T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
TOPLAM KALİTE YÖNETİMİ ANABİLİM DALI
KALİTE YÖNETİMİ PROGRAMI
YÜKSEK LİSANS TEZİ

VERİ KALİTESİNDE EKSİK VERİ SORUNLARININ DERİN ÖĞRENME YÖNTEMİ İLE ÇÖZÜLMESİ: ÜRETİCİ ÇEKİŞMELİ AĞLAR İLE BİR UYGULAMA

Şevhat DOGER

Danışman
Prof. Dr. Osman Avşar KURGUN

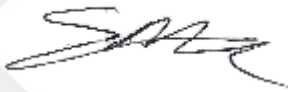
İZMİR – 2020

TEZ ONAY SAYFASI

YEMİN METNİ

Yüksek Lisans Tezi olarak sunduğum “Veri Kalitesinde Eksik Veri Sorunlarının Derin Öğrenme Yöntemi İle Çözülmesi: Üretici Çekişmeli Ağlar İle Bir Uygulama” adlı çalışmanın, tarafımdan, akademik kurallara ve etik değerlere uygun olarak yazıldığını ve yararlandığım eserlerin kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

Tarih
31/08/2020
Şevhat DOGER
İmza



ÖZET

Yüksek Lisans Tezi

Veri Kalitesinde Eksik Veri Sorunlarının Derin Öğrenme Yöntemi İle Çözülmesi: Üretici Çekişmeli Ağlar İle Bir Uygulama

Şevhat DOGER

Dokuz Eylül Üniversitesi

Sosyal BilimlerEnstitüsü

Toplam Kalite Yönetimi Anabilim Dalı

Kalite Yönetimi Programı

Büyük veri analizi ve makine öğrenimi algoritmaları alanında yapılan devrim niteliğindeki gelişmeler, bankacılık, finansal hizmetler, varlık yönetimi, yiyecek-içecek ve e-ticaret gibi endüstrilerin iş stratejilerini değiştirmiştir. Bu işletmelerin veriye dayalı karar alma stratejileri rekabet edebilirliklerini artırırken bazı yeni problemlerle karşılaşmalarına neden olmuştur. İşletmelerin verileri kullanırken karşılaştığı en yaygın problemlerden biri veri setlerinde eksik verilerin bulunmasıdır. Eksik veri problemi veri kalitesini etkileyen önemli unsurlardan birisidir. Araştırmanın amacı, veri kalitesini etkileyen eksik veri problemlerini çözmek için uygun yönteminin seçilmesi ve şarap üreten işletmeler için eksik veri problemleri karşısında başvurabilecekleri bir rehber oluşturmaktır. Bu amaç doğrultusunda uygulama, şarap kalitesi isimli veri setinde eksik değerler oluşturularak elde edilen yeni veri seti üzerine yapılmıştır. Bu bağlamda türevlenebilir üretici model deneysel olarak eksik veri tamamlamada kullanılmıştır. Üretici çekişmeli atama ağları (GAIN- Generative Adversial Imputation Networks) türevlenebilir üretici modeller sınıfını temsil etmektedir. GAIN modelinin performansı geleneksel yöntemlerden karar ağaçları, çoklu atama, beklenti maksimizasyonu ile kıyaslanmıştır. Ayrıca GAIN modelinin wasserstein ayarıyla geliştirilmiş versiyonu wasserstein üretici çekişmeli atama ağları (WGAIN-Wasserstein Generative Adversial Imputation Networks) algoritması tanıtılmıştır. Algoritmanın değerlendirilmesi orijinal veri seti ile tamamlanmış veri seti değerleri arasındaki farklar hesaplanarak

yapılmıştır. RMSE değerlendirme kriteri bunun için kullanılmıştır. Ampirik bulgular, GAIN algoritmasının özellikle WGAIN algoritmasının her eksiklik mekanizmasında ve %10' dan % 50 ye varan eksiklik oranlarında en başarılı performansı gösterdiği tespit edilmiştir. Şarap üreten işletme ölçeğinde eksik veri probleminin çözüm aşamaları anlatılmış en uygun yöntem olan WGAIN algoritmasının tercih edilmesi gerektiği belirtilmiştir. Analiz, eksik veri çözümleme yöntemleri arasında kullanımı Türkçe kaynaklarda nadir bulunan, GAIN yer verilmiş olmasından ötürü önemlidir.

Anahtar Kelimeler: Veri Kalitesi, Veri Kalitesi Boyutları, Eksik Veri, Eksik veri Tamamlama, Üretici Çekişmeli Atama Ağları, Wasserstein Üretici Çekişmeli Atama Ağları, Şarap Kalitesi.

ABSTRACT

Master's Thesis

Solving Missing Data Problems in Data Quality with Deep Learning Method: An Application with Generative Adversial Networks

Şevhat DOGER

Dokuz Eylül University
Graduate School of Social Sciences
Department of Total Quality Management
Quality Management Program

Revolutionary developments in the field of big data analysis and machine learning algorithms have changed the business strategies of industries such as banking, financial services, asset management, food and beverage and e-commerce. Data-driven decision-making strategies of these enterprises increased their competitiveness, causing them to face some new problems. One of the most common problems that businesses face when using data is missing data in data sets. Missing data problem is one of the important factors affecting data quality. The aim of the research is to choose the appropriate method to solve the missing data problems affecting the data quality and to create a guide for the wine producing enterprises that they can apply against the missing data problems. For this purpose, the application was made on the new data set obtained by creating missing values in the data set named wine quality. In this context, the differentiable generative model was used experimentally in completing missing data. The generative adversial imputation networks (GAIN) represent the class of differentiable generative models. The performance of the GAIN model was compared with traditional methods, Decision trees, multiple imputation, expectation maximization. In addition, the wasserstein generative advsersial imputation networks(WGAIN) algorithm, which was developed with the wasserstein setting of the GAIN model, was introduced. The evaluation of the algorithm was made by calculating the differences between the original data set and the completed data set values. The RMSE evaluation criterion was used

for this. Empirical findings, GAIN algorithm, especially WGAIN algorithm has been found to show the most successful performance in each missing mechanism and missing rates from 10% to 50%. The solution stages of the missing data problem in the wine producing enterprise scale are explained and it is stated that the most suitable method, WGAIN algorithm should be preferred. Analysis is important since GAIN is included among the missing data analysis methods, which is rarely found in Turkish sources.

Keywords: Data Quality, Data Quality Dimensions, Missing Data, Missing Data Imputation, Generative Adversial Imputation Networks, Wasserstein Generative Adversial Imputation Networks, Wine Quality.

VERİ KALİTESİNDE EKSİK VERİ SORUNLARININ DERİN ÖĞRENME YÖNTEMİ İLE ÇÖZÜLMESİ: ÜRETİCİ ÇEKİŞMELİ AĞLAR İLE BİR UYGULAMA

İÇİNDEKİLER

İKİNCİ BÖLÜM
EKSİK VERİ VE EKSİK VERİ TAMAMLAMA YÖNTEMLERİ

ÜÇÜNCÜ BÖLÜM

ŞARAP ÜRETİMİNDE VERİ KALİTESİNİ ETKİLEYEN EKSİK VERİ PROBLEMLERİNİN ÇÖZÜMLENMESİ ÜZERİNE BİR UYGULAMA

KISALTMALAR

TDWI	Verileri Zeka ile Dönüştürme
EFQM	Avrupa Kalite Yönetimi Kurumu
IP-MAP	Bilgi Ürün Haritası
ERC	Avrupa Araştırma Projesi
DWQ	Veri Depoları Kalite Kurumları
TDQM	Toplam Veri Kalitesi Yönetimi
DWQ	Veri Ambarı Kalite Metodolojisi
TIQM	Toplam Bilgi Kalitesi Yönetimi
AIMQ	Bilgi Kalitesi Değerlendirmesi İçin Metodoloji
CIHI	Kanada Sağlık Bilgisi Metodolojisi Enstitüsü
DQA	Veri Kalitesi Değerlendirmesi
IQM	Bilgi Kalitesi Ölçümü
ISTAT	İtalya Ulusal İstatistik Enstitüsü
AMEQ	Faaliyete Dayalı Ürün Bilgi Kalitesi Metodolojisinin Ölçülmesi ve Değerlendirilmesi
COLDQ	Düşük Veri Kalitesinin Maliyet Etkisi
QAFD	Finansal Verilerin Kalite Değerlendirmesi Metodolojisi
DaQuinCIS	Kooperatif Bilgi Sistemlerinde Veri Kalitesi
CDQ	Veri Kalitesi için Kapsamlı Metodoloji
EM	Beklenti Maksimizasyonu
ML	En Çok Olabilirlik
HKO	Hata Kare Ortalamaları
EKK	En Küçük Kareler
MCAR	Tamamen Rasgele Eksik Veri
MAR	Rasgele Eksik Veri
MNAR	Rasgele Olmayan Eksik Veri
YSA	Yapay Sinir Ağları
VAE	Varyasyonel Oto Kodlayıcılar
GAN	Üretici Çekişmeli Ağlar

GAIN	Üretici Çekişmeli Atama Ağları
WGAIN	Wasserstein Üretici Çekişmeli Atama Ağları
MI	Çoklu Atama
UCI	Kaliforniya Üniversitesi Irvin

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Bilgi işletmeler için kritik öneme sahip olan bir kavramdır. Bilginin analizi işletme körlüğünü ortadan kaldırarak rekabetçi üstünlüğü sağlarken üretim faktörlerine doğrudan etki ederek aynı zamanda bilgi ekonomisiyle de bütünleşmeyi sağlamaktadır. İşletmelerin kaynaklarının bir tek merkezden kolaylıkla yönetilmesine imkân sağlayarak hem verimlilikte olan artışı hem de atıl kapasitedeki azalışı temin etmektedir.

2000'li yılların başlamasıyla internetin, büyük veri tabanlarının ve karmaşık iletişim ağlarının bir araya gelmesi kurumların işleyişlerinde büyük değişiklikler yapmıştır. İşletmeler, hem hizmet kalitelerini iyileştirmek için hem de verimlilik artışını sağlamak, karar vermek ve planlama yapmak için bilgi teknolojilerine güvenmektedir. Ancak bilgi teknolojileri kuruluşların başarısını tek başına garantilememektedir. Çünkü veri kalitesini sağlamış işletme ya da kurumlar daha iyi rekabet edebilmektedirler.

Mevcut verilerin eksik olması, karar alma süreçleri üzerinde, kararların tam bilgiye dayanması nedeniyle olumsuz bir etkiye sahiptir. Verilerin bir veri kümesinde kaybolmasına ve bazılarının diğerlerinden daha baskın olmasının neden çeşitli nedenleri vardır. İlk bilinen neden, katılımcıların, örneğin aylık gelir gibi bazı kişisel ve hassas bilgileri vermeyi reddetmesidir. İkinci ana neden, veriyi veri tabanlarında depolamak için kullanılan sistemlerin başarısız olmasıdır. Diğer bir ana neden, sistemlerin birlikte çalışılabilirlik özelliği sonucu sistemler arasında değiş tokuş edilen bilgiler eksik verilere maruz kalabilmektedir. Bazı durumlarda, veri toplayıcıları herhangi bir veri toplama sistemini eksik verilerin varlığına karşı bağışık hale getirmek için sağlam prosedürler uygulmaktadır. Bununla birlikte, tüm bu girişimlere rağmen, eksik verilerin varlığının veri analizi görevlerinde önemli bir sorun olmaya devam etmesi sık görülen bir durumdur (Leke ve Marwala, 2019: 1-2).

Bu çalışma toplamda üç bölümden meydana gelmektedir. Çalışmanın birinci bölümünde öncelikle veri, kalite ve veri kalitesi kavramlarına açıklık getirilmiştir. Daha sonra veri kalitesi yaklaşımlarına değinilerek söz konusu yaklaşımların kullanım alanları belirtilmiştir. Son olarak veri kalitesinin boyutları tanıtılarak eksik veri problemlerinin veri kalitesinin bütünlük boyutu ile olan ilişkisine değinilmiştir.

İkinci bölümde eksik veriye ilişkin literatürde yer alan farklı tanımlar ve eksik veri mekanizmalarına, eksik veri tamamlama yöntemlerine ve teorilerine değinilerek incelenmiştir. Araştırmanın önemi, eksik veri çözümleme yöntemleri arasında kullanımı Türkçe kaynaklarda nadir bulunan, üretici çekişmeli atama ağlara yer verilmiş olmasıdır. İkinci bölümün son başlığı altında üretici çekişmeli atama ağların değiştirilmiş versiyonu wasserstein üretici çekişmeli atama ağlarından bahsedilmiştir.

Üçüncü bölümde ise öncelikle araştırma kapsamında incelenen veri seti ve veri setinin değişkenlerinin tanıtımı yapılmıştır. Kaliteli şarap üretmek isteyen bir işletmenin karşılaşabileceği problemlerden biri eksik değerlere sahip veri seti üzerinde karar almak zorunda kalabilmesidir. Araştırmanın amacı, veri kalitesini etkileyen eksik veri problemlerini çözmek için uygun yönteminin seçilmesi ve şarap üreten işletmeler için eksik veri problemleri karşısında başvurabilecekleri rehber oluşturmaktır. Bu amaç doğrultusunda şarapların kalite bakımından sınıflandırmasında kullanılan veri seti üzerinde bütünlüğü bozacak şekilde eksik veri problemi yaratılmış ve problemin çözümü için gerekli aşamalardan bahsedilmiştir. Eksik veri problemi, eksikliğin deseninin belirlenmesi, mekanizmasının anlaşılması ve rassallığın tespiti konu başlıklarında incelenmiş ve problemin giderilmesinde eksik veri çözümleme yöntemleri kullanılmıştır. Son olarak analiz sonuçlarına yer ayrılmış, elde edilen bulgular değerlendirilmiştir.

BİRİNCİ BÖLÜM
VERİ KALİTESİ

Veri kalitesi, günümüz kuruluşlarında önemli bir rol oynamaktadır, çünkü düşük veri kalitesi, düşük kurumsal üreticilikle sonuçlanan yanlış kararlara yol açabilmektedir.

Forrester, "şirketlerin çoğunluğu veri kalitesi yönetim olgunluğunun ortalamanın altında olduğuna inandığı için veri kalitesi yönetim teknolojilerine ve en iyi uygulamalara yatırım ve uygulama yönünde bir hareket olduğunu" savunmuştur (Forrester, 2011: 1).

Organizasyonlardaki düşük veri kalitesinin verimlilik üzerinde olumsuz bir etkisi olduğu özetlenmektedir (Fisher, Lauria, Smith, ve Wang, 2011: 4).

Verileri Zeka ile Dönüştürme (TDWI-Transforming Data With Intelligence), düşük veri kalitesinin işletmelere yıllık 600 milyar ABD Doları tutarında maliyet getirdiğini belirmiştir (Rockwell, 2012: 1).

Hatalı verilerden alınan hatalı kararlar sadece elverişsiz değil, aynı zamanda son derece maliyetlidir. Gartner araştırmasına göre, düşük veri kalitesinin kuruluşlar üzerindeki ortalama finansal etkisi yıllık 9,7 milyar dolardır. IBM, ABD'de işletmelerin, zayıf veri kalitesi nedeniyle yılda 3,1 trilyon dolar kaybettiğini açıklamıştır (Gartner, 2018:1; IBM, 2019: 1).

Veri kalitesi hem verileri kuruluş çapında kullanılabilir hale getirir hem de temizlemeye ve yönetmeye imkân vermektedir. Kalitesi yüksek olan veriler, stratejik sistemlerin hem kuruluşa hem de kuruluş içerisindeki bağımlılıklara yönelik tam (eksik olmayan) bir görünüm sunması amacıyla bütün söz konusu verileri entegre edebilmeye olanak sağlamaktadır. Ayrıca verinin kalitesi, karar almanın güvenilirliğinin belirlendiği temel niteliklerdendir. Veri, kurum/kuruluş içinde hareketlerinin yönetilmesinin gerekli olduğu değerli bir varlık olarak nitelendirilmektedir. Bilgi kaynakları çeşitlilik ve sayı açısından her geçen gün daha da artarken ve yasal düzenlemelere uyum sağlama çabaları daha odaklı hale gelirken, söz konusu çokça farklı kaynaklardan sağlanan bilgileri entegre etmek, bu bilgilere hem tutarlı hem de güvenilir bir şekilde erişmek ve bunları tekrar kullanmak ihtiyacı

da git gide önem kazanmaktadır. Veri kalitesini sağlamış kuruluşlar, tüm iş hedeflerine ulaşabilmek için iş kullanımına hazır halde olan bir veri hattı boyunca verinin kalitesinin korunmasını sağlayabilirler. Kuruluşlar, veri kalitesi çözümlerini uyguladıklarında, bilgi varlıklarından en iyi biçimde faydalanmak için veri bütünlüğünü iyileştirebilirler.

Bu bölümde veri kalitesi kavramının alt başlıklarından veri, kalite ve veri kalitesi kavramlarının tanımları yapılmış ve literatürde yer alan veri kalitesi yaklaşımlarından söz edilmiştir. Daha sonra, veri kalitesi boyutları arasından araştırmanın amacına uygun olarak seçilmiş altı veri kalitesi boyutuna kısaca yer verilmiş ve araştırma konusunun temel odağı olan eksik veri kavramı bütünlük kalitesi boyut içinde tanıtılmıştır.

1.1. VERİ KALİTESİ KAVRAMI

Veri kalitesi kavramı; yönetim-bilişim sistemleri ve istatistik gibi branşların öncü olduğu çeşitli alanlarda kullanılmıştır. Ancak 1960'lı yılların sonlarına doğru istatistiksel veri setlerinde aynı verinin birden fazla şekilde kaydedilmesiyle ortaya çıkan tekrarlı verileri bulabilmek amacıyla matematiksel bir teori geliştirerek veri kalitesiyle ilişkili olan bazı problemlerin çözümüne yönelik ilk araştırmaları gerçekleştirenler istatistikçiler olmuştur. 1980'li yılların başına gelindiğinde, veri kalitesi problemlerini belirleyebilmek için yönetim alanının araştırmacıları veri üretim sistemlerini kontrol etme konusuyla yakından ilgilenmişlerdir. 1990'lı yılların başındaysa, bilişim sistemlerinde araştırmalar yürüten uzmanlar veri ambarlarında ve veri depolarında depolanmış verilerin kalitesini tanımlamak, ölçmek ve iyileştirmek üzerine çalışmalar gerçekleştirmişlerdir (Batini ve Scannapieca, 2006: 4).

Veri kalitesi kavramı, giderek daha büyük elektronik veri setlerinin oluşmasıyla ve veriye dayanarak alınan yönetim kararlarından dolayı yeni boyutlar kazanmış ve böylelikle genişlemiştir. Bu bağlamda veri kalitesi problemleri de giderek daha önemli bir hal almıştır. Veri kalitesi problemlerinin çözümüne yönelik gerçekleştirilen bilimsel araştırmalar, farklı disiplinlerin bir araya gelmesini sağlamıştır (Karr, Sanil ve Banks, 2006: 137).

Veri kalitesi kavramına değinilmeden önce söz konusu kavramı meydana getiren veri ve kalite kavramlarından ayrı ayrı söz edilmesi yararlı olacaktır.

1.1.1. Veri Tanımı

Veri (İngilizce ve Latince datum; data), ham (işlem yapılmamış) gerçek enformasyon parçasına verilmiş olan terim olarak adlandırılmaktadır. Veriler, ölçümler, sayımlar, deneyler, gözlemler veya çeşitli araştırma yollarıyla edinilmektedir. Ölçüm ya da sayım yöntemiyle elde edilen ve sayısal bir değer ifade eden veriler nicel veriler olarak; sayısal bir değer ifade etmeyen veriler ise nitel veriler olarak isimlendirilmektedir. Sembolik her gösterim gibi aslında veri de belirli bir nesne, birey veya olguya yönelik bir soyutlama olarak ifade edilmektedir. Fakat enformasyon ve bilginin soyutluk seviyeleriyle kıyaslandığında, verilerin soyutluk seviyesi daha düşük kalmaktadır. Bir verinin tek başına bir anlamı olmadığı gibi herhangi bir işlevi de bulunmamaktadır (Bosij, Chafey, Greasley ve Hickie, 2003: 4).

Veriler toplandıktan sonra gruplanma, sıralanma ve özetlenme, el veya bilgisayar aracılığıyla işlem görüp enformasyona dönüştürme aşamalarıyla anlam kazanmaktadır. Böylelikle ait olduğu bağlamı açıklayabilme gücüne erişmektedir. Problem çözmek veya karar vermek gibi belli bir amaca hizmet edebilecek hale gelmektedir.

Checkland ve Holwell'a göre veri kavramı, gerçeklerin, kavramların ya da açıklamaların, bireylerin ya da otomatik araçların iletişimine, yorumlanmasına ve işlenmesine ilişkin biçimsel görünüşler olarak tanımlanmaktadır (Checkland ve Holwell, 1998: 32).

Veri kavramının tanımı yapılırken veri, enformasyon ve bilgi kavramlarının farklılıklarına değinilmiştir. Gerçek obje, birey ya da olayları ölçmek, gözlemlemek ya da sayma yöntemiyle edinilen sayılara veri denmektedir. Verilerin bir araya gelmesi ve düzenlenmesi sonucunda enformasyon elde edilirken, enformasyonun hacimce olduça küçülmüş fakat kullanım değeri çok artmış olan türü ise bilgi olarak adlandırılmıştır (Gürsakal, 2007: 9).

İstatistik veri kavramını, yorumlanma ve sunulma amacıyla toplanmış, çözümlenmiş ve özetlenmiş gerçekler olarak tanımlamaktadır. Nümerik (sayısal) verilere nicel veri, nitel verilere ise kategorik veri denilmektedir. Aynı zamanda nicel veriler de kendi içinde kesikli ve sürekli olmak üzere iki grupta incelenmektedir. Kesikli veriler sayım yöntemiyle toplanırlar ve çoğunlukla sayma sayıları türünden ifade edilirler. Sürekli veriler ise ölçüm yoluyla toplanıp, gerçel (reel) sayılar türünden ifade edilirler. Nitel veriler de ordinal ve nominal olmak üzere iki başlıkta ele alınmaktadır. Nominal veriler yazıyla ifade edilirler ve anlamlı herhangi bir sıralamaya tabi tutulmazlar. Ordinal veriler ise yazıyla ifade edilmesiyle beraber anlamlı bir sıralamaya da konabilen veri türleridir (Anderson, Sweeney ve Williams, 2011: 5).

İstatistikte, "doğum yılı" örneğinde gösterildiği şekilde, doğrudan elde edilebilen verilere basit veriler; "yaş = içinde bulunulan yıl – doğum yılı" örneğindeki gibi, basit verilerin işlenmesiyle sağlanan verilere ise türetilmiş veriler denmektedir.

Bilgisayar bilimleri açısından ele alındığında veri, hesaplama ya da manipülasyon amacıyla kullanılan bir gerçeği ifade etmektedir. Veri, makine düzeyinde ikili (binary) gösterimle; ya da karakter (rakam ya da harf) biçiminde kodlanmakta olduğu gibi her verinin de bir türü bulunmaktadır. Bu türler veri yapısı olarak adlandırılmaktadır. Veri yapıları karmaşık ya da basit olabilmektedir. Belirli bir bağlamda bir araya getirilmiş farklı türden ve çok sayıda veri, bir veri tabanında toplanabilmektedir. Veri tabanlarını oluşturabilmek ve yönetebilmek amacıyla veri tabanı yönetim sistemleri olarak adlandırılan yazılımlardan faydalanılmaktadır (Das, 2006: 1).

Daha farklı bir disiplin olan işletme yönetimi açısından değerlendirildiğinde ise veriler, finansal ve fiziksel hareketlilikleri kaydetmek, izlemek ve yönetimsel kararlar verebilmek amacıyla kullanılmaktadır (Pipino, Lee ve Wang, 2002: 211).

1.1.2. Kalite Tanımı

Kalite kavramı aslında yeni bir kavram değildir. Bu kavram insanlığın yerleşik düzene geçtiği çağlardan bu yana bilinen bir olgudur. Kaliteye ilişkin ilk

kayıtlar milattan öncesine (M.Ö) değin uzanmaktadır. Söz konusu dönemlerde, belli standartlara uygun bir şekilde üretim yapılması amacıyla bazı yaptırımların uygulandığı bilinmektedir. Eski Mısırlıların piramitleri, Yunanlıların ve Romalıların antik eserleri kalite anlayışının tarihi hakkında bize yol göstermektedir (Mitra, 1993: 4).

Çeşitli mesleki yayınlar, endüstri ve günlük yaşamımız da dâhil olmak üzere hem kalite kavramı hem de bu kavramla ilgili konularda bir birliğin olmadığı açıktır. Örnek olarak, bir kimse kaliteli ürün ya da hizmet deyimi ile ürün ya da hizmetin fiyatının yüksekliğini, başka bir kimse ise ürün ya da hizmetin güvenirliğini ifade etmeye çalışabilmektedir (Kobu, 2003: 544). Bu da bize kalite anlayışının kişiden kişiye değiştiğini göstermektedir.

Türk Dil Kurumu (TDK) tarafından yayınlanan Büyük Türkçe Sözlüğünde "nitelik" anlamına gelen kalite kelimesi, "bir malı, işi, hizmeti ya da başka bir şeyi diğeri/diğerlerinden ayırt eden özelliği, derecesi, cinsi, mükemmelliği ya da uygunluk standardı" olarak tanımlanmaktadır (Seyidoğlu, 1999: 305).

Latincede "Qualis" sözcüğünden türemiş olan ve "Qualitas" sözcüğüyle ifadeedilen kalite kavramı, "bir şeyin nasıl oluştuğu", "herhangi bir şey" anlamlarına gelmektedir (Halis, 2000: 42).

Yamak (2001: 322) kalite kavramının önceleri "standartlara uygunluk", sonraları "kullanıma uygunluk" olarak tanımlandığına, şimdiler de ise, Toplam Kalite Yönetimi yaklaşımı çerçevesinde "müşterilerinin istediğini verebilmek" olarak özetlendiğine değinmiştir. Bununla birlikte, yalnızca müşterinin ya da kullanıcının bugün ne istediğini tespit etmenin yeterli olmadığını, yarın ne isteyeceğinin de önceden tahmin edilmesinin gerektiğini savunmuştur.

Literatürde, kalite kavramının tanımları incelendiğinde, Crosby (1995), Deming (1986), Garvin (1988), Juran (1992) tarafından kalite konusunda yapılan tanımlar ve söz konusu tanımlar arasındaki farklılıklar göze çarpmaktadır.

Deming (1986: 16) kalite tanımını, yönetim sürecinin, planlaması, uygulanması, ve projede yapılan geliştirmeler üzerine tanımlamıştır. Ayrıca müşteri memnuniyetinin kaliteyi tanımlamada önemli olduğunu savunmuştur.

Deming'in kaliteye yaklaşımı daha çok istatistiksel süreç kontrolü ile ilişkilidir. Deming, ürünlerdeki ve süreçlerdeki değişimleri belirlemek,

değerlendirmek ve azaltmak için istatistiği kullanmasına rağmen kaliteye yaklaşımı bununla sınırlı kalmamıştır. Deming, kaliteyi "tekdüzelik ile tüketicilerin ihtiyaçlarını karşılamak" şeklinde tanımlamakta ve kalitenin tüketicilerin şimdiki ve gelecekteki ihtiyaçlarını karşılaması gerektiğini savunmaktadır. Bu tanımı ile kaliteyi açık ve belirgin bir şekilde tanımlamamasına rağmen, tekdüzelik ve elde edilebilirlik kavramlarını kullanarak sadece ürün değişkenliğini kontrol etmenin kaliteyi garantilemeyeceğini anlatmıştır. Bunun yanı sıra Deming'in kalite anlayışına göre, kalite, sistemin tüm seviyelerinde kontrol edilmelidir (Bovee, 1986: 32-33).

Juran kalite kavramını yöneticilere önemli gelen iki anahtar anlam; ürün özellikleri ve eksikliklerin olmaması olarak tanımlamıştır. Juran'a göre, müşteriler, daha fazla ürün özellikleri ve daha az eksikliği daha iyi kalite ile bağdaştırmaktadırlar. Bunun tersi olarak da tüketici memnuniyeti ürün ya da hizmetlerin özelliklerinde yatarken, memnuniyetsizliğin sebebi ise eksikliklerden kaynaklanmaktadır (Juran, 1991: 9).

Juran, kalite için basit ve tek bir tanım yapmaktan kaçınmıştır. Çünkü basit ve tek bir tanımlamanın kalite kavramının yanlış anlaşılmasına neden olacağından çoklu bir tanım yapmanın daha doğru olacağını iddia etmektedir. Bununla beraber başka yazarlar, Juran'ın kalite tanımını kısaca "kullanıma uygunluk" diye tanımlamaktadırlar (Ho, 1995: 51).

Garvin (1998: 42) kalite kavramının herhangi bir tanımının kesinlikle yapılamayacağını belirtmiştir. Garvin kalite kavramını beş farklı yaklaşımla tanımlamaktadır. Kalite nasıl tanımlanırsa tanımlansın yöneticilerin kaliteli ürün ya da hizmeti gördüklerinde kalitenin ne demek olduğunu bileceklerini ifade ederek kalite kavramının üstünlük boyutunu vurgulamaktadır. İkinci yaklaşım, kalitenin bir objenin sahip olduğu bir ya da daha fazla özelliğin derecesi ya da miktarı ile belirleneceğini savunarak özellik kavramına ağırlık vermektedir. Garvin güvenilirlik yaklaşımıyla, kalitenin bir ürünün ya da hizmetin üretimi aşamasında güvenirliği artırarak elde edileceğini; müşteri odaklı yaklaşımıyla kalitenin Juran gibi "kullanıma uygunluk" olarak tanımlanabileceğini; yararlılık yaklaşımıyla ise kalitenin bir kimsenin ödediği fiyatta alabileceği mükemmellik diye nitelendirilebileceğini belirtmiştir.

Son olarak Crosby (1995: 24-25) müşterileri tatmin etmek ve memnun etmek gibi tanımların hem ölçülemez hem de geliştirilemez olduklarını ve böylece hiçbir işe yaramayacaklarını düşünmektedir. Crosby, kalite kavramını çok farklı yönlerden ele almıştır. Bunların hepsinin ortak yanı gerekliliklere uygunluk ve sıfır hatadır. Crosby'e göre, kalite mükemmellik kavramı değil, "taahhüt edilen gerekliliklere uymak" anlamına gelen "yapılacağı söylenen şeyi yapmak"tır.

1.2. VERİ KALİTESİ

İş çevrelerinde başarılı olmak isteyen firmaların ya da kurumların, bilgiye karşı güçlü bir talebi bulunmaktadır. 2000'li yıllarla birlikte bu firmalar ya da kurumlar, dünyanın her yerinde faaliyetlerini sürdürebilmekte ve tedarikçileri, üretim süreçleri, piyasaları ve müşterileri hakkında kapsamlı bilgiye ihtiyaç duymaktadırlar. Ancak bu noktada "veri kalitesi" gibi önemli bir sorun karşımıza çıkmaktadır. Bilgi, veriden doğmaktadır. Ancak veriler genellikle hatalı olup kullanılmadan önce işleme tabi tutulmaları gerekmektedir. Bu nedenle verinin önemi arttıkça veri kalitesinin önemi de gün geçtikçe daha fazla artmaktadır (Schmid ve diğerleri, 2004: 69).

Redman'ın (2001:8) belirttiği gibi "eğer bilgi teknolojileri, Bilgi Çağının motorları ise, o zaman veri ve bilgi de bu motorların yakıtıdır".

İnsandan kaynaklı veri girişi hataları ve uygun veri standartlarının eksikliği, büyük kurumlarca kullanılan önemli verilerin % 25'inin hatalı olmasına neden olmaktadır. Veri kalitesi sorunu veri kavramının kendisi kadar eski olmasına rağmen, ancak 2000'li yılların başlarında akademik çevrelerde bu konu hakkında artan ilgiye, ilgili konferanslarda ve dergilerde veri kalitesi ile ilgili makalelere rastlanmaktadır (Gartner, 2018: 1).

Üniversiteler ve kütüphaneler gibi kar amacı gütmeyen kurumlar için de veri kalitesinin önemi yadsınamaz. Yüksek kaliteli veri, kurumların sadece güçlü kararlar vermelerine yardımcı olmaz, aynı zamanda sağladıkları hizmetlere de değer katmaktadır. Örneğin bir kütüphanede, güvenilir bir elektronik katalog taraması kullanıcıların ihtiyaç duydukları kaynakları bulmalarına yardımcı olmaktadır. Aynı zamanda kütüphane çalışanlarının kaynakların kullanımını daha doğru şekilde takip etmelerine yardımcı olabilmektedir. Diğer taraftan, hatalı bir katalog tarama sistemi

sadece kullanıcıları hayal kırıklığına uğratmaz, aynı zamanda hatalı veriyi temizlemek gibi çalışanlara ekstra iş yükü olarak geri dönmektedir (Huang, Lee ve Wang, 1999: 44).

Kar amacı güden ya da kar amacı gütmeyen kurumlarda ve hatta bireylerin günlük yaşantılarında veri kalitesinin önemi gün geçtikçe artarken, diğer taraftan veri kalitesi kavramının ne anlama geldiği sorusu da dikkat çekmektedir. Birçok yazar veri kalitesini "kullanıma uygunluk", diğer bir deyişle veri toplamanın kullanıcıların gereksinimlerini karşılama yeteneği olarak tanımlamaktadır (Orr, 1998: 66).

Karr ve diğerleri (2006: 138) veri kalitesi kavramını, verinin verimli, ekonomik ve hızlı bilgiye dönüştürülebilme ve kararları değerlendirmede kullanılabilme özelliği olarak tanımlamışlardır.

Veri kalitesinin çok boyutlu bir kavram olması ve bir standartının olmaması veri kalitesi boyutlarının sınıflandırılma biçimlerinin artmasına neden olmaktadır. Farklı araştırma alanlarındaki değişik yaklaşımlar, veri kalitesi boyutlarını farklı şekillerde tanımlamaktadırlar.

Veri kalitesi, nitel ya da nicel bilgi parçalarının durumunu ifade etmektedir. Veri kalitesinin birçok tanımı vardır, ancak veriler genellikle "operasyonlarda, karar verme ve planlamada amaçlanan kullanımlarına uygunsa" yüksek kalite olarak kabul edilmektedir (Redman, 2008: 16; Fadahunsi, 2019: 1). Dahası, veri, atıfta bulunduğu gerçek dünya yapısını doğru bir şekilde temsil ediyorsa yüksek kalitede kabul edilmektedir. Ayrıca, bu tanımların dışında, veri kaynaklarının sayısı arttıkça, herhangi bir dış amaç için uygun olup olmadığına bakılmaksızın iç veri tutarlılığı sorunu önem kazanmaktadır. İnsanların veri kalitesi hakkındaki görüşleri, aynı amaçla kullanılan aynı veri kümesini tartışırken bile çoğu zaman anlaşmazlık içinde olabilmektedir. Bu durumda veri yönetimi, veri kalitesi için tanımlar ve standartlar üzerinde anlaşmaya varmak için kullanılmaktadır. Bu gibi durumlarda, veri kalitesini sağlamak için standardizasyon da dahil olmak üzere veri temizleme gerekebilmektedir (Smallwood, 2014: 110).

1.3. VERİ KALİTESİ YAKLAŞIMLARI

Veri kalitesi tanımlarından bu kavramın çok boyutlu bir kavram olduğu görülmektedir. Veri kalitesi boyutları farklı disiplinlerde farklı şekillerde ele alınmaktadır.

Karr ve diğerleri (2006: 140-141), veri kalitesini üç disiplin altında incelemiştir: istatistik, toplam kalite yönetimi ve bilgisayar bilimleri. Bu üç disiplin altında veri kalitesi boyutlarını farklı yazarlar farklı şekillerde belirlemişlerdir. Bu bölümde veri kalitesi boyutlarının sözü edilen üç disiplin altında ne şekilde ele alındıkları gösterilmiştir. Veri kalitesi boyutları tanımlarından yalnızca bütünlük boyutunun tanımı tezin konusu olduğu için bir sonraki başlık altında verilecektir.

1.3.1. Toplam Kalite Yönetiminde Veri Kalitesi

Toplam kalite yönetimi, müşteri ve kullanıcıları memnun etmek ve olası hataları önlemek için insanlardaki yetenekleri ve davranışları ortaya koyarak mükemmelliğe ulaşmaya yönelik yapılan sürekli arayış ve araştırmalar olarak tanımlanmaktadır (Lakhe ve Mohanty, 1994: 9).

Bir başka tanıma göre toplam kalite yönetimi, "tüketicilerin isteklerini minimum ekonomik seviyede karşılamak amacıyla işletme organizasyonu içerisindeki çeşitli ünitelerin, kalitenin yaratılması, yaşatılması ve geliştirilmesi sürecindeki çabaları birleştirmek suretiyle koordine eden etkili bir sistem" şeklindedir (Kobu, 2003: 552).

Toplam kalite yönetimi, ürün ve hizmetlerin toplam kalitesine odaklanarak faaliyet göstermeyi amaçlayan bir yönetim felsefesidir. Avrupa Kalite Yönetimi Kurumu (EFQM-European Foundation for Quality Management) Modeli gibi modeller ile Malcolm Baldridge ve Japon Kalite ödülleri gibi ödüller toplam kalite yönetimi felsefesine dayanmaktadır. İsveç İstatistik Kurumunun kendi kalite yaklaşımı için yorumladığı toplam kalite yönetiminin ana değerlerinden bazıları müşteri odaklı olmak, herkesin liderliği ve katılımı, süreç odaklı olmak, değişkenliğin ölçümü, anlaşılması ve sürekli iyileştirme olarak sıralanabilir (Bergdahl, Booleman, Brakel ve Grünewald 2001: 1).

İstatistiksel kalite kontrolünün endüstriyel üretimde derin etkileri olmuştur. Bazı araştırmacılar, benzer fikirlerin ve tekniklerin veri kalitesine uygulanabileceğine inanmışlardır. Ancak, maliyetlerin modellenememesi, toplam kalite yönetimi felsefesinin veri kalitesine uygulanmasını zorlaştıran nedenlerden biridir (Karr ve diğerleri, 2006: 140).

Toplam kalite yönetimi felsefesi ile veri kalitesini ilişkilendirme mantığı çok açıktır. Veri; üreticileri ve müşterileri olan bir ürün olup, bundan dolayı verinin hem maliyeti hem de bir değeri vardır. Kavramsal olarak fiziksel ürünlere benzer şekilde, veriler de toplanma sürecinden başlayarak kalite özelliklerine sahiptir. Prensipte, veri kalitesi ölçülebilir ve iyileştirilebilir. Ayrıca yüksek kaliteli ürünler için verilen finansal teşvikler de aynı zamanda veri kalitesine uygulanabilir (Karr ve diğerleri, 2006: 140).

Veri kalitesi ile bilgi kalitesi toplam kalite yönetimi yaklaşımı altında şu şekilde ilişkilendirilmiştir. Toplanan veriler; bütünlük, doğruluk, güncellik gibi kalite özelliklerine sahip olduğu sürece, bu verilerin analizi sonucu elde edilen bilgi nihai müşterilerin beklentilerini karşılayabilmektedir (English, 1999: 13).

Veri kalitesi literatüründe yayınlanan ilk genel metodoloji olan toplam veri kalitesi yönetimi metodolojisinin temel amacı, toplam kalite yönetiminin prensiplerini veri kalitesine uygulamaktır. Toplam kalite yönetimi, üretim yönetiminde, üretim sürecinin çıktıları ile müşterilerin beklentileri arasındaki farklılıkları ortadan kaldırmayı amaçlamaktadır. Bu ilkeler ile tutarlı olarak, toplam veri kalitesi yönetimi de toplanan verilerin işlenmesi ile raporlanıp sunulan bilginin kullanıma uygun olması gerektiğini savunmuştur. Bu amaçla, bilgi üretim süreçlerinin tanımları için bir dil sunmuştur, ve bunu Bilgi Ürün Haritası (IP-MAP-Information Product Map) olarak adlandırmıştır (Batini, Cappiello, Francalanci ve Maurino, 2009: 34-35).

Toplam kalite yönetimi yaklaşımının veri kalitesi metodolojisine katkısı, veri kalitesi kavramının verilerin toplanmasından kullanıcıların kullanımına sunulmasına kadar geçen süreç içerisinde düşünülmesi gerekliliğidir. Aynı zamanda veri kalitesinin ölçülmesi ve iyileştirilmesi aşamalarında kullanıcıların ve müşterilerin ihtiyaçlarına ve beklentilerine cevap verecek şekilde gerekli düzenlemelerin yapılması gerekliliğini vurgulamıştır.

1.3.2. İstatistikte Veri Kalitesi

İstatistikçiler daima daha kaliteli veri için çalışmışlardır. İstatistikçiler için kaliteyi tanımlamadaki zorluk, kaliteyi hali hazırda yaptıkları bir şey olarak görmelerinde gizlidir. Tüm çalışmaları, istatistiksel tahminlerin kalitesini iyileştirmek, istatistiksel modellerin uygunluğunu denetlemek ve belirsizlik altında verdikleri kararların kalitesini yükseltmekle sınırlıdır. Standart hata, sapma, uyum ölçüsü ve hipotez testlerinde hata kavramlarını kullanarak; veri kalitesi kavramını yaptıkları tahminlerde ve analizlerde kullanmışlardır (Brackstone, 1999: 3).

Karr ve diğerleri (2006: 140), istatistikçilerin veri kalitesi metodolojisine yaptıkları önemli katkılardan bazılarını şu şekilde sıralamıştır: Veri kayıtlarını gözden geçirip önceden belirlenen sınırlılıklara uymayan verileri düzeltme süreci olarak veri düzenleme (data editing); bir dosyadaki kayıtların belli kurallar çerçevesinde başka bir dosyadaki kayıtlarla eşleştiren olasılıksal kayıt bağlantıları (probabilistic record linkage); anket yöntemi ile elde edilen verilerde eksik veri incelemesi, kapsam hatalarının ve ölçüm hatalarının belirlenmesi.

İstatistikçilerin kurumsal bazda veri kalitesine bakış açılarına bakıldığında Türkiye'de faaliyet gösteren Türkiye İstatistik Kurumu gibi birçok ülkede yer alan kurumların bazı standartlar yayınladıkları görülmektedir. Bunların her birinin amacı yayınladıkları istatistiksel verilerin kalitesini, toplam kalite yönetimi felsefesinin savunduğu fikirler çerçevesinde iyileştirmek ve yayınladıkları verileri kullananlara bu verilerin kaliteli olduğunu ve güvenle kullanabileceklerini garanti etmektir.

Aynı amaçla yola çıkan bu kurumlar, veri kalitesi boyutlarını farklı yönlerden ele almaktadırlar. Aşağıda maddeler halinde, bu kurumlardan bazılarının yayınladıkları verilerin kalitelerini ölçmek ve iyileştirmek için veri kalitesi kavramı altında inceledikleri başlıca veri kalitesi boyutları verilmiştir.

a) Eurostat, Lüksemburg : 1953 yılında kurulan Eurostat'ın amacı Avrupa Birliği ülkelerine yüksek kaliteli istatistiksel bilgi hizmeti sunmaktır. 2000 yılında yayınladıkları Standart Kalite Rapor'unda veri kalitesi boyutlarını yedi alt başlık halinde tanımlamışlardır. Bunlar; uygunluk (relevance), doğruluk (accuracy), güncellik ve

dakiklik (timeliness and punctuality), ulaşılabilirlik ve açıklık (accessibility and clarity), karşılaştırılabilirlik (comparability), tutarlılık (coherence) ve bütünlüktür (completeness). Eurostat bu yedi veri kalitesi boyutuna ek olarak verilerin maliyetinin de göz önünde bulundurulması gerektiğini savunmuştur (Eurostat, 2000: 2).

b) Birleşmiş Milletler (United Nations) : 1983 yılında Birleşmiş Milletler'in bünyesinde Avrupalı istatistikçilerin düzenlediği konferansta yayınlanan raporda Eurostat'ın söz ettiği yedi veri kalitesi boyutuna ek olarak zaman içinde karşılaştırılabilirlik ve diğer istatistikler ile karşılaştırılabilirlik boyutları ilave edilmiştir (Eurostat, 2000: 7).

c) Kanada İstatistik Kurumu (Statistics Canada) : Kanada'da kurulan bu kurum her beş yılda bir nüfus sayımı yapmakta ve buna ek olarak Kanada yaşamını yansıtmak için aktif olarak 350 farklı konuda anket yürütmektedir. 2003 yılında yayınladıkları Kalite Rehberi isimli kitaplarında, istatistiksel çıktılarının kullanıcılarının ihtiyaçlarını karşılamasını sağlayacak veri kalitesi yönetimi anlayışlarından bahsetmişlerdir. Bu kitapta sözü edilen veri kalitesi boyutları uygunluk (relevance), doğruluk (accuracy), güncellik (timeliness), ulaşılabilirlik (accessibility), yorumlanabilirlik (interpretability) ve tutarlılıktır (coherence) (Statistic Canada, 2003: 6-7).

d) Hollanda İstatistik Kurumu (Statistics Netherlands): 2004 yılında Hollanda'da kurulan bu kurumun amacı, politikacılar ve bilimsel araştırmalar için kullanılmak üzere veri toplamak, işlemek ve istatistikler yayınlamaktır. Avrupa İstatistik Sistemi'nin (European Statistical System) yayınladığı, istatistiki veriler yayınlayan tüm kurumların uyması gereken kalite kurallarını içeren rapora göre Hollanda İstatistik Kurumu ürettiği istatistiksel çıktılarının kalitesini şu boyutlara göre belirlemektedir: maliyet verimliliği (cost effectiveness), uygunluk (relevance), doğruluk ve güvenilirlik

(accuracy and reliability), güncellik ve dakiklik (timeliness and punctuality), tutarlılık ve karşılaştırılabilirlik (coherence and comparability) ve ulaşılabilirlik ve açıklık (accessibility and clarity) (Statistics Netherlands, 2008: 3).

e) İsveç İstatistik Kurumu (Statistics Sweden) : İsveç'te kurulan bu kurumun amacı, müşterilerine karar vermelerinde, tartışmalarında ve araştırmalarında yardımcı istatistikler sağlamaktır. Yayınladıkları istatistiklerin kalitelerini beş boyut altında değerlendirmektedirler. Bu boyutlar; doğruluk, güncellik, ulaşılabilirlik ve açıklık, karşılaştırılabilirlik ve tutarlılıktır (Statistic Sweden, 2000: 4).

İstatistiksel Bilimler Ansiklopedisi'nde ise resmi istatistiklerin kalitesini değerlendirmek ve iyileştirmek için veri kalitesi kavramı üç ana başlık altında incelenmiştir. Bunlar; raporların içeriği, doğruluk ve güncelliktir. Raporların içeriği ana başlığı altında; birimler, anakütle, değişkenler, istatistiksel ölçümler, araştırma alanı, referans zamanı ve kapsamlılık boyutu yer almaktadır. Yukarıda sözü edilen kurumların ayrı birer boyut olarak verdikleri dakiklik, karşılaştırılabilirlik, tutarlılık ve erişilebilirlik boyutları bu ansiklopedide güncellik boyutu altında yer almaktadır (Kotz, Johnson ve Read, 2005: 622).

1.3.3. Bilgisayar Bilimlerinde Veri Kalitesi

Kurumlar ve firmalar verilerini elektronik ortamda toplamaya ve saklamaya başladıklarından beri, bilgi teknolojileri departmanları verilerin muhafızları olarak hizmet etmeye başlamışlardır. Bunun doğal sonucu olarak, bilgisayar ve bilgi teknolojileri alanında uzman insanlar veri kalitesi kavramlarına önemli katkılarda bulunmuşlardır (Karr ve diğerleri, 2006: 141).

Tüm disiplinlerde; kalitenin, boyutların ve bunları değerlendirmek için kullanılan ölçümlerin tanımları ciddi bir iştir. Bilgisayar bilimlerindeki veri kalitesi literatürü, veri kalitesi boyutlarının tam sınıflandırmasını yapmıştır; ancak kalitenin kavramsal doğası gereği birçok boyutun tanımında farklılıklar vardır. Bilgisayar bilimlerinde ya da bilişim teknolojilerinde altı en önemli veri kalitesi boyutu

sınıflandırması Wand ve Wang (1996), Wang ve Strong (1996), Redman (1996), Jarke ve diğerleri (1995), Bovee ve diğerleri (2001) ve Naumann (2002) tarafından yapılmıştır (Batini ve diğerleri, 2009: 16).

a. Wand ve Wang (1996: 88): Bu yaklaşım bilişim sistemleri için resmi bir modele dayanmaktadır. Veri kalitesi boyutları, gerçek dünyadan bir bilgi sistemine haritalama fonksiyonları ile tanımlanmıştır. Örneğin; verinin doğru olmaması, bilgi sisteminin tanımlaması gerekenden farklı bir gerçek dünya durumunu simgelediği anlamına gelmektedir. Başka bir örnek olarak; bütünlük boyutu, gerçek dünya durumlarından bilgi sistemi durumlarına eksik haritalanma olarak tanımlanmaktadır. Bu yaklaşımda beş veri kalitesi boyutu önerilmiştir; doğruluk (accuracy), bütünlük (completeness), tutarlılık (consistent), güncellik (timeliness) ve güvenilirlik (reliability).
b. Wang ve Strong (1996: 21): Wang ve Strong'un önerdikleri yaklaşım deneysel bir çalışmadan türemiştir. Veri kalitesi boyutları veri kullanıcıları ile görüşmeler sonucunda seçilmiştir. Yazarlar 118 adet veri kalitesi boyutundan başlayarak, 15 farklı boyut seçmişler ve boyutları dört ana kategoride altında toplamışlardır. Bu kategoriler; gerçek *(intrinsic)* veri kalitesi, kavramsal (*contextual*) veri kalitesi, temsili (*representational*) veri kalitesi ve sonuncu ana başlık olarak ulaşılabilir (*accessibility*) veri kalitesidir. Bu ana kategoriler içinde yer alan veri kalitesi boyutları; doğruluk (*accuracy*), bütünlük (*completeness*), temsili tutarlılık (*consistent representation*), güncellik (*timeliness*), yorumlanabilirlik (*interpretability*), anlaşılabilirlik *(understandabilitiy)*, inanılabilirlik *(believability)*, itibar *(reputation)*, tarafsızlık *(objectivitiy)*, ulaşılabilirlik (*accessibility*), güvenlik *(security)*, katma değer *(value-added)*, açık sunum *(concise representation)* ve uygun veri miktarıdır *(appropriateamount of data)*.
c. Redman (1996: 25): Redman'ın önerdiği yaklaşım veri kalitesi boyutlarını üç ana kategoride gruplamaktadır. Bu ana kategoriler; verinin kavramsal görünümü (*conceptual view of data*), veri değerleri (*data values*) ve veri formatıdır (*data format*). Kavramsal görünümde

beş veri kalitesi boyutu, veri değerleri kategorisinde dört veri kalitesi boyutu ve veri formatı kategorisinde yedi veri kalitesi boyutu yer almaktadır. Kavramsal görünüm boyutları, doğruluk (*accuracy*), bütünlük (*completeness*), tutarlılık (*consistency*), temsili tutarlılık (*consistent representation*) ve geçerlilik *(validity).* Veri değerleri boyutları; yorumlanabilirlik (*interpretability)*, uygunluk (*relevance*), uygun veri miktarı *(appropriate amount of data)* ve taşınabilirliktir *(portability).* Veri formatı boyutları ise; açık sunum *(concise representation)*, değerlerin elde edilebilirliği *(obtainability of values),* kapsamlılık *(comprehensiveness),* minimum gereksizlik *(minimum redundancy),* format esnekliği *(format flexibility),* eksik değerleri temsil yeteneği *(ability torepresent null values),* ve kayıt cihazlarının etkin kullanımıdır *(efficient usage of recording media).*

d. Jarke ve diğerleri (1995: 10): Bu yaklaşım Avrupa Araştırma Projesi (ERC-European Research Project), Veri Depoları Kalite Kurumları (DWQ-Foundations of Data Warehouse Quality) kapsamında geliştirilmiştir. Tüm projenin amacı veritabanı dizayn faaliyetlerine rehberlik etmektir. Bu kapsamda, özel veri kalitesi boyutları önerilmiştir. Bu veri kalitesi boyutları veri tabanı çevresindeki kullanıcıların rollerine göre sınıflandırılmıştır. Dizayn ve Yönetim Kaliteleri için altı boyut, Yazılım Uygulama Kalitesi için altı boyut, Veri Kullanım Kalitesi için beş boyut ve Veri Saklama Kalitesi için beş boyut yer almaktadır. Bu boyutlardan en temel olanları şunlardır; doğruluk (*accuracy*), bütünlük (*completeness*), tutarlılık (*consistency*), güncellik (*timeliness*), geçerlilik (validity), geçicilik (*temporariness*), yorumlanabilirlik (*interpretability*), güvenilirlik (*reliability*), inandırıcılık *(credibility),* ulaşılabilirlik (*accesibility*), güvenlik (*security*), elde edilebilirlik (*attainability*), taşınabilirlik (*portability*) ve cevap verebilirliktir *(response time).*

e. Boveeve diğerleri (2001: 51): Veri kalitesi kavramının "kullanıma uygunluk" tanımını benimseyen bu yaklaşım bazı alt boyutlar ile birlikte dört veri kalitesi boyutu önermiştir. Bu yaklaşıma göre veri,

bir kullanıcı bilgiyi alabildiği (ulaşılabilirlik-accesibility), onu anlayabildiği (yorumlanabilirlik-*interpretability*), belli bir alanda uygulanabilir bulduğu (uygunluk-*relevance*) ve inandırıcı olduğuna güvendiği (inandırıcılık-*credibility*) sürece kullanıma uygundur. Bu yaklaşımın diğer boyutları, doğruluk (accuracy), bütünlük (*completeness*), tutarlılık (*consistency*), güncellik (*timeliness*) ve geçiciliktir (*temporariness*).

f. Naumann (2002: 160): Naumann'ın önerdiği yaklaşım; Web Bilgi Sistemleri için özel veri kalitesi boyutları tanımlamaktadır. Neumann, yirmi bir veri kalitesi boyutu için dört kategori belirlemiştir. Bu dört kategori; kurtarılan gerçek verileri içeren içerik ilişkili kategori, kaynak, ağ ve kullanıcı ile ilgili tarafları içeren teknik kategori, veri kaynakların sübjektif taraflarını içeren zihinsel kategori ve veri sunumunu içeren anlık kategoridir. Bu kategorilerde yer alan veri kalitesi boyutları ise; doğruluk, bütünlük, temsili tutarlılık, güncellik, yorumlanabilirlik, anlaşılabilirlik, inanılabilirlik, itibar, tarafsızlık, uygunluk, güvenlik, katma değer, açık tanım, uygun veri miktarı, elde edilebilirlik ve cevap verebilirliktir.

Batini ve diğerleri (2009: 19), bilgisayar bilimlerinde veri kalitesini değerlendiren ve iyileştiren birçok tekniğin olduğundan bahsetmişlerdir. Yaptıkları araştırmada, bu tekniklerin çeşitliliği ve karmaşıklığından dolayı, veri kalitesi değerlendirme ve iyileştirme tekniklerinin seçimine, özelleştirilmesine ve uygulanmasına yardımcı olan veri kalitesi metodolojilerini incelemişlerdir. Çalışmalarında on üç veri kalitesi metodolojisini dokuz ayrı kritere göre karşılaştırmışlardır. Bu dokuz kriterden biri olan veri kalitesi boyutlarının tanımları için yukarıda sözü edilen altı temel yaklaşımı referans göstererek on üç ayrı veri kalitesi metodolojisinde hangi boyutların incelendiğini Tablo 1'de göstermişlerdir.

Tablo 1: Metodojiler ve Veri Kalitesi Boyutları

Metodoloji		Veri Kalitesi Boyutları
TDQM	Total Data Quality Management (Toplam Veri Kalitesi Yönetimi)	Ulaşılabilirlik, Uygunluk, İnanılabilirlik, Bütünlük, Açık/Tutarlı sunum, İşlemede kolaylık, Katma değer, Hatasız olma, Yorumlanabilirlik, Tarafsızlık, Uygunluk, İtibar, Güvenlik, Güncellik, Anlaşılabilirlik
DWQ	The Datawarehouse Quality Methodology (Veri Ambarı Kalite Metodolojisi)	Doğruluk, Bütünlük, Asgarilik, İzlenebilirlik, Yorumlanabilirlik, Metaveri gelişimi, Ulaşılabilirlik (Sistem, İşlemsel, Güvenlik), Kullanılabilirlik (Yorumlanabilirlik), Güncellik (Yaygınlık, Geçicilik), Cevap verebilirlik, İnandırıcılık, Tutarlılık, Yorumlanabilirlik
TIQM	Total Information Quality Management (Toplam Bilgi Kalitesi Yönetimi)	Doğal boyutlar: Tutarlılık, Bütünlük, İşletme kurallarına uyum, Doğruluk (kaynağa ve gerçekliğe yönelik), Açıklık, Tekrarların olmaması, Gereğinden fazla verilerin denkliği, uyumu, Pragmatik boyutlar: Ulaşılabilirlik, Güncellik, Kavramsal açıklık, Kaynak sağlamlığı, Kullanılabilirlik, Doğruluk, Maliyet
AIMQ	A Methodolgy for Information Quality Assessment (Bilgi Kalitesi Değerlendirmesi İçin Metodoloji	Ulaşılabilirlik, Uygunluk, İnandırıcılık, Bütünlük, Açık/Tutarlı sunum, Faaliyetlerde kolaylık, Hatasız olma, Yorumlanabilirlik, Tarafsızlık, İtibar, Güvenlik, Güncellik, Anlaşılabilirlik
CIHI	Canadian Institute for Health Information Methodology (Kanada Sağlık Bilgisi Metodolojisi Enstitüsü)	Boyutlar: Doğruluk, Güncellik, Karşılaştırılabilirlik, Kullanılabilirlik, Uygunluk Özellikler: Kapsam fazlası, Kapsam altı, Basit/ilişkili cevap varyansı, Güvenilirlik, Birim/madde eksik verileri, Düzenleme ve yerine koyma, İşleme, Tahminleme, Güncellik, Kapsamlılık, Bütünlük, Standartlaştırma, Denklik, İlişkilendirebilme, Ürün/Tarih karşılaştırması, Ulaşılabilirlik, Raporlama, Yorumlanabilirlik, Adapte edebilirlik, Değer.
DQA	Data Quality Assessment (Veri Kalitesi Değerlendirmesi)	Ulaşılabilirlik, Uygun veri miktarı, İnanılabilirlik, Bütünlük, Hatasız olma, Tutarlılık, Açık sunum, Uygunluk, İşlemede kolaylık, Yorumlanabilirlik, Tarafsızlık, İtibar, Güvenlik, Güncellik, Anlaşılabilirlik, Katma Değer.
IQM	Information Quality Measurement	Ulaşılabilirlik, Tutarlılık, Güncellik, Açıklık, Onarabilirlik, Yaygınlık, Uygulanabilirlik, Uygunluk, Hız, Kapsamlılık, Akılcılık, Doğruluk, İzlenebilirlik, Güvenlik.

	(Bilgi Kalitesi Ölçümü)	
ISTAT	Italian National Institute of Statistics (İtalya Ulusal İstatistik Enstitüsü)	Doğruluk, Bütünlük, Tutarlılık
AMEQ	Activity-based Measuring and Evaluating of Product Information Quality Methodology (Faaliyete Dayalı Ürün Bilgi Kalitesi Metodolojisinin Ölçülmesi ve Değerlendirilmesi)	Tutarlı sunum, Yorumlanabilirlik, Anlaşılabilirlik, Açık sunum, Güncellik, Bütünlük, Katma değer, Uygunluk, Anlamlılık, Karışıklığın olmaması, Düzenleme, Okunabilirlik, Sebeplerinin olması, Açıklık, Güvenilirlik, Hatasız olma, Veri noksanlığı, Dizayn noksanlığı, İşleme, Noksanlıklar, Doğruluk, Maliyet, Tarafsızlık, İnanılabilirlik, İtibar, Ulaşılabilirlik, Belirlilik, Tutarlılık
COLDQ	Cost-effect of Low Data Quality (Düşük Veri Kalitesinin Maliyet Etkisi)	Şema: Açık tanım, Kapsamlılık, Esneklik, Sağlamlık, Esaslılık, Homojenlik, Elde edilebilirlik, Uygunluk, Basitlik/Karmaşıklık, Anlamsal tutarlılık, Sözdizimsel tutarlılık Veri: Doğruluk, Eksik değerler, Bütünlük, Tutarlılık, Yaygınlık, Güncellik, Uzmanlık, Aynı anda her yerde bulunma. Sunum: Uygunluk, Doğru yorumlanabilirlik, Esneklik, Şekilsel basitlik, Taşınabilirlik, Tutarlılık. Bilgi politikası: Ulaşılabilirlik, Güvenlik, Özerklik, Maliyet, Gereğinden fazla olma.
QAFD	Methodology for the Quality Assessment of Financial Data (Finansal Verilerin Kalite Değerlendirmesi Metodolojisi)	Sözdizimsel/Anlamsal doğruluk, İçsel/Dışsal tutarlılık, Bütünlük, Yaygınlık, Benzersiz olma
DaQuinCIS	Data Quality in Cooperative Information Systems (Kooperatif Bilgi Sistemlerinde Veri	Doğruluk, Bütünlük, Tutarlılık, Yaygınlık, Güvenilir olma

	Kalitesi)	
CDQ	Comprehensive Methodology for Data Quality (Veri Kalitesi için Kapsamlı Metodoloji)	Şema: Modele göre doğru olma, Gerekliliklere göre doğru olma, Bütünlük, Okunabilirlik, Normalleştirme Veri: Sözdizimsel/Anlamsal doğruluk, Bütünlük, Tutarlılık, Yaygınlık, Güncellik, İtibar, Ulaşılabilirlik, Maliyet

Kaynak: Batini ve diğerleri, 2009: 18.

Tablo 1 de birçok kalite boyutu değişkenin metodojilerde tanımlandığı gösterilmiştir. Bunun nedeni çok sayıda olayın veri olarak tanımlanmasıdır Metodolojiler, kalite boyutlarının birden fazla sınıflandırılmasını önermektedir. TIQM boyutları doğal ve pragmatik olarak sınıflandırır. COLDQ, şema, veri, sunum ve bilgi politikası boyutları arasında ayrım yapar. CIHI, boyutlar ve ilgili özellikler açısından iki seviyeli bir sınıflandırma sağlar. CDQ ise şema ve veri boyutlarını önerir.

1.4. VERİ KALİTESİ BOYUTLARI

Veri kalitesi kavramı, çok boyutlu bir kavram olduğundan Toplam Kalite Yönetimi, İstatistik ve Bilgisayar Bilimleri alanlarında farklı şekillerde incelenmiştir. Veri kalitesi yaklaşımları başlığı altında, bu üç disiplinin literatürde yer alan başlıca yazarların ve kurumların veri kalitesi kavramını tanımlamada, değerlendirmede ve ölçümünde kullandıkları farklı veri kalitesi boyutlarına değinildi.

Veri kalitesi boyutları birbirinden bağımsız değildir, yani aralarında ilişki vardır. Bir boyutun belirli bir uygulama için diğerlerinden daha önemli olduğu düşünülürse, onu tercih etme seçimi diğeri boyut için olumsuz sonuçlar doğurabilmektedir. Ayrıca veri kalitesi boyutları hiyerarşiktir; yani bütünlük gibi üst düzey ölçümler, doğruluk gibi alt düzey ölçümleri etkilemektedir. Doğruluk geçerliliğe bağlıdır: Ölçüm öncelikle geçerli değilse doğru olması imkansızdır. Geçerlilik uygunluğa bağlıdır. Yani ölçümün önce referans olarak ibraz edilen belgeye uyması daha sonra geçerli olabileceği anlamına gelmektedir. Örneğin bir sistem bilinen bir endüstri standardından üç karakterli bir kod değeri bekliyorsa ve

sunulan veriler dahili bir standarttan iki karakterli bir değerse, değer geçerli olmayacaktır. Benzer şekilde, uygunluk bütünlüğe boyutuna bağlıdır.

Verilerin kalitesini değerlendirmede, kullanılabilecek altı temel veri kalitesi boyutu seçilmiştir. Bu altı kalite boyutunun seçilmesinin nedeni tüm yaklaşımlarda ortak olması ve işletmelerde verilerin kalitesini belirlemede sıklıkla kullanılan boyutları olmasıdır. Ayrıca araştırma uygulaması için seçilen veri setine bu boyutların uygulanabilmesi seçilme nedenlerinden bir diğeridir. Seçilen altı boyuttan bütünlük boyutu, eksik veri konusunu kapsadığından ve eksik veri bir sonraki bölümde anlatılacağı üzere uzun yıllardır araştırma konusu olduğu için eksik veri konusuna geçmeden önce bu boyutun tanımına kapsamlı olarak bir sonraki başlıkta yer verilmiştir. Ayrıca bu başlık altında diğer boyutların tanımlarına yer verilmiştir. Seçilen boyutlar: doğruluk, güvenirlilik bütünlük, güncellik, tutarlılık, geçerlilik boyutlarıdır.

1.4.1. Bütünlük Dışındaki Veri Boyutları

a) **Geçerlilik:** kullanılan bir ölçüm aracı ölçülmek istenen özelliğe uygun ise, ölçülen veriler istenen özelliklerin niteliğini tam olarak yansıtıyorsa ve veriler amaca yönelik olarak yararlı ise geçerli olduğu kabul edilir. Araştırmacılar ölçülen verilere bakarak genelleme yapmadan önce araştırma süreci ve toplanan verilerin geçerliliği ile üstüne bilgi vermelidirler. Bir araştırmada başvurulan istatistiki analizler ve bunun sonuncuda elde eldilen sonuçların değeri, geçerliliğe bağlıdır (Şencan, 2005: 725). Geçerlilik; her bilim disiplininde farklı yaklaşımlar ile ele alınmakta ve literatürde farklı şekillerde tanımlanmaktadır. Geçerlilik boyutu ile ilgili yapılan tanımlarda genelde üç öğe vurgulanmaktadır. Birincisi kullanılacak ölçüm aracının ölçülmek istenen özelliklere uygun olması gerekmektedir. İkincisi yapılan ölçüm ya da ölçümlerin belirlenen kurallara uygun olarak doğru şekilde yapılması gerekmektedir. Üçüncüsü ise , ölçülen verilerin ölçülmesi istenen özellikleri yansıtması gerekmektedir (Şencan, 2005: 1).

Geçerlilik, belirli bir biçime uymayan veya iş kurallarına uymayan bilgileri ifade eden bir veri kalitesi boyutudur. Geçerlilik boyutuna verilen en yayın örneklerden biri doğum günleridir. Birçok sistem doğum günü tarihinin belirli bir

biçimde girilmesini istemektedir ve bu yapılmadığı takdirde veri girişi geçersiz sayılmaktadır (Sattler, 2009: 46).

b) **Güvenilirlik:** Hauser ve Rouse,güvenirliliği, bir veri değerinin geçerli olup olmadığına bakılmaksızın benzer değere dönüşme süreci şeklinde tanımlamaktadır. Ayrıca Geçerliliği, veri kaynağının doğru olması ve tam olarak belirttiği değeri yansıtmasını geçerli olma süreci olarak belirtmektedir. Bir veri kaynağının güvenilir olmasına rağmen geçerli bir veri kaynağı olmayabileceğini savunmaktadır. (Hauser ve Rouse, 2007: 164).

Üretilen bir bilginin bilimsel bir nitelik kazanması, bilginin doğru olmasına ve bu bilginin her defasında tekrarlanan deney ve gözlemlerle kanıtlanabilmesine bağlıdır. Belirli varsayım/varsayımların sınandığı, değişkenler arasında nedensellik ilişkisinin kurulabildiği veriler, güvenilirlik ve geçerlilik analizlerine dayanıyorsa güven sağlar (Pipino ve diğerleri, 2002: 215).

Hollanda İstatistik Kurumu, güvenilirlik boyutu ile doğruluk boyutunu bir arada incelemiştir ve istatistiklerin ölçmeyi amaçladıkları şeyi ölçebildikleri sürece güvenilir olacağını belirtmiştir. Böylelikle yayınlanan istatistiklerin gerçek değerlere oldukça yakın olacağını savunmaktadır (Statistics Netherlands, 2008: 5).

c) **Tutarlılık:** Bir veri kümesindeki veri değerlerini, başka bir veri kümesindeki değerlerle tutarlı olmasını ifade etmektedir. Kesin bir tutarlılık tanımı, ayrı veri kümelerinden çizilen iki veri değerinin birbiriyle çakışmaması gerektiğini belirtmektedir. Önceden tanımlanmış bir dizi kısıtlamayla tutarlılık kavramı daha da karmaşık olabilmektedir. Daha resmi tutarlılık kısıtlamaları, bir kayıt ya da mesaj boyunca ya da tek bir özelliğin tüm değerleri boyunca özellik değerleri arasındaki tutarlılık ilişkilerini belirten bir kurallar kümesi olarak kapsüllenebilir. Ancak, işlem hatalarının farklı platformlarda çoğaltılabilmesinin birçok yolu vardır, bu da bazen doğru olmasalar bile tutarlı olabilecek veri değerlerine yol açmaktadır. Tutarlılık kuralını doğrulayan bir örnek, kurumsal bir hiyerarşi yapısı içinde, her müşteri temsilcisine atanan müşteri sayısının, şirketin tamamının sahip olduğu müşteri sayısını geçmemesi gerektiği verilmektedir (Loshin, 2006: 9).

Modern veritabanı sistemleri verilerin bütünlüğünü ve tutarlılığını sağlamak için gelişmiş destek sağlamasına rağmen, tutarsızlığın bir başka önemli veri kalitesi sorunu olmasının birçok nedeni vardır. Bu nedenle, bir veri kalitesi boyutu olarak

tutarlılık, bir sistemde yönetilen verilerin belirtilen kısıtlamaları veya iş kurallarını ne ölçüde karşıladığı olarak tanımlanmaktadır. Bu tür kurallar, müşteri tanımlayıcılarının benzersizliği veya referans bütünlüğü (ör. "Her sipariş için bir müşteri kaydı bulunmalıdır") gibi klasik veritabanı bütünlüğü kısıtlamaları veya öznitelikler arasındaki ilişkileri açıklayan daha gelişmiş iş kuralları (örneğin, "yaş = içinde bulunulan tarih – doğum tarihi" ya da sürücü ehliyet numarası yaş > 16 ise vardır) olabilmektedir. Bu kurallar kullanıcı tarafından belirtilmelidir veya kural indüksiyonu uygulanarak eğitim verilerinden otomatik olarak türetilebilmektedir (Sattler, 2009: 12).

d) **Doğruluk:** Ulusal istatistik kurumları veri kalitesi kavramını, "kullanım uygunluğunu" etkileyebilen istatistiksel verilerin tüm niteliklerini kapsayacak şekilde tanımlamaktadır. Bu niteliklerin sayısı kurumlara göre değişiklik gösterebilmekte, ancak kurumların tümü de doğruluk boyutuna önemli bir boyut olarak yaklaşmaktadır. Doğruluk kavramsal olarak; istatistiksel tahminlerin, tahminledikleri kavramsal değerlere yakınlığı olarak tanımlanmaktadır. Veri kalitesini oluşturan birçok nitelikten sadece biri olan doğruluk, veri kalitesi kapsamında gerekli ancak yeterli olmayan boyuttur (Brackstone, 2001: 16).

Bir başka veri kalitesi problemi genellikle ölçüm hataları, gözlem yanlılıkları veya basitçe uygunsuz gösterimlerden kaynaklanmaktadır. Doğruluk, verilerin ne ölçüde doğru, güvenilir ve hatasız olarak sertifikalandırıldığı olarak tanımlanmaktadır. Doğruluk boyutu uygulamaya bağlı olabilmektedir. Gerçek dünya değerine olan mesafeyi veya bir özellik değerinin en uygun ayrıntı derecesini belirtmektedir (Sattler, 2009: 46).

e) **Güncellik:** Bir verinin güncelliğinin bilinmesi ve veri çıktısının zamanında ulaşılabilir olup olmadığı bilgisine sahip olma anlamlarına gelmektedir. Güncellik üç faktörden etkilenmektedir. Birinci faktör, gerçek dünya sistem değişikliği verilerinin bilgi sistemi tarafından ne kadarının elde edilip güncellendiği; ikinci faktör, gerçek dünya sisteminin değişme miktarı ve üçüncü faktör, verinin gerçek manada kullanıldığı zaman (Wand ve Wang, 1996: 93).

Batini ve Scannapieca, veri kalitesinin zaman boyutunu üç şekilde incelemişlerdir. Yaygınlık boyutu, verinin ne kadar güncel olduğunu göstermektedir. Geçicilik boyutu, verinin zaman içinde değişme frekansıdır. Güncellik ise üzerinde

çalışan görev için verinin elde edilebilirliğidir. Güncellik boyutunun, belli bir kullanım alanı için geç kalındığı için faydasız olan mevcut verilere sahip olunabilineceği gerçeğini yansıttığını savunmuşlardır. Örneğin, üniversite dersleri için zaman çizelgesi en güncel verileri içererek hazır olabilir, ancak sadece derslerin başlama zamanından sonra elde edilebilir olursa güncelliği geçersiz olur (Batini ve Scannapieca, 2016: 29).

1.4.2. Bütünlük Boyutu

Türk Dil Kurumu (TDK) tarafından yayınlanan Büyük Türkçe Sözlükte "eksik olmama durumu, bütün olma durumu" olarak tanımlanan bütünlük ya da tamlık veri kalitesi kavramının önemli boyutlarından biridir.

Eksik veri, veri yokluğu veya veri bulunamaması anlamına gelmektedir. Veri eksikliği; anket formundaki belli bir bölümde yer alan bazı sorularda, belirli bir bölümde veya anketin tamamında görülebilir. Örneklemde yer alan bir kişi anket formundaki bir bölümü veya bazı soruları cevaplandırmamışsa bu tür eksik veriler madde bazında eksik veri (item nonresponse) olarak adlandırılır. Bir anket formunun tamamı örneklemde yer alan bir kişi tarafından bilerek veya o kişiye ulaşılamadığından doldurulmamış ise bu tür eksik verilere birim bazında eksik veri (unit nonresponse) denir (Groves ve diğerleri, 2004: 13).

Batini ve Scannapieca'nın bilgisayar bilimleri alanındaki veri kalitesi literatüründe önemli bir yeri olan kitaplarında bütünlük kavramı, "verinin, üzerinde çalışılan görev için yeterli genişliğe, derinliğe ve kapsama sahip olma derecesi" olarak tanımlanmıştır. Üç tür bütünlük boyutu tanımlanmıştır. Bunlar, şema bütünlüğü, kavramların ve özelliklerinin şemada eksik olmaması derecesi,sütun bütünlüğü, tablodaki belirli bir özellik ya da sütun için eksik değerlerin ölçüsü,popülasyon bütünlüğü, eksik değerlerin referans popülasyona göre değerlendirilmesi, tablodaki bir sütunun ya da belli bir özelliğin eksik olmaması anlamına gelen sütun bütünlüğü ve referans bir anakütle ile ilişkili değerlerin eksik olmaması anlamına gelen anakütle bütünlüğüdür (Batini ve Scannepica, 2016: 28).

Tablo 2'de bütünlük boyutunun, veri yaklaşımları başlığı altında sözü edilen altı temel yaklaşımda nasıl tanımlandığı gösterilmektedir. Tablodan görüldüğü gibi,

bütünlük boyutunun özet tanımı konusunda fikir birliği bulunmaktadır. Tanımlar, bu yazarların referans aldıkları kavramlar bakımından farklılık göstermektedir, örneğin Wand ve Wang bilgi sistemlerini, Jarke ve diğerleri veri depolarını ve Bovee ve diğerleri ise varlıkları incelemiştir.

Tablo 2: Bütünlük Boyutunun Bilgisayar Bilimlerindeki Altı Temel Yaklaşımdaki Tanımları

Referans	**Tanım**
Wand ve Wang (1996)	Bir bilgi sisteminin gerçek dünya sisteminin her anlamlı durumunu temsil etme yeteneğidir.
Wang ve Strong (1996)	Verinin, üzerinde çalışılan görev için yeterli genişliğe, derinliğe ve kapsama sahip olma derecesidir.
Redman (1996)	Değerlerin bir veri topluluğuna dahil edilme derecesidir.
Jarke ve diğerleri (1995)	Veri kaynaklarında ve/ya da veri depolarında bulunan gerçek-yaşam bilgilerinin oranıdır.
Bovee ve diğerleri (2001)	Bilginin, bir varlığın tanımı için gerekli tüm parçaları taşımasıdır.
Naumann (2002)	Bir kaynaktaki eksik değerlerin sayısı ile evrensel ilişkinin boyutu arasındaki orandır.

Kaynak: Batini ve diğerleri, 2009: 16.

İstatistik alanında bütünlük boyutu, diğer boyutlar gibi belirsiz değildir. Her bir veri kaydının tam – eksik veri olmadan – ya da bütün olması anlamına gelmektedir. Buna rağmen bazı problemler çıkabilmektedir. Örneğin eksik veriler ile bilerek boş bırakılmış veri değerleri karıştırılabilmektedir (Karr ve diğerleri, 2006: 157).

Eurostat, bütünlük boyutunu başka türlü ele almıştır. Bütünlük boyutunun, bir bütün olarak Avrupa istatistik sisteminin, küresel boyutta kullanıcıların taleplerini ve mevcut istatistikleri karşılaştırarak ve istatistiksel kavramların uygunluğunu ve ihtiyaç duyulan istatistiklerin üretimi için gerekli zamanı dikkate alarak stratejik kullanıcıların, özellikle politikacıların ve medyanın, ihtiyaçlarına ve önceliklerine, ne derece cevap verdiğini ölçmek zorunda olduğunu belirtmiştir (Eurostat, 2000: 5).

Ancak eksik verilerin incelenmesi ve eksik verilerin tespit edilmesi durumunda yerine konulacak verilerin belirlenmesi ya da boş bırakılmasına karar

verilmesi aşamaları veri seti türlerine, veri setindeki değişkenlerin aralarındaki ilişkiye göre değişiklik göstereceğinden ve başlı başına bir yöntem olduğundan, bu çalışmada bütünlük boyutu, eksik verilerin incelenmesi olarak Bölüm-2 de ayrı olarak değerlendirilecektir.

1.4.3. Bütünlük Boyutunun Ölçümü

Ölçme, değişkenlere değer biçme veya aldıkları değerleri belirleme işlemidir. Ölçme işleminin önemi ise, verilere uygulanacak doğru istatistiksel tekniklerin seçilmesinden kaynaklanmaktadır (Bülbül, 2000: 21). Bütünlük boyutunun ölçümü de eksik verileri belirleme ve tanımlama işlemidir. Veri setinde yer alan eksik verileri belirlemeden ve tanımlamadan veri kalitesinde meydana getirdikleri eksiklikleri gidermeye çalışmak, veri kalitesinde daha önemli zararlara neden olabilir.

Batini ve Scannapieca'nın bütünlük boyutu tanımını referans göstererek Pipino ve diğerleri üç tür bütünlük boyutunun ölçümü için göstergeler önermişlerdir. Şema bütünlüğünü, şemada eksik olan girdilerin ve özelliklerin tüm girdi ve özelliklere oranının 1'den çıkarılması ile; sütun bütünlüğünü bir sütundaki eksik değerlerin tüm değerlere oranının 1'den çıkarılması ile; anakütle bütünlüğünü ise örneğin, araştırmaya dâhil edilmeyen eyaletlerin toplam eyalet sayısına oranının 1'den çıkarılması ile hesaplamayı önermişlerdir (Pipino ve diğerleri, 2002: 213).

İstatistik alanında ise bütünlük boyutunun tanımında bahsedildiği gibi Eurostat, Karr ve diğerleri hariç genellikle tüm kurum ve yazarlar bu boyutu doğruluk boyutunun içinde değerlendirmişlerdir. Eurostat, bütünlük boyutunun ölçümü için herhangi bir teknikten bahsetmemiştir. Karr ve diğerleri ise bu boyutun ölçümü için iki göstergenin kullanılabileceğini savunmuşlardır. Bu göstergeler; eksik olmayan özellik değerlerin sayısının tüm verilerin sayısına oranı ve eksik olan özellik değerlerin eksik olmayan özellik değerlerine oranıdır (Karr ve diğerleri, 2006: 157-158).

Shankaranarayanan ve Cai (2006: 302), yaptıkları çalışmada veri kalitesi boyutlarından sadece bütünlük boyutunu kapsam-bağımlı ve kapsamdan-bağımsız bütünlük olarak iki biçimde ele almışlardır. Bu boyutun ölçümü için, bilgi ürünü yaklaşıma dayalı bilgi ürünün üretimini simgeleyen bir sunum şeması olan IPMAP (Bilgi Ürünü Haritası) tekniğini kullanmışlardır.

İKİNCİ BÖLÜM
EKSİK VERİ VE EKSİK VERİ TAMAMLAMA YÖNTEMLERİ

Bilgi toplanan verilerden elde edilir. Toplanan verilerin eksik olması işletmelerin bilgiye dayalı elde ettikleri avantajlarını kaybetmesine sebep olabilmektedir. Eksik verilerin varlığı çoğu araştırma alanında bir biçimde ortaya çıkan bir sorundur. Ancak eksik verilerin ele alınmasına ihtiyaç duyulmasına rağmen, eksik veri problemi her zaman çözülmek istenmeyebilir ve bu problemi çözmek için hangi yöntemlerin kullanılacağı bazen belirsiz olabilmektedir (Buuren, 2012: 16).

Bu bölümde veri matirisi ve eksik veri ile ilgili genel bilgiler verildikten sonra eksik verinin tanımı, tarihsel gelişimi, eksik veri mekanizmaları, eksik veri yapıları, klasik eksik veri çözümleme yöntemleri, makine öğrenmesi yaklaşımlarıyla eksik veri tamamlama yöntemleri, derin öğrenme kavramı, derin üretici modellerden üretici çekişmeli ağlar, varyasyonel oto-kodlayıcılar ve eksik veri atamasında kullanılan üretici çekişmeli atama ağlar tanıtılmıştır.

2.1. VERİ MATRİSİ

Veri matrisi her istatistiksel çalışmanın temel unsurudur. İstatistiksel araçlar kullanılarak bir veri setinin analiz edilebilmesi için verilerin tablo biçiminde sunulması gerekir. Tablo biçimindeki gösterim şekline veri matrisi denir. Bir veri matrisindeki her sütun bir değişkeni (gösterge, ölçüm, bir anketteki sorular) ve her satırda bir gözlemi (ülkeler, bir anketteki kişiler, bir denemedeki konular) temsil eder. Veri matrisi, satırlar (gözlemler) ile sütunlardan (değişkenler) oluşan değerler (bilgiler) topluluğudur. Aşağıdaki Tablo 3 veri matrisine bir örnektir.

Tablo 3: Veri Matrisi Örneği

		Değişkenler											
		sabit asitlik	**uçucu asitlik**	**sitrik asit**	**artık şeker**	**klorür**	**serbest kükürt dioksit**	**toplam kükürt dioksit**	**yoğunluk**	**pH**	**sülfat**	**alkol**	**kalite**
	0	7,4	0,7	0	1,9	0,076	11	34	0,9978	3,51	0,56	9,4	5
Gözlemler	**1**	7,8	0,88	0	2,6	0,098	25	67	0,9968	3,2	0,68	9,8	5
	2	7,8	0,76	0,04	2,3	0,092	15	54	0,997	3,26	0,65	9,8	5
	3	11,2	0,28	0,56	1,9	0,075	17	60	0,998	3,16	0,58	9,8	6
	4	7,4	0,7	0	1,9	0,076	11	34	0,9978	3,51	0,56	9,4	5
	5	7,4	0,66	0	1,8	0,075	13	40	0,9978	3,51	0,56	9,4	5
	6	7,9	0,6	0,06	1,6	0,069	15	59	0,9964	3,3	0,46	9,4	5
	7	7,3	0,65	0	1,2	0,065	15	21	0,9946	3,39	0,47	10	7
	8	7,8	0,58	0,02	2	0,073	9	18	0,9968	3,36	0,57	9,5	7
	9	7,5	0,5	0,36	6,1	0,071	17	102	0,9978	3,35	0,8	10,5	5

Kaynak: Cortez ve diğerleri, 2009: 547.

Tablo 3 Portekiz "Vinho Verde" kırmızı şarap çeşitlerini üreten bir işletmenin şarap kalitesini sınıflandırmada kullandığı veri matrisinin ilk 9 gözlemini göstermektedir. Veri matrisinde yer alan değişkenler fizyokimyasal girişler; sabit asitlik, uçucu asitlik, sitrik asit, artık şeker, klorür, serbest kükürt dioksit, toplam kükürt dioksit, yoğunluk, pH, sülfat ve alkol değişkenlerinden oluşmaktadır.

Duyusal çıktı değişkeni kalite değişkeni, üretilen şarabın koku ve tadı üzerine şarap uzmanları tarafından verilen 0-10 arasındaki puanları ifade etmektedir.

2.2. EKSİK VERİ KAVRAMI

Belirli ya da tüm değişkenlerde eksik verilerinin olduğu veri setleri ve üzerinde çalışılan alanla ilgili problemlerin analizini zorlaştıran veri girişlerinin varlığı eksik veri olarak tanımlanmaktadır (Rubin, 1978: 20).

İşlemsel olarak (ya da sezgisel olarak), gerçekleştirilecek spesifik analiz için anlamlı bir değer gizlenirse, bir değişken için bir değer eksik olarak tanımlanır. Bu tür durumlara örnek olarak gelir; tansiyon, eğitim, yaş vb. değişkenler verilebilir (Raghunathan, 2016: 26).

Eksik veriler, gözlemlenmesi durumunda analiz için anlamlı olabilecek gözlemlenmemiş değerlerdir; diğer bir deyişle, eksik bir değer anlamlı bir değeri gizler (Little ve Rubin, 2020: 4).

2.2.1. Eksik Veri Yükleme Kavramı

İngilizce karşılığı "imputation" olan yükleme kavramı, "to impute" İngilizce fiilinden ve Latince "imputo" kökeninden gelmektedir. Hesaba katmak, atfetmek, yüklemek anlamlarında kullanımı bulunmaktadır.

Yükleme kavramı istatistiksel liteatürde "veriyi doldurmak" anlamında kullanılmaktadır. Amerika Nüfus Sayım Bürosu 1957 yılında yaptığı çalışmalarda yükleme kavramını bu anlamıyla kullanan ilk kurum olmuştur (Buuren, 2012: 25).

2.2.2. Eksik Veri Problemine Çözüm Yaklaşımlarının Tarihsel Gelişimi

Eksik veri problemlerinin kronolojisi incelendiğinde en az veri analizi kadar eskiye dayanmakta olduğu görülmektedir. Eksik veriye yönelik ilk çalışmalar 1900'lü yılların başına denk gelse de, Rubin (1976: 581), tarafından gerçekleştirilen eksik veri mekanizmalarının incelendiği çalışma bu alanın dönüm noktası olarak gösterilebilir.

Elderton ve Pearson (1910: 769) tarafından yapılan çalışma, ebeveynlerin alkol kullanımının etkilerini araştırmaktadır. Bu bağlamda araştırmacılar, eksik değere sahip değişkenleri ortalama ile veri yükleme metodu ile doldurmuşlardır.

Sturge ve Horsley (1911: 72) ise Elderton ve Pearson (1910)'un çalışmasında yer alan istatistiksel hatalar üzerine bir araştırma gerçekleştirmişlerdir.

McKendrick (1926: 98) tarafından, Beklenti Maksimizasyonu (EM-Expectation Maximization) algoritmasının daha farklı bir versiyonu olarak nitelendirilebilecek özyinelemeli tahmin ile ilgili ilk sistematik çalışma gerçekleştirilmiştir.

Allan ve Wishart (1930: 399) ise eksik verinin deney tasarımı üzerinde bulunduğu durumlar için çalışmalar gerçekleştirmişlerdir. Latin kare ve rastgele blok düzenindeki tasarımlara yönelik olarak bir (tek)eksik verinin değerini tahminlemeye çalışmışlardır.

Wilks (1932: 163) parçalı örneklerde anakütle parametrelerine ait dağılımların ve momentlerin bulunması üzerine yaptığı çalışmasında iki değişkenli normal dağılıma sahip olan bir anakütlenin parametrelerini eksik verileri kullanarak tahminlemeye çalışmıştır. Anakütle parametrelerini tahminleme aşamasında En Çok Olabilirlik (ML-Maximum Likelihood) yöntemini kullanmış ve büyük örneklerde varyans-kovaryans eşik değerlerini belirlemeye çalışmıştır.

Yates (1933: 129) ise bir (tek) eksik veriyi tahminleyebilmek amacıyla eksik verinin yerine atanabilen bir "x" değerinin hata sapmalarını minimize edecek genel bir matematiksel eşitlik önermiştir. Yates (1936: 121) tamamlanmamış Latin kare ve tamamlanmamış rastgele blok düzenine açıklık getirerek deney tasarımlarında yer alan eksik verileri tamamlamaya çalışmıştır.

Bose ve Mahalanobis (1938: 103) ve Bose (1938: 112), deney tasarımı üzerinde yalnızca incelenen değişkenlerin arasında doğrusal bir kombinasyon olduğu durumlara yönelik çalışmalar gerçekleştirmişlerdir.

Matthai (1951: 145), Wilks (1932: 163) tarafından iki değişkenli olarak incelenen formu üç değişken olacak şekilde ele almıştır. Daha sonra Lord (1955: 1227) tarafından gerçekleştirilen çalışmada da Wilks (1932: 163)'in iki değişkenli olarak incelediği durum üç değişkenli olarak incelenmiştir. Normal dağılıma gösteren üç değişkenli bir anakütleye ait parametrelerin eksik verilerden tahmin edilebilmesine yönelik olarak EM eşitliklerinin kolaylıkla elde edilebileceği özel bir yapıda incelemiştir.

Edgett (1956: 122), çok değişkenli regresyonda (Multivariate Regression) eksik veri içeren açıklayıcı değişkenlerle alakalı bir çalışma gerçekleştirmiştir. Normal dağılım sergileyen üç değişkenli bir anakütlenin parametrelerini tahminleyebilmek amacıyla eksik verinin olduğu bir değişkenin EM eşitliklerini elde etmeye çalışmıştır.

Healy ve Westmacott (1956: 03), gerçekleştirdikleri çalışmada deney tasarımı üzerinde, genelleştirilmiş EM algoritmasının temellerini oluşturmuşlardır. Bu bağlamda çalışmalarını üç adıma dayandırmışlardır:

a. Parametre değerleri (mevcut verilmiş olan) ile eksik veriler tahminlenir,

b. Güncellenmiş olan parametre değerlerini tahminlemek amacıyla eksik verilerin tahminlenmiş ve mevcut değerleri kullanılır,

c. En yaklaşık sonuç elde edilene kadar söz konusu işleme devam edilir (iterasyon).

Anderson (1957: 200) , Lord (1955: 1227) ve Edgett (1956: 122) tarafından gerçekleştirilen ve normal dağılım sergileyen üç değişkenli bir anakütle için yaptığı çalışmaları daha genel bir çerçeveye oturtarak bir çok durum için kolayca uygulanabilir hale getirmiştir.

Nicholson (1957: 523) ise, Wilks (1932: 163)'in iki değişkenli ve Lord (1955: 1227)'un üç değişkenli formda ele aldığı durumu "p" değişkenin içerildiği özel durumları incelemiştir.

Wilkinson (1957: 363), eksik verinin olduğu durumların kovaryans analizi ile çözümlenebilmesi amacıyla çalışmalar gerçekleştirmiştir. Eksik verilere karşılık gelen eşzamanlı ölçümlerin eldeki mevcut gözlemlerin analizi ile ilgili olmadığını ve tamamlanmış verilerin üzerinden kovaryans analizinin yapılması gerektiğini belirtmiştir. Daha sonra yaptığı çalışmada Wilkinson (1958: 360), varyans analizi gerçekleştirirken eksik veri olması halinde kullanılması için yeni eşitlikler geliştirmiştir.

Hartley (1958: 174), eksik veri olduğu durumlarda ML tahmini için genel bir metot ortaya koymuştur.

Dear (1959: 23), eksik veri olması halinde ML için temel bileşenler analizinin kullanılması üzerine bir çalışma gerçekleştirmiştir.

Buck (1960: 302) “k” değişken içeren bir anakütlenin eksik verilerini regresyon metotlarıyla tahminlenmesi ve bu tahmin değerlerinin kullanılarak bir varyans-kovaryans matrisinin elde edilmesi üzerine çalışmalarda bulunmuştur.

Glasser (1964: 834), eksik verinin olduğu durumlar için regresyon modelinin verimliliği ile, bağımsız değişkenler arasındaki korelasyon ve gözlemlerin eksik veri oranı arasında bir ilişki olduğunu belirtmiştir.

Afifi ve Elashoff (1966: 595), kendilerinden önce bu alanda yapılan çalışmaları (ortalama, varyans, korelasyon ve doğrusal regresyon) çalışmaları inceleyerek eksik verinin belli bir düzene göre oluşması halinde yapılacak tahminleme işlemlerini basitleştirecek yöntemler geliştirmişlerdir. Daha sonra Afifi ve Elashoff (1967: 10), iki rassal değişkenle tahminlenen doğrusal regresyonda eksik veri bulunması halinde bu değerleri tahminlemek amacıyla olasılık temelli yöntemlere başvurmuşlardır. Araştırmacılar elde ettikleri hata kare ortalamalarını (HKO) test etmek amacıyla en küçük kareler (EKK) metodunu uygulamışlardır.

Haitovsky (1968: 67) iki farklı yönteme dayanarak regresyon veri setlerindeki eksik verileri gidermek üzerine çalışmıştır. Uyguladığı ilk yöntemde eksik veri içeren tüm gözlemleri veri setinden dışlayarak EKK ile tahminleme gerçekleştirmiştir. İkinci yönteminde ise tüm değişkenlerin birbirleriyle olan kovaryanslarını normal denklemler aracılığıyla hesaplamıştır. Yalnızca tam gözlemler üzerinden EKK uyguladığı yaklaşımdan daha iyi sonuçlar sağladığını belirtmiştir.

Afifi ve Elashoff (1969a: 337), x ve y olmak üzere iki değişken üzerinde bazı gözlemler eksik olduğunda, y’nin x üzerine kurulan regresyon parametreleri için asimptotik dağılımlar türetmişlerdir. Sonraki aşamada bu tahminleri HKO ile kıyaslamışlardır.

Afifi ve Elashoff (1969b: 359), eksik veri olması halinde doğrusal regresyonda tahminlenen parametrelerin küçük örnek kullanıldığı durumlarda sapma ve verimliliklerini ölçümlemişlerdir. Ayrıca küçük ve büyük örnekler arasındaki etkinlik farkının korelasyon katsayısıyla (ρ) ve eksik verilerin yapısıyla alakalı olduğunu ifade etmişlerdir.

Hartley ve Hocking (1971: 783) ile Orchard ve Woodbury (1972: 691) tarafından yapılan çalışmalar tamamen geliştirilmiş EM algoritmasının temellerini oluşturmuştur.

Rubin (1976: 581)'in araştırmasında eksik veri üç sınıfa ayrılmış ve literatürde eksik veri mekanizmaları başlığıyla yer alan ayrım ortaya konmuştur. Bu sınıflandırma sayesinde eksik veri sorununun hangi yaklaşımlar kullanılarak giderilebileceği hususunda fikir sahibi olunabileceğine açıklık getirilmiştir.

Dempster ve diğerleri (1977: 1) tarafından yapılan çalışmada eksik veri problemlerinde kullanılması amacıyla ML tahminlerinin tekrarlı (iteratif) versiyonu olan EM algoritması tamamen geliştirilmiştir.

Louis (1982: 226) tarafından yapılan çalışmada, eksik veri problemlerinde ML tahminlerini bulmak amacıyla EM algoritması kullanıldığı durumlarda, gözlenen veri matrisinin çıkarılması için ve EM algoritmasının yakınsamasını hızlandırabilmek amacıyla bir metot geliştirilmiştir.

Little ve Rubin tarafından yayınlanan kitap, eksik veri alanının ilk sistematik çalışması olarak gösterilebilmektedir. Eksik veri varlığında meydana gelebilecek durumları örnekler üzerinden açıklık getirmeye çalışmışlardır. Eksik veri analizinin bu araştırmadan sonra daha yaygın bir hale geldiği belirtilmektedir (Little ve Rubin, 1987).

Daha sonra Little (1992: 1227), regresyon analizlerinde eşdeğişkenlerin eksik olduğu durumlarla ilgili kapsamlı bir çalışma gerçekleştirmiştir.

Jones (1996: 22) araştırmasında doğrusal regresyon analizinde eksik verinin olması halinde kukla değişkenler kullanılması önerilmekteyken, Vach (1994: 2)'ın çalışmasında ise yine kukla değişkenler lojistik regresyon üzerinde uygulamıştır.

1990'lı yıllardan itibaren çeşitli eksik veri tahmin metotları geliştirilmiş ve birçok farklı sektöre uygulanmıştır. 2000'li yıllarda ise araştırmacılar, eksik veri tahminleme işleminin sonuçlarından elde edilen değerlerin karar verme süreçlerinin üzerinde ne kadar hassas olduğunu irdelemeye başlamışlardır. Eksik veri girişlerini tahminlemek amacıyla yeni yöntemler bulabilmek için araştırmalar yapılmaktadır. Örneğin, sinir ağları gibi yapay zeka yöntemleri ve evrişimsel hesaplama algoritmaları gibi optimizasyon algoritmaları, eksik veri tahmini görevlerinde giderek daha fazla kullanılan yaklaşımlardan bazılarıdır (Dhlamini ve diğerleri, 2006: 280-287; Nelwamondo ve Marwala, 2007; 1297-1306)

Juhola ve Laurikkala (2013: 231) tarafından gerçekleştirilen araştırmada sınıflandırma sonuçlarını etkileyici etkisi olan kayıp veri miktarını ölçmeye

çalışmışlardır. %30, %20, %10 ve %0 kayıp veri içeren veri kümelerinde, sınıflandırmalar diskriminant analizi, en yakın komşu ve Bayesyen yaklaşımıyla yapılmıştır. İkiden fazla değişkenin olduğu %20-%30'lara yaklaşan ölçüde eksik veri gözlenen veri kümelerinde gerçekleştirilen analizlerden verimli sonuçlar sağlanırken, ikiden fazla değişken olması ve %10-%20 kayıp verinin olduğu durumların sorun oluşturabileceğini belirtmişlerdir.

2.2.3. Eksik Veri Oranı

Bir veri kümesinin içerdiği eksik verilerin kabul edilme yüzdesiyle ilgili farklı görüşler mevcuttur. Bennet (2001: 164), eksik veri miktarı eğer %10'dan fazlaysa eksik veri atama işlemlerinin yapılması gerektiğini; Schafer (1999: 3) ise %5 ya da daha azının önemli olmadığını belirtmiş ve veri kümesindeki eksikliğin %5'ten fazla olduğu durumlarda zaten eksik veri sorununun çözümlenmesi gerektiğini ifade etmiştir.

Veri kümelerindeki eksik veriler, bilgilere dayanarak ulaşılan analiz, çıkarımlar ve sonuçlara etki eder. Veri setindeki eksik verilerin oranının artışı istatistiksel analizin ve makine öğrenmesi algoritmalarının performansı üzerinde daha fazla belirgin hale gelmektedir. Araştırmacılar, büyük ölçekli veri setlerinde eksik veri oranı az olduğunda makine öğrenme algoritmaları üzerindeki etkinin önemli olmadığını göstermiştir (Polikar ve diğerleri, 2010: 3817; Ramoni ve Sebastiani 2001: 147; Tremblay ve diğerleri, 2010: 1). Bu, belirli makine öğrenimi algoritmalarının bazı eksik veri oranlarına rağmen çalışabilmeleri için doğal algoritmik tasarıma sahip olmasıyla ilişkilendirilebilir (Leke ve Twala, 2019: 3).

Twala, eksik veri oranlarındaki bir artışla, örneğin oranın % 25'in üzerinde olduğu durumlarda, makine öğrenme algoritmalarının hata payı ve performans seviyelerinin önemli ölçüde azaldığının gözlemlendiğini, hata payı ve performanstaki bu düşük seviyelerden dolayı, eksik veri sorununu çözmek için daha karmaşık ve güvenilir yaklaşımlar gerekli olduğunu söylemiştir (Twala, 2009: 373).

2.2.4. Eksikliğin Deseni ve Yapısı

Eksikliği karakterize etmek için eksikliğin deseni ve yapısı kullanılır. Eksiklik deseni, bir veri setinde değerlerin nasıl eksik olduğunu görsel olarak gösterirken, eksiklik yapısı eksik değerlerin olasılıklı tanımıyla ilgilenir. Eksik veri deseni, eksik değerlerin potansiyel bir tam veri matrisindeki (yani %100 yanıt oranına sahip veri matrisindeki) konumunu tanımlar.

Eksik verilerin ortaya çıkma şekli Tablo 4, 5, 6 tarafından verilen üç desende gruplandırılabilir. Tablo 4, sütun S7'de görüldüğü gibi, sadece bir değişkende eksik verilerin varlığıyla tanımlanan senaryo olan tek değişkenli bir deseni göstermektedir. Tablo 5, eksik verilerin dağıtılmış ve rasgele bir şekilde meydana geldiği bir senaryo olan rasgele bir eksik veri desenini göstermektedir. Son model, Tablo 6'da gösterilen monoton eksik veri desenidir. Bu desen, eksik verilerin birden fazla değişkende mevcut olabileceği ve anlaşılması kolay olduğu durumlarda tekdüzen bir desen olarak da adlandırılmaktadır (Ramoni ve Sebastiani, 2001: 156).

Tablo 4'te tek değişkenli eksik veri desenini gösteren örnek bir veri matrisi gösterilmiştir.

Tablo 4: Tek Değişkenli Eksik Veri Desenini Gösteren Örnek Bir Veri Matrisi

Gözlemler	S1	S2	S3	S4	S5	S6	S7
1	0.38	0.25	0.45	0.56	0.65	0.74	0.87
2	0.96	0.17	0.24	0.47	0.63	0.87	?
3	0.41	0.54	0.43	0.85	0.21	0.34	?
4	0.12	0.31	0.51	0.68	0.93	0.35	?

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 4 de değişken isimleri S1, S2, S3, S4, S5, S6, S7 ve gözlem numaraları 1, 2, 3, 4 olarak gösterilmiştir. S7 isimli değişkenin 2, 3 ve 4 numaralı gözlemlerinde "?" karakteri eksik değerleri belirtmek için kullanılmıştır. S7 sütununda eksik gözlemlerin oluşturduğu desen tek değişkenli eksik veri desenini göstermektedir. Tek değişkenli eksik veri desenine örnek olarak sıcaklık değerlerini ölçen sensörlerin

çok yüksek ya da çok düşük sıcaklıklarda değerleri kaydememesi sonucu ortaya çıkan eksik veri durumu verilebilir.

Tablo 5'te rasgele eksik veri desenini gösteren örnek bir veri matrisi gösterilmiştir.

Tablo 5: Rasgele eksik veri deseni(Ardışık)

Gözlemler	S1	S2	S3	S4	S5	S6	S7
1	0.38	0.25	0.45	0.56	?	0.74	0.87
2	0.96	0.17	0.24	?	0.63	?	0.04
3	0.41	0.54	?	0.85	0.21	0.34	0.054
4	?	0.31	0.51	0.68	0.93	0.35	?

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 5 te değişken isimleri S1, S2, S3, S4, S5, S6, S7 ve gözlem numaraları 1, 2, 3, 4 olarak gösterilmiştir. S1 isimli değişkenin 4 numaralı gözlemi, S3 isimli değişkenin 3 numaralı gözlemi, S4 isimli değişkenin 2 numaralı gözlemi, S5 isimli değişkenin 1 numaralı gözlemi, S6 isimli değişkenin 2 numaralı gözlemi ve S7 isimli değişkenin 4 numaralı gözlemindeki "?" karakteri eksik değerleri belirtmek için kullanılmıştır. Değişkenlerde görülen eksik değerlerin ardışık bir şekilde ortaya çıkmasıyla rasgele eksik veri deseni oluşmaktadır. Ardışık eksik veri deseni istatistiksel analizlerde ve araştırmalarda kullanılmak için toplanan verisetlerinde sıklıkla görülebilir. Örnek olarak, havayı oluşturan bileşenleri ve sıcaklık değerlerini toplayan sensörlerin herhangi bir zamanda arızalanması nedeniyle bileşenleri ve sıcaklık değerlerini kaydedemesi sonucu ortaya çıkan eksik veri durumu verilebilir.

Tablo 6'da monoton eksik veri desenini gösteren örnek bir veri matrisi gösterilmiştir.

Tablo 6: Monoton eksik veri deseni

Örnek	S1	S2	S3	S4	S5	S6	S7
1	0.38	0.25	0.45	0.56	0.65	0.74	?
2	0.96	0.17	0.24	0.47	0.63	?	?
3	0.41	0.54	0.43	0.85	?	?	?
4	0.12	0.31	0.51	?	?	?	?

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 6 da değişken isimleri S1, S2, S3, S4, S5, S6, S7 ve gözlem numaraları 1, 2, 3, 4 olarak gösterilmiştir. S4 isimli değişkenin 4 numaralı gözleminde, S5 isimli değişkenin 3 ve 4 numaralı gözlemlerinde, S6 isimli değişkenin 2, 3, 4. gözlemlerinde ve S7 isimli değişkenin 1, 2, 3 ve 4. gözleminde "?" karakteri eksik değerleri belirtmek için kullanılmıştır. Değişkenlerde görülen eksik değerler monoton eksik veri desenini oluşmaktadır. Monoton eksik veri desenine örnek olarak uzun dönemli yapılan araştırmalarda katılımcılardan birinin yada birkaçının araştırma tamamlanmadan (ölüm ya da kişiye ulaşılamaması, alerjik reaksiyonlar) çıkması verilebilir.

Tezin kapsamında ele alınan eksik veri deseni, rastgele eksik veri desenidir ve mekanizmalar tamamen rasgele eksik veri (MCAR-Missing Completly Random) ve rastgele eksik veri mekanizmasıdır(MAR-Missing At Random).

2.3. EKSİK VERİ MEKANİZMALARI (TÜRLERİ)

Veri setinde yer alan eksik verileri belirlemeden ve tanımlamadan veri kalitesinde meydana getirdikleri eksiklikleri gidermeye çalışmak, veri kalitesinde daha önemli zararlara neden olabilmektedir. Eksik verilere sahip bir veri setinde, verilerin eksik olmasının nedenini belirlemek önemlidir. Eksiklik ile ilgili açıklama bilindiğinde, eksik veri ataması için uygun bir yöntem seçilebilir ya da türetilebilir, bu da daha yüksek etkililik ve tahmin doğruluğu ile sonuçlanmaktadır. Çoğu durumda, veri toplayıcıları eksik verinin nedenini bilmektedirler. Oysa istatistikçiler ve veri kullanıcıları kendilerine sunulan bilgilerle analiz yaptıklarında bu nedenlerden haberdar olmayabilirler. Bu tür senaryolarda, veri kullanıcıları eksik verilerin gözlemlenen verilerle ilişkisini anlamak için veri analizine yardımcı olabilecek başka teknikler kullanmak zorunda kalabilmektedirler. Bunun sonucunda eksiklikle ilgili bazı olası nedenler elde edilebilmektedir. Veri kümesindeki bir değişken ya da özellik, diğer değişkenlerde bulunan eksikliğin nedenini açıklamaya yardımcı olması durumunda bir yapı olarak görüntülenmektedir.

Literatürde en çok kullanılan eksik veri tanımlama sistemi Rubin (1976: 582) tarafından geliştirilmiştir. Rubin; eksik veri türlerini tamamen rasgele eksik veri

(MCAR-Missing Completely At Random), rasgele eksik veri (MAR-Missing At Random) ve rassal olmayan eksik veri (MNAR-Missing Not At Random) olmak üzere üç grupta incelemiştir.

Anketler yoluyla toplanan veri kümelerinde, eksik veri girişleri olan değişkenler genellikle insanların ifşa etmeye utandıkları ayrıntılarla ilişkilendirilmektedir. Bununla birlikte, bu tür bilgiler genellikle eksik olmayan verilerden elde edilebilmektedir. Örneğin, X ve Y sırasıyla gelir ve eğitim seviyelerini gösteren değişkenler olsun. Düşük gelirli insanlar gelirlerini açıklamak istemeyebilirler, ancak eğitim seviyelerini açıklayabilirler. Veri kullanıcıları, X değişkeni hakkında bir fikir edinmek için Y değişkenindeki bilgileri kullanabilirler.

Rubin (1978: 20) Y isimli bir veri kümesindeki eksik ve tam değerlerden oluşan veri matrisinini , Y = $\{Y_0, Y_m\}$ şeklinde ifade etmiştir. Burada Y_o Y'nin gözlenen bileşeniyken Y_m eksik bileşenidir.

Eksik veri varlığı, farklı sektörlerde kaliteli ve tam verilerin mevcut olmasına bağlı olarak çeşitli sorunlara neden olmaktadır. Bu durum ise farklı disiplinlerdeki eksik veri problemini inceleyebilmek için birçok farklı yöntemin ortaya çıkmasını sağlamıştır. Eksik veri problemlerinin çözüm metotlarının kabul edilebilir metotlar olması ise eksikliğin niteliğine bağlı olduğu belirtilmektedir (Rubin, 1978:22; Allison, 2000: 301).

2.3.1. Tamamen Rasgele Eksik Veri(Missing Completely At Random-MCAR)

MCAR varsayımı, bir değişkendeki kayıp verilerin olasılığının, bu değişkenin kendi değeriyle ya da veri setindeki diğer herhangi bir değişkenin değeriyle ilişkili olmadığını ifade etmektedir. Bu ilişki matematiksel olarak Little ve Rubin (2014: 21) tarafından aşağıdaki denklem (2.1) ile ifade edilmiştir.

$$P(M|Y_0, Y_M) = P(M) \qquad (2.1)$$

M ∈ {0,1} ve Y veri matrisinde tam ve eksik verilerin girişlerini göstermek için oluşturulmuş ikili değerlerden (binary) oluşan matristir. Y veri matrisinde değişkenlerin bilinen ve bilinmeyen (eksik) değerleri için oluşturulacak M isimli matriste değerler biliniyorsa M = 1, bilinmiyorsa / eksikse M = 0 olur. Y_0, Y cinsinden gözlemlenen değerleri, Y_m ise Y'nin eksik değerlerini temsil etmektedir. Denklem (2.1) Y matrisindeki bir değişkende eksik verilerin bulunma olasılığının Y_0 ya da Y_m ile ilişkili olmadığını göstermektedir.

2.3.2. Rasgele Eksik Veri (Missing At Random-MAR)

MAR varsayımı, bir değişkendeki kayıp verilerin olasılığının, analizdeki diğer değişkenler kontrol altına alındığında, bu değişkenin kendi değeri ile ilişkisiz olduğunu ifade etmektedir (Schafer, 2000: 153). Bu ilişki matematiksel olarak (Marwala, 2009: 12) tarafından denklem (2.2) 'deki gibi ifade edilmiştir:

$$P(M|Y_0, Y_M) = P(M|Y_0) \tag{2.2}$$

Burada M ∈ {0,1} eksik veri göstergesidir ve Y matrisindeki değerler biliniyorsa M = 1, Y bilinmiyorsa / eksikse M = 0 olur. Y_0 Y cinsinden gözlemlenen değerleri temsil ederken, Y_m Y'nin eksik değerlerini temsil eder. Denklem (2.2) Y isimli veri kümesinde eksik veri bulunma olasılığı Y_0 yani veri kümesindeki eksik olmayan gözlemlere bağlı olduğunu göstermektedir. Örneğin, iki değişkeni olan bir veri kümesini, aylık harcama ve aylık geliri göz önünde bulunduralım. Yüksek gelirli kişiler aylık harcama bilgilerini vermek istemeyebilirler, ancak düşük gelirliler aylık harcama bilgilerini vermekten çekinmeyebilirler. Bu durumda aylık harcama değişkeninde ortaya çıkan eksik veri değerlerinin yüksek gelirli kişilerle ilişkisi olduğu gözlenebilir. Dolayısıyla aylık harcama değerlerinde eksik verilerin bulunmasını bireyin gelir seviyesiyle ilişkilendirmek mümkündür.

2.3.3. Rasgele Olmayan Eksik Veri(Missing Not At Random-MNAR)

Üçüncü eksik veri mekanizması, MNAR ya da göz ardı edilemeyen (non-ignorable) eksik veri mekanizmasıdır. MNAR mekanizması, bir değişkende eksik veri bulunma olasılığının değişkenin kendisine bağlı olduğunda ortaya çıkan eksiklik mekanizmasıdır (Allison, 2000: 302). Bu gibi senaryolarda, eksik verilerin ortaya çıkma olasılığı rastgele olmadığından, veri kümesindeki diğer değişkenleri kullanarak eksik verileri tahmin etmek imkansızdır. MNAR, modellemesi en zor olan eksik veri mekanizmasıdır ve bu değerlerin tahmin edilmesi oldukça zordur (Rubin, 1978: 20). Örneğin, bazı yüksek gelirli kişiler aylık harcamalarını açıklarken diğerleri açıklamak istemeyebilirler, aynı durum düşük gelirli kişiler için de gözlenebilir. MAR mekanizmasından farklı olarak, bu durumda aylık harcama değişkenindeki eksik girdiler, gelir değişkeni ya da başka bir değişkenle doğrudan bağlantılı olmadıkları için göz ardı edilemez.

Bu türdeki eksik verilerin bulunduğu verilerle yapılan model geliştirme tahminleri genellikle hatalı çalışır. Bu mekanizmanın olasılıklı bir matematiksel formu kolay değildir, çünkü bu mekanizmadaki veriler MAR ya da MCAR değildir (Marwala, 2019: 18).

2.4. KLASİK EKSİK VERİ ÇÖZÜMLEME YÖNTEMLERİ

Bir veri setinde verinin eksiklik durumuna bağlı olarak, istatiksel yazılım paketlerinde kullanılan çeşitli veri atama teknikleri bulunmaktadır. Bu teknikler, satır boyunca veri silme ve daha hassas yapay zeka ve istatistiksel yöntemlerin uygulanmasıyla karakterize edilen yaklaşımlara geçme gibi temel yaklaşımları içerir. Bir sonraki alt başlıkta en sık kullanılan eksik veri çözümleme yöntemlerinin bazıları sıralanmıştır. Temel ve sade yaklaşımlar ile başlanmış ve daha karmaşık ve etkili yöntemlerden bahsedilmiştir. Bu bölümde sunulan çözümleme yöntemleri liste halinde ya da satır bazında silme, ikili silme, ortalama atama, EM atama, sıcak ve soğuk deste atama, çoklu atama ve regresyon yöntemleridir.

2.4.1. Liste Bazında Ya da Satır Bazında Silme

Birçok istatistiksel yaklaşım veri setindeki sütunlardan herhangi birinin eksik veri girişine sahip olduğu durumlarda eksik verileri silme yoluna gitmektedir. Bu yaklaşım liste ya da satır bazında veri silme olarak adlandırılır ve eksik veri barındıran sütun ya da özellik değişkenindeki gözlemin veri setinden çıkarılması işlemidir. Liste bazında veri silme en basit ve en kolay eksik veri problemleri çözme yöntemi olmasına rağmen, veri analizi görevi için gerekli olan veri setindeki kayıt sayısının büyüklüğünü azaltma eğiliminde olduğu için en az tavsiye edilen yöntemdir. Bu yaklaşım veri setindeki eksik veri oranının az olduğu durumlarda uygulanabilir. Eksik veri oranının veri setindeki oranı yüksek ise liste bazında silme işlemi veri analizi sonuçlarının sapmalı olmasına sebep olmaktadır (Marwala, 2019: 7). Liste bazında silme yöntemine örnek olması için aşağıda Tablo 7 de bir veri seti verilmiştir.

Tablo 7: Liste Bazında Silme işlemi Öncesi Örnek Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
2	35	E	5000
3	30	E	4500
4	32	.	4250
5	42	K	4000
6	.	K	5000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur

Tablo 7 yaş, cinsiyet ve aylık gelir harcamalarını gösteren değişkenlerden oluşmaktadır. Tablo 7' de 4. ve 6. satırlarda eksik veriler bulunmaktadır. Bu durumda "Yaş" değişkeni için ortalamalar hesaplanmak istendiğinde hata ile karşılaşılacaktır. Liste bazında silme yöntemi kullanıldığında 4.ve 6. gözlem veri setinden çıkarılacaktır. Liste bazında silme yönteminde eksik veri içeren satırlar veri setinden çıkarıldıktan sonra elde edilen yeni veri seti Tablo 8 deki gibi olur.

Tablo 8: Liste Bazında Silme işlemi Sonrası Örnek Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
2	35	E	5000
3	30	E	4500
5	42	K	4000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 8' de eksik veriler bulunmamaktadır ve bu yeni oluşturulmuş veri seti ile analiz yapıldığında herhangi bir hata ile karşılaşılmayacaktır.

Liste bazında silme yöntemi veri setindeki faydalı olabilecek diğer verilerin silinmesine sebep olacağı için dezavantajları olan bir yöntemdir. Araştırma konusu kapsamında toplanan veri setinde, eksik veri yapısı rassal olmadığında problemlere yol açabilmektedir (Olinsky ve diğerleri, 2003: 53).

2.4.2. Çiftler Bazında Silme

Çiftler bazında silme, varyans-kovaryans matrisine odaklanan bir prosedürdür. Bu matrisin her elemanı, o eleman için mevcut olan tüm verilerden tahmin edilir.

Çiftler bazında silme, bir gözlemin belirli bir analiz için gerekli bir değişkeninde eksik gözlem olduğu, ancak bu gözlemdeki eksik olmayan verilerin gerekli tüm değişkenlerin mevcut olduğu analizlere dahil edilmesini içerir. Çiftler bazında silme kullanıldığında, analiz için kullanılan toplam N büyüklüğündeki veri setinin parametre tahminleri tutarlı olmayacaktır. Zaman içinde bazı noktalardaki eksik değerler nedeniyle, diğer parametreler için tam gözlem karşılaştırması yapılırken, çiftler bazında silme, % 100'ün üzerindeki korelasyonlar gibi imkansız matematiksel durumlar oluşturabilir (Enders, 2010 : 41-42).

Çiftler bazında silme tekniği korelasyon-kovaryans ilişkisinin, ortalama ve standart sapma gibi ölçülerin elde edilmesi için eksik veri barındırmayan gözlemleri kullanmaktadır. Analiz eksik gözleme sahip olmayan değişken ya da değişken çiftlerinin kullanılmasıyla yapılmaktadır. Çiftler bazında silme yöntemi ile iki

değişken arasındaki gözlem sayısı maksimum seviyede kullanılarak korelasyon katsayıları hesaplanır ve buna bağlı olarak ikili değişkenlerin tüm bilgisini kullanmak mümkün olmaktadır (Alpar, 2017: 33).

Bu yaklaşım ile iki değişkenli analizlerde eksik gözlem barındırmayan değişkenlerin kullanılmaktadır. Ancak bu yaklaşım, eksik veri mekanizması MAR ya da MNAR türünde olduğu veri setlerinde sapmalı tahminlere yol açacağından hatalı sonuçlara neden olabilmektedir. Buna karşılık çiftler bazında silme yaklaşımı eksik veri mekanizması MCAR türünde olduğunda uygulanabilmektedir (Marwala, 2019: 8).

Çiftler bazında silme yöntemine örnek olması için bir önceki başlıkta kullanılan örnek veri seti eksik girişleri değiştirilerek tekrar kullanılmıştır. Veri seti aşağıda Tablo 9' da gösterilmiştir.

Tablo 9: Çiftler Bazında Silme İşlemi Öncesi Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
2	35	.	.
3	30	E	4500
4	32	E	4250
5	42	K	4000
6	.	K	5000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur.

Çiftler bazında silme yönteminde veri setinde bulunan eksik verilere karşılık gelen satırların çıkarılması yerine üzerinde işlem yapılmak istenen sütundaki eksik veriler çıkarılır. Tablo 9 ' da 2. ve 6. gözlemlerde eksik veriler bulunmaktadır. Tablo 9' da ortalama aylık gelir hesaplanmak istenirse ve çiftler bazında silme yöntemi uygulanırsa 2. gözlem veri setinden çıkarılacaktır. İşlem sonrası elde edilen veri seti Tablo 10' da gösterilmiştir.

Tablo 10: Çiftler Bazında Silme İşlemi Sonrası Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
3	30	E	4500
4	32	E	4250
5	42	K	4000
6	.	K	5000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 10 çiftler bazında silme yöntemi sonrası elde edilen veri setidir. Veri setinde 2. gözlemin çıkarılması sonrasında aylık gelir değişkeni için eksik verinin bulunmadığı sütun elde edilmiştir. Böylece ortalama aylık gelir hesaplamasında herhangi bir hata ile karşılaşılmayacaktır.

Yaş değişkenindeki veriler kullanılarak ortalama yaş değeri hesaplanmadan önce çiftler bazında silme yöntemi uygulanmıştır. Yöntem ile 6. gözlem verisetinden çıkarılmıştır. Oluşan yeni veri seti Tablo 11' de gösterilmiştir.

Tablo 11: Veri Setindeki Ortalama Yaş Değerini Bulmak için Uygulanan Çiftler Bazında Silme İşlemi Sonrası Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
2	35	.	.
3	30	E	4500
4	32	E	4250
5	42	K	4000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 11 çiftler bazında silme yöntemi sonrası elde edilen yeni veri setidir. "Yaş" değişkeninde eksik veri girişi bulunmamaktadır. Ortalama yaş değeri hesaplanıldığında herhangi bir hata ile karşılaşılmayacaktır.

2.4.3. Aritmetik Ortalama – Modİle Veri Yükleme

Bu yaklaşım, değişkendeki eksik verilerin eksik olmayan değerlerin ortalamasının ya da modunun değeri ile değiştirerek çalışmaktadır. Çiftler bazında veri silme yaklaşımı gibi eksik verilerin yerine kullanılan değerler ile sapmalı tahminlerin elde edilmesi olasılığı yüksektir (Allison, 2002: 37).

Sürekli ya da sayısal değerleri olan veri setindeki değişkenler için eksik veri girişleri, ilgili değişkenlerin ortalaması ile değiştirilir. Diğer taraftan, kategorik ya da nominal değerlere sahip değişkenler için, eksik veriler, ilgili değişkeninin en yaygın ya da modsal değeri kullanılarak değiştirilir. Aritmetik ortalma-mod ile veri yükleme, eksik veriler MCAR mekanizması olduğunda etkilidir (Little ve Rubin, 2014: 22).

Ortalamanın eksik değer yerine yerleştirilmesi, değişken üzerindeki varyansı azaltırken kovaryanslar ve korelasyonları olumsuz etkiler. Ayrıca, bu yöntem standart hataları tahmin etmeyi zorlaştırır (Graham, 2012: 4).

Aritmetik ortalama-mod ile veri yükleme yönteminin kullanımını göstermek için Tablo 12' de örnek veri seti verilmiştir.

Tablo 12: Aritmetik Ortalama İle Veri Yükleme Öncesi Örnek Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
2	35	.	.
3	30	E	4500
4	32	E	4250
5	42	K	4000
6	.	K	5000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur.

$$Ortalama = \frac{29+35+30+32+42+50}{6} = 36.33 \qquad (2.3)$$

Tablo 12 de "Yaş" değişkeninde 6.gözlemde eksik veri bulunmaktadır. Eksik değer, aritmetik ortalama-mod ile veri yükleme yöntemi kullanıldığında denklem (2.3) ile elde edilden değeri alacaktır. Denklem (2.3), "Yaş" değişkenindeki tam değerlerin gözlemlerin sayısına (tam gözlem içeren) bölünmesiyle elde edilmiştir. Yaş değeri ondalıklı sayı olarak yazılamadığından tam değer olarak "36" yazılır. Elde edilen yeni veri seti Tablo 13' te gösterilmiştir.

Tablo 13: Aritmetik Ortalama İle Veri Yükleme Sonrası Veri Seti

Gözlem	Yaş	Cinsiyet	Aylık Gelir(TL)
1	29	E	4000
2	35	.	.
3	30	E	4500
4	32	E	4250
5	42	K	4000
6	36	K	5000
7	50	E	6000

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 13 aritmetik ortalama ile veri yükleme sonrası 6. gözlemin 36 değerini aldığı yeni veri setini göstermektedir. Yeni veri seti istatistiksel hesaplamalar sırasında hata vermeyecek duruma gelmiştir.

Büyük verilerde uygulanma zorluğunun olmasının yanında bütün verinin sistemin hafızasına yüklenip hesaplanması gerekliliği yöntemin dezavantajı olarak görülebilir (Şeker ve Eşmekaya, 2017: 13).

2.4.4. Eksik Veri Yerine Atama Yapmak

İstatistikte atama, eksik verilerin ikame edilmiş değerlerle değiştirilmesi, böylece eksik verilerin varlığından kaynaklanan olumsuzlukların ele alınması sürecidir. Atama teknikleri tekli ve çoklu atama olarak kategorize edilebilir. Tekli atama, eksik bir değerin sadece bir tahmini değerle değiştirilmesini içerirken, çoklu

atamada ise eksik olan her bir girişin yerine bir M tahmini değer seti yerleştirilir (Graham, 2012: 8).

2.4.4.1. Tekli Atama Yöntemleri

Tekli atama yöntemlerinde veri kümelerindeki eksik değerler bilinen başka gözlem değerleriyle değiştirilmektedir. Aşağıda tekli atama yöntemlerinden birkaçı sıralanmıştır.

a) Beklenti Maksimizasyon-Beklenti En Büyükleme (EM-Expectation-maximization), eksik veri içeren olasılıklı yöntemlerde parametre tahmini için tasarlanmış model tabanlı bir atama tekniğidir. Beklenti maksimizasyonu, beklenti adımı (E) ve maksimizasyon (M) adımı olarak iki adımın art arda tekrarlanmasıyla gerçekleşir. E-adımı modelde eksik verinin tamamlanması üzerine parametrelerin o anki tahminlerini kullanarak bir log-olabilirlik beklentisi fonksiyonu oluşturur. M adımı parametre değerlerini log-olabilirlik beklentisini maksimize edecek şekilde günceller. Yani bu iki adımın her biri diğerinin girdisini hesaplayarak birbirini besler. EM adımları tahmindeki hata miktarı belirli bir oranın altına düşene kadar yinelenir (Dempster ve diğerleri, 1977: 1).

b) Sıcak-Soğuk Deste Atama: Eksik verileri doldurmanın diğer bir yolu ise rastgele seçilen değerler kullanmaktır. Bu ise gözlenen veriden rastgele bir değerin seçilmesiyle sıcak deste (hot deck) veya benzer değişkenleri içeren başka bir veri setinden seçilmesiyle soğuk deste (cold deck) ile gerçekleştirilebilir. Aşağıda bu iki rastgele veri yükleme yöntemi ile ilgili bilgiler verilmiştir.

a. Sıcak Deste Atama: Bu atama tekniğinde, veri setindeki eksik veriler, ilgili değişkende yer alan benzer veriler ile doldurulur. Bu teknik üç farklı şekilde uygulanabilir (Çilingirtürk ve Aktaş, 2010: 76-77). Birinci uygulama şekli, tam verilerden rassal olarak bir değerin seçilmesi ve eksik verinin yerine atanmasıdır. rassal olarak seçilip yazılabilir. Tam olan herhangi bir değerin seçilme olasılığı, o değişken için verilen cevapların farklılığına bağlıdır. Örneğin 100 kişilik bir örneklemde sadece 3 kişi 15 yaşında ve 30 kişi de 35 yaşında ise, eksik veri için 35 değerinin yazılma

olasılığı 15 değerinin yazılma olasılığının 10 katıdır. Sıcak deste atama tekniğinin ikinci uygulanma biçimi ise, ilişkili değişkenleri kontrol altına alarak değer atamaktır. Örneğin yaş değişkenin eksik olması kişilerin cinsiyeti ile ilişkili ise, örneğin kadınlar erkeklere göre yaş sorusu için daha fazla eksik veriye sahip ise, eksik verilere atanacak değer kadınların yaş sorusuna verdikleri cevaplardan rassal olarak belirlenir. Sıcak deste atama tekniğinin üçüncü uygulanma biçiminde eksik veriye değer atamak için en yakın komşunun ilgili değişkene ait değeri kullanılır. Örneğin, 11. kişinin yaş sorusu eksik ise bu kişinin yaşı için 10. veya 12. kişinin yaş sorusuna verdiği cevap kullanılır.

b. **Soğuk Deste Atama:** Bu teknikte; aynı konuda yapılmış daha önceki araştırmalardan, aynı araştırmanın önceki uygulamalarından veya dış kaynaklardan elde edilen sabit bir değer, ilgili değişkendeki tüm eksik verilerin yerine atanır (Little ve Rubin, 2014: 25).

Sıcak ya da soğuk deste atama, basitliklerinden dolayı popülerdir ve verilere uyması için kullanılan model hakkında güçlü varsayımlar yapmaya gerek yoktur. Öte yandan, atama stratejisinin eksik veri kümesine göre sapmalarda bir azalmaya yol açmayacağı da belirtilmiştir (Marwala, 2019: 9).

2.4.4.2. Regresyon Yöntemleri

Regresyon ile atama yönteminde veri setindeki eksik veriler, eksik veri içermeyen diğer veriler kullanılarak kurulan bir regresyon denklemine göre bulunmaktadır. Regresyon denklemi, eksik veri içeren değişkenin tamamlanan, eksik veri içermeyen diğer değişkenlerin ise tamamlayıcı konumunda olacağı şekilde kurulur. Tamamlanan değişkende eksik olan gözlemler, diğer değişkenlere ait değerlerin bu denklemde yerine konulmasıyla tahmin edilirler (Buuren, 2012: 64).

Regresyon yöntemi, belirli bir değişken için mevcut tüm verileri içeren, gözleme bağlı regresyon denkleminin oluşturulmasını içerir. Regresyon denklemi oluşturmak için, eksik veriler içeren değişken denklemdeki bağımlı değişken olarak kabul edilirken, diğer tüm değişkenler bağımsız değişkenler (tahminleyiciler) olarak kabul edilir. Eksik verilere sahip gözlemlerin ilgili değişkeni çıktı olarak ve

diğerlerinin tümü de model denklem girdileri olarak regresyon denkleminde kullanılır. Bu denklem ile eksik veri yerine geçecek değerler tahmin edilir (Little ve Rubin, 2014: 24).

Tüm bu eksik giriş değerleri tahmin edilip değiştirilene kadar eksik veri girişleri olan değişkenler için regresyon denklemleri oluşturma süreci tekrar tekrar ve sırayla yapılır. Bu, eksik veri girişlerine sahip bir değişkenin, diğer değişkenler için bilinen değerlere sahip gözlemleri kullanarak kendisi için oluşturulmuş bir modeli olacağı anlamına gelir.

Doğrusal regresyon, bağımlı değişkenin bağımsız değişkenlerce doğrusal bir eşitlikle açıklanması yöntemidir. Geleneksel eksik veri yöntemlerinden farklı olarak, regresyon atama, MAR'daki eksikliği etkileyen değişkenlerin, kullanılan istatistiksel modellere dâhil edilmesi koşuluyla, hem MCAR hem de MAR altında yansız parametre tahminleri verir. Aynı zamanda eksik veriye yönelik atamaları iyileştirmek için mevcut verilerin tümünü kullanarak mevcut verilerin kullanımını en üst düzeye çıkarır. Regresyon tahmininin dezavantajı ise, eksik değerler için koşullu araçların kullanılmasından dolayı, tüm değerler regresyon çizgisi etrafında olacak ve bu durumda değişkenler arasındaki ilişkileri doğal olarak güçlendirecektir. Bu, bir regresyon denkleminde standart hataların küçümsenmesine (daha az değer vermeyi) yanı sıra, verilerin değişkenliğinin hafife alınmasına yol açacaktır (Lee ve diğerleri, 1994: 231).

Tablo 14 Regresyon yöntemi için kullanılmış örnek veri setini göstermektedir. Tablo 14'ün 4. gözlemindeki eksik veri girişini tahmin etmek için oluşturulan regresyon denklemi S2, S3, S4, S5, S6, S7 değişkenlerini dikkate alacaktır. Denklem (2.4) oluşturulan yeni regresyon denklemini ve i_2, i_3, i_4, i_5, i_6 ve i_7 değişken katsayılarını göstermektedir.

$$S1 = i_2S2 + i_3S3 + i_4S4 + i_5S5 + i_6S6 + i_7S7 \qquad (2.4)$$

Tablo 14: Regresyon Yöntemi için Örnek Veri Seti

Gözlemler	S1	S2	S3	S4	S5	S6	S7
1	0.38	0.25	0.45	0.56	?	0.74	0.87
2	0.96	0.17	0.24	?	0.63	?	0.04
3	0.41	0.54	?	0.85	0.21	0.34	0.054
4	?	0.31	0.51	0.68	0.93	0.35	0.75

Denklem (2.4) S2, S3, S4, S5, S6 ve S7 değişkeninin bilinen değerlerini kullanarak S1 değişkeninin eksik veri girişlerine yaklaşık değer olarak tahminleme yapar.

2.4.4.3. Çoklu Atama Yöntemi

Çoklu atamada (MI-Multiple Imputation) regresyonla atama işlemi birden fazla kez tekrarlanarak her bir veri kümesi daha sonra kullanılmak üzere saklanır. Elde edilen tüm verilerin hata miktarları hesaplanır. Ardından bu verilerin ortalamaları alınarak gerçek veri kümesi elde edilir. Buradaki tek ayrıntı standart hatanın hesaplanması sırasında iki veri kümesinin birleştirilmesi için iki standart hata miktarı toplanıp karekökleri alınır. Kısacası çoklu atama, daha iyi sonuç elde etmek için regresyonla atamanın birden fazla yapılması olarak düşünülebilir (Şeker ve Eşmekaya, 2017: 15).

Çoklu atama, eksik veri girişinin bir dizi M sayıda yaklaşık değerle değiştirildiği bir yaklaşımdır (Rubin, 1978: 23). MI, birbirini izleyen üç adımda tanımlanır. İlk adım, veri kümesindeki eksik veri girişlerini M farklı değerlerle değiştirmeyi içerir. Bu, M farklı veri kümelerinin tam kayıtlarla elde edilmesiyle sonuçlanır. İşlemin ikinci adımı, farklı M veri kümelerinin eksiksiz veri analiz tekniklerini uygulayarak tam kayıtlarla analiz edilmesini gerektirir. Son olarak, üçüncü adımda, M veri kümelerinden elde edilen sonuçlar, ikinci adımda yapılan analize dayanarak birleştirilir. İkinci adımda atıfta bulunulan bu sonuç, hangi M veri kümelerinin eksik veri girişlerinin en iyi durumunu aldığını ya da daha iyi sonuçlar ve çıkarımlar verdiğini gösterir. Bu yaklaşım, tekli atama yaklaşımlarından daha iyi bir yaklaşımdır. Ayrıca, işlenecek yeni veri matrislerini elde etmek için sıcak deste

atama yönteminin popüler özellikleri ile EM ve olasılık tahmini yaklaşımlarının avantajlarından yararlanmaktadır. Bu yöntem, verilerin MAR olduğu ve çok değişkenli tekdüze dağılımdan kaynaklandığı varsayımına bağlıdır (Schafer, 2000: 153; Allison 2002: 38).

2.5. MAKİNE ÖĞRENMESİ YAKLAŞIMLARIYLA EKSİK VERİ ÇÖZÜMLEME YÖNTEMLERİ

Makine öğrenimine dayanan atama yöntemleri, genellikle eksik verilerin yerine geçecek değerleri tahmin etmek için bir öngörücü model oluşturmayı içeren karmaşık prosedürlerdir. Bu yaklaşımlar, veri kümesindeki mevcut bilgilere dayanarak eksik veri tahminini modellemektedir. Bu tekniklerin bazıları ağaç (karar ağaçları, rasgele orman) temelli ya da biyolojik kavramlara (yapay sinir ağları) dayalıdır.

2.5.1. Karar Ağaçları

Karar ağaçları, verileri sınıflandırma ya da regresyon analizi için tutarlı kümelere ayırmayı amaçlayan denetimli öğrenme modelleridir. Karar ağacı varsayılan olarak döngüseldir ve bir kök düğümü, yaprak düğümleri, iç düğümler ve kenarlardan oluşur. Kök düğüm, ağacın başlangıcını, sonucu ya da sınıf etiketini sunan ağacın sonunu temsil eden yaprak düğümleriyle belirtir. Dahili düğüm, her bir düğümdeki verileri bölmek için kullanılan değişken hakkındaki ayrıntıları saklar. Kenarlar düğümler arasındaki bağlantılardır ve bölünmeler hakkında ayrıntılar içerir. Bir kaydın sonucu, ağaç boyunca kök düğümden yaprak düğüme kadar olan bilgilerin işlenmesiyle elde edilir (Twala ve Cartwright 2010: 299).

Eksik veri tahmini görevini gerçekleştirmek için karar ağaçlarının kullanılması, eksik veri girişlerine sahip her değişken için bir ağaç oluşturmayı gerektirir. Bu değişken, girdi değişken kümesinin bir parçasını oluşturan gerçek sınıf etiketine sahip sınıf etiketi olarak kabul edilir. Ağacın inşası bilinen sınıf etiketlerine sahip kayıtlar kullanılarak yapılır ve eksik veri girişlerne karşılık gelen bir ağaç ile değiştirilir (Twala ve Cartwright 2010: 300).

Örneğin, bir veri setinde I1, I2, I3 değişkenleri ve bu şekilde elde edilen bir sınıf etiketi L vardır: L (I1, I2, I3). I1'in eksik değerleri olduğunu varsayalım, I1 sınıf etiketi olarak kabul edilirken L, giriş değişkenlerinden biri olarak kabul edilir. Yeni sınıf etiketi şu şekilde elde edilir: I1 (L, I2, I3). I2'de eksik veriler varsa, yeni çıktı şu şekilde elde edilir: I2 (I1, L, I3) ve I2 yeni sınıf etiketi olarak kabul edilir. Bu prosedür, eksik verileri olan tüm özellik değişkenleri tamamlanana kadar yürütülür.

Eksik değerlerin ikame edildiği dizilerin tahminleri etkileyip etkilemediği bilinmemektedir ve yöntem, verilerin MCAR olduğu varsayıldığında en iyi sonucu verir (Marwala, 2019: 11).

2.5.2. Yapay Sinir Ağları

YSA'lar, verileri biyolojik sinir sisteminin yaptığı gibi işlemek için olasılıklı bir model kullanır (Abdella ve Marwala, 2005b: 598). İnsan beyni böyle bir sisteme çok iyi bir örnektir. Nöronlar birbirine bağlıdır ve bu bağlantıların doğası da ağın performansını etkiler. Bir sinir ağının temel işlem birimi, bir nöron olarak adlandırılan yapıdır (Ming-Hau 2010: 711).

Bir sinir ağı dört önemli bileşenden oluşur, bunlar:

a. Sinir ağının biyoritmindeki herhangi bir aşamada, her nöronun bir aktivasyon fonksiyonu değeri vardır;

b. Her bir nöron diğer bir nörona bağlanır ve bu bağlantılar, bir nöronun aktivasyon seviyesinin başka bir nöron için nasıl girdi haline geldiğini belirler. Bu bağlantıların her birine bir ağırlık değeri tahsis edilmiştir;

c. Bir nöronda, çıkış katmanındaki ya da sonraki gizli katmanlardaki nöronlar için yeni bir giriş oluşturmak üzere gelen tüm girişlere bir aktivasyon fonksiyonu uygulanır ve;

d. Giriş-çıkış eşleştirmesi verildiğinde nöronlar arasındaki ağırlıkları ayarlamak için bir öğrenme algoritması kullanılır (Haykin, 1999: 23).

Abdella ve Marwala (2005a: 577) YSA modeli kullanarak eksik verileri hesaplamıştır. YSA modeli veri setinin özel olarak modellemesinde yani giriş katman değeri ile çıkış katman değeri aynı olacak şekilde oluşturulmuştur.

2.6. DERİN ÜRETİCİ MODELLER İLE EKSİK VERİ ÇÖZÜMLEME

Hiyerarşik öğrenme ya da derin yapılandırılmış öğrenme olarak da adlandırılan derin öğrenme, geleneksel göreve özgü yöntemlerin aksine, veri temsillerini öğrenmeye dayanan geniş makine öğrenme teknikleri ailesinin bir parçasını oluşturur. Veri sunumlarının öğrenilmesi yarı-denetimli, denetimli ya da denetimsiz yaklaşımlar yoluyla olabilir. Derin öğrenme modelleri, biyolojik sinir sisteminde gözlemlenen ve beyinde çeşitli uyaranlar ve ilişkili nöronal yanıtlar arasındaki ilişkileri tanımlamaya ve açıklamaya çalışan nöral kodlama gibi bilgi işleme ve iletişim modellerini taklit eder. Derin öğrenme bir ya da daha fazla gizli katman içeren YSA'lar ve benzeri makine öğrenme algoritmalarını kapsayan çalışma alanıdır. Yani en az bir adet YSA'nın kullanıldığı ve bir çok algoritma ile, bilgisayarın eldeki verilerden yeni veriler elde etmesidir. Derin öğrenme gözetimli, yarı gözetimli ya da gözetimsiz olarak gerçekleştirilebilir (Bengio ve diğerleri, 2015: 436).

Derin yapay sinir ağları pekiştirmeli öğrenme yaklaşımıyla da başarılı sonuçlar vermiştir (Lavet ve diğerleri, 2018: 219).

Derin üretici modeller, genellikle gizli alanı temsil eden rastgele bir değişkenden veri noktalarının örneklenmesine izin veren tüm olası verilerin (örneğin, yüz görüntüleri) alanı üzerinde bir dağılımı tanımlamaktadır. Derin üretici modeller derin sinir ağları kullanarak veri dağılımını parametreleştirmekte ve görüntü üretmede etkileyici sonuçlar elde etmektedirler (Brock ve diğerleri, 2018: 2).

Eksik veri atamalarında iki ana tip derin üretici model kullanılmaktadır: varyasyonel oto-kodlayıcılar (VAE'ler) ve üretici çekişmeli ağlar (GAN'lar).

2.6.1. Varyasyonel Oto Kodlayıcılar

Varyasyonel otomatik kodlayıcı, etkin veri kodlamalarını denetimsiz bir şekilde öğrenmek için kullanılan bir tür yapay sinir ağıdır. Bir otomatik kodlayıcının amacı, ağı sinyal "gürültüsünü" göz ardı edecek şekilde eğiterek, tipik olarak boyutsallık azaltımı için bir veri kümesi için bir temsili öğrenmektir (Kingma ve Welling, 2014: 3).

VAE, eksik veri problemlerine uyarlamaya yönelik bir yaklaşım, modele keyfi şartlandırma getirmektedir. Gözlemlenen verilerin keyfi olarak bir alt kümede koşullandırıldığı varyasyonel oto-kodlayıcıya dayanan tek bir nöral olasılık modelini kullanmaktadır. Veri setindeki değerler hem sayısal hem de kategorik olabilmektedir. Modelin eğitimi stokastik varyasyonel Bayes tarafından gerçekleştirilmektedir. Sentetik veriler üzerindeki deneysel değerlendirme, özellik ataması ve görüntü boyama problemleri, önerilen yaklaşımın etkinliğini ve üretilen örneklerin çeşitliliğini göstermektedir (Ivanov ve diğerleri, 2019: 3).

2.6.2. Üretici Çekişmeli Ağlar

Makine öğrenmesi algoritmaları, mevcut verilerdeki kalıpları tanımada ve bu kavrama gücüyle sınıflandırma (bir örneği doğru kategoriyi atama) ve regresyon (çeşitli girdilere dayalı sayısal bir değer tahmin etme) görevlerinde çalışırlar. Bununla birlikte, bu algoritmalardan yeni veriler üretilmesi istendiğinde bilgisayarlar bu görevin üstesinden gelememekteler. Bir makine öğrenmesi algoritması büyük bir satranç ustasını yenebilir, hisse senedi fiyatı hareketlerini tahmin edebilir ve bir kredi kartı işleminin hileli olup olmadığını sınıflandırabilir. Buna karşılık, Amazon'un Alexa'sı ya da Apple'ın Siri'si gibi asistan görevlerinde kullanılan uygulamalar ile küçük bir konuşma yapma becerisi çok sınırlıdır. İnsanın en basit ve en temel yetkinlikleri, keyifli bir konuşma ya da orijinal bir ürün ortaya çıkarma, en karmaşık süper bilgisayarları bile dijital spazmlara maruz bırakabilir.

Bu durum üretici çekişmeli ağların (GAN-Generative Adversial Networks) icat edilmesiyle değişmiştir. GAN'lar, bilgisayarların bir değil iki ayrı sinir ağı kullanarak gerçekçi veriler üretmesini sağlayan bir tekniktir. GAN' lar veri oluşturmak için kullanılan ilk bilgisayar programları olmamasına karşın, sonuçları ve çok yönlülüğü bu tekniği diğer tekniklerden ayırmaktadır. GAN'lar eşzamanlı olarak eğitilmiş iki modelden oluşan bir makine öğrenmesi teknikleri sınıfıdır: Biri (üretici) sahte veriler üretmek için eğitilirken, diğeri (ayırıcı) sahte verileri gerçek örnek verilerinden ayırt etmek için eğitilmektedir (Goodfellow ve diğerleri, 2014: 2672).

GAN'lar, yapay sistemler için neredeyse imkânsız olduğu düşünülen, gerçek dünyaya benzer kalitede sahte görüntüler üretme, sıradan bir karalamayı fotoğraf

benzeri bir görüntüye dönüştürme ya da bir atın video görüntüsünü koşan bir zebra video görüntüsüne dönüştürmede çok iyi sonuçlar elde etmiş ve herhangi bir etiketlenmiş eğitim verisine ihtiyaç duymadan bunun sağlanmasına olanak tanımaktadır (Langr ve Bok, 2019: 21).

Üretici (Generative) kelimesi, modelin genel amacını belirtir; yeni bir veri oluşturmak. Bir GAN'ın üretmeyi öğreneceği veriler, modele yüklenen veri setinin seçimine bağlıdır. Bu veri seti, müzik, resim ya da video görüntülerinden oluşan veri setleri olabilir.

Çekişmeli (Adversial) kelimesi, GAN çerçevesini oluşturan iki model arasındaki oyun benzeri, rekabetçi dinamiğe işaret eder; üretici ve ayırıcı. Üreticinin amacı, eğitim setindeki gerçek verilerden ayırt edilemeyen örnekler oluşturmaktır. Ayırıcının amacı, üretici tarafından üretilen sahte örnekleri eğitim veri setinden gelen gerçek örneklerden ayırmaktır. Bu iki ağ sürekli olarak birbirini geçmeye çalışır: Üretici ikna edici veriler yaratmada ne kadar iyi olursa, ayırıcının gerçek örnekleri sahte olanlardan ayırt etmekte o kadar iyi olması gerekir.

Son olarak Ağ (network) kelimesi üretici ve ayırıcıyı temsil etmek için yaygın olarak kullanılan makine öğrenme modelleri sınıfını gösterir: Sinir ağları. GAN uygulamasının karmaşıklığına bağlı olarak, bunlar basit ileri beslemeli sinir ağlarından, evrişimli sinir ağlarına ya da daha karmaşık varyasyonlarına kadar değişebilir.

Özetle üreticinin amacı, modeli eğitmek için kullanılan veri setinin özelliklerini yakalayan örnekler üretmektir, öyle ki ürettiği örnekler eğitim verilerinden ayırt edilemez görünecektir. Ayırıcı, üreticinin tersine bir nesne tanıma modeli olarak tanımlanır. Nesne tanıma algoritmaları, bir görüntünün içeriğini ayırt etmek için görüntülerdeki desenleri öğrenebilirler. Üretici, kalıpları tanımak yerine, temelde sıfırdan yaratmayı öğrenir ve giriş verileri rastgele sayılardan oluşan bir vektör ile başlatılır.

Tablo 15, GAN mimarisini oluşturan iki alt ağı hakkında temel çıkarımları özetlemektedir.

Tablo 15: Üretici Ve Ayırıcı Ağın Bileşenleri

	ÜRETİCİ	AYIRICI
GİRDİ	Rasgele sayılardan oluşan vektör	Ayrıcı iki farklı kaynaktan girdi alır: -Eğitim veri setinden alınan gerçek örnekler -Üreticiden gelen sahte veriler
ÇIKTI	Mümkün olduğunca ikna edici olmaya çalışan sahte örnekler	Girdi örneğinin gerçek olma olasılığı
HEDEF	Eğitim veri kümesinin örneklerinden ayırt edilemeyen sahte veriler oluşturmak	Üreticiden gelen sahte örnekleri ve eğitim veri setinden gelen gerçek örnekleri ayırt etmek

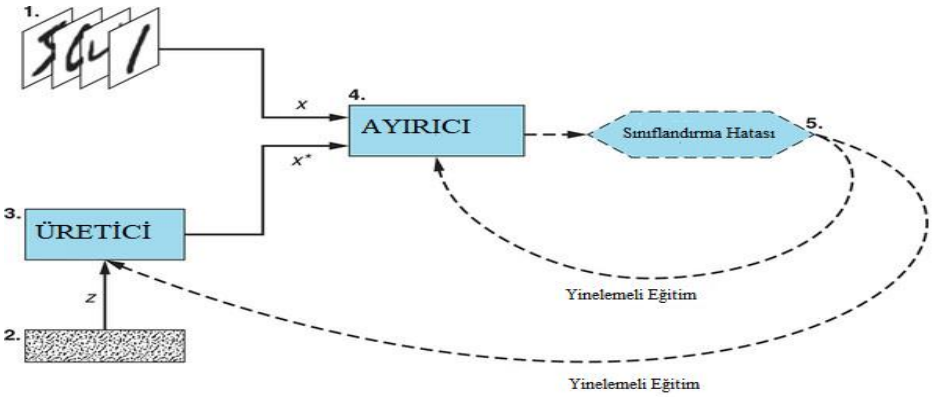
Kaynak: Langr ve Bok, 2019: 23

Tablo 15' te GAN mimarisini oluşturan bileşenlerin çalışma şekli: Üretici, ayırıcının sınıflandırmalarından aldığı geri bildirimler yoluyla öğrenir. Ayırıcı, belirli bir örneğin gerçek (eğitim veri kümesinden gelen) ya da sahte (üretici tarafından oluşturulan) olup olmadığını belirler. Buna göre, ayırıcı sahte bir görüntüyü gerçek olarak sınıflandırdığında, üretici, ayırıcının hangi verileri gerçek olarak sınıflandırdığını öğrenir. Tersine, ayırıcı, üretici tarafından üretilen bir görüntüyü sahte olarak her reddettiğinde, üretici iyileştirmesi gereken geri bildirimi alır. Eş zamanlı olarak ayırıcı gelişmeye devam eder. Herhangi bir sınıflandırıcı gibi, tahminlerinin gerçek etiketlerden (gerçek ya da sahte) ne kadar uzak olduğunu öğrenir. Bu nedenle, üretici gerçekçi görünümlü veriler üretme konusunda daha iyi hale geldikçe, ayırıcı sahte verileri gerçeklerden ayırt etmede daha iyi olur ve her iki ağ da aynı anda gelişmeye devam eder.

GAN'lar ve bileşenlerinin çalışma şeklinden bahsedildikten sonra GAN sisteminin çalışma prensibinin daha iyi anlaşılabilmesi için el yazısı ile yazılmış rakamlardan oluşan görüntü veri seti kullanılarak gerçekçi görüntüler elde edilmesi amacında bir sistem örneği ve sistemin çalışma prensibini gösteren görsel Şekil 1 de gösterilmiştir.

Şekil 1 GAN mimarisinin çekirdeğini göstermektedir.

Şekil 1: GAN mimarisinin Çekirdeği



Kaynak: Langr ve Bok, 2019: 24

Şekil 1 deki diyagram özetle;

a. **Eğitim Veri Kümesi (Training Data Set)**: Üretici ağın mükemmele yakın kalitede taklit etmeyi öğrenmesini istediğimiz gerçek örneklerin veri kümesi. Şekil 1 deki GAN mimarisini gösteren diyagramda, veri kümesi el yazısı rakamların görüntülerinden oluşturulmuştur. Bu veri kümesi, ayırıcı ağına giriş **(x)** işlevi görür.
b. **Rasgele Gürültü Vektörü (z):** Bu girdi, üretici ağın sahte örnekleri sentezlemek için bir başlangıç noktası olarak kullandığı rastgele sayıların bir vektörüdür.
c. **Üretici Ağ**: Üretici ağ giriş olarak rastgele sayıların **(z)** bir vektörünü alır ve sahte örnekler **(x *)** verir. Böylece, ürettiği sahte örnekleri eğitim veri setindeki gerçek örneklerden ayırt edilemez hale getirmeye çalışır.
d. **Ayırıcı Ağ:** Ayırıcı, girdi olarak ya eğitim setinden gelen gerçek bir örneği **(x)** ya da üretici tarafından üretilen sahte bir örneği **(x *)** alır. Her örnek için, ayırıcı, örneğin gerçek olup olmadığını belirler ve çıktısını verir.
e. **Yinelemeli eğitim / ayarlama:** Ayırıcının her bir tahmini için - normal bir sınıflandırıcıda yapıldığı gibi - ne kadar iyi olduğu belirlenir. Sonuçlar

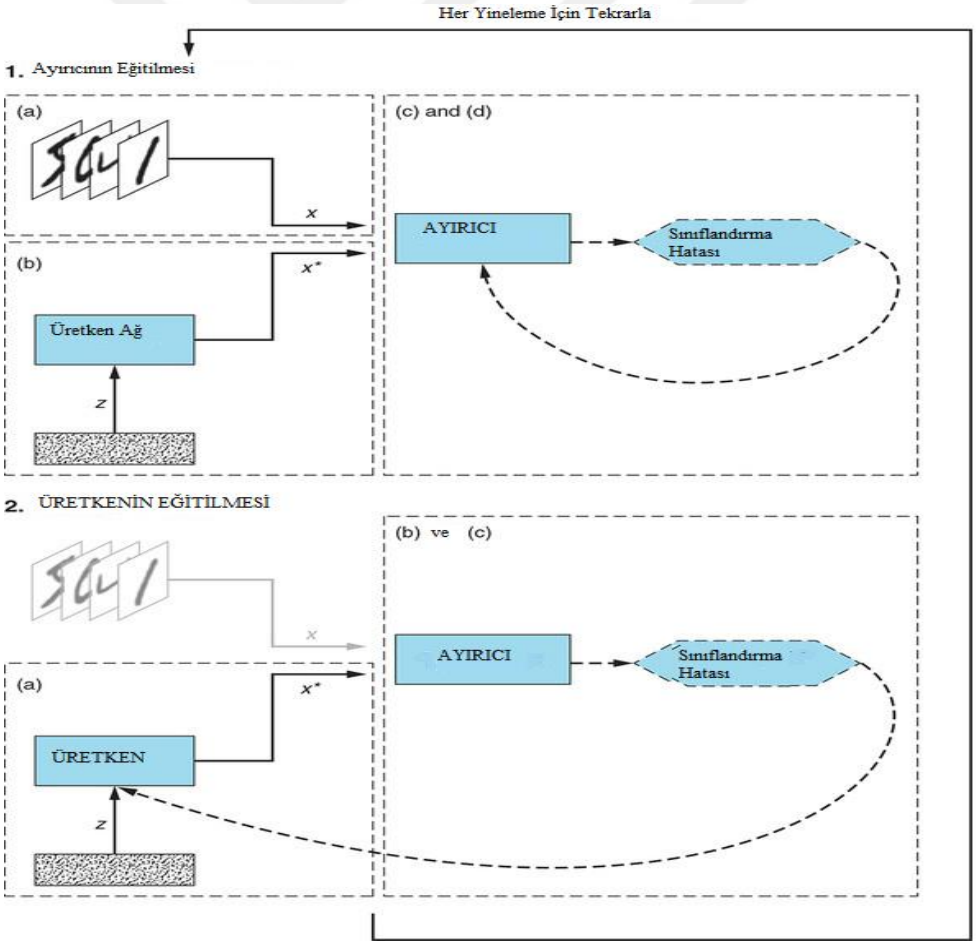
ayırıcı ve üretici ağlarını geri yayılım yoluyla tekrar tekrar ayarlamak için kullanılır:

f. Ayırıcının ağırlıkları ve sapma değerleri (bias), sınıflandırma doğruluğunu en üst düzeye çıkarmak için güncellenir (doğru tahmin olasılığını en üst düzeye çıkarır: **x: gerçek** ve **x *: sahte olarak**).

g. Üretici ağın ağırlıkları ve sapma değerleri (bias), ayırıcının **x *** 'ı (x tahmin) gerçek olarak yanlış sınıflandırma olasılığını en üst düzeye çıkarmak için güncellenir.

GAN bileşenlerini ve mimari şemayı çalışırken gösteren diyagram Şekil 2 de gösterilmiştir.

Şekil 2: GAN Mimarisi Ve Çalışma Şeklini Gösteren Diyagram



Kaynak: Langr ve Bok, 2019: 25

Şekil 2'de GAN eğitim algoritmasının iki ana bölümü gösterilmiştir. Bu iki bölüm, ayırıcı ağın eğitimi ve üretici ağın eğitimi, aynı GAN mimarisinin eğitim sürecinin ilgili aşamalarında farklı zamanda anlık görüntülerini gösterir.

Şekil 2 'deki sürecin işleyişi;

a. Ayırıcı Ağın Eğitilmesi:

- Eğitim veri kümesinden rastgele bir gerçek örnek **x** al.

- Yeni bir rastgele gürültü vektörü **z** edin ve üretici ağını kullanarak sahte bir örnek **x *** sentezle

- **x** ve **x *** 'ı sınıflandırmak için ayırıcı ağını kullan

- Sınıflandırma hatalarını hesapla ve sınıflandırma hatalarını en aza indirmek için ayırıcı ağın ağırlıklarını ve sapmalarını güncellemek için toplam hatayı geri yay

b. Üretici Ağı Eğit:

- Yeni bir rastgele gürültü vektörü **z** edin ve üretici ağını kullanarak sahte bir örnek **x *** sentezle.

- **x*** 'ı sınıflandırmak için ayırıcı ağını kullan.

- Sınıflandırma hatasını hesapla ve ayırıcı ağın hatasını en üst düzeye çıkarmak için üretici ağın ağırlıklarını ve sapmalarını güncellemek için hatayı geri yay.

GAN eğitim döngüsünün ne zaman durması gerektiği yani uygun sayıda eğitim yinelemesinin belirleyebilmek için Oyun Teorisi (Game Teory) isimli bir sıfır-toplamlı-oyun(zero-sum-game) senaryosu kullanılmaktadır. Birbiriyle rekabet halinde olan iki sinir ağı bu oyun senaryosu teorisi temelinde eğitilmektedir. Bir oyuncunun kazancı diğer oyuncunun kaybı anlamına gelen teorik oyun senaryosu, oyuncuların birbirlerinin kazancını etkilemeye çalıştığı ve Nash Dengesine (Nash equilibrium) ulaştıklarında eğitim yinelemesinin duracağı teorisi ile eğitilmektedirler (Langr ve Bok, 2019: 33).

GAN, aşağıdaki koşullar gerçekleştiğinde Nash dengesine ulaşır:

- Üretici ağ, eğitim veri setindeki gerçek verilerden ayırt edilemeyen sahte örnekler ürettiğinde,

- Ayırıcı ağ, belirli bir örneğin gerçek ya da sahte olup olmadığını rastgele olarak tahmin etmeye başladığında, (yani bir örneğin gerçek olup olmadığını 50/50 olasıkla tahmin etmeye başladığında).

Denge sağlandığında, GAN'ın yakınsadığı varsayılmaktadır. Ancak, uygulamada, öngörülemeyen karmaşıklıklar nedeniyle GAN'lar için Nash dengesini bulmak neredeyse imkânsızdır (Langr ve Bok, 2019: 34).

2.6.1.1. Kayıp Fonksiyonları

GAN'lar temelde kullanıcının seçmiş olduğu veri setindeki olasılık dağılımını kopyalamaya çalışmaktadır. Bu nedenle GAN tarafından üretilen verilerin dağılımı ile gerçek verilerin dağılımı arasındaki mesafeyi yansıtan kayıp fonksiyonları kullanılmaktadır (Goodfellow ve diğerleri, 2014: 2676).

Kayıp fonksiyonu, bir veya daha fazla değişkenin bir olayını veya değerlerini, sezgisel olarak olayla ilişkili bir "maliyeti-kaybı" temsil eden gerçek bir sayıya eşleyen bir fonksiyondur.

Bir GAN'ın iki kayıp fonksiyonu olabilir: biri üretici ağın eğitimi için ve diğeri ayırıcı ağın eğitimi için. Olasılık dağılımları arasındaki mesafe ölçüsünü yansıtmak için iki kayıp fonksiyonu birlikte çalışmaktadır (Goodfellow ve diğerleri, 2014: 2676). GAN mimarisi iki kayıp fonksiyonu ile tanıtılmıştır: İlki minimax GAN kaybı olarak bilinen ve ikincisi doygunluk vermeyen GAN kaybıdır. Bu araştırma konusu kapsamında yaygın olarak kullanılan iki GAN kayıp fonksiyonu kullanılmıştır:

a) **Mini-max Kayıp Fonksiyonu:** Mini-max GAN kaybı, ayırıcı ve üretici ağların minimax eşzamanlı optimizasyonunu ifade etmektedir. GAN'ları tanıtan makalede ilk önerilen kayıp fonksiyonu olma özelliğini taşımaktadır (Goodfellow ve diğerleri, 2014: 2677).

Minimax, iki oyunculu sıra tabanlı oyunlarda, diğer oyuncunun en kötü durumunun kaybını ya da maliyetini en aza indirmek için bir optimizasyon stratejisini ifade etmektedir (Maschler ve diğerleri, 2013: 176).

GAN için, üretici ve ayırıcı iki oyuncudur ve model ağırlıklarındaki güncellemeleri sırayla almalarını ifade etmektedir. Min ve maks, üretici kaybının minimize edilmesine ve ayırıcı kaybının maksimize edilmesine işaret etmektedir (Goodfellow ve diğerleri, 2014: 2679). GAN'larda kullanılan Minimax stratejisi denklem (2.5)'te gösterilmiştir.

$$\min \max (D, G) = E_x\left[\log\left(D(x)\right)\right] + E_z\left[\log\left(1 - D\left(G(z)\right)\right)\right] \qquad (2.5)$$

Denklem (2.5)'te belirtildiği gibi, ayırıcı gerçek görüntülerin logaritma olasılığının ortalamasını ve sahte görüntüler için ters olasılığın logaritmasını en üst düzeye çıkarmayı amaçlamaktadır. Bu denklemde,

- **D(x),** ayırıcının gerçek veri örneğinden gelen **x'in** gerçek olma olasılığına ilişkin tahminidir.
- $\boldsymbol{E_x}$, tüm gerçek veri örnekleri üzerinde beklenen değerdir.
- **G(z),** gürültü vektörü **z** verildiğinde üretici ağın çıktısıdır.
- **D(G(z))**, Ayırıcı'nın sahte bir örneğin gerçek olma olasılığına ilişkin tahminidir.
- $\boldsymbol{E_z}$, üretici ağa yapılan tüm rastgele girişler üzerinde beklenen değerdir (aslında, üretilen tüm sahte örnekler **G (z)** üzerinde beklenen değerdir).
- Formül, gerçek ve üretilen dağılımlar arasındaki çapraz entropiden türetilir.
- Üretici, fonksiyondaki **log(D(x))** terimini doğrudan etkileyemez, bu nedenle ayırıcı için, kaybı minimize etmek **log(1-D(G(z)))** minimize edilmesine eşdeğerdir (Langr ve Bok, 2019: 43).

Pratikte min-max kayıp fonksiyonunun kullanıldığı GAN mimarilerinde ayırıcı ağ için yavaş yakınsama problemi ortaya çıkmaktadır. Yavaş yakınsama problemi bilgisayarların hesaplama gücüne rağmen saatlerce ve bazen de günlerce sürebilen bir eğitim döngüsüne sebep olmaktadır. Bu durumun önüne geçebilmek için geliştirilmiş maliyet fonksiyonu önerilmiştir. Min-max maliyet fonksiyonu geliştirici tarafından önerilen ilk fonksiyon olmakla birlikte Nash dengesinin anlaşılmasını ve konu bütünlüğüne yardımcı olması bakımından önemlidir.

b) Doyumsuz Minimax Kaybı (Non-Saturating GAN Loss): denklem (2.5)'te ifade edilen minimaks kayıp fonksiyonunun, ayırıcının işi çok

kolay olduğunda GAN eğitiminin ilk aşamalarında modelin (saturating-model doyumu) sıkışmasına neden olabileceği belirtilmektedir (Goodfellow, 2016: 22).

Bu nedenle, üreticinin **logD(G(z))** fonksiyonunu maksimize etmeye çalışacak şekilde üreticinin kaybını değiştirmeyi önermektedir (Goodfellow, 2016: 22).

Üretici'nin kaybının değişmesi ve yeni denklemin gösterimi denklem (2.6)'da gösterilmiştir;

Ayırıcı: max log D(x) + log(1 − D(G(z))) (2.6)

c) **Wasserstein Kayıp Fonksiyonu:** Bu kayıp fonksiyonu, ayırıcının örnekleri gerçekten sınıflandırmadığı GAN şemasının değiştirilmesiyle elde edilebilmektedir. Bu yöntemde, her örnek için çıktı olarak bir sayı verilmektedir. Bu sayının 1'den az ya da 0' dan daha büyük olması gerekmez, bu nedenle bir örneğin gerçek ya da sahte olup olmadığına karar vermek için 0,5 eşik olarak kullanılmamaktadır. Ayrıcı'nın eğitimi, sahte örneklerin çıktılarını büyütmek yerine gerçek örneklerin çıktılarını (sınıflandırmalarını) büyütmeye çalışmaktadır (Arjovsky ve diğerleri, 2017: 21).

WGAN mimarisinde kayıp fonksiyonları denklem (2.7) ve denklem (2.8)'de gösterildiği gibi güncellenmiştir:

$$Eleştirmen = L_D^{WGAN} = E[D(x)] - E\left[D\left(G(z)\right)\right] \quad (2.7)$$

$$Üretici = L_G^{WGAN} = E[D(G(z))] \quad (2.8)$$

Denklem (2.7)'de Eleştirmen olarak isimlendirilen yeni bileşen GAN mimarisindeki ayırıcının yerine geçmektedir. Üretici denklem (2.8)'deki fonksiyonu en büyütmeye çalışır. Başka bir deyişle, sahte örnekler için ayırıcının çıktı değerini en büyütmeye çalışır.

Bu fonksiyonda;

- **D(x),** eleştirmenin gerçek bir örnek için çıktısıdır.
- **G(z),** gürültü **z** verildiğinde üreticinin çıktısıdır.

- D(G(z)), Eleştirmenin sahte bir örnek için çıktısıdır.

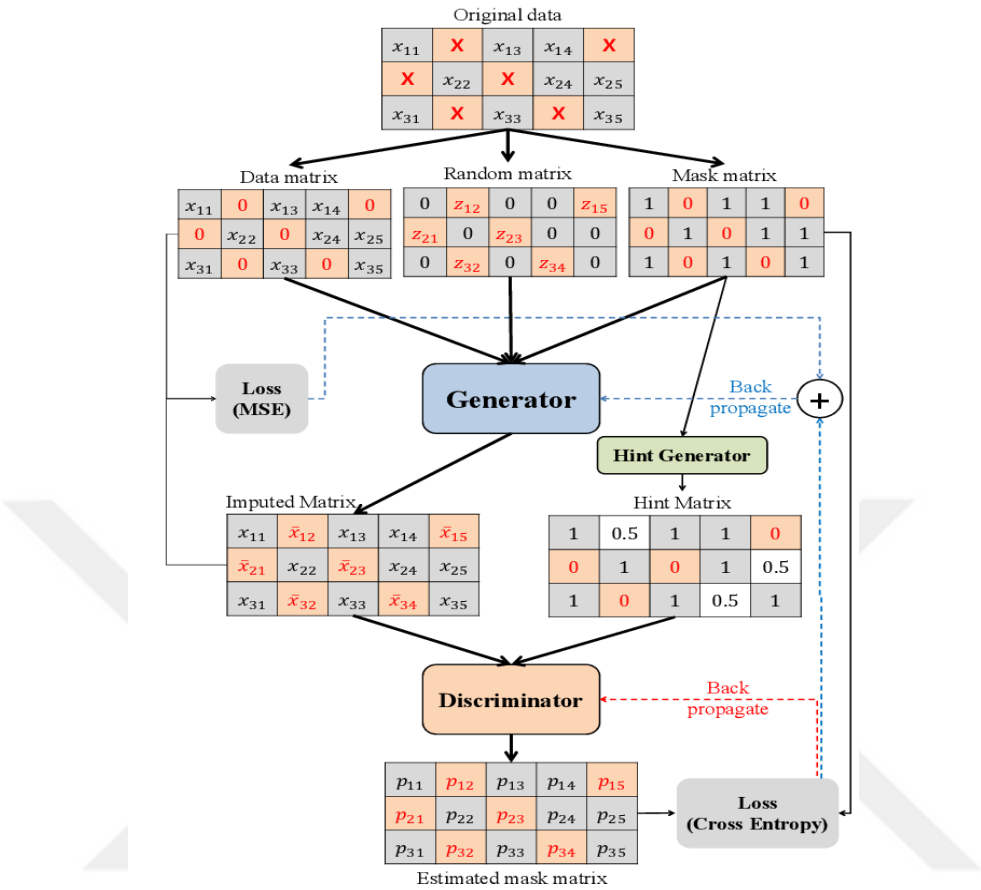
Wasserstein GAN'lar minimax tabanlı GAN'larda sıklıkla görülen modelin sıkışması problemlerine karşı daha az hassastır ve kaybolan gradyanlar ile ilgili sorunları önlemektedir (Arjovsky ve diğerleri, 2017: 21).

2.6.3. Üretici Çekişmeli Atama Ağları

GAIN, GAN mimarisini kullanarak eksik veri ataması yapılması sağlayan bir modeldir. GAIN modeli, bir veri matrisindeki eksik gözlemlerin yerine atama işlemini, eksik olmayan verilerden oluşan veri matrisinin şartlı olasılık dağılımına göre yapabilmektedir (Yoon ve diğerleri, 2018: 5689).

GAIN modeli, GAN mimarisindeki üretici ve çekişmeli ağları kullanır. Ancak üretici ve ayırıcı ağlarının veri üretme ve üretilen veriyi ayırt etme görevleri eksik veri atama ve atanan gözlemlerin ayırt edilmesi görevini yerine getirmek için değiştirilmiştir. GAIN mimarisi Şekil 3 te gösterilmiştir. Şekil 3 GAIN algoritmasının genel mimarisini, farklı matrislerin nasıl bağlandığını ve hangilerinin üretici ve ayırıcı üzerinde etkisi olduğunu gösterir.

Şekil 3: GAIN Algoritmasının Genel Mimarisi



Kaynak: Yoon ve diğerleri, 2018: 5692

Şekil 3 te GAIN mimarisini oluşturan bileşenler:

a. Original Data (Orjinal Veri Matrisi): Eksik gözlemlere sahip orijinal veri matrisini,

b. Data Matrix (Veri Matrisi): Veri matrisindeki gözlenen değerleri ve eksik değerlerin konumunun ” 0”' olarak belirtildiği eksik veri matrisini,

c. Mask Matrix (İşaretçi Matris): Original Data matrisindeki eksik olmayan değerlerin “1”, eksik olan değerlerin “0” ile belirtildiği işaretçi matrisini (indicator matrix),

d. Random Matrix (Rassal Matris): Mask matrix’teki eksik olmayan gözlemlerin yerinin “0” , eksik olan gözlemlerin yerine ise rasgele sayıların atandığı matrisi,

e. Genarator: Üretici ağını,

f. Loss (MSE-Kayıp Fonksiyonu): atama yapılan ve gözlenmiş değerler arasındaki sapmaları hesaplayan fonksiyonu,

g. Imputed Matrix (Atanmış Matris): Eksik gözlemlerin yerine atama yapılmış veri matrisini,

h. Hint Matrix (İpucu Matrisini): Ayırıcı ağın, hangi gözleminin atanmış, hangi gözleminin orijinal veri matrisine ait olduğunu ve karar vermesi gereken (0.5 ile kodlanmış gözlemler) gözlem değerlere sahip veri matrisini,

i. Discriminator: Ayırıcı ağını,

j. Estimated Mask Matrix (Tahminlenmiş İşaretçi Matrisi): Ayırıcı ağa gelen veri matrisindeki gözlenmiş değerlerin (gerçek veri matrisine ait ve işaretçi matriste 1 ve 0 olarak kodlanmış matrise göre) olasılıksal sonuç matrisini,

k. Loss (Cross Entropy-Kayıp Fonksiyonu): İşaretçi matris ve tahminlenmiş işaretçi matris arasındaki farkı hesaplayan kayıp fonksiyonu,

l. Backpropagate (Geriye yayılım): kayıp fonksiyonları kullanılarak elde edilen ve model parametrelerini güncelleyen geriye yayılım işlemini gösterir.

GAIN mimarisinin çalışma prensibi şu şekildedir;

GAN mimarisinde, üretici ağ rasgele değerlerden oluşan **z** vektörünü orijinal veri matrisindeki değerlere benzetmeye çalışırken, GAIN mimarisinde, üretici ağ yine rasgele değerlerden oluşan z vektörünü, ayrıca veri matrisini, rassal matrisi, ve işaretçi matrisi girdi bileşenleri olarak alır. Üretici ağ, bu bileşenleri GAN mimarisindeki parametrelendirilmiş fonksiyon aracılığıyla orijinal veri matrisine benzetmek yerine, eksik olmayan gözlemlerin oluşturduğu veri dağılımına benzeyen dağılımın elde edilmesini sağlayan atamalar yaparak ayırıcı ağın değerlendirmesi için eksik gözlemler yerine atama yapılmış veri matrisini oluşturur.

GAIN'de görünen sorunlar GAN kullanırken ortaya çıkan sorunlara benzer. GAIN' in eğitimi de yavaş ve kararsız olabilir. GAIN'de kayıp (loss) değerlerin aynı kalmasını sağlamak özellikle Ayırıcı'nın kaybının aynı kalmasını sağlamak zor olabilir. Bunun dışında GAIN, kaybolan gradyanlardan ve mimarinin(mod collapse) çökmesinden GAN gibi etkilenir. GAIN mimarisinin çökmesi farklı bir şekilde görünür. GAN'da mimarinin çökmesinin sonucu, Üretici her zaman aynı çıktıyı çok küçük ya da hiçbir değişiklik olmadan verir. GAIN mimarisinin çökmesi, üreticinin farklı gürültü vektör değerleri ve aynı giriş değerlerine rağmen eksik gözlemlerin

yerine aynı değeri atadığı şeklinde ortaya çıkar. GAN'daki sorunlar WGAN kullanılarak çözülebildiği için, GAIN algoritması Wasserstein ayarı ile birlikte kullanıldığındadaha iyi performans göstereceği, işletmelerin sahip olduğu veri setlerinin kalitesini etkileyen eksik veri problemlerinin çözümünde kullanımı ve diğer çözüm yöntemleriyle kıyaslanması araştırma konusunu oluşturmaktadır. Bu hibrit kullanım ise WGAIN olarak isimlendirilecektir.

ÜÇÜNCÜ BÖLÜM
ŞARAP ÜRETİMİNDEVERİ KALİTESİNİ ETKİLEYEN EKSİK VERİ PROBLEMLERİNİN ÇÖZÜMLENMESİ ÜZERİNE BİR UYGULAMA

Bu bölümde yapılan uygulamanın amacı ve kapsamından, uygulamada kullanılan veri setinden ve bu veri setinde eksiklik mekanizmalarının nasıl oluşturulduğundan, eksik veri yapısından ve eksiklik mekanizmalarının rassallığının test edilmesinden, eksik veri çözümleme yaklaşımlarının kıyaslanmasını sağlayan değerlendirme yönteminden, kullanılan yazılım ve ortamlardan son olarak da analiz ve bulgulardan bahsedilecektir.

3.1. ŞARAP, ŞARAP ÜRETİMİ VE ŞARAP KALİTESİNİ ETKİLEYEN DEĞİŞKENLER

Şarap, genellikle şeker, asit, enzim, su veya diğer besinler eklenmeden fermente edilmiş asma üzümlerinden yapılan alkollü bir içecektir. Farklı şarap çeşitleri arasında kırmızı şarap ve beyaz şarap dünya çapında popüler olan şarap çeşitlerindendir. Kırmızı şarap üretim süreci, üzüm kabuğundan renk ve lezzet bileşenlerinin çıkarılmasını içermektedir. Kırmızı şarap, koyu renkli üzüm çeşitlerinden yapılmaktadır. Şarabın gerçek rengi yeni üretilmiş şaraplarda menekçe renginde, birkaç yıllık kırmızı şaraplarda kırmızı ve yıllandırılmış şaraplarda kahverengi olabilmektedir. Beyaz şarap saman sarısı, sarı-yeşil ya da sarı-altın renklerinde olabilmektedir. Renksiz üzüm özünün fermantasyonu ile beyaz şarap üretilmektedir. Beyaz şarabın üretildiği üzümler tipik olarak yeşil veya sarı renklerinde olabilmektedir (Canbaş, 2006: 48).

Şarap üretim kapasitesinin artırılması için endüstriler hem şarap yapımı hem de satış süreci için yeni teknolojilere yatırım yapmaktadır. Kalite değerlendirmesi bu bağlamda önemli bir unsurdur. Örneğin, şarap üretimini, en etkili faktörleri belirleyerek iyileştirmek ve premium markalar oluşturarak şarap markaları arasında fiyat farklarının belirlenmesi kalite değerlendirmeleri sonucunda elde edilmektedir. Kalite değerlendirmesi fizikokimyasal değişkenlerin testleri (sabit asit, uçucu asitliği, sitrik asit, artık şeker, klorürler, serbest sülfür dioksit, toplam sülfür dioksit,

yoğunluk, pH, sülfatlar, alkol) ya da duyusal testlerle (renk, tat, koku) yapılmaktadır. Duyu testi şarap tadım uzmanlarının değerlendirmelerine dayanırken, şarabın yoğunluk, alkol oranı ya da pH değerlerinin belirlenmesinde fizikokimyasal laboratuvar testleri kullanılmaktadır (Cortez ve diğerleri, 2009: 549).

Şarabın kalitesi ve özellikleri, üzüm çeşidine, şaraba işlenen üzümlerin bileşimine, şarap yapımında uygulanan işlemlere ve fermantasyon sonrası dinlendirme ve olgunlaştırma koşullarına bağlıdır. Ayrıca şarabın kalitesi fizikokimyasal değişkenlerin oranına (sabit asit, uçucu asitliği, sitrik asit, artık şeker, klorürler, serbest sülfür dioksit, toplam sülfür dioksit, yoğunluk, pH, sülfatlar, alkol) ve duyusal değişkenlere (renk, tat, koku) bağlıdır (Jackson, 2008: 16).

Araştırma konusu kapsamında kullanılan veri seti, şarap kalitesini sınıflandırmak için toplanmış verilerden oluşmaktadır. Veri setindeki fizikokimyasal değişkenlerin şarabın kalitesi ileilişkisi şu şekildedir:

a) Sabit asitlik: Üretilen şaraplar içinde bulunan sabit ya da uçucu olmayan asit miktarını göstermektedir.

b) Uçucu asitlik: Şaraptaki asetik asit miktarını ifade etmektedir. Şaraptaki uçucu asitlik oranının yüksek olması hoş olmayan, sirke tadına neden olabilmektedir.

c) Sitrik asit: Şaraplara tazelik ve lezzet katan sitrik asit miktarını ifade etmektedir.

Yüksek asitli bir şarap damak tadında genellikle daha canlı ve daha fazla keskin bir tat bırakmaktadır. Düşük asitli şarap ise damakta daha pürüzsüz ve daha yuvarlak tat bırakmaktadır. Asitlik, şaraplarda uzun süreli yıllandırma için gerekli olan bazı temel yapıların oluşmasını sağlamaktadır, bu nedenle yüksek asitli şarapların zamanla daha az miktarda olanlara göre yıllandıkça tadının iyileşme olasılığı olduğu belirtilmektedir (Nierman, 2004: 1).

d) Artık şeker: Fermentasyon durduktan sonra kalan şeker miktarını ifade etmektedir. 1 litre içinde 1 gramdan az artık şeker bulunması üretilen şarabın nadir bulunan lezzete sahip olduğunu gösterir. Üretilen şarap için dikkate alınan değer litre başına kaç gram şeker içerdiğidir.

e) Klorür: Şaraptaki tuz miktarını ifade etmektedir. Şarabın tadını etkileyen değişkenlerden biridir.

f) Serbest kükürt dioksit: Serbest kükürt dioksit formu, moleküler kükürt dioksit (çözünmüş gaz olarak) ve bisülfit iyonu arasındaki dengede bulunur; şaraptaki serbest kükürt dioksit miktarı mikrobiyal büyümeyi ve şarabın oksidasyonunu (şarabın oksijen ile çok fazla karışması) önlemektedir. Doğru kükürt miktarının katılmaması şarapta renk koyulaşması ve kahverengileşmeye sebep olur. Kokusal olarak çürük elma, sirke, aseton kokuları hissedilir.

g) Toplam kükürt dioksit: Şarapta litre başına 50 miligramın üzerinde olduğunda burunda ve şarabın tadında istenmeyen etkiyi bırakır. Değişken istenmeyen etkinin önüne geçmek için gerekli sınır değerlerinin takip edilmesini sağlayan değerleri içermektedir.

h) Yoğunluk: Şaraptaki alkol ve şeker miktarına bağlı olarak şarabın içinde bulunan suyun yoğunluğunu ifade etmektedir.

i) pH: Bir şarabın asidik ya da bazik ölçeğindeki değerini ifade etmektedir. 0 çok asidik 14 çok bazik ölçeği temel alınarak asidik-bazik ölçeği arasında aldığı değerleri ifade etmektedir. pH değerinin belirlenmiş sınırlan üzerinde olması şarabın yavaş yıllanmasına ve şarap içinde mikroorganizmaların çoğalmasına sebep olmaktadır.

j) Sülfat: Antimikrobiyal ve antioksidan görevi gören kükürt dioksit gazı seviyelerine katkıda bulunabilecek sülfat değerlerini ifade etmektedir. Şarabın sirkeye dönüşmesini engellemektedir.

k) Alkol: Şaraptaki alkol miktarını ifade etmektedir.

Şarap kalitesinin sınıflandırılması, şarap üreticilerinin gelirini, pazarlama stratejisini ve karar verme sürecini potansiyel olarak iyileştirmeleri için yararlı olabilmektedir.

3.2. UYGULAMANIN AMACI VE KAPSAMI

Bu alt başlıkta; yapılan uygulamanın amacı ve kapsamı hakkında bilgiler verilecektir.

3.2.1. Uygulamanın Amacı Ve Önemi

Yapılan uygulamanın amacı; şarap üretimi gerçekleştiren bir işletmede kaliteli şarap üretmek için gerekli olan değişkenlerin eksik değerlere sahip olması durumunda bir derin öğrenme algoritması olan GAIN algoritması, maliyet fonksiyonu değiştirilerek oluşturulan WGAIN algoritması ile eksik değerlerin tamamlanması ve bu algoritmaların performansının ikinci bölümde bahsedilen diğer eksik veri çözümleme yöntemlerinden EM, MI ve Karar Ağaçları algoritmalarıyla kıyaslayarak işletmenin sahip olduğu veri setinin veri kalitesini hangi yöntem seçilirse iyileştirilebileceğini göstermeye çalışmaktır.

Çalışmanın önemi ise söz konusu işletmelerin şarap kalitesini artırmada kullanabileceği kaliteli veri setinin elde edilmesinde ilk defa GAN tabanlı algoritmaların kullanılması ve şarap üreten işletmelere eksik veri problemlerini çözmelerine yönelik olarak bir rehber sunulacak olmasıdır.

3.2.2. Uygulamanın Kapsamı

Yapılan uygulamanın kapsamı; Portekiz "Vinho Verde" şarabının kırmızı çeşidiyle ile ilgilidir. Bu veri seti, sınıflandırma ya da regresyon görevleri için kullanılmaktadır. Gizlilik ve lojistik konulardan dolayı, sadece fizikokimyasal (girişler) ve duyusal (çıktı) değişkenler mevcuttur (örneğin üzüm türleri, şarap markası, şarap satış fiyatı vb. hakkında veri yoktur (Cortez ve diğerleri 2009: 547). Ayrıca veri setinde yapay olarak oluşturulan eksiklik mekanizma türleri MCAR ve MAR türünde eksiklik mekanizmalarıdır. MNAR türünde eksiklik mekanizması çalışma kapsamı dışında bırakılmıştır.

Son olarak eksik veri çözümleme yöntemlerinden EM, MI ve Karar Ağaçları çözümleme yöntemleri kullanılmış, klasik eksik veri çözümleme yöntemleri ve

makine öğrenmesi algoritmaları başlığı altında tanıtılan YSA kapsam dışı bırakılmıştır. YSA ile yapılan eksi veri çözümleme çalışmaları veri setinin özel olarak modellemesinde yani giriş katman değeri ile çıkış katman değeri aynı olacak şekilde oluşturulması işlemlerinden geçer. Sinir ağlarının giriş ve çıkış katmanlarının ayarlanması, test ve eğitim veri setlerine ayrılması gibi özel ayarlar ek olarak yapılır. Ancak GAIN ve WGAIN algoritması veri setleri üzerinde modelleme yapılmadan kullanıldığından YSA kapsam dışında bırakılmıştır. Klasik eksik veri çözümleme yöntemleri başlığı altında bahsedilen çözümleme yöntemlerinin MCAR ve MAR eksiklik mekanizmalarında birlikte kullanılamaması ve MCAR eksiklik mekanizmasında performansları iyi iken MAR eksiklik mekanizmasında sıklıkla kötü performans gösterebilmelerinden bahsedilmişti. Bu bağlamda klasik eksik veri çözümleme yöntemleri kapsam dışında bırakılmıştır.

3.3. VERİ SETİ

Veri seti, UCI Makine Öğrenimi Havuzu' ndan (University of California Irvine Machine Learning Repository – Kaliforniya Üniversitesi Irvin Makine Öğrenmesi Havuzu-UCI) alınmıştır. UCI Makine Öğrenimi Havuzu, makine öğrenimi topluluğu tarafından makine öğrenimi algoritmalarının ampirik analizi için kullanılan veritabanları, etki alanı teorileri ve veri üreteçlerinin bir koleksiyonudur. Arşiv, 1987 yılında David Aha ve UC Irvine'deki yüksek lisans öğrencileri tarafından bir ftp arşivi olarak oluşturulmuştur (Cortez ve diğerleri, 2009: 547).

3.3.1. Veri Setinde MCAR Ve MAR Mekanizmalarında Eksikliğin Oluşturulması

Eksik veri mekanizmaları ve eksiklik oranları, R programlama dilinde bulunan "mice" isimli kütüphanenin "ampute" fonksiyonu kullanılarak yapılmıştır. Bu fonksiyon özellikle eksik veri yöntemlerinin değerlendirilmesinde yararlı olmakla birlikte, planlanan eksik veri tasarımlarının oluşturulması, ölçüm hatasının istatistiksel çıkarımlar üzerindeki etkisinin incelenmesi ve çoklu kaynak verileri alanındaki araştırmalar için de kullanılmaktadır (Schouten, Lugtig ve Vink, 2018: 1).

İki tür eksik veri mekanizmasında (MCAR, MAR) sırasıyla %10, %20, %30, %40, %50 oranlarında eksik değerler oluşturulmuştur.

3.4. EKSİK VERİ DESENİ

Veri kümesindeki eksik veri sorununun karmaşıklığı ve eksik değerlerin yeri hakkında bilgi edinmek için eksik veri deseni hakkında değerlendirme yapılmaktadır (Buuren, 20018: 95). Her eksik veri mekanizması için %10 oranında eksik veri yapısını gösteren tablolar ve grafikler gösterilmiştir. Eksik veri desenini göstermek için R programlama dilinde "naniar" kütüphanesinde bulunan "vis_miss" fonksiyonu ve SPSS'de Eksik Değer Analizi (Missing Value Analysis-MVA) prosedürünün seçenekleri kullanılmıştır.

Tablo 16'da oluşturulan MCAR türündeki eksiklik mekanizması değişkenlerde görülen eksik değerlerin, veri girişi hatası ya da ölçüm sensörlerinin çalışmaması neticesinde verilerin kaydedilmediği varsayımına göre oluşturulmuştur. Bu tür varsayımlarda eksiklik rassal olarak ortaya çıkmaktır.

Tablo 16: MCAR Veri Mekanizmasında ve %10 Eksiklik Oranında Veri Setinin İlk 10 Gözlemini Gösteren Eksik Veri Deseni

Gözlem No	sabit asit	uçucu asit	sitrik asit	artık şeker	klorür	serbest kükürt dioksit	toplam kükürt dioksit	yoğunluk	pH	sülfat	alkol
1	7,4	0,7	0	1,9	0,076	11	34	0,9978	3,51	0,56	9,4
2	NA	0,88	0	2,6	0,098	25	67	0,9968	3,2	0,68	9,8
3	7,8	NA	0,04	NA	0,092	15	54	0,997	3,26	NA	9,8
4	11,2	0,28	0,56	1,9	0,075	17	60	0,998	3,16	0,58	9,8
5	7,4	0,7	0	NA	0,076	11	NA	0,9978	3,51	0,56	9,4
6	7,4	0,66	0	1,8	0,075	13	40	0,9978	3,51	0,56	9,4
7	NA	0,6	0,06	1,6	0,069	15	59	0,9964	3,3	0,46	9,4
8	7,3	0,65	0	1,2	0,065	NA	21	0,9946	3,39	0,47	10
9	7,8	0,58	0,02	2	0,073	9	18	0,9968	3,36	0,57	9,5
10	7,5	0,5	0,36	6,1	0,071	17	102	0,9978	3,35	0,8	10,5

Kaynak: Yazar tarafından oluşturulmuştur.

Tablo 16'da şarap kalitesini sınıflandırmak için 11 değişkenden oluşan veri setinin ilk 10 gözlemi gösterilmiştir. Şarap kalitesini belirleyen değişkenlerden

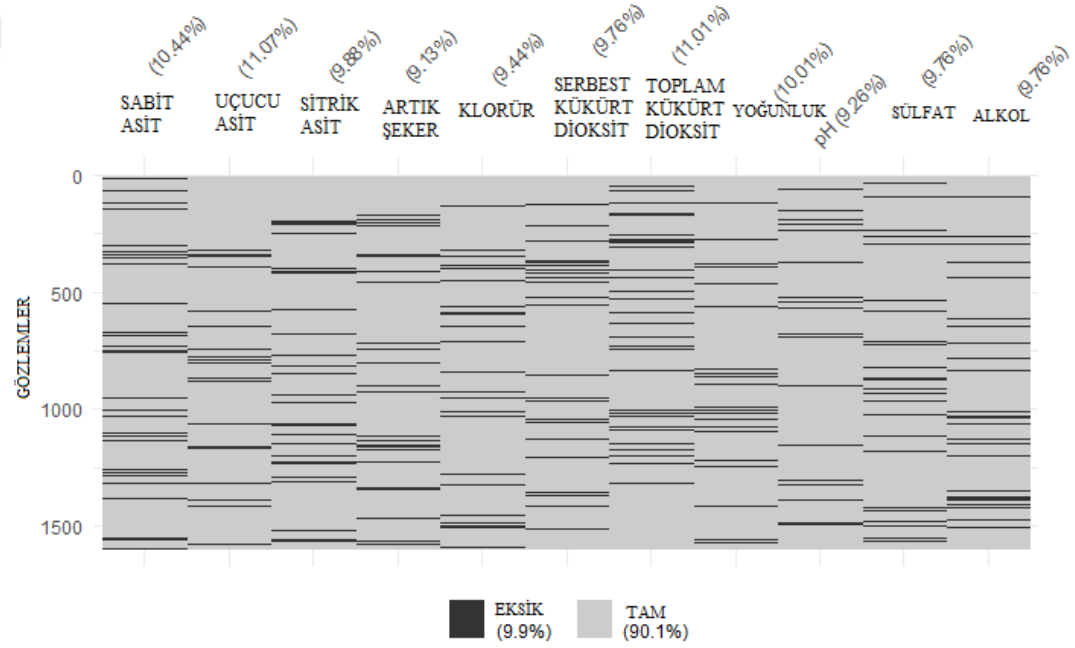
oluşan veri setinde eksik değerlerin varlığı üretilen şarabın kalite puanını analiz etmeyi zorlaştırmaktadır. Eksik verilerin varlığında, üretilen şarabın aldığı puanın, kalite puanlaması tadım uzmanları tarafından yapılan bir puanlama olduğu için, değişkenlerle ilişkisi belirlenemez. Örneğin kalite puanlaması 3 olan bir şarabın değerlendirmesi tadım uzmanı tarafından tat, koku ya da renk gibi duyusal değerlendirmeler üzerine olacak ve üretilen şarabın değişkenleri hakkında sayısal veriler sağlamayacaktır. Kalite puanlaması 7 olan bir şarap ise yine tadım uzmanı tarafından renk, tat ve koku gibi duyusal değerlendirmelerle kaliteli şarap sınıfına girmesine rağmen aynı şarabın üretilmesi için işletmeye herhangi bir sayısal veri vermeyecektir. Bu bağlamda değişkenlerde görülen eksik değerlerin yerinin belirlenmesi ve eksik değerlerin oluşma nedeninin bilinmesi şarap üretimi sürecinde veri seti kullanılarak yapılacak analizler için uygun yöntemlerin seçilmesini (değişkenlerin değerlerinde yapılacak değişiklikler ya da en etkili değişkenlerin seçilmesi gibi), seçilen yöntemlerle hedeflenen kaliteli şarap üretiminin gerçekleşmesini sağlayacaktır.

Eksik veri analizi öncesi ilk aşama olan eksikliğin deseninin tanımlanması sonraki aşamalar için neden önemli olduğundan bahsedildi. Tablo 16, eksik verilerin dağıtılmış ve rasgele bir şekilde meydana geldiği bir senaryo olan rasgele eksik veri desenini göstermektedir. Tablo 16'da NA (Not Available) ile doldurulmuş gözlem değerleri eksik verileri göstermektedir. Eksik değerlere sahip gözlemler incelenerek değişkenlerin gözlem bazında eksik değerlere sahip olup olmadığı ve eksik değerlerle ilgili rassallık bilgisi elde edilebilmektedir. Bu bağlamda, Tablo 16' da NA değerlerinin gözlem bazında rassal olarak ortaya çıktığı gözlenmiştir. Örneğin, 2 numaralı gözlemde "sabit asit" isimli değişkende gözlemlenen eksikliğin başka bir değişken ile aynı anda ortaya çıkmadığı gözlenmiştir. Bir başka eksiklik durumunun olduğu 3 numaralı gözlemdeki NA değerlerine sahip değişkenlerin (uçucu asit, artık şeker ve sülfat değişkenleri) 5 numaralı gözlemde NA değerlere sahip olmadığı gözlenmiştir. Bu, değişkenlerin sahip olduğu eksik değerlerin hem kendisinden hem de diğer değişkenlerden bağımsız olarak ortaya çıktığı yani MCAR varsayımına uyduğunu anlamına gelmektedir.

İlk 10 gözlemi gösteren Tablo 16 araştırmacılar ya da işletme için stratejik kararların alınmasına yardımcı olan veri bilimciler için yeterli görülmeyebilir. Veri

setindeki bütün eksiklik değerlerini gösteren tablo çıktısı ile eksik değerlerin rassallığını incelemek mümkündür. Ancak araştırma konusu kapsamında kullanılan veri seti hacim olarak küçük olmasına rağmen bu seçenek yine de kullanılmamıştır. Çünkü bu yöntem detaylı eksiklik analizleri için daha kullanışlıdır. Bunun yerine rassalık tespiti için tam veri setindeki eksiklik durumunu gösteren görsel çıktı kullanılmıştır. Bu görsel Şekil 4'te gösterilmektedir:

Şekil 4: MCAR Eksik Veri Mekanizmasında ve %10 Eksiklik Oranında Tüm Eksik Verilerin Veri Setindeki Yerini Gösteren Eksik Veri Deseni Yapısı

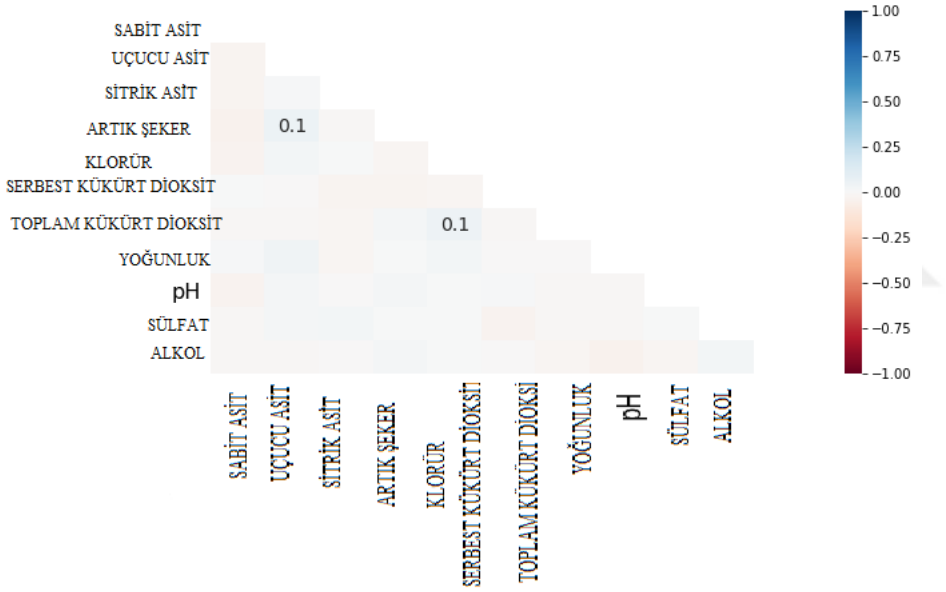


Şekil 4 MCAR eksik veri mekanizmasında ve %10 eksiklik oranında eksik verilerin veri setindeki yerlerini göstermektedir. "Gri" renkler gözlem bazında eksik olmayan (tam değerleri) değerleri gösterirken "siyah" renkler gözlem bazında eksik değerleri gösterir. Eksik değerlerin bağımsızlığı, her değişken için satır bazında görülen "siyah" renklerin (eksik gözlemlerin) diğer satırlarda görülmemesi durumunda anlaşılmaktadır. Bütün bir veri setindeki eksik değerlerin yerini gösteren görsel incelendiğinde şarap kalitesini belirleyen değişkenlerde eksiklik durumlarının ortaya çıkma durumu birbirinden ya da değişkenin kendisinden bağımsız olduğu görülmektedir. Ancak eksikliğin başka değişkenlerden dolayı olup olmadığı

rahatlıkla görülebilirken kendisinden kaynaklanıp kaynaklanmadığını anlamanın yolu ısı-haritalarını (heatmap) kullanmaktır. Isı haritaları aracılığıyla eksiklik durumunun rassallığı araştırılabilmektedir. Şekil 5'te ilgili yönteme ait görsel sunulmuştur.

Eksik veri deseninin tanımlanması, değişkenlerde gözlemlenen eksik veri problemleri için uygun çözümleme yöntemlerini seçilmesini sağlamaktadır. Uygun eksik veri tamamlama yönteminin seçilmesi kaliteli şarap üretimi için gerekli olan değişkenlerin değerlerinin doğru hesaplanmasını ve hedeflenen kaliteli şarap üretiminin gerçeklemesini sağlayabilmektedir.

Şekil 5: MCAR Eksik Veri Mekanizmasında ve %10 Eksiklik Oranında Eksik Verilere Sahip Değişkenlerin Birbiriyle İlişkisini Gösteren Isı Haritası (Heat-Map)



Isı haritaları değişkenlerin birbirleriyle ilişkisini göstermek için kullanılan bir görselleştirme tekniğidir. Isı haritası, bir olgunun büyüklüğünü iki boyutta renk olarak gösteren bir veri görselleştirme tekniğidir. Renkteki değişiklik, renk tonu ya da yoğunluğu ile olabilir, bu da okuyucuya fenomenin nasıl kümelendirildiği ya da uzayda (örnek uzayında) değiştiği hakkında açık görsel ipuçları vermektedir (Leland ve Micheal, 2009: 179).

Şekil 5 te MCAR eksiklik mekanizması ve %10 eksiklik oranına sahip veri setinin ısı haritası gösterilmiştir. Yukarıdaki tanımdan yola çıkarak ısı haritasını eksikliğin şarap kalitesini etkileyen değişkenler arasındaki ilişkisini görebilmek için kullanmak mümkündür. Isı haritalarının değişkenlerdeki eksiklik mekanizmalarının tespitinde kullanımı değişkenler arasındaki yokluk ilişkisini (nullity correlation) -1 ve +1 değerleri arasında sayısal olarak göstermektedir. Böylece eksikliğin ortaya çıkarken değişkenlerin birbirini etkileme gücünü gözlemek mümkün olmaktadır. Örneğin; "artık şeker" isimli değişkende gözlemlenen eksik değerlerin %10 u "uçucu asit" isimli değişken ile ilişkilidir. Bu bağlamda, % 10 ilişki oranı eksiklik mekanizmasının MAR ya da MNAR türünde eksiklik mekanizması olduğunu söylemek için yeterli bir değer değildir. Bir başka değişken olan " toplam kükürt dioksit" isimli değişkende gözlemlenen eksiklik durumunun " serbest kükürt dioksit" isimli değişken ile ilişkisi %10 oranındadır. Aynı şekilde bu oran eksiklik mekanizmasının MAR ya da MNAR olduğunu söylemek için yeterli bir oran değildir.

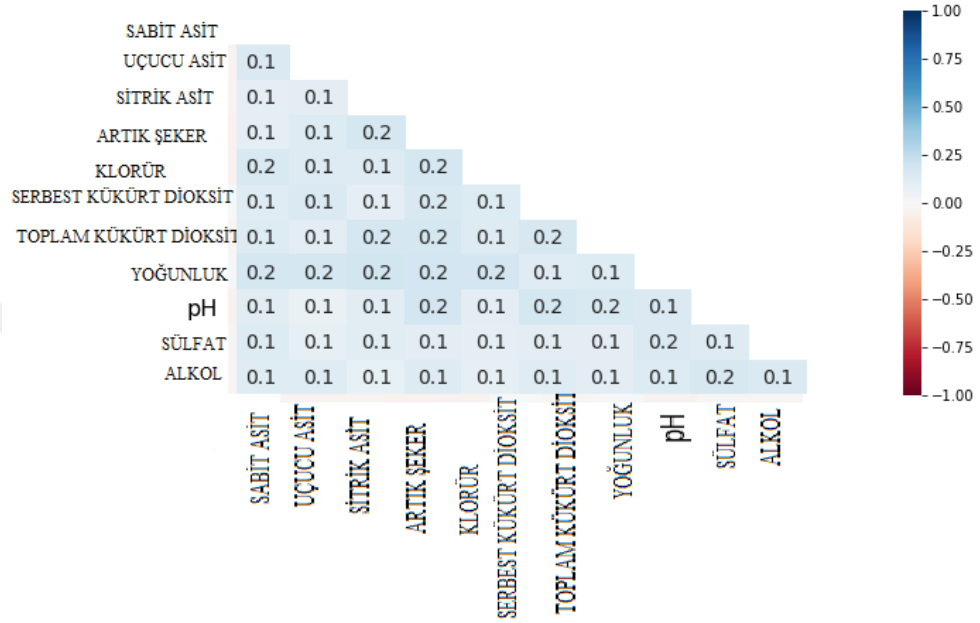
Diğer değişkenlerin birbirleriyle ilişkisi olmadığı için boş değerler açık renklerle gösterilmiştir. Şekil 5 te sağda gösterilen değer skalası değişkenlerin birbirleriyle ilişkisine göre alacakları renk ve sayısal değerleri temsil etmektedir. Dolayısıyla değişkenler arasında ilişki yoksa renk sıfıra yakın renkleri ve sayısal ifade olarak hiçbir değerin verilmemesi şeklinde olmaktadır. Bu durum eksiklik mekanizmasının MCAR varsayımını karşıladığını göstermektedir.

Isı haritaları aracılığıyla, şarap kalitesini sınıflandırmak için kullanılan değişkenlerden birinin ya da bir kaçının eksik olması ve eksikliğin ortaya çıkma nedeninin bilinmesi sınıflandırmaya etkisi olan değişkenler için uygun eksik veri tamamlama yöntemlerinin seçilmesi bakımından önemlidir. Uygun eksik veri tamamlama yönteminin seçilmesi değişkenlerin değerlerinin bilinmesini ve hedeflenen kaliteli şarap üretiminin gerçeklemesini sağlayabilmektedir.

MAR mekanizması için tablo gösterimi ya da bütün veri setini gösteren çıktıya yer verilmemiştir. Bunun nedeni MAR eksiklik mekanizmasının ısı haritalarında daha kolay görülebilmesidir Ayrıca tablo gösterimi ve tüm veri setini gösteren görsel çıktı MAR eksiklik mekanizması için yorumlamayı güçleştirmektedir. Bir önceki eksiklik mekanizmasında hem tablo gösterimi hem de

tüm veri setinin görsel çıktısına yer verilmesinin nedeni konu bütünlüğünün sağlanması amacına yardımcı olmasıdır.

Şekil 6: MAR Eksik Veri Mekanizmasında ve %10 Eksiklik Oranında Eksik Verilere Sahip Değişkenlerin Birbiriyle İlişkisini Gösteren Isı Haritası (Heat-Map)



Şekil 6 veri setindeki eksik gözlemlere sahip değişkenlerin birbiriyle ilişkisini göstermektedir. Örneğin, dikey kesitteki "artık şeker" isimli değişkendeki eksik değerler % 20 oranında yatay eksendeki "sitrik asit" isimli değişkenle ilişkilidir. Bir diğer eksiklik ilişkisi, dikey eksende yer alan"sülfat" değişkeni" ile yatay eksendeki "pH" değişkeni arasında % 20 oranındandır. Diğer bütün değişkenler dikey ve yatay kesişim değerlerine göre değerlerdirildiğinde % 10 - % 20 oranlarında ilişkiye sahiptir. Bu oranlar MAR türünde eksiklik mekanizmasına sahip oldukları anlamına gelmektedir.

Şarap kalitesini sınıflandırmak için kullanılan değişkenlerden birinin ya da bir kaçının eksik olması ve eksikliğin ortaya çıkma nedeninin bilinmesi sınıflandırmaya etkisi olan değişkenler için uygun eksik veri tamamlama yöntemlerin seçilmesi bakımından önemlidir. Eksikliğin MAR türünde eksikik mekanizmasına sahip olmasının belirlenmesi uygun eksik veri tamamlama yönteminin seçilmesini, kaliteli

şarap üretimi için gerekli olan değişkenlerin değerlerinin doğru hesaplanmasını ve hedeflenen kaliteli şarap üretiminin gerçeklemesini sağlayabilmektedir.

3.5. VERİ SETİNDE EKSİK VERİLERİN RASSALLIĞININ TEST EDİLMESİ

Eksik değerlere sahip çok değişkenli verilerle karşılaşıldığında sıklıkla ortaya çıkan bir sorun, eksik verilerin MCAR olup olmadığıdır; diğer bir deyişle, eksikliğin veri kümesindeki değişkenlere ilişkisine dair bilginin bilinmemesidir. Eksik veri çözümleme yöntemlerinin performansı, temeldeki eksik veri mekanizmasına bağlıdır (Little, 1988: 1198).

Daha önce bahsedildiği gibi, MCAR ve MCAR olmayan mekanizmalar arasındaki fark, eksik veri olasılığı ile gözlemlenen değişkenler arasındaki ilişkiye bağlıdır. Bu ilişki algılanamazsa (ve verilerin eksik olmasının belirli bir nedeni yoksa) verilerin MCAR olduğu varsayılır. Bir tür ilişki varsa, eksik veriler MAR ya da MNAR olarak değerlendirilir.

Uygulamada değişkenlerin birbirleriyle ilişkili sonuçları incelenmekte ve ölçülmektedir. Bu, MAR eksik veri mekanizması için çoğunlukla pratikte kabul edilmiş ve kullanılabilen eksik veri varsayımı yapmaktadır (Little, 1988: 1199).

Çalışma kapsamında veri setinde yapay olarak oluşturulan eksiklik mekanizmalarının türü Little'ın MCAR testi kullanılarak test edilmiştir (Little, 1988: 1200). İki eksik veri mekanizması için ve %10 seviyesindeki eksiklik oranları test edilmiş ve eksiklik oranının test sonuçları Tablo 17 ve Tablo 18'de gösterilmiştir.

Tablo 17, % 10 oranında eksikliğe sahip şarap kalitesini sınıflandırmak için kullanılan veri setine uygulanan Little' ın MCAR testi sonuçlarını göstermektedir.

Tablo 17: Little'ın MCAR Testi Bilgilerini İçeren Çıktı Tabloları. MCAR Türünde Ve %10 Eksiklik Oranında Veri Setine Ait Test Sonuçları

Beklenti Maksimizasyon Ortalalamarı(EM Means)										
Sabit asit	Uçucu asit	Sitrik asit	Artık Şeker	Klorür	Serbest Kükürt Dioksit	Toplam Kükürt Dioksit	Yoğunluk	pH	Sülfat	Alkol
8,3292	,52605	,2713	2,52793	,0878	15,91	46,55	,9968	3,3005	,6574	10,4141
a. Little' ın MCAR testi: Ki-kare= 1418,710, Serbestlik Derecesi = 1454**, Sig. = ,741**										

Tablo 17'de Little'ın MCAR testinin sonuçları EM ortalamaları tahmin tablosundan okunmaktadır. Little'ın MCAR testi için sıfır hipotezi (H_o), verilerin MCAR olmasıdır. Eksik değerler içeren veri setinde anlamlılık değeri **Sig. =0,741> 0,05'ten** büyük olduğu için H_o rededilemez. Yani eksik verilerin oluşturduğu mekanizma MCAR 'dır.

Little' ın MCAR testi eksiklik mekanizmasının belirlenmesinde kullanılan başka bir yöntemdir. MCAR türünde eksiklik mekanizmasının belirlendiği test aracılığıyla şarap kalitesini sınıflandırmak için kullanılan değişkenlerden birinin ya da bir kaçının eksik olması ve eksikliğin ortaya çıkma nedeninin bilinmesi sınıflandırmaya etkisi olan değişkenler için uygun eksik veri tamamlama yöntemlerin seçilmesi bakımından önemlidir. Uygun eksik veri tamamlama yönteminin seçilmesi kaliteli şarap üretimi için gerekli olan değişkenlerin değerlerinin doğru hesaplanmasını ve hedeflenen kaliteli şarap üretiminin gerçeklemesini sağlayabilmektedir.

Tablo 18 % 10 oranında eksikliğe sahip şarap kalitesini sınıflandırmak için kullanılan veri setine uygulanan Little' ın MCAR testi sonuçlarını göstermektedir.

Tablo 18: Little'ın MCAR Testi Bilgilerini İçeren Çıktı Tabloları. MAR Türünde ve %10 Eksiklik Oranında Veri Setine Ait Test Sonuçları

Beklenti Maksimizasyon Ortalalamarı(EM Means)										
Sabit asit	Uçucu asit	Sitrik asit	Artık Şeker	Klorür	Serbest Kükürt Dioksit	Toplam Kükürt Dioksit	Yoğunluk	pH	Sülfat	Alkol
8,3208	,5281	,2697	2,5117	,0866	15,91	46,54	,9968	3,3105	,6561	10,4112
a. Little'ın MCAR testi: Ki-kare = 3972,301, Serbestlik Derecesi = 1682**, Sig. = ,000**										

Tablo 18 de Little'ın MCAR testinin sonuçları EM ortalamaları tahmin tablosundan okunmaktadır. Little'ın MCAR testi için sıfır hipotezi (H_o), verilerin MCAR olmasıdır. Eksik değerler içeren veri setinde anlamlılık değeri i **Sig. = ,000<0,05'ten** küçük olduğu için eksik verilerin oluşturduğu mekanizma MCAR değildir. Yani sıfır hipotezi (H_o) rededilir. MAR türünde eksiklik mekanizması vardır.

Little' ın MCAR testi eksiklik mekanizmasının belirlenmesinin başka bir yoludur. MAR türünde eksiklik mekanizmasının belirlendiği test aracılığıyla şarap kalitesini sınıflandırmak için kullanılan değişkenlerden birinin ya da bir kaçının eksik olması ve eksikliğin ortaya çıkma nedeninin bilinmesi sınıflandırmaya etkisi olan değişkenler için uygun eksik veri tamamlama yöntemlerin seçilmesi bakımından önemlidir. Uygun eksik veri tamamlama yönteminin seçilmesi kaliteli şarap üretimi için gerekli olan değişkenlerin değerlerinin doğru hesaplanmasını ve hedeflenen kaliteli şarap üretiminin gerçeklemesini sağlayabilmektedir.

3.6. DEĞERLENDİRME YÖNTEMİ

Eksik veri çözümleme yöntemlerinin amacı eksik verilerin bulunduğu veri setlerinden istatistiksel olarak geçerli çıkarımların elde edilmesini sağlayacak tam veri setlerini oluşturmaktır. Dolayısıyla atama yöntemlerinin kalitesi bu amaç doğrultusunda değerlendirilir. Araştırma konusu için seçilmiş değerlendirme kriteri RMSE (Root Mean Square Error- Hata Karelerinin Kök Ortalaması) dir.

RMSE, tahmin edilen değerler ile gerçek değerler arasındaki uzaklığı ölçen bir metriktir.

RMSE değerlendirme kriterinin seçilmesinin nedeni eksik veri çözümleme yöntemlerinin performansını değerlendirmek için literatürde sıklıkla bu yöntemin kullanılmış olmasıdır (Buuren, 2018: 57). Ayrıca araştırma konusu kapsamında tanıtılan GAIN algoritmasının performansı RMSE kullanılarak test edilmiştir (Yoon ve diğerleri, 2018: 5696).

RMSE orijinal tam veri setini atanmış verilerin olduğu veri setleriyle kıyaslayarak gerçek değere ne kadar yakın olduğunu göstermektedir. Eksik veri çözümleme yöntemleriyle atanmış değerlerin orijinal veri setindeki değerlere yakın ya da aynı değerleri almasının tespiti kaliteli şarap üretimi için önemlidir. Olası eksik

değerlere sahip değişkenler kullanılarak kaliteli şarap üretimi yapılamayacağından daha önce elde edilmiş RMSE değeri referans alınarak uygun atama yönteminin seçimi yapılabilmektedir. Dolayısıyla istenen kaliteli şarap üretimi hedefine ulaşılması sağlanacaktır.

3.7. KULLANILAN YAZILIMLAR VE ORTAM

WGAIN algoritması GAIN algoritmasıyla ve diğer atama yöntemleriyle kıyaslanmıştır. Seçilen atama yöntemleri MissForest, MICE, ve Amelia II kütüphanelerindeki yöntemlerdir.

Amelia II, R programlama dilinde bulunan EM yaklaşımını kullanarak atama yapılmasını sağlayan bir paket yazılımdır. MissForest R programlama dilinde bulunan ve Karar Ağaçları temelli çalışan atama yöntemi paket yazılımıdır (Stekhoven, 2013: 1).

MICE ise çoklu atama yapılmasını sağlayan R programlama dilinde bulunan bir paket yazılımdır (Buuren, 2019: 1).

WGAIN ve GAIN algoritması Python programlama dili ve bir Derin Öğrenme kütüphanesi olan Tensorflow kullanılarak Google'ın makine öğrenmesi ve derin öğrenme uygulamalarının çalıştırılabildiği Colab isimli bulut tabanlı sanal makine ortamı kullanılarak çalıştırılmıştır.

3.8. EKSİK VERİ ÇÖZÜMLEME YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Eksik veri tamamlama yöntemleri MCAR ve MAR türündeki eksik veri mekanizmasında ve %10 %20 %30 %40 %50 eksiklik oranlarındaki veri seti üzerinde tekrarlı olarak denenmiştir. Toplamda 5 defa deneme yapılmış olup ortalama RMSE değerleri MCAR eksiklik mekanizması için Tablo 19 'da, MAR eksiklik mekanizması için Tablo 20'de gösterilmiştir. Düşük RMSE değeri daha iyi performans göstergesidir ve bunu belirtmek için kalın yazı tipinde gösterilmiştir.

Tablo 19: MCAR Eksik Veri Mekanizmasında Eksik Veri Tamamlama Yöntemlerinin RMSE Değerlerini Gösteren Sonuç Tablosu

Algoritmalar	**MCAR Türünde Eksiklik Yüzdeleri**				
	% 10	% 20	% 30	% 40	% 50
WGAIN	**0,0790**	**0,0870**	**0,1043**	**0,1137**	**0,1309**
GAIN	0,0900	0,0950	0,1108	0,1223	0,1314
MissForest	0,1210	0,1217	0,1229	0,1244	0,1318
MICE	0,1038	0,1397	0,1653	0,1665	0,2008
EM	0,1911	0,1856	0,191	0,1983	0,2013

Tablo 20: MAR Eksik Veri Mekanizmasında Eksik Veri Tamamlama Yöntemlerinin RMSE Değerlerini Gösteren Sonuç Tablosu

Algoritmalar	**MAR Türünde Eksiklik Yüzdeleri**				
	% 10	% 20	% 30	% 40	% 50
WGAIN	**0,1148**	**0,1209**	**0,1288**	**0,1369**	**0,1398**
GAIN	0,1204	0,1262	0,1307	0,1393	0,1465
MissForest	0,1392	0,13484	0,13246	0,1434	0,1502
MICE	0,1367	0,1345	0,1673	0,21	0,1918
EM	0,2015	0,1951	0,2034	0,2197	0,1983

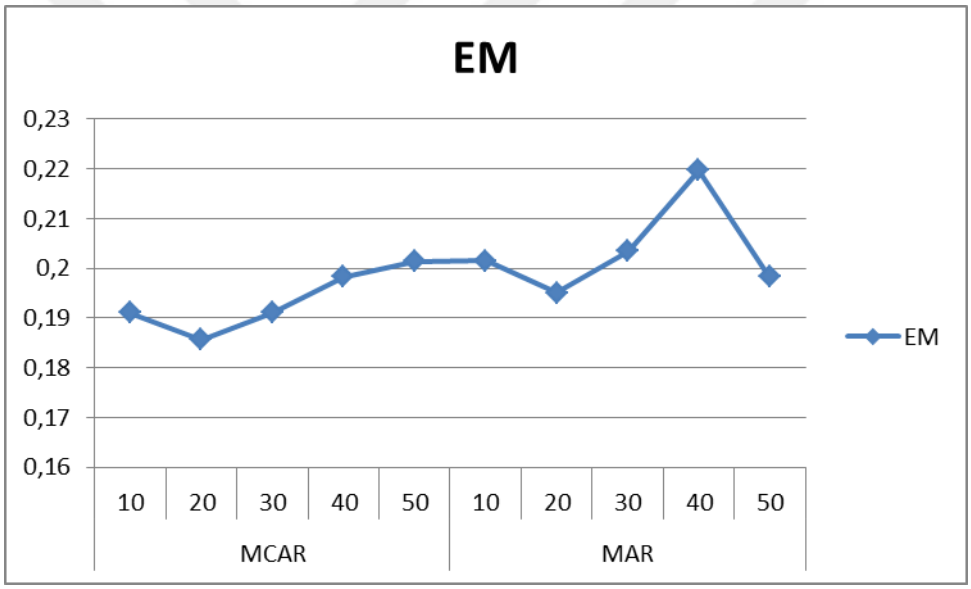
Tablo 19 ve 20 incelendiğinde atama yöntemlerinin performansının değişiklik gösterdiği ve MCAR eksiklik mekanizmasında ve %10, %20, %30, %40 ve %50 oranındaki eksiklik oranlarında WGAIN algoritmasının diğer çözümleme yöntemlerine göre daha düşük RMSE değerine sahip olduğu gözlenmiştir. MAR eksiklik mekanizmasında ve %10, %20, %30, %40 ve %50 eksiklik oranlarında WGAIN algoritmasının diğer çözümleme yöntemlerine göre daha düşük RMSE değerine sahip olduğu gözlenmiştir.

Yukarıda özetlenen RMSE değerleri, eksik değerlere sahip veri seti ile kaliteli şarap üretmek isteyen bir işletmenin, üretim öncesi eksiksiz veri setine ihtiyacı olacağından, seçmesi gereken algortimanın MCAR ve MAR eksiklik mekanizması

için WGAIN algoritması olması gerektiğini göstermektedir. Farklı eksiklik mekanizması ve eksiklik yüzdelerinde seçilmesi gereken algoritmalar kalite sınıflandırması için gerekli olan değişken değerlerinin belirlenmesini sağlayacaktır. Değişkenlerin eksiksiz ve az hata payı ile atanmış değerlere sahip olması olası kalitesiz şarap üretimini engelleyecektir. Veri setinde kullanılan atama yöntemlerinin RMSE değerlerine göre performansının görsel çıktıları Şekil 7, Şekil 8, Şekil 9, Şekil 10 ve Şekil 11'de gösterilmiştir.

Şekil 7'de EM algoritmasının algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında RMSE performansı gösterilmiştir:

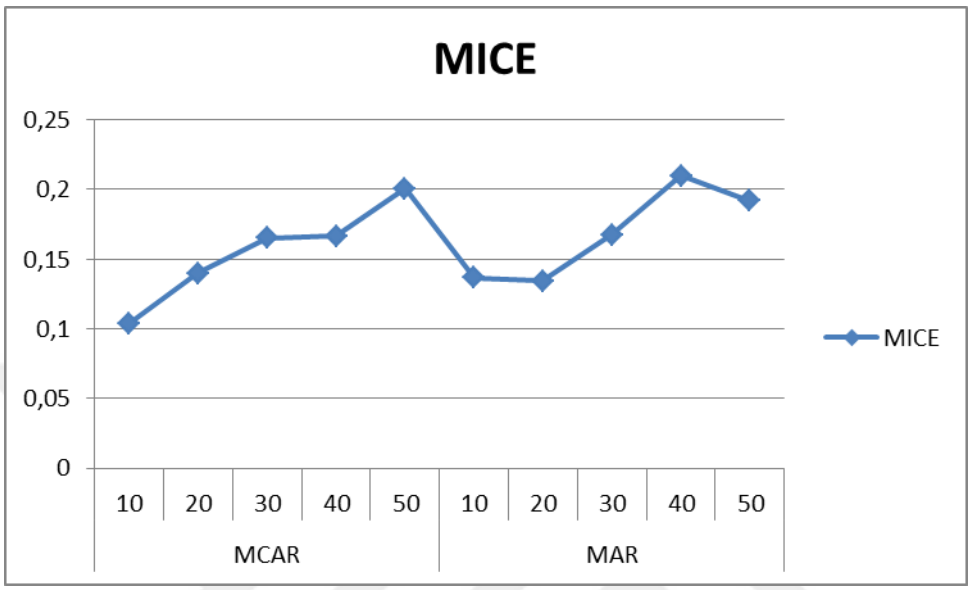
Şekil 7: EM Algoritmasının RMSE Değerine Göre Performansı



EM algoritmasının MCAR eksi veri mekanizmasında ve farklı eksiklik oranlarında önce düştüğü sonra ise yükseldiği gözlenmiştir. Aynı durum MAR eksik veri mekanizmasında da gözlenmiştir. Ancak eksiklik oranı %40 tan %50 ye çıktığında RMSE değerinde düşüş gözlenmiştir.

Şekil 8'de MICE algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında RMSE performansı gösterilmiştir.

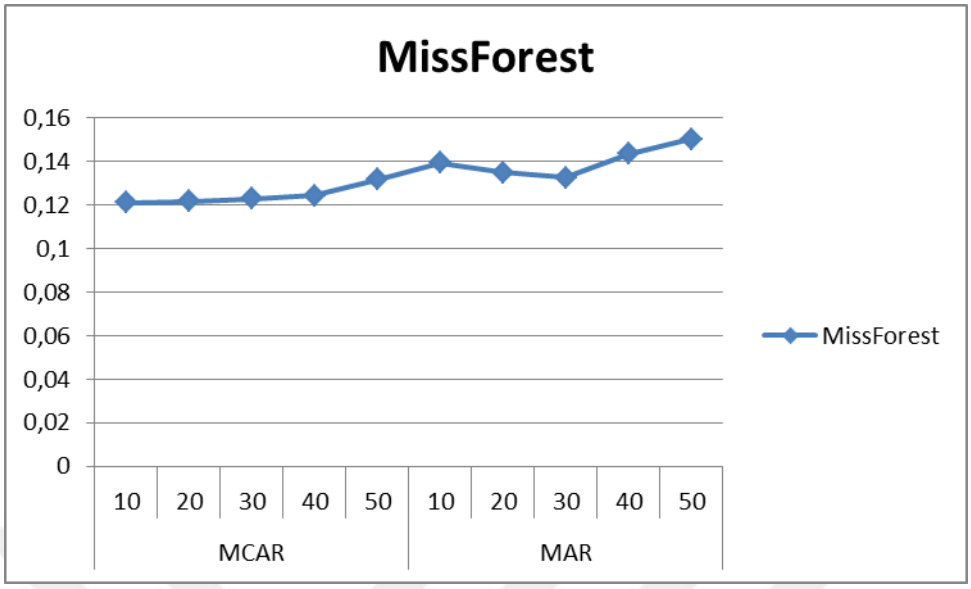
Şekil 8: MICE Algoritmasının RMSE Değerine Göre Performansı



Şekil 8'de MICE algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında performansı incelendiğinde, eksiklik yüzdesi artıkça RMSE değerinin yükseldiği gözlenmiştir. MAR türündeki eksiklik mekanizmasında ise önce çok az değiştiği sonra tekrar yükseldiği ve eksiklik yüzdesi %40'tan % 50 ye çıktığında tekrar düştüğü görülmektedir. Bununla birlikte MCAR ve MAR eksiklik mekanizmaları karşılaştırıldığında en düşük RMSE değerine MCAR türündeki eksiklik mekanizmasında sahip olduğu görülmektedir.

Şekil 9'da MissForest algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında RMSE performansı gösterilmiştir.

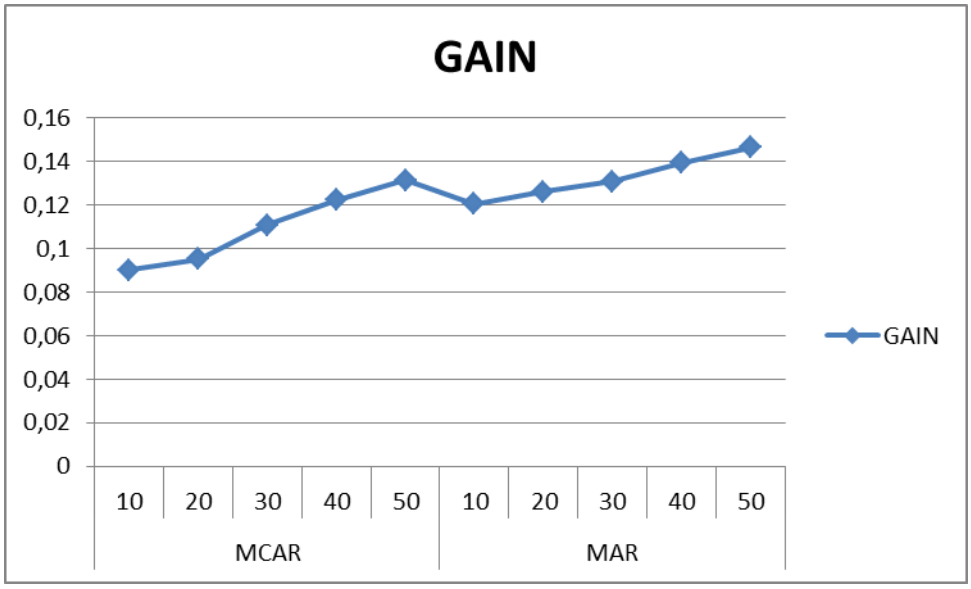
Şekil 9: MissForest Algoritmasının RMSE Değerine Göre Performansı



Şekil 9'da MissForest'ın (Karar Ağaçları) eksiklik yüzdesi artıkça ve farklı eksiklik mekanizmalarında RMSE değerinin yükseldiği görülmektedir. Bununla birlikte MCAR ve MAR eksiklik mekanizmaları karşılaştırıldığında en düşük RMSE değerine MCAR türündeki eksiklik mekanizmasında sahip olduğu görülmektedir.

Şekil 10'da GAIN algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında RMSE performansı gösterilmiştir.

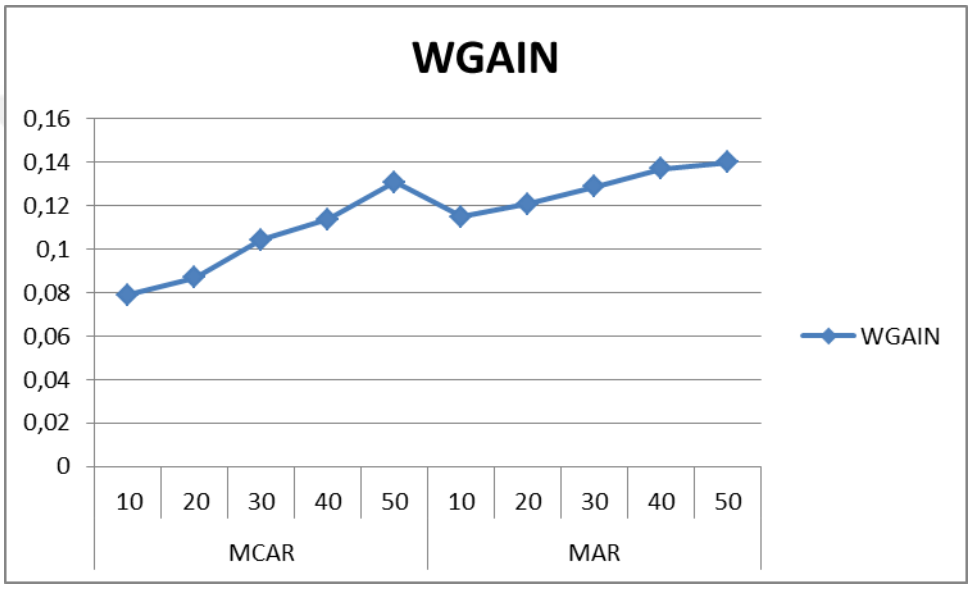
Şekil 10: GAIN Algoritmasının RMSE Değerine Göre Performansı



Şekil 10'da GAIN algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında performansı incelendiğinde, eksiklik yüzdesi artıkça RMSE değerinin yükseldiği görülmektedir. Bununla birlikte MCAR ve MAR eksiklik mekanizmaları karşılaştırıldığında en düşük RMSE değerine MCAR türündeki eksiklik mekanizmasında sahip olduğu görülmektedir.

Şekil 11'de WGAIN algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında RMSE performansı gösterilmiştir.

Şekil 11: WGAIN Algoritmasının RMSE Değerine Göre Performansı



Şekil 11'de WGAIN algoritmasının MCAR ve MAR türündeki eksiklik mekanizmasında ve farklı eksiklik oranlarında performansı incelendiğinde, eksiklik yüzdesi artıkça RMSE değerinin yükseldiği görülmektedir. Bu durum MAR türündeki eksiklik mekanizmasında da görülmektedir. Bununla birlikte MCAR ve MAR eksiklik mekanizmaları karşılaştırıldığında en düşük RMSE değerine MCAR türündeki eksiklik mekanizmasında sahip olduğu görülmektedir.

SONUÇ

Veri kalitesi, belirli bir veri kümesinin ne kadar güvenilir olduğunu göstermektedir. Veri kalitesi kuruluşların kampanyaları ve bütçeleri hakkında hızlı bir şekilde karar vermelerini sağlayabilmektedir. Kuruluşlar yanlış, eksik veya güncel olmayan veriler üzerine kararlar aldıklarında, tüketicilerinin tercihlerini yansıtmayan stratejiler veya politikalar yürütme riskiyle karşı karşıya kalabilmektedirler.

İşletmelerin veriye dayalı stratejilerinin hedefine ulaşması kullanılan verilerin kalitesine bağlıdır. Ancak, işletmelerde kullanılan verilerin çoğu hatalı olabilmektedir. Hatalı verilerin analizi sonucunda da hatalı bilgilere ulaşılmaktadır. Bu araştırmada istatistiksel analiz öncesi eksik veri analiz teknikleri ile bilgisayar bilimleri alanındaki veri kalitesi metodolojisi bütünlük boyutu kapsamında birleştirilerek, şarap kalitesi isimli veri setinin kalitesini iyileştirmek için yapılması gerekenler tüm aşamalar ile ele alınmaya çalışılmıştır.

Şarap sertifikasını değerlendirmek için fizikokimyasal ve duyusal testler kullanılmaktadır. Şarapların sınıflandırılması, üst alanının karmaşıklığı ve heterojenliği nedeniyle kolay bir süreç değildir. Şarapların sınıflandırılması farklı nedenlerden dolayı çok önemlidir. Bu nedenlerden bazıları; şarap ürünlerinin ekonomik değeri, şarapların kalitesini korumak, şarapların karıştırılmasını engellemek ve içecek işlemeyi kontrol etmektir. Şarap kalitesi isimli veri setini kullanan işletmenin amacı, bir uzmanın asitlik ve alkol bileşimi gibi bir dizi fizikokimyasal özellik kullanarak bir şarap örneğine vereceği derecelendirmeyi tahmin etmektir. Derecelendirmenin tahmin edilmesi şarap sınıflandırmasına yardımcı olmaktadır. Veri setindeki bileşen değerleri şarap üretimi yapmak için gerekli bilgileri barındırmaktadır. Şarap uzmanı tarafından 7 puan alan bir şarabın bileşen değerleri bilinmediğinde aynı üretimi yapmak mümkün olmayacaktır. Kalite puanı 3 olan şarap için de bileşen değerleri bilinmediğinde aynı kalitede şarap üretimi tekrarlanacaktır. Şarap kalitesini etkileyen bileşenlerden pH, alkol ve yoğunluk bileşen değerlerinin bilinmemesi işletmenin üzerinde kontrolünün olmadığı bir süreçle üretim yapmasına neden olacaktır. pH, alkol ve yoğunluk gibi kalite puanını etkileyen bileşenlerin değerinin bilinmesi üretimin hangi değerlerle yapıldığı

ve nasıl yapılacağı hakkında bilgi verirken hedeflenen kalitede şarap üretiminin gerçekleşmesini sağlayacaktır.

Bu çalışmanın konusunu oluşturan şarap veri kalitesinin iyileştirilmesi üzerine yapılan uygulamada ilk olarak; tam veri seti üzerinde yapay olarak MCAR ve MAR eksiklik mekanizmalarında %10' dan %50' ye kadar oranlarda eksik değerler yaratılmıştır. Böylece işletmenin karşılaşabileceği eksik veri problemlerinin türleri ve veri setindeki eksiklik oranları kurgulanmıştır.

Uygulamanın ikinci aşamasında veri setindeki eksik değerler, beklenti maksimizasyonu, karar ağaçları, çoklu atama yöntemleri, üretici çekişmeli atama ağları ve geliştirilmiş versiyonu wasserstein üretici çekişmeli atama ağları ile tamamlanmıştır. Elde edilen yeni veri seti orijinal veri setiyle gerçek değerlere yakınlığının ölçülmesi bakımından hesaplanmıştır. Bu hesaplama için RMSE değerlendirme yöntemine başvurulmuştur. Daha sonra elde edilen RMSE değerleri tablo gösterimi ve görsel çıktı yardımıyla uygun yöntemin MCAR ve MAR eksiklik mekanizmasında çalışma performansı gösterilmiştir. RMSE yönteminin kullanıldığı aşama, işletmelerin atama sonrası veri setinin kalitesini değerlendirmesi bakımından önemli bir aşamadır. Ayrıca RMSE, eksik değerler tam veri setlerinde yapay olarak oluşturulmasına rağmen, doğal nedenlerden eksik değerlere sahip veri setlerinde problemin çözümü için kullanılan yöntemin kalitesini ölçmek için de kullanılabilmektedir.

Analiz sonuçlarında WGAIN'in performansının en yakın iki yöntemden belirgin bir şekilde daha iyi performans gösterdiği tespit edilmiştir. Bunun nedeni gözlemlenen verilerde bilginin azalmasıyla (daha fazla eksik değerin olması nedeniyle) kullanılan yöntemlerde atama kalitesi daha önemli hale gelmekte ve WGAIN'ın değişen eksiklik oranlarına rağmen önemli ölçüde kaliteli atama yapmasıdır. WGAIN hem kalite puanı bilinen ancak değişken değerleri eksik bir şarabın üretim değerlerinin bilinmesini çok az hata ile sağlayacak hem de kalite puanı bilinmeyen ve değişken değerleri eksik şarabın üretim değerlerinin bilinmesini yine çok az hata ile sağlayarak şarap sınıflandırılmasını mümkün kılacaktır.

Araştırmanın önemi, eksik veri çözümleme yöntemleri arasında kullanımı Türkçe kaynaklarda nadir bulunan, üretici çekişmeli atama ağlara yer verilmiş

olmasıdır. İkinci bölümün son başlığı altında üretici çekişmeli atama ağların değiştirilmiş versiyonu wasserstein üretici çekişmeli atama ağlarından bahsedilmiştir.

Bu çalışmada eksik veri tamamlama görevi için üretici modeller sınıfına giren GAIN algoritmasının geliştirilmiş versiyonu WGAIN'in kullanımı önerilmiştir. Bu yeni mimari, GAN'larda sıklıkla görülen problemlere karşı geliştirilmiş maliyet fonksiyonunun değiştirilmesi fikriyle oluşturulmuştur ve atama probleminin benzersiz özellikleri ile başa çıkabileceği şekilde genelleştirmektedir. Gerçek dünya veri kümesiyle yapılan deneyde, WGAIN için elde edilen RMSE değerleri ile diğer atama tekniklerinden önemli ölçüde daha iyi performans gösterdiği tespit edilmiştir. Atama için yeni, en etkili atama tekniğinin geliştirilmesi tıpta ve diğer alanlarda (veri kümelerinin çoğunda veriler eksiktir) dönüştürücü etkilere sahip olabilir. Gelecekteki çalışmalarda, WGAIN'in daha büyük veri setlerinde ve farklı mimariler kullanılarak performansı araştırılacaktır.

KAYNAKÇA

Abdella, M. ve Marwala, T. (2005a). The Use Of Genetic Algorithms And Neural Networks To Approximate Missing Data In Database. *IEEE 3rd International Conference On Computational Cybernetics*. 24: 577–589.

Abdella, M. ve Marwala, T. (2005b). Treatment Of Missing Data Using Neural Networks. *In: Proceedings Of The IEEE International Joint Conference On Neural Networks.* 1: 598–603.

Afifi, A. A. ve Elashoff, R. M. (1966). Missing Observations in Multivariate Statistics: I. Review Of The Literature. *Journal Of The American Statistical Association.* 61(315): 595- 604.

Afifi, A. A. ve Elashoff, R. M. (1967). Missing Observations In Multivariate Statistics: II. Point Estimation In Simple Linear Regression. *Journal Of The American Statistical Association.* 62(317): 10-29.

Afifi, A. A. ve Elashoff, R. M. (1969a). Missing Observations In Multivariate Statistics. III: Large Sample Analysis Of Simple Linear Regression. *Journal Of The American Statistical Association*. 64(325): 337-358.

Afifi, A. A. ve Elashoff, R. M. (1969b). Missing Observations In Multivariate Statistics. IV: A Note On Simple Linear Regression. *Journal Of The American Statistical Association.* 64(325): 359-365.

Allan, F. E. ve Wishart, J. (1930). A Method Of Estimating The Yield Of A Missing Plot In Field Experimental Work. *The Journal Of Agricultural Science*. 20(3): 399–406.

Allison, P. D. (2000). Multiple Imputation For Missing Data: A cautionary Tale. *Sociological Methods and Research*. 28: 301-309.

Allison, P. D. (2002). *Missing Data*. University of Pennsylvania, USA: Sage Publications.

Alpar, R. (2017). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. Ankara: Detay Yayıncılık.

Anderson, D. R., Sweeney, D.J. ve Williams, T.A. (2011). *Statistics For Business And Economics*. Boston: Cengage Learning.

Anderson, T. W. (1957). Maximum Likelihood Estimates For A Multivariate Normal Distribution When Some Observations Are Missing. *Journal Of The American Statistical Association.* 52(278): 200-203.

Arjovsky, M., Chintala, S. ve Bottou, L. (2017). Wasserstein GAN. *Courant Institute Of Mathematical Sciences: Facebook AI Research*. 1-32

Batini, C. ve Scannapieca, M. (2006). *Data Quality: Concepts, Methodologies And Techniques*. Berlin: Springer Verlag.

Batini, C., Cappiello, C., Francalanci, C. ve Maurino, A. (2009). Methodologies For Data Quality Assessment And Improvement. *ACM Computing Surveys*. 41(3): 16-34.

Batini, C.ve Scannapieca, M. (2016). *Data And Information Quality: Dimensions, Principles And Techniques.* Switzerland: Springer International Publishing.

Bengio,Y., LeCun, Y. ve Hinton, G. (2015). Deep Learning. *Nature*. 521 (7553): 436–444.

Bennet, D. (2001). How Can I Deal With Missing Data?. *Australian And New Zealand Journal Of Public Health*. 25: 464-469.

Bergdahl, M., Booleman, M., Brakel , R. V. ve Grünewald, W. (2001). The Interrelationship Of Different Management Mainframes. *International Conference On Quality In Official Statistics.* 28: 1-6.

Bose, S. S.(1938). Appendix. The Estimation Of Mixed-Up Yields And Their Standard Errors. *Sankhyā: The Indian Journal of Statistics*. 4(1): 112-120.

Bose, S.S. ve Mahalanobis, P. C.(1938). On Estimating Individual Yields In The Case Of Mixed-Up Yields Of Two Or More Plots In Field Experiment. *Sankhyā: The Indian Journal Of Statistics*. 4(1): 103-111.

Bosij, P., Chafey, D., Greasley, A. ve Hickie, S. (2003). *Business Information Systems: Technology, Development and Management*. London: Pearson.

Bovee, M., Srivastava, R. ve Mak, B. (2001). A Conceptual Framework And Belief-Function Approach To Assessing Overall Information Quality. *International Journal Of Intelligent Systems.* 18(1): 51-74.

Bovee, M.W. (1986). *Information Quality: A Conceptual Framework And Empirical Validation.* (Yayınlanmamış Doktora Tezi). Lawrance: Kansas Üniversitesi, School of Business.

Brackstone, G. (1999). Managing Data Quality In A Statistical Agency. *Statistica Canada: Survey Methodology*. 25(2): 1-23.

Brackstone, G. (2001). Managing Data Quality: *The Accuracy Dimension. International Conference on Quality In Official Statistics*. 2(3): 16-32.

Brock, A., Donahue, J. ve Simonyan, K. (2018). Large Scale GAN Training For High Fidelity Natural Image Synthesis. *In: ArXiv abs/1809.11096*. 1-35.

Buck, S. F. (1960). A Method Of Estimation Of Missing Values In Multivariate Data Suitable For Use With An Electronic Computer. *Journal Of the Royal Statistical Society. Series B (Methodological)*. 22(2): 302-306.

Buuren, S. V. (2012). *Flexible Imputation Of Missing Data*. New York: CRC Press

Buuren, S. V. (2018). *Flexible Imputation of Missing Data*. New York: CRC Press.

Buuren, S. V. (2019). *Multivariate Imputation by Chained Equations*. Version: 3.7.0. https://cran.r-project.org/web/packages/mice/index.html. (25.04.2020).

Bülbül, Ş. (2000). *Tanımlayıcı İstatistik*. İstanbul: Der Yayınları.

Canbaş, A. (2006). *Şarap Teknolojisi Ders Notları* (Yayınlanmamış Ders Notları). Adana. Çukurova Üniversitesi Ziraat Fakültesi.

Checkland, P. ve Holwell, S. (1998). *Information, Systems And Information Systems: Making Sense Of The Field*. Chichester: Wiley.

Cihan, P ., Gökçe, E. ve Kalıpsız, O . (2019). A Review On Determination Of Computer Aid Diagnosis And/Or Risk Factors Using Data Mining Methods in Veterinary Field. *Atatürk Üniversitesi Veteriner Bilimleri Dergisi*. 14 (2): 209-220.

Crosby, P. B. (1995). *Quality Is Still Free; Making Quality Certain in Uncertain Times*. New York: McGraw-Hill.

Çilingirtürk, A. M. ve Altaş, D. (2010). Makro İktisat Verilerinde Kayıp Verilerin Regresyona Dayalı En Yakın Komşu 'Hot Deck' Yöntemi İle Tamamlanması. *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi.* 25(2): 76-77.

Das, V. V. (2006). *Principles Of Data Structures Using C and C++*. Delhi: New Age International.

Dear, R. E. (1959). *A principal Component Missing Data Method For Multiple Regression Models.* California: System Development Corporation.

Deming, W.E. (1986). *Out Of The Crisis*. USA: MIT Press.

Dempster, A. P., Laird, N. M. ve Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data Via The EM Algorithm. *Journal Of The Royal Statistical Society*. 39(1): 1-38.

Dhlamini, S. M., Nelwamondo, F. V. ve Marwala, T. (2006). Condition Monitoring Of HV Bushings In The Presence Of Missing Data Using Evolutionary Computing. *Transactions On Power Systems*. 1(2): 280–287.

Dong, Y. ve Peng, C. Y. (2013). Principled Missing Data Methods For Researchers. *SpringerPlus.* 2(1): 222.

Edgett, G. L. (1956). Multiple Regression With Missing Observations Among The Independent Variables. *Journal Of the American Statistical Association*. 51(273): 122- 131.

Elderton, E. M. ve Pearson, K. (1910). A First Study Of The Influence Of Parental Alcoholism On The Physique And Ability Of the Offspring. *Journal of the Royal Statistical Society*. 73(6/7): 769-773.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.

English, L. P. (1999). *Improving Data Warehouse And Business Information Quality: Methods For Reducing Costs And Increasing Profits*. New York: Wiley.

Eurostat. Standart Quality Report. (2000). http://www.unece.org/stats/documents/2000/11/metis/crp.3.e.pdf. (Erişim Tarihi: 17.12.2019)

Fadahunsi, K. P., Akinlua, J. T., O'Connor, S., Wark, P. A., Gallagher, J., Carroll, C., Majeed, A. ve O'Donoghue, J. (2019). Protocol For A Systematic Review And Qualitative Synthesis Of Information Quality Frameworks In eHealth. *BMJ Open*. 9(3): 1-5.

Fisher, C., Lauría, E., Chengalur-Smith, S. ve Wang, R. (2011). *Introduction To Information Quality*. Bloomington: AuthorHouse.

Forrester Consulting. (2011): *Trends In Data Quality And Business Process Alignment*. https://docplayer.net/5030274-Trends-in-data-quality-and-business-process-alignment.html, (Erişim Tarihi: 17.12.2019)

Gartner. (2018). *How to Create a Business Case for Data QualityImprovement*. https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/. (20.06.2020).

Garvin, D. A. (1998). *Managing Quality: The Strategic And Competitive Edge*. New York: The Free Press, Macmillen Inc.

Glasser, M. (1964). Linear Regression Analysis With Missing Observations Among the Independent Variables. *Journal of the American Statistical Association*. 59(307): 834- 844.

Goodfellow, I. (2016). Generative Adversarial Networks. *NIPS 2016 Tutorial*. 1-53

Goodfellow, I., Jean, P. A., Mehdi, M., Bing, X., David, W. F., Ozair, S. C. ve Aaron, B. Y. (2014). Generative Adversarial Networks. *Proceedings of the International Conference on Neural Information Processing Systems*. 2672–2680.

Graham, John W. (2012). *Missing Data: Analysis and Design*. Germany: Springer

Groves, R. M., Floyd, J. F., Mick P. C., James, M. L., Eleanor, S. ve Tourangeau, R. (2004). *Survey Methodology*. USA: John Wiley & Sons

Gürsakal, N. (2007). *Betimsel İstatistik*. Ankara: Nobel Yayın Dağıtım.

Haitovsky, Y. (1968). Missing Data In Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*. 30(1): 67-82.

Halis, M. (2000). *Paradigmadan Uygulamaya Toplam Kalite Yönetimi*. İstanbul: Beta Basım Yayım Dağıtım.

Hartley, H. O. (1958). Maximum Likelihood Estimation From Incomplete Data. *Biometrics*. 14(2): 174-194.

Hartley, H. O.,Hocking, R. R. (1971). The Analysis Of Incomplete Data. *Biometrics*. 27(4): 783-23.

Haykin, S. (1999). *Neural networks (2nd ed.).* New Jersey: Prentice-Hall.

Healy, M. ve Westmacott, M. (1956). Missing Values in Experiments Analysed On Automatic Computers. *Applied Statistics*. 5(3): 203-206.

Ho, S. K. M. (1995). Is The ISO 9000 Series For Total Quality Management?. *International Journal Of Phsical Distribution And Logistics Management*. 25(1): 51-66.

Huang, K. T., Lee, Y. W . ve Wang, R. Y. (1999). *Quality Information And Knowledge*. New Jersey: PrenticeHall

IBM. (2019). *Big Data And Analytics Hub*. https://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data. (20.06.2020).

International Organization For Standardization. (2009). *Geographic Information, Imagery, Gridded And Coverage Data Framework* (IS/TS 19129). https://cdb.iso.org/cdb/termentry!display.action?entry=165880&language=1, (Erişim Tarihi: 14.12.2019).

Ivanov, O., Figurnov, M. ve Vetrov, D. (2019). Variational Autoencoder with Arbitrary Conditioning. *In: International Conference On Learning Representations*. 1-25.

Jackson, R. (2008). *Wine Science: Principles and Applications*. Cambridge: Academic Press.

Jarke, M. M., Lenzerini, Y. ve Vassiliou, P. (1995). *Fundamentals of Data Warehouses*. Berlin: Springer Verlag.

Jones, M. P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. *Journal of the American Statistical Association*. 91(433): 222-230.

Juhola, M. ve Laurikkala, J. (2013). Missing Values: How Many Can They Be To Preserve Classification Reliability. *Artificial Intelligence Review*. 40(3). 231–245.

Juran, M. J. (1992). *Juran On Quality By Design: The New Steps For Planning Quality Into Goodsand Services*. New York: The FreePress, Macmillan Inc.

Karr, A. F., Sanil A. P. ve Banks, D. L. (2006). Data Quality: A Statistical Perspective. *Statistical Methodology*. 3: 137-173.

Kingma, D. ve Welling, M. (2014). Auto-Encoding Variational Bayes. *International Conference on Learning Representations. 1-14.*

Klein, B. D. ve Callahan, T.J. (2007). A Comparison of Information Technology Professionals And Data Comsumers: Perceptions Of The Importance Of the Dimensions Of Information Quality. *International Journal of Information Quality*. 1(4): 392-411.

Kobu, B. (2003). *Üretim Yönetimi*. İstanbul: Avcıol Basım Yayın.

Kotz, S., Johnson, N. L. ve Read, C. B. (2005). Quality Concepts For Official Statistics: *Encyclopedia of Statistical Sciences*. 3: 621-629.

Lakhe ,R. R. ve Mohanty, R.P. (1994). Total Quality Management Concepts, Evaluation And Acceptability in Developing Economies. *International Journal Of Quality And Reliability Management*. 11(9): 9-33.

Langr, J. ve Bok, V. (2019). *Gans In Acton*. New York: Manning Publications.

Lavet, V. F., Henderson, P., Islam, R., Bellemare, M. G. ve Pineau, J. (2018). An Introduction to Deep Reinforcement Learning. *Foundations And Trends In Machine Learning*. 11(3-4): 219-354.

Lee, H., Rancourt, E. ve Srndal, C. E., 1994. Experiments With Variance Estimation From Survey Data With Imputed Values. *Journal Of Official Statistics*. 10: 231-238.

Leke, C. A. ve Marwala., T. (2019), *Deep Learning and Missing Data in Engineering Systems, Studies in Big Data 48*. Switzerland: Springer Nature AG.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*. 83 (404): 1198–1202.

Little, R. J. A. (1992). Regression With Missing X's: A Review. *Journal of the American Statistical Association*. 87(420): 1227-1237.

Little, R. J. A. ve Rubin, D. B. (1987). *Statistical Analysis With Missing Data*. New York: Wiley.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Little, R. ve Rubin, D. (2014). *Statistical Analysis With Missing Data*. New York: Wiley.

Little, R. ve Rubin, D. (2020) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons

Longford, N. T. (2006). *Missing Data And Small-Area Estimation: Modern Analytical Equipment For The Survey Statistician*. New York: Springer.

Lord, F. M. (1955). Estimation of Parameters From Incomplete Data. *Journal of the American Statistical Association*. 50(271): 1227-1237.

Loshin, D. (2006). *Monitoring Data Quality Performance Using Data Quality Metrics*. USA: Informatica.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*. 44(2): 226-233.

Marwala, T. (2009). *Computational Intelligence For Missing Data Imputation: Estimation and Management Knowledge Optimization Techniques*. New York: Information Science Reference.

Maschler, M., Solan, E. ve Zamir, S. (2013). *Game Theory*. London: Cambridge University Press.

Matthai, A. (1951). Estimation Of Parameters From Incomplete Data With Application To Design Of Sample Surveys. *Sankhyā: The Indian Journal of Statistics*. 11(2): 145-152.

McKendrick, A. G. (1926). Applications Of Mathematics To Medical Problems. *Proceedings of Edinburgh Mathematical Society*. 44: 98–130.

Ming-Hau, C. (2010). Pattern Recognition Of Business Failure By Autoassociative Neural Networks In Considering The Missing Values. *International Computer Symposium*. 711–715.

Mitra, A. (1993). *Fundamentals Of Quality Control And Improvement*. New York: Macmillan Publishing Company.

Naumann, F. (2002). *Quality-Driven Query Answering For Integrated Information* Systems. Berlin: Springer.

Nelwamondo, F. V. ve Marwala, T. (2007). Handling Missing Data From Heteroskedastic And Nonstationary Data. *Lecture Notes In Computer Science*. 4491(1): 1297–1306.

Nicholson, G. E. (1957). Estimation of Parameters From Incomplete Multivariate Samples. *Journal of the American Statistical Association*. 52(280): 523-526.

Nierman, D. (2004). *Acidity Is A Fundamental Property Of Wine, Imparting Sourness And Resistance To Microbial Infection*. https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity#:~:text=The%20predominant%20fixed%20acids%20found,malic%2C%20citric%2C%20and%20succinic.&text=Wines%20produced%20from%20cool%20climate,or%20by%20malo%2Dlactic%20fermentation. (Erişim Tarihi: 12.06.2020).

Olinsky, A., Chen, S. ve Harlow, L. (2003). The Comparative Efficacy Of Imputations Methods For Missing Data In Structural Equation Modeling. *European Journal of Operational Research*. 151 (1): 53–79.

Orchard, T., Woodbury, M. a. (1972). A Missing Information Principle: Theory And Applications. *Proceedings Of The Sixth Berkeley Symposium On Mathematical Statistics and Probability*. 1: 691-715.

Orr, K. (1998). Data Quality And Systems Theory. *Communications of the ACM*. 41(2). 66-71.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos VE J. Reis.(2009). Modeling Wine Preferences By Data Mining From Physicochemical Properties. *In Decision Support Systems. Elsevier*. 47(4): 547-553.

Pipino, L. L., Lee, Y. W. ve Wang, R. Y. (2002). Data Quality Assessment. *Communications Of The ACM*. 45(4): 211-218.

Polikar, R., De Pasquale, J., Mohammed, H. S., Brown, G. ve Kuncheva, L. I. (2010). Learn++mf: *A Random Subspace Approach For The Missing Feature Problem. Pattern Recognition*. 43(11): 3817–3832.

Raghunathan, T. (2016). *Missing Data Analysis in Practice*. New York: Chapman and Hall/CRC.

Ramoni, M. ve Sebastiani, P. (2001). Robust Learning With Missing Data. *Journal of Machine Learning*. 45(2): 147–170.

Ramoni, M. ve Sebastiani, P. (2001). Robust Learning With Missing Data. *Journal of Machine Learning*. 45(2): 147–170.

Redman, T. C. (2001). *Data Quality: The Field Guide*. Boston: Digital Press.

Redman, T. C. (1996). *Data Quality For The Information Age*. Boston: Artech House.

Redman, T. C. (2008). *Data Driven: Profiting From Your Most Important Business Asset*. Massachusetts: Harvard Business Press.

Rockwell, D. (2012): *5 Tips for Cleaning Your Dirty Data*. https://tdwi.org/articles/2012/05/22/5-tips-cleaning-data.aspx. (Erişim Tarihi: 21.02.2020).

Rubin, D. (1978). Bayesian Inference For Causal Effects: The Role of Randomization. *The Annals of Statistics*. 6(1): 34-58.

Rubin, D. (1978). Multiple Imputations In Sample Surveys A Phenomenological Bayesian Approach To Nonresponse. *Proceedings Of The Survey Research Methods Section Of The American Statistical Association*. 1: 20–34.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*. 63(3): 581-592.

Satatistics Canada. (2003). *Statistics Canada Quality Guidelines*. Ottawa: Minister of Industry.

Sattler, K. (2009) *Data Quality Dimensions. Encyclopedia Of Database Systems*. Boston: Springer.

Saygili, A., Albayrak, S. (2017). A New Computer-Based Approach For Fully Automated Segmentation Of Knee Meniscus From Magnetic Resonance Image. *Biocybernetics And Biomedical Engineering*. 37: 432-442.

Scarisbrick-Hauser, A. ve Rouse, C. (2007). The Whole Truth And Nothing But The Truth? The Role of Data Quality Today. *Direct Marketing An International Journal*. 1(3): 161-171

Schafer, J. (2000). Dealing With Missing Data. *Research Letters in the Information and Mathematical Sciences.* 3: 153–160.

Schafer, J. L. (1999). Multiple Imputation: A primer. *Statistical Methods in Medical Research*. 8(1): 3–15.

Schmid, J. (2004). *The Main Steps To Data Quality*. Berlin: Springer

Schouten, Rianne., Lugtig, Peter. ve Vink, Gerko. (2018). Generating Missing Values For Simulation Purposes: A Multivariate Amputation Procedure. *Journal of Statistical Computation and Simulation*. 1-22.

Seyidoğlu, H. (1999). *Ekonomik Terimler Ansiklopedik Sözlük*. İstanbul: Güzem Can Yayınları.

Shankaranarayanan, G. ve Yu Cai, Y. (2006). Supporting Data Quality Management In Decision-making. *Decision Support Systems*. 42: 302-304.

Smallwood, R.F. (2014). *Information Governance: Concepts, Strategies, and Best Practices*. New Jersey: John Wiley And Sons.

Statistics Netherlands. (2008). *Quality Declarations of Statistics Netherlands*. http://www.cbs.nl/en-GB/menu/organisatie/kwaliteitsverklaring/default.htm (Erişim Tarihi: 16.12.2019).

Statistics Sweden. (2000). *Assessment of Quality in the Swedish CVTS2*. http://www.scb.se/statistik/UF/UF0502/_dokument/TekniskCVTS.pdf, (Erişim Tarihi: 16.12.2019).

Stekhoven, D. J . (2013). *Nonparametric Missing Value Imputation Using Random Forest*. https://rdrr.io/cran/missForest/man/missForest.html. (25.04.2020).

Sturge, M. ve Horsley, V. (1911). On Some Of The Biological And Statistical Errors In The Work On Parental Alcoholism By Miss Elderton And Professor Karl Pearson. *F.R.S. The British Medical Journal*. 1(2611): 72-82.

Sundström, H. J. (2001). New Tools For Strategic Quality Improvement – Statistics Finland's Experience. *International Conference On Quality in Official Statistics*. 28(2): 1-5.

Şeker, Ş. E. ve Eşmekaya, E. (2017). Eksik Verilerin Tamamlanması. *YBS Ansiklopedi*. 4(3): 10-17.

Şencan, H. (2005). *Sosyal ve Davranışsal Ölçümlerde Güvenilirlik ve Geçerlilik.* Ankara: Seçkin Kitabevi.

Tremblay, M. C., Dutta, K., ve Vandermeer, D. (2010). Using Data Mining Techniques To Discover Bias Patterns In Missing Data. *Journal Of Data And Information Quality*. 2(1): 1-19.

Twala, B. (2009). An Empirical Comparison Of Techniques For Handling Incomplete Data Using Decision Trees. *Applied Artificial Intelligence*. 23(5): 373–405.

Twala, B. ve Phorah, M. (2010). Predicting Incomplete Gene Microarray Data With The Use Of Supervised Learning Algorithms. *Pattern Recognition Letters*. 31: 2061–2069.

Twala, B. ve Cartwright,M. (2010). Ensemble Missing Data Techniques For Software Effort Prediction. *Intelligent Data Analysis*. 14(3): 299–331.

Twala, B., Jones, M. C. ve Hand, D. J. (2008). Good Methods For Coping With Missing Data In Decision Trees. *Pattern Recognition Letters*. 29(7): 950–956.

Vach, W. (1994). *Logistic Regression with Missing Values in the Covariates*. Switzerland: Springer.

Wand, Y. ve Wang, R. Y. (1996). Anchoring Data Quality Dimensions In Ontological Foundations. *Communications of the ACM*. 39(11): 86–95.

Wang, R. Y. ve Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means To Data Consumers. *Journal of Management Information Systems*. 12(4): 5-34.

Wilkinson, G. N. (1957). The Analysis Of Covariance With Incomplete Data. *Biometrics*. 13(3): 363- 372.

Wilkinson, G. N. (1958). The Analysis of Variance and Derivation of Standard Errors for Incomplete Data. *Biometrics*. 14(3): 360- 384.

Wilkinson, L. ve Friendly, M. (2009). The History of the Cluster Heat Map. *The American Statistician*. 63(2): 179–184.

Wilks, S. S. (1932). Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples. *The Annals of Mathematical Statistics. Institute of Mathematical Statistics*. 3(3): 163-195

Yamak, O. (2001). *Üretim Yönetimi*. İstanbul: Rema Matbaacılık.

Yates, F. (1933). The Analysis Of Replicated Experiments When The Field Results Are Incomplete. *Empire Jour Exp Agric*. 1(2): 129–142.

Yates, F. (1936). Incomplete Randomized Blocks. *Annals Of Eugenics*. 7(2): 121–140.

Yoon, J., Jordon, J. ve Van Der Schaar, M. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. *Proceedings of the 35th International Conference on Machine Learning*. 80: 5689-5698.