MAKİNE ÖĞRENMESİ YÖNTEMLERİNE DAYALI VERİ YÖNETİM SİSTEMİ

Ülgen AYDIN

Danışman: Doç. Dr. Gökay AKKAYA

Yüksek Lisans Tezi
Endüstri Mühendisliği Ana Bilim Dalı
2024

T.C.
ATATÜRK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANA BİLİM DALI

MAKİNE ÖĞRENMESİ YÖNTEMLERİNE DAYALI VERİ YÖNETİM SİSTEMİ

(Data Management System Based On Machine Learning Methods)

YÜKSEK LİSANS TEZİ

Ülgen AYDIN

Danışman: Doç. Dr. Gökay AKKAYA

Erzurum

Şubat, 2024

KABUL VE ONAY TUTANAĞI

Ülgen AYDIN tarafından hazırlanan "Makine Öğrenmesi Yöntemlerine Dayalı Veri Yönetim Sistemleri" başlıklı çalışması 5/ 2 / 2024 tarihinde yapılan tez savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Endüstri Mühendisliği Ana Bilim Dalı, Endüstri Mühendisliği Bilim Dalında yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	Doç. Dr. Mustafa Yılmaz *Atatürk Üniversitesi*	Aslı Islak İmzalıdır
Danışman:	Doç. Dr. Gökay AKKAYA *Atatürk Üniversitesi*	Aslı Islak İmzalıdır
Jüri Üyesi:	Dr. Öğr. Üyesi Şeyma Emeç *Erzurum Teknik Üniversitesi*	Aslı Islak İmzalıdır

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddelerinde belirtilen şartları yerine getirdiğini onaylarım.

Prof. Dr. Saltuk Buğrahan CEYHUN
Enstitü Müdürü
Aslı Islak İmzalıdır

Bu çalışma BAP projeleri kapsamında desteklenmiştir.
Proje No: FYL-2023-11813

ETİK BİLDİRİM VE İNTİHAL BEYAN FORMU

Yüksek Lisans Tezi olarak *Doç. Dr. Gökay AKKAYA* danışmanlığında sunulan "Makine Öğrenmesi Yöntemlerine Dayalı Veri Yönetim Sistemleri" başlıklı çalışmanın tarafımızdan bilimsel etik ilkelere uyularak yazıldığını, yararlanılan eserlerin kaynakçada gösterildiğini, Fen Bilimleri Enstitüsü tarafından belirlenmiş olan Turnitin Programı benzerlik oranlarının aşılmadığını ve aşağıdaki oranlarda olduğunu beyan ederiz.

Tez Bölümleri	**Tezin Benzerlik Oranı (%)**	**Maksimum Oran (%)**
Giriş	*4*	**30**
Kuramsal Temeller	*3*	**30**
Materyal ve Metot	*2*	**35**
Araştırma Bulguları ve Tartışma	*0*	**20**
Sonuçlar	*0*	**20**
Tezin Geneli	*4*	**25**

Not: *Yedi kelimeye kadar benzerlikler ile Başlık, Kaynakça, İçindekiler, Teşekkür, Dizin ve Ekler kısımları tarama dışı bırakılabilir. Yukarıdaki azami benzerlik oranları yanında tek bir kaynaktan olan benzerlik oranlarının* ***%5'den*** *büyük olmaması gerekir.*

Beyan edilen bilgilerin doğru olduğunu, aksi halde doğacak hukuki sorumlulukları kabul ve beyan ederiz.

Tez Yazarı (Öğrenci)	**Tez Danışmanı**
Ülgen AYDIN	Doç. Dr. Gökay AKKAYA
6.2.2024	**6.2.2024**
İmza: Aslı Islak İmzalıdır	İmza: Aslı Islak İmzalıdır

***** Tez ile ilgili YÖKTEZ'de yayınlamasına ilişkin bir engelleme var ise aşağıdaki alanı doldurunuz.

☒ Tezle ilgili patent başvurusu yapılması / patent alma sürecinin devam etmesi sebebiyle Enstitü Yönetim Kurulunun …./…/…. tarih ve ……. sayılı kararı ile teze erişim 2 (iki) yıl süreyle engellenmiştir.

☒ Enstitü Yönetim Kurulunun …./…/…. tarih ve ……. sayılı kararı ile teze erişim 6 (altı) ay süreyle engellenmiştir.

TEŞEKKÜR

Bu çalışmanın konusunun belirlenmesinde ve hazırlanma sürecinin her aşamasında değerli bilgilerini ve zamanını benden esirgemeyerek her fırsatta çalışmamla yakından ilgilenen, eleştirileriyle yol gösteren danışman hocam Doç. Dr. Gökay AKKAYA' ya teşekkür ve minnetimi özellikle belirtmek istiyorum.

Ayrıca Atatürk Üniversitesi Bilimsel Araştırma Projeleri (BAP) koordinasyon birimi tarafından desteklenen ve FYL-2023-11813 nolu proje kapsamında ihtiyaç duyulan bütçe tahsisini sağladıkları ve değerli katkılarından dolayı BAP koordinasyon birimine teşekkür ederim.

Ülgen AYDIN

ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ YÖNTEMLERİNE DAYALI VERİ YÖNETİM SİSTEMLERİ

Ülgen AYDIN

Danışman: Doç. Dr. Gökay AKKAYA

Amaç: Bu çalışmada eksik verilerin tahmini için makine öğrenmesi yöntemlerine dayalı eksik veri yönetim sistemi geliştirilmiştir. Veri setlerindeki eksik veri kavramını ele alıp eksik veri problemini silme, ortalama alma gibi yanlılığa neden olan klasik süreçlerden kurtararak alanlarında uzman olmasalar bile rahatlıkla kullanabilecekleri bir arayüz oluşturup, veri setlerindeki eksik nümerik verileri tamamlayan bir sistem geliştirmektir.

Yöntem: Bu çalışmada, belirtilen amaç doğrultusunda Python programlama dili kullanılarak eksik veri tamamlama sistemi geliştirilmiştir. Geliştirilen sistem masaüstü uygulaması haline dönüştürülmüştür. Eksik verilerin tahminini yapmak için yinelemeli atama ve tahminci olarak da rastgele orman algoritması kullanılmıştır. En iyi yineleme sayısını bulmak için çapraz doğrulama kullanılan sistemde, kullanıcı tarafından girilen eksik veri setinin raporlaması, eksik veri türleri ve ısı haritası olarak kullanıcıya sunulmuştur.

Bulgular: Geliştirilen algoritma 4 farklı veri setinde %5, %10 ve %15 olmak üzere 3 farklı eksiklik oranında tamamen rastgele eksik olarak azaltılmış ve test edilmiştir. Azaltılan nümerik veriler geliştirilen algoritma tarafından tamamlanmış ve sonuç olarak %57 ile %79 arasında değişen bir doğruluk oranıyla tahminler yapılmıştır.

Sonuç: Eksik veri problemine çözüm yaklaşımı sunan bu çalışmada amaçlanan hedefler doğrultusunda kullanıcılara kolay ve anlaşılabilir bir arayüz sunularak veri setlerindeki eksik nümerik verilerin tamamlanması sağlanmış, klasik olarak yapılan silme veya ortalama alma gibi işlemlerde meydana gelen yanlılık problemini aşmak için bir çözüm sunulmuştur. Ayrıca sözel verilerin tahmini ve kompakt bir sistem içinde literatüre farklı bir bakış kazandırılmıştır.

Anahtar Kelimeler: Makine öğrenmesi, Eksik veri, Yapay zekâ, Veri yönetim sistemi

Şubat 2024, 40 sayfa

ABSTRACT

MASTER'S THESIS

DATA MANAGEMENT SYSTHEM BASED ON MACHINE LEARNING METHODS

Ülgen AYDIN

Supervisor: Assoc. Prof. Dr. Gökay AKKAYA

Purpose: In this study, a missing data management system based on machine learning methods has been developed for predicting missing values. Addressing the concept of missing data in datasets and aiming to eliminate biases introduced by classical processes such as deletion or mean imputation, the goal is to create a user-friendly interface that can be easily utilized by individuals, even if they are not experts in the field. The system is designed to complete missing numerical data in datasets.

Method: In pursuit of the stated objective in this study, a missing data completion system has been developed using the Python programming language. The developed system has been transformed into a desktop application. For predicting missing data, the iterative imputer has been employed, with the random forest regressor algorithm serving as the predictor. Cross-validation has been utilized to determine the optimal number of iterations in the system. The system provides users with a report on the entered missing data set, including information on missing data types presented as a heatmap.

Findings: The developed algorithm was tested on four different datasets with three different missingness rates: 5%, 10%, and 15%, where the missing values were completely randomly reduced. The reduced numerical data were imputed by the developed algorithm, resulting in prediction accuracy ranging from 57% to 79%.

Results: In this study, which presents a solution approach to the missing data problem, the aim was to provide users with an easy and understandable interface to complete missing numerical data in datasets. A solution was offered to overcome the bias problem that arises in classical processes such as deletion or mean imputation. Additionally, the prediction of categorical data was addressed, contributing a novel perspective to the literature within a compact system.

Keywords: Artificial intelligent, Missing data, machine learning, Data management system

February 2024, 40 pages

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

KISALTMALAR VE SİMGELER DİZİNİ

DVM **:** Destek Vektör Makineleri

EM **:** Beklenti Maksimizasyonu

K-NN **:** K-En Yakın Komşu Algoritması

MAE **:** Ortalama Mutlak Hata

MAR **:** Rastgele Eksik

MCAR **:** Tamamen Rastgele Eksik

MNAR **:** Rastgele Olmayan Eksik

R^2 **:** Belirlilik Katsayısı

RMSE **:** Karekök Ortalama Hata

GİRİŞ

Veri bilgiyi oluşturan ana yapıdır. Verinin doğru analizi araştırmacıların ve işletmelerin bilgiyi elde etmek ve doğru kullanma sürecindeki en temel gerekliliğidir. Gerek araştırmacılar gerek de şirketler veriyi toplama, analiz etme, modelleme ve kontrol etme süreçlerinden geçirerek optimize etmeye çalışır. 2000'li yılların başlamasıyla internet, hayatımızın odak noktası haline gelmiş ve yapay zekâ kavramının hedeflerine adım adım ulaşmasıyla insanoğlunun oluşturduğu veri boyutları artmış ve veri analiz süreci hayati önem taşımaya başlamıştır. Araştırmacılar ve şirketler hem hizmet kalitelerini iyileştirmek için hem de verimlilik artışını sağlamak, karar vermek ve planlama yapmak için bilgi teknolojilerine daha bağımlı hale gelmiştir. Ancak hem veri toplama süreci zorlukları hem de veri analiz sürecinin büyük veri tabanları ile çalışırken daha da zorlaşmasıyla, bilgi teknolojileri tek başına ihtiyacı karşılayamaz hale gelmiştir. Çünkü artan teknoloji getirdiği kolaylıkların yanı sıra veri bilimi açısından bakacak olursak büyük veri yığınlarının artması sebebiyle analiz süreci zorlaşmış ve mevcut istatistiki yöntemler ihtiyacı karşılayamayacak hale gelmiştir. Rekabetin artık saniyeler boyutunda yaşandığı günümüzde büyük veri setlerinden anlamlı ve doğru bilgiyi elde etme süreci de zorlaşmıştır. Bu da hem şirketler hem de araştırmacılar için dezavantajlı bir durum haline gelmektedir. Çünkü veri kalitesini sağlamış işletme ya da kurumlar daha iyi rekabet edebilmektedirler (Osman Avşar KURGUN, nod.). Yukarıda bahsettiğimiz veri analiz sürecinin en önemli aşamalarından biri ise veri setlerindeki eksik veri varlığıdır. Bir veri setindeki verilerin eksik olması, karar alma süreçleri üzerinde olumsuz bir etkiye sahiptir. Eksik değerler, ölçülmesi ve kaydedilmesi gereken ancak çeşitli nedenlerden elde edilemeyen verilerdir. Eksik değerlerin nedenleri arasında sensör arızaları, iletişim hataları, güç kesintileri, sansürleme, örnekleme stratejileri, veri giriş hataları, veri kaybı, veri reddi, veri uyumsuzluğu vb. sayılabilir (Abdala & Marwala, n.d.). Eksik değerleri tamamlamanın ilk adımı, veri toplama sürecini iyileştirmeye ve uygun tamamlama yöntemini seçmeye yardımcı olmak için eksik verilerin oluşum nedenlerini anlamaktır. İlk olarak, eksik değerlerin tipleri, eksik değerin nedeninin diğer değişkenlerle ilişkili olup olmadığına bağlı olarak rastgeleliği göre tamamen rastgele eksik (MCAR), rastgele eksik (MAR), ve rastgele eksik olmayan (MCAR) olarak sınıflandırılabilir (Little & Rubin, 2019). MCAR, en yüksek derecede rastgelelik gösterir ve eksik verilerin değişkenler arasında rastgele dağıldığını ve oluşma olasılığının diğer değişkenlerle ilgili olmadığını gösterir. MAR, orta düzeyde rastgelelik temsil eder ve eksik veri oluşma olasılığının diğer değişkenlerle ilgili olmadığını gösterir (J. Kim *et al.* 2023a). Bu tez kapsamında veri nedir? Sorusu yanıtlanmış, eksik veri tanımı ve türleri hakkında bilgi verilmiş,

veri madenciliğinin temelleri tanıtılmış ve daha sonra veri setlerindeki eksik verileri tamamlama yöntemlerine dair bazı çalışmalara yer veren literatür taraması yapılıp, yapılan çalışmaya dair materyal ve metot tanıtılarak, tez kapsamında elde edilen sonuçlara yer verilmiştir.

Veri Nedir?

Veri, sözlük anlamı olarak bir araştırmada, bir tartışmada, bir akıl yürütmede sonuca ulaşabilmek için gereken ilk değer olarak sunulur. Veriler, ölçümler, sayımlar, deneyler, gözlemler veya çeşitli araştırma yollarıyla toplanan ve yapılan araştırma özelinde sonuca ulaştıran ve daha çok sayısal ve sözel formlardan oluşan mekanizmalardır. Veriler toplandıktan sonra yapılan çalışma özelinde gruplandırılır, derlenir ve işlenir. Böylelikle toplanmaktaki amacı olan bilgiyi vermesini bekleyerek araştırma, problem çözümü veya karar verme gibi belirli bir amaca hizmet edebilecek hale gelmektedir. Verinin önemini anlatmak için ortaya konulabilecek en önemli kriter, verinin sisteme kattığı değerdir (Korcan; ARSLANTEKİN, 2016). Bu değer, ister şirketlerin ister araştırmacıların kendi kriterlerine göre hedef haline getirdiği amaca ulaşmak için en önemli aşamadır. Bir çalışma için veri topladığımızda sayısal ve sözel ifadelerden oluşan belirli kategorik şartlara sahip yapılar olduğunu görürüz. Bu yapılar içinde eksik veriler, gürültülü veriler olup olmadığına bakılır. Bu değerleri anlamlı hale getirebilmek için de verileri ön işleme tabi tutma zorunluluğumuz bulunur.

Veri madenciliği

Tanım olarak veri madenciliği kavramına baktığımızda büyük veri setlerinden anlamlı bilgiyi elde etmek olduğunu görürüz. İlk olarak 1970'ler de ortaya çıkan veri tabanı yönetim sistemleri teknolojinin gelişimiyle birlikte artık farklı bir boyuta evrilmiştir. Günlük hayatımızın odak noktası haline gelen internet ve veri depolama kapasitesini arttıran teknolojik cihazlar veri boyutlarının büyümesine yol açmış ve anlamlı veriyi elde etme süreci de gittikçe daha fazla zorlaşmıştır. Veri madenciliğinde sistem için gerekli metotlar kullanılarak büyük miktarlı veriden bilgi çıkarımı hedeflenir. Veri madenciliği veriden modeller geliştirmek, ilgi çekici yapılar veya yinelenen temalar bulmak vb. için istatistik, yapay zekâ, bilgisayar bilimi gibi çeşitli bilim dallarından algoritmalar kullanmaktadır (Korcan; ARSLANTEKİN, 2016).

Klasik istatistik uygulamaları ile veri madenciliği arasındaki temel farklılıklardan biri, kullanılan veri kümesinin büyüklüğüdür. İstatistikçiler genellikle birkaç yüz veya bin veriyi ele alırken, veri madenciliği uzmanları için bu kavram milyonlarca veya milyarlarca veriyi içermektedir. İstatistikçiler için büyük olarak kabul edilen bir veri kümesi, veri madenciliği

alanında rutin bir ölçüttür. Bu tip büyük veri tabanları günümüzde sensörler hayatımıza girdiğinden beri sıkça ortaya çıkmaktadır (Oğuzlar, 2003).

Veri madenciliğinin temel süreçlerine baktığımıza 6 adımdan oluştuğunu görürüz:

1. Veri temizleme
2. Veri bütünleştirme (Birleştirme)
3. Veri indirgeme
4. Veri dönüştürme
5. Belirlenen algoritmaya göre işlem yapma
6. Sonuç ve değerlendirme

Bu adımlara genel olarak bakarsak veri temizleme adımı, veri setlerinde bulunan eksik veya gürültülü verilerin çıkarılması veya temizlenmesi, doldurulması adımlarıdır. Bu adım günümüze kadar istatistik yöntemleri baz alarak satır bazında silme, eksik verinin ortalamasını alma gibi işlemlerle atılmıştır. Günümüze gelindiğinde ise gelişen ve her gün katlanarak artan yapay zekâ teknolojisi ve onun alt dallarından makine öğrenmesi, derin öğrenme veya yapay sinir ağları gibi yöntemlerle eksik veya hatalı veriler mümkün olan en iyi algoritmalarla tamamlanıp, veri temizleme işlemi gerçekleşmektedir.

Veri bütünleştirme aşaması, farklı veri tabanlarından veya kaynaklardan elde edilen verilerin bir araya getirilip değerlendirilebilmesi için çeşitli türlerdeki verilerin tek bir formata dönüştürülmesi işlemidir (Akif *et al.* n.d.). Bu işlem farklı veri setlerindeki aynı durumu anlatan iki olayın niteliklerinin birleştirilmesi işlemidir. Aynı çalışma için veri toplandığında bir veri tabanında girişler farklı olabilir. Örneğin öğrenci bilgi sistemine ait 2 farklı veri tabanına öğrenci numaraları öğrenci_id ve öğrenci_no şeklinde yazılmış olabilir. Bu durumun önüne geçebilmek için veri bütünleştirme aşamasını uygulamamız gerekir.

Veri indirgeme aşamasında ise çalıştığımız veri setlerinin özelliklerine göre eğer veriler boyutlarına veya özelliklerine göre sıkıştırılabiliyorsa veya genelleştirilebiliyorsa bu adımı uygularız. Bu işlem araştırmacı için veri setini daha okunabilir bir hale getirir ve bu sayede veri setinin anlamlı bir hale gelmesi kolaylaşır.

Veri dönüştürme işlemi, verinin kullanılacak modele uygun bir şekilde şeklinin dönüştürülmesini içerir, bu süreçte verinin içeriği korunmaya çalışılır. Dönüştürme işlemi, kullanılacak model veya algoritmanın gereksinimlerine uygun bir biçimde gerçekleştirilmelidir, çünkü verinin model ve algoritma tarafından doğru bir şekilde anlaşılması hayati öneme sahiptir. Değişkenlerin ortalama veya varyanslarının önemli ölçüde farklı olduğu durumlarda, yüksek ortalamaya veya varyansa sahip değişkenlerin diğerlerini etkileme eğilimi daha fazla

olabilir, bu durum yanıltıcı sonuçlara neden olabilir. Bu nedenle, verideki bu tür yanlılıkları düzelten bir normalizasyon işlemi uygulanır. Beşinci aşama olan seçilen algoritmanın uygulanması, belirlenen niteliklere göre seçilen yöntemin uygulanması sürecini ifade eder. En son aşamada ise veri madenciliği uygulanan veri setinin uygun bir şekilde raporlanması ve sonuçlarının sunumu yapılır.

Eksik veri

Bilginin topladığımız verilerden elde edildiğini yukarıda anlattık. Veri toplama aşamasında karşılaştığımız zorlukların başında ölçüm hataları, kayıt hataları, sensör arızaları gelmektedir. Bu ve bunun gibi birçok sebepten dolayı veri setlerimizde eksik veriler olması sıkça karşılaştığımız bir durumdur. Bu eksik veriler veri madenciliği sürecinde başa çıkılması gereken temel sorundur. Eksik veriler ise 3 farklı şekilde sunulmuştur:

Tamamen rastgele eksik (MCAR)

Veri toplama süreçleri, veri setlerindeki eksik verilerin oluşumunu da açıklar. Eğer toplanan verilerdeki eksiklik tamamen rastgele ve kontrol altında olmayan bir şekilde oluşmuşsa bu tamamen rastgele eksik veri olarak adlandırılır (Little & Rubin, 2019). Bu duruma örnek olarak evlerimizde günlük olarak kullandığımız malzemelerin listesini bir anket olarak oluştururken bazı malzemelerin anketi yapanların gözünden kaçması verilebilir.

Rastgele eksik (MAR)

Verilerin rasgele eksik olduğu durum ise eksik verilerin varlığı diğer değişkenlerle ilişkili ve eksik verilerden bağımsızdır. Yani eksik veri miktarı diğer değişkenlerin veri toplama aşamasında eksik değerlere etkisi vardır. Bu duruma örnek olarak yukarıda verdiğimiz örnekten devam edecek olursak, evlerde kullanılan malzemelerin anketi yapılan bir çalışmada, anketi dolduran kişilerden bazıları sigara tütün gibi ürünleri tükettiğini ankette belirtmek istememesi olarak verilebilir.

Rastgele olmayan eksik (MNAR)

Rastgele olmayan eksik veri durumu ise eksik verinin eksik verilere bağlı oluşması durumudur. Yani eksik verilerin oluşum şekli katılımcıların inisiyatifine bağımlı olarak değişir. Eksik veri mekanizması MCAR ya da MAR olduğunda eksik veriler ihmal edilebilir. İhmal edilebilir mekanizmalarda araştırmacı, verilerin analizinde eksik verilerin nedenlerini görmezden gelebilir ve böylece eksik veriler için kullanılan yöntemleri basitleştirebilir. Eksik veri mekanizması MNAR olduğunda ise eksik veriler göz ardı edilemez. Araştırmacılar yanlı

analizlerden kaçınmak için eksik verinin nedenlerini araştırmalı ve bu mekanizmaya uygun yöntemler geliştirilmelidir.

Eksik verileri tamamlama yöntemleri

Eksik veri kavramı, veri setleriyle yapılan çalışmalar başladığından beri önemli bir sorun teşkil etmektedir. 1970'lerden itibaren bu sorunun çözümü için geliştirilen pek çok yöntem var. Bu yöntemlere genel olarak baktığımızda son 20 yıla kadar istatistik yöntemlerin ağırlıklı olduğu, son 20 yıllık süreçte de yapay zekâ ve alt dallarının bu görevi üstlendiğini görmekteyiz. Bu bağlamda baktığımızda eksik veriler geçmişten günümüze silinmiş, ortalama atanmış, veya tamamlama yöntemleri seçilmiştir. Silme yöntemlerine baktığımızda liste ya da satır bazında silme işlemini veya çiftler bazında silme işlemini görürüz (RJA Little, 2019). Liste ya da satır bazında silme işleminde eksik veri barındıran satır veya özelliğin veri setinden çıkarılması işlemidir. Aşağıda verilen tablo 1 de bilgisayar parçaları, envanter miktarı ve fiyatı örnek olarak verilmiştir.

Tablo 1. Satır Bazında Silme İşlemi

Ürün	Fiyat	Envanter miktarı
Monitör	110	103
Klavye	223	214
Anakart	------	120
Ekran kartı	134	121
İşlemci	202	------

Tablo 1'e baktığımızda hücrelerde bulunan eksik verileri silme işlemine tabi tutabiliriz. Satır bazlı silme işlemi yaptığımızda tablo aşağıdaki hale dönüşür:

Tablo 2. Satır Bazında Silme İşlemi Uygulanmış Tablo

Ürün	Fiyat	Envanter miktarı
Monitör	110	103
Klavye	223	214
Ekran kartı	134	121

Tablo 2'den de görüleceği üzere satır bazlı silme işlemi yaptığımızda eksik veri olan satırlar veri setinden çıkarılır. Örneklem sayısı az olan veri setlerinde bu işlemi yapabiliriz. Ancak büyük çaplı veri setlerinde satır bazında silme işlemi, yanlılığa yol açar (Leke & Marwala, 2019). Çiftler bazında silme işlemi, her bir değişken için mevcut olan kayıp değerlerin sayısı olarak belirlenen kayıp veri miktarını ifade eder. Bu yöntem, liste bazında veri silme yönteminin neden olduğu veri kaybını azaltmak ve kovaryans/korelasyon matrislerini

hesaplamak amacıyla geliştirilmiştir. Genel olarak bütün tablonun silinmesi yerine gerekli işlem çiftleri silinir (Evren ŞEKER *et al.* 2017). Bir diğer eksik verilerle başa çıkma yöntemi olarak ortalama alma yöntemi kullanılır. Bu yöntemi aşağıda tablo 3 yardımıyla anlatmak gerekirse:

Tablo 3. Ortalama Alma Yöntemi Öncesi

Ürün	Fiyat	Envanter miktarı
Monitör	110	103
Klavye	223	214
Anakart	------	120
Ekran kartı	134	121
İşlemci	203	144

Tablo 3'e baktığımızda fiyat sütunun altında Anakart bölmesinin boş olduğunu görürüz. Bu boş sütunu doldurmak için sütundaki sayıların ortalamasını alırız. Sütun değerlerine baktığımızda (110+223+0+134+203) /5 = 134 değerini buluruz. Eksik değeri de aynı şekilde 134 değeri ile doldururuz. Ortalama alma yöntemi de aynı şekilde küçük çaplı veri setlerinde işe yarasa da büyük çaplı veri setlerinde yanlılığa yol açar ve nümerik veri setleri haricinde kullanıma uygun değildir.

Tablo 4. Ortalama Alma Yöntemi Sonrası

Ürün	Fiyat	Envanter miktarı
Monitör	110	103
Klavye	223	214
Anakart	134	120
Ekran kartı	134	121
İşlemci	203	144

Yöntemlerden bir diğeri eksik veriler yerine atama yapmadır. Bu yöntemleri aşağıdaki gibidir:

Bayesci veri atama

Bayes teoremi, X ve Y gibi iki rastgele olay arasındaki koşullu olasılıklar ile marjinal olasılıkları, rassal bir süreçle ilişkilendiren bir matematik teoremidir. Bayesci veri atama yaklaşımı ise gözlenebilir veriler kullanılarak olasılıkların tahmin edildiği ve elde edilen bilgilerle eksik verilerin tamamlandığı bir yöntem olarak tanımlanabilir. (Kalaycıoğlu, 2017). Analiz tamamlandığında, regresyon parametrelerine ait nokta tahminleri, bu parametrelerin

sonsal beklenti dağılımlarının ortalamaları kullanılarak hesaplanır. Aralık tahminleri ise, yine aynı sonsal beklenti dağılımları kullanılarak hesaplanır. (ÇÜM *et al.* 2018).

Beklenti maksimizasyonu algoritması

Beklenti maksimizasyonu (EM) eksik veri içeren olasılıklı yöntemlerde parametre tahmini için tasarlanmış, bir objenin hangi kümeye ait olduğunu belirlemede tahminsel ölçütleri kullanan model tabanlı bir tekniktir. Beklenti ve maksimizasyon olarak 2 adımdan oluşur. Beklenti adımı eksik verinin tamamlanması için olabilirlik fonksiyonlarını değerlendirirken, maksimizasyon adımı da olabilecek en iyi sonuçlara odaklanır (Celik, 2013; ÇÜM *et al.* 2018).

Hot/Cold deck yöntemi

Bu metodolojide, eksik veriler önceki gözlemler ve diğer dış kaynaklardan türetilen sabit bir sayı ile doldurulur. Temel prensip, eksik gözlemleri doldurmak için kaydedilmiş güncel verilerin kullanılmasıdır. Esasen, bu yaklaşım, yerine ortalama koyma yöntemiyle benzerlik gösterebilir, ancak eksik gözlemi doldurmak için kullanılan değerin alındığı kaynak bakımından farklılık gösterir. Yerine ortalama koyma yönteminde karşılaşılabilecek aynı dezavantajlar, bu yöntem için de geçerlidir. Hot/Cold Deck yöntemi, araştırmacıya örnek ortalamasından daha geçerli bir değer ataması sağladığı için ortalama koymaktan daha avantajlı görülebilir. Bu yöntem, regresyon ataması ve yerine ortalama koyma gibi yöntemlerde olduğu gibi atanacak değerin kaynağına bağlı olarak farklı isimler alır. Örneğin, kendi grubundan türetilen bir sayı söz konusu olduğunda "Hot Deck Ataması" olarak adlandırılırken, diğer gruptan bir sabit türetildiğinde "Cold Deck Ataması" olarak adlandırılır. (MUSTAFA & DEMOGRAFİ, n.d.).

Regresyon analizi

Regresyon, değişkenler arasındaki ilişkileri araştırmak için bir istatistiksel araçtır. Genellikle, araştırmacı, bir değişkenin başka bir değişken üzerindeki nedensel etkisini belirlemek ister. Örneğin, bir şirket pazarlama çalışması için reklam harcamaları ile satışlar arasındaki inceleme için regresyon analizini kullanabilir. Burada bağımlı değişken satış, bağımsız değişken ise reklam harcamaları olarak değerlendirilebilir. Bu gibi konuları incelemek için, araştırmacı ilgi duyduğu değişkenlerin verilerini toplar ve regresyonu, nedensel değişkenlerin etkilenen değişken üzerindeki nicel etkisini tahmin etmek için kullanır. Araştırmacı, tahmin edilen ilişkilerin "istatistiksel anlamlılığını" da değerlendirir, yani, gerçek ilişkinin tahmin edilen ilişkiye yakın olma derecesini ele alır (Sykes, n.d.).

Doğrusal regresyon, bağımlı değişkenin bağımsız değişkenlerle açıklanması amacıyla kullanılan bir modeldir. Geleneksel eksik veri yöntemlerinden ayrılarak, regresyon ataması, eksik verinin rastgele olmayan (MCAR) ve bağımsız değişkenlere bağlı olarak rastgele (MAR) olması durumlarında, eksikliği etkileyen değişkenleri istatistiksel modellere dahil etme şartıyla, yansız parametre tahminleri sunar. Aynı zamanda mevcut verilerin tümünü kullanarak eksik veriye yönelik atamaları iyileştirmeye çalışır.

Regresyon tahmininin bir dezavantajı, eksik değerler için koşullu tahminlerin kullanılması nedeniyle tüm değerlerin regresyon çizgisi etrafında toplanacak olmasıdır. Bu durum, değişkenler arasındaki ilişkilerin doğal olarak güçlendirilmesine yol açabilir. Ayrıca regresyon denkleminde standart hataların küçümsenmesine ve verilen değişkenliğin hafife alınmasına yol açabilir (Hyunshik Lee & Särndal, 1994).

Çoklu atama (Multiple İmputation)

Çoklu atama yöntemleri yukarıda bahsedilen tekli atama yöntemlerinin kombinasyonu olarak karşımıza çıkar. Bu yöntem maksimum benzerliği kendine hedef alır. Çoklu atama yöntemi her bir veri seti için m adet tam veri seti elde etmeyi amaçlar. Örneğin ortalamalarının alınması gibi yöntemlerle birleştirilerek tek bir eksiksiz veri setine dönüştürülür. Bu şekilde oluşturulan tek bir veri seti analizlerde gerçek veri seti gibi değerlendirilir (Faktör & İncelenmesi, n.d.)

Makine öğrenmesi yöntemleri

Gelişen yapay zekâ teknolojileriyle birlikte mevcut veriden öğrenen makine öğrenimi yöntemleri eksik veri tamamlama alanında kullanılmaya başlanmıştır. Literatüre baktığımızda en çok kullanılan yöntemlerden bazılarını ve özelliklerine aşağıda verilmiştir.

K-En Yakın komşu (K-NN) algoritması

K-NN algoritması makine öğrenmesi algoritmaları arasında en çok kullanılan örnek tabanlı öğrenme algoritmalarından biridir. Geçmiş gözlemlere bağlı olarak bir olayın sonucu kendine en yakın olaylara en yakındır mantalitesi ile hareket eder. Hem uygulama kolaylığı hem de kolay matematiksel alt yapısı sayesinde tahmin problemlerinde sıkça kullanılır. Aranan veri noktasına en küçük Öklid mesafesini kendine ölçü alır (ALTUNKAYNAK *et al.* 2020). Algoritma adımları aşağıdaki gibidir:

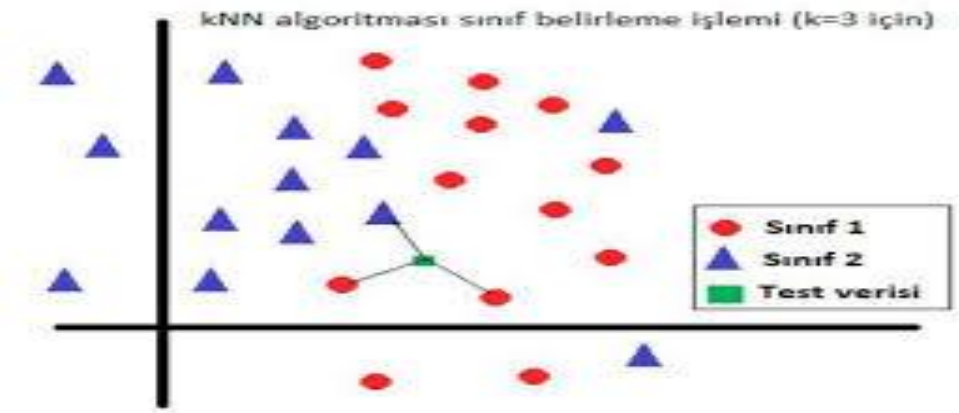
Adım 1: k (komşuluk) değeri belirlenir.

Adım 2: Hedef değişkene olan uzaklıklar belirlenir.

Adım 3: Uzaklık değerleri büyüklüklerine göre sıralanır ve en küçük eğer kabul edilir.

Adım 4: Bir önceki adımda belirlenen k tane komşuluğun sınıfları belirlenir.

Adım 5: Bir önceki adımda belirlenen sınıflardan en çok gözlemlenen sınıf seçilir (Özarı & Demirkale, 2022).



Şekil 1. K-en yakın komşu algoritması genel yapısı (Coşar & Deniz, 2021)

K-NN algoritmasında en önemli nokta k komşuluk değerinin doğru bir şekilde belirlenmesidir. Tahmin problemlerinde etkin olarak kullanılsa da büyük veri setlerinde yüksek bellek gereksinimine sahip olduğu için ve komşuluk gereksinimleri hassas olduğu için tek başına yetersizdir (Taşcı & Onan, 2016).

Karar ağacı regresyonu (Decision tree regressor)

Karar ağaçları, doğrusal olmayan hem sınıflandırma hem de tahmin problemlerinde kullanılabilen kategorik ve sürekli verilere sahip değişkenler ile kullanılabilen makine öğrenimi yöntemlerinden biridir. Sistemin öğrenme prensibi tümevarım mantığı ile aynıdır (*Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi Cilt: 6 Sayı: 2 Aralık 2005*, n.d.).

Hedef değişken özniteliklerin eşik değeri olarak tanımlanır. Algoritma her bir hedef değişken için eşik değer hesaplar ve özniteliğin eşik değerine göre alt gruplara ayrılır. Ağaca benzeyen bir akış şemasını andıran karar ağaçlarında bir kök ve boğumları, dallar ve yapraklardan oluşur (Harita Teknolojileri Elektronik Dergisi Cilt; Kavzoğlu *et al.* 2010). Örneğin hafta sonu aktivitesi için bir karar ağacı oluşturmamız gerekirse, hava durumunu iyi veya kötü olarak sınıflara ayırırız ve alt sınıf olarak eğer hava iyi ise pikniğe git veya hava kötü ise evde film izle şeklinde sınıflandırabiliriz. Bu örneğe ait algoritma aşağıda verilmiştir:

Sınıf 1: Hava güzel

Sınıf 1.1: Arkadaşlar müsait

Sonuç 1.1.1: Pikniğe git

Sınıf 1.2: Arkadaşlar müsait değil

Sonuç 1.2.1: Yürüyüş yap

Sınıf 2: Hava kötü

Sınıf 2.1: Yağmur yağıyor

Sonuç 2.1.1: Evde film izle

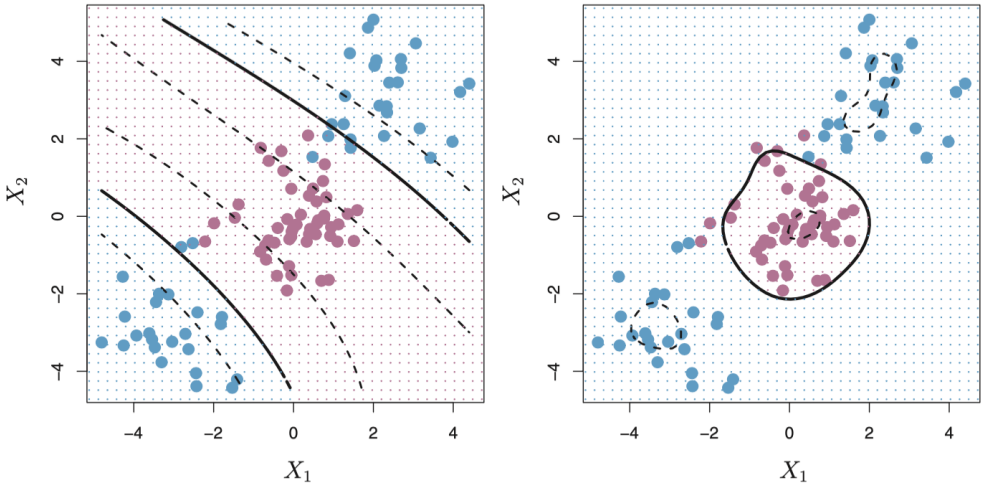
Sınıf 2.2: Yağmur yağmıyor

Sonuç 2.2.1: Alışverişe git.

Yukarıdaki algoritma örneği basit bir şekilde karar ağacı mantığını anlatmaktadır. Karar ağaçları ile eksik veri tahmini yapmanın avantajlarından biri yapının kolay anlaşılabilir olması ve değişkenlerin katkılarını ve önemlerini kategorize ederken yapısı sayesinde kolaylık sağlar. Ancak büyük ve dengesiz veri setlerinde aşırı öğrenme (overfitting) eğilimi gösterebilir.

Destek vektör makineleri (DVM)

Destek vektör makineleri sınıflandırma ve regresyon için kullanılan makine öğrenmesi yöntemlerinden biridir. İstatistiksel öğrenme teorisi ve yapısal risk minimizasyonu üzerine kuruludur. Destek vektör makinelerinin temel amacı eğitim verilerini sınıf etiketleriyle ayırabilen çok boyutlu uzayda bir fonksiyon bulabilmektir (HARMAN, 2021).



Şekil 2. Destek vektör makineleri (Cheng *et al.* 2017)

Destek vektör makineleri Kernel fonksiyonları kullanarak veri setini yüksek boyutlu uzaylarda temsil edebilir. Aykırı değerlere gösterdiği direnç sayesinde düşük boyutlu veri setlerinde oldukça iyi performans gösterir. Bu özellikleri sayesinde veri madenciliği alanında oldukça sık bir şekilde kullanılmaktadır (Güner *et al.* n.d.).

KURAMSAL TEMELLER

Teknolojik hareketliliğin artması, sosyal medyanın yaygınlaşması, sensör sistemlerinin teknolojik gelişmeleri, iletişim teknolojilerinin daha erişilebilir hale gelmesi ve birçok iş kolunun elektronik platformlara taşınması gibi faktörler, üretilen verinin çeşitliliğini artırmanın yanı sıra toplama hızı ve miktarını da önemli ölçüde artırmıştır. (Atalay & Çelik, 2017). Ancak meta grup tarafından yapılan bir araştırmaya göre, ham verilerin kalitesiz veya yetersiz olması nedeniyle yalnızca orijinal veriler kullanılırsa, ilgili projelerin %41'inden fazlası başarısız olacaktır(Shi *et al.* 2015). Yaşanan bu gelişmeler veri yığınlarının oluşmasına neden olmuş ve kaliteli veri ihtiyacını doğurmuştur. Geçmişte günümüze kıyasla daha küçük ölçekli veri setlerine yönelik geliştirilmiş algoritmalar, büyük veri işleme gereksinimlerini karşılamakta yetersiz kalmaktadır. Veri analizi süreci, bilimsel araştırma ve veri madenciliği süreçlerinin temel adımlarından birini oluşturur. Bu aşamada toplanan veriler, amaç doğrultusunda en uygun istatistiksel, matematiksel veya yapay zekâ teknikleriyle işlenerek analiz edilir. (Arslan *et al.* n.d.). Veri analizindeki problemlerin başında ise eksik/hatalı veri gelmektedir. Bir veri tabanındaki eksik veriler çeşitli nedenlerle ortaya çıkabilir. Veri girişi hataları, katılımcıların veri toplama sürecindeki bazı maddelere yanıt vermemesi, araçların arızalanması ve diğer çeşitli nedenlerden kaynaklanabilir (Abdella & Marwala, 2005). Geçmişten günümüze, eksik verilerle başa çıkabilmek amacıyla çeşitli yöntemler geliştirilmiştir. Eksik verilerle devam etme, eksik gözlemleri analiz dışına çıkarma, eksik gözlemler yerine veri atama veya çeşitli istatistiksel yöntemlerle eksik verileri tamamlama gibi yöntemler, bu tür durumlarla başa çıkmak için sıkça kullanılan yaklaşımlardandır. (Sezgin *et al.* n.d.). Ancak bu yöntemler günümüz veri yığınlarında yavaş yavaş işlevini kaybetmekte ve güvenilirlikleri azalmaktadır. Bu yüzden son zamanlarda eksik veri tamamlama çalışmaları için yapay zekâ teknikleri daha çok ilgi görmekte ve çalışmalar bu yöne kaymaktadır. Bu çalışmalardan bazıları; genetik algoritmalar ile veri tahmini (Kahraman *et al.* 2012), meta-sezgisel algoritmalar kullanılarak veri tahmini(Holland, 1992), makine öğrenmesi algoritmaları ile veri tahmini(Ray, 2019), ve yapay sinir ağları ile veri tahmini şeklindedir(Awad *et al.* 1990).

Bir diğer çalışmada ise orman ekolojik verilerinde uzun dizi eksik veri atama işlemi için yeni bir derin öğrenme modeli olan Bip-informer önerilmektedir. Bu model eksik verilerin ileri ve geri yönlerindeki bilgileri etkin bir şekilde kullanmak ve hata birikimini azaltmak için çift yönlü bilgi iletim mekanizması ve çift yönlü kayıp ceza stratejisi içermektedir (Wang *et al.* 2024). Daha farklı metotlardan biri de eksik verileri yerel uzayın bilgisini kullanarak doldurmak

için yeni bir yöntem öneren Do Gyun Kim ve Jin Young Choi'nin çalışmasıdır. Önerdikleri yönteme göre; yerel uzay hedef sınıfın verilerini tam olarak tanımlayan geometrik bir tek sınıflı sınıflandırıcı olan H-RTGL (Hiper- Dikdörtgen) ile tanımlamaktır. Bu yöntem eksik verilerin yakınındaki yerel uzaydaki bilgileri etkin bir şekilde kullanmak ve eksik veriler arasındaki ilişkiyi yansıtmak için bir bileşik bulanık model kullanmaktadır. Model eksik verileri hesaplamak için her yerel uzaydan elde edilen imputasyon modellerinin ağırlıklı bir toplamını alarak, birden fazla yerel uzayın etkisini dikkate almaktadır (D. G. Kim & Choi, 2024). Yapay zekâ ile birlikte bu alana birçok farklı çalışma ve yöntem de gelişmeye devam etmektedir. Bunlardan biri de eksik verileri doldurmak ve hedef değişkenin doğrudan nedenlerini ve doğrudan etkilerini ayırt etmek için yeni bir yöntem olan misLCS'idir (Local Casual Structure Learning with Missing Data). Bu yöntem eksik verileri yinelemeli olarak atayarak, hedef değişkenle yakından ilişkili olan değişkenlerin bir alt kümesini elde etmekte ve bu alt kümede yerel nedensel iskeleti inşa etmekte ve koşullu bağımsızlık testleri ve Meek kuralları kullanarak kenar yönlerini belirlemektedir (Sheng *et al.* 2024). R tabanlı yapılan bir çalışmada eksik verilere tahmin denklemleri uygulamak için MIIPV adlı bir paket geliştirilmiştir. Paket eksik kovaryans verilerini ortalama puan ve ters olasılık ağırlıklı yöntemlerin bir kombinasyonu ile imha etmektedir. Ayrıca eksik veriler için çoklu atama modeli kullanmaktadır (Bhattacharjee *et al.* 2024). Eksik veriler çalışılan her bir veri tabanına göre model değişkenliğini kaçınılmaz olarak göstermektedir. Bu durumda her bir veri seti yapılan çalışmalara özgü modellerde daha az hata ile doldurulmaktadır. Bunun bir örneğini göstermek için konut binalarında izlenen verilerde eksik değerlerin tamamlanması için 6 farklı yöntem üzerinde çalışılmış ve veri özelliklerine bağlı olarak enerji tüketimi verileri için veri tabanlı modeller, davranış ve çevre verileri için doğrusal enterpolasyon yönteminin daha iyi sonuç verdiğini gözlemlemişlerdir (J. Kim *et al.* 2023b). Veri seti özelinde yapılan çalışmalardan hareketle geliştirilen yöntemlerin biri de bir imputasyon algoritması olan OMIG'tir. Yüksek boyutlu karışık tip verilerde eksik değerleri tahmin etmek için genelleştirilmiş faktör modelleri temelli çevrimiçi bir imputasyon algoritmasıdır (Liu *et al.* 2023). Bu algoritmaların yanı sıra çeşitli imputasyon yöntemlerinin ortak çalışmasını konu eden çalışmalardan biri Shafig Alam ve arkadaşlarının yayınlamış olduğu eksik değerlerin sıralı verilerde nasıl ele alınacağına dair çeşitli imputasyon değerlerini araştıran ve tekniklerin kümeleme ve sınıflandırma analizlerinin geçerliliğini nasıl etkilediğini değerlendiren farklı veri kümeleri ve eksik değer yüzdeleri üzerinden karşılaştırmıştır. Bu karşılaştırmada eksik değerlerin atanması için en uygun yöntemi belirlemek ve atamanın kümeleme ve sınıflandırma performansı üzerindeki etkisini analiz etmek için k-ortalamalar, Partitioning Arounds Medoids (PAM), k-en yakın komşu (K-NN), Naive Bayes (NB) ve Multilayer Perceptron (MLP) gibi kümeleme ve sınıflandırma metotları kullanılmıştır. Çalışma

karar ağacı yönteminin eksik değerlerin atanması için en etkili yöntem olduğunu, orijinal veriye en yakın değerleri ürettiğini ve yüksek doğruluk sağladığını bulmuştur (Alam *et al.* 2023). Karşılaştırma çalışmalarından biri de k-nn ve yapay sinir ağı yöntemini yüzey sıcaklığı karakterizasyon verileri üzerinde karşılaştıran J.L. Lopez ve arkadaşlarının çalışmasıdır. Bu çalışmada iki yöntem kıyaslanmış ve k-nn yönteminin veri setinde daha iyi sonuç verdiğini göstermiştir (López *et al.* 2021). Literatüre baktığımızda bu çalışmaların farklı veri setleri, farklı eksik oranları ve farklı şartlarda çözümler verdiğini görmekteyiz.

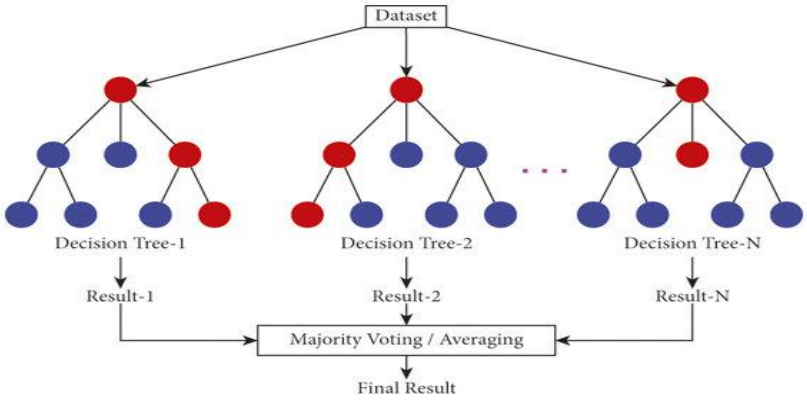
MATERYAL VE METOT

Tez çalışması kapsamında geliştirilen Python tabanlı uygulamanın işleyiş aşamaları ve çalışma mekanizması aşağıda detaylarıyla birlikte verilmiştir. Öncelikle kullanılan makine öğrenmesi yöntemleri verilmiş ve detaylıca anlatılmıştır.

Rastgele Orman Regresyonu (Random Forest Regression)

Rastgele orman algoritması, karar ağacı temelli bütünleşik bir öğrenme algoritmasıdır. Birden fazla karar ağacının birleşimi şeklinde olan bu yöntemde her ağaç rastgele bir veri ve özellik alt kümesi tarafından eğitilir. Her ağaç çekilen özellikleriyle paralel olarak tahmin sonuçlarına ulaşmaya çalışır. Elde edilen bütün sonuçlar bir araya getirilerek ortalama sonuç hesaplanır (Guo *et al.* 2023).

Rastgele orman algoritmasında birden fazla ağaç algoritması kullanıldığı için aşırı öğrenme (overfitting) riski minimize edilir. Şekil 3'te görüldüğü gibi araştırmacılar için analiz aşamasında anlaşılması kolaydır ve yüksek doğrulukla tahmin yeteneği gelişmiştir. Hem regresyon hem de sınıflandırma problemlerinde kullanılabilirliği ve büyük veri setlerinde rahatlıkla çalışabilme yeteneği sayesinde tercih edilen önemli makine öğrenmesi yöntemlerinden biridir (Li *et al.* 2018; Seki *et al.* 2023; Üniversitesi & Mühendisliği, 2024).



Şekil 3. Rastgele orman algoritması işleyişi (Khan *et al.* 2021)

Algoritmanın işleyişini daha net anlatabilmek adına şu örneği verebiliriz: bir şirketin kullanıcılarına hedeflenmiş reklamlar sunan bir uygulamaya sahip olduğunu ve bu uygulamanın gösterilen reklamları değerlendirerek bir reklamın tıklanma olasılığını hesaplayan ve bu hesap tahminini rastgele orman algoritması kullanarak yaptığını farz edelim. Bunun için öncelikle kullanıcı davranışları, demografik özellikler, sistemi kullanma süresi gibi veriler toplanır. Daha

sonra veriler ön işleme tabi tutulduktan sonra rastgele orman algoritması için eğitim veri seti, reklamların daha önceki performansına bağlı olarak oluşturulur. Oluşturulan eğitim veri seti üzerinden birçok karar ağacı oluşturularak rastgele örnekleme ve rastgele özellik seçimi ile model eğitilir. Bir kullanıcı uygulamayı kullanırken, sistem bu kullanıcının özelliklerinden her ağacın tahminini alır. Her ağacın tıklanma olasılığını bir oy alarak değerlendirmesiyle toplu bir tahmin elde edilir. Toplanan tahminler bir araya getirilerek bir sonuç elde edilir. Çoğunluk oyu alan bir reklamın tıklanma olasılığını daha yüksek kabul edersek, reklam yönetim sistemi bu oylamaya en uygun reklamı seçip göstermiş olur.

Yinelemeli atama (İterative İmputer)

Bu yöntem eksik verileri tahmin etmek ve doldurmak için kullanılan makine öğrenmesi tabanlı Python paketidir (Oberman *et al.* n.d.). İterative imputer farklı makine öğrenme modellerini kullanma esnekliği sağlar. Bu sağlamış olduğu esneklik sayesinde eksik değerleri tahmin etmek için veri setinin özelliklerini ve eksik olmayan gözlemlerle eğitilmesi sağlanır. Beraber çalıştığı yöntemle birlikte tahmin aşaması ve eksik verileri doldurma aşamasında bir döngü oluşturur ve bu döngü birkaç iterasyon boyunca devam ederek her bir iterasyonda tahminler iyileştirilir (Altukhova, 2020). Bu yöntem özellikle büyük ve karmaşık veri setlerinde beraberinde kullanılan algoritmayla beraber iyi sonuçlar üretir (Wan, 2022).

Çapraz doğrulama (Cross validation)

Çapraz doğrulama bir modelin performansını değerlendirmek için kullanılan ve veri kümesini eğitim ve test kümelerine bölerek modelin genelleme yeteneğini daha iyi ölçebilmek adına kullanılan bir yöntemdir (Malakouti, 2023; Narin *et al.* n.d.). En yaygın kullanılan biçimi ise k-katlı çapraz doğrulamadır. Bu yöntem de veri seti k adet parçaya bölünerek her bir parça öncelikle test kümesi olarak kullanılarak modelin genelleme performansı değerlendirilir (Bilgin, n.d.).

K-katlı çapraz doğrulama daha iyi açıklamak için elimizde bir veri seti olduğunu ve bu veri setinin 5 parçaya (k=5) böldüğümüzü düşünelim. İlk iterasyonda birinci parça hariç kalan dört parça kullanılarak model eğitilir ve birinci parça test kümesi olarak kullanılır. İkinci iterasyonda ise ikinci parça hariç kalan dört parça kullanılarak model test edilir ve ikinci parça test kümesi olarak kullanılır. Bu süreç her parça sırasıyla test kümesi olarak kullanılıncaya kadar devam eder. Her iterasyon sonunda elde edilen performans metrikleri kaydedilir ve bütün iterasyonlar bittiğinde ortalaması alınarak genel bir değerlendirme ölçütü elde edilir.

Bu yöntem modelin gerçek dünya verileri üzerine daha tutarlı ve güvenilir bir performans sağlamasına yardımcı olurken aşırı öğrenme riskini de en aza indirir.

Geliştirilen algoritma

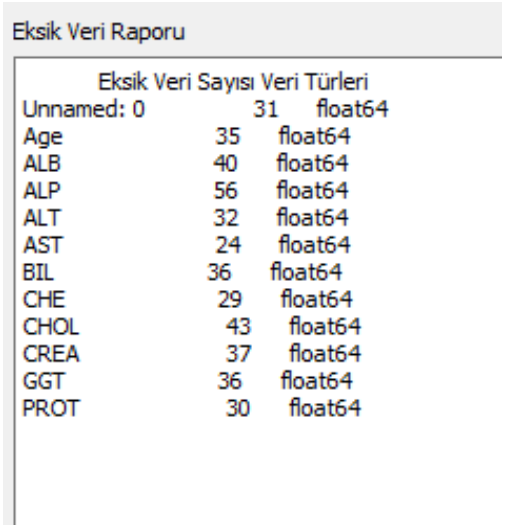
Eksik veri problemini ele almak ve veri madenciliği konusunda uzman olmasa bile her araştırmacının eksik verileri problemi ile başa çıkmasını kolaylaştırmak amacıyla Python tabanlı bir uygulama geliştirilmiş ve uygulama masaüstü uygulaması haline dönüştürülmüştür. Geliştirilen algoritma nümerik veriler üzerinde çalışmaktadır. Algoritma adımları ve detayları adım adım aşağıda anlatılmıştır.

Adım 1. Uygulama arayüzü

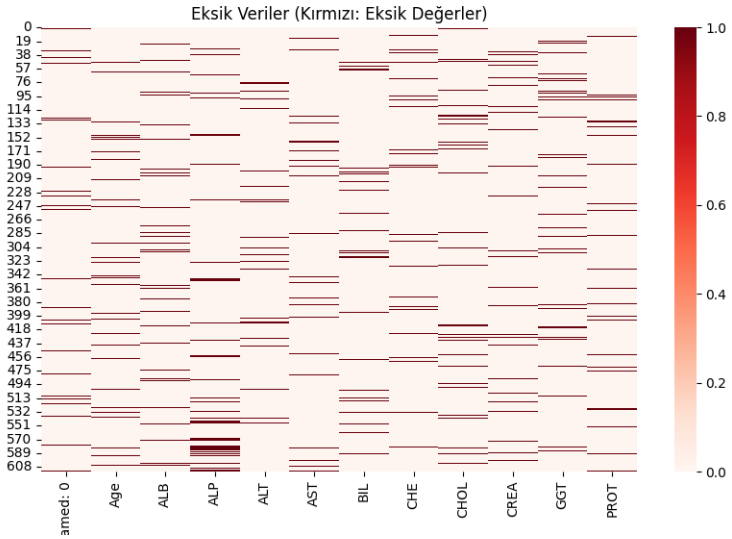
Python programlama dili kullanılarak PyQt5 kütüphanesi kullanılarak bir masaüstü uygulaması geliştirilmiştir. Bu uygulama dosya seçme, eksik veri raporu görüntüleme alanı, eksik verileri doldurma düğmesini içeren işlevselliğe sahiptir. Arayüz kullanıcı veri madenciliği alanında uzman olmasa bile kolaylıkla kullanabileceği bir basitlikte inşa edilmiştir.

Adım 2. Dosya seçme ve eksik veri analizi

Bu aşamada kullanıcı arayüz üzerinden eksik verilerin bulunduğu csv formatında dosya seçer. Seçilen dosya üzerinden eksik veri analizi yapılır. Eğer dosya da eksik veri bulunmuyorsa hata raporu kullanıcıya iletilir. Seçilen dosyadaki eksik nümerik veriler ve tipleri kullanıcıya rapor olarak iletilir. Raporun yanı sıra eksik verilerin dağılım oranları da ısı haritası yardımıyla kullanıcıya gösterilir ve kaydedilmesi için olanak sağlar.



Şekil 4. Eksik veri raporu



Şekil 5. Eksik verilere ait ısı haritası

Adım 3. Eksik verileri doldurma ve kaydetme

Bu adımda uygulama eksik verileri doldurmak için iterative imputer sınıfı kullanılmaktadır. Bu sınıf eksik değerleri tahmin etmek için her bir özniteliği diğer özniteliklere bağlı bir regresyon problemi olarak ele alır. Bu işlem çapraz doğrulama ile belirlenen iterasyon sayısına kadar devam eder. Her iterasyonda, tahminler daha iyi hale getirilir. İterative imputer sınıfı tahminci olarak random forest regressor makine öğrenimi yöntemini kullanır. Veri tamamlamanın alt adımlarına bakacak olursak veri seti rastgele, eğitim ve test verisi olarak iki parçaya bölünür. Eğitim veri seti, makine öğrenmesi yöntemini eğitmek için kullanılır. Test veri seti ise rastgele orman algoritmasının performansını değerlendirmek için kullanılır. Test veri seti olarak %20'lik kısım, eğitim veri seti olarak da %80'lik kısım kullanılmaktadır. Sistemi genelleştirmek ve gerçek hayata uyumluluk derecesini belirlemek için iterasyon sayısını belirlemek en önemli adımlardan biridir. Bu iterasyon sayısını belirlemek için çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama, eğitim veri setini daha küçük parçalara bölerek, her bir parçayı sırasıyla test veri seti olarak kullanıp, geri kalanını eğitim veri seti olarak kullanmayı sağlayan bir yöntemdir. Bu şekilde, her bir parça için bir performans ölçüsü elde edilir. Bu ölçülerin ortalaması alınarak, genel performans ölçüsü elde edilir. Geliştirilen algoritmada performans ölçüsü olarak ortalama mutlak hata (MAE) kullanılır. Ortalama mutlak hata, tahmin edilen değerler ile gerçek değerler arasındaki farkların mutlak değerlerinin ortalamasıdır. Ortalama mutlak hata ne kadar küçükse, performans o kadar iyidir. En iyi iterasyon sayısını belirlemek için, farklı iterasyon sayılarında ortalama mutlak hata hesaplanmaktadır. En küçük ortalama mutlak hata değerine sahip olan iterasyon seçilmektedir. Bu sayı da ideal iterasyon sayısı olarak belirlenmektedir.

En iyi iterasyon belirlendikten sonra iterative imputer sınıfı ve tahmin yöntemi olan random forest regressor aracılığı ile eğitim veri setindeki eksik veriler tahmin edilir. Doldurulan bu veri seti bir pandas veri çerçevesine dönüştürülür. Pandas, veri analizi için kullanılan bir Python kütüphanesidir. Veri çerçevesi, tablo şeklinde verileri tutan bir veri yapısıdır. Bu veri çerçevesi sayısal olmayan öznitelikleri de içerecek şekilde genişletilir. Bu şekilde eksik verileri doldurulmuş tam bir veri seti elde edilir. Doldurulan veri yine arayüz aracılığıyla kullanıcının seçeceği dosya konumuna csv dosyası şeklinde kaydedilir.

Adım 4. Sonuç ve raporlama

Doldurulan ve kaydedilen veri setine ait raporlama işlemi scatter plot fonksiyonu ile görselleştirilir. Yukarıda hesaplanan ortalama mutlak hata değeri ekrana yansıtılır.

Tez kapsamında kullanılan veri setleri

Bu çalışma kapsamında kullanılan 4 farklı veri setine ait bilgiler aşağıda verilmiştir.

Hepatit C veri seti

Hepatit C hastalığının tahmini için kullanılan veri seti 615 hastanın kan testi sonuçlarını ve hepatit tanısı durumlarını içeren bir bağımlı değişken ve 12 bağımsız değişkenden oluşan veri setidir (Hoffmann *et al.* 2018).

	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood D	32	m	38.5	52.5	7.Tem	22.Oca	7.May	Haz.93	Mar.23	106	12.Oca	69
2	0=Blood D	32	m	38.5	70.3	18	24.Tem	3.Eyl	Kas.17	4.Ağu	74	15.Haz	76.5
3	0=Blood D	32	m	46.9	74.7	36.2	52.6	6.Oca	Ağu.84	5.Şub	86	33.2	79.3
4	0=Blood D	32	m	43.2	52	30.Haz	22.Haz	18.Eyl	Tem.33	Nis.74	80	33.8	75.7
5	0=Blood D	32	m	39.2	74.1	32.6	24.Ağu	9.Haz	Eyl.15	Nis.32	76	29.Eyl	68.7
6	0=Blood D	32	m	41.6	43.3	18.May	19.Tem	12.Mar	Eyl.92	6.May	111	91	74
7	0=Blood D	32	m	46.3	41.3	17.May	17.Ağu	8.May	7.Oca	Nis.79	70	16.Eyl	74.5
8	0=Blood D	32	m	42.2	41.9	35.8	31.Oca	16.Oca	May.82	4.Haz	109	21.May	67.1
9	0=Blood D	32	m	50.9	65.5	23.Şub	21.Şub	6.Eyl	Ağu.69	4.Oca	83	13.Tem	71.3
10	0=Blood D	32	m	42.4	86.3	20.Mar	20	35.2	May.46	Nis.45	81	15.Eyl	69.9
11	0=Blood D	32	m	44.3	52.3	21.Tem	22.Nis	17.Şub	Nis.15	Mar.57	78	24.Oca	75.4
12	0=Blood D	33	m	46.4	68.2	10.Mar	20	5.Tem	Tem.36	4.Mar	79	18.Tem	68.6
13	0=Blood D	33	m	36.3	78.6	23.Haz	22	7	Ağu.56	May.38	78	19.Nis	68.7
14	0=Blood D	33	m	39	51.7	15.Eyl	24	6.Ağu	Haz.46	Mar.38	65	7	70.4
15	0=Blood D	33	m	38.7	39.8	22.May	23	4.Oca	Nis.63	Nis.97	63	15.Şub	71.9
16	0=Blood D	33	m	41.8	65	33.1	38	6.Haz	Ağu.83	Nis.43	71	24	72.7
17	0=Blood D	33	m	40.9	73	17.Şub	22.Eyl	10	Haz.98	May.22	90	14.Tem	72.4
18	0=Blood D	33	m	45.2	88.3	32.4	31.2	10.Oca	Eyl.78	May.51	102	48.5	76.5
19	0=Blood D	33	m	36.6	57.1	38.9	40.3	24.Eyl	Eyl.62	5.May	112	27.Haz	69.3

Şekil 6. Hepatit C veri setine ait genel bir görsel

Altın fiyatı tahmini veri seti

2011-2019 yılları arası toplanmış, Bükreş'e ait altın fiyatlarının yer aldığı 1718 satır ve 80 sütundan oluşan veri setidir (Bucharest House Price Dataset, 2024).

Bükreş konut fiyatları veri seti

Bükreş kentine ait 2019 yılına ait 6 bağımsız değişken ve 80 sütundan oluşan veri setidir (*Bucharest House Price Dataset*, 2024).

UV-visible spektrum veri seti

1202 satır ve 6 sütundan oluşan ve molekül yapısını gösteren veri setidir. UV-visible datası, yapılan çalışmada elde edilen nanoparçacığın optik özelliklerinin karakterize edilmesi amacıyla kullanılmıştır (Aydin *et al.* 2022).

Değerlendirme kriterleri

Belirlilik katsayısı testi (R^2 testi)

R^2 testi doğrusal regresyon modelleri için bir uyum iyiliği ölçüsüdür. Bu istatistik bağımlı değişkendeki varyansın bağımsız değişkenler tarafından ne kadar açıklandığını gösterir. Model ve bağımlı değişken arasındaki ilişkiyi %0 ile %100 arasında ölçer (Yang, 2005). Veri noktalarının uyduğu regresyon çizgisinin etrafındaki saçılımları değerlendirir ve aynı veri setinde gözlenen daha yüksek R^2 değerleri gözlenen veriler ve uyumlu değerler arasında daha küçük farkları temsil eder. R^2 testi, bağımlı değişkenin ortalaması etrafındaki varyansın yüzde kaçının doğrusal model tarafından açıklandığını anlatır. Genellikle R^2 değeri ne kadar büyükse, regresyon modeli o kadar iyi uyum sağlar (Quinino *et al.* 2012).

Karekök ortalama hata testi (RMSE testi)

Kök ortalama kare hata (RMSE) testi, bir modelin tahmin performansını ölçmek için kullanılan istatistiksel bir yöntemdir (Villalobos *et al.* 2006). Gözlenen ve tahmin edilen değerler arasındaki farkın karelerinin ortalamasının karekökünü hesaplar. RMSE değeri, ne kadar küçükse modelin tahminleri o kadar doğrudur (Orgaz *et al.* 2006).

RMSE, hataların birimini gerçek değerlerle aynı yaparak yorum yeteneğini kolaylaştırır ve aykırı değerlere duyarlılığı sayesinde yüksek hataları cezalandırır (Testi & Giorgetti, 2022).

Değerlendirme kriterlerinin algoritma sonuçlarına uygulanması

Bu aşamada R^2 ve RMSE değerlerini hesaplamak için, orijinal veri seti ile doldurulmuş veri seti arasındaki testleri yapacak yapay sinir ağı geliştirilmiştir. Geliştirilen yapay sinir ağı giriş katmanı, gizli katman ve çıkış katmanı olmak üzere 3 katmanlı bir yapıya sahiptir. Veri setinin orijinal hali ve doldurulmuş hali giriş katmanına iletilir. Giriş katmanına alınan veriler gizli katmanlara iletilir. 64 nöronlu gizli katman gelen verilere ağırlıklarını uygular ve bir çıkış üretir. Ardından karmaşıklığı arttırarak veri seti arasındaki ilişkiyi öğrenmeye çalışır. 32

nöronlu gizli katman ise modelin daha yüksek düzeyde özellikleri öğrenmesine ve daha genel bir temsil yapısı kurmasına olanak sağlar. Veriler çıkış katmanına iletilerek istenilen R^2 değeri ve RMSE değerlerine ulaşılır.

ARAŞTIRMA BULGULARI

Geliştirilen algoritmanın geçerliliğini test etmek için veri setleri sırasıyla %5, %10 ve %15 oranında tamamen rastgele eksik (MCAR) bir şekilde azaltılarak algoritma veri setlerine uygulanmıştır. Elde edilen sonuçlar tablo halinde aşağıda verilmiştir.

Tablo 5. Hepatit Veri Tabanı Test Sonuçları

EKSİLME ORANI	RMSE DEĞERİ	R^2 TESTİ
%5	2,432197	0,76871
%10	2,897318	0,75332
%15	3,344554	0,72297

Tablo 6. Altın Fiyatları Veri Set Test Sonuçları

EKSİLME ORANI	RMSE DEĞERİ	R^2 TESTİ
%5	4,476231	0,79414
%10	5,067433	0,77641
%15	5,949174	0,68134

Tablo 7. Bükreş Konut Fiyatı Datası Test Sonuçları

EKSİLME ORANI	RMSE DEĞERİ	R^2 TESTİ
%5	8,441538	0,69147
%10	9,178293	0,68881
%15	9,817835	0,67101

Tablo 8. UV-visible Spektrum Datası Test Sonuçları

EKSİLME ORANI	RMSE DEĞERİ	R^2 TESTİ
%5	0,032126	0,60976
%10	0,041114	0,59855
%15	0,049852	0,57890

SONUÇLAR VE ÖNERİLER

Eksik veri kavramına baktığımızda kaliteli bilgi elde edebilmek için çözülmesi gereken bir sorun olduğunu görürüz. Gün geçtikçe gelişen ve ilerleyen teknoloji getirdiği kolaylıkların yanı sıra rekabet dezavantajına da neden olmaktadır. Bu çalışma kapsamında veri setlerindeki eksik nümerik verilerin tahmini için makine öğrenmesi yöntemlerine dayalı veri yönetim sistemi geliştirilmiştir. Çalışmada kullanılan veri setleri tamamen rastgele eksik (MCAR) bir şekilde %5, %10, % 15 oranında nümerik veriler azaltılmış, belirli oranlarda azaltılan bu eksik veri setleri geliştirilen algoritma ile test edilmiştir. Çalışmanın amacı kapsamında alanlarında uzman olmasa bile her araştırmacının rahatlıkla kullanabileceği bir arayüz oluşturulmuş ve veri setlerindeki eksik nümerik veriler %57 ile %79 arasında doğruluk oranı ile tahmin edilmiştir. Çalışmanın ileri aşamalarında sözel verilerin tahmini üzerinde durulacak, makine öğrenmesi ya da derin öğrenme gibi yapay zekâ kavramının sunduğu yöntemlerle hem sözel hem de sayısal verileri, bilgiyi daha net elde edebilmek adına kompakt sistemler üzerinde çalışılacaktır.

KAYNAKLAR

Abdella, M., & Marwala, T. (2005). The Use Of Genetıc Algorıthms And Neural Networks To Approxımate Mıssıng Data In Database. In *Computing and Informatics* (Vol. 24).

Abdella, M., & Marwala, T. (n.d.). The use of genetic algorithms and neural networks to approximate missing data in database. *IEEE 3rd International Conference on Computational Cybernetics, 2005. ICCC 2005.*, 207–212. https://doi.org/10.1109/ICCCYB.2005.1511574

Akif, M., Üniversitesi, E., Bilişim, Y., & Bölümü, S. (n.d.). *Veri Madenciliği Eda Coşlu*.

Alam, S., Ayub, M. S., Arora, S., & Khan, M. A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, *9*. https://doi.org/10.1016/j.dajour.2023.100341

Altukhova, O. (2020). Choice of method imputation missing values for obstetrics clinical data. *Procedia Computer Science*, *176*, 976–984. https://doi.org/10.1016/j.procs.2020.09.093

Altunkaynak, A., Başakın, E. E., & Kartal, E. (2020). Dalgacık K-EN Yakın Komşuluk Yöntemi İle Hava Kirliliği Tahmini. *Uludağ University Journal of The Faculty of Engineering*, 1547–1556. https://doi.org/10.17482/uumfd.809938

Arslan, F., Kahraman, H. T., Kelimeler, A., Zekâ, Y., Veri, E., Gürültü, T., Ve, T., Yapay, O., Ağları K-En, S., Komşular, Y., & Knn, S. (n.d.). *Yapay Zekâ Tabanlı Büyük Veri Yönetim Aracı*.

Atalay, M., & Çelik, E. (2017). Büyük Veri Analizinde Yapay Zekâ Ve Makine Öğrenmesi Uygulamaları - Artıfıcıal Intellıgence And Machıne Learnıng Applıcatıons In Bıg Data Analysıs. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 155–172. https://doi.org/10.20875/makusobed.309727

Awad, E. M., Association for Computing Machinery. Special Interest Group on Business Data Processing., & ACM Digital Library. (1990). *Proceedings of the 1990 ACM SIGBDP Conference on Trends and Directions in Expert Systems : October 31-November 2, 1990, Orlando, Florida*. Special Interest Data Group on Business Data Processing of the Association for Computing Machinery.

Aydin, S., Ustun, O., Ghosigharehaghaji, A., Tavaci, T., Yilmaz, A., & Yilmaz, M. (2022). Hydrothermal Synthesis of Nitrogen-Doped and Excitation-Dependent Carbon Quantum Dots for Selective Detection of Fe3+ in Blood Plasma. *Coatings*, *12*(9). https://doi.org/10.3390/coatings12091311

Bhattacharjee, A., Vishwakarma, G. K., Rajbongshi, B. K., & Tripathy, A. (2024). MIIPW: An R package for Generalized Estimating Equations with missing data integration using a combination of mean score and inverse probability weighted approaches and multiple imputation. *Expert Systems with Applications*, *238*. https://doi.org/10.1016/j.eswa.2023.121973

Bilgin, M. (n.d.). *Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi*.

Bucharest House Price Dataset. (2024, January 10). Https://Www.Kaggle.Com/Datasets/Denisadutca/Bucharest-House-Price-Dataset.

Celik, Y. (2013). *Comparison of Data Used For Loss Of Data Mining Methods*. https://www.researchgate.net/publication/348787393

Cheng, X., Xu, H., Peng, J., Wang, Z., Wu, A., & Li, S. (2017). *Exploration of classification methods: SVM and KDE*.

Coşar, M., & Deniz, E. (2021). Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1012986

Çüm, S., Demir, E. K., Gelbal, S., & Kışla, T. (2018). Kayıp Veriler Yerine Yaklaşık Değer Atamak için Kullanılan Gelişmiş Yöntemlerin Farklı Koşullar Altında Karşılaştırılması. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi*, *0*(45), 230–249. https://doi.org/10.21764/maeuefd.332605

Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi Cilt: 6 Sayı: 2 Aralık 2005. (n.d.).

Evren Şeker, Ş., Eşmekaya, E., Şehir Üniversitesi, İ., Bilişim Sistemleri Bölümü, Y., Demirel Üniversitesi, S., & Mühendisliği Bölümü, B. (2017). YBS Ansiklopedi Eksik Verilerin Tamamlanması (Imputation). In *Cilt* (Vol. 4). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/

Faktör, A., & İncelenmesi, A. E. (n.d.). *Çoklu Atama Yönteminde Değişimlenen Atama Sayısının*.

Gold Price Prediction Dataset. (2023, December 10). Https://Www.Kaggle.Com/Datasets/Sid321axn/Gold-Price-Prediction-Dataset.

Guo, J., Zan, X., Wang, L., Lei, L., Ou, C., & Bai, S. (2023). A random forest regression with Bayesian optimization-based method for fatigue strength prediction of ferrous alloys. *Engineering Fracture Mechanics*, *293*. https://doi.org/10.1016/j.engfracmech.2023.109714

Güner, N., Çomak, E., Üniversitesi, P., Fakültesi, M., & Bölümü, B. M. (n.d.). *Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi Predicting Performance of First Year Engineering Students in Calculus by Using Support Vector Machines*.

Harita Teknolojileri Elektronik Dergisi Cilt ; Kavzoğlu, T., Çölkesen, İ., & Kavzoğlu, T. (2010). Classification of Satellite Images Using Decision Trees: Kocaeli Case. *Electronic Journal of Map Technologies*, *2*(1), 36–45. www.tuik.gov.tr

Harman, G. (2021). Destek Vektör Makineleri ve Naive Bayes Sınıflandırma Algoritmalarını Kullanarak Diabetes Mellitus Tahmini. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.1041186

Hoffmann, G., Bietenbeck, A., Lichtinghagen, R., & Klawonn, F. (2018). Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *Journal of Laboratory and Precision Medicine*, *3*, 58–58. https://doi.org/10.21037/jlpm.2018.06.01

Holland, J. H. (1992). *This excerpt from Adaptation in Natural and Artificial Systems*.

Hyunshik Lee, E. R., & Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, *10*(3), 231–243.

Kahraman, H. T., Bayindir, R., & Sagiroglu, S. (2012). A new approach to predict the excitation current and parameter weightings of synchronous machines based on genetic

algorithm-based k-NN estimator. *Energy Conversion and Management*, *64*, 129–138. https://doi.org/10.1016/j.enconman.2012.05.004

Kalaycıoğlu, O. (2017). Rastlantısal Olmayan Kayıp Veri Varlığında Seçim Modelleri ile bir Duyarlılık Analizi Uygulaması. *İstatistikçiler Dergisi:İstatistik ve Aktüerya*, *10*(2), 76–85.

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S., & Raazi, S. M. K. U. R. (2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*, *2021*. https://doi.org/10.1155/2021/2553199

Kim, D. G., & Choi, J. Y. (2024). Efficient imputation of missing data using the information of local space defined by the geometric one-class classifier. *Expert Systems with Applications*, *242*. https://doi.org/10.1016/j.eswa.2023.122775

Kim, J., Kwak, Y., Mun, S. H., & Huh, J. H. (2023a). Imputation of missing values in residential building monitored data: Energy consumption, behavior, and environment information. *Building and Environment*, *245*. https://doi.org/10.1016/j.buildenv.2023.110919

Kim, J., Kwak, Y., Mun, S. H., & Huh, J. H. (2023b). Imputation of missing values in residential building monitored data: Energy consumption, behavior, and environment information. *Building and Environment*, *245*. https://doi.org/10.1016/j.buildenv.2023.110919

Korcan; ARSLANTEKİN, D. (2016). Büyük veri: önemi, yapısı ve günümüzdeki durum. *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi - DTCF Dergisi*, *56*(1), 15–36. https://doi.org/10.1501/dtcfder_0000001461

Leke, C. A., & Marwala, T. (2019). *Deep learning and missing data in engineering systems*. Springer.

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C. W., van den Bossche, P., Van Mierlo, J., & Omar, N. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. *Applied Energy*, *232*, 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Liu, W., Luo, L., & Zhou, L. (2023). Online missing value imputation for high-dimensional mixed-type data via generalized factor models. *Computational Statistics and Data Analysis*, *187*. https://doi.org/10.1016/j.csda.2023.107822

López, J. L., Hernández, S., Urrutia, A., López-Cortés, X. A., Araya, H., & Morales-Salinas, L. (2021). Effect of missing data on short time series and their application in the characterization of surface temperature by detrended fluctuation analysis. *Computers and Geosciences*, *153*. https://doi.org/10.1016/j.cageo.2021.104794

Malakouti, S. M. (2023). Improving the prediction of wind speed and power production of SCADA system with ensemble method and 10-fold cross-validation. *Case Studies in Chemical and Environmental Engineering*, *8*. https://doi.org/10.1016/j.cscee.2023.100351

Mustafa, D. R., & Demografi, B. V. E. (n.d.). *Arzu Baygül*.

Narin, A., İşler, Y., & Özer, M. (n.d.). *DEÜ Mühendislik Fakültesi Mühendislik Bilimleri Dergisi Konjestif Kalp Yetmezliği Teşhisinde Kullanılan Çapraz Doğrulama*

Yöntemlerinin Sınıflandırıcı Performanslarının Belirlenmesine Olan Etkilerinin Karşılaştırılması (Comparıson Of The Effects Of Cross Valıdatıon Methods On Determınıng Performances Of Classıfiers Used In Dıagnosıng Congestıve Heart Faılure). http://www.physionet.org

Oberman, H., Vink, G., & De Jong, V. (n.d.). *Iterative Imputation in Python*.

Oğuzlar, A. (2003). *VERİ ÖN İŞLEME* (Issue 21).

Orgaz, F., Testi, L., Villalobos, F. J., & Fereres, E. (2006). Water requirements of olive orchards-II: Determination of crop coefficients for irrigation scheduling. *Irrigation Science*, *24*(2), 77–84. https://doi.org/10.1007/s00271-005-0012-x

Osman Avşar Kurgun, D. (n.d.). *Veri Kalitesinde Eksik Veri Sorunlarının Derin Öğrenme Yöntemi İle Çözülmesi: Üretici Çekişmeli Ağlar İle Bir Uygulama Şevhat DOGER*.

Özarı, Ç., & Demirkale, Ö. (2022). K-En Yakın Komşu Algoritması İle Dolar-TL Ve Euro-TL Kuru Kullanarak Borsa Endeks Tahmini. *Maliye ve Finans Yazıları*, *117*, 41–62. https://doi.org/10.33203/mfy.1034155

Quinino, R. C., Reis, E. A., & Bessegato, L. F. (2012). *Using the coefficient of determination R 2 to test the significance of multiple linear regression*.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39. https://doi.org/10.1109/COMITCon.2019.8862451

RJA Little, D. R. (2019). *Statistical analysis with missing data*.

Seki, T., Hsiao, Y. Y., Ishizawa, F., Sugano, Y., & Takahashi, Y. (2023). Establishment of a random forest regression model to estimate the age of bloodstains based on temporal colorimetric analysis. *Legal Medicine*. https://doi.org/10.1016/j.legalmed.2023.102343

Sezgin, E., Çelik, Y., Üniversitesi, A., Bölüm Başkanlığı, E., MehmetBey Üniversitesi, K., & Myo, K. (n.d.). *Veri Madenciliğinde Kayıp Veriler Đçin Kullanılan Yöntemlerin Karşılaştırılması*.

Sheng, S., Guo, X., Yu, K., & Wu, X. (2024). Local causal structure learning with missing data[Formula presented]. *Expert Systems with Applications*, *238*. https://doi.org/10.1016/j.eswa.2023.121831

Shi, W., Zhu, Y., Zhang, J., Tao, X., Sheng, G., Lian, Y., Wang, G., & Chen, Y. (2015). Improving Power Grid Monitoring Data Quality: An Efficient Machine Learning Framework for Missing Data Prediction. *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, 417–422. https://doi.org/10.1109/HPCC-CSS-ICESS.2015.16

Sykes, A. O. (n.d.). *An Introduction to Regression Analysis*. https://chicagounbound.uchicago.edu/law_and_economics

Taşcı, E., & Onan, A. (2016). K-en yakın komşu algoritması parametrelerinin sınıflandırma performansı üzerine etkisinin incelenmesi. *Akademik Bilişim*, *1*(1), 4–18.

Testi, E., & Giorgetti, A. (2022). RSS-Based Localization of Multiple Radio Transmitters via Blind Source Separation. *IEEE Communications Letters*, *26*(3), 532–536. https://doi.org/10.1109/LCOMM.2021.3137598

Üniversitesi, F., & Mühendisliği, B. (2024). YBS Ansiklopedi Eksik Veri Probleminin Çözümü Solving The Missing Data Problem Uğur CAN. In *Cilt* (Vol. 12).

Villalobos, F. J., Testi, L., Hidalgo, J., Pastor, M., & Orgaz, F. (2006). Modelling potential growth and yield of olive (Olea europaea L.) canopies. *European Journal of Agronomy*, *24*(4), 296–303. https://doi.org/10.1016/j.eja.2005.10.008

Wan, Z. (2022). *Discover Factors Which Have Effects on Airbnb's Stakeholders by Using Python Using Sydney Airbnb as an Example*.

Wang, Y., Zhao, Y., Song, R., Zhao, Y., & Han, Q. (2024). Forest ecological long sequence missing data imputation method based on BiP-Informer. *Measurement: Journal of the International Measurement Confederation*, *225*. https://doi.org/10.1016/j.measurement.2023.113972

Yang, L. J. (2005). A test methodology for the determination of wear coefficient. *Wear*, *259*(7–12), 1453–1461. https://doi.org/10.1016/j.wear.2005.01.026

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı:	Ülgen AYDIN
Doğum tarihi:	
Doğum Yeri:	
Uyruğu:	
Adres:	
Tel:	
E-mail:	
Eğitim	
Lise:	Ordu Başöğretmen Anadolu Lisesi
Lisans:	Atatürk Üniversitesi, Endüstri Mühendisliği
Yüksek lisans:	Atatürk Üniversitesi, Endüstri mühendisliği Anabilim Dalı (2021)
Yabancı Dil Bilgisi	
İngilizce:	İyi
Üye Olunan Mesleki Kuruluşlar	
Tezden Üretilmiş Yayınlar	