan, aynı zamanda kırgınlık ve yorgunluğu olan hastaların %100'ü İnsülin bağımlı olmayan diyabetes mellitustur. Destek oranı % 0,297'dir.
- Serebrovasküler hastalıkları olan ve metformin kullanan hastaların %100'ü İnsülin bağımlı olmayan diyabetes mellitustur. Destek oranı % 0,243'tür.

Tablo 31: Kural setleri 1

Consequent	Antecedent	Support %	Confidence %
Insulin Bagimli Olmayan Diyabetes Mellitus	Diyabetik Polinoropati Kirginlik ve Yorgunluk	0,297	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Serebrovaskuler Hastaliklar Metformin	0,243	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Aterosklerotik Kardiyovaskuler Hastalik Diyabetik Polinoropati	0,202	100,0
Metformin	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi	0,202	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi	0,202	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi Metformin	0,202	100,0
Metformin	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi Insulin Bagimli Olmayan Diyabetes Mellitus	0,202	100,0
Metformin	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi Hipertansiyon	0,175	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi Hipertansiyon	0,175	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi Hipertansiyon Metformin	0,175	100,0
Metformin	Aterosklerotik Kardiyovaskuler Hastalik Vitamin B12 Eksikligi Anemisi Hipertansiyon Insulin Bagimli Olmayan Diyabetes Mellitus	0,175	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Serebrovaskuler Hastaliklar Hipertansiyon Metformin	0,162	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Demir Eksikligi Anemileri Diyabetik Polinoropati	0,148	100,0

Tablo 32: Kural setleri 2

Consequent	Antecedent	Support %	Confidence %
Hipertansiyon	Demir Eksikligi Anemileri Serebrovaskuler Hastaliklar	0,067	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Demir Eksikligi Anemileri Serebrovaskuler Hastaliklar	0,067	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Demir Eksikligi Anemileri Menopoz	0,027	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Demir Eksikligi Anemileri Diyabetik Polinoropati	0,148	100,0
Metformin	Demir Eksikligi Anemileri Vitamin B12 Eksikligi Anemisi	0,094	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Demir Eksikligi Anemileri Kirginlik ve Yorgunluk	0,094	100,0
Hiperlipidemi	Serebrovaskuler Hastaliklar Osteoporoz	0,027	100,0
Hipertansiyon	Serebrovaskuler Hastaliklar Osteoporoz	0,027	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Serebrovaskuler Hastaliklar Osteoporoz	0,027	100,0
Hipertansiyon	Serebrovaskuler Hastaliklar Aterosklerotik Kardiyovaskuler Hastalik	0,027	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Serebrovaskuler Hastaliklar Aterosklerotik Kardiyovaskuler Hastalik	0,027	100,0
Diyabetik Polinoropati	Serebrovaskuler Hastaliklar Gastro Ozofajial Reflu Hastaligi	0,013	100,0
Hipertansiyon	Serebrovaskuler Hastaliklar Gastro Ozofajial Reflu Hastaligi	0,013	100,0
Insulin Bagimli Olmayan Diyabetes Mellitus	Serebrovaskuler Hastaliklar Gastro Ozofajial Reflu Hastaligi	0,013	100,0

4.11. Genel Değerlendirme ve Model Seçimi

Diyabetik polinöropati hastalığının öngörülebilmesi için sırasıyla KNN, Naive Bayes, Lojistik Regresyon ve C4.5 Karar Ağaçları algoritmaları kullanılmış ve bu algoritmaların performans değerlendirme ölçütleri kıyaslanmıştır.

Tablo 33: Genel değerlendirme ve model seçimi

	Doğruluk	Hata	TPR	SPC	PPV	NPV	FPR	FNR	F-ölçütü
KNN(k=1)	0.934	0,06	0.149	0.96	0.147	0.965	0.03	0.85	0.148
KNN(k=10)	0.9615	0.039	0.08	0.998	0.25	0.961	0.001	0.991	0.016
Naive Bayes Algoritması	0.949	0.05	0.043	0.985	0.108	0,962	0.013	0.956	0.062
Lojistik Regresyon	0.952	0.04	0.05	0.98	0.15	0.96	0.011	0.94	0.07
C4.5 Karar Ağacı Algoritması	0.9611	0.0388	0.026	0.998	0.428	0.962	0.001	0.973	0.049

Belirlenen performans değerlendirme oranlarına göre KNN k=10 ve C4.5 karar ağacı algoritması sonuçları birbirine çok yakın çıkmıştır. Doğruluk, hata oranı ve F ölçütü baz alınırsa en iyi performans veren algoritma C4.5 karar ağacı algoritmasıdır denilebilir.

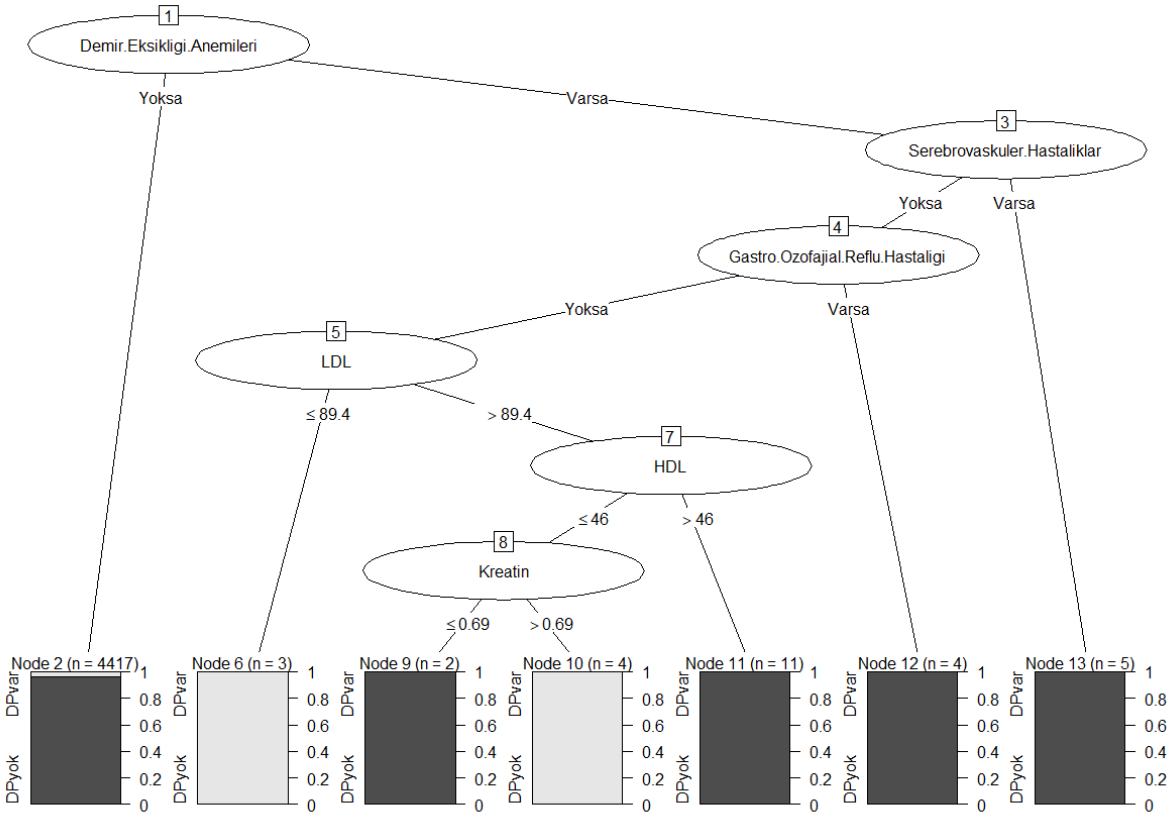
Lojistik regresyonda anlamlı olan değişkenler HiperlipidemiVar, MetforminVar, İnsülin bağımlı olmayan diyabetüs mellitüsVar(tip2), Eklem ağrısıVar, Demir eksikliği anemileriVar ve Serobrovasküler hastalıklarVar çıkmıştır.

C4.5 karar ağacı algoritması oluşturulurken modelin kullandığı belirleyici değişkenlerde demir eksikliği anemileri, serobrovasküler hastalıklar, gastro özofajial reflü hastalığı, LDL, HDL ve Kreatinindir.

Lojistik regresyon uygulaması ve C.45 karar ağacı uygulaması incelendiğinde demir eksikliği anemileri ve serobrovasküler hastalıkların belirleyici değişkenler olduğu görülmektedir.

Ayrıca birliktelik kuralları incelendiğinde demir eksikliği anemileri ve serobrovasküler hastalıklar belirleyici değişkenler olarak çıkmaktadır.

Şekil 40: C4.5 karar ağacı



SONUÇ

Diyabet kronik bir metabolizma bozukluğudur. İnsülin hormonunun eksik olması veya yeterince salgılanmasına karşın vücutta kullanılmaması sonucu oluşmaktadır. Diyabet hastalığına "Diabetes Mellitus" veya "Şeker Hastalığı" adı verilmektedir. Diyabet tipleri "Tip 1, Tip 2, Gebelik diyabeti ve diğer" olmak üzere gruplara ayrılmaktadır. Diyabet Dünya Sağlık Örgütü'ne göre geleceğin en yaygın görülecek olan hastalıklarından biridir. Diyabet yaşam kalitesini azaltmakta ve sağlık harcamalarını arttırmakta olan bir hastalıktır. Uluslararası Diyabet Federasyonu'nun yayınladığı diyabet atlasına göre, 2015 yılında 415 milyon diyabetli varken, 2040 yılında bu rakamın 642 milyona ulaşması tahmin edilmektedir. Bu durum diyabet ile ilişkili hastalıklara yapılacak harcamanın da artacağını göstermektedir.

Diyabetin dünya üzerinde yaygın olarak görülmesi, tedavi edilmediğinde veya tedavide geç kalınması durumunda organlara zarar vermesi, önemli bir tehdittir. Diyabetin dikkatle yönetilmesi önemli ve hayatı tehdit eden komplikasyonların önlenmesini de sağlamaktadır. Konunun önemi sebebiyle bu tez çalışmasında, diyabetik polinöropatiyi etkileyen faktörlerin belirlenmesi hedeflenmiştir.

Çalışmanın birinci bölümünde veri madenciliği kavramı, veri madenciliği süreci, veri madenciliğinin kullanıldığı alanlar ve veri madenciliği teknikleri ele alınmıştır.

İkinci bölümünde sağlık hizmetlerinde veri madenciliği uygulamalarına yer verilmiştir.

Tezin üçüncü bölümünde diyabet hastalığı hakkında bilgi verilmiştir. Tezin uygulama bölümünde kullanılan değişkenler açıklanmıştır.

Dördüncü bölüm uygulama bölümüdür. Veri madenciliği sürecine sadık kalınarak, uygulama aşamaları anlatılmış ve uygulamada kullanılan tekniklere yer verilmiştir. K-nn, Naive-Bayes, Lojistik regresyon, C4.5 ve Birliktelik kuralları analizi kullanılmıştır.

K-nn algoritması k değeri 1'den 10'a kadar değer verilerek tahminlenmeye çalışılmıştır. Bu algoritma için performans değerlendirme ölçütlerine bakıldığında k=8, k=9 ve k=10 algoritmalarının en iyi sonuçları verdiği gözlenmiştir. k=8 ve k=9

değerleri ile tahminlenen modeller birbirinin aynısı olduğu için k=10 algoritma sonuçlarının performans ölçüm değerlendirmesi bakımından en iyi sonuçları verdiği söylenebilir. Bunu göre doğruluk oranı %96 ve hata oranı %3.9 dur.

K-nn algoritması kontenjans tablosu incelendiğinde doğru pozitif, yanlış pozitif, yanlış negatif ve doğru negatif tahmin değerleri elde edilmektedir. Buna göre matrisler incelendiğinde k=1 değeri ile tahminlenen modelde gerçekte diyabetik polinöropati hastası olan hastaların, modelin diyabetik polnöropati hastasıdır diye tahmin ettiği 17 hasta olduğu görülmektedir. Yanlış pozitif olarak tahmin ettiği yani gerçekte diyabetik polinöropatisi olmayan hastalardan modelin diyebetik polinöropatisi vardır olarak tahmin ettiği 98 hasta olduğu görülmektedir. Yanlış negatif değerleri yani gerçekte diyabetik polinöropati olan hastaları modelin diyabetik polinöropati hastası değildir olarak yanlış tahmin ettiği 97 hasta vardır. Doğru negatif değeri ise gerçekte diyabetik polinöropati olmayan hastaların modelde de diyabetik polinöropati hastası değildir şeklinde tahmin edilmesidir. Bu sayı da 2751 kişidir.

K-nn algoritması k değerleri karışıklar matrisi için incelendiğinde k=2 için doğru pozitif tahmin sayısının 8' e düştüğü görülmektedir. k=3 değeri için doğru pozitif değeri 5 olmuştur ve incelendiğinde k=4, k=5, k=6, k=7, k=8, k=9 ve k=10 değerleri ile tahminlenen matrislerde doğru pozitif değerinin 1 olduğu görülmektedir.

Buna göre düşünüldüğünde model doğruluk oranı en yüksek ve hata oranını en düşük veren k=10 değeri mi, yoksa gerçek pozitif değeri 17 kişiyi doğru tahmin eden k=1 değeri mi alınmalıdır. Gerçek hasta verisi ile çalışıldığından ve sadece bir hastayı değil de daha çok hastayı doğru tahmin etmek istenildiğinden k=1 ile çözümlenen knn algoritması daha iyi performans göstermektedir şeklinde yorumlanabilir.

Naive Bayes algoritması sonuçları incelendiğinde gerçekte diyabetik polinöropati hastası olan hastalardan modelin 5 kişiyi diyabetik polinöropatisi vardır şeklinde doğru tahmin ettiği görülmektedir. Doğru pozitif değeri 5'tir.

Gerçekte diyabetik polinöropatisi olmayan hastalardan, modelin diyabetik polinöropatisi vardır şeklinde, "yanlış pozitif " tahmin ettiği 41 kişi vardır.

Gerçekte diyabetik polinöropatisi olan hastalardan, modelin diyabetik polinöropatisi yoktur şeklinde, "yanlış negatif " tahmin ettiği 109 kişi vardır.

Gerçekte diyabetik polinöropati hastası olmayan hastaların, modelde de diyabetik polinöropati hastası değildir şeklinde tahminlendiği 2808 kişi vardır. Doğru negatif değeri 2808'dir.

Naive Bayes algoritması performans ölçümlerine bakıldığında doğruluk oranı %94.9 ve hata oranı %5'tir.

Lojistik regresyon analizi sonuçları incelendiğinde modelin "doğru pozitif" tahmin ettiği 6 kişi, "yanlış pozitif" tahmin ettiği 33 kişi, "yanlış negatif" tahmin ettiği 108 kişi ve "doğru negatif" tahmin ettiği 2816 kişi vardır.

Modelin performans değerlendirme ölçütlerinde doğruluk oranı %95.2 ve hata oranı %4.7 çıkmıştır.

Modelin beta katsayıları incelendiğinde sabit değişken, hiperlipidemi var, kırgınlık ve yorgunluk var, metformin var, insülin bağımlı olmayan diabetes mellitüs var, eklem ağrısı var, demir eksikliği anemileri var ve serobrovasküler hastalıklar var değişkenleri anlamlı çıkmıştır. Katsayı değer sonuçlarına göre modele en fazla katkıyı veren değişken demir eksikliği anemileri olmuştur.

Bu değişkenlerin diyabetik polinöropati hastalığının varlığı için lojistik regresyon algoritmasında belirleyici değişkenler olduğu söylenebilir.

C4.5 karar ağacı algoritması karışık matrisi sonuçları incelendiğinde "doğru pozitif" 3 kişi, "yanlış pozitif" 4 kişi, "yanlış negatif" 111 kişi ve "doğru negatif" 2845 kişi tahmin edilmiştir.

Modelin performans ölçümlerine bakıldığında doğruluk oranı %96.1 ve hata oranı %3.8 çıkmıştır.

C4.5 karar ağacı algoritması ile ortaya çıkan 7 tane kural vardır. Bu kurallar içerisindeki belirleyici değişkenler demir eksikliği anemileri, serobrovasküler hastalıklar, gastro özofajial reflü hastalığı, LDL, HDL ve kreatinin değişkenidir.

Bunlara göre demir eksikliği anemisi olmayan hastada diyabetik polinöropati de yoktur. Demir eksikliği olan hastalardan serobrovasküler hastalıkları olan hastalarda diyabetik polinöropati yoktur.

Demir eksikliği olan, serobrovasküler hastalıkları olmayan ve gastro özofajial reflü hastalığı olan hastalarda diyabetik polinöropati yoktur.

Demir eksikliği olan, serobrovasküler hastalıkları olmayan, gastro özofajial reflü hastalığı olmayan ve LDL≤ 89.4 ise diyabetik polinöropati vardır. LDL kötü kolesterol olarak adlandırılmaktadır ve fazla yüksek olması istenmez. Literatürde 100 mg/dl'nin üzeri diyabet hastalığı için optimal seviyeyi aştığını gösterir. Bu çalışmada hedef nitelik diyabetik polinöropati değişkenidir. Model bize LDL değişkeni 89.4 mg/dl altında kalırsa diğer değişkenler ile birlikte diyabetik polinöropati vardır sonucu vermektedir.

Demir eksikliği olan, serobrovasküler hastalıkları olmayan, gastro özofajial reflü hastalığı olmayan ve LDL>89.4 ve HDL > 46 mg/dl ise diyabetik polinöropati yoktur. Buradaki HDL değişkeni iyi kolesterol olarak adlandırılmaktadır. Diyabet hastalığı için kötü seviyeler 40 mg/dl'nin altında olmasıdır. Modelde HDL değişkeni 46mg/dl üzerinde çıkmıştır ve literatür ile uyumludur.

Demir eksikliği olan, serobrovasküler hastalıkları olmayan, gastro özofajial reflü hastalığı olmayan ve LDL>89.4 ve HDL ≤ 46 mg/dl ve Kreatinin değeri ≤0.69 ise diyabetik polinöropati yoktur. Kreatinin değeri literatürde 0.6≤kreatinin ≤1.2 normal kabul edilmektedir.0.6 mg/dl değeri altında kalan değerler düşük Kreatinin değeridir. Modelde de diyabetik polinöropati yoktur sonucu çıkmıştır, sonuçlar literatür ile uyumludur.

Demir eksikliği olan, serobrovasküler hastalıkları olmayan, gastro özofajial reflü hastalığı olmayan ve LDL>89.4 ve HDL ≤ 46 mg/dl ve Kreatinin değeri >0.69 ise diyabetik polinöropati vardır. LDL kolesterolün yüksek olması istenen bir durum değildir, kötü kolesterolü temsil eder. Bu kuralda 89.4 mg/dl üzeri değerleri ifade ediyor. HDL kolesterol iyi kolesteroldür ve 40 mg/dl' nin altında olması istenir. Bu kuralda 46 mg/dl altında olmalı ifadesi çıkmış. Kreatinin değeri 0.69 ile normal

değerleri arasında gözükmektedir. Bu şartları sağlayan hastalarda diyabetik polinöropati vardır.

Birliktelik kuralları hangi değişkenlerin birlikte hareket ettiğini verir. Sonuçlar incelendiğinde insülin bağımlı olmayan diyabetes mellitüs yani tip 2 diyabet hastalığını, metformin ve hipertansiyonun etkilediği görülmektedir. Bulgular bu sonuçlar üzerinden dönmüştür.

Aynı zamanda demir eksikliği anemileri ve serobrovasküler hastalıklar belirleyici değişken çıkmıştır.

Tüm modeller birlikte değerlendirildiğinde performans ölçüm modelleri değerlendirme ölçütlerine göre en yüksek doğruluk ve en düşük hatayı veren C4.5 algoritması en uygun modeldir denilebilir.

Modellerin belirleyici değişkenlerine göre genel bir değerlendirme yapılırsa; lojistik regresyon algoritması sonuçlarında anlamlı değişken olarak demir eksikliği anemisi var çıkmıştır. C4.5 algortiması kurallarında da demir eksikliği yoksa diyabetik polinöropati yoktur sonucu çıkmıştır. Yine birliktelik kurallarına göre demir eksikliği anemisi belirleyici değişken çıkmıştır. Yani bu üç tekniğin verdiği sonuçlara göre demir eksikliği anemisi diyabetik polinöropati hastalığı üzerinde etkileyici değişkendir.

Yine lojistik regresyon algoritması sonuçlarına göre serobrovasküler hastalıklar var anlamlı değişken çıkmıştır. C4.5 algoritması kurallarında demir eksikliği olan, serobrovasküler hastalığı varsa diyabetik polinöropati yoktur demektedir. Birliktelik kuralarında serobrovasküler hastalık belirleyici değişken çıkmıştır. Bu üç yöntemde de serobrovasküler hastalık değişkeni belirleyici olmuştur yani diyabetik polinöropati hastalığının belirlenmesinde etkkileyici değişekendir denilebilir.

Bu çalışmayı diğer diyabet çalışmalarından ayıran unsur veri kümesi olarak diyabet şüphesi ile muayeneye gelen tüm hastaların alınmasıdır. Bunların arasından diyabetik polinöropati olanlar tahminlenmeye çalışılmış ve değişkenlerden hangilerinin belirleyici değişken olduğu tahmin edilmeye çalışılmıştır.

Gelecek çalışmalarda değişken sayısı arttırılıp azaltılırarak veya farklılaştırılarak daha farklı bulgular elde edilebilir. Uygulanan algoritma çıktılarında, kontenjans tablosu yanlış pozitif ve yanlış negatif değerleri yüksek çıkmıştır. Bu araştırmada gerçek veri seti ile bir hastalık teşhisini tahminlemeye çalışılmıştır. Gelecek çalışmalar için değişken sayısı arttırılarak bu değerlerin düşürülmesi sağlanabilir.

KAYNAKÇA

ADA: 2010	"Diagnosis And Classification Of Diabetes Mellitus", **Diabetes Care**, 33 (1): 62-69.
AKAY, A., DRAGOMİR, A., ve ERLANDSSON, B. E.: 2013	"A novel data-mining approach leveraging social media to monitor and respond to outcomes of diabetes drugs and treatment", **IEEE EMBS Special Topic Conference on Point-of-Care (POC) Healthcare Technologies: Synergy Towards Better Global Healthcare**, PHT2013, 264–266.
AKYÜZ, F. VE SOYER, Ö.M.: 2017	"Gastroözofageal Reflü Hastalığının Gelişmesi Açısından Risk Faktörü Oluşturan Hastalıklar Nelerdir?", **Turk J Gastroenterol**, 28(1), 544-547.
AL-NOZHA, M. M. ve ark.: 2004	"Diabetes mellitus in Saudi Arabia", **Saudi Med J**, 25(11), 1603–1610.
ALIC, B., GURBETA, L., ve BADNJEVIC, A.: 2017	"Machine Learning Techniques For Classification Of Diabetes And Cardiovascular Diseases", **2017 6th Mediterranean Conference on Embedded Computing (MECO)**, (June), 1–4.
ALJUMAH, A. A., AHAMAD, M. G. ve 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: 2015	"Diagnosis And Classification Of Diabetes Mellitus", **Diabetes Care**, 38(1), 8–16.
AYDIN, İ., KARAKÖSE, M.,	"Zaman Serilerinde Veri Madenciliği ve Destek Vektör Makinalar Kullanan Yeni Bir Akıllı Arıza Sınıflandırma

AKIN, M.: 2008 Yöntemi", **Gazi Üniv. Müh. Mim. Fak. Der.**, 23(2), 431–440.

BAGDI, R., ve PATIL, P.: 2012 "Diagnosis of Diabetes Using OLAP and Data Mining Integration", **International Journal of Computer Science & Communication Networks**, 2(3), 314- 322.

BANAEE, H., AHMED, M. U., ve LOUTFİ, A.: 2013 "Data Mining For Wearable Sensors in Health Monitoring Systems: A Review Of Recent Trends and Challenges", **Sensors (Switzerland)**, 13(12), 17472–17500.

BAYRAK, G., SERKANT, Ö.B., YILMAZER, T.T. ve SUHER, M.: 2009 "Demir Eksikliği Anemisi ve HbA1c Düzeyi Arasındaki İlişki", **Turkish Medical Journal**, 3(1), 29-33.

BECK-NIELSEN H, ve ark.: 1995 "Pathophysiology of noninsülin-dependent diyabetes mellitus (NIDDM)", **Diyabetes Res Clin Pract 28Suppl,** 1, 13-25.

BERNDT, D. J., ve ark.: 2001 "Healthcare Data Warehousing And Quality Assurance", **Computer**, 34(12), 33-42.

BHRAMARAMBA, R., ve ark.: 2011 "Application of Data Mining Techniques on Diabetes Related Proteins", **International Journal of Diabetes in Developing Countries**, 31(1), 22–25.

BİRCAN, H.: 2004. "Lojistik Regresyon Analizi : T ı p Verileri Üzerine Bir Uygulama", **Kocaeli Ünv Sosyal Bilimler Ens. Derg.**, 2, 185–208.

BRAMER, M.: 2007 **Principles of Data Mining**, Canada, Springer.

BRAMER, M.: 2016 **Principles of Data Mining**, Canada, Springer.

BREAULT, J. L., GOODALL, C. R., ve FOS, P. J.: 2002 "Data Mining a Diabetic Data Warehouse", **Artificial Intelligence in Medicine**, 26(1–2), 37–54.

BROCKMANN, D., HUFNAGEL, L., ve GEISEL, T.: 2006 **Data Mining and Knowledge Discovery Handbook**, London, Springer.

CABENA, P., ve ark.: 1998 **Discovering Data Mining:From Concept to Implementation**, NJ, Prentice Hall.

CANHOTO, A. I., ve ARP, S.: 2017 "Exploring The Factors That Support Adoption and Sustained Use of Health and Fitness Wearables", **Journal of Marketing Management**, 33(1–2), 32–60.

CASSELMAN, J., ONOPA, N., ve KHANSA, L.: 2017 "Wearable Healthcare: Lessons From the Past and a Peek Into The Future", **Telematics and Informatics**, 34(7), 1011–1023.

CHARPENTIER, G. ve ark.: 2003 "Control of Diabetes and Cardiovascular Risk Factors in Patients with Type 2 Diabetes: A Nationwide French Survey", **Diabetes & Metabolism**, 29(2 Pt 1), 152–158.

CRAIG, M. E., ve ark.: 2007 "Diabetes Care, Glycemic Control, And Complications in Children with Type 1 Diabetes From Asia and The Western Pacific Region", **Journal of Diabetes and Its Complications**, 21(5), 280–287.

ÇINKIR, Ü.: 2011 "Diyabetik Nefropatili Hastalarda Vitamin D Tedavisinin Proteinüri Üzerine Etkisi", Uzmanlık Tezi, Adana, **T.C. Çukurova Üniversitesi Tıp Fakültesi İç Hastalıkları Anabilim Dalı**.

DAGLIATI, A., ve ark.: 2018 "Careflow Mining Techniques to Explore Type 2 Diabetes Evolution", **Journal of Diabetes Science and Technology**, 12(2), 251–259.

DEVI, M.R. ve SHYLA, J.M.: 2016 "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", **International Journal of Applied Engineering Research**, 11(1), 727-730.

DIABETES UK: 2009 "Diabetes in the UK 2009: Key Statistics on Diabetes", (Çevrimiçi) http://www.nationalschool.gov.uk/policyhub/news_item/diabetes_uk09.asp, 22. 07. 2018.

DURNA, Z.: 2002 Diyabet Sınıflandırması ve Tanı Kriterleri, **Diyabet Hemşireliği**, Ed. by. Semra Erdoğan, İstanbul, Diyabet Hemşireliği Derneği.

EARLEY, S.: 2015 "The Promise of Healthcare Analytics", **IT Professional**, 17(2), 7–9.

EHRLICH A. ve SCHROEDER C.L.: 2009 **Medical Terminology for Health Professions**, NY, Delmar Cangage.

FRITSCHI, C. ve QUINN, L.: 2010 "Fatigue in Patients with Diabetes: a Review", **J Psychosom Res.**, 69(1), 33-41.

GANGULI, B.: 2011 "A Review of Multivariate Longitudinal Data Analysis Income Affluence in Poland", **Statistical Methods in Medical Research,** 20, 299–330.

GOLDBERG, D. E.: 1989 **Genetic Algorithms in Search, Optimization, Machine Learning**. Boston, Addison-Wesley.

GARTNER GROUP: 2007 "Data&Analytics", (Çevrimiçi) http://www.gartner.com/6_help/glossary/GlossaryD.jsp, 16.09.2007.

GORDON, D., ve ark.: 1998 "What is an Electronic Patient Record?", **Proceedings of the American Medical Informatics Association Symposium**, 240–244.

GROOP, L.C., ve ark.: 1989 "Glucose and Free Fatty Acid Metabolism in Non-Insulin-Dependent Diyabetes Mellitus, Evidence for Multiple Sites of Insulin Resistance", **J Clin Invest**, 84, 205–213.

GUYTON, A.C. ve HALL, J.E.: 2001 **Tıbbi Fizyoloji**. ÇAVUŞOĞLU, H. (Çeviren). 10. Baskı, İstanbul, Tavaslı Matbaacılık.

GÜRSOY, U.T.: 2009 **Veri Madenciliği ve Bilgi Keşfi**, Ankara: Pegem Akademi.

GÜRSOY, T.: 2017 **Veri Madenciliğinde Güncel Yaklaşımlar**. İstanbul, Çağlayan.

HAND, D., MANNILA, H. ve SMYTH, P.: 2001 **Principles of Data Mining**, Cambridge, The MIT Press.

HARLEEN, ve BHAMBRI, P.: 2016 "A Prediction Technique in Data Mining for Diabetes Mellitus", **Apeejay-Journal of Management Sciences and Technology**, 4(1), 1–12.

HASKINS, K. J.: 2017 "Wearable Technology and Implications for the Americans with Disabilities Act Genetic Information Nondiscrimination Act , and Health Privacy", **Journal of Labor &**

Employment Law, 69, 69–78.

HEALTHLINE: 2018 "Identifying and Treating Diabetes Joint Pain", (Çevrimiçi) https://www.healthline.com/health/diabetes/joint-pain, 23.07.2018.

HERLAND, M., KHOSHGOFTAAR, T. M., ve WALD, R.: 2014 "A Review of Data Mining Using Big Data in Health Informatics", **Journal of Big Data**, 1(1), 1-35.

IDF-a (International Diabetes Federation): 2017. "The IDF Diabetes Atlas, 8th Edition", (Çevrimiçi) http://diabetesatlas.org/resources/2017-atlas.html, 26.07.2018.

IDF-b (International Diabetes Federation): 2017 "The IDF Diabetes Atlas, 8th Edition, Turkey Country Report 2017&2045", (Çevrimiçi) https://www.diabete.qc.ca/en/understand-diabetes/resources/.../IDF-DA-8e-EN-finalR3.p... 26.07.2018.

JIAWEI, H., ve KAMBER, M.: 2006 **Data Mining: Concepts And Techniques**, San Francisco, Morgan Kaufmann.

KAPLAN, R.M., BUSH, J.W. ve BERRY, C.C.: 1976 "Health Status: Types of Validity and the Index of Well-being", **Health Services Research**, 11(4), 478-507.

KARVONEN-GUTIERREZ, C.A., PARK, S.K. ve KIM, C.: 2016 "Diabetes and Menopause", **Curr Diab Rep**, 16(20), 1-8.

KAUTZKY-WILLER, A., HARREITER, J. ve "Sex and Gender Differences in Risk, Pathophysiology and Complications of Type 2 Diabetes Mellitus", **Endocrine**

PACINI, G.: 2016 **Reviews**, 37(3), 278-316.

KAUR, G., ve CHHABRA, A.: 2014 "Improved J48 Classification Algorithm for the Prediction of Diabetes", **International Journal of Computer Applications**, 98(22), 13–17.

KAVAKIOTIS, I., ve ark.: 2017 "Machine Learning and Data Mining Methods in Diabetes Research", **Computational and Structural Biotechnology Journal**, 15, 104–116.

KHOSLA, V.: 2012 "Technology Will Replace 80% of What Doctors Do", (Çevrimiçi) http://fortune.com/2012/12/04/technology-will-replace-80-of-what-doctors-do/, 29.07.2018.

KOYUNCUGİL, A. S., ve ÖZGÜLBAŞ, N.: 2009 "Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları", **Bilişim Teknolojileri Dergisi**, 2(2), 21–32.

KÖKLÜ, M. ve ÜNAL, Y.: 2013 "Analysis of a Population of Diabetic Patients Databases with Classifiers", **Human Resources**, 1(2), 481-483.

KUMAR, V.V., KIRAN, J.S. ve MURTY, G.S.: 2013 "Pattern based Dimensionality Reduction Model for Age Classification", **International Journal of Computer Applications**, 79(13), 14-20.

LAKSHMI, K. R.: 2013 "Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability", **International Journal of Modern Education and Computer Science**, 5(8), 8–17.

LAROSE, D. T., 2005 **Discovering Knowledge in Data: An Introduction to Data Mining**. Canada, JohnWiley & Sons.

LAROSE, D. T.: 2006 **Data Mining Methods and Models**. Canada, JohnWiley & Sons.

LEDGER, D.: 2014 "Inside Waerables Part 2.", (Çevrimiçi) https://medium.com/@endeavourprtnrs/inside-wearables-part-2-july-2014-ef301d425cdd, 30.07.2018.

LI, Y., LI, H., & YAO, H.: 2018 "Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach in New Urban Area of Urumqi, Xinjiang, China 2016–2017", **Computational and Mathematical Methods in Medicine**, 2018, 1-9.

MACGILL, M.: 2017. "Everything You Need to Know About Hypertension.", (Çevrimiçi) https://www.medicalnewstoday.com/articles/150109.php, 20.07.2018.

MAIMON, O. ve ROKACH, L. 2010 **Data Mining and Knowledge Discovery Hand Book**, London, Springer.

MAINDONALD, J. H.: 2012 "Data Mining with Rattle and R, The Art of Excavating Data for Knowledge Discovery by Graham Williams", **International Statistical Review**, 80(1), 176-204.

MANYIKA, J., ve ark.: 2015 "The Internet of Things: Mapping The Value Beyond the Hype", (Çevrimiçi) https://www.mckinsey.com, 28.07.2018.

MANIRUZZAMAN, M., ve ark.: 2017 "Comparative Approaches For Classification Of Diabetes Mellitus Data: Machine Learning Paradigm", **Computer Methods and Programs in Biomedicine**, 152, 23–34.

MARDONOVA, M. ve CHOI, Y.: 2018 “Review of Wearable Device Technology and Its Applications to the Mining Industry”, **Energies**, 11(3), 1-14.

Marketsandmarkets. "Data Mining Tools Market", (Çevrimiçi)

com,: 2018 https://www.marketsandmarkets.com/Market-Reports/data-mining-tools-market-259286296.html, 26.07.2018.

MEDICINENET: 2016 “Medical Definition of Creatinine” (Çevrimiçi) https://www.medicinenet.com/script/main/art.asp?articlekey=12550, 23.07.2018.

MERİH, K.: 2017 "Naive Bayes Algoritması ve R Uygulaması", (Çevrimiçi) http://datalabtr.com/index.php/2017/03/31/naive-bayes-algoritmasi-ve-r-uygulamasi, 25.07.2018.

MESSAN, K. J. ve ZHAI,Y. X. Z.: 2017 "Application of Data Mining Methods in Diabetes Prediction", **2nd International Conference on Image, Vision and Computing**, ICIVC 2017, 7(1), 1006–1010.

MOHAMMADI, M., HOSSEINI, M., ve TABATABAEE, H.: 2015 "Using Bayesian Network for the Prediction and Diagnosis of Diabetes", **BEPLS Bull. Env. Pharmacol. Life Sci**, 4(49), 109–114.

MOTKA, R., ve PARMAR, V.: 2013 "Diabetes Mellitus Forecast Using Different Data Mining Techniques", **2013 4th International Conference on Computer and Communication Tecnology (ICCCT)**, 99–103.

NIMMAGADDA, S. L., NIMMAGADDA, S. K., ve DREHER, H.: 2011 "Multidimensional Data Warehousing & Mining Of Diabetes & Food-Domain Ontologies For E-Health", **IEEE International Conference on Industrial Informatics (INDIN)**, 682–687.

OBENSHAIN, M. K.: 2004 "Application of Data Mining Techniques to Healthcare Data", **MAT Source: Infection Control and Hospital Epidemiology**, 25(August), 690–695.

PERVEEN, S., ve ark.: 2016 "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", **Procedia Computer Science**, 82(March), 115–121.

PFLIPSEN, M.C., OB, R.C., SAGUIL, A., vd.: 2009 "The Prevalence of Vitamin B12 Deficiency in Patients with Type 2 Diabetes: A Cross-Sectional Study", **JABFM**, 22(5), 528-534.

POLONSKY, K.S., STURIS, J. ve BELL, G.I.: 1996 "Seminars in Medicine of the Beth Israel Hospital, Boston. Non-Insulin-Dependent Diyabetes Mellitus a Genetically Programmed Failure of The Beta Cell to Compensate for Insulin Resistance", **N Engl J Med**, 334, 777–783.

PRnewswire: 2018. "Global Data Mining Tools Market 2018-2023: The Need for Embedded Intelligence to Gain Competitive Advantage", (Çevrimiçi) https://www.prnewswire.com/news-releases/global-data-mining-tools-market-2018-2023-the-need-for-embedded-intelligence-to-gain-competitive-advantage-300650961.html, 26.07.2018.

PRADHAN, M., ve SAHU, R. K.: 2011 "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", **International Journal of Computer Science & Emerging** Technologies, 2(2), 303–311.

RADHA, P., ve SRINIVASAN, B.: 2014 "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques", **International Journal of Innovative Science, Engineering & Technology**. 1(6), 334–339.

RAĞBETLİ, C.: 2009 "Hiperlipidemi", **Van Tıp Dergisi**, 16(1), 43-47.

RAHMAN,M. M.: 2014 “Machine Learning Based Data Pre-processing for the Purpose of Medical Data Mining and Decision Support”,

Yayınlanmamış Doktora Tezi, University of Hull.

RAJA, U., ve ark.: 2008 "Text mining in Healthcare. Applications and Opportunities",. **Journal of Healthcare Information Management : JHIM**, 22(3), 52–56.

RAJESH, K., ve SANGEETHA, V.: 2012 "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", **(Sem Qualis) International Journal of Engineering Research and Innovative Technology (IJEIT)**, 2(3), 224–229.

RESSING, M., BLETTNER, M., ve KLUG, S.: 2010 "Data Analysis of Epidemiological Studies", **Dtsch Arztebl Int**, 107(11), 187–192.

RICHARDS, G., ve ark.: 2001 "Data Mining for Indicators of Early Mortality In a Database of Clinical Records", **Artificial Intelligence in Medicine**, 22(3), 215–231.

ROJAS, L.B.A. VE GOMESI M.B.: 2012 "Metformin: an Old But Still The Best Treatment For Type 2 Diabetes", **Diabetology& Metabolic Syndrome**, 5(6), 1-15.

ROMÁN-PINTOS, L.M. ve ark.: 2016 "Diabetic Polyneuropathy in Type 2 Diabetes Mellitus: Inflammation, Oxidative Stress, and Mitochondrial Function", **J Diabetes Res.**, 2016, 1-16.

ROMINE, M.F., RODIONOV, D.A., MAEZATO, Y. vd.: 2017 "Elucidation of Roles for Vitamin B_{12} in Regulation of Folate, Ubiquinone, and Methionine Metabolism", **Proc Natl Acad Sci U.S.A,** 114(7), E1205-E1214.

ROSENSTOCK, J. ve RASKIN, P.: "Hypertension in Diabetes Mellitus", **Cardiology Clinics**, 6(4), 547-560.

1988

ROŞU, M.M., GIRGAVU, S.R., CORICI, O.M. vd.: 2018 "Atherosclerotıc Cardıovascular Dısease In a Young Male wıth Dıabetes – Case Report", **Rom J Diabetes Nutr Metab Dis.**, 25(2), 181-186.

ROTHER, K.I.: 2007 "Diabetes Treatment-Bridging the Divide", **N. Engl. J. Med**., 356 (15), 1499-1501.

SACCHI, L., ve ark.: 2015. Improving Risk-Stratification of Diabetes Complications Using Temporal Data Mining, **2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**, 2131–2134.

SANKARANARAYANAN, S., ve PERUMAL, T. P.: 2014 "A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", **2014 World Congress on Computing and Communication Technologies**, 231–233.

SARAVANA KUMAR, N. M., ESWARI, T., SAMPATH, P., ve LAVANYA, S.: 2015 "Predictive Methodology for Diabetic Data Analysis In Big Data", **Procedia Computer Science**, 50, 203–208.

SATMAN, İ., İMAMOĞLU, G. ve YILMAZ, C.: 2009 **Diabetes Mellitus ve Komplikasyonlarının Tanı Tedavi ve İzlem Kılavuzu**, Ankara,Türkiye Endokrinoloji ve Metabolizma Derneği.

SATMAN, İ., İMAMOĞLU, G. ve YILMAZ, C.: 2010 "Physician Adherence to the SEMT Guidelines fort he Management of Type 2 Diabetes in Turkey: ADMIRE Study", **Turkish Journal of Endocrinology and**

Metaboslim, 14(3), 66-72.

SEALAND, R. RAVAZI, C. ve ADLER, R.A. : 2013 "Diabetes Mellitus and Osteoporosis", **Curr Diab Rep**., 13, 411-418.

SHARMA, R., SINGH, S. N., ve KHATRI, S.: 2016 "Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), 687–691.

SHAW, J.E., SICREE, R.A. ve ZIMMET, P.Z.: 2010 "Global Estimates of the Prevalence of Diabetes for 2010 and 2030", **Diabetes Res Clin Pr**, 87, 4-14.

SIGURDARDOTTIR, A. K., JONSDOTTIR, H., ve BENEDIKTSSON, R.: 2007 "Outcomes of Educational Interventions In Type 2 Diabetes: WEKA Data-Mining Analysis", **Patient Education and Counseling**, 67(1–2), 21–31.

SINGH, S.ve KAUR, K.: 2013 "A Review on Diagnosis of Diabetes in Data Mining", **International Journal of Science and Research (IJSR),** 4(6), 2406–2408.

SORIGUER-ESCOFET, F., ve ark.: 2002 "Prevalence of Latent Autoimmune Diabetes of Adults (LADA) in Southern Spain", **Diabetes Res Clin Pract**, 56(3), 213–220.

SOSYAL GÜVENLİK KURUMU: 2013 "SGK'nın Bakış Açısıyla Diyabet", (Çevrimiçi) http://www.sgk.gov.tr/bultenler/SGK_BULTEN_62/sgk_bulten_62.pdf, 25.07.2018.

SPERLING, M.A.: 2000 "Diyabetes Mellitus in children", in: **Behrman RE, Kliegman RM, Jenson HB. Nelson Textbook of Pediatrics**. 16th Ed, Philadelphia: WB Saunders Company, 1767-1787.

STRANIERI, A., ve ark.: 2015 "Data-Analytically Derived Flexible Hba1c Thresholds for Type 2 Diabetes Mellitus Diagnostic", **Artificial Intelligence Research**, 5(1), 111-134.

SUBBE, C. P., ve ark.: 2001 "Validation of a Modified Early Warning Score in Medical Admissions", **QJM : Monthly Journal of the Association of Physicians**, 94(10), 521–526.

SULTAN, N.: 2015 "Reflective Thoughts on the Potential and Challenges of Wearable Technology for Healthcare Provision and Medical Education", **International Journal of Information Management**, 35(5), 521–526.

ŞIKLAR, E.: 2000 **Regresyon Analizine Giriş**, Eskişehir, Anadolu Üniversitesi.

QUINLAN, J. R.: 1986 "Induction of Decision Trees", **Machine Learning**, 1(1), 81-106.

QUINLAN, J. R.: 1993 **C4.5: Programs for Machine Learning**, San Fransisco, Morgan Kaufmann.

TAN, P., STEINBACH, M. ve KUMAR, V.: 2006 **Introduction to Data Mining**, Pearson Addison-Wesley.

TANG, Z. ve MACLENNAN, J.: 2005 **Data Mining with SQL Server**, USA, Wiley.

TANIAR, D.: 2008 **Data Mining and Knowledge Discovery Technologies**,

U.K., I.G.I.

TALBOT, D.: 2001 "Detecting Bioterrorism", (Çevrimiçi) https://www.technologyreview.com/s/401301/detecting-bioterrorism, 22.07.2018.

TEHRANI, KIANA ve MICHAEL, A.: 2014 "Wearable Technology and Wearable Devices: Everything You Need to Know", **Wearable Devices Magazine**, March 26, (Çevrimiçi) www.wearabledevices.com/what-is-a-wearable-device, 20.07.2018.

TURKDIAB: 2017 "Diyabet Tanı ve Tedavi", (Çevrimiçi) http://www.turkdiab.org/admin/PICS/webfiles/Diyabet_tani_ve_tedavi__kitabi.pdf, 23.07.2018.

TURNEA, M. ve ILEA, M.: 2018 "Predictive Simulation for Type II Diabetes Using Data Mining Strategies Applied to Big Data", **The 14 th International Scientific Conference eLearning and Software for Education**, 481.

TÜRKİYE HALK SAĞLIĞI KURUMU: 2014. "Türkiye Diyabet Programı", (Çevrimiçi) http://beslenme.gov.tr/content/files/diyabet/turkiyediyabetprogrami.pdf, 19.07.2018.

VEENA, V., ve RAVIKUMAR, A.: 2014 "Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus", **International Journal of Computer Applications**, 95(17), 12–16.

VIJAYALAKSHMI, M. N. ve JENIFER, M. T.: 2017 "An Analysis of Risk Factors for Diabetes Using Data Mining Approach", **International Journal of Computer Science and Mobile Computing**, 6(7), 166–172.

VYAS, R., ve ark.: 2016 "Building and Analysis of Protein-Protein Interactions Related to Diabetes Mellitus Using Support Vector Machine,

Biomedical Text Mining and Network Analysis", **Computational Biology and Chemistry**, 65, 37–44.

WANG, H. VE WEIGEND, A. S.: 2004 "Data Mining for Financial Decision Making", **Decision Support Systems**, 37(4), 457-460.

WARD, W.K., BEARD J.C. ve PORTE D.: 1984 "Pathophysiology of İnsülin Secretion in Non-Insulindependent Diyabetes Mellitus", **Diyabetes Care**, 7, 491-502.

WHO: 1985. "Diabetes Mellitus", (Çevrimiçi) http://apps.who.int/iris/handle/10665/39592, 23.07.2018.

WHO: 2011 "Use of Glycated Heamoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus", (Çevrimiçi) http://www.who.int/diabetes/publications/report-hba1c_2011.pdf, 23.07.2018.

WITTEN, I. H. ve FRANK, E.: 2005 Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, San Francisco, Morgan Kaufmann.

WITTWER, I.E.: 2016 "Diabetes, Chronic Kidney Disease and Anemia", **European Dialysis and Transplant Nurses Association**, 1-10.

YAMAN, İ. ve KESKİN, A.: 2018 "Gastro Özofajial Reflü Hastalığı ve Tedavisi", (Çevrimiçi) http://www.ekolkbb.com/gastro-ozofajial-reflu-hastaligi-ve-tedavisi, 23.08.2018.

YILDIRIM, E. G., KARAHOCA, A., ve UÇAR, T.: 2011 "Dosage Planning For Diabetes Patients Using Data Mining Methods", **Procedia Computer Science**, 3, 1374–1380.

YÜKSEL, O.: 2016 "Biyoterörizm ve Sağlık", **Sağlık İdaresi Dergisi**, 19(2), 203–222.

ZHANG, M., ve ark.: 2015 "Drug Repositioning for Diabetes Based on “Omics” Data Mining", **PLoS ONE**, 10(5), 1–14.

ZHAOLI, C., ve ark.: 2014 "To discover the Traditional Chinese Medicine Techniques Applied in Diabetes Mellitus Through Data Mining", **2014 9th International Conference on Computer Science & Education (Iccse)**, 672–676.

ZHOU, H., ZHANG, X. ve LU, J.: 2014 "Progress on Diabetic Cerebrovascular Diseases", **Bosn J Basic Med Sci**, 14(4), 185-190.

EKLER

Ek 1: Veri Önişleme İçin Kullanılan R kodları

#Kullanılan veri seti dosyadan seçilir.

> veriler =read.table (file.choose(),header=T,sep=";")

#Veri yapısı incelenir.

> str(veriler)

'data.frame': 7409 obs. of 22 variables:
$ Yas : int 59 65 61 58 34 73 69 65 72 66 ...
$ Cinsiyet : int 2 1 2 2 1 1 1 1 2 2 ...
$ Hipertansiyon : int 0 0 0 0 0 0 0 0 0 0 ...
$ Hiperlipidemi : int 0 1 0 0 0 0 0 1 0 0 ...
$ Menapoz : int 0 0 0 0 0 0 0 0 0 0 ...
$ HbA1c : num 8 7 6.1 7.5 6.8 8.15 8.5 8.9 7 6.7 ...
$ Kreatin : num 0.8 1.37 0.74 0.72 0.87 1.02 0.95 0.93 0.88 0.71 ...
$ Total.Kolesterol : num 234 131 203 206 168 168 184 205 142 181 ...
$ HDL : num 50 39 64 56 40 61 44 45 41 56 ...
$ LDL : num 153.8 77.2 121.8 111.4 96.8 ...
$ Kırgınlık.ve.Yorgunluk : num 0 0 0 0 0 0 0 0 0 0 ...
$ Metformin : int 0 0 1 0 0 0 0 1 0 0 ...
$ Insulin.Bagimli.Olmayan.Diyabetes.Mellitus: int 0 0 0 0 0 0 0 0 0 0 ...
$ Gastro.Ozofajial.Reflu.Hastalıgı : int 0 0 0 0 0 0 0 0 0 0 ...
$ Eklem.Agrisi : int 0 0 0 0 0 0 0 0 0 0 ...
$ Demir.Eksikligi.Anemileri : int 0 0 0 0 0 0 0 0 0 0 ...
$ Vitamin.B12.Eksikligi.Anemisi : int 0 0 0 0 0 0 0 0 0 0 ...
$ Aterosklerotik.Kardiyovaskuler.Hastalık : int 0 0 0 0 0 0 0 0 0 0 ...
$ Insulin.Bagimli.Diyabetes.Mellitus : int 0 0 0 0 0 0 0 0 0 0 ...
$ Serebrovaskuler.Hastalıklar : int 0 0 0 0 0 0 0 0 0 0 ...

```
 $ Osteoporoz                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Diyabetik.Polinoropati        : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
#Veri yapısı nümerik ve faktör olarak tanımlanır.
> veriler[1]<-as.numeric(x=veriler[[1]])
> veriler[2]<-as.factor(x=veriler[[2]])
> veriler[3]<-as.factor(x=veriler[[3]])
> veriler[4]<-as.factor(x=veriler[[4]])
> veriler[5]<-as.factor(x=veriler[[5]])
> veriler[11]<-as.factor(x=veriler[[11]])
> veriler[12]<-as.factor(x=veriler[[12]])
> veriler[13]<-as.factor(x=veriler[[13]])
> veriler[14]<-as.factor(x=veriler[[14]])
> veriler[15]<-as.factor(x=veriler[[15]])
> veriler[16]<-as.factor(x=veriler[[16]])
> veriler[17]<-as.factor(x=veriler[[17]])
> veriler[18]<-as.factor(x=veriler[[18]])
> veriler[19]<-as.factor(x=veriler[[19]])
> veriler[20]<-as.factor(x=veriler[[20]])
> veriler[21]<-as.factor(x=veriler[[21]])
> veriler[22]<-as.factor(x=veriler[[22]])
# Veri yapısı tekrar incelenir
```

```
> str(veriler)

'data.frame':    7409 obs. of  22 variables:

 $ Yas                                : num  59 65 61 58 34 73 69 65 72 66 ...
 $ Cinsiyet                           : Factor w/ 2 levels "1","2": 2 1 2 2 1 1 1 1 2..
 $ Hipertansiyon                      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..
 $ Hiperlipidemi                      : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 1..
 $ Menapoz                            : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..
 $ HbA1c                              : num  8 7 6.1 7.5 6.8 8.15 8.5 8.9 7 6.7 ...
 $ Kreatin                            : num  0.8 1.37 0.74 0.72 0.87 1.02 0.95 0.93 0.88
0.71 ..
 $ Total.Kolesterol                   : num  234 131 203 206 168 168 184 205 142 181
..
 $ HDL                                : num  50 39 64 56 40 61 44 45 41 56 ...
 $ LDL                                : num  153.8 77.2 121.8 111.4 96.8 ...
 $ Kırgınlık.ve.Yorgunluk             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..
 $ Metformin                          : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 2 1..
 $ Insulin.Bagimli.Olmayan.Diyabetes.Mellitus: Factor w/ 2 levels "0","1": 1 1 1 1 1
1 1 1 1..
 $ Gastro.Ozofajial.Reflu.Hastalıgı   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1
1..
 $ Eklem.Agrisi                       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..
 $ Demir.Eksikligi.Anemileri          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1
1..
 $ Vitamin.B12.Eksikligi.Anemisi      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1
1..
 $ Aterosklerotik.Kardiyovaskuler.Hastalık  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1
1 1 1..
 $ Insulin.Bagimli.Diyabetes.Mellitus : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1
1..
 $ Serebrovaskuler.Hastalıklar        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..
 $ Osteoporoz                         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..
```

$ Diyabetik.Polinoropati : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1..

Hedef nitelik, Diyabetik Polinöropati değişkeninin değerleri 0= yok, 1= var şekline dönüştürülür.

> install.packages("plyr")

> library(plyr)

> veriler$Diyabetik.Polinoropati <-revalue(veriler$Diyabetik.Polinoropati, c("1"="var","0"="yok"))

Hedef nitelik, Diyabetik Polinöropatinin referans değerinin var ile başlaması için şu kodlar yazılır.

> veriler$Diyabetik.Polinoropati <- relevel (veriler$Diyabetik.Polinoropati, ref="var")
> table(veriler$Diyabetik.Polinoropati)

var yok
285 7124
#Veri setinin özetine bakılır

> summary(veriler)

#Nümerik değişkenlerin grafikleri çizilir.

> hist(veriler$Yas, col="red", main = "Yaş Histogram Grafiği")

> hist(veriler$HbA1c, col="red", main = "HbA1c Histogram Grafiği")

> hist(veriler$Kreatin, col="red", main = "Kreatin Histogram Grafiği")

> hist(veriler$Total.Kolesterol, col="red", main = "Total Kolesterol Histogram Grafiği")

```
> hist(veriler$HDL, col="red", main = "HDL Kolesterol Histogram Grafiği")

> hist(veriler$LDL, col="red", main = "LDL Kolesterol Histogram Grafiği")

#Kategorik değişkenlerin grafikleri çizilir.

#Öncesinde ,  Cinsiyet için 2 yerine “Kadın”, 1 yerine “Erkek”  diğer kategorik
değişkenler için 0 yerine “yok”, 1 yerine “var” şeklinde değişim yapılır.
#Hedef nitelik diyabetik polinöropati için de 1 yerine “DPvar”, 0 yerine “DPyok”
değişimi yapılır.
> veriler$Cinsiyet <- revalue(veriler$Cinsiyet, c(“2”= “Kadın”, 1= “Erkek”))
> veriler$Hipertansiyon <-revalue(veriler$Hipertansiyon, c("1"="Var", "0"="Yok"))
> veriler$Hiperlipidemi <-revalue(veriler$Hiperlipidemi, c("1"="Var", "0"="Yok"))
> veriler$Menopoz<-revalue(veriler$Menopoz, c("1"="Var", "0"="Yok"))
> veriler$Kirginlik.ve.Yorgunluk<-revalue(veriler$Kirginlik.ve.Yorgunluk,
c("1"="Var", "0"="Yok"))
> veriler$Metformin <-revalue(veriler$Metformin, c("1"="Var", "0"="Yok"))
> veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus <-
revalue(veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus, c("1"="Var",
"0"="Yok"))
> veriler$Gastro.Ozofajial.Reflu.Hastaligi <-
revalue(veriler$Gastro.Ozofajial.Reflu.Hastaligi, c("1"="Var", "0"="Yok"))
> veriler$Eklem.Agrisi <-revalue(veriler$Eklem.Agrisi, c("1"="Var", "0"="Yok"))
> veriler$Demir.Eksikligi.Anemileri <-revalue(veriler$Demir.Eksikligi.Anemileri,
c("1"="Var", "0"="Yok"))
> veriler$Vitamin.B12.Eksikligi.Anemisi <-
revalue(veriler$Vitamin.B12.Eksikligi.Anemisi, c("1"="Var", "0"="Yok"))
> veriler$Aterosklerotik.Kardiyovaskuler.Hastalik <-
revalue(veriler$Aterosklerotik.Kardiyovaskuler.Hastalik, c("1"="Var", "0"="Yok"))
> veriler$Insulin.Bagimli.Diyabetes.Mellitus <-
revalue(veriler$Insulin.Bagimli.Diyabetes.Mellitus, c("1"="Var", "0"="Yok"))
```

```
> veriler$Serebrovaskuler.Hastaliklar <-revalue(veriler$Serebrovaskuler.Hastaliklar,
c("1"="Var", "0"="Yok"))
> veriler$Osteoporoz <-revalue(veriler$Osteoporoz, c("1"="Var", "0"="Yok"))
> veriler$Diyabetik.Polinoropati <-revalue(veriler$Diyabetik.Polinoropati,
c("1"="DPvar","0"="DPyok"))

#Cinsiyet için grafik çizimi
> frekanscinsiyet <- table(veriler$Cinsiyet)
> barplot(frekanscinsiyet, col="purple" , main="Cinsiyet Dağılımları
Grafiği",xlab="Kişi Sayısı",ylab = "Cinsiyet" , horiz = TRUE)

# Hipertansiyon için grafik çizimi

> frekanshipertansiyon <- table(veriler$Hipertansiyon)
> barplot(frekanshipertansiyon, col="purple" , main="Hipertansiyon Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Hipertansiyon" , horiz = TRUE)

# Hiperlipidemi için grafik çizimi
> frekanshiperlipidemi <- table(veriler$Hiperlipidemi)
> barplot(frekanshiperlipidemi, col="purple" , main="Hiperlipidemi Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Hiperlipidemi" , horiz = TRUE)

#Menopoz için grafik çizimi
> frekansmenopoz <- table(veriler$Menopoz)
> barplot(frekansmenopoz, col="purple" , main="Menopoz Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Menopoz" , horiz = TRUE)

#Kırgınlık ve Yorgunluk için grafik çizimi
> frekanskırgınlık <- table(veriler$Kirginlik.ve.Yorgunluk)
> barplot(frekanskırgınlık, col="purple" , main="Kırgınlık ve Yorgunlk Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Kırgınlık ve Yorgunluk" , horiz = TRUE)
```

```
#Metformin için grafik çizimi
> frekansmetformin <- table(veriler$Metformin)
> barplot(frekansmetformin, col="purple" , main="Metformin Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Metformin" , horiz = TRUE)

# insülin bağımlı olmayan diabetes mellütüs için grafik çizimi
> frekanstip2 <- table(veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus)
> barplot(frekanstip2, col="purple" , main="İnsülin Bağımlı Olmayan Diabetes
Mellitüs Dağılım Grafiği",xlab="Kişi Sayısı",ylab = "Tip2" , horiz = TRUE)

#Gastro özofajial reflü hastalığı grafik çizimi
> frekansreflu <- table(veriler$Gastro.Ozofajial.Reflu.Hastaligi)
> barplot(frekansreflu, col="purple" , main=" Gastro Özofajial Reflü Hastalığı
Dağılım Grafiği",xlab="Kişi Sayısı",ylab = "Gastro Özofajial Reflü" , horiz =
TRUE)

#Eklem ağrısı grafik çizimi

> frekanseklem <- table(veriler$Eklem.Agrisi)
> barplot(frekanseklem, col="purple" , main="Eklem Ağrısı Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Eklem Ağrısı" , horiz = TRUE)

#Demir eksikliği anemileri grafik çizimi
> frekansdemir <- table(veriler$Demir.Eksikligi.Anemileri)
> barplot(frekansdemir, col="purple" , main="Demir Eksikliği Anemileri Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Demir Eksikliği Anemileri" , horiz = TRUE)

#Vitamin B12 eksikliği anemisi grafik çizimi
> frekansvitamin <- table(veriler$Vitamin.B12.Eksikligi.Anemisi)
```

```
> barplot(frekansvitamin, col="purple" , main="Vitamin B12 Eksikliği Anemisi
Dağılım Grafiği",xlab="Kişi Sayısı",ylab = "Vitamin B12 Eksikliği Anemisi" , horiz
= TRUE)

#Aterosklerotik kardiyovasküler hastalık grafik çizimi

> frekanskardiyo <- table(veriler$Aterosklerotik.Kardiyovaskuler.Hastalik)
> barplot(frekanskardiyo, col="purple" , main="Aterosklerotik Kardiyovasküler
Hastalık Dağılım Grafiği",xlab="Kişi Sayısı",ylab = "Aterosklerotik
Kardiyovasküler Hastalık" , horiz = TRUE)

# İnsülin bağımlı diyabetes mellitüs grafik çizimi
> frekanstip1 <- table(veriler$Insulin.Bagimli.Diyabetes.Mellitus)

> barplot(frekanstip1, col="purple" , main="İnsülin Bağımlı Diabetes Mellitüs
Dağılım Grafiği",xlab="Kişi Sayısı",ylab = "Tip1" , horiz = TRUE)

#Serobrovasküler hastalıklar grafik çizimi
> frekanserobro <- table(veriler$Serebrovaskuler.Hastaliklar)
> barplot(frekanserobro, col="purple" , main="Serobrovasküler Hastalıklar Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Serobrovasküler Hastalıklar" , horiz = TRUE)

#Osteoporoz grafik çizimi
> frekansosteo <- table(veriler$Osteoporoz)
> barplot(frekansosteo, col="purple" , main="Osteoporoz Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Osteoporoz" , horiz = TRUE)

#Diyabetik Polinöropati grafik çizimi ile gösterimi
> frekansDP <- table(veriler$Diyabetik.Polinoropati)
> barplot(frekansDP, col="purple" , main="Diyabetik Polinöropati Dağılım
Grafiği",xlab="Kişi Sayısı",ylab = "Diyabetik Polinöropati" , horiz = TRUE)
```

```
#kutu grafikleri çizimi
> boxplot(veriler$Yas, col="orange", main="Yas Kutu Grafiği")
> boxplot(veriler$HbA1c, col="orange", main="HbA1c Kutu Grafiği")
> boxplot(veriler$Kreatin, col="orange", main="Kreatin Kutu Grafiği")
> boxplot(veriler$Total.Kolesterol, col="orange", main="Total Kolesterol Kutu
Grafiği")
> boxplot(veriler$HDL, col="orange", main="HDL kolesterol Kutu Grafiği"

#serpilme diyagramı çizimi

> pairs( ~ Yas+ HbA1c + Kreatin + Total.Kolesterol + HDL + LDL  , data= veriler,
col=" dark green", main= "Serpilme Diyagramları")


#normalizasyon işlemleri min-maks yöntemi ile

>install.packages("clusterSim")
library(clusterSim)

#yas icin
>veriler$Yas <- data.Normalization(veriler$Yas, type = "n4", normalization =
"column")
,
#HbA1c icin
>veriler$HbA1c <- data.Normalization(veriler$HbA1c, type = "n4", normalization =
"column")

#Kreatinin icin
>veriler$Kreatin <- data.Normalization(veriler$Kreatin, type = "n4", normalization =
"column")
```

#Total Kolesterol icin

>veriler$Total.Kolesterol <- data.Normalization(veriler$Total.Kolesterol, type = "n4", normalization = "column")

#HDL icin

>veriler$HDL <- data.Normalization(veriler$HDL, type = "n4", normalization = "column")

#LDL icin

>veriler$LDL <- data.Normalization(veriler$LDL, type = "n4", normalization = "column")

Ek2: KNN Algortiması Uygulaması Kodları

Sadece nümerik değerler taşıyan ve hedef niteliğin olduğu bir alt küme oluşturuldu.

> n_veriler <- veriler [, c(1,6,7,8,9,10,22)]

#veri seti eğitim ve test veri seti olarak ayrılır.

> install.packages("caret")

> library(caret)

> set.seed(1)

> verisetibolme <- createDataPartition(y=n_veriler$Diyabetik.Polinöropati, p=0.6, list=FALSE)

> egitim <- n_veriler[verisetibolme,]

> test <- n_veriler[-verisetibolme,]

#Ayrılan setteki hedef nitelik hem eğitim nitelikleri hem test niteliklerine ayrılır.

#Yine hedef nitelik eğitim veri setinden oluşan eğitim hedef Niteliklerine hem de test hedef niteliklerine atanır.

> #oluşturdugumuz n_veriler setinde dp 7. sirada o yuzden

> testNitelikleri <- test[,-7]

> testHedefNitelik <- test[[7]]

> egitimNitelikleri <- egitim [,-7]

> egitimHedefNitelik <- egitim [[7]]

#k değeri 1 için

> #KNN uygulanabilmesi icin class() paketi yuklenir

> library(class)

> set.seed(1)

> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k = k_degeri))

#Modelin performans ölçümü için kontenjans tablosu kurulur.

> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini Siniflar", "Gercek Siniflar")))

	Gercek Siniflar	
Tahmini Siniflar	DPvar	DPyok
DPvar	17	98
DPyok	97	2751

```
> (TP <- karisiklikmatrisi [1])
[1] 17
> (FP <- karisiklikmatrisi [3])
[1] 98
> (FN <- karisiklikmatrisi [2])
[1] 97
> (TN <- karisiklikmatrisi [4])
[1] 2751

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.934188322645967"
> #performans degerlendirme olcutleri

> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.934188322645967"

> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.065811677354033"

> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.149122807017544"

> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.965601965601966"


> #PPV=kesinlik ya da pozitif ongoru degeri
```

```
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.147826086956522"

> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.965941011235955"

> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.0330745865676679"

> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.850877192982456"

> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.148471615720524"

# k değeri 2 için

> k_degeri =2

> #KNN uygulanabilmesi icin class() paketi yüklenir

> library(class)

> set.seed(1)

> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))
```

```
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))
              Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     8   112
         DPyok   106  2737

> (TP <- karisiklikmatrisi [1])
[1] 8

> (FP <- karisiklikmatrisi [3])
[1] 112

> (FN <- karisiklikmatrisi [2])
[1] 106

> (TN <- karisiklikmatrisi [4])
[1] 2737

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.926425919676004"

> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0735740803239959"

> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0701754385964912"
```

```
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.960687960687961"

> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.0666666666666667"

> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.962715441435104"

> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.0377995275059062"

> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.929824561403509"

> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0683760683760684"

#k değeri 3 için

> k_degeri =3

> #KNN uygulanabilmesi icin class() paketi yuklenir

> library(class)
```

```
> set.seed(1)

> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))

> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini Siniflar", "Gercek S

                Gercek Siniflar
Tahmini Siniflar  DPvar DPyok
          DPvar     5    26
          DPyok   109  2823

> (TP <- karisiklikmatrisi [1])
[1] 5

> (FP <- karisiklikmatrisi [3])
[1] 26

> (FN <- karisiklikmatrisi [2])
[1] 109

> (TN <- karisiklikmatrisi [4])
[1] 2823

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.954438069524131"

> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0455619304758691"
```

```
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.043859649122807"

> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.990873990873991"

> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.161290322580645"

> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.962824010914052"

> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00877489031387108"

> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.956140350877193"

> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0689655172413793"

>
```

```
#k değeri 4 için

> k_degeri =4

> #KNN uygulanabilmesi icin class() paketi yuklenir

> library(class)

> set.seed(1)

> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))

> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))

              Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     1    21
         DPyok   113  2828

> (TP <- karisiklikmatrisi [1])
[1] 1

> (FP <- karisiklikmatrisi [3])
[1] 21

> (FN <- karisiklikmatrisi [2])
[1] 113

> (TN <- karisiklikmatrisi [4])
[1] 2828
```

```
> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.954775565305434"

> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0452244346945663"

> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614"

> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.992628992628993"

> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.0454545454545455"

> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.96157769466168"

> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00708741140735741"

> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"
```

```
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0147058823529412"


#k değeri 5 için

> k_degeri =5

> #KNN uygulanabilmesi icin class() paketi yuklenir

> library(class)

> set.seed(1)

> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))

> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))

              Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     1     9
         DPyok   113  2840

> (TP <- karisiklikmatrisi [1])
[1] 1

> (FP <- karisiklikmatrisi [3])
[1] 9
```

```
> (FN <- karisiklikmatrisi [2])
[1] 113

> (TN <- karisiklikmatrisi [4])
[1] 2840

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.958825514681066"

> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0411744853189335"

> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614"

> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.996840996840997"

> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.1"

> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.961733830003386"

> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
```

```
[1] "FPR = 0.0030374620317246"

> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"

> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0161290322580645"


#k değeri 6 için

> k_degeri =6
> #KNN uygulanabilmesi icin class() paketi yuklenir
> library(class)
> set.seed(1)
> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))

              Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     1    11
         DPyok   113  2838

> (TP <- karisiklikmatrisi [1])
[1] 1
> (FP <- karisiklikmatrisi [3])
[1] 11
> (FN <- karisiklikmatrisi [2])
[1] 113
```

```
> (TN <- karisiklikmatrisi [4])
[1] 2838
> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.958150523118461"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0418494768815389"
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.996138996138996"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.0833333333333333"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.9617078956286"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00371245359433007"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0158730158730159"
```

k değeri 7 için

```
> k_degeri =7
```

```
> #KNN uygulanabilmesi icin class() paketi yuklenir
> library(class)
> set.seed(1)
> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))

                Gercek Siniflar
Tahmini Siniflar DPvar DPyok
          DPvar     1     8
          DPyok   113  2841

> (TP <- karisiklikmatrisi [1])
[1] 1
> (FP <- karisiklikmatrisi [3])
[1] 8
> (FN <- karisiklikmatrisi [2])
[1] 113
> (TN <- karisiklikmatrisi [4])
[1] 2841

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.959163010462369"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0408369895376308"
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614
```

```
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.997191997191997"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.111111111111111"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.961746784021666"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00269996625042187"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.016260162601626"
```

#k değeri 8 için

```
> k_degeri =8
> #KNN uygulanabilmesi icin class() paketi yuklenir
> library(class)
> set.seed(1)
> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))
```

```
                    Gercek Siniflar
Tahmini Siniflar DPvar DPyok
        DPvar     1     4
        DPyok   113  2845

> (TP <- karisiklikmatrisi [1])
[1] 1
> (FP <- karisiklikmatrisi [3])
[1] 4
> (FN <- karisiklikmatrisi [2])
[1] 113
> (TN <- karisiklikmatrisi [4])
[1] 2845

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.96051299358758"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0394870064124199"
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.998595998595999"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.2"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.961798512508452"
```

```
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00134998312521093"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0168067226890756"

#k değeri 9 için
> k_degeri =9
> #KNN uygulanabilmesi icin class() paketi yuklenir
> library(class)
> set.seed(1)
> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))

             Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     1     4
         DPyok   113  2845
> (TP <- karisiklikmatrisi [1])
[1] 1
> (FP <- karisiklikmatrisi [3])
[1] 4
> (FN <- karisiklikmatrisi [2])
[1] 113
```

```
> (TN <- karisiklikmatrisi [4])
[1] 2845

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.96051299358758"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0394870064124199"
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.998595998595999"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.2"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.961798512508452"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00134998312521093"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0168067226890756"
```

```
#k değeri 10 için

> k_degeri =10
> #KNN uygulanabilmesi icin class() paketi yuklenir
> library(class)
> #KNN uygulanabilmesi icin class() paketi yuklenir
> library(class)
> set.seed(1)
> (tahminiSiniflar =knn(egitimNitelikleri, testNitelikleri, egitimHedefNitelik , k =
k_degeri))
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))
              Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     1     3
         DPyok   113  2846

> (TP <- karisiklikmatrisi [1])
[1] 1
> (FP <- karisiklikmatrisi [3])
[1] 3
> (FN <- karisiklikmatrisi [2])
[1] 113
> (TN <- karisiklikmatrisi [4])
[1] 2846

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.960850489368883"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0391495106311172"
```

```
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0087719298245614"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.998946998946999"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.25"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.961811422777965"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.0010124873439082"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.991228070175439"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0169491525423729"
```

Ek 3: Naive – Bayes Algortiması İçin Kodlar

```
# Önce veri seti çağırıldı

> veriler =read.table (file.choose(),header=T,sep=";")
# veri seti incelenir, nümerik ve kategorik veriler tanımlanır

> str(veriler)

> veriler[1]<-as.numeric(x=veriler[[1]])
> veriler[2]<-as.factor(x=veriler[[2]])
> veriler[3]<-as.factor(x=veriler[[3]])
> veriler[4]<-as.factor(x=veriler[[4]])
```

```
> veriler[5]<-as.factor(x=veriler[[5]])
> veriler[11]<-as.factor(x=veriler[[11]])
> veriler[12]<-as.factor(x=veriler[[12]])
> veriler[13]<-as.factor(x=veriler[[13]])
> veriler[14]<-as.factor(x=veriler[[14]])
> veriler[15]<-as.factor(x=veriler[[15]])
> veriler[16]<-as.factor(x=veriler[[16]])
> veriler[17]<-as.factor(x=veriler[[17]])
> veriler[18]<-as.factor(x=veriler[[18]])
> veriler[19]<-as.factor(x=veriler[[19]])
> veriler[20]<-as.factor(x=veriler[[20]])
> veriler[21]<-as.factor(x=veriler[[21]])
> veriler[22]<-as.factor(x=veriler[[22]])
> str(veriler)
#hedef nitelik diyabetik polinöropati 1=DPvar, 0=DPyok şeklinde tanımlanır

> library("plyr")
> veriler$Diyabetik.Polinoropati <-revalue(veriler$Diyabetik.Polinoropati,
c("1"="DPvar","0"="DPyok"))
#referans değeri var ile başlatılır

> veriler$Diyabetik.Polinoropati <- relevel (veriler$Diyabetik.Polinoropati,
ref="DPvar")
#nümerik değişkenler normalize edilir.

> #normalizasyon
> #veri setinde bulunan numerik degerler normalize edilir
> #clusterSim paketinden data.Normalization() fonksiyonu kullanilir
> library(clusterSim)
> #yas icin
> veriler$Yas <- data.Normalization(veriler$Yas, type = "n4", normalization =
"column")
```

```
> #HbA1c icin
> veriler$HbA1c <- data.Normalization(veriler$HbA1c, type = "n4", normalization =
"column")
> #Kreatin icin
> veriler$Kreatin <- data.Normalization(veriler$Kreatin, type = "n4", normalization
= "column")
> #total kolesterol icin
> veriler$Total.Kolesterol <- data.Normalization(veriler$Total.Kolesterol, type =
"n4", normalization = "column")
> #HDL icin
> veriler$HDL <- data.Normalization(veriler$HDL, type = "n4", normalization =
"column")
> #LDL icin
> veriler$LDL <- data.Normalization(veriler$LDL, type = "n4", normalization =
"column")
#veri seti eğitim ve test veri seti olarak ayrılır.

> #veri seti egitim ve test seti olarak ikiye ayrilacak
> library(caret)
> set.seed(1)
> verisetibolme <- createDataPartition(y=veriler$Diyabetik.Polinoropati, p=0.6,
list=FALSE)
> #veri setini egitim ve test olarak rastgele ikiye ayiracagiz
> egitim <- veriler[verisetibolme,]
> test <- veriler[-verisetibolme,]
# Eğitim ve test veri setine tahmininde kullanılacak nitelik ve hedef nitelik(diyabetik
polinöropati) atanır. Diyabetik polinöropati 22. Sütunda olduğu için 22 kullanıldı.

> testNitelikleri <- test[,-22]
> testHedefNitelik <- test[[22]]
> egitimNitelikleri <- egitim [,-22]
> egitimHedefNitelik <- egitim [[22]]
```

Naive bayes için e1071 paketi çağrıldı. Bu paketteki naiveBayes() fonksiyonu kullanıldı.

```
> library(e1071)
> naiveBayes_modeli_kuruldu <- naiveBayes(egitimNitelikleri, egitimHedefNitelik)
> naiveBayes_modeli_kuruldu

#modelin tahminleri bulunur
> (tahminiSiniflar <- predict(naiveBayes_modeli_kuruldu, testNitelikleri))

> #gercek siniflar ile tahmini siniflarin kiyasi
> (karisiklikmatrisi <- table(tahminiSiniflar, testHedefNitelik, dnn =c ("Tahmini
Siniflar", "Gercek Siniflar")))
              Gercek Siniflar
Tahmini Siniflar DPvar DPyok
         DPvar     5    41
         DPyok   109  2808
> (TP <- karisiklikmatrisi [1])
[1] 5
> (FP <- karisiklikmatrisi [3])
[1] 41
> (FN <- karisiklikmatrisi [2])
[1] 109
> (TN <- karisiklikmatrisi [4])
[1] 2808

#Performans değerlendirme ölçütleri hesaplandı
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.94937563280459"
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0506243671954101"
```

```
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.043859649122807"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.985608985608986"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.108695652173913"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.962632841960919"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.0138373270334121"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.956140350877193"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0625"
```

Ek 4: Lojistik Regresyon Analizi Algoritmaları Kodları

```
#veriler çağrılır
> veriler =read.table (file.choose(),header=T,sep=";")
#verilerin yapısı incelenir, kategorik değişkenler faktöre olarak tanımlanır
> str(veriler)
> veriler[1]<-as.numeric(x=veriler[[1]])
> veriler[2]<-as.factor(x=veriler[[2]])
> veriler[3]<-as.factor(x=veriler[[3]])
> veriler[4]<-as.factor(x=veriler[[4]])
```

```
> veriler[5]<-as.factor(x=veriler[[5]])
> veriler[11]<-as.factor(x=veriler[[11]])
> veriler[12]<-as.factor(x=veriler[[12]])
> veriler[13]<-as.factor(x=veriler[[13]])
> veriler[14]<-as.factor(x=veriler[[14]])
> veriler[15]<-as.factor(x=veriler[[15]])
> veriler[16]<-as.factor(x=veriler[[16]])
> veriler[17]<-as.factor(x=veriler[[17]])
> veriler[18]<-as.factor(x=veriler[[18]])
> veriler[19]<-as.factor(x=veriler[[19]])
> veriler[20]<-as.factor(x=veriler[[20]])
> veriler[21]<-as.factor(x=veriler[[21]])
> veriler[22]<-as.factor(x=veriler[[22]])

#hedef nitelik diyabetik polinöropati değişeni değerleri 1=DPvar, 0=DPyok şeklinde
değiştirilir
> library("plyr")
> veriler$Diyabetik.Polinoropati <-revalue(veriler$Diyabetik.Polinoropati,
c("1"="DPvar","0"="DPyok"))

#diğer değişkenler için 0 yerine yok, 1 yerine var atanır. Cinsiyet değişkeni için 2
yerine kadın 1 yerine erkek atanır.
> veriler$Cinsiyet <- revalue(veriler$Cinsiyet, c("2"="Kadın", "1"="Erkek"))
> veriler$Hipertansiyon <-revalue(veriler$Hipertansiyon, c("1"="Var", "0"="Yok"))
> veriler$Hiperlipidemi <-revalue(veriler$Hiperlipidemi, c("1"="Var", "0"="Yok"))
> veriler$Menopoz<-revalue(veriler$Menopoz, c("1"="Var", "0"="Yok"))
> veriler$Kirginlik.ve.Yorgunluk<-revalue(veriler$Kirginlik.ve.Yorgunluk,
c("1"="Var", "0"="Yok"))
> veriler$Metformin <-revalue(veriler$Metformin, c("1"="Var", "0"="Yok"))
```

```
> veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus <-
revalue(veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus, c("1"="Var",
"0"="Yok"))
> veriler$Gastro.Ozofajial.Reflu.Hastaligi <-
revalue(veriler$Gastro.Ozofajial.Reflu.Hastaligi, c("1"="Var", "0"="Yok"))
> veriler$Eklem.Agrisi <-revalue(veriler$Eklem.Agrisi, c("1"="Var", "0"="Yok"))
> veriler$Demir.Eksikligi.Anemileri <-revalue(veriler$Demir.Eksikligi.Anemileri,
c("1"="Var", "0"="Yok"))
> veriler$Vitamin.B12.Eksikligi.Anemisi <-
revalue(veriler$Vitamin.B12.Eksikligi.Anemisi, c("1"="Var", "0"="Yok"))
> veriler$Aterosklerotik.Kardiyovaskuler.Hastalik <-
revalue(veriler$Aterosklerotik.Kardiyovaskuler.Hastalik, c("1"="Var", "0"="Yok"))
> veriler$Insulin.Bagimli.Diyabetes.Mellitus <-
revalue(veriler$Insulin.Bagimli.Diyabetes.Mellitus, c("1"="Var", "0"="Yok"))
> veriler$Serebrovaskuler.Hastaliklar <-revalue(veriler$Serebrovaskuler.Hastaliklar,
c("1"="Var", "0"="Yok"))
> veriler$Osteoporoz <-revalue(veriler$Osteoporoz, c("1"="Var", "0"="Yok"))

# nümerik değişkenler normalize edildi
> library(clusterSim)
> #yas icin
> veriler$Yas <- data.Normalization(veriler$Yas, type = "n4", normalization =
"column")
> #HbA1c icin
> veriler$HbA1c <- data.Normalization(veriler$HbA1c, type = "n4", normalization =
"column")
> #Kreatinin icin
> veriler$Kreatin <- data.Normalization(veriler$Kreatin, type = "n4", normalization
= "column")
> #total kolesterol icin
```

```
> veriler$Total.Kolesterol <- data.Normalization(veriler$Total.Kolesterol, type =
"n4", normalization = "column")
> #HDL icin
> veriler$HDL <- data.Normalization(veriler$HDL, type = "n4", normalization =
"column")
> #LDL icin
> veriler$LDL <- data.Normalization(veriler$LDL, type = "n4", normalization =
"column")

#Veri seti eğitim ve test eğitim seti olarak ayrıldı.
> library(caret)
> set.seed(1)
> verisetibolme <- createDataPartition(y=veriler$Diyabetik.Polinoropati, p=0.6,
list=FALSE)
> egitim <-veriler[verisetibolme,]
> test <- veriler[-verisetibolme,]

#lojisitik regresyon uygulamasına başlanır. Glm() fonksiyonu kullanılır.
> lr_modeli_kurulumu <- glm(Diyabetik.Polinoropati ~.,data=egitim,
family=binomial )
> print(summary(lr_modeli_kurulumu))
# β katsayılarını ayrıca görmek için

> print(lr_modeli_kurulumu$coefficients)
#Güven aralıkları bulunur
> confint.default(lr_modeli_kurulumu)
#üstünlük oranları bulunur
> print(exp(lr_modeli_kurulumu$coefficients))

# test veri seti için tahmini sınıflar oluşturulur
> tahminiSiniflar1 <- predict(lr_modeli_kurulumu, type = "response", newdata =
test[,-22])
```

```
> tahminiSiniflar1
karisiklikmatrisi <- table(tahminiSiniflar,testdata$Diyabetik.Polinoropati,
dnn=c("Tahmini Siniflar", "Gercek Siniflar")))
(TP <- karisiklikmatrisi [1])
(FP <- karisiklikmatrisi [3])
(FN <- karisiklikmatrisi [2])
(TN <- karisiklikmatrisi [4])

#performans degerlendirme olcutleri
paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
paste0("Hata = ",(Hata <- 1-Dogruluk))
#TPR=Duyarlilik orani
paste0("TPR= ", (TPR <- TP/(TP+FN)))
#SPC=Belirleyicilik orani
paste0("SPC= ", (SPC <- TN/(FP+TN)))
#PPV=kesinlik ya da pozitif ongoru degeri
paste0("PPV= ", (PPV <- TP/(TP+FP)))
#NPV=negatif ongoru degeri
paste0("NPV= ", (NPV <- TN/(TN+FN)))
#FPR=Yanlis pozitif orani
paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
#FNR=Yanlis negatif orani
paste0("FNR=",(FNR <- FN/(FN+TP)))
#F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))

#ROC egrisi cizimi
install.packages("pROC")
library(pROC)
roc_egrisi <- roc(Diyabetik.Polinoropati ~tahminiSiniflar1, data= testdata)
#grid izgara demek
```

plot(roc_egrisi, xlab="yanlis pozitif orani", ylab ="dogru pozitif orani", main="Roc Egrisi", col = "red", grid =c(0.1,0.2), grid.col=c("blue","orange"), print.thres=F, reuse.auc=T)

Ek 5: C4.5 Karar Ağacı Algoritması Kodları

#veri seti çağrılır

```
> veriler =read.table (file.choose(),header=T,sep=";")
```

veri yapılarına göre nümerik ve faktör olarak atanır

```
> veriler[1]<-as.numeric(x=veriler[[1]])
> veriler[2]<-as.factor(x=veriler[[2]])
> veriler[3]<-as.factor(x=veriler[[3]])
> veriler[4]<-as.factor(x=veriler[[4]])
> veriler[5]<-as.factor(x=veriler[[5]])
> veriler[11]<-as.factor(x=veriler[[11]])
> veriler[12]<-as.factor(x=veriler[[12]])
> veriler[13]<-as.factor(x=veriler[[13]])
> veriler[14]<-as.factor(x=veriler[[14]])
> veriler[15]<-as.factor(x=veriler[[15]])
> veriler[16]<-as.factor(x=veriler[[16]])
> veriler[17]<-as.factor(x=veriler[[17]])
> veriler[18]<-as.factor(x=veriler[[18]])
> veriler[19]<-as.factor(x=veriler[[19]])
> veriler[20]<-as.factor(x=veriler[[20]])
> veriler[21]<-as.factor(x=veriler[[21]])
> veriler[22]<-as.factor(x=veriler[[22]])
```

#hedef nitelik diyabetik polinöropati 1 ile gösterilen değer DPvar, 0 ile gösterilen değer DPvok şeklinde dönüştürüldü.

```
> library("plyr")
> veriler$Diyabetik.Polinoropati <-
revalue(veriler$Diyabetik.Polinoropati, c("1"="DPvar","0"="DPyok"))
```

cinsiyet değişkeni 2 yerine Kadın 1 yerine Erkek atandı

```
> veriler$Cinsiyet <- revalue(veriler$Cinsiyet, c("2"="Kadın",
"1"="Erkek"))
```

#kalan değişkenlerde 1 ile gösterilen değer yerine var , 0 ile gösterilen değer yerine yok atandı

```
> veriler$Hipertansiyon <-revalue(veriler$Hipertansiyon,
c("1"="Var", "0"="Yok"))
> veriler$Hiperlipidemi <-revalue(veriler$Hiperlipidemi,
c("1"="Var", "0"="Yok"))
> veriler$Menopoz<-revalue(veriler$Menopoz, c("1"="Var", "0"="Yok"))
> veriler$Kirginlik.ve.Yorgunluk<-
revalue(veriler$Kirginlik.ve.Yorgunluk, c("1"="Var", "0"="Yok"))
> veriler$Metformin <-revalue(veriler$Metformin, c("1"="Var",
"0"="Yok"))
> veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus <-
revalue(veriler$Insulin.Bagimli.Olmayan.Diyabetes.Mellitus,
c("1"="Var", "0"="Yok"))
> veriler$Gastro.Ozofajial.Reflu.Hastaligi <-
revalue(veriler$Gastro.Ozofajial.Reflu.Hastaligi, c("1"="Var",
"0"="Yok"))
> veriler$Eklem.Agrisi <-revalue(veriler$Eklem.Agrisi, c("1"="Var",
"0"="Yok"))
> veriler$Demir.Eksikligi.Anemileri <-
revalue(veriler$Demir.Eksikligi.Anemileri, c("1"="Var", "0"="Yok"))
> veriler$Vitamin.B12.Eksikligi.Anemisi <-
revalue(veriler$Vitamin.B12.Eksikligi.Anemisi, c("1"="Var",
"0"="Yok"))
> veriler$Aterosklerotik.Kardiyovaskuler.Hastalik <-
revalue(veriler$Aterosklerotik.Kardiyovaskuler.Hastalik,
c("1"="Var", "0"="Yok"))
> veriler$Insulin.Bagimli.Diyabetes.Mellitus <-
revalue(veriler$Insulin.Bagimli.Diyabetes.Mellitus, c("1"="Var",
"0"="Yok"))
> veriler$Serebrovaskuler.Hastaliklar <-
revalue(veriler$Serebrovaskuler.Hastaliklar, c("1"="Var",
"0"="Yok"))
> veriler$Osteoporoz <-revalue(veriler$Osteoporoz, c("1"="Var",
"0"="Yok"))
```

#diyabetik polinöropati değişkeninin referans değeri var olarak değiştitirildi

```
> veriler$Diyabetik.Polinoropati <- relevel
(veriler$Diyabetik.Polinoropati, ref="DPvar")
```

#test ve eğitim veri seti rastgele ayrıldı

```
> library(caret)
> set.seed(1)
```

```
> egitimIndisleri <-
createDataPartition(y=veriler$Diyabetik.Polinoropati, p=0.6,
list=FALSE)
> egitim <- veriler[egitimIndisleri,]
> test <- veriler[-egitimIndisleri,]
> testNitelikleri <- test[,-22]
> testHedefNitelik <- test[[22]]
> egitimNitelikleri <- egitim [,-22]
> egitimHedefNitelik <- egitim [[22]]
```

#C4.5 Karar Ağacı uygulaması

```
> library(RWeka)
> C45_modeli <- J48 (Diyabetik.Polinoropati ~ ., data =egitim)
> print(summary(C45_modeli))

> #niteliklere ait kazanc orani degerleri
> GainRatioAttributeEval(Diyabetik.Polinoropati ~., data=egitim)

> #modelin sinif tahminleri
> (tahminiSiniflar <- predict (C45_modeli, newdata= test[,-22]))
> #performans degerlendirmesi yapilir
> (karisiklikmatrisi <-
table(tahminiSiniflar,test$Diyabetik.Polinoropati, dnn=c("Tahmini
Siniflar", "Gercek Siniflar")))

               Gercek Siniflar
Tahmini Siniflar DPvar DPyok
           DPvar     3     4
           DPyok   111  2845

> (TP <- karisiklikmatrisi [1])
[1] 3
> (FP <- karisiklikmatrisi [3])
[1] 4
> (FN <- karisiklikmatrisi [2])
[1] 111
> (TN <- karisiklikmatrisi [4])
[1] 2845

> #performans degerlendirme olcutleri
> paste0("Dogruluk = ",(Dogruluk <- (TP+TN)/sum(karisiklikmatrisi)))
[1] "Dogruluk = 0.961187985150186"
```

```
> paste0("Hata = ",(Hata <- 1-Dogruluk))
[1] "Hata = 0.0388120148498143"
> #TPR=Duyarlilik orani
> paste0("TPR= ", (TPR <- TP/(TP+FN)))
[1] "TPR= 0.0263157894736842"
> #SPC=Belirleyicilik orani
> paste0("SPC= ", (SPC <- TN/(FP+TN)))
[1] "SPC= 0.998595998595999"
> #PPV=kesinlik ya da pozitif ongoru degeri
> paste0("PPV= ", (PPV <- TP/(TP+FP)))
[1] "PPV= 0.428571428571429"
> #NPV=negatif ongoru degeri
> paste0("NPV= ", (NPV <- TN/(TN+FN)))
[1] "NPV= 0.962449255751015"
> #FPR=Yanlis pozitif orani
> paste0("FPR = ", (FPR <- FP/sum(karisiklikmatrisi)))
[1] "FPR = 0.00134998312521093"
> #FNR=Yanlis negatif orani
> paste0("FNR=",(FNR <- FN/(FN+TP)))
[1] "FNR=0.973684210526316"
> #F olcutu kesinlik ve duyarlilik olcutlerinin harmonik ortalamasi
> paste0("F_measure = ", (F_measure <- (2*PPV*TPR)/(PPV+TPR)))
[1] "F_measure = 0.0495867768595041"
```

ÖZGEÇMİŞ

Öğrenim Bilgisi

Doktora 2014- Devam ediyor	İstanbul Üniversitesi Sosyal Bilimler Enstitüsü/ Sayısal Yöntemler (Dr) Tez adı: Sağlık Hizmetlerinde Veri Analitiği Tez Danışmanı: Umman Tuğba Şimşek Gürsoy
Yüksek Lisans 2011-2014	Anadolu Üniversitesi Sosyal Bilimler Enstitüsü/Sayısal Yöntemler Anabilim Dalı Tez adı: 2008-2013 Yılları Arasında Faaliyet Gösteren Mevduat Bankalarının Etkinliklerinin Ölçülmesi Tez danışmanı: Ali Özdemir
Lisans 2007-2011	Bilecik Üniversitesi İktisadi Ve İdari Bilimler Fakültesi/İşletme Bölümü

Görevler

Araştırma Görevlisi 2013- Devam ediyor	Bilecik Şeyh Edebali Üniversitesi İktisadi ve İdari Bilimler Fakültesi/İşletme Bölümü

Yer Aldığı Projeler

1-	Bilecik Yükseköğretiminin Gelecek Projeksiyonunun Ortaya Konulması, Yükseköğretim Kurumları tarafından destekli bilimsel araştırma projesi, Araştırmacı, 2015-2016, (Ulusal).

2- Bilecik Şeyh Edebali Üniversitesi Kurumsal İmajının Paydaşlar Tarafından Algılanışı, Yükseköğretim Kurumları tarafından destekli bilimsel araştırma projesi, Araştırmacı, 2015-2016, (Ulusal).

Eserler

A- Uluslararası Hakemli Dergilerde Yayımlanan Makaleler

1- Torun, N.K. ve Torun, T. (2015). Doğrudan Pazarlama Faaliyetlerinde Yatırım Kararının Tahmininde Sınıflandırma Algoritmalarının Karşılaştırılması, Balkan and Near Eastern Journal of Social Sciences, 1(1), 38-50.

B- Uluslararası Bilimsel Toplantılarda Sunulan Ve Bildiri Kitaplarında (Proceedings) Basılan Bildiriler

1- Torun, N.K. (2018). The Examination of Vaginal Birth after Caesarean (VBAC) Choice via Google Trends in Turkey, 11TH INTERNATIONAL CONFERENCE ON NEW CHALLENGES IN MANAGEMENT AND BUSINESS (28.03.2018 - 29.03.2018)

D- Ulusal Hakemli Dergilerde Yayımlanan Makaleler

1- Torun, N.K. ve Özdemir, A. (2015).

Türk Bankacılık Sektörünün 2008 Küresel Finansal Krizi Sürecinde Veri Zarflama Analizi İle Etkinlik Analiz, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 33, 129-142.

E- Ulusal Bilimsel Toplantılarda Sunulan Ve Bildiri Kitaplarında Basılan Bildiriler

1- Torun, N.K. ve Cengiz, E. (2018). Endüstri 4.0 Bakış Açısının Öğrenciler Gözünden Teknoloji Kabul Modeli (TKM) İle Ölçümü, 38. Yöneylem Araştırması Endüstri Mühendisliği Ulusal Kongresi (26.06.2018 - 29.06.2018).