International Journal of Contemporary Tourism Research

http://dergipark.gov.tr/ijctr

ŞARAP ÜRETİMİNDE VERİ KALİTESİNE İLİŞKİN EKSİK VERİ SORUNLARININ DERİN ÖĞRENME İLE ÇÖZÜLMESİ: ÜRETİCİ ÇEKİŞMECİ AĞLARLA BİR UYGULAMA

Araştırma Makalesi

Şevhat DOGER[1], Osman Avşar KURGUN[2]

ÖZET

Araştırmanın amacı şarap üretiminde veri kalitesini etkileyen eksik veri problemlerini çözmek için uygun yöntemin seçilmesi ve şarap üreten işletmeler için eksik veri problemleri karşısında başvurabilecekleri bir 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 aşamalar analiz edilmiştir. Çalışmada eksik veri tamamlama görevi için üretici modeller sınıfına giren Üretici Çekişmeli Ağlar (GAN-Generative Adversial Networks) algoritmasının geliştirilmiş versiyonu Wasserstein Üretici Çekişmeli Atama Ağları (WGAIN-Wasserstein Generative Adversial Imputation Networks) kullanımı önerilmiştir. Bu yeni mimari, GAN'larda sıklıkla görülen problemlere karşı geliştirilmiş maliyet fonksiyonunun değiştirilmesi fikriyle oluşturulmuş ve atama probleminin benzersiz özellikleri ile başa çıkabileceği şekilde genelleştirilmiştir. Gerçek dünya veri kümesiyle yapılan deneyde, WGAIN için elde edilen hata karelerinin kök ortalaması (RMSE-Root Mean Square Error) değerleri ile diğer atama tekniklerinden önemli ölçüde daha iyi performans gösterdiği tespit edilmiştir.
Anahtar Kelimeler: Veri Kalitesi, Eksik Veri, Wasserstein Üretici Çekişmeli Atama Ağları, Şarap Kalitesi.
JEL Sınıflama Kodları: C81, L15, L66

SOLVING MISSING DATA PROBLEMS RELATED TO DATA QUALITY IN WINE PRODUCTION BY DEEP LEARNING: AN APPLICATION WITH GENERATIVE ADVERSIAL NETWORKS

Research Article

ABSTRACT

The aim of the study is to select the appropriate method to solve the missing data problems affecting the data quality in wine production and to create a guide for wine producing businesses to refer to in the face of missing data

[1] Bilim Uzmanı, Dokuz Eylül Üniversitesi, Sosyal Bilimler Enstitüsü, Toplam Kalite Yönetimi Anabilim Dalı, Kalite Yönetimi Programı, sevhatdoger@gmail.com, orcid.org/0000-0001-9174-159X
[2] Prof. Dr., Dokuz Eylül Üniversitesi, Turizm Fakültesi, Turizm İşletmeciliği Bölümü, avsar.kurgun@deu.edu.tr, orcid.org/0000-0002-2092-5292
Not: Bu makale Şevhat Doger'in Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Toplam Kalite Yönetimi Anabilim Dalı Kalite Yönetimi Programında yazdığı "Veri Kalitesinde Eksik Veri Sorunlarının Derin Öğrenme Yöntemi İle Çözülmesi: Üretici Çekişmeli Ağlar İle Bir Uygulama" isimli Yüksek Lisans Tezinden üretilmiştir.

"Doger, Ş. ve Kurgun, O. A. (2021). Şarap Üretiminde Veri Kalitesine İlişkin Eksik Veri Sorunlarının Derin Öğrenme ile Çözülmesi: Üretici Çekişmeci Ağlarla Bir Uygulama, *International Journal of Contemporary Tourism Research*, Vol 5: No: 1, p.99-111, doi: 10.30625/ijctr.943818"
Makale Gönderim Tarihi:28.05.2021 *Kabul Tarihi:13.06.2021*

problems. For this purpose, an incomplete data problem was created on the data set used in classifying the wines in terms of quality, in a way that disrupts the integrity, and the necessary steps for the solution of the problem were analyzed. In the study, the use of Wasserstein Generative Adversial Networks (WGAIN), an improved version of the Generator Adversial Networks (GAN) algorithm, is proposed for the missing data completion task. This new architecture was created with the idea of changing the cost function developed against the problems common in GANs and generalized so that it can cope with the unique features of the assignment problem. In the experiment performed with the real world dataset, it was determined that the values of the Root Mean Square Error (RMSE) obtained for WGAIN performed significantly better than the other imputation techniques.
Keywords: Data Quality, Missing Data, Wasserstein Generative Adversial Imputation Networks, Wine Quality.
JEL Classification Code: C81, L15, L66

GİRİŞ

Veri kalitesi, belirli bir veri kümesinin ne kadar güvenilir olduğunu göstermektedir. İşletmeler yanlış, eksik ya da güncel olmayan verilerle karar aldıklarında, tüketicilerinin tercihlerini yansıtmayan strateji ve politikalarla yapılandırabilmektedirler. İşletmelerin veriye dayalı stratejilerinin hedefine ulaşması kullanılan verilerin kalitesine bağlıdır. Ancak işletmelerde kullanılan verilerin kalitesinin düşük ya da verilerin eksik olması önemli bir sorundur.

Şarap kalitesini değerlendirmek için fizikokimyasal ve duyusal testler kullanılmaktadır. Şarapların sınıflandırılması, sürecin karmaşıklığı ve heterojenliği nedeniyle zor bir süreçtir. Şarapların sınıflandırılması ekonomik değeri belirlemek, kaliteyi korumak, karıştırılmaları engellemek ve işlemeyi kontrol etmek açısından hayati önem taşımaktadır. Şarap kalitesinin korunmasında bir uzmanın asitlik ve alkol bileşimi gibi bir dizi fizikokimyasal özellik kullanarak bir şarap örneğine vereceği derecelendirmeyi tahmin edebilmesi kritik değerdedir. Derecelendirmenin tahmin edilmesi şarapların sınıflandırılmasına yardımcı olmaktadır.

Veri setindeki bileşen değerleri şarap üretimi yapmak için gerekli bilgileri barındırmaktadır. Şarap uzmanı tarafından yedi puan ile takdir edilen bir şarabın bileşen değerleri bilinmediğinde aynı üretimi sürdürmek mümkün değildir. Kalite puanı üç olan şarap için bileşen değerleri bilinmediğinde düşük kalite şarap üretimi devam ettirilecektir. Şarap kalitesini etkileyen bileşenlerden pH, alkol ve yoğunluk değerlerinin bilinmemesi işletmenin kalite üzerinde kontrolünün olmadığı bir süreçte 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.

KAVRAMSAL ÇERÇEVE

Veri Kalitesi

Veri (İngilizce ve Latince datum; data), ham (işlem yapılmamış) yani gerçek enformasyon parçasıdır. Ölçüm ya da sayım yöntemiyle elde edilen ve sayısal bir değer ifade eden veriler nicel, sayısal bir değer ifade etmeyen veriler nitel veriler olarak isimlendirilmektedir. Sembolik her gösterim gibi aslında veri de belirli bir nesne, birey ya da olguya yönelik bir soyutlama olarak ifade edilmektedir (Bosij, Chafey, Greasley ve Hickie, 2003: 4). Bir araya getirilmiş ve düzenlenmiş verilere enformasyon, enformasyonun hacim olarak küçük ancak kullanım değeri çok yükselmiş hali bilgi olarak ifade edilmektedir (Gürsakal, 2007: 9).

İstatistik veri kavramını, yorumlanma ve sunulma amacıyla toplanmış, çözümlenmiş ve özetlenmiş gerçekler olarak tanımlamaktadır. Nicel veriler kesikli ve sürekli olmak üzere iki türden oluşmaktadır. Kesikli veriler sayım yöntemiyle toplanır. Çoğunlukla sayma sayıları türünden ifade edilirler. Sürekli veriler ölçüm yoluyla toplanır. Gerçek (reel) sayılar türünden ifade edilirler. Nitel veriler ordinal ve nominal olmak üzere iki türden oluşmaktadır (Anderson, Sweeney ve Williams, 2011: 5).

Hatalı veriler nedeniyle alınan kararlar sadece yanıltıcı değil, aynı zamanda son derece maliyetlidir. Gartner'a göre, düşük veri kalitesinin kuruluşlar üzerindeki ortalama finansal etkisi yıllık 9,7 milyar dolardır. IBM, ABD'de işletmelerin, düşük veri kalitesi nedeniyle yılda 3,1 trilyon dolar kaybettiğini açıklamıştır (Gartner, 2018:1; IBM, 2019: 1). Veri kalitesi, temiz verilerin işletme için kullanılabilir ve yönetilebilir hale getirilmesi ile sağlanabilir. Yüksek kaliteli veriler veri entegrasyonunun gerçekleştirilmesine de olanak sağlamaktadır. Ayrıca veri kalitesi, kaliteli ve güvenilir kararların da anahtarı konumundadır.

Veri kalitesi kavramı; yönetim-bilişim sistemleri ve istatistik gibi branşların öncü olduğu çeşitli alanlarda kullanılmıştır. 1960'lı yılların sonlarına doğru veri kalitesiyle ilişkili problemlerin çözümüne yönelik ilk araştırmaları istatistikçiler gerçekleştirmişlerdir. 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: 5).

Veri kalitesi kavramı, giderek daha büyük elektronik veri setlerinin oluşması ve veriye dayanan 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, nitel ya da nicel bilgi parçalarının durumunu ifade etmektedir. Veriler operasyonlarda, karar verme ve planlamada amaçlanan kullanımlarına uygunsa yüksek kaliteli olarak kabul edilmektedir (Redman, 2008: 16; Fadahunsi vd., 2019: 1).

Veri Kalitesi Boyutları

Veri kalitesi boyutları farklı yönlerden ele alınmaktadırlar. Çeşitli kuruluşlar veri kalitesini ölçmek ve iyileştirmek için farklı veri kalitesi boyutları tanımlamıştır.

Eurostat'ın 2000 yılında yayınladığı Standart Kalite Rapor'unda yedi veri kalitesi boyutu tanımlanmıştır. Bunlar; uygunluk (relevance), doğruluk (accuracy), güncellik ve dakiklik (timeliness and punctuality), ulaşılabilirlik ve açıklık (accessibility and clarity), karşılaştırılabilirlik (comparability), tutarlılık (consistency) ve bütünlüktür (completeness) (Eurostat, 2000: 2). Hollanda İstatistik Kurumu ise veri kalitesi boyutlarını şu biçimde belirlemiştir; maliyet etkinliği (cost effectiveness), uygunluk (relevance), doğruluk ve güvenilirlik (accuracy and reliability), güncellik ve dakiklik (timeliness and punctuality), tutarlılık (consistency) ve karşılaştırılabilirlik (coherence and comparability) ve ulaşılabilirlik ve açıklık (accessibility and clarity) (Statistics Netherlands, 2008: 3). Wand ve Wang (1996: 88) veri kalitesi boyutlarını; doğruluk (accuracy), bütünlük (completeness), tutarlılık (consistency), güncellik (timeliness) ve güvenilirlik (reliability) olarak ifade etmişlerdir.

Çalışmada veri kalitesinin değerlendirilmesinde kullanılabilecek altı boyut seçilmiştir. Bu altı kalite boyutunun seçilmesinin nedeni yaklaşımların çoğunda ortak olması ve işletmelerde verilerin kalitesini belirlenmesinde sıklıkla kullanılan boyutlar olmasıdır. Araştırma için seçilen veri setine uygulanabiliyor olması boyutların seçilme nedenlerinden bir diğeridir. Söz konusu boyutlar şu biçimde açıklanabilir;

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 (Şencan, 2005: 725).

b. Güvenilirlik: Güvenilirlik veri değerinin geçerli olup olmadığına bakılmaksızın benzer değere dönüşme süreci şeklinde tanımlanmaktadır. Bir veri kaynağının güvenilir olmasına rağmen geçerli olmaması mümkün olabilir (Scarisbrick-Hauser ve Rouse, 2007: 164).

c. Tutarlılık: Bir veri kümesindeki veri değerlerinin, başka bir veri kümesindeki değerlerle tutarlı olmasını ifade etmektedir. Tutarlılık kuralını doğrulayan bir örnek, kurumsal bir hiyerarşi yapısı içinde, her müşteri temsilcisine atanan müşteri sayısının, şirketin tamamının sahip olduğu müşteri sayısını geçmemesi gerektiği şeklinde verilebilir (Loshin, 2006: 9). Tutarlılık, bir sistemde yönetilen verilerin belirtilen kısıtlamaları ya da iş kurallarını ne ölçüde karşıladığı olarak da tanımlanmaktadır (Sattler, 2009: 12).

d. Doğruluk: Doğruluk kavramsal olarak; istatistiksel tahminlerin, tahminledikleri kavramsal değerlere yakınlığı olarak tanımlanmaktadır (Brackstone, 2001: 16). Doğruluk, verilerin ne ölçüde doğru, güvenilir ve hatasız olarak sertifikalandırıldığı olarak da tanımlanmaktadır (Sattler, 2009: 46).

e. Güncellik: Bir veri çıktısının zamanında ulaşılabilir olup olmadığı bilgisine sahip olma anlamına gelmektedir (Wand ve Wang, 1996: 93). Güncellik, üzerinde çalışan görev için verinin elde edilebilirliğidir. Güncellik boyutunun, belli bir kullanım alanı için geç kalındığı için faydasız olan mevcut verilere sahip olunabilineceği gerçeğini

yansıttığı ifade edilebilir (Batini ve Scannapieca, 2016: 29).

f. Bütünlük: Verinin üzerinde çalışılan görev için yeterli genişliğe, derinliğe ve kapsama sahip olma derecesi olarak tanımlanabilir (Batini ve Scannepica, 2016: 28). İstatistik alanında bütünlük boyutu her bir veri kaydının tam (eksik veri olmadan) ya da bütün olması anlamına gelmektedir (Karr vd., 2006: 157).

Eksik Veri Kavramı ve Türleri

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 (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). Bu nedenle işletmelerde eksik veri sorununun çözümü katma değer içeren fırsatların ortaya çıkarılmasında rol oynanaktadır. Söz konusu rolün kritik önemi nedeniyle bu alanda araştırmacılar tarafından birçok bilimsel çalışma gerçekleştirilmiş ve çeşitli yaklaşımlar yapılandırılmıştır.

Literatürde en çok kullanılan eksik veri tanımlama sistemi Rubin (1976: 582) tarafından geliştirilmiştir. Rubin; eksik veri türlerini tamamen rastgele eksik veri (MCAR-Missing Completely At Random), rastgele eksik veri (MAR-Missing At Random) ve rassal olmayan eksik veri (MNAR-Missing Not At Random) olmak üzere üç grupta incelemiştir.

MCAR varsayımı, bir değişkende eksik veri bulunma olasılığının, bu değişkenin kendi değeriyle ya da veri setindeki diğer herhangi bir değişkenin değeriyle ilişkili olmadığında ortaya çıkan eksiklik türüdür. MAR varsayımı, bir değişkende eksik veri bulunma olasılığının, analizdeki diğer değişkenler kontrol altına alındığında, bu değişkenin kendi değeri ile ilişkisiz olduğunda ortaya çıkan eksiklik türüdür. 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 türüdür (Little ve Rubin, 2014: 21). Bu türdeki eksik verilerin bulunduğu verilerle yapılan model geliştirme tahminleri genellikle hatalı çalışır. Bu mekanizmanın olasılıklı bir matematiksel formu kolay değildir, çünkü bu türdeki veriler MAR ya da MCAR değildir (Leke ve Marwala, 2019: 18).

Little ve Rubin (1987) yayınladıkları kitapta eksik veri nedeniyle meydana gelebilecek durumlara örnekler yardımı ile açıklık getirmeye çalışmışlardır. Bu çalışma eksik veri alanının ilk sistematik çalışması olarak önem taşımaktadır. Eksik veri analizi bu araştırmadan sonra daha yaygın bir hale gelmiştir.

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

Klasik Eksik Veri Çözümleme Yöntemleri

Bir veri setinde verinin eksiklik durumuna bağlı olarak, istatiksel yazılım paketlerinde kullanılan çeşitli veri atama teknikleri bulunmaktadır.

a. 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. Bu yaklaşım veri setindeki eksik veri oranının az olduğu durumlarda uygulanabilir (Leke ve Marwala, 2019: 7).

b. Çiftler Bazında Silme: 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 (Enders, 2010: 41-42).

c. Aritmetik Ortalama-Mod ile 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).

d. 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).

d.a. Tekli Atama Yöntemleri: Tekli atama yöntemlerinde veri kümelerindeki eksik değerler bilinen başka gözlem değerleriyle değiştirilmektedir. Beklenti Maksimizasyon-Beklenti En Büyükleme (EM-Expectation-maximization) yöntemi, 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 logaritmik olasılık fonksiyonu oluşturur. M adımı parametre değerlerini logaritmik olasılık fonksiyonunu 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, Laird ve Rubin, 1977: 1). Sıcak-Soğuk Deste Atama yöntemi ise eksik verileri doldurmanın diğer bir yoludur ve rastgele seçilen değerler kullanmaktır. Atama gözlenen veriden rastgele bir değerin seçilmesiyle sıcak deste (hot deck) ya da benzer değişkenleri içeren başka bir veri setinden seçilmesiyle soğuk deste (cold deck) ile gerçekleştirilebilir (Çilingirtürk ve Aktaş, 2010: 76-77). 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 (Leke ve Marwala, 2019: 9).

d.b. Regresyon Yöntemi: 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).

d.c. Çoklu Atama Yöntemi (MI-Multiple Imputation): Çoklu atamada 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).

Eksik Veri Çözümleme Yöntemleri

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 zekâ ve istatistiksel yöntemlerin uygulanmasıyla karakterize edilen yaklaşımlara geçme gibi temel yaklaşımları içerir. Çalışmada eksik veri çözümleme yöntemleri olarak makine öğrenmesi ve derin öğrenme modelleri ele alınmıştır.

a. Makine Öğrenmesi Yaklaşımıyla

Makine öğrenimi kapsamında yer alan atama yöntemleri, eksik verilerinin yerine kullanılacak değeri tahminlemek için model oluşturmayı kapsayan karmaşık yöntemlerdir. Makine öğrenimi modelleri veri kümesindeki var olan bilgileri temel alarak eksik veriyi tahminlemek amacıyla bir model üretmektedirler. Bu yöntemler ağaç (karar ağaçları, rastgele orman) ya da biyolojik temelli (yapay sinir ağları) olabilirler.

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. 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 girişlerine karşılık gelen bir ağaç ile değiştirilir (Twala ve Cartwright 2010: 300).

Yapay sinir ağları, verileri biyolojik sinir sisteminin yaptığı gibi işlemek için olasılıklı bir model kullanmaktadır (Abdella ve Marwala, 2005a: 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). Abdella ve Marwala (2005b: 577), yapay sinir ağları modelini kullanarak eksik verileri hesaplamıştır. Yapay sinir ağları modeli veri setinin özel olarak modellemesinde yani giriş katman değeri ile çıkış katman değeri aynı olacak şekilde oluşturulmuştur.

b. Derin Üretici Model Yaklaşımıyla

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 (Bengio, LeCun ve Hinton, 2015: 436). 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, Donahue ve Simonyan, 2018: 2). Eksik veri atamalarında iki ana tip derin üretici model kullanılmaktadır: Varyasyonel Oto-Kodlayıcılar (VAE-Variational Autoencoder) ve Üretici Çekişmeli Ağlar (GAN-Generative Adversial Networks).

VAE, etkin veri kodlamalarını denetimsiz bir şekilde öğrenmek için kullanılan bir tür yapay sinir ağıdır (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 (Ivanov, Figurnov ve Vetrov, 2019: 3). GAN 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 vd., 2014: 2672).

Ayırıcının örnekleri gerçekten sınıflandırmadığı GAN şemasının değiştirilmesiyle Wasserstein Üretici Çekişmeli Ağ (Wasserstein GAN-WGAN) mimarisi 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. Ayrıcının eğitimi, sahte örneklerin çıktılarını büyütmek yerine gerçek örneklerin çıktılarını (sınıflandırmalarını) büyütmeye çalışmaktadır (Arjovsky, Chintala ve Bottou, 2017: 21). GAIN modeli, GAN mimarisindeki üretici ve çekişmeli ağları kullanmaktadır. 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 modelinde üretici; veri matrisindeki eksik gözlemlerin yerine atama işlemini, eksik olmayan gözlemlerden oluşan veri matrisinin şartlı olasılık dağılımına göre yapmaktadır. Ayırıcı, daha sonra tamamlanmış veri matrisini alır ve hangi değerlerin gerçek hangi değerlerin atanmış değerler olduğunu belirlemeye çalışır (Yoon, Jordon ve Van Der Schaar, 2018: 5689). GAIN modeli temel alınarak, Wasserstein GAN mimarisinin özelleştirilmiş türü eksik veri atama yönteminde kullanılmıştır. Bu yöntem WGAIN olarak isimlendirilmiştir.

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

Uygulamanın Amacı ve Önemi

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 ve maliyet fonksiyonu değiştirilerek oluşturulan WGAIN algoritması ile eksik değerlerin tamamlanması daha sonra bu algoritmaların performansının diğer eksik veri çözümleme yöntemlerinden Beklenti Maksimizasyonu (EM-Expectation Maximization), Çoklu Atama (MI- Multiple Imputation) ve Karar Ağaçları (DT-Decision Trees) algoritmalarıyla kıyaslayarak işletmenin sahip olduğu veri setinin veri kalitesini hangi yöntem seçilirse iyileştirilebileceğini göstermeye çalışmaktır. Uygulamanın önemi ise söz konusu işletmelerin şarap kalitesini artırmada yararlanabileceği kaliteli veri setinin elde edilmesinde GAN tabanlı algoritmaların kullanılması ve şarap üreten işletmelere eksik veri problemlerini çözmelerine yönelik olarak bir rehber sunulmasıdır.

Uygulamanın Kapsamı

Yapılan uygulamanın kapsamı; Portekiz "Vinho Verde" şarabının kırmızı çeşidiyle ile sınırlandırılmıştır. 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 bulunmamaktadır) (Cortez, Cerdeira, Almeida, Matos ve Reis 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, makine öğrenmesi algoritmaları başlığı altında tanıtılan Yapay Sinir Ağları (YSA-Artificial Neural Network) kapsam dışı bırakılmıştır. YSA ile yapılan eksik veri çözümleme çalışmaları veri setinin özel olarak modellemesinde yani giriş katman değeri ile çıkış katman değeri aynı olacak şekilde oluşturulması işlemlerinden oluşmaktadır. 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.

Veri Seti

Veri seti, UCI Makine Öğrenimi Havuzu'ndan (UCI-University of California Irvine Machine Learning Repository-Kaliforniya Üniversitesi Irvin Makine Öğrenmesi Havuzu) alınmıştır (Cortez vd., 2009: 547). 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 ve MAR) sırasıyla %10, %20, %30, %40, %50 oranlarında eksik değerler oluşturulmuştur.

Veri Setinde Eksik Verilerin Rassallığının Test Edilmesi

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). 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 1 ve Tablo 2'de gösterilmiştir. Tablo 1 %10 oranında eksikliğe sahip şarap kalitesini sınıflandırmak için kullanılan veri setine uygulanan Little'ın MCAR testi sonuçlarını göstermektedir.

Tablo 1: Little'ın MCAR Testi Bilgilerini İçeren Çıktı Tablo. 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 1'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. Tablo 2 %10 oranında eksikliğe sahip şarap kalitesini sınıflandırmak için kullanılan veri setine uygulanan Little' ın MCAR testi sonuçlarını göstermektedir.

Tablo 2: Little'ın MCAR Testi Bilgilerini İçeren Çıktı Tablo. 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. = ,001**										

Tablo 2'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.=,001<0,05'ten küçük olduğu için eksik verilerin oluşturduğu mekanizma MCAR değildir. Yani sıfır hipotezi (H_o) rededilir. MAR türünde eksiklik mekanizması vardır.

Değerlendirme Yöntemi

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'dir (Root Mean Square Error-Hata Karelerinin Kök Ortalaması).

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 GAIN algoritmasının performansı RMSE kullanılarak test edilmiştir (Yoon ve diğerleri, 2018: 5696).

Kullanılan Yazılımlar 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.

Eksik Veri Çözümleme Yöntemlerinin Karşılaştırılması

Eksik veri tamamlama yöntemleri MCAR ve MAR türündeki eksik veri mekanizmasında ve %10, %20, %30, %40 ve %50 eksiklik oranlarındaki veri seti üzerinde tekrarlı olarak denenmiştir. Toplamda beş deneme yapılmış olup ortalama RMSE değerleri MCAR eksiklik mekanizması için Tablo 3'de, MAR eksiklik mekanizması için Tablo 4'de gösterilmiştir. Düşük RMSE değeri daha iyi performans göstergesidir ve bunu belirtmek için kalın yazı tipinde gösterilmiştir.

Tablo 3: 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 4: 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 3 ve 4 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.

TARTIŞMA VE SONUÇ

Veri kalitesi, belirli bir veri kümesinin ne kadar güvenilir olduğunu göstermektedir. İşletmeler yanlış, eksik ya da güncel olmayan veriler üzerine kararlar alabilmektedirler. Bu kararlar sonucunda tüketicilerinin tercihlerini yansıtmayan stratejiler ya da politikalar yürütme riskiyle karşı karşıya kalabilmektedirler. İşletmelerin veriye dayalı stratejilerinin hedefine ulaşması kullanılan verilerin kalitesine bağlıdır. Ancak, işletmelerde

kullanılan verilerin kalitesinin düşük ve verilerin eksik olması önemli bir sorundur.

Bu çalışmada 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 gerçekleştirilmesi gereken tüm aşamalar incelenmiştir. Şarap kalitesini değerlendirmek için fizikokimyasal ve duyusal testler kullanılmaktadır. Şarapların sınıflandırılması kolay bir süreç değildir. Şarapların sınıflandırılması; şarap ürünlerinin ekonomik değerini belirlemek, şarapların kalitesini korumak, şarapların karıştırılmasını engellemek ve işlem süreçlerini kontrol etmek açısından büyük önem taşımaktadır. 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 yedi puan alan bir şarabın bileşen değerleri bilinmediğinde aynı üretimi yapmak mümkün olmayacaktır. Kalite puanı üç 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, GAIN ve geliştirilmiş versiyonu WGAIN 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.

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 genelleştirilmiştir. 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 farklı sektörde faaliyet gösteren işletmelerde dönüştürücü etkilere sahip olabilir. Gelecekteki çalışmalarda, WGAIN'in performansının daha büyük veri setlerinde ve farklı

mimariler kullanılarak araştırılması önemli katkılar sağlayacaktır.

KAYNAKÇA

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

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

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

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

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. (2016). Data And Information Quality: Dimensions, Principles And Techniques. Switzerland: Springer International Publishing.

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

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

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

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.

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

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

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

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

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.

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.

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

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.

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/, Erişim Tarihi: 20.06.2020.

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, J. W. (2012). Missing Data: Analysis and Design. Germany: Springer.

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

IBM. (2019). *Big Data And Analytics Hub*. https://www.ibmbigdatahub.com/infographic/

extracting-business-value-4-vs-big-data, Erişim Tarihi: 20.06.2020.

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

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.*

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. ve Rubin, D. (2020) 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. J. A. ve Rubin, D. B. (1987). Statistical Analysis With Missing Data. New York: Wiley.

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

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

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

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

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

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.

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

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.

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

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

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

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.

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

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

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.

Etik Onay

Bu çalışma, katılımcılardan birebir veri toplamayı gerektiren araştırma kapsamına girmediği ve veriler ikincil veri olarak elde edildiği için etik kurul onayı gerektirmeyen çalışmalar arasında yer almaktadır.

Araştırmacıların Katkı Oranı

Yazarlar çalışmaya eşit katkıda bulunmuştur.

Çıkar Çatışması

Bu çalışmada potansiyel bir çıkar çatışması yoktur.